Menskalakan infrastruktur Amazon EKS untuk mengoptimalkan komputasi, beban kerja, dan kinerja jaringan

AWS Panduan Preskriptif

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Panduan Preskriptif: Menskalakan infrastruktur Amazon EKS untuk mengoptimalkan komputasi, beban kerja, dan kinerja jaringan

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Merek dagang dan tampilan dagang Amazon tidak boleh digunakan sehubungan dengan produk atau layanan apa pun yang bukan milik Amazon, dengan cara apa pun yang dapat menyebabkan kebingungan di antara pelanggan, atau dengan cara apa pun yang merendahkan atau mendiskreditkan Amazon. Semua merek dagang lain yang tidak dimiliki oleh Amazon merupakan hak milik masing-masing pemiliknya, yang mungkin atau tidak terafiliasi, terkait dengan, atau disponsori oleh Amazon.

Table of Contents

Pengantar ... 1
Tujuan ... 2
Hitung penskalaan ... 4
Cluster AutoScaler ... 4
Cluster Autoscaler dengan penyediaan berlebih ... 5
Karpenter ... 5
Penskalaan beban kerja ... 7
Penskala Otomatis Pod Horizontal ... 7
Cluster Proporsional Autoscaler ... 8
Autoscaler Berbasis Event berbasis Kubernetes ... 9
Penskalaan jaringan ... 11
Plugin Amazon VPC CNI untuk Kubernetes ... 11
Jaringan kustom ... 12
Delegasi awalan ... 13
Kisi VPC Amazon ... 14
Optimalisasi biaya ... 16
Kubecost ... 16
Goldilocks ... 17
AWS Fargate ... 18
Instans Spot ... 18
Instans Terpesan ... 19
AWS Contoh graviton ... 20
Langkah selanjutnya ... 22
Sumber daya ... 23
Riwayat dokumen ... 24
Glosarium ... 25
... 25
A ... 26
B ... 29
C ... 31
D ... 34
E ... 38
F ... 40
G ... 42

H 43
I 44
L 47
M 48
O 53
P 55
Q 58
R 59
D 62
T 66
U 67
V 68
W 68
Z 69
..... lxxi

Menskalakan infrastruktur Amazon EKS untuk mengoptimalkan komputasi, beban kerja, dan kinerja jaringan

Aniket Dekate, Aniket Kurzadkar, dan Ishwar Chathaiwale, Amazon Web Services ()AWS

November 2024 (riwayat dokumen)

Amazon Elastic Kubernetes Service (Amazon EKS) adalah layanan Kubernetes yang dikelola. Dengan Amazon EKS, Anda dapat menjalankan pod Kubernetes di lingkungan cloud yang di-containerized tanpa perlu menginstal dan mengoperasikan control plane Anda sendiri. Dengan AWS mengelola bidang kontrol, Amazon EKS mengurangi manajemen operasional organisasi. Manfaat lain menggunakan Amazon EKS termasuk penskalaan, keandalan, dan keamanan di lingkungan cloud.

Panduan ini dirancang untuk membantu organisasi mengoptimalkan infrastruktur Amazon EKS mereka di bidang-bidang berikut:

- Compute scaling adalah komponen penting untuk performa aplikasi di lingkungan Kubernetes yang dinamis:
  - Alokasi sumber daya yang efisien — Pelajari tentang teknik untuk mengalokasikan sumber daya yang dihitung secara dinamis untuk memenuhi berbagai permintaan.
  - Alat otomatisasi — Dapatkan ikhtisar alat dan layanan yang mengotomatiskan penskalaan komputasi, mengurangi kebutuhan akan intervensi manual.
- Penskalaan beban kerja membantu memastikan bahwa aplikasi dapat menangani berbagai beban kerja tanpa penurunan kinerja:
  - Horizontal pod autoscaler — Lihat secara mendalam bagaimana HPA membantu dalam penskalaan beban kerja berdasarkan metrik real-time.
  - Cluster Proportional Autoscaler — Pelajari bagaimana CPA secara otomatis menskalakan dan mempertahankan hubungan proporsional antara node dan replika, meningkatkan atau menurunkan beban kerja saat ukuran klaster berubah.
  - Event-driven scaling — Meninjau strategi untuk aplikasi penskalaan dalam menanggapi peristiwa atau pemicu tertentu.

- Penskalaan jaringan membantu menjaga komunikasi yang mulus antara layanan dan aliran data yang efisien di lingkungan yang dinamis:
  - Plugin Amazon VPC CNI - Pelajari bagaimana plugin VPC CNI memungkinkan jaringan yang dapat diskalakan dalam kluster Amazon EKS.
  - Jaringan khusus - Tinjau manajemen alamat IP dan segregasi lalu lintas jaringan di kluster Amazon EKS.
  - Delegasi awalan - Dapatkan ikhtisar tentang merampingkan manajemen IP di kluster Amazon EKS yang besar dan dapat diskalakan.
  - Amazon VPC Lattice — Dapatkan ikhtisar tentang bagaimana VPC Lattice dapat mengelola cross-VPC dan jaringan untuk penskalaan yang mulus. service-to-service
- Optimalisasi biaya membantu bisnis melihat di mana sumber daya mereka dibelanjakan dan menetapkan pengeluaran dengan tepat ke departemen atau proyek:
  - Sumber daya ukuran kanan — Pertimbangkan teknik untuk mengukur sumber daya cloud dengan tepat untuk beban kerja.
  - Pemantauan dan pengendalian biaya - Tinjau alat dan praktik terbaik untuk melacak dan mengoptimalkan biaya cloud.

Setiap bagian berfokus pada tujuan tertentu yang diperlukan untuk menciptakan lingkungan cloud yang andal, efektif, dan terjangkau.

Tujuan

Panduan ini dapat membantu Anda dan organisasi Anda mencapai tujuan bisnis berikut:

- Efisiensi sumber daya yang ditingkatkan — Mencapai pemanfaatan sumber daya yang optimal dengan menskalakan komputasi, beban kerja, dan sumber daya jaringan secara dinamis berdasarkan permintaan waktu nyata.

  Tujuan ini menekankan pentingnya penskalaan sumber daya naik dan turun sebagai respons terhadap pola penggunaan aktual. Alat seperti penskalaan otomatis pod horizontal dan plugin Amazon VPC CNI membantu organisasi hanya menggunakan sumber daya yang mereka butuhkan, meminimalkan pemborosan, dan memaksimalkan kinerja.

- Peningkatan kinerja aplikasi - Pertahankan kinerja dan daya tanggap aplikasi yang tinggi, bahkan di bawah beban kerja dan pola lalu lintas yang berfluktuasi.

Tujuan ini berfokus pada strategi untuk membantu memastikan bahwa aplikasi dapat menangani lalu lintas puncak dan beban kerja yang berat tanpa mengorbankan kinerja. Teknik seperti penskalaan beban kerja berbasis peristiwa, alokasi komputasi yang efisien, dan arsitektur jaringan yang dapat diskalakan adalah kunci untuk mencapai tujuan ini.

- Skalabilitas yang mulus — Memungkinkan penskalaan komponen infrastruktur yang lancar, memungkinkan pertumbuhan dan adaptasi yang mudah terhadap perubahan kebutuhan bisnis.

  Skalabilitas yang mulus sangat penting bagi organisasi yang mengantisipasi pertumbuhan atau mengalami berbagai tingkat lalu lintas. Tujuan ini membahas pentingnya menerapkan solusi yang dapat diskalakan di seluruh komputasi, beban kerja, dan sumber daya jaringan, sehingga penskalaan dapat otomatis, efisien, dan transparan.

- Optimalisasi biaya — Minimalkan biaya cloud sambil mempertahankan atau meningkatkan kinerja dan skalabilitas.

  Optimalisasi biaya dapat mencakup pengurangan biaya, seperti sumber daya ukuran yang tepat, menggunakan solusi penskalaan yang hemat biaya, dan memantau pengeluaran. Tujuannya adalah untuk menyeimbangkan penghematan biaya dengan kebutuhan akan kinerja dan skalabilitas tinggi.

Hitung penskalaan

Compute scaling adalah komponen penting untuk kinerja aplikasi dalam lingkungan Kubernetes yang dinamis. Kubernetes mengurangi pemborosan melalui penyesuaian dinamis sumber daya komputasi (seperti CPU dan memori) sebagai respons terhadap permintaan real-time. Kemampuan ini membantu menghindari penyediaan yang berlebihan atau kurang, yang juga dapat menghemat biaya operasional. Kubernetes secara efektif menghilangkan kebutuhan akan intervensi manual dengan memungkinkan infrastruktur meningkat secara otomatis selama jam sibuk dan turun selama periode off-peak.

Penskalaan komputasi Kubernetes secara keseluruhan mengotomatiskan proses penskalaan, yang meningkatkan fleksibilitas dan skalabilitas aplikasi serta meningkatkan perilaku toleran kesalahan. Pada akhirnya, kemampuan Kubernetes meningkatkan keunggulan operasional dan produktivitas.

Bagian ini membahas jenis penskalaan komputasi berikut:

- Cluster Autoscaler
- Cluster Autoscaler dengan penyediaan berlebih
- Karpenter

Cluster AutoScaler

Bergantung pada kebutuhan pod, alat Cluster Autoscaler secara otomatis memodifikasi ukuran dengan menambahkan node bila diperlukan atau menghapus node saat tidak diperlukan dan kurang dimanfaatkan.

Pertimbangkan alat Cluster Autoscaler sebagai solusi penskalaan untuk beban kerja di mana permintaan meningkat secara bertahap dan latensi dalam penskalaan bukanlah masalah utama.

Alat Cluster Autoscaler menyediakan fitur-fitur utama berikut:

- Penskalaan — Menskalakan node naik dan turun secara dinamis sebagai respons terhadap permintaan sumber daya yang sebenarnya.
- Penjadwalan Pod — Membantu memastikan bahwa setiap pod beroperasi dan memiliki sumber daya yang dibutuhkan untuk berfungsi, mencegah kelangkaan sumber daya.
- Efisiensi biaya — Menghilangkan biaya yang tidak perlu untuk mengoperasikan node yang kurang dimanfaatkan dengan menghilangkannya.

Cluster Autoscaler dengan penyediaan berlebih

Cluster Autoscaler dengan fungsi over-provisioning yang mirip dengan Cluster Autoscaler karena ia menyebarkan node secara efisien dan menghemat waktu dengan menjalankan pod prioritas rendah pada node. Dengan teknik ini, lalu lintas dialihkan ke pod ini sebagai respons terhadap lonjakan permintaan yang tiba-tiba, memungkinkan aplikasi untuk terus beroperasi tanpa gangguan.

Cluster Autoscaler dengan over-provisioning menawarkan fitur dummy pod yang dapat digunakan untuk menyebarkan dan menjalankan node dengan mudah saat beban kerja sangat besar, latensi tidak diperlukan, dan penskalaan harus cepat.

Cluster Autoscaler dengan over-provisioning menyediakan fitur utama berikut:

- Responsif yang lebih baik — Dengan membuat kelebihan kapasitas dapat diakses secara konstan, dibutuhkan lebih sedikit waktu untuk meningkatkan klaster sebagai respons terhadap lonjakan permintaan.
- Reservasi sumber daya — Mengelola lonjakan lalu lintas yang tidak terduga secara efektif membantu manajemen yang benar dengan sedikit waktu henti.
- Penskalaan yang lancar — Meminimalkan penundaan alokasi sumber daya memfasilitasi proses penskalaan yang lebih mulus.

Karpenter

Karpenter for Kubernetes mengungguli alat Cluster Autoscaler tradisional dalam hal open source, kinerja, dan kemampuan penyesuaian. Dengan Karpenter, Anda dapat secara otomatis meluncurkan hanya sumber daya komputasi yang diperlukan untuk menangani permintaan cluster Anda secara real time. Karpenter dirancang untuk memberikan penskalaan yang lebih efisien dan responsif.

Aplikasi dengan beban kerja yang sangat bervariasi atau kompleks, di mana keputusan penskalaan cepat sangat penting, mendapat manfaat besar dari penggunaan Karpenter. Ini terintegrasi dengan AWS, menawarkan penerapan yang lebih baik dan optimasi pemilihan simpul.

Karpenter mencakup fitur-fitur utama berikut:

- Penyediaan dinamis — Karpenter menyediakan instance dan ukuran yang tepat untuk tujuan tersebut dan menyediakan node baru secara dinamis berdasarkan persyaratan tertentu dari pod.

- Penjadwalan lanjutan — Menggunakan penempatan pod yang cerdas, Karpenter mengatur node sedemikian rupa sehingga sumber daya seperti GPU, CPU, memori, dan penyimpanan digunakan seefektif mungkin.
- Penskalaan cepat — Karpenter dapat menskalakan dengan cepat, sering bereaksi dalam hitungan detik. Responsif ini sangat membantu untuk pola lalu lintas mendadak atau ketika beban kerja menuntut penskalaan segera
- Efisiensi biaya — Dengan memilih instans yang paling efektif dengan cermat, Anda dapat menurunkan biaya pengoperasian dan memanfaatkan alternatif penghematan biaya tambahan yang ditawarkan oleh AWS, seperti Instans Sesuai Permintaan, Instans Spot, dan Instans Cadangan.

Penskalaan beban kerja

Penskalaan beban kerja di Kubernetes sangat penting untuk menjaga kinerja aplikasi dan efisiensi sumber daya di lingkungan yang dinamis. Penskalaan membantu memastikan bahwa aplikasi dapat menangani berbagai beban kerja tanpa penurunan kinerja. Kubernetes menyediakan kemampuan untuk secara otomatis menskalakan sumber daya naik atau turun berdasarkan metrik real-time, yang memungkinkan organisasi merespons perubahan lalu lintas dengan cepat. Elastisitas ini tidak hanya meningkatkan pengalaman pengguna tetapi juga mengoptimalkan pemanfaatan sumber daya, membantu meminimalkan biaya yang terkait dengan sumber daya yang kurang digunakan atau terlalu banyak.

Selain itu, penskalaan beban kerja yang efektif mendukung ketersediaan tinggi, memastikan bahwa aplikasi tetap responsif bahkan selama periode permintaan puncak. Penskalaan beban kerja di Kubernetes memungkinkan organisasi memanfaatkan sumber daya cloud dengan lebih baik dengan menyesuaikan kapasitas secara dinamis untuk memenuhi kebutuhan saat ini.

Bagian ini membahas jenis penskalaan beban kerja berikut:

- Penskaler Otomatis Pod Horisontal
- Cluster Proporsional Autoscaler
- Event Driven Autoscaler berbasis Kubernetes

Penskala Otomatis Pod Horizontal

Horizontal Pod Autoscaler (HPA) adalah fitur Kubernetes yang secara otomatis menyesuaikan jumlah replika pod dalam deployment, replication controller, atau stateful set, berdasarkan pemanfaatan CPU yang diamati atau metrik tertentu lainnya. HPA memastikan bahwa aplikasi dapat mengelola lalu lintas dan tingkat beban kerja yang berfluktuasi tanpa perlu intervensi manual. HPA menawarkan sarana untuk menjaga kinerja optimal sambil memanfaatkan sumber daya yang tersedia secara efektif.

Dalam konteks di mana permintaan pengguna mungkin berfluktuasi jauh dari waktu ke waktu, aplikasi web seperti, layanan mikro, dan APIs, HPA sangat membantu.

Horizontal Pod Autoscaler menyediakan fitur-fitur utama berikut:

- Penskalaan otomatis — HPA secara otomatis menambah atau mengurangi jumlah replika pod sebagai respons terhadap metrik waktu nyata, memastikan bahwa aplikasi dapat menskalakan untuk memenuhi permintaan pengguna.
- Keputusan berbasis metrik — Secara default, skala HPA berdasarkan pemanfaatan CPU. Namun, itu juga dapat menggunakan metrik khusus, seperti penggunaan memori atau metrik khusus aplikasi, memungkinkan strategi penskalaan yang lebih disesuaikan.
- Parameter yang dapat dikonfigurasi — Anda dapat memilih jumlah replika minimum dan maksimum serta persentase pemanfaatan yang diinginkan, memberi Anda wewenang atas seberapa parah penskalaan seharusnya.
- Integrasi dengan Kubernetes — Untuk memantau dan memodifikasi sumber daya, HPA bekerja bersama-sama dengan elemen ekosistem Kubernetes lainnya, termasuk Metrics Server, Kubernetes API, dan adaptor metrik kustom.
- Pemanfaatan sumber daya yang lebih baik — HPA membantu memastikan bahwa sumber daya digunakan secara efektif, menurunkan biaya dan meningkatkan kinerja, dengan memodifikasi jumlah pod secara dinamis.

Cluster Proporsional Autoscaler

Cluster Proportional Autoscaler (CPA) adalah komponen Kubernetes yang dirancang untuk secara otomatis menyesuaikan jumlah replika pod dalam sebuah cluster berdasarkan jumlah node yang tersedia. Tidak seperti autoscaler tradisional yang menskalakan berdasarkan metrik pemanfaatan sumber daya (seperti CPU dan memori), CPA menskalakan beban kerja secara proporsional dengan ukuran cluster itu sendiri.

Pendekatan ini sangat berguna untuk aplikasi yang perlu mempertahankan tingkat redundansi atau ketersediaan tertentu relatif terhadap ukuran cluster, seperti CoreDNS dan layanan infrastruktur lainnya. Beberapa kasus penggunaan utama untuk CPA meliputi:

- Penyediaan berlebihan
- Skala layanan platform inti
- Skalakan beban kerja karena CPA tidak memerlukan server metrik atau Adaptor Prometheus

Dengan mengotomatiskan proses penskalaan, CPA membantu bisnis dalam mempertahankan distribusi beban kerja yang seimbang, meningkatkan efisiensi sumber daya, dan memastikan bahwa aplikasi disediakan dengan tepat untuk memenuhi permintaan pengguna.

Cluster Proportional Autoscaler menyediakan fitur-fitur utama berikut:

- Penskalaan berbasis node — CPA menskalakan replika sesuai dengan jumlah node cluster yang dapat dijadwalkan, memungkinkan aplikasi untuk memperluas atau berkontraksi secara proporsional dengan ukuran cluster.
- Penyesuaian proporsional — Untuk memastikan bahwa aplikasi dapat menskalakan sesuai dengan perubahan ukuran cluster, autoscaler menetapkan hubungan proporsional antara jumlah node dan jumlah replika. Hubungan ini digunakan untuk menghitung jumlah replika yang diinginkan untuk beban kerja.
- Integrasi dengan komponen Kubernetes — CPA bekerja dengan komponen Kubernetes standar seperti Horizontal Pod Autoscaler (HPA) tetapi berfokus secara khusus pada jumlah node daripada metrik pemanfaatan sumber daya. Integrasi ini memungkinkan strategi penskalaan yang lebih komprehensif.
- Klien API Golang — Untuk memantau jumlah node dan core yang tersedia, CPA menggunakan klien API Golang yang berjalan di dalam pod dan berbicara dengan server API Kubernetes.
- Parameter yang dapat dikonfigurasi — Dengan menggunakan a`ConfigMap`, pengguna dapat mengatur ambang batas dan parameter penskalaan yang digunakan CPA untuk memodifikasi perilakunya dan memastikannya mengikuti rencana penskalaan yang dimaksud.

Autoscaler Berbasis Event berbasis Kubernetes

Event Driven Autoscaler (KEDA) berbasis Kubernetes adalah proyek sumber terbuka yang memungkinkan beban kerja Kubernetes untuk diskalakan berdasarkan jumlah peristiwa yang perlu diproses. KEDA meningkatkan skalabilitas aplikasi dengan memungkinkan mereka merespons secara dinamis terhadap berbagai beban kerja, terutama yang didorong oleh peristiwa.

Dengan mengotomatiskan proses penskalaan berdasarkan peristiwa, KEDA membantu organisasi mengoptimalkan pemanfaatan sumber daya, meningkatkan kinerja aplikasi, dan mengurangi biaya yang terkait dengan penyediaan berlebih. Pendekatan ini sangat berharga untuk aplikasi yang mengalami berbagai pola lalu lintas, seperti layanan mikro, fungsi tanpa server, dan sistem pemrosesan data real-time.

KEDA menyediakan fitur-fitur utama berikut:

- Penskalaan berbasis peristiwa — KEDA memungkinkan Anda menentukan aturan penskalaan berdasarkan sumber peristiwa eksternal, seperti antrian pesan, permintaan HTTP, atau metrik

khusus. Kemampuan ini membantu memastikan bahwa skala aplikasi dalam menanggapi permintaan waktu nyata.

- Komponen ringan — KEDA adalah komponen ringan dengan tujuan tunggal yang tidak memerlukan banyak pengaturan atau overhead untuk mudah diintegrasikan ke dalam klaster Kubernetes yang ada.
- Integrasi dengan Kubernetes — KEDA memperluas kemampuan komponen asli Kubernetes, seperti Horizontal Pod Autoscaler (HPA). KEDA menambahkan kemampuan penskalaan berbasis peristiwa ke komponen ini, meningkatkan daripada menggantinya.
- Support untuk berbagai sumber acara — KEDA kompatibel dengan berbagai sumber acara, termasuk platform perpesanan populer seperti RabbitMQ, Apache Kafka, dan lainnya. Karena kemampuan beradaptasi ini, Anda dapat menyesuaikan penskalaan agar sesuai dengan arsitektur berbasis peristiwa unik Anda.
- Scaler kustom — Dengan menggunakan scaler khusus, Anda dapat menetapkan metrik spesifik yang dapat digunakan KEDA untuk memulai tindakan penskalaan sebagai respons terhadap logika atau persyaratan bisnis tertentu.
- Konfigurasi deklaratif — Sejalan dengan prinsip Kubernetes, Anda dapat menggunakan KEDA untuk mendeskripsikan perilaku penskalaan secara deklaratif dengan menggunakan sumber daya kustom Kubernetes untuk menentukan bagaimana penskalaan seharusnya terjadi.

Penskalaan jaringan

Penskalaan jaringan di Kubernetes sangat penting untuk menjaga komunikasi yang mulus antar layanan dan mendukung aliran data yang efisien di lingkungan yang dinamis. Penskalaan infrastruktur jaringan membantu memastikan bahwa cluster dapat menangani berbagai tingkat lalu lintas tanpa mengalami kemacetan atau masalah latensi. Kubernetes menyediakan alat dan mekanisme untuk menskalakan sumber daya jaringan, memungkinkan organisasi untuk mempertahankan kinerja optimal seiring perubahan pola lalu lintas.

Elastisitas dalam penskalaan jaringan ini meningkatkan pengalaman pengguna secara keseluruhan dengan memastikan koneksi yang cepat dan andal. Penskalaan jaringan juga mengoptimalkan penggunaan sumber daya jaringan, membantu mengurangi biaya yang terkait dengan komponen jaringan yang kurang dimanfaatkan atau dibebani.

Selain itu, penskalaan jaringan yang efektif sangat penting untuk mendukung ketersediaan dan ketahanan yang tinggi. Dengan menyesuaikan kapasitas jaringan dan routing secara dinamis, organisasi dapat memastikan bahwa layanan tetap dapat diakses dan responsif bahkan selama periode permintaan puncak atau lonjakan lalu lintas yang tidak terduga. Pendekatan ini memungkinkan pemanfaatan sumber daya jaringan cloud yang lebih baik, memastikan bahwa infrastruktur selalu selaras dengan persyaratan saat ini.

Bagian ini membahas jenis-jenis penskalaan jaringan berikut:

- Plugin Amazon VPC CNI untuk Kubernetes
- Jaringan kustom
- Delegasi awalan
- Kisi VPC Amazon

Plugin Amazon VPC CNI untuk Kubernetes

Plugin Amazon VPC Container Network Interface (CNI) untuk Kubernetes adalah komponen penting di Amazon EKS. Plugin VPC CNI menyediakan kemampuan jaringan tingkat lanjut dengan mengintegrasikan pod Kubernetes dengan Amazon VPC. Dengan plugin ini, setiap pod diberi alamat IP unik dari virtual private cloud (VPC), sehingga meningkatkan isolasi dan kinerja jaringan. Seiring pertumbuhan cluster dan permintaan jaringan berfluktuasi, plugin Amazon VPC CNI memainkan peran kunci dalam memastikan operasi jaringan yang efisien dan terukur.

Plugin secara otomatis mengelola alokasi dan perutean alamat IP dalam VPC, menyederhanakan manajemen jaringan dan mengurangi risiko konflik IP. Ini mendukung fitur seperti delegasi awalan, yang memungkinkan manajemen IP yang lebih fleksibel.

Plugin VPC CNI membantu organisasi mengoptimalkan kinerja jaringan, meningkatkan keamanan, dan mengurangi risiko kelelahan IP. Kemampuan ini sangat berharga untuk lingkungan dinamis skala besar di mana tuntutan jaringan berfluktuasi, seperti arsitektur layanan mikro, beban kerja kepadatan tinggi, dan aplikasi multi-penyewa.

Plugin Amazon VPC CNI menyediakan fitur-fitur utama berikut:

- Jaringan yang disempurnakan - Plugin VPC CNI memungkinkan setiap pod menerima alamat IP sendiri langsung dari VPC, memberikan isolasi yang kuat dan kinerja jaringan. Pendekatan ini sangat penting untuk beban kerja yang membutuhkan throughput jaringan tinggi dan latensi rendah.
- Delegasi awalan — Untuk mengatasi masalah kehabisan alamat IP dalam klaster besar, delegasi awalan secara dinamis mengalokasikan blok yang lebih besar ke node, yang kemudian dibagi lagi IPs untuk penggunaan pod. Pendekatan ini memastikan pemanfaatan IP yang efisien dan menyederhanakan penskalaan jaringan.
- Jaringan khusus - Pengguna dapat mengonfigurasi antarmuka jaringan kustom (ENIs) untuk pod, yang membantu mendistribusikan lalu lintas pod di beberapa antarmuka, mengurangi kemacetan jaringan dan meningkatkan skalabilitas.
- Support for IPv6 — IPv6 Dengan mengaktifkan klaster Amazon EKS, pengguna dapat secara signifikan memperluas ruang alamat IP yang tersedia, memfasilitasi penskalaan aplikasi besar dan terdistribusi tanpa batasan batasan. IPv4
- Integrasi dengan Kubernetes - Plugin VPC CNI bekerja dengan mulus dengan komponen jaringan Kubernetes, memastikan bahwa IPs plugin tersebut dikelola secara efisien di seluruh pod, layanan, dan endpoint eksternal, dan mendukung fitur-fitur canggih seperti grup keamanan untuk pod.

Jaringan kustom

Jaringan khusus di Amazon EKS memungkinkan penetapan antarmuka jaringan tertentu ke pod, memberikan kontrol yang ditingkatkan atas manajemen alamat IP dan lalu lintas jaringan. Pendekatan ini sangat berguna dalam skenario di mana kelelahan alamat IP menjadi perhatian atau ketika ada kebutuhan untuk memisahkan lalu lintas jaringan untuk alasan keamanan, kepatuhan,

atau kinerja. Jaringan khusus membantu organisasi mengelola ruang alamat IP secara efisien, memisahkan lalu lintas, dan memastikan kinerja jaringan yang dapat diskalakan.

Dengan jaringan khusus, administrator dapat mengelola sumber daya jaringan dengan lebih efisien. Admin dapat menggunakan jaringan khusus untuk membantu memastikan bahwa pod memiliki isolasi jaringan yang diperlukan dan klaster dapat menskalakan tanpa menghadapi batasan alamat IP.

Jaringan khusus menyediakan fitur-fitur utama berikut:

- Manajemen IP yang disempurnakan - Jaringan khusus memungkinkan penetapan antarmuka jaringan tertentu (ENIs) ke pod, membantu mengelola kelelahan alamat IP dengan mendistribusikan lalu lintas pod ke beberapa. ENIs Kemampuan ini sangat penting dalam kelompok dengan beban kerja kepadatan tinggi.
- Pemisahan lalu lintas - Dengan antarmuka jaringan khusus, Anda dapat memisahkan lalu lintas pod berdasarkan kriteria tertentu, seperti jenis aplikasi atau persyaratan keamanan. Pendekatan ini memberikan kontrol yang lebih besar atas bagaimana lalu lintas mengalir di dalam dan di luar cluster.
- Support for IPv6 - Jaringan khusus di Amazon EKS juga mendukung IPv6, menawarkan solusi untuk keterbatasan IPv4 alamat. Jaringan dapat menskalakan secara efisien tanpa konflik alamat IP, bahkan dalam penyebaran skala besar.
- Skalabilitas dan fleksibilitas — Sebagai skala cluster, jaringan kustom memungkinkan manajemen dinamis antarmuka jaringan. Pod baru diberi sumber daya jaringan yang sesuai tanpa intervensi manual. Pendekatan ini membantu menjaga lingkungan jaringan yang fleksibel dan terukur yang dapat beradaptasi dengan perubahan beban kerja.

Delegasi awalan

Delegasi awalan di Kubernetes, khususnya di Amazon EKS, dirancang untuk merampingkan dan mengoptimalkan manajemen alamat IP saat skala cluster. Dengan secara dinamis mengalokasikan blok alamat IP (awalan) yang lebih besar ke node, delegasi awalan mengurangi risiko kelelahan IP dan menyederhanakan pengelolaan ruang IP.

Pendekatan ini meningkatkan efisiensi jaringan, meminimalkan fragmentasi, dan membantu klaster menskalakan dengan lancar tanpa penyesuaian rentang IP manual. Delegasi awalan sangat berharga untuk penerapan skala besar, beban kerja kepadatan tinggi, dan lingkungan di mana manajemen IP dinamis yang fleksibel sangat penting untuk menjaga kinerja dan skalabilitas jaringan.

Delegasi awalan menyediakan fitur utama berikut:

- Manajemen alamat IP yang efisien - Delegasi awalan memungkinkan alokasi dinamis rentang IP, mengurangi risiko kelelahan IP dan memastikan penggunaan ruang IP yang tersedia secara efisien.
- Manajemen jaringan yang disederhanakan — Dengan memungkinkan node untuk menangani alokasi IP mereka sendiri, delegasi awalan meminimalkan fragmentasi jaringan dan menyederhanakan proses routing, membuatnya lebih mudah untuk skala cluster sesuai kebutuhan.
- Support untuk penerapan skala besar — Dalam cluster besar dengan beban kerja kepadatan tinggi, delegasi awalan memungkinkan penskalaan tanpa batas dengan memungkinkan node baru untuk bergabung dengan cluster tanpa penyesuaian rentang IP manual.

Kisi VPC Amazon

Amazon VPC Lattice memungkinkan service-to-service komunikasi yang efisien dan aman di dalam dan di seluruh VPCs, terutama dalam arsitektur layanan mikro. VPC Lattice menggunakan langkah-langkah keamanan seperti grup keamanan dan daftar kontrol akses jaringan (jaringan ACLs) selain integrasi AWS Identity and Access Management (IAM) untuk otentikasi aplikasi berbutir halus. Layanan proxy layer-7 di jantung VPC Lattice menawarkan koneksi, load balancing, otentikasi, otorisasi, observabilitas, manajemen lalu lintas, dan penemuan layanan.

Dengan menyederhanakan konfigurasi jaringan dan keamanan, VPC Lattice membantu organisasi mengoptimalkan manajemen lalu lintas, meningkatkan kinerja aplikasi, dan menskalakan secara mulus di beberapa dan. VPCs Wilayah AWS Ini sangat berharga untuk aplikasi terdistribusi yang membutuhkan jaringan yang konsisten dan andal, seperti layanan mikro, penyebaran lintas wilayah, dan lingkungan cloud-native yang kompleks.

Amazon VPC Lattice menyediakan fitur-fitur utama berikut:

- Service-to-service jaringan — VPC Lattice menyederhanakan jaringan dan konfigurasi keamanan antara layanan dalam arsitektur microservices. Ini menyediakan platform terpadu untuk mengelola komunikasi, sehingga layanan dapat skala secara independen sambil mempertahankan kinerja dan keamanan yang tinggi.
- Jaringan lintas-VPC — VPC Lattice sangat penting untuk mengelola lalu lintas di beberapa atau Wilayah. VPCs Ini menyediakan kerangka jaringan yang konsisten yang memungkinkan layanan untuk berkomunikasi dengan mulus, terlepas dari lokasi fisik mereka. Kemampuan ini sangat penting untuk aplikasi skala besar yang menjangkau beberapa Wilayah VPCs atau geografis.

- Manajemen keamanan yang ditingkatkan — Dengan mengintegrasikan kebijakan keamanan langsung ke lapisan jaringan, VPC Lattice service-to-service mendukung komunikasi yang aman dan efisien. Fitur ini mengurangi kompleksitas pengelolaan keamanan di seluruh lingkungan terdistribusi, memungkinkan penskalaan yang lebih mudah dan mengurangi overhead operasional.
- Manajemen lalu lintas yang disederhanakan - VPC Lattice menawarkan fitur manajemen lalu lintas canggih, termasuk perutean, penyeimbangan beban, dan mekanisme failover. Dengan fitur-fitur ini, lalu lintas didistribusikan secara efisien di seluruh layanan, mengoptimalkan kinerja jaringan dan meningkatkan skalabilitas aplikasi.

Optimalisasi biaya

Untuk mendukung pengendalian sumber daya yang efektif, minimalisasi biaya Kubernetes sangat penting bagi perusahaan yang menggunakan teknologi orkestrasi kontainer ini. Sulit untuk melacak pengeluaran dengan benar di pengaturan Kubernetes karena kompleksitasnya, yang mencakup beberapa komponen seperti pod dan node. Melalui penerapan teknik pengoptimalan biaya, bisnis dapat melihat di mana sumber daya mereka dihabiskan dan secara tepat menetapkan pengeluaran ke departemen atau proyek.

Meskipun penskalaan dinamis memiliki kelebihan, jika tidak dikelola dengan baik, hal itu dapat mengakibatkan pengeluaran yang tidak terduga. Manajemen biaya yang efisien membantu mengalokasikan sumber daya hanya ketika mereka benar-benar dibutuhkan, mencegah lonjakan pengeluaran yang tidak terduga.

Bagian ini membahas pendekatan berikut untuk optimalisasi biaya:

- Kubecost
- Goldilocks
- AWS Fargate
- Instans Spot
- Instans Cadangan
- AWS Contoh graviton

Kubecost

Kubecost adalah solusi manajemen biaya yang membantu bisnis melacak, mengontrol, dan memaksimalkan pengeluaran mereka pada infrastruktur cloud. Ini dibuat khusus untuk cluster Kubernetes. Kubecost memberi Anda wawasan tentang pemanfaatan sumber daya dan kesadaran biaya waktu nyata, memungkinkan Anda untuk lebih memahami di mana dan berapa banyak sumber daya cloud Anda digunakan. Dengan wawasan ini, Anda dapat mengoptimalkan pengeluaran infrastruktur, meningkatkan efisiensi sumber daya, dan membuat keputusan yang lebih tepat tentang investasi cloud Anda.

Kubecost menyediakan fitur-fitur utama berikut:

- Alokasi biaya — Kubecost menawarkan alokasi biaya menyeluruh untuk sumber daya Kubernetes, termasuk beban kerja, layanan, ruang nama, dan label. Fitur ini membantu tim memantau biaya berdasarkan lingkungan, proyek, atau tim.
- Pemantauan biaya waktu nyata — Ini menawarkan pemantauan biaya cloud secara real-time, memberikan wawasan langsung kepada organisasi tentang pola pengeluaran dan membantu mencegah pembengkakan biaya yang tidak terduga.
- Rekomendasi optimasi — Kubecost menawarkan saran praktis untuk meminimalkan pemanfaatan sumber daya, termasuk mengurangi sumber daya yang tidak digunakan, mengukur beban kerja yang tepat, dan memaksimalkan biaya penyimpanan.
- Penganggaran dan peringatan — Pengguna Kubecost dapat membuat anggaran dan menerima pengingat ketika pengeluaran mendekati atau melampaui kriteria yang telah ditentukan sebelumnya. Fitur ini membantu tim mematuhi kendala keuangan.

Goldilocks

Goldilocks adalah utilitas Kubernetes yang dirancang untuk membantu pengguna mengoptimalkan permintaan sumber daya dan batasan untuk beban kerja Kubernetes. Ini memberikan rekomendasi tentang cara mengkonfigurasi sumber daya CPU dan memori untuk kontainer yang berjalan di cluster Kubernetes. Rekomendasi ini membantu Anda memastikan bahwa aplikasi memiliki jumlah sumber daya yang tepat untuk bekerja secara efisien tanpa pemborosan. Optimalisasi ini dapat menghasilkan penghematan biaya, peningkatan kinerja, dan penggunaan klaster Kubernetes yang lebih efisien.

Goldilocks menyediakan fitur utama berikut:

- Rekomendasi sumber daya — Goldilocks menentukan pengaturan ideal untuk permintaan dan pembatasan sumber daya dengan menganalisis statistik konsumsi CPU dan memori sebelumnya untuk beban kerja Kubernetes. Dengan melakukan ini, menjadi lebih mudah untuk menghindari penyediaan yang kurang atau berlebihan, yang dapat mengakibatkan masalah kinerja dan pemborosan sumber daya.
- Integrasi VPA — Goldilocks memanfaatkan Kubernetes Vertical Pod Autoscaler (VPA) untuk mengumpulkan data dan memberikan rekomendasi. Ini berjalan dalam “mode rekomendasi,” yang berarti itu tidak benar-benar mengubah pengaturan sumber daya tetapi menawarkan panduan tentang pengaturan apa yang seharusnya.
- Analisis berbasis Namespace — Goldilocks memberi Anda kemampuan untuk mengatur dengan baik beban kerja mana yang dioptimalkan dan dipantau dengan memungkinkan Anda menargetkan ruang nama tertentu untuk dianalisis.

- Dasbor visual — Dasbor berbasis web menampilkan permintaan dan batasan sumber daya yang disarankan secara visual, yang membuatnya mudah bagi Anda untuk memahami dan mengambil tindakan terhadap data.
- Operasi non-intrusif — Goldilocks tidak mengubah pengaturan cluster karena beroperasi dalam mode rekomendasi. Jika mau, Anda dapat secara manual menerapkan pengaturan sumber daya yang disarankan setelah meninjau rekomendasi.

AWS Fargate

Dalam konteks Amazon EKS, Anda dapat https://docs.aws.amazon.com/eks/latest/userguide/fargate.htmlAWS Fargate menjalankan pod Kubernetes tanpa mengelola instans Amazon yang mendasarinya. EC2 Ini adalah mesin komputasi tanpa server yang memungkinkan Anda fokus pada penerapan dan penskalaan aplikasi kontainer tanpa mengkhawatirkan infrastruktur.

AWS Fargate menyediakan fitur utama berikut:

- Tidak ada manajemen infrastruktur — Fargate menghilangkan kebutuhan untuk menyediakan, mengelola, atau menskalakan EC2 instans Amazon atau node Kubernetes. AWS menangani semua manajemen infrastruktur, termasuk patching dan scaling.
- Isolasi tingkat POD — Tidak seperti node pekerja yang berbasis di Amazon, EC2 Fargate menyediakan isolasi tingkat tugas atau pod. Setiap pod berjalan di lingkungan komputasi terisolasinya sendiri, yang meningkatkan keamanan dan kinerja.
- Penskalaan otomatis — Fargate secara otomatis menskalakan pod Kubernetes berdasarkan permintaan. Anda tidak perlu mengelola kebijakan penskalaan atau kumpulan node.
- Penagihan per detik - Anda hanya membayar vCPU dan sumber daya memori yang dikonsumsi oleh setiap pod untuk durasi yang tepat yang dijalankan, yang merupakan opsi hemat biaya untuk beban kerja tertentu.
- Mengurangi overhead — Dengan menghilangkan kebutuhan untuk mengelola EC2 instans, Fargate memungkinkan Anda untuk fokus membangun dan mengelola aplikasi Anda daripada operasi infrastruktur.

Instans Spot

Instans Spot menawarkan penghematan yang signifikan atas harga Instans Sesuai Permintaan dan merupakan opsi yang terjangkau untuk menjalankan node EC2 pekerja Amazon di klaster Amazon

EKS. Namun, AWS dapat mengganggu Instans Spot jika kapasitas Instans Sesuai Permintaan diperlukan. AWS dapat merebut kembali Instans Spot dengan pemberitahuan 2 menit saat kapasitas diperlukan, membuatnya kurang dapat diandalkan untuk beban kerja yang kritis dan stateful.

Untuk beban kerja yang sensitif terhadap biaya dan dapat menahan gangguan, Instans Spot di Amazon EKS adalah pilihan yang baik. Menggunakan kombinasi Instans Spot dan Instans Sesuai Permintaan di klaster Kubernetes membantu Anda menghemat uang tanpa mengorbankan ketersediaan untuk beban kerja penting.

Instans Spot menyediakan fitur utama berikut:

- Penghematan biaya — Instans Spot bisa lebih murah daripada harga Instans Sesuai Permintaan, menjadikannya ideal untuk beban kerja yang sensitif terhadap biaya.
- Ideal untuk beban kerja yang toleran terhadap kesalahan - Sangat cocok untuk beban kerja tanpa kewarganegaraan, toleran kesalahan seperti pemrosesan batch, pekerjaan CI/CD, pembelajaran mesin, atau pemrosesan data skala besar di mana instance dapat diganti tanpa gangguan besar.
- Integrasi grup Penskalaan Otomatis — Amazon EKS mengintegrasikan Instans Spot dengan Kubernetes Cluster Autoscaler, yang dapat secara otomatis mengganti node Instance Spot yang terputus dengan Instans Spot atau Instans Sesuai Permintaan lainnya yang tersedia.

Instans Terpesan

Di Amazon EKS, Instans Cadangan adalah model harga untuk node EC2 pekerja Amazon yang menjalankan beban kerja Kubernetes Anda. Dengan menggunakan Instans Cadangan, Anda berkomitmen untuk menggunakan jenis instans tertentu untuk jangka waktu 1 tahun atau 3 tahun, dengan imbalan penghematan biaya dibandingkan dengan harga Instans Sesuai Permintaan. Pemesanan instans di Amazon EKS adalah cara yang terjangkau untuk melakukan beban kerja jangka panjang yang konsisten di node pekerja Amazon EC2 .

Instans Cadangan biasanya digunakan untuk Amazon EC2. Namun, node pekerja di kluster Amazon EKS Anda (yang merupakan EC2 instance) juga dapat memanfaatkan model penghematan biaya ini, asalkan beban kerja memerlukan penggunaan jangka panjang yang dapat diprediksi.

Layanan produksi, database, dan aplikasi stateful lainnya yang membutuhkan ketersediaan tinggi dan kinerja yang konsisten adalah contoh beban kerja stabil yang cocok untuk Instans Cadangan.

Instans Cadangan menyediakan fitur utama berikut:

- Penghematan biaya — Instans Cadangan menawarkan penghematan dibandingkan dengan instans Sesuai Permintaan, tergantung pada jangka waktu (1 atau 3 tahun) dan rencana pembayaran (Semua di Muka, Sebagian di Muka, atau Tidak Ada di Muka).
- Komitmen jangka panjang — Anda berkomitmen untuk jangka waktu 1 tahun atau 3 tahun untuk jenis, ukuran, dan contoh tertentu. Wilayah AWS Ini sangat ideal untuk beban kerja yang stabil dan berjalan terus menerus dari waktu ke waktu.
- Harga yang dapat diprediksi — Karena Anda berkomitmen untuk jangka waktu tertentu, Instans Cadangan memberikan biaya bulanan atau dimuka yang dapat diprediksi, sehingga memudahkan anggaran untuk beban kerja jangka panjang.
- Fleksibilitas instans — Dengan Instans Cadangan Konvertibel, Anda dapat mengubah jenis instans, keluarga, atau ukuran selama periode reservasi. Instans Cadangan Konvertibel menawarkan lebih banyak fleksibilitas daripada Instans Cadangan Standar, yang tidak mengizinkan perubahan.
- Kapasitas terjamin — Instans Cadangan memastikan bahwa kapasitas tersedia di Availability Zone tempat reservasi dibuat, yang sangat penting untuk beban kerja penting yang membutuhkan daya komputasi yang konsisten.
- Tidak ada risiko gangguan — Tidak seperti Instans Spot, Instans Cadangan tidak mengalami gangguan oleh. AWS Ini membuatnya ideal untuk menjalankan beban kerja mission-critical yang membutuhkan waktu aktif yang terjamin.

AWS Contoh graviton

AWS Graviton adalah keluarga prosesor berbasis ARM yang dirancang AWS untuk memberikan peningkatan kinerja dan efisiensi biaya untuk beban kerja cloud. Dalam konteks Amazon EKS, Anda dapat menggunakan instance Graviton sebagai node pekerja untuk menjalankan beban kerja Kubernetes Anda, menawarkan peningkatan kinerja yang signifikan dan penghematan biaya.

Instans Graviton adalah pilihan yang sangat baik untuk aplikasi cloud-native dan komputasi intensif karena mereka menawarkan rasio harga-kinerja yang lebih tinggi daripada instans x86. Namun, ketika Anda mempertimbangkan untuk mengadopsi instance Graviton, pertimbangkan kompatibilitas ARM.

AWS Instans Graviton menyediakan fitur utama berikut:

- Arsitektur berbasis ARM - Prosesor AWS Graviton dibangun di atas arsitektur ARM, yang berbeda dari arsitektur x86 tradisional tetapi sangat efisien untuk banyak beban kerja.

- Hemat biaya - EC2 Instans Amazon berdasarkan Graviton biasanya menawarkan kinerja harga yang lebih baik dibandingkan dengan instans berbasis x86. EC2 Ini menjadikannya opsi yang menarik untuk cluster Kubernetes yang menjalankan Amazon EKS.
- Kinerja - Prosesor Graviton2, generasi kedua AWS Graviton, menawarkan peningkatan yang signifikan dalam hal kinerja komputasi, throughput memori, dan efisiensi energi. Mereka ideal untuk beban kerja intensif CPU dan intensif memori.
- Jenis instans yang beragam - Instans Graviton datang dalam berbagai keluarga, seperti t4g, m7g, c7g, dan r7g, yang mencakup berbagai kasus penggunaan mulai dari tujuan umum hingga beban kerja yang dioptimalkan untuk komputasi, dioptimalkan untuk memori, dan burstable.
- Grup node Amazon EKS — Anda dapat mengonfigurasi grup node yang dikelola oleh Amazon EKS atau grup node yang dikelola sendiri untuk menyertakan instance berbasis Graviton. Dengan pendekatan ini, Anda dapat menjalankan beban kerja yang dioptimalkan untuk arsitektur ARM pada cluster Kubernetes yang sama bersama instance berbasis x86.

Langkah selanjutnya

Panduan ini memberikan informasi untuk membantu Anda mengoptimalkan Amazon EKS sehubungan dengan penskalaan komputasi, penskalaan beban kerja, penskalaan jaringan, dan pengoptimalan biaya. Dengan memahami dan menerapkan konsep-konsep ini, organisasi dapat mencapai lingkungan cloud yang sangat efisien, terukur, dan hemat biaya yang memenuhi kebutuhan dinamis mereka.

Implementasi komputasi dan penskalaan beban kerja yang efektif membantu memastikan bahwa sumber daya digunakan secara efisien dan aplikasi mempertahankan kinerja tinggi bahkan selama waktu puncak. Merangkul teknik penskalaan jaringan, seperti jaringan khusus dan delegasi awalan, mendukung pengelolaan sumber daya jaringan dan skalabilitas yang mulus. Menekankan optimalisasi biaya membantu organisasi menyeimbangkan kinerja dengan efisiensi keuangan.

Mengintegrasikan panduan ini ke dalam strategi cloud Anda dapat membantu Anda meningkatkan kinerja dan skalabilitas infrastruktur Anda dan mendorong penghematan biaya. Pendekatan komprehensif ini memungkinkan Anda untuk membangun lingkungan cloud yang kuat yang mendukung pertumbuhan organisasi Anda dan beradaptasi dengan tuntutan bisnis yang selalu berubah.

Sumber daya

AWS blog

- Membangun untuk Optimalisasi Biaya dan Ketahanan untuk EKS dengan Instans Spot
- Mencampur AWS Graviton dengan x86 CPUs untuk mengoptimalkan biaya dan ketahanan menggunakan Amazon EKS

AWS dokumentasi

- Amazon VPC CNI
- Amazon Elastic Kubernetes Service AWS (whitepaper: Ikhtisar Opsi Deployment aktif) AWS
- Panduan Praktik Terbaik Amazon EKS
- Karpenter
- Pelajari lebih lanjut tentang Kubecost
- Sederhanakan manajemen komputasi dengan AWS Fargate

Sumber daya lainnya

- Penskalaan Otomatis Cluster (dokumentasi Kubernetes)
- Goldilocks: Alat Sumber Terbuka untuk Merekomendasikan Permintaan Sumber Daya (Blog Fairwinds)
- Penskalaan Otomatis Pod Horisontal (dokumentasi Kubernetes)
- Kubecost (dokumentasi Kubecost)
- Kubernetes Event-driven Autoscaling (dokumentasi KEDA)

Riwayat dokumen

Tabel berikut menjelaskan perubahan signifikan pada panduan ini, Menskalakan infrastruktur Amazon EKS untuk mengoptimalkan komputasi, beban kerja, dan kinerja jaringan. Jika Anda ingin diberi tahu tentang pembaruan masa depan, Anda dapat berlangganan umpan RSS.

Perubahan	Deskripsi	Tanggal
Publikasi awal	—	November 11, 2024

AWS Glosarium Panduan Preskriptif

Berikut ini adalah istilah yang umum digunakan dalam strategi, panduan, dan pola yang disediakan oleh Panduan AWS Preskriptif. Untuk menyarankan entri, silakan gunakan tautan Berikan umpan balik di akhir glosarium.

Nomor

7 Rs

Tujuh strategi migrasi umum untuk memindahkan aplikasi ke cloud. Strategi ini dibangun di atas 5 Rs yang diidentifikasi Gartner pada tahun 2011 dan terdiri dari yang berikut:

- Refactor/Re-Architect — Memindahkan aplikasi dan memodifikasi arsitekturnya dengan memanfaatkan sepenuhnya fitur cloud-native untuk meningkatkan kelincahan, kinerja, dan skalabilitas. Ini biasanya melibatkan porting sistem operasi dan database. Contoh: Migrasikan database Oracle lokal Anda ke Amazon Aurora PostgreSQL Compatible Edition.
- Replatform (angkat dan bentuk ulang) — Pindahkan aplikasi ke cloud, dan perkenalkan beberapa tingkat pengoptimalan untuk memanfaatkan kemampuan cloud. Contoh: Memigrasikan database Oracle lokal Anda ke Amazon Relational Database Service (Amazon RDS) untuk Oracle di. AWS Cloud
- Pembelian kembali (drop and shop) - Beralih ke produk yang berbeda, biasanya dengan beralih dari lisensi tradisional ke model SaaS. Contoh: Migrasikan sistem manajemen hubungan pelanggan (CRM) Anda ke Salesforce.com.
- Rehost (lift dan shift) — Pindahkan aplikasi ke cloud tanpa membuat perubahan apa pun untuk memanfaatkan kemampuan cloud. Contoh: Migrasikan database Oracle lokal Anda ke Oracle pada instance EC2 di. AWS Cloud
- Relokasi (hypervisor-level lift and shift) — Pindahkan infrastruktur ke cloud tanpa membeli perangkat keras baru, menulis ulang aplikasi, atau memodifikasi operasi yang ada. Anda memigrasikan server dari platform lokal ke layanan cloud untuk platform yang sama. Contoh: Migrasikan Microsoft Hyper-V aplikasi ke AWS.
- Pertahankan (kunjungi kembali) - Simpan aplikasi di lingkungan sumber Anda. Ini mungkin termasuk aplikasi yang memerlukan refactoring besar, dan Anda ingin menunda pekerjaan itu sampai nanti, dan aplikasi lama yang ingin Anda pertahankan, karena tidak ada pembenaran bisnis untuk memigrasikannya.

- Pensiun — Menonaktifkan atau menghapus aplikasi yang tidak lagi diperlukan di lingkungan sumber Anda.

A

ABAC

Lihat kontrol akses berbasis atribut.

layanan abstrak

Lihat layanan terkelola.

ASAM

Lihat atomisitas, konsistensi, isolasi, daya tahan.

migrasi aktif-aktif

Metode migrasi database di mana database sumber dan target tetap sinkron (dengan menggunakan alat replikasi dua arah atau operasi penulisan ganda), dan kedua database menangani transaksi dari menghubungkan aplikasi selama migrasi. Metode ini mendukung migrasi dalam batch kecil yang terkontrol alih-alih memerlukan pemotongan satu kali. Ini lebih fleksibel tetapi membutuhkan lebih banyak pekerjaan daripada migrasi aktif-pasif.

migrasi aktif-pasif

Metode migrasi database di mana database sumber dan target disimpan dalam sinkron, tetapi hanya database sumber yang menangani transaksi dari menghubungkan aplikasi sementara data direplikasi ke database target. Basis data target tidak menerima transaksi apa pun selama migrasi.

fungsi agregat

Fungsi SQL yang beroperasi pada sekelompok baris dan menghitung nilai pengembalian tunggal untuk grup. Contoh fungsi agregat meliputi `SUM` dan`MAX`.

AI

Lihat kecerdasan buatan.

AIOps

Lihat operasi kecerdasan buatan.

anonimisasi

Proses menghapus informasi pribadi secara permanen dalam kumpulan data. Anonimisasi dapat membantu melindungi privasi pribadi. Data anonim tidak lagi dianggap sebagai data pribadi.

anti-pola

Solusi yang sering digunakan untuk masalah berulang di mana solusinya kontra-produktif, tidak efektif, atau kurang efektif daripada alternatif.

kontrol aplikasi

Pendekatan keamanan yang memungkinkan penggunaan hanya aplikasi yang disetujui untuk membantu melindungi sistem dari malware.

portofolio aplikasi

Kumpulan informasi rinci tentang setiap aplikasi yang digunakan oleh organisasi, termasuk biaya untuk membangun dan memelihara aplikasi, dan nilai bisnisnya. Informasi ini adalah kunci untuk penemuan portofolio dan proses analisis dan membantu mengidentifikasi dan memprioritaskan aplikasi yang akan dimigrasi, dimodernisasi, dan dioptimalkan.

kecerdasan buatan (AI)

Bidang ilmu komputer yang didedikasikan untuk menggunakan teknologi komputasi untuk melakukan fungsi kognitif yang biasanya terkait dengan manusia, seperti belajar, memecahkan masalah, dan mengenali pola. Untuk informasi lebih lanjut, lihat Apa itu Kecerdasan Buatan?

operasi kecerdasan buatan (AIOps)

Proses menggunakan teknik pembelajaran mesin untuk memecahkan masalah operasional, mengurangi insiden operasional dan intervensi manusia, dan meningkatkan kualitas layanan. Untuk informasi selengkapnya tentang cara AIOps digunakan dalam strategi AWS migrasi, lihat panduan integrasi operasi.

enkripsi asimetris

Algoritma enkripsi yang menggunakan sepasang kunci, kunci publik untuk enkripsi dan kunci pribadi untuk dekripsi. Anda dapat berbagi kunci publik karena tidak digunakan untuk dekripsi, tetapi akses ke kunci pribadi harus sangat dibatasi.

atomisitas, konsistensi, isolasi, daya tahan (ACID)

Satu set properti perangkat lunak yang menjamin validitas data dan keandalan operasional database, bahkan dalam kasus kesalahan, kegagalan daya, atau masalah lainnya.

kontrol akses berbasis atribut (ABAC)

Praktik membuat izin berbutir halus berdasarkan atribut pengguna, seperti departemen, peran pekerjaan, dan nama tim. Untuk informasi selengkapnya, lihat ABAC untuk AWS dokumentasi AWS Identity and Access Management (IAM).

sumber data otoritatif

Lokasi di mana Anda menyimpan versi utama data, yang dianggap sebagai sumber informasi yang paling dapat diandalkan. Anda dapat menyalin data dari sumber data otoritatif ke lokasi lain untuk tujuan memproses atau memodifikasi data, seperti menganonimkan, menyunting, atau membuat nama samaran.

Zona Ketersediaan

Lokasi berbeda di dalam Wilayah AWS yang terisolasi dari kegagalan di Availability Zone lainnya dan menyediakan konektivitas jaringan latensi rendah yang murah ke Availability Zone lainnya di Wilayah yang sama.

AWS Kerangka Adopsi Cloud (AWS CAF)

Kerangka pedoman dan praktik terbaik AWS untuk membantu organisasi mengembangkan rencana yang efisien dan efektif untuk bergerak dengan sukses ke cloud. AWS CAF mengatur panduan ke dalam enam area fokus yang disebut perspektif: bisnis, orang, tata kelola, platform, keamanan, dan operasi. Perspektif bisnis, orang, dan tata kelola fokus pada keterampilan dan proses bisnis; perspektif platform, keamanan, dan operasi fokus pada keterampilan dan proses teknis. Misalnya, perspektif masyarakat menargetkan pemangku kepentingan yang menangani sumber daya manusia (SDM), fungsi kepegawaian, dan manajemen orang. Untuk perspektif ini, AWS CAF memberikan panduan untuk pengembangan, pelatihan, dan komunikasi orang untuk membantu mempersiapkan organisasi untuk adopsi cloud yang sukses. Untuk informasi lebih lanjut, lihat situs web AWS CAF dan whitepaper AWS CAF.

AWS Kerangka Kualifikasi Beban Kerja (AWS WQF)

Alat yang mengevaluasi beban kerja migrasi database, merekomendasikan strategi migrasi, dan memberikan perkiraan kerja. AWS WQF disertakan dengan AWS Schema Conversion Tool ()AWS SCT. Ini menganalisis skema database dan objek kode, kode aplikasi, dependensi, dan karakteristik kinerja, dan memberikan laporan penilaian.

B

bot buruk

Bot yang dimaksudkan untuk mengganggu atau membahayakan individu atau organisasi.

BCP

Lihat perencanaan kontinuitas bisnis.

grafik perilaku

Pandangan interaktif yang terpadu tentang perilaku dan interaksi sumber daya dari waktu ke waktu. Anda dapat menggunakan grafik perilaku dengan Amazon Detective untuk memeriksa upaya logon yang gagal, panggilan API yang mencurigakan, dan tindakan serupa. Untuk informasi selengkapnya, lihat Data dalam grafik perilaku di dokumentasi Detektif.

sistem big-endian

Sistem yang menyimpan byte paling signifikan terlebih dahulu. Lihat juga endianness.

klasifikasi biner

Sebuah proses yang memprediksi hasil biner (salah satu dari dua kelas yang mungkin). Misalnya, model ML Anda mungkin perlu memprediksi masalah seperti “Apakah email ini spam atau bukan spam?” atau “Apakah produk ini buku atau mobil?”

filter mekar

Struktur data probabilistik dan efisien memori yang digunakan untuk menguji apakah suatu elemen adalah anggota dari suatu himpunan.

deployment biru/hijau

Strategi penyebaran tempat Anda membuat dua lingkungan yang terpisah namun identik. Anda menjalankan versi aplikasi saat ini di satu lingkungan (biru) dan versi aplikasi baru di lingkungan lain (hijau). Strategi ini membantu Anda dengan cepat memutar kembali dengan dampak minimal.

bot

Aplikasi perangkat lunak yang menjalankan tugas otomatis melalui internet dan mensimulasikan aktivitas atau interaksi manusia. Beberapa bot berguna atau bermanfaat, seperti perayap web yang mengindeks informasi di internet. Beberapa bot lain, yang dikenal sebagai bot buruk, dimaksudkan untuk mengganggu atau membahayakan individu atau organisasi.

botnet

Jaringan bot yang terinfeksi oleh malware dan berada di bawah kendali satu pihak, yang dikenal sebagai bot herder atau operator bot. Botnet adalah mekanisme paling terkenal untuk skala bot dan dampaknya.

cabang

Area berisi repositori kode. Cabang pertama yang dibuat dalam repositori adalah cabang utama. Anda dapat membuat cabang baru dari cabang yang ada, dan Anda kemudian dapat mengembangkan fitur atau memperbaiki bug di cabang baru. Cabang yang Anda buat untuk membangun fitur biasanya disebut sebagai cabang fitur. Saat fitur siap dirilis, Anda menggabungkan cabang fitur kembali ke cabang utama. Untuk informasi selengkapnya, lihat Tentang cabang (GitHub dokumentasi).

akses break-glass

Dalam keadaan luar biasa dan melalui proses yang disetujui, cara cepat bagi pengguna untuk mendapatkan akses ke Akun AWS yang biasanya tidak memiliki izin untuk mengaksesnya. Untuk informasi lebih lanjut, lihat indikator Implementasikan prosedur break-glass dalam panduan Well-Architected AWS .

strategi brownfield

Infrastruktur yang ada di lingkungan Anda. Saat mengadopsi strategi brownfield untuk arsitektur sistem, Anda merancang arsitektur di sekitar kendala sistem dan infrastruktur saat ini. Jika Anda memperluas infrastruktur yang ada, Anda dapat memadukan strategi brownfield dan greenfield.

cache penyangga

Area memori tempat data yang paling sering diakses disimpan.

kemampuan bisnis

Apa yang dilakukan bisnis untuk menghasilkan nilai (misalnya, penjualan, layanan pelanggan, atau pemasaran). Arsitektur layanan mikro dan keputusan pengembangan dapat didorong oleh kemampuan bisnis. Untuk informasi selengkapnya, lihat bagian Terorganisir di sekitar kemampuan bisnis dari Menjalankan layanan mikro kontainer di whitepaper. AWS

perencanaan kelangsungan bisnis (BCP)

Rencana yang membahas dampak potensial dari peristiwa yang mengganggu, seperti migrasi skala besar, pada operasi dan memungkinkan bisnis untuk melanjutkan operasi dengan cepat.

C

KAFE

Lihat Kerangka Adopsi AWS Cloud.

penyebaran kenari

Rilis versi yang lambat dan bertahap untuk pengguna akhir. Ketika Anda yakin, Anda menyebarkan versi baru dan mengganti versi saat ini secara keseluruhan.

CCoE

Lihat Cloud Center of Excellence.

CDC

Lihat mengubah pengambilan data.

ubah pengambilan data (CDC)

Proses melacak perubahan ke sumber data, seperti tabel database, dan merekam metadata tentang perubahan tersebut. Anda dapat menggunakan CDC untuk berbagai tujuan, seperti mengaudit atau mereplikasi perubahan dalam sistem target untuk mempertahankan sinkronisasi.

rekayasa kekacauan

Dengan sengaja memperkenalkan kegagalan atau peristiwa yang mengganggu untuk menguji ketahanan sistem. Anda dapat menggunakan AWS Fault Injection Service (AWS FIS) untuk melakukan eksperimen yang menekankan AWS beban kerja Anda dan mengevaluasi responsnya.

CI/CD

Lihat integrasi berkelanjutan dan pengiriman berkelanjutan.

klasifikasi

Proses kategorisasi yang membantu menghasilkan prediksi. Model ML untuk masalah klasifikasi memprediksi nilai diskrit. Nilai diskrit selalu berbeda satu sama lain. Misalnya, model mungkin perlu mengevaluasi apakah ada mobil dalam gambar atau tidak.

Enkripsi sisi klien

Enkripsi data secara lokal, sebelum target Layanan AWS menerimanya.

Pusat Keunggulan Cloud (CCoE)

Tim multi-disiplin yang mendorong upaya adopsi cloud di seluruh organisasi, termasuk mengembangkan praktik terbaik cloud, memobilisasi sumber daya, menetapkan jadwal migrasi, dan memimpin organisasi melalui transformasi skala besar. Untuk informasi selengkapnya, lihat posting CCo E di Blog Strategi AWS Cloud Perusahaan.

komputasi cloud

Teknologi cloud yang biasanya digunakan untuk penyimpanan data jarak jauh dan manajemen perangkat IoT. Cloud computing umumnya terhubung ke teknologi edge computing.

model operasi cloud

Dalam organisasi TI, model operasi yang digunakan untuk membangun, mematangkan, dan mengoptimalkan satu atau lebih lingkungan cloud. Untuk informasi selengkapnya, lihat Membangun Model Operasi Cloud Anda.

tahap adopsi cloud

Empat fase yang biasanya dilalui organisasi ketika mereka bermigrasi ke AWS Cloud:

- Proyek — Menjalankan beberapa proyek terkait cloud untuk bukti konsep dan tujuan pembelajaran
- Foundation — Melakukan investasi dasar untuk meningkatkan adopsi cloud Anda (misalnya, membuat landing zone, mendefinisikan CCo E, membuat model operasi)
- Migrasi — Migrasi aplikasi individual
- Re-invention — Mengoptimalkan produk dan layanan, dan berinovasi di cloud

Tahapan ini didefinisikan oleh Stephen Orban dalam posting blog The Journey Toward Cloud-First & the Stages of Adoption di blog Strategi Perusahaan. AWS Cloud Untuk informasi tentang bagaimana kaitannya dengan strategi AWS migrasi, lihat panduan kesiapan migrasi.

CMDB

Lihat database manajemen konfigurasi.

repositori kode

Lokasi di mana kode sumber dan aset lainnya, seperti dokumentasi, sampel, dan skrip, disimpan dan diperbarui melalui proses kontrol versi. Repositori cloud umum termasuk GitHub atau. Bitbucket Cloud Setiap versi kode disebut cabang. Dalam struktur layanan mikro, setiap repositori

dikhususkan untuk satu bagian fungsionalitas. Pipa CI/CD tunggal dapat menggunakan beberapa repositori.

cache dingin

Cache buffer yang kosong, tidak terisi dengan baik, atau berisi data basi atau tidak relevan. Ini mempengaruhi kinerja karena instance database harus membaca dari memori utama atau disk, yang lebih lambat daripada membaca dari cache buffer.

data dingin

Data yang jarang diakses dan biasanya historis. Saat menanyakan jenis data ini, kueri lambat biasanya dapat diterima. Memindahkan data ini ke tingkat atau kelas penyimpanan yang berkinerja lebih rendah dan lebih murah dapat mengurangi biaya.

visi komputer (CV)

Bidang AI yang menggunakan pembelajaran mesin untuk menganalisis dan mengekstrak informasi dari format visual seperti gambar dan video digital. Misalnya, Amazon SageMaker AI menyediakan algoritma pemrosesan gambar untuk CV.

konfigurasi drift

Untuk beban kerja, konfigurasi berubah dari status yang diharapkan. Ini dapat menyebabkan beban kerja menjadi tidak patuh, dan biasanya bertahap dan tidak disengaja.

database manajemen konfigurasi (CMDB)

Repositori yang menyimpan dan mengelola informasi tentang database dan lingkungan TI, termasuk komponen perangkat keras dan perangkat lunak dan konfigurasinya. Anda biasanya menggunakan data dari CMDB dalam penemuan portofolio dan tahap analisis migrasi.

paket kesesuaian

Kumpulan AWS Config aturan dan tindakan remediasi yang dapat Anda kumpulkan untuk menyesuaikan kepatuhan dan pemeriksaan keamanan Anda. Anda dapat menerapkan paket kesesuaian sebagai entitas tunggal di Akun AWS dan Region, atau di seluruh organisasi, dengan menggunakan templat YAMM. Untuk informasi selengkapnya, lihat Paket kesesuaian dalam dokumentasi. AWS Config

integrasi berkelanjutan dan pengiriman berkelanjutan (CI/CD)

Proses mengotomatiskan sumber, membangun, menguji, pementasan, dan tahap produksi dari proses rilis perangkat lunak. CI/CD is commonly described as a pipeline. CI/CDdapat membantu

Anda mengotomatiskan proses, meningkatkan produktivitas, meningkatkan kualitas kode, dan memberikan lebih cepat. Untuk informasi lebih lanjut, lihat Manfaat pengiriman berkelanjutan. CD juga dapat berarti penerapan berkelanjutan. Untuk informasi selengkapnya, lihat Continuous Delivery vs Continuous Deployment.

CV

Lihat visi komputer.

D

data saat istirahat

Data yang stasioner di jaringan Anda, seperti data yang ada di penyimpanan.

klasifikasi data

Proses untuk mengidentifikasi dan mengkategorikan data dalam jaringan Anda berdasarkan kekritisan dan sensitivitasnya. Ini adalah komponen penting dari setiap strategi manajemen risiko keamanan siber karena membantu Anda menentukan perlindungan dan kontrol retensi yang tepat untuk data. Klasifikasi data adalah komponen pilar keamanan dalam AWS Well-Architected Framework. Untuk informasi selengkapnya, lihat Klasifikasi data.

penyimpangan data

Variasi yang berarti antara data produksi dan data yang digunakan untuk melatih model ML, atau perubahan yang berarti dalam data input dari waktu ke waktu. Penyimpangan data dapat mengurangi kualitas, akurasi, dan keadilan keseluruhan dalam prediksi model ML.

data dalam transit

Data yang aktif bergerak melalui jaringan Anda, seperti antara sumber daya jaringan.

jala data

Kerangka arsitektur yang menyediakan kepemilikan data terdistribusi dan terdesentralisasi dengan manajemen dan tata kelola terpusat.

minimalisasi data

Prinsip pengumpulan dan pemrosesan hanya data yang sangat diperlukan. Mempraktikkan minimalisasi data di dalamnya AWS Cloud dapat mengurangi risiko privasi, biaya, dan jejak karbon analitik Anda.

perimeter data

Satu set pagar pembatas pencegahan di AWS lingkungan Anda yang membantu memastikan bahwa hanya identitas tepercaya yang mengakses sumber daya tepercaya dari jaringan yang diharapkan. Untuk informasi selengkapnya, lihat Membangun perimeter data pada AWS.

prapemrosesan data

Untuk mengubah data mentah menjadi format yang mudah diuraikan oleh model ML Anda. Preprocessing data dapat berarti menghapus kolom atau baris tertentu dan menangani nilai yang hilang, tidak konsisten, atau duplikat.

asal data

Proses melacak asal dan riwayat data sepanjang siklus hidupnya, seperti bagaimana data dihasilkan, ditransmisikan, dan disimpan.

subjek data

Individu yang datanya dikumpulkan dan diproses.

gudang data

Sistem manajemen data yang mendukung intelijen bisnis, seperti analitik. Gudang data biasanya berisi sejumlah besar data historis, dan biasanya digunakan untuk kueri dan analisis.

bahasa definisi database (DDL)

Pernyataan atau perintah untuk membuat atau memodifikasi struktur tabel dan objek dalam database.

bahasa manipulasi basis data (DHTML)

Pernyataan atau perintah untuk memodifikasi (memasukkan, memperbarui, dan menghapus) informasi dalam database.

DDL

Lihat bahasa definisi database.

ansambel yang dalam

Untuk menggabungkan beberapa model pembelajaran mendalam untuk prediksi. Anda dapat menggunakan ansambel dalam untuk mendapatkan prediksi yang lebih akurat atau untuk memperkirakan ketidakpastian dalam prediksi.

pembelajaran mendalam

Subbidang ML yang menggunakan beberapa lapisan jaringan saraf tiruan untuk mengidentifikasi pemetaan antara data input dan variabel target yang diinginkan.

defense-in-depth

Pendekatan keamanan informasi di mana serangkaian mekanisme dan kontrol keamanan dilapisi dengan cermat di seluruh jaringan komputer untuk melindungi kerahasiaan, integritas, dan ketersediaan jaringan dan data di dalamnya. Saat Anda mengadopsi strategi ini AWS, Anda menambahkan beberapa kontrol pada lapisan AWS Organizations struktur yang berbeda untuk membantu mengamankan sumber daya. Misalnya, defense-in-depth pendekatan mungkin menggabungkan otentikasi multi-faktor, segmentasi jaringan, dan enkripsi.

administrator yang didelegasikan

Di AWS Organizations, layanan yang kompatibel dapat mendaftarkan akun AWS anggota untuk mengelola akun organisasi dan mengelola izin untuk layanan tersebut. Akun ini disebut administrator yang didelegasikan untuk layanan itu. Untuk informasi selengkapnya dan daftar layanan yang kompatibel, lihat Layanan yang berfungsi dengan AWS Organizations AWS Organizations dokumentasi.

deployment

Proses pembuatan aplikasi, fitur baru, atau perbaikan kode tersedia di lingkungan target. Deployment melibatkan penerapan perubahan dalam basis kode dan kemudian membangun dan menjalankan basis kode itu di lingkungan aplikasi.

lingkungan pengembangan

Lihat lingkungan.

kontrol detektif

Kontrol keamanan yang dirancang untuk mendeteksi, mencatat, dan memperingatkan setelah suatu peristiwa terjadi. Kontrol ini adalah garis pertahanan kedua, memperingatkan Anda tentang peristiwa keamanan yang melewati kontrol pencegahan yang ada. Untuk informasi selengkapnya, lihat Kontrol Detektif dalam Menerapkan kontrol keamanan pada. AWS

pemetaan aliran nilai pengembangan (DVSM)

Sebuah proses yang digunakan untuk mengidentifikasi dan memprioritaskan kendala yang mempengaruhi kecepatan dan kualitas dalam siklus hidup pengembangan perangkat lunak. DVSM memperluas proses pemetaan aliran nilai yang awalnya dirancang untuk praktik

manufaktur ramping. Ini berfokus pada langkah-langkah dan tim yang diperlukan untuk menciptakan dan memindahkan nilai melalui proses pengembangan perangkat lunak.

kembar digital

Representasi virtual dari sistem dunia nyata, seperti bangunan, pabrik, peralatan industri, atau jalur produksi. Kembar digital mendukung pemeliharaan prediktif, pemantauan jarak jauh, dan optimalisasi produksi.

tabel dimensi

Dalam skema bintang, tabel yang lebih kecil yang berisi atribut data tentang data kuantitatif dalam tabel fakta. Atribut tabel dimensi biasanya bidang teks atau angka diskrit yang berperilaku seperti teks. Atribut ini biasanya digunakan untuk pembatasan kueri, pemfilteran, dan pelabelan set hasil.

musibah

Peristiwa yang mencegah beban kerja atau sistem memenuhi tujuan bisnisnya di lokasi utama yang digunakan. Peristiwa ini dapat berupa bencana alam, kegagalan teknis, atau akibat dari tindakan manusia, seperti kesalahan konfigurasi yang tidak disengaja atau serangan malware.

pemulihan bencana (DR)

Strategi dan proses yang Anda gunakan untuk meminimalkan downtime dan kehilangan data yang disebabkan oleh bencana. Untuk informasi selengkapnya, lihat Disaster Recovery of Workloads on AWS: Recovery in the Cloud in the AWS Well-Architected Framework.

DML~

Lihat bahasa manipulasi basis data.

desain berbasis domain

Pendekatan untuk mengembangkan sistem perangkat lunak yang kompleks dengan menghubungkan komponennya ke domain yang berkembang, atau tujuan bisnis inti, yang dilayani oleh setiap komponen. Konsep ini diperkenalkan oleh Eric Evans dalam bukunya, Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston: Addison-Wesley Professional, 2003). Untuk informasi tentang cara menggunakan desain berbasis domain dengan pola gambar pencekik, lihat Memodernisasi layanan web Microsoft ASP.NET (ASMX) lama secara bertahap menggunakan container dan Amazon API Gateway.

DR

Lihat pemulihan bencana.

deteksi drift

Melacak penyimpangan dari konfigurasi dasar. Misalnya, Anda dapat menggunakan AWS CloudFormation untuk mendeteksi penyimpangan dalam sumber daya sistem, atau Anda dapat menggunakannya AWS Control Tower untuk mendeteksi perubahan di landing zone yang mungkin memengaruhi kepatuhan terhadap persyaratan tata kelola.

DVSM

Lihat pemetaan aliran nilai pengembangan.

E

EDA

Lihat analisis data eksplorasi.

EDI

Lihat pertukaran data elektronik.

komputasi tepi

Teknologi yang meningkatkan daya komputasi untuk perangkat pintar di tepi jaringan IoT. Jika dibandingkan dengan komputasi awan, komputasi tepi dapat mengurangi latensi komunikasi dan meningkatkan waktu respons.

pertukaran data elektronik (EDI)

Pertukaran otomatis dokumen bisnis antar organisasi. Untuk informasi selengkapnya, lihat Apa itu Pertukaran Data Elektronik.

enkripsi

Proses komputasi yang mengubah data plaintext, yang dapat dibaca manusia, menjadi ciphertext.

kunci enkripsi

String kriptografi dari bit acak yang dihasilkan oleh algoritma enkripsi. Panjang kunci dapat bervariasi, dan setiap kunci dirancang agar tidak dapat diprediksi dan unik.

endianness

Urutan byte disimpan dalam memori komputer. Sistem big-endian menyimpan byte paling signifikan terlebih dahulu. Sistem little-endian menyimpan byte paling tidak signifikan terlebih dahulu.

titik akhir

Lihat titik akhir layanan.

layanan endpoint

Layanan yang dapat Anda host di cloud pribadi virtual (VPC) untuk dibagikan dengan pengguna lain. Anda dapat membuat layanan endpoint dengan AWS PrivateLink dan memberikan izin kepada prinsipal lain Akun AWS atau ke AWS Identity and Access Management (IAM). Akun atau prinsipal ini dapat terhubung ke layanan endpoint Anda secara pribadi dengan membuat titik akhir VPC antarmuka. Untuk informasi selengkapnya, lihat Membuat layanan titik akhir di dokumentasi Amazon Virtual Private Cloud (Amazon VPC).

perencanaan sumber daya perusahaan (ERP)

Sistem yang mengotomatiskan dan mengelola proses bisnis utama (seperti akuntansi, MES, dan manajemen proyek) untuk suatu perusahaan.

enkripsi amplop

Proses mengenkripsi kunci enkripsi dengan kunci enkripsi lain. Untuk informasi selengkapnya, lihat Enkripsi amplop dalam dokumentasi AWS Key Management Service (AWS KMS).

lingkungan

Sebuah contoh dari aplikasi yang sedang berjalan. Berikut ini adalah jenis lingkungan yang umum dalam komputasi awan:

- Development Environment — Sebuah contoh dari aplikasi yang berjalan yang hanya tersedia untuk tim inti yang bertanggung jawab untuk memelihara aplikasi. Lingkungan pengembangan digunakan untuk menguji perubahan sebelum mempromosikannya ke lingkungan atas. Jenis lingkungan ini kadang-kadang disebut sebagai lingkungan pengujian.
- lingkungan yang lebih rendah — Semua lingkungan pengembangan untuk aplikasi, seperti yang digunakan untuk build awal dan pengujian.
- lingkungan produksi — Sebuah contoh dari aplikasi yang berjalan yang pengguna akhir dapat mengakses. Dalam pipa CI/CD, lingkungan produksi adalah lingkungan penyebaran terakhir.

- lingkungan atas — Semua lingkungan yang dapat diakses oleh pengguna selain tim pengembangan inti. Ini dapat mencakup lingkungan produksi, lingkungan praproduksi, dan lingkungan untuk pengujian penerimaan pengguna.

epik

Dalam metodologi tangkas, kategori fungsional yang membantu mengatur dan memprioritaskan pekerjaan Anda. Epik memberikan deskripsi tingkat tinggi tentang persyaratan dan tugas implementasi. Misalnya, epos keamanan AWS CAF mencakup manajemen identitas dan akses, kontrol detektif, keamanan infrastruktur, perlindungan data, dan respons insiden. Untuk informasi selengkapnya tentang epos dalam strategi AWS migrasi, lihat panduan implementasi program.

ERP

Lihat perencanaan sumber daya perusahaan.

analisis data eksplorasi (EDA)

Proses menganalisis dataset untuk memahami karakteristik utamanya. Anda mengumpulkan atau mengumpulkan data dan kemudian melakukan penyelidikan awal untuk menemukan pola, mendeteksi anomali, dan memeriksa asumsi. EDA dilakukan dengan menghitung statistik ringkasan dan membuat visualisasi data.

F

tabel fakta

Tabel tengah dalam skema bintang. Ini menyimpan data kuantitatif tentang operasi bisnis. Biasanya, tabel fakta berisi dua jenis kolom: kolom yang berisi ukuran dan yang berisi kunci asing ke tabel dimensi.

gagal cepat

Filosofi yang menggunakan pengujian yang sering dan bertahap untuk mengurangi siklus hidup pengembangan. Ini adalah bagian penting dari pendekatan tangkas.

batas isolasi kesalahan

Dalam AWS Cloud, batas seperti Availability Zone, Wilayah AWS, control plane, atau data plane yang membatasi efek kegagalan dan membantu meningkatkan ketahanan beban kerja. Untuk informasi selengkapnya, lihat Batas Isolasi AWS Kesalahan.

cabang fitur

Lihat cabang.

fitur

Data input yang Anda gunakan untuk membuat prediksi. Misalnya, dalam konteks manufaktur, fitur bisa berupa gambar yang diambil secara berkala dari lini manufaktur.

pentingnya fitur

Seberapa signifikan fitur untuk prediksi model. Ini biasanya dinyatakan sebagai skor numerik yang dapat dihitung melalui berbagai teknik, seperti Shapley Additive Explanations (SHAP) dan gradien terintegrasi. Untuk informasi lebih lanjut, lihat Interpretabilitas model pembelajaran mesin dengan. AWS

transformasi fitur

Untuk mengoptimalkan data untuk proses ML, termasuk memperkaya data dengan sumber tambahan, menskalakan nilai, atau mengekstrak beberapa set informasi dari satu bidang data. Hal ini memungkinkan model ML untuk mendapatkan keuntungan dari data. Misalnya, jika Anda memecah tanggal “2021-05-27 00:15:37" menjadi “2021", “Mei”, “Kamis”, dan “15", Anda dapat membantu algoritme pembelajaran mempelajari pola bernuansa yang terkait dengan komponen data yang berbeda.

beberapa tembakan mendorong

Menyediakan LLM dengan sejumlah kecil contoh yang menunjukkan tugas dan output yang diinginkan sebelum memintanya untuk melakukan tugas serupa. Teknik ini adalah aplikasi pembelajaran dalam konteks, di mana model belajar dari contoh (bidikan) yang tertanam dalam petunjuk. Beberapa bidikan dapat efektif untuk tugas-tugas yang memerlukan pemformatan, penalaran, atau pengetahuan domain tertentu. Lihat juga bidikan nol.

FGAC

Lihat kontrol akses berbutir halus.

kontrol akses berbutir halus (FGAC)

Penggunaan beberapa kondisi untuk mengizinkan atau menolak permintaan akses.

migrasi flash-cut

Metode migrasi database yang menggunakan replikasi data berkelanjutan melalui pengambilan data perubahan untuk memigrasikan data dalam waktu sesingkat mungkin, alih-alih

menggunakan pendekatan bertahap. Tujuannya adalah untuk menjaga downtime seminimal mungkin.

FM

Lihat model pondasi.

model pondasi (FM)

Jaringan saraf pembelajaran mendalam yang besar yang telah melatih kumpulan data besar-besaran data umum dan tidak berlabel. FMs mampu melakukan berbagai tugas umum, seperti memahami bahasa, menghasilkan teks dan gambar, dan berbicara dalam bahasa alami. Untuk informasi selengkapnya, lihat Apa itu Model Foundation.

G

AI generatif

Subset model AI yang telah dilatih pada sejumlah besar data dan yang dapat menggunakan prompt teks sederhana untuk membuat konten dan artefak baru, seperti gambar, video, teks, dan audio. Untuk informasi lebih lanjut, lihat Apa itu AI Generatif.

pemblokiran geografis

Lihat pembatasan geografis.

pembatasan geografis (pemblokiran geografis)

Di Amazon CloudFront, opsi untuk mencegah pengguna di negara tertentu mengakses distribusi konten. Anda dapat menggunakan daftar izinkan atau daftar blokir untuk menentukan negara yang disetujui dan dilarang. Untuk informasi selengkapnya, lihat Membatasi distribusi geografis konten Anda dalam dokumentasi. CloudFront

Alur kerja Gitflow

Pendekatan di mana lingkungan bawah dan atas menggunakan cabang yang berbeda dalam repositori kode sumber. Alur kerja Gitflow dianggap warisan, dan alur kerja berbasis batang adalah pendekatan modern yang lebih disukai.

gambar emas

Sebuah snapshot dari sistem atau perangkat lunak yang digunakan sebagai template untuk menyebarkan instance baru dari sistem atau perangkat lunak itu. Misalnya, di bidang manufaktur,

gambar emas dapat digunakan untuk menyediakan perangkat lunak pada beberapa perangkat dan membantu meningkatkan kecepatan, skalabilitas, dan produktivitas dalam operasi manufaktur perangkat.

strategi greenfield

Tidak adanya infrastruktur yang ada di lingkungan baru. Saat mengadopsi strategi greenfield untuk arsitektur sistem, Anda dapat memilih semua teknologi baru tanpa batasan kompatibilitas dengan infrastruktur yang ada, juga dikenal sebagai brownfield. Jika Anda memperluas infrastruktur yang ada, Anda dapat memadukan strategi brownfield dan greenfield.

pagar pembatas

Aturan tingkat tinggi yang membantu mengatur sumber daya, kebijakan, dan kepatuhan di seluruh unit organisasi ()OUs. Pagar pembatas preventif menegakkan kebijakan untuk memastikan keselarasan dengan standar kepatuhan. Mereka diimplementasikan dengan menggunakan kebijakan kontrol layanan dan batas izin IAM. Detective guardrails mendeteksi pelanggaran kebijakan dan masalah kepatuhan, dan menghasilkan peringatan untuk remediasi. Mereka diimplementasikan dengan menggunakan AWS Config, AWS Security Hub, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector, dan pemeriksaan khusus AWS Lambda .

H

HA

Lihat ketersediaan tinggi.

migrasi database heterogen

Memigrasi database sumber Anda ke database target yang menggunakan mesin database yang berbeda (misalnya, Oracle ke Amazon Aurora). Migrasi heterogen biasanya merupakan bagian dari upaya arsitektur ulang, dan mengubah skema dapat menjadi tugas yang kompleks. AWS menyediakan AWS SCT yang membantu dengan konversi skema.

ketersediaan tinggi (HA)

Kemampuan beban kerja untuk beroperasi terus menerus, tanpa intervensi, jika terjadi tantangan atau bencana. Sistem HA dirancang untuk gagal secara otomatis, secara konsisten memberikan kinerja berkualitas tinggi, dan menangani beban dan kegagalan yang berbeda dengan dampak kinerja minimal.

modernisasi sejarawan

Pendekatan yang digunakan untuk memodernisasi dan meningkatkan sistem teknologi operasional (OT) untuk melayani kebutuhan industri manufaktur dengan lebih baik. Sejarawan adalah jenis database yang digunakan untuk mengumpulkan dan menyimpan data dari berbagai sumber di pabrik.

data penahanan

Sebagian dari data historis berlabel yang ditahan dari kumpulan data yang digunakan untuk melatih model pembelajaran mesin. Anda dapat menggunakan data penahanan untuk mengevaluasi kinerja model dengan membandingkan prediksi model dengan data penahanan.

migrasi database homogen

Memigrasi database sumber Anda ke database target yang berbagi mesin database yang sama (misalnya, Microsoft SQL Server ke Amazon RDS for SQL Server). Migrasi homogen biasanya merupakan bagian dari upaya rehosting atau replatforming. Anda dapat menggunakan utilitas database asli untuk memigrasi skema.

data panas

Data yang sering diakses, seperti data real-time atau data translasi terbaru. Data ini biasanya memerlukan tingkat atau kelas penyimpanan berkinerja tinggi untuk memberikan respons kueri yang cepat.

perbaikan terbaru

Perbaikan mendesak untuk masalah kritis dalam lingkungan produksi. Karena urgensinya, perbaikan terbaru biasanya dibuat di luar alur kerja DevOps rilis biasa.

periode hypercare

Segera setelah cutover, periode waktu ketika tim migrasi mengelola dan memantau aplikasi yang dimigrasi di cloud untuk mengatasi masalah apa pun. Biasanya, periode ini panjangnya 1-4 hari. Pada akhir periode hypercare, tim migrasi biasanya mentransfer tanggung jawab untuk aplikasi ke tim operasi cloud.

I

IAc

Lihat infrastruktur sebagai kode.

kebijakan berbasis identitas

Kebijakan yang dilampirkan pada satu atau beberapa prinsip IAM yang mendefinisikan izin mereka dalam lingkungan. AWS Cloud

aplikasi idle

Aplikasi yang memiliki penggunaan CPU dan memori rata-rata antara 5 dan 20 persen selama periode 90 hari. Dalam proyek migrasi, adalah umum untuk menghentikan aplikasi ini atau mempertahankannya di tempat.

IIoT

Lihat Internet of Things industri.

infrastruktur yang tidak dapat diubah

Model yang menyebarkan infrastruktur baru untuk beban kerja produksi alih-alih memperbarui, menambal, atau memodifikasi infrastruktur yang ada. Infrastruktur yang tidak dapat diubah secara inheren lebih konsisten, andal, dan dapat diprediksi daripada infrastruktur yang dapat berubah. Untuk informasi selengkapnya, lihat praktik terbaik Deploy using immutable infrastructure di AWS Well-Architected Framework.

masuk (masuknya) VPC

Dalam arsitektur AWS multi-akun, VPC yang menerima, memeriksa, dan merutekan koneksi jaringan dari luar aplikasi. Arsitektur Referensi AWS Keamanan merekomendasikan pengaturan akun Jaringan Anda dengan inbound, outbound, dan inspeksi VPCs untuk melindungi antarmuka dua arah antara aplikasi Anda dan internet yang lebih luas.

migrasi inkremental

Strategi cutover di mana Anda memigrasikan aplikasi Anda dalam bagian-bagian kecil alih-alih melakukan satu cutover penuh. Misalnya, Anda mungkin hanya memindahkan beberapa layanan mikro atau pengguna ke sistem baru pada awalnya. Setelah Anda memverifikasi bahwa semuanya berfungsi dengan baik, Anda dapat secara bertahap memindahkan layanan mikro atau pengguna tambahan hingga Anda dapat menonaktifkan sistem lama Anda. Strategi ini mengurangi risiko yang terkait dengan migrasi besar.

Industri 4.0

Sebuah istilah yang diperkenalkan oleh Klaus Schwab pada tahun 2016 untuk merujuk pada modernisasi proses manufaktur melalui kemajuan dalam konektivitas, data real-time, otomatisasi, analitik, dan AI/ML.

infrastruktur

Semua sumber daya dan aset yang terkandung dalam lingkungan aplikasi.

infrastruktur sebagai kode (IAc)

Proses penyediaan dan pengelolaan infrastruktur aplikasi melalui satu set file konfigurasi. IAc dirancang untuk membantu Anda memusatkan manajemen infrastruktur, menstandarisasi sumber daya, dan menskalakan dengan cepat sehingga lingkungan baru dapat diulang, andal, dan konsisten.

Internet of Things industri (IIoT)

Penggunaan sensor dan perangkat yang terhubung ke internet di sektor industri, seperti manufaktur, energi, otomotif, perawatan kesehatan, ilmu kehidupan, dan pertanian. Untuk informasi lebih lanjut, lihat Membangun strategi transformasi digital Internet of Things (IIoT) industri.

inspeksi VPC

Dalam arsitektur AWS multi-akun, VPC terpusat yang mengelola inspeksi lalu lintas jaringan antara VPCs (dalam yang sama atau berbeda Wilayah AWS), internet, dan jaringan lokal. Arsitektur Referensi AWS Keamanan merekomendasikan pengaturan akun Jaringan Anda dengan inbound, outbound, dan inspeksi VPCs untuk melindungi antarmuka dua arah antara aplikasi Anda dan internet yang lebih luas.

Internet of Things (IoT)

Jaringan objek fisik yang terhubung dengan sensor atau prosesor tertanam yang berkomunikasi dengan perangkat dan sistem lain melalui internet atau melalui jaringan komunikasi lokal. Untuk informasi selengkapnya, lihat Apa itu IoT?

interpretabilitas

Karakteristik model pembelajaran mesin yang menggambarkan sejauh mana manusia dapat memahami bagaimana prediksi model bergantung pada inputnya. Untuk informasi lebih lanjut, lihat Interpretabilitas model pembelajaran mesin dengan. AWS

IoT

Lihat Internet of Things.

Perpustakaan informasi TI (ITIL)

Serangkaian praktik terbaik untuk memberikan layanan TI dan menyelaraskan layanan ini dengan persyaratan bisnis. ITIL menyediakan dasar untuk ITSM.

Manajemen layanan TI (ITSM)

Kegiatan yang terkait dengan merancang, menerapkan, mengelola, dan mendukung layanan TI untuk suatu organisasi. Untuk informasi tentang mengintegrasikan operasi cloud dengan alat ITSM, lihat panduan integrasi operasi.

ITIL

Lihat perpustakaan informasi TI.

ITSM

Lihat manajemen layanan TI.

L

kontrol akses berbasis label (LBAC)

Implementasi kontrol akses wajib (MAC) di mana pengguna dan data itu sendiri masing-masing secara eksplisit diberi nilai label keamanan. Persimpangan antara label keamanan pengguna dan label keamanan data menentukan baris dan kolom mana yang dapat dilihat oleh pengguna.

landing zone

Landing zone adalah AWS lingkungan multi-akun yang dirancang dengan baik yang dapat diskalakan dan aman. Ini adalah titik awal dari mana organisasi Anda dapat dengan cepat meluncurkan dan menyebarkan beban kerja dan aplikasi dengan percaya diri dalam lingkungan keamanan dan infrastruktur mereka. Untuk informasi selengkapnya tentang zona pendaratan, lihat Menyiapkan lingkungan multi-akun AWS yang aman dan dapat diskalakan.

model bahasa besar (LLM)

Model AI pembelajaran mendalam yang dilatih sebelumnya pada sejumlah besar data. LLM dapat melakukan beberapa tugas, seperti menjawab pertanyaan, meringkas dokumen, menerjemahkan teks ke dalam bahasa lain, dan menyelesaikan kalimat. Untuk informasi lebih lanjut, lihat Apa itu LLMs.

migrasi besar

Migrasi 300 atau lebih server.

LBAC

Lihat kontrol akses berbasis label.

hak istimewa paling sedikit

Praktik keamanan terbaik untuk memberikan izin minimum yang diperlukan untuk melakukan tugas. Untuk informasi selengkapnya, lihat Menerapkan izin hak istimewa terkecil dalam dokumentasi IAM.

angkat dan geser

Lihat 7 Rs.

sistem endian kecil

Sebuah sistem yang menyimpan byte paling tidak signifikan terlebih dahulu. Lihat juga endianness.

LLM

Lihat model bahasa besar.

lingkungan yang lebih rendah

Lihat lingkungan.

M

pembelajaran mesin (ML)

Jenis kecerdasan buatan yang menggunakan algoritma dan teknik untuk pengenalan pola dan pembelajaran. ML menganalisis dan belajar dari data yang direkam, seperti data Internet of Things (IoT), untuk menghasilkan model statistik berdasarkan pola. Untuk informasi selengkapnya, lihat Machine Learning.

cabang utama

Lihat cabang.

malware

Perangkat lunak yang dirancang untuk membahayakan keamanan atau privasi komputer. Malware dapat mengganggu sistem komputer, membocorkan informasi sensitif, atau mendapatkan akses yang tidak sah. Contoh malware termasuk virus, worm, ransomware, Trojan horse, spyware, dan keyloggers.

layanan terkelola

Layanan AWS yang AWS mengoperasikan lapisan infrastruktur, sistem operasi, dan platform, dan Anda mengakses titik akhir untuk menyimpan dan mengambil data. Amazon Simple Storage Service (Amazon S3) dan Amazon DynamoDB adalah contoh layanan terkelola. Ini juga dikenal sebagai layanan abstrak.

sistem eksekusi manufaktur (MES)

Sistem perangkat lunak untuk melacak, memantau, mendokumentasikan, dan mengendalikan proses produksi yang mengubah bahan baku menjadi produk jadi di lantai toko.

PETA

Lihat Program Percepatan Migrasi.

mekanisme

Proses lengkap di mana Anda membuat alat, mendorong adopsi alat, dan kemudian memeriksa hasilnya untuk melakukan penyesuaian. Mekanisme adalah siklus yang memperkuat dan meningkatkan dirinya sendiri saat beroperasi. Untuk informasi lebih lanjut, lihat Membangun mekanisme di AWS Well-Architected Framework.

akun anggota

Semua Akun AWS selain akun manajemen yang merupakan bagian dari organisasi di AWS Organizations. Akun dapat menjadi anggota dari hanya satu organisasi pada suatu waktu.

MES

Lihat sistem eksekusi manufaktur.

Transportasi Telemetri Antrian Pesan (MQTT)

Protokol komunikasi ringan machine-to-machine (M2M), berdasarkan pola terbitkan/berlangganan, untuk perangkat IoT yang dibatasi sumber daya.

layanan mikro

Layanan kecil dan independen yang berkomunikasi dengan jelas APIs dan biasanya dimiliki oleh tim kecil yang mandiri. Misalnya, sistem asuransi mungkin mencakup layanan mikro yang memetakan kemampuan bisnis, seperti penjualan atau pemasaran, atau subdomain, seperti pembelian, klaim, atau analitik. Manfaat layanan mikro termasuk kelincahan, penskalaan yang fleksibel, penyebaran yang mudah, kode yang dapat digunakan kembali, dan ketahanan. Untuk informasi selengkapnya, lihat Mengintegrasikan layanan mikro dengan menggunakan layanan tanpa AWS server.

arsitektur microservices

Pendekatan untuk membangun aplikasi dengan komponen independen yang menjalankan setiap proses aplikasi sebagai layanan mikro. Layanan mikro ini berkomunikasi melalui antarmuka yang terdefinisi dengan baik dengan menggunakan ringan. APIs Setiap layanan mikro dalam arsitektur ini dapat diperbarui, digunakan, dan diskalakan untuk memenuhi permintaan fungsi tertentu dari suatu aplikasi. Untuk informasi selengkapnya, lihat Menerapkan layanan mikro di AWS.

Program Percepatan Migrasi (MAP)

AWS Program yang menyediakan dukungan konsultasi, pelatihan, dan layanan untuk membantu organisasi membangun fondasi operasional yang kuat untuk pindah ke cloud, dan untuk membantu mengimbangi biaya awal migrasi. MAP mencakup metodologi migrasi untuk mengeksekusi migrasi lama dengan cara metodis dan seperangkat alat untuk mengotomatisasi dan mempercepat skenario migrasi umum.

migrasi dalam skala

Proses memindahkan sebagian besar portofolio aplikasi ke cloud dalam gelombang, dengan lebih banyak aplikasi bergerak pada tingkat yang lebih cepat di setiap gelombang. Fase ini menggunakan praktik dan pelajaran terbaik dari fase sebelumnya untuk mengimplementasikan pabrik migrasi tim, alat, dan proses untuk merampingkan migrasi beban kerja melalui otomatisasi dan pengiriman tangkas. Ini adalah fase ketiga dari strategi AWS migrasi.

pabrik migrasi

Tim lintas fungsi yang merampingkan migrasi beban kerja melalui pendekatan otomatis dan gesit. Tim pabrik migrasi biasanya mencakup operasi, analis dan pemilik bisnis, insinyur migrasi, pengembang, dan DevOps profesional yang bekerja di sprint. Antara 20 dan 50 persen portofolio aplikasi perusahaan terdiri dari pola berulang yang dapat dioptimalkan dengan pendekatan pabrik. Untuk informasi selengkapnya, lihat diskusi tentang pabrik migrasi dan panduan Pabrik Migrasi Cloud di kumpulan konten ini.

metadata migrasi

Informasi tentang aplikasi dan server yang diperlukan untuk menyelesaikan migrasi. Setiap pola migrasi memerlukan satu set metadata migrasi yang berbeda. Contoh metadata migrasi termasuk subnet target, grup keamanan, dan akun. AWS

pola migrasi

Tugas migrasi berulang yang merinci strategi migrasi, tujuan migrasi, dan aplikasi atau layanan migrasi yang digunakan. Contoh: Rehost migrasi ke Amazon EC2 dengan Layanan Migrasi AWS Aplikasi.

Penilaian Portofolio Migrasi (MPA)

Alat online yang menyediakan informasi untuk memvalidasi kasus bisnis untuk bermigrasi ke. AWS Cloud MPA menyediakan penilaian portofolio terperinci (ukuran kanan server, harga, perbandingan TCO, analisis biaya migrasi) serta perencanaan migrasi (analisis data aplikasi dan pengumpulan data, pengelompokan aplikasi, prioritas migrasi, dan perencanaan gelombang). Alat MPA (memerlukan login) tersedia gratis untuk semua AWS konsultan dan konsultan APN Partner.

Penilaian Kesiapan Migrasi (MRA)

Proses mendapatkan wawasan tentang status kesiapan cloud organisasi, mengidentifikasi kekuatan dan kelemahan, dan membangun rencana aksi untuk menutup kesenjangan yang diidentifikasi, menggunakan CAF. AWS Untuk informasi selengkapnya, lihat panduan kesiapan migrasi. MRA adalah tahap pertama dari strategi AWS migrasi.

strategi migrasi

Pendekatan yang digunakan untuk memigrasikan beban kerja ke file. AWS Cloud Untuk informasi lebih lanjut, lihat entri 7 Rs di glosarium ini dan lihat Memobilisasi organisasi Anda untuk mempercepat migrasi skala besar.

ML

Lihat pembelajaran mesin.

modernisasi

Mengubah aplikasi usang (warisan atau monolitik) dan infrastrukturnya menjadi sistem yang gesit, elastis, dan sangat tersedia di cloud untuk mengurangi biaya, mendapatkan efisiensi, dan memanfaatkan inovasi. Untuk informasi selengkapnya, lihat Strategi untuk memodernisasi aplikasi di. AWS Cloud

penilaian kesiapan modernisasi

Evaluasi yang membantu menentukan kesiapan modernisasi aplikasi organisasi; mengidentifikasi manfaat, risiko, dan dependensi; dan menentukan seberapa baik organisasi dapat mendukung keadaan masa depan aplikasi tersebut. Hasil penilaian adalah cetak biru arsitektur target, peta jalan yang merinci fase pengembangan dan tonggak untuk proses modernisasi, dan rencana aksi untuk mengatasi kesenjangan yang diidentifikasi. Untuk informasi lebih lanjut, lihat Mengevaluasi kesiapan modernisasi untuk aplikasi di. AWS Cloud

aplikasi monolitik (monolit)

Aplikasi yang berjalan sebagai layanan tunggal dengan proses yang digabungkan secara ketat. Aplikasi monolitik memiliki beberapa kelemahan. Jika satu fitur aplikasi mengalami lonjakan permintaan, seluruh arsitektur harus diskalakan. Menambahkan atau meningkatkan fitur aplikasi monolitik juga menjadi lebih kompleks ketika basis kode tumbuh. Untuk mengatasi masalah ini, Anda dapat menggunakan arsitektur microservices. Untuk informasi lebih lanjut, lihat Menguraikan monolit menjadi layanan mikro.

MPA

Lihat Penilaian Portofolio Migrasi.

MQTT

Lihat Transportasi Telemetri Antrian Pesan.

klasifikasi multiclass

Sebuah proses yang membantu menghasilkan prediksi untuk beberapa kelas (memprediksi satu dari lebih dari dua hasil). Misalnya, model ML mungkin bertanya “Apakah produk ini buku, mobil, atau telepon?” atau “Kategori produk mana yang paling menarik bagi pelanggan ini?”

infrastruktur yang bisa berubah

Model yang memperbarui dan memodifikasi infrastruktur yang ada untuk beban kerja produksi. Untuk meningkatkan konsistensi, keandalan, dan prediktabilitas, AWS Well-Architected Framework merekomendasikan penggunaan infrastruktur yang tidak dapat diubah sebagai praktik terbaik.

O

OAC

Lihat kontrol akses asal.

OAI

Lihat identitas akses asal.

OCM

Lihat manajemen perubahan organisasi.

migrasi offline

Metode migrasi di mana beban kerja sumber diturunkan selama proses migrasi. Metode ini melibatkan waktu henti yang diperpanjang dan biasanya digunakan untuk beban kerja kecil dan tidak kritis.

OI

Lihat integrasi operasi.

OLA

Lihat perjanjian tingkat operasional.

migrasi online

Metode migrasi di mana beban kerja sumber disalin ke sistem target tanpa diambil offline. Aplikasi yang terhubung ke beban kerja dapat terus berfungsi selama migrasi. Metode ini melibatkan waktu henti nol hingga minimal dan biasanya digunakan untuk beban kerja produksi yang kritis.

OPC-UA

Lihat Komunikasi Proses Terbuka - Arsitektur Terpadu.

Komunikasi Proses Terbuka - Arsitektur Terpadu (OPC-UA)

Protokol komunikasi machine-to-machine (M2M) untuk otomasi industri. OPC-UA menyediakan standar interoperabilitas dengan enkripsi data, otentikasi, dan skema otorisasi.

perjanjian tingkat operasional (OLA)

Perjanjian yang menjelaskan apa yang dijanjikan kelompok TI fungsional untuk diberikan satu sama lain, untuk mendukung perjanjian tingkat layanan (SLA).

Tinjauan Kesiapan Operasional (ORR)

Daftar pertanyaan dan praktik terbaik terkait yang membantu Anda memahami, mengevaluasi, mencegah, atau mengurangi ruang lingkup insiden dan kemungkinan kegagalan. Untuk informasi lebih lanjut, lihat Ulasan Kesiapan Operasional (ORR) dalam Kerangka Kerja Well-Architected AWS .

teknologi operasional (OT)

Sistem perangkat keras dan perangkat lunak yang bekerja dengan lingkungan fisik untuk mengendalikan operasi industri, peralatan, dan infrastruktur. Di bidang manufaktur, integrasi sistem OT dan teknologi informasi (TI) adalah fokus utama untuk transformasi Industri 4.0.

integrasi operasi (OI)

Proses modernisasi operasi di cloud, yang melibatkan perencanaan kesiapan, otomatisasi, dan integrasi. Untuk informasi selengkapnya, lihat panduan integrasi operasi.

jejak organisasi

Jejak yang dibuat oleh AWS CloudTrail itu mencatat semua peristiwa untuk semua Akun AWS dalam organisasi di AWS Organizations. Jejak ini dibuat di setiap Akun AWS bagian organisasi dan melacak aktivitas di setiap akun. Untuk informasi selengkapnya, lihat Membuat jejak untuk organisasi dalam CloudTrail dokumentasi.

manajemen perubahan organisasi (OCM)

Kerangka kerja untuk mengelola transformasi bisnis utama yang mengganggu dari perspektif orang, budaya, dan kepemimpinan. OCM membantu organisasi mempersiapkan, dan transisi ke, sistem dan strategi baru dengan mempercepat adopsi perubahan, mengatasi masalah transisi, dan mendorong perubahan budaya dan organisasi. Dalam strategi AWS migrasi, kerangka kerja ini disebut percepatan orang, karena kecepatan perubahan yang diperlukan dalam proyek adopsi cloud. Untuk informasi lebih lanjut, lihat panduan OCM.

kontrol akses asal (OAC)

Di CloudFront, opsi yang disempurnakan untuk membatasi akses untuk mengamankan konten Amazon Simple Storage Service (Amazon S3) Anda. OAC mendukung semua bucket S3 di semua Wilayah AWS, enkripsi sisi server dengan AWS KMS (SSE-KMS), dan dinamis dan permintaan ke bucket S3. `PUT` `DELETE`

identitas akses asal (OAI)

Di CloudFront, opsi untuk membatasi akses untuk mengamankan konten Amazon S3 Anda. Saat Anda menggunakan OAI, CloudFront buat prinsipal yang dapat diautentikasi oleh Amazon S3. Prinsipal yang diautentikasi dapat mengakses konten dalam bucket S3 hanya melalui distribusi tertentu. CloudFront Lihat juga OAC, yang menyediakan kontrol akses yang lebih terperinci dan ditingkatkan.

ORR

Lihat tinjauan kesiapan operasional.

OT

Lihat teknologi operasional.

keluar (jalan keluar) VPC

Dalam arsitektur AWS multi-akun, VPC yang menangani koneksi jaringan yang dimulai dari dalam aplikasi. Arsitektur Referensi AWS Keamanan merekomendasikan pengaturan akun Jaringan Anda dengan inbound, outbound, dan inspeksi VPCs untuk melindungi antarmuka dua arah antara aplikasi Anda dan internet yang lebih luas.

P

batas izin

Kebijakan manajemen IAM yang dilampirkan pada prinsipal IAM untuk menetapkan izin maksimum yang dapat dimiliki pengguna atau peran. Untuk informasi selengkapnya, lihat Batas izin dalam dokumentasi IAM.

Informasi Identifikasi Pribadi (PII)

Informasi yang, jika dilihat secara langsung atau dipasangkan dengan data terkait lainnya, dapat digunakan untuk menyimpulkan identitas individu secara wajar. Contoh PII termasuk nama, alamat, dan informasi kontak.

PII

Lihat informasi yang dapat diidentifikasi secara pribadi.

buku pedoman

Serangkaian langkah yang telah ditentukan sebelumnya yang menangkap pekerjaan yang terkait dengan migrasi, seperti mengirimkan fungsi operasi inti di cloud. Buku pedoman dapat berupa skrip, runbook otomatis, atau ringkasan proses atau langkah-langkah yang diperlukan untuk mengoperasikan lingkungan modern Anda.

PLC

Lihat pengontrol logika yang dapat diprogram.

PLM

Lihat manajemen siklus hidup produk.

kebijakan

Objek yang dapat menentukan izin (lihat kebijakan berbasis identitas), menentukan kondisi akses (lihat kebijakan berbasis sumber daya), atau menentukan izin maksimum untuk semua akun di organisasi (lihat kebijakan kontrol layanan). AWS Organizations

ketekunan poliglot

Secara independen memilih teknologi penyimpanan data microservice berdasarkan pola akses data dan persyaratan lainnya. Jika layanan mikro Anda memiliki teknologi penyimpanan data yang sama, mereka dapat menghadapi tantangan implementasi atau mengalami kinerja yang buruk. Layanan mikro lebih mudah diimplementasikan dan mencapai kinerja dan skalabilitas yang lebih baik jika mereka menggunakan penyimpanan data yang paling sesuai dengan kebutuhan mereka. Untuk informasi selengkapnya, lihat Mengaktifkan persistensi data di layanan mikro.

penilaian portofolio

Proses menemukan, menganalisis, dan memprioritaskan portofolio aplikasi untuk merencanakan migrasi. Untuk informasi selengkapnya, lihat Mengevaluasi kesiapan migrasi.

predikat

Kondisi kueri yang mengembalikan `true` atau`false`, biasanya terletak di `WHERE` klausa.

predikat pushdown

Teknik optimasi kueri database yang menyaring data dalam kueri sebelum transfer. Ini mengurangi jumlah data yang harus diambil dan diproses dari database relasional, dan meningkatkan kinerja kueri.

kontrol preventif

Kontrol keamanan yang dirancang untuk mencegah suatu peristiwa terjadi. Kontrol ini adalah garis pertahanan pertama untuk membantu mencegah akses tidak sah atau perubahan yang tidak diinginkan ke jaringan Anda. Untuk informasi selengkapnya, lihat Kontrol pencegahan dalam Menerapkan kontrol keamanan pada. AWS

principal

Entitas AWS yang dapat melakukan tindakan dan mengakses sumber daya. Entitas ini biasanya merupakan pengguna root untuk Akun AWS, peran IAM, atau pengguna. Untuk informasi selengkapnya, lihat Prinsip dalam istilah dan konsep Peran dalam dokumentasi IAM.

privasi berdasarkan desain

Pendekatan rekayasa sistem yang memperhitungkan privasi melalui seluruh proses pengembangan.

zona yang dihosting pribadi

Container yang menyimpan informasi tentang bagaimana Anda ingin Amazon Route 53 merespons kueri DNS untuk domain dan subdomainnya dalam satu atau lebih. VPCs Untuk informasi selengkapnya, lihat Bekerja dengan zona yang dihosting pribadi di dokumentasi Route 53.

kontrol proaktif

Kontrol keamanan yang dirancang untuk mencegah penyebaran sumber daya yang tidak sesuai. Kontrol ini memindai sumber daya sebelum disediakan. Jika sumber daya tidak sesuai dengan kontrol, maka itu tidak disediakan. Untuk informasi selengkapnya, lihat panduan referensi Kontrol dalam AWS Control Tower dokumentasi dan lihat Kontrol proaktif dalam Menerapkan kontrol keamanan pada AWS.

manajemen siklus hidup produk (PLM)

Manajemen data dan proses untuk suatu produk di seluruh siklus hidupnya, mulai dari desain, pengembangan, dan peluncuran, melalui pertumbuhan dan kematangan, hingga penurunan dan penghapusan.

lingkungan produksi

Lihat lingkungan.

pengontrol logika yang dapat diprogram (PLC)

Di bidang manufaktur, komputer yang sangat andal dan mudah beradaptasi yang memantau mesin dan mengotomatiskan proses manufaktur.

rantai cepat

Menggunakan output dari satu prompt LLM sebagai input untuk prompt berikutnya untuk menghasilkan respons yang lebih baik. Teknik ini digunakan untuk memecah tugas yang kompleks menjadi subtugas, atau untuk secara iteratif memperbaiki atau memperluas respons awal. Ini membantu meningkatkan akurasi dan relevansi respons model dan memungkinkan hasil yang lebih terperinci dan dipersonalisasi.

pseudonimisasi

Proses penggantian pengenal pribadi dalam kumpulan data dengan nilai placeholder. Pseudonimisasi dapat membantu melindungi privasi pribadi. Data pseudonim masih dianggap sebagai data pribadi.

publish/subscribe (pub/sub)

Pola yang memungkinkan komunikasi asinkron antara layanan mikro untuk meningkatkan skalabilitas dan daya tanggap. Misalnya, dalam MES berbasis layanan mikro, layanan mikro dapat mempublikasikan pesan peristiwa ke saluran yang dapat berlangganan layanan mikro lainnya. Sistem dapat menambahkan layanan mikro baru tanpa mengubah layanan penerbitan.

Q

rencana kueri

Serangkaian langkah, seperti instruksi, yang digunakan untuk mengakses data dalam sistem database relasional SQL.

regresi rencana kueri

Ketika pengoptimal layanan database memilih rencana yang kurang optimal daripada sebelum perubahan yang diberikan ke lingkungan database. Hal ini dapat disebabkan oleh perubahan statistik, kendala, pengaturan lingkungan, pengikatan parameter kueri, dan pembaruan ke mesin database.

R

Matriks RACI

Lihat bertanggung jawab, akuntabel, dikonsultasikan, diinformasikan (RACI).

LAP

Lihat Retrieval Augmented Generation.

ransomware

Perangkat lunak berbahaya yang dirancang untuk memblokir akses ke sistem komputer atau data sampai pembayaran dilakukan.

Matriks RASCI

Lihat bertanggung jawab, akuntabel, dikonsultasikan, diinformasikan (RACI).

RCAC

Lihat kontrol akses baris dan kolom.

replika baca

Salinan database yang digunakan untuk tujuan read-only. Anda dapat merutekan kueri ke replika baca untuk mengurangi beban pada database utama Anda.

arsitek ulang

Lihat 7 Rs.

tujuan titik pemulihan (RPO)

Jumlah waktu maksimum yang dapat diterima sejak titik pemulihan data terakhir. Ini menentukan apa yang dianggap sebagai kehilangan data yang dapat diterima antara titik pemulihan terakhir dan gangguan layanan.

tujuan waktu pemulihan (RTO)

Penundaan maksimum yang dapat diterima antara gangguan layanan dan pemulihan layanan.

refactor

Lihat 7 Rs.

Wilayah

Kumpulan AWS sumber daya di wilayah geografis. Masing-masing Wilayah AWS terisolasi dan independen dari yang lain untuk memberikan toleransi kesalahan, stabilitas, dan ketahanan. Untuk informasi selengkapnya, lihat Menentukan Wilayah AWS akun yang dapat digunakan.

regresi

Teknik ML yang memprediksi nilai numerik. Misalnya, untuk memecahkan masalah “Berapa harga rumah ini akan dijual?” Model ML dapat menggunakan model regresi linier untuk memprediksi harga jual rumah berdasarkan fakta yang diketahui tentang rumah (misalnya, luas persegi).

rehost

Lihat 7 Rs.

melepaskan

Dalam proses penyebaran, tindakan mempromosikan perubahan pada lingkungan produksi.

memindahkan

Lihat 7 Rs.

memplatform ulang

Lihat 7 Rs.

pembelian kembali

Lihat 7 Rs.

ketahanan

Kemampuan aplikasi untuk melawan atau pulih dari gangguan. Ketersediaan tinggi dan pemulihan bencana adalah pertimbangan umum ketika merencanakan ketahanan di. AWS Cloud Untuk informasi lebih lanjut, lihat AWS Cloud Ketahanan.

kebijakan berbasis sumber daya

Kebijakan yang dilampirkan ke sumber daya, seperti bucket Amazon S3, titik akhir, atau kunci enkripsi. Jenis kebijakan ini menentukan prinsipal mana yang diizinkan mengakses, tindakan yang didukung, dan kondisi lain yang harus dipenuhi.

matriks yang bertanggung jawab, akuntabel, dikonsultasikan, diinformasikan (RACI)

Matriks yang mendefinisikan peran dan tanggung jawab untuk semua pihak yang terlibat dalam kegiatan migrasi dan operasi cloud. Nama matriks berasal dari jenis tanggung jawab yang

didefinisikan dalam matriks: bertanggung jawab (R), akuntabel (A), dikonsultasikan (C), dan diinformasikan (I). Tipe dukungan (S) adalah opsional. Jika Anda menyertakan dukungan, matriks disebut matriks RASCI, dan jika Anda mengecualikannya, itu disebut matriks RACI.

kontrol responsif

Kontrol keamanan yang dirancang untuk mendorong remediasi efek samping atau penyimpangan dari garis dasar keamanan Anda. Untuk informasi selengkapnya, lihat Kontrol responsif dalam Menerapkan kontrol keamanan pada AWS.

melestarikan

Lihat 7 Rs.

pensiun

Lihat 7 Rs.

Retrieval Augmented Generation (RAG)

Teknologi AI generatif di mana LLM merujuk sumber data otoritatif yang berada di luar sumber data pelatihannya sebelum menghasilkan respons. Misalnya, model RAG mungkin melakukan pencarian semantik dari basis pengetahuan organisasi atau data kustom. Untuk informasi lebih lanjut, lihat Apa itu RAG.

rotasi

Proses memperbarui rahasia secara berkala untuk membuatnya lebih sulit bagi penyerang untuk mengakses kredensil.

kontrol akses baris dan kolom (RCAC)

Penggunaan ekspresi SQL dasar dan fleksibel yang telah menetapkan aturan akses. RCAC terdiri dari izin baris dan topeng kolom.

RPO

Lihat tujuan titik pemulihan.

RTO

Lihat tujuan waktu pemulihan.

buku runbook

Satu set prosedur manual atau otomatis yang diperlukan untuk melakukan tugas tertentu. Ini biasanya dibangun untuk merampingkan operasi berulang atau prosedur dengan tingkat kesalahan yang tinggi.

D

SAML 2.0

Standar terbuka yang digunakan oleh banyak penyedia identitas (IdPs). Fitur ini memungkinkan sistem masuk tunggal gabungan (SSO), sehingga pengguna dapat masuk ke AWS Management Console atau memanggil operasi AWS API tanpa Anda harus membuat pengguna di IAM untuk semua orang di organisasi Anda. Untuk informasi lebih lanjut tentang federasi berbasis SAMP 2.0, lihat Tentang federasi berbasis SAMP 2.0 dalam dokumentasi IAM.

SCADA

Lihat kontrol pengawasan dan akuisisi data.

SCP

Lihat kebijakan kontrol layanan.

Rahasia

Dalam AWS Secrets Manager, informasi rahasia atau terbatas, seperti kata sandi atau kredensil pengguna, yang Anda simpan dalam bentuk terenkripsi. Ini terdiri dari nilai rahasia dan metadata-nya. Nilai rahasia dapat berupa biner, string tunggal, atau beberapa string. Untuk informasi selengkapnya, lihat Apa yang ada di rahasia Secrets Manager? dalam dokumentasi Secrets Manager.

keamanan dengan desain

Pendekatan rekayasa sistem yang memperhitungkan keamanan melalui seluruh proses pengembangan.

kontrol keamanan

Pagar pembatas teknis atau administratif yang mencegah, mendeteksi, atau mengurangi kemampuan pelaku ancaman untuk mengeksploitasi kerentanan keamanan. Ada empat jenis kontrol keamanan utama: preventif, detektif, responsif, dan proaktif.

pengerasan keamanan

Proses mengurangi permukaan serangan untuk membuatnya lebih tahan terhadap serangan. Ini dapat mencakup tindakan seperti menghapus sumber daya yang tidak lagi diperlukan, menerapkan praktik keamanan terbaik untuk memberikan hak istimewa paling sedikit, atau menonaktifkan fitur yang tidak perlu dalam file konfigurasi.

sistem informasi keamanan dan manajemen acara (SIEM)

Alat dan layanan yang menggabungkan sistem manajemen informasi keamanan (SIM) dan manajemen acara keamanan (SEM). Sistem SIEM mengumpulkan, memantau, dan menganalisis data dari server, jaringan, perangkat, dan sumber lain untuk mendeteksi ancaman dan pelanggaran keamanan, dan untuk menghasilkan peringatan.

otomatisasi respons keamanan

Tindakan yang telah ditentukan dan diprogram yang dirancang untuk secara otomatis merespons atau memulihkan peristiwa keamanan. Otomatisasi ini berfungsi sebagai kontrol keamanan detektif atau responsif yang membantu Anda menerapkan praktik terbaik AWS keamanan. Contoh tindakan respons otomatis termasuk memodifikasi grup keamanan VPC, menambal instans EC2 Amazon, atau memutar kredensil.

enkripsi sisi server

Enkripsi data di tujuannya, oleh Layanan AWS yang menerimanya.

kebijakan kontrol layanan (SCP)

Kebijakan yang menyediakan kontrol terpusat atas izin untuk semua akun di organisasi. AWS Organizations SCPs menentukan pagar pembatas atau menetapkan batasan pada tindakan yang dapat didelegasikan oleh administrator kepada pengguna atau peran. Anda dapat menggunakan SCPs daftar izin atau daftar penolakan, untuk menentukan layanan atau tindakan mana yang diizinkan atau dilarang. Untuk informasi selengkapnya, lihat Kebijakan kontrol layanan dalam AWS Organizations dokumentasi.

titik akhir layanan

URL titik masuk untuk file Layanan AWS. Anda dapat menggunakan endpoint untuk terhubung secara terprogram ke layanan target. Untuk informasi selengkapnya, lihat Layanan AWS titik akhir di Referensi Umum AWS.

perjanjian tingkat layanan (SLA)

Perjanjian yang menjelaskan apa yang dijanjikan tim TI untuk diberikan kepada pelanggan mereka, seperti uptime dan kinerja layanan.

indikator tingkat layanan (SLI)

Pengukuran aspek kinerja layanan, seperti tingkat kesalahan, ketersediaan, atau throughputnya.

tujuan tingkat layanan (SLO)

Metrik target yang mewakili kesehatan layanan, yang diukur dengan indikator tingkat layanan.

model tanggung jawab bersama

Model yang menjelaskan tanggung jawab yang Anda bagikan AWS untuk keamanan dan kepatuhan cloud. AWS bertanggung jawab atas keamanan cloud, sedangkan Anda bertanggung jawab atas keamanan di cloud. Untuk informasi selengkapnya, lihat Model tanggung jawab bersama.

SIEM

Lihat informasi keamanan dan sistem manajemen acara.

titik kegagalan tunggal (SPOF)

Kegagalan dalam satu komponen penting dari aplikasi yang dapat mengganggu sistem.

SLA

Lihat perjanjian tingkat layanan.

SLI

Lihat indikator tingkat layanan.

SLO

Lihat tujuan tingkat layanan.

split-and-seed model

Pola untuk menskalakan dan mempercepat proyek modernisasi. Ketika fitur baru dan rilis produk didefinisikan, tim inti berpisah untuk membuat tim produk baru. Ini membantu meningkatkan kemampuan dan layanan organisasi Anda, meningkatkan produktivitas pengembang, dan

mendukung inovasi yang cepat. Untuk informasi lebih lanjut, lihat Pendekatan bertahap untuk memodernisasi aplikasi di. AWS Cloud

SPOF

Lihat satu titik kegagalan.

skema bintang

Struktur organisasi database yang menggunakan satu tabel fakta besar untuk menyimpan data transaksional atau terukur dan menggunakan satu atau lebih tabel dimensi yang lebih kecil untuk menyimpan atribut data. Struktur ini dirancang untuk digunakan dalam gudang data atau untuk tujuan intelijen bisnis.

pola ara pencekik

Pendekatan untuk memodernisasi sistem monolitik dengan menulis ulang secara bertahap dan mengganti fungsionalitas sistem sampai sistem warisan dapat dinonaktifkan. Pola ini menggunakan analogi pohon ara yang tumbuh menjadi pohon yang sudah mapan dan akhirnya mengatasi dan menggantikan inangnya. Pola ini diperkenalkan oleh Martin Fowler sebagai cara untuk mengelola risiko saat menulis ulang sistem monolitik. Untuk contoh cara menerapkan pola ini, lihat Memodernisasi layanan web Microsoft ASP.NET (ASMX) lama secara bertahap menggunakan container dan Amazon API Gateway.

subnet

Rentang alamat IP dalam VPC Anda. Subnet harus berada di Availability Zone tunggal.

kontrol pengawasan dan akuisisi data (SCADA)

Di bidang manufaktur, sistem yang menggunakan perangkat keras dan perangkat lunak untuk memantau aset fisik dan operasi produksi.

enkripsi simetris

Algoritma enkripsi yang menggunakan kunci yang sama untuk mengenkripsi dan mendekripsi data.

pengujian sintetis

Menguji sistem dengan cara yang mensimulasikan interaksi pengguna untuk mendeteksi potensi masalah atau untuk memantau kinerja. Anda dapat menggunakan Amazon CloudWatch Synthetics untuk membuat tes ini.

sistem prompt

Teknik untuk memberikan konteks, instruksi, atau pedoman ke LLM untuk mengarahkan perilakunya. Permintaan sistem membantu mengatur konteks dan menetapkan aturan untuk interaksi dengan pengguna.

T

tag

Pasangan nilai kunci yang bertindak sebagai metadata untuk mengatur sumber daya Anda. AWS Tanda dapat membantu Anda mengelola, mengidentifikasi, mengatur, dan memfilter sumber daya. Untuk informasi selengkapnya, lihat Menandai AWS sumber daya Anda.

variabel target

Nilai yang Anda coba prediksi dalam ML yang diawasi. Ini juga disebut sebagai variabel hasil. Misalnya, dalam pengaturan manufaktur, variabel target bisa menjadi cacat produk.

daftar tugas

Alat yang digunakan untuk melacak kemajuan melalui runbook. Daftar tugas berisi ikhtisar runbook dan daftar tugas umum yang harus diselesaikan. Untuk setiap tugas umum, itu termasuk perkiraan jumlah waktu yang dibutuhkan, pemilik, dan kemajuan.

lingkungan uji

Lihat lingkungan.

pelatihan

Untuk menyediakan data bagi model ML Anda untuk dipelajari. Data pelatihan harus berisi jawaban yang benar. Algoritma pembelajaran menemukan pola dalam data pelatihan yang memetakan atribut data input ke target (jawaban yang ingin Anda prediksi). Ini menghasilkan model ML yang menangkap pola-pola ini. Anda kemudian dapat menggunakan model ML untuk membuat prediksi pada data baru yang Anda tidak tahu targetnya.

gerbang transit

Hub transit jaringan yang dapat Anda gunakan untuk menghubungkan jaringan Anda VPCs dan lokal. Untuk informasi selengkapnya, lihat Apa itu gateway transit dalam AWS Transit Gateway dokumentasi.

alur kerja berbasis batang

Pendekatan di mana pengembang membangun dan menguji fitur secara lokal di cabang fitur dan kemudian menggabungkan perubahan tersebut ke cabang utama. Cabang utama kemudian dibangun untuk pengembangan, praproduksi, dan lingkungan produksi, secara berurutan.

akses tepercaya

Memberikan izin ke layanan yang Anda tentukan untuk melakukan tugas di organisasi Anda di dalam AWS Organizations dan di akunnya atas nama Anda. Layanan tepercaya menciptakan peran terkait layanan di setiap akun, ketika peran itu diperlukan, untuk melakukan tugas manajemen untuk Anda. Untuk informasi selengkapnya, lihat Menggunakan AWS Organizations dengan AWS layanan lain dalam AWS Organizations dokumentasi.

penyetelan

Untuk mengubah aspek proses pelatihan Anda untuk meningkatkan akurasi model ML. Misalnya, Anda dapat melatih model ML dengan membuat set pelabelan, menambahkan label, dan kemudian mengulangi langkah-langkah ini beberapa kali di bawah pengaturan yang berbeda untuk mengoptimalkan model.

tim dua pizza

Sebuah DevOps tim kecil yang bisa Anda beri makan dengan dua pizza. Ukuran tim dua pizza memastikan peluang terbaik untuk berkolaborasi dalam pengembangan perangkat lunak.

U

waswas

Sebuah konsep yang mengacu pada informasi yang tidak tepat, tidak lengkap, atau tidak diketahui yang dapat merusak keandalan model ML prediktif. Ada dua jenis ketidakpastian: ketidakpastian epistemik disebabkan oleh data yang terbatas dan tidak lengkap, sedangkan ketidakpastian aleatorik disebabkan oleh kebisingan dan keacakan yang melekat dalam data. Untuk informasi lebih lanjut, lihat panduan Mengukur ketidakpastian dalam sistem pembelajaran mendalam.

tugas yang tidak terdiferensiasi

Juga dikenal sebagai angkat berat, pekerjaan yang diperlukan untuk membuat dan mengoperasikan aplikasi tetapi itu tidak memberikan nilai langsung kepada pengguna akhir atau

memberikan keunggulan kompetitif. Contoh tugas yang tidak terdiferensiasi termasuk pengadaan, pemeliharaan, dan perencanaan kapasitas.

lingkungan atas

Lihat lingkungan.

V

menyedot debu

Operasi pemeliharaan database yang melibatkan pembersihan setelah pembaruan tambahan untuk merebut kembali penyimpanan dan meningkatkan kinerja.

kendali versi

Proses dan alat yang melacak perubahan, seperti perubahan kode sumber dalam repositori.

Peering VPC

Koneksi antara dua VPCs yang memungkinkan Anda untuk merutekan lalu lintas dengan menggunakan alamat IP pribadi. Untuk informasi selengkapnya, lihat Apa itu peering VPC di dokumentasi VPC Amazon.

kerentanan

Kelemahan perangkat lunak atau perangkat keras yang membahayakan keamanan sistem.

W

cache hangat

Cache buffer yang berisi data saat ini dan relevan yang sering diakses. Instance database dapat membaca dari cache buffer, yang lebih cepat daripada membaca dari memori utama atau disk.

data hangat

Data yang jarang diakses. Saat menanyakan jenis data ini, kueri yang cukup lambat biasanya dapat diterima.

fungsi jendela

Fungsi SQL yang melakukan perhitungan pada sekelompok baris yang berhubungan dengan catatan saat ini. Fungsi jendela berguna untuk memproses tugas, seperti menghitung rata-rata bergerak atau mengakses nilai baris berdasarkan posisi relatif dari baris saat ini.

beban kerja

Kumpulan sumber daya dan kode yang memberikan nilai bisnis, seperti aplikasi yang dihadapi pelanggan atau proses backend.

aliran kerja

Grup fungsional dalam proyek migrasi yang bertanggung jawab atas serangkaian tugas tertentu. Setiap alur kerja independen tetapi mendukung alur kerja lain dalam proyek. Misalnya, alur kerja portofolio bertanggung jawab untuk memprioritaskan aplikasi, perencanaan gelombang, dan mengumpulkan metadata migrasi. Alur kerja portofolio mengirimkan aset ini ke alur kerja migrasi, yang kemudian memigrasikan server dan aplikasi.

CACING

Lihat menulis sekali, baca banyak.

WQF

Lihat AWS Kerangka Kualifikasi Beban Kerja.

tulis sekali, baca banyak (WORM)

Model penyimpanan yang menulis data satu kali dan mencegah data dihapus atau dimodifikasi. Pengguna yang berwenang dapat membaca data sebanyak yang diperlukan, tetapi mereka tidak dapat mengubahnya. Infrastruktur penyimpanan data ini dianggap tidak dapat diubah.

Z

eksploitasi zero-day

Serangan, biasanya malware, yang memanfaatkan kerentanan zero-day.

kerentanan zero-day

Cacat atau kerentanan yang tak tanggung-tanggung dalam sistem produksi. Aktor ancaman dapat menggunakan jenis kerentanan ini untuk menyerang sistem. Pengembang sering menyadari kerentanan sebagai akibat dari serangan tersebut.

bisikan zero-shot

Memberikan LLM dengan instruksi untuk melakukan tugas tetapi tidak ada contoh (tembakan) yang dapat membantu membimbingnya. LLM harus menggunakan pengetahuan pra-terlatih untuk menangani tugas. Efektivitas bidikan nol tergantung pada kompleksitas tugas dan kualitas prompt. Lihat juga beberapa bidikan yang diminta.

aplikasi zombie

Aplikasi yang memiliki CPU rata-rata dan penggunaan memori di bawah 5 persen. Dalam proyek migrasi, adalah umum untuk menghentikan aplikasi ini.

Terjemahan disediakan oleh mesin penerjemah. Jika konten terjemahan yang diberikan bertentangan dengan versi bahasa Inggris aslinya, utamakan versi bahasa Inggris.