



Penilaian beban kerja AI generatif

AWS Bimbingan Preskriptif



AWS Bimbingan Preskriptif: Penilaian beban kerja AI generatif

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

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

Table of Contents

Pengantar	1
Tujuan dari panduan ini	2
Target audiens dan manfaatnya	2
Cakupan	2
Hasil bisnis yang ditargetkan	4
Pertimbangan dan prasyarat penilaian	7
Mulailah dengan kasus penggunaan yang jelas	7
Pastikan keselarasan bisnis	8
Menerapkan tata kelola dan pengawasan	8
Data alamat dan prasyarat teknis	8
Pertimbangkan persyaratan sumber daya komputasi	8
Mengatasi implikasi privasi dan keamanan	9
Libatkan pemangku kepentingan lebih awal	9
Iterasi dan pelajari	9
Kuesioner penilaian beban kerja AI generatif	10
Kesiapan	11
Kasus penggunaan	13
Arsitektur	16
Penyimpanan	17
Peraturan dan kepatuhan	18
Integrasi	19
Pengujian	21
Penerapan dan otomatisasi	22
Strategi data	25
Menerjemahkan wawasan penilaian menjadi hasil yang dapat ditindaklanjuti	28
Langkah selanjutnya	30
Pertanyaan yang Sering Diajukan	31
Apa tujuan utamanya?	31
Siapa yang harus menggunakan penilaian ini?	31
Apa komponen kuncinya?	31
Bagaimana ini membantu mendefinisikan arsitektur?	31
Apa manfaatnya?	31
Bagaimana kita bisa mengimplementasikan ini dengan sukses?	32
Apa tantangannya?	32

Apa persyaratan peraturan dan kepatuhan?	32
Apa peran pemangku kepentingan?	32
Bagaimana kita bisa mengukur kesuksesan?	33
Bagaimana pendekatannya berbeda berdasarkan ukuran organisasi?	33
Sumber daya	34
Riwayat dokumen	35
Glosarium	36
#	36
A	37
B	40
C	42
D	45
E	49
F	51
G	53
H	54
I	55
L	58
M	59
O	63
P	66
Q	69
R	69
D	72
T	76
U	78
V	78
W	79
Z	80
.....	lxxxi

Penilaian beban kerja AI generatif

Tabby Ward dan Deepak Dixit, Amazon Web Services (AWS)

November 2024 ([riwayat dokumen](#))

Penilaian beban kerja AI generatif adalah metode strategis yang bertujuan untuk mengevaluasi dan meningkatkan kesiapan organisasi untuk membuat atau memperbarui beban kerja AI generatifnya. Penilaian ini penting karena memasukkan AI generatif ke dalam operasi bisnis dapat sangat mengubah cara kerja, dan dapat memberikan efisiensi dan kemampuan baru. Namun, untuk mengadopsi AI generatif dengan sukses, penting untuk benar-benar memahami sistem saat ini dan memiliki rencana yang jelas untuk masa depan.

Beban kerja AI generatif mengacu pada tugas komputasi yang melibatkan penggunaan model kecerdasan buatan yang dapat membuat konten baru, seperti teks, gambar, kode, atau tipe data lainnya. Beban kerja ini biasanya membutuhkan daya komputasi yang besar, perangkat keras khusus seperti GPUs, dan kumpulan data besar untuk pelatihan dan inferensi. Mengintegrasikan beban kerja AI generatif ke dalam operasi menghadirkan beberapa tantangan:

- Persyaratan infrastruktur: Menyediakan sumber daya komputasi yang signifikan dan perangkat keras khusus yang dibutuhkan model AI generatif.
- Manajemen data: Memastikan kualitas data, privasi, dan kepatuhan saat menangani kumpulan data besar.
- Kesenjangan keterampilan: Kurangnya keahlian dalam teknologi AI dan penerapan model.
- Pertimbangan etis: Mengatasi bias, keadilan, dan transparansi dalam konten yang dihasilkan AI.
- Kompleksitas integrasi: Menggabungkan AI generatif dengan mulus ke dalam alur kerja dan sistem lama yang ada.
- Manajemen biaya: Menyeimbangkan manfaat potensial dengan biaya implementasi dan operasi yang tinggi.

Mengatasi tantangan ini membutuhkan perencanaan yang matang, investasi dalam infrastruktur dan bakat, dan pendekatan strategis untuk implementasi.

Tujuan dari panduan ini

AI generatif dengan cepat menjadi komponen penting di banyak industri. Ini memberikan peluang transformatif tetapi juga menimbulkan tantangan dalam hal integrasi, kepatuhan, dan skalabilitas. Banyak organisasi berjuang untuk sepenuhnya memanfaatkan AI karena fondasi teknologi yang lemah, resistensi terhadap perubahan, dan masalah kualitas data. Penilaian beban kerja AI generatif mengatasi tantangan ini dengan mengidentifikasi persyaratan untuk modernisasi, mendefinisikan ruang lingkup implementasi, dan menantang sistem dan pemikiran warisan. Ini juga membantu dalam menentukan produk minimum yang layak (MVPs) dan membantu Anda mengembangkan arsitektur solusi target, memastikan pendekatan terstruktur dan strategis untuk adopsi AI.

Panduan ini berfungsi sebagai pendekatan terstruktur untuk membantu organisasi menavigasi kompleksitas mengadopsi teknologi AI generatif. Alih-alih mendefinisikan persyaratan dengan jelas sejak awal, panduan ini membantu dalam:

- Mengidentifikasi kasus penggunaan potensial untuk AI generatif dalam organisasi Anda.
- Menilai kesiapan organisasi Anda untuk adopsi AI generatif.
- Mendefinisikan dan menyempurnakan tujuan kasus penggunaan dan meregangkan tujuan.
- Menentukan ruang lingkup dan persyaratan untuk implementasi AI generatif.
- Mengembangkan arsitektur solusi target.

Target audiens dan manfaatnya

Penilaian ini dirancang khusus untuk arsitek solusi, arsitek perusahaan, dan arsitek aplikasi yang ingin mengevaluasi aspek teknis modernisasi beban kerja AI generatif. Ini juga berharga bagi manajer program dan orang yang ingin mengukur kesiapan tim mereka secara keseluruhan, alokasi sumber daya, dan persyaratan pemberdayaan. Praktik terbaik industri menekankan pentingnya penilaian komprehensif untuk memastikan kesiapan adopsi AI. Ini termasuk mengevaluasi arsitektur, penyimpanan, kepatuhan, integrasi, pengujian, penyebaran, dan otomatisasi.

Cakupan

Topik-topik berikut adalah cakupan metode penilaian beban kerja AI generatif:

- Teknologi dan model AI generatif saat ini (misalnya, model bahasa besar, model pembuatan gambar)

- Aplikasi AI sempit yang menggunakan teknik generatif
- Integrasi AI generatif dengan sistem dan alur kerja yang ada
- Strategi data untuk pelatihan dan penyempurnaan model AI generatif
- Pertimbangan etis dan praktik AI yang bertanggung jawab untuk aplikasi AI generatif saat ini
- Strategi pengujian dan penerapan untuk AI generatif di lingkungan produksi
- Pertimbangan keamanan dan privasi untuk implementasi AI generatif
- Optimalisasi kinerja dan skalabilitas beban kerja AI generatif
- Kasus penggunaan dan aplikasi AI generatif di berbagai industri
- Evaluasi output AI generatif dan proses jaminan kualitas

Topik-topik berikut berada di luar cakupan:

- Skenario kecerdasan umum buatan (AGI) dan kecerdasan super buatan (ASI)
- Kemajuan spekulatif future dalam AI di luar model generatif saat ini
- Aplikasi komputasi kuantum dalam AI
- Komputasi neuromorfik dan antarmuka otak-komputer
- Kesadaran dan kesadaran diri dalam sistem AI
- Dampak sosial jangka panjang dari AI canggih di luar aplikasi AI generatif saat ini
- Kerangka kerja regulasi untuk teknologi AI masa depan hipotetis
- Perdebatan filosofis tentang sifat kecerdasan dan kesadaran dalam mesin
- Kasus tepi ekstrim atau kasus penggunaan AI yang sangat spekulatif
- Spesifikasi teknis terperinci dari model atau arsitektur AI berpemilik

Hasil bisnis yang ditargetkan

Penilaian beban kerja AI generatif bertujuan untuk memberikan beberapa hasil yang ditargetkan yang sangat penting untuk berhasil memodernisasi beban kerja AI generatif. Hasil ini memastikan bahwa organisasi dipersiapkan dengan baik untuk mengintegrasikan teknologi AI secara efektif dan efisien.

Untuk setiap hasil yang ditargetkan, penilaian beban kerja AI generatif berfokus pada:

- **Interdependensi:** Identifikasi dan klarifikasi setiap ketergantungan antara hasil dan aspek lain dari proses modernisasi. Ini termasuk memahami bagaimana satu hasil dapat mempengaruhi atau dipengaruhi oleh orang lain, untuk memastikan pendekatan holistik terhadap modernisasi.
- **Penyelarasan pemangku kepentingan:** Garis besar strategi untuk menyelaraskan berbagai pemangku kepentingan dengan setiap hasil. Ini melibatkan mengkomunikasikan nilai dan dampak dari setiap hasil ke tingkat organisasi dan departemen yang berbeda, untuk mendorong dukungan dan dukungan.
- **Prioritas:** Dalam kasus di mana beberapa kasus penggunaan atau hasil diidentifikasi, sediakan kerangka kerja untuk memprioritaskannya berdasarkan faktor-faktor seperti dampak bisnis, persyaratan sumber daya, dan penyelarasan strategis.
- **Perbaikan berkelanjutan:** Untuk setiap hasil, buat mekanisme untuk evaluasi dan penyempurnaan yang berkelanjutan. Ini memastikan bahwa upaya modernisasi tetap adaptif dan responsif terhadap perubahan lanskap teknologi dan kebutuhan bisnis.

Berikut adalah diskusi rinci dari setiap hasil yang ditargetkan:

Arsitektur target

- **Definisi:** Penilaian membantu menentukan arsitektur target yang jelas dan dapat diskalakan untuk beban kerja AI generatif.
- **Komponen:** Ini termasuk memilih layanan cloud yang sesuai, merancang jaringan data, dan memastikan interoperabilitas sistem.
- **Manfaat:** Arsitektur yang terdefinisi dengan baik mendukung skalabilitas, keandalan, dan optimalisasi kinerja, dan memberikan dasar yang kuat untuk modernisasi.

Kesiapan pelanggan

- **Evaluasi:** Menilai keadaan infrastruktur, proses, dan budaya organisasi saat ini untuk menentukan kesiapan adopsi modernisasi AI generatif.
- **Kriteria:** Ini melibatkan evaluasi kemampuan teknis, kualitas data, dan kemauan organisasi untuk merangkul perubahan.
- **Hasil:** Mengidentifikasi kesenjangan dan area untuk perbaikan memastikan bahwa organisasi siap untuk transisi yang mulus ke solusi dan teknologi modern.

Gunakan tujuan kasus dan regangkan tujuan

- Tujuan kasus penggunaan menetapkan tujuan yang jelas untuk implementasi solusi target, dengan fokus pada masalah atau peluang bisnis tertentu.

Tujuan kasus penggunaan dalam konteks modernisasi AI generatif mengacu pada tujuan spesifik dan terukur yang ingin dicapai organisasi dengan menerapkan solusi AI generatif. Tujuan ini biasanya selaras dengan tujuan bisnis yang lebih luas dan fokus pada mengatasi tantangan atau peluang tertentu dalam organisasi. Contoh tujuan kasus penggunaan mungkin termasuk:

- Mengurangi waktu respons layanan pelanggan hingga 50 persen dengan menggunakan chatbots generatif yang didukung AI.
- Meningkatkan efisiensi peninjauan kode sebesar 30 persen melalui analisis kode berbantuan AI generatif.
- Meningkatkan akurasi deteksi penipuan sebesar 25 persen dengan menggunakan pengenalan pola AI generatif.
- Tujuan peregangan menentukan target ambisius yang mendorong batas-batas apa yang dapat dicapai oleh modernisasi AI generatif dalam organisasi.
- **Dampak:** Menetapkan tujuan yang dapat dicapai dan aspirasional membantu menyelaraskan inisiatif modernisasi AI generatif dengan tujuan bisnis strategis dan mendorong inovasi.

Estimasi upaya

- **Tujuan:** Estimasi upaya yang akurat membantu dalam perencanaan sumber daya dan memastikan bahwa proyek disampaikan tepat waktu dan sesuai anggaran.
- **Lingkup:** Perkirakan sumber daya, waktu, dan anggaran yang diperlukan untuk mengimplementasikan rencana modernisasi AI generatif.
- **Faktor:** Pertimbangkan kompleksitas teknis, tantangan integrasi, dan potensi risiko.

Kebutuhan pemberdayaan

- Pelatihan dan pengembangan: Identifikasi keterampilan dan pengetahuan yang diperlukan untuk adopsi modernisasi AI generatif yang sukses.
- Sumber Daya: Menentukan kebutuhan akan program pelatihan, lokakarya, dan kegiatan pemberdayaan lainnya.
- Hasil: Memastikan bahwa staf dilengkapi dengan keterampilan yang diperlukan meningkatkan efektivitas inisiatif modernisasi AI generatif dan mendukung kesuksesan jangka panjang.

Rencana implementasi

- Peta Jalan: Kembangkan rencana terperinci yang menguraikan langkah-langkah yang diperlukan untuk mencapai modernisasi AI generatif.
- Tonggak sejarah: Tentukan tonggak penting dan kiriman untuk melacak kemajuan.
- Manfaat: Rencana implementasi yang jelas memberikan arahan dan akuntabilitas, dan memfasilitasi pendekatan terstruktur untuk modernisasi AI generatif.

Pertimbangan dan prasyarat penilaian

Mulailah dengan kasus penggunaan yang jelas

Identifikasi masalah atau peluang bisnis tertentu yang dapat diatasi oleh AI generatif. Fokus pada kasus penggunaan yang selaras dengan tujuan bisnis strategis dan menawarkan manfaat yang terukur. Prioritaskan kasus penggunaan yang menargetkan tantangan yang sering dihadapi dalam organisasi untuk memastikan bahwa arsitektur solusi dapat berfungsi sebagai pola untuk beberapa skenario.

Memulai proses penilaian dengan pemahaman umum tentang aplikasi AI generatif potensial bermanfaat tetapi tidak wajib. [Kuesioner](#) yang disertakan dengan panduan ini mengakomodasi berbagai tingkat kesiapsiagaan, dari organisasi yang memiliki kasus penggunaan yang terdefinisi dengan baik hingga yang hanya memiliki gagasan luas. Proses penilaian berfungsi untuk:

- Sempurnakan dan klarifikasi ide kasus penggunaan awal ini.
- Identifikasi kasus penggunaan potensial baru.
- Kembangkan tujuan spesifik dan terukur untuk setiap kasus penggunaan.
- Menilai kelayakan dan dampak potensial dari setiap kasus penggunaan.

Mari kita pertimbangkan contoh hipotetis: Perusahaan jasa keuangan memutuskan untuk mengeksplorasi modernisasi AI generatif. Mereka mulai dengan ide luas untuk meningkatkan layanan pelanggan dan proses deteksi penipuan mereka.

- Penilaian awal: Kuesioner membantu mereka mengevaluasi sistem mereka saat ini, kualitas data, dan kesiapan organisasi untuk adopsi AI generatif.
- Penyempurnaan kasus penggunaan: Melalui proses penilaian, mereka menyempurnakan ide awal mereka menjadi dua kasus penggunaan spesifik:
 - Menerapkan chatbot bertenaga AI generatif untuk pertanyaan pelanggan
 - Menggunakan AI generatif untuk deteksi penipuan transaksi real-time
- Pengaturan tujuan: Untuk setiap kasus penggunaan, mereka menentukan tujuan tertentu:
 - Mengurangi waktu respons layanan pelanggan sebesar 40 persen dalam 6 bulan
 - Tingkatkan akurasi deteksi penipuan sebesar 20 persen dan kurangi false positive sebesar 15 persen

- Peregangan tujuan: Mereka juga menetapkan target ambisius ini:
 - Raih kepuasan pelanggan 80 persen dengan respons yang dibantu AI
 - Mengembangkan model deteksi penipuan prediktif yang mengidentifikasi pola penipuan baru
- Definisi MVP: Kuesioner membantu mereka menentukan MVP untuk setiap kasus penggunaan, dengan fokus pada fitur-fitur penting yang memberikan nilai langsung.
- Arsitektur target: Akhirnya, mereka mengembangkan arsitektur target yang mendukung satu atau kedua kasus penggunaan, dan memastikan skalabilitas dan integrasi dengan sistem yang ada.

Pastikan keselarasan bisnis

Selaraskan inisiatif AI generatif dengan strategi dan tujuan bisnis secara keseluruhan. Untuk setiap kasus penggunaan, kembangkan proposisi nilai yang jelas yang menunjukkan bagaimana AI generatif berkontribusi pada pertumbuhan, efisiensi, atau inovasi bisnis. Tetapkan metrik untuk mengukur dampak implementasi AI generatif pada indikator kinerja utama (). KPIs

Menerapkan tata kelola dan pengawasan

Buat komite pengarah lintas fungsi untuk mengawasi inisiatif AI generatif. Kembangkan kebijakan dan pedoman untuk penggunaan AI yang bertanggung jawab, mengatasi pertimbangan etis dan potensi bias. Menetapkan proses peninjauan untuk proyek AI generatif untuk memastikan kepatuhan dengan standar organisasi dan persyaratan peraturan.

Data alamat dan prasyarat teknis

Menilai dan meningkatkan kualitas data, dan menerapkan praktik tata kelola data untuk memastikan input yang andal untuk model AI generatif. Kembangkan strategi data yang membahas pengumpulan data, penyimpanan, dan manajemen yang khusus untuk kebutuhan AI generatif. Mengevaluasi dan meningkatkan infrastruktur data untuk mendukung volume dan kecepatan data yang diperlukan untuk beban kerja AI generatif.

Pertimbangkan persyaratan sumber daya komputasi

Menilai infrastruktur TI saat ini dan mengidentifikasi kesenjangan dalam kapasitas komputasi untuk beban kerja AI generatif. Rencanakan sumber daya komputasi yang dapat diskalakan, dengan mempertimbangkan opsi seperti layanan cloud atau kluster komputasi berkinerja tinggi lokal.

Optimalkan alokasi sumber daya untuk menyeimbangkan kinerja dan efektivitas biaya untuk beban kerja pelatihan dan inferensi.

Mengatasi implikasi privasi dan keamanan

Menerapkan langkah-langkah keamanan yang kuat untuk melindungi data sensitif yang digunakan dalam pelatihan dan operasi AI generatif. Pastikan kepatuhan terhadap peraturan perlindungan data seperti Peraturan Perlindungan Data Umum (GDPR) atau California Consumer Privacy Act (CCPA) saat menangani informasi pribadi. Kembangkan protokol untuk penerapan dan pemantauan model yang aman untuk mencegah akses yang tidak sah atau penyalahgunaan kemampuan AI generatif.

Libatkan pemangku kepentingan lebih awal

Libatkan pemangku kepentingan utama sejak awal untuk mendapatkan dukungan dan dukungan kepemimpinan. Komunikasikan dengan jelas manfaat dan dampak potensial dari inisiatif modernisasi, khususnya untuk beban kerja AI generatif. Memberikan pelatihan dan sumber daya untuk membantu para pemangku kepentingan memahami teknologi AI generatif dan implikasinya.

Iterasi dan pelajari

Mengadopsi pendekatan inkremental yang memungkinkan Anda menyempurnakan solusi target. Gunakan loop umpan balik untuk terus meningkatkan arsitektur dan proses beban kerja. Secara teratur menilai kinerja dan dampak implementasi AI generatif, dan sesuaikan strategi sesuai kebutuhan berdasarkan hasil dunia nyata dan kebutuhan bisnis yang berkembang.

Kuesioner penilaian beban kerja AI generatif

Bagian berikut memberikan pertanyaan yang dapat Anda gunakan untuk mengevaluasi berbagai aspek modernisasi beban kerja AI generatif untuk organisasi Anda. Kuesioner komprehensif ini mengevaluasi kesiapan organisasi Anda untuk mengadopsi dan menerapkan beban kerja AI generatif dengan pertanyaan di seluruh area utama, termasuk kasus penggunaan, arsitektur, penyimpanan, kepatuhan, integrasi, pengujian, penerapan, dan strategi data. Dengan menangani aspek-aspek penting dari implementasi AI generatif, mulai dari infrastruktur teknis hingga pertimbangan peraturan, kuesioner ini membantu Anda mengidentifikasi kekuatan, kesenjangan, dan peluang dalam perjalanan modernisasi AI Anda.

Bagian:

- [Kesiapan](#)
- [Kasus penggunaan](#)
- [Arsitektur](#)
- [Penyimpanan](#)
- [Peraturan dan kepatuhan](#)
- [Integrasi](#)
- [Pengujian](#)
- [Penerapan dan otomatisasi](#)
- [Strategi data](#)

Anda juga dapat mengunduh kuesioner dalam format Microsoft Excel dan menggunakannya untuk merekam informasi Anda.



[kuesioner](#)

Unduh

Kesiapan

Pertanyaan	Contoh respon
Apakah Anda memiliki AWS akun yang dapat dimanfaatkan untuk beban kerja ini?	Ya atau tidak.
Apakah Anda memiliki perjanjian perusahaan yang ada dengan AWS?	Ya atau tidak.
Seberapa skalabel infrastruktur cloud Anda saat ini untuk menangani beban kerja AI generatif?	Infrastruktur cloud kami sangat terukur, dengan kemampuan penskalaan otomatis untuk sumber daya komputasi dan sistem penyimpanan terdistribusi yang dirancang untuk menangani beban kerja AI generatif skala besar secara efisien.
Apakah Anda memiliki kemampuan pipa data untuk preprocessing dan rekayasa fitur dalam skala besar?	Pipa data kami menggunakan kerangka kerja pemrosesan terdistribusi seperti Apache Spark untuk preprocessing data skala besar dan rekayasa fitur, dengan dukungan untuk pemrosesan data batch dan streaming.
Apakah Anda memiliki kemampuan penyediaan dan manajemen akun?	Ya atau tidak.
Bagaimana Anda menggambarkan literasi AI organisasi Anda dan kesiapan untuk mengadopsi teknologi AI generatif?	Organisasi kami telah banyak berinvestasi dalam program pendidikan AI, dan sebagian besar staf teknis telah menyelesaikan pelatihan AI/ML dasar. Organisasi ini memiliki budaya inovasi yang merangkul teknologi baru, termasuk AI generatif.
Keahlian AI/ML apa yang ada dalam organisasi Anda, dan bagaimana cara mendistribusikannya?	Kami memiliki AI Center of Excellence yang berdedikasi dengan ilmuwan data berpengalaman dan insinyur ML. Kami meningkatkan keterampilan pakar domain di berbagai unit

Pertanyaan	Contoh respon
	bisnis untuk menjadi melek AI dan mengidentifikasi kasus penggunaan AI generatif.
Apakah Anda memiliki kasus bisnis tingkat tinggi yang mengartikulasikan tujuan, manfaat, dan biaya program cloud?	Ya atau tidak.
Apa garis waktu Anda untuk mengambil solusi untuk produksi?	Minggu, bulan, dan sebagainya.
Apakah komitmen pendanaan telah dibuat oleh pemangku kepentingan utama Anda (misalnya, CFO, CIT/CTO, COO)?	Ya atau tidak.
Bagaimana Anda memastikan kepatuhan terhadap peraturan perlindungan data dalam inisiatif AI generatif Anda?	Kami memiliki tim kepatuhan khusus yang bekerja sama dengan tim AI kami. Kami melakukan penilaian dampak privasi secara teratur, menerapkan perlindungan data berdasarkan prinsip desain, dan memelihara catatan pemrosesan data terperinci untuk semua proyek AI generatif.
Seberapa matang sistem Anda yang ada yang terintegrasi dengan teknologi AI generatif baru?	Arsitektur TI kami didasarkan pada layanan mikro dan APIs memungkinkan integrasi fleksibel teknologi AI generatif baru. Sistem ini distandarisasi pada format dan protokol data umum untuk memastikan interoperabilitas.
Pengalaman apa yang Anda miliki dalam mengoperasikan model ML, dan bagaimana hal ini dapat diterapkan pada sistem AI generatif?	Kami telah menetapkan MLOps praktik, termasuk jalur pipa penerapan model otomatis, sistem pemantauan, dan kerangka kerja pengujian A/B. Praktik ini sedang disesuaikan untuk menangani persyaratan unik model AI generatif skala besar.

Kasus penggunaan

Pertanyaan	Contoh respon
Apa tujuan utama atau kriteria keberhasilan kasus penggunaan?	Untuk meningkatkan waktu respons dukungan pelanggan, meningkatkan konversi penjualan , meningkatkan rekomendasi produk. Juga: Untuk meningkatkan kepuasan pengguna, tingkat penyelesaian tugas, kualitas respons, dan sebagainya.
Bagaimana kasus penggunaan ini selaras dengan tujuan strategis organisasi Anda?	Ini sejalan dengan tujuan strategis kami untuk meningkatkan kepuasan pelanggan dengan mengurangi waktu respons dalam layanan pelanggan.
Berapa volume data atau permintaan yang diharapkan untuk kasus penggunaan?	500 transaksi per detik (TPS).
Jenis sumber data apa yang diperlukan untuk mendukung beban kerja AI generatif Anda?	Database terstruktur internal (catatan pelanggan, data penjualan, dan sebagainya); data teks tidak terstruktur dari dokumen, email, dan media sosial; file audio dan video untuk tugas pengenalan ucapan dan gambar; data streaming waktu nyata dari perangkat dan sensor IoT; kumpulan data publik dan untuk pengayaan. APIs
Seberapa sering Anda perlu memperbarui atau menyegarkan data dari sumber-sumber ini?	Database transaksional: hampir pembaruan waktu nyata; repositori dokumen: pembaruan batch harian; umpan media sosial: pembaruan per jam; Data sensor IoT: streaming real-time berkelanjutan; kumpulan data publik: pembaruan bulanan atau triwulanan.
Format data apa yang dibutuhkan model AI generatif Anda sebagai input?	Data terstruktur: tabel database CSV, JSON, dan SQL; data teks: teks biasa, PDF, dan HTML; data gambar: JPEG, PNG, dan TIFF;

Pertanyaan	Contoh respon
	data audio: WAV dan; data video: dan AVI MP3. MP4
Apa masalah kualitas data utama Anda untuk beban kerja AI generatif?	Kelengkapan: memastikan bahwa tidak ada bidang kritis yang hilang; akurasi: memverifikasi kebenaran data dan menghilangkan kesalahan ; konsistensi: mempertahankan format dan nilai yang seragam di seluruh sumber; ketepatan waktu: memastikan bahwa data mutakhir untuk inferensi waktu nyata; relevansi: mengonfirmasi bahwa data selaras dengan tugas AI generatif tertentu.
Apa persyaratan kinerja utama (misalnya, waktu respons, throughput, akurasi)?	Akurasi 95%; < 500 ms waktu respons; kemampuan untuk menangani 1000 permintaan/detik. Akurasi tinggi (95% +), akurasi sedang (80-90%), upaya terbaik, dan sebagainya.
Apakah Anda memiliki yang lain KPIs untuk mengukur keberhasilan kasus penggunaan ini?	Kunci KPIs termasuk pengurangan tingkat kesalahan, penghematan waktu per transaksi, dan skor kepuasan pelanggan.
Berapa akurasi model yang diinginkan, dan bagaimana menyeimbangkannya dengan biaya?	Akurasi tinggi (> 90%) dengan biaya sedang, akurasi sedang (70-80%) dengan biaya rendah, dan sebagainya.
Apa kasus penggunaan utama atau skenario untuk solusi AI generatif?	Chatbot layanan pelanggan, pembuatan konten, rekomendasi produk, dan sebagainya.
Apa target pengguna atau persona untuk sistem AI generatif?	Agen layanan pelanggan, tim pemasaran, karyawan, pengguna akhir, dan sebagainya.
Berapa volume permintaan atau pengguna yang diharapkan?	1.000 permintaan per hari; 10.000 pengguna aktif bulanan.
Apakah ada batasan atau persyaratan kasus penggunaan khusus?	Respons real-time, dukungan multi-bahasa, privasi data, dan sebagainya.

Pertanyaan	Contoh respon
Apakah Anda memiliki anggaran yang dialokasikan untuk mengembangkan dan memelihara solusi AI generatif?	Biaya pengembangan awal diperkirakan \$200.000, dengan biaya pemeliharaan tahunan \$50.000.
Berapa proyeksi pengembalian investasi (ROI) dan periode pengembalian modal untuk kasus penggunaan ini?	ROI yang diharapkan sebesar 150% selama tiga tahun, dengan periode pengembalian modal 18 bulan.
Apakah ada biaya tersembunyi atau potensi penghematan yang harus dipertimbangkan?	Penghematan potensial termasuk pengurangan biaya lembur. Biaya tersembunyi mungkin melibatkan pelatihan tambahan untuk staf.
Apa skalabilitas dan kemungkinan ekspansi masa depan dari solusi AI generatif ini?	Solusi ini dirancang untuk skala dengan operasi kami, dengan kemungkinan memperluas ke departemen lain di masa depan.
Bagaimana Anda memastikan keadilan dan mengurangi bias dalam model AI generatif Anda?	Kami berencana untuk mengurangi bias melalui pengumpulan data yang beragam, audit bias reguler, dan implementasi teknik mitigasi bias.
Proses apa yang Anda miliki untuk mengatasi masalah etika atau konsekuensi yang tidak diinginkan?	Kami akan mengelola masalah etika melalui rencana respons insiden AI yang mapan, penilaian risiko etis reguler, sistem pelaporan anonim untuk karyawan, kolaborasi dengan pakar etika eksternal, dan pemantauan dan penyesuaian berkelanjutan dari model yang diterapkan berdasarkan umpan balik.

Pertanyaan	Contoh respon
Bagaimana Anda mendekati memprioritaskan dan mengurutkan penilaian beban kerja AI generatif di berbagai proyek dan departemen di organisasi Anda?	Dengan melakukan survei tingkat tinggi di semua departemen untuk mengidentifikasi potensi kasus penggunaan AI generatif dan mengevaluasinya berdasarkan tiga kriteria utama: dampak bisnis, kelayakan teknis, dan pertimbangan etis. Proyek dengan dampak potensial tinggi, hambatan teknis yang lebih rendah, dan masalah etika minimal diberikan prioritas.

Arsitektur

Pertanyaan	Contoh respon
Jenis model atau arsitektur AI generatif apa yang sedang dipertimbangkan?	Transformator, jaringan saraf konvolusional (CNN), jaringan saraf berulang (RNN), pohon keputusan, dan sebagainya.
Berapa skala atau volume data dan perhitungan yang diharapkan?	Jutaan pengguna, petabyte data, dan sebagainya.
Apa persyaratan perangkat keras (misalnya, CPUs atau GPUs) untuk pelatihan dan inferensi?	High-end GPUs, cluster CPU, instance cloud, dan sebagainya.
Bagaimana model AI generatif akan diperbarui atau dilatih ulang dari waktu ke waktu?	Melalui pembelajaran berkelanjutan, pelatihan ulang berkala, pembaruan manual, dan sebagainya.
Apa saja preprocessing data dan persyaratan rekayasa fitur?	Pembersihan teks, augmentasi gambar, pemilihan fitur, dan sebagainya.
Bagaimana sistem AI generatif menangani kasus tepi, outlier, atau input dengan kepercayaan rendah?	Melalui mundur ke pengawasan manusia, meminta klarifikasi, dan sebagainya.

Pertanyaan	Contoh respon
Apa persyaratan latensi untuk aplikasi AI generatif?	Real-time, mendekati waktu nyata, pemrosesan batch, dan sebagainya.

Penyimpanan

Pertanyaan	Contoh respon
Di mana data pelatihan akan disimpan?	Di penyimpanan cloud (misalnya, Amazon S3, penyimpanan file, penyimpanan blok, atau penyimpanan objek), di penyimpanan lokal, dan sebagainya.
Apa persyaratan penyimpanan untuk data pelatihan dan artefak model (misalnya, kapasitas, daya tahan, ketersediaan)?	Penyimpanan skala petabyte, daya tahan tinggi (daya tahan 99,999999999%), ketersediaan tinggi, dan sebagainya.
Apa persyaratan retensi dan cadangan data untuk data pelatihan dan artefak model?	Retensi data selama x tahun, backup harian, backup off-site, dan sebagainya.
Format file mana yang terutama digunakan untuk menyimpan kumpulan data pelatihan AI Anda (misalnya, CSV, JSON, Parquet,)? HDF5	File parquet untuk data terstruktur, dan HDF5 untuk array multidimensi besar dan data tidak terstruktur seperti gambar dan teks. Kami menggunakan format khusus seperti TFRecord untuk mengoptimalkan pemuatan data selama pelatihan.
Bagaimana kumpulan data pelatihan Anda diatur: sebagai file individual, dalam database, atau menggunakan format data AI khusus?	Kumpulan data kecil hingga menengah disimpan sebagai file Parquet individual dalam penyimpanan objek untuk fleksibilitas. Dataset besar disimpan dalam database terdistribusi (Cassandra) untuk menangani skala.
Apakah Anda menggunakan teknik kompresi atau pengkodean data khusus untuk data pelatihan AI generatif?	Untuk data tabular, kami menggunakan teknik pengkodean kamus dan bit-packing yang tersedia di Parquet. Untuk gambar, kami

Pertanyaan	Contoh respon
	menggunakan kompresi JPEG lossy dengan pengaturan kualitas yang dioptimalkan untuk model kami.
Bagaimana Anda menangani pembuatan versi dan penyimpanan berbagai iterasi kumpulan data pelatihan? Apa dampaknya terhadap kebutuhan penyimpanan Anda secara keseluruhan?	Kami menggunakan sistem versi data (DVC) yang terintegrasi dengan platform ML kami.

Peraturan dan kepatuhan

Pertanyaan	Contoh respon
Apa peraturan atau persyaratan kepatuhan yang relevan untuk solusi AI generatif (misalnya, GDPR, HIPAA, PCI-DSS)?	GDPR untuk menangani data pribadi, HIPAA untuk data perawatan kesehatan, PCI-DSS untuk data pembayaran, dan sebagainya.
Pedoman atau kerangka kerja AI generatif etis apa yang telah diadopsi organisasi Anda?	Kami menerapkan pedoman AI bertanggung jawab kami sendiri. Semua proyek AI generatif menjalani tinjauan etis sebelum persetujuan dan penerapan.
Apa persyaratan keamanan untuk sistem AI generatif?	Enkripsi data, komunikasi jaringan yang aman, audit keamanan reguler.
Apa persyaratan untuk privasi dan perlindungan data?	Anonimisasi data, enkripsi, kontrol akses, dan sebagainya.
Apa persyaratan solusi untuk menangani data sensitif atau rahasia?	Kontrol akses yang ketat, penyembunyian data, persyaratan residensi data, dan sebagainya.
Bagaimana otentikasi dan otorisasi pengguna ditangani?	Dengan menggunakan OAuth, kunci API, sistem masuk tunggal (SSO), dan kontrol akses berbasis peran (RBAC).

Pertanyaan	Contoh respon
Bagaimana solusi akan dipantau dan dikelola dalam produksi?	Dengan menggunakan alat pemantauan seperti Prometheus dan Datadog, alat logging seperti ELK Stack, sistem peringatan, dan sebagainya.

Integrasi

Pertanyaan	Contoh respon
Apa persyaratan untuk mengintegrasikan solusi AI generatif dengan sistem atau sumber data yang ada?	REST APIs, antrian pesan, konektor database, dan sebagainya.
Bagaimana data akan dicerna dan diproses sebelumnya untuk solusi AI generatif?	Dengan menggunakan pemrosesan batch, streaming data, transformasi data, dan rekayasa fitur.
Bagaimana output dari solusi AI generatif akan dikonsumsi atau diintegrasikan dengan sistem hilir?	Melalui endpoint API, antrian pesan, pembaruan database, dan sebagainya.
Pola integrasi berbasis peristiwa mana yang dapat digunakan untuk solusi AI generatif?	Antrian pesan (seperti Amazon SQS, Apache Kafka, RabbitMQ), pub/sub sistem, webhook, platform streaming acara.
Pendekatan integrasi berbasis API mana yang dapat digunakan untuk menghubungkan solusi AI generatif dengan sistem lain?	RESTful APIs, APIs GraphQL, APIs SOAP (untuk sistem warisan).
Komponen arsitektur layanan mikro mana yang dapat digunakan untuk integrasi solusi AI generatif?	Service mesh untuk komunikasi antar layanan, gateway API, orkestrasi kontainer (misalnya, Kubernetes).
Bagaimana integrasi hybrid dapat diterapkan untuk solusi AI generatif?	Dengan menggabungkan pola berbasis peristiwa untuk pembaruan waktu nyata,

Pertanyaan	Contoh respon
	pemrosesan batch untuk data historis, dan APIs untuk integrasi sistem eksternal.
Bagaimana output solusi AI generatif dapat diintegrasikan dengan sistem hilir?	Melalui titik akhir API, antrian pesan, pembaruan basis data, webhook, dan ekspor file.
Langkah-langkah keamanan mana yang harus dipertimbangkan untuk mengintegrasikan solusi AI generatif?	Mekanisme otentikasi (seperti OAuth atau JWT), enkripsi (dalam perjalanan dan saat istirahat), pembatasan laju API, dan daftar kontrol akses (). ACLs
Bagaimana Anda berencana untuk mengintegrasikan kerangka kerja open source seperti LlamaIndex atau LangChain ke dalam pipeline data yang ada dan alur kerja AI generatif?	Kami berencana menggunakannya LangChain untuk membangun aplikasi AI generatif yang kompleks, terutama untuk kemampuan agen dan manajemen memorinya. Kami bertujuan untuk memiliki 60% proyek AI generatif kami yang digunakan LangChain dalam 6 bulan ke depan.
Bagaimana Anda memastikan kompatibilitas antara kerangka kerja open source yang Anda pilih dan infrastruktur data yang ada?	Kami menciptakan tim integrasi khusus untuk memastikan kompatibilitas yang lancar. Pada kuartal ketiga, tujuan kami adalah memiliki pipa terintegrasi penuh yang digunakan LlamaIndex untuk pengindeksan dan pengambilan data yang efisien dalam struktur data lake kami saat ini.
Bagaimana Anda berencana untuk memanfaatkan komponen modular kerangka kerja seperti LangChain untuk pembuatan prototipe dan eksperimen cepat?	Kami sedang menyiapkan lingkungan kotak pasir di mana pengembang dapat dengan cepat membuat prototipe dengan menggunakan LangChain komponen.

Pertanyaan	Contoh respon
Apa strategi Anda untuk mengikuti pembaruan dan fitur baru dalam kerangka kerja open source yang berkembang pesat ini?	Kami telah menugaskan tim untuk memantau GitHub repositori dan forum komunitas untuk LangChain dan. LlamaIndex Kami berencana untuk mengevaluasi dan mengintegrasikan pembaruan besar setiap triwulan, dengan fokus pada peningkatan kinerja dan kemampuan baru.

Pengujian

Pertanyaan	Contoh respon
Apa persyaratan pengujian (misalnya, pengujian unit, pengujian integrasi, end-to-end pengujian)?	Unit pengujian untuk komponen individu, pengujian integrasi dengan sistem eksternal, end-to-end pengujian untuk skenario kritis, dan sebagainya.
Bagaimana Anda memastikan kualitas dan konsistensi data di berbagai sumber untuk pelatihan AI generatif?	Kami menjaga kualitas data melalui alat pembuatan profil data otomatis, audit data reguler, dan katalog data terpusat. Kami telah menerapkan kebijakan tata kelola data untuk memastikan konsistensi di seluruh sumber dan untuk mempertahankan garis keturunan data.
Bagaimana model AI generatif akan dievaluasi dan divalidasi?	Dengan menggunakan dataset holdout, evaluasi manusia, pengujian A/B, dan sebagainya.
Apa kriteria untuk mengevaluasi kinerja dan akurasi model AI generatif?	Presisi, ingatan, skor F1, kebingungan, evaluasi manusia, dan sebagainya.
Bagaimana kasus tepi dan kasus sudut diidentifikasi dan ditangani?	Dengan menggunakan rangkaian pengujian yang komprehensif, evaluasi manusia, pengujian permusuhan, dan sebagainya.

Pertanyaan	Contoh respon
Bagaimana Anda akan menguji potensi bias dalam model AI generatif?	Dengan menggunakan analisis paritas demografis, pengujian kesempatan yang sama, teknik de-biasing permusuhan, pengujian kontrafaktual, dan sebagainya.
Metrik mana yang akan digunakan untuk mengukur keadilan dalam output model?	Rasio dampak yang berbeda, peluang yang disamakan, paritas demografis, metrik keadilan individu, dan sebagainya.
Bagaimana Anda memastikan representasi yang beragam dalam kumpulan data pengujian Anda untuk deteksi bias?	Dengan menggunakan stratified sampling lintas kelompok demografis, kolaborasi dengan pakar keragaman, penggunaan data sintetis untuk mengisi kesenjangan, dan sebagainya.
Proses mana yang akan diterapkan untuk pemantauan berkelanjutan dari keadilan model pasca-penerapan?	Audit keadilan reguler, sistem deteksi bias otomatis, analisis umpan balik pengguna, pelatihan ulang berkala dengan kumpulan data yang diperbarui, dan sebagainya.
Bagaimana Anda akan mengatasi bias interseksional dalam model AI generatif?	Dengan menggunakan analisis keadilan interseksional, pengujian subkelompok, kolaborasi dengan pakar domain tentang interseksionalitas, dan sebagainya.
Bagaimana Anda akan menguji kinerja model di berbagai bahasa dan konteks budaya?	Dengan menggunakan set tes multibahasa, kolaborasi dengan pakar budaya, metrik keadilan lokal, studi perbandingan lintas budaya, dan sebagainya.

Penerapan dan otomatisasi

Pertanyaan	Contoh respon
Apa persyaratan untuk penskalaan dan penyeimbangan beban?	Perutean permintaan cerdas; sistem penskalaan otomatis; mengoptimalkan awal dingin yang

Pertanyaan	Contoh respon
	cepat dengan menggunakan teknik seperti caching model, pemuatan lambat, dan sistem penyimpanan terdistribusi; merancang sistem untuk menangani pola lalu lintas yang meledak dan tidak dapat diprediksi.
Apa persyaratan untuk memperbarui dan meluncurkan versi baru?	Penerapan biru/hijau, rilis kenari, pembaruan bergulir, dan sebagainya.
Apa persyaratan untuk pemulihan bencana dan kelangsungan bisnis?	Prosedur pencadangan dan pemulihan, mekanisme failover, konfigurasi ketersediaan tinggi, dan sebagainya.
Apa persyaratan untuk mengotomatiskan pelatihan, penyebaran, dan manajemen model AI generatif?	Pipa pelatihan otomatis, penerapan berkelanjutan, penskalaan otomatis, dan sebagainya.
Bagaimana model AI generatif akan diperbarui dan dilatih ulang saat data baru tersedia?	Melalui pelatihan ulang berkala, pembelajaran inkremental, pembelajaran transfer, dan sebagainya.
Apa persyaratan untuk mengotomatisasi pemantauan dan manajemen?	Peringatan otomatis, penskalaan otomatis, penyembuhan diri, dan sebagainya.
Apa lingkungan penerapan pilihan Anda untuk beban kerja AI generatif?	Pendekatan hybrid yang menggunakan AWS untuk pelatihan model dan infrastruktur lokal kami untuk inferensi guna memenuhi persyaratan residensi data.
Apakah ada platform cloud khusus yang Anda sukai untuk penerapan AI generatif?	Layanan AWS, khususnya Amazon SageMaker AI untuk pengembangan dan penerapan model, dan Amazon Bedrock untuk model pondasi.
Teknologi kontainerisasi apa yang Anda pertimbangkan untuk beban kerja AI generatif?	Kami ingin melakukan standarisasi pada kontainer Docker yang diatur dengan Kubernetes untuk memastikan portabilitas dan skalabilitas di seluruh lingkungan hybrid kami.

Pertanyaan	Contoh respon
Apakah Anda memiliki alat pilihan untuk CI/CD di saluran AI generatif Anda?	GitLab untuk kontrol versi dan pipa CI/CD, terintegrasi dengan Jenkins untuk pengujian dan penerapan otomatis.
Alat orkestrasi apa yang Anda pertimbangkan untuk mengelola alur kerja AI generatif?	Apache Airflow untuk orkestrasi alur kerja, terutama untuk pra-pemrosesan data dan jalur pelatihan model.
Apakah Anda memiliki persyaratan khusus untuk infrastruktur lokal untuk mendukung beban kerja AI generatif?	Kami berinvestasi di server berakselerasi GPU dan jaringan berkecepatan tinggi untuk mendukung beban kerja inferensi lokal.
Bagaimana Anda berencana untuk mengelola versi model dan penerapan di lingkungan yang berbeda?	Kami berencana untuk menggunakannya MLflow untuk pelacakan model dan pembuatan versi, dan mengintegrasikannya dengan infrastruktur Kubernetes kami untuk penerapan yang mulus di seluruh lingkungan.
Alat pemantauan dan observabilitas apa yang Anda pertimbangkan untuk penerapan AI generatif?	Prometheus untuk pengumpulan metrik dan Grafana untuk visualisasi, dengan solusi pencatatan khusus tambahan untuk pemantauan khusus model.
Bagaimana Anda menangani pergerakan dan sinkronisasi data dalam model penerapan hibrida?	Kami akan menggunakannya AWS DataSync untuk transfer data yang efisien antara penyimpanan lokal dan AWS, dengan pekerjaan sinkronisasi otomatis yang dijadwalkan berdasarkan siklus pelatihan kami.
Langkah-langkah keamanan apa yang Anda terapkan untuk penerapan AI generatif di berbagai lingkungan?	Kami akan menggunakan IAM untuk sumber daya cloud, terintegrasi dengan Active Directory lokal kami untuk mengimplementasikan end-to-end enkripsi dan segmentasi jaringan untuk mengamankan aliran data.

Strategi data

Pertanyaan	Contoh respon
Jenis data spesifik apa yang penting untuk beban kerja AI generatif Anda, dan berapa persentasenya yang saat ini dapat diakses?	Log panggilan pelanggan dan data ulasan produk sangat penting. Saat ini, 85% dari tipe data ini dapat diakses untuk proyek AI generatif kami.
Bagaimana Anda memastikan dan mengukur kualitas data Anda?	Kami telah menerapkan metrik kualitas data, termasuk kelengkapan, akurasi, konsistensi, dan ketepatan waktu. Kami menggunakan alat otomatis untuk menilai metrik ini secara teratur dan memiliki tim khusus untuk pembersihan dan pengayaan data.
Berapa persentase data Anda yang memenuhi standar kualitas Anda untuk penggunaan AI generatif?	Saat ini, 78% data kami memenuhi standar kualitas kami. Kami menargetkan 95% dalam 12 bulan ke depan melalui proses pembersihan data yang lebih baik.
Bagaimana Anda berencana untuk membangun kepercayaan tentang penggunaan data dalam AI generatif di antara para pemangku kepentingan Anda?	Kami menerapkan dewan etika AI, memberikan penjelasan yang jelas tentang keputusan AI, dan melakukan audit AI triwulanan untuk memastikan transparansi dan keadilan.
Seberapa komprehensif dokumentasi Anda untuk sumber data dan garis keturunan?	Kami memelihara katalog data terperinci yang mencakup metadata untuk semua sumber data kami, termasuk asal, frekuensi pembaruan, dan penggunaan. Kami menggunakan alat silsilah data untuk melacak bagaimana data mengalir dan berubah di seluruh sistem kami.
Bagaimana Anda memastikan keragaman dalam kumpulan data Anda untuk mencegah bias dalam model AI?	Kami secara aktif mengambil data dari beragam demografi dan secara teratur mengaudit kumpulan data kami untuk bias representasional. Kami juga menggunakan teknik pembuatan data sintetis untuk

Pertanyaan	Contoh respon
	menyeimbangkan kategori yang kurang terwakili.
Berapa kecepatan refresh data Anda untuk model AI generatif kritis, dan bagaimana Anda menentukan frekuensi ini?	Model kritis disegarkan setiap minggu. Frekuensi ini ditentukan oleh metrik kinerja pengujian A/B, dan kami bertujuan untuk tidak lebih dari 2% degradasi antara penyegaran.
Berapa banyak versi kumpulan data penting yang Anda pertahankan dan untuk berapa lama?	Kami mempertahankan lima versi terakhir dari setiap kumpulan data penting, dengan periode retensi 18 bulan untuk setiap versi.
Berapa banyak tim lintas fungsi yang terlibat dalam inisiatif AI generatif Anda dan memiliki akses ke data Anda?	Kami memiliki tiga tim lintas fungsi. Setiap tim mencakup ilmuwan data, pakar domain, ahli etika, dan analis bisnis.
Kebijakan dan praktik tata kelola data apa yang Anda miliki?	Kami memiliki komite tata kelola data lintas fungsi yang mengawasi kebijakan data kami. Kami telah menerapkan kontrol akses berbasis peran, skema klasifikasi data, dan audit reguler untuk memastikan kepatuhan terhadap kerangka tata kelola kami.
Tindakan apa yang Anda miliki untuk memastikan privasi data, mendapatkan persetujuan yang tepat, dan menjaga kerahasiaan?	Kami telah menerapkan kerangka kerja privasi data komprehensif yang selaras dengan GDPR dan CCPA. Ini termasuk mendapatkan persetujuan eksplisit untuk penggunaan data, menerapkan teknik anonimisasi data, dan penilaian dampak privasi reguler.
Berapa persentase kumpulan data pelatihan AI Anda yang telah diaudit untuk bias pada kuartal terakhir?	70% dari kumpulan data pelatihan AI kami diaudit untuk bias kuartal terakhir. Kami menerapkan alat deteksi bias otomatis untuk mencapai audit triwulanan 100%.

Pertanyaan	Contoh respon
Berapa kapasitas pemrosesan data Anda saat ini, dan berapa banyak proyek yang Anda butuhkan untuk beban kerja AI generatif masa depan?	Kapasitas kami saat ini adalah 10 TB/day. We project needing 30 TB/day dalam setahun dan meningkatkan infrastruktur kami untuk memenuhi permintaan ini.
Apa strategi Anda untuk menyeimbangkan privasi data dengan kebutuhan data model AI generatif?	Kami menerapkan teknik anonimisasi tingkat lanjut dan pembuatan data sintetis. Tujuan kami adalah untuk meningkatkan data yang dapat digunakan untuk AI sebesar 40% sekaligus mengurangi risiko privasi hingga 60% selama tahun depan.
Berapa persentase kumpulan data pembelajaran mesin (ML) Anda yang diberi label secara akurat, dan berapa tingkat akurasi target Anda?	Saat ini, 85% dari kumpulan data ML kami diberi label secara akurat. Kami menargetkan tingkat akurasi 95% dalam kuartal berikutnya dengan menggunakan teknik pelabelan manusia dan otomatis.

Menerjemahkan wawasan penilaian menjadi hasil yang dapat ditindaklanjuti

Bagian ini menyediakan kerangka kerja untuk menganalisis tanggapan kuesioner dan menggunakan wawasan tersebut untuk membentuk arsitektur target dan kiriman utama lainnya dari inisiatif modernisasi AI generatif. Kerangka kerja ini menjembatani kesenjangan antara pengumpulan dan implementasi data, dan memastikan bahwa penilaian secara langsung menginformasikan dan mendorong strategi modernisasi Anda.

Definisi arsitektur target:

- Gunakan tanggapan kuesioner untuk menginformasikan pemilihan layanan cloud dan desain pipa data.
- Pastikan bahwa desain arsitektur mendukung skalabilitas dan interoperabilitas seperti yang disorot dalam panduan.

Evaluasi kesiapan pelanggan:

- Menganalisis tanggapan kuesioner yang terkait dengan infrastruktur, proses, dan budaya organisasi saat ini.
- Identifikasi kesenjangan dan buat rencana untuk mengatasinya. Prioritaskan kesenjangan yang penting untuk keberhasilan MVP.

Gunakan tujuan kasing dan peregangan:

- Ekstrak masalah bisnis tertentu dari tanggapan kuesioner untuk menentukan tujuan kasus penggunaan yang jelas.
- Tetapkan tujuan yang selaras dengan visi jangka panjang organisasi Anda untuk modernisasi AI generatif.

Estimasi upaya:

- Gunakan data kuesioner untuk memperkirakan sumber daya, waktu, dan anggaran untuk MVP dan implementasi penuh.
- Buat pendekatan bertahap yang dimulai dengan MVP, dan uraikan fase berikutnya.

Kebutuhan pemberdayaan:

- Berdasarkan tanggapan kuesioner, identifikasi kesenjangan keterampilan dan kebutuhan pelatihan.
- Kembangkan rencana pelatihan yang mendukung kebutuhan MVP langsung dan adopsi AI generatif jangka panjang.

Rencana implementasi:

- Buat peta jalan komprehensif yang dimulai dengan MVP dan uraikan langkah-langkah menuju modernisasi AI generatif penuh.
- Tentukan tonggak dan kiriman yang jelas untuk setiap fase implementasi.

Langkah-langkah praktis:

- Matriks prioritas: Buat matriks yang memetakan tanggapan kuesioner terhadap [enam hasil](#) untuk membantu memprioritaskan fitur dan upaya.
- Pendekatan iteratif: Rancang MVP untuk menjadi iterasi pertama dalam serangkaian rilis yang direncanakan, di mana setiap rilis dibangun menuju arsitektur target penuh.
- Penyelarasan pemangku kepentingan: Gunakan hasil kuesioner untuk menyelaraskan pemangku kepentingan pada lingkup MVP dan pendekatan bertahap untuk mencapai semua hasil.
- Loop umpan balik berkelanjutan: Menerapkan mekanisme untuk mengumpulkan umpan balik setelah penerapan MVP, dan menggunakan wawasan untuk menyempurnakan rencana untuk fase berikutnya.
- Implementasi tangkas: Mengadopsi metodologi tangkas yang memungkinkan fleksibilitas dalam menangani semua hasil dari waktu ke waktu, dimulai dengan hasil paling kritis dalam MVP.

Langkah selanjutnya

Setelah Anda menyelesaikan penilaian beban kerja AI generatif, ikuti langkah-langkah berikut:

1. Memberikan arsitektur target yang terperinci

- Tujuan: Arsitek solusi menciptakan arsitektur target komprehensif yang selaras dengan tujuan organisasi dan hasil penilaian.
- Komponen: Arsitektur ini mencakup desain konsumsi data, titik integrasi, dan interoperabilitas sistem untuk memastikan skalabilitas, keandalan, dan optimalisasi kinerja.

2. Jelaskan seberapa spesifik Layanan AWS sesuai dengan kasus penggunaan

- Pemetaan layanan: Identifikasi dan petakan spesifik Layanan AWS yang paling sesuai dengan kasus penggunaan yang diidentifikasi.
- Manfaat: Sorot bagaimana layanan ini memenuhi kebutuhan bisnis tertentu, meningkatkan efisiensi, dan memberikan skalabilitas.

3. Berikan solusi alternatif opsional dengan pro dan kontra

- Alternatif: Menyajikan solusi alternatif yang juga dapat memenuhi persyaratan organisasi.
- Analisis: Menawarkan analisis terperinci tentang keuntungan dan kerugian dari setiap alternatif dengan mempertimbangkan faktor-faktor seperti biaya, kompleksitas, dan keselarasan dengan tujuan bisnis.

4. Memberikan estimasi harga rinci Layanan AWS

- Analisis biaya: Memberikan estimasi biaya terperinci untuk yang diusulkan Layanan AWS, termasuk skenario penggunaan potensial dan model penetapan harga.
- Penyelarasan anggaran: Pastikan bahwa biaya sejalan dengan kendala anggaran organisasi dan memberikan pemahaman yang jelas tentang implikasi keuangan.

5. Dapatkan umpan balik tentang arsitektur yang diusulkan

- Keterlibatan pemangku kepentingan: Terlibat dengan pemangku kepentingan untuk mempresentasikan arsitektur yang diusulkan dan mengumpulkan umpan balik.
- Peningkatan berulang: Gunakan umpan balik untuk menyempurnakan dan meningkatkan solusi, dan konfirmasi bahwa itu memenuhi kebutuhan dan harapan semua pemangku kepentingan.

Pertanyaan yang Sering Diajukan

Apa tujuan utama penilaian beban kerja AI generatif?

Tujuan utama dari penilaian ini adalah untuk mengevaluasi kesiapan organisasi untuk memodernisasi beban kerja AI generatif mereka, mengidentifikasi kasus penggunaan, dan mengembangkan arsitektur solusi target. Ini bertujuan untuk menentukan persyaratan modernisasi, menentukan ruang lingkup implementasi, dan mempersiapkan modernisasi AI generatif yang sukses.

Siapa yang harus menggunakan penilaian ini?

Penilaian ini untuk arsitek solusi, arsitek perusahaan, dan arsitek aplikasi yang ingin menilai aspek teknis modernisasi AI generatif. Hal ini juga berguna bagi manajer program dan manajer orang untuk mengukur kesiapan keseluruhan, alokasi sumber daya, dan kebutuhan pemberdayaan.

Apa komponen kunci yang dievaluasi dalam penilaian?

Penilaian tersebut mencakup kesiapan keseluruhan, kasus penggunaan, arsitektur, penyimpanan, peraturan dan kepatuhan, integrasi, pengujian, otomatisasi penyebaran, dan strategi data. Komponen-komponen ini sangat penting untuk menentukan kesiapan teknis dan organisasi untuk adopsi modernisasi AI generatif.

Bagaimana penilaian membantu menentukan arsitektur target?

Penilaian memberikan pendekatan terstruktur untuk mengevaluasi sistem saat ini dan mengidentifikasi perbaikan. Ini membantu Anda memilih teknologi yang tepat dan merancang arsitektur terukur yang selaras dengan tujuan bisnis dan persyaratan kasus penggunaan.

Apa manfaat melakukan penilaian beban kerja AI generatif?

Manfaat termasuk peningkatan efisiensi, peningkatan pengambilan keputusan, jaminan kepatuhan, pembinaan inovasi, dan persiapan skalabilitas. Penilaian tersebut menetapkan pendekatan strategis untuk modernisasi AI generatif, dan memaksimalkan potensi manfaat sambil mengurangi risiko.

Bagaimana organisasi dapat memastikan keberhasilan implementasi setelah penilaian?

Organizations harus mengembangkan rencana implementasi yang jelas yang mencakup tonggak yang ditentukan, melibatkan pemangku kepentingan lebih awal, dan mengadopsi pendekatan berulang. Mendirikan Center of Excellence (CoE) dan berfokus pada pengembangan bakat juga direkomendasikan praktik terbaik.

Tantangan apa yang mungkin dihadapi organisasi selama penilaian?

Tantangan mungkin termasuk resistensi terhadap perubahan, masalah kualitas data, dan kompleksitas kepatuhan. Mengatasi tantangan ini membutuhkan pengembangan budaya inovasi, memastikan kesiapan data, dan menerapkan langkah-langkah keamanan yang kuat.

Bagaimana penilaian menangani persyaratan peraturan dan kepatuhan?

Penilaian mengevaluasi langkah-langkah kepatuhan saat ini dan mengidentifikasi kesenjangan. Ini memastikan bahwa solusi target mematuhi peraturan yang relevan dan undang-undang privasi data, dan menggabungkan praktik terbaik keamanan untuk melindungi informasi sensitif.

Peran apa yang dimainkan oleh keterlibatan pemangku kepentingan dalam proses penilaian?

Keterlibatan pemangku kepentingan sangat penting untuk mendapatkan dukungan, menyelaraskan inisiatif modernisasi dengan tujuan bisnis, dan memastikan implementasi yang berhasil. Keterlibatan awal dan komunikasi manfaat yang jelas adalah kunci untuk mengatasi resistensi dan mendorong dukungan.

Bagaimana organisasi dapat mengukur keberhasilan inisiatif modernisasi AI generatif mereka setelah penilaian?

Sukses dapat diukur dengan menggunakan indikator kinerja utama (KPIs) yang selaras dengan tujuan bisnis. Pemantauan dan evaluasi metrik ini secara teratur membantu memandu pengambilan keputusan dan menunjukkan nilai modernisasi AI generatif kepada para pemangku kepentingan.

Bagaimana pendekatan penilaian berbeda untuk organisasi dengan berbagai ukuran (kecil, menengah, atau perusahaan) atau industri?

Organisasi kecil:

- Mungkin memiliki sumber daya dan keahlian yang terbatas untuk penilaian komprehensif
- Kemungkinan untuk fokus pada kasus penggunaan berdampak tinggi tertentu daripada adopsi di seluruh perusahaan
- Mungkin lebih bergantung pada alat dan layanan pihak ketiga untuk penilaian
- Proses penilaian mungkin kurang formal dan lebih gesit

Organisasi menengah:

- Seringkali memiliki tim TI atau data khusus tetapi mungkin tidak memiliki keahlian AI khusus
- Mungkin mengambil pendekatan bertahap, dimulai dengan proyek percontohan di departemen utama
- Perlu menyeimbangkan inovasi dengan sistem dan proses yang ada
- Penilaian kemungkinan melibatkan tim lintas fungsi

Organisasi perusahaan:

- Biasanya memiliki tim AI/ML yang berdedikasi dan lebih banyak sumber daya untuk penilaian komprehensif
- Perlu mempertimbangkan integrasi kompleks dengan sistem perusahaan yang ada
- Mungkin memiliki persyaratan peraturan khusus industri untuk dipertimbangkan
- Penilaian sering melibatkan proses tata kelola formal

Sumber daya

- [AI generatif aktif AWS](#)
- [AWS menawarkan kecerdasan buatan baru, pembelajaran mesin, dan panduan AI generatif untuk merencanakan strategi AI Anda](#) (posting AWS blog)
- [Praktik terbaik untuk membangun aplikasi AI generatif di AWS](#) (posting AWS blog)
- [Pembuat Aplikasi AI Generatif di AWS](#) (Perpustakaan AWS Solusi)
- [Kemampuan AI generatif](#) (Arsitektur Referensi AWS Keamanan)
- [AWS kerangka kerja praktik terbaik AI generatif](#) (Panduan AWS Audit Manager Pengguna)
- [Memilih layanan AI generatif](#) (panduan AWS keputusan)
- [Apa itu Amazon Bedrock?](#) (Panduan Pengguna Amazon Bedrock)
- [Apa itu Amazon SageMaker AI?](#) (Panduan Pengembang Amazon SageMaker AI)

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	—	6 November 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 and 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 menyebabkan kerugian bagi 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 login 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

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 biasanya digambarkan sebagai pipa. CI/CD dapat 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 kekritisannya 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 di tempat. 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 dapat diakses oleh pengguna akhir. Dalam sebuah CI/CD pipeline, 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 dari 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 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 (OU). 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. 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 pengoptimalan 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 pengenalan 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 prinsip 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 kredensial.

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 kredensial pengguna, yang Anda simpan dalam bentuk terenkripsi. Ini terdiri dari nilai rahasia dan metadatanya. 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 kredensial.

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 waktu kerja 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 (tembakkan) 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.