AWS Well-Architected Framework

Pilier Efficacité des performances

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Pilier Efficacité des performances: AWS Well-Architected Framework

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Les marques et la présentation commerciale d’Amazon ne peuvent être utilisées en relation avec un produit ou un service qui n’est pas d’Amazon, d’une manière susceptible de créer une confusion parmi les clients, ou d’une manière qui dénigre ou discrédite Amazon. Toutes les autres marques commerciales qui ne sont pas la propriété d’Amazon appartiennent à leurs propriétaires respectifs, qui peuvent ou non être affiliés ou connectés à Amazon, ou sponsorisés par Amazon.

Table of Contents

Résumé et introduction 1
Introduction 1
Efficacité des performances 3
Principes de conception 3
Définition 4
Choix d’architecture 5
PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles 5
Directives d’implémentation 6
Ressources 7
PERF01-BP02 Utilisation des recommandations de votre fournisseur de cloud ou d’un partenaire approprié pour en savoir plus sur les modèles d’architecture et les bonnes pratiques 8
Directives d’implémentation 6
Ressources 7
PERF01-BP03 Intégrer les coûts dans les décisions architecturales 10
Directives d’implémentation 6
Ressources 7
PERF01-BP04 Évaluation de l’impact des compromis sur les clients et l’efficacité de l’architecture 12
Directives d’implémentation 6
Ressources 7
PERF01-BP05 Politiques d'utilisation et architectures de référence 14
Directives d’implémentation 6
Ressources 7
PERF01-BP06 Utilisation du benchmarking pour éclairer vos décisions architecturales 16
Directives d’implémentation 6
Ressources 7
PERF01-BP07 Utilisation d’une approche orientée données pour les choix architecturaux 19
Directives d’implémentation 6
Ressources 7
Informatique et matériel 22
PERF02-BP01 Sélection des meilleures options de calcul pour votre charge de travail 22
Directives d’implémentation 6

Étapes d'implémentation 6
Ressources 7
PERF02-BP02 Compréhension des configurations et des fonctionnalités de calcul disponibles 26
Directives d'implémentation 6
Étapes d'implémentation 6
Ressources 7
PERF02-BP03 Collecter des métriques liées au calcul 30
Directives d'implémentation 6
Étapes d'implémentation 6
Ressources 7
PERF02-BP04 Configuration et dimensionnement corrects des ressources de calcul 33
Directives d'implémentation 6
Ressources 7
PERF02-BP05 Mettre à l'échelle vos ressources de calcul de manière dynamique 35
Directives d'implémentation 6
Ressources 7
PERF02-BP06 Utilisation d'accélérateurs de calcul matériels optimisés 39
Directives d'implémentation 6
Ressources 7
Gestion des données 42
PERF03-BP01 Utilisation d'un magasin de données dédié particulièrement adapté à vos besoins en matière de stockage des données et d'accès aux données 42
Directives d'implémentation 6
Ressources 7
PERF03-BP02 Évaluation des options de configuration disponibles pour un magasin de données 55
Directives d'implémentation 6
Ressources 7
PERF03-BP03 Collecte et archivage des métriques de performance du magasin de données 60
Directives d'implémentation 6
Étapes d'implémentation 6
Ressources 7
PERF03-BP04 Mise en œuvre de stratégies pour améliorer les performances des requêtes dans un magasin de données 63
Directives d'implémentation 6

Ressources 7
PERF03-BP05 Mise en œuvre de modèles d'accès aux données utilisant la mise en cache 65
Directives d'implémentation 6
Ressources 7
Réseau et diffusion de contenu 70
PERF04-BP01 Compréhension de l'impact de la mise en réseau sur les performances 70
Directives d'implémentation 6
Ressources 7
PERF04-BP02 Évaluation des fonctionnalités de mise en réseau disponibles 74
Directives d'implémentation 6
Ressources 7
PERF04-BP03 Choix d'une connectivité dédiée ou d'un VPN approprié pour votre charge de travail 81
Directives d'implémentation 6
Ressources 7
PERF04-BP04 Utilisation de l'équilibrage de charge pour répartir le trafic entre plusieurs ressources 84
Directives d'implémentation 6
Ressources 7
PERF04-BP05 Choix de protocoles réseau afin d'améliorer les performances 88
Directives d'implémentation 6
Ressources 7
PERF04-BP06 Choisissez l'emplacement de votre charge de travail en fonction des exigences du réseau 92
Directives d'implémentation 6
Ressources 7
PERF04-BP07 Optimisation de la configuration réseau en fonction de métriques 98
Directives d'implémentation 6
Ressources 7
Processus et culture 104
PERF05-BP01 Définition d'indicateurs de rendement clés (KPI) pour mesurer l'état et les performances de la charge de travail 106
Directives d'implémentation 6
Étapes d'implémentation 6
Ressources 7

PERF05-BP02 Utilisation de solutions de surveillance pour comprendre les domaines où les performances sont d'une importance critique 109
Directives d'implémentation 6
Ressources 7
PERF05-BP03 Définition d'un processus pour améliorer les performances des charges de travail 112
Directives d'implémentation 6
Ressources 7
PERF05-BP04 Testez votre charge de travail 114
Directives d'implémentation 6
Ressources 7
PERF05-BP05 Utilisation de l'automatisation pour résoudre de manière proactive les problèmes liés aux performances 116
Directives d'implémentation 6
Ressources 7
PERF05-BP06 Maintenez votre charge de travail et vos services up-to-date 119
Directives d'implémentation 6
Étapes d'implémentation 6
Ressources 7
PERF05-BP07 Vérification des métriques à intervalles réguliers 121
Directives d'implémentation 6
Ressources 7
Conclusion 124
Collaborateurs 125
Suggestions de lecture 126
Révisions du document 127
Avis 129
AWS Glossaire 130

Pilier Efficacité des performances - AWS Well-Architected Framework

Date de publication : 6 novembre 2024 (Révisions du document)

Ce livre blanc porte sur le pilier Efficacité des performances d'AWS Well-Architected Framework. Il fournit des conseils pour aider les clients à appliquer les bonnes pratiques de conception, de distribution et de maintenance des environnements AWS.

Introduction

AWS Well-Architected Framework vous aide à mesurer le pour et le contre des options qui se présentent lors de la création de charges de travail 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 cloud. Il vous permet d'évaluer systématiquement vos architectures par rapport aux bonnes pratiques et d'identifier les domaines à améliorer. Nous pensons que le fait d'avoir des charges de travail bien structurées augmente considérablement les chances de réussite métier.

Le cadre repose sur six piliers :

- Excellence opérationnelle
- Sécurité
- Fiabilité
- Efficacité des performances
- Optimisation des coûts
- Durabilité

Ce livre blanc porte sur l'application des principes du pilier Efficacité des performances à vos charges de travail. Dans les environnements sur site traditionnels, il est difficile de bénéficier de performances élevées et durables. En appliquant les principes de ce livre blanc, vous pourrez créer des architectures sur AWS qui fournissent avec efficacité des performances soutenues sur le long terme. Les conseils et les bonnes pratiques présentés dans ce document sont répartis dans cinq domaines clés qui servent de principes directeurs pour la création de solutions cloud performantes sur AWS. Ces domaines d'intérêt sont les suivants :

- Choix d’architecture
- Informatique et matériel
- Gestion des données
- Réseau et diffusion de contenu
- Processus et culture

Le présent document est conçu pour ceux et celles qui sont dépositaires de rôles technologiques, comme les directeurs de la technologie, les architectes, les développeurs et les membres de l’équipe d’exploitation. Après avoir lu ce document, vous allez vous familiariser avec les bonnes pratiques et les stratégies d’AWS à utiliser lors de la conception d’architectures cloud performantes.

Efficacité des performances

Le pilier Efficacité des performances englobe la capacité à utiliser efficacement les ressources du cloud pour satisfaire aux exigences système et à maintenir cette efficacité au fur et à mesure que la demande change et que les technologies évoluent.

Rubriques

- Principes de conception
- Définition

Principes de conception

Les principes de conception suivants peuvent vous aider à créer des charges de travail efficaces dans le cloud, tout en veillant à ce qu'elles le restent dans la durée.

- 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 technologies en tant que service. Par exemple, l'absence SQL de bases de données, le transcodage multimédia et l'apprentissage automatique sont autant de technologies qui nécessitent une expertise spécialisée. 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.
- Passez à l'international en quelques minutes : le déploiement de votre charge de travail dans plusieurs AWS régions du monde vous permet de réduire le temps de latence et d'offrir une meilleure expérience à vos clients à moindre coût.
- 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.

- Envisager la compréhension technique : utilisez l’approche technologique qui correspond le mieux à vos objectifs. Par exemple, tenez compte des modèles d’accès aux données lorsque vous sélectionnez les approches de stockage ou de base de données de votre charge de travail.

Définition

Concentrez-vous sur les domaines suivants pour assurer l’efficacité des performances dans le cloud :

- Choix d’architecture
- Informatique et matériel
- Gestion des données
- Réseau et diffusion de contenu
- Processus et culture

Adoptez une approche axée sur les données pour créer une architecture performante. 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 revoyant régulièrement vos choix, vous vous assurez de tirer parti de l'évolution constante du AWS Cloud. La surveillance vous offre la garantie d’être informé de tout écart par rapport aux performances attendues. Effectuer 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.

Choix d’architecture

La solution optimale pour une charge de travail peut varier, et les solutions combinent souvent plusieurs approches. Les charges de travail Well-Architected 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 qu’Amazon DynamoDB fournit une base de données NoSQL entièrement gérée avec une latence de moins de 10 millisecondes, quelle que soit l’échelle.

Ce domaine d’intérêt partage des conseils et des bonnes pratiques sur la manière de sélectionner des ressources cloud et des modèles d’architecture efficaces et performants.

Bonnes pratiques

- PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles
- PERF01-BP02 Utilisation des 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 Intégrer les coûts dans les décisions architecturales
- PERF01-BP04 Évaluation de l’impact des compromis sur les clients et l’efficacité de l’architecture
- PERF01-BP05 Politiques d'utilisation et architectures de référence
- PERF01-BP06 Utilisation du benchmarking pour éclairer vos décisions architecturales
- PERF01-BP07 Utilisation d’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 d'exposition au risque 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.
  - Amazon Web Services Cloud
  - AWS Academy
  - Quelles sont les nouveautés AWS ?
  - AWSBlog
  - AWS Skill Builder

  - Événements et webinaires AWS
  - AWS Training et certifications
  - Chaîne YouTube AWS
  - Ateliers AWS
  - Communautés AWS
- Utilisez Amazon Q pour obtenir des informations pertinentes et des conseils sur les services.
- Utilisez des environnements de test (hors production) pour découvrir et tester de nouveaux services sans frais supplémentaires.
- Découvrez en permanence les nouveaux services et fonctionnalités du cloud.

Ressources

Documents connexes :

- Présentation d'Amazon Web Services
- Fonctionnalités Amazon EC2
- Apprentissage étape par étape grâce à un Plan de formation pour les partenaires AWS
- Formation et certification AWS
- Mon parcours d'apprentissage pour devenir architecte de solutions AWS
- Centre d'architecture AWS
- AWS Partner Network
- AWS Bibliothèque de solutions
- Centre de connaissances AWS
- Créer des applications modernes sur AWS

Vidéos connexes :

- AWS re:Invent 2023 - What's new with Amazon EC2
- AWS re:Invent 2022 - Reduce your operational and infrastructure costs with Amazon ECS
- AWS re:Invent 2023 - Build with the efficiency, agility & innovation of the cloud with AWS
- AWS re:Invent 2022 - Deploy ML models for inference at high performance and low cost

- Voici mon architecture

Exemples connexes :

- AWS Exemples
- AWS Exemples de kit SDK

PERF01-BP02 Utilisation des 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.

Avantage de l'établissement 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 d'exposition au risque 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 auprès des experts AWS les bonnes pratiques d'utilisation des services AWS.
  - Apprentissage étape par étape grâce à un Plan de formation pour les partenaires 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. AWS Les architectes de solutions et les services professionnels AWS prodiguent des conseils pour la mise en œuvre de solutions. AWS Les partenaires AWS apportent une expertise pour vous aider à gagner en agilité et favoriser l'innovation au sein de votre entreprise.
- Utilisez 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:

- AWS Centre d'architecture
- AWS Partner Network

- AWS Bibliothèque de solutions
- Centre de connaissances AWS
- AWS Entreprise Support

Vidéos connexes :

- Voici mon architecture
- AWS re:Invent 2023 - Advanced event-driven patterns with Amazon EventBridge
- AWS re:Invent 2023 - Implementing distributed design patterns on AWS
- AWS re:Invent 2023 - Application architecture as code

Exemples connexes :

- AWS Exemples
- Exemples de kit SDK AWS
- Architecture de référence pour l'analytique AWS

PERF01-BP03 Intégrer les coûts dans les 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 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 encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

L'optimisation des charges de travail en matière 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. Calculateur de tarification AWS et AWS Cost Explorer vous permettent d'identifier les principaux facteurs de coûts dans votre charge de travail.
- Comprendre les modèles de tarification dans le cloud, tels que la demande, les instances réservées, les Savings Plans et les instances ponctuelles.
- Utilisez les bonnes pratiques d'optimisation des coûts de Well-Architected pour optimiser ces composants clés en matière de coûts.
- Surveillez et analysez en permanence les coûts afin d'identifier les opportunités d'optimisation des coûts dans votre charge de travail.
  - Utilisez les budgets AWS pour recevoir des alertes en cas de coûts inacceptables.
  - Utilisez Optimiseur de calcul AWS ou AWS Trusted Advisor pour obtenir des recommandations en matière d'optimisation des coûts.
  - Utilisez la détection des anomalies de coûts AWS pour obtenir une détection automatisée des anomalies de coûts et une analyse des causes profondes.

Ressources

Documents connexes :

- Qu'est-ce que AWS Billing and Cost Management ?
- Optimisation des coûts avec AWS
- Choix d'une stratégie de gestion des AWS coûts
- Guide de gestion des AWS coûts pour débutants

- Présentation détaillée du tableau de bord Cost Intelligence Dashboard
- Centre d'architecture AWS
- Bibliothèque de solutions AWS
- Centre de connaissances AWS

Vidéos connexes :

- This is my Architecture
- AWS re:Invent 2023 - Nouveautés en matière d'optimisation des coûts AWS
- AWS re:Invent 2023 - Optimisez les coûts et les performances et suivez les progrès en matière d'atténuation
- AWS re:Invent 2023 - meilleures pratiques en matière d'optimisation des coûts AWS de stockage
- AWS re:Invent 2023 - Optimisez les coûts dans vos environnements multi-comptes

Exemples connexes :

- Optimiseur de calcul AWS Code de démonstration
- Atelier d'optimisation des coûts
- Playbooks de mise en œuvre technique de la gestion financière dans le cloud
- Optimisation du démarrage : ajustement des performances des applications pour une efficacité maximale
- Atelier d'optimisation sans serveur (performances et coûts)
- Mise à l'échelle d'architectures rentables

PERF01-BP04 Évaluation de 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 en termes d'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 d'exposition au risque 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. Comprenez les compromis entre la complexité, les performances et les coûts de votre charge de travail. Généralement, la priorisation de deux des facteurs se fait au détriment du troisième.

Ressources

Documents connexes :

- Bibliothèque Amazon Builders' Library
- KPI de Quick
- Amazon CloudWatch RUM
- Documentation X-Ray
- Comprenez les modèles de résilience et les compromis pour concevoir une architecture efficace dans le cloud

Vidéos connexes :

- Optimize applications through via Amazon CloudWatch RUM
- AWSre:Invent 2023 - Capacity, availability, cost efficiency: Pick three
- AWS re:Invent 2023 - Advanced integration patterns & trade-offs for loosely coupled systems

Exemples connexes :

- Measure page load time with Amazon CloudWatch Synthetics
- Client Web Amazon CloudWatch RUM

PERF01-BP05 Politiques d'utilisation et 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 encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Le fait de disposer de stratégies 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 (sandbox).
- up-to-dateRespectez les normes et les AWS mises à jour du secteur pour vous assurer que vos politiques et architectures de référence contribuent à optimiser votre charge de travail dans le cloud.

Ressources

Documents connexes :

- Centre d’architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS

- AWS Blogue d'architecture

Vidéos connexes :

- This is my Architecture
- AWS re:Invent 2022 - Accélérez la création de valeur pour votre entreprise grâce à une architecture de SAP référence AWS

Exemples connexes :

- Exemples AWS
- AWS SDKExemples

PERF01-BP06 Utilisation du 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 de l'établissement de cette bonne pratique : le benchmarking de votre implémentation actuelle vous permet de mesurer les améliorations de performance.

Niveau d'exposition au risque 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. Le benchmarking est généralement plus rapide à configurer que les tests de charge. Il est utilisé pour évaluer la technologie pour un

composant en particulier. Le benchmarking est souvent utilisé 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 performances, ou bien utiliser un test conforme aux normes du secteur, comme le TPC-DS pour comparer 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 du benchmarking 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

- Planification et définition :
  - Définissez les objectifs, la base de référence, les scénarios de test, les métriques (telles que l'utilisation du processeur, la latence ou le débit) et les indicateurs de rendement clés de votre test de performances.
  - Concentrez-vous sur les exigences des utilisateurs en matière d'expérience utilisateur et sur des facteurs tels que le temps de réponse et l'accessibilité.
  - Identifiez un outil de benchmarking adapté à votre charge de travail. Vous pouvez utiliser des services AWS tels qu'Amazon CloudWatch ou un outil tiers compatible avec votre charge de travail.
- Configuration et instrumentation :
  - Configurez votre environnement et vos ressources.
  - Mettez en œuvre la surveillance et la journalisation pour capturer les résultats des tests.
- Comparaison et surveillance :
  - Effectuez vos tests de performances et surveillez les métriques pendant le test.
- Analyse et documentation :

  - Documentez votre processus de benchmarking et vos résultats.
  - Analysez les résultats pour identifier les goulots d'étranglement, les tendances et les domaines d'amélioration.
  - 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.
- Optimisation et répétition :
  - Ajustez les configurations et les allocations des ressources en fonction de vos critères de référence.
  - Testez à nouveau votre charge de travail après ajustement pour valider vos améliorations.
  - Documentez vos conclusions et répétez le processus pour identifier d'autres domaines d'amélioration.

Ressources

Documents connexes :

- Centre d'architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Flux de travail génomiques, partie 5 : benchmarking automatisé
- Évaluation et optimisation du déploiement des points de terminaison dans Amazon SageMaker AI JumpStart

Vidéos connexes :

- AWS re:Invent 2023 - Benchmarking AWS Lambda cold starts
- Benchmarking stateful services in the cloud
- This is my Architecture
- Optimize applications through via Amazon CloudWatch RUM

- Présentation d'Amazon CloudWatch Synthetics

Exemples connexes :

- AWS Exemples
- Exemples de kit SDK AWS
- Tests de charge distribuée
- Mesure du temps de chargement des pages avec Amazon CloudWatch Synthetics
- Client Web Amazon CloudWatch RUM

PERF01-BP07 Utilisation d'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 d'exposition au risque 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 choisir les 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 Utilisation des 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, comme AWS Solutions Architects, AWSArchitecture Center et AWS Partner Network, pour vous aider à choisir l'architecture adaptée à votre charge de travail.

- Définissez des métriques de performances 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
- Architectural Patterns to Build End-to-End Data Driven Applications on AWS

Vidéos connexes :

- This is my Architecture
- AWS re:Invent 2021 - Data-driven enterprise: Going from vision to value
- AWS re:Invent 2022 – Delivering sustainable, high-performing architectures
- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results

Exemples connexes :

- AWS Exemples
- Exemples de kit SDK AWS

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. Le choix d'une solution de calcul inadaptée à une architecture peut nuire à ses performances.

Ce domaine d'intérêt partage des conseils et de bonnes pratiques sur la manière d'identifier et d'optimiser les options de calcul pour l'efficacité des performances dans le cloud.

Bonnes pratiques

- PERF02-BP01 Sélection des meilleures options de calcul pour votre charge de travail
- PERF02-BP02 Compréhension des configurations et des fonctionnalités de calcul disponibles
- PERF02-BP03 Collecter des métriques liées au calcul
- PERF02-BP04 Configuration et dimensionnement corrects des ressources de calcul
- PERF02-BP05 Mettre à l'échelle vos ressources de calcul de manière dynamique
- PERF02-BP06 Utilisation d'accélérateurs de calcul matériels optimisés

PERF02-BP01 Sélection des 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 cloud computing 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 d'exposition au risque 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 à gérer de serveurs, Amazon ECS ou Amazon EKS pour exécuter et gérer des conteneurs ou encore AWS Batch pour traiter d'importants 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

- 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.
- Découvrez les différents services AWS informatiques adaptés à votre charge de travail. Pour de plus amples informations, consultez 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
Amazon Elastic Compute Cloud (Amazon EC2)	Possède une option dédiée pour le matériel, les exigences de licence, une	Migration « lift-and-shift », application monolithique,

Service AWS	Principales caractéristiques	Cas d'utilisation courants
	large sélection de différent es familles d'instances, les types de processeurs et les accélérateurs de calcul	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	Service de calcul sans serveur qui exécute du code en réponse à des événement s et gère automatiquement les ressources de calcul sous-jacentes.	Microservices, applications basées sur les événements
AWS Batch	Approvisionnement et mise à l'échelle efficaces et dynamiques d'Amazon Elastic Container Service (Amazon ECS), d'Amazon Elastic Kubernetes Service (Amazon EKS) et des ressources de calcul AWS Fargate, avec la possibilité d'utiliser des instances à la demande ou des instances ponctuelles en fonction des exigences de votre travail.	HPC, entraînement des modèles de ML
Amazon Lightsail	Application Linux et Windows préconfigurée pour exécuter de petites charges de travail	Applications web simples, site web personnalisé

- 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.
- 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.
- Après avoir testé et identifié votre nouvelle solution de calcul, planifiez votre migration et validez vos métriques de performance.
- Utilisez les outils de surveillance AWS tels qu'Amazon CloudWatch et les services d'optimisation tels que Optimiseur de calcul AWS pour optimiser en continu votre calcul en fonction de modèles d'utilisation réels.

Ressources

Documents connexes:

- Cloud Compute with AWS
- Types d'instances Amazon EC2
- Conteneurs Amazon EKS : composants master Amazon EKS
- Conteneurs Amazon ECS : instances de conteneur Amazon ECS
- Fonctions : configuration des fonctions Lambda
- Recommandations pour les conteneurs
- Recommandations pour les modèles sans serveur

Vidéos connexes :

- AWS re:Invent 2023 AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 - New Amazon Elastic Compute Cloud generative AI capabilities in AMS
- AWS re:Invent 2023 - What's new with Amazon Elastic Compute Cloud
- AWS re:Invent 2023 - Smart savings: Amazon Elastic Compute Cloud cost-optimization strategies
- AWS re:Invent 2021 - Powering next-gen Amazon Elastic Compute Cloud: Deep dive on the Nitro System
- AWS re:Invent 2019 - Optimize performance and cost for your AWS compute
- AWS re:Invent 2019 - Amazon Elastic Compute Cloud foundations
- AWS re:Invent 2022 - Deploy ML models for inference at high performance and low cost

- AWS re:Invent 2019 - Optimize performance and cost for your AWS compute
- Amazon EC2 foundations
- Deploy ML models for inference at high performance and low cost

Exemples connexes :

- Migration de l'application Web vers des conteneurs
- Exécution d'un modèle Hello World sans serveur
- Atelier Amazon EKS
- Atelier Amazon EC2
- Charges de travail efficaces et résilientes avec l'autoscaling Amazon EC2 Auto Scaling
- Migration vers Graviton AWS avec Container Services

PERF02-BP02 Compréhension des configurations et des 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 ni 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 d'exposition au risque 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 prévoyez que les caractéristiques resteront les mêmes à l'avenir, vous pouvez utiliser Optimiseur de calcul AWS 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

- Comprenez les exigences de la charge de travail (comme les besoins en UC, la mémoire et la latence).
- 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 processeur virtuel/mémoire plus élevé. • Les instances à mémoire optimisée 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	• 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

Option de configuration	Exemples
	dépasser les besoins de base en matière de performances. • Les Savings Plans permettent de réaliser des économies importantes par rapport aux instances à la demande, en échange d'un engagement à utiliser une quantité spécifiqu e de puissance de calcul pour une période d'un ou de trois ans. • Les 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 la configuration Auto Scaling pour faire correspondre les ressources de calcul aux modèles de trafic.
Dimensionnement	• Utilisez Compute Optimizer pour obtenir des recommandations basées sur 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 recherche de modèles de données plus efficacement que les alternatives basées sur le processeur. • Pour les charges de travail de machine learning, tirez parti d'un matériel conçu spécialement pour votre charge de travail, par exemple AWS Trainium, AWS Inferentia et Amazon EC2 DL1

Ressources

Documents connexes:

- Cloud Compute with AWS
- Types d'instances Amazon EC2
- Contrôle des états du processeur pour votre instance Amazon EC2
- Conteneurs Amazon EKS : composants master Amazon EKS
- Conteneurs Amazon ECS : instances de conteneur Amazon ECS
- Fonctions : configuration des fonctions Lambda

Vidéos connexes :

- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in Console de gestion AWS
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations
- AWSre:Invent 2022 – Optimizing Amazon EKS for performance and cost on AWS

Exemples connexes :

- Code de démonstration de Compute Optimizer
- Atelier sur les instances spot Amazon EC2
- Charges de travail efficaces et résilientes avec Amazon EC2 AWS Auto Scaling
- Atelier pour développeurs Graviton
- Journée d'immersion AWS pour les charges de travail Microsoft
- Journée d'immersion AWS pour les charges de travail Linux
- Code de démonstration Optimiseur de calcul AWS
- Atelier Amazon EKS

PERF02-BP03 Collecter des 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 d'exposition au risque 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 ce contexte AWS Cloud, la collecte de métriques est une

étape cruciale pour améliorer la sécurité, la rentabilité, les performances et la durabilité. AWS fournit un large éventail de mesures liées aux performances à l'aide de services de surveillance tels qu'Amazon CloudWatch pour vous fournir des informations précieuses. Des indicateurs tels que CPU l'utilisation, l'utilisation de la mémoire, les E/S du disque et les entrées et sorties du réseau peuvent fournir des informations sur les niveaux d'utilisation ou les goulots d'étranglement 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

- 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).
  - Métriques EC2 par défaut d'Amazon
  - Métriques ECS par défaut d'Amazon
  - Métriques EKS par défaut d'Amazon
  - Métriques par défaut de Lambda
  - Métriques relatives à EC2 la mémoire et au disque Amazon
- Choisissez et configurez la solution de journalisation et de surveillance adaptée à votre charge de travail.
  - Observabilité native AWS
  - AWS Distro pour OpenTelemetry
  - Amazon Managed Service for Prometheus
- Définissez le filtre et l'agrégation requis pour les métriques en fonction de vos exigences en matière de charge de travail.
  - Quantifiez les métriques personnalisées des applications avec Amazon CloudWatch Logs et les filtres métriques
  - Collectez des statistiques personnalisées grâce au balisage CloudWatch stratégique d'Amazon
- Configurez des stratégies de conservation des données pour vos métriques afin qu'elles correspondent à vos objectifs sécuritaires et opérationnels.
  - Conservation des données par défaut pour les CloudWatch métriques
  - Conservation des données par défaut pour les CloudWatch journaux

- 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.
  - Créez des alarmes pour des métriques personnalisées à l'aide de la détection des CloudWatch anomalies Amazon
  - Créez des métriques et des alarmes pour des pages Web spécifiques avec Amazon CloudWatch RUM
- Utilisez l’automatisation pour déployer vos agents d’agrégation de métriques et de journaux.
  - AWS Systems Manager automatisation
  - OpenTelemetryCollectionneur

Ressources

Documents connexes :

- Surveillance et observabilité
- Bonnes pratiques : mise en œuvre de l'observabilité avec AWS
- CloudWatch Documentation Amazon
- Collectez des métriques et des journaux à partir d'EC2instances Amazon et de serveurs sur site avec l'agent CloudWatch
- Accès à Amazon CloudWatch Logs pour AWS Lambda
- Utilisation CloudWatch des journaux avec des instances de conteneur
- Publier des métriques personnalisées
- AWS Réponse : journalisation centralisée
- AWS Services qui publient CloudWatch des métriques
- Surveillance d'Amazon EKS sur AWS Fargate

Vidéos connexes :

- AWS re:Invent 2023 — [LAUNCH] Surveillance des applications pour les charges de travail modernes
- AWS re:Invent 2023 — Mise en œuvre de l'observabilité des applications
- AWS re:Invent 2023 — Élaborer une stratégie d'observabilité efficace
- AWS re:Invent 2023 — Une observabilité sans faille avec Distro pour AWS OpenTelemetry

- Gestion des performances des applications sur AWS

Exemples connexes :

- AWS Journée d'immersion pour les charges de travail Linux - Amazon CloudWatch
- Surveillance des ECS clusters et des conteneurs Amazon
- Surveillance à l'aide des tableaux de CloudWatch bord Amazon
- EKSAtelier Amazon

PERF02-BP04 Configuration et dimensionnement corrects des 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 de 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 : un dimensionnement correct des ressources de calcul garantit un 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 des performances renforcées, une meilleure expérience client et une baisse des coûts.

Niveau d’exposition au risque 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 entraîner des performances médiocres et une expérience client négative. AWS fournit des outils tels que Optimiseur de calcul AWS et AWS Trusted Advisor qui utilisent des données historiques pour fournir des recommandations afin de dimensionner correctement vos ressources informatiques.

Étapes d'implémentation

- Choisissez le type d'instance qui correspond le mieux à vos besoins :
  - Comment choisir le type d'instance EC2 approprié pour mon application ?
  - Sélection de type d'instance basée sur des attributs pour la flotte d'Amazon EC2
  - Création d'un groupe Auto Scaling en utilisant la sélection du type d'instance basée sur des attributs
  - Optimisation de vos coûts de calcul Kubernetes avec la consolidation Karpenter
- 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 qu'Amazon CloudWatch.
- Sélectionnez la configuration adaptée à vos ressources de calcul.
  - Pour les charges de travail éphémères, évaluez les métriques Amazon CloudWatch de l'instance, `CPUUtilization` afin de déterminer si l'instance est sous-utilisée ou surutilisée.
  - Pour les charges de travail stables, vérifiez les outils de redimensionnement AWS tels que Optimiseur de calcul AWS et AWS Trusted Advisor à intervalles réguliers pour identifier les opportunités d'optimisation et de redimensionnement des ressources de calcul.
- 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:

- Cloud Compute with AWS
- Types d'instances Amazon EC
- Conteneurs Amazon ECS : instances de conteneur Amazon ECS
- Conteneurs Amazon EKS : composant master Amazon EKS
- Fonctions : configuration des fonctions Lambda
- Contrôle des états du processeur pour votre instance Amazon EC2

Vidéos connexes :

- Amazon EC2 foundations
- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in Console de gestion AWS
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations

Exemples connexes :

- Code de démonstration Optimiseur de calcul AWS
- Atelier Amazon EKS
- Recommandations en matière de redimensionnement

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 d'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 d'exposition au risque 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 de suivi des objectifs : 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 le calendrier : définissez 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 d'autoscaling	Où utiliser
Amazon EC2 Auto Scaling	Permet de s'assurer que vous disposez du nombre adéquat d'instances Amazon EC2 pour gérer la charge de votre application.
Application Autoscaling	Pour mettre à l'échelle automatiquement les ressources pour des services AWS individuels au-delà d'Amazon EC2, tels que les fonctions AWS Lambda ou les services Amazon Elastic Container Service (Amazon ECS).
Outil Cluster Autoscaler/Karpenter de Kubernetes	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 pour 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. Le cas échéant, vous pouvez utiliser une métrique personnalisée pour votre politique de dimensionnement. 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 métrique doit augmenter ou diminuer en proportion du nombre d'instances présentes dans le groupe Auto Scaling.
- Assurez-vous d'utiliser une mise à l'échelle dynamique plutôt qu'une mise à l'échelle manuelle pour votre groupe Auto Scaling. Nous vous recommandons également d'utiliser des politiques de dimensionnement pour le suivi des cibles dans votre dimensionnement 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). Par exemple, vous pouvez utiliser l'historique des activités pour vérifier une activité de mise à l'échelle dans 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, reportez-vous à Mise à l'échelle prédictive pour Amazon EC2 Auto Scaling.

Ressources

Documents connexes :

- Cloud Compute with AWS
- Types d'instances Amazon EC2
- Conteneurs Amazon ECS : instances de conteneur Amazon ECS
- Conteneurs Amazon EKS : composant master Amazon EKS
- Fonctions : configuration des fonctions Lambda
- Contrôle des états du processeur pour votre instance Amazon EC2
- Présentation approfondie d'Amazon ECS Cluster Auto Scaling
- Présentation de Karpenter, un Cluster Autoscaler de Kubernetes hautement performant et open source

Vidéos connexes :

- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations

Exemples connexes :

- Exemples de groupes Amazon EC2 Auto Scaling

- Atelier Amazon EKS
- Mettez à l'échelle vos charges de travail Amazon EKS en les exécutant sur IPv6

PERF02-BP06 Utilisation d'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 moindre 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 processeurs graphiques (GPU) et une matrice de portes programmables sur site (FPGA), vous pouvez exécuter certaines fonctions de traitement de manière plus efficace.

Niveau d'exposition au risque 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 matière 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 les instances de calcul accéléré qui 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, par exemple AWS Trainum, AWS Inferentia et Amazon EC2 DL1. AWS Les instances Inferentia telles que les instances Inf2 offrent des performances/watt jusqu'à 50 % supérieures à celles des instances Amazon EC2 comparables.
- Collectez des métriques d'utilisation pour vos instances de calcul accéléré. Par exemple, vous pouvez utiliser l'agent CloudWatch pour collecter des métriques telles que `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 l'IA 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 :

- Utilisation de GPU sur Amazon Elastic Container Service
- instances GPU
- Instances avec AWS Trainium
- Instances avec AWS Inferentia
- Passons à l'architecture ! Architecture avec des puces personnalisées et des accélérateurs

- Calcul accéléré
- Instances Amazon EC2 VT1
- Comment choisir le type d'instance EC2 approprié pour mon application ?
- Choix du meilleur accélérateur d'IA et de la meilleure compilation de modèles pour l'inférence de vision par ordinateur avec l'IA Amazon SageMaker

Vidéos connexes :

- AWSre:Invent 2021 - How to select Amazon Elastic Compute Cloud GPU instances for deep learning
- AWSre:INVENT 2022 - [NEW LAUNCH!] Introducing AWS Inferentia2-based Amazon EC2 Inf2 instances
- AWSre:Invent 2022 - Accelerate deep learning and innovate faster with AWS Trainium
- AWS re:Invent 2022 - Deep learning on AWS with NVIDIA: From training to deployment

Exemples connexes :

- IA Amazon SageMaker et NVIDIA GPU Cloud (NGC)
- Utilisation de l'IA SageMaker avec Trainium et Inferentia pour optimiser les charges de travail d’inférence et d'entraînement du deep learning
- Optimisation des modèles NLP avec les instances Amazon Elastic Compute Cloud Inf1 dans l'IA Amazon SageMaker

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 Well-Architected utilisent des magasins de données sur mesure qui intègrent différentes fonctionnalités pour améliorer les performances.

Ce domaine d'intérêt partage des conseils et de bonnes pratiques pour optimiser le stockage de données, les modèles de déplacement et d'accès, ainsi que l'efficacité des performances des magasins de données.

Bonnes pratiques

- PERF03-BP01 Utilisation d'un magasin de données dédié particulièrement adapté à vos besoins en matière de stockage des données et d'accès aux données
- PERF03-BP02 Évaluation des options de configuration disponibles pour un magasin de données
- PERF03-BP03 Collecte et archivage des métriques de performance du magasin de données
- PERF03-BP04 Mise en œuvre de 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 Utilisation d'un magasin de données dédié particulièrement 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 implémenté 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 d'exposition au risque 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)

  - Période de conservation des données
- Découvrez les différents magasins de données (services de stockage et de base de données) disponibles pour votre charge de travail sur AWS qui peuvent répondre à vos caractéristiques de 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	AWS Services	Principales caractéristiques
Stockage d'objets	Amazon S3 dans l3	Capacité de mise à l'échelle illimitée, haute disponibilité et plusieurs options d'accessi bilité. Le transfert et l'accès à des objets à l'intérieur et à l'extérieur d'Amazon S3 peuvent utiliser un service, tel que Transfer Accelerat ion ou Access Points (points d'accès), pour répondre à votre localisation, à vos besoins en matière de sécurité et à vos modèles d'accès.
Archivage et stockage	Amazon Glacier	Conçu pour l'archivage des données.
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.

Type	AWS Services	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 l'IOPS d'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 en mode bloc	Amazon Elastic Block Store (Amazon EBS)	Service de stockage en blocs évolutif et à hautes performances conçu pour Amazon Elastic Compute Cloud (Amazon EC2). Amazon EBS inclut un stockage SSD pour les charges de travail transactionnelles intensives en IOPS et un stockage sur disque dur pour les charges de travail gourmandes en débit.

Type	AWS Services	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 l'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.

Type	AWS Services	Principales caractéristiques
Base de données documentaire	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.
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

Type	AWS Services	Principales caractéristiques
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.
Base de données de séries temporelles	Amazon Timestream	Utilisées pour collecter, synthétiser et extraire efficacement des informations 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 de séries temporelles.

Type	AWS Services	Principales caractéristiques
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. Généralement, vous voyez un magasin de colonnes larges dans les applications industrielles à grande échelle pour la maintenance des équipements, la gestion des parcs et l'optimisation des itinéraires.
Registre	Amazon Quantum Ledger Database (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 construisez une plateforme de données, tirez parti d'une architecture de données moderne sur AWS pour intégrer votre lac de données, votre entrepôt de données et vos magasins de données spécifiques.
- 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 structuré es, envisagez un magasin d'objets tel qu'Amazon S3 ou une base de données NoSQL telle qu'Amazon DocumentDB • Pour les données clé-valeur, pensez à DynamoDB, Amazon ElastiCache (Redis OSS) ou Amazon MemoryDB
Quel niveau d'intégrité référentielle est requis ?	• En ce qui concerne les contraintes liées aux clés étrangères, les bases de données relationnelles telles qu'Amazon RDS et Aurora peuvent fournir ce niveau d'intégrité. • 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 cohérence forte est requise pour une base de données NoSQL, vous pouvez utiliser des lectures fortement cohérentes avec DynamoDB.

Question	Éléments à prendre en compte
Comment les exigences de stockage vont-elles évoluer au fil du temps ? Comment cela affectera-t-il la capacité de mise à l'échelle ?	• Les bases de données sans serveur telles que 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 performances ?	• Les charges de travail lourdes en lecture 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.
Le stockage et la modification (OLTP - Online Transaction Processing) ou la récupération 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 d'utiliser une base de données en colonnes telle qu'Amazon Redshift ou d'exporter les données vers Amazon S3 et d'effectuer des analyses à l'aide d'Athena ou d'Amazon Quick.

Question	Éléments à prendre en compte
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.
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. • Tirez parti de 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-hé bergement d'une base de données NoSQL contribue à réduire les frais généraux opérationnels.

Question	Éléments à prendre en compte
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 maintenir la compatibilité avec les bases de données qu'ils prennent en charge. Amazon RDS est entièrement compatible avec les différentes versions du 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 volumes Amazon EBS
- Stockage Amazon EC2
- Amazon EFS : performances d'Amazon EFS
- Performances d'Amazon FSx pour Lustre
- Performances d'Amazon FSx for Windows File Server
- Amazon Glacier : documentation Amazon 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 d'Amazon Redshift

- Amazon Athena top 10 de conseils en matière de performance
- Bonnes pratiques Amazon Redshift Spectrum
- Bonnes pratiques Amazon DynamoDB
- Choix entre Amazon EC2 et Amazon RDS
- Bonnes pratiques de mise en œuvre d'Amazon ElastiCache

Vidéos connexes :

- AWS re:Invent 2023: Improve Amazon Elastic Block Store efficiency and be more cost-efficient
- AWS re:Invent 2023: Optimizing storage price and performance with Amazon Simple Storage Service
- AWS re:Invent 2023: Building and optimizing a data lake on Amazon Simple Storage Service
- AWS re:Invent 2022: Building modern data architectures on AWS
- AWS re:Invent 2022: Building data mesh architectures on AWS
- AWS re:Invent 2023: Deep dive into Amazon Aurora and its innovations
- AWS re:Invent 2023: Advanced data modeling with Amazon DynamoDB
- AWS re:Invent 2022: Modernize apps with purpose-built databases
- Amazon DynamoDB deep dive: Advanced design patterns

Exemples connexes :

- Atelier sur les bases de données sur mesure AWS
- Bases de données pour développeurs
- Journée d'immersion dans l'architecture de données moderne AWS
- Création d'un maillage de données sur AWS
- Exemples Amazon S3
- Optimisation du modèle de données à l'aide du partage de données Amazon Redshift
- Migrations des bases de données
- MS SQL Server – AWS Database Migration Service (AWS DMS) Démonstration de réplication
- Atelier pratique sur la modernisation des bases de données
- Échantillons Amazon Neptune

PERF03-BP02 Évaluation des 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 provisionnées 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 d'exposition au risque 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, ainsi que 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 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 provisionnées 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 ce problème, choisissez une clé de partition distribuée plus uniformément ou partitionnez en écriture la clé de partition.

Option de configuration	Exemples
Politiques de gestion du cycle de vie de vos jeux de données	• Vous pouvez utiliser Amazon S3 Lifecycle afin de gérer vos objets au cours de leur cycle de vie. Si vos schémas 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 vers des niveaux d'accès moins coûteux. Vous pouvez tirer parti des métriques Amazon S3 Storage Lens pour identifier les opportunités d'optimisation et les lacunes dans la gestion du cycle de vie. • La fonction de gestion du cycle de vie Amazon EFS gère automatiquement le stockage de vos systèmes de fichiers.
Gestion et regroupement des connexions	• Proxy Amazon RDS 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 stratégies 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 (tel qu'Amazon CloudWatch) et d'optimisation (tel qu'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 volumes Amazon EBS
- Stockage Amazon EC2
- Amazon EFS : Performances d'Amazon EFS
- Amazon FSx pour Lustre Performance
- Amazon FSx for Windows File Server Performance
- Amazon Glacier : documentation Amazon Glacier
- Amazon S3 : directives en matière de débit de demandes et de performances
- 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
- Amazon Athena top 10 performance tips
- Bonnes pratiques Amazon Redshift Spectrum
- Bonnes pratiques Amazon DynamoDB

Vidéos connexes :

- AWS re:Invent 2023: Improve Amazon Elastic Block Store efficiency and be more cost-efficient
- AWS re:Invent 2023: Optimize storage price and performance with Amazon Simple Storage Service
- AWS re:Invent 2023: Building and optimizing a data lake on Amazon Simple Storage Service
- AWS re:Invent 2023: What's new with AWS file storage
- AWSre:Invent 2023: Dive deep into Amazon DynamoDB

Exemples connexes :

- AWS Atelier sur les bases de données sur mesure
- Bases de données pour développeurs
- Journée d'immersion dans l'architecture de données moderne AWS
- Mise à l'échelle automatique d'Amazon EBS
- 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 Collecte et archivage des 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 d'exposition au risque 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

- Identifiez les métriques de performances clés que votre magasin de données doit suivre.
  - Métriques et dimensions d'Amazon S3
  - Surveillance des métriques dans une instance Amazon RDS
  - Surveillance de la charge de la base de données avec Performance Insights sur Amazon RDS
  - Vue d'ensemble de la surveillance améliorée
  - Métriques et dimensions DynamoDB
  - Surveillance de l'accélérateur DynamoDB
  - Surveillance d'Amazon IVS à l'aide d'Amazon CloudWatch
  - Quelles métriques dois-je surveiller?
  - Surveillance des performances de cluster Amazon Redshift
  - Métriques et dimensions Timestream

    - Métriques Amazon CloudWatch pour Amazon Aurora
    - Journalisation et surveillance dans Amazon Keyspaces (pour Apache Cassandra)
    - Surveillance des ressources Amazon Neptune
- 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.
- Vérifiez si la surveillance du magasin de données peut bénéficier d'une solution de machine learning qui détecte les anomalies de performance.
    - Amazon DevOps Guru pour Amazon RDS assure la visibilité des problèmes de performances et suggère des actions correctives.
- Configurez la conservation des données dans votre solution de surveillance et de journalisation en fonction de vos objectifs sécuritaires et opérationnels.
    - Conservation des données par défaut pour les métriques CloudWatch
    - Conservation des données par défaut pour les journaux CloudWatch

Ressources

Documents connexes :

- Mise en cache de bases de données AWS
- Amazon Athena top 10 de conseils en matière de performance
- Bonnes pratiques Amazon Aurora
- DynamoDB Accelerator
- Bonnes pratiques Amazon DynamoDB
- Bonnes pratiques Amazon Redshift Spectrum
- Performances d'Amazon Redshift
- Bases de données cloud avec AWS
- Analyse des performances d'Amazon RDS

Vidéos connexes :

- AWS re:Invent 2022 - Performance monitoring with Amazon RDS and Aurora, featuring Autodesk
- Database Performance Monitoring and Tuning with Amazon DevOps Guru for Amazon RDS
- AWS re:Invent 2023 - What's new with AWS file storage
- AWSre:Invent 2023 - Dive deep into Amazon DynamoDB
- AWS re:Invent 2023 - Building and optimizing a data lake on Amazon S3
- AWS re:Invent 2023 - What's new with AWS file storage
- AWSre:Invent 2023 - Dive deep into Amazon DynamoDB
- Bonnes pratiques pour la surveillance des charges de travail Redis sur Amazon ElastiCache

Exemples connexes :

- Cadre de collecte de métriques pour l'ingestion des jeux de données AWS
- Atelier de surveillance Amazon RDS
- Atelier sur les bases de données sur mesure AWS

PERF03-BP04 Mise en œuvre de 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 d'exposition au risque 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 de 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 requêtes dans Amazon Redshift
  - Utilisation d’EXPLAIN et EXPLAIN ANALYZE sur Athena
- 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 de données dans Athena
    - Partitions et distribution des données
  - L’indexation des données sur les colonnes communes de la requête.
  - Utilisez des vues matérialisées pour les requêtes fréquentes.
    - Compréhension des vues matérialisées
    - Création de vues matérialisées dans Amazon Redshift

  - 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 distribué pour améliorer la latence et réduire le nombre d’opérations d’E/S dans la base de données.
  - Maintenance régulière, telle que l’aspiration, la réindexation et les statistiques d’exécution.
- Expérimentez et testez les stratégies dans un environnement hors production.

Ressources

Documents connexes :

- Bonnes pratiques Amazon Aurora
- Performances d’Amazon Redshift
- Amazon Athena top 10 de conseils en matière de performance
- Mise en cache de bases de données AWS
- Bonnes pratiques de mise en œuvre d’Amazon ElastiCache
- Partitionnement de données dans Athena

Vidéos connexes :

- AWS re:Invent 2023 - AWS storage cost-optimization best practices
- AWS re:Invent 2022 - Performance monitoring with Amazon RDS and Aurora, featuring Autodesk
- Optimize Amazon Athena Queries with New Query Analysis Tools

Exemples connexes :

- AWS Atelier sur les bases de données sur mesure

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 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 de l'application, par exemple au sein de l'application en effectuant des appels à distance ou mise en cache côté client ou en utilisant un service secondaire rapide pour stocker les données 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 l'entrepôt de données dorsales) peut stocker les résultats de ses requêtes uniques localement pendant une durée spécifiée. Cela permet de réduire le nombre de requêtes adressées à un entrepôt de données sur le réseau en vérifiant d'abord le cache du client local. En l'absence de résultats, l'application peut alors interroger l'entrepôt 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 l'entrepôt de données dorsales 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 de données entre clients, un service externe rapide ou un cache distant, peut être utilisé pour stocker les données demandées. Au lieu de vérifier un magasin de données local, chaque client vérifie le cache distant avant d'interroger l'entrepôt de données dorsales. 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

- 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 mettre à l'échelle conviennent à la mise en cache.
  - Mise en cache de bases de données
  - Activation de la mise en cache des API pour améliorer la réactivité
- 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
- Suivez les bonnes pratiques de mise en cache pour votre banque de données.
- 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 l'entrepôt de données dorsales.

- 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 Redis et Amazon ElastiCache (Redis OSS)
- 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
  - Surveillance des bonnes pratiques avec Amazon ElastiCache (Redis OSS) à l'aide d'Amazon CloudWatch
- Mettre en œuvre 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 du cadre Amazon ElastiCache Well-Architected
- Surveillance des bonnes pratiques avec Amazon ElastiCache (Redis OSS) à l'aide d'Amazon CloudWatch
- Quelles métriques dois-je surveiller?
- Livre blanc sur les performances à grande échelle avec Amazon ElastiCache
- Défis et stratégies en matière de mise en cache

Vidéos connexes :

- Parcours de formation Amazon ElastiCache
- Concevoir pour réussir avec les bonnes pratiques Amazon ElastiCache
- AWS re:Invent 2020 – Concevoir pour réussir avec les bonnes pratiques Amazon ElastiCache
- AWS re:Invent 2023 - [LAUNCH] Introducing Amazon ElastiCache Serverless
- AWS re:Invent 2022 - 5 great ways to reimagine your data layer with Redis
- AWS re:Invent 2021 – Présentation approfondie d'Amazon ElastiCache (Redis OSS)

Exemples connexes :

- Améliorer les performances des bases de données MySQL avec Amazon ElastiCache (Redis OSS)

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 de répondre à vos besoins en matière de réseau. AWS propose des fonctionnalités de produit (par exemple, Enhanced Networking, Amazon EC2 networking optimized instances, Amazon S3 transfer acceleration et Amazon CloudFront dynamique) pour optimiser le trafic réseau. AWS propose également des fonctionnalités de mise en réseau (par exemple, Amazon Route 53 latency routing, des points de terminaison Amazon VPC, AWS Direct Connect et AWS Global Accelerator) pour réduire la distance ou la gigue du réseau.

Ce domaine d'intérêt partage des conseils et de bonnes pratiques pour concevoir, configurer et exploiter des solutions de mise en réseau et de diffusion de contenu efficaces dans le cloud.

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 Choisissez l'emplacement de votre charge de travail en fonction des exigences du 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 aide 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.

Veillez à 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 :

- Définissez et documentez les exigences de performance réseau, y compris les métriques telles 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.
- Découvrez les principaux services réseau AWS tels que les VPC, AWS Direct Connect, Elastic Load Balancing (ELB) et Amazon Route 53.
- 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 • AWS Transit Gateway métriques • AWS PrivateLink métriques
Caractéristiques de mise en réseau des applications	• Elastic Fabric Adapter • AWS App Mesh métriques • Métriques pour Amazon API Gateway
Caractéristiques de mise en réseau à la périphérie	• Métriques Amazon CloudFront • Métriques Amazon Route 53 • AWS Global Accelerator métriques
Caractéristiques de mise en réseau hybride	• Direct Connect métriques • AWS Site-to-Site VPN métriques • AWS Client VPN métriques • Métriques WAN AWS Cloud
Caractéristiques de mise en réseau de la sécurité	• Métriques AWS Shield, AWS WAF et AWS Network Firewall
Caractéristiques de traçage	• AWS X-Ray • VPC Reachability Analyzer • Analyseur d'accès réseau

Caractéristiques	Outils et métriques
	• Amazon Inspector • Amazon CloudWatch RUM

- Définition de points de référence et test des performances du réseau :
  - Étalonnez 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.
  - 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
- Passerelle de transit
- Transition vers le routage basé sur la latence dans Amazon Route 53
- Points de terminaison d'un VPC

Vidéos connexes :

- AWS re:Invent 2023 - AWS networking foundations
- AWS re:Invent 2023 - What can networking do for your application?
- AWS re:Invent 2023 - Advanced VPC designs and new capabilities
- AWS re:Invent 2023 - A developer's guide to cloud networking
- AWS re:Invent 2019 - Connectivity to AWS and hybrid AWS network architectures

- AWS re:Invent 2019 - Optimizing Network Performance for Amazon EC2 Instances
- AWS Summit Online - Improve Global Network Performance for Applications
- AWS re:Invent 2020 - Networking best practices and tips with the Well-Architected Framework
- AWS re:Invent 2020 - AWS networking best practices in large-scale migrations

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité de mise à l'échelle
- Ateliers sur la mise en réseau AWS
- Atelier pratique sur le pare-feu réseau
- Observation et diagnostic de votre réseau sur AWS
- Détection et résolution des erreurs de configuration du réseau sur 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 tels que AWS Global Accelerator et Amazon CloudFront qui peuvent contribuer à améliorer les performances du réseau, tandis que la plupart des services AWS proposent des fonctionnalités (telles que la fonctionnalité Amazon S3 Transfer Acceleration) permettant d'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.
  - Pensez à utiliser AWS Cloud WAN pour créer, gérer et surveiller le réseau de votre organisation lors de la création d'un réseau mondial unifié.
  - Surveiller vos réseaux mondiaux et principaux avec les métriques Amazon CloudWatch Logs Tirer parti d'Amazon CloudWatch RUM, qui fournit des informations permettant d'identifier, de comprendre et d'améliorer l'expérience numérique des utilisateurs.
  - Visualisez la latence agrégée du réseau entre Régions AWS et les zones de disponibilité, ainsi qu'à l'intérieur de chaque zone de disponibilité, à l'aide de AWS Network Manager pour mieux comprendre comment la performance de votre application est liée à la performance du réseau AWS sous-jacent.
  - Utilisez un outil de base de données de gestion de la configuration (CMDB) existant ou un service tel que AWS Config pour créer un inventaire de votre charge de travail et de la manière dont elle est configurée.
- 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, effectuez des tests de charge pour identifier les goulots d'étranglement liés aux 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 l'analyseur d'accès réseau pour identifier les chemins ou les itinéraires.
Protocoles réseau	Consultez PERF04-BP05 Choix de protocoles réseau afin d'améliorer les performances.
Topologie du réseau	Évaluez vos compromis opérationnels et de performance entre VPC Peering et AWS Transit Gateway lors de la connexion de plusieurs comptes. AWS Transit Gateway facilite 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. Consultez PERF04-BP03 Choix d'une connectivité dédiée ou d'un VPN approprié pour votre charge de travail.
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

Opportunité d’amélioration	Solution
	diffusion de contenu et de latence à l’échelle mondiale. Lambda@edge vous permet d’exécuter des fonctions qui personnalisent le contenu diffusé par CloudFront au plus près des utilisate urs, de réduire la latence et d’améliorer les performances. Amazon Route 53 propose des options de routage basées sur la latence, de routage de géolocalisation, de routage de géolocali sation et de routage basé sur IP pour vous aider à améliorer les performances de votre charge de travail auprès d’un public mondial. 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	L’accélération de transfert Amazon S3 est une fonctionnalité 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 Amazon S3 répliquent le contenu vers plusieurs régions et simplifient la charge de travail en fournissa nt 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 compartiment 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. Passez en revue vos groupes de placement pour optimiser le débit de votre réseau EC2. Pour éviter un goulot d'étrangl ement par flux, créez votre application 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 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 métrique personnalisée dans CloudWatch. Les adaptateurs réseau élastiques (ENA) d'Amazon permettent d'optimiser davantage le débit de vos instances au sein d'un groupe de placement du cluster. 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 communica tion entre les nœuds à grande échelle sur AWS. Les instances Amazon EBS optimisées utilisent une pile de configuration optimisée et fournissent une capacité supplémentaire dédiée pour augmenter les capacités d'E/S d'Amazon EBS.

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
- Transition vers le routage basé sur la latence dans Amazon Route 53
- Points de terminaison d'un VPC
- Journaux de flux VPC

Vidéos connexes :

- AWS re:Invent 2023 – Ready for what's next? Designing networks for growth and flexibility
- AWS re:Invent 2023 – Advanced VPC designs and new capabilities
- AWS re:Invent 2023 – A developer's guide to cloud networking
- AWS re:Invent 2022 – Dive deep on AWS networking infrastructure
- AWS re:Invent 2019 – Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2018 – Optimizing Network Performance for Amazon EC2 Instances
- AWS Global Accelerator

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité de mise à l'échelle
- Ateliers sur la mise en réseau AWS
- Observation et diagnostic de votre réseau
- Détection et résolution des erreurs de configuration du réseau sur 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 réseau hybride en fonction de vos besoins en bande passante. 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’ Direct Connect.

Si vos besoins en bande passante sont élevés, vous pouvez envisager divers services Direct Connect ou VPN. Le trafic peut être équilibré entre les services, mais nous ne recommandons pas l’équilibrage de charge entre Direct Connect et le VPN en raison des différences de latence et de bande passante.

Étapes d'implémentation

- Évaluez les besoins en bande passante et en latence de vos applications existantes.
  - Pour les charges de travail existantes qui migrent vers AWS, exploitez les données de vos systèmes internes de surveillance du réseau.
  - 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.
- Sélectionnez une connexion dédiée ou un VPN comme option de connectivité. En fonction de toutes les exigences de la charge de travail (besoins en matière de chiffrement, de bande passante et de trafic), vous pouvez choisir AWS Direct Connect ou Site-to-Site VPN (ou les deux). Le schéma suivant peut vous aider à choisir le type de connexion approprié.
  - 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).
  - Le VPN de site à site AWS fournit un service VPN géré prenant en charge la sécurité du protocole Internet (IPsec). Lorsqu'une connexion VPN est créée, chaque connexion VPN comprend deux tunnels pour une haute disponibilité.
- Consultez la documentation AWS pour choisir l'option de connectivité appropriée :
  - Si vous décidez d'utiliser Direct Connect, sélectionnez la bande passante adaptée à votre connectivité.
  - Si vous utilisez un réseau AWS Site-to-Site VPN sur plusieurs sites pour vous connecter à une Région AWS, utilisez une connexion VPN de site à site accélérée pour améliorer les performances du réseau.
  - Si la conception de votre réseau consiste en une connexion VPN IPSec sur AWS Direct Connect, pensez à utiliser un VPN IP privé pour améliorer la sécurité et réaliser une segmentation. AWS Le VPN IP privé de site à site est déployé au-dessus de l'interface virtuelle de transit (VIF).

  - 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 sites AWS Direct Connect, en contournant Régions AWS.
- 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.
- Surveillez régulièrement les performances et l'utilisation de votre connectivité et optimisez-les si nécessaire.

Organigramme des performances déterministes

Ressources

Documents connexes :

- Mise en réseau de produits avec AWS
- AWS Transit Gateway
- Points de terminaison d'un VPC
- Création d'une infrastructure réseau AWS multi-VPC évolutive et sécurisée
- VPN client

Vidéos connexes :

- AWS re:Invent 2023 – Building hybrid network connectivity with AWS
- AWS re:Invent 2023 – Secure remote connectivity to AWS
- AWS re:Invent 2022 – Optimizing performance with Amazon CloudFront
- AWS re:Invent 2019 – Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2020 – AWS Transit Gateway Connect

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité de mise à l'échelle
- AWS Ateliers sur la mise en réseau

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’équilibreur 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. Application Load Balancer est davantage adapté à l'équilibrage de charge du trafic HTTP et HTTPS et fournit un routage avancé des demandes, 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 dorsal 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 dorsales 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 AWS Local Zones 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 dorsales. Pour ce faire, vous pouvez tirer parti des intégrations d'équilibreurs de charge pour les ressources cibles du système dorsal. 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
- Répartition de la charge des applications sur Amazon EKS
- Répartition de charge réseau sur Amazon EKS

Étapes d'implémentation

- Définissez vos exigences en matière d'équilibrage de charge, notamment en matière 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 (ALB comme cible de 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, consultez la comparaison des produits ELB.
- Utilisez le déchargement SSL/TLS si possible.
  - Configurez les écouteurs HTTPS/TLS avec Application Load Balancer et 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 connaître les bonnes pratiques en matière de sécurité, consultez SEC09-BP02 Appliquer le 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 dorsales et donc dans leur impact sur les performances. Par exemple, ALB propose deux options pour les algorithmes de routage :
  - Demandes en suspens les moins nombreuses : permet d'obtenir une meilleure répartition de la charge sur vos cibles dorsales dans les cas où les requêtes de votre application varient en complexité ou vos cibles varient en capacité de traitement.
  - Tour de rôle : 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 la désactivation des zones croisées (isolation zonale) pour améliorer la latence et les domaines de défaillance zonale. Elle est désactivée par défaut dans NLB et dans ALB, vous pouvez la désactiver par groupe cible.
  - Utilisez les zones croisées activées pour une disponibilité et une flexibilité accrues. Par défaut, la zone croisée est activée pour ALB et dans NLB, vous pouvez l'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 dorsales 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 dorsales. Pour savoir comment procéder pour Apache et Nginx, consultez Quels sont les paramètres optimaux pour utiliser Apache ou NGINX en tant que serveur dorsal pour ELB ?
- Activez la surveillance pour votre équilibreur de charge.
  - Activez les journaux d'accès pour votre Application Load Balancer et Network Load Balancer.
  - Les principaux champs à prendre en compte pour ALB sont `request_processing_time`, `request_processing_time`, et `response_processing_time`.
  - Les principaux champs à prendre en compte pour NLB sont `connection_time` et `tls_handshake_time`.
  - Soyez prêt à interroger les journaux lorsque vous en aurez besoin. Vous pouvez utiliser Amazon Athena pour interroger les journaux ALB et les journaux NLB.
  - Créez les alarmes pour les métriques liées aux performances, telles que `TargetResponseTime` pour ALB.

Ressources

Documents connexes :

- Comparaison des produits ELB
- AWSInfrastructure mondiale
- Amélioration des performances et réduction des coûts grâce à l'affinité des zones de disponibilité
- Procédure détaillée d'analyse des journaux avec Amazon Athena
- Interrogation des journaux de l'application Load Balancer
- Surveillance de vos Application Load Balancers
- Surveillance de vos Network Load Balancers
- Utiliser Elastic Load Balancing pour répartir le trafic sur les instances dans votre groupe Auto Scaling.

Vidéos connexes :

- AWS re:Invent 2023: qu'est-ce que la mise en réseau peut faire pour votre application ?
- AWS re:Inforce 20 : comment utiliser Elastic Load Balancing pour améliorer votre posture de sécurité à l'échelle
- AWS re:Invent 2018 : Elastic Load Balancing : Plongée en profondeur et bonnes pratiques
- AWS re:Invent 2021 - How to choose the right load balancer for your AWS workloads
- AWS re:Invent 2019: Get the most from Elastic Load Balancing for different workloads

Exemples connexes :

- Passerelle équilibreur de charge
- CDK et CloudFormation exemples 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 protocole SRD est donc le mieux adapté aux charges de travail de 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 le transfert d'un gros volume de données entre les nœuds.

Étapes d'implémentation

- 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 mettre à l'échelle et gérer en toute sécurité vos transferts de fichiers vers des services de stockage AWS.

- 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 cliente 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 et retour (RTT) peuvent être utilisées pour mesurer la latence du réseau. Si votre charge de travail sert des contenus dynamiques aux utilisateurs, envisagez d'utiliser Amazon CloudFront, qui établit une connexion persistante avec chaque origine de contenu dynamique afin de supprimer le temps d'établissement de la connexion qui, sinon, ralentirait chaque demande du client.
- 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 le délestage SSL/TLS sur Elastic Load Balancing pour améliorer les performances de la charge de travail en permettant à l'équilibreur de charge de gérer le processus de cryptage et de décryptage SSL/TLS au lieu de laisser les instances dorsales s'en charger. Cela peut contribuer à réduire l'utilisation du processeur sur les instances dorsales, ce qui peut améliorer les performances et augmenter la capacité.
- Utilisez le Network Load Balancer (NLB) pour déployer des services reposant sur le protocole UDP, tels que l'authentification et l'autorisation, la journalisation, DNS, IoT et le média de streaming, 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 mettre à l'échelle votre charge de travail horizontalement, d'augmenter la capacité et de réduire les frais généraux associés à une seule cible.
- Pour vos charges de travail de calcul haute performance (HPC), pensez à utiliser la fonctionnalité Adaptateur réseau élastique (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.
- Utilisez Application Load Balancer (ALB) pour acheminer et équilibrer la charge de 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 matière 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 :

- Comment router le trafic UDP dans Kubernetes ?
- Application Load Balancer
- Mise en réseau améliorée d'EC2 sous Linux
- Capacité réseau améliorée d'EC2 sous Windows
- Groupes de placement EC
- 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
- Transition vers le routage basé sur la latence dans Amazon Route 53
- Points de terminaison d'un VPC

Vidéos connexes :

- AWS re:Invent 2022 – Scaling network performance on next-gen Amazon Elastic Compute Cloud instances
- AWS re:Invent 2022 – Application networking foundations

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité de mise à l'échelle
- Ateliers sur la mise en réseau AWS

PERF04-BP06 Choisissez l'emplacement de votre charge de travail en fonction des exigences du 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 un réseau mondial AWS privé Régions AWS et approprié, 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 EC2 les instances Amazon, sont placées dans des zones de disponibilité Régions AWSinternes AWS Outposts, des zones AWS locales ou AWS Wavelengthdes zones. 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 tels qu'Amazon CloudFront AWS Global Acceleratorpeuvent également être utilisés pour améliorer les performances du réseau soit en mettant en cache le contenu sur des sites périphériques, soit en fournissant aux utilisateurs un chemin optimal vers la charge de travail via le réseau AWS mondial.

Amazon EC2 propose 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 dotés de types d'instances compatibles et d'un adaptateur réseau élastique (ENA) permet aux charges de travail de participer à un réseau de 25 Gbit/s à faible latence et à 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 sites périphériques via un réseau AWS mondial, tel qu'Amazon. CloudFront Ces emplacements périphériques fournissent généralement des services tels que le réseau de diffusion de contenu (CDN) et le système de noms de domaine (DNS). En disposant de ces services à la périphérie, les charges de travail peuvent répondre avec une faible latence aux demandes de contenu ou de DNS résolution. 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 deux DNS etHTTP/HTTPSafin de tirer le meilleur parti 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 de VPC Flow Logs
  - 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, tels qu'Amazon CloudWatch 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 :
  - Lieu de stockage 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.
  - Lieu de stockage de vos données : pour les applications orientées utilisateur, choisissez une région (ou des régions) proche des utilisateurs de votre charge de travail.
  - Autres contraintes : tenez compte des contraintes telles que le coût et la conformité, comme expliqué dans la section Éléments à prendre en compte lors de la sélection d'une région pour vos charges de travail.
- Utilisez des zones locales AWS pour exécuter des charges de travail telles que le rendu vidéo. Les 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 haute résolution, le son haute fidélité et la réalité augmentée ou virtuelle (AR/VR) nécessitent ultra-low-latency des appareils 5G. Pour de telles applications, considérez AWS Wavelength. AWS Wavelength intègre des services de AWS

calcul et de stockage dans les réseaux 5G, fournissant une infrastructure informatique de pointe mobile pour le développement, le déploiement et la mise à l'échelle d' ultra-low-latencyapplications.

- Utilisez des solutions de mise en cache locale ou proposées par AWS pour les ressuorces fréquemment utilisées afin d'améliorer les performances, de réduire les déplacements de données et de diminuer l'impact environnemental.

Service	Utilisation
Amazon CloudFront	Utilisez-le pour mettre en cache du contenu statique tel que des images, des scripts et des vidéos, ainsi que du contenu dynamique tel que API des réponses 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	Utilisation
Lambda@Edge	Destiné aux opérations exigeantes en puissance de calcul qui sont lancées lorsque des objets ne sont pas dans le cache.
CloudFront Fonctions Amazon	À utiliser pour des cas d'utilisation simples tels que HTTP des requêtes ou des manipulations de réponses qui peuvent être initiées 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 anycast statiques et des TCP terminaisons aux emplacements périphériques. AWS Global Acceleratorpeut améliorer les performances de vos applications jusqu'à 60 % et permettre un basculement rapide pour les architectures multirégionales. AWS Global Accelerator vous fournit des adresses IP anycast statiques qui servent de point d'entrée fixe pour vos applications hébergées dans une ou plusieurs d' Régions AWS entre elles. Ces adresses IP permettent au trafic de pénétrer sur le réseau AWS mondial aussi près que possible de vos utilisateurs. AWS Global Accelerator réduit le temps de configuration de la connexion initiale en établissant une TCP connexion entre le client et l'emplacement AWS périphérique le plus proche du client. Passez en revue l'utilisation de AWS Global Accelerator pour améliorer les performances de vosTCP/UDPworkloads et permettre un basculement rapide pour les architectures multirégionales.

Ressources

Bonnes pratiques associées :

- COST07-BP02 Mettre en œuvre les régions en fonction des coûts
- COST08-BP03 Mettre en œuvre des services pour réduire les coûts de transfert de données
- REL10-BP01 Déployer la charge de travail sur plusieurs sites
- REL10-BP02 Sélectionnez les emplacements appropriés pour votre déploiement multisite
- SUS01-BP01 Choisissez la région en fonction des exigences commerciales et des objectifs de durabilité
- SUS02-BP04 Optimiser le placement géographique des charges de travail en fonction de leurs exigences en matière de réseau
- SUS04-BP07 Minimiser le mouvement des données sur les réseaux

Documents connexes :

- AWS Infrastructure mondiale
- AWS Zones locales et AWS Outposts choix de la technologie adaptée à votre charge de travail périphérique
- Groupes de placement
- AWS Zones Locales

- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Vidéos connexes :

- AWS Vidéo explicative sur les Zones Locales
- AWS Outposts: Overview and How it Works
- AWS re:Invent 2023 - Une stratégie de migration pour les charges de travail en périphérie et sur site
- AWS re:INVENT 2021 - AWS Outposts : Apporter l' AWS expérience sur site
- AWS re:Invent 2020 : AWS Wavelength : Exécutez des applications avec une latence extrêmement faible à la périphérie de la 5G
- AWS re:Invent 2022 - Zones AWS locales : création d'applications pour une périphérie distribuée
- AWS re:Invent 2021 - Création de sites Web à faible latence avec Amazon CloudFront
- AWS re:Invent 2022 - Améliorez les performances et la disponibilité avec AWS Global Accelerator
- AWS re:Invent 2022 - Construisez votre réseau étendu mondial en utilisant AWS
- AWS re:Invent 2020 : gestion du trafic mondial avec Amazon Route 53

Exemples connexes :

- AWS Global Accelerator Atelier de routage personnalisé
- 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 : bas

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 (IPAM).	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 pouvez optimiser les configurations DNS en comprenant le domaine ou le sous-doma ine qui a été demandé ou les emplacements périphériques Route 53 qui ont répondu aux requêtes DNS.

Outil AWS	Où utiliser
Reachability Analyzer	Reachability Analyzer vous aide à analyser et à déboguer l'accessibilité du réseau. Reachability Analyzer est un outil d'analyse de configuration qui vous permet d'effectuer des tests de connectivité entre une ressource source et une ressource de destination dans vos clouds privés virtuels (VPC). Cet outil vous aide à vérifier que votre configuration réseau correspond à la connectivité souhaitée.
Analyseur d'accès réseau	Vous pouvez utiliser l'Analyseur d'accès réseau pour comprendre l'accès réseau à vos ressources. Vous pouvez utiliser l'analyseur d'accès réseau 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 optimisan t la configuration de votre réseau correspon dant, vous pouvez comprendre et vérifier l'état de votre réseau et démontrer si votre réseau sur AWS répond à vos exigences de conformit é.

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, l'équilibreur de charge Elastic 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 vous permet de surveiller les performances historiques et en temps réel du réseau mondial AWS à des fins opérationnelles et de planification. Network Manager fournit la latence agrégée du réseau entre Régions AWS et les zones de disponibilité et à l'intérieur de chaque zone de disponibilité, ce qui vous permet de mieux comprendre comment la performance de votre application est liée à la performance 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. L'Analyseur d'accès réseau vous aide à le faire.

Ressources

Documents connexes :

- Journalisation des requêtes DNS publiques
- Qu'est-ce qu'IPAM ?
- Définir Reachability Analyzer
- Définir l'Analyseur d'accès réseau
- Métriques CloudWatch pour vos VPC
- Optimiser les performances et réduire les coûts de l'analyse 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 :

- AWS re:Invent 2023 – A developer's guide to cloud networking
- AWS re:Invent 2023 – Ready for what's next? Designing networks for growth and flexibility
- AWS re:Invent 2023 - Advanced VPC designs and new capabilities
- AWS re:Invent 2022 – Dive deep on AWS networking infrastructure
- AWS re:Invent 2020 – Networking best practices and tips with the AWS Well-Architected Framework
- AWS re:Invent 2020 – Monitoring and troubleshooting network traffic

Exemples connexes :

- Ateliers sur la mise en réseau AWS
- Surveillance réseau AWS
- Observation et diagnostic de votre réseau sur AWS
- Détection et résolution des erreurs de configuration du réseau sur AWS

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. Ce domaine d'intérêt propose les bonnes pratiques pour l'adoption d'une culture qui favorise l'efficacité des performances des charges de travail dans le cloud.

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. Vous pouvez ainsi appliquer les pratiques utilisées pour développer des logiciels à votre infrastructure et 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. Vous pouvez ainsi déployer de manière reproductible et cohérente, le tout à un faible coût, à mesure que vous itérez.
- Métriques bien définies : configurez vos métriques et votre solution de surveillance 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. D'autres mesures généralement applicables comprennent le nombre de threads, le taux de récupérateur de 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 bien effectués. 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. Vous pouvez ainsi 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'examen 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 de procéder à 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.

Bonnes pratiques

- PERF05-BP01 Définition d'indicateurs de rendement clés (KPI) pour mesurer l'état et les performances de la charge de travail
- PERF05-BP02 Utilisation de solutions de surveillance pour comprendre les domaines où les performances sont d'une importance critique
- PERF05-BP03 Définition d'un processus pour améliorer les performances des charges de travail
- PERF05-BP04 Testez votre charge de travail
- PERF05-BP05 Utilisation de l'automatisation pour résoudre de manière proactive les problèmes liés aux performances
- PERF05-BP06 Maintenez votre charge de travail et vos services up-to-date
- PERF05-BP07 Vérification des métriques à intervalles réguliers

PERF05-BP01 Définition d’indicateurs de rendement clés (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 de ces métriques.
- Vous supposez que vos KPI sont déjà en cours de publication et de partage 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 d’exposition au risque 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

- Identification des parties prenantes : identifiez et documentez les principales parties prenantes de l'entreprise, y compris les équipes de développement et d'exploitation.
- Définition d'objectifs : collaborez avec ces parties prenantes pour définir et documenter les objectifs de votre charge de travail. Tenez compte des aspects critiques des performances de vos charges de travail, tels que le débit, le temps de réponse et le coût, ainsi que des objectifs métier, tels que la satisfaction des utilisateurs.
- Passage en revue des bonnes pratiques du secteur : passez en revue les bonnes pratiques du secteur pour identifier les KPI pertinents qui correspondent à vos objectifs en matière de charge de travail.
- Identification des métriques : identifiez les métriques qui correspondent aux objectifs de votre charge de travail et qui peuvent vous aider à mesurer les performances et les objectifs commerciaux. Établissez des KPI sur la base de ces métriques. Les mesures telles que le temps de réponse moyen ou le nombre d'utilisateurs simultanés sont des exemples de métriques.
- Définition et documentation des KPI : utilisez les bonnes pratiques du secteur et les objectifs de votre 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. Identifiez et documentez le risque et l'impact du non-respect d'un KPI.
- Mise en œuvre de la surveillance : utilisez des outils de surveillance tels qu'Amazon CloudWatch ou AWS Config pour collecter des métriques et mesurer les KPI.
- Communication visuelle des KPI : utilisez des outils de tableau de bord tels qu'Amazon Quick pour visualiser et communiquer les indicateurs de performance clés aux parties prenantes.

- Analyse et optimisation : passez en revue et analysez régulièrement les KPI pour identifier les domaines de votre charge de travail qui doivent être améliorés. Collaborez avec les parties prenantes pour mettre en œuvre ces améliorations.
- Révision et affinage : passez régulièrement en revue les indicateurs et les indicateurs de performance clés pour évaluer leur efficacité, en particulier lorsque les objectifs commerciaux ou les performances de la charge de travail changent.

Ressources

Documents connexes:

- Documentation CloudWatch
- Surveillance, journalisation et performances AWS Partner
- Outils d'observabilité d'AWS
- L'importance des indicateurs de rendement clés (KPI) pour les migrations vers le cloud à grande échelle
- Suivi des KPI d'optimisation des coûts avec KPI Dashboard
- Documentation X-Ray
- Utilisation des tableaux de bord Amazon CloudWatch
- KPI de Quick

Vidéos connexes :

- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2023 - Manage resource lifecycle events at scale with AWS Health
- AWS re:Invent 2023 - Performances & efficiency at Pinterest: Optimizing the latest instances
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2023 – Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2023 - Scaling on AWS for the first 10 million users
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- Creating an Effective Metrics Strategy for Your Business | AWS Events

Exemples connexes :

- Création d'un tableau de bord avec Quick

PERF05-BP02 Utilisation de 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 de clients peut 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.

Améliorez l'efficacité des performances en comprenant votre architecture, vos modèles de trafic et d'accès aux données, et identifiez vos temps de latence et de traitement. Identifiez 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

- 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
Surveillance des utilisateurs réels (RUM) avec Amazon CloudWatch	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.
Informations sur les performances du service de base de données relationnelle Amazon	Pour consulter les métriques de performan ces de la base de données et identifier les améliorations des performances.
Surveillance améliorée Amazon RDS	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.

- 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 les scripts Canary synthétiques CloudWatch pour imiter par programmation les activités des utilisateurs basées sur le navigateur à l'aide de tâches cron Linux ou de valeurs de déclenchement afin de générer des métriques cohérentes au fil du temps.
  - Utiliser le test de charge distribué AWS afin de générer un trafic de pointe ou de tester la charge de travail au taux de croissance attendu.
- 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.
- Expérimentez des améliorations des performances et mesurez ces changements avec des données. Par exemple, vous pouvez utiliser CloudWatch Evidently pour tester les nouvelles améliorations et les impacts sur les performances de votre charge de travail.

Ressources

Documents connexes :

- Les nouveautés en matière d'observabilité AWS à re:Invent 2023
- Bibliothèque Amazon Builders' Library
- Documentation X-Ray
- Amazon CloudWatch RUM
- Amazon DevOps Guru

Vidéos connexes :

- AWS re:Invent 2023 - [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 - Implementing application observability
- AWS re:Invent 2023 – Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2022 – The Amazon Builders' Library: 25 years of Amazon operational excellence
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- Visual Monitoring of Applications with Amazon CloudWatch Synthetics

Exemples connexes :

- Measure page load time with Amazon CloudWatch Synthetics
- Client Web Amazon CloudWatch RUM
- X-Ray SDK pour Python
- Test de charge distribuée sur AWS

PERF05-BP03 Définition d’un processus pour améliorer les performances des charges de travail

Définissez un processus d’évaluation de nouveaux services, modèles de conception, types de ressources et configurations au fur et à mesure qu’ils 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 d’exposition au risque 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

- Identification des KPI : identifiez les KPI de performance de votre charge de travail comme indiqué dans la section pour établir une base PERF05-BP01 Définition d'indicateurs de rendement clés (KPI) pour mesurer l'état et les performances de la charge de travail de référence de votre charge de travail.
- Mise en œuvre du suivi : utilisez des outils AWS d'observabilité pour collecter des indicateurs de performance et mesurer les KPI.
- Réalisation d'une analyse : 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 Utilisation de 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 les stratégies d'amélioration des performances.
- Validation des améliorations : utilisez des environnements de test (sandbox) ou en préproduction pour valider l'efficacité des stratégies d'amélioration.
- Mise en œuvre des modifications : 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.
- Révision et affinage : passez régulièrement en revue votre processus d'amélioration des performances afin d'identifier les domaines à améliorer.

Ressources

Documents connexes :

- AWS Blog
- Nouveautés avec AWS
- AWS Skill Builder

Vidéos connexes :

- AWS re:Invent 2022 – Delivering sustainable, high-performing architectures

- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2022 - Optimize your AWS workloads with best-practice guidance

Exemples connexes :

- AWS Github

PERF05-BP04 Testez 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 EC2 test d'Amazon et sans soumettre de formulaire de soumission d'événements simulés. Cela entraîne l'échec de votre test, car il ressemble à un denial-of-service événement.

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 d'exposition au risque 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 par rapport à des seuils KPIs et à des seuils prédéfinis. Ce processus vous permet de continuer à atteindre les performances requises.

Étapes d'implémentation

- Définition de vos objectifs de test : identifiez les aspects de performance de votre charge de travail que vous souhaitez évaluer, tels que le débit et le temps de réponse.
- Sélection d'un outil de test : choisissez et configurez l'outil de test de charge adapté à votre charge de travail.
- Configuration de votre environnement : configurez l'environnement de test en fonction de votre environnement de production. Vous pouvez utiliser AWS les services pour exécuter des environnements de production afin de tester votre architecture.
- Mettez en œuvre la surveillance : utilisez des outils de surveillance tels qu'Amazon CloudWatch pour collecter des métriques sur les ressources de votre architecture. Vous pouvez également collecter et publier des métriques personnalisées.
- Définition de des scénarios : définissez les scénarios et les paramètres de test de charge (tels que la durée du test et le nombre d'utilisateurs).
- Tests de charge : réalisez des scénarios de test à grande échelle. Profitez-en AWS Cloud pour tester votre charge de travail afin de découvrir où elle ne parvient pas à évoluer 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.
- Analyse des résultats des tests : analysez les résultats pour identifier les goulots d'étranglement en matière de performances et les domaines à améliorer.
- Documentation et partage des résultats : documentez et rendez compte des résultats et des recommandations. Partagez ces informations avec les parties prenantes pour les aider à prendre des décisions éclairées concernant les stratégies d'optimisation des performances.
- Itération continuelle : les tests de charge doivent être effectués à une cadence régulière, en particulier après un changement ou une mise à jour du système.

Ressources

Documents connexes :

- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Test de charge distribué sur AWS

Vidéos connexes :

- AWS Sommet ANZ 2023 : Accélérez en toute confiance grâce AWS aux tests de charge distribués
- AWS re:Invent 2022 - Tirez parti AWS de vos 10 premiers millions d'utilisateurs
- Résoudre avec AWS des solutions : tests de charge distribués

- AWS re:Invent 2021 - Optimisez les applications grâce aux informations des utilisateurs finaux avec Amazon CloudWatch RUM
- Démo d'Amazon CloudWatch Synthetics

Exemples connexes :

- Test de charge distribué sur AWS

PERF05-BP05 Utilisation de 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 d'exposition au risque si cette bonne pratique n'est pas respectée : bas

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

- Identification du processus de remédiation : identifiez et comprenez le problème lié aux performances qui peut être résolu automatiquement. Utilisez des solutions de surveillance AWS telles qu'Amazon CloudWatch ou AWS X-Ray pour vous aider à mieux comprendre la cause première du problème.
- Définition du processus d'automatisation : créez un plan et un processus de résolution étape par étape qui peuvent être utilisés pour résoudre automatiquement le problème.
- Configuration de l'événement d'initiation : configurez l'événement pour lancer automatiquement le processus de correction. Par exemple, vous pouvez définir un déclencheur pour redémarrer automatiquement une instance lorsqu'elle atteint un certain seuil d'utilisation de l'UC.
- Automatisation de la remédiation : 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. Veillez à utiliser une logique d'auto-réparation pour annuler les modifications si elles ne permettent pas de résoudre le problème.
- Test du flux de travail : testez le processus de résolution automatisé dans un environnement de pré-production.
- Mise en œuvre du flux de travail : implémentez la correction automatique dans l'environnement de production.

- Élaboration d'un manuel : élaborez et documentez un manuel qui décrit les étapes du plan de remédiation, y compris les événements initiateurs, la logique de remédiation et les mesures prises. Veillez à former les parties prenantes pour les aider à répondre efficacement aux événements de résolution automatisée.
- Révision et affinage : évaluez régulièrement l'efficacité du flux de travail de correction automatisé. Ajustez les événements de lancement et la logique de résolution, si nécessaire.

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
- Build a Cloud Automation Practice for Operational Excellence: Best Practices from AWS Managed Services
- Automatisez le réglage des performances de votre Amazon Redshift grâce à l'optimisation automatique des tables

Vidéos connexes :

- AWS re:Invent 2023 - Strategies for automated scaling, remediation, and smart self-healing
- AWS re:Invent 2023 - [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 - Implementing application observability
- AWS re:Invent 2021 - Intelligently automating cloud operations
- AWS re:Invent 2022 - Setting up controls at scale in your AWS environment
- AWS re:Inforce 2022 - Automating patch management and compliance using AWS
- AWSre:Invent 2022 - How Amazon uses better metrics for improved website performance
- AWSre:Invent 2023 - Take a load off: Diagnose & resolve performance issues with Amazon RDS
- AWSre:Invent 2021 - {New Launch} Automatically detect and resolve issues with Amazon DevOps Guru
- AWSre:Invent 2023 - Centralize your operations

Exemples connexes :

- Personnalisation des alarmes Cloudwatch Logs

PERF05-BP06 Maintenez votre charge de travail et vos services up-to-date

Restez up-to-date sur les nouveaux services et fonctionnalités du 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 de la mise en place de cette meilleure pratique : en établissant un processus pour rester à up-to-date jour avec les nouveaux services et offres, vous pouvez adopter de nouvelles fonctionnalités, résoudre les problèmes et améliorer les performances de la charge de travail.

Niveau d’exposition au risque 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, créez une démonstration de faisabilité qui utilise les nouvelles technologies ou consultez 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

- Inventaire de votre charge de travail : établissez l’inventaire de votre logiciel de charge de travail et de l’architecture, et identifiez les composants pouvant être mis à jour.
- Identification des sources de mise à jour : identifiez les actualités et mettez à jour les sources liées aux composants de votre charge de travail. Par exemple, vous pouvez vous abonner au AWS blog What's New at pour découvrir les produits correspondant à votre composante de charge de travail. Vous pouvez vous abonner au RSS flux ou gérer vos abonnements par e-mail.
- Définition d’un calendrier de mise à jour : 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 collecter les métadonnées du système d'exploitation (OS), des applications et des instances à partir de vos EC2 instances Amazon et comprendre rapidement quelles instances exécutent le logiciel et les configurations requises par votre politique logicielle et quelles instances doivent être mises à jour.
- Évaluation de la nouvelle mise à jour : comprenez comment mettre à jour les composants de votre charge de travail. Profitez de l’agilité du cloud pour tester rapidement la façon dont les nouvelles fonctionnalités peuvent améliorer votre charge de travail afin de gagner en efficacité.
- Utiliser l’automatisation : 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 à jour AMIs automatiquement des images de conteneur et d'autres artefacts liés à votre application cloud.
  - Vous pouvez utiliser des outils tels que AWS Systems Manager Patch Manager pour automatiser le processus de mise à jour du système et planifier l’activité à l’aide de AWS Systems Manager Maintenance Windows.
- Documentation du processus : 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 commerciales documentées et aidez KPIs à hiérarchiser les mises à jour qui auront un impact commercial positif.

Ressources

Documents connexes :

- Blog AWS
- Quoi de neuf avec AWS
- up-to-dateImplémentation d'images avec des pipelines EC2 Image Builder automatisés

Vidéos connexes :

- AWS Re:inForce 2022 - Automatisation de la gestion des correctifs et de la conformité à l'aide de AWS
- All Things Patch : AWS Systems Manager | AWS Événements

Exemples connexes :

- Gestion de l’inventaire et des correctifs
- Un atelier sur l’observabilité

PERF05-BP07 Vérification des 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 : passez 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 d’exposition au risque 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

- Définition de métriques : définissez des métriques de performance critiques à surveiller qui correspondent à votre objectif de charge de travail, notamment des métriques telles que le temps de réponse et l’utilisation des ressources.
- Établissement de bases de référence : définissez une base de référence et une valeur souhaitable pour chaque métrique. La base de référence doit fournir des points de référence pour identifier les écarts ou les anomalies.
- Établissement d’une cadence : définissez une cadence (hebdomadaire ou mensuelle, par exemple) pour examiner les métriques critiques.
- Identification des problèmes de performance : 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. Pour les problèmes identifiés, effectuez une analyse détaillée des causes profondes afin de comprendre la raison principale du problème.
- Identification des actions correctives : utilisez votre analyse pour identifier les actions correctives. Cela peut inclure le réglage des paramètres, la correction de bogues et la mise à l’échelle des ressources.
- Documentation des résultats : documentez vos conclusions, y compris les problèmes identifiés, les causes profondes et les mesures correctives.
- Répétition et amélioration : évaluez et améliorez en permanence le processus de révision des métriques. Utilisez les enseignements tirés de la révision précédente pour améliorer le processus au fil du temps.

Ressources

Documents connexes :

- Documentation CloudWatch
- Collecte de métriques et de journaux à partir d'instances Amazon EC2 et de serveurs sur site avec l'agent CloudWatch
- Interrogation de vos métriques avec CloudWatch Metrics Insights
- Surveillance, journalisation et performances : partenaires AWS Partner Network
- Documentation X-Ray

Vidéos connexes :

- AWS re:Invent 2022 - Setting up controls at scale in your AWS environment
- AWSre:Invent 2022 - How Amazon uses better metrics for improved website performance
- AWS re:Invent 2023 – Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWSre:Invent 2023 - Take a load off: Diagnose & resolve performance issues with Amazon RDS

Exemples connexes :

- Création d'un tableau de bord avec Quick
- Tableaux de bord CloudWatch

Conclusion

Pour atteindre et maintenir l’efficacité des performances, il est nécessaire d’avoir une approche axée sur les données. Vous devriez sérieusement envisager des modèles d’accès et des compromis qui vous permettront d’optimiser les performances. L’utilisation d’un processus d’évaluation basé sur des comparatifs et des tests de charge vous permet de sélectionner les configurations et types de ressources appropriés. En traitant votre infrastructure comme du code, vous pouvez faire évoluer votre architecture rapidement et en toute sécurité tout en utilisant les données pour prendre des décisions basées sur des faits en ce qui concerne votre architecture. La mise en place d’une surveillance à la fois active et passive permet de s’assurer que les performances de votre architecture ne se dégradent pas au fil du temps.

AWS s'efforce de vous aider à créer des architectures performantes tout en apportant une valeur commerciale. Utilisez les outils et techniques présentés dans ce document pour garantir votre réussite.

Collaborateurs

Les personnes et organisations suivantes ont contribué à l'élaboration du présent document :

- Sam Mokhtari, architecte principal de solutions en matière d'efficacité, Amazon Web Services
- Josh Hart, architecte de solutions, Amazon Web Services
- Richard Trabing, architecte de solutions, Amazon Web Services
- Brett Looney, architecte principal de solutions, Amazon Web Services
- Nina Vogl, architecte principal de solutions, Amazon Web Services
- Eric Pullen, architecte de solutions, Amazon Web Services
- Julien Lépine, responsable des architectes de solutions spécialisés, Amazon Web Services
- Ronnen Slasky, architecte de solutions, Amazon Web Services

Suggestions de lecture

Pour obtenir de l’aide, consultez les ressources suivantes :

- Framework AWS Well-Architected
- Centre d’architecture AWS

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 mineure des bonnes pratiques	La bonne pratique PERF03-BP04 a été mise à jour avec de nouvelles recommand ations de service.	6 novembre 2024
Mises à jour des conseils sur les bonnes pratiques	Plusieurs petites mises à jour dans l'ensemble du pilier.	27 juin 2024
Mise à jour et restructuration majeures	Ce pilier a été restructuré pour inclure cinq domaines de bonnes pratiques (contre huit auparavant). Le contenu a été regroupé dans ces cinq domaines et a été mis à jour. Les nouveaux domaines de bonnes pratiques sont la sélection de l'architecture, le calcul et le matériel, la gestion des données, lamise en réseau et la diffusion de contenu, ainsi que les processus et la culture.	3 octobre 2023
Mise à jour mineure	Suppression du langage non inclusif.	13 avril 2023
Mises à jour du nouveau cadre	Les bonnes pratiques ont été mises à jour avec des recommandations et de nouvelles bonnes pratiques.	10 avril 2023

Livre blanc mis à jour	Les bonnes pratiques ont été mises à jour avec de nouvelles recommandations en matière d'implémentation.	15 décembre 2022
Livre blanc mis à jour	Développement des bonnes pratiques et ajout de plans d'amélioration.	20 octobre 2022
Mise à jour mineure	Suppression du langage non inclusif.	22 avril 2022
Mises à jour mineures	Mise à jour des liens.	10 mars 2021
Mises à jour mineures	Le délai d'attente AWS Lambda a été modifié à 900 secondes et le nom d'Amazon Keyspaces (pour Apache Cassandra) a été corrigé.	5 octobre 2020
Mise à jour mineure	Correction d'un lien rompu.	15 juillet 2020
Mises à jour du nouveau cadre	Révision et mise à jour majeures du contenu	8 juillet 2020
Livre blanc mis à jour	Mise à jour mineure pour les problèmes grammaticaux	1er juillet 2018
Livre blanc mis à jour	Actualisation du livre blanc pour refléter les modifications apportées à AWS	1er novembre 2017
Publication initiale	Pilier Efficacité des performances – AWS Well-Architected Framework publié.	1er novembre 2016

Avis

Il incombe aux clients de procéder à une évaluation indépendante des informations contenues dans le présent document. Ce document : (a) est fourni à titre informatif uniquement, (b) représente les offres de AWS produits et les pratiques actuelles, qui sont susceptibles d'être modifiées sans préavis, et (c) ne crée aucun engagement ni aucune assurance de la part de AWS ses filiales, fournisseurs ou concédants de licence. AWS les produits ou services sont fournis « tels quels » sans garanties, déclarations ou conditions d'aucune sorte, qu'elles soient explicites ou implicites. Les responsabilités et obligations AWS de ses clients sont régies par AWS des accords, et ce document ne fait partie d'aucun accord conclu entre AWS et ses clients et ne les modifie pas.

© 2023, Amazon Web Services, Inc. ou ses sociétés apparentées. Tous droits réservés.

AWS Glossaire

Pour la AWS terminologie la plus récente, consultez le AWS glossaire dans la Glossaire AWS référence.