Pengambilan opsi dan arsitektur Augmented Generation di AWS

AWS Bimbingan Preskriptif

AWS Bimbingan Preskriptif: Pengambilan opsi dan arsitektur Augmented Generation di AWS

Table of Contents

Pengambilan opsi dan arsitektur Augmented Generation di AWS

Mithil Shah, Rajeev Muralidhar, dan Benteng Natacha, Amazon Web Services

Oktober 2024 (sejarah dokumen)

Generative AI mengacu pada subset model AI yang dapat membuat konten dan artefak baru, seperti gambar, video, teks, dan audio, dari prompt teks sederhana. Model AI generatif dilatih pada sejumlah besar data yang mencakup berbagai subjek dan tugas. Hal ini memungkinkan mereka untuk menunjukkan keserbagunaan yang luar biasa dalam melakukan berbagai tugas, bahkan yang belum dilatih secara eksplisit. Karena kemampuan model tunggal untuk melakukan banyak tugas, model ini sering disebut sebagai model dasar (FMs).

Salah satu aplikasi penting dari model AI generatif adalah kemahiran mereka dalam menjawab pertanyaan. Namun, ada tantangan khusus yang muncul ketika model ini digunakan untuk menjawab pertanyaan berdasarkan dokumen khusus. Dokumen khusus dapat mencakup informasi kepemilikan, situs web internal, dokumentasi internal, Confluence halaman, SharePoint halaman, dan lain-lain. Salah satu opsi adalah menggunakan Retrieval Augmented Generation (RAG). Dengan RAG, model foundation mereferensikan sumber data otoritatif yang berada di luar sumber data pelatihannya (seperti dokumen kustom Anda) sebelum menghasilkan respons.

Panduan ini menjelaskan opsi AI generatif berbeda yang tersedia untuk menjawab pertanyaan dari dokumentasi khusus, termasuk sistem Retrieval Augmented Generation (RAG). Ini juga memberikan gambaran umum tentang membangun sistem RAG di Amazon Web Services (AWS). Dengan meninjau opsi dan arsitektur RAG, Anda dapat memilih antara layanan yang dikelola sepenuhnya AWS dan arsitektur RAG khusus.

Audiens yang dituju

Audiens yang dituju untuk panduan ini adalah arsitek dan manajer AI generatif yang ingin membangun solusi RAG, untuk meninjau arsitektur yang tersedia, dan untuk memahami manfaat dan kerugian dari setiap opsi.

Tujuan

Panduan ini membantu Anda melakukan hal berikut:

- Memahami opsi AI generatif yang tersedia untuk menjawab pertanyaan dari dokumen khusus
- Tinjau opsi arsitektur untuk sistem RAG AWS
- Pahami kelebihan dan kekurangan masing-masing opsi RAG
- Pilih arsitektur RAG untuk lingkungan Anda AWS

Opsi AI generatif untuk menanyakan dokumen khusus

Organizations sering memiliki berbagai sumber data terstruktur dan tidak terstruktur. Panduan ini berfokus pada bagaimana Anda dapat menggunakan AI generatif untuk menjawab pertanyaan dari data yang tidak terstruktur.

Data yang tidak terstruktur dalam organisasi Anda dapat berasal dari berbagai sumber. Ini mungkin PDFs, file teks, wiki internal, dokumen teknis, situs web publik, basis pengetahuan, atau lainnya. Jika Anda menginginkan model dasar yang dapat menjawab pertanyaan tentang data tidak terstruktur, opsi berikut tersedia:

- Latih model yayasan baru dengan menggunakan dokumen khusus Anda dan data pelatihan lainnya
- Sempurnakan model foundation yang ada dengan menggunakan data dari dokumen kustom Anda
- Gunakan pembelajaran dalam konteks untuk meneruskan dokumen ke model yayasan saat Anda mengajukan pertanyaan
- Gunakan pendekatan Retrieval Augmented Generation (RAG)

Melatih model fondasi baru dari awal yang menyertakan data kustom Anda adalah usaha yang ambisius. Beberapa perusahaan telah berhasil melakukannya, seperti Bloomberg dengan mereka BloombergGPTmodel. Contoh lainnya adalah multimodal EXAONEmodel oleh LG AI Research, yang dilatih dengan menggunakan 600 miliar karya seni dan 250 juta gambar resolusi tinggi, disertai dengan teks. Menurut Biaya AI: Haruskah Anda Membangun atau Membeli Model Foundation Anda (LinkedIn), model yang mirip dengan Meta Llama 2 biaya sekitar USD $4,8 juta untuk melatih. Ada dua prasyarat utama untuk melatih model dari awal: akses ke sumber daya (keuangan, teknis, waktu) dan pengembalian investasi yang jelas. Jika ini tampaknya tidak cocok, maka opsi selanjutnya adalah menyempurnakan model pondasi yang ada.

Menyesuaikan model yang ada melibatkan pengambilan model, seperti Amazon Titan, Mistral, atau model Llama, dan kemudian mengadaptasi model ke data kustom Anda. Ada berbagai teknik untuk fine-tuning, yang sebagian besar melibatkan memodifikasi hanya beberapa parameter alih-alih memodifikasi semua parameter dalam model. Ini disebut fine-tuning parameter-efisien. Ada dua metode utama untuk fine-tuning:

- Penyetelan halus yang diawasi menggunakan data berlabel dan membantu Anda melatih model untuk jenis tugas baru. Misalnya, jika Anda ingin membuat laporan berdasarkan formulir PDF,

maka Anda mungkin harus mengajarkan model bagaimana melakukannya dengan memberikan contoh yang cukup.

- Penyetelan halus tanpa pengawasan adalah agnostik tugas dan menyesuaikan model fondasi dengan data Anda sendiri. Ini melatih model untuk memahami konteks dokumen Anda. Model fine-tuned kemudian membuat konten, seperti laporan, dengan menggunakan gaya yang lebih kustom organisasi Anda.

Namun, fine-tuning mungkin tidak ideal untuk kasus penggunaan tanya jawab. Untuk informasi selengkapnya, lihat Membandingkan RAG dan fine-tuning dalam panduan ini.

Ketika Anda mengajukan pertanyaan, Anda dapat meneruskan dokumen model dasar dan menggunakan pembelajaran dalam konteks model untuk mengembalikan jawaban dari dokumen. Opsi ini cocok untuk kueri ad-hoc dari satu dokumen. Namun, solusi ini tidak berfungsi dengan baik untuk menanyakan beberapa dokumen atau untuk sistem kueri dan aplikasi, seperti Microsoft SharePoint atau Atlassian Confluence.

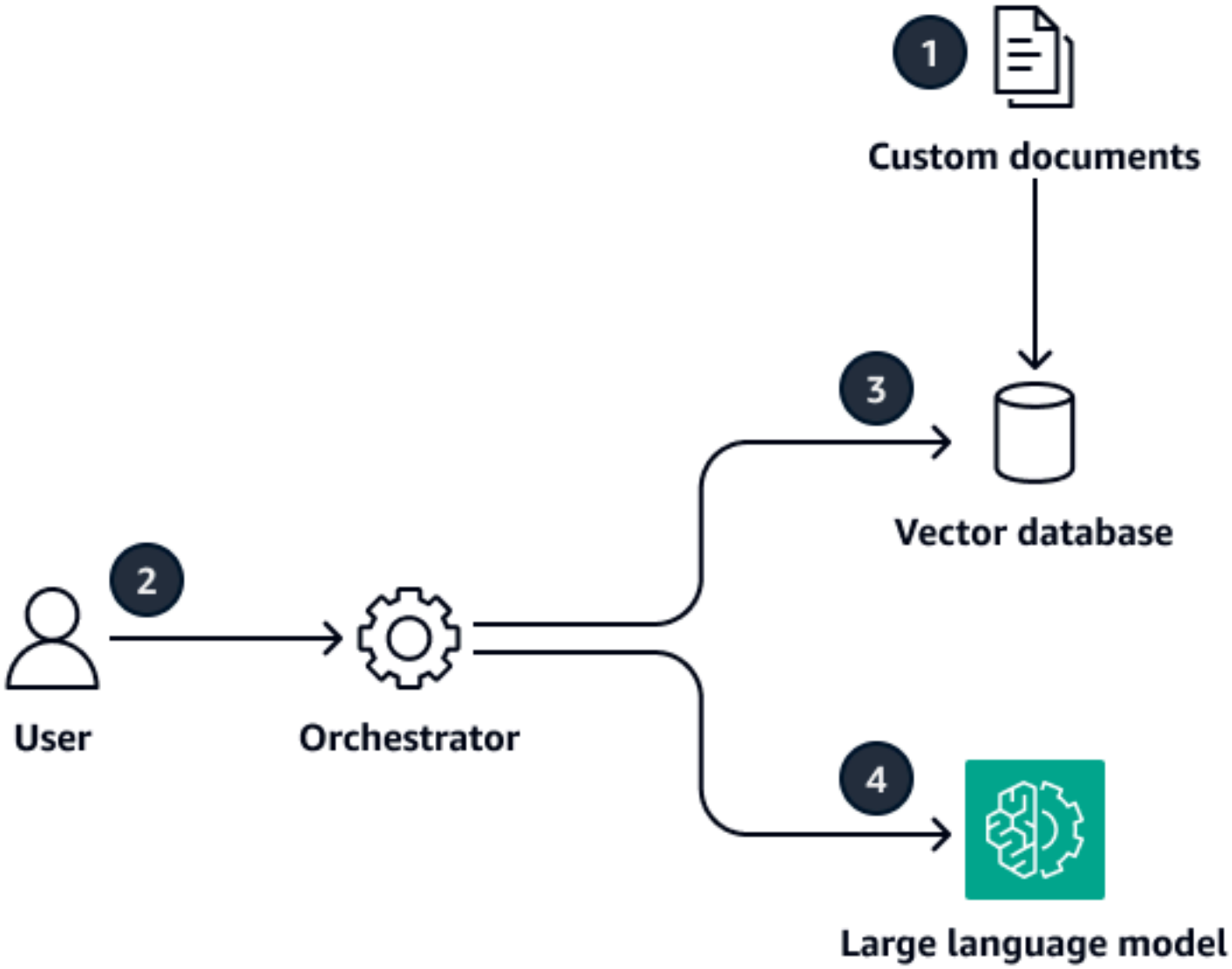
Opsi terakhir adalah menggunakan RAG. Dengan RAG, model foundation mereferensikan dokumen kustom Anda sebelum menghasilkan respons. RAG memperluas kemampuan model ke basis pengetahuan internal organisasi Anda, semua tanpa perlu melatih kembali model. Ini adalah pendekatan hemat biaya untuk meningkatkan output model sehingga tetap relevan, akurat, dan berguna dalam berbagai konteks.

Topik di bagian ini:

- Memahami Generasi Augmented Retrieval
- Membandingkan Retrieval Augmented Generation dan fine-tuning
- Kasus penggunaan untuk Retrieval Augmented Generation

Memahami Generasi Augmented Retrieval

Retrieval Augmented Generation (RAG) adalah teknik yang digunakan untuk menambah model bahasa besar (LLM) dengan data eksternal, seperti dokumen internal perusahaan. Ini memberikan model dengan konteks yang dibutuhkan untuk menghasilkan output yang akurat dan berguna untuk kasus penggunaan spesifik Anda. RAG adalah pendekatan pragmatis dan efektif untuk digunakan LLMs dalam suatu perusahaan. Diagram berikut menunjukkan gambaran tingkat tinggi tentang cara kerja pendekatan RAG.



Secara garis besar, proses RAG adalah empat langkah. Langkah pertama dilakukan sekali, dan tiga langkah lainnya dilakukan sebanyak yang diperlukan:

1. Anda membuat embeddings untuk menelan dokumen internal ke dalam database vektor. Embeddings adalah representasi numerik teks dalam dokumen yang menangkap makna semantik atau kontekstual dari data. Database vektor pada dasarnya adalah database dari embeddings ini, dan kadang-kadang disebut penyimpanan vektor atau indeks vektor. Langkah ini membutuhkan pembersihan data, pemformatan, dan chunking, tetapi ini adalah aktivitas satu kali di muka.
2. Seorang manusia mengirimkan kueri dalam bahasa alami.
3. Orkestrator melakukan pencarian kesamaan dalam database vektor dan mengambil data yang relevan. Orkestrator menambahkan data yang diambil (juga dikenal sebagai konteks) ke prompt yang berisi kueri.

4. Orkestrator mengirimkan kueri dan konteks ke LLM. LLM menghasilkan respons terhadap kueri dengan menggunakan konteks tambahan.

Dari perspektif pengguna, RAG terlihat seperti berinteraksi dengan LLM apa pun. Namun, sistem tahu lebih banyak tentang konten yang dimaksud dan memberikan jawaban yang disesuaikan dengan basis pengetahuan organisasi.

Untuk informasi selengkapnya tentang cara kerja pendekatan RAG, lihat Apa itu RAG di situs web. AWS

Komponen sistem RAG tingkat produksi

Membangun sistem RAG tingkat produksi membutuhkan pemikiran melalui beberapa aspek yang berbeda dari alur kerja RAG. Secara konseptual, alur kerja RAG tingkat produksi memerlukan kemampuan dan komponen berikut, terlepas dari implementasi spesifiknya:

- Konektor — Ini menghubungkan sumber data perusahaan yang berbeda dengan database vektor. Contoh sumber data terstruktur termasuk database transaksional dan analitis. Contoh sumber data yang tidak terstruktur termasuk penyimpanan objek, basis kode, dan platform perangkat lunak sebagai layanan (SaaS). Setiap sumber data mungkin memerlukan pola konektivitas, lisensi, dan konfigurasi yang berbeda.
- Pemrosesan data — Data datang dalam berbagai bentuk dan bentuk, seperti PDFs, gambar yang dipindai, dokumen, presentasi, dan Microsoft SharePoint berkas. Anda harus menggunakan teknik pemrosesan data untuk mengekstrak, memproses, dan menyiapkan data untuk pengindeksan.
- Embeddings — Untuk melakukan pencarian relevansi, Anda harus mengonversi dokumen dan kueri pengguna Anda ke dalam format yang kompatibel. Dengan menggunakan menyematkan model bahasa, Anda mengonversi dokumen menjadi representasi numerik. Ini pada dasarnya adalah input untuk model pondasi yang mendasarinya.
- Database vektor — Database vektor adalah indeks dari embeddings, teks terkait, dan metadata. Indeks dioptimalkan untuk pencarian dan pengambilan.
- Retriever — Untuk kueri pengguna, retriever mengambil konteks yang relevan dari database vektor dan memberi peringkat tanggapan berdasarkan persyaratan bisnis.
- Model pondasi — Model dasar untuk sistem RAG biasanya LLM. Dengan memproses konteks dan prompt, model pondasi menghasilkan dan memformat respons bagi pengguna.
- Pagar pembatas dirancang untuk memastikan bahwa kueri, konteks yang cepat, diambil, dan respons LLM akurat, bertanggung jawab, etis, dan bebas dari halusinasi dan bias.

- Orkestrator - Orkestrator bertanggung jawab untuk menjadwalkan dan mengelola alur kerja. end-to-end
- Pengalaman pengguna — Biasanya, pengguna berinteraksi dengan antarmuka obrolan percakapan yang memiliki fitur yang kaya, termasuk menampilkan riwayat obrolan dan mengumpulkan umpan balik pengguna tentang tanggapan.
- Identitas dan manajemen pengguna - Sangat penting untuk mengontrol akses pengguna ke aplikasi dengan perincian yang baik. Dalam AWS Cloud, kebijakan, peran, dan izin biasanya dikelola melalui AWS Identity and Access Management (IAM).

Jelas, ada sejumlah besar pekerjaan untuk merencanakan, mengembangkan, merilis, dan mengelola sistem RAG. Layanan yang dikelola sepenuhnya, seperti Amazon Bedrock atau Amazon Q Business, dapat membantu Anda mengelola beberapa angkat berat yang tidak berdiferensiasi. Namun, arsitektur RAG kustom dapat memberikan kontrol lebih besar atas komponen, seperti retriever atau database vektor.

Membandingkan Retrieval Augmented Generation dan fine-tuning

Tabel berikut menjelaskan keuntungan dan kerugian dari pendekatan fine-tuning dan berbasis RAG.

Pendekatan	Keuntungan	Kekurangan
Penyetelan halus	• Jika model yang disetel dengan baik dilatih menggunakan pendekatan tanpa pengawasan, maka ia dapat membuat konten yang lebih cocok dengan gaya organisasi Anda. • Model yang disetel dengan baik yang dilatih tentang data kepemilikan atau peraturan dapat membantu organisasi Anda mengikuti data internal atau khusus industri serta standar kepatuhan.	• Fine-tuning dapat memakan waktu beberapa jam hingga berhari-hari, tergantung pada ukuran model. Oleh karena itu, ini bukan solusi yang baik jika dokumen khusus Anda sering berubah. • Fine-tuning membutuhk an pemahaman tentang teknik, seperti adaptasi peringkat rendah (LoRa) dan fine-tuning hemat parameter (PEFT). Fine-tuni

Pendekatan	Keuntungan	Kekurangan
		ng mungkin membutuhkan ilmuwan data. • Fine-tuning mungkin tidak tersedia untuk semua model. • Model yang disetel dengan baik tidak memberikan referensi ke sumber dalam tanggapannya. • Mungkin ada peningkat an risiko halusinasi saat menggunakan model yang disetel dengan baik untuk menjawab pertanyaan.

Pendekatan	Keuntungan	Kekurangan
LAP	• RAG memungkinkan Anda membangun sistem penjawab pertanyaan untuk dokumen kustom Anda tanpa fine-tuning. • RAG dapat memasukkan dokumen terbaru dalam beberapa menit. • AWS menawarkan solusi RAG yang dikelola sepenuhnya. Oleh karena itu, tidak diperlukan ilmuwan data atau pengetahuan khusus tentang pembelajaran mesin. • Dalam tanggapannya, model RAG memberikan referensi ke sumber informasi. • Karena RAG menggunakan konteks dari pencarian vektor sebagai dasar jawaban yang dihasilkan, ada penurunan risiko halusinasi.	• RAG tidak berfungsi dengan baik saat merangkum informasi dari seluruh dokumen.

Jika Anda perlu membangun solusi penjawab pertanyaan yang mereferensikan dokumen kustom Anda, maka kami sarankan Anda memulai dari pendekatan berbasis RAG. Gunakan fine-tuning jika Anda membutuhkan model untuk melakukan tugas tambahan, seperti meringkas.

Anda dapat menggabungkan pendekatan fine-tuning dan RAG dalam satu model. Dalam kasus ini, arsitektur RAG tidak berubah, tetapi LLM yang menghasilkan jawaban juga disesuaikan dengan dokumen khusus. Ini menggabungkan yang terbaik dari kedua dunia, dan ini mungkin solusi optimal

untuk kasus penggunaan Anda. Untuk informasi lebih lanjut tentang cara menggabungkan fine-tuning yang diawasi dengan RAG, lihat penelitian RAFT: Adapting Language Model to Domain Specific RAG dari University of California, Berkeley.

Kasus penggunaan untuk Retrieval Augmented Generation

Berikut ini adalah kasus penggunaan umum untuk menggunakan pendekatan RAG:

- Mesin pencari — Mesin pencari yang mendukung RAG dapat memberikan cuplikan yang lebih akurat dan up-to-date berfitur dalam hasil pencarian mereka.
- Sistem tanya jawab — RAG dapat meningkatkan kualitas tanggapan dalam sistem penjawab pertanyaan. Model berbasis pengambilan menggunakan pencarian kesamaan untuk menemukan bagian atau dokumen yang relevan yang berisi jawabannya. Kemudian, menghasilkan respons yang ringkas dan relevan berdasarkan informasi itu.
- Ritel atau e-commerce — RAG dapat meningkatkan pengalaman pengguna dalam e-commerce dengan memberikan rekomendasi produk yang lebih relevan dan personal. Dengan mengambil dan menggabungkan informasi tentang preferensi pengguna dan detail produk, RAG dapat menghasilkan rekomendasi yang lebih akurat dan bermanfaat bagi pelanggan.
- Industri atau manufaktur - Di bidang manufaktur, RAG membantu Anda mengakses informasi penting dengan cepat, seperti operasi pabrik pabrik. Ini juga dapat membantu proses pengambilan keputusan, pemecahan masalah, dan inovasi organisasi. Untuk produsen yang beroperasi dalam kerangka peraturan yang ketat, RAG dapat dengan cepat mengambil peraturan dan standar kepatuhan yang diperbarui dari sumber internal dan eksternal, seperti dari standar industri atau badan pengatur.
- Perawatan Kesehatan — RAG memiliki potensi dalam industri perawatan kesehatan, di mana akses ke informasi yang akurat dan tepat waktu sangat penting. Dengan mengambil dan menggabungkan pengetahuan medis yang relevan dari sumber eksternal, RAG dapat memberikan tanggapan yang lebih akurat dan sadar konteks dalam aplikasi perawatan kesehatan. Aplikasi semacam itu menambah informasi yang dapat diakses oleh dokter manusia, yang pada akhirnya membuat panggilan dan bukan model.
- Legal — RAG dapat diterapkan secara kuat dalam skenario hukum, seperti merger dan akuisisi, di mana dokumen hukum yang kompleks memberikan konteks untuk pertanyaan. Ini dapat membantu para profesional hukum dengan cepat menavigasi masalah peraturan yang kompleks.

Opsi Retrieval Augmented Generation yang dikelola sepenuhnya pada AWS

Untuk mengelola alur kerja Retrieval Augmented Generation (RAG) AWS, Anda dapat menggunakan pipeline RAG khusus atau menggunakan beberapa kemampuan layanan terkelola penuh yang ditawarkan. AWS Karena mereka mencakup banyak komponen inti dari sistem berbasis RAG, layanan yang dikelola sepenuhnya dapat membantu Anda mengelola beberapa angkat berat yang tidak berdiferensiasi. Namun, layanan ini memberikan lebih sedikit kesempatan untuk penyesuaian.

Konektor Layanan AWS penggunaan yang dikelola sepenuhnya untuk menyerap data dari sumber data eksternal, seperti situs web, Atlassian Confluence, atau Microsoft. SharePoint Sumber data yang didukung berbeda-beda Layanan AWS.

Bagian ini mengeksplorasi opsi terkelola penuh berikut untuk membangun alur kerja RAG pada: AWS

- Basis pengetahuan untuk Amazon Bedrock
- Amazon Q Bisnis
- Kanvas Amazon SageMaker AI

Untuk informasi selengkapnya tentang cara memilih di antara opsi-opsi ini, lihat Memilih opsi Retrieval Augmented Generation di AWS di panduan ini.

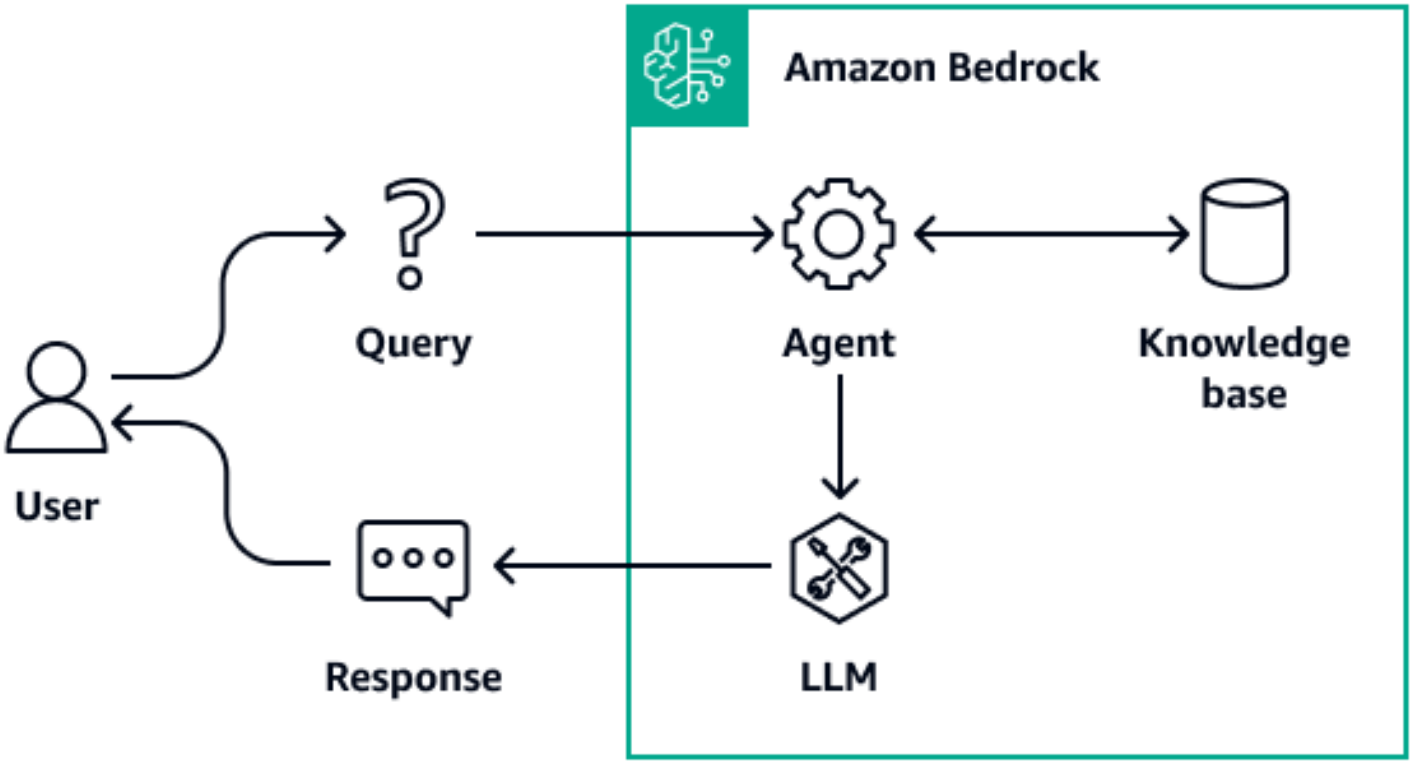
Basis pengetahuan untuk Amazon Bedrock

Amazon Bedrock adalah layanan yang dikelola sepenuhnya yang membuat model foundation berkinerja tinggi (FMs) dari startup AI terkemuka dan Amazon tersedia untuk Anda gunakan melalui API terpadu. Basis pengetahuan adalah kemampuan Amazon Bedrock yang membantu Anda menerapkan seluruh alur kerja RAG, mulai dari konsumsi hingga pengambilan dan augmentasi yang cepat. Tidak perlu membangun integrasi kustom ke sumber data atau untuk mengelola aliran data. Manajemen konteks sesi dibangun sehingga aplikasi AI generatif Anda dapat dengan mudah mendukung percakapan multi-putaran.

Setelah Anda menentukan lokasi data Anda, basis pengetahuan untuk Amazon Bedrock secara internal mengambil dokumen, memotongnya menjadi blok teks, mengubah teks menjadi embeddings, dan kemudian menyimpan embeddings dalam database vektor pilihan Anda. Amazon Bedrock mengelola dan memperbarui embeddings, menjaga database vektor tetap sinkron dengan data.

Untuk informasi selengkapnya tentang cara kerja basis pengetahuan, lihat Cara kerja basis pengetahuan Amazon Bedrock.

Jika Anda menambahkan basis pengetahuan ke agen Amazon Bedrock, agen mengidentifikasi basis pengetahuan yang sesuai berdasarkan masukan pengguna. Agen mengambil informasi yang relevan dan menambahkan informasi ke prompt input. Prompt yang diperbarui menyediakan model dengan lebih banyak informasi konteks untuk menghasilkan respons. Untuk meningkatkan transparansi dan meminimalkan halusinasi, informasi yang diambil dari basis pengetahuan dapat dilacak ke sumbernya.



Amazon Bedrock mendukung dua berikut APIs untuk RAG:

- RetrieveAndGenerate— Anda dapat menggunakan API ini untuk menanyakan basis pengetahuan Anda dan menghasilkan tanggapan dari informasi yang diambilnya. Secara internal, Amazon Bedrock mengubah kueri menjadi embeddings, menanyakan basis pengetahuan, menambah prompt dengan hasil pencarian sebagai informasi konteks, dan mengembalikan respons yang dihasilkan LLM. Amazon Bedrock juga mengelola memori jangka pendek percakapan untuk memberikan hasil yang lebih kontekstual.
- Ambil - Anda dapat menggunakan API ini untuk menanyakan basis pengetahuan Anda dengan informasi yang diambil langsung dari basis pengetahuan. Anda dapat menggunakan informasi

yang dikembalikan dari API ini untuk memproses teks yang diambil, mengevaluasi relevansinya, atau mengembangkan alur kerja terpisah untuk pembuatan respons. Secara internal, Amazon Bedrock mengubah kueri menjadi embeddings, mencari basis pengetahuan, dan mengembalikan hasil yang relevan. Anda dapat membuat alur kerja tambahan di atas hasil pencarian. Misalnya, Anda dapat menggunakan LangChain`AmazonKnowledgeBasesRetriever`plugin untuk mengintegrasikan alur kerja RAG ke dalam aplikasi AI generatif.

Untuk contoh pola arsitektur dan step-by-step instruksi untuk menggunakan APIs, lihat Pangkalan Pengetahuan sekarang memberikan pengalaman RAG yang dikelola sepenuhnya di Amazon Bedrock (AWS posting blog). Untuk informasi selengkapnya tentang cara menggunakan `RetrieveAndGenerate` API untuk membangun alur kerja RAG untuk aplikasi berbasis obrolan cerdas, lihat Membangun aplikasi chatbot kontekstual menggunakan Pangkalan Pengetahuan Amazon Bedrock (posting blog).AWS

Sumber data untuk basis pengetahuan

Anda dapat menghubungkan data kepemilikan Anda ke basis pengetahuan. Setelah mengonfigurasi konektor sumber data, Anda dapat menyinkronkan atau memperbarui data dengan basis pengetahuan Anda dan membuat data Anda tersedia untuk kueri. Basis pengetahuan Amazon Bedrock mendukung koneksi ke sumber data berikut:

- Amazon Simple Storage Service (Amazon S3) - Anda dapat menghubungkan bucket Amazon S3 ke basis pengetahuan Amazon Bedrock dengan menggunakan konsol atau API. Basis pengetahuan mencerna dan mengindeks file dalam ember. Jenis sumber data ini mendukung fitur-fitur berikut:
  - Bidang metadata dokumen — Anda dapat menyertakan file terpisah untuk menentukan metadata file di bucket Amazon S3. Anda kemudian dapat menggunakan bidang metadata ini untuk memfilter dan meningkatkan relevansi tanggapan.
  - Filter penyertaan atau pengecualian — Anda dapat menyertakan atau mengecualikan konten tertentu saat merayapi.
  - Sinkronisasi tambahan — Perubahan konten dilacak, dan hanya konten yang telah berubah sejak sinkronisasi terakhir dirayapi.
- Confluence— Anda dapat menghubungkan Atlassian Confluence instance ke basis pengetahuan Amazon Bedrock dengan menggunakan konsol atau API. Jenis sumber data ini mendukung fitur-fitur berikut:

  - Deteksi otomatis bidang dokumen utama - Bidang metadata secara otomatis terdeteksi dan dirayapi. Anda dapat menggunakan bidang ini untuk pemfilteran.
  - Filter konten penyertaan atau pengecualian — Anda dapat menyertakan atau mengecualikan konten tertentu dengan menggunakan awalan atau pola ekspresi reguler pada spasi, judul halaman, judul blog, komentar, nama lampiran, atau ekstensi.
  - Sinkronisasi tambahan - Perubahan konten dilacak, dan hanya konten yang telah berubah sejak sinkronisasi terakhir dirayapi.
  - OAuth 2.0 otentikasi, otentikasi dengan Confluence Token API — Kredensi otentikasi disimpan di. AWS Secrets Manager
- Microsoft SharePoint— Anda dapat menghubungkan SharePoint instance ke basis pengetahuan dengan menggunakan konsol atau API. Jenis sumber data ini mendukung fitur-fitur berikut:
  - Deteksi otomatis bidang dokumen utama - Bidang metadata secara otomatis terdeteksi dan dirayapi. Anda dapat menggunakan bidang ini untuk pemfilteran.
  - Filter konten penyertaan atau pengecualian — Anda dapat menyertakan atau mengecualikan konten tertentu dengan menggunakan awalan atau pola ekspresi reguler pada judul halaman utama, nama acara, dan nama file (termasuk ekstensinya).
  - Sinkronisasi tambahan - Perubahan konten dilacak, dan hanya konten yang telah berubah sejak sinkronisasi terakhir dirayapi.
  - OAuth Otentikasi 2.0 — Kredensyal otentikasi disimpan di. AWS Secrets Manager
- Salesforce— Anda dapat menghubungkan Salesforce instance ke basis pengetahuan dengan menggunakan konsol atau API. Jenis sumber data ini mendukung fitur-fitur berikut:
  - Deteksi otomatis bidang dokumen utama - Bidang metadata secara otomatis terdeteksi dan dirayapi. Anda dapat menggunakan bidang ini untuk pemfilteran.
  - Filter konten penyertaan atau pengecualian — Anda dapat menyertakan atau mengecualikan konten tertentu dengan menggunakan awalan atau pola ekspresi reguler. Untuk daftar jenis konten yang dapat Anda terapkan filter, lihat Filter inklusi/pengecualian di dokumentasi Amazon Bedrock.
  - Sinkronisasi tambahan — Perubahan konten dilacak, dan hanya konten yang telah berubah sejak sinkronisasi terakhir dirayapi.
  - OAuth Otentikasi 2.0 — Kredensyal otentikasi disimpan di. AWS Secrets Manager
- Web Crawler — Perayap Web Amazon Bedrock menghubungkan dan merayapi yang Anda berikan. URLs Fitur-fitur berikut didukung:
  - Pilih beberapa URLs untuk dirayapi

- Hormati arahan robots.txt standar, seperti dan `Allow` `Disallow`
- Kecualikan URLs yang cocok dengan pola
- Batasi tingkat perayapan
- Di Amazon CloudWatch, lihat status setiap URL yang dirayapi

Untuk informasi selengkapnya tentang sumber data yang dapat Anda sambungkan ke basis pengetahuan Amazon Bedrock, lihat Membuat konektor sumber data untuk basis pengetahuan Anda.

Database vektor untuk basis pengetahuan

Ketika Anda mengatur koneksi antara basis pengetahuan dan sumber data, Anda harus mengkonfigurasi database vektor, juga dikenal sebagai penyimpanan vektor. Basis data vektor adalah tempat Amazon Bedrock menyimpan, memperbarui, dan mengelola embeddings yang mewakili data Anda. Setiap sumber data mendukung berbagai jenis database vektor. Untuk menentukan database vektor mana yang tersedia untuk sumber data Anda, lihat tipe sumber data.

Jika Anda lebih suka Amazon Bedrock untuk secara otomatis membuat database vektor di Amazon OpenSearch Tanpa Server untuk Anda, Anda dapat memilih opsi ini saat membuat basis pengetahuan. Namun, Anda juga dapat memilih untuk mengatur database vektor Anda sendiri. Jika Anda menyiapkan database vektor Anda sendiri, lihat Prasyarat untuk penyimpanan vektor Anda sendiri untuk basis pengetahuan. Setiap jenis database vektor memiliki prasyaratnya sendiri.

Bergantung pada tipe sumber data Anda, basis pengetahuan Amazon Bedrock mendukung database vektor berikut:

- Amazon Tanpa OpenSearch Server
- Edisi Kompatibel dengan Amazon Aurora PostgreSQL
- Pinecone (Pinecone dokumentasi)
- Redis Enterprise Cloud (Redis dokumentasi)
- MongoDB Atlas (MongoDB dokumentasi)

Amazon Q Bisnis

Amazon Q Business adalah asisten bertenaga Generatif-AI yang dikelola sepenuhnya yang dapat Anda konfigurasikan untuk menjawab pertanyaan, memberikan ringkasan, menghasilkan konten, dan

menyelesaikan tugas berdasarkan data perusahaan Anda. Ini memungkinkan pengguna akhir untuk menerima tanggapan langsung dan sadar izin dari sumber data perusahaan dengan kutipan.

Fitur utama

Kemampuan Amazon Q Business berikut ini dapat membantu Anda membangun aplikasi AI generatif berbasis RAG tingkat produksi:

- Konektor bawaan - Amazon Q Business mendukung lebih dari 40 jenis konektor, seperti konektor untuk Adobe Experience Manager (AEM), Salesforce, Jira, dan Microsoft SharePoint. Untuk daftar selengkapnya, lihat Konektor yang didukung. Jika Anda memerlukan konektor yang tidak didukung, Anda dapat menggunakan Amazon AppFlow untuk menarik data dari sumber data Anda ke Amazon Simple Storage Service (Amazon S3) dan kemudian menghubungkan Amazon Q Business ke bucket Amazon S3. Untuk daftar lengkap sumber data yang AppFlow didukung Amazon, lihat Aplikasi yang didukung.
- Pipeline pengindeksan bawaan — Amazon Q Business menyediakan pipeline bawaan untuk mengindeks data dalam database vektor. Anda dapat menggunakan AWS Lambda fungsi untuk menambahkan logika preprocessing untuk pipeline pengindeksan Anda.
- Opsi indeks — Anda dapat membuat dan menyediakan indeks asli di Amazon Q Business, dan Anda menggunakan retriever Amazon Q Business untuk menarik data dari indeks tersebut. Atau, Anda dapat menggunakan indeks Amazon Kendra yang telah dikonfigurasi sebelumnya sebagai retriever. Untuk informasi selengkapnya, lihat Membuat retriever untuk aplikasi Amazon Q Business.
- Model Foundation — Amazon Q Business menggunakan model foundation yang didukung di Amazon Bedrock. Untuk daftar lengkapnya, lihat Model foundation yang didukung di Amazon Bedrock.
- Plugin - Amazon Q Business menyediakan kemampuan untuk menggunakan plugin untuk berintegrasi dengan sistem target, seperti cara otomatis untuk meringkas informasi tiket dan pembuatan tiket Jira. Setelah dikonfigurasi, plugin dapat mendukung tindakan baca dan tulis yang dapat membantu Anda meningkatkan produktivitas pengguna akhir. Amazon Q Business mendukung dua jenis plugin: plugin bawaan dan plugin khusus.
- Pagar pembatas - Amazon Q Business mendukung kontrol global dan kontrol tingkat topik. Misalnya, kontrol ini dapat mendeteksi informasi identitas pribadi (PII), penyalahgunaan, atau informasi sensitif dalam permintaan. Untuk informasi selengkapnya, lihat Kontrol admin dan pagar pembatas di Amazon Q Business.

- Manajemen identitas — Dengan Amazon Q Business, Anda dapat mengelola pengguna dan akses mereka ke aplikasi AI generatif berbasis RAG. Untuk informasi selengkapnya, lihat Identitas dan manajemen akses untuk Amazon Q Business. Selain itu, konektor Amazon Q Business mengindeks informasi daftar kontrol akses (ACL) yang dilampirkan ke dokumen bersama dengan dokumen itu sendiri. Kemudian, Amazon Q Business menyimpan informasi ACL yang diindeksnya di Amazon Q Business User Store untuk membuat pemetaan pengguna dan grup serta memfilter respons obrolan berdasarkan akses pengguna akhir ke dokumen. Untuk informasi selengkapnya, lihat Konsep konektor sumber data.
- Pengayaan dokumen — Fitur pengayaan dokumen membantu Anda mengontrol dokumen dan atribut dokumen apa yang dicerna ke dalam indeks Anda dan juga bagaimana mereka dicerna. Ini dapat dicapai melalui dua pendekatan:
  - Konfigurasikan operasi dasar — Gunakan operasi dasar untuk menambah, memperbarui, atau menghapus atribut dokumen dari data Anda. Misalnya, Anda dapat menggosok data PII dengan memilih untuk menghapus atribut dokumen apa pun yang terkait dengan PII.
  - Konfigurasikan fungsi Lambda — Gunakan fungsi Lambda yang telah dikonfigurasi sebelumnya untuk melakukan logika manipulasi atribut dokumen lanjutan yang lebih disesuaikan ke data Anda. Misalnya, data perusahaan Anda mungkin disimpan sebagai gambar yang dipindai. Dalam hal ini, Anda dapat menggunakan fungsi Lambda untuk menjalankan pengenalan karakter optik (OCR) pada dokumen yang dipindai untuk mengekstrak teks darinya. Kemudian, setiap dokumen yang dipindai diperlakukan sebagai dokumen teks selama konsumsi. Terakhir, selama obrolan, Amazon Q akan memfaktorkan data tekstual yang diekstraksi dari dokumen yang dipindai saat menghasilkan respons.

  Ketika Anda menerapkan solusi Anda, Anda dapat memilih untuk menggabungkan kedua pendekatan pengayaan dokumen. Anda dapat menggunakan operasi dasar untuk melakukan parse pertama data Anda dan kemudian menggunakan fungsi Lambda untuk operasi yang lebih kompleks. Untuk informasi selengkapnya, lihat Pengayaan dokumen di Amazon Q Business.
- Integrasi - Setelah Anda membuat aplikasi Amazon Q Business, Anda dapat mengintegrasikannya ke aplikasi lain, seperti Slack atau Microsoft Teams. Misalnya, lihat Menerapkan Slack gateway untukBisnis Amazon Q dan Menyebarkan Microsoft Teams gateway untuk Amazon Q Business (posting AWS blog).

Kustomisasi pengguna akhir

Amazon Q Business mendukung pengunggahan dokumen yang mungkin tidak disimpan dalam sumber data dan indeks organisasi Anda. Dokumen yang diunggah tidak disimpan. Mereka

tersedia untuk digunakan hanya untuk percakapan di mana dokumen diunggah. Amazon Q Business mendukung jenis dokumen tertentu untuk diunggah. Untuk informasi selengkapnya, lihat Mengunggah file dan mengobrol di Amazon Q Business.

Amazon Q Business menyertakan fitur atribut pemfilteran berdasarkan dokumen. Baik administrator maupun pengguna akhir dapat menggunakan fitur ini. Administrator dapat menyesuaikan dan mengontrol respons obrolan untuk pengguna akhir dengan menggunakan atribut. Misalnya, jika tipe sumber data adalah atribut yang dilampirkan ke dokumen Anda, Anda dapat menentukan bahwa respons obrolan dibuat hanya dari sumber data tertentu. Atau, Anda dapat mengizinkan pengguna akhir untuk membatasi cakupan respons obrolan dengan menggunakan filter atribut yang telah Anda pilih.

Pengguna akhir dapat membuat Aplikasi Amazon Q yang ringan dan dibuat khusus dalam lingkungan aplikasi Amazon Q Business Anda yang lebih luas. Aplikasi Amazon Q memungkinkan otomatisasi tugas untuk domain tertentu, seperti aplikasi yang dibuat khusus untuk tim pemasaran.

Kanvas Amazon SageMaker AI

Amazon SageMaker AI Canvas membantu Anda menggunakan pembelajaran mesin untuk menghasilkan prediksi tanpa perlu menulis kode apa pun. Ini menyediakan antarmuka visual tanpa kode yang memberdayakan Anda untuk menyiapkan data, membangun, dan menerapkan model ML, merampingkan siklus hidup end-to-end ML dalam lingkungan terpadu. Kompleksitas persiapan data, pengembangan model, deteksi bias, penjelasan, dan pemantauan diabstraksikan jauh di belakang antarmuka yang intuitif. Pengguna tidak perlu ahli SageMaker AI atau operasi pembelajaran mesin (MLOps) untuk mengembangkan, mengoperasionalkan, dan memantau model dengan SageMaker AI Canvas.

Dengan SageMaker AI Canvas, fungsionalitas RAG disediakan melalui fitur kueri dokumen tanpa kode. Anda dapat memperkaya pengalaman obrolan di SageMaker AI Canvas dengan menggunakan indeks Amazon Kendra sebagai pencarian perusahaan yang mendasarinya. Untuk informasi selengkapnya, lihat Mengekstrak informasi dari dokumen dengan kueri dokumen.

Menghubungkan SageMaker AI Canvas ke indeks Amazon Kendra memerlukan pengaturan satu kali. Sebagai bagian dari konfigurasi domain, administrator cloud dapat memilih satu atau lebih indeks Kendra yang dapat ditanyakan pengguna saat berinteraksi dengan Canvas. SageMaker Untuk petunjuk tentang cara mengaktifkan fitur kueri dokumen, lihat Memulai menggunakan Amazon SageMaker AI Canvas.

SageMaker AI Canvas mengelola komunikasi yang mendasari antara Amazon Kendra dan model pondasi yang dipilih. Untuk informasi lebih lanjut tentang model foundation yang didukung SageMaker AI Canvas, lihat Model foundation Generative AI di SageMaker AI Canvas. Diagram berikut menunjukkan cara kerja fitur kueri dokumen setelah administrator cloud menghubungkan SageMaker AI Canvas ke indeks Amazon Kendra.

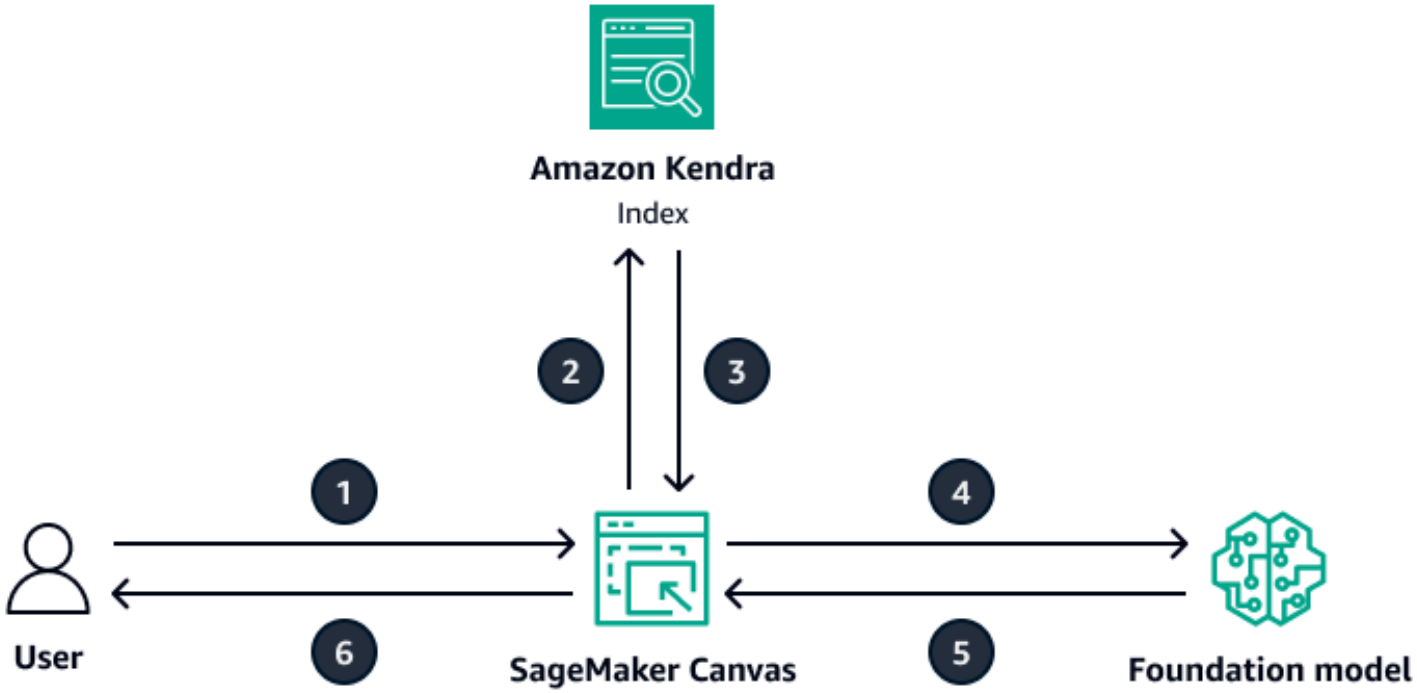


Diagram menunjukkan alur kerja berikut:

1. Pengguna memulai obrolan baru di SageMaker AI Canvas, mengaktifkan dokumen Kueri, memilih indeks target, dan kemudian mengirimkan pertanyaan.
2. SageMaker AI Canvas menggunakan kueri untuk mencari indeks Amazon Kendra untuk data yang relevan.
3. SageMaker AI Canvas mengambil data dan sumbernya dari indeks Amazon Kendra.
4. SageMaker AI Canvas memperbarui prompt untuk menyertakan konteks yang diambil dari indeks Amazon Kendra dan mengirimkan prompt ke model yayasan.
5. Model dasar menggunakan pertanyaan asli dan konteks yang diambil untuk menghasilkan jawaban.
6. SageMaker AI Canvas memberikan jawaban yang dihasilkan kepada pengguna. Ini termasuk referensi ke sumber data, seperti dokumen, yang digunakan untuk menghasilkan respons.

Arsitektur Generasi Tambahan Pengambilan Kustom di AWS

Bagian sebelumnya menjelaskan cara menggunakan Retrieval Augmented Generation (RAG) yang dikelola Layanan AWS sepenuhnya. Namun, beberapa kasus penggunaan memerlukan kontrol lebih besar atas komponen sistem, seperti retriever atau LLM (juga disebut generator). Misalnya, Anda mungkin memerlukan fleksibilitas untuk memilih database vektor Anda sendiri atau mengakses sumber data yang tidak didukung. Untuk kasus penggunaan ini, Anda dapat membangun arsitektur RAG kustom.

Bagian ini berisi topik berikut:

- Retriever untuk alur kerja RAG
- Generator untuk alur kerja RAG

Untuk informasi selengkapnya tentang cara memilih antara opsi retriever dan generator di bagian ini, lihat Memilih opsi Retrieval Augmented Generation di AWS di panduan ini.

Retriever untuk alur kerja RAG

Bagian ini menjelaskan cara membuat retriever. Anda dapat menggunakan solusi pencarian semantik yang dikelola sepenuhnya, seperti Amazon Kendra, atau Anda dapat membuat pencarian semantik kustom dengan menggunakan database vektor. AWS

Sebelum Anda meninjau opsi retriever, pastikan Anda memahami tiga langkah proses pencarian vektor:

1. Anda memisahkan dokumen yang perlu diindeks menjadi bagian-bagian yang lebih kecil. Ini disebut chunking.
2. Anda menggunakan proses yang disebut embedding untuk mengubah setiap potongan menjadi vektor matematika. Kemudian, Anda mengindeks setiap vektor dalam database vektor. Pendekatan yang Anda gunakan untuk mengindeks dokumen memengaruhi kecepatan dan akurasi pencarian. Pendekatan pengindeksan tergantung pada database vektor dan opsi konfigurasi yang disediakannya.
3. Anda mengubah kueri pengguna menjadi vektor dengan menggunakan proses yang sama. Retriever mencari database vektor untuk vektor yang mirip dengan vektor kueri pengguna. Kesamaan dihitung dengan menggunakan metrik seperti jarak Euclidean, jarak kosinus, atau produk titik.

Panduan ini menjelaskan cara menggunakan layanan berikut Layanan AWS atau pihak ketiga untuk membuat lapisan pengambilan kustom pada AWS:

- Amazon Kendra
- OpenSearch Layanan Amazon
- Amazon Aurora PostgreSQL dan pgvector
- Analisis Amazon Neptunus
- Amazon MemoryDB
- Amazon DocumentDB
- Pinecone
- MongoDB Atlas
- Weaviate

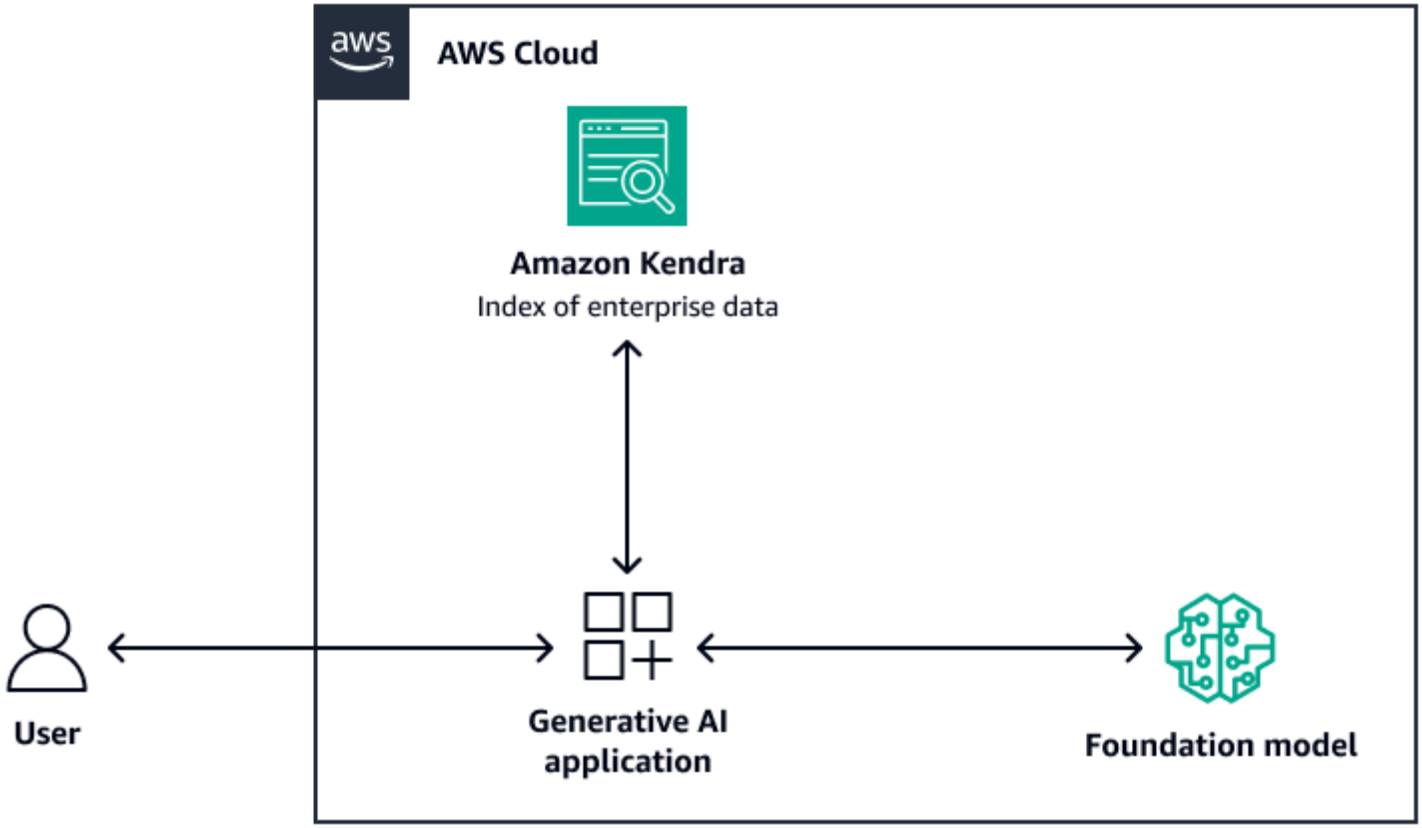
Amazon Kendra

Amazon Kendra adalah layanan pencarian cerdas yang dikelola sepenuhnya yang menggunakan pemrosesan bahasa alami dan algoritme pembelajaran mesin canggih untuk mengembalikan jawaban spesifik atas pertanyaan penelusuran dari data Anda. Amazon Kendra membantu Anda secara langsung menyerap dokumen dari berbagai sumber dan menanyakan dokumen setelah berhasil disinkronkan. Proses sinkronisasi menciptakan infrastruktur yang diperlukan untuk membuat pencarian vektor pada dokumen yang dicerna. Oleh karena itu, Amazon Kendra tidak memerlukan tiga langkah tradisional dari proses pencarian vektor. Setelah sinkronisasi awal, Anda dapat menggunakan jadwal yang ditentukan untuk menangani konsumsi yang sedang berlangsung.

Berikut ini adalah keuntungan menggunakan Amazon Kendra untuk RAG:

- Anda tidak perlu memelihara database vektor karena Amazon Kendra menangani seluruh proses pencarian vektor.
- Amazon Kendra berisi konektor pra-bangun untuk sumber data populer, seperti database, perayap situs web, bucket Amazon S3, Microsoft SharePoint contoh, dan Atlassian Confluence contoh. Konektor yang dikembangkan oleh AWS Mitra tersedia, seperti konektor untuk Box and GitLab.
- Amazon Kendra menyediakan pemfilteran daftar kontrol akses (ACL) yang hanya mengembalikan dokumen yang dapat diakses pengguna akhir.
- Amazon Kendra dapat meningkatkan respons berdasarkan metadata, seperti tanggal atau repositori sumber.

Gambar berikut menunjukkan contoh arsitektur yang menggunakan Amazon Kendra sebagai lapisan pengambilan sistem RAG. Untuk informasi selengkapnya, lihat Membuat aplikasi AI Generatif dengan akurasi tinggi dengan cepat pada data perusahaan menggunakan Amazon Kendra, LangChain, dan model bahasa besar (posting AWS blog).



Untuk model foundation, Anda dapat menggunakan Amazon Bedrock atau LLM yang digunakan melalui Amazon AI. SageMaker JumpStart Anda dapat menggunakan AWS Lambda dengan LangChainuntuk mengatur aliran antara pengguna, Amazon Kendra, dan LLM. Untuk membangun sistem RAG yang menggunakan Amazon Kendra, LangChain, dan berbagai LLMs, lihat Amazon Kendra LangChain GitHub Repositori ekstensi.

OpenSearch Layanan Amazon

Amazon OpenSearch Service menyediakan algoritme HTML bawaan untuk pencarian k-near neighbors (k-NN) untuk melakukan pencarian vektor. OpenSearch Layanan juga menyediakan mesin vektor untuk Amazon EMR Tanpa Server. Anda dapat menggunakan mesin vektor ini untuk membangun sistem RAG yang memiliki kemampuan penyimpanan dan pencarian vektor yang dapat diskalakan dan berkinerja tinggi. Untuk informasi selengkapnya tentang cara membangun sistem RAG menggunakan OpenSearch Tanpa Server, lihat Membangun alur kerja RAG yang dapat

diskalakan dan tanpa server dengan mesin vektor untuk model Amazon OpenSearch Tanpa Server dan Amazon Bedrock Claude (posting blog).AWS

Berikut ini adalah keuntungan menggunakan OpenSearch Service untuk pencarian vektor:

- Ini memberikan kontrol penuh atas database vektor, termasuk membangun pencarian vektor terukur dengan menggunakan OpenSearch Serverless.
- Ini memberikan kontrol atas strategi chunking.
- Ini menggunakan algoritma perkiraan tetangga terdekat (ANN) dari perpustakaan Non-Metric Space Library (NMSLIB), Faiss, dan Apache Lucene untuk menyalakan pencarian K-nn. Anda dapat mengubah algoritme berdasarkan kasus penggunaan. Untuk informasi selengkapnya tentang opsi untuk menyesuaikan pencarian vektor melalui OpenSearch Layanan, lihat kemampuan database vektor Amazon OpenSearch Service dijelaskan (posting AWS blog).
- OpenSearch Tanpa server terintegrasi dengan basis pengetahuan Amazon Bedrock sebagai indeks vektor.

Amazon Aurora PostgreSQL dan pgvector

Amazon Aurora PostgreSQL Compatible Edition adalah mesin database relasional terkelola sepenuhnya yang membantu Anda mengatur, mengoperasikan, dan menskalakan penerapan PostgreSQL. pgvector adalah ekstensi PostgreSQL open-source yang menyediakan kemampuan pencarian kesamaan vektor. Ekstensi ini tersedia untuk Aurora PostgreSQL yang kompatibel dan untuk Amazon Relational Database Service (Amazon RDS) untuk PostgreSQL. Untuk informasi selengkapnya tentang cara membangun sistem berbasis RAG yang menggunakan Aurora PostgreSQL kompatibel dan pgvector, lihat posting blog berikut: AWS

- Membangun pencarian bertenaga AI di PostgreSQL menggunakan Amazon AI dan pgvector SageMaker
- Manfaatkan pgvector dan Amazon Aurora PostgreSQL untuk Pemrosesan Bahasa Alami, Chatbots, dan Analisis Sentimen

Berikut ini adalah keuntungan menggunakan pgvector dan Aurora PostgreSQL kompatibel:

- Ini mendukung pencarian tetangga terdekat yang tepat dan perkiraan. Ini juga mendukung metrik kesamaan berikut: jarak L2, produk dalam, dan jarak kosinus.

- Ini mendukung Inverted File dengan Flat Compression (IVFFlat) dan Hierarchical Navigable Small Worlds (HNSW) pengindeksan.
- Anda dapat menggabungkan pencarian vektor dengan kueri atas data spesifik domain yang tersedia dalam instance PostgreSQL yang sama.
- Aurora PostgreSQL kompatibel dioptimalkan untuk I/O dan menyediakan caching berjenjang. Untuk beban kerja yang melebihi memori instans yang tersedia, pgvector dapat meningkatkan kueri per detik untuk pencarian vektor hingga 8 kali.

Analisis Amazon Neptunus

Amazon Neptune Analytics adalah mesin database grafik yang dioptimalkan untuk memori untuk analitik. Ini mendukung perpustakaan algoritma analitik grafik yang dioptimalkan, kueri grafik latensi rendah, dan kemampuan pencarian vektor dalam traversal grafik. Ini juga memiliki pencarian kesamaan vektor bawaan. Ini menyediakan satu titik akhir untuk membuat grafik, memuat data, memanggil kueri, dan melakukan pencarian kesamaan vektor. Untuk informasi selengkapnya tentang cara membangun sistem berbasis RAG yang menggunakan Neptunus Analytics, lihat Menggunakan grafik pengetahuan untuk membangun aplikasi GraphRag dengan Amazon Bedrock dan Amazon Neptunus (posting blog).AWS

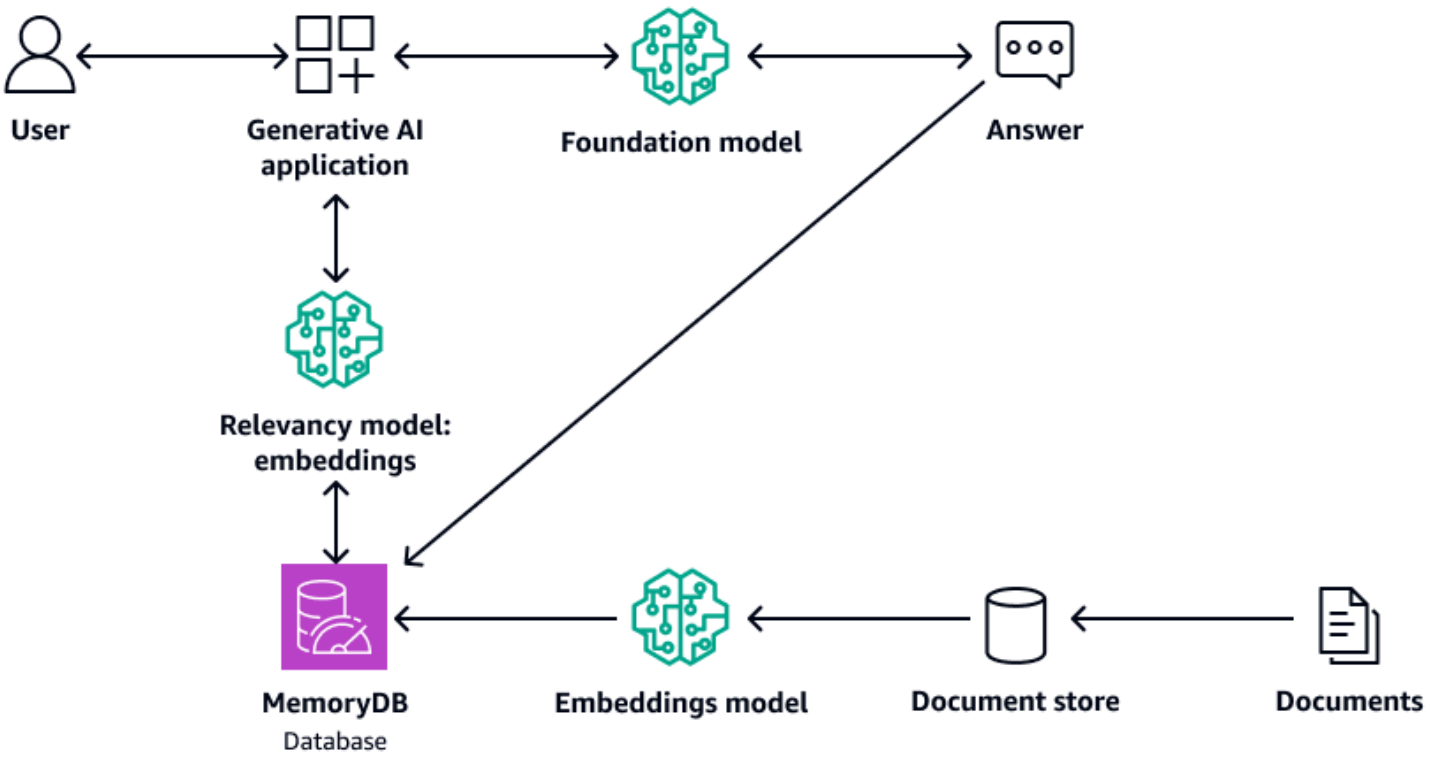
Berikut ini adalah keuntungan menggunakan Neptune Analytics:

- Anda dapat menyimpan dan mencari embeddings dalam kueri grafik.
- Jika Anda mengintegrasikan Neptune Analytics dengan LangChain, arsitektur ini mendukung kueri grafik bahasa alami.
- Arsitektur ini menyimpan dataset grafik besar dalam memori.

Amazon MemoryDB

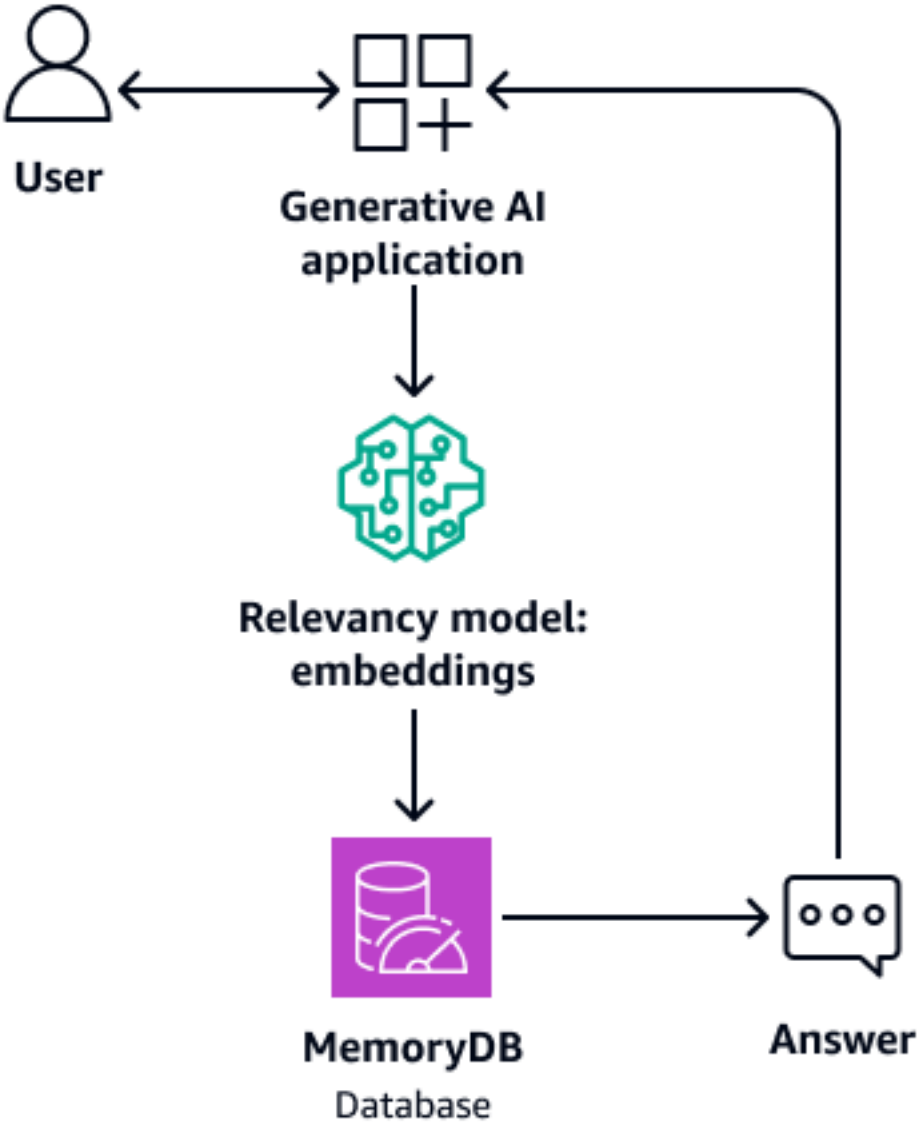
Amazon MemoryDB adalah layanan database dalam memori yang tahan lama yang memberikan kinerja sangat cepat. Semua data Anda disimpan dalam memori, yang mendukung pembacaan mikrodetik, latensi tulis milidetik satu digit, dan throughput tinggi. Pencarian vektor untuk MemoryDB memperluas fungsionalitas MemoryDB dan dapat digunakan bersama dengan fungsionalitas MemoryDB yang ada. Untuk informasi lebih lanjut, lihat Pertanyaan menjawab dengan LLM dan repositori RAG di. GitHub

Diagram berikut menunjukkan arsitektur sampel yang menggunakan MemoryDB sebagai database vektor.



Berikut ini adalah keuntungan menggunakan MemoryDB:

- Ini mendukung algoritma pengindeksan Flat dan HNSW. Untuk informasi lebih lanjut, lihat Pencarian vektor untuk Amazon MemoryDB sekarang tersedia secara umum di Blog Berita AWS
- Ini juga dapat bertindak sebagai memori penyangga untuk model pondasi. Ini berarti bahwa pertanyaan yang dijawab sebelumnya diambil dari buffer alih-alih melalui proses pengambilan dan pembuatan lagi. Diagram berikut menunjukkan proses ini.



- Karena menggunakan database dalam memori, arsitektur ini menyediakan waktu kueri milidetik satu digit untuk pencarian semantik.
- Ini menyediakan hingga 33.000 kueri per detik pada penarikan 95-99% dan 26.500 kueri per detik dengan penarikan lebih dari 99%. Untuk informasi lebih lanjut, lihat AWS re:Invent 2023 - Pencarian vektor latensi ultra-rendah untuk video Amazon MemoryDB di YouTube.

Amazon DocumentDB

Amazon DocumentDB (dengan kompatibilitas MongoDB) adalah layanan database yang cepat, andal, dan dikelola sepenuhnya. Itu membuatnya mudah untuk mengatur, mengoperasikan, dan menskalakan MongoDB-database yang kompatibel di cloud. Pencarian vektor untuk Amazon

DocumentDB menggabungkan fleksibilitas dan kemampuan query yang kaya dari database dokumen berbasis JSON dengan kekuatan pencarian vektor. Untuk informasi lebih lanjut, lihat Pertanyaan menjawab dengan LLM dan repositori RAG di. GitHub

Diagram berikut menunjukkan contoh arsitektur yang menggunakan Amazon DocumentDB sebagai database vektor.

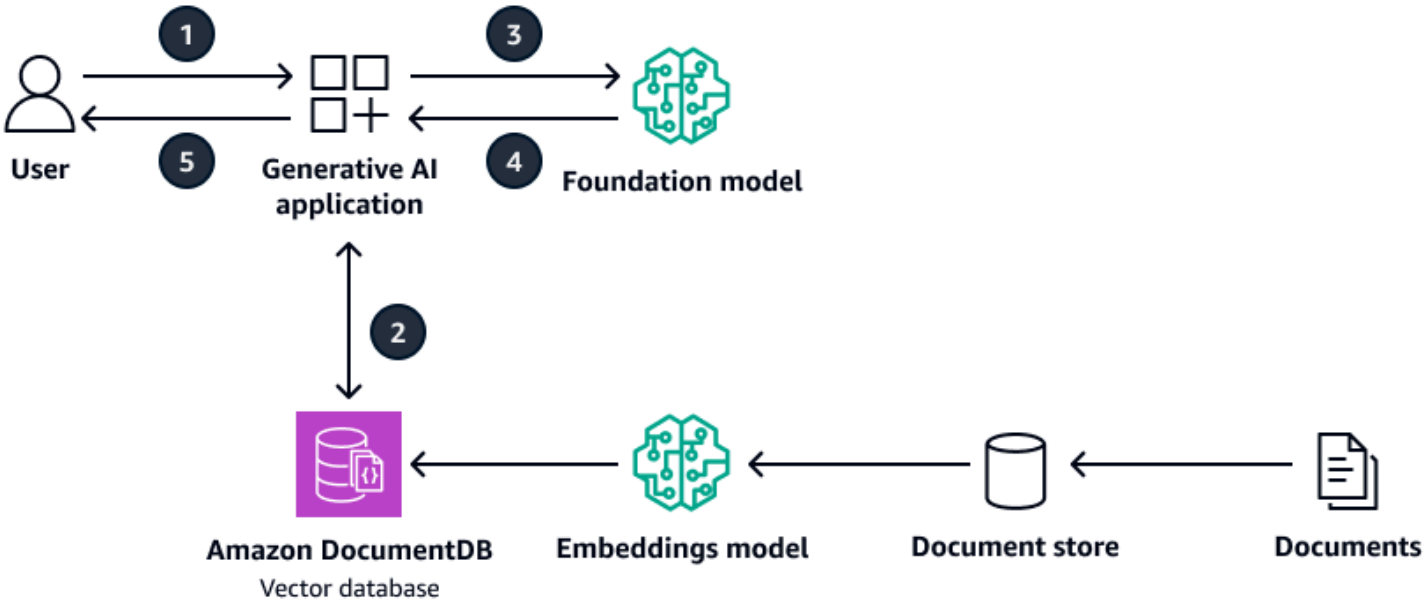


Diagram menunjukkan alur kerja berikut:

1. Pengguna mengirimkan kueri ke aplikasi AI generatif.
2. Aplikasi AI generatif melakukan pencarian kesamaan di database vektor Amazon DocumentDB dan mengambil ekstrak dokumen yang relevan.
3. Aplikasi AI generatif memperbarui kueri pengguna dengan konteks yang diambil dan mengirimkan prompt ke model pondasi target.
4. Model dasar menggunakan konteks untuk menghasilkan respons terhadap pertanyaan pengguna dan mengembalikan respons.
5. Aplikasi AI generatif mengembalikan respons kepada pengguna.

Berikut ini adalah keuntungan menggunakan Amazon DocumentDB:

- Ini mendukung metode HNSW dan IVFFlat pengindeksan.
- Ini mendukung hingga 2.000 dimensi dalam data vektor dan mendukung metrik jarak produk Euclidean, cosinus, dan titik.

- Ini memberikan waktu respons milidetik.

Pinecone

Pineconeadalah database vektor yang dikelola sepenuhnya yang membantu Anda menambahkan pencarian vektor ke aplikasi produksi. Ini tersedia melalui AWS Marketplace. Penagihan didasarkan pada penggunaan, dan biaya dihitung dengan mengalikan harga pod dengan jumlah pod. Untuk informasi lebih lanjut tentang cara membangun sistem berbasis RAG yang menggunakan Pinecone, lihat posting AWS blog berikut:

- Mengurangi halusinasi melalui RAG menggunakan Pinecone database vektor & Llama-2 dari Amazon AI SageMaker JumpStart
- Gunakan Amazon SageMaker AI Studio untuk membuat solusi penjawab pertanyaan RAG dengan Llama 2, LangChain, dan Pinecone untuk eksperimen cepat

Diagram berikut menunjukkan contoh arsitektur yang menggunakan Pinecone sebagai database vektor.

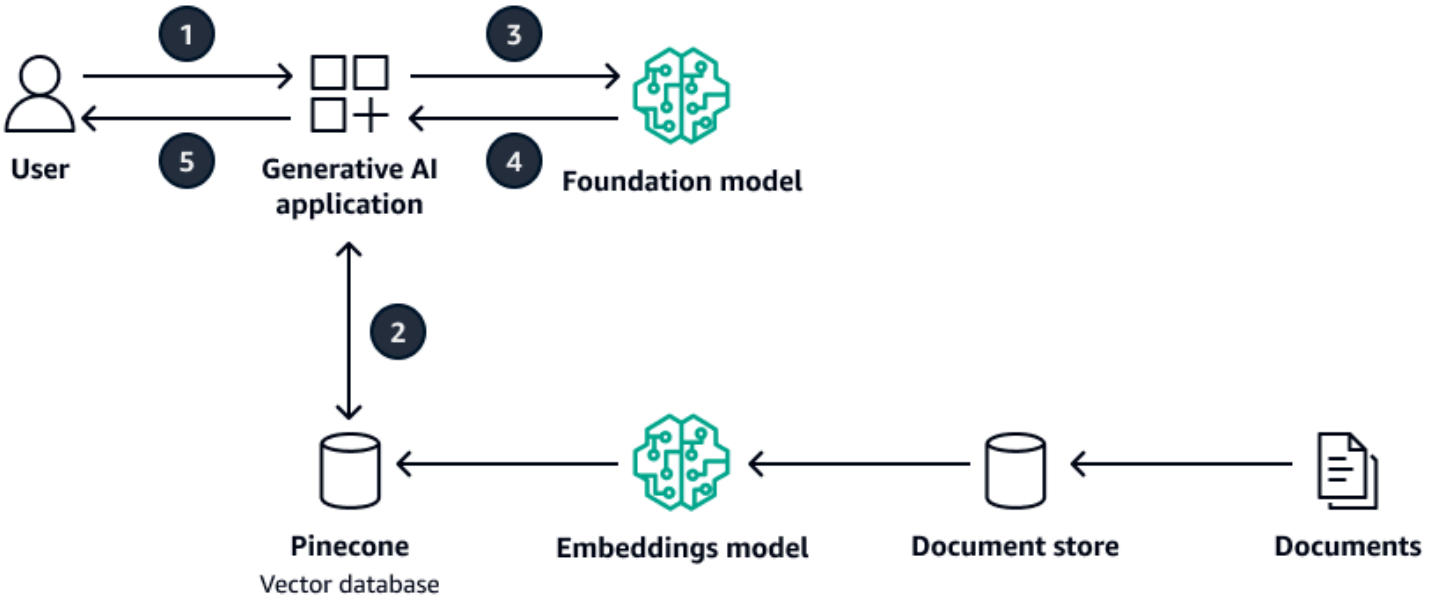


Diagram menunjukkan alur kerja berikut:

1. Pengguna mengirimkan kueri ke aplikasi AI generatif.
2. Aplikasi AI generatif melakukan pencarian kesamaan di Pinecone database vektor dan mengambil ekstrak dokumen yang relevan.

3. Aplikasi AI generatif memperbarui kueri pengguna dengan konteks yang diambil dan mengirimkan prompt ke model pondasi target.
4. Model dasar menggunakan konteks untuk menghasilkan respons terhadap pertanyaan pengguna dan mengembalikan respons.
5. Aplikasi AI generatif mengembalikan respons kepada pengguna.

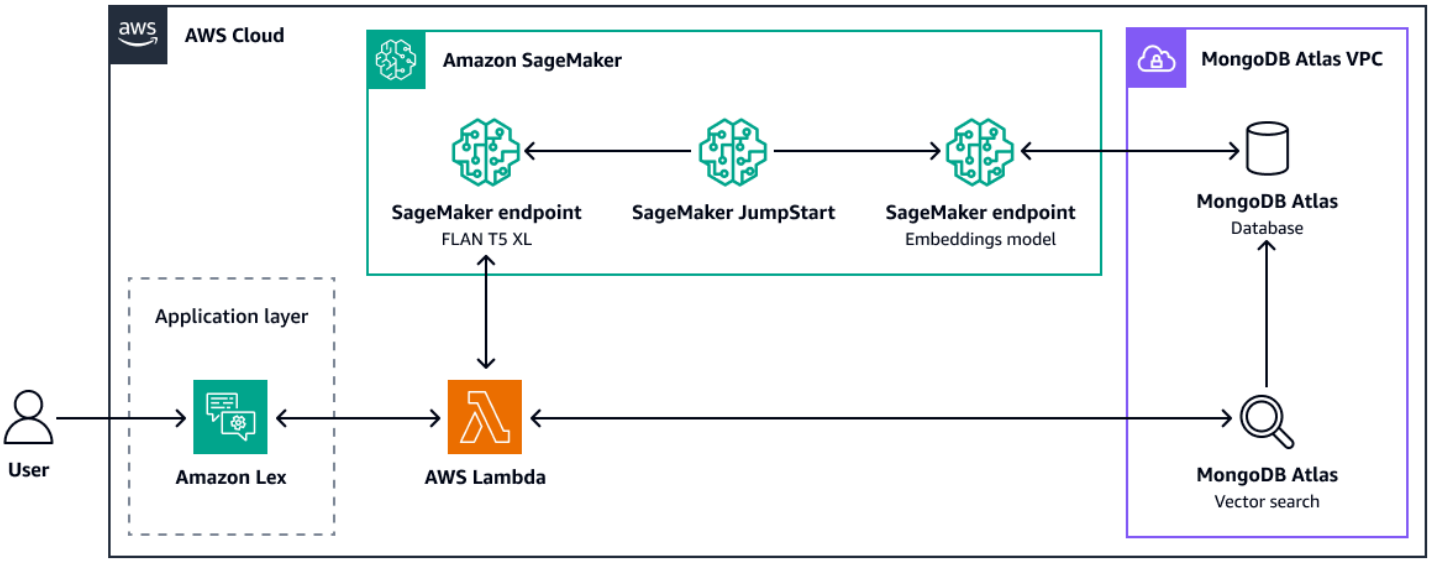
Berikut ini adalah keuntungan menggunakan Pinecone:

- Ini adalah database vektor yang dikelola sepenuhnya dan menghilangkan biaya pengelolaan infrastruktur Anda sendiri.
- Ini menyediakan fitur tambahan penyaringan, pembaruan indeks langsung, dan peningkatan kata kunci (pencarian hibrida).

MongoDB Atlas

MongoDB Atlasadalah database cloud yang dikelola sepenuhnya yang menangani semua kompleksitas penerapan dan pengelolaan penerapan Anda. AWS Anda dapat menggunakan pencarian Vektor untuk MongoDB Atlasuntuk menyimpan embeddings vektor di MongoDB basis data. Basis pengetahuan Amazon Bedrock mendukung MongoDB Atlas untuk penyimpanan vektor. Untuk informasi selengkapnya, lihat Memulai Integrasi Basis Pengetahuan Amazon Bedrock di MongoDB dokumentasi.

Untuk informasi lebih lanjut tentang cara menggunakan MongoDB Atlas pencarian vektor untuk RAG, lihat Retrieval-Augmented Generation dengan LangChain, Amazon SageMaker AI JumpStart, dan MongoDB Atlas Pencarian Semantik (posting AWS blog). Diagram berikut menunjukkan arsitektur solusi rinci dalam posting blog ini.



Berikut ini adalah keuntungan menggunakan MongoDB Atlas pencarian vektor:

- Anda dapat menggunakan implementasi yang sudah ada MongoDB Atlas untuk menyimpan dan mencari embeddings vektor.
- Anda dapat menggunakan MongoDB Kueri API untuk menanyakan embeddings vektor.
- Anda dapat secara mandiri menskalakan pencarian vektor dan database.
- Penyematan vektor disimpan di dekat data sumber (dokumen), yang meningkatkan kinerja pengindeksan.

Weaviate

Weaviateadalah sumber terbuka populer, database vektor latensi rendah yang mendukung jenis media multimodal, seperti teks dan gambar. Basis data menyimpan objek dan vektor, yang menggabungkan pencarian vektor dengan penyaringan terstruktur. Untuk informasi lebih lanjut tentang penggunaan Weaviate dan Amazon Bedrock untuk membangun alur kerja RAG, lihat Membangun solusi AI generatif siap perusahaan dengan model foundation Cohere di Amazon Bedrock dan Weaviate database vektor pada AWS Marketplace (posting AWS blog).

Berikut ini adalah keuntungan menggunakan Weaviate:

- Ini adalah open source dan didukung oleh komunitas yang kuat.
- Itu dibangun untuk pencarian hibrida (baik vektor dan kata kunci).

- Anda dapat menerapkannya sebagai penawaran perangkat lunak terkelola AWS sebagai layanan (SaaS) atau sebagai klaster Kubernetes.

Generator untuk alur kerja RAG

Model bahasa besar (LLMs) adalah model pembelajaran mendalam yang sangat besar yang dilatih sebelumnya pada sejumlah besar data. Mereka sangat fleksibel. LLMs dapat melakukan berbagai tugas, seperti menjawab pertanyaan, meringkas dokumen, menerjemahkan bahasa, dan menyelesaikan kalimat. Mereka memiliki potensi untuk mengganggu pembuatan konten dan cara orang menggunakan mesin pencari dan asisten virtual. Meskipun tidak sempurna, LLMs tunjukkan kemampuan luar biasa untuk membuat prediksi berdasarkan prompt atau jumlah input yang relatif kecil.

LLMs adalah komponen penting dari solusi RAG. Untuk arsitektur RAG kustom, ada dua Layanan AWS yang berfungsi sebagai opsi utama:

- Amazon Bedrock adalah layanan yang dikelola sepenuhnya yang membuat LLMs dari perusahaan AI terkemuka dan Amazon tersedia untuk Anda gunakan melalui API terpadu.
- Amazon SageMaker AI JumpStart adalah hub ML yang menawarkan model dasar, algoritme bawaan, dan solusi HTML bawaan. Dengan SageMaker AI JumpStart, Anda dapat mengakses model yang telah dilatih sebelumnya, termasuk model pondasi. Anda juga dapat menggunakan data Anda sendiri untuk menyempurnakan model yang telah dilatih sebelumnya.

Amazon Bedrock

Amazon Bedrock menawarkan model industri terkemuka dari Anthropic, Stability AI, Meta, Cohere, AI21 Labs, Mistral AI, dan Amazon. Untuk daftar lengkapnya, lihat Model foundation yang didukung di Amazon Bedrock. Amazon Bedrock juga memungkinkan Anda untuk menyesuaikan model dengan data Anda sendiri.

Anda dapat mengevaluasi kinerja model untuk menentukan mana yang paling cocok untuk kasus penggunaan RAG Anda. Anda dapat menguji model terbaru dan juga menguji untuk melihat kemampuan dan fitur mana yang memberikan hasil terbaik dan dengan harga terbaik. Bagian Anthropic Model Claude Sonnet adalah pilihan umum untuk aplikasi RAG karena unggul dalam berbagai tugas dan memberikan tingkat keandalan dan prediktabilitas yang tinggi.

SageMaker AI JumpStart

SageMaker AI JumpStart menyediakan model open source terlatih untuk berbagai jenis masalah. Anda dapat melatih dan menyempurnakan model-model ini secara bertahap sebelum penerapan. Anda dapat mengakses model, templat solusi, dan contoh yang telah dilatih sebelumnya melalui halaman JumpStart arahan SageMaker AI di Amazon SageMaker AI Studio atau menggunakan AI Python SageMaker SDK.

SageMaker AI JumpStart menawarkan model state-of-the-art dasar untuk kasus penggunaan seperti penulisan konten, pembuatan kode, penjawab pertanyaan, copywriting, ringkasan, klasifikasi, pengambilan informasi, dan banyak lagi. Gunakan model JumpStart foundation untuk membuat solusi AI generatif Anda sendiri dan mengintegrasikan solusi khusus dengan fitur SageMaker AI tambahan. Untuk informasi selengkapnya, lihat Memulai dengan Amazon SageMaker AI JumpStart.

SageMaker AI melakukan JumpStart onboard dan memelihara model foundation yang tersedia untuk umum agar Anda dapat mengakses, menyesuaikan, dan mengintegrasikan ke dalam siklus hidup ML Anda. Untuk informasi lebih lanjut, lihat Model dasar yang tersedia untuk umum. SageMaker AI JumpStart juga mencakup model yayasan berpemilik dari penyedia pihak ketiga. Untuk informasi selengkapnya, lihat Model foundation proprietary.

Memilih opsi Retrieval Augmented Generation di AWS

Opsi RAG yang dikelola sepenuhnya dan bagian arsitektur RAG Kustom dari panduan ini menjelaskan berbagai pendekatan untuk membangun solusi pencarian berbasis RAG. AWS Bagian ini menjelaskan cara memilih di antara opsi-opsi ini berdasarkan kasus penggunaan Anda. Dalam beberapa situasi, lebih dari satu opsi mungkin berhasil. Dalam skenario itu, pilihan tergantung pada kemudahan implementasi, keterampilan yang tersedia di organisasi Anda, dan kebijakan dan standar perusahaan Anda.

Kami menyarankan Anda mempertimbangkan opsi RAG yang dikelola sepenuhnya dan kustom dalam urutan berikut dan memilih opsi pertama yang sesuai dengan kasus penggunaan Anda:

1. Gunakan Amazon Q Business kecuali:
   - Layanan ini tidak tersedia di Anda Wilayah AWS, dan data Anda tidak dapat dipindahkan ke Wilayah yang tersedia
   - Anda memiliki alasan khusus untuk menyesuaikan alur kerja RAG
   - Anda ingin menggunakan database vektor yang ada atau LLM tertentu
2. Gunakan basis pengetahuan untuk Amazon Bedrock kecuali:
   - Anda memiliki database vektor yang tidak didukung
   - Anda memiliki alasan khusus untuk menyesuaikan alur kerja RAG
3. Gabungkan Amazon Kendra dengan generator pilihan Anda kecuali:
   - Anda ingin memilih database vektor Anda sendiri
   - Anda ingin menyesuaikan strategi chunking
4. Jika Anda ingin lebih banyak kontrol atas retriever dan ingin memilih database vektor Anda sendiri:
   - Jika Anda tidak memiliki database vektor yang ada dan tidak memerlukan latensi rendah atau kueri grafik, pertimbangkan untuk menggunakan Amazon OpenSearch Service.
   - Jika Anda memiliki yang sudah ada PostgreSQL database vektor, pertimbangkan untuk menggunakan Amazon Aurora PostgreSQL dan pgvectorpilihan.
   - Jika Anda membutuhkan latensi rendah, pertimbangkan opsi dalam memori, seperti Amazon MemoryDB atau Amazon DocumentDB.
   - Jika Anda ingin menggabungkan pencarian vektor dengan kueri grafik, pertimbangkan Amazon Neptunus Analytics.
   - Jika Anda sudah menggunakan database vektor pihak ketiga atau menemukan manfaat tertentu darinya, pertimbangkan Pinecone, MongoDB Atlas, dan Weaviate.

5. Jika Anda ingin memilih LLM:
   - Jika Anda menggunakan Amazon Q Business, Anda tidak dapat memilih LLM.
   - Jika Anda menggunakan Amazon Bedrock, Anda dapat memilih salah satu model pondasi yang didukung.
   - Jika Anda menggunakan Amazon Kendra atau database vektor kustom, Anda dapat menggunakan salah satu generator yang dijelaskan dalam panduan ini atau menggunakan LLM kustom.

   > **Note**
   >
   > Anda juga dapat menggunakan dokumen kustom Anda untuk menyempurnakan LLM yang ada untuk meningkatkan akurasi tanggapannya. Untuk informasi selengkapnya, lihat Membandingkan RAG dan fine-tuning dalam panduan ini.

6. Jika Anda memiliki implementasi Amazon SageMaker AI Canvas yang ingin Anda gunakan atau jika Anda ingin membandingkan respons RAG dari yang berbeda LLMs, pertimbangkan Amazon SageMaker AI Canvas.

Kesimpulan

Panduan ini menjelaskan berbagai opsi untuk membangun sistem Retrieval Augmented Generation (RAG). AWS Anda dapat memulai dengan layanan yang dikelola sepenuhnya, seperti Amazon Q Business dan basis pengetahuan Amazon Bedrock. Jika Anda ingin lebih banyak kontrol atas alur kerja RAG, Anda dapat memilih retriever kustom. Untuk generator, Anda dapat menggunakan API untuk memanggil LLM yang didukung di Amazon Bedrock, atau Anda dapat menerapkan LLM Anda sendiri dengan menggunakan Amazon AI. SageMaker JumpStart Tinjau rekomendasi dalam Memilih opsi RAG untuk menentukan opsi mana yang paling cocok untuk kasus penggunaan Anda. Setelah Anda memilih opsi terbaik untuk kasus penggunaan Anda, gunakan referensi yang disediakan dalam panduan ini untuk mulai membangun aplikasi berbasis RAG Anda.

Riwayat dokumen

Tabel berikut menjelaskan perubahan signifikan pada panduan ini. Jika Anda ingin diberi tahu tentang pembaruan masa depan, Anda dapat berlangganan umpan RSS.

Perubahan	Deskripsi	Tanggal
Publikasi awal	—	Oktober 28, 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.