Evaluación generativa de la carga de trabajo de IA

AWS Guía prescriptiva

AWS Guía prescriptiva: Evaluación generativa de la carga de trabajo de IA

Table of Contents

Evaluación generativa de la carga de trabajo de IA

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

Noviembre de 2024 (historial del documento)

La evaluación de la carga de trabajo de IA generativa es un método estratégico destinado a evaluar y mejorar la preparación de una organización para crear o actualizar sus cargas de trabajo de IA generativa. Esta evaluación es importante porque la incorporación de la IA generativa en las operaciones empresariales puede cambiar considerablemente el funcionamiento de las cosas y proporcionar nuevas eficiencias y capacidades. Sin embargo, para adoptar la IA generativa con éxito, es esencial comprender a fondo los sistemas actuales y tener un plan claro para el futuro.

Las cargas de trabajo de la IA generativa se refieren a las tareas computacionales que implican el uso de modelos de inteligencia artificial que pueden crear contenido nuevo, como texto, imágenes, código u otros tipos de datos. Estas cargas de trabajo suelen requerir una potencia informática considerable, hardware especializado y grandes conjuntos de datos para el entrenamiento y la inferencia. GPUs La integración de las cargas de trabajo generativas de IA en las operaciones presenta varios desafíos:

- Requisitos de infraestructura: aprovisionar los importantes recursos computacionales y el hardware especializado que requieren los modelos de IA generativa.
- Gestión de datos: garantizar la calidad, la privacidad y el cumplimiento de los datos al gestionar grandes conjuntos de datos.
- Falta de habilidades: falta de experiencia en tecnologías de IA y despliegue de modelos.
- Consideraciones éticas: abordar los prejuicios, la imparcialidad y la transparencia en el contenido generado por la IA.
- Complejidad de la integración: incorpora sin problemas la IA generativa a los flujos de trabajo existentes y a los sistemas heredados.
- Gestión de costes: equilibrar los beneficios potenciales con los altos costes de implementación y operación.

Superar estos desafíos requiere una planificación cuidadosa, una inversión en infraestructura y talento, y un enfoque estratégico de la implementación.

Propósito de esta guía

La IA generativa se está convirtiendo rápidamente en un componente fundamental en muchos sectores. Ofrece oportunidades transformadoras, pero también plantea desafíos en términos de integración, cumplimiento y escalabilidad. Muchas organizaciones tienen dificultades para aprovechar al máximo la IA debido a la debilidad de la base tecnológica, la resistencia al cambio y los problemas de calidad de los datos. La evaluación generativa de la carga de trabajo de la IA aborda estos desafíos al identificar los requisitos de modernización, definir el alcance de la implementación y cuestionar los sistemas y el pensamiento heredados. También ayuda a determinar los productos mínimos viables (MVPs) y le ayuda a desarrollar una arquitectura de solución específica, lo que garantiza un enfoque estructurado y estratégico para la adopción de la IA.

Esta guía sirve como un enfoque estructurado para ayudar a las organizaciones a sortear las complejidades de la adopción de tecnologías de IA generativa. En lugar de definir claramente los requisitos desde el principio, la guía ayuda a:

- Identificar posibles casos de uso de la IA generativa en su organización.
- Evaluar la preparación de su organización para la adopción de la IA generativa.
- Definir y refinar los objetivos de los casos de uso y los objetivos ambiciosos.
- Determinar el alcance y los requisitos de la implementación de la IA generativa.
- Desarrollar una arquitectura de solución objetivo.

Público objetivo y beneficios

Esta evaluación está diseñada específicamente para arquitectos de soluciones, arquitectos empresariales y arquitectos de aplicaciones que desean evaluar los aspectos técnicos de la modernización generativa de las cargas de trabajo de IA. También es útil para los gerentes de programas y personal que desean evaluar la preparación general de su equipo, la asignación de recursos y los requisitos de habilitación. Las mejores prácticas del sector hacen hincapié en la importancia de una evaluación exhaustiva para garantizar que estén preparados para la adopción de la IA. Esto incluye evaluar la arquitectura, el almacenamiento, el cumplimiento, la integración, las pruebas, el despliegue y la automatización.

Alcance

El método de evaluación generativa de la carga de trabajo de la IA incluye los siguientes temas:

- Tecnologías y modelos actuales de IA generativa (por ejemplo, modelos de lenguaje de gran tamaño, modelos de generación de imágenes)
- Aplicaciones de IA restringidas que utilizan técnicas generativas
- Integración de la IA generativa con los sistemas y flujos de trabajo existentes
- Estrategias de datos para entrenar y ajustar los modelos de IA generativa
- Consideraciones éticas y prácticas de IA responsables para las aplicaciones actuales de IA generativa
- Estrategias de prueba e implementación de la IA generativa en entornos de producción
- Consideraciones de seguridad y privacidad para las implementaciones de IA generativa
- Optimización del rendimiento y escalabilidad de las cargas de trabajo de IA generativa
- Casos de uso y aplicaciones de la IA generativa en varios sectores
- Evaluación de los resultados de la IA generativa y de los procesos de control de calidad

Los siguientes temas están fuera del ámbito de aplicación:

- Escenarios de inteligencia artificial general (AGI) y superinteligencia artificial (ASI)
- Los avances especulativos del futuro en la IA van más allá de los modelos generativos actuales
- Aplicaciones de computación cuántica en IA
- Computación neuromórfica e interfaces cerebro-ordenador
- Conciencia y autoconciencia en los sistemas de IA
- Los impactos sociales a largo plazo de la IA avanzada van más allá de las aplicaciones generativas actuales de la IA
- Marcos regulatorios para hipotéticas tecnologías de IA del futuro
- Debates filosóficos sobre la naturaleza de la inteligencia y la conciencia en las máquinas
- Casos extremos o casos de uso de la IA altamente especulativos
- Especificaciones técnicas detalladas de modelos o arquitecturas de IA patentados

Resultados empresariales específicos

La evaluación de la carga de trabajo de la IA generativa tiene como objetivo ofrecer varios resultados específicos que son cruciales para modernizar con éxito las cargas de trabajo de la IA generativa. Estos resultados garantizan que las organizaciones estén bien preparadas para integrar las tecnologías de IA de forma eficaz y eficiente.

Para cada resultado objetivo, la evaluación generativa de la carga de trabajo de la IA se centra en:

- Interdependencias: identifique y aclare cualquier interdependencia entre el resultado y otros aspectos del proceso de modernización. Esto incluye comprender cómo un resultado puede influir o ser influenciado por otros, para garantizar un enfoque holístico de la modernización.
- Alineación de las partes interesadas: describa las estrategias para alinear a las diversas partes interesadas con cada resultado. Esto implica comunicar el valor y el impacto de cada resultado a los diferentes niveles y departamentos de la organización, a fin de fomentar la aceptación y el apoyo.
- Priorización: en los casos en que se identifiquen varios casos de uso o resultados, proporcione un marco para priorizarlos en función de factores como el impacto empresarial, los requisitos de recursos y la alineación estratégica.
- Mejora continua: para cada resultado, establezca mecanismos de evaluación y perfeccionamiento continuos. Esto garantiza que los esfuerzos de modernización sigan adaptándose y respondiendo a los cambios en el panorama tecnológico y las necesidades empresariales.

A continuación, se presenta un análisis detallado de cada uno de los resultados previstos:

Arquitectura de destino

- Definición: La evaluación ayuda a definir una arquitectura objetivo clara y escalable para las cargas de trabajo generativas de IA.
- Componentes: Esto incluye seleccionar los servicios en la nube adecuados, diseñar las canalizaciones de datos y garantizar la interoperabilidad del sistema.
- Ventajas: una arquitectura bien definida respalda la escalabilidad, la confiabilidad y la optimización del rendimiento, y proporciona una base sólida para la modernización.

Preparación para el cliente

- Evaluación: evalúe el estado actual de la infraestructura, los procesos y la cultura de la organización para determinar si está preparada para la adopción generativa de la modernización de la IA.
- Criterios: Esto implica evaluar las capacidades técnicas, la calidad de los datos y la voluntad de la organización de aceptar el cambio.
- Resultado: La identificación de las brechas y las áreas de mejora garantiza que la organización esté preparada para una transición sin problemas a las soluciones y tecnologías modernas.

Metas basadas en casos de uso y objetivos ambiciosos

- Los objetivos de los casos de uso establecen objetivos claros para la implementación de la solución objetivo, centrándose en problemas u oportunidades empresariales específicos.

  Un objetivo de caso de uso en el contexto de la modernización de la IA generativa se refiere a un objetivo específico y mensurable que una organización pretende alcanzar mediante la implementación de soluciones de IA generativa. Estos objetivos suelen estar alineados con objetivos empresariales más amplios y se centran en abordar desafíos u oportunidades particulares dentro de la organización. Algunos ejemplos de objetivos de casos de uso pueden incluir:

  - Reducir el tiempo de respuesta del servicio de atención al cliente en un 50 por ciento mediante el uso de chatbots generativos impulsados por IA.
  - Mejorar la eficiencia de la revisión de código en un 30 por ciento mediante el análisis generativo de código asistido por IA.
  - Mejorando la precisión de la detección de fraudes en un 25 por ciento mediante el reconocimiento generativo de patrones mediante IA.
- Los objetivos ambiciosos definen objetivos ambiciosos que amplían los límites de lo que la modernización generativa de la IA puede lograr en la organización.
- Impacto: establecer objetivos alcanzables y ambiciosos ayuda a alinear las iniciativas generativas de modernización de la IA con los objetivos empresariales estratégicos y fomenta la innovación.

Estimación del esfuerzo

- Propósito: La estimación precisa del esfuerzo ayuda a planificar los recursos y garantiza que los proyectos se entreguen a tiempo y dentro del presupuesto.

- Alcance: Calcule los recursos, el tiempo y el presupuesto necesarios para implementar el plan de modernización de la IA generativa.
- Factores: tenga en cuenta la complejidad técnica, los desafíos de integración y los posibles riesgos.

Necesidades de habilitación

- Capacitación y desarrollo: identifique las habilidades y los conocimientos necesarios para adoptar con éxito la modernización de la IA generativa.
- Recursos: determine la necesidad de programas de capacitación, talleres y otras actividades de habilitación.
- Resultado: garantizar que el personal cuente con las habilidades necesarias mejora la eficacia de las iniciativas generativas de modernización de la IA y contribuye al éxito a largo plazo.

Plan de implementación

- Hoja de ruta: desarrolle un plan detallado que describa los pasos necesarios para lograr la modernización generativa de la IA.
- Hitos: defina los hitos y resultados clave para hacer un seguimiento del progreso.
- Beneficios: Un plan de implementación claro proporciona orientación y responsabilidad, y facilita un enfoque estructurado para la modernización generativa de la IA.

Consideraciones y requisitos previos de la evaluación

Comience con casos de uso claros

Identifique problemas u oportunidades empresariales específicos que la IA generativa pueda abordar. Céntrese en los casos de uso que se ajusten a los objetivos empresariales estratégicos y ofrezcan beneficios cuantificables. Priorice los casos de uso que se centren en los desafíos más comunes de la organización para garantizar que la arquitectura de la solución pueda servir de patrón para múltiples escenarios.

Iniciar el proceso de evaluación con un conocimiento general de las posibles aplicaciones generativas de la IA es beneficioso, pero no obligatorio. El cuestionario que se incluye en esta guía se adapta a varios niveles de preparación, desde organizaciones que tienen casos de uso bien definidos hasta aquellas que solo tienen ideas generales. El proceso de evaluación sirve para:

- Perfeccione y aclare estas ideas iniciales de casos de uso.
- Identifique nuevos casos de uso potenciales.
- Desarrolle objetivos específicos y medibles para cada caso de uso.
- Evalúe la viabilidad y el impacto potencial de cada caso de uso.

Consideremos un ejemplo hipotético: una empresa de servicios financieros decide explorar la modernización de la IA generativa. Empiezan con una idea amplia de mejorar su servicio al cliente y sus procesos de detección de fraudes.

- Evaluación inicial: el cuestionario les ayuda a evaluar sus sistemas actuales, la calidad de los datos y la preparación organizativa para la adopción generativa de la IA.
- Perfeccionamiento de los casos de uso: a lo largo del proceso de evaluación, refinan sus ideas iniciales en dos casos de uso específicos:
  - Implementación de un chatbot generativo impulsado por IA para las consultas de los clientes
  - Uso de IA generativa para la detección de fraudes en transacciones en tiempo real
- Establecimiento de objetivos: para cada caso de uso, definen objetivos específicos:
  - Reduzca el tiempo de respuesta del servicio de atención al cliente en un 40 por ciento en un plazo de 6 meses
  - Mejore la precisión de la detección de fraudes en un 20 por ciento y reduzca los falsos positivos en un 15 por ciento

- Objetivos ambiciosos: también establecen estos ambiciosos objetivos:
  - Logre un 80 por ciento de satisfacción de los clientes con respuestas asistidas por IA
  - Desarrolle un modelo predictivo de detección de fraudes que identifique nuevos patrones de fraude
- Definición de MVP: el cuestionario les ayuda a determinar un MVP para cada caso de uso, centrándose en las características esenciales que ofrecen un valor inmediato.
- Arquitectura objetivo: Por último, desarrollan una arquitectura objetivo que admite uno o ambos casos de uso y garantiza la escalabilidad y la integración con los sistemas existentes.

Garantice la alineación empresarial

Alinee las iniciativas generativas de IA con la estrategia y los objetivos empresariales generales. Para cada caso de uso, desarrolle una propuesta de valor clara que demuestre cómo la IA generativa contribuye al crecimiento, la eficiencia o la innovación empresarial. Establezca métricas para medir el impacto de las implementaciones de IA generativa en los indicadores clave de rendimiento (). KPIs

Implemente la gobernanza y la supervisión

Cree un comité directivo multifuncional para supervisar las iniciativas de IA generativa. Desarrolle políticas y directrices para un uso responsable de la IA, abordando las consideraciones éticas y los posibles sesgos. Establezca un proceso de revisión para los proyectos de IA generativa a fin de garantizar el cumplimiento de las normas organizativas y los requisitos reglamentarios.

Aborde los requisitos técnicos y de datos

Evalúe y mejore la calidad de los datos e implemente prácticas de gobierno de datos para garantizar la fiabilidad de los insumos para los modelos de IA generativa. Desarrolle una estrategia de datos que aborde la recopilación, el almacenamiento y la gestión de datos específicos de las necesidades generativas de inteligencia artificial. Evalúe y mejore la infraestructura de datos para soportar el volumen y la velocidad de los datos necesarios para las cargas de trabajo de IA generativa.

Tenga en cuenta los requisitos de recursos informáticos

Evalúe la infraestructura de TI actual e identifique las brechas en la capacidad computacional para las cargas de trabajo generativas de IA. Planifique recursos informáticos escalables, considerando

opciones como los servicios en la nube o los clústeres de computación de alto rendimiento locales. Optimice la asignación de recursos para equilibrar el rendimiento y la rentabilidad de las cargas de trabajo de formación y de inferencia.

Aborde las implicaciones de privacidad y seguridad

Implemente medidas de seguridad sólidas para proteger los datos confidenciales que se utilizan en la formación y las operaciones de IA generativa. Garantice el cumplimiento de las normas de protección de datos, como el Reglamento General de Protección de Datos (GDPR) o la Ley de Privacidad del Consumidor de California (CCPA), al gestionar la información personal. Desarrolle protocolos para el despliegue y la supervisión seguros de los modelos a fin de evitar el acceso no autorizado o el uso indebido de las capacidades de IA generativa.

Interactúe pronto con las partes interesadas

Involucre a las principales partes interesadas desde el principio para lograr la aceptación y el apoyo de los líderes. Comunique con claridad los beneficios y el impacto potencial de las iniciativas de modernización, específicamente para las cargas de trabajo generativas de IA. Proporcione formación y recursos para ayudar a las partes interesadas a comprender las tecnologías de IA generativa y sus implicaciones.

Repite y aprende

Adopte un enfoque gradual que le permita refinar las soluciones objetivo. Utilice los circuitos de retroalimentación para mejorar continuamente la arquitectura y los procesos de la carga de trabajo. Evalúe periódicamente el rendimiento y el impacto de las implementaciones de IA generativa y ajuste las estrategias según sea necesario en función de los resultados del mundo real y de las necesidades empresariales en evolución.

Cuestionario de evaluación generativa de carga de trabajo de IA

En las siguientes secciones se incluyen preguntas que puede utilizar para evaluar diferentes aspectos de la modernización de la carga de trabajo de IA generativa para su organización. Este exhaustivo cuestionario evalúa la preparación de su organización para adoptar e implementar cargas de trabajo de IA generativa con preguntas sobre áreas clave, como los casos de uso, la arquitectura, el almacenamiento, el cumplimiento, la integración, las pruebas, el despliegue y la estrategia de datos. Al abordar los aspectos fundamentales de la implementación de la IA generativa, desde la infraestructura técnica hasta las consideraciones regulatorias, este cuestionario le ayuda a identificar los puntos fuertes, las brechas y las oportunidades en su proceso de modernización de la IA.

Secciones:

También puede descargar el cuestionario en formato Microsoft Excel y utilizarlo para registrar su información.

Descargar el cuestionario

Preparar

Pregunta	Ejemplo de respuesta
¿Tiene AWS cuentas que puedan utilizarse para estas cargas de trabajo?	Sí o no.
¿Tiene un acuerdo empresarial vigente con usted AWS?	Sí o no.
¿Hasta qué punto es escalable su infraestr uctura de nube actual para gestionar las cargas de trabajo generativas de IA?	Nuestra infraestructura de nube es altamente escalable, con capacidades de escalado automático para los recursos de cómputo y los sistemas de almacenamiento distribuido que están diseñados para gestionar de manera eficiente las cargas de trabajo de IA generativa a gran escala.
¿Cuenta con capacidades de canalización de datos para el preprocesamiento y la ingeniería de características a escala?	Nuestros flujos de datos utilizan marcos de procesamiento distribuido, como Apache Spark, para el preprocesamiento de datos a gran escala y la ingeniería de funciones, con soporte para el procesamiento de datos por lotes y en streaming.
¿Tiene capacidad de aprovisionamiento y administración de cuentas?	Sí o no.
¿Cómo describiría los conocimientos de IA de su organización y su disposición a adoptar tecnologías de IA generativa?	Nuestra organización ha realizado grandes inversiones en programas educativos sobre IA y la mayoría del personal técnico ha completad o una formación básica en IA y aprendizaje automático. La organización tiene una cultura de innovación que abarca las nuevas tecnologí as, incluida la IA generativa.

Pregunta	Ejemplo de respuesta
¿Qué experiencia en inteligencia artificial y aprendizaje automático existe en su organizac ión y cómo se distribuye?	Tenemos un centro de excelencia dedicado a la IA con científicos de datos e ingenieros de aprendizaje automático con experiencia. Capacitamos a expertos en diferentes unidades de negocio para que adquieran conocimientos de IA e identifiquen casos de uso generativos de IA.
¿Tiene un modelo de negocio de alto nivel que articule los objetivos, los beneficios y el costo del programa de nube?	Sí o no.
¿Cuál es su plazo para llevar la solución a producción?	Semanas, meses, etc.
¿Sus principales partes interesadas (por ejemplo, el director financiero, el CIT/CTO o el director de operaciones) han asumido un compromiso de financiación?	Sí o no.
¿Cómo garantiza el cumplimiento de las normas de protección de datos en sus iniciativ as de IA generativa?	Contamos con un equipo de cumplimie nto especializado que trabaja en estrecha colaboración con nuestros equipos de IA. Realizamos evaluaciones periódicas del impacto en la privacidad, implementamos principios de protección de datos desde el diseño y mantenemos registros detallados del procesamiento de datos para todos los proyectos de IA generativa.
¿Qué grado de madurez tienen sus sistemas actuales que se integran con las nuevas tecnologías de IA generativa?	Nuestra arquitectura de TI se basa en microservicios y APIs permiten una integraci ón flexible de las nuevas tecnologías de IA generativa. Estos sistemas están estandari zados en formatos y protocolos de datos comunes para garantizar la interoperabilidad.

Pregunta	Ejemplo de respuesta
¿Qué experiencia tiene en la operacion alización de modelos de aprendizaje automátic o y cómo podría aplicarse esto a los sistemas de IA generativa?	Tenemos MLOps prácticas establecidas, que incluyen procesos de implementación de modelos automatizados, sistemas de monitoreo y marcos de pruebas A/B. Estas prácticas se están adaptando para cumplir con los requisito s únicos de los modelos de IA generativa a gran escala.

Casos de uso

Pregunta	Ejemplo de respuesta
¿Cuál es el objetivo principal o el criterio de éxito del caso de uso?	Para mejorar el tiempo de respuesta del servicio de atención al cliente, aumentar las conversiones de ventas y mejorar las recomendaciones de productos. Además: para mejorar la satisfacción del usuario, la tasa de finalización de las tareas, la calidad de la respuesta, etc.
¿Cómo se alinea este caso de uso con los objetivos estratégicos de su organización?	Esto se alinea con nuestro objetivo estratégico de mejorar la satisfacción del cliente mediante la reducción de los tiempos de respuesta en el servicio de atención al cliente.
¿Cuál es el volumen esperado de datos o solicitudes para el caso de uso?	500 transacciones por segundo (TPS).
¿Qué tipos de fuentes de datos se necesitan para respaldar sus cargas de trabajo generativ as de IA?	Bases de datos estructuradas internas (registro s de clientes, datos de ventas, etc.); datos de texto no estructurados de documentos, correos electrónicos y redes sociales; archivos de audio y vídeo para tareas de reconocimiento de voz e imágenes; datos de transmisión en tiempo real

Pregunta	Ejemplo de respuesta
	de dispositivos y sensores de IoT; conjuntos de datos públicos y APIs para su enriquecimiento.
¿Con qué frecuencia necesita actualizar o refrescar los datos de estas fuentes?	Bases de datos transaccionales: actualizaciones casi en tiempo real; repositorios de documentos: actualizaciones diarias por lotes; feeds de redes sociales: actualizaciones por hora; datos de sensores de IoT: transmisión continua en tiempo real; conjuntos de datos públicos: actualizaciones mensuales o trimestrales.
¿Qué formatos de datos requieren como entrada sus modelos de IA generativa?	Datos estructurados: tablas de bases de datos CSV, JSON y SQL; datos de texto: texto sin formato, PDF y HTML; datos de imagen: JPEG, PNG y TIFF; datos de audio: WAV y MP3; datos de vídeo: MP4 y AVI.
¿Cuáles son sus principales preocupaciones sobre la calidad de los datos para las cargas de trabajo generativas de IA?	Integridad: garantizar que no falte ningún campo crítico; precisión: verificar la exactitud de los datos y eliminar errores; coherencia: mantener formatos y valores uniformes en todas las fuentes; puntualidad: garantizar que los datos estén actualizados para poder inferirlos en tiempo real; relevancia: confirmar que los datos se alinean con la tarea específica de IA generativa.
¿Cuáles son los requisitos clave de rendimiento (por ejemplo, tiempo de respuesta, rendimiento y precisión)?	Precisión del 95%; tiempo de respuesta inferior a 500 ms; capacidad para gestionar 1000 solicitudes/seg. Alta precisión (95% +), precisión moderada (80-90%), mejor esfuerzo, etc.

Pregunta	Ejemplo de respuesta
¿Tiene alguna otra opción KPIs para medir el éxito de este caso de uso?	Entre las principales se KPIs incluyen la reducción de la tasa de errores, el ahorro de tiempo por transacción y las puntuaciones de satisfacción de los clientes.
¿Cuál es la precisión del modelo que se desea y cómo se equilibra con el coste?	Alta precisión (> 90%) con un coste moderado, precisión moderada (70-80%) con un coste bajo, etc.
¿Cuáles son los principales casos de uso o escenarios de la solución de IA generativa?	Chatbot de servicio al cliente, generación de contenido, recomendación de productos, etc.
¿Cuáles son los usuarios o personas objetivo del sistema de IA generativa?	Agentes de servicio al cliente, equipo de marketing, empleados, usuarios finales, etc.
¿Cuál es el volumen esperado de solicitudes o usuarios?	1000 solicitudes al día; 10 000 usuarios activos al mes.
¿Existen restricciones o requisitos específicos para cada caso de uso?	Respuesta en tiempo real, soporte multilingüe, privacidad de datos, etc.
¿Tiene un presupuesto asignado para desarrollar y mantener la solución de IA generativa?	El coste inicial de desarrollo se estima en 200 000$, con unos costes de mantenimiento anuales de 50 000$.
¿Cuáles son el retorno de la inversión (ROI) y el período de recuperación proyectados para este caso de uso?	Se espera un ROI del 150% en tres años, con un período de amortización de 18 meses.
¿Hay costes ocultos o posibles ahorros que deban tenerse en cuenta?	Los posibles ahorros incluyen la reducción de los costos de horas extra. Los costos ocultos pueden implicar una formación adicional para el personal.
¿Cuáles son las posibilidades de escalabilidad y expansión futura de esta solución de IA generativa?	La solución está diseñada para ampliarse con nuestras operaciones, con la posibilidad de expandirse a otros departamentos en el futuro.

Pregunta	Ejemplo de respuesta
¿Cómo puede garantizar la imparcialidad y mitigar los sesgos en sus modelos de IA generativa?	Planeamos mitigar los sesgos mediante una recopilación de datos diversa, auditorías periódicas de sesgos y la implementación de técnicas de mitigación de sesgos.
¿De qué procesos dispone para abordar las preocupaciones éticas o las consecuencias imprevistas?	Gestionaremos las preocupaciones éticas mediante un plan establecido de respuesta a los incidentes relacionados con la IA, evaluacio nes éticas periódicas de los riesgos, un sistema de denuncias anónimas para los empleados, la colaboración con expertos en ética externos y la supervisión y el ajuste continuos de los modelos implementados en función de los comentarios.
¿Cómo aborda la priorización y la secuencia ción de las evaluaciones generativas de la carga de trabajo de la IA en los diferentes proyectos y departamentos de su organizac ión?	Mediante la realización de una encuesta de alto nivel en todos los departamentos para identific ar posibles casos de uso de la IA generativa y evaluarlos en función de tres criterios clave: el impacto empresarial, la viabilidad técnica y las consideraciones éticas. Se da prioridad a los proyectos con un alto impacto potencial, menores barreras técnicas y preocupaciones éticas mínimas.

Arquitectura

Pregunta	Ejemplo de respuesta
¿Qué tipo de modelo o arquitectura de IA generativa se está considerando?	Transformador, red neuronal convolucional (CNN), red neuronal recurrente (RNN), árboles de decisión, etc.

Pregunta	Ejemplo de respuesta
¿Cuál es la escala o el volumen esperados de datos y cálculos?	Millones de usuarios, petabytes de datos, etc.
¿Cuáles son los requisitos de hardware (por ejemplo, CPUs o GPUs) para la formación y la inferencia?	De gama alta GPUs, clústeres de CPU, instancias en la nube, etc.
¿Cómo se actualizará o reentrenará el modelo de IA generativa a lo largo del tiempo?	Mediante el aprendizaje continuo, el reentrena miento periódico, las actualizaciones manuales, etc.
¿Cuáles son los requisitos de preproces amiento de datos y de ingeniería de caracterí sticas?	Limpieza de texto, aumento de imágenes, selección de funciones, etc.
¿Cómo gestionará el sistema de IA generativ a los casos extremos, los valores atípicos o las entradas de baja confianza?	Recurriendo a la supervisión humana, solicitan do aclaraciones, etc.
¿Cuáles son los requisitos de latencia para la aplicación de IA generativa?	Procesamiento en tiempo real, casi en tiempo real, por lotes, etc.

Almacenamiento

Pregunta	Ejemplo de respuesta
¿Dónde se almacenarán los datos de entrenamiento?	En el almacenamiento en la nube (por ejemplo, Amazon S3, almacenamiento de archivos, almacenamiento en bloques o almacenamiento de objetos), en el almacenamiento local, etc.
¿Cuáles son los requisitos de almacenamiento de los datos de entrenamiento y los artefactos del modelo (por ejemplo, capacidad, durabilid ad, disponibilidad)?	Almacenamiento a escala de petabytes, alta durabilidad (99,99999% de durabilidad), alta disponibilidad, etc.

Pregunta	Ejemplo de respuesta
¿Cuáles son los requisitos de retención y respaldo de datos para los datos de entrenamiento y los artefactos del modelo?	Retención de datos durante x años, copias de seguridad diarias, copias de seguridad externas, etc.
¿Qué formatos de archivo se utilizan principalmente para almacenar tus conjuntos de datos de entrenamiento de IA (por ejemplo, CSV, JSON, HDF5 Parquet)?	Archivos tipo parquet para datos estructurados y HDF5 para matrices multidimensionales de gran tamaño y datos no estructurados, como imágenes y texto. Utilizamos formatos especializados, por ejemplo, TFRecord para optimizar la carga de datos durante el entrenamiento.
¿Cómo se organizan sus conjuntos de datos de entrenamiento: como archivos individuales, en bases de datos o utilizando formatos de datos de IA especializados?	Los conjuntos de datos pequeños y medianos se almacenan como archivos Parquet individuales en el almacenamiento de objetos para mayor flexibilidad. Los conjuntos de datos grandes se almacenan en una base de datos distribuida (Cassandra) para gestionar la escalabilidad.
¿Utilizas alguna técnica de compresión o codificación de datos específica para generar datos de entrenamiento con IA?	Para los datos tabulares, utilizamos técnicas de codificación de diccionario y empaquetado de bits que están disponibles en Parquet. Para las imágenes, utilizamos la compresión JPEG con pérdidas con ajustes de calidad optimizados para nuestros modelos.
¿Cómo se gestiona el control de versiones y el almacenamiento de las diferentes iteraciones de los conjuntos de datos de entrenamiento? ¿Qué impacto tiene esto en sus necesidades generales de almacenamiento?	Usamos un sistema de control de versiones de datos (DVC) que está integrado con nuestra plataforma de aprendizaje automático.

Regulaciones y cumplimiento

Pregunta	Ejemplo de respuesta
¿Cuáles son las normas o los requisitos de conformidad pertinentes para la solución de IA generativa (por ejemplo, el RGPD, la HIPAA o el PCI-DSS)?	El GDPR para el manejo de datos personales, la HIPAA para los datos de salud, el PCI-DSS para los datos de pago, etc.
¿Qué directrices o marcos éticos de IA generativa ha adoptado su organización?	Implementamos nuestras propias directrices de IA responsable. Todos los proyectos de IA generativa se someten a una revisión ética antes de su aprobación e implementación.
¿Cuáles son los requisitos de seguridad del sistema de IA generativa?	Cifrado de datos, comunicación de red segura, auditorías de seguridad periódicas.
¿Cuáles son los requisitos de privacidad y protección de los datos?	Anonimización de datos, cifrado, control de acceso, etc.
¿Cuáles son los requisitos de la solución para gestionar datos sensibles o confidenciales?	Controles de acceso estrictos, enmascara miento de datos, requisitos de residencia de los datos, etc.
¿Cómo se gestionarán la autenticación y la autorización de los usuarios?	Mediante el uso de claves de API OAuth, inicio de sesión único (SSO) y control de acceso basado en roles (RBAC).
¿Cómo se supervisará y gestionará la solución en producción?	Mediante el uso de herramientas de monitoreo como Prometheus y Datadog, herramientas de registro como ELK Stack, sistemas de alerta, etc.

Integración

Pregunta	Ejemplo de respuesta
¿Cuáles son los requisitos para integrar la solución de IA generativa con los sistemas o las fuentes de datos existentes?	REST APIs, colas de mensajes, conectores de bases de datos, etc.
¿Cómo se ingerirán y preprocesarán los datos para la solución de IA generativa?	Mediante el procesamiento por lotes, la transmisión de datos, las transformaciones de datos y la ingeniería de funciones.
¿Cómo se consumirán los resultados de la solución de IA generativa o cómo se integrarán en los sistemas posteriores?	A través de puntos finales de API, colas de mensajes, actualizaciones de bases de datos, etc.
¿Qué patrones de integración basados en eventos se pueden utilizar para la solución de IA generativa?	Colas de mensajes (como Amazon SQS, Apache Kafka o RabbitMQ), sistemas pub/sub, webhooks y plataformas de streaming de eventos.
¿Qué enfoques de integración basados en API se pueden utilizar para conectar la solución de IA generativa con otros sistemas?	RESTful APIs, GraphQL APIs, SOAP APIs (para sistemas heredados).
¿Qué componentes de la arquitectura de microservicios se pueden utilizar para la integración generativa de la solución de IA?	Malla de servicios para la comunicación entre servicios, las pasarelas de API y la organizac ión de contenedores (por ejemplo, Kubernetes).
¿Cómo se puede implementar la integración híbrida para la solución de IA generativa?	Combinando patrones basados en eventos para actualizaciones en tiempo real, procesami ento por lotes de datos históricos y APIs para la integración de sistemas externos.
¿Cómo se puede integrar el resultado de la solución de IA generativa con los sistemas posteriores?	Mediante puntos de enlace de API, colas de mensajes, actualizaciones de bases de datos, webhooks y exportaciones de archivos.

Pregunta	Ejemplo de respuesta
¿Qué medidas de seguridad deberían tenerse en cuenta para integrar la solución de IA generativa?	Mecanismos de autenticación (como OAuth el JWT), el cifrado (en tránsito y en reposo), la limitación de la velocidad de las API y las listas de control de acceso (ACLs).
¿Cómo piensa integrar marcos de código abierto, como LlamaIndex su flujo de datos actual y su flujo de trabajo generativo de IA? LangChain	Tenemos previsto utilizarlas LangChain para crear aplicaciones de IA generativa complejas, especialmente por sus funciones de gestión de agentes y memoria. Nuestro objetivo es que el 60% de nuestros proyectos de IA generativa se utilicen LangChain en los próximos 6 meses.
¿Cómo garantizará la compatibilidad entre los marcos de código abierto que haya elegido y su infraestructura de datos existente?	Estamos creando un equipo de integración dedicado para garantizar una compatibilidad fluida. Para el tercer trimestre, nuestro objetivo es contar con una canalización totalmente integrada que utilice LlamaIndex la indexació n y recuperación de datos eficientes dentro de nuestra estructura actual de lagos de datos.
¿Cómo piensa aprovechar los component es modulares de los marcos, por ejemplo, LangChain para la creación rápida de prototipo s y la experimentación?	Estamos configurando un entorno aislado en el que los desarrolladores pueden crear prototipo s rápidamente utilizando sus componentes LangChain.
¿Cuál es su estrategia para mantenerse al día con las actualizaciones y las nuevas funciones en estos marcos de código abierto en rápida evolución?	Hemos asignado un equipo para que supervise GitHub los repositorios y los foros comunitarios durante LangChain y LlamaIndex. Planeamos evaluar e integrar las principales actualiza ciones trimestralmente, centrándonos en las mejoras de rendimiento y las nuevas capacidad es.

Testeo

Pregunta	Ejemplo de respuesta
¿Cuáles son los requisitos de las pruebas (por ejemplo, pruebas unitarias, pruebas de integración, end-to-end pruebas)?	Pruebas unitarias para componentes individuales, pruebas de integración con sistemas externos, end-to-end pruebas para escenarios críticos, etc.
¿Cómo se garantiza la calidad y la coherencia de los datos en las diferentes fuentes para la formación generativa en IA?	Mantenemos la calidad de los datos mediante herramientas automatizadas de creación de perfiles de datos, auditorías de datos periódicas y un catálogo de datos centralizado. Hemos implementado políticas de gobierno de datos para garantizar la coherencia entre las fuentes y mantener el linaje de datos.
¿Cómo se evaluará y validará el modelo de IA generativa?	Mediante el uso de un conjunto de datos de reserva, una evaluación humana, pruebas A/B, etc.
¿Cuáles son los criterios para evaluar el rendimiento y la precisión del modelo de IA generativa?	Precisión, memoria, puntuación de F1, perplejidad, evaluación humana, etc.
¿Cómo se identificarán y gestionarán los casos extremos y los casos extremos?	Mediante el uso de un conjunto de pruebas completo, evaluación humana, pruebas contradictorias, etc.
¿Cómo probará los posibles sesgos en el modelo de IA generativa?	Mediante el uso de análisis de paridad demográfica, pruebas de igualdad de oportunidades, técnicas de eliminación de sesgos contradictorios, pruebas contrafácticas, etc.
¿Qué métricas se utilizarán para medir la imparcialidad de los resultados del modelo?	Coeficiente de impacto dispar, cuotas igualadas, paridad demográfica, métricas de equidad individual, etc.

Pregunta	Ejemplo de respuesta
¿Cómo garantizará una representación diversa en los conjuntos de datos de sus pruebas para la detección de sesgos?	Mediante el uso de un muestreo estratifi cado para todos los grupos demográficos, la colaboración con expertos en diversidad, el uso de datos sintéticos para subsanar las carencias , etc.
¿Qué proceso se implementará para el monitoreo continuo de la imparcialidad del modelo después del despliegue?	Auditorías de imparcialidad periódicas, sistemas automatizados de detección de sesgos, análisis de los comentarios de los usuarios, reentrenamiento periódico con conjuntos de datos actualizados, etc.
¿Cómo abordará los sesgos interseccionales en el modelo de IA generativa?	Mediante el análisis de equidad interseccional, las pruebas de subgrupos, la colaboración con expertos en el campo de la interseccionalidad, etc.
¿Cómo probará el rendimiento del modelo en diferentes idiomas y contextos culturales?	Mediante el uso de conjuntos de pruebas multilingües, la colaboración con expertos culturales, métricas de equidad localizadas, estudios comparativos interculturales, etc.

Despliegue y automatización

Pregunta	Ejemplo de respuesta
¿Cuáles son los requisitos de escalado y equilibrio de carga?	Enrutamiento inteligente de solicitudes; sistema de escalado automático; optimización para arranques rápidos en frío mediante el empleo de técnicas como el almacenamiento en caché de modelos, la carga diferida y los sistemas de almacenamiento distribuido; diseño del sistema para gestionar patrones de tráfico impredeci bles y a ráfagas.

Pregunta	Ejemplo de respuesta
¿Cuáles son los requisitos para actualizar e implementar nuevas versiones?	Implementaciones azules y verdes, versiones preliminares, actualizaciones progresivas, etc.
¿Cuáles son los requisitos para la recuperación ante desastres y la continuidad empresarial?	Procedimientos de backup y restauración, mecanismos de conmutación por error, configuraciones de alta disponibilidad, etc.
¿Cuáles son los requisitos para automatizar la formación, el despliegue y la gestión del modelo de IA generativa?	Proceso de formación automatizado, despliegu e continuo, escalado automático, etc.
¿Cómo se actualizará y reentrenará el modelo de IA generativa a medida que haya nuevos datos disponibles?	Mediante el reentrenamiento periódico, el aprendizaje incremental, el aprendizaje por transferencia, etc.
¿Cuáles son los requisitos para automatizar la supervisión y la gestión?	Alertas automatizadas, escalado automático, recuperación automática, etc.
¿Cuál es su entorno de despliegue preferido para las cargas de trabajo generativas de IA?	Un enfoque híbrido que utiliza AWS para el entrenamiento de modelos y nuestra infraestr uctura local para realizar inferencias a fin de cumplir con los requisitos de residencia de los datos.
¿Prefiere alguna plataforma en la nube específica para las implementaciones generativ as de IA?	Servicios de AWS, en particular Amazon SageMaker AI para el desarrollo e implement ación de modelos y Amazon Bedrock para los modelos básicos.
¿Qué tecnologías de contenedorización está considerando para las cargas de trabajo generativas de IA?	Queremos estandarizar los contenedores Docker organizados con Kubernetes para garantizar la portabilidad y la escalabilidad en nuestro entorno híbrido.

Pregunta	Ejemplo de respuesta
¿Tiene alguna herramienta preferida para la CI/CD en su cartera de IA generativa?	GitLab para el control de versiones y las canalizaciones de CI/CD, integradas con Jenkins para automatizar las pruebas y el despliegue.
¿Qué herramientas de organización está considerando para gestionar los flujos de trabajo generativos de IA?	Apache Airflow para la organización del flujo de trabajo, especialmente para el preproces amiento de datos y los procesos de formación de modelos.
¿Tiene algún requisito específico para que la infraestructura local soporte las cargas de trabajo generativas de IA?	Estamos invirtiendo en servidores acelerado s por GPU y redes de alta velocidad para soportar las cargas de trabajo de inferencia locales.
¿Cómo piensa gestionar el control de versiones y la implementación de modelos en diferentes entornos?	Planeamos utilizarlos MLflow para el seguimien to de modelos y el control de versiones, e integrarlos con nuestra infraestructura de Kubernetes para una implementación perfecta en todos los entornos.
¿Qué herramientas de monitoreo y observabi lidad está considerando para los despliegues de IA generativa?	Prometheus para la recopilación de métricas y Grafana para la visualización, con soluciones de registro personalizadas adicionales para el monitoreo específico del modelo.
¿Cómo está abordando el movimiento y la sincronización de datos en un modelo de implementación híbrido?	Lo utilizaremos AWS DataSync para una transferencia de datos eficiente entre el almacenamiento local y AWS, con tareas de sincronización automatizadas que se programarán en función de nuestros ciclos de formación.

Pregunta	Ejemplo de respuesta
¿Qué medidas de seguridad están implement ando para los despliegues generativos de IA en diferentes entornos?	Usaremos la IAM para los recursos en la nube y la integraremos con nuestro Active Directory local para implementar el end-to-end cifrado y la segmentación de la red a fin de proteger los flujos de datos.

Estrategia de datos

Pregunta	Ejemplo de respuesta
¿Qué tipos de datos específicos son cruciales para sus cargas de trabajo de IA generativa y a qué porcentaje de ellos se puede acceder actualmente?	Los registros de llamadas de los clientes y los datos de reseñas de productos son cruciales. En la actualidad, se puede acceder al 85% de estos tipos de datos para nuestros proyectos de IA generativa.
¿Cómo garantiza y mide la calidad de sus datos?	Hemos implementado métricas de calidad de los datos, que incluyen la integridad, la precisión, la coherencia y la puntualidad. Utilizamos herramientas automatizadas para evaluar estas métricas con regularidad y contamos con un equipo dedicado a la limpieza y el enriquecimiento de los datos.
¿Qué porcentaje de sus datos cumple con sus estándares de calidad para el uso generativo de la IA?	En la actualidad, el 78% de nuestros datos cumplen con nuestros estándares de calidad. Nuestro objetivo es alcanzar el 95% en los próximos 12 meses mediante la mejora de los procesos de limpieza de datos.
¿Cómo piensa generar confianza entre sus partes interesadas sobre el uso de datos en la IA generativa?	Estamos creando un consejo de ética sobre la IA, proporcionando explicaciones claras de las decisiones en materia de IA y realizand

Pregunta	Ejemplo de respuesta
	o auditorías trimestrales sobre la IA para garantizar la transparencia y la imparcialidad.
¿Qué tan completa es su documentación sobre las fuentes y el linaje de los datos?	Mantenemos un catálogo de datos detallado que incluye los metadatos de todas nuestras fuentes de datos, incluidos el origen, la frecuencia de actualización y el uso. Usamos herramientas de linaje de datos para rastrear cómo fluyen y se transforman los datos en nuestros sistemas.
¿Cómo garantiza la diversidad en sus conjuntos de datos para evitar sesgos en los modelos de IA?	Obtenemos activamente datos de diversos grupos demográficos y auditamos periódica mente nuestros conjuntos de datos para detectar sesgos representativos. También utilizamos técnicas de generación de datos sintéticos para equilibrar las categorías subrepresentadas.
¿Cuál es su frecuencia de actualización de datos para los modelos de IA generativa críticos y cómo se determina esta frecuencia?	Los modelos críticos se actualizan semanalme nte. Esta frecuencia viene determinada por las métricas de rendimiento de las pruebas A/B, y nuestro objetivo es que la degradación entre actualizaciones no supere el 2%.
¿Cuántas versiones de los conjuntos de datos críticos mantiene y durante cuánto tiempo?	Mantenemos las últimas cinco versiones de cada conjunto de datos críticos, con un período de retención de 18 meses para cada versión.
¿Cuántos equipos multifuncionales participa n en sus iniciativas de IA generativa y tienen acceso a sus datos?	Tenemos tres equipos multifuncionales. Cada equipo incluye científicos de datos, expertos en el campo, especialistas en ética y analistas de negocios.

Pregunta	Ejemplo de respuesta
¿Qué políticas y prácticas de gobierno de datos tiene implementadas?	Tenemos un comité de gobierno de datos multifuncional que supervisa nuestras políticas de datos. Hemos implementado controles de acceso basados en funciones, esquemas de clasificación de datos y auditorías periódica s para garantizar el cumplimiento de nuestro marco de gobierno.
¿Qué medidas ha adoptado para garantizar la privacidad de los datos, obtener el consentim iento adecuado y mantener la confidencialidad?	Hemos implementado un marco integral de privacidad de datos alineado con el GDPR y la CCPA. Esto incluye obtener el consentimiento explícito para el uso de los datos, implement ar técnicas de anonimización de los datos y realizar evaluaciones periódicas del impacto en la privacidad.
¿Qué porcentaje de sus conjuntos de datos de entrenamiento de IA se auditaron para detectar sesgos en el último trimestre?	El 70% de nuestros conjuntos de datos de entrenamiento de IA se auditaron para detectar sesgos el trimestre pasado. Estamos implementando herramientas automatizadas de detección de sesgos para realizar auditorías trimestrales al 100%.
¿Cuál es su capacidad de procesamiento de datos actual y cuánto prevé que necesitará para las futuras cargas de trabajo generativas de IA?	Nuestra capacidad actual es de 10 TB/day. We project needing 30 TB/day en un año y estamos ampliando nuestra infraestructura para satisfacer esta demanda.
¿Cuál es su estrategia para equilibrar la privacidad de los datos con las necesidades de datos de los modelos de IA generativa?	Estamos implementando técnicas avanzadas de anonimización y generación de datos sintéticos. Nuestro objetivo es aumentar los datos utilizables para la IA en un 40% y, al mismo tiempo, reducir los riesgos de privacidad en un 60% durante el próximo año.

Pregunta	Ejemplo de respuesta
¿Qué porcentaje de sus conjuntos de datos de aprendizaje automático (ML) están etiquetados con precisión y cuál es su tasa de precisión objetivo?	Actualmente, el 85% de nuestros conjuntos de datos de aprendizaje automático están etiquetados con precisión. Nuestro objetivo es lograr una tasa de precisión del 95% en el próximo trimestre mediante el empleo de técnicas de etiquetado tanto humanas como automatizadas.

Traducir los conocimientos de la evaluación en resultados procesables

Esta sección proporciona un marco para analizar las respuestas al cuestionario y utilizar esos conocimientos para configurar la arquitectura objetivo y otros resultados clave de la iniciativa de modernización generativa de la IA. Este marco cierra la brecha entre la recopilación de datos y la implementación y garantiza que la evaluación sirva de base e impulse directamente su estrategia de modernización.

Definición de la arquitectura objetivo:

- Utilice las respuestas al cuestionario para fundamentar la selección de servicios en la nube y el diseño de las canalizaciones de datos.
- Asegúrese de que el diseño de la arquitectura sea compatible con la escalabilidad y la interoperabilidad, tal como se destaca en la guía.

Evaluación de la preparación del cliente:

- Analice las respuestas al cuestionario relacionadas con la infraestructura, los procesos y la cultura organizacional actuales.
- Identifique las brechas y cree un plan para abordarlas. Priorice las brechas que son fundamentales para el éxito del MVP.

Caso de uso y objetivos ambiciosos:

- Extraiga problemas empresariales específicos de las respuestas al cuestionario para definir objetivos claros para los casos de uso.
- Establezca objetivos ambiciosos que se ajusten a la visión a largo plazo de su organización para la modernización generativa de la IA.

Estimación del esfuerzo:

- Utilice los datos del cuestionario para estimar los recursos, el tiempo y el presupuesto tanto para el MVP como para la implementación completa.
- Cree un enfoque gradual que comience con el MVP y describa las fases posteriores.

Necesidades de habilitación:

- Con base en las respuestas al cuestionario, identifique las brechas de habilidades y las necesidades de capacitación.
- Desarrolle un plan de formación que respalde tanto las necesidades inmediatas de los MVP como la adopción generativa a largo plazo de la IA.

Plan de implementación:

- Cree una hoja de ruta integral que comience con el MVP y describa los pasos hacia la modernización total de la IA generativa.
- Defina hitos y resultados claros para cada fase de la implementación.

Pasos prácticos:

- Matriz de priorización: cree una matriz que asigne las respuestas del cuestionario a los seis resultados para ayudar a priorizar las características y los esfuerzos.
- Enfoque iterativo: diseñe el MVP para que sea la primera iteración de una serie de versiones planificadas, en las que cada versión se base en la arquitectura objetivo completa.
- Alineación de las partes interesadas: utilice los resultados del cuestionario para alinear a las partes interesadas según el alcance del MVP y el enfoque gradual para lograr todos los resultados.
- Ciclo de retroalimentación continuo: Implemente mecanismos para recopilar comentarios después del despliegue del MVP y utilice la información para refinar los planes para las fases posteriores.
- Implementación ágil: adopte una metodología ágil que ofrezca flexibilidad a la hora de abordar todos los resultados a lo largo del tiempo, empezando por los resultados más importantes del MVP.

Pasos a seguir a continuación

Tras completar la evaluación generativa de la carga de trabajo de IA, siga estos pasos:

1. Ofrezca una arquitectura de destino detallada
   - Objetivo: el arquitecto de soluciones crea una arquitectura de destino integral que se alinea con los objetivos de la organización y los resultados de la evaluación.
   - Componentes: esta arquitectura incluye el diseño de la ingesta de datos, los puntos de integración y la interoperabilidad del sistema para garantizar la escalabilidad, la confiabilidad y la optimización del rendimiento.
2. Explique en qué medida se Servicios de AWS ajusta específicamente al caso de uso
   - Mapeo de servicios: identifique y mapee los aspectos específicos Servicios de AWS que mejor se adapten a los casos de uso identificados.
   - Ventajas: destaque la forma en que estos servicios abordan necesidades empresariales específicas, mejoran la eficiencia y proporcionan escalabilidad.
3. Ofrezca soluciones alternativas opcionales con ventajas y desventajas
   - Alternativas: presente soluciones alternativas que también puedan cumplir con los requisitos de la organización.
   - Análisis: ofrezca un análisis detallado de las ventajas y desventajas de cada alternativa teniendo en cuenta factores como el costo, la complejidad y la alineación con los objetivos comerciales.
4. Proporcione una estimación de precios detallada de Servicios de AWS
   - Análisis de costes: proporcione una estimación detallada de los costes de la propuesta Servicios de AWS, incluidos los posibles escenarios de uso y los modelos de precios.
   - Alineación del presupuesto: asegúrese de que el costo se ajuste a las restricciones presupuestarias de la organización y proporcione una comprensión clara de las implicaciones financieras.
5. Obtenga comentarios sobre la arquitectura propuesta
   - Participación de las partes interesadas: interactúe con las partes interesadas para presentar la arquitectura propuesta y recopilar comentarios.
   - Mejora iterativa: utilice los comentarios para refinar y mejorar la solución y confirmar que cumple con las necesidades y expectativas de todas las partes interesadas.

Preguntas frecuentes

¿Cuál es el objetivo principal de la evaluación de la carga de trabajo de la IA generativa?

El objetivo principal de la evaluación es evaluar la preparación de una organización para modernizar sus cargas de trabajo de IA generativa, identificar los casos de uso y desarrollar una arquitectura de solución específica. Su objetivo es definir los requisitos de modernización, determinar el alcance de la implementación y prepararse para una modernización exitosa de la IA generativa.

¿Quién debe utilizar esta evaluación?

Esta evaluación está destinada a arquitectos de soluciones, arquitectos empresariales y arquitectos de aplicaciones que desean evaluar los aspectos técnicos de la modernización generativa de la IA. También es útil para que los directores de programas y los gerentes de personal evalúen las necesidades generales de preparación, asignación de recursos y habilitación.

¿Cuáles son los componentes clave que se evalúan en la evaluación?

La evaluación abarca la preparación general, el caso de uso, la arquitectura, el almacenamiento, las normas y el cumplimiento, la integración, las pruebas, la automatización del despliegue y la estrategia de datos. Estos componentes son cruciales para determinar la preparación técnica y organizativa necesaria para la adopción generativa de la modernización de la IA.

¿Cómo ayuda la evaluación a definir la arquitectura objetivo?

La evaluación proporciona un enfoque estructurado para evaluar los sistemas actuales e identificar las mejoras. Le ayuda a seleccionar las tecnologías adecuadas y a diseñar arquitecturas escalables que se ajusten a los objetivos empresariales y a los requisitos de los casos de uso.

¿Cuáles son las ventajas de realizar una evaluación generativa de la carga de trabajo de la IA?

Los beneficios incluyen una mayor eficiencia, una mejor toma de decisiones, la garantía del cumplimiento, el fomento de la innovación y la preparación para la escalabilidad. La evaluación establece un enfoque estratégico para la modernización generativa de la IA y maximiza los beneficios potenciales al tiempo que mitiga los riesgos.

¿Cómo pueden las organizaciones garantizar una implementación exitosa tras la evaluación?

Las organizaciones deben desarrollar un plan de implementación claro que incluya hitos definidos, involucrar a las partes interesadas desde el principio y adoptar un enfoque iterativo. Establecer un centro de excelencia (CoE) y centrarse en el desarrollo del talento también son prácticas recomendadas.

¿A qué desafíos podrían enfrentarse las organizaciones durante la evaluación?

Los desafíos pueden incluir la resistencia al cambio, los problemas de calidad de los datos y las complejidades del cumplimiento. Para hacer frente a estos desafíos es necesario fomentar una cultura de innovación, garantizar la disponibilidad de los datos e implementar medidas de seguridad sólidas.

¿Cómo aborda la evaluación los requisitos normativos y de cumplimiento?

La evaluación evalúa las medidas de cumplimiento actuales e identifica las brechas. Garantiza que las soluciones objetivo cumplan con las normativas y leyes de privacidad de datos pertinentes e incorporen las mejores prácticas de seguridad para proteger la información confidencial.

¿Qué papel desempeña la participación de las partes interesadas en el proceso de evaluación?

La participación de las partes interesadas es crucial para lograr la aceptación, alinear las iniciativas de modernización con los objetivos empresariales y garantizar una implementación exitosa. La participación temprana y la comunicación clara de los beneficios son fundamentales para superar la resistencia y fomentar el apoyo.

¿Cómo pueden las organizaciones medir el éxito de sus iniciativas generativas de modernización de la IA tras la evaluación?

El éxito se puede medir mediante el uso de indicadores clave de rendimiento (KPIs) que se alinean con los objetivos empresariales. El seguimiento y la evaluación periódicos de estas métricas ayudan a guiar la toma de decisiones y a demostrar a las partes interesadas el valor de la modernización generativa de la IA.

¿En qué se diferencia el enfoque de evaluación para las organizaciones de distintos tamaños (pequeñas, medianas o empresas) o sectores?

Organizaciones pequeñas:

- Es posible que tengan recursos y experiencia limitados para realizar evaluaciones integrales
- Es probable que se centre en casos de uso específicos de alto impacto en lugar de en su adopción en toda la empresa
- Es posible que dependa más de herramientas y servicios de terceros para la evaluación
- El proceso de evaluación puede ser menos formal y más ágil

Organizaciones medianas:

- A menudo cuentan con equipos especializados en TI o datos, pero es posible que carezcan de experiencia especializada en IA
- Podrían adoptar un enfoque gradual, empezando por proyectos piloto en departamentos clave
- ¿Es necesario equilibrar la innovación con los sistemas y procesos existentes

- Es probable que la evaluación involucre a equipos multifuncionales

Organizaciones empresariales:

- Por lo general, cuentan con equipos dedicados a la IA y el aprendizaje automático y más recursos para una evaluación integral
- ¿Necesita considerar integraciones complejas con los sistemas empresariales existentes
- Es posible que deban tenerse en cuenta los requisitos reglamentarios específicos de la industria
- La evaluación a menudo implica procesos de gobierno formales

Recursos

- La IA generativa está activada AWS
- AWS ofrece nuevas guías sobre inteligencia artificial, aprendizaje automático e IA generativa para planificar tu estrategia de IA (AWS entrada del blog)
- Mejores prácticas para crear aplicaciones de IA generativa a partir de ellas AWS(AWS entrada del blog)
- El generador de aplicaciones de IA generativa está activado AWS(biblioteca de AWS soluciones)
- Capacidades de IA generativa (arquitectura de referencia AWS de seguridad)
- AWS marco de mejores prácticas de IA generativa (guía AWS Audit Manager del usuario)
- Elección de un servicio de IA generativa (guía deAWS decisiones)
- ¿Qué es Amazon Bedrock? (Guía del usuario de Amazon Bedrock)
- ¿Qué es Amazon SageMaker AI?(Guía para desarrolladores de Amazon SageMaker AI)

Historial de documentos

En la siguiente tabla, se describen cambios significativos de esta guía. Si quiere recibir notificaciones de futuras actualizaciones, puede suscribirse a las notificaciones RSS.

Cambio	Descripción	Fecha
Publicación inicial	—	6 de noviembre de 2024

AWS Glosario de orientación prescriptiva

Los siguientes son términos de uso común en las estrategias, guías y patrones proporcionados por la Guía AWS prescriptiva. Para sugerir entradas, utilice el enlace Enviar comentarios al final del glosario.

Números

Las 7 R

Siete estrategias de migración comunes para trasladar aplicaciones a la nube. Estas estrategias se basan en las 5 R que Gartner identificó en 2011 y consisten en lo siguiente:

- Refactorizar/rediseñar: traslade una aplicación y modifique su arquitectura mediante el máximo aprovechamiento de las características nativas en la nube para mejorar la agilidad, el rendimiento y la escalabilidad. Por lo general, esto implica trasladar el sistema operativo y la base de datos. Ejemplo: migre su base de datos Oracle local a la edición compatible con PostgreSQL de Amazon Aurora.
- Redefinir la plataforma (transportar y redefinir): traslade una aplicación a la nube e introduzca algún nivel de optimización para aprovechar las capacidades de la nube. Ejemplo: migre su base de datos Oracle local a Amazon Relational Database Service (Amazon RDS) para Oracle en el. Nube de AWS
- Recomprar (readquirir): cambie a un producto diferente, lo cual se suele llevar a cabo al pasar de una licencia tradicional a un modelo SaaS. Ejemplo: migre su sistema de gestión de relaciones con los clientes (CRM) a Salesforce.com.
- Volver a alojar (migrar mediante lift-and-shift): traslade una aplicación a la nube sin realizar cambios para aprovechar las capacidades de la nube. Ejemplo: migre su base de datos Oracle local a Oracle en una EC2 instancia del. Nube de AWS
- Reubicar: (migrar el hipervisor mediante lift and shift): traslade la infraestructura a la nube sin comprar equipo nuevo, reescribir aplicaciones o modificar las operaciones actuales. Los servidores se migran de una plataforma local a un servicio en la nube para la misma plataforma. Ejemplo: migrar una Microsoft Hyper-V aplicación a AWS.
- Retener (revisitar): conserve las aplicaciones en el entorno de origen. Estas pueden incluir las aplicaciones que requieren una refactorización importante, que desee posponer para más adelante, y las aplicaciones heredadas que desee retener, ya que no hay ninguna justificación empresarial para migrarlas.

- Retirar: retire o elimine las aplicaciones que ya no sean necesarias en un entorno de origen.

A

ABAC

Consulte control de acceso basado en atributos.

servicios abstractos

Consulte servicios gestionados.

ACID

Consulte atomicidad, consistencia, aislamiento y durabilidad.

migración activa-activa

Método de migración de bases de datos en el que las bases de datos de origen y destino se mantienen sincronizadas (mediante una herramienta de replicación bidireccional o mediante operaciones de escritura doble) y ambas bases de datos gestionan las transacciones de las aplicaciones conectadas durante la migración. Este método permite la migración en lotes pequeños y controlados, en lugar de requerir una transición única. Es más flexible, pero requiere más trabajo que la migración activa-pasiva.

migración activa-pasiva

Método de migración de bases de datos en el que las bases de datos de origen y destino se mantienen sincronizadas, pero solo la base de datos de origen gestiona las transacciones de las aplicaciones conectadas mientras los datos se replican en la base de datos de destino. La base de datos de destino no acepta ninguna transacción durante la migración.

función de agregación

Función SQL que opera en un grupo de filas y calcula un único valor de retorno para el grupo. Entre los ejemplos de funciones agregadas se incluyen `SUM` y `MAX`.

IA

Véase inteligencia artificial.

AIOps

Consulte las operaciones de inteligencia artificial.

anonimización

El proceso de eliminar permanentemente la información personal de un conjunto de datos. La anonimización puede ayudar a proteger la privacidad personal. Los datos anonimizados ya no se consideran datos personales.

antipatrones

Una solución que se utiliza con frecuencia para un problema recurrente en el que la solución es contraproducente, ineficaz o menos eficaz que una alternativa.

control de aplicaciones

Un enfoque de seguridad que permite el uso únicamente de aplicaciones aprobadas para ayudar a proteger un sistema contra el malware.

cartera de aplicaciones

Recopilación de información detallada sobre cada aplicación que utiliza una organización, incluido el costo de creación y mantenimiento de la aplicación y su valor empresarial. Esta información es clave para el proceso de detección y análisis de la cartera y ayuda a identificar y priorizar las aplicaciones que se van a migrar, modernizar y optimizar.

inteligencia artificial (IA)

El campo de la informática que se dedica al uso de tecnologías informáticas para realizar funciones cognitivas que suelen estar asociadas a los seres humanos, como el aprendizaje, la resolución de problemas y el reconocimiento de patrones. Para más información, consulte ¿Qué es la inteligencia artificial?

operaciones de inteligencia artificial (AIOps)

El proceso de utilizar técnicas de machine learning para resolver problemas operativos, reducir los incidentes operativos y la intervención humana, y mejorar la calidad del servicio. Para obtener más información sobre cómo AIOps se utiliza en la estrategia de AWS migración, consulte la guía de integración de operaciones.

cifrado asimétrico

Algoritmo de cifrado que utiliza un par de claves, una clave pública para el cifrado y una clave privada para el descifrado. Puede compartir la clave pública porque no se utiliza para el descifrado, pero el acceso a la clave privada debe estar sumamente restringido.

atomicidad, consistencia, aislamiento, durabilidad (ACID)

Conjunto de propiedades de software que garantizan la validez de los datos y la fiabilidad operativa de una base de datos, incluso en caso de errores, cortes de energía u otros problemas.

control de acceso basado en atributos (ABAC)

La práctica de crear permisos detallados basados en los atributos del usuario, como el departamento, el puesto de trabajo y el nombre del equipo. Para obtener más información, consulte ABAC AWS en la documentación AWS Identity and Access Management (IAM).

origen de datos fidedigno

Ubicación en la que se almacena la versión principal de los datos, que se considera la fuente de información más fiable. Puede copiar los datos del origen de datos autorizado a otras ubicaciones con el fin de procesarlos o modificarlos, por ejemplo, anonimizarlos, redactarlos o seudonimizarlos.

Zona de disponibilidad

Una ubicación distinta dentro de una Región de AWS que está aislada de los fallos en otras zonas de disponibilidad y que proporciona una conectividad de red económica y de baja latencia a otras zonas de disponibilidad de la misma región.

AWS Marco de adopción de la nube (AWS CAF)

Un marco de directrices y mejores prácticas AWS para ayudar a las organizaciones a desarrollar un plan eficiente y eficaz para migrar con éxito a la nube. AWS CAF organiza la orientación en seis áreas de enfoque denominadas perspectivas: negocios, personas, gobierno, plataforma, seguridad y operaciones. Las perspectivas empresariales, humanas y de gobernanza se centran en las habilidades y los procesos empresariales; las perspectivas de plataforma, seguridad y operaciones se centran en las habilidades y los procesos técnicos. Por ejemplo, la perspectiva humana se dirige a las partes interesadas que se ocupan de los Recursos Humanos (RR. HH.), las funciones del personal y la administración de las personas. Desde esta perspectiva, AWS CAF proporciona orientación para el desarrollo, la formación y la comunicación de las personas a fin de preparar a la organización para una adopción exitosa de la nube. Para obtener más información, consulte la Página web de AWS CAF y el Documento técnico de AWS CAF.

AWS Marco de calificación de la carga de trabajo (AWS WQF)

Herramienta que evalúa las cargas de trabajo de migración de bases de datos, recomienda estrategias de migración y proporciona estimaciones de trabajo. AWS WQF se incluye con AWS

Schema Conversion Tool ().AWS SCT Analiza los esquemas de bases de datos y los objetos de código, el código de las aplicaciones, las dependencias y las características de rendimiento y proporciona informes de evaluación.

B

Un bot malo

Un bot destinado a interrumpir o causar daño a personas u organizaciones.

BCP

Consulte la planificación de la continuidad del negocio.

gráfico de comportamiento

Una vista unificada e interactiva del comportamiento de los recursos y de las interacciones a lo largo del tiempo. Puede utilizar un gráfico de comportamiento con Amazon Detective para examinar los intentos de inicio de sesión fallidos, las llamadas sospechosas a la API y acciones similares. Para obtener más información, consulte Datos en un gráfico de comportamiento en la documentación de Detective.

sistema big-endian

Un sistema que almacena primero el byte más significativo. Véase también endianness.

clasificación binaria

Un proceso que predice un resultado binario (una de las dos clases posibles). Por ejemplo, es posible que su modelo de ML necesite predecir problemas como “¿Este correo electrónico es spam o no es spam?” o “¿Este producto es un libro o un automóvil?”.

filtro de floración

Estructura de datos probabilística y eficiente en términos de memoria que se utiliza para comprobar si un elemento es miembro de un conjunto.

implementación azul/verde

Una estrategia de despliegue en la que se crean dos entornos separados pero idénticos. La versión actual de la aplicación se ejecuta en un entorno (azul) y la nueva versión de la aplicación en el otro entorno (verde). Esta estrategia le ayuda a revertirla rápidamente con un impacto mínimo.

bot

Una aplicación de software que ejecuta tareas automatizadas a través de Internet y simula la actividad o interacción humana. Algunos bots son útiles o beneficiosos, como los rastreadores web que indexan información en Internet. Algunos otros bots, conocidos como bots malos, tienen como objetivo interrumpir o causar daños a personas u organizaciones.

botnet

Redes de bots que están infectadas por malware y que están bajo el control de una sola parte, conocida como pastor u operador de bots. Las botnets son el mecanismo más conocido para escalar los bots y su impacto.

branch

Área contenida de un repositorio de código. La primera rama que se crea en un repositorio es la rama principal. Puede crear una rama nueva a partir de una rama existente y, a continuación, desarrollar características o corregir errores en la rama nueva. Una rama que se genera para crear una característica se denomina comúnmente rama de característica. Cuando la característica se encuentra lista para su lanzamiento, se vuelve a combinar la rama de característica con la rama principal. Para obtener más información, consulte Acerca de las sucursales (GitHub documentación).

acceso con cristales rotos

En circunstancias excepcionales y mediante un proceso aprobado, un usuario puede acceder rápidamente a un sitio para el Cuenta de AWS que normalmente no tiene permisos de acceso. Para obtener más información, consulte el indicador Implemente procedimientos de rotura de cristales en la guía Well-Architected AWS .

estrategia de implementación sobre infraestructura existente

La infraestructura existente en su entorno. Al adoptar una estrategia de implementación sobre infraestructura existente para una arquitectura de sistemas, se diseña la arquitectura en función de las limitaciones de los sistemas y la infraestructura actuales. Si está ampliando la infraestructura existente, puede combinar las estrategias de implementación sobre infraestructuras existentes y de implementación desde cero.

caché de búfer

El área de memoria donde se almacenan los datos a los que se accede con más frecuencia.

capacidad empresarial

Lo que hace una empresa para generar valor (por ejemplo, ventas, servicio al cliente o marketing). Las arquitecturas de microservicios y las decisiones de desarrollo pueden estar impulsadas por las capacidades empresariales. Para obtener más información, consulte la sección Organizado en torno a las capacidades empresariales del documento técnico Ejecutar microservicios en contenedores en AWS.

planificación de la continuidad del negocio (BCP)

Plan que aborda el posible impacto de un evento disruptivo, como una migración a gran escala en las operaciones y permite a la empresa reanudar las operaciones rápidamente.

C

CAF

Consulte el marco AWS de adopción de la nube.

despliegue canario

El lanzamiento lento e incremental de una versión para los usuarios finales. Cuando se tiene confianza, se despliega la nueva versión y se reemplaza la versión actual en su totalidad.

CCoE

Consulte Cloud Center of Excellence.

CDC

Consulte la captura de datos de cambios.

captura de datos de cambio (CDC)

Proceso de seguimiento de los cambios en un origen de datos, como una tabla de base de datos, y registro de los metadatos relacionados con el cambio. Puede utilizar los CDC para diversos fines, como auditar o replicar los cambios en un sistema de destino para mantener la sincronización.

ingeniería del caos

Introducir intencionalmente fallos o eventos disruptivos para poner a prueba la resiliencia de un sistema. Puedes usar AWS Fault Injection Service (AWS FIS) para realizar experimentos que estresen tus AWS cargas de trabajo y evalúen su respuesta.

CI/CD

Consulte la integración continua y la entrega continua.

clasificación

Un proceso de categorización que permite generar predicciones. Los modelos de ML para problemas de clasificación predicen un valor discreto. Los valores discretos siempre son distintos entre sí. Por ejemplo, es posible que un modelo necesite evaluar si hay o no un automóvil en una imagen.

cifrado del cliente

Cifrado de datos localmente, antes de que el objetivo los Servicio de AWS reciba.

Centro de excelencia en la nube (CCoE)

Equipo multidisciplinario que impulsa los esfuerzos de adopción de la nube en toda la organización, incluido el desarrollo de las prácticas recomendadas en la nube, la movilización de recursos, el establecimiento de plazos de migración y la dirección de la organización durante las transformaciones a gran escala. Para obtener más información, consulte las publicaciones de CCo E en el blog de estrategia Nube de AWS empresarial.

computación en la nube

La tecnología en la nube que se utiliza normalmente para la administración de dispositivos de IoT y el almacenamiento de datos de forma remota. La computación en la nube suele estar conectada a la tecnología de computación perimetral.

modelo operativo en la nube

En una organización de TI, el modelo operativo que se utiliza para crear, madurar y optimizar uno o más entornos de nube. Para obtener más información, consulte Creación de su modelo operativo de nube.

etapas de adopción de la nube

Las cuatro fases por las que suelen pasar las organizaciones cuando migran a Nube de AWS:

- Proyecto: ejecución de algunos proyectos relacionados con la nube con fines de prueba de concepto y aprendizaje
- Fundamento: realizar inversiones fundamentales para escalar su adopción de la nube (p. ej., crear una landing zone, definir una CCo E, establecer un modelo de operaciones)
- Migración: migración de aplicaciones individuales
- Reinvención: optimización de productos y servicios e innovación en la nube

Stephen Orban definió estas etapas en la entrada del blog The Journey Toward Cloud-First & the Stages of Adoption en el blog Nube de AWS Enterprise Strategy. Para obtener información sobre su relación con la estrategia de AWS migración, consulte la guía de preparación para la migración.

CMDB

Consulte la base de datos de administración de la configuración.

repositorio de código

Una ubicación donde el código fuente y otros activos, como documentación, muestras y scripts, se almacenan y actualizan mediante procesos de control de versiones. Los repositorios en la nube más comunes incluyen GitHub oBitbucket Cloud. Cada versión del código se denomina rama. En una estructura de microservicios, cada repositorio se encuentra dedicado a una única funcionalidad. Una sola canalización de CI/CD puede utilizar varios repositorios.

caché en frío

Una caché de búfer que está vacía no está bien poblada o contiene datos obsoletos o irrelevantes. Esto afecta al rendimiento, ya que la instancia de la base de datos debe leer desde la memoria principal o el disco, lo que es más lento que leer desde la memoria caché del búfer.

datos fríos

Datos a los que se accede con poca frecuencia y que suelen ser históricos. Al consultar este tipo de datos, normalmente se aceptan consultas lentas. Trasladar estos datos a niveles o clases de almacenamiento de menor rendimiento y menos costosos puede reducir los costos.

visión artificial (CV)

Campo de la IA que utiliza el aprendizaje automático para analizar y extraer información de formatos visuales, como imágenes y vídeos digitales. Por ejemplo, Amazon SageMaker AI proporciona algoritmos de procesamiento de imágenes para CV.

desviación de configuración

En el caso de una carga de trabajo, un cambio de configuración con respecto al estado esperado. Puede provocar que la carga de trabajo deje de cumplir las normas y, por lo general, es gradual e involuntario.

base de datos de administración de configuración (CMDB)

Repositorio que almacena y administra información sobre una base de datos y su entorno de TI, incluidos los componentes de hardware y software y sus configuraciones. Por lo general, los

datos de una CMDB se utilizan en la etapa de detección y análisis de la cartera de productos durante la migración.

paquete de conformidad

Conjunto de AWS Config reglas y medidas correctivas que puede reunir para personalizar sus comprobaciones de conformidad y seguridad. Puede implementar un paquete de conformidad como una entidad única en una región Cuenta de AWS y, o en una organización, mediante una plantilla YAML. Para obtener más información, consulta los paquetes de conformidad en la documentación. AWS Config

integración y entrega continuas (CI/CD)

El proceso de automatización de las etapas de origen, compilación, prueba, puesta en escena y producción del proceso de publicación del software. CI/CD se describe comúnmente como una canalización. CI/CD puede ayudarlo a automatizar los procesos, mejorar la productividad, mejorar la calidad del código y entregar más rápido. Para obtener más información, consulte Beneficios de la entrega continua. CD también puede significar implementación continua. Para obtener más información, consulte Entrega continua frente a implementación continua.

CV

Vea la visión artificial.

D

datos en reposo

Datos que están estacionarios en la red, como los datos que se encuentran almacenados.

clasificación de datos

Un proceso para identificar y clasificar los datos de su red en función de su importancia y sensibilidad. Es un componente fundamental de cualquier estrategia de administración de riesgos de ciberseguridad porque lo ayuda a determinar los controles de protección y retención adecuados para los datos. La clasificación de datos es un componente del pilar de seguridad del AWS Well-Architected Framework. Para obtener más información, consulte Clasificación de datos.

desviación de datos

Una variación significativa entre los datos de producción y los datos que se utilizaron para entrenar un modelo de machine learning, o un cambio significativo en los datos de entrada

a lo largo del tiempo. La desviación de los datos puede reducir la calidad, la precisión y la imparcialidad generales de las predicciones de los modelos de machine learning.

datos en tránsito

Datos que se mueven de forma activa por la red, por ejemplo, entre los recursos de la red.

malla de datos

Un marco arquitectónico que proporciona una propiedad de datos distribuida y descentralizada con administración y gobierno centralizados.

minimización de datos

El principio de recopilar y procesar solo los datos estrictamente necesarios. Practicar la minimización de los datos Nube de AWS puede reducir los riesgos de privacidad, los costos y la huella de carbono de la analítica.

perímetro de datos

Un conjunto de barreras preventivas en su AWS entorno que ayudan a garantizar que solo las identidades confiables accedan a los recursos confiables desde las redes esperadas. Para obtener más información, consulte Crear un perímetro de datos sobre. AWS

preprocesamiento de datos

Transformar los datos sin procesar en un formato que su modelo de ML pueda analizar fácilmente. El preprocesamiento de datos puede implicar eliminar determinadas columnas o filas y corregir los valores faltantes, incoherentes o duplicados.

procedencia de los datos

El proceso de rastrear el origen y el historial de los datos a lo largo de su ciclo de vida, por ejemplo, la forma en que se generaron, transmitieron y almacenaron los datos.

titular de los datos

Persona cuyos datos se recopilan y procesan.

almacenamiento de datos

Un sistema de administración de datos que respalde la inteligencia empresarial, como el análisis. Los almacenes de datos suelen contener grandes cantidades de datos históricos y, por lo general, se utilizan para consultas y análisis.

lenguaje de definición de datos (DDL)

Instrucciones o comandos para crear o modificar la estructura de tablas y objetos de una base de datos.

lenguaje de manipulación de datos (DML)

Instrucciones o comandos para modificar (insertar, actualizar y eliminar) la información de una base de datos.

DDL

Consulte el lenguaje de definición de bases de datos.

conjunto profundo

Combinar varios modelos de aprendizaje profundo para la predicción. Puede utilizar conjuntos profundos para obtener una predicción más precisa o para estimar la incertidumbre de las predicciones.

aprendizaje profundo

Un subcampo del ML que utiliza múltiples capas de redes neuronales artificiales para identificar el mapeo entre los datos de entrada y las variables objetivo de interés.

defense-in-depth

Un enfoque de seguridad de la información en el que se distribuyen cuidadosamente una serie de mecanismos y controles de seguridad en una red informática para proteger la confidencialidad, la integridad y la disponibilidad de la red y de los datos que contiene. Al adoptar esta estrategia AWS, se añaden varios controles en diferentes capas de la AWS Organizations estructura para ayudar a proteger los recursos. Por ejemplo, un defense-in-depth enfoque podría combinar la autenticación multifactorial, la segmentación de la red y el cifrado.

administrador delegado

En AWS Organizations, un servicio compatible puede registrar una cuenta de AWS miembro para administrar las cuentas de la organización y gestionar los permisos de ese servicio. Esta cuenta se denomina administrador delegado para ese servicio. Para obtener más información y una lista de servicios compatibles, consulte Servicios que funcionan con AWS Organizations en la documentación de AWS Organizations .

Implementación

El proceso de hacer que una aplicación, características nuevas o correcciones de código se encuentren disponibles en el entorno de destino. La implementación abarca implementar

cambios en una base de código y, a continuación, crear y ejecutar esa base en los entornos de la aplicación.

entorno de desarrollo

Consulte entorno.

control de detección

Un control de seguridad que se ha diseñado para detectar, registrar y alertar después de que se produzca un evento. Estos controles son una segunda línea de defensa, ya que lo advierten sobre los eventos de seguridad que han eludido los controles preventivos establecidos. Para obtener más información, consulte Controles de detección en Implementación de controles de seguridad en AWS.

asignación de flujos de valor para el desarrollo (DVSM)

Proceso que se utiliza para identificar y priorizar las restricciones que afectan negativamente a la velocidad y la calidad en el ciclo de vida del desarrollo de software. DVSM amplía el proceso de asignación del flujo de valor diseñado originalmente para las prácticas de fabricación ajustada. Se centra en los pasos y los equipos necesarios para crear y transferir valor a través del proceso de desarrollo de software.

gemelo digital

Representación virtual de un sistema del mundo real, como un edificio, una fábrica, un equipo industrial o una línea de producción. Los gemelos digitales son compatibles con el mantenimiento predictivo, la supervisión remota y la optimización de la producción.

tabla de dimensiones

En un esquema en estrella, tabla más pequeña que contiene los atributos de datos sobre los datos cuantitativos de una tabla de hechos. Los atributos de la tabla de dimensiones suelen ser campos de texto o números discretos que se comportan como texto. Estos atributos se utilizan habitualmente para restringir consultas, filtrar y etiquetar conjuntos de resultados.

desastre

Un evento que impide que una carga de trabajo o un sistema cumplan sus objetivos empresariales en su ubicación principal de implementación. Estos eventos pueden ser desastres naturales, fallos técnicos o el resultado de acciones humanas, como una configuración incorrecta involuntaria o un ataque de malware.

recuperación de desastres (DR)

La estrategia y el proceso que se utilizan para minimizar el tiempo de inactividad y la pérdida de datos ocasionados por un desastre. Para obtener más información, consulte Recuperación ante desastres de cargas de trabajo en AWS: Recovery in the Cloud in the AWS Well-Architected Framework.

DML

Consulte el lenguaje de manipulación de bases de datos.

diseño basado en el dominio

Un enfoque para desarrollar un sistema de software complejo mediante la conexión de sus componentes a dominios en evolución, o a los objetivos empresariales principales, a los que sirve cada componente. Este concepto lo introdujo Eric Evans en su libro, Diseño impulsado por el dominio: abordando la complejidad en el corazón del software (Boston: Addison-Wesley Professional, 2003). Para obtener información sobre cómo utilizar el diseño basado en dominios con el patrón de higos estranguladores, consulte Modernización gradual de los servicios web antiguos de Microsoft ASP.NET (ASMX) mediante contenedores y Amazon API Gateway.

DR

Consulte recuperación ante desastres.

detección de deriva

Seguimiento de las desviaciones con respecto a una configuración de referencia. Por ejemplo, puedes usarlo AWS CloudFormation para detectar desviaciones en los recursos del sistema o puedes usarlo AWS Control Tower para detectar cambios en tu landing zone que puedan afectar al cumplimiento de los requisitos de gobierno.

DVSM

Consulte el mapeo del flujo de valor del desarrollo.

E

EDA

Consulte el análisis exploratorio de datos.

EDI

Véase intercambio electrónico de datos.

computación en la periferia

La tecnología que aumenta la potencia de cálculo de los dispositivos inteligentes en la periferia de una red de IoT. En comparación con la computación en nube, la computación perimetral puede reducir la latencia de la comunicación y mejorar el tiempo de respuesta.

intercambio electrónico de datos (EDI)

El intercambio automatizado de documentos comerciales entre organizaciones. Para obtener más información, consulte Qué es el intercambio electrónico de datos.

cifrado

Proceso informático que transforma datos de texto plano, legibles por humanos, en texto cifrado.

clave de cifrado

Cadena criptográfica de bits aleatorios que se genera mediante un algoritmo de cifrado. Las claves pueden variar en longitud y cada una se ha diseñado para ser impredecible y única.

endianidad

El orden en el que se almacenan los bytes en la memoria del ordenador. Los sistemas big-endianos almacenan primero el byte más significativo. Los sistemas Little-Endian almacenan primero el byte menos significativo.

punto de conexión

Consulte el punto final del servicio.

servicio de punto de conexión

Servicio que puede alojar en una nube privada virtual (VPC) para compartir con otros usuarios. Puede crear un servicio de punto final AWS PrivateLink y conceder permisos a otros directores Cuentas de AWS o a AWS Identity and Access Management (IAM). Estas cuentas o entidades principales pueden conectarse a su servicio de punto de conexión de forma privada mediante la creación de puntos de conexión de VPC de interfaz. Para obtener más información, consulte Creación de un servicio de punto de conexión en la documentación de Amazon Virtual Private Cloud (Amazon VPC).

planificación de recursos empresariales (ERP)

Un sistema que automatiza y gestiona los procesos empresariales clave (como la contabilidad, el MES y la gestión de proyectos) de una empresa.

cifrado de sobre

El proceso de cifrar una clave de cifrado con otra clave de cifrado. Para obtener más información, consulte el cifrado de sobres en la documentación de AWS Key Management Service (AWS KMS).

entorno

Una instancia de una aplicación en ejecución. Los siguientes son los tipos de entornos más comunes en la computación en la nube:

- entorno de desarrollo: instancia de una aplicación en ejecución que solo se encuentra disponible para el equipo principal responsable del mantenimiento de la aplicación. Los entornos de desarrollo se utilizan para probar los cambios antes de promocionarlos a los entornos superiores. Este tipo de entorno a veces se denomina entorno de prueba.
- entornos inferiores: todos los entornos de desarrollo de una aplicación, como los que se utilizan para las compilaciones y pruebas iniciales.
- entorno de producción: instancia de una aplicación en ejecución a la que pueden acceder los usuarios finales. En un CI/CD proceso, el entorno de producción es el último entorno de implementación.
- entornos superiores: todos los entornos a los que pueden acceder usuarios que no sean del equipo de desarrollo principal. Esto puede incluir un entorno de producción, entornos de preproducción y entornos para las pruebas de aceptación por parte de los usuarios.

epopeya

En las metodologías ágiles, son categorías funcionales que ayudan a organizar y priorizar el trabajo. Las epopeyas brindan una descripción detallada de los requisitos y las tareas de implementación. Por ejemplo, las epopeyas AWS de seguridad de CAF incluyen la gestión de identidades y accesos, los controles de detección, la seguridad de la infraestructura, la protección de datos y la respuesta a incidentes. Para obtener más información sobre las epopeyas en la estrategia de migración de AWS , consulte la Guía de implementación del programa.

PERP

Consulte planificación de recursos empresariales.

análisis de datos de tipo exploratorio (EDA)

El proceso de analizar un conjunto de datos para comprender sus características principales. Se recopilan o agregan datos y, a continuación, se realizan las investigaciones iniciales para

encontrar patrones, detectar anomalías y comprobar las suposiciones. El EDA se realiza mediante el cálculo de estadísticas resumidas y la creación de visualizaciones de datos.

F

tabla de datos

La tabla central de un esquema en forma de estrella. Almacena datos cuantitativos sobre las operaciones comerciales. Normalmente, una tabla de hechos contiene dos tipos de columnas: las que contienen medidas y las que contienen una clave externa para una tabla de dimensiones.

fallan rápidamente

Una filosofía que utiliza pruebas frecuentes e incrementales para reducir el ciclo de vida del desarrollo. Es una parte fundamental de un enfoque ágil.

límite de aislamiento de fallas

En el Nube de AWS, un límite, como una zona de disponibilidad Región de AWS, un plano de control o un plano de datos, que limita el efecto de una falla y ayuda a mejorar la resiliencia de las cargas de trabajo. Para obtener más información, consulte Límites de AWS aislamiento de errores.

rama de característica

Consulte la sucursal.

características

Los datos de entrada que se utilizan para hacer una predicción. Por ejemplo, en un contexto de fabricación, las características pueden ser imágenes que se capturan periódicamente desde la línea de fabricación.

importancia de las características

La importancia que tiene una característica para las predicciones de un modelo. Por lo general, esto se expresa como una puntuación numérica que se puede calcular mediante diversas técnicas, como las explicaciones aditivas de Shapley (SHAP) y los gradientes integrados. Para obtener más información, consulte Interpretabilidad del modelo de aprendizaje automático con AWS.

transformación de funciones

Optimizar los datos para el proceso de ML, lo que incluye enriquecer los datos con fuentes adicionales, escalar los valores o extraer varios conjuntos de información de un solo campo de datos. Esto permite que el modelo de ML se beneficie de los datos. Por ejemplo, si divide la fecha del “27 de mayo de 2021 00:15:37” en “jueves”, “mayo”, “2021” y “15”, puede ayudar al algoritmo de aprendizaje a aprender patrones matizados asociados a los diferentes componentes de los datos.

indicaciones de pocos pasos

Proporcionar a un LLM un pequeño número de ejemplos que demuestren la tarea y el resultado deseado antes de pedirle que realice una tarea similar. Esta técnica es una aplicación del aprendizaje contextual, en el que los modelos aprenden a partir de ejemplos (planos) integrados en las instrucciones. Las indicaciones con pocas tomas pueden ser eficaces para tareas que requieren un formato, un razonamiento o un conocimiento del dominio específicos. Consulte también el apartado de mensajes sin intervención.

FGAC

Consulte el control de acceso detallado.

control de acceso preciso (FGAC)

El uso de varias condiciones que tienen por objetivo permitir o denegar una solicitud de acceso.

migración relámpago

Método de migración de bases de datos que utiliza la replicación continua de datos mediante la captura de datos modificados para migrar los datos en el menor tiempo posible, en lugar de utilizar un enfoque gradual. El objetivo es reducir al mínimo el tiempo de inactividad.

FM

Consulte el modelo básico.

modelo de base (FM)

Una gran red neuronal de aprendizaje profundo que se ha estado entrenando con conjuntos de datos masivos de datos generalizados y sin etiquetar. FMs son capaces de realizar una amplia variedad de tareas generales, como comprender el lenguaje, generar texto e imágenes y conversar en lenguaje natural. Para obtener más información, consulte Qué son los modelos básicos.

G

IA generativa

Un subconjunto de modelos de IA que se han entrenado con grandes cantidades de datos y que pueden utilizar un simple mensaje de texto para crear contenido y artefactos nuevos, como imágenes, vídeos, texto y audio. Para obtener más información, consulte Qué es la IA generativa.

bloqueo geográfico

Consulta las restricciones geográficas.

restricciones geográficas (bloqueo geográfico)

En Amazon CloudFront, una opción para impedir que los usuarios de países específicos accedan a las distribuciones de contenido. Puede utilizar una lista de permitidos o bloqueados para especificar los países aprobados y prohibidos. Para obtener más información, consulta la sección Restringir la distribución geográfica del contenido en la CloudFront documentación.

Flujo de trabajo de Gitflow

Un enfoque en el que los entornos inferiores y superiores utilizan diferentes ramas en un repositorio de código fuente. El flujo de trabajo de Gitflow se considera heredado, y el flujo de trabajo basado en enlaces troncales es el enfoque moderno preferido.

imagen dorada

Instantánea de un sistema o software que se utiliza como plantilla para implementar nuevas instancias de ese sistema o software. Por ejemplo, en la fabricación, una imagen dorada se puede utilizar para aprovisionar software en varios dispositivos y ayuda a mejorar la velocidad, la escalabilidad y la productividad de las operaciones de fabricación de dispositivos.

estrategia de implementación desde cero

La ausencia de infraestructura existente en un entorno nuevo. Al adoptar una estrategia de implementación desde cero para una arquitectura de sistemas, puede seleccionar todas las tecnologías nuevas sin que estas deban ser compatibles con una infraestructura existente, lo que también se conoce como implementación sobre infraestructura existente. Si está ampliando la infraestructura existente, puede combinar las estrategias de implementación sobre infraestructuras existentes y de implementación desde cero.

barrera de protección

Una regla de alto nivel que ayuda a regular los recursos, las políticas y el cumplimiento en todas las unidades organizativas (OUs). Las barreras de protección preventivas aplican políticas para garantizar la alineación con los estándares de conformidad. Se implementan mediante políticas de control de servicios y límites de permisos de IAM. Las barreras de protección de detección detectan las vulneraciones de las políticas y los problemas de conformidad, y generan alertas para su corrección. Se implementan mediante Amazon AWS Config AWS Security Hub GuardDuty AWS Trusted Advisor, Amazon Inspector y AWS Lambda cheques personalizados.

H

HA

Consulte la alta disponibilidad.

migración heterogénea de bases de datos

Migración de la base de datos de origen a una base de datos de destino que utilice un motor de base de datos diferente (por ejemplo, de Oracle a Amazon Aurora). La migración heterogénea suele ser parte de un esfuerzo de rediseño de la arquitectura y convertir el esquema puede ser una tarea compleja. AWS ofrece AWS SCT, lo cual ayuda con las conversiones de esquemas.

alta disponibilidad (HA)

La capacidad de una carga de trabajo para funcionar de forma continua, sin intervención, en caso de desafíos o desastres. Los sistemas de alta disponibilidad están diseñados para realizar una conmutación por error automática, ofrecer un rendimiento de alta calidad de forma constante y gestionar diferentes cargas y fallos con un impacto mínimo en el rendimiento.

modernización histórica

Un enfoque utilizado para modernizar y actualizar los sistemas de tecnología operativa (TO) a fin de satisfacer mejor las necesidades de la industria manufacturera. Un histórico es un tipo de base de datos que se utiliza para recopilar y almacenar datos de diversas fuentes en una fábrica.

datos retenidos

Parte de los datos históricos etiquetados que se ocultan de un conjunto de datos que se utiliza para entrenar un modelo de aprendizaje automático. Puede utilizar los datos de reserva para evaluar el rendimiento del modelo comparando las predicciones del modelo con los datos de reserva.

migración homogénea de bases de datos

Migración de la base de datos de origen a una base de datos de destino que comparte el mismo motor de base de datos (por ejemplo, Microsoft SQL Server a Amazon RDS para SQL Server). La migración homogénea suele formar parte de un esfuerzo para volver a alojar o redefinir la plataforma. Puede utilizar las utilidades de bases de datos nativas para migrar el esquema.

datos recientes

Datos a los que se accede con frecuencia, como datos en tiempo real o datos traslacionales recientes. Por lo general, estos datos requieren un nivel o una clase de almacenamiento de alto rendimiento para proporcionar respuestas rápidas a las consultas.

hotfix

Una solución urgente para un problema crítico en un entorno de producción. Debido a su urgencia, las revisiones suelen realizarse fuera del flujo de trabajo habitual de las versiones. DevOps

periodo de hiperatención

Periodo, inmediatamente después de la transición, durante el cual un equipo de migración administra y monitorea las aplicaciones migradas en la nube para solucionar cualquier problema. Por lo general, este periodo dura de 1 a 4 días. Al final del periodo de hiperatención, el equipo de migración suele transferir la responsabilidad de las aplicaciones al equipo de operaciones en la nube.

I

IaC

Vea la infraestructura como código.

políticas basadas en identidades

Política asociada a uno o más directores de IAM que define sus permisos en el Nube de AWS entorno.

aplicación inactiva

Aplicación que utiliza un promedio de CPU y memoria de entre 5 y 20 por ciento durante un periodo de 90 días. En un proyecto de migración, es habitual retirar estas aplicaciones o mantenerlas en las instalaciones.

IIoT

Consulte Internet de las cosas industrial.

infraestructura inmutable

Un modelo que implementa una nueva infraestructura para las cargas de trabajo de producción en lugar de actualizar, parchear o modificar la infraestructura existente. Las infraestructuras inmutables son intrínsecamente más consistentes, fiables y predecibles que las infraestructuras mutables. Para obtener más información, consulte las prácticas recomendadas para implementar con una infraestructura inmutable en Well-Architected Framework AWS .

VPC entrante (de entrada)

En una arquitectura de AWS cuentas múltiples, una VPC que acepta, inspecciona y enruta las conexiones de red desde fuera de una aplicación. La arquitectura AWS de referencia de seguridad recomienda configurar la cuenta de red con entradas, salidas e inspección VPCs para proteger la interfaz bidireccional entre la aplicación y el resto de Internet.

migración gradual

Estrategia de transición en la que se migra la aplicación en partes pequeñas en lugar de realizar una transición única y completa. Por ejemplo, puede trasladar inicialmente solo unos pocos microservicios o usuarios al nuevo sistema. Tras comprobar que todo funciona correctamente, puede trasladar microservicios o usuarios adicionales de forma gradual hasta que pueda retirar su sistema heredado. Esta estrategia reduce los riesgos asociados a las grandes migraciones.

Industria 4.0

Un término que Klaus Schwab introdujo en 2016 para referirse a la modernización de los procesos de fabricación mediante avances en la conectividad, los datos en tiempo real, la automatización, el análisis y la inteligencia artificial/aprendizaje automático.

infraestructura

Todos los recursos y activos que se encuentran en el entorno de una aplicación.

infraestructura como código (IaC)

Proceso de aprovisionamiento y administración de la infraestructura de una aplicación mediante un conjunto de archivos de configuración. La IaC se ha diseñado para ayudarlo a centralizar la administración de la infraestructura, estandarizar los recursos y escalar con rapidez a fin de que los entornos nuevos sean repetibles, fiables y consistentes.

Internet de las cosas industrial (T) IIo

El uso de sensores y dispositivos conectados a Internet en los sectores industriales, como el productivo, el eléctrico, el automotriz, el sanitario, el de las ciencias de la vida y el de la agricultura. Para obtener más información, consulte Creación de una estrategia de transformación digital de la Internet de las cosas (IIoT) industrial.

VPC de inspección

En una arquitectura de AWS cuentas múltiples, una VPC centralizada que gestiona las inspecciones del tráfico de red VPCs entre Internet y las redes locales (en una misma o Regiones de AWS diferente). La arquitectura AWS de referencia de seguridad recomienda configurar su cuenta de red con entrada, salida e inspección VPCs para proteger la interfaz bidireccional entre la aplicación e Internet en general.

Internet de las cosas (IoT)

Red de objetos físicos conectados con sensores o procesadores integrados que se comunican con otros dispositivos y sistemas a través de Internet o de una red de comunicación local. Para obtener más información, consulte ¿Qué es IoT?.

interpretabilidad

Característica de un modelo de machine learning que describe el grado en que un ser humano puede entender cómo las predicciones del modelo dependen de sus entradas. Para obtener más información, consulte Interpretabilidad del modelo de aprendizaje automático con. AWS

IoT

Consulte Internet de las cosas.

biblioteca de información de TI (ITIL)

Conjunto de prácticas recomendadas para ofrecer servicios de TI y alinearlos con los requisitos empresariales. La ITIL proporciona la base para la ITSM.

administración de servicios de TI (ITSM)

Actividades asociadas con el diseño, la implementación, la administración y el soporte de los servicios de TI para una organización. Para obtener información sobre la integración de las operaciones en la nube con las herramientas de ITSM, consulte la Guía de integración de operaciones.

ITIL

Consulte la biblioteca de información de TI.

ITSM

Consulte Administración de servicios de TI.

L

control de acceso basado en etiquetas (LBAC)

Una implementación del control de acceso obligatorio (MAC) en la que a los usuarios y a los propios datos se les asigna explícitamente un valor de etiqueta de seguridad. La intersección entre la etiqueta de seguridad del usuario y la etiqueta de seguridad de los datos determina qué filas y columnas puede ver el usuario.

zona de aterrizaje

Una landing zone es un AWS entorno multicuenta bien diseñado, escalable y seguro. Este es un punto de partida desde el cual las empresas pueden lanzar e implementar rápidamente cargas de trabajo y aplicaciones con confianza en su entorno de seguridad e infraestructura. Para obtener más información sobre las zonas de aterrizaje, consulte Configuración de un entorno de AWS seguro y escalable con varias cuentas.

modelo de lenguaje grande (LLM)

Un modelo de IA de aprendizaje profundo que se entrena previamente con una gran cantidad de datos. Un LLM puede realizar múltiples tareas, como responder preguntas, resumir documentos, traducir textos a otros idiomas y completar oraciones. Para obtener más información, consulte Qué son. LLMs

migración grande

Migración de 300 servidores o más.

LBAC

Consulte control de acceso basado en etiquetas.

privilegio mínimo

La práctica recomendada de seguridad que consiste en conceder los permisos mínimos necesarios para realizar una tarea. Para obtener más información, consulte Aplicar permisos de privilegio mínimo en la documentación de IAM.

migrar mediante lift-and-shift

Ver 7 Rs.

sistema little-endian

Un sistema que almacena primero el byte menos significativo. Véase también endianness.

LLM

Véase un modelo de lenguaje amplio.

entornos inferiores

Véase entorno.

M

machine learning (ML)

Un tipo de inteligencia artificial que utiliza algoritmos y técnicas para el reconocimiento y el aprendizaje de patrones. El ML analiza y aprende de los datos registrados, como los datos del Internet de las cosas (IoT), para generar un modelo estadístico basado en patrones. Para más información, consulte Machine learning.

rama principal

Ver sucursal.

malware

Software diseñado para comprometer la seguridad o la privacidad de la computadora. El malware puede interrumpir los sistemas informáticos, filtrar información confidencial u obtener acceso no autorizado. Algunos ejemplos de malware son los virus, los gusanos, el ransomware, los troyanos, el spyware y los keyloggers.

servicios gestionados

Servicios de AWS para los que AWS opera la capa de infraestructura, el sistema operativo y las plataformas, y usted accede a los puntos finales para almacenar y recuperar datos. Amazon Simple Storage Service (Amazon S3) y Amazon DynamoDB son ejemplos de servicios gestionados. También se conocen como servicios abstractos.

sistema de ejecución de fabricación (MES)

Un sistema de software para rastrear, monitorear, documentar y controlar los procesos de producción que convierten las materias primas en productos terminados en el taller.

MAP

Consulte Migration Acceleration Program.

mecanismo

Un proceso completo en el que se crea una herramienta, se impulsa su adopción y, a continuación, se inspeccionan los resultados para realizar ajustes. Un mecanismo es un ciclo que se refuerza y mejora a sí mismo a medida que funciona. Para obtener más información, consulte Creación de mecanismos en el AWS Well-Architected Framework.

cuenta de miembro

Todas las Cuentas de AWS demás cuentas, excepto la de administración, que forman parte de una organización. AWS Organizations Una cuenta no puede pertenecer a más de una organización a la vez.

MES

Consulte el sistema de ejecución de la fabricación.

Transporte telemétrico de Message Queue Queue (MQTT)

Un protocolo de comunicación ligero machine-to-machine (M2M), basado en el patrón de publicación/suscripción, para dispositivos de IoT con recursos limitados.

microservicio

Un servicio pequeño e independiente que se comunica a través de una red bien definida APIs y que, por lo general, es propiedad de equipos pequeños e independientes. Por ejemplo, un sistema de seguros puede incluir microservicios que se adapten a las capacidades empresariales, como las de ventas o marketing, o a subdominios, como las de compras, reclamaciones o análisis. Los beneficios de los microservicios incluyen la agilidad, la escalabilidad flexible, la facilidad de implementación, el código reutilizable y la resiliencia. Para obtener más información, consulte Integrar microservicios mediante AWS servicios sin servidor.

arquitectura de microservicios

Un enfoque para crear una aplicación con componentes independientes que ejecutan cada proceso de la aplicación como un microservicio. Estos microservicios se comunican a través de una interfaz bien definida mediante un uso ligero. APIs Cada microservicio de esta arquitectura se puede actualizar, implementar y escalar para satisfacer la demanda de funciones específicas de una aplicación. Para obtener más información, consulte Implementación de microservicios en. AWS

Programa de aceleración de la migración (MAP)

Un AWS programa que proporciona soporte de consultoría, formación y servicios para ayudar a las organizaciones a crear una base operativa sólida para migrar a la nube y para ayudar a compensar el costo inicial de las migraciones. El MAP incluye una metodología de migración para ejecutar las migraciones antiguas de forma metódica y un conjunto de herramientas para automatizar y acelerar los escenarios de migración más comunes.

migración a escala

Proceso de transferencia de la mayoría de la cartera de aplicaciones a la nube en oleadas, con más aplicaciones desplazadas a un ritmo más rápido en cada oleada. En esta fase, se utilizan las prácticas recomendadas y las lecciones aprendidas en las fases anteriores para implementar una fábrica de migración de equipos, herramientas y procesos con el fin de agilizar la migración de las cargas de trabajo mediante la automatización y la entrega ágil. Esta es la tercera fase de la estrategia de migración de AWS.

fábrica de migración

Equipos multifuncionales que agilizan la migración de las cargas de trabajo mediante enfoques automatizados y ágiles. Los equipos de las fábricas de migración suelen incluir a analistas y propietarios de operaciones, empresas, ingenieros de migración, desarrolladores y DevOps profesionales que trabajan a pasos agigantados. Entre el 20 y el 50 por ciento de la cartera de aplicaciones empresariales se compone de patrones repetidos que pueden optimizarse mediante un enfoque de fábrica. Para obtener más información, consulte la discusión sobre las fábricas de migración y la Guía de fábricas de migración a la nube en este contenido.

metadatos de migración

Información sobre la aplicación y el servidor que se necesita para completar la migración. Cada patrón de migración requiere un conjunto diferente de metadatos de migración. Algunos ejemplos de metadatos de migración son la subred de destino, el grupo de seguridad y AWS la cuenta.

patrón de migración

Tarea de migración repetible que detalla la estrategia de migración, el destino de la migración y la aplicación o el servicio de migración utilizados. Ejemplo: realoje la migración a Amazon EC2 con AWS Application Migration Service.

Migration Portfolio Assessment (MPA)

Una herramienta en línea que proporciona información para validar el modelo de negocio para migrar a. Nube de AWS La MPA ofrece una evaluación detallada de la cartera (adecuación del

tamaño de los servidores, precios, comparaciones del costo total de propiedad, análisis de los costos de migración), así como una planificación de la migración (análisis y recopilación de datos de aplicaciones, agrupación de aplicaciones, priorización de la migración y planificación de oleadas). La herramienta MPA (requiere iniciar sesión) está disponible de forma gratuita para todos los AWS consultores y consultores asociados de APN.

Evaluación de la preparación para la migración (MRA)

Proceso que consiste en obtener información sobre el estado de preparación de una organización para la nube, identificar sus puntos fuertes y débiles y elaborar un plan de acción para cerrar las brechas identificadas mediante el AWS CAF. Para obtener más información, consulte la Guía de preparación para la migración. La MRA es la primera fase de la estrategia de migración de AWS.

estrategia de migración

El enfoque utilizado para migrar una carga de trabajo a. Nube de AWS Para obtener más información, consulte la entrada de las 7 R de este glosario y consulte Movilice a su organización para acelerar las migraciones a gran escala.

ML

Consulte el aprendizaje automático.

modernización

Transformar una aplicación obsoleta (antigua o monolítica) y su infraestructura en un sistema ágil, elástico y de alta disponibilidad en la nube para reducir los gastos, aumentar la eficiencia y aprovechar las innovaciones. Para obtener más información, consulte Estrategia para modernizar las aplicaciones en el Nube de AWS.

evaluación de la preparación para la modernización

Evaluación que ayuda a determinar la preparación para la modernización de las aplicaciones de una organización; identifica los beneficios, los riesgos y las dependencias; y determina qué tan bien la organización puede soportar el estado futuro de esas aplicaciones. El resultado de la evaluación es un esquema de la arquitectura objetivo, una hoja de ruta que detalla las fases de desarrollo y los hitos del proceso de modernización y un plan de acción para abordar las brechas identificadas. Para obtener más información, consulte Evaluación de la preparación para la modernización de las aplicaciones en el Nube de AWS.

aplicaciones monolíticas (monolitos)

Aplicaciones que se ejecutan como un único servicio con procesos estrechamente acoplados. Las aplicaciones monolíticas presentan varios inconvenientes. Si una característica de la

aplicación experimenta un aumento en la demanda, se debe escalar toda la arquitectura. Agregar o mejorar las características de una aplicación monolítica también se vuelve más complejo a medida que crece la base de código. Para solucionar problemas con la aplicación, puede utilizar una arquitectura de microservicios. Para obtener más información, consulte Descomposición de monolitos en microservicios.

MAPA

Consulte la evaluación de la cartera de migración.

MQTT

Consulte Message Queue Queue Telemetría y Transporte.

clasificación multiclase

Un proceso que ayuda a generar predicciones para varias clases (predice uno de más de dos resultados). Por ejemplo, un modelo de ML podría preguntar "¿Este producto es un libro, un automóvil o un teléfono?" o "¿Qué categoría de productos es más interesante para este cliente?".

infraestructura mutable

Un modelo que actualiza y modifica la infraestructura existente para las cargas de trabajo de producción. Para mejorar la coherencia, la fiabilidad y la previsibilidad, el AWS Well-Architected Framework recomienda el uso de una infraestructura inmutable como práctica recomendada.

O

OAC

Consulte el control de acceso de origen.

OAI

Consulte la identidad de acceso de origen.

OCM

Consulte gestión del cambio organizacional.

migración fuera de línea

Método de migración en el que la carga de trabajo de origen se elimina durante el proceso de migración. Este método implica un tiempo de inactividad prolongado y, por lo general, se utiliza para cargas de trabajo pequeñas y no críticas.

OI

Consulte integración de operaciones.

OLA

Véase el acuerdo a nivel operativo.

migración en línea

Método de migración en el que la carga de trabajo de origen se copia al sistema de destino sin que se desconecte. Las aplicaciones que están conectadas a la carga de trabajo pueden seguir funcionando durante la migración. Este método implica un tiempo de inactividad nulo o mínimo y, por lo general, se utiliza para cargas de trabajo de producción críticas.

OPC-UA

Consulte Open Process Communications: arquitectura unificada.

Comunicaciones de proceso abierto: arquitectura unificada (OPC-UA)

Un protocolo de comunicación machine-to-machine (M2M) para la automatización industrial. El OPC-UA proporciona un estándar de interoperabilidad con esquemas de cifrado, autenticación y autorización de datos.

acuerdo de nivel operativo (OLA)

Acuerdo que aclara lo que los grupos de TI operativos se comprometen a ofrecerse entre sí, para respaldar un acuerdo de nivel de servicio (SLA).

revisión de la preparación operativa (ORR)

Una lista de preguntas y las mejores prácticas asociadas que le ayudan a comprender, evaluar, prevenir o reducir el alcance de los incidentes y posibles fallos. Para obtener más información, consulte las Revisiones de preparación operativa (ORR) en AWS Well-Architected Framework.

tecnología operativa (OT)

Sistemas de hardware y software que funcionan con el entorno físico para controlar las operaciones, los equipos y la infraestructura industriales. En la industria manufacturera, la integración de los sistemas de TO y tecnología de la información (TI) es un enfoque clave para las transformaciones de la industria 4.0.

integración de operaciones (OI)

Proceso de modernización de las operaciones en la nube, que implica la planificación de la preparación, la automatización y la integración. Para obtener más información, consulte la Guía de integración de las operaciones.

registro de seguimiento organizativo

Un registro creado por el AWS CloudTrail que se registran todos los eventos para todos Cuentas de AWS los miembros de una organización AWS Organizations. Este registro de seguimiento se crea en cada Cuenta de AWS que forma parte de la organización y realiza un seguimiento de la actividad en cada cuenta. Para obtener más información, consulte Crear un registro para una organización en la CloudTrail documentación.

administración del cambio organizacional (OCM)

Marco para administrar las transformaciones empresariales importantes y disruptivas desde la perspectiva de las personas, la cultura y el liderazgo. La OCM ayuda a las empresas a prepararse para nuevos sistemas y estrategias y a realizar la transición a ellos, al acelerar la adopción de cambios, abordar los problemas de transición e impulsar cambios culturales y organizacionales. En la estrategia de AWS migración, este marco se denomina aceleración de personal, debido a la velocidad de cambio que requieren los proyectos de adopción de la nube. Para obtener más información, consulte la Guía de OCM.

control de acceso de origen (OAC)

En CloudFront, una opción mejorada para restringir el acceso y proteger el contenido del Amazon Simple Storage Service (Amazon S3). El OAC admite todos los buckets de S3 Regiones de AWS, el cifrado del lado del servidor AWS KMS (SSE-KMS) y las solicitudes dinámicas `PUT` y `DELETE` dirigidas al bucket de S3.

identidad de acceso de origen (OAI)

En CloudFront, una opción para restringir el acceso y proteger el contenido de Amazon S3. Cuando utiliza OAI, CloudFront crea un principal con el que Amazon S3 puede autenticarse. Los directores autenticados solo pueden acceder al contenido de un bucket de S3 a través de una distribución específica. CloudFront Consulte también el OAC, que proporciona un control de acceso más detallado y mejorado.

ORR

Consulte la revisión de la preparación operativa.

OT

Consulte la tecnología operativa.

VPC saliente (de salida)

En una arquitectura de AWS cuentas múltiples, una VPC que gestiona las conexiones de red que se inician desde una aplicación. La arquitectura AWS de referencia de seguridad recomienda configurar la cuenta de red con entradas, salidas e inspección VPCs para proteger la interfaz bidireccional entre la aplicación e Internet en general.

P

límite de permisos

Una política de administración de IAM que se adjunta a las entidades principales de IAM para establecer los permisos máximos que puede tener el usuario o el rol. Para obtener más información, consulte Límites de permisos en la documentación de IAM.

información de identificación personal (PII)

Información que, vista directamente o combinada con otros datos relacionados, puede utilizarse para deducir de manera razonable la identidad de una persona. Algunos ejemplos de información de identificación personal son los nombres, las direcciones y la información de contacto.

PII

Consulte la información de identificación personal.

manual de estrategias

Conjunto de pasos predefinidos que capturan el trabajo asociado a las migraciones, como la entrega de las funciones de operaciones principales en la nube. Un manual puede adoptar la forma de scripts, manuales de procedimientos automatizados o resúmenes de los procesos o pasos necesarios para operar un entorno modernizado.

PLC

Consulte controlador lógico programable.

PLM

Consulte la gestión del ciclo de vida del producto.

policy

Un objeto que puede definir los permisos (consulte la política basada en la identidad), especifique las condiciones de acceso (consulte la política basada en los recursos) o defina los permisos máximos para todas las cuentas de una organización AWS Organizations (consulte la política de control de servicios).

persistencia políglota

Elegir de forma independiente la tecnología de almacenamiento de datos de un microservicio en función de los patrones de acceso a los datos y otros requisitos. Si sus microservicios tienen la misma tecnología de almacenamiento de datos, pueden enfrentarse a desafíos de implementación o experimentar un rendimiento deficiente. Los microservicios se implementan más fácilmente y logran un mejor rendimiento y escalabilidad si utilizan el almacén de datos que mejor se adapte a sus necesidades. Para obtener más información, consulte Habilitación de la persistencia de datos en los microservicios.

evaluación de cartera

Proceso de detección, análisis y priorización de la cartera de aplicaciones para planificar la migración. Para obtener más información, consulte la Evaluación de la preparación para la migración.

predicate

Una condición de consulta que devuelve `true` o`false`, normalmente, se encuentra en una cláusula. `WHERE`

pulsar un predicado

Técnica de optimización de consultas de bases de datos que filtra los datos de la consulta antes de transferirlos. Esto reduce la cantidad de datos que se deben recuperar y procesar de la base de datos relacional y mejora el rendimiento de las consultas.

control preventivo

Un control de seguridad diseñado para evitar que ocurra un evento. Estos controles son la primera línea de defensa para evitar el acceso no autorizado o los cambios no deseados en la red. Para obtener más información, consulte Controles preventivos en Implementación de controles de seguridad en AWS.

entidad principal

Una entidad AWS que puede realizar acciones y acceder a los recursos. Esta entidad suele ser un usuario raíz para un Cuenta de AWS rol de IAM o un usuario. Para obtener más información, consulte Entidad principal en Términos y conceptos de roles en la documentación de IAM.

privacidad desde el diseño

Un enfoque de ingeniería de sistemas que tiene en cuenta la privacidad durante todo el proceso de desarrollo.

zonas alojadas privadas

Un contenedor que contiene información sobre cómo desea que Amazon Route 53 responda a las consultas de DNS de un dominio y sus subdominios dentro de uno o más VPCs. Para obtener más información, consulte Uso de zonas alojadas privadas en la documentación de Route 53.

control proactivo

Un control de seguridad diseñado para evitar el despliegue de recursos no conformes. Estos controles escanean los recursos antes de aprovisionarlos. Si el recurso no cumple con el control, significa que no está aprovisionado. Para obtener más información, consulte la guía de referencia de controles en la AWS Control Tower documentación y consulte Controles proactivos en Implementación de controles de seguridad en AWS.

gestión del ciclo de vida del producto (PLM)

La gestión de los datos y los procesos de un producto a lo largo de todo su ciclo de vida, desde el diseño, el desarrollo y el lanzamiento, pasando por el crecimiento y la madurez, hasta el rechazo y la retirada.

entorno de producción

Consulte el entorno.

controlador lógico programable (PLC)

En la fabricación, una computadora adaptable y altamente confiable que monitorea las máquinas y automatiza los procesos de fabricación.

encadenamiento rápido

Utilizar la salida de una solicitud de LLM como entrada para la siguiente solicitud para generar mejores respuestas. Esta técnica se utiliza para dividir una tarea compleja en subtareas o para

refinar o ampliar de forma iterativa una respuesta preliminar. Ayuda a mejorar la precisión y la relevancia de las respuestas de un modelo y permite obtener resultados más detallados y personalizados.

seudonimización

El proceso de reemplazar los identificadores personales de un conjunto de datos por valores de marcadores de posición. La seudonimización puede ayudar a proteger la privacidad personal. Los datos seudonimizados siguen considerándose datos personales.

publish/subscribe (pub/sub)

Un patrón que permite las comunicaciones asíncronas entre microservicios para mejorar la escalabilidad y la capacidad de respuesta. Por ejemplo, en un MES basado en microservicios, un microservicio puede publicar mensajes de eventos en un canal al que se puedan suscribir otros microservicios. El sistema puede añadir nuevos microservicios sin cambiar el servicio de publicación.

Q

plan de consulta

Serie de pasos, como instrucciones, que se utilizan para acceder a los datos de un sistema de base de datos relacional SQL.

regresión del plan de consulta

El optimizador de servicios de la base de datos elige un plan menos óptimo que antes de un cambio determinado en el entorno de la base de datos. Los cambios en estadísticas, restricciones, configuración del entorno, enlaces de parámetros de consultas y actualizaciones del motor de base de datos PostgreSQL pueden provocar una regresión del plan.

R

Matriz RACI

Véase responsable, responsable, consultado, informado (RACI).

RAG

Consulte Recuperación y generación aumentada.

ransomware

Software malicioso que se ha diseñado para bloquear el acceso a un sistema informático o a los datos hasta que se efectúe un pago.

Matriz RASCI

Véase responsable, responsable, consultado, informado (RACI).

RCAC

Consulte control de acceso por filas y columnas.

réplica de lectura

Una copia de una base de datos que se utiliza con fines de solo lectura. Puede enrutar las consultas a la réplica de lectura para reducir la carga en la base de datos principal.

rediseñar

Ver 7 Rs.

objetivo de punto de recuperación (RPO)

La cantidad de tiempo máximo aceptable desde el último punto de recuperación de datos. Esto determina qué se considera una pérdida de datos aceptable entre el último punto de recuperación y la interrupción del servicio.

objetivo de tiempo de recuperación (RTO)

La demora máxima aceptable entre la interrupción del servicio y el restablecimiento del servicio.

refactorizar

Ver 7 Rs.

Región

Una colección de AWS recursos en un área geográfica. Cada uno Región de AWS está aislado e independiente de los demás para proporcionar tolerancia a las fallas, estabilidad y resiliencia. Para obtener más información, consulte Regiones de AWS Especificar qué cuenta puede usar.

regresión

Una técnica de ML que predice un valor numérico. Por ejemplo, para resolver el problema de “¿A qué precio se venderá esta casa?”, un modelo de ML podría utilizar un modelo de regresión lineal para predecir el precio de venta de una vivienda en función de datos conocidos sobre ella (por ejemplo, los metros cuadrados).

volver a alojar

Consulte 7 Rs.

versión

En un proceso de implementación, el acto de promover cambios en un entorno de producción.

trasladarse

Ver 7 Rs.

redefinir la plataforma

Ver 7 Rs.

recompra

Ver 7 Rs.

resiliencia

La capacidad de una aplicación para resistir las interrupciones o recuperarse de ellas. La alta disponibilidad y la recuperación ante desastres son consideraciones comunes a la hora de planificar la resiliencia en el. Nube de AWS Para obtener más información, consulte Nube de AWS Resiliencia.

política basada en recursos

Una política asociada a un recurso, como un bucket de Amazon S3, un punto de conexión o una clave de cifrado. Este tipo de política especifica a qué entidades principales se les permite el acceso, las acciones compatibles y cualquier otra condición que deba cumplirse.

matriz responsable, confiable, consultada e informada (RACI)

Una matriz que define las funciones y responsabilidades de todas las partes involucradas en las actividades de migración y las operaciones de la nube. El nombre de la matriz se deriva de los tipos de responsabilidad definidos en la matriz: responsable (R), contable (A), consultado (C) e informado (I). El tipo de soporte (S) es opcional. Si incluye el soporte, la matriz se denomina matriz RASCI y, si la excluye, se denomina matriz RACI.

control receptivo

Un control de seguridad que se ha diseñado para corregir los eventos adversos o las desviaciones con respecto a su base de seguridad. Para obtener más información, consulte Controles receptivos en Implementación de controles de seguridad en AWS.

retain

Consulte 7 Rs.

jubilarse

Ver 7 Rs.

Generación aumentada de recuperación (RAG)

Tecnología de inteligencia artificial generativa en la que un máster hace referencia a una fuente de datos autorizada que se encuentra fuera de sus fuentes de datos de formación antes de generar una respuesta. Por ejemplo, un modelo RAG podría realizar una búsqueda semántica en la base de conocimientos o en los datos personalizados de una organización. Para obtener más información, consulte Qué es el RAG.

rotación

Proceso de actualizar periódicamente un secreto para dificultar el acceso de un atacante a las credenciales.

control de acceso por filas y columnas (RCAC)

El uso de expresiones SQL básicas y flexibles que tienen reglas de acceso definidas. El RCAC consta de permisos de fila y máscaras de columnas.

RPO

Consulte el objetivo del punto de recuperación.

RTO

Consulte el objetivo de tiempo de recuperación.

manual de procedimientos

Conjunto de procedimientos manuales o automatizados necesarios para realizar una tarea específica. Por lo general, se diseñan para agilizar las operaciones o los procedimientos repetitivos con altas tasas de error.

S

SAML 2.0

Un estándar abierto que utilizan muchos proveedores de identidad (IdPs). Esta función permite el inicio de sesión único (SSO) federado, de modo que los usuarios pueden iniciar sesión AWS

Management Console o llamar a las operaciones de la AWS API sin tener que crear un usuario en IAM para todos los miembros de la organización. Para obtener más información sobre la federación basada en SAML 2.0, consulte Acerca de la federación basada en SAML 2.0 en la documentación de IAM.

SCADA

Consulte el control de supervisión y la adquisición de datos.

SCP

Consulte la política de control de servicios.

secreta

Información confidencial o restringida, como una contraseña o credenciales de usuario, que almacene de forma cifrada. AWS Secrets Manager Se compone del valor secreto y sus metadatos. El valor secreto puede ser binario, una sola cadena o varias cadenas. Para obtener más información, consulta ¿Qué hay en un secreto de Secrets Manager? en la documentación de Secrets Manager.

seguridad desde el diseño

Un enfoque de ingeniería de sistemas que tiene en cuenta la seguridad durante todo el proceso de desarrollo.

control de seguridad

Barrera de protección técnica o administrativa que impide, detecta o reduce la capacidad de un agente de amenazas para aprovechar una vulnerabilidad de seguridad. Existen cuatro tipos principales de controles de seguridad: preventivos, de detección, con capacidad de respuesta y proactivos.

refuerzo de la seguridad

Proceso de reducir la superficie expuesta a ataques para hacerla más resistente a los ataques. Esto puede incluir acciones, como la eliminación de los recursos que ya no se necesitan, la implementación de prácticas recomendadas de seguridad consistente en conceder privilegios mínimos o la desactivación de características innecesarias en los archivos de configuración.

sistema de información sobre seguridad y administración de eventos (SIEM)

Herramientas y servicios que combinan sistemas de administración de información sobre seguridad (SIM) y de administración de eventos de seguridad (SEM). Un sistema de SIEM

recopila, monitorea y analiza los datos de servidores, redes, dispositivos y otras fuentes para detectar amenazas y brechas de seguridad y generar alertas.

automatización de la respuesta de seguridad

Una acción predefinida y programada que está diseñada para responder automáticamente a un evento de seguridad o remediarlo. Estas automatizaciones sirven como controles de seguridad detectables o adaptables que le ayudan a implementar las mejores prácticas AWS de seguridad. Algunos ejemplos de acciones de respuesta automatizadas incluyen la modificación de un grupo de seguridad de VPC, la aplicación de parches a una EC2 instancia de Amazon o la rotación de credenciales.

cifrado del servidor

Cifrado de los datos en su destino, por parte de quien Servicio de AWS los recibe.

política de control de servicio (SCP)

Política que proporciona un control centralizado de los permisos de todas las cuentas de una organización en AWS Organizations. SCPs defina barreras o establezca límites a las acciones que un administrador puede delegar en usuarios o roles. Puede utilizarlas SCPs como listas de permitidos o rechazados para especificar qué servicios o acciones están permitidos o prohibidos. Para obtener más información, consulte las políticas de control de servicios en la AWS Organizations documentación.

punto de enlace de servicio

La URL del punto de entrada de un Servicio de AWS. Para conectarse mediante programación a un servicio de destino, puede utilizar un punto de conexión. Para obtener más información, consulte Puntos de conexión de Servicio de AWS en Referencia general de AWS.

acuerdo de nivel de servicio (SLA)

Acuerdo que aclara lo que un equipo de TI se compromete a ofrecer a los clientes, como el tiempo de actividad y el rendimiento del servicio.

indicador de nivel de servicio (SLI)

Medición de un aspecto del rendimiento de un servicio, como la tasa de errores, la disponibilidad o el rendimiento.

objetivo de nivel de servicio (SLO)

Una métrica objetivo que representa el estado de un servicio, medido mediante un indicador de nivel de servicio.

modelo de responsabilidad compartida

Un modelo que describe la responsabilidad que compartes con respecto a la seguridad y AWS el cumplimiento de la nube. AWS es responsable de la seguridad de la nube, mientras que usted es responsable de la seguridad en la nube. Para obtener más información, consulte el Modelo de responsabilidad compartida.

SIEM

Consulte la información de seguridad y el sistema de gestión de eventos.

punto único de fallo (SPOF)

Una falla en un único componente crítico de una aplicación que puede interrumpir el sistema.

SLA

Consulte el acuerdo de nivel de servicio.

SLI

Consulte el indicador de nivel de servicio.

SLO

Consulte el objetivo de nivel de servicio.

split-and-seed modelo

Un patrón para escalar y acelerar los proyectos de modernización. A medida que se definen las nuevas funciones y los lanzamientos de los productos, el equipo principal se divide para crear nuevos equipos de productos. Esto ayuda a ampliar las capacidades y los servicios de su organización, mejora la productividad de los desarrolladores y apoya la innovación rápida. Para obtener más información, consulte Enfoque gradual para modernizar las aplicaciones en. Nube de AWS

SPOT

Consulte el punto único de falla.

esquema en forma de estrella

Estructura organizativa de una base de datos que utiliza una tabla de datos grande para almacenar datos transaccionales o medidos y una o más tablas dimensionales más pequeñas para almacenar los atributos de los datos. Esta estructura está diseñada para usarse en un almacén de datos o con fines de inteligencia empresarial.

patrón de higo estrangulador

Un enfoque para modernizar los sistemas monolíticos mediante la reescritura y el reemplazo gradual de las funciones del sistema hasta que se pueda desmantelar el sistema heredado. Este patrón utiliza la analogía de una higuera que crece hasta convertirse en un árbol estable y, finalmente, se apodera y reemplaza a su host. El patrón fue presentado por Martin Fowler como una forma de gestionar el riesgo al reescribir sistemas monolíticos. Para ver un ejemplo con la aplicación de este patrón, consulte Modernización gradual de los servicios web antiguos de Microsoft ASP.NET (ASMX) mediante contenedores y Amazon API Gateway.

subred

Un intervalo de direcciones IP en la VPC. Una subred debe residir en una sola zona de disponibilidad.

supervisión, control y adquisición de datos (SCADA)

En la industria manufacturera, un sistema que utiliza hardware y software para monitorear los activos físicos y las operaciones de producción.

cifrado simétrico

Un algoritmo de cifrado que utiliza la misma clave para cifrar y descifrar los datos.

pruebas sintéticas

Probar un sistema de manera que simule las interacciones de los usuarios para detectar posibles problemas o monitorear el rendimiento. Puede usar Amazon CloudWatch Synthetics para crear estas pruebas.

indicador del sistema

Una técnica para proporcionar contexto, instrucciones o pautas a un LLM para dirigir su comportamiento. Las indicaciones del sistema ayudan a establecer el contexto y las reglas para las interacciones con los usuarios.

T

etiquetas

Pares clave-valor que actúan como metadatos para organizar los recursos. AWS Las etiquetas pueden ayudarle a administrar, identificar, organizar, buscar y filtrar recursos. Para obtener más información, consulte Etiquetado de los recursos de AWS.

variable de destino

El valor que intenta predecir en el ML supervisado. Esto también se conoce como variable de resultado. Por ejemplo, en un entorno de fabricación, la variable objetivo podría ser un defecto del producto.

lista de tareas

Herramienta que se utiliza para hacer un seguimiento del progreso mediante un manual de procedimientos. La lista de tareas contiene una descripción general del manual de procedimientos y una lista de las tareas generales que deben completarse. Para cada tarea general, se incluye la cantidad estimada de tiempo necesario, el propietario y el progreso.

entorno de prueba

Consulte entorno.

entrenamiento

Proporcionar datos de los que pueda aprender su modelo de ML. Los datos de entrenamiento deben contener la respuesta correcta. El algoritmo de aprendizaje encuentra patrones en los datos de entrenamiento que asignan los atributos de los datos de entrada al destino (la respuesta que desea predecir). Genera un modelo de ML que captura estos patrones. Luego, el modelo de ML se puede utilizar para obtener predicciones sobre datos nuevos para los que no se conoce el destino.

puerta de enlace de tránsito

Un centro de tránsito de red que puede usar para interconectar sus VPCs redes con las locales. Para obtener más información, consulte Qué es una pasarela de tránsito en la AWS Transit Gateway documentación.

flujo de trabajo basado en enlaces troncales

Un enfoque en el que los desarrolladores crean y prueban características de forma local en una rama de característica y, a continuación, combinan esos cambios en la rama principal. Luego, la rama principal se adapta a los entornos de desarrollo, preproducción y producción, de forma secuencial.

acceso de confianza

Otorgar permisos a un servicio que especifique para realizar tareas en su organización AWS Organizations y en sus cuentas en su nombre. El servicio de confianza crea un rol vinculado al servicio en cada cuenta, cuando ese rol es necesario, para realizar las tareas de administración

por usted. Para obtener más información, consulte AWS Organizations Utilización con otros AWS servicios en la AWS Organizations documentación.

ajuste

Cambiar aspectos de su proceso de formación a fin de mejorar la precisión del modelo de ML. Por ejemplo, puede entrenar el modelo de ML al generar un conjunto de etiquetas, incorporar etiquetas y, luego, repetir estos pasos varias veces con diferentes ajustes para optimizar el modelo.

equipo de dos pizzas

Un DevOps equipo pequeño al que puedes alimentar con dos pizzas. Un equipo formado por dos integrantes garantiza la mejor oportunidad posible de colaboración en el desarrollo de software.

U

incertidumbre

Un concepto que hace referencia a información imprecisa, incompleta o desconocida que puede socavar la fiabilidad de los modelos predictivos de ML. Hay dos tipos de incertidumbre: la incertidumbre epistémica se debe a datos limitados e incompletos, mientras que la incertidumbre aleatoria se debe al ruido y la aleatoriedad inherentes a los datos. Para más información, consulte la guía Cuantificación de la incertidumbre en los sistemas de aprendizaje profundo.

tareas indiferenciadas

También conocido como tareas arduas, es el trabajo que es necesario para crear y operar una aplicación, pero que no proporciona un valor directo al usuario final ni proporciona una ventaja competitiva. Algunos ejemplos de tareas indiferenciadas son la adquisición, el mantenimiento y la planificación de la capacidad.

entornos superiores

Ver entorno.

V

succión

Una operación de mantenimiento de bases de datos que implica limpiar después de las actualizaciones incrementales para recuperar espacio de almacenamiento y mejorar el rendimiento.

control de versión

Procesos y herramientas que realizan un seguimiento de los cambios, como los cambios en el código fuente de un repositorio.

Emparejamiento de VPC

Una conexión entre dos VPCs que le permite enrutar el tráfico mediante direcciones IP privadas. Para obtener más información, consulte ¿Qué es una interconexión de VPC? en la documentación de Amazon VPC.

vulnerabilidad

Defecto de software o hardware que pone en peligro la seguridad del sistema.

W

caché caliente

Un búfer caché que contiene datos actuales y relevantes a los que se accede con frecuencia. La instancia de base de datos puede leer desde la caché del búfer, lo que es más rápido que leer desde la memoria principal o el disco.

datos templados

Datos a los que el acceso es infrecuente. Al consultar este tipo de datos, normalmente se aceptan consultas moderadamente lentas.

función de ventana

Función SQL que realiza un cálculo en un grupo de filas que se relacionan de alguna manera con el registro actual. Las funciones de ventana son útiles para procesar tareas, como calcular una media móvil o acceder al valor de las filas en función de la posición relativa de la fila actual.

carga de trabajo

Conjunto de recursos y código que ofrece valor comercial, como una aplicación orientada al cliente o un proceso de backend.

flujo de trabajo

Grupos funcionales de un proyecto de migración que son responsables de un conjunto específico de tareas. Cada flujo de trabajo es independiente, pero respalda a los demás flujos de trabajo del proyecto. Por ejemplo, el flujo de trabajo de la cartera es responsable de priorizar las aplicaciones, planificar las oleadas y recopilar los metadatos de migración. El flujo de trabajo de la cartera entrega estos recursos al flujo de trabajo de migración, que luego migra los servidores y las aplicaciones.

GUSANO

Mira, escribe una vez, lee muchas.

WQF

Consulte el marco AWS de calificación de la carga de trabajo.

escribe una vez, lee muchas (WORM)

Un modelo de almacenamiento que escribe los datos una sola vez y evita que se eliminen o modifiquen. Los usuarios autorizados pueden leer los datos tantas veces como sea necesario, pero no pueden cambiarlos. Esta infraestructura de almacenamiento de datos se considera inmutable.

Z

ataque de día cero

Un ataque, normalmente de malware, que aprovecha una vulnerabilidad de día cero.

vulnerabilidad de día cero

Un defecto o una vulnerabilidad sin mitigación en un sistema de producción. Los agentes de amenazas pueden usar este tipo de vulnerabilidad para atacar el sistema. Los desarrolladores suelen darse cuenta de la vulnerabilidad a raíz del ataque.

aviso de tiro cero

Proporcionar a un LLM instrucciones para realizar una tarea, pero sin ejemplos (imágenes) que puedan ayudar a guiarla. El LLM debe utilizar sus conocimientos previamente entrenados para

realizar la tarea. La eficacia de las indicaciones cero depende de la complejidad de la tarea y de la calidad de las indicaciones. Consulte también las indicaciones de pocos pasos.

aplicación zombi

Aplicación que utiliza un promedio de CPU y memoria menor al 5 por ciento. En un proyecto de migración, es habitual retirar estas aplicaciones.

Las traducciones son generadas a través de traducción automática. En caso de conflicto entre la traducción y la version original de inglés, prevalecerá la version en inglés.