Unable to locate subtitle

AWS Well-Architected Framework

Copyright © 2024 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Well-Architected Framework: ***Unable to locate subtitle***

Copyright © 2024 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Table of Contents

Résumé et introduction 1
Introduction 1
Définitions 2
Sur l'architecture 5
Principes généraux de conception 6
Les piliers du framework 8
Excellence opérationnelle 8
Principes de conception 9
Définition 10
Bonnes pratiques 11
Ressources 21
Sécurité 21
Principes de conception 21
Définition 22
Bonnes pratiques 23
Ressources 30
Fiabilité 31
Principes de conception 31
Définition 32
Bonnes pratiques 33
Ressources 38
Efficacité en matière de performance 39
Principes de conception 39
Définition 40
Bonnes pratiques 41
Ressources 47
Optimisation des coûts 47
Principes de conception 48
Définition 49
Bonnes pratiques 49
Ressources 56
Durabilité 56
Principes de conception 57
Définition 58

Bonnes pratiques 58
Processus de vérification 67
Conclusion 70
Participants 71
Autres lectures 72
Révisions du document 73
Annexe : questions et bonnes pratiques 77
Excellence opérationnelle 77
Organisation 77
Préparation 116
Exploiter 188
Évoluer 224
Sécurité 241
Bases du pilier Sécurité 242
Identity and Access Management 262
Détection 317
Protection de l'infrastructure 327
Protection des données 347
Réponse aux incidents 380
Sécurité des applications 403
Fiabilité 424
Fondations 424
Architecture de charge de travail 466
Gestion des modifications 513
Gestion des défaillances 554
Efficacité en matière de performance 662
Choix d'architecture 663
Informatique et matériel 677
Gestion des données 695
Mise en réseau et diffusion de contenu 722
Processus et culture 754
Optimisation des coûts 770
Pratiques en matière de gestion financière du cloud 770
Sensibilisation aux dépenses et à l'utilisation 795
Ressources rentables 840
Gérer la demande et les sources d'approvisionnement 881

Optimiser dans le temps 895
Durabilité 903
Choix de la région 904
Alignement sur la demande 906
Logiciels et architecture 920
Données 931
Matériel et services 952
Processus et culture 962
Mentions légales 971

AWS Well-Architected Framework

Date de publication : 3 octobre 2023 (Révisions du document)

AWS Well-Architected Framework vous aide à comprendre les avantages et les inconvénients des décisions que vous prenez lors du développement de systèmes sur AWS. En utilisant ce cadre, vous apprendrez les bonnes pratiques architecturales pour concevoir et exploiter des systèmes fiables, sécurisés, efficaces, économiques et durables dans le cloud.

Introduction

AWS Well-Architected Framework vous aide à comprendre les avantages et les inconvénients des décisions que vous prenez lors du développement de systèmes sur AWS. En utilisant ce cadre, vous apprenez les bonnes pratiques architecturales en matière de conception et d'exploitation de charges de travail fiables, sécurisées, efficaces, économiques et durables dans le AWS Cloud. Il vous permet d'évaluer systématiquement vos architectures par rapport aux bonnes pratiques et d'identifier les domaines à améliorer. Le processus d'examen d'une architecture repose sur une discussion constructive sur les décisions architecturales, et non un mécanisme d'audit. Nous pensons que l'adoption de systèmes Well-Architected augmente considérablement les chances de réussite d'une entreprise.

Les architectes de solutions AWS ont des années d'expérience en architecture de produits sur une très grande variété de segments verticaux commerciaux et de cas d'utilisation. Nous avons contribué à la conception et à la vérification de milliers d'architectures client sur AWS. Sur la base de cette expérience, nous avons identifié les bonnes pratiques et les principales stratégies d'architecture de systèmes dans le Cloud.

Well-Architected Framework AWS documente un ensemble de questions de base afin de vous aider à évaluer si une architecture spécifique respecte bien les bonnes pratiques du cloud. Le cadre offre une approche cohérente pour évaluer les systèmes par rapport aux qualités attendues des systèmes modernes basés sur le Cloud, ainsi que les corrections requises pour atteindre ces qualités. Comme AWS évolue en permanence et que nous ne cessons d'apprendre en collaborant avec nos clients, nous continuerons à affiner la définition de « well-architected ».

Le présent outil est conçu pour ceux qui sont occupent des postes technologiques, comme les directeurs de la technologie, les architectes, les développeurs et les membres de l'équipe d'exploitation. Il décrit les stratégies et les bonnes pratiques AWS à utiliser lors de la conception et de l'exécution d'une charge de travail dans le cloud, et fournit des liens vers d'autres détails de

mise en œuvre et modèles d'architecture. Pour en savoir plus, consultez Page d'accueil AWS Well-Architected.

AWS fournit également un service gratuit pour vérifier vos charges de travail. La AWS Well-Architected Tool (AWS WA Tool) est un service dans le cloud qui fournit un processus uniforme pour que vous puissiez vérifier et mesurer votre architecture à l'aide d'AWS Well-Architected Framework. AWS WA Tool offre des recommandations qui vous permettent de renforcer la fiabilité, la sécurité, l'efficacité et la rentabilité de vos charges de travail.

Pour vous aider à appliquer les bonnes pratiques, nous avons créé des ateliers AWS Well-Architected,qui vous fournissent un référentiel de code et de documentation afin de vous donner une expérience pratique de mise en œuvre des bonnes pratiques. Nous avons également fait équipe avec des partenaires du réseau de partenaires AWS (APN) triés sur le volet et eux-mêmes membres du programme de partenariat AWS Well-Architected. Ces partenaires AWS disposent de connaissances approfondies sur AWS et peuvent vous aider à vérifier et améliorer vos charges de travail.

Définitions

Chaque jour, les experts AWS aident les clients à concevoir des systèmes afin de tirer parti des bonnes pratiques dans le cloud. Nous collaborons avec vous pour parvenir à des compromis architecturaux au fur et à mesure que vos conceptions évoluent. Lorsque vous déployez ces systèmes dans des environnements réels, nous découvrons leurs performances réelles ainsi que les conséquences de ces compromis.

Grâce aux enseignements acquis, nous avons créé AWS Well-Architected Framework, qui fournit un ensemble cohérent de bonnes pratiques pour les clients et les partenaires, afin d'évaluer les architectures. Il inclut également un ensemble de questions dont vous pouvez vous inspirer pour évaluer le degré de conformité d'une architecture aux bonnes pratiques AWS.

Le cadre AWS Well-Architected Framework repose sur six piliers, à avoir l'Excellence opérationnelle, la Sécurité, la Fiabilité, l'Efficacité des performances, l'Optimisation des coûts et la Durabilité.

Tableau 1. Les piliers d'AWS Well-Architected Framework

Nom	Description
Excellence opérationnelle	Capacité de soutenir le développement et de gérer efficacement les charges de travail, de

Nom	Description
	recueillir des informations sur leurs opération s et d'améliorer continuellement les processus et procédures de soutien afin de fournir de la valeur ajoutée.
Sécurité	Le pilier Sécurité décrit comment tirer parti des technologies du cloud pour protéger les données, les systèmes et les ressources de manière à améliorer votre niveau de sécurité.
Fiabilité	Le pilier Fiabilité englobe la capacité d'une charge de travail à exécuter sa fonction de manière correcte et cohérente et ce, en temps utile. Cela inclut la possibilité d'exploiter et de tester la charge de travail tout au long de son cycle de vie. Ce livre blanc fournit des bonnes pratiques détaillées pour la mise en œuvre de charges de travail fiables sur AWS.
Efficacité en matière de performance	Capacité à utiliser efficacement les ressource s informatiques pour satisfaire aux exigences système et à maintenir cette efficacité au fur et à mesure que la demande change et que les technologies évoluent.
Optimisation des coûts	Possibilité d'exécuter des systèmes pour proposer une valeur commerciale au prix le plus bas.
Durabilité	Capacité d'améliorer continuellement les impacts sur la durabilité via la réduction de la consommation d'énergie et l'amélioration de l'efficacité de tous les composants d'une charge de travail en maximisant les avantages des ressources allouées et en minimisant les ressources totales requises.

Dans le cadre AWS Well-Architected Framework, nous utilisons les termes suivants :

- A Un composant représente le code, la configuration et les ressources AWS qui répondent ensemble à une exigence commerciale. Un composant est souvent une unité de propriété technique. Il est découplé des autres composants.
- Le terme scalable est utilisé pour désigner un ensemble de composants qui collaborent pour apporter une valeur métier. L'application représente généralement le niveau de détails dont discutent les responsables métier et techniques.
- Nous considérons l'architecture comme la façon dont les composants fonctionnent ensemble dans une charge de travail. Les schémas d'architecture se concentrent souvent sur la manière dont les composants communiquent et interagissent entre eux.
- Les Milestones (Événements) signalent les modifications importantes de votre architecture à mesure de son évolution tout au long du cycle de vie du produit (conception, mise en place, tests, pré-production et production).
- Au sein d'une entreprise, le portefeuille technologique est l'ensemble des charges de travail qui sont nécessaires au bon fonctionnement de l'entreprise.
- La niveau d'effort consiste à catégoriser le temps, les efforts et la complexité qu'une tâche nécessite pour sa mise en œuvre. Chaque organisation doit tenir compte de la taille et de l'expertise de l'équipe et de la complexité de la charge de travail comme contexte supplémentaire afin de catégoriser correctement son niveau d'effort.
  - Élevé : Le projet peut prendre plusieurs semaines, voire plusieurs mois. Il pourrait être divisé en plusieurs scénarios, versions et tâches.
  - Moyen : Le projet peut prendre plusieurs jours, voire plusieurs semaines. Il pourrait être divisé en plusieurs versions et tâches.
  - Faible : Le projet peut prendre plusieurs heures, voire plusieurs jours. Il pourrait être divisé en plusieurs tâches.

Lorsque vous concevez des charges de travail, vous faites des compromis entre des piliers en fonction de votre contexte commercial. Ces décisions métier peuvent vous aider à gérer vos priorités techniques. Vous pouvez opter pour l'optimisation afin d'améliorer l'impact sur la durabilité et de réduire les coûts au détriment de la fiabilité dans les environnements de développement, ou, pour les solutions stratégiques, vous pouvez optimiser la fiabilité avec des coûts plus élevés et un impact plus important sur la durabilité. Dans les solutions d'e-commerce, les performances peuvent affecter les revenus et la propension des clients à acheter les produits. La sécurité et l'excellence opérationnelle ne sont généralement pas la contrepartie des autres piliers.

Sur l'architecture

Dans les environnements sur site, les clients possèdent souvent une équipe centrale pour l'architecture technologique, qui supervise les autres équipes dédiées aux produits ou aux fonctionnalités afin de vérifier qu'elles ont adopté les bonnes pratiques. Les équipes d'architecture technologique comptent généralement un ensemble de rôles, notamment : architecte technique (infrastructure), architecte de solutions (logiciel), architecte de données, architecte de mise en réseau et architecte de sécurité. Souvent, ces équipes utilisent TOGAF ou le cadre Zachman en tant qu'élément de la capacité d'architecture de l'entreprise.

Chez AWS, nous préférons répartir les fonctionnalités entre plusieurs équipes dédiées plutôt que de ne recourir qu'à une seule équipe centralisée pour toutes les fonctionnalités. Il existe des risques lorsque vous choisissez de répartir les décisions. Par exemple, il faut vérifier que les équipes respectent les normes internes. Nous réduisons ces risques de deux manières. Tout d'abord, nous avons les pratiques (façons de procéder, processus, règles et autres normes acceptées) qui visent à permettre à chaque équipe de se charger de cette fonctionnalité comme il se doit, et nous avons recours à des experts afin de vérifier que les équipes dépassent les exigences. Ensuite, nous mettons en place des mécanismes qui effectuent des vérifications automatisées pour vérifier que les normes sont respectées.

> « Les bonnes intentions ne suffisent pas, il faut de bons mécanismes pour tout rendre possible », Jeff Bezos.

Cela implique d'avoir recours à des mécanismes (souvent automatisés) qui vérifient la conformité aux règles ou aux processus plutôt que de faire appel à la bonne volonté des employés. Cette approche distribuée s'appuie sur les principes de leadership d'Amazonet établit une culture pour tous les rôles qui travaillent à rebours à partir du client. Le travail à rebours est un élément fondamental de notre processus d'innovation. Nous commençons par les clients et ce dont ils ont besoin, afin de définir et de guider nos efforts. Les équipes obsédées par le client fabriquent des produits en s'appuyant sur les besoins des clients.

Pour l'architecture, cela signifie que nous prévoyons que chaque équipe soit capable de créer des architectures et de suivre les bonnes pratiques. Pour aider les nouvelles équipes à s'approprier ces fonctionnalités, ou les équipes existantes à mettre la barre plus haut, nous activons l'accès à une communauté virtuelle d'ingénieurs en chef qui peuvent vérifier leurs conceptions et les aider à comprendre ce que sont les bonnes pratiques AWS. La communauté des ingénieurs principaux

travaille pour rendre les bonnes pratiques visibles et accessibles. Une solution pourrait être, par exemple, d'organiser des discussions durant le déjeuner qui se concentrent sur l'application de bonnes pratiques à de véritables exemples. Ces discussions sont enregistrées et peuvent être utilisées dans le cadre de documents d'accueil pour les nouveaux membres de l'équipe.

Les bonnes pratiques AWS naissent de notre expérience dans l'exécution de milliers de systèmes à l'échelle d'Internet. Nous préférons utiliser des données pour définir les bonnes pratiques, mais utilisons également des experts fonctionnels en tant qu'ingénieurs principaux pour les définir. Lorsque les ingénieurs en chef voient l'émergence de nouvelles bonnes pratiques, ils collaborent afin de vérifier que les équipes les suivent. Avec le temps, ces bonnes pratiques sont officialisées dans nos processus d'évaluation internes, ainsi que dans les mécanismes qui renforcent la conformité. Le cadre Well-Architected est l'implémentation destinée aux clients de notre processus d'évaluation interne, où nous avons codifié notre réflexion sur l'ingénierie principale à travers les rôles tels que l'architecture de solutions et les équipes d'ingénieurs internes. Le cadre Well-Architected est un mécanisme évolutif qui vous permet de tirer parti de ces connaissances.

En suivant l'approche d'une communauté d'ingénierie principale avec une propriété d'architecture distribuée, nous pensons qu'une architecture d'entreprise Well-Architected reposant sur les besoins du client peut émerger. Nos leaders en matière de technologie (tels que les directeurs techniques ou les directeurs de développement), qui effectuent des évaluations Well-Architected sur toutes vos charges de travail, vous permettront de mieux comprendre les risques au sein de votre portefeuille de technologies. Cette approche vous permet d'identifier les thèmes, pour différentes équipes, que votre organisation pourrait aborder via les mécanismes, les formations, ou les discussions de midi où vos ingénieurs principaux peuvent partager leurs idées sur les domaines spécifiques avec plusieurs équipes.

Principes généraux de conception

Le cadre Well-Architected identifie un ensemble de principes généraux de conception destinés à faciliter la bonne conception dans le Cloud :

- Une capacité réellement adaptée à vos besoins : si vous prenez une mauvaise décision en matière de capacité lors du déploiement d'une charge de travail, il se peut que vous vous retrouviez face à des ressources inutilisées onéreuses, ou que vous deviez traiter les implications relatives aux performances d'une capacité limitée. Grâce au cloud computing, vous n'avez plus de soucis à vous faire. Vous pouvez utiliser autant ou aussi peu de capacité que vous le souhaitez, et l'augmenter ou la réduire automatiquement.

- Tester les systèmes à l'échelle de la production : dans le cloud, vous pouvez créer un environnement de tests à l'échelle de la production et à la demande, exécuter les tests, puis mettre les ressources hors service. Puisque vous ne payez l'environnement de test que lorsqu'il s'exécute, vous pouvez simuler votre environnement réel pour une fraction du coût d'un test sur site.
- Automatiser en gardant à l'esprit l'expérimentation architecturale: l'automatisation vous permet de créer et de répliquer vos charges de travail à un coût peu élevé et d'éviter les frais de main-d'œuvre. Vous pouvez suivre les modifications apportées à l'automatisation, auditer l'impact et rétablir les paramètres antérieurs si nécessaire.
- Tenir compte des architectures évolutives : dans un environnement traditionnel, les décisions d'architecture sont souvent mises en place comme des événements statiques et fixes, avec quelques versions majeures d'un système pendant sa durée de vie. Tandis que l'entreprise et son contexte continuent à évoluer, ces décisions initiales peuvent entraver la capacité du système à satisfaire des exigences métier variables. Dans le cloud, la capacité d'automatiser et de tester les éléments à la demande réduit le risque d'impact des modifications de conception. Les systèmes peuvent ainsi évoluer au fil du temps, de telle sorte que les entreprises peuvent tirer profit des innovations dans le cadre d'une pratique standard.
- Créer des architectures basées sur des données : dans le cloud, vous pouvez collecter des données sur la façon dont vos choix architecturaux affectent le comportement de votre charge de travail. Cela vous permet de prendre des décisions basées sur les faits sur la façon d'améliorer votre charge de travail. Votre infrastructure Cloud est codée. Vous pouvez donc utiliser ces données pour alimenter vos choix architecturaux et des améliorations au fil du temps.
- Améliorer les systèmes grâce aux journées jeu de rôle : testez les performances de votre architecture et de vos processus en programmant régulièrement des tests de simulation de pannes, pour simuler des événements durant la production. Cela vous aidera à comprendre où apporter des améliorations et à développer une expérience de gestion des événements au sein de votre organisation.

Les piliers du framework

La création d'un logiciel est similaire à celle d'un bâtiment. Si la fondation n'est pas solide, des problèmes structurels peuvent saper l'intégrité et la fonction du bâtiment. Lorsque vous concevez des solutions technologiques, si vous négligez les six piliers de l'excellence opérationnelle, à savoir la sécurité, la fiabilité, l'efficacité des performances, l'optimisation des coûts et la durabilité, il peut s'avérer difficile de créer un système qui répond à vos attentes et à vos exigences. Le fait d'intégrer ces domaines à votre architecture vous aidera à produire des systèmes stables et efficaces. Vous pouvez ainsi vous concentrer sur d'autres aspects de la conception, tels que les exigences fonctionnelles.

Piliers

- Excellence opérationnelle
- Sécurité
- Fiabilité
- Efficacité en matière de performance
- Optimisation des coûts
- Durabilité

Excellence opérationnelle

Le pilier Excellence opérationnelle comprend la capacité à soutenir le développement et à gérer efficacement les charges de travail, à recueillir des informations sur leurs opérations et à améliorer continuellement les processus et procédures de soutien afin de fournir de la valeur métier.

Le pilier Excellence opérationnelle fournit une vue d'ensemble des principes de conception, des bonnes pratiques et des questions. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Excellence opérationnelle.

Rubriques

- Principes de conception
- Définition
- Bonnes pratiques
- Ressources

Principes de conception

Voici les principes de conception pour l’excellence opérationnelle dans le cloud :

- Exécuter les opérations sous la forme de code : dans le cloud, vous pouvez appliquer la même discipline d'ingénierie que celle que vous utilisez pour le code d'application dans l'ensemble de l'environnement. Vous pouvez définir l'ensemble de votre charge de travail (applications, infrastructure, etc.) en tant que code et la mettre à jour avec du code. Vous pouvez créer des scripts pour vos procédures opératoires et automatiser leur exécution en les lançant en réponse à des événements. En effectuant les opérations en tant que code, vous limitez les erreurs humaines et créez des réponses cohérentes aux événements.
- Effectuer des modifications fréquentes, mineures et réversibles : concevez des charges de travail évolutives et couplées faiblement pour permettre la mise à jour régulière des composants. Les techniques de déploiement automatisé associées à des modifications mineures et incrémentielles réduisent le rayon d'impact et permettent de faire marche arrière plus rapidement en cas de problème. Cela renforce la confiance dans la possibilité d'apporter des modifications positives à votre charge de travail tout en maintenant la qualité et en s'adaptant rapidement à l'évolution des conditions du marché.
- Affiner régulièrement les procédures opérationnelles : au fur et à mesure que vos charges de travail évoluent, faites évoluer vos opérations en conséquence. Tout en utilisant des procédures opérationnelles, cherchez le moyen de les améliorer. Passez régulièrement en revue les procédures et assurez-vous qu'elles sont efficaces et maîtrisées par les équipes. Lorsque des lacunes sont identifiées, actualisez les procédures en conséquence. Communiquez les mises à jour des procédures à toutes les parties prenantes et équipes. Transformez vos opérations en jeu pour partager les bonnes pratiques et former les équipes.
- Anticiper les pannes : par exemple, effectuez des exercices « pre-mortem » afin d'identifier les causes possibles de défaillances, et ainsi les éliminer ou les atténuer. Testez vos scénarios de pannes et confirmez votre compréhension de leur impact. Testez vos procédures de réponse pour vous assurer qu'elles sont efficaces et que les équipes sont familiarisées avec leur exécution. Planifiez des simulations de pannes pour tester les réponses des charges de travail et de l'équipe face à des événements simulés.
- Tirer des leçons de toutes les pannes opérationnelles : visez l'amélioration grâce aux leçons apprises de tous les événements et pannes liés aux opérations. Communiquez ce qui a été appris aux équipes et à l'ensemble de l'entreprise.

- Utiliser des services gérés : réduisez la charge opérationnelle en utilisant des services AWS gérés dans la mesure du possible. Élaborez des procédures opérationnelles autour des interactions avec ces services.
- Mettre en œuvre l'observabilité pour obtenir des informations exploitables : faites-vous une idée précise du comportement, des performances, de la fiabilité, des coûts et de l'état de la charge de travail. Établissez des indicateurs de performance clés (KPI) et tirez parti de la télémétrie de l'observabilité pour prendre des décisions éclairées et agir rapidement lorsque les résultats de l'entreprise sont menacés. Améliorez de manière proactive les performances, la fiabilité et les coûts sur la base de données d'observabilité exploitables.

Définition

Il existe quatre domaines de bonnes pratiques pour l'excellence opérationnelle dans le cloud :

- Organisation
- Préparation
- Exploiter
- Évolution

La direction de votre organisation définit les objectifs opérationnels. Votre organisation doit comprendre les besoins et les priorités et les utiliser pour organiser et mener des travaux visant à soutenir l'obtention des résultats opérationnels. Votre charge de travail doit émettre les informations nécessaires pour la prendre en charge. La mise en œuvre de services permettant l'intégration, le déploiement et la distribution de votre charge de travail générera un flux accru de changements bénéfiques dans la production en automatisant les processus répétitifs.

Il peut exister des risques inhérents à l'exploitation de votre charge de travail. Vous devez comprendre ces risques et prendre une décision avisée lors de la mise en production. Vos équipes doivent pouvoir soutenir votre charge de travail. Les métriques économiques et opérationnelles dérivées des résultats économiques souhaités vous permettront de comprendre l'état de votre charge de travail et de vos activités opérationnelles, et de réagir aux incidents. Vos priorités évolueront en fonction des besoins de votre entreprise et des changements dans l'environnement de votre entreprise. Utilisez-les comme une boucle de rétroaction afin d'améliorer continuellement votre organisation et le fonctionnement de votre charge de travail.

Bonnes pratiques

Rubriques

- Organisation
- Préparer
- Exploiter
- Évolution

Organisation

Vos équipes doivent avoir une compréhension commune de l'ensemble de votre charge de travail, de leur rôle dans celle-ci et de leurs objectifs économiques communs afin de fixer les priorités qui permettent la réussite de l'entreprise. Des priorités bien définies maximiseront les bénéfices tirés de vos efforts. Évaluer les besoins des clients internes et externes en impliquant les principales parties prenantes, notamment les équipes commerciales, de développement et d'exploitation, afin de déterminer où il est nécessaire de concentrer les efforts. L'évaluation des besoins des clients vous permet de vous assurer que vous avez une compréhension approfondie du soutien nécessaire pour atteindre les résultats économiques. Assurez-vous de connaître les lignes directrices ou les obligations définies par la gouvernance de votre entreprise, ainsi que les facteurs externes, tels que les exigences de conformité réglementaire et les normes sectorielles, qui peuvent imposer un objectif spécifique ou mettre l'accent sur ce dernier. Vérifiez que vous disposez de mécanismes permettant d'identifier les changements apportés à la gouvernance interne et aux exigences de conformité externe. Si aucune exigence n'est identifiée, assurez-vous d'avoir effectué les vérifications préalables dans cette détermination. Revoyez régulièrement vos priorités afin qu'elles puissent être mises à jour en fonction de l'évolution des besoins.

Évaluez les menaces pesant sur l'entreprise (par exemple, les risques et les responsabilités de l'entreprise, et les menaces sur la sécurité des informations) et conservez ces informations dans un registre des risques. Évaluez l'impact des risques et les compromis entre des intérêts concurrents ou des approches alternatives. Par exemple, l'accélération de la mise sur le marché de nouvelles fonctionnalités peut être privilégiée par rapport à l'optimisation des coûts, ou vous pouvez choisir une base de données relationnelle pour les données non relationnelles afin de simplifier la migration d'un système sans restructuration. Gérez les avantages et les risques afin de prendre des décisions éclairées lorsqu'il s'agit de déterminer où il est nécessaire de concentrer les efforts. Certains risques ou choix peuvent être acceptables pendant un certain temps, il peut être possible d'atténuer les

risques associés, ou il peut devenir inacceptable de laisser un risque subsister, auquel cas vous prendrez des mesures pour y remédier.

Vos équipes doivent comprendre leur rôle dans l'obtention des résultats de l'entreprise. Les équipes doivent comprendre leur rôle dans la réussite des autres équipes, le rôle des autres équipes dans leur réussite, et avoir des objectifs communs. Comprendre la responsabilité, la manière dont les décisions sont prises et qui a le pouvoir de prendre des décisions vous aide à concentrer les efforts et à maximiser les avantages de vos équipes. Les besoins d'une équipe seront déterminés par le client qu'elle soutient, son organisation, la composition de l'équipe et les caractéristiques de sa charge de travail. Il n'est pas raisonnable de s'attendre à ce qu'un modèle d'exploitation unique puisse soutenir toutes les équipes et leurs charges de travail dans votre entreprise.

Assurez-vous qu'il existe des propriétaires identifiés pour chaque application, charge de travail, plateforme et composant d'infrastructure, et que chaque processus et procédure a un propriétaire identifié responsable de sa définition, et des propriétaires responsables de leur performance.

La compréhension de la valeur ajoutée de chaque composant, processus et procédure, de la raison pour laquelle ces ressources sont en place ou ces activités exécutées, et de la raison pour laquelle cette propriété existe, éclaire les actions des membres de votre équipe. Définissez clairement les responsabilités des membres de l'équipe afin qu'ils puissent agir de manière appropriée et disposer de mécanismes permettant d'identifier la responsabilité et la propriété. Mettez en œuvre des mécanismes permettant de demander des ajouts, des modifications et des exceptions afin de ne pas entraver l'innovation. Définissez des accords entre les équipes décrivant la manière dont elles travaillent ensemble pour se soutenir mutuellement et soutenir les résultats de votre entreprise.

Fournissez un soutien aux membres de votre équipe afin qu'ils puissent être plus efficaces dans leur action et soutenir les résultats de votre entreprise. Les dirigeants engagés doivent fixer des attentes et mesurer le succès. Les principaux dirigeants devraient être le parrain, l'avocat et le moteur de l'adoption des bonnes pratiques et de l'évolution de l'organisation. Permettez aux membres de l'équipe d'agir lorsque les résultats sont menacés afin de minimiser l'impact et de les encourager à remonter jusqu'aux décideurs et aux parties prenantes lorsqu'ils estiment qu'il existe un risque afin de pouvoir le traiter et éviter les incidents. Fournissez en temps utile des communications claires et exploitables sur les risques connus et les événements prévus afin que les membres de l'équipe puissent prendre des mesures appropriées en temps opportun.

Encouragez l'expérimentation pour accélérer la formation et maintenir l'intérêt et l'engagement des membres de l'équipe. Les équipes doivent développer leurs compétences pour adopter les nouvelles technologies, et pour soutenir l'évolution des besoins et des responsabilités. Soutenez et

encouragez cette démarche en accordant du temps structurel à la formation. Assurez-vous que les membres de votre équipe disposent des ressources, à la fois des outils et des membres de l'équipe, nécessaires à la réussite et à la mise à l'échelle pour soutenir les résultats de l'entreprise. Exploitez la diversité inter-organisationnelle pour rechercher des perspectives multiples et uniques. Utilisez cette perspective pour accroître l'innovation, remettre en question vos hypothèses et réduire le risque de biais de confirmation. Développez l'inclusion, la diversité et l'accessibilité au sein de vos équipes afin d'obtenir des perspectives bénéfiques.

Si des exigences réglementaires ou de conformité externes s'appliquent à votre organisation, vous devez utiliser les ressources fournies par AWS Cloud Compliance pour former vos équipes afin qu'elles puissent déterminer l'impact sur vos priorités. Le cadre Well-Architected met l'accent sur la formation, la mesure et l'amélioration. Il offre une approche cohérente pour évaluer les architectures et mettre en œuvre des conceptions qui seront mises à l'échelle dans le temps. AWS fournit l'outil AWS Well-Architected Tool pour vous aider à vérifier votre approche avant le développement et l'état de vos charges de travail avant et pendant la production. Vous pouvez comparer les charges de travail aux bonnes pratiques architecturales AWS les plus récentes, surveiller leur état global et obtenir des informations sur les risques potentiels. AWS Trusted Advisor est un outil qui permet d'accéder à un ensemble de vérifications de base qui recommandent des optimisations pouvant vous aider à définir vos priorités. Les clients du Business and Enterprise Support ont accès à des contrôles supplémentaires axés sur la sécurité, la fiabilité, les performances, l'optimisation des coûts et la durabilité qui peuvent les aider à définir leurs priorités.

AWS peut vous aider à former vos équipes à AWS et à ses services afin qu'elles comprennent mieux comment leurs choix peuvent avoir un impact sur votre charge de travail. Utilisez les ressources fournies par AWS Support (Centre de connaissances AWS, forums de discussion AWSet AWS Support Center) et la documentation AWS pour former vos équipes. Contactez AWS Support via AWS Support Center pour obtenir des réponses à vos questions AWS. AWS partage également les bonnes pratiques et les modèles que nous avons appris grâce à l'exploitation d'AWS dans Amazon Builders' Library. Un grand nombre d'autres informations utiles sont disponibles sur le blog AWS et sur le podcast officiel AWS. AWS Training and Certification offre une formation par le biais de cours en ligne d'auto-formation sur les principes fondamentaux d'AWS. Vous pouvez également vous inscrire à une formation dirigée par un formateur afin de soutenir le développement des compétences AWS de vos équipes.

Utilisez des outils ou des services qui permettent de gérer de manière centralisée vos environnements dans plusieurs comptes, comme AWS Organizations, pour gérer vos modèles d'exploitation. Des services tels que AWS Control Tower élargissent cette fonctionnalité de gestion en vous permettant de définir des plans (soutenant vos modèles d'exploitation) pour la configuration

des comptes, d'appliquer une gouvernance continue en utilisant AWS Organizations et d'automatiser l'allocation de nouveaux comptes. Les fournisseurs de services gérés tels que AWS Managed Services, les partenaires AWS Managed Services ou les fournisseurs de services gérés du réseau de partenaires AWS offrent une expertise dans la mise en œuvre des environnements cloud et soutiennent vos exigences de sécurité et de conformité, ainsi que vos objectifs métier. L'ajout de services gérés à votre modèle d'exploitation peut vous faire gagner du temps et économiser des ressources, et maintenir vos équipes internes réduites et concentrées sur les résultats stratégiques qui différencieront votre entreprise, plutôt que de développer de nouvelles compétences et capacités.

Les questions suivantes sont axées sur ces quelques considérations relatives à l'excellence opérationnelle. (Pour obtenir la liste des questions et des bonnes pratiques d'excellence opérationnelle, consultez l' Appendix.)

OPS 1: How do you determine what your priorities are?

Everyone must understand their part in achieving business success. Have shared goals in order to set priorities for resources. This will maximize the benefits of your efforts.

OPS 2: How do you structure your organization to support your business outcomes?

Your teams must understand their part in achieving business outcomes. Teams must understand their roles in the success of other teams, the role of other teams in their success, and have shared goals. Understanding responsibility, ownership, how decisions are made, and who has authority to make decisions will help focus efforts and maximize the benefits from your teams.

OPS 3: How does your organizational culture support your business outcomes?

Provide support for your team members so that they can be more effective in taking action and supporting your business outcome.

Vous pouvez décider à un moment donné de mettre l'accent sur un petit sous-ensemble de priorités opérationnelles. Utilisez une approche équilibrée sur le long terme pour garantir le développement des capacités nécessaires et de la gestion des risques. Vérifiez régulièrement les priorités opérationnelles et mettez-les à jour en fonction de l'évolution de vos besoins. Lorsque la

responsabilité et la propriété sont indéfinies ou inconnues, vous risquez à la fois de ne pas effectuer les actions nécessaires en temps utile et de déployer des efforts redondants et potentiellement conflictuels pour répondre à ces besoins. La culture organisationnelle a un impact direct sur la satisfaction professionnelle et la fidélisation des membres de l'équipe. Stimulez l'engagement et l'exploitation des capacités des membres de votre équipe pour assurer la réussite de votre entreprise. L'expérimentation est nécessaire pour que l'innovation se produise et transforme les idées en résultats. Admettez qu'un résultat non désiré est une expérience positive qui a identifié un chemin qui ne mène pas au succès.

Préparer

Pour vous préparer à l'excellence opérationnelle, il est nécessaire de comprendre vos charges de travail et les comportements attendus. Vous pourrez ensuite les concevoir pour fournir des informations sur leur statut et créer les procédures nécessaires pour les prendre en charge.

Concevez votre charge de travail de manière à ce qu'elle vous fournisse les informations nécessaires pour comprendre son état interne (par exemple, les mesures, les journaux, les événements et les traces) dans tous ses composants à des fins d'observation et de résolution des problèmes. L'observabilité va au-delà de la simple surveillance. Elle fournit une compréhension complète du fonctionnement interne d'un système sur la base de ses résultats externes. Enracinée dans les métriques, les journaux et les données de suivi, l'observabilité propose des informations approfondies sur le comportement et la dynamique du système. Grâce à une observabilité efficace, les équipes peuvent identifier les modèles, les anomalies et les tendances, ce qui leur permet de résoudre les problèmes potentiels de manière proactive et de maintenir un état optimal du système. L'identification des indicateurs clés de performance (KPI) est essentielle pour garantir l'alignement entre les activités de surveillance et les objectifs commerciaux. Cet alignement garantit que les équipes prennent des décisions basées sur les données en utilisant des indicateurs réellement importants, optimisant à la fois les performances du système et les résultats commerciaux. En outre, l'observabilité permet aux entreprises d'être proactives plutôt que réactives. Les équipes peuvent comprendre les relations de cause à effet au sein de leurs systèmes, prévoir et prévenir les problèmes au lieu de simplement y réagir. À mesure que les charges de travail évoluent, il est essentiel de revoir et d'affiner la stratégie d'observabilité, afin de s'assurer qu'elle reste pertinente et efficace.

Adoptez des approches qui améliorent le flux des changements en production et qui permettent la restructuration, un retour d'information rapide sur la qualité et la correction des bugs. Ces approches accélèrent l'entrée des modifications bénéfiques dans l'environnement de production, limitent les problèmes déployés et permettent d'identifier et de corriger rapidement les problèmes introduits par les activités de déploiement ou découverts dans vos environnements.

Adoptez des approches qui fournissent un retour d'information rapide sur la qualité et permettent une reprise rapide à la suite de changements qui n'offrent pas les résultats escomptés. L'utilisation de ces pratiques diminue l'impact des problèmes découlant du déploiement des modifications. Prévoyez les modifications qui échouent afin de pouvoir réagir plus rapidement si nécessaire, et testez et validez les changements que vous apportez. Tenez compte des activités planifiées dans vos environnements afin de pouvoir gérer le risque des modifications affectant les activités planifiées. Mettez l'accent sur les modifications fréquentes, minimes et réversibles pour limiter leur portée. Ainsi, vous facilitez la résolution des problèmes et les corrections avec la possibilité d'annuler une modification. Cela signifie également que vous pouvez tirer profit plus souvent de modifications importantes.

Évaluez l'état de préparation opérationnelle de votre charge de travail, de vos processus, de vos procédures et de votre personnel afin de comprendre les risques opérationnels liés à votre charge de travail. Utilisez un processus cohérent (y compris des listes de contrôle manuelles ou automatisées) pour déterminer quand vous êtes prêt à mettre en service votre charge de travail ou un changement. Cela vous permet également d'identifier tous les domaines d'amélioration nécessaire. Dotez-vous de runbooks qui documentent vos activités de routine, et de playbooks qui guident vos processus pour la résolution des problèmes.. Déterminez les avantages et les risques afin de prendre des décisions éclairées pour autoriser les changements dans l'environnement de production.

AWS vous permet de visualiser l'ensemble de votre charge de travail (applications, infrastructure, politique, gouvernance et opérations) en tant que code. Cela signifie que vous pouvez appliquer la même discipline d'ingénierie que celle que vous utilisez pour le code d'application à chaque élément de votre pile et partager ces éléments entre les équipes ou les organisations afin d'amplifier les avantages des efforts de développement. Utilisez les opérations en tant que code dans le cloud et testez-les en toute sécurité pour développer votre charge de travail, vos procédures d'opérations et la pratique de l'échec. L'utilisation de AWS CloudFormation vous permet de disposer d'environnements de développement, de test et de production cohérents et modélisés, avec des niveaux de contrôle des opérations toujours plus élevés.

Les questions suivantes sont axées sur ces quelques considérations relatives à l'excellence opérationnelle.

OPS 4: How do you implement observability in your workload?

Implement observability in your workload so that you can understand its state and make data-driv en decisions based on business requirements.

OPS 5: How do you reduce defects, ease remediation, and improve flow into production?

Adopt approaches that improve flow of changes into production that achieve refactoring fast feedback on quality, and bug fixing. These accelerate beneficial changes entering production, limit issues deployed, and achieve rapid identification and remediation of issues introduced through deployment activities.

OPS 6: How do you mitigate deployment risks?

Adopt approaches that provide fast feedback on quality and achieve rapid recovery from changes that do not have desired outcomes. Using these practices mitigates the impact of issues introduce d through the deployment of changes.

OPS 7: How do you know that you are ready to support a workload?

Evaluate the operational readiness of your workload, processes and procedures, and personnel to understand the operational risks related to your workload.

Investissez dans la mise en œuvre des activités opérationnelles en tant que code pour maximiser la productivité du personnel opérationnel, minimiser les taux d'erreur et automatiser les réponses. Adoptez des « pre-mortems » pour anticiper les échecs, et créez des procédures si nécessaire. Appliquez des métadonnées à l'aide des balises de ressource et de AWS Resource Groups en suivant une stratégie de balisage cohérente pour permettre l'identification de vos ressources. Balisez vos ressources pour l'organisation, la comptabilité analytique, les contrôles d'accès et le ciblage de l'exécution des activités d'opérations automatisées. Adoptez des pratiques de déploiement qui tirent parti de l'élasticité du cloud pour faciliter les activités de développement, et le pré-déploiement des systèmes pour accélérer les mises en œuvre. Lorsque vous apportez des modifications aux listes de contrôle que vous utilisez pour évaluer votre charge de travail, planifiez les opérations que vous allez exécuter pour les systèmes en service qui ne sont plus conformes.

Exploiter

L'observabilité vous permet de vous concentrer sur les données pertinentes et de comprendre les interactions et les résultats de votre charge de travail. En vous concentrant sur les informations

essentielles et en éliminant les données inutiles, vous maintenez une approche simple pour comprendre les performances des charges de travail. Il est essentiel non seulement de collecter des données, mais également de les interpréter correctement. Définissez des bases de référence claires, spécifiez des seuils d'alerte appropriés et surveillez activement tout écart. Un changement au niveau d'une métrique clé, en particulier lorsqu'elle est corrélée à d'autres données, contribue à identifier des problèmes spécifiques. Grâce à l'observabilité, vous êtes mieux équipé pour prévoir et relever les défis potentiels, veillant ainsi à ce que votre charge de travail fonctionne sans heurts et réponde aux besoins de l'entreprise.

Le bon fonctionnement d'une charge de travail se mesure à l'aune des résultats obtenus par les entreprises et les clients. Définissez les résultats attendus, déterminez comment le succès sera mesuré et identifiez les paramètres qui seront utilisés dans ces calculs pour déterminer le succès de votre charge de travail et des opérations. L'état opérationnel comprend à la fois l'état de la charge de travail et l'état et le succès des activités opérationnelles menées pour soutenir la charge de travail (par exemple, déploiement et réponse aux incidents). Établissez des métriques de référence pour l'amélioration, l'investigation et l'intervention, collectez et analysez vos métriques, puis validez votre compréhension du succès des opérations et de leur évolution dans le temps. Utilisez les métriques collectées pour déterminer si vous satisfaites vos clients et vos besoins commerciaux, et pour identifier les points à améliorer.

Une efficacité opérationnelle et une gestion efficace des événements sont requises pour atteindre une excellence opérationnelle. Cela s'applique à la fois aux événements opérationnels planifiés et imprévus. Utilisez les runbooks établis pour les événements bien compris, et utilisez les playbooks pour faciliter l'investigation et la résolution des problèmes. Prioriser les réponses aux événements en fonction de leur impact sur l'entreprise et les clients. Assurez-vous que, si une alerte est générée en réponse à un événement, il existe un processus associé à exécuter, avec un propriétaire spécifiquement identifié. Définissez à l'avance le personnel requis pour résoudre un événement et inclure des processus de remontée pour engager du personnel supplémentaire, si nécessaire, en fonction de l'urgence et de l'impact. Identifiez et engagez des personnes habilitées à prendre une décision sur les mesures à prendre lorsqu'une réponse à un événement non traité auparavant a un impact opérationnel.

Communiquez l'état opérationnel des charges de travail au moyen de tableaux de bord et de notifications adaptés au public cible (par exemple, clients, entreprises, développeurs, opérations) afin qu'il puisse prendre les mesures appropriées, que leurs attentes soient gérées et qu'il soit informé lorsque les opérations normales reprennent.

Dans AWS, vous pouvez générer des vues de tableau de bord de vos métriques collectées à partir des charges de travail et nativement depuis AWS. Vous pouvez tirer profit de CloudWatch ou d'applications tierces pour regrouper et présenter des perspectives d'opérations au niveau de l'entreprise, de la charge de travail ou des opérations. AWS fournit des informations de charges de travail par le biais de fonctionnalités de journalisation, notamment AWS X-Ray, CloudWatch, CloudTrail et les journaux de flux VPC pour identifier les problèmes de charges de travail en soutien à l'analyse des causes racines et à la résolution.

Les questions suivantes sont axées sur ces quelques considérations relatives à l'excellence opérationnelle.

OPS 8: How do you utilize workload observability in your organization?
Ensure optimal workload health by leveraging observability. Utilize relevant metrics, logs, and traces to gain a comprehensive view of your workload's performance and address issues efficient ly.

OPS 9: How do you understand the health of your operations?
Define, capture, and analyze operations metrics to gain visibility to operations events so that you can take appropriate action.

OPS 10: How do you manage workload and operations events?
Prepare and validate procedures for responding to events to minimize their disruption to your workload.

Toutes les métriques que vous recueillez doivent être alignées sur un besoin métier et les résultats qu'elles prennent en charge. Développez des réponses scriptées aux événements bien compris et automatisez leur exécution en réponse à la reconnaissance de l'événement.

Évolution

Apprenez, partagez et progressez continuellement pour maintenir l'excellence opérationnelle. Consacrez des cycles de travail à la réalisation quasi continuelle d'améliorations supplémentaires.

Effectuez une analyse post-incident de tous les événements ayant un impact sur le client. Identifiez les facteurs contributifs et les mesures préventives pour limiter ou empêcher la récurrence. Communiquez les facteurs contributifs aux communautés concernées, le cas échéant. Évaluez régulièrement et priorisez les possibilités d'amélioration (par exemple, les demandes de fonctionnalités, la correction des problèmes et les exigences de conformité), y compris la charge de travail et les procédures opérationnelles.

Introduisez des boucles de rétroaction au sein de vos procédures pour identifier rapidement les domaines d'amélioration et de tirer des enseignements de l'exécution d'opérations.

Partagez les leçons retenues et leurs avantages entre les équipes. Analysez les tendances dans les leçons apprises et effectuez une analyse rétrospective entre les équipes des opérations de métriques pour identifier les opportunités et les méthodes d'amélioration. Mettez en œuvre les changements destinés à apporter des améliorations et évaluez les résultats pour déterminer le succès.

Sur AWS, vous pouvez exporter vos données de journaux vers Amazon S3 ou envoyer les journaux directement vers Amazon S3 pour un stockage longue durée. Avec AWS Glue, vous pouvez découvrir et préparer vos données de journaux dans Amazon S3 à des fins d'analyse et stocker les métadonnées associées dans le AWS Glue Data Catalog. Amazon Athena, grâce à son intégration native dans AWS Glue, peut ensuite être utilisé pour analyser vos données de journaux, en les interrogeant à l'aide des requêtes SQL standard. En utilisant un outil de Business Intelligence comme Amazon QuickSight, vous pouvez visualiser, explorer et analyser vos données. Découvrez les tendances et les événements d'intérêt qui peuvent entraîner une amélioration.

La question suivante est axée sur ces considérations relatives à l'excellence opérationnelle.

OPS 11: How do you evolve operations?

Dedicate time and resources for nearly continuous incremental improvement to evolve the effectiveness and efficiency of your operations.

L'évolution réussie des opérations repose sur de fréquentes améliorations minimes, la fourniture d'environnements sûrs et le temps pour expérimenter, développer, tester les améliorations, et les environnements dans lesquels on encourage à tirer les leçons des échecs. La prise en charge des opérations pour les environnements sandbox, de développement, de test et de production, avec un niveau croissant de contrôles opérationnels, facilite le développement et augmente la prévisibilité des résultats positifs des changements déployés en production.

Ressources

Veuillez vous référer aux ressources suivantes pour en savoir plus sur les bonnes pratiques en matière d'excellence opérationnelle.

Documentation

- DevOps et AWS

Livre blanc

- Pilier Excellence opérationnelle

Vidéo

- DevOps chez Amazon

Sécurité

Le pilier Sécurité présente la capacité de protéger les données ainsi que les systèmes et les ressources pour tirer parti des technologies du cloud et améliorer votre sécurité.

Il fournit une vue d'ensemble des principes de conception, des bonnes pratiques et des questions. Vous trouverez des recommandations sur l'implémentation dans le Livre blanc du pilier Sécurité.

Rubriques

- Principes de conception
- Définition
- Bonnes pratiques
- Ressources

Principes de conception

Il existe sept principes de conception pour la sécurité dans le cloud :

- Mettre en place une solide base pour le contrôle des identités : mettez en œuvre le principe du moindre privilège et appliquez la séparation des responsabilités avec l'autorisation appropriée

pour chaque interaction avec vos ressources AWS. Centralisez la gestion des identités et visez l'élimination de la dépendance aux informations d'identification statiques de longue durée.

- Mettre en place la traçabilité : surveillez, alertez et contrôlez les actions et les modifications apportées à votre environnement en temps réel. Intégrez la collecte des journaux et des métriques aux systèmes pour effectuer des analyses et prendre des mesures automatiquement.
- Appliquer la sécurité à toutes les couches : appliquez une approche approfondie de protection avec plusieurs contrôles de sécurité. Appliquez-les à toutes les couches (par exemple, périphérie du réseau, VPC, équilibrage de charge, chaque instance et service de calcul, système d'exploitation, application et code).
- Automatiser les bonnes pratiques en matière de sécurité : les mécanismes de sécurité automatisés basés sur les logiciels améliorent votre capacité à vous mettre à l'échelle de manière plus rapide et plus rentable en toute sécurité. Créez des architectures sécurisées, y compris avec mise en œuvre des contrôles définis et gérés en tant que code dans les modèles de contrôle de versions.
- Protéger les données en transit et au repos : classez vos données selon différents niveaux de sensibilité et utilisez des mécanismes, tels que le chiffrement, la création de jetons et le contrôle d'accès, si nécessaire.
- Éviter les interventions humaines sur les données : utilisez des mécanismes et outils pour réduire ou éliminer le besoin d'accès direct ou le traitement manuel des données. Cette approche permet de réduire les risques de mauvaise manipulation ou de modification ainsi que les erreurs humaines lors d'interventions sur des données sensibles.
- Se préparer aux incidents impliquant la sécurité : préparez-vous à un incident en mettant en œuvre une stratégie et des processus de gestion et d'investigation des incidents conformes aux besoins de l'entreprise. Exécutez des simulations de réponse aux incidents et utilisez des outils d'automatisation pour améliorer votre vitesse de détection, d'investigation et de récupération.

Définition

Il existe six domaines de bonnes pratiques en matière de sécurité dans le cloud :

- Sécurité
- Identity and Access Management
- Détection
- Protection de l'infrastructure
- Protection des données

- Réponse aux incidents

Vous devez mettre en place des pratiques qui influent sur la sécurité avant de concevoir l'architecture d'une charge de travail. Vous voudrez contrôler qui peut faire quoi. De plus, vous voulez être en mesure d'identifier les incidents de sécurité, de protéger vos systèmes et services, et de maintenir la confidentialité et l'intégrité des données via la protection de données. Vous devez disposer d'un processus bien défini et utilisé pour répondre aux incidents de sécurité. Ces outils et techniques sont importants, car ils soutiennent des objectifs tels que la prévention des pertes financières ou le respect des obligations réglementaires.

Le modèle de responsabilité partagée AWS permet aux organisations qui adoptent le cloud d'atteindre leurs objectifs de sécurité et de conformité. Puisqu'AWS sécurise physiquement l'infrastructure qui prend en charge nos services cloud, en tant que client AWS vous pouvez vous concentrer sur l'utilisation de services pour accomplir vos objectifs. Le cloud AWS inclut également un meilleur accès aux données de sécurité et offre une approche automatisée pour répondre aux événements de sécurité.

Bonnes pratiques

Rubriques

- Sécurité
- Identity and Access Management
- Détection
- Protection de l'infrastructure
- Protection des données
- Réponse aux incidents

Sécurité

Vous devez appliquer de bonnes pratiques générales à chaque domaine de la sécurité pour réussir à gérer votre charge de travail en toute sécurité. Appliquez à tous les domaines les conditions et les processus que vous avez définis en matière d'excellence opérationnelle au niveau de l'organisation et de la charge de travail.

La connaissance des recommandations actuelles d'AWS et du secteur ainsi que des renseignements sur les menaces vous aide à faire évoluer votre modèle de menace et vos objectifs de contrôle.

L'automatisation des processus de sécurité, des tests et de la validation, vous permet de mettre à l'échelle vos opérations de sécurité.

La question suivante est axée sur ces considérations relatives à la sécurité. (Pour obtenir la liste des questions et des bonnes pratiques liées à la sécurité, consultez l' Annexe.)

SEC 1 : Comment gérer votre charge de travail en toute sécurité ?
Vous devez appliquer de bonnes pratiques générales à chaque domaine de la sécurité pour réussir à gérer votre charge de travail en toute sécurité. Appliquez à tous les domaines les conditions et les processus que vous avez définis en matière d'excellence opérationnelle au niveau de l'organisation et de la charge de travail. La connaissance des recommandations actuelles d'AWS pour votre secteur et des informations sur les menaces vous aidera à faire évoluer votre modèle de menaces et vos objectifs de contrôle. L'automatisation des processus de sécurité, des tests et de la validation, vous permet de mettre à l'échelle vos opérations de sécurité.

La séparation des différentes charges de travail par compte selon leur fonction et les exigences de conformité ou de sensibilité des données est une approche recommandée dans AWS.

Identity and Access Management

Identity and access management est un élément essentiel d'un programme de protection des informations. En effet, il garantit que seuls les utilisateurs et les composants autorisés et authentifiés sont en mesure d'accéder à vos ressources, et uniquement comme vous le décidez. Par exemple, vous devez définir des mandataires (c'est-à-dire les comptes, les utilisateurs, les rôles et les services qui peuvent effectuer des actions dans votre compte), élaborer des stratégies conformes à ces mandataires et mettre en œuvre une gestion solide des informations d'identification. Ces éléments de gestion des privilèges constituent la base de l'authentification et de l'autorisation.

Dans AWS, la gestion des privilèges est principalement prise en charge par le service AWS Identity and Access Management (IAM), qui vous permet de contrôler l'accès des utilisateurs et l'accès par programmation aux services et ressources AWS. Vous devez appliquer des politiques précises qui attribuent des autorisations à un utilisateur, un groupe, un rôle ou une ressource. Vous avez également la possibilité d'exiger des pratiques de mot de passe rigoureuses, telles que le niveau de complexité, qui vous évitent de réutiliser les mêmes et permet d'appliquer l'authentification multi-facteurs (MFA). Vous pouvez utiliser la fédération avec votre service d'annuaire existant. Pour les charges de travail qui exigent des systèmes qu'ils disposent d'un accès à AWS, IAM permet un

accès sécurisé via des rôles, des profils d'instance, une fédération d'identité et des informations d'identification temporaires.

Les questions suivantes sont axées sur ces quelques considérations liées à la sécurité.

SEC 2 : Comment gérer les identités des personnes et des machines ?

Il existe deux types d'identités à gérer dans le cadre de l'exploitation de charges de travail AWS sécurisées. Comprendre le type d'identité à gérer et dont vous devez autoriser l'accès vous permet de garantir l'accès aux ressources adéquates, dans les bonnes conditions.

Identités humaines : vos administrateurs, développeurs, opérateurs et utilisateurs finaux ont besoin d'une identité pour accéder à vos environnements et applications AWS. Il s'agit des membres de votre organisation ou des utilisateurs externes avec lesquels vous collaborez et qui interagissent avec vos ressources AWS via un navigateur Web, une application cliente ou des outils de ligne de commande interactifs.

Identités de machines : vos applications de service, outils opérationnels et charges de travail nécessitent une identité pour envoyer des demandes aux services AWS, par exemple pour lire des données. Ces identités comprennent des machines s'exécutant dans votre environnement AWS, telles que des instances Amazon EC2 ou des fonctions AWS Lambda. Vous pouvez également gérer les identités de machines pour les tiers qui ont besoin d'un accès. De plus, certaines machines en dehors d'AWS peuvent avoir besoin d'accéder à votre environnement AWS.

SEC 3 : Comment gérer les autorisations des personnes et des machines ?

Gérez les autorisations des identités de personnes et de machines qui nécessitent un accès à AWS ainsi qu'à votre charge de travail. Les autorisations régissent les ressources accessibles et les conditions d'accès.

Les informations d'identification ne doivent être partagées entre aucun utilisateur ou système. L'accès utilisateur doit être accordé via une approche de moindre privilège, en suivant différentes bonnes pratiques comme les exigences de mot de passe et l'application de l'authentification multi-facteurs. L'accès par programmation, y compris les appels d'API adressés aux services AWS, doit être exécuté à l'aide d'informations d'identification temporaires et à privilèges limités, telles que celles émises par le service AWS Security Token Service.

AWS fournit des ressources qui peuvent vous aider avec Identity and Access Management. Pour vous aider à apprendre les bonnes pratiques, explorez nos ateliers pratiques sur la gestion des informations d'identification et de l'authentification, le contrôle de l'accès humainet le contrôle de l'accès par programmation.

Détection

Vous pouvez utiliser les contrôles de détection pour identifier une menace ou un incident de sécurité potentiel. Ils constituent un élément essentiel des cadres de gouvernance et peuvent être utilisés pour soutenir un processus de qualité, une obligation légale ou de conformité et pour identifier les menaces et renforcer les moyens d'intervention. Il existe différents types de contrôles de détection. Par exemple, effectuer un inventaire des ressources et de leurs attributs détaillés favorise une prise de décision plus efficace (et des contrôles du cycle de vie) pour contribuer à établir des bases de référence opérationnelles. Vous pouvez également utiliser un audit interne, un examen des contrôles liés aux systèmes d'informations, pour vous assurer que les pratiques répondent aux politiques et aux exigences, et que vous avez défini les notifications d'alerte automatique correctes en fonction des conditions définies. Ces contrôles sont des facteurs réactifs importants qui peuvent aider votre organisation à identifier et à comprendre la portée des activités anormales.

Dans AWS, vous pouvez implémenter des contrôles de détection en traitant les journaux, les événements et la surveillance qui permettent d'effectuer des audits, des analyses automatisées et de configurer des alarmes. Les journaux CloudTrail, les appels d'API AWS et CloudWatch assurent une surveillance des métriques avec des alarmes, et AWS Config fournit un historique de configuration. Amazon GuardDuty est un service géré de détection des menaces qui surveille en permanence les comportements malveillants ou non autorisés pour vous aider à protéger vos comptes et charges de travail AWS. Les journaux de niveau de service sont également disponibles. Par exemple, vous pouvez utiliser Amazon Simple Storage Service (Amazon S3) pour journaliser les demandes d'accès.

La question suivante est axée sur ces considérations relatives à la sécurité.

SEC 4 : Comment détecter et enquêter sur les événements de sécurité ?

Capturez et analysez les événements de journaux et les métriques pour obtenir une meilleure visibilité. Prenez des mesures en réaction aux événements de sécurité et aux menaces potentiel les afin de contribuer à sécuriser votre charge de travail.

La gestion des journaux est essentielle dans le cadre d'une charge de travail well-architected, pour des raisons allant de sécurité, d'analyse ou d'exigences réglementaires ou légales. Il est essentiel

que vous analysiez les fichiers journaux et que vous y répondiez pour pouvoir identifier les problèmes de sécurité éventuels. AWS fournit des fonctionnalités qui simplifient l'implémentation de la gestion des journaux en vous offrant la possibilité de définir un cycle de vie de conservation des données ou de définir à quel emplacement les données seront conservées, archivées ou éventuellement supprimées. La gestion des données fiables et prévisibles en devient plus simple et plus économique.

Protection de l'infrastructure

La protection de l'infrastructure comprend des méthodologies de contrôle, telles que la défense en profondeur, nécessaires au respect des bonnes pratiques et des obligations organisationnelles ou réglementaires. L'utilisation de ces méthodologies est essentielle au succès des opérations en cours, que ce soit dans le cloud ou sur site.

Dans AWS, vous pouvez implémenter une inspection des paquets avec état et sans état, en utilisant des technologies AWS natives ou des produits et services partenaires disponibles via AWS Marketplace. Vous devez utiliser Amazon Virtual Private Cloud (Amazon VPC) pour créer un environnement privé, sécurisé et scalable dans lequel vous pouvez définir votre topologie, notamment les passerelles, les tables de routage et les sous-réseaux publics et privés.

Les questions suivantes sont axées sur ces quelques considérations liées à la sécurité.

SEC 5 : Comment protéger vos ressources réseau ?

Pour toute charge de travail ayant une forme quelconque de connectivité réseau, qu'il s'agisse d'Internet ou d'un réseau privé, plusieurs couches de défense sont nécessaires pour vous protéger contre les menaces externes et internes basées sur le réseau.

SEC 6 : Comment protéger vos ressources de calcul ?

Les ressources de calcul de votre charge de travail nécessitent plusieurs couches de défense pour vous aider à vous protéger des menaces externes et internes. Les ressources de calcul incluent les instances EC2, les conteneurs, les fonctions AWS Lambda, les services de bases de données, les appareils IoT, etc.

Plusieurs couches de défense sont conseillées dans tout type d'environnement. Dans le cas de la protection de l'infrastructure, la plupart des concepts et méthodes sont valides pour les modèles

cloud et sur site. L'application d'une protection de la périphérie, la surveillance des points d'entrée et de sortie, la journalisation complète, la supervision et les alertes, sont toutes essentielles à un plan de sécurité de l'information efficace.

Les clients AWS peuvent adapter ou renforcer la configuration d'un conteneur Amazon Elastic Compute Cloud (Amazon EC2), Amazon Elastic Container Service (Amazon ECS) ou d'une instance AWS Elastic Beanstalk et conserver cette configuration sur une Amazon Machine Image (AMI) immuable. Puis, qu'ils soient déclenchés par Auto Scaling ou lancés manuellement, tous les nouveaux serveurs virtuels (instances) lancés avec cet AMI reçoivent la configuration renforcée.

Protection des données

Avant de concevoir l'architecture d'un système, les pratiques de base qui influent sur la sécurité doivent être en place. Par exemple, la classification des données permet de classer les données organisationnelles en fonction des niveaux de sensibilité, et le chiffrement protège les données en les rendant incompréhensibles en cas d'accès non autorisé. Ces outils et techniques sont importants, car ils soutiennent des objectifs tels que la prévention des pertes financières ou le respect des obligations réglementaires.

Dans AWS, les pratiques suivantes facilitent la protection des données :

- En tant que client AWS, vous conservez la maîtrise totale de vos données.
- AWS facilite le chiffrement de vos données et la gestion des clés, y compris la rotation régulière des clés, qui peut être facilement automatisée par AWS ou gérée par vous-même.
- La journalisation détaillée qui contient des informations importantes, telles que l'accès aux fichiers et les modifications apportées, est disponible.
- AWS a conçu des systèmes de stockage pour une résilience exceptionnelle. Par exemple, Amazon S3 Standard, S3 Standard – Accès peu fréquent, S3 unizone – Accès peu fréquent et Amazon Glacier sont tous conçus pour fournir une durabilité des objets de 99,999999999 % sur une période d'un an. Ce niveau de durabilité correspond à une perte moyenne annuelle de 0,000000001 % des objets.
- La gestion des versions, qui peut faire partie d'un processus de gestion du cycle de vie des données plus étendu, assure une protection contre les remplacements ou suppressions accidentels et les dommages similaires.
- AWS n'initie jamais de transfert de données entre les régions. Le contenu affecté à une région restera dans celle-ci, à moins que vous n'activiez explicitement une fonction ou que vous n'exploitiez un service qui fournit cette fonctionnalité.

Les questions suivantes sont axées sur ces quelques considérations liées à la sécurité.

SEC 7 : Comment classer vos données ?

La classification des données fournit un moyen de classer les données en fonction de leur importance et de leur sensibilité afin de vous aider à déterminer les contrôles de protection et de conservation appropriés.

SEC 8 : Comment protéger les données au repos ?

Protégez vos données au repos en mettant en œuvre plusieurs contrôles, afin de réduire le risque d'accès non autorisé ou de mauvaise gestion.

SEC 9 : Comment protéger vos données en transit ?

Protégez vos données en transit en mettant en œuvre plusieurs contrôles afin de réduire le risque d'accès non autorisé ou de perte.

AWS fournit plusieurs moyens de chiffrement des données au repos et en transit. Nous intégrons à nos services des fonctionnalités qui facilitent le chiffrement de vos données. Par exemple, nous avons implémenté le chiffrement côté serveur (SSE) pour Amazon S3 afin de faciliter le stockage de vos données sous une forme chiffrée. Vous pouvez aussi prendre les dispositions nécessaires pour que la totalité du processus de chiffrement et de déchiffrement HTTPS (généralement appelé terminaison SSL) soit gérée par Elastic Load Balancing (ELB).

Réponse aux incidents

Même avec des contrôles de détection et de prévention extrêmement matures, votre organisation doit toujours mettre en place des processus pour répondre et atténuer l'impact potentiel d'incidents de sécurité. L'architecture de votre charge de travail affecte fortement la capacité de vos équipes à fonctionner efficacement lors d'un incident, à isoler ou à contenir des systèmes et à restaurer les opérations dans un état correct connu. La mise en place des outils et de l'accès avant un incident de sécurité, puis la mise en pratique régulière de la réponse aux incidents pendant les journées de jeu, vous aideront à vous assurer que votre architecture peut prendre en charge des enquêtes et des restaurations rapides.

Dans AWS, les pratiques suivantes facilitent la gestion efficace face des incidents :

- La journalisation détaillée qui contient des informations importantes, telles que l'accès aux fichiers et les modifications apportées, est disponible.
- Les événements peuvent être traités automatiquement et déclencher des outils qui automatisent les réponses via l'utilisation d'API AWS.
- Vous pouvez préprovisionner des outils et une « salle blanche » à l'aide de AWS CloudFormation. Cela vous permet d'effectuer des analyses dans un environnement sécurisé et isolé.

La question suivante est axée sur ces considérations relatives à la sécurité.

SEC 10 : Comment anticiper les incidents, y répondre et reprendre les activités par la suite ?

La préparation est essentielle pour une enquête rapide et efficace, une réponse et une reprise en cas d'incidents de sécurité, afin de minimiser les perturbations pour votre organisation.

Assurez-vous de disposer d'un moyen permettant d'accorder rapidement l'accès à votre équipe de sécurité. Automatisez également l'isolation des instances ainsi que la saisie de données et l'état des analyses.

Ressources

Consultez les ressources suivantes pour en savoir plus sur nos bonnes pratiques relatives à la sécurité.

Documentation

- Sécurité du cloud AWS
- Conformité AWS
- Blog sur la sécurité AWS

Livre blanc

- Pilier Sécurité
- Présentation de la sécurité sur AWS
- Risque et conformité AWS

Vidéo

- AWS Security State of the Union
- Présentation de la responsabilité partagée

Fiabilité

Le pilier Fiabilité englobe la capacité d'une charge de travail à exécuter sa fonction de manière correcte et cohérente et ce, en temps utile. Cela inclut la possibilité d'exploiter et de tester la charge de travail tout au long de son cycle de vie. Ce livre blanc fournit des bonnes pratiques détaillées pour la mise en œuvre de charges de travail fiables sur AWS.

Le pilier fiabilité fournit une vue d'ensemble des principes de conception, des bonnes pratiques et des questions. Vous trouverez des recommandations sur l'implémentation dans le Livre blanc du pilier Fiabilité.

Rubriques

- Principes de conception
- Définition
- Bonnes pratiques
- Ressources

Principes de conception

Il existe cinq principes de conception pour la fiabilité dans le cloud :

- Reprise automatique après une panne : en contrôlant les indicateurs de performance clés (KPI) d'une charge de travail, vous pouvez déclencher l'automatisation en cas de dépassement d'un seuil. Ces KPI doivent couvrir la valeur commerciale et non des aspects techniques du fonctionnement du service. Cela permet la création de notifications automatiques, le suivi des pannes et l'exécution de processus de récupération automatique qui contournent ou corrigent les pannes. Une automatisation plus sophistiquée rend possible l'anticipation et la correction des pannes avant qu'elles ne se produisent.
- Test des procédures de reprise : dans un environnement sur site, des tests sont souvent nécessaires pour prouver que la charge de travail fonctionne dans un scénario particulier. Ces tests ne sont généralement pas utilisés pour valider les stratégies de récupération. Dans le cloud,

vous pouvez tester de quelle façon votre charge de travail cesse de fonctionner et valider vos procédures de récupération. Vous pouvez utiliser l'automatisation pour simuler différentes pannes ou recréer les scénarios qui ont déjà conduit à des pannes. Cette approche permet de réduire les risques en exposant les chemins de défaillance que vous pouvez tester et corriger avant qu'un scénario de défaillance réelle ne se produise.

- Mise à l'échelle horizontale pour augmenter la disponibilité de la charge de travail agrégée : remplacez une ressource volumineuse par plusieurs petites ressources pour réduire l'impact d'une panne unique sur la charge de travail globale. Répartissez les demandes entre plusieurs ressources plus petites pour garantir qu'elles ne partagent pas un point de panne commun.
- Une capacité réellement adaptée à vos besoins : une cause courante de défaillance des charges de travail sur site est la saturation des ressources, c'est-à-dire lorsque les demandes imposées à une charge de travail dépassent la capacité de cette dernière (c'est souvent l'objectif des attaques par déni de service). Dans le cloud, vous pouvez contrôler la demande et l'utilisation de la charge de travail. Vous pouvez aussi automatiser l'ajout ou la suppression de ressources afin de maintenir le niveau optimal de satisfaction de la demande sans surallocation ou sous-allocation. Des limites demeurent, mais certains quotas peuvent être contrôlés et d'autres gérés (consultez Gestion des Service Quotas et contraintes de service).
- Gestion des modifications de l'automatisation : les modifications apportées à votre infrastructure doivent être faites à l'aide de l'automatisation. Les modifications qui doivent être gérées incluent celles apportées à l'automatisation et qui peuvent ensuite être suivies et vérifiées.

Définition

Il existe quatre domaines de bonnes pratiques en matière de fiabilité dans le cloud :

- Bases
- Architecture de charge de travail
- Gestion des modifications
- Gestion des pannes

Pour la fiabilité, vous devez commencer par les bases, c'est-à-dire un environnement où les quotas de service et la topologie réseau s'adaptent à la charge de travail. L'architecture de la charge de travail du système distribué doit être conçue pour prévenir et atténuer les défaillances. La charge de travail doit gérer les modifications au niveau de la demande ou des exigences. Elle doit être conçue pour détecter les défaillances et se réparer automatiquement.

Bonnes pratiques

Rubriques

- Fondations
- Architecture de charge de travail
- Gestion des modifications
- Gestion des défaillances

Fondations

Les exigences de base sont celles dont le champ d'application s'étend au-delà d'une seule charge de travail ou d'un seul projet. Avant de concevoir l'architecture d'un système, les exigences de base qui influent sur la fiabilité doivent être mises en place. Par exemple, vous devez avoir une bande passante réseau suffisante pour votre centre de données.

Avec AWS, la plupart de ces exigences élémentaires sont déjà intégrées ou sont gérées au cas par cas. Le cloud est conçu pour être presque illimité. Il est donc de la responsabilité d'AWS de satisfaire l'exigence d'une capacité suffisante de mise en réseau et de calcul, ce qui vous laisse la liberté de modifier la taille des ressources et les allocations à la demande.

Les questions suivantes sont axées sur ces quelques considérations relatives à la fiabilité. (Pour obtenir la liste des questions et bonnes pratiques liées à la fiabilité, consultez l' Annexe.)

REL 1 : Comment gérer les quotas et les contraintes de service ?

Pour les architectures de charge de travail basées sur le cloud, il existe des quotas de service (aussi appelés Service Limits). Le rôle de ces quotas est d'empêcher la mise en service accidente lle de plus de ressources que nécessaire et de limiter les taux de demandes sur les opération s d'API afin de protéger les services contre les abus. Il existe également des contraintes de ressource. Par exemple, la vitesse à laquelle vous pouvez transmettre des bits sur un câble de fibre optique, ou la quantité de stockage sur un disque physique.

REL 2 : Comment planifier la topologie de votre réseau ?

Les charges de travail existent souvent dans plusieurs environnements. Il s'agit notamment de plusieurs environnements cloud (accessibles publiquement et privés) et éventuellement de votre

REL 2 : Comment planifier la topologie de votre réseau ?

infrastructure de centre de données existante. Les plans doivent inclure des considérations réseau telles que la connectivité intrasystème et intersystème, la gestion des adresses IP publiques, la gestion des adresses IP privées et la résolution des noms de domaine.

Pour les architectures de charge de travail basées sur le cloud, il existe des quotas de service (aussi appelés Service Limits). Ces quotas visent à empêcher la mise en service accidentelle de plus de ressources que nécessaire et à limiter les taux de requêtes sur les opérations d'API pour protéger les services contre les abus. Les charges de travail existent souvent dans plusieurs environnements. Vous devez surveiller et gérer ces quotas pour tous les environnements de charge de travail.
Il s'agit notamment de plusieurs environnements cloud (accessibles publiquement et privés) et peuvent inclure votre infrastructure de centre de données existante. Les plans doivent tenir compte de facteurs liés au réseau : connectivité intrasystème et intersystème, gestion des adresses IP publiques, gestion des adresses IP privées et résolution des noms de domaine.

Architecture de charge de travail

Pour garantir la fiabilité d'une charge de travail, il faut commencer par choisir le bon logiciel et la bonne infrastructure. Vos choix d'architecture ont un impact sur le comportement des charges de travail sur les différents piliers Well-Architected. Pour des raisons de fiabilité, vous devez suivre des modèles spécifiques.

Avec AWS, les développeurs de charges de travail peuvent choisir les langages et les technologies à utiliser. Les kits SDK AWS éliminent la complexité du codage en fournissant des API propres au langage pour les services AWS. Ces kits SDK, ainsi que le choix des langages, permettent aux développeurs de mettre en œuvre les bonnes pratiques de fiabilité répertoriées ici. Les développeurs peuvent également découvrir comment Amazon conçoit et exploite des logiciels dans l'Amazon Builders' Library.

Les questions suivantes sont axées sur ces quelques considérations relatives à la fiabilité.

REL 3 : Comment concevoir l'architecture de service de votre charge de travail ?

Créez des charges de travail hautement évolutives et fiables à l'aide d'une architecture orientée service (SOA) ou d'une architecture de microservices. La SOA consiste à rendre les composants logiciels réutilisables via les interfaces de service. L'architecture des microservices va plus loin, en particulier en rendant les composants plus petits et plus simples.

REL 4 : Comment concevoir des interactions dans un système distribué pour éviter les défaillances ?

Les systèmes distribués s'appuient sur des réseaux de communication pour interconnecter des composants comme des serveurs ou des services. Votre charge de travail doit fonctionner de manière fiable malgré la perte de données ou la latence dans ces réseaux. Les composants du système distribué doivent fonctionner d'une manière qui n'a pas d'impact négatif sur les autres composants ou la charge de travail. Ces bonnes pratiques empêchent les défaillances et améliorent le temps moyen entre les défaillances (MTBF).

REL 5 : Comment concevoir des interactions dans un système distribué pour atténuer ou résister aux défaillances ?

Les systèmes distribués s'appuient sur des réseaux de communication pour interconnecter des composants (tels que des serveurs ou des services). Votre charge de travail doit fonctionner de manière fiable malgré la perte de données ou la latence sur ces réseaux. Les composants du système distribué doivent fonctionner d'une manière qui n'a pas d'impact négatif sur les autres composants ou la charge de travail. Ces bonnes pratiques permettent aux charges de travail de résister aux contraintes ou aux défaillances, de s'en remettre plus rapidement et d'atténuer l'impact de ces déficiences. Il en résulte une amélioration du temps moyen de récupération (MTTR).

Gestion des modifications

Les modifications apportées à votre charge de travail ou à son environnement doivent être anticipées et prises en compte pour assurer un fonctionnement fiable de la charge de travail. Les modifications incluent celles imposées à votre charge de travail telles que les pics de demande, ainsi que celles venant de l'intérieur comme les déploiements de fonctions et l'installation de correctifs de sécurité.

Avec AWS, vous pouvez surveiller le comportement d'une charge de travail et automatiser la réponse aux KPI. Par exemple, votre charge de travail peut ajouter des serveurs supplémentaires à mesure que des utilisateurs supplémentaires s'y ajoutent. Vous pouvez contrôler les personnes qui ont l'autorisation d'apporter des modifications à la charge de travail et auditer l'historique de ces modifications.

Les questions suivantes sont axées sur ces quelques considérations relatives à la fiabilité.

REL 6 : Comment surveiller les ressources de charges de travail ?

Les journaux et les métriques sont de puissants outils pour obtenir informations sur l'état de votre charge de travail. Vous pouvez configurer votre charge de travail de sorte à surveiller les journaux et les métriques et envoyer des notifications lorsque les seuils sont franchis ou que des événements significatifs se produisent. La surveillance permet à votre charge de travail de reconnaître quand des seuils de faible performance sont franchis ou quand des défaillances se produisent, en vue de sa reprise automatique.

REL 7 : Comment concevoir votre charge de travail pour qu'elle s'adapte aux changements de demande ?

Une charge de travail évolutive fournit l'élasticité nécessaire pour ajouter ou supprimer automatiq uement des ressources de telle sorte qu'elles correspondent étroitement à tout moment à la demande en cours.

REL 8 : Comment implémenter les modifications ?

Des modifications contrôlées sont nécessaires pour au moins deux raisons : déployer de nouvelles fonctionnalités et s'assurer que les charges de travail et l'environnement d'exploitation fonctionn ent avec des logiciels connus et peuvent être corrigés ou remplacés de manière prévisible. Si les modifications ne sont pas contrôlées, il est difficile de prédire leur effet ou de résoudre les problèmes qui en découlent.

Lorsque vous concevez l'architecture d'une charge de travail de manière à ajouter ou supprimer automatiquement des ressources en réponse à l'évolution de la demande, cela accroît la fiabilité et garantit que la réussite commerciale ne devient pas un poids. Une fois la surveillance en place, votre équipe est automatiquement avertie lorsque les KPI cessent de correspondre aux valeurs attendues. La journalisation automatique des modifications apportées à votre environnement vous permet d'auditer et d'identifier rapidement les actions susceptibles d'avoir un impact sur la fiabilité. Les contrôles de la gestion des modifications permettent de s'assurer que vous appliquez les règles offrant la fiabilité dont vous avez besoin.

Gestion des défaillances

Les pannes sont à prévoir dans tout système présentant un niveau de complexité raisonnable. Pour que votre charge de travail soit fiable, vous devez avoir connaissance des défaillances au moment où elles se produisent et prendre des mesures pour éviter qu'elles aient un impact sur la disponibilité des services. Les charges de travail doivent être en mesure de résister aux défaillances et de résoudre automatiquement les problèmes.

Avec AWS, vous pouvez tirer profit de l'automatisation pour réagir aux données de surveillance. Par exemple, lorsqu'une métrique particulière franchit un seuil, vous pouvez déclencher une action automatique pour corriger le problème. De même, plutôt que de tenter de diagnostiquer et de corriger une ressource défaillante qui fait partie de votre environnement de production, vous pouvez la remplacer par une nouvelle ressource et exécuter l'analyse de cette ressource hors production. Comme le Cloud vous permet de maintenir les versions temporaires d'un système complet à bas coût, vous pouvez utiliser les tests automatiques pour vérifier les processus complets de récupération.

Les questions suivantes sont axées sur ces quelques considérations relatives à la fiabilité.

REL 9 : Comment sauvegarder des données ?

Sauvegardez les données, les applications et la configuration pour répondre à vos exigences en matière d'objectifs de temps de récupération (RTO) et d'objectifs de point de récupération (RPO).

REL 10 : Comment utiliser l'isolation des pannes pour protéger votre charge de travail ?

Les limites isolées pour les défaillances limitent l'effet d'une défaillance au sein d'une charge de travail à un nombre limité de composants. Les composants en dehors de la limite ne sont pas affectés par la défaillance. En utilisant plusieurs limites isolées par défaut, vous pouvez limiter l'impact sur votre charge de travail.

REL 11 : Comment concevoir votre charge de travail pour la rendre résistante aux défaillances de composants ?

Les charges de travail exigeant une haute disponibilité et un faible temps moyen de récupération (MTTR) doivent être conçues pour être résilientes.

REL 12 : Comment tester la fiabilité ?

Une fois que vous avez conçu votre charge de travail pour qu'elle soit résiliente aux sollicitations de la production, les tests sont le seul moyen de s'assurer qu'elle fonctionne comme prévu et d'obtenir la résilience voulue.

REL 13 : Comment planifier la reprise après sinistre (DR) ?

La mise en place de sauvegardes et de composants de charge de travail redondants constitue le début de votre stratégie de DR. RTO et RPO sont vos objectifs pour la restauration de votre charge de travail. Définissez-les en fonction des besoins de l'entreprise. Mettez en œuvre une stratégie pour atteindre ces objectifs, en particulier en tenant compte de l'emplacement et de la fonction des données et des ressources de charge de travail. La probabilité d'une perturbation et le coût de la reprise sont également des facteurs clés qui permettent de déterminer la valeur opérationnelle de la reprise après sinistre d'une charge de travail.

Sauvegardez régulièrement vos données et testez vos fichiers de sauvegarde pour vous assurer de pouvoir récupérer aussi bien après des erreurs logiques que des erreurs physiques. La clé de la gestion des pannes réside dans des tests réguliers et automatiques des charges de travail afin de créer des pannes, et dans l'observation de la façon dont ces charges reprennent. Effectuez ces opérations régulièrement et assurez-vous que de tels tests sont également déclenchés après des modifications significatives de la charge de travail. Suivez activement les KPI, ainsi que l'objectif de délai de reprise (RTO) et l'objectif de point de reprise (RPO) pour évaluer la résilience d'une charge de travail (notamment au cours de scénarios de test de panne). Le suivi des KPI vous aidera à identifier et à atténuer les points de défaillance uniques. L'objectif est de tester intégralement vos processus de reprise de charge de travail de telle sorte que vous soyez assuré de récupérer l'ensemble de vos données et de continuer à servir vos clients, même en présence de problèmes persistants. Vos processus de reprise doivent être aussi bien maîtrisés que vos processus de production habituels.

Ressources

Consultez les ressources suivantes pour en savoir plus sur nos bonnes pratiques relatives à la fiabilité.

Documentation

- Documentation AWS
- Infrastructure mondiale AWS
- AWS Auto Scaling : Fonctionnement des plans de mise à l'échelle
- Qu'est-ce que AWS Backup ?

Livre blanc

- Pilier Fiabilité : AWS Well-Architected
- Implémentation des microservices sur AWS

Efficacité en matière de performance

Le pilier Efficacité des performances comprend la capacité à utiliser efficacement les ressources de calcul pour répondre aux exigences du système et à maintenir cette efficacité à mesure que la demande change et les technologies évoluent.

Le pilier Efficacité des performances fournit une vue d'ensemble des principes de conception, des bonnes pratiques et des questions. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Efficacité en matière de performance.

Rubriques

- Principes de conception
- Définition
- Bonnes pratiques
- Ressources

Principes de conception

Il existe cinq principes de conception pour l'efficacité des performances dans le cloud :

- Démocratiser les technologies avancées : simplifiez la mise en œuvre de technologies avancées pour votre équipe en déléguant des tâches complexes à votre fournisseur de cloud. Plutôt que de demander à votre équipe informatique de s'informer sur l'hébergement et l'exploitation de

nouvelles technologies, envisagez de consommer les technologie en tant que service. Par exemple, les bases de données NoSQL, le transcodage multimédia et le machine learning sont trois technologies qui requièrent des compétences spécialisées. Dans le cloud, ces technologies deviennent des services que votre équipe peut consommer, ce qui lui permet de se dédier au développement de produits plutôt qu'à l'allocation et à la gestion des ressources.

- Une portée mondiale en quelques minutes : le déploiement de votre charge de travail dans plusieurs régions AWS du monde vous permet d'offrir une latence plus faible et une meilleure expérience à vos clients, à un coût réduit.
- Utilisation d'architectures sans serveur : les architectures sans serveur vous évitent d'exécuter et de gérer des serveurs physiques pour les activités traditionnelles de calcul. Par exemple, les services de stockage sans serveur peuvent agir comme des sites Web statiques (éliminant le besoin de serveurs Web), et les services d'événements peuvent héberger du code. Ainsi, vous supprimez la charge opérationnelle de gestion des serveurs physiques et réduisez les coûts des transactions, car les services gérés fonctionnent à l'échelle du cloud.
- Expérimentation plus fréquente : avec des ressources virtuelles et automatisables, vous pouvez rapidement exécuter des tests comparatifs à l'aide de différents types d'instances, de stockages ou de configurations.
- Tenir compte de la « sympathie mécanique » : comprenez comment les services cloud sont consommés et utilisez toujours l'approche technologique qui correspond le mieux à vos objectifs de charges de travail. Par exemple, tenez compte des modèles d'accès aux données lorsque vous sélectionnez les approches de stockage ou de base de données.

Définition

Les bonnes pratiques en matière d'efficacité des performances dans le cloud sont au nombre de cinq :

- Choix d'architecture
- Informatique et matériel
- Gestion des données
- Mise en réseau et diffusion de contenu
- Processus et culture

Optez pour une approche orientée données lors de la création d'une architecture à hautes performances. Collectez des données sur tous les aspects de l'architecture, depuis la conception générale jusqu'à la sélection et la configuration des types de ressources.

En réexaminant vos choix régulièrement, vous faites en sorte de tirer parti de l'évolution constante du cloud AWS. La surveillance vous offre la garantie d'être informé de tout écart par rapport aux performances attendues. Effectuez des compromis dans votre architecture pour améliorer les performances, comme l'utilisation de la compression, la mise en cache ou l'abaissement des exigences de cohérence.

Bonnes pratiques

Rubriques

- Choix d'architecture
- Informatique et matériel
- Gestion des données
- Mise en réseau et diffusion de contenu
- Processus et culture

Choix d'architecture

La solution optimale pour une charge de travail peut varier, et les solutions combinent souvent plusieurs approches. Les charges de travail bien architecturées utilisent plusieurs solutions et permettent d'exploiter différentes fonctionnalités pour améliorer les performances.

De nombreux types et configurations de ressources AWS sont proposés. Il est ainsi plus facile de trouver l'approche qui correspond le mieux à vos besoins. Vous pouvez également rechercher des options qui ne sont pas facilement accessibles avec une infrastructure sur site. Par exemple, un service géré tel que Amazon DynamoDB fournit une base de données NoSQL entièrement gérée avec une latence de moins de dix millisecondes, quelle que soit l'échelle.

La question suivante est axée sur ces quelques considérations relatives à l'efficacité des performances. (Pour obtenir la liste des questions et bonnes pratiques en matière d'efficacité des performances, consultez l' Appendix.)

PERF 1: How do you select appropriate cloud resources and architecture patterns for your workload?

Often, multiple approaches are required for more effective performance across a workload. Well-Architected systems use multiple solutions and features to improve performance.

Informatique et matériel

Le choix d'une solution de calcul optimale pour une charge de travail particulière peut varier selon la conception de l'application, les modèles d'utilisation et les paramètres de configuration. Les architectures peuvent utiliser différentes solutions de calcul pour divers composants et permettent différentes fonctionnalités pour améliorer les performances. Choisir une solution de calcul inadaptée pour une architecture peut nuire à ses performances.

Dans AWS, les capacités de calcul sont disponibles sous trois formes, à savoir les instances, les conteneurs et les fonctions :

- Les instances sont des serveurs virtualisés, ce qui vous permet de modifier leurs capacités à l'aide d'un bouton ou d'un appel d'API. Comme les décisions relatives aux ressources dans le cloud sont pas figées, vous pouvez expérimenter avec différents types de serveurs. Dans AWS, ces instances de serveurs virtuels se déclinent en différentes familles et tailles, et elles offrent une grande variété de fonctionnalités, notamment des disques Solid-State Drives (SSD) et des unités de traitement graphique (GPU).
- Les Conteneurs sont une méthode de virtualisation du système d'exploitation vous permettant d'exécuter une application et ses dépendances dans des processus isolés par les ressources. AWS Fargate est un système de calcul sans serveur pour les conteneurs, ou Amazon EC2 peut être utilisé si vous avez besoin de contrôler l'installation, la configuration et la gestion de votre environnement de calcul. Vous pouvez également choisir parmi plusieurs plateformes d'orchestration de conteneurs : Amazon Elastic Container Service (ECS) ou Amazon Elastic Kubernetes Service (EKS).
- Les fonctions permettent d'abstraire l'environnement d'exécution du code que vous souhaitez appliquer. Par exemple, AWS Lambda vous permet d'exécuter du code sans exécuter d'instance.

La question suivante est axée sur ces quelques considérations relatives à l'efficacité des performances.

PERF 2: How do you select and use compute resources in your workload?

The more efficient compute solution for a workload varies based on application design, usage patterns, and configuration settings. Architectures can use different compute solutions for various components and turn on different features to improve performance. Selecting the wrong compute solution for an architecture can lead to lower performance efficiency.

Gestion des données

La solution optimale de gestion des données pour un système particulier varie en fonction du type de données (bloc, fichier ou objet), des modèles d'accès (aléatoire ou séquentiel), du débit requis, de la fréquence d'accès (en ligne, hors ligne, archivage), de la fréquence de mise à jour (WORM, dynamique), ainsi que des contraintes de disponibilité et de durabilité. Les charges de travail bien architecturées utilisent des magasins de données sur mesure qui intègrent différentes fonctionnalités pour améliorer les performances.

Dans AWS, le stockage est disponible sous trois formes : par objet, par bloc ou par fichier.

- Le stockage d'objet fournit une plateforme évolutive et durable pour rendre les données accessibles depuis n'importe quel emplacement Internet pour le contenu généré par l'utilisateur, l'archivage actif, le calcul sans serveur, le stockage du big data ou la sauvegarde et la restauration. Amazon Simple Storage Service (Amazon S3) est un service de stockage d'objets qui offre une capacité de mise à l’échelle, une disponibilité des données, une sécurité et des performances de pointe. Amazon S3 est conçu pour offrir une durabilité de 99,999999999 % (à 11 9) et stocke les données de millions d'applications pour des entreprises du monde entier.
- Le stockage par blocs fournit un stockage par blocs hautement disponible, cohérent et à faible latence pour chaque hôte virtuel. Il est analogue au stockage à connexion directe (DAS) ou à un réseau SAN (Storage Area Network). Amazon Elastic Block Store (Amazon EBS) est conçu pour les charges de travail qui nécessitent un stockage permanent accessible par les instances EC2, ce qui vous aide à ajuster les applications avec la capacité de stockage, les performances et le coût appropriés.
- Le stockage de fichiers permet d'accéder à un système de fichiers partagé sur plusieurs systèmes. Les solutions de stockage de fichiers comme Amazon Elastic File System (Amazon EFS) sont idéales pour les cas d'utilisation tels que les référentiels de contenu volumineux, les environnements de développement, les magasins multimédias ou les répertoires de base d'utilisateurs. Amazon FSx facilite et rentabilise le lancement et l'exécution de systèmes de

fichiers populaires. Vous pouvez ainsi tirer parti des ensembles de fonctionnalités riches et des performances rapides des systèmes open source et sous licence commerciale très répandus.

La question suivante est axée sur ces quelques considérations relatives à l'efficacité des performances.

PERF 3: How do you store, manage, and access data in your workload?

The more efficient storage solution for a system varies based on the kind of access operation (block, file, or object), patterns of access (random or sequential), required throughput, frequency of access (online, offline, archival), frequency of update (WORM, dynamic), and availability and durability constraints. Well-architected systems use multiple storage solutions and turn on different features to improve performance and use resources efficiently.

Mise en réseau et diffusion de contenu

La solution de mise en réseau optimale pour une charge de travail varie en fonction de la latence, des exigences de débit, de l'instabilité et de la bande passante. Le choix des options d'emplacement est tributaire des contraintes physiques telles que les ressources pour utilisateur ou sur site. Ces contraintes peuvent être compensées avec les emplacements périphériques ou le placement des ressources.

Sur AWS, la mise en réseau est virtualisée et disponible dans plusieurs types et configurations. Il est ainsi plus facile d'adapter vos méthodes de mise en réseau à vos besoins. AWS propose des fonctionnalités de produit (par exemple, la mise en réseau améliorée, les instances optimisées pour Amazon EC2, Amazon S3 Transfer Acceleration et Amazon CloudFront dynamique) pour optimiser le trafic réseau. AWS propose également des fonctionnalités de mise en réseau (par exemple, le routage de latence Amazon Route 53, les points de terminaison du Amazon VPC, AWS Direct Connect et AWS Global Accelerator) pour réduire l'instabilité ou la distance du réseau.

La question suivante est axée sur ces quelques considérations relatives à l'efficacité des performances.

PERF 4: How do you select and configure networking resources in your workload?

This question includes guidance and best practices to design, configure, and operate efficient networking and content delivery solutions in the cloud.

Processus et culture

Lors de la création de l'architecture des charges de travail, vous pouvez adopter certains principes et certaines pratiques pour optimiser l'exécution de charges de travail cloud efficaces et performantes. Pour adopter une culture qui favorise l'efficacité des performances des charges de travail dans le cloud, tenez compte des principes et pratiques clés suivants.

Tenez compte de ces principes clés pour développer cette culture :

- Infrastructure en tant que code : définissez votre infrastructure en tant que code à l'aide de méthodes telles que les modèles AWS CloudFormation. L'utilisation de modèles vous permet de placer votre infrastructure en mode de contrôle de code source parallèlement au code et aux configurations de votre application. Ceci vous permet d'appliquer les pratiques utilisées pour développer des logiciels à votre infrastructure et ainsi d'itérer rapidement.
- Pipeline de déploiement : utilisez un pipeline de déploiement d'intégration continue (CI) et de livraison continue (CD) (par exemple, référentiel de code source, systèmes de génération, déploiement et automatisation des tests) pour déployer votre infrastructure. Cela vous permet de déployer de manière reproductible et cohérente, le tout à un faible coût, à mesure que vous itérez.
- Métriques bien définies : configurez et surveillez vos métriques pour capturer les indicateurs de performances clés (KPI). Nous vous recommandons d'utiliser des métriques techniques, mais aussi des métriques commerciales. Pour les sites web ou les applications mobiles, les indicateurs clés capturent le temps jusqu'au premier octet ou le rendu. Les autres métriques applicables de manière générale comprennent le nombre de threads, la vitesse de nettoyage de la mémoire et les états d'attente. Les métriques commerciales, telles que les coûts cumulés agrégés par demande, peuvent vous permettre d'identifier des solutions pour réduire vos coûts. Réfléchissez bien à la façon dont vous prévoyez d'interpréter les métriques. Par exemple, vous pouvez choisir le maximum ou le 99e centile plutôt que la moyenne.
- Tests de performance automatiques : dans le cadre de votre processus de déploiement, des tests de performance peuvent se déclencher automatiquement une fois les tests en cours d'exécution effectués avec succès. L'automatisation doit créer un environnement, configurer des conditions initiales (comme des données de test), puis exécuter une série d'analyses comparatives et de tests de charge. Les résultats de ces tests doivent être rattachés à la version de génération afin que vous puissiez suivre l'évolution des performances dans le temps. Pour les tests de longue durée, vous pouvez rendre cette partie du pipeline asynchrone par rapport au reste de la compilation. Sinon, vous pouvez exécuter des tests de performances pendant la nuit en utilisant les instances Spot Amazon EC2.

- Génération de charge : vous devez créer une série de scripts qui reproduisent des parcours utilisateur synthétiques ou préenregistrés. Ces scripts doivent être idempotents et non couplés. Il se peut que vous deviez aussi inclure à cette série des scripts de préparation pour obtenir des résultats valides. Dans la mesure du possible, vos scripts de test doivent pouvoir répliquer le comportement d'utilisation en production. Vous pouvez utiliser un logiciel ou des solutions de logiciel en tant que service (SaaS) pour générer la charge. Envisagez d'utiliser les solutions AWS Marketplace et les instances Spot : elles peuvent être des moyens économiques de générer la charge.
- Visibilité des performances : les métriques clés doivent être visibles pour votre équipe, en particulier pour chaque version. Cela vous permet d'identifier les tendances positives ou négatives significatives au fil du temps. Vous devez également afficher les métriques sur le nombre d'erreurs ou d'exceptions pour vous assurer que vous testez un système fonctionnel.
- Visualisation : utilisez des techniques de visualisation qui permettent d'identifier clairement l'origine des problèmes de performances, les points chauds, les états d'attente ou les taux d'utilisation faibles. Superposez les métriques de performance sur les schémas d'architecture, des graphiques ou codes d'appel qui peuvent vous aider à identifier rapidement les problèmes.
- Processus d’évaluation régulier : les architectures qui présentent des performances médiocres sont généralement le résultat d'un processus d'évaluation des performances inexistant ou interrompu. Si votre architecture est peu performante, la mise en œuvre d'un processus d'évaluation des performances vous permet d'apporter des améliorations itératives.
- Optimisation continue : adoptez une culture permettant d'optimiser en permanence l'efficacité des performances de votre charge de travail dans le cloud.

La question suivante est axée sur ces quelques considérations relatives à l'efficacité des performances.

PERF 5: What process do you use to support more performance efficiency for your workload?

When architecting workloads, there are principles and practices that you can adopt to help you better run efficient high-performing cloud workloads. To adopt a culture that fosters performance efficiency of cloud workloads, consider these key principles and practices.

Ressources

Consultez les ressources suivantes pour en savoir plus sur nos bonnes pratiques relatives à l'efficacité des performances.

Documentation

- Optimisation des performances Amazon S3
- Performances des volumes Amazon EBS

Livre blanc

- Pilier Efficacité des performances

Vidéo

- AWS re:Invent 2019: foundations Amazon EC2 (CMP211-R2)
- AWS re:Invent 2019: Leadership session: Storage state of the union (STG201-L)
- AWS re:Invent 2019: Leadership session: AWS purpose-built databases (DAT209-L)
- AWS re:Invent 2019: Connectivity to AWS and hybrid AWS network architectures (NET317-R1)
- AWS re:Invent 2019: Powering next-gen Amazon EC2: Deep dive into the Nitro system (CMP303-R2)
- AWS re:Invent 2019: Scaling up to your first 10 million users (ARC211-R)

Optimisation des coûts

Le pilier Optimisation des coûts comprend la possibilité d'exécuter des systèmes pour offrir une valeur métier au prix le plus bas.

Le pilier que représente l'optimisation des coûts fournit une vue d'ensemble des principes de conception, des bonnes pratiques et des interrogations. Vous trouverez des recommandations sur l'implémentation dans le Livre blanc du pilier Optimisation des coûts.

Rubriques

- Principes de conception

- Définition
- Bonnes pratiques
- Ressources

Principes de conception

Il existe cinq principes de conception pour l'optimisation des coûts dans le cloud :

- Mettre en œuvre la gestion financière dans le cloud : pour obtenir un succès financier et accélérer la génération de valeur métier dans le cloud, vous devez investir dans la gestion financière du cloud/l'optimisation des coûts. Votre organisation doit consacrer du temps et des ressources à la création de fonctionnalités dans ce nouveau domaine de la technologie et de la gestion de l'utilisation. Comme pour les piliers Sécurité ou Excellence opérationnelle, vous devez renforcer vos capacités par l'acquisition de connaissances, de programmes, de ressources et de processus pour devenir une entreprise rentable.
- Adopter un modèle de consommation : payez uniquement pour les ressources de calcul dont vous avez besoin et augmentez ou diminuez l'utilisation en fonction des exigences métier et non à l'aide de prévisions sophistiquées. Par exemple, les environnements de développement et de test ne sont généralement utilisés que huit heures par jour pendant la semaine de travail. Vous pouvez désactiver ces ressources lorsqu'elles ne sont pas utilisées et réduire les coûts de 75 % (40 heures au lieu de 168 heures).
- Mesurer l'efficacité globale : mesurez les résultats métiers de vos charges de travail et les coûts associés à leur exécution. Utilisez ces informations pour déterminer les avantages que vous pouvez tirer de l'augmentation du rendement et de la réduction des coûts.
- Stop aux dépenses pour le gros du travail : AWS se charge des opérations lourdes des centres de données, telles que la mise en rack, l'empilement et l'alimentation des serveurs. Cela supprime également la charge opérationnelle liée à la gestion des systèmes d'exploitation et des applications avec des services gérés. Ainsi, vous pouvez vous dédier aux clients et aux projets professionnels plutôt qu'à l'infrastructure informatique.
- Analyser et répartir les dépenses : le cloud permet d'appréhender plus facilement et avec précision l'utilisation et les coûts des systèmes, ce qui permet ensuite de répartir de manière transparente les frais informatiques entre les différents propriétaires de charges de travail. Cela permet de mesurer le retour sur investissement et offre la possibilité aux propriétaires de programmes d'optimiser leurs ressources et de réduire les coûts.

Définition

Il existe cinq domaines de bonnes pratiques pour l'optimisation des coûts dans le cloud :

- Pratiques en matière de gestion financière du cloud
- Sensibilisation aux dépenses et à l'utilisation
- Ressources rentables
- Gérer la demande et les sources d'approvisionnement
- Optimiser dans le temps

Comme pour les autres piliers du cadre Well-Architected, il est nécessaire d'établir des compromis en se demandant, par exemple, s'il est préférable d'optimiser pour accélérer la mise sur le marché ou pour réduire les coûts. Dans certains cas, plutôt que d'investir dans une optimisation des coûts initiaux, il est préférable d'optimiser la vitesse, avec une mise sur le marché rapide, la livraison de nouvelles fonctions ou le simple respect d'une échéance. Les décisions de conception sont parfois prises avec vitesse et non selon les données, et il est tentant de surcompenser « au cas où », plutôt que de consacrer du temps à des essais comparatifs pour un déploiement le plus optimal en termes de coût. Cela peut entraîner des déploiements sur-approvisionnés et sous-optimisés. Cependant, il s'agit d'un choix raisonnable lorsque vous avez besoin de « basculer » depuis votre environnement sur site vers le cloud, puis d'optimiser par la suite. Investir dès le départ la bonne quantité d'efforts dans une stratégie d'optimisation des coûts, permet de réaliser plus facilement les avantages économiques du cloud en garantissant une adhésion constante aux bonnes pratiques et en évitant une sur-allocation inutile. Les sections suivantes présentent les techniques et les bonnes pratiques pour la mise en œuvre initiale et continue de la gestion financière dans le cloud et l'optimisation des coûts de vos charges de travail.

Bonnes pratiques

Rubriques

- Pratiques en matière de gestion financière du cloud
- Sensibilisation aux dépenses et à l'utilisation
- Ressources rentables
- Gérer la demande et les sources d'approvisionnement
- Optimiser dans le temps

Pratiques en matière de gestion financière du cloud

Avec l'adoption du cloud, les équipes technologiques innovent plus rapidement grâce à la réduction des cycles d'approbation, d'achat et de déploiement des infrastructures. Une nouvelle approche de la gestion financière dans le cloud est nécessaire pour générer de la valeur ajoutée et connaître le succès financier. Cette approche, appelée « gestion financière dans le cloud », permet de renforcer les capacités de votre organisation en mettant en œuvre des programmes, des ressources et des processus de renforcement des connaissances à l'échelle de l'organisation.

De nombreuses organisations sont composées de nombreuses unités différentes avec des priorités différentes. La capacité d'aligner votre organisation sur un ensemble d'objectifs financiers convenus et de lui fournir les mécanismes nécessaires pour les atteindre, crée une organisation plus efficace. Une organisation sera capable d'innover et de créer plus rapidement, d'être plus agile et de s'adapter à tous les facteurs internes ou externes.

Dans AWS, vous pouvez utiliser Cost Explorer et, éventuellement, Amazon Athena et Amazon QuickSight avec le rapport de coût et d'utilisation (CUR) pour mieux connaître les coûts et l'utilisation dans l'ensemble de votre entreprise. AWS Budgets fournit des notifications proactives concernant les coûts et l'utilisation. Les blogs AWS fournissent des informations sur les nouveaux services et fonctionnalités afin de vous tenir informé des nouvelles versions de service.

La question suivante est axée sur ces quelques considérations relatives à l'optimisation des coûts. (Pour obtenir la liste des questions et bonnes pratiques en matière d'optimisation des coûts, consultez l' Annexe.)

COST 1 : Comment mettre en œuvre la gestion financière du cloud ?

La gestion financière du cloud (CFM) permet aux organisations de générer de la valeur ajoutée et d'être financièrement performantes en optimisant leurs coûts et l'utilisation, et de se développer sur AWS.

Lors de la création d'une fonction d'optimisation des coûts, ayez recours aux membres et complétez l'équipe avec des experts en gestion financière du cloud et en optimisation des coûts. Les membres actuels de l'équipe comprendront comment l'entreprise fonctionne actuellement et détermineront la manière de mettre en œuvre rapidement des améliorations. Pensez également à inclure des personnes disposant de compétences supplémentaires ou spécialisées, telles que dans les domaines de l'analyse et de la gestion de projet.

Lorsque vous mettez en œuvre la sensibilisation aux coûts dans votre entreprise, améliorez ou appuyez-vous sur les programmes et processus existants. Il est beaucoup plus rapide d'ajouter des intégrations aux processus et programmes existants, que d'en créer de nouveaux. Ainsi, les résultats sont beaucoup plus rapides.

Sensibilisation aux dépenses et à l'utilisation

La flexibilité et la souplesse accrues que permet le cloud favorisent l'innovation, ainsi que le développement et le déploiement à un rythme soutenu. Le Cloud élimine les processus manuels et les délais associés au provisionnement d'une infrastructure locale, y compris l'identification des spécifications matérielles, la négociation des devis, la gestion des bons de commande, la planification des livraisons et le déploiement des ressources. Cependant, la facilité d'utilisation et la capacité illimitée et à la demande nécessitent une nouvelle façon d'envisager les dépenses.

De nombreuses entreprises sont composées de plusieurs systèmes, dirigés par diverses équipes. La possibilité de répartir les coûts des ressources entre les différentes organisations ou les différents responsables de produits permet un comportement d'utilisation efficace et contribue à réduire le gaspillage. La répartition précise des coûts permet d'identifier les produits réellement rentables, et de prendre des décisions en connaissance de cause quant à la répartition du budget.

Dans AWS, vous créez une structure de compte avec AWS Organizations ou AWS Control Tower, ce qui assure la séparation et facilite la répartition de vos coûts et de votre utilisation. Vous pouvez également utiliser le balisage des ressources pour appliquer les informations de l'entreprise à votre utilisation et à vos coûts. Utilisez AWS Cost Explorer pour améliorer la visibilité de vos coûts et de votre utilisation, ou créez des tableaux de bord et des analyses personnalisés avec Amazon Athena et Amazon QuickSight. Le contrôle de vos coûts et de votre utilisation est effectué par des notifications via AWS Budgets et par des contrôles à l'aide d'AWS Identity and Access Management (IAM), et de Service Quotas.

Les questions suivantes sont axées sur ces quelques considérations relatives à l'optimisation des coûts.

COST 2 : Comment gérer l'utilisation ?

Définissez des stratégies et des mécanismes pour vous assurer que les coûts appropriés sont facturés tout en atteignant les objectifs. En adoptant une approche d'équilibre des pouvoirs, vous pouvez innover sans dépense excessive.

COST 3 : Comment surveiller l'utilisation et les coûts ?

Définissez des stratégies et des procédures pour surveiller et allouer vos coûts de manière appropriée. Cela vous permet d'évaluer et d'améliorer la rentabilité de cette charge de travail.

COST 4 : Comment mettre les ressources hors service ?

Mettez en œuvre le contrôle des modifications et la gestion des ressources depuis le début du projet jusqu'à la fin. Cela garantit que vous arrêtiez ou résiliez les ressources inutilisées pour réduire le gaspillage.

Vous pouvez utiliser des balises de répartition des coûts pour catégoriser et suivre votre utilisation et vos coûts AWS. Lorsque vous appliquez des balises à vos ressources AWS (telles que les instances EC2 ou les compartiments S3), AWS génère un rapport d'utilisation et de coût avec votre utilisation et vos balises. Vous pouvez appliquer des balises qui représentent des catégories de l'organisation (telles que les centres de coûts, les noms des charges de travail ou les propriétaires) pour organiser vos coûts dans plusieurs services.

Veillez à utiliser le niveau de détail et la granularité appropriés dans les rapports et la surveillance des coûts et de l'utilisation. Pour obtenir des informations de haut niveau et des tendances générales, utilisez la granularité quotidienne avec AWS Cost Explorer. Pour une analyse et une inspection plus approfondies, utilisez la granularité horaire dans AWS Cost Explorer, ou Amazon Athena et Amazon QuickSight avec le rapport de coût et d'utilisation (CUR) à une granularité horaire.

La combinaison de ressources balisées et d'une fonction suivi du cycle de vie des entités (employés, projets) permet d'identifier les ressources orphelines ou les projets qui ne génèrent plus de valeur pour l'organisation et qui doivent être mis hors service. Vous pouvez configurer des alertes de facturation pour être averti des dépassements de dépenses prévisibles.

Ressources rentables

L'utilisation d'instances et de ressources adaptées à votre charge de travail est l'élément essentiel à la réalisation d'économies. Par exemple, un processus de reporting peut prendre jusqu'à cinq heures pour s'exécuter sur un petit serveur, mais seulement une heure sur un serveur plus grand et deux fois plus cher. Vous obtiendrez les mêmes résultats avec les deux serveurs, mais le plus petit implique un coût plus élevé au fil du temps.

Une charge de travail Well-Architected utilise les ressources les plus rentables, ce qui peut avoir un impact économique positif et significatif. Vous pouvez également utiliser des services gérés pour réduire les coûts. Par exemple, plutôt que d'entretenir des serveurs pour envoyer des e-mails, vous pouvez utiliser un service effectuant une facturation au message.

AWS propose une grande variété d'options de tarification flexibles et rentables pour acquérir des instances d'Amazon EC2 et d'autres services et ce, de la façon qui correspond le mieux à vos besoins. À la demande Instances vous permettent de payer la capacité de calcul à l'heure, sans aucun engagement minimum. Savings Plans et les instances réservées vous permettent d'économiser jusqu'à 75 % sur la tarification à la demande. Avec les instances Spot, vous pouvez tirer profit de la capacité Amazon EC2 non utilisée et économiser jusqu'à 90 % sur la tarification à la demande. Instances Spot sont adaptées lorsque le système est en mesure de prendre en charge l'utilisation d'une flotte de serveurs où les serveurs individuels peuvent aller et venir de manière dynamique, comme les serveurs Web sans état, le traitement par lots ou lors de l'utilisation du calcul hautes performances (HPC) et du Big Data.

La sélection du service approprié peut également réduire l'utilisation et les coûts (tel que CloudFront pour minimiser le transfert de données), ou éliminer totalement les coûts (tel que l'utilisation d'Amazon Aurora sur RDS pour supprimer les coûts élevés de licence de bases de données).

Les questions suivantes sont axées sur ces quelques considérations relatives à l'optimisation des coûts.

COST 5 : Comment évaluer les coûts lorsque vous sélectionnez des services ?

Amazon EC2, Amazon EBS et Amazon S3 sont des services fondamentaux d'AWS. Les services gérés, tels que Amazon RDS et Amazon DynamoDB, sont des services AWS de plus haut niveau, ou au niveau de l'application. En sélectionnant les services fondamentaux et les services gérés appropriés, vous pouvez optimiser cette charge de travail en termes de coûts. Par exemple, en utilisant des services gérés, vous pouvez réduire ou supprimer une grande partie de votre traitement administratif et opérationnel, et vous dégagez ainsi du temps pour travailler sur les applications et les activités liées aux activités.

COST 6 : Comment atteindre les objectifs de coût lorsque vous sélectionnez le type, la taille et le nombre de ressources ?

Veillez à choisir la taille et le nombre de ressources qui conviennent pour la tâche à accomplir. En choisissant le type, la taille et le nombre les plus rentables, vous réduisez le gaspillage.

COST 7 : Comment utiliser les modèles de tarification pour réduire les coûts ?

Utilisez le modèle de tarification qui convient le mieux à vos ressources pour réduire les dépenses.

COST 8 : Comment planifier les frais de transfert de données ?

Veillez à planifier et à surveiller les frais de transfert de données afin de pouvoir prendre des décisions architecturales pour minimiser les coûts. Une modification architecturale minime, mais efficace, peut réduire de façon spectaculaire vos coûts d'exploitation.

En prenant en compte les coûts lors de la sélection du service et en utilisant des outils tels que Cost Explorer et AWS Trusted Advisor pour vérifier régulièrement votre utilisation AWS, vous pouvez surveiller activement votre utilisation et ajuster vos déploiements en conséquence.

Gérer la demande et les sources d'approvisionnement

Lorsque vous migrez vers le cloud, vous ne payez que ce dont vous avez besoin. Vous pouvez fournir des ressources pour répondre à la demande de la charge de travail au moment où elles sont nécessaires, ce qui élimine une sur-allocation coûteuse et inutile. Vous pouvez également modifier la demande à l'aide d'une limitation, d'une mémoire-tampon ou d'une file d'attente pour la lisser et la gérer avec moins de ressources, ce qui réduit les coûts, ou la traiter ultérieurement avec un service de traitement par lots.

Dans AWS, vous pouvez allouer automatiquement des ressources pour répondre à la demande de charge de travail. Auto Scaling utilisant des approches basées sur la demande ou sur le temps vous permet d'ajouter et de supprimer des ressources selon les besoins. Si vous pouvez anticiper l'évolution de la demande, vous pouvez économiser plus et faire en sorte que vos ressources répondent aux besoins de la charge de travail. Vous pouvez utiliser Amazon API Gateway pour

mettre en place une limitation ou Amazon SQS pour mettre en place une file d'attente dans votre charge de travail. Ces deux éléments vous permettent de modifier la demande sur les composants de votre charge de travail.

La question suivante est axée sur ces quelques considérations relatives à l'optimisation des coûts.

COST 9 : Comment gérer les ressources de demande et d'offre ?

Pour une charge de travail dont les dépenses et les performances sont équilibrées, assurez-v ous que tout ce que vous payez est utilisé et évitez une sous-utilisation importante des instances . Une métrique d'utilisation faussée dans un sens ou dans l'autre a un impact négatif sur votre organisation, que ce soit en termes de coûts d'exploitation (dégradation des performances due à une surutilisation) ou de gaspillage de dépenses AWS (en raison d'une surallocation).

Lorsque vous concevez dans le but de modifier la demande et l'offre de ressources, pensez activement aux modèles d'utilisation, au temps nécessaire pour allouer de nouvelles ressources et à la prévisibilité du modèle de la demande. Lors de la gestion de la demande, veillez à disposer d'une file d'attente ou d'une mémoire tampon correctement dimensionnée et à répondre à la demande de la charge de travail dans le délai requis.

Optimiser dans le temps

Tandis qu'AWS propose de nouveaux services et de nouvelles fonctionnalités, une bonne pratique consiste à vérifier vos choix architecturaux existants afin d'être sûr qu'ils continuent à être les plus rentables. Lorsque vous besoins changent, n'hésitez pas à mettre hors service les ressources, les services entiers et les systèmes devenus inutiles.

La mise en œuvre de nouvelles fonctionnalités ou de nouveaux types de ressources peut optimiser votre charge de travail de façon progressive, tout en minimisant les efforts requis pour mettre en œuvre la modification. Ainsi, vous améliorez continuellement l'efficacité au fil du temps et vous restez à la pointe de la technologie pour réduire les coûts d'exploitation. Vous pouvez également remplacer ou ajouter de nouveaux composants à la charge de travail avec de nouveaux services. Cela peut accroître considérablement l'efficacité. Il est donc essentiel de vérifier régulièrement votre charge de travail et de mettre en œuvre de nouveaux services et de nouvelles fonctionnalités.

La question suivante est axée sur ces quelques considérations relatives à l'optimisation des coûts.

COST 10 : Comment évaluer les nouveaux services ?

Tandis qu'AWS propose de nouveaux services et de nouvelles fonctionnalités, une bonne pratique consiste à vérifier vos choix architecturaux existants afin d'être sûr qu'ils continuent à être les plus rentables.

En réexaminant régulièrement vos déploiements, évaluez dans quelle mesure des services plus récents peuvent vous permettre de réaliser des économies. Par exemple, Amazon Aurora sur RDS peut réduire les coûts des bases de données relationnelles. L'utilisation des technologies serverless, telle que Lambda, peut éviter d'exploiter et de gérer des instances pour exécuter du code.

Ressources

Consultez les ressources suivantes pour en savoir plus sur nos bonnes pratiques relatives à l'optimisation des coûts.

Documentation

- Documentation AWS

Livre blanc

- Pilier Optimisation des coûts

Durabilité

Le pilier Durabilité se concentre sur les impacts environnementaux, notamment la consommation et l'efficacité énergétiques, qui sont des leviers importants permettant aux architectes de recueillir des informations sur les actions directes afin d'utiliser moins de ressources. Vous trouverez des recommandations sur l'implémentation dans le Livre blanc du pilier Durabilité.

Rubriques

- Principes de conception
- Définition
- Bonnes pratiques

Principes de conception

Il existe six principes de conception pour la durabilité dans le cloud :

- Déterminer votre impact : Mesurez l'impact de votre charge de travail sur le cloud et modélisez l'impact futur de votre charge de travail. Incluez toutes les sources d'impact, y compris les impacts résultant de l'utilisation de vos produits par les clients, et ceux découlant de leur éventuelle mise hors service. Comparez le rendement productif à l'impact total de vos charges de travail sur le cloud en évaluant les ressources et les émissions nécessaires par unité de travail. Utilisez ces données pour établir des indicateurs clés de performance (KPI), évaluer des moyens d'améliorer la productivité tout en réduisant l'impact et estimer l'impact des changements proposés au fil du temps.
- Définir des objectifs de durabilité : Pour chaque charge de travail dans le cloud, établissez des objectifs de durabilité à long terme, tels que la réduction des ressources de calcul et de stockage nécessaires par transaction. Modélisez le retour sur investissement des améliorations durables pour les charges de travail existantes et donnez aux propriétaires les ressources dont ils ont besoin pour investir dans leurs objectifs de durabilité. Planifiez en vue d'une croissance et concevez l'architecture de vos charges de travail afin que la croissance entraîne une intensité de l'impact moindre mesurée par rapport à une unité appropriée, par utilisateur ou par transaction par exemple. Les objectifs vous aident à soutenir les cibles de durabilité plus larges de votre entreprise ou organisation, identifier les régressions et privilégier les zones pouvant être améliorées.
- Optimiser l'utilisation : Dimensionnez correctement les charges de travail et intégrez une conception efficace pour assurer une forte utilisation et optimiser l'efficacité énergétique du matériel sous-jacent. Deux hôtes s'exécutant à 30 % de leur utilisation sont moins efficaces qu'un seul hôte s'exécutant à 60 % du fait de la consommation énergétique de base par hôte. Éliminez ou minimisez également les ressources, le traitement et le stockage inactifs afin de réduire l'énergie totale nécessaire pour alimenter votre charge de travail.
- Anticiper et adopter des offres matérielles et logicielles neuves et plus efficaces : Soutenez les améliorations en amont de vos partenaires et fournisseurs pour aider à réduire l'impact de vos charges de travail sur le cloud. Contrôlez et évaluez de façon continue des offres matérielles et logicielles neuves et plus efficaces. Concevez de manière flexible afin de permettre l'adoption rapide de nouvelles technologies efficaces.
- Utiliser des services gérés : Le partage des services auprès d'une clientèle importante permet d'optimiser l'utilisation des ressources, ce qui réduit la quantité d'infrastructure nécessaire pour soutenir les charges de travail dans le cloud. Par exemple, les clients peuvent partager l'impact de composants courants du centre de données tels que l'énergie et les réseaux en migrant des

charges de travail vers le cloud AWS Cloud et en adoptant des services gérés, pour les conteneurs sans serveur, comme AWS Fargate. AWS y opère à grande échelle et est responsable de son fonctionnement efficace. Utilisez des services gérés qui peuvent aider à réduire votre impact, tels que le déplacement automatique des données à accès peu fréquent vers un stockage froid avec les configurations du cycle de vie Amazon S3 ou Amazon EC2 Auto Scaling pour ajuster la capacité à répondre à la demande.

- Réduire l'impact en aval de vos charges de travail dans le cloud : Réduisez la quantité d'énergie ou de ressources nécessaires pour utiliser vos services. Réduisez ou supprimez la nécessité pour les clients de mettre à niveau leurs appareils afin d'utiliser vos services. Réalisez des tests à l'aide de Device Farms pour comprendre l'impact attendu et auprès de clients pour comprendre l'impact réel que représente l'utilisation de vos services.

Définition

Il existe six domaines de bonnes pratiques en matière de durabilité dans le cloud :

- Choix de la région
- Modèles de comportement des utilisateurs
- Modèles logiciels et d'architecture
- Modèles de données
- Modèles matériels
- Processus de développement et de déploiement

La durabilité dans le cloud est un effort continu axé principalement sur la réduction d'énergie et l'efficacité de tous les composants d'une charge de travail en tirant le meilleur parti possible des ressources allouées et en minimisant les ressources totales requises. Cet effort peut inclure la sélection initiale d'un langage de programmation efficace, l'adoption d'algorithmes modernes, l'utilisation de techniques de stockage de données performantes, le déploiement sur une infrastructure de calcul correctement dimensionnée et efficace, et la réduction des besoins en matériel de grande puissance pour les utilisateurs finaux.

Bonnes pratiques

Rubriques

- Choix de la région

- Modèles de comportement des utilisateurs
- Modèles logiciels et d'architecture
- Modèles de données
- Modèles matériels
- Modèles de développement et de déploiement
- Ressources

Choix de la région

Choisissez les régions où mettre en œuvre vos charges de travail en fonction des exigences et des objectifs de durabilité de votre entreprise.

La question suivante est axée sur les considérations relatives à la durabilité. (Pour obtenir la liste des questions et bonnes pratiques liées à la durabilité, consultez l' Annexe.)

SUS 1 : Comment choisir les régions afin de prendre en charge vos objectifs de durabilité ?

Choisissez des régions proches des projets d'énergie renouvelable d'Amazon et des régions où le réseau a une intensité en carbone publique inférieure aux autres sites (ou régions).

Modèles de comportement des utilisateurs

La façon dont les utilisateurs consomment vos charges de travail et d'autres ressources peut vous aider à identifier les améliorations nécessaires pour atteindre vos objectifs de durabilité. Mettez votre infrastructure à l'échelle afin qu'elle corresponde toujours à la charge de l'utilisateur et garantir que seul le minimum de ressources nécessaires pour soutenir les utilisateurs est déployé. Alignez les niveaux de service sur les besoins des clients. Positionnez des ressources afin de limiter le réseau nécessaire aux utilisateurs pour les consommer. Supprimez toute ressources existante inutilisée. Identifiez les ressources créées et inutilisées, et arrêtez de les générer. Fournissez des appareils aux membres de votre équipe qui répondent à leurs besoins avec un impact minimal en matière de durabilité.

La question suivante est axée sur les considérations relatives à la durabilité :

SUS 2 : Comment tirer parti des modèles de comportement des utilisateurs afin de soutenir vos objectifs de durabilité ?

La façon dont les utilisateurs consomment vos charges de travail et d'autres ressources peut vous aider à identifier les améliorations nécessaires pour atteindre vos objectifs de durabilité. Mettez votre infrastructure à l'échelle afin qu'elle corresponde toujours à la charge de l'utilisateur et garantir que seul le minimum de ressources nécessaires pour soutenir les utilisateurs est déployé. Alignez les niveaux de service sur les besoins des clients. Positionnez des ressources afin de limiter le réseau nécessaire aux utilisateurs pour les consommer. Supprimez toute ressource s existante inutilisée. Identifiez les ressources créées et inutilisées, et arrêtez de les générer. Fournissez des appareils aux membres de votre équipe qui répondent à leurs besoins avec un impact minimal en matière de durabilité.

Mettre à l'échelle l'infrastructure avec la charge de l'utilisateur : identifiez les périodes d'utilisation faible ou nulle, et mettez vos ressources à l'échelle afin de supprimer toute capacité excédentaire et de gagner en efficacité.

Aligner les SLA sur vos objectifs de durabilité : définissez et mettez à jour les contrats de niveau de service (SLA), tels que la durabilité ou les périodes de conservation des données, afin de réduire le nombre de ressources nécessaires pour assurer votre charge de travail tout en continuant à répondre à vos exigences métier.

Éliminer la création et la gestion des ressources inutilisées : analysez les ressources de l'application (telles que les rapports pré-compilés, les jeux de données et les images statiques) et les modèles d'accès aux ressources pour identifier des redondances, une sous-utilisation et d'éventuelles cibles de mise hors service. Consolidez les ressources générées avec le contenu redondant (par exemple, des rapports mensuels avec des jeux de données et des résultats se chevauchant ou courants) pour éliminer les ressources consommées lors de la duplication des résultats. Mettez hors service les ressources inutilisées (par exemple, des images de produits qui ne sont plus vendus) afin de libérer des ressources consommées et réduire le nombre de ressources utilisées afin de soutenir la charge de travail.

Optimiser l'emplacement géographique des charges de travail en fonction de la localisation des utilisateurs : analysez les modèles d'accès au réseau pour identifier les lieux de connexion de vos clients. Choisissez des régions et des services qui réduisent la distance que le trafic du réseau doit parcourir afin de diminuer le nombre total de ressources réseau nécessaires pour assurer votre charge de travail.

Optimiser les ressources des membres de l'équipe pour les activités réalisées : optimisez les ressources fournies aux membres de l'équipe pour réduire l'impact sur la durabilité tout en répondant à leurs besoins. Par exemple, réalisez des opérations complexes, telles que le rendu et la compilation, sur des bureaux partagés sur le cloud et hautement utilisés au lieu de systèmes à utilisateur unique sous-utilisés et très puissants.

Modèles logiciels et d'architecture

Mettez en œuvre des modèles permettant de lisser les charges et de conserver une haute utilisation constante des ressources déployées afin de réduire les ressources consommées. Les composants peuvent devenir inactifs s'ils ne sont pas utilisés à la suite de changements de comportement des utilisateurs dans le temps. Révisez les modèles et l'architecture afin de consolider les composants sous-utilisés et d'augmenter l'utilisation globale. Mettez hors service les composants qui ne sont plus nécessaires. Comprenez les performances des composants de vos charges de travail et optimisez les composants qui consomment le plus de ressources. Soyez au courant des appareils que vos clients utilisent pour accéder à vos services et mettez en œuvre des modèles qui réduisent le besoin de mettre à niveau les appareils.

Les questions suivantes sont axées sur les considérations relatives à la durabilité :

SUS 3 : Comment tirer parti des modèles logiciels et d'architecture afin de soutenir vos objectifs de durabilité ?

Mettez en œuvre des modèles permettant de lisser les charges et de conserver une haute utilisation constante des ressources déployées afin de réduire les ressources consommées. Les composants peuvent devenir inactifs s'ils ne sont pas utilisés à la suite de changements de comportement des utilisateurs dans le temps. Révisez les modèles et l'architecture afin de consolider les composants sous-utilisés et d'augmenter l'utilisation globale. Mettez hors service les composants qui ne sont plus nécessaires. Comprenez les performances des composants de vos charges de travail et optimisez les composants qui consomment le plus de ressources. Soyez au courant des appareils que vos clients utilisent pour accéder à vos services et mettez en œuvre des modèles qui réduisent le besoin de mettre à niveau les appareils.

Optimiser les logiciels et l'architecture pour les tâches asynchrones et prévues : Utilisez des conceptions et des architectures logicielles efficaces pour réduire les ressources moyennes nécessaires par unité de travail. Mettez en œuvre des mécanismes qui entraînent une utilisation

uniforme des composants pour réduire les ressources inactives entre deux tâches et réduire l'impact des pics de charge.

Supprimer ou refactoriser les composants de charges de travail faiblement utilisés ou inutilisés : surveillez l'activité de la charge de travail pour identifier des changements dans l'utilisation des composants individuels dans le temps. Supprimez les composants utilisés et qui ne sont plus nécessaires, et refactorisez les composants peu utilisés afin de limiter le gaspillage des ressources.

Optimiser les sections de votre code les plus longues ou qui consomment le plus de ressources : contrôlez l'activité de la charge de travail pour identifier les composants de l'application qui consomment le plus de ressources. Optimisez le code exécuté dans ces composants pour réduire l'utilisation des ressources tout en optimisant la performance.

Optimiser l'impact sur les appareils et les équipements des clients : ayez une compréhension des appareils et du matériel utilisés par vos clients pour consommer vos services, leur cycle de vie prévu et l'impact financier et durable que représente le remplacement de ces composants. Mettez en œuvre des modèles et des architectures logiciels pour réduire le besoin pour les clients de remplacer les appareils et de mettre à niveau leur matériel. Par exemple, mettez en œuvre de nouvelles fonctions en utilisant du code compatible avec du matériel et des versions de systèmes d'exploitation plus anciens, ou gérez la taille des charges utiles afin qu'elles n'excèdent pas la capacité de stockage de l'appareil cible.

Utiliser des modèles logiciels et des architectures qui soutiennent au mieux l'accès aux données et les modèles de stockage : comprenez comment les données sont utilisées au sein de votre charge de travail, comment elles sont consommées par vos utilisateurs, transférées et stockées. Sélectionnez des technologies afin de réduire le traitement des données et les exigences de stockage.

Modèles de données

Mettez en œuvre des modèles permettant de lisser les charges et de conserver une haute utilisation constante des ressources déployées afin de réduire les ressources consommées. Les composants peuvent devenir inactifs s'ils ne sont pas utilisés à la suite de changements de comportement des utilisateurs dans le temps. Révisez les modèles et l'architecture afin de consolider les composants sous-utilisés et d'augmenter l'utilisation globale. Mettez hors service les composants qui ne sont plus nécessaires. Comprenez les performances des composants de vos charges de travail et optimisez les composants qui consomment le plus de ressources. Soyez au courant des appareils que vos clients utilisent pour accéder à vos services et mettez en œuvre des modèles qui réduisent le besoin de mettre à niveau les appareils.

La question suivante est axée sur les considérations relatives à la durabilité :

SUS 4 : Comment profiter de l'accès aux données et des modèles d'utilisation afin de soutenir vos objectifs de durabilité ?

Mettez en œuvre des pratiques de gestion des données afin de réduire le stockage alloué nécessaire pour assurer votre charge de travail et les ressources nécessaires à son utilisation. Comprenez vos données et utilisez des technologies et des configurations de stockage qui soutiennent au mieux la valeur opérationnelle des données et leur utilisation. Adoptez un cycle de vie des données offrant un stockage plus efficace et moins performant quand les exigences baissent et supprimez les données qui ne sont plus nécessaires.

Mettre en œuvre une politique de classification des données : classez les données afin de déterminer leur importance pour les résultats commerciaux. Utilisez ces informations afin de déterminer quand déplacer vos données vers un stockage plus économe en énergie ou les supprimer en toute sécurité.

Utiliser des technologies qui prennent en charge les modèles d'accès aux données et de stockage : exploitez le stockage qui prend le mieux en charge l'accès à vos données et leur stockage afin de réduire les ressources allouées tout en soutenant votre charge de travail. Par exemple, les appareils SSD sont plus gourmands en énergie que les disques magnétiques et doivent uniquement être utilisés pour les cas d'utilisation de données actives. Utilisez un stockage de classe d'archivage économe en énergie pour les données rarement consultées.

Utiliser des politiques de cycle de vie pour supprimer les données inutiles : Gérez le cycle de vie de toutes vos données et appliquez automatiquement des délais de suppression pour réduire l'ensemble des besoins de stockage de votre charge de travail.

Réduire le sur-approvisionnement dans le stockage par bloc : pour réduire au minimum le stockage alloué total, créez un stockage par bloc avec des allocations de taille adaptées à la charge de travail. Utilisez des volumes Elastic pour agrandir le stockage au fur et à mesure que les données augmentent sans avoir à redimensionner le stockage attaché aux ressources de calcul. Examinez régulièrement les volumes Elastic et réduisez les volumes sur-alloués pour qu'ils correspondent à la taille actuelle des données.

Supprimer les données inutiles ou redondantes : dupliquez les données uniquement lorsque cela s'avère nécessaire pour réduire le stockage total consommé. Utilisez des technologies de sauvegarde qui dédupliquent les données au niveau du fichier et du bloc. Limitez l'utilisation de configurations RAID (Redundant Array of Independent Drives), sauf si nécessaire pour respecter les SLA.

Utiliser des systèmes de fichiers partagés ou le stockage d'objets pour accéder aux données courantes : adoptez le stockage partagé et des sources uniques de confiance pour éviter la duplication des données et réduire l'ensemble des besoins en stockage pour votre charge de travail. Récupérez les données à partir du stockage partagé uniquement en fonction des besoins. Détachez les volumes inutilisés afin de libérer des ressources. Réduisez au minimum les déplacements des données entre les réseaux : utilisez le stockage partagé et accédez aux données des magasins de données régionaux pour réduire les ressources de réseaux totales nécessaires à la prise en charge des mouvements des données pour votre charge de travail.

Sauvegarder des données uniquement lorsqu'elles sont difficiles à recréer : Afin de réduire la consommation de stockage, sauvegardez uniquement les données ayant une valeur opérationnelle ou nécessaires pour répondre aux exigences en matière de conformité. Examinez les politiques de sauvegarde et excluez tout magasin éphémère n'apportant aucune valeur dans un scénario de récupération.

Modèles matériels

Recherchez des possibilités de réduire les impacts en matière de durabilité de la charge de travail en modifiant vos pratiques de gestion du matériel. Réduisez la quantité de matériel nécessaire à allouer et à déployer, et sélectionnez le matériel le plus efficace pour votre charge de travail individuelle.

La question suivante est axée sur les considérations relatives à la durabilité :

SUS 5 : Comment vos pratiques de gestion et d'utilisation du matériel soutiennent-elles vos objectifs en matière de durabilité ?

Recherchez des possibilités de réduire les impacts en matière de durabilité de la charge de travail en modifiant vos pratiques de gestion du matériel. Réduisez la quantité de matériel nécessaire à allouer et à déployer, et sélectionnez le matériel le plus efficace pour votre charge de travail individuelle.

Utiliser la quantité minimale de matériel pour répondre à vos besoins : en utilisant les fonctionnalités du cloud, vous pouvez apporter régulièrement des modifications à vos mises en œuvre de charges de travail. Mettez à jour les composants déployés à mesure que vos besoins évoluent.

Utiliser les types d'instance ayant le moins d'impact : contrôlez de façon continue le lancement de nouveaux types d'instances et profitez d'améliorations de l'efficacité énergétique, y compris les

types d'instances conçus pour soutenir des charges de travail spécifiques comme l'entraînement et l'inférence du machine learning et le transcodage vidéo.

Utiliser des services gérés : les services gérés permettent de déléguer la responsabilité liée au maintien d'une utilisation moyenne élevée et à l'optimisation de la durabilité du matériel déployé à AWS. Utilisez des services gérés pour distribuer l'impact de la durabilité du service sur tous les locataires du service, ce qui réduit votre contribution individuelle.

Optimiser l'utilisation des unités GPU : les unités de traitement graphique (GPU) peuvent constituer une source de consommation énergétique élevée, et de nombreuses charges de travail de GPU sont très variables, comme le rendu, le transcodage, ainsi que l'entraînement et la modélisation du machine learning. Exécutez uniquement les instances GPU pendant le temps nécessaire et mettez-les hors service grâce à l'automatisation lorsque ce n'est plus le cas afin de réduire les ressources consommées.

Modèles de développement et de déploiement

Recherchez des opportunités de réduire votre impact en matière de durabilité modifiant vos pratiques de développement, de test et de déploiement.

La question suivante est axée sur les considérations relatives à la durabilité :

SUS 6 : Comment vos processus de développement et de déploiement soutiennent-ils vos objectifs de durabilité ?

Recherchez des opportunités de réduire votre impact en matière de durabilité modifiant vos pratiques de développement, de test et de déploiement.

Adopter des méthodes qui peuvent rapidement présenter des améliorations en matière de durabilité : testez et validez les améliorations potentielles avant de les déployer en production. Tenez compte du coût des tests lors du calcul des avantages futurs potentiels d'une amélioration. Développez des méthodes de test à faible coût pour permettre l'apport de petites améliorations.

Garder votre charge de travail à jour : les systèmes d'exploitation, les bibliothèques et les applications à jour peuvent améliorer l'efficacité de la charge de travail et permettent une adoption plus facile des technologies plus efficaces. Les logiciels à jour peuvent également inclure des fonctions permettant de mesurer plus précisément l'impact en matière de durabilité de votre charge de travail, car les fournisseurs proposent des fonctions pour atteindre leurs propres objectifs de durabilité.

Augmenter l'utilisation de vos environnements de création : utilisez l'automatisation et l'infrastructure en tant que code pour mettre en place des environnements de pré-production lorsque cela est nécessaire et les arrêter lorsqu'ils ne sont pas utilisés. Un modèle courant consiste à planifier des périodes de disponibilité qui coïncident avec les heures de travail des membres de votre équipe de développement. La mise en veille prolongée est un outil pratique pour préserver l'état et mettre rapidement des instances en ligne uniquement lorsque cela est nécessaire. Utilisez des types d'instance pouvant transmettre en rafales, des instances Spot, des services de base de données Elastic, des conteneurs et d'autres technologies pour harmoniser la capacité de développement et de test avec l'utilisation.

Utiliser des tests Device Farms gérés pour effectuer les tests : les tests Device Farms gérés répartissent l'impact en matière de durabilité de la fabrication de matériel et de l'utilisation des ressources sur plusieurs locataires. Les tests Device Farms gérés proposent divers types d'appareils afin de vous permettre de prendre en charge du matériel plus ancien et moins courant, et d'éviter que les mises à niveau inutiles d'appareils impactent la durabilité des clients.

Ressources

Consultez les ressources suivantes pour en savoir plus sur nos bonnes pratiques relatives à la durabilité.

Livre blanc

- Pilier Durabilité

Vidéo

- The Climate Pledge

Processus de vérification

La vérification des architectures doit être effectuée de manière cohérente, avec une approche sans faute qui invite à étudier la situation en profondeur. Il doit s'agir d'un processus léger (durant des heures et non des jours), car il s'agit d'une conversation et non d'un audit. L'objectif de l'évaluation d'une architecture est d'identifier les problèmes critiques qui doivent être gérés, ou les domaines qui peuvent être améliorés. Le résultat de la révision est un ensemble d'actions ayant pour objectif d'améliorer l'expérience d'un client à l'aide de la charge de travail.

Comme indiqué dans la section « Sur l'architecture », il faut que chaque membre de l'équipe assume la responsabilité de la qualité de son architecture. Nous recommandons que les membres de l'équipe qui construisent une architecture utilisent le cadre Well-Architected pour revoir continuellement leur architecture, au lieu d'organiser une réunion formelle de révision. Une approche continue permet aux membres de votre équipe de mettre à jour des réponses au fur et à mesure que l'architecture évolue, et d'améliorer l'architecture lorsque vous fournissez des fonctions.

AWS Well-Architected Framework est aligné sur la façon dont AWS vérifie les systèmes et les services en interne. Il repose sur un ensemble de principes de conception qui influent sur l'approche architecturale, et les questions qui garantissent que les gens ne négligent pas des zones qui sont souvent présentées dans l'analyse des causes racines. Chaque fois qu'un problème important lié à un système interne, à un service AWS ou à un client survient, nous nous référons à l'analyse des causes racines pour déterminer s'il est possible d'améliorer les processus de vérification que nous utilisons.

Les vérifications doivent être effectuées aux étapes clés du cycle de vie du produit, dès le début de la phase de conception afin d'éviter les portes unidirectionnelles qui sont difficiles à modifier, puis avant la date de mise en ligne. De nombreuses décisions sont des portes bidirectionnelles réversibles. Ces décisions peuvent utiliser un processus léger. Les portes unidirectionnelles sont difficiles voire impossibles à inverser et nécessitent une inspection plus approfondie avant leur création. Après le lancement de la production, votre charge de travail continuera à évoluer à mesure que de nouvelles fonctions seront ajoutées et que les implémentations technologiques seront modifiées. L'architecture d'une charge de travail change au fil du temps. Vous devez suivre les bonnes pratiques d'hygiène pour empêcher ses caractéristiques architecturales de se dégrader au fur et à mesure de son évolution. Lorsque vous apportez des modifications d'architecture significatives, vous devez suivre un ensemble de processus d'hygiène et procéder à une évaluation Well-Architected.

Si vous souhaitez utiliser l'évaluation en tant qu'instantané unique ou mesure indépendante, vous devez vous assurer que vous avez inclus toutes les bonnes personnes dans la conversation.

Souvent, nous constatons que c'est lors des évaluations qu'une équipe comprend vraiment, pour la première fois, ce qu'elle a mis en œuvre. Une approche qui fonctionne bien lors de l'évaluation d'une autre charge de travail d'équipe est d'avoir une série de conversations informelles sur leur architecture, durant lesquelles vous pouvez recueillir les réponses à la plupart des questions. Vous pouvez ensuite effectuer un suivi avec une ou deux réunions où vous pouvez gagner en clarté, ou des informations complètes sur les domaines d'ambiguïté ou les risques perçus.

Voici quelques suggestions pour faciliter vos réunions :

- Une salle de réunion avec des tableaux blancs
- Des impressions de tous les schémas ou notes de conception
- Une liste de questions qui exigent une recherche hors-bande pour répondre (par exemple, « Avons-nous activé le chiffrement ou pas ? »)

Une fois que vous avez effectué une vérification, vous devez avoir une liste de questions que vous pouvez hiérarchiser en fonction de votre environnement métier. Vous devrez également tenir compte de l'impact de ces problèmes sur le travail quotidien de votre équipe. Si vous traitez ces problèmes tôt, vous pourrez libérer du temps pour travailler sur la création d'une valeur métier au lieu de résoudre des problèmes récurrents. Au fur et à mesure que vous traitez les problèmes, vous pouvez mettre à jour votre vérification pour voir comment l'architecture s'améliore.

Bien que la valeur ajoutée d'une évaluation d'architecture est claire une fois l'exercice terminé, il est possible que vous rencontriez de la résistance de la part d'une nouvelle équipe au début. Voici quelques objections qui peuvent être traitées grâce à la sensibilisation de l'équipe sur les avantages d'une révision :

- « Nous sommes trop occupés ! » (Souvent dit lorsque l'équipe se prépare pour un grand lancement.)
  - Si vous préparez un grand lancement, vous voudrez qu'il se passe bien. La révision vous permettra de comprendre les problèmes que vous pourriez avoir manqués.
  - Nous vous recommandons de réaliser des révisions tôt dans le cycle de vie du produit pour découvrir les risques et développer un plan d'atténuation aligné avec la fonctionnalité de route de livraison.
- « Nous n'avons pas le temps de faire quoi que ce soit avec les résultats ! » (Souvent dit lorsqu'ils ciblent un événement fixe, tel que le Super Bowl.)
  - Ces événements ne peuvent pas être déplacés. Voulez-vous vraiment vous aventurer dans cette entreprise sans connaître les risques liés à votre architecture ? Même si vous ne traitez pas

tous ces problèmes, vous pouvez toujours avoir des stratégies en place pour les gérer s'ils se concrétisent.

- « Nous ne voulons pas que les autres connaissent les secrets de l'implémentation de notre solution ! »
  - Si vous orientez l'équipe vers les questions qui figurent dans l'annexe du livre blanc sur Well-Architected Framework, elle verra qu'aucune des questions ne révèle d'informations propriétaires de nature technique ou commerciale.

Au fur et à mesure que vous effectuez des vérifications avec les équipes au sein de votre entreprise, vous pourrez identifier les problèmes récurrents. Par exemple, vous pourrez voir qu'un groupe d'équipes a rencontré divers problèmes liés à un pilier ou sujet particulier. Il est conseillé d'examiner toutes vos révisions d'une manière globale, et d'identifier tous les mécanismes, formations, ou les discussions d'ingénierie principale qui pourraient aider à traiter ces questions thématiques.

Conclusion

AWS Well-Architected Framework fournit de bonnes pratiques architecturales pour les six piliers de la conception et de l'exploitation de systèmes fiables, sécurisés, efficaces, rentables et durables dans le cloud. Ce cadre fournit un ensemble de questions qui vous permettent d'évaluer une architecture existante ou proposée. Il propose également un ensemble de bonnes pratiques AWS pour chaque pilier. L'utilisation du cadre dans votre architecture vous aidera à produire des systèmes stables et efficaces, qui vous permettent de vous concentrer sur vos exigences fonctionnelles.

Participants

Les personnes et organisations suivantes ont participé à la préparation du présent document :

- Brian Carlson, responsable des opérations de l'équipe Well-Architected, Amazon Web Services
- Ben Potter, responsable sécurité de l'équipe Well-Architected, Amazon Web Services
- Seth Eliot, responsable de la fiabilité de l'équipe Well-Architected, Amazon Web Services
- Eric Pullen, architecte de solutions principal, Amazon Web Services
- Rodney Lester, architecte principal des solutions, Amazon Web Services
- Jon Steele, senior des comptes techniques, Amazon Web Services
- Max Ramsay, architecte principal de solutions de sécurité, Amazon Web Services
- Callum Hughes, architecte de solutions, Amazon Web Services
- Aden Leirer, responsable des programmes de contenu de l'équipe Well-Architected, Amazon Web Services

Autres lectures

Centre d'architecture AWS

Conformité du cloud AWS

Programme de partenariat AWS Well-Architected

AWS Well-Architected Tool

Page d'accueil AWS Well-Architected

livre blanc du pilier Excellence opérationnelle

Livre blanc du pilier Sécurité

Livre blanc du pilier Fiabilité

Livre blanc du pilier Efficacité des performances

Livre blanc du pilier Optimisation des coûts

Livre blanc du pilier Durabilité

Bibliothèque Amazon Builders' Library

Révisions du document

Pour être informé des mises à jour de ce livre blanc, abonnez-vous au flux RSS.

Modification	Description	Date
Mise à jour majeure	Restructuration majeure du pilier Performance pour ramener à cinq le nombre de domaines de bonnes pratiques . Mise à jour importante des bonnes pratiques et des conseils relatifs au pilier Sécurité dans Réponse aux incidents (SÉC 10). Changements de contenu majeurs et consolidation dans les domaines d'excellence opérationnelle OPS 04, 05, 06, 08 et 09. Mises à jour des conseils dans les piliers Optimisation des coûts et Fiabilité . Mises à jour mineures des niveaux de risque du pilier Durabilité .	October 3, 2023
Mises à jour du nouveau cadre	Les bonnes pratiques ont été mises à jour avec des recommandations et de nouvelles bonnes pratiques. De nouvelles questions ont été ajoutées aux piliers Sécurité et Optimisation des coûts.	April 10, 2023
Mise à jour mineure	Ajout d'une définition du niveau d'effort et mise à jour	October 20, 2022

	des bonnes pratiques dans l'annexe.	
Livre blanc mis à jour	Ajout du pilier Durabilité et mise à jour des liens.	December 2, 2021
Mise à jour majeure	Ajout du pilier Durabilité dans le framework.	November 20, 2021
Mise à jour mineure	Suppression du langage non inclusif.	April 22, 2021
Mise à jour mineure	Correction de plusieurs liens.	March 10, 2021
Mise à jour mineure	Modifications rédactionnelles mineures du document.	July 15, 2020
Mises à jour pour le nouveau cadre	Vérification et réécriture de la plupart des questions et réponses.	July 8, 2020
Livre blanc mis à jour	Ajout d'AWS Well-Arch itected Tool et de liens vers les ateliers AWS Well-Arch itected et les partenaires AWS Well-Architected, correctif s mineurs pour prendre en charge la version multilingue du framework.	July 1, 2019

Livre blanc mis à jour	Révision et reformulation de la plupart des questions et réponses, afin que les questions soient axées sur un thème à la fois. Certaines questions précédentes ont été divisées en plusieurs questions. Ajout de termes courants aux définitions (charge de travail, composant , etc.). Modification de la présentation des questions dans le corps principal pour inclure un texte descriptif.	November 1, 2018
Livre blanc mis à jour	Mises à jour pour simplifier le texte, normaliser les réponses et améliorer la lisibilité.	June 1, 2018
Livre blanc mis à jour	L'excellence opérationnelle est déplacée devant les piliers et est récrite afin qu'elle encadre les autres domaines. Actualisa tion des autres piliers pour refléter l'évolution d'AWS.	November 1, 2017
Livre blanc mis à jour	Mise à jour du cadre pour inclure un domaine d'excelle nce opérationnelle, révision et mise à jour des autres piliers pour réduire le dédoublement et intégrer les découvertes obtenues à partir des révisions avec des milliers de clients.	November 1, 2016

Mises à jour mineures	Mise à jour de l'annexe avec les informations Amazon CloudWatch Logs actuelles.	November 1, 2015
Publication initiale	Publication d'AWS Well-Architected Framework.	October 1, 2015

Annexe : questions et bonnes pratiques

Cette annexe résume toutes les questions et les bonnes pratiques du cadre AWS Well-Architected.

Piliers

- Excellence opérationnelle
- Sécurité
- Fiabilité
- Efficacité en matière de performance
- Optimisation des coûts
- Durabilité

Excellence opérationnelle

Le pilier Excellence opérationnelle comprend la capacité à soutenir le développement et à gérer efficacement les charges de travail, à recueillir des informations sur vos opérations, et à améliorer continuellement les processus et procédures de soutien afin de fournir de la valeur métier. Vous trouverez des recommandations sur l'implémentation dans le livre blanc du pilier Excellence opérationnelle.

Domaines de bonnes pratiques

- Organisation
- Préparation
- Exploiter
- Évolution

Organisation

Questions

- OPS 1. Comment déterminer vos priorités ?
- OPS 2. Comment structurez-vous votre organisation pour soutenir les résultats de l'entreprise ?
- OPS 3. Comment votre culture organisationnelle soutient-elle vos résultats opérationnels ?

OPS 1. Comment déterminer vos priorités ?

Chacun doit comprendre le rôle qu'il a à jouer dans la réussite de l'entreprise. Établissez des objectifs partagés afin de définir des priorités pour les ressources. Cela permet de maximiser le fruit de vos efforts.

Bonnes pratiques

- OPS01-BP01 Évaluer les besoins des clients externes
- OPS01-BP02 Évaluer les besoins des clients internes
- OPS01-BP03 Évaluer les exigences de gouvernance
- OPS01-BP04 Évaluer les exigences de conformité
- OPS01-BP05 Évaluer les menaces existantes
- OPS01-BP06 Évaluer les compromis
- OPS01-BP07 Gérer les avantages et les risques

OPS01-BP01 Évaluer les besoins des clients externes

Impliquez les principales parties prenantes, notamment les équipes commerciales, de développement et d'exploitation, pour déterminer où il est nécessaire de concentrer les efforts sur les besoins des clients externes. Vous aurez ainsi une compréhension approfondie du soutien opérationnel nécessaire pour atteindre les résultats opérationnels souhaités.

Anti-modèles courants :

- Vous avez décidé de ne pas bénéficier du service client en dehors des heures de bureau, mais vous n'avez pas examiné les données historiques des demandes d'assistance. Vous ne savez pas si cela aura un impact sur vos clients.
- Vous développez une nouvelle fonctionnalité, mais n'avez pas contacté vos clients pour déterminer si elle est souhaitée, sous quelle forme, et sans expérimentation pour valider le besoin et la méthode de distribution.

Avantages liés au respect de cette bonne pratique : Les clients dont les besoins sont satisfaits sont beaucoup plus susceptibles de rester fidèles. L'évaluation et la compréhension des besoins des clients externes vous permettent d'établir des priorités dans vos efforts pour apporter de la valeur ajoutée à votre entreprise.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Comprendre les besoins de l'entreprise : la réussite s'obtient en établissant des objectifs et une compréhension communs entre les différents acteurs, y compris les équipes commerciales, de développement et d'opérations.
  - Analyser les objectifs de l'entreprise, les besoins et les priorités des clients externes : impliquez les acteurs clés, notamment, les équipes commerciales, du développement et des opérations, pour discuter des objectifs, besoins et priorités des clients externes. Cela permet de vérifier que vous comprenez bien le soutien opérationnel requis pour atteindre les résultats de l'entreprise et des clients.
  - Établir une compréhension commune : établissez une compréhension commune des fonctions opérationnelles de la charge de travail, des rôles de chacune des équipes dans l'exploitation de la charge de travail, et de la manière dont ces facteurs soutiennent les objectifs opérationnels partagés chez les clients internes et externes.

Ressources

Documents connexes :

- Concepts AWS Well-Architected Framework – Boucle de rétroaction

OPS01-BP02 Évaluer les besoins des clients internes

Impliquez les principales parties prenantes, notamment les équipes commerciales, de développement et d'exploitation, lorsqu'il s'agit de déterminer où il est nécessaire de concentrer les efforts sur les besoins des clients internes. Ainsi, vous aurez une connaissance approfondie du soutien opérationnel requis pour atteindre les résultats opérationnels.

Tenez compte des priorités que vous avez établies pour concentrer vos efforts d'amélioration là où ils auront le plus d'impact (par exemple, le développement des compétences de l'équipe, l'amélioration des performances des charges de travail, la réduction des coûts, l'automatisation des runbooks ou encore l'amélioration de la surveillance). Mettez à jour vos priorités en fonction de vos besoins.

Anti-modèles courants :

- Vous avez décidé de modifier l'attribution des adresses IP de vos équipes de produits sans les consulter, afin de faciliter la gestion de votre réseau. Vous ne connaissez pas l'impact que cela aura sur vos équipes de produits.

- Vous mettez en place un nouvel outil de développement, mais vous n'avez pas demandé à vos clients internes s'ils en ont besoin ou s'il est compatible avec leurs pratiques existantes.
- Vous mettez en place un nouveau système de surveillance, mais vous demandez à vos clients internes s'ils ont des besoins en matière de surveillance ou de rapports à prendre en compte.

Avantages liés au respect de cette bonne pratique : L'évaluation et la compréhension des besoins des clients internes vous permettent d'établir des priorités dans vos efforts pour apporter de la valeur ajoutée à votre entreprise.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Comprendre les besoins de l'entreprise : la réussite s'obtient en établissant des objectifs et une compréhension communs entre les différents acteurs, y compris les équipes commerciales, de développement et d'opérations.
  - Analyser les objectifs de l'entreprise, les besoins et les priorités des clients internes : impliquez les acteurs clés, notamment, les équipes commerciales, du développement et des opérations, pour discuter des objectifs, besoins et priorités des clients internes. Cela permet de vérifier que vous comprenez bien le soutien opérationnel requis pour atteindre les résultats de l’entreprise et des clients.
  - Établir une compréhension commune : établissez une compréhension commune des fonctions opérationnelles de la charge de travail, des rôles de chacune des équipes dans l'exploitation de la charge de travail, et de la manière dont ces facteurs soutiennent les objectifs opérationnels partagés chez les clients internes et externes.

Ressources

Documents connexes :

- Concepts AWS Well-Architected Framework – Boucle de rétroaction

OPS01-BP03 Évaluer les exigences de gouvernance

La gouvernance désigne l'ensemble des politiques, règles ou cadres qu'une entreprise utilise pour atteindre ses objectifs commerciaux. Les exigences en matière de gouvernance sont générées au sein de votre organisation. Elles peuvent affecter les types de technologies que vous choisissez ou

influencer la façon dont vous gérez votre charge de travail. Incorporez les exigences de gouvernance organisationnelle dans votre charge de travail. La conformité désigne la capacité à prouver que vous avez mis en œuvre les exigences de gouvernance.

Résultat souhaité :

- Les exigences de gouvernance sont intégrées à la conception architecturale et au fonctionnement de votre charge de travail.
- Vous pouvez fournir la preuve que vous avez suivi les exigences de gouvernance.
- Les exigences en matière de gouvernance sont régulièrement revues et mises à jour.

Anti-modèles courants :

- Votre organisation exige que le compte racine dispose d'une authentification multi-facteur. Vous n'avez pas mis en œuvre cette exigence et le compte racine est compromis.
- Lors de la conception de votre charge de travail, vous choisissez un type d'instance qui n'est pas approuvé par le service informatique. Vous ne parvenez pas à lancer votre charge de travail et devez procéder à une refonte.
- Vous êtes tenu de préparer un plan de reprise après sinistre. Vous n'en avez pas créé et votre charge de travail subit une interruption prolongée.
- Votre équipe souhaite utiliser de nouvelles instances mais vos exigences de gouvernance n'ont pas été mises à jour pour les autoriser.

Avantages liés au respect de cette bonne pratique :

- Le respect des exigences de gouvernance permet d'aligner votre charge de travail sur les politiques de l'organisation dans son ensemble.
- Les exigences en matière de gouvernance reflètent les normes industrielles et les bonnes pratiques de votre organisation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Identifiez les besoins en matière de gouvernance en travaillant avec les parties prenantes et les organisations de gouvernance. Incorporez les exigences de gouvernance à votre charge de travail. Soyez en mesure de prouver que vous avez respecté les exigences de gouvernance.

Exemple de client

Chez AnyCompany Retail, l'équipe chargée des opérations dans le cloud collabore avec les parties prenantes de toute l'organisation pour élaborer des exigences de gouvernance. Par exemple, ils interdisent l'accès SSH aux instances Amazon EC2. Si les équipes doivent accéder au système, elles doivent utiliser AWS Systems Manager Session Manager. L'équipe chargée des opérations dans le cloud met régulièrement à jour les exigences de gouvernance à mesure que de nouveaux services sont disponibles.

Étapes d'implémentation

1. Identifiez les parties prenantes de votre charge de travail, y compris toute équipe centralisée.
2. Travaillez avec les parties prenantes pour identifier les exigences de gouvernance.
3. Une fois que vous avez dressé une liste, classez les points à améliorer par ordre de priorité et commencez à les mettre en œuvre dans votre charge de travail.
   a. Utilisez des services tels que AWS Config pour créer un code de gouvernance et pour valider le respect des exigences de gouvernance.
   b. Si vous utilisez AWS Organizations, vous pouvez tirer parti des politiques de contrôle des services pour mettre en œuvre les exigences de gouvernance.
4. Fournissez la documentation qui valide la mise en œuvre.

Niveau d'effort du plan d'implémentation : moyen. La mise en œuvre des exigences de gouvernance manquantes peut entraîner une refonte de votre charge de travail.

Ressources

Bonnes pratiques associées :

- OPS01-BP04 Évaluer les exigences de conformité : la conformité est comparable à la gouvernance mais vient de l'extérieur de l'organisation.

Documents connexes :

- AWS Management and Governance Cloud Environment Guide (Guide de gestion et de gouvernance de l'environnement cloud AWS)

- Best Practices for AWS Organizations Service Control Policies in a Multi-Account Environment (Bonnes pratiques pour les stratégies de contrôle des services d'AWS Organizations dans un environnement multi-comptes)
- Governance in the AWS Cloud: The Right Balance Between Agility and Safety (La gouvernance dans le cloud AWS : le juste équilibre entre agilité et sécurité)
- Qu'est-ce que la gouvernance, le risque et la conformité (GRC) ?

Vidéos connexes :

- Gestion et gouvernance AWS : configuration, conformité et audit - AWS Online Tech Talks
- AWS re:Inforce 2019: Governance for the Cloud Age (DEM12-R1) [AWS re:Inforce 2019 : la gouvernance à l'ère du cloud (DEM12-R1)]
- AWS re:Invent 2020: Achieve compliance as code using AWS Config (AWS re:Invent 2020 : mettre en œuvre la conformité en tant que code en utilisant AWS Config)
- AWS re:Invent 2020: Agile governance on AWS GovCloud (US) [AWS re:Invent 2020 : la gouvernance agile sur AWS GovCloud (US)]

Exemples connexes :

- Exemples de packs de conformité AWS Config

Services associés :

- AWS Config
- AWS Organizations : politiques de contrôle de services

OPS01-BP04 Évaluer les exigences de conformité

Les exigences en matière de conformité réglementaire, sectorielle et interne constituent un facteur important pour définir les priorités de votre organisation. Votre cadre de conformité peut vous empêcher d'utiliser des technologies ou des emplacements géographiques spécifiques. Appliquez les principes de diligence raisonnable si aucun cadre de conformité externe n'est identifié. Générez des audits ou des rapports qui valident la conformité.

Si vous mettez en avant le fait que votre produit respecte des normes de conformité spécifiques, vous devez mettre en place un processus interne pour assurer une conformité constante. Les

normes PCI DSS, FedRAMP et HIPAA sont des exemples de normes de conformité. Les normes de conformité applicables sont déterminées par divers facteurs, tels que les types des données stockées ou transmises par la solution et les régions géographiques prises en charge par la solution.

Résultat souhaité :

- Les exigences en matière de conformité réglementaire, industrielle et interne sont intégrées dans le choix de l'architecture.
- Vous pouvez valider la conformité et générer des rapports d'audit.

Anti-modèles courants :

- Certaines parties de votre charge de travail relèvent du cadre de la norme de sécurité des données de l'industrie des cartes de paiement (PCI DSS), mais votre charge de travail stocke les données des cartes de crédit sans les chiffrer.
- Vos développeurs et architectes de logiciels ne connaissent pas le cadre de conformité auquel votre organisation doit se conformer.
- L'audit annuel SOC2 (Systems and Organizations Control) de type II aura lieu prochainement et vous n'êtes pas en mesure de vérifier que les contrôles sont en place.

Avantages liés au respect de cette bonne pratique :

- L'évaluation et la compréhension des exigences de conformité qui s'appliquent à votre charge de travail détermineront la façon dont vous priorisez vos efforts pour produire de la valeur ajoutée.
- Vous choisissez les bons sites et les bonnes technologies, en accord avec votre cadre de conformité.
- La conception de votre charge de travail en vue de son auditabilité vous permet de prouver que vous adhérez à votre cadre de conformité.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

La mise en œuvre de cette bonne pratique signifie que vous intégrez les exigences de conformité dans votre processus de conception de l'architecture. Les membres de votre équipe connaissent le cadre de conformité requis. Vous validez la conformité conformément au cadre.

Exemple de client

AnyCompany Retail stocke les informations relatives aux cartes de crédit des clients. Les développeurs de l'équipe chargée du stockage des cartes comprennent qu'ils doivent se conformer au cadre PCI-DSS. Ils ont pris des mesures pour vérifier que les informations relatives aux cartes de crédit sont stockées et accessibles en toute sécurité, conformément au cadre PCI-DSS. Chaque année, ils travaillent avec leur équipe de sécurité pour valider la conformité.

Étapes d'implémentation

1. Travaillez avec vos équipes de sécurité et de gouvernance pour déterminer les cadres de conformité sectoriels, réglementaires ou internes auxquels votre charge de travail doit se conformer. Incorporez les cadres de conformité à votre charge de travail.
   a. Validez la conformité constante des ressources AWS avec des services comme AWS Compute Optimizer et AWS Security Hub.
2. Informez les membres de votre équipe sur les exigences de conformité afin qu'ils puissent travailler et faire évoluer la charge de travail en fonction de celles-ci. Les exigences de conformité doivent être incorporées aux choix architecturaux et technologiques.
3. En fonction du cadre de conformité, vous pouvez être amené à générer un audit ou un rapport de conformité. Travaillez avec votre organisation pour automatiser ce processus autant que possible.
   a. Utilisez des services tels que AWS Audit Manager pour valider la conformité et générer des rapports d'audit.
   b. Vous pouvez télécharger des documents sur la sécurité et la conformité d'AWS avec AWS Artifact.

Niveau d'effort du plan d'implémentation : moyen. La mise en œuvre de cadres de conformité peut s'avérer difficile. La production de rapports d'audit ou de documents de conformité ajoute un niveau de complexité supplémentaire.

Ressources

Bonnes pratiques associées :

- SEC01-BP03 Identifier et valider les objectifs de contrôle : les objectifs de contrôle de la sécurité sont une part importante de la conformité globale.

- SEC01-BP06 Automatiser les tests et la validation des contrôles de sécurité dans les pipelines : dans le cadre de vos pipelines, validez les contrôles de sécurité. Vous pouvez également générer des documents de conformité pour les nouvelles modifications.
- SEC07-BP02 Définir les contrôles de protection des données : de nombreux cadres de conformité reposent sur des politiques de traitement et de stockage des données.
- SEC10-BP03 Préparer les fonctionnalités d'analyse poussée : les capacités d'analyse permettent parfois de vérifier la conformité.

Documents connexes :

- Centre de conformité AWS
- Ressources de conformité AWS
- Livre blanc Risques et conformité AWS
- Modèle de responsabilité partagée d'AWS
- Services AWS dans l'étendue par programmes de conformité

Vidéos connexes :

- AWS re:Invent 2020: Achieve compliance as code using AWS Compute Optimizer (AWS re:Invent 2020 : mettre en œuvre la conformité en tant que code en utilisant AWS Config)
- AWS re:Invent 2021 - Cloud compliance, assurance, and auditing (AWS re:Invent 2021 : conformité, assurance et audit du cloud)
- AWS Summit ATL 2022 - Implementing compliance, assurance, and auditing on AWS (COP202) [Sommet AWS ATL 2022 : mise en œuvre de la conformité, de l'assurance et de l'audit sur AWS (COP202)]

Exemples connexes :

- PCI DSS and AWS Foundational Security Best Practices on AWS (Bonnes pratiques en matière de sécurité de base PCI DSS et AWS sur le cloud AWS)

Services associés :

- AWS Artifact
- AWS Audit Manager

- AWS Compute Optimizer
- AWS Security Hub

OPS01-BP05 Évaluer les menaces existantes

Évaluez les menaces pesant sur l'entreprise (par exemple, la concurrence, les risques commerciaux et les responsabilités, les risques opérationnels et les menaces sur la sécurité des informations) et tenez à jour les informations dans un registre des risques. Incluez l'impact des risques pour déterminer où concentrer les efforts.

La version Le cadre AWS Well-Architected met l'accent sur l'apprentissage, les évaluations et l'amélioration. Il vous fournit une approche cohérente pour évaluer les architectures et mettre en œuvre des conceptions qui évoluent dans le temps. AWS fournit l'outil AWS Well-Architected Tool pour vous aider à vérifier votre approche avant le développement et l'état de vos charges de travail avant et pendant la production. Vous pouvez les comparer aux dernières bonnes pratiques architecturales AWS, surveiller l'état général de vos charges de travail et avoir un aperçu des risques potentiels.

Les clients AWS peuvent bénéficier d'une vérification guidée Well-Architected de leurs charges de travail stratégiques afin d'évaluer la conformité de leurs architectures par rapport aux bonnes pratiques AWS. Les clients ayant souscrit au programme Enterprise Support peuvent bénéficier d'une vérification des opérations,conçue pour les aider à identifier les failles de leur approche d'exploitation dans le cloud.

L'implication des équipes dans ces vérifications contribue à établir une compréhension partagée de vos charges de travail et de la façon dont les rôles de chacun contribuent à la réussite de l'équipe. Les besoins identifiés par la vérification peuvent vous aider à définir vos priorités.

AWS Trusted Advisor est un outil qui donne accès à un ensemble de base de vérifications qui recommandent des optimisations pouvant vous aider à définir vos priorités. Clients ayant un plan de support Business ou Enterprise Les clients ayant un plan de support Business ou Enterprise ont accès à des vérifications supplémentaires axées sur la sécurité, la fiabilité, les performances et l'optimisation des coûts, qui peuvent les aider à définir leurs priorités.

Anti-modèles courants :

- Vous utilisez une ancienne version d'une bibliothèque de logiciels dans votre produit. Vous n'êtes pas au courant des mises à jour de sécurité de la bibliothèque pour les questions qui peuvent avoir un impact involontaire sur votre charge de travail.

- Votre concurrent vient de lancer une version de son produit qui répond aux nombreuses plaintes de vos clients concernant votre produit. Vous n'avez pas priorisé la résolution des problèmes connus.
- Les régulateurs ont poursuivi des entreprises comme la vôtre qui ne respectaient pas les exigences légales de conformité réglementaire. Vous n'avez pas priorisé la résolution des vos exigences de conformité en suspens.

Avantages liés au respect de cette bonne pratique : L'identification et la compréhension des menaces qui pèsent sur votre organisation et votre charge de travail vous permettent de déterminer les menaces à traiter, leur priorité et les ressources nécessaires pour y parvenir.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Évaluer les menaces existantes : évaluez les menaces qui pèsent sur l'entreprise (par exemple, la concurrence, les risques commerciaux et les responsabilités, les risques opérationnels et les menaces sur la sécurité des données) afin de pouvoir tenir compte de leur impact lorsque vous déterminez où concentrer vos efforts.
  - Derniers bulletins de sécurité AWS
  - AWS Trusted Advisor
  - Gérer un modèle de menace : établissez et gérez un modèle de menace identifiant les menaces potentielles, les mesures d'atténuation prévues et en place, et leur priorité. Examinez la probabilité que les menaces se manifestent par des incidents, le coût de la récupération après ces incidents, le préjudice attendu et le coût de la prévention de ces incidents. Modifiez les priorités au fur et à mesure que le contenu du modèle de menace change.

Ressources

Documents connexes :

- Conformité du AWS Cloud
- Derniers bulletins de sécurité AWS
- AWS Trusted Advisor

OPS01-BP06 Évaluer les compromis

Évaluez l'impact des compromis entre des intérêts concurrents ou des approches alternatives pour prendre des décisions éclairées au moment de déterminer où concentrer les efforts ou choisir une ligne de conduite. Par exemple, l'accélération de la mise sur le marché de nouvelles fonctionnalités peut être privilégiée par rapport à l'optimisation des coûts, ou vous pouvez choisir une base de données relationnelle pour les données non relationnelles afin de simplifier l'effort de migration d'un système, plutôt que de migrer vers une base de données optimisée pour votre type de données et de mettre à jour votre application.

AWS peut vous aider à former vos équipes à AWS et à ses services afin qu'elles comprennent mieux comment leurs choix peuvent avoir un impact sur votre charge de travail. Vous pouvez utiliser les ressources fournies par AWS Support (le centre de connaissances AWS, les forums de discussion AWSet le centre AWS Support) et la documentation AWS pour former vos équipes. Contactez AWS Support via le centre AWS Support pour obtenir des réponses à vos questions AWS.

AWS partage également les bonnes pratiques et les modèles tirés de l'expérience AWS dans la Bibliothèque Amazon Builders' Library. Une grande variété d'autres informations utiles sont disponibles via le blog AWS et le podcast AWS officiel.

Anti-modèles courants :

- Vous utilisez une base de données relationnelle pour gérer les séries chronologiques et les données non relationnelles. Il existe des options de base de données qui sont optimisées pour prendre en charge les types de données que vous utilisez, mais vous ne connaissez pas les avantages, car vous n'avez pas évalué les compromis entre les solutions.
- Vos investisseurs vous demandent de prouver que vous respectez les normes de sécurité des données du secteur des cartes de paiement (PCI DSS). Vous n'envisagez pas les compromis entre la satisfaction de leur demande et la poursuite de vos efforts de développement actuels. Au lieu de cela, vous poursuivez vos efforts de développement sans en démontrer la conformité. Vos investisseurs cessent de soutenir votre entreprise en raison de préoccupations concernant la sécurité de votre plate-forme et de leurs investissements.

Avantages liés au respect de cette bonne pratique : Comprendre les implications et les conséquences de vos choix permet de prioriser vos options.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Évaluer les compromis : évaluez l'impact des compromis entre des intérêts concurrents afin de prendre des décisions éclairées lorsqu'il s'agit de déterminer où concentrer les efforts. Par exemple, l'accélération de la mise sur le marché de nouvelles fonctionnalités pourrait être privilégiée par rapport à l'optimisation des coûts.
- AWS peut vous aider à former vos équipes à AWS et à ses services afin qu'elles comprennent mieux comment leurs choix peuvent avoir un impact sur votre charge de travail. Vous devez utiliser les ressources fournies par AWS Support (Centre de connaissances AWS, forums de discussion AWSet AWS Support Center) et la documentation AWS pour former vos équipes. Contactez AWS Support via le centre AWS Support pour obtenir des réponses à vos questions AWS.
- AWS partage également les bonnes pratiques et les modèles que nous avons appris grâce à l'exploitation d'AWS dans Amazon Builders' Library. Un grand nombre d'autres informations utiles sont disponibles sur le blog AWS et sur le podcast officiel AWS.

Ressources

Documents connexes :

- blog AWS
- Conformité du AWS Cloud
- les forums de discussion AWS
- documentation AWS
- le centre de connaissances AWS
- AWS Support
- le centre AWS Support
- Bibliothèque Amazon Builders' Library
- le podcast AWS officiel

OPS01-BP07 Gérer les avantages et les risques

Gérez les avantages et les risques afin de prendre des décisions éclairées lorsqu'il s'agit de déterminer où il est nécessaire de concentrer les efforts. Par exemple, il peut être avantageux de déployer une charge de travail comportant des problèmes non résolus afin que de nouvelles

fonctionnalités importantes puissent être mises à la disposition des clients. Il peut être possible d'atténuer les risques associés, ou il peut devenir inacceptable de laisser un risque subsister, auquel cas vous prendrez des mesures pour y remédier.

Vous pouvez décider à un moment donné de mettre l'accent sur un petit sous-ensemble de priorités opérationnelles. Utilisez une approche équilibrée sur le long terme pour garantir le développement des capacités nécessaires et de la gestion des risques. Mettez à jour vos priorités en fonction de vos besoins.

Anti-modèles courants :

- Vous avez décidé d'inclure une bibliothèque qui fait « tout ce dont vous avez besoin », une bibliothèque que l'un de vos développeurs « a trouvé sur Internet ». Vous n'avez pas évalué les risques d'adoption de cette bibliothèque d’une source inconnue et ne savez pas si elle contient des failles ou du code malveillant.
- Vous avez décidé de développer et de déployer une nouvelle fonctionnalité au lieu de résoudre un problème existant. Vous n'avez pas évalué les risques de ne pas traiter le problème jusqu'au déploiement de la fonctionnalité et ne savez pas quel sera l'impact sur vos clients.
- Vous avez décidé de ne pas déployer une fonctionnalité fréquemment demandée par les clients en raison de préoccupations non spécifiées de votre équipe de conformité.

Avantages liés au respect de cette bonne pratique : Identifier les avantages de vos choix et connaître les risques pour votre organisation permet de prendre des décisions en connaissance de cause.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Gérer les avantages et les risques : trouvez un juste milieu entre les avantages des décisions et les risques impliqués.
  - Identifier les avantages : identifiez les avantages en fonction des objectifs, des besoins et des priorités de l'entreprise. Les exemples incluent les délais de commercialisation, la sécurité, la fiabilité, la performance et les coûts.
  - Identifier les risques : identifiez les risques en fonction des objectifs, des besoins et des priorités de l'entreprise. Les exemples incluent les délais de commercialisation, la sécurité, la fiabilité, la performance et les coûts.
  - Évaluer les avantages par rapport aux risques et prendre des décisions avisées : déterminez l'impact des avantages et des risques en fonction des objectifs, des besoins et des priorités de

vos acteurs clés, notamment les équipes commerciales, le développement et les opérations. Évaluez la valeur ajoutée de l'avantage par rapport à la probabilité de réalisation du risque et au coût de son impact. Par exemple, mettre l'accent sur la rapidité de mise sur le marché plutôt que sur la fiabilité pourrait fournir un avantage concurrentiel. Toutefois, cela peut entraîner une réduction du temps de fonctionnement en cas de problèmes de fiabilité.

OPS 2. Comment structurez-vous votre organisation pour soutenir les résultats de l'entreprise ?

Vos équipes doivent comprendre leur rôle dans l'obtention des résultats de l'entreprise. Les équipes doivent comprendre leur rôle dans la réussite des autres équipes, le rôle des autres équipes dans leur réussite, et avoir des objectifs communs. Comprendre la responsabilité, la manière dont les décisions sont prises et qui a le pouvoir de prendre des décisions vous aide à concentrer les efforts et à maximiser les avantages de vos équipes.

Bonnes pratiques

- OPS02-BP01 Les ressources ont des propriétaires identifiés
- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés
- OPS02-BP03 Les activités opérationnelles ont des propriétaires identifiés responsables de leurs performances
- OPS02-BP04 Les membres de l'équipe savent de quoi ils sont responsables
- OPS02-BP05 Des mécanismes sont en place pour identifier la responsabilité et la propriété
- OPS02-BP06 Des mécanismes sont en place pour demander des ajouts, des modifications et des dérogations
- OPS02-BP07 Les responsabilités entre les équipes sont prédéfinies ou négociées

OPS02-BP01 Les ressources ont des propriétaires identifiés

Les ressources de votre charge de travail doivent disposer de propriétaires identifiés pour le contrôle des modifications, le dépannage et d'autres fonctions. Des propriétaires sont désignés pour les charges de travail, les comptes, l'infrastructure, les plateformes et les applications. La propriété est enregistrée à l'aide d'outils tels qu'un registre central ou des métadonnées attachées aux ressources. La valeur commerciale des composants informe les processus et les procédures qui leur sont appliqués.

Résultat souhaité :

- Les ressources disposent de propriétaires identifiés à l'aide de métadonnées ou d'un registre central.
- Les membres de l'équipe peuvent identifier qui est propriétaire des ressources.
- Les comptes disposent d'un propriétaire unique dans la mesure du possible.

Anti-modèles courants :

- Les contacts alternatifs pour vos Comptes AWS ne sont pas remplis.
- Les ressources manquent de balises permettant d'identifier les équipes qui les possèdent.
- Vous avez une file d'attente ITSM sans mappage d'e-mail.
- Deux équipes se partagent la propriété d'un élément d'infrastructure critique.

Avantages liés au respect de cette bonne pratique :

- Le contrôle des modifications pour les ressources est simple et la propriété est attribuée.
- Vous pouvez impliquer les bons propriétaires lors du dépannage des problèmes.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Définissez ce que signifie la propriété pour les cas d'utilisation des ressources dans votre environnement. La propriété peut signifier qui supervise les modifications apportées à la ressource, qui prend en charge la ressource pendant le dépannage, ou qui est financièrement responsable. Précisez et enregistrez les propriétaires des ressources, y compris, le nom, les coordonnées, l'organisation et l'équipe.

Exemple de client

AnyCompany Retail définit la propriété comme l'équipe ou l'individu responsable des modifications apportées aux ressources et de leur prise en charge. Ces personnes utilisent AWS Organizations pour gérer leurs Comptes AWS. Les contacts des comptes alternatifs sont configurés via des boîtes de réception de groupe. Chaque file d'attente ITSM correspond à un alias e-mail. Les balises permettent d'identifier les propriétaires des ressources AWS. Pour les autres plateformes et infrastructures, ces personnes disposent d'une page wiki qui identifie les propriétaires et les informations de contact.

Étapes d'implémentation

1. Commencez par définir la propriété dans votre organisation. La propriété peut impliquer qui est responsable du risque pour la ressource, qui est responsable des modifications apportées à la ressource, ou qui prend en charge la ressource lors du dépannage. La propriété peut également impliquer la propriété financière ou administrative de la ressource.
2. Utilisez AWS Organizations pour gérer les comptes. Vous pouvez gérer les contacts alternatifs de vos comptes de manière centralisée.
   a. Grâce aux adresses e-mail et aux numéros de téléphone appartenant à l'entreprise, vous pourrez y accéder même si les personnes qui les consultent ne font plus partie de votre entreprise. Par exemple, créez des listes de distribution d'e-mails distinctes pour la facturation, les opérations et la sécurité, et configurez-les en tant que contact Facturation, Sécurité et Opérations dans chaque Compte AWS actif. Plusieurs personnes recevront les notifications AWS et seront en mesure de répondre, même si l'une de ces personnes est en congés, change de poste ou quitte la société.
   b. Si un compte n'est pas géré par AWS Organizations, les contacts alternatifs de compte aident AWS à contacter le personnel approprié si nécessaire. Configurez les contacts alternatifs du compte pour qu'ils pointent vers un groupe plutôt que vers un individu.
3. Utilisez des balises pour identifier les propriétaires des ressources AWS. Vous pouvez indiquer les deux propriétaires et leurs coordonnées dans des balises distinctes.
   a. Vous pouvez utiliser des règles AWS Config pour faire en sorte que les ressources disposent des balises de propriété requises.
   b. Pour obtenir des conseils détaillés sur la manière d'élaborer une stratégie de balisage pour votre organisation, consultez le livre blanc des bonnes pratiques de balisage AWS.
4. Pour les autres ressources, plateformes et infrastructures, créez une documentation qui identifie la propriété. Tous les membres de l'équipe doivent y avoir accès.

Niveau d'effort du plan d'implémentation : faible. Exploitez les informations de contact et les balises des comptes pour attribuer la propriété des ressources AWS. Pour les autres ressources, vous pouvez utiliser quelque chose de simple comme un tableau dans un wiki pour enregistrer la propriété et les informations de contact, ou utiliser un outil ITSM pour cartographier la propriété.

Ressources

Bonnes pratiques associées :

- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés : les processus et procédures de prise en charge des ressources dépendent de la propriété des ressources.
- OPS02-BP04 Les membres de l'équipe savent de quoi ils sont responsables : les membres de l'équipe doivent comprendre de quelles ressources ils sont propriétaires.
- OPS02-BP05 Des mécanismes sont en place pour identifier la responsabilité et la propriété : la propriété doit pouvoir être découverte à l'aide de mécanismes tels que les balises ou les contacts de compte.

Documents connexes :

- AWS Account Management - Updating contact information (Gestion des comptes AWS - Mise à jour des informations de contact)
- AWS Config Rules - required-tags (Règles AWS Config - balises obligatoires)
- AWS Organizations - Mise à jour d'autres contacts de votre organisation
- Livre blanc des bonnes pratiques de balisage AWS

Exemples connexes :

- AWS Config Rules - Amazon EC2 with required tags and valid values (Règles AWS Config - EC2 avec les bonnes balises et valeurs)

Services associés :

- AWS Config
- AWS Organizations

OPS02-BP02 Les processus et procédures ont des propriétaires identifiés

Déterminez qui est propriétaire de la définition des différents processus et procédures individuels, pourquoi ces processus et procédures sont utilisés et pourquoi cette propriété existe. La détermination des raisons pour lesquelles des processus et des procédures spécifiques sont utilisés permet d'identifier les possibilités d'amélioration.

Avantages liés au respect de cette bonne pratique : La détermination de la propriété permet d'identifier qui peut approuver les améliorations, les mettre en œuvre ou les deux.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Les processus et procédures ont des propriétaires identifiés responsables de leur définition : capturez les processus et procédures utilisés dans votre environnement, ainsi que la personne ou l'équipe responsable de leur définition.
  - Identifier les processus et les procédures : identifiez les activités opérationnelles réalisées à l'aide de vos charges de travail. Documentez ces activités dans un emplacement détectable.
  - Déterminer qui est responsable de la définition d'un processus ou d'une procédure : identifiez de façon unique l'individu ou l'équipe responsable de la spécification d'une activité. Il leur incombe de s'assurer qu'elle peut être exécutée avec succès par un membre de l'équipe disposant des autorisations, des accès et des outils appropriés. En cas de problème lié à l'exécution de l'activité, les membres de l'équipe qui l'exécutent sont tenus de fournir les commentaires détaillés nécessaires à son amélioration.
  - Capturer la propriété dans les métadonnées de l'artefact d'activité : les procédures automatisées dans des services tels qu'AWS Systems Manager, via des documents et AWS Lambda, en tant que fonctions, prennent en charge la capture des informations de métadonnées sous forme de balises. Capturez la propriété des ressources à l'aide de balises ou de groupes de ressources, en spécifiant les informations de propriété et de contact. Utilisez AWS Organizations pour créer des politiques de balisage et vous assurer que les informations de propriété et de contact sont capturées.

OPS02-BP03 Les activités opérationnelles ont des propriétaires identifiés responsables de leurs performances

Déterminez qui est responsable d'exécuter des activités spécifiques sur des charges de travail définies et pourquoi cette responsabilité existe. La détermination de qui est responsable de l'exécution des activités indique qui va mener l'activité, valider le résultat et fournir des commentaires au propriétaire de l'activité.

Avantages liés au respect de cette bonne pratique : La détermination de qui est responsable de l'exécution d'une activité indique qui doit être notifier quand une action est nécessaire et qui doit exécuter l'action, valider le résultats et fournir des commentaires au propriétaire de l'activité.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Les activités opérationnelles ont des propriétaires identifiés responsables de leurs performances . capturez la responsabilité de l'exécution des processus et procédures utilisés dans votre environnement.
  - Identifier les processus et les procédures : identifiez les activités opérationnelles réalisées à l'aide de vos charges de travail. Documentez ces activités dans un emplacement détectable.
  - Définir qui est responsable de chaque activité : identifiez l'équipe responsable d'une activité. Assurez-vous qu'elle dispose des informations de l'activité, des compétences nécessaires et des autorisations et outils corrects pour exécuter l'activité. Elle doit comprendre la condition de son exécution (par exemple, en cas d'événement ou selon un calendrier). Rendez ces informations accessibles afin que les membres de votre organisation puissent identifier les personnes qu'ils doivent contacter, équipe ou personne, pour des besoins spécifiques.

OPS02-BP04 Les membres de l'équipe savent de quoi ils sont responsables

La compréhension des responsabilités de votre rôle et de la manière dont vous contribuez aux résultats de l'entreprise permet de définir les priorités de vos tâches et de comprendre pourquoi votre rôle est important. Cela permet aux membres d'équipe de reconnaître les besoins et de répondre de manière appropriée.

Avantages liés au respect de cette bonne pratique : la compréhension de vos responsabilités indique les décisions que vous prenez, les actions que vous exécutez et vos activités à leurs propriétaires.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

- Veiller à ce que les membres de l'équipe comprennent leurs rôles et responsabilités : identifiez les rôles et responsabilités des membres de l'équipe et assurez-vous qu'ils comprennent les attentes liées à leur rôle. Rendez ces informations accessibles afin que les membres de votre organisation puissent identifier les personnes qu'ils doivent contacter, équipe ou personne, pour des besoins spécifiques.

OPS02-BP05 Des mécanismes sont en place pour identifier la responsabilité et la propriété

Lorsque aucune personne ou équipe n'est identifiée, il existe des chemins de remontée vers une personne ayant le pouvoir d'attribuer la propriété ou le plan pour traiter le besoin.

Avantages liés au respect de cette bonne pratique : Savoir qui est responsable ou propriétaire permet de faire appel à l'équipe ou au membre de l'équipe approprié pour faire une demande ou transférer une tâche. Le fait d'avoir une personne identifiée qui est autorisée à attribuer la responsabilité ou la propriété ou à planifier pour répondre aux besoins réduit le risque d'inaction et de voir les besoins non satisfaits.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Des mécanismes sont en place pour identifier la responsabilité et la propriété : proposez des mécanismes accessibles aux membres de votre organisation pour découvrir et identifier la propriété et la responsabilité. Ils leur permettront d'identifier les personnes à contacter, équipe ou individu, pour des besoins spécifiques.

OPS02-BP06 Des mécanismes sont en place pour demander des ajouts, des modifications et des dérogations

Vous pouvez adresser des demandes aux propriétaires des processus, des procédures et des ressources. Les demandes comprennent les ajouts, les modifications et les exceptions. Ces demandes sont soumises à un processus de gestion des modifications. Prenez des décisions avisées pour approuver les demandes lorsque celles-ci sont viables et appropriées après une évaluation des avantages et des risques.

Résultat souhaité :

- Vous pouvez faire des demandes de modification des processus, des procédures et des ressources en fonction de la propriété attribuée.
- Les modifications sont réalisées de manière délibérée, en pesant les avantages et les risques.

Anti-modèles courants :

- Vous devez mettre à jour la façon dont vous déployez votre application, mais il n'existe aucun moyen de demander à l'équipe chargée des opérations de modifier le processus de déploiement.
- Le plan de reprise après sinistre doit être mis à jour, mais il n'y a aucun propriétaire désigné à qui demander des modifications.

Avantages liés au respect de cette bonne pratique :

- Les processus, les procédures et les ressources peuvent évoluer au fur et à mesure que les exigences évoluent.
- Les propriétaires peuvent décider en connaissance de cause du moment où il convient d'apporter des modifications.
- Les modifications sont réalisées de manière délibérée.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Pour mettre en œuvre cette bonne pratique, vous devez être en mesure de demander des modifications des processus, des procédures et des ressources. Le processus de gestion des modifications peut être léger. Documenter le processus de gestion des modifications.

Exemple de client

AnyCompany Retail utilise une matrice d'attribution des responsabilités (RACI) pour identifier qui est responsable des modifications des processus, des procédures et des ressources. La société dispose d'un processus de gestion des modifications documenté, léger et facile à suivre. En utilisant la matrice RACI et le processus, n'importe qui peut soumettre des demandes de modification.

Étapes d'implémentation

1. Identifiez les processus, les procédures et les ressources pour votre charge de travail et les responsables de chacun d'entre eux. Documentez-les dans votre système de gestion des connaissances.
   a. Si vous n'avez pas encore mis en œuvre OPS02-BP01 Les ressources ont des propriétaires identifiés, OPS02-BP02 Les processus et procédures ont des propriétaires identifiés, ou OPS02-BP03 Les activités opérationnelles ont des propriétaires identifiés responsables de leurs performances, commencez par là.
2. Travaillez avec les parties prenantes de votre organisation pour élaborer un processus de gestion des modifications. Le processus doit couvrir les ajouts, les modifications et les exceptions pour les ressources, les processus et les procédures.
   a. Vous pouvez utiliser le Gestionnaire des modifications AWS Systems Manager comme une plateforme de gestion des modifications pour les ressources de la charge de travail.
3. Documentez le processus de gestion des modifications dans votre système de gestion des connaissances.

Niveau d'effort du plan d'implémentation : moyen. L'élaboration d'un processus de gestion des modifications nécessite un alignement avec les multiples parties prenantes de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS02-BP01 Les ressources ont des propriétaires identifiés : il faut identifier les propriétaires des ressources avant de mettre en place un processus de gestion des modifications.
- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés : il faut identifier les propriétaires des processus avant de mettre en place un processus de gestion des modifications.
- OPS02-BP03 Les activités opérationnelles ont des propriétaires identifiés responsables de leurs performances : il faut identifier les propriétaires des activités opérationnelles avant de mettre en place un processus de gestion des modifications.

Documents connexes :

- AWS Prescriptive Guidance - Foundation playbook for AWS large migrations: Creating RACI matrices (Recommandations - Foundation playbook for AWS large migrations : création de matrices RACI)
- Livre blanc sur la gestion des modifications dans le cloud

Services associés :

- Gestionnaire des modifications AWS Systems Manager

OPS02-BP07 Les responsabilités entre les équipes sont prédéfinies ou négociées

Utilisez des accords définis ou négociés entre les équipes, accords qui décrivent la manière dont elles travaillent ensemble et se soutiennent mutuellement (par exemple, les temps de réponse, les objectifs de niveau de service ou les contrats de niveau de service). Les canaux de communication inter-équipes sont documentés. La compréhension de l'impact du travail des équipes sur les résultats opérationnels et les résultats des autres équipes et organisations indique la priorité de leurs tâches et les aide à répondre de manière appropriée.

Lorsque la responsabilité et la propriété ne sont pas définies ou sont inconnues, vous risquez de ne pas traiter les activités nécessaires en temps opportun et de déployer des efforts redondants et potentiellement contradictoires pour répondre à ces besoins.

Résultat souhaité :

- Des accords de travail ou de soutien inter-équipes sont convenus et documentés.
- Les équipes qui se soutiennent ou travaillent les unes avec les autres ont défini des canaux de communication et des attentes en matière de réponse.

Anti-modèles courants :

- Un problème survient en production et deux équipes distinctes commencent à le résoudre indépendamment l'une de l'autre. Leurs efforts cloisonnés prolongent la panne.
- L'équipe chargée des opérations a besoin de l'aide de l'équipe de développement mais aucun délai de réponse n'a été convenu. La demande est bloquée dans le backlog.

Avantages liés au respect de cette bonne pratique :

- Les équipes savent comment interagir et se soutenir mutuellement.
- Les attentes en matière de réactivité sont connues.
- Les canaux de communication sont clairement définis.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

La mise en œuvre de cette bonne pratique signifie qu'il n'y a aucune ambiguïté sur la façon dont les équipes travaillent les unes avec les autres. Les accords formels codifient la manière dont les équipes travaillent ensemble ou se soutiennent mutuellement. Les canaux de communication inter-équipes sont documentés.

Exemple de client

L'équipe SRE d'AnyCompany Retail a conclu un contrat de niveau de service avec son équipe de développement. Chaque fois que l'équipe de développement émet une demande dans son système de tickets, elle peut s'attendre à recevoir une réponse dans les quinze minutes. En cas de panne du site, l'équipe SRE mène l'enquête avec le soutien de l'équipe de développement.

Étapes d'implémentation

1. En collaboration avec les parties prenantes de votre organisation, élaborez des accords entre les équipes sur la base de processus et de procédures.
   a. Si un processus ou une procédure est partagé entre deux équipes, élaborez un runbook sur la manière dont les équipes travailleront ensemble.
   b. S'il existe des dépendances entre les équipes, convenez d'un accord de niveau de service pour la réponse aux demandes.
2. Documentez les responsabilités dans votre système de gestion des connaissances.

Niveau d'effort du plan d'implémentation : moyen. Si rien n'est convenu entre les équipes, il peut être difficile de parvenir à un accord avec les parties prenantes de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés : la propriété du processus doit être identifiée avant de fixer des accords entre les équipes.
- OPS02-BP03 Les activités opérationnelles ont des propriétaires identifiés responsables de leurs performances : la propriété des activités d'opérations doit être identifiée avant d'établir des accords entre les équipes.

Documents connexes :

- AWS Executive Insights - Empowering Innovation with the Two-Pizza Team (AWS Executive Insights : stimuler l'innovation et la rapidité avec les équipes à deux pizzas d'Amazon)
- Introduction to DevOps on AWS - Two-Pizza Teams (Introduction à DevOps sur AWS : équipes à deux pizzas)

OPS 3. Comment votre culture organisationnelle soutient-elle vos résultats opérationnels ?

Offrez un support aux membres de votre équipe afin qu'ils puissent agir plus efficacement et soutenir les résultats opérationnels.

Bonnes pratiques

- OPS03-BP01 Parrainage de la direction

- OPS03-BP02 Les membres de l'équipe sont habilités à agir lorsque les résultats sont remis en cause
- OPS03-BP03 La remontée hiérarchique est encouragée
- OPS03-BP04 Les communications sont opportunes, claires et exploitables
- OPS03-BP05 L'expérimentation est encouragée
- OPS03-BP06 Les membres de l'équipe sont invités à maintenir et à développer leurs compétences
- OPS03-BP07 Fournir aux équipes les ressources appropriées
- OPS03-BP08 La diversité des opinions est encouragée et recherchée au sein des équipes et entre elles

OPS03-BP01 Parrainage de la direction

Les principaux dirigeants définissent clairement les attentes de l'organisation et évaluent le succès. Les principaux dirigeants sont le parrain, l’avocat et le moteur de l'adoption des bonnes pratiques et de l'évolution de l'organisation

Avantages liés au respect de cette bonne pratique : Une direction engagée, des attentes clairement communiquées et des objectifs partagés permettent aux membres de l'équipe de savoir ce qu'on attend d'eux. L'évaluation du succès permet d'identifier les obstacles au succès afin qu'ils puissent être éliminés par l'intervention du parrain, de l’avocat ou de leurs délégués.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Parrainage de la direction : les équipes de direction définissent clairement les attentes de l'organisation et évaluent la réussite. Les principaux dirigeants sont le parrain, l’avocat et le moteur de l'adoption des bonnes pratiques et de l'évolution de l'organisation
  - Définir les attentes : spécifiez et publiez des objectifs pour vos organisations, y compris la façon dont elles sont évaluées.
  - Suivre la réalisation des objectifs : évaluez régulièrement la réalisation progressive des objectifs et partagez les résultats afin que les mesures appropriées puissent être prises si les résultats sont remis en cause.
  - Fournir les ressources nécessaires pour atteindre vos objectifs : vérifiez régulièrement si les ressources sont toujours appropriées ou si des ressources supplémentaires sont nécessaires en fonction des nouvelles informations, des modifications des objectifs, des responsabilités ou de l'environnement de l'entreprise.

- Défendre vos équipes : restez impliqué avec vos équipes afin que vous compreniez comment elles évoluent et s'il existe des facteurs externes qui les affectent. Lorsque vos équipes sont affectées par des facteurs externes, réévaluez les objectifs et ajustez les cibles le cas échéant. Identifiez les obstacles qui entravent la progression de vos équipes. Agissez au nom de vos équipes pour surmonter les obstacles et éliminer les charges inutiles.
- Être un moteur de l'adoption des bonnes pratiques : acceptez les bonnes pratiques qui apporte des avantages quantifiables et montrez de la reconnaissance pour les créateurs et les adoptants. Encouragez une adoption plus large pour amplifier les avantages obtenus.
- Être un moteur d'évolution pour vos équipes : instaurez une culture d'amélioration continue. Encouragez la croissance et le développement personnels et organisationnels. Fixez des objectifs à long terme qui nécessiteront une réalisation progressive dans le temps. Adaptez cette vision à vos besoins et à vos objectifs et votre environnement opérationnels à mesure qu'ils évoluent.

OPS03-BP02 Les membres de l'équipe sont habilités à agir lorsque les résultats sont remis en cause

Le responsable de la charge de travail a défini des orientations et un champ d'action permettant aux membres de l'équipe de réagir lorsque les résultats sont menacés. Des mécanismes de remontée sont utilisés pour obtenir des orientations lorsque les événements outrepassent le champ d'action défini.

Avantages liés au respect de cette bonne pratique : En testant et validant rapidement les modifications, vous êtes en mesure de résoudre les problèmes avec des coûts réduits, et de limiter l'impact sur vos clients. En testant avant le déploiement, vous minimisez l'introduction d'erreurs.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Les membres de l'équipe sont habilités à agir lorsque les résultats sont remis en cause : fournissez aux membres de votre équipe les autorisations, les outils et la possibilité de mettre en pratique les compétences nécessaires pour réagir efficacement.
  - Donner aux membres de votre équipe la possibilité de mettre en pratique les compétences nécessaires pour réagir : fournissez d'autres environnements sûrs où les processus et les procédures peuvent être testés et entraînés en toute sécurité. Effectuez des simulations pour permettre aux membres de l'équipe d'acquérir de l'expérience dans la gestion d'incidents concrets dans des environnements simulés et sûrs.

- Définir et reconnaître l'autorité des membres de l'équipe pour agir : définissez spécifiquement le pouvoir des membres de l'équipe d'agir en leur attribuant des autorisations et un accès aux charges de travail et aux composants qu'ils prennent en charge. Acceptez qu'ils sont habilités à agir lorsque les résultats sont menacés.

OPS03-BP03 La remontée hiérarchique est encouragée

Les membres de l'équipe disposent de mécanismes et sont encouragés à faire part de leurs préoccupations aux décideurs et aux parties prenantes s'ils estiment que les résultats sont menacés. Les remontées doivent être effectuées tôt et souvent afin que les risques puissent être identifiés et les incidents évités.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Encourager une remontée hiérarchique précoce et fréquente : reconnaissez sur le plan organisationnel qu'une remontée hiérarchique précoce et fréquente est conseillée. Reconnaissez et acceptez sur le plan organisationnel que les remontées peuvent s'avérer non fondées et qu'il est préférable d'avoir la possibilité d'éviter un incident que de manquer cette opportunité en ne la faisant pas remonter.
  - Disposer d'un mécanisme de remontée : élaborez des procédures documentées définissant quand et comment la remontée doit avoir lieu. Documentez les séries de personnes ayant un pouvoir croissant pour prendre ou approuver des mesures, et leurs informations de contact. La remontée doit se poursuivre jusqu'à ce que le membre de l'équipe soit convaincu qu'il a transféré le risque à une personne capable d'y faire face, ou qu'il a contacté la personne qui assume le risque et la responsabilité de l'exploitation de la charge de travail. C'est cette personne qui, en fin de compte, détient toutes les décisions concernant sa charge de travail. Les remontées doivent inclure la nature du risque, la criticité de la charge de travail, les personnes impactées, l'impact et l'urgence, c'est-à-dire le moment où l'impact est attendu.
  - Protéger les employés qui font remonter les informations : créez une politique qui protège les membres de l'équipe contre les représailles s'ils font remonter des informations autour d'un décideur ou d'une partie prenante non réceptifs. Mettez en place des mécanismes permettant d'identifier si cela se produit et de répondre de manière appropriée.

OPS03-BP04 Les communications sont opportunes, claires et exploitables

Des mécanismes existent et sont utilisés pour informer en temps opportun les membres de l'équipe des risques connus et des événements planifiés. Le contexte, les informations et le temps nécessaires (dans la mesure du possible) sont communiqués pour déterminer si une action est nécessaire, quelle action est requise et prendre des mesures en temps opportun. Par exemple, la notification des failles logicielles afin d'accélérer l'application des correctifs, ou la notification des promotions de vente prévues afin qu'un gel des modifications puisse être mis en œuvre pour éviter le risque d'interruption de service. Les événements planifiés peuvent être reportés dans un calendrier de modifications ou un calendrier de maintenance afin que les membres de l'équipe puissent identifier les activités en attente.

Résultat souhaité :

- Les communications fournissent le contexte, les détails et les délais prévus.
- Les membres de l'équipe comprennent clairement quand et comment agir en réponse aux communications.
- Tirez parti des calendriers de changement pour notifier les modifications attendues.

Anti-modèles courants :

- Une alerte survient plusieurs fois par semaine et c'est un faux positif. Vous mettez la notification en sourdine à chaque fois qu'elle se produit.
- On vous demande d'apporter une modification à vos groupes de sécurité mais on ne vous indique pas quand cela doit se produire.
- Vous recevez constamment des notifications dans la conversation instantanée lorsque les systèmes montent en capacité, mais aucune action n'est nécessaire. Vous évitez le canal de conversation instantanée et manquez une notification importante.
- Une modification est apportée à la production sans en informer l'équipe chargée des opérations. Cette modification déclenche une alerte et l'équipe d'astreinte est contactée.

Avantages liés au respect de cette bonne pratique :

- Votre organisation évite la lassitude liée aux alertes.
- Les membres de l'équipe peuvent agir en tenant compte du contexte et des attentes nécessaires.
- Les modifications peuvent intervenir pendant les fenêtres de changement, ce qui réduit les risques.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour mettre en œuvre cette bonne pratique, vous devez travailler avec les parties prenantes de votre organisation pour convenir de normes de communication. Diffusez ces normes dans votre organisation. Identifiez et supprimez les alertes qui sont faussement positives ou toujours actives. Utilisez des calendriers de changement pour que les membres de l'équipe sachent quand les actions peuvent être entreprises et quelles activités sont en attente. Vérifiez que les communications mènent à des actions claires avec le contexte nécessaire.

Exemple de client

AnyCompany Retail utilise la conversation instantanée comme principal moyen de communication. Les alertes et autres informations alimentent des canaux spécifiques. Lorsque quelqu'un doit agir, le résultat attendu est clairement énoncé et, dans de nombreux cas, on lui donne un manuel ou un guide à utiliser. Les équipes utilisent un calendrier des changements pour planifier les principales modifications apportées aux systèmes de production.

Étapes d'implémentation

1. Analysez vos alertes pour détecter les faux positifs ou les alertes qui sont constamment déclenchées. Supprimez-les ou modifiez-les de manière à ce qu'elles se déclenchent lorsqu'une intervention humaine est nécessaire. Si une alerte se déclenche, fournissez un runbook ou un guide à l'opérateur.
   a. Vous pouvez utiliser les documents AWS Systems Manager pour créer des guides et des runbooks pour les alertes.
2. Deux mécanismes sont en place pour notifier les risques ou les événements prévus d'une manière claire et exploitable, avec un préavis suffisant pour permettre des réponses appropriées. Utilisez des listes d'e-mails ou des canaux de conversation instantanée pour envoyer des notifications avant les événements prévus.
   a. AWS Chatbot permet d'envoyer des alertes et de répondre à des événements au sein de la plateforme de messagerie de votre organisation.
3. Fournissez une source d'informations accessible où les événements planifiés peuvent être découverts. Envoyez des notifications d'événements planifiés à partir du même système.
   a. Le calendrier des modifications AWS Systems Manager permet de créer des fenêtres de changement lorsque des modifications peuvent intervenir. Cela permet aux membres de l'équipe de savoir quand ils peuvent apporter des modifications en toute sécurité.

4. Surveillez les notifications de failles et les informations sur les correctifs pour comprendre les failles dangereuses et les risques potentiels associés aux éléments de votre charge de travail. Envoyez une notification aux membres de l'équipe afin qu'ils puissent agir.
   a. Vous pouvez vous abonner aux bulletins de sécurité AWS pour recevoir des notifications de failles sur AWS.

Ressources

Bonnes pratiques associées :

- OPS07-BP03 Utiliser des runbooks pour effectuer des procédures : rendez les communications exploitables en fournissant un runbook lorsque le résultat est connu.
- OPS07-BP04 Utiliser des playbooks pour analyser les problèmes : lorsque le résultat est inconnu, les guides peuvent rendre les communications exploitables.

Documents connexes :

- Bulletins de sécurité AWS
- Open CVE (CVE ouvertes)

Exemples connexes :

- Ateliers Well-Architected : inventaire et gestion des correctifs (niveau 100)

Services associés :

- AWS Chatbot
- Modifier le calendrier AWS Systems Manager
- Documents AWS Systems Manager

OPS03-BP05 L'expérimentation est encouragée

L'expérimentation est un catalyseur qui permet de transformer de nouvelles idées en produits et en fonctionnalités. Elle accélère la formation et permet aux membres de l'équipe de s'intéresser et d'être engagés. Les membres de l'équipe sont encouragés à expérimenter souvent pour stimuler l'innovation. Même lorsqu'un résultat indésirable se produit, il est bon de savoir ce qu'il ne faut

pas faire. Les membres de l'équipe ne sont pas sanctionnés pour les expérimentations réussies produisant des résultats indésirables.

Résultat souhaité :

- Votre organisation encourage l'expérimentation pour favoriser l'innovation.
- Les expériences sont utilisées comme une occasion d'apprendre.

Anti-modèles courants :

- Vous souhaitez effectuer un test A/B mais il n'existe aucun mécanisme pour réaliser l'expérience. Vous déployez une modification de l'interface utilisateur sans pouvoir la tester. Il en résulte une expérience négative pour le client.
- Votre entreprise ne dispose que d'un environnement d'étape et de production. Il n'existe pas d'environnement de test (sandbox) pour expérimenter de nouvelles fonctionnalités ou de nouveaux produits. Vous devez donc expérimenter dans l'environnement de production.

Avantages liés au respect de cette bonne pratique :

- L'expérimentation est le moteur de l'innovation.
- Vous pouvez réagir plus rapidement aux commentaires des utilisateurs grâce à l'expérimentation.
- Votre organisation développe une culture de l'apprentissage.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les expériences doivent être menées en toute sécurité. Exploitez plusieurs environnements pour expérimenter sans mettre en péril les ressources de production. Utilisez les tests A/B et les indicateurs de fonctionnalités pour tester les expériences. Donnez aux membres de l'équipe la possibilité de mener des expériences dans un environnement de test (sandbox).

Exemple de client

AnyCompany Retail encourage l'expérimentation. Les membres de l'équipe peuvent utiliser 20 % de leur semaine de travail pour expérimenter ou apprendre de nouvelles technologies. Ils disposent d'un environnement de test (sandbox) où ils peuvent innover. Les tests A/B sont utilisés pour les nouvelles fonctionnalités afin de les valider en fonction des commentaires réels des utilisateurs.

Étapes d'implémentation

1. Travaillez avec les dirigeants de votre organisation pour soutenir l'expérimentation. Les membres de l'équipe doivent être encouragés à réaliser des expériences en toute sécurité.
2. Offrez aux membres de votre équipe un environnement où ils peuvent expérimenter en toute sécurité. Ils doivent avoir accès à un environnement similaire à celui de la production.
   a. Vous pouvez utiliser un Compte AWS distinct pour créer un environnement de test (sandbox) pour l'expérimentation. AWS Control Tower peut être utilisé pour provisionner ces comptes.
3. Utilisez des indicateurs de fonctions et des tests A/B pour expérimenter en toute sécurité et recueillir les commentaires des utilisateurs.
   a. Les indicateurs de fonctions AWS AppConfig offrent la possibilité de créer des indicateurs de fonctionnalités.
   b. Amazon CloudWatch Evidently permet d'effectuer des tests A/B sur un déploiement limité.
   c. Vous pouvez utiliser les versions AWS Lambda pour déployer une nouvelle version d'une fonction pour un test bêta.

Niveau d'effort du plan d'implémentation : élevé. Fournir aux membres de l'équipe un environnement dans lequel expérimenter et un moyen sûr de mener des expériences peut nécessiter un investissement important. Il se peut également que vous deviez modifier le code de l'application pour utiliser des indicateurs de fonctions ou prendre en charge les tests A/B.

Ressources

Bonnes pratiques associées :

- OPS11-BP02 Effectuer une analyse post-incident : les leçons tirées des incidents sont un moteur important pour l'innovation, au même titre que l'expérimentation.
- OPS11-BP03 Mettre en œuvre des boucles de rétroaction : les boucles de rétroaction sont une composante importante de l'expérimentation.

Documents connexes :

- An Inside Look at the Amazon Culture: Experimentation, Failure, and Customer Obsession (Un examen interne de la culture Amazon : expérimentation, échec et obsession du client)
- Best practices for creating and managing sandbox accounts in AWS (Bonnes pratiques pour la création et la gestion des comptes sandbox dans AWS)

- Create a Culture of Experimentation Enabled by the Cloud (Créer une culture de l'expérimentation grâce au cloud)
- Enabling experimentation and innovation in the cloud at SulAmérica Seguros (Favoriser l'expérimentation et l'innovation dans le cloud chez SulAmérica Seguros)
- Experiment More, Fail Less (Expérimenter davantage, moins échouer)
- Organisation de votre environnement AWS à l'aide de plusieurs comptes - Sandbox OU
- Using AWS AppConfig Feature Flags (Utilisation des indicateurs de fonctions AWS AppConfig)

Vidéos connexes :

- AWS On Air ft. Amazon CloudWatch Evidently | AWS Events
- AWS On Air San Fran Summit 2022 ft. AWS AppConfig Feature Flags integration with Jira
- AWS re:Invent 2022 - A deployment is not a release: Control your launches w/feature flags (BOA305-R) [AWS re:Invent 2022 - Un déploiement n'est pas un lancement : contrôlez vos lancements avec des indicateurs de fonctions (BOA305-R)]
- Programmatically Create an Compte AWS with AWS Control Tower (Créer programmatiquement un compte AWS avec WS Control Tower)
- Set Up a Multi-Account AWS Environment that Uses Best Practices for AWS Organizations

Exemples connexes :

- AWS Innovation Sandbox
- End-to-end Personalization 101 for E-Commerce (Bases de la personnalisation de bout en bout pour le e-commerce)

Services associés :

- Amazon CloudWatch Evidently
- AWS AppConfig
- AWS Control Tower

OPS03-BP06 Les membres de l'équipe sont invités à maintenir et à développer leurs compétences

Les équipes doivent accroître leurs compétences pour adopter les nouvelles technologies, et pour faire face à l'évolution de la demande et des responsabilités afin de supporter votre charge de travail. Le développement des compétences dans les nouvelles technologies est souvent une source de satisfaction pour les membres de l'équipe et favorise l'innovation. Soutenez les membres de votre équipe dans la recherche et le maintien de certifications sectorielles qui valident et reconnaissent leurs compétences croissantes. Mettez en place la formation croisée pour promouvoir le transfert de connaissances et réduire le risque d'impact significatif lorsque vous perdez des membres d'équipe qualifiés et expérimentés ayant un savoir institutionnel. Mettez en place des créneaux dédiés à la formation.

AWS fournit des ressources, y compris le Centre de ressources de démarrage AWS, les blogs AWS, les conférences techniques en ligne AWS, les événements et webinaires AWSet les ateliers AWS Well-ArchitectedAWS, qui fournissent des conseils, des exemples et des démonstrations détaillées pour former vos équipes.

AWS partage également les bonnes pratiques et les modèles tirés de l'expérience AWS dans la Bibliothèque Amazon Builders' Library et d'une grande variété d'autres supports pédagogiques utiles via le Blog AWS et Podcast AWS officiel.

Tirez parti des ressources pédagogiques fournies par AWS, telles que les ateliers Well-Architected, AWS Support (Centre de connaissances AWS, les forums de discussion AWSet Centre AWS Support) et la Documentation AWS pour former vos équipes. Contactez AWS Support via le centre AWS Support pour obtenir des réponses à vos questions AWS.

AWS Training and Certification offre une formation gratuite par le biais de cours en ligne d'auto-formation sur les principes fondamentaux d'AWS. Vous pouvez également vous inscrire à une formation dirigée par un formateur afin de soutenir le développement des compétences AWS de vos équipes.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Les membres de l'équipe sont invités à maintenir et à développer leurs compétences : pour adopter les nouvelles technologies, soutenir l'innovation et accompagner l'évolution de la demande et des responsabilités à l'appui de vos charges de travail, la formation continue est nécessaire.
  - Fournir des ressources de formation : fournissez un temps structuré dédié, l'accès à des supports de formation, des ressources d'atelier et la possibilité de se joindre à des conférences

et à des organisations professionnelles qui offrent des possibilités de formation auprès de formateurs et de pairs. Offrez aux membres juniors de l'équipe l'accès aux membres chevronnés en tant que mentors ou autorisez-les à suivre leur travail et à être exposés à leurs méthodes et compétences. Encouragez l'apprentissage du contenu qui n’est pas directement lié au travail afin d'avoir une perspective plus large.

- Formation des équipes et engagement inter-équipe : planifiez les besoins de formation continue des membres de votre équipe. Offrir aux membres de l'équipe la possibilité de rejoindre d'autres équipes (temporairement ou définitivement) pour partager les compétences et les bonnes pratiques au profit de l'ensemble de votre organisation
- Soutenir l'obtention et la conservation des certifications du secteur : aidez les membres de votre équipe à acquérir et à conserver les certifications sectorielles qui valident ce qu'ils ont appris, et reconnaissez leurs réalisations.

Ressources

Documents connexes :

- Centre de ressources de démarrage AWS
- les blogs AWS
- Conformité du AWS Cloud
- les forums de discussion AWS
- Documentation AWS
- les conférences techniques en ligne AWS
- les événements et webinaires AWS
- Centre de connaissances AWS
- AWS Support
- AWS Training and Certification
- ateliers AWS Well-Architected,
- Bibliothèque Amazon Builders' Library
- Podcast AWS officiel.

OPS03-BP07 Fournir aux équipes les ressources appropriées

Maintenez les aptitudes des membres de l'équipe et fournissez les outils et les ressources nécessaires pour répondre aux besoins de votre charge de travail. Surmener les membres d'équipe augmente le risque d'incidents résultant d'une erreur humaine. Les investissements dans les outils et les ressources (par exemple, l'automatisation des activités fréquemment exécutées) peuvent accroître l'efficacité de votre équipe, lui permettant de soutenir des activités supplémentaires.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Fournir aux équipes les ressources appropriées : veillez à bien comprendre le succès de vos équipes et les facteurs qui contribuent à ce succès ou à leur échec. Agissez pour soutenir les équipes avec les ressources appropriées.
  - Comprendre les performances de l'équipe : mesurez la réalisation des résultats opérationnels et le développement des atouts par vos équipes. Suivez l'évolution de la production et du taux d'erreur dans le temps. Collaborez avec les équipes pour comprendre les défis liés au travail qui les impactent (par exemple, augmentation des responsabilités, modifications technologiques, perte de personnel ou augmentation du nombre de clients pris en charge).
  - Comprendre les impacts sur les performances des équipes : restez impliqué avec vos équipes afin que vous compreniez comment elles évoluent et s'il existe des facteurs externes qui les affectent. Lorsque vos équipes sont affectées par des facteurs externes, réévaluez les objectifs et ajustez les cibles le cas échéant. Identifiez les obstacles qui entravent la progression de vos équipes. Agissez au nom de vos équipes pour surmonter les obstacles et éliminer les charges inutiles.
  - Fournir les ressources nécessaires pour assurer le succès des équipes : vérifiez régulièrement si les ressources sont toujours appropriées et si des ressources supplémentaires sont nécessaires, et procédez aux ajustements appropriés pour les équipes de support.

OPS03-BP08 La diversité des opinions est encouragée et recherchée au sein des équipes et entre elles

Exploitez la diversité inter-organisationnelle pour rechercher des perspectives multiples et uniques. Utilisez cette perspective pour accroître l'innovation, remettre en question vos hypothèses et réduire le risque de biais de confirmation. Développez l'inclusion, la diversité et l'accessibilité au sein de vos équipes afin d'obtenir des perspectives bénéfiques.

La culture organisationnelle a un impact direct sur la satisfaction professionnelle et la fidélisation des membres de l'équipe. Favorisez l'engagement et l'exploitation des capacités des membres de votre équipe pour assurer la réussite de votre entreprise.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Rechercher des opinions et des perspectives variées : encouragez les contributions de chacun. Donnez la parole aux groupes sous-représentés. Effectuez une rotation des rôles et des responsabilités lors des réunions.
  - Élargir les rôles et les responsabilités : offrez aux membres de l'équipe l'opportunité d'assumer des rôles qu'ils n'auraient pas autrement. Ils acquièrent de l'expérience et du point de vue du rôle et des interactions avec les nouveaux membres de l'équipe avec lesquels ils n'interagissent pas autrement. Ils apporteront leur expérience et leur point de vue au nouveau rôle et aux membres de l'équipe avec lesquels ils interagissent. À mesure que les perspectives s'élargissent, de nouvelles opportunités commerciales peuvent se présenter, ou de nouvelles possibilités d'amélioration peuvent être identifiées. Demandez aux membres d'une équipe d'effectuer à tour de rôle des tâches communes que les autres exécutent habituellement afin de comprendre les exigences et l'impact de leur exécution.
  - Fournir un environnement sûr et accueillant : mettez en place une politique et des contrôles qui protègent la sécurité mentale et physique des membres de l'équipe au sein de votre organisation. Les membres de l'équipe doivent être en mesure d'interagir sans craindre de représailles. Lorsque les membres de l'équipe se sentent en sécurité et sont les bienvenus, ils sont plus susceptibles d'être impliqués et productifs. Plus votre organisation est diversifiée, mieux vous pouvez comprendre les personnes que vous soutenez, y compris vos clients. Lorsque les membres de votre équipe sont à l'aise, se sentent libres de parler et sont sûrs d'être entendus, ils sont plus susceptibles de partager des informations précieuses (par exemple, les possibilités de marketing, les besoins d'accessibilité, les segments de marché délaissés, les risques non reconnus dans votre environnement).
  - Permettre aux membres de l'équipe de participer pleinement : fournissez les ressources nécessaires pour que vos employés puissent participer pleinement à toutes les activités liées à leur travail. Les membres de l'équipe qui font face à des défis quotidiens ont développé des compétences pour s’y atteler. Ces compétences développées de manière unique peuvent apporter des avantages considérables à votre organisation. L'accompagnement des membres de l'équipe avec les ajustements nécessaires augmente les avantages que vous pouvez tirer de leurs contributions.

Préparation

Questions

- OPS 4. Comment mettre en œuvre l'observabilité dans votre charge de travail ?
- OPS 5. Comment réduire les défauts, faciliter les corrections et améliorer le flux dans la production ?
- OPS 6. Comment réduire les risques liés au déploiement ?
- OPS 7. Comment savoir si vous êtes prêt à assurer une charge de travail ?

OPS 4. Comment mettre en œuvre l'observabilité dans votre charge de travail ?

Intégrez l'observabilité à votre charge de travail afin de comprendre son état et de prendre des décisions basées sur les données en fonction des exigences de l'entreprise.

Bonnes pratiques

- OPS04-BP01 Identification des indicateurs clés de performance
- OPS04-BP02 Mise en œuvre de la télémétrie de l'application
- OPS04-BP03 Mise en œuvre de la télémétrie de l'expérience utilisateur
- OPS04-BP04 Mise en œuvre de la télémétrie des dépendances
- OPS04-BP05 Mise en œuvre du suivi distribué

OPS04-BP01 Identification des indicateurs clés de performance

La mise en œuvre de l'observabilité dans votre charge de travail commence par la compréhension de son état et par la prise de décisions basées sur les données en fonction des exigences de l'entreprise. L'un des moyens les plus efficaces de garantir l'alignement entre les activités de surveillance et les objectifs commerciaux consiste à définir des indicateurs clés de performance (KPI) et à les suivre de près.

Résultat souhaité : Pratiques d'observabilité efficaces qui sont étroitement alignées sur les objectifs commerciaux, garantissant que les efforts de surveillance sont toujours au service de résultats commerciaux tangibles.

Anti-modèles courants :

- KPI non définis : travailler sans indicateurs clés de performance bien définis peut entraîner une surveillance excessive ou insuffisante, ce qui peut faire passer à côté de signaux essentiels.
- KPI statiques : ne pas revisiter ou affiner les indicateurs clés de performance au fur et à mesure de l'évolution de la charge de travail ou des objectifs commerciaux.
- Désalignement : se concentrer sur des métriques techniques qui ne sont pas directement corrélées aux résultats commerciaux ou qui sont plus difficiles à corréler aux problèmes réels.

Avantages liés au respect de cette bonne pratique :

- Facilité d'identification des problèmes : les KPI commerciaux permettent souvent de détecter les problèmes plus clairement que les métriques techniques. La baisse d'un KPI commercial permet d'identifier un problème plus efficacement que de passer au crible de nombreuses métriques techniques.
- Cohérence des activités : garantit que les activités de surveillance soutiennent directement les objectifs commerciaux.
- Efficacité : la priorité est donnée à la surveillance des ressources et l'attention est concentrée sur les métriques déterminantes.
- Proactivité : identifiez et traitez les problèmes avant qu'ils n'aient des implications commerciales plus larges.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Pour définir efficacement les indicateurs clés de performance de la charge de travail :

1. Commencez par les résultats commerciaux : Avant de vous plonger dans les métriques, déterminez les résultats commerciaux souhaités. S'agit-il d'une augmentation des ventes, d'un engagement plus élevé des utilisateurs ou d'une réduction des temps de réponse ?
2. Corrélez les métriques techniques aux objectifs commerciaux : Les métriques techniques n'ont pas toutes un impact direct sur les résultats commerciaux. Identifiez celles qui ont un impact direct, mais il est souvent plus simple d'identifier un problème à l'aide d'un KPI commercial.
3. Utilisez Amazon CloudWatch : Utilisez CloudWatch pour définir et surveiller les métriques qui représentent vos KPI.
4. Passez régulièrement en revue les KPI et actualisez-les : À mesure que votre charge de travail et votre activité évoluent, veillez à ce que vos KPI restent pertinents.

5. Impliquez les parties prenantes : Impliquez les équipes techniques et commerciales dans la définition et la révision des KPI.

Niveau d'effort du plan d'implémentation : Moyen

Ressources

Bonnes pratiques associées :

- the section called “OPS04-BP02 Mise en œuvre de la télémétrie de l'application”
- the section called “OPS04-BP03 Mise en œuvre de la télémétrie de l'expérience utilisateur”
- the section called “OPS04-BP04 Mise en œuvre de la télémétrie des dépendances”
- the section called “OPS04-BP05 Mise en œuvre du suivi distribué”

Documents connexes :

- Bonnes pratiques AWS en matière d'observabilité
- Guide de l'utilisateur CloudWatch
- Cours de renforcement des compétences en observabilité AWS

Vidéos connexes :

- Developing an observability strategy

Exemples connexes :

- Un atelier sur l'observabilité

OPS04-BP02 Mise en œuvre de la télémétrie de l'application

La télémétrie de l'application est la pierre angulaire de l'observabilité de votre charge de travail. Il est essentiel de diffuser des données télémétriques fournissant des informations exploitables sur l'état de votre application et sur son taux de réussite par rapport aux résultats techniques et commerciaux. Qu'il s'agisse de résoudre des problèmes, de mesurer l'impact d'une nouvelle fonctionnalité ou de garantir l'alignement sur les indicateurs clés de performance (KPI) de l'entreprise, la télémétrie de l'application vous permet de créer, d'exploiter et de faire évoluer votre charge de travail.

Les métriques, les journaux et les données de suivi constituent les trois principaux piliers de l'observabilité. Ils servent d'outils de diagnostic qui décrivent l'état de votre application. Au fil du temps, ils contribuent à créer des points de référence et à identifier les anomalies. Cependant, pour garantir l'alignement entre les activités de surveillance et les objectifs commerciaux, il est essentiel de définir et de surveiller les KPI. Les KPI commerciaux facilitent souvent l'identification des problèmes par rapport aux seules métriques techniques.

D'autres types de télémétrie, tels que la surveillance des utilisateurs réels (RUM) et les transactions synthétiques, complètent ces sources de données principales. RUM fournit des informations sur les interactions des utilisateurs en temps réel, tandis que les transactions synthétiques simulent les comportements potentiels des utilisateurs, ce qui contribue à détecter les goulets d'étranglement avant que les utilisateurs réels ne soient affectés.

Résultat souhaité : Obtenez des informations exploitables sur les performances de votre charge de travail. Ces informations vous permettront de prendre des décisions proactives concernant l'optimisation des performances, d'accroître la stabilité de la charge de travail, de rationaliser les processus CI/CD et d'utiliser efficacement les ressources.

Anti-modèles courants :

- Observabilité incomplète : le fait de négliger d'intégrer l'observabilité à chaque niveau de la charge de travail entraîne des angles morts susceptibles de masquer des informations essentielles sur les performances et le comportement du système.
- Vue fragmentée des données : lorsque les données sont dispersées entre plusieurs outils et systèmes, il devient difficile de conserver une vision globale de l'état et des performances de la charge de travail.
- Problèmes signalés par les utilisateurs : indique que la détection proactive des problèmes par le biais de la télémétrie et de la surveillance des indicateurs clés de performance de l'entreprise fait défaut.

Avantages liés au respect de cette bonne pratique :

- Prise de décision éclairée : grâce aux informations issues de la télémétrie et des KPI commerciaux, vous pouvez prendre des décisions basées sur les données.
- Efficacité opérationnelle améliorée : l'utilisation des ressources axée sur les données est source de rentabilité.

- Stabilité accrue de la charge de travail : détection et résolution plus rapides des problèmes, ce qui améliore la disponibilité.
- Processus CI/CD rationalisés : les informations issues des données de télémétrie facilitent l'affinement des processus et la livraison fiable du code.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Pour implémenter la télémétrie d'une application pour votre charge de travail, utilisez des services AWS comme Amazon CloudWatch et AWS X-Ray. Amazon CloudWatch fournit une suite complète d'outils de surveillance, vous permettant d'observer vos ressources et vos applications dans AWS et dans les environnements sur site. Il collecte, suit et analyse les métriques, consolide et surveille les données des journaux, et répond à l'évolution de vos ressources, vous permettant ainsi de mieux comprendre le fonctionnement de votre charge de travail. En tandem, AWS X-Ray vous permet de suivre, d'analyser et de déboguer vos applications, ce qui vous permet aussi de mieux comprendre le comportement de votre charge de travail. Avec des fonctionnalités telles que les cartographies des services, les distributions de latence et les chronologies de suivi, X-Ray fournit des informations sur les performances de votre charge de travail et les obstacles qui l'affectent.

Étapes d'implémentation

1. Identifiez les données à collecter : déterminez les métriques, les journaux et les données de suivi essentiels qui fourniraient des informations substantielles sur l'état, les performances et le comportement de votre charge de travail.
2. Déployez l'agent CloudWatch : L'agent CloudWatch joue un rôle essentiel dans l'obtention des métriques et des journaux liés au système et aux applications à partir de votre charge de travail et de son infrastructure sous-jacente. L'agent CloudWatch peut également être utilisé pour collecter des données de suivi OpenTelemetry ou X-Ray et les envoyer à X-Ray.
3. Définissez et surveillez les KPI commerciaux : déterminez des métriques personnalisées qui correspondent à vos résultats commerciaux..
4. Instrumentez votre application avec AWS X-Ray : outre le déploiement de l'agent CloudWatch, il est essentiel d'instrumenter votre application pour émettre des données de suivi. Ce processus peut fournir des informations supplémentaires sur le comportement et les performances de votre charge de travail.
5. Standardisez la collecte de données dans l'ensemble de votre application : standardisez les pratiques de collecte de données dans l'ensemble de votre application. L'uniformité facilite la

corrélation et l'analyse des données, fournissant ainsi une vue complète du comportement de votre application.

6. Analysez les données et agissez en conséquence : une fois que la collecte et la normalisation des données sont en place, utilisez Amazon CloudWatch pour l'analyse des métriques et des journaux, et AWS X-Ray pour l'analyse des données de suivi. Une telle analyse peut fournir des informations cruciales sur l'état, les performances et le comportement de votre charge de travail, orientant ainsi votre processus décisionnel.

Niveau d'effort du plan d'implémentation : Élevé

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identification des indicateurs clés de performance
- OPS04-BP03 Mise en œuvre de la télémétrie de l'expérience utilisateur
- OPS04-BP04 Mise en œuvre de la télémétrie des dépendances
- OPS04-BP05 Mise en œuvre du suivi distribué

Documents connexes :

- Bonnes pratiques AWS en matière d'observabilité
- Guide de l'utilisateur CloudWatch
- Guide du développeur AWS X-Ray
- Instrumentation des systèmes distribués au profit de la visibilité opérationnelle
- Cours de renforcement des compétences en observabilité AWS
- Nouveautés avec Amazon CloudWatch
- Nouveautés avec AWS X-Ray

Vidéos connexes :

- AWS re:Invent 2022 - Observability best practices at Amazon
- AWS re:Invent 2022 - Developing an observability strategy

Exemples connexes :

- Un atelier sur l'observabilité
- Bibliothèque de solutions AWS : surveillance des applications avec Amazon CloudWatch

OPS04-BP03 Mise en œuvre de la télémétrie de l'expérience utilisateur

Il est essentiel d'obtenir des informations approfondies sur les expériences des clients et leurs interactions avec votre application. La surveillance des utilisateurs réels (ou RUM) et les transactions synthétiques constituent de puissants outils à cette fin. RUM fournit des données sur les interactions des utilisateurs réels, offrant une perspective non filtrée de la satisfaction des utilisateurs, tandis que les transactions synthétiques simulent les interactions des utilisateurs, ce qui contribue à détecter les problèmes potentiels avant même qu'ils n'affectent les utilisateurs réels.

Résultat souhaité : une vision globale de l'expérience client, une détection proactive des problèmes et une optimisation des interactions avec les utilisateurs pour proposer des expériences numériques fluides.

Anti-modèles courants :

- Applications sans surveillance des utilisateurs réels (RUM) :
  - Détection différée des problèmes : sans RUM, il est possible que vous ne vous rendiez compte de l'existence de goulets d'étranglement ou de problèmes de performances que lorsque les utilisateurs se plaignent. Cette approche réactive peut entraîner l'insatisfaction des clients.
  - Manque d'informations sur l'expérience utilisateur : si vous n'utilisez pas RUM, vous passez à côté de données cruciales qui montrent comment les utilisateurs réels interagissent avec votre application. Vous limitez ainsi votre capacité à optimiser l'expérience utilisateur.
- Applications sans transactions synthétiques :
  - Cas marginaux manqués : les transactions synthétiques vous aident à tester des chemins et des fonctions qui ne sont pas toujours fréquemment utilisés par les utilisateurs ordinaires, mais qui sont essentiels à certaines fonctions commerciales. Sans ces transactions synthétiques, ces chemins pourraient mal fonctionner et passer inaperçus.
  - Recherche de problèmes lorsque l'application n'est pas utilisée : des tests synthétiques réguliers permettent de simuler les situations où les utilisateurs réels n'interagissent pas activement avec votre application, garantissant ainsi le bon fonctionnement du système.

Avantages liés au respect de cette bonne pratique :

- Détection proactive des problèmes : identifiez et résolvez les problèmes potentiels avant qu'ils n'affectent les utilisateurs réels.
- Expérience utilisateur optimisée : les retours continus issus de RUM contribuent à affiner et à améliorer l'expérience utilisateur globale.
- Informations sur les performances de l'appareil et du navigateur : comprenez le fonctionnement de votre application sur différents appareils et navigateurs, afin de l'affiner davantage.
- Flux de travail validés : des transactions synthétiques régulières garantissent que les fonctionnalités de base et les chemins critiques restent opérationnels et efficaces.
- Performances améliorées des applications : exploitez les informations recueillies à partir de données sur les utilisateurs réels pour améliorer la réactivité et la fiabilité des applications.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Pour tirer parti de RUM et des transactions synthétiques pour la télémétrie de l'activité des utilisateurs, AWS propose des services comme Amazon CloudWatch RUM et Amazon CloudWatch Synthetics. Les métriques, les journaux et les données de suivi, associés aux données d'activité des utilisateurs, fournissent une vue complète de l'état de fonctionnement de l'application et de l'expérience utilisateur.

Étapes d'implémentation

1. Déployez Amazon CloudWatch RUM : intégrez votre application à CloudWatch RUM pour collecter, analyser et présenter des données sur les utilisateurs réels.
   a. Utilisez la bibliothèque JavaScript CloudWatch RUM pour intégrer RUM à votre application.
   b. Configurez des tableaux de bord pour visualiser et surveiller les données sur les utilisateurs réels.
2. Configurez CloudWatch Synthetics : créez des canarys, ou des routines scriptées, qui simulent les interactions des utilisateurs avec votre application.
   a. Définissez les flux de travail et les chemins d'application critiques.
   b. Concevez des canarys en utilisant des scripts CloudWatch Synthetics afin de simuler les interactions des utilisateurs pour ces chemins.
   c. Planifiez et surveillez les canarys pour qu'ils fonctionnent à des intervalles spécifiés, afin de garantir des contrôles de performance cohérents.

3. Analysez les données et agissez en conséquence : utilisez les données issues de RUM et des transactions synthétiques pour obtenir des informations exploitables et prendre des mesures correctives lorsque des anomalies sont détectées. Utilisez des tableaux de bord et des alarmes CloudWatch pour rester informé.

Niveau d'effort du plan d'implémentation : Moyen

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identification des indicateurs clés de performance
- OPS04-BP02 Mise en œuvre de la télémétrie de l'application
- OPS04-BP04 Mise en œuvre de la télémétrie des dépendances
- OPS04-BP05 Mise en œuvre du suivi distribué

Documents connexes :

- Guide d'Amazon CloudWatch RUM
- Guide d'Amazon CloudWatch Synthetics

Vidéos connexes :

- Optimize applications through end user insights with Amazon CloudWatch RUM
- AWS on Air ft. Real-User Monitoring for Amazon CloudWatch

Exemples connexes :

- Un atelier sur l'observabilité
- Référentiel Git pour le client web Amazon CloudWatch RUM
- Utilisation d'Amazon CloudWatch Synthetics pour mesurer le temps de chargement des pages

OPS04-BP04 Mise en œuvre de la télémétrie des dépendances

La télémétrie des dépendances est essentielle pour surveiller l'état et les performances des services et composants externes sur lesquels repose votre charge de travail. Elle fournit des

informations précieuses sur l'accessibilité, les délais d'attente et d'autres événements critiques liés aux dépendances tels que le DNS, les bases de données ou les API tierces. En instrumentant votre application de sorte à émettre des métriques, des journaux et des données de suivi concernant ces dépendances, vous identifiez plus facilement les goulets d'étranglement potentiels, les problèmes de performances ou les défaillances susceptibles d'avoir un impact sur votre charge de travail.

Résultat souhaité : Les dépendances sur lesquelles repose votre charge de travail fonctionnent comme prévu, ce qui vous permet de résoudre les problèmes de manière proactive et de garantir des performances de charge de travail optimales.

Anti-modèles courants :

- Omission des dépendances externes : se concentrer uniquement sur les métriques internes des applications tout en négligeant les métriques liées aux dépendances externes.
- Absence de surveillance proactive : attendre l'apparition de problèmes au lieu de surveiller en permanence l'état et les performances des dépendances.
- Surveillance cloisonnée : utiliser des outils de surveillance divers et variés qui peuvent donner lieu à des visions fragmentées et incohérentes de l'état des dépendances.

Avantages liés au respect de cette bonne pratique :

- Fiabilité améliorée de la charge de travail : en garantissant que les dépendances externes sont constamment disponibles et fonctionnent de manière optimale.
- Détection et résolution plus rapides des problèmes : en identifiant et en résolvant de manière proactive les problèmes liés aux dépendances avant qu'ils n'affectent la charge de travail.
- Vue globale : grâce à une visibilité complète des composants internes et externes qui influencent l'état de la charge de travail.
- Meilleure capacité de mise à l'échelle de la charge de travail : grâce à une meilleure compréhension des limites de la capacité de mise à l'échelle et des caractéristiques de performance des dépendances externes.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Mettez en œuvre la télémétrie des dépendances en commençant par identifier les services, l'infrastructure et les processus sur lesquels repose votre charge de travail. Quantifiez ce à quoi

les conditions favorables ressemblent lorsque ces dépendances fonctionnent comme prévu, puis déterminez les données nécessaires pour les mesurer. Ces informations vous permettront de créer des tableaux de bord et des alertes qui fourniront à vos équipes opérationnelles des informations sur l'état de ces dépendances. Utilisez les outils AWS pour découvrir et quantifier les impacts lorsque les dépendances ne répondent pas aux besoins. Revoyez continuellement votre stratégie en tenant compte de l'évolution des priorités, des objectifs et des connaissances acquises.

Étapes d'implémentation

Pour implémenter efficacement la télémétrie des dépendances :

1. Identifiez les dépendances externes : Collaborez avec les parties prenantes pour identifier les dépendances externes sur lesquelles repose votre charge de travail. Les dépendances externes peuvent inclure des services tels que des bases de données externes, des API tierces, des routes de connectivité réseau vers d'autres environnements et des services DNS. La première étape à suivre pour assurer l'efficacité de la télémétrie des dépendances consiste à comprendre parfaitement ce que sont ces dépendances.
2. Élaborez une stratégie de surveillance : Une fois que vous avez une idée précise de vos dépendances externes, élaborez une stratégie de surveillance qui leur est adaptée. Cela implique de comprendre le caractère critique de chaque dépendance, son comportement attendu et tous les contrats ou tous les objectifs de niveau de service associés (SLA ou SLT). Configurez des alertes proactives pour vous informer des changements d'état ou des écarts de performance.
3. Exploitez le Moniteur Internet Amazon CloudWatch : Il fournit un aperçu de l'Internet mondial, ce qui vous aide à comprendre les pannes ou les perturbations susceptibles d'avoir un impact sur vos dépendances externes.
4. Restez informé avec AWS Health Dashboard : Il fournit des alertes et des conseils de résolution en cas d'événements AWS susceptibles d'avoir un impact sur vos services.
5. Instrumentez votre application avec AWS X-Ray : AWS X-Ray fournit des informations sur les performances des applications et de leurs dépendances sous-jacentes. En suivant les requêtes du début à la fin, vous pouvez identifier les goulets d'étranglement ou les défaillances des services ou composants externes sur lesquels repose votre application.
6. Utilisez Amazon DevOps Guru : Ce service basé sur le machine learning identifie les problèmes opérationnels, prédit quand des problèmes critiques peuvent survenir et recommande des mesures spécifiques à prendre. Il s'agit d'un outil inestimable qui permet de mieux comprendre les dépendances et de déterminer qu'elles ne sont pas à l'origine de problèmes opérationnels.

7. À surveiller régulièrement : Surveillez en permanence les métriques et les journaux liés aux dépendances externes. Configurez des alertes en cas de comportement inattendu ou de dégradation des performances.
8. À valider après toute modification : Chaque fois qu'une dépendance externe est mise à jour ou modifiée, validez ses performances et vérifiez qu'elle correspond aux exigences de votre application.

Niveau d'effort du plan d'implémentation : Moyen

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identification des indicateurs clés de performance
- OPS04-BP02 Mise en œuvre de la télémétrie de l'application
- OPS04-BP03 Mise en œuvre de la télémétrie de l'expérience utilisateur
- OPS04-BP05 Mise en œuvre du suivi distribué

Documents connexes :

- Qu'est-ce qu'AWS Health ?
- Utilisation du Moniteur Internet Amazon CloudWatch
- Guide du développeur AWS X-Ray
- Guide de l'utilisateur Amazon DevOps Guru

Vidéos connexes :

- Visibility into how internet issues impact app performance
- Introduction to Amazon DevOps Guru

Exemples connexes :

- Obtenir des informations opérationnelles grâce à l'AIOps en utilisant Amazon DevOps Guru
- AWS Health Aware

OPS04-BP05 Mise en œuvre du suivi distribué

Le suivi distribué permet de surveiller et de visualiser les requêtes lorsqu'elles traversent les différents composants d'un système distribué. En capturant les données de suivi provenant de plusieurs sources et en les analysant dans une vue unifiée, les équipes peuvent mieux comprendre le flux des requêtes, les endroits où les goulots d'étranglement ont lieu et les domaines dans lesquels les efforts d'optimisation doivent se concentrer.

Résultat souhaité : Bénéficiez d'une vue globale des requêtes circulant dans votre système distribué, ce qui permet un débogage précis, des performances optimisées et une meilleure expérience utilisateur.

Anti-modèles courants :

- Instrumentation incohérente : les services d'un système distribué ne sont pas tous instrumentés pour le suivi.
- Ignorer la latence : se concentrer uniquement sur les erreurs et ne pas tenir compte de la latence ou de la dégradation progressive des performances.

Avantages liés au respect de cette bonne pratique :

- Vue d'ensemble complète du système : visualisation du parcours complet des requêtes, de l'entrée à la sortie.
- Débogage amélioré : identification rapide des défaillances ou des problèmes de performance.
- Expérience utilisateur améliorée : surveillance et optimisation basées sur des données sur les utilisateurs réels, afin de garantir que le système répond aux exigences du monde réel.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Commencez par identifier tous les éléments de votre charge de travail qui nécessitent de l'instrumentation. Une fois que tous les composants sont pris en compte, utilisez des outils tels qu'OpenTelemetry et AWS X-Ray pour collecter des données de suivi à des fins d'analyse à l'aide d'outils comme X-Ray et Amazon CloudWatch ServiceLens Map. Participez à des révisions régulières avec les développeurs et complétez ces discussions avec des outils comme Amazon DevOps Guru, X-Ray Analytics et X-Ray Insights pour découvrir des résultats plus approfondis.

Définissez des alertes à partir des données de suivi pour envoyer une notification lorsque les résultats, tels que décrits dans le plan de surveillance de la charge de travail, sont menacés.

Étapes d'implémentation

Pour mettre en œuvre efficacement le suivi distribué :

1. Adoptez AWS X-Ray : Intégrez X-Ray à votre application pour mieux comprendre son comportement, interpréter ses performances et identifier les goulots d'étranglement. Utilisez X-Ray Insights pour l'analyse automatique des données de suivi.
2. Instrumentez vos services : Vérifiez que chaque service, d'une fonction AWS Lambda à une instance EC2, envoie des données de suivi. Plus vous instrumentez de services, plus la vue de bout en bout est claire.
3. Incorporez la surveillance des utilisateurs réels CloudWatch et la surveillance synthétique : Intégrez la surveillance des utilisateurs réels (RUM) et la surveillance synthétique avec X-Ray. Cela permet de capturer des expériences utilisateur réelles et de simuler les interactions des utilisateurs afin d'identifier les problèmes potentiels.
4. Utilisez l'agent CloudWatch : Cet agent peut envoyer des données de suivi depuis X-Ray ou OpenTelemetry, améliorant ainsi la profondeur des informations obtenues.
5. Utilisez Amazon DevOps Guru : DevOps Guru utilise des données provenant de X-Ray, CloudWatch, AWS Config et AWS CloudTrail pour fournir des recommandations exploitables.
6. Analysez les données de suivi : Passez régulièrement en revue les données de suivi pour identifier les tendances, les anomalies ou les goulots d'étranglement susceptibles d'avoir un impact sur les performances de votre application.
7. Configurez des alertes : Configurez des alarmes dans CloudWatch pour les tendances inhabituelles ou les latences prolongées, ce qui permet de résoudre les problèmes de manière proactive.
8. Amélioration continue : Revoyez votre stratégie de suivi au fur et à mesure que des services sont ajoutés ou modifiés afin de capturer tous les points de données pertinents.

Niveau d'effort du plan d'implémentation : Moyen

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identification des indicateurs clés de performance

- OPS04-BP02 Mise en œuvre de la télémétrie de l'application
- OPS04-BP03 Mise en œuvre de la télémétrie de l'expérience utilisateur
- OPS04-BP04 Mise en œuvre de la télémétrie des dépendances

Documents connexes :

- Guide du développeur AWS X-Ray
- Guide de l'utilisateur de l'agent Amazon CloudWatch
- Guide de l'utilisateur Amazon DevOps Guru

Vidéos connexes :

- Use AWS X-Ray Insights
- AWS on Air ft. Observability: Amazon CloudWatch and AWS X-Ray

Exemples connexes :

- Instrumentation de votre application avec AWS X-Ray

OPS 5. Comment réduire les défauts, faciliter les corrections et améliorer le flux dans la production ?

Adoptez des approches qui améliorent l'entrée des modifications en production et qui permettent une refactorisation, des retours rapides sur la qualité et la correction de bogues. Cela permet d'accélérer l'entrée des modifications bénéfiques en production, de limiter le déploiement de problèmes et d'identifier et de corriger rapidement les problèmes introduits par les activités de déploiement.

Bonnes pratiques

- OPS05-BP01 Utilisation du contrôle de version
- OPS05-BP02 Tester et valider les modifications
- OPS05-BP03 Utiliser des systèmes de gestion de la configuration
- OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement
- OPS05-BP05 Procéder à la gestion des correctifs
- OPS05-BP06 Partager les normes de conception

- OPS05-BP07 Mettre en œuvre des pratiques visant à améliorer la qualité du code
- OPS05-BP08 Utiliser plusieurs environnements
- OPS05-BP09 Effectuer des modifications fréquentes, légères et réversibles
- OPS05-BP10 Automatiser complètement l'intégration et le déploiement

OPS05-BP01 Utilisation du contrôle de version

Utilisez le contrôle de version pour activer le suivi des modifications et des versions.

De nombreux services AWS offrent des fonctionnalités de contrôle de version. Utilisez un système de contrôle de source ou de révision comme AWS CodeCommit pour gérer le code et d'autres artefacts, tels que les modèles AWS CloudFormation avec contrôle de version de votre infrastructure.

Résultat souhaité : Vos équipes collaborent sur le code. Une fois fusionné, le code est cohérent et aucune modification n'est perdue. Les erreurs sont facilement corrigées grâce à une gestion des versions appropriée.

Anti-modèles courants :

- Vous avez développé et stocké le code sur votre poste de travail. Un problème de stockage s'est produit sur le poste de travail et vous avez perdu le code.
- Après avoir remplacé le code existant par vos modifications, vous redémarrez votre application et elle n'est plus utilisable. Vous ne pouvez pas annuler la modification.
- Vous disposez d'un verrou d'écriture sur un fichier de rapport que quelqu'un d'autre doit modifier. Il vous contacte pour vous demander d'arrêter de travailler dessus afin qu'il puisse effectuer ses tâches.
- Votre équipe de recherche a travaillé sur une analyse détaillée qui façonnera vos futurs travaux. Quelqu'un a accidentellement enregistré sa liste d'achats sur le rapport final. Vous ne pouvez pas annuler la modification et vous devrez recréer le rapport.

Avantages liés au respect de cette bonne pratique : En utilisant les fonctionnalités de contrôle de version, vous pouvez revenir facilement aux bons états connus et aux versions précédentes, et limiter le risque de perte de ressources.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Maintenez les ressources dans des référentiels avec contrôle de version. Cela permet le suivi des modifications, le déploiement de nouvelles versions, la détection des modifications apportées aux versions existantes, et le retour à des versions antérieures (par exemple, la restauration à un état correct connu en cas de défaillance). Intégrez les fonctionnalités de contrôle de version de vos systèmes de gestion de la configuration dans vos procédures.

Ressources

Bonnes pratiques associées :

- OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement

Documents connexes :

- Qu'est-ce qu'AWS CodeCommit ?

Vidéos connexes :

- Présentation d'AWS CodeCommit

OPS05-BP02 Tester et valider les modifications

Chaque changement déployé doit être testé pour éviter des erreurs de production. Cette bonne pratique est axée sur les tests des changements du contrôle des versions à la création d'artefacts. En plus des changements du code de l'application, les tests doivent inclure l'infrastructure, la configuration, les contrôles de sécurité et les procédures opérationnelles. Les tests peuvent prendre de nombreuses formes, des tests unitaires à l'analyse des composants d'un logiciel (SCA). Pousser les tests encore plus loin dans le processus d'intégration et de livraison de logiciels entraîne une plus grande certitude de la qualité des artefacts.

Votre organisation doit développer des normes de test pour tous les artefacts logiciels. Les tests automatisés réduisent la quantité de travail et évitent les erreurs de test manuels. Des tests manuels peuvent être nécessaires dans certains cas. Les développeurs doivent avoir accès aux résultats des tests automatisés pour créer des boucles de rétroaction qui améliorent la qualité du logiciel.

Résultat souhaité : Les changements apportés au logiciel sont testés avant d'être livrés. Les développeurs ont accès aux résultats des tests et aux validations. Votre organisation a une norme de test qui s'applique à tous les changements apportés au logiciel.

Anti-modèles courants :

- Vous déployez un nouveau changement apporté au logiciel sans aucun test. Il ne s'exécute pas en production, ce qui entraîne une panne.
- De nouveaux groupes de sécurité sont déployés avec AWS CloudFormation sans être testés dans un environnement de préproduction. Les groupes de sécurité empêchent les clients d'atteindre votre application.
- Une méthode est modifiée mais il n'existe aucun test d'unité. Le logiciel échoue quand il est déployé en production.

Avantages liés au respect de cette bonne pratique : Le taux d'échec des changements lors des déploiements de logiciel diminue. La qualité du logiciel s'améliore. Les développeurs ont une meilleure connaissance de la viabilité de leur code. Des politiques de sécurité peuvent être déployées en toute confiance pour soutenir la conformité de l'organisation. Les changements apportés à l'infrastructure, tels que les mises à jour de la politique de mise à l'échelle automatique, sont testés à l'avance pour répondre aux besoins du trafic.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Des tests sont réalisés sur tous les changements, du code de l'application à l'infrastructure, dans le cadre de votre pratique d'intégration continue. Les résultats des tests sont publiés afin que les développeurs disposent d'une rétroaction rapide. Votre organisation a une norme de test que tous les changements doivent respecter.

Exemple client

Dans le cadre de leur pipeline d'intégration continue, la société XYZ réalise plusieurs types de test sur tous les artefacts logiciels. L'entreprise pratique le développement axé sur les tests afin que tous les logiciels bénéficient de tests d'unités. Une fois l'artefact créé, elle exécute des tests de bout en bout. Une fois cette première série de tests terminée, elle exécute une analyse de la sécurité des applications statiques qui cherchent des vulnérabilités connues. Les développeurs reçoivent des messages indiquant que chaque pallier de test est validé. Une fois tous les tests terminés, l'artefact logiciel est stocké dans un référentiel d'artefacts.

Étapes d'implémentation

1. Collaborez avec les parties prenantes dans votre organisation pour développer une norme de test pour les artefacts logiciels. Quels tests standards tous les artefacts doivent-ils valider ? Des exigences en termes de conformité ou de réglementation doivent-elles être incluses dans la couverture des tests ? Faut-il réaliser des tests de qualité du code ? Qui doit être informé de la fin des tests ?
   a. L'architecture de référence des pipelines de déploiement d'AWS contient une liste officielle des types de tests qui peuvent être réalisés sur des artefacts logiciels dans le cadre d'un pipeline d'intégration.
2. Instrumentez votre application avec les tests nécessaires en fonction de la norme de test de votre logiciel. Chaque ensemble de tests doit être réalisé en moins de dix minutes. Les tests doivent être exécutés dans le cadre d'un pipeline d'intégration.
   a. Amazon CodeGuru Reviewer peut tester le code de votre application afin de détecter les défauts.
   b. Vous pouvez utiliser AWS CodeBuild pour réaliser des tests sur les artefacts logiciels.
   c. AWS CodePipeline peut orchestrer vos tests logiciels dans un pipeline.

Ressources

Bonnes pratiques associées :

- OPS05-BP01 Utilisation du contrôle de version
- OPS05-BP06 Partager les normes de conception
- OPS05-BP10 Automatiser complètement l'intégration et le déploiement

Documents connexes :

- Adopter une approche de développement axé sur les tests
- Pipeline de test AWS CloudFormation automatisé avec TaskCat et CodePipeline
- Création d'un pipeline CI/CD AWS DevSecOps de bout en bout avec les outils open source SCA, SAST et DAST
- Démarrer avec les applications de test sans serveur
- Le pipeline d'intégration et de livraison continues comme pierre angulaire de la cohérence du code
- Livre blanc Mise en pratique de l'intégration continue/livraison continue sur AWS

Vidéos connexes :

- AWS re:Invent 2020: Testable infrastructure: Integration testing on AWS
- AWS Summit ANZ 2021 - Driving a test-first strategy with CDK and test driven development
- Testing Your Infrastructure as Code with AWS CDK

Ressources connexes :

- Architecture de référence des pipelines de déploiement d'AWS : application
- Pipeline AWS Kubernetes DevSecOps
- Atelier Politique en tant que code : développement axé sur les tests
- Exécution de tests d'unités pour une application Node.js de GitHub à l'aide d'AWS CodeBuild
- Utilisation de Serverspec pour le développement axé sur les tests du code d'infrastructure

Services associés :

- Amazon CodeGuru Reviewer
- AWS CodeBuild
- AWS CodePipeline

OPS05-BP03 Utiliser des systèmes de gestion de la configuration

Utilisez des systèmes de gestion de la configuration pour effectuer et suivre les modifications de la configuration. Ces systèmes réduisent les erreurs causées par les processus manuels et diminuent le niveau d'effort nécessaire au déploiement des modifications.

La gestion de la configuration statique définit des valeurs lors de l'initialisation d'une ressource. Elles doivent rester cohérentes tout au long de la durée de vie de cette ressource. Certains exemples incluent la définition de la configuration d'un serveur web ou d'applications sur une instance, ou la définition de la configuration d'un service AWS dans AWS Management Console ou via l'interface AWS CLI.

La gestion dynamique de la configuration définit des valeurs à l'initialisation qui peuvent ou sont censées changer pendant la durée de vie d'une ressource. Par exemple, vous pouvez définir un mécanisme d'activation et de désactivation d'une fonctionnalité dans votre code via un changement de configuration, ou modifier le niveau de détail des journaux pendant un incident pour capturer plus

de données, puis revenir en arrière après l'incident en éliminant les journaux désormais inutiles et les dépenses associées.

Sur AWS, vous pouvez utiliser AWS Config pour surveiller en permanence vos configurations de ressources AWS entre les comptes et les régions. Il vous permet de suivre leur historique de configuration, de comprendre comment une modification de la configuration affecterait d'autres ressources et de les auditer par rapport aux configurations attendues ou souhaitées avec AWS Config Rules et les packs de conformité AWS Config.

Si des configurations dynamiques sont appliquées à vos applications exécutées sur des instances Amazon EC2, AWS Lambda, des conteneurs, des applications mobiles ou des appareils IoT, vous pouvez utiliser AWS AppConfig pour les configurer, les valider, les déployer et les surveiller dans l'ensemble de vos environnements.

Dans AWS, vous pouvez créer des pipelines d'intégration continue/de déploiement continu (CI/CD) à l'aide de services tels que les Outils pour développeurs AWS (par exemple, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeployet AWS CodeStar).

Résultat souhaité : Vous effectuez la configuration, la validation et le déploiement dans le cadre de votre pipeline d'intégration et de livraison continues (CI/CD). Vous assurez la surveillance pour vérifier que les configurations sont correctes. Cela permet de minimiser l'impact sur les utilisateurs finaux et les clients.

Anti-modèles courants :

- Vous mettez manuellement à jour la configuration des serveurs Web de votre flotte, et un certain nombre de serveurs ne répondent plus en raison d'erreurs de mise à jour.
- Vous mettez à jour manuellement votre flotte de serveurs d'applications pendant plusieurs heures. L'incohérence de la configuration pendant la modification entraîne des comportements inattendus.
- Quelqu'un a mis à jour vos groupes de sécurité et vos serveurs Web ne sont plus accessibles. Sans savoir ce qui a changé, vous passez beaucoup de temps à enquêter sur la question, ce qui prolonge votre temps de reprise.
- Vous mettez en production une configuration de pré-production via le pipeline CI/CD sans validation. Vous exposez les utilisateurs et les clients à des données et à des services incorrects.

Avantages liés au respect de cette bonne pratique : L'adoption de systèmes de gestion de la configuration réduit le niveau d'effort nécessaire pour effectuer et suivre les changements, ainsi que la fréquence des erreurs causées par les procédures manuelles. Les systèmes de gestion de la

configuration fournissent des garanties en matière de gouvernance, de conformité et d'exigences réglementaires.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Les systèmes de gestion de la configuration sont utilisés pour suivre et mettre en œuvre les modifications apportées aux configurations des applications et de l'environnement. Ils sont également utilisés pour réduire les erreurs causées par les processus manuels, pour rendre les modifications de configuration reproductibles et vérifiables, et pour réduire le niveau d'effort.

Étapes d'implémentation

1. Identifiez les responsables de la configuration.
   a. Informez les responsables de la configuration de tout besoin en matière de conformité, de gouvernance ou de réglementation.
2. Identifiez les éléments de configuration et les livrables.
   a. Les éléments de configuration sont toutes les configurations d'application et d'environnement concernées par un déploiement au sein de votre pipeline CI/CD.
   b. Les livrables incluent les critères de réussite, la validation et ce qui doit être surveillé.
3. Sélectionnez les outils de gestion de la configuration en fonction des besoins de votre entreprise et de votre pipeline de livraison.
4. Envisagez des déploiements pondérés tels que les déploiements canary pour les modifications de configuration importantes, afin de minimiser l'impact des configurations incorrectes.
5. Intégrez la gestion de votre configuration dans votre pipeline CI/CD.
6. Validez toutes les modifications apportées.

Ressources

Bonnes pratiques associées :

- OPS06-BP01 Planifier les modifications infructueuses
- OPS06-BP02 Déploiements de tests
- OPS06-BP03 Adopter des stratégies de déploiement sûres
- OPS06-BP04 Automatiser les tests et les restaurations

Documents connexes :

- AWS Control Tower
- Accélérateur de zone de destination AWS
- AWS Config
- Qu'est-ce qu'AWS Config ?
- AWS AppConfig
- Qu'est-ce qu'AWS CloudFormation ?
- Outils pour développeurs AWS

Vidéos connexes :

- AWS re:Invent 2022 - Proactive governance and compliance for AWS workloads
- AWS re:Invent 2020: Achieve compliance as code using AWS Config
- Manage and Deploy Application Configurations with AWS AppConfig

OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement

Utilisez des systèmes de gestion du développement et du déploiement. Ces systèmes réduisent les erreurs causées par les processus manuels et diminuent le niveau d'effort nécessaire au déploiement des modifications.

Dans AWS, vous pouvez créer des pipelines d'intégration continue/de déploiement continu (CI/CD) à l'aide de services tels que les Outils pour développeurs AWS (par exemple, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeployet AWS CodeStar).

Résultat souhaité : Vos systèmes de gestion du développement et du déploiement prennent en charge le système d'intégration et de livraison continues (CI/CD) de votre entreprise, qui fournit des fonctionnalités permettant d'automatiser des déploiements sécurisés avec les configurations appropriées.

Anti-modèles courants :

- Après avoir compilé votre code sur votre système de développement, vous copiez l'exécutable sur vos systèmes de production et il ne démarre pas. Les fichiers journaux locaux indiquent qu'il n'a pas fonctionné en raison de dépendances manquantes.

- Vous créez avec succès votre application avec de nouvelles fonctionnalités dans votre environnement de développement et soumettez le code à l'assurance qualité (QA). L'assurance qualité échoue, car il manque des ressources statiques.
- Vendredi, après de nombreux efforts, vous avez réussi à créer manuellement votre application dans votre environnement de développement, y compris vos nouvelles fonctionnalités codées. Lundi, vous ne pouvez pas répéter les étapes qui vous ont permis de créer votre application avec succès.
- Vous effectuez les tests que vous avez créés pour votre nouvelle version. Ensuite, vous passez la semaine suivante à configurer un environnement de test et à exécuter tous les tests d'intégration existants, suivis des tests de performances. Le nouveau code a un impact inacceptable sur les performances et doit être redéveloppé, puis retesté.

Avantages liés au respect de cette bonne pratique : En fournissant des mécanismes pour gérer les activités de construction et de déploiement, vous réduisez le niveau d'effort nécessaire pour effectuer des tâches répétitives, vous libérez les membres de votre équipe pour qu'ils puissent se concentrer sur leurs tâches créatives de grande valeur et vous limitez l'introduction d'erreurs provenant des procédures manuelles.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Les systèmes de gestion du développement et du déploiement sont utilisés pour suivre et mettre en œuvre les modifications, réduire les erreurs causées par les processus manuels et limiter le niveau d'effort requis pour des déploiements sûrs. Automatisez entièrement le pipeline d'intégration et de déploiement à partir du code d'enregistrement et par le biais du développement, des tests, du déploiement et de la validation. Cela permet de réduire les délais, de diminuer les coûts, d'augmenter la fréquence des modifications, de limiter le niveau d'effort et d'accroître la collaboration.

Étapes d'implémentation

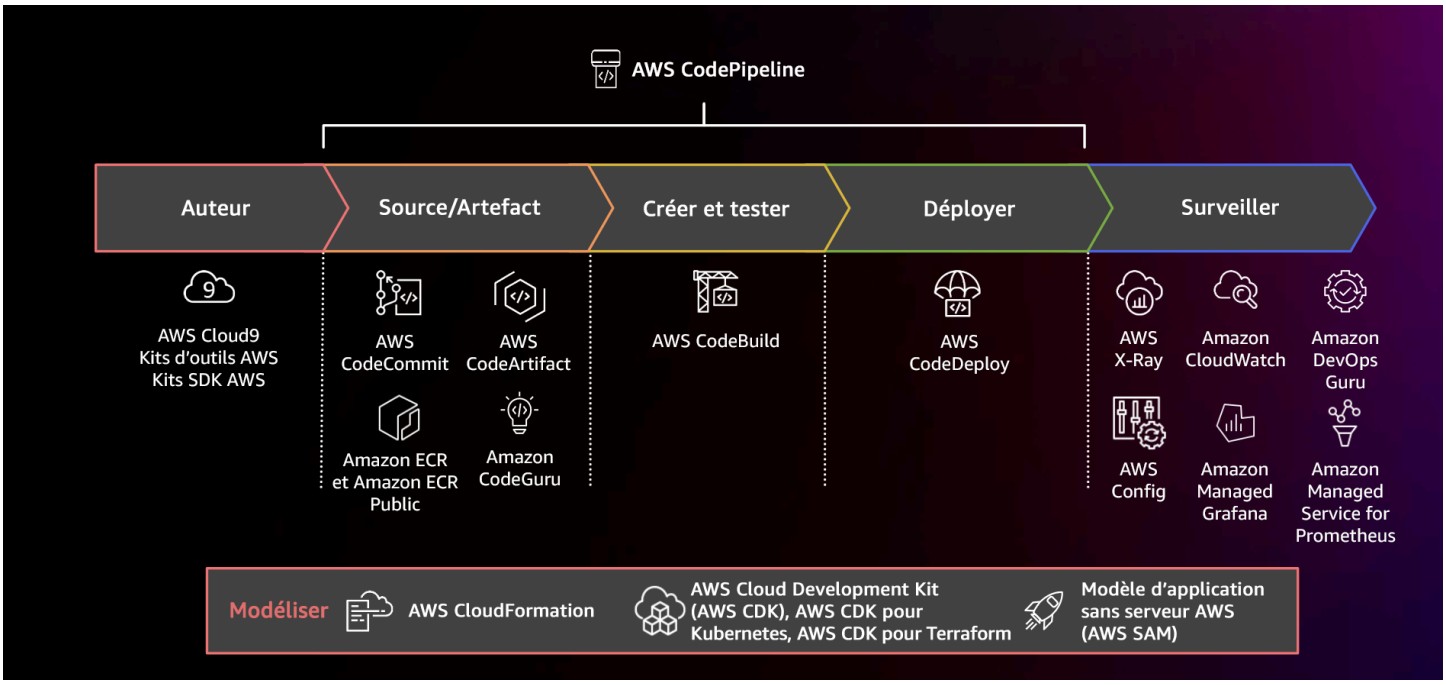


Schéma illustrant un pipeline CI/CD utilisant AWS CodePipeline et des services connexes

1. Utilisez AWS CodeCommit pour contrôler les versions, stocker et gérer les ressources (tels que des documents, du code source et des fichiers binaires).
2. Utilisez CodeBuild pour compiler votre code source, exécuter des tests unitaires et produire des artefacts prêts à être déployés.
3. Utilisez CodeDeploy en tant que service de déploiement qui automatise les déploiements d'applications pour les instances Amazon EC2 , les instances sur site, les fonctions AWS Lambda sans serveurou Amazon ECS.
4. Surveillez vos déploiements.

Ressources

Bonnes pratiques associées :

- OPS06-BP04 Automatiser les tests et les restaurations

Documents connexes :

- Outils pour développeurs AWS

- Qu'est-ce qu'AWS CodeCommit ?
- Qu'est-ce qu'AWS CodeBuild ?
- AWS CodeBuild
- Qu'est-ce qu'AWS CodeDeploy ?

Vidéos connexes :

- AWS re:Invent 2022 - AWS Well-Architected best practices for DevOps on AWS

OPS05-BP05 Procéder à la gestion des correctifs

Procédez à la gestion des correctifs afin de profiter des fonctionnalités, de résoudre les problèmes et de rester conforme à la gouvernance. Automatisez la gestion des correctifs pour réduire les erreurs causées par les processus manuels, pour permettre la mise à l'échelle et pour réduire le niveau d'efforts nécessaire aux correctifs.

La gestion des correctifs et des vulnérabilités fait partie de vos activités de gestion des bénéfices et des risques. Il est préférable d'avoir des infrastructures immuables et de déployer des charges de travail dans des états de bon fonctionnement connus et vérifiés. Lorsque cela n'est pas viable, l'application de correctifs est la seule solution.

Amazon EC2 Image Builder fournit des pipelines pour mettre à jour les images des machines. Dans le cadre de la gestion des correctifs, envisagez d'utiliser des Amazon Machine Images (AMI) à l'aide d'un pipeline d'images AMI ou des images de conteneurs avec un pipeline d'images Docker, tandis qu'AWS Lambda fournit des modèles pour des exécutions personnalisées et des bibliothèques supplémentaires pour supprimer les vulnérabilités.

Vous devez gérer les mises à jour des Amazon Machine Images pour les images Linux ou Windows Server à l'aide de Amazon EC2 Image Builder. Vous pouvez utiliser Amazon Elastic Container Registry (Amazon ECR) avec votre pipeline existant pour gérer les images Amazon ECS et Amazon EKS. Lambda inclut des fonctions de gestion des versions..

L'application de correctifs ne doit pas être effectuée sur les systèmes de production sans avoir effectué un test préalable dans un environnement sûr. Les correctifs ne doivent être appliqués que s'ils favorisent la réalisation d'un résultat opérationnel ou métier. Sur AWS, vous pouvez utiliser le gestionnaire de correctifs AWS Systems Manager pour automatiser le processus d'application des correctifs sur les systèmes gérés et planifier l'activité à l'aide des fenêtres de maintenance Systems Manager..

Résultat souhaité : Vos images AMI et images de conteneur sont corrigées, mises à jour et prêtes à être lancées. Vous pouvez suivre l'état de toutes les images déployées et déterminer la conformité des correctifs. Vous êtes en mesure de rendre compte de l'état actuel et de disposer d'un processus pour répondre à vos besoins en matière de conformité.

Anti-modèles courants :

- On vous demande d'appliquer tous les nouveaux correctifs de sécurité dans un délai de deux heures, ce qui entraîne de multiples pannes dues à l'incompatibilité de l'application avec les correctifs.
- Une bibliothèque non corrigée entraîne des conséquences imprévues, car des parties inconnues y utilisent des failles pour accéder à votre charge de travail.
- Vous corrigez automatiquement les environnements de développement sans en informer les développeurs. Vous recevez plusieurs réclamations des développeurs indiquant que leur environnement ne fonctionne plus correctement.
- Vous n'avez pas corrigé le logiciel sur une instance persistante. Lorsque vous rencontrez un problème avec le logiciel et que vous contactez le fournisseur, celui-ci vous informe que la version n'est pas prise en charge et que vous devez effectuer appliquer un correctif à un niveau spécifique pour recevoir de l'aide.
- Un correctif récemment publié pour le logiciel de chiffrement que vous avez utilisé présente des améliorations significatives de performances. Votre système non corrigé présente des problèmes de performances qui persistent suite à l'absence de correctifs.
- Vous êtes averti d'une vulnérabilité de type « jour zéro » nécessitant une correction d'urgence et vous devez corriger manuellement tous vos environnements.

Avantages liés au respect de cette bonne pratique : En établissant un processus de gestion des correctifs, y compris vos critères d'application des correctifs et la méthodologie de distribution dans vos environnements, vous pouvez adapter les niveaux de correctifs et créer des rapports sur ces niveaux. Cela fournit des garanties concernant les correctifs de sécurité et assure une visibilité claire sur l'état des correctifs connus en cours de mise en place. Cela encourage aussi l'adoption des fonctions et fonctionnalités désirées, l'élimination rapide des problèmes et le respect durable de la gouvernance. Mettez en œuvre des systèmes de gestion des correctifs et d'automatisation pour réduire le niveau d'effort nécessaire au déploiement des correctifs et limiter les erreurs causées par les processus manuels.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Appliquez des correctifs aux systèmes pour corriger les problèmes, obtenir des fonctionnalités souhaitées et rester conforme à la politique de gouvernance et aux exigences d'assistance du fournisseur. Dans les systèmes immuables, déployez avec l'ensemble de correctifs approprié pour obtenir le résultat souhaité. Automatisez le mécanisme de gestion des correctifs afin de réduire le temps écoulé avant l'application des correctifs, d'éviter les erreurs causées par les processus manuels et de limiter le niveau d'efforts nécessaire pour appliquer les correctifs.

Étapes d'implémentation

Pour Amazon EC2 Image Builder :

1. Avec Amazon EC2 Image Builder, spécifiez les détails du pipeline :
    a. Créez un pipeline d'images et nommez-le.
    b. Définissez le calendrier et le fuseau horaire du pipeline.
    c. Configurez toutes les dépendances.
2. Choisissez une recette :
    a. Sélectionnez une recette existante ou créez-en une.
    b. Sélectionnez le type d'image.
    c. Donnez un nom et une version à votre recette.
    d. Sélectionnez votre image de base.
    e. Ajoutez des composants de build et incluez-les dans le registre cible.
3. Facultatif : définissez la configuration de votre infrastructure.
4. Facultatif : définissez les paramètres de configuration.
5. Vérifiez les paramètres.
6. Gérez régulièrement l'hygiène des recettes.

Pour le gestionnaire de correctifs d'Systems Manager:

1. Créez un référentiel de correctifs.
2. Sélectionnez une méthode d'opérations de cheminement.
3. Activez le reporting et l'analyse de conformité.

Ressources

Bonnes pratiques associées :

- OPS06-BP04 Automatiser les tests et les restaurations

Documents connexes :

- Qu'est-ce qu'Amazon EC2 Image Builder ?
- Création d'un pipeline d'images à l'aide d'Amazon EC2 Image Builder
- Création d'un pipeline d'images de conteneurs
- AWS Systems Manager Patch Manager
- Utilisation du gestionnaire de correctifs
- Utilisation des rapports de conformité des correctifs
- Outils pour développeurs AWS

Vidéos connexes :

- CI/CD for Serverless Applications on AWS
- Design with Ops in Mind

Exemples connexes :

- Ateliers Well-Architected : inventaire et gestion des correctifs
- Tutoriels du gestionnaire de correctifs d'AWS Systems Manager

OPS05-BP06 Partager les normes de conception

Partagez les bonnes pratiques entre les équipes pour sensibiliser et maximiser les bénéfices des efforts de développement. Documentez-les et mettez-les à jour au fur et à mesure de l'évolution de votre architecture. Si votre organisation applique des normes partagées, il est essentiel de prévoir des mécanismes permettant de demander des ajouts, des modifications et des exceptions aux normes. Sans cette possibilité, les normes deviennent une contrainte à l'innovation.

Résultat souhaité : Les normes de conception sont partagées par toutes les équipes de votre organisation. Elles sont documentées et mises à jour au fur et à mesure de l'évolution des bonnes pratiques.

Anti-modèles courants :

- Deux équipes de développement ont chacune créé un service d'authentification des utilisateurs. Vos utilisateurs doivent conserver un ensemble distinct d'informations d'identification pour chaque partie du système à laquelle ils veulent accéder.
- Chaque équipe gère sa propre infrastructure. Une nouvelle exigence de conformité impose une modification de votre infrastructure et chaque équipe la met en œuvre de manière différente.

Avantages liés au respect de cette bonne pratique : L'utilisation de normes communes favorise l'adoption de bonnes pratiques et maximise les avantages des efforts de développement. La documentation et la mise à jour des normes de conception permettent à votre organisation de rester à jour par rapport aux bonnes pratiques et aux exigences de sécurité et de conformité.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Partagez les bonnes pratiques existantes, les normes de conception, les listes de contrôle, les procédures d'exploitation, les conseils et les exigences de gouvernance entre les équipes. Prévoyez des procédures pour demander des modifications, des ajouts et des exceptions aux normes de conception afin de favoriser l'amélioration et l'innovation. Assurez-vous que les équipes sont au courant du contenu publié. Prévoyez un mécanisme permettant de mettre à jour les normes de conception au fur et à mesure que de nouvelles bonnes pratiques apparaissent.

Exemple client

La société XYZ dispose d'une équipe d'architecture interfonctionnelle qui crée des modèles d'architecture logicielle. Cette équipe construit l'architecture en y intégrant les aspects de conformité et de gouvernance. Les équipes qui adoptent ces normes communes bénéficient des avantages de la conformité et de la gouvernance intégrées. Elles peuvent rapidement s'appuyer sur la norme de conception. L'équipe d'architecture se réunit tous les trimestres pour évaluer les modèles d'architecture et les mettre à jour si nécessaire.

Étapes d'implémentation

1. Identifiez une équipe interfonctionnelle qui sera chargée de développer et de mettre à jour les normes de conception. Cette équipe travaillera avec les parties prenantes de votre organisation pour élaborer des normes de conception, des procédures d'exploitation, des listes de contrôle, des

conseils et des exigences de gouvernance. Documentez les normes de conception et partagez-les au sein de votre organisation.

   a. AWS Service Catalog permet de créer des portefeuilles représentant les normes de conception en utilisant l'infrastructure en tant que code. Vous pouvez partager des portefeuilles entre plusieurs comptes.

2. Prévoyez un mécanisme permettant de mettre à jour les normes de conception au fur et à mesure que de nouvelles bonnes pratiques sont identifiées.
3. Si les normes de conception sont appliquées de manière centralisée, il faut prévoir un processus pour demander des modifications, des mises à jour et des exemptions.

Niveau d'effort du plan d'implémentation : moyen. L'élaboration d'un processus de création et de partage des normes de conception peut nécessiter une coordination et une coopération avec les parties prenantes de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS01-BP03 Évaluer les exigences de gouvernance - Les exigences de gouvernance influencent les normes de conception.
- OPS01-BP04 Évaluer les exigences de conformité - La conformité est un élément essentiel de la création de normes de conception.
- OPS07-BP02 Assurer un examen cohérent de l'état de préparation opérationnelle - Les listes de contrôle de la disponibilité opérationnelle constituent un mécanisme de mise en œuvre des normes de conception lors de la conception de votre charge de travail.
- OPS11-BP01 Définir un processus d'amélioration continue - La mise à jour des normes de conception fait partie de l'amélioration continue.
- OPS11-BP04 Gérer les connaissances - Dans le cadre de votre pratique de gestion des connaissances, documentez et partagez les normes de conception.

Documents connexes :

- Automatiser les sauvegardes AWS Backup avec AWS Service Catalog
- AWS Service Catalog Account Factory-Enhanced
- How Expedia Group built Database as a Service (DBaaS) offering using AWS Service Catalog

- Assurer la visibilité sur l'utilisation des modèles d'architecture cloud
- Simplify sharing your AWS Service Catalog portfolios in an AWS Organizations setup

Vidéos connexes :

- AWS Service Catalog – Getting Started
- AWS re:Invent 2020: Manage your AWS Service Catalog portfolios like an expert

Exemples connexes :

- Architecture de référence AWS Service Catalog
- Atelier AWS Service Catalog

Services associés :

- AWS Service Catalog

OPS05-BP07 Mettre en œuvre des pratiques visant à améliorer la qualité du code

Mettez en place des pratiques pour améliorer la qualité du code et limiter les failles. Parmi les exemples, citons le développement piloté par les tests, les révisions de code, l'adoption de normes et la programmation en binôme. Incorporez ces pratiques dans votre processus d'intégration et de livraison continues.

Résultat souhaité : Votre organisation utilise des bonnes pratiques comme les révisions de code ou la programmation en binôme pour améliorer la qualité du code. Les développeurs et les opérateurs adoptent les bonnes pratiques en matière de qualité du code dans le cadre du cycle de vie du développement logiciel.

Anti-modèles courants :

- Vous livrez du code à la branche principale de votre application sans effectuer de révision du code. La modification est automatiquement déployée en production et provoque une panne.
- Une nouvelle application est développée sans aucun test d'unité, de bout en bout ou d'intégration. Il n'y a aucun moyen de tester l'application avant son déploiement.

- Vos équipes procèdent à des modifications manuelles en production pour corriger les défauts. Les modifications ne sont pas soumises à des tests ou à des révisions de code et ne sont pas saisies ou enregistrées dans le cadre des processus d'intégration et de livraison continues.

Avantages liés au respect de cette bonne pratique : En adoptant des pratiques visant à améliorer la qualité du code, vous contribuez à minimiser les problèmes introduits dans la production. La qualité du code s'améliore grâce aux bonnes pratiques telles que la programmation en binôme et les révisions de code.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Mettez en œuvre des pratiques visant à améliorer la qualité du code afin de minimiser les défauts avant leur déploiement. Utilisez des pratiques telles que le développement piloté par les tests, les révisions de code et la programmation en binôme pour améliorer la qualité de votre développement.

Exemple client

La société XYZ adopte plusieurs pratiques pour améliorer la qualité du code. La société a adopté le développement piloté par les tests comme norme d'écriture des applications. Pour certaines nouvelles fonctionnalités, elle demande aux développeurs de programmer en binôme pendant un sprint. Chaque demande d'extraction est soumise à une révision du code par un développeur principal avant d'être intégrée et déployée.

Étapes d'implémentation

1. Adoptez des pratiques de qualité du code telles que le développement piloté par les tests, les révisions de code et la programmation en binôme dans votre processus d'intégration et de livraison continues. Utilisez ces techniques pour améliorer la qualité des logiciels.
   a. Amazon CodeGuru Reviewer peut fournir des recommandations de programmation pour le code Java et Python en utilisant le machine learning.
   b. Vous pouvez créer des environnements de développement partagés avec AWS Cloud9 où vous pouvez collaborer au développement du code.

Niveau d'effort du plan d'implémentation : moyen. Il existe de nombreuses façons de mettre en œuvre cette bonne pratique, mais il peut être difficile de la faire adopter par l'organisation.

Ressources

Bonnes pratiques associées :

- OPS05-BP06 Partager les normes de conception - Vous pouvez partager les normes de conception dans le cadre de votre pratique de la qualité du code.

Documents connexes :

- Guide du logiciel Agile
- Le pipeline d'intégration et de livraison continues comme pierre angulaire de la cohérence du code
- Automatiser les révisions de code avec Amazon CodeGuru Reviewer
- Adopter une approche de développement axé sur les tests
- Comment DevFactory crée de meilleures applications avec Amazon CodeGuru
- Concernant la programmation en binôme
- RENGA Inc. automatise les révisions de code avec Amazon CodeGuru
- L'art du développement agile : le développement piloté par les tests
- Pourquoi les révisions de code sont importantes (et font gagner du temps !)

Vidéos connexes :

- AWS re:Invent 2020: Continuous improvement of code quality with Amazon CodeGuru
- AWS Summit ANZ 2021 - Driving a test-first strategy with CDK and test driven development

Services associés :

- Amazon CodeGuru Reviewer
- Amazon CodeGuru Profiler
- AWS Cloud9

OPS05-BP08 Utiliser plusieurs environnements

Utilisez plusieurs environnements pour expérimenter, développer et tester votre charge de travail. Utilisez des niveaux de contrôle croissants lorsque les environnements approchent de la production pour vous assurer que votre charge de travail fonctionnera correctement une fois déployée.

Résultat souhaité : Vous disposez de plusieurs environnements qui répondent à vos besoins en matière de conformité et de gouvernance. Vous testez et promouvez le code dans les différents environnements jusqu'à la production.

Anti-modèles courants :

- Vous effectuez un développement dans un environnement de développement partagé et un autre développeur remplace vos modifications de code.
- Les contrôles de sécurité restrictifs sur votre environnement de développement partagé vous empêchent d'expérimenter de nouveaux services et fonctionnalités.
- Vous effectuez des tests de charge sur vos systèmes de production et provoquez une panne pour vos utilisateurs.
- Une erreur critique entraînant une perte de données s'est produite en production. Dans votre environnement de production, vous essayez de recréer les conditions qui ont conduit à la perte de données afin de pouvoir identifier comment elle s'est produite et empêcher qu'elle ne se reproduise. Pour éviter toute perte de données supplémentaire pendant les tests, vous devez rendre l'application indisponible aux utilisateurs.
- Vous explorez un service multi-locataire et n'êtes pas en mesure de répondre à la demande d'un client pour un environnement dédié.
- Il se peut que vous ne réalisiez pas toujours des tests, mais lorsque vous le faites, vous procédez dans votre environnement de production.
- Vous pensez que la simplicité d'un environnement unique l'emporte sur la portée de l'impact des modifications au sein de l'environnement.

Avantages liés au respect de cette bonne pratique : Vous pouvez prendre en charge plusieurs environnements de développement, de test et de production simultanément sans créer de conflits entre les développeurs ou les communautés d'utilisateurs.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Utilisez plusieurs environnements et fournissez aux développeurs des environnements de test (sandbox) avec des contrôles réduits au minimum pour faciliter l'expérimentation. Fournissez des environnements de développement individuels pour permettre le travail en parallèle, ce qui augmente l'agilité du développement. Mettez en œuvre davantage de contrôles rigoureux dans les environnements proches de la production pour offrir aux développeurs la liberté d'innover. Utilisez

l'infrastructure en tant que code et les systèmes de gestion de la configuration pour déployer des environnements configurés de manière cohérente par rapport aux contrôles de production pour veiller au bon fonctionnement des systèmes lorsqu'ils sont déployés. Lorsque les environnements ne sont pas en cours d'utilisation, désactivez-les pour éviter les coûts associés à des ressources inutilisées (par exemple, les systèmes de développement en soirée et les week-ends). Déployez des environnements équivalents à la production lors des tests de charge pour accroître les résultats valides.

Ressources

Documents connexes :

- Instance Scheduler sur AWS
- Qu'est-ce qu'AWS CloudFormation ?

OPS05-BP09 Effectuer des modifications fréquentes, légères et réversibles

Les modifications fréquentes, légères et réversibles limitent la portée et l'impact d'une modification. Lorsqu'elles sont utilisées conjointement avec des systèmes de gestion des modifications, des systèmes de gestion de configuration et des systèmes de construction et de livraison, les modifications fréquentes, mineures et réversibles limitent la portée et l'impact d'une modification. Cela se traduit par une résolution plus efficace des problèmes et par des corrections plus rapides avec la possibilité d'annuler les modifications effectuées.

Anti-modèles courants :

- Vous déployez une nouvelle version de votre application tous les trimestres avec une fenêtre de modification qui signifie qu'un service principal est désactivé.
- Vous modifiez fréquemment le schéma de votre base de données sans suivre les modifications apportées à vos systèmes de gestion.
- Vous effectuez des mises à jour manuelles sur place, en remplaçant les installations et les configurations existantes, sans aucun plan de restauration clair.

Avantages liés au respect de cette bonne pratique : Les efforts de développement sont accélérés en déployant fréquemment de petites modifications. Lorsque les changements sont minimes, il est beaucoup plus facile d'identifier s'ils ont des conséquences inattendues et ils sont plus faciles à annuler. Lorsque les changements sont réversibles, les risques de mise en œuvre d'une modification

sont minimes, car la récupération est simplifiée. Le processus de modification présente un risque réduit et l'impact de l'échec d'une modification est réduit.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

Ayez recours à des modifications fréquentes, légères et réversibles pour limiter leur portée et leur impact. Cela facilite la résolution des problèmes, contribue à accélérer les corrections et offre la possibilité d'annuler une modification. Cela augmente également la vitesse à laquelle vous pouvez apporter de la valeur à votre entreprise.

Ressources

Bonnes pratiques associées :

- OPS05-BP03 Utiliser des systèmes de gestion de la configuration
- OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement
- OPS06-BP04 Automatiser les tests et les restaurations

Documents connexes :

- Implémentation des microservices sur AWS
- Microservices – Observabilité

OPS05-BP10 Automatiser complètement l'intégration et le déploiement

Automatisez la création, le déploiement et le test de la charge de travail. Cela permet de réduire les erreurs découlant des processus manuels, ainsi que les efforts nécessaires au déploiement des modifications.

Appliquez des métadonnées à l'aide des balises de ressources et AWS Resource Groups en respectant une stratégie de balisage cohérente pour faciliter l'identification de vos ressources. Balisez vos ressources pour l'organisation, la comptabilité analytique, les contrôles d'accès et le ciblage de l'exécution des activités d'opérations automatisées.

Résultat souhaité : Les développeurs utilisent des outils pour fournir du code et le promouvoir jusqu'à la production. Les développeurs n'ont pas besoin de se connecter à la AWS Management Console pour fournir des mises à jour. Il existe une piste d'audit complète des modifications et

de la configuration, répondant aux besoins de gouvernance et de conformité. Les processus sont reproductibles et standardisés entre les équipes. Les développeurs sont libres de se concentrer sur le développement et les envois de code, ce qui augmente la productivité.

Anti-modèles courants :

- Vendredi, vous avez fini de créer le code de votre branche de fonctionnalité. Lundi, après avoir exécuté vos scripts de test de la qualité du code et chacun de vos scripts de tests unitaires, vous vérifiez votre code pour la prochaine version prévue.
- Vous êtes chargé de coder un correctif pour un problème critique affectant un grand nombre de clients en production. Après avoir testé le correctif, vous validez votre code et envoyez un e-mail à l'équipe de gestion des modifications pour demander l'autorisation de le déployer en production.
- En tant que développeur, vous vous connectez à la AWS Management Console pour créer un environnement de développement à l'aide de méthodes et de systèmes non standard.

Avantages liés au respect de cette bonne pratique : En mettant en œuvre des systèmes automatisés de gestion de la création et du déploiement, vous réduisez les erreurs causées par les processus manuels et diminuez l'effort de déploiement des changements, ce qui permet aux membres de votre équipe de se concentrer sur la création de valeur ajoutée. Vous accélérez la vitesse de livraison au fur et à mesure que vous progressez jusqu'à la production.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

Utilisez des systèmes de gestion du développement et du déploiement afin de suivre et de mettre en œuvre des modifications, de réduire les erreurs causées par les processus manuels et de réduire le niveau d'efforts. Automatisez entièrement le pipeline d'intégration et de déploiement à partir du code d'enregistrement et par le biais du développement, des tests, du déploiement et de la validation. Cela permet de raccourcir les délais, d'augmenter la fréquence des modifications, de réduire le niveau d'effort, d'accélérer la mise sur le marché, d'augmenter la productivité et de renforcer la sécurité de votre code jusqu'à la production.

Ressources

Bonnes pratiques associées :

- OPS05-BP03 Utiliser des systèmes de gestion de la configuration
- OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement

Documents connexes :

- Qu'est-ce qu'AWS CodeBuild ?
- Qu'est-ce qu'AWS CodeDeploy ?

Vidéos connexes :

- AWS re\:Invent 2022 - AWS Well-Architected best practices for DevOps on AWS

OPS 6. Comment réduire les risques liés au déploiement ?

Adoptez des approches qui fournissent un retour d'information rapide sur la qualité et permettent une reprise rapide à la suite de changements qui n'offrent pas les résultats escomptés. L'utilisation de ces pratiques diminue l'impact des problèmes découlant du déploiement des modifications.

Bonnes pratiques

- OPS06-BP01 Planifier les modifications infructueuses
- OPS06-BP02 Déploiements de tests
- OPS06-BP03 Adopter des stratégies de déploiement sûres
- OPS06-BP04 Automatiser les tests et les restaurations

OPS06-BP01 Planifier les modifications infructueuses

Prévoyez de revenir à un état correct connu ou de remédier à la situation dans l'environnement de production si le déploiement entraîne des résultats indésirables. L'existence d'une politique visant à établir un tel plan aide toutes les équipes à développer des stratégies de récupération en cas d'échec des modifications. Parmi les exemples de politiques, citons les étapes de déploiement et de restauration, les politiques de changement, les indicateurs de fonction, l'isolation du trafic et le déplacement du trafic. Une seule version peut inclure plusieurs modifications de composants connexes. La stratégie doit permettre de résister ou de se remettre d'une défaillance de tout changement de composant.

Résultat souhaité : Vous avez préparé un plan de reprise détaillé pour votre modification en cas d'échec. En outre, vous avez réduit la taille de votre version afin de minimiser l'impact potentiel sur d'autres composants de la charge de travail. Vous avez ainsi réduit l'impact sur l'entreprise en diminuant le temps d'arrêt potentiel causé par une modification ratée et en augmentant la flexibilité et l'efficacité des temps de récupération.

Anti-modèles courants :

- Vous avez effectué un déploiement et votre application est devenue instable, mais il semble qu'il y ait des utilisateurs actifs sur le système. Vous devez décider entre annuler la modification et avoir un impact sur les utilisateurs actifs et attendre pour annuler la modification en sachant que les utilisateurs peuvent être impactés de toute façon.
- Après avoir modifié la routine, vos nouveaux environnements sont accessibles, mais l'un de vos sous-réseaux est devenu inaccessible. Vous devez décider de tout annuler ou d'essayer de réparer le sous-réseau inaccessible. Pendant cette période de détermination, le sous-réseau reste inaccessible.
- Vos systèmes ne sont pas conçus de manière à pouvoir être mis à jour avec de petites versions. Par conséquent, il est difficile d'annuler ces modifications en bloc en cas d'échec du déploiement.
- Vous n'utilisez pas l'infrastructure en tant que code (IaC) et vous avez effectué des mises à jour manuelles de votre infrastructure, ce qui a entraîné une configuration indésirable. Vous n'êtes pas en mesure de suivre et d'annuler efficacement les modifications manuelles.
- Parce que vous n'avez pas mesuré l'augmentation de la fréquence de vos déploiements, votre équipe n'est pas incitée à réduire la taille de ses changements et à améliorer ses plans de restauration pour chaque modification, ce qui entraîne une augmentation des risques et des taux d'échec.
- Vous ne mesurez pas la durée totale d'une panne causée par des modifications infructueuses. Votre équipe n'est pas en mesure d'établir des priorités et d'améliorer l'efficacité de son processus de déploiement et de son plan de reprise.

Avantages liés au respect de cette bonne pratique : Disposer d'un plan de reprise en cas de modifications infructueuses permet de minimiser le temps moyen de récupération (MTTR) et de réduire l'impact sur votre entreprise.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Une politique et une pratique cohérentes et documentées, adoptées par les équipes de publication des versions, permettent à une organisation de planifier ce qui doit se passer en cas d'échec des modifications. La politique devrait permettre la correction à l'avance dans des circonstances spécifiques. Dans les deux cas, un plan de correction à l'avance ou de restauration doit être bien documenté et testé avant d'être déployé dans la production réelle, afin de réduire au minimum la durée nécessaire pour restaurer une modification.

Étapes d'implémentation

1. Documentez les politiques qui exigent des équipes qu'elles disposent de plans efficaces pour restaurer les modifications dans un délai donné.
   a. Les politiques doivent préciser les cas où une situation de correction à l'avance est autorisée.
   b. Exigez qu'un plan de restauration documenté soit accessible à toutes les personnes concernées.
   c. Précisez les conditions de restauration (par exemple, lorsqu'il s'avère que des modifications non autorisées ont été déployées).
2. Analysez le niveau d'impact de toutes les modifications liées à chaque composante d'une charge de travail.
   a. Autorisez les modifications répétitives à être normalisées, modélisées et préautorisées si elles suivent un flux de travail cohérent qui applique les politiques de modification.
   b. Réduisez l'impact potentiel de toute modification en en réduisant la taille, de sorte que la reprise prenne moins de temps et ait moins d'impact sur l'entreprise.
   c. Veillez à ce que les procédures de restauration ramènent le code à l'état correct connu afin d'éviter les incidents dans la mesure du possible.
3. Intégrez des outils et des flux de travail pour appliquer vos politiques de manière programmée.
4. Faites en sorte que les données relatives aux modifications soient visibles pour les autres propriétaires de charges de travail afin d'améliorer la rapidité du diagnostic en cas de modification défaillante impossible à annuler.
   a. Mesurez le degré de réussite de cette pratique à l'aide de données sur les modifications visibles et identifiez les améliorations itératives.
5. Utilisez des outils de surveillance pour vérifier le succès ou l'échec d'un déploiement afin d'accélérer la prise de décision concernant la restauration.
6. Mesurez la durée de l'interruption lors d'un changement infructueux afin d'améliorer continuellement vos plans de reprise.

Niveau d'effort du plan d'implémentation : Moyen

Ressources

Bonnes pratiques associées :

- OPS06-BP04 Automatiser les tests et les restaurations

Documents connexes :

- Builders Library AWS | Exécuter des annulations sûres pendant les déploiements
- Livre blanc AWS | Gestion des modifications dans le cloud

Vidéos connexes :

- re:Invent 2019 | Amazon's approach to high-availability deployment

OPS06-BP02 Déploiements de tests

Testez les procédures de mise à disposition en pré-production en utilisant la même configuration de déploiement, les mêmes contrôles de sécurité, les mêmes étapes et les mêmes procédures qu'en production. Confirmez que toutes les étapes du déploiement se sont déroulées comme prévu, par exemple en inspectant les fichiers, les configurations et les services. Testez ensuite toutes les modifications à l'aide de tests fonctionnels, d'intégration et de charge, ainsi que de contrôles tels que les surveillances de l'état. En effectuant ces tests, vous pouvez identifier rapidement les problèmes de déploiement et avoir la possibilité de les planifier et de les atténuer avant la mise en production.

Vous pouvez créer des environnements parallèles temporaires pour tester chaque modification. Automatisez le déploiement des environnements de test à l'aide de l'infrastructure en tant que code (IaC) afin de réduire la quantité de travail nécessaire et d'assurer la stabilité, la cohérence et une livraison plus rapide des fonctionnalités.

Résultat souhaité : Votre organisation adopte une culture de développement piloté par les tests qui inclut des déploiements de tests. Cela permet de veiller à ce que les équipes se concentrent sur la création de valeur pour l'entreprise plutôt que sur la gestion des versions. Les équipes sont impliquées dès l'identification des risques de déploiement afin de déterminer les mesures d'atténuation appropriées.

Anti-modèles courants :

- Pendant les mises en production, les déploiements non testés entraînent des problèmes fréquents qui nécessitent un dépannage et une remontée.
- Votre version contient une infrastructure sous forme de code (IaC) qui met à jour les ressources existantes. Vous n'êtes pas certain que l'IaC s'exécute correctement ou qu'elle a un impact sur les ressources.

- Vous déployez une nouvelle fonctionnalité dans votre application. Elle ne fonctionne pas comme prévu et il n'y a aucune visibilité jusqu'à ce qu'elle soit signalée par les utilisateurs concernés.
- Vous mettez à jour vos certificats. Vous installez accidentellement les certificats sur les mauvais composants, ce qui passe inaperçu et a un impact sur les visiteurs du site web parce qu'il est impossible d'établir une connexion sécurisée avec le site web.

Avantages liés au respect de cette bonne pratique : Des tests approfondis en pré-production des procédures de déploiement et des modifications qu'elles introduisent minimisent l'impact potentiel sur la production causé par les étapes de déploiement. Cela permet d'accroître la confiance lors de la mise en production et de minimiser l'assistance opérationnelle sans ralentir la vitesse des changements apportés.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Il est tout aussi important de tester votre processus de déploiement que les modifications qui en découlent. Pour ce faire, vous pouvez tester vos étapes de déploiement dans un environnement de pré-production qui reflète le plus fidèlement possible l'environnement de production. Les problèmes courants, tels que les étapes de déploiement incomplètes ou incorrectes, ou les mauvaises configurations, peuvent être détectés avant la mise en production. De plus, vous pouvez tester vos étapes de reprise.

Exemple client

Dans le cadre de son pipeline d'intégration et de livraison continues (CI/CD), AnyCompany Retail exécute les étapes définies nécessaires au lancement de l'infrastructure et des mises à jour logicielles pour ses clients dans un environnement de type production. Le pipeline comprend des contrôles préalables pour détecter les altérations (détection des changements apportés aux ressources en dehors de votre IaC) dans les ressources avant le déploiement, ainsi que pour valider les actions que l'IaC entreprend lors de son lancement. Il valide les étapes du déploiement, en vérifiant par exemple que certains fichiers et configurations sont en place, que les services sont en cours d'exécution et qu'ils répondent correctement aux surveillances de l'état sur l'hôte local avant de s'enregistrer à nouveau auprès de l'équilibreur de charge. En outre, toutes les modifications font l'objet d'un certain nombre de tests automatisés, tels que des tests fonctionnels, de sécurité, de régression, d'intégration et de charge.

Étapes d'implémentation

1. Effectuez des contrôles avant l'installation pour reproduire l'environnement de pré-production en production.
   a. Utilisez la détection des altérations pour détecter si des ressources ont été modifiées en dehors de AWS CloudFormation.
   b. Utilisez des jeux de modifications pour vérifier que l'intention de la mise à jour de la pile correspond aux actions entreprises par AWS CloudFormation lorsque le jeu de modifications est lancé.
2. Cela déclenche une étape d'approbation manuelle dans AWS CodePipeline afin d'autoriser le déploiement dans l'environnement de pré-production.
3. Utilisez les configurations de déploiement telles que les fichiers AppSpec AWS CodeDeploy pour définir les étapes de déploiement et de validation.
4. Le cas échéant, intégrez AWS CodeDeploy à d'autres services AWS ou intégrez AWS CodeDeploy aux produits et services des partenaires. »
5. Surveillez les déploiements à l'aide de Amazon CloudWatch, de AWS CloudTrail et des notifications d'événements Amazon SNS.
6. Réalisez des tests automatisés après déploiement, y compris des tests fonctionnels, de sécurité, de régression, d'intégration et de charge.
7. Résolvez les problèmes de déploiement.
8. La validation réussie des étapes précédentes devrait lancer un mécanisme d'autorisation manuel pour autoriser le déploiement en production.

Niveau d'effort du plan d'implémentation : Élevé

Ressources

Bonnes pratiques associées :

- OPS05-BP02 Tester et valider les modifications

Documents connexes :

- Builders' Library AWS | Automatisation de déploiements sécurisés sans intervention | Déploiements tests
- Livre blanc AWS | Mise en pratique de l'intégration continue et de la livraison continue sur AWS

- The Story of Apollo - Amazon's Deployment Engine
- Comment tester et déboguer AWS CodeDeploy localement avant d'expédier votre code
- Integrating Network Connectivity Testing with Infrastructure Deployment

Vidéos connexes :

- re:Invent 2020 | Testing software and systems at Amazon

Exemples connexes :

- Tutoriel | Déploiement et maintenance Amazon ECS à l'aide d'un test de validation

OPS06-BP03 Adopter des stratégies de déploiement sûres

Les déploiements de production sécurisés contrôlent le flux des modifications bénéfiques dans le but de minimiser l'impact perçu de ces modifications sur les clients. Les contrôles de sécurité fournissent des mécanismes d'inspection permettant de valider les résultats souhaités et de limiter l'étendue de l'impact des défaillances introduites par les modifications ou des échecs de déploiement. Les déploiements sûrs peuvent inclure des stratégies telles que les indicateurs de fonctions, les déploiements sur un seul hôte, les déploiements continus (versions canary), les déploiements immuables, la division du trafic et les déploiements bleu/vert.

Résultat souhaité : Votre organisation utilise un système d'intégration continue et de livraison continue (CI/CD) qui permet d'automatiser des déploiements sûrs. Les équipes sont tenues d'utiliser des stratégies de déploiement sûres et appropriées.

Anti-modèles courants :

- Vous déployez une modification infructueuse dans l'ensemble de l'environnement de production en une seule fois. Par conséquent, tous les clients sont touchés simultanément.
- Une défaillance introduite lors d'un déploiement simultané dans tous les systèmes nécessite un lancement d'urgence. La correction pour tous les clients prend plusieurs jours.
- La gestion des versions de production nécessite la planification et la participation de plusieurs équipes. Cela limite votre capacité à mettre fréquemment à jour les fonctionnalités pour vos clients.
- Vous effectuez un déploiement mutable en modifiant vos systèmes existants. Après avoir découvert que la modification n'a pas abouti, vous devez modifier à nouveau les systèmes pour restaurer l'ancienne version, ce qui prolonge votre délai de récupération.

Avantages liés au respect de cette bonne pratique : Les déploiements automatisés permettent de concilier la rapidité des déploiements et la cohérence des modifications apportées aux clients. Limiter l'impact permet d'éviter des échecs de déploiement coûteux et de maximiser la capacité des équipes à répondre efficacement aux défaillances.

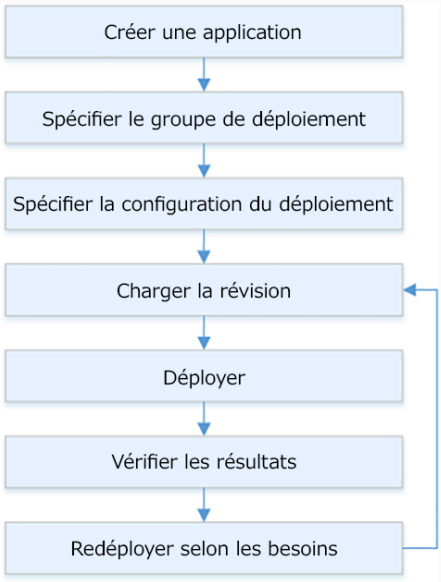
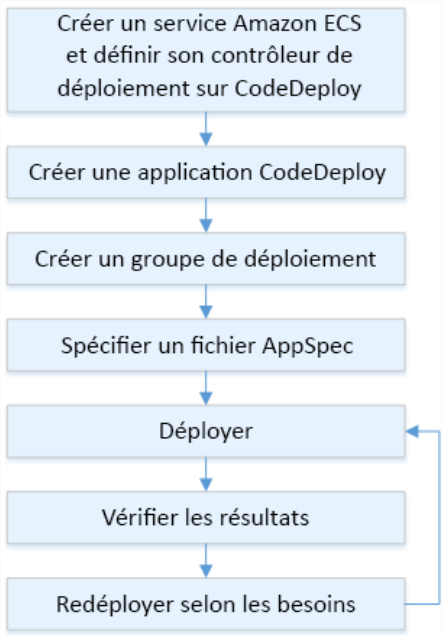
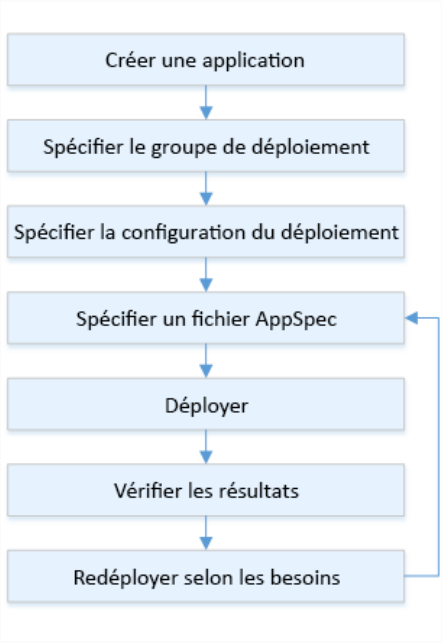
Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Les défaillances de la livraison en continu peuvent entraîner une réduction de la disponibilité des services et de mauvaises expériences pour les clients. Pour maximiser le taux de réussite des déploiements, mettez en œuvre des contrôles de sécurité dans le processus de lancement de bout en bout afin de minimiser les erreurs de déploiement ; l'objectif étant de parvenir à zéro échec de déploiement.

Exemple client

AnyCompany Retail a pour mission de réaliser des déploiements avec un temps d'arrêt minimal ou nul, ce qui signifie qu'il n'y a pas d'impact perceptible pour ses utilisateurs pendant les déploiements. Pour ce faire, l'entreprise a établi des modèles de déploiement (voir le diagramme de flux de travail suivant), tels que les déploiements continus et les déploiements bleu/vert. Toutes les équipes adoptent un ou plusieurs de ces modèles dans leur pipeline CI/CD.

Flux de travail CodeDeploy pour Amazon EC2	Flux de travail CodeDeploy pour Amazon ECS	Flux de travail CodeDeploy pour Lambda
		

Étapes d'implémentation

1. Utilisez un flux de travail d'approbation pour lancer la séquence des étapes de déploiement de la production lors de la promotion en production.
2. Utilisez un système de déploiement automatisé tel que AWS CodeDeploy. Les options de déploiement AWS CodeDeploy comprennent les déploiements sur place pour EC2/sur site et les déploiements bleu/vert pour EC2/sur site, AWS Lambda et Amazon ECS (voir le diagramme de flux de travail précédent).
   a. Le cas échéant, intégrez AWS CodeDeploy à d'autres services AWS ou intégrez AWS CodeDeploy aux produits et services des partenaires. »
3. Utilisez les déploiements bleu/vert pour les bases de données telles que Amazon Aurora et Amazon RDS. »
4. Surveillez les déploiements à l'aide de Amazon CloudWatch, de AWS CloudTrail et des notifications d'événements Amazon SNS.
5. Effectuez des tests automatisés post-déploiement, y compris des tests fonctionnels, de sécurité, de régression, d'intégration et tout test de charge.
6. Résolvez les problèmes de déploiement.

Niveau d'effort du plan d'implémentation : Moyen

Ressources

Bonnes pratiques associées :

- OPS05-BP02 Tester et valider les modifications
- OPS05-BP09 Effectuer des modifications fréquentes, légères et réversibles
- OPS05-BP10 Automatiser complètement l'intégration et le déploiement

Documents connexes :

- Builders' Library AWS | Automatisation de déploiements sécurisés sans intervention | Déploiements en production
- Builders Library AWS | Mon pipeline CI/CD est mon capitaine de versions | Versions de production automatiques et sécurisées
- Livre blanc AWS | Mise en pratique de l'intégration continue et de la livraison continue sur AWS | Méthodes de déploiement
- Guide de l'utilisateur AWS CodeDeploy
- Utilisation des configurations de déploiement AWS CodeDeploy
- Configurer un déploiement de version API Gateway Canary
- Types de déploiement Amazon ECS
- Déploiements bleu/vert entièrement gérés dans Amazon Aurora et Amazon RDS
- Déploiements bleu/vert avec AWS Elastic Beanstalk

Vidéos connexes :

- re:Invent 2020 | Hands-off: Automating continuous delivery pipelines at Amazon
- re:Invent 2019 | Amazon's Approach to high-availability deployment

Exemples connexes :

- Essayer un exemple de déploiement bleu/vert dans AWS CodeDeploy
- Atelier | Création de pipelines CI/CD pour les déploiements Lambda Canary à l'aide de AWS CDK

- Atelier | Blue/Green and Canary Deployment for EKS and ECS
- Atelier | Building a Cross-account CI/CD Pipeline

OPS06-BP04 Automatiser les tests et les restaurations

Pour accroître la rapidité, la fiabilité et la confiance de votre processus de déploiement, mettez en place une stratégie de tests automatisés et de restauration dans les environnements de pré-production et de production. Automatisez les tests lors du déploiement en production afin de simuler les interactions entre l'homme et le système et de vérifier les modifications déployées. Automatisez la restauration pour revenir rapidement à un état antérieur sain connu. La restauration doit être déclenchée automatiquement dans des conditions prédéfinies, par exemple lorsque le résultat souhaité de la modification n'est pas atteint ou lorsque le test automatisé échoue. L'automatisation de ces deux activités améliore le taux de réussite de vos déploiements, minimise le temps de reprise et réduit l'impact potentiel sur l'entreprise.

Résultat souhaité : Vos tests automatisés et vos stratégies de restauration sont intégrés dans votre pipeline d'intégration continue et de livraison continue (CI/CD). Votre surveillance est en mesure de valider vos critères de réussite et de déclencher une restauration automatique en cas d'échec. Cela permet de minimiser l'impact sur les utilisateurs finaux et les clients. Par exemple, lorsque tous les résultats des tests ont été satisfaits, vous transférez votre code dans l'environnement de production où des tests de régression automatisés sont lancés, en utilisant les mêmes cas de test. Si les résultats des tests de régression ne correspondent pas aux attentes, une restauration automatisée est lancée dans le flux de travail du pipeline.

Anti-modèles courants :

- Vos systèmes ne sont pas conçus de manière à pouvoir être mis à jour avec de petites versions. Par conséquent, il est difficile d'annuler ces modifications en bloc en cas d'échec du déploiement.
- Votre processus de déploiement consiste en une série d'étapes manuelles. Après avoir apporté des modifications à votre charge de travail, vous commencez les tests de post-déploiement. Après les tests, vous vous rendez compte que votre charge de travail est inopérante et que les clients sont déconnectés. Vous commencez les opérations pour restaurer la version précédente. Toutes ces étapes manuelles retardent la reprise globale du système et ont un impact prolongé sur vos clients.
- Vous avez passé du temps à développer des cas de tests automatisés pour des fonctionnalités qui ne sont pas fréquemment utilisées dans votre application, minimisant ainsi le retour sur investissement de votre capacité de tests automatisés.

- Votre version est composée de mises à jour d'applications, d'infrastructures, de correctifs et de configurations qui sont indépendantes les unes des autres. Cependant, vous disposez d'un seul pipeline CI/CD qui fournit toutes les modifications en une seule fois. La défaillance d'un composant vous oblige à annuler toutes les modifications, ce qui rend votre restauration complexe et inefficace.
- Votre équipe termine le travail de codage au cours du premier sprint et commence le travail du deuxième sprint, mais votre plan ne prévoyait pas de tests avant le troisième sprint. En conséquence, les tests automatisés ont révélé des défauts du premier sprint qui ont dû être résolus avant que les tests des produits livrables du deuxième sprint puissent commencer et la version entière est retardée, ce qui dévalorise vos tests automatisés.
- Vos tests de régression automatisés pour la version de production sont terminés, mais vous ne surveillez pas l'état de la charge de travail. Comme vous ne savez pas si le service a redémarré ou non, vous ne savez pas si la restauration est nécessaire ou si elle a déjà eu lieu.

Avantages liés au respect de cette bonne pratique : L'automatisation des tests accroît la transparence de votre processus de test et votre capacité à couvrir davantage de fonctionnalités dans un laps de temps plus court. En testant et en validant les modifications en production, vous êtes en mesure d'identifier immédiatement les problèmes. L'amélioration de la cohérence avec les outils de test automatisés permet une meilleure détection des défauts. En restaurant automatiquement la version précédente, vous réduisez l'impact sur vos clients. La restauration automatisée inspire finalement plus de confiance dans vos capacités de déploiement en réduisant l'impact sur l'entreprise. Dans l'ensemble, ces capacités permettent de réduire les délais de livraison tout en garantissant la qualité.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Automatisez le test des environnements déployés pour confirmer les résultats souhaités plus rapidement. Automatisez la restauration du dernier état connu de bonne qualité lorsque les résultats prédéfinis ne sont pas atteints, afin de minimiser les temps de récupération et de réduire les erreurs causées par les processus manuels. Intégrez des outils de test au flux de travail de votre pipeline afin de tester de manière cohérente et de minimiser les saisies manuelles. Privilégiez l'automatisation des cas de test, tels que ceux qui atténuent les risques les plus importants et qui doivent être testés fréquemment à chaque modification. En outre, vous pouvez automatiser la restauration en fonction de conditions spécifiques prédéfinies dans votre plan de test.

Étapes d'implémentation

1. Établissez un cycle de test pour votre cycle de développement qui définit chaque étape du processus de test, de la planification des exigences au développement des cas de test, en passant par la configuration des outils, les tests automatisés et la clôture des cas de test.
   a. Créez une approche de test spécifique à la charge de travail à partir de votre stratégie de test globale.
   b. Envisagez, le cas échéant, une stratégie de tests continus tout au long du cycle de développement.
2. Choisissez des outils automatisés pour les tests et la restauration en fonction des besoins de votre entreprise et des investissements dans le pipeline.
3. Décidez des cas de test que vous souhaitez automatiser et de ceux qui doivent être exécutés manuellement. Ceux-ci peuvent être définis en fonction de la priorité de la valeur commerciale de la fonctionnalité testée. Alignez chaque membre de l'équipe sur ce plan et vérifiez leur responsabilité en ce qui concerne l'exécution des tests manuels.
   a. Appliquez les capacités de test automatisé à des cas de test spécifiques qui se prêtent à l'automatisation, tels que les cas répétables ou fréquemment exécutés, ceux qui nécessitent des tâches répétitives ou ceux qui sont requis dans plusieurs configurations.
   b. Définissez les scripts d'automatisation des tests ainsi que les critères de réussite dans l'outil d'automatisation afin que l'automatisation continue du flux de travail puisse être lancée lorsque des cas spécifiques échouent.
   c. Définissez des critères d'échec spécifiques pour la restauration automatisée.
4. Donnez la priorité à l'automatisation des tests afin d'obtenir des résultats cohérents grâce à un développement approfondi des cas de test où la complexité et l'interaction humaine présentent un risque d'échec plus élevé.
5. Intégrez vos outils de tests automatisés et de restauration dans votre pipeline CI/CD.
   a. Définissez des critères de réussite clairs pour vos modifications.
   b. Surveillez et observez pour détecter ces critères et annuler automatiquement les modifications lorsque des critères de restauration spécifiques sont remplis.
6. Procédez à différents types de tests de production automatisés, tels que :
   a. des tests A/B pour afficher les résultats par rapport à la version actuelle entre deux groupes d'utilisateurs ;
   b. des tests Canary qui vous permettent de déployer votre modification auprès d'un sous-ensemble d'utilisateurs avant de la diffuser à tous ;

   c. des tests d'indicateur de fonctions qui permettent d'activer et de désactiver une seule fonctionnalité de la nouvelle version depuis l'extérieur de l'application, de sorte que chaque nouvelle fonctionnalité puisse être validée une à la fois ;
   d. des tests de régression pour vérifier les nouvelles fonctionnalités avec les composants interdépendants existants.
7. Contrôlez les aspects opérationnels de l'application, les transactions et les interactions avec d'autres applications et composants. Élaborez des rapports pour illustrer le degré de réussite des modifications en fonction de la charge de travail, afin que vous puissiez identifier les parties de l'automatisation et du flux de travail qui peuvent être encore optimisées.
   a. Élaborez des rapports sur les résultats des tests qui vous aideront à prendre des décisions rapides sur le fait d'appeler ou non les procédures de restauration.
   b. Mettez en œuvre une stratégie permettant une restauration automatisée sur la base de conditions d'échec prédéfinies résultant d'une ou de plusieurs de vos méthodes de test.
8. Développez vos cas de test automatisés pour permettre leur réutilisation dans le cadre de futures modifications répétées.

Niveau d'effort du plan d'implémentation : Moyen

Ressources

Bonnes pratiques associées :

- OPS06-BP01 Planifier les modifications infructueuses
- OPS06-BP02 Déploiements de tests

Documents connexes :

- Builders Library AWS Builders Library | Exécuter des annulations sûres pendant les déploiements
- Redéployer et annuler un déploiement avec AWS CodeDeploy
- 8 bonnes pratiques pour automatiser vos déploiements avec AWS CloudFormation

Exemples connexes :

- Tests d'interface utilisateur sans serveur à l'aide de Selenium, AWS Lambda, AWS Fargate (Fargate) et AWS Developer Tools

Vidéos connexes :

- re:Invent 2020 | Hands-off: Automating continuous delivery pipelines at Amazon
- re:Invent 2019 | Amazon's Approach to high-availability deployment

OPS 7. Comment savoir si vous êtes prêt à assurer une charge de travail ?

Évaluez la disponibilité opérationnelle de votre charge de travail, des processus et des procédures, ainsi que le personnel pour comprendre les risques opérationnels liés à votre charge de travail.

Bonnes pratiques

- OPS07-BP01 Garantir les compétences du personnel
- OPS07-BP02 Assurer un examen cohérent de l'état de préparation opérationnelle
- OPS07-BP03 Utiliser des runbooks pour effectuer des procédures
- OPS07-BP04 Utiliser des playbooks pour analyser les problèmes
- OPS07-BP05 Prendre des décisions avisées pour déployer des systèmes et des modifications
- OPS07-BP06 Activer les formules de support pour les charges de travail de production

OPS07-BP01 Garantir les compétences du personnel

Prévoyez un mécanisme pour valider que vous disposez du nombre approprié de personnes formées pour supporter la charge de travail. Elles doivent être formées à la plateforme et aux services qui constituent votre charge de travail. Donnez-leur les connaissances nécessaires pour exploiter la charge de travail. Vous devez former un nombre suffisant de membres du personnel pour assurer le fonctionnement normal de la charge de travail et résoudre les incidents qui surviennent. Prévoyez suffisamment de personnel pour pouvoir effectuer une rotation pendant les astreintes et les vacances afin d'éviter l'épuisement professionnel.

Résultat souhaité :

- Il y a suffisamment de membres du personnel formés pour supporter la charge de travail aux moments où celle-ci est disponible.
- Vous assurez la formation de votre personnel sur les logiciels et services qui constituent votre charge de travail.

Anti-modèles courants :

- Déploiement d'une charge de travail sans que les membres de l'équipe soient qualifiés pour gérer la plateforme et les services utilisés.
- Ne pas disposer d'un personnel suffisant pour assurer les rotations d'astreinte ou les congés du personnel.

Avantages liés au respect de cette bonne pratique :

- Le fait de disposer de membres d'équipe compétents vous permet de prendre efficacement en charge votre charge de travail.
- Avec un nombre suffisant de membres de l'équipe, vous pouvez supporter la charge de travail et les rotations d'astreinte tout en diminuant le risque d'épuisement professionnel.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Confirmez qu'il y a suffisamment de personnel formé pour soutenir la charge de travail. Vérifiez que vous avez suffisamment de membres de l'équipe pour couvrir les activités opérationnelles normales, y compris les rotations d'astreinte.

Exemple de client

AnyCompany Retail veille à ce que les équipes qui supportent la charge de travail soient correctement dotées en personnel et formées. Elles disposent de suffisamment d'ingénieurs pour assurer une rotation d'astreinte. Le personnel reçoit une formation sur le logiciel et la plateforme sur lesquels repose la charge de travail et est encouragé à obtenir des certifications. Il y a suffisamment de membres du personnel pour que les gens puissent prendre des congés tout en supportant la charge de travail et la rotation des astreintes.

Étapes d'implémentation

1. Affectez un nombre suffisant d'employés à l'exploitation et au soutien de votre charge de travail, y compris aux fonctions d'astreinte.
2. Formez votre personnel aux logiciels et aux plateformes qui composent votre charge de travail.
   a. AWS Training and Certification dispose d'une bibliothèque de cours sur AWS. Le service propose des cours gratuits et payants, en ligne et en personne.
   b. AWS organise des événements et des webinaires au cours desquels vous apprendrez auprès d'experts AWS.

3. Évaluez régulièrement la taille et les compétences de l'équipe en fonction de l'évolution des conditions d'exploitation et de la charge de travail. Adaptez la taille et les compétences de l'équipe aux besoins opérationnels.

Niveau d'effort du plan d'implémentation : élevé. L'embauche et la formation d'une équipe pour soutenir une charge de travail peuvent demander des efforts considérables, mais présentent des avantages importants à long terme.

Ressources

Bonnes pratiques associées :

- OPS11-BP04 Gérer les connaissances - Les membres de l'équipe doivent disposer des informations nécessaires au fonctionnement et au soutien de la charge de travail. La gestion des connaissances est la clé pour y parvenir.

Documents connexes :

- Événements et webinaires AWS
- Formation et certification AWS

OPS07-BP02 Assurer un examen cohérent de l'état de préparation opérationnelle

Utilisez les examens de disponibilité opérationnelle (ORR) afin de vous assurer que vous pouvez gérer votre charge de travail. L'ORR est un mécanisme élaboré par Amazon afin de s'assurer que les équipes peuvent exécuter leurs charges de travail en toute sécurité. Un ORR est un processus d'examen et d'inspection qui utilise une liste de contrôle des exigences. Un ORR est une expérience en libre-service que les équipes utilisent pour certifier leurs charges de travail. Les ORR comprennent les bonnes pratiques tirées des enseignements liés aux années que nous avons consacrées à la création de logiciels.

La liste de contrôle d'un ORR est composée de recommandations architecturales, de processus opérationnels, de gestion d'événements et de qualité de version. Notre processus de correction des erreurs (CoE) est l'un des principaux moteurs de ces éléments. Votre propre analyse post-incident doit orienter l'évolution de votre propre ORR. Un ORR consiste non seulement à suivre les bonnes pratiques, mais permet également d'éviter la répétition d'événements que vous avez déjà vus. Enfin, les exigences en matière de sécurité, de gouvernance et de conformité peuvent également être incluses dans un ORR.

Exécutez les ORR avant qu'une charge de travail ne soit généralement disponible, puis tout au long du cycle de développement du logiciel. L'exécution d'un ORR avant le lancement augmente votre capacité de gestion de la charge de travail en toute sécurité. Réexécutez régulièrement votre ORR sur la charge de travail afin de détecter toute dérive par rapport aux bonnes pratiques. Vous pouvez avoir des listes de contrôle des ORR pour les lancements de nouveaux services et des ORR pour les examens périodiques. Cela vous permet de vous tenir au courant des nouvelles bonnes pratiques et d'intégrer les leçons tirées de l'analyse après incident. Au fur et à mesure que votre utilisation du cloud évolue, vous pouvez intégrer les exigences des ORR dans votre architecture par défaut.

Résultat souhaité : vous avez une liste de contrôle de l'ORR avec les bonnes pratiques pour votre organisation. Les ORR sont effectuées avant le lancement des charges de travail. Les ORR sont exécutés périodiquement tout au long du cycle de vie de la charge de travail.

Anti-modèles courants :

- Vous lancez une charge de travail sans savoir si vous pouvez l'utiliser.
- Les exigences en matière de gouvernance et de sécurité ne sont pas incluses dans la certification d'une charge de travail pour le lancement.
- Les charges de travail ne sont pas réévaluées périodiquement.
- Les charges de travail sont lancées sans procédures requises en place.
- Vous voyez la répétition de la même cause racine de défaillances dans plusieurs charges de travail.

Avantages liés au respect de cette bonne pratique :

- Vos charges de travail comprennent les bonnes pratiques en matière d'architecture, de processus et de gestion.
- Les enseignements tirés sont intégrés à votre processus d'ORR.
- Les procédures requises sont en place lors du lancement des charges de travail.
- Les ORR sont exécutés tout au long du cycle de vie logiciel de vos charges de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Un ORR est composé de deux éléments : un processus et une liste de contrôle. Votre processus d'ORR doit être adopté par votre organisation et soutenu par un responsable exécutif. Au minimum,

les ORR doivent être effectués avant qu'une charge de travail ne soit généralement disponible. Exécutez l'ORR tout au long du cycle de développement du logiciel afin de l'actualiser avec les bonnes pratiques ou les nouvelles exigences. La liste de contrôle d'un ORR doit comprendre les éléments de configuration, les exigences en matière de sécurité et de gouvernance et les bonnes pratiques de votre organisation. Au fil du temps, vous pouvez utiliser des services tels qu' AWS Config, AWS Security Hubet les barrières de protection AWS Control Towerafin d'intégrer les bonnes pratiques de l'ORR aux barrières de protection pour la détection automatique des bonnes pratiques.

Exemple client

Après plusieurs incidents de production, AnyCompany Retail a décidé de mettre en place un processus d'ORR. L'entreprise a élaboré une liste de contrôle composée de bonnes pratiques, d'exigences en matière de gouvernance et de conformité et d'enseignements tirés des pannes. De nouvelles charges de travail effectuent des ORR avant leur lancement. Chaque charge de travail effectue un ORR annuel avec un sous-ensemble de bonnes pratiques pour intégrer de nouvelles bonnes pratiques et des exigences qui sont ajoutées à la liste de contrôle de l'ORR. Au fil du temps, AnyCompany Retail a utilisé AWS Config afin de détecter certaines bonnes pratiques, en accélérant le processus d'ORR.

Étapes d'implémentation

Pour en savoir plus sur les ORR, lisez le livre blanc Operational Readiness Reviews (ORR). Il fournit des informations détaillées sur l'historique du processus d'ORR, sur la façon d'établir votre propre pratique d'ORR et sur la façon d'élaborer votre liste de contrôle pour les ORR. Les étapes suivantes sont une version abrégée de ce document. Pour une compréhension approfondie des ORR et de la façon dont vous pouvez créer les vôtres, nous vous recommandons de lire ce livre blanc.

1. Réunissez les parties prenantes clés, notamment les représentants de la sécurité, des opérations et du développement.
2. Demandez à chaque partie prenante de fournir au moins une exigence. Pour la première itération, essayez de limiter le nombre d'éléments à trente ou moins.
    - L'Annexe A, Example ORR questions, du livre blanc Operational Readiness Reviews (ORR) contient des exemples de questions que vous pouvez utiliser pour démarrer.
3. Regroupez vos exigences dans une feuille de calcul.
    - Vous pouvez utiliser des approches personnalisées dans l' AWS Well-Architected Tool afin de développer votre ORR et de le partager avec vos comptes et votre AWS Organization.
4. Identifiez une charge de travail pour effectuer l'ORR. Il est recommandé d'utiliser une charge de travail avant le lancement ou une charge de travail interne.

5. Parcourez la liste de contrôle de l'ORR et notez toutes vos découvertes. Les découvertes peuvent ne pas être acceptables si une mesure d'atténuation est en place. Pour toute découverte qui ne comporte pas de mesures d'atténuation, ajoutez ces dernières à votre liste de tâches en attente et implémentez-les avant le lancement.
6. Continuez d'ajouter des bonnes pratiques et des exigences à votre liste de contrôle de l'ORR au fil du temps.

Les clients AWS Support disposant d'un Enterprise Support peuvent demander l'atelier Operational Readiness Review Workshop à leur gestionnaire de compte technique. Cet atelier est une session de travail à rebours permettant de développer votre propre liste de contrôle pour un ORR.

Niveau d'effort du plan d'implémentation : élevé L'adoption d'une pratique d'ORR dans votre organisation nécessite un parrainage de la haute direction et l'adhésion des parties prenantes. Créez et mettez à jour la liste de contrôle à l'aide des commentaires de l'ensemble de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS01-BP03 Évaluer les exigences de gouvernance – Les exigences en matière de gouvernance conviennent naturellement à la liste de contrôle d'un ORR.
- OPS01-BP04 Évaluer les exigences de conformité – Les exigences de conformité sont parfois incluses dans la liste de contrôle d'un ORR. Parfois, il s'agit d'un processus distinct.
- OPS03-BP07 Fournir aux équipes les ressources appropriées – La capacité de l'équipe peut faire partie des exigences d'un ORR.
- OPS06-BP01 Planifier les modifications infructueuses – Un plan de restauration ou de retour en arrière doit être établi avant le lancement de votre charge de travail.
- OPS07-BP01 Garantir les compétences du personnel – Pour gérer une charge de travail, vous devez disposer du personnel requis.
- SEC01-BP03 Identifier et valider les objectifs de contrôle – Les objectifs de contrôle de sécurité constituent d'excellentes exigences d'ORR.
- REL13-BP01 Définir les objectifs de reprise pour les temps d'arrêt et les pertes de données – Les plans de reprise après sinistre constituent une exigence appropriée dans le cadre d'un ORR.
- COST02-BP01 Développer des stratégies en fonction des exigences de votre organisation – Il est recommandé d'inclure les politiques de gestion des coûts dans la liste de contrôle d'un ORR.

Documents connexes :

- AWS Control Tower - Guardrails in AWS Control Tower
- AWS Well-Architected Tool - Custom Lenses
- Operational Readiness Review Template par Adrian Hornsby
- Livre blanc Operational Readiness Reviews (ORR)

Vidéos connexes :

- AWS Supports You | Building an Effective Operational Readiness Review (ORR)

Exemples connexes :

- Sample Operational Readiness Review (ORR) Lens

Services associés :

- AWS Config
- AWS Control Tower
- AWS Security Hub
- AWS Well-Architected Tool

OPS07-BP03 Utiliser des runbooks pour effectuer des procédures

A runbook est un processus documenté pour atteindre un résultat spécifique. Les runbooks consistent en une série d'étapes permettant à la personne qui les suit d'obtenir des résultats concrets. L'utilisation des runbooks dans les opérations remonte aux débuts de l'aviation. Dans les opérations de cloud, nous utilisons des runbooks pour réduire les risques et obtenir les résultats souhaités. Dans sa forme la plus simple, un runbook est une liste de contrôle pour exécuter une tâche.

Les runbooks représentent une part essentielle du fonctionnement de votre charge de travail. De l'intégration d'un nouveau membre de l'équipe au déploiement d'une version majeure, les runbooks sont des processus codifiés qui fournissent des résultats cohérents quelle que soit la personne qui les utilise. Les runbooks doivent être publiés dans un emplacement central et mis à jour à mesure que le processus évolue, car la mise à jour des runbooks est un composant essentiel du processus

de gestion des changements. Ils doivent également inclure des conseils sur la gestion des erreurs, les outils, les autorisations, les exceptions et les remontées en cas de problème.

À mesure que votre entreprise évolue, commencez à automatiser les runbooks. Prenez tout d'abord les runbooks courts et fréquemment utilisés. Utilisez des langages de scripts pour automatiser les étapes ou les rendre plus faciles. À mesure que vous automatiserez les premiers runbooks, vous consacrerez du temps à l'automatisation de runbooks plus complexes. Au fil du temps, la plupart de vos runbooks seront automatisés d'une certaine façon.

Résultat souhaité : Votre équipe dispose de plusieurs guides détaillés pour exécuter des tâches de charge de travail. Les runbooks contiennent le résultat souhaité, les outils et autorisations nécessaires, ainsi que les instructions pour gérer les erreurs. Ils sont stockés dans un emplacement central et mis à jour fréquemment.

Anti-modèles courants :

- Utilisation de la mémoire pour exécuter chaque étape d'un processus.
- Déploiement manuel des changements sans liste de contrôle.
- Différents membres de l'équipe exécutant le même processus, mais avec des étapes ou résultats différents.
- Désynchronisation des runbooks avec les changements du système et l'automatisation.

Avantages liés au respect de cette bonne pratique :

- Réduction du taux d'erreur pour les tâches manuelles.
- Exécution cohérente des opérations.
- Exécution des tâches plus tôt par les nouveaux membres de l'équipe.
- Automatisation des runbooks pour diminuer la quantité de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Les runbooks peuvent prendre plusieurs formes selon le niveau de maturité de votre entreprise. Au minimum, ils doivent consister en un document texte détaillé. Le résultat souhaité doit être clairement indiqué. Ils documentent explicitement les autorisations spéciales ou outils nécessaires. Ils fournissent des conseils sur la gestion des erreurs et les remontées en cas de problème. Recherchez le propriétaire du runbook et publiez-le dans un emplacement central. Une fois votre runbook

documenté, validez-le en demandant à un membre de votre équipe de l'exécuter. À mesure que les procédures évoluent, mettez à jour vos runbooks conformément à votre processus de gestion des changements.

Vos runbooks texte doivent être automatisés à mesure que votre entreprise évolue. Grâce à des services tels que les automatisations AWS Systems Manager, vous pouvez transformer un fichier texte en automatisations pouvant être exécutées sur votre charge de travail. Ces automatisations peuvent être exécutées en réponse aux événements, tout en réduisant la charge opérationnelle pour maintenir votre charge de travail.

Exemple client

AnyCompany Retail doit mettre à jour des schémas de bases de données lors de déploiements logiciels. L'équipe en charge des opérations de cloud en collaboration avec l'équipe responsable de l'administration des bases de données ont créé un runbook, pour déployer manuellement ces changements. Le runbook répertoriait chacune des étapes du processus sous forme de liste de contrôle. Il comprenait une section sur la gestion des erreurs en cas de problème. Les équipes ont publié le runbook sur leur wiki interne contenant leurs autres runbooks. L'équipe en charge des opérations de cloud envisage d'automatiser le runbook dans un prochain sprint.

Étapes d'implémentation

Si vous ne disposez pas d'un référentiel de documents, un référentiel de contrôle de version est un emplacement idéal pour commencer à créer votre bibliothèque de runbooks. Vous pouvez créer vos runbooks en utilisant le format Markdown. Voici un exemple de modèle de runbook que vous pouvez utiliser pour commencer à créer vos runbooks.

```
# Runbook Title ## Runbook Info | Runbook ID | Description | Tools Used
 | Special Permissions | Runbook Author | Last Updated | Escalation POC |
 |-------|-------|-------|-------|-------|-------|-------| | RUN001 | What is this
 runbook for? What is the desired outcome? | Tools | Permissions | Your Name |
 2022-09-21 | Escalation Name | ## Steps 1. Step one 2. Step two
```

1. Si vous ne possédez pas de référentiel de documentation ou de wiki existant, créer un référentiel de contrôle de version dans votre système de contrôle de version.
2. Identifier un processus ne possédant pas de runbook. Le processus idéal doit être réalisé de manière semi-régulière, contenir peu d'étapes et avoir des échecs à faible impact.
3. Dans votre référentiel de documents, créer un brouillon au format Markdown en utilisant le modèle. Remplissez le champ `Runbook Title` et les champs obligatoires sous `Runbook Info`.

4. En commençant par la première, remplir la partie `Étapes` du runbook.
5. Donner le runbook à un membre de l'équipe. Demandez-lui d'utiliser le runbook pour valider les étapes. En cas d'élément manquant ou de besoin de clarification, mettez à jour le runbook.
6. Publier le runbook sur votre référentiel de documentation interne. Une fois publié, partagez l'information avec votre équipe et les autres parties prenantes.
7. Au fil du temps, vous créerez une bibliothèque de runbooks. À mesure que cette bibliothèque s'étoffe, commencez à travailler sur l'automatisation des runbooks.

Niveau d'effort du plan d'implémentation : faible. La norme minimum pour un runbook est un guide texte détaillé. L'automatisation des runbooks peut augmenter l'effort d'implémentation.

Ressources

Bonnes pratiques associées :

- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés : les runbooks doivent avoir un propriétaire chargé d'en assurer la maintenance.
- OPS07-BP04 Utiliser des playbooks pour analyser les problèmes : les runbooks et les playbooks sont identiques à une différence près : un runbook a un résultat souhaité. Dans de nombreux cas, les runbooks sont déclenchés suite à l'identification d'une cause profonde par un playbook.
- OPS10-BP01 Utiliser un processus pour la gestion des événements, des incidents et des problèmes : les runbooks sont une part essentielle de la pratique de la gestion d'un bon déroulement, d'un incident et d'un problème
- OPS10-BP02 Disposer d'un processus par alerte : les runbooks et les playbooks doivent être utilisés pour répondre aux alertes. Avec le temps, ces réactions doivent être automatisées.
- OPS11-BP04 Gérer les connaissances : la maintenance des runbooks représente une part essentielle de la gestion des connaissances.

Documents connexes :

- Atteindre l'excellence opérationnelle grâce à l'automatisation de playbooks et de runbooks
- AWS Systems Manager : travailler avec des runbooks
- Playbook d'atténuation des risques pour les importantes migrations AWS – Tâche 4 : amélioration de vos runbooks de migration
- Utiliser les runbooks AWS Systems Manager Automation pour résoudre des tâches opérationnelles

Vidéos connexes :

- AWS re:Invent 2019: DIY guide to runbooks, incident reports, and incident response (SEC318-R1)
- Comment automatiser des opérations informatiques sur AWS | Amazon Web Services
- Intégrations de scripts dans AWS Systems Manager

Exemples connexes :

- AWS Systems Manager : procédure étape par étape pour l'automatisation
- AWS Systems Manager : restaurer un volume racine à partir du dernier runbook d'instantanés
- Créer un runbook de réponse d'incident AWS à l'aide des blocs-notes Jupyter et CloudTrail Lake
- Gitlab – Runbooks
- Rubix – Une bibliothèque Python pour créer des runbooks dans les blocs-notes Jupyter
- Utilisation d'un créateur de documents pour créer un runbook personnalisé
- Ateliers Well-Architected : automatisation des opérations avec les playbooks et les runbooks

Services associés :

- AWS Systems Manager Automation

OPS07-BP04 Utiliser des playbooks pour analyser les problèmes

Les playbooks sont des guides étape par étape utilisés pour analyser un incident. Lorsque des incidents se produisent, les playbooks sont utilisés pour analyser, évaluer l'impact et identifier une cause racine. Les playbooks sont utilisés dans le cadre de différents scénarios allant des échecs de déploiement aux incidents de sécurité. Dans la plupart des cas, les playbooks identifient la cause racine qui est atténuée par l'utilisation d'un runbook. Les playbooks sont une composante essentielle des plans de réponse de votre organisation en cas d'incident.

Un playbook efficace comporte plusieurs fonctionnalités clés. Il guide l'utilisateur, étape par étape, dans le processus de découverte. Si vous optez pour un point de vue extérieur, quelles étapes devez-vous suivre pour diagnostiquer un incident ? Définissez clairement dans le playbook si des outils spéciaux ou des autorisations élevées sont nécessaires. Il est essentiel d'élaborer un plan de communication pour informer les parties prenantes du statut de l'analyse. Lorsqu'il est impossible de déterminer la cause racine, le playbook doit comporter un plan de remontée des informations vers la

hiérarchie. Si la cause racine est identifiée, le playbook doit faire référence à un runbook décrivant une solution pour la résoudre. Les playbooks doivent être stockés dans un emplacement central et mis à jour régulièrement. Si des playbooks sont utilisés pour des alertes précises, donnez aux membres de votre équipe des indications relatives au playbook dans le cadre de l'alerte.

Au fur et à mesure que votre organisation évolue, automatisez vos playbooks. Commencez par des playbooks qui couvrent les incidents à faible risque. Utilisez des scripts pour automatiser les étapes de découverte. Veillez à créer des runbooks complémentaires destinés à atténuer les causes racine courantes.

Résultat souhaité : votre organisation dispose de playbooks pour les incidents courants. Les playbooks sont stockés dans un emplacement central et mis à la disposition des membres de votre équipe. Les playbooks sont souvent mis à jour. Pour toute cause racine connue, des runbooks complémentaires sont créés.

Anti-modèles courants :

- Il n'existe pas de façon standard d'analyser un incident.
- Les membres de l'équipe comptent sur la mémoire musculaire ou les connaissances institutionnelles pour résoudre un échec de déploiement.
- Les nouveaux membres de l'équipe apprennent à analyser les problèmes par un procédé de tâtonnement.
- Les bonnes pratiques d'analyse des problèmes ne sont pas partagées entre les équipes.

Avantages liés au respect de cette bonne pratique :

- Les playbooks dynamisent les efforts nécessaires pour atténuer les incidents.
- Différents membres de l'équipe peuvent utiliser le même playbook pour identifier une cause racine de façon cohérente.
- Les causes racine connues peuvent être associées à des runbooks développés spécialement pour leur résolution, ce qui permet d'accélérer le délai de récupération.
- Les playbooks permettent aux membres de l'équipe de commencer à apporter leur contribution plus tôt.
- Les équipes peuvent adapter leurs processus à l'aide de playbooks reproductibles.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

La façon dont vous créez et utilisez les playbooks dépend de la maturité de votre organisation. Si vous débutez dans le cloud, créez des playbooks sous forme de texte dans un référentiel de documents centralisé. Au fur et à mesure que votre organisation évolue, les playbooks peuvent devenir semi-automatisés avec des langages de script comme Python. Ces scripts peuvent être exécutés dans un bloc-notes Jupyter afin d'accélérer la découverte. Les organisations avancées ont des playbooks entièrement automatisés pour les problèmes courants qui sont corrigés automatiquement avec des runbooks.

Pour commencer à créer vos playbooks, répertoriez les incidents qui affectent couramment votre charge de travail. Pour commencer, choisissez des playbooks pour les incidents à faible risque dont la cause racine a été réduite à quelques problèmes. Une fois que vous disposez de playbooks pour des scénarios plus simples, passez aux scénarios à risque élevé ou à ceux dont la cause racine est peu connue.

Vos playbooks sous forme de texte doivent être automatisés à mesure que votre entreprise évolue. Grâce à des services tels que AWS Systems Manager Automation, le texte brut peut être transformé en automatisations. Ces automatisations peuvent être exécutées en fonction de votre charge de travail pour accélérer les analyses. Ces automatisations peuvent être activées en réponse à des événements, ce qui réduit le temps nécessaire pour découvrir et résoudre les incidents.

Les clients peuvent utiliser AWS Systems Manager Incident Manager afin de répondre aux incidents. Ce service offre une interface unique pour trier les incidents, informer les parties prenantes pendant la découverte et l'atténuation, et collaborer tout au long de l'incident. Il utilise AWS Systems Manager Automation afin d'accélérer la détection et la récupération.

Exemple client

AnyCompany Retail a dû faire face à un incident de production. L'ingénieur d'astreinte a utilisé un playbook pour analyser le problème. À mesure qu'il effectuait les différentes étapes, il a informé les parties prenantes identifiées dans le playbook de l'évolution de la situation. L'ingénieur a identifié que la cause racine était une condition de concurrence dans un service back-end. À l'aide d'un runbook, il a relancé le service et a permis à AnyCompany Retail d'être à nouveau en ligne.

Étapes d'implémentation

Si vous n'avez pas de référentiel de documents existant, nous vous suggérons de créer un référentiel de contrôle de version pour votre bibliothèque de playbooks. Vous pouvez créer vos playbooks en

utilisant Markdown, qui est compatible avec la plupart des systèmes d'automatisation de playbook. Si vous démarrez de zéro, utilisez l'exemple de modèle de playbook suivant.

```
# Titre du playbook ## Informations sur le playbook | ID du playbook | Description
 | Outils utilisés | Autorisations spéciales | Auteur du playbook | Dernière mise à
 jour | POC de remontée hiérarchique | Parties prenantes | Plan de communication |
 |-------|-------|-------|-------|-------|-------|-------|-------|-------| | RUN001
 | À quoi sert ce playbook ? Pour quel type d'incident est-il utilisé ? | Outils |
 Autorisations | Votre nom | 2022-09-21 | Contact pour la remontée des informations
 vers la hiérarchie | Nom de la partie prenante | Comment les dernières informations
 seront-elles communiquées au cours de l'analyse ? | ## Étapes 1. Première étape 2.
 Deuxième étape
```

1. Si vous ne possédez pas de référentiel de documents ni de wiki existant, créez un référentiel de contrôle de version pour vos playbooks dans votre système de contrôle de version.
2. Identifiez un problème courant qui doit être analysé. Il doit s'agir d'un scénario où la cause racine se limite à quelques problèmes et où la résolution présente peu de risques.
3. À l'aide du modèle Markdown, remplissez la section `Nom du playbook` et les champs situés sous `Informations sur le playbook`.
4. Remplissez les étapes de résolution du problème. Soyez aussi clair que possible sur les actions à effectuer ou les domaines à analyser.
5. Remettez le playbook à un membre de l'équipe et demandez-lui de le passer en revue afin de le valider. S'il manque quelque chose ou si un point n'est pas clair, mettez à jour le playbook.
6. Publiez le playbook dans votre référentiel de documents et informez votre équipe et les parties prenantes.
7. Cette bibliothèque de playbooks s'enrichira à mesure que vous ajouterez d'autres playbooks. Une fois que vous avez plusieurs playbooks, commencez à les automatiser en utilisant des outils comme AWS Systems Manager Automation afin de garantir la synchronisation entre l'automatisation et les playbooks.

Niveau d'effort du plan d'implémentation : faible. Vos playbooks doivent être des documents texte stockés dans un emplacement central. Les organisations plus avancées évolueront vers l'automatisation des playbooks.

Ressources

Bonnes pratiques associées :

- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés : un propriétaire doit être désigné pour les playbooks et être chargé d'en assurer la gestion.
- OPS07-BP03 Utiliser des runbooks pour effectuer des procédures : les runbooks et les playbooks sont similaires, mais se distinguent par le fait qu'un résultat souhaité est défini pour un runbook. Dans de nombreux cas, les runbooks sont utilisés après qu'un playbook a identifié une cause racine.
- OPS10-BP01 Utiliser un processus pour la gestion des événements, des incidents et des problèmes : les runbooks constituent un élément important d'une bonne pratique de gestion des événements, des incidents et des problèmes.
- OPS10-BP02 Disposer d'un processus par alerte : les runbooks et les playbooks doivent être utilisés pour répondre aux alertes. Avec le temps, ces réactions doivent être automatisées.
- OPS11-BP04 Gérer les connaissances : la gestion des playbooks est un élément clé de la gestion des connaissances.

Documents connexes :

- Achieving Operational Excellence using automated playbook and runbook
- AWS Systems Manager – Utilisation de runbooks
- Use AWS Systems Manager Automation runbooks to resolve operational tasks

Vidéos connexes :

- AWS re:Invent 2019: DIY guide to runbooks, incident reports, and incident response (SEC318-R1)
- AWS Systems Manager Incident Manager - AWS Virtual Workshops
- Integrate Scripts into AWS Systems Manager

Exemples connexes :

- AWS Customer Playbook Framework
- AWS Systems Manager : Procédures d'automatisation
- Building an AWS incident response runbook using Jupyter notebooks and CloudTrail Lake
- Rubix – Une bibliothèque Python pour créer des runbooks dans les bloc-notes Jupyter
- Utilisation de Document Builder pour créer un runbook personnalisé

- Ateliers Well-Architected : automatisation des opérations avec les playbooks et les runbooks
- Ateliers Well-Architected : playbook de réponse aux incidents avec Jupyter

Services associés :

- AWS Systems Manager Automation
- AWS Systems Manager Incident Manager

OPS07-BP05 Prendre des décisions avisées pour déployer des systèmes et des modifications

Mettez en place des processus pour les modifications réussies et ratées de votre charge de travail. Un pré-mortem est un exercice où une équipe simule un échec pour développer des stratégies d'atténuation. Utilisez des pré-mortems pour anticiper les échecs et créer des procédures le cas échéant. Évaluez les avantages et les risques liés au déploiement de modifications dans votre charge de travail. Vérifiez que toutes les modifications sont conformes à la gouvernance.

Résultat souhaité :

- Vous prenez des décisions éclairées lorsque vous déployez des modifications dans votre charge de travail.
- Les modifications sont conformes à la gouvernance.

Anti-modèles courants :

- Déployer une modification dans notre charge de travail sans disposer de processus pour gérer un déploiement raté.
- Apporter des modifications à votre environnement de production qui ne sont pas conformes aux exigences de gouvernance.
- Déployer une nouvelle version de votre charge de travail sans établir une base de référence pour l'utilisation des ressources.

Avantages liés au respect de cette bonne pratique :

- Vous êtes préparé à des modifications ratées de votre charge de travail.
- Les modifications apportées à votre charge de travail sont conformes aux politiques de gouvernance.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Utilisez des pré-mortems pour développer des processus pour les modifications ratées. Documentez vos processus pour les modifications ratées. Veillez à ce que toutes les modifications soient conformes à la gouvernance. Évaluez les avantages et les risques liés au déploiement de modifications dans votre charge de travail.

Exemple de client

AnyCompany Retail effectue régulièrement des pré-mortems pour valider ses processus en cas de modification ratée. La société documente ses processus dans un Wiki partagé et le met à jour fréquemment. Toutes les modifications sont conformes aux exigences de gouvernance.

Étapes d'implémentation

1. Prenez des décisions éclairées lorsque vous déployez des modifications dans votre charge de travail. Définissez et révisez les critères d'un déploiement réussi. Développez des scénarios ou des critères qui déclencheraient la restauration d'une modification. Comparez les avantages du déploiement des modifications avec les risques associés à l'échec d'une modification.
2. Vérifiez que toutes les modifications sont conformes aux politiques de gouvernance.
3. Utilisez les pré-mortems pour planifier les modifications ratées et documenter les stratégies d'atténuation. Réalisez un exercice théorique pour modéliser une modification qui n'a pas abouti et valider les procédures de restauration.

Niveau d'effort du plan d'implémentation : modéré. La mise en œuvre d'une pratique de pré-mortems nécessite une coordination et des efforts de la part des parties prenantes de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS01-BP03 Évaluer les exigences de gouvernance - Les exigences de gouvernance sont un facteur clé pour déterminer s'il faut déployer une modification.
- OPS06-BP01 Planifier les modifications infructueuses - Établissez des plans pour atténuer les effets d'un déploiement raté et utilisez des pré-mortems pour les valider.
- OPS06-BP02 Déploiements de tests - Chaque modification apportée à un logiciel doit être correctement testée avant le déploiement afin de réduire les défauts en production.

- OPS07-BP01 Garantir les compétences du personnel - Il est essentiel de disposer de suffisamment de membres du personnel formés pour supporter la charge de travail afin de prendre une décision éclairée quant au déploiement d'une modification du système.

Documents connexes :

- Amazon Web Services : risques et conformité
- Modèle de responsabilité partagée d'AWS
- Governance in the AWS Cloud: The Right Balance Between Agility and Safety (La gouvernance dans le cloud AWS : le juste équilibre entre agilité et sécurité)

OPS07-BP06 Activer les formules de support pour les charges de travail de production

Activez la prise en charge de tous les logiciels et services sur lesquels repose votre charge de travail de production. Sélectionnez un niveau de support approprié pour répondre à vos besoins en matière de niveau de service de production. Il convient de prévoir des formules de support pour ces dépendances en cas d'interruption de service ou de problème logiciel. Documentez les formules de support et les procédures de demande de support pour tous les fournisseurs de services et de logiciels. Mettez en œuvre des mécanismes permettant de vérifier que les points de contact du support sont tenus à jour.

Résultat souhaité :

- Mettre en œuvre des formules de support pour les logiciels et les services sur lesquels reposent les charges de travail de production.
- Choisir une formule de support appropriée en fonction des besoins du niveau de service.
- Documenter les formules de support, les niveaux de support et les procédures de demande de support.

Anti-modèles courants :

- Vous n'avez pas de formule de support pour un fournisseur de logiciels critiques. Votre charge de travail en est affectée et vous ne pouvez rien faire pour accélérer la mise en place d'une solution ou obtenir des mises à jour en temps voulu de la part du fournisseur.
- Un développeur qui était le principal point de contact pour un fournisseur de logiciels a quitté l'entreprise. Vous n'arrivez pas à joindre directement le support du fournisseur. Vous devez passer

du temps à rechercher et à naviguer dans des systèmes de contact génériques, ce qui augmente le temps nécessaire pour répondre en cas de besoin.

- Un fournisseur de logiciels connaît un arrêt de production. Il n'existe pas de documentation sur la manière de déposer un dossier de support.

Avantages liés au respect de cette bonne pratique :

- En adoptant le niveau de support approprié, vous êtes en mesure d'obtenir une réponse dans le délai nécessaire pour répondre aux besoins du niveau de service.
- En tant que client bénéficiant du support, vous pouvez faire remonter les problèmes de production.
- Les fournisseurs de logiciels et de services peuvent contribuer au dépannage pendant un incident.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Activez les formules de support pour tous les fournisseurs de logiciels et de services sur lesquels repose votre charge de travail de production. Mettez en place des formules de support appropriées pour répondre aux besoins du niveau de service. Pour les clients AWS, cela signifie qu'il faut activer l'offre AWS Business Support ou supérieure sur tous les comptes où vous avez des charges de travail de production. Rencontrez régulièrement les fournisseurs de services de support afin d'obtenir des informations actualisées sur les offres de support, les processus et les contacts. Documentez les procédures de demande de support auprès des fournisseurs de logiciels et de services, y compris la manière de faire remonter les informations en cas de panne. Mettez en œuvre des mécanismes permettant de tenir à jour les contacts du support.

Exemple de client

Chez AnyCompany Retail, toutes les dépendances des logiciels et services commerciaux disposent de formules de support. Par exemple, l'offre AWS Enterprise Support est activée sur tous les comptes comportant des charges de travail de production. Tout développeur peut soulever un incident auprès du support en cas de problème. Il existe une page wiki contenant des informations sur la manière de demander de l'aide, sur les personnes à prévenir et sur les bonnes pratiques pour accélérer le traitement d'un incident.

Étapes d'implémentation

1. Travaillez avec les parties prenantes de votre organisation pour identifier les fournisseurs de logiciels et de services sur lesquels repose votre charge de travail. Documentez ces dépendances.
2. Déterminez les besoins en matière de niveau de service pour votre charge de travail. Sélectionnez un plan de support qui leur corresponde.
3. Pour les logiciels et services commerciaux, mettez en place une formule de support avec les fournisseurs.
    a. Nous vous conseillons vivement de souscrire à AWS Business Support ou à un niveau supérieur pour tous les comptes de production, ce qui vous permettra de bénéficier de temps de réponse plus courts de la part d'AWS Support. Si vous ne disposez pas d'une offre de support premium, mettez en place un plan d'action pour gérer les problèmes qui nécessitent l'aide de AWS Support. AWS Support fournit une combinaison d'outils et de technologies, de personnes et de programmes conçus pour vous aider à optimiser vos performances, à réduire vos coûts et à innover plus rapidement. AWS Business Support offre des avantages supplémentaires, notamment l'accès à AWS Trusted Advisor et à AWS Personal Health Dashboard, ainsi que des délais de réponse plus rapides.
4. Documentez la formule de support dans votre outil de gestion des connaissances. Il s'agit notamment de savoir comment demander de l'aide, qui avertir en cas de demande de support et comment faire remonter l'information pendant un incident. Un wiki constitue un bon mécanisme pour permettre à quiconque d'apporter les mises à jour nécessaires à la documentation lorsqu'il prend connaissance de changements dans les processus ou les contacts de support.

Niveau d'effort du plan d'implémentation : faible. La plupart des fournisseurs de logiciels et de services proposent des formules de support à l'inscription. La documentation et le partage des bonnes pratiques en matière de support sur votre système de gestion des connaissances permettent de vérifier que votre équipe sait ce qu'il faut faire en cas d'incident de production.

Ressources

Bonnes pratiques associées :

- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés

Documents connexes :

- AWS Support Plans

Services associés :

- AWS Business Support
- AWS Enterprise Support

Exploiter

Questions

- OPS 8. Comment exploitez-vous l'observabilité de la charge de travail dans votre organisation ?
- OPS 9. Comment comprendre l'état de vos opérations ?
- OPS 10. Comment gérer les événements relatifs à la charge de travail et aux opérations ?

OPS 8. Comment exploitez-vous l'observabilité de la charge de travail dans votre organisation ?

Garantissez un état optimal de la charge de travail en tirant parti de l'observabilité. Utilisez des métriques, des journaux et des données de suivi pertinents pour obtenir une vue complète des performances de votre charge de travail et résoudre les problèmes de manière efficace.

Bonnes pratiques

- OPS08-BP01 Analyse des métriques de charge de travail
- OPS08-BP02 Analyse des journaux de charge de travail
- OPS08-BP03 Analyse des données de suivi de la charge de travail
- OPS08-BP04 Création d'alertes exploitables
- OPS08-BP05 Création de tableaux de bord

OPS08-BP01 Analyse des métriques de charge de travail

Après avoir implémenté la télémétrie des applications, analysez régulièrement les métriques collectées. Bien que la latence, les requêtes, les erreurs et la capacité (ou les quotas) fournissent des informations sur les performances du système, il est essentiel de donner la priorité à l'examen des métriques liées aux résultats commerciaux. Vous vous assurez ainsi de prendre des décisions basées sur des données conformes aux objectifs de votre entreprise.

Résultat souhaité : Informations précises sur les performances des charges de travail afin de prendre des décisions éclairées par les données, garantissant ainsi l'alignement avec les objectifs de votre entreprise.

Anti-modèles courants :

- Analyser les métriques de manière isolée sans tenir compte de leur impact sur les résultats commerciaux.
- Se fier de manière excessive aux métriques techniques tout en mettant de côté les métriques commerciales.
- Ne pas examiner les métriques assez souvent, ce qui vous fait passer à côté de possibilités de prise de décision en temps réel.

Avantages liés au respect de cette bonne pratique :

- Meilleure compréhension de la corrélation entre les performances techniques et les résultats commerciaux.
- Processus décisionnel amélioré grâce à des données en temps réel.
- Identification et atténuation proactives des problèmes avant qu'ils n'affectent les résultats commerciaux.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Tirez parti d'outils comme Amazon CloudWatch pour effectuer l'analyse des métriques. Des services AWS comme AWS Cost Anomaly Detection et Amazon DevOps Guru peuvent être utilisés pour détecter les anomalies, en particulier lorsque les seuils statiques sont inconnus ou lorsque les modèles de comportement sont davantage adaptés à la détection d'anomalies.

Étapes d'implémentation

1. Analysez et vérifiez : Passez régulièrement en revue les métriques de votre charge de travail et interprétez-les.
   a. Donnez la priorité aux métriques liées aux résultats commerciaux par rapport aux métriques purement techniques.
   b. Comprenez l'importance des pics, des baisses ou des tendances dans vos données.

2. Utilisez Amazon CloudWatch : Utilisez Amazon CloudWatch pour obtenir une vue centralisée et une analyse approfondie.
   a. Configurez des tableaux de bord CloudWatch pour visualiser vos métriques et les comparer au fil du temps.
   b. Utilisez les centiles dans CloudWatch pour avoir une vision claire de la distribution des métriques, ce qui peut aider à définir les SLA et à interpréter les valeurs aberrantes.
   c. Configurez AWS Cost Anomaly Detection pour identifier des tendances inhabituelles sans se fier aux seuils statiques.
   d. Implémentez l'observabilité CloudWatch entre comptes pour surveiller et dépanner les applications qui couvrent plusieurs comptes au sein d'une même région.
   e. Utilisez CloudWatch Metrics Insights pour interroger et analyser les données des métriques entre les comptes et les régions, en identifiant les tendances et les anomalies.
   f. Appliquez CloudWatch Metric Math pour transformer, agréger ou effectuer des calculs sur vos métriques afin d'obtenir des informations plus approfondies.
3. Ayez recours à Amazon DevOps Guru : Incorporez Amazon DevOps Guru pour sa détection des anomalies améliorée par le machine learning afin d'identifier les premiers signes de problèmes opérationnels pour vos applications sans serveur et de les résoudre avant qu'ils n'affectent vos clients.
4. Optimisez en fonction des informations : Prenez des décisions éclairées grâce à l'analyse de vos métriques afin d'ajuster et d'améliorer vos charges de travail.

Niveau d'effort du plan d'implémentation : Moyen

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identification des indicateurs clés de performance
- OPS04-BP02 Mise en œuvre de la télémétrie de l'application

Documents connexes :

- The Wheel Blog : souligner l'importance de revoir continuellement les métriques
- Importance des centiles
- Utiliser AWS Cost Anomaly Detection

- Observabilité CloudWatch entre comptes
- Interroger vos métriques avec CloudWatch Metrics Insights

Vidéos connexes :

- Enable Cross-Account Observability in Amazon CloudWatch
- Introduction to Amazon DevOps Guru
- Continuously Analyze Metrics using AWS Cost Anomaly Detection

Exemples connexes :

- Un atelier sur l'observabilité
- Obtenir des informations opérationnelles grâce à l'AIOps en utilisant Amazon DevOps Guru

OPS08-BP02 Analyse des journaux de charge de travail

L'analyse régulière des journaux de charge de travail est essentielle pour mieux comprendre les aspects opérationnels de votre application. En analysant, en visualisant et en interprétant efficacement les données des journaux, vous pouvez optimiser en permanence les performances et la sécurité des applications.

Résultat souhaité : Informations détaillées sur le comportement et le fonctionnement des applications grâce à une analyse approfondie des journaux, garantissant une détection et une atténuation proactives des problèmes.

Anti-modèles courants :

- Négliger l'analyse des journaux jusqu'à ce qu'un problème critique survienne.
- Ne pas utiliser la suite complète d'outils disponibles pour l'analyse des journaux, ce qui fait passer à côté d'informations critiques.
- Se fier uniquement à l'examen manuel des journaux sans tirer parti des fonctionnalités d'automatisation et de requête.

Avantages liés au respect de cette bonne pratique :

- Identification proactive des goulots d'étranglement opérationnels, des menaces de sécurité et d'autres problèmes potentiels.

- Utilisation efficace des données de journal pour une optimisation continue des applications.
- Meilleure compréhension du comportement des applications, ce qui aide au débogage et au dépannage.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Amazon CloudWatch Logs est un outil puissant pour l'analyse des journaux. Des fonctionnalités intégrées telles que CloudWatch Logs Insights et Contributor Insights rendent intuitif et efficace le processus d'obtention d'informations pertinentes à partir des journaux.

Étapes d'implémentation

1. Configurez CloudWatch Logs : configurez les applications et les services auxquels envoyer les journaux CloudWatch Logs.
2. Configurez CloudWatch Logs Insights : Utilisez CloudWatch Logs Insights pour rechercher et analyser de manière interactive vos données de journaux.
   a. Créez des requêtes pour extraire des modèles, visualiser les données des journaux et obtenir des informations exploitables.
3. Exploitez Contributor Insights Utilisez CloudWatch Contributor Insights pour identifier les personnes qui participent le plus dans des dimensions à cardinalité élevée, telles que les adresses IP ou les user-agents.
4. Implémentez des filtres de métriques CloudWatch Logs : configurez des filtres de métriques de journaux CloudWatch pour convertir les données de journaux en métriques exploitables. Cela vous permettra de définir des alarmes ou d'analyser davantage les modèles.
5. Remettez-vous en question régulièrement et affinez les stratégies en conséquence : passez régulièrement en revue vos stratégies d'analyse des journaux afin de recueillir toutes les informations pertinentes et d'optimiser en permanence les performances des applications.

Niveau d'effort du plan d'implémentation : moyen.

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identification des indicateurs clés de performance

- OPS04-BP02 Mise en œuvre de la télémétrie de l'application
- OPS08-BP01 Analyse des métriques de charge de travail

Documents connexes :

- Analyse des données des journaux avec CloudWatch Logs Insights
- Utilisation de CloudWatch Contributor Insights
- Création et gestion de filtres de métriques de journaux CloudWatch Logs

Vidéos connexes :

- Analyze Log Data with CloudWatch Logs Insights
- Use CloudWatch Contributor Insights to Analyze High-Cardinality Data

Exemples connexes :

- Exemples de requêtes CloudWatch Logs
- Un atelier sur l'observabilité

OPS08-BP03 Analyse des données de suivi de la charge de travail

L'analyse des données de suivi est essentielle pour obtenir une vue complète du parcours opérationnel d'une application. En visualisant et en comprenant les interactions entre les différents composants, il est possible d'affiner les performances, d'identifier les goulots d'étranglement et d'améliorer l'expérience utilisateur.

Résultat souhaité : Vous bénéficiez d'une visibilité claire sur les opérations distribuées de votre application, ce qui permet de résoudre les problèmes plus rapidement et d'améliorer l'expérience utilisateur.

Anti-modèles courants :

- Négliger les données de suivi, en s'appuyant uniquement sur les journaux et les métriques.
- Aucune corrélation entre les données de suivi et les journaux associés.
- Ignorer les métriques dérivées des données de suivi, telles que la latence et les taux de défaillance.

Avantages liés au respect de cette bonne pratique :

- Améliorez le dépannage et réduisez le temps moyen de résolution (MTTR).
- Obtenez des informations exploitables sur les dépendances et leur impact.
- Accélérez l'identification et la résolution des problèmes de performance.
- Tirez parti des métriques dérivées des données de suivi pour une prise de décision éclairée.
- Améliorez les expériences utilisateur grâce à des interactions optimisées entre les composants.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

AWS X-Ray propose une suite complète pour l'analyse des données de suivi. Il fournit une vue globale des interactions entre les services, surveille les activités des utilisateurs et détecte les problèmes de performance. Des fonctionnalités telles que ServiceLens, X-Ray Insights, X-Ray Analytics et Amazon DevOps Guru améliorent la profondeur des informations exploitables dérivées des données de suivi.

Étapes d'implémentation

Les étapes suivantes proposent une approche structurée pour mettre en œuvre efficacement l'analyse des données de suivi à l'aide des services AWS :

1. Intégrez AWS X-Ray : Assurez-vous qu'X-Ray est intégré à vos applications pour capturer les données de suivi.
2. Analysez les métriques X-Ray : Explorez les métriques dérivées des données de suivi X-Ray telles que la latence, les taux de requêtes, les taux d'erreur et les distributions de temps de réponse à l'aide de la cartographie des services pour surveiller l'état des applications.
3. Utilisez ServiceLens : Tirez parti de la cartographie ServiceLens pour une meilleure observabilité de vos services et applications. Cela permet une visualisation intégrée des données de suivi, des métriques, des journaux, des alarmes et d'autres informations liées à l'état.
4. Activez X-Ray Insights :
   a. Activez X-Ray Insights pour la détection automatique des anomalies dans les données de suivi.
   b. Examinez les informations pour identifier les tendances et en déterminer les causes profondes, telles que l'augmentation des taux de défaillance ou des latences.
   c. Consultez la chronologie des informations pour une analyse temporelle des problèmes détectés.

5. Utilisez X-Ray Analytics : X-Ray Analytics vous permet d'explorer en profondeur les données de suivi, d'identifier des modèles et d'en extraire des informations.
6. Utilisez des groupes dans X-Ray : Créez des groupes X-Ray pour filtrer les données de suivi en fonction de critères tels qu'une latence élevée, afin de permettre une analyse plus ciblée.
7. Incorporez Amazon DevOps Guru : Impliquez Amazon DevOps Guru pour bénéficier de modèles de machine learning identifiant les anomalies opérationnelles dans les données de suivi.
8. Utilisez CloudWatch Synthetics : Utilisez CloudWatch Synthetics pour créer des canarys permettant de surveiller en permanence vos points de terminaison et vos flux de travail. Ces canarys peuvent s'intégrer à X-Ray pour fournir des données de suivi permettant une analyse approfondie des applications testées.
9. Utilisez la surveillance des utilisateurs réels (RUM) : Avec AWS X-Ray et CloudWatch RUM, vous pouvez analyser et déboguer le chemin de la requête en commençant par les utilisateurs finaux de votre application via les services AWS gérés en aval. Cela vous permet d'identifier les tendances de latence et les erreurs qui ont un impact sur vos utilisateurs.
10. Corrélez les données de suivi aux journaux : Corrélez les données de suivi aux journaux associés dans la vue de suivi X-Ray pour une perspective granulaire du comportement des applications. Cela vous permet de visualiser les événements de journal directement associés aux transactions suivies.

Niveau d'effort du plan d'implémentation : moyen.

Ressources

Bonnes pratiques associées :

- OPS08-BP01 Analyse des métriques de charge de travail
- OPS08-BP02 Analyse des journaux de charge de travail

Documents connexes :

- Utilisation de ServiceLens pour surveiller l'état des applications
- Exploration des données de suivi grâce à X-Ray Analytics
- Détection des anomalies dans les données de suivi grâce à X-Ray Insights
- Surveillance continue avec CloudWatch Synthetics

Vidéos connexes :

- Analyze and Debug Applications Using Amazon CloudWatch Synthetics and AWS X-Ray
- Use AWS X-Ray Insights

Exemples connexes :

- Un atelier sur l'observabilité
- Mise en œuvre d'X-Ray avec AWS Lambda
- Modèles canary CloudWatch Synthetics

OPS08-BP04 Création d'alertes exploitables

Il est crucial de détecter rapidement les écarts de comportement de votre application et d'y réagir rapidement. Il est particulièrement important de savoir quand les résultats basés sur les indicateurs clés de performance (KPI) sont menacés ou lorsque des anomalies inattendues surviennent. Le fait de baser les alertes sur les KPI garantit que les signaux que vous recevez sont directement liés à l'impact commercial ou opérationnel. Cette approche des alertes exploitables favorise les réponses proactives et contribue à maintenir les performances et la fiabilité du système.

Résultat souhaité : Vous recevez des alertes opportunes, pertinentes et exploitables permettent d'identifier et d'atténuer rapidement les problèmes potentiels, en particulier lorsque les résultats basés sur les KPI sont menacés.

Anti-modèles courants :

- Configurer un trop grand nombre d'alertes non critiques, ce qui entraîne de la lassitude.
- Ne pas hiérarchiser les alertes en fonction des KPI, ce qui complique la compréhension de l'impact commercial des problèmes.
- Négliger de traiter les causes profondes, ce qui entraîne des alertes répétitives pour le même problème.

Avantages liés au respect de cette bonne pratique :

- Réduction de la lassitude liée aux alertes grâce à des alertes pertinentes et exploitables.
- Disponibilité et fiabilité du système améliorées grâce à la détection et à l'atténuation proactives des problèmes.

- Collaboration d'équipe améliorée et résolution plus rapide des problèmes grâce à l'intégration à des outils connus d'alerte et de communication.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Pour créer un mécanisme d'alerte efficace, il est essentiel d'utiliser des métriques, des journaux et des données de suivi qui signalent les risques liés aux résultats basés sur les KPI ou les anomalies détectées.

Étapes d'implémentation

1. Déterminez les indicateurs clés de performance (KPI) : Identifiez les KPI de votre application. Les alertes doivent être liées à ces KPI afin de refléter avec précision l'impact commercial.
2. Mettez en œuvre la détection des anomalies :
   - Utilisez AWS Cost Anomaly Detection : Configurez AWS Cost Anomaly Detection pour détecter automatiquement les modèles inhabituels, en veillant à ce que les alertes ne soient générées que pour les anomalies réelles.
   - Utilisez X-Ray Insights :
     a. Configurez X-Ray Insights pour détecter les anomalies dans les données de suivi.
     b. Configurez les notifications pour X-Ray Insights pour être alerté des problèmes détectés.
   - Intégrez DevOps Guru :
     a. Exploitez Amazon DevOps Guru pour ses fonctionnalités de machine learning permettant de détecter les anomalies opérationnelles avec des données existantes.
     b. Accédez aux paramètres de notification dans DevOps Guru pour configurer des alertes d'anomalie.
3. Implémentez des alertes exploitables : Concevez des alertes qui fournissent des informations adéquates pour une action immédiate.
4. Réduisez la lassitude liée aux alarmes : Minimisez les alertes non critiques. Le fait de surcharger les équipes avec de nombreuses alertes insignifiantes peut les inciter à négliger des problèmes critiques et diminuer l'efficacité globale du mécanisme d'alerte.
5. Configurez des alarmes composites : Utilisez des alarmes composites Amazon CloudWatch pour regrouper plusieurs alarmes.
6. Intégrez des outils d'alerte : Intégrez des outils tels que Ops Genie et PagerDuty.

7. Impliquez AWS Chatbot Intégrez AWS Chatbotpour transmettre des alertes à Chime, Microsoft Teams et Slack.
8. Alerte basée sur les journaux : Utilisez des filtres de métriques de journaux dans CloudWatch pour créer des alarmes basées sur des événements de journal spécifiques.
9. Passez en revue et répétez : Revoyez et affinez régulièrement les configurations des alertes.

Niveau d'effort du plan d'implémentation : moyen.

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identification des indicateurs clés de performance
- OPS04-BP02 Mise en œuvre de la télémétrie de l'application
- OPS04-BP03 Mise en œuvre de la télémétrie de l'expérience utilisateur
- OPS04-BP04 Mise en œuvre de la télémétrie des dépendances
- OPS04-BP05 Mise en œuvre du suivi distribué
- OPS08-BP01 Analyse des métriques de charge de travail
- OPS08-BP02 Analyse des journaux de charge de travail
- OPS08-BP03 Analyse des données de suivi de la charge de travail

Documents connexes :

- Utilisation des alarmes Amazon CloudWatch
- Création d'une alarme composite
- Création d'une alarme CloudWatch basée sur la détection d'anomalies
- Notifications DevOps Guru
- Notifications X-Ray Insights
- Surveiller, gérer et dépanner vos ressources AWS grâce au ChatOps interactif
- Guide d'intégration Amazon CloudWatch | PagerDuty
- Intégrer OpsGenie à Amazon CloudWatch

Vidéos connexes :

- Create Composite Alarms in Amazon CloudWatch
- AWS Chatbot Overview
- AWS on Air ft. Mutative Commands in AWS Chatbot

Exemples connexes :

- Alarmes, gestion des incidents et résolution dans le cloud avec Amazon CloudWatch
- Tutoriel : création d'une règle Amazon EventBridge qui envoie des notifications à AWS Chatbot
- Un atelier sur l'observabilité

OPS08-BP05 Création de tableaux de bord

Les tableaux de bord offrent une vue centrée sur l'humain des données télémétriques de vos charges de travail. Bien qu'ils fournissent une interface visuelle essentielle, ils ne doivent pas remplacer les mécanismes d'alerte, mais les compléter. Lorsqu'ils sont conçus avec soin, ils peuvent non seulement fournir des informations rapides sur l'état et les performances du système, mais ils peuvent également présenter aux parties prenantes des informations en temps réel sur les résultats commerciaux et l'impact des problèmes.

Résultat souhaité : Informations claires et exploitables sur l'état du système et de l'entreprise à l'aide de représentations visuelles.

Anti-modèles courants :

- Tableaux de bord trop compliqués avec trop de métriques.
- Utilisation de tableaux de bord sans alertes pour détecter les anomalies.
- Pas de mise à jour des tableaux de bord à mesure que les charges de travail évoluent.

Avantages liés au respect de cette bonne pratique :

- Visibilité immédiate sur les métriques critiques du système et les KPI.
- Amélioration de la communication et de la compréhension avec les parties prenantes.
- Aperçu rapide de l'impact des problèmes opérationnels.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Tableaux de bord centrés sur l'entreprise

Les tableaux de bord adaptés aux indicateurs clés de performance de l'entreprise mobilisent un plus large éventail de parties prenantes. Bien que ces personnes ne soient pas intéressées par les métriques du système, elles souhaitent comprendre les implications commerciales de ces chiffres. Un tableau de bord centré sur l'entreprise garantit que toutes les métriques techniques et opérationnelles surveillées et analysées sont synchronisées avec les objectifs globaux de l'entreprise. Cet alignement apporte de la clarté et garantit que tout le monde est d'accord sur ce qui est essentiel et sur ce qui ne l'est pas. En outre, les tableaux de bord qui mettent en évidence les KPI commerciaux ont tendance à être plus exploitables. Les parties prenantes peuvent rapidement comprendre l'état des opérations, les domaines nécessitant une attention particulière et l'impact potentiel sur les résultats commerciaux.

Dans cette optique, lors de la création de vos tableaux de bord, assurez-vous qu'il existe un juste milieu entre les métriques techniques et les KPI commerciaux. Les deux sont essentiels, mais ils s'adressent à des publics différents. Idéalement, vous devriez disposer de tableaux de bord offrant une vue globale de l'état et des performances du système tout en mettant l'accent sur les principaux résultats commerciaux et leurs implications.

Les tableaux de bord Amazon CloudWatch sont des pages d'accueil personnalisables de la console CloudWatch. Vous pouvez les utiliser pour surveiller vos ressources dans une seule fenêtre, y compris les ressources réparties sur différents comptes et différentes Régions AWS.

Étapes d'implémentation

1. Créez un tableau de bord de base : Créez un tableau de bord dans CloudWatch, en lui donnant un nom descriptif.
2. Utilisez les widgets Markdown : avant de vous plonger dans les métriques, utilisez les widgets Markdown pour ajouter un contexte textuel en haut de votre tableau de bord. Expliquez ce que couvre le tableau de bord et l'importance des métriques représentées, et ajoutez éventuellement des liens vers d'autres tableaux de bord et outils de résolution des problèmes.
3. Créez des variables de tableau de bord : le cas échéant, incorporez des variables pour permettre des vues de tableau de bord dynamiques et flexibles.
4. Créez des widgets de statistiques : ajoutez des widgets de métriques pour visualiser les différentes métriques émises par votre application, en personnalisant ces widgets pour représenter efficacement l'état du système et les résultats commerciaux.

5. Requêtes Logs Insights : utilisez CloudWatch Logs Insights pour obtenir des métriques exploitables à partir de vos journaux et afficher ces informations sur votre tableau de bord.
6. Configurez les alarmes : intégrez des alarmes CloudWatch dans votre tableau de bord pour un aperçu rapide de toutes les métriques dépassant leurs seuils.
7. Utilisez Contributor Insights : incorporez CloudWatch Contributor Insights pour analyser les champs à forte cardinalité et mieux comprendre les principaux contributeurs de votre ressource.
8. Concevez des widgets personnalisés : pour les besoins spécifiques non satisfaits par les widgets standard, pensez à créer des widgets personnalisés. Ils peuvent être extraits de différentes sources de données ou représenter les données de manière unique.
9. Répétez et affinez : au fur et à mesure que votre application évolue, revoyez régulièrement votre tableau de bord pour vous assurer de sa pertinence.

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identification des indicateurs clés de performance
- OPS08-BP01 Analyse des métriques de charge de travail
- OPS08-BP02 Analyse des journaux de charge de travail
- OPS08-BP03 Analyse des données de suivi de la charge de travail
- OPS08-BP04 Création d'alertes exploitables

Documents connexes :

- Création de tableaux de bord pour une visibilité opérationnelle
- Fonctionnement des tableaux de bord Amazon CloudWatch

Vidéos connexes :

- Create Cross Account & Cross Region CloudWatch Dashboards
- AWS re:Invent 2021 - Gain enterprise visibility with AWS Cloud operation dashboards

Exemples connexes :

- Un atelier sur l'observabilité

- Surveillance des applications avec Amazon CloudWatch

OPS 9. Comment comprendre l'état de vos opérations ?

Définissez, capturez et analysez les métriques des opérations pour obtenir une visibilité sur les événements des opérations afin de pouvoir prendre des mesures appropriées.

Bonnes pratiques

- OPS09-BP01 Mesure des objectifs opérationnels et des KPI à l'aide de métriques
- OPS09-BP02 Communication de l'état et des tendances pour garantir la visibilité des opérations
- OPS09-BP03 Vérification des métriques des opérations et définition de la priorité des améliorations

OPS09-BP01 Mesure des objectifs opérationnels et des KPI à l'aide de métriques

Obtenez des objectifs et des indicateurs clés de performance qui définissent le succès des opérations de votre organisation et déterminez les métriques qui les reflètent. Définissez des points de référence et réévaluez-les régulièrement. Développez des mécanismes permettant de recueillir ces métriques auprès des équipes à des fins d'évaluation.

Résultat souhaité :

- Les objectifs et les KPI des équipes opérationnelles de l'organisation ont été publiés et partagés.
- Des métriques reflétant ces KPI sont établies. Exemples :
  - Profondeur de la file d'attente ou âge moyen des demandes d'assistance
  - Nombre de demandes d'assistance regroupées par type de problème
  - Temps passé à résoudre les problèmes avec ou sans procédure opérationnelle normalisée (SOP)
  - Délai de récupération après un échec d'envoi de code
  - Volume d'appels

Anti-modèles courants :

- Les délais de déploiement ne sont pas respectés, car les développeurs sont contraints d'effectuer des tâches de dépannage. Les équipes de développement plaident en faveur d'une augmentation du personnel, mais ne peuvent pas quantifier le nombre de collaborateurs dont elles ont besoin, car le temps perdu ne peut pas être mesuré.

- Un bureau de niveau 1 a été mis en place pour traiter les appels des utilisateurs. Au fil du temps, de nouvelles charges de travail ont été ajoutées, mais aucun effectif n'a été affecté au bureau de niveau 1. La satisfaction des clients en pâtit alors que le temps d'appel augmente et que la résolution des problèmes ralentit, mais la direction n'en voit aucun signe, empêchant toute action.
- Une charge de travail problématique a été confiée à une équipe opérationnelle distincte. Contrairement aux autres charges de travail, celle-ci n'a pas été fournie avec la documentation et les runbooks appropriés. Les équipes consacrent donc plus de temps au dépannage et à la résolution des défaillances. Cependant, aucune métrique ne permet de documenter ces efforts, ce qui empêche les équipes de rendre compte de la situation.

Avantages liés au respect de cette bonne pratique : Lorsque la surveillance de la charge de travail indique l'état des applications et services, les équipes chargées de la surveillance des opérations fournissent aux parties prenantes un aperçu des changements survenus chez les consommateurs de ces charges de travail, tels que l'évolution des besoins commerciaux. Mesurez l'efficacité de ces équipes et évaluez-les par rapport aux objectifs commerciaux en créant des métriques qui reflètent l'état des opérations. Ces métriques peuvent mettre en évidence les problèmes de support ou identifier les cas où des écarts se produisent par rapport à un objectif de niveau de service.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Planifiez du temps avec les responsables et les parties prenantes afin de déterminer les objectifs généraux du service. Déterminez quelles devraient être les tâches des différentes équipes opérationnelles et quels défis elles pourraient rencontrer. Sur la base de ces informations, réfléchissez à des indicateurs clés de performance (KPI) susceptibles de refléter ces objectifs opérationnels. Il peut s'agir de la satisfaction du client, du délai entre la conception des fonctionnalités et le déploiement, du temps moyen de résolution des problèmes, etc.

À partir de ces KPI, identifiez les métriques et les sources de données qui pourraient mieux refléter ces objectifs. La satisfaction des clients peut être une combinaison de diverses métriques telles que les temps d'attente ou de réponse aux appels, les scores de satisfaction et les types de problèmes soulevés. Les temps de déploiement peuvent être la somme du temps nécessaire aux tests et au déploiement, plus les correctifs à ajouter après le déploiement lui-même. Les statistiques indiquant le temps consacré à différents types de problèmes (ou le nombre de ces problèmes) peuvent fournir un aperçu des domaines dans lesquels des efforts ciblés sont nécessaires.

Ressources

Documents connexes :

- Amazon QuickSight – Utilisation des KPI
- Amazon CloudWatch – Utilisation des métriques
- Création de tableaux de bord
- How to track your cost optimization KPIs with KPI Dashboard

OPS09-BP02 Communication de l'état et des tendances pour garantir la visibilité des opérations

Connaître l'état de vos opérations et leurs tendances est nécessaire pour identifier les cas où les résultats peuvent être menacés, pour déterminer si des efforts supplémentaires sont justifiés ou non, ou pour identifier les effets des modifications sur vos équipes. Lors d'événements opérationnels, le fait de disposer de pages d'état auxquelles les utilisateurs et les équipes opérationnelles peuvent se référer pour obtenir des informations contribue à réduire la pression sur les canaux de communication et à diffuser les informations de manière proactive.

Résultat souhaité :

- Les responsables des opérations ont un aperçu rapide des volumes d'appels auxquels leurs équipes sont confrontées et des initiatives en cours, telles que les déploiements.
- Des alertes sont diffusées aux parties prenantes et aux communautés d'utilisateurs lorsque des répercussions sur les opérations normales se produisent.
- La direction de l'organisation et les parties prenantes peuvent consulter une page d'état en réponse à une alerte ou à un impact, et obtenir des informations concernant un événement opérationnel, telles que les points de contact, des informations sur les demandes d'assistance et les délais de reprise estimés.
- Des rapports sont mis à la disposition de la direction et des autres parties prenantes pour présenter des statistiques opérationnelles telles que le volume d'appels sur une période donnée, les scores de satisfaction des utilisateurs, le nombre de demandes d'assistance en attente et leur ancienneté.

Anti-modèles courants :

- Une charge de travail tombe en panne, ce qui rend un service indisponible. Le volume d'appels augmente lorsque les utilisateurs demandent à savoir ce qui se passe. Les responsables ajoutent

au volume en demandant à savoir qui est à l'origine du problème. Les différentes équipes opérationnelles redoublent leurs efforts pour tenter d'identifier la cause première.

- Pour répondre à un nouveau besoin, plusieurs membres du personnel sont réaffectés à un effort d'ingénierie. Les postes vacants ne sont pas pourvus, et les délais de résolution des problèmes augmentent. Ces informations ne sont pas capturées, et ce n'est qu'après plusieurs semaines et après avoir reçu des commentaires insatisfaits des utilisateurs que les dirigeants prennent conscience du problème.

Avantages liés au respect de cette bonne pratique : Lors d'événements opérationnels affectant l'entreprise, beaucoup de temps et d'énergie peuvent être gaspillés à demander des informations aux différentes équipes qui tentent de comprendre la situation. En mettant en place des pages d'état et des tableaux de bord largement diffusés, les parties prenantes peuvent rapidement se procurer les informations nécessaires et déterminer, par exemple, si un problème a été détecté ou non, qui est responsable du problème ou quand un retour à une activité normale est attendu. Cela évite aux membres de l'équipe d'avoir à passer trop de temps à communiquer la situation aux autres. Ils peuvent ainsi consacrer plus de temps à la résolution des problèmes.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Créez des tableaux de bord qui présentent les métriques clés actuelles pour vos équipes opérationnelles et mettez-les à disposition des responsables des opérations et de la direction.

Créez des pages d'état qui peuvent être mises à jour rapidement pour indiquer quand un incident ou un événement se produit, qui en est le responsable et qui coordonne la réponse. Partagez sur cette page les étapes ou les solutions que les utilisateurs doivent envisager et diffusez largement son emplacement. Encouragez les utilisateurs à vérifier d'abord cet emplacement lorsqu'ils sont confrontés à un problème inconnu.

Collectez et fournissez des rapports qui présentent l'état des opérations au fil du temps, et distribuez-les aux dirigeants et aux décideurs pour illustrer le travail des opérations ainsi que les défis et les besoins.

Partagez entre les équipes les métriques et rapports qui reflètent au mieux les objectifs et les KPI, ainsi que les domaines où ils ont contribué au changement. Consacrez du temps à ces activités afin de renforcer l'importance des opérations au sein des équipes et entre elles.

Ressources

Documents connexes :

- Mesurer les progrès
- Création de tableaux de bord pour une visibilité opérationnelle

Solutions associées :

- Opérations de données

OPS09-BP03 Vérification des métriques des opérations et définition de la priorité des améliorations

Le fait de consacrer du temps et des ressources à l'examen de l'état des opérations garantit que le service quotidien des activités demeure une priorité. Réunissez les responsables des opérations et les parties prenantes pour vérifier régulièrement les métriques, réaffirmer ou modifier les objectifs et prioriser les améliorations.

Résultat souhaité :

- Les responsables des opérations et le personnel se rencontrent régulièrement pour vérifier les métriques au cours d'une période de référence donnée. Les défis sont communiqués, les victoires sont célébrées et les leçons tirées sont partagées.
- Les parties prenantes et les responsables sont régulièrement informés de l'état des opérations et sont invités à donner leur avis concernant les objectifs, les KPI et les initiatives futures. Les compromis entre la prestation de services, les opérations et la maintenance font l'objet de discussions et sont mis en contexte.

Anti-modèles courants :

- Un nouveau produit est lancé, mais les équipes opérationnelles de niveau 1 et de niveau 2 ne sont pas suffisamment formées pour fournir l'assistance nécessaire ou n'ont pas de personnel supplémentaire. Les métriques qui montrent une dégradation des délais de résolution des demandes d'assistance et l'augmentation du volume d'incidents ne sont pas pris en compte par les dirigeants. Des mesures sont prises des semaines plus tard lorsque le nombre d'abonnements commence à baisser alors que les utilisateurs mécontents quittent la plateforme.
- Un processus manuel pour effectuer la maintenance d'une charge de travail est en place depuis longtemps. Bien que le désir d'automatiser soit présent, cela n'est pas une priorité compte tenu de

la faible importance du système. Cependant, au fil du temps, le système gagne de l'importance et ces processus manuels occupent désormais la majeure partie du temps des opérations. Aucune ressource n'est prévue pour assister les opérations, ce qui entraîne un épuisement du personnel à mesure que la charge de travail augmente. La direction n'en prend conscience que lorsqu'on lui signale que le personnel démissionne pour aller travailler pour d'autres concurrents.

Avantages liés au respect de cette bonne pratique : Dans certaines organisations, il peut être difficile de consacrer le même temps et la même attention à la prestation de services et aux nouveaux produits ou offres. Lorsque cela a lieu, le secteur d'activité peut en pâtir, car le niveau de service attendu se détériore lentement. Cela s'explique par le fait que les opérations ne changent pas et n'évoluent pas avec la croissance de l'entreprise, et peuvent se retrouver à la traîne. Sans un examen régulier des informations recueillies par les opérations, le risque pour l'entreprise peut ne devenir visible que lorsqu'il sera trop tard. En allouant du temps à l'examen des métriques et des procédures à la fois au sein des équipes opérationnelles et auprès de la direction, le rôle crucial joué par les opérations reste visible, et les risques peuvent être identifiés bien avant qu'ils n'atteignent des niveaux critiques. Les équipes opérationnelles ont une meilleure idée des changements et initiatives commerciaux imminents, ce qui permet de lancer des initiatives proactives. La visibilité qu'ont les dirigeants sur les métriques opérationnelles met en évidence le rôle que jouent ces équipes dans la satisfaction des clients, à la fois en interne et en externe, et leur permet de mieux évaluer les choix en fonction des priorités, ou de s'assurer que les opérations disposent du temps et des ressources nécessaires pour changer et évoluer avec de nouvelles initiatives stratégiques et de charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Consacrez du temps à la vérification des métriques opérationnelles entre les parties prenantes et les équipes opérationnelles et à l'examen des données des rapports. Placez ces rapports dans le contexte des objectifs de l'organisation afin de déterminer s'ils sont atteints. Identifiez les sources d'ambiguïté lorsque les objectifs ne sont pas clairs ou lorsqu'il peut y avoir des conflits entre ce qui est demandé et ce qui est fourni.

Identifiez les domaines où plus de temps, plus de personnel et plus d'outils peuvent contribuer à de meilleurs résultats des opérations. Déterminez les KPI sur lesquels cela aurait un impact et quels devraient être les objectifs de réussite. Révisez-les régulièrement pour vous assurer que les opérations disposent de ressources suffisantes pour soutenir le secteur d'activité.

Ressources

Documents connexes :

- Amazon Athena
- Référence aux dimensions et métriques Amazon CloudWatch
- Amazon QuickSight
- AWS Glue
- AWS Glue Data Catalog
- Collecter des métriques et des journaux auprès d'instances Amazon EC2 et de serveurs sur site avec l'agent Amazon CloudWatch
- Utilisation des métriques Amazon CloudWatch

OPS 10. Comment gérer les événements relatifs à la charge de travail et aux opérations ?

Préparez et validez des procédures de réponse aux événements afin de réduire leur effet disruptif sur votre charge de travail.

Bonnes pratiques

- OPS10-BP01 Utiliser un processus pour la gestion des événements, des incidents et des problèmes
- OPS10-BP02 Disposer d'un processus par alerte
- OPS10-BP03 Hiérarchiser les événements opérationnels en fonction de leur impact sur l'activité
- OPS10-BP04 Définir l'acheminement hiérarchique
- OPS10-BP05 Définissez un plan de communication avec les clients en cas d'interruption de service
- OPS10-BP06 Communiquer l'état grâce aux tableaux de bord
- OPS10-BP07 Automatiser les réponses aux événements

OPS10-BP01 Utiliser un processus pour la gestion des événements, des incidents et des problèmes

Votre entreprise dispose de processus pour gérer les événements, les incidents et les problèmes. Événements se produisent dans votre charge de travail, mais ne nécessitent pas d'intervention. Les incidents sont des événements qui nécessitent une intervention. Les problèmes sont des événements

récurrents qui nécessitent une intervention ou ne peuvent pas être résolus. Vous avez besoin de processus pour réduire l'impact de ces événements sur votre entreprise et répondre de manière adaptée.

Lorsque des incidents et des problèmes se produisent dans votre charge de travail, vous avez besoin de processus pour les gérer. Comment communiquer le statut de l'événement aux parties prenantes ? Qui supervise l'intervention ? Quels sont les outils à utiliser pour réduire l'impact de ces événements ? Voici des exemples de questions auxquelles vous devez répondre pour renforcer votre processus de réponse.

Les processus doivent être documentés dans un emplacement central et accessible à toute personne impliquée dans votre charge de travail. Si vous ne disposez pas d'un wiki central ou d'un magasin de documents, un référentiel de contrôle de version peut être utilisé. Vous devez garder ces plans à jour à mesure que vos processus évoluent.

Les problèmes sont de parfaits candidats à l'automatisation. Ces événements empiètent sur votre temps passé à innover. Commencez par créer un processus reproductible pour réduire l'impact du problème. Avec le temps, concentrez-vous sur l'automatisation de la réduction ou de la résolution du problème sous-jacent. Cela permet de libérer du temps pour vous consacrer à l'amélioration de votre charge de travail.

Résultat souhaité : Votre entreprise dispose d'un processus pour gérer les événements, les incidents et les problèmes. Ces processus sont documentés et stockés dans un emplacement central. Ils sont mis à jour à mesure que les processus évoluent.

Anti-modèles courants :

- Un incident se produit pendant le week-end et l'ingénieur de garde ne sait pas quoi faire.
- Un client vous envoie un e-mail pour vous informer que l'application ne fonctionne plus. Vous redémarrez le serveur pour résoudre le problème. Cela arrive fréquemment.
- Un incident se produit et plusieurs équipes travaillent indépendamment pour essayer de le résoudre.
- Des déploiements se produisent dans votre charge de travail sans être enregistrés.

Avantages liés au respect de cette bonne pratique :

- Vous disposez d'une piste d'audit des événements dans votre charge de travail.
- Votre temps de récupération après un incident diminue.

- Les membres de l'équipe peuvent résoudre des incidents et des problèmes de manière cohérente.
- L'effort est plus consolidé lorsqu'on enquête sur un incident.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

L'implémentation de cette bonne pratique signifie que vous suivez les événements de charge de travail. Vous disposez de processus pour gérer les incidents et les problèmes. Les processus sont documentés, partagés et mis à jour fréquemment. Les problèmes sont identifiés, hiérarchisés et résolus.

Exemple client

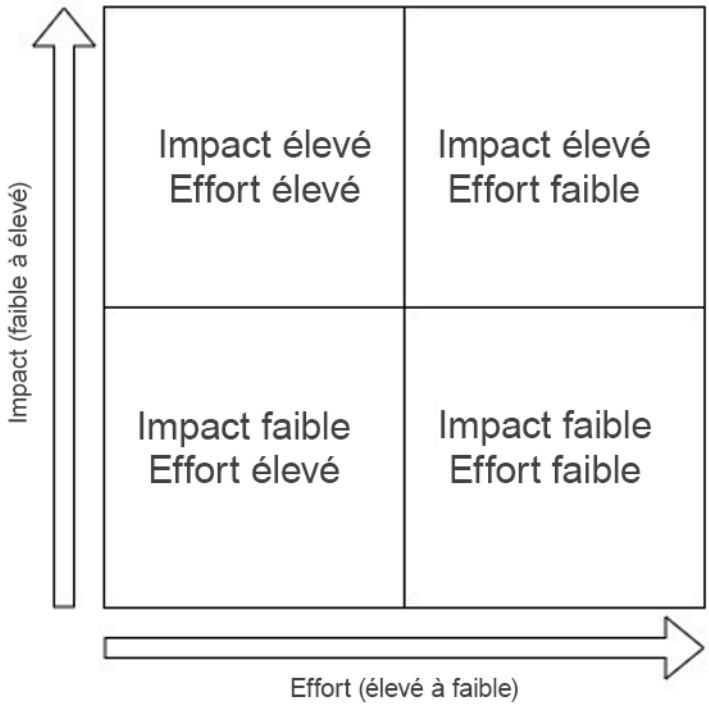
Une partie du wiki interne d'AnyCompany Retail est consacrée au processus pour la gestion de l'événement, de l'incident et du problème. Tous les événements sont envoyés à Amazon EventBridge. Les problèmes sont identifiés en tant qu'OpsItems dans AWS Systems Manager OpsCenter et hiérarchisés pour être résolus, ce qui réduit la main d'œuvre indifférenciée. À mesure que les processus évoluent, ils sont mis à jour dans son wiki interne. L'entreprise utilise AWS Systems Manager Incident Manager pour gérer les incidents et coordonner les efforts de réduction de l'impact des événements.

Étapes d'implémentation

1. Événements
    - Suivez les événements qui se produisent dans votre charge de travail, même si aucune intervention humaine n'est requise.
    - Collaborez avec les parties prenantes de la charge de travail pour développer une liste des événements devant être suivis. Certains exemples sont des déploiements terminés ou des correctifs réussis.
    - Vous pouvez utiliser des services comme Amazon EventBridge ou Amazon Simple Notification Service pour générer des événements personnalisés pour le suivi.
2. Les incidents
    - Commencez par définir le plan de communication pour les incidents. Quelles parties prenantes doivent être informées ? Comment les tiendrez-vous informées ? Qui supervise les efforts de coordination ? Nous recommandons de mettre en place un canal de chat interne pour la communication et la coordination.

- Définissez les chemins de remontée pour les équipes prenant en charge votre charge de travail, notamment si l'équipe n'a pas de système de rotation de garde. Selon votre niveau de prise en charge, vous pouvez également créer un ticket avec AWS Support.
- Créez un playbook pour enquêter sur l'incident. Il doit inclure le plan de communication et les étapes détaillées de l'enquête. Incluez la vérification du AWS Health Dashboard dans votre enquête.
- Documentez votre plan de réponse aux incidents. Communiquez le plan de gestion des incidents afin que les clients internes et externes comprennent les règles d'engagement et ce qu'on attend d'eux. Entraînez les membres de votre équipe à l'utiliser.
- Les clients peuvent utiliser Incident Manager pour configurer et gérer leur plan de réponse aux incidents.
- Les clients ayant un plan de support Business peuvent demander l' atelier Gestion des incidents auprès de leur gestionnaire de compte technique. Cet atelier guidé teste votre plan de réponse aux incidents existant et vous aide à identifier les domaines à améliorer.

3. Problèmes
    - Les problèmes doivent être identifiés et suivis dans votre système ITSM.
    - Identifiez tous les problèmes connus et hiérarchisez-les par effort de résolution et impact sur la charge de travail.



- Résolvez d'abord les problèmes ayant un impact élevé et un effort faible. Une fois ces problèmes résolus, passez à ceux ayant un impact faible et un effort faible.
- Vous pouvez utiliser Systems Manager OpsCenter pour identifier ces problèmes, leur attacher des runbooks et les suivre.

Niveau d'effort du plan d'implémentation : moyen. Vous avez besoin d'un processus et d'outils pour implémenter cette bonne pratique. Documentez vos processus et rendez-les accessibles à toute personne associée à la charge de travail. Mettez-les à jour fréquemment. Vous disposez d'un processus pour gérer les problèmes et les réduire ou les résoudre.

Ressources

Bonnes pratiques associées :

- OPS07-BP03 Utiliser des runbooks pour effectuer des procédures : les problèmes connus ont besoin d'un runbook associé pour que les efforts de réduction de l'impact soient cohérents.
- OPS07-BP04 Utiliser des playbooks pour analyser les problèmes : les incidents doivent faire l'objet d'une enquête en utilisant des playbooks.
- OPS11-BP02 Effectuer une analyse post-incident : procédez toujours à une analyse à froid suite à une reprise après un incident.

Documents connexes :

- Atlassian : gestion des incidents à l'âge du DevOps
- Guide des réponses aux incidents de sécurité AWS
- Gestion des incidents à l'âge du DevOps et de SRE
- PagerDuty : qu'est-ce que la gestion des incidents ?

Vidéos connexes :

- AWS re:Invent 2020: Incident management in a distributed organization
- AWS re:Invent 2021 - Building next-gen applications with event-driven architectures
- AWS Supports You | Exploring the Incident Management Tabletop Exercise
- AWS Systems Manager Incident Manager : ateliers virtuels AWS
- AWS What's Next ft. Incident Manager | Événements AWS

Exemples connexes :

- Atelier Outils de gestion et de gouvernance AWS : OpsCenter
- Services proactifs AWS : atelier de gestion des incidents
- Création d'une application gérée par les événements avec Amazon EventBridge
- Création d'architectures gérées par les événements sur AWS

Services associés :

- Amazon EventBridge
- Amazon SNS
- AWS Health Dashboard
- AWS Systems Manager Incident Manager
- AWS Systems Manager OpsCenter

OPS10-BP02 Disposer d'un processus par alerte

Répondez de manière bien définie (un runbook ou un playbook), avec un responsable spécifiquement identifié, à tout événement pour lequel vous déclenchez une alerte. Cela permet de répondre efficacement et rapidement aux événements liés aux opérations et d'éviter que les événements donnant lieu à une action ne soient occultés par des notifications de moindre valeur.

Anti-modèles courants :

- Votre système de surveillance vous présente un flux de connexions approuvées et d'autres messages. Le volume des messages est si important que vous manquez des messages d'erreur réguliers qui nécessitent votre intervention.
- Vous recevez une alerte indiquant que le site Web est en panne. Il n'existe aucun processus défini lorsque cela se produit. Vous êtes contraint d'adopter une approche ponctuelle pour diagnostiquer et résoudre le problème. Le développement de ce processus au fur et à mesure allonge le délai de reprise.

Avantages liés au respect de cette bonne pratique : En n'envoyant une alerte que lorsqu'une action est nécessaire, vous évitez que des alertes de faible importance ne dissimulent des alertes plus importantes. En ayant un processus en place pour toutes les alertes nécessitant une action, vous permettez une réponse cohérente et rapide aux événements dans votre environnement.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Processus par alerte : tout événement pour lequel vous déclenchez une alerte doit avoir une réponse bien définie (un runbook ou un playbook) avec un responsable spécifiquement identifié (par exemple, une personne, une équipe ou un rôle), garant du bon déroulement du processus. L'intervention peut être automatisée ou effectuée par une autre équipe, mais le responsable doit veiller à ce que le processus transmette les résultats attendus. En disposant de ces processus, vous garantissez des réponses efficaces et rapides aux événements opérationnels et vous pouvez empêcher que les événements concrets soient masqués par des notifications moins importantes. Par exemple, la mise à l'échelle automatique pourrait être appliquée pour mettre à l'échelle un front-end Web, mais l'équipe des opérations pourrait être responsable de s'assurer que les règles et les limites de mise à l'échelle automatique sont appropriées aux besoins de la charge de travail.

Ressources

Documents connexes :

- Fonctionnalités d'Amazon CloudWatch
- Qu'est-ce que Amazon CloudWatch Events ?

Vidéos connexes :

- Élaborer un plan de surveillance

OPS10-BP03 Hiérarchiser les événements opérationnels en fonction de leur impact sur l'activité

Assurez-vous que, lorsque plusieurs événements nécessitent une intervention, les plus importants pour l'activité sont traités en premier. Les impacts peuvent inclure la mort ou une blessure, une perte financière ou l'atteinte à la réputation ou à la confiance.

Anti-modèles courants :

- Vous recevez une demande de support pour ajouter une configuration d'imprimante pour un utilisateur. Alors que vous travaillez sur le problème, vous recevez une demande de support indiquant que votre site de vente au détail est en panne. Après avoir terminé la configuration de l'imprimante pour votre utilisateur, vous commencez à travailler sur le problème du site Web.

- Vous êtes averti que votre site Web de vente au détail et votre système de paie sont en panne. Vous ne savez pas auquel des deux vous devez accorder la priorité.

Avantages liés au respect de cette bonne pratique : La priorisation des réponses aux incidents ayant le plus grand impact sur l'entreprise vous permet de gérer cet impact.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Hiérarchiser les événements opérationnels en fonction de leur impact sur l'activité : établissez la priorité des événements en fonction de l'impact métier pour vous assurer que, lorsque plusieurs événements nécessitent une intervention, les plus importants pour l'activité sont traités en premier. Les impacts peuvent inclure un décès ou une blessure, des pertes financières, des violations réglementaires ou une atteinte à la réputation ou à la confiance.

OPS10-BP04 Définir l'acheminement hiérarchique

Définissez l'acheminement hiérarchique dans vos runbooks et playbooks, y compris ce qui le déclenche et les procédures qui le régissent. Identifiez spécifiquement les propriétaires de chaque action afin de garantir des réponses efficaces et rapides aux événements liés aux opérations.

Déterminez quand une décision humaine est nécessaire avant d'effectuer une action. Collaborez avec les décideurs pour que cette décision soit prise à l'avance et que l'action soit préapprouvée, afin que le temps moyen de résolution ne soit pas étendu.

Anti-modèles courants :

- Votre site de vente au détail est en panne. Vous ne comprenez pas le runbook pour restaurer le site. Vous commencez à appeler vos collègues en espérant que quelqu'un sera en mesure de vous aider.
- Vous recevez une demande de support pour une application inaccessible. Vous n'êtes pas autorisé à administrer le système. Vous ne savez pas qui le fait. Vous essayez de contacter le propriétaire du système qui a ouvert le dossier et il ne répond pas. Vous n'avez aucun contact pour le système et vos collègues ne le connaissent pas.

Avantages liés au respect de cette bonne pratique : En définissant des remontées, des déclencheurs de remontée et des procédures de remontée, vous permettez l'ajout systématique de ressources à un incident à un rythme adapté à l'impact.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Définir l'acheminement hiérarchique : définissez des chemins hiérarchiques dans vos runbooks et playbooks, y compris ce qui les déclenche et les procédures qui régissent cet acheminement. Par exemple, l'acheminement hiérarchique d'un problème des ingénieurs support aux ingénieurs support seniors lorsque les runbooks ne peuvent pas résoudre le problème, ou lorsqu'un laps de temps prédéfini s'est écoulé. Un autre exemple d'acheminement hiérarchique approprié est l'acheminement des ingénieurs support seniors à l'équipe de développement pour une charge de travail lorsque les playbooks ne sont pas en mesure d'identifier une méthode de correction, ou lorsqu'un laps de temps prédéfini s'est écoulé. Identifiez spécifiquement les propriétaires de chaque action afin de garantir des réponses efficaces et rapides aux événements liés aux opérations. Les acheminements hiérarchiques peuvent inclure des tiers. Par exemple, un fournisseur de connectivité réseau ou un fournisseur de logiciels. Les acheminements hiérarchiques peuvent inclure des décideurs autorisés identifiés pour les systèmes impactés.

OPS10-BP05 Définissez un plan de communication avec les clients en cas d'interruption de service

Définissez et testez un plan fiable de communication pour les pannes système pour tenir vos clients et vos parties prenantes informés pendant les pannes. Communiquez directement avec vos utilisateurs lorsque les services qu'ils utilisent sont touchés et lorsqu'ils reviennent à la normale.

Résultat souhaité :

- Vous disposez d'un plan de communication pour les situations allant de la maintenance programmée aux grandes pannes imprévues, y compris l'appel de plans de reprise après sinistre.
- Dans vos communications, vous fournissez des informations claires et transparentes sur les problèmes liés aux systèmes afin d'aider les clients à ne pas douter des performances de leurs systèmes.
- Vous utilisez des messages d'erreur personnalisés et des pages d'état pour réduire le pic des demandes au service d'assistance et tenir les utilisateurs informés.
- Le plan de communication est régulièrement testé pour vérifier qu'il fonctionnera comme prévu en cas de panne réelle.

Anti-modèles courants :

- Une interruption de la charge de travail survient mais vous n'avez aucun plan de communication. Les utilisateurs inondent votre système de tickets (TT) de demandes parce qu'ils n'ont aucune information sur la panne.
- Vous envoyez un e-mail de notification à vos utilisateurs pendant une panne. Il ne contient pas de calendrier de rétablissement du service, de sorte que les utilisateurs ne peuvent pas planifier en fonction de la panne.
- Il existe un plan de communication pour les pannes, mais il n'a jamais été testé. Une panne survient et le plan de communication échoue parce qu'une étape critique a été omise, alors qu'elle aurait pu être détectée lors des tests.
- Pendant une panne, vous envoyez une notification aux utilisateurs contenant trop de détails techniques et d'informations appartenant à votre accord de confidentialité AWS.

Avantages liés au respect de cette bonne pratique :

- Le maintien de la communication pendant les pannes garantit que les clients sont informés de l'état d'avancement des problèmes et du temps estimé pour les résoudre.
- L'élaboration d'un plan de communication bien défini permet de vérifier que vos clients et utilisateurs finaux sont bien informés afin qu'ils puissent prendre les mesures supplémentaires nécessaires pour atténuer l'impact des pannes.
- Grâce à une communication adéquate et à une sensibilisation accrue aux interruptions planifiées et non planifiées, vous pouvez améliorer la satisfaction des clients, limiter les réactions involontaires et favoriser la fidélisation des clients.
- Une communication opportune et transparente sur les pannes de système permet d'instaurer la confiance nécessaire au maintien des relations entre vous et vos clients.
- Une stratégie de communication éprouvée lors d'une panne ou d'une crise réduit les spéculations et les commérages qui pourraient entraver votre capacité de récupération.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les plans de communication qui permettent de tenir vos clients informés pendant les pannes sont complets et couvrent plusieurs interfaces, notamment les pages d'erreur destinées aux clients, les messages d'erreur API personnalisés, les bannières d'état du système et les pages d'état de santé.

Si votre système comprend des utilisateurs enregistrés, vous pouvez communiquer via des canaux de messagerie tels que l'e-mail, les SMS ou les notifications push pour envoyer des messages personnalisés à vos clients.

Outils de communication avec les clients

En guise de première ligne de défense, les applications web et mobiles doivent fournir des messages d'erreur conviviaux et informatifs pendant une panne et avoir la possibilité de rediriger le trafic vers une page d'état.Amazon CloudFront est un réseau de diffusion de contenu (CDN) entièrement géré qui permet de définir et de diffuser un contenu d'erreur personnalisé. Les pages d'erreur personnalisées dans CloudFront constituent une bonne première couche de messagerie client pour les pannes au niveau des composants. CloudFront peut également simplifier la gestion et l'activation d'une page d'état pour intercepter toutes les demandes pendant les pannes planifiées ou non.

Les messages d'erreur d'API personnalisés peuvent aider à détecter et à réduire l'impact lorsque les pannes sont limitées à des services distincts. Amazon API Gateway vous permet de configurer des réponses personnalisées pour vos API REST. Cela vous permet de fournir des messages clairs et pertinents aux consommateurs d'API lorsque API Gateway n'est pas en mesure d'atteindre les services dorsaux. Les messages personnalisés peuvent également être utilisés pour prendre en charge le contenu des bannières d'interruption de service et les notifications lorsqu'une fonctionnalité particulière du système est dégradée en raison d'interruptions de service.

La messagerie directe est le type de messagerie client le plus personnalisé. Amazon Pinpoint est un service géré pour des communications multicanales évolutives. Amazon Pinpoint vous permet de créer des campagnes qui peuvent diffuser des messages à grande échelle dans votre base de clients affectés par SMS, e-mail, voix, notifications push, ou des canaux personnalisés que vous définissez. Lorsque vous gérez la messagerie avec Amazon Pinpoint, les campagnes de messages sont bien définies, testables et peuvent être appliquées intelligemment à des segments de clientèle ciblés. Une fois établies, les campagnes peuvent être programmées ou déclenchées par des événements et elles peuvent facilement être testées.

Exemple de client

Lorsque la charge de travail est réduite, AnyCompany Retail envoie un e-mail de notification à ses utilisateurs. L'e-mail décrit les fonctionnalités de l'entreprise qui ont été affectées et donne une estimation réaliste de la date à laquelle le service sera rétabli. En outre, la société dispose d'une page d'état qui affiche des informations en temps réel sur l'état de sa charge de travail. Le plan de communication est testé dans un environnement de développement deux fois par an pour valider son efficacité.

Étapes d'implémentation

1. Déterminez les canaux de communication pour votre stratégie de messagerie. Tenez compte des aspects architecturaux de votre application et déterminez la meilleure stratégie pour fournir des commentaires à vos clients. Il peut s'agir d'une ou plusieurs des stratégies d'orientation décrites, notamment les pages d'erreur et d'état, les réponses personnalisées aux erreurs de l'API ou la messagerie directe.
2. Concevez des pages d'état pour votre application. Si vous avez déterminé que les pages d'état ou les pages d'erreur personnalisées conviennent à vos clients, vous devrez concevoir votre contenu et votre message pour ces pages. Les pages d'erreur expliquent aux utilisateurs pourquoi une application n'est pas disponible, quand elle pourrait le redevenir et ce qu'ils peuvent faire entre-temps. Si votre application utilise Amazon CloudFront, vous pouvez envoyer des réponses d'erreur personnalisées ou utiliser Lambda en périphérie pour traduire les erreurs et réécrire le contenu des pages. CloudFront permet également d'échanger les destinations du contenu de votre application contre une origine de contenu Amazon S3 statique contenant votre page d'état de maintenance ou d'interruption.
3. Concevez l'ensemble correct de statuts d'erreur API pour votre service. Les messages d'erreur produits par API Gateway lorsqu'il ne peut pas atteindre les services dorsaux, ainsi que les exceptions du niveau de service, peuvent ne pas contenir de messages conviviaux adaptés aux utilisateurs finaux. Sans avoir à modifier le code de vos services dorsaux, vous pouvez configurer des réponses d'erreur personnalisées API Gateway afin de faire correspondre les codes de réponse HTTP aux messages d'erreur de l'API.
4. Concevez les messages dans une optique commerciale afin qu'ils soient pertinents pour les utilisateurs finaux de votre système et ne contiennent pas de détails techniques. Tenez compte de votre public et alignez vos messages. Par exemple, vous pouvez orienter les utilisateurs internes vers une solution de contournement ou un processus manuel qui s'appuie sur d'autres systèmes. Les utilisateurs externes peuvent être invités à attendre que le système soit rétabli ou à s'inscrire aux mises à jour pour recevoir une notification lorsque le système sera rétabli. Définissez les messages approuvés pour de multiples scénarios, y compris les pannes inattendues, la maintenance planifiée et les pannes partielles du système où une fonction particulière peut être dégradée ou indisponible.
5. Modélisez et automatisez vos messages destinés aux clients. Une fois que vous avez préparé le contenu de votre message, vous pouvez utiliser Amazon Pinpoint ou d'autres outils pour automatiser votre campagne de messagerie. Avec Amazon Pinpoint, vous pouvez créer des segments cibles pour les utilisateurs concernés et transformer les messages en modèles.

Consultez le tutoriel Amazon Pinpoint pour comprendre comment configurer une campagne de messagerie.

6. Évitez de fortement coupler les capacités de messagerie à votre système de contact avec la clientèle. Votre stratégie de messagerie ne doit pas avoir de dépendance stricte vis-à-vis des magasins de données ou des services du système pour vérifier que vous pouvez envoyer des messages en cas de panne. Envisagez de créer la possibilité d'envoyer des messages à partir de plus d'une zone de disponibilité ou d'une région pour la disponibilité de la messagerie. Si vous utilisez des services AWS pour envoyer des messages, privilégiez les opérations du plan de données à celles du plan de contrôle pour appeler votre messagerie.

Niveau d'effort du plan d'implémentation : élevé. L'élaboration d'un plan de communication, et des mécanismes pour l'envoyer, peut demander un effort important.

Ressources

Bonnes pratiques associées :

- OPS07-BP03 Utiliser des runbooks pour effectuer des procédures - Votre plan de communication doit être associé à un cahier des charges afin que votre personnel sache comment réagir.
- OPS11-BP02 Effectuer une analyse post-incident - Après une panne, effectuez une analyse post-incident pour identifier les mécanismes permettant d'éviter une nouvelle panne.

Documents connexes :

- Modèles de gestion des erreurs dans Amazon API Gateway et AWS Lambda
- Réponses Amazon API Gateway

Exemples connexes :

- Tableau de bord AWS Health
- Summary of the AWS Service Event in the Northern Virginia (US-EAST-1) Region [Résumé de l'événement de service AWS dans la région de Virginie du Nord (US-EAST-1)]

Services associés :

- AWS Support

- Contrat Client AWS
- Amazon CloudFront
- Amazon API Gateway
- Amazon Pinpoint
- Amazon S3

OPS10-BP06 Communiquer l'état grâce aux tableaux de bord

Fournissez des tableaux de bord adaptés à leurs publics cibles (par exemple, équipes techniques internes, dirigeants et clients) pour communiquer l'état de fonctionnement actuel de l'entreprise et fournir des métriques d'intérêt.

Vous pouvez créer des tableaux de bord Amazon CloudWatch sur les pages d'accueil personnalisables de la console CloudWatch. Grâce aux services d'informatique décisionnelle tels que Amazon QuickSight vous pouvez créer et publier des tableaux de bord interactifs de votre charge de travail et de votre état opérationnel (par exemple, les taux de commande, les utilisateurs connectés et les heures de transaction). Créez des tableaux de bord qui présentent des affichages de vos métriques au niveau du système et de l'entreprise.

Anti-modèles courants :

- Sur demande, vous exécutez un rapport sur l'utilisation actuelle de votre application pour la gestion.
- Lors d'un incident, vous êtes contacté toutes les 20 minutes par un propriétaire de système concerné qui souhaite savoir si le problème est résolu.

Avantages liés au respect de cette bonne pratique : En créant des tableaux de bord, vous autorisez un accès en libre-service aux informations permettant à vos clients de s'informer et de déterminer s'ils doivent prendre des mesures.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Communiquer l'état grâce aux tableaux de bord : créez des tableaux de bord adaptés à leurs groupes cibles (par exemple, les équipes techniques internes, la direction et les clients) pour communiquer l'état d'exploitation actuel de l'entreprise et pour fournir des indicateurs d'intérêt.

Fournir une option en libre-service pour les informations d'état réduit l'impact disruptif lié aux demandes d'état de la part de l'équipe opérationnelle. Citons, par exemple, les tableaux de bord Amazon CloudWatch et AWS Health Dashboard.

- Les tableaux de bord CloudWatch créent et utilisent des vues de métriques personnalisées

Ressources

Documents connexes :

- Amazon QuickSight
- Les tableaux de bord CloudWatch créent et utilisent des vues de métriques personnalisées

OPS10-BP07 Automatiser les réponses aux événements

Automatisez les réponses aux événements pour réduire les erreurs causées par les processus manuels, et pour garantir des réponses rapides et cohérentes.

Il existe plusieurs façons d'automatiser les actions de runbooks et de playbooks sur AWS. Pour répondre à un événement à partir d'un changement d'état dans vos ressources AWS, ou à partir de vos propres événements personnalisés, vous devez créer des règles CloudWatch Events afin de déclencher des réponses via les cibles CloudWatch (par exemple, les fonctionsLambda, les rubriques Amazon Simple Notification Service (Amazon SNS), les tâches Amazon ECS et AWS Systems Manager Automation).

Pour répondre à une métrique qui dépasse un seuil pour une ressource (par exemple, le temps d'attente), vous devez créer des alarmes CloudWatch pour effectuer une ou plusieurs actions à l'aide des actions Amazon EC2, des actions Auto Scaling, ou pour envoyer une notification à une rubrique Amazon SNS. Si vous avez besoin d'effectuer des actions personnalisées en réponse à une alarme, appelez Lambda par le biais d'une notification Amazon SNS. Utilisez Amazon SNS pour publier des notifications d'événements et des messages de remontée pour que les personnes restent informées.

AWS prend également en charge les systèmes tiers via les API et les kits SDK de service AWS. Il existe divers outils fournis par les partenaires AWS et des tiers qui permettent la surveillance, les notifications et les réponses. Ces outils incluent notamment New Relic, Splunk, Loggly, SumoLogic et Datadog.

Vous devriez maintenir à disposition des procédures manuelles critiques pouvant être utilisées lorsque les procédures automatisées échouent.

Anti-modèles courants :

- Un développeur vérifie son code. Cet événement aurait pu être utilisé pour démarrer une génération, puis effectuer des tests, mais rien ne se passe.
- Votre application consigne une erreur spécifique avant de cesser de fonctionner. La procédure de redémarrage de l'application est bien comprise et peut être scriptée. Vous pouvez utiliser l'événement de journal pour appeler un script et redémarrer l'application. Au lieu de cela, lorsque l'erreur se produit à 3 h le dimanche matin, vous êtes réveillé en tant que ressource de garde chargée de résoudre le système.

Avantages liés au respect de cette bonne pratique : En utilisant des réponses automatisées aux événements, vous réduisez le temps de réponse et limitez l'introduction d'erreurs provenant d'activités manuelles.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Automatiser les réponses aux événements : automatisez les réponses aux événements pour réduire les erreurs causées par les processus manuels, et pour garantir des réponses rapides et cohérentes.
  - Qu'est-ce que Amazon CloudWatch Events ?
  - Création d'une règle CloudWatch Events qui se déclenche sur un événement
  - Création d'une règle CloudWatch Events qui se déclenche sur un appel d'API AWS avec AWS CloudTrail
  - Exemples d'événements CloudWatch Events tirés des services pris en charge

Ressources

Documents connexes :

- Fonctionnalités d'Amazon CloudWatch
- Exemples d'événements CloudWatch Events tirés des services pris en charge
- Création d'une règle CloudWatch Events qui se déclenche sur un appel d'API AWS avec AWS CloudTrail
- Création d'une règle CloudWatch Events qui se déclenche sur un événement

- Qu'est-ce que Amazon CloudWatch Events ?

Vidéos connexes :

- Élaborer un plan de surveillance

Exemples connexes :

Évolution

Question

- OPS 11. Comment faire évoluer vos opérations ?

OPS 11. Comment faire évoluer vos opérations ?

Consacrez du temps et des ressources à l'amélioration incrémentielle presque continue pour contribuer à l'évolution de l'efficacité et de l'efficience de vos opérations.

Bonnes pratiques

- OPS11-BP01 Définir un processus d'amélioration continue
- OPS11-BP02 Effectuer une analyse post-incident
- OPS11-BP03 Mettre en œuvre des boucles de rétroaction
- OPS11-BP04 Gérer les connaissances
- OPS11-BP05 Définir les facteurs d'amélioration
- OPS11-BP06 Valider les informations
- OPS11-BP07 Examiner les métriques des opérations
- OPS11-BP08 Documenter et partager des enseignements
- OPS11-BP09 Allouer du temps aux améliorations

OPS11-BP01 Définir un processus d'amélioration continue

Évaluez votre charge de travail par rapport aux bonnes pratiques d'architecture internes et externes. Vérifiez votre charge de travail au moins une fois par an. Priorisez les opportunités d'amélioration dans la cadence de développement de votre logiciel.

Résultat souhaité :

- Vous analysez votre charge de travail par rapport aux bonnes pratiques d'architecture au moins une fois par an.
- Les opportunités d'amélioration reçoivent la même priorité dans votre processus de développement logiciel.

Anti-modèles courants :

- Vous n'avez pas vérifié l'architecture de votre charge de travail depuis qu'elle a été déployée il y a plusieurs années.
- Les opportunités d'amélioration reçoivent une priorité moindre et restent dans la liste de suivi.
- Il n'existe aucune norme pour mettre en œuvre des modifications issues des bonnes pratiques pour l'organisation.

Avantages liés au respect de cette bonne pratique :

- Votre charge de travail est conforme aux bonnes pratiques d'architecture.
- L'évolution de votre charge de travail est réalisée de manière délibérée.
- Vous pouvez tirer profit des bonnes pratiques de l'organisation pour améliorer toutes les charges de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Vous vérifiez l'architecture de votre charge de travail au moins une fois par an. Évaluez votre charge de travail et identifier les opportunités d'amélioration grâce aux bonnes pratiques internes et externes. Priorisez les opportunités d'amélioration dans la cadence de développement de votre logiciel.

Exemple de client

Toutes les charges de travail d'AnyCompany Retail suivent un processus annuel de vérification de l'architecture. L'entreprise a développé sa propre liste de vérification des bonnes pratiques qui s'appliquent à toutes les charges de travail. Grâce à la fonction Custom Lens d'AWS Well-Architected Tool, elle réalise des vérifications avec l'outil et son filtre personnalisé des bonnes pratiques. Les opportunités d'amélioration générées à partir des vérifications sont prioritaires dans leurs sprints (échéances de vérification) logiciels.

Étapes d'implémentation

1. Réalisez des vérifications régulières de l'architecture de votre charge de travail de production au moins une fois par an. Utilisez une norme d'architecture documentée qui comprend des bonnes pratiques spécifiques à AWS.
   a. Nous vous recommandons d'utiliser vos propres normes définies en interne pour ces vérifications. Si vous n'avez pas de norme interne, nous vous recommandons d'utiliser le cadre AWS Well-Architected.
   b. Vous pouvez utiliser AWS Well-Architected Tool pour créer un filtre personnalisé de vos bonnes pratiques internes et vérifier votre architecture.
   c. Les clients peuvent contacter leur architecte de solutions AWS pour réaliser une vérification guidée grâce au cadre Well-Architected de leur charge de travail.
2. Priorisez les opportunités d'amélioration identifiées pendant la vérification au sein de votre processus de développement logiciel.

Niveau d'effort du plan d'implémentation : faible. Vous pouvez utiliser le cadre AWS Well-Architected pour réaliser la vérification annuelle de votre architecture.

Ressources

Bonnes pratiques associées :

- OPS11-BP02 Effectuer une analyse post-incident : l'analyse post-incident est un autre générateur d'éléments d'amélioration. Ajoutez les leçons apprises à votre liste interne de bonnes pratiques d'architecture.
- OPS11-BP08 Documenter et partager des enseignements : au fur et à mesure que vous développez vos bonnes pratiques d'architecture, partagez-les au sein de votre organisation.

Documents connexes :

- AWS Well-Architected Tool - Custom lenses (AWS Well-Architected Tool : filtres personnalisés)
- AWS Well-Architected Whitepaper : processus de vérification
- Customize Well-Architected Reviews using Custom Lenses and the AWS Well-Architected Tool (Personnaliser les vérification Well-Architected avec les filtres personnalisés et AWS Well-Architected Tool)

- Implementing the AWS Well-Architected Custom Lens lifecycle in your organization (Mise en œuvre du cycle de vie du filtre personnalisé AWS Well-Architected dans votre organisation)

Vidéos connexes :

- Well-Architected Labs - Level 100: Custom Lenses on AWS Well-Architected Tool (Ateliers Well-Architected : filtres personnalisés sur AWS Well-Architected Tool)

Exemples connexes :

- AWS Well-Architected Tool

OPS11-BP02 Effectuer une analyse post-incident

Examinez les événements ayant un impact sur les clients, et identifiez les facteurs contributifs et les actions préventives. Utilisez ces informations pour développer des mesures d'atténuation afin de limiter ou d’empêcher la récurrence. Développez des procédures pour fournir des réponses rapides et efficaces. Publiez, le cas échéant, les facteurs adjuvants et les mesures correctives adaptées au public ciblé.

Anti-modèles courants :

- Vous administrez un serveur d'applications. Toutes vos séances actives sont interrompues toutes les 23 heures et 55 minutes environ. Vous avez essayé d'identifier le problème sur votre serveur d'applications. Vous pensez qu'il pourrait s'agir d'un problème de réseau, mais vous ne pouvez pas obtenir la coopération de l'équipe réseau, car elle est trop occupée pour vous aider. Vous n'avez pas de processus prédéfini à suivre pour obtenir de l'aide et collecter les informations nécessaires pour déterminer ce qui se passe.
- Vous avez subi une perte de données au sein de votre charge de travail. C'est la première fois que cela se produit et la cause n'est pas évidente. Vous décidez que ce n'est pas important, car vous pouvez recréer les données. La perte de données se reproduit plus fréquemment en affectant vos clients. Cela vous impose également une charge opérationnelle supplémentaire lorsque vous restaurez les données manquantes.

Avantages liés au respect de cette bonne pratique : Le fait de disposer d'un processus prédéfini pour déterminer les composants, les conditions, les actions et les événements qui ont contribué à un incident vous permet d'identifier les possibilités d'amélioration.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Passez en revue tous les incidents ayant un impact sur le client. Dotez-vous d'un processus pour identifier et documenter les facteurs contributifs d'un incident afin de pouvoir mettre au point des mesures d'atténuation pour limiter ou empêcher la récurrence, et élaborez des procédures pour fournir des réponses rapides et efficaces. Indiquez la cause racine, si nécessaire, de manière appropriée et adaptée aux publics cibles.

OPS11-BP03 Mettre en œuvre des boucles de rétroaction

Les boucles de rétroaction fournissent des informations exploitables qui orientent la prise de décision. Créez des boucles de rétroaction dans vos procédures et vos charges de travail. Elles vous permettent d'identifier les problèmes et les points à améliorer. Elles valident également les investissements dans les améliorations. Ces boucles de rétroaction sont à la base de l'amélioration continue de votre charge de travail.

Il existe deux catégories de boucles de rétroaction : les rétroactions immédiates et les analyses rétrospectives. Les rétroactions immédiates sont collectées via l'examen des performances et des résultats des activités opérationnelles. Ces rétroactions proviennent des membres de l'équipe, des clients ou de la sortie automatisée de l'activité. Les rétroactions immédiates proviennent notamment de tests A/B et de la mise à disposition de nouvelles fonctionnalités, et il est essentiel d'échouer rapidement.

Les analyses rétrospectives doivent être effectuées régulièrement pour recueillir des rétroactions concernant l'évaluation des métriques et des résultats opérationnels au fil du temps. Ces analyses rétrospectives se déroulent à la fin d'un sprint, sur une cadence, ou après des versions ou des événements majeurs. Ce type de boucle de rétroaction valide les investissements dans les opérations ou votre charge de travail. Il vous permet de mesurer la réussite et valide votre stratégie.

Résultat souhaité : Les rétroactions immédiates et les analyses rétrospectives permettent d'apporter des améliorations. Il existe un mécanisme pour recueillir les rétroactions des utilisateurs et des membres de l'équipe. Les analyses rétrospectives sont utilisées pour déterminer les tendances qui entraînent des améliorations.

Anti-modèles courants :

- Vous lancez une nouvelle fonctionnalité, mais vous n'avez aucun moyen de recevoir les rétroactions des clients à ce sujet.

- Après avoir investi dans des améliorations opérationnelles, vous n'effectuez pas d'analyse rétrospective pour les valider.
- Vous recueillez les rétroactions des clients, mais ne les examinez pas régulièrement.
- Les boucles de rétroaction mènent à des mesures de suivi proposées, mais elles ne sont pas incluses dans le processus de développement de logiciels.
- Les clients ne reçoivent pas de rétroactions sur les améliorations qu'ils ont proposées.

Avantages liés au respect de cette bonne pratique :

- Vous pouvez travailler à rebours en partant du client pour générer de nouvelles fonctionnalités.
- Votre culture organisationnelle peut réagir plus rapidement face aux changements.
- Les tendances sont utilisées afin d'identifier des possibilités d'amélioration.
- Les analyses rétrospectives valident les investissements effectués dans votre charge de travail et vos opérations.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

L'implémentation de cette bonne pratique signifie que vous utilisez à la fois les rétroactions immédiates et les analyses rétrospectives. Ces boucles de rétroaction stimulent les améliorations. Il existe de nombreux mécanismes de rétroaction immédiate, notamment des enquêtes, des sondages auprès des clients ou des formulaires de rétroaction. Votre organisation utilise également des analyses rétrospectives afin d'identifier les possibilités d'amélioration et de valider les initiatives.

Exemple client

AnyCompany Retail a créé un formulaire web via lequel les clients peuvent transmettre une rétroaction ou signaler les problèmes. Au cours du scrum hebdomadaire, les rétroactions des utilisateurs sont évaluées par l'équipe de développement logiciel. Les rétroactions sont régulièrement utilisées pour orienter l'évolution de la plateforme de l'entreprise. Elle effectue une analyse rétrospective à la fin de chaque sprint afin d'identifier les éléments qu'elle souhaite améliorer.

Étapes d'implémentation

1. Rétroaction immédiate

- Vous avez besoin d'un mécanisme pour recevoir les rétroactions des clients et des membres de l'équipe. Vos activités opérationnelles peuvent également être configurées de façon à fournir une rétroaction automatisée.
- Votre organisation a besoin d'un processus pour examiner cette rétroaction, déterminer ce qui doit être amélioré et planifier l'amélioration.
- La rétroaction doit être ajoutée à votre processus de développement logiciel.
- Lorsque vous apportez des améliorations, effectuez un suivi auprès de l'auteur de la rétroaction.
  - Vous pouvez utiliser AWS Systems Manager OpsCenter pour créer et suivre ces améliorations en tant qu' OpsItems.

2. Analyse rétrospective

- Effectuez des analyses rétrospectives à la fin d'un cycle de développement, sur une cadence définie ou après une version majeure.
- Réunissez les parties prenantes impliquées dans la charge de travail pour une réunion rétrospective.
- Créez trois colonnes sur un tableau blanc ou une feuille de calcul : Arrêter, Commencer et Conserver.
  - La colonne Arrêter comportera tout ce que votre équipe doit arrêter de faire.
  - La colonne Commencer comportera tout ce que votre équipe doit commencer à faire.
  - La colonne Conserver comportera tout ce que vous souhaitez continuer à faire.
- Faites le tour de la salle et recueillez les rétroactions des parties prenantes.
- Privilégiez les rétroactions. Attribuez les actions et les parties prenantes aux points que vous souhaitez commencer ou conserver.
- Ajoutez les actions à votre processus de développement logiciel et communiquez les mises à jour de statut aux parties prenantes à mesure que vous apportez les améliorations.

Niveau d'effort du plan d'implémentation : moyen. Pour implémenter cette bonne pratique, vous avez besoin d'une solution pour recevoir une rétroaction immédiate et l'analyser. En outre, vous devez établir un processus d'analyse rétrospective.

Ressources

Bonnes pratiques associées :

- OPS01-BP01 Évaluer les besoins des clients externes : les boucles de rétroaction sont un mécanisme qui permet de recueillir les besoins des clients externes.
- OPS01-BP02 Évaluer les besoins des clients internes : les parties prenantes internes peuvent utiliser les boucles de rétroaction afin de communiquer les besoins et les exigences.
- OPS11-BP02 Effectuer une analyse post-incident : les analyses post-incident sont une forme importante d'analyse rétrospective menée après les incidents.
- OPS11-BP07 Examiner les métriques des opérations : les examens des métriques opérationnelles permettent d'identifier les tendances et les points à améliorer.

Documents connexes :

- 7 Pitfalls to Avoid When Building a CCOE
- Playbook de l'équipe Atlassian – Rétrospectives
- Email Definitions: Feedback Loops
- Establishing Feedback Loops Based on the AWS Well-Architected Framework Review
- IBM Garage Methodology - Hold a retrospective
- Investopedia – The PDCS Cycle
- Maximizing Developer Effectiveness de Tim Cochran
- Operations Readiness Reviews (ORR) Whitepaper - Iteration
- TIL CSI - Continual Service Improvement
- When Toyota met e-commerce: Lean at Amazon

Vidéos connexes :

- Building Effective Customer Feedback Loops

Exemples connexes :

- Astuto - Open source customer feedback tool
- AWS Solutions - QnABot on AWS
- Fider - A platform to organize customer feedback

Services associés :

- AWS Systems Manager OpsCenter

OPS11-BP04 Gérer les connaissances

La gestion des connaissances aide les membres de l'équipe à trouver les informations nécessaires à l'accomplissement de leur tâche. Dans les organisations qui fonctionnent selon le principe de l'apprentissage, les informations sont librement partagées, ce qui donne du pouvoir aux individus. Les informations peuvent être découvertes ou recherchées. Les informations sont exactes et à jour. Il existe des mécanismes permettant de générer de nouvelles informations, de mettre à jour les informations existantes et d'archiver les informations obsolètes. L'exemple le plus courant de plateforme de gestion des connaissances est un système de gestion de contenu comme un wiki.

Résultat souhaité :

- Les membres de l'équipe ont accès à des informations précises et opportunes.
- Les informations sont consultables.
- Il existe des mécanismes pour ajouter, mettre à jour et archiver les informations.

Anti-modèles courants :

- Il n'y a pas de stockage centralisé des connaissances. Les membres de l'équipe gèrent leurs propres notes sur leurs machines locales.
- Vous disposez d'un wiki auto-hébergé mais ne disposez d'aucun mécanisme de gestion des informations, ce qui se traduit par des informations obsolètes.
- Quelqu'un identifie des informations manquantes mais il n'existe aucun processus pour demander leur ajout dans le wiki de l'équipe. Les personnes l'ajoutent elles-mêmes mais manquent une étape clé, ce qui entraîne une panne.

Avantages liés au respect de cette bonne pratique :

- Les membres de l'équipe sont responsabilisés car les informations sont partagées librement.
- Les nouveaux membres de l'équipe sont intégrés plus rapidement car la documentation est à jour et consultable.
- Les informations sont opportunes, précises et exploitables.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

La gestion des connaissances est une facette importante des organisations qui fonctionnent selon le principe de l'apprentissage. Pour commencer, vous avez besoin d'un référentiel central pour stocker vos connaissances (par exemple, un wiki auto-hébergé). Vous devez développer des processus pour ajouter, mettre à jour et archiver les connaissances. Développez des normes pour ce qui doit être documenté et laissez chacune et chacun contribuer.

Exemple de client

AnyCompany Retail héberge un Wiki interne où toutes les connaissances sont stockées. Les membres de l'équipe sont encouragés à enrichir la base de connaissances dans l'exercice de leurs fonctions quotidiennes. Chaque trimestre, une équipe interfonctionnelle évalue les pages les moins mises à jour et détermine si elles doivent être archivées ou mises à jour.

Étapes d'implémentation

1. Commencez par identifier le système de gestion de contenu dans lequel les connaissances seront stockées. Obtenez l'accord des parties prenantes de votre organisation.
   a. Si vous ne disposez pas d'un système de gestion de contenu, envisagez d'utiliser un wiki hébergé par vos soins ou un référentiel de contrôle de version comme point de départ.
2. Développez des runbooks pour l'ajout, la mise à jour et l'archivage des informations. Formez votre équipe à ces processus.
3. Identifiez les connaissances qui doivent être stockées dans le système de gestion de contenu. Commencez par les activités quotidiennes (runbooks et playbooks) que les membres de l'équipe effectuent. Travaillez avec les parties prenantes pour prioriser les connaissances à ajouter.
4. Travaillez périodiquement avec les parties prenantes pour identifier les informations obsolètes et les archiver ou les mettre à jour.

Niveau d'effort du plan d'implémentation : moyen. Si vous ne disposez pas d'un système de gestion de contenu, vous pouvez mettre en place un wiki auto-hébergé ou un référentiel de documents contrôlé par version.

Ressources

Bonnes pratiques associées :

- OPS11-BP08 Documenter et partager des enseignements - La gestion des connaissances facilite le partage des informations sur les enseignements tirés.

Documents connexes :

- Atlassian : gestion des connaissances

Exemples connexes :

- DokuWiki
- Gollum
- MediaWiki
- Wiki.js

OPS11-BP05 Définir les facteurs d'amélioration

Identifiez les facteurs d'amélioration pour évaluer et prioriser les possibilités.

Sur AWS, vous pouvez regrouper les journaux de toutes vos activités opérationnelles, de vos charges de travail et de votre infrastructure pour créer un historique d'activité détaillé. Vous pouvez ensuite utiliser les outils AWS pour analyser l'état de vos opérations et de votre charge de travail au fil du temps (par exemple, identifier des tendances, mettre en corrélation des événements et des activités avec des résultats et comparer les environnements et les systèmes) pour identifier les possibilités d'amélioration en fonction de vos facteurs.

Vous devez utiliser CloudTrail pour suivre l'activité des API (via AWS Management Console, l'interface de ligne de commande (CLI), les kits SDK et les API) afin de savoir ce qu'il se passe sur l'ensemble de vos comptes. Suivez les activités de déploiement de vos outils pour développeurs AWS avec CloudTrail et CloudWatch. Ceci permet d'ajouter un historique d'activité détaillé de vos déploiements et de leurs résultats à vos données de journaux CloudWatch Logs.

Exportez vos données de journaux vers Amazon S3 pour un stockage à long terme. Avec AWS Glue, vous détectez et préparez vos données de journaux dans Amazon S3 à des fins d'analyse. Utilisez Amazon Athena, par le biais de son intégration native à AWS Glue, pour analyser vos données de journaux. Utilisez un outil d'informatique décisionnelle comme Amazon QuickSight pour visualiser, explorer et analyser vos données.

Anti-modèles courants :

- Vous disposez d'un script qui fonctionne, mais qui n'est pas élégant. Vous consacrez du temps à sa réécriture. Il s'agit désormais d'une œuvre d'art.

- Votre start-up essaie d'obtenir d’autres financements auprès d'un investisseur en capital-risque. Ils veulent que vous prouviez votre conformité à la norme PCI DSS. Vous voulez les contenter, et vous documentez votre conformité, mais manquez une date de livraison pour un client et perdez le client. Ce n'était pas une mauvaise chose à faire, mais maintenant vous vous demandez si c'était opportun.

Avantages liés au respect de cette bonne pratique : En déterminant les critères que vous voulez utiliser pour l'amélioration, vous pouvez minimiser l'impact des motivations liées aux événements ou de l'investissement émotionnel.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Comprendre les moteurs de l'amélioration : avant d'apporter des modifications à un système, il faut s'assurer que le résultat souhaité est bien pris en charge par celui-ci.
  - Fonctionnalités souhaitées : évaluez les fonctionnalités souhaitées lorsque vous étudiez les possibilités d'amélioration.
    - Nouveautés AWS
  - Problèmes inadmissibles : évaluez les problèmes inadmissibles, les bogues et les vulnérabilités lorsque vous étudiez les possibilités d'amélioration.
    - Derniers bulletins de sécurité AWS
    - AWS Trusted Advisor
  - Exigences de conformité : évaluez les mises à jour et les changements nécessaires pour assurer la conformité avec la réglementation ou une politique, ou pour continuer à bénéficier du soutien d'un tiers, lors de l'examen des possibilités d'amélioration.
    - Conformité AWS
    - Programmes de conformité AWS
    - Dernières actualités sur la conformité AWS

Ressources

Documents connexes :

- Amazon Athena
- Amazon QuickSight

- Conformité AWS
- Dernières actualités sur la conformité AWS
- Programmes de conformité AWS
- AWS Glue
- Derniers bulletins de sécurité AWS
- AWS Trusted Advisor
- Exportez vos données de journaux vers Amazon S3
- Nouveautés AWS

OPS11-BP06 Valider les informations

Vérifiez vos résultats d'analyse et les réponses avec les équipes interfonctionnelles et les responsables métier. Utilisez ces analyses pour établir la compréhension, identifier des impacts supplémentaires et déterminer des lignes de conduite. Ajustez les réponses si nécessaire.

Anti-modèles courants :

- Vous constatez que l'utilisation de la CPU est de 95 % sur un système et vous en faites une priorité pour trouver un moyen de réduire la charge sur le système. Vous déterminez que la meilleure action consiste à monter en charge. Le système est un transcodeur et dimensionné pour utiliser à 95 % en permanence la CPU. Le propriétaire du système aurait pu vous expliquer la situation si vous l'aviez contacté. Vous avez perdu du temps.
- Le propriétaire d'un système indique que son système est stratégique. Le système n'a pas été placé dans un environnement hautement sécurisé. Pour améliorer la sécurité, vous mettez en œuvre les contrôles de détection et de prévention supplémentaires requis pour les systèmes stratégiques. Vous informez le propriétaire du système que le travail est terminé et que les ressources supplémentaires lui seront facturées. Dans la discussion qui suit cette notification, le propriétaire du système apprend qu'il existe une définition formelle des systèmes stratégiques qui ne s'applique pas à son système.

Avantages liés au respect de cette bonne pratique : En validant les informations avec les propriétaires d'entreprises et les experts du domaine, vous pouvez établir une compréhension commune et orienter plus efficacement les améliorations.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Valider les informations : collaborez avec les propriétaires d'entreprise et les experts du domaine pour vous assurer qu'il existe une compréhension et un accord communs sur la signification des données que vous avez recueillies. Identifiez les autres préoccupations, les impacts potentiels et déterminez les mesures à prendre.

OPS11-BP07 Examiner les métriques des opérations

Régulièrement, faites des analyses rétrospectives des métriques opérationnelles avec des intervenants provenant de différents services de l'entreprise. Utilisez ces examens pour identifier les possibilités d'amélioration, les pistes d'action potentielles et pour partager les enseignements tirés.

Recherchez des opportunités d'amélioration dans l'ensemble de vos environnements (par exemple, le développement, le test et la production).

Anti-modèles courants :

- Une promotion de vente au détail importante a été interrompue par votre fenêtre de maintenance. L'entreprise continue d’ignorer qu'il existe une fenêtre de maintenance standard qui peut être retardée si d'autres événements ont un impact sur l'activité.
- Vous avez subi une panne prolongée en raison de votre utilisation d'une bibliothèque défaillante couramment utilisée dans votre organisation. Depuis, vous avez migré vers une bibliothèque fiable. Les autres équipes de votre organisation ne savent pas qu'elles sont exposées à des risques. Si vous vous réunissiez régulièrement et examiniez cet incident, elles seraient averties des risques.
- Les performances de votre transcodeur n'ont cessé de diminuer, ce qui a eu un impact sur l'équipe multimédia. Ce n'est pas encore catastrophique. Vous ne pourrez le savoir que quand la situation se sera suffisamment dégradée au point de provoquer un incident. Si vous examiniez vos métriques opérationnelles avec l'équipe multimédia, il serait possible d'identifier le changement dans les métriques et sa situation, et vous pourriez traiter le problème.
- Vous ne passez pas en revue votre respect des SLA des clients. Vous avez tendance à ne pas respecter les SLA de vos clients. Des pénalités financières sont liées au non-respect des SLA de vos clients. Si vous vous réunissiez régulièrement pour examiner les métriques de ces SLA, vous pourriez identifier et résoudre le problème.

Avantages liés au respect de cette bonne pratique : En vous réunissant régulièrement pour examiner les métriques des opérations, les événements et les incidents, vous maintenez une

compréhension commune entre les équipes, partagez les leçons apprises et pouvez prioriser et cibler les améliorations.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Examiner les métriques des opérations : effectuez régulièrement des analyses rétrospectives des métriques opérationnelles avec des intervenants provenant de différentes équipes et de différents services de l'entreprise. Faites appel à différents intervenants, y compris des membres de l'équipe commerciale, de l'équipe de développement et de l'équipe opérationnelle, pour qu'ils valident vos résultats par l'intermédiaire de rétroactions immédiates et d'analyses rétrospectives et pour partager les leçons apprises. Utilisez leurs informations pour identifier les possibilités d'amélioration et les plans d'action possibles.
  - Amazon CloudWatch
  - Utilisation des métriques Amazon CloudWatch
  - Publier des métriques personnalisées
  - Référence aux dimensions et métriques Amazon CloudWatch

Ressources

Documents connexes :

- Amazon CloudWatch
- Référence aux dimensions et métriques Amazon CloudWatch
- Publier des métriques personnalisées
- Utilisation des métriques Amazon CloudWatch

OPS11-BP08 Documenter et partager des enseignements

Documenter et partager des enseignements : documentez et partagez les enseignements que vous tirez des activités opérationnelles afin de pouvoir les utiliser en interne et entre les équipes.

Vous devriez partager les enseignements tirés par vos équipes afin d'en retirer un bénéfice accru pour toute votre organisation. Vous devriez partager des informations et des ressources pour prévenir les erreurs évitables et faciliter les efforts de développement. Cela vous permettra de vous concentrer sur la publication des fonctionnalités souhaitées.

Utilisez AWS Identity and Access Management (IAM) pour définir les autorisations permettant de contrôler l'accès aux ressources que vous souhaitez partager au sein des comptes et entre les comptes. Vous devez ensuite utiliser des référentiels AWS CodeCommit avec contrôle de version pour partager des bibliothèques d'application, des procédures scriptées, de la documentation de procédure et d'autres documentations systèmes. Partagez vos normes de calcul en partageant l'accès à vos AMI et en autorisant l'utilisation de vos fonctions Lambda entre les comptes. Vous devez également partager vos normes d'infrastructure comme modèles AWS CloudFormation.

Grâce aux API et aux kits SDK AWS, vous pouvez intégrer des outils et référentiels tiers et externes (par exemple, GitHub, BitBucket et SourceForge). Lorsque vous partagez ce que vous avez appris et développé, veillez à structurer les autorisations de manière à garantir l'intégrité des référentiels partagés.

Anti-modèles courants :

- Vous avez subi une panne prolongée en raison de votre utilisation d'une bibliothèque défaillante couramment utilisée dans votre organisation. Depuis, vous avez migré vers une bibliothèque fiable. Les autres équipes de votre organisation ne savent pas qu'elles sont exposées à des risques. Si vous documentiez et partagiez votre expérience concernant cette bibliothèque, elles seraient avertis des risques.
- Vous avez identifié un cas limite dans un microservice partagé en interne qui entraîne l'abandon des séances. Vous avez mis à jour vos appels au service pour éviter ce cas limite. Les autres équipes de votre organisation ne savent pas qu'elles sont exposées à des risques. Si vous documentiez et partagiez votre expérience concernant cette bibliothèque, elles seraient avertis des risques.
- Vous avez trouvé un moyen de réduire considérablement les besoins d'utilisation du processeur pour l'un de vos microservices. Vous ne savez pas si d'autres équipes peuvent tirer parti de cette technique. Si vous documentiez et partagiez votre expérience concernant cette bibliothèque, elles pourraient le faire.

Avantages liés au respect de cette bonne pratique : Partagez les enseignements que vous avez tirés pour soutenir l'amélioration et pour optimiser les bénéfices de l'expérience.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Documenter et partager des enseignements : mettez en place des procédures pour documenter les enseignements que vous tirez de l'exécution des activités opérationnelles et des analyses rétrospectives, afin que ceux-ci puissent être utilisés par d'autres équipes.
  - Partager les enseignements : imaginez des procédures permettant de partager ces enseignements, ainsi que les artefacts qui y sont associés, avec les autres équipes. Partagez par exemple les mises à jour concernant les procédures, les conseils, la gouvernance et les bonnes pratiques par l'intermédiaire d'un wiki accessible. Partagez des scripts, du code et des bibliothèques grâce à un référentiel commun.
    - Déléguer l'accès à votre environnement AWS
    - Partager un référentiel AWS CodeCommit
    - Facilité d'autorisation des fonctions AWS Lambda
    - Partager une AMI avec des comptes AWS spécifiques
    - Accélérez le partage de modèles avec une URL de concepteur AWS CloudFormation
    - Utilisation d'AWS Lambda avec Amazon SNS

Ressources

Documents connexes :

- Facilité d'autorisation des fonctions AWS Lambda
- Partager un référentiel AWS CodeCommit
- Partager une AMI avec des comptes AWS spécifiques
- Accélérez le partage de modèles avec une URL de concepteur AWS CloudFormation
- Utilisation d'AWS Lambda avec Amazon SNS

Vidéos connexes :

- Déléguer l'accès à votre environnement AWS

OPS11-BP09 Allouer du temps aux améliorations

Consacrez du temps et des ressources à vos processus pour permettre des améliorations progressives continues.

Sur AWS, vous pouvez créer des copies temporaires d'environnements, ce qui permet de réduire le risque, les efforts, et les coûts d'expérimentation et de test. Ces copies d'environnements peuvent être utilisées pour tester les conclusions de votre analyse, expérimenter, et développer et tester des améliorations planifiées.

Anti-modèles courants :

- Il existe un problème de performances connu sur votre serveur d'applications. Il s'ajoute au retard accumulé dans la mise en œuvre de chaque fonctionnalité planifiée. Si le rythme d'ajout des fonctionnalités prévues reste constant, la question des performances ne sera jamais abordée.
- Pour permettre l'amélioration continue, vous autorisez les administrateurs et les développeurs à utiliser tout leur temps supplémentaire pour sélectionner et mettre en œuvre les améliorations. Aucune amélioration n'est effectuée.

Avantages liés au respect de cette bonne pratique : Ainsi, vous permettez d'apporter des améliorations progressives continues.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Allouer du temps aux améliorations : ce sont les ressources et le temps consacrés à vos processus qui vous permettent d'apporter des améliorations incrémentielles continues. Mettez en œuvre des modifications afin d'améliorer et d'évaluer les résultats, mais également de déterminer le taux de réussite qu'ils représentent. Si les résultats sont en deçà des objectifs et que l'amélioration constitue toujours une priorité, exécutez d'autres plans d'action.

Sécurité

Le pilier Sécurité présente la capacité de protéger les données ainsi que les systèmes et les ressources pour tirer parti des technologies du cloud et améliorer votre sécurité. Vous trouverez des recommandations sur l'implémentation dans le Livre blanc du pilier Sécurité.

Domaines de bonnes pratiques

- Bases du pilier Sécurité
- Identity and Access Management
- Détection

- Protection de l'infrastructure
- Protection des données
- Réponse aux incidents
- Sécurité des applications

Bases du pilier Sécurité

Question

- SÉC 1. Comment gérer votre charge de travail en toute sécurité ?

SÉC 1. Comment gérer votre charge de travail en toute sécurité ?

Vous devez appliquer de bonnes pratiques générales à chaque domaine de la sécurité pour réussir à gérer votre charge de travail en toute sécurité. Appliquez à tous les domaines les conditions et les processus que vous avez définis en matière d'excellence opérationnelle au niveau de l'organisation et de la charge de travail. La connaissance des recommandations actuelles d'AWS et du secteur ainsi que des renseignements sur les menaces vous aide à faire évoluer votre modèle de menace et vos objectifs de contrôle. L'automatisation des processus de sécurité, des tests et de la validation vous permet de mettre à l'échelle vos opérations de sécurité.

Bonnes pratiques

- SEC01-BP01 Séparer les charges de travail à l'aide de comptes
- SEC01-BP02 Sécuriser l'utilisateur root et les propriétés du compte
- SEC01-BP03 Identifier et valider les objectifs de contrôle
- SEC01-BP04 Rester informé des menaces de sécurité
- SEC01-BP05 Connaître les recommandations de sécurité
- SEC01-BP06 Automatiser les tests et la validation des contrôles de sécurité dans les pipelines
- SEC01-BP07 Identifier les menaces et hiérarchiser les atténuations à l'aide d'un modèle de menaces
- SEC01-BP08 Évaluer et mettre en œuvre régulièrement de nouveaux services et de nouvelles fonctions de sécurité

SEC01-BP01 Séparer les charges de travail à l'aide de comptes

Établissez des barrières de protection et un isolement communs entre les environnements (par exemple, production, développement et test) et les charges de travail grâce à une stratégie multicompte. La séparation au niveau des comptes est vivement recommandée, car elle fournit une solide limite d'isolement pour la sécurité, la facturation et les accès.

Résultat souhaité : une structure de compte qui isole les opérations cloud, les charges de travail non liées et les environnements dans des comptes séparés, ce qui permet de renforcer la sécurité dans l'infrastructure cloud.

Anti-modèles courants :

- Placer plusieurs charges de travail non liées avec différents niveaux de sensibilité des données dans le même compte.
- Structure d'unité d'organisation mal définie.

Avantages liés à l'instauration de cette bonne pratique :

- Réduction de la portée des répercussions si un utilisateur accède à une charge de travail par inadvertance.
- Gouvernance centralisée des services, ressources et régions AWS.
- Maintien de la sécurité de l'infrastructure cloud avec des politiques et une administration centralisée des services de sécurité.
- Processus automatisé de création et de gestion des comptes.
- Audit centralisé de votre infrastructure pour les exigences en matière de conformité et de réglementation.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

Les Comptes AWS établissent une limite d'isolement de sécurité entre les charges de travail ou les ressources qui fonctionnent à différents niveaux de sensibilité. AWS fournit des outils permettant de gérer vos charges de travail cloud à grande échelle grâce à une stratégie multicompte pour tirer parti de cette limite d'isolement. Pour obtenir des conseils sur les concepts, les modèles et l'implémentation d'une stratégie multicompte sur AWS, consultez Organizing Your AWS Environment Using Multiple Accounts (Organisation de votre environnement AWS à l'aide de plusieurs comptes).

Lorsque plusieurs Comptes AWS sont gérés de façon centralisée, ils doivent être organisés selon une hiérarchie définie par des couches d'unités d'organisation. Les contrôles de sécurité peuvent ensuite être organisés et appliqués aux unités d'organisation et aux comptes membres, ce qui permet d'établir des contrôles préventifs uniformes sur les comptes membres au sein de l'organisation. Les contrôles de sécurité sont hérités, vous pouvez donc filtrer les autorisations disponibles pour les comptes membres situés aux niveaux inférieurs d'une hiérarchie d'unités d'organisation. Une bonne conception tire parti de cet héritage pour réduire le nombre et la complexité des politiques de sécurité nécessaires afin de mettre en place les contrôles de sécurité souhaités pour chaque compte membre.

Les services AWS Organizations et AWS Control Tower peuvent être utilisés pour implémenter et gérer cette structure multicompte dans votre environnement AWS. AWS Organizations vous permet d'organiser les comptes dans une hiérarchie définie par une ou plusieurs couches d'unités d'organisation, chacune de ces dernières contenant un certain nombre de comptes membres. Les politiques de contrôle des services (SCP) permettent à l'administrateur de l'organisation d'établir des contrôles préventifs granulaires sur les comptes membres et AWS Config peut être utilisé pour établir des contrôles proactifs et de détection sur les comptes membres. Un grand nombre de services AWS s'intègrent à AWS Organizations pour fournir des contrôles administratifs délégués et effectuer des tâches propres aux services dans tous les comptes membres de l'organisation.

Ajouté au-dessus d'AWS Organizations, AWS Control Tower fournit une configuration en un clic des bonnes pratiques pour un environnement AWS multicompte avec une zone de destination. La zone de destination est le point d'entrée de l'environnement multicompte établi par Control Tower. Control Tower offre plusieurs avantages par rapport à AWS Organizations. Les trois avantages qui permettent d'améliorer la gouvernance des comptes sont les suivants :

- Des barrières de protection obligatoires intégrées qui sont automatiquement appliquées aux comptes admis dans l'organisation.
- Des barrières de protection facultatives qui peuvent être activées ou désactivées pour un ensemble donné d'unités d'organisation.
- AWS Control Tower Account Factory fournit un déploiement automatisé des comptes contenant des bases de référence préapprouvées et des options de configuration au sein de votre organisation.

Étapes d'implémentation

1. Concevez une structure d'unités d'organisation : une structure d'unités d'organisation bien conçue réduit la charge de gestion liée à la création et à l'application des politiques de contrôle des

services et d'autres contrôles de sécurité. Votre structure d'unités d'organisation doit être alignée sur les besoins opérationnels, la sensibilité des données et la structure des charges de travail.

2. Créez une zone de destination pour votre environnement multicompte : une zone de destination fournit une base cohérente de sécurité et d'infrastructure à partir de laquelle votre organisation peut rapidement développer, lancer et déployer des charges de travail. Vous pouvez utiliser une zone de destination personnalisée ou AWS Control Tower pour orchestrer votre environnement.
3. Établissez des barrières de protection : implémentez des barrières de protection de sécurité uniformes pour votre environnement dans votre zone de destination. AWS Control Tower fournit une liste de contrôles obligatoires et facultatifs qui peuvent être déployés. Les contrôles obligatoires sont déployés automatiquement lors de l'implémentation de Control Tower. Passez en revue la liste des contrôles hautement recommandés et facultatifs, puis implémentez les contrôles adaptés à vos besoins.
4. Limitez l'accès aux régions qui viennent d'être ajoutées : pour les nouvelles Régions AWS, les ressources IAM telles que les utilisateurs et les rôles sont uniquement propagées vers les régions que vous spécifiez. Cette action peut être effectuée via la console lorsque vous utilisez Control Tower ou en modifiant les politiques d'autorisations IAM dans AWS Organizations.
5. Envisagez l'utilisation d'AWS CloudFormation StackSets : les StackSets vous permettent de déployer des ressources, dont les politiques, rôles et groupes IAM dans différentes régions et différents Comptes AWS à partir d'un modèle approuvé.

Ressources

Bonnes pratiques associées :

- SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé

Documents connexes :

- AWS Control Tower
- AWS Security Audit Guidelines (Consignes pour les audits de sécurité AWS)
- Bonnes pratiques dans IAM
- Use CloudFormation StackSets to provision resources across multiple Comptes AWS and regions (Utiliser CloudFormation StackSets pour provisionner les ressources sur plusieurs comptes et régions AWS)
- FAQ sur AWS Organizations

- Terminologie et concepts relatifs à AWS Organizations
- Best Practices for Service Control Policies in an AWS Organizations Multi-Account Environment (Bonnes pratiques pour les politiques de contrôle des services d'AWS Organizations dans un environnement multicompte)
- AWS Account Management Reference Guide (Guide de référence de la gestion des comptes AWS)
- Organizing Your AWS Environment Using Multiple Accounts (Organisation de votre environnement AWS à l'aide de plusieurs comptes)

Vidéos connexes :

- Enable AWS adoption at scale with automation and governance
- Security Best Practices the Well-Architected Way
- Building and Governing Multiple Accounts using AWS Control Tower
- Enable Control Tower for Existing Organizations

Ateliers connexes :

- Journée d'immersion Control Tower

SEC01-BP02 Sécuriser l'utilisateur root et les propriétés du compte

L'utilisateur root est celui qui dispose du plus de privilèges dans un Compte AWS, avec un accès administratif complet à toutes les ressources du compte. De plus, dans certains cas, il ne peut pas être limité par les politiques de sécurité. Si vous désactivez l'accès par programmation pour l'utilisateur root, établissez des contrôles appropriés pour l'utilisateur root et évitez l'utilisation de routine de l'utilisateur root, vous réduirez le risque d'exposition accidentelle des informations d'identification root et de compromission ultérieure de l'environnement cloud.

Résultat souhaité : la sécurisation de l'utilisateur root permet de réduire les risques de dommages accidentels ou intentionnels en raison de l'utilisation inappropriée des informations d'identification de l'utilisateur root. La mise en place de contrôles de détection permet également d'alerter le personnel approprié lorsque des mesures sont prises à l'aide de l'utilisateur root.

Anti-modèles courants :

- Se servir de l'utilisateur root pour des tâches autres que celles nécessitant des informations d'identification de l'utilisateur root.
- Omettre de tester régulièrement des plans d'urgence pour vérifier le fonctionnement de l'infrastructure, des processus et du personnel essentiels dans les situations d'urgence.
- Ne tenir compte que du flux de connexion type du compte et omettre d'envisager ou de tester d'autres méthodes de récupération de compte.
- Ne pas gérer les DNS, les serveurs de messagerie et les fournisseurs de services téléphoniques dans le cadre du périmètre de sécurité critique, car ils sont utilisés dans le flux de récupération de compte.

Avantages liés à l'instauration de cette bonne pratique : la sécurisation de l'accès à l'utilisateur root permet de garantir le contrôle et la vérification des actions effectuées dans votre compte.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

AWS propose plusieurs outils afin de vous aider à sécuriser votre compte. Toutefois, étant donné que certaines de ces mesures ne sont pas activées par défaut, vous devez intervenir directement pour les implémenter. Considérez ces recommandations comme des étapes fondamentales pour sécuriser votre Compte AWS. À mesure que vous mettez en œuvre ces étapes, il est important d'établir un processus permettant d'évaluer et de surveiller continuellement les contrôles de sécurité.

Lorsque vous créez un compte Compte AWS, vous commencez avec une seule identité disposant d'un accès complet à toutes les ressources et à tous les services AWS de ce compte. Cette identité est appelée l'utilisateur root du Compte AWS. Vous pouvez vous connecter en tant qu'utilisateur root en utilisant l'adresse e-mail et le mot de passe utilisés pour créer le compte. En raison de l'accès élevé accordé à l'utilisateur root AWS, vous devez limiter l'utilisation de cet utilisateur root AWS à l'exécution de tâches qui en ont spécifiquement besoin. Les informations d'identification de l'utilisateur root doivent être étroitement protégées et l'authentification multifactorielle (MFA) doit toujours être activée pour l'utilisateur root du Compte AWS.

Outre le flux d'authentification normal pour vous connecter à votre utilisateur root en utilisant un nom d'utilisateur, un mot de passe et un dispositif d'authentification multifactorielle (MFA), il y a des flux de récupération de compte pour vous connecter à l'utilisateur root de votre Compte AWS, à condition de disposer d'un accès à l'adresse e-mail et au numéro de téléphone associés à votre compte. Par conséquent, il est tout aussi important de sécuriser le compte de messagerie de l'utilisateur root là

où l'e-mail de récupération est envoyé, ainsi que le numéro de téléphone associé au compte. Il est également nécessaire de tenir compte des dépendances circulaires possibles lorsque l'adresse e-mail associée à l'utilisateur root est hébergée sur des serveurs de messagerie ou des ressources du service de noms de domaine (DNS) à partir du même Compte AWS.

Lorsque vous utilisez AWS Organizations, il y a plusieurs Comptes AWS, chacun d'entre eux ayant un utilisateur root. Un compte est désigné comme compte de gestion et plusieurs couches de comptes membres peuvent alors être ajoutées sous le compte de gestion. Privilégiez la sécurisation de l'utilisateur root de votre compte de gestion, puis occupez-vous des utilisateurs root des comptes membres. La stratégie de sécurisation de l'utilisateur root de votre compte de gestion peut différer de celle des utilisateurs root des comptes membres et vous pouvez placer des contrôles de sécurité préventifs sur les utilisateurs root des comptes membres.

Étapes d'implémentation

Les étapes d'implémentation suivantes sont recommandées afin d'établir des contrôles pour l'utilisateur root. Le cas échéant, les recommandations comportent des renvois vers les contrôles de référence CIS AWS Foundations version 1.4.0. Outre ces étapes, consultez Consignes en matière de bonnes pratiques avec AWS pour sécuriser vos ressources et votre Compte AWS.

Contrôles préventifs

1. Configurez des coordonnées exactes pour le compte.
    a. Ces informations sont utilisées pour le flux de récupération de mot de passe perdu, le flux de récupération de compte d'authentification multifactorielle perdu et pour les communications critiques liées à la sécurité avec votre équipe.
    b. Utilisez une adresse e-mail hébergée par votre domaine d'entreprise, de préférence une liste de distribution, comme adresse e-mail de l'utilisateur root. L'utilisation d'une liste de distribution plutôt que d'un compte de messagerie individuel fournit une redondance et une continuité supplémentaires pour l'accès au compte root sur de longues périodes.
    c. Le numéro de téléphone indiqué pour les coordonnées doit correspondre à un téléphone dédié et sécurisé à cette fin. Ce numéro de téléphone ne doit figurer sur aucune liste ni être communiqué à personne.
2. Ne créez pas de clés d'accès pour l'utilisateur root. Si des clés d'accès existent, retirez-les (CIS 1.4).
    a. Éliminez les informations d'identification par programmation de longue durée (clés d'accès et secrètes) pour l'utilisateur root.

   b. S'il existe déjà des clés d'accès pour l'utilisateur root, vous devez effectuer la transition des processus en utilisant ces clés afin de vous servir de clés d'accès temporaires issues d'un rôle AWS Identity and Access Management (IAM), puis supprimer les clés d'accès de l'utilisateur root.

3. Déterminez si vous devez stocker les informations d'identification de l'utilisateur root.
   a. Si vous utilisez AWS Organizations pour créer de nouveaux comptes membres, le mot de passe initial pour l'utilisateur root sur ces nouveaux comptes est une valeur aléatoire à laquelle vous n'avez pas accès. Envisagez d'utiliser le flux de réinitialisation du mot de passe à partir de votre compte de gestion AWS Organization pour accéder au compte membre si nécessaire.
   b. Pour les Comptes AWS autonomes ou le compte de gestion AWS Organization, envisagez de créer et de stocker en toute sécurité les informations d'identification de l'utilisateur root. Activez l'authentification multifactorielle pour l'utilisateur root.
4. Activez les contrôles préventifs pour les utilisateurs root des comptes membres dans les environnements AWS multicomptes.
   a. Envisagez d'activer la barrière de protection préventive Désactiver la création des clés d'accès root pour l'utilisateur root pour les comptes membres.
   b. Envisagez d'activer la barrière de protection préventive Désactiver les actions en tant qu'utilisateur root pour les comptes membres.
5. Si vous avez besoin d'informations d'identification pour l'utilisateur root :
   a. Utilisez un mot de passe complexe.
   b. Activez l'authentification multifactorielle (MFA) pour l'utilisateur root, plus particulièrement pour les comptes de gestion (payeur) AWS Organizations (CIS 1.5).
   c. Envisagez l'utilisation des appareils d'authentification multifactorielle pour la résilience et la sécurité, car les appareils à usage unique peuvent réduire les risques de réutilisation des appareils contenant vos codes d'authentification multifactorielle à d'autres fins. Vérifiez que les appareils d'authentification multifactorielle alimentés par une batterie sont remplacés régulièrement. (CIS 1.6)
      - Si vous souhaitez configurer l'authentification multifactorielle pour l'utilisateur root, suivez les instructions d'activation d'une authentification multifactorielle virtuelle ou d'un appareil d'authentification multifactorielle.
   d. Envisagez d'inscrire plusieurs appareils d'authentification multifactorielle pour la sauvegarde. Jusqu'à 8 appareils d'authentification multifactorielle sont autorisés par compte.

- Notez que l'inscription de plusieurs appareils d'authentification multifactorielle pour l'utilisateur root désactive automatiquement le flux de récupération de votre compte si l'appareil d'authentification multifactorielle est perdu.

e. Stockez le mot de passe en sécurité et tenez compte des dépendances circulaires si vous le stockez électroniquement. Ne stockez pas le mot de passe de manière à ce qu'il nécessite un accès au même Compte AWS pour l'obtenir.

6. Facultatif : envisagez d'établir un calendrier périodique de rotation des mots de passe pour l'utilisateur root.
   - Les bonnes pratiques relatives à la gestion des informations d'identification dépendent de vos exigences en matière de réglementation et de politiques. Les utilisateurs root protégés par l'authentification multifactorielle ne dépendent pas du mot de passe comme facteur d'authentification unique.
   - La modification périodique du mot de passe de l'utilisateur root réduit le risque d'utilisation inappropriée d'un mot de passe exposé par inadvertance.

Contrôles de détection

- Créez des alarmes pour détecter l'utilisation des informations d'identification root (CIS 1.7). L'activation d'Amazon GuardDuty permettra de surveiller et d'alerter sur l'utilisation des informations d'identification de l'API de l'utilisateur root via la recherche RootCredentialUsage.
- Évaluez et implémentez les contrôles de détection inclus dans le pack de conformité du pilier Sécurité AWS Well-Architected pour AWS Config ou, si vous utilisez AWS Control Tower, les contrôles vivement recommandés disponibles dans Control Tower.

Conseils opérationnels

- Déterminez qui, au sein de l'organisation, doit avoir accès aux informations d'identification de l'utilisateur root.
  - Utilisez la règle des deux personnes pour éviter qu'une seule personne ait accès à toutes les informations d'identification et à l'authentification multifactorielle nécessaires pour obtenir l'accès à l'utilisateur root.
  - Vérifiez que l'organisation, et non une seule personne, conserve le contrôle du numéro de téléphone et de l'alias d'e-mail associés au compte (qui sont utilisés pour la réinitialisation du mot de passe et l'authentification multifactorielle).
- Utilisez l'utilisateur root uniquement de façon exceptionnelle (CIS 1.7).

  - L'utilisateur root AWS ne doit pas être employé pour des tâches quotidiennes, même les tâches d'administration. Connectez-vous en tant qu'utilisateur root uniquement pour effectuer des tâches AWS qui requièrent l'utilisateur root. Toutes les autres actions doivent être effectuées par d'autres utilisateurs assumant les rôles appropriés.
- Vérifiez régulièrement que l'accès à l'utilisateur root fonctionne afin que les procédures soient testées avant une situation d'urgence nécessitant l'utilisation des informations d'identification de l'utilisateur root.
- Vérifiez régulièrement que l'adresse e-mail associée au compte et les adresses répertoriées sous Autres contacts fonctionnent. Vérifiez dans ces boîtes de réception si vous avez reçu des notifications de sécurité de la part de <abuse@amazon.com>. Assurez-vous également que les numéros de téléphone associés au compte fonctionnent.
- Préparez les procédures d'intervention en cas d'incident pour réagir face à une utilisation inappropriée du compte root. Consultez le guide AWS Security Incident Response Guide (Guide d'intervention en cas d'incident de sécurité) et les bonnes pratiques dans la section sur le pilier Sécurité du livre blanc consacré aux réponses face aux incidents pour plus d'informations sur l'élaboration d'une stratégie de réponse face aux incidents pour votre Compte AWS.

Ressources

Bonnes pratiques associées :

- SEC01-BP01 Séparer les charges de travail à l'aide de comptes
- SEC02-BP01 Utiliser de solides mécanismes d'authentification
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège
- SEC03-BP03 Établir un processus d'accès d'urgence
- SEC10-BP05 Préallouer les accès

Documents connexes :

- AWS Control Tower
- AWS Security Audit Guidelines (Consignes pour les audits de sécurité AWS)
- Bonnes pratiques dans IAM
- Amazon GuardDuty – root credential usage alert (Alerte d'utilisation des informations d'identification root)

- Step-by-step guidance on monitoring for root credential use through CloudTrail (Conseils étape par étape sur la surveillance de l'utilisation des informations d'identification root via Control Tower)
- MFA tokens approved for use with AWS (Jetons d'authentification multifactorielle approuvés pour une utilisation avec AWS)
- Implementing break glass access on AWS
- Top 10 security items to improve in your Compte AWS
- Que faire si je remarque une activité non autorisée dans mon Compte AWS ?

Vidéos connexes :

- Enable AWS adoption at scale with automation and governance
- Bonnes pratiques de sécurité : une approche Well-Architected
- Limitation de l'utilisation des AWS informations d'identification root de AWS re:inforce 2022 – Bonnes pratiques de sécurité avec AWS IAM

Exemples et ateliers connexes :

- Atelier : Compte AWS and root user

SEC01-BP03 Identifier et valider les objectifs de contrôle

Fixez et validez les objectifs de contrôle et les contrôles que vous devez appliquer à votre charge de travail en fonction de vos exigences de conformité et des risques identifiés à partir de votre modèle de menace. La validation continue des objectifs de contrôle et des contrôles permet de mesurer l'efficacité de l'atténuation des risques.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Identifier les exigences de conformité : découvrez les exigences organisationnelles, juridiques et de conformité que votre charge de travail doit nécessairement respecter.
- Identifier les ressources de conformité AWS : identifiez les ressources que propose AWS pour vous aider à rester conforme.
  - https://aws.amazon.com/compliance/
  - https://aws.amazon.com/artifact/

Ressources

Documents connexes :

- Directives d'audit de sécurité AWS
- Bulletins de sécurité

Vidéos connexes :

- AWS Security Hub : gérer les alertes de sécurité et automatiser la conformité
- Bonnes pratiques de sécurité : une approche Well-Architected

SEC01-BP04 Rester informé des menaces de sécurité

Identifiez les vecteurs d'attaque en restant informé des dernières menaces de sécurité afin de définir et de mettre en œuvre les contrôles appropriés. Utilisez AWS Managed Services pour faciliter la réception des notifications de comportement inattendu ou inhabituel dans vos comptes AWS. Réalisez vos investigations à l'aide d'outils partenaires AWS ou de flux d'informations sur les menaces tiers dans le cadre de votre flux d'informations de sécurité. La liste des vulnérabilités et risques communs (CVE) contient des vulnérabilités de cybersécurité divulguées publiquement, que vous pouvez utiliser pour rester à jour.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- S'abonner aux sources d'informations sur les menaces : consultez régulièrement les informations sur les menaces issues de plusieurs sources spécifiques aux technologies utilisées dans votre charge de travail.
  - Liste des vulnérabilités et risques communs (CVE)
- Envisager le service AWS Shield Advanced : offre une visibilité quasiment en temps réel sur les sources d'informations si votre charge de travail est accessible sur Internet.

Ressources

Documents connexes :

- Directives d'audit de sécurité AWS

- AWS Shield
- Bulletins de sécurité

Vidéos connexes :

- Bonnes pratiques de sécurité : une approche Well-Architected

SEC01-BP05 Connaître les recommandations de sécurité

Tenez-vous au courant des recommandations AWS et des recommandations de sécurité pertinentes pour faire évoluer le niveau de sécurité de votre charge de travail. Les bulletins de sécurité AWS contiennent des informations importantes sur les notifications de sécurité et de confidentialité.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Suivre l'actualité AWS : abonnez-vous ou consultez régulièrement les nouvelles recommandations, ainsi que les conseils et astuces.
  - Ateliers AWS Well-Architected
  - Blog sur la sécurité AWS
  - Documentation des services AWS
- S'abonner aux sources d'actualité sur le secteur : consultez régulièrement les flux d'actualités issus de plusieurs sources pertinentes pour les technologies qui sont utilisées dans votre charge de travail.
  - Exemple : liste des vulnérabilités et risques communs (CVE)

Ressources

Documents connexes :

- Bulletins de sécurité

Vidéos connexes :

- Bonnes pratiques de sécurité : une approche Well-Architected

SEC01-BP06 Automatiser les tests et la validation des contrôles de sécurité dans les pipelines

Établissez des bases et des modèles sécurisés pour les mécanismes de sécurité qui sont testés et validés dans le cadre de votre version, de vos pipelines et de vos processus. Utilisez des outils et l'automatisation pour tester et valider en continu tous les contrôles de sécurité. Par exemple, analysez des éléments tels que les images machine et les modèles d'infrastructure en tant que de code pour détecter les failles, les irrégularités et les dérives de sécurité par rapport à des points de référence établis à chaque étape. AWS CloudFormation Guard permet de vérifier que les modèles CloudFormation sont sûrs, vous font gagner du temps et réduisent le risque d'erreur de configuration.

Il est essentiel de réduire le nombre d'erreurs de configuration de sécurité introduites dans un environnement de production : plus vous contrôlez la qualité et réduisez les défauts dans le processus de génération, mieux c'est. Concevez des pipelines d'intégration et de déploiement continus (CI/CD, continuous integration and continuous deployment) pour tester la sécurité dans la mesure du possible. Les pipelines CI/CD offrent la possibilité d'améliorer la sécurité à chaque étape de la création et de la distribution. Les outils de sécurité CI/CD doivent être également maintenus à jour pour atténuer l'évolution des menaces.

Suivez les modifications apportées à la configuration de votre charge de travail pour faciliter les audits de conformité, la gestion des modifications et les enquêtes susceptibles de vous concerner. Vous pouvez utiliser AWS Config pour enregistrer et évaluer vos ressources AWS et tierces. Il vous permet d'auditer et d'évaluer en continu la conformité globale avec les règles et les packs de conformité, qui sont des ensembles de règles avec des actions correctives.

Le suivi des modifications doit inclure les modifications planifiées, qui font partie du processus de contrôle des modifications de votre organisation (parfois appelé MACD), les modifications non planifiées et les modifications inattendues, telles que les incidents. Des modifications peuvent se produire sur l'infrastructure, mais elles peuvent également être liées à d'autres catégories, telles que des changements dans les référentiels de code, des modifications au niveau des images machine et de l'inventaire d'applications, des modifications de processus et de politique ou des modifications de documentation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Automatiser la gestion de la configuration : appliquez et validez des configurations sécurisées automatiquement à l'aide d'un service ou d'un outil de gestion de la configuration.
  - AWS Systems Manager

- AWS CloudFormation
- Configuration d'un pipeline CI/CD sur AWS

Ressources

Documents connexes :

- Comment utiliser des politiques de contrôle des services pour définir des protections par autorisation dans les comptes de votre organisation AWS

Vidéos connexes :

- Managing Multi-Account AWS Environments Using AWS Organizations
- Bonnes pratiques de sécurité : une approche Well-Architected

SEC01-BP07 Identifier les menaces et hiérarchiser les atténuations à l'aide d'un modèle de menaces

Effectuez une modélisation des menaces pour identifier et gérer un registre actualisé des menaces potentielles et des mesures d'atténuation connexes pour votre charge de travail. Hiérarchisez vos menaces et adaptez vos atténuations des contrôles de sécurité pour les prévenir, les détecter et y répondre. Ajustez et maintenez ces mesures en fonction de votre charge de travail et de l'évolution de l'environnement de sécurité.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

Qu'est-ce que la modélisation des menaces ?

« La modélisation des menaces permet d'identifier, de communiquer et de comprendre les menaces et les atténuations dans le contexte de la protection de quelque chose de valeur. » – The Open Web Application Security Project (OWASP) Application Threat Modeling

Quel est l'intérêt de la modélisation des menaces ?

Les systèmes sont complexes et deviennent de plus en plus complexes et compétents au fil du temps, offrant plus de valeur opérationnelle, ainsi qu'une satisfaction et un engagement client accrus. Cela signifie que les décisions de conception informatique doivent tenir compte d'un nombre toujours

croissant de cas d'utilisation. Cette complexité et ce nombre de permutations des cas d'utilisation nuisent généralement à l'efficacité des approches non structurées pour trouver et atténuer les menaces. Dans ces conditions, il est préférable d'adopter une approche systématique pour recenser les menaces potentielles qui pèsent sur le système, concevoir les atténuations et d'établir la priorité de ces atténuations afin de veiller à ce que les ressources limitées de votre organisation aient un impact maximal sur l'amélioration de la posture de sécurité globale du système.

La modélisation des menaces est conçue pour fournir cette approche systématique, dans le but de trouver et de régler les problèmes au début du processus de conception, lorsque les atténuations impliquent un coût relatif et des efforts limités par rapport à plus tard dans le cycle de vie. Cette approche est conforme au principe de sécurité shift left. Au final, la modélisation des menaces s'intègre au processus de gestion des risques d'une organisation et aide à prendre des décisions sur les contrôles à mettre en œuvre en utilisant une approche axée sur les menaces.

Quand la modélisation des menaces doit-elle être effectuée ?

Commencez la modélisation des menaces le plus tôt possible dans le cycle de vie de votre charge de travail, afin de bénéficier de plus de flexibilité pour la gestion des menaces identifiées. Comme pour les bogues logiciels, plus vous identifiez les menaces rapidement, plus leur résolution est économique. Un modèle de menace est un document évolutif et il doit continuer à évoluer avec vos charges de travail. Revoyez vos modèles de menaces au fil du temps, y compris lorsqu'il y a un changement majeur, une évolution du contexte des menaces ou lorsque vous adoptez une nouvelle fonctionnalité ou un nouveau service.

Étapes d'implémentation

Comment pouvons-nous modéliser les menaces ?

Il existe de nombreuses façons de modéliser les menaces. Comme pour les langages de programmation, chaque méthode a ses avantages et ses inconvénients. À vous de choisir celle qui fonctionne le mieux pour votre organisation. Une approche consiste à commencer par le cadre des 4 questions de Shostack pour la modélisation des menaces, qui pose des questions ouvertes afin de structurer votre exercice de modélisation des menaces :

1. Sur quoi travaillons-nous ?

   Le but de cette question est de vous aider à comprendre et à vous mettre d'accord sur le système que vous créez et les détails associés qui sont pertinents pour la sécurité. La création d'un modèle ou d'un diagramme est la solution la plus populaire pour répondre à cette question, car elle vous aide à visualiser ce que vous créez, par exemple en utilisant un diagramme de flux des

données. Le fait de noter les hypothèses et les détails importants sur votre système vous aide également à définir ce qui est inclus dans le champ d'application. Cela permet à tous ceux qui contribuent au modèle de menaces de se concentrer sur la même chose et d'éviter les détours fastidieux pour étudier des sujets qui ne rentrent pas dans le champ d'application (y compris les versions obsolètes de votre système). Par exemple, si vous créez une application web, il n'est probablement pas intéressant de consacrer du temps à la modélisation de la séquence de démarrage autorisé du système d'exploitation pour les clients du navigateur, car vous ne pouvez pas avoir un impact sur ce point avec votre conception.

2. Quels problèmes pouvez-vous rencontrer ?

   C'est là que vous identifiez les menaces qui pèsent sur votre système. Les menaces sont des actions ou des événements accidentels ou intentionnels qui ont des impacts indésirables et pourraient affecter la sécurité de votre système. Sans une compréhension claire de ce qui pourrait poser un problème, vous n'avez aucun moyen de faire quoi que ce soit.

   Il n'existe pas de liste standard des problèmes potentiels. La création de cette liste nécessite un brainstorming et une collaboration entre tous les membres de votre équipe et les décideurs pertinents impliqués dans l'exercice de modélisation des menaces. Vous pouvez faciliter votre brainstorming en utilisant un modèle pour identifier les menaces, par exemple STRIDE, qui suggère différentes catégories à évaluer : Usurpation d'identité, Altération, Répudiation, Divulgation d'informations, Déni de service et Élévation de privilège. De plus, vous pouvez faciliter le brainstorming en examinant les listes et les recherches existantes afin de vous en inspirer, y compris l'OWASP Top 10, le HiTrust Threat Catalog, ainsi que le catalogue des menaces de votre organisation.

3. Qu'allons-nous faire à ce sujet ?

   Comme pour la question précédente, il n'existe pas de liste standard avec toutes les atténuations possibles. Lors de cette étape, les informations utilisées sont les menaces, les acteurs et les domaines d'amélioration identifiés par rapport à l'étape précédente.

   La sécurité et la conformité sont une responsabilité partagée entre vous et AWS. Il est important de comprendre que lorsque vous demandez « Qu'allons-nous faire à ce sujet ? », vous demandez également qui est responsable de ce qui doit être fait. En comprenant l'équilibre des responsabilités entre vous-même et AWS, vous pouvez évaluer votre exercice de modélisation des menaces en fonction des atténuations qui sont sous votre contrôle, c'est-à-dire, en règle générale, une combinaison des options de configuration du service AWS et vos propres atténuations spécifiques au système.

Pour la partie AWS de la responsabilité partagée, vous constaterez que les services AWS sont couverts par de nombreux programmes de conformité. Ces programmes vous aident à comprendre les contrôles rigoureux en place chez AWS afin de garantir la sécurité et la conformité du cloud. Les rapports d'audit de ces programmes peuvent être téléchargés pour les clients AWS à partir d'AWS Artifact.

Quels que soient les services AWS utilisés, il y a toujours un élément de responsabilité client et les atténuations correspondant à ces responsabilités doivent être incluses dans votre modèle de menaces. En ce qui concerne les atténuations en matière de contrôle de sécurité pour les services AWS eux-mêmes, envisagez l'implémentation de contrôles de sécurité dans tous les domaines, y compris la gestion des identités et des accès (authentification et autorisation), la protection des données (au repos et en transit), la sécurité de l'infrastructure, la journalisation et la surveillance. La documentation de chaque service AWS comporte un chapitre dédié à la sécurité qui fournit des conseils sur les contrôles de sécurité à implémenter à des fins d'atténuation. Il est surtout important de réfléchir au code que vous écrivez et à ses dépendances, ainsi que de penser aux contrôles que vous pourriez mettre en place pour résoudre ces menaces. Ces contrôles peuvent notamment prendre les formes suivantes : validation des entrées, gestion des sessions et gestion des limites. La plupart des vulnérabilités sont souvent introduites dans le code personnalisé, c'est pourquoi il est important de se concentrer sur ce domaine.

4. Avons-nous fait du bon travail ?

   L'objectif est que votre équipe et votre organisation améliorent la qualité des modèles de menaces et la vitesse à laquelle vous effectuez la modélisation des menaces au fil du temps. Ces améliorations découlent d'une combinaison de pratique, d'apprentissage, d'enseignement et de révision. Pour approfondir ces notions et vous exercer, votre équipe et vous-même pouvez suivre le cours de formation ou l'atelier sur les bons principes de modélisation des menaces pour les créateurs. De plus, si vous souhaitez obtenir des conseils sur l'intégration de la modélisation des menaces dans le cycle de développement des applications de votre organisation, consultez la publication How to approach threat modeling sur le Blog de sécurité d'AWS.

Threat Composer

Pour vous aider et vous guider dans la modélisation des menaces, pensez à utiliser l'outil Threat Composer, qui vise à réduire le délai de modélisation des menaces. L'outil vous permet d'effectuer les opérations suivantes :

- Rédiger des déclarations de menaces utiles, qui respectent la grammaire des menaces et fonctionnent dans un flux de travail naturel et non linéaire
- Générer un modèle de menaces lisible par l'homme
- Générer un modèle de menaces lisible par machine pour vous permettre de traiter les modèles de menaces comme du code
- Identifier rapidement les domaines dans lesquels la qualité et la couverture peuvent être améliorées à l'aide du tableau de bord

Pour en savoir plus, accédez à Threat Composer et basculez vers l'exemple d'espace de travail défini par le système.

Ressources

Bonnes pratiques associées :

- SEC01-BP03 Identifier et valider les objectifs de contrôle
- SEC01-BP04 Rester informé des menaces de sécurité
- SEC01-BP05 Connaître les recommandations de sécurité
- SEC01-BP08 Évaluer et mettre en œuvre régulièrement de nouveaux services et de nouvelles fonctions de sécurité

Documents connexes :

- How to approach threat modeling (AWS Security Blog)
- NIST: Guide to Data-Centric System Threat Modelling

Vidéos connexes :

- AWS Summit ANZ 2021 - How to approach threat modelling
- AWS Summit ANZ 2022 - Scaling security – Optimise for fast and secure delivery

Formations associées :

- Threat modeling the right way for builders – AWS Skill Builder virtual self-paced training
- Threat modeling the right way for builders – AWS Workshop

Outils associés :

- Threat Composer

SEC01-BP08 Évaluer et mettre en œuvre régulièrement de nouveaux services et de nouvelles fonctions de sécurité

Évaluez et mettez en œuvre les services et fonctions de sécurité proposés par AWS et les partenaires AWS qui vous permettent de faire évoluer le niveau de sécurité de votre charge de travail. Le blog sur la sécurité AWSmet en évidence les nouveaux services et fonctionnalités AWS, les guides de mise en œuvre et des conseils généraux de sécurité. Les nouveautés AWS représentent un excellent moyen de se tenir au courant de tous les nouveaux services, fonctionnalités et annonces AWS.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Planifier des vérifications régulières : créez un calendrier d'activités de révision qui inclut les exigences de conformité, l'évaluation des nouvelles fonctionnalités et des nouveaux services de sécurité AWS et des créneaux pour rester informé des actualités du secteur.
- Découvrir les services et fonctions AWS : découvrez les fonctions de sécurité qui sont disponibles pour les services que vous utilisez. Évaluez les nouvelles fonctions au fur et à mesure qu'elles sont publiées.
  - Blog sur la sécurité AWS
  - Bulletins de sécurité AWS
  - Documentation des services AWS
- Définir le processus d'intégration des services AWS : définissez les processus d'intégration des nouveaux services AWS. Incluez la manière d'évaluer le fonctionnement des nouveaux services AWS et les exigences en matière de conformité pour votre charge de travail.
- Tester les nouveaux services et les nouvelles fonctions : testez les nouveaux services et les nouvelles fonctions au fil de leur publication dans un environnement hors production qui réplique fidèlement votre environnement de production.
- Mettre en place d'autres mécanismes de défense : implémentez des mécanismes automatisés pour défendre votre charge de travail et explorez les options disponibles.
  - Correction des ressources AWS non conformes à l'aide de règles AWS Config Rules

Ressources

Vidéos connexes :

- Bonnes pratiques de sécurité : une approche Well-Architected

Identity and Access Management

Questions

- SÉC 2. Comment gérez-vous l'authentification pour les personnes et les machines ?
- SÉC 3. Comment gérez-vous les autorisations des personnes et des machines ?

SÉC 2. Comment gérez-vous l'authentification pour les personnes et les machines ?

Il existe deux types d'identités à gérer dans le cadre de l'exploitation de charges de travail AWS sécurisées. Comprendre le type d'identité que vous devez gérer et pour lequel vous devez autoriser l'accès vous permet de garantir l'accès aux ressources adéquates, dans les bonnes conditions.

Identités humaines : vos administrateurs, développeurs, opérateurs et utilisateurs finaux ont besoin d'une identité pour accéder à vos environnements et applications AWS. Il s'agit des membres de votre organisation ou des utilisateurs externes avec lesquels vous collaborez et qui interagissent avec vos ressources AWS via un navigateur web, une application cliente ou des outils de ligne de commande interactifs.

Identités des machines : vos applications de service, outils opérationnels et charges de travail nécessitent une identité pour envoyer des demandes aux services AWS, par exemple pour lire des données. Ces identités comprennent des machines s'exécutant dans votre environnement AWS, telles que des instances Amazon EC2 ou des fonctions AWS Lambda. Vous pouvez également gérer les identités de machines pour les tiers qui ont besoin d'un accès. De plus, certaines machines en dehors d'AWS peuvent avoir besoin d'accéder à votre environnement AWS.

Bonnes pratiques

- SEC02-BP01 Utiliser de solides mécanismes d'authentification
- SEC02-BP02 Utiliser des informations d'identification temporaires
- SEC02-BP03 Stocker et utiliser des secrets en toute sécurité
- SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé

- SEC02-BP05 Contrôler et effectuer régulièrement une rotation des informations d'identification
- SEC02-BP06 Tirer parti des groupes d'utilisateurs et des attributs

SEC02-BP01 Utiliser de solides mécanismes d'authentification

Les connexions (authentification au moyen d'informations d'identification de connexion) peuvent présenter des risques lorsque l'on n'utilise pas des mécanismes tels que l'authentification multifactorielle (MFA), surtout dans les situations où les informations d'identification de connexion ont été divulguées par inadvertance ou peuvent être devinées facilement. Vous devez utiliser de solides mécanismes d'authentification pour réduire ces risques en exigeant l'authentification multifactorielle (MFA) et des politiques strictes de gestion des mots de passe.

Résultat souhaité : réduire les risques d'accès involontaire aux informations d'identification dans AWS en utilisant des mécanismes de connexion solides pour les utilisateurs AWS Identity and Access Management (IAM), l'utilisateur root Compte AWS, AWS IAM Identity Center (successeur d'AWS Single Sign-On [AWS SSO]), et les fournisseurs d'identité tiers. Cela signifie que vous devez exiger une authentification multifactorielle, appliquer des politiques strictes de gestion des mots de passe et détecter les comportements de connexion anormaux.

Anti-modèles courants :

- Ne pas appliquer de politique stricte de gestion des mots de passe pour vos identités, notamment des mots de passe complexes et l'authentification multifactorielle (MFA).
- Utiliser les mêmes informations d'identification pour différents utilisateurs.
- Ne pas utiliser de contrôles de détection pour les connexions suspectes.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

Les identités humaines peuvent se connecter à AWS de plusieurs façons. Une bonne pratique AWS consiste à faire appel à un fournisseur d'identité centralisé utilisant la fédération (fédération directe ou AWS IAM Identity Center) lors de l'authentification auprès d'AWS. Dans ce cas, vous devez établir une connexion sécurisée avec votre fournisseur d'identité ou Microsoft Active Directory.

Lorsque vous ouvrez pour la première fois un Compte AWS, vous commencez avec un utilisateur root de Compte AWS. Vous devez utiliser uniquement l'utilisateur root du compte afin de configurer l'accès pour vos utilisateurs (et pour les tâches qui requièrent l'utilisateur root). Il est important

d'activer l'authentification multifactorielle pour l'utilisateur root du compte immédiatement après l'ouverture de votre Compte AWS et de sécuriser l'utilisateur root en s'appuyant sur le Guide des bonnes pratiques AWS.

Si vous créez des utilisateurs dans AWS IAM Identity Center, sécurisez le processus de connexion dans ce service. Pour les identités des consommateurs, vous pouvez utiliser Amazon Cognito user pools et sécuriser le processus de connexion dans ce service, ou utiliser l'un des fournisseurs d'identité pris en charge par Amazon Cognito user pools.

Si vous employez des utilisateurs AWS Identity and Access Management (IAM), vous devez sécuriser le processus de connexion à l'aide d'IAM.

Quelle que soit la méthode de connexion, il est essentiel d'appliquer une politique de connexion rigoureuse.

Étapes d'implémentation

Voici des recommandations générales en matière de connexion. Les paramètres réels que vous configurez doivent être définis par votre politique d'entreprise ou utiliser une norme telle que NIST 800-63.

- Exigez l'authentification multifactorielle. Dans le cadre des bonnes pratiques IAM, il est recommandé d'exiger l'authentification multifactorielle pour les identités et les charges de travail humaines. L'activation de l'authentification multifactorielle fournit une couche de sécurité supplémentaire en exigeant que les utilisateurs fournissent des informations d'identification et un mot de passe unique (OTP) ou une chaîne de caractères générée et vérifiée cryptographiquement à partir d'un appareil physique.
- Mettez en place une longueur de mot de passe minimale, il s'agit d'un facteur essentiel pour garantir la force du mot de passe.
- Appliquez la complexité des mots de passe pour les rendre plus difficiles à deviner.
- Permettez aux utilisateurs de changer leurs propres mots de passe.
- Créez des identités individuelles plutôt que des informations d'identification partagées. En créant des identités individuelles, vous pouvez attribuer à chaque utilisateur un ensemble unique d'informations d'identification de sécurité. Les utilisateurs individuels offrent la possibilité d'auditer l'activité de chaque utilisateur.

Recommandations IAM Identity Center :

- IAM Identity Center fournit une politique de mot de passe prédéfinie lorsque vous utilisez le répertoire par défaut qui établit la longueur, la complexité et les exigences de réutilisation du mot de passe.
- Activez l'authentification multifactorielle et configurez le paramètre contextuel ou toujours activé pour l'authentification multifactorielle lorsque la source d'identité est le répertoire par défaut, AWS Managed Microsoft AD ou AD Connector.
- Autorisez les utilisateurs à enregistrer leurs propres appareils d'authentification multifactorielle (MFA).

Recommandations pour les répertoires Amazon Cognito user pools :

- Configurez les paramètres de force des mots de passe.
- Exigez l'authentification multifactorielle pour les utilisateurs.
- Utilisez les paramètres de sécurité avancés Amazon Cognito user pools pour les fonctionnalités telles que l'authentification adaptative qui peut bloquer les connexions suspectes.

Recommandations pour les utilisateurs IAM :

- Idéalement, vous utilisez IAM Identity Center ou la fédération directe. Cependant, vous aurez peut-être besoin d'utilisateurs IAM. Le cas échéant, définissez une politique de mot de passe pour les utilisateurs IAM. Vous pouvez utiliser la politique de gestion des mots de passe pour définir des exigences telles que la longueur minimale ou la nécessité d'utiliser des caractères non alphabétiques.
- Créez une politique IAM pour appliquer la connexion avec authentification multifactorielle afin que les utilisateurs puissent gérer leurs propres mots de passe et appareils d'authentification multifactorielle.

Ressources

Bonnes pratiques associées :

- SEC02-BP03 Stocker et utiliser des secrets en toute sécurité
- SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé
- SEC03-BP08 Partager des ressources en toute sécurité au sein de votre organisation

Documents connexes :

- AWS IAM Identity Center (successor to AWS Single Sign-On)
- Définition d'une politique de mot de passe du compte pour les utilisateurs IAM
- Utilisateur root Compte AWS
- Amazon Cognito password policy
- AWS Credentials (Informations d'identification AWS)
- Bonnes pratiques de sécurité dans IAM

Vidéos connexes :

- Managing user permissions at scale with AWS IAM Identity Center
- Mastering identity at every layer of the cake

SEC02-BP02 Utiliser des informations d'identification temporaires

Lors de tout type d'authentification, il est préférable d'utiliser des informations d'identification temporaires plutôt que des informations d'identification à long terme afin de réduire ou d'éliminer les risques, tels que la divulgation, le partage ou le vol des informations d'identification par inadvertance.

Résultat souhaité : réduire les risques liés aux informations d'identification à long terme, utiliser des informations d'identification temporaires dès que possible pour les identités humaines et machine. Les informations d'identification à long terme créent de nombreux risques, par exemple lorsqu'ils sont téléchargés dans du code dans des référentiels GitHub publics. En utilisant des informations d'identification temporaires, vous réduisez considérablement les risques de compromission de ces informations.

Anti-modèles courants :

- Les développeurs utilisent des clés d'accès à long terme issues des IAM users au lieu d'obtenir des informations d'identification temporaires de la CLI à l'aide de la fédération.
- Les développeurs intègrent des clés d'accès à long terme dans leur code et téléchargent ce code dans des référentiels Git publics.
- Les développeurs intègrent des clés d'accès à long terme dans les applications mobiles qui sont ensuite disponibles dans les boutiques d'applications.

- Les utilisateurs partagent des clés d'accès à long terme avec d'autres utilisateurs ou des employés quittent l'entreprise avec des clés d'accès à long terme toujours en leur possession.
- Utilisation des clés d'accès à long terme pour les identités machine lorsque des informations d'identification temporaires peuvent être utilisées.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

Utilisez des informations d'identification de sécurité temporaires plutôt que des informations d'identification à long terme pour toutes les demandes d'API et de CLI AWS. Les demandes d'API et CLI transmises aux services AWS doivent, dans presque tous les cas, être signées en utilisant des clés d'accès AWS. Ces demandes peuvent être signées avec des informations d'identification temporaires ou à long terme. Le seul moment où vous devez utiliser des informations d'identification à long terme, également connues sous le nom de clés d'accès à long terme, est si vous employez un utilisateur IAM ou l'utilisateur root du Compte AWS. Lorsque vous fédérez sur AWS ou si vous assumez un rôle IAM via d'autres méthodes, des informations d'identification temporaires sont générées. Même lorsque vous accédez à la AWS Management Console à l'aide des informations d'identification de connexion, des informations d'identification temporaires sont gérées pour vous permettre d'appeler les services AWS. Vous avez rarement besoin d'informations d'identification à long terme et vous pouvez accomplir presque toutes les tâches en utilisant des informations d'identification temporaires.

Privilégiez les informations d'identification temporaires plutôt que les informations d'identification à long terme et, parallèlement, mettez en place une stratégie de réduction des utilisateurs IAM au profit de la fédération et des rôles IAM. Bien que les utilisateurs IAM aient été employés pour les identités humaines et machine dans le passé, nous recommandons désormais de ne plus procéder ainsi afin d'éviter les risques liés à l'utilisation de clés d'accès à long terme.

Étapes d'implémentation

Pour les identités humaines comme les employés, les administrateurs, les développeurs, les opérateurs et les clients :

- faites appel à un fournisseur d'identité centralisé et exigez des utilisateurs humains qu'ils se servent de la fédération avec un fournisseur d'identité pour accéder à AWS à l'aide d'informations d'identification temporaires. La fédération pour vos utilisateurs peut être mise en place soit avec une fédération directe à chaque Compte AWS, soit en utilisant AWS IAM Identity Center

(successeur d'AWS IAM Identity Center) et le fournisseur d'identité de votre choix. La fédération offre un certain nombre d'avantages par rapport aux utilisateurs IAM, outre l'élimination des informations d'identification à long terme. Les utilisateurs peuvent également demander des informations d'identification temporaires à partir de la ligne de commande pour une fédération directe ou en utilisant IAM Identity Center. Cela signifie que peu de cas d'utilisation nécessitent des utilisateurs IAM ou des informations d'identification à long terme pour vos utilisateurs.

- Lors de l'octroi d'accès aux ressources à des tiers, par exemple les fournisseurs de logiciels en tant que service (SaaS) dans votre Compte AWS, vous pouvez utiliser des rôles intercomptes et des politiques basées sur les ressources.
- Si vous devez accorder à des consommateurs ou des clients des autorisations d'accès à vos ressources AWS, vous pouvez utiliser des groupes d'identités Amazon Cognito ou Amazon Cognito user pools pour fournir des informations d'identification temporaires. Les autorisations pour les informations d'identification sont configurées via des rôles IAM. Vous pouvez également définir un rôle IAM distinct avec des autorisations limitées pour les utilisateurs invités qui ne sont pas authentifiés.

Pour les identités machine, vous devrez peut-être utiliser des informations d'identification à long terme. Le cas échéant, vous devez exiger que les charges de travail utilisent des informations d'identification temporaires avec des rôles IAM pour accéder à AWS.

- Pour Amazon Elastic Compute Cloud (Amazon EC2), vous pouvez utiliser des rôles pour Amazon EC2.

- AWS Lambda vous permet de configurer un rôle d'exécution Lambda afin d'octroyer au service les autorisations permettant d'effectuer des actions AWS en utilisant des informations d'identification temporaires. Il existe de nombreux modèles similaires pour permettre aux services AWS d'octroyer des informations d'identification temporaires à l'aide des rôles IAM.
- Pour les appareils IoT, vous pouvez utiliser le fournisseur d'informations d'identification AWS IoT Core afin de demander des informations d'identification temporaires.
- Pour les systèmes sur site ou les systèmes qui fonctionnent en dehors d'AWS et qui ont besoin d'accéder aux ressources AWS, vous pouvez utiliser IAM Roles Anywhere.

Dans certains cas, il est impossible d'utiliser des informations d'identification temporaires et vous devrez alors opter pour des informations d'identification à long terme. Le cas échéant, auditez et effectuez une rotation des informations d'identification périodiquement et effectuez une rotation des

clés d'accès régulièrement pour les cas d'utilisation qui requièrent des informations d'identification à long terme. Parmi les exemples qui peuvent exiger des informations d'identification à long terme, citons notamment les plug-ins WordPress et les clients AWS tiers. Dans les situations où vous devez utiliser des informations d'identification à long terme ou des informations d'identification autres que les clés d'accès AWS, comme les connexions aux bases de données, vous pouvez utiliser un service conçu pour gérer la gestion des secrets, par exemple AWS Secrets Manager. Secrets Manager facilite la gestion, la rotation et le stockage sécurisé des secrets chiffrés à l'aide des services pris en charge. Pour plus d'informations sur la rotation des informations d'identification à long terme, consultez Rotation des clés d'accès.

Ressources

Bonnes pratiques associées :

- SEC02-BP03 Stocker et utiliser des secrets en toute sécurité
- SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé
- SEC03-BP08 Partager des ressources en toute sécurité au sein de votre organisation

Documents connexes :

- Informations d'identification de sécurité temporaires
- AWS Credentials (Informations d'identification AWS)
- Bonnes pratiques de sécurité dans IAM
- Rôles IAM
- IAM Identity Center
- Fournisseurs d'identité et fédération
- Rotating Access Keys (Rotation des clés d'accès)
- Solutions partenaires de sécurité : accès et contrôle d'accès
- Utilisateur root Compte AWS

Vidéos connexes :

- Managing user permissions at scale with AWS IAM Identity Center (successor to AWS IAM Identity Center)
- Mastering identity at every layer of the cake

SEC02-BP03 Stocker et utiliser des secrets en toute sécurité

Une charge de travail nécessite une capacité automatisée pour prouver son identité aux bases de données, aux ressources et aux services tiers. Cela se fait à l'aide d'identifiants d'accès secrets, tels que des clés d'accès à l'API, des mots de passe et des jetons OAuth. L'utilisation d'un service spécialement conçu pour stocker, gérer et faire tourner ces informations d'identification permet de réduire les risques de compromission de ces informations d'identification.

Résultat souhaité : implémentation d'un mécanisme de gestion sécurisée des informations d'identification des applications qui atteint les objectifs suivants :

- Identification des secrets nécessaires pour la charge de travail.
- Réduction du nombre d'informations d'identification à long terme requis en les remplaçant par des informations d'identification à court terme, dans la mesure du possible.
- Établissement d'un stockage sécurisé et d'une rotation automatisée des informations d'identification à long terme restantes.
- Audit de l'accès aux secrets qui existent dans la charge de travail.
- Surveillance continue pour vérifier qu'aucun secret n'est intégré dans le code source pendant le processus de développement.
- Réduction des risques de divulgation des informations d'identification par inadvertance.

Anti-modèles courants :

- Aucune rotation des informations d'identification.
- Stockage des informations d'identification à long terme dans le code source ou les fichiers de configuration.
- Stockage des informations d'identification au repos non chiffrées.

Avantages liés à l'instauration de cette bonne pratique :

- Les secrets sont chiffrés au repos et en transit.
- L'accès aux informations d'identification est sécurisé par une API (il s'agit plus ou moins d'un distributeur d'informations d'identification).
- L'accès à une information d'identification (en lecture et en écriture) est audité et consigné.
- Séparation des préoccupations : la rotation des informations d'identification est effectuée par un composant distinct, qui peut être séparé du reste de l'architecture.

- Les secrets sont distribués automatiquement à la demande aux composants logiciels et la rotation se produit dans un emplacement central.
- L'accès aux informations d'identification peut être contrôlé de façon précise.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

Dans le passé, les informations d'identification utilisées pour s'authentifier auprès des bases de données, des API tierces, des jetons et d'autres secrets pouvaient être intégrées dans du code source ou des fichiers d'environnement. AWS fournit plusieurs mécanismes pour stocker ces informations d'identification en toute sécurité, en effectuer la rotation automatiquement et vérifier leur utilisation.

Pour gérer les secrets de façon optimale, la meilleure solution consiste à suivre les directives de suppression, de remplacement et de rotation. Les informations d'identification les plus sûres sont celles que vous n'avez pas à stocker, gérer ou manipuler. Certaines informations d'identification qui ne sont plus nécessaires au fonctionnement de la charge de travail peuvent être supprimées en toute sécurité.

Pour les informations d'identification qui restent nécessaires au bon fonctionnement de la charge de travail, il peut être possible d'opter pour une solution temporaire ou à court terme au lieu d'utiliser des informations à long terme. Par exemple, au lieu de coder en dur une clé d'accès secrète AWS, envisagez de remplacer les informations d'identification à long terme par des informations d'identification temporaires à l'aide de rôles IAM.

Certains secrets de longue durée ne peuvent pas être supprimés ni remplacés. Ces secrets peuvent être stockés dans un service tel qu'AWS Secrets Manager, où ils peuvent être stockés, gérés et mis en rotation de façon centralisée.

Un audit du code source de la charge de travail et des fichiers de configuration peut révéler de nombreux types d'informations d'identification. Le tableau suivant résume les stratégies de traitement des types courants d'informations d'identification :

Credential type	Description	Suggested strategy
IAM access keys	AWS IAM access and secret keys used to assume IAM roles inside of a workload	Replace: Use Rôles IAM assigned to the compute instances (such as Amazon

Credential type	Description	Suggested strategy
		EC2 or AWS Lambda) instead. For interoperability with third parties that require access to resources in your Compte AWS, ask if they support Accès intercompte AWS. For mobile apps, consider using temporary credentials through Groupes d'identités Amazon Cognito (identités fédérées). For workloads running outside of AWS, consider IAM Roles Anywhere or AWS Systems Manager Hybrid Activations.
SSH keys	Secure Shell private keys used to log into Linux EC2 instances, manually or as part of an automated process	Replace: Use AWS Systems Manager or EC2 Instance Connect to provide programmatic and human access to EC2 instances using IAM roles.
Application and database credentials	Passwords – plain text string	Rotate: Store credentials in AWS Secrets Manager and establish automated rotation if possible.

Credential type	Description	Suggested strategy
Amazon RDS and Aurora Admin Database credentials	Passwords – plain text string	Replace: Use the Intégration d'Secrets Manager à Amazon RDS or Amazon Aurora. In addition, some RDS database types can use IAM roles instead of passwords for some use cases (for more detail, see Authentification de base de données IAM).
OAuth tokens	Secret tokens – plain text string	Rotate: Store tokens in AWS Secrets Manager and configure automated rotation.
API tokens and keys	Secret tokens – plain text string	Rotate: Store in AWS Secrets Manager and establish automated rotation if possible.

Parmi les anti-modèles courants, citons l'intégration des clés d'accès IAM dans le code source, les fichiers de configuration ou les applications mobiles. Lorsqu'une clé d'accès IAM est requise pour communiquer avec un service AWS, utilisez des informations d'identification de sécurité temporaires (à court terme). Ces informations d'identification à court terme peuvent être fournies via des rôles IAM pour les instances EC2, des rôles d'exécution pour les fonctions Lambda, des rôles IAM Cognito pour l'accès des utilisateurs mobiles et des politiques IoT Core pour les appareils IoT. Lorsque vous interagissez avec des tiers, privilégiez la délégation des accès à un rôle IAM avec l'accès nécessaire aux ressources de votre compte au lieu de configurer un utilisateur IAM et d'envoyer au tiers la clé d'accès secrète pour cet utilisateur.

Dans de nombreux cas, la charge de travail exige le stockage de secrets nécessaires pour interagir avec d'autres services et ressources. AWS Secrets Manager est conçu pour gérer en toute sécurité ces informations d'identification, ainsi que le stockage, l'utilisation et la rotation des jetons d'API, mots de passe et autres informations d'identification.

AWS Secrets Manager fournit cinq capacités clés pour assurer le stockage et la manipulation sécurisés des informations d'identification sensibles : chiffrement au repos, chiffrement en transit,

audit complet, contrôle d'accès détaillé et rotation extensible des informations d'identification. D'autres services de gestion des secrets créés par des partenaires AWS ou des solutions développées localement qui offrent des capacités et des assurances similaires sont également acceptables.

Étapes d'implémentation

1. Identifiez les chemins de code contenant des informations d'identification codées en dur à l'aide d'outils automatisés tels que Amazon CodeGuru.
   - Utilisez Amazon CodeGuru pour analyser vos référentiels de code. Une fois la vérification terminée, filtrez sur `Type=Secrets` dans CodeGuru afin de trouver les lignes de code qui posent problème.
2. Identifiez les informations d'identification qui peuvent être supprimées ou remplacées.
   a. Identifiez les informations d'identification qui ne sont plus nécessaires et marquez-les en vue de leur suppression.
   b. Pour les clés secrètes AWS qui sont intégrées au code source, remplacez-les par des rôles IAM associés aux ressources nécessaires. Si une partie de votre charge de travail se trouve en dehors d'AWS mais requiert des informations d'identification IAM pour accéder aux ressources AWS, envisagez l'utilisation d'IAM Roles Anywhere ou d'AWS Systems Manager Hybrid Activations.
3. Pour les autres secrets tiers de longue durée qui nécessitent l'utilisation de la stratégie de rotation, intégrez Secrets Manager dans votre code afin d'extraire les secrets tiers au moment de l'exécution.
   a. La console CodeGuru peut créer automatiquement un secret dans Secrets Manager à l'aide des informations d'identification découvertes.
   b. Intégrez l'extraction des secrets d'Secrets Manager dans votre code d'application.
      - Les fonctions Lambda sans serveur peuvent utiliser une extension Lambda qui ne dépend pas du langage.
      - Pour les instances ou conteneurs EC2, AWS fournit un exemple de code côté client permettant d'extraire les secrets d'Secrets Manager dans plusieurs langages de programmation populaires.
4. Examinez régulièrement votre base de code et effectuez une nouvelle analyse afin de vérifier qu'aucun nouveau secret n'a été ajouté au code.
   - Envisagez d'utiliser un outil tel que git-secrets pour éviter d'intégrer de nouveaux secrets dans votre référentiel de code source.

5. Surveillez l'activité d'Secrets Manager afin de détecter toute utilisation inattendue, tout accès aux secrets inapproprié ou toute tentative de suppression de secrets.
6. Réduisez l'exposition humaine aux informations d'identification. Limitez l'accès à la lecture, à l'écriture et à la modification des informations d'identification à un rôle IAM dédié à cette fin et fournissez un accès uniquement pour assumer le rôle à un petit sous-ensemble d'utilisateurs opérationnels.

Ressources

Bonnes pratiques associées :

- SEC02-BP02 Utiliser des informations d'identification temporaires
- SEC02-BP05 Contrôler et effectuer régulièrement une rotation des informations d'identification

Documents connexes :

- Getting Started with AWS Secrets Manager (Démarrer avec AWS Secrets Manager)
- Fournisseurs d'identité et fédération
- Amazon CodeGuru Introduces Secrets Detector (Amazon DevOps Guru présente le détecteur de secrets)
- Comment AWS Secrets Manager utilise AWS Key Management Service
- Chiffrement et déchiffrement de secrets dans Secrets Manager
- Entrées de blog sur Secrets Manager
- Amazon RDS annonce l'intégration avec AWS Secrets Manager

Vidéos connexes :

- Best Practices for Managing, Retrieving, and Rotating Secrets at Scale
- Find Hard-Coded Secrets Using Amazon CodeGuru Secrets Detector
- Securing Secrets for Hybrid Workloads Using AWS Secrets Manager

Ateliers connexes :

- Store, retrieve, and manage sensitive credentials in AWS Secrets Manager

- AWS Systems Manager Hybrid Activations

SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé

Pour les identités du personnel (employés et sous-traitants), faites confiance à un fournisseur d'identité qui vous permet de gérer les identités de manière centralisée. Cela facilite la gestion de l'accès entre plusieurs applications et systèmes, car vous créez, attribuez, gérez, révoquez et auditez l'accès depuis un seul emplacement.

Résultat souhaité : Vous disposez d'un fournisseur d'identité centralisé dans lequel vous gérez de manière centralisée les utilisateurs faisant partie du personnel, les politiques d'authentification (telles que l'exigence d'authentification multifactorielle (MFA)) et les autorisations accordées aux systèmes et aux applications (telles que l'attribution de l'accès en fonction de l'appartenance à un groupe ou des attributs d'un utilisateur). Les utilisateurs en interne se connectent au fournisseur d'identité central et se fédèrent (authentification unique) avec les applications internes et externes, ce qui leur évite d'avoir à mémoriser différentes informations d'identification. Votre fournisseur d'identité est intégré à vos systèmes de ressources humaines (RH) afin que les changements de personnel soient automatiquement synchronisés avec lui. Par exemple, si quelqu'un quitte votre organisation, vous pouvez automatiquement révoquer l'accès aux applications et systèmes fédérés (y compris AWS). Vous avez activé la journalisation détaillée des audits dans votre fournisseur d'identité et vous surveillez ces journaux pour détecter tout comportement inhabituel des utilisateurs.

Anti-modèles courants :

- Vous n'utilisez pas la fédération ni l'authentification unique. Les utilisateurs en interne créent des comptes utilisateur et des informations d'identification distincts dans plusieurs applications et systèmes.
- Vous n'avez pas automatisé le cycle de vie des identités pour les utilisateurs en interne, par exemple en intégrant votre fournisseur d'identité à vos systèmes RH. Lorsqu'un utilisateur quitte votre organisation ou change de rôle, vous suivez un processus manuel pour supprimer ou mettre à jour ses enregistrements dans plusieurs applications et systèmes.

Avantages liés au respect de cette bonne pratique : En utilisant un fournisseur d'identité centralisé, vous disposez d'un emplacement unique pour gérer les identités et les politiques des utilisateurs en interne, de la possibilité d'attribuer l'accès aux applications, aux utilisateurs et aux groupes, et de la capacité de surveiller l'activité de connexion des utilisateurs. Grâce à l'intégration du fournisseur d'identité dans vos systèmes de ressources humaines (RH), lorsqu'un utilisateur change de rôle,

ces modifications sont synchronisées avec le fournisseur d'identité et mettent automatiquement à jour les applications et les autorisations qui lui ont été attribuées. Lorsqu'un utilisateur quitte votre organisation, son identité est automatiquement désactivée dans le fournisseur d'identité, révoquant ainsi son accès aux applications et systèmes fédérés.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Conseils pour les utilisateurs en interne accédant à AWS

Les utilisateurs en interne, tels que les employés et les sous-traitants de votre organisation, peuvent avoir besoin d'accéder à AWS avec la AWS Management Console ou AWS Command Line Interface (AWS CLI) pour exécuter leurs tâches. Vous pouvez accorder l'accès AWS aux utilisateurs en interne en les fédérant avec AWS à deux niveaux à partir de votre fournisseur d'identité centralisé : fédération directe vers chaque Compte AWS ou fédération vers plusieurs comptes dans votre organisation AWS.

- Pour fédérer les utilisateurs en interne directement avec chaque Compte AWS, vous pouvez utiliser un fournisseur d'identité centralisé afin de les fédérer à AWS Identity and Access Management sur ce compte. La flexibilité d'IAM vous permet d'activer un fournisseur d'identité SAML 2.0 ou OpenID Connect (OIDC) distinct pour chaque Compte AWS et d'utiliser les attributs des utilisateurs fédérés pour le contrôle de l'accès. Les utilisateurs en interne utiliseront leur navigateur web pour se connecter au fournisseur d'identité en indiquant leurs informations d'identification (telles que des mots de passe et des codes de jeton MFA). Le fournisseur d'identité enverra à son navigateur une assertion SAML soumise à l'URL de connexion de la AWS Management Console pour permettre à l'utilisateur de s'authentifier de manière unique auprès de la AWS Management Console en assumant un rôle IAM. Vos utilisateurs peuvent également obtenir des informations d'identification d'API AWS temporaires à utiliser dans AWS CLI ou les kits SDK AWS depuis AWS STS en endossant le rôle IAM à l'aide d'une assertion SAML auprès du fournisseur d'identité.
- Pour fédérer les utilisateurs en interne disposant de plusieurs comptes dans votre organisation AWS, vous pouvez utiliser AWS IAM Identity Center afin de gérer de manière centralisée l'accès des utilisateurs en interne aux Comptes AWS et aux applications. Activez Identity Center pour votre organisation et configurez votre source d'identité. IAM Identity Center fournit un annuaire source d'identités par défaut que vous pouvez utiliser pour gérer vos utilisateurs et vos groupes. Vous pouvez également choisir une source d'identité externe en vous connectant à votre fournisseur d'identité externe à l'aide de SAML 2.0 et en approvisionnant automatiquement les utilisateurs et les groupes avec SCIM, ou en vous connectant à votre annuaire Microsoft AD avec AWS Directory

Service. Une fois qu'une source d'identité est configurée, vous pouvez attribuer aux utilisateurs et aux groupes l'accès aux Comptes AWS en définissant des politiques de moindre privilège dans vos ensembles d'autorisations. Les utilisateurs en interne peuvent s'authentifier par le biais de votre fournisseur d'identité central pour se connecter au portail d'accès AWS et s'authentifier de manière unique aux Comptes AWS et aux applications cloud qui leur sont attribués. Vos utilisateurs peuvent configurer AWS CLI v2 pour s'authentifier auprès d'Identity Center et obtenir des informations d'identification pour exécuter des commandes AWS CLI. Identity Center permet également l'accès par authentification unique à des applications AWS comme Amazon SageMaker Studio et les portails AWS IoT Sitewise Monitor.

Une fois que vous aurez suivi les instructions précédentes, vos utilisateurs en interne n'auront plus besoin d'utiliser des IAM users et des groupes pour les opérations normales lors de la gestion des charges de travail sur AWS. Au lieu de cela, vos utilisateurs et vos groupes seront gérés en dehors d'AWS, et les utilisateurs pourront accéder aux ressources AWS en tant qu'identité fédérée. Les identités fédérées utilisent les groupes définis par votre fournisseur d'identité centralisé. Vous devez identifier et supprimer les groupes IAM, les IAM users et les informations d'identification utilisateur de longue durée (mots de passe et clés d'accès) dont vous n'avez plus besoin dans vos Comptes AWS. Vous pouvez trouver les informations d'identification non utilisées avec des rapports sur les informations d'identification IAM, supprimer les IAM users correspondants et supprimer les groupes IAM. Vous pouvez appliquer une politique de contrôle des services (SCP) à votre organisation afin d'empêcher la création d'autres groupes et IAM users, en vous assurant que cet accès à AWS se fasse via des identités fédérées.

Conseils pour les utilisateurs de vos applications

Vous pouvez gérer l'identité des utilisateurs de vos applications, telles qu'une application mobile, en utilisant Amazon Cognito en tant que fournisseur d'identité centralisé. Amazon Cognito permet l'authentification, l'autorisation et la gestion des utilisateurs pour vos applications web et mobiles. Amazon Cognito fournit une banque d'identités adaptée à des millions d'utilisateurs, prend en charge la fédération des identités sociales et d'entreprise et propose des fonctionnalités de sécurité avancées pour protéger vos utilisateurs et votre entreprise. Vous pouvez intégrer votre application web ou mobile personnalisée avec Amazon Cognito pour ajouter l'authentification des utilisateurs et le contrôle d'accès à vos applications en quelques minutes. Fondé sur des normes d'identité ouvertes telles que SAML et OpenID Connect (OIDC), Amazon Cognito prend en charge diverses réglementations de conformité et s'intègre aux ressources de développement frontend et backend.

Étapes d'implémentation

Étapes à suivre pour permettre aux utilisateurs en interne d'accéder à AWS

- Fédérez les utilisateurs en interne avec AWS pour qu'ils utilisent un fournisseur d'identité centralisé en utilisant l'une des approches suivantes :
  - Utilisez IAM Identity Center pour activer l'authentification unique à plusieurs Comptes AWS dans votre organisation AWS en vous fédérant avec votre fournisseur d'identité.
  - Utilisez IAM pour connecter votre fournisseur d'identité directement à chaque Compte AWS afin de permettre un accès fédéré précis.
- Identifiez et supprimez les groupes et IAM users qui seront remplacés par des identités fédérées.

Étapes à suivre pour les utilisateurs de vos applications

- Utilisez Amazon Cognito comme fournisseur d'identité centralisé pour vos applications.
- Intégrez vos applications personnalisées à Amazon Cognito à l'aide d'OpenID Connect et d'OAuth. Vous pouvez développer vos applications personnalisées à l'aide des bibliothèques Amplify qui fournissent des interfaces simples à intégrer à divers services AWS, tels que l'authentification Amazon Cognito.

Ressources

Bonnes pratiques Well-Architected connexes :

- SEC02-BP06 Tirer parti des groupes d'utilisateurs et des attributs
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège
- SEC03-BP06 Gérer l'accès en fonction du cycle de vie

Documents connexes :

- Fédération d'identité dans AWS
- Bonnes pratiques de sécurité dans IAM
- Bonnes pratiques AWS Identity and Access Management
- Commencer à utiliser l'administration déléguée IAM Identity Center
- Comment utiliser les politiques gérées par le client dans IAM Identity Center pour les cas d'utilisation avancés

- AWS CLI v2 : fournisseur d'informations d'identification IAM Identity Center

Vidéos connexes :

- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive
- AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center
- AWS re:Invent 2018: Mastering Identity at Every Layer of the Cake

Exemples connexes :

- Atelier : utiliser AWS IAM Identity Center pour assurer une gestion solide de l'identité
- Atelier : identité sans serveur

Outils associés :

- Partenaires AWS disposant de la compétence Sécurité : gestion des identités et des accès
- saml2aws

SEC02-BP05 Contrôler et effectuer régulièrement une rotation des informations d'identification

Contrôlez et effectuez régulièrement une rotation des informations d'identification afin de limiter leur durée d'utilisation pour accéder à vos ressources. Les informations d'identification à long terme créent de nombreux risques qui peuvent être réduits par une rotation régulière de ces informations.

Résultat souhaité : implémenter la rotation des informations d'identification afin de réduire les risques associés à l'utilisation d'informations d'identification à long terme. Auditez et corrigez régulièrement toute non-conformité avec les politiques de rotation des informations d'identification.

Anti-modèles courants :

- Ne pas auditer l'utilisation des informations d'identification.
- Utiliser inutilement des informations d'identification à long terme.
- Utiliser des informations d'identification à long terme et ne pas effectuer de rotation régulièrement.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : moyen

Directives d'implémentation

Lorsque l'utilisation d'informations d'identification temporaires est impossible et que vous avez besoin d'informations d'identification à long terme, vérifiez les informations d'identification pour vous assurer que les contrôles définis, tels que l'authentification multifactorielle (MFA) sont appliqués, changés régulièrement et disposent du niveau d'accès approprié.

La validation régulière, de préférence via un outil automatisé, est nécessaire pour vérifier que les contrôles corrects sont appliqués. Pour les identités humaines, vous devez obliger les utilisateurs à modifier leurs mots de passe régulièrement et à abandonner les clés d'accès au profit d'informations d'identification temporaires. En passant des utilisateurs AWS Identity and Access Management (IAM) aux identités centralisées, vous pouvez générer un rapport des informations d'identification afin d'auditer vos utilisateurs.

Nous vous recommandons également d'appliquer les paramètres d'authentification multifactorielle dans votre fournisseur d'identité. Vous pouvez configurer AWS Config Rules ou utiliser les normes de sécurité AWS Security Hub, afin de surveiller si l'authentification multifactorielle est activée pour les utilisateurs. Envisagez d'utiliser IAM Roles Anywhere afin de fournir des informations d'identification temporaires pour les identités machine. Lorsque l'utilisation de rôles IAM et d'informations d'identification temporaires n'est pas possible, il est nécessaire de réaliser fréquemment des audits et la rotation des clés d'accès.

Étapes d'implémentation

- Auditez régulièrement les informations d'identification : l'audit des identités configurées dans votre fournisseur d'identité et dans IAM permet de s'assurer que seules les identités autorisées ont accès à votre charge de travail. Ces identités peuvent inclure, sans s'y limiter, des utilisateurs IAM, des utilisateurs AWS IAM Identity Center, des utilisateurs Active Directory ou des utilisateurs dans un autre fournisseur d'identité en amont. Par exemple, supprimez les personnes qui quittent l'organisation et supprimez les rôles intercomptes qui ne sont plus requis. Mettez en place un processus pour auditer périodiquement les autorisations aux services auxquels accède une entité IAM. Cela vous permet d'identifier les politiques à modifier afin de supprimer les autorisations inutilisées. Utilisez les rapports d'informations d'identification et AWS Identity and Access Management Access Analyzer pour auditer les informations d'identification et les autorisations IAM. Vous pouvez utiliser Amazon CloudWatch afin de configurer des alarmes pour des appels d'API spécifiques au sein de votre environnement AWS. Amazon GuardDuty peut également vous alerter en cas d'activité inattendue, ce qui peut indiquer un accès trop permissif ou involontaire à des informations d'identification IAM.

- Effectuez une rotation régulière des informations d'identification : lorsque vous ne pouvez pas utiliser d'informations d'identification temporaires, effectuez une rotation régulière des clés d'accès IAM (au maximum tous les 90 jours). Si une clé d'accès est divulguée involontairement à votre insu, cela limite la durée pendant laquelle les informations d'identification peuvent être utilisées pour accéder à vos ressources. Pour plus d'informations sur la rotation des clés d'accès pour les utilisateurs IAM, consultez Rotation des clés d'accès.
- Passez en revue les autorisations IAM : pour améliorer la sécurité de votre Compte AWS, passez en revue et surveillez régulièrement chacune de vos politiques IAM. Vérifiez que les politiques respectent le principe du moindre privilège.
- Envisagez d'automatiser la création et les mises à jour des ressources IAM : IAM Identity Center automatise de nombreuses tâches IAM, telles que la gestion des rôles et des politiques. Sinon, AWS CloudFormation peut être utilisé afin d'automatiser le déploiement des ressources IAM, y compris les rôles et les politiques, afin de réduire le risque d'erreur humaine, car les modèles peuvent être vérifiés et la version contrôlée.
- Utilisez IAM Roles Anywhere pour remplacer les utilisateurs IAM par des identités machine : IAM Roles Anywhere vous permet d'utiliser des rôles dans des domaines où cela était impossible auparavant, par exemple avec les serveurs sur site. IAM Roles Anywhere utilise un certificat X.509 autorisé afin de s'authentifier auprès d'AWS et de recevoir des informations d'identification temporaires. L'utilisation d'IAM Roles Anywhere vous évite d'avoir à effectuer des rotations de ces informations d'identification, car les informations d'identification à long terme ne sont plus stockées dans votre environnement sur site. Veuillez noter que vous devrez surveiller et faire tourner le certificat X.509 à l'approche de son expiration.

Ressources

Bonnes pratiques associées :

- SEC02-BP02 Utiliser des informations d'identification temporaires
- SEC02-BP03 Stocker et utiliser des secrets en toute sécurité

Documents connexes :

- Getting started with AWS Secrets Manager (Démarrer avec Amazon SQS)
- Bonnes pratiques de sécurité dans IAM
- Fournisseurs d'identité et fédération
- Solutions partenaires de sécurité : accès et contrôle d'accès

- Informations d'identification de sécurité temporaires
- Obtenir des rapports d'informations d'identification pour votre Compte AWS

Vidéos connexes :

- Best Practices for Managing, Retrieving, and Rotating Secrets at Scale
- Managing user permissions at scale with AWS IAM Identity Center
- Mastering identity at every layer of the cake

Exemples connexes :

- Atelier Well-Architected – Automated IAM User Cleanup
- Atelier Well-Architected – Automated Deployment of IAM Groups and Roles

SEC02-BP06 Tirer parti des groupes d'utilisateurs et des attributs

Au fur et à mesure que le nombre d'utilisateurs que vous gérez augmente, vous devez déterminer les moyens de les organiser afin de pouvoir les gérer à grande échelle. Placez les utilisateurs ayant des exigences de sécurité communes dans des groupes définis par votre fournisseur d'identité et mettez en place des mécanismes pour s'assurer que les attributs pouvant être utilisés pour le contrôle d'accès (par exemple, service ou emplacement) sont corrects et mis à jour. Utilisez ces groupes et attributs pour contrôler l'accès plutôt que des utilisateurs individuels. Cela vous permet de gérer l'accès de manière centralisée en modifiant une fois l'appartenance à un groupe ou les attributs d'un utilisateur avec un jeu d'autorisations, plutôt que de mettre à jour de nombreuses stratégies individuelles lorsque les besoins d'accès d'un utilisateur changent. Vous pouvez utiliser AWS IAM Identity Center (IAM Identity Center) pour gérer les groupes d'utilisateurs et les attributs. IAM Identity Center prend en charge les attributs les plus couramment utilisés, qu'ils soient saisis manuellement lors de la création de l'utilisateur ou alloués automatiquement à l'aide d'un moteur de synchronisation, tel que défini dans la spécification SCIM (Cross-Domain Identity Management).

Placez les utilisateurs ayant des exigences de sécurité communes dans des groupes définis par votre fournisseur d'identité et mettez en place des mécanismes pour s'assurer que les attributs pouvant être utilisés pour le contrôle d'accès (par exemple, service ou emplacement) sont corrects et mis à jour. Utilisez ces groupes et attributs, plutôt que des utilisateurs individuels, pour contrôler l'accès. Cela vous permet de gérer l'accès de manière centralisée en modifiant l'appartenance à un

groupe ou les attributs d'un utilisateur une fois, plutôt que de mettre à jour de nombreuses stratégies individuelles lorsque les besoins d'accès d'un utilisateur changent.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Si vous utilisez AWS IAM Identity Center (IAM Identity Center), configurez des groupes : IAM Identity Center vous permet de configurer des groupes d'utilisateurs et d'attribuer aux groupes le niveau d'autorisation souhaité.
  - Authentification unique AWS : gérer les identités
- Familiarisez-vous avec le contrôle d'accès basé sur les attributs (ABAC) : l'ABAC est une stratégie d'autorisation qui définit les autorisations en fonction des attributs.
  - Qu'est-ce que l'ABAC pour AWS ?
  - Atelier : Contrôle d'accès basé sur les balises IAM pour EC2

Ressources

Documents connexes :

- Démarrer avec AWS Secrets Manager
- Bonnes pratiques IAM
- Fournisseurs d'identité et fédération
- Utilisateur racine d'un compte AWS

Vidéos connexes :

- Bonnes pratiques de gestion, d'extraction et de renouvellement des secrets à grande échelle
- Gestion des autorisations utilisateur à grande échelle avec AWS IAM Identity Center
- Maîtrise des identités dans chaque couche

Exemples connexes :

- Atelier : Contrôle d'accès basé sur les balises IAM pour EC2

SÉC 3. Comment gérez-vous les autorisations des personnes et des machines ?

Gérez les autorisations des identités de personnes et de machines qui nécessitent un accès à AWS ainsi qu'à votre charge de travail. Les autorisations régissent les ressources accessibles et les conditions d'accès.

Bonnes pratiques

- SEC03-BP01 Définir les conditions d'accès
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège
- SEC03-BP03 Établir un processus d'accès d'urgence
- SEC03-BP04 Limiter les autorisations au minimum requis en permanence
- SEC03-BP05 Définir des protections par autorisation pour votre organisation
- SEC03-BP06 Gérer l'accès en fonction du cycle de vie
- SEC03-BP07 Analyser l'accès public et entre les comptes
- SEC03-BP08 Partager des ressources en toute sécurité au sein de votre organisation
- SEC03-BP09 Partager des ressources en toute sécurité avec un tiers

SEC03-BP01 Définir les conditions d'accès

Les administrateurs, utilisateurs finaux ou autres composants doivent pouvoir accéder à chaque composant ou ressource de votre charge de travail. Définissez clairement qui ou quoi doit avoir accès à chaque composant, choisissez le type d'identité et la méthode d'authentification et d'autorisation appropriés.

Anti-modèles courants :

- Codage en dur ou stockage de secrets dans votre application.
- Octroi d'autorisations personnalisées à chaque utilisateur.
- Utilisation d'informations d'identification durables.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les administrateurs, utilisateurs finaux ou autres composants doivent pouvoir accéder à chaque composant ou ressource de votre charge de travail. Définissez clairement qui ou quoi doit avoir accès

à chaque composant, choisissez le type d'identité et la méthode d'authentification et d'autorisation appropriés.

Un accès standard aux Comptes AWS de l'organisation doit être fourni à l'aide d'un accès fédéré ou d'un fournisseur d'identité centralisé. Vous devez également centraliser la gestion des identités et vous assurer qu'il existe une pratique établie pour intégrer l'accès à AWS au cycle de vie de l'accès des employés. Par exemple, lorsqu'un employé change de poste et de niveau d'accès, son appartenance à un groupe doit également évoluer de façon à refléter les nouvelles conditions d'accès qui lui sont associées.

Lorsque vous définissez des conditions d'accès pour des identités non humaines, déterminez quels applications et composants ont besoin d'un accès et comment les autorisations sont accordées. Dans cette optique, il est recommandé d'utiliser les rôles IAM créés avec le modèle d'accès du moindre privilège. Les politiques gérées par AWS établissent des politiques IAM prédéfinies qui couvrent les cas d'utilisation les plus courants.

Les services AWS, tels qu' AWS Secrets Manager et AWS Systems Manager Parameter Store,peuvent permettre de dissocier les secrets de l'application ou de la charge de travail en toute sécurité lorsqu'il est impossible d'utiliser des rôles IAM. Dans Secrets Manager, vous pouvez établir une rotation automatique de vos informations d'identification. Vous pouvez utiliser Systems Manager de façon à référencer les paramètres dans vos scripts, commandes, documents SSM, configuration et flux de travail d'automatisation en utilisant le nom unique que vous avez spécifié lors de la création du paramètre.

Vous pouvez utiliser des rôles AWS Identity and Access Management partout de façon à obtenir des informations d'identification de sécurité temporaires dans IAM pour les charges de travail exécutées en dehors d'AWS. Vos charges de travail peuvent utiliser les mêmes politiques IAM et rôles IAM que ceux utilisés avec les applications AWS pour accéder aux ressources AWS.

Dans la mesure du possible, privilégiez les informations d'identification temporaires à court terme plutôt que les informations d'identification statiques à long terme. Pour les scénarios dans le cadre desquels les utilisateurs IAM doivent disposer d'un accès par programmation et d'informations d'identification à long terme, utilisez les dernières informations de clé d'accès utilisées pour effectuer la rotation des clés d'accès et supprimer ces dernières.

Ressources

Documents connexes :

- Contrôle d'accès basé sur les attributs (ABAC)

- AWS IAM Identity Center
- IAM Roles Anywhere
- AWS politiques gérées pour IAM Identity Center
- AWS IAM policy conditions
- IAM use cases
- Remove unnecessary credentials
- Gestion des politiques IAM
- How to control access to AWS resources based on Compte AWS, OU, or organization
- Identify, arrange, and manage secrets easily using enhanced search in AWS Secrets Manager

Vidéos connexes :

- Devenir un expert en stratégie IAM en 60 minutes maximum
- Separation of Duties, Least Privilege, Delegation, and CI/CD
- Streamlining identity and access management for innovation

SEC03-BP02 Accorder un accès selon le principe du moindre privilège

Une bonne pratique consiste à accorder uniquement l'accès dont les identités ont besoin pour effectuer des actions spécifiques sur des ressources spécifiques dans des conditions spécifiques. Faites appel à des groupes et des attributs d'identité pour définir de façon dynamique des autorisations à grande échelle, plutôt que pour des utilisateurs individuels. Par exemple, vous pouvez autoriser un groupe de développeurs à gérer uniquement les ressources de leur projet. Ainsi, si un développeur quitte le projet, son accès est automatiquement révoqué sans que les stratégies d'accès sous-jacentes soient modifiées.

Résultat souhaité : les utilisateurs doivent uniquement disposer des autorisations requises pour faire leur travail. Les utilisateurs ne doivent avoir accès qu'aux environnements de production pour effectuer une tâche précise dans un délai limité et cet accès doit être révoqué une fois la tâche terminée. Les autorisations doivent être révoquées lorsqu'elles ne sont plus nécessaires, y compris lorsqu'un utilisateur passe à un autre projet ou à une autre fonction. Les privilèges d'administrateur ne doivent être accordés qu'à un petit groupe d'administrateurs approuvés. Les autorisations doivent être examinées régulièrement pour éviter toute dérive. Les comptes des machines ou des systèmes doivent recevoir le plus petit ensemble d'autorisations nécessaires pour effectuer leurs tâches.

Anti-modèles courants :

- Octroi par défaut des autorisations d'administrateur aux utilisateurs.
- Utilisation de l'utilisateur root pour les activités quotidiennes.
- Création de politiques trop permissives, mais sans tous les privilèges d'administrateur.
- Absence de révision des autorisations pour comprendre si elles autorisent l'accès selon le principe du moindre privilège.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Le principe du moindre privilège établit que les identités ne doivent être autorisées à effectuer que le plus petit ensemble d'actions nécessaires pour accomplir une tâche spécifique. Il permet d'atteindre un équilibre entre la convivialité, l'efficacité et la sécurité. Le respect de ce principe permet de limiter l'accès non intentionnel et de déterminer qui a accès aux ressources. Les utilisateurs et les rôles IAM n'ont aucune autorisation par défaut. L'utilisateur root dispose d'une accès complet par défaut et doit être étroitement contrôle et surveillé. De plus, il doit être utilisé uniquement pour des tâches qui requièrent un accès root.

Les politiques IAM sont utilisées pour octroyer explicitement des autorisations aux rôles IAM ou à des ressources spécifiques. Par exemple, les politiques basées sur l'identité peuvent être attachées à des groupes IAM, tandis que les compartiments S3 peuvent être contrôlés par des politiques basées sur les ressources.

Lorsque vous créez une politique IAM, vous pouvez spécifier les actions de service, les ressources et les conditions qui doivent être remplies pour qu'AWS autorise ou refuse l'accès. AWS prend en charge diverses conditions pour vous aider à limiter l'accès. Par exemple, en utilisant la clé de condition `PrincipalOrgID`, vous pouvez refuser des actions si le demandeur ne fait pas partie de votre AWS Organization.

Vous pouvez également contrôler les demandes effectuées par les services AWS en votre nom, par exemple AWS CloudFormation qui crée une fonction AWS Lambda, en utilisant la clé de condition `CalledVia`. Vous devez superposer différents types de politiques pour établir une défense en profondeur et limiter les autorisations globales de vos utilisateurs. Vous pouvez également limiter les autorisations qui peuvent être accordées et sous quelles conditions. Par exemple, vous pouvez autoriser les équipes de votre application à créer leurs propres politiques IAM pour les systèmes qu'elles créent, mais vous devez également appliquer une limite d'autorisation afin de restreindre les autorisations maximum que le système peut recevoir.

Étapes d'implémentation

- Implémentez des politiques du moindre privilège : attribuez des politiques d'accès avec le moins de privilèges possibles aux groupes et rôles IAM pour rester cohérent avec le rôle ou la fonction de l'utilisateur que vous avez défini.
  - Politiques de base sur l'utilisation des API : pour déterminer les autorisations nécessaires, vous pouvez notamment passer en revue les journaux AWS CloudTrail. Cela vous permet de créer des autorisations adaptées aux actions généralement effectuées par l'utilisateur dans AWS. IAM Access Analyzer peut générer automatiquement une politique IAM basée sur l'activité. Vous pouvez utiliser IAM Access Advisor au niveau de l'organisation ou du compte pour suivre les dernières informations consultées pour une politique particulière.
- Envisagez d'utiliser des politiques gérées par AWS pour les fonctions professionnelles. Lorsque vous commencez à créer des politiques d'autorisations détaillées, il peut être difficile de savoir par où commencer. AWS dispose de politiques gérées pour les rôles professionnels courants, par exemple la facturation, les administrateurs de bases de données et les scientifiques des données. Ces politiques peuvent permettre de restreindre l'accès des utilisateurs en déterminant comment mettre en œuvre les politiques reposant sur le principe du moindre privilège.
- Supprimez les autorisations inutiles : supprimez les autorisations qui ne sont pas nécessaires et réduisez les politiques trop permissives. La génération de politique IAM Access Analyzer peut vous aider à affiner les politiques d'autorisations.
- Assurez-vous que les utilisateurs ont un accès limité aux environnements de production : les utilisateurs ne doivent avoir accès aux environnements de production qu'avec un cas d'utilisation valide. Une fois que l'utilisateur a effectué les tâches précises qui nécessitent un accès en production, cet accès doit être révoqué. Le fait de limiter l'accès aux environnements de production permet de prévenir les événements imprévus ayant une incidence sur la production et réduit la portée des répercussions de l'accès involontaire.
- Envisagez des limites d'autorisations : une limite des autorisations est une fonction qui permet d'utiliser une stratégie gérée définissant les autorisations maximales qu'une entité IAM peut recevoir d'une politique basée sur une identité. La limite des autorisations d'une entité lui permet d'exécuter uniquement les actions autorisées par ses stratégies basées sur l'identité et ses limites d'autorisations.
- Envisagez les balises de ressources pour les autorisations : un modèle de contrôle d'accès basé sur des attributs utilisant des balises de ressources vous permet d'accorder l'accès en fonction de l'objectif de la ressource, du propriétaire, de l'environnement ou d'autres critères. Par exemple, vous pouvez utiliser des balises de ressources pour différencier les environnements de développement et de production. En utilisant ces balises, vous pouvez limiter les développeurs

à l'environnement de développement. En combinant les politiques de balisage et d'autorisations, vous pouvez obtenir un accès précis aux ressources sans avoir à définir des politiques compliquées et personnalisées pour chaque fonction professionnelle.

- Utilisez les politiques de contrôle des services pour AWS Organizations. Les politiques de contrôle des services contrôlent de façon centralisée les autorisations disponibles maximum pour les comptes membres de votre organisation. Il est important de noter que les politiques de contrôle des services vous permettent de limiter les autorisations des utilisateurs root dans les comptes membres. Envisagez également d'utiliser AWS Control Tower, qui fournit des contrôles gérés normatifs permettant d'enrichir AWS Organizations. Vous pouvez également définir vos propres contrôles dans Control Tower.
- Établissez une politique de cycle de vie de l'utilisateur pour votre organisation : les politiques du cycle de vie de l'utilisateur définissent les tâches à effectuer lorsque les utilisateurs sont intégrés à AWS, changent de rôle ou de fonctions, ou qu'ils n'ont plus besoin d'accéder à AWS. Les autorisations doivent être vérifiées à chaque étape du cycle de vie d'un utilisateur pour s'assurer qu'elles sont suffisamment restrictives et éviter les dérives.
- Établissez un calendrier régulier pour passer en revue les autorisations et supprimer les autorisations inutiles : vous devez régulièrement passer en revue les accès utilisateur afin de vérifier que les utilisateurs ne disposent pas d'un accès trop permissif. AWS Config et IAM Access Analyzer peuvent être utiles pour auditer les autorisations utilisateur.
- Établissez une matrice des fonctions : une matrice des fonctions permet de visualiser les différents rôles et les niveaux d'accès requis pour votre empreinte AWS. À l'aide d'une matrice des fonctions, vous pouvez définir et séparer les autorisations en fonction des responsabilités des utilisateurs au sein de votre organisation. Utilisez des groupes au lieu d'appliquer des autorisations directement aux utilisateurs ou rôles individuels.

Ressources

Documents connexes :

- Accorder un accès selon le principe du moindre privilège
- Limites d'autorisations pour les entités IAM
- Techniques for writing least privilege IAM policies
- IAM Access Analyzer makes it easier to implement least privilege permissions by generating IAM policies based on access activity
- Delegate permission management to developers by using IAM permissions boundaries

- Ajustement des autorisations dans AWS à l'aide des dernières informations consultées
- Politiques et autorisations dans IAM
- Test des politiques IAM avec le simulateur de politiques IAM
- Guardrails in AWS Control Tower
- Zero Trust architectures: An AWS perspective
- How to implement the principle of least privilege with CloudFormation StackSets
- Contrôle d'accès basé sur les attributs (ABAC)
- Réduction de la portée de la politique en affichant l'activité des utilisateurs
- Afficher l'accès du rôle
- Use Tagging to Organize Your Environment and Drive Accountability
- AWS Tagging Strategies
- Tagging AWS resources

Vidéos connexes :

- Next-generation permissions management
- Zero Trust: An AWS perspective
- How can I use permissions boundaries to limit users and roles to prevent privilege escalation?

Exemples connexes :

- Atelier : IAM permissions boundaries delegating role creation
- Atelier : IAM tag based access control for EC2

SEC03-BP03 Établir un processus d'accès d'urgence

Élaborez un processus permettant un accès d'urgence à vos charges de travail dans le cas peu probable où un problème avec votre fournisseur d'identité centralisé surviendrait.

Vous devez concevoir des processus pour les différents modes de défaillance susceptibles de provoquer un événement d'urgence. Par exemple, dans des circonstances normales, les utilisateurs en interne se fédèrent au cloud à l'aide d'un fournisseur d'identité centralisé (SEC02-BP04) pour gérer leur charge de travail. Toutefois, si votre fournisseur d'identité centralisé échoue ou si la configuration de la fédération dans le cloud est modifiée, les utilisateurs en interne risquent de ne

pas parvenir à se fédérer dans le cloud. Un processus d'accès d'urgence permet aux administrateurs autorisés d'accéder à vos ressources cloud par d'autres moyens (tels qu'une autre forme de fédération ou un accès utilisateur direct) afin de résoudre les problèmes liés à la configuration de la fédération ou à vos charges de travail. Le processus d'accès d'urgence est utilisé jusqu'à ce que le mécanisme de fédération normal soit rétabli.

Résultat souhaité :

- Vous avez défini et documenté les modes de défaillance considérés comme une urgence : envisagez les circonstances habituelles et les systèmes dont dépendent vos utilisateurs pour gérer leurs charges de travail. Réfléchissez à la façon dont chacune de ces dépendances peut échouer et provoquer une situation d'urgence. Les questions et les bonnes pratiques du pilier Fiabilité vous aideront à identifier les modes de défaillance et à concevoir des systèmes plus résilients afin de minimiser le risque de défaillance.
- Vous avez documenté les étapes à suivre pour confirmer qu'une défaillance est une urgence. Par exemple, vous pouvez demander aux administrateurs d'identité de vérifier l'état des fournisseurs d'identité principal et secondaire et, si les deux ne sont pas disponibles, de déclarer un événement d'urgence pour cause de défaillance du fournisseur d'identité.
- Vous avez défini un processus d'accès d'urgence spécifique à chaque type d'urgence ou de mode de défaillance. En étant aussi précis que possible, vous éviterez que les utilisateurs abusent d'un processus général pour tous les types d'urgence. Vos processus d'accès d'urgence décrivent les circonstances dans lesquelles chaque processus doit être utilisé, et inversement les situations dans lesquelles le processus ne doit pas être utilisé et renvoie à d'autres processus qui peuvent s'appliquer.
- Vos processus sont bien documentés avec des instructions détaillées et des playbooks qui peuvent être suivis rapidement et efficacement. N'oubliez pas qu'un événement d'urgence peut être stressant pour vos utilisateurs et qu'ils peuvent être soumis à des contraintes de temps extrêmes. Concevez donc votre processus de manière à ce qu'il soit aussi simple que possible.

Anti-modèles courants :

- Vous ne disposez pas de processus d'accès d'urgence bien documentés et bien testés. Vos utilisateurs ne sont pas préparés à une situation d'urgence et suivent des processus improvisés lorsqu'une situation d'urgence survient.
- Vos processus d'accès d'urgence dépendent des mêmes systèmes (tels qu'un fournisseur d'identité centralisé) que vos mécanismes d'accès habituels. Autrement dit, la défaillance d'un

système de ce type peut avoir un impact à la fois sur vos mécanismes d'accès habituels et sur les mécanismes d'accès d'urgence, et nuire à votre capacité à vous remettre de la panne.

- Vos processus d'accès d'urgence sont utilisés dans des situations non urgentes. Par exemple, vos utilisateurs utilisent fréquemment à mauvais escient les processus d'accès d'urgence, car ils trouvent qu'il est plus facile d'apporter des modifications directement que de les soumettre par le biais d'un pipeline.
- Vos processus d'accès d'urgence ne génèrent pas suffisamment de journaux pour auditer les processus, ou les journaux ne sont pas surveillés pour signaler une éventuelle utilisation inappropriée des processus.

Avantages liés au respect de cette bonne pratique :

- En disposant de processus d'accès d'urgence bien documentés et bien testés, vous réduisez le temps nécessaire à vos utilisateurs pour répondre à un événement d'urgence et le résoudre. Cela peut se traduire par une réduction des temps d'arrêt et une meilleure disponibilité des services que vous offrez à vos clients.
- Vous pouvez suivre chaque demande d'accès d'urgence, détecter les tentatives non autorisées d'utilisation abusive du processus pour des événements non urgents et les signaler.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Cette section fournit des conseils pour créer des processus d'accès d'urgence pour plusieurs modes de défaillance liés aux charges de travail déployées sur AWS, en commençant par des conseils communs applicables à tous les modes de défaillance, suivies de directives spécifiques basées sur le type de mode de défaillance.

Conseils communs pour tous les modes de défaillance

Envisagez les points suivants lorsque vous concevez un processus d'accès d'urgence pour un mode de défaillance :

- Documentez les conditions préalables et les hypothèses du processus : situations dans lesquelles le processus doit être utilisé et situations dans lesquelles il ne doit pas être utilisé. Il est utile de détailler le mode de défaillance et de documenter les hypothèses, telles que l'état d'autres systèmes connexes. Par exemple, le processus du mode de défaillance 2 suppose que le fournisseur d'identité est disponible, mais que la configuration sur AWS est modifiée ou a expiré.

- Créez au préalable les ressources nécessaires au processus d'accès d'urgence (SEC10-BP05). Par exemple, créez au préalable le Compte AWS d'accès d'urgence avec les rôles et IAM users, ainsi que les rôles IAM entre comptes dans tous les comptes de la charge de travail. Vous pourrez ainsi vérifier que ces ressources sont prêtes et disponibles en cas d'urgence. En créant des ressources au préalable, vous n'êtes pas tributaire des API de plan de contrôle AWS (utilisées pour créer et modifier les ressources AWS) qui peuvent ne pas être disponibles en cas d'urgence. De plus, en créant au préalable des ressources IAM, vous n'avez pas besoin de prendre en compte les retards potentiels dus à la cohérence à terme.
- Incluez les processus d'accès d'urgence dans vos plans de gestion des incidents (SEC10-BP02). Documentez la manière dont les événements d'urgence sont suivis et communiqués aux autres membres de votre organisation (tels que vos pairs et la direction) et, le cas échéant, à vos clients et partenaires commerciaux.
- Définissez le processus de demande d'accès d'urgence dans votre système de flux de travail des demandes de service existant, si vous en avez un. Généralement, ces systèmes de flux de travail vous permettent de créer des formulaires de réception pour collecter des informations sur la demande, de suivre la demande à chaque étape du flux de travail et d'ajouter des étapes d'approbation automatisées et manuelles. Associez chaque demande à un événement d'urgence correspondant suivi dans votre système de gestion des incidents. Le fait de disposer d'un système uniforme pour les accès d'urgence vous permet de suivre ces demandes dans un seul système, d'analyser les tendances d'utilisation et d'améliorer vos processus.
- Vérifiez que vos processus d'accès d'urgence ne peuvent être initiés que par des utilisateurs autorisés et nécessitent l'approbation de pairs ou de la direction de l'utilisateur, le cas échéant. Le processus d'approbation doit fonctionner efficacement pendant les heures de bureau et au-delà. Définissez comment les demandes d'approbation autorisent les approbateurs secondaires si les approbateurs principaux ne sont pas disponibles et comment elles remontent dans la chaîne de gestion jusqu'à ce qu'elles soient approuvées.
- Vérifiez que le processus génère des journaux d'audit et des événements détaillés pour les tentatives d'accès d'urgence qui aboutissent et pour celles qui échouent. Surveillez à la fois le processus de demande et le mécanisme d'accès d'urgence pour détecter les abus ou les accès non autorisés. Corrélez l'activité avec les événements d'urgence en cours depuis votre système de gestion des incidents et signalez les situations où des actions se produisent en dehors des périodes prévues. Par exemple, vous devez surveiller le Compte AWS d'accès d'urgence et signaler toute activité, car il ne doit jamais être utilisé dans le cadre des opérations habituelles.
- Testez régulièrement les processus d'accès d'urgence pour vérifier que les étapes sont claires et accorder le niveau d'accès approprié rapidement et efficacement. Vos processus d'accès

d'urgence doivent être testés dans le cadre de simulations de réponse aux incidents (SEC10-BP07) et de tests de reprise après sinistre (REL13-BP03).

Mode de défaillance 1 : le fournisseur d'identité utilisé pour la fédération à AWS n'est pas disponible

Comme décrit dans SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé, nous vous recommandons de faire appel à un fournisseur d'identité centralisé pour fédérer les utilisateurs en interne et accorder l'accès aux Comptes AWS. Vous pouvez fédérer les utilisateurs à plusieurs Comptes AWS au sein de votre organisation AWS à l'aide de IAM Identity Center, ou vous pouvez les fédérer à des Comptes AWS individuels avec IAM. Dans les deux cas, les utilisateurs en interne s'authentifient auprès de votre fournisseur d'identité centralisé avant d'être redirigés vers un point de terminaison de connexion AWS pour l'authentification unique.

Dans le cas peu probable où votre fournisseur d'identité centralisé ne serait pas disponible, les utilisateurs en interne ne pourraient pas se fédérer aux Comptes AWS ni gérer leurs charges de travail. Dans ce cas d'urgence, vous pouvez fournir un processus d'accès d'urgence permettant à un petit groupe d'administrateurs d'accéder aux Comptes AWS pour effectuer des tâches critiques qui ne peuvent pas attendre que vos fournisseurs d'identité centralisés soient de nouveau disponibles. Par exemple, votre fournisseur d'identité n'est pas disponible pendant 4 heures et, durant cette période, vous devez modifier les limites supérieures d'un groupe Amazon EC2 Auto Scaling dans un compte de production pour faire face à un pic inattendu du trafic client. Vos administrateurs d'urgence doivent suivre le processus d'accès d'urgence pour accéder au Compte AWS de production spécifique et apporter les modifications nécessaires.

Le processus d'accès d'urgence repose sur un Compte AWS d'accès d'urgence créé au préalable, qui est utilisé uniquement pour l'accès d'urgence et dispose de ressources AWS (comme les rôles IAM et les IAM users) pour soutenir le processus d'accès d'urgence. Pendant les opérations normales, personne ne doit accéder au compte d'accès d'urgence et vous devez surveiller et signaler tout cas d'utilisation abusive de ce compte (pour plus de détails, consultez la section précédente consacrée aux conseils communs).

Le compte d'accès d'urgence possède des rôles IAM d'accès d'urgence autorisés à endosser des rôles entre comptes dans les Comptes AWS nécessitant un accès d'urgence. Ces rôles IAM sont créés au préalable et configurés avec des politiques d'approbation qui assurent la validité des rôles IAM du compte d'urgence.

Le processus d'accès d'urgence peut utiliser l'une des approches suivantes :

- Vous pouvez créer au préalable un ensemble d'IAM users pour vos administrateurs d'urgence dans le compte d'accès d'urgence avec des mots de passe forts et des jetons MFA associés. Ces IAM users seront autorisés à endosser les rôles IAM qui autoriseront ensuite l'accès intercompte au Compte AWS où un accès d'urgence est requis. Nous vous recommandons de créer le moins d'utilisateurs possible et d'affecter chaque utilisateur à un seul administrateur d'urgence. En cas d'urgence, un utilisateur administrateur d'urgence se connecte au compte d'accès d'urgence à l'aide de son mot de passe et de son code de jeton MFA, passe au rôle IAM d'accès d'urgence dans le compte d'urgence, puis passe au rôle IAM d'accès d'urgence dans le compte de la charge de travail pour effectuer l'action de modification d'urgence. L'avantage de cette approche est que chaque IAM user est associé à un seul administrateur d'urgence et que vous pouvez savoir quel utilisateur s'est connecté en consultant les événements CloudTrail. L'inconvénient est que vous devez gérer plusieurs IAM users avec leurs mots de passe de longue durée de vie et leurs jetons MFA associés.
- Vous pouvez utiliser l'utilisateur root du Compte AWS d'accès d'urgence pour vous connecter au compte d'accès d'urgence, endosser le rôle IAM d'accès d'urgence et endosser le rôle entre comptes dans le compte de la charge de travail. Nous recommandons de définir un mot de passe fort et plusieurs jetons MFA pour l'utilisateur root. Nous conseillons également de stocker le mot de passe et les jetons MFA dans un coffre-fort d'informations d'identification d'entreprise sécurisé qui applique des mécanismes solides d'authentification et d'autorisation. Vous devez sécuriser les facteurs de réinitialisation des mots de passe et des jetons MFA : configurez l'adresse e-mail du compte sur une liste de distribution surveillée par vos administrateurs de sécurité cloud, et le numéro de téléphone du compte doit être un numéro partagé également surveillé par ces administrateurs. L'avantage de cette approche est qu'il n'existe qu'un seul ensemble d'informations d'identification d'utilisateur root à gérer. L'inconvénient est qu'étant donné qu'il s'agit d'un utilisateur partagé, plusieurs administrateurs ont la possibilité de se connecter en tant qu'utilisateur root. Vous devez auditer les événements de journal de votre coffre-fort d'entreprise pour identifier quel administrateur a extrait le mot de passe de l'utilisateur root.

Mode de défaillance 2 : la configuration du fournisseur d'identité sur AWS est modifiée ou a expiré

Pour permettre aux utilisateurs en interne de se fédérer aux Comptes AWS, vous pouvez configurer la IAM Identity Center auprès d'un fournisseur d'identité externe ou créer un fournisseur d'identité IAM (SEC02-BP04). Généralement, vous les configurez en important un document XML de métadonnées SAML fourni par votre fournisseur d'identité. Ce document XML de métadonnées inclut un certificat X.509 correspondant à une clé privée que le fournisseur d'identité utilise pour signer ses assertions SAML.

Ces configurations côté AWS peuvent être modifiées ou supprimées par erreur par un administrateur. Dans un autre scénario, le certificat X.509 importé dans AWS peut expirer, et aucun nouveau fichier XML de métadonnées contenant un nouveau certificat n'a encore été importé dans AWS. Ces deux scénarios peuvent désactiver la fédération des utilisateurs en interne à AWS, ce qui peut entraîner une situation d'urgence.

Dans un tel cas d'urgence, vous pouvez fournir à vos administrateurs d'identité un accès à AWS pour résoudre les problèmes de fédération. Par exemple, votre administrateur d'identité utilisera le processus d'accès d'urgence pour se connecter au Compte AWS d'accès d'urgence, passera à un rôle dans le compte administrateur d'Identity Center et mettra à jour la configuration du fournisseur d'identité externe en important le dernier document XML de métadonnées SAML de votre fournisseur d'identité afin de réactiver la fédération. Une fois la fédération rétablie, les utilisateurs en interne pourront continuer à utiliser le processus d'exploitation habituel pour se fédérer aux comptes de leur charge de travail.

Vous pouvez suivre les approches détaillées dans le précédent mode de défaillance 1 pour créer un processus d'accès d'urgence. Vous pouvez accorder des autorisations de moindre privilège aux administrateurs d'identité pour qu'ils ne puissent accéder qu'au compte administrateur d'Identity Center et effectuer des actions sur Identity Center dans ce compte uniquement.

Mode de défaillance 3 : interruption d'Identity Center

Dans le cas peu probable où une Région AWS ou une connexion IAM Identity Center serait interrompue, nous vous recommandons de créer une configuration que vous pourrez utiliser pour assurer un accès temporaire à la AWS Management Console.

Le processus d'accès d'urgence utilise une fédération directe entre votre fournisseur d'identité et IAM dans un compte d'urgence. Pour plus de détails sur le processus et les considérations de conception, voir Configurer un accès d'urgence à la AWS Management Console.

Étapes d'implémentation

Étapes communes pour tous les modes de défaillance

- Créez un Compte AWS dédié aux processus d'accès d'urgence. Créez au préalable les ressources IAM nécessaires dans le compte, telles que les rôles IAM ou les IAM users et, éventuellement, les fournisseurs d'identité IAM. En outre, créez au préalable des rôles IAM entre comptes dans les Comptes AWS de la charge de travail avec des relations d'approbation avec les rôles IAM correspondants dans le compte d'accès d'urgence. Vous pouvez utiliser AWS CloudFormation

StackSets avec AWS Organizations pour créer ces ressources dans les comptes membres de votre organisation.

- Créez des politiques de contrôle des services AWS Organizations (SCP) pour refuser la suppression et la modification des rôles IAM entre comptes dans les Comptes AWS membres.
- Activez CloudTrail pour le Compte AWS d'accès d'urgence et envoyez les événements de suivi vers un compartiment S3 central du Compte AWS de collecte de journaux. Si vous utilisez AWS Control Tower pour configurer et gérer votre environnement AWS multi-comptes, chaque compte que vous créez avec AWS Control Tower ou que vous inscrivez dans AWS Control Tower est activé pour CloudTrail par défaut et envoyé vers un compartiment S3 dans un Compte AWS d'archive de journal dédié.
- Surveillez l'activité du compte d'accès d'urgence en créant des règles EventBridge qui correspondent lors de la connexion à la console et de l'activité de l'API en fonction des rôles IAM d'urgence. Envoyez des notifications à votre centre des opérations de sécurité lorsque des activités se produisent en dehors d'un événement d'urgence en cours suivi dans votre système de gestion des incidents.

Étapes supplémentaires pour le mode de défaillance 1 : le fournisseur d'identité utilisé pour la fédération à AWS n'est pas disponible et pour le mode de défaillance 2 : la configuration du fournisseur d'identité sur AWS est modifiée ou a expiré

- Créez des ressources au préalable en fonction du mécanisme que vous avez choisi pour l'accès d'urgence :
  - Avec les IAM users: créez au préalable les IAM users avec des mots de passe forts et les dispositifs MFA associés.
  - Avec l'utilisateur root du compte d'urgence : configurez l'utilisateur root avec un mot de passe fort et stockez ce mot de passe dans le coffre-fort d'informations d'identification de votre entreprise. Associez plusieurs appareils MFA physiques à l'utilisateur root et stockez-les à des emplacements auxquels les membres de votre équipe d'administrateurs d'urgence peuvent accéder rapidement.

Étapes supplémentaires pour le mode de défaillance 3 : interruption d'Identity Center

- Comme indiqué dans Configurer un accès d'urgence à la AWS Management Console, dans le Compte AWS d'accès d'urgence, créez un fournisseur d'identité IAM pour activer la fédération SAML directe à partir de votre fournisseur d'identité.
- Créez des groupes d'opérations d'urgence dans votre fournisseur d'identité sans aucun membre.
- Créez des rôles IAM correspondant aux groupes d'opérations d'urgence dans le compte d'accès d'urgence.

Ressources

Bonnes pratiques Well-Architected connexes :

- SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège
- SEC10-BP02 Développer des plans de gestion des incidents
- SEC10-BP07 Organiser des jeux de rôle

Documents connexes :

- Configurer un accès d'urgence à la AWS Management Console
- Permettre aux utilisateurs fédérés SAML 2.0 d'accéder à la AWS Management Console
- Accès en mode « bris de glace »

Vidéos connexes :

- AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center
- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive

Exemples connexes :

- AWS Break Glass Role
- AWS customer playbook framework
- AWS incident response playbook samples

SEC03-BP04 Limiter les autorisations au minimum requis en permanence

Au fur et à mesure que vos équipes déterminent les accès nécessaires, supprimez les autorisations inutiles et mettez en place des processus de révision afin d'obtenir des autorisations de moindre privilège. Surveillez et supprimez en permanence les identités et autorisations inutilisées pour les accès humains et machine.

Résultat souhaité : les politiques d'autorisation doivent respecter le principe du moindre privilège. Au fur et à mesure que les tâches et les rôles sont mieux définis, vos politiques d'autorisation doivent être revues de façon à supprimer les autorisations inutiles. Cette approche réduit l'impact si les informations d'identification sont exposées par inadvertance ou autrement consultées sans autorisation.

Anti-modèles courants :

- Octroi par défaut des autorisations d'administrateur aux utilisateurs.
- Création de politiques trop permissives, mais sans tous les privilèges d'administrateur.
- Maintien des politiques d'autorisation une fois qu'elles ne sont plus nécessaires.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : moyen

Directives d'implémentation

Lorsque les équipes et les projets ne font que commencer, des politiques d'autorisation permissives peuvent être utilisées pour favoriser l'innovation et l'agilité. Par exemple, dans un environnement de développement ou de test, les développeurs peuvent se voir octroyer un accès à un large éventail de services AWS. Nous vous recommandons d'évaluer l'accès en continu et de restreindre l'accès aux services et aux actions de service nécessaires pour effectuer le travail en cours. Nous recommandons cette évaluation pour les identités humaines et machine. Les identités machine, parfois appelées comptes de système ou de service, donnent un accès AWS aux applications ou aux serveurs. Cet accès est particulièrement important dans un environnement de production, où des autorisations trop permissives peuvent avoir un impact important et exposer les données des clients.

AWS fournit plusieurs méthodes pour identifier les utilisateurs, les rôles, les autorisations et les informations d'identification inutilisés. AWS peut également faciliter l'analyse de l'activité d'accès des utilisateurs et rôles IAM, notamment des clés associées, ainsi que l'accès aux ressources AWS telles que les objets dans les compartiments Amazon S3. La génération de politiques AWS Identity and Access Management Access Analyzer peut vous aider à créer des politiques d'autorisations restrictives basées sur les services et les actions réels avec lesquels un principal

interagit. Le contrôle d'accès basé sur les attributs (ABAC) peut permettre de simplifier la gestion des autorisations, car vous pouvez fournir des autorisations aux utilisateurs en utilisant leurs attributs au lieu d'associer des politiques d'autorisations directement à chaque utilisateur.

Étapes d'implémentation

- Utilisez AWS Identity and Access Management Access Analyzer : IAM Access Analyzer permet d'identifier les ressources de votre organisation et des comptes, comme les compartiments Amazon Simple Storage Service (Amazon S3) ou les rôles IAM qui sont partagés avec une entité externe.
- Utilisez la génération de politiques IAM Access Analyzer : la génération de politiques IAM Access Analyzer vous permet de créer des politiques d'autorisation détaillées basées sur l'activité d'accès d'un utilisateur ou d'un rôle IAM.
- Déterminez un calendrier et une politique d'utilisation acceptables pour les utilisateurs et les rôles IAM : utilisez l'horodatage des derniers accès pour identifier les utilisateurs et les rôles inutilisés et les supprimer. Passez en revue les informations relatives aux services et actions consultés en dernier afin d'identifier et de restreindre les autorisations à des utilisateurs et des rôles spécifiques. Par exemple, vous pouvez utiliser les dernières informations consultées pour identifier les actions Amazon S3 spécifiques dont votre rôle d'application a besoin et limiter l'accès du rôle à celles-ci uniquement. Ces fonctionnalités relatives aux informations sur les derniers accès sont disponibles dans la AWS Management Console et par programmation pour vous permettre de les intégrer dans vos flux de travail d'infrastructure et vos outils automatisés.
- Envisagez de consigner les événements de données dans AWS CloudTrail : par défaut, CloudTrail ne consigne pas les événements de données tels que l'activité au niveau des objets Amazon S3 (par exemple, `GetObject` et `DeleteObject`) ou les activités de table Amazon DynamoDB (par exemple, `PutItem` et `DeleteItem`). Envisagez d'autoriser la journalisation de ces événements afin de déterminer quels utilisateurs et rôles ont besoin d'accéder à des objets Amazon S3 ou des éléments de table DynamoDB spécifiques.

Ressources

Documents connexes :

- Grant least privilege
- Remove unnecessary credentials
- Qu'est-ce qu'AWS CloudTrail ?
- Gestion des politiques IAM

- Journalisation et surveillance dans DynamoDB
- Activation de la journalisation des événements CloudTrail pour les compartiments et les objets Amazon S3
- Obtenir des rapports d'informations d'identification pour votre Compte AWS

Vidéos connexes :

- Become an IAM Policy Master in 60 Minutes or Less
- Separation of Duties, Least Privilege, Delegation, and CI/CD
- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive

SEC03-BP05 Définir des protections par autorisation pour votre organisation

Établissez des contrôles communs qui limitent l'accès à toutes les identités de votre organisation. Par exemple, vous pouvez restreindre l'accès à des Régions AWS spécifiques ou empêcher vos techniciens de supprimer des ressources communes, telles qu'un rôle IAM utilisé pour votre équipe de sécurité centrale.

Anti-modèles courants :

- Exécution des charges de travail dans votre compte d'administrateur organisationnel.
- Exécution des charges de travail de production et autres dans le même compte.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Au fur et à mesure que vous développez et gérez des charges de travail supplémentaires dans AWS, vous devez séparer ces charges de travail à l'aide de comptes et gérer ces comptes à l'aide d'AWS Organizations. Nous vous recommandons d'établir des protections par autorisation communes qui limitent l'accès à toutes les identités de votre organisation. Par exemple, vous pouvez restreindre l'accès à des Régions AWS spécifiques ou empêcher votre équipe de supprimer des ressources communes, telles qu'un rôle IAM utilisé par votre équipe de sécurité centrale.

Vous pouvez commencer en implémentant des exemples de politiques de contrôle des services, par exemple en empêchant les utilisateurs de désactiver les services clés. Les SCP utilisent le langage de politique IAM et vous permettent d'établir des contrôles auxquels tous les principaux (utilisateurs

et rôles) IAM adhèrent. Vous pouvez restreindre l'accès à des actions de service spécifiques, à des ressources et en fonction de conditions spécifiques pour répondre aux besoins de contrôle d'accès de votre organisation. Si nécessaire, vous pouvez définir des exceptions à vos barrières de protection. Par exemple, vous pouvez restreindre les actions de service pour toutes les entités IAM du compte, à l'exception d'un rôle d'administrateur spécifique.

Nous vous déconseillons l'exécution de vos charges de travail dans votre compte de gestion. Le compte de gestion doit être utilisé afin de gouverner et déployer des barrières de protection en matière de sécurité qui affecteront les comptes des membres. Certains services AWS prennent en charge l'utilisation d'un compte d'administrateur délégué. Lorsqu'il est disponible, nous vous recommandons d'utiliser ce compte délégué plutôt que le compte de gestion. Vous devez limiter fortement l'accès au compte d'administrateur organisationnel.

La mise en place d'une stratégie multicompte vous permet de bénéficier d'une plus grande flexibilité dans l'application de barrières de protection à vos charges de travail. L'AWS Security Reference Architecture propose des conseils normatifs en ce qui concerne la conception de la structure de votre compte. Les services AWS tels qu'AWS Control Tower offrent des capacités de gestion centralisée des contrôles préventifs et de détection au sein de votre organisation. Définissez un objectif clair pour chaque compte ou unité opérationnelle au sein de votre organisation et limitez les contrôles conformément à cet objectif.

Ressources

Documents connexes :

- AWS Organizations
- Politiques de contrôle de service (SCP)
- Get more out of service control policies in a multi-account environment
- AWS Architecture de référence de sécurité (AWS SRA)

Vidéos connexes :

- Enforce Preventive Guardrails using Service Control Policies
- Building governance at scale with AWS Control Tower
- AWS Identity and Access Management deep dive

SEC03-BP06 Gérer l'accès en fonction du cycle de vie

Intégrez les contrôles d'accès au cycle de vie des opérateurs et des applications et à votre fournisseur de fédération centralisée. Par exemple, supprimez l'accès d'un utilisateur lorsqu'il quitte l'organisation ou change de poste.

Lorsque vous gérez des charges de travail à l'aide de comptes distincts, vous devrez partager des ressources entre ces comptes. Nous vous recommandons de partager des ressources à l'aide d' AWS Resource Access Manager (AWS RAM). Ce service vous permet de partager facilement et en toute sécurité des ressources AWS au sein de votre AWS Organizations et de vos unités d'organisation. Avec AWS RAM, l'accès aux ressources partagées est automatiquement accordé ou révoqué lorsque les comptes sont déplacés vers et hors de l'organisation ou de l'unité d'organisation avec laquelle ils sont partagés. Cela permet de vous assurer que les ressources sont uniquement partagées avec les comptes que vous souhaitez.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

Cycle de vie de l'accès utilisateur : implémentez une stratégie de cycle de vie d'accès utilisateur pour les nouveaux utilisateurs qui rejoignent l'entreprise, les changements de poste et les utilisateurs qui quittent l'entreprise, afin que seuls les utilisateurs actifs disposent d'un accès approprié.

Ressources

Documents connexes :

- Contrôle d'accès basé sur les attributs (ABAC)
- Accorder le privilège le plus faible
- IAM Access Analyzer
- Suppression des informations d'identification inutiles
- Travailler avec des stratégies

Vidéos connexes :

- Devenir un expert en stratégie IAM en 60 minutes maximum
- Séparation des responsabilités, moindre privilège, délégation et CI/CD

SEC03-BP07 Analyser l'accès public et entre les comptes

Surveillez en continu les résultats qui mettent en évidence l'accès public et intercompte. Limitez l'accès public et intercompte uniquement aux ressources spécifiques qui requièrent ce type d'accès.

Résultat souhaité : savoir quelles ressources AWS sont partagées et avec qui. Surveillez et auditez continuellement vos ressources partagées afin de vérifier qu'elles ne sont partagées qu'avec les principaux autorisés.

Anti-modèles courants :

- Ne pas tenir un inventaire des ressources partagées.
- Ne pas suivre de processus pour régir l'accès intercompte et public aux ressources.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : faible

Directives d'implémentation

Si votre compte est dans AWS Organizations, vous pouvez accorder l'accès aux ressources à l'ensemble de l'organisation, à des unités d'organisation spécifiques ou à des comptes individuels. Si votre compte n'est pas membre d'une organisation, vous pouvez partager des ressources avec des comptes individuels. Vous pouvez accorder un accès direct intercompte à l'aide de politiques axées sur les ressources, par exemple les politiques de compartiment Amazon Simple Storage Service (Amazon S3), ou en autorisant un principal dans un autre compte à assumer un rôle IAM dans votre compte. Lorsque vous utilisez des politiques de ressources, vérifiez que l'accès n'est accordé qu'aux principaux autorisés. Définissez un processus d'approbation de toutes les ressources qui doivent être accessibles au public.

AWS Identity and Access Management Access Analyzer utilise une sécurité prouvable pour identifier tous les chemins d'accès à une ressource depuis l'extérieur de son compte. Il passe en revue les stratégies de ressources en continu et présente les résultats d'accès public et intercompte pour vous permettre d'analyser facilement un accès potentiellement étendu. Envisagez de configurer IAM Access Analyzer avec AWS Organizations afin de vérifier que vous avez une visibilité sur tous vos comptes. IAM Access Analyzer vous permet également de prévisualiser les résultats avant de déployer les autorisations des ressources. Vous pouvez ainsi vérifier que vos modifications de politique n'accordent que l'accès public et intercompte prévu à vos ressources. Lors de la conception pour un accès multicompte, vous pouvez utiliser les politiques d'approbation afin de contrôler dans quels cas un rôle peut être assumé. Par exemple, vous pouvez utiliser la clé de condition

`PrincipalOrgId` pour refuser une tentative d'assumer un rôle depuis l'extérieur de votre AWS Organizations.

AWS Config peut signaler les ressources qui sont mal configurées et, via des contrôles de politique AWS Config, il peut détecter les ressources pour lesquelles un accès public est configuré. Des services tels que AWS Control Tower et AWS Security Hub simplifient le déploiement des contrôles de détection et des barrières de protection dans AWS Organizations afin d'identifier et de résoudre les problèmes des ressources exposées au public. Par exemple, AWS Control Tower a une barrière de protection gérée qui peut détecter si des instantanés Amazon EBS peuvent être restaurés par des Comptes AWS.

Étapes d'implémentation

- Envisagez d'activer AWS Config pour AWS Organizations : AWS Config vous permet de regrouper les résultats de plusieurs comptes d'un AWS Organizations dans un compte d'administrateur délégué. Cela fournit une vue d'ensemble et vous permet de déployer AWS Config Rules sur plusieurs comptes afin de détecter les ressources accessibles publiquement.
- Configurez AWS Identity and Access Management Access Analyzer IAM Access Analyzer vous permet d'identifier les ressources de votre organisation et les comptes, par exemple les compartiments Amazon S3 ou les rôles IAM qui sont partagés avec une entité externe.
- Utilisez l'atténuation automatique dans AWS Config pour répondre aux changements apportés à la configuration de l'accès public des compartiments Amazon S3 : Vous pouvez réactiver automatiquement les paramètres d'accès public du bloc pour les compartiments Amazon S3.
- Implémentez la surveillance et les alertes afin de déterminer si les compartiments Amazon S3 sont devenus publics : vous devez mettre en place la surveillance et les alertes pour identifier si le Blocage de l'accès public Amazon S3 est désactivé et si les compartiments Amazon S3 deviennent publics. De plus, si vous utilisez AWS Organizations, vous pouvez créer une politique de contrôle des services qui empêche les modifications des politiques d'accès public Amazon S3. AWS Trusted Advisor vérifie les compartiments Amazon S3 qui ont des autorisations d'accès ouvert. Les autorisations de compartiment qui accordent à tous un accès au chargement ou à la suppression créent des problèmes de sécurité potentiels, en permettant à quiconque d'ajouter, de modifier ou de supprimer les éléments d'un compartiment. La vérification Trusted Advisor examine les autorisations explicites de compartiment et les politiques associées de compartiment susceptibles de remplacer les autorisations du compartiment. Vous pouvez également utiliser AWS Config pour surveiller l'accès public de vos compartiments Amazon S3. Pour plus d'informations, consultez How to Use AWS Config to Monitor for and Respond to Amazon S3 Buckets Allowing Public Access. Lors de l'examen de l'accès, il est important de tenir compte des types de données contenus dans

les compartiments Amazon S3. Amazon Macie permet de découvrir et de protéger les données sensibles, comme les PII, les PHI et les informations d'identification, dont les clés privées ou AWS.

Ressources

Documents connexes :

- Utiliser AWS Identity and Access Management Access Analyzer
- Bibliothèque des contrôles AWS Control Tower
- AWS Foundational Security Best Practices standard (Normes concernant les bonnes pratiques de sécurité de base AWS)
- AWS Config Managed Rules (Règles gérées AWS Config)
- Référence de la vérification AWS Trusted Advisor
- Surveillance des résultats de vérification AWS Trusted Advisor avec Amazon EventBridge
- Managing AWS Config Rules Across All Accounts in Your Organization (Gestion des règles AWS Config pour tous les comptes de votre organisation)
- AWS Config et AWS Organizations

Vidéos connexes :

- Best Practices for securing your multi-account environment
- Dive Deep into IAM Access Analyzer

SEC03-BP08 Partager des ressources en toute sécurité au sein de votre organisation

À mesure que le nombre de charges de travail augmente, vous devrez peut-être partager l'accès aux ressources de ces charges de travail ou fournir les ressources plusieurs fois pour plusieurs comptes. Vous pouvez utiliser des constructions pour compartimenter votre environnement, par exemple des environnements de développement, de test et de production. Cependant, le fait d'avoir des constructions distinctes ne vous empêche pas de partager en toute sécurité. En partageant des composants qui se chevauchent, vous pouvez réduire les frais d'exploitation et offrir une expérience cohérente sans avoir à deviner ce que vous avez pu manquer en créant la même ressource plusieurs fois.

Résultat souhaité : limiter autant que possible les accès involontaires en utilisant des méthodes sécurisées pour partager des ressources au sein de votre organisation, et vous aider dans le cadre

de votre initiative de prévention de la perte de données. Réduisez vos frais généraux opérationnels par rapport à la gestion de composants individuels, réduisez les erreurs liées à la création manuelle du même composant plusieurs fois et augmentez la capacité de mise à l'échelle de vos charges de travail. Vous pouvez bénéficier d'une réduction du délai de résolution dans les scénarios de défaillance multipoints et augmenter votre confiance dans l'évaluation du moment où un composant n'est plus nécessaire. Pour des conseils normatifs sur l'analyse des ressources partagées en externe, consultez SEC03-BP07 Analyser l'accès public et entre les comptes.

Anti-modèles courants :

- Manque de processus pour surveiller continuellement et alerter automatiquement sur un partage externe inattendu.
- Manque de référence sur ce qui doit être partagé et ce qui ne doit pas l'être.
- Adoption par défaut d'une politique largement ouverte au lieu de la partager explicitement lorsque c'est nécessaire.
- Création manuelle des ressources de base qui se chevauchent si nécessaire.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : moyen

Directives d'implémentation

Concevez vos contrôles et modèles d'accès de façon à régir la consommation de ressources partagées en toute sécurité et uniquement avec des entités approuvées. Surveillez les ressources partagées et examinez l'accès aux ressources partagées en permanence, et soyez alerté sur les partages inappropriés ou inattendus. Consultez Analyser l'accès public et intercompte pour vous aider à établir une gouvernance afin de réduire l'accès externe aux seules ressources qui en ont besoin, et d'établir un processus de surveillance continue et d'alerte automatique.

Le partage intercompte dans AWS Organizations est pris en charge par un certain nombre de services AWS, comme AWS Security Hub, Amazon GuardDuty et AWS Backup. Ces services permettent de partager les données vers un compte central, de rendre les données accessibles à partir d'un compte central ou de gérer les ressources et les données à partir d'un compte central. Par exemple, AWS Security Hub peut transférer les découvertes des comptes individuels vers un compte central où vous pouvez voir toutes ces informations. AWS Backup peut prendre une sauvegarde pour une ressource et la partager entre les comptes. Vous pouvez utiliser AWS Resource Access Manager (AWS RAM) afin de partager d'autres ressources communes, telles que les sous-réseaux VPC et les pièces jointes Transit Gateway, AWS Network Firewall ou les pipelines Amazon SageMaker.

Pour limiter votre compte au partage de ressources au sein de votre organisation, utilisez les politiques de contrôle des services (SCP) afin d'empêcher l'accès aux principaux externes. Lorsque vous partagez des ressources, combinez les contrôles basés sur l'identité et les contrôles réseau afin de créer un périmètre de données pour votre organisation dans le but de contribuer à la protection contre les accès involontaires. Un périmètre de données est un ensemble de barrières de protection préventives qui vous permettent de vous assurer que seules les identités approuvées accèdent aux ressources approuvées à partir des réseaux attendus. Ces contrôles doivent placer des limites appropriées pour les ressources susceptibles d'être partagées et empêcher le partage ou l'exposition de ressources qui ne doivent pas être autorisées. Par exemple, dans le cadre de votre périmètre de données, vous pouvez utiliser les politiques de point de terminaison VPC et la condition `AWS:PrincipalOrgId` afin de vous assurer que les identités qui accèdent aux compartiments Amazon S3 appartiennent à votre organisation. Il est important de noter que les SCP ne s'appliquent pas aux rôles liés aux services (LSR) ni aux principaux de services AWS.

Lorsque vous utilisez Amazon S3, désactivez les listes de contrôle d'accès pour votre compartiment Amazon S3 et utilisez les politiques IAM pour définir le contrôle des accès. Pour limiter l'accès à une origine Amazon S3 depuis Amazon CloudFront, migrez depuis l'identité d'accès d'origine (OAI) vers le contrôle d'accès d'origine (OAC) qui prend en charge des fonctionnalités supplémentaires, y compris le chiffrement côté serveur avec AWS Key Management Service.

Dans certains cas, vous pouvez autoriser le partage des ressources à l'extérieur de votre organisation ou accorder à un tiers l'accès à vos ressources. Pour des conseils normatifs sur la gestion des autorisations de partage des ressources à l'externe, consultez Gestion des autorisations.

Étapes d'implémentation

1. Utilisez AWS Organizations.

   AWS Organizations est un service de gestion des comptes qui vous permet de regrouper plusieurs Comptes AWS dans une organisation que vous créez et gérez de façon centralisée. Vous pouvez regrouper vos comptes en unités d'organisation (UO) et joindre différentes politiques à chacune d'entre elles afin de vous aider à répondre à vos besoins en matière de budget, de sécurité et de conformité. Vous pouvez également contrôler la façon dont l'intelligence artificielle (IA) et le machine learning (ML) d'AWS peuvent collecter et stocker des données, et utiliser la gestion multicompte des services AWS intégrés à Organizations.
2. Intégrez AWS Organizations aux services AWS.

   Lorsque vous permettez à un service AWS d'effectuer des tâches en votre nom dans les comptes membres de votre organisation, AWS Organizations crée un rôle IAM lié à ce service dans chaque

compte membre. Gérez l'accès approuvé à l'aide de la AWS Management Console, des API AWS ou d'AWS CLI. Pour obtenir des conseils normatifs sur la mise en place d'un accès approuvé, consultez Utilisation d'AWS Organizations avec d'autres services AWS et Services AWS que vous pouvez utiliser avec Organizations.

3. Établissez un périmètre de données.

   Le périmètre AWS est généralement représenté comme une organisation gérée par AWS Organizations. Avec les réseaux et les systèmes sur site, l'accès aux ressources AWS est ce que beaucoup considèrent comme le périmètre de mon AWS. L'objectif du périmètre est de vérifier que l'accès est autorisé si l'identité est approuvée, si la ressource est approuvée et si le réseau est attendu.

   a. Définissez et implémentez les périmètres.

      Suivez les étapes décrites dans Perimeter implementation (Implémentation du périmètre) dans le livre blanc Building a Perimeter on AWS (Créer un périmètre sur AWS) pour chaque condition d'autorisation. Pour des conseils normatifs sur la protection de la couche réseau, consultez Protection des réseaux.

   b. Surveillez et alertez en continu.

      AWS Identity and Access Management Access Analyzer vous permet d'identifier les ressources de votre organisation et les comptes qui sont partagés avec des entités externes. Vous pouvez intégrer IAM Access Analyzer à AWS Security Hub pour envoyer et regrouper les découvertes d'une ressource d'IAM Access Analyzer vers Security Hub afin de contribuer à l'analyse de la situation de sécurité de votre environnement. Pour mettre en place l'intégration, activez IAM Access Analyzer et Security Hub dans chaque région de chaque compte. Vous pouvez utiliser AWS Config Rules pour auditer la configuration et alerter la partie appropriée à l'aide d'AWS Chatbot avec AWS Security Hub. Vous pouvez ensuite utiliser les documents AWS Systems Manager Automation pour résoudre les problèmes des ressources non conformes.

   c. Pour des conseils normatifs sur la surveillance et les alertes en continu relatives aux ressources partagées en externe, consultez Analyser l'accès public et entre les comptes.

4. Utilisez le partage des ressources dans les services AWS et limitez l'accès en conséquence.

   De nombreux services AWS vous permettent de partager des ressources avec un autre compte ou de cibler une ressource dans un autre compte, par exemple Amazon Machine Images (AMI) et AWS Resource Access Manager (AWS RAM). Limitez l'API `ModifyImageAttribute` à la spécification des comptes approuvés pour partager l'AMI. Spécifiez la condition `ram:RequestedAllowsExternalPrincipals` lorsque vous utilisez AWS RAM pour limiter le

partage à votre organisation seulement, pour empêcher l'accès à des identités non approuvées. Pour des conseils et des considérations normatifs, consultez Partage des ressources et cibles externes.

5. Utilisez AWS RAM pour partager en toute sécurité dans un compte ou avec d'autres Comptes AWS.

   AWS RAM vous aide à partager en toute sécurité les ressources que vous avez créées avec les rôles et les utilisateurs de votre compte et avec d'autres Comptes AWS. Dans un environnement multicompte, AWS RAM vous permet de créer une ressource une fois et de la partager avec d'autres comptes. Cette approche permet de réduire vos frais généraux opérationnels tout en assurant la cohérence, la visibilité et l'auditabilité grâce à des intégrations avec Amazon CloudWatch et AWS CloudTrail, que vous ne recevez pas lorsque vous utilisez l'accès intercompte.

   Si vous avez déjà partagé des ressources à l'aide d'une politique basée sur les ressources, vous pouvez utiliser l'API `PromoteResourceShareCreatedFromPolicy` ou un équivalent pour faire passer le partage des ressources vers un partage AWS RAM complet.

   Dans certains cas, vous devrez peut-être prendre des mesures supplémentaires pour partager les ressources. Par exemple, pour partager un instantané chiffré, vous devez partager une clé AWS KMS.

Ressources

Bonnes pratiques associées :

- SEC03-BP07 Analyser l'accès public et entre les comptes
- SEC03-BP09 Partager des ressources en toute sécurité avec un tiers
- SEC05-BP01 Créer des couches réseau

Documents connexes :

- Propriétaire d'un compartiment accordant des autorisations entre comptes à des objets qu'il ne possède pas
- How to use Trust Policies with IAM
- Building Data Perimeter on AWS

- Procédure d'utilisation d'un ID externe lorsque vous accordez l'accès à vos ressources AWS à un tiers
- Services AWS que vous pouvez utiliser avec AWS Organizations
- Establishing a data perimeter on AWS: Allow only trusted identities to access company data

Vidéos connexes :

- Granular Access with AWS Resource Access Manager
- Securing your data perimeter with VPC endpoints
- Establishing a data perimeter on AWS

Outils associés :

- Exemples de politiques de périmètre de données

SEC03-BP09 Partager des ressources en toute sécurité avec un tiers

La sécurité de votre environnement cloud ne s'arrête pas à votre organisation. Votre organisation peut faire appel à un tiers pour gérer une partie de vos données. La gestion des autorisations pour le système géré par un tiers doit suivre la pratique de l'accès juste à temps en utilisant le principe du moindre privilège avec des informations d'identification temporaires. En travaillant en étroite collaboration avec un tiers, vous pouvez réduire ensemble l'étendue de l'impact et le risque d'accès involontaire.

Résultat souhaité : des informations d'identification AWS Identity and Access Management (IAM) à long terme, des clés d'accès IAM et des clés secrètes qui sont associées à un utilisateur peuvent être utilisées par n'importe qui tant que les informations d'identification sont valides et actives. L'utilisation d'un rôle IAM et d'informations d'identification temporaires vous permettent d'améliorer votre situation globale en matière de sécurité en réduisant l'effort de gestion des informations d'identification à long terme, y compris la gestion et les frais généraux opérationnels de ces détails sensibles. En utilisant un identifiant unique universel (UUID) pour l'ID externe dans la politique d'approbation IAM et en veillant à ce que les politiques IAM restent attachées au rôle IAM sous votre contrôle, vous pouvez auditer et vérifier que l'accès accordé au tiers n'est pas trop permissif. Pour des conseils normatifs sur l'analyse des ressources partagées en externe, consultez SEC03-BP07 Analyser l'accès public et entre les comptes.

Anti-modèles courants :

- Utilisation de la politique d'approbation IAM sans aucune condition.
- Utilisation d'informations d'identification IAM et de clés d'accès à long terme.
- Réutilisation des ID externes.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : moyen

Directives d'implémentation

Vous pouvez autoriser le partage des ressources en dehors d'AWS Organizations ou accorder un accès tiers à votre compte. Par exemple, un tiers peut fournir une solution de surveillance qui doit accéder aux ressources de votre compte. Dans ces cas de figure, vous devez créer un rôle intercompte IAM en lui attribuant uniquement les privilèges requis par le tiers. De plus, vous devez définir une politique d'approbation à l'aide de la condition d'ID externe. Lorsque vous utilisez un identifiant externe, vous pouvez (ou le tiers peut) générer un identifiant unique pour chaque client, tiers ou location. L'ID unique ne doit être contrôlé que par vous après sa création. Le tiers doit implémenter un processus pour relier l'ID externe au client de manière sécurisée, auditable et reproductible.

Vous pouvez également utiliser IAM Roles Anywhere afin de gérer les rôles IAM pour les applications en dehors d'AWS qui utilisent les API AWS.

Si le tiers n'a plus besoin d'accéder à votre environnement, supprimez le rôle. Évitez de fournir à des tiers des informations d'identification à long terme. Gardez un œil sur les autres services AWS qui prennent en charge le partage. Par exemple, AWS Well-Architected Tool autorise le partage d'une charge de travail avec d'autres Comptes AWS et AWS Resource Access Manager vous aide à partager en toute sécurité une ressource AWS que vous possédez avec d'autres comptes.

Étapes d'implémentation

1. Utilisez des rôles intercomptes pour donner accès aux comptes externes.

   Les rôles intercomptes réduisent la quantité d'informations sensibles stockées par des comptes externes et des tiers pour servir leurs clients. Les rôles intercomptes vous permettent d'octroyer l'accès aux ressources AWS de votre compte en toute sécurité à un tiers, par exemple aux AWS Partner ou à d'autres comptes de votre organisation, tout en préservant la capacité de gérer et de vérifier cet accès.

   Il se peut que le tiers vous fournisse des services à partir d'une infrastructure hybride ou qu'il extraie des données hors site pour les transférer dans un emplacement hors site. IAM Roles

Anywhere permet aux charges de travail tierces d'interagir en toute sécurité avec vos charges de travail AWS et réduit davantage la nécessité d'utiliser des informations d'identification à long terme.

Vous ne devez pas utiliser d'informations d'identification à long terme ni de clés d'accès associées aux utilisateurs pour fournir un accès à un compte externe. Utilisez plutôt les rôles intercomptes pour fournir l'accès intercompte.

2. Utilisez un ID externe avec des tiers.

   L'utilisation d'un ID externe vous permet de désigner qui peut assumer un rôle dans une politique d'approbation IAM. La politique d'approbation peut exiger que l'utilisateur qui assume le rôle fasse valoir la condition et la cible dans lesquelles il opère. Elle permet également au propriétaire du compte d'accepter que le rôle soit endossé uniquement dans des circonstances spécifiques. La fonction principale de l'ID externe est de traiter et de prévenir le problème de confusion de principal.

   Utilisez un ID externe si vous êtes propriétaire d'un Compte AWS et vous avez configuré un rôle pour un tiers qui accède à d'autres Comptes AWS en plus des vôtres, ou lorsque vous assumez des rôles au nom de différents clients. Collaborez avec votre tiers ou AWS Partner pour établir une condition d'identification externe à inclure dans la politique d'approbation IAM.

3. Utilisez des ID externes universellement uniques.

   Implémentez un processus qui génère une valeur unique aléatoire pour un ID externe, comme un identifiant unique universel (UUID). Un tiers qui réutilise des ID externes entre différents clients ne règle pas le problème de confusion du principal, car le client A pourrait être en mesure de consulter les données du client B en utilisant le rôle ARN du client B avec l'ID externe dupliqué. Dans un environnement multilocataire, où un tiers prend en charge plusieurs clients avec différents Comptes AWS, le tiers doit utiliser un ID unique différent comme ID externe pour chaque Compte AWS. Le tiers est responsable de la détection des ID externes dupliqués et de la correspondance sécurisée entre chaque client et son ID externe respectif. Le tiers doit vérifier qu'il peut uniquement assumer ce rôle lorsqu'il indique l'ID externe. Le tiers doit s'abstenir de stocker le rôle du client ARN et l'ID externe jusqu'à ce que l'ID externe soit requis.

   L'ID externe n'est pas traité comme un secret, mais il ne doit pas être facile à deviner, comme un numéro de téléphone, un nom ou un numéro de compte. Faites de l'ID externe un champ en lecture seule afin qu'il ne puisse pas être modifié dans le but de se faire passer pour la configuration.

Le tiers ou vous-même pouvez générer l'ID externe. Définissez un processus pour déterminer qui est responsable de la génération de l'ID. Quelle que soit l'entité qui crée l'ID externe, le tiers applique l'unicité et les formats de façon uniforme parmi les clients.

4. Rendez obsolètes les informations d'identification à long terme fournis par le client.

   Rendez obsolète l'utilisation d'informations d'identification à long terme et utilisez des rôles intercomptes ou IAM Roles Anywhere. Si vous devez utiliser des informations d'identification à long terme, établissez un plan pour migrer vers un accès basé sur les rôles. Pour plus d'informations sur la gestion des clés, consultez Gestion des identités. Collaborez également avec votre équipe Compte AWS et le tiers pour établir un runbook d'atténuation des risques. Pour obtenir des conseils normatifs sur la façon d'intervenir et d'atténuer les répercussions potentielles d'un incident de sécurité, consultez Réponse aux incidents.

5. Vérifiez que la configuration est conforme aux conseils normatifs ou qu'elle est automatisée.

   La politique créée pour l'accès intercompte doit suivre le principe du moindre privilège. Le tiers doit fournir un document de politique de rôle ou un mécanisme de configuration automatisé qui utilise un modèle AWS CloudFormation ou un équivalent pour vous. Cela réduit le risque d'erreurs associées à la création manuelle de politiques et offre une piste auditable. Pour plus d'informations sur l'utilisation d'un modèle AWS CloudFormation afin de créer des rôles intercomptes, consultez Rôles intercomptes.

   Le tiers doit fournir un mécanisme de configuration automatisé et auditable. Cependant, si vous utilisez le document de politique de rôle décrivant l'accès nécessaire, vous devez automatiser la configuration du rôle. Si vous utilisez un modèle AWS CloudFormation ou un équivalent, vous devrez surveiller les changements liés à la détection des dérives dans le cadre de la pratique d'audit.

6. Planifiez les modifications.

   Votre structure de compte, la nécessité de faire appel à un tiers, ou son offre de service peuvent changer. Vous devez anticiper les changements et les défaillances, et planifier en conséquence avec les personnes, processus et technologies appropriés. Auditez régulièrement le niveau d'accès que vous fournissez et implémentez des méthodes de détection pour vous avertir des changements imprévus. Surveillez et auditez l'utilisation du rôle et de l'entrepôt de données des ID externes. Vous devez être prêt à révoquer l'accès tiers, de façon temporaire ou permanente, en raison de changements ou de tendances d'accès imprévus. De plus, mesurez l'impact sur votre opération de révocation, y compris le temps nécessaire pour l'exécution, les personnes impliquées, le coût et l'impact sur d'autres ressources.

Pour des conseils normatifs sur les méthodes de détection, consultez Bonnes pratiques de détection.

Ressources

Bonnes pratiques associées :

- SEC02-BP02 Utiliser des informations d'identification temporaires
- SEC03-BP05 Définir des protections par autorisation pour votre organisation
- SEC03-BP06 Gérer l'accès en fonction du cycle de vie
- SEC03-BP07 Analyser l'accès public et entre les comptes
- SEC04 Détection

Documents connexes :

- Propriétaire d'un compartiment accordant des autorisations entre comptes à des objets qu'il ne possède pas
- How to use trust policies with IAM roles
- Déléguer l'accès entre les Comptes AWS à l'aide des rôles IAM
- How do I access resources in another Compte AWS using IAM?
- Bonnes pratiques de sécurité dans IAM
- Logique d'évaluation de politiques intercomptes
- Procédure d'utilisation d'un ID externe lorsque vous accordez l'accès à vos ressources AWS à un tiers
- Collecting Information from AWS CloudFormation Resources Created in External Accounts with Custom Resources
- Securely Using External ID for Accessing AWS Accounts Owned by Others
- Extend IAM roles to workloads outside of IAM with IAM Roles Anywhere

Vidéos connexes :

- How do I allow users or roles in a separate Compte AWS access to my Compte AWS?
- AWS re:Invent 2018: Become an IAM Policy Master in 60 Minutes or Less

- AWS Knowledge Center Live: IAM Best Practices and Design Decisions

Exemples connexes :

- Atelier Well-Architected – Lambda cross account IAM role assumption (Level 300)
- Configure cross-account access to Amazon DynamoDB (Configuration de l'accès intercompte à Amazon DynamoDB)
- AWS STS Network Query Tool

Détection

Question

- SÉC 4. Comment détectez-vous et enquêtez-vous sur les événements de sécurité ?

SÉC 4. Comment détectez-vous et enquêtez-vous sur les événements de sécurité ?

Capturez et analysez les événements de journaux et les métriques pour obtenir une meilleure visibilité. Prenez des mesures en réaction aux événements de sécurité et aux menaces potentielles afin de contribuer à sécuriser votre charge de travail.

Bonnes pratiques

- SEC04-BP01 Configurer une journalisation de service et d'application
- SEC04-BP02 Analyse centralisée des journaux, des résultats et des métriques
- SEC04-BP03 Automatiser la réponse aux événements
- SEC04-BP04 Implémenter des événements de sécurité exploitables

SEC04-BP01 Configurer une journalisation de service et d'application

Conservez les journaux d'événements de sécurité des services et des applications. Il s'agit d'un principe de sécurité fondamental pour les cas d'audit, d'enquête et d'utilisation opérationnelle, et d'une exigence de sécurité commune dictée par les normes, politiques et procédures de gouvernance, de risque et de conformité (GRC).

Résultat souhaité : une organisation doit être en mesure de récupérer de façon fiable et cohérente les journaux d'événements de sécurité à partir de services et d'applications AWS rapidement lorsqu'il est nécessaire de réaliser un processus ou une obligation interne, par exemple une intervention en

cas d'incident de sécurité. Envisagez de centraliser les journaux pour obtenir de meilleurs résultats opérationnels.

Anti-modèles courants :

- Les journaux sont stockés à perpétuité ou supprimés trop tôt.
- Tout le monde peut accéder aux journaux.
- Se fier entièrement aux processus manuels pour la gouvernance et l'utilisation des journaux.
- Stocker chaque type de journal au cas où il serait nécessaire.
- Vérifier l'intégrité des journaux uniquement lorsque cela s'avère nécessaire.

Avantages liés à l'instauration de cette bonne pratique : implémentation d'un mécanisme d'analyse des causes fondamentales (RCA) pour les incidents de sécurité et une source de preuve pour vos obligations en matière de gouvernance, de risque et de conformité.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

Au cours d'une enquête de sécurité ou d'autres cas d'utilisation en fonction de vos besoins, vous devez être en mesure d'examiner les journaux pertinents pour consigner et comprendre la portée et la chronologie complètes de l'incident. Des journaux sont également requis pour la génération d'alertes, indiquant que certaines actions intéressantes ont eu lieu. Il est essentiel de sélectionner, d'activer, de stocker et de configurer les mécanismes d'interrogation et de récupération ainsi que les alertes.

Étapes d'implémentation

- Sélectionnez et activez les sources du journal. Avant une enquête de sécurité, vous devez saisir les journaux pertinents pour reconstruire rétroactivement l'activité dans un Compte AWS. Sélectionnez et activez les sources de journaux pertinentes pour vos charges de travail.

  Les critères de sélection des sources de journalisation doivent être fondés sur les cas d'utilisation requis par votre entreprise. Établissez une piste pour chaque Compte AWS en utilisant AWS CloudTrail ou une piste AWS Organizations, puis configurez un compartiment Amazon S3 pour cette piste.

  AWS CloudTrail est un service de journalisation qui suit les appels API sur un Compte AWS pour capturer l'activité de service AWS. Il est activé par défaut avec une conservation de 90 jours des

événements de gestion qui peuvent être extraits via l'historique des événements CloudTrail à l'aide de la AWS Management Console, de l'AWS CLI ou d'un kit AWS SDK. Pour une conservation et une visibilité plus longues des données, créez une piste CloudTrail et associez-la à un compartiment Amazon S3, ainsi qu'à un groupe de journaux Amazon CloudWatch si nécessaire. Vous pouvez également créer un CloudTrail Lake, qui conserve les journaux CloudTrail jusqu'à sept ans et fournit une fonction de requête SQL.

AWS recommande aux clients qui utilisent un VPC d'activer les journaux réseau trafic et DNS en utilisant les journaux de flux VPC et les journaux de requête du résolveur de requêtes Amazon Route 53, respectivement, et de les diffuser en continu dans un compartiment Amazon S3 ou un groupe de journaux CloudWatch. Vous pouvez créer un journal de flux VPC pour un VPC, un sous-réseau ou une interface réseau. Pour les journaux de flux VPC, vous pouvez choisir la façon dont et l'endroit où vous les utilisez pour réduire les coûts.

Les journaux AWS CloudTrail, les journaux de flux VPC et les journaux de requêtes du résolveur Route 53 sont les sources de journalisation de base qui soutiennent les enquêtes de sécurité dans AWS. Vous pouvez également utiliser Amazon Security Lake pour collecter, normaliser et stocker ces données de journaux au format Apache Parquet et Open Cybersecurity Schema Framework (OCSF), qui est prêt pour l'interrogation. Security Lake prend également en charge d'autres journaux AWS et des journaux de sources tierces.

Les services AWS peuvent générer des journaux non capturés par les sources de journaux de base, comme les journaux Elastic Load Balancing, les journaux AWS WAF, les journaux de l'enregistreur AWS Config, les découvertes Amazon GuardDuty, les journaux d'audit Amazon Elastic Kubernetes Service (Amazon EKS) et les journaux d'application et de système d'exploitation des instances Amazon EC2. Pour une liste complète des options de journalisation et de surveillance, consultez Annexe A : Définitions de la capacité cloud – Journalisation et événements dans le Guide de réponse aux incidents de sécurité AWS.

- Recherchez les capacités de journalisation pour chaque service et application AWS : chaque service et application AWS vous offre des options de stockage des journaux, chacune avec ses propres capacités de conservation et de cycle de vie. Les deux services de stockage de journaux les plus courants sont Amazon Simple Storage Service (Amazon S3) et Amazon CloudWatch. Pour de longues périodes de conservation, il est recommandé d'utiliser Amazon S3 pour sa rentabilité et ses capacités de cycle de vie flexibles. Si l'option de journalisation principale est Journaux Amazon CloudWatch, en tant qu'option, vous devez envisager d'archiver les journaux les moins consultés dans Amazon S3.

- Sélectionnez le stockage des journaux : le choix du stockage des journaux est généralement lié à l'outil d'interrogation que vous utilisez, aux capacités de conservation, à la familiarité et au coût. Les options principales du stockage de journaux sont un compartiment Amazon S3 ou un groupe de journaux CloudWatch.

  Un compartiment Amazon S3 offre un stockage durable et rentable avec une politique de cycle de vie facultative. Les journaux stockés dans des compartiments Amazon S3 peuvent être interrogés à l'aide de services tels que Amazon Athena.

  Un groupe de journaux CloudWatch offre un stockage durable et une installation de requête intégrée via CloudWatch Logs Insights.
- Identifiez la conservation appropriée des journaux : lorsque vous utilisez un compartiment Amazon S3 ou un groupe de journaux CloudWatch pour stocker des journaux, vous devez établir des cycles de vie adéquats pour chaque source de journaux afin d'optimiser les coûts de stockage et de récupération. Les clients ont généralement entre trois mois et un an de journaux facilement disponibles pour la recherche, avec une conservation de sept ans maximum. Le choix de la disponibilité et de la conservation doit correspondre à vos exigences en matière de sécurité et à un ensemble d'obligations statutaires, réglementaires et opérationnelles.
- Activez la journalisation pour chaque service et application AWS avec des politiques de conservation et de cycle de vie appropriées : pour chaque service ou application AWS dans votre organisation, recherchez les conseils de configuration de journalisation spécifiques :
  - Configurer AWS CloudTrail Trail
  - Configurer des journaux de flux VPC
  - Configurer l'exportation des découvertes Amazon GuardDuty
  - Configurer l'enregistrement AWS Config
  - Configurer le trafic d'ACL web AWS WAF
  - Configurer les journaux de trafic réseau AWS Network Firewall
  - Configurer les journaux d'accès Elastic Load Balancing
  - Configurer les journaux de requêtes du résolveur Amazon Route 53
  - Configurer les journaux Amazon RDS
  - Configurer les journaux de plan de contrôle Amazon EKS
  - Configurer l'agent Amazon CloudWatch pour les instances Amazon EC2 et les serveurs sur site
- Sélectionnez et implémentez des mécanismes d'interrogation pour les journaux : pour les interrogations de journaux, vous pouvez utiliser CloudWatch Logs Insights pour les données

stockées dans les groupes de journaux CloudWatch, et Amazon Athena et Amazon OpenSearch Service pour les données stockées dans Amazon S3. Vous pouvez également utiliser des outils d'interrogation tiers tels qu'un service de gestion des informations de sécurité et des événements (SIEM).

Le processus de sélection d'un outil d'interrogation de journaux doit tenir compte des aspects humains, technologiques et de processus de vos opérations de sécurité. Choisissez un outil qui répond aux exigences opérationnelles, métier et de sécurité, tout en étant accessible et gérable à long terme. Gardez à l'esprit que les outils d'interrogation de journaux fonctionnent de manière optimale lorsque le nombre de journaux à analyser est maintenu dans les limites de l'outil. Il n'est pas rare d'avoir plusieurs outils d'interrogation en raison de contraintes de coût ou techniques.

Par exemple, vous pouvez utiliser un outil de gestion des événements et des informations de sécurité tiers pour effectuer des requêtes sur les 90 derniers jours de données, mais utiliser Athena pour effectuer des requêtes au-delà de 90 jours en raison du coût d'ingestion du journal d'un SIEM. Quelle que soit l'implémentation choisie, assurez-vous que votre approche réduit au minimum le nombre d'outils requis pour maximiser l'efficacité opérationnelle, en particulier pendant une enquête sur un événement de sécurité.

- Utilisez des journaux pour les alertes : AWS fournit des alertes par l'intermédiaire de plusieurs services de sécurité :
  - AWS Config surveille et enregistre vos configurations de ressources AWS et vous permet d'automatiser l'évaluation et la correction par rapport aux configurations souhaitées.
  - Amazon GuardDuty est un service de détection des menaces qui surveille continuellement les activités malveillantes et les comportements non autorisés pour protéger votre Comptes AWS et vos charges de travail. GuardDuty ingère, regroupe et analyse les informations provenant de sources telles que la gestion et les événements de données AWS CloudTrail, les journaux DNS, les flux de journaux VPC et les journaux d'audit Amazon EKS. GuardDuty extrait des flux de données indépendants directement depuis CloudTrail, les journaux de flux VPC, les journaux de requêtes DNS et Amazon EKS. Vous n'avez pas besoin de gérer les politiques de compartiment Amazon S3 ni de modifier la façon dont vous collectez et stockez les journaux. Il est toujours recommandé de conserver ces journaux à des fins d'enquête et de conformité.
  - AWS Security Hub fournit un emplacement unique qui regroupe, organise et priorise vos alertes de sécurité ou vos résultats provenant de plusieurs services AWS et de produits tiers en option pour vous donner une vue complète des alertes de sécurité et du statut de conformité.

  Vous pouvez également utiliser des moteurs de génération d'alertes personnalisés pour les alertes de sécurité non couvertes par ces services ou pour les alertes spécifiques pertinentes à votre

environnement. Pour plus d'informations sur la création de ces alertes, consultez Détection dans le guide des réponses aux incidents de sécurité AWS.

Ressources

Bonnes pratiques associées :

- SEC04-BP02 Analyse centralisée des journaux, des résultats et des métriques
- SEC07-BP04 Définir la gestion du cycle de vie des données
- SEC10-BP06 Prédéployer les outils

Documents connexes :

- AWS Security Incident Response Guide
- Démarrer avec Amazon Security Lake
- Mise en route avec Amazon CloudWatch Logs
- Security Partner Solutions: Logging and Monitoring (Solutions partenaires de sécurité : journalisation et surveillance)

Vidéos connexes :

- AWS re:Invent 2022 - Introducing Amazon Security Lake

Exemples connexes :

- Assisted Log Enabler for AWS
- AWS Security Hub Findings Historical Export

Outils associés :

- Snowflake pour la cybersécurité

SEC04-BP02 Analyse centralisée des journaux, des résultats et des métriques

les équipes responsables des opérations de sécurité s'appuient sur la collecte de journaux et l'utilisation d'outils de recherche pour découvrir les événements d'intérêt potentiels, susceptibles

d'indiquer une activité non autorisée ou une modification non intentionnelle. Cependant, la simple analyse des données collectées et le traitement manuel des informations ne suffisent pas pour faire face au volume d'informations provenant d'architectures complexes. L'analyse et les rapports ne facilitent pas l'affectation des ressources appropriées pour traiter un événement dans les délais impartis.

Une bonne pratique pour constituer une équipe d'opérations de sécurité mature consiste à intégrer profondément le flux d'événements et de résultats de sécurité dans un système de notification et de flux de travail, tel qu'un système de tickets, un système de gestion des bogues et problèmes ou un autre système de gestion des informations et des événements de sécurité (SIEM). Ainsi, le flux de travail n'est plus intégré aux rapports par e-mail et statiques, ce qui permet d'acheminer, de transférer et de gérer les événements ou les résultats. De nombreuses organisations intègrent également des alertes de sécurité dans leurs plateformes de discussion instantanée et collaborative et de productivité des développeurs. Pour les organisations qui se lancent dans l'automatisation, un système de tickets à faible latence axé sur les API, offre une flexibilité considérable pour planifier ce qu'il faut automatiser en premier.

Cette bonne pratique s'applique non seulement aux événements de sécurité générés par les messages du journal décrivant l'activité des utilisateurs ou les événements du réseau, mais aussi aux changements détectés dans l'infrastructure elle-même. La capacité à détecter les changements, à déterminer si un changement était approprié, puis à acheminer ces informations vers le processus de correction approprié est essentielle pour gérer et valider une architecture sécurisée, dans le contexte de changements dont la nature indésirable est suffisamment subtile pour que leur exécution ne puisse être actuellement empêchée par une combinaison de configuration AWS Identity and Access Management(IAM) et AWS Organizations.

Amazon GuardDuty et AWS Security Hub fournissent des mécanismes d'agrégation, de déduplication et d'analyse pour les enregistrements de journaux qui sont également mis à votre disposition via d'autres services AWS. GuardDuty ingère, agrège et analyse les informations provenant de sources telles que les événements de gestion et de données AWS CloudTrail, les journaux DNS VPC et les journaux de flux VPC. Security Hub peut ingérer, agréger et analyser les résultats provenant de GuardDuty, AWS Config, Amazon Inspector, Amazon Macie, AWS Firewall Manager et d'un nombre important de produits de sécurité tiers disponibles dans AWS Marketplace et, s'il a été conçu en conséquence, votre propre code. GuardDuty et Security Hub ont tous les deux un modèle Administrateur/Maître qui peut agréger les résultats et les informations sur plusieurs comptes. Security Hub est souvent utilisé par les clients qui ont un système de gestion des informations et des événements de sécurité (SIEM) sur site comme préprocesseur et agrégateur de journaux et d'alertes

côté AWS à partir duquel ils peuvent ensuite ingérer Amazon EventBridge via un processeur et un redirecteur basé sur AWS Lambda.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Évaluer les fonctionnalités de traitement des journaux : évaluez les options disponibles pour le traitement des journaux.
  - Utilisez Amazon OpenSearch Service pour enregistrer et surveiller (presque) tout.
  - Trouvez un partenaire spécialisé dans les solutions de journalisation et de surveillance.
- Pour commencer à analyser les journaux CloudTrail, testez Amazon Athena.
  - Configuration d'Athena pour analyser les journaux CloudTrail
- Implémenter la journalisation centralisée dans AWS : consultez l'exemple de solution AWS suivant pour centraliser la journalisation à partir de plusieurs sources.
  - Centraliser la solution de journalisation
- Implémenter la journalisation centralisée avec le partenaire : les partenaires APN disposent de solutions pour vous aider à analyser les journaux de manière centralisée.
  - Journalisation et surveillance

Ressources

Documents connexes :

- AWS Answers : journalisation centralisée
- AWS Security Hub
- Amazon CloudWatch
- Amazon EventBridge
- Premiers pas : Amazon CloudWatch Logs
- Solutions partenaires de sécurité : journalisation et surveillance

Vidéos connexes :

- Surveillance centralisée de la configuration et de la conformité des ressources
- Résolution des problèmes identifiés par Amazon GuardDuty et AWS Security Hub

- Gestion des menaces dans le cloud : Amazon GuardDuty et AWS Security Hub

SEC04-BP03 Automatiser la réponse aux événements

L'utilisation de l'automatisation pour enquêter et corriger les événements réduit les efforts humains et les erreurs. Elle vous permet aussi de mettre à l'échelle les capacités d'investigation. Les vérifications régulières vous aideront à ajuster les outils d'automatisation et à itérer en continu.

Dans AWS, l'analyse des événements d'intérêt et des informations utiles sur des changements potentiellement inattendus dans un flux de travail automatisé peut être réalisée avec Amazon EventBridge. Ce service fournit un moteur de règles évolutif conçu pour négocier les formats d'événements AWS natifs (tels que les événements AWS CloudTrail) ainsi que les événements personnalisés que vous pouvez générer vous-même. Amazon GuardDuty vous permet également d'acheminer les événements vers un système de flux de travail pour ceux qui mettent en place des systèmes de réponse aux incidents (AWS Step Functions), vers un compte de sécurité central ou encore vers un compartiment pour une analyse plus approfondie.

La détection des changements et l'acheminement de ces informations vers le flux de travail approprié peuvent également être réalisés via AWS Config Rules et les packs de conformité. AWS Config détecte les modifications apportées aux services concernés (bien que sa latence soit supérieure à celle d'EventBridge) et génère des événements qui peuvent être analysés avec AWS Config Rules pour la restauration, l'application de la politique de conformité et le transfert d'informations aux systèmes, tels que les plateformes de gestion des modifications et les systèmes de tickets opérationnels. En plus d'écrire vos propres fonctions Lambda pour répondre aux événements AWS Config, vous pouvez tirer parti du kit de développement AWS Config Rulesset d'une bibliothèque de règles open source AWS Config Rules. Les packs de conformité sont une collection d'actions correctives et de règles AWS Config Rules que vous déployez en tant qu'entité unique créée en tant que modèle YAML. A exemple de modèle de pack de conformité est disponible pour le pilier Sécurité Well-Architected.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Implémenter des alertes automatisées avec GuardDuty : GuardDuty est un service de détection des menaces qui surveille en permanence les activités malveillantes et les comportements non autorisés afin de protéger vos Comptes AWS et vos charges de travail. Activez GuardDuty et configurez des alertes automatiques.

- Automatiser les processus d'investigation : développez des processus automatisés qui enquêtent sur un événement et rapportent les informations à un administrateur pour gagner du temps.
  - Atelier : Amazon GuardDuty dans la pratique

Ressources

Documents connexes :

- AWS Answers : journalisation centralisée
- AWS Security Hub
- Amazon CloudWatch
- Amazon EventBridge
- Premiers pas : Amazon CloudWatch Logs
- Solutions partenaires de sécurité : journalisation et surveillance
- Configuration d'Amazon GuardDuty

Vidéos connexes :

- Surveillance centralisée de la configuration et de la conformité des ressources
- Résolution des problèmes identifiés par Amazon GuardDuty et AWS Security Hub
- Gestion des menaces dans le cloud : Amazon GuardDuty et AWS Security Hub

Exemples connexes :

- Atelier : Déploiement automatisé des contrôles de détection

SEC04-BP04 Implémenter des événements de sécurité exploitables

Créez des alertes qui sont envoyées à votre équipe et qui peuvent être exécutées par celle-ci. Assurez-vous que les alertes incluent des informations pertinentes pour que l'équipe agisse en conséquence. pour chaque mécanisme de détection dont vous disposez, vous devez également disposer d'un processus, sous la forme d'un runbook ou d'un playbookpour enquêter. Par exemple, lorsque vous activez Amazon GuardDuty, il génère des résultats différents. Vous devez avoir une entrée de runbook pour chaque type de résultat. Par exemple, si un cheval de Troie est détecté, votre runbook comporte des instructions simples qui demandent à quelqu'un d'enquêter et de corriger.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Découvrir les métriques disponibles pour les services AWS : découvrez les métriques disponibles via Amazon CloudWatch pour les services que vous utilisez.
  - Documentation des services AWS
  - Utilisation des métriques Amazon CloudWatch
- Configurez des alarmes Amazon CloudWatch.
  - Utilisation des alarmes Amazon CloudWatch

Ressources

Documents connexes :

- Amazon CloudWatch
- Amazon EventBridge
- Solutions partenaires de sécurité : journalisation et surveillance

Vidéos connexes :

- Surveillance centralisée de la configuration et de la conformité des ressources
- Résolution des problèmes identifiés par Amazon GuardDuty et AWS Security Hub
- Gestion des menaces dans le cloud : Amazon GuardDuty et AWS Security Hub

Protection de l'infrastructure

Questions

- SÉC 5. Comment protégez-vous vos ressources réseau ?
- SÉC 6. De quelle façon pouvez-vous assurer la protection de vos ressources de calcul ?

SÉC 5. Comment protégez-vous vos ressources réseau ?

Pour toute charge de travail ayant une forme quelconque de connectivité réseau, qu'il s'agisse d'Internet ou d'un réseau privé, plusieurs couches de défense sont nécessaires pour vous protéger contre les menaces externes et internes basées sur le réseau.

Bonnes pratiques

- SEC05-BP01 Créer des couches réseau
- SEC05-BP02 Contrôler le trafic sur toutes les couches
- SEC05-BP03 Automatiser la protection du réseau
- SEC05-BP04 Mettre en œuvre l'inspection et la protection

SEC05-BP01 Créer des couches réseau

Créez des groupes multicouches pour les composants qui partagent des exigences en matière de sensibilité afin de réduire au minimum la portée potentielle des répercussions d'un accès non autorisé. Par exemple, un cluster de bases de données dans un cloud privé virtuel (VPC) n'ayant pas besoin d'accès à Internet doit être placé dans des sous-réseaux sans routage vers ou depuis Internet. Le trafic ne doit provenir que de la ressource adjacente suivante la moins sensible. Réfléchissez au cas d'une application web se trouvant derrière un équilibreur de charge. Votre base de données ne doit pas être accessible directement depuis l'équilibreur de charge. La logique métier ou le serveur web doivent être les seuls à avoir un accès direct à votre base de données.

Résultat souhaité : créer un réseau multicouche. Les réseaux multicouches permettent de regrouper logiquement des composants de réseau similaires. Ils réduisent également la portée potentielle de l'impact de l'accès non autorisé au réseau. Un réseau multicouche approprié complique les choses pour les utilisateurs non autorisés qui souhaitent accéder à des ressources supplémentaires au sein de votre environnement AWS. En plus de sécuriser les chemins réseau internes, vous devez également protéger votre périphérie de réseau, comme les applications web et les points de terminaison d'API.

Anti-modèles courants :

- Créer toutes les ressources dans un seul VPC ou sous-réseau.
- Utiliser des groupes de sécurité trop permissifs.
- Ne pas utiliser de sous-réseaux.
- Autoriser un accès direct aux stockages de données tels que les bases de données.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

Les composants, tels que les instances Amazon Elastic Compute Cloud (Amazon EC2), les clusters de bases de données Amazon Relational Database Service (Amazon RDS) et les fonctions AWS Lambda qui partagent les exigences d'accessibilité, peuvent être segmentés en couches formées par des sous-réseaux. Envisagez de déployer des charges de travail sans serveur, comme les fonctions Lambda dans un VPC ou derrière un Amazon API Gateway. Les tâches AWS Fargate (Fargate) qui n'ont pas besoin d'accès à Internet doivent être placées dans des sous-réseaux sans routage vers ou depuis Internet. Cette approche multicouche réduit l'impact d'une mauvaise configuration de couche unique, ce qui pourrait autoriser un accès involontaire. Pour AWS Lambda, vous pouvez exécuter vos fonctions dans votre VPC pour anticiper les contrôles basés sur un VPC.

Pour une connectivité réseau pouvant inclure des milliers de VCP, des Comptes AWS et des réseaux sur site, vous devez utiliser AWS Transit Gateway. Transit Gateway agit comme un hub qui contrôle la façon dont le trafic est acheminé entre tous les réseaux connectés, qui agissent comme des rayons. Le trafic entre Amazon Virtual Private Cloud (Amazon VPC) et Transit Gateway reste sur le réseau privé AWS, ce qui réduit l'exposition externe aux utilisateurs non autorisés et les problèmes de sécurité potentiels. L'appairage inter-région Transit Gateway chiffre également le trafic inter-région sans point unique de défaillance ni goulot d'étranglement sur la bande passante.

Étapes d'implémentation

- Utilisez Reachability Analyzer pour analyser le chemin entre une source et une destination en fonction de la configuration : Reachability Analyzer vous permet d'automatiser la vérification de la connectivité vers et depuis les ressources connectées VPC. Notez que cette analyse se fait en examinant la configuration (aucun paquet réseau n'est envoyé lors de l'analyse).
- Utilisez l'analyseur d'accès réseau Amazon VPC pour identifier l'accès involontaire aux ressources du réseau : l'analyseur d'accès réseau Amazon VPC vous permet de spécifier vos besoins d'accès réseau et d'identifier les chemins réseau potentiels.
- Déterminez si les ressources doivent être dans un sous-réseau public : ne placez pas les ressources dans les sous-réseaux publics de votre VPC à moins qu'elles doivent absolument recevoir du trafic réseau entrant de sources publiques.
- Créez des sous-réseaux dans vos VPC : créez des sous-réseaux pour chaque couche réseau (dans les groupes qui comprennent plusieurs zones de disponibilité) pour améliorer la micro-segmentation. Vérifiez également que vous avez associé les tables de routage appropriées à vos sous-réseaux pour contrôler le routage et la connectivité Internet.

- Utilisez AWS Firewall Manager pour gérer vos groupes de sécurité VPC : AWS Firewall Manager permet d'alléger la complexité de gestion liée à l'utilisation de plusieurs groupes de sécurité.
- Utilisez AWS WAF pour vous protéger contre les vulnérabilités web courantes : AWS WAF peut permettre d'améliorer la sécurité de la périphérie en inspectant le trafic à la recherche des vulnérabilités web courantes, telles que l'injection de SQL. Il vous permet également de limiter le trafic des adresses IP provenant de certains pays ou emplacements géographiques.
- Utilisez Amazon CloudFront comme réseau de distribution de contenu (CDN) : Amazon CloudFront peut vous aider à accélérer votre application web en stockant les données plus près de vos utilisateurs. Il peut également améliorer la sécurité de la périphérie en appliquant HTTPS, en limitant l'accès aux zones géographiques et en veillant à ce que le trafic réseau ne puisse accéder aux ressources que lorsqu'il est acheminé via CloudFront.
- Utilisez Amazon API Gateway lors de la création d'interfaces de programmation d'applications (API) : Amazon API Gateway permet de publier, surveiller et sécuriser les API REST, HTTPS et WebSocket.

Ressources

Documents connexes :

- AWS Firewall Manager
- Amazon Inspector
- Sécurité dans Amazon VPC
- Reachability Analyzer
- Amazon VPC Network Access Analyzer (Analyseur d'accès réseau VPC)

Vidéos connexes :

- AWS Transit Gateway reference architectures for many VPCs
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield
- AWS re:Inforce 2022 - Validate effective network access controls on AWS
- AWS re:Inforce 2022 - Advanced protections against bots using AWS WAF

Exemples connexes :

- Atelier Well-Architected – Automated Deployment of VPC

- Atelier : Amazon VPC Network Access Analyzer

SEC05-BP02 Contrôler le trafic sur toutes les couches

lorsque vous créez l'architecture de votre topologie réseau, vous devez examiner les exigences de connectivité de chaque composant. Par exemple, si un composant nécessite d'accéder à Internet (en entrée et en sortie), une connectivité aux VPC, aux services périphériques et aux centres de données externes.

Un VPC vous permet de définir votre topologie de réseau qui s'étend sur une Région AWS avec une plage d'adresses IPv4 privée que vous définissez, ou une plage d'adresses IPv6 que sélectionne AWS. Vous devez appliquer des contrôles multiples avec une approche de défense en profondeur pour le trafic entrant et sortant, y compris l'utilisation de groupes de sécurité (pare-feu à inspection permanente), de listes de contrôle d'accès (ACL) réseau, de sous-réseaux et de tables de routage. Au sein d'un VPC, vous pouvez créer des sous-réseaux dans une zone de disponibilité. Chaque sous-réseau peut avoir une table de routage associée qui définit les règles de routage pour gérer les chemins que le trafic emprunte au sein du sous-réseau. Vous pouvez définir un sous-réseau routable Internet en ayant une route qui accède à une passerelle Internet ou NAT connectée au VPC, ou passant par un autre VPC.

Lorsqu'une instance, une base de données Amazon Relational Database Service(Amazon RDS) ou un autre service sont lancés au sein d'un VPC, ils disposent de leur propre groupe de sécurité sur chaque interface réseau. Ce pare-feu se situe en dehors de la couche du système d'exploitation et peut être utilisé pour définir des règles pour le trafic entrant et sortant autorisé. Vous pouvez également définir les relations entre les groupes de sécurité. Par exemple, les instances d'un groupe de sécurité de la couche base de données n'acceptent que le trafic des instances de la couche application, par référence aux groupes de sécurité appliqués aux instances concernées. Si vous utilisez des protocoles non-TCP, il n'est pas nécessaire de laisser une instance Amazon Elastic Compute Cloud(Amazon EC2) directement accessible par Internet (même avec des ports restreints par des groupes de sécurité) sans utiliser d'équilibreur de charge ou CloudFront. Cela permet de la protéger contre un accès involontaire en cas de problème de système d'exploitation ou d'application. Un sous-réseau peut également avoir une liste ACL réseau qui fait office de pare-feu sans état. Vous devez configurer la liste de contrôle d'accès(ACL) au réseau de manière à réduire l'étendue du trafic autorisé entre les couches. Notez que vous devez définir des règles à la fois pour les flux entrants et sortants.

Certains services AWS nécessitent que des composants accèdent à Internet pour effectuer des appels d'API, là où les points de terminaison d'API AWS sont situés. D'autres services AWS utilisent

les Points de terminaison d'un VPC dans vos Amazon VPC. De nombreux services AWS, notamment Amazon S3 et Amazon DynamoDB, prennent en charge les points de terminaison d'un VPC, et cette technologie a été généralisée dans . AWS PrivateLink. Nous vous recommandons d'utiliser cette approche pour accéder en toute sécurité aux services AWS, aux services tiers et à vos propres services hébergés dans d'autres VPC. Tout le trafic réseau sur AWS PrivateLink reste sur la dorsale mondiale AWS et ne traverse jamais Internet. La connectivité ne peut être initiée que par le consommateur du service, et non par le fournisseur du service. Utiliser AWS PrivateLink pour l'accès au service externe vous permet de créer des VPC isolés sans accès à Internet et contribue à protéger vos VPC contre les vecteurs de menace externes. Les services tiers peuvent utiliser AWS PrivateLink pour permettre à leurs clients de se connecter aux services à partir de leurs VPC via des adresses IP privées. Pour les ressources VPC qui ont besoin d'établir des connexions sortantes à Internet, celles-ci peuvent être établies en mode sortant uniquement (unidirectionnel) via une passerelle NAT gérée par AWS, une passerelle Internet en mode sortant uniquement ou des proxys Web que vous créez et gérez.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Contrôler le trafic réseau dans un VPC : mettez en œuvre les bonnes pratiques liées aux VPC pour contrôler le trafic.
  - Sécurité des Amazon VPC
  - Points de terminaison d'un VPC
  - Groupe de sécurité de Amazon VPC
  - ACL réseau
- Contrôler le trafic en périphérie : implémentez des services périphériques comme Amazon CloudFront pour fournir une couche supplémentaire de protection et d'autres fonctions.
  - Cas d'utilisation d'Amazon CloudFront
  - AWS Global Accelerator
  - AWS Web Application Firewall (AWS WAF)
  - Amazon Route 53
  - Amazon VPC Ingress Routing
- Contrôler le trafic réseau privé : implémentez des services qui protègent le trafic privé pour votre charge de travail.
  - Appairage des Amazon VPC

- Amazon VPC Endpoint Services (AWS PrivateLink)
- Amazon VPC Transit Gateway
- AWS Direct Connect
- AWS Site-to-Site VPN
- AWS Client VPN
- Points d'accès Amazon S3

Ressources

Documents connexes :

- AWS Firewall Manager
- Amazon Inspector
- Démarrer avec AWS WAF

Vidéos connexes :

- AWS Transit Gateway reference architectures for many VPCs
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield

Exemples connexes :

- Atelier : Déploiement automatisé d'un VPC

SEC05-BP03 Automatiser la protection du réseau

Automatisez les mécanismes de protection pour disposer d'un réseau capable de se défendre lui-même grâce à l'intelligence des menaces et à la détection des anomalies. Par exemple, des outils de détection et de prévention des intrusions capables de s'adapter aux menaces actuelles et de réduire leur impact. Un pare-feu d'application web est un scénario dans lequel vous pouvez automatiser la protection du réseau, par exemple, en utilisant la solution AWS WAF Security Automations (https://github.com/awslabs/aws-waf-security-automations) pour bloquer automatiquement les requêtes provenant d'adresses IP associées à des acteurs de menaces connus.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Automatiser la protection du trafic web : AWS propose une solution qui utilise AWS CloudFormation pour déployer automatiquement un ensemble de règles AWS WAF conçues pour filtrer les attaques courantes sur le web. Les utilisateurs peuvent choisir parmi des fonctions de protection préconfigurées qui définissent les règles incluses dans une liste de contrôle d'accès (ACL web) AWS WAF.
  - Automatisations de sécurité AWS WAF
- Envisager les solutions AWS Partner : les partenaires AWS proposent des centaines de produits leaders du secteur qui sont équivalents, identiques ou s'intègrent aux contrôles existants dans vos environnements sur site. Ces produits complètent les services AWS existants pour vous permettre de déployer une architecture de sécurité exhaustive et une expérience plus homogène dans vos environnements cloud et sur site.
  - Sécurité de l'infrastructure

Ressources

Documents connexes :

- AWS Firewall Manager
- Amazon Inspector
- Sécurité des VPC Amazon
- Démarrer avec AWS WAF

Vidéos connexes :

- AWS Transit Gateway reference architectures for many VPCs
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield

Exemples connexes :

- Atelier : Déploiement automatisé d'un VPC

SEC05-BP04 Mettre en œuvre l'inspection et la protection

Inspectez et filtrez votre trafic au niveau de chaque couche. Vous pouvez inspecter les configurations de vos VPC pour détecter tout accès involontaire potentiel à l'aide de VPC Network Access Analyzer. Vous pouvez spécifier vos exigences d'accès au réseau et identifier les chemins réseau potentiels qui ne les satisfont pas. Pour les composants effectuant des transactions via des protocoles basés sur HTTP, un pare-feu d'application Web peut protéger contre les attaques courantes. AWS WAF est un pare-feu d'application Web qui permet de surveiller et de bloquer les requêtes HTTP correspondant à vos règles configurables qui sont transmises à une API Amazon API Gateway, à Amazon CloudFront ou à un Application Load Balancer. Pour commencer à utiliser AWS WAF, vous pouvez utiliser des AWS Managed Rules en combinaison avec les vôtres ou utiliser des intégrations de partenaires existantes.

Pour gérer AWS WAF, les protections AWS Shield Advanced et les groupes de sécurité Amazon VPC dans AWS Organizations, vous pouvez utiliser AWS Firewall Manager. Il vous permet de configurer et de gérer de manière centralisée les règles de pare-feu de l'ensemble de vos comptes et applications, ce qui facilite l'application à grande échelle des règles communes. Il permet également de répondre rapidement aux attaques, à l'aide d' AWS Shield Advancedou de solutions qui peuvent bloquer automatiquement les demandes indésirables adressées à vos applications Web. Firewall Manager fonctionne également avec AWS Network Firewall. AWS Network Firewall est un service géré qui utilise un moteur de règles pour vous donner un contrôle précis sur le trafic réseau avec et sans état. Il prend en charge les spécifications du système de prévention des intrusions (IPS) open source compatible avec Suricata pour les règles afin de protéger plus efficacement votre charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Configurer Amazon GuardDuty : GuardDuty est un service de détection des menaces qui surveille en permanence les activités malveillantes et les comportements non autorisés afin de protéger vos Comptes AWS et vos charges de travail. Activez GuardDuty et configurez des alertes automatiques.
  - Amazon GuardDuty
  - Atelier : Déploiement automatisé des contrôles de détection
- Configurer des flux de journaux de cloud privé virtuel (VPC) : les journaux de flux de VPC sont une fonction qui vous permet de capturer des informations sur le trafic IP allant et venant des interfaces réseau de votre VPC. Les données des journaux de flux peuvent être publiées sur

Amazon CloudWatch Logs et Amazon Simple Storage Service (Amazon S3). Une fois que vous avez créé un journal de flux, vous pouvez extraire et afficher ses données dans la destination de votre choix.

- Envisager la mise en miroir du trafic VPC : la mise en miroir du trafic est une fonction Amazon VPC que vous pouvez utiliser pour copier le trafic réseau à partir d'une interface réseau Elastic d'instances Amazon Elastic Compute Cloud (Amazon EC2), puis l'envoyer à des appareils de sécurité et de surveillance hors bande pour l'inspection du contenu, la surveillance des menaces et le dépannage.
  - Mise en miroir du trafic VPC

Ressources

Documents connexes :

- AWS Firewall Manager
- Amazon Inspector
- Sécurité du Amazon VPC
- Démarrer avec AWS WAF

Vidéos connexes :

- AWS Transit Gateway reference architectures for many VPCs
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield

Exemples connexes :

- Atelier : Déploiement automatisé d'un VPC

SÉC 6. De quelle façon pouvez-vous assurer la protection de vos ressources de calcul ?

Les ressources de calcul de votre charge de travail nécessitent plusieurs couches de défense pour vous aider à vous protéger des menaces externes et internes. Les ressources de calcul incluent les instances EC2, les conteneurs, les fonctions AWS Lambda, les services de bases de données, les appareils IoT, etc.

Bonnes pratiques

- SEC06-BP01 Gérer les failles
- SEC06-BP02 Réduire la surface d'attaque
- SEC06-BP03 Mettre en œuvre des services gérés
- SEC06-BP04 Automatiser la protection du calcul
- SEC06-BP05 Permettre aux utilisateurs d'effectuer des actions à distance
- SEC06-BP06 Valider l'intégrité des logiciels

SEC06-BP01 Gérer les failles

Analysez et éliminez fréquemment les failles de sécurité dans votre code, vos dépendances et votre infrastructure afin de vous protéger contre les nouvelles menaces.

Résultat souhaité : créer et gérer un programme de gestion des failles. Analysez et corrigez régulièrement les ressources telles que les instances Amazon EC2, les conteneurs Amazon Elastic Container Service (Amazon ECS) et les charges de travail Amazon Elastic Kubernetes Service (Amazon EKS). Configurez des fenêtres de maintenance pour les ressources gérées par AWS, par exemple les bases de données Amazon Relational Database Service (Amazon RDS). Utilisez l'analyse de code statique pour rechercher des problèmes courants dans le code source de l'application. Envisagez de tester la pénétration des applications web si votre organisation possède les compétences requises ou peut recruter de l'aide externe.

Anti-modèles courants :

- L'absence de programme de gestion des failles.
- L'application de correctifs système sans tenir compte de la gravité ni de l'évitement des risques.
- Utilisation d'un logiciel dont la date de fin de vie (EOL) a été dépassée.
- Déploiement du code en production avant de l'analyser afin de détecter tout problème de sécurité.

Avantages liés à l'instauration de cette bonne pratique :

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

Un programme de gestion des failles comprend l'évaluation de la sécurité, l'identification des problèmes, l'établissement des priorités et la mise en œuvre des correctifs dans le cadre de la

résolution des problèmes. L'automatisation est la clé pour analyser continuellement les charges de travail afin de détecter les problèmes, l'exposition involontaire du réseau et la mise en œuvre de mesures correctives. L'automatisation de la création et de la mise à jour des ressources permet de gagner du temps et de réduire le risque d'erreurs de configuration, ce qui crée d'autres problèmes. Un programme de gestion des failles bien conçu doit également tenir compte des tests de vulnérabilité pendant les étapes de développement et de déploiement du cycle de vie du logiciel. L'implémentation de la gestion des failles pendant le développement et le déploiement réduit les risques d'avoir une vulnérabilité dans votre environnement de production.

L'implémentation d'un programme de gestion des failles nécessite une bonne compréhension du modèle de responsabilité partagée AWS et de la façon dont cela affecte vos charges de travail spécifiques. Dans le cadre du modèle de responsabilité partagée, AWS est responsable de la protection de l'infrastructure du AWS Cloud. Cette infrastructure se compose de matériels, de logiciels, de réseaux et d'installations exécutant les services du AWS Cloud. Vous êtes responsable de la sécurité dans le cloud, par exemple des données réelles, de la configuration de la sécurité et des tâches de gestion des instances Amazon EC2. Vous devez également vérifier que les objets Amazon S3 sont classés et configurés correctement. Votre approche en matière de gestion des failles peut également varier selon les services que vous utilisez. Par exemple, AWS gère l'application des correctifs pour notre service de base de données relationnelle gérée, Amazon RDS, mais vous êtes responsable de l'application des correctifs dans les bases de données auto-hébergées.

AWS offre une gamme de services pour vous aider dans le cadre de votre programme de gestion des failles. Amazon Inspector analyse continuellement les charges de travail AWS afin d'identifier les problèmes de logiciels et les accès réseau involontaires. Le Gestionnaire de correctifs d'AWS Systems Manager permet de gérer l'application des correctifs sur vos instances Amazon EC2. Amazon Inspector et Systems Manager peuvent être consultés dans AWS Security Hub, un service de gestion de la situation de sécurité dans le cloud qui aide à automatiser les contrôles de sécurité AWS et à centraliser les alertes de sécurité.

Amazon CodeGuru permet d'identifier les problèmes potentiels dans les applications Java et Python en utilisant l'analyse de code statique.

Étapes d'implémentation

- Configurez Amazon Inspector : Amazon Inspector détecte automatiquement les instances Amazon EC2 qui viennent d'être lancées, les fonctions Lambda et les images de conteneur éligibles envoyées dans Amazon ECR, et il les analyse immédiatement afin d'identifier les problèmes de logiciels, les défauts potentiels et tout exposition involontaire du réseau.

- Analysez le code source : rechercher les problèmes et les défauts dans les bibliothèques et les dépendances. Amazon CodeGuru peut analyser et recommander des mesures correctives pour les problèmes de sécurité courants dans les applications Java et Python. La Fondation OWASP publie une liste des outils d'analyse de code source (également appelés outils SAST).
- Implémentez un mécanisme permettant d'analyser et de corriger votre environnement existant, ainsi qu'une analyse dans le cadre d'un processus de création de pipeline CI/CD : implémentez un mécanisme pour analyser et corriger les problèmes dans vos dépendances et systèmes d'exploitation afin de garantir votre protection contre les nouvelles menaces. Exécutez ce mécanisme régulièrement. La gestion des failles logicielles est essentielle pour comprendre où vous devez appliquer les correctifs ou résoudre les problèmes logiciels. Privilégiez la correction des problèmes de sécurité potentiels en intégrant rapidement les évaluations des vulnérabilités à votre pipeline d'intégration continue et de livraison continue (CI/CD). Votre approche peut varier en fonction des services AWS que vous utilisez. Pour vérifier l'absence de problèmes potentiels dans le logiciel exécuté dans les instances Amazon EC2, ajoutez Amazon Inspector à votre pipeline pour vous alerter et arrêter le processus de construction si des problèmes ou des défauts potentiels sont détectés. Amazon Inspector surveille les ressources en continu. Vous pouvez également utiliser des produits open source tels que OWASP Dependency-Check, Snyk, OpenVAS, des gestionnaires de packages et des outils AWS Partner pour la gestion des failles.
- Utilisez AWS Systems Manager : vous êtes responsable de la gestion des correctifs pour vos ressources AWS, notamment les instances Amazon Elastic Compute Cloud (Amazon EC2), les Amazon Machine Images (AMI) et d'autres ressources de calcul. Le Gestionnaire de correctifs AWS Systems Manager automatise le processus d'application de correctifs aux instances gérées avec des mises à jour liées à la sécurité et d'autres types de mises à jour. Le Gestionnaire de correctifs peut être utilisé pour appliquer des correctifs sur les instances Amazon EC2 pour les systèmes d'exploitation et les applications, dont les applications Microsoft, les packs de service Windows et les mises à niveau mineures pour les instances basées sur Linux. Outre Amazon EC2, le Gestionnaire de correctifs peut également être utilisé pour corriger les serveurs sur site.

  Pour obtenir la liste des systèmes d'exploitation compatibles, consultez Systèmes d'exploitation compatibles dans le guide utilisateur Systems Manager. Vous pouvez analyser les instances pour afficher uniquement un rapport sur les correctifs manquants, ou vous pouvez analyser et installer automatiquement tous les correctifs manquants.
- Utilisez AWS Security Hub : Security Hub fournit une vue complète de votre situation de sécurité dans AWS. Il collecte des données de sécurité dans plusieurs services AWS et communique ces découvertes dans un format normalisé, ce qui vous permet d'établir l'ordre de priorité des découvertes en matière de sécurité dans les services AWS.

- Utilisez AWS CloudFormation : AWS CloudFormation est un service d'infrastructure en tant que code (IaC) qui permet la gestion des failles en automatisant le déploiement des ressources et en normalisant l'architecture des ressources sur plusieurs comptes et environnements.

Ressources

Documents connexes :

- AWS Systems Manager
- Security Overview of AWS Lambda
- Amazon CodeGuru
- Improved, Automated Vulnerability Management for Cloud Workloads with a New Amazon Inspector
- Automate vulnerability management and remediation in AWS using Amazon Inspector and AWS Systems Manager – Part 1

Vidéos connexes :

- Securing Serverless and Container Services
- Security best practices for the Amazon EC2 instance metadata service

SEC06-BP02 Réduire la surface d'attaque

Réduisez votre exposition aux accès imprévus en renforçant les systèmes d'exploitation et en limitant au minimum l'utilisation des composants, des bibliothèques et des services consommables en externe. Commencez par réduire les composants inutilisés, qu'il s'agisse de packages de système d'exploitation ou d'applications pour les charges de travail basées sur Amazon Elastic Compute Cloud (Amazon EC2), ou de modules logiciels externes dans votre code (pour toutes les charges de travail). Il existe de nombreux guides sur le renforcement et la configuration de la sécurité pour les systèmes d'exploitation et les logiciels serveur courants. Par exemple, vous pouvez commencer par le Center for Internet Security et partir de là.

Dans Amazon EC2, vous pouvez créer vos propres Amazon Machine Images (AMI), auxquelles vous avez appliqué des correctifs et dont vous avez renforcé la sécurité, pour vous aider à répondre aux exigences de sécurité spécifiques de votre organisation. Les correctifs et autres contrôles de sécurité que vous appliquez sur l'AMI sont effectifs au moment où ils ont été créés. Ils ne sont

pas dynamiques, sauf si vous les modifiez après le lancement, par exemple, avec AWS Systems Manager.

Vous pouvez simplifier le processus de création d'AMI sécurisées avec EC2 Image Builder. EC2 Image Builder réduit considérablement l'effort requis pour créer et préserver des images de référence sans avoir à écrire ni à gérer l'automatisation. Lorsque des mises à jour logicielles sont disponibles, Image Builder génère automatiquement une nouvelle image sans obliger les utilisateurs à lancer manuellement les créations d'image. EC2 Image Builder vous permet de valider facilement la fonctionnalité et la sécurité de vos images avant de les utiliser en production avec les tests fournis par AWS et vos propres tests. Vous pouvez également appliquer les paramètres de sécurité fournis par AWS pour sécuriser davantage vos images afin de répondre aux critères de sécurité internes. Par exemple, vous pouvez créer des images conformes à la norme Security Technical Implementation Guide (STIG) à l'aide de modèles fournis par AWS.

Grâce à des outils d'analyse de code statique tiers, vous pouvez identifier les problèmes de sécurité courants tels que les limites d'entrée de fonction non contrôlées, ainsi que les vulnérabilités et expositions courantes applicables. Vous pouvez utiliser Amazon CodeGuru pour les langues prises en charge. Les outils de vérification des dépendances peuvent également être utilisés pour déterminer si les bibliothèques avec lesquelles votre code est lié sont les dernières versions, sont elles-mêmes exemptes de vulnérabilités et d'expositions courantes et ont des conditions de licence qui répondent aux exigences de votre politique logicielle.

À l'aide d'Amazon Inspector, vous pouvez effectuer des évaluations de configuration de vos instances pour identifier les vulnérabilités et expositions communes connues, les évaluer par rapport à des points de référence en matière de sécurité et automatiser la notification des défauts. Amazon Inspector s'exécute sur des instances de production ou dans un pipeline de conception et notifie les développeurs et ingénieurs lorsque les résultats sont prêts. Vous pouvez accéder aux résultats par programmation et diriger votre équipe vers les systèmes de suivi des bugs et des retards. EC2 Image Builder peut être utilisé pour gérer les images de serveur (AMI) avec l'application de correctifs automatisée, l'application de politiques de sécurité fournies par AWS et d'autres personnalisations. Lorsque vous utilisez des conteneurs, mettez en œuvre l' analyse d'image ECR dans votre pipeline de génération et régulièrement par rapport à votre référentiel d'images pour rechercher les failles CVE dans vos conteneurs.

Si Amazon Inspector et d'autres outils sont efficaces pour identifier les configurations et les failles CVE présentes, d'autres méthodes sont nécessaires pour tester votre charge de travail au niveau de l'application. Le fuzzing est une méthode bien connue pour trouver des bugs en utilisant

l'automatisation pour injecter des données malformées dans les champs de saisie et d'autres parties de votre application.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Renforcer la sécurité du système d'exploitation : configurez les systèmes d'exploitation de manière à respecter les bonnes pratiques.
  - Sécurisation d'Amazon Linux
  - Sécurisation de Microsoft Windows Server
- Renforcer la sécurité des ressources conteneurisées : configurez les ressources conteneurisées de manière à respecter les bonnes pratiques en matière de sécurité.
- Implémentez les bonnes pratiques AWS Lambda.
  - Bonnes pratiques AWS Lambda

Ressources

Documents connexes :

- AWS Systems Manager
- Remplacement d'un hôte bastion par Amazon EC2 Systems Manager
- Présentation de la sécurité d'AWS Lambda

Vidéos connexes :

- Exécution de charges de travail à haute sécurité sur Amazon EKS
- Sécurisation des services sans serveur et de conteneur
- Bonnes pratiques de sécurité pour le service des métadonnées d'instance Amazon EC2

Exemples connexes :

- Atelier : Déploiement automatisé d'un pare-feu d'application Web

SEC06-BP03 Mettre en œuvre des services gérés

Mettez en œuvre des services qui gèrent les ressources, comme Amazon Relational Database Service (Amazon RDS), AWS Lambda, and Amazon Elastic Container Service (Amazon ECS), afin de réduire vos tâches de maintenance de la sécurité dans le cadre du modèle de responsabilité partagée. Par exemple, Amazon RDS vous aide à configurer, exploiter et dimensionner une base de données relationnelle, automatise les tâches d'administration telles que la mise en service du matériel, la configuration de base de données, l'application de correctifs et les sauvegardes. Cela signifie que vous pouvez consacrer plus de temps à la sécurisation de votre application selon les autres méthodes décrites dans le cadre AWS Well-Architected Framework. Lambda vous permet d'exécuter le code sans avoir à mettre en service ni à gérer des serveurs. Par conséquent, vous pouvez vous concentrer uniquement sur la connectivité, les appels et la sécurité au niveau du code, et non pas sur l'infrastructure ou le système d'exploitation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Explorer les services disponibles : explorez, testez et implémentez des services qui gèrent des ressources, notamment Amazon RDS, AWS Lambda et Amazon ECS.

Ressources

Documents connexes :

- Site web AWS
- AWS Systems Manager
- Remplacement d'un hôte bastion par Amazon EC2 Systems Manager
- Présentation de la sécurité d'AWS Lambda

Vidéos connexes :

- Exécution de charges de travail à haute sécurité sur Amazon EKS
- Sécurisation des services sans serveur et de conteneur
- Bonnes pratiques de sécurité pour le service des métadonnées d'instance Amazon EC2

Exemples connexes :

- Atelier : Demander un certificat public avec AWS Certificate Manager

SEC06-BP04 Automatiser la protection du calcul

Automatisez vos mécanismes de protection du calcul, en particulier la gestion des failles, la réduction de la surface d'attaque et la gestion des ressources. L'automatisation vous aide à investir du temps pour sécuriser d'autres aspects de votre charge de travail, et à réduire le risque d'erreur humaine.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Automatiser la gestion de la configuration : appliquez et validez des configurations sécurisées automatiquement à l'aide d'un service ou d'un outil de gestion de la configuration.
  - AWS Systems Manager
  - AWS CloudFormation
  - Atelier : Déploiement automatisé d'un VPC
  - Atelier : Déploiement automatisé d'une application Web avec EC2

- Automatiser l'application de correctifs aux instances Amazon Elastic Compute Cloud (Amazon EC2) : AWS Systems Manager Patch Manager automatise le processus de correction des instances gérées avec des mises à jour liées à la sécurité et d'autres types de mises à jour. Vous pouvez utiliser le gestionnaire de correctifs pour appliquer des correctifs aux systèmes d'exploitation et aux applications.
  - le gestionnaire de correctifs AWS Systems Manager
  - Application centralisée de correctifs multicomptes et multirégions avec AWS Systems Manager Automation

- Mettre en place une détection et une prévention des intrusions : mettez en place un outil de détection et de prévention des intrusions pour surveiller et arrêter les opérations malveillantes au niveau des instances.
- Envisager les solutions AWS Partner : les partenaires AWS proposent des centaines de produits leaders du secteur qui sont équivalents, identiques ou s'intègrent aux contrôles existants dans vos environnements sur site. Ces produits complètent les services AWS existants pour vous permettre de déployer une architecture de sécurité exhaustive et une expérience plus homogène dans vos environnements cloud et sur site.

- Sécurité de l'infrastructure

Ressources

Documents connexes :

- AWS CloudFormation
- AWS Systems Manager
- le gestionnaire de correctifs AWS Systems Manager
- Application centralisée de correctifs multicomptes et multirégions avec AWS Systems Manager Automation
- Sécurité de l'infrastructure
- Remplacement d'un hôte bastion par Amazon EC2 Systems Manager
- Présentation de la sécurité d'AWS Lambda

Vidéos connexes :

- Exécution de charges de travail à haute sécurité sur Amazon EKS
- Sécurisation des services sans serveur et de conteneur
- Bonnes pratiques de sécurité pour le service des métadonnées d'instance Amazon EC2

Exemples connexes :

- Atelier : Déploiement automatisé d'un pare-feu d'application Web
- Atelier : Déploiement automatisé d'une application Web avec EC2

SEC06-BP05 Permettre aux utilisateurs d'effectuer des actions à distance

La suppression de la possibilité d'accès interactif réduit le risque d'erreur humaine et le potentiel de configuration ou de gestion manuelle. Par exemple, utilisez un flux de travail de gestion des modifications pour déployer des instances Amazon Elastic Compute Cloud (Amazon EC2) à l'aide d'une infrastructure en tant que code, puis gérez les instances Amazon EC2 avec des outils tels qu'AWS Systems Manager au lieu d'autoriser un accès direct ou via un hôte bastion. AWS Systems Manager automatise diverses tâches de maintenance et de déploiement à l'aide de fonctionnalités telles que l' automatisation (flux de travail), les documents (playbooks) et la Run Command. Les

piles AWS CloudFormation sont construites à partir de pipelines et peuvent automatiser les tâches de déploiement et de gestion de votre infrastructure sans utiliser directement AWS Management Console ni les API.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Remplacer l'accès à la console : remplacez l'accès (protocole SSH ou RDP) aux instances depuis la console par AWS Systems Manager Run Command pour automatiser les tâches de gestion.

- AWS Systems Manager Run Command

Ressources

Documents connexes :

- AWS Systems Manager
- AWS Systems Manager Run Command
- Remplacement d'un hôte bastion par Amazon EC2 Systems Manager
- Présentation de la sécurité d'AWS Lambda

Vidéos connexes :

- Exécution de charges de travail à haute sécurité sur Amazon EKS
- Sécurisation des services sans serveur et de conteneur
- Bonnes pratiques de sécurité pour le service des métadonnées d'instance Amazon EC2

Exemples connexes :

- Atelier : Déploiement automatisé d'un pare-feu d'application Web

SEC06-BP06 Valider l'intégrité des logiciels

Mettez en place des mécanismes (signature de code) pour confirmer que les logiciels, le code et les bibliothèques utilisés dans la charge de travail proviennent de sources fiables et n'ont pas été altérés. Par exemple, vous devez vérifier le certificat de signature de code des fichiers binaires et des scripts

pour vérifier l'auteur, et vous assurer qu'il n'a pas été altéré depuis sa création par l'auteur. AWS Signer contribue à garantir la confiance et l'intégrité de votre code en gérant de manière centralisée le cycle de vie de la signature de code, y compris la certification de la signature et les clés publiques et privées. Vous pouvez apprendre à utiliser des modèles avancés et les bonnes pratiques en matière de signature de code avec AWS Lambda. En outre, un total de contrôle du logiciel que vous téléchargez, comparé à celui du fournisseur, peut garantir qu'il n'a pas été altéré.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Étudier les mécanismes : la signature de code est un mécanisme qui peut être utilisé pour valider l'intégrité des logiciels.
  - NIST : considérations de sécurité pour la signature de code

Ressources

Documents connexes :

- AWS Signer
- Nouveau : la signature de code, un contrôle de confiance et d'intégrité pour AWS Lambda

Protection des données

Questions

- SÉC 7. Comment classer vos données ?
- SÉC 8. Comment protéger les données au repos ?
- SÉC 9. Comment protéger vos données en transit ?

SÉC 7. Comment classer vos données ?

La classification des données fournit un moyen de classer les données en fonction de leur importance et de leur sensibilité afin de vous aider à déterminer les contrôles de protection et de conservation appropriés.

Bonnes pratiques

- SEC07-BP01 Identifier les données au sein de votre charge de travail

- SEC07-BP02 Définir les contrôles de protection des données
- SEC07-BP03 Automatiser l'identification et la classification
- SEC07-BP04 Définir la gestion du cycle de vie des données

SEC07-BP01 Identifier les données au sein de votre charge de travail

Il est essentiel de comprendre le type et la classification des données que votre charge de travail traite, les processus d'entreprise associés, l'endroit où les données sont stockées et qui est le propriétaire des données. Vous devez également connaître les exigences légales et de conformité applicables à votre charge de travail, et savoir quels contrôles de données doivent être appliqués. L'identification des données est la première étape du processus de classification des données.

Avantages liés à l'instauration de cette bonne pratique :

La classification des données permet aux responsables de la charge de travail d'identifier les emplacements qui stockent des données sensibles et de déterminer comment ces données doivent être consultées et partagées.

La classification des données vise à répondre aux questions suivantes :

- De quel type de données disposez-vous ?

  Il peut s'agir de données telles que :

  - Propriété intellectuelle, comme les secrets commerciaux, les brevets ou les contrats.
  - Informations de santé protégées (PH), comme les dossiers médicaux qui contiennent des informations sur les antécédents médicaux d'une personne.
  - Données d'identification personnelle (PII), comme le nom, l'adresse, la date de naissance et le numéro d'identification ou d'enregistrement national.
  - Données de carte bancaire, comme le numéro de compte principal (PAN), le nom du titulaire de la carte, la date d'expiration et le numéro de code de service.
  - Où les données sensibles sont-elles stockées ?
  - Qui peut consulter, modifier et supprimer des données ?
  - Il est essentiel de comprendre les autorisations des utilisateurs pour éviter toute manipulation inappropriée des données.
- Qui peut effectuer des opérations de création, de lecture, de mise à jour et de suppression (CRUD) ?

  - Tenez compte de l'escalade potentielle des privilèges en comprenant qui peut gérer les autorisations d'accès aux données.
- Quelles sont les répercussions sur les activités si les données sont divulguées involontairement, modifiées ou supprimées ?
  - Comprenez les conséquences du risque si des données sont modifiées, supprimées ou divulguées par inadvertance.

En connaissant les réponses à ces questions, vous pouvez prendre les mesures suivantes :

- Réduire la portée des données sensibles (comme le nombre d'emplacements de données sensibles) et limiter l'accès aux données sensibles aux utilisateurs approuvés uniquement.
- Comprendre les différents types de données afin de pouvoir implémenter des mécanismes et des techniques de protection des données appropriés, tels que le chiffrement, la prévention de la perte de données et la gestion des identités et des accès.
- Optimiser les coûts en fournissant les bons objectifs de contrôle pour les données.
- Répondre en toute confiance aux questions des organismes de réglementation et des vérificateurs concernant les types et la quantité de données, et la façon dont les données de sensibilités différentes sont isolées les unes des autres.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

La classification des données consiste à déterminer la sensibilité des données. Il peut s'agir de baliser les données pour les rendre facilement interrogeables et traçables. La classification des données réduit également la duplication des données, ce qui peut permettre de réduire les coûts de stockage et de sauvegarde, tout en accélérant le processus de recherche.

Utilisez des services tels que Amazon Macie pour automatiser à grande échelle la découverte et la classification des données sensibles. D'autres services, comme Amazon EventBridge et AWS Config, peuvent être utilisés pour automatiser la correction des problèmes de sécurité des données, comme les compartiments Amazon Simple Storage Service (Amazon S3) non chiffrés et les volumes EBS Amazon EC2 ou les ressources de données non balisées. Pour obtenir une liste complètes des intégrations de service AWS, consultez la documentation EventBridge.

La détection des PII dans les données non structurées, comme les e-mails des clients, les tickets de support, les examens de produits et les réseaux sociaux, peut être effectuée en utilisant Amazon

Comprehend, qui est un service de traitement du langage naturel (NLP) qui utilise le machine learning (ML) pour trouver des idées et des relations, comme les personnes, les lieux, les sentiments et les sujets dans un texte non structuré. Pour une liste des services AWS qui peuvent faciliter l'identification des données, consultez Techniques courantes pour détecter les données PHI et PII à l'aide des services AWS.

Une autre méthode qui prend en charge la classification et la protection des données est le balisage des ressources AWS. Le balisage vous permet d'attribuer des métadonnées à vos ressources AWS pour vous aider à gérer, identifier, organiser, rechercher et filtrer ces dernières.

Dans certains cas, vous pouvez choisir de baliser des ressources entières (comme un compartiment S3), surtout lorsqu'une charge de travail ou un service particulier est censé stocker des processus ou des transmissions d'une classification de données déjà connue.

Le cas échéant, vous pouvez baliser un compartiment S3 au lieu d'objets individuels pour faciliter l'administration et la maintenance de la sécurité.

Étapes d'implémentation

Détectez les données sensibles dans Amazon S3 :

1. Avant de commencer, assurez-vous de disposer des autorisations appropriées pour accéder à la console Amazon Macie et aux opérations d'API. Pour plus d'informations, consultez Mise en route de Amazon Macie.
2. Utilisez Amazon Macie pour effectuer la découverte automatisée des données lorsque vos données sensibles résident dans Amazon S3.
   - Utilisez le guide Mise en route de Amazon Macie afin de configurer un référentiel pour les résultats de découverte de données sensibles et de créer une tâche de découverte des données sensibles.
   - Comment utiliser Amazon Macie pour prévisualiser les données sensibles dans les compartiments S3.

     Par défaut, Macie analyse les objets en utilisant l'ensemble d'identifiants de données gérés que nous recommandons pour la découverte automatisée de données sensibles. Vous pouvez personnaliser l'analyse en configurant Macie de façon à utiliser des identifiants de données gérés spécifiques, des identifiants de données personnalisés et des listes d'autorisations lorsqu'il effectue la découverte automatisée des données sensibles pour votre compte ou votre organisation. Vous pouvez ajuster la portée de l'analyse en excluant des compartiments

spécifiques (par exemple, les compartiments S3 qui stockent généralement les données de journalisation AWS).

3. Pour configurer et utiliser la découverte automatisée de données sensibles, consultez Effectuer la découverte automatisée des données sensibles avec Amazon Macie.
4. Consultez également Découverte automatisée des données pour Amazon Macie.

Détectez les données sensibles dans Amazon RDS :

pour plus d'informations sur la découverte des données dans les bases de données Amazon Relational Database Service (Amazon RDS), consultez Classification des données pour la base de données Amazon RDS avec Macie.

Détectez les données sensibles dans DynamoDB :

- la section Détection des données sensibles dans DynamoDB avec Macie explique comment utiliser Amazon Macie pour détecter les données sensibles dans les tables Amazon DynamoDB en exportant les données dans Amazon S3 en vue de leur analyse.

Solutions de partenaires AWS :

- Envisagez d'utiliser notre AWS Partner Network étendu. Les partenaires AWS disposent d'outils complets et de frameworks de conformité qui s'intègrent directement aux services AWS. Les partenaires peuvent vous fournir une solution de gouvernance et de conformité adaptée à vos besoins organisationnels.
- Pour plus d'informations sur les solutions personnalisées de classification des données, consultez Gouvernance des données à l'ère de la réglementation et exigences de conformité.

Vous pouvez appliquer automatiquement les normes de balisage que votre organisation adopte en créant et en déployant des politiques avec AWS Organizations. Les politiques de balises vous permettent de spécifier les règles qui définissent les noms de clés valides et les valeurs valides pour chaque clé. Vous pouvez choisir de surveiller uniquement, ce qui vous donne la possibilité d'évaluer et de nettoyer vos balises existantes. Une fois que vos balises sont conformes aux normes que vous avez choisies, vous pouvez activer l'application des politiques relatives aux balises pour empêcher la création de balises non conformes. Pour plus d'informations, consultez Sécurisation des balises de ressources utilisées pour l'autorisation à l'aide d'une politique de contrôle des services dans AWS

Organizations et l'exemple de politique sur les solutions pour éviter la modification des balises, sauf par les principaux autorisés..

- Pour commencer à utiliser des politiques de balises dans AWS Organizations, il est vivement recommandé de suivre le workflow dans Mise en route des politiques de balises avant de passer à des politiques de balises plus avancées. Comprendre les effets d'une simple politique de balise sur un seul compte avant de l'étendre à toute une unité d'organisation (UO) ou une organisation vous permet de voir les effets d'une politique de balise avant d'appliquer la politique de balise. Mise en route des politiques de balises fournit des liens vers des instructions relatives à des tâches plus avancées en matière de politiques.
- Envisagez d'évaluer d'autres services et fonctionnalités AWS qui prennent en charge la classification des données, comme indiqué dans le livre blanc sur la classification des données.

Ressources

Documents connexes :

- Getting started with Amazon Macie (Mise en route de Amazon Macie)
- Automated data discovery with Amazon Macie
- Mise en route des politiques de balises
- Detecting PII entities (Détecter les entités PII)

Blogs connexes :

- Comment utiliser Amazon Macie pour prévisualiser les données sensibles dans les compartiments S3.
- Performing automated sensitive data discovery with Amazon Macie.
- Common techniques to detect PHI and PII data using AWS Services
- Detecting and redacting PII using Amazon Comprehend
- Securing resource tags used for authorization using a service control policy in AWS Organizations
- Enabling data classification for Amazon RDS database with Macie
- Detecting sensitive data in DynamoDB with Macie
-

Vidéos connexes :

- Event-driven data security using Amazon Macie
- Amazon Macie for data protection and governance
- Fine-tune sensitive data findings with allow lists

SEC07-BP02 Définir les contrôles de protection des données

Protégez les données en fonction de leur niveau de classification. Par exemple, sécurisez les données classées comme publiques à l'aide des recommandations pertinentes tout en protégeant les données sensibles grâce à des contrôles supplémentaires.

En utilisant des balises de ressource, des comptes AWS séparés par sensibilité (et éventuellement aussi pour chaque mise en garde, isolement ou communauté d'intérêt), les politiques IAM, les politiques de contrôle des services (SCP) AWS Organizations, AWS Key Management Service (AWS KMS) et AWS CloudHSM, vous pouvez définir et mettre en œuvre vos politiques de classification et de protection des données avec chiffrement. Par exemple, si vous disposez d'un projet avec des compartiments S3 qui contiennent des données hautement critiques ou des instances Amazon Elastic Compute Cloud (Amazon EC2) qui traitent des données confidentielles, ils peuvent être marqués avec une balise `Project=ABC` . Seule votre équipe immédiate sait ce que le code du projet signifie, et cela permet d'utiliser un contrôle d'accès basé sur les attributs. Vous pouvez définir des niveaux d'accès aux clés de chiffrement AWS KMS par le biais de politiques de clés et d'autorisations afin de garantir que seuls les services appropriés ont accès au contenu sensible par un mécanisme sécurisé. Si vous prenez des décisions d'autorisation basées sur des balises, vous devez vous assurer que les autorisations sur les balises sont définies de manière appropriée en utilisant des politiques de balises dans AWS Organizations.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Définir votre schéma d'identification et de classification des données : l'identification et la classification de vos données sont faites pour évaluer l'impact potentiel et le type de données que vous stockez et qui peut y accéder.
  - Documentation AWS
- Découvrir les contrôles AWS disponibles : découvrez les contrôles de sécurité pour les services AWS que vous utilisez ou prévoyez d'utiliser. De nombreux services disposent d'une section relative à la sécurité dans leur documentation.
  - Documentation AWS

- Identifier les ressources de conformité AWS : identifiez les ressources que propose AWS pour vous aider.
  - https://aws.amazon.com/compliance/

Ressources

Documents connexes :

- Documentation AWS
- Livre blanc sur la classification des données
- Démarrer avec Amazon Macie
- Texte manquant

Vidéos connexes :

- Introducing the New Amazon Macie

SEC07-BP03 Automatiser l'identification et la classification

l'automatisation de l'identification et de la classification des données peut vous aider à mettre en œuvre les contrôles appropriés. Le recours à l'automatisation en la circonstance plutôt qu'à l'accès direct d'une personne réduit le risque d'erreur humaine et d'exposition. Vous devez évaluer, en utilisant un outil tel qu' Amazon Macie, qui utilise le machine learning pour découvrir, catégoriser et protéger les données sensibles dans AWS. Amazon Macie reconnaît les données sensibles en tant que données d'identification personnelle (PII) ou propriété intellectuelle, et génère des tableaux de bord et des alertes pour vous offrir de la visibilité sur les méthodes de déplacement ou d'accès à ces données.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Utiliser l'inventaire Amazon Simple Storage Service (Amazon S3) : l'inventaire Amazon S3 est l'un des outils que vous pouvez utiliser pour auditer et signaler le statut de réplication et de chiffrement de vos objets.
  - Inventaire Amazon S3

- Envisager Amazon Macie : Amazon Macie utilise le machine learning pour détecter et classer automatiquement les données stockées dans Amazon S3.
  - Amazon Macie

Ressources

Documents connexes :

- Amazon Macie
- Inventaire Amazon S3
- Livre blanc sur la classification des données
- Mise en route avec Amazon Macie

Vidéos connexes :

- Introducing the New Amazon Macie

SEC07-BP04 Définir la gestion du cycle de vie des données

votre stratégie de cycle de vie définie doit être basée sur le niveau de sensibilité, ainsi que sur les exigences légales et organisationnelles. Des aspects, tels que la durée de conservation des données, les processus de destruction des données, la gestion de l'accès aux données, la transformation des données et le partage des données, doivent être pris en compte. Lorsque vous choisissez une méthodologie de classification des données, équilibrez l'utilisabilité par rapport à l'accès. Vous devez également tenir compte des multiples niveaux d'accès et des nuances pour mettre en œuvre une approche sécurisée, mais toujours utilisable, pour chaque niveau. Utilisez toujours une approche de défense en profondeur et réduisez l'accès humain aux données et aux mécanismes de transformation, de suppression ou de copie des données. Par exemple, exigez que les utilisateurs s'authentifient d'une manière forte auprès d'une application, et donnez à l'application, plutôt qu'aux utilisateurs, l'autorisation d'accès requise pour effectuer une « action à distance ». En outre, veillez à ce que les utilisateurs proviennent d'un chemin de réseau approuvé et aient besoin d'un accès aux clés de déchiffrement. Utiliser des outils, tels que des tableaux de bord et des rapports automatisés, pour donner aux utilisateurs des informations à partir des données plutôt que de leur fournir un accès direct aux données.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Identifier les types de données : identifiez les types de données que vous stockez ou traitez dans votre charge de travail. Ces données peuvent être du texte, des images, des bases de données binaires, etc.

Ressources

Documents connexes :

- Livre blanc sur la classification des données
- Mise en route avec Amazon Macie

Vidéos connexes :

- Introducing the New Amazon Macie

SÉC 8. Comment protéger les données au repos ?

Protégez vos données au repos en mettant en œuvre plusieurs contrôles, afin de réduire le risque d'accès non autorisé ou de mauvaise gestion.

Bonnes pratiques

- SEC08-BP01 Mise en œuvre de la gestion sécurisée des clés
- SEC08-BP02 Appliquer le chiffrement au repos
- SEC08-BP03 Automatiser la protection des données au repos
- SEC08-BP04 Appliquer le contrôle d'accès
- SEC08-BP05 Utiliser des mécanismes pour protéger l'accès aux données

SEC08-BP01 Mise en œuvre de la gestion sécurisée des clés

La gestion sécurisée des clés inclut le stockage, la rotation, le contrôle d'accès et la surveillance des informations sur les clés nécessaires pour sécuriser les données inactives adaptées à votre charge de travail.

Résultat souhaité : Un mécanisme de gestion des clés évolutif, reproductible et automatisé. Ce mécanisme devrait permettre de faire respecter le principe du moindre privilège d'accès aux

informations sur les clés et de trouver le juste équilibre entre la disponibilité des clés, la confidentialité et l'intégrité. L'accès aux clés doit être surveillé et les informations sur les clés doivent être alternées par le biais d'un processus automatisé. Les clés ne doivent jamais être accessibles aux identités humaines.

Anti-modèles courants :

- Accès humain à des informations sur les clés non chiffrées.
- Création d'algorithmes cryptographiques personnalisés.
- Autorisations trop larges pour accéder aux informations sur les clés.

Avantages liés au respect de cette bonne pratique : En établissant un mécanisme sécurisé de gestion des clés pour votre charge de travail, vous contribuez à protéger votre contenu contre tout accès non autorisé. En outre, vous pouvez être soumis à des exigences réglementaires en matière de chiffrement de vos données. Une solution efficace de gestion des clés peut fournir des mécanismes techniques conformes à ces réglementations afin de protéger les informations sur les clés.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

De nombreuses exigences réglementaires et bonnes pratiques incluent le chiffrement des données au repos en tant que contrôle de sécurité fondamental. Afin de respecter ce contrôle, votre charge de travail a besoin d'un mécanisme permettant de stocker et de gérer en toute sécurité les informations sur les clés utilisées pour chiffrer vos données au repos.

AWS propose AWS Key Management Service (AWS KMS) pour fournir un stockage durable, sécurisé et redondant pour les clés AWS KMS. De nombreux services AWS s'intègrent à AWS KMS pour prendre en charge le chiffrement de vos données. AWS KMS utilise des modules de sécurité matériels validés FIPS 140-2 niveau 3 pour protéger vos clés. Il n'existe aucun mécanisme permettant d'exporter les clés AWS KMS en texte brut.

Lors du déploiement de charges de travail à l'aide d'une stratégie multi-comptes, il est conseillé de conserver les clés AWS KMS dans le même compte que la charge de travail qui les utilise. Dans ce modèle distribué, la responsabilité de la gestion des clés AWS KMS incombe à l'équipe chargée de l'application. Dans d'autres cas d'utilisation, les entreprises peuvent choisir de stocker les clés AWS KMS dans un compte centralisé. Cette structure centralisée nécessite des politiques supplémentaires pour permettre l'accès intercompte requis afin que le compte de la charge de travail puisse accéder

aux clés stockées dans le compte centralisé, mais elle s'applique peut-être plus aux cas d'utilisation où une seule clé est partagée entre plusieurs Comptes AWS.

Quel que soit l'endroit où les informations sur les clés sont stockées, l'accès à la clé doit être étroitement contrôlé grâce à l'utilisation de politiques de clés et de stratégies IAM. Les politiques de clés constituent le principal moyen de contrôler l'accès à une clé AWS KMS. En outre, les octrois de clés AWS KMS peuvent donner accès à des services AWS permettant de chiffrer et de déchiffrer les données en votre nom. Prenez le temps de consulter les bonnes pratiques en matière de contrôle d'accès à vos clés AWS KMS.

Il est recommandé de surveiller l'utilisation des clés de chiffrement afin de détecter les modèles d'accès inhabituels. Les opérations effectuées à l'aide de clés gérées par AWS et de clés gérées par le client stockées dans AWS KMS peuvent être journalisées dans AWS CloudTrail et doivent être examinées périodiquement. Une attention particulière doit être accordée à la surveillance des événements de destruction des clés. Pour limiter la destruction accidentelle ou malveillante des informations sur les clés, les événements de destruction des clés ne suppriment pas immédiatement ces informations. Les tentatives de suppression de clés AWS KMS sont soumises à une période d'attente dont la durée par défaut est de 30 jours, ce qui laisse aux administrateurs le temps de vérifier ces actions et d'annuler la requête si nécessaire.

La plupart des services AWS utilisent AWS KMS de manière transparente pour vous. Vous n'avez qu'à décider si vous souhaitez utiliser une clé gérée par AWS ou une clé gérée par le client. Si votre charge de travail nécessite l'utilisation directe de AWS KMS pour chiffrer ou déchiffrer des données, il est conseillé de recourir au chiffrement d'enveloppe pour protéger vos données. Le kit SDK AWS Encryption peut fournir à vos applications des primitives de chiffrement côté client pour implémenter le chiffrement d'enveloppe et l'intégrer à AWS KMS.

Étapes d'implémentation

1. Déterminez les options appropriées de gestion des clés (gérées par AWS ou par le client).
    - Pour faciliter l'utilisation, AWS propose des clés AWS qui appartiennent au client et des clés gérées par AWS pour la plupart des services. Elles fournissent une fonctionnalité de chiffrement au repos sans qu'il soit nécessaire de gérer les informations sur les clés ou les politiques les concernant.
    - Lorsque vous utilisez des clés gérées par le client, pensez au key store par défaut afin de trouver le meilleur équilibre entre agilité, sécurité, souveraineté des données et disponibilité. D'autres cas d'utilisation peuvent nécessiter l'utilisation de key stores personnalisés avec AWS CloudHSM ou le key store externe.

2. Consultez la liste des services que vous utilisez pour votre charge de travail afin de comprendre comment AWS KMS s'y intègre. Par exemple, les instances EC2 peuvent utiliser des volumes EBS chiffrés. Elles vérifient ainsi que les instantanés Amazon EBS créés à partir de ces volumes sont également chiffrés à l'aide d'une clé gérée par le client et limitent la divulgation accidentelle des données instantanées non chiffrées.
    - Comment les services AWS utilisent AWS KMS
    - Pour obtenir des informations détaillées sur les options de chiffrement proposées par un service AWS, consultez la rubrique Chiffrement au repos du guide de l'utilisateur ou du guide du développeur du service.
3. Mettez en œuvre AWS KMS : AWS KMS simplifie la création et la gestion des clés et le contrôle de l'utilisation du chiffrement dans un large éventail de services AWS et dans vos applications.
    - Premiers pas : AWS Key Management Service (AWS KMS)
    - Consultez les bonnes pratiques en matière de contrôle d'accès à vos clés AWS KMS.
4. Envisagez AWS Encryption SDK : utilisez le kit AWS Encryption SDK avec l'intégration AWS KMS lorsque votre application doit chiffrer des données côté client.
    - AWS Encryption SDK
5. Activez IAM Access Analyzer pour examiner et envoyer automatiquement des notifications si les politiques des clés AWS KMS sont trop génériques.
6. Activez Security Hub pour recevoir des notifications en cas de mauvaise configuration des politiques relatives aux clés, de clés dont la suppression est prévue ou de clés dont la rotation automatique est activée.
7. Déterminez le niveau de journalisation approprié pour vos clés AWS KMS. Étant donné que les appels à AWS KMS, y compris les événements en lecture seule, sont journalisés, les journaux CloudTrail associés à AWS KMS peuvent devenir volumineux.
    - Certaines organisations préfèrent séparer les activités de journalisation AWS KMS à un emplacement distinct. Pour en savoir plus, consultez la section Journalisation des appels d'API AWS KMS avec CloudTrail dans le guide du développeur AWS KMS.

Ressources

Documents connexes :

- AWS Key Management Service
- Services et outils cryptographiques AWS

- Protection des données Amazon S3 à l'aide du chiffrement
- Chiffrement d'enveloppe
- L'engagement de souveraineté numérique
- Démystifier les opérations de clés AWS KMS, apporter votre propre clé, key store personnalisé et portabilité du texte chiffré
- Informations cryptographiques AWS Key Management Service

Vidéos connexes :

- How Encryption Works in AWS
- Securing Your Block Storage on AWS
- AWS data protection: Using locks, keys, signatures, and certificates

Exemples connexes :

- Mettre en œuvre des mécanismes de contrôle d'accès avancés avec AWS KMS

SEC08-BP02 Appliquer le chiffrement au repos

Vous devez appliquer l'utilisation du chiffrement pour les données au repos. Le chiffrement permet de préserver la confidentialité des données sensibles en cas d'accès non autorisé ou de divulgation accidentelle.

Résultat souhaité : les données privées doivent être chiffrées par défaut au repos. Le chiffrement permet de préserver la confidentialité des données et offre une protection supplémentaire contre la divulgation ou l'exfiltration intentionnelle ou involontaire des données. Les données chiffrées ne peuvent pas être lues ni consultées si elles n'ont pas été déchiffrées au préalable. Toutes les données stockées non chiffrées doivent être inventoriées et contrôlées.

Anti-modèles courants :

- Ne pas utiliser les configurations chiffrées par défaut.
- Fournir un accès trop permissif aux clés de déchiffrement.
- Ne pas surveiller l'utilisation des clés de chiffrement et de déchiffrement.
- Stocker des données non chiffrées.

- Utiliser la même clé de chiffrement pour toutes les données, quels que soient l'utilisation, le type et la classification des données.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

Mappez les clés de chiffrement aux classifications de données dans vos charges de travail. Cette approche offre une protection contre les accès trop permissifs lors de l'utilisation d'une seule, ou d'un très petit nombre de clés de chiffrement pour vos données (consultez SEC07-BP01 Identifier les données au sein de votre charge de travail).

AWS Key Management Service (AWS KMS) s'intègre à de nombreux services AWS afin de faciliter le chiffrement des données au repos. Par exemple, dans Amazon Simple Storage Service (Amazon S3), vous pouvez définir un chiffrement par défaut sur un compartiment pour que tous les nouveaux objets soient chiffrés automatiquement. Lorsque vous utilisez AWS KMS, tenez compte du degré de restriction des données. Les clés AWS KMS par défaut et contrôlées par le service sont gérées et utilisées en votre nom par AWS. Pour les données sensibles qui nécessitent un accès précis à la clé de chiffrement sous-jacente, envisagez les clés gérées par le client (CMK). Vous disposez d'un contrôle total sur les CMK, y compris la rotation et la gestion des accès grâce à l'utilisation de politiques de clés.

De plus, Amazon Elastic Compute Cloud (Amazon EC2) et Amazon S3 prennent en charge la mise en application du chiffrement en définissant le chiffrement par défaut. Vous pouvez utiliser AWS Config Rules pour vérifier automatiquement que vous utilisez le chiffrement, par exemple, pour les volumes Amazon Elastic Block Store (Amazon EBS), les instances Amazon Relational Database Service (Amazon RDS) et les compartiments Amazon S3.

AWS fournit également des options de chiffrement côté client, ce qui vous permet de chiffrer les données avant de les télécharger dans le cloud. Le AWS Encryption SDK fournit une solution pour chiffrer vos données à l'aide du chiffrement d'enveloppe. Vous fournissez la clé de wrapping et le AWS Encryption SDK génère une clé de données unique pour chaque objet de données qu'il chiffre. Envisagez AWS CloudHSM si vous avez besoin d'un module de sécurité du matériel géré à un seul locataire (HSM). AWS CloudHSM vous permet de générer, d'importer et de gérer des clés de chiffrement sur un HSM validé FIPS 140-2 de niveau 3. Certains cas d'utilisation pour AWS CloudHSM incluent la protection des clés privées pour l'émission d'une autorité de certification (CA) et le chiffrement transparent des données (TDE) pour les bases de données Oracle. Le kit SDK client AWS CloudHSM fournit un logiciel qui vous permet de chiffrer des données côté client à l'aide

de clés stockées dans AWS CloudHSM avant de télécharger vos données dans AWS. Amazon DynamoDB Encryption Client vous permet également de chiffrer et de signer les éléments avant de les télécharger dans une table DynamoDB.

Étapes d'implémentation

- Imposez le chiffrement au report pour Amazon S3 : implémentez le chiffrement par défaut des compartiments Amazon S3.

  Configurez le chiffrement par défaut pour les nouveaux volumes Amazon EBS : indiquez que vous souhaitez que tous les nouveaux volumes Amazon EBS soient chiffrés, avec la possibilité d'utiliser la clé par défaut fournie par AWS ou une clé que vous créez.

  Configurez des Amazon Machine Images (AMI) chiffrées : la copie d'une AMI existante avec le chiffrement activé chiffrera automatiquement les volumes racine et les instantanés.

  Configurez le chiffrement Amazon RDS : configurez le chiffrement de vos clusters de bases de données Amazon RDS et de vos instantanés au repos en utilisant l'option de chiffrement.

  Créez et configurez des clés AWS KMS avec des politiques qui limitent l'accès aux principaux appropriés pour chaque classification des données : par exemple, créez une clé AWS KMS pour chiffrer les données de production et une clé différente pour chiffrer les données de développement ou de test. Vous pouvez également fournir un accès de clé à d'autres Comptes AWS. Envisagez d'avoir différents comptes pour vos environnements de développement et de production. Si votre environnement de production a besoin de déchiffrer des artefacts dans le compte de développement, vous pouvez modifier la politique de CMK utilisée pour chiffrer les artefacts de développement afin de permettre au compte de production de déchiffrer ces artefacts. L'environnement de production peut ensuite ingérer les données déchiffrées afin de les utiliser en production.

  Configurez le chiffrement dans des services AWS supplémentaires : pour les autres services AWS que vous utilisez, consultez la documentation sur la sécurité associée aux services concernés afin de déterminer vos options de chiffrement.

Ressources

Documents connexes :

- Documentation sur AWS Crypto Tools

- Documentation sur AWS
- AWS Encryption SDK
- Livre blanc Présentation des détails cryptographiques de AWS KMS
- AWS Key Management Service
- AWS cryptographic services and tools (Services et outils de chiffrement AWS)
- Chiffrement Amazon EBS
- Default encryption for Amazon EBS volumes
- Chiffrement des ressources Amazon RDS
- How do I enable default encryption for an Amazon S3 bucket?
- Protection des données Amazon S3 à l'aide du chiffrement

Vidéos connexes :

- How Encryption Works in AWS
- Securing Your Block Storage on AWS

SEC08-BP03 Automatiser la protection des données au repos

utilisez des outils automatisés pour valider et faire respecter en permanence les contrôles des données au repos, par exemple en vérifiant qu'il n'y a que des ressources de stockage chiffrées. Vous pouvez automatiser la validation du chiffrement de tous les volumes de données EBS en utilisant AWS Config Rules. AWS Security Hub peut également vérifier plusieurs contrôles différents via des vérifications automatisées par rapport aux normes de sécurité. De plus, AWS Config Rules peut corriger les ressources non conformes automatiquement.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Données au repos représentent toutes les données que vous conservez dans un stockage non volatil pendant toute la durée de votre charge de travail. Cela comprend le stockage par bloc, le stockage d'objets, les bases de données, les archives, les appareils IoT et tout autre support de stockage sur lequel les données sont conservées. La protection de vos données inactives permet de réduire le risque d'accès non autorisé, lorsque le chiffrement et les contrôles d'accès appropriés sont mis en place.

Appliquer le chiffrement au repos : vous devez faire en sorte que le seul moyen de stocker des données est de les chiffrer. AWS KMS s'intègre en toute transparence à de nombreux services AWS pour vous permettre de chiffrer plus facilement toutes vos données au repos. Par exemple, dans Amazon Simple Storage Service (Amazon S3), vous pouvez définir un chiffrement par défaut sur un compartiment pour que tous les nouveaux objets soient chiffrés automatiquement. En outre, Amazon EC2 et Amazon S3 prennent en charge la mise en application du chiffrement en définissant le chiffrement par défaut. Vous pouvez utiliser des règles de configuration gérées AWS pour vérifier automatiquement que vous utilisez le chiffrement, par exemple, pour les volumes EBS, les instances Amazon Relational Database Service (Amazon RDS)et des compartiments Amazon S3.

Ressources

Documents connexes :

- Outils de chiffrement AWS
- Kit SDK AWS

Vidéos connexes :

- How Encryption Works in AWS
- Securing Your Block Storage on AWS

SEC08-BP04 Appliquer le contrôle d'accès

Pour vous aider à protéger vos données au repos, appliquez le contrôle d'accès à l'aide de mécanismes tels que l'isolement et la gestion des versions, et appliquez le principe du moindre privilège. Empêchez l'octroi d'un accès public à vos données.

Résultat souhaité : vérifiez que seuls les utilisateurs autorisés peuvent accéder aux données en fonction de la nécessité de les connaître. Protégez vos données avec des sauvegardes et des versions régulières pour éviter toute modification ou suppression intentionnelle ou involontaire des données. Isolez les données critiques des autres données afin de protéger leur confidentialité et leur intégrité.

Anti-modèles courants :

- Stocker ensemble des données ayant différentes exigences en termes de sensibilité ou de classification.

- Utiliser des autorisations trop permissives sur les clés de déchiffrement.
- Classer les données de façon incorrecte.
- Ne pas conserver les sauvegardes détaillées des données importantes.
- Fournir un accès permanent aux données de production.
- Ne pas auditer l'accès aux données ni examiner régulièrement les autorisations

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : faible

Directives d'implémentation

L'utilisation de plusieurs contrôles permet de protéger vos données au repos, y compris l'accès (en utilisant le moindre privilèges), l'isolement et la gestion des versions. L'accès à vos données doit être vérifié à l'aide des mécanismes de détection, notamment AWS CloudTrail, et le journal des niveaux de service, comme les journaux d'accès Amazon Simple Storage Service (Amazon S3). Vous devez faire l'inventaire des données accessibles au public et créer un plan permettant de réduire la quantité de données disponibles publiquement au fil du temps.

Amazon S3 Glacier Vault Lock et Amazon S3 Object Lock sont des fonctionnalités qui fournissent un contrôle d'accès obligatoire pour les objets dans Amazon S3. Lorsqu'une politique de coffre est verrouillée avec l'option de conformité, même l'utilisateur root ne peut pas la modifier avant l'expiration du verrouillage.

Étapes d'implémentation

- Appliquez le contrôle d'accès : appliquez le contrôle d'accès avec le principe du moindre privilège, y compris l'accès aux clés de chiffrement.
- Séparez les données selon différents niveaux de classification : utilisez différents Comptes AWS pour les niveaux de classification des données et gérez ces comptes en utilisant AWS Organizations.
- Vérifiez les politiques AWS Key Management Service (AWS KMS) : vérifiez le niveau d'accès octroyé dans les politiques AWS KMS.
- Examinez les autorisations de compartiment et d'objet Amazon S3 : examinez régulièrement le niveau d'accès octroyé dans les politiques de compartiment S3. Une bonne pratique consiste à éviter d'utiliser des compartiments publiquement accessibles en lecture ou en écriture. Envisagez d'utiliser AWS Config pour détecter les compartiments publiquement disponibles et Amazon CloudFront pour diffuser du contenu depuis Amazon S3. Vérifiez que les compartiments qui ne

doivent pas permettre l'accès public sont configurés correctement pour empêcher l'accès public. Par défaut, tous les compartiments S3 sont privés et ne sont accessibles qu'aux utilisateurs auxquels l'accès a été explicitement accordé.

- Activez l'Analyseur d'accèsAWS IAM : l'Analyseur d'accès IAM analyse les compartiments Amazon S3 et génère une découverte quand une politique S3 octroie un accès à une entité externe.
- Activez la gestion des versions Amazon S3 et le le verrouillage des objets lorsque c'est nécessaire.
- Utilisez l'inventaire Amazon S3 : l'inventaire Amazon S3 peut être utilisé pour auditer et rendre compte du statut de réplication et de chiffrement de vos objets S3.
- Passez en revue les autorisations de partage Amazon EBS et AMI : le partage des autorisations peut permettre le partage des images et des volumes avec des Comptes AWS externes à votre charge de travail.
- Vérifiez régulièrement les partages du Gestionnaire des accès aux ressources AWS afin de déterminer si des ressources doivent encore être partagées. Le Gestionnaire des accès aux ressources vous permet de partager des ressources, telles que les politiques AWS Network Firewall, les règles du résolveur Amazon Route 53 et les sous-réseaux avec vos Amazon VPC. Auditez régulièrement les ressources partagées et cessez de partager les ressources qui n'ont plus besoin de l'être.

Ressources

Bonnes pratiques associées :

- SEC03-BP01 Définir les conditions d'accès
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège

Documents connexes :

- Livre blanc Présentation des détails cryptographiques de AWS KMS
- Introduction to Managing Access Permissions to Your Amazon S3 Resources (Introduction à la gestion des autorisations d'accès à vos ressources Amazon S3)
- Overview of managing access to your AWS KMS resources (Aperçu de la gestion de l'accès à vos ressources KMS AWS)
- AWS Config Rules
- Amazon S3 + Amazon CloudFront: A Match Made in the Cloud
- Utilisation de la gestion des versions

- Utilisation du verrouillage des objets Amazon S3
- Partager un instantané Amazon EBS
- AMI partagées
- Hosting a single-page application on Amazon S3 (Hébergement d'une application à une page sur Amazon S3)

Vidéos connexes :

- Securing Your Block Storage on AWS

SEC08-BP05 Utiliser des mécanismes pour protéger l'accès aux données

Empêchez tous les utilisateurs d'accéder directement aux données et systèmes sensibles dans des circonstances opérationnelles normales. Par exemple, utilisez un flux de travail de gestion des changements pour gérer les instances Amazon Elastic Compute Cloud (Amazon EC2) avec des outils au lieu d'autoriser un accès direct ou un hôte bastion. Pour ce faire, recourez à AWS Systems Manager Automation, qui utilise des documents d'automatisation contenant les étapes nécessaires pour effectuer les tâches. Ces documents peuvent être stockés dans un système de contrôle de source, être examinés par des pairs avant l'exécution et être testés minutieusement pour minimiser les risques par rapport à un accès shell. Les utilisateurs de l'entreprise peuvent disposer d'un tableau de bord au lieu d'un accès direct à un magasin de données afin d'effectuer des requêtes. Lorsque des pipelines CI/CD ne sont pas utilisés, identifiez les contrôles et processus nécessaires pour fournir de manière adéquate un mécanisme alternatif normalement désactivé.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Implémenter des mécanismes pour protéger l'accès aux données : ces mécanismes incluent l'utilisation de tableaux de bord comme Amazon QuickSight pour présenter les données aux utilisateurs au lieu d'envoyer des requêtes directement.
  - Amazon QuickSight
- Automatisez la gestion de la configuration : effectuez des actions à distance, appliquez et validez automatiquement des configurations sécurisées en utilisant un service ou un outil de gestion de configuration. Évitez d'utiliser des hôtes bastion ou d'accéder directement aux instances EC2.
  - AWS Systems Manager

- AWS CloudFormation
- Pipeline CI/CD pour les modèles AWS CloudFormation sur AWS

Ressources

Documents connexes :

- Livre blanc sur les informations cryptographiques AWS KMS

Vidéos connexes :

- How Encryption Works in AWS
- Securing Your Block Storage on AWS

SÉC 9. Comment protéger vos données en transit ?

Protégez vos données en transit en mettant en œuvre plusieurs contrôles afin de réduire le risque d'accès non autorisé ou de perte.

Bonnes pratiques

- SEC09-BP01 Implémenter la gestion sécurisée des clés et des certificats
- SEC09-BP02 Appliquer le chiffrement en transit
- SEC09-BP03 Automatiser la détection des accès involontaires aux données
- SEC09-BP04 Authentifier les communications réseau

SEC09-BP01 Implémenter la gestion sécurisée des clés et des certificats

Les certificats du protocole TLS (Transport Layer Security) permettent de sécuriser les communications réseau et établir l'identité des sites web, des ressources et des charges de travail sur Internet, ainsi que sur les réseaux privés.

Résultat souhaité : Un système de gestion des certificats sécurisé qui peut provisionner, déployer, stocker et renouveler des certificats dans une infrastructure à clé publique (PKI). Un mécanisme sécurisé de gestion des clés et des certificats empêche la divulgation de la clé privée du certificat et renouvelle automatiquement et périodiquement le certificat. Il s'intègre également à d'autres services pour fournir des communications réseau et une identité sécurisées pour les ressources de la

machine au sein de votre charge de travail. Les clés ne doivent jamais être accessibles aux identités humaines.

Anti-modèles courants :

- Exécuter des étapes manuelles au cours des processus de déploiement ou de renouvellement des certificats.
- Ne pas accorder suffisamment d'attention à la hiérarchie de l'autorité de certification (AC) lors de la conception d'une AC privée.
- Utiliser des certificats auto-signés pour les ressources publiques.

Avantages liés au respect de cette bonne pratique :

- Simplifiez la gestion des certificats en automatisant leur déploiement et leur renouvellement
- Encouragez le chiffrage des données en transit à l'aide de certificats TLS
- Amélioration de la sécurité et de l'auditabilité des actions de certification entreprises par l'autorité de certification
- Organisation des tâches de gestion à différents niveaux de la hiérarchie de l'AC

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les charges de travail modernes font un usage intensif des communications réseau chiffrées à l'aide de protocoles PKI tels que le protocole TLS. La gestion des certificats PKI peut être complexe, mais le provisionnement, le déploiement et le renouvellement automatisés des certificats peuvent réduire les inconvénients liés à la gestion des certificats.

AWS fournit deux services pour gérer les certificats PKI à usage général : AWS Certificate Manager et AWS Private Certificate Authority (AWS Private CA). ACM est le principal service que les clients utilisent pour provisionner, gérer et déployer des certificats destinés à être utilisés dans des charges de travail AWS publiques et privées. ACM émet des certificats en utilisant AWS Private CA et intègre avec de nombreux autres services gérés AWS pour fournir des certificats TLS sécurisés pour les charges de travail.

AWS Private CA vous permet d'établir votre propre autorité de certification racine ou subordonnée et d'émettre des certificats TLS par l'intermédiaire d'une API. Vous pouvez utiliser ce type de certificats dans des scénarios où vous contrôlez et gérez la chaîne de confiance du côté client de la connexion

TLS. En plus des cas d'utilisation TLS, AWS Private CA peut émettre des certificats à des pods Kubernetes, des attestations produits pour appareils Matter, une signature de code et d'autres cas d'utilisation avec un modèle personnalisé. » Vous pouvez également utiliser IAM Roles Anywhere pour fournir des informations d'identification IAM temporaires aux charges de travail sur site qui ont reçu des certificats X.509 signés par votre autorité de certification privée.

En plus de ACM et AWS Private CA, AWS IoT Core fournit un support spécialisé pour provisionner, gérer et déployer des certificats PKI sur des appareils de l'Internet des objets. AWS IoT Core fournit des mécanismes spécialisés pour l'intégration des appareils IoT dans votre infrastructure de clés publiques à grande échelle.

Considérations relatives à l'établissement d'une hiérarchie d'autorités de certification privées

Lorsque vous devez établir une autorité de certification privée, il est important de prendre soin de concevoir correctement la hiérarchie de l'autorité de certification dès le départ. La bonne pratique consiste à déployer chaque niveau de votre hiérarchie d'autorité de certification dans des Comptes AWS distincts lorsque vous créez une hiérarchie d'autorité de certification privée. Cette étape intentionnelle réduit la surface de chaque niveau de la hiérarchie de l'autorité de certification, ce qui facilite la découverte d'anomalies dans les données de journalisation CloudTrail et réduit l'étendue de l'accès ou l'impact en cas d'accès non autorisé à l'un des comptes. L'autorité de certification racine doit résider dans son propre compte et ne doit être utilisée que pour émettre un ou plusieurs certificats d'autorité de certification intermédiaire.

Créez ensuite une ou plusieurs autorités de certification intermédiaires dans des comptes distincts du compte de l'autorité de certification racine afin d'émettre des certificats pour les utilisateurs finaux, les appareils ou d'autres charges de travail. Enfin, émettez des certificats à partir de votre autorité de certification racine vers les autorités de certification intermédiaires, qui émettront à leur tour des certificats vers vos utilisateurs finaux ou vos appareils. Pour plus d'informations sur la planification du déploiement des AC et la conception de la hiérarchie des AC, y compris la planification de la résilience, la réplication interrégionale, le partage des AC au sein de votre organisation et plus encore, voir Planification de votre déploiement AWS Private CA. »

Étapes d'implémentation

1. Déterminez les services AWS pertinents requis pour votre cas d'utilisation :
    - De nombreux cas d'utilisation peuvent s'appuyer sur l'infrastructure de clés publiques existante d'AWS à l'aide de AWS Certificate Manager. ACM peut déployer des certificats TLS pour les serveurs web, les équilibreurs de charge ou d'autres utilisations pour des certificats publiquement approuvés.

- Envisagez AWS Private CA si vous devez établir votre propre hiérarchie d'autorité de certification privée ou si vous avez besoin d'accéder à des certificats exportables. ACM peut ensuite émettre de nombreux types de certificats d'entités finales à l'aide de AWS Private CA.
- Pour les cas d'utilisation où les certificats doivent être provisionnés à grande échelle pour les appareils de l'Internet des objets (IoT) embarqués, envisagez AWS IoT Core. »

2. Mettez en œuvre le renouvellement automatisé des certificats dans la mesure du possible :
   - Utilisez le renouvellement géré de ACM pour les certificats émis par ACM, ainsi que les services intégrés gérés par AWS.
3. Établissez des journaux et des pistes d'audit :
   - Activez les journaux CloudTrail pour suivre l'accès aux comptes détenant des autorités de certification. Envisagez de configurer la validation de l'intégrité des fichiers journaux dans CloudTrail pour vérifier l'authenticité des données du journal.
   - Générez et révisez périodiquement des rapports d'audit répertoriant les certificats émis ou révoqués par votre autorité de certification privée. Ces rapports peuvent être exportés vers un compartiment S3.
   - Lors du déploiement d'une autorité de certification privée, vous devrez également créer un compartiment S3 pour stocker la liste de révocation des certificats (CRL). Pour obtenir des conseils sur la configuration de ce compartiment S3 en fonction des exigences de votre charge de travail, voir Planification d'une liste de révocation de certifications (CRL). »

Ressources

Bonnes pratiques associées :

- SEC02-BP02 Utiliser des informations d'identification temporaires
- SEC08-BP01 Mise en œuvre de la gestion sécurisée des clés
- SEC09-BP04 Authentifier les communications réseau

Documents connexes :

- Comment héberger et gérer une infrastructure complète de certificats privés dans AWS
- Comment garantir une hiérarchie d'autorités de certification privées ACM à l'échelle de l'entreprise pour l'automobile et la fabrication
- Private CA best practices

- Comment utiliser AWS RAM pour partager votre compte croisé Private CA ACM

Vidéos connexes :

- Activer Private CA AWS Certificate Manager (atelier)

Exemples connexes :

- Atelier sur Private CA
- Atelier sur IOT Device Management (y compris l'allocation des appareils)

Outils associés :

- Plugin pour Kubernetes cert-manager à utiliser AWS Private CA

SEC09-BP02 Appliquer le chiffrement en transit

Appliquez vos exigences de chiffrement définies en fonction des politiques, des obligations réglementaires et des normes de votre entreprise afin de répondre aux exigences organisationnelles, juridiques et de conformité. Utilisez uniquement les protocoles avec chiffrement lors de la transmission de données sensibles en dehors de votre cloud privé virtuel (VPC). Le chiffrement permet de préserver la confidentialité des données, même lorsque celles-ci transitent par des réseaux non fiables.

Résultat souhaité : toutes les données doivent être chiffrées en transit à l'aide de protocoles TLS sécurisés et de suites de chiffrement. Le trafic réseau entre vos ressources et Internet doit être chiffré pour limiter l'accès non autorisé aux données. Le trafic réseau uniquement au sein de votre environnement AWS doit être chiffré à l'aide de TLS dès que possible. Le réseau interne AWS est chiffré par défaut et le trafic réseau au sein d'un VPC ne peut pas être falsifié ni reniflé à moins qu'une partie non autorisée n'ait accès à quelque ressource que ce soit qui génère du trafic (comme les instances Amazon EC2 et les conteneurs Amazon ECS). Envisagez de protéger le trafic réseau à réseau avec un réseau privé virtuel (VPN) IPsec.

Anti-modèles courants :

- Utiliser des versions obsolètes de composants SSL, TLS et de suite de chiffrement (par exemple, SSL v3.0, clés RSA 1024 bits et chiffrement RC4).

- Autoriser le trafic non chiffré (HTTP) vers ou depuis des ressources publiques.
- Ne pas surveiller et ne pas remplacer les certificats X.509 avant leur expiration.
- Utiliser des certificats X.509 auto-signés pour TLS.

Niveau de risque exposé si cette bonne pratique n'est pas instaurée : élevé

Directives d'implémentation

Les services AWS fournissent des points de terminaison HTTPS utilisant TLS pour la communication, ce qui assure le chiffrement en transit lors de la communication avec les API AWS. Les protocoles non sécurisés comme HTTP peuvent être contrôlés et bloqués dans un VPC à l'aide de groupes de sécurité. Les requêtes HTTP peuvent également être redirigées automatiquement vers HTTPS dans Amazon CloudFront ou sur un Application Load Balancer. Vous disposez d'un contrôle total sur vos ressources de calcul pour mettre en œuvre le chiffrement en transit dans l'ensemble de vos services. De plus, vous pouvez utiliser la connectivité VPN dans votre VPC à partir d'un réseau externe ou d'AWS Direct Connect pour faciliter le chiffrement du trafic. Vérifiez que vos clients effectuent des appels vers des API AWS en utilisant au moins TLS 1.2, car AWS cessera d'utiliser TLS 1.0 et 1.1 en juin 2023. Des solutions tierces sont disponibles sur AWS Marketplace si vous avez des exigences particulières.

Étapes d'implémentation

- Appliquez le chiffrement en transit : vos exigences en matière de chiffrement doivent être définies selon les dernières normes et bonnes pratiques en matière de sécurité, et doivent autoriser uniquement des protocoles sécurisés. Par exemple, configurez un groupe de sécurité afin d'autoriser uniquement le protocole HTTPS pour un Application Load Balancer ou une instance Amazon EC2.
- Configurez des protocoles sécurisés dans les services périphériques : configurez HTTPS avec Amazon CloudFront et utilisez un profil de sécurité approprié pour votre situation de sécurité et votre cas d'utilisation.
- Utilisez un VPN pour la connectivité externe : envisagez d'utiliser un VPN IPsec pour sécuriser les connexions point à point ou réseau à réseau afin d'assurer à la fois la confidentialité et l'intégrité des données.
- Configurez les protocoles de sécurité dans les équilibreurs de charge : sélectionnez une politique de sécurité qui fournit les suites de chiffrement les plus solides prises en charge par les clients qui se connecteront à l'écouteur. Créez un écouteur HTTPS pour votre Application Load Balancer.

- Configurez des protocoles sécurisés dans Amazon Redshift : configurez votre cluster de façon à exiger une connexion SSL ou TLS.
- Configurez des protocoles de sécurité : consultez la documentation des services AWS afin de déterminer les capacités de chiffrement en transit.
- Configurez un accès sécurisé lors du téléchargement vers les compartiments Amazon S3 : utilisez les contrôles des politiques de compartiments Amazon S3 pour appliquer un accès sécurisé aux données.
- Envisagez d'utiliser AWS Certificate Manager : ACM vous permet de mettre en service, de gérer et de déployer des certificats TLS publics en vue de leur utilisation avec des services AWS.
- Envisagez d'utiliser AWS Private Certificate Authority pour les besoins PKI privés : AWS Private CA vous permet de créer des hiérarchies d'autorités de certification privées afin d'émettre des certificats X.509 d'entité finale qui peuvent être utilisés afin de créer des canaux TLS chiffrés.

Ressources

Documents connexes :

- Documentation sur AWS
- Utilisation du protocole HTTP avec CloudFront
- Connexion de votre VPC à des réseaux distants utilisant AWS Virtual Private Network
- Create an HTTPS listener for your Application Load Balancer (Créer un écouteur HTTPS pour votre Application Load Balancer)
- Didacticiel : Configurer SSL/TLS sur Amazon Linux 2
- Using SSL/TLS to encrypt a connection to a DB instance (Utilisation de SSL/TLS pour chiffrer une connexion à une instance de base de données)
- Configuration des options de sécurité des connexions

SEC09-BP03 Automatiser la détection des accès involontaires aux données

Utilisez des outils comme Amazon GuardDuty pour détecter automatiquement les activités suspectes ou les tentatives de déplacement de données en dehors des limites définies. Par exemple, GuardDuty peut détecter une activité de lecture Amazon Simple Storage Service (Amazon S3) inhabituelle avec le résultat Exfiltration:S3/AnomalousBehavior. Outre GuardDuty, les journaux de flux Amazon VPC, qui capture des informations sur le trafic réseau, peuvent être utilisés avec

Amazon EventBridge pour déclencher la détection des connexions anormales, qu'elles aboutissent ou non. Amazon S3 Access Analyzer peut vous aider à déterminer les données accessibles aux utilisateurs de vos compartiments Amazon S3.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Automatiser la détection de l'accès involontaire aux données : utilisez un outil ou un mécanisme de détection pour identifier automatiquement les tentatives de déplacement des données en dehors des limites définies, par exemple, pour détecter un système de base de données qui copie des données vers un hôte non reconnu.
  - Journaux de flux VPC
- Envisager Amazon Macie : Amazon Macie est un service de sécurité et de confidentialité des données entièrement géré qui utilise le machine learning et la comparaison de modèles pour découvrir et protéger vos données sensibles dans AWS.
  - Amazon Macie

Ressources

Documents connexes :

- Journaux de flux VPC
- Amazon Macie

SEC09-BP04 Authentifier les communications réseau

Vérifiez l'identité des communications à l'aide de protocoles comme TLS (Transport Layer Security) ou IPsec qui prennent en charge l'authentification.

Concevez votre charge de travail de manière à utiliser des protocoles réseau sécurisés et authentifiés lors de la communication entre les services, les applications ou avec les utilisateurs. L'utilisation de protocoles réseau qui prennent en charge l'authentification et l'autorisation permet de mieux contrôler les flux du réseau et de réduire l'impact des accès non autorisés.

Résultat souhaité : une charge de travail avec des flux de trafic bien définis entre les services au niveau du plan de données et du plan de contrôle. Les flux de trafic utilisent des protocoles réseau authentifiés et chiffrés lorsque cela est techniquement possible.

Anti-modèles courants :

- Flux de trafic non chiffrés ou non authentifiés au sein de votre charge de travail.
- Réutilisation des informations d'authentification par plusieurs utilisateurs ou entités.
- S'appuyer uniquement sur les contrôles réseau pour contrôler les accès.
- Créer un mécanisme d'authentification personnalisé au lieu d'utiliser des mécanismes d'authentification standard.
- Flux de trafic trop permissifs entre les composants des services ou d'autres ressources dans le VPC.

Avantages liés à l'instauration de cette bonne pratique :

- Limite l'impact des accès non autorisés à une partie de la charge de travail.
- Offre la garantie que les actions ne sont effectuées que par des entités authentifiées.
- Améliore le découplage des services en définissant clairement et en appliquant les interfaces de transfert de données prévues.
- Améliore la surveillance, la journalisation et la réponse aux incidents grâce à l'attribution des demandes et à des interfaces de communication bien définies.
- Assure une défense approfondie de vos charges de travail en combinant des contrôles réseau avec des contrôles d'authentification et d'autorisation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Les modèles de trafic réseau de votre charge de travail peuvent être classés en deux catégories :

- Le trafic est-ouest représente le trafic entre les services qui constituent une charge de travail.
- Le trafic nord-sud représente le trafic entre votre charge de travail et les consommateurs.

Le chiffrement du trafic nord-sud est courant, mais la sécurisation du trafic est-ouest à l'aide de protocoles authentifiés l'est moins. Les pratiques modernes de sécurité recommandent que la conception du réseau ne permette pas à elle seule d'établir une relation de confiance entre deux entités. Lorsque deux services peuvent résider dans les limites d'un réseau commun, il est toujours recommandé de chiffrer, d'authentifier et d'autoriser les communications entre ces services.

Par exemple, les API des services AWS utilisent le protocole de signature des demandes d'API AWS Version 4 (SigV4) pour authentifier l'appelant, quel que soit le réseau d'origine de la demande. Cette authentification garantit que les API AWS peuvent vérifier l'identité de la personne qui a demandé l'action, et cette identité peut ensuite être combinée avec des stratégies pour décider si l'action doit être autorisée ou non.

Des services comme Amazon VPC Lattice et Amazon API Gateway vous permettent d'utiliser le même protocole de signature SigV4 pour ajouter une authentification et une autorisation au trafic est-ouest dans vos propres charges de travail. Si des ressources extérieures à votre environnement AWS ont besoin de communiquer avec des services qui nécessitent une authentification et une autorisation basées sur SigV4, vous pouvez utiliser AWS Identity and Access Management (IAM) Roles Anywhere sur la ressource non AWS pour obtenir des informations d'identification AWS temporaires. Ces informations d'identification peuvent être utilisées pour signer les demandes de services utilisant SigV4 pour autoriser l'accès.

L'authentification mutuelle TLS (mTLS) est un autre mécanisme courant pour authentifier le trafic est-ouest. De nombreuses applications IoT (Internet des objets) et B2B, ainsi que des microservices utilisent mTLS pour valider l'identité des deux côtés d'une communication TLS à l'aide de certificats X.509 côté client et côté serveur. Ces certificats peuvent être émis par AWS Private Certificate Authority (AWS Private CA). Vous pouvez utiliser des services comme Amazon API Gateway et AWS App Mesh pour fournir une authentification mTLS pour les communications entre les charges de travail ou à l'intérieur de celles-ci. mTLS fournit des informations d'authentification pour les deux côtés d'une communication TLS, mais elle ne fournit pas de mécanisme d'autorisation.

Enfin, OAuth 2.0 et OpenID Connect (OIDC) sont deux protocoles généralement utilisés pour contrôler l'accès aux services par les utilisateurs, mais ils sont également de plus en plus populaires pour le trafic de service à service. API Gateway fournit un mécanisme d'autorisation JSON Web Token (JWT), permettant aux charges de travail de restreindre l'accès aux routes API à l'aide de JWT émis par les fournisseurs d'identité OIDC et OAuth 2.0. Les champs d'application OAuth2 peuvent être utilisés comme source pour les décisions d'autorisation de base, mais les contrôles d'autorisation doivent encore être mis en œuvre dans la couche applicative, et les champs d'application OAuth2 ne peuvent pas à eux seuls répondre à des besoins d'autorisation plus complexes.

Étapes d'implémentation

- Définir et documenter les flux de réseau de votre charge de travail : la première étape de la mise en œuvre d'une stratégie de défense en profondeur consiste à définir les flux de trafic de votre charge de travail.

  - Créez un diagramme de flux de données qui définit clairement la transmission des données entre les différents services qui constituent votre charge de travail. Ce schéma constitue la première étape de l'application de ces flux par le biais de réseaux authentifiés.
  - Instrumentez votre charge de travail lors des phases de développement et de test pour vérifier que le diagramme de flux de données reflète avec précision le comportement de la charge de travail lors de l'exécution.
  - Un diagramme de flux de données peut également être utile lors d'un exercice de modélisation des menaces, comme décrit dans SEC01-BP07 Identifier les menaces et hiérarchiser les atténuations à l'aide d'un modèle de menaces.
- Mettre en place des contrôles de réseau : tenez compte de capacités d'AWS pour mettre en place des contrôles réseau alignés sur vos flux de données. Les limites du réseau ne doivent pas représenter le seul contrôle de sécurité, mais elles constituent une couche de la stratégie de défense en profondeur visant à protéger votre charge de travail.
  - Utilisez des groupes de sécurité pour établir, définir et restreindre les flux de données entre les ressources.
  - Envisagez d'utiliser AWS PrivateLink pour communiquer avec les services d'assistance AWS et tiers qui prennent en charge AWS PrivateLink. Les données envoyées via un point de terminaison d'interface AWS PrivateLink restent dans le réseau AWS et ne transitent pas par l'Internet public.
- Mettre en œuvre un système d'authentification et d'autorisation pour tous les services de votre charge de travail : choisissez l'ensemble de services AWS le plus approprié pour authentifier et chiffrer les flux de trafic de votre charge de travail.
  - Envisagez d'utiliser Amazon VPC Lattice pour sécuriser les communications de service à service. VPC Lattice peut utiliser l'authentification SigV4 et des stratégies d'authentification pour contrôler les accès de service à service.
  - Pour la communication de service à service à l'aide de mTLS, envisagez d'utiliser API Gateway ou App Mesh. AWS Private CA peut être utilisé pour établir une hiérarchie des autorités de certification privées capables d'émettre des certificats à utiliser avec mTLS.
  - Pour l'intégration à des services utilisant OAuth 2.0 ou OIDC, envisagez d'utiliser API Gateway avec les mécanismes d'autorisation JWT.
  - Pour les communications entre votre charge de travail et des appareils IoT, AWS IoT Core propose plusieurs options de chiffrement et d'authentification du trafic réseau.

- Surveiller les accès non autorisés : surveillez en permanence les canaux de communication involontaires, les personnes non autorisées qui tentent d'accéder à des ressources protégées et autres schémas d'accès inappropriés.
  - Si vous utilisez VPC Lattice pour gérer l'accès à vos services, envisagez d'activer et de surveiller les journaux d'accès VPC Lattice. Ces journaux contiennent des informations sur le demandeur et le réseau, notamment le VPC source et de destination, et les métadonnées des demandes.
  - Envisagez d'activer les journaux de flux VPC pour capturer des métadonnées sur les flux du réseau et passer régulièrement en revue les anomalies.
  - Consultez le guide de réponse aux incidents de sécurité AWS et la section Réponse aux incidents du livre blanc Pilier de sécurité - AWS Well-Architected Framework pour plus de conseils sur la planification et la simulation des incidents de sécurité, ainsi que la réponse qui y est apportée.

Ressources

Bonnes pratiques associées :

- SEC03-BP07 Analyser l'accès public et entre les comptes
- SEC02-BP02 Utiliser des informations d'identification temporaires
- SEC01-BP07 Identifier les menaces et hiérarchiser les atténuations à l'aide d'un modèle de menaces

Documents connexes :

- Evaluating access control methods to secure Amazon API Gateway APIs
- Configuration de l'authentification TLS mutuelle pour une API REST
- How to secure API Gateway HTTP endpoints with JWT authorizer
- Authorizing direct calls to AWS services using AWS IoT Core credential provider
- AWS Security Incident Response Guide

Vidéos connexes :

- AWS re:invent 2022: Introducing VPC Lattice
- AWS re:invent 2020: Serverless API authentication for HTTP APIs on AWS

Exemples connexes :

- Amazon VPC Lattice Workshop
- Zero-Trust Episode 1 – The Phantom Service Perimeter workshop

Réponse aux incidents

Question

- SÉC 10. Comment anticipez-vous les incidents, y répondez et y parvenez-vous ?

SÉC 10. Comment anticipez-vous les incidents, y répondez et y parvenez-vous ?

Même avec des contrôles préventifs et de détection matures, votre organisation doit mettre en place des mécanismes pour répondre aux incidents de sécurité et en atténuer l'impact potentiel. Votre préparation affectera fortement la capacité de vos équipes à opérer efficacement lors d'un incident, à analyser, isoler et contenir les problèmes, et à rétablir les opérations à un état de fonctionnement correct. La mise en place des outils et des accès avant un incident de sécurité, puis la pratique régulière de la réponse aux incidents pendant des exercices de simulation, vous permettent de rétablir les opérations tout en minimisant les interruptions d'activité.

Bonnes pratiques

- SEC10-BP01 Identifier les postes clés internes ainsi que les principales ressources externes
- SEC10-BP02 Développer des plans de gestion des incidents
- SEC10-BP03 Préparer les fonctionnalités d'analyse poussée
- SEC10-BP04 Développer et tester des playbooks de réponse aux incidents de sécurité
- SEC10-BP05 Préallouer les accès
- SEC10-BP06 Prédéployer les outils
- SEC10-BP07 Exécuter des simulations
- SEC10-BP08 Établir un cadre pour tirer des enseignements des incidents

SEC10-BP01 Identifier les postes clés internes ainsi que les principales ressources externes

Identifiez les postes clés internes et externes, les ressources et les obligations légales qui aideront votre organisation à réagir en cas d'incident.

Lorsque vous définissez votre approche de la réponse aux incidents dans le cloud, à l'unisson avec d'autres équipes (telles que votre conseiller juridique, vos dirigeants, les parties prenantes de l'entreprise, les services AWS Support, etc.), vous devez identifier le personnel clé, les parties prenantes et les contacts pertinents. Pour réduire la dépendance et le temps de réponse, veillez à ce que votre équipe, les équipes de sécurité spécialisées et les intervenants soient formés aux services que vous utilisez et aient la possibilité d'effectuer des exercices pratiques.

Nous vous encourageons à identifier des partenaires de sécurité AWS externes qui peuvent vous fournir une expertise extérieure et une perspective différente pour augmenter vos capacités d'intervention. Vos partenaires de sécurité de confiance peuvent vous aider à identifier des risques ou des menaces potentiels que vous ne connaissez peut-être pas.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

- Identifier les postes clés de votre organisation : Tenez à jour une liste des employés au sein de votre organisation que vous devez impliquer pour réagir et récupérer après un incident.
- Identifier les partenaires externes : Collaborez le cas échéant avec des partenaires externes qui pourront vous aider à réagir et à reprendre après un incident.

Ressources

Documents connexes :

- Guide de réponse aux incidents AWS

Vidéos connexes :

- Prepare for and respond to security incidents in your AWS environment

Exemples connexes :

SEC10-BP02 Développer des plans de gestion des incidents

Le premier document à élaborer pour la réponse aux incidents est le plan de réponse aux incidents. Le plan de réponse aux incidents est conçu pour servir de base à votre programme et à votre stratégie de réponse aux incidents.

Avantages liés au respect de cette bonne pratique : Le développement de processus de réponse aux incidents complets et clairement définis est essentiel à la réussite d'un programme de réponse aux incidents évolutif. Lorsqu'un incident de sécurité se produit, des étapes et des flux de travail clairs peuvent vous aider à réagir rapidement. Vous disposez peut-être déjà de processus de réponse aux incidents. Quel que soit votre état actuel, il est important de mettre à jour, d'itérer et de tester régulièrement vos processus de réponse aux incidents.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Un plan de gestion des incidents est essentiel pour réagir, atténuer et se remettre des répercussions potentielles des incidents de sécurité. Un plan de gestion des incidents est un processus structuré qui permet d'identifier les incidents de sécurité, d'y remédier et d'y répondre rapidement.

Le cloud comporte un grand nombre de rôles et exigences opérationnels identiques à ceux d'un environnement sur site. Lorsque vous créez un plan de gestion des incidents, il est important de tenir compte des stratégies d'intervention et de récupération qui correspondent le mieux aux résultats opérationnels et aux exigences de conformité. Par exemple, si vous exécutez des charges de travail dans AWS qui sont conformes à FedRAMP aux États-Unis, il est utile de respecter NIST SP 800-61 Computer Security Handling Guide. De la même manière, lorsque vous exécutez des charges de travail avec des données européennes personnellement identifiables, envisagez des scénarios tels que la façon dont vous pourriez protéger les données et résoudre des problèmes liés à la résidence des données, comme l'exige le Règlement Général sur la Protection des Données (RGPD). »

Lorsque vous élaborez un plan de gestion des incidents pour vos charges de travail dans AWS, commencez par le Modèle de responsabilité partagée AWS, afin de créer une approche de défense en profondeur en matière de réponse aux incidents. Dans le cadre de ce modèle, AWS gère la sécurité du cloud et vous êtes responsable de la sécurité dans le cloud. Cela signifie que vous conservez le contrôle et que vous êtes responsable des contrôles de sécurité que vous choisissez d'implémenter. L' AWS Security Incident Response Guide détaille les concepts clés et les conseils de base pour l'élaboration d'un plan de gestion des incidents axé sur le cloud.

Un plan de gestion des incidents efficace doit être répété constamment, tout en poursuivant votre objectif d'opérations dans le cloud. Envisagez d'utiliser les plans d'implémentation décrits ci-dessous pour créer et faire évoluer votre plan de gestion des incidents.

Étapes d'implémentation

Définissez les rôles et les responsabilités

La gestion des événements de sécurité exige une discipline interorganisationnelle et une volonté d'action. Au sein de votre structure organisationnelle, de nombreuses personnes doivent être responsables, tenues de rendre des comptes, consultées ou tenues informées lors d'un incident. Il peut notamment s'agir de représentants des ressources humaines (RH), de l'équipe de direction et du service juridique. Tenez compte de ces rôles et responsabilités et déterminez si des tiers doivent être impliqués. Notez que de nombreuses zones géographiques ont des lois locales qui régissent ce qui doit et ne doit pas être fait. Bien qu'il puisse sembler bureaucratique de créer un tableau RACI (réalisateur, approbateur, consulté et informé) pour vos plans de réponse en matière de sécurité, cela facilite une communication rapide et directe et définit clairement le leadership à chaque étape de l'événement.

Lors d'un incident, il est essentiel d'inclure les propriétaires et les développeurs des applications et des ressources concernées, car ce sont des experts en la matière (SME) qui peuvent fournir des informations et un contexte afin d'aider à mesurer l'impact. Assurez-vous d'établir et de maintenir des relations avec les développeurs et les propriétaires d'applications avant de vous fier à leur expertise pour répondre aux incidents. Les propriétaires d'applications ou SME, tels que vos administrateurs ou ingénieurs cloud, peuvent avoir besoin d'agir dans des situations où l'environnement est inconnu ou complexe, ou lorsque les intervenants n'y ont pas accès.

Enfin, des partenaires de confiance peuvent être impliqués dans l'enquête ou la réponse car ils peuvent apporter une expertise supplémentaire et un examen minutieux. Si vous ne possédez pas ces compétences au sein de votre propre équipe, vous pouvez faire appel à un tiers pour obtenir de l'aide.

Comprenez les équipes d'intervention et le support AWS

- AWS Support
  - AWS Support propose une gamme de plans qui donnent accès à des outils et à une expertise qui contribuent à la réussite et à l'intégrité opérationnelle de vos solutions AWS. Si vous avez besoin d'un support technique et de ressources supplémentaires pour planifier, déployer et optimiser votre environnement AWS, vous pouvez sélectionner le plan de support le plus adapté à votre cas d'utilisation AWS.
  - Envisagez le Centre de support dans AWS Management Console (connexion requise) en tant que point de contact central pour obtenir de l'aide en cas de problèmes affectant vos ressources AWS. L'accès à AWS Support est contrôlé par AWS Identity and Access Management. Pour plus d'informations sur l'accès aux fonctionnalités AWS Support, consultez Mise en route avec AWS Support.
- Équipe de réponse aux incidents clients (CIRT) AWS

  - L'équipe de réponse aux incidents clients (CIRT) AWS est une équipe AWS internationale spécialisée et disponible 24 heures sur 24, 7 jours sur 7, qui fournit une assistance aux clients lors d'événements de sécurité actifs côté client du Modèle de responsabilité partagée AWS.
  - Lorsque la CIRT AWS vous accompagne, elle fournit une assistance en matière de triage et de récupération en cas d'événement de sécurité actif sur AWS. Elle peut vous aider à analyser les causes profondes à l'aide des journaux de service AWS et vous fournir des recommandations pour la récupération. Elle peut également fournir des recommandations de sécurité et des bonnes pratiques pour vous aider à éviter des incidents de sécurité à l'avenir.
  - Les clients AWS peuvent contacter la CIRT AWS par le biais d'un cas AWS Support.
- Support de réponse aux attaques DDoS
  - Offres AWS AWS Shield, qui fournit un service géré de protection contre le déni de service distribué (DDoS) protégeant les applications Web exécutées sur AWS. Shield fournit une détection permanente et des mesures d'atténuation automatiques en ligne capables de minimiser les temps d'arrêt et la latence des applications, de sorte qu'il n'est pas nécessaire d'engager AWS Support pour bénéficier de la protection DDoS. Il existe deux niveaux de Shield : AWS Shield Standard et AWS Shield Advanced. Pour en savoir plus sur les différences entre ces deux niveaux, consultez la documentation sur les fonctionnalités Shield.
- AWS Managed Services (AMS)
  - AWS Managed Services (AMS) fournit une gestion continue de votre infrastructure AWS afin que vous puissiez vous concentrer sur vos applications. En mettant en œuvre les meilleures pratiques pour gérer votre infrastructure, AMS permet de réduire vos frais généraux et vos risques opérationnels. AMS automatise les activités courantes telles que les demandes de modification, la surveillance, la gestion des correctifs, la sécurité et les services de sauvegarde, et fournit des services de cycle de vie complets pour provisionner, exécuter et prendre en charge votre infrastructure.
  - AMS prend la responsabilité de déployer une suite de contrôles de sécurité et fournit une réponse de première ligne 24 heures sur 24, 7 jours sur 7 aux alertes. Lorsqu'une alerte est déclenchée, AMS suit un ensemble standard de playbooks automatisés et manuels pour vérifier une réponse cohérente. Ces playbooks sont partagés avec les clients AMS lors de l'intégration afin qu'ils puissent développer et coordonner une réponse avec AMS.

Élaborez le plan de réponse aux incidents

Le plan de réponse aux incidents est conçu pour servir de base à votre programme et à votre stratégie de réponse aux incidents. Le plan de réponse aux incidents doit figurer dans un document formel. Un plan de réponse aux incidents comprend généralement les sections suivantes :

- Présentation de l'équipe de réponse aux incidents : décrit les objectifs et les fonctions de l'équipe de réponse aux incidents.
- Rôles et responsabilités : répertorie les parties prenantes de la réponse aux incidents et détaille leurs rôles en cas d'incident.
- Un plan de communication : détaille les coordonnées et la manière dont vous communiquez lors d'un incident.
- Méthodes de communication de secours : il est recommandé d'utiliser une communication hors bande comme solution de secours pour les communications en cas d'incident. Un exemple d'application qui fournit un canal de communication hors bande sécurisé est AWS Wickr.
- Phases de la réponse aux incidents et mesures à prendre : énumère les phases de la réponse aux incidents (par exemple, détection, analyse, éradication, maîtrise et récupération), y compris les mesures de haut niveau à prendre au cours de ces phases.
- Définitions de la gravité et de la hiérarchisation des incidents : décrit en détail comment classer la gravité d'un incident, comment hiérarchiser l'incident, puis comment les définitions de gravité affectent les procédures de remontée.

Bien que ces sections soient communes à des entreprises de tailles et de secteurs différents, le plan de réponse aux incidents de chaque organisation est unique. Vous devez élaborer un plan de réponse aux incidents qui convient le mieux à votre organisation.

Ressources

Bonnes pratiques connexes :

- SEC04 (Comment détecter et enquêter sur les événements de sécurité ?)

Documents connexes :

- AWS Security Incident Response Guide
- NIST: Computer Security Incident Handling Guide

SEC10-BP03 Préparer les fonctionnalités d'analyse poussée

Pour anticiper un incident de sécurité, envisagez de développer des fonctionnalités d'analyse poussée pour faciliter les enquêtes sur les événements de sécurité.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Les concepts issus de l'analyse poussée traditionnelle sur site s'appliquent à AWS. Pour obtenir des informations clés sur la manière de commencer à renforcer les capacités d'analyse poussée dans le AWS Cloud, consultez Stratégies d'environnement d'enquête pour les analyses poussées dans le AWS Cloud (langue française non garantie). »

Une fois que vous avez configuré la structure de votre environnement et de votre compte Compte AWS en vue de l'analyse poussée, définissez les technologies requises pour appliquer efficacement les méthodologies d'analyse poussée en quatre phases :

- Collecte : Collectez des journaux AWS pertinents, tels que AWS CloudTrail, AWS Config, les journaux de flux VPC et les journaux au niveau de l'hôte. Collectez des instantanés, des sauvegardes et des fichiers de vidage de mémoire des ressources AWS concernées, le cas échéant.
- Examen : Examinez les données collectées en extrayant et en évaluant les informations pertinentes.
- Analyse : Analysez les données collectées afin de comprendre l'incident et d'en tirer des conclusions.
- Reporting : Présentez les informations issues de la phase d'analyse.

Étapes d'implémentation

Préparation de votre environnement d'analyse poussée

AWS Organizations vous permet de gérer et de gouverner de manière centralisée un environnement AWS à mesure que vous vous développez et que vous mettez à l'échelle vos ressources AWS. Une organisation AWS consolide vos Comptes AWS pour que vous puissiez les administrer en tant qu'unité unique. Vous pouvez utiliser des unités d'organisation (UO) pour regrouper des comptes afin de les administrer en tant qu'unité unique.

Pour réagir face aux incidents, il est utile de disposer d'une structure de Compte AWS prenant en charge les fonctions de réponse aux incidents, qui comprend une unité d'organisation de sécurité

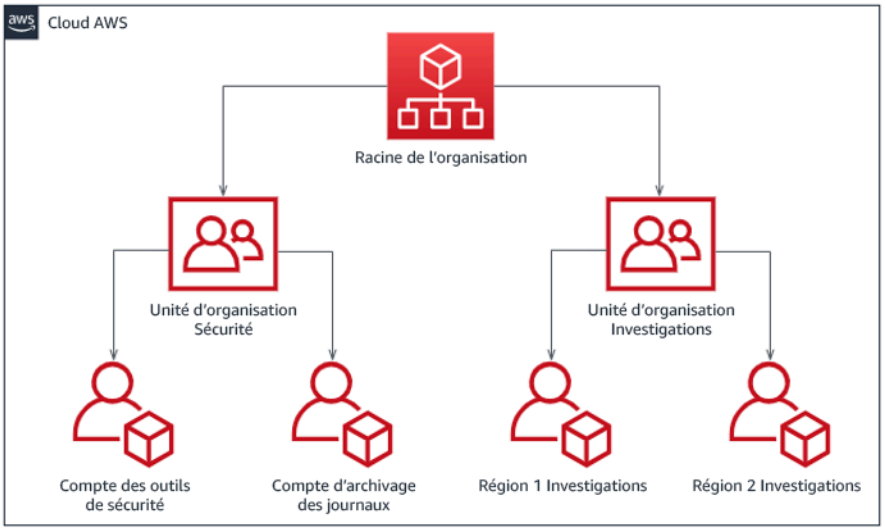
et une unité d'organisation d'analyse poussée.. » Au sein de l'unité d'organisation de sécurité, vous devez disposer de comptes pour :

- Archivage des journaux : Regroupez les journaux dans un Compte AWS d'archivage de journaux avec des autorisations limitées.
- Outils de sécurité : Centralisez les services de sécurité dans un Compte AWS d'outil de sécurité. Ce compte joue le rôle d'administrateur délégué pour les services de sécurité.

Au sein de l'unité d'organisation d'analyse poussée, vous avez la possibilité de mettre en place un ou plusieurs comptes d'analyse poussée pour chaque région dans laquelle vous opérez, selon ce qui convient le mieux à votre entreprise et à votre modèle opérationnel. Si vous créez un compte d'analyse poussée par région, vous pouvez bloquer la création des ressources AWS en dehors de cette région et réduire le risque que les ressources soient copiées vers une région non prévue. Par exemple, si vous opérez uniquement dans US East (N. Virginia) Region (`us-east-1`) et US West (Oregon) (`us-west-2`), vous auriez alors deux comptes dans l'unité d'organisation d'analyse poussée : une pour `us-east-1` et une pour `us-west-2`. »

Vous pouvez créer un Compte AWS d'analyse poussée pour plusieurs régions. Soyez prudent lorsque vous copiez des ressources AWS sur ce compte afin de vérifier que vous respectez vos exigences en matière de souveraineté des données. Étant donné que la mise en place de nouveaux comptes prend du temps, il est impératif de créer et d'instrumenter les comptes d'analyse poussée bien avant un incident afin que les intervenants puissent être prêts à les utiliser efficacement pour intervenir.

Le diagramme suivant présente un exemple de structure de compte, y compris une unité d'organisation d'analyse poussée avec des comptes d'analyse poussée par région :



Structure de compte par région pour la réponse aux incidents

Capture de sauvegardes et d'instantanés

La configuration de sauvegardes des systèmes et des bases de données clés s'avère essentielle pour récupérer d'un incident de sécurité et à des fins d'analyse poussée. Une fois les sauvegardes en place, vous pouvez restaurer vos systèmes à leur état stable antérieur. Sur AWS, vous pouvez prendre des instantanés de différentes ressources. Les instantanés fournissent des sauvegardes ponctuelles de ces ressources. De nombreux services AWS peuvent vous aider en matière de sauvegarde et de restauration. Pour plus de détails sur ces services et approches en matière de sauvegarde et de restauration, consultez Conseils normatifs sur la sauvegarde et la restauration et Utilisez des sauvegardes pour restaurer après des incidents de sécurité. »

Il est essentiel que vos sauvegardes soient bien protégées, en particulier dans le cas de rançongiciels. Pour obtenir des conseils sur la sécurisation de vos sauvegardes, consultez Les 10 meilleures pratiques de sécurité pour sécuriser les sauvegardes dans AWS. » Outre la sécurisation de vos sauvegardes, vous devez régulièrement tester vos processus de sauvegarde et de restauration pour vérifier que la technologie et les processus que vous avez mis en place fonctionnent comme prévu.

Automatisation de l'analyse poussée

Lors d'un événement de sécurité, votre équipe de réponse aux incidents doit être en mesure de collecter et d'analyser rapidement les preuves tout en préservant l'exactitude de la période pendant laquelle s'est produit l'événement (par exemple, en capturant les journaux relatifs à un événement ou à une ressource spécifique ou en collectant les fichiers de vidage de mémoire d'une instance

Amazon EC2). Il est à la fois difficile et fastidieux pour l'équipe de réponse aux incidents de collecter manuellement les preuves pertinentes, en particulier sur un grand nombre d'instances et de comptes. De plus, la collecte manuelle peut faire l'objet d'erreurs humaines. Pour ces raisons, vous devez développer et mettre en œuvre autant que possible l'automatisation de l'analyse poussée.

AWS propose un certain nombre de ressources d'automatisation pour l'analyse poussée, qui sont répertoriées dans la section Ressources suivante. Ces ressources sont des exemples de modèles d'analyse poussée que nous avons développés et que les clients ont mis en œuvre. Bien qu'elles puissent constituer une architecture de référence utile au départ, envisagez de les modifier ou de créer de nouveaux modèles d'automatisation de l'analyse poussée en fonction de votre environnement, de vos exigences, de vos outils et de vos processus d'analyse poussée.

Ressources

Documents connexes :

- Guide AWS de réponse aux incidents de sécurité - Développement des fonctionnalités d'analyse poussée (langue française non garantie)
- Guide AWS de réponse aux incidents de sécurité - Ressources d'analyse poussée (langue française non garantie)
- Stratégies d'environnement d'enquête pour les analyses poussées dans le AWS Cloud (langue française non garantie)
- Comment automatiser la collecte de disques d'analyse dans AWS
- Recommandations AWS - Automatiser la réponse aux incidents et l'analyse poussée (langue française non garantie)

Vidéos connexes :

- Automatisation de la réponse aux incidents et investigations

Exemples connexes :

- Cadre d'automatisation de la réponse aux incidents et de l'analyse poussée (langue française non garantie)
- Orchestrateur d'analyse poussée automatisée pour Amazon EC2 (langue française non garantie)

SEC10-BP04 Développer et tester des playbooks de réponse aux incidents de sécurité

L'élaboration de playbooks est une étape clé de la préparation de vos processus de réponse aux incidents. Les playbooks de réponse aux incidents fournissent une série de recommandations et d'étapes à suivre en cas d'événement de sécurité. Le fait de disposer d'une structure et d'étapes claires simplifie la réponse et réduit le risque d'erreur humaine.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Il est recommandé de créer des playbooks dans les scénarios d'incidents suivants :

- Incidents attendus: créez des playbooks pour les incidents que vous anticipez. Cela inclut des menaces telles que le déni de service (DoS), les rançongiciels et la mise en péril des informations d'identification.
- Résultats ou alertes de sécurité connus: créez des playbooks pour vos résultats et alertes de sécurité connus, tels que les résultats GuardDuty. Vous pourriez recevoir un résultat GuardDuty et vous demander ce que vous devez en faire. Pour éviter de mal gérer ou d'ignorer un résultat GuardDuty, créez un playbook pour chaque résultat GuardDuty potentiel. Certains détails et conseils de résolution sont disponibles dans la documentation GuardDuty. » Il convient de noter que GuardDuty n'est pas activé par défaut et que son activation n'entraîne aucun frais. Pour plus de détails sur GuardDuty, consultez Annexe A : Définition de la capacité du cloud - Visibilité et alertes (langue française non garantie). »

Les playbooks doivent contenir les étapes techniques qu'un analyste de sécurité doit suivre afin d'enquêter de manière adéquate et de répondre à un éventuel incident de sécurité.

Étapes d'implémentation

Les éléments à inclure dans un playbook incluent :

- Présentation du playbook: quel scénario de risque ou d'incident ce playbook aborde-t-il ? Quel est l'objectif du playbook ?
- Conditions préalables: quels journaux, mécanismes de détection et outils automatisés sont requis pour ce scénario d'incident ? Quelle est la notification attendue ?
- Informations sur la communication et les remontées: qui sont les personnes impliquées et quelles sont leurs coordonnées ? Quelles sont les responsabilités de chacune des parties prenantes ?

- Étapes de réponse: quelles sont les mesures tactiques à prendre au cours des différentes phases de la réponse à un incident ? Quelles requêtes un analyste doit-il exécuter ? Quel code doit être exécuté pour obtenir le résultat souhaité ?
  - Détecter: comment l'incident sera-t-il détecté ?
  - Analyser: comment l'étendue de l'impact sera-t-elle déterminée ?
  - Contenir: comment l'incident sera-t-il isolé pour en limiter la portée ?
  - Éradiquer: comment éliminer la menace de l'environnement ?
  - Récupérer: comment le système ou la ressource concernés seront-ils remis en production ?
- Résultats attendus: une fois les requêtes et le code exécutés, quel est le résultat attendu du playbook ?

Ressources

Bonnes pratiques Well-Architected connexes :

- SEC10-BP02 - Développer des plans de gestion des incidents

Documents connexes :

- Cadre des playbooks de réponse aux incidents (langue française non garantie)
- Développer vos propres playbooks de réponse aux incidents (langue française non garantie)
- Exemples de playbooks de réponse aux incidents (langue française non garantie)
- Création d'un runbook de réponse aux incidents AWS à l'aide de playbooks Jupyter et CloudTrail Lake (langue française non garantie)

SEC10-BP05 Préallouer les accès

Vérifiez que les intervenants en cas d'incident disposent du bon accès préalablement alloué dans AWS afin de réduire le temps d'investigation jusqu'à la reprise.

Anti-modèles courants :

- Utilisation du compte racine pour la réponse aux incidents.
- Modification des comptes utilisateur existants.

- Manipulation des autorisations IAM directement lors de la fourniture d'une élévation de privilèges juste-à-temps.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

AWS recommande de réduire ou de supprimer l'utilisation des informations d'identification durables dans la mesure du possible et de privilégier les informations d'identification temporaire à la place, ainsi que les mécanismes d'élévation des privilèges juste-à-temps. Les informations d'identification durables sont sujettes aux risques de sécurité et augmentent les frais généraux opérationnels. Pour la plupart des tâches de gestion et de réponse aux incidents, nous vous recommandons de mettre en œuvre la fédération d'identité parallèlement à l'élévation temporaire pour l'accès administratif. Dans le cadre de ce modèle, un utilisateur demande une élévation à un niveau de privilège plus élevé (par exemple un rôle de réponse aux incidents) et, si l'utilisateur est admissible à cette élévation, une demande est envoyée à un approbateur. Si la demande est approuvée, l'utilisateur reçoit un ensemble d'informations d'identification AWS temporaires qui peuvent être utilisées afin d'exécuter les tâches. Une fois que ces informations d'identification ont expiré, l'utilisateur doit soumettre une nouvelle demande d'élévation.

Nous vous recommandons d'utiliser une élévation temporaire des privilèges dans la plupart des cas de réponse aux incidents. Dans cette optique, la meilleure solution consiste à utiliser AWS Security Token Service et les politiques de session afin de délimiter l'accès.

Dans certains cas, les identités fédérées ne sont pas disponibles, par exemple :

- Panne liée à la compromission d'un fournisseur d'identité (IdP).
- Mauvaise configuration ou erreur humaine entraînant la panne d'un système de gestion d'accès fédéré.
- Activité malveillante, par exemple un déni de service distribué (DDoS) ou une indisponibilité du système.

Dans les cas précédents, il doit y avoir un accès d'urgence de type Break Glass configuré afin de permettre l'analyse et la correction rapide des incidents. Nous vous recommandons également d'utiliser un utilisateur IAM disposant des autorisations appropriées pour effectuer des tâches et accéder aux ressources AWS. Utilisez des informations d'identification racine uniquement pour les tâches qui requièrent un accès en tant qu'utilisateur root. Pour vérifier que les intervenants en cas d'incident disposent d'un niveau d'accès approprié à AWS et aux autres systèmes pertinents, nous

vous recommandons de pré-allouer des comptes utilisateur dédiés. Les comptes utilisateur requièrent un accès privilégié et doivent être étroitement contrôlés et surveillés. Les comptes doivent être créés avec le moins de privilèges requis pour effectuer les tâches nécessaires et le niveau d'accès doit être basé sur les playbooks créés dans le cadre du plan de gestion des incidents.

Utilisez des utilisateurs et des rôles spécialement conçus et dédiés au titre de bonne pratique. L'élévation temporaire de l'accès des utilisateurs ou des rôles via l'ajout de politiques IAM ne permet pas de savoir clairement de quel type d'accès bénéficiaient les utilisateurs pendant l'incident et peut empêcher la révocation des privilèges élevés au niveau supérieur.

Il est important de supprimer autant de dépendances que possible afin de vérifier que l'accès peut être obtenu dans le plus grand nombre possible de scénarios de défaillance. Afin de vous faciliter la tâche, créez un playbook permettant de vérifier que les utilisateurs chargés des réponses en cas d'incident ont été créés en tant qu'utilisateurs AWS Identity and Access Management dans un compte de sécurité dédié et qu'ils ne sont pas gérés via une solution d'authentification unique ou de fédération existante. Chaque intervenant en cas d'incident doit avoir son propre compte nommé. La configuration du compte doit appliquer une politique stricte de gestion des mots de passe et une authentification multifactorielle (MFA). Si les playbooks de réponse aux incidents ne nécessitent qu'un accès à la AWS Management Console, l'utilisateur ne doit pas avoir de clés d'accès configurées et il doit lui être explicitement interdit de créer des clés d'accès. Cela peut être configuré avec des politiquesIAM ou des politiques de contrôle de service (SCP), comme mentionné dans les bonnes pratiques de sécurité AWS pour les SCP AWS Organizations. Les utilisateurs ne doivent pas avoir d'autres privilèges que la capacité d'assumer des rôles de réponse aux incidents dans d'autres comptes.

Pendant un incident, il peut être nécessaire d'accorder l'accès à d'autres personnes internes ou externes afin de prendre en charge les activités d'analyse, de correction ou de reprise. Dans ce cas, utilisez le mécanisme de playbook mentionné précédemment. Celui-ci doit comporter un processus permettant de s'assurer que tout accès supplémentaire est révoqué immédiatement après l'incident.

Pour s'assurer que l'utilisation des rôles de réponse aux incidents peut être correctement surveillée et vérifiée, il est essentiel que les comptes utilisateur IAM créés à cette fin ne soient pas partagés entre les personnes et que l'utilisateur root Compte AWS ne soit pas utilisé, sauf si cela s'avère nécessaire pour une tâche spécifique. Si l'utilisateur root est requis (par exemple, l'accès IAM à un compte spécifique n'est pas disponible), utilisez un processus distinct avec un playbook disponible afin de vérifier la disponibilité du mot de passe utilisateur root et du jeton d'authentification multifactorielle.

Pour configurer les politiques IAM des rôles de réponse aux incidents, pensez à utiliser IAM Access Analyzer pour générer des politiques basées sur les journaux AWS CloudTrail. Pour cela, accordez à

l'administrateur l'accès au rôle de réponse aux incidents sur un compte hors production et exécutez vos playbooks. Une fois que vous aurez terminé, vous pourrez créer une politique autorisant uniquement les mesures prises. Cette politique peut ensuite être appliquée à tous les rôles de réponse aux incidents dans tous les comptes. Vous pouvez éventuellement créer une politique IAM distincte pour chaque playbook afin de faciliter la gestion et la vérification. Les exemples de playbooks peuvent comprendre des plans d'intervention pour les rançongiciels, les atteintes à la protection des données, la perte d'accès à la production et d'autres scénarios.

Utilisez les comptes utilisateur de réponse aux incidents pour assumer les rôles IAM dédiés de réponse aux incidents dans d'autres Comptes AWS. Ces rôles doivent être configurés de façon à pouvoir être assumés uniquement par les utilisateurs du compte de sécurité et la relation de confiance doit exiger que le principal appelant ait été authentifié au moyen de l'authentification multifactorielle. Les rôles doivent utiliser des politiques IAM à portée limitée afin de contrôler l'accès. Assurez-vous que toutes les demandes `AssumeRole` pour ces rôles sont consignées dans CloudTrail et font l'objet d'une alerte, et que toutes les mesures prises en utilisant ces rôles sont consignées.

Il est vivement recommandé de nommer les comptes utilisateur IAM et les rôles IAM afin d'en faciliter la recherche dans les journaux CloudTrail. Par exemple, les comptes IAM pourraient être nommés *<USER_ID>*`-BREAK-GLASS` et les rôles IAM pourraient être nommés `BREAK-GLASS-ROLE`.

CloudTrail est utilisé pour consigner l'activité de l'API dans vos comptes AWS et doit être utilisé pour configurer les alertes sur l'utilisation des rôles de réponse aux incidents. Consultez la publication de blog sur la configuration des alertes lorsque les clés racine sont utilisées. Les instructions peuvent être modifiées de façon à configurer la métrique Amazon CloudWatch de filtre à filtre sur les événements `AssumeRole` liés au rôle IAM de réponse aux incidents :

```
{ $.eventName = "AssumeRole" && $.requestParameters.roleArn =
 "<INCIDENT_RESPONSE_ROLE_ARN>" && $.userIdentity.invokedBy NOT EXISTS && $.eventType !
= "AwsServiceEvent" }
```

Dans la mesure où les rôles de réponse aux incidents sont susceptibles d'avoir un niveau d'accès élevé, il est important que ces alertes soient transmises à un vaste groupe et qui y donnera suite rapidement.

Lors d'un incident, il est possible qu'un intervenant ait besoin d'accéder à des systèmes qui ne sont pas sécurisés directement par IAM. Il peut notamment s'agir d'instances Amazon Elastic Compute Cloud, de bases de données Amazon Relational Database Service ou de plateformes de logiciel en tant que service (SaaS). Il est vivement recommandé d'utiliser AWS Systems Manager Session

Manager plutôt que des protocoles natifs tels que SSH ou RDP pour tous les accès administratifs aux instances Amazon EC2. Cet accès peut être contrôlé à l'aide d'IAM, qui est sécurisé et vérifié. Il est également possible d'automatiser certaines parties de vos playbooks en utilisant des documents AWS Systems Manager,qui permettent de réduire les erreurs utilisateur et d'améliorer le temps de récupération. Pour accéder aux bases de données et aux outils tiers, nous recommandons de stocker les informations d'identification dans AWS Secrets Manager et d'accorder l'accès aux rôles des intervenants en cas d'incident.

En dernier lieu, la gestion des comptes utilisateur IAM de réponse aux incidents doit être ajoutée à vos processus d'entrées, de changements et de sorties. Elle doit également être vérifiée et testée régulièrement afin de vous assurer que seul l'accès prévu est autorisé.

Ressources

Documents connexes :

- Managing temporary elevated access to your AWS environment
- AWS Security Incident Response Guide
- AWS Elastic Disaster Recovery
- AWS Systems Manager Incident Manager
- Définition d'une politique de mot de passe du compte pour les utilisateurs IAM
- Utilisation de l'authentification multifactorielle (MFA) dans l'interface AWS
- Configuring Cross-Account Access with MFA
- Using IAM Access Analyzer to generate IAM policies
- Best Practices for AWS Organizations Service Control Policies in a Multi-Account Environment
- How to Receive Notifications When Your AWS Account's Root Access Keys Are Used
- Create fine-grained session permissions using IAM managed policies

Vidéos connexes :

- Automating Incident Response and Forensics in AWS
- DIY guide to runbooks, incident reports, and incident response
- Prepare for and respond to security incidents in your AWS environment

Exemples connexes :

- Atelier : AWS Account Setup and Root User
- Atelier : Incident Response with AWS Console and CLI

SEC10-BP06 Prédéployer les outils

Vérifiez que le personnel de sécurité dispose des outils appropriés préalablement déployés pour accélérer l'enquête jusqu'à la récupération.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Pour automatiser les fonctions de réponse et d'exploitation de la sécurité, vous pouvez utiliser un ensemble complet d'API et d'outils d'AWS. Vous pouvez automatiser entièrement la gestion des identités, la sécurité des réseaux, la protection des données et les fonctionnalités de surveillance, et les mettre en œuvre en utilisant les méthodes de développement de logiciel les plus courantes que vous avez déjà mises en place. Lorsque vous automatisez la sécurité, votre système peut surveiller, examiner et déclencher une réponse, plutôt que d'avoir à demander à des personnes de surveiller votre niveau de sécurité et de réagir manuellement aux événements.

Si vos équipes de réponse aux incidents continuent de répondre aux alertes de la même manière, elles risquent de se lasser des alertes. Au fil du temps, l'équipe peut faire moins attention aux alertes et soit faire des erreurs en gérant des situations ordinaires, soit manquer des alertes inhabituelles. L'automatisation permet d'éliminer la lassitude liée aux alertes en utilisant des fonctions qui traitent les alertes répétitives et ordinaires, laissant aux personnes le soin de gérer les incidents sensibles et uniques. L'intégration de systèmes de détection d'anomalies, comme Amazon GuardDuty, AWS CloudTrail Insights et Amazon CloudWatch Anomaly Detection, peut alléger les alertes courantes basées sur des seuils.

Vous pouvez améliorer les processus manuels en automatisant par programmation les étapes du processus. Une fois que vous avez défini le modèle de correction d'un événement, vous pouvez le décomposer en logique exploitable et écrire le code pour exécuter cette logique. Les intervenants peuvent ensuite exécuter ce code pour corriger le problème. Au fil du temps, vous pouvez automatiser un nombre croissant d'étapes et, enfin, gérer automatiquement des catégories entières d'incidents courants.

Au cours d'une enquête de sécurité, vous devez être en mesure d'examiner les journaux pertinents pour consigner et comprendre la portée et la chronologie complètes de l'incident. Des journaux sont également requis pour la génération d'alertes, indiquant que certaines actions intéressantes

ont eu lieu. Il est essentiel de sélectionner, d'activer, de stocker et de configurer les mécanismes d'interrogation et de récupération et de configurer les alertes. En outre, une solution efficace qui fournit des outils de recherche dans les données des journaux est Amazon Detective. »

AWS propose plus de 200 services cloud et des milliers de fonctionnalités. Nous vous recommandons de passer en revue les services susceptibles de prendre en charge et de simplifier votre stratégie de réponse aux incidents.

Outre la journalisation, vous devez développer et mettre en œuvre une cohérente. » Le balisage peut aider à mettre en contexte l'objectif d'une ressource AWS. Le balisage peut également être utilisé à des fins d'automatisation.

Étapes d'implémentation

Sélection et configuration de journaux à des fins d'analyse et d'alerte

Consultez la documentation suivante relative à la configuration de la journalisation pour la réponse aux incidents :

- Stratégies de journalisation pour la réponse aux incidents de sécurité (langue française non garantie)
- SEC04-BP01 Configurer une journalisation de service et d'application

Activation de la prise en charge de la détection et de la réponse pour les services de sécurité

AWS fournit des fonctionnalités natives de détection, de prévention et de réponse et d'autres services peuvent être utilisés pour concevoir des solutions de sécurité personnalisées. Pour obtenir la liste des services les plus pertinents en matière de réponse aux incidents de sécurité, consultez Définition de la capacité du cloud (langue française non garantie). »

Élaboration et mise en œuvre d'une stratégie de balisage

Il peut être difficile d'obtenir des informations contextuelles sur le cas d'utilisation métier et les parties prenantes internes pertinentes concernant une ressource AWS. Pour ce faire, vous pouvez utiliser des balises qui attribuent des métadonnées à vos ressources AWS. Ces balises comprennent une clé et une valeur définies par l'utilisateur. Vous pouvez créer des balises pour classer les ressources par objectif, propriétaire, environnement, type de données traitées et d'autres critères de votre choix.

Le fait de disposer d'une stratégie de balisage cohérente peut accélérer les temps de réponse et réduire le temps consacré au contexte organisationnel en vous permettant d'identifier et de discerner

rapidement les informations contextuelles relatives à une ressource AWS. Les balises peuvent également servir de mécanisme pour initier l'automatisation des réponses. Pour plus de détails sur les éléments à étiqueter, consultez Balisage de vos ressources AWS. » Vous devez d'abord définir les balises que vous souhaitez implémenter dans votre organisation. Ensuite, vous mettez en œuvre et appliquez cette stratégie de balisage. Pour plus de détails sur la mise en œuvre et l'application, consultez Mise en œuvre d'une stratégie de balisage des ressources AWS à l'aide de politiques de balisage AWS et de politiques de contrôle des services (SCP) (langue française non garantie).. »

Ressources

Bonnes pratiques Well-Architected connexes :

- SEC04-BP01 Configurer une journalisation de service et d'application
- SEC04-BP02 Analyse centralisée des journaux, des résultats et des métriques

Documents connexes :

- Stratégies de journalisation pour la réponse aux incidents de sécurité (langue française non garantie)
- Définition de la capacité du cloud pour la réponse aux incidents (langue française non garantie)

Exemples connexes :

- Détection et réponse aux menaces avec Amazon GuardDuty et Amazon Detective (langue française non garantie)
- Atelier Security Hub (langue française non garantie)
- Gestion des vulnérabilités avec Amazon Inspector (langue française non garantie)

SEC10-BP07 Exécuter des simulations

À mesure que les organisations se développent et évoluent au fil du temps, le paysage des menaces change. Il est donc important de revoir en permanence vos capacités de réponse aux incidents. L'organisation de simulations (également appelées « journées de jeu ») est une méthode qui peut être utilisée pour effectuer cette évaluation. Les simulations utilisent des scénarios d'événements de sécurité réels conçus pour imiter les tactiques, techniques et procédures (TTP) d'un acteur de la menace et permettre à une organisation d'exercer et d'évaluer ses capacités de réponse aux incidents en réagissant à ces cyberévénements fictifs tels qu'ils peuvent se produire dans la réalité.

Avantages liés au respect de cette bonne pratique : les simulations présentent de nombreux avantages :

- Validation de l'état de préparation à la cybersécurité et renforcement de la confiance de vos intervenants en cas d'incident.
- Test de la précision et de l'efficacité des outils et des flux de travail.
- Amélioration des méthodes de communication et de remontées en fonction de votre plan de réponse aux incidents.
- Possibilité de répondre à des vecteurs moins courants.

Niveau de risque exposé si cette bonne pratique n'est pas établie: moyen

Directives d'implémentation

Il existe trois principaux types de simulations :

- Exercices de simulation : l'approche théorique des simulations est une session basée sur des discussions auxquelles participent les différentes parties prenantes de la réponse aux incidents afin de mettre en pratique leurs rôles et leurs responsabilités et d'utiliser des outils de communication et des manuels établis. L'animation d'exercices peut généralement être réalisée en une journée complète dans un lieu virtuel, un lieu physique ou une combinaison des deux. Dans la mesure où il repose sur la discussion, l'exercice théorique met l'accent sur les processus, les personnes et la collaboration. La technologie fait partie intégrante de la discussion, mais l'utilisation effective d'outils ou de scripts de réponse aux incidents ne fait généralement pas partie de l'exercice théorique.
- Exercices de l'équipe violette : les exercices de l'équipe violette augmentent le niveau de collaboration entre les intervenants en cas d'incident (équipe bleue) et les acteurs de menaces simulées (équipe rouge). L'équipe bleue est composée de membres du centre des opérations de sécurité (SOC), mais peut également inclure d'autres parties prenantes qui seraient impliquées lors d'un véritable cyberévénement. L'équipe rouge est composée d'une équipe de tests de pénétration ou de parties prenantes clés formées à la sécurité offensive. L'équipe rouge travaille en collaboration avec les animateurs de l'exercice lors de la conception d'un scénario afin que celui-ci soit précis et réalisable. Lors des exercices de l'équipe violette, l'accent est principalement mis sur les mécanismes de détection, les outils et les procédures opérationnelles standard (SOP) qui soutiennent les efforts de réponse aux incidents.
- Exercices de l'équipe rouge : au cours d'un exercice de l'équipe rouge, l'attaque (l'équipe rouge) effectue une simulation pour atteindre un objectif donné ou un ensemble d'objectifs à partir d'une

portée prédéterminée. Les défenseurs (équipe bleue) ne seront pas nécessairement au courant de la portée ni de la durée de l'exercice, ce qui permet d'évaluer de manière plus réaliste la manière dont ils réagiraient en cas d'incident réel. Étant donné que les exercices de l'équipe rouge peuvent être des tests invasifs, soyez prudent et mettez en œuvre des contrôles pour vérifier que l'exercice ne cause pas de dommages réels à votre environnement.

Envisagez d'animer des simulations cybernétiques à intervalles réguliers. Chaque type d'exercice peut apporter des avantages uniques aux participants et à l'organisation dans son ensemble. Vous pouvez donc choisir de commencer par des types de simulation moins complexes (tels que des exercices théoriques) et de passer ensuite à des types de simulation plus complexes (exercices de l'équipe rouge). Vous devez sélectionner un type de simulation en fonction de la maturité de votre sécurité, de vos ressources et des résultats souhaités. Certains clients peuvent décider de ne pas effectuer les exercices de l'équipe rouge en raison de leur complexité et de leur coût.

Étapes d'implémentation

Quel que soit le type de simulation que vous choisissez, les simulations suivent généralement les étapes de mise en œuvre suivantes :

1. Définissez les principaux éléments de l'exercice : définissez le scénario de simulation et les objectifs de la simulation. Les deux doivent être acceptés par les dirigeants.
2. Identifiez les principales parties prenantes : un exercice nécessite au minimum des animateurs et des participants. Selon le scénario, d'autres parties prenantes telles que les services juridiques, l'équipe de communication ou la direction, peuvent être impliquées.
3. Créez et testez le scénario : le scénario devra peut-être être redéfini au fur et à mesure de sa création si des éléments spécifiques ne sont pas réalisables. Un scénario finalisé est attendu à l'issue de cette étape.
4. Animez la simulation : le type de simulation détermine l'animation utilisée (un scénario papier par rapport à un scénario simulé hautement technique). Les animateurs doivent adapter leurs tactiques d'animation aux objectifs de l'exercice et impliquer tous les participants dans l'exercice dans la mesure du possible afin d'en tirer le meilleur parti.
5. Effectuez un rapport après action (AAR) : identifiez les domaines qui se sont bien déroulés, ceux qui peuvent être améliorés et les lacunes potentielles. L'AAR doit mesurer l'efficacité de la simulation ainsi que la réponse de l'équipe à l'événement simulé afin que les progrès puissent être suivis au fil du temps lors de futures simulations.

Ressources

Documents connexes :

- Guide AWS de réponse aux incidents (langue française non garantie)

Vidéos connexes :

- AWS GameDay - Security Edition

SEC10-BP08 Établir un cadre pour tirer des enseignements des incidents

La mise en œuvre d'un cadre d'enseignements tirés et d'une capacité d'analyse des causes profondes peut non seulement contribuer à améliorer les capacités de réponse aux incidents, mais également à empêcher que l'incident ne se reproduise. En tirant les leçons de chaque incident, vous pouvez éviter de répéter les mêmes erreurs, expositions ou erreurs de configuration, non seulement en améliorant votre posture de sécurité, mais également en réduisant le temps perdu dans des situations évitables.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Il est important de mettre en œuvre un cadre d'enseignements tirés qui établit et atteint, à un niveau élevé, les points suivants :

- Quand se déroule un processus des enseignements tirés ?
- En quoi consiste le processus des enseignements tirés ?
- Comment se déroule un processus des enseignements tirés ?
- Qui est impliqué dans le processus et comment ?
- Comment les domaines à améliorer seront-ils identifiés ?
- Comment allez-vous vérifier que les améliorations sont suivies et mises en œuvre de manière efficace ?

Le cadre ne doit pas se concentrer sur les individus ni les blâmer, mais doit plutôt se concentrer sur l'amélioration des outils et des processus.

Étapes d'implémentation

Outre les résultats de haut niveau énumérés ci-dessus, il est important de poser les bonnes questions afin de tirer le meilleur parti (informations menant à des améliorations réalisables) du processus. Posez-vous les questions suivantes pour commencer à développer vos discussions sur les enseignements tirés :

- Quel a été l'incident ?
- Quand l'incident a-t-il été identifié pour la première fois ?
- Comment a-t-il été identifié ?
- Quels systèmes ont alerté sur l'activité ?
- Quels systèmes, services et données étaient concernés ?
- Que s'est-il passé précisément ?
- Qu'est-ce qui a bien fonctionné ?
- Qu'est-ce qui n'a pas bien fonctionné ?
- Quels processus ou procédures ont échoué ou n'ont pas pu être mis à l'échelle pour répondre à l'incident ?
- Qu'est-ce qui peut être amélioré dans les domaines suivants :
  - Les collaborateurs
    - Les personnes à contacter étaient-elles réellement disponibles et la liste de contacts était-elle à jour ?
    - Les personnes manquaient-elles de formation ou n'avaient-elles pas les capacités nécessaires pour intervenir et enquêter efficacement sur l'incident ?
    - Les ressources appropriées étaient-elles prêtes et disponibles ?
  - Les processus
    - Les processus et procédures ont-ils été suivis ?
    - Les processus et procédures étaient-ils documentés et disponibles pour cet incident ou ce type d'incident ?
    - Les processus et procédures requis étaient-ils absents ?
    - Les intervenants ont-ils pu accéder en temps opportun aux informations requises pour répondre au problème ?
  - La technologie

- Les systèmes d'alerte existants ont-ils identifié l'activité et ont-ils envoyé des alertes efficaces ?
- Comment aurions-nous pu réduire le délai de détection de 50 % ?
- Les alertes existantes doivent-elles être améliorées ou de nouvelles alertes doivent-elles être créées pour cet incident ou ce type d'incident ?
- Les outils existants ont-ils permis d'enquêter efficacement (recherche/analyse) sur l'incident ?
- Que peut-on faire pour identifier cet incident ou ce type d'incident plus rapidement ?
- Que peut-on faire pour éviter que cet incident ou ce type d'incident ne se reproduise ?
- À qui appartient le plan d'amélioration et comment allez-vous vérifier qu'il a été mis en œuvre ?
- Quel est le calendrier des contrôles et processus de surveillance ou de prévention supplémentaires à mettre en œuvre et à tester ?

Cette liste n'est pas exhaustive, mais vise à servir de point de départ pour identifier les besoins de l'organisation et de l'entreprise et la manière dont vous pouvez les analyser afin de tirer les meilleurs enseignements des incidents et d'améliorer en permanence votre posture de sécurité. Le plus important est de commencer par intégrer les enseignements tirés dans le cadre standard de votre processus de réponse aux incidents, de la documentation et des attentes des parties prenantes.

Ressources

Documents connexes :

- Guide de réponse aux incidents de sécurité AWS - Établir un cadre pour tirer des enseignements des incidents (langue française non garantie)
- Recommandations CAF du NCSC - Enseignements tirés (langue française non garantie)

Sécurité des applications

Question

- SÉC 11. Comment intégrer et valider les propriétés de sécurité des applications tout au long du cycle de vie de la conception, du développement et du déploiement ?

SÉC 11. Comment intégrer et valider les propriétés de sécurité des applications tout au long du cycle de vie de la conception, du développement et du déploiement ?

La formation du personnel, l'automatisation des tests, la compréhension des dépendances et la validation des propriétés de sécurité des outils et des applications contribuent à réduire la probabilité de problèmes de sécurité dans les charges de travail de production.

Bonnes pratiques

- SEC11-BP01 Formation à la sécurité des applications
- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication
- SEC11-BP03 Réalisation de tests de pénétration réguliers
- SEC11-BP04 Révisions de code manuelles
- SEC11-BP05 Centralisation des services pour les packages et les dépendances
- SEC11-BP06 Déploiement programmatique de logiciels
- SEC11-BP07 Évaluation régulière des caractéristiques de sécurité des pipelines
- SEC11-BP08 Création d'un programme permettant aux équipes responsables de la charge de travail de s'approprier les mécanismes de sécurité

SEC11-BP01 Formation à la sécurité des applications

Formez les concepteurs de votre organisation aux pratiques courantes de développement et d'exploitation sécurisés des applications. L'adoption de pratiques de développement axées sur la sécurité permet de réduire la probabilité d'apparition de problèmes décelés uniquement au stade de l'examen de la sécurité.

Résultat souhaité : les logiciels doivent être conçus et construits en tenant compte de la sécurité. Lorsque les concepteurs d'une organisation sont formés à des pratiques de développement sécurisées qui commencent par un modèle de menace, la qualité et la sécurité globales des logiciels produits s'en trouvent améliorées. Cette approche peut réduire le délai de livraison des logiciels ou des fonctionnalités, car moins de retouches sont nécessaires après la phase d'examen de la sécurité.

Aux fins de cette bonne pratique, le développement sécurisé désigne le logiciel conçu et les outils ou systèmes qui soutiennent le cycle du développement logiciel (SDLC).

Anti-modèles courants :

- Attendre un examen de la sécurité, puis réfléchir aux propriétés de sécurité d'un système.

- Laisser toutes les décisions en matière de sécurité à l'équipe de sécurité.
- Ne pas communiquer sur la manière dont les décisions prises au cours du cycle de développement du logiciel sont liées aux attentes ou aux politiques générales de l'organisation en matière de sécurité.
- S'impliquer trop tard dans le processus d'examen de la sécurité.

Avantages liés au respect de cette bonne pratique :

- Meilleure connaissance des exigences organisationnelles en matière de sécurité dès le début du cycle de développement.
- Possibilité d'identifier les problèmes de sécurité potentiels et d'y remédier plus rapidement, ce qui se traduit par une mise à disposition plus rapide des fonctionnalités.
- Amélioration de la qualité des logiciels et des systèmes.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Formez les concepteurs de votre organisation. Une formation à la modélisation des menaces constitue une bonne base pour se former à la sécurité. Idéalement, les concepteurs devraient pouvoir accéder en libre-service aux informations pertinentes pour leur charge de travail. Cet accès leur permet de prendre des décisions éclairées sur les propriétés de sécurité des systèmes qu'ils construisent, sans avoir à solliciter une autre équipe. Le processus de participation de l'équipe de sécurité aux révisions doit être clairement défini et simple à suivre. Les étapes du processus de révision doivent être ajoutées à la formation à la sécurité. Lorsque des modèles d'implémentation connus sont disponibles, ils doivent être faciles à trouver et à relier aux exigences de sécurité globales. Envisagez d'utiliser AWS CloudFormation, AWS Cloud Development Kit (AWS CDK) Constructs, Service Catalog, ou d'autres outils de modélisation pour réduire la nécessité d'une configuration personnalisée.

Étapes d'implémentation

- Commencez par donner aux concepteurs un cours sur la modélisation des menaces afin de leur donner une bonne base et de les aider à penser à la sécurité.
- Fournissez un accès aux formations AWS Training and Certification, de l'industrie ou des partenaires AWS.

- Dispensez une formation sur le processus d'examen de la sécurité de votre organisation, qui clarifie la répartition des responsabilités entre l'équipe chargée de la sécurité, les équipes responsables de la charge de travail et les autres parties prenantes.
- Publiez des conseils en libre-service sur la manière de répondre à vos exigences en matière de sécurité, y compris des exemples de code et des modèles, s'ils sont disponibles.
- Recueillez régulièrement les commentaires des équipes de concepteurs sur leur expérience du processus d'examen de la sécurité et de la formation, et utilisez ces commentaires pour vous améliorer.
- Utilisez des tests de simulation de pannes ou des campagnes de chasse aux bogues pour réduire le nombre de problèmes et améliorer les compétences de vos concepteurs.

Ressources

Bonnes pratiques associées :

- SEC11-BP08 Création d'un programme permettant aux équipes responsables de la charge de travail de s'approprier les mécanismes de sécurité

Documents connexes :

- AWS Training and Certification
- How to think about cloud security governance
- How to approach threat modeling
- Accelerating training – The AWS Skills Guild (Accélérer la formation – AWS Skills Guild)

Vidéos connexes :

- Proactive security: Considerations and approaches

Exemples connexes :

- Atelier sur la modélisation des menaces
- Sensibilisation des développeurs à l'industrie

Services associés :

- AWS CloudFormation
- AWS Cloud Development Kit (AWS CDK) (AWS CDK) Constructs
- Service Catalog
- AWS BugBust

SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Automatisez les tests des propriétés de sécurité tout au long du cycle de développement et de publication. L'automatisation facilite l'identification systématique et répétée des problèmes potentiels dans les logiciels avant leur diffusion, ce qui réduit le risque de problèmes de sécurité dans les logiciels fournis.

Résultat souhaité : l'objectif des tests automatisés est de fournir un moyen programmatique de détecter les problèmes potentiels de manière précoce et fréquente tout au long du cycle de développement. Lorsque vous automatisez les tests de régression, vous pouvez exécuter à nouveau les tests fonctionnels et non fonctionnels pour vérifier que le logiciel testé précédemment fonctionne toujours comme prévu après une modification. Lorsque vous définissez des tests d'unités de sécurité pour vérifier les erreurs de configuration courantes, telles qu'une authentification défaillante ou manquante, vous pouvez identifier et résoudre ces problèmes dès le début du processus de développement.

L'automatisation des tests utilise des cas de test spécifiques pour la validation de l'application, sur la base des exigences de l'application et de la fonctionnalité souhaitée. Le résultat du test automatisé est basé sur la comparaison entre le résultat du test généré et le résultat attendu, ce qui accélère le cycle de vie global du test. Les méthodologies de test telles que les tests de régression et les suites de tests d'unités sont les mieux adaptées à l'automatisation. L'automatisation des tests des propriétés de sécurité permet aux concepteurs de recevoir des commentaires automatisés sans avoir à attendre un examen de sécurité. Les tests automatisés sous forme d'analyse statique ou dynamique du code peuvent améliorer la qualité du code et aider à détecter les problèmes logiciels potentiels dès le début du cycle de développement.

Anti-modèles courants :

- Ne pas communiquer les cas de test et les résultats des tests automatisés.
- Effectuer uniquement les tests automatisés juste avant la mise en production.
- Automatiser les cas de test avec des exigences qui changent fréquemment.
- Ne pas fournir de recommandations sur la manière de traiter les résultats des tests de sécurité.

Avantages liés au respect de cette bonne pratique :

- Réduction de la dépendance à l'égard des personnes qui évaluent les propriétés de sécurité des systèmes.
- Le fait de disposer de résultats cohérents dans plusieurs domaines de travail améliore la cohérence.
- Réduction de la probabilité d'introduire des problèmes de sécurité dans les logiciels de production.
- Un délai plus court entre la détection et la remédiation grâce à une détection plus précoce des problèmes logiciels.
- Visibilité accrue des comportements systémiques ou répétés dans plusieurs domaines de travail, ce qui peut être utilisé pour apporter des améliorations à l'échelle de l'organisation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Au fur et à mesure du développement de votre logiciel, adoptez divers mécanismes de test pour vous assurer que vous testez votre application à la fois pour les exigences fonctionnelles, basées sur la logique commerciale de votre application, et pour les exigences non fonctionnelles, qui sont axées sur la fiabilité, la performance et la sécurité de l'application.

Les tests statiques de sécurité des applications (SAST) analysent votre code source à la recherche de schémas de sécurité anormaux et fournissent des indications sur le code sujet aux défauts. Les tests SAST s'appuient sur des données statiques, telles que la documentation (spécifications des exigences, documentation de conception et spécifications de conception) et le code source de l'application, pour tester une série de problèmes de sécurité connus. Les analyseurs de code statique permettent d'accélérer l'analyse de gros volumes de code. Le groupe de qualité NIST propose une comparaison des analyseurs de sécurité du code source, qui comprend des outils open source pour les lecteurs de codes à octets et les lecteurs de code binaire.

Complétez vos tests statiques par des méthodes de sécurité des applications (DAST), qui consistent à effectuer des tests sur l'application en cours d'exécution afin d'identifier les comportements potentiellement inattendus. Les tests dynamiques peuvent détecter des problèmes potentiels qui ne sont pas détectables par l'analyse statique. Les tests effectués aux stades du référentiel de code, de la build et du pipeline vous permettent de vérifier différents types de problèmes potentiels avant qu'ils ne s'introduisent dans votre code. Amazon CodeWhisperer fournit des recommandations sur le

code, y compris l'analyse de la sécurité, dans l'IDE du créateur. Amazon CodeGuru Reviewer peut identifier les problèmes critiques, les problèmes de sécurité et les bogues difficiles à trouver pendant le développement de l'application, et fournit des recommandations pour améliorer la qualité du code.

L'atelier sur la sécurité pour les développeurs utilise des outils de développement AWS, tels que AWS CodeBuild, AWS CodeCommit, et AWS CodePipeline, pour l'automatisation de la chaîne de production qui comprend les méthodologies de test SAST et DAST.

Au fur et à mesure que vous progressez dans votre cycle de développement du logiciel, mettez en place un processus itératif qui comprend des révisions périodiques des applications avec votre équipe de sécurité. Les commentaires recueillis lors de ces examens de sécurité doivent être traités et validés dans le cadre de l'examen de l'état de préparation à la mise en production. Ces examens permettent de définir un solide niveau de sécurité des applications et fournissent aux concepteurs des commentaires exploitables pour résoudre les problèmes potentiels.

Étapes d'implémentation

- Implémentez des outils cohérents d'IDE, de révision du code et de CI/CD qui incluent des tests de sécurité.
- Réfléchissez à l'étape du cycle de développement du logiciel où il convient de bloquer les pipelines au lieu de simplement avertir les concepteurs que des problèmes doivent être résolus.
- L'atelier sur la sécurité pour les développeurs fournit un exemple d'intégration des tests statiques et dynamiques dans un pipeline de publication.
- La réalisation de tests ou d'analyses de code à l'aide d'outils automatisés, tels que Amazon CodeWhisperer intégré aux IDE des développeurs et Amazon CodeGuru Reviewer pour l'analyse du code lors de la validation, aide les concepteurs à obtenir des commentaires au bon moment.
- Lorsque vous utilisez AWS Lambda pour votre conception, vous pouvez utiliser Amazon Inspector pour analyser le code de l'application dans vos fonctions.
- L'Atelier CI/CD AWS fournit un point de départ pour construire des pipelines CI/CD sur AWS.
- Lorsque les tests automatisés sont inclus dans les pipelines CI/CD, vous devez utiliser un système de tickets pour suivre la notification et la résolution des problèmes logiciels.
- Pour les tests de sécurité susceptibles de donner lieu à des conclusions, un lien vers des conseils pour remédier à la situation aide les concepteurs à améliorer la qualité du code.
- Analysez régulièrement les résultats des outils automatisés afin de donner la priorité à la prochaine automatisation, à la formation des concepteurs ou à la campagne de sensibilisation.

Ressources

Documents connexes :

- Livraison et déploiement continus

- Partenaires de compétence AWS DevOps
- Partenaires de compétence en matière de sécurité d'AWS pour la sécurité des applications
- Choosing a Well-Architected CI/CD approach
- Monitoring CodeCommit events in Amazon EventBridge and Amazon CloudWatch Events (Surveillance des événements CodeCommit sur Amazon EventBridge et Amazon CloudWatch Events)
- Secrets detection in Amazon CodeGuru Review
- Accelerate deployments on AWS with effective governance
- Comment AWS aborde l'automatisation de déploiement sans intervention et en toute sécurité

Vidéos connexes :

- Hands-off: Automating continuous delivery pipelines at Amazon
- Automating cross-account CI/CD pipelines

Exemples connexes :

- Sensibilisation des développeurs au secteur
- Gouvernance AWS CodePipeline (GitHub)
- Atelier sur la sécurité pour les développeurs
- Atelier CI/CD AWS

SEC11-BP03 Réalisation de tests de pénétration réguliers

Effectuez régulièrement des tests de pénétration de votre logiciel. Ce mécanisme permet d'identifier les problèmes logiciels potentiels impossibles à détecter par des tests automatisés ou une révision manuelle du code. Il peut également vous aider à comprendre l'efficacité de vos contrôles de détection. Les tests de pénétration doivent tenter de déterminer si le logiciel peut être amené

à fonctionner de manière inattendue, par exemple en exposant des données qui devraient être protégées ou en accordant des autorisations plus étendues que prévu.

Résultat souhaité : les tests de pénétration permettent de détecter, de remédier et de valider les propriétés de sécurité de votre application. Des tests de pénétration réguliers et programmés doivent être effectués dans le cadre du cycle de développement des logiciels (SDLC). Les résultats des tests de pénétration doivent être pris en compte avant le lancement du logiciel. Vous devez analyser les résultats des tests de pénétration pour déterminer s'il existe des problèmes qui pourraient être détectés grâce à l'automatisation. Le fait de disposer d'un processus de test de pénétration régulier et reproductible, qui comprend un mécanisme de commentaires actif, permet d'éclairer les conseils donnés aux concepteurs et d'améliorer la qualité des logiciels.

Anti-modèles courants :

- Les tests de pénétration ne concernent que les problèmes de sécurité connus ou répandus.
- Tests de pénétration d'applications sans outils et bibliothèques tiers dépendants.
- Uniquement des tests de pénétration pour les problèmes de sécurité des packages, et non l'évaluation de la logique métier implémentée.

Avantages liés au respect de cette bonne pratique :

- Confiance accrue dans les propriétés de sécurité du logiciel avant sa diffusion.
- Possibilité d'identifier des modèles d'application privilégiés, ce qui permet d'améliorer la qualité des logiciels.
- Une boucle de rétroaction permettant d'identifier plus tôt dans le cycle de développement où l'automatisation ou une formation supplémentaire peuvent améliorer les propriétés de sécurité des logiciels.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Le test de pénétration est un exercice de test de sécurité structuré dans lequel vous exécutez des scénarios de faille de sécurité planifiés afin de détecter des problèmes, d'y remédier et de valider les contrôles de sécurité. Les tests de pénétration commencent par une reconnaissance, au cours de laquelle des données sont recueillies sur la base de la conception actuelle de l'application et de

ses dépendances. Une liste de scénarios de test spécifiques à la sécurité est élaborée et exécutée. L'objectif principal de ces tests est de découvrir les problèmes de sécurité de votre application, qui pourraient être exploités pour obtenir un accès involontaire à votre environnement ou un accès non autorisé aux données. Vous devez effectuer des tests de pénétration lorsque vous lancez de nouvelles fonctionnalités, ou chaque fois que votre application a subi des changements majeurs en matière de fonction ou d'implémentation technique.

Vous devez identifier l'étape la plus appropriée du cycle de développement pour effectuer des tests de pénétration. Ces tests doivent avoir lieu suffisamment tard pour que la fonctionnalité du système soit proche de l'état final prévu, mais avec suffisamment de temps pour remédier aux éventuels problèmes.

Étapes d'implémentation

- Prévoyez un processus structuré pour l'étendue des tests de pénétration. Le fait de baser ce processus sur le modèle de menace est un bon moyen de maintenir le contexte.
- Identifiez l'endroit approprié dans le cycle de développement pour effectuer des tests de pénétration. Ce délai doit être respecté lorsque les changements attendus dans l'application sont minimes, mais qu'il reste suffisamment de temps pour mettre en œuvre des mesures correctives.
- Formez vos créateurs sur ce qu'il faut attendre des résultats des tests de pénétration et sur la manière d'obtenir des informations sur les mesures correctives.
- Utilisez des outils pour accélérer le processus de test de pénétration en automatisant les tests courants ou reproductibles.
- Analysez les résultats des tests de pénétration afin d'identifier les problèmes de sécurité systémiques et utilisez ces données pour effectuer des tests automatisés supplémentaires et former en permanence les créateurs.

Ressources

Bonnes pratiques associées :

- SEC11-BP01 Formation à la sécurité des applications
- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Documents connexes :

- Le test de pénétration AWS fournit des conseils détaillés pour les tests de pénétration sur AWS

- Accelerate deployments on AWS with effective governance
- Partenaires AWS disposant de la compétence Sécurité
- Modernize your penetration testing architecture on AWS Fargate
- AWS Fault Injection Simulator

Exemples connexes :

- Automate API testing with AWS CodePipeline (GitHub)
- Automated security helper (GitHub)

SEC11-BP04 Révisions de code manuelles

Procédez à une révision manuelle du code des logiciels que vous produisez. Ce processus permet de vérifier que la personne qui a écrit le code n'est pas la seule à en contrôler la qualité.

Résultat souhaité : l'inclusion d'une étape de révision manuelle du code au cours du développement permet d'améliorer la qualité du logiciel conçu, de renforcer les compétences des membres les moins expérimentés de l'équipe et d'identifier les domaines dans lesquels l'automatisation peut être utilisée. Les révisions de code manuelles peuvent être soutenues par des outils et des tests automatisés.

Anti-modèles courants :

- Ne pas effectuer de révision de code avant le déploiement.
- Faire rédiger et réviser le code par la même personne.
- Ne pas utiliser l'automatisation pour assister ou orchestrer les révisions de code.
- Ne pas former les créateurs à la sécurité des applications avant qu'ils ne procèdent à la révision du code.

Avantages liés au respect de cette bonne pratique :

- Amélioration de la qualité du code.
- Amélioration de la cohérence de développement du code grâce à la réutilisation d'approches communes.
- Réduction du nombre de problèmes découverts lors des tests de pénétration et des étapes ultérieures.

- Amélioration du transfert de connaissances au sein de l'équipe.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

L'étape de révision doit être implémentée dans le cadre du flux global de gestion du code. Les spécificités dépendent de l'approche utilisée pour les branches, les demandes d'extraction et la fusion. Vous utilisez peut-être AWS CodeCommit ou des solutions tierces telles que GitHub, GitLab ou Bitbucket. Quelle que soit la méthode utilisée, il est important de vérifier que vos processus exigent la révision du code avant qu'il ne soit déployé dans un environnement de production. L'utilisation d'outils tels que Amazon CodeGuru Reviewer peut faciliter l'orchestration du processus de révision de code.

Étapes d'implémentation

- Implémentez une étape de révision manuelle dans le cadre de votre flux de gestion du code et procédez à cette révision avant de poursuivre.
- Envisagez Amazon CodeGuru Reviewer pour gérer et vous aider dans les révisions de code.
- Implémentez un flux d'approbation qui exige qu'une révision de code soit achevée avant que le code puisse passer à l'étape suivante.
- Vérifiez qu'il existe un processus permettant d'identifier les problèmes découverts lors des révisions manuelles du code et qui pourraient être détectés automatiquement.
- Intégrez l'étape de révision manuelle du code d'une manière qui s'aligne sur vos pratiques de développement du code.

Ressources

Bonnes pratiques associées :

- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Documents connexes :

- Utilisation des demandes d'extraction dans les référentiels AWS CodeCommit
- Utilisation des modèles de règles d'approbation dans AWS CodeCommit
- À propos des demandes de tirage (pull requests)

- Automate code reviews with Amazon CodeGuru Reviewer
- Automating detection of security vulnerabilities and bugs in CI/CD pipelines using Amazon CodeGuru Reviewer CLI

Vidéos connexes :

- Continuous improvement of code quality with Amazon CodeGuru

Exemples connexes :

- Atelier sur la sécurité pour les développeurs

SEC11-BP05 Centralisation des services pour les packages et les dépendances

Fournissez des services centralisés aux équipes de créateurs pour l'obtention de packages logiciels et d'autres dépendances. Cela permet de valider les packages avant qu'ils ne soient incorporés au logiciel que vous écrivez, et fournit une source de données pour l'analyse du logiciel utilisé dans votre entreprise.

Résultat souhaité : un logiciel est composé d'un ensemble d'autres packages logiciels en plus du code qui est en train d'être écrit. Cela simplifie la consommation des implémentations de fonctionnalités utilisées de manière répétée, telles qu'un analyseur JSON ou une bibliothèque de chiffrement. La centralisation logique des sources de ces packages et dépendances offre aux équipes de sécurité un mécanisme de validation des propriétés des packages avant leur utilisation. Cette approche réduit également le risque qu'un problème inattendu soit causé par une modification d'un package existant, ou par des équipes de créateurs incluant des packages arbitraires provenant directement d'Internet. Utilisez cette approche en conjonction avec les flux de tests manuels et automatisés pour accroître la confiance dans la qualité du logiciel en cours de développement.

Anti-modèles courants :

- Extraction de packages à partir de référentiels arbitraires sur Internet.
- Ne pas tester les nouveaux packages avant de les mettre à la disposition des créateurs.

Avantages liés au respect de cette bonne pratique :

- Meilleure compréhension des packages utilisés dans le logiciel en cours de création

- Possibilité d'informer les équipes responsables de la charge de travail lorsqu'un package doit être mis à jour en fonction de la compréhension de qui utilise quoi.
- Réduire le risque qu'un package présentant des problèmes soit inclus dans votre logiciel.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Fournissez des services centralisés pour les packages et les dépendances d'une manière simple à utiliser pour les créateurs. Les services centralisés sont logiquement centraux plutôt que d'être implémentés sous la forme d'un système monolithique. Cette approche vous permet de fournir des services de manière à répondre aux besoins de vos concepteurs. Vous devez implémenter une méthode efficace pour ajouter des packages au référentiel lorsque des mises à jour sont effectuées ou que de nouvelles exigences apparaissent. Des services AWS tels que AWS CodeArtifact ou des solutions de partenaires AWS similaires permettent de fournir cette capacité.

Étapes d'implémentation :

- Implémentez un service de référentiel centralisé et logique, disponible dans tous les environnements où des logiciels sont développés.
- Prévoir l'accès au référentiel dans le cadre de la procédure d'attribution du Compte AWS.
- Concevez une automatisation pour tester les packages avant qu'ils ne soient publiés dans un référentiel.
- Conservez des métriques concernant les packages, les langages et les équipes les plus couramment utilisés et ayant subi le plus grand nombre de changements.
- Prévoyez un mécanisme automatisé permettant aux équipes de créateurs de demander de nouveaux packages et de fournir des commentaires.
- Analysez régulièrement les packages de votre référentiel afin d'identifier l'impact potentiel des problèmes récemment découverts.

Ressources

Bonnes pratiques associées :

- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Documents connexes :

- Accelerate deployments on AWS with effective governance
- Tighten your package security with CodeArtifact Package Origin Control toolkit
- Detecting security issues in logging with Amazon CodeGuru Reviewer
- Supply chain Levels for Software Artifacts (SLSA)

Vidéos connexes :

- Proactive security: Considerations and approaches
- The AWS Philosophy of Security (re:Invent 2017)
- When security, safety, and urgency all matter: Handling Log4Shell

Exemples connexes :

- Pipeline de publication de packages multirégionaux (GitHub)
- Publication de modules Node.js sur AWS CodeArtifact à l'aide de AWS CodePipeline (GitHub)
- Publication de packages AWS CDK Java CodeArtifact (GitHub)
- Distribuer des packages .NET NuGet privés avec AWS CodeArtifact (GitHub)

SEC11-BP06 Déploiement programmatique de logiciels

Dans la mesure du possible, procédez à des déploiements de logiciels par programme. Cette approche réduit la probabilité qu'un déploiement échoue ou qu'une erreur humaine entraîne un problème inattendu.

Résultat souhaité : éloigner les personnes des données est un principe clé pour construire en toute sécurité dans le AWS Cloud. Ce principe s'applique également à la manière dont vous déployez votre logiciel.

Ne pas dépendre d'individus pour déployer un logiciel vous permet d'être certain que ce que vous déployez correspond à ce que vous avez testé, et que le déploiement est effectué de manière cohérente à chaque fois. Il ne doit pas être nécessaire de modifier un logiciel afin qu'il fonctionne dans différents environnements. L'utilisation des principes du développement d'applications à douze facteurs, en particulier l'externalisation de la configuration, vous permet de déployer le même code dans plusieurs environnements sans avoir à le modifier. Le chiffrement de la signature des packages logiciels permet de vérifier que rien n'a changé d'un environnement à l'autre. Le résultat global de

cette approche est de réduire les risques dans votre processus de changement et d'améliorer la cohérence des versions du logiciel.

Anti-modèles courants :

- Déploiement manuel d'un logiciel en production.
- Modification manuelle d'un logiciel pour l'adapter à des environnements différents.

Avantages liés au respect de cette bonne pratique :

- Confiance accrue dans le processus de lancement des logiciels.
- Réduction du risque que l'échec d'une modification affecte l'entreprise.
- Augmentation de la cadence de lancement en raison de la diminution du risque de changement.
- Capacité de restauration automatique en cas d'événements inattendus au cours du déploiement.
- Capacité à prouver par chiffrement que le logiciel testé est celui qui est déployé.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Créez la structure de votre Compte AWS de manière à supprimer la récurrence de l'accès de personnes à partir d'environnements et utilisez des outils CI/CD pour effectuer des déploiements. Concevez vos applications de manière à ce que les données de configuration spécifiques à l'environnement soient obtenues à partir d'une source externe, telle que AWS Systems Manager Parameter Store. Signez les packages après les avoir testés et validez ces signatures lors du déploiement. Configurez vos pipelines CI/CD pour transmettre le code de l'application et utilisez des tests Canary pour confirmer le succès du déploiement. Utilisez des outils tels que AWS CloudFormation ou AWS CDK pour définir votre infrastructure, puis utilisez AWS CodeBuild et AWS CodePipeline pour effectuer des opérations CI/CD.

Étapes d'implémentation

- Créez des pipelines CI/CD bien définis pour rationaliser le processus de déploiement.
- Utilisez AWS CodeBuild et AWS Code Pipeline pour fournir une capacité CI/CD afin de faciliter l'intégration des tests de sécurité dans vos pipelines.
- Suivez les conseils sur la séparation des environnements dans le livre blanc Organisation de votre environnement AWS à l'aide de plusieurs comptes.

- Vérifiez que personne n'a accès aux environnements dans lesquels des charges de travail de production sont en cours d'exécution.
- Architecturez vos applications de manière à prendre en charge l'externalisation des données de configuration.
- Envisagez un modèle de déploiement bleu/vert.
- Implémentez des tests Canary pour valider la réussite du déploiement du logiciel.
- Utilisez des outils cryptographiques tels AWS Signer ou AWS Key Management Service (AWS KMS) pour signer et vérifier les packages logiciels que vous déployez.

Ressources

Bonnes pratiques associées :

- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Documents connexes :

- Atelier CI/CD AWS
- Accelerate deployments on AWS with effective governance
- Automatisation de déploiements sécurisés sans intervention
- Code signing using AWS Certificate Manager Private CA and AWS Key Management Service asymmetric keys
- Code Signing, a Trust and Integrity Control for AWS Lambda

Vidéos connexes :

- Hands-off: Automating continuous delivery pipelines at Amazon

Exemples connexes :

- Déploiements bleu/vert avec AWS Fargate

SEC11-BP07 Évaluation régulière des caractéristiques de sécurité des pipelines

Appliquez les principes du pilier Sécurité Well-Architected à vos pipelines, en accordant une attention particulière à la séparation des autorisations. Évaluez régulièrement les caractéristiques de sécurité de votre infrastructure de pipelines. Une gestion efficace de la sécurité des pipelines vous permet d'assurer la sécurité des logiciels qui transitent par ces pipelines.

Résultat souhaité : les pipelines utilisés pour construire et déployer votre logiciel doivent suivre les mêmes pratiques recommandées que toute autre charge de travail dans votre environnement. Les tests implémentés dans les pipelines ne doivent pas être modifiables par les créateurs qui les utilisent. Les pipelines ne doivent disposer que des autorisations nécessaires aux déploiements qu'ils effectuent et doivent implémenter des protections pour éviter de déployer dans les mauvais environnements. Les pipelines ne devraient pas s'appuyer sur des informations d'identification à long terme et devraient être configurés pour émettre un état afin que l'intégrité des environnements de création puisse être validée.

Anti-modèles courants :

- Tests de sécurité qui peuvent être contournés par les créateurs.
- Des autorisations trop larges pour les pipelines de déploiement.
- Les pipelines ne sont pas configurés pour valider les entrées.
- Ne pas passer régulièrement en revue les autorisations associées à votre infrastructure CI/CD.
- Utilisation d'informations d'identification à long terme ou codées en dur.

Avantages liés au respect de cette bonne pratique :

- Une plus grande confiance dans l'intégrité du logiciel conçu et déployé par le biais des pipelines.
- Possibilité d'interrompre un déploiement en cas d'activité suspecte.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Le fait de débuter avec des services CI/CD gérés qui prennent en charge les rôles IAM réduit le risque de fuite d'informations d'identification. L'application des principes de Pilier Sécurité à l'infrastructure de votre pipeline CI/CD peut vous aider à déterminer les améliorations à apporter en matière de sécurité. Suivre les recommandations de l'Architecture de référence des pipelines de déploiement d'AWS constitue un bon point de départ pour construire vos environnements CI/

CD. L'examen régulier de l'implémentation des pipelines et l'analyse des journaux à la recherche de comportements inattendus peuvent vous aider à comprendre les schémas d'utilisation des pipelines utilisés pour déployer des logiciels.

Étapes d'implémentation

- Commencez par suivre les recommandations de l'Architecture de référence des pipelines de déploiement d'AWS.
- Envisagez d'utiliser AWS IAM Access Analyzer pour générer de manière programmatique des politiques IAM de moindre privilège pour les pipelines.
- Intégrez des fonctions de surveillance et d'alerte à vos pipelines afin d'être informé des activités inattendues ou anormales, car les services Amazon EventBridge gérés AWS vous permettent d'acheminer les données vers des cibles telles que AWS Lambda ou Amazon Simple Notification Service (Amazon SNS).

Ressources

Documents connexes :

- Architecture de référence des pipelines de déploiement d'AWS
- Surveillance d'AWS CodePipeline
- Bonnes pratiques de sécurité pour AWS CodePipeline

Exemples connexes :

- DevOps monitoring dashboard (GitHub)

SEC11-BP08 Création d'un programme permettant aux équipes responsables de la charge de travail de s'approprier les mécanismes de sécurité

Créez un programme ou un mécanisme qui permette aux équipes de créateurs de prendre des décisions en matière de sécurité pour les logiciels qu'ils créent. Votre équipe de sécurité doit bien sûr procéder à un examen de ces décisions afin de les valider. Mais le fait de rendre les équipes de créateurs responsables de la sécurité permet d'élaborer des charges de travail plus rapides et plus sûres. Ce mécanisme favorise également une culture de responsabilisation qui a un impact positif sur le fonctionnement des systèmes que vous construisez.

Résultat souhaité : pour rendre les équipes de créateurs responsables de la sécurité et décisionnaires, vous pouvez soit former les créateurs aux implications de la sécurité, soit compléter leur formation par des personnes chargées de la sécurité, intégrées ou associées aux équipes de créateurs. Les deux approches sont pertinentes et permettent à l'équipe de prendre des décisions de meilleure qualité en matière de sécurité plus tôt dans le cycle de développement. Ce modèle de responsabilité repose sur la formation à la sécurité des applications. En commençant par le modèle de menace correspondant à une charge de travail donnée, il est possible d'axer le design thinking sur le contexte approprié. Disposer d'une communauté de créateurs axés sur la sécurité ou d'un groupe d'ingénieurs en sécurité travaillant avec des équipes de créateurs présente un autre avantage : la possibilité de comprendre plus en profondeur comment les logiciels sont écrits. Cette compréhension vous aide à déterminer les prochains domaines d'amélioration de votre capacité d'automatisation.

Anti-modèles courants :

- Laisser à une équipe de sécurité le soin de prendre toutes les décisions relatives à la conception de la sécurité.
- Ne pas tenir compte des exigences de sécurité suffisamment tôt dans le processus de développement.
- Ne pas recueillir de commentaires des créateurs et des responsables de la sécurité sur le fonctionnement du programme.

Avantages liés au respect de cette bonne pratique :

- Réduction du temps nécessaire à la réalisation des examens de sécurité.
- Réduction des problèmes de sécurité qui ne sont détectés qu'au stade de l'examen de la sécurité.
- Amélioration de la qualité globale du logiciel en cours d'écriture.
- Possibilité d'identifier et de comprendre les problèmes systémiques ou les domaines d'amélioration à forte valeur ajoutée.
- Réduction de la quantité de travail à refaire en raison des conclusions de l'examen de sécurité.
- Amélioration de la perception de la fonction de sécurité.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Commencez par suivre les conseils de SEC11-BP01 Formation à la sécurité des applications. Identifiez ensuite le modèle opérationnel du programme qui vous semble le plus adapté à votre organisation. Les deux principaux modèles consistent à former les créateurs ou à intégrer les responsables de la sécurité dans les équipes de créateurs. Une fois que vous avez décidé de l'approche initiale, vous devez mener un projet pilote avec une seule équipe ou un petit groupe d'équipes de charge de travail afin de prouver que le modèle fonctionne pour votre organisation. Le soutien de la direction de l'organisation en matière de construction et de sécurité contribue à la mise en œuvre et à la réussite du programme. Lors de la création de ce programme, il est important de choisir des métriques qui peuvent être utilisées pour montrer la valeur du programme. Apprendre de la manière dont AWS les autres ont abordé ce problème est une bonne expérience d'apprentissage. Cette bonne pratique est très axée sur le changement organisationnel et la culture. Les outils que vous utilisez doivent favoriser la collaboration entre les créateurs et les responsables de la sécurité.

Étapes d'implémentation

- Commencez par former vos créateurs à la cybersécurité des applications.
- Créer une communauté et un programme d'intégration pour former les créateurs.
- Choisissez un nom pour le programme. Les termes « tuteur », « champion » ou « défenseur » sont couramment utilisés.
- Identifier le modèle à utiliser : former des créateurs, intégrer des ingénieurs en sécurité ou avoir des rôles de sécurité connexes.
- Identifier les sponsors du projet parmi les responsables de la sécurité, les créateurs et éventuellement d'autres groupes concernés.
- Suivez les métriques concernant le nombre de personnes impliquées dans le programme, le temps nécessaire aux examens et les commentaires des créateurs et des responsables de la sécurité. Utilisez ces métriques pour apporter des améliorations.

Ressources

Bonnes pratiques associées :

- SEC11-BP01 Formation à la sécurité des applications
- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Documents connexes :

- How to approach threat modeling
- How to think about cloud security governance

Vidéos connexes :

- Proactive security: Considerations and approaches

Fiabilité

Le pilier Fiabilité englobe la capacité d'une charge de travail à exécuter sa fonction de manière correcte et cohérente et ce, en temps utile. Vous trouverez des recommandations sur l'implémentation dans le Livre blanc du pilier Fiabilité.

Domaines de bonnes pratiques

- Fondations
- Architecture de charge de travail
- Gestion des modifications
- Gestion des défaillances

Fondations

Questions

- FIA 1. Comment gérez-vous les Service Quotas et les contraintes de service ?
- FIA 2. Comment planifiez-vous la topologie de votre réseau ?

FIA 1. Comment gérez-vous les Service Quotas et les contraintes de service ?

Pour les architectures de charge de travail basées sur le cloud, il existe des Service Quotas (également appelés limites de services). Le rôle de ces quotas est d'empêcher la mise en service accidentelle de plus de ressources que nécessaire et de limiter les taux de demandes sur les opérations d'API afin de protéger les services contre les abus. Il existe également des contraintes de ressource. Par exemple, la vitesse à laquelle vous pouvez transmettre des bits sur un câble de fibre optique, ou la quantité de stockage sur un disque physique.

Bonnes pratiques

- REL01-BP01 Connaissance des quotas de service et des contraintes
- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP03 Tenir compte des quotas et des contraintes de service fixes dans l'architecture
- REL01-BP04 Surveiller et gérer les quotas
- REL01-BP05 Automatiser la gestion des quotas
- REL01-BP06 Garantir un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement

REL01-BP01 Connaissance des quotas de service et des contraintes

Connaissez vos quotas par défaut et gérez vos demandes d'augmentation de quota pour votre architecture de charge de travail. Connaissez également les contraintes de ressources, comme le disque ou le réseau, qui sont susceptibles d'avoir un impact.

Résultat souhaité : les clients peuvent prévenir la dégradation ou l'interruption des services dans leurs Comptes AWS en mettant en œuvre des directives appropriées pour la surveillance des métriques clés, les vérifications de l'infrastructure et l'automatisation des étapes de remédiation pour vérifier que les quotas des services et les contraintes ne sont pas atteints, ce qui pourrait entraîner une dégradation ou une interruption des services.

Anti-modèles courants :

- Déployer une charge de travail sans comprendre les quotas matériels ou logiciels et leurs limites pour les services utilisés.
- Déployer une charge de travail de remplacement sans analyser ni reconfigurer les quotas nécessaires ou contacter d'abord l'assistance.
- Supposer que les services cloud sont sans limite et que les services peuvent être utilisés sans prendre en compte les taux, les limites, les nombres et les quantités.
- Supposer que les quotas augmenteront automatiquement.
- Ne pas connaître le processus et la chronologie des demandes de quotas.
- Supposer que le quota du service cloud pas défaut est le même pour chaque service par rapport à d'autres régions.
- Supposer que les contraintes de service peuvent être enfreintes et que les systèmes se mettront automatiquement à l'échelle ou augmenteront la limite au-delà des contraintes de la ressource.

- Ne pas tester l'application sur des pics de trafic pour tester la résistance de l'utilisation de ces ressources.
- Provisionner les ressources sans analyser la taille de ressource nécessaire.
- Surprovisionner la capacité en choisissant des types de ressources largement supérieures aux besoins réels ou aux pics attendus.
- Ne pas évaluer les exigences de capacité pour les nouveaux niveaux de trafic avant un nouvel événement client ou le déploiement d'une nouvelle technologie.

Avantages liés au respect de cette bonne pratique : la surveillance et la gestion automatisée des quotas de service et des contraintes de ressources peuvent proactivement réduire les échecs. Les changements dans les modèles de trafic d'un service client peuvent entraînement une interruption ou une dégradation si les bonnes pratiques ne sont pas suivies. En surveillante et en gérant ces valeurs sur toutes les régions et tous les comptes, les applications peuvent bénéficier d'une meilleure résilience lors d'événements indésirables ou imprévus.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Service Quotas est un service AWS qui vous aide à gérer vos quotas pour plus de 250 services AWS depuis un seul et même emplacement. En plus de rechercher les valeurs de quota, vous pouvez également demander et suivre les augmentations de quota à partir de la console Service Quotas ou via le kit SDK AWS. AWS Trusted Advisor propose un contrôle des quotas de service qui affiche votre utilisation et les quotas de différents aspects de certains services. Les quotas de service par défaut par service figurent également dans la documentation AWS de chaque service (par exemple, consultez Quotas Amazon VPC).

Certaines limites de services, comme les limites de taux sur les API limitées sont définies dans Amazon API Gateway en configurant un plan d'utilisation. Certaines limites définies en tant que configuration sur leurs services respectifs incluent les IOPS provisionnés, le stockage Amazon RDS alloué et les allocations de volume Amazon EBS. Amazon Elastic Compute Cloud dispose de son propre tableau de bord des limites de service qui peut vous aider à gérer votre instance, Amazon Elastic Block Store et les limites d'adresses IP Elastic. Si vous possédez un cas d'utilisation pour lequel les quotas de service affectent les performances de votre application sans être ajustables à vos besoins, contactez AWS Support pour déterminer si des mesures d'atténuation peuvent être implémentées.

Les quotas de service peuvent être spécifiques à une région ou mondiaux par nature. Un service AWS qui atteint son quota ne se comportera pas comme lors d'une utilisation normale et peut entraîner une interruption ou une dégradation du service. Par exemple, un quota de service limité le nombre de DL Amazon EC2 qui peuvent être utilisés dans une région et cette limite peut être atteinte lors d'un événement de mise à l'échelle du trafic avec des groupes Auto Scaling (ASG).

L'utilisation des quotas de service pour chaque compte doit être évaluée régulièrement pour déterminer quelles seraient les limites de service appropriées pour ce compte. Ces quotas de service existent en tant que barrières de protection opérationnelles pour empêcher le provisionnement accidentel de plus de ressources que nécessaire. Ils servent également à limiter les taux de requêtes sur les opérations d'API pour protéger les services des abus.

Les contraintes de service sont différentes des quotas de service. Les contraintes de service représentent les limites d'une ressource spécifique, telles que définies par ce type de ressource. Il peut s'agir de la capacité de stockage (par exemple, gp2 a une limite de 1 Go à 16 To) ou du débit du disque (10 000 IOPS). Il est essentiel qu'une contrainte d'un type de ressource soit optimisée et constamment évaluée par rapport à une utilisation qui pourrait atteindre ses limites. Si une contrainte est atteinte de manière inattendue, les applications ou les services du compte peuvent être dégradés ou interrompus.

S'il existe un cas d'utilisation pour lequel les quotas de service affectent les performances d'une application sans être ajustables à vos besoins, contactez AWS Support pour déterminer si des améliorations sont possibles. Pour plus d'informations sur l'ajustement des quotas fixes, consultez REL01-BP03 Tenir compte des quotas et des contraintes de service fixes dans l'architecture.

Il existe un grand nombre de services et d'outils AWS pour vous aider à surveiller et gérer Service Quotas. Les services et les outils doivent être exploités pour vérifier automatiquement ou manuellement les niveaux de quotas.

- AWS Trusted Advisor propose un contrôle des quotas de service qui affiche votre utilisation et les quotas de différents aspects de certains services. Il peut aider à identifier des services proches du quota.
- AWS Management Console fournit des méthodes permettant d'afficher les valeurs des quotas de service, de les gérer, de demander de nouveaux quotas, de surveiller le statut des demandes de quotas et d'afficher l'historique des quotas.
- AWS CLI et CDK offrent des méthodes par programmation pour gérer et surveiller automatiquement les niveaux et l'utilisation des quotas de service.

Étapes d'implémentation

Pour Service Quotas :

- Vérifier AWS Service Quotas.
- Pour connaître vos quotas de service existant, déterminez les services (comme IAM Access Analyzer) utilisés. Il existe environ 250 services AWS contrôlés par des quotas de service. Ensuite, déterminez le nom du quota de service spécifique qui pourrait être utilisé au sein de chaque compte et région. Il existe environ 3 000 noms de quotas de service par région.
- Augmentez cette analyse des quotas avec AWS Config pour trouver toutes les ressources AWS utilisées dans vos Comptes AWS.
- Utilisez les données AWS CloudFormation pour déterminer vos ressources AWS utilisées. Examinez les ressources créées dans AWS Management Console ou via la commande `list-stack-resources` de l'AWS CLI. Vous pouvez également voir les ressources configurées pour être déployées directement dans le modèle.
- Déterminez tous les services indispensables à votre charge de travail en prenant en compte le code de déploiement.
- Identifiez les quotas de service pertinents. Utilisez les informations accessibles par programmation par le biais de Trusted Advisor et Service Quotas.
- Établissez une méthode de surveillance automatisée (consultez REL01-BP02 Gérer les quotas de service entre les comptes et les régions et REL01-BP04 Surveiller et gérer les quotas) pour vous alerter et vous informer si des quotas de service sont sur le point d’atteindre ou ont atteint leur limite.
- Établissez une méthode automatisée et par programmation pour vérifier si un quota de service a été modifié dans une région mais pas dans d'autres régions au sein du même compte (consultez REL01-BP02 Gérer les quotas de service entre les comptes et les régions et REL01-BP04 Surveiller et gérer les quotas).
- Automatisez l'analyse des journaux et des métriques de l'application pour déterminer s'il existe des erreurs de quotas ou de contraintes de service. Si de telles erreurs existent, envoyez des alertes au système de surveillance.
- Établissez des procédures d'ingénierie pour calculer le changement de quota nécessaire (consultez REL01-BP05 Automatiser la gestion des quotas) une fois qu'il a été identifié que des quotas plus importants sont nécessaires pour des services spécifiques.

- Créez une flux de travail de provisionnement et d'approbation pour demander des modifications des quotas de service. Cela doit inclure un flux de travail d'exception en cas de refus d'une demande ou d'une approbation partielle.
- Créez une méthode d'ingénierie pour vérifier les quotas de service avant le provisionnement et utilisez de nouveaux services AWS avant le déploiement dans des environnements de production ou chargés (par exemple, un compte de test de charge).

Pour les contraintes de service :

- Établissez des méthodes de surveillance et des métriques pour alerter quand les ressources sont proches de leurs contraintes. Tirez profit de CloudWatch tel que nécessaire pour la surveillance des métriques ou des journaux.
- Établissez des seuils d'alertes pour chaque ressource ayant une contrainte importante pour l'application ou le système.
- Créez des procédures de gestion des flux de travail et de l'infrastructure pour changer le type de ressource si la contrainte est proche. Ce flux de travail doit inclure le test de charge comme une bonne pratique pour vérifier que ce nouveau type est le bon type de ressource avec les nouvelles contraintes.
- Procédez à la migration des ressources identifiées vers le nouveau type de ressource recommandé avec les procédures et les processus existants.

Ressources

Bonnes pratiques associées :

- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP03 Tenir compte des quotas et des contraintes de service fixes dans l'architecture
- REL01-BP04 Surveiller et gérer les quotas
- REL01-BP05 Automatiser la gestion des quotas
- REL01-BP06 Garantir un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement
- REL03-BP01 Choisir comment segmenter votre charge de travail
- REL10-BP01 Déployer la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la réparation sur toutes les couches

- REL12-BP05 Tester la résilience à l'aide de l'ingénierie du chaos

Documents connexes :

- Pilier Fiabilité du cadre AWS Well-Architected : Disponibilité
- AWS Service Quotas (anciennement appelés limites de service)
- Vérifications des bonnes pratiques AWS Trusted Advisor (voir la section Limites de service)
- AWS limit monitor on AWS answers (Surveillance de limites AWS sur les réponses AWS)
- Amazon EC2 Service Limits (Limites de service EC2)
- Qu'est-ce qu'Service Quotas ?
- How to Request quota increase (Comment demander une augmentation du quota)
- Service endpoints and quotas (Points de terminaison et quotas de service)
- Guide de l'utilisateur Service Quotas
- Quota Monitor for AWS (Surveillance de quotas pour AWS)
- AWS Fault Isolation Boundaries (Limites d'isolement des pannes AWS)
- Availability with redundancy (Disponibilité avec redondance)
- AWS pour les données
- Qu'est-ce que l'intégration continue ?
- Qu'est-ce que la livraison continue ?
- Partenaire APN : partenaires facilitant la gestion de la configuration
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS (Gestion du cycle de vie des comptes dans les environnements SaaS de type compte par locataire sur AWS)
- Gestion et surveillance de la limitation des API dans vos charges de travail
- View AWS Trusted Advisor recommendations at scale with AWS Organizations (Examiner les recommandations AWS Trusted Advisor à grande échelle avec AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower (Automatisation de l'augmentation des limites de service et Enterprise Support avec AWS Control Tower)

Vidéos connexes :

- AWS Live re:Inforce 2019 - Service Quotas

- View and Manage Quotas for AWS Services Using Service Quotas (Examiner et gérer les quotas pour les services AWS avec les quotas de service)
- AWS IAM Quotas Demo (Démonstration sur les quotas IAM d'AWS)

Outils associés :

- Amazon CodeGuru Reviewer
- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP02 Gérer les quotas de service entre les comptes et les régions

Si vous utilisez plusieurs comptes ou régions, demandez les quotas appropriés dans tous les environnements où vos charges de travail de production s'exécutent.

Résultat souhaité : Les services et les applications ne doivent pas être affectés par un épuisement de quota de service pour les configurations qui englobent les comptes ou les régions ou dont les conceptions de résilience s'appuient sur le basculement de zone, de région ou de compte.

Anti-modèles courants :

- Laisser l'utilisation des ressources dans une région d'isolement se développer sans aucun mécanisme pour maintenir de la capacité dans les autres zones.
- Définir manuellement tous les quotas de manière indépendante dans les régions d'isolement.
- Ne pas prendre en considération l'effet des architectures de résilience (par ex., actives ou passives) dans les futurs besoins de quotas alors qu'une dégradation est observée dans la région non principale.

- Ne pas évaluer les quotas régulièrement et ne pas apporter les changements qui s'imposent dans chaque région et chaque compte où la charge de travail s'exécute.
- Ne pas tirer parti des modèles de demande de quota pour demander des augmentations dans plusieurs régions et comptes.
- Ne pas mettre à jour les quotas de service pensant à tort que l'augmentation de quotas a des répercussions sur les coûts comme les demandes de réservation de capacité de calcul.

Avantages liés au respect de cette bonne pratique : Possibilité de vérifier que vous êtes en mesure de gérer votre charge actuelle dans les régions ou comptes secondaires au cas où les services régionaux viendraient à être indisponibles. Cela peut contribuer à limiter le nombre d'erreurs ou les niveaux de dégradations observés lors d'une perte de région.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les quotas de service sont suivis par compte. Sauf indication contraire, chaque quota est propre à une Région AWS. En plus des environnements de production, tâchez également de gérer les quotas dans tous les autres environnements applicables de façon à ne pas entraver les tests et le développement. Pour maintenir un haut niveau de résilience, il convient d'évaluer constamment les quotas de service (que ce soit de façon automatisée ou manuelle).

Compte tenu du plus grand nombre de charges de travail englobant les régions du fait de l'implémentation de conceptions utilisant les approches Actif/Actif, Actif/Passif – Chaud, Actif/Passif – Froid et Actif/Passif – Veilleuse, il est essentiel de comprendre tous les niveaux de quota de région et de compte. Les modèles de trafic passés ne permettent pas toujours de déterminer correctement si le quota de service est bien défini.

Tout aussi important, la longueur limite des noms de quota de service n'est pas toujours identique d'une région à l'autre. Ainsi, cette valeur peut être égale à cinq dans une région et à dix dans une autre. La gestion de ces quotas doit englober tous les services, comptes et régions identiques pour offrir une résilience cohérente dans des conditions de charge.

Rapprochez toutes les différences de quota de service entre les différentes régions (région active ou région passive) et créez des processus permettant de rapprocher constamment ces différences. Les plans de test de basculements de régions passives sont rarement mis à l'échelle pour atteindre une capacité active de pointe, ce qui signifie que les exercices de simulation (« game day ») et les exercices de table (« table top ») ne permettent pas nécessairement d'identifier les différences dans les quotas de service entre les régions et donc de maintenir les limites adéquates.

La dérive de quota de service, condition où les limites de quota de service pour un quota nommé spécifique changent dans une région mais pas dans toutes, est très importante pour le suivi et l'évaluation. Il doit être envisagé de changer le quota dans les régions qui présentent du trafic ou qui pourraient potentiellement en véhiculer.

- Sélectionnez les comptes et les régions appropriés en fonction de vos exigences de service, de latence, de réglementation et de reprise après sinistre (DR).
- Identifiez les quotas de services dans l'ensemble des comptes, régions et zones de disponibilité appropriés. Les limites s'appliquent au compte et à la région. Ces valeurs doivent être comparées pour repérer les différences.

Étapes d'implémentation

- Examinez les valeurs Service Quotas susceptibles d'avoir transgressé le niveau d'utilisation à risque. AWS Trusted Advisor propose des alertes pour les violations de seuil de 80 % et 90 %.
- Examinez les valeurs de quotas de service dans les régions passives (dans une conception de type Actif/Passif). Vérifiez que la charge s'exécutera correctement dans les régions secondaires en cas de défaillance dans la région principale.
- Automatisez l'évaluation pour identifier une éventuelle dérive de quota de service entre des régions d'un même compte et agissez en conséquence pour changer les limites.
- Si la structure des unités d'organisation (UO) est prise en charge, les modèles de quota de service doivent être mis à jour en fonction des changements apportés aux quotas qui doivent s'appliquer à plusieurs régions et comptes.
  - Créez un modèle et associez les régions au changement de quota.
  - Examinez tous les modèles de quota de service existants pour y apporter les changements nécessaires (région, limites et comptes).

Ressources

Bonnes pratiques associées :

- REL01-BP01 Connaissance des quotas de service et des contraintes
- REL01-BP03 Tenir compte des quotas et des contraintes de service fixes dans l'architecture
- REL01-BP04 Surveiller et gérer les quotas
- REL01-BP05 Automatiser la gestion des quotas

- REL01-BP06 Garantir un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement
- REL03-BP01 Choisir comment segmenter votre charge de travail
- REL10-BP01 Déployer la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la réparation sur toutes les couches
- REL12-BP05 Tester la résilience à l'aide de l'ingénierie du chaos

Documents connexes :

- Pilier Fiabilité du cadre AWS Well-Architected : Disponibilité
- AWS Service Quotas (anciennement appelés limites de service)
- Vérifications des bonnes pratiques AWS Trusted Advisor (voir la section Limites de service)
- AWS limit monitor on AWS answers (Surveillance de limites AWS sur les réponses AWS)
- Amazon EC2 Service Limits (Limites de service EC2)
- Qu'est-ce qu'Service Quotas ?
- How to Request quota increase (Comment demander une augmentation du quota)
- Service endpoints and quotas (Points de terminaison et quotas de service)
- Guide de l'utilisateur Service Quotas
- Quota Monitor for AWS (Surveillance de quotas pour AWS)
- AWS Fault Isolation Boundaries (Limites d'isolement des pannes AWS)
- Availability with redundancy (Disponibilité avec redondance)
- AWS pour les données
- Qu'est-ce que l'intégration continue ?
- Qu'est-ce que la livraison continue ?
- Partenaire APN : partenaires facilitant la gestion de la configuration
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS (Gestion du cycle de vie des comptes dans les environnements SaaS de type compte par locataire sur AWS)
- Gestion et surveillance de la limitation des API dans vos charges de travail
- View AWS Trusted Advisor recommendations at scale with AWS Organizations (Examiner les recommandations AWS Trusted Advisor à grande échelle avec AWS Organizations)

- Automating Service Limit Increases and Enterprise Support with AWS Control Tower (Automatisation de l'augmentation des limites de service et Enterprise Support avec AWS Control Tower)

Vidéos connexes :

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Examiner et gérer les quotas pour les services AWS avec les quotas de service)
- AWS IAM Quotas Demo (Démonstration sur les quotas IAM d'AWS)

Services associés :

- Amazon CodeGuru Reviewer
- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP03 Tenir compte des quotas et des contraintes de service fixes dans l'architecture

Ayez conscience des quotas de service non modifiables, des contraintes de service et des limites de ressources physiques. Concevez des architectures pour les applications et les services afin d'éviter que ces limites n'aient un impact sur la fiabilité.

Par exemple, la bande passante du réseau, la taille de la charge utile des appels de fonctions sans serveur, le taux d'accélération d'une passerelle API et les connexions simultanées d'utilisateurs à une base de données.

Résultat souhaité : l'application ou le service fonctionne comme prévu dans des conditions de trafic normal et élevé. Ils ont été conçus pour fonctionner dans les limites des contraintes fixes ou des quotas de service de cette ressource.

Anti-modèles courants :

- Choix d'une conception qui utilise une ressource d'un service, sans savoir qu'il existe des contraintes de conception qui entraîneront l'échec de cette conception au fil des mises à l'échelle.
- Effectuer une évaluation comparative qui n'est pas réaliste et qui atteindra les quotas fixés par le service pendant les tests. Par exemple, l'exécution de tests à une limite de débordement mais pendant une durée prolongée.
- Le choix d'une conception qui ne peut pas évoluer ou être modifiée si des quotas de service fixes doivent être dépassés. Par exemple, une taille de charge utile SQS de 256 KB.
- L'observabilité n'a pas été conçue et mise en œuvre pour surveiller et alerter sur les seuils des quotas de service qui pourraient être compromis lors d'événements à fort trafic

Avantages liés au respect de cette bonne pratique : vérifier que l'application fonctionnera sous tous les niveaux de charge des services projetés sans perturbation ni dégradation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Contrairement aux quotas de services souples ou aux ressources qui peuvent être remplacées par des unités de plus grande capacité, les quotas fixes des services AWS ne peuvent pas être modifiés. Cela signifie que tous ces types de services AWS doivent être évalués en fonction des limites potentielles de capacité matérielle lorsqu'ils sont utilisés dans la conception d'une application.

Les limites strictes sont affichées dans la console Service Quotas. Si les colonnes affichent la valeur `ADJUSTABLE = No`, alors le service comporte une limite stricte. Des limites strictes sont également indiquées dans les pages de configuration de certaines ressources. Par exemple, Lambda possède des limites strictes spécifiques qui ne peuvent pas être ajustées.

À titre d'exemple, lors de la conception d'une application python destinée à être exécutée dans une fonction Lambda, l'application doit être évaluée pour déterminer si Lambda risque de s'exécuter pendant plus de 15 minutes. Si le code peut fonctionner au-delà de cette limite de quota de service, il faut envisager d'autres technologies ou conceptions. Si cette limite est atteinte après le déploiement de la production, l'application subira une dégradation et des perturbations jusqu'à ce qu'il soit

possible d'y remédier. Contrairement aux quotas souples, il n'existe aucune méthode permettant de passer à ces limites, même en cas d'événements d'urgence de gravité 1.

Une fois que l'application a été déployée dans un environnement de test, il convient d'utiliser des stratégies pour déterminer si des limites strictes peuvent être atteintes. Les tests de résistance, les tests de charge et les tests de chaos doivent faire partie du plan de test d'introduction.

Étapes d'implémentation

- Examinez la liste complète des services AWS qui pourraient être utilisés dans la phase de conception de l'application.
- Examinez les limites de quota logiciel et de quota matériel pour tous ces services. Toutes les limites ne sont pas affichées dans la console Service Quotas. Certains services détaillent ces limites à d'autres endroits.
- Lors de la conception de votre application, examinez les facteurs opérationnels et technologiques de votre charge de travail, tels que les résultats opérationnels, le cas d'utilisation, les systèmes dépendants, les objectifs de disponibilité et les objets de reprise après sinistre. Laissez vos facteurs commerciaux et technologiques guider le processus d'identification du système distribué qui convient à votre charge de travail.
- Analysez la charge de service dans les régions et les comptes. De nombreuses limites strictes se basent sur la région pour les services. Cependant, certaines limites sont basées sur le compte.
- Analysez les architectures de résilience pour l'utilisation des ressources lors d'une panne de zone et d'une panne régionale. Dans la progression des conceptions multirégionales utilisant des approches actives/actives, actives/passives – à chaud, actives/passives – à froid, et actives/passives – environnement de veille, ces cas de panne entraîneront une utilisation plus importante. Cela crée un cas d'utilisation potentiel pour atteindre des limites strictes.

Ressources

Bonnes pratiques associées :

- REL01-BP01 Connaissance des quotas de service et des contraintes
- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP04 Surveiller et gérer les quotas
- REL01-BP05 Automatiser la gestion des quotas
- REL01-BP06 Garantir un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement

- REL03-BP01 Choisir comment segmenter votre charge de travail
- REL10-BP01 Déployer la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la réparation sur toutes les couches
- REL12-BP05 Tester la résilience à l'aide de l'ingénierie du chaos

Documents connexes :

- Pilier Fiabilité du cadre AWS Well-Architected : Disponibilité
- AWS Service Quotas (anciennement appelés limites de service)
- Vérifications des bonnes pratiques AWS Trusted Advisor (voir la section Limites de service)
- AWS limit monitor on AWS answers (Surveillance de limites AWS sur les réponses AWS)
- Amazon EC2 Service Limits (Limites de service EC2)
- Qu'est-ce qu'Service Quotas ?
- How to Request quota increase (Comment demander une augmentation du quota)
- Service endpoints and quotas (Points de terminaison et quotas de service)
- Guide de l'utilisateur Service Quotas
- Quota Monitor for AWS (Surveillance de quotas pour AWS)
- AWS Fault Isolation Boundaries (Limites d'isolement des pannes AWS)
- Availability with redundancy (Disponibilité avec redondance)
- AWS pour les données
- Qu'est-ce que l'intégration continue ?
- Qu'est-ce que la livraison continue ?
- Partenaire APN : partenaires facilitant la gestion de la configuration
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS (Gestion du cycle de vie des comptes dans les environnements SaaS de type compte par locataire sur AWS)
- Gestion et surveillance de la limitation des API dans vos charges de travail
- View AWS Trusted Advisor recommendations at scale with AWS Organizations (Examiner les recommandations AWS Trusted Advisor à grande échelle avec AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower (Automatisation de l'augmentation des limites de service et Enterprise Support avec AWS Control Tower)

- Actions, ressources et clés de condition pour les Service Quotas

Vidéos connexes :

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Examiner et gérer les quotas pour les services AWS avec les quotas de service)
- AWS IAM Quotas Demo (Démonstration sur les quotas IAM d'AWS)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small

Outils associés :

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP04 Surveiller et gérer les quotas

Évaluez votre utilisation potentielle et augmentez vos quotas de manière appropriée afin d'assurer une croissance planifiée de l'utilisation.

Résultat souhaité : des systèmes actifs et automatisés de gestion et de suivi ont été déployés. Ces solutions opérationnelles permettent de s'assurer que les seuils d'utilisation des quotas sont sur le point d'être atteints. Les changements de quotas demandés permettraient de remédier à ces problèmes de manière proactive.

Anti-modèles courants :

- Ne pas configurer la surveillance pour vérifier les seuils de quota de service
- Ne pas configurer la surveillance pour les limites strictes, même si ces valeurs ne peuvent pas être modifiées.
- En supposant que le délai nécessaire pour demander et obtenir un changement de quota souple soit immédiat ou de courte durée.
- Configuration d'alarmes d'approche des quotas de service, mais sans processus sur la façon de répondre à une alerte.
- Configuration d'alarmes uniquement pour les services pris en charge par les AWS Service Quotas, sans surveiller les autres services AWS.
- Ne pas prendre en compte la gestion des quotas pour les conceptions de résilience à régions multiples, comme les approches actives/actives, actives/passives – à chaud, actives/passives – à froid, et actives/passives – environnement de veille.
- Ne pas évaluer les différences de quotas entre les régions.
- Ne pas évaluer les besoins de chaque région pour une demande spécifique d'augmentation de quota.
- Ne pas utiliser les modèles pour la gestion des quotas multirégionaux.

Avantages liés au respect de cette bonne pratique : le suivi automatique des Service Quotas AWS et la surveillance de votre utilisation par rapport à ces quotas vous permettront de voir quand vous approchez de la limite du quota. Vous pouvez également utiliser ces données de surveillance pour limiter les dégradations dues à l'épuisement des quotas.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Pour les services pris en charge, vous pouvez surveiller vos quotas en configurant différents services qui peuvent évaluer et ensuite envoyer des alertes ou des alarmes. Cela peut aider à surveiller l'utilisation et vous alerter sur l'approche des quotas. Ces alarmes peuvent être déclenchées à partir de AWS Config, de fonctions Lambda, de Amazon CloudWatch, ou de AWS Trusted Advisor. Vous pouvez également utiliser des filtres de métriques sur les journaux CloudWatch pour rechercher et extraire des modèles dans les journaux afin de déterminer si l'utilisation approche des seuils de quota.

Étapes d'implémentation

Pour la surveillance :

- Enregistrez la consommation des ressources actuelles (par exemple, les compartiments, ou les instances). Utilisez les opérations de l'API de service, telles que l'API Amazon EC2, `DescribeInstances` pour recueillir la consommation actuelle des ressources.
- Saisissez vos quotas actuels qui sont essentiels et applicables aux services utilisés :
  - AWS Service Quotas
  - AWS Trusted Advisor
  - documentation AWS
  - Pages spécifiques aux services AWS
  - AWS Command Line Interface (AWS CLI)
  - AWS Cloud Development Kit (AWS CDK)
- Utilisez AWS Service Quotas, un service AWS qui vous aide à gérer vos quotas pour plus de 250 services AWS à partir d'un seul emplacement.
- Utilisez les limites de service Trusted Advisor pour surveiller vos limites de service actuelles à différents seuils.
- Utilisez l'historique des quotas de service (console ou AWS CLI) pour vérifier les augmentations régionales.
- Comparez les changements de quotas de service dans chaque région et chaque compte pour créer une équivalence, si nécessaire.

Pour la gestion :

- Automatisé : configurez une règle AWS Config personnalisée pour analyser les quotas de service dans les régions et comparer les différences.
- Automatisé : configurez une fonction programmée Lambda pour analyser les quotas de service dans les régions et comparer les différences.
- Manuel : analysez les quotas de services par le biais de la AWS CLI, d'API ou de la console AWS pour analyser les quotas de services dans les régions et comparer les différences. Signalez les différences.
- Si des différences de quotas sont identifiées entre les régions, demandez un changement de quota, si nécessaire.
- Passez en revue le résultat de toutes les demandes.

Ressources

Bonnes pratiques associées :

- REL01-BP01 Connaissance des quotas de service et des contraintes
- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP03 Tenir compte des quotas et des contraintes de service fixes dans l'architecture
- REL01-BP05 Automatiser la gestion des quotas
- REL01-BP06 Garantir un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement
- REL03-BP01 Choisir comment segmenter votre charge de travail
- REL10-BP01 Déployer la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la réparation sur toutes les couches
- REL12-BP05 Tester la résilience à l'aide de l'ingénierie du chaos

Documents connexes :

- Pilier Fiabilité du cadre AWS Well-Architected : Disponibilité
- AWS Service Quotas (anciennement appelés limites de service)
- Vérifications des bonnes pratiques AWS Trusted Advisor (voir la section Limites de service)
- AWS limit monitor on AWS answers (Surveillance de limites AWS sur les réponses AWS)
- Amazon EC2 Service Limits (Limites de service EC2)
- Qu'est-ce qu'Service Quotas ?
- How to Request quota increase (Comment demander une augmentation du quota)
- Service endpoints and quotas (Points de terminaison et quotas de service)
- Guide de l'utilisateur Service Quotas
- Quota Monitor for AWS (Surveillance de quotas pour AWS)
- AWS Fault Isolation Boundaries (Limites d'isolement des pannes AWS)
- Availability with redundancy (Disponibilité avec redondance)
- AWS pour les données
- Qu'est-ce que l'intégration continue ?

- Qu'est-ce que la livraison continue ?
- Partenaire APN : partenaires facilitant la gestion de la configuration
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS (Gestion du cycle de vie des comptes dans les environnements SaaS de type compte par locataire sur AWS)
- Gestion et surveillance de la limitation des API dans vos charges de travail
- View AWS Trusted Advisor recommendations at scale with AWS Organizations (Examiner les recommandations AWS Trusted Advisor à grande échelle avec AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower (Automatisation de l'augmentation des limites de service et Enterprise Support avec AWS Control Tower)
- Actions, ressources et clés de condition pour les Service Quotas

Vidéos connexes :

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Examiner et gérer les quotas pour les services AWS avec les quotas de service)
- AWS IAM Quotas Demo (Démonstration sur les quotas IAM d'AWS)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small

Outils associés :

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager

- AWS Marketplace

REL01-BP05 Automatiser la gestion des quotas

Mettez en place des outils pour être informé à lorsque l'atteinte des seuils est proche. Vous pouvez automatiser les demandes d'augmentation de quota à l'aide des API AWS Service Quotas.

Si vous intégrez votre base de données de gestion de configuration (CMDB) ou votre système de tickets avec Service Quotas, vous pouvez automatiser le suivi des requêtes d'augmentation de quotas et des quotas actuels. En plus du kit SDK AWS, Service Quotas propose une automatisation avec AWS Command Line Interface (AWS CLI).

Anti-modèles courants :

- Suivi des quotas et de l'utilisation dans les feuilles de calcul.
- Exécution de rapports sur l'utilisation quotidienne, hebdomadaire ou mensuelle, puis comparaison de l'utilisation aux quotas.

Avantages liés au respect de cette bonne pratique : Le suivi automatisé des quotas de service AWS et la surveillance de votre utilisation par rapport à ce quota vous permettent de voir quand vous vous rapprochez d'un quota. Vous pouvez configurer l'automatisation pour vous aider à demander une augmentation de quota si nécessaire. Vous pouvez envisager de réduire certains quotas lorsque votre utilisation évolue dans le sens inverse pour profiter des avantages d'une réduction des risques (en cas d'informations d'identification corrompues) et des économies de coûts.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Configurer la surveillance automatisée : mettez en place des outils à l'aide de kits SDK pour être informé lorsque des seuils sont sur le point d'être atteints.
  - Utilisez les Service Quotas et complétez le service avec une solution automatisée de surveillance des quotas, tels qu'AWS Limit Monitor ou une offre sur AWS Marketplace.
    - Qu'est-ce que Service Quotas ?
    - Surveillance des quotas sur AWS - Solution AWS
  - Configurez des réponses déclenchées en fonction des seuils de quota, à l'aide des API Amazon SNS et AWS Service Quotas .

- Testez l'automatisation.
  - Configurez les seuils de limites.
  - Intégrez les événements de modification provenant d'AWS Config, des pipelines de déploiement, d'Amazon EventBridge ou de tiers.
  - Définissez des seuils de limites artificiellement bas pour tester les réponses.
  - Configurez des déclencheurs pour prendre les mesures appropriées en cas de notifications et contactez AWS Support, le cas échéant
  - Déclenchez manuellement les événements de modifications.
  - Organisez un jeu de rôle pour tester le processus d'augmentation des quotas.

Ressources

Documents connexes :

- Partenaire APN : partenaires facilitant la gestion de la configuration
- AWS Marketplace : produits CMDB facilitant le suivi des limites
- AWS Service Quotas (anciennement appelés Service Limits)
- Vérifications des bonnes pratiques AWS Trusted Advisor (voir la section Service Limits)
- Surveillance des quotas sur AWS - Solution AWS
- Amazon EC2 Service Limits
- Qu'est-ce que Service Quotas ?

Vidéos connexes :

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP06 Garantir un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement

En cas de panne ou d'inaccessibilité d'une ressource, celle-ci peut être comptabilisée dans un quota jusqu'à ce qu'elle soit correctement terminée. Vérifiez que vos quotas couvrent le chevauchement des ressources défaillantes ou inaccessibles et de leurs remplacements. Vous devez prendre en compte les cas d'utilisation tels que les pannes de réseau, la panne de la zone de disponibilité ou les pannes régionales lorsque vous calculez cet écart.

Résultat souhaité : les défaillances, petites ou grandes, des ressources ou de leur accessibilité peuvent être gérées par les seuils de service actuels. Les pannes de zone, les pannes de réseau, voire les pannes régionales ont été prises en compte dans la planification des ressources.

Anti-modèles courants :

- Définition de quotas de service en fonction des quotas actuels sans tenir compte des scénarios de basculement.
- Ne pas tenir compte des principes de stabilité statique lors du calcul du quota de pointe pour un service.
- Ne pas tenir compte du potentiel des ressources inaccessibles dans le calcul du quota total nécessaire pour chaque région.
- Ne pas prendre en compte les limites d'isolement des pannes de service AWS pour certains services et leurs potentiels schémas d'utilisation anormaux.

Avantages liés au respect de cette bonne pratique : lorsqu'une interruption de service a un impact sur la disponibilité des applications, le cloud vous permet de mettre en œuvre des stratégies pour atténuer ou récupérer ces événements. Ces stratégies incluent souvent la création de ressources supplémentaires pour remplacer celles qui ont échoué ou celles qui sont inaccessibles. Votre stratégie de quotas devrait tenir compte de ces conditions de basculement et ne pas ajouter de dégradations supplémentaires dues à l'épuisement des limites de service.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Lors de l'évaluation des limites de quotas, il faut tenir compte des cas de basculement qui pourraient survenir en raison d'une certaine dégradation. Les types de basculement suivants doivent être pris en compte :

- Un VPC perturbé ou inaccessible.
- Un sous-réseau inaccessible.
- Une zone de disponibilité suffisamment dégradée pour avoir un impact sur l'accessibilité de nombreuses ressources.
- Divers itinéraires de mise en réseau ou points d'entrée et de sortie sont bloqués ou modifiés.
- Une région a été suffisamment dégradée pour avoir un impact sur l'accessibilité de nombreuses ressources.

- Il existe plusieurs ressources, mais toutes ne sont pas affectées par une panne dans une région ou une zone de disponibilité.

Des pannes comme celles énumérées ci-dessus peuvent être le déclencheur d'un événement de basculement. La décision de basculement est unique selon la situation et le client, car l'impact sur l'entreprise peut varier considérablement. Toutefois, lorsque l'on décide, sur le plan opérationnel, de basculer une application ou des services, la planification de la capacité des ressources dans l'emplacement de basculement et les quotas correspondants doivent être définis avant l'événement.

Passez en revue les quotas de service pour chaque service en tenant compte des pics supérieurs à la normale qui pourraient se produire. Ces pics peuvent être liés à des ressources qui ne peuvent être atteintes en raison de la mise en réseau ou des autorisations, mais qui sont toujours actives. Les ressources actives non résiliées seront toujours comptabilisées dans la limite du quota de service.

Étapes d'implémentation

- Vérifiez que l'écart entre votre quota de service et votre utilisation maximale est suffisant pour faire face à un basculement ou à une perte d'accessibilité.
- Identifiez les quotas de service en tenant compte de vos modèles de déploiement, de vos exigences en matière de disponibilité et de la croissance de votre consommation.
- Demandez des augmentations de quota si nécessaire Prévoyez le temps nécessaire pour l'approbation des demandes d'augmentation des quotas.
- Déterminez vos exigences de fiabilité (également connues sous le nom de « nombre de neuf »).
- Définissez vos scénarios de défaillance (par exemple, perte d'un composant, d'une zone de disponibilité ou d'une région).
- Définissez votre méthodologie de déploiement (par exemple, canary, bleu/vert, rouge/noir ou par propagation).
- Ajoutez une mémoire tampon appropriée (par exemple, 15 %) à la limite actuelle.
- Ajoutez les calculs de stabilité statique (zonale et régionale), le cas échéant.
- Anticipez la croissance de la consommation (par exemple, surveillance de vos tendances de consommation).
- Songez à l'impact de la stabilité statique pour vos charges de travail les plus critiques. Évaluez les ressources conformes à un système statiquement stable dans toutes les régions et zones de disponibilité.

- Envisagez l'utilisation de réservations de capacité à la demande pour programmer la capacité avant tout basculement. Cette stratégie peut s'avérer utile lors des calendriers d'activité les plus critiques afin de réduire les risques potentiels liés à l'obtention de la bonne quantité et du bon type de ressources lors du basculement.

Ressources

Bonnes pratiques associées :

- REL01-BP01 Connaissance des quotas de service et des contraintes
- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP03 Tenir compte des quotas et des contraintes de service fixes dans l'architecture
- REL01-BP04 Surveiller et gérer les quotas
- REL01-BP05 Automatiser la gestion des quotas
- REL03-BP01 Choisir comment segmenter votre charge de travail
- REL10-BP01 Déployer la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la réparation sur toutes les couches
- REL12-BP05 Tester la résilience à l'aide de l'ingénierie du chaos

Documents connexes :

- Pilier Fiabilité du cadre AWS Well-Architected : Disponibilité
- AWS Service Quotas (anciennement appelés limites de service)
- Vérifications des bonnes pratiques AWS Trusted Advisor (voir la section Limites de service)
- AWS limit monitor on AWS answers (Surveillance de limites AWS sur les réponses AWS)
- Amazon EC2 Service Limits (Limites de service EC2)
- Qu'est-ce qu'Service Quotas ?
- How to Request quota increase (Comment demander une augmentation du quota)
- Service endpoints and quotas (Points de terminaison et quotas de service)
- Guide de l'utilisateur Service Quotas
- Quota Monitor for AWS (Surveillance de quotas pour AWS)

- AWS Fault Isolation Boundaries (Limites d'isolement des pannes AWS)
- Availability with redundancy (Disponibilité avec redondance)
- AWS pour les données
- Qu'est-ce que l'intégration continue ?
- Qu'est-ce que la livraison continue ?
- Partenaire APN : partenaires facilitant la gestion de la configuration
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS (Gestion du cycle de vie des comptes dans les environnements SaaS de type compte par locataire sur AWS)
- Gestion et surveillance de la limitation des API dans vos charges de travail
- View AWS Trusted Advisor recommendations at scale with AWS Organizations (Examiner les recommandations AWS Trusted Advisor à grande échelle avec AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower (Automatisation de l'augmentation des limites de service et Enterprise Support avec AWS Control Tower)
- Actions, ressources et clés de condition pour les Service Quotas

Vidéos connexes :

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Examiner et gérer les quotas pour les services AWS avec les quotas de service)
- AWS IAM Quotas Demo (Démonstration sur les quotas IAM d'AWS)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small

Outils associés :

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config

- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

FIA 2. Comment planifiez-vous la topologie de votre réseau ?

Les charges de travail existent souvent dans plusieurs environnements. Il s'agit notamment de plusieurs environnements cloud (accessibles publiquement et privés) et éventuellement de votre infrastructure de centre de données existante. Les plans doivent inclure des considérations réseau telles que la connectivité intrasystème et intersystème, la gestion des adresses IP publiques, la gestion des adresses IP privées et la résolution des noms de domaine.

Bonnes pratiques

- REL02-BP01 Utiliser une connectivité réseau hautement disponible pour vos points de terminaison publics de charge de travail
- REL02-BP02 Mettre en service une connectivité redondante entre les réseaux privés dans le cloud et les environnements sur site
- REL02-BP03 S'assurer que l'allocation des sous-réseaux IP tient compte de l'expansion et de la disponibilité
- REL02-BP04 Préférer les topologies en étoile au maillage « many-to-many »
- REL02-BP05 Appliquer des plages d'adresses IP privées sans chevauchement dans tous les espaces d'adressage privés où ils sont connectés

REL02-BP01 Utiliser une connectivité réseau hautement disponible pour vos points de terminaison publics de charge de travail

La mise en place d'une connectivité réseau hautement disponible aux points de terminaison publics de vos charges de travail peut vous aider à réduire les temps d'arrêt dus à la perte de connectivité et à améliorer la disponibilité et le SLA de votre charge de travail. Pour ce faire, utilisez le DNS hautement disponible, les réseaux de diffusion de contenu (CDN), des passerelles API, l'équilibrage de charge ou les proxys inverses.

Résultat souhaité : il est essentiel de planifier, de construire et de rendre opérationnelle une connectivité réseau hautement disponible pour vos points de terminaison publics. Si votre charge

de travail devient inaccessible en raison d'une perte de connectivité, même si elle est en cours d'exécution et disponible, vos clients verront votre système comme étant en panne. En combinant une connectivité réseau hautement disponible et résiliente pour les points de terminaison publics de votre charge de travail, ainsi qu'une architecture résiliente pour votre charge de travail elle-même, vous pouvez offrir la meilleure disponibilité et le meilleur niveau de service possible à vos clients.

Les services AWS Global Accelerator, Amazon CloudFront, Amazon API Gateway, les URL de fonction AWS Lambda, les API AWS AppSync et Elastic Load Balancing (ELB) fournissent tous des points de terminaison publics hautement disponibles. Amazon Route 53 fournit un service DNS hautement disponible pour la résolution des noms de domaine afin de vérifier que les adresses de vos points de terminaison publics peuvent être résolues.

Vous pouvez également évaluer des appliances logicielles AWS Marketplace pour l'équilibrage de charge et les proxys.

Anti-modèles courants :

- Concevoir une charge de travail hautement disponible sans planifier le DNS et la connectivité réseau pour la haute disponibilité.
- Utilisation d'adresses Internet publiques sur des instances ou des conteneurs individuels et gestion de la connectivité à ces adresses avec le DNS.
- Utilisation des adresses IP au lieu des noms de domaine pour localiser les services.
- Ne pas tester des scénarios où la connectivité à vos points de terminaison publics est perdue.
- Ne pas analyser les besoins en débit du réseau et les modèles de distribution.
- Ne pas tester et planifier des scénarios dans lesquels la connectivité du réseau Internet à vos points de terminaison publics de votre charge de travail pourrait être interrompue.
- Fourniture du contenu (comme les pages web, les ressources statiques ou les fichiers multimédias) à une grande zone géographique sans utilisation d'un réseau de diffusion de contenu.
- Ne pas se préparer aux attaques par déni de service distribué (DDoS). Les attaques DDoS risquent d'interrompre le trafic légitime et de réduire la disponibilité pour vos utilisateurs.

Avantages liés au respect de cette bonne pratique : la conception d'une connectivité réseau hautement disponible et résiliente garantit que votre charge de travail est accessible et disponible pour vos utilisateurs.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Le routage du trafic est au cœur de la mise en place d'une connectivité réseau hautement disponible pour vos points de terminaison publics. Pour vérifier que votre trafic est en mesure d'atteindre les points de terminaison, le DNS doit être capable de résoudre les noms de domaine à leurs adresses IP correspondantes. Utilisez un système de nom de domaine (DNS) hautement disponible et évolutif tel que Amazon Route 53 pour gérer les enregistrements DNS de votre domaine. Vous pouvez également utiliser les surveillances de l'état fournies par Amazon Route 53. Les surveillances de l'état permettent de s'assurer que votre application est accessible, disponible et fonctionnelle. Elles peuvent être configurées de manière à imiter le comportement de l'utilisateur, comme la demande d'une page web ou d'une URL spécifique. En cas de panne, Amazon Route 53 répond aux demandes de résolution DNS et dirige uniquement le trafic vers les points de terminaison en bonne santé. Vous pouvez également envisager d'utiliser les fonctionnalités de Geo DNS et de routage basé sur la latence offertes par Amazon Route 53.

Pour vérifier que votre charge de travail elle-même est hautement disponible, utilisez Elastic Load Balancing (ELB).Amazon Route 53 peut cibler le trafic vers ELB, qui distribue le trafic vers les instances de calcul cibles. Vous pouvez également utiliser Amazon API Gateway avec AWS Lambda pour une solution sans serveur. Les clients peuvent également exécuter des charges de travail dans plusieurs Régions AWS. Avec le modèle multisite actif/actif, la charge de travail peut servir le trafic de plusieurs régions. Avec un modèle multisite actif/passif, la charge de travail sert le trafic de la région active tandis que les données sont répliquées dans la région secondaire et deviennent actives en cas de panne de la région principale. Les surveillances de l'état Route 53 peuvent alors contrôler le basculement DNS de n'importe quel point de terminaison dans une région principale vers un point de terminaison dans une région secondaire, vérifiant que votre charge de travail est accessible et disponible pour vos utilisateurs.

Amazon CloudFront fournit une API simple pour distribuer du contenu avec une faible latence et des taux de transfert de données élevés en répondant aux demandes à l'aide d'un réseau d'emplacements périphériques dans le monde entier. Les réseaux de diffusion de contenu (CDN) servent les clients en proposant un contenu situé ou mis en cache à un endroit proche de l'utilisateur. La disponibilité de votre application s'en trouve également améliorée, car la charge de contenu est déplacée de vos serveurs vers lesemplacements périphériques de CloudFront. Les emplacements périphériques et les caches périphériques régionaux conservent des copies en cache de votre contenu à proximité de vos utilisateurs, ce qui permet une récupération rapide et augmente l'accessibilité et la disponibilité de votre charge de travail.

Pour les charges de travail avec des utilisateurs dispersés géographiquement, AWS Global Accelerator améliore la disponibilité et les performances des applications. AWS Global Accelerator fournit des adresses IP statiques Anycast qui servent de point d'entrée fixe à votre application hébergée dans une ou plusieurs Régions AWS. Cela permet au trafic d'entrer sur le réseau mondial AWS aussi près que possible de vos utilisateurs, améliorant ainsi l'accessibilité et la disponibilité de votre charge de travail. AWS Global Accelerator surveille également l'état de santé des points de terminaison de vos applications en utilisant la surveillance de l'état TCP, HTTP et HTTPS. Toute modification de l'état ou de la configuration de vos points de terminaison déclenche la redirection du trafic utilisateur vers des points de terminaison sains qui offrent les meilleures performances et la meilleure disponibilité à vos utilisateurs. De plus, AWS Global Accelerator est conçu pour être isolé des pannes et utilise deux adresses IPv4 statiques qui sont desservies par des zones réseau indépendantes, ce qui augmente la disponibilité de vos applications.

Pour aider à protéger les clients contre les attaques DDoS, AWS propose AWS Shield Standard. Shield Standard est automatiquement activé et protège contre les attaques d'infrastructure courantes (couches 3 et 4) telles que les inondations SYN/UDP et les attaques par réflexion pour assurer la haute disponibilité de vos applications sur AWS. Pour bénéficier de protections supplémentaires contre des attaques plus sophistiquées et plus importantes (comme les inondations UDP), les attaques par épuisement d'état (comme les inondations TCP SYN), et pour aider à protéger vos applications fonctionnant sur Amazon Elastic Compute Cloud (Amazon EC2), Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator, et Route 53, envisagez d'utiliser AWS Shield Advanced. Pour se protéger contre les attaques au niveau de la couche application, comme les inondations HTTP POST ou GET, utilisez AWS WAF. AWS WAF peut utiliser les adresses IP, les en-têtes HTTP, le corps HTTP, les chaînes URI, l'injection SQL et les conditions de script intersite pour déterminer si une requête doit être bloquée ou autorisée.

Étapes d'implémentation

1. Configurez un DNS hautement disponible : Amazon Route 53 est un service web de système de nom de domaine (DNS) hautement disponible et évolutif. Route 53 connecte les demandes des utilisateurs aux applications Internet fonctionnant sur AWS ou sur site. Pour obtenir plus d'informations, consultez Configuration d'Amazon Route 53 en tant que service DNS.
2. Configurez la surveillance de l'état : lorsque vous utilisez Route 53, vérifiez que seules les cibles saines sont résolubles. Commencez par créer des surveillances de l'état Route 53 et par configurer le basculement DNS. Il est important de tenir compte des aspects suivants lors de la mise en place des surveillances de l'état :
   a. Comment Amazon Route 53 détermine si une vérification de l'état est saine

   b. Création, mise à jour et suppression de vérifications de l'état
   c. Statut de la surveillance de l'état et réception de notifications
   d. Bonnes pratiques relatives à Amazon Route 53 DNS
3. Connectez votre service DNS à vos points de terminaison.
   a. Lorsque vous utilisez Elastic Load Balancing comme cible pour votre trafic, créez un enregistrement d'alias utilisant Amazon Route 53 qui pointe vers le point de terminaison régional de votre équilibreur de charge. Pendant la création de l'enregistrement de l'alias, réglez l'option « Évaluer l'état de la cible » sur « Oui ».
   b. Pour les charges de travail sans serveur ou les API privées, lorsque vous utilisez API Gateway, utilisez Route 53 pour router le trafic vers API Gateway.
4. Choisissez un réseau de diffusion de contenu.
   a. Pour diffuser du contenu en utilisant des emplacements périphériques plus proches de l'utilisateur, il faut commencer par comprendre comment CloudFront diffuse le contenu.
   b. Commencez par une simple distribution CloudFront. CloudFront comprend alors l'endroit d'où vous souhaitez que le contenu soit diffusé, ainsi que les détails concernant le suivi et la gestion de la diffusion du contenu. Il est important de comprendre et de prendre en compte les aspects suivants lors de la mise en place de la distribution CloudFront :
      i. Fonctionnement de la mise en cache avec les emplacements périphériques CloudFront
      ii. Augmentation de la proportion de demandes servies directement à partir des caches CloudFront (taux d'accès au cache)
      iii. Utilisation Amazon CloudFront Origin Shield
      iv. Optimisation de la haute disponibilité avec le basculement d'origine CloudFront
5. Configurez la protection de la couche d'application : AWS WAF vous aide à vous protéger contre les exploits et les bots web courants qui peuvent affecter la disponibilité, compromettre la sécurité ou consommer des ressources excessives. Pour mieux comprendre, examinez comment AWS WAF fonctionne et quand vous êtes prêt à mettre en œuvre les protections de la couche application HTTP POST ET GET, consultez Getting started with AWS WAF (Démarrer avec AWS WAF). Vous pouvez également utiliser AWS WAF avec CloudFront. Consultez la documentation pour comprendre comment AWS WAF fonctionne avec les fonctionnalités Amazon CloudFront.
6. Configurez une protection DDoS supplémentaire : par défaut, tous les clients AWS bénéficient d'une protection contre les attaques DDoS les plus fréquentes au niveau de la couche réseau et de la couche transport qui ciblent votre site web ou votre application avec AWS Shield Standard, et ce sans frais supplémentaires. Pour bénéficier d'une protection supplémentaire des applications

accessibles sur Internet et fonctionnant sur Amazon EC2, Elastic Load Balancing, Amazon CloudFront, AWS Global Accelerator, et Amazon Route 53, vous pouvez envisager AWS Shield Advanced et passer en revue des exemples d'architectures résistantes aux attaques DDoS. Pour protéger votre charge de travail et vos points de terminaison publics contre les attaques DDoS, consultez Getting started with AWS Shield Advanced (Démarrer avec AWS Shield Advanced).

Ressources

Bonnes pratiques associées :

- REL10-BP01 Déployer la charge de travail sur plusieurs emplacements
- REL10-BP02 Sélectionner les emplacements appropriés pour votre déploiement multisite
- REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle pendant la récupération
- REL11-BP06 Envoyer des notifications lorsque des événements affectent la disponibilité

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à planifier votre mise en réseau
- AWS Marketplace pour l'infrastructure réseau
- Qu'est-ce que AWS Global Accelerator ?
- Qu'est-ce qu'Amazon CloudFront ?
- Qu'est-ce qu'Amazon Route 53 ?
- Qu'est-ce qu'Elastic Load Balancing ?
- Capacité de connectivité du réseau : établir les fondements de votre cloud
- Qu'est-ce que Amazon API Gateway ?
- Que sont AWS WAF, AWS Shield, et AWS Firewall Manager ?
- What is Amazon Route 53 Application Recovery Controller? (Qu'est-ce que le contrôleur de récupération d'application d'Amazon Route 53 ?)
- Configurer des surveillances de l'état personnalisées pour le basculement DNS

Vidéos connexes :

- AWS re:Invent 2022 - Improve performance and availability with AWS Global Accelerator

- AWS re:Invent 2020: Global traffic management with Amazon Route 53
- AWS re:Invent 2022 - Operating highly available Multi-AZ applications
- AWS re:Invent 2022 - Dive deep on AWS networking infrastructure
- AWS re:Invent 2022 - Building resilient networks

Exemples connexes :

- Disaster Recovery with Amazon Route 53 Application Recovery Controller (ARC) [Reprise après sinistre avec le contrôleur de récupération d'application (ARC) d'Amazon Route 53]
- Ateliers sur la fiabilité
- Atelier AWS Global Accelerator

REL02-BP02 Mettre en service une connectivité redondante entre les réseaux privés dans le cloud et les environnements sur site

Utilisez plusieurs connexions AWS Direct Connect ou tunnels VPN entre des réseaux privés déployés séparément. Utilisez plusieurs emplacements Direct Connect pour une plus haute disponibilité. Si vous utilisez plusieurs Régions AWS, assurez la redondance dans au moins deux d'entre elles.
Il serait souhaitable d'évaluer les appliances AWS Marketplace qui mettent fin aux VPN. Si vous utilisez des appliances AWS Marketplace, déployez des instances redondantes pour une plus haute disponibilité dans différentes zones de disponibilité.

AWS Direct Connect est un service cloud qui permet d'établir facilement une connexion réseau dédiée depuis votre environnement sur site vers AWS. Grâce à la passerelle Direct Connect, votre centre de données sur site peut être connecté à plusieurs VPC AWS répartis sur plusieurs Régions AWS.

Cette redondance permet de faire face aux pannes possibles ayant un impact sur la résilience de la connectivité :

- Comment allez-vous résister aux pannes dans votre topologie ?
- Que se passe-t-il si un composant est mal configuré et perd sa connectivité ?
- Pourrez-vous gérer une augmentation inattendue du trafic ou de l'utilisation de vos services ?
- Serez-vous en mesure d'absorber une tentative d'attaque par déni de service distribué ?

Lors de la connexion de votre VPC à votre centre de données sur site via un VPN, tenez compte des exigences en matière de résilience et de bande passante lorsque vous sélectionnez le fournisseur et la taille d'instance dont vous avez besoin pour exécuter l'appliance. Si vous utilisez une appliance VPN qui n'est pas résiliente dans son implémentation, vous devez avoir une connexion redondante via une seconde appliance. Pour tous ces scénarios, vous devez définir les temps de récupération acceptables et faire des tests pour vous assurer que vous pouvez satisfaire ces exigences.

Si vous choisissez de connecter votre VPC à votre centre de données à l'aide d'une connexion Direct Connect et que vous avez besoin que cette connexion soit hautement disponible, vous devez disposer de connexions Direct Connect redondantes à partir de chaque centre de données. La connexion redondante doit utiliser une deuxième connexion Direct Connect à partir d'un emplacement différent de la première. Si vous avez plusieurs centres de données, assurez-vous que les connexions se terminent à différents emplacements. Utilisez la boîte à outils Direct Connect Resiliency Toolkit pour faciliter la mise en œuvre de cette configuration.

Si vous choisissez de basculer sur un VPN via Internet à l'aide d'un AWS VPN, il est important de comprendre qu'il prend en charge jusqu'à 1,25 Gbit/s de débit par tunnel VPN, mais pas le protocole ECMP (Equal Cost Multi Path) pour le trafic sortant dans le cas de plusieurs tunnels VPN gérés par AWS aboutissant à la même passerelle virtuelle. Nous vous déconseillons d'utiliser un VPN géré par AWS comme système de secours pour les connexions Direct Connect, sauf si vous pouvez tolérer des vitesses inférieures à 1 Gbit/s lors du basculement.

Vous pouvez également utiliser les points de terminaison d'un VPC pour connecter de manière privée votre VPC aux services AWS pris en charge et aux services de point de terminaison d'un VPC à technologie AWS PrivateLink sans traverser l'Internet public. Les points de terminaison sont des dispositifs virtuels. Il s'agit de composants VPC mis à l’échelle horizontalement, redondants et hautement disponibles. Ils permettent la communication entre les instances de votre VPC et les services sans imposer de risques de disponibilité ou de contraintes de bande passante sur votre trafic réseau.

Anti-modèles courants :

- Un seul fournisseur de connectivité entre votre réseau sur site et AWS.
- Utilisation des capacités de connectivité de votre connexion AWS Direct Connect, mais en ayant qu'une seule connexion.
- Un seul chemin pour votre connectivité VPN.

Avantages liés au respect de cette bonne pratique : En implémentant une connectivité redondante entre votre environnement cloud et votre environnement d'entreprise/sur site, vous pouvez vous assurer que les services dépendants entre les deux environnements peuvent communiquer de manière fiable.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

- Assurez-vous que vous disposez d’une connectivité hautement disponible entre AWS et l'environnement sur site. Utilisez plusieurs connexions AWS Direct Connect ou tunnels VPN entre des réseaux privés déployés séparément. Utilisez plusieurs emplacements Direct Connect pour une plus haute disponibilité. Si vous utilisez plusieurs Régions AWS, assurez la redondance dans au moins deux d'entre elles. Il serait souhaitable d'évaluer les appliances AWS Marketplace qui mettent fin aux VPN. Si vous utilisez des appliances AWS Marketplace, déployez des instances redondantes pour une plus haute disponibilité dans différentes zones de disponibilité.
  - Assurez-vous que vous avez une connexion redondante à votre environnement sur site. Vous pouvez avoir besoin de connexions redondantes à plusieurs Régions AWS pour répondre à vos besoins de disponibilité.
    - Recommandations de résilience AWS Direct Connect
    - Utilisation de connexions VPN de site à site redondantes pour assurer le basculement
      - Utilisez des opérations d'API de service pour confirmer la bonne utilisation des circuits Direct Connect.
        - DescribeConnections
        - DescribeConnectionsOnInterconnect
        - DescribeDirectConnectGatewayAssociations
        - DescribeDirectConnectGatewayAttachments
        - DescribeDirectConnectGateways
        - DescribeHostedConnections
        - DescribeInterconnects
      - S'il existe une seule connexion Direct Connect ou si vous n'en avez aucune, configurez des tunnels VPN redondants vers vos passerelles réseau privées virtuelles.
        - Qu'est-ce qu'AWS Site-to-Site VPN ?
- Capturez votre connectivité actuelle (par exemple, Direct Connect, passerelles réseau privées virtuelles, appliances AWS Marketplace).

- Utilisez des opérations d'API de service pour interroger la configuration des connexions Direct Connect.
  - DescribeConnections
  - DescribeConnectionsOnInterconnect
  - DescribeDirectConnectGatewayAssociations
  - DescribeDirectConnectGatewayAttachments
  - DescribeDirectConnectGateways
  - DescribeHostedConnections
  - DescribeInterconnects
- Utilisez des opérations d'API de service pour collecter les passerelles réseau privées virtuelles là où les tables de routage les utilisent.
  - DescribeVpnGateways
  - DescribeRouteTables
- Utilisez des opérations d'API de service pour collecter les applications AWS Marketplace là où les tables de routage les utilisent.
  - DescribeRouteTables

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à planifier votre mise en réseau
- Recommandations de résilience AWS Direct Connect
- AWS Marketplace pour l'infrastructure réseau
- Livre blanc sur les options de connectivité Amazon Virtual Private Cloud
- Connectivité réseau haute disponibilité de plusieurs centres de données
- Utilisation de connexions VPN de site à site redondantes pour assurer le basculement
- Utilisation de la boîte à outils Direct Connect Resiliency Toolkit pour démarrer
- Points de terminaison d'un VPC et services de point de terminaison de VPC (AWS PrivateLink)
- Qu'est-ce qu'Amazon VPC ?
- Qu'est-ce que Transit Gateway ?
- Qu'est-ce qu'AWS Site-to-Site VPN ?

- Utilisation des passerelles Direct Connect

Vidéos connexes :

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (NET406-R1)

REL02-BP03 S'assurer que l'allocation des sous-réseaux IP tient compte de l'expansion et de la disponibilité

Les plages d'adresses IP de Amazon VPC doivent être assez grandes pour répondre aux exigences de charges de travail, y compris en prévision d'une expansion et de l'allocation d'adresses IP aux sous-réseaux sur les zones de disponibilité. Cela inclut les équilibreurs de charge, les instances EC2 et les applications basées sur conteneur.

Lorsque vous planifiez votre topologie de réseau, la première étape consiste à définir l'espace d'adressage IP lui-même. Des plages d'adresses IP privées (conformément aux directives RFC 1918) doivent être allouées pour chaque VPC. Vous devez remplir les exigences suivantes dans le cadre de ce processus :

- Autorisez un espace d'adressage IP pour plus d'un VPC par région.
- Au sein d'un VPC, prévoyez de l'espace supplémentaire pour plusieurs sous-réseaux couvrant plusieurs zones de disponibilité.
- Laissez toujours de l'espace de bloc CIDR non utilisé au sein d'un VPC pour une future expansion.
- Assurez-vous qu'il existe un espace d'adressage IP pour répondre aux besoins de tout parc transitoire d'instances EC2 que vous pourriez utiliser, comme les parcs d'instances Spot pour Machine Learning, les clusters Amazon EMR ou Amazon Redshift.
- Remarque : les quatre premières adresses IP et la dernière adresse IP dans le bloc CIDR de chaque sous-réseau sont réservées et ne peuvent pas être utilisées.
- Vous devez planifier le déploiement de grands blocs CIDR pour votre VPC. Notez que le bloc CIDR initial du VPC alloué à votre VPC ne peut pas être modifié ou supprimé, mais vous pouvez ajouter des blocs CIDR non superposés au VPC. Les CIDR IPv4 de sous-réseau ne sont pas modifiables, mais les CIDR IPv6 le sont. Gardez à l'esprit que le déploiement du plus grand VPC possible (/16) représente plus de 65 000 adresses IP. Dans l'espace d'adressage IP 10.x.x.x de base uniquement, vous pouvez mettre en service 255 VPC de ce type. Par conséquent, il est

préférable d’avoir un système surdimensionné que sous-dimensionné pour faciliter la gestion de vos VPC.

Anti-modèles courants :

- Création de petits VPC.
- Création de petits sous-réseaux, puis ajout de sous-réseaux aux configurations au fur et à mesure que vous développez.
- Estimation incorrecte du nombre d'adresses IP qu'un Elastic Load Balancer peut utiliser.
- Déploiement de nombreux équilibreurs de charge à trafic élevé dans les mêmes sous-réseaux.

Avantages liés au respect de cette bonne pratique : Ces tailles sont la garantie que vous pouvez prendre en charge la croissance de vos charges de travail et continuer à fournir une disponibilité au fur et à mesure de votre mise à l'échelle.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Planifiez votre réseau en prévision de votre croissance, de la conformité réglementaire et de son intégration avec d'autres composants. La croissance peut être sous-estimée, la conformité réglementaire peut changer, et les acquisitions ou les connexions à des réseaux privés peuvent être difficiles à implémenter sans une planification appropriée.
  - Sélectionnez les régions et Comptes AWS pertinents en fonction de vos exigences de services, de la latence, des exigences réglementaires et de reprise après sinistre (DR).
  - Identifiez vos besoins pour les déploiements VPC régionaux.
  - Identifiez la taille des VPC.
    - Déterminez si vous allez déployer une connectivité multi-VPC.
      - Qu'est-ce que Transit Gateway ?
      - Connectivité multi-VPC dans un seule région
    - Déterminez si vous avez besoin d'une mise en réseau séparée pour les exigences réglementaires.
    - Rendez vos VPC aussi larges que possible. Le bloc d'adresse CIDR du VPC initial alloué à votre VPC ne peut pas être modifié ou supprimé, mais vous pouvez ajouter des blocs

d'adresse CIDR non superposés au VPC. Cela peut toutefois fragmenter vos plages d'adresses.

- Rendez vos VPC aussi larges que possible. Le bloc d'adresse CIDR du VPC initial alloué à votre VPC ne peut pas être modifié ou supprimé, mais vous pouvez ajouter des blocs d'adresse CIDR non superposés au VPC. Cela peut toutefois fragmenter vos plages d'adresses.

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à planifier votre mise en réseau
- AWS Marketplace pour l'infrastructure réseau
- Livre blanc sur les options de connectivité Amazon Virtual Private Cloud
- Connectivité réseau haute disponibilité de plusieurs centres de données
- Connectivité multi-VPC dans un seule région
- Qu'est-ce qu'Amazon VPC ?

Vidéos connexes :

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (NET406-R1)

REL02-BP04 Préférer les topologies en étoile au maillage « many-to-many »

Si plus de deux espaces d'adresses réseau (par exemple, des VPC et des réseaux sur site) sont connectés via l'appairage de VPC, AWS Direct Connect ou un VPN, utilisez un modèle en étoile, comme celui fourni par AWS Transit Gateway.

Si vous n'avez que deux réseaux de ce type, vous pouvez simplement les connecter l'un à l'autre, mais à mesure que le nombre de réseaux augmente, la complexité de ces connexions maillées devient intenable. AWS Transit Gateway fournit un modèle en étoile facile à gérer, permettant d'acheminer le trafic sur vos différents réseaux.

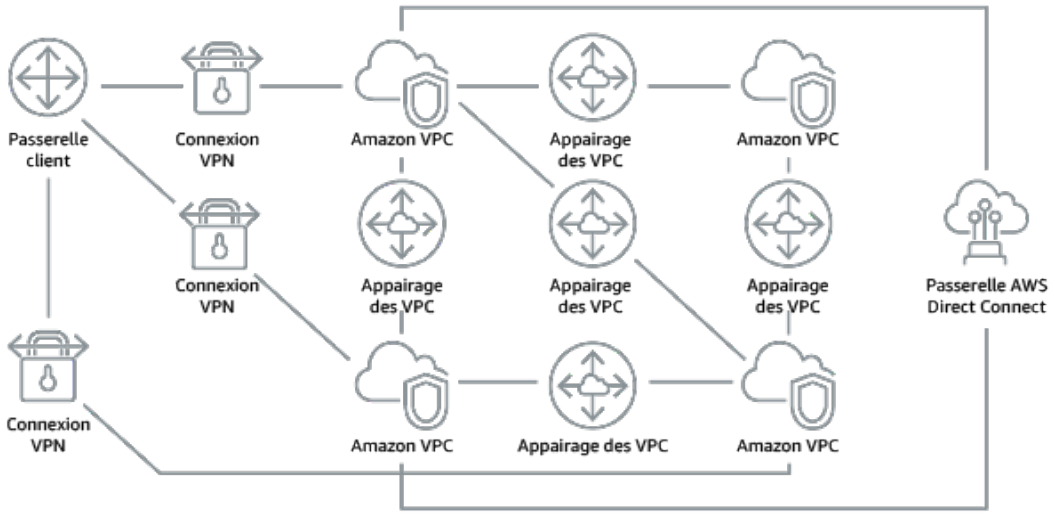


Figure 1 : Sans AWS Transit Gateway, vous devez appairer chaque Amazon VPC l'un à l'autre et à chaque emplacement sur site à l'aide d'une connexion VPN, ce qui peut devenir complexe à mesure que votre système se développe.

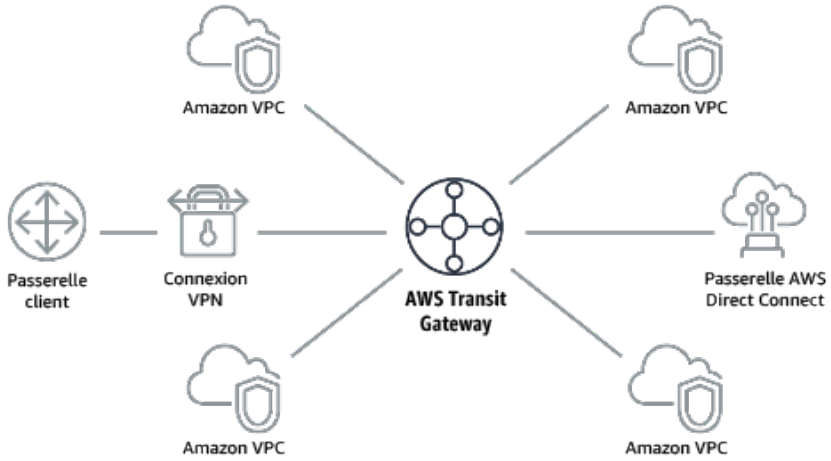


Figure 2 : Avec AWS Transit Gateway, il vous suffit de connecter chaque Amazon VPC ou VPN à AWS Transit Gateway pour acheminer le trafic vers et depuis chaque VPC ou VPN.

Anti-modèles courants :

- Utilisation de l'appairage de VPC pour connecter plus de deux VPC.
- Établissement de plusieurs séances BGP pour chaque VPC afin d'établir une connectivité couvrant les Virtual Private Cloud (VPC) répartis sur plusieurs Régions AWS.

Avantages liés au respect de cette bonne pratique : La complexité de ces connexions maillées devient intenable au fur et à mesure que le nombre de réseaux augmente. AWSTransit Gateway fournit un modèle en étoile facile à gérer, permettant d'acheminer le trafic entre vos différents réseaux.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Préférez les topologies en étoile au maillage « many-to-many ». Si plus de deux espaces d'adresses réseau (par exemple, des VPC et des réseaux sur site) sont connectés via l'appairage de VPC, AWS Direct Connect ou un VPN, utilisez un modèle en étoile, comme celui fourni par AWS Transit Gateway.
  - Quand il s’agit de seulement deux de ces réseaux, vous pouvez simplement les connecter l'un à l'autre, mais à mesure que le nombre de réseaux augmente, la complexité de ces connexions maillées devient intenable. AWS Transit Gateway fournit un modèle en étoile facile à gérer, permettant d'acheminer le trafic sur vos différents réseaux.
    - Qu'est-ce que Transit Gateway ?

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à planifier votre mise en réseau
- AWS Marketplace pour l'infrastructure réseau
- Connectivité réseau haute disponibilité de plusieurs centres de données
- Points de terminaison d'un VPC et services de point de terminaison de VPC (AWS PrivateLink)
- Qu'est-ce qu'Amazon VPC ?
- Qu'est-ce que Transit Gateway ?

Vidéos connexes :

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (NET406-R1)

REL02-BP05 Appliquer des plages d'adresses IP privées sans chevauchement dans tous les espaces d'adressage privés où ils sont connectés

Les plages d'adresses IP de chacun de vos VPC ne doivent pas se chevaucher lorsqu'elles sont appairées ou connectées via VPN. De même, vous devez éviter les conflits d'adresses IP entre un VPC et des environnements sur site, ou avec d'autres fournisseurs de cloud que vous utilisez. Vous devez également disposer d'un moyen d'allouer des plages d'adresses IP privées lorsque cela est nécessaire.

Un système de gestion des adresses IP (IPAM) peut vous y aider. Plusieurs IPAM sont disponibles sur AWS Marketplace.

Anti-modèles courants :

- Utilisation de la même plage d'adresses IP dans votre VPC que sur site ou dans votre réseau d'entreprise.
- Non suivi des plages d'adresses IP des VPC utilisés pour déployer vos charges de travail.

Avantages liés au respect de cette bonne pratique : La planification active de votre réseau garantit que vous n'avez pas plusieurs occurrences de la même adresse IP dans les réseaux interconnectés. Cela empêche les problèmes de routage de se produire dans certaines parties de la charge de travail qui utilisent les différentes applications.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Surveillez et gérez votre utilisation CIDR. Évaluez votre utilisation potentielle sur AWS, ajoutez des plages CIDR à des VPC existants et créez des VPC pour autoriser une croissance d'utilisation planifiée.
  - Capturez votre consommation CIDR actuelle (par exemple, les VPC et les sous-réseaux).
    - Utilisez des opérations d'API de service pour collecter la consommation CIDR actuelle.
  - Capturez l'utilisation actuelle de votre sous-réseau.
    - Utilisez des opérations d'API de service pour collecter les sous-réseaux par VPC dans chaque région.
      - DescribeSubnets
    - Enregistrez l'utilisation actuelle.
    - Déterminez si vous avez créé des plages d'adresses IP se chevauchant.

- Calculez la capacité inutilisée.
- Identifiez les plages d'adresses IP qui se chevauchent. Vous pouvez soit migrer vers une nouvelle plage d'adresses, soit utiliser les appliances de traduction de port et de réseau (NAT) d'AWS Marketplace si vous avez besoin de connecter les plages qui se chevauchent.

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à planifier votre mise en réseau
- AWS Marketplace pour l'infrastructure réseau
- Livre blanc sur les options de connectivité Amazon Virtual Private Cloud
- Connectivité réseau haute disponibilité de plusieurs centres de données
- Qu'est-ce qu'Amazon VPC ?
- Qu'est-ce qu'IPAM ?

Vidéos connexes :

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (NET406-R1)

Architecture de charge de travail

Questions

- FIA 3. Comment concevez-vous l'architecture de service de votre charge de travail ?
- FIA 4. Comment concevez-vous des interactions dans un système distribué pour éviter les défaillances ?
- FIA 5. Comment concevez-vous des interactions dans un système distribué pour atténuer ou résister aux défaillances ?

FIA 3. Comment concevez-vous l'architecture de service de votre charge de travail ?

Créez des charges de travail hautement évolutives et fiables à l'aide d'une architecture orientée service (SOA) ou d'une architecture de microservices. La SOA consiste à rendre les composants

logiciels réutilisables via les interfaces de service. L'architecture des microservices va plus loin, en particulier en rendant les composants plus petits et plus simples.

Bonnes pratiques

- REL03-BP01 Choisir comment segmenter votre charge de travail
- REL03-BP02 Créer des services axés sur des domaines d'activité et la fonctionnalité
- REL03-BP03 Fournir des contrats de service par API

REL03-BP01 Choisir comment segmenter votre charge de travail

La segmentation de la charge de travail est importante lorsqu'il s'agit de déterminer les exigences de résilience de votre application. L'architecture monolithique doit être évitée dans la mesure du possible. À la place, réfléchissez bien aux composants de l'application capables d'être divisés en microservices. Selon les exigences de votre application, il peut s'agir d'une combinaison d'une architecture orientée services et de microservices dans la mesure du possible. Les charges de travail capables d'absence d'état sont davantage en mesure d'être déployées en tant que microservices.

Résultat souhaité : Les charges de travail doivent être supportables, évolutives et aussi faiblement couplées que possible.

Lorsque vous choisissez comment segmenter votre charge de travail, comparez les avantages aux complexités. Ce qui convient pour un nouveau produit en course pour un premier lancement est différent de ce dont a besoin une charge de travail conçue pour augmenter d'échelle. Lors de la refactorisation d'une architecture monolithique existante, vous devez évaluer comment l'application prendra en charge une décomposition vers l'absence d'état. La division de services en microservices permet aux petites équipes bien définies de les développer et les gérer. Toutefois, les services plus petits peuvent créer des complexités dont une latence supérieure, un débogage plus complexe et une charge opérationnelle accrue.

Anti-modèles courants :

- La version microservice Death Star est une situation dans laquelle les composants atomiques deviennent si interdépendants que l'échec de l'un d'entre eux résulte en un échec encore plus important, ce qui rend les composants aussi rigides et fragiles qu'une architecture monolithique.

Avantages liés au respect de cette pratique :

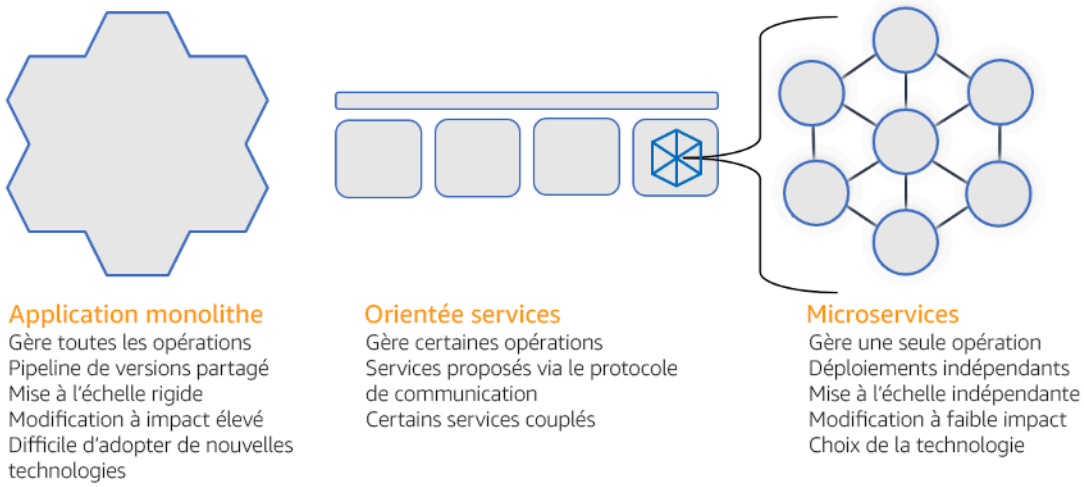
- L'utilisation de segments plus petits permet une plus grande agilité, une plus grande flexibilité organisationnelle et une évolutivité.
- L'impact réduit des interruptions de service.
- Les composants de l'application peuvent avoir différentes exigences de disponibilité, pouvant être pris en charge par une segmentation plus atomique.
- Des responsabilités bien définies pour les équipes prenant en charge la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Choisissez votre type d'architecture en fonction de la façon dont vous segmenterez votre charge de travail. Choisissez une architecture orientée service (SOA) ou une architecture de microservices (ou, dans de rares cas, une architecture monolithique). Même si vous choisissez de commencer avec une architecture monolithe, vous devez vous assurer qu'elle est modulaire et peut évoluer vers une SOA ou vers des microservices à mesure que votre produit se développe avec son adoption par les utilisateurs. Une SOA et une architecture de microservices offrent une segmentation plus petite. Si elle est préférable en tant qu'architecture moderne évolutive et fiable, il faut prendre en compte des compromis, notamment lors du déploiement d'une architecture de microservices.

Le principal compromis est que vous avez maintenant une architecture pour le calcul distribué qui peut compliquer le respect des exigences en matière de latence des utilisateurs et qui complexifie le suivi et le débogage des interactions des utilisateurs. AWS X-Ray peut vous aider à résoudre ce problème. Un autre effet à prendre en compte est la hausse de la complexité opérationnelle à mesure que vous augmentez le nombre d'applications que vous gérez, ce qui nécessite le déploiement de plusieurs composants indépendants.



Architectures monolithique, orientée services et de microservices

Étapes d'implémentation

- Déterminer l'architecture adaptée pour refactoriser ou créer votre application. SOA et les microservices offrent respectivement une segmentation plus petite, ce qui est préférable pour une architecture moderne évolutive et fiable. SOA peut constituer un bon compromis pour parvenir à une segmentation plus réduite tout en évitant certaines des complexités des microservices. Pour en savoir plus, voir Compromis des microservices.
- Si votre charge de travail est appropriée et que votre organisation peut la prendre en charge, vous devez utiliser une architecture de microservices pour obtenir la meilleure agilité et la meilleure fiabilité. Pour en savoir plus, voir Implémentation des microservices sur AWS.
- Tenir compte du modèle Figuier étrangleur pour refactoriser une architecture monolithique en composants plus petits. Cela implique de remplacer petit à petit des composants d'une application spécifique par de nouveaux services et applications. AWS Migration Hub Refactor Spaces agit comme le point de départ de la refactorisation incrémentielle. Pour en savoir plus, voir Migrer sans interruption vers des charges de travail existantes sur site à l'aide d'un modèle Figuier étrangleur.
- L'implémentation d'une architecture de microservices peut exiger un mécanisme de découverte de service pour permettre à ces services distribués de communiquer entre eux. AWS App Mesh peut être utilisé avec des architectures orientées service afin de fournir une découverte et un accès fiables aux services. AWS Cloud Map peut également être utilisé pour la découverte dynamique de service basée sur un DNS.

- Si vous migrez d'une architecture monolithique vers une architecture orientée service, Amazon MQ peut combler le fossé en tant que bus de services lors de la reconception des applications existantes dans le cloud.
- Pour les architectures monolithiques existantes avec une base de données partagée unique, choisissez comment réorganiser les données en segments plus petits. Vous pouvez les réorganiser par unité commerciale, modèle d'accès ou structure de données. À ce stade du processus de refactorisation, vous devez choisir d'utiliser un type de base de données relationnelle ou non relationnelle. Pour en savoir plus, voir De SQL à NoSQL.

Niveau d'effort du plan d'implémentation : Élevé

Ressources

Bonnes pratiques associées :

- REL03-BP02 Créer des services axés sur des domaines d'activité et la fonctionnalité

Documents connexes :

- Amazon API Gateway : configuration d'une API REST à l'aide d'OpenAPI
- Qu'est-ce que l'architecture orientée service ?
- Contexte délimité (modèle central dans la conception pilotée par domaine)
- Implémentation des microservices sur AWS
- Compromis des microservices
- Microservices : une définition de ce nouveau terme architectural
- Microservices sur AWS
- Qu'est-ce qu'AWS App Mesh ?

Exemples connexes :

- Atelier sur la modernisation itérative des applications

Vidéos connexes :

- Offrir l'excellence avec l'architecture de microservices sur AWS

REL03-BP02 Créer des services axés sur des domaines d'activité et la fonctionnalité

Une architecture orientée services (SOA) définit des services avec des fonctions bien déterminées dictées par les besoins métier. Les microservices utilisent des modèles de domaine et un contexte limité pour définir les limites des services en fonction des limites du contexte métier. En se concentrant sur les domaines d'activité et les fonctionnalités, les équipes peuvent définir des exigences de fiabilité indépendantes pour leurs services. Les contextes limités isolent et encapsulent la logique métier, ce qui permet aux équipes de mieux raisonner sur la manière de gérer les défaillances.

Résultat souhaité : les ingénieurs et les parties prenantes de l'entreprise définissent conjointement des contextes délimités et les utilisent pour concevoir des systèmes en tant que services remplissant des fonctions commerciales spécifiques. Ces équipes utilisent des pratiques établies telles que l'event storming pour définir les exigences. Les nouvelles applications sont conçues comme des services dont les limites sont bien définies et qui possèdent un couplage faible. Les monolithes existants sont décomposés en contextes limités et la conception des systèmes évolue vers des architectures SOA ou de microservices. Lorsque les monolithes sont refactorisés, des approches établies telles que les contextes de bulles et les modèles de décomposition des monolithes sont appliquées.

Les services orientés domaine sont exécutés sous la forme d'un ou de plusieurs processus qui ne partagent pas d'état. Ils répondent de manière indépendante aux fluctuations de la demande et gèrent les scénarios de panne à la lumière des exigences spécifiques du domaine.

Anti-modèles courants :

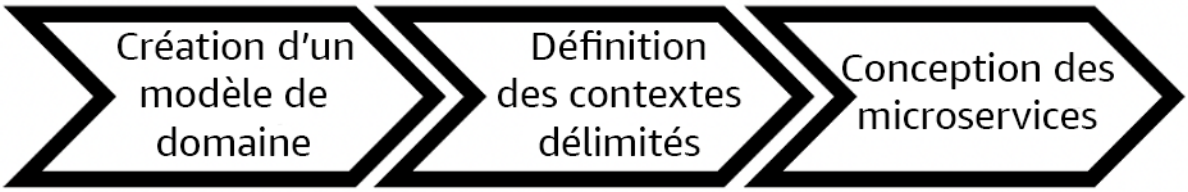
- Les équipes sont constituées autour de domaines techniques spécifiques tels que l'interface utilisateur et l'expérience utilisateur, les intergiciels ou les bases de données plutôt que de domaines commerciaux spécifiques.
- Les applications couvrent des responsabilités de domaine. Les services qui couvrent des contextes limités peuvent être plus difficiles à gérer, nécessiter des efforts de test plus importants et nécessiter la participation de plusieurs équipes de domaine aux mises à jour logicielles.
- Les dépendances de domaine, telles que les bibliothèques d'entités de domaine, sont partagées entre les services de telle sorte que les modifications apportées à un domaine de service nécessitent des modifications apportées à d'autres domaines de service.
- Les contrats de service et la logique métier n'expriment pas les entités dans un langage de domaine commun et cohérent, ce qui crée des couches de traduction qui compliquent les systèmes et augmentent les efforts de débogage.

Avantages liés au respect de cette bonne pratique : Les applications sont conçues comme des services indépendants délimités par domaines d'activité et utilisent un langage métier commun. Les services peuvent être testés et déployés indépendamment. Les services répondent aux exigences de résilience spécifiques au domaine mis en œuvre.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

La décision pilotée par domaine (DDD) est l'approche fondamentale de la conception et de la création de logiciels autour de domaines métier. Il est utile de travailler avec un cadre existant lorsque vous créez des services axés sur des domaines métier. Lorsque vous travaillez avec des applications monolithiques existantes, vous pouvez tirer parti des modèles de décomposition qui fournissent des techniques éprouvées pour moderniser les applications en services.



Décision pilotée par domaine

Étapes d'implémentation

- Les équipes peuvent organiser des ateliers d'event storming pour identifier rapidement les événements, les commandes, les agrégats et les domaines dans un format léger de notes autocollantes.
- Une fois que les entités et les fonctions de domaine ont été créées dans un contexte de domaine, vous pouvez diviser votre domaine en services en utilisant un contexte limitédans lequel les entités qui partagent des fonctions et des attributs similaires sont regroupées. La division en contextes permet de faire émerger un modèle de délimitation des microservices.
  - Par exemple, les entités du site Amazon.com peuvent inclure le colis, la livraison, le calendrier, le prix, la remise et la devise.
  - Le colis, la livraison et le calendrier sont regroupés dans le contexte d'expédition, tandis que le prix, la remise et la devise sont regroupés dans le contexte de tarification.

- La décomposition des monolithes en microservices décrit les modèles de refactorisation des microservices. L'utilisation de modèles de décomposition par capacité métier, sous-domaine ou transaction s'inscrit parfaitement dans les approches axées sur le domaine.
- Des techniques tactiques telles que le contexte de bulles vous permettent d'introduire le DDD dans des applications existantes ou héritées sans devoir procéder à des réécritures préalables et sans engagement total envers le DDD. Dans une approche de contexte à bulles, un petit contexte limité est établi à l'aide d'un mappage et d'une coordination des services, ou d'une couche anticorruption, qui protège le modèle de domaine nouvellement défini des influences extérieures.

Une fois que les équipes ont effectué une analyse du domaine et défini des entités et des contrats de service, elles peuvent tirer parti des services AWS pour mettre en œuvre leur conception axée sur le domaine sous forme de services basés sur le cloud.

- Commencez votre développement en définissant des tests qui appliquent les règles métier de votre domaine. Le développement piloté par les tests (TDD) et le développement piloté par le comportement (BDD) aident les équipes à concentrer leurs services sur la résolution des problèmes commerciaux.
- Sélectionnez les services AWS qui répondent le mieux aux exigences de votre domaine d'activité et à l'architecture de microservices :
  - AWS sans serveur permet à votre équipe de se concentrer sur une logique de domaine spécifique au lieu de gérer les serveurs et l'infrastructure.
  - Les conteneurs AWS simplifient la gestion de votre infrastructure afin que vous puissiez vous concentrer sur les exigences de votre domaine.
  - Les bases de données sur mesure vous permettent d'adapter les exigences de votre domaine au type de base de données le mieux adapté.
- Création d'architectures hexagonales sur AWS décrit un cadre permettant d'intégrer une logique métier à des services en procédant de manière rétroactive à partir d'un domaine métier afin de répondre à des exigences fonctionnelles, puis d'associer des adaptateurs d'intégration. Les modèles qui séparent les détails de l'interface de la logique métier avec les services AWS aident les équipes à se concentrer sur les fonctionnalités du domaine et à améliorer la qualité des logiciels.

Ressources

Bonnes pratiques associées :

- REL03-BP01 Choisir comment segmenter votre charge de travail
- REL03-BP03 Fournir des contrats de service par API

Documents connexes :

- Microservices AWS
- Implémentation des microservices sur AWS
- Comment convertir un monolithe en microservices
- Premiers pas avec DDD en présence de systèmes hérités
- Conception axée sur le domaine : aborder la complexité au cœur du logiciel
- Création d'architectures hexagonales sur AWS
- Décomposition des monolithes en microservices
- Event Storming
- Messages entre contextes limités
- Microservices
- Développement piloté par les tests
- Développement axé sur le comportement

Exemples connexes :

- Atelier natif cloud en entreprise
- Conception de microservices natifs cloud sur AWS (extrait de DDD/EventStormingWorkshop)

Outils associés :

- Bases de données AWS Cloud
- Sans serveur sur AWS
- Conteneurs AWS

REL03-BP03 Fournir des contrats de service par API

Les contrats de service sont des accords documentés entre les producteurs d'API et les consommateurs définis dans une définition d'API lisible par machine. Une stratégie de gestion

des versions permet aux consommateurs de continuer à utiliser l'API existante et de procéder à la migration de leurs applications vers la nouvelle API lorsqu'ils sont prêts. Le déploiement du producteur peut avoir lieu à tout moment tant que le contrat est respecté. Les équipes de service peuvent utiliser la pile technologique de leur choix pour satisfaire aux clauses du contrat d'API.

Résultat souhaité :

Anti-modèles courants : Les applications créées à l'aide d'architectures orientées services ou microservices peuvent fonctionner de manière indépendante tout en bénéficiant d'une dépendance intégrée à l'exécution. Les modifications apportées à un consommateur ou à un producteur d'API n'interrompent pas la stabilité de l'ensemble du système lorsque les deux parties respectent un contrat d'API commun. Les composants qui communiquent via des API de service peuvent exécuter des versions fonctionnelles indépendantes, effectuer des mises à niveau vers des dépendances d'exécution ou effectuer un basculement vers un site de reprise après sinistre (DR) avec peu ou pas d'impact mutuel. En outre, les services discrets sont capables d'évoluer de manière indépendante en absorbant la demande de ressources sans que les autres services évoluent à l'unisson.

- Création d'API de service sans schémas fortement typés. Cela se traduit par des API qui ne peuvent pas être utilisées pour générer des liaisons d'API et des charges utiles qui ne peuvent pas être validées par programmation.
- Absence d'adoption d'une stratégie de gestion des versions, qui oblige les utilisateurs d'API à les mettre à jour et à les publier, sous peine de défaillance lorsque les contrats de service évoluent.
- Messages d'erreur qui divulguent les détails de l'implémentation du service sous-jacent au lieu de décrire les échecs d'intégration dans le contexte et le langage du domaine.
- Absence d'utilisation des contrats d'API pour développer des scénarios de test et simuler des implémentations d'API afin de permettre des tests indépendants des composants du service.

Avantages liés au respect de cette bonne pratique : les systèmes distribués constitués de composants qui communiquent via des contrats de service d'API peuvent améliorer la fiabilité. Les développeurs peuvent détecter les problèmes potentiels au début du processus de développement en vérifiant les types lors de la compilation afin de s'assurer que les demandes et les réponses respectent le contrat d'API et que les champs obligatoires sont présents. Les contrats d'API fournissent une interface d'autodocumentation claire pour les API et assurent une meilleure interopérabilité entre les différents systèmes et langages de programmation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Une fois que vous avez identifié les domaines d'activité et déterminé la segmentation de votre charge de travail, vous pouvez développer vos API de service. Définissez d'abord des contrats de service lisibles par machine pour les API, puis mettez en œuvre une stratégie de gestion des versions des API. Lorsque vous êtes prêt à intégrer des services via des protocoles courants tels que REST, GraphQL ou des événements asynchrones, vous pouvez intégrer des services AWS à votre architecture pour intégrer vos composants à l'aide de contrats d'API clairement typés.

Services AWS pour les contrats d'API de service

Intégrez des services AWS, notamment Amazon API Gateway, AWS AppSyncet Amazon EventBridge dans votre architecture pour utiliser les contrats de service d'API dans votre application. Amazon API Gateway vous aide à intégrer directement des services AWS natifs et d'autres services Web. API Gateway prend en charge la spécification et la gestion des versions OpenAPI. AWS AppSyncest un point de terminaison GraphQL géré que vous configurez en définissant un schéma GraphQL pour définir une interface de service pour les requêtes, les mutations et les abonnements. Amazon EventBridge utilise des schémas d'événements pour définir des événements et générer des liaisons de code pour vos événements.

Étapes d'implémentation

- Tout d'abord, définissez un contrat pour votre API. Un contrat exprimera les fonctionnalités d'une API et définira des objets de données et des champs fortement typés pour l'entrée et la sortie de l'API.
- Lorsque vous configurez des API dans API Gateway, vous pouvez importer et exporter les spécifications OpenAPI pour vos points de terminaison.
  - L'importation d'une définition OpenAPI simplifie la création de votre API et peut être intégrée à l'infrastructure AWS sous forme d'outils de code tels qu' AWS Serverless Application Model et AWS Cloud Development Kit (AWS CDK).
  - L'exportation d'une définition d'API simplifie l'intégration aux outils de test d'API et fournit aux consommateurs de services une spécification d'intégration.
- Vous pouvez définir et gérer les API GraphQL avec AWS AppSync en définissant un fichier de schéma GraphQL afin de générer votre interface de contrat et de simplifier l'interaction avec des modèles REST complexes, plusieurs tables de base de données ou des services existants.
- Les projets AWS Amplify intégrés à AWS AppSync génèrent des fichiers de requête JavaScript fortement typés à utiliser dans votre application ainsi qu'une bibliothèque cliente AWS AppSync GraphQL pour les tables Amazon DynamoDB .

- Lorsque vous consommez des événements de service provenant d'Amazon EventBridge, les événements adhèrent à des schémas qui existent déjà dans le registre des schémas ou que vous définissez à l'aide de la spécification OpenAPI. Avec un schéma défini dans le registre, vous pouvez également générer des liaisons client à partir du contrat de schéma afin d'intégrer votre code aux événements.
- Extension ou gestion des versions de votre API. L'extension d'une API est une option plus simple lorsque vous ajoutez des champs qui peuvent être configurés avec des champs facultatifs ou des valeurs par défaut pour les champs obligatoires.
  - Les contrats basés sur JSON pour des protocoles tels que REST et GraphQL peuvent être une bonne solution pour l'extension de contrat.
  - Les contrats basés sur XML pour des protocoles tels que SOAP doivent être testés auprès des consommateurs de services afin de déterminer la faisabilité d'une extension du contrat.
- Lors de la gestion des versions d'une API, pensez à implémenter la gestion des versions par proxy lorsqu'une façade est utilisée pour prendre en charge les versions afin que la logique puisse être maintenue dans une base de code unique.
  - Avec API Gateway vous pouvez utiliser les mappages de demandes et de réponses afin de simplifier l'absorption des modifications du contrat en établissant une façade permettant de fournir des valeurs par défaut pour les nouveaux champs ou de supprimer les champs retirés d'une demande ou d'une réponse. Avec cette approche, le service sous-jacent peut gérer une base de code unique.

Ressources

Bonnes pratiques associées :

- REL03-BP01 Choisir comment segmenter votre charge de travail
- REL03-BP02 Créer des services axés sur des domaines d'activité et la fonctionnalité
- REL04-BP02 Implémenter des dépendances couplées faiblement
- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative
- REL05-BP05 Définir les délais d'attente du client

Documents connexes :

- Qu'est-ce qu'une API (interface de programmation d'applications) ?
- Implémentation des microservices sur AWS

- Compromis des microservices
- Microservices : une définition de ce nouveau terme architectural
- Microservices sur AWS
- Utilisation des extensions API Gateway pour OpenAPI
- Spécification OpenAPI
- GraphQL : schémas et types
- Liaisons de code Amazon EventBridge

Exemples connexes :

- Amazon API Gateway : configuration d'une API REST à l'aide d'OpenAPI
- Amazon API Gateway vers une application Amazon DynamoDB CRUD à l'aide d'OpenAPI
- Modèles modernes d'intégration des applications à l'ère de l'informatique sans serveur : intégration des services API Gateway
- Implémentation de la gestion des versions API Gateway basée sur les en-têtes avec Amazon CloudFront
- AWS AppSync : création d'une application client

Vidéos connexes :

- Using OpenAPI in AWS SAM to manage API Gateway

Outils associés :

- Amazon API Gateway
- AWS AppSync
- Amazon EventBridge

FIA 4. Comment concevez-vous des interactions dans un système distribué pour éviter les défaillances ?

Les systèmes distribués s'appuient sur des réseaux de communication pour interconnecter des composants comme des serveurs ou des services. Votre charge de travail doit fonctionner de manière fiable malgré la perte de données ou la latence dans ces réseaux. Les composants du

système distribué doivent fonctionner d'une manière qui n'a pas d'impact négatif sur les autres composants ou la charge de travail. Ces bonnes pratiques empêchent les défaillances et améliorent le temps moyen entre les défaillances (MTBF).

Bonnes pratiques

- REL04-BP01 Identifier le type de système distribué requis
- REL04-BP02 Implémenter des dépendances couplées faiblement
- REL04-BP03 Effectuer un travail constant
- REL04-BP04 Rendre toutes les réponses idempotentes

REL04-BP01 Identifier le type de système distribué requis

Les systèmes matériels distribués en temps réel exigent la fourniture des réponses de manière synchrone et rapide, alors que les systèmes en temps réel souples disposent d'une fenêtre de temps plus importante (en minutes ou plus). Les systèmes hors connexion gèrent les réponses via un traitement par lots ou asynchrone. Les systèmes matériels distribués en temps réel ont les exigences de fiabilité les plus strictes.

Les problèmes les plus complexes inhérents aux systèmes distribués concernent les systèmes distribués en temps réel stricts, également appelés services de requête/réponse. Ce qui les rend difficiles, c'est que les requêtes arrivent de façon imprévisible et que les réponses doivent être données rapidement (par exemple, le client attend activement la réponse). Les serveurs web front-end, le pipeline de commandes, les transactions par carte de crédit, chaque API AWS et la téléphonie en sont des exemples.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Identifiez le type de système distribué requis. Les défis posés par les systèmes distribués sont la latence, la mise à l'échelle, la compréhension des API de réseau, le regroupement et le dégroupement des données et la complexité des algorithmes tels que Paxos. Des cas jadis marginaux et théoriques deviennent monnaie courante au fur et à mesure que les systèmes deviennent de plus en plus grands et distribués.
  - L'Amazon Builders' Library : défis liés aux systèmes distribués
    - Des réponses données de manière synchrone et rapide sont nécessaires pour les systèmes matériels distribués en temps.

- Les systèmes logiciels en temps réel ont un créneau de temps plus généreux de plusieurs minutes ou plus pour la réponse.
- Les systèmes hors connexion gèrent les réponses via un traitement par lots ou asynchrone.
- Les systèmes matériels distribués en temps réel ont les exigences de fiabilité les plus strictes.

Ressources

Documents connexes :

- Amazon EC2 : garantir l'idempotence
- L'Amazon Builders' Library : défis liés aux systèmes distribués
- L'Amazon Builders' Library : fiabilité, travail constant et une bonne tasse de café
- Qu'est-ce qu'Amazon EventBridge ?
- Qu'est-ce que Amazon Simple Queue Service ?

Vidéos connexes :

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (inclut un couplage faible, un travail constant et une stabilité statique)
- AWS re:Invent 2019: Moving to event-driven architectures (SVS308)

REL04-BP02 Implémenter des dépendances couplées faiblement

Des dépendances telles que des systèmes de file d'attente, des systèmes de streaming, des flux de travail et des équilibreurs de charge sont couplées faiblement. Le couplage faible permet d'isoler le comportement d'un composant des autres composants qui en dépendent, ce qui augmente la résilience et l'agilité.

Dans les systèmes couplés fortement, la modification d'un composant peut nécessiter de modifier d'autres composants qui en dépendent, ce qui entraîne une dégradation des performances de tous les composants. Le couplage faible rompt cette dépendance de sorte que les composants dépendants n'ont besoin que de connaître l'interface publiée et sa version. La mise en œuvre d'un couplage faible entre les dépendances permet d'isoler une défaillance dans l'une afin de ne pas en impacter une autre.

Le couplage faible vous permet de modifier le code ou d'ajouter des fonctionnalités à un composant tout en minimisant les risques pour les autres composants qui en dépendent. Il offre également une résilience granulaire au niveau des composants, ce qui vous permet de mettre à l'échelle voire de modifier la mise en œuvre sous-jacente de la dépendance.

Pour améliorer encore la résilience par un couplage faible, dans la mesure du possible, rendez asynchrones les interactions des composants. Ce modèle convient à toute interaction qui ne nécessite pas une réponse immédiate et pour laquelle une confirmation de l'enregistrement d'une requête suffira. Il implique un composant qui génère des événements et un autre qui les consomme. Les deux composants ne s'intègrent pas via une interaction directe point à point, mais généralement via une couche de stockage durable intermédiaire, telle qu'une file d'attente Amazon SQS ou une plateforme de données de streaming comme Amazon Kinesis ou AWS Step Functions.

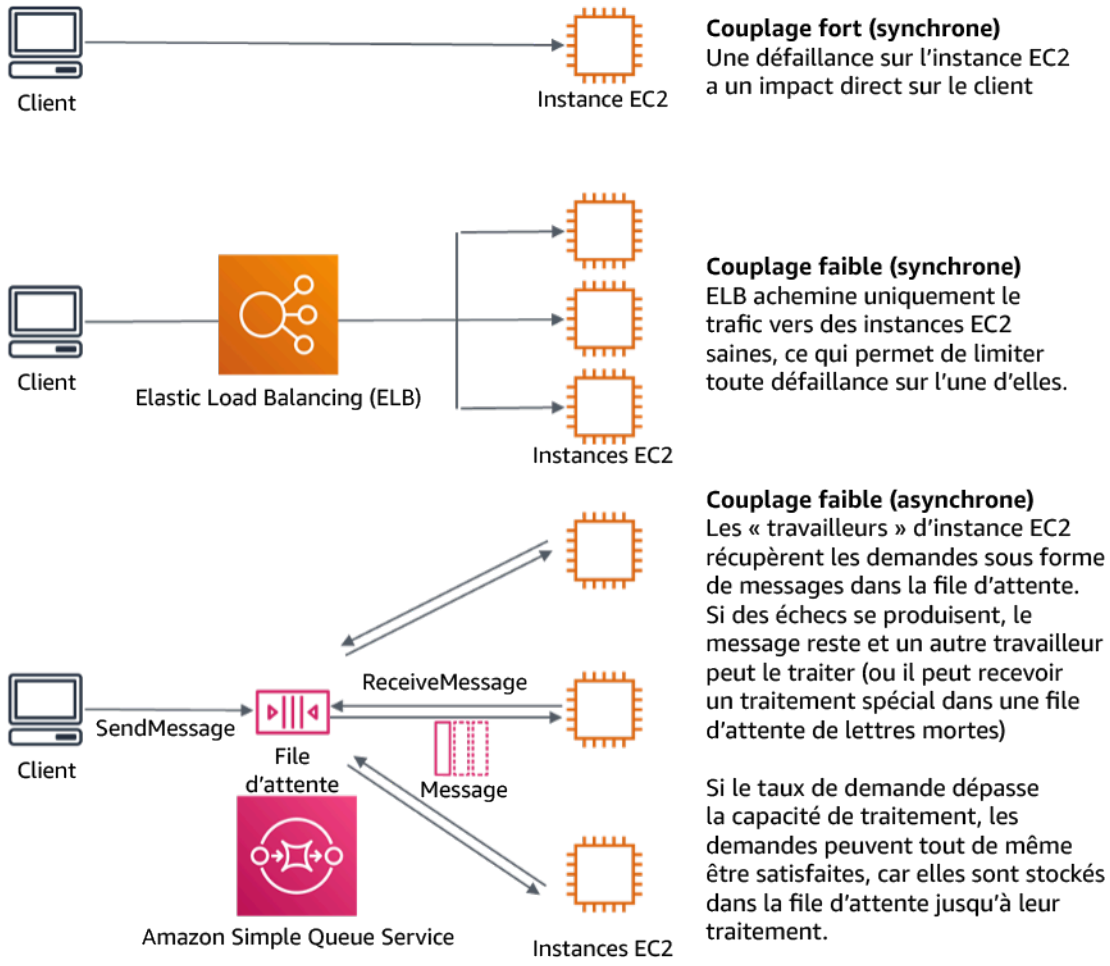


Figure 4 : Les dépendances telles que des systèmes de file d'attente et des équilibreurs de charge sont couplées faiblement

Les files d'attente Amazon SQS et les programmes Elastic Load Balancer ne sont que deux façons d'ajouter une couche intermédiaire pour un couplage faible. Les architectures guidées par les événements peuvent également être conçues dans le AWS Cloud à l'aide d'Amazon EventBridge, qui peut extraire des clients (producteurs d'événements) des services sur lesquels ils s'appuient (clients d'événements). Amazon Simple Notification Service (Amazon SNS) est une solution efficace lorsque vous avez besoin d'une messagerie de type « many-to-many », à haut débit et en mode push. Grâce aux rubriques Amazon SNS, vos systèmes d'édition peuvent diffuser des messages vers un grand nombre de points de terminaison abonnés pour un traitement parallèle.

Bien que les files d'attente offrent plusieurs avantages, dans la plupart des systèmes en temps réel stricts, les requêtes antérieures à un seuil (souvent en secondes) sont considérées comme obsolètes (le client a abandonné et n'attend plus de réponse). En conséquence, elles ne sont pas traitées. De cette façon, les requêtes plus récentes (et probablement toujours valides) peuvent être traitées à la place.

Résultat souhaité : la mise en œuvre de dépendances couplées faiblement réduit la surface de défaillance au niveau des composants, ce qui permet de diagnostiquer et de résoudre les problèmes. Elle simplifie également les cycles de développement en permettant aux équipes de mettre en œuvre des modifications à un niveau modulaire sans affecter les performances des autres composants qui en dépendent. Avec cette approche, il est possible de mettre à l'échelle un composant en fonction des besoins en ressources et de l'utilisation de ce composant, ce qui contribue à améliorer la rentabilité.

Anti-modèles courants :

- Déploiement d'une charge de travail monolithique.
- Appel direct d'API entre les niveaux de charge de travail sans possibilité de basculement ou de traitement asynchrone de la demande.
- Couplage fort à l'aide de données partagées. Les systèmes couplés faiblement évitent de partager des données par le biais de bases de données partagées ou d'autres formes de stockage de données couplées fortement, qui peuvent réintroduire un couplage fort et entraver la capacité de mise à l'échelle.
- Ignorer la contre-pression. Votre charge de travail doit être capable de ralentir ou d'arrêter les données entrantes lorsqu'un composant ne peut pas les traiter au même rythme.

Avantages liés au respect de cette bonne pratique : le couplage faible permet d'isoler le comportement d'un composant de celui des autres composants qui en dépendent, et d'augmenter ainsi la résilience et l'agilité. La défaillance d'un composant est isolée des autres.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Implémentez des dépendances couplées faiblement. Différentes solutions permettent de créer des applications couplées faiblement. Ces services incluent notamment la mise en œuvre de files d'attente entièrement gérées, des flux de travail automatisés, des réactions aux événements et des API, qui peuvent aider à isoler un composant des autres composants et, par conséquent, à accroître la résilience et l'agilité.

- Créer des architectures basées sur les événements : Amazon EventBridge vous aide à créer des architectures basées sur les événements, qui sont couplées faiblement et distribuées.
- Mettre en œuvre des files d'attente dans des systèmes distribués : vous pouvez utiliser Amazon Simple Queue Service (Amazon SQS) pour intégrer et découpler des systèmes distribués.
- Conteneuriser les composants en tant que microservices : les microservices permettent aux équipes de créer des applications composées de petits composants indépendants qui communiquent via des API bien définies. Amazon Elastic Container Service(Amazon ECS) et Amazon Elastic Kubernetes Service (Amazon EKS) peuvent vous aider à faire plus rapidement vos premiers pas avec les conteneurs.
- Gérer les flux de travail avec Step Functions : Step Functions vous aide à coordonner plusieurs services AWS dans des flux de travail flexibles.
- Tirer parti des architectures de messagerie pub/sub : Amazon Simple Notification Service (Amazon SNS) transmet les messages des diffuseurs de publication aux abonnés (également appelés producteurs et consommateurs).

Étapes d'implémentation

- Les composants d'une architecture basée sur les événements sont initiés par des événements. Les événements sont des actions qui se produisent dans un système (par exemple, un utilisateur ajoute un article à un panier). Lorsque l'action aboutit, un événement est généré et active le composant suivant du système.
  - Building Event-driven Applications with Amazon EventBridge
  - AWS re:Invent 2022 - Designing Event-Driven Integrations using Amazon EventBridge

- Les systèmes de messagerie distribuée comportent trois parties principales qui doivent être mises en œuvre pour une architecture basée sur des files d'attente : les composants du système distribué, la file d'attente utilisée pour le découplage (distribuée sur des serveurs Amazon SQS) et les messages de la file d'attente. Dans un système classique, les producteurs envoient le message dans la file d'attente et le consommateur reçoit le message de la file d'attente. La file d'attente stocke les messages sur plusieurs serveurs Amazon SQS à des fins de redondance.
  - Basic Amazon SQS architecture
  - Send Messages Between Distributed Applications with Amazon Simple Queue Service
- Lorsqu'ils sont bien utilisés, les microservices améliorent la maintenabilité et la capacité de mise à l'échelle, car les composants couplés faiblement sont gérés par des équipes indépendantes. Ils permettent également d'isoler les comportements d'un composant en cas de changement.
  - Implémentation des microservices sur AWS
  - Let's Architect! Architecting microservices with containers
- Avec AWS Step Functions, vous pouvez notamment créer des applications distribuées, automatiser des processus et orchestrer des microservices. L'orchestration de plusieurs composants dans un flux de travail automatisé vous permet de découpler des dépendances dans votre application.
  - Create a Serverless Workflow with AWS Step Functions and AWS Lambda
  - Mise en route avec AWS Step Functions

Ressources

Documents connexes :

- Amazon EC2: Ensuring Idempotency
- L'Amazon Builders' Library : défis liés aux systèmes distribués
- The Amazon Builders' Library: Reliability, constant work, and a good cup of coffee
- What Is Amazon EventBridge?
- What Is Amazon Simple Queue Service?
- Break up with your monolith
- Orchestrate Queue-based Microservices with AWS Step Functions and Amazon SQS
- Basic Amazon SQS architecture
- Architecture basée sur des files d'attente

Vidéos connexes :

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (inclut le couplage faible, le travail constant et la stabilité statique)
- AWS re:Invent 2019: Moving to event-driven architectures (SVS308)
- AWS re:Invent 2019: Scalable serverless event-driven applications using Amazon SQS and Lambda (API304)
- AWS re:Invent 2019: Scalable serverless event-driven applications using Amazon SQS and Lambda
- AWS re:Invent 2022 - Designing event-driven integrations using Amazon EventBridge
- AWS re:Invent 2017: Elastic Load Balancing Deep Dive and Best Practices

REL04-BP03 Effectuer un travail constant

Les systèmes peuvent échouer en cas de modifications importantes et rapides de la charge. Par exemple, si votre charge de travail effectue une surveillance de l'état de milliers de serveurs, elle doit envoyer chaque fois une charge utile de la même taille (un instantané complet de l'état actuel). Qu'aucun des serveurs ne présente de problème ou qu'ils en connaissent tous, le système de surveillance de l'état effectue un travail constant sans modifications importantes ni rapides.

Par exemple, si le système de vérification de l'état surveille 100 000 serveurs, la charge sur celui-ci est nominale sous le taux de défaillance normalement faible du serveur. En revanche, si un événement majeur rendait la moitié de ces serveurs défectueux, le système de vérification de l'état serait submergé en tentant de mettre à jour les systèmes de notification et de communiquer l'état à ses clients. Le système de vérification de l'état doit donc plutôt envoyer à chaque fois l'instantané complet de l'état actuel. 100 000 états d'intégrité du serveur, chacun représenté par un bit, ne seraient qu'une charge utile de 12,5 Ko. Qu'aucun des serveurs ne présente de problème ou qu'ils en connaissent tous, le système de vérification de l'état effectue un travail constant, et les modifications importantes et rapides ne menacent pas la stabilité du système. C'est ainsi qu'Amazon Route 53 gère les vérifications de l'état des points de terminaison (tels que les adresses IP) pour déterminer comment les utilisateurs finaux sont acheminés vers eux.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

- Effectuer un travail constant : les systèmes peuvent échouer lorsque la charge connaît des changements rapides et importants.
- Implémentez des dépendances couplées faiblement. Des dépendances telles que des systèmes de file d'attente, des systèmes de streaming, des flux de travail et des équilibreurs de charge sont couplées faiblement. Le couplage faible permet d'isoler le comportement d'un composant des autres composants qui en dépendent, ce qui augmente la résilience et l'agilité.
  - L'Amazon Builders' Library : fiabilité, travail constant et une bonne tasse de café
  - AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (includes constant work)
    - Pour l'exemple d'un système de vérification de l'état surveillant 100 000 serveurs, concevez les charges de travail de manière à ce que les tailles de charge utile restent constantes, quel que soit le nombre de réussites ou d'échecs.

Ressources

Documents connexes :

- Amazon EC2 : garantir l'idempotence
- L'Amazon Builders' Library : défis liés aux systèmes distribués
- L'Amazon Builders' Library : fiabilité, travail constant et une bonne tasse de café

Vidéos connexes :

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (includes constant work)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (inclut un couplage faible, un travail constant et une stabilité statique)
- AWS re:Invent 2019: Moving to event-driven architectures (SVS308)

REL04-BP04 Rendre toutes les réponses idempotentes

Un service idempotent promet que chaque demande est traitée une seule fois et exactement de la même façon, de sorte que l'exécution de plusieurs demandes identiques ait le même effet qu'une seule demande. Un service idempotent permet à un client d'implémenter plus facilement les nouvelles tentatives sans craindre qu'une demande soit traitée plusieurs fois par erreur. Pour ce faire, les clients peuvent émettre des demandes d'API avec un jeton d'idempotence. Le même jeton est utilisé chaque fois que la demande est répétée. Une API de service idempotente utilise le jeton pour renvoyer une réponse identique à la réponse qui a été renvoyée la première fois que la demande a été traitée.

Dans un système distribué, il est facile d'effectuer une action au maximum une fois (le client n'effectue qu'une seule demande) ou au moins une fois (continuer à demander jusqu'à ce que le client reçoive la confirmation de la réussite). En revanche, il est difficile de garantir qu'une action est idempotente, c'est-à-dire exécutée une seule fois, de sorte que l'exécution de plusieurs demandes identiques a le même effet qu'une seule demande. En utilisant des jetons d'idempotence dans les API, les services peuvent recevoir une demande de mutation une ou plusieurs fois sans créer d'enregistrements dupliqués ou induire des effets secondaires.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Rendez toutes les réponses idempotentes. Un service idempotent promet que chaque demande est traitée une seule fois et exactement de la même façon, de sorte que l'exécution de plusieurs demandes identiques ait le même effet qu'une seule demande.
  - Les clients peuvent émettre des demandes d'API avec un jeton d'idempotence. Le même jeton est utilisé chaque fois que la demande est répétée. Une API de service idempotente utilise le jeton pour renvoyer une réponse identique à la réponse qui a été renvoyée la première fois que la demande a été traitée.
    - Garantir l'idempotence Amazon EC2

Ressources

Documents connexes :

- Garantir l'idempotence Amazon EC2
- L'Amazon Builders' Library : défis liés aux systèmes distribués

- L'Amazon Builders' Library : fiabilité, travail constant et une bonne tasse de café

Vidéos connexes :

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (inclut un couplage faible, un travail constant et une stabilité statique)
- AWS re:Invent 2019: Moving to event-driven architectures (SVS308)

FIA 5. Comment concevez-vous des interactions dans un système distribué pour atténuer ou résister aux défaillances ?

Les systèmes distribués s'appuient sur des réseaux de communication pour interconnecter des composants (tels que des serveurs ou des services). Votre charge de travail doit fonctionner de manière fiable malgré la perte de données ou la latence sur ces réseaux. Les composants du système distribué doivent fonctionner d'une manière qui n'a pas d'impact négatif sur les autres composants ou la charge de travail. Ces bonnes pratiques permettent aux charges de travail de résister aux contraintes ou aux défaillances, de s'en remettre plus rapidement et d'atténuer l'impact de ces altérations. Il en résulte une amélioration du temps moyen de récupération (MTTR).

Bonnes pratiques

- REL05-BP01 Implémenter une dégradation appropriée pour transformer les dépendances matérielles applicables en dépendances logicielles
- REL05-BP02 Limiter les demandes
- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative
- REL05-BP04 Procéder à une interruption immédiate et limiter les files d'attente
- REL05-BP05 Définir les délais d'attente du client
- REL05-BP06 Rendre les services sans état dans la mesure du possible
- REL05-BP07 Mettre en place des leviers d'urgence

REL05-BP01 Implémenter une dégradation appropriée pour transformer les dépendances matérielles applicables en dépendances logicielles

Les composants de l'application doivent continuer à exécuter leur fonction principale même si les dépendances deviennent indisponibles. Ils peuvent fournir des données légèrement obsolètes, des données alternatives ou même aucune donnée. Cela garantit que le fonctionnement global du système n'est que très peu entravé par des défaillances localisées tout en fournissant une valeur commerciale centrale.

Résultat souhaité : Lorsque les dépendances d'un composant ne sont pas saines, le composant lui-même peut continuer de fonctionner, bien que de manière dégradée. Les modes de défaillance des composants doivent être considérés comme un fonctionnement normal. Les flux de travail doivent être conçus de manière à ce que ces défaillances n'aboutissent pas à une défaillance complète ou qu'elles aboutissent au moins à des états prévisibles et récupérables.

Anti-modèles courants :

- Ne pas identifier les fonctionnalités métier essentielles nécessaires. Ne pas tester le fonctionnement des composants, même en cas de défaillance des dépendances.
- Aucune donnée n'est diffusée en cas d'erreur ou lorsqu'une seule dépendance parmi plusieurs n'est pas disponible et que des résultats partiels peuvent toujours être renvoyés.
- Création d'un état incohérent lorsqu'une transaction échoue partiellement.
- Ne pas disposer d'un autre moyen d'accéder à un magasin de paramètres central.
- Invalider ou vider l'état local à la suite d'un échec d'actualisation sans prendre en compte les conséquences d'une telle opération.

Avantages liés au respect de cette bonne pratique : une dégradation progressive améliore la disponibilité du système dans son ensemble et maintient la fonctionnalité des fonctions les plus importantes même en cas de panne.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

La mise en œuvre d'une dégradation progressive permet de minimiser l'impact des défaillances de dépendance sur le fonctionnement des composants. Idéalement, un composant détecte les défaillances liées aux dépendances et les contourne de manière à avoir un impact minimal sur les autres composants ou les clients.

L'architecture permettant une dégradation progressive implique de prendre en compte les modes de défaillance potentiels lors de la conception des dépendances. Pour chaque mode de défaillance, déterminez un moyen de fournir la plupart des fonctionnalités, ou les plus critiques d'entre elles, du composant aux appelants ou aux clients. Ces considérations peuvent devenir des exigences supplémentaires qui peuvent être testées et vérifiées. Idéalement, un composant est capable d'exécuter sa fonction principale de manière acceptable, même en cas de défaillance d'une ou de plusieurs dépendances.

Il s'agit tout autant d'une discussion commerciale que technique. Toutes les exigences commerciales sont importantes et doivent être satisfaites dans la mesure du possible. Cependant, il est tout de même logique de se demander ce qui doit se passer lorsque toutes les exigences ne peuvent pas être satisfaites. Un système peut être conçu pour être disponible et cohérent, mais lorsqu'une exigence doit être supprimée, laquelle est la plus importante ? Pour le traitement des paiements, il peut s'agir de la cohérence. Pour une application en temps réel, il peut s'agir de la disponibilité. Pour un site Web orienté client, la réponse peut dépendre des attentes du client.

Ce que cela signifie dépend des exigences du composant et de ce qui doit être considéré comme sa fonction principale. Par exemple :

- un site Web d'e-commerce peut afficher des données provenant de plusieurs systèmes différents, par exemple des recommandations personnalisées, les produits les mieux classés et l'état des commandes des clients sur la page de destination. Lorsqu'un système en amont est défaillant, il est tout de même judicieux d'afficher tout le reste au lieu d'afficher une page d'erreur à un client.
- Un composant effectuant des écritures par lots peut toujours continuer à traiter un lot si l'une des opérations individuelles échoue. La mise en œuvre d'un mécanisme de nouvelle tentative doit être simple. Cela peut être fait en renvoyant à l'appelant des informations indiquant quelles opérations ont réussi, lesquelles ont échoué et pourquoi elles ont échoué, ou en plaçant les demandes ayant échoué dans une file d'attente de lettres mortes pour implémenter des tentatives asynchrones. Les informations relatives aux opérations ayant échoué doivent également être consignées.
- Un système qui traite les transactions doit vérifier que toutes les mises à jour individuelles sont exécutées ou qu'aucune d'entre elles ne l'est. Pour les transactions distribuées, le modèle Saga peut être utilisé pour annuler les opérations précédentes en cas d'échec d'une opération ultérieure de la même transaction. Ici, la fonction principale est de maintenir la cohérence.
- Les systèmes soumis à des contraintes de temps doivent être en mesure de gérer les dépendances qui ne répondent pas en temps voulu. Dans ces cas de figure, le modèle du disjoncteur peut être utilisé. Lorsque les réponses d'une dépendance commencent à expirer, le système peut passer à un état fermé où aucun appel supplémentaire n'est effectué.

- Une application peut lire des paramètres à partir d'un magasin de paramètres. Il peut être utile de créer des images de conteneur avec un ensemble de paramètres par défaut et de les utiliser si le magasin de paramètres n'est pas disponible.

Notez que les chemins empruntés en cas de défaillance d'un composant doivent être testés et doivent être nettement plus simples que le chemin principal. D'une manière générale, les stratégies de repli doivent être évitées.

Étapes d'implémentation

Identifiez les dépendances externes et internes. Déterminez quels types de défaillances peuvent y survenir. Réfléchissez à des moyens de minimiser l'impact négatif sur les systèmes en amont et en aval, ainsi que sur les clients lors de ces défaillances.

Vous trouverez ci-dessous une liste des dépendances et la manière de les dégrader de façon appropriée en cas d'échec :

1. Défaillance partielle des dépendances : un composant peut adresser plusieurs demandes à des systèmes en aval, soit sous la forme de demandes multiples adressées à un système, soit sous celle d'une demande adressée à plusieurs systèmes. Selon le contexte métier, différentes méthodes de gestion peuvent être appropriées (pour plus de détails, voir les exemples précédents dans le guide de mise en œuvre).
2. Un système en aval est incapable de traiter les demandes en raison d'une charge élevée : si les demandes adressées à un système en aval échouent régulièrement, il n'est pas logique de continuer à réessayer. Cela peut créer une charge supplémentaire sur un système déjà surchargé et rendre la récupération plus difficile. Le modèle du disjoncteur peut être utilisé ici afin de surveiller les appels en échec vers un système en aval. Si un grand nombre d'appels échouent, il cessera d'envoyer d'autres demandes au système en aval et n'autorisera les appels qu'occasionnellement pour vérifier si le système en aval est à nouveau disponible.
3. Un magasin de paramètres n'est pas disponible : pour transformer un magasin de paramètres, vous pouvez utiliser la mise en cache des dépendances souples ou des valeurs par défaut saines incluses dans les images de conteneur ou de machine. Notez que ces valeurs par défaut doivent être tenues à jour et incluses dans les suites de tests.
4. Aucun service de surveillance ou aucune autre dépendance non fonctionnelle n'est disponible : si un composant ne peut pas envoyer par intermittence des journaux, des métriques ou des traces à un service de surveillance central, il est souvent préférable de continuer à exécuter les fonctions métier comme d'habitude. Il est souvent inacceptable de ne pas enregistrer ni de pousser des

métriques pendant une longue période. En outre, certains cas d'utilisation peuvent nécessiter des entrées d'audit complètes pour répondre aux exigences de conformité.

5. Une instance principale d'une base de données relationnelle n'est peut-être pas disponible : Amazon Relational Database Service, comme presque toutes les bases de données relationnelles, ne peut comporter qu'une seule instance d'écriture principale. Cela crée un point de défaillance unique pour les charges de travail d'écriture et complique la mise à l'échelle. Ce problème peut être partiellement atténué en utilisant une configuration multi-AZ pour une haute disponibilité ou Amazon Aurora sans serveur pour une meilleure évolutivité. Pour des exigences de très haute disponibilité, il peut être judicieux de ne pas se fier du tout au rédacteur principal. Pour les requêtes qui se limitent à la lecture, des répliques de lecture peuvent être utilisées, ce qui assure la redondance et la possibilité d'une mise à l'échelle, et pas seulement d'une augmentation. Les écritures peuvent être mises en mémoire tampon, par exemple dans une file d'attente Amazon Simple Queue Service, afin que les demandes d'écriture des clients puissent toujours être acceptées même si le serveur principal est temporairement indisponible.

Ressources

Documents connexes :

- Amazon API Gateway : limiter les demandes d'API pour un meilleur débit
- Coupe-circuit (présentation du coupe-circuit, ouvrage « Release It! »)
- Nouvelles tentatives après erreur et interruptions exponentielles dans AWS
- Michael Nygard « Release It! Design and Deploy Production-Ready Software »
- L'Amazon Builders' Library : éviter le basculement dans les systèmes distribués
- L'Amazon Builders' Library : éviter les retards de file d'attente insurmontables
- L'Amazon Builders' Library : défis et stratégies de mise en cache
- L'Amazon Builders' Library : délais d'attente, nouvelles tentatives et backoff avec instabilité

Vidéos connexes :

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

Exemples connexes :

- Atelier Well-Architected : niveau 300 : implémentation de la surveillance de l'état et gestion des dépendances pour améliorer la fiabilité

REL05-BP02 Limiter les demandes

Limitez les demandes pour atténuer l'épuisement des ressources en cas d'augmentation imprévue de la demande. Les demandes inférieures à la limite sont traitées tandis que celles dépassant la limite définie sont rejetées et présentent un message de retour indiquant que la demande a dépassé la limite.

Résultat souhaité : les pics de volume importants, qu'ils soient dus à une augmentation soudaine du trafic client, à des inondations ou à des tempêtes de nouvelles tentatives, sont atténués par la limitation des demandes, ce qui permet aux charges de travail de poursuivre le traitement normal du volume de demandes pris en charge.

Anti-modèles courants :

- Les limitations des points de terminaison de l'API ne sont pas implémentées ou leurs valeurs par défaut sont conservées sans tenir compte des volumes attendus.
- Les points de terminaison de l'API ne sont pas testés en termes de charge ou les limites de régulation ne sont pas testées.
- Limiter les taux de demandes sans tenir compte de la taille ou de la complexité des demandes.
- Tester les taux de demande maximaux ou la taille maximale des demandes, mais pas les deux simultanément.
- Les ressources ne sont pas provisionnées selon les mêmes limites établies lors des tests.
- Aucun plan d'utilisation n'a été configuré ni envisagé pour les utilisateurs d'API d'application à application (A2A).
- Les utilisateurs de files d'attente qui redimensionnent horizontalement ne disposent pas de paramètres de simultanéité maximaux configurés.
- La limitation du débit par adresse IP n'a pas été mise en œuvre.

Avantages liés au respect de cette bonne pratique : Les charges de travail qui fixent des limites peuvent fonctionner normalement et traiter correctement le chargement des demandes acceptées en cas de pics de volume inattendus. Les pics soudains ou soutenus de demandes adressées aux API et aux files d'attente sont limités et n'épuisent pas les ressources de traitement des demandes.

Les limites de débit limitent les requêtes individuelles afin que les volumes élevés de trafic provenant d'une seule adresse IP ou d'un seul utilisateur d'API n'épuisent pas les ressources et n'aient pas d'impact sur les autres consommateurs.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les services doivent être conçus pour traiter une capacité connue de demandes ; cette capacité peut être établie par des tests de charge. Si les taux d'arrivée des demandes dépassent les limites, la réponse appropriée indique qu'une demande a été limitée. Cela permet au consommateur de gérer l'erreur et de réessayer ultérieurement.

Lorsque votre service nécessite une implémentation de limitation, pensez à implémenter l'algorithme du compartiment à jetons, dans lequel un jeton compte pour une demande. Les jetons sont rechargés à une vitesse limitée par seconde et vidés de manière asynchrone à raison d'un jeton par demande.

L'algorithme du compartiment à jetons.

Amazon API Gateway implémente l'algorithme du compartiment à jetons en fonction des limites du compte et de la région et il peut être configuré par client avec des plans d'utilisation. En outre, Amazon Simple Queue Service (Amazon SQS) et Amazon Kinesis peuvent mettre en mémoire tampon les demandes afin de lisser le taux de demandes et de permettre des taux de limitation plus élevés pour les demandes pouvant être traitées. Enfin, vous pouvez implémenter la limitation du débit avec AWS WAF afin de limiter les utilisateurs d'API spécifiques qui génèrent une charge anormalement élevée.

Étapes d'implémentation

Vous pouvez configurer API Gateway avec des limites de régulation pour vos API et retourner `des erreurs 429 Demandes trop nombreuses` lorsque les limites sont dépassées. Vous pouvez utiliser AWS WAF avec vos points de terminaison API Gateway et AWS AppSync pour activer la limitation du débit par adresse IP. En outre, lorsque votre système peut tolérer un traitement asynchrone, vous pouvez placer les messages dans une file d'attente ou un flux pour accélérer les réponses aux clients du service, ce qui vous permet d'atteindre des taux d'accélération plus élevés.

Avec le traitement asynchrone, lorsque vous avez configuré Amazon SQS en tant que source d'événements pourAWS Lambda, vous pouvez configurer la simultanéité maximale afin d'éviter que des taux d'événements élevés n'épuisent le quota d'exécution simultanée disponible du compte requis pour d'autres services de votre charge de travail ou de votre compte.

Bien qu'API Gateway propose une implémentation gérée du compartiment à jetons, lorsque vous ne pouvez pas utiliserAPI Gateway, vous pouvez tirer parti des implémentations open source spécifiques à la langue (voir les exemples associés dans Ressources) du compartiment à jetons pour vos services.

- Comprenez et configurez les limitations API Gateway au niveau du compte par région, de l'API par étape et de la clé d'API par niveau de plan d'utilisation.
- Appliquez les règles de limitation de débit AWS WAF vers les points de terminaison API Gateway et AWS AppSync pour vous protéger contre les inondations et bloquer les adresses IP malveillantes. Les règles de limitation de débit peuvent également être configurées sur les clés d'API AWS AppSync pour les consommateurs A2A.
- Déterminez si vous avez besoin d'un contrôle plus limitant qu'une limitation du débit pour les API AWS AppSync et, si c'est le cas, configurez un API Gateway devant votre point de terminaison AWS AppSync.
- Lorsque des files d'attente Amazon SQS sont configurées comme déclencheurs pour les utilisateurs de files d'attente Lambda, définissez la simultanéité maximale sur une valeur capable d'effectuer un traitement suffisant pour vous permettre d'atteindre vos objectifs de niveau de service sans pour autant dépasser les limites de simultanéité affectant les autres fonctions Lambda. Envisagez de définir une concurrence réservée pour d'autres fonctions Lambda du même compte et de la même région lorsque vous utilisez des files d'attente avec Lambda.
- Utilisez API Gateway avec des intégrations de services natives vers Amazon SQS ou Kinesis pour mettre des demandes en mémoire tampon.

- Si vous ne pouvez pas utiliser API Gateway, examinez les bibliothèques spécifiques à la langue pour implémenter l'algorithme de compartiment à jetons adapté à votre charge de travail. Consultez la section des exemples et faites vos propres recherches pour trouver une bibliothèque appropriée.
- Testez les limites que vous envisagez de définir ou d'autoriser à augmenter, et documentez les limites testées.
- N'augmentez pas les limites au-delà de ce que vous avez établi lors des tests. Lorsque vous augmentez une limite, vérifiez que les ressources provisionnées sont déjà équivalentes ou supérieures à celles des scénarios de test avant d'appliquer l'augmentation.

Ressources

Bonnes pratiques associées :

- REL04-BP03 Effectuer un travail constant
- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative

Documents connexes :

- Amazon API Gateway : limiter les demandes d'API pour un meilleur débit
- AWS WAF : déclaration de règle basée sur le débit
- Introduction d'une simultanéité maximale de AWS Lambda en utilisant Amazon SQS comme source d'événements
- AWS Lambda : simultanéité maximale

Exemples connexes :

- Les trois principales règles AWS WAF basées sur le débit
- Bucket4j Java
- Jeton-compartiment Python
- Nœud jeton-compartiment
- Limitation du débit de threading du système .NET

Vidéos connexes :

- Implementing GraphQL API security best practices with AWS AppSync

Outils associés :

- Amazon API Gateway
- AWS AppSync
- Amazon SQS
- Amazon Kinesis
- AWS WAF

REL05-BP03 Contrôler et limiter les appels de nouvelle tentative

Utilisez le backoff exponentiel pour relancer les demandes à des intervalles de plus en plus longs entre chaque nouvelle tentative. Introduisez un décalage entre les tentatives afin de randomiser les intervalles entre les tentatives. Limitez le nombre maximal de tentatives.

Résultat souhaité : les composants types d'un système logiciel distribué incluent les serveurs, les équilibreurs de charge, les bases de données et les serveurs DNS. Pendant le fonctionnement normal, ces composants peuvent répondre aux demandes par des erreurs temporaires ou limitées, ainsi que par des erreurs qui persisteraient indépendamment des nouvelles tentatives. Lorsque des clients adressent des demandes à des services, celles-ci consomment des ressources, notamment de la mémoire, des threads, des connexions, des ports ou toute autre ressource limitée. Le contrôle et la limitation des nouvelles tentatives constituent une stratégie visant à libérer et à minimiser la consommation de ressources afin que les composants du système soumis à des contraintes ne soient pas surchargés.

Lorsque le client demande une expiration du délai ou reçoit des réponses d'erreur, il doit décider de réessayer ou non. S'il recommence, il le fait avec un backoff exponentiel avec une instabilité et une valeur de nouvelle tentative maximale. Par conséquent, les services et processus back-end sont moins sollicités et le temps nécessaire pour s'autoréparer est réduit, ce qui se traduit par une récupération plus rapide et un traitement efficace des demandes.

Anti-modèles courants :

- Implémentation de nouvelles tentatives sans ajouter de valeurs de backoff exponentiel, d'instabilité et de nouvelles tentatives maximales. Le backoff et l'instabilité permettent d'éviter les pics de trafic artificiels dus à des tentatives involontaires coordonnées à intervalles réguliers.
- Implémentation de nouvelles tentatives sans tester leurs effets ou en supposant que les nouvelles tentatives sont déjà intégrées dans un kit SDK sans tester de scénarios de nouvelles tentatives.

- Incapacité à comprendre les codes d'erreur publiés à partir des dépendances, ce qui entraîne une nouvelle tentative pour toutes les erreurs, y compris celles dont la cause claire indique un manque d'autorisation, une erreur de configuration ou toute autre condition qui, comme on pouvait s'y attendre, ne sera pas résolue sans intervention manuelle.
- Ne pas aborder les pratiques d'observabilité, notamment la surveillance et l'envoi d'alertes en cas de pannes de service répétées afin que les problèmes sous-jacents soient connus et puissent être résolus.
- Développement de mécanismes de nouvelle tentative personnalisés lorsque des fonctionnalités de nouvelle tentative intégrées ou tierces suffisent.
- Réessayer à plusieurs couches de votre pile d'applications d'une manière qui complique les nouvelles tentatives et augmente la consommation de ressources lors d'une tempête de nouvelles tentatives. Assurez-vous de comprendre comment ces erreurs affectent votre application et les dépendances sur lesquelles vous vous appuyez, puis implémentez les nouvelles tentatives à un seul niveau.
- Réessayer les appels de service qui ne sont pas idempotents, ce qui peut entraîner des effets secondaires inattendus tels que des résultats dupliqués.

Avantages liés au respect de cette bonne pratique : les nouvelles tentatives aident les clients à obtenir les résultats souhaités lorsque les requêtes échouent, mais elles font également perdre plus de temps au serveur pour obtenir les réponses souhaitées. Lorsque les défaillances sont rares ou transitoires, les nouvelles tentatives fonctionnent bien. Lorsque les défaillances sont causées par une surcharge de ressources, les nouvelles tentatives peuvent aggraver la situation. L'ajout d'un backoff exponentiel avec instabilité aux nouvelles tentatives des clients permet aux serveurs de se rétablir en cas de défaillance provoquée par une surcharge de ressources. L'instabilité permet d'éviter l'alignement des demandes en pics, tandis que le backoff réduit l'escalade de charge provoquée par l'ajout de nouvelles tentatives au chargement normal des demandes. Enfin, il est important de configurer un nombre maximal de nouvelles tentatives ou un temps écoulé afin d'éviter de créer des backlogs susceptibles d'entraîner des échecs métastables.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Contrôler et limiter les appels de nouvelle tentative. Utilisez le backoff exponentiel pour réessayer après des intervalles progressivement plus longs. Introduisez l'instabilité pour randomiser les intervalles de nouvelle tentative et limiter le nombre maximal de nouvelles tentatives.

Les kits AWS SDK implémentent les nouvelles tentatives et le backoff exponentiel par défaut. Utilisez ces implémentations AWS intégrées là où cela s'avère nécessaire dans votre charge de travail. Implémentez une logique similaire dans votre charge de travail lorsque vous appelez des services qui sont idempotents et où les nouvelles tentatives améliorent la disponibilité de vos clients. Déterminez quels sont les délais d'expiration et quand les nouvelles tentatives doivent s'arrêter en fonction de votre cas d'utilisation. Créez et mettez en pratique des scénarios de test pour ces cas d'utilisation impliquant de nouvelles tentatives.

Étapes d'implémentation

- Déterminez la couche optimale de votre pile d'applications pour implémenter de nouvelles tentatives pour les services sur lesquels repose votre application.
- Soyez conscient des SDK existants qui mettent en œuvre des stratégies de nouvelle tentative éprouvées avec un backoff exponentiel et une instabilité pour la langue de votre choix, et privilégiez-les par rapport à l'écriture de vos propres implémentations de nouvelles tentatives.
- Vérifiez que les services sont idempotents avant de mettre en œuvre de nouvelles tentatives. Une fois les nouvelles tentatives mises en œuvre, assurez-vous qu'elles sont à la fois testées et régulièrement mises en œuvre en production.
- Lorsque vous appelez des API de service AWS, utilisez les kits AWS SDK et AWS CLI et comprenez les options de configuration d'une nouvelle tentative. Déterminez si les valeurs par défaut conviennent à votre cas d'utilisation, testez-les et ajustez-les si nécessaire.

Ressources

Bonnes pratiques associées :

- REL04-BP04 Rendre toutes les réponses idempotentes
- REL05-BP02 Limiter les demandes
- REL05-BP04 Procéder à une interruption immédiate et limiter les files d'attente
- REL05-BP05 Définir les délais d'attente du client
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances

Documents connexes :

- Nouvelles tentatives après erreur et backoff exponentiel dans AWS
- Amazon Builders' Library : délais d'attente, nouvelles tentatives et backoff avec instabilité

- Backoff exponentiel et instabilité
- Sécuriser les nouvelles tentatives avec des API idempotentes

Exemples connexes :

- Nouvelle tentative Spring
- Nouvelle tentative Resilience4j

Vidéos connexes :

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

Outils associés :

- Kits AWS SDK et outils : comportement en cas de nouvelle tentative
- AWS Command Line Interface : nouvelles tentatives AWS CLI

REL05-BP04 Procéder à une interruption immédiate et limiter les files d'attente

Lorsqu'un service n'est pas en mesure de répondre correctement à une demande, procédez à son interruption immédiate. Cela permet la libération des ressources associées à une demande et donne la possibilité au service de récupérer s'il lui manque des ressources. L'interruption immédiate est un modèle de conception logicielle bien établi qui peut être exploité pour créer des charges de travail hautement fiables dans le cloud. La mise en file d'attente est également un modèle d'intégration d'entreprise bien établi qui permet de faciliter le chargement et de permettre aux clients de libérer des ressources lorsque le traitement asynchrone peut être toléré. Lorsqu'un service est capable de répondre correctement dans des conditions normales, mais échoue lorsque le taux de demandes est trop élevé, utilisez une file d'attente pour mettre les demandes en mémoire tampon. Toutefois, ne permettez pas l'accumulation de longs backlogs de files d'attente susceptibles d'entraîner le traitement de demandes obsolètes auxquelles un client a déjà renoncé.

Résultat souhaité : lorsque les systèmes sont confrontés à des problèmes de ressources, à des dépassements de délai, à des exceptions ou à des pannes grises qui rendent les objectifs de niveau de service irréalisables, les stratégies d'interruption immédiate permettent d'accélérer la récupération du système. Les systèmes qui doivent absorber les pics de trafic et peuvent prendre en charge le traitement asynchrone peuvent améliorer la fiabilité en permettant aux clients de lancer rapidement

des demandes grâce à l'utilisation des files d'attente pour mettre en mémoire tampon les demandes envoyées aux services back-end. Lors de la mise en mémoire tampon des demandes dans des files d'attente, des stratégies de gestion des files d'attente sont mises en œuvre pour éviter des backlogs insurmontables.

Anti-modèles courants :

- Mise en œuvre de files de messages sans configurer de files d'attente de lettres mortes (DLQ) ni d'alarmes sur les volumes DLQ pour détecter les défaillances d'un système.
- Il ne s'agit pas de mesurer l'ancienneté des messages dans une file d'attente, mais de mesurer la latence pour comprendre quand les utilisateurs prennent du retard ou si des erreurs entraînent de nouvelles tentatives.
- Conservation des messages en attente dans une file d'attente, alors qu'il n'est plus utile de traiter ces messages si l'entreprise n'en a plus besoin.
- La configuration de files d'attente du premier entré, premier sorti (FIFO) au moment du dernier entré, premier sorti (LIFO) permettrait de mieux répondre aux besoins des clients, par exemple lorsqu'un ordre strict n'est pas requis et que le traitement du backlog retarde toutes les nouvelles demandes urgentes, ce qui entraîne une violation des niveaux de service pour tous les clients.
- Exposition des files d'attente internes aux clients au lieu d'exposer les API qui gèrent la prise de travail et placent les demandes dans les files d'attente internes.
- La combinaison d'un trop grand nombre de types de demandes de travail dans une seule file d'attente peut aggraver les problèmes de backlog en répartissant la demande de ressources entre les types de demandes.
- Traitement de demandes complexes et simples dans la même file d'attente, malgré la nécessité d'une surveillance, de délais d'expiration et d'allocations de ressources différents.
- Absence de validation des entrées ou utilisation des assertions pour implémenter des mécanismes rapides dans les logiciels qui génèrent des exceptions vers des composants de niveau supérieur capables de gérer les erreurs de façon appropriée.
- Absence de suppression des ressources défectueuses du routage des requêtes, en particulier lorsque les défaillances sont grises, ce qui indique à la fois des réussites et des échecs en raison d'un plantage et d'un redémarrage, d'une panne de dépendance intermittente, d'une capacité réduite ou d'une perte de paquets réseau.

Avantages liés au respect de cette bonne pratique : les systèmes qui tombent rapidement en panne sont plus faciles à déboguer et à corriger, et présentent souvent des problèmes de codage et de

configuration avant la publication des versions en production. Les systèmes qui intègrent des stratégies de mise en file d'attente efficaces offrent une résilience et une fiabilité accrues face aux pics de trafic et aux pannes intermittentes du système.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les stratégies infaillibles peuvent être codées dans des solutions logicielles ou configurées dans l'infrastructure. En plus de leur capacité d'interruption immédiate, les files d'attente constituent une technique architecturale simple mais puissante qui permet de découpler les composants du système en douceur. Amazon CloudWatch fournit des fonctionnalités de surveillance et d'alerte en cas de défaillance. Une fois que l'on sait qu'un système est défaillant, des stratégies d'atténuation peuvent être invoquées, notamment en cas de défaillance de ressources altérées. Lorsque les systèmes implémentent des files d'attente avec Amazon SQS et d'autres technologies de file d'attente pour faciliter le chargement, ils doivent tenir compte de la gestion des backlogs de files d'attente, ainsi que des défaillances de consommation de messages.

Étapes d'implémentation

- Implémentez des assertions par programmation ou des métriques spécifiques dans votre logiciel et utilisez-les pour émettre des alertes explicites en cas de problèmes système. Amazon CloudWatch vous permet de créer des métriques et des alarmes en fonction du modèle de journal des applications et de l'instrumentation du kit SDK.
- Utilisez les métriques et les alarmes CloudWatch pour éviter les problèmes liés à l'altération des ressources qui ajoutent de la latence au traitement ou qui échouent à plusieurs reprises à traiter les demandes.
- Utilisez le traitement asynchrone en concevant des API pour accepter les demandes et les ajouter aux files d'attente internes en utilisant Amazon SQS, puis en répondant au client émetteur du message par un message de réussite afin que le client puisse libérer des ressources et passer à autre chose pendant que les utilisateurs de la file d'attente back-end traitent les demandes.
- Mesurez et surveillez la latence de traitement des files d'attente en produisant une métrique CloudWatch chaque fois que vous retirez un message d'une file d'attente en comparant l'heure actuelle à l'horodatage du message.
- Lorsque des défaillances empêchent le bon traitement des messages ou que des pics de trafic concernent des volumes qui ne peuvent pas être traités conformément aux contrats de niveau de service, mettez de côté le trafic ancien ou excédentaire vers une file d'attente de débordement. Cela permet de traiter en priorité les nouvelles tâches et les tâches plus anciennes lorsque la

capacité est disponible. Cette technique est une approximation du traitement LIFO et permet un traitement normal du système pour toutes les nouvelles tâches.

- Utilisez des lettres mortes ou réadaptez des files d'attente afin de déplacer les messages qui ne peuvent pas être traités hors du backlog vers un emplacement pouvant faire l'objet de recherches et de résolutions ultérieures.
- Réessayez ou, si cela est acceptable, supprimez les anciens messages en comparant l'heure actuelle à l'horodatage du message et en supprimant les messages qui ne sont plus pertinents pour le client demandeur.

Ressources

Bonnes pratiques associées :

- REL04-BP02 Implémenter des dépendances couplées faiblement
- REL05-BP02 Limiter les demandes
- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative
- REL06-BP02 Définir et calculer des métriques (agrégation)
- REL06-BP07 Surveiller la traçabilité complète des demandes via votre système

Documents connexes :

- Éviter les backlogs insurmontables dans les files d'attente
- Interruption immédiate
- Comment puis-je empêcher la croissance d'un backlog de messages dans ma file d'attente Amazon SQS ?
- Elastic Load Balancing : changement de zone
- Amazon Route 53Application Recovery Controller : contrôle du routage pour le basculement du trafic

Exemples connexes :

- Modèles d'intégration d'entreprise : canal des lettres mortes

Vidéos connexes :

- AWS re:Invent 2022 - Operating highly available Multi-AZ applications

Outils associés :

- Amazon SQS
- Amazon MQ
- AWS IoT Core
- Amazon CloudWatch

REL05-BP05 Définir les délais d'attente du client

Définissez les délais d'expiration de manière appropriée pour les connexions et les demandes, vérifiez-les systématiquement et ne vous fiez pas aux valeurs par défaut, car elles ne tiennent pas compte des spécificités de la charge de travail.

Résultat souhaité : les délais d'expiration du client doivent prendre en compte le coût pour le client, le serveur et la charge de travail associés à l'attente des demandes dont le traitement prend un temps anormal. Dans la mesure où il est impossible de connaître la cause exacte d'un délai d'attente, les clients doivent utiliser leur connaissance des services pour établir des attentes relatives aux causes probables et aux délais d'expiration appropriés.

Le délai d'expiration des connexions client dépend des valeurs configurées. Après avoir dépassé le délai imparti, les clients décident de revenir en arrière et de réessayer ou d'ouvrir un disjoncteur. Ces modèles évitent d'émettre des demandes susceptibles d'exacerber un problème d'erreur sous-jacent.

Anti-modèles courants :

- Ne pas connaître les délais d'expiration du système ou les délais d'expiration par défaut.
- Ne pas connaître le délai normal d'exécution des demandes.
- Ne pas connaître les raisons pour lesquelles les demandes peuvent prendre un temps anormalement long à traiter, ni les coûts pour le client, le service ou les performances de la charge de travail associés à l'attente de ces traitements.
- Ne pas connaître la probabilité qu'un réseau défaillant entraîne l'échec d'une requête uniquement lorsque le délai d'expiration est atteint, ainsi que les coûts pour les performances du client et de la charge de travail si l'on n'adopte pas un délai d'expiration plus court.
- Ne pas tester les scénarios de délai d'expiration à la fois pour les connexions et les demandes.

- Définir des délais d'expiration trop élevés, ce qui peut entraîner de longs temps d'attente et augmenter l'utilisation des ressources.
- Définir des délais d'attente trop bas, ce qui entraîne des défaillances artificielles.
- Oublier les modèles pour gérer les erreurs de temporisation des appels distants, par exemple les disjoncteurs et les nouvelles tentatives.
- Ne pas envisager de surveiller les taux d'erreur des appels de service, les objectifs de niveau de service en matière de latence et les valeurs aberrantes en matière de latence. Ces métriques peuvent fournir des informations sur les délais d'attente agressifs ou permissifs.

Avantages liés au respect de cette bonne pratique : les délais d'expiration des appels distants sont configurés et les systèmes sont conçus pour gérer les délais d'expiration de façon appropriée afin de préserver les ressources lorsque les appels distants répondent de manière anormalement lente et que les erreurs de délai d'expiration sont gérées de façon appropriée par les clients du service.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Définissez un délai d'expiration de la connexion et un délai d'expiration de la demande pour tout appel de dépendance de service et, plus généralement, pour tout appel entre processus. De nombreux cadres proposent des capacités de délai d'expiration intégrées, mais soyez prudent, car certains ont des valeurs par défaut infinies ou supérieures à ce qui est acceptable pour vos objectifs de service. Une valeur trop élevée réduit l'utilité du délai d'attente, car les ressources continuent d'être consommées pendant que le client attend l'expiration du délai. Une valeur trop faible peut générer un trafic accru sur le back-end et une latence accrue en raison du nombre excessif de demandes réessayées. Dans certains cas, cela peut entraîner des interruptions complètes, car toutes les demandes font l'objet d'une nouvelle tentative.

Tenez compte des points suivants lorsque vous déterminez des stratégies de délai d'expiration :

- Le traitement des demandes peut prendre plus de temps que d'habitude en raison de leur contenu, de défaillances d'un service cible ou d'une panne de partition réseau.
- Les demandes dont le contenu est anormalement coûteux peuvent consommer des ressources inutiles du serveur et du client. Dans ce cas, le fait d'avoir un délai d'expiration pour ces demandes et de ne pas réessayer peut préserver les ressources. Les services doivent également se protéger contre les contenus anormalement coûteux avec des limites et des délais d'expiration côté serveur.

- Les demandes qui prennent anormalement longtemps en raison d'une défaillance du service peuvent être interrompues et réessayées. Il convient de tenir compte des coûts de service liés à la demande et à la nouvelle tentative, mais si la cause est une déficience localisée, une nouvelle tentative ne sera probablement pas coûteuse et réduira la consommation de ressources du client. Le délai d'expiration peut également libérer des ressources du serveur en fonction de la nature de la déficience.
- Les demandes dont l'exécution prend beaucoup de temps parce que la demande ou la réponse n'a pas été envoyée par le réseau peuvent être interrompues et réessayées. La demande ou la réponse n'ayant pas été envoyée, il en aurait résulté un échec indépendamment de la durée du délai imparti. Dans ce cas, l'expiration du délai ne libérera pas les ressources du serveur, mais des ressources client et cela améliorera les performances de la charge de travail.

Tirez parti de modèles de conception bien établis, tels que les nouvelles tentatives et les disjoncteurs, pour gérer les délais d'expiration de façon appropriée et adopter des approches d'interruption immédiate. Les kits AWS SDK et AWS CLI permettent de configurer les délais d'expiration de la connexion et des demandes, ainsi que d'effectuer de nouvelles tentatives avec des backoffs exponentiels et des instabilités. Les fonctions AWS Lambda prennent en charge la configuration des délais d'expiration et, avec AWS Step Functions, vous pouvez créer des disjoncteurs low-code qui tirent parti des intégrations prédéfinies avec des services AWS et des kits de développement logiciel (SDK). AWS App Mesh Envoy fournit des fonctionnalités de délai d'expiration et de disjoncteur.

Étapes d'implémentation

- Configurez les délais d'expiration pour les appels de service à distance et profitez des fonctionnalités de délai d'expiration spécifique à la langue intégrées ou des bibliothèques de délai d'expiration open source.
- Lorsque votre charge de travail passe des appels avec un kit AWS SDK, consultez la documentation pour la configuration du délai d'expiration spécifique à la langue.
  - Python
  - PHP
  - .NET
  - Ruby
  - Java
  - Go
  - Node.js

  - C++
- Lorsque vous utilisez des kits AWS SDK ou des commandes AWS CLI dans votre charge de travail, configurez les valeurs de délai d'expiration par défaut en définissant les valeurs de configuration AWS par défaut pour `connectTimeoutInMillis` et `tlsNegotiationTimeoutInMillis`.
- Appliquez les options de ligne de commande `cli-connect-timeout` et `cli-read-timeout` pour contrôler les commandes AWS CLI ponctuelles destinées à des services AWS.
- Surveillez les appels de service à distance pour détecter les délais d'expiration et définissez des alarmes en cas d'erreurs persistantes afin de pouvoir gérer de manière proactive les scénarios d'erreur.
- Implémentez les métriques CloudWatch et la détection des anomalies CloudWatch sur les taux d'erreur d'appel, les objectifs de niveau de service en matière de latence et les valeurs aberrantes pour la latence afin de fournir des informations sur la gestion des délais d'attente trop agressifs ou permissifs.
- Configurez les délais d'expiration sur les fonctions Lambda.
- Les clients API Gateway doivent implémenter leurs propres tentatives lors de la gestion des délais d'expiration. API Gateway prend en charge un délai d'expiration de l'intégration de 50 millisecondes à 29 secondes pour les intégrations en aval et ne réessaie pas lorsque l'intégration demande un délai d'expiration.
- Implémentez le modèle de disjoncteur pour éviter de passer des appels à distance lorsque le délai d'expiration est écoulé. Ouvrez le circuit pour éviter les échecs d'appels et fermez-le lorsque les appels répondent normalement.
- Pour les charges de travail basées sur des conteneurs, consultez les fonctionnalités App Mesh Envoy permettant de tirer parti des délais d'expiration et des disjoncteurs intégrés.
- Utilisez AWS Step Functions pour créer des disjoncteurs low-code pour les appels de service à distance, en particulier lorsque vous appelez des kits AWS SDK natifs et des intégrations Step Functions compatibles afin de simplifier votre charge de travail.

Ressources

Bonnes pratiques associées :

- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative
- REL05-BP04 Procéder à une interruption immédiate et limiter les files d'attente

- REL06-BP07 Surveiller la traçabilité complète des demandes via votre système

Documents connexes :

- Kits SDK AWS : nouvelles tentatives et délais d'attente
- Amazon Builders' Library : délais d'attente, nouvelles tentatives et backoff avec instabilité
- Quotas Amazon API Gateway et notes importantes
- AWS Command Line Interface : options de ligne de commande
- AWS SDK for Java 2.x : configuration des délais d'expiration de l'API
- AWSBotocore utilisant l'objet de configuration et la référence de configuration
- AWS SDK for .NET : nouvelles tentatives et délais d'expiration
- AWS Lambda : configuration des options de fonction Lambda

Exemples connexes :

- Utilisation du modèle de disjoncteur avec AWS Step Functions et Amazon DynamoDB
- Martin Fowler : disjoncteur

Outils associés :

- Les kits AWS SDK
- Les fonctions AWS Lambda
- Amazon SQS
- AWS Step Functions
- AWS Command Line Interface

REL05-BP06 Rendre les services sans état dans la mesure du possible

Les services ne doivent pas exiger d'état ou doivent décharger un état de telle sorte qu'entre les différentes demandes client, il n'y ait pas de dépendance vis-à-vis des données stockées localement sur disque et en mémoire. Cela permet aux serveurs d'être remplacés à volonté sans avoir d'impact sur la disponibilité. Amazon ElastiCache ou Amazon DynamoDB sont de bonnes destinations pour l'état déchargé.

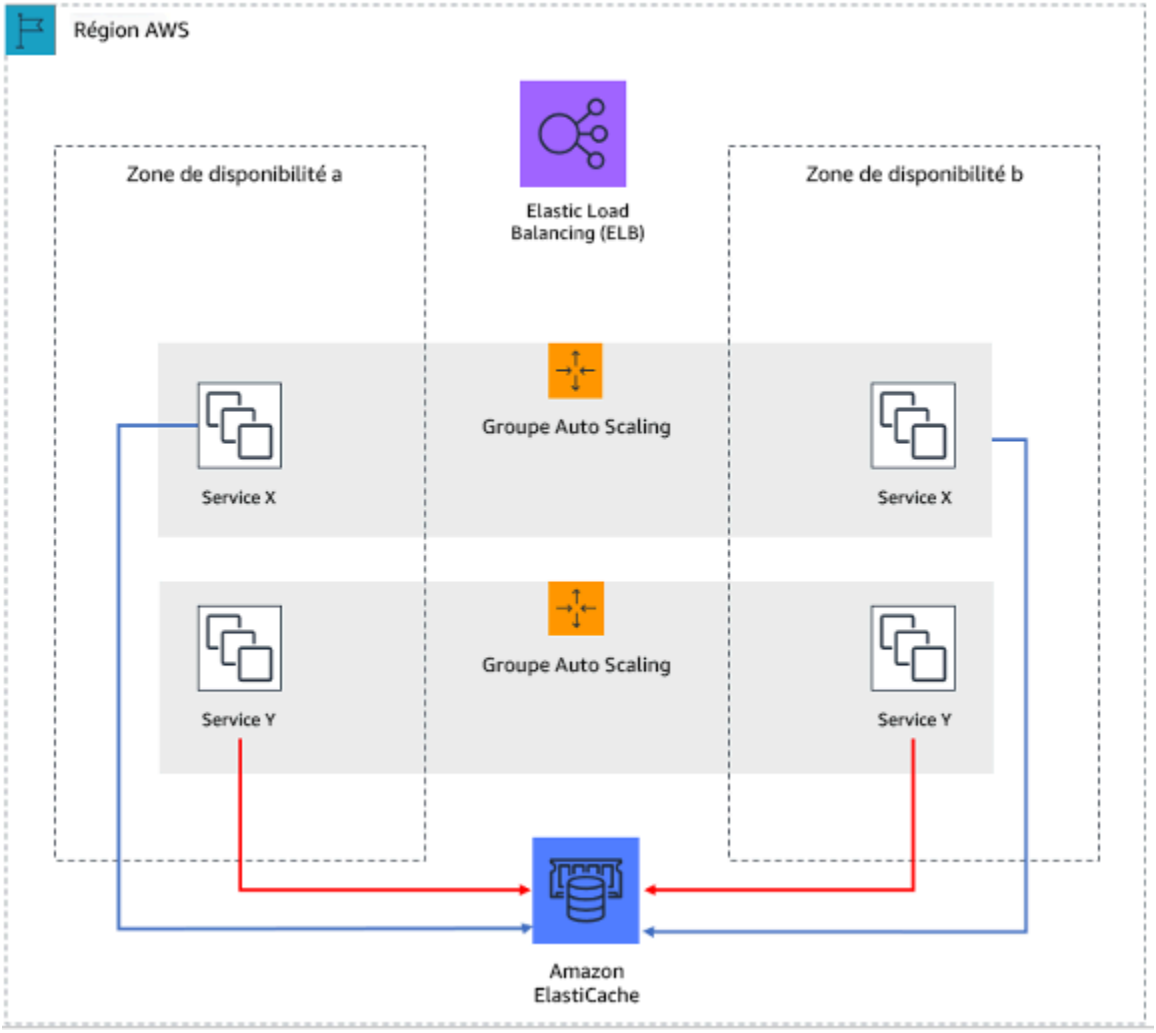


Figure 7 : Dans cette application Web sans état, l'état de la session est déchargé vers Amazon ElastiCache.

Lorsque des utilisateurs ou des services interagissent avec une application, ils exécutent souvent une série d'interactions qui forment une session. Une session correspond aux données uniques des utilisateurs qui persistent entre les requêtes pendant l'utilisation de l'application. Une application sans état n'a pas besoin de connaître les interactions précédentes et ne stocke pas d'informations de session.

Lorsqu'une application est conçue pour être sans état, vous pouvez utiliser des services de calcul sans serveur, comme AWS Lambda ou AWS Fargate.

Outre le remplacement du serveur, les applications sans état ont également pour avantage de pouvoir être mises à l'échelle horizontalement, car toutes les ressources de calcul disponibles (telles que les instances EC2 et les fonctions AWS Lambda) peuvent répondre à n'importe quelle requête.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Rendre vos applications sans état. Les applications sans état permettent une mise à l'échelle horizontale et tolèrent la défaillance d'un nœud individuel.
  - Supprimez les états qui peuvent être stockés dans les paramètres de demande.
  - Après avoir vérifié si l'état est requis, déplacez n'importe quel suivi d'état vers un cache ou un magasin de données multizone résistant comme Amazon ElastiCache, Amazon RDS, Amazon DynamoDB ou une autre solution de données distribuée. Enregistrez un état qui n'a pas pu être déplacés vers des magasins de données résilient.
    - Certaines données (comme les cookies) peuvent être transmises dans les en-têtes ou les paramètres de requête
    - Réfactorisez afin de supprimer l'état qui peut rapidement être transmis dans les requêtes
    - Certaines données ne sont pas forcément nécessaires pour certaines requêtes et peuvent être récupérées à la demande.
    - Supprimez les données qui peuvent être récupérées de manière asynchrone.
    - Choisissez un magasin de donnée qui répond aux exigences relatives à l'état requis.
    - Pensez à avoir une base de données NoSQL pour les données non relationnelles.

Ressources

Documents connexes :

- L'Amazon Builders' Library : éviter le basculement dans les systèmes distribués
- L'Amazon Builders' Library : éviter les retards de file d'attente insurmontables
- L'Amazon Builders' Library : défis et stratégies de mise en cache

REL05-BP07 Mettre en place des leviers d'urgence

Les leviers d'urgence sont des processus rapides qui peuvent réduire l'impact sur la disponibilité de votre charge de travail.

Les leviers d'urgence fonctionnent en désactivant, en limitant ou en modifiant le comportement des composants ou des dépendances à l'aide de mécanismes connus et testés. Ils permettent d'atténuer les perturbations de la charge de travail causées par l'épuisement des ressources dû à une

augmentation inattendue de la demande et de réduire l'impact des défaillances des composants non stratégiques de votre charge de travail.

Résultat souhaité : en mettant en place des leviers d'urgence, vous pouvez établir des processus dont le fonctionnement a été vérifié pour préserver la disponibilité des composants stratégiques de votre charge de travail. La charge de travail devrait se dégrader de manière appropriée et continuer à remplir ses fonctions stratégiques durant l'activation d'un levier d'urgence. Pour plus d'informations sur la dégradation appropriée, consultez REL05-BP01 Implémenter une dégradation appropriée pour transformer les dépendances matérielles applicables en dépendances logicielles.

Anti-modèles courants :

- La défaillance des dépendances non stratégiques a un impact sur la disponibilité de votre charge de travail principale.
- Le comportement des composants stratégiques n'est pas testé ou vérifié lors d'une défaillance d'un composant non stratégique.
- Aucun critère clair et déterministe n'a été défini pour l'activation ou la désactivation d'un levier d'urgence.

Avantages liés au respect de cette bonne pratique : la mise en place de leviers d'urgence peut améliorer la disponibilité des composants stratégiques de votre charge de travail en fournissant à vos résolveurs des processus établis pour répondre à des pics de demande imprévus ou à des défaillances des dépendances non stratégiques.

Niveau de risque exposé si cette bonne pratique n'est pas établie: moyen

Directives d'implémentation

- Identifier les composants stratégiques de votre charge de travail.
- Concevoir et construire les composants stratégiques de votre charge de travail de manière à ce qu'ils résistent aux défaillances des composants non stratégiques.
- Effectuer des tests pour valider le comportement de vos composants stratégiques en cas de défaillance des composants non stratégiques.
- Définir et surveiller des métriques ou des déclencheurs pertinents pour lancer des procédures de levier d'urgence.
- Définir les procédures (manuelles ou automatisées) qui comprennent le levier d'urgence.

Étapes d'implémentation

- Identifier les composants stratégiques de votre charge de travail.
  - Chaque composant technique de votre charge de travail doit être associé à la fonction commerciale correspondante et classé comme stratégique ou non stratégique. Pour des exemples de fonctionnalités stratégiques et non stratégiques chez Amazon, consultez Any Day Can Be Prime Day : How Amazon.com Search Uses Chaos Engineering to handle over 84K requests per second.
  - Il s'agit d'une décision à la fois technique et commerciale, qui varie en fonction de l'organisation et de la charge de travail.
- Concevoir et construire les composants stratégiques de votre charge de travail de manière à ce qu'ils résistent aux défaillances des composants non stratégiques.
  - Lors de l'analyse des dépendances, tenez compte de tous les modes de défaillance potentiels et vérifiez que vos mécanismes de levier d'urgence fournissent les fonctionnalités stratégiques aux composants en aval.
- Effectuer des tests pour valider le comportement de vos composants stratégiques pendant l'activation de vos leviers d'urgence.
  - Éviter les comportements bimodaux. Pour plus d'informations, consultez REL11-BP05 Utiliser la stabilité statique pour éviter les comportements bimodaux.
- Définir et surveiller des métriques pertinentes pour lancer des procédures de levier d'urgence.
  - La recherche des bonnes métriques à surveiller dépend de votre charge de travail. Parmi les métriques, citons la latence ou le nombre de demandes infructueuses à une dépendance.
- Définir les procédures (manuelles ou automatisées) qui comprennent le levier d'urgence.
  - Il peut s'agir de mécanismes tels que le délestage de charge, la limitation des requêtes ou la mise en œuvre d'une dégradation appropriée.

Ressources

Bonnes pratiques associées :

- REL05-BP01 Implémenter une dégradation appropriée pour transformer les dépendances matérielles applicables en dépendances logicielles
- REL05-BP02 Limiter les demandes
- REL11-BP05 Utiliser la stabilité statique pour éviter les comportements bimodaux

Documents connexes :

- Automatisation de déploiements sécurisés sans intervention
- Any Day Can Be Prime Day: How Amazon.com Search Uses Chaos Engineering to Handle Over 84K Requests Per Second

Vidéos connexes :

- AWS re:Invent 2020: Reliability, consistency, and confidence through immutability

Gestion des modifications

Questions

- FIA 6. Comment surveillez-vous les ressources de charge de travail ?
- FIA 7. Comment concevez-vous votre charge de travail pour s'adapter aux changements de demande ?
- FIA 8. Comment implémenter les modifications ?

FIA 6. Comment surveillez-vous les ressources de charge de travail ?

Les journaux et les métriques sont de puissants outils pour obtenir informations sur l'état de votre charge de travail. Vous pouvez configurer votre charge de travail de sorte à surveiller les journaux et les métriques et envoyer des notifications lorsque les seuils sont franchis ou que des événements significatifs se produisent. La surveillance permet à votre charge de travail de reconnaître quand des seuils de faibles performances sont franchis ou quand des défaillances se produisent, afin d'y répondre par une récupération automatique.

Bonnes pratiques

- REL06-BP01 Surveiller tous les composants de la charge de travail (génération)
- REL06-BP02 Définir et calculer des métriques (agrégation)
- REL06-BP03 Envoyer des notifications (traitement et alarmes en temps réel)
- REL06-BP04 Automatiser les réponses (traitement et alarmes en temps réel)
- REL06-BP05 Analytique
- REL06-BP06 Procéder à des examens réguliers
- REL06-BP07 Surveiller la traçabilité complète des demandes via votre système

REL06-BP01 Surveiller tous les composants de la charge de travail (génération)

Surveillez les composants de la charge de travail avec Amazon CloudWatch ou des outils tiers. Surveillez les services AWS avec le tableau de bord AWS Health.

Tous les composants de votre charge de travail doivent être surveillés, y compris le côté utilisateur, la logique métier et les niveaux de stockage. Au besoin, définissez des métriques clés, décrivez leur procédure d'extraction des journaux, puis spécifiez des seuils de déclenchement pour les événements d'alarme correspondants Assurez-vous que les métriques sont pertinentes pour les indicateurs clés de performance (KPI) de votre charge de travail, et utilisez des métriques et des journaux pour identifier les signes avant-coureurs de la dégradation du service. Par exemple, une métrique liée aux résultats commerciaux, telle que le nombre de commandes traitées avec succès par minute, peut indiquer des problèmes de charge de travail plus rapidement qu'une métrique technique, telle que l'utilisation du processeur. Utilisez le tableau de bord AWS Health pour obtenir une vue personnalisée des performances et de la disponibilité des services AWS sous-jacents à vos ressources AWS.

La surveillance dans le cloud offre de nouvelles opportunités. La plupart des fournisseurs de cloud ont développé des hooks personnalisables et peuvent fournir des informations pour vous aider à surveiller plusieurs couches de votre charge de travail. Des services AWS comme Amazon CloudWatch appliquent des algorithmes statistiques et de machine learning pour analyser en continu les métriques des systèmes et des applications, déterminer les points de référence normaux et détecter les anomalies avec une intervention minimale de l'utilisateur. Les algorithmes de détection d'anomalies tiennent compte de la saisonnalité et des changements de tendance des métriques.

AWS met à disposition une multitude d'informations de surveillance et de journalisation qui peuvent être utilisées pour définir des métriques spécifiques à la charge de travail et des processus de changement de la demande, et pour adopter des techniques de machine learning, quelle que soit l'expertise en ML.

En outre, surveillez l'ensemble de vos points de terminaison externes afin de vous assurer qu'ils sont indépendants de votre implémentation de base. Cette surveillance active peut être effectuée avec des transactions synthétiques (parfois appelées tests canary utilisateur, mais à ne pas confondre avec les déploiements canary), qui exécutent périodiquement plusieurs tâches communes correspondant aux actions effectuées par les clients de la charge de travail. Maintenez ces tâches de courte durée et veillez à ne pas surcharger votre charge de travail pendant les tests. Amazon CloudWatch Synthetics vous permet de créer des tests canary synthétiques pour surveiller vos points de terminaison et vos API. Vous pouvez également combiner les nœuds de clients synthétiques

Canari avec la console AWS X-Ray pour identifier les scripts Canari synthétiques qui rencontrent des erreurs, des pannes ou des taux de limitation au cours de la période sélectionnée.

Résultat souhaité :

Collectez et utilisez des métriques critiques de tous les composants de la charge de travail pour garantir la fiabilité de la charge de travail et une expérience utilisateur optimale. Détecter qu'une charge de travail n'atteint pas les résultats vous permet de déclarer rapidement un sinistre et de vous remettre d'un incident.

Anti-modèles courants :

- Surveillance limitée aux interfaces externes de votre charge de travail.
- Ne pas générer de métriques spécifiques à la charge de travail et s'appuyer uniquement sur les métriques qui vous sont fournies par les services AWS utilisés par votre charge de travail.
- Utiliser uniquement des métriques techniques dans votre charge de travail et ne surveiller aucune métrique liée aux KPI non techniques auxquels la charge de travail contribue.
- S'appuyer sur le trafic de production et de simples vérifications de l'état pour surveiller et évaluer l'état de la charge de travail.

Avantages liés au respect de cette bonne pratique : La surveillance à tous les niveaux de votre charge de travail vous permet d'anticiper et de résoudre plus rapidement les problèmes dans les composants qui constituent la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

1. Activer la journalisation si disponible. Les données de surveillance doivent être obtenues à partir de tous les composants des charges de travail. Activez la journalisation supplémentaire, telle que les journaux d'accès S3, et autorisez votre charge de travail à consigner des données qui lui sont spécifiques. Collectez des métriques pour les moyennes d'UC, d'E/S réseau et d'E/S disque à partir de services comme Amazon ECS, Amazon EKS, Amazon EC2, Elastic Load Balancing, AWS Auto Scaling et Amazon EMR. Consulter Services AWS publiant des métriques CloudWatch pour obtenir une liste des services AWS qui publient des métriques CloudWatch.
2. Passez en revue toutes les métriques par défaut et explorez toutes les lacunes de collecte de données. Chaque service génère des métriques par défaut. La collecte des métriques par défaut vous permet de mieux comprendre les dépendances entre les composants de charge de travail et

sur la manière dont la fiabilité et les performances des composants affectent la charge de travail. Vous pouvez également créer et publier vos propres métriques vers CloudWatch à l'aide d'AWS CLI ou d'une API. Évaluez

3. toutes les métriques pour décider celles pour lesquelles envoyer des alertes pour chaque service AWS de votre charge de travail. Vous pouvez choisir de sélectionner un sous-ensemble de métriques qui ont un impact majeur sur la fiabilité de la charge de travail. Se concentrer sur les métriques critiques et le seuil vous permet d'affiner le nombre d'information et peut aider à minimiser les faux positifs.
4. Définissez des alertes et le processus de récupération de votre charge de travail après le déclenchement de l'alerte. La définition d'alertes vous permet de notifier, de faire remonter et de suivre rapidement les étapes nécessaires pour vous remettre d'un incident et atteindre votre objectif de temps de récupération (RTO) prescrit. Vous pouvez utiliser des alarmes Amazon CloudWatch pour appeler des flux de travail automatisés et lancer des procédures de récupération en fonction de seuils définis.
5. Explorez l'utilisation de transactions synthétiques pour collecter des données pertinentes sur l'état des charges de travail. La surveillance synthétique suit les mêmes routes et effectue les mêmes actions qu'un client, ce qui vous permet de vérifier en permanence l'expérience client même lorsque vous n'avez aucun trafic client sur vos charges de travail. En utilisant les transactions synthétiques, vous pouvez découvrir les problèmes avant vos clients.

Ressources

Bonnes pratiques associées :

- REL11-BP03 Automatiser la réparation sur toutes les couches

Documents connexes :

- Premiers pas avec le tableau de bord AWS Health : état de votre compte
- Services AWS publiant des métriques CloudWatch
- Journaux d'accès pour votre Network Load Balancer
- Journaux d'accès pour votre Application Load Balancer
- Accès à Amazon CloudWatch Logs pour AWS Lambda
- Journalisation de l'accès au serveur Amazon S3
- Activer les journaux d'accès pour votre Classic Load Balancer

- Exportation de données de journal vers Amazon S3
- Installer l'agent CloudWatch sur une instance Amazon EC2
- Publication des métriques personnalisées
- Fonctionnement des tableaux de bord Amazon CloudWatch
- Utilisation des métriques Amazon CloudWatch
- Utilisation de scripts Canary (Amazon CloudWatch Synthetics)
- Que sont les Amazon CloudWatch Logs ?

Guides de l'utilisateur :

- Création d'un journal d'activité
- Surveillance des métriques de mémoire et de disque pour les instances Linux Amazon Amazon EC2
- Utiliser CloudWatch Logs avec des instances de conteneur
- Journaux de flux VPC
- Qu'est-ce qu'Amazon DevOps Guru ?
- Qu'est-ce que AWS X-Ray ?

Blogs connexes :

- Débogage avec Amazon CloudWatch Synthetics et AWS X-Ray

Exemples et ateliers connexes :

- Ateliers AWS Well-Architected : Excellence opérationnelle - Surveillance des dépendances
- L'Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle
- Atelier sur l'observabilité

REL06-BP02 Définir et calculer des métriques (agrégation)

Stockez les données des journaux et appliquez des filtres si nécessaire pour calculer les métriques, en particulier le décompte d'un événement de journal spécifique ou la latence calculée à partir des horodatages des événements de journaux.

Amazon CloudWatch et Amazon S3 servent de couches principales pour l'agrégation et le stockage. Pour certains services, comme AWS Auto Scaling et Elastic Load Balancing, les métriques par défaut sont fournies par défaut pour la charge du processeur ou la latence moyenne des demandes au sein d'un cluster ou d'une instance. Pour les services de streaming, comme les journaux de flux VPC et AWS CloudTrail, les données d'événement sont transmises à CloudWatch Logs et vous devez définir et appliquer des filtres de métrique pour extraire des métriques à partir des données d'événement. Cela vous donne des données de séries chronologiques, qui peuvent servir d'entrées aux alarmes CloudWatch que vous définissez pour déclencher les alertes.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Définir et calculer des métriques (agrégation). Stockez les données des journaux et appliquez des filtres si nécessaire pour calculer les métriques, en particulier le décompte d'un événement de journal spécifique ou la latence calculée à partir des horodatages des événements de journaux
  - Les filtres de métrique définissent les termes et les modèles à rechercher dans les données de journal lorsqu'elles sont envoyées à CloudWatch Logs. CloudWatch Logs utilise ces filtres de métriques pour transformer les données de journal en métriques numériques CloudWatch que vous pouvez représenter graphiquement ou au niveau desquelles vous pouvez définir une alarme.
    - Recherche et filtrage des données de journaux
  - Utiliser un agent tiers de confiance pour agréger les journaux
    - Suivez les instructions de l'agent tiers. La plupart des produits tiers s'intègrent à CloudWatch et Amazon S3.
  - Certains services AWS peuvent publier des journaux directement dans Amazon S3. Ainsi, si votre principale exigence pour les journaux est le stockage dans Amazon S3, vous pouvez facilement faire en sorte que le service produisant les journaux les envoie directement à Amazon S3 sans configurer d'infrastructure supplémentaire.
    - Envoi des journaux directement à Amazon S3

Ressources

Documents connexes :

- Exemples de requêtes Amazon CloudWatch Logs Insights
- Débogage avec Amazon CloudWatch Synthetics et AWS X-Ray

- Un atelier sur l'observabilité
- Recherche et filtrage des données de journaux
- Envoi des journaux directement à Amazon S3
- L'Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle

REL06-BP03 Envoyer des notifications (traitement et alarmes en temps réel)

Lorsque les organisations détectent des problèmes potentiels, elles envoient des notifications et des alertes en temps réel au personnel et aux systèmes appropriés afin de résoudre rapidement et efficacement ces problèmes.

Résultat souhaité : Il est possible d'obtenir des réponses rapides aux événements opérationnels en configurant des alarmes pertinentes basées sur les métriques de service et d'application. Lorsque les seuils d'alarme sont dépassés, le personnel et les systèmes appropriés sont avertis afin de résoudre les problèmes sous-jacents.

Anti-modèles courants :

- Vous configurez les alarmes avec un seuil trop élevé, ce qui fait échouer l'envoi des notifications vitales.
- Vous configurez les alarmes avec un seuil trop bas, ce qui empêche la prise en compte des alertes importantes à cause du bruit généré par un trop grand nombre de notifications.
- Vous ne mettez pas à jour les alarmes et leur seuil en cas de changement d'utilisation.
- Pour les alarmes qu'il est préférable de traiter par des actions automatisées, vous envoyez la notification au personnel au lieu de générer l'action automatisée, ce qui entraîne l'envoi d'un trop grand nombre de notifications.

Avantages liés au respect de cette bonne pratique : En envoyant des notifications et des alertes en temps réel au personnel et aux systèmes appropriés, vous pouvez détecter rapidement les problèmes et réagir rapidement face aux incidents opérationnels.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les charges de travail doivent être équipées d'un système de traitement et d'avertissement en temps réel afin d'améliorer la détectabilité des problèmes susceptibles d'affecter la disponibilité de

l'application et de déclencher une réponse automatique. Les organisations peuvent procéder au traitement et à l'avertissement en temps réel en créant des alertes avec des métriques définies afin de recevoir des notifications chaque fois que des événements importants se produisent ou qu'une métrique dépasse un seuil.

Amazon CloudWatch vous permet de créer des alarmes de métrique et des alarmes composites au moyen d'alarmes CloudWatch basées sur le seuil statique, la détection des anomalies et d'autres critères. Pour plus de détails sur les types d'alarmes que vous pouvez configurer avec CloudWatch, consultez la section relative aux alarmes dans la documentation CloudWatch. »

Vous pouvez créer des vues personnalisées des métriques et des alertes de vos ressources AWS pour vos équipes en utilisant les tableaux de bord CloudWatch. » Les pages d'accueil personnalisables de la console CloudWatch vous permettent de surveiller vos ressources dans une vue unique sur plusieurs régions.

Les alarmes peuvent effectuer une ou plusieurs actions, comme envoyer une notification à une rubrique Amazon SNS, exécuter une action Amazon EC2 ou une action Amazon EC2 Auto Scaling , ou créer un OpsItem ou de réponse dans AWS Systems Manager.

Amazon CloudWatch utilise Amazon SNS pour envoyer des notifications lorsque l'alarme change d'état, ce qui permet de transmettre les messages des diffuseurs de publication (producteurs) aux abonnés (consommateurs). Pour plus de détails sur la configuration des notifications Amazon SNS, consultez Configuration d'Amazon SNS. »

CloudWatch envoie des événements EventBridge 43 % chaque fois qu'une alarme CloudWatch est créée, mise à jour, supprimée ou que son état change. Vous pouvez utiliser EventBridge avec ces événements pour créer des règles qui exécutent des actions, comme vous avertir chaque fois que l'état d'une alarme change ou déclencher automatiquement des événements sur votre compte à l'aide de l'automatisation Systems Manager. »

Quand utiliser EventBridge ou Amazon SNS ?

EventBridge et Amazon SNS peuvent être utilisés pour développer des applications pilotées par des événements. Votre choix dépendra de vos besoins spécifiques.

Amazon EventBridge est recommandé lorsque vous souhaitez créer une application qui réagit aux événements de vos propres applications, des applications SaaS et des services AWS. EventBridge est le seul service basé sur les événements qui s'intègre directement aux partenaires SaaS tiers. EventBridge ingère également automatiquement les événements de plus de 200 services AWS sans que les développeurs n'aient à créer de ressources sur leur compte.

EventBridge utilise une structure définie basée sur JSON pour les événements et vous aide à créer des règles qui s'appliquent à l'ensemble du corps de l'événement afin de sélectionner les événements à transférer vers une cible. EventBridge prend actuellement en charge plus de 20 services AWS en tant que cibles, y compris AWS Lambda, Amazon SQS, Amazon SNS, Amazon Kinesis Data Streamset Amazon Data Firehose. »

Amazon SNS est recommandé pour les applications qui nécessitent un niveau de diffusion élevé (des milliers, voire des millions de points de terminaison). Il est courant d'observer que les clients utilisent Amazon SNS comme cible pour leur règle afin de filtrer les événements dont ils ont besoin et de les diffuser sur plusieurs points de terminaison.

Les messages ne sont pas structurés et peuvent être de n'importe quel format. Amazon SNS prend en charge le transfert de messages vers six types de cibles différents, notamment Lambda, Amazon SQS, les points de terminaison HTTP/S, les SMS, les notifications push mobiles et les e-mails. La latence type de Amazon SNS est inférieure à 30 millisecondes. » Un large éventail de services AWS envoie des messages Amazon SNS en configurant le service dans cet objectif (plus de 30, y compris Amazon EC2, Amazon S3et Amazon RDS).

Étapes d'implémentation

1. Créez une alarme à l'aide d'alarmes Amazon CloudWatch. »
   a. Une alarme de métrique surveille une seule métrique CloudWatch ou une expression qui dépend de métriques CloudWatch. L'alarme déclenche une ou plusieurs actions en fonction de la valeur de la métrique ou de l'expression par rapport à un seuil sur un certain nombre d'intervalles de temps. L'action peut consister à envoyer une notification à une rubrique Amazon SNS, exécuter une action Amazon EC2 ou une action Amazon EC2 Auto Scaling , ou créer un OpsItem ou de réponse dans AWS Systems Manager.
   b. Une alarme composite est une expression de règle qui prend en compte les conditions d'alarme des autres alarmes que vous avez créées. L'alarme composite ne passe en état d'alarme que si toutes les conditions de la règle sont satisfaites. Les alarmes spécifiées dans l'expression de règle d'une alarme composite peuvent inclure des alarmes de métrique et des alarmes composites supplémentaires. Les alarmes composites peuvent envoyer des notifications Amazon SNS lorsque leur état change et peuvent créer des Systems Manager OpsItems ou des incidents lorsqu'ils entrent en état d'alarme, mais qu'ils ne peuvent effectuer aucune action Amazon EC2 ou Auto Scaling.
2. Configurez les notifications Amazon SNS. » Lorsque vous créez une alarme CloudWatch, vous pouvez inclure une rubrique Amazon SNS pour envoyer une notification lorsque l'alarme change d'état.

3. Créez des règles dans EventBridge qui correspondent aux alarmes CloudWatch spécifiées. Chaque règle prend en charge plusieurs cibles, y compris des fonctions Lambda. Par exemple, vous pouvez définir une alarme qui se déclenche lorsque l'espace disque disponible est insuffisant, ce qui déclenche une fonction Lambda par le biais d'une règle EventBridge permettant de libérer de l'espace. Pour plus de détails sur les cibles EventBridge, consultez Cibles EventBridge (langue française non garantie).. »

Ressources

Bonnes pratiques Well-Architected connexes :

- REL06-BP01 Surveiller tous les composants de la charge de travail (génération)
- REL06-BP02 Définir et calculer des métriques (agrégation)
- REL12-BP01 Utiliser des playbooks pour enquêter sur les causes des défaillances

Documents connexes :

- Amazon CloudWatch
- Informations CloudWatch Logs
- Utilisation des alarmes Amazon CloudWatch
- Fonctionnement des tableaux de bord Amazon CloudWatch
- Utilisation des métriques Amazon CloudWatch
- Configuration de notifications Amazon SNS (langue française non garantie)
- la détection des anomalies CloudWatch
- Protection des données CloudWatch Logs
- Amazon EventBridge
- Amazon Simple Notification Service

Vidéos connexes :

- reinvent 2022 observability videos
- AWS re:Invent 2022 - Observability best practices at Amazon

Exemples connexes :

- Un atelier sur l'observabilité
- Amazon EventBridge à AWS Lambda avec contrôle du feedback par des alarmes Amazon CloudWatch

REL06-BP04 Automatiser les réponses (traitement et alarmes en temps réel)

utilisez l'automatisation pour agir en cas de détection d'événement, par exemple, pour remplacer les composants défectueux.

Un traitement automatique en temps réel des alarmes est mis en œuvre afin que les systèmes puissent prendre rapidement des mesures correctives et tenter d'éviter les pannes ou une dégradation du service lorsque les alarmes se déclenchent. Les réponses automatisées aux alarmes peuvent inclure le remplacement des composants défaillants, l'ajustement de la capacité de calcul, la redirection du trafic vers des hôtes, des zones de disponibilité ou d'autres régions en bonne santé, et la notification des opérateurs.

Résultat souhaité : des alarmes en temps réel sont identifiées et un traitement automatisé des alarmes est mis en place pour déclencher les actions appropriées afin de maintenir les objectifs de niveau de service et les contrats de niveau de service (SLA). L'automatisation peut aller de l'autoréparation de composants individuels au basculement complet du site.

Anti-modèles courants :

- Pas d'inventaire ou de catalogue clair des principales alarmes en temps réel.
- Aucune réponse automatique aux alarmes critiques (par exemple, lorsque la capacité de calcul est presque épuisée, une mise à l'échelle automatique se produit).
- Réponses aux alarmes contradictoires.
- Pas de procédure opérationnelle standard (SOP) que les opérateurs doivent suivre lorsqu'ils reçoivent des notifications d'alerte.
- Pas de surveillance des modifications de configuration, alors que des changements de configuration non détectés peuvent entraîner des temps d'arrêt pour les charges de travail.
- Pas de stratégie pour annuler les modifications de configuration involontaires.

Avantages liés au respect de cette bonne pratique : l'automatisation du traitement des alarmes peut améliorer la résilience du système. Le système prend automatiquement des mesures correctives, réduisant ainsi les activités manuelles qui nécessitent des interventions humaines sujettes aux

erreurs. L'exécution de la charge de travail permet d'atteindre les objectifs de disponibilité et de réduire les interruptions de service.

Niveau de risque exposé si cette bonne pratique n'est pas établie: moyen

Directives d'implémentation

Pour gérer efficacement les alertes et automatiser leur réponse, classez les alertes en fonction de leur criticité et de leur impact, documentez les procédures de réponse et planifiez les réponses avant de classer les tâches.

Identifiez les tâches nécessitant des actions spécifiques (souvent détaillées dans les runbooks) et examinez tous les runbooks et playbooks pour déterminer les tâches qui peuvent être automatisées. Si les actions peuvent être définies, alors elles sont souvent automatisables. Si elles ne le sont pas, documentez les étapes manuelles dans une SOP et formez les opérateurs à ces étapes. Remettez continuellement en question les processus manuels pour trouver des opportunités d'automatisation où vous pouvez établir et maintenir un plan d'automatisation des réponses aux alertes.

Étapes d'implémentation

1. Dresser un inventaire des alarmes : pour obtenir la liste de toutes les alarmes, vous pouvez utiliser AWS CLI à l'aide de la commande Amazon CloudWatch `describe-alarms`. Selon le nombre d'alarmes que vous avez configurées, vous devrez peut-être utiliser la pagination pour récupérer un sous-ensemble d'alarmes pour chaque appel, ou vous pouvez utiliser le SDK AWS pour obtenir les alarmes à l'aide d'un appel d'API.
2. Documenter les actions de toutes les alarmes : tenez à jour un runbook avec toutes les alarmes et leurs actions, qu'elles soient manuelles ou automatisées. AWS Systems Manager fournit des runbooks prédéfinis. Pour plus d'informations sur les runbooks, consultez Créer vos propres runbooks. Pour plus d'informations sur l'affichage du contenu du runbook, consultez Afficher le contenu du runbook.
3. Configurer et gérer les actions des alarmes : pour toutes les alarmes nécessitant une action, spécifiez l'action automatisée à l'aide du SDK CloudWatch. Par exemple, vous pouvez modifier automatiquement l'état de vos instances Amazon EC2 basées sur une alarme CloudWatch en créant des actions sur l'alarme et en les activant ou désactivant.

   Vous pouvez également utiliser Amazon EventBridge pour répondre automatiquement aux événements du système (problèmes de disponibilité d'une application ou modifications de ressources, par exemple). Vous pouvez créer des règles pour indiquer les événements qui vous

intéressent et les mesures à prendre lorsqu'un événement correspond à une règle. Les actions qui peuvent être lancées automatiquement incluent l'appel d'une fonction AWS Lambda, l'appel d'Amazon EC2 `Run Command`, le relais de l'événement à Amazon Kinesis Data Streams et l'affichage d'Automatiser Amazon EC2 en utilisant EventBridge.

4. Procédures opérationnelles standard (SOP) : en fonction des composants de votre application, AWS Resilience Hub recommande plusieurs modèles de SOP. Vous pouvez utiliser ces SOP pour documenter tous les processus qu'un opérateur doit suivre en cas d'alerte. Vous pouvez également créer une SOP en fonction des recommandations d'Resilience Hub, lorsque vous avez besoin d'une application Resilience Hub avec une stratégie de résilience associée, ainsi que d'une évaluation historique de la résilience de cette application. Les recommandations de SOP sont générées par l'évaluation de la résilience.

   Resilience Hub travaille de pair avec Systems Manager pour automatiser les étapes de vos SOP en fournissant un certain nombre de documents SSM que vous pouvez utiliser comme référence pour ces SOP. Par exemple, Resilience Hub peut recommander une procédure SOP pour ajouter de l'espace disque sur la base d'un document d'automatisation SSM existant.
5. Effectuer des actions automatisées avec Amazon DevOps Guru : vous pouvez utiliser Amazon DevOps Guru pour surveiller automatiquement les ressources de votre application afin de détecter des comportements anormaux et fournir des recommandations ciblées afin d'accélérer l'identification des problèmes et les délais de résolution. Avec DevOps Guru, vous pouvez surveiller les flux de données opérationnelles de plusieurs sources en temps quasi réel, notamment des métriques Amazon CloudWatch, AWS ConfigAWS CloudFormation et AWS X-Ray. Vous pouvez également utiliser DevOps Guru pour créer automatiquement des OpsItems dans OpsCenter et envoyer des événements à EventBridge pour une automatisation supplémentaire.

Ressources

Bonnes pratiques associées :

- REL06-BP01 Surveiller tous les composants de la charge de travail (génération)
- REL06-BP02 Définir et calculer des métriques (agrégation)
- REL06-BP03 Envoyer des notifications (traitement et alarmes en temps réel)
- REL08-BP01 Utiliser des runbooks pour les activités standard telles que le déploiement

Documents connexes :

- AWS Systems Manager Automation
- Creating an EventBridge Rule That Triggers on an Event from an AWS Resource
- One Observability Workshop
- Amazon Builders' Library : Instrumenter les systèmes distribués pour une visibilité opérationnelle
- What is Amazon DevOps Guru?
- Utilisation des documents d'automatisation (playbooks)

Vidéos connexes :

- AWS re:Invent 2022 - Observability best practices at Amazon
- AWS re:Invent 2020: Automate anything with AWS Systems Manager
- Introduction to AWS Resilience Hub
- Create Custom Ticket Systems for Amazon DevOps Guru Notifications
- Enable Multi-Account Insight Aggregation with Amazon DevOps Guru

Exemples connexes :

- Ateliers sur la fiabilité
- Amazon CloudWatch and Systems Manager Workshop

REL06-BP05 Analytique

collectez les fichiers journaux et les historiques de métriques, puis analysez-les pour obtenir des informations plus générales sur les tendances et la charge de travail.

Amazon CloudWatch Logs Insights prend en charge un langage de requête simple et puissant que vous pouvez utiliser pour analyser les données des journaux. Amazon CloudWatch Logs prend également en charge les abonnements qui permettent aux données de circuler en toute transparence vers Amazon S3, où vous pouvez utiliser Amazon Athena pour interroger les données. Il prend également en charge les requêtes dans une grande variétés de formats. Consulter Formats de données et SerDes pris en charge dans le guide de l'utilisateur Amazon Athena pour plus d'informations. Pour analyser des ensembles de fichiers journaux volumineux, vous pouvez exécuter un cluster Amazon EMR pour exécuter des analyses à l'échelle du pétaoctet.

Il existe divers outils fournis par les partenaires AWS et les tiers qui permettent l'agrégation, le traitement, le stockage et l'analyse. Parmi ces outils figurent New Relic, Splunk, Loggly,

Logstash, CloudHealth et Nagios. Cependant, la génération en dehors du système et des journaux d'applications est propre à chaque fournisseur de cloud, et généralement, spécifique à chaque service.

Un partie souvent négligée de la surveillance des processus concerne la gestion des données. Vous devez déterminer les exigences de rétention des données de surveillance, puis appliquer des stratégies de cycle de vie en conséquence. Amazon S3 prend en charge la gestion du cycle de vie au niveau du compartiment S3. Cette gestion du cycle de vie peut être appliquée différemment à d'autres chemins dans le compartiment. Vers la fin du cycle de vie, vous pouvez transférer des données dans Amazon S3 Glacier pour un stockage à long terme, puis les laisser expirer une fois la fin de la période de rétention terminée. La classe de stockage S3 Intelligent-Tiering est conçue pour optimiser les coûts en transférant automatiquement les données vers le niveau d'accès le plus économique, sans impact sur les performances ni surcharge opérationnelle.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- CloudWatch Logs vous permet de rechercher et d'analyser de manière interactive vos données de journaux dans Amazon CloudWatch Logs.
  - Analyse des données des journaux avec CloudWatch Logs Insights
  - Exemples de requêtes Amazon CloudWatch Logs Insights
- Utiliser Amazon CloudWatch Logs pour envoyer des journaux vers Amazon S3 où vous pouvez les exploiter ou utiliser Amazon Athena pour interroger les données
  - Comment analyser mes journaux d'accès au serveur Amazon S3 à l'aide d'Athena ?
    - Créez une stratégie de cycle de vie S3 pour votre compartiment de journaux d'accès au serveur. Configurez la stratégie de cycle de vie de sorte à supprimer régulièrement les fichiers journaux. Cette suppression permet de réduire la quantité de données analysées par Athena pour chaque requête.
      - Comment créer une stratégie de cycle de vie pour un compartiment S3 ?

Ressources

Documents connexes :

- Exemples de requêtes Amazon CloudWatch Logs Insights
- Analyse des données des journaux avec CloudWatch Logs Insights

- Débogage avec Amazon CloudWatch Synthetics et AWS X-Ray
- Comment créer une stratégie de cycle de vie pour un compartiment S3 ?
- Comment analyser mes journaux d'accès au serveur Amazon S3 à l'aide d'Athena ?
- Un atelier sur l'observabilité
- L'Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle

REL06-BP06 Procéder à des examens réguliers

Examinez fréquemment comment la surveillance de la charge de travail est mise en œuvre et mettez-la à jour en fonction des événements et des modifications majeurs.

Une surveillance efficace repose sur des métriques commerciales clés. Assurez-vous que ces métriques sont prises en compte dans votre charge de travail au fur et à mesure que les priorités de l'entreprise évoluent.

Auditer votre surveillance vous permet de savoir avec certitude quand une application est conforme à ses objectifs de disponibilité. Pour pouvoir analyser les causes premières, il faut pouvoir découvrir ce qui se passe lorsque des défaillances se produisent. AWS fournit des services qui vous permettent de suivre l'état de vos services lors d'un incident :

- Amazon CloudWatch Logs : vous pouvez stocker vos journaux dans ce service et inspecter leur contenu.
- Amazon CloudWatch Logs Insights : service entièrement géré qui vous permet d'analyser des journaux volumineux en quelques secondes. Il permet des requêtes et des visualisations rapides et interactives.
- AWS Config : permet de voir quelle infrastructure AWS a été utilisée à différents moments.
- AWS CloudTrail : permet de voir quelles API AWS ont été appelées à quel moment et par quel mandataire.

Chez AWS, nous organisons des réunions hebdomadaires pour examiner les performances opérationnelles et partager les enseignements entre les équipes. Compte tenu du nombre conséquent d'équipes AWS, nous avons créé The Wheel pour choisir de façon aléatoire une charge de travail à examiner. Le respect d'un rythme régulier pour l'examen des performances opérationnelles et le partager des connaissances améliore la capacité de vos équipes opérationnelles à atteindre des performances supérieures.

Anti-modèles courants :

- Collecte limitée aux métriques par défaut.
- Définition d'une stratégie de surveillance et sans examen.
- Absence de discussion relative à la surveillance lorsque des modifications majeures sont déployées.

Avantages liés au respect de cette bonne pratique : La vérification régulière de votre surveillance permet d'anticiper les problèmes potentiels, au lieu de réagir aux notifications lorsqu'un problème anticipé se produit réellement.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Créer plusieurs tableaux de bord pour la charge de travail. Vous devez disposer d'un tableau de bord de niveau supérieur qui contient les principales métriques commerciales, ainsi que les métriques techniques que vous avez identifiées comme étant les plus pertinentes pour l'état projeté de la charge de travail au fil de la variation de l'utilisation. Vous devez également avoir des tableaux de bord pour différents niveaux et dépendances d'application qui peuvent être inspectés.
  - Fonctionnement des tableaux de bord Amazon CloudWatch
- Planifier et effectuer des vérifications régulières des tableaux de bord de charge de travail. Effectuez une inspection régulière des tableaux de bord. Vous pouvez avoir des cadences différentes selon la profondeur à laquelle vous inspectez.
  - Inspecter les tendances dans les métriques. Comparez les valeurs des métriques aux valeurs historiques pour voir s'il existe des tendances qui peuvent indiquer que quelque chose doit faire l'objet d'une enquête. Voici quelques exemples : augmentation de la latence, diminution de la fonction principale de l'entreprise et augmentation des réponses aux échecs.
  - Inspecter les valeurs atypiques ou les anomalies dans vos métriques. Les moyennes ou les médianes peuvent masquer des valeurs atypiques et des anomalies. Observez les valeurs les plus élevées et les plus faibles pendant la période et examinez les causes des scores extrêmes. L'abaissement de votre définition de l'extrême vous permet de continuer à améliorer l'homogénéité des performances de votre charge de travail au fur et à mesure que vous continuez à éliminer ces causes.
  - Rechercher des changements importants de comportement. Un changement immédiat de quantité ou de direction d'une métrique peut indiquer qu'il y a eu un changement dans

l'application ou la présence de facteurs externes pour le suivi desquels vous devez ajouter des métriques supplémentaires.

Ressources

Documents connexes :

- Exemples de requêtes Amazon CloudWatch Logs Insights
- Débogage avec Amazon CloudWatch Synthetics et AWS X-Ray
- Un atelier sur l'observabilité
- L'Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle
- Fonctionnement des tableaux de bord Amazon CloudWatch

REL06-BP07 Surveiller la traçabilité complète des demandes via votre système

Suivez les demandes au fur et à mesure qu'elles sont traitées dans les composants du service afin que les équipes produits puissent plus facilement analyser et résoudre les problèmes et améliorer les performances.

Résultat souhaité : les charges de travail dotées d'un suivi complet de tous les composants sont faciles à déboguer, ce qui permet d'améliorer le temps moyen de résolution (MTTR) des erreurs et la latence en simplifiant la découverte de la cause première. Le traçage de bout en bout réduit le temps nécessaire à la découverte des composants concernés et à l'analyse détaillée des causes profondes des erreurs ou de la latence.

Anti-modèles courants :

- Le traçage est utilisé pour certains composants, mais pas pour tous. Par exemple, sans traçage pour AWS Lambda, les équipes risquent de ne pas comprendre clairement la latence provoquée par les démarrages à froid dans le cadre d'une charge de travail irrégulière.
- Les canaris synthétiques ou la surveillance des utilisateurs réels (RUM) ne sont pas configurés avec le traçage. Sans canaris ni RUM, la télémétrie des interactions avec le client est omise de l'analyse des traces, ce qui donne un profil de performance incomplet.
- Les charges de travail hybrides incluent à la fois des outils de suivi natifs du cloud et des outils tiers, mais aucune mesure n'a été prise pour intégrer pleinement une solution de traçage unique.

En fonction de la solution de traçage sélectionnée, les kits SDK de traçage natifs du cloud doivent être utilisés pour instrumenter des composants qui ne sont pas natifs du cloud ou des outils tiers doivent être configurés pour ingérer la télémétrie de suivi native du cloud.

Avantages liés au respect de cette bonne pratique : lorsque les équipes de développement sont alertées de problèmes, elles peuvent obtenir une image complète des interactions entre les composants du système, y compris la corrélation composant par composant avec la journalisation, les performances et les défaillances. Dans la mesure où le traçage permet d'identifier visuellement les causes profondes, vous passez moins de temps à les étudier. Les équipes qui comprennent en détail les interactions entre les composants prennent de meilleures décisions plus rapidement lors de la résolution des problèmes. L'analyse des traces des systèmes permet d'améliorer la prise de décisions, par exemple quand il convient de recourir à la reprise après sinistre (DR) ou de choisir le meilleur endroit pour mettre en œuvre des stratégies d'autoréparation, ce qui permet d'améliorer la satisfaction des clients envers vos services.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Les équipes qui exploitent des applications distribuées peuvent utiliser des outils de traçage pour établir un identifiant de corrélation, collecter des traces de demandes et créer des cartes de service pour les composants connectés. Tous les composants de l'application doivent être inclus dans les traces des demandes, notamment les clients de service, les passerelles d'intergiciels et les bus d'événements, les composants de calcul et le stockage, y compris les magasins de clés-valeurs et les bases de données. Incluez des canaris synthétiques et une surveillance des utilisateurs réels dans votre configuration de traçage de bout en bout pour mesurer les interactions avec les clients distants et la latence afin d'évaluer avec précision les performances de vos systèmes par rapport à vos contrats et objectifs de niveau de service.

Vous pouvez utiliser les services d'instrumentation d'AWS X-Ray et de la surveillance des applications Amazon CloudWatch pour fournir une vue complète des demandes au fur et à mesure qu'elles transitent par votre application. X-Ray collecte la télémétrie des applications et vous permet de la visualiser et de la filtrer en fonction des charges utiles, des fonctions, des traces, des services et des API. Il peut être activé pour les composants du système sans code ou avec peu de code. La surveillance des applications CloudWatch inclut ServiceLens pour intégrer vos traces aux métriques, aux journaux et aux alarmes. La surveillance des applications CloudWatch inclut également des outils synthétiques pour surveiller vos points de terminaison et vos API, ainsi que la surveillance des utilisateurs réels pour instrumenter vos clients d'applications Web.

Étapes d'implémentation

- Utilisez AWS X-Ray sur tous les services natifs pris en charge, comme Amazon S3, AWS Lambda et Amazon API Gateway. Ces services AWS permettent l'utilisation de X-Ray grâce à des options de configuration utilisant l'infrastructure sous forme de code, de kits AWS SDK ou de la AWS Management Console.
- Applications des instruments AWS Distro for Open Telemetry et X-Ray ou des agents de collecte tiers.
- Consultez le Guide du développeur AWS X-Ray pour plus d'informations sur l'implémentation spécifique au langage de programmation. Ces sections de la documentation expliquent comment instrumenter les requêtes HTTP, les requêtes SQL et d'autres processus spécifiques à votre langage de programmation d'application.
- Utilisez le traçage X-Ray pour les canaris synthétiques Amazon CloudWatch et Amazon CloudWatch RUM afin d'analyser le chemin de requête depuis votre client utilisateur final via votre infrastructure AWS en aval.
- Configurez les métriques et les alarmes CloudWatch en fonction de l'intégrité des ressources et de la télémétrie canari afin que les équipes soient rapidement alertées des problèmes, puis qu'elles puissent analyser en profondeur les traces et les cartographies de services avec ServiceLens.
- Activez l'intégration de X-Ray pour les outils de traçage tiers tels que Datadog, New Relicou Dynatrace si vous utilisez des outils tiers pour votre solution de traçage principale.

Ressources

Bonnes pratiques associées :

- REL06-BP01 Surveiller tous les composants de la charge de travail (génération)
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances

Documents connexes :

- Qu'est-ce qu'AWS X-Ray ?
- Amazon CloudWatch : surveillance des applications
- Débogage avec Amazon CloudWatch Synthetics et AWS X-Ray
- Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle

- Intégration d'AWS X-Ray à d'autres services AWS
- AWS Distro for OpenTelemetry et AWS X-Ray
- Amazon CloudWatch : utilisation de la surveillance synthétique
- Amazon CloudWatch : utilisation de CloudWatch RUM
- Configuration d'un canari synthétique Amazon CloudWatch et d'une alarme Amazon CloudWatch
- Disponibilité et plus encore : comprendre et améliorer la résilience des systèmes distribués sur AWS

Exemples connexes :

- Un atelier sur l'observabilité

Vidéos connexes :

- AWS re:Invent 2022 - How to monitor applications across multiple accounts
- How to Monitor your AWS Applications

Outils associés :

- d'AWS X-Ray
- Amazon CloudWatch
- Amazon Route 53

FIA 7. Comment concevez-vous votre charge de travail pour s'adapter aux changements de demande ?

Une charge de travail évolutive fournit l'élasticité nécessaire pour ajouter ou supprimer automatiquement des ressources de telle sorte qu'elles correspondent étroitement à tout moment à la demande en cours.

Bonnes pratiques

- REL07-BP01 Utiliser l'automatisation lors de l'obtention des ressources ou de leur mise à l'échelle
- REL07-BP02 Obtenir des ressources après la détection d'un problème sur une charge de travail

- REL07-BP03 Obtenir des ressources après avoir réalisé qu'un plus grand nombre de ressources est nécessaire pour une charge de travail
- REL07-BP04 Effectuer un test de charge de votre charge de travail

REL07-BP01 Utiliser l'automatisation lors de l'obtention des ressources ou de leur mise à l'échelle

Lorsque vous remplacez des ressources compromises ou que vous mettez à l'échelle votre charge de travail, automatisez le processus à l'aide de services AWS gérés comme Amazon S3 et AWS Auto Scaling. Vous pouvez également utiliser des outils tiers et les kits SDK AWS pour automatiser la mise à l'échelle.

Les services AWS gérés comprennent Amazon S3, Amazon CloudFront, AWS Auto Scaling, AWS Lambda, Amazon DynamoDB, AWS Fargate et Amazon Route 53.

AWS Auto Scaling vous permet de détecter et de remplacer les instances dégradées. Il offre également la possibilité de créer des plans de mise à l'échelle pour les ressources, notamment les instances Amazon EC2 et les parcs Spot, les tâches Amazon ECS les tables et index Amazon DynamoDB et les réplicas Amazon Aurora .

Lors de la mise à l'échelle d'instances EC2, veillez à utiliser plusieurs zones de disponibilité (de préférence au moins trois) et à ajouter ou supprimer de la capacité pour maintenir l'équilibre entre ces zones de disponibilité. Les tâches ECS ou les pods Kubernetes (lors de l'utilisation d'Amazon Elastic Kubernetes Service) doivent également être répartis sur plusieurs zones de disponibilité.

Lorsque vous utilisez AWS Lambda, la mise à l'échelle est automatique. Chaque fois qu'une notification d'événement est reçue pour votre fonction, AWS Lambda localise rapidement la capacité disponible dans son parc de calcul et exécute votre code jusqu'à la simultanéité allouée. Vous devez vous assurer que la simultanéité nécessaire est configurée sur la fonction Lambda spécifique et dans Service Quotas.

Amazon S3 se met automatiquement à l'échelle pour gérer les débits de requêtes élevés. Par exemple, votre application peut obtenir au moins 3 500 demandes PUT/COPY/POST/DELETE ou 5 500 requêtes GET/HEAD par seconde et par préfixe partitionné dans un compartiment. Le nombre de préfixes dans un compartiment est illimité. Vous pouvez augmenter vos performances de lecture ou d'écriture en parallélisant les lectures. Par exemple, si vous créez 10 préfixes dans un compartiment Amazon S3 pour paralléliser les lectures, vous pouvez mettre à l'échelle vos performances de lecture sur 55 000 requêtes de lecture par seconde.

Configurez et utilisez Amazon CloudFront ou un réseau de diffusion de contenus (CDN) de confiance Un CDN fournit des temps de réponse plus rapides à l'utilisateur final et répond aux demandes de contenu à partir du cache, ce qui vous évite (dans une certaine mesure) de devoir adapter votre charge de travail.

Anti-modèles courants :

- Implémentation de groupes Auto Scaling pour la réparation automatique sans implémentation de l'élasticité
- Utilisation de la mise à l'échelle automatique pour répondre aux pics importants du trafic.
- Déploiement d'applications hautement dynamiques avec élimination de l'option d'élasticité.

Avantages liés au respect de cette bonne pratique : L'automatisation élimine le risque d'erreur manuelle lors du déploiement et de la mise hors service des ressources. Elle élimine aussi le risque de dépassement de coûts et de déni de service en raison d'une réponse lente aux besoins de déploiement ou de mise hors service.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Configurer et utiliser AWS Auto Scaling. AWS Auto Scaling permet de surveiller vos applications et d'ajuster automatiquement la capacité pour maintenir des performances stables et prévisibles au coût le plus bas possible. Avec AWS Auto Scaling, vous pouvez configurer la mise à l'échelle des applications pour plusieurs ressources sur plusieurs services.
  - Qu'est-ce qu'AWS Auto Scaling ?
    - Configurez Auto Scaling sur vos instances Amazon EC2 et vos parcs d'instances Spot, vos tâches Amazon ECS, vos tables et index Amazon DynamoDB, vos réplicas Amazon Aurora et vos appliances AWS Marketplace, le cas échéant.
      - Gestion automatique de la capacité de débit avec DynamoDB Auto Scaling.
        - Utiliser les opérations d'API de service pour spécifier les alarmes, les stratégies de mise à l'échelle, ainsi que les temps de montée et de baisse de charge
- Utiliser Elastic Load Balancing. Les équilibreurs de charge peuvent répartir la charge par chemin d'accès ou par connectivité réseau.
  - Qu'est-ce qu'Elastic Load Balancing ?
    - Les Application Load Balancers peuvent répartir la charge par chemin.

    - Qu'est-ce qu'un Application Load Balancer ?
      - Configurer un Application Load Balancer pour répartir le trafic sur différentes charges de travail selon le chemin d'accès du nom de domaine
      - Les Application Load Balancers peuvent être utilisés pour répartir les charges d'une manière qui s'intègre à AWS Auto Scaling pour gérer la demande.
        - Utiliser un équilibreur de charge avec un groupe Auto Scaling
    - Les Network Load Balancers peuvent répartir la charge de travail par connexion.
      - Qu'est-ce qu'un Network Load Balancer ?
        - Configurer un Network Load Balancer pour répartir le trafic sur différentes charges de travail à l'aide du TCP ou disposer constamment d'un jeu d'adresses IP pour votre charge de travail
        - Les Network Load Balancers peuvent être utilisés pour répartir les charges d'une manière qui s'intègre à AWS Auto Scaling pour gérer la demande.
- Utiliser un fournisseur DNS à haut niveau de disponibilité. Les noms DNS permettent à vos utilisateurs de saisir des noms plutôt que des adresses IP pour accéder à vos charges de travail et distribuer ces informations sur une portée précise, en général mondiale, pour les utilisateurs de ces charges de travail.
  - Utiliser Amazon Route 53 ou un fournisseur DNS de confiance.
    - Qu'est-ce qu'Amazon Route 53 ?
  - Utilisez Route 53 pour gérer vos distributions CloudFront et vos équilibreurs de charge.
    - Déterminer les domaines et les sous-domaines que vous allez à gérer
    - Créez des jeux d'enregistrements appropriés à l'aide d'enregistrements ALIAS ou CNAME.
      - Utilisation des enregistrements
- Utilisez le réseau mondial AWS pour optimiser le chemin de vos utilisateurs vers vos applications. AWS Global Accelerator surveille en permanence l'état des points de terminaison de votre application et redirige le trafic vers des points de terminaison sains en moins de 30 secondes.
  - AWS Global Accelerator est un service qui améliore la disponibilité et les performances de vos applications auprès d'utilisateurs locaux ou internationaux. Il fournit des adresses IP statiques qui font office de point d'entrée fixe aux points de terminaison de votre application dans une ou plusieurs Régions AWS, telles que vos Application Load Balancers, vos Network Load Balancers ou vos instances Amazon EC2.
    - Qu'est-ce qu'AWS Global Accelerator ?

- Configurez et utilisez Amazon CloudFront ou un réseau de diffusion de contenus (CDN) de confiance Un réseau de diffusion de contenus peut fournir des temps de réponse des utilisateurs finaux plus rapides et traiter les requêtes de contenu susceptibles de causer une mise à l'échelle inutile de vos charges de travail.
  - Qu'est-ce que Amazon CloudFront ?
    - Configurez les distributions Amazon CloudFront pour vos charges de travail ou utilisez un CDN tiers.
      - Vous pouvez limiter l'accès à vos charges de travail de sorte qu'elles ne soient accessibles qu'à partir de CloudFront. Pour ce faire, vous pouvez utiliser les plages d'adresses IP pour CloudFront dans vos groupes de sécurité ou vos politiques d'accès des points de terminaison.

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à créer des solutions de calcul automatisées
- AWS Auto Scaling : Fonctionnement des plans de mise à l'échelle
- AWS Marketplace : produits utilisables avec Auto Scaling
- Gestion automatique de la capacité de débit avec DynamoDB Auto Scaling
- Utiliser un équilibreur de charge avec un groupe Auto Scaling
- Qu'est-ce qu'AWS Global Accelerator ?
- Qu'est-ce que Amazon EC2 Auto Scaling ?
- Qu'est-ce qu'AWS Auto Scaling ?
- Qu'est-ce que Amazon CloudFront ?
- Qu'est-ce qu'Amazon Route 53 ?
- Qu'est-ce qu'Elastic Load Balancing ?
- Qu'est-ce qu'un Network Load Balancer ?
- Qu'est-ce qu'un Application Load Balancer ?
- Utilisation des enregistrements

REL07-BP02 Obtenir des ressources après la détection d'un problème sur une charge de travail

Si la disponibilité est affectée, mettez à l'échelle les ressources de manière réactive si nécessaire, afin de restaurer la disponibilité de la charge de travail.

Vous devez commencer par configurer les vérifications de l'état et les critères de ces vérifications pour indiquer quand la disponibilité est affectée par le manque de ressources. Informez ensuite le personnel approprié qu'il doit mettre à l'échelle manuellement la ressource ou lancer l'automatisation pour procéder à une mise à l'échelle automatique.

La mise à l'échelle peut être ajustée manuellement en fonction de votre charge de travail. Par exemple, vous pouvez modifier le nombre d'instances EC2 dans un groupe Auto Scaling ou le débit d'une table DynamoDB via la AWS Management Console ou AWS CLI). Toutefois, l'automatisation doit être utilisée à chaque fois que c'est possible (voir Utiliser l'automatisation lors de l'obtention des ressources ou de leur mise à l'échelle).

Résultat souhaité : des opérations de mise à l'échelle (automatique ou manuelle) sont lancées pour rétablir la disponibilité dès la détection d'une panne ou d'une dégradation de l'expérience client.

Niveau de risque exposé si cette bonne pratique n'est pas établie: moyen

Directives d'implémentation

Mettez en œuvre l'observabilité et la surveillance de tous les composants de votre charge de travail, afin de surveiller l'expérience client et de détecter les défaillances. Définissez les procédures, manuelles ou automatisées, de mise à l'échelle des ressources requises. Pour plus d'informations, consultez REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances.

Étapes d'implémentation

- Définissez les procédures, manuelles ou automatisées, de mise à l'échelle des ressources requises.
  - Les procédures de mise à l'échelle dépendent de la conception des différents composants de votre charge de travail.
  - Les procédures de mise à l'échelle varient également en fonction de la technologie sous-jacente utilisée.
    - Les composants qui utilisent AWS Auto Scaling peuvent utiliser des plans de dimensionnement pour configurer un ensemble d'instructions pour la mise à l'échelle de vos

ressources. Si vous travaillez avec AWS CloudFormation ou que vous ajoutez des balises à des ressources AWS, vous pouvez configurer des plans de dimensionnement pour différents ensembles de ressources par application. Auto Scaling offre des recommandations de stratégies de mise à l'échelle personnalisées pour chaque ressource. Une fois que vous avez créé votre plan de dimensionnement, Auto Scaling combine des méthodes de mise à l'échelle dynamique et prédictive pour prendre en charge votre stratégie de mise à l'échelle. Pour plus d'informations, consultez Fonctionnement des plans de dimensionnement.

- Amazon EC2 Auto Scaling vérifie que vous disposez du nombre adéquat d'instances Amazon EC2 disponibles pour gérer la charge de votre application. Vous créez des collections d'instances EC2, appelées groupes Auto Scaling. Vous pouvez spécifier le nombre minimal et maximal d'instances dans chaque groupe Auto Scaling et Amazon EC2 Auto Scaling s'assure que votre groupe ne dépasse jamais ces limites. Pour plus d'informations, consultez What is Amazon EC2 Auto Scaling?
- La mise à l'échelle automatique d'Amazon DynamoDB utilise le service Application Auto Scaling pour ajuster de manière dynamique la capacité de débit alloué en votre nom, en fonction des modèles de trafic réels. Cela permet à une table ou à un index secondaire global d'augmenter sa capacité de lecture et d'écriture allouée afin de gérer sans limitations les augmentations soudaines du trafic. Pour plus d'informations, consultez Gestion automatique de la capacité de débit avec la scalabilité automatique de DynamoDB.

Ressources

Bonnes pratiques associées :

- REL07-BP01 Utiliser l'automatisation lors de l'obtention des ressources ou de leur mise à l'échelle
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances

Documents connexes :

- AWS Auto Scaling: Fonctionnement des plans de dimensionnement
- Gestion automatique de la capacité de débit avec la scalabilité automatique de DynamoDB
- What Is Amazon EC2 Auto Scaling?

REL07-BP03 Obtenir des ressources après avoir réalisé qu'un plus grand nombre de ressources est nécessaire pour une charge de travail

Mettez à l'échelle les ressources de manière proactive pour répondre à la demande et éviter l'impact sur la disponibilité.

De nombreux services AWS sont automatiquement mis à l'échelle pour répondre à la demande. Si vous utilisez des instances Amazon EC2 ou des clusters Amazon ECS, vous pouvez configurer la mise à l'échelle automatique de ces instances pour qu'elle intervienne en fonction des métriques d'utilisation qui correspondent à la demande de votre charge de travail. Pour Amazon EC2, l'utilisation moyenne du CPU, le nombre de requêtes de l'équilibreur de charge ou la bande passante du réseau peuvent être utilisés pour augmenter (ou diminuer) les instances EC2. Pour Amazon ECS, l'utilisation moyenne du CPU, le nombre de requêtes de l'équilibreur de charge et l'utilisation de la mémoire peuvent être utilisés pour augmenter (ou diminuer) les tâches ECS. En utilisant Target Auto Scaling sur AWS, l'Autoscaler agit comme un thermostat domestique, en ajoutant ou en supprimant des ressources pour maintenir la valeur cible (par exemple, 70 % d'utilisation du CPU) que vous spécifiez.

AWS Auto Scaling peut également exécuter Predictive Auto Scaling, qui s'appuie sur le machine learning pour analyser la charge de travail historique de chaque ressource et anticipe régulièrement la charge future des deux prochains jours.

Little's Law permet de calculer le nombre d'instances de calcul (instances EC2, fonctions Lambda simultanées, etc.) dont vous avez besoin.

$L = \lambda W$

L = nombre d'instances (ou simultanéité moyenne dans le système)

λ = vitesse moyenne à laquelle les requêtes arrivent (demande/s.)

W = temps moyen que chaque requête passe dans le système (s.)

Par exemple, à 100 rps, si le traitement de chaque requête prend 0,5 seconde, vous aurez besoin de 50 instances pour prendre en charge la requête.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Obtenir des ressources après avoir réalisé qu'un plus grand nombre de ressources est nécessaire pour une charge de travail. Mettez à l'échelle les ressources de manière proactive pour répondre à la demande et éviter l'impact sur la disponibilité.
  - Déterminez le nombre de ressources de calcul dont vous aurez besoin (simultanéité de calcul) pour gérer un débit de demandes donné.
    - Telling Stories About Little's Law
  - Configurez la mise à l'échelle planifiée pour Amazon EC2 Auto Scaling lorsque vous disposez d'un modèle d'utilisation historique.
    - Mise à l'échelle planifiée pour Amazon EC2 Auto Scaling
  - Utilisez la mise à l'échelle prédictive AWS.
    - Scalabilité prédictive pour EC2 alimentée par le machine learning

Ressources

Documents connexes :

- AWS Auto Scaling : Fonctionnement des plans de mise à l'échelle
- AWS Marketplace : produits utilisables avec Auto Scaling
- Gestion automatique de la capacité de débit avec DynamoDB Auto Scaling
- Scalabilité prédictive pour EC2 alimentée par le machine learning
- Mise à l'échelle planifiée pour Amazon EC2 Auto Scaling
- Telling Stories About Little's Law
- Qu'est-ce que Amazon EC2 Auto Scaling ?

REL07-BP04 Effectuer un test de charge de votre charge de travail

Adoptez une méthodologie de test de charge pour déterminer si la mise à l'échelle répond aux exigences de la charge de travail.

Il est important d'exécuter régulièrement des tests de charge. Les tests de charge devraient découvrir le point de rupture et tester les performances de votre charge de travail. AWS facilite la configuration d'environnements de test temporaires qui modélisent l'échelle de votre charge de travail de production. Dans le Cloud, vous pouvez créer un environnement d'essai à l'échelle de la

production et à la demande, exécuter les tests, puis désactiver les ressources. Puisque vous ne payez l'environnement de test que lorsqu'il s'exécute, vous pouvez simuler votre environnement réel pour une fraction du coût d'un test sur site.

Les tests de charge en production doivent également être intégrés aux tests de simulation de pannes, lors desquels le système de production est mis sous tension pendant les périodes où le client est moins utilisé et tout le personnel est disponible pour interpréter les résultats et résoudre les problèmes qui surviennent.

Anti-modèles courants :

- Exécution de tests de charge sur des déploiements qui ne n'ont pas la même configuration que votre production.
- Effectuer un test de charge uniquement sur des éléments individuels de votre charge de travail, et non sur l'ensemble de la charge de travail.
- Exécution de tests de charge avec un sous-ensemble de demandes et non un ensemble représentatif de demandes réelles.
- Exécution de tests de charge avec un faible facteur de sécurité au-dessus de la charge prévue.

Avantages liés au respect de cette bonne pratique : Vous savez quels composants de votre architecture échouent sous charge et vous pouvez identifier les métriques à surveiller qui indiquent suffisamment à temps que vous approchez de cette charge pour que vous résolviez le problème et empêchiez ainsi l'impact de cette défaillance.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Exécutez des tests de charge pour identifier l'aspect de votre charge de travail qui indique que vous devez ajouter ou supprimer de la capacité. Les tests de charge doivent avoir un trafic représentatif similaire à ce que vous recevez en production. Augmentez la charge tout en surveillant les métriques que vous avez instrumentées pour déterminer quelle métrique indique quand vous devez ajouter ou supprimer des ressources.
  - Test de charge distribuée sur AWS : simulation de milliers d'utilisateurs connectés
    - Identifiez le mélange de demandes. Comme vous pouvez avoir divers mélanges de demandes, vous devez examiner les différentes périodes lors de l'identification de la combinaison de trafic.

- Implémentez un pilote de charge. Vous pouvez utiliser un code personnalisé, un logiciel open source ou un logiciel commercial pour implémenter un pilote de charge.
- Effectuez un test de charge initial avec une faible capacité. Vous constatez des effets immédiats en entraînant une charge moindre, éventuellement aussi petite qu'une instance ou un conteneur.
- Effectuez un test de charge par rapport à une capacité plus importante. Étant donné que les effets seront différents sur une charge distribuée, vous devez procéder à des essais dans un environnement aussi proche que possible de celui du produit.

Ressources

Documents connexes :

- Test de charge distribuée sur AWS : simulation de milliers d'utilisateurs connectés

FIA 8. Comment implémenter les modifications ?

Des modifications contrôlées sont nécessaires pour déployer de nouvelles fonctionnalités et s'assurer que les charges de travail et l'environnement d'exploitation fonctionnent avec des logiciels connus et peuvent être corrigés ou remplacés de manière prévisible. Si les modifications ne sont pas contrôlées, il est difficile de prédire leur effet ou de résoudre les problèmes qui en découlent.

Bonnes pratiques

- REL08-BP01 Utiliser des runbooks pour les activités standard telles que le déploiement
- REL08-BP02 Intégrer les tests fonctionnels dans le cadre de votre déploiement
- REL08-BP03 Intégrer les tests de résilience dans le cadre de votre déploiement
- REL08-BP04 Effectuer le déploiement à l'aide d'une infrastructure immuable
- REL08-BP05 Déployer les modifications avec l'automatisation

REL08-BP01 Utiliser des runbooks pour les activités standard telles que le déploiement

Les runbooks sont les procédures prédéfinies destinées à parvenir à un résultat spécifique. Utilisez des runbooks pour effectuer des tâches manuelles ou automatiques standard. Il peut s'agir du déploiement d'une charge de travail, de l'application de correctifs à une charge de travail ou de la modification du DNS.

Par exemple, mettez en place des processus pour assurer la sécurité des restaurations pendant les déploiements. Pour garantir la fiabilité d'un service, il est essentiel de s'assurer que vous pouvez restaurer un déploiement sans interruption pour vos clients.

Concernant les procédures de runbook, commencez par un processus manuel efficace valide, mettez-le en œuvre dans le code et, le cas échéant, déclenchez son exécution automatique.

Même pour les charges de travail sophistiquées hautement automatisées, les runbooks restent utiles pour exécuter des tests de simulation de pannes ou répondre à des exigences rigoureuses en matière de rapports et d'audit.

Notez que les playbooks sont utilisés en réponse à des incidents spécifiques et que les runbooks le sont pour obtenir des résultats spécifiques. En règle générale, les runbooks sont destinés aux activités de routine, tandis que les playbooks sont utilisés pour répondre à des événements non réguliers.

Anti-modèles courants :

- Exécution de modifications imprévues de la configuration en production.
- Ignorer les étapes de votre plan afin d'accélérer le déploiement, ce qui entraîne un échec du déploiement.
- Effectuez des modifications sans tester l'annulation de la modification.

Avantages liés au respect de cette bonne pratique : Une planification efficace des modifications augmente votre capacité à exécuter correctement la modification, car vous êtes conscient de tous les systèmes concernés. Vous gagnez en confiance si vous réussissez à valider des modifications que vous apportez aux environnements de test.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Obtenez des réponses cohérentes et rapides à des événements bien compris en documentant les procédures dans des runbooks.
  - Concepts AWS Well-Architected Framework : runbook
- Utilisez le principe de l'infrastructure en tant que code pour définir votre infrastructure. En ayant recours à AWS CloudFormation ou à un tiers de confiance pour définir votre infrastructure, vous pouvez utiliser le contrôle de version et suivre les modifications apportées à la version du logiciel.

- Utilisez AWS CloudFormation ou un fournisseur tiers de confiance pour définir votre infrastructure.
  - Qu'est-ce qu'AWS CloudFormation ?
- Créez des modèles qui sont singuliers et découplés, en utilisant de bons principes de conception de logiciels.
  - Déterminez les autorisations, les modèles et les responsables de l'implémentation
    - Contrôle de l'accès avec AWS Identity and Access Management
  - Utilisez le contrôle de code source, comme AWS CodeCommit ou un outil tiers de confiance, pour le contrôle de version.
    - Qu'est-ce qu'AWS CodeCommit ?

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à créer des solutions de déploiement automatisées
- AWS Marketplace : produits pouvant être utilisés pour automatiser vos déploiements
- Concepts AWS Well-Architected Framework : runbook
- Qu'est-ce qu'AWS CloudFormation ?
- Qu'est-ce qu'AWS CodeCommit ?

Exemples connexes :

- Automatisation des opérations avec les playbooks et les runbooks

REL08-BP02 Intégrer les tests fonctionnels dans le cadre de votre déploiement

Les tests fonctionnels sont exécutés dans le cadre du déploiement automatisé. Si les critères de réussite ne sont pas respectés, le pipeline est arrêté ou annulé.

Ces tests sont exécutés dans un environnement de préproduction, qui est mis en place avant la production dans le pipeline. Idéalement, cela s'effectue dans le cadre d'un pipeline de déploiement.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Intégrez les tests fonctionnels dans le cadre de votre déploiement. Les tests fonctionnels sont exécutés dans le cadre du déploiement automatisé. Si les critères de réussite ne sont pas respectés, le pipeline est arrêté ou annulé.
  - Appelez AWS CodeBuild lors de « l'action de test » de vos pipelines de publication de logiciels modélisés dans AWS CodePipeline. Cette fonctionnalité vous permet d'exécuter facilement divers tests sur votre code, en particulier des tests unitaires, des analyses de code statique et des tests d'intégration.
    - AWS CodePipeline ajoute la prise en charge des tests unitaires et des tests d'intégration personnalisés avec AWS CodeBuild
  - Utilisez les solutions AWS Marketplace pour exécuter des tests automatisés dans le cadre de votre pipeline de distribution de logiciels.
    - Automatisation des tests logiciels

Ressources

Documents connexes :

- AWS CodePipeline ajoute la prise en charge des tests unitaires et des tests d'intégration personnalisés avec AWS CodeBuild
- Automatisation des tests logiciels
- Qu'est-ce que AWS CodePipeline ?

REL08-BP03 Intégrer les tests de résilience dans le cadre de votre déploiement

Les tests de résilience (basés sur les principes de l'ingénierie du chaos) sont exécutés dans le cadre du pipeline de déploiement automatisé dans un environnement de préproduction.

Ces tests sont mis en place et exécutés dans le pipeline dans un environnement de préproduction. Ils doivent également être exécutés en production, mais dans le cadre des tests de simulation de pannes.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Intégrez les tests de résilience dans le cadre de votre déploiement. Utilisez l'ingénierie du chaos qui est la discipline d'expérimentation d'une charge de travail afin d'améliorer votre confiance en sa capacité à supporter des conditions instables en production.
  - Les tests de résilience injectent des défaillances ou une dégradation des ressources pour vérifier que votre charge de travail répond avec le niveau de résilience prévue à la conception.
    - Atelier Well-Architected : niveau 300 : test de la résilience d’EC2 RDS et de S3
  - Ces tests peuvent être exécutés régulièrement dans des environnements de préproduction dans des pipelines de déploiement automatisés.
  - Ils doivent également être exécutés en production, dans le cadre des tests de simulation de panne.
  - Si vous utilisez les principes de l'ingénierie du chaos, proposez des hypothèses sur les performances de votre charge de travail dans différentes situations, puis testez vos hypothèses à l'aide de tests de résilience.
    - Principes de l'ingénierie du chaos

Ressources

Documents connexes :

- Principes de l'ingénierie du chaos
- Qu'est qu'AWS Fault Injection Simulator (AWS FIS) ?

Exemples connexes :

- Atelier Well-Architected : niveau 300 : test de la résilience d’EC2 RDS et de S3

REL08-BP04 Effectuer le déploiement à l'aide d'une infrastructure immuable

Une infrastructure immuable est un modèle qui exige qu'aucune mise à jour, aucune application de correctifs de sécurité ni aucun changement de configuration ne se produise sur place sur les charges de travail de production. Lorsqu'un changement est nécessaire, l'architecture est intégrée à la nouvelle infrastructure et déployée en production.

Suivez une stratégie de déploiement d'infrastructure immuable pour améliorer la fiabilité, la cohérence et la reproductibilité de vos déploiements de charges de travail.

Résultat souhaité : avec une infrastructure immuable, aucune modification sur place n'est autorisée pour exécuter les ressources d'infrastructure au sein d'une charge de travail. Lorsqu'une modification est nécessaire, un nouvel ensemble de ressources d'infrastructure contenant toutes les modifications nécessaires est déployé parallèlement à vos ressources existantes. Ce déploiement est validé automatiquement et, en cas de succès, le trafic est progressivement transféré vers ce nouvel ensemble de ressources.

Cette stratégie de déploiement s'applique notamment aux mises à jour logicielles, aux correctifs de sécurité, aux modifications de l'infrastructure, ainsi qu'aux mises à jour de la configuration et des applications.

Anti-modèles courants :

- Modifications sur place des ressources d'infrastructure en cours d'exécution.

Avantages liés à l'instauration de cette bonne pratique :

- Plus grande cohérence entre les environnements : comme les ressources d'infrastructure ne diffèrent pas d'un environnement à l'autre, la cohérence est renforcée et les tests sont simplifiés.
- Réduction des écarts de configuration : en remplaçant les ressources d'infrastructure par une configuration connue et dont la version est contrôlée, l'infrastructure se trouve dans un état connu, testé et fiable, ce qui permet d'éviter les écarts de configuration.
- Déploiements atomiques fiables : soit les déploiements se déroulent avec succès, soit rien ne change, ce qui accroît la cohérence et la fiabilité du processus de déploiement.
- Déploiements simplifiés : les déploiements sont simplifiés, car ils n'ont pas besoin de prendre en charge les mises à niveau. Les mises à niveau sont simplement de nouveaux déploiements.
- Déploiements plus sûrs avec des processus de restauration et de récupération rapides : les déploiements sont plus sûrs, car la version de travail précédente n'est pas modifiée. Vous pouvez la restaurer si des erreurs sont détectées.
- Niveau de sécurité renforcé : l'impossibilité de modifier l'infrastructure permet de désactiver les mécanismes d'accès à distance (comme SSH). Vous pouvez ainsi réduire les vecteurs d'attaque tout en renforçant la sécurité de votre organisation.

Niveau de risque exposé si cette bonne pratique n'est pas établie: moyen

Directives d'implémentation

Automatisation

Lors de la définition d'une stratégie de déploiement d'infrastructure immuable, il est recommandé d'utiliser l'automatisation autant que possible afin d'améliorer la reproductibilité et de minimiser le risque d'erreur humaine. Pour plus d'informations, consultez REL08-BP05 Déployer les modifications avec l'automatisation et Automatisation de déploiements sécurisés sans intervention.

Avec l'infrastructure en tant que code (IaC), les étapes de provisionnement, d'orchestration et de déploiement de l'infrastructure sont définies de manière programmatique, descriptive et déclarative et stockées dans un système de contrôle de source. L'utilisation de l'infrastructure en tant que code simplifie l'automatisation du déploiement de l'infrastructure et contribue à garantir l'immuabilité de cette dernière.

Schémas de déploiement

Lorsqu'une modification de la charge de travail est requise, la stratégie de déploiement d'infrastructure immuable impose le déploiement d'un nouvel ensemble de ressources d'infrastructure comprenant toutes les modifications nécessaires. Il est important que ce nouvel ensemble de ressources suive un schéma de déploiement qui minimise l'impact sur les utilisateurs. Il existe deux stratégies principales pour ce type de déploiement :

Déploiement canary : il consiste à diriger un petit nombre de vos clients vers la nouvelle version, généralement exécutée sur une seule instance de service (canary). Examinez ensuite en profondeur les modifications de comportement ou les erreurs générées. Vous pouvez supprimer le trafic du canary si vous rencontrez des problèmes critiques et faire basculer les utilisateurs vers la version précédente. Si le déploiement réussit, vous pouvez le poursuivre à la vitesse souhaitée, tout en surveillant les modifications afin de détecter les erreurs, jusqu'à ce qu'il soit terminé. AWS CodeDeploy peut être configuré avec une configuration de déploiement autorisant un déploiement canary.

Déploiement bleu/vert : il est semblable au déploiement canary, à la différence qu'un parc complet de l'application est déployé en parallèle. Vos déploiements alternent entre deux piles (bleu et vert). Une fois encore, vous pouvez faire basculer le trafic vers la nouvelle version et revenir à l'ancienne si vous rencontrez des problèmes lors du déploiement. Généralement, tout le trafic est basculé en même temps, mais vous pouvez également orienter des fractions de votre trafic vers chaque version pour accélérer l'adoption de la nouvelle version en utilisant les capacités de routage DNS pondéré d'Amazon Route 53. AWS CodeDeploy et AWS Elastic Beanstalk peuvent être configurés avec une configuration de déploiement autorisant un déploiement bleu/vert.

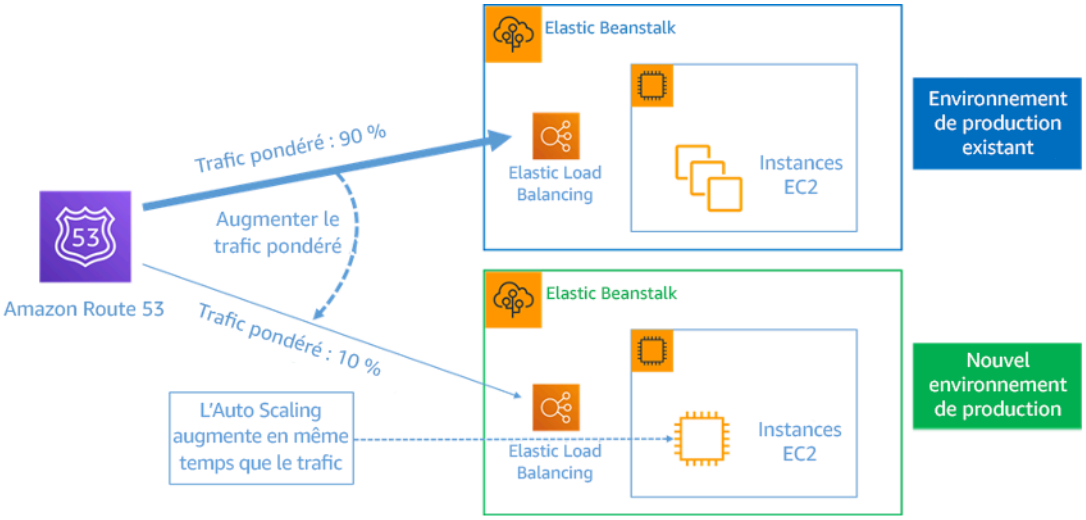


Figure 8 : Déploiement bleu/vert avec AWS Elastic Beanstalk et Amazon Route 53

Détection d'écart

Un écart est un changement qui entraîne un état ou une configuration d'une ressource d'infrastructure différent de celui attendu. Toute modification de configuration non gérée va à l'encontre de la notion d'infrastructure immuable et doit être détectée et corrigée afin de garantir la mise en œuvre d'une infrastructure immuable.

Étapes d'implémentation

- Interdisez la modification sur place des ressources d'infrastructure en cours d'exécution.
  - Vous pouvez utiliser AWS Identity and Access Management (IAM) pour spécifier les personnes ou les entités qui peuvent accéder aux services et aux ressources dans AWS, centraliser la gestion des autorisations précises et analyser les accès pour affiner les autorisations dans AWS
- Automatisez le déploiement des ressources d'infrastructure pour améliorer la reproductibilité et minimiser le risque d'erreur humaine.
  - Comme décrit dans le livre blanc Présentation de DevOps sur AWS, l'automatisation est la pierre angulaire des services AWS et elle est prise en charge en interne dans l'ensemble des services, fonctionnalités et offres.
  - La préparation de votre Amazon Machine Image (AMI) peut accélérer leur lancement. EC2 Image Builder est un service AWS entièrement géré qui vous aide à automatiser la création, la maintenance, la validation, le partage et le déploiement d'une AMI Linux ou Windows personnalisée, sécurisée et à jour.
  - Les services qui prennent en charge l'automatisation incluent :

    - AWS Elastic Beanstalk est un service permettant de déployer et de mettre à l'échelle rapidement des applications Web développées avec Java, .NET, PHP, Node.js, Python, Ruby, Go et Docker sur des serveurs familiers tels qu'Apache, NGINX, Passenger et IIS.
    - AWS Proton aide les équipes de plateforme à connecter et à coordonner tous les différents outils dont vos équipes de développement ont besoin pour le provisionnement de l'infrastructure, les déploiements de code, la surveillance et les mises à jour. AWS Proton permet d'automatiser le provisionnement de l'infrastructure en tant que code et le déploiement d'applications sans serveur et basées sur des conteneurs.
- L'utilisation d'une infrastructure en tant que code facilite l'automatisation du déploiement de l'infrastructure et contribue à garantir l'immuabilité de l'infrastructure. AWS fournit des services de création, de déploiement et de maintenance programmatique, descriptive et déclarative de l'infrastructure.
    - AWS CloudFormation aide les développeurs à créer des ressources AWS de façon ordonnée et prévisible. Les ressources sont écrites dans des fichiers texte au format JSON ou YAML. Les modèles nécessitent une syntaxe et une structure spécifiques, qui dépendent des types de ressources créées et gérées. Vous créez vos ressources au format JSON ou YAML avec n'importe quel éditeur de code AWS Cloud9, vous les archivez dans un système de contrôle de version, puis CloudFormation crée les services spécifiés de manière sûre et reproductible.
    - AWS Serverless Application Model (AWS SAM) est un cadre open source que vous pouvez utiliser pour créer des applications sans serveur sur AWS. AWS SAM s'intègre à d'autres services AWS, et est une extension de AWS CloudFormation.
    - AWS Cloud Development Kit (AWS CDK) est un cadre de développement logiciel open source permettant de modéliser et de provisionner les ressources de vos applications cloud à l'aide de langages de programmation familiers. Vous pouvez utiliser AWS CDK pour modéliser l'infrastructure d'applications avec TypeScript, Python, Java et .NET. AWS CDK utilise AWS CloudFormation en arrière-plan pour provisionner les ressources de manière sécurisée et reproductible.
    - AWS Cloud Control API introduit un ensemble commun d'API CRUDL (Create, Read, Update, Delete, and List) pour aider les développeurs à gérer leur infrastructure cloud de façon simple et cohérente. Les API courantes de Cloud Control API permettent aux développeurs de gérer de manière uniforme le cycle de vie des services AWS et tiers.
- Mettez en œuvre des modèles de déploiement qui minimisent l'impact sur les utilisateurs.
    - Déploiements canary :
        - Configuration d'un déploiement de la version canary API Gateway

    - Create a pipeline with canary deployments for Amazon ECS using AWS App Mesh
  - Déploiements bleu/vert : le livre blanc Blue/Green Deployments on AWS décrit des exemples de techniques pour mettre en œuvre des stratégies de déploiement bleu/vert.
- Détectez les écarts de configuration ou d'état. Pour plus d'informations, consultez Détection de modifications non gérées de la configuration des piles et des ressources.

Ressources

Bonnes pratiques associées :

- REL08-BP05 Déployer les modifications avec l'automatisation

Documents connexes :

- Automatisation de déploiements sécurisés sans intervention
- Leveraging AWS CloudFormation to create an immutable infrastructure at Nubank
- Infrastructure en tant que code
- Implementing an alarm to automatically detect drift in AWS CloudFormation stacks

Vidéos connexes :

- AWS re:Invent 2020: Reliability, consistency, and confidence through immutability

REL08-BP05 Déployer les modifications avec l'automatisation

Les déploiements et l'application de correctifs sont automatisés pour éliminer l'impact négatif.

Les modifications apportées aux systèmes de production sont l'un des secteurs de risque les plus importants pour de nombreuses organisations. Nous considérons les déploiements comme un problème de premie ordre à résoudre, tout comme les problèmes opérationnels que le logiciel rencontre. Aujourd'hui, il convient d'appliquer l'automatisation dès que les opérations le permettent, y compris lors des tests et du déploiement de modifications, lors de l'ajout ou de la suppression de capacités et lors de la migration des données. AWS CodePipeline vous permet de gérer les étapes nécessaires à la libération de votre charge de travail. Cela englobe un état de déploiement utilisant AWS CodeDeploy pour automatiser le déploiement du code d'application sur les instances Amazon EC2, les instances sur site, les fonctions Lambda sans serveur ou les services Amazon ECS.

> **Recommandations**
>
> Bien que les principes traditionnels suggèrent de garder les interventions humaines dans la boucle des procédures opérationnelles les plus complexes, nous vous conseillons justement d'automatiser ces mêmes procédures pour cette raison.

Anti-modèles courants :

- Modifications manuelles
- Saut des étapes de votre automatisation via les flux de travail d'urgence.
- Non suivi de vos plans.

Avantages liés au respect de cette bonne pratique : L'utilisation de l'automatisation pour déployer toutes les modifications élimine le risque d'introduction d'erreurs humaines et permet de tester avant de changer la production afin de s'assurer que vos plans sont suivis.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Automatisez votre pipeline de déploiement. Le déploiement des pipelines vous permet d'une part d'invoquer des tests automatisés et la détection des anomalies et, d'autre part, d'arrêter le pipeline à une certaine étape avant le déploiement en production ou de restaurer automatiquement l'environnement d'avant la modification.
  - L'Amazon Builders' Library : Garantir la sécurité des restaurations pendant les déploiements
  - L'Amazon Builders' Library : Aller plus vite avec la distribution continue
    - Utilisez AWS CodePipeline ou un produit tiers de confiance pour définir et exécuter vos pipelines.
      - Configurez le pipeline pour démarrer lorsqu'une modification est apportée à votre référentiel de code.
        - Qu'est-ce qu'AWS CodePipeline ?
      - Utilisez Amazon Simple Notification Service (Amazon SNS) and Amazon Simple Email Service (Amazon SES) pour envoyer des notifications sur les problèmes dans le pipeline ou pour intégrer un outil de chat d'équipe, comme Amazon Chime.
        - Qu'est-ce qu'Amazon Simple Notification Service ?

- Qu'est-ce que Amazon SES ?
- Qu'est-ce qu'Amazon Chime ?
- Automatisez les messages de chat avec les webhooks.

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à créer des solutions de déploiement automatisées
- AWS Marketplace : produits pouvant être utilisés pour automatiser vos déploiements
- Automatisez les messages de chat avec les webhooks.
- L'Amazon Builders' Library : Garantir la sécurité des restaurations pendant les déploiements
- L'Amazon Builders' Library : Aller plus vite avec la distribution continue
- Qu'est-ce que AWS CodePipeline ?
- Qu'est-ce que CodeDeploy ?
- le gestionnaire de correctifs AWS Systems Manager
- Qu'est-ce que Amazon SES ?
- Qu'est-ce qu'Amazon Simple Notification Service ?

Vidéos connexes :

- AWS Summit 2019: CI/CD on AWS

Gestion des défaillances

Questions

- FIA 9. Comment sauvegarder des données ?
- FIA 10. Comment utilisez-vous l'isolement des pannes pour protéger votre charge de travail ?
- FIA 11. Comment concevez-vous votre charge de travail pour la rendre résistante aux défaillances de composants ?
- FIA 12. Comment tester la fiabilité ?
- FIA 13. Comment planifier la reprise après sinistre (DR) ?

FIA 9. Comment sauvegarder des données ?

Sauvegardez les données, les applications et la configuration pour répondre à vos exigences en matière d'objectifs de temps de récupération (RTO) et d'objectifs de point de récupération (RPO).

Bonnes pratiques

- REL09-BP01 Identifier et sauvegarder toutes les données qui doivent être sauvegardées, ou reproduire les données à partir de sources
- REL09-BP02 Sécuriser et chiffrer les sauvegardes
- REL09-BP03 Effectuer automatiquement la sauvegarde des données
- REL09-BP04 Effectuer une récupération périodique des données pour vérifier l'intégrité et les processus de sauvegarde

REL09-BP01 Identifier et sauvegarder toutes les données qui doivent être sauvegardées, ou reproduire les données à partir de sources

Identifiez et utilisez les fonctionnalités de sauvegarde des services et ressources de données utilisés par votre charge de travail. La plupart des services offrent des fonctionnalités permettant de sauvegarder vos données de charge de travail.

Résultat souhaité : les sources de données ont été identifiées et classées en fonction de leur ordre d'importance. Définissez ensuite une stratégie de récupération des données basée sur le RPO. Cette stratégie implique soit de sauvegarder ces sources de données, soit d'avoir la capacité de reproduire des données provenant d'autres sources. En cas de perte de données, la stratégie mise en place permet la récupération ou la reproduction des données dans les RPO et RTO définis.

Phase de maturité du cloud : fondamentale

Anti-modèles courants :

- Ne pas connaître toutes les sources de données pour la charge de travail ni leur ordre d'importance.
- Ne pas effectuer de sauvegardes des sources de données critiques.
- Sauvegarder uniquement certaines sources de données sans utiliser leur ordre d'importance comme critère.
- Aucun RPO défini, ou la fréquence de sauvegarde ne parvient pas à atteindre le RPO.

- Ne pas évaluer si une sauvegarde est nécessaire ou si les données peuvent être reproduites à partir d'autres sources.

Avantages liés au respect de cette bonne pratique : identifier les emplacements où les sauvegardes sont nécessaires et mettre en place un mécanisme pour créer des sauvegardes, ou être capable de reproduire les données à partir d'une source externe améliore la capacité de restauration et de récupération des données lors d'une panne.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Tous les magasins de données AWS offrent des fonctionnalités de sauvegarde. Des services comme Amazon RDS et Amazon DynamoDB prennent également en charge la sauvegarde automatisée qui permet la récupération ponctuelle (PITR). Vous pouvez ainsi restaurer une sauvegarde remontant jusqu'à cinq minutes ou moins avant l'heure actuelle. De nombreux services AWS offrent la possibilité de copier les sauvegardes vers une autre Région AWS. AWS Backup est un outil qui vous permet de centraliser et d'automatiser la protection des données entre les services AWS. AWS Elastic Disaster Recovery vous permet de copier des charges de travail complètes de serveurs et de maintenir une protection continue des données à partir d'un site, d'une zone géographique ou d'une région, avec un objectif de point de reprise (RPO) mesuré en secondes.

Amazon S3 peut être utilisé comme destination de sauvegarde pour les sources de données autogérées et gérées par AWS. Les services AWS tels qu'Amazon EBS, Amazon RDS et Amazon DynamoDB ont des fonctionnalités intégrées permettant de créer des sauvegardes. Vous pouvez aussi utiliser des logiciels de sauvegarde tiers.

Les données sur site peuvent être sauvegardées sur le AWS Cloud à l'aide de AWS Storage Gateway ou de AWS DataSync. Les compartiments Amazon S3 permettent de stocker ces données sur AWS. Amazon S3 offre plusieurs niveaux de stockage tels que Amazon S3 Glacier ou S3 Glacier Deep Archive pour réduire le coût du stockage des données.

Il se peut que vous puissiez répondre aux besoins de récupération de données en reproduisant les données à partir d'autres sources. Par exemple, les nœuds de réplica Amazon ElastiCache ou les réplicas en lecture Amazon RDS peuvent reproduire des données en cas de perte de la source principale. Dans les cas où des sources de ce type peuvent être utilisées pour atteindre votre objectif de point de reprise (RPO) et votre objectif de délai de reprise (RTO), il se peut que vous n'ayez pas besoin d'une sauvegarde. Autre exemple, si vous travaillez avec Amazon EMR, il n'est peut-être pas

nécessaire de sauvegarder votre magasin de données HDFS, tant que vous pouvez reproduire les données dans Amazon EMR à partir de Amazon S3.

Lors de la sélection d'une stratégie de sauvegarde, tenez compte du temps nécessaire pour récupérer les données. Le temps nécessaire pour récupérer les données dépend du type de sauvegarde (dans le cas d'une stratégie de sauvegarde) ou de la complexité du mécanisme de reproduction des données. Cette durée doit être conforme au RTO de la charge de travail.

Étapes d'implémentation

1. Identifiez toutes les sources de données pour la charge de travail. Les données peuvent être stockées sur un certain nombre de ressources telles que les bases de données, les volumes, les systèmes de fichiers, les systèmes de journalisation et le stockage d'objets. Reportez-vous à la section Ressources pour trouver des documents connexes sur les différents services AWS où les données sont stockées, et la capacité de sauvegarde que ces services fournissent.
2. Classez les sources de données en fonction de leur ordre d'importance. Différents jeux de données ont différents niveaux d'importance pour une charge de travail, et donc différentes exigences en matière de résilience. Par exemple, certaines données peuvent être critiques et nécessiter un RPO proche de zéro, tandis que d'autres données peuvent être moins critiques et peuvent tolérer un RPO plus élevé et la perte de certaines données. De même, différents jeu de données peuvent également avoir des exigences de RTO différentes.
3. Utilisez AWS ou des services tiers pour créer des sauvegardes des données. AWS Backup est un service géré qui permet de créer des sauvegardes de diverses sources de données sur AWS. AWS Elastic Disaster Recovery gère la réplication automatisée des données à la seconde près vers une Région AWS. La plupart des services AWS ont également des fonctionnalités natives permettant de créer des sauvegardes. AWS Marketplace inclut de nombreuses solutions qui offrent également ces fonctionnalités. Reportez-vous à la section Ressources ci-dessous pour découvrir comment créer des sauvegardes de données à partir de divers services AWS.
4. Pour les données non sauvegardées, définissez un mécanisme de reproduction des données. Vous pouvez choisir de ne pas sauvegarder les données qui peuvent être reproduites à partir d'autres sources pour diverses raisons. Il peut arriver qu'il soit moins coûteux de reproduire des données à partir de sources en cas de besoin plutôt que de créer une sauvegarde, car le stockage des sauvegardes peut impliquer un coût. Ou peut-être la restauration à partir d'une sauvegarde prend-elle plus de temps que la reproduction des données à partir des sources, ce qui entraîne une violation du RTO. Dans de telles situations, envisagez les avantages et inconvénients de chaque approche et définissez un processus clair sur la façon dont les données peuvent être reproduites à partir de ces sources lorsque la récupération des données est nécessaire. Si vous

avez chargé des données depuis Amazon S3 vers un entrepôt de données (comme Amazon Redshift) ou un cluster MapReduce (comme Amazon EMR) pour les analyser, vous disposez d'un exemple de données reproductibles à partir d'autres sources. Tant que les résultats de ces analyses sont stockés quelque part ou reproductibles, vous ne perdrez pas données en cas de défaillance de l'entrepôt de données ou du cluster MapReduce. Parmi les autres exemples reproductibles à partir de sources, figurent les caches (comme Amazon ElastiCache) ou les réplicas en lecture RDS.

5. Spécifiez un rythme de sauvegarde des données. La création de sauvegardes de sources de données est un processus périodique, et la fréquence doit dépendre du RPO.

Niveau d'effort du plan d'implémentation : modéré

Ressources

Bonnes pratiques associées :

REL13-BP01 Définir les objectifs de reprise pour les temps d'arrêt et les pertes de données

REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise

Documents connexes :

- Qu'est-ce que AWS Backup ?
- What is AWS DataSync? (Qu'est-ce qu'AWS DataSync ?)
- Qu'est-ce que la passerelle de volume ?
- Partenaire APN : partenaires pouvant faciliter la sauvegarde
- AWS Marketplace : produits pouvant être utilisés pour la sauvegarde
- Instantanés Amazon EBS
- Backing Up Amazon EFS (Sauvegarde Amazon EFS)
- Sauvegarde d'Amazon FSx for Windows File Server
- Sauvegarde et restauration d'ElastiCache pour Redis
- Création d'un instantané de cluster de base de données dans Neptune
- Création d'un instantané de base de données
- Création d'une règle EventBridge qui se déclenche selon un calendrier
- Réplication entre régions avec Amazon S3

- EFS-to-EFS AWS Backup
- Exportation de données de journal vers Amazon S3
- Gestion du cycle de vie des objets
- Sauvegarde et restauration à la demande pour DynamoDB
- Restauration à un instant dans le passé pour DynamoDB
- Utilisation des instantanés d'index Amazon OpenSearch Service
- Qu'est-ce que AWS Elastic Disaster Recovery ?

Vidéos connexes :

- AWS re:Invent 2021 - Backup, disaster recovery, and ransomware protection with AWS
- AWS Backup Demo: Cross-Account and Cross-Region Backup (Démonstration de la sauvegarde AWS : sauvegarde intercompte et inter-régions)
- AWS re:Invent 2019: Deep dive on AWS Backup, ft. Rackspace (STG341)

Exemples connexes :

- Atelier Well-Architected : implémentation de la réplication bidirectionnelle entre régions pour Amazon S3
- Atelier Well-Architected : test de la sauvegarde et de la restauration de données
- Atelier Well-Architected : sauvegarde et restauration avec basculement automatique pour la charge de travail d'analyse
- Atelier Well-Architected : reprise après sinistre - sauvegarde et restauration

REL09-BP02 Sécuriser et chiffrer les sauvegardes

Contrôlez et détectez l'accès aux sauvegardes à l'aide de l'authentification et de l'autorisation. Assurez la prévention et détectez si l'intégrité des données des sauvegardes est compromise à l'aide du chiffrement.

Anti-modèles courants :

- Avoir le même accès aux sauvegardes et à l'automatisation de la restauration que vous le faites pour les données.
- Absence de chiffrement de vos sauvegardes.

Avantages liés au respect de cette bonne pratique : la sécurisation de vos sauvegardes empêche la falsification des données. De même, le chiffrement des données empêche l'accès à ces données si elles sont accidentellement exposées.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Contrôlez et détectez l'accès aux sauvegardes à l'aide de l'authentification et de l'autorisation, telles qu'AWS Identity and Access Management (IAM). Assurez la prévention et détectez si l'intégrité des données des sauvegardes est compromise à l'aide du chiffrement.

Amazon S3 prend en charge plusieurs méthodes de chiffrement de vos données inactives. Grâce au chiffrement côté serveur, Amazon S3 accepte vos objets sous forme de données non chiffrées, puis les chiffre lors de leur stockage. Avec le chiffrement côté client, votre application de charge de travail s'occupe du chiffrement des données avant leur transmission à Amazon S3. Ces deux méthodes vous permettent d'utiliser AWS Key Management Service (AWS KMS) pour créer et stocker la clé de données. Vous pouvez également fournir votre propre clé (vous en assumez alors la responsabilité). Avec AWS KMS, vous pouvez définir des stratégies via IAM pour déterminer qui peut ou non accéder à vos clés de données et à vos données déchiffrées.

Pour Amazon RDS, si vous avez choisi de chiffrer vos bases de données, vos sauvegardes sont également chiffrées. Les sauvegardes DynamoDB sont toujours chiffrées. En utilisant AWS Elastic Disaster Recovery, toutes les données en transit et au repos sont chiffrées. Avec Elastic Disaster Recovery, les données au repos peuvent être chiffrées à l'aide de la clé Amazon EBS de chiffrement de volume par défaut ou d'une clé personnalisée gérée par le client.

Étapes d'implémentation

1. Utilisez le chiffrement sur chacun de vos magasins de données. La sauvegarde est également chiffrée si vos données sources le sont.
   - Utilisez le chiffrement dans Amazon RDS. Vous pouvez configurer le chiffrement au repos à l'aide d'AWS Key Management Service lorsque vous créez une instance RDS.
   - Utilisez le chiffrement sur les volumes Amazon EBS. Vous pouvez configurer le chiffrement par défaut ou spécifier une clé unique lors de la création du volume.
   - Utilisez le chiffrement Amazon DynamoDB requis. DynamoDB chiffre toutes les données au repos. Vous pouvez utiliser une clé AWS KMS détenue par AWS ou une clé KMS gérée par AWS, en spécifiant une clé stockée dans votre compte.

- Chiffrez vos données stockées dans Amazon EFS. Configurez le chiffrement lorsque vous créez votre système de fichiers.
- Configurez le chiffrement dans les régions source et de destination. Vous pouvez configurer le chiffrement au repos dans Amazon S3 à l'aide de clés stockées dans KMS, mais les clés sont spécifiques à la région. Vous pouvez spécifier les clés de destination lorsque vous configurez la réplication.
- Choisissez d'utiliser le chiffrement par défaut ou le chiffrement Amazon EBS personnalisé pour Elastic Disaster Recovery. Cette option permet de chiffrer les données répliquées au repos sur les disques du sous-réseau de la zone de transit et sur les disques répliqués.

2. Mettez en œuvre les autorisations de moindre privilège pour accéder à vos sauvegardes. Suivez les bonnes pratiques pour limiter l'accès aux sauvegardes, instantanés et réplicas conformément aux bonnes pratiques de sécurité.

Ressources

Documents connexes :

- AWS Marketplace : produits pouvant être utilisés pour la sauvegarde
- Chiffrement Amazon EBS
- Protection des données Amazon S3 à l'aide du chiffrement
- Configuration supplémentaire de la réplication entre régions : réplication d'objets créés avec le chiffrement côté serveur (SSE) à l'aide de clés de chiffrement stockées dans AWS KMS
- Chiffrement DynamoDB au repos
- Chiffrement des ressources Amazon RDS
- Chiffrement des données et métadonnées dans Amazon EFS
- Chiffrement des sauvegardes dans AWS
- Gestion des tables chiffrées
- Pilier Sécurité - CadreAWS Well-Architected
- Qu'est-ce que AWS Elastic Disaster Recovery ?

Exemples connexes :

- Atelier Well-Architected : implémentation de la réplication bidirectionnelle entre régions pour Amazon S3

REL09-BP03 Effectuer automatiquement la sauvegarde des données

Configurez les sauvegardes à effectuer automatiquement en fonction d'un calendrier périodique informé par l'objectif de point de récupération (RPO) ou par les modifications du jeu de données. Les jeux de données critiques dont le seuil de tolérance pour la perte de données est faible doivent être sauvegardés automatiquement et fréquemment, tandis que les données moins critiques où certaines données peuvent être perdues peuvent être sauvegardées moins fréquemment.

Résultat souhaité : un processus automatisé qui crée des sauvegardes de sources de données à une cadence établie.

Anti-modèles courants :

- Exécution manuelle des sauvegardes
- Utilisation de ressources qui ont une capacité de sauvegarde, mais sans inclure la sauvegarde dans votre automatisation.

Avantages liés au respect de cette bonne pratique : l'automatisation des sauvegardes permet de vérifier qu'elles sont effectuées régulièrement en fonction de votre RPO, et vous alerte si elles ne sont pas effectuées.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

AWS Backup peut être utilisé pour créer des sauvegardes de données automatisées de diverses sources de données AWS. Les instances Amazon RDS peuvent être sauvegardées presque en continu toutes les cinq minutes, et les objets Amazon S3 peuvent être sauvegardés presque en continu toutes les quinze minutes, ce qui permet une récupération ponctuelle (PITR) dans l'historique de sauvegarde. Pour les autres sources de données AWS, telles que les volumes Amazon EBS, les tables Amazon DynamoDB ou les systèmes de fichiers Amazon FSx, AWS Backup peut exécuter une sauvegarde automatique qui peut avoir lieu toutes les heures. Ces services offrent également des capacités de sauvegarde natives. Les services AWS qui offrent une sauvegarde automatisée avec une récupération ponctuelle comprennent Amazon DynamoDB , Amazon RDS, et Amazon Keyspaces (pour Apache Cassandra) : ceux-ci peuvent être restaurés à un moment spécifique dans l'historique de sauvegarde. La plupart des autres services de stockage de données AWS offrent la possibilité de planifier des sauvegardes périodiques, à une fréquence de sauvegarde pouvant atteindre toutes les heures.

Amazon RDS et Amazon DynamoDB offrent une sauvegarde continue avec une récupération ponctuelle. La gestion des versions Amazon S3, une fois activée, est automatique.Amazon Data Lifecycle Manager peut automatiser la création, la copie et la suppression d'instantanés Amazon EBS. Il peut également automatiser la création, la copie, l'abandon et le désenregistrement des Amazon Machine Images (AMI) basées sur Amazon EBS et de leurs instantanés Amazon EBS sous-jacents.

AWS Elastic Disaster Recovery fournit une réplication continue au niveau des blocs depuis l'environnement source (sur site ou sur AWS) vers la région de reprise cible. Des instantanés Amazon EBS ponctuels sont automatiquement créés et gérés par le service.

AWS Backup fournit une solution de sauvegarde entièrement gérée basée sur des stratégies afin d'offrir une vue centralisée de l'automatisation et de l'historique de vos sauvegardes. Cette solution centralise et automatise la sauvegarde des données sur plusieurs services AWS dans le cloud et sur site à l'aide d'AWS Storage Gateway.

Outre la gestion des versions, Amazon S3 intègre la réplication. L'intégralité du compartiment S3 peut être automatiquement répliquée vers un autre compartiment d'une autre Région AWS.

Étapes d'implémentation

1. Identifiez les sources de données qui sont actuellement sauvegardées manuellement. Pour en savoir plus, consultez REL09-BP01 Identifier et sauvegarder toutes les données qui doivent être sauvegardées, ou reproduire les données à partir de sources. »
2. Déterminez le RPO de la charge de travail. Pour en savoir plus, consultez REL13-BP01 Définir les objectifs de reprise pour les temps d'arrêt et les pertes de données. »
3. Utilisez une solution de sauvegarde automatisée ou un service géré. AWS Backup est un service entièrement géré qui permet de centraliser et d'automatiser facilement la protection des données entre les services AWS, dans le cloud et sur site. En utilisant les plans de sauvegarde dans AWS Backup, créez des règles qui définissent les ressources à sauvegarder, et la fréquence à laquelle ces sauvegardes doivent être créées. Cette fréquence doit être informée par le RPO établi à l'étape 2. Pour obtenir des conseils pratiques sur la création de sauvegardes automatisées à l'aide de AWS Backup, consultez Test de la sauvegarde et de la restauration de données. Les fonctionnalités de sauvegarde natives sont offertes par la plupart des services AWS qui stockent des données. Par exemple, RDS peut être exploité pour les sauvegardes automatisées avec une récupération ponctuelle (PITR).
4. Pour les sources de données non prises en charge par une solution de sauvegarde automatisée ou un service géré tel que des sources de données sur site ou des files d'attente de messages,

envisagez d'utiliser une solution tierce de confiance pour créer des sauvegardes automatisées. Vous pouvez également créer une automatisation pour ce faire avec AWS CLI ou les kits SDK. Vous pouvez utiliser les fonctions AWS Lambda ou AWS Step Functions pour définir la logique impliquée dans la création d'une sauvegarde de données, et exploiter Amazon EventBridge pour l'exécuter à une fréquence basée sur votre RPO.

Niveau d'effort du plan d'implémentation : faible

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant faciliter la sauvegarde
- AWS Marketplace : produits pouvant être utilisés pour la sauvegarde
- Création d'une règle EventBridge qui se déclenche selon un calendrier
- Qu'est-ce que AWS Backup ?
- Qu'est-ce que AWS Step Functions ?
- Qu'est-ce que AWS Elastic Disaster Recovery ?

Vidéos connexes :

- AWS re:Invent 2019: Deep dive on AWS Backup, ft. Rackspace (STG341)

Exemples connexes :

- Atelier Well-Architected : test de la sauvegarde et de la restauration de données

REL09-BP04 Effectuer une récupération périodique des données pour vérifier l'intégrité et les processus de sauvegarde

Confirmez que l'implémentation de votre processus de sauvegarde répond à vos objectifs de délai de reprise (RTO) et à vos objectifs de point de reprise (RPO) en effectuant un test de reprise.

Résultat souhaité : les données des sauvegardes sont périodiquement récupérées à l'aide de mécanismes bien définis pour vérifier que la récupération est conforme à l'objectif de temps de récupération (RTO) établi pour la charge de travail. Vérifiez que la restauration à partir d'une

sauvegarde aboutit à une ressource qui contient les données d'origine sans qu'aucune d'entre elles ne soit corrompue ou inaccessible, et avec une perte de données conforme à l'objectif de point de récupération (RPO).

Anti-modèles courants :

- Restauration d'une sauvegarde, mais sans interroger ou récupérer des données pour vérifier l'utilisation de la restauration
- Supposer qu'une sauvegarde existe.
- Supposer que la sauvegarde d'un système est pleinement opérationnelle et que les données peuvent être récupérées à partir de celle-ci.
- Supposer que le temps de restauration ou de récupération des données à partir d'une sauvegarde est conforme au RTO de la charge de travail.
- Supposer que les données contenues dans la sauvegarde sont conformes au RPO de la charge de travail.
- Effectuez une restauration si nécessaire, sans utiliser de runbook ou en dehors d'une procédure automatisée établie.

Avantages liés au respect de cette bonne pratique : le test de la récupération des sauvegardes vérifie que les données peuvent être restaurées en cas de besoin sans craindre qu'elles soient manquantes ou corrompues, que la restauration et la récupération sont possibles conformément au RTO de la charge de travail et que toute perte de données est conforme au RPO de la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Tester la fonctionnalité de sauvegarde et de restauration permet de garantir que ces actions peuvent être effectuées pendant une panne. Restaurez périodiquement les sauvegardes vers un nouvel emplacement et exécutez des tests pour vérifier l'intégrité des données. Certains tests courants doivent être effectués pour vérifier que toutes les données sont disponibles, qu'elles ne sont pas corrompues, qu'elles sont accessibles et que toute perte de données ne dépasse pas le RPO de la charge de travail. Ces tests peuvent également contribuer à déterminer si les mécanismes de récupération sont suffisamment rapides pour s'adapter au RTO de la charge de travail.

Avec AWS, vous pouvez mettre en place un environnement de test et restaurer vos sauvegardes afin d'évaluer le RTO et le RPO, et exécuter des tests sur le contenu et l'intégrité des données.

De plus, Amazon RDS et Amazon DynamoDB permettent une restauration à un instant donné dans le passé (PITR). Grâce à la sauvegarde continue, vous pouvez restaurer votre jeu de données à l'état dans lequel il était à une date et une heure spécifiées.

si toutes les données sont disponibles, si elles ne sont pas corrompues, si elles sont accessibles et si toute perte de données est conforme au RPO de la charge de travail. Ces tests peuvent également contribuer à déterminer si les mécanismes de récupération sont suffisamment rapides pour s'adapter au RTO de la charge de travail.

AWS Elastic Disaster Recovery offre des instantanés de récupération ponctuelle et continue des volumes Amazon EBS. Au fur et à mesure que les serveurs sources sont répliqués, les états ponctuels sont consignés dans le temps en fonction de la stratégie configurée. Elastic Disaster Recovery vous aide à vérifier l'intégrité de ces instantanés en lançant des instances à des fins de test et d'exercice sans rediriger le trafic.

Étapes d'implémentation

1. Identifiez les sources de données qui sont actuellement sauvegardées et où ces sauvegardes sont stockées. Pour obtenir des conseils de mise en œuvre, consultez REL09-BP01 Identifier et sauvegarder toutes les données qui doivent être sauvegardées, ou reproduire les données à partir de sources. »
2. Établissez des critères de validation des données pour chaque source de données. Différents types de données ont des propriétés différentes qui pourraient nécessiter des mécanismes de validation distincts. Réfléchissez à la manière dont ces données pourraient être validées avant de vous assurer que vous pouvez les utiliser en production. Certaines méthodes courantes de validation des données consistent à utiliser des propriétés de données et de sauvegarde telles que le type de données, le format, la somme de contrôle, la taille ou une combinaison de ces propriétés avec une logique de validation personnalisée. Par exemple, il peut s'agir d'une comparaison des valeurs de somme de contrôle entre la ressource restaurée et la source de données au moment de la création de la sauvegarde.
3. Établissez le RTO et le RPO pour la restauration des données en fonction de leur criticité. Pour obtenir des conseils de mise en œuvre, consultez REL13-BP01 Définir les objectifs de reprise pour les temps d'arrêt et les pertes de données. »
4. Évaluez votre capacité de récupération. Passez en revue votre stratégie de sauvegarde et de restauration pour déterminer si elle peut répondre au RTO et au RPO, et ajustez la stratégie si nécessaire. En utilisant AWS Resilience Hub, vous pouvez effectuer une évaluation de votre charge de travail. Celle-ci se concentre sur la configuration de votre application par rapport à la stratégie de résilience et signale si vos objectifs RTO et RPO peuvent être atteints.

5. Effectuez un test de restauration avec les processus établis utilisés en production pour la restauration des données. Ces processus dépendent de la façon dont la source de données d'origine a été sauvegardée, du format et de l'emplacement de stockage de la sauvegarde elle-même, ou ils varient selon que les données sont reproduites à partir d'autres sources. Par exemple, si vous utilisez un service géré tel que AWS Backup, il peut suffire de restaurer la sauvegarde en tant que nouvelle ressource. Si vous avez utilisé AWS Elastic Disaster Recovery, vous pouvez lancer une opération de récupération.
6. Validez la récupération des données à partir de la ressource restaurée en fonction des critères que vous avez précédemment établis pour la validation des données. Les données restaurées et récupérées contiennent-elles l'enregistrement/l'élément le plus récent au moment de la sauvegarde ? Ces données sont-elles conformes au RPO de la charge de travail ?
7. Mesurez le temps nécessaire à la restauration et à la récupération et comparez-le au RTO que vous avez défini. Ce processus est-il conforme au RTO de la charge de travail ? Par exemple, comparez les horodatages du début du processus de restauration et de la fin de la validation de la récupération pour calculer la durée de ce processus. Tous les appels d'API AWS sont horodatés, et ces informations sont disponibles dans AWS CloudTrail. Bien que ces informations puissent fournir des détails sur le début du processus de restauration, l'horodatage indiquant la fin de la validation doit être enregistré par votre logique de validation. Si vous utilisez un processus automatisé, des services tels que Amazon DynamoDB permettent de stocker ces informations. En outre, de nombreux services AWS fournissent un historique des événements qui contient des informations horodatées indiquant quand certaines actions se sont produites. Au sein de AWS Backup, les actions de sauvegarde et de restauration sont appelées tâches, et ces tâches contiennent des informations d'horodatage dans le cadre de leurs métadonnées qui peuvent mesurer le temps nécessaire à la restauration et à la récupération.
8. Informez les parties prenantes si la validation des données échoue ou si le temps requis pour la restauration et la récupération dépasse le RTO établi pour la charge de travail. Lors de la mise en œuvre de l'automatisation, comme dans cet atelier, des services tels que Amazon Simple Notification Service (Amazon SNS) peuvent envoyer des notifications push, par e-mail ou SMS, aux parties prenantes. Ces messages peuvent également être publiés dans des applications de messagerie telles que Amazon Chime, Slack ou Microsoft Teams ou utilisés pour créer des tâches en tant qu'OpsItems à l'aide d'AWS Systems Manager OpsCenter.
9. Automatisez ce processus pour qu'il s'exécute périodiquement. Par exemple, des services comme AWS Lambda ou une machine d'état dans AWS Step Functions peuvent être utilisés pour automatiser les processus de restauration et de récupération, et Amazon EventBridge peut être utilisé pour déclencher périodiquement ce flux de travail d'automatisation, comme indiqué dans le diagramme d'architecture ci-dessous. Apprenez à automatiser la validation de la récupération

des données avec AWS Backup. En outre, cet atelier Well-Architected permet d'acquérir une expérience pratique sur la façon d'automatiser plusieurs des étapes décrites ici.

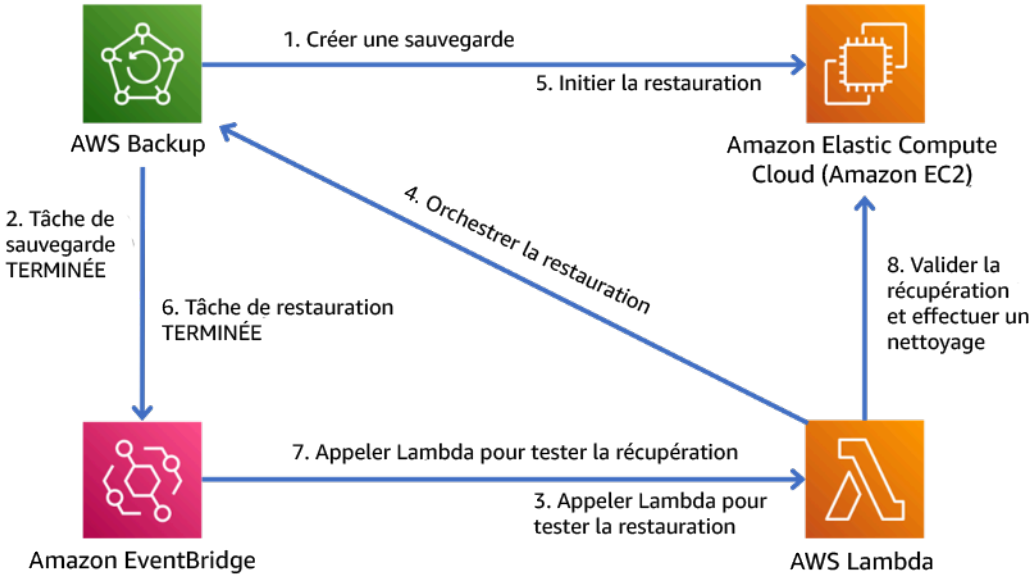


Figure 9 : Processus de sauvegarde et de restauration automatisé

Niveau d'effort du plan d'implémentation : modéré à élevé selon la complexité des critères de validation.

Ressources

Documents connexes :

- Automatiser la validation de la récupération des données avec AWS Backup
- Partenaire APN : partenaires pouvant faciliter la sauvegarde
- AWS Marketplace : produits pouvant être utilisés pour la sauvegarde
- Création d'une règle EventBridge qui se déclenche selon un calendrier
- Sauvegarde et restauration à la demande pour DynamoDB
- Qu'est-ce que AWS Backup ?
- Qu'est-ce que AWS Step Functions ?
- Qu'est-ce que AWS Elastic Disaster Recovery
- AWS Elastic Disaster Recovery

Exemples connexes :

- Atelier Well-Architected : test de la sauvegarde et de la restauration de données

FIA 10. Comment utilisez-vous l'isolement des pannes pour protéger votre charge de travail ?

Les limites isolées pour les défaillances limitent l'effet d'une défaillance au sein d'une charge de travail à un nombre limité de composants. Les composants en dehors de la limite ne sont pas affectés par la défaillance. En utilisant plusieurs limites isolées par défaut, vous pouvez limiter l'impact sur votre charge de travail.

Bonnes pratiques

- REL10-BP01 Déployer la charge de travail sur plusieurs emplacements
- REL10-BP02 Sélectionner les emplacements appropriés pour votre déploiement multisite
- REL10-BP03 Automatiser la récupération pour les composants limités à un seul emplacement
- REL10-BP04 Utiliser des architectures cloisonnées pour limiter la portée de l'impact

REL10-BP01 Déployer la charge de travail sur plusieurs emplacements

Distribuez les données et les ressources de charge de travail sur plusieurs zones de disponibilité ou, si nécessaire, entre Régions AWS. Ces emplacements peuvent être aussi variés que nécessaire.

L'un des principes fondamentaux de la conception de services dans AWS est d'éviter les points de défaillance uniques dans l'infrastructure physique sous-jacente. Cela nous motive à développer des logiciels et des systèmes qui utilisent plusieurs zones de disponibilité et sont résilients à la défaillance d'une seule zone. De la même manière, les systèmes sont conçus pour être résilients à la défaillance d'un seul nœud de calcul, d'un seul volume de stockage ou d'une seule instance de base de données. Lors de la création d'un système qui s'appuie sur des composants redondants, il est important de s'assurer que les composants fonctionnent de manière indépendante et, dans le cas de Régions AWS, de façon autonome. Les avantages obtenus grâce aux calculs de disponibilité théoriques avec des composants redondants ne sont valides qu'à cette condition.

Zones de disponibilité (AZ)

Les Régions AWS sont composées de plusieurs zones de disponibilité, toutes conçues pour être indépendantes les unes des autres. Chaque zone de disponibilité est séparée par une distance physique logique des autres zones afin d'éviter les scénarios de défaillance corrélées liés à des risques environnementaux comme les incendies, les inondations et les tornades. Chaque zone de

disponibilité comporte également une infrastructure physique indépendante : connexions dédiées à l'alimentation, sources d'énergie d'appoint indépendantes, services mécaniques indépendants et connectivité réseau indépendante dans et au-delà de la zone de disponibilité. Ce modèle limite les défaillances susceptibles de se produire dans l'un de ces systèmes à la seule zone de disponibilité affectée. Bien que géographiquement séparées, les zones de disponibilité sont situées dans la même zone régionale, ce qui permet une mise en réseau à haut débit et à faible latence. L'intégralité de la Région AWS (dans toutes les zones de disponibilité, composées de plusieurs centres de données physiquement indépendants) peut être traitée comme une cible de déploiement logique unique pour votre charge de travail, y compris pour répliquer les données de manière synchrone (par exemple, entre les bases de données). Cela vous permet d'utiliser les zones de disponibilité dans une configuration active/active ou active/veille.

Les zones de disponibilité sont indépendantes et, par conséquent, la disponibilité de la charge de travail est augmentée lorsque la charge de travail est conçue pour utiliser plusieurs zones. Certains services AWS (y compris le plan de données d'instance Amazon EC2) sont déployés en tant que services strictement locaux où leur sort dépend de la zone de disponibilité dans laquelle ils se trouvent. Cependant, les instances Amazon EC2 dans les autres AZ ne sont pas affectées et continuent à fonctionner . De même, si une défaillance dans une zone de disponibilité entraîne l'échec d'une base de données Amazon Aurora, une instance de réplica en lecture Aurora dans une AZ non affectée peut être automatiquement promue comme instance principale. D'autre part, les services régionaux AWS, comme Amazon DynamoDB, utilisent en interne plusieurs zones de disponibilité dans une configuration active/active pour atteindre les objectifs de conception de disponibilité de ce service, sans que vous ayez besoin de configurer le placement AZ.

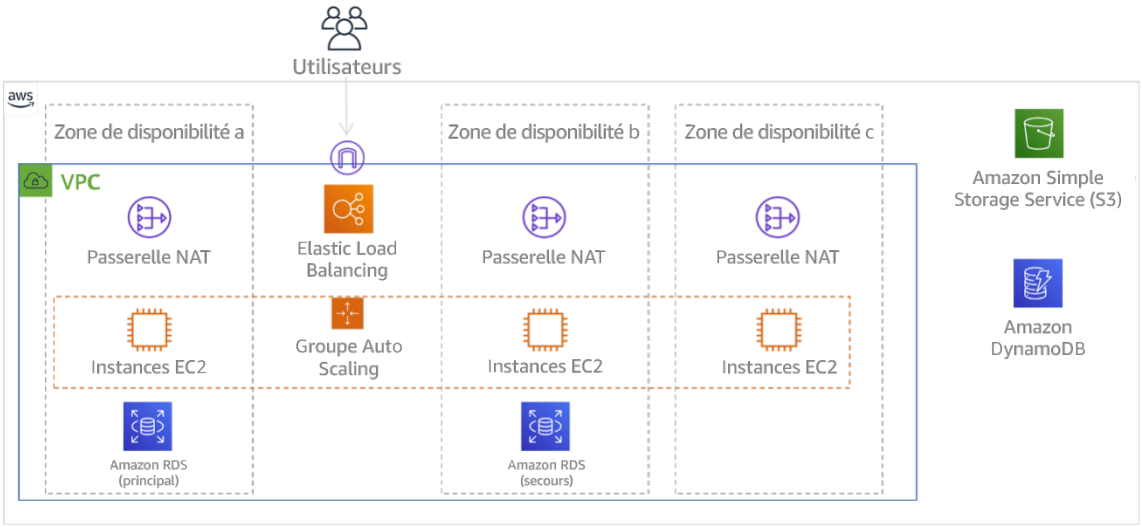


Figure 9 : Architecture multiniveau déployée sur trois zones de disponibilité. Notez qu'Amazon S3 et Amazon DynamoDB comportent toujours automatiquement plusieurs zones de disponibilité. L'ELB est également déployé dans les trois zones.

Bien que les plans de contrôle AWSoffrent généralement la possibilité de gérer les ressources sur l'ensemble de la région (plusieurs zones de disponibilité), certains plans de contrôle (y compris Amazon EC2 et Amazon EBS) ont la capacité de filtrer les résultats en une seule zone de disponibilité. Lorsque c'est le cas, la requête est traitée uniquement dans la zone de disponibilité spécifiée, ce qui réduit l'exposition aux perturbations dans les autres zones de disponibilité. Cet exemple AWS CLI illustre l'obtention d'informations sur l'instance Amazon EC2 uniquement à partir de la zone de disponibilité us-east-2c :

```
AWS ec2 describe-instances --filters Name=availability-zone,Values=us-east-2c
```

AWS Local Zones

Les AWS Local Zones agissent de la même manière que les zones de disponibilité au sein de leur Région AWS respective, car elles sont sélectionnables en tant qu'emplacement de placement des ressources AWS de la zone comme les sous-réseaux et les instances EC2. Elles sont spéciales, car elles ne sont pas situées dans la Région AWS associée, mais près de grands centres de population, industriels et informatiques où aucune Région AWS n'existe actuellement. Cependant, elles conservent une connexion sécurisée et à bande passante élevée entre les charges de travail locales dans la zone locale et celles exécutées dans la Région AWS. Utilise les AWS Local Zones pour déployer des charges de travail plus près de vos utilisateurs afin de répondre aux exigences de faible latence.

Réseau périphérique mondial d'Amazon

Le réseau périphérique mondial d'Amazon se compose d'emplacements périphériques dans des villes du monde entier. Amazon CloudFront utilise ce réseau pour fournir du contenu aux utilisateurs finaux avec une latence plus faible. AWS Global Accelerator vous permet de créer vos points de terminaison de charge de travail dans ces emplacements périphériques pour assurer l'intégration au réseau mondial AWS à proximité de vos utilisateurs. Amazon API Gateway active les points de terminaison d'API optimisés en périphérie à l'aide d'une distribution CloudFront pour faciliter l'accès client via l'emplacement périphérique le plus proche.

Régions AWS

Les Régions AWS sont conçues pour être autonomes. Par conséquent, pour utiliser une approche multirégion, vous devez déployer des copies dédiées des services dans chaque région.

Une approche multirégion est courante pour que les stratégies de reprise après sinistre atteignent les objectifs de récupération lorsque des événements ponctuels à grande échelle se produisent. Consulter Planification de la reprise après sinistre pour en savoir plus sur ces stratégies. Ici cependant, nous nous concentrons plutôt sur la disponibilité, qui vise à atteindre un objectif de disponibilité moyenne dans le temps. Pour des objectifs de haute disponibilité, une architecture multirégion est généralement conçue pour être active/active, où chaque copie de service (dans ses régions respectives) est active (en répondant aux requêtes).

> **Recommandations**
>
> Les objectifs de disponibilité pour la plupart des charges de travail peuvent être satisfaits à l'aide d'une stratégie multi-AZ dans une seule Région AWS. Envisagez les architectures multirégion uniquement lorsque les charges de travail ont des exigences de disponibilité extrêmes ou en cas d'autres objectifs commerciaux nécessitant une architecture multirégion.

AWS vous offre la possibilité de gérer des services entre régions. Par exemple, AWS fournit une réplication continue et asynchrone des données via la réplication Amazon Simple Storage Service (Amazon S3), les réplicas en lecture Amazon RDS (y compris les réplicas en lecture Aurora) et les tables globales Amazon DynamoDB. Grâce à la réplication continue, des versions de vos données sont disponibles pour une utilisation quasi immédiate dans chacune de vos régions actives.

Avec AWS CloudFormation, vous pouvez définir votre infrastructure et la déployer de manière cohérente sur les Comptes AWS et dans les Régions AWS. AWS CloudFormation StackSets étend cette fonctionnalité en vous permettant de créer, mettre à jour ou supprimer des piles AWS CloudFormation sur plusieurs comptes et régions en une seule opération. Pour les déploiements d'instance Amazon EC2, une AMI (Amazon Machine Image) est utilisée pour fournir des informations telles que la configuration matérielle et les logiciels installés. Vous pouvez implémenter un pipeline Amazon EC2 Image Builder qui crée les AMI dont vous avez besoin et les copie dans vos régions actives. Cela garantit que ces AMI approuvées disposent de tout ce dont vous avez besoin pour déployer et faire évoluer votre charge de travail dans chaque nouvelle région.

Pour acheminer le trafic, Amazon Route 53 et AWS Global Accelerator permettent la définition de politiques qui déterminent quels utilisateurs accèdent à quel point de terminaison régional actif. Avec Global Accelerator, vous définissez une option de trafic pour contrôler le pourcentage de trafic

dirigé vers chaque point de terminaison d'application. Route 53 prend en charge cette approche de pourcentage, ainsi que plusieurs autres politiques disponibles, notamment celles basées sur la géoproximité et la latence. Global Accelerator exploite automatiquement le vaste réseau de serveurs périphériques AWS pour intégrer le trafic à la dorsale du réseau AWS dès que possible, ce qui réduit la latence des demandes.

L'ensemble de ces fonctionnalités opèrent de manière à préserver l'autonomie de chaque région. Il existe quelques rares exceptions à cette approche, y compris nos services qui fonctionnent en périphérie au niveau mondial (comme Amazon CloudFront et Amazon Route 53), ainsi que le plan de contrôle pour le service AWS Identity and Access Management (IAM) La plupart des services fonctionnent entièrement au sein d'une seule région.

Centre de données sur site

Pour les charges de travail exécutées dans un centre de données sur site, concevez une expérience hybride dans la mesure du possible. AWS Direct Connect fournit une connexion réseau dédiée entre vos locaux et AWS, ce qui vous permet d'utiliser les deux.

Une autre option consiste à exécuter l'infrastructure et les services AWS sur site à l'aide d' AWS Outposts. AWS Outposts est un service entièrement géré qui étend l'infrastructure AWS, les services AWS, les API et les outils à votre centre de données. La même infrastructure matérielle utilisée dans le AWS Cloud est installée dans votre centre de données. Les services AWS Outposts sont alors connectés à la Région AWS la plus proche. Vous pouvez ensuite utiliser les services AWS Outposts pour prendre en charge vos charges de travail à faible latence ou vos exigences en matière de traitement des données locales.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Utilisez plusieurs zones de disponibilité et Régions AWS. Distribuez les données et les ressources de charge de travail sur plusieurs zones de disponibilité ou, si nécessaire, entre Régions AWS. Ces emplacements peuvent être aussi variés que nécessaire.
  - Les services régionaux sont, par nature, déployés entre les zones de disponibilité.
    - Cela inclut Amazon S3, Amazon DynamoDB et AWS Lambda (lorsqu'ils ne sont pas connectés à un VPC).
  - Déployez votre conteneur, votre instance et vos charges de travail basées sur des fonctions dans plusieurs zones de disponibilité. Utilisez les magasins de données multi-AZ, y compris les caches. Utilisez les fonctionnalités d'EC2 Auto Scaling, le placement de tâches ECS,

la configuration de fonctions AWS Lambda lors de l'exécution de votre VPC et les clusters ElastiCache.

- Utilisez les sous-réseaux situés dans des zones de disponibilité distinctes lorsque vous déployez des groupes Auto Scaling.
  - Exemple : Distribution d'instances dans des zones de disponibilité
  - Stratégies de placement des tâches Amazon ECS
  - Configuration d'une fonction AWS Lambda pour accéder aux ressources dans un Amazon VPC
  - Choix des régions et des zones de disponibilité
- Utilisez des sous-réseaux situés dans des zones de disponibilité distinctes lorsque vous déployez des groupes Auto Scaling.
  - Exemple : Distribution d'instances dans des zones de disponibilité
- Utilisez les paramètres de placement des tâches ECS, en spécifiant des groupes de sous-réseaux de base de données.
  - Stratégies de placement des tâches Amazon ECS
- Utilisez des sous-réseaux dans plusieurs zones de disponibilité lorsque vous configurez une fonction à exécuter dans votre VPC.
  - Configuration d'une fonction AWS Lambda pour accéder aux ressources dans un Amazon VPC
- Utilisez plusieurs zones de disponibilité avec des clusters ElastiCache.
  - Choix des régions et des zones de disponibilité

- Choisissez une stratégie sur plusieurs régions si votre charge de travail doit être déployée dans plusieurs régions. La plupart des besoins de fiabilité peuvent être satisfaits au sein d'une même Région AWS à l'aide d'une stratégie à plusieurs zones de disponibilité. Utilisez une stratégie sur plusieurs régions si nécessaire pour répondre aux besoins de votre entreprise.
  - AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
    - La sauvegarde dans une autre Région AWS peut donner des garanties supplémentaires que les données seront disponibles en cas de besoin.
    - Il existe, pour certaines charges de travail, des exigences réglementaires qui imposent l'utilisation d'une stratégie sur plusieurs régions.
- Évaluez AWS Outposts pour votre charge de travail. Si votre charge de travail nécessite une faible latence de connexion à votre centre de données sur site ou si elle a des exigences locales en

matière de traitement des données, Exécutez ensuite l'infrastructure et les services AWS sur site à l'aide d'AWS Outposts.

  - Qu'est-ce qu'AWS Outposts ?
- Déterminez si AWS Local Zones vous permet de fournir un service à vos utilisateurs. Si vous avez des exigences de faible latence, vérifiez si AWS Local Zones est situé près de vos utilisateurs. Si oui, utilisez-le pour déployer des charges de travail plus près de ces utilisateurs.
  - FAQ sur AWS Local Zones

Ressources

Documents connexes :

- Infrastructure mondiale AWS
- FAQ sur AWS Local Zones
- Stratégies de placement des tâches Amazon ECS
- Choix des régions et des zones de disponibilité
- Exemple : Distribution d'instances dans des zones de disponibilité
- Tables globales : réplication multirégions avec DynamoDB
- Utilisation des bases de données mondiales Amazon Aurora
- Série de blog sur la création d'une application multirégion avec les services AWS
- Qu'est-ce qu'AWS Outposts ?

Vidéos connexes :

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- AWS re:Invent 2019: Innovation and operation of the AWS global network infrastructure (NET339)

REL10-BP02 Sélectionner les emplacements appropriés pour votre déploiement multisite

Résultat souhaité

Pour une haute disponibilité, déployez toujours (si possible) vos composants de charge de travail sur plusieurs zones de disponibilité (AZ), comme illustré à la figure 10. Pour les charges de travail

avec des exigences de résilience extrêmes, évaluez soigneusement les options pour une architecture multirégion.

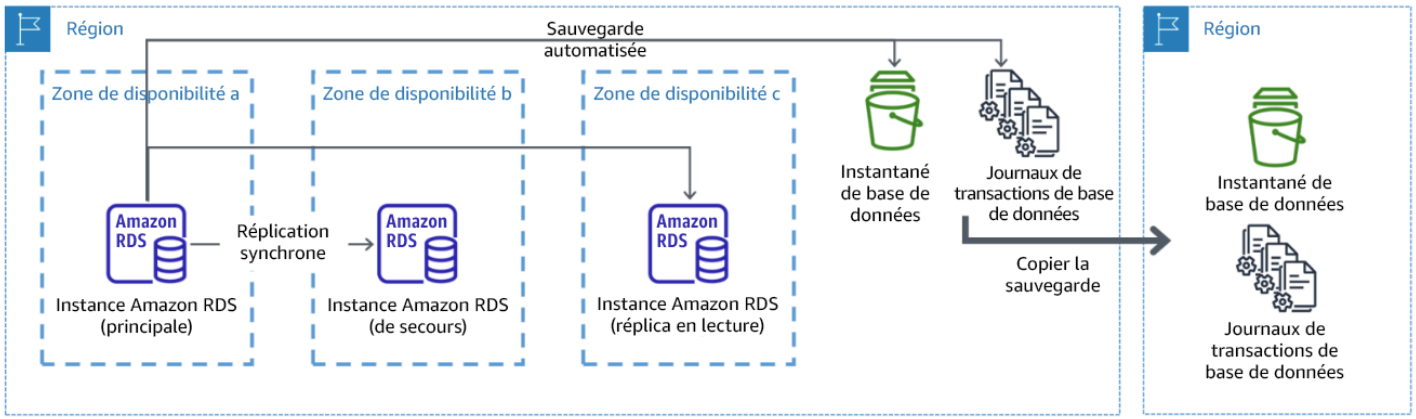


Figure 10 : déploiement d'une base de données multi-AZ résiliente avec sauvegarde dans une autre région AWS

Anti-modèles courants

- Choisir de concevoir une architecture multirégion alors qu'une architecture multi-AZ répond aux exigences.
- Ne pas tenir compte des dépendances entre les composants de l'application si les exigences en matière de résilience et d'emplacements multiples diffèrent entre ces composants.

Avantages liés au respect de cette bonne pratique :

Pour la résilience, vous devez adopter une approche qui repose sur des couches de défense. Une couche protège contre les perturbations de petite envergure et courantes en créant une architecture hautement disponible à l'aide de plusieurs AZ. Une autre couche de défense est destinée à protéger contre les événements rares tels que les catastrophes naturelles généralisées et les perturbations au niveau régional. Cette deuxième couche implique de concevoir l'architecture de votre application pour qu'elle s'étende sur plusieurs Régions AWS.

- La différence entre une disponibilité de 99,5 % et une disponibilité de 99,99 % est de plus de 3,5 heures par mois. La disponibilité attendue d'une charge de travail ne peut atteindre les « quatre neuf » que si elle se trouve dans plusieurs zones de disponibilité.
- En exécutant votre charge de travail dans plusieurs AZ, vous pouvez isoler les pannes d'alimentation, de refroidissement et de mise en réseau, ainsi que la plupart des catastrophes naturelles telles que les incendies et les inondations.

- La mise en œuvre d'une stratégie multirégion pour votre charge de travail permet de la protéger contre les catastrophes naturelles généralisées qui affectent une grande région géographique d'un pays, ou les défaillances techniques à l'échelle régionale. Sachez que la mise en œuvre d'une architecture multirégion peut être très complexe et n'est généralement pas requise pour la plupart des charges de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

Dans le cas d'une catastrophe basée sur l'interruption ou la perte partielle d'une zone de disponibilité, la mise en œuvre d'une charge de travail hautement disponible dans plusieurs zones de disponibilité au sein d'une seule Région AWS permet d'atténuer les effets des catastrophes naturelles et techniques. Chaque Région AWS est composé de plusieurs zones de disponibilité, chacune isolée des défaillances des autres zones et séparée par une distance significative. Cependant, dans le cas d'une catastrophe incluant le risque de perdre plusieurs composants de la zone de disponibilité (composants qui sont à une distance importante les uns des autres), vous devez implémenter des options de reprise après sinistre pour vous prémunir contre les défaillances à l'échelle de la région. Pour les charges de travail nécessitant une résilience extrême (infrastructure critique, applications liées à la santé, infrastructure du système financier, etc.), une stratégie multirégion peut être nécessaire.

Étapes d'implémentation

1. Évaluez votre charge de travail et déterminez si les besoins de résilience peuvent être satisfaits par une approche multi-AZ (Région AWS unique) ou s'ils nécessitent une approche multirégion. La mise en œuvre d'une architecture multirégion pour répondre à ces exigences ajoute de la complexité. Par conséquent, examinez attentivement votre cas d'utilisation et ses exigences. Les exigences de résilience peuvent presque toujours être satisfaites avec une seule Région AWS. Tenez compte des exigences possibles suivantes lorsque vous déterminez si vous devez utiliser plusieurs régions :
   a. Reprise après sinistre : dans le cas d'une catastrophe basée sur l'interruption ou la perte partielle d'une zone de disponibilité, la mise en œuvre d'une charge de travail hautement disponible dans plusieurs zones de disponibilité au sein d'une seule Région AWS permet d'atténuer les effets des catastrophes naturelles et techniques. Dans le cas d'une catastrophe incluant le risque de perdre plusieurs composants de la zone de disponibilité (composants qui sont à une distance importante les uns des autres), vous devez implémenter des options

de reprise après sinistre entre plusieurs régions pour vous prémunir contre les catastrophes naturelles et les incidents techniques à l'échelle de la région.

b. Haute disponibilité : une architecture multirégion (utilisant plusieurs AZ dans chaque région) peut être utilisée pour atteindre une disponibilité supérieure à quatre 9 (> 99,99 %).

c. Localisation des piles : lors du déploiement d'une charge de travail auprès d'une audience mondiale, vous pouvez déployer des piles localisées dans différentes Régions AWS pour répondre aux besoins des audiences dans ces régions. La localisation peut inclure la langue, la devise et les types de données stockées.

d. Proximité avec les utilisateurs : lors du déploiement d'une charge de travail auprès d'une audience mondiale, vous pouvez réduire la latence en déployant des piles dans les Régions AWS à proximité de l'endroit où se trouvent les utilisateurs finaux.

e. Résidence des données : certaines charges de travail sont soumises à des exigences en matière de situation géographique des données, où les données de certains utilisateurs doivent rester à l'intérieur des frontières d'un pays spécifique. En fonction de la réglementation en question, vous pouvez choisir de déployer une pile entière, ou uniquement les données, dans la Région AWS au sein de ces frontières.

2. Voici quelques exemples de fonctionnalités multi-AZ fournies par les services AWS :

a. Pour protéger les charges de travail à l'aide d'EC2 ou d'ECS, déployez un Elastic Load Balancer devant les ressources de calcul. Elastic Load Balancing fournira ensuite la solution pour détecter les instances dans les zones non saines et acheminer le trafic vers les zones qui le sont.

   i. Démarrer avec Application Load Balancers

   ii. Démarrer avec les Network Load Balancers

b. Dans le cas d'instances EC2 exécutant des logiciels commerciaux prêts à l'emploi qui ne prennent pas en charge l'équilibrage de charge, vous pouvez bénéficier d'une forme de tolérance aux pannes via la mise en œuvre d'une méthodologie de reprise après sinistre multi-AZ.

   i. the section called “REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise”

c. Pour les tâches Amazon ECS, déployez votre service uniformément sur trois AZ pour atteindre un juste équilibre entre disponibilité et coût.

   i. Bonnes pratiques de disponibilité Amazon ECS | Conteneurs

d. Pour les tâches qui n'entrent pas dans le cadre d'Aurora Amazon RDS, vous pouvez choisir Multi-AZ comme option de configuration. En cas de défaillance de l'instance de base de

données principale, Amazon RDS promeut automatiquement une base de données de secours pour recevoir le trafic dans une autre zone de disponibilité. Des réplicas en lecture multirégion peuvent également être créés pour améliorer la résilience.

   i. Déploiements multi-AZ Amazon RDS

   ii. Création d'un réplica en lecture dans une autre Région AWS

3. Voici quelques exemples de fonctionnalités multirégions fournies par les services AWS :

   a. Pour les charges de travail Amazon S3, où la disponibilité multi-AZ est fournie automatiquement par le service, envisagez des points d'accès multirégions si un déploiement multirégion est nécessaire.

      i. Points d'accès multirégions dans Amazon S3

   b. Pour les tables DynamoDB, où la disponibilité multi-AZ est fournie automatiquement par le service, vous pouvez facilement convertir les tables existantes en tables globales pour tirer parti de plusieurs régions.

      i. Convertir vos tables Amazon DynamoDB à région unique en tables globales

   c. Si votre charge de travail repose sur des Application Load Balancers ou des Network Load Balancers, utilisez AWS Global Accelerator pour améliorer la disponibilité de votre application en dirigeant le trafic vers plusieurs régions qui contiennent des points de terminaison sains.

      i. Points de terminaison pour les accélérateurs standards dans AWS Global Accelerator - AWS Global Accelerator (amazon.com)

   d. Pour les applications qui tirent parti d'AWS EventBridge, envisagez des bus interrégionaux pour transférer les événements vers d'autres régions que vous sélectionnez.

      i. Envoi et réception d'événements Amazon EventBridge entre les Régions AWS

   e. Pour les bases de données Amazon Aurora, considérez les bases de données globales Aurora, qui couvrent plusieurs régions AWS. Les clusters existants peuvent également être modifiés pour ajouter de nouvelles régions.

      i. Premiers pas avec les avec les bases de données globales Amazon Aurora

   f. Si votre charge de travail comprend des clés de chiffrement AWS Key Management Service (AWS KMS), déterminez si les clés multirégions sont appropriées pour votre application.

      i. Clés multirégions dans AWS KMS

   g. Pour d'autres fonctionnalités de service AWS, consultez cette série de blog sur la création d'une application multirégion avec les services AWS

Niveau d'effort du plan d'implémentation : De Modéré à Élevé

Ressources

Documents connexes :

- Création d'une application multirégion avec les services AWS
- Architecture de reprise après sinistre sur AWS, partie 4 : multisite actif/actif
- Infrastructure mondiale AWS
- FAQ sur AWS Local Zones
- Architecture de reprise après sinistre (DR) sur AWS, première partie : stratégies de reprise dans le cloud
- La reprise après sinistre est différente dans le cloud
- Tables globales : réplication multirégions avec DynamoDB

Vidéos connexes :

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- Auth0 : architecture à haute disponibilité sur plusieurs régions pouvant atteindre plus de 1,5 milliard de connexions par mois avec basculement automatique

Exemples connexes :

- Architecture de reprise après sinistre (DR) sur AWS, première partie : stratégies de reprise dans le cloud
- DTCC atteint une résilience bien supérieure à ce qu'elle peut obtenir sur site
- Expedia Group utilise une architecture reposant sur plusieurs zones de disponibilité et plusieurs régions avec un service DNS propriétaire pour renforcer la résilience des applications
- Uber : reprise après sinistre pour une plateforme Kafka sur plusieurs régions
- Netflix : stratégie active-active pour une résilience sur plusieurs régions
- Comment nous créons la résidence des données pour le cloud Atlassian
- Intuit TurboTax fonctionne sur deux régions

REL10-BP03 Automatiser la récupération pour les composants limités à un seul emplacement

Si les composants de la charge de travail ne peuvent s'exécuter que dans une seule zone de disponibilité ou un centre de données sur site, implémentez la capacité permettant d'effectuer une reconstruction complète de la charge de travail dans le cadre de vos objectifs de reprise définis.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Si la bonne pratique de déploiement de la charge de travail sur plusieurs emplacements n'est pas possible en raison de contraintes technologiques, vous devez implémenter une autre solution de résilience. Vous devez automatiser la possibilité de recréer l'infrastructure nécessaire, de redéployer les applications et de recréer les données nécessaires pour ces situations.

Par exemple, Amazon EMR lance tous les nœuds d'un cluster donné dans la même zone de disponibilité, car l'exécution d'un cluster dans la même zone améliore les performances des flux de travail en fournissant un taux d'accès aux données plus élevé. Si ce composant est requis pour la résilience de la charge de travail, vous devez pouvoir redéployer le cluster et ses données. De même, pour Amazon EMR, vous devez assurer la redondance autrement qu'en utilisant plusieurs zones de disponibilité. Vous pouvez provisionner plusieurs nœuds. Avec le système de fichiers EMR (EMRFS), les données EMR peuvent être conservées dans Amazon S3, et ainsi être répliquées sur plusieurs zones de disponibilité ou Régions AWS.

De même, Amazon Redshift met en service, par défaut, votre cluster dans une zone de disponibilité sélectionnée de façon aléatoire au sein de la Région AWS que vous sélectionnez. Tous les nœuds de cluster sont mis en service dans la même zone.

Pour les charges de travail basées sur des serveurs avec état déployés dans un centre de données sur site, vous pouvez utiliser AWS Elastic Disaster Recovery pour protéger vos charges de travail dans AWS. Si votre charge de travail est déjà hébergée dans AWS, vous pouvez utiliser Elastic Disaster Recovery pour la protéger dans une autre zone ou région de disponibilité. Elastic Disaster Recovery utilise une réplication continue au niveau des blocs vers une zone de transit légère pour fournir une restauration rapide et fiable des applications sur site et dans le cloud.

Étapes d'implémentation

1. Implémentation de l'autorégénération Dans la mesure du possible, déployez vos instances ou vos conteneurs en utilisant la scalabilité automatique. Si vous ne pouvez pas utiliser la scalabilité

automatique, utilisez la récupération automatique pour les instances EC2 ou mettez en place un mécanisme d'autoréparation basé sur Amazon EC2 ou des événements de cycle de vie de conteneur ECS.

- Utilisez les groupes Amazon EC2 Auto Scaling pour les instances et les charges de travail de conteneur qui n'ont aucune exigence en matière d'adresse IP d'instance, d'adresse IP privée, d'adresse IP élastique et de métadonnées d'instance.
  - Les données utilisateur du modèle de lancement peuvent être utilisées pour mettre en place un mécanisme permettant la récupération automatique de la plupart des charges de travail.
- Utilisez la récupération automatique des instances Amazon EC2 pour les charges de travail nécessitant une seule adresse d'ID d'instance, une seule adresse IP privée, une seule adresse IP élastique, et des métadonnées d'instance.
  - La récupération automatique envoie des alertes de statut de récupération à une rubrique SNS lorsque la défaillance de l'instance est détectée.
- Utilisez les événements du cycle de vie de l'instance Amazon EC2 ou les événements Amazon ECS pour automatiser l'autoréparation lorsque la scalabilité automatique ou la récupération de votre instance EC2 ne peuvent pas être utilisées.
  - Utilisez les événements pour appeler le mécanisme vous permettant de réparer votre composant selon la logique de processus dont vous avez besoin.
- Protégez les charges de travail avec état qui sont limitées à un seul emplacement en utilisant AWS Elastic Disaster Recovery.

Ressources

Documents connexes :

- Événements Amazon ECS
- Hooks de cycle de vie Amazon EC2 Auto Scaling
- Récupérer votre instance.
- Scalabilité automatique des services
- Qu'est-ce que Amazon EC2 Auto Scaling ?
- AWS Elastic Disaster Recovery

REL10-BP04 Utiliser des architectures cloisonnées pour limiter la portée de l'impact

Mettez en œuvre des architectures de cloisonnement (également connues sous le nom d'architectures cellulaires) pour restreindre l'effet d'une panne au sein d'une charge de travail à un nombre limité de composants.

Résultat souhaité : une architecture cellulaire utilise plusieurs instances isolées d'une charge de travail, où chaque instance est appelée cellule. Chaque cellule est indépendante, ne partage pas d'état avec les autres cellules et traite un sous-ensemble des demandes de la charge de travail globale. Cela réduit l'impact potentiel d'une panne, telle qu'une mauvaise mise à jour logicielle, sur une cellule individuelle et les demandes qu'elle traite. Si une charge de travail utilise 10 cellules pour traiter 100 demandes, lorsqu'une panne survient, 90 % des demandes globales ne sont pas affectées par la panne.

Anti-modèles courants :

- Permettre aux cellules de se développer sans limites.
- Appliquer des mises à jour ou des déploiements de code à toutes les cellules en même temps.
- Partage de l'état ou des composants entre les cellules (à l'exception de la couche routeur).
- Ajout d'une logique métier ou de routage complexe à la couche routeur.
- Ne pas minimiser les interactions entre les cellules.

Avantages liés au respect de cette bonne pratique : avec les architectures cellulaires, de nombreux types de pannes courants sont contenus dans la cellule elle-même, ce qui permet un isolement supplémentaire des pannes. Ces limites de défaillance peuvent fournir une résilience contre des types de panne qui seraient autrement difficiles à contenir, tels que des déploiements de code infructueux ou des requêtes corrompues ou déclenchant un mode de défaillance spécifique (également connues sous le nom de requêtes empoisonnées).

Directives d'implémentation

Sur un navire, les cloisons permettent de contenir une brèche dans la coque dans une seule section de la coque. Dans les systèmes complexes, ce modèle est souvent répliqué pour permettre d'isoler des pannes. Les limites isolées pour les défaillances restreignent l'effet d'une panne au sein d'une charge de travail à un nombre limité de composants. Les composants en dehors de la limite ne sont pas affectés par la défaillance. En utilisant plusieurs limites isolées par défaut, vous pouvez limiter l'impact sur votre charge de travail. Sur AWS, les clients peuvent utiliser plusieurs zones

de disponibilité et régions pour isoler des pannes, mais le concept d'isolement des pannes peut également être étendu à l'architecture de votre charge de travail.

La charge de travail globale est divisée en cellules par une clé de partition. Cette clé doit s'aligner sur la base de granularité du service, ou sur la manière naturelle dont la charge de travail d'un service peut être subdivisée avec un minimum d'interactions entre les cellules. Des exemples de clés de partition sont l'ID du client, l'ID de la ressource ou tout autre paramètre facilement accessible dans la plupart des appels d'API. Une couche de routage des cellules distribue les requêtes aux cellules individuelles en fonction de la clé de partition et présente un point de terminaison unique aux clients.

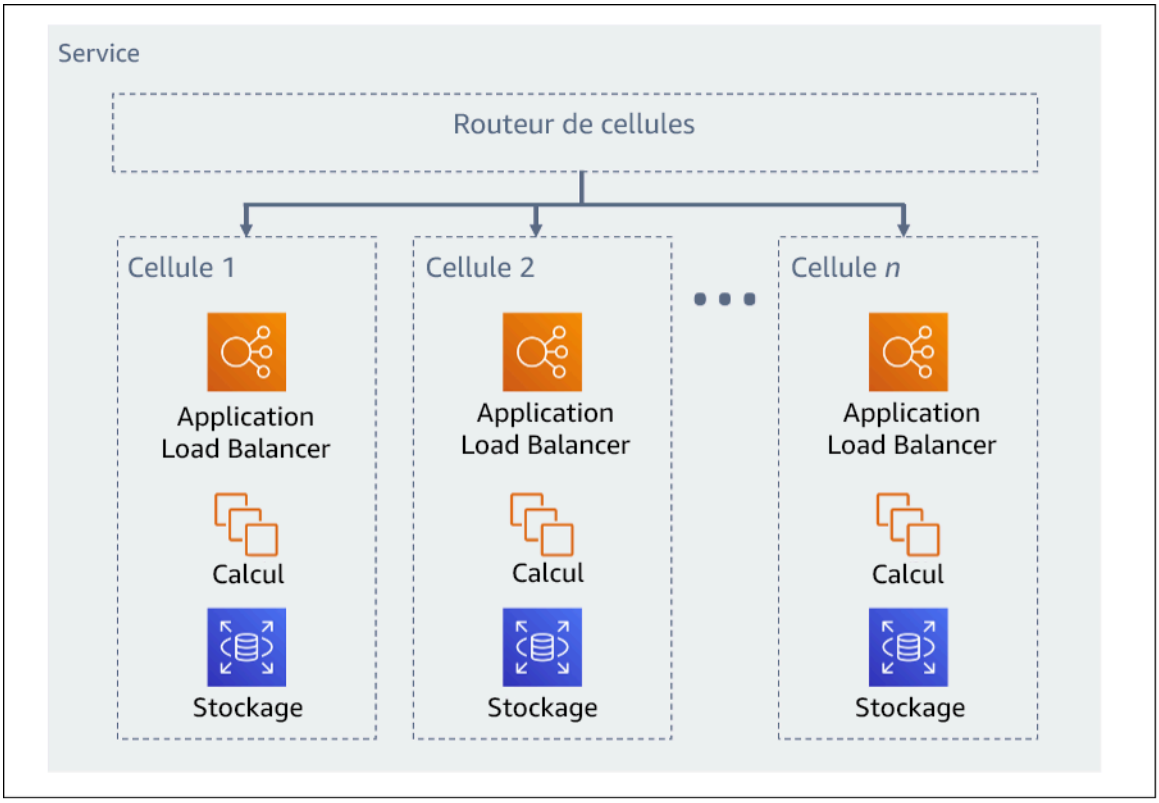


Figure 11 : Architecture cellulaire

Étapes d'implémentation

Lors de la conception d'une architecture cellulaire, vous devez tenir compte de plusieurs éléments :

1. Clé de partition : il convient d'accorder une attention particulière au choix de la clé de partition.
    - Celle-ci doit s'aligner sur la base de granularité du service, ou sur la manière naturelle dont la charge de travail d'un service peut être subdivisée avec un minimum d'interactions entre les cellules. Voici quelques exemples : `ID du client` ou `ID de la ressource`. »

- La clé de partition doit être disponible dans toutes les requêtes, soit directement, soit d'une manière qui pourrait être facilement déduite de façon déterministe par d'autres paramètres.

2. Mappage persistant des cellules : les services en amont ne doivent interagir qu'avec une seule cellule pendant le cycle de vie de leurs ressources.
   - En fonction de la charge de travail, vous devrez peut-être concevoir une stratégie de migration de cellules pour faire migrer les données d'une cellule à l'autre. La migration d'une cellule peut s'avérer nécessaire si un utilisateur ou une ressource particulière de votre charge de travail devient trop importante et nécessite une cellule dédiée.
   - Les cellules ne doivent pas partager d'état ou de composants entre elles.
   - Par conséquent, les interactions entre cellules doivent être évitées ou réduites au minimum, car elles créent des dépendances entre les cellules et diminuent donc les bénéfices de l'isolement des défaillances.
3. Couche routeur : la couche routeur est un composant partagé entre les cellules, et ne peut donc pas suivre la stratégie de compartimentage des cellules.
   - Nous recommandons de paramétrer la couche routeur pour distribuer les requêtes à des cellules individuelles à l'aide d'un algorithme de mappage de partition d'une manière efficace sur le plan des calculs. Par exemple, en combinant des fonctions de hachage cryptographiques et de l'arithmétique modulaire pour mapper les clés de partition aux cellules.
   - Pour éviter les impacts sur plusieurs cellules, la couche de routage doit rester aussi simple et évolutive horizontalement que possible, ce qui nécessite d'éviter toute logique métier complexe au sein de cette couche. Cela présente l'avantage supplémentaire de faciliter la compréhension de son comportement attendu à tout moment, favorisant ainsi une testabilité approfondie. Comme l'explique Colm MacCárthaigh dans son article Reliability, constant work, and a good cup of coffee (Fiabilité, travail constant, et une bonne tasse de café), les conceptions simples et les modèles de travail constants produisent des systèmes fiables et réduisent le phénomène d'antifragilité.
4. Taille des cellules : les cellules doivent comporter une taille maximale définie et ne doivent pas être autorisées à se développer au-delà.
   - La taille maximale doit être identifiée en effectuant des tests approfondis, jusqu'à ce que les points de rupture soient atteints et que des marges de fonctionnement sûres soient établies. Pour obtenir plus de détails sur la mise en œuvre des pratiques de test, consultez REL07-BP04 Effectuer un test de charge de votre charge de travail
   - La charge de travail globale doit se développer en ajoutant des cellules supplémentaires, ce qui lui permet de s'adapter à l'augmentation de la demande.

5. Stratégies multi-AZ ou multirégion : il convient de tirer parti de plusieurs couches de résilience pour se protéger contre différents domaines de panne.
   - Pour la résilience, vous devez adopter une approche qui repose sur des couches de défense. Une couche protège contre les perturbations de petite envergure et courantes en créant une architecture hautement disponible à l'aide de plusieurs AZ. Une autre couche de défense est destinée à protéger contre les événements rares tels que les catastrophes naturelles généralisées et les perturbations au niveau régional. Cette deuxième couche implique de concevoir l'architecture de votre application pour qu'elle s'étende sur plusieurs Régions AWS. La mise en œuvre d'une stratégie multirégion pour votre charge de travail permet de la protéger contre les catastrophes naturelles généralisées qui affectent une grande région géographique d'un pays, ou les défaillances techniques à l'échelle régionale. Sachez que la mise en œuvre d'une architecture multirégion peut être très complexe et n'est généralement pas requise pour la plupart des charges de travail. Pour en savoir plus, consultez REL10-BP02 Sélectionner les emplacements appropriés pour votre déploiement multisite. »
6. Déploiement du code : il faut privilégier une stratégie de déploiement de code échelonnée plutôt que de déployer les modifications de code dans toutes les cellules en même temps.
   - Cela permettra de minimiser les risques de panne de plusieurs cellules en raison d'un mauvais déploiement ou d'une erreur humaine. Pour obtenir plus de détails, consultez Automatisation de déploiements sécurisés sans intervention.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Ressources

Bonnes pratiques associées :

- REL07-BP04 Effectuer un test de charge de votre charge de travail
- REL10-BP02 Sélectionner les emplacements appropriés pour votre déploiement multisite

Documents connexes :

- Reliability, constant work, and a good cup of coffee (Fiabilité, travail constant, et une bonne tasse de café)
- AWS and Compartmentalization (AWS et le cloisonnement)
- Isolation de la charge de travail à l'aide du partitionnement aléatoire
- Automatisation de déploiements sécurisés sans intervention

Vidéos connexes :

- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small
- AWS re:Invent 2018: How AWS Minimizes the Blast Radius of Failures (ARC338)
- Shuffle-sharding: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)
- AWS Summit ANZ 2021 - Everything fails, all the time: Designing for resilience (Sommet AWS ANZ 2021 - tout échoue, tout le temps : concevoir pour la résilience)

Exemples connexes :

- Atelier Well-Architected : isolement des pannes avec le partitionnement aléatoire

FIA 11. Comment concevez-vous votre charge de travail pour la rendre résistante aux défaillances de composants ?

Les charges de travail exigeant une haute disponibilité et un faible temps moyen de récupération (MTTR) doivent être conçues pour être résilientes.

Bonnes pratiques

- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP02 Basculer vers des ressources saines
- REL11-BP03 Automatiser la réparation sur toutes les couches
- REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle pendant la récupération
- REL11-BP05 Utiliser la stabilité statique pour éviter les comportements bimodaux
- REL11-BP06 Envoyer des notifications lorsque des événements affectent la disponibilité
- REL11-BP07 Concevoir votre produit pour atteindre les objectifs de disponibilité et les accords de niveau de service (SLA)

REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances

Surveillez en continu l'état de votre charge de travail afin que vous et vos systèmes automatisés ayez connaissance des dégradations ou des défaillances dès qu'elles se produisent. Surveillez les indicateurs clés de performance (KPI) en fonction de la valeur commerciale.

Tous les mécanismes de récupération et de réparation doivent commencer par la capacité à détecter rapidement les problèmes. Les défaillances techniques doivent être détectées au préalable pour être résolues. Cependant, la disponibilité repose sur la capacité de votre charge de travail à fournir une valeur commerciale. Il doit donc s'agir d'indicateurs clés de performance (KPI) de votre stratégie de détection et de correction.

Résultat souhaité : Les composants essentiels d'une charge de travail sont surveillés de manière indépendante afin de détecter les défaillances et de les signaler au moment et à l'emplacement où elles se produisent.

Anti-modèles courants :

- Aucune alarme n'a été configurée. Il n'y a donc pas de notification lorsque des interruptions se produisent.
- Des alarmes existent, mais les seuils ne laissent pas assez de temps pour réagir.
- Les métriques ne sont pas collectées à une fréquence suffisante pour atteindre l'objectif de délai de reprise (RTO).
- Seules les interfaces de la charge de travail axées directement sur le client sont activement surveillées.
- Collecte uniquement des métriques techniques et non des métriques de fonction commerciale.
- Aucune métrique ne mesure l'expérience utilisateur de la charge de travail.
- Trop de contrôleurs sont créés.

Avantages liés au respect de cette bonne pratique : La surveillance appropriée à tous les niveaux vous permet de raccourcir le délai de reprise en réduisant le temps de détection.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Identifiez toutes les charges de travail qui seront examinées à des fins de surveillance. Une fois que vous avez identifié tous les composants de la charge de travail à surveiller, déterminez l'intervalle de surveillance. Cet intervalle a un impact direct sur la rapidité avec laquelle la restauration peut être initiée en fonction du temps nécessaire pour détecter une panne. Le délai moyen de détection (MTTD) est le délai entre le moment où une panne survient et le moment où les opérations de réparation commencent. La liste des services doit être longue et complète.

La surveillance doit couvrir toutes les couches de la pile d'applications, y compris l'application, la plate-forme, l'infrastructure et le réseau.

Votre stratégie de surveillance doit tenir compte de l'impact des défaillances grises. Pour plus de détails sur les défaillances grises, voir Défaillances grises dans le livre blanc sur les modèles de résilience multi-AZ avancés.

Étapes d'implémentation

- Votre intervalle de surveillance dépend de la vitesse à laquelle vous devez effectuer la récupération. Votre délai de reprise dépend du temps nécessaire à la récupération. Vous devez donc déterminer la fréquence de collecte en tenant compte de cette durée et de votre objectif de délai de reprise (RTO).
- Configurez la surveillance détaillée des composants et des services gérés.
  - Déterminez si la surveillance détaillée des instances EC2 et Auto Scaling est nécessaire. La surveillance détaillée fournit des métriques à intervalle d'une minute, et la surveillance par défaut fournit des métriques à intervalle de cinq minutes.
  - Déterminez si la surveillance améliorée pour RDS est nécessaire. La surveillance améliorée utilise un agent sur les instances RDS pour obtenir des informations utiles sur différents processus ou threads.
  - Déterminez les exigences de surveillance des composants sans serveur critiques pour Lambda, API Gateway, Amazon EKS, Amazon ECSet tous les types d'équilibreurs de charge.
  - Déterminez les exigences de surveillance des composants de stockage pour Amazon S3, Amazon FSx, Amazon EFSet Amazon EBS.
- Créez des métriques personnalisées pour mesurer les indicateurs clés de performance (KPI) de l'entreprise. Les charges de travail mettent en œuvre des fonctions commerciales stratégiques, qui doivent être utilisées comme indicateurs clés de performance permettant d'identifier les problèmes indirects.
- Surveillez l'expérience utilisateur pour détecter les défaillances à l'aide de tests canary utilisateur. Les tests de transactions synthétiques (également appelés « tests canary », à ne pas confondre avec les déploiements canary) qui peuvent exécuter et simuler le comportement des clients font partie des processus de test les plus importants. Exécutez ces tests en permanence sur vos points de terminaison de charge de travail à partir de divers emplacements distants.
- Créez des métriques personnalisées qui suivent l'expérience de l'utilisateur. Si vous pouvez analyser l'expérience du client, vous pouvez savoir à quel moment l'expérience du consommateur se dégrade.

- Définissez des alarmes pour détecter quand une partie de votre charge de travail ne fonctionne pas correctement et pour indiquer quand mettre automatiquement à l'échelle les ressources. Le système peut afficher les alarmes sur des tableaux de bord, envoyer des alertes via Amazon SNS ou par e-mail et fonctionner avec Auto Scaling pour une mise à l'échelle à la hausse ou à la baisse des ressources de la charge de travail.
- Créez des tableaux de bord pour visualiser vos métriques. Les tableaux de bord peuvent être utilisés pour afficher visuellement des tendances, des valeurs aberrantes et d'autres indicateurs de problèmes potentiels, ou pour fournir une indication des problèmes que vous pourriez vouloir examiner.
- Créez une surveillance distribuée pour vos services. Avec la surveillance distribuée, vous pouvez analyser les performances de votre application et de ses services sous-jacents, afin d'identifier et de dépanner la cause première des problèmes et des erreurs de performances.
- Créez les systèmes de surveillance (en utilisant CloudWatch ou X-Ray), les tableaux de bord et la collecte de données dans une région et un compte distincts.
- Créez une intégration pour la surveillance Amazon Health Aware pour permettre d'identifier les ressources AWS susceptibles de subir des dégradations. Pour les charges de travail essentielles à l'entreprise, cette solution permet d'accéder à des alertes proactives et en temps réel pour les services AWS.

Ressources

Bonnes pratiques associées :

- Définition de la disponibilité
- REL11-BP06 Envoi de notifications lorsque des événements affectent la disponibilité

Documents connexes :

- Amazon CloudWatch Synthetics vous permet de créer des tests canary utilisateur
- Activer ou désactiver la surveillance détaillée pour votre instance
- Surveillance améliorée
- Surveillance de vos groupes et instances Auto Scaling à l'aide d'Amazon CloudWatch
- Publication des métriques personnalisées
- Utilisation des alarmes Amazon CloudWatch

- Fonctionnement des tableaux de bord CloudWatch
- Utilisation de tableaux de bord CloudWatch interrégionaux entre comptes
- Utilisation du suivi X-Ray interrégional entre comptes
- Compréhension de la disponibilité
- Implémentation d'Amazon Health Aware (AHA)

Vidéos connexes :

- Mitigating gray failures

Exemples connexes :

- Atelier Well-Architected : niveau 300 : implémentation de la surveillance de l'état et gestion des dépendances pour améliorer la fiabilité
- Un atelier sur l'observabilité : explorer X-Ray

Outils associés :

- CloudWatch
- CloudWatch X-Ray

REL11-BP02 Basculer vers des ressources saines

En cas de défaillance des ressources, les ressources saines doivent continuer à répondre aux requêtes. Pour les altérations liées à l'emplacement (par exemple, zone de disponibilité ou Région AWS), vérifiez que des systèmes sont en place pour basculer vers des ressources saines dans des emplacements intacts.

Lorsque vous concevez un service, répartissez la charge entre les ressources, les zones de disponibilité ou les régions. Ainsi, la défaillance d'une ressource individuelle ou l'altération peut être atténuée en déplaçant le trafic vers les ressources saines restantes. Réfléchissez à la manière dont les services sont découverts et acheminés en cas de défaillance.

Concevez vos services en tenant compte de la restauration après panne. Chez AWS, nous concevons les services afin de minimiser le temps de restauration en cas de défaillance, ainsi que l'impact sur les données. Nos services utilisent principalement des magasins de données qui valident

les requêtes uniquement lorsque les données sont stockées durablement sur plusieurs réplicas au sein d'une région. Ils sont élaborés de manière à utiliser l'isolation basée sur les cellules et à faire appel à l'isolement des pannes fourni par des zones de disponibilité. Nous utilisons largement l'automatisation dans nos procédures opérationnelles. Nous optimisons également notre fonction de remplacement et redémarrage afin de récupérer rapidement en cas d'interruptions.

Les modèles et les conceptions qui permettent le basculement varient pour chaque service de plateforme AWS. De nombreux services natifs gérés par AWS sont dotés de plusieurs zones de disponibilité (comme Lambda ou API Gateway). D'autres services AWS (comme EC2 et EKS) nécessitent des conceptions répondant à des bonnes pratiques spécifiques pour prendre en charge le basculement des ressources ou le stockage des données entre les zones de disponibilité.

La surveillance doit être configurée pour vérifier que la ressource de basculement est saine, suivre la progression du basculement des ressources et surveiller le rétablissement des processus métier.

Résultat souhaité : Les systèmes sont capables d'utiliser automatiquement ou manuellement de nouvelles ressources pour se remettre d'une dégradation.

Anti-modèles courants :

- La planification des défaillances ne fait pas partie de la phase de planification et de conception.
- Le RTO et le RPO ne sont pas établis.
- Surveillance insuffisante pour détecter les ressources défaillantes.
- Isolement approprié des domaines de défaillance.
- Le basculement multirégional n'est pas pris en compte.
- La détection des défaillances est trop sensible ou trop agressive lors de la décision de basculer.
- Pas de test ni de validation de la conception du basculement.
- Automatisation de la réparation automatique sans notification indiquant que la réparation était nécessaire.
- Pas de période d'attente pour éviter un failback trop précoce.

Avantages liés au respect de cette bonne pratique : Vous pouvez créer des systèmes plus résilients qui préservent leur fiabilité en cas de défaillance en se dégradant progressivement et en se rétablissant rapidement.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les services AWS, comme Elastic Load Balancing et Amazon EC2 Auto Scalingcontribuent à répartir la charge entre les ressources et les zones de disponibilité. Par conséquent, la défaillance d'une ressource individuelle (telle qu'une instance EC2) ou l'altération d'une zone de disponibilité peut être atténuée en déplaçant le trafic vers les ressources saines restantes.

Pour les charges de travail multirégionales, les conceptions sont plus complexes. Par exemple, les réplicas en lecture entre régions vous permettent de déployer vos données sur plusieurs Régions AWS. Cependant, le basculement est toujours nécessaire pour faire passer le réplica en lecture au niveau principal, puis pour rediriger le trafic vers le nouveau point de terminaison. Amazon Route 53, Route 53 ARC, CloudFront et AWS Global Accelerator aident à acheminer le trafic entre les Régions AWS.

Les services AWS, comme Amazon S3, Lambda, API Gateway, Amazon SQS, Amazon SNS, Amazon SES, Amazon Pinpoint, Amazon ECR, AWS Certificate Manager, EventBridge ou Amazon DynamoDB, sont automatiquement déployés dans plusieurs zones de disponibilité par AWS. En cas de défaillance, ces services AWS acheminent automatiquement le trafic vers des emplacements sains. Les données sont stockées de manière redondante dans plusieurs zones de disponibilité et restent disponibles.

Pour Amazon RDS, Amazon Aurora, Amazon Redshift, Amazon EKS ou Amazon ECS, Multi-AZ est une option de configuration. AWS peut diriger le trafic vers l'instance saine si le basculement est initié. Cette action de basculement peut être prise par AWS ou conformément aux exigences du client.

Pour les instances Amazon EC2, Amazon Redshift, les tâches Amazon ECS ou les pods Amazon EKS, vous choisissez les zones de disponibilité dans lesquelles vous souhaitez effectuer le déploiement. Pour certaines conceptions, Elastic Load Balancing fournit la solution permettant de détecter les instances dans les zones défectueuses et d'acheminer le trafic vers les zones saines. Elastic Load Balancing peut également acheminer le trafic vers les composants de votre centre de données sur site.

Pour le basculement du trafic multirégional, le réacheminement peut tirer parti d'Amazon Route 53, ARC, AWS Global Accelerator, Route 53 Private DNS for VPCs ou CloudFront pour fournir un moyen de définir des domaines Internet et d'attribuer des politiques de routage, y compris des surveillances de l'état, afin d'acheminer le trafic vers des régions saines. AWS Global Accelerator fournit des adresses IP statiques qui agissent comme point d'entrée fixe vers votre application, puis acheminent

le trafic vers les points de terminaison des Régions AWS de votre choix, en utilisant le réseau mondial AWS plutôt qu'Internet pour de meilleures performances et une meilleure fiabilité.

Étapes d'implémentation

- Créez des modèles de basculement pour toutes les applications et tous les services appropriés. Isolez chaque composant de l'architecture et créez des conceptions de basculement respectant le RTO et le RPO pour chacun d'eux.
- Configurez les environnements de bas niveau (tels que le développement ou les tests) avec tous les services requis pour disposer d'un plan de basculement. Déployez les solutions en utilisant l'infrastructure en tant que code (IaC) pour garantir la reproductibilité.
- Configurez un site de reprise tel qu'une deuxième région pour implémenter et tester les modèles de basculement. Si nécessaire, les ressources pour les tests peuvent être configurées temporairement afin de limiter les coûts supplémentaires.
- Déterminez quels plans de basculement seront automatisés par AWS, ceux qui peuvent être automatisés par un processus DevOps et ceux qui peuvent être manuels. Documentez et mesurez le RTO et le RPO de chaque service.
- Créez un manuel de basculement et incluez toutes les étapes nécessaires au basculement de chaque ressource, application et service.
- Créez un manuel de failback et incluez toutes les étapes nécessaires (avec calendrier) pour chaque ressource, application et service
- Créez un plan pour lancer et répéter le manuel. Utilisez des simulations et des tests de chaos pour tester les étapes du manuel et l'automatisation.
- Pour toute altération liée à l'emplacement (par exemple, zone de disponibilité ou Région AWS), vérifiez que des systèmes sont en place pour basculer vers des ressources saines dans des emplacements intacts. Vérifiez le quota, les niveaux de mise à l'échelle automatique et les ressources en cours d'exécution avant le test de basculement.

Ressources

Bonnes pratiques Well-Architected connexes :

- REL13 – Plan de reprise après sinistre
- REL10 – Utilisation de l'isolation des défaillances pour protéger votre charge de travail

Documents connexes :

- Définition des objectifs RTO et RPO
- Configuration ARC avec des Application Load Balancers
- Basculement à l'aide du routage pondéré Route 53
- Reprise après sinistre avec ARC
- EC2 avec mise à l'échelle automatique
- Déploiements EC2 – Multi-AZ
- Déploiements ECS – Multi-AZ
- Changer de trafic avec ARC
- Lambda avec un Application Load Balancer et un basculement
- Réplication et basculement ACM
- Réplication et basculement du magasin de paramètres
- Réplication entre régions ECR et basculement
- Configuration de la réplication entre régions du gestionnaire de secrets
- Activation de la réplication entre régions pour EFS et basculement
- Réplication entre régions EFS et basculement
- Basculement du réseau
- Basculement des points de terminaison S3 à l'aide du protocole MRAP
- Création d'une réplication entre régions pour S3
- API Gateway régional de basculement avec ARC
- Basculement à l'aide d'un accélérateur mondial multirégional
- Basculement avec DRS
- Création de mécanismes de reprise après sinistre à l'aide d'Amazon Route 53

Exemples connexes :

- Reprise après sinistre sur AWS
- Elastic Disaster Recovery sur AWS

REL11-BP03 Automatiser la réparation sur toutes les couches

Utilisez des capacités automatisées pour effectuer des actions correctives en cas de détection d'une défaillance. Les dégradations peuvent être automatiquement corrigées par le biais de mécanismes

de service internes ou peuvent nécessiter le redémarrage ou la suppression des ressources par le biais d'actions correctives.

Pour les applications autogérées et la réparation interrégionale, les modèles de restauration et les processus de réparation automatisés peuvent être extraits des bonnes pratiques existantes.

Pouvoir redémarrer ou supprimer une ressource est important pour remédier aux défaillances. Une bonne pratique consiste à rendre les services sans état dans la mesure du possible. Cela évite toute perte de données ou de disponibilité au redémarrage des ressources. Dans le cloud, vous pouvez (et devriez généralement) remplacer la totalité de la ressource (par exemple, une instance de calcul ou une fonction sans serveur) dans le cadre du redémarrage. Le redémarrage proprement dit est un moyen simple et fiable de récupération après une défaillance. De nombreux types de défaillances différents se produisent dans les charges de travail. Les défaillances peuvent se produire au niveau du matériel, des logiciels, des communications et des opérations.

Le redémarrage ou la nouvelle tentative s'appliquent également aux requêtes réseau. Appliquez la même approche de récupération à la fois pour un délai d'expiration réseau et une défaillance de la dépendance, si la dépendance renvoie une erreur. Comme ces deux événements ont un effet semblable sur le système, plutôt que d'essayer de traiter l'un ou l'autre comme un « cas particulier », appliquez une stratégie semblable de nouvelle tentative limitée avec un backoff exponentiel et une instabilité. La possibilité d'exécuter un redémarrage est un mécanisme de récupération présenté dans l'informatique orientée récupération et dans les architectures de cluster haute disponibilité.

Résultat souhaité : Des actions automatisées sont effectuées pour remédier à la détection d'une panne.

Anti-modèles courants :

- Allocation des ressources sans mise à l'échelle automatique.
- Déploiement d'applications une par une dans des instances ou des conteneurs.
- Déploiement d'applications qui ne peuvent pas être déployées dans plusieurs emplacements sans utiliser la récupération automatique.
- Réparation manuelle des applications impossible à réparer par la mise à l'échelle et la récupération automatiques.
- Aucune automatisation pour le basculement des bases de données.
- Absence de méthodes automatisées pour rediriger le trafic vers de nouveaux points de terminaison.

- Aucune réplication du stockage.

Avantages liés au respect de cette bonne pratique : La réparation automatisée contribue à réduire le temps moyen de récupération et à améliorer votre disponibilité.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les conceptions pour Amazon EKS ou les autres services Kubernetes doivent inclure à la fois un minimum et un maximum de réplicas ou d'ensembles dynamiques, ainsi que le dimensionnement minimal des clusters et des groupes de nœuds. Ces mécanismes mettent à disposition un minimum de ressources de traitement en permanence tout en corrigeant automatiquement les défaillances à l'aide du plan de contrôle Kubernetes.

Les modèles de conception accessibles via un équilibreur de charge utilisant des clusters de calcul doivent tirer parti des groupes Auto Scaling. Elastic Load Balancing (ELB) distribue automatiquement le trafic applicatif entrant sur plusieurs cibles et appareils virtuels dans une ou plusieurs zones de disponibilité (AZ).

La taille des conceptions basées sur le calcul en cluster qui n'utilisent pas l'équilibrage de charge doit être conçue pour la perte d'au moins un nœud. Cela permet au service de continuer à fonctionner avec une capacité potentiellement réduite pendant la restauration d'un nouveau nœud. Exemples de services : Mongo, DynamoDB Accelerator, Amazon Redshift, Amazon EMR, Cassandra, Kafka, MSK-EC2, Couchbase, ELK et Amazon OpenSearch Service. Bon nombre de ces services peuvent être conçus avec des fonctionnalités supplémentaires de réparation automatique. Certaines technologies de cluster doivent générer une alerte en cas de perte d'un nœud, déclenchant un flux de travail automatique ou manuel pour recréer un nœud. Ce flux de travail peut être automatisé avec AWS Systems Manager pour résoudre rapidement les problèmes.

Amazon EventBridge peut être utilisé pour surveiller et filtrer les événements tels que les alarmes CloudWatch ou les changements d'état dans d'autres services AWS. Sur la base des informations relatives aux événements, il peut ensuite appeler AWS Lambda, Systems Manager Automation ou d'autres cibles pour exécuter une logique de correction personnalisée sur votre charge de travail. Amazon EC2 Auto Scaling peut être configuré pour vérifier l'état de l'instance EC2. Si l'instance est dans un état autre que celui en cours d'exécution, ou si le statut du système est dégradé, Amazon EC2 Auto Scaling considère l'instance comme défectueuse et lance une instance de remplacement. Pour les remplacements à grande échelle (comme la perte d'une zone de disponibilité complète), il est préférable d'opter pour la stabilité statique pour une haute disponibilité.

Étapes d'implémentation

- Utilisez des groupes Auto Scaling pour déployer des niveaux dans une charge de travail. Auto Scaling peut effectuer une autoréparation sur les applications sans état et ajouter ou supprimer de la capacité.
- Pour les instances de calcul mentionnées précédemment, utilisez l'équilibrage de charge et choisissez le type d'équilibreur de charge approprié.
- Envisagez la réparation pour Amazon RDS. Avec les instances de secours, configurez le basculement automatique vers l'instance de secours. Pour les réplicas en lecture Amazon RDS, un flux de travail automatisé est nécessaire pour rendre un réplica en lecture principal.
- Implémentez la restauration automatique sur les instances EC2 dont les applications déployées ne peuvent pas être déployées sur plusieurs sites et qui peuvent tolérer un redémarrage en cas de panne. La récupération automatique peut être utilisée pour remplacer du matériel défaillant et redémarrer l'instance lorsque l'application ne peut pas être déployée sur plusieurs emplacements. Les métadonnées de l'instance et les adresses IP associées sont conservées, ainsi que les volumes EBS et les points de montage vers Amazon Elastic File System ou le systèmes de fichiers pour Lustre et Windows. Avec AWS OpsWorks, vous pouvez configurer la réparation automatique des instances EC2 au niveau de la couche.
- Mettez en œuvre une restauration automatique avec AWS Step Functions et AWS Lambda lorsque vous ne pouvez pas utiliser la mise à l'échelle automatique ou la récupération automatique, ou lorsque la récupération automatique échoue. Lorsque vous ne pouvez pas utiliser la scalabilité automatique, que vous ne pouvez pas utiliser la récupération automatique ou que la récupération automatique échoue, vous pouvez automatiser la réparation à l'aide d'AWS Step Functions et d'AWS Lambda.
- Amazon EventBridge peut être utilisé pour surveiller et filtrer des événements tels que des alarmes CloudWatch ou des changements d'état dans d'autres services AWS. En fonction des informations d'événement, il peut ensuite déclencher AWS Lambda (ou d'autres cibles) pour exécuter une logique de correction personnalisée sur votre charge de travail.

Ressources

Bonnes pratiques associées :

- Définition de la disponibilité
- REL11-BP01 Surveillance de tous les composants de la charge de travail pour détecter les défaillances

Documents connexes :

- Fonctionnement d'AWS Auto Scaling
- Récupération automatique Amazon EC2
- Amazon Elastic Block Store (Amazon EBS)
- Amazon Elastic File System (Amazon EFS)
- Qu'est-ce que Amazon FSx for Lustre ?
- Qu'est-ce que Amazon FSx for Windows File Server ?
- AWS OpsWorks : utilisation de la réparation automatique pour remplacer des instances défectueuses
- Qu'est-ce qu'AWS Step Functions ?
- Qu'est-ce qu'AWS Lambda ?
- Qu'est-ce qu'Amazon EventBridge ?
- Utilisation des alarmes Amazon CloudWatch
- Basculement Amazon RDS
- SSM – Systems Manager Automation
- Bonnes pratiques en matière d'architecture résiliente

Vidéos connexes :

- Automatically Provision and Scale OpenSearch Service
- Amazon RDS Failover Automatically

Exemples connexes :

- Atelier sur Auto Scaling
- Atelier sur le basculement Amazon RDS

Outils associés :

- CloudWatch
- CloudWatch X-Ray

REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle pendant la récupération

Les plans de contrôle fournissent les API administratives utilisées pour créer, lire et décrire, mettre à jour, supprimer et répertorier (CRUDL) les ressources, tandis que les plans de données gèrent le trafic quotidien des services. Lorsque vous mettez en œuvre des réponses de restauration ou d'atténuation en cas d'événements susceptibles d'avoir un impact sur la résilience, concentrez-vous sur l'utilisation d'un nombre minimal d'opérations du plan de contrôle pour récupérer, redimensionner, restaurer, réparer ou basculer le service. L'action du plan de données doit remplacer toute activité lors de ces événements de dégradation.

Par exemple, les actions suivantes font toutes partie du plan de contrôle : lancement d'une nouvelle instance de calcul, création d'un stockage par blocs et description des services de file d'attente. Lorsque vous lancez des instances de calcul, le plan de contrôle doit effectuer plusieurs tâches, telles que la recherche d'un hôte physique avec la capacité suffisante, l'allocation d'interfaces réseau, la préparation de volumes locaux de stockage par blocs, la génération d'informations d'identification et l'ajout de règles de sécurité. Les plans de contrôle relèvent souvent d'une orchestration complexe.

Résultat souhaité : Lorsqu'une ressource passe à un état altéré, le système peut être rétabli automatiquement ou manuellement en transférant le trafic des ressources altérées vers des ressources saines.

Anti-modèles courants :

- Nécessité de modifier les enregistrements DNS pour rediriger le trafic.
- Nécessité de réaliser des opérations de dimensionnement du plan de contrôle pour remplacer les composants endommagés en raison de ressources sous-provisionnées.
- Utilisation d'actions de plan de contrôle étendues, multiservices et multi-API pour remédier à toute catégorie d'altération.

Avantages liés au respect de cette bonne pratique : L'augmentation du taux de réussite de la correction automatisée contribue à réduire le temps moyen de récupération et à améliorer la disponibilité de la charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen : pour certains types de dégradations de service, les plans de contrôle sont concernés. La nécessité d'utiliser de manière intensive le plan de contrôle pour la correction peut augmenter le délai de reprise (RTO) et le temps moyen de récupération (MTTR).

Directives d'implémentation

Pour limiter les actions du plan de données, évaluez chaque service pour déterminer les actions nécessaires afin de restaurer le service.

Tirez parti d'Amazon Application Recovery Controller (ARC) pour déplacer le trafic DNS. Ces fonctionnalités surveillent en permanence la capacité de votre application à se rétablir suite à des défaillances et vous permettent de contrôler la reprise de votre application dans plusieurs Régions AWS, plusieurs zones de disponibilité et sur site.

Les politiques de routage Route 53 utilisent le plan de contrôle. Ne vous fiez donc pas à celui-ci pour la récupération. Les plans de données Route 53 répondent aux requêtes DNS et effectuent et évaluent des surveillances de l'état. Ils sont distribués dans le monde entier et conçus pour un contrat de niveau de service (SLA) de 100 % de disponibilité..

Les API et consoles de gestion Route 53 dans lesquelles vous créez, mettez à jour et supprimez des ressources Route 53 s'exécutent sur des plans de contrôle conçus pour donner la priorité à la cohérence forte et à la durabilité dont vous avez besoin lors de la gestion du DNS. Pour ce faire, les plans de contrôle sont situés dans une seule région : USA Est (Virginie du Nord). Bien que les deux systèmes soient conçus pour être très fiables, les plans de contrôle ne sont pas inclus dans le SLA. Dans de rares cas, la conception résiliente du plan de données permet de maintenir la disponibilité alors que les plans de contrôle ne le font pas. Pour les mécanismes de reprise après sinistre et de basculement, utilisez les fonctions du plan de données pour assurer la meilleure fiabilité possible.

Pour Amazon EC2, utilisez des modèles de stabilité statique pour limiter les actions du plan de contrôle. Les actions du plan de contrôle incluent l'augmentation des ressources individuellement ou à l'aide de groupes Auto Scaling (ASG). Pour obtenir les niveaux de résilience les plus élevés, allouez une capacité suffisante dans le cluster utilisé pour le basculement. Si ce seuil de capacité doit être limité, définissez des limitations sur l'ensemble du système de bout en bout afin de restreindre en toute sécurité le trafic total atteignant l'ensemble limité de ressources.

Pour des services comme Amazon DynamoDB, Amazon API Gateway, les équilibreurs de charge et AWS Lambda sans serveur, leur utilisation permet de tirer parti du plan de données. Cependant, la création de fonctions, d'équilibreurs de charge, de passerelles d'API ou de tables DynamoDB est une action du plan de contrôle qui doit être terminée avant la dégradation afin de préparer un événement et de répéter les actions de basculement. Pour Amazon RDS, les actions du plan de données permettent d'accéder aux données.

Pour plus d'informations sur les plans de données, les plans de contrôle et la manière dont AWS crée des services pour atteindre les objectifs de haute disponibilité, consultez le livre blanc sur la stabilité statique avec les zones de disponibilité.

Comprendre quelles opérations relèvent du plan de données et quelles opérations relèvent du plan de contrôle

Étapes d'implémentation

Pour chaque charge de travail qui doit être restaurée après un événement de dégradation, évaluez le runbook de basculement, la conception de la haute disponibilité, la conception de la réparation automatique ou le plan de restauration des ressources haute disponibilité. Identifiez chaque action qui pourrait être considérée comme une action du plan de contrôle.

Envisagez de remplacer l'action du plan de contrôle par une action de plan de données :

- Auto Scaling (plan de contrôle) par rapport aux ressources Amazon EC2 pré-dimensionnées (plan de données)
- Migration vers Lambda et ses méthodes de mise à l'échelle (plan de données) ou Amazon EC2 et ASG (plan de contrôle)
- Évaluez toutes les conceptions utilisant Kubernetes, ainsi que la nature des actions du plan de contrôle. L'ajout de pods est une action du plan de données dans Kubernetes. Les actions doivent se limiter à l'ajout de pods et non à l'ajout de nœuds. L'utilisation de nœuds surprovisionnés est la méthode préférée pour limiter les actions du plan de contrôle.

Envisagez d'autres approches qui permettent aux actions du plan de données d'affecter les mêmes mesures correctives.

- Modification d'enregistrement Route 53 (plan de contrôle) ou ARC (plan de données)
- Surveillance de l'état Route 53 pour des mises à jour plus automatisées

Envisagez certains services dans une région secondaire, s'ils sont critiques, afin de permettre davantage d'actions du plan de contrôle et du plan de données dans une région non affectée.

- Amazon EC2 Auto Scaling ou Amazon EKS dans une région principale par rapport à Amazon EC2 Auto Scaling ou Amazon EKS dans une région secondaire et acheminement du trafic vers la région secondaire (action du plan de contrôle)

- Réalisez un réplica en lecture dans la région secondaire ou tentez la même action dans la région principale (action du plan de contrôle).

Ressources

Bonnes pratiques associées :

- Définition de la disponibilité
- REL11-BP01 Surveillance de tous les composants de la charge de travail pour détecter les défaillances

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à automatiser votre tolérance aux pannes
- AWS Marketplace : produits pouvant être utilisés pour la tolérance aux pannes
- L'Amazon Builders' Library : éviter la surcharge des systèmes distribués en plaçant sous contrôle le plus petit service
- API Amazon DynamoDB (plan de contrôle et plan de données)
- Exécutions AWS Lambda (réparties entre le plan de contrôle et le plan de données)
- Plan de données AWS Elemental MediaStore
- Création d'applications hautement résilientes à l'aide d'Amazon Route 53 Application Recovery Controller, partie 1 : pile dans une seule région
- Création d'applications hautement résilientes à l'aide d'Amazon Route 53 Application Recovery Controller, partie 2 : pile multirégion
- Création de mécanismes de reprise après sinistre à l'aide d'Amazon Route 53
- Qu'est-ce que Route 53 Application Recovery Controller ?
- Plan de contrôle et plan de données Kubernetes

Vidéos connexes :

- Back to Basics - Using Static Stability
- Building resilient multi-site workloads using AWS global services

Exemples connexes :

- Qu'est-ce qu'Amazon Route 53 Application Recovery Controller ?
- L'Amazon Builders' Library : éviter la surcharge des systèmes distribués en plaçant sous contrôle le plus petit service
- Création d'applications hautement résilientes à l'aide d'Amazon Route 53 Application Recovery Controller, partie 1 : pile dans une seule région
- Création d'applications hautement résilientes à l'aide d'Amazon Route 53 Application Recovery Controller, partie 2 : pile multirégion
- stabilité statique avec les zones de disponibilité

Outils associés :

- Amazon CloudWatch
- AWS X-Ray

REL11-BP05 Utiliser la stabilité statique pour éviter les comportements bimodaux

Les charges de travail doivent être statiquement stables et ne fonctionner que dans un seul mode normal. On parle de comportement bimodal lorsque la charge de travail présente un comportement différent en mode normal et en mode d'échec.

Par exemple, vous pouvez essayer de récupérer une défaillance de la zone de disponibilité en lançant de nouvelles instances dans une zone de disponibilité différente. Il peut en résulter une réponse bimodale lors d'un mode de défaillance. Pour éviter ce type de comportement, vous devez créer des charges de travail stables statiquement et qui fonctionnent dans un seul mode. Dans cet exemple, ces instances auraient dû être provisionnées dans la deuxième zone de disponibilité avant la panne. Ce modèle de stabilité statique permet de vérifier que la charge de travail ne fonctionne que dans un seul mode.

Résultat souhaité : Les charges de travail ne présentent pas de comportement bimodal en mode normal et en mode d'échec.

Anti-modèles courants :

- Supposer que les ressources peuvent toujours être provisionnées quelle que soit l'étendue de la défaillance.
- Essayer d'acquérir dynamiquement des ressources lors d'une panne.

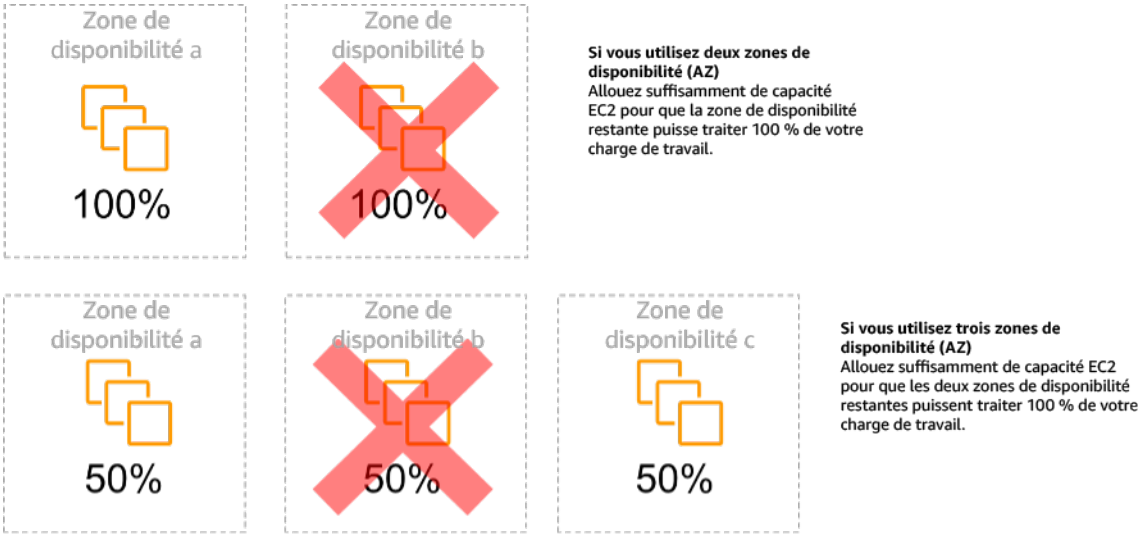
- Ne pas provisionner les ressources adéquates dans les zones ou les régions jusqu'à ce qu'une panne se produise.
- Envisager des modèles statiques et stables pour les ressources informatiques uniquement.

Avantages liés au respect de cette bonne pratique : Les charges de travail exécutées avec des modèles statiquement stables sont capables d'avoir des résultats prévisibles lors d'événements normaux et de défaillances.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Un comportement bimodal survient lorsque votre charge de travail adopte un comportement différent en mode normal et en mode de défaillance (par exemple, en s'appuyant sur le lancement de nouvelles instances en cas de défaillance d'une zone de disponibilité). Exemple de comportement bimodal : les modèles Amazon EC2 stables provisionnent suffisamment d'instances dans chaque zone de disponibilité pour gérer la charge de travail si une zone de disponibilité était supprimée. L'état de Elastic Load Balancing ou de Amazon Route 53 effectuerait une vérification pour déplacer une charge des instances déficientes. Une fois le trafic déplacé, utilisez AWS Auto Scaling pour remplacer de manière asynchrone les instances de la zone défaillante et les lancer dans les zones saines. La stabilité statique du déploiement de calcul (par exemple, des conteneurs ou des instances EC2) garantit une fiabilité optimale.



Stabilité statique des instances EC2 dans les zones de disponibilité

Cela doit être comparé au coût de ce modèle et à la valeur commerciale du maintien de la charge de travail dans tous les cas de résilience. Il est moins coûteux de provisionner moins de capacité de

calcul et de compter sur le lancement de nouvelles instances en cas de panne. Cependant, pour les pannes à grande échelle (comme une zone de disponibilité ou une panne régionale), cette approche se révèle moins efficace, car elle repose à la fois sur un plan opérationnel et sur la disponibilité de ressources suffisantes dans les zones ou les régions non affectées.

Votre solution doit également tenir compte de la fiabilité par rapport aux coûts nécessaires pour votre charge de travail. Les architectures de stabilité statique s'appliquent à différentes architectures, notamment aux instances de calcul réparties dans les zones de disponibilité, les modèles de réplicas en lecture de bases de données, les modèles de clusters Kubernetes (Amazon EKS) et les architectures de basculement multi-régions.

Il est également possible de mettre en œuvre un modèle plus stable sur le plan statique en utilisant davantage de ressources dans chaque zone. En ajoutant davantage de zones, vous réduisez la quantité de calcul supplémentaire nécessaire à la stabilité statique.

Autre exemple de comportement bimodal : un délai d'expiration du réseau peut amener un système à tenter d'actualiser l'état de configuration de l'ensemble du système. Cela ajouterait une charge inattendue à un autre composant et pourrait provoquer sa défaillance, entraînant d'autres conséquences inattendues. Cette boucle de rétroaction négative a un impact sur la disponibilité de votre charge de travail. Vous pourriez donc créer des systèmes stables statiquement et fonctionnant dans un seul mode. Un modèle statiquement stable consisterait à effectuer un travail constant et à toujours actualiser l'état de la configuration selon une cadence fixe. Lorsqu'un appel échoue, la charge de travail utilise la valeur précédemment mise en cache et déclenche une alarme.

Un autre exemple de comportement bimodal consiste à autoriser les clients à contourner votre cache de charge de travail lorsque des défaillances se produisent. Cette solution peut sembler répondre aux besoins des clients, mais elle peut modifier considérablement les exigences de votre charge de travail et risque d'entraîner des échecs.

Évaluez les charges de travail critiques afin de déterminer celles qui nécessitent ce type de modèle de résilience. Pour celles qui sont jugées critiques, chaque composant de l'application doit être examiné. Voici quelques exemples de services nécessitant une évaluation de la stabilité statique :

- Calcul: Amazon EC2, EKS-EC2, ECS-EC2, EMR-EC2
- Bases de données: Amazon Redshift, Amazon RDS, Amazon Aurora
- Stockage: Amazon S3 (zone unique), Amazon EFS (supports), Amazon FSx (supports)
- Équilibreurs de charge : Selon certains modèles

Étapes d'implémentation

- Créez des systèmes stables statiquement et qui fonctionnent dans un seul mode. Dans ce cas, provisionnez suffisamment d'instances dans chaque zone de disponibilité ou région pour gérer la capacité de la charge de travail si une zone de disponibilité ou une région était supprimée. Plusieurs services peuvent être utilisés pour l'acheminement vers des ressources saines, par exemple :
  - Routage DNS entre régions
  - Routage multi-régions MRAP Amazon S3
  - AWS Global Accelerator
  - Amazon Application Recovery Controller (ARC)
- Configurez les réplicas en lecture de la base de données pour tenir compte de la perte d'une instance primaire unique ou d'un réplica en lecture. Si le trafic est desservi par des réplicas en lecture, la quantité dans chaque zone de disponibilité et chaque région doit correspondre au besoin global en cas de défaillance de la zone ou de la région.
- Configurez les données critiques dans un stockage Amazon S3 conçu pour être statiquement stable pour les données stockées en cas de défaillance d'une zone de disponibilité. Si la classe de stockage Amazon S3 One Zone-IA est utilisée, elle ne doit pas être considérée comme statiquement stable, car la perte de cette zone minimise l'accès aux données stockées.
- Les équilibreurs de charge sont parfois configurés de manière incorrecte ou sciemment pour desservir une zone de disponibilité spécifique. Dans ce cas, le modèle statiquement stable peut consister à répartir une charge de travail sur plusieurs zones de disponibilité dans le cadre d'un modèle plus complexe. Le modèle original peut être utilisé pour réduire le trafic interzone pour des raisons de sécurité, de latence ou de coût.

Ressources

Bonnes pratiques Well-Architected connexes :

- Définition de la disponibilité
- REL11-BP01 Surveillance de tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle pendant la récupération

Documents connexes :

- Minimiser les dépendances dans un plan de reprise après sinistre
- L'Amazon Builders' Library : stabilité statique avec les zones de disponibilité
- Fault Isolation Boundaries
- stabilité statique avec les zones de disponibilité
- RDS multi-zones
- Minimiser les dépendances dans un plan de reprise après sinistre
- Routage DNS entre régions
- Routage multi-régions MRAP Amazon S3
- AWS Global Accelerator
- ARC
- Zone unique Amazon S3
- Équilibrage de charge entre zones

Vidéos connexes :

- Stabilité statique dans AWS : AWS re:Invent 2019 : présentation de la bibliothèque Amazon Builders' Library (DOP328)

Exemples connexes :

- L'Amazon Builders' Library : stabilité statique avec les zones de disponibilité

REL11-BP06 Envoyer des notifications lorsque des événements affectent la disponibilité

Des notifications sont envoyées en cas de détection de dépassement de seuils, même si l'événement à l'origine du problème a été automatiquement résolu.

La réparation automatisée permet à votre charge de travail d'être fiable. Cependant, elle peut également masquer les problèmes sous-jacents à résoudre. Implémentez une surveillance et des événements appropriés afin de pouvoir détecter les schémas de problèmes, y compris ceux résolus par la réparation automatique, afin de pouvoir résoudre les problèmes de cause racine.

Les systèmes résilients sont conçus de manière à ce que les événements de dégradation soient immédiatement communiqués aux équipes concernées. Ces notifications doivent être envoyées par un ou plusieurs canaux de communication.

Résultat souhaité : Des alertes sont immédiatement envoyées aux équipes chargées des opérations lorsque des seuils sont dépassés, tels que les taux d'erreur, la latence ou d'autres métriques d'indicateurs clés de performance (KPI) critiques, afin que ces problèmes soient résolus dès que possible et que l'impact sur les utilisateurs soit évité ou minimisé.

Anti-modèles courants :

- Envoyer un trop grand nombre d'alarmes.
- Envoyer des alarmes non exploitables.
- Régler les seuils d'alarme à un niveau trop élevé (sensibilité excessive) ou trop faible (sensibilité insuffisante).
- Ne pas envoyer d'alarmes pour les dépendances externes.
- Ne pas tenir compte des défaillances grises lors de la conception de la surveillance et des alarmes.
- Effectuer des réparations automatisées, mais ne pas notifier l'équipe appropriée que des réparations étaient nécessaires.

Avantages liés au respect de cette bonne pratique : Les notifications de reprise permettent aux équipes commerciales et chargées des opérations d'être informées des dégradations de service, de sorte qu'elles peuvent réagir immédiatement pour minimiser à la fois le temps moyen de détection (MTTD) et le temps moyen de réparation (MTTR). Les notifications d'événements de reprise vous permettent également de ne pas ignorer les problèmes qui se produisent peu fréquemment.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen. L'absence de mise en œuvre de mécanismes appropriés de surveillance et de notification des événements peut entraîner l'incapacité à détecter des schémas de problèmes, y compris ceux traités par la réparation automatisée. Une équipe ne sera informée de la dégradation du système que lorsque les utilisateurs contacteront le service clientèle ou par hasard.

Directives d'implémentation

Lors de la définition d'une stratégie de surveillance, le déclenchement d'une alarme est un événement courant. Cet événement contiendra probablement un identifiant pour l'alarme, l'état de l'alarme (comme `IN ALARM` ou `OK`) et des détails sur ce qui l'a déclenchée. Dans de nombreux cas,

un événement d'alarme doit être détecté et une notification par e-mail doit être envoyée. Voici un exemple d'action sur une alarme. La notification d'alarme est essentielle pour l'observabilité, car elle permet d'informer les bonnes personnes de l'existence d'un problème. Cependant, lorsque l'action sur les événements arrive à maturité dans votre solution d'observabilité, elle peut automatiquement remédier au problème sans nécessiter d'intervention humaine.

Une fois que les alarmes de suivi des KPI ont été établies, des alertes doivent être envoyées aux équipes concernées lorsque les seuils sont dépassés. Ces alertes peuvent également être utilisées pour déclencher des processus automatisés qui tenteront de remédier à la dégradation.

Pour une surveillance plus complexe des seuils, des alarmes composites doivent être envisagées. Les alarmes composites utilisent un certain nombre d'alarmes de surveillance des KPI pour créer une alerte basée sur la logique métier opérationnelle. Les alarmes CloudWatch peuvent être configurées pour envoyer des e-mails afin de consigner des incidents dans des systèmes tiers de suivi des incidents à l'aide de l'intégration d'Amazon SNS ou de Amazon EventBridge.

Étapes d'implémentation

Créez différents types d'alarmes en fonction des charges de travail surveillées, par exemple :

- Les alarmes d'application permettent de détecter si une partie de votre charge de travail ne fonctionne pas correctement.
- Les alarmes d'infrastructure indiquent à quel moment il convient de mettre les ressources à l'échelle. Le système peut afficher les alarmes sur des tableaux de bord, envoyer des alertes via Amazon SNS ou par e-mail et fonctionner avec Auto Scaling pour une mise à l'échelle verticale ou horizontale des ressources de la charge de travail.
- Simple Des alarmes statiques simples peuvent être créées pour surveiller le dépassement d'un seuil statique par une métrique pendant un nombre spécifié de périodes d'évaluation.
- Les alarmes composites peuvent prendre en compte des alarmes complexes provenant de sources multiples.
- Une fois l'alarme créée, créez les événements de notification appropriés. Vous pouvez directement appeler une API Amazon SNS pour envoyer des notifications et lier toute automatisation pour la remédiation ou la communication.
- Intégrez Amazon Health Aware pour permettre d'identifier les ressources AWS susceptibles de subir des dégradations. Pour les charges de travail essentielles à l'entreprise, cette solution permet d'accéder à des alertes proactives et en temps réel pour les services AWS.

Ressources

Bonnes pratiques Well-Architected connexes :

- Définition de la disponibilité

Documents connexes :

- Création d’une alarme CloudWatch basée sur un seuil statique
- Qu'est-ce qu'Amazon EventBridge ?
- Qu'est-ce qu'Amazon Simple Notification Service ?
- Publication des métriques personnalisées
- Utilisation des alarmes Amazon CloudWatch
- Amazon Health Aware (AHA)
- Configuration d'alarmes composites CloudWatch
- Les nouveautés en matière d'AWSobservabilité à re:Invent 2022

Outils associés :

- CloudWatch
- CloudWatch X-Ray

REL11-BP07 Concevoir votre produit pour atteindre les objectifs de disponibilité et les accords de niveau de service (SLA)

Concevez votre produit de manière à atteindre les objectifs de disponibilité et les accords de niveau de service (SLA). Si vous publiez ou convenez en privé d'objectifs de disponibilité ou d'accords de niveau de service, vérifiez que votre architecture et vos processus opérationnels sont conçus pour les prendre en charge.

Résultat souhaité : chaque application a un objectif défini pour la disponibilité et un accord de niveau de service pour les métriques de performance. Ces éléments peuvent être surveillés et maintenus afin d'atteindre les résultats opérationnels.

Anti-modèles courants :

- Concevoir et déployer des charges de travail sans fixer d'accords de niveau de service.

- Les métriques des SLA sont fixées à un niveau élevé sans justification ni exigences commerciales.
- Fixer des accords de niveau de service sans tenir compte des dépendances et des accords de niveau de service sous-jacents.
- Les conceptions d'applications sont créées sans tenir compte du modèle de responsabilité partagée pour la résilience.

Avantages liés au respect de cette bonne pratique : la conception d'applications basées sur des objectifs clés de résilience vous aide à atteindre les objectifs commerciaux et à répondre aux attentes des clients. Ces objectifs contribuent à orienter le processus de conception de l'application qui évalue les différentes technologies et envisage divers compromis.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

La conception des applications doit tenir compte d'un ensemble diversifié d'exigences découlant d'objectifs commerciaux, opérationnels et financiers. Dans le cadre des exigences opérationnelles, les charges de travail doivent avoir des objectifs spécifiques en matière de métriques de résilience afin qu'elles puissent être correctement surveillées et prises en charge. Les métriques de résilience ne doivent pas être définies ou déduites après le déploiement de la charge de travail. Elles doivent être définies pendant la phase de conception et aider à guider les diverses décisions et compromis.

- Chaque charge de travail doit disposer de son propre ensemble de métriques de résilience. Ces métriques peuvent être différentes de celles d'autres applications commerciales.
- La réduction des dépendances peut avoir un impact positif sur la disponibilité. Chaque charge de travail doit tenir compte de ses dépendances et de leurs accords de niveau de service. En général, sélectionnez les dépendances dont les objectifs de disponibilité sont égaux ou supérieurs à ceux de votre charge de travail.
- Envisagez des conceptions faiblement couplées afin que votre charge de travail puisse fonctionner correctement malgré l'altération des dépendances, lorsque cela est possible.
- Réduisez les dépendances du plan de contrôle, notamment lors de la reprise ou d'une dégradation. Évaluez les conceptions statiques stables pour les charges de travail critiques. Utilisez le partage des ressources pour augmenter la disponibilité de ces dépendances dans une charge de travail.
- L'observabilité et l'instrumentation sont essentielles pour respecter les accords de niveau de service en réduisant le temps moyen de détection (MTTD) et le temps moyen de réparation (MTTR).

- Des défaillances moins fréquentes (MTBF plus long), des temps de détection des défaillances plus courts (MTTD plus court) et des temps de réparation plus courts (MTTR plus court) sont les trois facteurs utilisés pour améliorer la disponibilité des systèmes distribués.
- L'établissement et le respect des métriques de résilience pour une charge de travail sont à la base de toute conception efficace. Ces conceptions doivent tenir compte des compromis entre la complexité de la conception, les dépendances des services, les performances, la mise à l'échelle et les coûts.

Étapes d'implémentation

- Examinez et documentez la conception de la charge de travail en tenant compte des questions suivantes :
  - Où les plans de contrôle sont-ils utilisés dans la charge de travail ?
  - Comment la charge de travail met-elle en œuvre la tolérance aux pannes ?
  - Quels sont les modèles de conception pour la mise à l'échelle, la scalabilité automatique, la redondance et les composants hautement disponibles ?
  - Quelles sont les exigences en matière de cohérence et de disponibilité des données ?
  - Y a-t-il des considérations relatives à l'économie des ressources ou à la stabilité statique des ressources ?
  - Quelles sont les dépendances des services ?
- Définissez les métriques SLA en fonction de l'architecture de la charge de travail tout en travaillant avec les parties prenantes. Tenez compte des SLA de toutes les dépendances utilisées par la charge de travail.
- Une fois l'objectif du SLA fixé, optimisez l'architecture pour le respecter.
- Une fois que la conception a été définie de manière à respecter l'accord de niveau de service, il faut mettre en œuvre les changements opérationnels, l'automatisation des processus et les runbooks qui visent également à réduire les délais d'attente et les temps de réponse.
- Une fois le déploiement effectué, surveillez et rendez compte de l'accord de niveau de service.

Ressources

Bonnes pratiques associées :

- REL03-BP01 Choisir comment segmenter votre charge de travail
- REL10-BP01 Déployer la charge de travail sur plusieurs emplacements

- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la réparation sur toutes les couches
- REL12-BP05 Tester la résilience à l'aide de l'ingénierie du chaos
- REL13-BP01 Définir les objectifs de reprise pour les temps d'arrêt et les pertes de données
- Comprendre l'état de la charge de travail

Documents connexes :

- Availability with redundancy (Disponibilité avec redondance)
- Pilier Fiabilité : disponibilité
- Measuring availability (Mesurer la disponibilité)
- AWS Fault Isolation Boundaries (Limites d'isolement des pannes AWS)
- Modèle de responsabilité partagée pour la résilience
- stabilité statique avec les zones de disponibilité
- Accords de niveau de service (SLA) AWS
- Guidance for Cell-based Architecture on AWS (Guide de l'architecture cellulaire sur AWS)
- Infrastructure AWS
- Advanced Multi-AZ Resiliance Patterns whitepaper (Livre blanc sur les modèles de résilience multi-AZ avancés)

Services associés :

- Amazon CloudWatch
- AWS Config
- AWS Trusted Advisor

FIA 12. Comment tester la fiabilité ?

Une fois que vous avez conçu votre charge de travail pour qu'elle soit résiliente aux sollicitations de la production, les tests sont le seul moyen de s'assurer qu'elle fonctionne comme prévu et d'obtenir la résilience voulue.

Bonnes pratiques

- REL12-BP01 Utiliser des playbooks pour enquêter sur les causes des défaillances

- REL12-BP02 Effectuer une analyse post-incident
- REL12-BP03 Tester les exigences fonctionnelles
- REL12-BP04 Tester les exigences de mise à l'échelle et de performance
- REL12-BP05 Tester la résilience à l'aide de l'ingénierie du chaos
- REL12-BP06 Organiser régulièrement des tests de simulation de panne

REL12-BP01 Utiliser des playbooks pour enquêter sur les causes des défaillances

Consignez le processus d'enquête dans des playbooks afin de faciliter l'application de réponses cohérentes et rapides face aux scénarios de défaillance qui ne sont pas bien compris. Les playbooks sont les étapes prédéfinies suivies pour identifier les facteurs adjuvants à un scénario défaillance. Les résultats des étapes du processus sont utilisés pour déterminer les prochaines mesures à prendre jusqu'à ce que la question soit identifiée ou remontée.

Le playbook est une planification proactive que vous devez appliquer afin de pouvoir prendre efficacement des mesures réactives. Lorsque des scénarios de défaillance ne figurant pas dans le playbook sont rencontrés en production, commencez par résoudre le problème (éteindre l'incendie). Procédez ensuite à une rétrospective en examinant les étapes suivies pour résoudre le problème et utilisez-les pour ajouter une nouvelle entrée dans le playbook.

Notez que les playbooks sont utilisés en réponse à des incidents spécifiques, tandis que les runbooks le sont pour obtenir des résultats spécifiques. En règle générale, les runbooks sont employés pour les activités de routine et les playbooks pour répondre à des événements non réguliers.

Anti-modèles courants :

- Planification du déploiement d'une charge de travail sans connaître les processus permettant de diagnostiquer les problèmes ou de répondre aux incidents.
- Décisions imprévues sur les systèmes à partir desquels peut se faire la collecte des journaux et métriques lors de l'examen d'un événement.
- Non-conservation des métriques et événements pendant suffisamment longtemps pour pouvoir récupérer les données.

Avantages liés au respect de cette bonne pratique : La capture des playbooks garantit que les processus peuvent être suivis de manière cohérente. La codification de vos playbooks limite l'introduction d'erreurs à partir de l'activité manuelle. L'automatisation des playbooks accélère le

temps de réponse à un événement en évitant aux membres de l'équipe d’intervenir ou en leur fournissant des informations supplémentaires lorsque leur intervention commence.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Utilisez des playbooks pour identifier les problèmes. Les playbooks sont des processus documentés pour enquêter sur les problèmes. Mettez en œuvre des réponses cohérentes et rapides aux échecs en documentant les processus dans des playbooks. Les playbooks doivent contenir les informations et les instructions nécessaires pour permettre à une personne compétente de recueillir les informations pertinentes, identifier les causes potentielles de défaillance, isoler les pannes et déterminer les facteurs adjuvants (c'est-à-dire effectuer une analyse post-incident).
  - Mettez en œuvre les playbooks en tant que code Effectuez vos opérations en tant que code scriptant vos playbooks afin d'en assurer la cohérence et de limiter les erreurs causées par les processus manuels. Les playbooks peuvent être composés de plusieurs scripts représentant les différentes étapes qui pourraient être nécessaires pour identifier les facteurs contribuant à un problème. Les activités Runbook peuvent être déclenchées ou effectuées dans le cadre d'activités playbook, ou peuvent demander l'exécution d'un playbook en réponse à des événements identifiés.
    - Automatisez vos playbooks opérationnels avec AWS Systems Manager
    - AWS Systems Manager Run Command
    - AWS Systems Manager Automation
    - Qu'est-ce qu'AWS Lambda ?
    - Qu'est-ce qu'Amazon EventBridge ?
    - Utilisation des alarmes Amazon CloudWatch

Ressources

Documents connexes :

- AWS Systems Manager Automation
- AWS Systems Manager Run Command
- Automatisez vos playbooks opérationnels avec AWS Systems Manager
- Utilisation des alarmes Amazon CloudWatch
- Utilisation de scripts Canary (Amazon CloudWatch Synthetics)

- Qu'est-ce qu'Amazon EventBridge ?
- Qu'est-ce qu'AWS Lambda ?

Exemples connexes :

- Automatisation des opérations avec les playbooks et les runbooks

REL12-BP02 Effectuer une analyse post-incident

Passez en revue les événements ayant un impact sur le client et identifiez les facteurs adjuvants et les mesures préventives. Utilisez ces informations pour prendre des mesures d'atténuation afin de limiter ou de remédier aux problèmes. Développez des procédures pour fournir des réponses rapides et efficaces. Publiez, le cas échéant, les facteurs adjuvants et les mesures correctives adaptées au public ciblé. Vous devez disposer d'une méthode pour communiquer ces causes à d'autres si nécessaire.

Évaluez pourquoi les tests existants n'ont pas permis de résoudre le problème. Ajoutez des tests pour ce cas si aucun test correspondant n'existe.

Résultat souhaité : vos équipes ont adopté une approche cohérente et concertée pour gérer l'analyse post-incident. L'un des mécanismes est le processus de correction des erreurs (COE). Celui-ci aide vos équipes à identifier, comprendre et traiter les causes profondes des incidents, tout en mettant en place des mécanismes et des barrières de protection pour limiter la probabilité qu'un incident se reproduise.

Anti-modèles courants :

- Trouver des facteurs adjuvants sans pour autant continuer à chercher plus en profondeur d'autres problèmes et approches potentiels pour atténuer les risques.
- Se contenter d'identifier les causes des erreurs humaines et ne pas proposer de formation ou d'automatisation susceptibles d'empêcher les erreurs humaines.
- Se concentrer sur les reproches plutôt que sur la compréhension des causes profondes, ce qui crée une culture de la peur et empêche de communiquer ouvertement
- Absence de partage d'informations, qui entrave la circulation des résultats de l'analyse de l'incident et empêche les autres de bénéficier des enseignements tirés
- Absence de mécanisme permettant de capturer les connaissances institutionnelles, ce qui engendre une perte d'informations précieuses en ne conservant pas les enseignements tirés sous

la forme de bonnes pratiques actualisées, et entraîne la répétition d'incidents ayant des causes profondes identiques ou similaires

Avantages liés au respect de cette bonne pratique : la réalisation d'une analyse post-incident et le partage des résultats permettent à d'autres charges de travail d'atténuer les risques si elles ont les mêmes facteurs contributifs. Cela permet aussi de mettre en œuvre la mesure d'atténuation ou de récupération automatisée avant qu'un incident ne se produise.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Une bonne analyse post-incident permet de proposer des solutions courantes pour les problèmes avec des modèles d'architecture utilisés dans d'autres compartiments de vos systèmes.

La documentation et la résolution des problèmes sont l'une des pierres angulaires du processus COE. Il est recommandé de définir une méthode normalisée pour documenter les causes profondes et de veiller à ce qu'elles soient examinées et traitées. Attribuez clairement la responsabilité du processus d'analyse post-incident. Désignez une équipe ou une personne chargée de superviser les enquêtes et le suivi de l'incident.

Encouragez une culture axée sur l'apprentissage et l'amélioration plutôt que sur les reproches. Insistez sur le fait que l'objectif est de prévenir de futurs incidents, et non de pénaliser des individus.

Élaborez des procédures bien définies pour mener les analyses post-incident. Ces procédures doivent décrire les étapes à suivre, les informations à collecter et les principales questions à aborder lors de l'analyse. Enquêtez en profondeur sur les incidents, en allant au-delà des causes immédiates afin d'identifier les causes profondes et les facteurs contributifs. Utilisez des techniques telles que les cinq pourquoi pour approfondir les problèmes sous-jacents.

Tenez un répertoire des enseignements tirés des analyses des incidents. Ces connaissances institutionnelles peuvent servir de référence pour les incidents futurs et les efforts de prévention. Partagez les conclusions et les réflexions tirées des analyses post-incident, et envisagez d'organiser des réunions de synthèse post-incident ouvertes à tous pour discuter des enseignements tirés.

Étapes d'implémentation

- Veillez à ce que l'analyse post-incident soit exempte de tout reproche. Cela permet aux personnes impliquées dans l'incident de faire preuve d'objectivité quant aux actions correctives proposées, et de promouvoir une auto-évaluation et une collaboration honnêtes entre les équipes.

- Définissez une méthode standardisée pour documenter les problèmes critiques. Voici un exemple de structure :
  - Que s'est-il passé ?
  - Quel a été l'impact sur vos clients et votre activité ?
  - Quelle était la cause profonde ?
  - Quelles sont les données à votre disposition pour étayer votre raisonnement ?
    - Par exemple, des métriques et des graphiques.
  - Quelles ont été les principales répercussions, notamment en termes de sécurité ?
    - Lors de la conception des charges de travail, vous faites un compromis entre les piliers en fonction de votre activité. Ces décisions métier peuvent vous aider à gérer vos priorités techniques. Vous pouvez opter pour l'optimisation afin de réduire les coûts au détriment de la fiabilité dans les environnements de développement ou, pour les solutions stratégiques, vous pouvez optimiser la fiabilité pour des coûts plus importants. La sécurité est toujours une priorité, car vous devez protéger vos clients.
  - Quelles leçons avez-vous apprises ?
  - Quelles mesures correctives allez-vous prendre ?
    - Éléments d'action
    - Articles connexes
- Élaborez des procédures bien définies pour mener les analyses post-incident.
- Mettez en place un processus standardisé de signalement des incidents. Documentez tous les incidents de manière exhaustive, y compris le rapport d'incident initial, les journaux, les communications et les mesures prises pendant l'incident.
- N'oubliez pas qu'un incident n'est pas forcément une panne. Il peut s'agir d'un accident évité de justesse ou d'un système qui fonctionne de manière inattendue tout en remplissant sa fonction.
- Améliorez sans cesse votre processus d'analyse post-incident en fonction des retours et des enseignements tirés.
- Capturez les principales conclusions dans un système de gestion des connaissances et examinez les modèles qui devraient être ajoutés aux guides du développeur ou aux listes de contrôle de prédéploiement.

Ressources

Documents connexes :

- Pourquoi mettre en place la correction des erreurs (COE)

Vidéos connexes :

- Amazon’s approach to failing successfully
- AWS re:Invent 2021 - Amazon Builders’ Library: Operational Excellence at Amazon

REL12-BP03 Tester les exigences fonctionnelles

Utilisez des techniques telles que les tests unitaires et les tests d'intégration qui valident les fonctionnalités requises.

Vous obtenez les meilleurs résultats lorsque ces tests sont exécutés automatiquement dans le cadre d'actions de génération et de déploiement. Par exemple, grâce à AWS CodePipeline, les développeurs valident les modifications apportées à un référentiel source dans lequel CodePipeline détecte automatiquement les modifications. Ces modifications sont générées et des tests sont exécutés. Une fois les tests terminés, le code généré est déployé sur des serveurs intermédiaires à des fins de test. Depuis le serveur intermédiaire, CodePipeline exécute d'autres tests, tels que des tests d'intégration ou de chargement. Une fois ces tests terminés avec succès, CodePipeline déploie le code testé et approuvé sur les instances de production.

De plus, l'expérience montre que les tests de transactions synthétiques (également appelés tests canaryà ne pas confondre avec les déploiements canary) qui peuvent exécuter et simuler le comportement des clients font partie des processus de test les plus importants. Exécutez ces tests en permanence sur vos points de terminaison de charge de travail à partir de divers emplacements distants. Amazon CloudWatch Synthetics vous permet de créer des scripts Canari pour surveiller vos points de terminaison et vos API.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Testez les exigences fonctionnelles. Il s'agit, entre autres, des tests unitaires et des tests d'intégration qui valident les fonctionnalités requises.
  - Utilisez CodePipeline avec AWS CodeBuild pour tester le code et exécuter des générations
  - AWS CodePipeline ajoute la prise en charge des tests unitaires et des tests d'intégration personnalisés avec AWS CodeBuild
  - Livraison et intégration continues (CI/CD)

- Utilisation de scripts Canary (Amazon CloudWatch Synthetics)
- Automatisation des tests logiciels

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant faciliter l'implémentation d'un pipeline d'intégration continue
- AWS CodePipeline ajoute la prise en charge des tests unitaires et des tests d'intégration personnalisés avec AWS CodeBuild
- AWS Marketplace : produits pouvant être utilisés pour une intégration continue
- Livraison et intégration continues (CI/CD)
- Automatisation des tests logiciels
- Utilisez CodePipeline avec AWS CodeBuild pour tester le code et exécuter des générations
- Utilisation de scripts Canary (Amazon CloudWatch Synthetics)

REL12-BP04 Tester les exigences de mise à l'échelle et de performance

Utilisez des techniques telles que les tests de charge pour valider que la charge de travail répond aux exigences de mise à l'échelle et de performances.

Dans le cloud, vous pouvez créer un environnement de test à la demande à l'échelle de la production pour votre charge de travail. Si vous exécutez ces tests sur une infrastructure réduite, vous devez mettre vos résultats observés à l'échelle en fonction de ce que vous pensez qu'il se produira en production. Les tests de charge et de performances peuvent également être réalisés en production si vous veillez à ne pas affecter les utilisateurs réels et à baliser vos données de test afin qu'elles ne correspondent pas aux données utilisateur réelles et ne corrompent pas les statistiques d'utilisation ni les rapports de production.

Utilisez les tests pour vous assurer que vos ressources de base, vos paramètres de mise à l'échelle, vos quotas de service et votre conception de la résilience fonctionnent comme prévu sous charge.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

- Testez les exigences de mise à l'échelle et de performance. Effectuez un test de charge pour vérifier que la charge de travail répond aux exigences de mise à l'échelle et de performance.

- Test de charge distribuée sur AWS : simulation de milliers d'utilisateurs connectés
- Apache JMeter
  - Déployez votre application dans un environnement identique à votre environnement de production et effectuez un test de charge.
    - Utilisez les concepts d'Infrastructure as Code pour créer un environnement aussi similaire que possible à votre environnement de production.

Ressources

Documents connexes :

- Test de charge distribuée sur AWS : simulation de milliers d'utilisateurs connectés
- Apache JMeter

REL12-BP05 Tester la résilience à l'aide de l'ingénierie du chaos

Exécutez des expériences de chaos dans des environnements dont les conditions se rapprochent autant que possible de la production pour comprendre comment nos systèmes réagissent à des conditions défavorables.

Résultat souhaité :

La résilience de la charge de travail est régulièrement vérifiée en appliquant l'ingénierie du chaos sous la forme d'expériences d'injection de défaillances ou de charge inattendue, en plus des tests de résilience qui confirment le comportement attendu connu de votre charge de travail lors d'un événement. Associez l'ingénierie du chaos aux tests de résilience pour avoir l'assurance que votre charge de travail peut résister en cas de défaillance des composants et récupérer suite à des perturbations inattendues avec peu ou pas d'impact.

Anti-modèles courants :

- Conception à des fins de résilience, mais pas de vérification du fonctionnement de la charge de travail dans son ensemble en cas de défaillances.
- Pas d'expériences dans des conditions concrètes et pour la charge prévue.
- Pas de traitement de vos expériences en tant que code ou de maintenance de vos expériences tout au long du cycle de développement.

- Pas d'exécution d'expériences de chaos dans le cadre de votre pipeline CI/CD, ainsi qu'en dehors des déploiements.
- Pas d'utilisation des analyses passées post-incident pour déterminer les défaillances à tester.

Avantages liés au respect de cette bonne pratique : L'injection de défaillances pour vérifier la résilience de votre charge de travail vous permet d'avoir l'assurance que les procédures de récupération de votre conception résiliente fonctionneront en cas de défaillances réelles.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

L'ingénierie du chaos offre la possibilité à vos équipes d'injecter en continu des perturbations concrètes (simulations) de manière contrôlée au niveau du fournisseur de services, de l'infrastructure, de la charge de travail et des composants, avec peu ou pas d'impact pour vos clients. Ainsi, vos équipes tirent les leçons de ces défaillances et observent, mesurent et améliorent la résilience de vos charges de travail, tout en confirmant que les alertes se déclenchent et que les équipes sont informées en cas d'événement.

Une pratique de l'ingénierie du chaos en continu peut mettre en évidence des défaillances dans vos charges de travail qui, si elles ne sont pas résolues, peuvent impacter de manière négative la disponibilité et le fonctionnement.

> **Note**
>
> L'ingénierie du chaos est la discipline d'expérimentation d'un système. Elle permet de s'assurer de la capacité du système à résister à des conditions de production difficiles. – Principes de l'ingénierie du chaos

Si un système est capable de résister à ces perturbations, l'expérience de chaos doit être maintenue en tant que test de régression automatisé. De cette façon, les expériences de chaos doivent être réalisées dans le cadre de votre cycle de développement des systèmes et de votre pipeline CI/CD.

Pour veiller à ce que votre charge de travail résiste en cas de défaillance des composants, injectez des événements concrets dans le cadre de vos expériences. Par exemple, expérimentez une perte des instances Amazon EC2 ou un basculement de l'instance de base de données Amazon RDS principale, puis vérifiez que votre charge de travail n'est pas impactée (ou très peu). Utilisez plusieurs

défaillances des composants pour simuler des événements capables de causer une perturbation dans une zone de disponibilité.

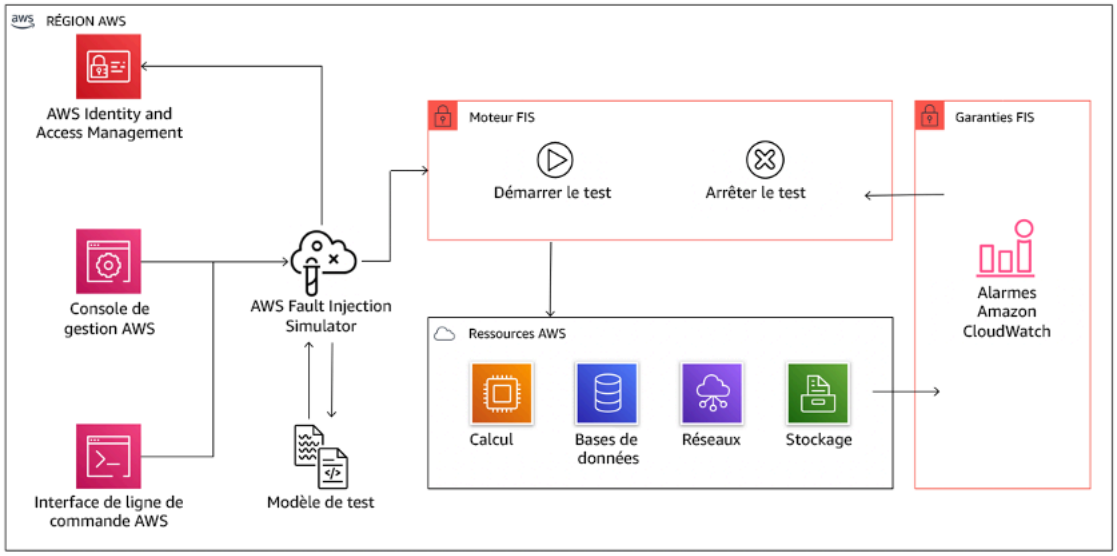
Pour les défaillances de niveau application (telles que les plantages), commencez par des tests de stress comme l'épuisement de la mémoire et du processeur.

Afin de valider les solutions de secours ou les mécanismes de basculement pour les dépendances externes dues aux pannes réseau intermittentes, vos composants doivent simuler un tel événement en bloquant l'accès aux fournisseurs tiers pendant une durée spécifiée pouvant aller de quelques secondes à plusieurs heures.

D'autres modes de dégradation peuvent entraîner des fonctionnalités limitées et ralentir les réponses, ce qui se traduit par une perturbation de vos services. Généralement, cette dégradation résulte d'une latence accrue sur les services critiques et d'une communication réseau peu fiable (perte de paquets). Les expériences avec ces défaillances, dont les effets de mise en réseau tels que la latence, les messages supprimés et les défaillances DNS, peuvent inclure l'incapacité de résoudre un nom, d'atteindre un service DNS ou de se connecter aux services dépendants.

Outils de l'ingénierie du chaos :

AWS Fault Injection Service (AWS FIS) est un service entièrement géré permettant l'exécution d'expériences d'injection de défaillances qui peuvent être utilisées dans le cadre de votre pipeline CD, ou en dehors du pipeline. AWS FIS s'impose donc comme un choix judicieux lors des tests de simulation de pannes. Il prend en charge de manière simultanée l'injection de défaillances sur différents types de ressources dont Amazon EC2, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) et Amazon RDS. Ces défaillances incluent l'arrêt des ressources, les basculements forcés, le stress du processeur ou de la mémoire, la limitation, la latence et la perte de paquets. Comme il est intégré aux alarmes Amazon CloudWatch, vous pouvez définir des conditions d'arrêt comme barrières de protection pour annuler une expérience si elle provoque un impact inattendu.



AWS Fault Injection Service s'intègre aux ressources AWS pour vous permettre d'exécuter des expériences d'injection de défaillances pour vos charges de travail.

Il existe également plusieurs options tierces pour les expériences d'injection de défaillances. Il existe notamment des outils open source tels que Chaos Toolkit, Chaos Meshet Litmus Chaos, ainsi que des options commerciales comme Gremlin. Pour élargir le champ des défaillances pouvant être injectées sur AWS, AWS FIS prend désormais en charge Chaos Mesh et Litmus Chaos, ce qui vous permet de coordonner les flux de travail d'injection des défaillances entre plusieurs outils. Par exemple, vous pouvez exécuter un test de stress sur un processeur de pod à l'aide des défaillances Chaos Mesh ou Litmus, tout en arrêtant un pourcentage de nœuds de cluster sélectionné de façon aléatoire grâce aux actions des défaillances AWS FIS.

Étapes d'implémentation

- Déterminer les défaillances à utiliser pour les expériences.

  Évaluez la conception de votre charge de travail à des fins de résilience. De telles conceptions (créées à l'aide des bonnes pratiques de Le cadre AWS Well-Architected) tiennent compte des risques en se basant sur des dépendances critiques, des événements passés, des erreurs connues et des exigences de conformité. Répertoriez chaque élément de la conception destiné à maintenir la résilience et les défaillances qu'il entend réduire. Pour plus d'informations sur la création de telles listes, consultez le livre blanc Examens de disponibilité opérationnelle qui vous guidera dans la création d'un processus capable de prévenir la répétition des incidents précédents. Le processus de Failure Modes and Effects Analysis (FMEA) ou d'analyse des modes de défaillance et de leurs effets vous propose un framework pour réaliser une analyse de niveau

composant des défaillances et de leur impact sur votre charge de travail. Le processus FMEA est décrit plus en détail par Adrian Cockcroft dans l'article Failure Modes and Continuous Resilience (modes de défaillance et résilience continue).

- Attribuer une priorité à chaque défaillance.

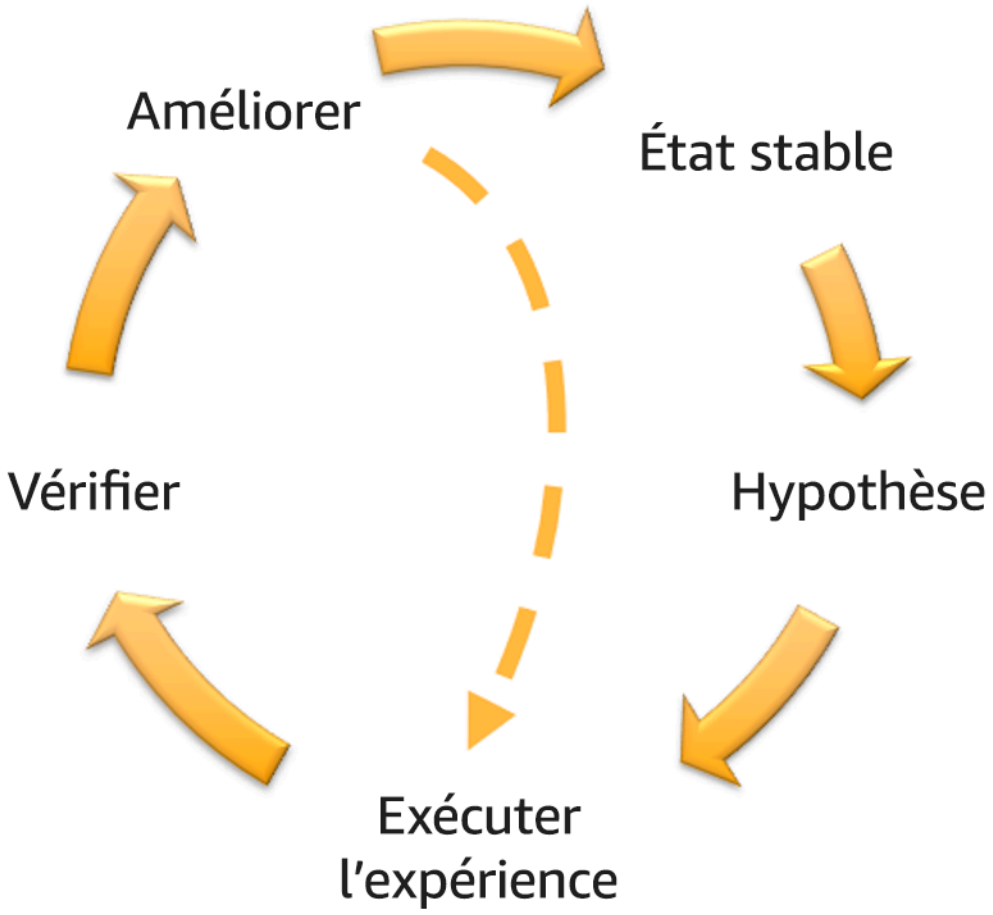
  Commencez par définir une classification grossière telle que élevée, moyenne et basse. Pour évaluer les priorités, tenez compte de la fréquence de la défaillance et de son impact sur la charge de travail dans son ensemble.

  Lors de la prise en compte de la fréquence d'une défaillance donnée, analysez les données passées de cette charge de travail, le cas échéant. Si aucune donnée passée n'est disponible, utilisez les données des autres charges de travail s'exécutant dans un environnement semblable.

  Lors de la prise en compte de l'impact d'une défaillance donnée, souvenez-vous qu'en général plus le champ de la défaillance est large, plus grand est l'impact. Tenez compte également de la conception de la charge de travail et de son objectif. Par exemple, la capacité à accéder aux magasins de données sources est essentielle pour une charge de travail effectuant des transformations et de l'analyse de données. Dans ce cas, vous donnerez la priorité aux expériences liées aux défaillances d'accès ainsi qu'aux accès limités et à l'insertion de la latence.

  Les analyses post-incident constituent une excellente source de données pour comprendre à la fois la fréquence et l'impact des modes de défaillance.

  Utilisez la priorité attribuée pour déterminer les défaillances à expérimenter en premier lieu, puis l'ordre dans lequel développer de nouvelles expériences d'injection de défaillances.
- Suivre le volant d'inertie de l'ingénierie du chaos et de la résilience continue pour chaque expérience réalisée.



Volant d'inertie de l'ingénierie du chaos et de la résilience continue réalisé grâce à la méthode scientifique d'Adrian Hornsby.

- Définir l'état stable comme le résultat mesurable d'une charge de travail qui indique un comportement normal.

  Votre charge de travail présente un état stable si elle fonctionne de manière fiable et comme prévu. Par conséquent, confirmez que charge de travail est saine avant de définir un état stable. L'état stable ne signifie pas forcément sans impact pour la charge de travail en cas de défaillance, car un certain pourcentage des défaillances n'excède pas des limites supportables. L'état stable constitue le repère que vous observerez pendant l'expérience, qui mettra en évidence des anomalies si votre hypothèse formulée dans l'étape suivante ne donne pas les résultats escomptés.

  Par exemple, un état stable d'un système de paiements peut être défini comme le traitement de 300 TPS avec un taux de réussite de 99 % et un temps de transmission aller-retour de 500 ms.

- Formuler une hypothèse sur la façon dont la charge de travail réagira à la défaillance.

  Une bonne hypothèse repose sur la façon dont la charge de travail est destinée à réduire la défaillance pour maintenir l'état stable. L'hypothèse indique que vu qu'il s'agit d'une défaillance d'un type particulier, le système ou la charge de travail maintiendra un état stable, car la charge de travail a été conçue avec une atténuation des risques spécifique. Le type particulier de défaillance et d'atténuation des risques doit être spécifié dans l'hypothèse.

  Le modèle suivant peut être utilisé pour l'hypothèse (mais une autre formulation est aussi acceptable) :

  > **Note**
  >
  > Si *`défaillance spécifique`* se produit, la charge de travail *`nom de la charge de travail`* va *`décrire les contrôles pour atténuer les risques`* afin de limiter *`impact métier ou technique`*.

  Par exemple :

  - Si 20 % des nœuds du node-group Amazon EKS sont supprimés, l'API Transaction Create API continue de répondre au 99e centile des demandes en moins de 100 ms (état stable). Les nœuds Amazon EKS seront opérationnels dans les cinq minutes, et les pods seront planifiés et traiteront le trafic huit minutes après le début de l'expérience. Les alertes se déclencheront sous trois minutes.
  - En cas de défaillance d'une seule instance Amazon EC2, la surveillance de l'état Elastic Load Balancing du système de commandes permet à Elastic Load Balancing d'envoyer uniquement des demandes aux instances saines restantes, tandis qu'Amazon EC2 Auto Scaling remplace l'instance en échec, tout en maintenant une augmentation des erreurs (5xx) côté serveur (état stable) inférieure à 0,01 %.
  - Si l'instance de base de données Amazon RDS principale échoue, la charge de travail de collecte des données Chaîne d'approvisionnement basculera et se connectera à l'instance de base de données Amazon RDS de secours pour maintenir les erreurs de lecture ou d'écriture de base de données (état stable) inférieures à 1 minute.
- Exécuter l'expérience en injectant la défaillance.

  Une expérience doit par défaut être sécurisée et tolérée par la charge de travail. Si vous savez que la charge de travail va échouer, n'exécutez pas l'expérience. L'ingénierie du chaos doit être

utilisée pour rechercher les risques connus ou inconnus. Les risques connus sont les choses dont vous êtes conscient mais que vous ne comprenez pas bien, et les risques inconnus sont les choses dont vous n'êtes pas conscient ou que vous ne comprenez pas bien. Exécuter une expérience sur une charge de travail que vous savez défaillante ne vous apportera rien de plus. Votre expérience doit être soigneusement préparée, disposer d'un champ d'impact défini et fournir un mécanisme de protection pouvant être appliqué en cas de perturbations inattendues. Si votre vérification préalable indique que votre charge de travail doit résister à l'expérience, exécutez cette dernière. Il existe plusieurs moyens d'injecter les défaillances. Pour les charges de travail sur AWS, AWS FIS propose de nombreuses simulations de défaillances prédéfinies appelées des mesures. Vous pouvez également définir des actions personnalisées qui s'exécutent dans AWS FIS à l'aide des documents AWS Systems Manager.

Nous déconseillons l'utilisation de scripts personnalisés pour les expériences de chaos, sauf si ces derniers sont capables de comprendre l'état actuel de la charge de travail, d'émettre des journaux, de fournir des mécanismes de protection pour annuler une expérience et des conditions d'arrêt dans la mesure du possible.

Un framework ou des outils efficaces capables de prendre en charge l'ingénierie du chaos doivent suivre l'état actuel d'une expérience, émettre des journaux et fournir des mécanismes de protection pour prendre en charge l'exécution contrôlée d'une expérience. Commencez par un service établi comme AWS FIS qui vous permet d'exécuter des expériences avec un champ clairement défini et des mécanismes de sécurité capables de protéger l'expérience en cas de perturbations inattendues. Pour découvrir plusieurs expériences utilisant AWS FIS, consultez également l' atelier Applications résilientes et Well-Architected avec l'ingénierie du chaos. Par ailleurs, AWS Resilience Hub analysera votre charge de travail et créera des expériences que vous pourrez choisir d'implémenter et d'exécuter dans AWS FIS.

> **Note**
>
> Pour chaque expérience, vous devez bien comprendre le champ et son impact. Nous recommandons que les défaillances soient d'abord simulées sur un environnement hors production avant d'être exécutées en production.

Les expériences doivent s'exécuter en production sous une charge concrète à l'aide des déploiements canary qui mettent en place des déploiements de système de contrôles et d'expériences, sous réserve de faisabilité. L'exécution d'expériences pendant les heures creuses

est une bonne pratique pour réduire l'impact potentiel de la première expérience en production. De plus, si l'utilisation du trafic client réel s'avère trop risquée, vous pouvez exécuter des expériences à l'aide du trafic synthétique sur l'infrastructure de production pour des déploiements de système de contrôles et d'expériences. Lorsqu'une exécution en production n'est pas possible, exécutez les expériences dans des environnements de pré-production aussi proches que possible de la production.

Vous devez définir des barrières de protection pour veiller à ce que l'expérience n'impacte pas le trafic de la production ou d'autres systèmes au-delà des limites acceptables. Définissez des conditions d'arrêt pour stopper une expérience si elle atteint le seuil d'une métrique de barrière protection défini par vos soins. Ces conditions doivent inclure les métriques de l'état stable de la charge de travail, ainsi que celles sur les composants dans lesquels vous injectez la défaillance. A surveillance synthétique (également appelée un utilisateur canary) est une métrique que vous devez généralement inclure en tant que proxy utilisateur. Les conditions d'arrêt pour AWS FIS sont prises en charge dans le cadre d'un modèle de test, autorisant jusqu'à cinq conditions d'arrêt par modèle.

L'un des principes de l'ingénierie du chaos est de minimiser le champ de l'expérience et son impact :

Bien qu'un impact négatif à court terme soit autorisé, l'ingénieur du chaos a la responsabilité et l'obligation de minimiser et de maîtriser les conséquences des expériences.

Pour vérifier le champ et l'impact potentiel, vous pouvez dans un premier temps exécuter l'expérience dans un environnement hors production, en vérifiant que les seuils des conditions d'arrêt s'activent comme prévu pendant l'expérience et que l'observabilité est en place pour détecter une exception, plutôt que d'exécuter l'expérience directement en production.

Lorsque vous exécutez des expériences d'injection de défaillances, vérifiez que toutes les parties responsables sont bien informées. Communiquez avec les équipes appropriées, telles que les équipes en charge des opérations, les équipes chargées de la fiabilité du service et le service client pour leur indiquer quand les expériences seront exécutées et à quoi ils doivent s'attendre. Donnez à ces équipes les outils de communication nécessaires pour informer les personnes en charge de l'exécution de l'expérience si elles constatent des effets négatifs.

Vous devez restaurer la charge de travail et ses systèmes sous-jacents dans leur état fonctionnel et connu d'origine. En général, la conception résiliente de la charge de travail lui permet de s'auto-réparer. Cependant, certaines conceptions défaillantes ou échecs d'expériences peuvent

laisser votre charge de travail dans un état d'échec inattendu. À la fin de l'expérience, vous devez en être conscient et restaurer la charge de travail et les systèmes. Avec AWS FIS, vous pouvez définir une configuration de barrière de protection (également appelée post action) dans les paramètres d'action. Une post action restaure la cible dans l'état dans lequel elle se trouvait avant l'exécution de l'action. Qu'elles soient automatisées (comme lorsque vous utilisez AWS FIS) ou manuelles, ces post actions doivent faire partie d'un playbook décrivant la façon de détecter et de gérer les échecs.

- Vérifier l'hypothèse.

  Principes de l'ingénierie du chaos donne des conseils sur la façon de vérifier l'état stable de votre charge de travail :

  Concentrez-vous sur le résultat mesurable d'un système, plutôt que sur les attributs internes du système. Les mesures de ce résultat sur une courte période de temps constituent un proxy pour l'état stable du système. Le débit général du système, les taux d'erreur et les centiles de latence peuvent tous être des métriques d'intérêt représentant un comportement d'état stable. En se focalisant sur les modèles de comportement systémique pendant les expériences, l'ingénierie du chaos vérifie que le système fonctionne, au lieu d'essayer de confirmer qu'il fonctionne.

  Dans nos deux exemples précédents, nous incluons la métrique de l'état stable inférieure à 0,01 % d'augmentation des erreurs (5xx) côté serveur et la métrique inférieure à 1 minute d'erreurs de lecture ou d'écriture de base de données.

  Les erreurs 5xx constituent une bonne métrique, car elles sont une conséquence du mode de défaillance dont le client de la charge de travail fera l'expérience directement. La mesure des erreurs de base de données est correcte en tant que conséquence directe de la défaillance, mais doit être également complétée par une mesure d'impact, telle que les échecs de demandes client ou les erreurs remontées. Par ailleurs, incluez une surveillance synthétique (également appelée utilisateur canary) sur n'importe quelle API ou URI directement accessible par le client de votre charge de travail.

- Améliorer la conception de la charge de travail à des fins de résilience.

  Si l'état stable n'a pas été maintenu, enquêtez sur les moyens d'améliorer la conception de la charge de travail afin de réduire la défaillance, tout en appliquant les bonnes pratiques du pilier Fiabilité du cadre AWS Well-Architected. Des conseils et ressources supplémentaires sont disponibles dans la bibliothèque des créateurs AWS, qui héberge des articles sur la façon d' améliorer vos surveillances de l'état ou utiliser de nouvelles tentatives avec interruption dans votre code d'application, entre autres.

Une fois ces changements implémentés, exécutez de nouveau l'expérience (illustrée par la ligne pointillée dans le volant d'inertie de l'ingénierie du chaos) pour déterminer son efficacité. Si l'étape de vérification indique que l'hypothèse est vraie, alors la charge de travail sera en état stable et le cycle continuera.

- Exécuter régulièrement des expériences.

  Une expérience de chaos est un cycle, et les expériences doivent être exécutées régulièrement dans le cadre de l'ingénierie du chaos. Lorsqu'une charge de travail correspond à l'hypothèse d'une expérience, cette dernière doit être automatisée pour s'exécuter en continu en tant que test de régression de votre pipeline CI/CD. Pour savoir comment faire, consultez ce blog sur l'exécution d'expériences AWS FIS en utilisant AWS CodePipeline. Cet atelier sur les expériences AWS FIS récurrentes dans un pipeline CI/CD vous permet de mettre la main à la pâte.

  Les expériences d'injection de défaillance font également partie des tests de simulation de pannes (consultez REL12-BP06 Organiser régulièrement des tests de simulation de panne). Les tests de simulation de pannes simulent une défaillance ou un événement pour vérifier les systèmes, les processus et la réponse de l'équipe. L'objectif est d'effectuer les actions que l'équipe effectuerait si un événement exceptionnel se produisait.

- Enregistrer et stocker les résultats des expériences.

  Les résultats des expériences d'injection de défaillance doivent être enregistrés et conservés. Incluez toutes les données nécessaires (telles que l'heure, la charge de travail et les conditions) afin de pouvoir analyser ultérieurement les résultats de l'expérience et les tendances. Les exemples de résultats peuvent inclure des captures d'écran des tableaux de bord, des fichiers CSV de la base de données de votre/vos métriques ou un registre manuscrit des événements et observations pendant l'expérience. La journalisation des expériences avec AWS FIS peut faire partie de la collecte des données.

Ressources

Bonnes pratiques associées :

- REL08-BP03 Intégrer les tests de résilience dans le cadre de votre déploiement
- REL13-BP03 Effectuer un test de validation de la mise en œuvre de la reprise après sinistre

Documents connexes :

- Qu'est-ce qu'AWS Fault Injection Service ?
- Qu'est-ce qu'AWS Resilience Hub ?
- Principes de l'ingénierie du chaos
- Ingénierie du chaos : préparation de votre première expérience
- Ingénierie de résilience : apprendre à intégrer les pannes
- Témoignages d'utilisation de l'ingénierie du chaos
- Éviter les solutions de secours dans les systèmes distribués
- Déploiement canary pour des expériences de chaos

Vidéos connexes :

- AWS re:Invent 2020: Testing resiliency using chaos engineering (ARC316)
- AWS re:Invent 2019: Improving resiliency with chaos engineering (DOP309-R1)
- AWS re:Invent 2019: Performing chaos engineering in a serverless world (CMY301)

Exemples connexes :

- Atelier Well-Architected : niveau 300 : test de la résilience d'Amazon EC2, Amazon RDS et Amazon S3
- Atelier L'ingénierie du chaos dans AWS
- atelier Applications résilientes et Well-Architected avec l'ingénierie du chaos
- Atelier Chaos sans serveur
- Atelier Mesurer et améliorer la résilience de votre application avec AWS Resilience Hub

Outils associés :

- AWS Fault Injection Service
- AWS Marketplace : Gremlin Chaos Engineering Platform
- Chaos Toolkit
- Chaos Mesh
- Litmus

REL12-BP06 Organiser régulièrement des tests de simulation de panne

Utilisez des tests de simulation de panne pour exercer régulièrement vos procédures de réponse aux événements et aux défaillances aussi près que possible de la production (y compris dans les environnements de production) avec les personnes qui seront impliquées dans les scénarios de défaillance réels. Les tests de simulation de panne appliquent des mesures pour s'assurer que les événements de production n'affectent pas les utilisateurs.

Ils simulent une défaillance ou un événement pour tester les systèmes, les processus et la réponse de l'équipe. L'objectif est d'effectuer les actions que l'équipe effectuerait si un événement exceptionnel se produisait. Cela vous aidera à comprendre où apporter des améliorations et à développer une expérience de gestion des événements au sein de votre organisation. Des tests de simulation de panne doivent être effectués régulièrement afin que votre équipe se forge une « mémoire musculaire » quant à la façon de réagir.

Une fois votre conception de résilience en place et testée dans des environnements non liés à la production, un test de simulation de panne permet de s'assurer que tout fonctionne comme prévu en production. Un test de simulation de pannes, particulièrement le premier, est une activité « exploitant toutes les ressources ». L’intégralité des ingénieurs et des opérations est informée de ce qui se passera et quand. Les playbooks sont en place. Des événements simulés sont exécutés, y compris des événements de défaillance possibles, dans les systèmes de production de la manière prescrite, et l'impact est évalué. Si tous les systèmes fonctionnent comme prévu, la détection et l'auto-réparation se produiront avec peu, voire aucun impact. En revanche, si un impact négatif est observé, le test est annulé et les problèmes de charge de travail sont résolus, manuellement au besoin (à l'aide du runbook). Étant donné que les tests de simulation de pannes se déroulent souvent en production, toutes les précautions doivent être prises pour s'assurer de l'absence d’impact sur la disponibilité pour vos clients.

Anti-modèles courants :

- Documenter vos procédures sans jamais les exercer.
- Non-inclusion des décideurs commerciaux dans les exercices de test.

Avantages liés au respect de cette bonne pratique : L'organisation régulière de tests de simulation de panne a un double avantage. D'une part, elle permet de s'assurer que tout le personnel suit les stratégies et les procédures lorsqu'un incident réel se produit. D'autre part, elle facilite la validation de l'adéquation de ces stratégies et procédures.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Planifiez des tests de simulation de panne pour tester régulièrement vos runbooks et vos playbooks. Les tests de simulation de panne doivent impliquer tous ceux qui seraient affectés par une interruption de production : le propriétaire de l'entreprise, les développeurs, le personnel d'exploitation et les équipes d'interventions.
  - Effectuez vos tests de charge ou de performances et mettez en œuvre l'injection de défaillances.
  - Recherchez des anomalies dans vos runbooks et des possibilités de test de vos playbooks.
    - Si vous vous écartez de vos runbooks, affinez-les ou corrigez le comportement. Lors des tests de votre playbook,, identifiez les runbooks qui auraient dû être utilisés ou créez-en de nouveaux.

Ressources

Documents connexes :

- Qu'est-ce qu'AWS GameDay ?

Vidéos connexes :

- AWS re:Invent 2019: Improving resiliency with chaos engineering (DOP309-R1)

  Exemples connexes :

- AWS Well-Architected Labs - Testing Resiliency

FIA 13. Comment planifier la reprise après sinistre (DR) ?

La mise en place de sauvegardes et de composants de charge de travail redondants constitue le début de votre stratégie de DR. RTO et RPO sont vos objectifs pour la restauration de votre charge de travail. Définissez-les en fonction des besoins de l'entreprise. Mettez en œuvre une stratégie pour atteindre ces objectifs, en particulier en tenant compte de l'emplacement et de la fonction des données et des ressources de charge de travail. La probabilité d'une perturbation et le coût de la reprise sont également des facteurs clés qui permettent de déterminer la valeur opérationnelle de la reprise après sinistre d'une charge de travail.

Bonnes pratiques

- REL13-BP01 Définir les objectifs de reprise pour les temps d'arrêt et les pertes de données
- REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise
- REL13-BP03 Effectuer un test de validation de la mise en œuvre de la reprise après sinistre
- REL13-BP04 Gérer l'écart de configuration au niveau du site ou de la région de reprise après sinistre
- REL13-BP05 Automatiser la reprise

REL13-BP01 Définir les objectifs de reprise pour les temps d'arrêt et les pertes de données

La charge de travail est associée à un objectif de délai de reprise (RTO) et à un objectif de point de reprise (RPO).

La durée maximale d'interruption admissible (RTO) correspond au délai maximum acceptable entre l'interruption du service et la restauration du service. Elle détermine ce qui est considéré comme étant un créneau de temps acceptable d'indisponibilité du service.

L'objectif de point de reprise (RPO) correspond au temps maximal acceptable depuis le dernier point de reprise des données. Il détermine ce qui est considéré comme étant une perte de données acceptable entre le dernier point de reprise et l'interruption du service.

Les valeurs RTO et RPO sont des considérations importantes lors de la sélection d'une stratégie de reprise après sinistre adaptée à votre charge de travail. Ces objectifs sont déterminés par l'entreprise, puis utilisés par les équipes techniques pour sélectionner et mettre en œuvre une stratégie de reprise après sinistre.

Résultat souhaité :

Un RTO et un RPO, définis en fonction de l'impact sur l'entreprise, sont attribués à chaque charge de travail. Un niveau prédéfini, définissant la disponibilité du service et une perte de données acceptable, avec un RTO et un RPO associés est assigné à la charge de travail. Si cette hiérarchisation n'est pas possible, elle peut être attribuée sur mesure pour chaque charge de travail, dans l'intention de créer des niveaux ultérieurement. Le RTO et le RPO font partie des principaux éléments pris en compte pour la sélection de la mise en œuvre d'une stratégie de reprise après sinistre pour la charge de travail. D'autres considérations dans le choix d'une stratégie de reprise après sinistre sont les contraintes de coût, les dépendances de la charge de travail et les exigences opérationnelles.

Pour le RTO, identifiez l'impact en fonction de la durée d'une panne. Est-il linéaire ou non (par exemple, après quatre heures, vous arrêtez une ligne de fabrication jusqu'au début du prochain quart de travail) ?

Une matrice de reprise après sinistre, comme la suivante, peut vous aider à comprendre dans quelle mesure l'ordre d'importance de la charge de travail est lié aux objectifs de reprise. Notez que les valeurs réelles des axes X et Y doivent être personnalisées en fonction des besoins de votre organisation.

Matrice de reprise après sinistre						
		Objectif de point de reprise				
		Moins de 1 minute	Moins de 1 heure	Moins de 6 heures	Moins de 1 jour	+ de 1 jour
Durée maximale d'interruption	Moins de 10 minutes	Critique	Critique	Débit	Moyenne entreprise	Moyenne entreprise
	Moins de 2 heures	Critique	Débit	Moyenne entreprise	Moyenne entreprise	Faible
	Moins de 8 heures	Débit	Moyenne entreprise	Moyenne entreprise	Faible	Faible
	Moins de 24 heures	Moyenne entreprise	Moyenne entreprise	Faible	Faible	Faible
	+ de 24 heures	Moyenne entreprise	Faible	Faible	Faible	Faible

Figure 16 : matrice de reprise après sinistre

Anti-modèles courants :

- Aucun objectif de reprise défini.
- Sélection d'objectifs arbitraires de reprise.
- Sélection d'objectifs de reprise trop lents et qui ne répondent pas aux objectifs de l'entreprise.
- Ne pas comprendre l'impact des temps d'arrêt et de la perte de données.
- Sélection d'objectifs de reprise irréalistes, tels que zéro temps de reprise et zéro perte de données, qui peuvent ne pas être réalisables pour la configuration de votre charge de travail.
- Sélection d'objectifs de reprise plus rigoureux que les objectifs commerciaux réels. Cela impose des implémentations de reprise après sinistre qui sont plus coûteuses et plus compliquées que ce dont a besoin la charge de travail.
- Sélection d'objectifs de reprise incompatibles avec ceux d'une charge de travail dépendante.
- Vos objectifs de reprise ne tiennent pas compte des exigences de conformité réglementaire.
- Définition de RTO et RPO jamais testés pour une charge de travail.

Avantages liés au respect de cette bonne pratique : Vos objectifs de reprise en cas de perte de temps et de données sont nécessaires pour guider votre implémentation de DR.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

Pour la charge de travail donnée, vous devez comprendre l'impact des temps d'arrêt et de la perte de données sur votre entreprise. L'impact augmente généralement avec les temps d'arrêt ou les pertes de données plus importants, mais son ampleur peut varier en fonction du type de charge de travail. Par exemple, vous pouvez tolérer des temps d'arrêt pouvant atteindre une heure avec peu d'impact, mais au-delà de ce délai, l'impact augmente rapidement. L'impact sur l'entreprise se manifeste sous de nombreuses formes, notamment le coût (tel que la perte de revenus), la confiance des clients (et l'impact sur la réputation), les problèmes opérationnels (tels que l'absence d'employés ou la baisse de productivité) et le risque réglementaire. Utilisez les étapes suivantes pour comprendre ces impacts et définir le RTO et le RPO pour votre charge de travail.

Étapes d'implémentation

1. Identifiez les parties prenantes spécifiques à cette charge de travail et collaborez avec elles pour mettre en œuvre ces étapes. Les objectifs de reprise d'une charge de travail relèvent d'une décision de l'entreprise. Les équipes techniques travaillent ensuite avec les parties prenantes de l'entreprise pour utiliser ces objectifs afin de sélectionner une stratégie de reprise après sinistre.

   > **Note**
   >
   > Pour les étapes 2 et 3, vous pouvez utiliser the section called “Fiche d'implémentation”.

2. Répondez aux questions ci-dessous pour rassembler les informations nécessaires pour prendre une décision.
3. Utilisez-vous des catégories ou des niveaux de criticité pour déterminer l'impact de la charge de travail dans votre organisation ?
   a. Si oui, affectez cette charge de travail à une catégorie.
   b. Dans le cas contraire, définissez ces catégories. Créez cinq catégories ou moins et affinez la plage de vos objectifs de délai et de point de reprise. Exemples de catégories : critique, élevé, moyen, faible. Pour comprendre comment les charges de travail correspondent aux catégories, déterminez si la charge de travail est stratégique, importante pour l'entreprise ou non commerciale.

   c. Définissez le RTO et le RPO de la charge de travail en fonction de sa catégorie. Choisissez toujours une catégorie plus stricte (RTO et RPO inférieurs) que les valeurs brutes calculées au début de cette étape. Si cela entraîne une variation de valeur trop importante, envisagez de créer une autre catégorie.

4. En fonction de ces réponses, attribuez des valeurs de RTO et RPO à la charge de travail. Cela peut se faire directement ou en affectant la charge de travail à un niveau de service prédéfini.
5. Documentez le plan de reprise après sinistre (DRP) pour cette charge de travail, qui fait partie du plan de continuité d'activité (BCP), à un emplacement accessible à l'équipe responsable de la charge de travail et aux parties prenantes.
   a. Enregistrez le RTO et le RPO, ainsi que les informations utilisées pour déterminer ces valeurs. Spécifiez la stratégie utilisée pour évaluer l'impact de la charge de travail sur l'entreprise.
   b. Enregistrez d'autres métriques que le RTO et le RPO que vous suivez ou prévoyez de suivre pour les objectifs de reprise après sinistre.
   c. Vous ajouterez les détails de votre stratégie de reprise après sinistre et de votre runbook à ce plan lorsque vous les créerez.
6. En recherchant la criticité de la charge de travail dans une matrice telle que celle de la figure 15, vous pouvez commencer à établir des niveaux de service prédéfinis pour votre organisation.
7. Après avoir mis en œuvre une stratégie de reprise après sinistre (ou une preuve de concept pour une stratégie de reprise après sinistre) conformément à the section called “REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise”, testez cette stratégie pour déterminer les valeurs RTC (temps de reprise possible) et RPC (point de reprise possible) réelles de la charge de travail. Si ceux-ci n'atteignent pas les objectifs de reprise cibles, vous pouvez soit collaborer avec les parties prenantes de votre entreprise pour les ajuster, soit apporter des modifications à la stratégie de reprise après sinistre, le cas échéant, pour atteindre ces objectifs.

Questions principales

1. Quelle est la durée maximale pendant laquelle la charge de travail peut être interrompue avant qu'un impact grave n'affecte l'entreprise ?
   a. Déterminez le coût (impact financier direct) pour l'entreprise par minute où la charge de travail est interrompue.
   b. Considérez que l'impact n'est pas toujours linéaire. L'impact peut être limité au début, puis augmenter rapidement au-delà d'un point critique dans le temps.

2. Quelle est la quantité maximale de données pouvant être perdues avant qu'un impact grave n'affecte l'entreprise ?
   a. Déterminez cette valeur en fonction de votre magasin de données le plus critique. Identifiez la criticité respective pour les autres magasins de données.
   b. Les données de la charge de travail peuvent-elles être recréées en cas de perte ? Si cette approche est plus facile sur le plan opérationnel que la sauvegarde et la restauration, choisissez le RPO en fonction de la criticité des données sources utilisées pour recréer les données de la charge de travail.
3. Quels sont les objectifs de reprise et les attentes de disponibilité des charges de travail dont celle-ci dépend (en aval) ou des charges de travail qui dépendent de celle-ci (en amont) ?
   a. Choisissez des objectifs de reprise qui permettent à cette charge de travail de répondre aux exigences des dépendances en amont.
   b. Choisissez des objectifs de reprise réalisables compte tenu des capacités de reprise des dépendances en aval. Les dépendances en aval non critiques (celles que vous pouvez « contourner ») peuvent être exclues. Ou, si nécessaire, traitez les dépendances critiques en aval pour améliorer leurs capacités de reprise.

Questions supplémentaires

Envisagez ces questions et dans quelle mesure elles s'appliquent à cette charge de travail :

4. Avez-vous des RTO et des RPO différents selon le type de panne (région ou AZ, etc.) ?
5. Existe-t-il un moment précis (saisonnalité, événements commerciaux, lancements de produits) où votre RTO/RPO peut changer ? Si oui, en quoi diffèrent-ils et quelle est leur limite de temps ?
6. Combien de clients seront touchés si la charge de travail est interrompue ?
7. Quel sera l'impact sur la réputation si la charge de travail est interrompue ?
8. Quels autres impacts opérationnels peuvent entrer en jeu si la charge de travail est interrompue ? Par exemple, la productivité des employés sera affectée si les systèmes de messagerie ne sont pas disponibles ou si les systèmes de paie ne sont pas en mesure de soumettre des transactions.
9. Comment le RTO et le RPO de la charge de travail s'alignent-ils sur la stratégie de reprise après sinistre de la succursale et de l'organisation ?
10. Existe-t-il des obligations contractuelles internes régissant la prestation d'un service ? Des sanctions sont-elles appliquées en cas de non-respect ?
11. Quelles sont les contraintes réglementaires ou de conformité liées aux données ?

Fiche d'implémentation

Vous pouvez utiliser cette feuille de calcul pour les étapes d'implémentation 2 et 3. Vous pouvez l'ajuster en fonction de vos besoins spécifiques, par exemple en ajoutant des questions supplémentaires.

Étape 2 : Questions principales	S'applique à la charge de travail ?	RPO de la charge de travail	RPO de la charge de travail	Ajustement de RTO.	Ajustement de RPO.	Instructions
[1] durée maximale pendant laquelle la charge de travail peut être inactive						mesuré en temps depuis le début de la panne jusqu'à la récupération
[2] quantité maximale de données pouvant être perdues						mesuré dans le temps depuis le dernier jeu de données restaurable
[3a] dépendances en amont						saisissez les objectifs de récupération en amont les plus stricts
[3b] dépendances en aval						saisissez les objectifs de récupération en aval les moins stricts
[3a] dépendances en amont rapprochées						Si la valeur en amont est inférieure aux valeurs actuelles et la valeur en aval supérieure,
[3b] dépendances en aval rapprochées						manipulez les dépendances pour les rapprocher et entrez les valeurs rapprochées ici
[3] dépendances						réduisez les valeurs pour répondre aux dépendances en amont ou augmentez-les selon les fonctionnalités de dépendance en aval
Étape 2 : Questions supplémentaires						Indiquez si la question s'applique. Dans le cas contraire, ignorez-la
RTO/RPO de base						Transférez les valeurs RTO et RPO d'en haut jusqu'ici
[4] type de panne	[ ]O / [ ]N					Indiquez les objectifs de récupération pour les événements avec les exigences les plus strictes
[5] objectifs temporels spécifiques	[ ]O / [ ]N					Indiquez les objectifs de récupération pour les durées avec les exigences les plus strictes
[6] clients perturbés	[ ]O / [ ]N					Représentez graphiquement les clients impactés en fonction du temps d'arrêt ou de la perte de données. Utilisez ces informations pour saisir le RTO et le RPO maximum autorisés en fonction de l'impact sur le client
[7] impact sur la réputation	[ ]O / [ ]N					Déterminez avec l'entreprise le RTO et le RPO maximum en fonction de l'impact sur la réputation
[8] impact opérationnel	[ ]O / [ ]N					Indiquez un RTO et un RPO maximum en fonction de l'impact opérationnel
[9] alignement organisationnel	[ ]O / [ ]N					Indiquez le RTO et le RPO maximum pour les charges de travail de ce type selon les exigences LOB et organisationnelles
[10] obligations contractuelles	[ ]O / [ ]N					Indiquez un RTO et un RPO maximum en fonction des obligations contractuelles
[11] conformité réglementaire	[ ]O / [ ]N					Indiquez le RTO et le RPO maximum en fonction de la conformité réglementaire applicable
cible basée sur des questions supplémentaires						Prenez la valeur minimale (valeur plus stricte) des Q 11-4 et entrez-la ici
cible ajustée						Si les objectifs de la ligne ci-dessus ne peuvent pas être atteints, collaborez avec les parties prenantes pour assouplir les contraintes et entrez un nouveau minimum ici
RTO/RPO ajusté						Indiquez les valeurs RPO/RTO de base, ou la cible ajustée, selon la valeur la plus basse
Étape 3						
Mapper vers une catégorie ou un niveau prédéfini						Ajustez les deux valeurs vers le bas (méthode plus stricte) pour vous aligner sur le niveau défini le plus proche

Fiche

Niveau d'effort du plan d'implémentation : Faible

Ressources

Bonnes pratiques associées :

- the section called “REL09-BP04 Effectuer une récupération périodique des données pour vérifier l'intégrité et les processus de sauvegarde”
- the section called “REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise”
- the section called “REL13-BP03 Effectuer un test de validation de la mise en œuvre de la reprise après sinistre”

Documents connexes :

- Blog d'architecture AWS : série sur la reprise après sinistre
- Reprise après sinistre des charges de travail sur AWS : reprise dans le cloud (livre blanc AWS)
- Gestion des politiques de résilience avec AWS Resilience Hub
- Partenaire APN : partenaires pouvant faciliter la reprise après sinistre
- AWS Marketplace : produits pouvant être utilisés pour la reprise après sinistre

Vidéos connexes :

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- Disaster Recovery of Workloads on AWS

REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise

Définissez une stratégie de reprise après sinistre qui répond aux objectifs de reprise de votre charge de travail. Choisissez une stratégie telle que : sauvegarde et restauration, mode secours (actif/passif) ou actif/actif.

Résultat souhaité : pour chaque charge de travail, il existe une stratégie de reprise après sinistre définie et implémentée qui permet à cette charge de travail d'atteindre les objectifs de reprise. Les stratégies de reprise après sinistre entre les charges de travail utilisent des modèles réutilisables (comme les stratégies décrites précédemment).

Anti-modèles courants :

- Mettre en œuvre des procédures de récupération incohérentes pour les charges de travail avec des objectifs de reprise après sinistre similaires.
- Conserver l'implémentation ad hoc de la stratégie de reprise après sinistre lorsqu'un sinistre se produit.
- Ne pas avoir de plan de reprise après sinistre.
- Être dépendant des opérations du plan de contrôle pendant la récupération.

Avantages liés au respect de cette bonne pratique :

- L'utilisation de stratégies de reprise définies vous permet d'utiliser des outils et des procédures de test courantes.
- L'utilisation de stratégies de reprise définies améliore le partage des connaissances entre les équipes et la mise en œuvre de la reprise après sinistre sur les charges de travail qu'elles possèdent.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé. Sans une stratégie de reprise après sinistre planifiée, mise en œuvre et testée, il est peu probable que vous atteigniez vos objectifs de reprise en cas de sinistre.

Directives d'implémentation

Une stratégie de reprise après sinistre repose sur la capacité à rétablir votre charge de travail sur un site de reprise si votre emplacement principal ne parvient plus à exécuter cette charge de travail. Les objectifs de récupération les plus courants sont le RTO et le RPO, comme indiqué dans REL13-BP01 Définir les objectifs de reprise pour les temps d'arrêt et les pertes de données.

Une stratégie de reprise après sinistre sur plusieurs zones de disponibilité (AZ) au sein d'une seule Région AWS peut vous prémunir contre les événements catastrophiques tels que les incendies, les inondations et les pannes de courant majeures. S'il est nécessaire de mettre en œuvre une protection contre un événement improbable qui empêcherait votre charge de travail de s'exécuter dans une Région AWS donnée, optez pour une stratégie de reprise après sinistre qui utilise plusieurs régions.

Lors de la conception d'une stratégie de reprise après sinistre dans plusieurs régions, vous devez choisir l'une des approches suivantes. Elles sont répertoriées par ordre croissant de coûts et de complexité et par ordre décroissant de RTO et RPO. La région de reprise fait référence à une Région AWS autre que la région principale utilisée pour votre charge de travail.

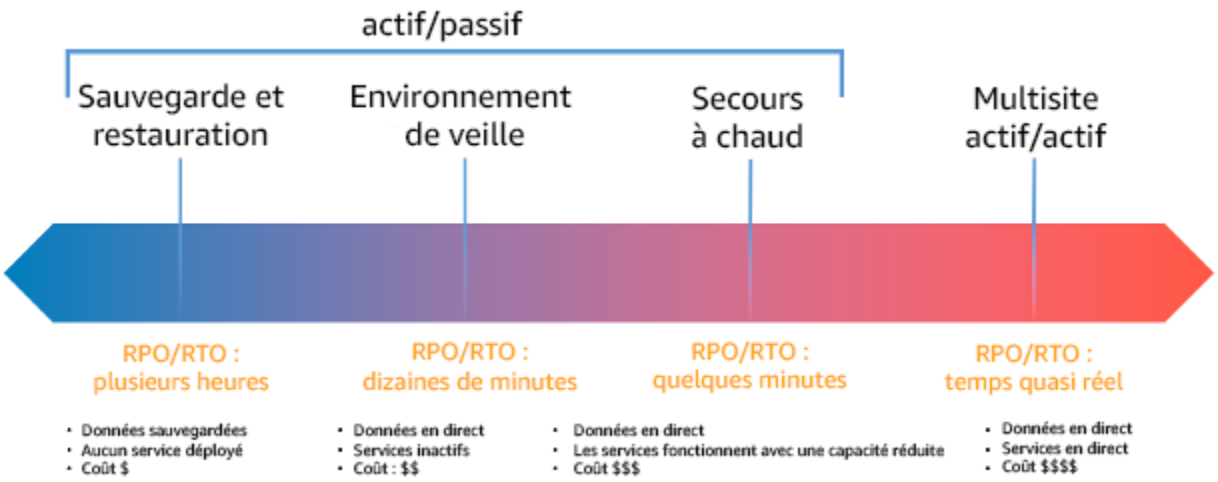


Figure 17 : stratégies de reprise après sinistre

- Sauvegarde et restauration (RPO en heures, RTO de 24 heures maximum) : sauvegardez vos données et applications dans la région de reprise après sinistre. L'utilisation de sauvegardes automatisées ou continues permet une récupération ponctuelle, ce qui peut réduire le RPO à seulement 5 minutes dans certains cas. En cas de sinistre, vous déployez votre infrastructure (en utilisant l'infrastructure en tant que code pour réduire le RTO), déployez votre code et restaurez les données sauvegardées pour vous remettre d'un sinistre dans la région de reprise.
- Environnement en veille (RPO de quelques minutes, RTO de dizaines de minutes) : allouez une copie de votre infrastructure de charge de travail principale dans la région de reprise. Répliquez vos données dans la région de reprise et créez-y des sauvegardes. Les ressources requises pour prendre en charge la réplication et la sauvegarde des données, telles que les bases de données et le stockage d'objets, sont toujours actives. D'autres éléments tels que les serveurs d'applications ou le calcul sans serveur ne sont pas déployés, mais peuvent être créés si nécessaire avec la configuration et le code d'application requis.
- Secours à chaud (RPO de quelques secondes, RTO de quelques minutes) : maintenez une version réduite d'une charge de travail entièrement fonctionnelle qui s'exécute toujours dans la région de reprise. Les systèmes stratégiques sont entièrement dupliqués et sont toujours opérationnels, mais avec une flotte réduite. Les données sont répliquées dans la région de reprise et y sont hébergées. Lorsque vient le moment de la reprise, le système est rapidement mis à l'échelle pour gérer la charge de production. Plus l'échelle du secours à chaud est élevée, plus la dépendance au RTO et au plan de contrôle est faible. Lorsqu'elle est complètement graduée, on parle de zone hébergée.

- Multi-région (multi-site) actif-actif (RPO proche de zéro, RTO potentiellement nul) : votre charge de travail est déployée et dessert activement le trafic à partir de plusieurs Régions AWS. Cette stratégie vous oblige à synchroniser les données entre les régions. Il est important d'éviter ou de gérer les éventuels conflits causés par des écritures sur le même enregistrement dans deux réplicas régionaux différents, ce qui peut être complexe. La réplication des données est utile pour la synchronisation des données et vous protège contre certains types de sinistres. Toutefois, elle ne vous protège pas contre la corruption ou la destruction des données à moins que votre solution n'inclue également des options de récupération ponctuelle.

> **Note**
>
> La différence entre l'environnement en veille et le secours à chaud est parfois difficile à cerner. Ces deux stratégies incluent un environnement dans votre région de reprise avec des copies des ressources de votre région principale. L'environnement en veille diffère en ce qu'il ne peut pas traiter les demandes sans qu'une action supplémentaire soit entreprise au préalable, tandis que le secours à chaud peut gérer le trafic (à des niveaux de capacité réduits) immédiatement. L'environnement en veille vous oblige à allumer des serveurs, à déployer éventuellement une infrastructure supplémentaire (non essentielle) et à augmenter l'échelle, tandis que le secours à chaud nécessite uniquement une augmentation de l'échelle (tout est déjà déployé et en cours d'exécution). Choisissez entre ces options en fonction de vos besoins en termes de RTO et de RPO.
>
> Si le coût est un problème et que vous souhaitez atteindre des objectifs de RPO et RTO similaires à ceux définis dans la stratégie de secours à chaud, vous pouvez envisager des solutions natives du cloud, comme AWS Elastic Disaster Recovery, qui adoptent l'approche de l'environnement de veille et offrent des objectifs de RPO et RTO améliorés.

Étapes d'implémentation

1. Déterminez une stratégie de reprise après sinistre qui répond aux exigences de récupération pour cette charge de travail.

Le choix d'une stratégie de reprise après sinistre vise à trouver un juste milieu entre la réduction des temps d'arrêt et de la perte de données (RTO et RPO) et le coût et la complexité liées à la mise en œuvre de cette stratégie. Évitez de mettre en œuvre une stratégie plus stricte que nécessaire, car cela entraînerait des coûts inutiles.

Par exemple, dans le diagramme suivant, l'entreprise a déterminé son RTO maximal autorisé ainsi que la limite de dépenses possible pour sa stratégie de restauration de service. Compte tenu des objectifs de l'entreprise, les stratégies environnement en veille et secours à chaud satisfont à la fois aux critères de RTO et de coût.

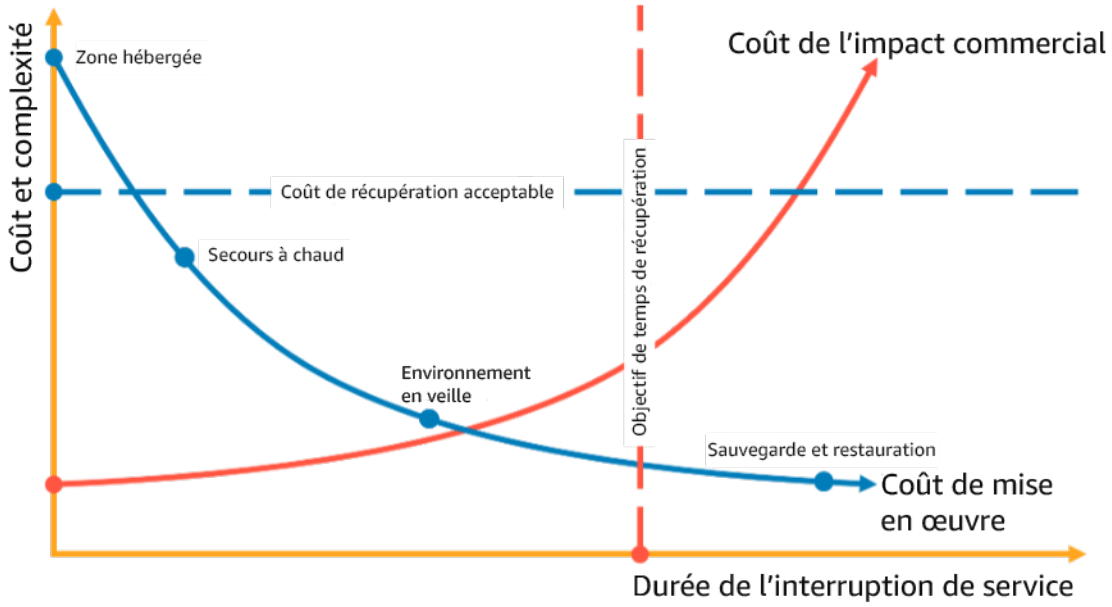


Figure 18 : choix d'une stratégie de reprise après sinistre basée sur le RTO et le coût

Pour en savoir plus, consultez Business Continuity Plan (BCP) [Plan de continuité d'activité (PCA)].

2. Passez en revue les modèles de mise en œuvre de la stratégie de reprise après sinistre sélectionnée.

Cette étape consiste à comprendre comment mettre en œuvre la stratégie sélectionnée. Les stratégies reposent sur l'utilisation de Régions AWS comme site principal et site de reprise. Cependant, vous pouvez également choisir d'utiliser des zones de disponibilité dans une seule région comme stratégie de reprise après sinistre, ce qui permet d'exploiter des éléments de plusieurs de ces stratégies.

Dans les étapes suivantes, vous pouvez appliquer la stratégie à votre charge de travail spécifique.

Sauvegarde et restauration

La sauvegarde et restauration est la stratégie la moins complexe à mettre en œuvre, mais nécessite plus de temps et d'efforts pour la restauration de la charge de travail, ce qui entraîne un RTO et un

RPO plus élevés. Il est conseillé de toujours faire des sauvegardes de vos données et de les copier sur un autre site (comme une autre Région AWS).

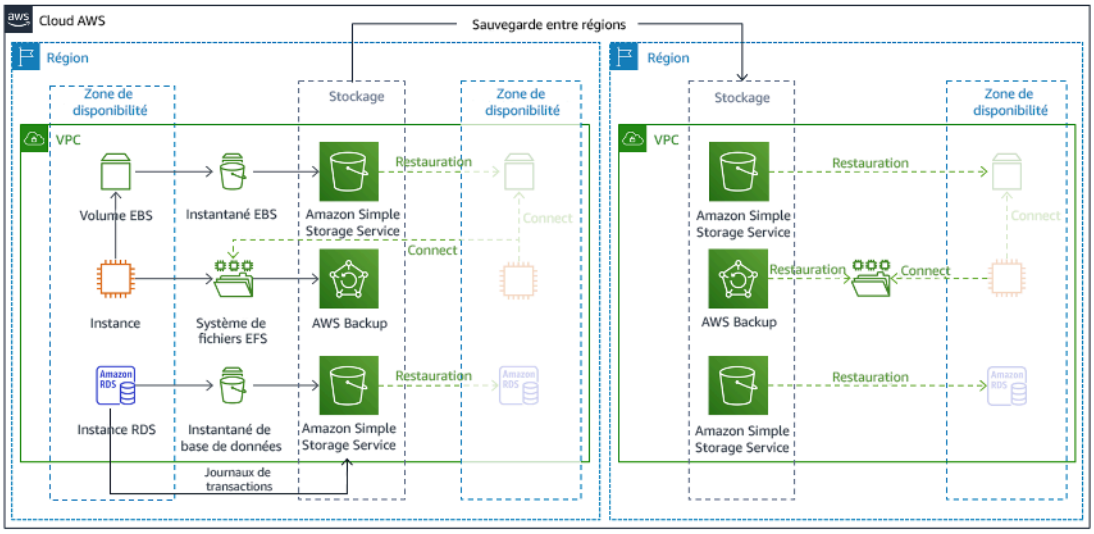


Figure 19 : architecture de sauvegarde et de restauration

Pour obtenir plus de détails sur cette stratégie, consultez Disaster Recovery (DR) Architecture on AWS, Part II: Backup and Restore with Rapid Recovery [Architecture de reprise après sinistre (DR) sur AWS, partie II : sauvegarde et restauration avec récupération rapide].

Environnement en veille

Avec l'approche de l'environnement en veille, vous répliquez vos données depuis la région principale vers la région de reprise. Les ressources principales utilisées pour l'infrastructure de charge de travail sont déployées dans la région de reprise, mais des ressources supplémentaires et toutes les dépendances sont toujours nécessaires pour en faire une pile fonctionnelle. Par exemple, dans la figure 20, aucune instance de calcul n'est déployée.

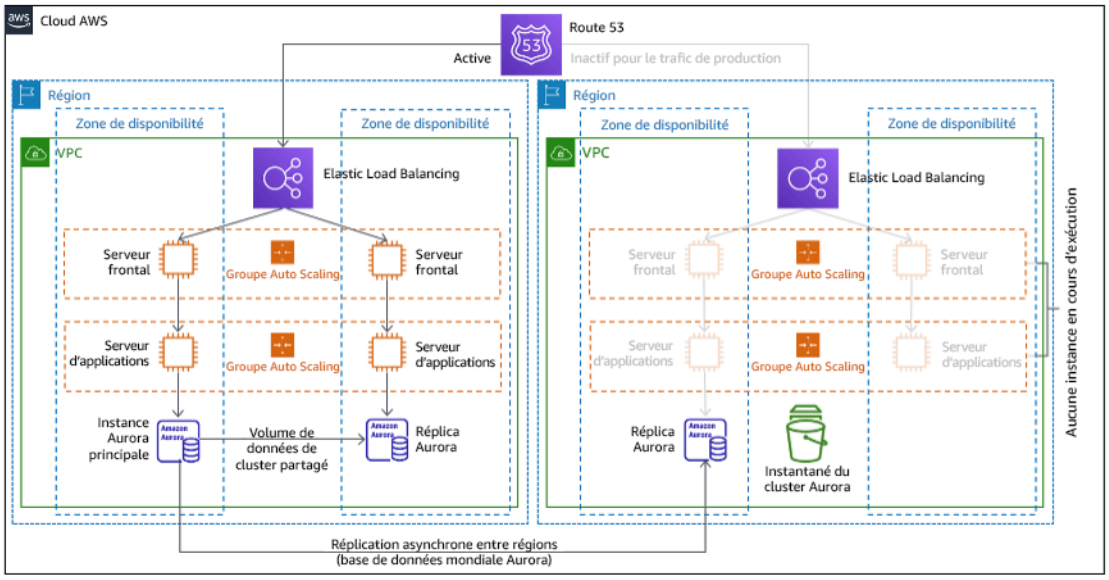


Figure 20 : architecture avec environnement en veille

Pour obtenir plus de détails sur cette stratégie, consultez Disaster Recovery (DR) Architecture on AWS, Part III: Pilot Light and Warm Standby [Architecture de reprise après sinistre (DR) sur AWS, partie III : environnement de veille et secours à chaud].

Secours à chaud

Le secours à chaud consiste à s'assurer qu'il existe une copie réduite, mais entièrement fonctionnelle, de votre environnement de production dans une autre région. Cette approche étend le concept d'environnement en veille et réduit le temps de récupération, car votre charge de travail reste active dans une autre région. Si la région de reprise est déployée à pleine capacité, on parle de zone hébergée.

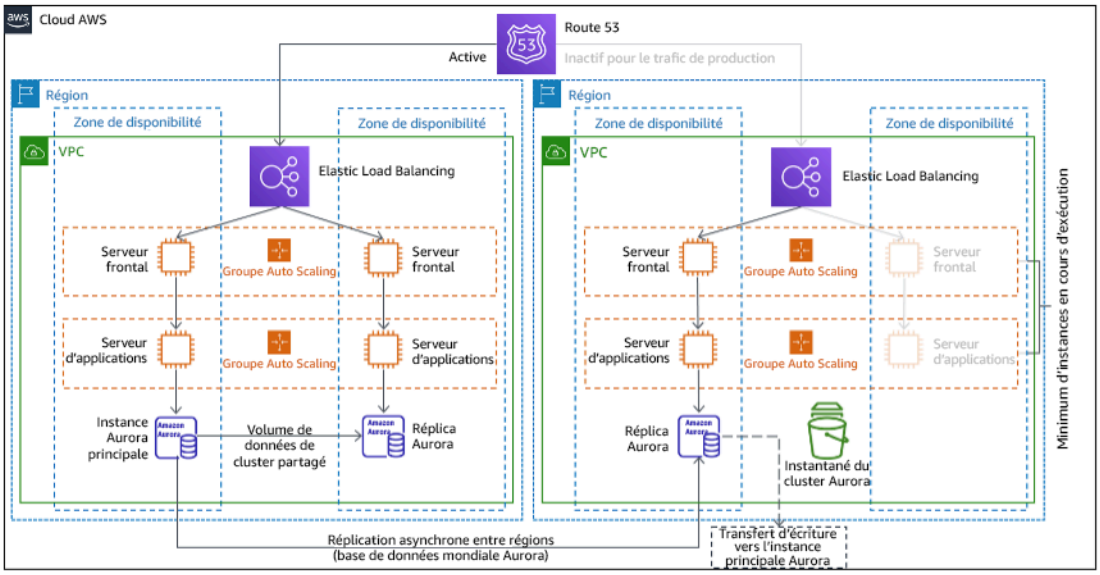


Figure 21 : Architecture de secours à chaud

L'utilisation du secours à chaud ou de l'environnement en veille nécessite une augmentation des ressources dans la région de reprise. Pour vérifier que la capacité est disponible en cas de besoin, envisagez l'utilisation des réserves de capacité pour les instances EC2. Si vous utilisez AWS Lambda, alors la simultanéité allouée peut provisionner des environnements d'exécution afin qu'ils soient prêts à répondre immédiatement aux appels de votre fonction.

Pour obtenir plus de détails sur cette stratégie, consultez Disaster Recovery (DR) Architecture on AWS, Part III: Pilot Light and Warm Standby [Architecture de reprise après sinistre (DR) sur AWS, partie III : environnement de veille et secours à chaud].

Multisite actif/actif

Vous pouvez exécuter votre charge de travail simultanément dans plusieurs régions dans le cadre d'une stratégie multisite actif/actif. Une stratégie multisite actif/actif dessert le trafic de toutes les régions dans lesquelles il est déployé. Les clients peuvent sélectionner cette stratégie pour des raisons autres que la reprise après sinistre. Elle peut être utilisée pour augmenter la disponibilité ou lors du déploiement d'une charge de travail auprès d'une audience mondiale (pour rapprocher le point de terminaison des utilisateurs et/ou déployer des piles localisées pour l'audience de cette région). En tant que stratégie de reprise après sinistre, si la charge de travail ne peut pas être prise en charge dans l'une des Régions AWS vers lesquelles elle est déployée, cette région est évacuée, et les régions restantes sont utilisées pour assurer la disponibilité. La stratégie de reprise après sinistre multisite actif/actif est la plus complexe sur le plan opérationnel et ne doit être sélectionnée que lorsque les besoins de l'entreprise l'exigent.

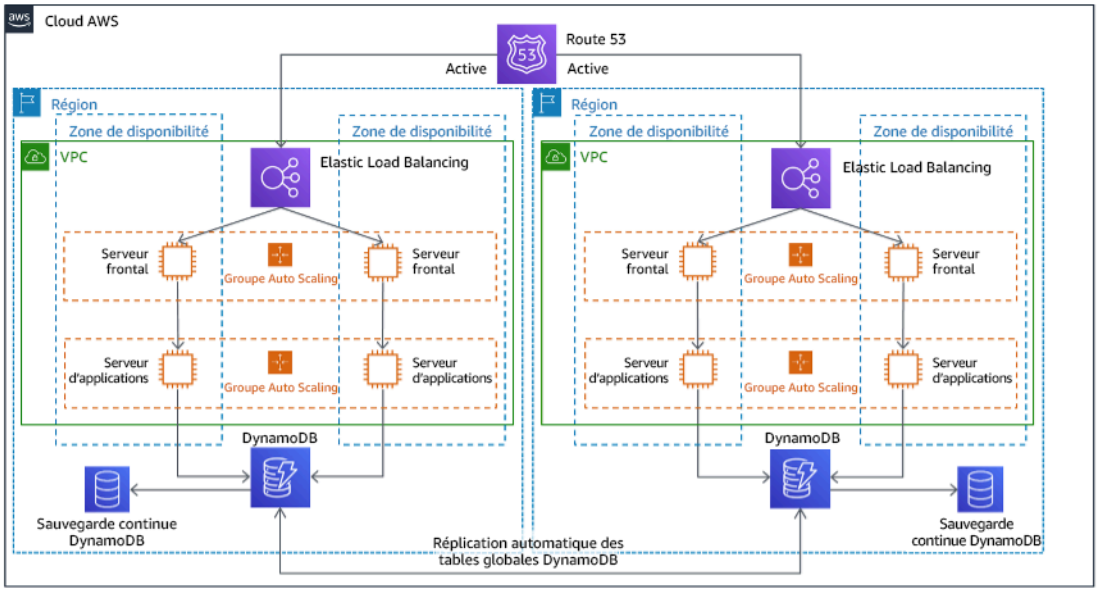


Figure 22 : architecture multisite de type actif/actif

Pour obtenir plus de détails sur cette stratégie, consultez Disaster Recovery (DR) Architecture on AWS, Part IV: Multi-site Active/Active [Architecture de reprise après sinistre (DR) sur AWS, partie IV : multisite actif/actif].

AWS Elastic Disaster Recovery

Si vous envisagez une stratégie d'environnement de veille ou de secours à chaud pour la reprise après sinistre, AWS Elastic Disaster Recovery pourrait constituer une approche alternative offrant de meilleurs avantages.Elastic Disaster Recovery peut offrir un objectif de RPO et de RTO similaire à celui du secours à chaud, tout en conservant l'approche économique de l'environnement de veille. Elastic Disaster Recovery réplique vos données de votre région principale vers votre région de reprise, en utilisant la protection continue des données pour atteindre un RPO mesuré en secondes et un RTO qui peut être mesuré en minutes. Seules les ressources nécessaires à la réplication des données sont déployées dans la région de reprise, ce qui permet de limiter les coûts, à l'instar de la stratégie de l'environnement de veille. En cas d'utilisation de Elastic Disaster Recovery, le service coordonne et orchestre la récupération des ressources informatiques lorsqu'elle est initiée dans le cadre d'un basculement ou d'une opération.

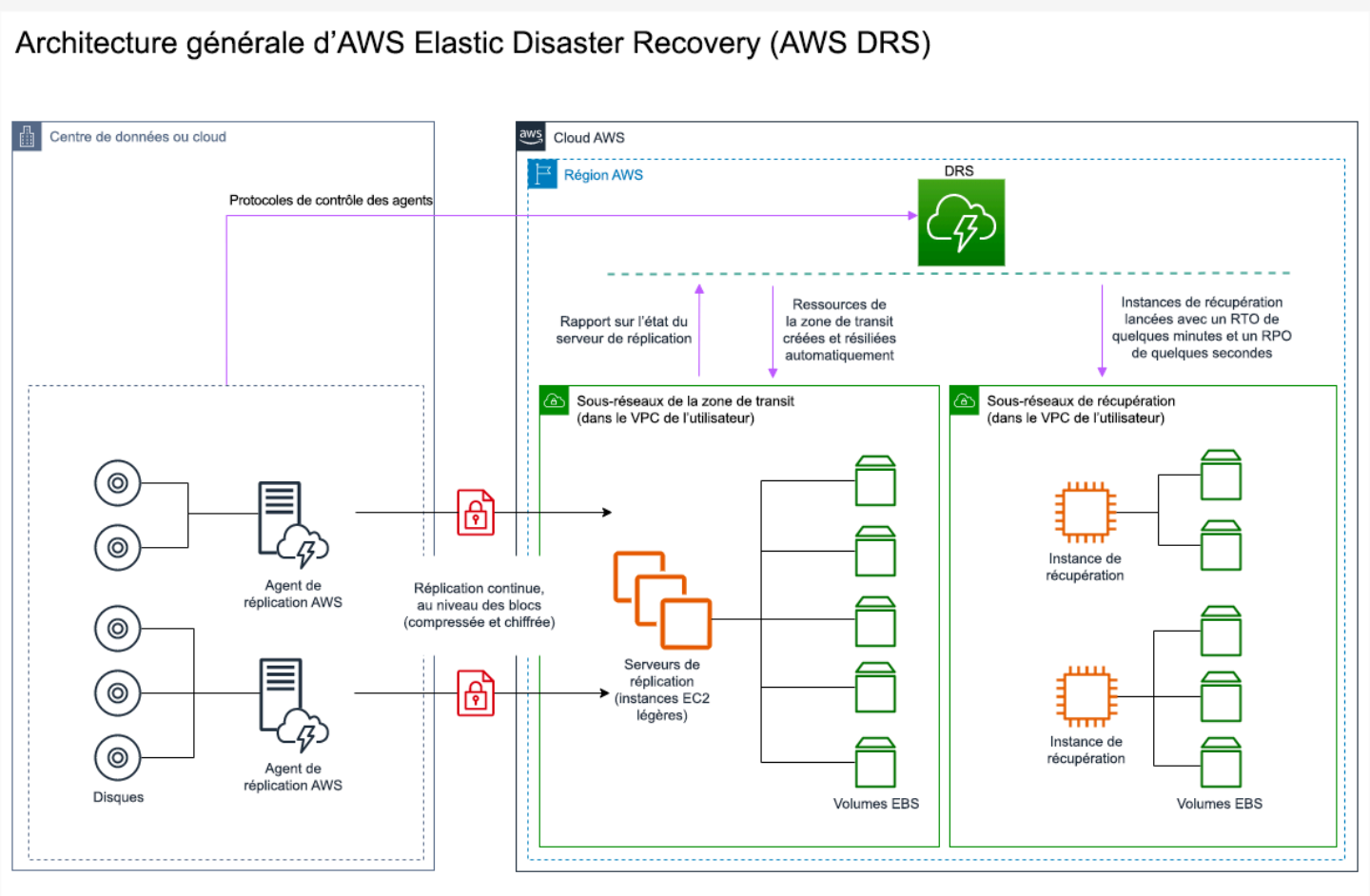


Figure 23 : architecture AWS Elastic Disaster Recovery

Pratiques supplémentaires de protection des données

Avec toutes les stratégies, vous devez également vous prémunir contre les catastrophes liées aux données La réplication continue des données vous protège contre certains types de sinistres, mais ne vous protège pas toujours contre la corruption ou la destruction des données, à moins que votre stratégie n'inclue également la gestion des versions des données stockées ou des options de récupération ponctuelle. Vous devez également sauvegarder les données répliquées sur le site de reprise pour créer des sauvegardes ponctuelles en plus des réplicas.

Utilisation de plusieurs zones de disponibilité (AZ) dans une seule Région AWS

Lorsque vous utilisez plusieurs AZ dans une même région, l'implémentation de la reprise après sinistre exploite plusieurs éléments des stratégies ci-dessus. Vous devez d'abord créer une architecture haute disponibilité (HA), en utilisant plusieurs AZ, comme illustré à la figure 23. Cette

architecture utilise une approche multisite actif/actif, car les instances Amazon EC2 et l' Elastic Load Balancer disposent de ressources déployées dans plusieurs zones de disponibilité, qui gèrent activement les requêtes. L'architecture présente également un système de zone hébergée qui permet, en cas de panne de l'instance principale Amazon RDS (ou de la zone de disponibilité elle-même), de faire passer l'instance de secours au rang d'instance principale.

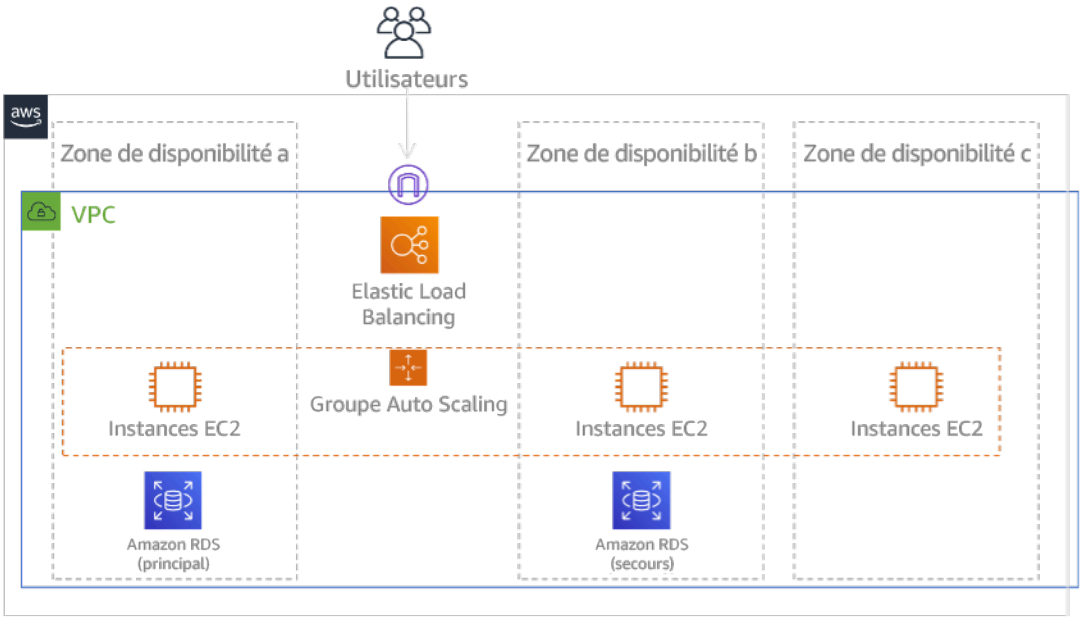


Figure 24 : architecture de multi-AZ

En plus de cette architecture haute disponibilité, vous devez ajouter des sauvegardes de toutes les données requises pour exécuter votre charge de travail. Ceci est particulièrement important pour les données limitées à une seule zone, telles que les Amazon EBS volumes ou les Amazon Redshift clusters. Si une zone de disponibilité tombe en panne, vous devrez restaurer ces données dans une autre zone de disponibilité. Dans la mesure du possible, vous devez également copier les sauvegardes de données dans une autre Région AWS comme couche de protection supplémentaire.

Une approche alternative moins courante de la reprise après sinistre à région unique et multi-AZ est illustrée dans l'article de blog, Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 1: Single-Region stack (Création d'applications hautement résilientes à l'aide du contrôleur de récupération d'application d'Amazon Route 53, partie 1 : pile à région unique). Dans ce cas, la stratégie consiste à maintenir autant que possible l'isolement entre les zones de disponibilité, à l'instar du fonctionnement des régions. Avec cette stratégie alternative, vous pouvez choisir une approche active/active ou active/passive.

> **Note**
>
> Certaines charges de travail sont soumises à des exigences réglementaires en matière de situation géographique des données. Si cela s'applique à votre charge de travail dans une localité qui n'a actuellement qu'une seule Région AWS, plusieurs régions ne répondront pas aux besoins de votre entreprise. Les stratégies multi-AZ assurent une bonne protection contre la plupart des catastrophes.

3. Évaluez les ressources de votre charge de travail et déterminez quelle sera leur configuration dans la région de reprise avant le basculement (pendant le fonctionnement normal).

Pour l'infrastructure et les ressources AWS, utilisez l'infrastructure en tant que code telle que AWS CloudFormation ou des outils tiers comme Hashicorp Terraform. Pour un déploiement sur plusieurs comptes et régions en une seule opération, vous pouvez utiliser AWS CloudFormation StackSets. Pour les stratégies « Multisite actif/actif » et « Zone hébergée », l'infrastructure déployée dans la région de reprise dispose des mêmes ressources que la région principale. Pour les stratégies « Environnement en veille » et « Secours à chaud », l'infrastructure déployée nécessitera des actions supplémentaires pour être prête pour la production. En utilisant les paramètres de CloudFormation et la logique conditionnelle , vous pouvez contrôler si une pile déployée est active ou en veille avec un seul modèle. En utilisant Elastic Disaster Recovery, le service répliquera et orchestrera la restauration des configurations d'applications et des ressources informatiques.

Toutes les stratégies de reprise après sinistre exigent que les sources de données soient sauvegardées dans la Région AWS, puis que ces sauvegardes soient copiées dans la région de reprise. AWS Backup fournit une vue centralisée où vous pouvez configurer, planifier et surveiller les sauvegardes de ces ressources. Pour les stratégies « Environnement en veille », « Secours à chaud » et « Multisite actif/actif », vous devez également répliquer les données de la région principale vers les ressources de données de la région de reprise, telles que des instances de base de données Amazon Relational Database Service (Amazon RDS) ou des tables Amazon DynamoDB. Ces ressources de données sont donc actives et prêtes à répondre aux demandes dans la région de reprise.

Pour en savoir plus sur comment fonctionnent les services AWS dans les régions, consultez cette série de blogs intitulée Creating a Multi-Region Application with AWS Services (Création d'une application multirégion avec les services AWS).

4. Déterminez et mettez en œuvre la manière dont vous préparerez votre région de reprise pour le basculement en cas de besoin (lors d'un sinistre).

Pour la stratégie multisite actif/actif, le basculement consiste à évacuer une région et à s'appuyer sur les régions actives restantes. En général, ces régions sont prêtes à accepter du trafic. Pour les stratégies Environnement en veille et Secours à chaud, vos actions de reprise devront déployer les ressources manquantes, telles que les instances EC2 de la figure 20, ainsi que toute autre ressource manquante.

Pour toutes les stratégies ci-dessus, vous devrez peut-être promouvoir les instances en lecture seule des bases de données au rang d'instances principales en lecture/écriture.

Pour la sauvegarde et la restauration, la restauration des données à partir de la sauvegarde crée des ressources pour ces données, telles que des volumes EBS, des instances de base de données RDS et des tables DynamoDB. Vous devez également restaurer l'infrastructure et déployer le code. Vous pouvez utiliser AWS Backup pour restaurer les données dans la région de reprise. Consulter REL09-BP01 Identifier et sauvegarder toutes les données qui doivent être sauvegardées, ou reproduire les données à partir de sources pour en savoir plus. La reconstruction de l'infrastructure comprend la création de ressources telles que les instances EC2, en plus des Amazon Virtual Private Cloud(Amazon VPC), sous-réseaux et groupes de sécurité nécessaires. Vous pouvez automatiser une grande partie du processus de restauration. Pour savoir comment procéder, consultez cet article de blog.

5. Déterminez et mettez en œuvre la manière dont vous redirigerez le trafic vers le basculement en cas de besoin (lors d'un sinistre).

Cette opération de basculement peut être lancée automatiquement ou manuellement. Le basculement lancé automatiquement sur la base de vérifications de l'état ou d'alarmes doit être utilisé avec prudence, car un basculement inutile (fausse alerte) entraînerait des coûts tels que l'indisponibilité et la perte de données. Le basculement manuel est donc souvent utilisé. Dans ce cas, nous vous conseillons tout de même d'automatiser les étapes de basculement, de sorte que vous n'ayez à appuyer que sur un bouton pour lancer le basculement.

Il existe plusieurs options de gestion du trafic à prendre en compte lors de l'utilisation des services AWS. Une option consiste à utiliser Amazon Route 53. Avec Amazon Route 53, vous pouvez associer plusieurs points de terminaison IP dans une ou plusieurs Régions AWS avec un nom de domaine Route 53. Pour mettre en œuvre un basculement manuel, vous pouvez utiliser le Contrôleur de récupération d'application Amazon Route 53, qui fournit une API de plan de données

hautement disponible pour réacheminer le trafic vers la région de reprise. Lors de la mise en œuvre du basculement, utilisez les opérations du plan de données et évitez celles du plan de contrôle, comme décrit dans REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle pendant la récupération. »

Pour en savoir plus sur cette option et d'autres, consultez cette section du livre blanc sur la reprise après sinistre.

6. Élaborez un plan pour déterminer la façon dont votre charge de travail se rétablira.

La restauration consiste à renvoyer l'exploitation de la charge de travail à la région principale, après qu'un événement de sinistre s'est atténué. La mise en service de l'infrastructure et du code dans la région principale suit généralement les mêmes étapes que celles utilisées initialement. Elle s'appuie notamment sur l'infrastructure en tant que code et les pipelines de déploiement de code. Le défi posé par la restauration consiste à restaurer les magasins de données et à garantir leur cohérence avec la région de reprise en cours d'exécution.

Lors de l'état de basculement, les bases de données de la région de reprise sont actives et disposent des données à jour. L'objectif est alors de resynchroniser les données de la région de reprise vers la région principale, en s'assurant qu'elle est à jour.

Certains services AWS effectuent cette opération automatiquement. Si vous utilisiez les tables globales Amazon DynamoDB, même si la table de la région principale devenait indisponible, DynamoDB reprendrait la propagation de toutes les écritures en attente lorsqu'elle se reconnecterait. Si vous utilisez Amazon Aurora Global Database et un basculement planifié géré, la topologie de réplication existante de la base de données globale Aurora est maintenue. Par conséquent, l'ancienne instance en lecture/écriture de la région principale deviendra un réplica et recevra les mises à jour de la région de reprise.

Dans les cas où cela n'est pas automatique, vous devrez rétablir la base de données dans la région principale en tant que réplica de la base de données dans la région de reprise. Dans de nombreux cas, cela implique la suppression de l'ancienne base de données principale et la création de nouveaux réplicas. Par exemple, pour obtenir des instructions sur la marche à suivre avec Amazon Aurora Global Database, en supposant un basculement non planifié, consultez cet atelier : Fail Back a Global Database (Failback d'une base de données globale).

Après un basculement, si vous pouvez poursuivre l'exécution dans la région de reprise, envisagez d'en faire la nouvelle région principale. Vous devriez alors suivre toutes les étapes ci-dessus pour convertir l'ancienne région principale en région de reprise. Certaines organisations effectuent une

rotation planifiée, en échangeant périodiquement leurs régions principale et de reprise (par exemple tous les trois mois).

Toutes les étapes nécessaires au basculement et au rétablissement doivent être conservées dans un playbook accessible à tous les membres de l'équipe et révisé périodiquement.

En utilisant Elastic Disaster Recovery, le service aidera à orchestrer et à automatiser le processus de failback. Pour obtenir plus de détails, consultez Effectuer un failback.

Niveau d'effort du plan d'implémentation : élevé

Ressources

Bonnes pratiques associées :

- the section called “REL09-BP01 Identifier et sauvegarder toutes les données qui doivent être sauvegardées, ou reproduire les données à partir de sources”
- the section called “REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle pendant la récupération”
- the section called “REL13-BP01 Définir les objectifs de reprise pour les temps d'arrêt et les pertes de données”

Documents connexes :

- Blog d'architecture AWS : série sur la reprise après sinistre
- Reprise après sinistre des charges de travail sur AWS : reprise dans le cloud (livre blanc AWS)
- Options de reprise après sinistre dans le cloud
- Créer une solution backend active-active sans serveur sur plusieurs régions en une heure
- Backend sans serveur sur plusieurs régions - rechargé
- RDS : réplication d'un réplica en lecture entre les régions
- Route 53 : configuration du basculement DNS
- S3 : réplication entre régions
- Qu'est-ce que AWS Backup ?
- What is Route 53 Application Recovery Controller? (Qu'est-ce que le contrôleur de récupération d'application d'Amazon Route 53 ?)

- AWS Elastic Disaster Recovery
- HashiCorp Terraform : Premiers pas - AWS
- Partenaire APN : partenaires pouvant faciliter la reprise après sinistre
- AWS Marketplace : produits pouvant être utilisés pour la reprise après sinistre

Vidéos connexes :

- Disaster Recovery of Workloads on AWS (Reprise après sinistre des charges de travail sur AWS)
- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- Démarrer avec AWS Elastic Disaster Recovery | Amazon Web Services

Exemples connexes :

- Atelier Well-Architected - Reprise après sinistre - Série d'ateliers illustrant les stratégies de reprise après sinistre

REL13-BP03 Effectuer un test de validation de la mise en œuvre de la reprise après sinistre

Testez régulièrement le basculement vers votre site de reprise pour vérifier qu'il fonctionne correctement et que les RTO et RPO sont respectés.

Anti-modèles courants :

- Ne jamais exécuter de basculements en production.

Avantages liés au respect de cette bonne pratique : en testant régulièrement votre plan de reprise après sinistre, vous vérifiez qu'il fonctionnera en cas de besoin et que votre équipe sait comment exécuter la stratégie.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

S'il y a bien un modèle à éviter, c'est celui qui consiste à développer des chemins de récupération rarement testés. Par exemple, vous pouvez avoir un magasin de données secondaire qui est utilisé pour les requêtes en lecture seule. Lorsque vous écrivez dans un magasin de données et que

l'instance principale connaît une défaillance, vous pouvez basculer vers le magasin de données secondaire. Si vous ne testez pas fréquemment ce basculement, vous constaterez peut-être que vos hypothèses sur les capacités du magasin de données secondaire sont incorrectes. La capacité du magasin de données secondaire, qui peut avoir été suffisante lors de votre dernier test, peut ne plus être en mesure de tolérer la charge dans le cadre de ce scénario. Notre expérience nous a montré que seul un chemin de récupération après erreur testé fréquemment fonctionne réellement. C'est pourquoi l'idéal est de n'avoir qu'un petit nombre de chemins de récupération. Vous pouvez établir des modèles de reprise et tester ceux-ci régulièrement. Si vous avez un chemin de récupération complexe ou critique, vous devez toujours exécuter régulièrement cette panne en production pour vous assurer du bon fonctionnement de ce chemin de récupération. Dans l'exemple que nous venons de présenter, vous devez procéder régulièrement au basculement vers l'instance de secours, quel que soit le besoin.

Étapes d'implémentation

1. Préparez vos charges de travail pour la reprise. Testez régulièrement vos chemins de récupération. L'informatique orientée récupération identifie les caractéristiques des systèmes qui améliorent la récupération : isolement et redondance, capacité de l'ensemble du système à réduire les modifications, capacité à surveiller et déterminer l'état de santé, capacité à fournir des diagnostics, reprise automatique, conception modulaire et capacité à redémarrer. Entraînez votre chemin de reprise pour vérifier qu'il peut s'effectuer au moment et à l'état spécifiés. Utilisez vos runbooks au cours de cette reprise pour documenter les problèmes et trouver des solutions pour les résoudre avant le prochain test.
2. Pour les charges de travail basées sur Amazon EC2, utilisez AWS Elastic Disaster Recovery pour mettre en œuvre et lancer des instances d'opérations dans le cadre de votre stratégie de reprise après sinistre. AWS Elastic Disaster Recovery permet d'exécuter efficacement des opérations, ce qui vous aide à vous préparer à un basculement. Vous pouvez également lancer fréquemment vos instances en utilisant Elastic Disaster Recovery à des fins de test et d'opération sans rediriger le trafic.

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant faciliter la reprise après sinistre
- Blog d'architecture AWS : série sur la reprise après sinistre
- AWS Marketplace : produits pouvant être utilisés pour la reprise après sinistre

- AWS Elastic Disaster Recovery
- Reprise après sinistre des charges de travail sur AWS : reprise dans le cloud (livre blanc AWS)
- AWS Elastic Disaster Recovery Preparing for Failover (Préparation au basculement : préparation au basculement)
- Projet informatique orientée reprise Berkeley/Stanford
- Qu'est qu'AWS Fault Injection Simulator (AWS FIS) ?

Vidéos connexes :

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications
- AWS re:Invent 2019: Backup-and-restore and disaster-recovery solutions with AWS

Exemples connexes :

- Atelier Well-Architected : tester la résilience

REL13-BP04 Gérer l'écart de configuration au niveau du site ou de la région de reprise après sinistre

Assurez-vous que l'infrastructure, les données et la configuration sont conformes aux besoins du site ou de la région de reprise après sinistre. Par exemple, vérifiez que les AMI et les quotas de service sont à jour.

AWS Config surveille et enregistre en permanence les configurations de vos ressources AWS. Il peut détecter tout écart et déclencher AWS Systems Manager Automation pour le corriger et activer les alarmes. AWS CloudFormation peut également détecter tout écart dans les piles que vous avez déployées.

Anti-modèles courants :

- Ne pas effectuer les mises à jour dans vos emplacements de récupération, lorsque vous apportez des modifications à la configuration ou à l'infrastructure sur les emplacements principaux.
- Ne pas prendre en compte des limitations potentielles (comme les différences de service) sur le site principal et le site de reprise.

Avantages liés au respect de cette bonne pratique : Pour une reprise complète, veillez à ce que votre environnement de reprise après sinistre soit cohérent avec votre environnement existant.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Assurez-vous que vos pipelines de diffusion assurent effectivement cette diffusion au niveau de votre site principal ainsi qu'au niveau de vos sites de sauvegarde. Les pipelines de diffusion pour le déploiement d'applications en production doivent être distribués à tous les emplacements spécifiés de la stratégie de DR, y compris les environnements de développement et de test.
- Activez AWS Config pour suivre les écart potentiels au niveau des emplacements. Utilisez les règles AWS Config pour créer des systèmes qui appliquent vos stratégies de reprise après sinistre et génèrent des alertes lorsqu'elles détectent un écart.
  - Correction des ressources AWS non conformes à l'aide des règles AWS Config Rules
  - AWS Systems Manager Automation
- Utilisez AWS CloudFormation pour déployer votre infrastructure. AWS CloudFormation peut détecter l'écart entre ce que vos modèles CloudFormation spécifient et ce qui est réellement déployé.
  - AWS CloudFormation : détection de tout écart à l'échelle d'une pile CloudFormation

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant faciliter la reprise après sinistre
- Blog d'architecture AWS : série sur la reprise après sinistre
- AWS CloudFormation : détection de tout écart à l'échelle d'une pile CloudFormation
- AWS Marketplace : produits pouvant être utilisés pour la reprise après sinistre
- AWS Systems Manager Automation
- Reprise après sinistre des charges de travail sur AWS : reprise dans le cloud (livre blanc AWS)
- Comment mettre en œuvre une solution de gestion de configuration d'infrastructure sur AWS ?
- Correction des ressources AWS non conformes à l'aide des règles AWS Config Rules

Vidéos connexes :

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)

REL13-BP05 Automatiser la reprise

Utilisez AWS ou des outils tiers pour automatiser la reprise du système et acheminer le trafic vers le site ou la région de reprise après sinistre.

En fonction des vérifications de l'état configurées, les services AWS tels qu'Elastic Load Balancing et AWS Auto Scaling peuvent répartir la charge vers des zones de disponibilité saines, tandis que les services tels qu'AWS et Global Accelerator peuvent acheminer la charge vers des Régions AWS saines. Amazon Route 53 Application Recovery Controller vous aide à gérer et à coordonner le basculement à l'aide de vérifications de l'état de préparation et fonctionnalités de contrôle du routage. Ces fonctionnalités surveillent en permanence la capacité de votre application à se rétablir après une défaillance et vous permettent de contrôler la reprise de votre application dans plusieurs Régions AWS, zones de disponibilité et sur site.

Pour les charges de travail sur des centres de données physiques ou virtuels existants ou des clouds privés, AWS Elastic Disaster Recoverydisponible via AWS Marketplace, permet aux organisations de configurer une stratégie de reprise après sinistre automatisée pour AWS. CloudEndure prend également en charge la reprise après sinistre entre régions et zones de disponibilité dans AWS.

Anti-modèles courants :

- La mise en œuvre d'un système de basculement et de restauration automatisés identique peut entraîner une oscillation de chemin lorsqu'une défaillance se produit.

Avantages liés au respect de cette bonne pratique : La reprise automatique réduit le temps de reprise en éliminant les risques d'erreurs manuelles.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

- Automatisez les chemins de récupération. Pour les temps de reprise courts, le jugement et les actions de l'humain ne peuvent pas être utilisés pour des scénarios à haute disponibilité. Le système doit absolument reprendre automatiquement, quelle que soit la situation.
  - Utilisez CloudEndure Disaster Recovery pour un basculement et une restauration automatisés. CloudEndure Disaster Recovery réplique en continu vos machines (notamment le système d'exploitation, la configuration d'état du système, les bases de données, les applications et les fichiers) dans une zone intermédiaire économique de votre Compte AWS cible et de votre région préférée. En cas de sinistre, vous pouvez demander à CloudEndure Disaster Recovery de lancer

automatiquement des milliers de vos machines dans leur état entièrement mis en service en quelques minutes.

- Exécution d'un basculement et d'une reprise après sinistre
- CloudEndure Disaster Recovery

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant faciliter la reprise après sinistre
- Blog d'architecture AWS : série sur la reprise après sinistre
- AWS Marketplace : produits pouvant être utilisés pour la reprise après sinistre
- AWS Systems Manager Automation
- CloudEndure Disaster Recovery vers AWS
- Reprise après sinistre des charges de travail sur AWS : reprise dans le cloud (livre blanc AWS)

Vidéos connexes :

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)

Efficacité en matière de performance

Le pilier Efficacité des performances comprend la capacité à utiliser efficacement les ressources de calcul pour répondre aux exigences du système et à maintenir cette efficacité à mesure que la demande change et les technologies évoluent. Vous trouverez des recommandations sur l'implémentation dans le Livre blanc du pilier Efficacité des performances.

Domaines de bonnes pratiques

- Choix d'architecture
- Informatique et matériel
- Gestion des données
- Mise en réseau et diffusion de contenu
- Processus et culture

Choix d'architecture

Questions

- PERF 1. Comment sélectionnez-vous les ressources et l'architecture cloud adaptées à votre charge de travail ?

PERF 1. Comment sélectionnez-vous les ressources et l'architecture cloud adaptées à votre charge de travail ?

La solution optimale pour une charge de travail peut varier, et les solutions combinent souvent plusieurs approches. Les charges de travail bien architecturées utilisent plusieurs solutions et permettent d'exploiter différentes fonctionnalités pour améliorer les performances.

Bonnes pratiques

- PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles
- PERF01-BP02 Utiliser les recommandations de votre fournisseur de cloud ou d'un partenaire approprié pour en savoir plus sur les modèles d'architecture et les bonnes pratiques
- PERF01-BP03 Tenir compte des coûts dans vos décisions architecturales
- PERF01-BP04 Évaluer l'impact des compromis sur les clients et l'efficacité de l'architecture
- PERF01-BP05 Utiliser des stratégies et des architectures de référence
- PERF01-BP06 Utiliser le benchmarking pour éclairer vos décisions architecturales
- PERF01-BP07 Utiliser une approche orientée données pour les choix architecturaux

PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles

Découvrez en continu les services et configurations disponibles qui vous aident à prendre de meilleures décisions architecturales et à améliorer l'efficacité des performances de votre architecture de charge de travail.

Anti-modèles courants :

- Vous utilisez le cloud comme centre de données hébergé.
- Vous ne modernisez pas votre application après la migration vers le cloud.
- Vous n'utilisez qu'un seul type de stockage pour tout ce que vous devez conserver.

- Vous utilisez les types d'instances qui correspondent le plus à vos standards actuels. Elles peuvent être de plus grande taille au besoin.
- Vous déployez et gérez les technologies disponibles en tant que services gérés.

Avantages liés au respect de cette bonne pratique : En envisageant de nouveaux services et de nouvelles configurations, vous pourriez être en mesure d'améliorer considérablement vos performances, de réduire les coûts et d'optimiser les efforts requis pour maintenir votre charge de travail. Elle peut également vous aider à accélérer le délai de valorisation des produits compatibles avec le cloud.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

AWS publie en permanence de nouveaux services et fonctionnalités susceptibles d'améliorer les performances et de réduire le coût des charges de travail dans le cloud. Il est essentiel de se tenir informé de ces nouveaux services et fonctionnalités pour maintenir l'efficacité des performances dans le cloud. La modernisation de votre architecture de charge de travail vous permet également d'accélérer la productivité, de stimuler l'innovation et de générer de nouvelles opportunités de croissance.

Étapes d'implémentation

- Faites l'inventaire de vos charges de travail logicielles et de l'architecture des services connexes. Déterminez la catégorie de produits sur laquelle vous souhaitez en savoir plus.
- Explorez les offres AWS pour identifier et découvrir les services et les options de configuration pertinents qui peuvent vous aider à améliorer les performances et à réduire les coûts et la complexité opérationnelle.
  - Nouveautés avec AWS
  - Blog AWS
  - AWS Skill Builder
  - Événements et webinaires AWS
  - AWS Training et certifications
  - Chaîne YouTube AWS
  - Ateliers AWS
  - Communautés AWS

- Utilisez des environnements de test (hors production) pour découvrir et tester de nouveaux services sans frais supplémentaires.

Ressources

Documents connexes :

- Centre d'architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- Créer des applications modernes sur AWS

Vidéos connexes :

- Voici mon architecture

Exemples connexes :

- Exemples AWS
- Exemples de kits SDK AWS

PERF01-BP02 Utiliser les recommandations de votre fournisseur de cloud ou d'un partenaire approprié pour en savoir plus sur les modèles d'architecture et les bonnes pratiques

Utilisez les ressources cloud de l'entreprise, telles que la documentation, les architectes de solutions, les services professionnels ou les partenaires appropriés pour éclairer vos décisions architecturales. Ces ressources vous aident à vérifier et à améliorer votre architecture pour obtenir des performances optimales.

Anti-modèles courants :

- Vous utilisez AWS en tant que fournisseur de cloud ordinaire.
- Vous utilisez les services AWS de manière non conforme à leur utilisation prévue.
- Vous suivez toutes les recommandations sans tenir compte du contexte de votre entreprise.

Avantages liés au respect de cette bonne pratique : En suivant les recommandations d'un fournisseur de cloud ou d'un partenaire approprié, vous pouvez faire les bons choix architecturaux pour votre charge de travail et vous avez confiance dans vos décisions.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

AWS propose un large éventail de recommandations, documentations et ressources qui peuvent vous aider à générer et à gérer des charges de travail cloud efficaces. La documentation AWS fournit des exemples de code, des tutoriels et des explications détaillées sur les services. Outre la documentation, AWS propose des programmes de formation et de certification, des architectes de solutions et des services professionnels qui peuvent aider les clients à explorer différents aspects des services cloud et à mettre en œuvre une architecture cloud efficace sur AWS.

Tirez parti de ces ressources pour obtenir des informations précieuses et des bonnes pratiques, gagner du temps et obtenir de meilleurs résultats dans le AWS Cloud.

Étapes d'implémentation

- Consultez la documentation et les recommandations AWS et suivez les bonnes pratiques. Ces ressources peuvent vous aider à choisir et à configurer efficacement les services, ainsi qu'à améliorer les performances.
  - documentation AWS (comme les guides d'utilisation et les livres blancs)
  - Blog AWS
  - AWS Training et certifications
  - Chaîne YouTube AWS
- Participez à des événements partenaires AWS (tels que les sommets mondiaux AWS, les groupes d'utilisateurs, re:Invent AWS et les ateliers) pour découvrir les bonnes pratiques d'utilisation des services AWS auprès des experts AWS.
  - Événements et webinaires AWS
  - Ateliers AWS
  - Communautés AWS
- Contactez AWS pour obtenir de l'aide lorsque vous avez besoin de conseils ou d'informations supplémentaires sur le produit. Les architectes de solutions AWS et les services professionnels AWS fournissent des conseils pour la mise en œuvre de solutions. les partenaires AWS apportent

une expertise AWS pour vous aider à gagner en agilité et favoriser l'innovation au sein de votre entreprise.

- Utilisez AWS Support si vous avez besoin d'une assistance technique pour utiliser un service de manière efficace. Nos plans de support sont conçus pour vous fournir la bonne combinaison d'outils et l'accès à une expertise afin que vous puissiez réussir avec AWS tout en optimisant les performances, en gérant les risques et en maîtrisant les coûts.

Ressources

Documents connexes :

- Centre d'architecture AWS
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- Support aux entreprises AWS

Vidéos connexes :

- Voici mon architecture

Exemples connexes :

- Exemples AWS
- Exemples de kits SDK AWS

PERF01-BP03 Tenir compte des coûts dans vos décisions architecturales

Tenez compte des coûts dans vos décisions architecturales afin d'améliorer l'utilisation des ressources et l'efficacité des performances de votre charge de travail cloud. Lorsque vous êtes conscient des implications financières de votre charge de travail cloud, vous êtes plus susceptible de tirer parti de ressources efficaces et de réduire les pratiques inutiles.

Anti-modèles courants :

- Vous n'utilisez qu'une seule famille d'instances.
- Vous n'évaluez pas les solutions sous licence par rapport aux solutions open source.

- Vous ne définissez pas de stratégies de cycle de vie pour le stockage.
- Vous ne passez pas en revue les nouveaux services et les nouvelles fonctionnalités du AWS Cloud.
- Vous utilisez uniquement le stockage par blocs.

Avantages liés au respect de cette bonne pratique : En tenant compte des coûts dans vos prises de décision, vous pouvez utiliser des ressources plus efficaces et explorer d'autres investissements.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

L'optimisation des charges de travail en termes de coûts peut améliorer l'utilisation des ressources et éviter le gaspillage dans une charge de travail cloud. La prise en compte des coûts dans les décisions architecturales implique généralement de dimensionner correctement les composants de la charge de travail et de renforcer l'élasticité, ce qui se traduit par une amélioration de l'efficacité des performances de la charge de travail cloud.

Étapes d'implémentation

- Fixez des objectifs de coûts tels que des limites budgétaires pour votre charge de travail cloud.
- Identifiez les composants clés (tels que les instances et le stockage) qui augmentent le coût de votre charge de travail. Vous pouvez utiliser AWS Pricing Calculator et AWS Cost Explorer pour identifier les principaux facteurs de coûts dans votre charge de travail.
- Utilisez les Bonnes pratiques d'optimisation des coûts Well-Architected afin d'optimiser ces composants clés en termes de coûts.
- Surveillez et analysez en permanence les coûts afin d'identifier les opportunités d'optimisation des coûts dans votre charge de travail.
  - Utilisez Budgets AWS pour recevoir des alertes en cas de coûts inadmissibles.
  - Utilisez AWS Compute Optimizer ou AWS Trusted Advisor pour obtenir des recommandations en matière d'optimisation des coûts.
  - Utilisez Cost Anomaly Detection AWS pour détecter automatiquement les anomalies de coûts et analyser les causes racines.

Ressources

Documents connexes :

- Présentation détaillée du tableau de bord Cost Intelligence Dashboard (langue française non garantie)
- Centre d'architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS

Vidéos connexes :

- Voici mon architecture
- Optimisez les performances et les coûts de votre calcul AWS

Exemples connexes :

- Exemples AWS
- Exemples de kits SDK AWS
- Dimensionnement avec Compute Optimizer et activation de l'utilisation de la mémoire
- Code de démonstration AWS Compute Optimizer

PERF01-BP04 Évaluer l'impact des compromis sur les clients et l'efficacité de l'architecture

Lors de l'évaluation des améliorations liées à la performance, identifiez les choix qui affectent vos clients et l'efficacité de la charge de travail. Par exemple, si l'utilisation d'un magasin de données clé-valeur augmente les performances du système, il est important d'évaluer l'impact de la nature constante de cette modification à terme sur les clients.

Anti-modèles courants :

- Vous supposez que tous les gains de performances doivent être mis en œuvre, même s'il existe des compromis pour ce qui est de l'implémentation.
- Vous n'évaluez les modifications apportées aux charges de travail que lorsqu'un problème de performances a atteint un point critique.

Avantages liés au respect de cette bonne pratique : Lorsque vous évaluez les améliorations potentielles liées aux performances, vous devez décider si les compromis concernant les modifications sont compatibles avec les exigences de charge de travail. Dans certains cas, vous devrez peut-être mettre en place des contrôles supplémentaires pour compenser les compromis.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Identifiez les domaines critiques de votre architecture en termes de performances et d'impact sur les clients. Déterminez la façon dont vous pouvez apporter des améliorations ainsi que les compromis que ces améliorations entraînent et la façon dont ils affectent le système et l'expérience de l'utilisateur. Par exemple, la mise en œuvre de la mise en cache des données permet d'améliorer de manière significative les performances, mais nécessite une stratégie précise concernant la manière et le moment où mettre à jour ou invalider les données mises en cache pour empêcher un comportement incorrect du système.

Étapes d'implémentation

- Comprenez vos exigences en matière de charge de travail et vos SLA.
- Définissez clairement les facteurs d'évaluation. Les facteurs peuvent être liés au coût, à la fiabilité, à la sécurité et aux performances de votre charge de travail.
- Sélectionnez l'architecture et les services qui répondent à vos besoins.
- Menez des expériences et des démonstrations de faisabilité (POC) afin d'évaluer les facteurs de compromis et l'impact sur les clients et l'efficacité de l'architecture. En général, les charges de travail hautement disponibles, performantes et sécurisées consomment davantage de ressources cloud tout en offrant une meilleure expérience client.

Ressources

Documents connexes :

- Bibliothèque Amazon Builders' Library
- KPI Amazon QuickSight
- Amazon CloudWatch RUM
- Documentation X-Ray
- Comprenez les modèles de résilience et les compromis pour concevoir une architecture efficace dans le cloud

Vidéos connexes :

- Build a Monitoring Plan
- Optimize applications through Amazon CloudWatch RUM
- Demo of Amazon CloudWatch Synthetics

Exemples connexes :

- Mesurer le temps de chargement des pages avec Amazon CloudWatch Synthetics
- Client web Amazon CloudWatch RUM

PERF01-BP05 Utiliser des stratégies et des architectures de référence

Utilisez les stratégies internes et les architectures de référence existantes lors de la sélection des services et des configurations en vue d'augmenter votre efficacité lorsque vous concevez et mettez en œuvre votre charge de travail.

Anti-modèles courants :

- Vous autorisez un large éventail de technologies qui peuvent avoir un impact sur les frais généraux de gestion de votre entreprise.

Avantages liés au respect de cette bonne pratique : L'établissement d'une stratégie pour les choix d'architecture, de technologie et de fournisseur permet de prendre des décisions rapidement.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Le fait de disposer de politiques internes en matière de sélection des ressources et de l'architecture fournit des normes et des directives à suivre lors des choix architecturaux. Ces directives simplifient le processus de prise de décision lors du choix du bon service cloud et peuvent contribuer à améliorer l'efficacité des performances. Déployez votre charge de travail à l'aide de stratégies ou d'architectures de référence. Intégrez les services à votre déploiement dans le cloud. Utilisez ensuite vos tests de performance pour vérifier que vous pouvez continuer à répondre à vos exigences de performance.

Étapes d'implémentation

- Comprenez clairement les exigences de votre charge de travail cloud.
- Passez en revue les stratégies internes et externes pour identifier les plus pertinentes.
- Utilisez les architectures de référence appropriées fournies par AWS ou les bonnes pratiques de votre secteur.
- Créez un continuum composé de stratégies, de normes, d'architectures de référence et de directives normatives pour les situations courantes. Vos équipes pourront ainsi agir plus rapidement. Adaptez les ressources à votre secteur d'activité, le cas échéant.
- Validez ces stratégies et architectures de référence pour votre charge de travail dans les environnements de test.
- Restez informé des normes du secteur et des mises à jour AWS pour garantir que vos stratégies et architectures de référence contribuent à optimiser votre charge de travail cloud.

Ressources

Documents connexes :

- Centre d'architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS

Vidéos connexes :

- Voici mon architecture

Exemples connexes :

- Exemples AWS
- Exemples de kits SDK AWS

PERF01-BP06 Utiliser le benchmarking pour éclairer vos décisions architecturales

Définissez des points de référence pour les performances d'une charge de travail existante afin de comprendre ses performances sur le cloud et prendre des décisions architecturales sur la base de ces données.

Anti-modèles courants :

- Vous comptez sur des points de référence courants qui ne reflètent pas les caractéristiques de votre charge de travail.
- Vous utilisez les commentaires et la perception des clients comme seule référence.

Avantages liés au respect de cette bonne pratique : La définition des points de référence de votre implémentation actuelle vous permet de mesurer l'amélioration des performances.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Utilisez la définition de points de référence avec des tests synthétiques pour évaluer les performances des composants de votre charge de travail. La définition de points de référence est généralement plus rapide à configurer que les tests de charge. Elle est utilisée pour évaluer la technologie pour un composant en particulier. La définition de points de référence est souvent utilisée au début d'un nouveau projet, lorsque vous n'avez pas de solution complète pour le test de charge.

Vous pouvez créer vos propres tests de définition de points de référence, ou utiliser un test conforme aux normes du secteur, comme le TPC-DS, pour évaluer vos charges de travail. Les points de référence du secteur sont utiles lorsque vous comparez différents environnements. Les points de référence personnalisés sont utiles pour cibler certains types d'opérations que vous souhaitez effectuer dans votre architecture.

Avec le benchmarking, il est important de préparer votre environnement de test pour obtenir des résultats valides. Exécutez plusieurs fois le même point de référence pour vous assurer d'avoir capturé toute variabilité au fil du temps.

Étant donné que les points de référence sont généralement plus rapides à exécuter que les tests de charge, ils peuvent être utilisés plus tôt dans le pipeline de déploiement et fournir un retour rapide sur les écarts de performances. Lorsque vous évaluez un changement important dans un composant ou un service, un point de référence peut être un moyen rapide pour voir si la modification a un intérêt.

L'utilisation de la définition de points de référence avec un test de charge est essentielle, car un test de charge vous indique comment votre charge de travail se comporte dans un environnement de production.

Étapes d'implémentation

- Définissez les métriques (telles que l'utilisation de l'UC, la latence ou le débit) pour évaluer les performances de votre charge de travail.
- Identifiez et configurez un outil de benchmarking adapté à votre charge de travail. Vous pouvez utiliser des services AWS (tels que Amazon CloudWatch) ou un outil tiers compatible avec votre charge de travail.
- Effectuez vos tests comparatifs et surveillez les métriques pendant le test.
- Analysez et documentez les résultats de benchmarking afin d'identifier les éventuels goulots d'étranglement et problèmes.
- Utilisez les résultats des tests pour prendre des décisions architecturales et ajuster votre charge de travail. Cet ajustement peut impliquer la modification des services ou l'adoption de nouvelles fonctionnalités.
- Testez à nouveau votre charge de travail après l'ajustement.

Ressources

Documents connexes :

- Centre d'architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics

Vidéos connexes :

- Voici mon architecture
- Optimisez les applications grâce à Amazon CloudWatch RUM
- Demo of Amazon CloudWatch Synthetics

Exemples connexes :

- Exemples AWS
- Exemples de kits SDK AWS
- Tests de charge distribuée
- Mesurer le temps de chargement des pages avec Amazon CloudWatch Synthetics
- Client web Amazon CloudWatch RUM

PERF01-BP07 Utiliser une approche orientée données pour les choix architecturaux

Définissez une approche orientée données claire pour les choix architecturaux afin de vérifier que les services et configurations cloud appropriés sont utilisés pour répondre aux besoins spécifiques de votre entreprise.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne devrait pas être mise à jour au fil du temps.
- Vos choix architecturaux sont basés sur des suppositions et des hypothèses.
- Vous introduisez des modifications d'architecture au fil du temps sans justification.

Avantages liés au respect de cette bonne pratique : En adoptant une approche bien définie pour les choix architecturaux, vous utilisez les données pour influencer la conception de votre charge de travail et prendre des décisions éclairées au fil du temps.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Mobilisez l'expérience et l'expertise des ressources cloud internes ou faites appel à des ressources externes, comme des cas d'utilisation publiés ou des livres blancs pour définir un processus de sélection des ressources et services dans votre architecture. Vous devriez disposer d'un processus bien défini qui encourage l'expérimentation et le benchmarking avec les services qui pourraient être utilisés dans votre charge de travail.

Les backlogs relatifs aux charges de travail critiques doivent non seulement comprendre des témoignages d'utilisateurs proposant des fonctionnalités pertinentes pour les entreprises et les

utilisateurs, mais également des récits techniques qui constituent une piste architecturale pour la charge de travail. Cette piste s'inspire des nouvelles avancées technologiques et des nouveaux services et les adopte sur la base de données et de justifications appropriées. Cela permet de vérifier que l'architecture reste pérenne et ne stagne pas.

Étapes d'implémentation

- Collaborez avec les principales parties prenantes pour définir les exigences en matière de charge de travail, y compris les considérations relatives aux performances, à la disponibilité et aux coûts. Tenez compte de facteurs tels que le nombre d'utilisateurs et le modèle d'utilisation de votre charge de travail.
- Créez une piste architecturale ou un backlog technologique qui est axé en priorité sur le backlog fonctionnel.
- Évaluez les différents services cloud (pour en savoir plus, consultez PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles).
- Explorez les différents modèles architecturaux, tels que les microservices ou le modèle sans serveur, qui répondent à vos exigences en termes de performances (pour en savoir plus, consultez PERF01-BP02 Utiliser les recommandations de votre fournisseur de cloud ou d'un partenaire approprié pour en savoir plus sur les modèles d'architecture et les bonnes pratiques).

- Consultez d'autres équipes, des diagrammes d'architecture et des ressources, tels que les architectes de solutions AWS, Centre d'architecture AWSet AWS Partner Networkpour vous aider à choisir l'architecture adaptée à votre charge de travail.

- Définissez des métriques de performance telles que le débit et le temps de réponse qui peuvent vous aider à évaluer les performances de votre charge de travail.
- Testez et utilisez des métriques définies pour valider les performances de l'architecture sélectionnée.
- Surveillez en continu les performances et effectuez les ajustements nécessaires pour maintenir un niveau optimal de performance pour votre architecture.
- Documentez l'architecture que vous avez sélectionnée et les décisions que vous avez prises comme référence pour les futures mises à jour et les futurs apprentissages.
- Vérifiez en permanence l'approche de sélection de l'architecture et mettez-la à jour en fonction des apprentissages, des nouvelles technologies et des métriques indiquant un changement nécessaire ou un problème dans l'approche actuelle.

Ressources

Documents connexes :

- Bibliothèque de solutions AWS
- Centre de connaissances AWS

Vidéos connexes :

- Voici mon architecture

Exemples connexes :

- Exemples AWS
- Exemples de kits SDK AWS

Informatique et matériel

PERF 2. Comment sélectionnez-vous et utilisez-vous les ressources de calcul dans votre charge de travail ?

Le choix d'une solution de calcul optimale pour une charge de travail particulière peut varier selon la conception de l'application, les modèles d'utilisation et les paramètres de configuration. Les architectures peuvent utiliser différentes solutions de calcul pour divers composants et permettent différentes fonctionnalités pour améliorer les performances. Choisir une solution de calcul inadaptée pour une architecture peut nuire à ses performances.

Bonnes pratiques

- PERF02-BP01 Sélectionner les meilleures options de calcul pour votre charge de travail
- PERF02-BP02 Comprendre les configurations et les fonctionnalités de calcul disponibles
- PERF02-BP03 Collecter les métriques liées au calcul
- PERF02-BP04 Configurer et dimensionner correctement les ressources de calcul
- PERF02-BP05 Mettre à l'échelle vos ressources de calcul de manière dynamique
- PERF02-BP06 Utiliser des accélérateurs de calcul matériels optimisés

PERF02-BP01 Sélectionner les meilleures options de calcul pour votre charge de travail

La sélection de l'option de calcul la mieux adaptée à votre charge de travail vous permet d'améliorer les performances, de réduire les coûts d'infrastructure inutiles et de diminuer les efforts opérationnels nécessaires pour maintenir votre charge de travail.

Anti-modèles courants :

- Vous utilisez la même option de calcul que celle utilisée sur site.
- Vous manquez de connaissances sur les options, les fonctionnalités et les solutions de calcul cloud et sur la manière dont elles pourraient améliorer vos performances de calcul.
- Vous surprovisionnez une option de calcul existante pour répondre aux exigences de mise à l'échelle ou de performances, alors qu'une autre option de calcul s'alignerait plus précisément sur les caractéristiques de votre charge de travail.

Avantages liés au respect de cette bonne pratique : En identifiant les exigences de calcul et en les comparant aux options disponibles, vous pouvez optimiser votre charge de travail en termes de ressources.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Pour optimiser vos charges de travail cloud afin d'améliorer l'efficacité des performances, il est important de sélectionner les options de calcul les mieux adaptées à votre cas d'utilisation et à vos exigences de performances. AWS fournit une variété d'options de calcul qui sont adaptées aux différentes charges de travail dans le cloud. Par exemple, vous pouvez utiliser Amazon EC2 pour lancer et gérer des serveurs virtuels, AWS Lambda pour exécuter du code sans avoir à provisionner ou gérer des serveurs, Amazon ECS ou Amazon EKS pour exécuter et gérer des conteneurs, ou AWS Batch pour traiter de gros volumes de données en parallèle. En fonction de vos besoins en termes de mise à l'échelle et de calcul, vous devez choisir et configurer la solution de calcul optimale pour votre situation. Vous pouvez également envisager d'utiliser plusieurs types de solutions de calcul dans une seule charge de travail, car chacune présente ses avantages et ses inconvénients.

Les étapes suivantes vous guident dans la sélection des options de calcul adaptées aux caractéristiques de votre charge de travail et à vos exigences de performances.

Étapes d'implémentation

1. Comprenez les exigences de calcul de votre charge de travail. Les exigences clés à prendre en compte incluent les besoins de traitement, les modèles de trafic, les modèles d'accès aux données, les besoins de mise à l'échelle et les exigences de latence.
2. Découvrez les différentes options de calcul disponibles pour votre charge de travail sur AWS (comme décrit dans PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles. » Voici quelques options de calcul AWS clés, leurs caractéristiques et leurs cas d'utilisation courants :

Service AWS	Principales caractéristiques	Cas d'utilisation courants
Les instances de serveur virtuel Amazon Elastic Compute Cloud (Amazon EC2)	Possède une option dédiée pour le matériel, les exigences de licence, une large sélection de différent es familles d'instances, les types de processeurs et les accélérateurs de calcul	Migration « lift-and-shift », application monolithique, environnements hybrides, applications d'entreprise
Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	Déploiement facile, environne ments cohérents, évolutivité	Microservices, environne ments hybrides
AWS Lambda	Calcul sans serveur service qui exécute du code en réponse à des événements et gère automatiquement les ressources de calcul sous-jacentes.	Microservices, applications basées sur les événements
AWS Batch	Provisionne et met à l’échelle de manière efficace et dynamique Amazon Elastic Container Service (Amazon ECS), Amazon Elastic	HPC, entraîner les modèles de ML

Service AWS	Principales caractéristiques	Cas d'utilisation courants
	Kubernetes Service (Amazon EKS)et AWS Fargate ressources de calcul, avec la possibilité d'utiliser des instances Spot ou à la demande en fonction des exigences de votre travail	
Amazon Lightsail	Application Linux et Windows préconfigurée pour exécuter de petites charges de travail	Applications Web simples, site Web personnalisé

3. Évaluez les coûts (tels que le tarif horaire ou le transfert de données) et les frais de gestion (tels que l'application de correctifs et la mise à l'échelle) associés à chaque option de calcul.
4. Réalisez des tests et procédez au benchmarking dans un environnement hors production afin d'identifier l'option de calcul la mieux adaptée à vos exigences en termes de charge de travail.
5. Après avoir testé et identifié votre nouvelle solution de calcul, planifiez votre migration et validez vos métriques de performance.
6. Utilisez des outils de surveillance AWS tels que Amazon CloudWatch et des services d'optimisation tels que AWS Compute Optimizer pour optimiser en permanence vos ressources de calcul en fonction des modèles d'utilisation réels.

Ressources

Documents connexes :

- Calcul sur le cloud avec AWS
- Types d'instances Amazon EC2
- Conteneurs Amazon EKS : composants master Amazon EKS
- Conteneurs Amazon ECS : instances de conteneur Amazon Amazon ECS
- Fonctions : configuration des fonctions Lambda
- Conseils prescriptifs pour les conteneurs
- Conseils prescriptifs pour les modèles sans serveur

Vidéos connexes :

- Comment choisir l'option de calcul pour les startups
- Optimisez les performances et les coûts de votre calcul AWS
- Fondations Amazon EC2
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Déployez des modèles de ML à des fins d'inférence avec des performances élevées et à faible coût
- Un calcul de meilleure qualité, plus rapide et moins cher : rentabiliser Amazon EC2

Exemples connexes :

- Migrer l'application Web vers des conteneurs (langue française non garantie)
- Exécuter un modèle Hello World sans serveur

PERF02-BP02 Comprendre les configurations et les fonctionnalités de calcul disponibles

Découvrez les options et les fonctionnalités de configuration disponibles pour votre service de calcul qui vous aideront à allouer la quantité de ressources appropriée et à améliorer l'efficacité des performances.

Anti-modèles courants :

- Vous ne comparez pas les options de calcul ou les familles d'instances disponibles avec les caractéristiques de la charge de travail.
- Vous surprovisionnez les ressources de calcul pour répondre aux pics de demande.

Avantages liés au respect de cette bonne pratique : Familiarisez-vous avec les fonctionnalités et les configurations de calcul d'AWS pour pouvoir utiliser une solution de calcul optimisée qui répond aux caractéristiques et aux besoins de votre charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Chaque solution de calcul dispose de configurations et de fonctionnalités uniques pour prendre en charge différentes caractéristiques et exigences de charge de travail. Découvrez comment

ces options soutiennent votre charge de travail et déterminez celles qui sont optimales pour votre système. Parmi ces options, citons, par exemple) la famille d'instances, les tailles, les fonctionnalités (GPU, E/S), la capacité de débordement (bursting), les délais d'attente, les tailles de fonction, les instances de conteneur et la simultanéité. Si votre charge de travail utilise la même option de calcul depuis plus de quatre semaines et que vous anticipez que les caractéristiques resteront les mêmes à l'avenir, vous pouvez utiliser AWS Compute Optimizer pour déterminer si votre option de calcul actuelle est adaptée aux charges de travail du point de vue du processeur et de la mémoire.

Étapes d'implémentation

1. Comprenez les exigences de la charge de travail (comme les besoins en UC, la mémoire et la latence).
2. Consultez la documentation AWS et les bonnes pratiques pour en savoir plus sur les options de configuration recommandées qui peuvent vous aider à améliorer vos performances de calcul. Voici quelques options de configuration clés à prendre en compte :

Option de configuration	Exemples
Type d'instance	• Les instances optimisées pour le calcul sont idéales pour les charges de travail qui exigent un ratio vCPU/mémoire plus élevé. • Les instances optimisées pour la mémoire offrent de grandes quantités de mémoire pour soutenir les charges de travail gourmandes en mémoire. • Les instances optimisées pour le stockage sont conçues pour les charges de travail nécessitant un accès séquentiel élevé en lecture et en écriture (IOPS) au stockage local.
Modèle de tarification	• Instances à la demande Les instances à la demande vous permettent d'utiliser la capacité de calcul à l'heure ou à la seconde sans engagement à long terme. Ces instances sont idéales pour dépasser les

Option de configuration	Exemples
	besoins de base en matière de performances. • Savings Plans permettent de réaliser des économies importantes par rapport aux instances à la demande, en échange d'un engagement à utiliser une quantité spécifique de puissance de calcul pour une période d'un ou de trois ans. • instances Spot. vous permettent de tirer parti de la capacité d'instance inutilisée à un prix réduit pour vos charges de travail sans état et tolérantes aux pannes.
Auto Scaling	Utilisez Auto Scaling pour faire correspondre les ressources de calcul aux modèles de trafic.
Dimensionnement	• Utilisez Compute Optimizer pour recevoir des recommandations optimisées par le machine learning sur la configuration de calcul qui correspond le mieux à vos caractéristiques de calcul. • Utilisez AWS Lambda Power Tuning pour sélectionner la meilleure configuration pour votre fonction Lambda.

Option de configuration	Exemples
Accélérateurs de calcul matériels	• Les instances de calcul accéléré exécutent des fonctions telles que le traitement graphique ou la mise en correspondance de modèles de données de manière plus efficace que les alternatives basées sur le CPU. • Pour les charges de travail de machine learning, tirez parti d'un matériel conçu spécialement pour votre charge de travail, comme AWS Trainium, AWS Inferentiaet Amazon EC2 DL1

Ressources

Documents connexes :

- Calcul sur le cloud avec AWS
- Types d'instances Amazon EC2
- Contrôle de l'état du processeur pour votre instance Amazon EC2 (langue française non garantie)
- Conteneurs Amazon EKS : composants master Amazon EKS
- Conteneurs Amazon ECS : instances de conteneur Amazon Amazon ECS
- Fonctions : configuration des fonctions Lambda

Vidéos connexes :

- Fondations Amazon EC2
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Optimisez les performances et les coûts de votre calcul AWS

Exemples connexes :

- Dimensionnement avec Compute Optimizer et activation de l'utilisation de la mémoire
- Code de démonstration AWS Compute Optimizer

PERF02-BP03 Collecter les métriques liées au calcul

Enregistrez et suivez les métriques liées au calcul pour mieux comprendre comment fonctionnent vos ressources de calcul et améliorer leurs performances et leur utilisation.

Anti-modèles courants :

- Vous utilisez uniquement la recherche manuelle des fichiers journaux pour les métriques.
- Vous n'utilisez que les métriques par défaut enregistrées par votre logiciel de surveillance.
- Vous n'examinez les métriques qu'en cas de problème.

Avantages liés au respect de cette bonne pratique : En collectant des métriques liées aux performances, vous pouvez aligner les performances des applications sur les exigences de l'entreprise afin de garantir que vous répondez à vos besoins en matière de charge de travail. Cela peut également vous aider à améliorer en continu les performances et l'utilisation des ressources de votre charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les charges de travail cloud peuvent générer de gros volumes de données telles que des métriques, des journaux et des événements. Dans le AWS Cloud, la collecte de métriques est une étape cruciale qui permet d'améliorer la sécurité, la rentabilité, les performances et la durabilité. AWS fournit un large éventail de métriques liées aux performances à l'aide de services de surveillance, tels que Amazon CloudWatch pour vous fournir des informations précieuses. Les métriques telles que l'utilisation de l'UC, l'utilisation de la mémoire, les E/S de disque et les métriques entrantes et sortantes du réseau peuvent fournir des informations sur les niveaux d'utilisation ou les goulots d'étranglement au niveau des performances. Utilisez ces métriques dans le cadre d'une approche fondée sur les données pour ajuster activement et optimiser les ressources de votre charge de travail. Dans un scénario idéal, vous devriez collecter toutes les métriques relatives à vos ressources de calcul sur une plateforme unique, avec des stratégies de conservation mises en œuvre pour atteindre les objectifs financiers et opérationnels.

Étapes d'implémentation

1. Identifiez les métriques liées aux performances qui sont pertinentes pour votre charge de travail. Vous devriez collecter des métriques relatives à l'utilisation des ressources et au fonctionnement de votre charge de travail cloud (comme le temps de réponse et le débit).

   a. Métriques par défaut Amazon EC2
   b. Métriques Amazon ECS par défaut
   c. Métriques par défaut Amazon EKS
   d. Métriques Lambda par défaut
   e. Métriques de mémoire et de disque Amazon EC2
2. Choisissez et configurez la solution de journalisation et de surveillance adaptée à votre charge de travail.
   a. Observabilité native AWS
   b. AWS Distro for OpenTelemetry
   c. Amazon Managed Service for Prometheus
3. Définissez le filtre et l'agrégation requis pour les métriques en fonction de vos exigences en matière de charge de travail.
   a. Quantifier les métriques d'application personnalisées avec Amazon CloudWatch Logs et les filtres de métrique (langue française non garantie)
   b. Collecter des métriques personnalisées avec le balisage stratégique Amazon CloudWatch (langue française non garantie)
4. Configurez des stratégies de conservation des données pour vos métriques afin qu'elles correspondent à vos objectifs sécuritaires et opérationnels.
   a. Métriques de conservation des données pour CloudWatch
   b. Conservation des données pour CloudWatch Logs
5. Si nécessaire, créez des alarmes et des notifications pour vos métriques afin de vous aider à résoudre de manière proactive les problèmes liés aux performances.
   a. Créer des alarmes pour les métriques personnalisées à l'aide de la détection d'anomalies Amazon CloudWatch (langue française non garantie)
   b. Créer des métriques et des alarmes pour certaines pages Web avec RUM Amazon CloudWatch (langue française non garantie)
6. Utilisez l'automatisation pour déployer vos agents d'agrégation de métriques et de journaux.
   a. Automation AWS Systems Manager
   b. Collecteur OpenTelemetry

Ressources

Documents connexes :

- Documentation Amazon CloudWatch
- Collecte des métriques et des journaux des instances Amazon EC2 et serveurs sur site avec l'agent CloudWatch
- Accès à Amazon CloudWatch Logs pour AWS Lambda
- Utiliser CloudWatch Logs avec des instances de conteneur
- Publier des métriques personnalisées
- AWS Answers : journalisation centralisée
- Services AWS publiant des métriques CloudWatch
- Surveillance d'Amazon EKS sur AWS Fargate

Vidéos connexes :

- Application Performance Management on AWS

Exemples connexes :

- Niveau 100 : surveillance avec les tableaux de bord CloudWatch
- Niveau 100 : surveillance d'une instance Windows EC2 avec les tableaux de bord CloudWatch
- Niveau 100 : surveillance d'une instance Amazon Linux EC2 avec les tableaux de bord CloudWatch

PERF02-BP04 Configurer et dimensionner correctement les ressources de calcul

Configurez et dimensionnez correctement les ressources de calcul en fonction des exigences de performance de votre charge de travail et évitez de sous-utiliser ou de surexploiter les ressources.

Anti-modèles courants :

- Vous ignorez les exigences de performance de votre charge de travail, ce qui entraîne un surprovisionnement ou un sous-provisionnement des ressources de calcul.
- Vous ne choisissez que la plus grande ou la plus petite instance disponible pour toutes les charges de travail.

- Vous n'utilisez qu'une seule famille d'instances pour faciliter la gestion.
- Vous ignorez les recommandations d'AWS Cost Explorer ou de Compute Optimizer concernant le redimensionnement.
- Vous ne réévaluez pas la charge de travail pour voir si de nouveaux types d'instances pourraient convenir.
- Vous ne certifiez qu'un petit nombre de configurations d'instance pour votre organisation.

Avantages liés au respect de cette bonne pratique : Dimensionner correctement les ressources de calcul garantit le fonctionnement optimal dans le cloud en évitant le surprovisionnement et le sous-provisionnement des ressources. Le dimensionnement correct des ressources de calcul se traduit généralement par de meilleures performances, une meilleure expérience client et une baisse des coûts.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Le dimensionnement correct permet aux organisations d'exploiter leur infrastructure cloud de manière efficace et rentable tout en répondant aux besoins de l'entreprise. Le surprovisionnement des ressources cloud peut entraîner des coûts supplémentaires, tandis que le sous-provisionnement peut générer de faibles performances et une expérience client négative. AWS fournit des outils tels que AWS Compute Optimizer et AWS Trusted Advisor qui utilisent des données historiques pour fournir des recommandations de redimensionnement de vos ressources de calcul.

Étapes d'implémentation

- Choisissez le type d'instance qui correspond le mieux à vos besoins :
  - Comment choisir le type d'instance Amazon EC2 approprié pour ma charge de travail ?
  - Sélection de type d'instance basée sur des attributs pour la flotte Amazon EC2
  - Créez un groupe Auto Scaling en utilisant la sélection du type d'instance basée sur des attributs (langue française non garantie)
  - Optimisation de vos coûts de calcul Kubernetes avec la consolidation Karpenter (langue française non garantie)
- Analysez les différentes caractéristiques de performances de votre charge de travail et la façon dont ces caractéristiques se rapportent à la mémoire, au réseau et à l'utilisation du processeur. Utilisez ces données pour choisir les ressources qui correspondent le mieux aux objectifs de votre charge de travail en termes de profil et de performance.

- Surveillez l'utilisation de vos ressources à l'aide des outils de surveillance d'AWS tels que Amazon CloudWatch.
- Sélectionnez la configuration adaptée à vos ressources de calcul.
  - Pour les charges de travail éphémères, évaluez les métriques d'instance Amazon CloudWatch telles que `CPUUtilization` pour identifier si l'instance est sous-utilisée ou surexploitée.
  - Pour les charges de travail stables, vérifiez les outils de redimensionnement AWS tels que AWS Compute Optimizer et AWS Trusted Advisor à intervalles réguliers pour identifier les opportunités d'optimisation et de redimensionnement des ressources de calcul.
    - Atelier Well-Architected : recommandations de redimensionnement
    - Atelier Well-Architected : redimensionnement avec Compute Optimizer
- Testez les changements de configuration dans un environnement hors production avant de les implémenter dans un environnement réel.
- Réévaluez en permanence les nouvelles offres de calcul et comparez-les aux besoins de votre charge de travail.

Ressources

Documents connexes :

- Calcul sur le cloud avec AWS
- Types d'instances Amazon EC2
- Conteneurs Amazon ECS : instances de conteneur Amazon Amazon ECS
- Conteneurs Amazon EKS : composants master Amazon EKS
- Fonctions : configuration des fonctions Lambda
- Contrôle de l'état du processeur pour votre instance Amazon EC2 (langue française non garantie)

Vidéos connexes :

- Fondations Amazon EC2
- Un calcul de meilleure qualité, plus rapide et moins cher : rentabiliser Amazon EC2
- Déployez des modèles de ML à des fins d'inférence avec des performances élevées et à faible coût
- Optimisez les performances et les coûts de votre calcul AWS

- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Simplifier le traitement des données pour améliorer l'innovation grâce à des outils sans serveur

Exemples connexes :

- Dimensionnement avec Compute Optimizer et activation de l'utilisation de la mémoire
- Code de démonstration AWS Compute Optimizer

PERF02-BP05 Mettre à l'échelle vos ressources de calcul de manière dynamique

Utilisez l'élasticité du cloud pour mettre à l'échelle vos ressources de calcul de manière dynamique afin de répondre à vos besoins et éviter de surprovisionner ou de sous-provisionner la capacité de votre charge de travail.

Anti-modèles courants :

- Vous réagissez aux alertes en augmentant manuellement la capacité.
- Vous utilisez les mêmes recommandations de dimensionnement (généralement, infrastructure statique) que sur site.
- Vous conservez une capacité accrue après un événement de mise à l'échelle au lieu de la réduire.

Avantages liés au respect de cette bonne pratique : En configurant et en testant l'élasticité des ressources de calcul, vous pouvez économiser de l'argent, maintenir les points de référence des performances et améliorer la fiabilité en fonction de l'évolution du trafic.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

AWS apporte la flexibilité nécessaire pour mettre à l'échelle vos ressources de manière dynamique grâce à divers mécanismes de mise à l'échelle afin de répondre à l'évolution de la demande. Combinée aux métriques liées au calcul, la mise à l'échelle dynamique permet aux charges de travail de réagir automatiquement aux changements et d'utiliser l'ensemble optimal de ressources de calcul pour atteindre son objectif.

Vous pouvez utiliser plusieurs approches pour rapprocher l'offre de ressources de la demande.

- Approche visant à suivre les cibles: surveillez votre métrique de capacité de mise à l'échelle et augmentez ou réduisez automatiquement votre capacité selon vos besoins.
- Mise à l'échelle prédictive: mettez à l'échelle en prévision des tendances quotidiennes et hebdomadaires.
- Approche basée sur un calendrier: planifiez votre propre calendrier de mise à l'échelle en fonction de changements de charge prévisibles.
- Mise à l'échelle des services: choisissez des services (sans serveur, par exemple) conçus pour se mettre à l'échelle automatiquement.

Vous devez vous assurer que les déploiements de charge de travail peuvent gérer les événements de mise à l'échelle ascendante et descendante.

Étapes d'implémentation

- Les instances de calcul, les conteneurs et les fonctions fournissent des mécanismes d'élasticité, soit en combinaison avec l'autoscaling, soit en tant que fonctionnalité du service. Voici des exemples de mécanismes d'autoscaling :

Mécanisme de mise à l'échelle automatique	Où utiliser
Amazon EC2 Auto Scaling	Pour vous assurer que vous disposez du nombre adéquat d'instances Amazon EC2 disponibles pour gérer la charge utilisateur de votre application.
Application Auto Scaling	Pour mettre à l'échelle automatiquement les ressources pour les services AWS individuels au-delà d'Amazon EC2, tels que les fonctions AWS Lambda ou les services Amazon Elastic Container Service (Amazon ECS) .
Kubernetes Cluster Autoscaler/Karpenter	Pour mettre à l'échelle automatiquement les clusters Kubernetes.

- La mise à l'échelle est souvent abordée pour les services de calcul, tels que les instances Amazon EC2 ou les fonctions AWS Lambda. Assurez-vous également de prendre en compte la configuration des services non liés au calcul tels que AWS Glue afin de répondre à la demande.

- Vérifiez que les métriques de mise à l'échelle correspondent aux caractéristiques de la charge de travail en cours de déploiement. Si vous déployez une application de transcodage vidéo, une utilisation de 100 % du processeur est attendue. N'en faites pas votre métrique principale. Utilisez plutôt la profondeur de la file d'attente des tâches de transcodage. Vous pouvez utiliser une métrique personnalisée pour votre politique de mise à l'échelle si besoin. Pour choisir les bonnes métriques, tenez compte des conseils suivants pour Amazon EC2 :
  - La métrique doit être une métrique d'utilisation valide et décrire à quel point l'instance est occupée.
  - La valeur de la métrique doit augmenter ou diminuer proportionnellement au nombre d'instances dans le groupe Auto Scaling.
- Assurez-vous d'utiliser la mise à l'échelle dynamique plutôt que la mise à l'échelle manuelle pour votre groupe Auto Scaling. Nous vous recommandons également d'utiliser des politiques de mise à l'échelle du suivi des cibles dans votre mise à l'échelle dynamique.
- Vérifiez que les déploiements de charges de travail peuvent gérer les deux événements de mise à l'échelle (augmentation et diminution des charges de travail). À titre d'exemple, vous pouvez utiliser l'historique d'activité pour vérifier une activité de mise à l'échelle pour un groupe Auto Scaling.
- Évaluez votre charge de travail pour les modèles prédictifs et mettez-la à l'échelle de manière proactive pour anticiper les changements prévisibles et prévus de la demande. Avec la mise à l'échelle prédictive, vous pouvez supprimer le besoin de surprovisionner de la capacité. Pour en savoir plus, consultez Mise à l'échelle prédictive avec Amazon EC2 Auto Scaling. »

Ressources

Documents connexes :

- Calcul sur le cloud avec AWS
- Types d'instances Amazon EC2
- Conteneurs Amazon ECS : instances de conteneur Amazon Amazon ECS
- Conteneurs Amazon EKS : composants master Amazon EKS
- Fonctions : configuration des fonctions Lambda
- Contrôle de l'état du processeur pour votre instance Amazon EC2 (langue française non garantie)
- En savoir plus sur la Auto Scaling d'un cluster Amazon ECS
- Présentation de Karpenter, un Kubernetes Cluster Autoscaler hautement performant et open source (langue française non garantie)

Vidéos connexes :

- Fondations Amazon EC2
- Un calcul de meilleure qualité, plus rapide et moins cher : rentabiliser Amazon EC2
- Optimisez les performances et les coûts de votre calcul AWS
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Concevoir un environnement de calcul rentable, économe en énergie et en ressources

Exemples connexes :

- Exemples de groupes Amazon EC2 Auto Scaling
- Implémenter la mise à l'échelle automatique avec Karpenter

PERF02-BP06 Utiliser des accélérateurs de calcul matériels optimisés

Utilisez des accélérateurs matériels pour exécuter certaines fonctions de manière plus efficace que les alternatives basées sur l'UC.

Anti-modèles courants :

- En ce qui concerne votre charge de travail, vous n'avez pas comparé une instance à usage général à une instance dédiée capable de fournir de meilleures performances à faible coût.
- Vous utilisez des accélérateurs de calcul matériels pour les tâches qui peuvent être plus efficaces en utilisant des alternatives basées sur l'UC.
- Vous ne surveillez pas l'utilisation du GPU.

Avantages liés au respect de cette bonne pratique : En utilisant des accélérateurs matériels, tels que des unités de traitement graphique (GPU) et des circuits logiques programmables (FPGA), vous pouvez exécuter certaines fonctions de traitement de manière plus efficace.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Les instances de calcul accéléré donnent accès à des accélérateurs de calcul matériels tels que les GPU et les FPGA. Ces accélérateurs matériels exécutent certaines fonctions comme le traitement graphique ou la correspondance de modèles de données plus efficacement que les alternatives

basées sur l'UC. De nombreuses charges de travail accélérées, telles que le rendu, le transcodage et le machine learning, sont très variables en termes d'utilisation des ressources. Exécutez ce matériel uniquement pendant le temps nécessaire et mettez-le hors service grâce à l'automatisation lorsque vous n'en avez plus besoin afin d'améliorer l'efficacité globale des performances.

Étapes d'implémentation

- Identifiez quelles instances informatiques accélérées peuvent répondre à vos besoins.
- Pour les charges de travail de machine learning, tirez parti d'un matériel conçu spécialement pour votre charge de travail, comme AWS Trainium, AWS Inferentiaet Amazon EC2 DL1. » Les instances Inferentia AWS, telles que les instances Inf2, offrent des performances par watt jusqu'à 50 % supérieures à celles des instances comparables Amazon EC2. »
- Collectez des métriques d'utilisation pour vos instances de calcul accéléré. Par exemple, vous pouvez utiliser un agent CloudWatch pour collecter des métriques comme `utilization_gpu` et `utilization_memory` pour vos GPU, comme indiqué dans Collecter les métriques des GPU NVIDIA avec Amazon CloudWatch.
- Optimisez le code, le fonctionnement du réseau et les paramètres des accélérateurs matériels pour veiller à ce que le matériel sous-jacent soit pleinement utilisé.
  - Optimisation des paramètres GPU
  - Surveillance et optimisation des GPU dans l'AMI Deep Learning
  - Optimisation des E/S pour le réglage des performances de GPU pour l'entraînement du deep learning dans Amazon SageMaker
- Utilisez les dernières bibliothèques performantes et les pilotes GPU.
- Utilisez l'automatisation pour libérer les instances GPU lorsqu'elles ne sont pas utilisées.

Ressources

Documents connexes :

- Instances GPU (langue française non garantie)
- Instances avec Trainium AWS (langue française non garantie)
- Instances avec Inferentia AWS (langue française non garantie)
- Passons à l'architecture ! Architecture avec des puces personnalisées et des accélérateurs

- Calcul accéléré

- Instances VT1 Amazon EC2
- Comment choisir le type d'instance Amazon EC2 approprié pour ma charge de travail ?
- Choisissez le meilleur accélérateur d'IA et la meilleure compilation de modèles pour l'inférence de vision par ordinateur avec Amazon SageMaker

Vidéos connexes :

- How to select Amazon EC2 GPU instances for deep learning
- Deploying Cost-Effective Deep Learning Inference

Gestion des données

PERF 3. Comment stockez-vous aux données de votre charge de travail, comment les gérez-vous et comment y accédez-vous ?

La solution optimale de gestion des données pour un système particulier varie en fonction du type de données (bloc, fichier ou objet), des modèles d'accès (aléatoire ou séquentiel), du débit requis, de la fréquence d'accès (en ligne, hors ligne, archivage), de la fréquence de mise à jour (WORM, dynamique), ainsi que des contraintes de disponibilité et de durabilité. Les charges de travail bien architecturées utilisent des magasins de données sur mesure qui intègrent différentes fonctionnalités pour améliorer les performances.

Bonnes pratiques

- PERF03-BP01 Utiliser un magasin de données dédié le mieux adapté à vos besoins en matière de stockage des données et d'accès aux données
- PERF03-BP02 Évaluer les options de configuration disponibles pour un magasin de données
- PERF03-BP03 Collecter et archiver les métriques de performance du magasin de données
- PERF03-BP04 Mettre en œuvre des stratégies pour améliorer les performances des requêtes dans un magasin de données
- PERF03-BP05 Mise en œuvre de modèles d'accès aux données utilisant la mise en cache

PERF03-BP01 Utiliser un magasin de données dédié le mieux adapté à vos besoins en matière de stockage des données et d'accès aux données

Comprenez les caractéristiques des données (telles que la possibilité de partage, la taille, la taille du cache, les modèles d'accès, la latence, le débit et la persistance des données) afin de sélectionner les magasins de données dédiés (stockage ou base de données) adaptés à votre charge de travail.

Anti-modèles courants :

- Vous vous en tenez à un magasin de données, car l'équipe interne sait comment tirer parti de ce type de solution en particulier.
- Vous partez du principe que toutes les charges de travail ont des exigences similaires en termes de stockage de données et d'accès aux données.
- Vous n'avez pas implémentée de catalogue de données pour inventorier vos ressources de données.

Avantages liés au respect de cette bonne pratique : En comprenant l'importance des caractéristiques et des exigences des données, vous pouvez déterminer la technologie de stockage la plus efficace et la plus performante adaptée à vos besoins en matière de charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Lors de la sélection et de la mise en œuvre du stockage des données, assurez-vous que les caractéristiques d'interrogation, de mise à l'échelle et de stockage répondent aux exigences en matière de données de charge de travail. AWS fournit de nombreuses technologies de stockage de données et de base de données, notamment le stockage par blocs, le stockage d'objets, le stockage en continu, le système de fichiers et les bases de données relationnelles, clé-valeur, document, en mémoire, orientées graphe, de séries chronologiques et de registre. Chaque solution de gestion de données propose des options et des configurations pour prendre en charge vos cas d'utilisation et vos modèles de données. En comprenant les caractéristiques et les exigences des données, vous pouvez vous affranchir de la technologie de stockage monolithique et des approches restrictives et universelles pour vous concentrer sur la gestion appropriée des données.

Étapes d'implémentation

- Procédez à l'inventaire des différents types de données qui existent dans votre charge de travail.

- Comprenez et documentez les caractéristiques et les exigences des données, notamment :
  - Type de données (non structurées, semi-structurées, relationnelles)
  - Volume et croissance des données
  - Durabilité des données : persistantes, éphémères, temporaires
  - Exigences ACID (atomicité, cohérence, isolement, durabilité)
  - Modèles d'accès aux données (à lecture intensive ou à écriture intensive)
  - Latence
  - débit
  - IOPS (opérations d'entrée/sortie par seconde)
  - Durée de conservation des données
- Découvrez les différents magasins de données disponibles pour votre charge de travail sur AWS qui peuvent répondre aux caractéristiques de vos données (comme indiqué dans PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles). Voici quelques exemples de technologies de stockage AWS et leurs principales caractéristiques :

Type	Services AWS	Principales caractéristiques
Stockage d'objets	Amazon S3	Capacité de mise à l'échelle illimitée, haute disponibilité et plusieurs options d'accessi bilité. Pour transférer des objets et accéder à des objets dans et en dehors d'Amazon S3, utilisez un service, tel que Transfer Accelerat ion ou Points d'accèspour prendre en charge votre emplacement, vos besoins en sécurité et les modèles d'accès.
Archivage et stockage	Amazon S3 Glacier	Conçu pour l'archivage des données.

Type	Services AWS	Principales caractéristiques
Stockage en streaming	Amazon Kinesis Amazon Managed Streaming for Apache Kafka (Amazon MSK)	Ingestion et stockage efficaces des données de streaming.
Système de fichiers partagé	Amazon Elastic File System (Amazon EFS)	Système de fichiers montable auquel plusieurs types de solutions informatiques peuvent accéder.
Système de fichiers partagé	Amazon FSx	Créé sur les dernières solutions de calcul AWS pour prendre en charge quatre systèmes de fichiers fréquemment utilisés : NetApp ONTAP, OpenZFS, Windows File Serve et Lustre. Amazon FSx La latence, le débit et les IOPS Amazon FSx varient par système de fichiers et doivent être pris en compte lorsque vous sélectionnez le système de fichiers adapté aux besoins de vos charges de travail.

Type	Services AWS	Principales caractéristiques
Stockage par blocs	Amazon Elastic Block Store (Amazon EBS)	Service de stockage par bloc hautement performant et capable de mise à l'échelle conçu pour Amazon Elastic Compute Cloud (Amazon EC2). Amazon EBS propose un stockage basé sur SSD pour les charges de travail transactionnelles et exigeantes en termes d'IOPS ainsi qu'un stockage basé sur HDD pour les charges de travail exigeantes en termes de débit.
Base de données relationnelle	Amazon Aurora, Amazon RDS, Amazon Redshift. »	Conçues pour prendre en charge les transactions ACID (atomicité, cohérence, isolation et durabilité) et maintenir l'intégrité référentielle et la cohérence des données. De nombreuses applications traditionnelles, la planification des ressources d'entreprise (ERP), la gestion de la relation client (CRM) et l'e-commerce utilisent des bases de données relationnelles pour stocker leurs données.

Type	Services AWS	Principales caractéristiques
Base de données clé-valeur	Amazon DynamoDB	Optimisées pour les modèles d'accès courants, généralement pour stocker et récupérer de gros volumes de données. Les applications Web à trafic élevé, les systèmes d'e-commerce et les applications de jeu sont des cas d'utilisation typiques pour les bases de données de valeurs-clés.
Base de données de documents	Amazon DocumentDB	Conçues pour stocker des données semi-structurées sous forme de documents de type JSON. Ces bases de données aident les développeurs à créer et mettre à jour rapidement des applications telles que la gestion de contenu, les catalogues et les profils utilisateur.

Type	Services AWS	Principales caractéristiques
Base de données en mémoire	Amazon ElastiCache , Amazon MemoryDB for Redis	Utilisées pour les applications qui nécessitent un accès en temps réel aux données, la latence la plus faible et le débit le plus élevé. Vous pouvez utiliser des bases de données en mémoire pour la mise en cache des applications, la gestion des sessions, les classements des jeux, le magasin de fonctionnalités ML à faible latence, le système de messagerie à microservices et un mécanisme de streaming à haut débit
Base de données orientée graphe	Amazon Neptune	Destinées aux applications qui doivent parcourir et interroger des millions de relations entre des jeux de données graphiques hautement connectés avec une latence de millisecondes à grande échelle. De nombreuses entreprises utilisent des bases de données de graphiques pour la détection des fraudes, les réseaux sociaux et les moteurs de recommandation.

Type	Services AWS	Principales caractéristiques
Base de données de séries temporelles	Amazon Timestream	Utilisées pour collecter , synthétiser et extraire efficacement des informati ons à partir de données qui changent au fil du temps. Les applications IoT, les DevOps et la télémétrie industrielle peuvent utiliser des bases de données en séries chronolog iques.
Larges colonnes	Amazon Keyspaces (pour Apache Cassandra)	Utilise des tables, des lignes et des colonnes, mais contrairement à une base de données relationnelle, les noms et le format des colonnes peuvent varier d'une ligne à l'autre dans la même table. Généralem ent, vous voyez un magasin de colonnes larges dans les applications industrie lles à grande échelle pour la maintenance des équipemen ts, la gestion des parcs et l'optimisation des itinéraires.

Type	Services AWS	Principales caractéristiques
Registre	Amazon Quantum Ledger Datab (Amazon QLDB)	Fournit une autorité centralisée et fiable pour conserver un enregistrement évolutif, immuable et vérifiable grâce au chiffrement des transactions pour chaque application. Il n'est pas rare de voir des bases de données de registre utilisées pour les systèmes d'enregistrement, la chaîne d'approvisionnement, les inscriptions et même les transactions bancaires.

- Si vous créez une plateforme de données, tirez parti de l'architecture de données moderne sur AWS pour intégrer votre lac de données, votre entrepôt de données et vos magasins de données spécialement conçus.
- Les principales questions que vous devez vous poser lors du choix d'un magasin de données pour votre charge de travail sont les suivantes :

Question	Éléments à prendre en compte
Comment sont structurées les données ?	• Si les données ne sont pas structurées, envisagez un stockage d'objets comme Amazon S3 ou une base de données NoSQL telle qu' Amazon DocumentDB • Pour les données clé-valeur, envisagez DynamoDB, Amazon ElastiCache (Redis OSS) ou Amazon MemoryDB
Quel niveau d'intégrité référentielle est requis ?	• Pour les contraintes de clé étrangère, les bases de données relationnelles comme Amazon RDS et Aurora peuvent fournir ce niveau d'intégrité.

Question	Éléments à prendre en compte
	• En règle générale, dans un modèle de données NoSQL, vous dénormalisez les données en un seul document ou en une collection de documents à récupérer en une seule requête au lieu de joindre des documents ou des tables.
La conformité ACID (atomicité, cohérence, isolement, durabilité) est-elle requise ?	• Si les propriétés ACID associées aux bases de données relationnelles sont requises, envisagez une base de données relationn elle comme Amazon RDS et Aurora. » • Si une forte cohérence est requise pour la base de données NoSQL, vous pouvez utiliser des lectures fortement cohérentes avec DynamoDB. »
Comment les exigences de stockage vont-elles évoluer au fil du temps ? Comment cela affectera-t-il l'évolutivité ?	• Les bases de données sans serveur comme DynamoDB et Amazon Quantum Ledger Database (Amazon QLDB) se mettront à l'échelle de manière dynamique. • Les bases de données relationnelles ont des limites supérieures sur le stockage alloué et doivent souvent être partitionnées horizontalement à l'aide de mécanismes tels que le partitionnement une fois qu'elles atteignent ces limites.
Quelle est la proportion de requêtes en lecture par rapport aux requêtes en écriture ? La mise en cache pourrait-elle améliorer les performan ces ?	• Les charges de travail à lecture intensive peuvent bénéficier d'une couche de mise en cache, comme ElastiCache ou DAX si la base de données est DynamoDB. • Les lectures peuvent également être déchargées pour lire des réplicas avec des bases de données relationnelles comme Amazon RDS.

Question	Éléments à prendre en compte
Le stockage et la modification (OLTP - Online Transaction Processing) ou la récupérat ion et le reporting (OLAP - Online Analytical Processing) ont-ils une priorité plus élevée ?	• Pour un traitement transactionnel des lectures en l'état à haut débit, envisagez d'utiliser une base de données NoSQL comme DynamoDB. • Pour des modèles de lecture complexes à haut débit (tels que la jointure) avec cohérence, utilisez Amazon RDS. • Pour les requêtes analytiques, envisagez une base de données en colonnes comme Amazon Redshift ou exportez les données vers Amazon S3 et effectuez des analyses à l'aide d' Athena ou Amazon QuickSight. »
Quel est le niveau de durabilité requis pour les données ?	• Aurora réplique automatiquement vos données sur trois zones de disponibilité au sein d'une région. Autrement dit, vos données sont très durables avec moins de risque de perte de données. • DynamoDB est automatiquement répliqué sur plusieurs zones de disponibilité, assurant ainsi la haute disponibilité et la durabilité des données. • Amazon S3 offre une durabilité de 99,999999999 %. De nombreux services de base de données, tels que Amazon RDS et DynamoDB, prennent en charge l'exporta tion des données vers Amazon S3 pour une conservation et un archivage à long terme.

Question	Éléments à prendre en compte
Souhaitez-vous vous éloigner des moteurs de base de données commerciaux ou des coûts de licence ?	• Envisagez d'utiliser des moteurs open source tels que PostgreSQL et MySQL sur Amazon RDS ou Aurora. • Exploitez AWS Database Migration Service et AWS Schema Conversion Tool pour passer des moteurs de bases de données commerciaux vers des moteurs open source.
Qu'attendez-vous de la base de données du point de vue opérationnel ? Le passage aux services gérés est-il une préoccupation majeure ?	• L'utilisation d'Amazon RDS au lieu d'Amazon EC2 et de DynamoDB ou d'Amazon DocumentDB au lieu de l'auto-hébergement d'une base de données NoSQL contribue à réduire les frais généraux opérationnels.
Comment accédez-vous actuellement à la base de données ? S'agit-il uniquement d'un accès via une application, ou y a-t-il des utilisateurs BI et d'autres applications prêtes à l'emploi qui y sont connectées ?	• Si vous dépendez d'outils externes, vous devrez peut-être préserver la compatibilité avec les bases de données qu'ils prennent en charge. Amazon RDS est entièrement compatible avec les différentes versions de moteur qu'il prend en charge, notamment Microsoft SQL Server, Oracle, MySQL et PostgreSQL.

- Réalisez des tests et procédez au benchmarking dans un environnement hors production afin d'identifier le magasin de données qui peut répondre à vos exigences en termes de charge de travail.

Ressources

Documents connexes :

- Types de volume Amazon EBS
- Stockage Amazon EC2

- Amazon EFS : Amazon EFS Performance
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance
- Amazon S3 Glacier : documentation S3 Glacier
- Amazon S3 : directives en matière de débit de demandes et de performances
- Stockage cloud avec AWS
- Caractéristiques d'E/S Amazon EBS
- Bases de données cloud avec AWS
- Mise en cache de bases de données AWS
- DynamoDB Accelerator
- Bonnes pratiques Amazon Aurora
- Performances Amazon Redshift
- Les 10 meilleures techniques pour améliorer les performances d'Amazon Athena
- Bonnes pratiques Amazon Redshift Spectrum
- Bonnes pratiques Amazon DynamoDB
- Choisir entre Amazon EC2 et Amazon RDS (langue française non garantie)
- Bonnes pratiques d'implémentation pour Amazon ElastiCache

Vidéos connexes :

- En savoir plus sur Amazon EBS
- Optimisez vos performances de stockage avec Amazon S3
- Moderniser les applications avec les bases de données sur mesure
- Stockage démystifié : comment tout cela fonctionne Amazon Aurora
- Plongée approfondie : modèles de conception avancés Amazon DynamoDB

Exemples connexes :

- Pilote CSI Amazon EFS
- Pilote CSI Amazon EBS
- Utilitaires Amazon EFS
- Amazon EBS Autoscale

- Exemples Amazon S3
- Optimiser le modèle de données à l'aide du partage de données Amazon Redshift
- Migrations des bases de données
- MS SQL Server - Démonstration de réplication AWS Database Migration Service (AWS DMS) (langue française non garantie)
- Atelier pratique sur la modernisation des bases de données
- Exemples Amazon Neptune

PERF03-BP02 Évaluer les options de configuration disponibles pour un magasin de données

Comprenez et évaluez les différentes fonctionnalités et options de configuration disponibles pour vos magasins de données afin d'optimiser l'espace de stockage et les performances de votre charge de travail.

Anti-modèles courants :

- Vous n'utilisez qu'un seul type de stockage (comme Amazon EBS) pour toutes les charges de travail.
- Vous utilisez les IOPS provisionnés pour toutes les charges de travail sans effectuer de test en situation réelle sur tous les niveaux de stockage.
- Vous ne connaissez pas les options de configuration de la solution de gestion de données que vous avez choisie.
- Vous vous concentrez uniquement sur l'augmentation de la taille de l'instance sans examiner les autres options de configuration disponibles.
- Vous ne testez pas les caractéristiques de mise à l'échelle de votre magasin de données.

Avantages liés au respect de cette bonne pratique : En explorant et en expérimentant les configurations de magasin de données, vous pourriez réduire le coût de l'infrastructure, améliorer les performances et réduire l'effort requis pour maintenir vos charges de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Une charge de travail peut comporter un ou plusieurs magasins de données utilisés en fonction des exigences de stockage des données et d'accès aux données. Pour optimiser l'efficacité et le coût

de vos performances, vous devez évaluer les modèles d'accès aux données afin de déterminer les configurations de magasin de données appropriées. Pendant que vous explorez les options de magasin de données, tenez compte de divers aspects tels que les options de stockage, la mémoire, le calcul, le réplica en lecture, les exigences de cohérence, le regroupement de connexions et les options de mise en cache. Testez ces différentes options de configuration pour améliorer les métriques d'efficacité des performances.

Étapes d'implémentation

- Comprenez les configurations actuelles (comme le type d'instance, la taille de stockage ou la version du moteur de base de données) de votre magasin de données.
- Consultez la documentation AWS et les bonnes pratiques pour en savoir plus sur les options de configuration recommandées qui peuvent vous aider à améliorer les performances de votre magasin de données. Les principales options de magasin de données à prendre en compte sont les suivantes :

Option de configuration	Exemples
Déchargement des lectures (comme les réplicas en lecture et la mise en cache)	• Pour les tables DynamoDB, vous pouvez décharger les lectures à l'aide de DAX pour la mise en cache. • Vous pouvez créer un cluster Amazon ElastiCache (Redis OSS) pour Redis et configurer votre application pour qu'elle lise d'abord les données à partir du cache, en revenant à la base de données si l'élément demandé n'est pas présent. • Les bases de données relationnelles comme Amazon RDS et Aurora et les bases de données NoSQL allouées telles que Neptune et Amazon DocumentDB prennent toutes en charge l'ajout de réplicas en lecture pour décharger les parties lues de la charge de travail. • Les bases de données sans serveur comme DynamoDB se mettent à l'échelle automatiq

Option de configuration	Exemples
	uement. Assurez-vous que vous disposez de suffisamment d'unités de capacité de lecture (RCU) allouées pour gérer la charge de travail.

Option de configuration	Exemples
Mise à l'échelle des écritures (comme le partitionnement des clés de partition ou l'introduction d'une file d'attente)	• Pour les bases de données relationnelles, vous pouvez augmenter la taille de l'instanc e pour qu'elle s'adapte à une charge de travail accrue ou augmenter les IOPS provisionnés pour permettre un débit accru vers le stockage sous-jacent. • Vous pouvez également ajouter une file d'attente devant votre base de données plutôt que d'écrire directement dans la base de données. Ce modèle vous permet de dissocier l'ingestion de la base de données et de contrôler le débit afin que la base de données ne soit pas submergée. • Regrouper vos demandes d'écriture plutôt que de créer de nombreuses transactions de courte durée contribue à améliorer le débit dans les bases de données relationn elles à volume d'écriture élevé. • Les bases de données sans serveur comme DynamoDB peuvent mettre à l'échelle le débit d'écriture automatiquement ou en ajustant les unités de capacité d'écriture allouées (WCU) en fonction du mode de capacité. • Vous pouvez toujours rencontrer des problèmes avec les partitions à chaud lorsque vous atteignez les limites de débit pour une clé de partition donnée. Pour pallier à ce problème, choisissez une clé de partition distribuée plus uniformément ou partitionnez en écriture la clé de partition.

Option de configuration	Exemples
Politiques pour gérer le cycle de vie de vos jeux de données	• Vous pouvez utiliser Amazon S3 Lifecycle afin de gérer vos objets tout au long de leur cycle de vie. Si vos modèles d'accès sont inconnus, changeants ou imprévisibles, vous pouvez utiliser Amazon S3 Intelligent-Tiering, qui surveille les modèles d'accès et déplace automatiquement les objets auxquels il n'a pas été possible d'accéder vers des niveaux d'accès moins coûteux. Vous pouvez utiliser les métriques Amazon S3 Storage Lens afin d'identifier les possibilités d'optimisation et les écarts dans la gestion du cycle de vie. • Gestion du cycle de vie Amazon EFS gère automatiquement le stockage des fichiers pour vos systèmes de fichiers.
Gestion et regroupement des connexions	• Amazon RDS Proxy peut être utilisé avec Amazon RDS et Aurora pour gérer les connexions à la base de données. • Les bases de données sans serveur comme DynamoDB n'ont pas de connexions associées, mais tenez compte de la capacité allouée et des politiques de mise à l'échelle automatique pour faire face aux pics de charge.

- Réalisez des tests et procédez au benchmarking dans un environnement hors production afin d'identifier l'option de configuration qui répond à vos exigences en termes de charge de travail.
- Après avoir réalisé vos tests, planifiez votre migration et validez vos métriques de performance.
- Utilisez les outils de surveillance AWS (tels que Amazon CloudWatch) et d'optimisation (tels que Amazon S3 Storage Lens) pour optimiser en continu votre magasin de données à l'aide d'un modèle d'utilisation réel.

Ressources

Documents connexes :

- Stockage cloud avec AWS
- Types de volume Amazon EBS
- Stockage Amazon EC2
- Amazon EFS : Amazon EFS Performance
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance
- Amazon S3 Glacier : documentation S3 Glacier
- Amazon S3 : directives en matière de débit de demandes et de performances
- Stockage cloud avec AWS
- Stockage cloud avec AWS
- Caractéristiques d'E/S Amazon EBS
- Bases de données cloud avec AWS
- Mise en cache de bases de données AWS
- DynamoDB Accelerator
- Bonnes pratiques Amazon Aurora
- Performances Amazon Redshift
- Les 10 meilleures techniques pour améliorer les performances d'Amazon Athena
- Bonnes pratiques Amazon Redshift Spectrum
- Bonnes pratiques Amazon DynamoDB

Vidéos connexes :

- En savoir plus sur Amazon EBS
- Optimisez vos performances de stockage avec Amazon S3
- Moderniser les applications avec les bases de données sur mesure
- Stockage démystifié : comment tout cela fonctionne Amazon Aurora
- Plongée approfondie : modèles de conception avancés Amazon DynamoDB

Exemples connexes :

- Pilote CSI Amazon EFS
- Pilote CSI Amazon EBS
- Utilitaires Amazon EFS
- Amazon EBS Autoscale
- Exemples Amazon S3
- Exemples Amazon DynamoDB
- Exemples de migration de base de données AWS
- Atelier sur la modernisation des bases de données
- Utilisation des paramètres de votre instance de base de données Amazon RDS for PostgreSQL

PERF03-BP03 Collecter et archiver les métriques de performance du magasin de données

Suivez et archivez les métriques de performance pertinentes pour votre magasin de données afin de comprendre comment fonctionnent vos solutions de gestion des données. Ces métriques peuvent vous aider à optimiser votre magasin de données, à vérifier que les exigences de votre charge de travail sont satisfaites et à fournir une vue d'ensemble claire sur le fonctionnement de la charge de travail.

Anti-modèles courants :

- Vous utilisez uniquement la recherche manuelle des fichiers journaux pour les métriques.
- Vous publiez uniquement des métriques sur les outils internes utilisés par votre équipe et vous n'avez pas une visibilité complète de votre charge de travail.
- Vous n'utilisez que les métriques par défaut enregistrées par le logiciel de surveillance que vous avez sélectionné.
- Vous n'examinez les métriques qu'en cas de problème.
- Vous ne surveillez que les métriques au niveau du système et vous ne capturez pas les métriques d'accès aux données ou d'utilisation des données.

Avantages liés au respect de cette bonne pratique : La définition de points de référence pour les performances vous permet de mieux comprendre le comportement normal et les exigences des charges de travail. Les modèles anormaux peuvent être identifiés et débogués plus rapidement, ce qui améliore les performances et la fiabilité du magasin de données.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

L'enregistrement de plusieurs métriques de performance sur une période donnée est nécessaire pour la surveillance des performances de vos magasins de données. Cette surveillance vous permet non seulement de détecter les anomalies, mais aussi d'évaluer les performances par rapport aux métriques métier afin de vérifier que vous répondez aux besoins de votre charge de travail.

Ces métriques doivent inclure à la fois le système sous-jacent qui prend en charge le magasin de données et les métriques de la base de données. Les métriques système sous-jacentes peuvent inclure l'utilisation du processeur, la mémoire, le stockage sur disque disponible, les E/S de disque, le taux d'accès au cache et les métriques entrantes et sortantes du réseau, tandis que les métriques du magasin de données peuvent inclure les transactions par seconde, les principales requêtes, les taux de requêtes moyens, les temps de réponse, l'utilisation de l'index, les verrouillages de table, les délais d'expiration des requêtes et le nombre de connexions ouvertes. Ces données sont essentielles pour comprendre comment fonctionne la charge de travail et comment la solution de gestion des données est utilisée. Utilisez ces métriques dans le cadre d'une approche fondée sur les données pour ajuster et optimiser les ressources de votre charge de travail.

Utilisez des outils, des bibliothèques et des systèmes qui enregistrent des mesures de performances liées aux performances de la base de données.

Étapes d'implémentation

1. Identifiez les métriques de performance clés que votre magasin de données doit suivre.
    a. Amazon S3 Métriques et dimensions
    b. Surveillance des métriques pour une instance Amazon RDS
    c. Surveillance de la charge de base de données avec Performance Insights sur Amazon RDS
    d. Présentation de la surveillance améliorée
    e. DynamoDB Métriques et dimensions
    f. Surveillance de DynamoDB Accelerator
    g. Surveillance de Amazon MemoryDB avec Amazon CloudWatch
    h. Quelles métriques dois-je surveiller ?
    i. Surveillance des performances du cluster Amazon Redshift
    j. Timestream Métriques et dimensions
    k. Métriques Amazon CloudWatch pour Amazon Aurora
    l. Journalisation et surveillance dans Amazon Keyspaces (for Apache Cassandra)

   m. Surveillance des ressources Amazon Neptune
2. Utilisez une solution de journalisation et de surveillance approuvée pour collecter ces métriques. Amazon CloudWatch peut récupérer des métriques à partir des ressources de votre architecture. Vous pouvez également récupérer et publier des métriques personnalisées pour faire apparaître des métriques d'entreprise ou des métriques dérivées. Utilisez CloudWatch ou des solutions tierces pour définir des alarmes qui indiquent les dépassements de seuils.
3. Vérifiez si la surveillance du magasin de données peut bénéficier d'une solution de machine learning qui détecte les anomalies de performance.
   a. Amazon DevOps Guru pour Amazon RDS assure la visibilité des problèmes de performance et suggère des actions correctives.
4. Configurez la conservation des données dans votre solution de surveillance et de journalisation en fonction de vos objectifs sécuritaires et opérationnels.
   a. Métriques de conservation des données pour CloudWatch
   b. Conservation des données pour CloudWatch Logs

Ressources

Documents connexes :

- Mise en cache de bases de données AWS
- Les 10 meilleures techniques pour améliorer les performances d'Amazon Athena
- Bonnes pratiques Amazon Aurora
- DynamoDB Accelerator
- Bonnes pratiques Amazon DynamoDB
- Bonnes pratiques Amazon Redshift Spectrum
- Performances Amazon Redshift
- Bases de données cloud avec AWS
- Amazon RDS Performance Insights

Vidéos connexes :

- Bases de données sur mesure AWS
- Stockage démystifié : comment tout cela fonctionne Amazon Aurora
- Plongée approfondie : modèles de conception avancés Amazon DynamoDB

- Bonnes pratiques pour la surveillance des charges de travail Redis sur Amazon ElastiCache

Exemples connexes :

- Niveau 100 : surveillance avec les tableaux de bord CloudWatch
- Infrastructure de collecte de métriques d'ingestion de jeux de données AWS
- Atelier sur la surveillance Amazon RDS

PERF03-BP04 Mettre en œuvre des stratégies pour améliorer les performances des requêtes dans un magasin de données

Mettez en œuvre des stratégies pour optimiser les données et améliorer les requêtes sur les données afin de renforcer la capacité de mise à l'échelle et l'efficacité des performances pour votre charge de travail.

Anti-modèles courants :

- Vous ne partitionnez pas les données dans votre magasin de données.
- Vous ne stockez les données que dans un seul format de fichier dans votre magasin de données.
- Vous n'utilisez pas d'index dans votre magasin de données.

Avantages liés au respect de cette bonne pratique : En optimisant les performances des données et des requêtes, vous augmentez leur efficacité, vous réduisez les coûts et vous améliorez l'expérience utilisateur.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

L'optimisation des données et des requêtes sont des aspects essentiels de l'efficacité des performances d'un magasin de données, car ils ont un impact sur les performances et la réactivité de l'ensemble de la charge de travail dans le cloud. Les données non optimisées peuvent augmenter l'utilisation des ressources et les goulots d'étranglement, ce qui réduit l'efficacité globale d'un magasin de données.

L'optimisation des données inclut plusieurs techniques pour garantir un stockage des données et un accès aux données efficaces. Cela permet également d'améliorer les performances des requêtes dans un magasin de données. Les principales stratégies incluent le partitionnement des données, la

compression des données et la dénormalisation des données, qui permettent d'optimiser les données à la fois pour le stockage et l'accès.

Étapes d'implémentation

- Comprenez et analysez les requêtes essentielles sur les données effectuées dans votre magasin de données.
- Identifiez les requêtes lentes dans votre magasin de données et utilisez des plans de requêtes pour comprendre leur état actuel.
  - Analyse du plan de requête dans Amazon Redshift
  - Utilisation d'EXPLAIN et EXPLAIN ANALYZE dans Athena (langue française non garantie)
- Mettez en œuvre des stratégies pour améliorer les performances des requêtes. Les stratégies clés incluent :
  - L'utilisation d'un format de fichier en colonnes (comme Parquet ou ORC).
  - La compression des données dans le magasin de données pour réduire l'espace de stockage et les opérations d'E/S.
  - Le partitionnement des données pour diviser les données en parties plus petites et réduire le temps d'analyse des données.
    - Partitionnement des données dans Athena (langue française non garantie)
    - Partitions et distribution de données
  - L'indexation des données sur les colonnes communes de la requête.
  - Choisissez l'opération de jointure appropriée pour la requête. Lorsque vous joignez deux tables, spécifiez la table la plus grande sur le côté gauche de la jointure et la plus petite sur le côté droit de la jointure.
  - La solution de mise en cache distribuée pour améliorer la latence et réduire le nombre d'opérations d'E/S dans la base de données.
  - La maintenance régulière, comme l'exécution de statistiques.
- Expérimentez et testez les stratégies dans un environnement hors production.

Ressources

Documents connexes :

- Bonnes pratiques Amazon Aurora
- Performances Amazon Redshift

- Les 10 meilleures techniques pour améliorer les performances d'Amazon Athena
- Mise en cache de bases de données AWS
- Bonnes pratiques d'implémentation pour Amazon ElastiCache
- Partitionnement des données dans Athena (langue française non garantie)

Vidéos connexes :

- Optimiser le modèle de données à l'aide du partage de données Amazon Redshift
- Optimisez les requêtes Amazon Athena grâce aux nouveaux outils d'analyse des requêtes

Exemples connexes :

- Pilote CSI Amazon EFS

PERF03-BP05 Mise en œuvre de modèles d'accès aux données utilisant la mise en cache

Mettez en œuvre des modèles d'accès qui peuvent tirer parti de la mise en cache des données pour une récupération rapide des données fréquemment consultées.

Anti-modèles courants :

- Vous mettez en cache des données qui changent fréquemment.
- Vous utilisez les données mises en cache comme si elles étaient stockées de manière durable et toujours disponibles.
- Vous ne tenez pas compte de la cohérence de vos données mises en cache.
- Vous ne surveillez pas l'efficacité de la mise en œuvre de la mise en cache.

Avantages liés au respect de cette bonne pratique : Le stockage des données dans un cache contribue à améliorer la latence et le débit de lecture, l'expérience utilisateur et l'efficacité globale, tout en réduisant les coûts.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Un cache est un composant logiciel ou matériel destiné à stocker des données afin que les requêtes futures portant sur les mêmes données puissent être traitées plus rapidement ou plus efficacement.

Les données stockées dans un cache peuvent être reconstruites en cas de perte en répétant un calcul antérieur ou en les récupérant dans un autre magasin de données.

La mise en cache des données peut être l'une des stratégies les plus efficaces pour améliorer les performances globales de votre application et réduire la charge qui pèse sur vos sources de données principales sous-jacentes. Les données peuvent être mises en cache à plusieurs niveaux dans l'application, par exemple dans l'application effectuant des appels à distance et également connue sous le nom de mise en cache côté client, ou en utilisant un service secondaire rapide pour stocker les données, ce que l'on appelle aussi mise en cache à distance.

Mise en cache côté client

Grâce à la mise en cache côté client, chaque client (une application ou un service qui interroge le magasin de données backend) peut stocker les résultats de ses requêtes uniques localement pendant une durée spécifiée. Cela permet de réduire le nombre de requêtes adressées à un magasin de données sur le réseau en vérifiant d'abord le cache du client local. En l'absence de résultats, l'application peut alors interroger le magasin de données et stocker ces résultats localement. Ce modèle permet à chaque client de stocker les données dans l'emplacement le plus proche possible (le client lui-même), ce qui se traduit par la latence la plus faible possible. Les clients peuvent également continuer à répondre à certaines requêtes lorsque le magasin de données backend n'est pas disponible, ce qui augmente la disponibilité de l'ensemble du système.

L'un des inconvénients de cette approche est que lorsque plusieurs clients sont impliqués, ils peuvent stocker les mêmes données mises en cache localement. Cela entraîne à la fois une double utilisation du stockage et une incohérence des données entre ces clients. Un client peut mettre en cache les résultats d'une requête et, une minute plus tard, un autre client peut exécuter la même requête et obtenir un résultat différent.

Mise en cache à distance

Pour résoudre le problème de duplication des données entre clients, un service externe rapide, ou cache distant, peut être utilisé pour stocker les données interrogées. Au lieu de vérifier un magasin de données local, chaque client vérifie le cache distant avant d'interroger le magasin de données backend. Cette stratégie permet d'obtenir des réponses plus cohérentes entre les clients, d'améliorer l'efficacité des données stockées et d'augmenter le volume de données mises en cache, car l'espace de stockage évolue indépendamment des clients.

L'inconvénient d'un cache distant est que l'ensemble du système peut connaître une latence plus élevée, car un saut de réseau à réseau supplémentaire est nécessaire pour vérifier le cache distant.

La mise en cache côté client peut être utilisée parallèlement à la mise en cache à distance pour une mise en cache à plusieurs niveaux afin d'améliorer la latence.

Étapes d'implémentation

1. Identifiez les bases de données, les API et les services réseau susceptibles de bénéficier de la mise en cache. Les services dont la charge de travail de lecture est importante, qui ont un ratio lecture/écriture élevé ou qui sont coûteux à adapter conviennent à la mise en cache.
   - Mise en cache de bases de données
   - Activez la mise en cache des API pour améliorer la réactivité.
2. Identifiez le type de stratégie de mise en cache le mieux adapté à votre modèle d'accès.
   - Stratégies de mise en cache
   - Solutions de mise en cache AWS
3. Suivez les bonnes pratiques en matière de mise en cache pour votre magasin de données.
4. Configurez une stratégie d'invalidation du cache, telle qu'une durée de vie (TTL), pour toutes les données afin d'équilibrer la fraîcheur des données et de réduire la pression qui pèse sur le magasin de données backend.
5. Activez des fonctionnalités telles que les nouvelles tentatives de connexion automatiques, le backoff exponentiel, les délais d'attente côté client et le regroupement des connexions dans le client, le cas échéant, car elles peuvent améliorer les performances et la fiabilité.
   - Bonnes pratiques : clients de Redis et Amazon ElastiCache (Redis OSS)
6. Surveillez le taux d'accès au cache en visant un objectif de 80 % ou plus. Des valeurs inférieures peuvent indiquer une taille de cache insuffisante ou un modèle d'accès qui ne bénéficie pas de la mise en cache.
   - Quelles métriques dois-je surveiller ?
   - Bonnes pratiques pour la surveillance des charges de travail Redis sur Amazon ElastiCache
   - Bonnes pratiques de surveillance avec Amazon ElastiCache (Redis OSS) via Amazon CloudWatch
7. Implémentez la réplication des données pour transférer les lectures vers plusieurs instances et améliorer les performances et la disponibilité de lecture des données.

Ressources

Documents connexes :

- Utilisation d'Amazon ElastiCache Well-Architected Lens
- Bonnes pratiques de surveillance avec Amazon ElastiCache (Redis OSS) via Amazon CloudWatch
- Quelles métriques dois-je surveiller ?
- Livre blanc « Performances à grande échelle avec Amazon ElastiCache »
- Défis et stratégies en matière de mise en cache

Vidéos connexes :

- Amazon ElastiCache Learning Path
- Design for success with Amazon ElastiCache best practices

Exemples connexes :

- Améliorer les performances des bases de données MySQL avec Amazon ElastiCache (Redis OSS)

Mise en réseau et diffusion de contenu

PERF 4. Comment sélectionnez-vous et configurez-vous les ressources de mise en réseau de votre charge de travail ?

La solution de base de données la plus efficace pour un système varie en fonction des exigences de cohérence, de disponibilité, de tolérance des partitions, de latence, de durabilité, d'évolutivité et de capacités de requête. De nombreux systèmes utilisent des solutions de base de données différentes pour divers sous-systèmes et activent des fonctionnalités distinctes pour améliorer les performances. La sélection d'une solution de base de données et de fonctionnalités incorrectes pour un système peut conduire à une efficacité moindre des performances.

Bonnes pratiques

- PERF04-BP01 Compréhension de l'impact de la mise en réseau sur les performances
- PERF04-BP02 Évaluation des fonctionnalités de mise en réseau disponibles
- PERF04-BP03 Choix d'une connectivité dédiée ou d'un VPN approprié pour votre charge de travail
- PERF04-BP04 Utilisation de l'équilibrage de charge pour répartir le trafic entre plusieurs ressources
- PERF04-BP05 Choix de protocoles réseau afin d'améliorer les performances

- PERF04-BP06 Choix du placement de votre charge de travail en fonction des exigences réseau
- PERF04-BP07 Optimisation de la configuration réseau en fonction de métriques

PERF04-BP01 Compréhension de l'impact de la mise en réseau sur les performances

Analysez et comprenez l'impact des décisions liées au réseau sur votre charge de travail afin de fournir des performances efficaces et une meilleure expérience utilisateur.

Anti-modèles courants :

- Tout le trafic passe par vos centres de données existants.
- Vous acheminez l'ensemble du trafic via des pare-feux centralisés au lieu d'utiliser des outils de sécurité réseau natifs cloud.
- Vous configurez des connexions AWS Direct Connect sans connaître les exigences d'utilisation réelles.
- Vous ne tenez pas compte des caractéristiques de la charge de travail et de la surcharge de chiffrement lors de la définition de vos solutions de mise en réseau.
- Vous utilisez des concepts et des stratégies sur site pour les solutions de mise en réseau dans le cloud.

Avantages liés au respect de cette bonne pratique : Comprendre comment la mise en réseau affecte les performances de la charge de travail vous aidera à identifier les goulots d'étranglement potentiels, à améliorer l'expérience utilisateur, à accroître la fiabilité et à réduire la maintenance opérationnelle à mesure que la charge de travail évolue.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Le réseau est responsable de la connectivité entre les composants d'application, les services cloud, les réseaux périphériques et les données sur site et, par conséquent, il peut avoir un impact majeur sur les performances de la charge de travail. Outre les performances de la charge de travail, l'expérience utilisateur peut également être affectée par la latence du réseau, la bande passante, les protocoles, l'emplacement, la congestion du réseau, l'instabilité, le débit et les règles de routage.

Avoir une liste documentée des exigences de mise en réseau de la charge de travail, y compris la latence, la taille des paquets, les règles de routage, les protocoles et les modèles de trafic pris

en charge. Passez en revue les solutions de mise en réseau disponibles et identifiez le service qui répond aux caractéristiques de mise en réseau de votre charge de travail. Les réseaux basés sur le cloud peuvent être rapidement recréés. L'évolution de votre architecture réseau au fil du temps est donc nécessaire pour améliorer l'efficacité des performances.

Étapes d'implémentation :

1. Définissez et documentez les exigences de performance réseau, y compris les métriques tels que la latence du réseau, la bande passante, les protocoles, les emplacements, les modèles de trafic (pics et fréquence), le débit, le chiffrement, l'inspection et les règles de routage.
2. Découvrez les principaux services AWS de mise en réseau tels que les VPC, AWS Direct Connect, Elastic Load Balancing (ELB)et Amazon Route 53.
3. Capturez les principales caractéristiques réseau suivantes :

Caractéristiques	Outils et métriques
Caractéristiques de mise en réseau fondamentales	• Journaux de flux VPC • Journaux de flux AWS Transit Gateway • Métriques AWS Transit Gateway • Métriques AWS PrivateLink
Caractéristiques de mise en réseau des applications	• Elastic Fabric Adapter • Métriques AWS App Mesh • Métriques Amazon API Gateway
Caractéristiques de mise en réseau à la périphérie	• Métriques Amazon CloudFront • Métriques Amazon Route 53 • Métriques AWS Global Accelerator
Caractéristiques de mise en réseau hybride	• Métriques AWS Direct Connect • Métriques AWS Site-to-Site VPN • Métriques AWS Client VPN • Métriques AWS Cloud WAN
Caractéristiques de mise en réseau de la sécurité	• Métriques AWS Shield, AWS WAF et AWS Network Firewall

Caractéristiques	Outils et métriques
Caractéristiques de traçage	• AWS X-Ray • VPC Reachability Analyzer • Network Access Analyzer • Amazon Inspector • Amazon CloudWatch RUM

4. Définir des points de référence et et tester les performances du réseau :
   a. Évaluez le débit du réseau, car certains facteurs peuvent affecter les performances du réseau Amazon EC2 lorsque les instances se trouvent dans le même VPC. Mesurez la bande passante du réseau entre les instances Amazon EC2 Linux dans le même VPC.
   b. Effectuez des tests de charge pour expérimenter des solutions et des options de mise en réseau.

Ressources

Documents connexes :

- Application Load Balancer
- Mise en réseau améliorée d'EC2 sous Linux
- Capacité réseau améliorée d'EC2 sous Windows
- Groupes de placement EC2
- Activation de la mise en réseau améliorée avec un adaptateur réseau élastique (ENA) sur les instances de Linux
- Network Load Balancer
- Mise en réseau de produits avec AWS
- Transit Gateway
- Transition vers le routage basé sur la latence dans Amazon Route 53
- Points de terminaison d'un VPC
- Journaux de flux VPC

Vidéos connexes :

- Connectivity to AWS and hybrid AWS network architectures
- Optimizing Network Performance for Amazon EC2 Instances
- Improve Global Network Performance for Applications
- EC2 Instances and Performance Optimization Best Practices
- Optimizing Network Performance for Amazon EC2 Instances
- Networking best practices and tips with the Well-Architected Framework
- AWS networking best practices in large-scale migrations

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- Ateliers sur la mise en réseau AWS

PERF04-BP02 Évaluation des fonctionnalités de mise en réseau disponibles

Évaluez les fonctions de mise en réseau dans le cloud qui peuvent améliorer les performances. Mesurez l'impact de ces fonctions au moyen de tests, de métriques et de l'analyse. Par exemple, tirez parti des fonctionnalités au niveau du réseau qui sont disponibles pour réduire la latence, la distance réseau ou l'instabilité.

Anti-modèles courants :

- Vous restez au sein d'une même région, car c'est là que votre siège social se trouve physiquement.
- Vous utilisez des pare-feux plutôt que des groupes de sécurité pour filtrer le trafic.
- Vous enfreignez le protocole TLS pour inspecter le trafic plutôt que de vous fier aux groupes de sécurité, aux politiques relatives aux points de terminaison et à d'autres fonctionnalités natives cloud.
- Vous utilisez uniquement la segmentation basée sur un sous-réseau au lieu des groupes de sécurité.

Avantages liés au respect de cette bonne pratique : L'évaluation de toutes les options et fonctionnalités de service peut augmenter les performances de vos charges de travail, baisser le coût d'infrastructure, réduire les efforts nécessaires à la maintenance de vos charges de travail et améliorer votre posture générale en matière de sécurité. Vous pouvez utiliser la couverture mondiale d'AWS pour fournir à vos clients une expérience de mise en réseau optimale.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

AWS propose des services comme AWS Global Accelerator et Amazon CloudFront qui contribuent à améliorer les performances du réseau, alors que la plupart des services AWS comportent des fonctionnalités de produit (telles que Amazon S3 Transfer Acceleration) pour optimiser le trafic réseau.

Examinez les options de configuration liées au réseau disponibles et leur impact potentiel sur votre charge de travail. L'optimisation des performances dépend de la compréhension de la manière dont ces options interagissent avec votre architecture et de l'impact qu'elles auront à la fois sur les performances mesurées et sur l'expérience utilisateur.

Étapes d'implémentation

- Créer une liste des composants de la charge de travail.
  - Envisagez d'utiliser AWS Cloud WAN pour créer, gérer et surveiller le réseau de votre organisation lors de la création d'un réseau mondial unifié.
  - Surveillez vos réseaux mondiaux et principaux avec les métriques Amazon CloudWatch Logs. Exploitez Amazon CloudWatch RUM, qui fournit des informations permettant d'identifier, de comprendre et d'améliorer l'expérience numérique des utilisateurs.
  - Visualisez la latence réseau globale entre les Régions AWS et les zones de disponibilité, et au sein de chaque zone de disponibilité, avec AWS Network Manager pour mieux comprendre comment les performances de votre application sont liées aux performances du réseau AWS sous-jacent.
  - Utilisez un outil ou un service de base de données de gestion de la configuration (CMDB) comme AWS Config pour créer un inventaire de votre charge de travail et de sa configuration.
- Identifier et documenter le test comparatif pour vos métriques de performances s'il s'agit d'une charge de travail existante, en vous concentrant sur les goulots d'étranglement et les zones à améliorer. Les métriques de mise en réseau liées aux performances diffèrent par charge de travail en fonction des exigences métier et des caractéristiques de charge de travail. Pour commencer, il pourrait être important d'examiner ces métriques pour votre charge de travail : bande passante, latence, perte de paquets, instabilité et retransmissions.
- S'il s'agit d'une nouvelle charge de travail, réaliser des tests de charge pour identifier les goulots d'étranglement au niveau des performances.

- Concernant l'identification des goulots d'étranglement au niveau des performances, examiner les options de configuration pour les solutions afin d'identifier les opportunités d'amélioration des performances. Découvrez les principales options et fonctionnalités de mise en réseau suivantes :

Opportunité d'amélioration	Solution
Chemin ou itinéraires réseau	Utilisez Network Access Analyzer pour identifier des chemins ou des itinéraires.
Protocoles réseau	Voir PERF04-BP05 Choix de protocoles réseau afin d'améliorer les performances
Topologie du réseau	Évaluez vos compromis de performances et opérationnels entre Appairage des VPC et AWS Transit Gateway lors de la connexion de plusieurs comptes. AWS Transit Gateway simplifie la façon dont vous interconnectez tous vos VPC, qui peuvent s'étendre sur des milliers de Comptes AWS et sur les réseaux sur site. Partagez votre AWS Transit Gateway entre plusieurs comptes à l'aide de AWS Resource Access Manager. Voir PERF04-BP03 Choix d'une connectivité dédiée ou d'un VPN approprié pour votre charge de travail

Opportunité d'amélioration	Solution
Services de réseau	AWS Global Accelerator est un service qui améliore de 60 % les performances du trafic réseau de vos utilisateurs grâce à l'infrast ructure réseau mondiale AWS. Amazon CloudFront contribue à améliorer les performances de votre charge de travail, de diffusion de contenu et de latence à l'échelle mondiale. Utilisez Lambda@edge pour exécuter des fonctions qui personnalisent le contenu diffusé par CloudFront au plus près des utilisateurs, réduire la latence et améliorer les performan ces. Amazon Route 53 offre des options de routage basé sur la latence, routage de géolocalisation, routage de proximité géographiqueet routage basé sur IP pour vous permettre d'améliorer les performan ces de votre charge de travail pour satisfair e un public international. Identifiez l'option de routage qui optimiserait les performances de votre charge de travail en examinant le trafic de votre charge de travail et la localisation des utilisateurs lorsque votre charge de travail est distribuée dans le monde entier.

Opportunité d'amélioration	Solution
Fonctionnalités des ressources de stockage	Amazon S3 Transfer Acceleration est une fonction qui permet aux utilisateurs externes de bénéficier des optimisations de mise en réseau de CloudFront pour charger des données dans Amazon S3. Cela améliore le transfert d'importants volumes de données à partir d'emplacements distants qui n'ont pas de connectivité dédiée au AWS Cloud. Les points d'accès multi-régions dans Amazon S3 répliquent le contenu vers plusieurs régions et simplifient la charge de travail en fournissant un point d'accès. Lorsqu'un point d'accès multi-région est utilisé, vous pouvez demander ou écrire des données à Amazon S3 tandis que le service identifie le compartim ent à la latence la plus faible.

Opportunité d'amélioration	Solution
Fonctionnalités des ressources informatiques	Les interfaces réseau Elastic (ENI) utilisées par des instances Amazon EC2, des conteneurs et des fonctions Lambda sont limitées par flux. Examinez vos groupes de placement pour optimiser votre débit de mise en réseau EC2. Pour éviter un goulot d'étranglement par flux, créez votre applicati on pour qu'elle utilise plusieurs flux. Pour surveiller et disposer d'une visibilité sur vos métriques de mise en réseau liée au calcul, utilisez les métriques CloudWatch et ethtool. La commande `ethtool` est incluse dans le pilote ENA et expose des métriques liées au réseau supplémentaires qui peuvent être publiées en tant que métriques personnalisées sur CloudWatch. Les adaptateurs réseau élastiques (ENA) fournissent une optimisation supérieure en offrant un meilleur débit pour vos instances dans un groupe de placement du cluster. Un Elastic Fabric Adapter (EFA) est une interface réseau pour les instances Amazon EC2 qui vous permet d'exécuter des charges de travail nécessitant des niveaux élevés de communication entre les nœuds à grande échelle sur AWS. Les instances optimisées Amazon EBS utilisent une pile de configuration optimisée et offrent une capacité dédiée supplémentaire pour les E/S Amazon EBS.

Ressources

Documents connexes :

- Amazon EBS – Instances optimisées
- Application Load Balancer
- Mise en réseau améliorée d'EC2 sous Linux
- Capacité réseau améliorée d'EC2 sous Windows
- Groupes de placement EC2
- Activation de la mise en réseau améliorée avec un adaptateur réseau élastique (ENA) sur les instances de Linux
- Network Load Balancer
- Mise en réseau de produits avec AWS
- AWS Transit Gateway
- Transition vers le routage basé sur la latence dans Amazon Route 53
- Points de terminaison d'un VPC
- Journaux de flux VPC

Vidéos connexes :

- Connectivity to AWS and hybrid AWS network architectures
- Optimizing Network Performance for Amazon EC2 Instances
- AWS Global Accelerator

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- Ateliers sur la mise en réseau AWS

PERF04-BP03 Choix d'une connectivité dédiée ou d'un VPN approprié pour votre charge de travail

Lorsque la connectivité hybride est requise pour connecter des ressources sur site et dans le cloud, allouez une bande passante adéquate pour répondre à vos exigences de performance. Estimez les exigences en matière de bande passante et de latence pour votre charge de travail hybride. Ces chiffres détermineront vos exigences en matière de dimensionnement.

Anti-modèles courants :

- Vous n'évaluez les solutions VPN que pour les exigences de chiffrement de votre réseau.
- Vous n'évaluez pas les options de sauvegarde ni de connectivité redondante.
- Vous n'identifiez pas toutes les exigences de la charge de travail (chiffrement, protocole, bande passante et trafic requis).

Avantages liés au respect de cette bonne pratique : La sélection et la configuration de solutions de connectivité appropriées renforcent la fiabilité de votre charge de travail et optimisent les performances. En identifiant les exigences de la charge de travail, en effectuant une planification appropriée et en évaluant les solutions hybrides, vous pouvez minimiser les modifications coûteuses du réseau physique et les frais généraux opérationnels tout en accélérant le délai de rentabilisation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Développez une architecture de mise en réseau hybride basée sur vos besoins en bande passante. AWS Direct Connect vous permet de connecter votre réseau sur site en privé à AWS. Cette solution convient lorsque vous avez besoin d'une bande passante élevée et d'une faible latence tout en conservant des performances constantes. Une connexion VPN établit une connexion sécurisée sur Internet. Elle sert uniquement lorsque seule une connexion temporaire est requise, lorsque le coût est un facteur, ou en cas d'urgence en attendant qu'une connectivité réseau physique résiliente soit établie lors de l'utilisation d'AWS Direct Connect.

Si vos besoins en bande passante sont élevés, vous pouvez envisager divers services AWS Direct Connect ou VPN. Le trafic peut être équilibré entre les services, mais nous ne recommandons pas l'équilibrage de charge entre AWS Direct Connect et le VPN en raison des différences de latence et de bande passante.

Étapes d'implémentation

1. Évaluez les besoins en bande passante et en latence de vos applications existantes.
    a. Pour les charges de travail existantes qui migrent vers AWS, exploitez les données de vos systèmes internes de surveillance du réseau.
    b. Pour les nouvelles charges de travail ou pour les charges de travail existantes pour lesquelles vous ne disposez pas de données de suivi, contactez les propriétaires du produit pour obtenir des métriques de performance adéquates et offrir une bonne expérience utilisateur.

2. Sélectionnez une connexion dédiée ou un VPN comme option de connectivité. En fonction de toutes les exigences de la charge de travail (chiffrement, bande passante et trafic requis), vous pouvez choisir AWS Direct Connect ou AWS VPN (ou les deux). Le schéma suivant peut vous aider à choisir le type de connexion approprié.
    a. AWS Direct Connect fournit une connectivité dédiée à l'environnement AWS, de 50 Mbit/s à 100 Gbit/s, en utilisant des connexions dédiées ou des connexions hébergées. Cela vous permet de gérer et de contrôler la latence et de profiter d'une bande passante provisionnée. Ainsi, vos charges de travail peuvent se connecter efficacement à d'autres environnements. Grâce aux partenaires AWS Direct Connect, vous bénéficiez d'une connectivité de bout en bout à partir de plusieurs environnements, ce qui vous permet de disposer d'un réseau étendu aux performances constantes. AWS offre une bande passante de connexion directe évolutive en utilisant soit le débit 100 Gbit/s natif, soit le protocole LAG (Link Aggregation Group), soit le protocole BGP ECMP (Equal-cost multipath).
    b. AWS Site-to-Site VPN fournit un service VPN géré prenant en charge le protocole de sécurité du protocole Internet (IPsec). Lorsqu'une connexion VPN est créée, chaque connexion VPN comprend deux tunnels pour une haute disponibilité.
3. Consultez la documentation AWS pour choisir l'option de connectivité appropriée :
    a. Si vous décidez d'utiliser AWS Direct Connect, sélectionnez la bande passante adaptée à votre connectivité.
    b. Si vous utilisez un AWS Site-to-Site VPN sur plusieurs emplacements pour vous connecter à une Région AWS, utilisez une connexion Site-to-Site VPN accélérée pour pouvoir améliorer les performances du réseau.
    c. Si la conception de votre réseau consiste en une connexion VPN IPsec via AWS Direct Connect, pensez à utiliser un VPN IP privé pour améliorer la sécurité et réaliser une segmentation. Un VPN IP privé AWS Site-to-Site est déployé au-dessus de l'interface virtuelle de transport (VIF).
    d. AWS Direct Connect SiteLink permet de créer des connexions redondantes et à faible latence entre vos centres de données à travers le monde en envoyant les données sur le chemin le plus rapide entre les emplacements AWS Direct Connect, en contournant les Régions AWS.
4. Validez votre configuration de connectivité avant le déploiement en production. Effectuez des tests de sécurité et de performance pour vous assurer qu'elle répond à vos exigences en matière de bande passante, de fiabilité, de latence et de conformité.
5. Surveillez régulièrement les performances et l'utilisation de votre connectivité et optimisez-les si nécessaire.

Organigramme des performances déterministes

Ressources

Documents connexes :

- Network Load Balancer
- Mise en réseau de produits avec AWS
- AWS Transit Gateway
- Transition vers le routage basé sur la latence dans Amazon Route 53
- Points de terminaison d'un VPC
- Site-to-Site VPN
- Création d'une infrastructure réseau AWS multi-VPC évolutive et sécurisée
- AWS Direct Connect
- Client VPN

Vidéos connexes :

- Connectivity to AWS and hybrid AWS network architectures
- Optimizing Network Performance for Amazon EC2 Instances
- AWS Global Accelerator
- AWS Direct Connect
- AWS Transit Gateway Connect
- Solutions VPN
- Sécurité avec les solutions VPN

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- Ateliers sur la mise en réseau AWS

PERF04-BP04 Utilisation de l'équilibrage de charge pour répartir le trafic entre plusieurs ressources

Répartissez le trafic sur plusieurs ressources ou services pour permettre à votre charge de travail de tirer parti de l'élasticité fournie par le cloud. Vous pouvez également utiliser l'équilibrage de charge afin de décharger la terminaison du chiffrement en vue d'améliorer les performances, d'assurer la fiabilité et de gérer et acheminer efficacement le trafic.

Anti-modèles courants :

- Vous ne tenez pas compte des exigences de votre charge de travail lorsque vous choisissez le type d'équilibreur de charge.
- Vous ne tirez pas parti des fonctions d'équilibrage de charge pour optimiser les performances.
- La charge de travail est exposée directement à Internet sans équilibreur de charge.
- Vous acheminez tout le trafic Internet via des équilibreurs de charge existants.
- Vous utilisez l'équilibrage de charge TCP générique et faites en sorte que chaque nœud de calcul gère le chiffrement SSL.

Avantages liés au respect de cette bonne pratique : Un équilibreur de charge gère la charge variable du trafic de votre application dans une seule zone de disponibilité ou entre plusieurs zones de disponibilité et permet une haute disponibilité, une mise à l'échelle automatique et une meilleure utilisation de votre charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les équilibreurs de charge constituent le point d'entrée de votre charge de travail, à partir duquel ils distribuent le trafic vers vos cibles principales, telles que les instances de calcul ou les conteneurs, afin d'améliorer l'utilisation.

Le choix du bon type d'équilibreur de charge est la première étape de l'optimisation de votre architecture. Commencez par énumérer les caractéristiques de votre charge de travail, telles que le protocole (TCP, HTTP, TLS ou WebSockets), le type de cible (instances, conteneurs ou sans serveur), les exigences de l'application (connexions de longue durée, authentification de l'utilisateur ou permanence) et le placement (région, zone locale, Outpost ou isolement de zone).

AWS fournit plusieurs modèles permettant à vos applications d'utiliser l'équilibrage de charge. L'Application Load Balancer est davantage adapté à l'équilibrage de charge du trafic HTTP et HTTPS

et fournit un routage avancé des requêtes, axé sur la diffusion d'architectures d'application modernes, notamment de microservices et de conteneurs.

Le Network Load Balancer est tout indiqué pour l'équilibrage de charge du trafic TCP, qui nécessite des performances extrêmes. Il est capable de traiter des millions de requêtes par seconde tout en maintenant de très faibles latences. Il est optimisé pour gérer les tendances soudaines et instables du trafic.

Elastic Load Balancing assure la gestion intégrée des certificats et le déchiffrement SSL/TLS, ce qui vous permet de gérer de façon centralisée les paramètres SSL de l'équilibreur de charge et de décharger les tâches gourmandes en CPU de votre charge de travail.

Après avoir choisi le bon équilibreur de charge, vous pouvez commencer à tirer parti de ses fonctionnalités pour réduire les efforts que votre système backend doit fournir pour servir le trafic.

Par exemple, en utilisant à la fois Application Load Balancer (ALB) et Network Load Balancer (NLB), vous pouvez effectuer un déchargement du chiffrement SSL/TLS. Cela permet d'éviter que la liaison TLS, très gourmande en ressources CPU, ne soit effectuée par vos cibles, et permet également d'améliorer la gestion des certificats.

Lorsque vous configurez le déchargement SSL/TLS dans votre équilibreur de charge, celui-ci se charge du chiffrement du trafic en provenance et à destination des clients, tout en acheminant le trafic non chiffré vers vos systèmes backend. Cela libère vos ressources backend et améliore le temps de réponse pour les clients.

Application Load Balancer peut également servir le trafic HTTP/2 sans avoir besoin de le prendre en charge sur vos cibles. Cette simple décision peut améliorer le temps de réponse de votre application, car HTTP/2 utilise plus efficacement les connexions TCP.

Les exigences de latence de votre charge de travail doivent être prises en compte lors de la définition de l'architecture. Par exemple, si vous avez une application sensible à la latence, vous pouvez décider d'utiliser Network Load Balancer, qui offre des latences extrêmement faibles. Vous pouvez également décider de rapprocher votre charge de travail de vos clients en tirant parti d'Application Load Balancer dans les zones locales AWS ou même AWS Outposts.

L'équilibrage de charge entre zones est un autre élément à prendre en compte pour les charges de travail sensibles à la latence. Avec l'équilibrage de charge inter-zone, chaque nœud d'équilibreur de charge distribue le trafic sur les cibles enregistrées dans toutes les zones de disponibilité activées.

Intégrez Auto Scaling à votre équilibreur de charge. L'un des aspects essentiels d'un système performant est le dimensionnement adéquat de vos ressources backend. Pour ce faire, vous pouvez

tirer parti des intégrations d'équilibreurs de charge pour les ressources cibles du système backend. Grâce à l'intégration de l'équilibreur de charge avec les groupes Auto Scaling, les cibles seront ajoutées ou retirées de l'équilibreur de charge selon les besoins en fonction du trafic entrant. Les équilibreurs de charge peuvent également s'intégrer à Amazon ECS et Amazon EKS pour les charges de travail conteneurisées.

- Amazon ECS : équilibrage de charge des services
- Équilibrage de charge d'application sur Amazon EKS
- Équilibrage de la charge réseau sur Amazon EKS

Étapes d'implémentation

- Définissez vos exigences en matière d'équilibrage de charge, notamment en termes de volume de trafic, de disponibilité et de capacité de mise à l'échelle des applications.
- Choisissez le type d'équilibreur de charge adapté à votre application.
  - Utilisez Application Load Balancer pour les charges de travail HTTP/HTTPS.
  - Utilisez Network Load Balancer pour les charges de travail non HTTP qui fonctionnent sur TCP ou UDP.
  - Utilisez une combinaison des deux (l'ALB en tant que cible du NLB) si vous souhaitez tirer parti des fonctionnalités des deux produits. Par exemple, vous pouvez le faire si vous voulez utiliser les IP statiques du NLB avec le routage basé sur l'en-tête HTTP de l'ALB, ou si vous voulez exposer votre charge de travail HTTP à un AWS PrivateLink.
  - Pour une comparaison complète des équilibreurs de charge, voir Comparaison des produits ELB.
- Utilisez le déchargement SSL/TLS si possible.
  - Configurez les écouteurs HTTPS/TLS avec l'Application Load Balancer et le Network Load Balancer intégrés à AWS Certificate Manager.
  - Notez que certaines charges de travail peuvent nécessiter un chiffrement de bout en bout pour des raisons de conformité. Dans ce cas, il est nécessaire de permettre le chiffrement au niveau des cibles.
  - Pour les bonnes pratiques de sécurité utilisateur, voir SEC09-BP02 Application du chiffrement en transit.
- Sélectionnez le bon algorithme de routage (ALB uniquement).

  - L'algorithme de routage peut faire une réelle différence dans la manière d'utiliser vos cibles backend et donc dans leur impact sur les performances. Par exemple, l'ALB fournit deux options pour les algorithmes de routage :
  - Demandes les moins en attente : permet d'obtenir une meilleure répartition de la charge sur vos cibles backend dans les cas où les requêtes de votre application varient en complexité ou vos cibles varient en capacité de traitement.
  - Tourniquet : utilisez cette méthode lorsque les requêtes et les cibles sont similaires, ou si vous devez distribuer les requêtes de manière égale entre les cibles.
- Envisagez un isolement inter-zone ou par zone.
  - Utilisez les zones croisées désactivées (isolement par zone) pour améliorer la latence et les domaines de panne par zone. Elles sont désactivées par défaut dans le NLB et, dans l'ALB, vous pouvez les désactiver par groupe cible.
  - Utilisez les zones croisées activées pour une disponibilité et une flexibilité accrues. Par défaut, les zones croisées sont activées pour l'ALB et, dans le NLB, vous pouvez les activer par groupe cible.
- Activez l'option de persistance HTTP pour vos charges de travail HTTP (ALB uniquement). Grâce à cette fonction, l'équilibreur de charge peut réutiliser les connexions backend jusqu'à l'expiration du délai de persistance, ce qui améliore les temps de demande et de réponse HTTP et réduit également l'utilisation des ressources sur vos cibles backend. Pour plus de détails sur la procédure à suivre pour Apache et Nginx, voir Quels sont les paramètres optimaux pour utiliser Apache ou NGINX en tant que serveur principal pour ELB ?
- Activez la surveillance pour votre équilibreur de charge.
  - Activez les journaux d'accès pour l'Application Load Balancer et le Network Load Balancer.
  - Les principaux champs à prendre en compte pour l'ALB sont les suivants : `request_processing_time`, `request_processing_timeet` `response_processing_t`
  - Les principaux champs à prendre en compte pour le NLB sont les suivants : `connection_time` et `tls_handshake_time`. »
  - Soyez prêt à interroger les journaux lorsque vous en aurez besoin. Vous pouvez utiliser Amazon Athena pour interroger à la fois les journaux ALB et les journaux NLB.
  - Créez des alarmes pour les métriques liées aux performances, telles que `TargetResponseTime` pour ALB.

Ressources

Documents connexes :

- Comparaison des produits ELB
- Infrastructure mondiale AWS
- Amélioration des performances et réduction des coûts grâce à l'affinité des zones de disponibilité
- Procédure détaillée d'analyse des journaux avec Amazon Athena
- Interroger les journaux Application Load Balancer
- Surveiller vos Application Load Balancers
- Surveiller votre Network Load Balancer
- Utiliser Elastic Load Balancing pour distribuer le trafic entre les instances de votre groupe Auto Scaling

Vidéos connexes :

- AWS re:Invent 2018: Elastic Load Balancing: Deep Dive and Best Practices
- AWS re:Invent 2021 - How to choose the right load balancer for your AWS workloads
- AWS re:Inforce 2022 - How to use Elastic Load Balancing to enhance your security posture at scale
- AWS re:Invent 2019: Get the most from Elastic Load Balancing for different workloads

Exemples connexes :

- Exemples de CDK et de AWS CloudFormation pour l'analyse des journaux avec Amazon Athena

PERF04-BP05 Choix de protocoles réseau afin d'améliorer les performances

Prenez des décisions concernant les protocoles de communication entre les systèmes et les réseaux en fonction de l'impact sur les performances de la charge de travail.

Il existe une relation entre la latence et la bande passante pour atteindre le débit. Si votre transfert de fichiers utilise le protocole de contrôle de transmission (TCP), des latences plus élevées réduiront très probablement le débit global. Il existe des approches pour résoudre ce problème avec le réglage du protocole TCP et les protocoles de transfert optimisés. Le protocole UDP (User Datagram Protocol) est une solution possible.

Anti-modèles courants :

- Vous utilisez TCP pour toutes les charges de travail, quelles que soient les exigences de performance.

Avantages liés au respect de cette bonne pratique : Vérifiez que vous utilisez un protocole approprié pour la communication entre les utilisateurs et les composants de la charge de travail, afin d'améliorer l'expérience globale des utilisateurs de vos applications. Par exemple, le protocole UDP sans connexion permet d'obtenir une vitesse élevée, mais sans retransmission ni fiabilité élevée. Quoique complet, le protocole TCP nécessite une surcharge plus importante pour le traitement des paquets.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Si vous avez la possibilité de choisir différents protocoles pour votre application et que vous possédez l'expertise nécessaire dans ce domaine, optimisez votre application et l'expérience de l'utilisateur final en utilisant un autre protocole. Notez que cette approche présente des difficultés importantes et ne doit être tentée que si vous avez d'abord optimisé votre application à d'autres égards.

Pour améliorer les performances de votre charge de travail, il est essentiel de comprendre les exigences en matière de latence et de débit, puis de choisir des protocoles réseau qui optimisent les performances.

Quand envisager l'utilisation du protocole TCP

Le protocole TCP assure une livraison fiable des données et peut être utilisé pour la communication entre les composants de la charge de travail où la fiabilité et la livraison garantie des données sont importantes. De nombreuses applications web reposent sur des protocoles basés sur le protocole TCP, tels que HTTP et HTTPS, pour ouvrir des sockets TCP pour la communication entre les composants de l'application. Les e-mails et le transfert de données de fichiers sont des applications courantes qui utilisent également le protocole TCP, car il s'agit d'un mécanisme de transfert simple et fiable entre les composants de l'application. L'utilisation de TLS avec TCP peut ajouter une certaine surcharge à la communication, ce qui peut entraîner une augmentation de la latence et une réduction du débit, mais elle présente l'avantage de la sécurité. La surcharge provient principalement de la charge supplémentaire du processus de liaison, qui peut prendre plusieurs allers-retours pour se

terminer. Une fois la liaison établie, la charge de chiffrement et de déchiffrement des données devient relativement faible.

Quand envisager l'utilisation du protocole UDP

UDP est un protocole orienté sans connexion et convient donc aux applications qui nécessitent une transmission rapide et efficace, comme les données de journal, de surveillance et de VoIP. En outre, envisagez d'utiliser UDP si vous avez des composants de charge de travail qui répondent à de petites requêtes provenant d'un grand nombre de clients, afin de garantir des performances optimales de la charge de travail. Le protocole DTLS (Datagram Transport Layer Security) est l'équivalent UDP du protocole TLS (Transport Layer Security). Lors de l'utilisation de DTLS avec UDP, la charge provient du chiffrement et du déchiffrement des données, car le processus de liaison est simplifié. DTLS ajoute également une petite quantité de charge aux paquets UDP, car il inclut des champs supplémentaires pour indiquer les paramètres de sécurité et pour détecter la falsification.

Quand envisager l'utilisation du protocole SRD

Le protocole SRD (Scalable reliable datagram) est un protocole de transport en réseau optimisé pour les charges de travail à haut débit en raison de sa capacité à répartir le trafic sur plusieurs chemins et à se rétablir rapidement en cas de perte de paquets ou de défaillance d'un lien. Le SRD est donc le mieux adapté aux charges de travail du calcul haute performance (HPC) qui nécessitent un débit élevé et une communication à faible latence entre les nœuds de calcul. Il peut s'agir de tâches de traitement parallèle telles que la simulation, la modélisation et l'analyse de données qui impliquent un grand nombre de transferts de données entre les nœuds.

Étapes d'implémentation

1. Utilisez les services AWS Global Accelerator et AWS Transfer Family pour améliorer le débit de vos applications de transfert de fichiers en ligne. Le service AWS Global Accelerator vous aide à réduire la latence entre vos appareils clients et votre charge de travail sur AWS. Avec AWS Transfer Family, vous pouvez utiliser des protocoles basés sur TCP tels que le protocole de transfert de fichiers Secure Shell (SFTP) et le protocole de transfert de fichiers sur SSL (FTPS) pour dimensionner et gérer en toute sécurité vos transferts de fichiers vers des services de stockage AWS.
2. Utilisez la latence du réseau pour déterminer si le protocole TCP est adapté à la communication entre les composants de la charge de travail. Si la latence du réseau entre votre application client et le serveur est élevée, la liaison tripartite TCP peut prendre un certain temps, ce qui a un impact sur la réactivité de votre application. Des métriques telles que le délai jusqu'au premier octet (TTFB) et le temps de propagation aller-retour (RTT) peuvent être utilisées pour mesurer la

latence du réseau. Si votre charge de travail fournit du contenu dynamique aux utilisateurs, pensez à utiliser Amazon CloudFront, qui établit une connexion persistante avec chaque origine pour le contenu dynamique afin d'éliminer le temps d'établissement de la connexion qui, autrement, ralentirait chaque demande du client.

3. L'utilisation de TLS avec TCP ou UDP peut entraîner une augmentation de la latence et une réduction du débit de votre charge de travail en raison de l'impact du chiffrement et du déchiffrement. Pour de telles charges de travail, envisagez d'activer le déchargement SSL/TLS sur Elastic Load Balancing pour améliorer les performances de la charge de travail en permettant à l'équilibreur de charge de gérer les processus de chiffrement et de déchiffrement SSL/TLS au lieu de laisser les instances backend s'en charger. Cela peut contribuer à réduire l'utilisation du CPU sur les instances backend, ce qui peut améliorer les performances et augmenter la capacité.
4. Utilisez le Network Load Balancer (NLB) pour déployer des services qui reposent sur le protocole UDP, tels que l'authentification et l'autorisation, la journalisation, le DNS, l'IoT et le streaming média, afin d'améliorer les performances et la fiabilité de votre charge de travail. Le NLB distribue le trafic UDP entrant sur plusieurs cibles, ce qui vous permet de faire évoluer votre charge de travail horizontalement, d'augmenter la capacité et de réduire les frais généraux associés à une seule cible.
5. Pour vos charges de travail liées au calcul haute performance (HPC), pensez à utiliser la fonctionnalité Elastic Network Adapter (ENA) Express qui utilise le protocole SRD pour améliorer les performances du réseau en fournissant une bande passante à flux unique plus élevée (25 Gbit/s) et une latence de queue plus faible (99,9 centile) pour le trafic réseau entre les instances EC2.
6. Utilisez l' Application Load Balancer (ALB) pour router et répartir votre trafic gRPC (Remote Procedure Calls) entre les composants de la charge de travail ou entre les clients et les services gRPC. gRPC utilise le protocole HTTP/2 basé sur TCP pour le transport et offre des avantages en termes de performances, tels qu'une empreinte réseau plus légère, la compression, une sérialisation binaire efficace, la prise en charge de nombreux langages et le streaming bidirectionnel.

Ressources

Documents connexes :

- Amazon EBS – Instances optimisées
- Application Load Balancer
- Mise en réseau améliorée d'EC2 sous Linux
- Capacité réseau améliorée d'EC2 sous Windows

- Groupes de placement EC2
- Activation de la mise en réseau améliorée avec un adaptateur réseau élastique (ENA) sur les instances de Linux
- Network Load Balancer
- Mise en réseau de produits avec AWS
- AWS Transit Gateway
- Transition vers le routage basé sur la latence dans Amazon Route 53
- Points de terminaison d'un VPC
- Journaux de flux VPC

Vidéos connexes :

- Connectivity to AWS and hybrid AWS network architectures
- Optimizing Network Performance for Amazon EC2 Instances

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- Ateliers sur la mise en réseau AWS

PERF04-BP06 Choix du placement de votre charge de travail en fonction des exigences réseau

Évaluez les options de placement des ressources afin de réduire la latence du réseau et d'améliorer le débit, offrant ainsi une expérience utilisateur optimale en réduisant les temps de chargement des pages et de transfert des données.

Anti-modèles courants :

- Vous regroupez toutes les ressources de charge de travail dans un seul emplacement géographique.
- Vous avez choisi la région la plus proche de votre emplacement, pas celle de l'utilisateur final de la charge de travail.

Avantages liés au respect de cette bonne pratique : L'expérience utilisateur est fortement affectée par le temps de latence entre l'utilisateur et votre application. En utilisant les Régions AWS appropriées

et le réseau mondial AWS privé, vous pouvez réduire le temps de latence et offrir une meilleure expérience aux utilisateurs distants.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Les ressources, telles que les instances Amazon EC2, sont placées dans des zones de disponibilité dans les Régions AWS, les zones locales AWS, AWS Outpostsou les zones AWS Wavelength . Le choix de cet emplacement influence la latence et le débit du réseau à partir d'un emplacement donné de l'utilisateur. Les services périphériques comme Amazon CloudFront et AWS Global Accelerator peuvent également améliorer les performances du réseau, soit par la mise en cache du contenu aux emplacements périphériques, soit en fournissant aux utilisateurs un chemin optimal vers la charge de travail à travers le réseau global AWS.

Amazon EC2 fournit des groupes de placement pour la mise en réseau. Un groupe de placement est un regroupement logique d'instances permettant de réduire la latence. L'utilisation de groupes de placement avec des types d'instance pris en charge et un adaptateur réseau élastique (ENA) permet aux charges de travail de participer à un réseau 25 Gbit/s à faible latence avec une instabilité réduite. Les groupes de placement sont recommandés pour les charges de travail nécessitant une latence réseau faible, un débit réseau élevé, ou les deux.

Les services sensibles à la latence sont fournis sur des emplacements périphériques via le réseau mondial AWS, comme Amazon CloudFront. Ces emplacements périphériques fournissent généralement des services tels que les réseaux de diffusion de contenu (CDN) et les systèmes de noms de domaine (DNS). Placer ces services en périphérie permet aux charges de travail de répondre avec une faible latence aux requêtes de contenu ou de résolution DNS. Ces services fournissent également des services géographiques tels que le ciblage géographique du contenu (qui fournit des contenus différents en fonction de l'emplacement des utilisateurs finaux) ou le routage en fonction de la latence pour diriger les utilisateurs finaux vers la région plus proche (latence minimum).

Utilisez des services en périphérie pour réduire la latence et permettre la mise en cache de contenu. Configurez correctement le contrôle du cache pour les services DNS et HTTP/HTTPS afin de tirer le plus grand bénéfice de ces approches.

Étapes d'implémentation

- Capturez des informations sur le trafic IP entrant et sortant des interfaces réseau.
  - Enregistrement du trafic IP à l'aide des journaux de flux VPC
  - Comment l'adresse IP du client est-elle préservée dans AWS Global Accelerator

- Analysez les modèles d'accès au réseau dans votre charge de travail afin d'identifier comment les utilisateurs utilisent votre application.
  - Utilisez des outils de surveillance, comme Amazon CloudWatch et AWS CloudTrail, pour recueillir des données sur les activités du réseau.
  - Analysez les données pour identifier le modèle d'accès au réseau.
- Choisissez les régions pour le déploiement de votre charge de travail en fonction des éléments clés suivants :
  - Emplacement de vos données : pour les applications utilisant de grandes quantités de données (telles que le big data et le machine learning). Le code de l'application doit s'exécuter aussi près que possible des données.
  - Emplacement de vos utilisateurs : pour les applications orientées utilisateur, choisissez une région ou des régions proches des utilisateurs de votre charge de travail.
  - Autres contraintes : tenez compte de contraintes telles que le coût et la conformité comme indiqué dans Éléments à prendre en compte lors de la sélection d'une région pour vos charges de travail.
- Utilisez les zones locales AWS pour exécuter des charges de travail telles que le rendu vidéo. Ces zones locales vous permettent de profiter des avantages liés à la présence de ressources de calcul et de stockage plus proches des utilisateurs finaux.
- Utilisez AWS Outposts pour les charges de travail qui doivent rester sur site et dont vous souhaitez qu'elles fonctionnent de manière transparente avec le reste de vos charges de travail dans AWS.
- Les applications telles que le streaming vidéo en direct à haute résolution, l'audio haute fidélité et la réalité augmentée ou virtuelle (RA/RV) exigent une latence ultra-faible pour les appareils 5G. Pour de telles applications, envisagez d'utiliser AWS Wavelength. AWS Wavelength intègre les services de calcul et de stockage AWS aux réseaux 5G, tout en offrant une infrastructure de calcul périphérique mobile pour le développement, le déploiement et la mise à l'échelle des applications avec une latence ultra-faible.
- Utilisez la mise en cache locale ou des solutions de mise en cache AWS pour les ressources fréquemment utilisées afin d'améliorer les performances, de limiter les mouvements de données et de réduire l'impact sur l'environnement.

Service	Quand l'utiliser
Amazon CloudFront	Permet de mettre en cache du contenu statique comme des images, des scripts et

Service	Quand l'utiliser
	des vidéos, ainsi que du contenu dynamique comme des réponses API ou des applications Web.
Amazon ElastiCache	Permet de mettre en cache du contenu pour les applications Web.
DynamoDB Accelerator	Permet d'ajouter une accélération en mémoire à vos tables DynamoDB.

- Utilisez des services capables de vous aider à exécuter le code plus près des utilisateurs de votre charge de travail, tels que les suivants :

Service	Quand l'utiliser
Lambda@edge	Destiné aux opérations exigeantes en puissance de calcul qui sont lancées lorsque des objets ne sont pas dans le cache.
Fonctions Amazon CloudFront	Destiné aux cas d'utilisation simples comme les requêtes HTTP(S) ou les manipulations de réponse pouvant être lancées par des fonctions de courte durée.
AWS IoT Greengrass	Permet d'exécuter du calcul local, une messagerie et une mise en cache de données pour les appareils connectés.

- Certaines applications nécessitent des points d'entrée fixes ou des performances plus élevées en réduisant la latence et l'instabilité du premier octet et en augmentant le débit. Ces applications peuvent bénéficier de services réseau qui fournissent des adresses IP statiques anycast et une terminaison TCP aux emplacements périphériques. AWS Global Accelerator peut améliorer les performances de vos applications jusqu'à 60 % et assurer un basculement rapide pour les architectures multirégionales. AWS Global Accelerator vous fournit des adresses IP statiques anycast qui servent de point d'entrée fixe à vos applications hébergées dans une ou plusieurs Régions AWS. Ces adresses IP permettent au trafic d'entrer sur le réseau global AWS aussi près que possible de vos utilisateurs. AWS Global Accelerator réduit le temps d'établissement de la

connexion initiale en établissant une connexion TCP entre le client et l'emplacement périphérique AWS le plus proche du client. Examinez l'utilisation de AWS Global Accelerator pour améliorer les performances de vos charges de travail TCP/UDP et fournir un basculement rapide pour les architectures multirégionales.

Ressources

Bonnes pratiques associées :

- COST07-BP02 Mise en œuvre de régions en fonction des coûts
- COST08-BP03 Mise en œuvre de services pour réduire les coûts de transfert de données
- REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements
- REL10-BP02 Sélection des emplacements appropriés pour votre déploiement multisite
- SUS01-BP01 Choix d'une région en fonction des exigences et des objectifs de durabilité de l'entreprise
- SUS02-BP04 Optimisation du placement géographique des charges de travail en fonction de leurs exigences réseau
- SUS04-BP07 Réduction du mouvement des données entre les réseaux

Documents connexes :

- Infrastructure mondiale AWS
- Zones locales AWS et AWS Outposts, choisir la bonne technologie pour votre charge de travail périphérique
- Groupes de placement
- les zones locales AWS
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Vidéos connexes :

- AWS Local Zones Explainer Video
- AWS Outposts: Overview and How it Works
- AWS re:Invent 2021 - AWS Outposts: Bringing the AWS experience on premises
- AWS re:Invent 2020: AWS Wavelength: Run apps with ultra-low latency at 5G edge
- AWS re:Invent 2022 - AWS Local Zones: Building applications for a distributed edge
- AWS re:Invent 2021 - Building low-latency websites with Amazon CloudFront
- AWS re:Invent 2022 - Improve performance and availability with AWS Global Accelerator
- AWS re:Invent 2022 - Build your global wide area network using AWS
- AWS re:Invent 2020: Global traffic management with Amazon Route 53

Exemples connexes :

- Atelier AWS Global Accelerator
- Gestion des réécritures et des redirections à l'aide des fonctions de périphérie

PERF04-BP07 Optimisation de la configuration réseau en fonction de métriques

Utilisez les données collectées et analysées pour prendre des décisions avisées concernant l'optimisation de votre configuration réseau.

Anti-modèles courants :

- Vous supposez que tous les problèmes liés aux performances sont liés à l'application.
- Vous testez uniquement les performances de votre réseau à partir d'un emplacement proche de l'endroit où vous avez déployé la charge de travail.
- Vous utilisez des configurations par défaut pour tous les services du réseau.
- Vous surdimensionnez la ressource réseau afin de fournir une capacité suffisante.

Avantages liés au respect de cette bonne pratique : La collecte des métriques nécessaires de votre réseau AWS et la mise en œuvre d'outils de surveillance du réseau vous permettent de comprendre les performances du réseau et d'optimiser les configurations du réseau.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

La surveillance du trafic en provenance et à destination des VPC, des sous-réseaux ou des interfaces réseau est essentielle pour comprendre comment utiliser les ressources réseau AWS et comment optimiser les configurations réseau. Les outils de mise en réseau AWS suivants vous permettent d'obtenir des informations supplémentaires sur l'utilisation du trafic, l'accès au réseau et les journaux.

Étapes d'implémentation

- Identifiez les indicateurs clés de performance tels que la latence ou la perte de paquets à collecter. AWS fournit plusieurs outils qui peuvent vous aider à collecter ces métriques. Les outils suivants vous permettent d'obtenir des informations supplémentaires sur l'utilisation du trafic, l'accès au réseau et les journaux.

Outil AWS	Où utiliser
Amazon VPC IP Address Manager.	Utilisez IPAM pour planifier, suivre et surveille r les adresses IP pour vos charges de travail AWS et sur site. Il s'agit d'une bonne pratique pour optimiser l'utilisation et l'allocation des adresses IP.
Journaux de flux VPC	Utilisez les journaux de flux VPC pour capturer des informations détaillées sur le trafic en provenance et à destination des interface s réseau de vos VPC. Grâce aux journaux de flux VPC, vous pouvez diagnostiquer les règles de groupes de sécurité trop restrictives ou trop permissives et déterminer la direction du trafic vers et depuis les interfaces réseau.
Journaux de flux AWS Transit Gateway	Utilisez les journaux de flux AWS Transit Gateway pour capturer des informations sur le trafic IP à destination et en provenance de vos passerelles de transit.
Journalisation des requêtes DNS	Enregistrez les informations relatives aux requêtes DNS publiques ou privées reçues par Route 53. Grâce aux journaux DNS, vous

Outil AWS	Où utiliser
	pouvez optimiser les configurations DNS en comprenant le domaine ou le sous-domaine qui a été demandé ou les emplacements périphériques Route 53 qui ont répondu aux requêtes DNS.
Reachability Analyzer	Utilisez Reachability Analyzer pour analyser et déboguer l'accessibilité du réseau. Reachability Analyzer est un outil d'analyse de la configuration qui vous permet d'effectuer des tests de connectivité entre une ressource source et une ressource de destination dans vos VPC. Cet outil vous aide à vérifier que votre configuration réseau correspond à la connectivité souhaitée.
Network Access Analyzer	Network Access Analyzer vous aide à comprendre l'accès réseau à vos ressources. Vous pouvez utiliser Network Access Analyzer pour spécifier vos exigences en matière d'accès au réseau et identifier les chemins d'accès potentiels qui ne répondent pas à vos exigences spécifiées. En optimisant la configuration de votre réseau correspondant, vous pouvez comprendre et vérifier l'état de votre réseau et démontrer si votre réseau sur AWS répond à vos exigences de conformité.

Outil AWS	Où utiliser
Amazon CloudWatch	Utilisez Amazon CloudWatch et activez les métriques appropriées pour les options réseau. Veillez à choisir la métrique de réseau adaptée à votre charge de travail. Par exemple, vous pouvez activer des métriques pour l'utilisation d'adresses réseau VPC, la passerelle VPC NAT, AWS Transit Gateway, le tunnel VPN, AWS Network Firewall, Elastic Load Balancing et AWS Direct Connect. La surveillance continue des métriques est une bonne pratique pour observer et comprendre l'état et l'utilisation de votre réseau. Elle vous aide à optimiser la configuration du réseau en fonction de vos observations.
AWS Network Manager	AWS Network Manager permet de surveiller les performances historiques et en temps réel du réseau mondial AWS à des fins opération nelles et de planification. Network Manager fournit une latence réseau globale entre les Régions AWS et les zones de disponibilité et au sein de chaque zone de disponibilité, ce qui vous permet de mieux comprendre le lien entre les performances de votre application et les performances du réseau AWS sous-jacent.
Amazon CloudWatch RUM	Utilisez Amazon CloudWatch RUM pour collecter les métriques fournissant les informations qui vous aideront à identifier, à comprendre et à améliorer l'expérience utilisateur.

- Identifiez les principaux intervenants et les modèles de trafic des applications à l'aide des journaux de flux VPC et AWS Transit Gateway.

- Évaluez et optimisez votre architecture réseau actuelle, y compris les VPC, les sous-réseaux et le routage. À titre d'exemple, vous pouvez évaluer l'impact de l'appairage de VPC ou d'AWS Transit Gateway sur l'amélioration de la mise en réseau de votre architecture.
- Évaluez les chemins de routage de votre réseau pour vérifier que le chemin le plus court entre les destinations est toujours utilisé. Network Access Analyzer peut vous aider à le faire.

Ressources

Documents connexes :

- Journaux de flux VPC
- Journalisation des requêtes DNS publiques
- Qu'est-ce qu'IPAM ?
- Qu'est-ce que Reachability Analyzer ?
- Qu'est-ce que Network Access Analyzer ?
- Métriques CloudWatch pour vos VPC
- Optimiser les performances et réduire les coûts analytiques des réseaux grâce aux journaux de flux VPC au format Apache Parquet
- Surveillance de vos réseaux mondiaux et principaux avec les métriques Amazon CloudWatch
- Surveiller en permanence le trafic et les ressources du réseau

Vidéos connexes :

- Networking best practices and tips with the AWS Well-Architected Framework
- Monitoring and troubleshooting network traffic

Exemples connexes :

- Ateliers sur la mise en réseau AWS
- Surveillance réseau AWS

Processus et culture

PERF 5. Comment vos pratiques et votre culture organisationnelles contribuent-elles à l'efficacité des performances dans le cadre de votre charge de travail ?

Lors de la création de l'architecture des charges de travail, vous pouvez adopter certains principes et certaines pratiques pour optimiser l'exécution de charges de travail cloud efficaces et performantes. Pour adopter une culture qui favorise l'efficacité des performances des charges de travail dans le cloud, tenez compte des principes et pratiques clés suivants :

Bonnes pratiques

- PERF05-BP01 Définir des indicateurs clés de performance (KPI) pour mesurer l'état et les performances de la charge de travail
- PERF05-BP02 Utiliser des solutions de surveillance pour comprendre les domaines où les performances sont d'une importance critique
- PERF05-BP03 Définir un processus pour améliorer les performances des charges de travail
- PERF05-BP04 Effectuer un test de charge de votre charge de travail
- PERF05-BP05 Utiliser l'automatisation pour résoudre de manière proactive les problèmes liés aux performances
- PERF05-BP06 Maintenir votre charge de travail et vos services à jour
- PERF05-BP07 Vérifier les métriques à intervalles réguliers

PERF05-BP01 Définir des indicateurs clés de performance (KPI) pour mesurer l'état et les performances de la charge de travail

Identifiez les KPI qui mesurent les performances de la charge de travail de manière quantitative et qualitative. Les KPI vous aident à mesurer l'état et les performances d'une charge de travail par rapport à un objectif métier.

Anti-modèles courants :

- Vous surveillez uniquement les métriques au niveau du système pour avoir un aperçu de votre charge de travail et ne comprenez pas les impacts commerciaux possibles.
- Vous supposez que vos KPI sont déjà publiés et partagés en tant que données de métriques standard.
- Vous ne définissez pas de KPI quantitatif et mesurable.

- Vous ne tenez pas compte des objectifs ni des stratégies de l'entreprise pour définir vos KPI.

Avantages liés au respect de cette bonne pratique : En identifiant les KPI spécifiques qui représentent l'état et les performances de la charge de travail, vous pouvez aligner les équipes sur leurs priorités et définir des résultats commerciaux atteignables. Le partage de ces métriques avec tous les départements offre une visibilité et un alignement sur les seuils, les attentes et l'impact commercial.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les KPI permettent aux équipes commerciales et d'ingénierie de s'aligner sur la mesure des objectifs et des stratégies et sur la façon dont ces facteurs se combinent pour générer des résultats commerciaux. Par exemple, une charge de travail de site Web peut utiliser le temps de chargement de la page comme indication des performances globales. Cette métrique serait l'un des éléments de données pris en compte qui mesure l'expérience d'un utilisateur. En plus d'identifier les temps limites de chargement des pages, vous devez documenter le résultat attendu ou le risque commercial si les performances idéales ne sont pas atteintes. Un temps de chargement long des pages affecte directement vos utilisateurs finaux, nuit à leur expérience utilisateur et peut entraîner une perte de clients. Lorsque vous définissez vos seuils de KPI, combinez à la fois les points de référence en vigueur dans votre secteur et les attentes de vos utilisateurs finaux. Par exemple, si le point de référence actuel établi par votre secteur d'activité pour le chargement d'une page Web est un délai de deux secondes, mais que vos utilisateurs finaux s'attendent à ce qu'une page Web se charge dans un délai d'une seconde, vous devez prendre en compte ces deux éléments de données lors de la définition des KPI.

Votre équipe doit évaluer les KPI de votre charge de travail à l'aide de données précises en temps réel et de données historiques à titre de référence et créer des tableaux de bord qui effectuent des calculs de métriques par rapport à vos données de KPI pour générer des informations opérationnelles et d'utilisation. Les KPI doivent être documentés et inclure les seuils qui soutiennent les objectifs et les stratégies de l'entreprise et doivent être mappés aux métriques surveillées. Les KPI doivent être revus lorsque les objectifs commerciaux, les stratégies ou les exigences des utilisateurs finaux changent.

Étapes d'implémentation

1. Identifiez et documentez les principales parties prenantes de l'entreprise.

2. Collaborez avec ces parties prenantes pour définir et documenter les objectifs de votre charge de travail.
3. Passez en revue les bonnes pratiques du secteur pour identifier les KPI pertinents qui correspondent à vos objectifs en matière de charge de travail.
4. Utilisez les bonnes pratiques du secteur et vos objectifs de charge de travail pour définir des cibles pour votre KPI de charge de travail. Utilisez ces informations pour définir les seuils de KPI pour les niveaux de gravité ou d'alarme.
5. Identifiez et documentez le risque et l'impact si le KPI n'est pas atteint.
6. Identifiez et documentez les métriques qui peuvent vous aider à établir les KPI.
7. Utilisez des outils de surveillance, comme Amazon CloudWatch ou AWS Config pour collecter des métriques et mesurer des KPI.
8. Utilisez des tableaux de bord pour visualiser et communiquer les KPI aux parties prenantes.
9. Passez en revue et analysez régulièrement les métriques pour identifier les domaines de charge de travail qui doivent être améliorés.
10. Revoyez les KPI lorsque les objectifs métier ou les performances de la charge de travail changent.

Ressources

Documents connexes :

- Documentation CloudWatch
- Surveillance, journalisation et performances AWS Partner
- Documentation X-Ray
- Fonctionnement des tableaux de bord Amazon CloudWatch
- KPI Amazon QuickSight

Vidéos connexes :

- AWS re:Invent 2019 : passez à vos 10 premiers millions d'utilisateurs
- Mettez fin au chaos : gagnez en visibilité et en informations opérationnelles
- Build a Monitoring Plan

Exemples connexes :

- Création d'un tableau de bord avec Amazon QuickSight

PERF05-BP02 Utiliser des solutions de surveillance pour comprendre les domaines où les performances sont d'une importance critique

Comprenez et identifiez les domaines où l'augmentation des performances de votre charge de travail aura un impact positif sur l'efficacité ou l'expérience client. Par exemple, un site Web qui comporte un grand nombre d'interactions clients pourrait gagner à utiliser des services de périphérie pour rapprocher la diffusion de contenus des clients.

Anti-modèles courants :

- Vous supposez que les métriques de calcul standard telles que l'utilisation du processeur ou la pression de mémoire, suffisent pour détecter les problèmes de performances.
- Vous n'utilisez que les métriques par défaut enregistrées par le logiciel de surveillance que vous avez sélectionné.
- Vous n'examinez les métriques qu'en cas de problème.

Avantages liés au respect de cette bonne pratique : la compréhension des domaines critiques de performances aide les propriétaires des charges de travail à surveiller les KPI et à prioriser les améliorations à impact élevé.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Mettez en place un suivi de bout en bout afin d'identifier les tendances du trafic, la latence et les domaines de performances critiques. Surveillez vos modèles d'accès aux données afin d'identifier les requêtes lentes ou les données mal fragmentées et partitionnées. Identifiez les zones de charge de travail limitées à l'aide de tests ou de surveillance des charges.

améliorer l'efficacité des performances en comprenant votre architecture, vos modèles de trafic et d'accès aux données, et identifier vos temps de latence et de traitement. Identifier les goulots d'étranglement potentiels qui pourraient avoir une incidence sur l'expérience client à mesure que la charge de travail augmente. Après avoir enquêté sur ces domaines, déterminez quelle solution vous pouvez déployer afin de surmonter ces problèmes de performances.

Étapes d'implémentation

1. Mettez en place une surveillance de bout en bout pour capturer tous les composants et métriques de la charge de travail. Voici des exemples de solutions de surveillance sur AWS.

Service	Où utiliser
Amazon CloudWatch Real-User Monitoring (RUM)	Pour capturer les métriques de performances des applications à partir de sessions réelles côté client et front-end.
AWS X-Ray	Pour tracer le trafic à travers les couches applicatives et identifier la latence entre les composants et les dépendances. Utilisez les cartographies de services X-Ray afin de voir les relations et la latence entre les composant s de la charge de travail.
Amazon Relational Database Service Performance Insights	Pour consulter les métriques de performan ces de la base de données et identifier les améliorations des performances.
Amazon RDS Enhanced Monitoring	Pour consulter les métriques de performan ces du système d'exploitation de la base de données.
Amazon DevOps Guru	Pour détecter les modèles de fonctionnement anormaux afin que vous puissiez identifier les problèmes opérationnels avant qu'ils n'affecte nt vos clients.

2. Effectuez des tests afin de générer des métriques, d'identifier les tendances de trafic, les goulots d'étranglement et les domaines de performance critiques. Voici quelques exemples de méthodes de test :
   - Configurez des tests canary synthétiques CloudWatch pour imiter par programmation les activités des utilisateurs basées sur le navigateur à l'aide de tâches cron Linux ou d'expressions de taux afin de générer des métriques cohérentes au fil du temps.

- Utilisez l' Test de charge distribuée sur AWS afin de générer un trafic de pointe ou de tester la charge de travail au taux de croissance attendu.

3. Évaluez les métriques et la télémétrie pour identifier vos domaines de performances critiques. Examinez ces domaines avec votre équipe afin de discuter de la surveillance et des solutions pour éviter les goulots d'étranglement.
4. Expérimentez des améliorations des performances et mesurez ces changements avec des données. À titre d'exemple, vous pouvez utiliser CloudWatch Evidently afin de tester les nouvelles améliorations et les impacts sur votre charge de travail.

Ressources

Documents connexes :

- Bibliothèque Amazon Builders' Library
- Documentation X-Ray
- Amazon CloudWatch RUM
- Amazon DevOps Guru

Vidéos connexes :

- La bibliothèque Amazon Builders’ Library : 25 ans d'excellence opérationnelle d'Amazon
- Surveillance visuelle des applications avec Synthetics Amazon CloudWatch

Exemples connexes :

- Mesurer le temps de chargement des pages avec Amazon CloudWatch Synthetics
- Client web Amazon CloudWatch RUM
- Kit SDK X-Ray pour Node.js
- Kit SDK X-Ray pour Python
- Kit SDK X-Ray pour Java
- Kit SDK X-Ray pour .Net
- Kit SDK X-Ray pour Ruby
- Démon X-Ray
- Test de charge distribuée sur AWS

PERF05-BP03 Définir un processus pour améliorer les performances des charges de travail

Définissez un processus d'évaluation de nouveaux services, les modèles de conception, les types de ressources et les configurations au fur et à mesure qu'elles deviennent disponibles. Par exemple, exécutez des tests de performances existants sur de nouvelles offres d'instances afin de déterminer leur potentiel d'amélioration de votre charge de travail.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne sera pas mise à jour au fil du temps.
- Vous introduisez des modifications d'architecture au fil du temps sans justification basée sur les métriques.

Avantages liés au respect de cette bonne pratique : Un processus défini pour les modifications d'architecture rend possible l'utilisation des données collectées pour influencer la conception de votre charge de travail au fil du temps.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Les performances de votre charge de travail présentent quelques contraintes clés. Documentez-les pour connaître les types d'innovations qui pourraient améliorer les performances de votre charge de travail. Utilisez ces informations lors de l'apprentissage de nouveaux services ou la technologie au fur et à mesure de leur disponibilité afin d'identifier les moyens d'atténuer des contraintes ou des goulets d'étranglement.

Identifiez les principales contraintes de performance pour votre charge de travail. Documentez les contraintes environnementales de votre charge de travail pour connaître les types d'innovations qui pourraient améliorer les performances de celle-ci.

Étapes d'implémentation

- Identifiez les KPI de performance de votre charge de travail, comme indiqué dans PERF05-BP01 Définir des indicateurs clés de performance (KPI) pour mesurer l'état et les performances de la charge de travail pour établir un point de comparaison pour votre charge de travail.
- Utilisez les outils d'observabilité d'AWS pour collecter des métriques de performance et mesurer les KPI.

- Effectuez une analyse approfondie pour identifier les domaines (tels que la configuration et le code d'application) de votre charge de travail qui ne sont pas performants, comme indiqué dans PERF05-BP02 Utiliser des solutions de surveillance pour comprendre les domaines où les performances sont d'une importance critique. »
- Utilisez vos outils d'analyse et de performance pour identifier la stratégie d'optimisation des performances.
- Utilisez des environnements de test ou de pré-production pour valider l'efficacité de la stratégie.
- Mettez en œuvre les modifications en production et surveillez en permanence les performances de la charge de travail.
- Documentez les améliorations et communiquez-les aux parties prenantes.

Ressources

Documents connexes :

- Blog AWS
- Nouveautés AWS

Vidéos connexes :

- Chaîne YouTube AWS Events
- Chaîne YouTube AWS Online Tech Talks
- Chaîne YouTube Amazon Web Services

Exemples connexes :

- AWS Github
- AWS Skill Builder

PERF05-BP04 Effectuer un test de charge de votre charge de travail

Effectuez un test de charge de votre charge de travail pour vérifier qu'elle peut supporter la charge de production et identifier les éventuels goulots d'étranglement en termes de performances.

Anti-modèles courants :

- Vous testez les différentes parties et non la totalité de votre charge de travail.
- Vous testez la charge sur une infrastructure qui n'est pas la même que votre environnement de production.
- Vous n'effectuez le test de charge que pour la charge prévue sans aller au-delà, avec pour but de prévoir où vous pourriez rencontrer des problèmes à l'avenir.
- Vous effectuez des tests de charge sans consulter la politique de test Amazon EC2 et en soumettant un formulaire de soumission d'événements simulés. Cela entraîne l'échec de votre test, car cela ressemble à un événement de déni de service.

Avantages liés au respect de cette bonne pratique : La mesure de vos performances dans le cadre d'un test de charge vous indiquera où vous serez affecté au fil de l'augmentation de la charge. Cela peut vous permettre d'anticiper les changements nécessaires avant qu'ils n'affectent votre charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

Les tests de charge dans le cloud sont un processus visant à mesurer les performances de la charge de travail cloud dans des conditions réalistes avec la charge utilisateur attendue. Ce processus implique la mise en service d'un environnement cloud de type production, l'utilisation d'outils de test de charge pour générer la charge et l'analyse de métriques pour évaluer la capacité de votre charge de travail à gérer une charge réaliste. Pour effectuer un test de charge, vous devez exécuter des versions de données de production factices ou légèrement altérées (supprimez les données sensibles ou les informations d'identification). Effectuez automatiquement des tests de charge dans le cadre de votre pipeline de livraison et comparez les résultats aux indicateurs de performance clés et aux seuils prédéfinis. Ce processus vous permet de continuer à atteindre les performances requises.

Étapes d'implémentation

- Configurez l'environnement de test en fonction de votre environnement de production. Vous pouvez utiliser les services AWS pour exécuter des environnements à l'échelle de la production afin de tester votre architecture.
- Choisissez et configurez l'outil de test de charge adapté à votre charge de travail.
- Définissez les scénarios et les paramètres de test de charge (tels que la durée du test et le nombre d'utilisateurs).

- Réalisez des scénarios de test à grande échelle. Utilisez le AWS Cloud pour tester votre charge de travail et découvrir où elle ne parvient pas à se dimensionner ou si elle évolue de manière non linéaire. Par exemple, utilisez les instances Spot pour générer des charges à faible coût et découvrir les goulots d'étranglement avant de les rencontrer en production.
- Surveillez et enregistrez les métriques de performance (comme le débit et le temps de réponse). Amazon CloudWatch peut récupérer des métriques à partir des ressources de votre architecture. Vous pouvez également récupérer et publier des métriques personnalisées pour faire apparaître des métriques d'entreprise ou des métriques dérivées.
- Analysez les résultats pour identifier les goulots d'étranglement en matière de performances et les domaines à améliorer.
- Documentez et rendez compte du processus et des résultats des tests de charge.

Ressources

Documents connexes :

- AWS CloudFormation
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Test de charge distribuée sur AWS

Vidéos connexes :

- Solving with AWS Solutions: Distributed Load Testing

- Optimize applications through Amazon CloudWatch RUM
- Demo of Amazon CloudWatch Synthetics

Exemples connexes :

- Test de charge distribuée sur AWS

PERF05-BP05 Utiliser l'automatisation pour résoudre de manière proactive les problèmes liés aux performances

Utilisez les KPI en combinaison avec des systèmes de surveillance et d'alarme pour traiter de manière proactive les problèmes liés aux performances.

Anti-modèles courants :

- Vous autorisez uniquement le personnel des opérations à apporter des modifications opérationnelles à la charge de travail.
- Vous confiez toutes les activités de filtre des alarmes à l'équipe des opérations sans correction proactive.

Avantages liés au respect de cette bonne pratique : La correction proactive des actions d'alarme permet au personnel d'assistance de se concentrer sur les éléments qui ne sont pas exploitables automatiquement. Cela permet au personnel des opérations de gérer toutes les alarmes sans être submergé et de se concentrer uniquement sur les alarmes critiques.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

Utilisez des alarmes pour déclencher des actions automatisées afin de corriger les problèmes dans la mesure du possible. Faites remonter l'alarme aux personnes qui peuvent répondre si une réponse automatique n'est pas possible. Par exemple, vous pourriez disposer d'un système capable de prédire les valeurs attendues de KPI et qui déclenche une alarme lorsqu'elles dépassent certains seuils. Vous pouvez aussi disposer d'un outil capable d'arrêter ou de restaurer automatiquement des déploiements si les valeurs des KPI dépassent celles attendues.

Mettez en place des processus qui rendent visibles les performances pendant que votre charge de travail est en cours d'exécution. Créez des tableaux de bord de surveillance et établissez des normes de référence pour les attentes en matière de performances pour déterminer si les performances de la charge de travail sont optimales.

Étapes d'implémentation

- Identifiez et comprenez le problème lié aux performances qui peut être résolu automatiquement. Utilisez les solutions de surveillance d'AWS telles que Amazon CloudWatch ou AWS X-Ray pour vous aider à mieux comprendre la cause profonde du problème.

- Créez un plan et un processus de résolution étape par étape qui peuvent être utilisés pour résoudre automatiquement le problème.
- Configurez le déclencheur pour lancer automatiquement le processus de résolution. Par exemple, vous pouvez définir un déclencheur pour redémarrer automatiquement une instance lorsqu'elle atteint un certain seuil d'utilisation de l'UC.
- Utilisez les services et technologies AWS pour automatiser le processus de résolution. Par exemple, AWS Systems Manager Automation fournit une solution sécurisée et évolutive d'automatisation du processus de résolution.
- Testez le processus de résolution automatisé dans un environnement de pré-production.
- Après les tests, mettez en œuvre le processus de résolution dans l'environnement de production et effectuez une surveillance continue pour identifier les domaines à améliorer.

Ressources

Documents connexes :

- Documentation CloudWatch
- Surveillance, journalisation et performances – Partenaires AWS Partner Network
- Documentation X-Ray
- Utilisation des alarmes et des actions d'alarme dans CloudWatch

Vidéos connexes :

- Intelligently automating cloud operations
- Configuration de contrôles à grande échelle dans votre environnement AWS
- Automatiser la gestion des correctifs et la conformité à l'aide de AWS
- How Amazon uses better metrics for improved website performance

Exemples connexes :

- Personnalisation des alarmes CloudWatch Logs (langue française non garantie)

PERF05-BP06 Maintenir votre charge de travail et vos services à jour

Restez informé des nouveaux services et des nouvelles fonctionnalités cloud pour adopter des fonctionnalités efficaces, résoudre les problèmes et améliorer l'efficacité globale des performances de votre charge de travail.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne sera pas mise à jour au fil du temps.
- Vous ne disposez pas de systèmes ou de rythme régulier pour évaluer la compatibilité des packages et des logiciels mis à jour avec votre charge de travail.

Avantages liés au respect de cette bonne pratique : En mettant en place un processus permettant de rester informé des nouveaux services et des nouvelles offres, vous pouvez adopter de nouvelles fonctionnalités et capacités, résoudre les problèmes et améliorer les performances de la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

Évaluez les méthodes d'amélioration des performances au fur et à mesure que de nouveaux services, modèles de conception et fonctionnalités de produits entrent en scène. Identifiez celles de ces méthodes qui sont susceptibles d'améliorer les performances ou d'accroître l'efficacité de la charge de travail via l'évaluation, la discussion interne ou l'analyse externe. Mettez en place un processus permettant d'évaluer les mises à jour, les nouvelles fonctions et les services pertinents pour votre charge de travail. Par exemple, la création d'une démonstration de faisabilité qui utilise les nouvelles technologies ou la consultation d'un groupe interne. Lorsque vous essayez de nouvelles idées ou services, exécutez des tests de performances pour mesurer leur impact sur les performances de la charge de travail.

Étapes d'implémentation

- Établissez l'inventaire de votre logiciel de charge de travail et de l'architecture, et identifiez les composants pouvant être mis à jour.
- Identifiez les actualités et mettez à jour les sources liées aux composants de votre charge de travail. À titre d'exemple, vous pouvez vous abonner au blog Quelles sont les nouveautés

AWS ? pour les produits qui correspondent à votre charge de travail. Vous pouvez vous abonner au flux RSS ou gérer vos abonnements aux e-mails. »

- Définissez un calendrier pour évaluer les nouveaux services et les nouvelles fonctionnalités adaptés à votre charge de travail.
  - Vous pouvez utiliser AWS Systems Manager Inventory pour récupérer les métadonnées des systèmes d'exploitation, des applications et des instances issues de vos instances Amazon EC2 et rapidement connaître les instances exécutant le logiciel, les configurations requises par votre politique de logiciel et les instances devant être mises à jour.
- Comprenez comment mettre à jour les composants de votre charge de travail. Profitez de l'agilité du cloud pour tester rapidement en quoi les nouvelles fonctionnalités peuvent permettre à votre charge de travail de gagner en efficacité.
- Utilisez l'automatisation pour le processus de mise à jour afin de réduire le niveau d'effort nécessaire au déploiement des nouvelles fonctionnalités et de limiter les erreurs causées par les processus manuels.
  - Vous pouvez utiliser CI/CD pour mettre automatiquement à jour les AMI, les images de conteneurs et d'autres artefacts liés à votre application cloud.
  - Vous pouvez utiliser des outils tels que le gestionnaire de correctifs AWS Systems Manager pour automatiser le processus de mise à jour du système, et planifier l'activité à l'aide des fenêtres de maintenance AWS Systems Manager. »
- Documentez votre processus d'évaluation des mises à jour et des nouveaux services. Donnez aux propriétaires le temps et l'espace nécessaires pour rechercher, tester, expérimenter et valider les mises à jour et les nouveaux services. Reportez-vous aux exigences opérationnelles documentées et aux KPI pour établir l'ordre de priorité des mises à jour qui auront un impact positif sur les activités.

Ressources

Documents connexes :

- Blog AWS
- Nouveautés AWS

Vidéos connexes :

- Chaîne YouTube AWS Events

- Chaîne YouTube AWS Online Tech Talks
- Chaîne YouTube Amazon Web Services

Exemples connexes :

- Ateliers Well-Architected : inventaire et gestion des correctifs
- Atelier : AWS Systems Manager

PERF05-BP07 Vérifier les métriques à intervalles réguliers

Vérifiez les métriques qui sont collectées dans le cadre de la maintenance de routine ou en réponse à des événements ou des incidents. Utilisez ces vérifications pour identifier d'une part les métriques qui ont été essentielles pour traiter les problèmes, et d'autre part les métriques supplémentaires, si elles ont été suivies, qui pourraient aider à identifier, traiter ou empêcher les problèmes.

Anti-modèles courants :

- Vous autorisez les métriques à rester dans un état d'alarme pendant longtemps.
- Vous créez des alarmes qui ne sont pas exploitables par un système d'automatisation.

Avantages liés au respect de cette bonne pratique : Passer en revue en permanence les métriques qui sont collectées pour vérifier qu'elles identifient, résolvent ou préviennent correctement les problèmes. Les métriques peuvent également devenir caduques si vous les laissez dans un état d'alarme pendant longtemps.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Améliorez constamment la surveillance et la collecte des métriques. Lorsque vous répondez aux incidents ou aux événements, évaluez les métriques qui ont été utiles dans la gestion du problème et les métriques qui auraient pu aider mais ne sont pas suivies actuellement. Utilisez cette méthode pour améliorer la qualité des métriques que vous collectez afin de pouvoir prévenir ou résoudre plus rapidement les incidents futurs.

Lorsque vous répondez aux incidents ou aux événements, évaluez les métriques qui ont été utiles dans la gestion du problème et les métriques qui auraient pu aider mais ne sont pas suivies

actuellement. Utilisez ce processus pour améliorer la qualité des métriques que vous collectez afin de pouvoir prévenir ou résoudre plus rapidement les incidents futurs.

Étapes d'implémentation

1. Définissez des métriques de performances critiques à surveiller qui correspondent à votre objectif de charge de travail.
2. Définissez une base de référence et une valeur souhaitable pour chaque métrique.
3. Définissez une cadence (hebdomadaire ou mensuelle, par exemple) pour examiner les métriques critiques.
4. Au cours de chaque examen, évaluez les tendances et les écarts par rapport aux valeurs de référence. Recherchez les goulots d'étranglement ou les anomalies au niveau des performances.
5. Pour les problèmes identifiés, effectuez une analyse détaillée des causes profondes afin de comprendre la raison principale du problème.
6. Documentez vos résultats et utilisez des stratégies pour résoudre les problèmes et les goulots d'étranglement identifiés.
7. Évaluez et améliorez en permanence le processus de révision des métriques.

Ressources

Documents connexes :

- Documentation CloudWatch
- Collecte des métriques et des journaux des instances Amazon EC2 et serveurs sur site avec l'agent CloudWatch
- Surveillance, journalisation et performances – Partenaires AWS Partner Network
- Documentation X-Ray

Vidéos connexes :

- Configuration de contrôles à grande échelle dans votre environnement AWS
- How Amazon uses better metrics for improved website performance

Exemples connexes :

- Création d'un tableau de bord avec Amazon QuickSight

- Niveau 100 : surveillance avec les tableaux de bord CloudWatch

Optimisation des coûts

Le pilier Optimisation des coûts comprend la possibilité d'exécuter des systèmes pour offrir une valeur métier au prix le plus bas. Vous trouverez des recommandations sur l'implémentation dans le Livre blanc du pilier Optimisation des coûts.

Domaines de bonnes pratiques

- Pratiques en matière de gestion financière du cloud
- Sensibilisation aux dépenses et à l'utilisation
- Ressources rentables
- Gérer la demande et les sources d'approvisionnement
- Optimiser dans le temps

Pratiques en matière de gestion financière du cloud

Question

- COÛT 1. Comment mettre en œuvre la gestion financière du cloud ?

COÛT 1. Comment mettre en œuvre la gestion financière du cloud ?

La gestion financière du cloud (CFM) permet aux organisations de générer de la valeur ajoutée et d'être financièrement performantes en optimisant leurs coûts et l'utilisation, et en se mettant à l'échelle sur AWS.

Bonnes pratiques

- COST01-BP01 S'approprier l'optimisation des coûts
- COST01-BP02 Établir un partenariat entre les équipes financières et technologiques
- COST01-BP03 Établir des budgets et des prévisions cloud
- COST01-BP04 Mettre en œuvre la sensibilisation aux coûts dans les processus organisationnels
- COST01-BP05 Rendre compte de l'optimisation des coûts
- COST01-BP06 Surveiller les coûts de manière proactive
- COST01-BP07 Suivre les nouvelles versions des services

- COST01-BP08 Créer une culture de sensibilisation aux coûts
- COST01-BP09 Quantifier la valeur ajoutée générée par l'optimisation des coûts

COST01-BP01 S'approprier l'optimisation des coûts

Créez une équipe (Bureau d'affaires du cloud, Centre d'excellence cloud ou équipe FinOps) chargée d'établir et de gérer la sensibilisation aux coûts dans toute votre organisation. Le propriétaire de l'optimisation des coûts peut être un individu ou une équipe (nécessite des personnes des équipes financières, technologiques et commerciales) qui comprend l'ensemble de l'organisation et la partie finance du cloud.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Il s'agit de la présentation d'une fonction ou d'une équipe (Bureau d'affaires du cloud ou Centre d'excellence cloud) chargée d'établir et de gérer une culture de sensibilisation aux coûts dans le cloud computing. Cette fonction peut être une personne existante, une équipe au sein de votre organisation ou une nouvelle équipe composée des principales parties prenantes de la finance, de la technologie et de l'organisation issues de toute l'entreprise.

La fonction (individu ou équipe) établit des priorités et consacre le pourcentage de temps requis aux activités de gestion et d'optimisation des coûts. Pour une petite organisation, la fonction peut consacrer un pourcentage de temps plus faible qu'une fonction à temps plein pour une grande entreprise.

Cette fonction (individu ou équipe) établit des priorités et consacre le pourcentage de temps requis aux activités de gestion et d'optimisation des coûts. Pour une petite organisation, la fonction peut consacrer un pourcentage de temps plus faible sur des activités de gestion et d'optimisation par rapport à une fonction à temps plein pour une grande entreprise.

La fonction exige une approche pluridisciplinaire, avec des capacités en gestion de projet, en science des données, en analyse financière et en développement de logiciels ou d'infrastructures. Elle peut améliorer l'efficacité de la charge de travail en procédant à des optimisations de coûts au sein de trois propriétés différentes :

- Centralisée : Grâce à des équipes désignées telles que l'équipe FinOps, l'équipe de gestion financière du cloud (CFM), le Bureau d'affaires du cloud (CBO) ou le Centre d'excellence cloud (CCoE), les clients peuvent concevoir et mettre en œuvre des mécanismes de gouvernance et mettre en place les bonnes pratiques à l'échelle de l'entreprise.

- Décentralisée : Influencer les équipes technologiques pour qu'elles optimisent les coûts.
- Hybride : Une combinaison des équipes centralisée et décentralisée peut collaborer pour exécuter les optimisations de coûts.

La fonction peut être mesurée par rapport à sa capacité à exécuter et à atteindre les objectifs d'optimisation des coûts (par exemple, les métriques d'efficacité de la charge de travail).

Vous devez obtenir un parrainage de la direction pour cette fonction, ce qui est un facteur de réussite clé. Le parrain est considéré comme un défenseur d'une consommation efficace du cloud et apporte son soutien dans le cadre de la remontée pour l'équipe afin de garantir que les activités d'optimisation des coûts sont traitées avec le niveau de priorité défini par l'organisation. Sinon, les conseils peuvent être ignorés et les opportunités d'économies ne seront pas prioritaires. Ensemble, le sponsor et l'équipe aident votre organisation à utiliser le cloud de manière efficace et apportent une valeur ajoutée.

Si vous disposez d'un plan Business, Enterprise-On-Ramp ou Enterprise Support et que vous avez besoin d'aide pour constituer cette équipe ou cette fonction, contactez vos experts en gestion financière dans le cloud (CFM) par l'intermédiaire de votre équipe de compte.

Étapes d'implémentation

- Définir les principaux membres : Toutes les parties concernées de votre organisation doivent contribuer à la gestion des coûts et s'y intéresser. En règle générale, les équipes comprennent la finance, les responsables d'application ou de produit, la direction et les équipes techniques (DevOps). Certaines sont impliquées à temps plein (finance ou technique), tandis que d'autres le sont périodiquement, en fonction des besoins. Les individus ou les équipes qui effectuent la CFM ont besoin des compétences suivantes :
  - Développement de logiciels : Lorsque des scripts et l'automatisation sont créés.
  - Ingénierie d'infrastructure : Pour déployer des scripts, automatiser des processus, et comprendre comment les services et les ressources sont provisionnés.
  - Perspicacité des opérations : La gestion financière de cloud signifie opérer efficacement sur le cloud en mesurant, en surveillant, en modifiant, en planifiant et en mettant à l'échelle l'utilisation efficace du cloud.
- Définir des objectifs et des métriques : La fonction doit apporter de la valeur à l'organisation de différentes manières. Ces objectifs sont définis et évoluent continuellement au rythme de l'organisation. Les activités courantes incluent la création et l'exécution de programmes de formation sur l'optimisation des coûts au sein de l'organisation, le développement de normes à

l'échelle de l'organisation, telles que la surveillance et la création de rapports pour l'optimisation des coûts, et la définition d'objectifs de charge de travail pour l'optimisation. Cette fonction doit également rendre compte régulièrement à l'organisation de sa capacité à optimiser les coûts.

Vous pouvez définir des indicateurs de performance clés (KPI) basés sur la valeur ou le coût. Lorsque vous définissez les indicateurs de performance clés, vous pouvez calculer le coût prévu en matière d'efficacité et les résultats métier attendus. Les KPI basés sur la valeur relient les métriques de coût et d'utilisation aux facteurs de valeur de l'entreprise et aident à rationaliser les changements dans les dépenses d'AWS. La première étape pour tirer profit des indicateurs de performance clés basés sur la valeur consiste à collaborer, d'un point de vue trans-organisationnel, pour sélectionner et convenir d'un ensemble standard d'indicateurs de performance clés.

- Définir une fréquence : Le groupe (équipes financières, technologiques et commerciales) doit se réunir régulièrement pour examiner ses objectifs et métriques. Une fréquence type implique d'examiner l'état de l'organisation, de passer en revue les programmes en cours, puis de vérifier les métriques financières et d'optimisation globales. Par la suite, les principales charges de travail font l'objet d'un rapport plus détaillé.

  Pendant ces examens réguliers, vous pouvez examiner l'efficacité (le coût) de la charge de travail et les résultats métier. Par exemple, une hausse de 20 % du coût d'une charge de travail peut correspondre avec une utilisation client accrue. Dans ce cas, cette hausse de 20 % du coût peut être interprétée comme un investissement. Ces réunions régulières peuvent aider les équipes à identifier les indicateurs de performance clés de valeur qui ont une signification pour toute l'entreprise.

Ressources

Documents connexes :

- Blog Centre d'excellence cloud AWS
- Création d'un bureau d'affaires du cloud
- Centre d'excellence cloud (CCoE)

Vidéos connexes :

- Témoignage de réussite : centre d'excellence cloud Vanguard

Exemples connexes :

- Utilisation d'un centre d'excellence cloud pour transformer l'entreprise entière
- Création d'un centre d'excellence cloud pour transformer l'entreprise entière
- 7 pièges à éviter lors de la création d'un centre d'excellence cloud

COST01-BP02 Établir un partenariat entre les équipes financières et technologiques

Impliquez les équipes financières et technologiques aux discussions sur les coûts et l'utilisation à toutes les étapes de votre transition vers le cloud. Les équipes se réunissent régulièrement et discutent de sujets tels que les objectifs et les cibles organisationnels, l'état actuel des coûts et l'utilisation et les pratiques financières et comptables.

Niveau d’exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Dans le cloud, les équipes technologiques innovent plus rapidement grâce à la réduction de la durée des cycles d'approbation, d'achat et de déploiement des infrastructures. Il peut s'agir d'un ajustement pour les organisations financières auparavant habituées à exécuter des processus longs et gourmands en ressources pour l'acquisition et le déploiement de capitaux dans les centres de données et les environnements sur site, et la répartition des coûts uniquement lors de l'approbation du projet.

Du point de vue d'un organisme financier et d'acquisition, le processus de budgétisation des capitaux, de demandes de capitaux, d'approbations, d'acquisitions et d'installation d'une infrastructure physique a été appris et standardisé durant des décennies :

- Les équipes ingénierie ou informatiques sont généralement les demandeurs
- Plusieurs équipes financières agissent en tant qu'approbateurs et acheteurs
- Les équipes opérations installent, intègrent et confient une infrastructure prête à l'emploi

Avec l'adoption du cloud, la consommation et l'acquisition d'infrastructure n'obéissent plus à une chaîne de dépendances. Dans le modèle cloud, les équipes technologiques et de produits ne se contentent plus de créer, elles sont les opérateurs et les propriétaires de leurs produits, responsables de la plupart des activités historiquement associées aux équipes financières et d'opérations, y compris l'acquisition et le déploiement.

Pour acquérir des ressources cloud, il suffit d'un compte utilisateur et des bonnes autorisations. C'est aussi ce qui réduit les risques informatiques et financiers ; ce qui signifie que les équipes ne sont qu'à quelques clics ou appels API d'arrêter des ressources cloud inactives ou inutiles. C'est également ce qui permet aux équipes technologiques d'innover plus rapidement : l'agilité et la capacité à mettre en place et à supprimer des expériences. Bien que la nature variable de la consommation cloud puisse

impacter la prévisibilité du point de vue de la prévision et de la budgétisation du capital, le cloud offre aux entreprises la possibilité de réduire les coûts de sur-approvisionnement, tout en diminuant les coûts d'opportunités associés au sous-approvisionnement conservateur.

Établissez un partenariat entre les principaux acteurs financiers et technologiques afin de créer une compréhension commune des objectifs organisationnels et de développer des mécanismes pour réussir financièrement dans le modèle de dépenses variables du cloud computing. Les équipes concernées au sein de votre organisation doivent être impliquées dans les discussions sur les coûts et l'utilisation à toutes les étapes de votre voyage dans le cloud, y compris :

- Responsables financiers : Les directeurs, contrôleurs, planificateurs financiers, les analystes métier, les responsables des achats, de l'approvisionnement et des comptes fournisseurs doivent comprendre le modèle de consommation du cloud, les options d'achat et le processus de facturation mensuelle. Les services financiers doivent s'associer aux équipes technologiques pour créer et socialiser une histoire de la valeur des TI et, ainsi, aider les équipes commerciales

à comprendre le lien entre les dépenses en technologie et les résultats commerciaux. Prises sous cet angle, les dépenses technologiques ne sont pas considérées comme des coûts, mais plutôt comme des investissements. En raison des différences fondamentales entre le cloud (telles que le taux de changement d'utilisation, la tarification à l'utilisation, la tarification progressive, les modèles de tarification et les informations détaillées sur la facturation et l'utilisation) par rapport à l'exploitation sur site, il est essentiel que l'organisme financier comprenne comment l'utilisation du cloud peut influer sur les aspects commerciaux, notamment les processus d'acquisition, le suivi des incitations, la répartition des coûts et les états financiers.

- Responsables technologiques : Les responsables technologiques (y compris les propriétaires de produits et d'applications) doivent être conscients des exigences financières (par exemple, les contraintes budgétaires), ainsi que des exigences métier (par exemple, les contrats de niveau de service). Cela permet de mettre en œuvre la charge de travail pour atteindre les objectifs souhaités de l'organisation.

Le partenariat entre la finance et la technologie offre les avantages suivants :

- Les équipes financières et technologiques bénéficient d'une visibilité quasiment en temps réel sur les coûts et l'utilisation.
- Les équipes financières et technologiques établissent une procédure d'exploitation standard pour gérer les variations des dépenses liées au cloud.
- Les acteurs financiers jouent le rôle de conseillers stratégiques en ce qui concerne la manière dont le capital est utilisé pour acheter des réductions sur engagement (par exemple, les instances réservées ou le modèle tarification flexible Savings Plans AWS), et la manière dont le cloud est utilisé pour développer l'organisation.
- Les processus existants de comptes fournisseurs et d'acquisition sont utilisés avec le cloud.
- Les équipes financières et technologiques collaborent à la prévision des coûts et de l'utilisation d'AWS afin d'aligner et de consolider les budgets de l'entreprise.
- Une meilleure communication inter-entreprise grâce à un langage partagé et une compréhension commune des concepts financiers.

Les autres parties prenantes au sein de votre organisation qui doivent être impliquées dans les discussions sur les coûts et l'utilisation sont notamment :

- Propriétaires de Business Units : Les propriétaires de Business Units doivent comprendre le modèle économique du cloud afin de pouvoir orienter les unités commerciales et l'entreprise dans

son ensemble. Cette connaissance du cloud est essentielle lorsqu'il est nécessaire de prévoir la croissance et l'utilisation de la charge de travail, et d'évaluer les options d'achat à plus long terme, telles que les instances réservées ou les Savings Plans.

- Équipe ingénierie : La mise en place d'un partenariat entre les équipes financières et technologiques est essentielle pour créer une culture de sensibilisation aux coûts capable d'encourager les ingénieurs à prendre des mesures sur la gestion financière du cloud. L'un des problèmes courants de la gestion financière du cloud ou des professionnels des opérations financières et des équipes financières est de faire comprendre aux ingénieurs l'ensemble de l'activité sur le cloud, de leur faire suivre les bonnes pratiques et de leur faire prendre des mesures recommandées.
- Tiers : Si votre entreprise fait appel à des tiers (par exemple, des consultants ou des outils), assurez-vous qu'ils sont en phase avec vos objectifs financiers et qu'ils peuvent le démontrer à la fois par leurs modèles d'engagement et par un retour sur investissement. En règle générale, les tiers contribueront à l'établissement de rapports et à l'analyse de toute charge de travail qu'ils gèrent, et ils fourniront une analyse des coûts de toute charge de travail qu'ils conçoivent.

La collaboration entre les équipes financières, technologiques et commerciales ainsi qu'un changement dans la manière dont les dépenses liées au cloud sont communiquées et évaluées au sein de l'organisation sont des préalables à la mise en œuvre et la réussite de la gestion financière du cloud. Incluez les équipes ingénierie afin qu'elles participent aux discussions sur le coût et l'utilisation à chaque étape, et les encourager à suivre les bonnes pratiques ainsi qu'à prendre les mesures convenues en conséquence.

Étapes d'implémentation

- Définir les principaux membres : Veillez à ce que tous les membres concernés de vos équipes financières et technologiques s'impliquent dans le partenariat. Les membres concernés dans l'équipe financière sont ceux qui interagissent avec le projet de loi sur le cloud. Il s'agit généralement de directeurs financiers, de contrôleurs financiers, de planificateurs financiers, d'analystes commerciaux et des financeurs. Les membres technologiques sont généralement les propriétaires de produit et d'application les responsables techniques et les représentants de toutes les équipes qui s'appuient sur le cloud. Les autres membres peuvent inclure les propriétaires d'unité commerciale, tels que le marketing qui influencera l'utilisation des produits. Il y a également des tiers, tels que des consultants afin d’assurer l'adéquation avec vos objectifs et vos mécanismes ainsi qu’une assistance pour les rapports d'activité.

- Définir les sujets de discussion : Définissez les sujets communs aux équipes ou qui nécessitent une compréhension commune. Suivez le coût à partir de sa création jusqu'au paiement de la facture. Notez tous les membres impliqués, ainsi que les processus organisationnels qui doivent être appliqués. Ayez une compréhension de chacune de ses étapes ou de chacun de ses processus et des informations associées, telles que les modèles de tarification disponibles, la tarification progressive, les modèles de réduction, la budgétisation et les exigences financières.
- Définir une fréquence : Pour créer un partenariat financier et technologique, mettez en place une cadence de communication régulière pour créer et maintenir un alignement. Le groupe doit se réunir régulièrement par rapport à ses objectifs et métriques. Une fréquence type implique d'examiner l'état de l'organisation, de passer en revue les programmes en cours, puis de vérifier les métriques financières et d'optimisation globales. Les principales charges de travail font l'objet d'un rapport plus détaillé.

Ressources

Documents connexes :

- Blog des actualités AWS

COST01-BP03 Établir des budgets et des prévisions cloud

Ajuster les processus existants de budgétisation et de prévision organisationnels afin qu'ils soient compatibles avec la nature hautement variable des coûts et de l'utilisation du cloud. Les processus doivent être dynamiques en utilisant des algorithmes basés sur les tendances ou les facteurs d'activité, ou une combinaison des deux.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les clients utilisent le cloud pour son efficacité, sa rapidité et son agilité, ce qui crée une quantité très variable de coûts et d'utilisation. Les coûts peuvent diminuer (ou parfois augmenter) avec l'augmentation de l'efficacité de la charge de travail ou à mesure que de nouvelles charges de travail et fonctionnalités sont déployées. Les charges de travail peuvent être mises à l'échelle pour servir un plus grand nombre de vos clients, ce qui augmentera l'utilisation du cloud et les coûts. Aujourd'hui, les ressources sont plus facilement accessibles que jamais. L'élasticité du cloud apporte également une élasticité des coûts et des prévisions. Les processus de budgétisation organisationnels existants doivent être modifiés pour intégrer cette variabilité.

Le budget est généralement préparé pour une seule année et reste fixe, exigeant un respect strict de la part de toutes les personnes concernées. En revanche, les prévisions sont plus souples, permettant des réajustements tout au long de l'année et fournissant des projections dynamiques sur une période d'un, deux ou trois ans. La budgétisation et les prévisions jouent toutes deux un rôle crucial dans l'établissement des attentes financières des différents acteurs technologiques et commerciaux. Des prévisions et une mise en œuvre précises permettent également de responsabiliser les parties prenantes qui sont directement en charge des coûts de provisionnement. De plus, les parties prenantes peuvent ainsi être sensibilisées aux coûts en général.

Ajustez les processus de budgétisation et de prévision existants pour les rendre plus dynamiques en utilisant soit un algorithme basé sur les tendances (utilisant les coûts historiques comme données d'entrée), soit des algorithmes basés sur les facteurs (par exemple, le lancement de nouveaux produits, l'expansion régionale ou de nouveaux environnements pour les charges de travail), ce qui est idéal pour un environnement de dépenses dynamique et variable, ou une combinaison de tendances et de facteurs commerciaux.

Vous pouvez utiliser AWS Cost Explorer pour effectuer des prévisions basées sur les tendances dans une plage temporelle future définie en fonction de vos dépenses passées. Le moteur de prévision de AWS Cost Explorer segmente vos données historiques en fonction des types de frais (par exemple, instances réservées) et utilise une combinaison de modèles de machine learning et de modèles basés sur des règles pour prédire les dépenses sur tous les types de frais individuellement.

Identifiez les facteurs commerciaux susceptibles d'avoir un impact sur votre coût d'utilisation et établissez des prévisions pour chacun d'entre eux séparément afin de veiller à ce que l'utilisation prévue soit calculée à l'avance. Certains de ces facteurs sont liés aux équipes informatiques et aux équipes chargées des produits au sein de l'organisation. D'autres facteurs commerciaux, tels que les événements commerciaux, les promotions, les fusions et les acquisitions, sont connus de vos responsables des ventes, du marketing et de l'entreprise, et il est important de collaborer et de tenir compte de tous ces moteurs de la demande également. Vous devez travailler en étroite collaboration avec eux pour comprendre l'impact sur les nouveaux facteurs internes.

Une fois que vous avez déterminé vos prévisions basées sur les tendances à l'aide de Cost Explorer ou d'un autre outil, utilisez l' AWS Pricing Calculator pour évaluer votre cas d'utilisation AWS et les coûts futurs en fonction de l'utilisation attendue (trafic, demandes par seconde ou instance requise Amazon EC2). Vous pouvez également les utiliser pour vous aider à prévoir vos dépenses, rechercher des opportunités de réaliser des économies et prendre des décisions éclairées lorsque vous utilisez AWS. Il est important de contrôler l'exactitude de ces prévisions, car les budgets doivent être établis sur la base de ces calculs et estimations prévisionnels.

Utilisez AWS Budgets pour établir des budgets personnalisés détaillés en spécifiant la période, la récurrence ou le montant (fixe ou variable), et en ajoutant des filtres tels que le service, la Région AWS et des balises. Pour rester informé des performances de vos budgets existants, vous pouvez créer et programmer des Rapports AWS Budgets qui sont régulièrement envoyés par e-mail à vous même ainsi qu'à vos parties prenantes. Vous pouvez également créer des alertes AWS Budgets basées sur les coûts réels, réactives par essence, ou sur les coûts prévus, ce qui vous donne le temps de mettre en place des mesures d'atténuation contre les dépassements de coûts potentiels. Vous pouvez être alerté lorsque votre coût ou votre utilisation dépassera ou devrait dépasser le montant prévu au budget.

Utilisez AWS Cost Anomaly Detection pour prévenir ou réduire les coûts inopinés et améliorer le contrôle sans ralentir le processus d'innovation. AWS Cost Anomaly Detection exploite le machine learning pour identifier les dépenses irrégulières et en déterminer les causes profondes, ce qui vous permet d'agir rapidement. En trois étapes simples, vous pouvez créer votre propre surveillance contextualisée et recevoir des alertes en cas de dépense irrégulière détectée.

Comme indiqué dans la section Partenariat financier et technologiquedu pilier Optimisation des coûts du cadre Well-Architected, il est important de mettre en place un partenariat et des cadences entre les services informatiques, les secteurs financiers et les autres parties prenantes afin de vérifier qu'ils utilisent tous les mêmes outils ou processus dans un souci de cohérence. Dans les cas où les budgets doivent être modifiés, une augmentation des points de contact de cadence peut permettre de réagir plus rapidement à ces changements.

Étapes d'implémentation

- Analysez les prévisions basées sur les tendances : Utilisez des outils de prévision basés sur les tendances, tels que AWS Cost Explorer et Amazon Forecast. Analysez votre coût d'utilisation en fonction de différentes dimensions comme le service, le compte, les balisages et les catégories de coûts. Si des prévisions avancées sont nécessaires, importez vos données AWS Cost and Usage Report dans Amazon Forecast (qui applique la régression linéaire comme forme de machine learning pour établir des prévisions).
- Analysez les prévisions basées sur les facteurs opérationnels : Identifiez l'impact des facteurs commerciaux sur votre utilisation du cloud et établissez des prévisions pour chacun d'entre eux séparément afin de calculer à l'avance le coût d'utilisation prévu. Travaillez en étroite collaboration avec les propriétaires d'unités commerciales et les parties prenantes pour comprendre l'impact sur les nouveaux facteurs et calculer les changements de coûts attendus afin de définir des budgets précis.

- Actualisez les processus de prévision et de budgétisation existants : Définissez vos processus de prévision budgétaires en vous basant sur les méthodes de prévision adoptées, telles que les méthodes basées sur les tendances, sur les facteurs commerciaux ou une combinaison de ces deux méthodes. Les budgets doivent être calculés et réalistes, sur la base de ces processus de prévision.
- Configurer des alertes et des notifications : Utilisez les alertes AWS Budgets et AWS Cost Anomaly Detection pour recevoir des alertes et des notifications.
- Effectuer des examens réguliers avec les principales parties prenantes : Par exemple, les parties prenantes des secteurs informatiques et des secteurs financiers, les équipes de plateforme et d'autres secteurs de l'entreprise doivent s'aligner avec les nouvelles orientations opérationnelles et les changements d'utilisation dans l'entreprise.

Ressources

Documents connexes :

- AWS Cost Explorer
- AWS Cost and Usage Report
- Prévision Amazon QuickSight
- Amazon Forecast
- AWS Budgets
- Blog des actualités AWS

Vidéos connexes :

- Comment utiliser AWS Budgets pour suivre mes dépenses et mon utilisation
- Série sur l'optimisation des coûts AWS : AWS Budgets

Exemples connexes :

- Understand and build driver-based forecasting
- How to establish and drive a forecasting culture
- How to improve your cloud cost forecasting
- Using the right tools for your cloud cost forecasting

COST01-BP04 Mettre en œuvre la sensibilisation aux coûts dans les processus organisationnels

Mettez en œuvre la sensibilisation aux coûts, créez une transparence et intégrez une sensibilisation à l'égard des coûts dans les processus nouveaux ou existants qui ont une incidence sur l'utilisation, et tirez parti des processus existants pour la sensibilisation aux coûts. Intégrez la sensibilisation aux coûts dans la formation des employés.

Niveau d’exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

La prise en compte des coûts doit être mise en œuvre dans les processus organisationnels nouveaux et existants. Il s'agit de l'une des capacités prérequises fondamentales pour les autres bonnes pratiques. Il est recommandé de réutiliser et de modifier les processus existants dans la mesure du possible, ce qui réduit l'impact sur l'agilité et la vitesse. Signalez les coûts de cloud aux équipes technologiques, aux décideurs de l'entreprise et aux équipes financières pour sensibiliser aux coûts, et établir des indicateurs de performance clés d'efficacité pour les parties prenantes financières et commerciales. Les recommandations suivantes vous aideront à mettre en œuvre la prise en compte des coûts dans votre charge de travail :

- Vérifiez que la gestion des modifications comprenne une mesure des coûts pour quantifier l'impact financier des modifications. Cela permet de répondre de manière proactive aux préoccupations liées aux coûts et de mettre en évidence les économies réalisées.
- Vérifiez que l'optimisation des coûts est une composante essentielle de vos capacités d'exploitation. Par exemple, vous pouvez tirer parti des processus de gestion des incidents existants pour investiguer et identifier les causes racines des anomalies de coût et d'utilisation ou surcoûts.
- Accélérez la réduction des coûts et la génération de valeur métier avec l'automatisation ou l'utilisation d'outils. Lorsque vous réfléchissez au coût de la mise en œuvre, encadrez la conversation pour y inclure une composante de retour sur investissement afin de justifier l'investissement en temps ou en argent.
- Allouez les coûts de cloud en implémentant des showbacks (récupération des données de facturation) ou des chargebacks (facturation) pour les dépenses de cloud, y compris les options d'achat basées sur l'engagement, les services partagés et les achats marketplace afin de stimuler la plupart de la consommation de cloud sensible aux coûts.
- Étendez les programmes de formation et de développement existants afin d'y inclure une formation de sensibilisation aux coûts dans toute votre entreprise. Il est recommandé d'inclure une

formation et une certification continues. Cela permettra de créer une organisation capable de gérer automatiquement les coûts et l'utilisation.

- Profitez des outils natifs AWS gratuits tels qu' AWS Cost Anomaly Detection, AWS Budgetset Rapports AWS Budgets.

Lorsque les entreprises adoptent de manière cohérente les pratiques de Gestion financière du cloud (CFM), ces comportements deviennent ancrés dans la façon de travailler et de prendre des décisions. Il en résulte une culture plus soucieuse des coûts, depuis les développeurs qui conçoivent une nouvelle application « née dans le cloud » jusqu'aux responsables financiers qui analysent le retour sur investissement de ces nouveaux investissements dans le cloud.

Étapes d'implémentation

- Identifier les processus organisationnels pertinents : Chaque unité organisationnelle passe en revue ses processus et identifie les processus qui ont un impact sur les coûts et l'utilisation. Tous les processus qui entraînent la création ou l'arrêt d'une ressource doivent être inclus dans la vérification. Recherchez des processus qui peuvent soutenir la prise en compte des coûts dans votre entreprise, tels que la gestion des incidents et la formation.
- Mettre en place une culture de sensibilisation aux coûts autonome : veillez à ce que toutes les parties prenantes pertinentes s'alignent avec la cause du changement et l'impact en tant que coût, afin qu'elles comprennent le coût du cloud. Cela permettra à votre entreprise de mettre en place une culture de sensibilisation aux coûts autonome de l'innovation.
- Mettre à jour les processus avec la sensibilisation aux coûts : Chaque processus est modifié pour tenir compte des coûts. Le processus peut nécessiter des contrôles préalables supplémentaires, tels que l'évaluation de l'impact du coût, ou des contrôles a posteriori validant que les changements attendus en matière de coût et d'utilisation se sont produits. Les processus de soutien, tels que la formation et la gestion des incidents, peuvent être étendus pour inclure des éléments relatifs au coût et à l'utilisation.

Pour obtenir de l'aide, contactez les experts de la gestion financière du cloud par le biais de l'équipe chargée de votre compte, ou parcourez les ressources et les documents associés ci-dessous.

Ressources

Documents connexes :

- Gestion financière du cloud AWS

Exemples connexes :

- Stratégie pour une gestion des coûts de cloud efficace
- Série de blog sur le contrôle des coûts #3 : comment gérer les augmentations de coûts
- Guide du débutant : AWS Cost Management

COST01-BP05 Rendre compte de l'optimisation des coûts

Mettez en place des budgets pour le cloud et configurez des mécanismes pour détecter les anomalies d'utilisation. Configurez les outils connexes pour les alertes de coût et d'utilisation par rapport à des objectifs prédéfinis et recevez des notifications lorsqu'une utilisation dépasse ces objectifs. Organisez des réunions régulières pour analyser la rentabilité de vos charges de travail et promouvoir la sensibilisation aux coûts.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

Vous devez régulièrement signaler l'optimisation des coûts et de l'utilisation au sein de votre entreprise. Vous pouvez mettre en œuvre des sessions dédiées pour discuter des performances en matière de coûts, ou inclure l'optimisation des coûts dans vos cycles de rapports opérationnels réguliers pour vos charges de travail. Utilisez des services et des outils pour contrôler régulièrement vos performances en matière de coûts et mettre en œuvre des possibilités d'économies.

Visualisez vos coûts et votre utilisation avec plusieurs filtres et de manière détaillée en utilisant AWS Cost Explorer, qui fournit des tableaux de bord et des rapports tels que les coûts par service ou par compte, les coûts quotidiens ou les coûts du marketplace. Suivez l'évolution de vos coûts et de votre utilisation par rapport aux budgets configurés avec les Rapports AWS Budgets.

Utilisez AWS Budgets pour définir des budgets personnalisés pour suivre vos coûts et votre utilisation, mais aussi répondre rapidement aux alertes reçues par e-mail ou notifications Amazon Simple Notification Service (Amazon SNS) en cas de dépassement de votre seuil. Définissez votre budget sur quotidien, mensuel, trimestriel ou annuel, puis définissez des limites budgétaires spécifiques pour rester informé de la progression des coûts et de l'utilisation réels et prévus par rapport à votre seuil budgétaire. Vous pouvez également définir des d'information et des mesures automatiques par rapport à ces alertes, ou via un processus d'approbation en cas de dépassement d'une cible budgétaire.

Mettez en œuvre des notifications sur les coûts et l'utilisation afin que les modifications puissent être rapidement prises en compte si elles sont imprévues. AWS Cost Anomaly Detection vous permet de réduire les coûts inopinés et améliorer le contrôle sans ralentir le processus d'innovation. AWS Cost Anomaly Detection identifie les dépenses irrégulières et les causes profondes, ce qui permet de réduire le risque de facturations inopinées. En trois étapes simples, vous pouvez créer votre propre surveillance contextualisée et recevoir des alertes en cas de dépense irrégulière détectée.

Vous pouvez également utiliser Amazon QuickSight avec des données (CUR) AWS Cost and Usage Report, afin de fournir des rapports hautement personnalisés avec des données plus détaillées. Amazon QuickSight vous permet de planifier des rapports et de recevoir des e-mails périodiques sur le rapport de coût pour connaître le coût et l'utilisation historiques, ou les opportunités d'économies. Consultez notre solution Cost Intelligence Dashboard (CID) intégrée sur Amazon QuickSight, qui vous offre une visibilité avancée.

Utilisez AWS Trusted Advisor, qui offre des conseils pour vérifier si les ressources allouées sont alignées avec les bonnes pratiques AWS pour des questions d'optimisation des coûts.

Vérifiez vos recommandations en matière de Savings Plans à l'aide de graphiques visuels en fonction de vos coûts et de votre utilisation. Des graphiques horaires présentent les dépenses à la demande en regard de l'engagement recommandé des Savings Plans, fournissant un aperçu des économies estimées, de la couverture des Savings Plans et de l'utilisation des Savings Plans. Cela permet aux organisations de comprendre les dépenses Savings Plans heure par heure sans avoir à investir du temps et des ressources dans l'élaboration de modèles pour analyser leurs dépenses.

Créez des rapports périodiques contenant les éléments clés des Savings Plans et des instances réservées, ainsi que les recommandations de redimensionnement d'Amazon EC2 (AWS Cost Explorer) pour commencer à réduire le coût associé aux charges de travail à état stable, ainsi qu'aux ressources inactives et sous-exploitées. Identifiez et récupérez les dépenses inutiles liées au cloud pour les ressources déployées. Les dépenses inutiles liées au cloud se produisent lorsque des ressources de taille inappropriée sont créées, ou des modèles d'utilisation différents sont observés au lieu de ce qui était prévu. Suivez les bonnes pratiques AWS pour réduire vos pertes ou demandez à l'équipe qui gère votre compte et à votre partenaire de vous aider à optimiser et économiser vos coûts de cloud.

Générez des rapports réguliers pour profiter de meilleures options d'achat pour vos ressources afin de réduire les coûts unitaires de vos charges de travail. Les options d'achat telles que les Savings Plans, les instances réservées ou les instances Spot Amazon EC2 offrent les meilleures économies pour les charges de travail tolérantes aux pannes et permettent aux parties prenantes (propriétaires d'entreprise, équipes financières et technologiques) de participer à ces discussions sur l'engagement.

Partagez les rapports contenant des opportunités ou des annonces de lancement capables de vous aider à réduire le coût total de possession (TCO) du cloud. Adoptez de nouveaux services, régions, fonctionnalités, solutions ou moyens de réduire davantage les coûts.

Étapes d'implémentation

- Configurer AWS Budgets : Configurez AWS Budgets sur tous les comptes de votre charge de travail. Définissez un budget pour les dépenses globales des comptes et un budget pour la charge de travail à l'aide de balises.
  - Ateliers Well-Architected : utilisation des coûts et de la gouvernance
- Rendre compte de l'optimisation des coûts : Définissez un cycle régulier pour discuter de l'efficacité de la charge de travail et l'analyser. À l'aide des métriques définies, rendez compte des métriques atteintes et du coût associé. Identifiez et corrigez les tendances négatives, tout en ciblant les tendances positives que vous pouvez promouvoir dans votre organisation. Les rapports devraient impliquer des représentants des finances, des équipes d'application et des propriétaires, ainsi que des décideurs clés en ce qui concerne les dépenses liées au cloud.

Ressources

Documents connexes :

- AWS Cost Explorer
- AWS Trusted Advisor
- AWS Budgets
- AWS Cost and Usage Report
- Bonnes pratiques AWS Budgets
- Amazon S3 Analytics

Exemples connexes :

- Ateliers Well-Architected : utilisation des coûts et de la gouvernance
- Méthodes clés pour commencer à optimiser vos coûts de cloud AWS

COST01-BP06 Surveiller les coûts de manière proactive

Mettez en œuvre des outils et des tableaux de bord pour surveiller de manière proactive les coûts de la charge de travail. Vérifiez régulièrement les coûts grâce aux outils configurés ou prêts à l'emploi. Ne vous contentez pas d'examiner les coûts et les catégories lorsque vous recevez des notifications. La surveillance et l'analyse des coûts de manière proactive permettent d'identifier les tendances positives et de les promouvoir dans toute votre organisation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Il est recommandé de surveiller les coûts et l'utilisation de manière proactive au sein de votre entreprise et pas seulement lorsque des exceptions ou des anomalies se présentent. Des tableaux de bord très visibles dans votre bureau ou votre environnement de travail garantissent que les personnes clés ont accès aux informations dont elles ont besoin et indiquent que l'organisation se concentre sur l'optimisation des coûts. Des tableaux de bord visibles vous permettent de promouvoir activement les résultats positifs et de les mettre en œuvre dans toute votre organisation.

Créez une routine quotidienne ou fréquente pour utiliser AWS Cost Explorer ou tout autre tableau de bord tel que Amazon QuickSight afin d'observer les coûts et de les analyser de manière proactive. Analysez les coûts et l'utilisation des services AWS au niveau du compte AWS, de la charge de travail ou des services AWS spécifiques avec le filtrage et le regroupement, et confirmez s'ils sont attendus ou non. Utilisez les balises ainsi que la granularité horaire et au niveau des ressources pour filtrer et identifier les coûts facturés pour les ressources principales. Vous pouvez également créer vos propres rapports grâce au Cost Intelligence Dashboard, une solution Amazon QuickSight créée par AWS Solutions Architects, et comparer vos budgets avec le coût et l'utilisation réels.

Étapes d'implémentation

- Rendre compte de l'optimisation des coûts : Définissez un cycle régulier pour discuter de l'efficacité de la charge de travail et l'analyser. À l'aide des métriques définies, rendez compte des métriques atteintes et du coût associé. Identifiez et corrigez les tendances négatives, et ciblez les tendances positives à promouvoir dans votre organisation. Les rapports doivent impliquer des représentants des équipes et des propriétaires d'application, de la finance et de la gestion.
- Créer et activer la granularité quotidienne AWS Budgets pour le coût et l'utilisation afin de prendre des mesures au moment opportun pour prévenir tout dépassement potentiel des coûts : AWS Budgets vous permet de configurer des notifications d'alerte, afin de rester informé si l'un de vos types de budget dépasse les seuils pré-configurés. Le meilleur moyen d'exploiter

AWS Budgets est de définir votre coût et votre utilisation prévus comme vos limites, afin que tout ce qui se situe au-dessus de vos budgets soit considéré comme un dépassement.

- Créer AWS Cost Anomaly Detection pour surveiller les coûts : AWS Cost Anomaly Detection utilise la technologie avancée de machine learning pour identifier les dépenses anormales et les causes profondes, afin que vous puissiez rapidement prendre des mesures. Cela vous permet de configurer des surveillances de coûts qui définissent les segments de dépenses que vous souhaitez évaluer (par exemple, services AWS individuels, comptes membres, balises de répartition des coûts et catégories de coûts),mais aussi de définir quand, où et comment vous recevez vos notifications d'alerte. Pour chaque surveillance, attachez plusieurs abonnements d'alerte pour les propriétaires d'entreprise et les équipes technologiques, notamment un nom, un seuil d'impact du coût et une fréquence d'alerte (alertes individuelles, résumé quotidien, résumé hebdomadaire) pour chaque abonnement.
- Utiliser AWS Cost Explorer ou intégrer vos données AWS Cost and Usage Report (CUR) aux tableaux de bord Amazon QuickSight pour visualiser les coûts de votre entreprise : AWS Cost Explorer possède une interface facile à utiliser qui vous permet de visualiser, de comprendre et de gérer vos coûts ainsi que l'utilisation AWS au fil du temps. La version Cost Intelligence Dashboard est un tableau de bord personnalisable et accessible pour aider à poser les bases de votre propre outil de gestion et d'optimisation des coûts.

Ressources

Documents connexes :

- AWS Budgets
- AWS Cost Explorer
- Coûts et budgets d'utilisation au quotidien
- AWS Cost Anomaly Detection

Exemples connexes :

- Ateliers Well-Architected : visualisation
- Ateliers Well-Architected : visualisation avancée
- Ateliers Well-Architected : Cloud Intelligence Dashboards
- Ateliers Well-Architected : visualisation des coûts
- Alerte AWS Cost Anomaly Detection avec Slack

COST01-BP07 Suivre les nouvelles versions des services

Travaillez régulièrement avec des experts ou des partenaires AWS pour identifier les services et les fonctionnalités qui offrent les coûts les plus bas. Examinez les blogs AWS et d'autres sources d'informations.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

AWS ajoute constamment de nouvelles capacités pour tirer parti des dernières technologies pour expérimenter et innover plus rapidement. Vous pourrez mettre en œuvre de nouveaux services et fonctionnalités AWS pour accroître la rentabilité de votre charge de travail. Consultez régulièrement la Gestion des coûts AWSle Blog des actualités AWSle blog sur la gestion des coûts AWSet les nouveautés AWS pour en savoir plus sur les nouvelles versions de services et de fonctionnalités. Les articles Nouveautés offrent un rapide aperçu de l'ensemble des services AWS, fonctionnalités et annonces d'expansion de régions dès leur lancement.

Étapes d'implémentation

- S'abonner aux blogs : Accédez aux pages des blogs AWS et abonnez-vous aux blogs des nouveautés et aux autres blogs pertinents. Vous pouvez vous inscrire sur la page des préférences de communication avec votre adresse e-mail.
- S'abonner aux Actualités AWS : Consultez régulièrement le Blog des actualités AWS et les nouveautés AWS pour en savoir plus sur les nouvelles versions de services et de fonctionnalités. Abonnez-vous au flux RSS ou avec votre e-mail pour suivre les annonces et les lancements.
- Suivre les réductions de prix AWS : les baisses de prix régulières sur tous nos services se sont imposées comme une méthode standard pour AWS permettant de transmettre l'efficacité économique obtenue grâce à notre mise à l'échelle. À compter d'avril 2022, AWS aura réduit les prix 115 fois depuis son lancement en 2006. Si vous avez des décisions métier en attente en raison d'inquiétudes concernant les prix, vous pouvez les examiner de nouveau après les réductions de prix et l'intégration de nouveaux services. Vous pouvez en savoir plus sur les précédentes réductions de prix, notamment les instances Amazon Elastic Compute Cloud (Amazon EC2), dans la catégorie réduction des prix du blog Actualités AWS.
- Événements et réunions AWS : Participez à votre conférence AWS et à toutes les réunions locales avec les autres organisations de votre région. Si vous ne pouvez pas y assister en personne, participez aux événements virtuels pour entendre les experts AWS et en savoir plus sur les cas métier des autres clients.

- Se réunir avec l'équipe chargée de votre compte : Planifiez un rythme régulier avec l'équipe chargée de votre compte, réunissez-vous et discutez des tendances du secteur et des services AWS. Parlez avec votre gestionnaire de compte, votre architecte de solutions et votre équipe de support.

Ressources

Documents connexes :

- Gestion des coûts AWS
- les nouveautés AWS
- Blog des actualités AWS

Exemples connexes :

- Amazon EC2 : 15 ans d'optimisation et de réduction des coûts informatiques
- Blog Actualités AWS : réduction des prix

COST01-BP08 Créer une culture de sensibilisation aux coûts

Mettez en œuvre des modifications ou des programmes dans toute votre entreprise afin de créer une culture de sensibilisation aux coûts. Il est recommandé de commencer petit, puis, au fur et à mesure que vos capacités augmentent et que votre organisation utilise le cloud, de mettre en œuvre des programmes de grande envergure.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

Une culture de sensibilisation aux coûts vous permet de déployer à grande échelle l'optimisation des coûts et la gestion financière dans le cloud (opérations financières, centre d'excellence cloud, équipes des opérations cloud, etc.) grâce aux bonnes pratiques qui sont appliquées de manière organique et décentralisée dans toute votre entreprise. La sensibilisation aux coûts permet de créer de hauts niveaux de capacité dans toute votre organisation avec un minimum d'efforts, par rapport à une approche centralisée et descendante stricte.

La création d'une sensibilisation aux coûts dans le cloud computing, notamment pour les principaux facteurs de coût, permet aux équipes de comprendre les résultats attendus de n'importe quel

changement en matière de coût. Les équipes qui accèdent aux environnements de cloud doivent connaître les modèles de tarification et la différence entre les centres de données sur site traditionnels et le cloud computing.

Le principal avantage d'une culture de sensibilisation aux coûts est que les équipes technologiques optimisent les coûts de manière proactive et continue (par exemple, elle est considérée comme une exigence non fonctionnelle lors de la création de l'architecture des nouvelles charges de travail ou de la modification de charges de travail existantes) au lieu de procéder à des optimisations de coûts réactives si nécessaire.

De petits changements de culture peuvent avoir de grandes répercussions sur l'efficacité de votre charge de travail actuelle et future. Voici quelques exemples :

- Donnez de la visibilité et créez de la sensibilisation dans les équipes ingénierie pour comprendre ce qu'elles font et leur impact en termes de coûts.
- Ludification des coûts et de l'utilisation dans votre entreprise. Cela peut se faire au moyen d'un tableau de bord visible du public ou d'un rapport qui compare les coûts normalisés et l'utilisation par les différentes équipes (par exemple, le coût par charge de travail et le coût par transaction).
- Reconnaissance de la rentabilité. Récompenser les réalisations volontaires ou non sollicitées en matière d'optimisation des coûts, publiquement ou en privé, et tirer les leçons des erreurs pour éviter de les répéter à l'avenir.
- Créez des exigences organisationnelles hiérarchisées pour que les charges de travail soient exécutées selon des budgets prédéfinis.
- Questionnez les exigences métier en matière de changements, et l'impact du coût des changements demandés apportés à l'infrastructure de l'architecture ou la configuration de charge de travail, pour veiller à payer uniquement ce dont vous avez besoin.
- Veillez à ce que le planificateur de changements soit informé des changements attendus ayant un impact sur le coût, et qu'ils soient confirmés par les parties prenantes pour fournir des résultats métier de manière rentable.

Étapes d'implémentation

- Signaler les coûts de cloud aux équipes technologiques : pour sensibiliser aux coûts, et établir des indicateurs de performance clés d'efficacité pour les parties prenantes financières et commerciales.
- Informer les parties prenantes ou les membres de l'équipe des changements planifiés : créez un point à l'ordre du jour pour discuter des changements planifiés et de l'impact coûts-avantages sur la charge de travail lors des réunions hebdomadaires sur les changements.

- Se réunir avec l'équipe chargée de votre compte : planifiez une réunion régulière avec l'équipe chargée de votre compte et discutez des tendances du secteur et des services AWS. Parlez avec votre gestionnaire de compte, architecte et équipe de support.
- Partager des témoignages de réussite : partagez des témoignages de réussite sur la réduction des coûts pour n'importe quelle charge de travail, Compte AWS ou entreprise afin de créer une attitude positive et des encouragements autour de l'optimisation des coûts.
- Entraîner : veillez à ce que les équipes techniques ou les membres de l'équipe soient entraînés pour la sensibilisation des coûts liés aux ressources sur AWS Cloud.
- Événements et réunions AWS : participez aux conférences AWS et à toutes les réunions locales avec les autres organisations de votre région.
- S'abonner aux blogs : Accédez aux pages des blogs AWS et abonnez-vous au blog sur les nouveautés et aux autres blogs pertinents pour suivre les lancements, les implémentations, les exemples et les changements partagés par AWS.

Ressources

Documents connexes :

- Blog AWS
- Gestion des coûts AWS
- Blog des actualités AWS

Exemples connexes :

- Gestion financière du cloud AWS
- Ateliers AWS Well-Architected : gestion financière du cloud

COST01-BP09 Quantifier la valeur ajoutée générée par l'optimisation des coûts

Quantifier la valeur métier obtenue grâce à l'optimisation des coûts permet de comprendre l'ensemble des avantages pour votre entreprise. Parce que l'optimisation des coûts est un investissement nécessaire, la quantification de la valeur ajoutée vous permet d'expliquer le retour sur investissement aux parties prenantes. La quantification de la valeur ajoutée peut vous aider à obtenir une meilleure adhésion des parties prenantes aux investissements futurs en matière d'optimisation des coûts, et fournit un cadre pour mesurer les résultats des activités d'optimisation des coûts de votre organisation.

Niveau de risque exposé si cette bonne pratique n'est pas établie: moyen

Directives d'implémentation

Quantifier la valeur métier signifie mesurer le bénéfice que les entreprises retirent des actions et des décisions qu'elles prennent. La valeur métier peut être tangible (comme la réduction des dépenses ou l'augmentation des profits) ou intangible (comme l'amélioration de la réputation de la marque ou l'augmentation de la satisfaction client).

Quantifier la valeur métier résultant de l'optimisation des coûts signifie déterminer la valeur ou le bénéfice que vous retirez de vos efforts pour dépenser plus efficacement. Par exemple, si une entreprise dépense 100 000 USD pour déployer une charge de travail sur AWS et l'optimise par la suite, le nouveau coût n'est que 80 000 USD sans sacrifier la qualité ou le résultat. Dans ce scénario, la valeur métier quantifiée résultant de l'optimisation des coûts représenterait une économie de 20 000 USD. Mais au-delà des simples économies, l'entreprise peut également quantifier la valeur en termes de rapidité de livraison, d'amélioration de la satisfaction client ou d'autres indicateurs résultant des efforts d'optimisation des coûts. Les parties prenantes doivent prendre des décisions concernant la valeur potentielle de l'optimisation des coûts, le coût de l'optimisation de la charge de travail et la valeur de retour.

En plus de faire état des économies réalisées grâce à l'optimisation des coûts, il est recommandé de quantifier la valeur métier générée. Les avantages de l'optimisation des coûts sont généralement quantifiés en termes de réduction des coûts par résultat commercial. Par exemple, vous pouvez quantifier les économies réalisées avec Amazon Elastic Compute Cloud(Amazon EC2) lorsque vous achetez des Savings Plans, qui permettent de réduire les coûts et de maintenir les niveaux de production de la charge de travail. Vous pouvez quantifier les économies réalisées dans au niveau des dépenses AWS lorsque des instances Amazon EC2 inactives sont mises hors service ou que des volumes Amazon Elastic Block Store (Amazon EBS) non attachés sont supprimés.

Les avantages de l'optimisation des coûts vont toutefois au-delà de la réduction ou de l'évitement des coûts. Envisagez de saisir des données supplémentaires pour mesurer les améliorations de l'efficacité et la valeur ajoutée.

Étapes d'implémentation

- Évaluer le bénéfice pour l'entreprise : il s'agit d'analyser et d'ajuster les coûts AWS Cloud de manière à maximiser le bénéfice tiré de chaque dollar dépensé. Au lieu de vous concentrer sur la réduction des coûts sans valeur métier, tenez compte du bénéfice et du retour sur investissement de l'optimisation des coûts, ce qui peut vous permettre de rentabiliser davantage l'argent que vous

dépensez. Il s'agit de dépenser judicieusement et de réaliser des investissements et des dépenses dans les secteurs qui génèrent le meilleur retour.

- Analyser les coûts AWS prévisionnels : la prévision des coûts permet aux acteurs financiers de définir des attentes avec d'autres acteurs internes et externes de l'organisation, et contribue à améliorer la prévisibilité financière de votre organisation. AWS Cost Explorer peut être utilisé pour effectuer des prévisions concernant vos coûts et votre utilisation.

Ressources

Documents connexes :

- Centre de rentabilité du Cloud
- Blog AWS
- AWS Cost Management
- Blog des actualités AWS
- Livre blanc du pilier Fiabilité de Well-Architected
- AWS Cost Explorer

Vidéos connexes :

- Unlock Business Value with Windows on AWS

Exemples connexes :

- Measuring and Maximizing the Business Value of Customer 360
- The Business Value of Adopting Amazon Web Services Managed Databases
- The Business Value of Amazon Web Services for Independent Software Vendors
- Business Value of Cloud Modernization
- The Business Value of Migration to Amazon Web Services

Sensibilisation aux dépenses et à l'utilisation

Questions

- COÛT 2. Comment gérer l'utilisation ?

- COÛT 3. Comment surveillez-vous vos coûts et votre utilisation ?
- COÛT 4. Comment mettez-vous les ressources hors service ?

COÛT 2. Comment gérer l'utilisation ?

Définissez des stratégies et des mécanismes pour vous assurer que les coûts appropriés sont facturés tout en atteignant les objectifs. En adoptant une approche d'équilibre des pouvoirs, vous pouvez innover sans dépense excessive.

Bonnes pratiques

- COST02-BP01 Développer des stratégies en fonction des exigences de votre organisation
- COST02-BP02 Mettre en œuvre des objectifs et des cibles
- COST02-BP03 Mettre en œuvre une structure de compte
- COST02-BP04 Mettre en œuvre des groupes et des rôles
- COST02-BP05 Mettre en œuvre des contrôles de coûts
- COST02-BP06 Suivre le cycle de vie du projet

COST02-BP01 Développer des stratégies en fonction des exigences de votre organisation

Développez des politiques qui définissent la manière dont les ressources sont gérées par votre organisation et inspectez-les régulièrement. Les stratégies doivent couvrir les aspects de coût des ressources et des charges de travail, y compris la création, la modification et la mise hors service pendant la durée de vie des ressources.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Il est essentiel de comprendre les coûts et les facteurs de coûts de votre entreprise pour les gérer efficacement et identifier les possibilités de réduction. Les organisations exécutent généralement plusieurs charges de travail exécutées par plusieurs équipes. Ces équipes peuvent appartenir à différentes unités commerciales, chacune ayant ses propres sources de revenus. La possibilité d'attribuer le coût des ressources aux charges de travail, à l'organisation individuelle ou aux propriétaires de produits permet d'adopter un comportement d'utilisation efficace et contribue à réduire les pertes. La surveillance précise des coûts et de l'utilisation vous aide à comprendre dans quelle mesure une charge de travail est optimisée, ainsi que le degré de rentabilité des unités

organisationnelles et des produits. Ces connaissances permettent de prendre des décisions plus éclairées quant à l'affectation des ressources au sein de votre organisation. La sensibilisation à l'utilisation à tous les niveaux de l'organisation est la clé du changement, car les changements d'utilisation entraînent des changements dans les coûts. Envisagez d'adopter une approche multidimensionnelle pour prendre conscience de votre utilisation et de vos dépenses.

La première étape de la gouvernance consiste à utiliser les exigences de votre entreprise pour élaborer des stratégies d'utilisation du cloud. Ces politiques définissent la façon dont votre organisation utilise le cloud et dont les ressources sont gérées. Les politiques doivent couvrir tous les aspects des ressources et des charges de travail qui ont trait au coût ou à l'utilisation, y compris la création, la modification et la mise hors service pendant la durée de vie de la ressource. Vérifiez que les politiques et les procédures sont suivies et mises en œuvre en cas de changement dans un environnement cloud. Lors de vos réunions sur la gestion des changements informatiques, posez des questions afin de connaître l'impact du coût des changements prévus, qu'il s'agisse d'une augmentation ou d'une baisse, la justification opérationnelle et le résultat attendu.

Les politiques doivent être simples afin qu'elles soient aisément compréhensibles et puissent être mises en œuvre efficacement dans toute l'organisation. Les politiques doivent également être faciles à suivre et à interpréter (afin qu'elles soient utilisées) et être spécifiques (aucune mauvaise interprétation entre les équipes). En outre, elles doivent être inspectées périodiquement (comme nos mécanismes) et mises à jour à mesure que les conditions commerciales ou les priorités des clients évoluent, ce qui rendrait la politique obsolète.

Commencez par des politiques générales de haut niveau, telles que la région géographique à utiliser ou les moments de la journée où les ressources doivent fonctionner. Affinez progressivement les politiques des différentes unités organisationnelles et des charges de travail. Les politiques communes comprennent les services et les fonctionnalités qui peuvent être utilisés (par exemple, un stockage moins performant dans les environnements de test et de développement), les types de ressources qui peuvent être utilisés par différents groupes (par exemple, la plus grande taille de ressource dans un compte de développement est moyenne) et la durée d'utilisation de ces ressources (qu'elle soit temporaire, courte ou spécifique).

Exemple de politique

Vous trouverez ci-dessous un exemple de politique que vous pouvez consulter pour créer vos propres politiques de gouvernance du cloud, axées sur l'optimisation des coûts. Assurez-vous d'ajuster la politique en fonction des exigences de votre organisation et des demandes de vos parties prenantes.

- Nom de la politique : définissez un nom de politique clair, par exemple Politique d'optimisation des ressources et de réduction des coûts.
- Objectif : expliquez pourquoi cette politique doit être utilisée et quel est le résultat attendu. L'objectif de cette politique est de vérifier qu'il existe un coût minimum requis pour déployer et exécuter la charge de travail souhaitée afin de répondre aux exigences de l'organisation.
- Champ d'application : définissez clairement qui doit utiliser cette politique et quand elle doit être utilisée, par exemple DevOps X Team doit utiliser cette politique pour les clients US, côte est, pour l'environnement X (production ou hors production).

Déclaration de politique

1. Sélectionnez 1 ou plusieus région US, côte est, en fonction de l'environnement de votre charge de travail et des exigences métier (développement, tests d'acceptation par les utilisateurs, préproduction ou production).
2. Programmez des instances Amazon EC2 et Amazon RDS qui devront être exécutées entre six heures et vingt heures (heure normale de l'Est (EST)).
3. Arrêtez toutes les instances Amazon EC2 non utilisées après huit heures et les instances Amazon RDS non utilisées après 24 heures d'inactivité.
4. Mettez fin à toutes les instances Amazon EC2 non utilisées après 24 heures d'inactivité dans les environnements hors production. Rappelez au propriétaire de l'instance Amazon EC2 (en fonction des balises) de revoir ses instances Amazon EC2 arrêtées en production et de l'informer que ses instances Amazon EC2 seront résiliées dans les 72 heures si elles ne sont pas utilisées.
5. Utilisez une famille et une taille d'instance génériques, telles que m5.large, puis redimensionnez l'instance en fonction de l'utilisation du processeur et de la mémoire avec AWS Compute Optimizer.
6. Priorisez l'utilisation de la mise à l'échelle automatique pour ajuster dynamiquement le nombre d'instances en cours d'exécution en fonction du trafic.
7. Utilisez des instances Spot pour les charges de travail non critiques.
8. Passez en revue les exigences en matière de capacité pour valider des plans d'épargne ou des instances réservées pour des charges de travail prévisibles et informez l'équipe de gestion financière du cloud.
9. Utilisez des politiques de cycle de vie Amazon S3 pour déplacer les données rarement consultées vers des niveaux de stockage moins coûteux. Si aucune politique de rétention n'est définie, utilisez

la hiérarchisation Amazon S3 intelligente pour déplacer automatiquement les objets vers le niveau archivé.

10. Surveillez l'utilisation des ressources et définissez des alarmes pour déclencher des événements de dimensionnement à l'aide d'Amazon CloudWatch.
11. Pour chaque Compte AWS, utilisez AWS Budgets pour définir les budgets de coûts et d'utilisation de votre compte en fonction du centre de coûts et des unités commerciales.
12. L'utilisation d'AWS Budgets pour définir les budgets de coûts et d'utilisation de votre compte peut vous aider à maîtriser vos dépenses et à éviter les factures imprévues, ce qui vous permet de mieux contrôler vos coûts.

Procédure : fournissez des procédures détaillées pour la mise en œuvre de cette politique ou consultez d'autres documents qui décrivent comment mettre en œuvre chaque déclaration de politique. Cette section doit fournir des instructions détaillées pour la mise en œuvre des exigences de la politique.

Pour mettre en œuvre cette politique, vous pouvez utiliser divers outils tiers ou règles AWS Config afin de vérifier la conformité avec la déclaration de politique et de déclencher des actions correctives automatisées à l'aide des fonctions AWS Lambda. Vous pouvez également utiliser AWS Organizations pour appliquer la politique. En outre, vous devez régulièrement revoir votre utilisation des ressources et ajuster la politique si nécessaire pour vérifier qu'elle continue de répondre aux besoins de votre organisation.

Étapes d'implémentation

- Rencontrez les parties prenantes : pour élaborer des politiques, demandez aux parties prenantes (bureaux commerciaux du cloud, ingénieurs ou décideurs fonctionnels chargés de l'application des politiques) au sein de votre organisation de spécifier leurs exigences et de les documenter. Adoptez une approche itérative en commençant par les grandes lignes et en affinant continuellement jusqu'aux plus petites unités à chaque étape. Les membres de l'équipe incluent ceux qui sont directement impliqués dans la charge de travail, tels que les unités d'organisation ou les propriétaires d'application, ainsi que les groupes de soutien, tels que les équipes de sécurité et financières.
- Obtenez une confirmation : assurez-vous que les équipes s'accordent sur les politiques décrivant qui peut accéder au AWS Cloud et y faire des déploiements. Vérifiez qu'elles suivent les politiques de votre organisation et confirmez que leurs créations de ressources s'alignent sur les politiques et les procédures convenues.

- Créez des sessions de formation d'intégration : demandez aux nouveaux membres de l'organisation de suivre des cours de formation d'intégration afin de les sensibiliser aux coûts et aux exigences de l'organisation. Ils peuvent supposer des politiques différentes issues de leur expérience passée ou ne pas en connaître du tout.
- Définissez des emplacements pour votre charge de travail : Définissez l'emplacement d'exécution de votre charge de travail, y compris le pays et la zone du pays. Ces informations seront utilisées pour l'association aux Régions AWS et aux zones de disponibilité AWS.
- Définir et regrouper les services et les ressources : Définissez les services dont les charges de travail ont besoin. Pour chaque service, spécifiez les types, la taille et le nombre de ressources requis. Définissez des groupes pour les ressources par fonction, tels que les serveurs d'applications ou le stockage de base de données. Les ressources peuvent appartenir à plusieurs groupes.
- Définir et regrouper les utilisateurs par fonction : Définissez les utilisateurs qui interagissent avec la charge de travail, en vous concentrant sur ce qu'ils font et sur la façon dont ils l'utilisent, et non pas sur leur identité ou leur poste au sein de l'organisation. Regroupez les utilisateurs ou fonctions similaires. Vous pouvez utiliser les politiques gérées par AWS comme guide.
- Définir les actions : En utilisant les emplacements, les ressources et les utilisateurs identifiés précédemment, définissez les actions requises par chacun pour atteindre les résultats de la charge de travail pendant sa durée de vie (développement, exploitation et mise hors service). Identifiez les actions en fonction des groupes, et non pas des éléments individuels des groupes, dans chaque emplacement. Commencez globalement avec la lecture ou l'écriture, puis affinez vers des actions spécifiques pour chaque service.
- Définir la période de vérification : Les charges de travail et les exigences organisationnelles peuvent changer au fil du temps. Définissez le calendrier de révision de la charge de travail pour qu'il reste conforme aux priorités de l'organisation.
- Documentez les politiques : assurez-vous que les politiques définies sont accessibles en fonction des besoins de votre organisation. Ces politiques sont utilisées pour mettre en œuvre, gérer et auditer l'accès de vos environnements.

Ressources

Documents connexes :

- Gestion des changements dans le cloud
- Stratégies gérées par AWS pour les fonctions professionnelles

- Stratégie de facturation multicompte AWS
- Actions, ressources et clés de condition pour les services AWS
- Gestion et gouvernance AWS
- Contrôler l'accès aux Régions AWS avec des politiques IAM
- Régions et AZ des Infrastructures mondiales

Vidéos connexes :

- AWS Management and Governance at Scale

Exemples connexes :

- VMware – Quelles sont les politiques cloud ?

COST02-BP02 Mettre en œuvre des objectifs et des cibles

Mettez en œuvre les objectifs et cibles de coût et d'utilisation de votre charge de travail. Les objectifs fournissent une orientation à votre organisation sur les résultats attendus et les cibles fournissent des résultats mesurables spécifiques à atteindre pour vos charges de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Développez des objectifs et des cibles de coût et d'utilisation pour votre entreprise. En tant qu'organisation en pleine croissance sur AWS, il est important de définir et de suivre des objectifs d'optimisation des coûts. Ces objectifs ou indicateurs clés de performance (KPI) peuvent inclure des éléments comme le pourcentage des dépenses à la demande ou l'adoption de certains services optimisés tels que les instances AWS Graviton ou les types de volumes EBS gp3. La définition d'objectifs mesurables et réalisables peut vous aider à continuer à mesurer les améliorations de l'efficacité, ce qui est important pour les opérations métier en cours. Les objectifs fournissent des conseils et des directives à votre organisation en ce qui concerne les résultats attendus. Les cibles fournissent des résultats mesurables spécifiques à atteindre. En résumé, un objectif est la direction que vous voulez prendre et une cible correspond à la distance dans cette direction et au moment où cet objectif doit être atteint (à l'aide de conseils SMART, c'est-à-dire spécifiques, mesurables, acceptables, réalistes et temporellement définis). Voici un exemple d'objectif : l'utilisation de la plateforme doit augmenter de manière significative, avec seulement une augmentation mineure

(non linéaire) des coûts. Voici un exemple de cible : une augmentation de 20 % de l'utilisation de la plateforme, avec une augmentation des coûts inférieure à 5 %. Voici un autre exemple d'objectif commun : les charges de travail doivent être plus efficaces tous les six mois. La cible qui correspondrait à cet objectif serait de faire en sorte que les indicateurs du coût par entreprise diminuent de 5 % tous les six mois.

L'un des objectifs de l'optimisation des coûts est d'accroître l'efficacité de la charge de travail, c'est-à-dire de réduire le coût par résultat métier de la charge de travail au fil du temps. Il est recommandé de mettre en œuvre cet objectif pour toutes les charges de travail et de définir également un objectif tel qu'une augmentation de 5 % de l'efficacité tous les six mois à un an. Cela peut être réalisé dans le cloud en développant des capacités d'optimisation des coûts et en publiant de nouveaux services et fonctionnalités.

Il est important de disposer d'une visibilité en temps quasi réel sur vos KPI et les opportunités d'économies associées, et de suivre vos progrès au fil du temps. Pour commencer à définir et à suivre les objectifs des KPI, nous vous recommandons d'utiliser le tableau de bord des KPI du cadre Cloud Intelligence Dashboards (CID). Sur la base des données issues d'AWS Cost and Usage Report, le tableau de bord des KPI fournit une série de KPI recommandés pour l'optimisation des coûts, avec la possibilité de définir des objectifs personnalisés et de suivre les progrès au fil du temps.

Si vous disposez d'une autre solution qui vous permet de définir et de suivre des objectifs de KPI, assurez-vous qu'elle est adoptée par toutes les parties prenantes de votre organisation en matière de gestion financière dans le cloud.

Étapes d'implémentation

- Définissez les niveaux d'utilisation attendus : pour commencer, concentrez-vous sur les niveaux d'utilisation. Collaborez avec les propriétaires d'application, les équipes marketing et les équipes commerciales plus importantes afin de comprendre quels seront les niveaux d'utilisation attendus pour la charge de travail. Comment la demande des clients évoluera-t-elle dans le temps et existera-t-il des changements dus à des augmentations saisonnières ou à des campagnes marketing ?
- Définissez les coûts et les ressources de la charge de travail : une fois les niveaux d'utilisation définis, quantifiez les modifications des ressources de charge de travail nécessaires pour atteindre ces niveaux d'utilisation. Il sera peut-être nécessaire d'augmenter la taille des ressources d'un composant de la charge de travail ou leur nombre, d'accroître le transfert de données ou de remplacer les composants de la charge de travail par un service différent à un niveau spécifique.

Précisez quels seront les coûts à chacun de ces points majeurs, et quels seront les changements de coûts lorsqu'il existera des changements d'utilisation.

- Définir les objectifs commerciaux : Combinez les résultats des modifications attendues en termes d'utilisation et de coût aux modifications technologiques attendues ou aux programmes que vous exécutez, et développez des objectifs pour la charge de travail. Les objectifs doivent tenir compte de l'utilisation, du coût et de la relation entre les deux. Les objectifs doivent être simples, généraux et aider les personnes à comprendre les attentes de l'entreprise en termes de résultats (par exemple, s'assurer que les ressources inutilisées restent en dessous d'un certain niveau de coût). Vous n'avez pas besoin de définir des objectifs pour chaque type de ressource inutilisé ni de définir des coûts qui entraînent des pertes pour les objectifs et les cibles. Assurez-vous qu'il existe des programmes organisationnels (par exemple le renforcement des capacités avec la formation et l'éducation) si des variations de coûts sont attendues sans changement dans l'utilisation.
- Définissez les cibles : pour chacun des objectifs définis, spécifiez une cible mesurable. Si l'objectif est d'augmenter l'efficacité de la charge de travail, la cible quantifiera le degré d'amélioration (généralement en termes de résultat économique par dollar dépensé) et le moment où il sera atteint. Par exemple, si votre objectif est de réduire le gaspillage issu du surprovisionnement, alors votre cible peut être que le gaspillage dû au surprovisionnement de calcul pour le premier niveau des charges de travail de production ne doit pas dépasser 10 % des coûts de calcul du niveau, et que le gaspillage dû au surprovisionnement de calcul pour le deuxième niveau des charges de travail de production ne doit pas dépasser 5 % des coûts de calcul du niveau.

Ressources

Documents connexes :

- Politiques gérées par AWS pour les fonctions professionnelles
- Stratégie AWS multicompte pour votre zone de destination AWS Control Tower
- Contrôler l'accès aux Régions AWS avec des politiques IAM
- Objectifs SMART

Vidéos connexes :

- Ateliers Well-Architected : objectifs et cibles (niveau 100)

Exemples connexes :

- Ateliers Well-Architected : mettre hors service des ressources (objectifs et cibles)
- Ateliers Well-Architected : type de ressource, taille et quantité (objectifs et cibles)

COST02-BP03 Mettre en œuvre une structure de compte

Implémentez une structure de comptes mappée sur votre organisation. Cela vous aide à répartir et à gérer les coûts dans toute votre organisation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

AWS Organizations vous permet de créer plusieurs Comptes AWS qui peuvent vous aider à gérer votre environnement de manière centralisée lorsque vous augmentez vos charges de travail sur AWS. Vous pouvez modéliser votre hiérarchie organisationnelle en regroupant les Comptes AWS dans une structure d'unité d'organisation (UO) et en créant plusieurs Comptes AWS sous chaque UO. Pour créer une structure de compte, vous devez d'abord décider lequel de vos Comptes AWS sera le compte de gestion. Ensuite, vous pouvez créer de nouveaux Comptes AWS ou sélectionner des comptes existants en tant que comptes membres basés sur votre structure de compte désignée en suivant les bonnes pratiques du compte de gestion et les bonnes pratiques des comptes membres.

Il est recommandé de toujours lier au moins un compte membre au compte de gestion, quelle que soit la taille de votre entreprise ou l'utilisation prévue. Toutes les ressources liées aux charges de travail doivent se trouver uniquement dans les comptes membres et aucune ressource ne doit être créée dans le compte de gestion. Il n'existe pas de nombre prédéfini de Comptes AWS dont vous devez disposer. Évaluez vos modèles opérationnels et de coûts actuels et futurs pour vous assurer que la structure de vos Comptes AWS reflète les objectifs de votre organisation. Certaines entreprises créent plusieurs comptes Comptes AWS pour des raisons professionnelles, par exemple :

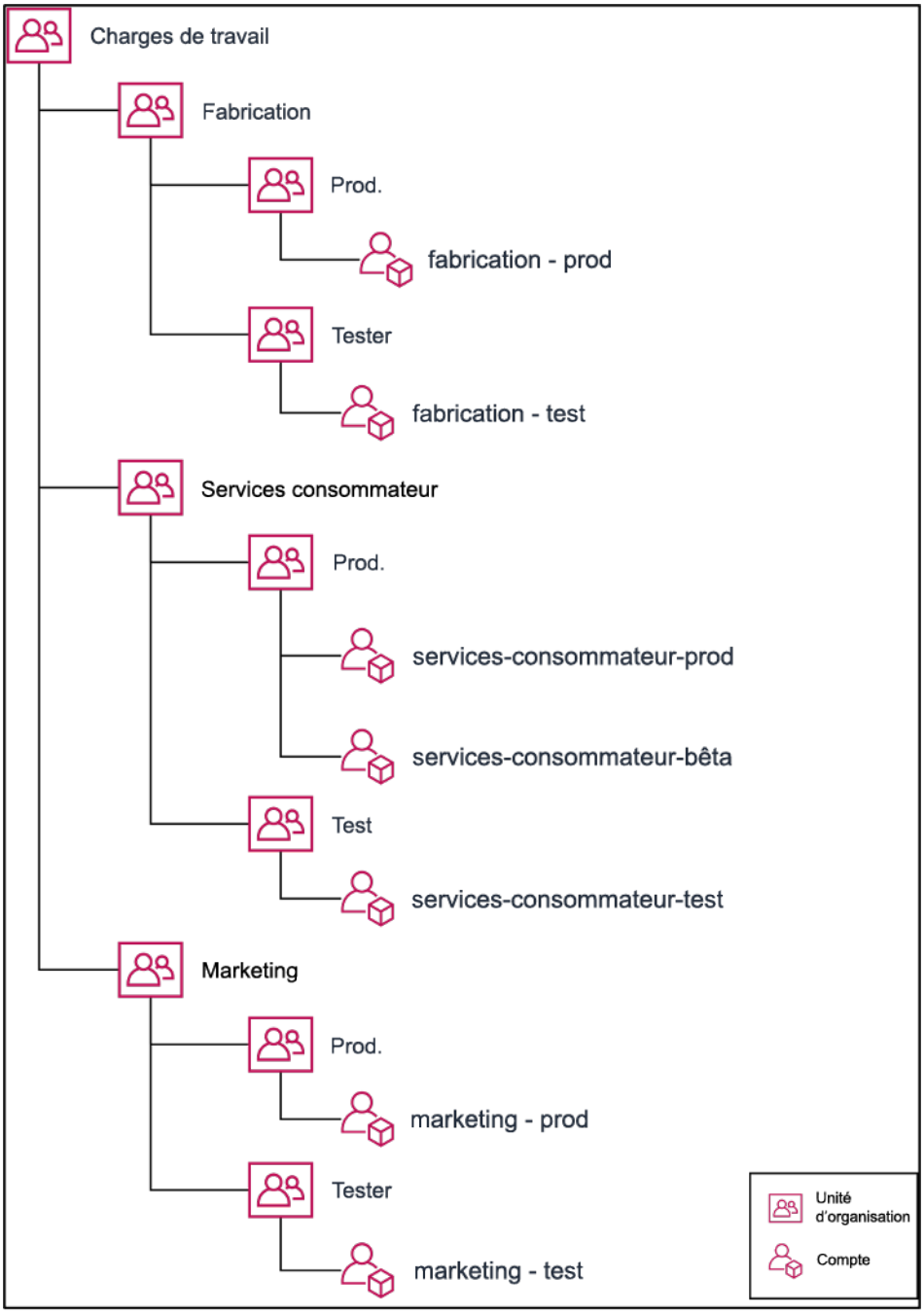
- Une isolation administrative, fiscale ou en matière de facturation est nécessaire entre les unités d'organisation, les centres de coûts ou les charges de travail spécifiques.
- Les limites du service AWS sont fixées pour être spécifiques à des charges de travail particulières.
- Il existe une exigence d'isolation et de séparation entre les charges de travail et les ressources.

Dans AWS Organizations, la facturation consolidée crée la structure entre un ou plusieurs comptes membres et le compte de gestion. Les comptes membres vous permettent d'isoler et de distinguer

votre coût et votre utilisation par groupes. Une pratique courante consiste à avoir des comptes membres séparés pour chaque unité d'organisation (comme les finances, le marketing et les ventes), ou pour chaque cycle de vie de l'environnement (comme le développement, les tests et la production), ou pour chaque charge de travail (charge de travail a, b et c), puis à regrouper ces comptes liés en utilisant la facturation consolidée.

La facturation consolidée vous permet de regrouper le paiement de plusieurs membres Comptes AWS sous un seul compte de gestion, tout en assurant la visibilité de l'activité de chaque compte lié. Comme les coûts et l'utilisation sont regroupés dans le compte de gestion, cela vous permet de maximiser vos réductions sur le volume de services et l'utilisation de vos remises sur engagement (Savings Plans et instances réservées) pour obtenir les remises les plus élevées.

Le diagramme suivant illustre l'utilisation d'AWS Organizations avec les unités d'organisation (UO) afin de regrouper plusieurs comptes et de placer de plusieurs Comptes AWS sous chaque UO. Il est recommandé d'utiliser des UO pour divers cas d'utilisation et charges de travail qui fournissent des modèles pour l'organisation des comptes.



Exemple de regroupement de plusieurs Comptes AWS sous des unités d'organisation.

AWS Control Tower peut rapidement installer et configurer plusieurs comptes AWS de façon à ce que la gouvernance soit conforme aux exigences de votre entreprise.

Étapes d'implémentation

- Définir les exigences de séparation : les exigences de séparation combinent plusieurs facteurs, notamment la sécurité, la fiabilité et les structures financières. Examinez chaque facteur dans

l'ordre et précisez si la charge de travail ou son environnement doivent être séparés des autres. La sécurité favorise le respect des exigences en matière d'accès et de données. La fiabilité gère les limites afin que les environnements et les charges de travail n'affectent pas les autres. Examinez périodiquement les piliers de sécurité et de fiabilité du cadre Well-Architected et suivez les bonnes pratiques fournies. Les structures financières créent une séparation financière stricte (pour les multiples centres de coûts, et les différentes responsabilités et propriétés liées aux charges de travail). Parmi les exemples courants de séparation, citons : les charges de travail de production et de test exécutées dans des comptes distincts ou l'utilisation d'un compte distinct afin que les données de facturation soient fournies aux unités commerciales, aux services individuels au sein de l'organisation ou à la partie prenante qui détient le compte.

- Définir les exigences de regroupement : les exigences de regroupement ne remplacent pas les exigences de séparation, mais sont utilisées pour faciliter la gestion. Regroupez les environnements ou les charges de travail similaires qui ne nécessitent pas de séparation. Par exemple, regroupez plusieurs environnements de test ou de développement d'une ou de plusieurs charges de travail.
- Définir la structure des comptes : à l'aide de ces séparations et regroupements, spécifiez un compte pour chaque groupe et gérez les exigences de séparation. Ces comptes sont vos comptes membres ou liés. En regroupant ces comptes membres au sein d'un seul compte de gestion ou compte payeur, vous rassemblez les données d'utilisation, ce qui vous permet d'obtenir des remises plus importantes sur le volume, en générant une seule facture pour tous les comptes. Il est possible de séparer les données de facturation et de créer une vue individuelle par compte membre. Si les données d'utilisation ou de facturation d'un compte membre ne doivent pas être visibles des autres comptes ou que la facturation séparée d'AWS est nécessaire, définissez plusieurs comptes de gestion ou comptes payeurs. Dans ce cas, chaque compte membre possède son propre compte de gestion ou compte payeur. Les ressources doivent toujours être placées dans des comptes membres ou comptes liés. Les comptes de gestion ou comptes payeurs doivent être uniquement utilisés pour la gestion.

Ressources

Documents connexes :

- Using Cost Allocation Tags (Utilisation des balises de répartition des coûts)
- Stratégies gérées par AWS pour les fonctions de tâches
- Stratégie de facturation multi-comptes AWS
- Contrôler l'accès aux Régions AWS avec des politiques IAM

- AWS Control Tower
- AWS Organizations
- Bonnes pratiques relatives aux comptes de gestion et aux comptes membres
- Organisation de votre environnement AWS à l'aide de plusieurs comptes
- Activation des remises de Savings Plans et sur instances réservées partagées
- Consolidated billing (Facturation consolidée)
- Consolidated billing (Facturation consolidée)

Exemples connexes :

- Splitting the CUR and Sharing Access (Fractionner le rapport d'utilisation et de coût (CUR) et partager l'accès)

Vidéos connexes :

- Introducing AWS Organizations
- Set Up a Multi-Account AWS Environment that Uses Best Practices for AWS Organizations

Exemples connexes :

- Well-Architected Labs: Create an AWS Organization (Level 100) (Ateliers Well-Architected : créer une organisation AWS (niveau 100))
- Splitting the AWS Cost and Usage Report and Sharing Access (Fractionner le rapport d'utilisation et de coût (CUR) et partager l'accès)
- Defining an AWS Multi-Account Strategy for telecommunications companies
- Best Practices for Optimizing Comptes AWS (Bonnes pratiques pour optimiser les comptes AWS)
- Best Practices for Organizational Units with AWS Organizations

COST02-BP04 Mettre en œuvre des groupes et des rôles

Mettez en œuvre des groupes et des rôles conformes à vos politiques et qui indiquent qui crée, modifie ou met hors service des instances et des ressources dans chaque groupe. Par exemple, mettez en place des groupes de développement, de test et de production. Cela s'applique aux services AWS et aux solutions tierces.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Les rôles et les groupes d'utilisateurs sont des éléments fondamentaux de la conception et de l'implémentation de systèmes sécurisés et efficaces. Les rôles et les groupes aident les organisations à trouver un équilibre entre le besoin de contrôle et le besoin de flexibilité et de productivité, et répondent ainsi aux objectifs de l'organisation et aux besoins des utilisateurs. Comme recommandé dans la section Identity and Access Management du livre blanc Pilier Sécurité - AWS Well-Architected Framework, vous devez mettre en place une gestion des identités et des autorisations solides pour garantir que les bonnes personnes ont accès aux bonnes ressources dans les bonnes conditions. Les utilisateurs disposent uniquement de l'accès dont ils ont besoin pour effectuer leurs tâches. Le risque d'accès non autorisé ou d'utilisation abusive s'en trouve ainsi réduit.

Après avoir élaboré des stratégies, vous pouvez créer des groupes logiques et des rôles d'utilisateurs au sein de votre organisation. Vous pouvez alors attribuer des autorisations, contrôler les utilisations et mettre en œuvre des mécanismes robustes de contrôle des accès pour empêcher tout accès non autorisé à des informations sensibles. Commencez par des groupes de personnes généraux. Ils correspondent généralement à des unités organisationnelles et à des fonctions (par exemple, administrateur système au sein du service informatique, contrôleur financier ou analyste commercial). Les groupes classent par catégories les personnes qui effectuent des tâches similaires et ont besoin d'un accès similaire. Les rôles définissent ce qu'un groupe doit faire. Il est plus facile de gérer les autorisations pour les groupes et les rôles que pour les utilisateurs individuels. Les rôles et les groupes attribuent des autorisations de manière systématique à tous les utilisateurs, ce qui permet d'éviter les erreurs et les incohérences.

Lorsqu'un utilisateur voit son rôle changer, les administrateurs peuvent modifier son accès au niveau du rôle ou du groupe au lieu de reconfigurer ses comptes individuels. Par exemple, un administrateur de système du service informatique a besoin d'un accès pour créer toutes les ressources, mais un membre de l'équipe d'analyse n'a besoin que de créer des ressources analytiques.

Étapes d'implémentation

- Mettre en place des groupes : utilisez les groupes d'utilisateurs définis dans les stratégies de votre organisation pour mettre en œuvre les groupes correspondants, si nécessaire. Pour connaître les bonnes pratiques en matière d'utilisateurs, de groupes et d'authentification, consultez le livre blanc Pilier Sécurité - AWS Well-Architected Framework.
- Mettre en œuvre des rôles et des stratégies : utilisez les actions définies dans les stratégies de votre organisation pour créer les rôles et stratégies d'accès requis. Pour connaître les bonnes

pratiques en matière de rôles et de stratégies, consultez le livre blanc Pilier Sécurité - AWS Well-Architected Framework.

Ressources

Documents connexes :

- Politiques gérées par AWS pour les fonctions de tâches
- Stratégie de facturation multi-comptes AWS
- Pilier Sécurité - AWS Well-Architected Framework.
- AWS Identity and Access Management (IAM)
- Politiques AWS Identity and Access Management

Vidéos connexes :

- Why use Identity and Access Management

Exemples connexes :

- Identité et accès de base de l'atelier Well-Architected
- Easier way to control access to Régions AWS using IAM policies
- Starting your Cloud Financial Management journey: Cloud cost operations

COST02-BP05 Mettre en œuvre des contrôles de coûts

Mettez en œuvre des contrôles reposant sur des politiques organisationnelles et les groupes et rôles définis. Il s'agit de s'assurer que les coûts encourus sont toujours conformes aux exigences de l'organisation, notamment en termes de contrôle d'accès aux régions ou aux types de ressources.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

En matière de contrôle des coûts, la première étape consiste le plus souvent à configurer l'envoi de notifications lorsque des événements liés aux coûts ou à l'utilisation sortant du cadre des politiques en vigueur se produisent. Vous pouvez agir rapidement et vérifier si une action corrective est nécessaire, sans restreindre ni affecter négativement les charges de travail ou la nouvelle activité.

Une fois que vous connaissez les limites de la charge de travail et de l'environnement, vous pouvez appliquer la gouvernance. AWS Budgets vous permet d'établir des notifications et de définir des budgets mensuels pour vos coûts, votre utilisation et vos remises sur engagement (Savings Plans et instances réservées) AWS. Vous pouvez créer des budgets à un niveau de coût global (intégralité des coûts) ou plus précis, si vous n'incluez que les dimensions spécifiques pertinentes, telles que les comptes, services, balises ou zones de disponibilité.

Une fois que vous avez fixé vos limites budgétaires avec AWS Budgets, utilisez AWS Cost Anomaly Detection pour réduire vos coûts imprévus. AWS Cost Anomaly Detection est un service de gestion des coûts qui utilise le machine learning pour surveiller continuellement vos coûts et votre utilisation afin de détecter les dépenses inhabituelles. Il vous permet d'identifier les dépenses anormales et les causes profondes, afin que vous puissiez prendre des mesures rapidement. Créez tout d'abord une surveillance des coûts dans AWS Cost Anomaly Detection, puis choisissez votre préférence d'alerte en établissant un seuil monétaire (par exemple une alerte sur les anomalies ayant un impact supérieur à 1 000 USD). Une fois les alertes reçues, vous pouvez analyser la cause profonde de l'anomalie et son impact sur vos coûts. Vous pouvez également surveiller et analyser les anomalies dans AWS Cost Explorer.

Appliquez les politiques de gouvernance dans AWS via AWS Identity and Access Management et les Politiques de contrôle de service (SCP) AWS Organizations. IAM vous permet de gérer en toute sécurité l'accès aux services et aux ressources AWS. Avec IAM, vous pouvez contrôler qui peut créer et gérer les ressources AWS ainsi que le type de ressources disponibles et leurs emplacements respectifs. Cela réduit au minimum les risques que des ressources soient créées en dehors du cadre de la politique définie. Utilisez les rôles et groupes créés précédemment et attribuez des politiques IAM pour garantir une utilisation correcte. Une politique SCP offre un contrôle central des autorisations maximales disponibles pour tous les comptes de votre organisation, afin que ces derniers respectent à tout moment vos directives de contrôle d'accès. Les politiques SCP ne sont disponibles que dans les organisations où toutes les fonctionnalités sont activées. Vous pouvez configurer les politiques SCP afin qu'elles refusent ou autorisent par défaut les actions des comptes membres. Pour plus de détails sur la mise en œuvre de la gestion des accès, consultez le livre blanc Well-Architected - Pilier Sécurité.

La gouvernance peut également être mise en œuvre grâce à la gestion deAWS quotas de service. En vous assurant que les quotas de service sont fixés avec un coût minimum et gérés avec précision, vous pouvez minimiser la création de ressources en dehors du cadre des exigences de votre organisation. Pour ce faire, vous devez comprendre à quel point vos exigences peuvent rapidement changer, appréhender les projets en cours (tant la création que la mise hors service des ressources),

et tenir compte de l'accélération des délais de mise en œuvre de ces quotas. Service Quotas est un service qui permet d'augmenter vos quotas en fonction de vos besoins.

Étapes d'implémentation

- Mettre en œuvre des notifications sur les dépenses : à l'aide des politiques organisationnelles que vous avez définies, créez des budgets AWS Budgets pour envoyer des notifications lorsque les dépenses dépassent les seuils fixés. Configurez plusieurs budgets de coûts, un pour chaque compte, afin d'être averti des dépenses globales du compte. Configurez des budgets de coûts supplémentaires dans chaque compte pour les plus petites unités du compte. Ces unités varient en fonction de la structure de votre compte. Quelques exemples courants sont les Régions AWS, les charges de travail (avec les balises) ou les services AWS. Configurez une liste de distribution comme destinataire des notifications au lieu d'utiliser un e-mail individuel. Vous pouvez définir un budget réel en cas de dépassement du montant ou utiliser un budget prévisionnel pour notifier de l'utilisation prévue. Vous pouvez également préconfigurer des actions AWS Budget pour appliquer des politiques IAM ou SCP spécifiques, voire arrêter les instances cibles Amazon EC2 ou Amazon RDS. Les actions Budget peuvent être automatiquement exécutées ou nécessiter l'approbation du flux de travail.
- Mettre en œuvre des notifications sur les dépenses anormales : utilisez AWS Cost Anomaly Detection pour réduire les coûts imprévus dans votre organisation et analyser la cause profonde des éventuelles dépenses anormales. Une fois que vous avez créé le suivi des coûts pour identifier les dépenses inhabituelles à la granularité spécifiée et configuré les notifications dans AWS Cost Anomaly Detection, vous recevez une alerte lorsque des dépenses inhabituelles sont détectées. Cela vous permettra d'analyser le cas racine de l'anomalie et de comprendre l'impact sur votre coût. Utilisez les catégories de coûts AWS en configurant AWS Cost Anomaly Detection afin de déterminer quelle équipe de projet ou d'unité commerciale peut analyser la cause profonde des coûts imprévus et prendre rapidement les mesures nécessaires.
- Mettre en œuvre des contrôles sur l'utilisation : à l'aide des politiques organisationnelles que vous avez définies, mettez en œuvre des politiques et des rôles IAM pour spécifier les actions que les utilisateurs peuvent et ne peuvent pas effectuer. Plusieurs politiques organisationnelles peuvent être incluses dans une politique AWS. De la même manière que vous avez défini les politiques, commencez par une approche générale et appliquez ensuite des contrôles plus fins à chaque étape. Les limites de service constituent également un contrôle efficace de l'utilisation. Mettez en œuvre les limites de service correctes sur tous vos comptes.

Ressources

Documents connexes :

- Politiques gérées AWS pour les fonctions professionnelles
- Stratégie de facturation multi-comptes AWS
- Contrôler l'accès aux Régions AWS avec des politiques IAM
- AWS Budgets
- AWS Cost Anomaly Detection
- Maîtriser vos coûts AWS

Vidéos connexes :

- How can I use AWS Budgets to track my spending and usage?

Exemples connexes :

- Example IAM access management policies (Exemples de politiques de gestion des accès IAM)
- Example service control policies (Exemples de politiques de contrôle de service)
- Actions AWS Budgets
- Comment créer une politique IAM pour contrôler l'accès aux ressources Amazon EC2 à l'aide de balises ?
- Est-il possible de limiter l'accès d'une identité IAM à des ressources Amazon EC2 spécifiques ?
- Create an IAM Policy to restrict Amazon EC2 usage by family (Créer une politique IAM pour limiter l'utilisation d'EC2 par famille)
- Well-Architected Labs: Cost and Usage Governance (Level 100) (Ateliers Well-Architected : gouvernance des coûts et de l'utilisation (niveau 100))
- Well-Architected Labs: Cost and Usage Governance (Level 200) (Ateliers Well-Architected : gouvernance des coûts et de l'utilisation (niveau 200))
- Slack integrations for Cost Anomaly Detection using AWS Chatbot (Intégrations de Slack pour Cost Anomaly Detection avec AWS Chatbot)

COST02-BP06 Suivre le cycle de vie du projet

Suivez, mesurez et auditez le cycle de vie des projets, des équipes et des environnements pour éviter d'utiliser et de payer des ressources superflues.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Un suivi efficace du cycle de vie des projets permet aux organisations de mieux contrôler leurs coûts grâce à une planification, une gestion et une optimisation améliorées des ressources, du temps et de la qualité. Les informations obtenues dans le cadre du suivi sont précieuses pour prendre des décisions éclairées qui contribuent à la rentabilité et à la réussite globale du projet.

Le suivi du cycle de vie complet de la charge de travail vous aide à comprendre quand les charges de travail ou leurs composants ne sont plus nécessaires. Les charges de travail et les composants existants peuvent sembler utilisés, mais ils peuvent être mis hors service ou adoptés lorsqu'AWS publie de nouveaux services ou fonctionnalités. Consultez les précédentes étapes des charges de travail. Une fois qu'une charge de travail est en production, les environnements précédents peuvent être mis hors service ou leur capacité fortement réduite jusqu'à ce qu'ils soient de nouveau requis.

AWS fournit des services de gestion et de gouvernance que vous pouvez utiliser pour le suivi du cycle de vie des entités. Vous pouvez utiliser AWS Config ou AWS Systems Manager pour fournir un inventaire détaillé de vos ressources et de votre configuration AWS. Il est recommandé de l'intégrer à vos systèmes de gestion de projets ou ressources existants pour assurer le suivi des projets et produits actifs au sein de votre organisation. La combinaison de votre système actuel avec le riche ensemble d'événements et de mesures fourni par AWS vous permet de construire une vue des événements importants du cycle de vie et de gérer les ressources de manière proactive afin de réduire les coûts inutiles.

À l'instar de la gestion du cycle de vie des applications (ALM), le suivi du cycle de vie des projets doit impliquer plusieurs processus, outils et équipes (conception et développement, tests, production, support, redondance des charges de travail, etc.).

En surveillant attentivement chaque phase du cycle de vie d'un projet, les organisations obtiennent des informations cruciales et un meilleur contrôle, ce qui facilite la planification, la mise en œuvre et la réalisation des projets. Cette surveillance attentive permet de vérifier que les projets répondent non seulement aux normes de qualité, mais également qu'ils sont livrés dans les délais et dans les limites du budget, ce qui favorise la rentabilité globale.

Pour plus d'informations sur la mise en œuvre du suivi du cycle de vie des entités, consultez le livre blanc Pilier Excellence opérationnelle - AWS Well-Architected Framework.

Étapes d'implémentation

- Établir un processus de suivi du cycle de vie des projets : l'équipe du Centre d'excellence cloud doit créer un processus de suivi du cycle de vie des projets. Établissez une approche structurée et systématique pour surveiller les charges de travail afin d'améliorer le contrôle, la visibilité et les performances des projets. Rendez le processus de suivi transparent, collaboratif et axé sur l'amélioration continue afin d'en maximiser l'efficacité et la valeur.
- Vérifier la charge de travail : conformément aux politiques de votre organisation, définissez une fréquence régulière pour auditer vos projets existants et vérifier la charge de travail. Le niveau d'effort consacré à l'audit doit être proportionnel au risque, à la valeur ou au coût approximatif pour l'organisation. Les principaux domaines à inclure dans l'audit sont le risque pour l'organisation d'un incident ou d'une panne, la valeur ou la contribution à l'organisation (mesurée en termes de chiffre d'affaires ou de réputation de la marque), le coût de la charge de travail (mesuré en tant que coût total des ressources et coûts opérationnels) et l'utilisation de la charge de travail (mesurée en nombre de résultats de l'organisation par unité de temps). Si ces domaines changent au cours du cycle de vie, des ajustements de la charge de travail sont nécessaires, tels que la mise hors service complète ou partielle.

Ressources

Documents connexes :

- Guidance for Tagging on AWS
- Qu'est-ce que l'ALM (gestion du cycle de vie de l'application) ?
- Politiques gérées AWS pour les fonctions professionnelles

Exemples connexes :

- Contrôler l'accès aux Régions AWS avec des politiques IAM

Outils associés :

- AWS Config
- AWS Systems Manager

- AWS Budgets
- AWS Organizations
- AWS CloudFormation

COÛT 3. Comment surveillez-vous vos coûts et votre utilisation ?

Définissez des stratégies et des procédures pour surveiller et allouer vos coûts de manière appropriée. Cela vous permet d'évaluer et d'améliorer la rentabilité de cette charge de travail.

Bonnes pratiques

- COST03-BP01 Configurer des sources d'informations détaillées
- COST03-BP02 Ajouter des informations organisationnelles aux coûts et à l'utilisation
- COST03-BP03 Identifier les catégories de répartition des coûts
- COST03-BP04 Établir des métriques organisationnelles
- COST03-BP05 Configurer des outils de facturation et de gestion des coûts
- COST03-BP06 Répartir les coûts selon les métriques de la charge de travail

COST03-BP01 Configurer des sources d'informations détaillées

Configurez les outils de gestion des coûts et de reporting pour une granularité horaire afin de fournir des informations détaillées sur les coûts et l'utilisation, ce qui permet de renforcer l'analyse et la transparence. Configurez votre charge de travail de manière à générer ou à disposer des entrées de journal pour chaque résultat opérationnel fourni.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Des informations de facturation détaillées telles que la granularité horaire dans les outils de gestion des coûts permettent aux organisations de suivre leurs consommations plus en détail et les aident à identifier certaines des raisons de l'augmentation des coûts. Ces sources de données offrent la vue la plus précise des coûts et de l'utilisation dans l'ensemble de votre organisation.

AWS Cost and Usage Report fournit une granularité d'utilisation quotidienne ou horaire, des tarifs, des coûts et des attributs d'utilisation pour tous les services AWS payants. Toutes les dimensions

possibles sont dans le rapport CUR, y compris le balisage, l'emplacement, les attributs des ressources et les ID de compte.

Configurez votre CUR avec les personnalisations suivantes :

- Inclure les ID de ressources
- Actualiser automatiquement le CUR
- Granularité horaire
- Gestion des versions : Remplacer le rapport existant
- Intégration des données : Athena (format et compression Parquet)

Utilisez AWS Glue pour préparer les données pour l'analyse et Amazon Athena pour effectuer l'analyse des données, à l'aide de SQL pour interroger les données. Vous pouvez également utiliser Amazon QuickSight pour créer des visualisations personnalisées et complexes et les distribuer dans l'ensemble de votre entreprise.

Étapes d'implémentation

- Configurer le rapport de coût et d'utilisation : Dans la console de facturation, configurez au moins un rapport d'utilisation et de coût. Configurez un rapport avec une granularité horaire incluant tous les identifiants et les ID de ressource. Vous pouvez également créer d'autres rapports avec différentes granularités pour fournir des informations récapitulatives générales.
- Configurer la granularité horaire dans Cost Explorer : Activez Horaire et Données au niveau des ressources pour accéder aux données de coût et d'utilisation à une granularité horaire pour les 14 derniers jours et à une granularité au niveau des ressources.
- Configurer la journalisation de l'application : Vérifiez que votre application consigne chaque résultat opérationnel qu'elle produit afin de le suivre et le mesurer. Veillez à ce que la granularité de ces données est au moins horaire pour être mise en correspondance avec les données de coût et d'utilisation. Pour plus d'informations sur la journalisation et la surveillance, voir au pilier Excellence opérationnelle Well-Architected. »

Ressources

Documents connexes :

- AWS Cost and Usage Report

- AWS Glue
- Amazon QuickSight
- Tarification de la gestion des coûts AWS
- Balisage des ressources AWS
- Analyse des coûts avec AWS Budgets
- Analyse des coûts avec Cost Explorer
- Gestion des AWS Cost and Usage Report
- au pilier Excellence opérationnelle Well-Architected

Exemples connexes :

- Configuration de compte AWS
- Cas d'utilisation courants et nouvel aspect de AWS Cost Explorer

COST03-BP02 Ajouter des informations organisationnelles aux coûts et à l'utilisation

Définissez un schéma de balisage en fonction de votre organisation, des attributs de la charge de travail et des catégories de répartition des coûts afin de pouvoir filtrer et rechercher des ressources ou surveiller les coûts et l'utilisation dans les outils de gestion des coûts. Mettez en œuvre le balisage cohérent sur toutes les ressources, dans la mesure du possible, par objectif, équipe, environnement ou tout autre critère pertinent pour votre entreprise.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Mettez en œuvre le balisage dans AWS pour ajouter des informations organisationnelles à vos ressources, qui seront ensuite ajoutées à vos informations de coût et d'utilisation. Une balise est une paire clé-valeur. La clé est définie et doit être unique dans votre organisation, et la valeur est unique à un groupe de ressources. Voici un exemple de paire clé-valeur : la clé est `Environment` (Environnement), avec une valeur `Production` (Production). Toutes les ressources de l'environnement de production auront cette paire clé-valeur. Le balisage permet de catégoriser et de suivre vos coûts à l'aide d'informations significatives pertinentes sur l'organisation. Vous pouvez appliquer des balises qui représentent des catégories d'organisations (telles que les centres de coûts, les noms d'application, les projets ou les propriétaires), et identifier les charges de travail et les

caractéristiques des charges de travail (telles que les tests ou la production) pour attribuer vos coûts et votre utilisation dans toute votre organisation.

Lorsque vous appliquez des balises à vos ressources AWS (telles que les instances Amazon Elastic Compute Cloud ou les compartiments Amazon Simple Storage Service) et que vous les activez, AWS ajoute ces informations à vos rapports de coûts et d'utilisation. Vous pouvez exécuter des rapports et effectuer des analyses sur les ressources balisées et non balisées pour permettre une meilleure conformité avec les politiques internes de gestion des coûts et assurer une attribution précise.

La création et la mise en œuvre d'une norme de balisage AWS dans les comptes de votre entreprise permettent de gérer et de gouverner vos environnements AWS de manière cohérente et uniforme. Utilisez Tag Policies (Politiques de balise) dans AWS Organizations pour définir les règles d'utilisation des balises au niveau des ressources AWS de vos comptes dans AWS Organizations. Les politiques de balises permettent d'adopter facilement une approche normalisée pour le balisage des ressources AWS.

AWS Tag Editor permet d'ajouter, de supprimer et de gérer des balises de plusieurs ressources. Avec Tag Editor, vous recherchez les ressources que vous voulez baliser et vous gérez ensuite les balises pour les ressources dans vos résultats de recherche.

Les catégories de coûts AWS permettent d'attribuer une signification organisationnelle à vos coûts, sans nécessiter de balises sur les ressources. Vous pouvez associer vos informations de coût et d'utilisation à des structures d'organisation internes uniques. Vous définissez des règles de catégorie pour associer et catégoriser les coûts à l'aide des dimensions de facturation, telles que les comptes et les balises. Cela fournit un autre niveau de fonctionnalité de gestion en plus du balisage. Vous pouvez également associer des comptes et des balises spécifiques à plusieurs projets.

Étapes d'implémentation

- Définir un schéma de balisage : réunissez toutes les parties prenantes de votre entreprise pour définir un schéma. Il s'agit généralement de membres du personnel technique, de l'équipe financière et de la direction. Définissez une liste de balises que toutes les ressources doivent avoir, ainsi qu'une liste de balises que des ressources doivent avoir. Veillez à ce que les noms et les valeurs des balises soient cohérents dans l'ensemble de votre organisation.
- Baliser des ressources : en utilisant vos catégories de répartition des coûts définies, placez des balises sur toutes les ressources de vos charges de travail en fonction des catégories. Utilisez des outils tels que l'interface de ligne de commande (CLI), Tag Editor ou AWS Systems Manager pour améliorer l'efficacité.

- Mettre en œuvre les catégories de coûts AWS : vous pouvez créer des catégories de coûts sans mettre en œuvre le balisage. Les catégories de coûts utilisent les dimensions de coûts et d'utilisation existantes. Créez des règles de catégorie à partir de votre schéma et mettez-les en œuvre dans les catégories de coûts.
- Automatiser le balisage : pour veiller à maintenir des niveaux élevés de balisage sur toutes les ressources, automatisez le balisage afin que les ressources soient automatiquement balisées lorsqu'elles sont créées. Utilisez des services tels que AWS CloudFormation pour vérifier que les ressources sont balisées lors de leur création. Vous pouvez également créer une solution personnalisée pour baliser automatiquement grâce aux fonctions Lambda ou utiliser un microservice qui analyse régulièrement la charge de travail et supprime toutes les ressources qui ne sont pas balisées, ce qui est pratique pour les environnements de test et de développement.
- Surveiller et créer des rapports sur le balisage : pour veiller à maintenir des niveaux élevés de balisage dans votre organisation, surveillez les balises de vos charges de travail et créez des rapports associés. Vous pouvez utiliser AWS Cost Explorer pour afficher le coût des ressources balisées et non balisées, ou recourir à des services tels que Tag Editor. Examinez régulièrement le nombre de ressources non balisées et prenez les mesures nécessaires pour ajouter des balises jusqu'à ce que vous atteigniez le niveau de balisage souhaité.

Ressources

Documents connexes :

- Bonnes pratiques de balisage
- Balise d'une ressource AWS CloudFormation
- Catégories de coûts AWS
- Balisage des ressources AWS
- Analyse des coûts avec AWS Budgets
- Analyse des coûts avec Cost Explorer
- Gestion des rapports d'utilisation et de coûts AWS

Vidéos connexes :

- Comment puis-je baliser mes ressources AWS pour répartir ma facture par centre de coûts ou par projet ?
- Balisage des ressources AWS

Exemples connexes :

- Automatiquement baliser des nouvelles ressources AWS en fonction de l'identité ou du rôle

COST03-BP03 Identifier les catégories de répartition des coûts

Identifiez les catégories d'organisation telles que les unités commerciales, les services ou les projets qui pourraient être utilisés pour répartir les coûts au sein de votre organisation entre les entités consommatrices internes. Utilisez ces catégories pour renforcer la responsabilité en matière de dépenses, sensibiliser aux coûts et encourager des comportements de consommation efficaces.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Le processus de catégorisation des coûts est crucial pour la budgétisation, la comptabilité, les rapports financiers, la prise de décision, l'analyse comparative et la gestion de projet. En classant et en catégorisant les dépenses, les équipes peuvent mieux comprendre les types de coûts qu'elles doivent supporter tout au long de leur transition vers le cloud, ce qui les aide à prendre des décisions éclairées et à gérer les budgets de manière efficace.

La responsabilité des dépenses liées au cloud incite fortement à une gestion disciplinée de la demande et des coûts. Il en résulte des économies importantes sur les coûts liés au cloud pour les organisations qui allouent la majeure partie de leurs dépenses en matière de cloud à des unités commerciales ou à des équipes consommatrices. En outre, l'affectation des dépenses liées au cloud aide les organisations à adopter davantage de bonnes pratiques en matière de gouvernance centralisée du cloud.

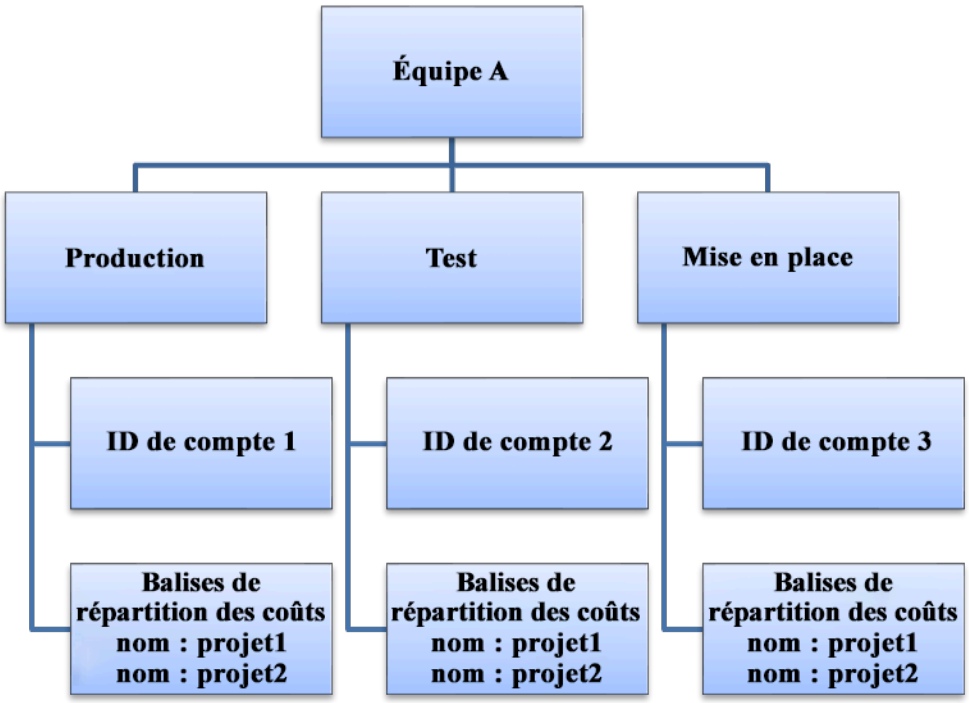
Travaillez avec votre équipe financière et les autres parties prenantes concernées pour comprendre les exigences relatives à la répartition des coûts au sein de votre entreprise lors de vos appels périodiques. Les coûts de la charge de travail doivent être répartis sur l'ensemble du cycle de vie, y compris le développement, les tests, la production et la mise hors service. Vous devez comprendre comment les coûts engagés pour l'apprentissage, le développement du personnel et la création d'idées sont attribués dans l'organisation. Cela peut être utile pour affecter correctement les comptes utilisés à cette fin aux budgets de formation et de développement, au lieu des budgets génériques de coûts informatiques.

Après avoir défini vos catégories d'attribution des coûts avec les parties prenantes de votre organisation, utilisez Catégories de coûts AWS pour regrouper vos informations de coût et

d'utilisation dans des catégories pertinentes dans AWS Cloud, par exemple les coûts relatifs à un projet spécifique ou à des Comptes AWS pour des départements ou des unités commerciales. Vous pouvez créer des catégories personnalisées et mapper vos informations de coût et d'utilisation dans ces catégories en fonction des règles que vous définissez grâce à différentes dimensions, telles que le compte, la balise, le service ou le type de frais. Une fois les catégories de coûts définies, vous pouvez afficher vos informations de coût et d'utilisation pour chacune d'entre elles pour permettre à votre organisation de prendre de meilleures décisions stratégiques et d'achat. Ces catégories sont également visibles dans AWS Cost Explorer, AWS Budgets et AWS Cost and Usage Report.

Par exemple, créez des catégories de coûts pour vos unités commerciales (équipe DevOps) et, sous chaque catégorie, créez plusieurs règles (des règles pour chaque sous-catégorie) avec plusieurs dimensions (Comptes AWS, balises de répartition des coûts, services ou type de frais) selon les regroupements que vous avez définis. Avec les catégories de coûts, vous pouvez organiser vos coûts grâce à un moteur basé sur des règles. Les règles que vous configurez organisent vos coûts en catégories. Vous pouvez filtrer ces règles sous plusieurs dimensions pour chaque catégorie, telles que des Comptes AWS, des services AWS ou des types de frais spécifiques. Vous pouvez ensuite utiliser ces catégories dans la console AWS Billing and Cost Management et gestion des coûts sur plusieurs produits. Cela inclut AWS Cost Explorer, AWS Budgets, AWS Cost and Usage Report et AWS Cost Anomaly Detection.

À titre d'exemple, le diagramme suivant vous montre comment regrouper vos coûts et vos informations d'utilisation dans votre organisation en ayant plusieurs équipes (catégorie de coûts), plusieurs environnements (règles), et chaque environnement ayant plusieurs ressources ou actifs (dimensions).



Graphique des coûts et de l'utilisation de l'organisation

Vous pouvez également regrouper les coûts avec les catégories de coûts. Une fois que vous avez créé les catégories de coûts (jusqu'à 24 heures après la création d'une catégorie de coût peuvent être nécessaires pour que les valeurs soient mises à jour dans vos relevés d'utilisation), elles apparaissent dans AWS Cost Explorer, AWS Budgets, AWS Cost and Usage Reportet AWS Cost Anomaly Detection. » Dans AWS Cost Explorer et AWS Budgets, une catégorie de coût apparaît dans une dimension de facturation supplémentaire. Vous pouvez l'utiliser pour filtrer la valeur de catégorie de coût spécifique ou effectuer un regroupement sous la catégorie de coût.

Étapes d'implémentation

- Définir les catégories de votre organisation : Rencontrez les parties prenantes internes et les unités commerciales pour définir les catégories qui reflètent la structure et les besoins de votre organisation. Ces catégories devraient correspondre directement à la structure des catégories financières existantes, telles que l'unité commerciale, le budget, le centre de coûts ou le service. Examinez les résultats que le cloud apporte à votre entreprise, tels que la formation ou l'éducation, car il s'agit également de catégories organisationnelles.
- Définir vos catégories fonctionnelles : Rencontrez les parties prenantes internes et les unités commerciales pour définir des catégories qui reflètent les fonctions de votre entreprise. Il peut

s'agir de la charge de travail ou des noms d'application, ainsi que du type d'environnement, comme la production, les tests ou le développement.

- Définissez les Catégories de coûts AWS : Créez des catégories de coûts pour organiser vos informations de coût et d'utilisation grâce à Catégories de coûts AWS et cartographiez vos coûts et votre utilisation d'AWS dans des catégories significatives. » Plusieurs catégories peuvent être attribuées à une ressource, et une ressource peut se trouver dans plusieurs catégories. Par conséquent, définissez autant de catégories que nécessaire afin de pouvoir gérer vos coûts dans la structure catégorisée à l'aide des Catégories de coûts AWS.

Ressources

Documents connexes :

- Balisage des ressources AWS
- Utilisation des balises de répartition des coûts
- Analyse des coûts avec AWS Budgets
- Analyse des coûts avec Cost Explorer
- Gestion des AWS Cost and Usage Report
- Catégories de coûts AWS
- Gestion de vos coûts grâce aux Catégories de coûts AWS
- Création de catégories de coûts
- Balisage des catégories de coûts
- Division des frais dans les catégories de coûts
- Fonctionnalités des Catégories de coûts AWS

Exemples connexes :

- Organisation de vos données de coûts et d'utilisation à l'aide des Catégories de coûts AWS
- Gestion de vos coûts grâce aux Catégories de coûts AWS
- Ateliers Well-Architected : visualisation des coûts et de l'utilisation
- Ateliers Well-Architected : catégories de coûts

COST03-BP04 Établir des métriques organisationnelles

Établissez les métriques de l'organisation qui sont requises pour cette charge de travail. Les rapports des clients produits ou les pages Web diffusées aux clients sont des exemples de métriques d'une charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Comprenez comment le rendement de votre charge de travail est mesuré par rapport à la réussite métier. Chaque charge de travail comporte généralement un petit ensemble de résultats majeurs qui indiquent les performances. Si votre charge de travail est complexe et comporte de nombreux éléments, vous pouvez en dresser la liste par ordre de priorité ou définir et suivre les paramètres de chaque élément. Travaillez avec vos équipes pour savoir quelles métriques vous devez utiliser. Cette unité sera utilisée pour comprendre l'efficacité de la charge de travail, ou le coût de chaque production commerciale.

Étapes d'implémentation

- Définir les résultats de la charge de travail : rencontrez les parties prenantes de l'entreprise et définissez les résultats de la charge de travail. Il s'agit des mesures principales de l'utilisation des clients, qui doivent être des métriques économiques et non pas techniques. Il doit exister un petit nombre de métriques générales (moins de cinq) par charge de travail. Si la charge de travail produit plusieurs résultats pour différents cas d'utilisation, regroupez-les dans une seule métrique.
- Définir les résultats des composants de la charge de travail : le cas échéant, si la charge de travail est volumineuse et complexe ou que vous pouvez facilement la diviser en composants (tels que des microservices) avec des entrées et des sorties bien définies, définissez des métriques pour chaque composant. L'effort doit refléter la valeur et le coût du composant. Procédez du plus grand au plus petit composant.

Ressources

Documents connexes :

- Balisage des ressources AWS
- Analyse des coûts avec AWS Budgets
- Analyse des coûts avec Cost Explorer
- Gestion des rapports d'utilisation et de coûts AWS

COST03-BP05 Configurer des outils de facturation et de gestion des coûts

Configurez les outils de gestion des coûts conformément aux politiques de votre organisation en matière de gestion et d'optimisation des dépenses du cloud. Ils incluent les services, les outils et les ressources pour organiser et suivre les données de coûts et d'utilisation, avoir plus de contrôle grâce à la facturation consolidée et les autorisations d'accès, améliorer la planification via des budgets et des prévisions, recevoir des notifications ou des alertes, et réduire davantage les coûts grâce aux optimisations des ressources et de la tarification.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Pour établir une solide responsabilisation, vous devez d'abord considérer la stratégie de votre compte comme faisant partie de votre stratégie de répartition des coûts. Faites les choses correctement et vous n'aurez peut-être pas besoin d'aller plus loin. Sinon, certains points risquent d'être omis et d'autres problèmes pourraient survenir par la suite.

Pour encourager la responsabilisation en matière de dépenses liées au cloud, les utilisateurs doivent avoir accès à des outils offrant une visibilité sur leurs coûts et leur utilisation. Il est recommandé de configurer des outils pour toutes les charges de travail et pour toutes les équipes pour les objectifs et détails suivants :

- Organiser : établissez votre répartition des coûts et votre base de référence de la gouvernance avec votre propre stratégie de balisage et vos catégorisations.
- Organiser : établissez votre répartition des coûts et votre base de référence de la gouvernance avec votre propre stratégie de balisage et votre taxonomie. Balisez les ressources AWS prises en charge et classez-les de manière significative en fonction de la structure de votre organisation (unités commerciales, départements ou projets). Balisez les noms de comptes pour des centres de coûts spécifiques et associez-les à des Catégories de coûts AWS afin de regrouper les comptes d'unités commerciales spécifiques pour leurs centres de coûts et permettre au propriétaire de l'unité commerciale de voir la consommation de plusieurs comptes dans un seul endroit.
- Accéder : suivez les informations de facturation à l'échelle de l'organisation dans la facturation consolidée et vérifiez que les bonnes parties prenantes et les propriétaires d'entreprise y ont accès.
- Contrôler : créez des mécanismes de gouvernance efficaces avec les bonnes barrières de protection pour éviter les scénarios inattendus lors de l'utilisation des politiques de contrôle des services (SCP), des politiques de balisage et des alertes budgétaires. Par exemple, vous pouvez autoriser les équipes à créer des ressources dans des régions privilégiées uniquement en utilisant des mécanismes de contrôle efficaces.

- État actuel : configurez un tableau de bord affichant les niveaux actuels de coûts et d'utilisation. Le tableau de bord doit être disponible dans un endroit hautement visible dans l'environnement de travail (similaire à un tableau de bord d'opérations). Vous pouvez utiliser le Cloud Intelligence Dashboard (CID) ou tout autre produit pris en charge pour créer cette visibilité.
- Notifications : fournissez des notifications lorsque le coût ou l'utilisation dépasse les limites définies et lorsque des anomalies surviennent avec AWS Budgets ou AWS Cost Anomaly Detection.
- Rapports : récapitulez toutes les informations de coût et d'utilisation, et sensibilisez et responsabilisez quant à vos dépenses cloud avec des données de coûts détaillées et attribuables. Les rapports doivent être pertinents pour l'équipe qui les utilise et, idéalement, contenir des recommandations.
- Suivi : affichez les coûts et l'utilisation actuels par rapport aux objectifs ou cibles configurés.
- Analyse : permettez aux membres de l'équipe d'effectuer des analyses personnalisées et approfondies jusqu'à la précision horaire, avec toutes les dimensions possibles.
- Inspecter : informez-vous sur vos opportunités de déploiement des ressources et d'optimisation des coûts. Recevez des notifications (en utilisant Amazon CloudWatch, Amazon SNS ou Amazon SES) pour les déploiements de ressources au niveau de l'organisation et consultez les recommandations d'optimisation des coûts (par exemple, AWS Compute Optimizer ou AWS Trusted Advisor).
- Tendance : affichez la variabilité des coûts et de l'utilisation sur la période requise avec la granularité nécessaire.
- Prévisions : affichez les coûts futurs prévus, estimez votre utilisation des ressources et dépensez en fonction des tableaux de bord des prévisions que vous créez.

Vous pouvez utiliser des outils AWS tels qu' AWS Cost Explorer, AWS Billing and Cost Managementou AWS Budgets pour les éléments essentiels ou vous pouvez intégrer les données CUR à Amazon Athena et Amazon QuickSight afin de fournir cette fonctionnalité pour des vues plus détaillées. Si vous ne disposez pas des compétences ni de la bande passante nécessaires au sein de votre organisation, vous pouvez utiliser AWS ProServ, AWS Managed Services (AMS)ou les AWS Partner et vous servir de leurs outils. Vous pouvez également utiliser des outils tiers, mais vérifiez d'abord que le coût offre de la valeur à votre organisation.

Étapes d'implémentation

- Autorisez l'accès des équipes aux outils : configurez vos comptes et créez des groupes qui ont accès aux rapports de coûts et d'utilisation requis pour leurs consommations et utilisez AWS Identity and Access Management pour contrôler l'accès aux outils tels qu'AWS Cost Explorer.

Ces groupes doivent inclure des représentants de toutes les équipes qui possèdent ou gèrent une application. Chaque équipe a ainsi accès à ses informations de coût et d'utilisation pour suivre sa consommation.

- Configurez AWS Budgets : configurez AWS Budgets sur tous les comptes de vos charges de travail. Définissez des budgets pour les dépenses globales des comptes et des budgets pour les charges de travail à l'aide de balises. Configurez des notifications dans AWS Budgets pour recevoir des alertes quand vous dépassez vos montants budgétisés ou quand vos coûts estimés dépassent vos budgets.
- Configurez AWS Cost Explorer : configurez AWS Cost Explorer pour votre charge de travail et vos comptes afin de visualiser vos données de coût pour une analyse plus approfondie. Créez un tableau de bord pour la charge de travail qui suit les dépenses globales et les principales métriques d'utilisation de la charge de travail, et qui prévoit les futurs coûts en fonction de vos anciennes données de coût.
- Configurez AWS Cost Anomaly Detection : utilisez AWS Cost Anomaly Detection pour vos comptes, vos services principaux ou les catégories de coûts que vous avez créées pour surveiller vos coûts et votre utilisation, et détecter les dépenses inhabituelles. Vous pouvez recevoir des alertes individuelles dans les rapports agrégés, et recevoir des alertes dans un e-mail ou une rubrique Amazon SNS qui vous permet d'analyser et de déterminer la cause profonde de l'anomalie et d'identifier le facteur responsable de l'augmentation des coûts.
- Configurez des outils avancés : si vous voulez, vous pouvez créer des outils personnalisés pour votre organisation, qui fournissent des informations et une granularité supplémentaires. Vous pouvez mettre en œuvre une fonction d'analyse avancée avec Amazon Athenaet les tableaux de bord avec Amazon QuickSight. » Envisagez d'utiliser la solution CID qui possède des tableaux de bord avancés préconfigurés. Vous pouvez également collaborer avec AWS Partner et adopter leurs solutions de gestion du cloud afin de surveiller et d'optimiser vos factures cloud dans un seul emplacement pratique.

Ressources

Documents connexes :

- Gestion des coûts AWS
- Balisage Ressources AWS
- Analyse des coûts avec AWS Budgets
- Analyse des coûts avec Cost Explorer

- Gestion de AWS Cost and Usage Report
- Catégories de coûts AWS
- Gestion financière du cloud avec AWS
- Example service control policies (Exemples de politiques de contrôle de service)
- Partenaires APN AWS – Gestion des coûts

Vidéos connexes :

- Deploying Cloud Intelligence Dashboards
- Get Alerts on any FinOps or Cost Optimization Metric or KPI

Exemples connexes :

- Ateliers Well-Architected : configuration de compte AWS
- Ateliers Well-Architected : visualisation de la facturation
- Ateliers Well-Architected : utilisation des coûts et de la gouvernance
- Ateliers Well-Architected : utilisation des coûts et de la gouvernance
- Ateliers Well-Architected : visualisation des coûts et de l'utilisation
- Ateliers Well-Architected : Cloud Intelligence Dashboards
- How to use SCPs to set permission guardrails across accounts

COST03-BP06 Répartir les coûts selon les métriques de la charge de travail

Répartissez les coûts de la charge de travail en fonction des métriques d'utilisation ou des résultats économiques afin de mesurer la rentabilité de la charge de travail. Mettez en œuvre un processus pour analyser les données de coût et d'utilisation avec les services d'analytique, ce qui peut fournir des informations et des fonctionnalités de refacturation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

L'optimisation des coûts consiste à obtenir des résultats métier au prix le plus bas, ce qui ne peut être obtenu qu'en allouant les coûts de la charge de travail en fonction des métriques de charge de travail (mesurées par l'efficacité de la charge de travail). Surveillez les métriques de charge de travail définies via des fichiers journaux ou une autre surveillance d'application. Combinez ces

données avec les coûts de la charge de travail qui peuvent être obtenus en examinant les coûts avec une valeur de balise spécifique ou un ID de compte. Il est recommandé d'effectuer cette analyse au niveau horaire. Votre efficacité changera généralement si certains composants de coût sont statiques (par exemple, une base de données principale fonctionnant en permanence) avec un taux de demandes variable (par exemple, des pics d'utilisation entre neuf heures et dix-sept heures, avec peu de demandes la nuit). La compréhension de la relation entre les coûts statiques et variables vous aidera à cibler vos activités d'optimisation.

La création de métriques de charge de travail pour les ressources partagées peut s'avérer difficile par rapport à des ressources telles que les applications conteneurisées sur Amazon Elastic Container Service (Amazon ECS) et Amazon API Gateway. Cependant, il existe certains moyens de catégoriser l'utilisation et de suivre les coûts. Si vous avez besoin de suivre Amazon ECS et les ressources partagées AWS Batch, vous pouvez activer le partage des données de répartition des coûts dans AWS Cost Explorer. Grâce au partage des données de répartition des coûts, vous pouvez comprendre et optimiser le coût et l'utilisation de vos applications conteneurisées et répartir les coûts des applications entre les différentes entités commerciales en fonction de la manière dont les ressources de calcul et de mémoire partagées sont consommées. Si vous avez partagé API Gateway et l'utilisation de la fonction AWS Lambda, vous pouvez utiliser AWS Application Cost Profiler pour catégoriser leur consommation en fonction de leur `ID locataire` ou `ID client`.

Étapes d'implémentation

- Répartissez les coûts dans les métriques de la charge de travail : à l'aide des métriques définies et des balises configurées, créez une métrique qui combine la sortie de la charge de travail et son coût. Utilisez les services d'analytique, tels qu'Amazon Athena et Amazon QuickSight, pour créer un tableau de bord d'efficacité de la charge de travail globale et des composants.

Ressources

Documents connexes :

- Balisage des ressources AWS
- Analyse des coûts avec AWS Budgets
- Analyse des coûts avec Cost Explorer
- Gestion des rapports d'utilisation et de coûts AWS

Exemples connexes :

- Améliorer la visibilité des coûts d'Amazon ECS et de AWS Batch avec les données de répartition des coûts fractionnés AWS

COÛT 4. Comment mettez-vous les ressources hors service ?

Mettez en œuvre le contrôle des modifications et la gestion des ressources depuis le début du projet jusqu'à la fin. Cela garantit que vous arrêtez ou résiliez les ressources inutilisées pour réduire le gaspillage.

Bonnes pratiques

- COST04-BP01 Suivre les ressources pendant toute leur durée de vie
- COST04-BP02 Mettre en œuvre un processus de mise hors service
- COST04-BP03 Mettre hors service des ressources
- COST04-BP04 Mettre hors service des ressources automatiquement
- COST04-BP05 Appliquer les politiques de conservation des données

COST04-BP01 Suivre les ressources pendant toute leur durée de vie

Définissez et mettez en œuvre une méthode pour suivre les ressources et leurs associations avec les systèmes, tout au long de leur durée de vie. Vous pouvez utiliser le balisage pour identifier la charge de travail ou la fonction de la ressource.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Mettez hors service les ressources de charge de travail qui ne sont plus requises. Cela concerne notamment les ressources utilisées pour les tests. Une fois les tests terminés, les ressources peuvent être supprimées. Le suivi des ressources avec des balises (et l'exécution de rapports sur ces balises) peut vous aider à identifier les actifs à mettre hors service, du fait de leur non-utilisation ou de l'expiration de la licence. Les balises constituent un moyen efficace de suivre les ressources, car elles identifient une ressource avec sa fonction ou une date connue à laquelle elle peut être mise hors service. Des rapports sur ces balises peuvent ensuite être exécutés. Par exemple, les valeurs utilisées pour baliser les ressources selon leurs fonctionnalités apparaissent sous la forme `feature-X testing` afin d'identifier l'objectif de la ressource en matière de cycle de vie de la charge de travail. Il est également possible d'utiliser `LifeSpan` ou `TTL` pour les ressources, de même le nom et

la valeur la clé de la balise « to-be-deleted » afin de définir la période ou l'heure spécifique de la mise hors service.

Étapes d'implémentation

- Mettre en œuvre un schéma de balisage : mettez en œuvre un schéma de balisage qui identifie la charge de travail à laquelle appartient la ressource, en veillant à ce que toutes les ressources de la charge de travail soient balisées en conséquence. Le balisage vous aide à catégoriser les ressources par objectif, équipe, environnement ou autres critères pertinents pour votre entreprise. Pour plus de détails sur le balisage des cas d'utilisation, des stratégies et des techniques, consultez Bonnes pratiques de balisage AWS.
- Mettre en œuvre la surveillance du débit ou de la réponse de la charge de travail : mettez en œuvre la surveillance ou les alarmes de contrôle du débit de la charge de travail, qui se déclenchent en fonction des demandes en entrée ou des réponses en sortie. Configurez-la pour envoyer des notifications lorsque les demandes ou les réponses de la charge de travail sont nulles, ce qui indique que ses ressources ne sont plus utilisées. Intégrez un facteur temporel si la charge de travail est régulièrement nulle dans des conditions normales. Pour plus de détails sur les ressources inutilisées ou sous-utilisées, consultez AWS Trusted Advisor Cost Optimization checks(Vérifications de l'optimisation des coûts AWS).
- Regrouper les ressources AWS : créez des groupes pour les ressources AWS. Vous pouvez utiliser AWS Resource Groups pour organiser et gérer vos ressources AWS qui se trouvent dans la même Région AWS. Vous pouvez ajouter des balises à la plupart de vos ressources afin d'identifier et de trier plus facilement ces dernières au sein de votre organisation. Utilisez Tag Editor pour ajouter des balises aux ressources prises en charge en bloc. Envisagez d'utiliser AWS Service Catalog pour créer, gérer et distribuer des portefeuilles de produits approuvés aux utilisateurs finaux et gérer le cycle de vie des produits.

Ressources

Documents connexes :

- AWS Auto Scaling
- AWS Trusted Advisor
- AWS Trusted Advisor Cost Optimization Checks (Vérifications de l'optimisation des coûts AWS)
- Tagging AWS resources (Balisage de ressources AWS)
- Publication des métriques personnalisées

Vidéos connexes :

- How to optimize costs using AWS Trusted Advisor ?

Exemples connexes :

- Comment organiser mes ressources AWS ?
- Comment puis-je optimiser les coûts à l'aide d'AWS Trusted Advisor ?

COST04-BP02 Mettre en œuvre un processus de mise hors service

Mettez en œuvre un processus pour identifier et mettre hors service les ressources inutilisées.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Mettez en place un processus normalisé dans l'ensemble de votre entreprise pour identifier et supprimer les ressources inutilisées. Ce processus doit définir la fréquence à laquelle les recherches sont effectuées, ainsi que les processus de suppression de la ressource, afin de s'assurer que toutes les exigences de l'organisation sont respectées.

Étapes d'implémentation

- Créer et mettre en œuvre un processus de mise hors service : travaillez avec les développeurs et les propriétaires pour créer un processus de mise hors service de la charge de travail et de ses ressources. Le processus doit couvrir la méthode pour vérifier que la charge de travail et chacune de ses ressources sont en cours d'utilisation. Détaillez également les étapes nécessaires à la mise hors service de la ressource, en la supprimant et en garantissant la conformité aux exigences réglementaires. Toutes les ressources associées doivent être incluses, notamment les licences ou le stockage dédié. Notifiez les propriétaires de la charge de travail que le processus de mise hors service a été exécuté.

  Suivez les étapes de mise hors service suivantes pour effectuer une à une les vérifications requises dans le cadre de votre processus :

  - Identifier les ressources à mettre hors service : identifiez les ressources qui sont éligibles pour la mise hors service dans votre AWS Cloud. Enregistrez toutes les informations nécessaires et planifiez la mise hors service. Dans votre chronologie, assurez-vous d’envisager la survenue de

problèmes imprévus et d'identifier les étapes les plus propices à cette éventualité au cours du processus.

- Coordonner et communiquer : collaborez avec les propriétaires de la charge de travail pour confirmer la ressource à mettre hors service.
- Enregistrer des métadonnées et créer des sauvegardes : enregistrez des métadonnées (telles que les IP publiques, la région, la zone de disponibilité, le VPC, le sous-réseau et les groupes de sécurité) et créez des sauvegardes (telles que des instantanés Amazon Elastic Block Store ou des AMI, l'exportation des clés et des certificats) si elles sont nécessaires pour les ressources dans l'environnement de production ou s'il s'agit de ressources essentielles.
- Valider l'infrastructure en tant que code : déterminez si les ressources ont été déployées avec AWS CloudFormation, Terraform, AWS Cloud Development Kit (AWS CDK) ou tout autre outil de déploiement d'infrastructure en tant que code afin de pouvoir les redéployer si nécessaire.
- Éviter l'accès : appliquez des contrôles restrictifs pendant un certain temps, afin d'empêcher l'utilisation des ressources pendant que vous déterminez si la ressource est nécessaire. Vérifiez que l'environnement des ressources peut être rétabli à son état d'origine si nécessaire.
- Suivre votre processus de mise hors service interne : suivez les tâches administratives et le processus de mise hors service de votre organisation, comme le retrait de la ressource du domaine de votre organisation, la suppression de l'enregistrement DNS et la suppression de la ressource de votre outil de gestion de la configuration, de votre outil de surveillance, de votre outil d'automatisation et d'autres outils de sécurité.

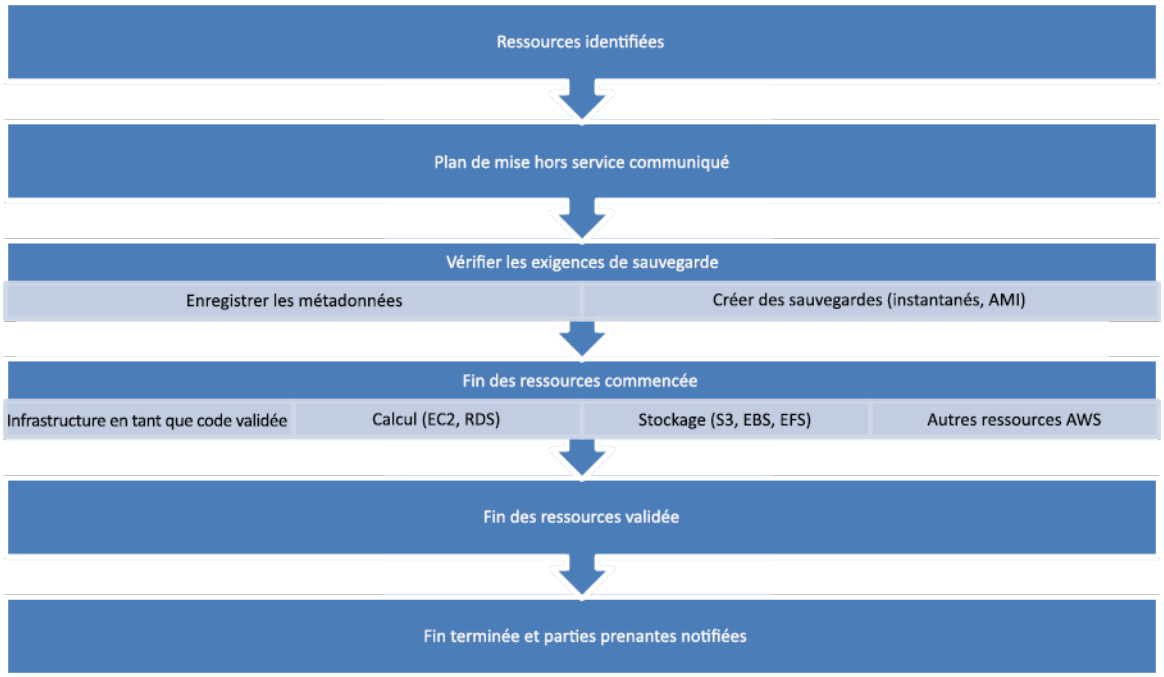
Si la ressource est une instance Amazon EC2, consultez la liste suivante. Pour plus de détails, consultez Comment supprimer ou résilier mes ressources Amazon EC2 ?

- Arrêtez ou résiliez toutes vos instances Amazon EC2 et vos équilibreurs de charge. Les instances Amazon EC2 sont visibles dans la console pendant une courte période après leur résiliation. Vous n'êtes pas facturé pour les instances qui ne sont pas en cours d'exécution
- Supprimez votre infrastructure Auto Scaling.
- Libérez tous les hôtes dédiés.
- Supprimez tous les volumes Amazon EBS et tous les instantanés Amazon EBS.
- Libérez toutes les adresses Elastic IP.
- Annulez l'enregistrement de toutes les Amazon Machine Images (AMI).
- Résiliez tous les environnements AWS Elastic Beanstalk.

Si la ressource est un objet dans le stockage Amazon S3 Glacier et si vous supprimez une archive avant d'atteindre la durée minimale de stockage, vous serez facturé une taxe de suppression

anticipée au prorata. La durée minimale de stockage Amazon S3 Glacier dépend de la classe de stockage utilisée. Si vous souhaitez accéder à un résumé de la durée minimale de stockage pour chaque classe de stockage, consultez Performances des classes de stockage Amazon S3. Pour plus de détails sur les frais de suppression précoce, consultez la section Tarification Amazon S3.

Le diagramme suivant du processus de mise hors service simple décrit les étapes de la mise hors service. Avant de mettre hors service des ressources, vérifiez que les ressources identifiées pour la mise hors service ne sont pas utilisées par l'organisation.



Flux de mise hors service des ressources.

Ressources

Documents connexes :

- AWS Auto Scaling
- AWS Trusted Advisor
- AWS CloudTrail

Vidéos connexes :

- Delete CloudFormation stack but retain some resources
- Find out which user launched Amazon EC2 instance

Exemples connexes :

- Comment supprimer ou résilier mes ressources Amazon EC2 ?
- Comment savoir quel utilisateur a lancé une instance Amazon EC2 dans mon compte ?

COST04-BP03 Mettre hors service des ressources

Mettez hors service les ressources déclenchées par des événements tels que les audits périodiques ou les modifications d'utilisation. La mise hors service est généralement effectuée régulièrement et elle peut être manuelle ou automatisée.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

La fréquence et l'effort de recherche des ressources inutilisées doivent refléter les économies potentielles, de sorte qu'un compte ayant un faible coût doit être analysé moins fréquemment qu'un compte ayant des coûts plus importants. Les recherches et les événements de mise hors service peuvent être déclenchés par des changements d'état dans la charge de travail, comme un produit en fin de vie ou en cours de remplacement. Les recherches et les événements de mise hors service peuvent également être déclenchés par des événements externes, tels que des changements dans les conditions du marché ou l'arrêt d'un produit.

Étapes d'implémentation

- Mettre hors service les ressources : il s'agit de l'étape où les ressources AWS deviennent obsolètes, car elles ne sont plus nécessaires ou car leur contrat de licence a expiré. Effectuez toutes les vérifications finales avant de passer à l'étape de l'élimination et de mettre hors service les ressources afin d'éviter toute perturbation indésirable, comme la réalisation d'instantanés ou de sauvegardes. En utilisant le processus dédié, mettez hors service chaque ressource ayant été identifiée comme inutilisée.

Ressources

Documents connexes :

- AWS Auto Scaling
- AWS Trusted Advisor

Exemples connexes :

- Well-Architected Labs: Decommission resources (Level 100) (Ateliers Well-Architected : mettre hors service des ressources (niveau 100))

COST04-BP04 Mettre hors service des ressources automatiquement

Concevez votre charge de travail de manière à gérer proprement l'arrêt des ressources lorsque vous identifiez et mettez hors service des ressources non critiques, des ressources qui ne sont pas nécessaires ou des ressources peu utilisées.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Utilisez l'automatisation pour réduire ou supprimer les coûts associés au processus de mise hors service. La conception de votre charge de travail pour effectuer une mise hors service automatisée réduira le coût global de la charge de travail pendant sa durée de vie. Vous pouvez utiliser AWS Auto Scaling pour effectuer le processus de mise hors service. Vous pouvez également implémenter du code personnalisé à l'aide de l'API ou du SDK pour mettre automatiquement hors service des ressources de charge de travail.

Les applications modernes sont d'abord créées sans serveur, une stratégie qui privilégie l'adoption de services sans serveur. AWS a développé des services sans serveur pour les trois couches de votre pile : calcul, intégration et magasins de données. L'utilisation d'une architecture sans serveur vous permettra de réduire les coûts pendant les périodes de faible trafic avec une mise à l'échelle automatique.

Étapes d'implémentation

- Implémenter AWS Auto Scaling : configurez les ressources prises en charge avec AWS Auto Scaling. AWS Auto Scaling peut vous aider à optimiser votre efficacité en termes d'utilisation et de coûts lors de l'utilisation des services AWS. Lorsque la demande baisse, AWS Auto Scaling supprime automatiquement toute capacité excédentaire afin d'éviter les dépenses excessives.
- Configurer CloudWatch pour résilier les instances : les instances peuvent être configurées de façon à être résiliées à l'aide d'alarmes CloudWatch. En utilisant les métriques du processus de mise hors service, mettez en œuvre une alarme avec une action Amazon Elastic Compute Cloud. Veillez à vérifier l'opération dans un environnement hors production avant le déploiement.

- Implémenter du code au sein de la charge de travail : vous pouvez utiliser AWS SDK ou l'AWS CLI pour mettre hors service les ressources de charge de travail. Mettez en œuvre le code d'application qui s'intègre à AWS et qui résilie ou supprime les ressources qui ne sont plus utilisées.
- Utiliser des services sans serveur : privilégiez la création d'architectures sans serveur et d'une architecture axée sur les événements sur AWS afin de créer et d'exécuter vos applications. AWS offre plusieurs services technologiques sans serveur qui, de façon inhérente, permettent d'optimiser automatiquement l'utilisation des ressources et la mise hors service automatisée (augmentation et diminution de l'échelle). Avec des applications sans serveur, l'utilisation des ressources est optimisée automatiquement et vous ne payez jamais un approvisionnement excessif.

Ressources

Documents connexes :

- AWS Auto Scaling
- AWS Trusted Advisor
- Serverless on AWS (Sans serveur sur AWS)
- Création d'alarmes qui arrêtent, mettent hors service, redémarrent ou récupèrent une instance EC2
- Commencer avec Amazon EC2 Auto Scaling
- Ajouter des actions d'arrêt aux alarmes Amazon CloudWatch

Exemples connexes :

- Scheduling automatic deletion of AWS CloudFormation stacks
- Well-Architected Labs – Decommission resources automatically (Level 100) (Ateliers Well-Architected : mettre hors service les ressources automatiquement (niveau 100))
- Servian AWS Auto Cleanup (Nettoyage automatique AWS Servian)

COST04-BP05 Appliquer les politiques de conservation des données

Définissez des politiques de conservation des données sur les ressources prises en charge pour traiter la suppression des objets conformément aux exigences de votre organisation. Identifiez et supprimez les ressources et les objets inutiles ou orphelins qui ne sont plus nécessaires.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Utilisez des politiques de conservation des données et des politiques de cycle de vie pour réduire les coûts associés au processus de mise hors service et les coûts de stockage des ressources identifiées. Le fait de définir une migration de classe de stockage ainsi qu'une suppression automatisées dans vos politiques de conservation des données et de cycle de vie réduira les frais de stockage généraux pendant leur durée de vie. Vous pouvez utiliser Amazon Data Lifecycle Manager pour automatiser la création et la suppression d'instantanés Amazon Elastic Block Store et d'Amazon Machine Images (AMI) basées sur Amazon EBS, et utiliser Amazon S3 Intelligent-Tiering ou une configuration du cycle de vie Amazon S3 pour gérer le cycle de vie de vos objets Amazon S3. Vous pouvez également mettre en œuvre du code personnalisé avec l'API ou le kit SDK pour créer des politiques de cycle de vie et des règles de politique pour des objets à supprimer automatiquement.

Étapes d'implémentation

- Utiliser Amazon Data Lifecycle Manager : utilisez des politiques de cycle de vie sur Amazon Data Lifecycle Manager pour automatiser la suppression d'instantanés Amazon EBS et d'AMI basées sur Amazon EBS.
- Définir une configuration du cycle de vie sur un compartiment : utilisez une configuration du cycle de vie Amazon S3 sur un compartiment afin de définir des actions que Amazon S3 doit réaliser au cours du cycle de vie d'un objet, ainsi que la suppression à la fin du cycle de vie de l'objet, selon les exigences de votre entreprise.

Ressources

Documents connexes :

- AWS Trusted Advisor
- Amazon Data Lifecycle Manager
- Comment définir la configuration du cycle de vie sur un compartiment Amazon S3

Vidéos connexes :

- Automate Amazon EBS Snapshots with Amazon Data Lifecycle Manager (Automatiser les instantanés Amazon EBS avec Amazon Data Lifecycle Manager)
- Empty an Amazon S3 bucket using a lifecycle configuration rule (Vider un compartiment Amazon S3 à l'aide d'une règle de configuration du cycle de vie)

Exemples connexes :

- Empty an Amazon S3 bucket using a lifecycle configuration rule (Vider un compartiment Amazon S3 à l'aide d'une règle de configuration du cycle de vie)
- Well-Architected Labs: Decommission resources automatically (Level 100) (Ateliers Well-Architected : mettre hors service les ressources automatiquement (niveau 100))

Ressources rentables

Questions

- COÛT 5. Comment évaluer les coûts lorsque vous sélectionnez des services ?
- COÛT 6. Comment atteindre les objectifs de coût lorsque vous sélectionnez le type, la taille et le nombre de ressources ?
- COÛT 7. Comment utiliser les modèles de tarification pour réduire les coûts ?
- COÛT 8. Comment planifiez-vous les frais de transfert de données ?

COÛT 5. Comment évaluer les coûts lorsque vous sélectionnez des services ?

Amazon EC2, Amazon EBS et Amazon S3 sont des services fondamentaux d'AWS. Les services gérés, tels que Amazon RDS et Amazon DynamoDB, sont des services AWS de plus haut niveau, ou au niveau de l'application. En sélectionnant les services fondamentaux et les services gérés appropriés, vous pouvez optimiser cette charge de travail en termes de coûts. Par exemple, en utilisant des services gérés, vous pouvez réduire ou supprimer une grande partie de votre traitement administratif et opérationnel, et vous dégagez ainsi du temps pour travailler sur les applications et les activités liées aux activités.

Bonnes pratiques

- COST05-BP01 Identifier les exigences de l'organisation en matière de coûts
- COST05-BP02 Analyser tous les composants de la charge de travail
- COST05-BP03 Effectuer une analyse approfondie de chaque composant
- COST05-BP04 Sélectionner des logiciels avec des licences rentables
- COST05-BP05 Sélectionner les composants de cette charge de travail afin d'optimiser les coûts en fonction des priorités de l'organisation
- COST05-BP06 Analyser les coûts d'une utilisation différente dans le temps

COST05-BP01 Identifier les exigences de l'organisation en matière de coûts

Collaborez avec les membres de l'équipe pour définir l'équilibre entre l'optimisation des coûts et les autres piliers, tels que Performance et Fiabilité, pour cette charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Dans la plupart des organisations, le département des technologies de l'information (TI) est composé de plusieurs petites équipes, chacune ayant son propre programme et son propre domaine d'intervention, qui reflètent les spécialités et les compétences des membres de son équipe. Vous devez comprendre les objectifs généraux, les priorités et les buts de votre organisation et la manière dont chaque département ou projet contribue à ces objectifs. La catégorisation de toutes les ressources essentielles, notamment le personnel, les équipements, les technologies, le matériel et les services externes, est cruciale pour atteindre les objectifs de l'organisation et mettre en place une planification budgétaire exhaustive. L'adoption de cette approche systématique de l'identification et de la compréhension des coûts est fondamentale pour établir un plan de coûts réaliste et solide pour l'organisation.

Lorsque vous sélectionnez des services pour votre charge de travail, il est essentiel que vous compreniez les priorités de votre entreprise. Créez un équilibre entre l'optimisation des coûts et les autres piliers d'AWS Well-Architected Framework, tels que les performances et la fiabilité. Ce processus doit être mené de manière systématique et régulière afin de refléter l'évolution des objectifs de l'organisation, des conditions du marché et de la dynamique opérationnelle. Une charge de travail entièrement optimisée en matière de coûts est la solution la plus conforme aux besoins de votre organisation, et pas nécessairement la moins coûteuse. Rencontrez toutes les équipes de votre organisation (équipes produits, commerciales, techniques et financières) pour recueillir des informations. Évaluez l'impact des compromis entre des intérêts concurrents ou des approches alternatives pour prendre des décisions éclairées au moment de déterminer où concentrer les efforts ou de choisir une ligne de conduite.

Par exemple, l'accélération de la mise sur le marché de nouvelles fonctionnalités peut être privilégiée par rapport à l'optimisation des coûts, ou vous pouvez choisir une base de données relationnelle pour les données non relationnelles afin de simplifier l'effort de migration d'un système, plutôt que de migrer vers une base de données optimisée pour votre type de données et de mettre à jour votre application.

Étapes d'implémentation

- Identifier les exigences de l'organisation en matière de coûts : rencontrez les équipes de votre organisation, y compris les chefs de produits, les responsables d'applications, les équipes de développement et d'exploitation, la direction et les services financiers. Hiérarchisez les piliers Well-Architected pour cette charge de travail et ses composants. Vous devriez obtenir un classement des piliers par ordre de priorité. Vous pouvez également attribuer une pondération pour indiquer le degré de priorité supplémentaire d'un pilier ou une similarité de priorité entre deux piliers.
- Réglez la dette technique et documentez-la : lors de l'examen de la charge de travail, réglez la dette technique. Documentez un élément en attente pour réexaminer la charge de travail à l'avenir dans le but de la refactoriser ou de la réorganiser pour l'optimiser davantage. Il est essentiel de communiquer clairement les concessions qui ont été faites aux autres parties prenantes.

Ressources

Bonnes pratiques associées :

- REL11-BP07 Concevoir votre produit pour atteindre les objectifs de disponibilité et les accords de niveau de service (SLA)
- OPS01-BP06 Évaluer les compromis

Documents connexes :

- Calculateur du coût total de possession (TCO) d'AWS
- Classes de stockage Amazon S3
- Produits cloud

COST05-BP02 Analyser tous les composants de la charge de travail

Assurez-vous que chaque composant de la charge de travail est analysé, peu importe la taille ou les coûts actuels. L'effort de vérification doit tenir compte des avantages potentiels, tels que les coûts actuels et prévus.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les composants de la charge de travail, qui sont conçus pour apporter une valeur métier à l'organisation, peuvent englober différents services. Pour chaque composant, il est possible de choisir des services AWS Cloud spécifiques répondant aux besoins de l'entreprise. Cette sélection peut être influencée par des facteurs tels que la connaissance ou l'expérience antérieure de ces services.

Après avoir identifié les besoins de votre organisation (comme indiqué dans COST05-BP01 Identifier les exigences de l'organisation en matière de coûts), analysez en profondeur tous les composants de votre charge de travail. Analysez chaque composant en tenant compte des coûts et des tailles actuels et prévus. Examinez le coût de l'analyse par rapport aux économies potentielles de la charge de travail au cours de son cycle de vie. L'effort d'analyse de tous les composants de cette charge de travail doit correspondre aux économies ou aux améliorations potentielles escomptées grâce à l'optimisation de ce composant spécifique. Par exemple, si le coût de la ressource proposée est de 10 USD par mois et que les charges prévues ne dépassent pas 15 USD par mois, une journée d'effort pour réduire les coûts de 50 % (5 USD par mois) pourrait dépasser le bénéfice potentiel sur la durée de vie du système. L'utilisation d'une estimation plus rapide et plus efficace basée sur des données permet d'obtenir le meilleur résultat global pour ce composant.

Les charges de travail peuvent évoluer dans le temps, et un ensemble de services qui est actuellement adapté peut ne pas être optimal si l'architecture ou l'utilisation de la charge de travail évolue. L'analyse pour la sélection des services doit intégrer les états de charge de travail et les niveaux d'utilisation actuels et futurs. La mise en œuvre d'un service pour un état ou un usage futur de la charge de travail peut réduire les coûts globaux en diminuant ou en supprimant l'effort nécessaire pour effectuer des changements futurs. Par exemple, Amazon EMR Serverless pourrait être un bon choix au départ. Toutefois, à mesure que la consommation de ce service augmente, le passage à Amazon EMR sur Amazon EC2 pourrait réduire les coûts liés à ce composant de la charge de travail.

L'examen stratégique de tous les composants de la charge de travail, quelles que soient leurs caractéristiques actuelles, est susceptible d'apporter des améliorations notables et des économies financières au fil du temps. L'effort déployé dans ce processus d'évaluation doit être délibéré, et tenir dûment compte des bénéfices potentiels qui pourraient en découler.

AWS Cost Explorer et AWS Cost and Usage Report (CUR) permettent d'analyser le coût d'une démonstration de faisabilité (Proof of Concept, PoC) ou d'un environnement en cours d'exécution. Vous pouvez également utiliser AWS Pricing Calculator pour estimer les coûts de la charge de travail.

Étapes d'implémentation

- Répertorier les composants de la charge de travail : dressez la liste des composants de votre charge de travail afin de vérifier que chaque composant a été analysé. L'effort déployé doit refléter la sévérité de la charge de travail telle que définie par les priorités de l'organisation. Le regroupement fonctionnel des ressources améliore l'efficacité, notamment du stockage des bases de données de production s'il existe plusieurs bases de données.
- Classer les composants de la liste par ordre de priorité : prenez la liste des composants et classez-la par ordre d'effort. Elle est généralement classée par ordre de coût du composant (du plus cher au moins cher) ou par ordre de criticité (telle qu'elle est définie par les priorités de votre organisation).
- Effectuer l'analyse : pour chaque élément de la liste, passez en revue les options et les services disponibles et choisissez l'option qui correspond le mieux aux priorités de votre organisation.

Ressources

Documents connexes :

- AWS Pricing Calculator
- AWS Cost Explorer
- Classes de stockage Amazon S3
- Produits cloud

COST05-BP03 Effectuer une analyse approfondie de chaque composant

Examinez le coût global de chaque composant pour l'organisation. Calculez le coût total de possession en tenant compte du coût des opérations et de la gestion, en particulier lorsque vous utilisez des services gérés par un fournisseur de cloud. L'effort d'examen doit refléter les avantages potentiels (par exemple, la durée de l'analyse est proportionnelle au coût du composant).

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Débit

Directives d'implémentation

Tenez compte du gain de temps qui permettra à votre équipe de se concentrer sur le remboursement de la dette technique, l'innovation, les fonctionnalités à valeur ajoutée et la création de votre avantage différentiel. Par exemple, il peut être nécessaire de procéder à un lift and shift (également appelé

réhébergement) de vos bases de données depuis votre environnement sur site vers le cloud aussi rapidement que possible et de l'optimiser ultérieurement. Il est intéressant d'explorer les économies possibles réalisées en utilisant des services gérés sur AWS qui peuvent supprimer ou réduire les coûts de licence. Les services gérés sur AWS suppriment la charge opérationnelle et administrative liée à la gestion d'un service, comme la correction ou la mise à niveau du système d'exploitation, et vous permettre de vous consacrer à l'innovation et l'entreprise.

Étant donné que les services gérés fonctionnent sur le cloud, ils peuvent réduire le coût par transaction ou par service. Vous pouvez effectuer des optimisations potentielles afin d'obtenir des bénéfices concrets, sans pour autant changer l'architecture de base de l'application. Par exemple, vous pouvez chercher à réduire la quantité de temps passé à gérer les instances de bases de données en migrant vers une plateforme de base de données en tant que service, telle qu'Amazon Relational Database Service (Amazon RDS), ou en procédant à la migration de votre application vers une plateforme entièrement gérée, telle qu'AWS Elastic Beanstalk.

En général, les services gérés ont des attributs que vous pouvez définir pour assurer une capacité suffisante. Vous devez définir et surveiller ces attributs afin que votre capacité excédentaire soit réduite au minimum et que vos performances soient maximisées. Vous pouvez modifier les attributs des AWS Managed Services à l'aide d'AWS Management Console ou des API et kits SDK AWS pour aligner les besoins en ressources sur l'évolution de la demande. Par exemple, vous pouvez augmenter ou diminuer le nombre de nœuds sur un cluster Amazon EMR (ou un cluster Amazon Redshift) pour monter ou descendre en puissance.

Vous pouvez également regrouper plusieurs instances sur une ressource AWS pour permettre une utilisation de plus haute densité. Par exemple, vous pouvez mettre en service plusieurs petites bases de données sur une seule instance de base de données Amazon Relational Database Service (Amazon RDS). Alors que l’utilisation augmente, vous pouvez migrer l'une des bases de données vers une instance de base de données Amazon RDS dédiée en utilisant un processus d'instantané et de restauration.

Lors de la mise en service de charges de travail sur des services gérés, vous devez connaître les exigences d'ajustement de la capacité du service. Ces exigences sont généralement le temps, l'effort et toute incidence sur le fonctionnement normal de la charge de travail. La ressource allouée doit laisser le temps à tout changement de se produise, en allouant la surcharge requise pour le permettre. L'effort continu nécessaire pour modifier les services peut être réduit à pratiquement zéro en utilisant des API et des kits SDK intégrés à des outils système et de surveillance, tels qu'Amazon CloudWatch.

Amazon RDS, Amazon Redshift et Amazon ElastiCache fournissent un service de base de données géré. Amazon Athena, Amazon EMR et Amazon OpenSearch Service fournissent un service d'analyse géré.

AMS est un service qui exploite l'infrastructure AWS pour le compte des entreprises clientes et partenaires. Il fournit un environnement sécurisé et conforme sur lequel vous pouvez déployer vos charges de travail. AMS utilise des modèles d'exploitation de cloud d'entreprise avec l'automatisation pour permettre de répondre aux exigences de votre organisation, de migrer plus rapidement vers le cloud et de réduire vos coûts de gestion continue.

Étapes d'implémentation

- Réaliser une analyse complète : à l'aide de la liste des composants, examinez chaque composant de la plus haute priorité à la plus basse. Pour les composants les plus prioritaires et les plus coûteux, effectuez une analyse supplémentaire et évaluez toutes les options disponibles et leur impact sur le long terme. Pour les composants de moindre priorité, évaluez si des changements d'utilisation modifieraient la priorité du composant, puis analysez l'effort approprié.
- Comparer les ressources gérées et non gérées : prenez en compte le coût opérationnel des ressources que vous gérez et comparez-le aux ressources gérées par AWS. Par exemple, évaluez vos bases de données s'exécutant sur des instances Amazon EC2 et comparez-les aux options Amazon RDS (un service géré par AWS) ou Amazon EMR par rapport à l'exécution d'Apache Spark sur Amazon EC2. Étudiez soigneusement vos options quand vous passez d'une charge de travail autogérée à une charge de travail entièrement gérée par AWS. Les trois facteurs les plus importants à prendre en compte sont le type de service géré que vous voulez utiliser, le processus que vous utiliserez pour procéder à la migration de vos données et le fait de comprendre le modèle de responsabilité partagée d'AWS.

Ressources

Documents connexes :

- Calculateur du coût total de possession (TCO) d'AWS
- Classes de stockage Amazon S3
- Produits AWS Cloud
- Modèle de responsabilité partagée d'AWS

Vidéos connexes :

- Why move to a managed database? (Pourquoi déplacer une base de données gérée ?)
- What is Amazon EMR and how can I use it for processing data? (Qu'est-ce qu'Amazon EMR et comment l'utiliser pour traiter des données ?)

Exemples connexes :

- Why move to a managed database? (Pourquoi déplacer une base de données gérée ?)
- Consolidate data from identical SQL Server databases into a single Amazon RDS for SQL Server database using AWS DMS (Consolider des données de bases de données SQL Server identiques en une seule base de données Amazon RDS for SQL Server avec AWS DMS)
- Deliver data at scale to Amazon Managed Streaming for Apache Kafka (Amazon MSK) (Livrer des données à grande échelle à Amazon Managed Streaming for Apache Kafka (Amazon MSK))
- Migrate an ASP.NET web application to AWS Elastic Beanstalk (Procédez à la migration d'une application web ASP.NET vers AWS Elastic Beanstalk)

COST05-BP04 Sélectionner des logiciels avec des licences rentables

Les logiciels open source éliminent les coûts de licences logicielles, qui peuvent entraîner des coûts significatifs pour la charge de travail. Lorsque des logiciels sous licence sont nécessaires, évitez les licences liées à des attributs arbitraires tels que les CPU. Recherchez les licences qui sont liées à des résultats. Le coût de ces licences est plus proche de l'avantage qu'elles procurent.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

L'open source est né dans le contexte du développement de logiciels pour indiquer que le logiciel est conforme à certains critères de distribution gratuite. Les logiciels open source sont composés de code source que tout le monde peut inspecter, modifier et améliorer. En fonction des exigences métier, des compétences des ingénieurs, de l'utilisation prévue et d'autres dépendances technologiques, les organisations peuvent envisager d'utiliser des logiciels open source sur AWS afin de minimiser leurs coûts de licence. En d'autres termes, le coût des licences logicielles peut être réduit grâce à l'utilisation de logiciels open source. Cela peut avoir un impact significatif sur les coûts de charge de travail à mesure que la taille de la charge de travail évolue.

Mesurez les avantages des logiciels sous licence par rapport au coût total pour optimiser votre charge de travail. Modélisez les modifications apportées aux licences et leur impact sur vos coûts de charge de travail. Si un fournisseur modifie le coût de votre licence de base de données, examinez

en quoi cela affecte l'efficacité globale de votre charge de travail. Prenez en compte l'historique des annonces de tarification de vos fournisseurs pour connaître les tendances des changements de licence pour leurs produits. Les coûts de licence peuvent également évoluer indépendamment du débit ou de l'utilisation, comme les licences qui évoluent en fonction du matériel (licences liées à l'UC). Ces licences doivent être évitées, car les coûts peuvent rapidement augmenter sans résultats correspondants.

Par exemple, l'exécution d'une instance Amazon EC2 sous Linux dans us-east-1 vous permet de réduire les coûts d'environ 45 % par rapport à l'exécution d'une autre instance Amazon EC2 sous Windows.

Le AWS Pricing Calculator offre un moyen complet de comparer les coûts de différentes ressources avec différentes options de licence, telles que des instances Amazon RDS et des moteurs de base de données. En outre, AWS Cost Explorer fournit une vue précieuse des coûts des charges de travail existantes, en particulier celles proposées avec différentes licences. Pour la gestion des licences, AWS License Manager propose une méthode rationalisée pour superviser et gérer les licences logicielles. Les clients peuvent déployer et utiliser leur logiciel open source préféré dans AWS Cloud.

Étapes d'implémentation

- Analyser les options de licence : passez en revue les conditions de licence des logiciels disponibles. Recherchez les versions open source qui ont les fonctionnalités requises, et déterminez si les avantages des logiciels sous licence l'emportent sur le coût. Des conditions favorables permettent d'aligner le coût du logiciel sur les avantages qu'il procure.
- Analyser l'éditeur du logiciel : passez en revue l'historique des changements de prix ou de licence de l'éditeur. Recherchez les changements qui ne s'alignent pas sur les résultats, tels que les conditions pénalisantes de l'exécution sur des matériels ou des plates-formes spécifiques à un fournisseur. Déterminez également comment ils effectuent les audits et les sanctions qui pourraient être imposées.

Ressources

Documents connexes :

- Open Source at AWS
- Calculateur du coût total de possession (TCO) d'AWS
- Classes de stockage Amazon S3
- Produits cloud

Exemples connexes :

- Open Source Blog
- Open Source at AWS
- Évaluation de l'optimisation et des licences

COST05-BP05 Sélectionner les composants de cette charge de travail afin d'optimiser les coûts en fonction des priorités de l'organisation

Tenez compte du coût lorsque vous sélectionnez tous les composants de votre charge de travail. Cela inclut l'utilisation de services gérés au niveau des applications et des services sans serveur, de conteneurs ou d'une architecture basée sur les événements pour réduire le coût global. Réduisez les coûts de licence en utilisant des logiciels open source, des logiciels qui ne comportent pas de frais de licence ou des alternatives pour réduire les dépenses.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Tenez compte du coût des services et des options lorsque vous sélectionnez tous les composants. Cela inclut l'utilisation de services gérés au niveau de l'application, comme Amazon Relational Database Service (Amazon RDS), Amazon DynamoDB, Amazon Simple Notification Service (Amazon SNS) et Amazon Simple Email Service (Amazon SES) pour réduire le coût global de l'organisation.

Utilisez les services sans serveur et les conteneurs pour le calcul, comme AWS Lambda et Amazon Simple Storage Service (Amazon S3) pour les sites web statiques. Placez votre application dans un conteneur, si possible, et utilisez des services de conteneurs gérés AWS comme Amazon Elastic Container Service (Amazon ECS) ou Amazon Elastic Kubernetes Service (Amazon EKS).

Réduisez les coûts de licence en utilisant des logiciels open source, ou des logiciels qui n'impliquent pas de frais de licence (par exemple, Amazon Linux pour les charges de travail de calcul ou la migration des bases de données vers Amazon Aurora).

Vous pouvez utiliser des services serverless ou de niveau application tels que Lambda, Amazon Simple Queue Service (Amazon SQS), Amazon SNSet Amazon SES. » Ces services vous dispensent de gérer une ressource et assurent les fonctions d'exécution de code, de mise en file d'attente et de distribution de messages. L'autre avantage est qu'ils sont mis à l'échelle en termes

de performances et de coûts en fonction de l'utilisation, ce qui permet une répartir et d'attribuer efficacement les coûts.

L'utilisation de l'architecture basée sur les événements est également possible avec les services sans serveur. Les architectures basées sur les événements reposent sur la technologie push, ce qui signifie que tout se passe à la demande au fur et à mesure que l'événement se présente dans le routeur. Ainsi, vous ne payez pas pour qu'une interrogation continue vérifie un événement. Il en résulte moins de consommation de bande passante du réseau, moins d'utilisation du processeur, moins de capacité de flotte inactive ou moins de liaisons SSL/TLS.

Pour en savoir plus sur les services sans serveur, consultez le livre blanc de la lentille d'application serverless Well-Architected.

Étapes d'implémentation

- Sélectionner chaque service pour optimiser les coûts : À l'aide de votre liste de priorités et d'analyse, sélectionnez chaque option qui correspond le mieux à vos priorités organisationnelles. Au lieu d'augmenter la capacité pour répondre à la demande, envisagez d'autres options qui peuvent vous offrir de meilleures performances à moindre coût. Par exemple, si vous devez évaluer le trafic attendu pour vos bases de données sur AWS, envisagez d'augmenter la taille d'instance ou d'utiliser des services Amazon ElastiCache (Redis ou Memcached) afin de fournir des mécanismes mis en cache à vos bases de données.
- Évaluer l'architecture basée sur les événements : Une architecture sans serveur vous permet également de créer une architecture basée sur les événements pour les applications distribuées reposant sur des micro-services, ce qui vous aide à créer des solutions évolutives, résilientes, flexibles et rentables.

Ressources

Documents connexes :

- Calculateur AWS de coût total de possession (TCO)
- AWS sans serveur
- Qu'est-ce qu'une architecture basée sur les événements ?
- Classes de stockage de fichiers Amazon S3
- Produits cloud
- Amazon ElastiCache (Redis OSS)

Exemples connexes :

- Getting started with event-driven architecture (Démarrer avec une architecture basée sur les événements)
- Architecture basée sur les événements
- Comment Statsig fonctionne de manière 100 fois plus économique grâce à Amazon ElastiCache (Redis OSS)
- Bonnes pratiques d'utilisation des fonctions AWS Lambda

COST05-BP06 Analyser les coûts d'une utilisation différente dans le temps

Les charges de travail peuvent changer au fil du temps. Certains services ou fonctionnalités sont plus rentables à différents niveaux d'utilisation. En analysant chaque composant dans le temps et en fonction de l'utilisation prévue, la charge de travail reste rentable pendant toute sa durée de vie.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Au fur et à mesure qu'AWS lance de nouveaux services et de nouvelles fonctionnalités, les services optimaux pour votre charge de travail peuvent évoluer. L'effort requis doit refléter les avantages potentiels. La fréquence de révision de la charge de travail dépend des exigences de votre organisation. S'il s'agit d'une charge de travail d'un coût important, la mise en œuvre de nouveaux services plus tôt permettra de maximiser les économies, de sorte qu'un examen plus fréquent peut être avantageux. Un autre déclencheur à vérifier est le changement des modèles d'utilisation. D'importants changements d'utilisation peuvent indiquer que d'autres services seraient plus optimaux.

Si vous devez déplacer des données vers AWS Cloud, vous pouvez sélectionner un large éventail de services offerts par AWS et d'outils de partenaires afin de vous aider pour la migration de vos jeux de données, qu'il s'agisse de fichiers, de bases de données, d'images de machine, de volumes en bloc ou même de sauvegardes sur bande. Par exemple, pour déplacer une importante quantité de données vers et depuis AWS ou traiter des données en périphérie, vous pouvez utiliser l'un des appareils sur mesure AWS pour déplacer des données hors ligne de façon rentable. Autre exemple, pour des vitesses de transfert de données plus élevées, un service de connexion directe peut être moins cher qu'un VPN et fournir la connectivité constante requise pour votre entreprise.

Évaluez votre activité de mise à l'échelle en fonction de l'analyse des coûts pour une utilisation différente au fil du temps. Analysez le résultat pour voir si la politique de mise à l'échelle peut être

ajustée pour ajouter des instances avec plusieurs types d'instances et d'options d'achat. Vérifiez vos paramètres pour voir si le minimum peut être réduit pour satisfaire les demandes des utilisateurs avec une plus petite taille de flotte et ajouter davantage de ressources pour répondre à la demande élevée attendue.

Analysez les coûts pour une utilisation différente au fil du temps en discutant avec les parties prenantes de votre organisation et utilisez la fonction de prévision d'AWS Cost Explorer pour prédire l'impact possible des changements de service. Surveillez les déclencheurs de niveau d'utilisation avec AWS Budgets, les alarmes de facturation CloudWatch et AWS Cost Anomaly Detection pour identifier et mettre en œuvre les services les plus rentables plus rapidement.

Étapes d'implémentation

- Définir des modèles d'utilisation prévisibles : en collaboration avec votre organisation, par exemple, les responsables du marketing et les propriétaires de produit, documentez les modes d'utilisation attendus et prévus de la charge de travail. Discutez avec les parties prenantes de votre entreprise des augmentations de coûts et d'utilisation historiques et prévues et assurez-vous que les augmentations s'alignent sur les exigences de votre entreprise. Identifiez les jours, les semaines ou les mois au cours desquels vous vous attendez à ce que davantage d'utilisateurs utilisent vos ressources AWS, indiquant que vous devriez augmenter la capacité des ressources existantes ou adopter des services supplémentaires pour réduire les coûts et augmenter les performances.
- Analyser les coûts au niveau d'utilisation prévue : à l'aide des modèles d'utilisation définis, analysez chacun de ces points. L'effort d'analyse doit refléter le résultat potentiel. Par exemple, si le changement d'utilisation est important, une analyse approfondie doit être effectuée pour vérifier les coûts et les changements éventuels. En d'autres termes, quand les coûts augmentent, l'utilisation de l'entreprise doit également augmenter.

Ressources

Documents connexes :

- Calculateur du coût total de possession (TCO) d'AWS
- Classes de stockage Amazon S3
- Produits cloud
- Amazon EC2 Auto Scaling
- Migration des données dans le cloud
- AWS Snow Family

Vidéos connexes :

- AWS OpsHub for Snow Family

COÛT 6. Comment atteindre les objectifs de coût lorsque vous sélectionnez le type, la taille et le nombre de ressources ?

Veillez à choisir la taille et le nombre de ressources qui conviennent pour la tâche à accomplir. En choisissant le type, la taille et le nombre les plus rentables, vous réduisez le gaspillage.

Bonnes pratiques

- COST06-BP01 Réaliser une modélisation des coûts
- COST06-BP02 Sélectionner le type, la taille et le nombre de ressources en fonction des données
- COST06-BP03 Sélectionner automatiquement le type, la taille et le nombre de ressources en fonction des métriques

COST06-BP01 Réaliser une modélisation des coûts

Identifiez les exigences de l'organisation (telles que les besoins métier et les engagements existants) et réalisez une modélisation des coûts (coût global) de la charge de travail et de chacun de ses composants. Procédez à des évaluation de la charge de travail en fonction de diverses charges prévues et comparez les coûts. L'effort de modélisation doit refléter les avantages potentiels. Par exemple, le temps passé est proportionnel au coût des composants.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Effectuez une modélisation des coûts de votre charge de travail et de chacun de ses composants, afin de comprendre l'équilibre entre les ressources et de déterminer la taille correcte de chaque ressource dans la charge de travail, compte tenu d'un niveau de performance spécifique. Comprendre les considérations relatives aux coûts peut éclairer le cas d'utilisation et le processus de prise de décision de votre organisation lors de l'évaluation des résultats de réalisation de valeur pour le déploiement d'une charge de travail planifié.

Procédez à des évaluation de la charge de travail en fonction de diverses charges prévues et comparez les coûts. L'effort de modélisation doit refléter les avantages potentiels. Par exemple, le temps passé est proportionnel au coût des composants ou aux économies prévues. Pour connaître

les bonnes pratiques, consultez la section Vérifiez du pilier Efficacité des performances du cadre AWS Well-Architected.

Par exemple, afin de créer une modélisation des coûts d'une charge de travail comprenant des ressources de calcul, AWS Compute Optimizer peut vous aider à modéliser les coûts pour l'exécution des charges de travail. Il fournit des recommandations de dimensionnement des ressources de calcul basées sur l'utilisation historique. Assurez-vous que des agents CloudWatch sont déployés sur les instances Amazon EC2 pour collecter des métriques de mémoire qui vous offrent des recommandations plus précises au sein d'AWS Compute Optimizer. Il s'agit de la source de données idéale pour les ressources de calcul, car c'est un service gratuit qui utilise le machine learning pour faire plusieurs recommandations en fonction des niveaux de risque.

Vous pouvez utiliser plusieurs services avec des journaux personnalisés comme sources de données pour les opérations de dimensionnement d'autres services et composants de charge de travail, tels qu'AWS Trusted Advisor, Amazon CloudWatch et Amazon CloudWatch Logs. AWS Trusted Advisor vérifie et signale les ressources peu utilisées, ce qui peut aider à mieux les dimensionner ainsi qu'à créer une modélisation des coûts.

Voici des recommandations pour les données et métriques de modélisation des coûts :

- Le suivi doit refléter l'expérience utilisateur avec précision. Choisissez le niveau de précision correct pour la période et choisissez judicieusement le maximum ou le 99e centile au lieu de la moyenne.
- Sélectionnez la granularité appropriée pour la période d'analyse qui couvre tous les cycles de charge de travail. Par exemple, si une analyse de deux semaines est effectuée, vous pourriez négliger un cycle mensuel de forte utilisation, ce qui pourrait conduire à une sous-allocation.
- Choisissez les bons services AWS pour votre charge de travail prévue en prenant en compte vos engagements existants, les modèles de tarification sélectionnés pour vos autres charges de travail et votre capacité à innover rapidement et à vous concentrer sur votre valeur métier principale.

Étapes d'implémentation

- Réaliser une modélisation des coûts pour les ressources : déployez la charge de travail ou une démonstration de faisabilité dans un compte séparé avec les types et tailles de ressources spécifiques à tester. Exécutez la charge de travail avec les données de test et enregistrez les résultats, ainsi que les données de coût pour la période où le test a été effectué. Redéployez ensuite la charge de travail ou modifiez les types et les tailles des ressources et relancez le test. Incluez les frais de licence de tous les produits que vous pourriez utiliser avec ces ressources et

les frais d'opérations (main-d'œuvre ou ingénierie) estimés pour le déploiement et la gestion de ces ressources pendant la création de la modélisation des coûts. Envisagez une modélisation des coûts par période (heure, jour, mois, année ou trois ans).

Ressources

Documents connexes :

- AWS Auto Scaling
- Identifying Opportunities to Right Size (Identification des opportunités de dimensionnement)
- Fonctions d'Amazon CloudWatch
- Cost Optimization: Amazon EC2 Right Sizing (Optimisation des coûts : dimensionnement Amazon EC2)
- AWS Compute Optimizer
- Calculateur de tarification AWS

Exemples connexes :

- Perform a Data-Driven Cost Modelling (Réaliser une modélisation des coûts axée sur les données)
- Estimate the cost of planned AWS resource configurations (Estimer le coût des configurations de ressources AWS prévues)
- Choose the right AWS tools (Choisir les bons outils AWS)

COST06-BP02 Sélectionner le type, la taille et le nombre de ressources en fonction des données

Sélectionnez la taille ou le type de ressources en fonction des données relatives à la charge de travail et aux caractéristiques des ressources (par exemple, le calcul, la mémoire, le débit ou l'accès intensif en écriture). Cette sélection est généralement effectuée en utilisant une version précédente (sur site) de la charge de travail, en utilisant de la documentation ou d'autres sources d'information sur la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas établie : moyen

Directives d'implémentation

Amazon EC2 fournit une large sélection de types d'instances avec différents niveaux de capacité de processeur, de mémoire, de stockage et de réseau pour répondre à divers cas d'utilisation. Ces

types d'instances proposent différentes combinaisons de capacités de processeur, de mémoire, de stockage et de réseau pour vous permettre de sélectionner la combinaison de ressources adaptée à vos projets. Chaque type d'instance est disponible dans plusieurs tailles afin que vous puissiez ajuster vos ressources en fonction des exigences de votre charge de travail. Pour déterminer le type d'instance dont vous avez besoin, rassemblez des informations sur la configuration système requise de l'application ou du logiciel que vous envisagez d'exécuter sur votre instance. Ces détails doivent comprendre les éléments suivants :

- Système d'exploitation
- Nombre de cœurs de processeur
- Cœurs GPU
- Quantité de mémoire système (RAM)
- Type et espace de stockage
- Bande passante réseau nécessaire

Identifiez la finalité des besoins en calcul et l'instance requise, puis explorez les différentes familles d'instances Amazon EC2. Amazon propose les familles de types d'instances suivantes :

- Polyvalente
- Optimisée pour le calcul
- À mémoire optimisée
- Optimisée pour le stockage
- Calcul accéléré
- Optimisée pour HPC

Pour mieux comprendre les objectifs et les cas d'utilisation spécifiques qu'une famille d'instances Amazon EC2 donnée peut remplir, consultez Types d'instances AWS.

La collecte de la configuration système requise est essentielle pour sélectionner la famille d'instances et le type d'instance les mieux adaptés à vos besoins. Les noms de types d'instance sont composés du nom de famille et de la taille de l'instance. Par exemple, l'instance t2.micro appartient à la famille T2 et a une micro-taille.

Sélectionnez la taille ou le type de ressources en fonction des caractéristiques de la charge de travail et des ressources (calcul, mémoire, débit ou accès en écriture intensif, par exemple).

Cette sélection est généralement effectuée à l'aide d'une modélisation des coûts, d'une version antérieure de la charge de travail (version sur site, par exemple), d'une documentation ou d'autres sources d'informations sur la charge de travail (livres blancs ou solutions publiées). L'utilisation de calculateurs de prix ou d'outils de gestion des coûts AWS peut vous aider à prendre des décisions éclairées quant aux types, aux tailles et aux configurations des instances.

Étapes d'implémentation

- Sélectionner les ressources en fonction des données : utilisez vos données de modélisation des coûts pour sélectionner le niveau d'utilisation prévu de la charge de travail, et choisissez le type et la taille de la ressource spécifiée. Sur la base de vos données de modélisation des coûts, déterminez le nombre de processeurs virtuels, la mémoire totale (Gio), le volume de stockage de l'instance locale (Go), les volumes Amazon EBS et le niveau de performances du réseau, en tenant compte du taux de transfert de données requis pour l'instance. Effectuez toujours vos choix en vous appuyant sur des analyses détaillées et des données précises afin d'optimiser les performances tout en gérant efficacement les coûts.

Ressources

Documents connexes :

- Types d'instances AWS
- AWS Auto Scaling
- Fonctions d'Amazon CloudWatch
- Optimisation des coûts : dimensionnement EC2

Vidéos connexes :

- Selecting the right Amazon EC2 instance for your workloads
- Right Size Your Services

Exemples connexes :

- It just got easier to discover and compare Amazon EC2 instance types

COST06-BP03 Sélectionner automatiquement le type, la taille et le nombre de ressources en fonction des métriques

Utilisez les métriques de la charge de travail en cours pour sélectionner la taille et le type appropriés afin d'optimiser les coûts. Mettez en service de manière appropriée le débit, le dimensionnement et le stockage pour les services de calcul, de stockage, de données et de mise en réseau. Pour ce faire, utilisez une boucle de rétroaction, telle que la mise à l'échelle automatique ou du code personnalisé dans la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Créez une boucle de rétroaction qui utilise des métriques actives de la charge de travail en cours pour apporter des modifications à cette dernière. Vous pouvez utiliser un service géré, comme AWS Auto Scaling, qui sera configuré pour effectuer les opérations de dimensionnement à votre place. AWS fournit également des API et SDK, ainsi que des fonctionnalités qui permettent aux ressources d'être modifiées avec un minimum d'effort. Vous pouvez programmer une charge de travail pour arrêter et démarrer une instance Amazon EC2 afin de modifier la taille ou le type d'instance. De cette manière, vous tirez parti des avantages d'un redimensionnement tout en supprimant presque tous les coûts opérationnels nécessaires pour effectuer la modification.

Certains services AWS sont dotés d'une sélection intégrée et automatique de type ou de taille, par exemple Amazon Simple Storage Service Intelligent-Tiering. Amazon S3 Intelligent-Tiering déplace automatiquement vos données entre deux niveaux d'accès, accès fréquent et accès peu fréquent, en fonction de vos modèles d'utilisation.

Étapes d'implémentation

- Augmenter votre observabilité en configurant les métriques de la charge de travail : capturez les métriques clés de la charge de travail. Ces métriques donnent une indication de l'expérience client, comme le rendement de la charge de travail, et s'alignent sur les différences entre les types et les tailles de ressources, comme l'utilisation de l'UC et de la mémoire. Pour calculer les ressources, analysez les données de performances afin d'adapter la taille de vos instances Amazon EC2. Identifiez les instances inactives et celles qui sont sous-utilisées. Les métriques clés à rechercher sont l'utilisation de l'UC et l'utilisation de la mémoire (par exemple, 40 % d'utilisation de l'UC à 90 % du temps, comme expliqué dans Rightsizing with AWS Compute Optimizer and Memory Utilization Enabled (Dimensionnement avec activation d'AWS Compute Optimizer et de l'utilisation de la mémoire). Identifier les instances avec une utilisation maximale de l'UC et une utilisation de

la mémoire inférieure à 40 % sur une période de quatre semaines. Ce sont les instances dont la taille doit être adaptée pour réduire les coûts. Pour les ressources de stockage telles que Amazon S3, vous pouvez utiliser Amazon S3 Storage Lens, ce qui vous permet de voir 28 métriques dans différentes catégories au niveau du compartiment et 14 jours de données historiques dans le tableau de bord par défaut. Vous pouvez filtrer votre tableau de bord Amazon S3 Storage Lens par récapitulatif et optimisation des coûts ou événements pour analyser des métriques spécifiques.

- Afficher les recommandations de dimensionnement : utilisez les recommandations de dimensionnement dans AWS Compute Optimizer et l'outil de dimensionnement Amazon EC2 dans la console Gestion des coûts ou examinez le redimensionnement de vos ressources par AWS Trusted Advisor afin d'effectuer des ajustements sur votre charge de travail. Il est important d'utiliser les bons outils lors du dimensionnement des différentes ressources et de suivre les directives de dimensionnement, qu'il s'agisse d'une instance Amazon EC2, de classes de stockage AWS ou de types d'instance Amazon RDS. Pour les ressources de stockage, vous pouvez utiliser Amazon S3 Storage Lens qui vous donne une visibilité sur l'utilisation du stockage d'objets et les tendances d'activité en plus de faire des recommandations exploitables afin d'optimiser les coûts et d'appliquer les bonnes pratiques en matière de protection des données. À l'aide des recommandations contextuelles déduites par Amazon S3 Storage Lens de l'analyse des métriques au sein de votre organisation, vous pouvez prendre des mesures immédiates afin d'optimiser votre stockage.
- Sélectionner le type de ressources et les dimensionner automatiquement en fonction des métriques : à l'aide des métriques de la charge de travail, sélectionnez manuellement ou automatiquement vos ressources de charge de travail. Pour les ressources de calcul, la configuration d'AWS Auto Scaling ou la mise en œuvre du code dans votre application peut limiter l'effort requis si des changements fréquents sont nécessaires. De plus, l'application des modifications peut ainsi survenir de manière plus précoce qu'avec un processus manuel. Vous pouvez lancer et mettre à l'échelle automatiquement une flotte d'instances à la demande et d'instances Spot au sein d'un seul groupe Auto Scaling. En plus de recevoir des remises pour l'utilisation des instances Spot, vous pouvez utiliser des instances réservées ou un Savings Plan qui vous permettront de bénéficier de taux réduits par rapport à la tarification standard des instances à la demande. Tous ces facteurs combinés vous aident à optimiser vos économies de coûts pour les instances Amazon EC2 et à déterminer l'échelle et les performances souhaitées pour votre application. Vous pouvez également utiliser une stratégie de sélection de type d'instance basée sur les attributs (ABS) dans les Groupes Auto Scaling (ASG), afin d'exprimer vos exigences en matière d'instances sous forme d'un ensemble d'attributs, tels que le processeur virtuel, la mémoire et le stockage. Vous pouvez utiliser automatiquement les types d'instance de nouvelle génération lorsqu'ils sont disponibles et accéder à une plus large gamme de capacités avec les

instances Spot Amazon EC2. La flotte Amazon EC2 et Amazon EC2 Auto Scaling sélectionnent et lancent les instances qui correspondent aux attributs spécifiés, en éliminant le besoin de sélectionner manuellement les types d'instance. Pour les ressources de stockage, vous pouvez utiliser les fonctionnalités Amazon S3 Intelligent Tiering et Amazon EFS Infrequent Access. Celles-ci permettent de sélectionner des classes de stockage et de réaliser des économies sur les coûts de stockage de manière automatique chaque fois que les modèles d'accès aux données changent, le tout sans impacter les performances ou les frais généraux opérationnels.

Ressources

Documents connexes :

- AWS Auto Scaling
- Dimensionnement approprié AWS
- AWS Compute Optimizer
- Fonctions d'Amazon CloudWatch
- Configuration d'CloudWatch
- Publication des métriques personnalisées CloudWatch
- Premier pas avec Amazon EC2 Auto Scaling
- Amazon S3 Storage Lens
- Amazon S3 Intelligent-Tiering
- Amazon EFS Infrequent Access
- Launch an Amazon EC2 Instance Using the SDK (Lancement d'une instance EC2 à l'aide du SDK)

Vidéos connexes :

- Right Size Your Services

Exemples connexes :

- Attribute based Instance Type Selection for Auto Scaling for Amazon EC2 Fleet
- Optimizing Amazon Elastic Container Service for cost using scheduled scaling
- Mise à l'échelle prédictive pour Amazon EC2 Auto ScalingAmazon EC2 Auto Scaling

- Optimize Costs and Gain Visibility into Usage with Amazon S3 Storage Lens (Optimiser les coûts et gagner de la visibilité sur l'utilisation avec Amazon S3 Storage Lens)
- Well-Architected Labs: Rightsizing Recommendations (Level 100) (Ateliers Well-Architected : recommandations en matière de dimensionnement (niveau 100))
- Well-Architected Labs: Rightsizing with AWS Compute Optimizer and Memory Utilization Enabled (Level 200) (Ateliers Well-Architected : dimensionnement avec activation de Compute Optimizer et de l'utilisation de la mémoire (niveau 200))

COÛT 7. Comment utiliser les modèles de tarification pour réduire les coûts ?

Utilisez le modèle de tarification qui convient le mieux à vos ressources pour réduire les dépenses.

Bonnes pratiques

- COST07-BP01 Analyser le modèle de tarification
- COST07-BP02 Choix des régions en fonction du coût
- COST07-BP03 Sélectionner des accords avec des tiers avec des conditions rentables
- COST07-BP04 Mettre en œuvre des modèles de tarification pour tous les composants de cette charge de travail
- COST07-BP05 Analyser le modèle de tarification au niveau du compte de gestion

COST07-BP01 Analyser le modèle de tarification

Analysez chaque composant de la charge de travail. Déterminez si le composant et les ressources fonctionneront pendant des périodes prolongées (pour les réductions d'engagement), ou dynamiques et de courte durée (pour les instances Spot ou à la demande). Effectuez une analyse de la charge de travail à l'aide des recommandations des outils de gestion des coûts et appliquez des règles métier à ces recommandations pour obtenir des rendements élevés.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

AWS dispose de plusieurs modèles de tarification qui permettent de payer vos ressources de la manière la plus rentable qui répond aux besoins de votre organisation. Travaillez avec vos équipes pour déterminer le modèle de tarification le plus approprié. Souvent, votre modèle de tarification consiste en une combinaison de plusieurs options, en fonction de votre disponibilité.

Les instances à la demande vous permettent de payer la capacité de calcul ou de base de données à l'heure ou à la seconde (60 secondes minimum) en fonction des instances que vous utilisez, sans engagement à long terme ni avance sur paiement.

Savings Plans sont un modèle de tarification flexible qui offre des prix bas sur l'utilisation d'Amazon EC2, de Lambda et de AWS Fargate (Fargate). En échange, vous vous engagez à un volume d'utilisation régulier (mesuré en dollars par heure) sur un ou trois ans.

Les instances Spot sont un mécanisme de tarification Amazon EC2 vous permettant de demander de la capacité de calcul supplémentaire à un taux horaire réduit (jusqu'à 90 % du prix à la demande) sans engagement préalable.

Les instances réservées vous permettent d'obtenir jusqu'à 75 % de réduction en payant la capacité à l'avance. Pour plus de détails, consultez Optimizing costs with reservations (Optimisation des coûts avec les réservations).

Vous pouvez choisir d'inclure une Savings Plan pour les ressources associées aux environnements de production, de qualité et de développement. Par ailleurs, les ressources de l'environnement de test (sandbox) n'étant provisionnées qu'en cas de besoin, vous pouvez opter pour un modèle à la demande pour les ressources de cet environnement. Utilisez les instances Spot d'Amazon pour réduire les coûts Amazon EC2 ou utilisez les Savings Plans de calcul pour réduire les coûts d'Amazon EC2, de Fargate et de Lambda. L'outil de recommandations AWS Cost Explorer offre des possibilités de remises d'engagement avec les Savings Plans.

Si vous avez acheté des instances réservées pour Amazon EC2 par le passé ou si vous avez établi des pratiques de répartition des coûts au sein de votre organisation, vous pouvez continuer à utiliser des instances réservées Amazon EC2 pour le moment. Cependant, nous recommandons une stratégie visant à utiliser Savings Plans à l'avenir comme un mécanisme plus flexible de réduction des coûts. Vous pouvez actualiser les recommandations Savings Plans (SP) dans AWS Cost Management pour générer de nouvelles recommandations de Savings Plans à tout moment. Utilisez les instances réservées (RI) pour réduire les coûts de Amazon RDS, de Amazon Redshift, d'Amazon ElastiCache, et d'Amazon OpenSearch Service. Les Savings Plans et les instances réservées sont disponibles en trois options : paiement intégral à l'avance, avance sur le paiement et aucun paiement initial. Utilisez les recommandations fournies dans les recommandations d'achat de RI et SP AWS Cost Explorer.

Pour trouver des opportunités de charges de travail Spot, utilisez une vue horaire de votre utilisation globale et recherchez des périodes régulières d'évolution d'utilisation ou d'élasticité. Vous pouvez utiliser des instances Spot pour diverses applications flexibles et tolérantes aux pannes. Il s'agit par

exemple de serveurs Web sans état, de points de terminaison d'API, d'applications de big data et d'analytique, de charges de travail conteneurisées, de CI/CD et d'autres charges de travail flexibles.

Analysez vos instances Amazon EC2 et Amazon RDS pour déterminer si elles peuvent être désactivées lorsque vous ne les utilisez pas (après les heures de travail et le week-end). Cette approche vous permettra de réduire les coûts de 70 % ou plus par rapport à leur utilisation 24 heures sur 24 et 7 jours sur 7. Si vous avez des clusters Amazon Redshift qui ne doivent être disponibles qu'à des moments précis, vous pouvez mettre le cluster en pause et le reprendre plus tard. Lorsque le cluster Amazon Redshift ou l'instance Amazon EC2 et Amazon RDS est arrêté(e), la facturation du calcul s'arrête et seuls les frais de stockage s'appliquent.

Notez que les réservations de capacité à la demande (ODCR) ne constituent pas une réduction de prix. Les réservations de capacité sont facturées au tarif à la demande équivalent, que vous exécutiez des instances en capacité réservée ou non. Pensez à cette option lorsque vous devez fournir une capacité suffisante pour les ressources que vous prévoyez d'exploiter. Les ODCR ne doivent pas nécessairement être liées à des engagements à long terme, puisqu'elles peuvent être annulées lorsque vous n'en avez plus besoin. Cependant, elles peuvent également bénéficier des réductions offertes par les Savings Plans et les instances réservées.

Étapes d'implémentation

- Analysez l'élasticité de la charge de travail : en utilisant la granularité horaire dans Cost Explorer ou dans un tableau de bord personnalisé, analysez l'élasticité de votre charge de travail. Recherchez les modifications régulières du nombre d'instances en cours d'exécution. Les instances de courte durée sont de bonnes candidates pour les instances Spot ou les parcs d'instances Spot.
  - Atelier Well-Architected : Cost Explorer
  - Atelier Well-Architected : visualisation des coûts
- Passer en revue la tarification des contrats existants : passez en revue les contrats ou engagements actuels pour les besoins à long terme. Analysez ce dont vous disposez actuellement et le degré d'utilisation de ces engagements. Tirez parti des remises contractuelles ou des accords d'entreprise préexistants. Les accords d'entreprise donnent aux clients la possibilité de personnaliser les accords qui répondent le mieux à leurs besoins. Pour les engagements à long terme, envisagez des réductions de prix réservées, des instances réservées ou des Savings Plans pour le type d'instance, la famille d'instances, la Région AWS et les zones de disponibilité spécifiques.
- Analyser les réductions pour engagement : en utilisant Cost Explorer dans votre compte, examinez les recommandations de Savings Plans et d'instances réservées. Pour mettre en œuvre les

recommandations correctes avec les réductions et les risques requis, suivez les recommandations des ateliers Well-Architected.

Ressources

Documents connexes :

- Accessing Reserved Instance recommendations (Accès aux recommandations des instances réservées)
- Options d'achat d'instance
- AWS Enterprise

Vidéos connexes :

- Save up to 90% and run production workloads on Spot (Économisez jusqu'à 90 % et exécutez des charges de travail de production sur des instances Spot)

Exemples connexes :

- Atelier Well-Architected : Cost Explorer
- Atelier Well-Architected : visualisation des coûts
- Atelier Well-Architected : modèles de tarification

COST07-BP02 Choix des régions en fonction du coût

La tarification des ressources peut être différente dans chaque région. Identifiez les différences de coûts entre régions et déployez uniquement dans les régions aux coûts plus élevés afin de répondre aux exigences de latence, de résidence des données et de souveraineté des données. En intégrant le coût de la région, vous payez le prix global le plus bas pour cette charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

L'infrastructure du AWS Cloud est mondiale, hébergée dans plusieurs sites dans le mondeet créée autour de Régions AWS, de zones de disponibilité, de zones locales, d'AWS Outposts et de zones Wavelength. Une région est un emplacement physique dans le monde et chaque région représente

une zone géographique distincte dans laquelle AWS a plusieurs zones de disponibilité. Les zones de disponibilité, qui sont plusieurs emplacements isolés dans chaque région, consistent en un ou plusieurs centres de données discrets, chacun disposant d'une alimentation, d'un réseau et d'une connectivité redondants.

Chaque Région AWS fonctionne selon les conditions du marché local et la tarification des ressources est différente dans chaque région compte tenu des différences de coûts des terrains, de la fibre, de l'électricité et des taxes, par exemple. Choisissez une région spécifique pour exploiter un composant ou l'ensemble de votre solution afin que vous puissiez fonctionner au prix le plus bas possible au niveau mondial. Utilisez le calculateur AWS pour estimer les coûts de votre charge de travail dans différentes régions en cherchant des services par type d'emplacement (région, zone Wavelength et zone locale) et par région.

Lorsque vous concevez vos solutions, une bonne pratique consiste à placer les ressources de calcul au plus près de l'utilisateur pour fournir une latence plus faible et une importante souveraineté des données. Sélectionner le lieu géographique en fonction de votre entreprise, votre confidentialité des données, vos performances et vos exigences en matière de sécurité. Pour les applications avec utilisateurs finaux internationaux, utilisez plusieurs emplacements.

Utilisez les régions qui offrent des services AWS à plus bas prix pour déployer vos charges de travail si vous n'avez aucune obligation en termes de confidentialité des données, de sécurité et d'exigences au niveau de l'entreprise. Par exemple, si votre région par défaut est ap-southeasth-2 (Sydney) et qu'il n'existe aucune restriction (par exemple, confidentialité des données ou sécurité) liée à l'utilisation d'autres régions, le déploiement d'instances Amazon EC2 non critiques (développement et test) dans la région north-east-1 (Virginie du Nord) vous coûtera moins d'argent.

	Conformité	Latence	Coût	Services/Fonctionnalités
Région 1	✓	15 ms	$$	✓
Région 2	✓	20 ms	$$$	X
Région 3	✓	80 ms	$	✓
Région 4	✓	15 ms	$$	✓
Région 5	✓	20 ms	$$$	X
Région 6	✓	15 ms	$	✓
Région 7	✓	80 ms	$	✓
Région 8	✓	15 ms	$	X

Tableau de la matrice des caractéristiques des régions

Le tableau matriciel précédent nous montre que la région 4 est la meilleure option pour ce scénario donné car la latence y est faible comparé aux autres régions, le service y est disponible et il s'agit de la région la moins chère.

Étapes d'implémentation

- Passez en revue la tarification de la Région AWS : analysez les coûts de charge de travail dans la région actuelle. En commençant par les coûts les plus élevés par service et par type d'utilisation, calculez les coûts dans les autres régions disponibles. Si l'économie prévue est supérieure au coût du déplacement du composant ou de la charge de travail, migrez vers la nouvelle région.
- Vérifiez les exigences pour les déploiements sur plusieurs régions : analysez les exigences et les obligations de votre entreprise (confidentialité des données, sécurité ou performances) pour découvrir s'il existe des restrictions vous empêchant d'utiliser plusieurs régions. Si aucune obligation ne vous restreint à utiliser une seule région, alors utilisez-en plusieurs.
- Analysez le transfert de données requis : tenez compte des coûts de transfert de données lors de la sélection des régions. Rapprochez vos données de votre client et des ressources. Sélectionnez des Régions AWS moins coûteuses où les données circulent et où il existe un transfert de données minimum. En fonction des besoins de votre entreprise en matière de transfert de données, vous pouvez utiliser Amazon CloudFront, AWS PrivateLink, AWS Direct Connectet AWS Virtual Private

Network pour réduire vos coûts de mise en réseau, améliorer les performances et renforcer la sécurité.

Ressources

Documents connexes :

- Accès aux recommandations des instances réservées
- Tarification Amazon EC2
- Options d'achat d'instance
- Tableau des régions

Vidéos connexes :

- Économisez jusqu'à 90 % et exécutez des charges de travail de production sur des instances Spot

Exemples connexes :

- Présentation des coûts de transfert des données pour les architectures courantes
- Considérations des coûts pour les déploiements mondiaux
- Éléments à prendre en compte lors de la sélection d'une région pour vos charges de travail
- Ateliers Well-Architected : restreindre l'utilisation d'un service par région (niveau 200)

COST07-BP03 Sélectionner des accords avec des tiers avec des conditions rentables

Les accords et conditions rentables garantissent que le coût de ces services évolue en fonction des avantages qu'ils offrent. Choisissez des accords et une tarification qui évoluent lorsqu'ils apportent des avantages supplémentaires à votre organisation.

Niveau de risque exposé si cette bonne pratique n'est pas établie: moyen

Directives d'implémentation

De nombreux produits du marché peuvent vous aider à gérer les coûts de vos environnements cloud. Ils peuvent présenter des différences en termes de fonctionnalités selon les exigences des clients. Certains privilégieront la gouvernance ou la visibilité des coûts et d'autres l'optimisation des coûts, par exemple. L'un des facteurs clés pour une optimisation et une gouvernance efficaces des coûts

consiste à utiliser le bon outil avec les bonnes fonctionnalités et le bon modèle de tarification. Ces produits ont des modèles de tarification différents. Certains correspondent à un certain pourcentage de votre facture mensuelle, et d'autres à un pourcentage des économies réalisées. Idéalement, vous ne devriez payer que ce dont vous avez besoin.

Lorsque vous utilisez des solutions ou des services tiers dans le cloud, il est important que les structures de tarification soient alignées sur les résultats souhaités. La tarification doit évoluer en fonction des résultats et de la valeur qu'elle fournit. Par exemple, dans le cas d'un logiciel facturé à un pourcentage des économies réalisées, plus vous économisez (résultat), plus le logiciel est cher. Les contrats de licence qui prévoient un paiement proportionnel à vos dépenses ne sont pas toujours dans votre intérêt pour optimiser les coûts. Toutefois, si l'éditeur offre des avantages clairs pour toutes les parties de votre facture, ces frais progressifs peuvent être justifiés.

Par exemple, une solution qui fournit des recommandations pour Amazon EC2 moyennant un pourcentage de votre facture totale peut devenir chère si vous utilisez d'autres services qui n'apportent aucun avantage. Prenons également l'exemple d'un service géré facturé à un pourcentage du coût des ressources gérées. Une instance de plus grande taille ne nécessite pas nécessairement plus d'efforts de gestion, mais elle peut être facturée plus cher. Vérifiez que ces accords de tarification de service incluent un programme ou des fonctions d'optimisation des coûts dans leur service afin d'améliorer leur rentabilité.

Les clients peuvent trouver ces produits du marché plus avancés ou plus faciles à utiliser. Vous devez prendre en compte le coût de ces produits et réfléchir aux possibilités d'optimisation des coûts à long terme.

Étapes d'implémentation

- Analyser les accords et conditions des tiers : passez en revue la tarification des accords avec des tiers. Effectuez une modélisation pour différents niveaux d'utilisation et tenez compte des nouveaux coûts tels que l'utilisation de nouveaux services ou l'augmentation des services actuels en raison de la croissance de la charge de travail. Déterminez si les coûts supplémentaires apportent les avantages requis à votre entreprise.

Ressources

Documents connexes :

- Accessing Reserved Instance recommendations
- Options d'achat d'instance

Vidéos connexes :

- Save up to 90% and run production workloads on Spot

COST07-BP04 Mettre en œuvre des modèles de tarification pour tous les composants de cette charge de travail

Les ressources fonctionnant en permanence doivent utiliser des capacités réservées telles que des Savings Plans ou des instances réservées. La capacité à court terme est configurée pour utiliser des instances Spot ou un parc d'instances Spot. Les instances à la demande ne sont utilisées que pour les charges de travail de courte durée qui ne peuvent pas être interrompues et qui ne durent pas assez longtemps pour la capacité réservée, entre 25 et 75 % de la période, selon le type de ressource.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

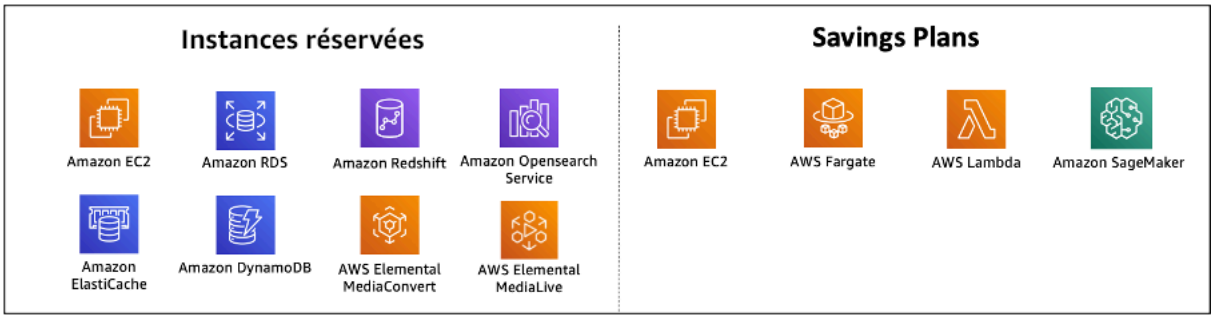
Directives d'implémentation

Pour améliorer votre rentabilité, AWS recommande plusieurs engagements en fonction de votre consommation passée. Vous pouvez utiliser ces recommandations pour comprendre les économies que vous pouvez réaliser et comment l'engagement sera utilisé. Vous pouvez utiliser ces services à la demande, ponctuellement ou vous engager pendant un certain temps et réduire vos coûts à la demande grâce aux instances réservées (RI) et aux Savings Plans (SP). Vous devez non seulement comprendre les composants de chaque charge de travail et les différents services AWS, mais également les remises sur engagement, les options d'achat et les instances Spot de ces services pour optimiser votre charge de travail.

Tenez compte des exigences des composants de votre charge de travail et maîtrisez les différents modèles de tarification de ces services. Définissez les besoins de disponibilité de ces composants. Déterminez s'il existe plusieurs ressources indépendantes qui remplissent la fonction dans la charge de travail, et quelles sont les exigences de la charge de travail au fil du temps. Comparez le coût des ressources à l'aide du modèle de tarification à la demande par défaut et à celui des autres modèles applicables. Tenez compte de toute modification éventuelle des ressources ou des éléments de la charge de travail.

Par exemple, examinons cette architecture d'application Web sur AWS. Cet exemple de charge de travail comprend plusieurs services AWS, comme Amazon Route 53, AWS WAF, Amazon CloudFront, des instances Amazon EC2, des instances Amazon RDS, des équilibreurs de charge, une capacité de stockage Amazon S3 et Amazon Elastic File System (Amazon EFS). Vous devez

passer en revue chacun de ces services et identifier les opportunités de réduction de coûts des différents modèles de tarification. Certains d'entre eux peuvent être éligibles à des RI ou à des SP, tandis que d'autres peuvent être disponibles uniquement à la demande. Comme le montre l'image suivante, des engagements peuvent être pris sur certains services AWS à l'aide de RI ou de SP.



Engagement sur des services AWS à l'aide d'instances réservées et de Savings Plans

Étapes d'implémentation

- Mettre en œuvre des modèles de tarification : utilisez les résultats de votre analyse pour acheter des Savings Plans, des instances réservées ou mettre en œuvre des instances Spot. S'il s'agit de votre premier achat avec engagement, choisissez les cinq ou dix meilleures recommandations de la liste, puis surveillez et analysez les résultats au cours des deux prochains mois. AWS Cost Management Console vous guide tout au long du processus. Consultez les recommandations RI ou SP de la console, personnalisez les recommandations (type, paiement et durée), passez en revue l'engagement horaire (par exemple, 20 USD/heure), puis ajoutez-les au panier. Les remises s'appliquent automatiquement à l'utilisation éligible. Achetez régulièrement un petit nombre d'engagements avec remise, par exemple toutes les deux semaines ou tous les mois. Mettez en œuvre des instances Spot pour les charges de travail qui peuvent être interrompues ou qui sont sans état. Enfin, sélectionnez des instances Amazon EC2 à la demande et allouez les ressources aux besoins restants.
- Cycle de vérification de la charge de travail : mettez en œuvre un cycle de vérification de la charge de travail, qui analyse spécifiquement la couverture du modèle de tarification. Une fois que la charge de travail dispose de la couverture requise, achetez des engagements avec remise supplémentaires régulièrement (tous les deux ou trois mois) ou en fonction de l'évolution de la consommation de votre organisation.

Ressources

Documents connexes :

- Understanding your Savings Plans recommendations
- Accessing Reserved Instance recommendations (Accès aux recommandations des instances réservées)
- Comment acheter des instances réservées
- Options d'achat d'instance
- Instances Spot
- Modèles de réservation pour d'autres services AWS
- Savings Plans Supported Services

Vidéos connexes :

- Save up to 90% and run production workloads on Spot

Exemples connexes :

- Que dois-je prendre en compte avant d'acheter des Savings Plans ?
- Comment puis-je utiliser l'Cost Explorer pour analyser les dépenses et l'utilisation ?

COST07-BP05 Analyser le modèle de tarification au niveau du compte de gestion

Vérifiez les outils de facturation et de gestion des coûts et consultez les remises recommandées avec les engagements et les réservations pour mener une analyse régulière au niveau du compte de gestion.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

La modélisation régulière des coûts vous aide à mettre en œuvre les possibilités d'optimisation sur plusieurs charges de travail. Par exemple, si plusieurs charges de travail utilisent des instances à la demande au niveau agrégé, le risque de changement est moindre, et la mise en œuvre d'une réduction basée sur un engagement permet d'obtenir un coût global plus faible. Il est recommandé d'effectuer les analyses selon des cycles réguliers de deux semaines à un mois. Cela vous permet de faire de petits achats d'ajustement, de sorte que la couverture de vos modèles de tarification continue à évoluer en fonction de l'évolution de vos charges de travail et de leurs composants.

Utilisez l'outil de recommandations AWS Cost Explorer pour identifier des opportunités de remises sur engagement dans votre compte de gestion. Les recommandations au niveau du compte de gestion sont calculées en tenant compte de l'utilisation de tous les comptes de votre organisation AWS qui ont des instances réservées (RI) ou des (SP) Savings Plans. Elles sont également calculées lorsque le partage des remises est activé afin de recommander un engagement qui maximise les économies sur tous les comptes.

Bien que les achats au niveau du compte de gestion permettent de réaliser des économies maximales dans de nombreux cas, il peut arriver que vous envisagiez d'acheter des SP au niveau du compte associé, par exemple lorsque vous souhaitez que les remises s'appliquent d'abord à l'utilisation sur ce compte associé en particulier. Les recommandations pour les comptes des membres sont calculées au niveau du compte individuel, afin de maximiser les économies pour chaque compte isolé. Si votre compte contient à la fois des engagements RI et SP, ils seront appliqués dans cet ordre :

1. RI zonale
2. RI standard
3. RI convertible
4. Instance Savings Plan
5. Compute Savings Plan

Si vous achetez un SP au niveau du compte de gestion, les économies seront appliquées en fonction du pourcentage de remise du plus élevé au plus bas. Au niveau des comptes de gestion, les SP examinent tous les comptes liés et appliquent les économies là où la remise sera la plus élevée. Si vous souhaitez limiter les domaines dans lesquels les économies sont appliquées, vous pouvez souscrire à un Savings Plan au niveau du compte associé. Dans ce cas, chaque fois que ce compte utilisera des services de calcul éligibles, la réduction sera appliquée en premier sur ce compte. Lorsque le compte n'exécute pas de services informatiques éligibles, la réduction est partagée entre les autres comptes liés sous le même compte de gestion. Le partage des remises est activé par défaut, mais il peut être désactivé si nécessaire.

Dans une famille de facturation consolidée, les Savings Plans s'appliquent d'abord à l'utilisation du compte du propriétaire, puis à l'utilisation des autres comptes. Cela se produit uniquement si le partage est activé. Vos Savings Plans sont d'abord appliqués à votre pourcentage d'économies le plus élevé. S'il y a plusieurs utilisations avec des pourcentages d'économie équivalents, les Savings Plans sont appliqués à la première utilisation avec le taux de Savings Plans le plus bas. Les Savings Plans continuent à s'appliquer jusqu'à ce qu'il n'y ait plus d'utilisations restantes ou que votre

engagement soit épuisé. Toute utilisation restante est facturée aux taux à la demande. Vous pouvez actualiser les recommandations de Savings Plans dans la gestion des coûts AWS afin de générer de nouvelles recommandations Savings Plans à tout moment.

Après avoir analysé la flexibilité des instances, choisissez un niveau d'engagement selon les recommandations. Créez une modélisation des coûts en analysant les coûts à court terme de la charge de travail avec différentes options de ressources potentielles, en analysant les modèles de tarification AWS et en les alignant sur vos exigences métier pour mettre en lumière le coût total de possession et les optimisation des coûts d'optimisation des coûts.

Étapes d'implémentation

Analyser les réductions d’engagement: en utilisant Cost Explorer dans votre compte, examinez les recommandations de Savings Plans et d'instances réservées. Assurez-vous de comprendre les recommandations du Savings Plan et estimez vos dépenses et les économies que vous réalisez chaque mois. Examinez les recommandations au niveau du compte de gestion, qui sont calculées en tenant compte de l'utilisation de tous les comptes membres de votre organisation AWS qui comportent des IR ou des Savings Plans pour lesquelles le partage des remises est activé. Ainsi, vous réaliserez un maximum d'économies sur tous les comptes. Vous pouvez confirmer que vous avez mis en œuvre les bonnes recommandations avec les remises et les risques requis en suivant les ateliers Well-Architected.

Ressources

Documents connexes :

- Comment fonctionne la tarification AWS ?
- Options d'achat d'instance
- Présentation du Saving Plan
- Recommandations en matière de Saving Plan
- Accès aux recommandations des instances réservées
- Comprendre les recommandations de vos Savings Plans
- Comment les Savings Plans s'appliquent-ils à votre utilisation AWS ?
- Savings Plans avec facturation consolidée
- Activation des instances réservées partagées et des remises Savings Plans

Vidéos connexes :

- Économisez jusqu'à 90 % et exécutez des charges de travail de production sur des instances Spot

Exemples connexes :

- Atelier Well-Architected AWS : modèles de tarification (niveau 200)
- Ateliers Well-Architected AWS : analyse des modèles de tarification (niveau 200)
- Que dois-je prendre en considération avant de souscrire un Savings Plan ?
- Comment puis-je utiliser le déploiement des Savings Plans pour réduire le risque d'engagement ?
- Quand utiliser les instances Spot ?

COÛT 8. Comment planifiez-vous les frais de transfert de données ?

Veillez à planifier et à surveiller les frais de transfert de données afin de pouvoir prendre des décisions architecturales pour minimiser les coûts. Une modification architecturale minime, mais efficace, peut réduire de façon spectaculaire vos coûts d'exploitation.

Bonnes pratiques

- COST08-BP01 Modéliser le transfert de données
- COST08-BP02 Sélectionner des composants pour optimiser les coûts de transfert de données
- COST08-BP03 Mettre en œuvre des services pour réduire les coûts de transfert de données

COST08-BP01 Modéliser le transfert de données

Recueillez les exigences de l'organisation et procédez à la modélisation du transfert de données de la charge de travail et de chacun de ses composants. Vous identifiez ainsi le coût le plus bas pour ses besoins de transfert de données actuels.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

L'habitude de concevoir une architecture utilisant des centres de données sur site ou le manque de connaissances peut conduire à négliger les frais de transfert de données lors de la conception d'une solution dans le cloud. Les frais de transfert de données d'AWS sont déterminés par la source, la destination et le volume du trafic. La prise en compte de ces frais lors de la phase de conception peut permettre de réaliser des économies. Pour estimer avec précision le coût total de possession (TCO),

il est très important de comprendre où a lieu le transfert de données dans votre charge de travail, son coût et les avantages qui en découlent. Cela vous permet de prendre une décision avisée pour modifier ou accepter la décision architecturale. Par exemple, vous pouvez avoir une configuration multi-zones de disponibilité dans laquelle vous répliquez les données entre les zones de disponibilité.

Vous modélisez les composants des services qui transfèrent les données de votre charge de travail, et décidez qu'il s'agit d'un coût acceptable (semblable au paiement de la capacité calcul et du stockage dans les deux zones de disponibilité) pour atteindre la fiabilité et la résilience requises. Modélisez les coûts sur différents niveaux d'utilisation. L'utilisation de la charge de travail peut changer dans le temps, et différents services peuvent être plus rentables à différents niveaux.

Lorsque vous modélisez votre transfert de données, réfléchissez au volume de données ingérées et à leur provenance. Tenez également compte de la quantité de données traitées et de la capacité de stockage ou de calcul requise. Lors de la modélisation, suivez les bonnes pratiques de mise en réseau pour l'architecture de votre charge de travail afin d'optimiser vos coûts potentiels de transfert de données.

AWS Pricing Calculator peut vous aider à estimer le coût des services AWS et du transfert de données prévu. Si une charge de travail est déjà en cours d'exécution (à des fins de test ou dans un environnement de préproduction), utilisez l'AWS Cost Explorer ou le AWS Cost and Usage Report(CUR) pour comprendre et modéliser les coûts de vos transferts de données. Configurez une preuve de concept (PoC) ou testez votre charge de travail et exécutez un test avec une charge simulée réaliste. Vous pouvez modéliser vos coûts selon différentes demandes de charge de travail.

Étapes d'implémentation

- Identifier les exigences : quels sont l'objectif principal et les besoins de l'entreprise pour le transfert de données prévu entre la source et la destination ? Quel est le résultat commercial attendu ? Définissez les besoins de l'entreprise et les résultats attendus.
- Identifier la source et la destination : quelles sont la source et la destination des données transférées (Régions AWS, services AWS, Internet, etc.) ?
  - Transfert de données dans une Région AWS
  - Transfert de données entre Régions AWS
  - Transfert de données à destination d'Internet
- Identifier les classifications des données : quelle est la classification des données de ce transfert ? De quel type de données s'agit-il ? Quelle est la taille des données ? À quelle fréquence les données doivent-elles être transférées ? Les données sont-elles sensibles ?

- Identifier les services ou les outils AWS à utiliser : quels sont les services AWS utilisés pour ce transfert de données ? Est-il possible d'utiliser un service déjà provisionné pour une autre charge de travail ?
- Calculer les coûts de transfert de données : utilisez la tarification AWSde la modélisation du transfert de données que vous avez précédemment créée pour calculer les coûts de transfert de données de la charge de travail. Calculez les coûts de transfert de données à différents niveaux d'utilisation, tant pour l'augmentation que pour la réduction de la charge de travail. Lorsqu'il existe plusieurs options pour l'architecture de la charge de travail, calculez le coût de chaque option à titre de comparaison.
- Relier les coûts aux résultats : pour chaque coût de transfert de données engagé, précisez le résultat qu'il permet d'obtenir pour la charge de travail. S'il s'agit d'un transfert entre composants, ce peut être pour le découplage. S'il s'agit d'un transfert entre zones de disponibilité, ce peut être pour la redondance.
- Créer une modélisation du transfert de données : après avoir collecté toutes les informations, créez une base conceptuelle de modélisation du transfert de données pour plusieurs cas d'utilisation et charges de travail.

Ressources

Documents connexes :

- Solutions de mise en cache AWS
- Tarification AWS
- Tarification d'Amazon EC2 à la demande
- Tarification d'Amazon VPC
- Understanding data transfer charges

Vidéos connexes :

- Monitoring and Optimizing Your Data Transfer Costs
- S3 Transfer Acceleration

Exemples connexes :

- Overview of Data Transfer Costs for Common Architectures

- AWS Prescriptive Guidance for Networking

COST08-BP02 Sélectionner des composants pour optimiser les coûts de transfert de données

Tous les composants sont sélectionnés, et l'architecture est conçue pour réduire les coûts de transfert des données. Cela inclut l'utilisation de composants tels que l'optimisation WAN et les configurations Multi-AZ

Niveau de risque exposé si cette bonne pratique n'est pas établie: moyen

Directives d'implémentation

L'architecture pour le transfert de données minimise les coûts de transfert de données. Cela peut impliquer l'utilisation de réseaux de diffusion de contenu pour localiser les données plus près des utilisateurs, ou l'utilisation de liaisons réseau dédiées depuis vos sites vers AWS. Vous pouvez également utiliser l'optimisation WAN et l'optimisation des applications pour réduire la quantité de données transférée entre les composants.

Lors du transfert de données vers ou au sein d'AWS Cloud, il est essentiel de connaître la destination en fonction des différents cas d'utilisation, de la nature des données et des ressources réseau disponibles afin de sélectionner les bons services AWS pour optimiser le transfert de données. AWS propose une gamme de services de transfert de données adaptés à divers besoins en matière de migration de données. Sélectionnez les bonnes options de stockage et de transfert de données en fonction des besoins de votre organisation.

Lorsque vous planifiez ou passez en revue l'architecture de votre charge de travail, tenez compte des points suivants :

- Utiliser des points de terminaison de VPC dans AWS : les points de terminaison de VPC autorisent les connexions privées entre votre VPC et les services AWS pris en charge. Cela vous évite d'utiliser l'Internet public, qui peut engendrer des coûts de transfert de données.
- Utiliser une passerelle NAT : utilisez une passerelle NAT de manière à ce que les instances d'un sous-réseau privé puissent se connecter à Internet ou aux services en dehors de votre VPC. Vérifiez si les ressources situées derrière la passerelle NAT qui envoient le plus de trafic se trouvent dans la même zone de disponibilité que la passerelle NAT. Si ce n'est pas le cas, créez des passerelles NAT dans la zone de disponibilité de la ressource pour réduire les frais de transfert de données entre zones de disponibilité.

- Utiliser AWS Direct Connect : AWS Direct Connect contourne l'Internet public et établit une connexion privée directe entre votre réseau sur site et AWS. Cela peut être plus rentable et plus cohérent que de transférer de gros volumes de données sur Internet.
- Éviter de transférer des données au-delà des frontières régionales : les transferts de données entre Régions AWS (d'une région à l'autre) entraînent généralement des frais. La décision de poursuivre dans une voie multirégionale doit être mûrement réfléchie. Pour plus d'informations, consultez Scénarios multirégion.
- Surveiller le transfert de données : utilisez Amazon CloudWatch et les journaux de flux VPC pour obtenir des informations sur votre transfert de données et l'utilisation du réseau. Analysez les informations capturées sur le trafic réseau dans vos VPC, telles que l'adresse IP ou la plage d'adresses IP à destination et en provenance des interfaces réseau.
- Analyser l'utilisation de votre réseau : utilisez des outils de mesure et de reporting comme AWS Cost Explorer, CUDOS Dashboards ou CloudWatch pour comprendre le coût de transfert de données de votre charge de travail.

Étapes d'implémentation

- Sélectionner les composants pour le transfert de données : à l'aide de la modélisation de transfert de données expliquée dans COST08-BP01 Modéliser le transfert de données, concentrez-vous sur les coûts de transfert de données les plus élevés ou sur ce qu'ils seraient si l'utilisation de la charge de travail changeait. Recherchez d'autres architectures ou des composants supplémentaires qui suppriment ou réduisent la nécessité d'un transfert de données, ou en diminuent le coût.

Ressources

Bonnes pratiques associées :

- COST08-BP01 Modéliser le transfert de données
- COST08-BP03 Mettre en œuvre des services pour réduire les coûts de transfert de données

Documents connexes :

- Migration des données dans le cloud
- Solutions de mise en cache AWS
- Deliver content faster with Amazon CloudFront

Exemples connexes :

- Overview of Data Transfer Costs for Common Architectures
- AWS Network Optimization Tips
- Optimize performance and reduce costs for network analytics with VPC Flow Logs in Apache Parquet format

COST08-BP03 Mettre en œuvre des services pour réduire les coûts de transfert de données

Mettez en œuvre des services pour réduire le transfert de données. Par exemple, utilisez des emplacements périphériques ou des réseaux de diffusion de contenu (CDN) pour fournir du contenu aux utilisateurs finaux, construisez des couches de mise en cache devant vos serveurs d'application ou vos bases de données, et utilisez des connexions réseau dédiées au lieu de VPN pour la connectivité au cloud.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Il existe différents services AWS qui peuvent vous aider à optimiser l'utilisation de votre réseau pour le transfert de données. En fonction des composants de votre charge de travail, du type et de l'architecture du cloud, ces services peuvent vous aider à la compression, à la mise en cache, ainsi qu'au partage et à la distribution de votre trafic sur le cloud.

- Amazon CloudFront est un réseau mondial de diffusion de contenu qui fournit des données avec une faible latence et des vitesses de transfert élevées. Il place les données en cache au niveau des emplacements périphériques dans le monde entier, ce qui réduit la charge sur vos ressources. En utilisant CloudFront, vous pouvez réduire les tâches d'administration liées à la diffusion du contenu à un grand nombre d'utilisateurs dans le monde entier, avec une latence minimale. La version forfait sécurité peut vous aider à économiser jusqu'à 30 % sur votre consommation CloudFront si vous prévoyez d'augmenter votre utilisation au fil du temps.
- AWS Direct Connect facilite la mise en place d'une connexion réseau dédiée depuis vos sites vers AWS. Cela peut réduire les coûts de réseau, augmenter la bande passante et fournir une expérience réseau plus constante que les connexions Internet.
- AWS VPN permet d'établir une connexion sécurisée et privée entre votre réseau privé et le réseau mondial AWS. Il est idéal pour les petits bureaux ou les partenaires commerciaux, car il fournit une connectivité simplifiée, et il s'agit d'un service entièrement géré et élastique.

- Points de terminaison d'un VPC permettent la connectivité entre les services AWS sur une mise en réseau privée et peuvent être utilisés pour réduire les coûts de transfert de données publiques et de Passerelle NAT coûts. Les points de terminaison de VPC de passerelle n'ont pas de frais horaires et prennent en charge Amazon S3 et Amazon DynamoDB. Les points de terminaison de VPC d'interface sont fournis par AWS PrivateLink et ont un tarif horaire et un coût d'utilisation par Go.
- Passerelles NAT permettent une mise à l'échelle et une gestion intégrées, ce qui réduit les coûts par rapport à une instance NAT autonome. Placez les passerelles NAT dans les mêmes zones de disponibilité que les instances à fort trafic et envisagez d'utiliser des points de terminaison VPC pour les instances qui ont besoin d'accéder à Amazon S3 ou à Amazon DynamoDB afin de réduire les coûts de transfert et de traitement des données.
- Utilisez AWS Snow Family dotés de ressources informatiques pour collecter et traiter des données en périphérie. Les appareils AWS Snow Family (Snowcone, Snowball et Snowmobile) vous permettent de déplacer des pétaoctets de données vers le AWS Cloud de manière rentable et hors ligne.

Étapes d'implémentation

- Mettre en œuvre des services : Sélectionnez les services réseau AWS applicables en fonction du type de charge de travail de votre service, en utilisant la modélisation du transfert de données et en examinant les journaux de flux VPC. Regardez où se situent les coûts les plus élevés et les flux les plus importants. Examinez les services AWS et évaluez s'il existe un service qui réduit ou supprime le transfert, en particulier, la mise en réseau et la diffusion de contenu. Recherchez également les services de mise en cache où il existe une répétition d'accès aux données, ou de grands volumes de données.

Ressources

Documents connexes :

- AWS Direct Connect
- Explorer les produits AWS
- Solutions de mise en cache AWS
- Amazon CloudFront
- AWS Snow Family
- Forfait sécurité Amazon CloudFront

Vidéos connexes :

- Monitoring and Optimizing Your Data Transfer Costs
- Série sur l'optimisation des coûts AWS : CloudFront
- Comment réduire les frais de transfert de données pour ma passerelle NAT ?

Exemples connexes :

- Comment rétrofacturer des services partagés : un exemple AWS Transit Gateway
- Bien comprendre les détails du transfert de données AWS à partir du rapport sur les coûts et l'utilisation à l'aide de Query et de QuickSight Athena
- Présentation des coûts de transfert des données pour les architectures courantes
- Utiliser AWS Cost Explorer pour analyser les coûts de transfert de données
- Optimisation du coût de vos architectures AWS en utilisant des fonctionnalités Amazon CloudFront
- Comment réduire les frais de transfert de données pour ma passerelle NAT ?

Gérer la demande et les sources d'approvisionnement

Question

- COÛT 9. Comment gérez-vous la demande et l'offre des ressources ?

COÛT 9. Comment gérez-vous la demande et l'offre des ressources ?

Pour une charge de travail dont les dépenses et les performances sont équilibrées, assurez-vous que tout ce que vous payez est utilisé et évitez une sous-utilisation importante des instances. Une métrique d'utilisation faussée dans un sens ou dans l'autre a un impact négatif sur votre organisation, que ce soit en termes de coûts d'exploitation (dégradation des performances due à une sur-utilisation) ou de gaspillage de dépenses AWS (en raison d'une sur-allocation).

Bonnes pratiques

- COST09-BP01 Effectuer une analyse de la demande de charge de travail
- COST09-BP02 Mettre en œuvre une mémoire tampon ou une limitation pour gérer la demande
- COST09-BP03 Fournir dynamiquement les ressources

COST09-BP01 Effectuer une analyse de la demande de charge de travail

Analysez la demande sur la charge de travail au fil du temps. Veillez à ce que l'analyse couvre les tendances saisonnières et représente avec précision les conditions d'exploitation pendant toute la durée de la charge de travail. L'effort d'analyse doit refléter les avantages potentiels : par exemple, le temps passé est proportionnel au coût de la charge de travail.

Niveau d’exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

L'analyse de la demande de charge de travail pour le cloud computing implique de comprendre les modèles et les caractéristiques des tâches informatiques qui sont lancées dans l'environnement de calcul. Cette analyse aide les utilisateurs à optimiser l'affectation des ressources, à gérer les coûts et à vérifier que les performances sont conformes aux niveaux requis.

Ayez connaissance des exigences de la charge de travail. Les exigences de votre organisation doivent indiquer les délais de réponse de la charge de travail aux demandes. Le temps de réponse peut être utilisé pour déterminer si la demande est gérée, ou si l'offre de ressources doit changer pour répondre à la demande.

L'analyse doit inclure la prévisibilité et la répétabilité de la demande ainsi que le taux et l'ampleur de variation de la demande. Effectuez l'analyse sur une période suffisamment longue pour tenir compte des variations saisonnières, telles que les traitements de fin de mois ou les pics d'activité pendant les vacances.

L'effort d'analyse doit refléter les avantages potentiels de la mise à l'échelle. Examinez le coût total attendu du composant, ainsi que les augmentations ou diminutions d'utilisation et de coût au cours de la durée de vie de la charge de travail.

Voici quelques aspects clés dont il faut tenir compte lors de l'analyse de la demande de charge de travail pour le cloud computing :

1. Indicateurs d'utilisation des ressources et de performance: analysez la manière dont les ressources AWS sont utilisées au fil du temps. Déterminez les schémas d'utilisation en période de pointe et en période creuse afin d'optimiser l'affectation des ressources et les stratégies de mise à l'échelle. Surveillez les métriques de performance telles que les temps de réponse, la latence, le débit et les taux d'erreur. Ces métriques permettent d'évaluer l'état et l'efficacité globales de l'infrastructure cloud.

2. Comportement des utilisateurs et des applications: comprenez le comportement de l'utilisateur et comment il affecte la demande de charge de travail. L'examen des schémas de trafic des utilisateurs permet d'améliorer la diffusion du contenu et la réactivité des applications. Analysez l'évolution des charges de travail en fonction de l'augmentation de la demande. Déterminez si les paramètres d'autoscaling sont configurés correctement et efficacement pour gérer les fluctuations de charge.
3. Types de charges de travail: identifiez les différents types de charges de travail s'exécutant dans le cloud, comme le traitement par lots, le traitement des données en temps réel, les applications web, les bases de données ou le machine learning. Chaque type de charge de travail peut avoir des besoins en ressources et des profils de performance différents.
4. Contrats de niveau de service (SLA): comparez les performances réelles aux contrats de niveau de service afin de garantir la conformité et d'identifier les domaines nécessitant une amélioration.

Vous pouvez utiliser Amazon CloudWatch pour collecter et suivre les métriques, surveiller les fichiers journaux, définir des alarmes et réagir automatiquement aux modifications de vos ressources AWS. Vous pouvez également utiliser Amazon CloudWatch pour gagner une visibilité à l'échelle du système sur l'utilisation des ressources, la performance de l'application et la santé opérationnelle.

Avec AWS Trusted Advisor, vous pouvez provisionner vos ressources en suivant les bonnes pratiques pour améliorer les performances et la fiabilité du système, renforcer la sécurité et rechercher des possibilités d'économies. Vous pouvez également désactiver les instances hors production et utiliser Amazon CloudWatch et Auto Scaling pour répondre aux augmentations ou aux réductions de la demande.

Enfin, vous pouvez utiliser AWS Cost Explorer ou Amazon QuickSight avec le fichier (CUR) AWS Cost and Usage Report ou les journaux de votre application pour effectuer une analyse avancée de la demande de charge de travail.

Globalement, une analyse complète de la demande de charge de travail permet aux entreprises de prendre des décisions éclairées sur le provisionnement, la mise à l'échelle et l'optimisation des ressources, ce qui se traduit par une amélioration des performances, de la rentabilité et de la satisfaction des utilisateurs.

Étapes d'implémentation

- Analyser les données de charge de travail existantes : Analysez les données de la charge de travail existante, des versions précédentes de la charge de travail ou des modèles d'utilisation prévus. Utilisez Amazon CloudWatch, les fichiers journaux et les données de surveillance pour

comprendre comment la charge de travail a été utilisée. Analysez un cycle complet de la charge de travail et recueillez des données pour tout changement saisonnier tel que les événements de fin de mois ou de fin d'année. L'effort reflété dans l'analyse doit refléter les caractéristiques de la charge de travail. L'effort le plus important doit porter sur les charges de travail à forte valeur ajoutée qui subissent les plus grandes variations dans la demande. Le moindre effort doit porter sur les charges de travail de faible valeur ajoutée qui subissent des variations minimes dans la demande.

- Prévoir l'influence extérieure : Rencontrez les membres des équipes de toute l'organisation qui peuvent influencer ou modifier la demande dans la charge de travail. Les équipes communes sont celles des ventes, du marketing ou du développement commercial. Collaborez avec elles pour connaître les cycles qu'elles appliquent et déterminer s'il existe des événements susceptibles de modifier la demande de la charge de travail. Prévoyez la demande de la charge de travail à l'aide de ces données.

Ressources

Documents connexes :

- Amazon CloudWatch
- AWS Trusted Advisor
- AWS X-Ray
- AWS Auto Scaling
- AWS Instance Scheduler
- Démarrer avec Amazon SQS
- AWS Cost Explorer
- Amazon QuickSight

Vidéos connexes :

Exemples connexes :

- Surveillance, suivi et analyse pour optimiser les coûts
- Recherche et analyse des journaux dans CloudWatch

COST09-BP02 Mettre en œuvre une mémoire tampon ou une limitation pour gérer la demande

La mise en mémoire tampon et la limitation modifient la charge de travail en atténuant les pics éventuels. Mettez en œuvre une limitation lorsque vos clients effectuent de nouveaux essais. Mettez en œuvre une mémoire tampon pour stocker la demande et reporter le traitement. Veillez à ce que vos limitations et mémoires tampon soient conçues de manière à ce que les clients reçoivent une réponse dans les délais requis.

Niveau de risque exposé si cette bonne pratique n'est pas établie: moyen

Directives d'implémentation

Dans le cloud computing, la mise en place d'une réserve ou d'une limitation est cruciale pour gérer la demande et réduire la capacité allouée requise pour votre charge de travail. Pour des performances optimales, il est essentiel d'évaluer la demande totale, y compris les pics, le taux de variation des demandes et le temps de réponse nécessaire. Lorsque les clients ont la possibilité de renvoyer leurs demandes, il devient pratique d'appliquer la limitation. À l'inverse, pour les clients qui ne disposent pas de fonctionnalités de nouvelle tentative, l'approche idéale consiste à mettre en œuvre une mémoire tampon. Ces mémoires tampons rationalisent l'afflux de demandes et optimisent l'interaction des applications avec des vitesses opérationnelles variées.

Courbe de demande avec deux pics distincts qui nécessitent un provisionnement de capacité élevé

Prenons l'exemple d'une charge de travail dont la courbe de demande est représentée dans l'image précédente. Cette charge de travail a deux pics, et pour gérer ces pics, la capacité des ressources comme indiqué par la ligne orange est provisionnée. Les ressources et l'énergie utilisées pour cette charge de travail ne sont pas indiquées par la zone sous la courbe de la demande, mais par la zone sous la ligne de la capacité provisionnée, car cette dernière est nécessaire pour gérer ces deux pics. L'aplanissement de la courbe de demande de la charge de travail peut vous aider à réduire la capacité provisionnée pour une charge de travail et à réduire son impact environnemental. Pour atténuer le pic, envisagez de mettre en œuvre une limitation ou une mise en mémoire tampon.

Pour mieux les comprendre, examinons les notions de limitation et de mise en mémoire tampon.

Limitation : si la source de la demande peut exécuter de nouvelles tentatives, vous pouvez mettre en place une limitation. La limitation indique à la source qu'elle doit réessayer ultérieurement si elle ne peut répondre à la demande actuellement. La source attend un certain temps, puis relance la demande. L'implémentation de la limitation a l'avantage de limiter la quantité maximale de ressources

et les coûts maximaux de la charge de travail. Dans AWS, vous pouvez utiliser l'Amazon API Gateway pour mettre en place une limitation.

Mémoire tampon : une mémoire tampon utilise des producteurs (composants qui envoient des messages à la file d'attente), des consommateurs (composants qui reçoivent des messages de la file d'attente) et une file d'attente (qui contient des messages) pour stocker les messages. Les messages sont lus par les consommateurs et traités, ce qui permet aux messages de fonctionner au rythme qui répond aux besoins des entreprises. À l'aide d'une mémoire tampon, les messages des producteurs sont hébergés dans des files d'attente ou des flux, prêts à être consultés par les consommateurs en fonction de leurs besoins opérationnels.

Dans AWS, vous pouvez choisir parmi plusieurs services pour mettre en place une mémoire tampon. Amazon Simple Queue Service(Amazon SQS) est un service géré qui fournit des files d'attente permettant à un seul consommateur de lire des messages individuels. Amazon Kinesis fournit un flux qui permet à plusieurs consommateurs de lire les mêmes messages.

La mise en mémoire tampon et la limitation peuvent atténuer les pics éventuels en modifiant la sollicitation de votre charge de travail. Utilisez la limitation lorsque les clients retentent des actions, et la mise en mémoire tampon pour conserver la demande et la traiter ultérieurement. Si vous utilisez une mise en mémoire tampon, créez votre charge de travail de manière à ce qu'elle réponde à la demande dans les délais requis et assurez-vous que vous êtes en mesure de traiter les demandes de travail en double. Analysez la demande globale, le taux de variation et le temps de réponse requis pour dimensionner correctement la limitation ou le tampon nécessaire.

Étapes d'implémentation

- Analyser les besoins du client : analysez les demandes du client pour déterminer s'il peut effectuer de nouvelles tentatives. S'il ne le peut pas, des mémoires tampon doivent être mises en œuvre. Analysez la demande globale, le taux de variation et le temps de réponse requis pour déterminer la taille de limitation ou de mémoire tampon nécessaire.
- Mettre en place une mémoire tampon ou une limitation : mettez en place une mémoire tampon ou une limitation dans la charge de travail. Une file d'attente comme Amazon Simple Queue Service (Amazon SQS) peut fournir une mémoire tampon à vos composants de charge de travail. Amazon API Gateway peut fournir une limitation pour vos composants de charge de travail.

Ressources

Bonnes pratiques associées :

- SUS02-BP06 Mise en œuvre de la mise en mémoire tampon ou de la limitation pour aplanir la courbe de la demande
- REL05-BP02 Limiter les demandes

Documents connexes :

- AWS Auto Scaling
- AWS Instance Scheduler
- Amazon API Gateway
- Amazon Simple Queue Service
- Getting started with Amazon SQS
- Amazon Kinesis

Vidéos connexes :

- Choosing the Right Messaging Service for Your Distributed App

Exemples connexes :

- Gestion et surveillance de la limitation des API dans vos charges de travail
- Throttling a tiered, multi-tenant REST API at scale using API Gateway
- Enabling Tiering and Throttling in a Multi-Tenant Amazon EKS SaaS Solution Using Amazon API Gateway
- Application integration Using Queues and Messages

COST09-BP03 Fournir dynamiquement les ressources

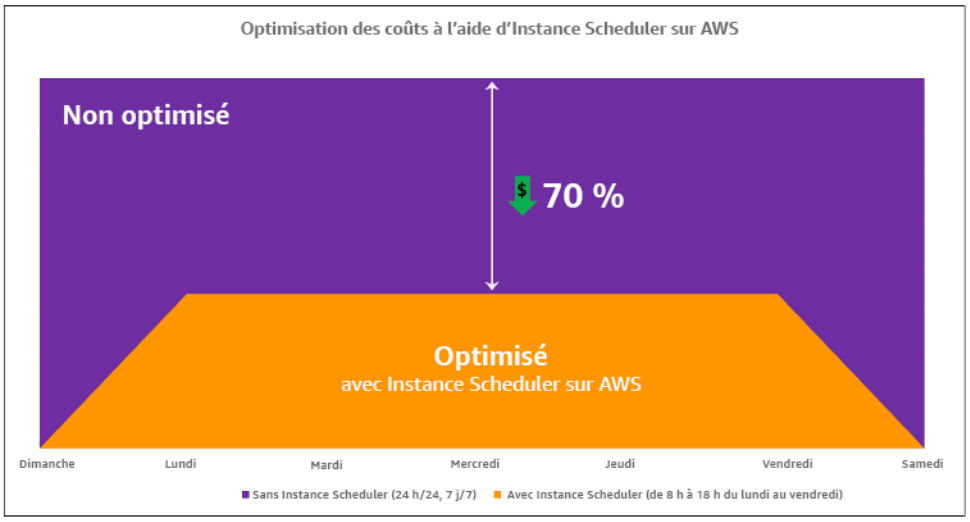
Les ressources sont allouées de façon planifiée. Cela peut reposer sur la demande, par exemple, via une mise à l'échelle automatique, ou sur le temps, lorsque la demande est prévisible et que les ressources sont fournies en fonction de la durée. Ces méthodes permettent de réduire au minimum la sur- ou sous-allocation.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

Les clients AWS peuvent augmenter les ressources disponibles pour leurs applications et fournir des ressources pour répondre à la demande de plusieurs manières. L'une de ces options consiste à utiliser AWS Instance Scheduler, qui automatise le démarrage et l'arrêt des instances Amazon Elastic Compute Cloud (Amazon EC2) et Amazon Relational Database Service (Amazon RDS). L'autre option consiste à utiliser AWS Auto Scaling, ce qui vous permet de dimensionner automatiquement vos ressources informatiques en fonction de la demande de votre application ou de votre service. La fourniture de ressources en fonction de la demande vous permettra de payer uniquement les ressources que vous utilisez, de réduire les coûts en lançant des ressources lorsqu'elles sont nécessaires et d'y mettre fin lorsqu'elles ne le sont pas.

AWS Instance Scheduler vous permet de configurer l'arrêt et le redémarrage de vos instances Amazon EC2 et Amazon RDS à des moments précis afin que vous puissiez répondre à la demande pour les mêmes ressources au sein d'un modèle de temps constant, comme le fait que des utilisateurs quotidiens accèdent à des instances Amazon EC2 à 8 h pour ne plus en avoir besoin après 18 h. Cette solution permet de réduire les coûts opérationnels en arrêtant des ressources qui ne sont pas utilisées et en les redémarrant quand il le faut.



Optimisation des coûts avec AWS Instance Scheduler.

Vous pouvez également configurer facilement les planifications pour vos instances Amazon EC2 sur l'ensemble de vos comptes et régions avec une interface utilisateur (UI) simple à l'aide de la configuration AWS Systems Manager rapide. Vous pouvez planifier des instances Amazon EC2 ou Amazon RDS avec AWS Instance Scheduler et vous pouvez arrêter et démarrer des instances existantes. Cependant, vous ne pouvez pas arrêter et démarrer des instances qui font partie de

votre groupe Auto Scaling (ASG) ou qui gèrent des services comme Amazon Redshift ou Amazon OpenSearch Service. Les groupes Auto Scaling disposent de leur propre planification pour les instances du groupe et ces instances sont créées.

AWS Auto Scaling vous permet d'ajuster votre capacité pour maintenir des performances stables et prévisibles au coût le plus bas possible. Il s'agit d'un service gratuit et entièrement géré permettant de mettre à l'échelle la capacité de votre application et qui s'intègre avec les instances Amazon EC2 et les parc d'instances Spot, Amazon ECS, Amazon DynamoDB et Amazon Aurora. Auto Scaling permet de découvrir automatiquement les ressources de votre charge de travail qui peuvent être configurées. Le service est doté de stratégies de mise à l'échelle intégrées pour optimiser les performances, les coûts ou un équilibre entre les deux et offre une mise à l'échelle prédictive pour faire face aux pics réguliers.

Plusieurs options de mise à l'échelle sont disponibles pour redimensionner votre groupe Auto Scaling :

- Maintien des niveaux d'instance actuels à tout moment
- Mise à l'échelle manuelle
- Mise à l'échelle basée sur un calendrier
- Mise à l'échelle basée sur la demande
- Utilisation de la mise à l'échelle prédictive

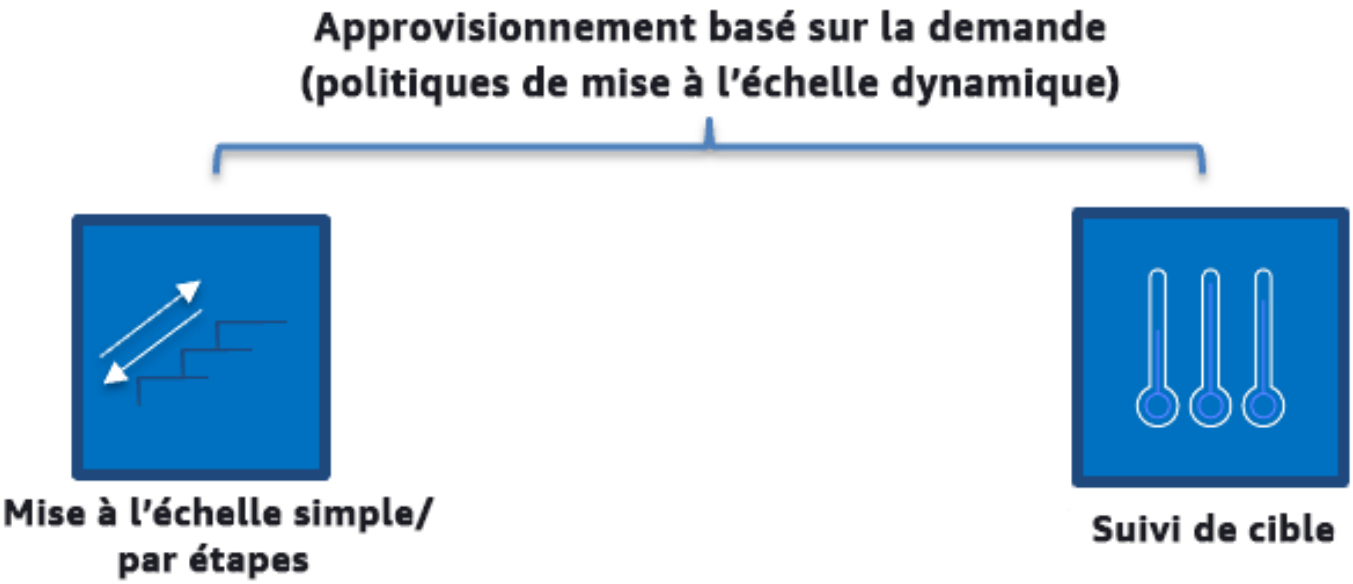
Les politiques Auto Scaling diffèrent et peuvent être classées dans la catégorie des politiques de mise à l'échelle dynamiques et planifiées. Les politiques dynamiques sont une mise à l'échelle manuelle ou dynamique, une mise à l'échelle planifiée ou prédictive. Vous pouvez utiliser des politiques de mise à l'échelle pour une mise à l'échelle dynamique, planifiée et prédictive. Vous pouvez également utiliser les métriques et les alarmes d' Amazon CloudWatch pour déclencher des événements de mise à l'échelle pour votre charge de travail. Nous vous recommandons d'utiliser des modèles de lancement, qui vous permettent d'accéder aux dernières fonctionnalités et améliorations. Toutes les fonctionnalités Auto Scaling ne sont pas disponibles lorsque vous utilisez les configurations de lancement. Par exemple, vous ne pouvez pas créer de groupe Auto Scaling qui lance à la fois des instances Spot et à la demande ou qui spécifie plusieurs types d'instances. Vous devez utiliser un modèle de lancement pour configurer ces fonctionnalités. Lorsque vous utilisez des modèles de lancement, nous vous recommandons de créer une version pour chacun d'entre eux. La gestion des versions des modèles de lancement vous permet de créer un sous-ensemble de l'ensemble complet de paramètres. Vous pouvez ensuite le réutiliser pour créer d'autres versions du même modèle de lancement.

Vous pouvez utiliser AWS Auto Scaling ou intégrer la mise à l'échelle dans votre code avec les API ou kits AWS SDK. Cela réduit le coût global de votre charge de travail en supprimant le coût opérationnel lié à la modification manuelle de votre environnement et les modifications peuvent être réalisées beaucoup plus rapidement. Cela adapte également les ressources de votre charge de travail à votre demande à tout moment. Afin de suivre cette bonne pratique et de fournir des ressources de façon dynamique à votre organisation, vous devez comprendre la mise à l'échelle horizontale et verticale dans le AWS Cloud, ainsi que la nature des applications exécutées sur des instances Amazon EC2. Il est préférable que votre équipe de gestion financière du cloud travaille avec les équipes techniques afin de suivre cette bonne pratique.

Elastic Load Balancing (Elastic Load Balancing) vous aide à mettre à l'échelle en répartissant la demande sur plusieurs ressources. En utilisant ASG et Elastic Load Balancing, vous pouvez gérer les demandes entrantes en acheminant le trafic de manière optimale afin qu'aucune instance ne soit surchargée au sein d'un groupe Auto Scaling. Les demandes seraient réparties entre toutes les cibles d'un groupe cible selon une procédure circulaire sans tenir compte de la capacité ou de l'utilisation.

Les métriques types peuvent être des métriques Amazon EC2 standard, telles que l'utilisation du processeur, le débit réseau et la latence de demande/réponse observée par Elastic Load Balancing. Dans la mesure du possible, vous devez utiliser une métrique qui indique l'expérience du client, généralement une métrique personnalisée qui peut provenir du code d'application au sein de votre charge de travail. Pour expliquer comment répondre à la demande de manière dynamique dans ce document, nous allons les regrouper en deux catégories Auto Scaling, à savoir les modèles d'approvisionnement basés sur la demande et les modèles d'approvisionnement basés sur le temps, et nous allons approfondir chacune d'entre elles.

Offre basée sur la demande : tirez parti de l'élasticité du cloud pour fournir les ressources nécessaires à l'évolution de la demande en vous appuyant sur l'état de la demande en temps quasi réel. Pour l'offre basée sur la demande, utilisez des API ou des fonctions de service pour faire varier par programmation et de façon dynamique la quantité de ressources cloud dans votre architecture. Cela vous permet de mettre à l'échelle les composants de votre architecture, d'augmenter le nombre de ressources pendant les pics de demande pour maintenir les performances, et de diminuer la capacité lorsque la demande diminue pour réduire les coûts.

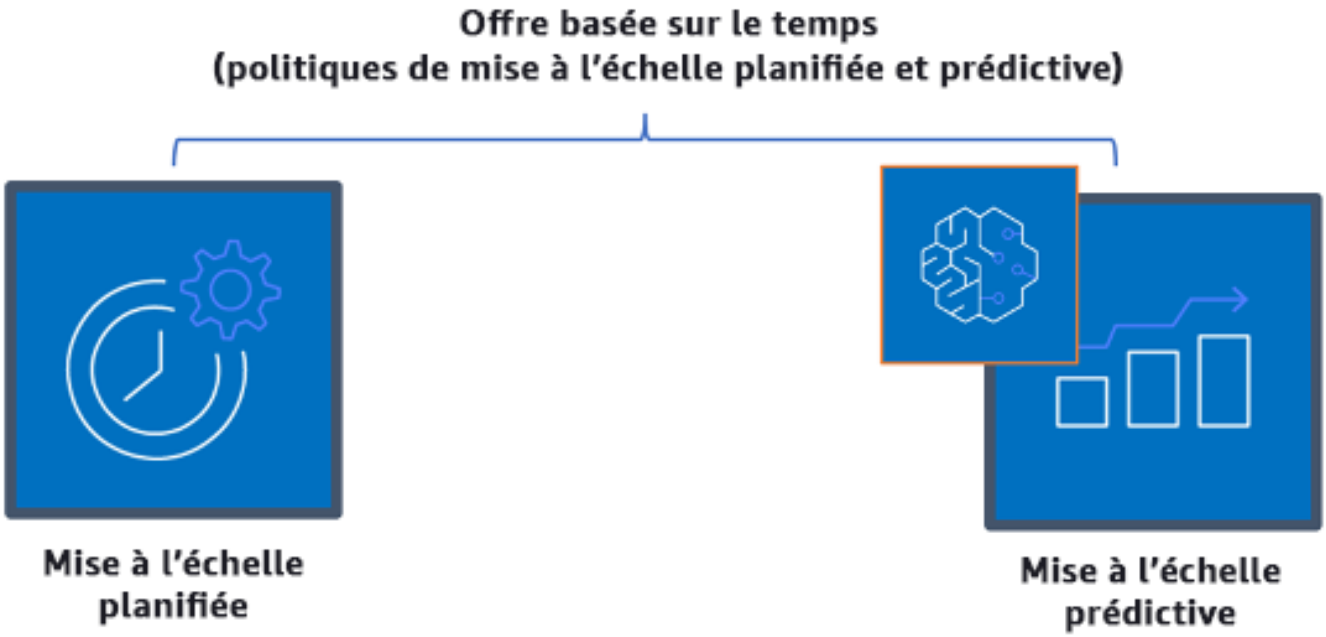


Politiques de mise à l'échelle dynamique basées sur la demande

- Mise à l'échelle simple/par étapes : surveille les métriques et ajoute/supprime des instances selon les étapes définies manuellement par les clients.
- Suivi des cibles : mécanisme de contrôle semblable à un thermostat qui ajoute ou supprime automatiquement des instances afin de maintenir les métriques à un objectif défini par le client.

Lorsque vous concevez une architecture en adoptant une approche basée sur la demande, gardez à l'esprit deux considérations clés. Premièrement, vous devez comprendre à quelle vitesse vous devez fournir de nouvelles ressources. Deuxièmement, vous devez comprendre que l'importance de la marge entre l'offre et la demande variera. Vous devez être prêt à faire face au taux de variation de la demande, ainsi qu'aux défaillances de ressources.

Offre basée sur le temps : Une approche basée sur le temps aligne la capacité des ressources avec une demande prévisible ou bien définie en fonction du temps. Cette approche ne dépend généralement pas des niveaux d'utilisation des ressources. Une approche basée sur le temps garantit que les ressources sont disponibles au moment précis où elles sont nécessaires et peuvent être fournies sans aucun retard dû à des procédures de démarrage et aux vérifications du système ou de la cohérence. Grâce à une approche basée sur le temps, vous pouvez fournir des ressources supplémentaires ou augmenter la capacité pendant les périodes de pointe.



Politiques de mise à l'échelle basées sur le temps

Vous pouvez utiliser l'autoscaling planifié pour mettre en place une approche basée sur le temps. Les charges de travail peuvent être programmées de manière à être réduites ou augmentées à des moments définis (par exemple, au début des heures de travail), ce qui rend les ressources disponibles lorsque les utilisateurs arrivent ou que la demande augmente. La mise à l'échelle prédictive utilise des modèles pour monter en puissance, tandis que la mise à l'échelle planifiée utilise des heures prédéfinies pour monter en puissance. Vous pouvez également utiliser la stratégie de sélection du type d'instance basée sur les attributs (ABS) dans les groupes Auto Scaling, ce qui vous permet d'exprimer les besoins de votre instance sous la forme d'un ensemble d'attributs, tels que le processeur virtuel, la mémoire et le stockage. Cela vous permet d'utiliser automatiquement les types d'instance de nouvelle génération lorsqu'ils sont disponibles et d'accéder à une plus large gamme de capacités avec les instances Spot Amazon EC2. La flotte Amazon EC2 et Amazon EC2 Auto Scaling sélectionnent et lancent les instances qui correspondent aux attributs spécifiés, en éliminant le besoin de sélectionner manuellement les types d'instance.

Vous pouvez également tirer parti des API et des kits SDK AWS et AWS CloudFormation pour automatiquement mettre en service ou hors service des environnements complets, selon vos besoins. Cette approche est idéale pour les environnements de développement ou de test qui s'exécutent uniquement pendant des heures ou des périodes de travail définies. Vous pouvez utiliser

les API pour mettre à l'échelle la taille des ressources au sein d'un environnement (mise à l'échelle verticale). Par exemple, vous pouvez monter en charge une charge de travail de production en modifiant la taille ou la catégorie d'instance. Cela peut être réalisé en arrêtant et en redémarrant l'instance, puis en sélectionnant une taille ou une catégorie différente. Cette technique peut être également appliquée à d'autres ressources, telles que les volumes Amazon EBS Elastic, qui peuvent être modifiés pour augmenter la taille, ajuster les performances (IOPS) ou changer le type de volume en cours d'utilisation.

Lorsque vous concevez une architecture en adoptant une approche basée sur le temps, gardez à l'esprit deux considérations clés. Premièrement, dans quelle mesure le modèle d'utilisation est-il cohérent ? Deuxièmement, quel est l'impact d'un changement de modèle ? Vous pouvez augmenter la précision des prédictions en surveillant vos charges de travail et en utilisant l'informatique décisionnelle. Si vous constatez des modifications importantes dans le modèle d'utilisation, vous pouvez ajuster les heures pour vous assurer que la couverture est fournie.

Étapes d'implémentation

- Configurez la mise à l'échelle planifiée : pour des changements prévisibles de la demande, une mise à l'échelle temporelle peut fournir le nombre correct de ressources en temps utile. Elle est également utile si la création et la configuration des ressources ne sont pas assez rapides pour répondre à l'évolution de la demande. À l'aide de l'analyse de la charge de travail, configurez la mise à l'échelle programmée via AWS Auto Scaling. Pour configurer une planification temporelle, vous pouvez utiliser la mise à l'échelle prédictive ou la mise à l'échelle planifiée pour augmenter le nombre d'instances Amazon EC2 dans votre groupe Auto Scaling à l'avance en fonction des changements de charge attendus ou prévisibles.
- Configurez la mise à l'échelle prédictive : la mise à l'échelle prédictive vous permet d'augmenter le nombre d'instances Amazon EC2 dans votre groupe Auto Scaling avant les modèles quotidiens ou hebdomadaires dans les flux de trafic. Si vous avez des pics de trafic réguliers et des applications lentes au démarrage, vous devez envisager la mise à l'échelle prédictive. La mise à l'échelle prédictive vous permet d'évoluer plus rapidement en initialisant de la capacité avant d'atteindre la charge projetée par comparaison avec la mise à l'échelle dynamique seule, qui est réactive par nature. Par exemple, si les utilisateurs commencent à utiliser votre charge de travail au début des heures de bureau mais pas pendant les heures qui suivent, la mise à l'échelle prédictive peut ajouter de la capacité avant le début des heures de bureau, ce qui supprime le retard lié au fait d'attendre que la mise à l'échelle dynamique réagisse au changement de trafic.
- Configurez la mise à l'échelle automatique dynamique : pour configurer la mise à l'échelle en fonction des métriques de charge de travail actives, utilisez Auto Scaling. Utilisez l'analyse et

configurez Auto Scaling pour déclencher les bons niveaux de ressources, et vérifiez que la charge de travail est mise à l'échelle dans les délais requis. Vous pouvez lancer et mettre à l'échelle automatiquement une flotte d'instances à la demande et d'instances Spot au sein d'un seul groupe Auto Scaling. En plus de recevoir des remises pour l'utilisation d'instances Spot, vous pouvez utiliser des instances réservées ou un Savings Plan qui vous permettront de bénéficier de taux réduits par rapport à la tarification standard des instances à la demande. Tous ces facteurs combinés vous aident à optimiser vos économies de coûts pour les instances Amazon EC2 et à obtenir l'échelle et les performances souhaitées pour votre application.

Ressources

Documents connexes :

- AWS Auto Scaling
- AWS Instance Scheduler
- Mettre à l'échelle la taille de votre groupe Auto Scaling
- Démarrer avec Amazon EC2 Auto Scaling
- Démarrer avec Amazon SQS
- Mise à l'échelle planifiée pour Amazon EC2 Auto Scaling
- Mise à l'échelle prédictive pour Amazon EC2 Auto Scaling

Vidéos connexes :

- Target Tracking Scaling Policies for Auto Scaling
- AWS Instance Scheduler

Exemples connexes :

- Sélection du type d'instance basée sur les attributs pour Auto Scaling pour la flotte Amazon EC2
- Optimisation d'Amazon Elastic Container Service pour les coûts à l'aide de la mise à l'échelle planifiée
- Mise à l'échelle prédictive avec Amazon EC2 Auto Scaling
- Comment utiliser Instance Scheduler avec AWS CloudFormation pour planifier des instances Amazon EC2 ?

Optimiser dans le temps

Questions

- COÛT 10. Comment évaluez-vous les nouveaux services?
- COÛT 11. Comment évaluer le coût de l'effort ?

COÛT 10. Comment évaluez-vous les nouveaux services?

Tandis qu'AWS propose de nouveaux services et de nouvelles fonctionnalités, une bonne pratique consiste à vérifier vos choix architecturaux existants afin d'être sûr qu'ils continuent à être les plus rentables.

Bonnes pratiques

- COST10-BP01 Développer un processus de révision de charge de travail
- COST10-BP02 Vérifier et analyser régulièrement cette charge de travail

COST10-BP01 Développer un processus de révision de charge de travail

Développez un processus qui définit les critères et le processus de révision de la charge de travail. La révision doit refléter les bénéfices potentiels. Par exemple, les charges de travail principales ou celles qui représentent plus de 10 % de la facture sont analysées chaque trimestre ou semestre, tandis que les charges de travail qui comptent pour moins de 10 % des frais sont examinées une fois par an.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour avoir la charge de travail la plus rentable, vous devez régulièrement la vérifier pour déterminer s'il existe des possibilités de mettre en œuvre de nouveaux services, fonctionnalités et composants. Pour réduire globalement les coûts, le processus doit être proportionnel au montant potentiel des économies. Par exemple, les charges de travail qui représentent 50 % de vos dépenses totales doivent être examinées plus régulièrement et plus en profondeur que les charges de travail qui représentent 5 % de vos dépenses totales. Facteur dans tous les facteurs externes ou volatilité. Si la charge de travail dessert une région géographique ou un segment de marché spécifique, et que l'on prévoit des changements dans ce domaine, des révisions plus fréquentes peuvent permettre de réaliser des économies. Un autre facteur à prendre en compte est l'effort de mise en œuvre des

modifications. Si des coûts importants sont associés au test et à la validation des modifications, les révisions doivent être moins fréquentes.

Tenez compte du coût à long terme de la maintenance des composants et ressources obsolètes et hérités ainsi que de l'impossibilité d'y intégrer de nouvelles fonctionnalités. Le coût actuel des tests et de la validation peut dépasser l'avantage proposé. Toutefois, au fil du temps, le coût du changement peut augmenter de manière significative, car l'écart entre la charge de travail et les technologies actuelles s'accroît, ce qui entraîne des coûts encore plus élevés. Par exemple, le coût du passage à un nouveau langage de programmation peut ne pas être actuellement rentable. Toutefois, en cinq ans, le coût des personnes compétentes dans cette langue pourrait augmenter et, en raison de l'accroissement de la charge de travail, vous transféreriez un système encore plus important vers la nouvelle langue, ce qui nécessiterait encore plus d'efforts qu'auparavant.

Décomposez votre charge de travail en composants, attribuez le coût du composant (une estimation est suffisante), puis énumérez les facteurs (par exemple, l'effort et les marchés extérieurs) à côté de chaque composant. Utilisez ces indicateurs pour déterminer une fréquence de révision pour chaque charge de travail. Par exemple, vous pouvez avoir des serveurs Web représentant un coût élevé, un faible effort de changement et des facteurs externes élevés, ce qui entraîne une fréquence de vérification révision. Une base de données centrale peut être un coût moyen, impliquer un effort de modification élevé et représenter des facteurs externes faibles, ce qui se traduit par une fréquence de révision moyenne.

Définissez un processus d'évaluation de nouveaux services, les modèles de conception, les types de ressources et les configurations pour optimiser le coût de votre charge de travail au fur et à mesure qu'elles deviennent disponibles. Comme pour les processus Vérifiez du pilier Performances et Vérifiez du pilier Fiabilité, identifiez, validez et priorisez l'optimisation et l'amélioration des activités et la résolution des problèmes, et intégrez-les à votre liste de suivi (backlog).

Étapes d'implémentation

- Définir la fréquence de vérification : définissez la fréquence à laquelle la charge de travail et ses composants doivent être vérifiés. Allouez du temps et des ressources pour une amélioration continue et une fréquence de vérification afin d'améliorer l'efficacité et l'optimisation de votre charge de travail. Il s'agit d'une combinaison de facteurs qui peut varier en fonction de la charge de travail dans votre organisation et d'un composant à l'autre dans la charge de travail. Les facteurs courants incluent l'importance pour l'organisation mesurée en termes de chiffre d'affaires ou de marque, le coût total d'exécution de la charge de travail (y compris les coûts d'exploitation et des ressources), la complexité de la charge de travail, la facilité de mise en œuvre d'un changement, les contrats de licence de logiciel et l'augmentation significative des coûts de licences pénalisants

en cas de changement. Les composants peuvent être définis de manière fonctionnelle ou technique, tels que les serveurs Web et les bases de données, ou les ressources de calcul et de stockage. Équilibrez les facteurs en conséquence et développez une période pour la charge de travail et ses composants. Vous pouvez décider de vérifier la charge de travail complète tous les 18 mois, les serveurs web tous les 6 mois, la base de données tous les 12 mois, le stockage de calcul et de courte durée tous les 6 mois et le stockage de longue durée tous les 12 mois.

- Définir la rigueur de l'examen : définissez les efforts consacrés à l'examen de la charge de travail ou des composants de la charge de travail. Similaire à la fréquence de vérification, il s'agit d'un équilibre entre plusieurs facteurs. Évaluez et hiérarchisez vos possibilités d'amélioration afin de concentrer les efforts là où ils permettent d'obtenir les plus grands avantages, tout en estimant la quantité d'efforts nécessaire pour ces activités. Si les résultats attendus sont en deçà des objectifs et que les efforts requis sont plus coûteux, itérez alors avec d'autres plans d'action. Vos processus de vérification doivent dédier du temps et des ressources pour permettre d'effectuer des améliorations progressives continues. Par exemple, vous pouvez décider d'analyser le composant de base de données pendant une semaine, les ressources de calcul pendant une semaine et le stockage pendant quatre heures.

Ressources

Documents connexes :

- Blog des actualités AWS
- Types de cloud computing
- Quelles sont les nouveautés AWS ?

Exemples connexes :

- AWS Support Proactive Services (Services d’assistance proactifs)
- Regular workload reviews for SAP workloads (Vérifications régulières des charges de travail SAP)

COST10-BP02 Vérifier et analyser régulièrement cette charge de travail

Les charges de travail existantes sont régulièrement passées en revue sur la base de chaque processus défini afin de déterminer si de nouveaux services peuvent être adoptés, si les services existants peuvent être remplacés ou si les charges de travail peuvent être réarchitecturées.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

AWS ajoute constamment de nouvelles fonctionnalités afin que vous puissiez expérimenter et innover plus rapidement avec les dernières technologies. Le blog Nouveautés AWS détaille la manière dont AWS procède et fournit un aperçu rapide des services, des fonctionnalités et des annonces d'expansion régionale d'AWS au fur et à mesure de leur publication. Consultez les lancements annoncés et utilisez-les pour examiner et analyser vos charges de travail existantes. Pour profiter des avantages des nouveaux services et fonctionnalités d'AWS, vous devez passer en revue vos charges de travail et mettre en œuvre de nouveaux services et fonctionnalités selon les besoins. Cela signifie que vous devrez peut-être remplacer les services existants que vous utilisez pour votre charge de travail, ou moderniser votre charge de travail pour adopter ces nouveaux services AWS. Par exemple, vous pouvez vérifier vos charges de travail et remplacer le composant de messagerie par Amazon Simple Email Service. Cela élimine le coût d'exploitation et de maintenance d'un parc d'instances, tout en fournissant toutes les fonctionnalités à un coût réduit.

Pour analyser votre charge de travail et mettre en évidence les opportunités potentielles, vous devez envisager non seulement de nouveaux services mais aussi de nouvelles façons de construire des solutions. Consultez la série de vidéos This is My Architecture (Voici mon architecture) sur AWS pour découvrir les conceptions architecturales d'autres clients, leurs défis et leurs solutions. Consultez All-In series pour découvrir les applications concrètes des services AWS et les témoignages de clients. Vous pouvez également regarder la série de vidéos Back to Basics (Retour à l'essentiel) qui explique, examine et décompose les bonnes pratiques de base en matière de modèles d'architecture de cloud. Autre source, les vidéos How to Build This (Comment construire ça) sont conçues pour aider les personnes ayant de grandes idées à donner vie à leur produit minimum viable (MVP) à l'aide des services AWS. Cela permet aux créateurs du monde entier ayant une idée forte de bénéficier des conseils d'architectes de solutions AWS expérimentés. Enfin, vous pouvez consulter les ressources documentaires de la page Démarrer avec AWS, qui contient des tutoriels étape par étape.

Avant de passer en revue votre architecture, suivez les exigences de votre entreprise en matière de charge de travail, de sécurité et de confidentialité des données afin d'utiliser un service ou une région spécifique, et les exigences de performance tout en déroulant votre processus d'examen.

Étapes d'implémentation

- Passer régulièrement en revue la charge de travail : en utilisant votre processus défini, effectuez des révisions à la fréquence spécifiée. Veillez à passer le temps approprié sur chaque composant. Ce processus est similaire au processus de conception initial dans lequel vous avez sélectionné des services pour l'optimisation des coûts. Analysez les services et les avantages qu'ils

apporteraient, cette fois-ci en tenant compte du coût du changement, et pas seulement des avantages à long terme.

- Mettre en œuvre de nouveaux services : si le résultat de l'analyse consiste à mettre en œuvre des modifications, effectuez d'abord une analyse de base de la charge de travail pour connaître le coût actuel de chaque sortie. Mettez en œuvre les modifications, puis effectuez une analyse pour vérifier le nouveau coût de chaque sortie.

Ressources

Documents connexes :

- Blog des actualités AWS
- Quelles sont les nouveautés AWS ?
- Documentation AWS
- Démarrer avec AWS
- Ressources générales AWS

Vidéos connexes :

- AWS - This is My Architecture (Voici mon architecture)
- AWS - Back to Basics (Retour à l'essentiel)
- AWS - All-In series
- How to Build This (Comment construire ça)

COÛT 11. Comment évaluer le coût de l'effort ?

Bonnes pratiques

- COST11-BP01 Réaliser des automatisations pour les opérations

COST11-BP01 Réaliser des automatisations pour les opérations

Évaluer le coût de l'effort pour les opérations sur le cloud. Quantifier la réduction du temps et des efforts consacrés aux tâches administratives, au déploiement et à d'autres opérations grâce à l'automatisation. Évaluer le temps et le coût nécessaires à l'effort d'exploitation et automatiser les tâches administratives pour réduire l'effort humain lorsque cela est possible.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

L'automatisation des opérations améliore la cohérence et la capacité de mise à l'échelle, offre davantage de visibilité, de fiabilité et de flexibilité, réduit les coûts et accélère l'innovation en libérant des ressources humaines tout en améliorant les métriques. Elle réduit la fréquence des tâches manuelles, améliore l'efficacité et profite aux entreprises en offrant une expérience cohérente et fiable lors du déploiement, de l'administration ou de l'exploitation des charges de travail. Vous pouvez libérer les ressources d'infrastructure des tâches opérationnelles manuelles et les utiliser pour des tâches et des innovations de plus grande valeur, améliorant ainsi les résultats opérationnels. Les entreprises ont besoin d'un moyen éprouvé et testé pour gérer leurs charges de travail dans le cloud. Cette solution doit être sûre, rapide et rentable, affichant un niveau de risque minimal et une fiabilité maximale.

Commencez par hiérarchiser vos opérations en fonction de l'effort requis en examinant le coût global des opérations dans le cloud. Par exemple, combien de temps faut-il pour déployer de nouvelles ressources dans le cloud, apporter des modifications d'optimisation aux ressources existantes ou mettre en œuvre les configurations nécessaires ? Examinez le coût total des actions humaines en tenant compte du coût des opérations et de la gestion. Privilégiez l'automatisation des tâches administratives afin de réduire l'effort humain. L'effort de révision doit refléter le bénéfice potentiel. Par exemple, le temps passé à effectuer des tâches manuellement plutôt qu'automatiquement. Donnez la priorité à l'automatisation des activités répétitives à forte valeur ajoutée. Les activités qui présentent un risque élevé d'erreur humaine sont généralement celles qu'il vaut mieux commencer à automatiser, car le risque représente souvent un coût opérationnel supplémentaire non souhaité (par exemple, l'équipe chargée des opérations fait des heures supplémentaires).

En utilisant des services AWS, des outils ou des produits tiers, vous pouvez choisir les automatisations AWS à mettre en œuvre et les adapter à vos besoins spécifiques. Le tableau suivant présente certaines des fonctions et des capacités d'exploitation de base que vous pouvez réaliser avec des services AWS pour automatiser l'administration et l'exploitation :

- AWS Audit Manager : auditez en permanence votre utilisation AWS pour simplifier l'appréciation du risque et de la conformité.
- AWS Backup : gérez et automatisez la protection des données de manière centralisée.
- AWS Config : configurez les ressources de calcul, évaluez, auditez et évaluez les configurations et l'inventaire des ressources.
- AWS CloudFormation : lancez des ressources hautement disponibles avec l'infrastructure en tant que code.

- AWS CloudTrail : gérez les modifications, la conformité et le contrôle des ressources informatiques.
- Amazon EventBridge : planifiez des événements et déclenchez des actions AWS Lambda.
- AWS Lambda : automatisez les processus répétitifs en les déclenchant avec des événements ou en les exécutant selon un calendrier fixe avec Amazon EventBridge.
- AWS Systems Manager : démarrez et arrêtez les charges de travail, corrigez les systèmes d'exploitation, automatisez la configuration et assurez la gestion continue.
- AWS Step Functions : planifiez les tâches et automatisez les flux de travail.
- AWS Service Catalog : utilisez des modèles et l'infrastructure en tant que code en bénéficiant de capacités de conformité et contrôle.

Tenez compte du gain de temps qui permettra à votre équipe de se concentrer sur le remboursement de la dette technique, l'innovation et les fonctionnalités à valeur ajoutée. Par exemple, il peut être nécessaire de procéder à un lift and shift de votre environnement sur site dans le cloud aussi rapidement que possible et de l'optimiser ultérieurement. Il est judicieux d'étudier les économies que vous pourriez réaliser en utilisant des services entièrement gérés par AWS qui suppriment ou réduisent les coûts de licence tels que Amazon Relational Database Service, Amazon EMR, Amazon WorkSpaces et Amazon SageMaker. Les services gérés suppriment la charge opérationnelle et administrative liée à la gestion d'un service pour vous permettre de vous dédier à l'innovation. En outre, comme les services gérés interviennent à l'échelle du cloud, ils peuvent offrir un coût moindre par transaction ou service.

Si vous souhaitez adopter immédiatement des automatisations en utilisant des produits et des services AWS et si vous ne disposez pas des compétences nécessaires dans votre organisation, contactez AWS Managed Services (AMS), les Services professionnels AWS ou les Partenaires AWS pour accroître l'adoption de l'automatisation et améliorer votre excellence opérationnelle dans le cloud.

AWS Managed Services (AMS) est un service qui exploite l'infrastructure AWS pour le compte des entreprises clientes et partenaires. Il fournit un environnement sécurisé et conforme sur lequel vous pouvez déployer vos charges de travail. AMS utilise des modèles d'exploitation de cloud d'entreprise avec l'automatisation pour permettre de répondre aux exigences de votre organisation, de migrer plus rapidement vers le cloud et de réduire vos coûts de gestion continue.

Les Services professionnels AWS peuvent également vous aider à atteindre les résultats opérationnels souhaités et à automatiser les opérations avec AWS. Les services professionnels AWS proposent des pratiques spécialisées mondiales pour soutenir vos efforts dans des domaines ciblés

du cloud computing d'entreprise. Les pratiques spécialisées fournissent des conseils ciblés par le biais de bonnes pratiques, de cadres, d'outils et de services dans les domaines des solutions, des technologies et des secteurs d'activité. Elles aident les clients à déployer des activités informatiques automatisées, robustes, agiles et des capacités de gouvernance optimisées pour le centre cloud.

Étapes d'implémentation

- Créer une seule fois et déployer à grande échelle : utilisez l'infrastructure en tant que code, comme AWS CloudFormation, le kit AWS SDK ou AWS Command Line Interface (AWS CLI), pour déployer une seule fois et utiliser votre modèle plusieurs fois dans le même environnement ou pour des scénarios de reprise après sinistre. Balisez lors du déploiement pour suivre votre consommation comme défini dans d'autres bonnes pratiques. Utilisez AWS Launch Wizard pour réduire la durée de déploiement de nombreuses charges de travail professionnelles populaires. AWS Launch Wizard vous guide dans le dimensionnement, la configuration et le déploiement de charges de travail professionnelles en suivant les bonnes pratiques AWS. Vous pouvez également utiliser AWS Service Catalog, qui vous aide à créer et à gérer des modèles approuvés d'infrastructure en tant que code à utiliser sur AWS, afin que tous les utilisateurs puissent découvrir des ressources approuvées en libre-service.
- Automatisez les opérations : exécutez les opérations de routine automatiquement sans intervention humaine. Grâce aux services et aux outils AWS, vous pouvez choisir les automatisations AWS à mettre en œuvre et les adapter à vos besoins spécifiques. Par exemple, utilisez EC2 Image Builder pour la création, le test et le déploiement d'images de machines virtuelles et de conteneurs pour une utilisation sur AWS ou sur site. Si l'action souhaitée ne peut pas être réalisée avec les services AWS ou si vous avez besoin de tâches plus complexes avec des ressources de filtrage, alors automatisez vos opérations en utilisant les outils AWS CLI ou le kit AWS SDK. AWS CLI permet d'automatiser l'ensemble du processus de contrôle et de gestion des services AWS via des scripts sans utiliser la console AWS. Sélectionnez vos kits AWS SDK préférés pour interagir avec les services AWS. Pour obtenir d'autres exemples de code, consultez le référentiel d'exemples de code du kit AWS SDK.

Ressources

Documents connexes :

- Modernizing operations in the AWS Cloud (Modernisation des opérations dans le cloud AWS)
- AWS Services for Automation (Services AWS pour l'automatisation)
- Automatisation AWS Systems Manager

- AWS automations for SAP administration and operations (Automatisations AWS pour l'administration et les opérations SAP)
- AWS Managed Services
- Services professionnels AWS
- Infrastructure et automatisation

Exemples connexes :

- Reinventing automated operations (Part I) [Réinventer les opérations automatisées (1ère partie)]
- Reinventing automated operations (Part II) [Réinventer les opérations automatisées (2ème partie)]
- AWS automations for SAP administration and operations (Automatisations AWS pour l'administration et les opérations SAP)
- Automatisations informatiques avec AWS Lambda
- Référentiel d'exemples de code AWS
- Exemples AWS

Durabilité

Lorsque des charges de travail sont créées dans le cloud, le pilier Durabilité consiste à comprendre les impacts des services utilisés, à mesurer les impacts tout au long du cycle de vie de la totalité de la charge de travail et à appliquer des principes de conception et de bonnes pratiques afin de réduire ces impacts. Vous trouverez des recommandations sur l'implémentation dans le Livre blanc du pilier Durabilité.

Domaines de bonnes pratiques

- Choix de la région
- Alignement sur la demande
- Logiciels et architecture
- Données
- Matériel et services
- Processus et culture

Choix de la région

Question

- SUS 1 Comment choisissez-vous les régions pour votre charge de travail ?

SUS 1 Comment choisissez-vous les régions pour votre charge de travail ?

Le choix de la région en fonction de votre charge de travail influe considérablement sur ses indicateurs de performance clés, y compris les performances, les coûts et l'empreinte carbone. Pour améliorer efficacement ces indicateurs de performance clés, vous devez choisir les régions pour vos charges de travail en fonction des exigences et des objectifs de durabilité de votre entreprise.

Bonnes pratiques

- SUS01-BP01 Choisir une région en fonction des exigences et des objectifs de durabilité de l'entreprise

SUS01-BP01 Choisir une région en fonction des exigences et des objectifs de durabilité de l'entreprise

Choisissez une région pour votre charge de travail en fonction des exigences et des objectifs de durabilité de votre entreprise afin d'optimiser ses KPI, dont les performances, les coûts et l'empreinte carbone.

Anti-modèles courants :

- Vous sélectionnez la région de la charge de travail en fonction de votre propre emplacement.
- Vous regroupez toutes les ressources de charge de travail dans un seul emplacement géographique.

Avantages liés au respect de cette bonne pratique : une charge de travail placée à proximité des projets d'énergie renouvelable d'Amazon ou des régions reconnues à faible intensité de carbone peut contribuer à la réduction de l'empreinte carbone d'une charge de travail cloud.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Le AWS Cloud est un réseau en constante expansion de régions et de points de présence (PoP), avec une infrastructure de réseau mondiale les reliant entre eux. Le choix de la région en fonction de votre charge de travail influe considérablement sur ses indicateurs de performance clés, y compris les performances, les coûts et l'empreinte carbone. Pour améliorer efficacement ces indicateurs de performance clés, vous devez choisir les régions pour votre charge de travail en fonction des exigences et des objectifs de durabilité de votre entreprise.

Étapes d'implémentation

- Suivez ces étapes pour évaluer et présélectionner les régions potentielles pour votre charge de travail en fonction des exigences de votre entreprise, y compris la conformité, les fonctionnalités disponibles, le coût et la latence :
  - Vérifiez que ces régions sont conformes, en fonction des réglementations locales à respecter.
  - Utilisez les listes de services régionaux AWS afin de vérifier si les régions proposent les services et fonctionnalités dont vous avez besoin pour exécuter votre charge de travail.
  - Calculez le coût de la charge de travail dans chaque région en utilisant AWS Pricing Calculator.
  - Testez la latence du réseau entre les emplacements des utilisateurs finaux et chaque Région AWS.
- Choisissez des régions proches des projets d'énergie renouvelable d'Amazon et des régions où le réseau a une intensité en carbone publique inférieure aux autres sites (ou régions).
  - Identifiez vos lignes directrices pertinentes en matière de durabilité pour suivre et comparer les émissions de carbone d'une année à l'autre en fonction du Protocole des GES (méthodes basées sur le marché et sur l'emplacement).
  - Choisissez une région en fonction de la méthode que vous utilisez pour suivre les émissions de carbone. Pour plus de détails sur le choix d'une région en fonction de vos directives de durabilité, consultez How to select a Region for your workload based on sustainability goals (Comment sélectionner une région pour votre charge de travail en fonction de vos objectifs de durabilité).

Ressources

Documents connexes :

- Comprendre les estimations de vos émissions de carbone
- Amazon à travers le monde

- Renewable Energy Methodology(Méthodologie de l'énergie renouvelable)
- What to Consider when Selecting a Region for your Workloads

Vidéos connexes :

- Architecting sustainably and reducing your AWS carbon footprint

Alignement sur la demande

Question

- SUS 2 Comment aligner les ressources du cloud sur votre demande ?

SUS 2 Comment aligner les ressources du cloud sur votre demande ?

La façon dont les utilisateurs et les applications consomment vos charges de travail et d'autres ressources peut vous aider à identifier les améliorations nécessaires pour atteindre vos objectifs de durabilité. Faites évoluer l'infrastructure pour répondre en permanence à la demande et vérifiez que vous n'utilisez que les ressources minimales requises pour prendre en charge vos utilisateurs. Alignez les niveaux de service sur les besoins des clients. Positionnez des ressources afin de limiter le réseau nécessaire aux utilisateurs et aux applications pour les consommer. Supprimez les ressources inutilisées. Fournissez aux membres de votre équipe des appareils qui répondent à leurs besoins et minimisent leur impact en matière de durabilité.

Bonnes pratiques

- SUS02-BP01 Mettre à l'échelle l'infrastructure de la charge de travail de façon dynamique
- SUS02-BP02 Aligner les SLA sur vos objectifs de durabilité
- SUS02-BP03 Arrêter la création et la maintenance des ressources inutilisées
- SUS02-BP04 Optimiser le placement géographique des charges de travail en fonction de leurs exigences réseau
- SUS02-BP05 Optimiser les ressources des membres de l'équipe pour les activités réalisées
- SUS02-BP06 Mise en œuvre de la mise en mémoire tampon ou de la limitation pour aplanir la courbe de la demande

SUS02-BP01 Mettre à l'échelle l'infrastructure de la charge de travail de façon dynamique

Utilisez l'élasticité du cloud et mettez à l'échelle votre infrastructure de façon dynamique afin de rapprocher l'offre de ressources cloud de la demande et d'éviter de surprovisionner une capacité dans votre charge de travail.

Anti-modèles courants :

- Vous ne mettez pas à l'échelle votre infrastructure avec la charge de l'utilisateur.
- Vous mettez à l'échelle manuellement votre infrastructure en permanence.
- Vous conservez une capacité accrue après un événement de mise à l'échelle au lieu de la réduire.

Avantages à établir cette meilleure pratique : configurer et tester l'élasticité de la charge de travail permet de rapprocher de façon efficace l'offre des ressources cloud de la demande et d'éviter de surprovisionner une capacité. Vous pouvez profiter de l'élasticité du cloud pour mettre à l'échelle automatiquement la capacité pendant et après les pics de demande, afin d'utiliser uniquement le bon nombre de ressources nécessaires pour répondre aux exigences de votre entreprise.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Le cloud vous apporte la flexibilité dont vous avez besoin pour développer ou réduire vos ressources de manière dynamique via une grande variété de mécanismes afin de répondre aux fluctuations de la demande. Rapprocher de façon optimale l'offre de la demande a le plus faible impact environnemental pour une charge de travail.

La demande peut être fixe ou variable, ce qui nécessite des métriques et une automatisation pour que la gestion ne devienne pas contraignante. Les applications peuvent se mettre à l'échelle de façon verticale (dans les deux sens) en modifiant la taille de l'instance, de façon horizontale (dans les deux sens) en modifiant le nombre d'instances, ou une combinaison des deux.

Vous pouvez utiliser plusieurs approches pour rapprocher l'offre de ressources de la demande.

- Approche visant à suivre les cibles : surveillez votre métrique de capacité de mise à l'échelle et augmentez ou réduisez automatiquement votre capacité selon vos besoins.
- Mise à l'échelle prédictive : mettez à l'échelle en prévision des tendances quotidiennes et hebdomadaires.

- Approche basée sur un calendrier : planifiez votre propre calendrier de mise à l'échelle en fonction de changements de charge prévisibles.
- Mise à l'échelle des services : sélectionnez des services (par exemple sans serveur) conçus pour se mettre à l'échelle ou fournissez une fonction de mise à l'échelle automatique.

Identifiez les périodes d'utilisation faible ou nulle, et mettez vos ressources à l'échelle afin de supprimer toute capacité excédentaire et améliorer l'efficacité.

Étapes d'implémentation

- L'élasticité correspond à l'offre de ressources dont vous disposez et à la demande pour ces ressources. Les instances, les conteneurs et les fonctions fournissent les mécanismes pour l'élasticité, soit en combinaison avec la mise à l'échelle automatique, soit en tant que fonction du service. AWS fournit une gamme de mécanismes de mise à l'échelle automatique pour veiller à ce que les charges de travail puissent réduire rapidement et facilement pendant les périodes de faible charge utilisateur. Voici des exemples de mécanismes de mise à l'échelle automatique :

Auto scaling mechanism	Where to use
Amazon EC2 Auto Scaling	Utilisez-le pour vous assurer que vous disposez du nombre adéquat d'instances Amazon EC2 disponibles pour gérer la charge utilisateur de votre application.
Application Auto Scaling	Utilisez-le pour mettre à l'échelle automatiq uement les ressources pour les services AWS individuels au-delà d'Amazon EC2, tels que les fonctions Lambda ou les services Amazon Elastic Container Service (Amazon ECS).
un Kubernetes Cluster Autoscaler	Utilisez-le pour mettre à l'échelle automatiq uement les clusters Kubernetes sur AWS.

- La mise à l'échelle est souvent discutée pour les services de calcul, tels que les instances Amazon EC2 ou les fonctions AWS Lambda. Envisagez la configuration de services non liés au calcul, tels que les unités de capacité de lecture et d'écriture Amazon DynamoDB ou les partitions Amazon Kinesis Data Streams pour rapprocher la demande.

- Vérifiez que les métriques de l'augmentation ou de la diminution sont validées par rapport au type de charge de travail déployée. Si vous déployez une application de transcodage vidéo, une utilisation de 100 % du processeur est attendue. N'en faites pas votre métrique principale. Vous pouvez utiliser une métrique personnalisée (telle que l'utilisation de la mémoire) pour votre politique de mise à l'échelle, le cas échéant. Pour choisir les bonnes métriques, tenez compte des conseils suivants pour Amazon EC2 :
  - La métrique doit être une métrique d'utilisation valide et décrire à quel point l'instance est occupée.
  - La valeur de la métrique doit augmenter ou diminuer proportionnellement au nombre d'instances dans le groupe Auto Scaling.
- Utilisez la mise à l'échelle dynamique au lieu de la mise à l'échelle manuelle pour votre groupe Auto Scaling. Nous vous recommandons également d'utiliser des politiques de mise à l'échelle en suivant les cibles pour votre mise à l'échelle dynamique.
- Vérifiez que les déploiements de charges de travail peuvent gérer à la fois les événements d'augmentation et de diminution des charges de travail. Créez des scénarios de test pour les événements de diminution afin de vérifier que la charge de travail se comporte comme prévu et n'a aucun impact sur l'expérience utilisateur (comme la perte de sessions permanentes). Vous pouvez utiliser Activity history (Historique de l'activité) pour vérifier une activité de mise à l'échelle pour un groupe Auto Scaling.
- Évaluez votre charge de travail pour les modèles prédictifs et mettez-la à l'échelle de manière proactive pour anticiper les changements prévisibles et prévus de la demande. Avec la mise à l'échelle prédictive, vous pouvez supprimer le besoin de surprovisionner de la capacité. Pour plus de détails, consultez Predictive Scaling with Amazon EC2 Auto Scaling (Mise à l'échelle prédictive avec Amazon EC2 Auto Scaling).

Ressources

Documents connexes :

- Démarrer avec Amazon EC2 Auto Scaling
- Scalabilité prédictive pour EC2 alimentée par le machine learning
- Analyser le comportement des utilisateurs avec Amazon OpenSearch Service, Amazon Data Firehose et Kibana
- Qu'est-ce qu'Amazon CloudWatch ?
- Surveillance de la charge de base de données avec Performance Insights sur Amazon RDS

- Présentation de la prise en charge native pour la mise à l'échelle prédictive avec Amazon EC2 Auto Scaling
- Présentation de Karpenter, un Kubernetes Cluster Autoscaler hautement performant et open source
- En savoir plus sur la Auto Scaling d'un cluster Amazon ECS

Vidéos connexes :

- Concevoir un environnement de calcul rentable, économe en énergie et en ressources
- Un calcul de meilleure qualité, plus rapide et moins cher : rentabiliser Amazon EC2 (CMP202-R1)

Exemples connexes :

- Atelier : exemples de groupes Amazon EC2 Auto Scaling
- Atelier : implémenter la mise à l'échelle automatique avec Karpenter

SUS02-BP02 Aligner les SLA sur vos objectifs de durabilité

Vérifiez et optimisez les contrats de niveau de service (SLA) de la charge de travail en fonction de vos objectifs de durabilité pour réduire les ressources nécessaires afin de prendre en charge votre charge de travail tout en continuant à répondre aux besoins de l'entreprise.

Anti-modèles courants :

- Les contrats de niveau de service (SLA) de la charge de travail ne sont pas connus ou ambigus.
- Vous définissez votre contrat de niveau de service (SLA) uniquement pour la disponibilité et les performances.
- Vous utilisez le même modèle de conception (comme une architecture multi-AZ) pour toutes vos charges de travail.

Avantages liés au respect de cette bonne pratique : l'alignement des contrats de niveau de service (SLA) sur les objectifs de durabilité entraîne une utilisation optimale des ressources tout en répondant aux besoins métier.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Les contrats de niveau de service (SLA) définissent le niveau de service attendu d'une charge de travail sur le cloud, comme le temps de réponse, la disponibilité et la conservation des données. Ils influencent l'architecture, l'utilisation des ressources et l'impact environnemental d'une charge de travail sur le cloud. À intervalles réguliers, vérifiez les contrats de niveau de service (SLA) et faites des compromis qui réduisent grandement l'utilisation des ressources en échange de baisses acceptables des niveaux de service.

Étapes d'implémentation

- Définissez ou rétablissez des contrats de niveau de service (SLA) qui soutiennent vos objectifs de durabilité tout en répondant aux exigences de l'entreprise, sans les dépasser.
- Faites des compromis qui réduisent grandement les impacts sur la durabilité en échange de baisses acceptables des niveaux de service.
  - Durabilité et fiabilité : les charges de travail hautement disponibles ont tendance à consommer plus de ressources.
  - Durabilité et performances : l'utilisation de plus de ressources pour booster les performances pourrait avoir un impact environnemental plus important.
  - Durabilité et sécurité : la sécurité trop importante des charges de travail pourrait avoir un impact environnemental plus important.
- Utilisez des modèles de conception, tels que des microservices sur AWS qui privilégient les fonctions essentielles à l'entreprise et permettent des niveaux de service inférieurs (tels que le temps de réponse ou les objectifs de temps pour la récupération) pour les fonctions non essentielles.

Ressources

Documents connexes :

- Accords de niveau de service (SLA) AWS
- Importance du contrat de niveau de service pour les fournisseurs de SaaS

Vidéos connexes :

- Delivering sustainable, high-performing architectures (Offre d'architectures durables hautement performantes)

- Concevoir un environnement de calcul rentable, économe en énergie et en ressources

SUS02-BP03 Arrêter la création et la maintenance des ressources inutilisées

Mettez hors service les ressources inutilisées de votre charge de travail afin de réduire le nombre de ressources cloud nécessaires pour répondre à votre demande et minimiser le gaspillage.

Anti-modèles courants :

- Vous n'analysez pas votre application pour détecter les ressources redondantes ou qui ne sont plus nécessaires.
- Vous ne supprimez pas les ressources redondantes ou qui ne sont plus nécessaires.

Avantages liés au respect de cette bonne pratique : la suppression des éléments inutilisés libère des ressources et améliore l'efficacité globale de la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Les ressources inutilisées consomment les ressources du cloud telles que l'espace de stockage et la puissance de calcul. En identifiant et en éliminant ces ressources, vous pouvez les libérer, ce qui se traduit par une architecture cloud plus efficace. Analysez régulièrement les ressources de l'application telles que les rapports pré-compilés, les jeux de données, les images statiques et les modèles d'accès aux ressources pour identifier des redondances, une sous-utilisation et d'éventuelles cibles de mise hors service. Supprimez ces ressources redondantes pour réduire le gaspillage de ressources dans votre charge de travail.

Étapes d'implémentation

- Utilisez des outils de surveillance pour identifier les ressources statiques qui ne sont plus nécessaires.
- Avant de supprimer une ressource, évaluez l'impact de sa suppression sur l'architecture.
- Élaborez un plan et supprimez les ressources qui ne sont plus nécessaires.
- Consolidez les ressources générées qui se chevauchent afin de supprimer tout traitement redondant.
- Mettez à jour vos applications pour ne plus produire et stocker les ressources qui ne sont pas nécessaires.

- Demandez aux tiers d'arrêter de produire et de stocker les ressources gérées en votre nom qui ne sont plus nécessaires.
- Demandez aux tiers d'arrêter de consolider les ressources redondantes produites en votre nom.
- Examinez régulièrement votre charge de travail pour identifier et supprimer les ressources inutilisées.

Ressources

Documents connexes :

- Optimisation de votre infrastructure AWS pour la durabilité, partie 2 : stockage
- Comment résilier les ressources actives dont je n'ai plus besoin sur mon Compte AWS ?

Vidéos connexes :

- Comment vérifier et supprimer les ressources actives dont je n'ai plus besoin sur mon Compte AWS ?

SUS02-BP04 Optimiser le placement géographique des charges de travail en fonction de leurs exigences réseau

Pour votre charge de travail, sélectionnez un emplacement et des services cloud qui réduisent la distance que le trafic réseau doit parcourir et diminuent les ressources réseau totales requises pour gérer votre charge de travail.

Anti-modèles courants :

- Vous sélectionnez la région de la charge de travail en fonction de votre propre emplacement.
- Vous regroupez toutes les ressources de charge de travail dans un seul emplacement géographique.
- Tout le trafic passe par vos centres de données existants.

Avantages liés au respect de cette bonne pratique : placer une charge de travail à proximité de ses clients fournit une faible latence, tout en réduisant les mouvements de données sur le réseau ainsi que l'impact sur l'environnement.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

L'infrastructure AWS Cloud est conçue autour d'options d'emplacement telles que les régions, les zones de disponibilité, les groupes de placement et les emplacements périphériques comme AWS Outposts et les zones locales AWS. Ces options d'emplacement sont responsables de la connectivité entre les composants d'application, les services cloud, les réseaux périphériques et les centres de données sur site.

Analysez les modèles d'accès au réseau dans votre charge de travail pour identifier comment utiliser ces options de localisation dans le cloud et réduire la distance que le trafic réseau doit parcourir.

Étapes d'implémentation

- Analysez les modèles d'accès au réseau dans votre charge de travail afin d'identifier comment les utilisateurs utilisent votre application.
  - Utilisez des outils de surveillance, comme Amazon CloudWatch et AWS CloudTrail, pour recueillir des données sur les activités du réseau.
  - Analysez les données pour identifier le modèle d'accès au réseau.
- Choisissez les régions pour votre déploiement de charge de travail en fonction des éléments clés suivants :
  - Objectif de durabilité comme indiqué dans Choix de la région.
  - Emplacement de vos données : pour les applications utilisant de grandes quantités de données (telles que le big data et le machine learning). Le code de l'application doit s'exécuter aussi près que possible des données.
  - Emplacement de vos utilisateurs : pour les applications orientées utilisateur, choisissez une région ou des régions proches des utilisateurs de votre charge de travail.
  - Autres contraintes : tenez compte de contraintes telles que le coût et la conformité comme indiqué dans Éléments à prendre en compte lors de la sélection d'une région pour vos charges de travail.
- Utilisez la mise en cache locale ou des solutions de mise en cache AWS pour les ressources fréquemment utilisées afin d'améliorer les performances, de limiter les mouvements de données et de réduire l'impact sur l'environnement.

Service	Quand l'utiliser
Amazon CloudFront	Permet de mettre en cache du contenu statique comme des images, des scripts et des vidéos, ainsi que du contenu dynamique comme des réponses API ou des applications Web.
Amazon ElastiCache	Permet de mettre en cache du contenu pour les applications Web.
DynamoDB Accelerator	Permet d'ajouter une accélération en mémoire à vos tables DynamoDB.

- Utilisez des services capables de vous aider à exécuter du code plus proche des utilisateurs de votre charge de travail :

Service	Quand l'utiliser
Lambda@Edge	Destiné aux opérations exigeantes en puissance de calcul qui sont lancées lorsque des objets ne sont pas dans le cache.
Fonctions Amazon CloudFront	Destiné aux cas d'utilisation simples comme une demande HTTP(S) ou des manipulat ions de réponse pouvant être lancées par des fonctions brèves.
AWS IoT Greengrass	Permet d'exécuter du calcul local, une messagerie et une mise en cache de données pour les appareils connectés.

- Utilisez le regroupement de connexions afin de pouvoir réutiliser les connexions et réduire les ressources nécessaires.
- Utilisez des magasins de données distribués qui ne s'appuient pas sur des connexions persistantes ni sur des mises à jour synchrones pour des raisons de cohérence afin de servir les populations régionales.

- Remplacez la capacité du réseau statique pré-allouée par une capacité dynamique partagée, et partagez l'impact en matière de durabilité de la capacité du réseau avec d'autres abonnés.

Ressources

Documents connexes :

- Optimisation de votre infrastructure AWS pour la durabilité, partie 3 : mise en réseau
- Documentation Amazon ElastiCache
- Qu'est-ce que Amazon CloudFront ?
- Fonctionnalités clés d'Amazon CloudFront

Vidéos connexes :

- Demystifying data transfer on AWS
- Scaling network performance on next-gen Amazon EC2 instances

Exemples connexes :

- Ateliers sur la mise en réseau AWS
- Une architecture axée sur la durabilité : réduisez les mouvements de données sur les réseaux

SUS02-BP05 Optimiser les ressources des membres de l'équipe pour les activités réalisées

Optimisez les ressources fournies aux membres de l'équipe pour réduire l'impact sur la durabilité environnementale tout en répondant à leurs besoins.

Anti-modèles courants :

- Vous ignorez l'impact des appareils utilisés par les membres de votre équipe sur l'efficacité globale de votre application cloud.
- Vous gérez et mettez à jour manuellement les ressources utilisées par les membres de l'équipe.

Avantages liés au respect de cette bonne pratique : l'optimisation des ressources des membres de l'équipe améliore l'efficacité globale des applications basées sur le cloud.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Comprenez les ressources que les membres de votre équipe utilisent pour consommer vos services, leur cycle de vie prévu et l'impact financier et sur la durabilité. Mettez en œuvre des stratégies pour optimiser ces ressources. Par exemple, effectuez des opérations complexes, telles que le rendu et la compilation, sur une infrastructure évolutive hautement utilisée plutôt que sur des systèmes mono-utilisateurs puissants et sous-utilisés.

Étapes d'implémentation

- Allouez des postes de travail et d'autres appareils conformément à leur utilisation.
- Utilisez des bureaux virtuels et le streaming d'applications pour limiter les exigences liées aux mises à niveau et aux appareils.
- Déplacez les tâches gourmandes en processeur ou en mémoire vers le cloud pour profiter de son élasticité.
- Évaluez l'impact des processus et des systèmes sur le cycle de vie de votre appareil et choisissez des solutions qui réduisent au minimum le besoin de remplacer celui-ci tout en répondant aux exigences de l'entreprise.
- Intégrez la gestion à distance des appareils afin de réduire les déplacements professionnels nécessaires.
  - AWS Systems Manager Fleet Manager est une interface utilisateur (IU) unifiée qui vous aide à gérer à distance vos nœuds fonctionnant sur site ou dans AWS.

Ressources

Documents connexes :

- Qu'est-ce qu'Amazon WorkSpaces ?
- Cost Optimizer for Amazon WorkSpaces (Optimiseur de coûts pour Amazon WorkSpaces)
- Documentation Amazon AppStream 2.0
- NICE DCV

Vidéos connexes :

- Managing cost for Amazon WorkSpaces on AWS (Gestion des coûts pour Amazon WorkSpaces sur AWS)

SUS02-BP06 Mise en œuvre de la mise en mémoire tampon ou de la limitation pour aplanir la courbe de la demande

La mise en mémoire tampon et la limitation aplatissent la courbe de la demande et réduisent la capacité provisionnée requise pour votre charge de travail.

Anti-modèles courants :

- Vous traitez les demandes des clients immédiatement alors que ce n'est pas nécessaire.
- Vous n'analysez pas les exigences des demandes des clients.

Avantages liés au respect de cette bonne pratique : l'aplatissement de la courbe de la demande réduit la capacité provisionnée requise pour la charge de travail. En réduisant la capacité provisionnée, on réduit la consommation d'énergie et l'impact environnemental.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

L'aplatissement de la courbe de demande de la charge de travail peut vous aider à réduire la capacité provisionnée pour une charge de travail et à réduire son impact environnemental. Prenons l'exemple une charge de travail dont la courbe de demande est représentée dans la figure ci-dessous. Cette charge de travail a deux pics, et pour gérer ces pics, la capacité des ressources comme indiqué par la ligne orange est provisionnée. Les ressources et l'énergie utilisées pour cette charge de travail ne sont pas indiquées par la zone sous la courbe de la demande, mais par la zone sous la ligne de la capacité provisionnée, car la capacité provisionnée est nécessaire pour gérer ces deux pics.

Courbe de demande avec deux pics distincts qui nécessitent une capacité provisionnée élevée.

Vous pouvez utiliser la mise en mémoire tampon ou la limitation pour modifier la courbe de la demande et lisser les pics, ce qui signifie moins de capacité provisionnée et moins d'énergie consommée. Mettez en œuvre la limitation lorsque vos clients peuvent effectuer de nouvelles tentatives. Mettez en œuvre une mémoire tampon pour stocker la demande et reporter le traitement.

Effet de la limitation sur la courbe de la demande et sur les capacités provisionnées.

Étapes d'implémentation

- Analysez les demandes des clients pour déterminer comment y répondre. Les questions à se poser sont les suivantes :
  - Cette demande peut-elle être traitée de manière asynchrone ?
  - Le client a-t-il la possibilité de lancer de nouvelles tentatives ?
- Si le client a la possibilité de lancer de nouvelles tentatives, vous pouvez mettre en œuvre un système de limitation, qui indique à la source que si elle ne peut pas répondre à la demande au moment même, elle doit réessayer plus tard.
  - Vous pouvez utiliser Amazon API Gateway pour la mise en œuvre de la limitation.
- Pour les clients qui ne peuvent pas effectuer de nouvelles tentatives, il faut mettre en œuvre un tampon pour aplanir la courbe de demande. Un tampon diffère le traitement des demandes, ce qui permet aux applications qui s'exécutent à différents débits de communiquer efficacement. Une approche basée sur la mémoire tampon utilise une file d'attente ou un flux pour accepter les messages des producteurs. Les messages sont lus par les consommateurs et traités, ce qui permet aux messages de fonctionner au rythme qui répond aux besoins des entreprises.
  - Amazon Simple Queue Service (Amazon SQS) est un service géré qui fournit des files d'attente permettant à un seul consommateur de lire des messages individuels.
  - Amazon Kinesis fournit un flux de données qui permet à de nombreux consommateurs de lire les mêmes messages.
- Analysez la demande globale, le taux de variation et le temps de réponse requis pour dimensionner correctement la limitation ou le tampon nécessaire.

Ressources

Documents connexes :

- Getting started with Amazon SQS (Démarrer avec Amazon SQS)
- Application integration Using Queues and Messages (Intégration des applications à l'aide de files d'attente et de messages)

Vidéos connexes :

- Choosing the Right Messaging Service for Your Distributed App (Choisir le bon service de messagerie pour votre application distribuée)

Logiciels et architecture

Question

- SUS 3 Comment tirer parti des modèles logiciels et d'architecture afin de soutenir vos objectifs de durabilité ?

SUS 3 Comment tirer parti des modèles logiciels et d'architecture afin de soutenir vos objectifs de durabilité ?

Mettez en œuvre des modèles permettant de lisser les charges et de conserver une haute utilisation constante des ressources déployées afin de réduire les ressources consommées. Les composants peuvent devenir inactifs s'ils ne sont pas utilisés à la suite de changements de comportement des utilisateurs dans le temps. Révisez les modèles et l'architecture afin de consolider les composants sous-utilisés et d'augmenter l'utilisation globale. Mettez hors service les composants qui ne sont plus nécessaires. Comprenez les performances des composants de vos charges de travail et optimisez les composants qui consomment le plus de ressources. Soyez conscient des appareils que vos clients utilisent pour accéder à vos services et mettez en œuvre des modèles qui réduisent le besoin de mettre à niveau ces appareils.

Bonnes pratiques

- SUS03-BP01 Optimiser les logiciels et l'architecture pour les tâches asynchrones et planifiées
- SUS03-BP02 Supprimer ou refactoriser les composants de charges de travail faiblement utilisés ou inutilisés
- SUS03-BP03 Optimiser les sections de votre code qui consomment le plus de temps ou de ressources
- SUS03-BP04 Optimiser l'impact sur les appareils et les équipements
- SUS03-BP05 Utiliser des modèles logiciels et des architectures qui soutiennent au mieux l'accès aux données et les modèles de stockage.

SUS03-BP01 Optimiser les logiciels et l'architecture pour les tâches asynchrones et planifiées

Utilisez des modèles d'architecture et de logiciels efficaces comme ceux axés sur les files d'attente afin de maintenir une utilisation élevée et constante des ressources déployées.

Anti-modèles courants :

- Vous mettez en service trop de ressources dans votre charge de travail cloud pour répondre aux pics imprévus de la demande.
- Votre architecture ne découple pas les expéditeurs et les destinataires de messages asynchrones par un composant de messagerie.

Avantages liés au respect de cette bonne pratique :

- Des modèles de logiciels et d'architecture efficaces réduisent les ressources inutilisées dans votre charge de travail et améliorent l'efficacité globale.
- Vous pouvez mettre à l'échelle le traitement indépendamment de la réception de messages asynchrones.
- Par le biais d'un composant de messagerie, vous avez assoupli les exigences de disponibilité auxquelles vous pouvez répondre avec moins de ressources.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Utilisez des modèles d'architecture efficaces tels que l'architecture axée sur les événements afin de bénéficier d'une utilisation uniforme des composants et réduire autant que possible le surprovisionnement dans votre charge de travail. L'utilisation de modèles d'architecture efficaces réduit au minimum les ressources inutilisées en raison des changements de la demande au fil du temps.

Comprenez les exigences des composants de votre charge de travail et adoptez des modèles d'architecture qui augmentent l'utilisation globale des ressources. Retirez les composants qui ne sont plus nécessaires.

Étapes d'implémentation

- Analysez la demande pour votre charge de travail afin de déterminer comment y répondre.
- Pour les demandes ou les tâches qui ne nécessitent pas de réponses synchrones, utilisez des architectures axées sur les files d'attente et des agents de travail de mise à l'échelle automatique afin de maximiser l'utilisation. Voici quelques exemples de situations où vous pourriez envisager une architecture axée sur les files d'attente :

Queuing mechanism	Description
Files d'attente de tâches AWS Batch	Les tâches AWS Batch sont soumises à une file d'attente de travail là où elles résident jusqu'à ce que leur exécution puisse être planifiée dans un environnement de calcul.
Amazon Simple Queue Service et instances Spot Amazon EC2	Associez Amazon SQS et les instances Spot pour créer une architecture tolérante aux pannes et efficace.

- Pour les demandes ou les tâches qui peuvent être traitées à tout moment, utilisez les mécanismes de planification afin de traiter les tâches par lots pour plus d'efficacité. Voici quelques exemples de planification des mécanismes sur AWS :

Scheduling mechanism	Description
Amazon EventBridge Scheduler	Une capacité d'Amazon EventBridge qui vous permet de créer, d'exécuter et de gérer des tâches planifiées à grande échelle.
Planification temporelle pour AWS Glue	Définissez une planification temporelle pour les crawlers et les tâches dans AWS Glue.
Tâches planifiées Amazon Elastic Container Service (Amazon ECS)	Amazon ECS prend en charge la création de tâches planifiées. Les tâches planifiée s utilisent des règles Amazon EventBridge pour exécuter des tâches dans le cadre d'une planification ou en réponse à un événement EventBridge.
Instance Scheduler	Configurez des calendriers pour les débuts et les arrêts des instances Amazon EC2 et Amazon Relational Database Service.

- Si vous utilisez des mécanismes d'interrogation et de webhooks dans votre architecture, remplacez-les par des événements. Utilisez des architectures axées sur les événements pour créer des charges de travail hautement efficaces.
- Utilisez le sans serveur sur AWS afin d'éliminer une infrastructure surprovisionnée.
- Dimensionnez les composants individuels afin d'éviter les ressources inactives attendant une entrée.

Ressources

Documents connexes :

- Qu'est-ce qu'Amazon Simple Queue Service ?
- Qu'est-ce qu'Amazon MQ ?
- Mise à l'échelle basée sur Amazon SQS
- Qu'est-ce qu'AWS Step Functions ?
- Qu'est-ce qu'AWS Lambda ?
- Utilisation d'AWS Lambda avec Amazon SQS
- Qu'est-ce qu'Amazon EventBridge ?

Vidéos connexes :

- Moving to event-driven architectures

SUS03-BP02 Supprimer ou refactoriser les composants de charges de travail faiblement utilisés ou inutilisés

Supprimez les composants inutilisés et devenus inutiles, et refactorisez les composants peu utilisés afin de minimiser le gaspillage dans votre charge de travail.

Anti-modèles courants :

- Vous ne vérifiez pas régulièrement le niveau d'utilisation des différents composants de votre charge de travail.
- Vous ne vérifiez pas et n'analysez pas les recommandations des outils de redimensionnement AWS tels que AWS Compute Optimizer.

Avantages liés au respect de cette bonne pratique : la suppression des composants inutilisés minimise le gaspillage et améliore l'efficacité globale de votre charge de travail dans le cloud.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Passez en revue votre charge de travail pour identifier les composants inutilisés. Il s'agit d'un processus d'amélioration itératif qui peut être déclenché par l'évolution de la demande ou le lancement d'un nouveau service de cloud. Par exemple, une baisse significative du temps d'exécution des fonctions AWS Lambda peut être un indicateur de la nécessité de réduire la taille de la mémoire. De plus, à mesure que AWS publie de nouveaux services et de nouvelles fonctionnalités, les services et l'architecture optimaux pour votre charge de travail peuvent changer.

Surveillez en permanence l'activité de la charge de travail et recherchez les possibilités d'améliorer le niveau d'utilisation des différents composants. En supprimant les composants inutilisés et en effectuant des activités de redimensionnement, vous répondez aux besoins de votre entreprise avec le moins de ressources cloud possible.

Étapes d'implémentation

- Surveillez et capturez les métriques d'utilisation des composants critiques de votre charge de travail (comme l'utilisation du CPU, l'utilisation de la mémoire ou le débit du réseau dans les métriques Amazon CloudWatch).
- Pour les charges de travail stables, vérifiez les outils de redimensionnement AWS tels que AWS Compute Optimizer à intervalles réguliers pour identifier les composants inactifs, inutilisés ou sous-utilisés.
- Pour les charges de travail éphémères, évaluez les métriques d'utilisation pour identifier les composants inactifs, inutilisés ou sous-utilisés.
- Retirez les composants et les ressources associées (comme les images Amazon ECR) qui ne sont plus nécessaires.
- Refactorisez ou consolidez les composants sous-utilisés avec d'autres ressources pour améliorer l'efficacité de l'utilisation. Par exemple, vous pouvez provisionner plusieurs petites bases de données sur une seule instance de base de données Amazon RDS au lieu d'exécuter des bases de données sur des instances individuelles sous-utilisées.
- Identifiez les ressources provisionnées par votre charge de travail pour mener à bien une unité de travail.

Ressources

Documents connexes :

- AWS Trusted Advisor
- Qu'est-ce qu'Amazon CloudWatch ?
- Automated Cleanup of Unused Images in Amazon ECR (Nettoyage automatisé des images inutilisées dans Amazon ECR)

Exemples connexes :

- Atelier Well-Architected : redimensionnement avec AWS Compute Optimizer
- Atelier Well-Architected : optimiser les modèles matériels et observer les indicateurs de performance clés de durabilité

SUS03-BP03 Optimiser les sections de votre code qui consomment le plus de temps ou de ressources

Optimisez votre code qui s'exécute dans les différents composants de votre architecture afin de minimiser l'utilisation des ressources tout en maximisant les performances.

Anti-modèles courants :

- Vous ignorez l'optimisation de votre code pour l'utilisation des ressources.
- Vous répondez généralement aux problèmes de performance en augmentant les ressources.
- Votre processus de révision et de développement du code ne permet pas de suivre les variations de performance.

Avantages liés au respect de cette bonne pratique : l'utilisation d'un code efficace permet de minimiser l'utilisation des ressources et d'améliorer les performances.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Il est essentiel d'examiner chaque domaine fonctionnel, y compris le code d'une application conçue dans le cloud, pour optimiser l'utilisation des ressources et les performances. Surveillez en permanence les performances de votre charge de travail dans les environnements de construction et

de production et identifiez les possibilités d'améliorer les extraits de code qui utilisent particulièrement bien les ressources. Adoptez un processus de révision régulier pour identifier les bogues ou les anti-modèles dans votre code qui utilisent les ressources de manière inefficace. Exploitez des algorithmes simples et efficaces qui produisent les mêmes résultats pour votre cas d'utilisation.

Étapes d'implémentation

- Pendant le développement de vos charges de travail, adoptez un processus de révision automatique du code pour améliorer la qualité et identifier les bogues et les anti-modèles.
  - Automatiser les révisions de code avec Amazon CodeGuru Reviewer
  - Détection des bogues de concurrence avec Amazon CodeGuru
  - Améliorer la qualité du code des applications Python grâce à Amazon CodeGuru
- Au fur et à mesure que vous exécutez vos charges de travail, surveillez les ressources afin d'identifier les composants dont les besoins en ressources par unité de travail sont élevés et qui peuvent faire l'objet d'une révision du code.
- Pour les révisions de code, utilisez un profileur de code pour identifier les sections du code les plus longues ou qui consomment le plus de ressources dans le but de les optimiser.
  - Réduire l'empreinte carbone de votre organisation avec Amazon CodeGuru Profiler
  - Comprendre l'utilisation de la mémoire dans votre application Java avec Amazon CodeGuru Profiler
  - Améliorer l'expérience client et réduire les coûts avec Amazon CodeGuru Profiler
- Utilisez le système d'exploitation et le langage de programmation les plus efficaces pour la charge de travail. Pour plus de détails sur les langages de programmation économes en énergie (y compris Rust), consultez Durabilité avec Rust.
- Remplacez les algorithmes à forte intensité de calcul par des versions plus simples et plus efficaces qui produisent le même résultat.
- Supprimez le code inutile tel que le tri et le formatage.

Ressources

Documents connexes :

- Qu'est-ce qu'Amazon CodeGuru Profiler ?
- Instances FPGA
- Les kits de développement logiciel (SDK) AWS sur les outils pour créer sur AWS

Vidéos connexes :

- Improve Code Efficiency Using Amazon CodeGuru Profiler
- Automate Code Reviews and Application Performance Recommendations with Amazon CodeGuru

SUS03-BP04 Optimiser l'impact sur les appareils et les équipements

Comprenez les appareils et les équipements utilisés dans votre architecture et employez des stratégies pour réduire leur utilisation. Cela peut minimiser l'impact environnemental global de votre charge de travail dans le cloud.

Anti-modèles courants :

- Vous ignorez l'impact environnemental des appareils utilisés par vos clients.
- Vous gérez et mettez à jour manuellement les ressources utilisées par les clients.

Avantages liés au respect de cette bonne pratique : la mise en œuvre de modèles et de fonctionnalités logicielles optimisés pour l'appareil du client peut réduire l'impact environnemental global de la charge de travail dans le cloud.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

La mise en œuvre de modèles et de fonctionnalités logicielles optimisés pour les appareils des clients peut réduire l'impact environnemental de plusieurs façons :

- La mise en œuvre de nouvelles fonctionnalités qui sont rétrocompatibles peut réduire le nombre de remplacements de matériel.
- L'optimisation d'une application pour qu'elle fonctionne efficacement sur les appareils peut contribuer à réduire leur consommation d'énergie et à prolonger leur durée de vie (s'ils sont alimentés par une batterie).
- L'optimisation d'une application pour les appareils peut également réduire le transfert de données sur le réseau.

Comprenez les appareils et les équipements utilisés dans votre architecture, leur cycle de vie prévu et l'impact du remplacement de ces composants. Mettez en œuvre des modèles et des

fonctionnalités logicielles qui minimisent la consommation d'énergie de l'appareil, réduisent la nécessité pour les clients de remplacer l'appareil et aussi de le mettre à niveau manuellement.

Étapes d'implémentation

- Dressez l'inventaire des appareils utilisés dans votre architecture. Il peut s'agir d'appareils mobiles, de tablettes, d'appareils IoT, de lampes intelligentes ou même d'appareils intelligents dans une usine.
- Optimisez l'application fonctionnant sur les appareils :
  - utilisez des stratégies telles que l'exécution de tâches en arrière-plan pour réduire leur consommation d'énergie.
  - Prenez en compte la bande passante et la latence du réseau lorsque vous créez des charges utiles et intégrez des capacités qui aident vos applications à fonctionner correctement sur des liens à faible bande passante et à latence élevée.
  - Convertissez les charges utiles et les fichiers dans les formats optimisés requis par les appareils. Par exemple, vous pouvez utiliser Amazon Elastic Transcoder ou AWS Elemental MediaConvert pour convertir des fichiers multimédias numériques volumineux et de haute qualité dans des formats que les utilisateurs peuvent lire sur des appareils mobiles, des tablettes, des navigateurs Web et des téléviseurs connectés.
  - Réalisez des activités gourmandes en calcul côté serveur (comme le rendu d'images) ou utilisez le streaming d'applications pour améliorer l'expérience utilisateur sur des appareils plus anciens.
  - Segmentez et paginez la sortie, en particulier, pour les séances interactives, afin de gérer les charges utiles et limiter les exigences en matière de stockage local.
- Utilisez le mécanisme automatisé par voie hertzienne (OTA) pour déployer les mises à jour sur un ou plusieurs appareils.
  - Vous pouvez utiliser un pipeline CI/CD pour mettre à jour les applications mobiles.
  - Vous pouvez utiliser AWS IoT Device Management pour gérer à distance les appareils connectés à grande échelle.
- Pour tester les nouvelles fonctionnalités et les mises à jour, utilisez AWS Device Farm avec des ensembles représentatifs de matériel et itérez le développement pour maximiser les dispositifs pris en charge. Pour en savoir plus, consultez SUS06-BP04 Utiliser des tests Device Farms gérés.

Ressources

Documents connexes :

- Qu'est-ce qu'AWS Device Farm ?
- Documentation Amazon AppStream 2.0
- NICE DCV
- OTA tutorial for updating firmware on devices running FreeRTOS (Tutoriel OTA pour la mise à jour du firmware sur les appareils fonctionnant sous FreeRTOS)

Vidéos connexes :

- Présentation d'AWS Device Farm

SUS03-BP05 Utiliser des modèles logiciels et des architectures qui soutiennent au mieux l'accès aux données et les modèles de stockage.

Comprenez comment les données sont utilisées au sein de votre charge de travail, comment elles sont consommées par vos utilisateurs, transférées et stockées. Utilisez des modèles et des architectures logicielles qui prennent le mieux en charge l'accès et le stockage des données afin de minimiser les ressources de calcul, de mise en réseau et de stockage nécessaires pour supporter la charge de travail.

Anti-modèles courants :

- Vous partez du principe que toutes les charges de travail ont des modèles de stockage de données et d'accès similaires.
- Vous n'utilisez qu'un seul niveau de stockage, partant du principe que toutes les charges de travail s'intègrent dans ce niveau.
- Vous partez du principe que les modèles d'accès aux données n'évolueront pas dans le temps.
- Votre architecture prend en charge un potentiel pic important d'accès aux données, ce qui fait que les ressources restent inactives la plupart du temps.

Avantages liés au respect de cette bonne pratique : la sélection et l'optimisation de votre architecture en fonction des modèles d'accès et de stockage des données permettront de réduire la complexité du développement et d'augmenter l'utilisation globale. Savoir quand utiliser les tables globales, le partitionnement des données et la mise en cache vous aidera à réduire les frais généraux opérationnels et à évoluer en fonction des besoins de votre charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Utilisez les modèles de logiciels et d'architecture qui correspondent le mieux aux caractéristiques de vos données et à vos modèles d'accès. Par exemple, utilisez une architecture de données moderne sur AWS qui vous permet d'utiliser des services spécialisés optimisés pour vos cas d'utilisation analytiques uniques. Ces modèles d'architecture permettent un traitement efficace des données et réduisent l'utilisation des ressources.

Étapes d'implémentation

- Analysez les caractéristiques de vos données et les modèles d'accès afin d'identifier la bonne configuration pour vos ressources cloud. Les caractéristiques clés à prendre en considération sont les suivantes :
  - Type de données : structuré, semi-structuré, non structuré
  - Croissance des données : limitée, illimitée
  - Durabilité des données : persistantes, éphémères, temporaires
  - Modèles d'accès en lecture ou écriture, fréquence de mise à jour, irrégularité, constance
- Utilisez les modèles d'architecture qui prennent le mieux en charge les modèles d'accès et de stockage des données.
  - Let's Architect! Architectures de données modernes
  - Bases de données sur AWS : le bon outil pour la bonne tâche
- Utilisez des technologies qui peuvent fonctionner en natif avec les données compressées.
- Utilisez des services d'analytique spécialisés pour le traitement des données dans votre architecture.
- Utilisez le moteur de base de données qui prend le mieux en charge votre modèle de requête dominant. Gérez vos index de bases de données afin de garantir l'exécution efficace de vos requêtes. Pour plus de détails, consultez Bases de données AWS.
- Sélectionnez des protocoles réseaux qui réduisent la quantité de capacité réseau consommée dans votre architecture.

Ressources

Documents connexes :

- Formats de fichiers prenant en charge la compression Athena
- COPIE de formats de données en colonnes avec Amazon Redshift

- Conversion de votre format de registre d'entrée dans Firehose
- Options de format pour les entrées et les sorties ETL dans AWS Glue
- Améliorer la performance des requêtes sur Amazon Athena grâce à une conversion en formats de colonnes
- Chargement de fichiers de données compressés depuis Amazon S3 vers Amazon Redshift
- Surveillance de la charge de base de données avec Performance Insights sur Amazon Aurora
- Surveillance de la charge de base de données avec Performance Insights sur Amazon RDS
- Amazon S3 Intelligent-Tiering storage class (Classe de stockage Amazon S3 Intelligent-Tiering)

Vidéos connexes :

- Building modern data architectures on AWS

Données

Question

- SUS 4 Comment tirez-vous parti des politiques et des modèles de gestion des données pour soutenir vos objectifs de durabilité ?

SUS 4 Comment tirez-vous parti des politiques et des modèles de gestion des données pour soutenir vos objectifs de durabilité ?

Mettez en œuvre des pratiques de gestion des données afin de réduire le stockage alloué nécessaire pour assurer votre charge de travail et les ressources nécessaires à son utilisation. Comprenez vos données et utilisez des technologies et des configurations de stockage qui soutiennent plus efficacement la valeur métier des données et leur utilisation. Adoptez un cycle de vie des données offrant un stockage plus efficace et moins performant quand les exigences baissent et supprimez les données qui ne sont plus nécessaires.

Bonnes pratiques

- SUS04-BP01 Mettre en œuvre une politique de classification des données
- SUS04-BP02 Utiliser les technologies qui prennent en charge les modèles d'accès aux données et les modèles de stockage
- SUS04-BP03 Utiliser des politiques pour gérer le cycle de vie de vos ensembles de données

- SUS04-BP04 Utiliser l'élasticité et l'automatisation pour étendre le stockage par blocs ou le système de fichiers
- SUS04-BP05 Supprimer les données inutiles ou redondantes
- SUS04-BP06 Utiliser des systèmes de fichiers partagés ou le stockage pour accéder aux données courantes
- SUS04-BP07 Réduire le mouvement des données entre les réseaux
- SUS04-BP08 Sauvegarder des données uniquement lorsqu'elles sont difficiles à recréer

SUS04-BP01 Mettre en œuvre une politique de classification des données

Classifiez les données pour identifier leur criticité vis-à-vis des résultats opérationnels, et choisissez le niveau de stockage économe en énergie approprié pour stocker les données.

Anti-modèles courants :

- Vous n'identifiez pas les ressources de données actuellement traitées ou stockées ayant des caractéristiques similaires (comme la sensibilité, la criticité métier ou les exigences réglementaires).
- Vous n'avez pas implémentée de catalogue de données pour inventorier vos ressources de données.

Avantages liés au respect de cette bonne pratique : La mise en œuvre d'une politique de classification des données vous permet d'identifier le niveau de stockage le plus économe en énergie pour les données.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

La classification des données implique d'identifier les types de données actuellement traitées ou stockées dans un système d'information détenu ou exploité par une organisation. Elle implique également de déterminer la criticité des données et l'impact possible d'une compromission, d'une perte ou d'une mauvaise utilisation de ces données.

Mettez en œuvre la politique de classification des données en partant de l'utilisation contextuelle des données et en créant un schéma de catégorisation qui prend en compte le niveau de criticité d'un jeu de données déterminé vis-à-vis des opérations d'une organisation.

Étapes d'implémentation

- Procédez à l'inventaire des différents types de données qui existent pour votre charge de travail.
  - Pour en savoir plus sur les catégories de classification des données, consultez le livre blanc Classification des données.
- Déterminez la criticité, la confidentialité, l'intégrité et la disponibilité des données en fonction du risque vis-à-vis de l'organisation. Prenez en compte ces exigences pour regrouper les données dans l'un des niveaux de classification des données que vous adoptez.
  - En guise d'exemple, consultez Four simple steps to classify your data and secure your startup.
- Auditez régulièrement votre environnement pour identifier les données non identifiées et non classifiées, puis classifiez et identifiez-les correctement.
  - En guise d'exemple, consultez Catalogue de données et crawlers (logiciels d'indexation) dans AWS Glue.
- Établissez un catalogue de données qui propose des capacités d'audit et de gouvernance.
- Déterminez et documentez les procédures de gestion pour chaque classe de données.
- Faites appel à l'automatisation pour auditer de façon continue votre environnement à la recherche de données non identifiées et non classifiées, puis classifiez et identifiez ces données en bonne et due forme.

Ressources

Documents connexes :

- Utilisation du AWS Cloud pour la prise en charge de la classification des données
- Politiques de balises d'AWS Organizations

Vidéos connexes :

- Enabling agility with data governance on AWS

SUS04-BP02 Utiliser les technologies qui prennent en charge les modèles d'accès aux données et les modèles de stockage

Utilisez les technologies de stockage qui prennent le mieux en charge l'accès à vos données et leur stockage pour limiter le provisionnement de ressources tout en soutenant votre charge de travail.

Anti-modèles courants :

- Vous partez du principe que toutes les charges de travail ont des modèles de stockage de données et d'accès similaires.
- Vous n'utilisez qu'un seul niveau de stockage, partant du principe que toutes les charges de travail s'intègrent dans ce niveau.
- Vous partez du principe que les modèles d'accès aux données n'évolueront pas dans le temps.

Avantages liés au respect de cette bonne pratique : en choisissant et en optimisant vos technologies de stockage en fonction des modèles d'accès aux données et de stockage, vos besoins métier demanderont moins de ressources cloud et vous améliorerez l'efficace globale de votre charge de travail cloud.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

Choisissez la solution de stockage la mieux adaptée à vos modèles d'accès ou envisagez de changer vos modèles d'accès en fonction de la solution de stockage pour optimiser les performances.

- Évaluez les caractéristiques de vos données et votre modèle d'accès pour déterminer les caractéristiques clés de vos besoins de stockage. Les caractéristiques clés à prendre en considération sont les suivantes :
  - Type de données : structurées, semi-structurées, non structurées
  - Croissance des données : limitée, illimitée
  - Durabilité des données : persistantes, éphémères, transitoires
  - Modèles d'accès : lecture ou écriture, fréquence, irrégularité, constance
- Procédez à la migration des données vers une technologie de stockage appropriée prenant en charge les caractéristiques de vos données ainsi que votre modèle d'accès. Voici quelques exemples de technologies de stockage AWS et leurs caractéristiques clés :

Type	Technologie	Principales caractéristiques
Stockage d'objets	Amazon S3	Service de stockage d'objets offrant une capacité de mise à l'échelle illimitée, une haute disponibilité et plusieurs

Type	Technologie	Principales caractéristiques
		options d'accessibilité. Pour transférer des objets et accéder à des objets dans et en dehors d'Amazon S3, utilisez un service, tel que Transfer Acceleration ou Points d'accèspour prendre en charge votre emplaceme nt, vos besoins en sécurité et les modèles d'accès.
Archivage et stockage	Amazon S3 Glacier	Classe de stockage d'Amazon S3 conçue pour l'archivage de données.
Système de fichiers partagé	Amazon Elastic File System (Amazon EFS)	Système de fichiers montable accessible à plusieurs types de solutions de calcul. Amazon EFS augmente et diminue automatiquement la capacité de stockage et est optimisé pour offrir des latences faibles et constante s.

Type	Technologie	Principales caractéristiques
Système de fichiers partagé	Amazon FSx	Créé sur les dernières solutions de calcul AWS pour prendre en charge quatre systèmes de fichiers fréquemment utilisés : NetApp ONTAP, OpenZFS, Windows File Serve et Lustre. La latence, le débit et les IOPS Amazon FSx varient par système de fichiers et doivent être pris en compte lorsque vous sélectionnez le système de fichiers adapté aux besoins de vos charges de travail.
Stockage par blocs	Amazon Elastic Block Store (Amazon EBS)	Service de stockage par bloc hautement performant et capable de mise à l'échelle conçu pour Amazon Elastic Compute Cloud (Amazon EC2). Amazon EBS propose un stockage basé sur SSD pour les charges de travail transactionnelles et exigeante s en termes d'IOPS ainsi qu'un stockage basé sur HDD pour les charges de travail exigeantes en termes de débit.

Type	Technologie	Principales caractéristiques
Base de données relationnelle	Amazon Aurora, Amazon RDS, Amazon Redshift	Conçues pour prendre en charge les transactions ACID (atomicité, cohérence, isolation et durabilité) et maintenir l'intégrité référentielle et la cohérence des données. De nombreuses applications traditionnelles, la planification des ressources d'entreprise (ERP), la gestion de la relation client (CRM) et des systèmes d'e-commerce utilisent des bases de données relationnelles pour stocker leurs données.
Base de données clé-valeur	Amazon DynamoDB	Optimisées pour les modèles d'accès courants, généralement pour stocker et récupérer de gros volumes de données. Les applications Web à trafic élevé, les systèmes d'e-commerce et les applications de jeu sont des cas d'utilisation typiques pour les bases de données de valeurs-clés.

- Pour les systèmes de stockage à taille fixe, comme Amazon EBS ou Amazon FSx, surveillez l'espace de stockage disponible et automatisez l'allocation de stockage dès qu'un seuil est atteint. Vous pouvez utiliser Amazon CloudWatch pour collecter et analyser différentes métriques pour Amazon EBS et Amazon FSx.
- Il est possible de configurer les classes de stockage Amazon S3 au niveau de l'objet, et un même compartiment peut contenir des objets stockés dans toutes les classes de stockage.
- Vous pouvez également utiliser des stratégies de cycle de vie Amazon S3 pour faire passer automatiquement des objets d'une classe de stockage vers une autre ou supprimer des données

sans aucune modification au niveau de l'application. Ces mécanismes de stockage vous imposent généralement de faire un compromis entre l'efficacité des ressources, la latence d'accès et la fiabilité.

Ressources

Documents connexes :

- Types de volume Amazon EBS
- Stockage d'instance Amazon EC2
- Amazon S3 Intelligent-Tiering
- Caractéristiques d'E/S Amazon EBS
- Utilisation des classes de stockage Amazon S3
- Qu'est-ce qu'Amazon S3 Glacier ?

Vidéos connexes :

- Architectural Patterns for Data Lakes on AWS
- Deep dive on Amazon EBS (STG303-R1)
- Optimize your storage performance with Amazon S3 (STG343)
- Building modern data architectures on AWS

Exemples connexes :

- Pilote CSI Amazon EFS
- Pilote CSI Amazon EBS
- Utilitaires Amazon EFS
- Amazon EBS Autoscale
- Exemples Amazon S3

SUS04-BP03 Utiliser des politiques pour gérer le cycle de vie de vos ensembles de données

Gérez le cycle de vie de toutes vos données et appliquez automatiquement la suppression pour réduire au minimum le stockage total requis pour votre charge de travail.

Anti-modèles courants :

- Vous supprimez manuellement les données.
- Vous ne supprimez aucune donnée de vos charges de travail.
- Vous ne déplacez pas les données vers des niveaux de stockage plus écoénergétiques en fonction de leurs exigences de conservation et d'accès.

Avantages liés au respect de cette bonne pratique : l'utilisation de politiques de cycle de vie des données assure un accès et une conservation efficaces des données dans une charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Les exigences en matière de conservation et d'accès des jeux de données varient généralement au cours de leur cycle de vie. Par exemple, votre application peut nécessiter un accès fréquent à certains jeux de données pendant une période limitée. Après cela, ces jeux de données sont rarement consultés.

Pour gérer efficacement vos jeux de données tout au long de leur cycle de vie, configurez les politiques de cycle de vie, des règles qui définissent la gestion de vos jeux de données.

Avec les règles de configuration du cycle de vie, vous pouvez demander au service de stockage spécifique de transférer un jeu de données vers des niveaux de stockage plus écoénergétiques, de l'archiver ou de le supprimer.

Étapes d'implémentation

- Classifiez les jeux de données dans votre charge de travail.
- Définissez les procédures de gestion pour chaque classe de données.
- Définissez des politiques de cycle de vie automatisées pour appliquer des règles de cycle de vie. Voici quelques exemples de la configuration des politiques de cycle de vie automatisé pour différents services de stockage AWS :

Storage service	How to set automated lifecycle policies
Amazon S3	Vous pouvez utiliser le cycle de vie Amazon S3 afin de gérer vos objets au cours de leur

Storage service	How to set automated lifecycle policies
	cycle de vie. Si vos modèles d'accès sont inconnus, changeants ou imprévisibles, vous pouvez utiliser Amazon S3 Intelligent-Tiering, qui surveille les modèles d'accès et déplace automatiquement les objets qui n'ont pas été consultés aux niveaux d'accès à moindre coût. Vous pouvez utiliser les métriques de Amazon S3 Storage Lens afin d'identifier les possibilités d'optimisation et les écarts dans la gestion du cycle de vie.
Amazon Elastic Block Store	Vous pouvez utiliser Amazon Data Lifecycle Manager afin d'automatiser la création, la conservation et la suppression des instantanés Amazon EBS et des AMI basées sur Amazon EBS.
Amazon Elastic File System	La gestion du cycle de vie Amazon EFS gère automatiquement le stockage des fichiers pour vos systèmes de fichiers.
Amazon Elastic Container Registry	Les politiques de cycle de vie Amazon ECR automatisent le nettoyage de vos images de conteneur en faisant expirer des images en fonction de leur ancienneté ou de leur nombre.
AWS Elemental MediaStore	Vous pouvez utiliser une politique de cycle de vie des objets qui régit la façon dont les objets longs doivent être stockés dans un conteneur MediaStore.

- Supprimez les volumes, les instantanés et les données inutilisés qui dépassent leur période de conservation. Utilisez des fonctionnalités de service natives telles que la durée de vie Amazon DynamoDB ou la conservation de journal Amazon CloudWatch pour la suppression.
- Regroupez et compressez les données le cas échéant en fonction des règles de cycle de vie.

Ressources

Documents connexes :

- Optimize your Amazon S3 Lifecycle rules with Amazon S3 Storage Class Analysis (Optimiser vos règles de cycle de vie Amazon S3 avec Amazon S3)
- Evaluating Resources with AWS Config Rules (Évaluation des ressources avec les règles AWS Config)

Vidéos connexes :

- Simplify Your Data Lifecycle and Optimize Storage Costs With Amazon S3 Lifecycle
- Reduce Your Storage Costs Using Amazon S3 Storage Lens

SUS04-BP04 Utiliser l'élasticité et l'automatisation pour étendre le stockage par blocs ou le système de fichiers

Utilisez l'élasticité et l'automatisation pour étendre le stockage par blocs ou le système de fichiers au fur et à mesure que le volume de données augmente afin de minimiser le stockage total provisionné.

Anti-modèles courants :

- Vous provisionnez un grand bloc de stockage ou un grand système de fichiers pour vos besoins futurs.
- Vous surprovisionnez les opérations d'entrée et de sortie par seconde (IOPS) de votre système de fichiers.
- Vous ne contrôlez pas l'utilisation de vos volumes de données.

Avantages liés au respect de cette bonne pratique : minimiser le provisionnement excessif du système de stockage réduit les ressources inutilisées et améliore l'efficacité globale de votre charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Créez des systèmes de stockage par blocs et des systèmes de fichiers avec une allocation de taille, un débit et une latence adaptés à votre charge de travail. Utilisez l'élasticité et l'automatisation pour

étendre le stockage par blocs ou le système de fichiers en fonction de la croissance des données sans avoir à provisionner ces services de stockage de manière excessive.

Étapes d'implémentation

- Pour le stockage à taille fixe, comme par exemple Amazon EBS, vérifiez que vous surveillez la quantité de stockage utilisée par rapport à la taille globale du stockage et créez une automatisation, si possible, pour augmenter la taille du stockage lorsqu'un seuil est atteint.
- Utilisez des volumes Elastic et des services de données par bloc gérés pour automatiser l'allocation de stockage supplémentaire à mesure que vos données persistantes augmentent. Par exemple, vous pouvez utiliser lesVolumes Amazon EBS Elastic pour modifier la taille et le type de volume, ou ajuster les performances de vos volumes Amazon EBS.
- Choisissez la bonne classe de stockage, le bon mode de performance et le mode de débit adapté à votre système de fichiers afin de répondre aux besoins de votre entreprise, sans les dépasser.
  - Performances Amazon EFS
  - Performances des volumes Amazon EBS sur les instances Linux
- Définissez des niveaux cibles d'utilisation des volumes de données et redimensionnez les volumes en dehors des plages attendues.
- Dimensionnez correctement les volumes en lecture seule en fonction des données.
- Migrez les données vers des magasins d'objets pour éviter d'allouer la capacité excédentaire des tailles de volume fixes vers le stockage par bloc.
- Examinez régulièrement les volumes Elastic et les systèmes de fichiers pour mettre fin aux volumes inutilisés et réduire les ressources surprovisionnées pour les adapter à la taille actuelle des données.

Ressources

Documents connexes :

- Documentation Amazon FSx
- Qu'est-ce qu'Amazon Elastic File System ?

Vidéos connexes :

- Deep Dive on Amazon EBS Elastic Volumes [Découverte approfondie d'Elastic Block Storage (Amazon EBS)]

- Amazon EBS and Snapshot Optimization Strategies for Better Performance and Cost Savings (Stratégies d'optimisation d'Amazon EBS et des instantanés pour de meilleures performances et des économies de coûts)
- Optimiser Amazon EFS en matière de coût et de performance, à l'aide des bonnes pratiques

SUS04-BP05 Supprimer les données inutiles ou redondantes

Supprimez les données inutiles ou redondantes pour minimiser les ressources de stockage requises pour stocker vos jeux de données.

Anti-modèles courants :

- Vous dupliquez des données qui peuvent être facilement obtenues ou recréées.
- Vous sauvegardez toutes les données sans tenir compte de leur criticité.
- Vous ne supprimez les données que de façon irrégulière, sur les événements opérationnels ou pas du tout.
- Vous stockez les données de manière redondante, quelle que soit la durabilité du service de stockage.
- Vous activez la gestion des versions Amazon S3 sans justification professionnelle.

Avantages liés au respect de cette pratique : la suppression des données inutiles réduit la taille de stockage requise pour votre charge de travail et l'impact environnemental de la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Ne stockez pas les données dont vous n'avez pas besoin. Automatisez la suppression des données inutiles. Utilisez des technologies qui dédupliquent les données au niveau du fichier et du bloc. Utilisez la réplication des données native et les fonctionnalités de redondance des services.

Étapes d'implémentation

- Évaluez si vous pouvez éviter de stocker les données en utilisant des jeux de données disponibles pour le public dans AWS Data Exchange et des données ouvertes sur AWS.
- Utilisez des mécanismes qui peuvent dédupliquer les données au niveau du bloc et de l'objet. Voici quelques exemples de déduplication des données sur AWS :

Storage service	Deduplication mechanism
Amazon S3	Utilisez AWS Lake Formation FindMatches afin de trouver des enregistrements correspondants dans un jeu de données (y compris ceux sans identifiants) en utilisant la nouvelle transformation ML FindMatches.
Amazon FSx	Activez la déduplication des données sur Amazon FSx for Windows.
Instantanés Amazon Elastic Block Store	Les instantanés sont des sauvegardes incrémentielles, ce qui signifie que seuls les blocs de l'appareil qui ont changé après votre instantané le plus récent sont enregistrés.

- Analysez l'accès aux données pour identifier les données inutiles. Automatisez les politiques de cycle de vie. Utilisez des fonctionnalités de service natives telles que la durée de vie Amazon DynamoDB, le cycle de vie Amazon S3 ou la conservation des journaux Amazon CloudWatch pour la suppression.
- Utilisez les capacités de virtualisation des données sur AWS afin de maintenir les données à leur source et d'éviter leur duplication.
  - Virtualisation des données natives du cloud sur AWS
  - Lab: Optimize Data Pattern Using Amazon Redshift Data Sharing (Atelier : optimiser le modèle de données à l'aide du partage de données)
- Utilisez une technologie de sauvegarde qui peut réaliser des sauvegardes incrémentielles.
- Utilisez la durabilité de Amazon S3 et la réplication d'Amazon EBS pour atteindre vos objectifs de durabilité au lieu des technologies autogérées (comme un tableau redondant de disques indépendants (RAID)).
- Centralisez les données de journalisation et de suivi, dédupliquez les entrées de journal identiques et établissez des mécanismes pour ajuster le niveau d'informations transmises, le cas échéant.
- Préremplissez les caches uniquement lorsque cela est justifié.
- Établissez la surveillance et l'automatisation des caches pour redimensionner correctement les caches.

- Supprimez les déploiements et les ressources obsolètes des magasins d'objets et des caches périphériques lors de la transmission des nouvelles versions de votre charge de travail.

Ressources

Documents connexes :

- Modification de la conservation des données de journaux dans CloudWatch Logs
- Data deduplication on Amazon FSx for Windows File Server(Déduplication des données sur Amazon Fsx for Windows File Server)
- Features of Amazon FSx for ONTAP including data deduplication (Fonctions d'Amazon FSx pour ONTAP qui incluent la déduplication des données)
- Invalidation de fichiers sur Amazon CloudFront
- Using AWS Backup to back up and restore Amazon EFS file systems (Utilisation d'AWS Backup pour sauvegarder et restaurer les systèmes de fichiers Amazon EFS)
- Qu'est-ce qu'Amazon CloudWatch Logs ?
- Utilisation des sauvegardes sur Amazon RDS

Vidéos connexes :

- Fuzzy Matching and Deduplicating Data with ML Transforms for AWS Lake Formation

Exemples connexes :

- Comment analyser les journaux d'accès au serveur Amazon S3 à l'aide d'Amazon Athena ?

SUS04-BP06 Utiliser des systèmes de fichiers partagés ou le stockage pour accéder aux données courantes

Adoptez des systèmes de fichiers ou de stockage partagés pour éviter la duplication des données et permettre une infrastructure plus efficace pour votre charge de travail.

Anti-modèles courants :

- Vous mettez en service le stockage pour chaque client individuel.
- Vous ne détachez pas le volume de données des clients inactifs.

- Vous ne fournissez pas d'accès au stockage pour les plateformes et les systèmes.

Avantages liés au respect de cette bonne pratique : l'utilisation de systèmes de fichiers ou de stockage partagés permet de partager des données à un ou plusieurs consommateurs sans avoir à copier les données. Cela permet de réduire les ressources de stockage nécessaires à la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Si plusieurs utilisateurs ou applications accèdent aux mêmes jeux de données, l'utilisation de la technologie de stockage partagé est cruciale pour mettre en œuvre une infrastructure efficace pour votre charge de travail. La technologie de stockage partagé fournit un emplacement central pour stocker et gérer les jeux de données et éviter la duplication des données. Elle assure également la cohérence des données entre les différents systèmes. En outre, la technologie de stockage partagé permet d'utiliser plus efficacement la puissance de calcul, car plusieurs ressources informatiques peuvent accéder aux données et les traiter simultanément en parallèle.

Ne récupérez les données de ces services de stockage partagé qu'en fonction des besoins et détachez les volumes inutilisés pour libérer des ressources.

Étapes d'implémentation

- Migrez les données vers le stockage partagé lorsque les données ont plusieurs consommateurs. Voici quelques exemples de technologie de stockage partagé sur AWS :

Storage option	When to use
Amazon EBS Multi-Attach	Amazon EBS Multi-Attach vous permet d'attacher un seul volume SSD IOPS provisionnés (io1 ou io2) à plusieurs instances qui se trouvent dans la même zone de disponibilité.
Amazon EFS	Consultez When to Choose Amazon EFS (Quand choisir Amazon EFS).
Amazon FSx	Consultez Choisir un système de fichiers Amazon FSx.

Storage option	When to use
Amazon S3	Les applications qui ne nécessitent pas de structure de système de fichiers et qui sont conçues pour fonctionner avec le stockage d'objet peuvent utiliser Amazon S3 comme une solution de stockage d'objet massivement évolutive, durable et peu coûteuse.

- Copiez des données vers ou récupérez des données depuis des systèmes de fichiers partagés uniquement si nécessaire. Par exemple, vous pouvez créer un système de fichiers Amazon FSx for Lustre soutenu par Amazon S3 et ne charger que le sous-ensemble de données nécessaires au traitement des tâches vers Amazon FSx.
- Supprimez les données selon vos modèles d'utilisation, comme indiqué dans SUS04-BP03 Utiliser des politiques pour gérer le cycle de vie de vos ensembles de données.
- Détachez les volumes des clients qui ne les utilisent pas activement.

Ressources

Documents connexes :

- Linking your file system to an Amazon S3 bucket (Liaison de votre système de fichiers à un compartiment S3)
- Using Amazon EFS for AWS Lambda in your serverless applications (Utilisation d'Amazon EFS pour AWS Lambda dans vos applications sans serveur)
- Amazon EFS Intelligent-Tiering Optimizes Costs for Workloads with Changing Access Patterns (Amazon EFS Intelligent-Tiering optimise les coûts pour les charges de travail avec des modèles d'accès évolutifs)
- Using Amazon FSx with your on-premises data repository (Utilisation d'Amazon FSx avec votre référentiel de données sur site)

Vidéos connexes :

- Optimisation des coûts de stockage avec Amazon EFS

SUS04-BP07 Réduire le mouvement des données entre les réseaux

Utilisez des systèmes de fichiers partagés ou un stockage objet pour accéder aux données communes et minimiser les ressources réseau totales requises pour prendre en charge le déplacement des données de votre charge de travail.

Anti-modèles courants :

- Vous stockez toutes les données dans la même Région AWS, indépendamment de l'endroit où se trouvent les utilisateurs des données.
- Vous n'optimisez ni la taille ni le format des données avant de les déplacer sur le réseau.

Avantages liés au respect de cette bonne pratique : L'optimisation du déplacement des données sur le réseau réduit les ressources réseau totales nécessaires à la charge de travail et diminue son impact environnemental.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Le déplacement des données dans votre entreprise nécessite des ressources de calcul, de réseau et de stockage. Utilisez des techniques pour minimiser les déplacements de données et améliorer l'efficacité globale de votre charge de travail.

Étapes d'implémentation

- Considérez la proximité des données ou des utilisateurs comme un facteur de décision lors de la sélection d'une région pour votre charge de travail.
- Divisez les services consommés par région afin que les données spécifiques à une région soient stockées dans celle où elles sont consommées.
- Utilisez des formats de fichiers efficaces (tels que Parquet ou ORC) et compressez les données avant de les déplacer sur le réseau.
- Ne déplacez pas les données inutilisées. Voici quelques exemples qui peuvent vous aider à éviter de déplacer des données inutilisées :
  - Réduisez les réponses de l'API aux seules données pertinentes.
  - Agrégez les données lorsqu'elles sont détaillées (les informations au niveau de l'enregistrement ne sont pas requises).

  - Consultez Atelier Well-Architected : optimiser le modèle de données à l'aide du partage de données Amazon Redshift.
  - Envisagez le partage de données entre comptes dans AWS Lake Formation.
- Utilisez des services qui peuvent vous aider à exécuter du code au plus près des utilisateurs de votre charge de travail.

Service	Quand l'utiliser
Lambda@Edge	Utilisez ce service pour les opérations exigeantes en puissance de calcul qui sont exécutées lorsque des objets ne sont pas dans le cache.
Fonctions CloudFront	Utilisez ce service pour les cas d'utilisation simples comme une demande HTTP(S) ou des manipulations de réponse pouvant être lancées par des fonctions brèves.
AWS IoT Greengrass	Utilisez ce service pour exécuter du calcul local, une messagerie et une mise en cache de données pour les appareils connectés.

Ressources

Documents connexes :

- Optimisation de votre infrastructure AWS pour la durabilité, partie 3 : mise en réseau
- Infrastructure mondiale AWS
- Fonctionnalités clés d'Amazon CloudFront, y compris le réseau périphérique mondial CloudFront
- Compression des requêtes HTTP dans Amazon OpenSearch Service
- Compression intermédiaire de données avec Amazon EMR
- Chargement de fichiers de données compressés depuis Amazon S3 vers Amazon Redshift
- Diffusion de fichiers compressés avec Amazon CloudFront

Vidéos connexes :

- Demystifying data transfer on AWS

Exemples connexes :

- Une architecture axée sur la durabilité : réduisez les mouvements de données sur les réseaux

SUS04-BP08 Sauvegarder des données uniquement lorsqu'elles sont difficiles à recréer

Évitez de sauvegarder les données qui n'ont aucune valeur commerciale afin de minimiser les besoins en ressources de stockage pour votre charge de travail.

Anti-modèles courants :

- Vous n'avez aucune stratégie de sauvegarde en place pour vos données.
- Vous sauvegardez des données qui peuvent être facilement recréées.

Avantages liés au respect de cette bonne pratique : le fait d'éviter la sauvegarde de données non critiques réduit les ressources de stockage nécessaires à la charge de travail et diminue son impact environnemental.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Le fait d'éviter la sauvegarde de données inutiles peut contribuer à réduire les coûts et les ressources de stockage utilisées par la charge de travail. Sauvegardez uniquement les données ayant une valeur opérationnelle ou nécessaires pour répondre aux exigences en matière de conformité. Examinez les politiques de sauvegarde et excluez tout magasin éphémère n'apportant aucune valeur dans un scénario de récupération.

Étapes d'implémentation

- Mettez en œuvre la politique de classification des données comme indiqué dans SUS04-BP01 Mettre en œuvre une politique de classification des données.
- Utilisez la criticité de la classification de vos données et concevez une stratégie de sauvegarde basée sur votre objectif de délai de reprise (RTO) et votre objectif de point de reprise (RPO). Évitez de sauvegarder les données non critiques.
  - Excluez les données qui peuvent être facilement recréées.

  - Excluez les données éphémères de vos sauvegardes.
  - Excluez les copies locales des données, sauf si le temps nécessaire pour restaurer ces données à partir d'un emplacement commun dépasse vos accords de niveau de service (SLA).
- Utilisez une solution automatisée ou un service géré pour sauvegarder les données essentielles à l'entreprise.
  - AWS Backup est un service entièrement géré qui permet de centraliser et d'automatiser facilement la protection des données entre les services AWS, dans le cloud et sur site. Pour obtenir des conseils pratiques sur la façon de créer des sauvegardes automatisées à l'aide de AWS Backup, consultez Well-Architected Labs - Testing Backup and Restore of Data (Ateliers Well-Architected : test de sauvegarde et de restauration des données).
  - Automate backups and optimize backup costs for Amazon EFS using AWS Backup (Automatisation des sauvegardes et optimisation des coûts de sauvegarde pour Amazon EFS avec AWS Backup).

Ressources

Bonnes pratiques associées :

- REL09-BP01 Identifier et sauvegarder toutes les données qui doivent être sauvegardées, ou reproduire les données à partir de sources
- REL09-BP03 Effectuer automatiquement la sauvegarde des données
- REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise

Documents connexes :

- Using AWS Backup to back up and restore Amazon EFS file systems (Utilisation d'AWS Backup pour sauvegarder et restaurer les systèmes de fichiers Amazon EFS)
- Instantanés Amazon EBS
- Utilisation des sauvegardes sur Amazon Relational Database Service
- Partenaire APN : partenaires pouvant faciliter la sauvegarde
- AWS Marketplace : produits pouvant être utilisés pour la sauvegarde
- Sauvegarde d'Amazon EFS
- Sauvegarde d'Amazon FSx for Windows File Server
- Sauvegarde et restauration pour Amazon ElastiCache (Redis OSS)

Vidéos connexes :

- AWS re:Invent 2021 - Backup, disaster recovery, and ransomware protection with AWS (AWS re:Invent 2021 - Sauvegarde, reprise après sinistre et protection contre les rançongiciels avec AWS)
- AWS Backup Demo: Cross-Account and Cross-Region Backup (Démonstration de la sauvegarde AWS : sauvegarde intercompte et inter-régions)
- AWS re:Invent 2019: Deep dive on AWS Backup, ft. (AWS re:Invent 2019 : immersion dans AWS Backup, ft.) Rackspace (STG341)

Exemples connexes :

- Atelier Well-Architected : test de la sauvegarde et de la restauration de données
- Atelier Well-Architected : sauvegarde et restauration avec basculement automatique pour la charge de travail d'analyse
- Atelier Well-Architected : reprise après sinistre - sauvegarde et restauration

Matériel et services

Question

- SUS 5 Comment choisissez-vous et utilisez-vous le matériel et les services du cloud dans votre architecture pour soutenir vos objectifs de durabilité ?

SUS 5 Comment choisissez-vous et utilisez-vous le matériel et les services du cloud dans votre architecture pour soutenir vos objectifs de durabilité ?

Recherchez des possibilités de réduire les impacts en matière de durabilité de la charge de travail en modifiant vos pratiques de gestion du matériel. Réduisez la quantité de matériel nécessaire à allouer et à déployer, et sélectionnez le matériel et les services les plus efficaces pour votre charge de travail individuelle.

Bonnes pratiques

- SUS05-BP01 Utiliser la quantité minimale de matériel pour répondre à vos besoins
- SUS05-BP02 Utiliser des types d'instance ayant le moins d'impact
- SUS05-BP03 Utiliser des services gérés

- SUS05-BP04 Optimiser votre utilisation des accélérateurs de calcul matériels

SUS05-BP01 Utiliser la quantité minimale de matériel pour répondre à vos besoins

Utilisez la quantité minimale de matériel pour votre charge de travail afin de répondre efficacement aux besoins de votre entreprise.

Anti-modèles courants :

- Vous ne surveillez pas l'utilisation des ressources.
- Vous disposez de ressources avec un faible niveau d'utilisation dans votre architecture.
- Vous n'examinez pas l'utilisation du matériel statique pour déterminer s'il doit être redimensionné.
- Vous ne fixez pas d'objectifs d'utilisation du matériel pour votre infrastructure informatique en fonction des indicateurs clés de performance de l'entreprise.

Avantages liés au respect de cette bonne pratique : le redimensionnement de vos ressources cloud permet de réduire l'impact environnemental d'une charge de travail, d'économiser de l'argent et de maintenir les références de performance.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Sélectionnez de manière optimale le nombre total de composants matériels requis pour votre charge de travail afin d'améliorer son efficacité globale. Le AWS Cloud vous apporte la flexibilité dont vous avez besoin pour développer ou réduire le nombre de ressources de manière dynamique par le biais de divers mécanismes, tels que AWS Auto Scaling, et de répondre aux variations de la demande. Il fournit également des API et des kits SDK qui permettent de modifier les ressources avec un minimum d'effort. Utilisez ces capacités pour apporter des modifications fréquentes à vos mises en œuvre de charges de travail. En outre, utilisez les directives de redimensionnement des outils AWS pour exploiter efficacement votre ressource cloud et répondre aux besoins de votre entreprise.

Étapes d'implémentation

- Choisissez le type d'instances qui correspond le mieux à vos besoins.
  - Comment choisir le type d'instance Amazon EC2 EC2 approprié pour mon application ?
  - Sélection de type d'instance basée sur des attributs pour la flotte Amazon EC2.

    - Créer un groupe Auto Scaling en utilisant la sélection du type d'instance basée sur des attributs.
- Diminuez les charges de travail variables par petits paliers.
- Utilisez plusieurs options d'achat de calcul afin d'équilibrer la flexibilité, la capacité de mise à l'échelle et la réduction des coûts des instances.
    - Les instances à la demande sont les mieux adaptées aux charges de travail nouvelles, à état constant et fluctuantes qui ne peuvent pas être flexibles en termes de type d'instance, de lieu ou de temps.
    - Les instances Spot sont un excellent moyen de compléter les autres options pour les applications qui sont tolérantes aux pannes et flexibles.
    - Tirez parti des Compute Savings Plans pour les charges de travail stables qui permettent une certaine flexibilité si vos besoins (comme une AZ, une région, des familles d'instances ou des types d'instances) changent.
- Utilisez la diversité des instances et des zones de disponibilité pour maximiser la disponibilité des applications et tirer parti de la capacité excédentaire lorsque cela est possible.
- Utilisez les recommandations de redimensionnement des outils AWS pour faire des ajustements sur votre charge de travail.
    - AWS Compute Optimizer
    - AWS Trusted Advisor
- Négociez des SLA qui permettent une réduction temporaire de la capacité, et laissez l'automatisation déployer des ressources de remplacement.

Ressources

Documents connexes :

- Optimisation de votre infrastructure AWS pour la durabilité, partie 1 : calcul
- Sélection de type d'instance basée sur des attributs pour Auto Scaling pour la flotte Amazon EC2
- Documentation AWS Compute Optimizer
- Utilisation de Lambda : optimisation de la performance
- Documentation sur la scalabilité automatique

Vidéos connexes :

- Concevoir un environnement de calcul rentable, économe en énergie et en ressources

Exemples connexes :

- Well-Architected Lab: Rightsizing with AWS Compute Optimizer and Memory Utilization Enabled (Level 200) [Atelier Well-Architected : dimensionnement avec activation de Compute Optimizer et de l'utilisation de la mémoire (niveau 200)]

SUS05-BP02 Utiliser des types d'instance ayant le moins d'impact

Contrôlez et utilisez en permanence de nouveaux types d'instances pour tirer parti des améliorations de l'efficacité énergétique.

Anti-modèles courants :

- Vous n'utilisez qu'une seule famille d'instances.
- Vous n'utilisez que des instances x86.
- Vous spécifiez un type d'instance dans votre configuration Amazon EC2 Auto Scaling.
- Vous utilisez des instances AWS de manière non conforme à leur utilisation prévue (par exemple, vous utilisez des instances optimisées pour le calcul pour une charge de travail exigeante en mémoire).
- Vous n'évaluez pas régulièrement de nouveaux types d'instance.
- Vous ne vérifiez pas les recommandations des outils de redimensionnement AWS tels que AWS Compute Optimizer.

Avantages liés au respect de cette bonne pratique : En utilisant des instances économes en énergie et dimensionnées, vous pouvez grandement réduire l'impact sur l'environnement et le coût de votre charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

L'utilisation d'instances efficaces dans les charges de travail du cloud est cruciale pour réduire l'utilisation des ressources et pour une meilleure rentabilité. Contrôlez de façon continue le lancement de nouveaux types d'instances et profitez d'améliorations de l'efficacité énergétique, dont ces types d'instances conçus pour soutenir des charges de travail spécifiques comme l'entraînement et l'inférence du machine learning et le transcodage vidéo.

Étapes d'implémentation

- Découvrez et explorez les types d'instance capables de réduire l'impact sur l'environnement de votre charge de travail.
  - Abonnez-vous à Nouveautés AWS pour vous tenir informé des dernières technologies et instances AWS.
  - Découvrez les différents types d'instance AWS.
  - Découvrez les instances AWS basées sur Graviton qui offrent les meilleures performances en matière de consommation énergétique dans Amazon EC2 en regardant re:Invent 2020 - Deep dive on AWS Graviton2 processor-powered Amazon EC2 instances et Deep dive into AWS Graviton3 and Amazon EC2 C7g instances.
- Planifiez et migrez votre charge de travail vers les types d'instance avec le moins d'impact.
  - Définissez un processus pour évaluer les nouvelles fonctionnalités ou instances pour votre charge de travail. Profitez de l'agilité du cloud pour tester rapidement en quoi les nouveaux types d'instance peuvent améliorer la durabilité environnementale de votre charge de travail. Utilisez des métriques de proxy pour mesurer le nombre de ressources nécessaires pour mener à bien une unité de travail.
  - Si possible, modifiez votre charge de travail pour qu'elle fonctionne avec différents nombres de processeurs et différentes quantités de mémoire afin de maximiser votre choix de type d'instance.
  - Envisagez de migrer votre charge de travail vers des instances basées sur Graviton pour améliorer l'efficacité des performances de votre charge de travail.
    - AWS Graviton Fast Start
    - Éléments à considérer lors de la migration des charges de travail vers les instances AWS basées sur Amazon Elastic Compute Cloud Graviton
    - AWS Graviton2 for ISVs
  - Envisagez de sélectionner l'option AWS Graviton lorsque vous utilisez des services gérés par AWS.
  - Migrez votre charge de travail vers des régions qui offrent des instances ayant un impact moindre en matière de durabilité et qui répondent à vos exigences métier.
  - Pour les charges de travail de machine learning, tirez parti d'un matériel conçu spécialement pour votre charge de travail, comme AWS Trainium, AWS Inferentiaet Amazon EC2 DL1. Les instances AWS Inferentia telles que les instances Inf2 offrent des performances par watt jusqu'à 50 % supérieures à celles des instances Amazon EC2 comparables.

    - Utilisez Amazon SageMaker Inference Recommender pour redimensionner le point de terminaison de l'inférence de machine learning.
    - Pour les pics de charges de travail (charges de travail aux besoins de capacité supplémentaire irréguliers), utilisez des instances à performances extensibles.
    - Pour les charges de travail sans état et tolérantes aux pannes, utilisez des instances Spot Amazon EC2 pour augmenter l'utilisation globale du cloud et réduire l'impact en matière de durabilité des ressources inutilisées.
- Exploitez et optimisez votre instance de charge de travail.
    - Pour les charges de travail éphémères, évaluez les métriques d'instance Amazon CloudWatch telles que `CPUUtilization` pour identifier si l'instance est inactive ou sous-exploitée.
    - Pour les charges de travail stables, vérifiez les outils de redimensionnement AWS tels que AWS Compute Optimizer à intervalles réguliers pour identifier les possibilités d'optimiser et de redimensionner les instances.
        - Atelier Well-Architected : recommandations de redimensionnement
        - Atelier Well-Architected : redimensionnement avec Compute Optimizer
        - Atelier Well-Architected : optimiser les modèles matériels et observer les indicateurs de performance clés de durabilité

Ressources

Documents connexes :

- Optimisation de votre infrastructure AWS pour la durabilité, partie 1 : calcul
- AWS Graviton
- Amazon EC2 DL1
- Parcs de réserve de capacité Amazon EC2
- Parc d'instances Spot Amazon EC2
- Fonctions : configuration des fonctions Lambda
- Sélection de type d'instance basée sur des attributs pour la flotte Amazon EC2
- Création d'applications durables, efficaces et optimisées en termes de coûts sur AWS
- Comment le tableau de bord de durabilité de Contino aide les clients à optimiser leur empreinte carbone

Vidéos connexes :

- Deep dive on AWS Graviton2 processer-powered Amazon EC2 instances
- Deep dive into AWS Graviton3 and Amazon EC2 C7g instances
- Concevoir un environnement de calcul rentable, économe en énergie et en ressources

Exemples connexes :

- Solution : conseils pour l'optimisation des charges de travail de deep learning pour atteindre la durabilité sur AWS
- Atelier Well-Architected : recommandations de redimensionnement
- Atelier Well-Architected : redimensionnement avec Compute Optimizer
- Atelier Well-Architected : optimiser les modèles matériels et observer les indicateurs de performance clés de durabilité
- Atelier Well-Architected : migration des services vers Graviton

SUS05-BP03 Utiliser des services gérés

Utilisez les services gérés pour fonctionner plus efficacement dans le cloud.

Anti-modèles courants :

- Vous utilisez des instances Amazon EC2 à faible utilisation pour exécuter vos applications.
- Votre équipe interne ne fait que gérer la charge de travail, sans avoir le temps de se concentrer sur l'innovation ou les simplifications.
- Vous déployez et maintenez des technologies pour des tâches qui peuvent être exécutées plus efficacement sur des services gérés.

Avantages liés au respect de cette bonne pratique :

- L'utilisation de services gérés transfère la responsabilité vers AWS qui dispose d'informations sur des millions de clients pouvant contribuer à de nouvelles innovations et à des gains d'efficacité.
- Le service géré répartit l'impact environnemental du service entre de nombreux utilisateurs grâce aux plans de contrôle multi-réseaux.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyenne entreprise

Directives d'implémentation

Les services gérés transfèrent à AWS la responsabilité de maintenir une utilisation élevée et d'optimiser la durabilité du matériel déployé. Les services gérés suppriment également la charge opérationnelle et administrative liée à la maintenance d'un service, ce qui permet à votre équipe de disposer de plus de temps et de se concentrer sur l'innovation.

Examinez votre charge de travail pour identifier les composants qui peuvent être remplacés par des services gérés AWS. Par exemple, Amazon RDS, Amazon Redshift et Amazon ElastiCache fournissent un service de base de données géré. Amazon Athena, Amazon EMR et Amazon OpenSearch Service fournissent un service d'analytique géré.

Étapes d'implémentation

1. Dressez l'inventaire de votre charge de travail pour les services et les composants.
2. Évaluez et identifiez les composants qui peuvent être remplacés par des services gérés. Voici quelques exemples de situations dans lesquelles vous pourriez envisager de recourir à un service géré :

Task	What to use on AWS
Hébergement d'une base de données	Utilisez les instances Amazon Relational Database Service (Amazon RDS) gérées au lieu de gérer vos propres instances Amazon RDS sur Amazon Elastic Compute Cloud (Amazon EC2).
Héberger une charge de travail en conteneur	Utilisez AWS Fargate au lieu de mettre en œuvre votre propre infrastructure de conteneurs.
Hébergement d'applications Web	Utilisez AWS Amplify Hosting comme service entièrement géré de CI/CD et d'hébergement pour les sites Web statiques et les applicati ons Web rendues côté serveur.

3. Identifiez les dépendances et créez un plan de migration. Mettez à jour les runbooks et les playbooks en conséquence.

    - AWS Application Discovery Service rassemble et présente automatiquement les informations détaillées sur les dépendances et l'utilisation des applications pour vous aider à prendre des décision en connaissance de cause pour votre programme de migration
4. Testez le service avant de migrer vers le service géré.
5. Utilisez le plan de migration pour remplacer les services auto-hébergés par des services gérés.
6. Surveillez continuellement le service une fois la migration terminée afin d'apporter les modifications nécessaires et d'optimiser le service.

Ressources

Documents connexes :

- Produits AWS Cloud
- Calculateur du coût total de possession (TCO) d'AWS
- Amazon DocumentDB
- Amazon Elastic Kubernetes Service (EKS)
- Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Vidéos connexes :

- Cloud operations at scale with AWS Managed Services (Opérations de cloud à grande échelle avec AWS Managed Services)

SUS05-BP04 Optimiser votre utilisation des accélérateurs de calcul matériels

Optimisez votre utilisation des instances de calcul accéléré pour réduire les exigences d'infrastructure physique de votre charge de travail.

Anti-modèles courants :

- Vous ne surveillez pas l'utilisation du GPU.
- Vous utilisez une instance à usage général pour la charge de travail alors qu'une instance spécialement conçue peut fournir des performances supérieures, des coûts plus faibles et de meilleures performances par watt.
- Vous utilisez des accélérateurs de calcul matériels pour les tâches où ils sont plus efficaces en utilisant des alternatives basées sur l'UC.

Avantages liés au respect de cette bonne pratique : en optimisant l'utilisation des accélérateurs matériels, vous pouvez réduire les exigences de votre charge de travail en termes d'infrastructure physique.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Si vous avez besoin d'une capacité de traitement élevée, vous pouvez bénéficier de l'utilisation d'instances de calcul accéléré, qui vous donnent accès à des accélérateurs de calcul matériels tels que des unités de traitement graphique (GPU) et des matrices de portes programmables sur site (FPGA). Ces accélérateurs matériels exécutent certaines fonctions comme le traitement graphique ou la correspondance de modèles de données plus efficacement que les alternatives basées sur l'UC. De nombreuses charges de travail accélérées, telles que le rendu, le transcodage et le machine learning, sont très variables en termes d'utilisation des ressources. Exécutez ce matériel uniquement pendant le temps nécessaire et mettez-le hors service grâce à l'automatisation lorsque vous n'en avez plus besoin afin de limiter les ressources consommées.

Étapes d'implémentation

- Identifiez quelles instances informatiques accélérées peuvent répondre à vos besoins.
- Pour les charges de travail de machine learning, tirez parti d'un matériel conçu spécialement pour votre charge de travail, comme AWS Trainium, AWS Inferentiaet Amazon EC2 DL1. Les instances AWS Inferentia, telles que les instances Inf2, offrent des performances par watt supérieures de 50 % à celles des instances Amazon EC2 comparables.
- Collectez des métriques d'utilisation pour vos instances de calcul accéléré. Par exemple, vous pouvez utiliser un agent CloudWatch pour collecter des métriques comme `utilization_gpu` et `utilization_memory` pour vos GPU, comme indiqué dans Collecter les métriques des GPU NVIDIA avec Amazon CloudWatch.
- Optimisez le code, le fonctionnement du réseau et les paramètres des accélérateurs matériels pour veiller à ce que le matériel sous-jacent soit pleinement utilisé.
  - Optimisation des paramètres GPU
  - Surveillance et optimisation des GPU dans l'AMI Deep Learning
  - Optimisation des E/S pour le réglage des performances de GPU pour l'entraînement du deep learning dans Amazon SageMaker
- Utilisez les dernières bibliothèques performantes et les pilotes GPU.
- Utilisez l'automatisation pour libérer les instances GPU lorsqu'elles ne sont pas utilisées.

Ressources

Documents connexes :

- Calcul accéléré
- Passons à l'architecture Architecture avec des puces personnalisées et des accélérateurs
- Comment choisir le type d'instance Amazon EC2 approprié pour ma charge de travail ?
- Instances VT1 Amazon EC2
- Choisissez le meilleur accélérateur d'IA et la meilleure compilation de modèles pour l'inférence de vision par ordinateur avec Amazon SageMaker

Vidéos connexes :

- How to select Amazon EC2 GPU instances for deep learning
- Deploying Cost-Effective Deep Learning Inference

Processus et culture

Question

- SUS 6 Comment vos processus organisationnels soutiennent-ils vos objectifs de durabilité ?

SUS 6 Comment vos processus organisationnels soutiennent-ils vos objectifs de durabilité ?

Recherchez des opportunités de réduire votre impact en matière de durabilité modifiant vos pratiques de développement, de test et de déploiement.

Bonnes pratiques

- SUS06-BP01 Adopter des méthodes qui peuvent rapidement présenter des améliorations en matière de durabilité
- SUS06-BP02 Garder votre charge de travail à jour
- SUS06-BP03 Augmenter l'utilisation de vos environnements de compilation
- SUS06-BP04 Utiliser des tests Device Farms gérés

SUS06-BP01 Adopter des méthodes qui peuvent rapidement présenter des améliorations en matière de durabilité

Adoptez des méthodes et des processus pour valider les améliorations potentielles, minimiser les coûts des tests et apporter de petites améliorations.

Anti-modèles courants :

- L'examen de la durabilité de votre application est une tâche qui n'est effectuée qu'une seule fois au début d'un projet.
- Votre charge de travail est devenue obsolète, car le processus de lancement est trop lourd pour introduire des changements mineurs dans un souci d'efficacité des ressources.
- Vous ne disposez pas de mécanismes pour améliorer votre charge de travail afin d'atteindre davantage de durabilité.

Avantages liés au respect de cette bonne pratique : en établissant un processus pour introduire et suivre les améliorations de la durabilité, vous serez en mesure d'adopter continuellement de nouvelles fonctionnalités et capacités, de supprimer les problèmes et d'améliorer l'efficacité de la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Testez et validez les améliorations potentielles en matière de durabilité avant de les déployer en production. Tenez compte du coût des tests lors du calcul des avantages futurs potentiels d'une amélioration. Développez des méthodes d'essai à faible coût pour apporter de petites améliorations.

Étapes d'implémentation

- Ajoutez les exigences relatives à l'amélioration de la durabilité à votre liste de tâches de développement en attente.
- Utilisez un processus d'amélioration itératif pour identifier, évaluer, hiérarchiser, tester et déployer ces améliorations.
- Améliorez et rationalisez en permanence vos processus de développement. À titre d'exemple, automatisez votre processus de livraison de logiciels en utilisant des pipelines d'intégration et de livraison continues (CI/CD) pour tester et déployer les améliorations potentielles afin de réduire le niveau d'effort et de limiter les erreurs causées par les processus manuels.

- Développez et testez les améliorations potentielles en utilisant les composants représentatifs viables minimum afin de réduire le coût des tests.
- Évaluez en permanence l'impact des améliorations et procédez aux ajustements nécessaires.

Ressources

Documents connexes :

- AWS active des solutions de durabilité
- Scalable agile development practices based on AWS CodeCommit (Pratiques de développement agiles et évolutives basées sur AWS CodeCommit)

Vidéos connexes :

- Delivering sustainable, high-performing architectures (Offre d'architectures durables hautement performantes)

Exemples connexes :

- Well-Architected Lab - Turning cost & usage reports into efficiency reports (Atelier Well-Architected : transformer les rapports de coût et d'utilisation en rapports d'efficacité)

SUS06-BP02 Garder votre charge de travail à jour

Maintenez votre charge de travail à jour pour adopter des fonctionnalités efficaces, supprimer les problèmes et améliorer l'efficacité globale de votre charge de travail.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne sera pas mise à jour au fil du temps.
- Vous ne disposez pas de systèmes ou de rythme régulier pour évaluer la compatibilité des packages et des logiciels mis à jour avec votre charge de travail.

Avantages liés au respect de cette bonne pratique : en mettant en place un processus pour maintenir votre charge de travail à jour, vous pouvez adopter de nouvelles fonctionnalités et capacités, résoudre les problèmes et améliorer l'efficacité de la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Des systèmes d'exploitation, des moteurs d'exécution, des logiciels intermédiaires, des bibliothèques et des applications à jour peuvent améliorer l'efficacité de la charge de travail et faciliter l'adoption de technologies plus efficaces. Les logiciels à jour peuvent également inclure des fonctions permettant de mesurer plus précisément l'impact en matière de durabilité de votre charge de travail, car les fournisseurs proposent des fonctions pour atteindre leurs propres objectifs de durabilité. Adoptez une cadence régulière pour maintenir votre charge de travail à jour avec les dernières fonctionnalités et versions.

Étapes d'implémentation

- Définissez un processus et un calendrier pour évaluer les nouvelles fonctionnalités ou instances pour votre charge de travail. Profitez de l'agilité du cloud pour tester rapidement en quoi les nouvelles fonctionnalités peuvent permettre à votre charge de travail de :
  - Réduire les impacts sur la durabilité.
  - Gagner en efficacité de la performance.
  - Supprimer les obstacles à une amélioration planifiée.
  - Améliorer votre capacité à mesurer et à gérer les impacts en matière de durabilité.
- Établissez l'inventaire de votre logiciel de charge de travail et de l'architecture, et identifiez les composants pouvant être mis à jour.
  - Utilisez AWS Systems Manager Inventory pour récupérer les métadonnées des systèmes d'exploitation, des applications et des instances issues de vos instances Amazon EC2 et rapidement connaître les instances exécutant le logiciel, les configurations requises par votre politique de logiciel et les instances devant être mises à jour.
- Comprenez comment mettre à jour les composants de votre charge de travail.

Workload component	How to update
Images machine	Utilisez EC2 Image Builder pour gérer les mises à jour des Amazon Machine Images (AMI) pour les images de serveur Linux ou Windows.

Workload component	How to update
Images de conteneurs	Utilisez Amazon Elastic Container Registry (Amazon ECR) avec votre pipeline existant pour gérer les images Amazon Elastic Container Service (Amazon ECS).
AWS Lambda	AWS Lambda comprend des fonctions de gestion des versions.

- Utilisez l'automatisation pour le processus de mise à jour afin de réduire le niveau d'effort nécessaire au déploiement des nouvelles fonctionnalités et de limiter les erreurs causées par les processus manuels.
  - Vous pouvez utiliser CI/CD pour mettre automatiquement à jour les AMI, les images de conteneurs et d'autres artefacts liés à votre application cloud.
  - Vous pouvez utiliser des outils tels que AWS Systems ManagerPatch Manager pour automatiser le processus de mise à jour du système, et programmer l'activité à l'aide des Fenêtres de maintenance AWS Systems Manager.

Ressources

Documents connexes :

- Centre d'architecture AWS
- Quelles sont les nouveautés AWS ?
- Outils pour développeurs AWS

Exemples connexes :

- Ateliers Well-Architected : inventaire et gestion des correctifs
- Atelier : AWS Systems Manager

SUS06-BP03 Augmenter l'utilisation de vos environnements de compilation

Augmentez l'utilisation des ressources pour développer, tester et construire vos charges de travail.

Anti-modèles courants :

- Vous provisionnez ou résiliez manuellement vos environnements de construction.
- Vous faites fonctionner vos environnements de construction indépendamment des activités de test, de construction ou de lancement (par exemple, en faisant fonctionner un environnement en dehors des heures de travail des membres de votre équipe de développement).
- Vous provisionnez trop de ressources pour vos environnements de construction.

Avantages liés au respect de cette bonne pratique : en augmentant l'utilisation des environnements de construction, vous pouvez améliorer l'efficacité globale de votre charge de travail dans le cloud tout en allouant les ressources aux constructeurs pour développer, tester et construire efficacement.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Exploitez l'automatisation et l'infrastructure en tant que code pour mettre en place des environnements de construction lorsque cela est nécessaire et les arrêter lorsqu'ils ne sont pas utilisés. Un modèle courant consiste à planifier des périodes de disponibilité qui coïncident avec les heures de travail des membres de votre équipe de développement. Vos environnements de test doivent ressembler de près à la configuration de production. Toutefois, recherchez les possibilités d'utiliser des types d'instance avec une capacité de débordement, des instances Amazon EC2 Spot, des services de base de données à scalabilité automatique, des conteneurs et des technologies sans serveur pour aligner la capacité de développement et de test sur l'utilisation. Limitez le volume de données pour répondre aux exigences du test. Si vous utilisez des données de production dans les tests, étudiez les possibilités de partager les données de production et de ne pas déplacer les données à un autre emplacement.

Étapes d'implémentation

- Utilisez l'infrastructure en tant que code pour provisionner vos environnements de construction.
- Utilisez l'automatisation pour gérer le cycle de vie de vos environnements de développement et de test et maximiser l'efficacité de vos ressources de construction.
- Utilisez des stratégies pour optimiser l'utilisation des environnements de développement et de test.
  - Utilisez des environnements représentatifs viables minimum pour développer et tester les améliorations potentielles.
  - Utilisez les technologies sans serveur si possible.
  - Utilisez des instances à la demande pour compléter les appareils de vos développeurs.

- Utilisez des types d'instance à capacité de débordement, des instances Spot et d'autres technologies pour harmoniser la capacité de création et l'utilisation.
- Adoptez des services natifs du cloud pour l'accès à un shell d'instance sécurisé plutôt que de déployer des flottes d'hôtes bastion.
- Mettez automatiquement à l'échelle vos ressources de construction en fonction de vos tâches de construction.

Ressources

Documents connexes :

- Gestionnaire de sessions AWS Systems Manager
- Instances Amazon EC2 de performance à capacité extensible
- Qu'est-ce qu'AWS CloudFormation ?
- Qu'est-ce que AWS CodeBuild ?
- Instance Scheduler sur AWS

Vidéos connexes :

- Continuous Integration Best Practices (Bonnes pratiques d'intégration continue)

SUS06-BP04 Utiliser des tests Device Farms gérés

Utilisez les Device Farms gérés pour tester efficacement une nouvelle fonctionnalité sur un ensemble représentatif de matériel.

Anti-modèles courants :

- Vous testez et déployez manuellement votre application sur des appareils physiques individuels.
- Vous n'utilisez pas le service de test d'applications pour tester et interagir avec vos applications (par exemple, les applications Android, iOS et Web) sur des appareils physiques réels.

Avantages liés au respect de cette bonne pratique : l'utilisation de Device Farms gérés pour tester les applications basées sur le cloud présente un certain nombre d'avantages :

- la solution comprend des fonctionnalités plus efficaces pour tester l'application sur de nombreux appareils différents.
- Elle élimine la nécessité d'une infrastructure interne pour les essais.
- Elle permet l'utilisation de divers types d'appareils, y compris des matériels plus anciens et moins populaires, ce qui élimine le besoin de mises à niveau inutiles des appareils.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

L'utilisation de Device Farms gérés peut vous aider à rationaliser le processus de test des nouvelles fonctionnalités sur un ensemble représentatif de matériel. Les tests Device Farms gérés proposent divers types d'appareils , notamment du matériel plus ancien et moins courant, et permettent d'éviter que les mises à niveau inutiles d'appareils affectent la durabilité des clients.

Étapes d'implémentation

- Définissez vos exigences et votre plan de test (comme le type de test, les systèmes d'exploitation et le calendrier des tests).
  - Vous pouvez utiliser Amazon CloudWatch RUM pour collecter et analyser les données côté client et élaborer votre plan de test.
- Sélectionnez le Device farm géré qui peut prendre en charge vos exigences de test. Par exemple, vous pouvez utiliser AWS Device Farm pour tester et comprendre l'impact de vos changements sur un ensemble représentatif de matériel.
- Utilisez l'intégration continue/déploiement continu (CI/CD) pour programmer et exécuter vos tests.
  - Integrating AWS Device Farm with your CI/CD pipeline to run cross-browser Selenium tests (Intégrer AWS Device Farm à votre pipeline CI/CD pour exécuter des tests Selenium inter-navigateurs)
  - Building and testing iOS and iPadOS apps with AWS DevOps and mobile services (Créer et tester des applications iOS et iPadOS avec AWS DevOps et les services mobiles)
- Examinez continuellement les résultats de vos tests et apportez les améliorations nécessaires.

Ressources

Documents connexes :

- Liste des appareils AWS Device Farm

- Affichage du tableau de bord CloudWatch RUM

Exemples connexes :

- AWS Device Farm Sample App for Android (Application type Device Farm pour Android)
- AWS Device Farm Sample App for iOS (Application type Device Farm pour iOS)
- Appium Web tests for AWS Device Farm (Tests Web Appium pour AWS Device Farm)

Vidéos connexes :

- Optimiser les applications grâce à la connaissance de l'utilisateur final avec Amazon CloudWatch RUM

Mentions légales

Les clients sont responsables de leur propre évaluation indépendante des informations contenues dans ce document. Le présent document : (a) est fourni à titre informatif uniquement, (b) représente les offres et pratiques actuelles de produits AWS, qui sont susceptibles d'être modifiées sans préavis, et (c) ne crée aucun engagement ou assurance de la part d'AWSet de ses affiliés, fournisseurs ou concédants de licences. Les produits ou services AWS sont fournis « en l'état » sans garantie, représentation ou condition, de quelque nature que ce soit, explicite ou implicite. Les responsabilités et obligations d'AWSenvers ses clients sont déterminées par les contrats AWS. Le présent document ne fait pas partie d'un contrat entre AWS et ses clients, et ne le modifie pas.

Copyright © 2021 Amazon Web Services, Inc. ou ses sociétés apparentées.