Creación de arquitecturas multiusuario para la IA de los agentes en AWS

AWS Guía prescriptiva

AWS Guía prescriptiva: Creación de arquitecturas multiusuario para la IA de los agentes en AWS

Table of Contents

Creación de arquitecturas multiusuario para la IA de los agentes en AWS

Aaron Sempf y Tod Golding, Amazon Web Services

Julio de 2025 (historial del documento)

La IA de Agentic representa un cambio de paradigma disruptivo que requiere que las organizaciones se replanteen la forma de construir, entregar y operar sus sistemas. El modelo de agencia hace que los equipos exploren nuevas formas de descomponer los sistemas en uno o más agentes que creen nuevos caminos, posibilidades y valores.

Gran parte del debate entre los agentes se centra en las herramientas, los marcos y los patrones que se utilizan para crear e implementar agentes. No solo debemos adoptar buenas herramientas para crear agentes, sino también nuevos protocolos de integración, estrategias de autenticación y mecanismos de descubrimiento que puedan servir de base para las arquitecturas de los agentes.

A medida que aumenta la cantidad de herramientas de los agentes, los equipos también deben considerar la forma en que sus agentes abordan los desafíos de la arquitectura más tradicional. La escalabilidad, la proximidad ruidosa, la resiliencia, el coste y la eficiencia operativa son aspectos fundamentales que deben evaluarse a la hora de diseñar, crear e implementar agentes. Independientemente de lo autónomos e inteligentes que sean los agentes, también debemos asegurarnos de que logren economías de escala, eficiencia y agilidad que se ajusten a las necesidades empresariales.

El objetivo de esta guía es explorar las diversas dimensiones de la presencia de los agentes. Esto incluye revisar varios patrones de despliegue y consumo de agentes y destacar diferentes estrategias para crear agentes que aborden los objetivos arquitectónicos. También implica analizar cómo se podrían consumir los agentes en un entorno con varios inquilinos mediante la introducción de estructuras internas que normalmente se requieren en un entorno con varios inquilinos.

Destinatarios previstos

Esta guía está dirigida a arquitectos, desarrolladores y líderes tecnológicos que desean crear sistemas multiusuario basados en la IA.

Objetivos

Esta guía lo ayuda a hacer lo siguiente:

- Conozca los despliegues de agentes multiusuario y explore los modelos aislados y agrupados, y cómo el contexto del inquilino afecta a la implementación de los agentes
- Explore la administración de agentes, incluida la incorporación, el aislamiento de inquilinos y la administración de recursos en entornos de uno o varios proveedores
- Evalúe los aspectos de los agentes multiusuario, como la propiedad de los datos, la supervisión y las pruebas

Acerca de esta serie de contenido

Esta guía forma parte de un conjunto de publicaciones que proporcionan planos arquitectónicos y orientación técnica para crear agentes de software basados en la IA. AWS La serie de guías AWS prescriptivas incluye las siguientes guías:

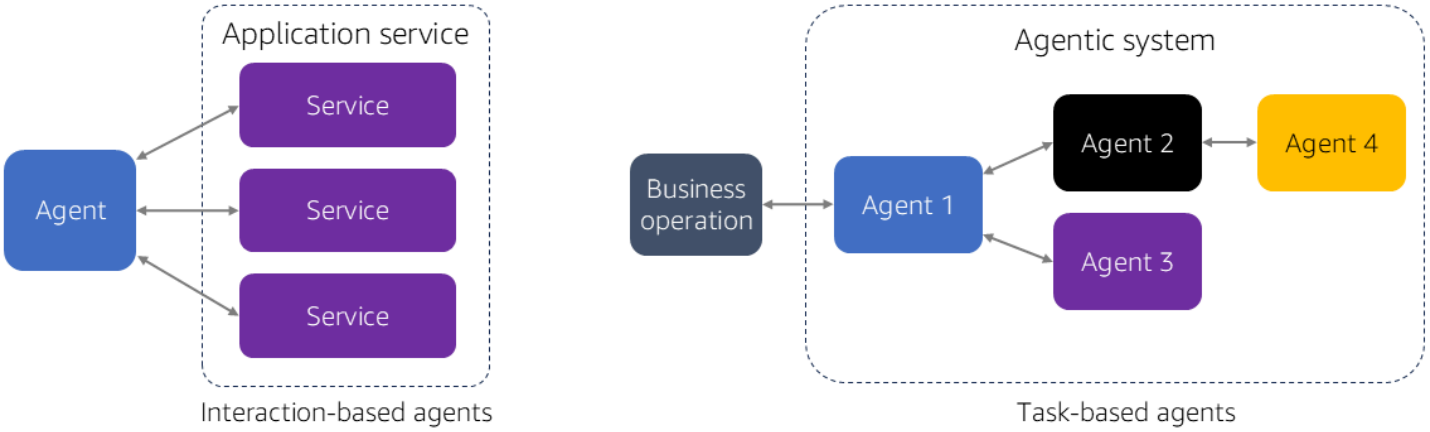
- Operacionalización de la IA de los agentes en AWS
- Fundamentos de la IA agencial en AWS
- Los patrones y flujos de trabajo de la IA de los agentes están activos AWS
- Los marcos, protocolos y herramientas de inteligencia artificial de las agencias están disponibles AWS
- Creación de arquitecturas sin servidor para la IA de los agentes en AWS
- Creación de arquitecturas multiusuario para la IA de los agentes (esta guía) AWS

Para obtener más información sobre esta serie de contenido, consulte Agentic AI.

Fundamentos de los agentes

Antes de analizar los detalles de la arquitectura, debemos describir las diferentes funciones que desempeñan los agentes, ya que «agente» es un término sobrecargado que se puede aplicar a muchos casos de uso. Empecemos con algunos términos generales que pueden ayudar a clasificarlos.

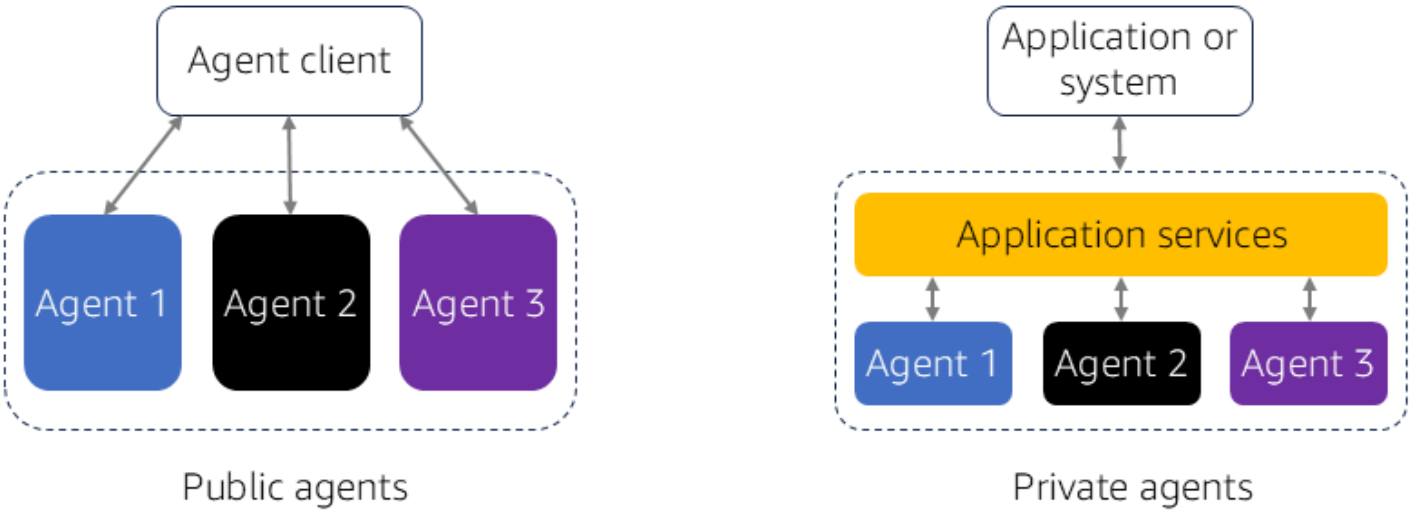
En el nivel más externo, debemos empezar por clasificar el papel y la naturaleza de los agentes. Esto es un desafío porque hay una amplia gama de escenarios en los que los agentes se pueden aplicar a cualquier tipo de problemas. Sin embargo, en este análisis nos centramos en lo que significa introducir un agente en una aplicación o un sistema. En este modelo, hacemos hincapié en cómo y dónde los agentes pueden enriquecer mejor la experiencia de su sistema. Las opciones que elija influyen en la forma en que sus agentes se crean, integran y aplican a los diferentes dominios y casos de uso. El siguiente diagrama muestra dos patrones de agentes que utilizan los desarrolladores.



En la parte izquierda del diagrama hay un agente basado en interacciones. En este modo, un agente crea una vista de un sistema existente para organizar las interacciones con los servicios subyacentes a fin de lograr un objetivo o un resultado. La clave es que el agente se añada a un sistema como un enfoque alternativo para impulsar las funciones y capacidades del sistema. Imagine, por ejemplo, que un proveedor de software independiente (ISV) tiene un sistema de contabilidad con una experiencia de usuario que se utiliza para realizar operaciones. El agente basado en interacciones simplifica la interacción con estas capacidades existentes. Se trata menos de aprender a alcanzar un objetivo poco definido y más de proporcionar una forma de organizar las vías conocidas.

Por el contrario, el sistema basado en tareas que aparece en la parte derecha del diagrama representa un enfoque diferente. Los agentes de ese sistema utilizan sus conocimientos y habilidades para aprender a completar las tareas e impulsar los resultados empresariales. Se podría argumentar que ambos modelos logran resultados empresariales, pero un modelo basado en tareas depende de los propios agentes para determinar cómo lograr un resultado. Estos agentes son menos deterministas y, en cambio, se basan en su capacidad para aprender y evolucionar. Por el contrario, los agentes basados en interacciones están diseñados principalmente para orquestar un conjunto de capacidades conocidas. Estas diferencias afectan a la forma en que crea, selecciona e integra los agentes para respaldar su empresa.

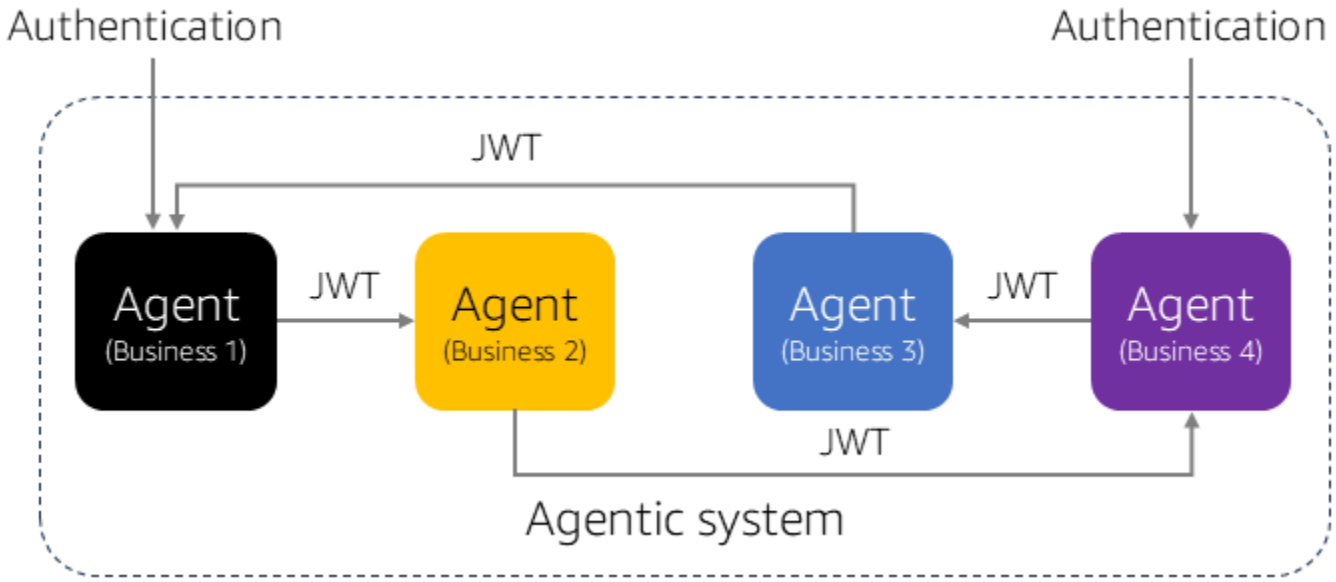
También necesitamos términos que describan cómo y dónde desplegamos los agentes. El lugar en el que se encuentre un agente dentro del espacio físico de su sistema puede influir en la forma en que se construye, se enfoca y se protege. El siguiente diagrama describe dos modelos distintos que podrían aplicarse a los agentes.



En la parte izquierda del diagrama hay un sistema de despliegue con tres agentes diferentes. Los agentes están expuestos a clientes externos que pueden ser otros agentes o aplicaciones. Para este modelo, los agentes se denominan agentes públicos.

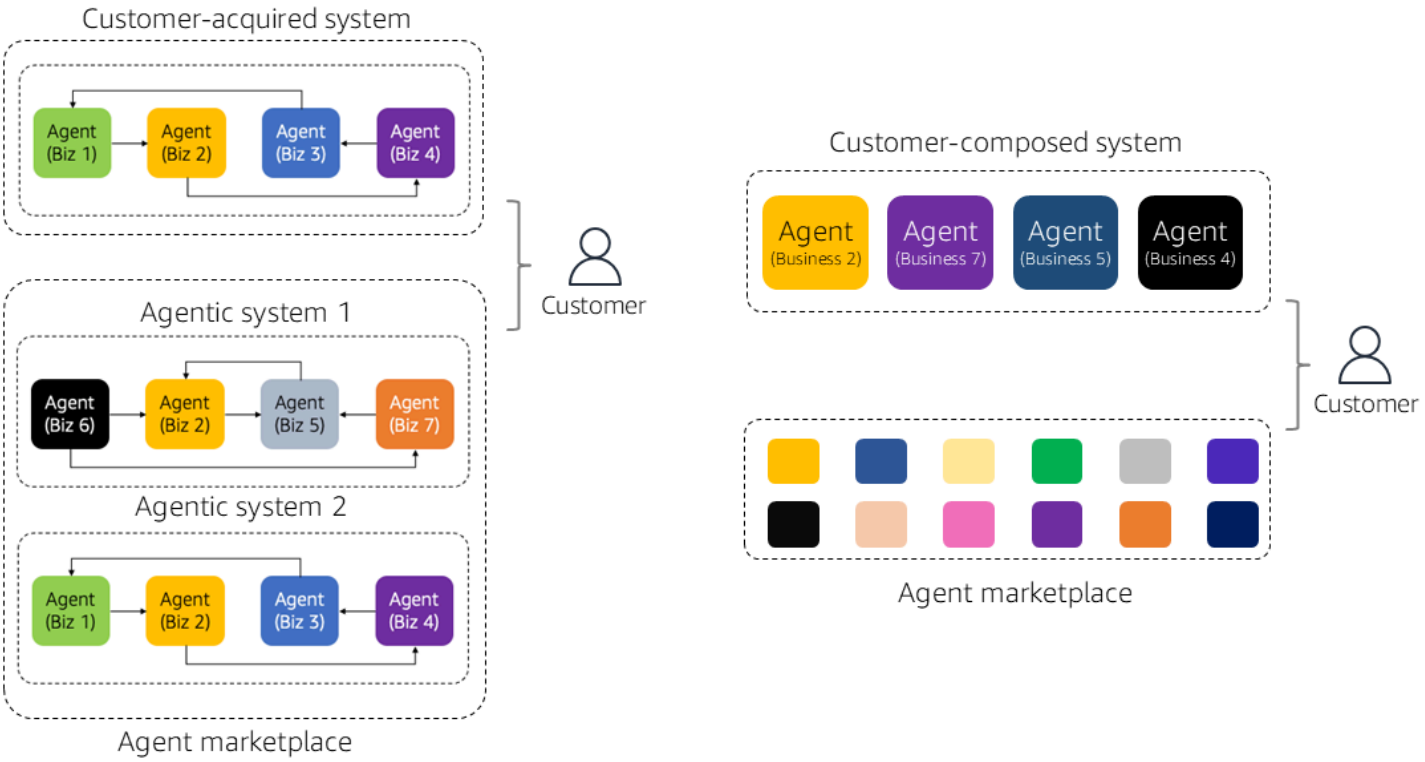
Por el contrario, el diagrama de la derecha muestra los agentes dentro de la implementación de la solución. En este caso, hay una serie de servicios de aplicaciones que consumen los usuarios o los sistemas. Estos usuarios interactúan con la aplicación sin darse cuenta de que los agentes son parte de la experiencia. A continuación, los servicios del sistema subyacente invocan y organizan los agentes. Los agentes desplegados de esta manera se denominan agentes privados.

Gran parte del valor de un agente se centra en el modelo público, en el que los proveedores pueden publicar a sus agentes con la intención de integrarlos con otros agentes externos. De este modo, los agentes formarían parte de una malla o red de servicios interconectados que, en conjunto, podrían abordar muchos casos de uso. Si bien estos agentes podrían usarse en muchos dominios, el caso de business-to-business uso es una opción natural. El siguiente diagrama proporciona una vista conceptualizada de lo que sería ensamblar un agente de recolección que resuelva un problema específico.



El diagrama muestra cuatro agentes comerciales que trabajan juntos para lograr un conjunto de objetivos. Cuando los agentes están compuestos de esta manera, representan un sistema agencial, y hay muchos tipos de sistemas de este tipo. Podrían ser un conjunto preempaquetado de agentes colaboradores que, por lo general, se consumen como una sola unidad. O bien, el sistema podría ser ensamblado dinámicamente por los clientes que deseen seleccionar la combinación de agentes que mejor se adapte a sus necesidades.

Ambos enfoques ofrecen vías viables para la integración de los agentes. Algunos agentes se crean con la expectativa de que se integren en sistemas específicos donde puedan maximizar su valor, alcance e impacto. Esta noción de sistemas de agencia también plantea dudas sobre cómo se adquieren los agentes, y podría haber muchas formas de abordarlo. El siguiente diagrama proporciona ejemplos de cómo se pueden crear estos agentes y sistemas a través de experiencias transaccionales.

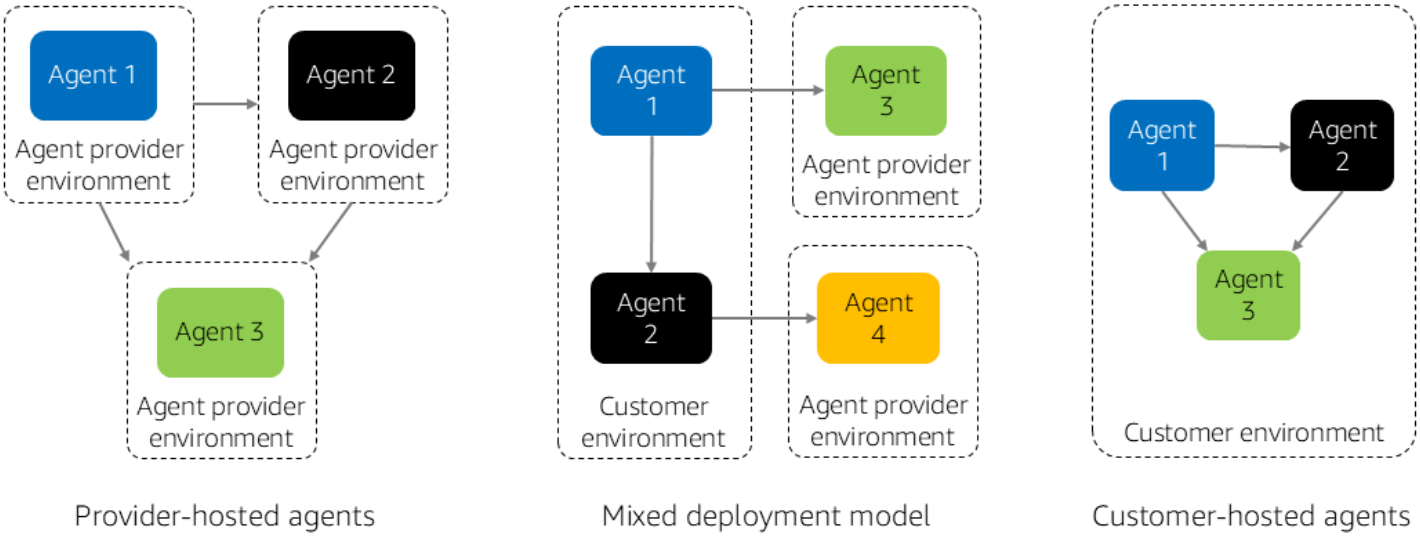


Se muestran dos ejemplos de experiencias en el mercado. En el lado izquierdo, se utiliza un mercado para adquirir sistemas preempaquetados. En este escenario, el mercado descubre e incorpora sistemas que abordan objetivos más amplios que requieren la integración y la organización de varios agentes.

El ejemplo de la derecha muestra un mercado en el que los agentes son descubiertos y agrupados en sistemas de agentes. En este escenario, los clientes pueden crear cualquier sistema de agentes integrados y compatibles que satisfaga sus necesidades. La capacidad de ensamblar agentes de esta manera depende del modelo de compatibilidad y de los requisitos de integración de los agentes individuales.

Consideraciones sobre el alojamiento de agentes

Ahora que ya tiene una idea más amplia de los conceptos de agencia, analicemos qué significa alojar y administrar estos agentes. Debemos pensar en cómo y dónde se ejecutan los cálculos, cómo se escalan, cómo funcionan y cómo se administran. Al mismo tiempo, algunos patrones que esperamos ver como agentes se están aplicando y adoptando más ampliamente. El siguiente diagrama muestra un ejemplo de posibles permutaciones.



Aquí se representan tres estrategias distintas. En la parte izquierda del diagrama, puede ver un modelo en el que nuestros agentes se alojan, escalan y administran dentro de los entornos de cada proveedor de agentes. Estos agentes se publican y consumen como servicios, y funcionan según lo que se denomina un modelo de agente como servicio (AAaS). En el lado derecho hay un modelo en el que todos los agentes de un proveedor están alojados en un entorno dedicado al cliente.

En la mitad del diagrama hay un modelo de implementación mixto que combina estas dos estrategias: aloja a algunos agentes de forma local en el entorno del cliente e interactúa con algunos agentes que están alojados de forma remota en el entorno de un proveedor.

Una cuarta opción (que no se muestra) podría consistir en que los agentes se diseñen como servicios con poco código o sin código, escalados y gestionados mediante servicios de infraestructura de agentes. No los abordaremos en detalle porque la arquitectura y el alojamiento de los agentes gestionados dependen principalmente de la organización propietaria de los servicios.

Puede imaginarse la variedad de factores que podrían influir en la adopción de uno de estos modelos. Las restricciones de conformidad, reglamentarias y de seguridad, por ejemplo, podrían

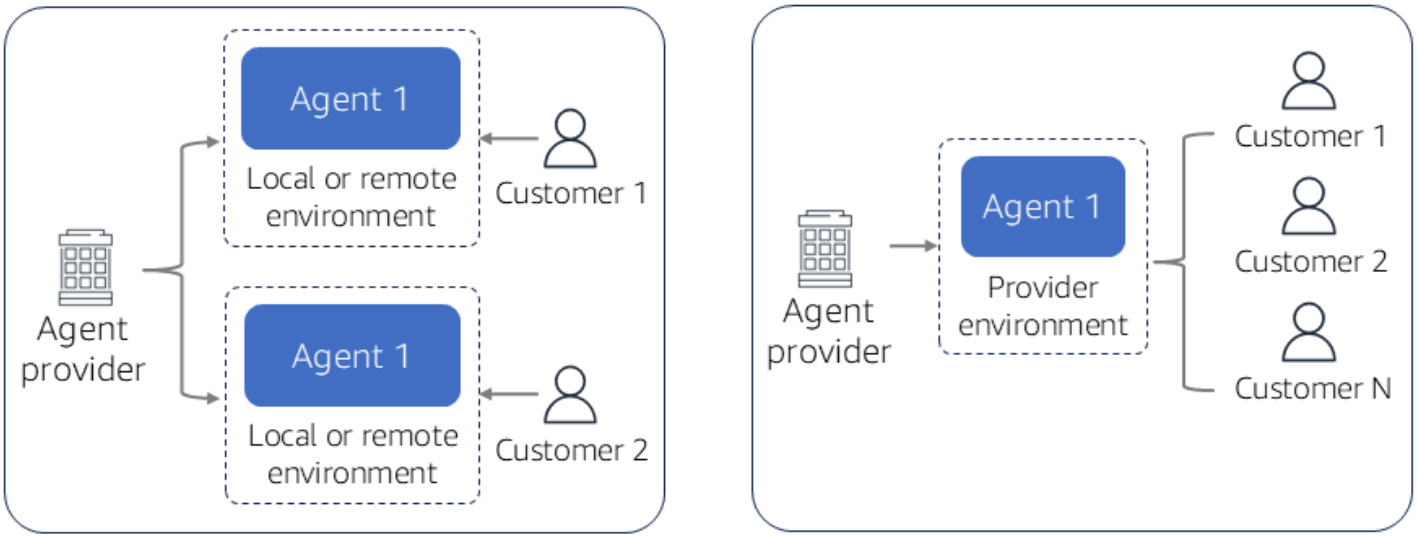
llevar a alguien a optar por agentes alojados por el cliente. La escala, la agilidad y la eficiencia podrían impulsar a las organizaciones a adoptar más el modelo AAA.

El concepto clave en este caso es que los agentes pueden y están desplegados y alojados de muchas maneras. Su trabajo consiste en determinar cuál es la mejor manera de utilizar los agentes. El espacio, la seguridad y el despliegue, entre otros factores, afectan considerablemente a la forma en que se abordan la creación y el funcionamiento de los agentes. Los agentes públicos y privados, por ejemplo, pueden tener diseños y ciclos de vida de lanzamiento diferentes.

Los agentes cumplen con la multitenencia

Es fácil pensar en los agentes como componentes básicos, ya que los agentes se consideran una serie de componentes autónomos que se ensamblan para satisfacer las necesidades de un dominio o problema empresarial específico. Lo que resulta más interesante es cuando empezamos a pensar en cómo los proveedores empaquetan y consumen estos agentes. En muchos sentidos, un agente se convierte en una fuente de costes e ingresos para una empresa. Los agentes proveedores deben tener en cuenta las diferentes personas que consumen sus servicios, el perfil de consumo de las personas y las estrategias de monetización que les permiten crear modelos de precios y niveles que se adapten a los de los consumidores.

Los proveedores de agentes podrían admitir varios modelos de despliegue de sus agentes para satisfacer las necesidades de los clientes. El siguiente diagrama muestra una vista conceptual de los dos modelos principales de despliegue de agentes.



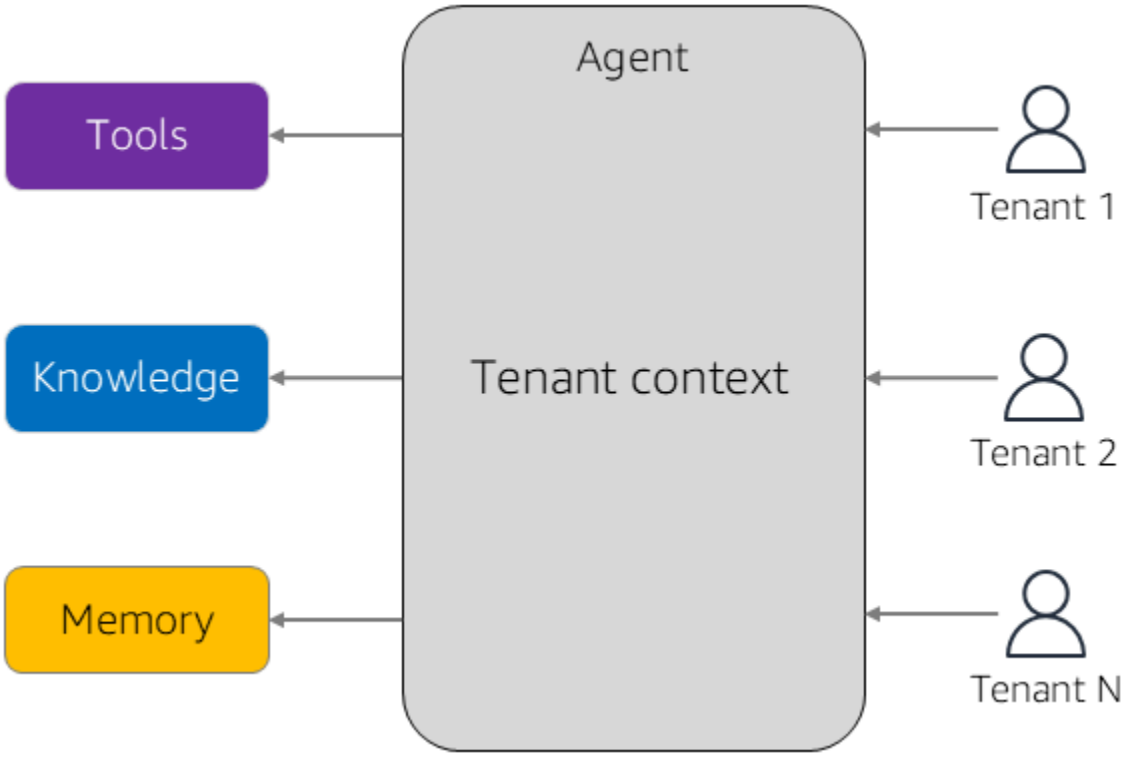
En la parte izquierda del diagrama se muestra el modelo de agente dedicado al cliente. Un proveedor de agentes crea un agente mediante la implementación de una instancia de agente independiente para cada cliente incorporado. Con este enfoque, las capacidades del agente y su capacidad para adquirir conocimientos se limitarían al ámbito del entorno de un cliente determinado. Esto acaba representando una experiencia por cliente que hereda algunas de las complejidades y ventajas de ofrecer un entorno de cliente específico.

Por el contrario, el diagrama de la parte derecha del diagrama muestra un único agente que se implementa en el entorno del proveedor. El agente procesa las solicitudes de varios clientes,

evolucionando y aprendiendo en función de la experiencia colectiva de todos los clientes. Cada nuevo cliente que se añada simplemente representará a otro cliente válido del agente. El agente funciona como un modelo de agente como servicio (AAA), utilizando estructuras compartidas para satisfacer las necesidades del cliente. En ambos casos, los agentes consumidores pueden ser aplicaciones, sistemas o incluso otros agentes.

Hay dos maneras de analizar el modelo AAA. El modelo anterior ofrece la misma experiencia a todos los clientes. Esto significa que los aspectos internos del agente no incluirán ningún nivel de especialización que tenga en cuenta el contexto del cliente solicitante. En general, en este modo, se parte del supuesto de que la naturaleza del alcance, los objetivos y el valor de un agente se centra en un conjunto compartido de recursos, conocimientos y resultados que se aplican de forma universal a todos los clientes.

El enfoque alternativo al AAA es aquel en el que el contexto de los clientes influye en la experiencia y la implementación del agente. El siguiente diagrama proporciona una visión conceptual de la presencia de un agente de la AAA en este contexto.



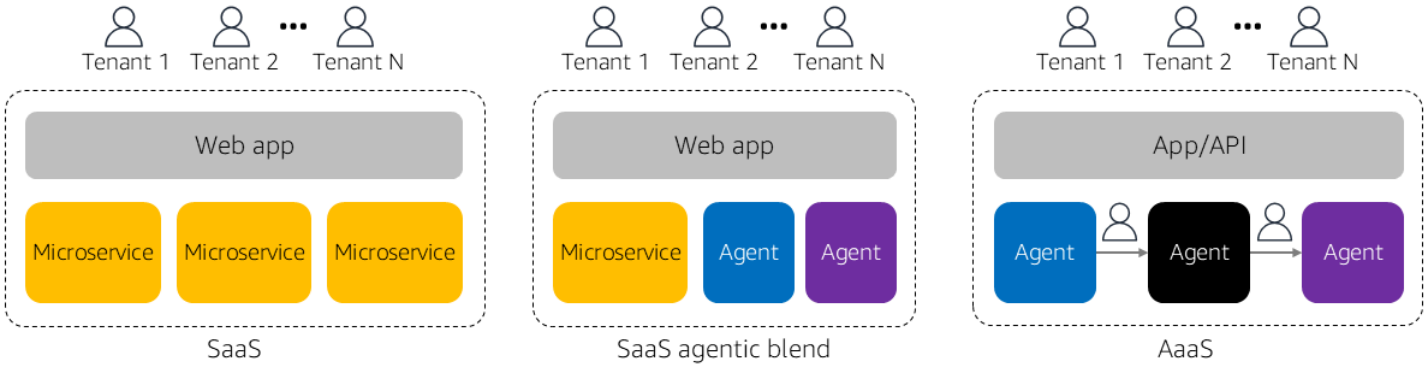
Desde el punto de vista de la AAA, el origen y el contexto de las solicitudes entrantes afectan significativamente a la presencia del agente. Los recursos, las acciones y las herramientas que forman parte de la implementación subyacente del agente pueden variar para cada solicitud entrante del inquilino. El valor de un agente está relacionado con su capacidad de utilizar el contexto del

inquilino para determinar acciones y resultados en los que influyen la situación del inquilino, sus conocimientos y otros factores. Algunas solicitudes pueden generar un resultado único para cada inquilino, y otras pueden generar resultados más personalizados para cada inquilino. Esto añade una nueva dimensión a la capacidad de aprendizaje del agente, que podría incluir ser más contextual y adquirir y aplicar conocimientos que mejoren los resultados esperados.

Para los proveedores, el modelo AAA ofrece muchas ventajas. Dado que varios clientes consumen un solo agente, el proveedor tiene una mejor oportunidad de lograr economías de escala, impulsar la eficiencia operativa, controlar los costos y crear una experiencia de administración unificada. Esto tiene el potencial de aumentar la agilidad, la innovación y el crecimiento del negocio de los agentes.

Estas cualidades se superponen con los mismos principios que impulsan la adopción del modelo de software como servicio (SaaS). Básicamente, el modelo SaaS está diseñado como un servicio multiusuario que hereda muchos de los mismos atributos de escala, resiliencia, aislamiento, incorporación y operativos que se encuentran en un entorno SaaS. En muchos aspectos, la experiencia de SaaS se basa en gran medida en las estrategias y prácticas utilizadas por los proveedores de SaaS, pero es razonable separar estos términos. Para nuestros propósitos, el énfasis se centra principalmente en las implicaciones que conlleva la creación y operación de agentes que requieren el apoyo de múltiples inquilinos.

En el caso de un sistema que pueda tratar a todos los usuarios por igual y no requiera la gestión de datos persistentes, confidenciales o específicos de los clientes, la noción de arrendamiento afectaría mínimamente a sus agentes. En el caso de los sistemas que se espera que atiendan a varios clientes y, al mismo tiempo, preserven el aislamiento de los datos, la personalización y el conocimiento del contexto, dar soporte a varios usuarios podría ser un elemento esencial del diseño, la estrategia y el objetivo de un agente. El siguiente diagrama muestra cómo se puede utilizar la multitenencia en entornos de agencias.



En el lado izquierdo de este diagrama hay una arquitectura multiusuario clásica. Incluye una aplicación web y una serie de microservicios que implementan la lógica empresarial. Varios inquilinos

consumen la infraestructura compartida de este entorno y se amplía para adaptarse a las cargas de trabajo cambiantes de una población de inquilinos en constante evolución. El entorno se gestiona y gestiona a través de un único panel de vidrio para todos los inquilinos.

Imagínese cómo este modelo mental representa al agente situado en el lado derecho de este diagrama. Un agente utiliza un modelo AAA que utilizan uno o más inquilinos. Los agentes pueden provenir de varios proveedores y el contexto de los inquilinos fluya entre ellos, ya que una sola instancia de un agente debe procesar las solicitudes de varios inquilinos.

El ejemplo que aparece en la mitad de este diagrama es un modelo híbrido en el que los agentes forman parte de la experiencia general de SaaS. Algunas partes del sistema se implementan en un modelo más tradicional y otras partes del sistema se basan en los agentes. Es probable que este patrón sea común en muchas ofertas de SaaS, especialmente para las organizaciones que están haciendo la transición a una experiencia de agente. Es habitual que este modelo persista, ya que no todos los sistemas se suministran únicamente con tecnología AAA. Tenga en cuenta también que la multitenencia sigue aplicándose a los agentes del modelo. Si bien los agentes pueden estar integrados en un sistema, aún pueden procesar las solicitudes de varios inquilinos.

Es natural preguntarse si la multitenencia es realmente importante. Se podría argumentar que un agente procesa las solicitudes, por lo que respaldar el arrendamiento puede tener poco efecto. Sin embargo, a medida que profundizamos en las implicaciones de la gestión de múltiples inquilinos, el arrendamiento puede afectar directamente a la forma en que los agentes influyen en la forma en que los agentes acceden, despliegan y configuran las herramientas, la memoria, los datos y otras partes del agente para ayudar a los inquilinos individuales. El arrendamiento también influye en la forma en que el escalado, la regulación, los precios, la estratificación y otros aspectos empresariales se aplican a la arquitectura de su agente.

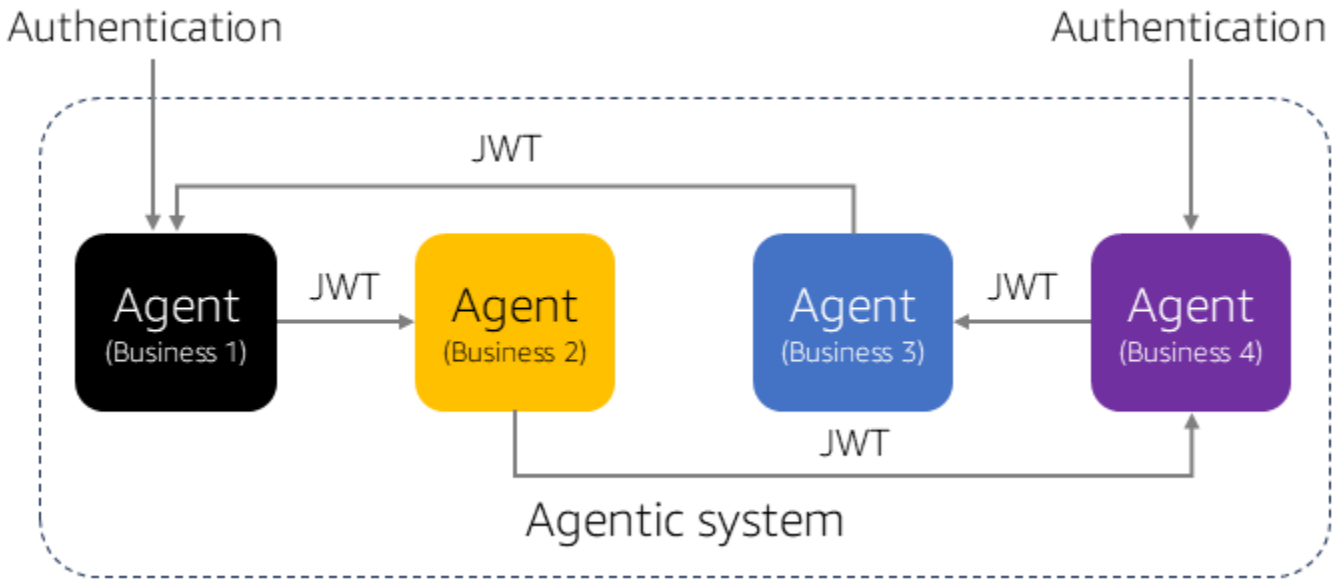
Una de las conclusiones de todo esto es que hay casos de uso de agentes que requieren el apoyo de varios inquilinos. El desafío consiste en determinar cómo la multitenencia da forma al diseño y la arquitectura generales de la experiencia de su agencia. Para algunos agentes, el soporte multiusuario representa una capacidad diferenciadora, ya que permite a los agentes aplicar el contexto específico del inquilino a los agentes para obtener los resultados esperados.

En las siguientes secciones, verá la utilidad de la terminología y los patrones de diseño que creamos para describir las arquitecturas SaaS multiusuario. El modelo AAA puede adoptar estos conceptos tomando prestados aspectos útiles, que introducen nuevos conceptos específicos de los agentes cuando son necesarios.

Identidad, contexto del inquilino y sistemas de agencia

Añadir el contexto de los inquilinos a los agentes individuales no es particularmente difícil. En muchos casos, los equipos pueden confiar en los mecanismos típicos que vinculan a los usuarios y los sistemas con los inquilinos y transfieren a los agentes los tokens compatibles con los inquilinos. Esto es relevante si tenemos en cuenta cómo el contexto y la identidad de los inquilinos ayudan a varios agentes. En este modelo, los inquilinos deben estar vinculados a una identidad que abarque a todos los agentes colaboradores.

En general, el ámbito de las agencias requiere un modelo de identidad más transversal que se ajuste a las necesidades actuales y emergentes de los sistemas de las agencias. Los proveedores de agentes requieren mecanismos de identidad que respalden los modelos únicos de seguridad, cumplimiento y autorización que vienen con los sistemas de agentes operativos. Esto es especialmente difícil en entornos en los que los sistemas están compuestos por clientes u otros agentes. Cada agente incorporado debe conectar su identidad y el contexto del inquilino con las interacciones entre los agentes. El siguiente diagrama destaca los posibles desafíos relacionados con la identidad y el contexto del inquilino que forman parte de las interacciones agent-to-agent (a2a).



Este diagrama muestra una serie de agentes creados por un proveedor que interactúan como parte del sistema de agentes que analizamos. Ahora está modernizado con el contexto de identidad y arrendatario. Este escenario es un ejemplo de un sistema de agencia que admite varios puntos de entrada. Suponemos que cada agente de este sistema requiere su propio mecanismo de autenticación para resolver el problema del sistema o el usuario ante un inquilino determinado. A

medida que estos agentes interactúan, el contexto del inquilino pasa a un token web JSON (JWT) que se utilizará para autorizar el acceso e inyectar el contexto del inquilino en el agente.

Conceptualmente, la principal diferencia con este escenario es que los agentes se despliegan y funcionan de forma independiente, lo que significa que cada agente debe poder determinar su identidad y autorizar el acceso. La clave es que su identidad debe tener alguna capacidad distribuida para gestionar las necesidades del sistema de agentes en general. También debe haber una alineación en la forma en que los agentes comparten el contexto de los inquilinos.

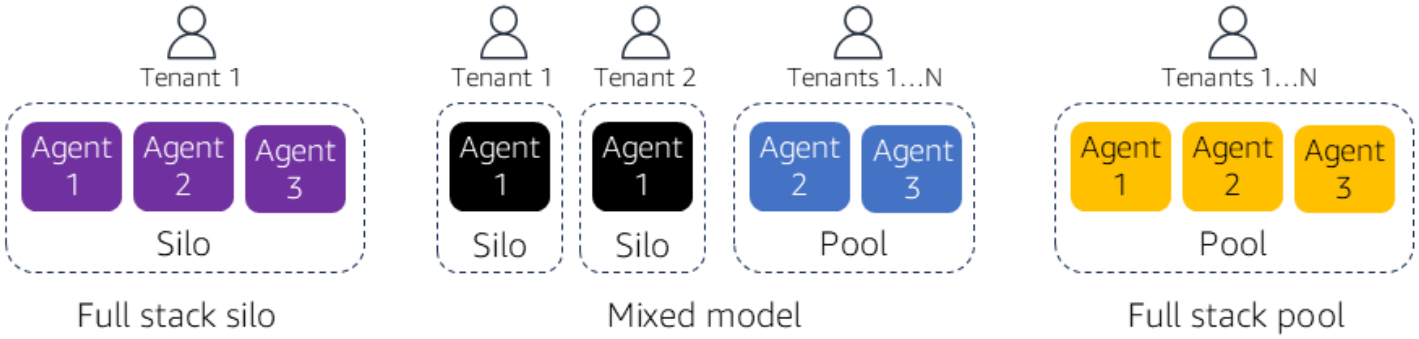
Aplicando el valor empresarial de SaaS a SaaS

Por lo general, cuando analizamos la ejecución de cualquier sistema en un as-a-service modelo, tenemos en cuenta la naturaleza de la experiencia y la forma en que su huella técnica y operativa impulsa los resultados empresariales. Al adoptar el SaaS, por ejemplo, las organizaciones utilizan las economías de escala, la eficiencia operativa, los perfiles de costes y la agilidad para impulsar el crecimiento, los márgenes y la innovación.

Es probable que los agentes que se entreguen como AAA tengan como objetivo obtener resultados empresariales similares. Al dar soporte a varios inquilinos, un agente puede alinear el consumo de recursos con las actividades de los inquilinos. Esto produce economías de escala propias de los entornos SaaS tradicionales. El AAAs también permite a las organizaciones gestionar, operar e implementar agentes de una forma que posibilita la publicación frecuente de versiones e impulsa la agilidad de los proveedores de agentes. La clave es que el modelo AAAs no depende de la tecnología. Crea e impulsa estrategias empresariales que promueven el crecimiento, agilizan la adopción y simplifican las operaciones.

Modelos de despliegue de agentes

En una experiencia de AAA básica, un proveedor puede implementar agentes siguiendo varios patrones. Existen innumerables factores que influyen en la forma en que se despliegan los agentes para satisfacer las necesidades de los clientes, el rendimiento, el cumplimiento, la geografía y la seguridad. Las diferentes estrategias de despliegue afectan a la forma en que se diseña, implementa y consume un agente. Es aquí donde podemos introducir los términos clásicos de varios usuarios para etiquetar las diferentes estrategias de despliegue. En el siguiente diagrama, se muestran diferentes permutaciones para implementar agentes en un entorno AAA.



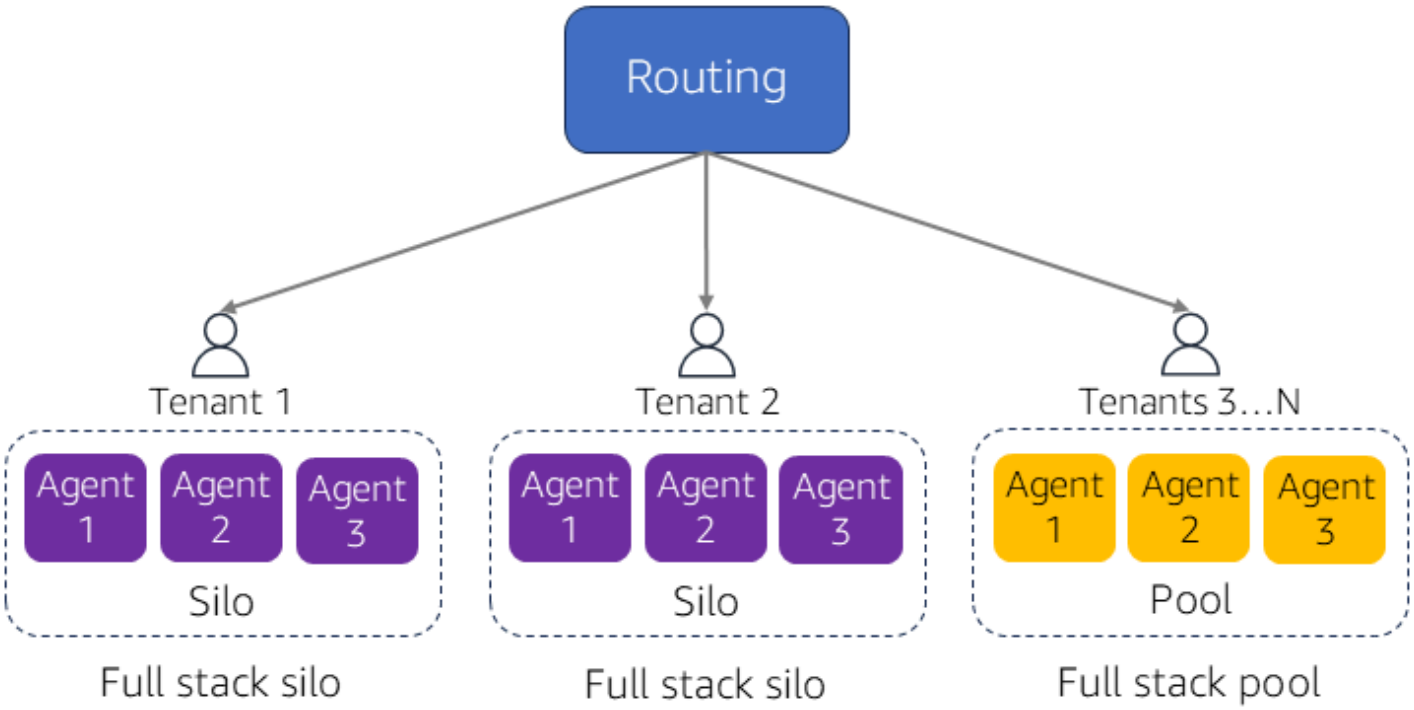
Este diagrama representa tres modos de despliegue de agentes. En el lado izquierdo hay un modelo aislado, en el que cada inquilino cuenta con una experiencia totalmente aislada y un conjunto exclusivo de agentes. En este escenario, los agentes no comparten los entornos de cómputo, recursos o ejecución entre los inquilinos.

El ejemplo intermedio ilustra un modelo híbrido, en el que los inquilinos utilizan una combinación de agentes agrupados y en silos. Por ejemplo, el agente 1 se despliega en modo aislado (cada inquilino recibe una instancia dedicada), mientras que los agentes 2 y 3 funcionan en un modelo agrupado y comparten los recursos entre los inquilinos.

En el lado derecho hay un modelo totalmente agrupado, en el que todos los agentes se comparten entre los inquilinos, lo que ofrece una implementación clásica con varios inquilinos. En este escenario, los inquilinos aprovechan una infraestructura común de procesamiento, memoria y servicios para la ejecución de los agentes.

La idea es que los agentes puedan operar en diferentes modelos de implementación, con recursos informáticos y dependientes dedicados (en silos) o compartidos (agrupados) entre los inquilinos. Estas estrategias de despliegue no se excluyen mutuamente. Los servicios de agente suelen satisfacer una variedad de necesidades de los clientes y combinan ambos modelos para equilibrar el

rendimiento, el aislamiento, el costo y la escalabilidad. El siguiente diagrama muestra un sistema de agencia que admite múltiples configuraciones de implementación en el mismo entorno operativo.



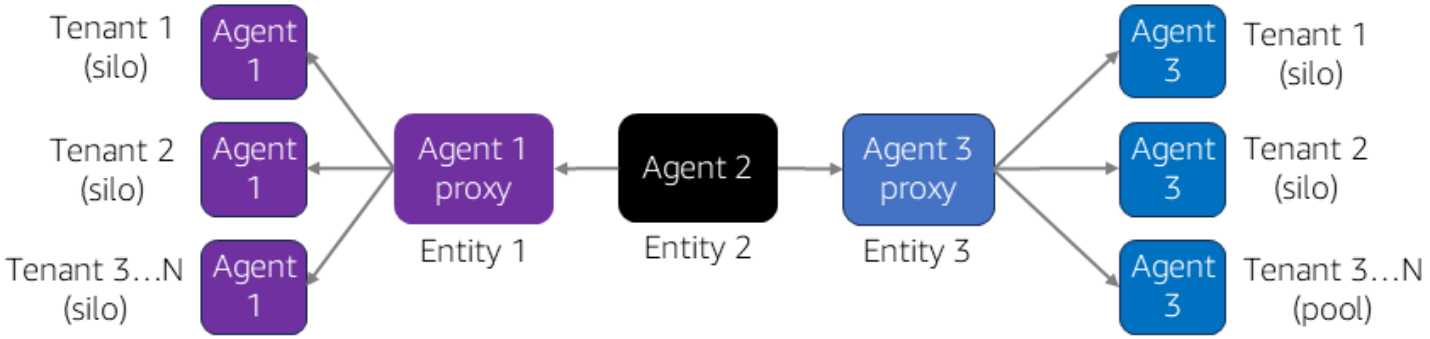
En este diagrama, un proveedor de agentes tiene tres agentes que se implementan mediante el agente como servicio (AAAs). Son compatibles con dos tipos de inquilinos. En el lado izquierdo, dos inquilinos tienen requisitos de cumplimiento y rendimiento que deben cumplir mediante un modelo de silo completo. El inquilino restante, a la derecha, utiliza un modelo agrupado en el que los inquilinos comparten los recursos.

Si el objetivo es la agilidad y la eficiencia operativa, intente limitar los efectos asociados a la compatibilidad con los modelos de implementación por inquilino. Esto significa implementar mecanismos de enrutamiento y otros mecanismos de experiencia que permitan administrar, operar e implementar los agentes desde un solo panel de control.

Si crea un agente en un entorno con poco código o sin él, no habrá ningún concepto de agentes aislados o agrupados. En su lugar, los agentes pueden ser gestionados completamente por otro agente. Los modelos aislados y agrupados se aplican más a los entornos en los que una organización controla la estructura y la presencia del agente. En este caso, los equipos deben considerar qué modelo de implementación respaldar.

A primera vista, estos modelos de despliegue no afectan directamente al funcionamiento de un agente en un sistema más amplio. Es posible que un agente no conozca directamente a otros

agentes que están desplegados en un modelo silo o agrupado. En cambio, estas estrategias de implementación se pueden implementar como parte de una estructura de enrutamiento dentro de un entorno. El siguiente diagrama muestra un ejemplo de cómo se pueden implementar los modelos agrupados y en silos mediante una estrategia de enrutamiento.



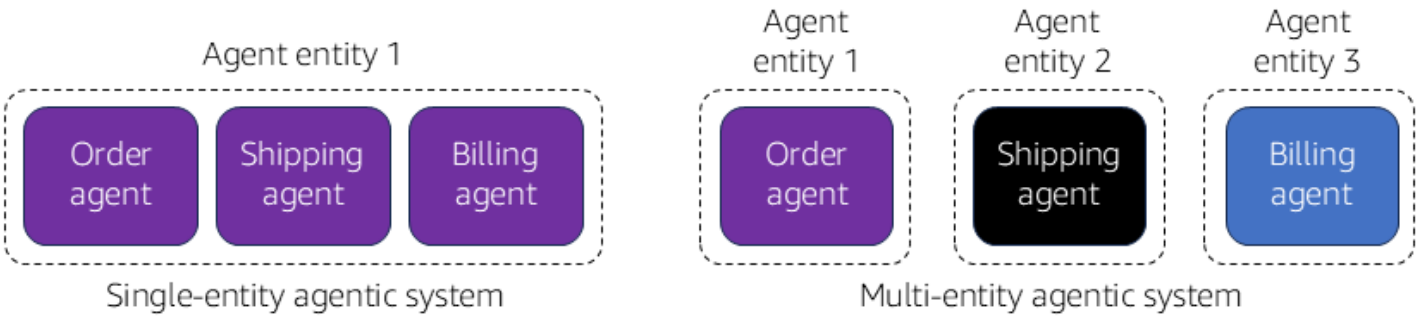
Este ejemplo incluye tres agentes de tres proveedores diferentes. Cada proveedor de agentes tiene la opción de implementar su propia estrategia de despliegue. Por ejemplo, el agente 1 usa un proxy para distribuir las solicitudes entrantes a un conjunto de agentes arrendatarios aislados. El agente 2 no requiere enrutamiento y admite todas las solicitudes de los inquilinos a través de un agente agrupado. El Agente 3 es una implementación de modelo híbrido en la que algunos inquilinos están agrupados en silos y otros agrupados.

La decisión de admitir estos modelos de implementación y la forma en que lo haga depende de la naturaleza de la solución. Es posible que no necesite admitir ninguno de los dos modelos. Sin embargo, es posible que haya casos en los que deba considerar la posibilidad de respaldar esta estrategia, por ejemplo, con el cumplimiento, la proximidad ruidosa, el rendimiento o la estratificación.

Introducción y aplicación del contexto del inquilino

Si creamos agentes que apoyen la contratación múltiple, debemos empezar por considerar cómo configurar el contexto del inquilino, que se utilizará para aplicar políticas, estrategias y mecanismos específicos para cada inquilino dentro de la implementación del agente.

En el nivel más básico, puede introducir el contexto del inquilino en los agentes mediante las herramientas y los mecanismos habituales que utilizamos en las arquitecturas multiusuario clásicas. Esto podría realizarse mediante una clave de API o mediante otros mecanismos de validación. OAuth Muchos ejemplos de esto se centran en convertir un sistema o usuario autenticado en una clave de token web JSON (JWT) que contiene el contexto del inquilino. A continuación, el JWT se propaga por el sistema. Esto se vuelve más interesante cuando consideramos cómo componer los sistemas de agentes. El siguiente diagrama muestra un ejemplo de dos tipos de entornos de agentes.



En este diagrama, el modelo de la izquierda representa un sistema de agencias en el que todos los agentes son propiedad de una sola entidad, están gestionados y alojados por ella. Cuando tenga el control total de toda la experiencia, podrá utilizar estrategias típicas para transferir los inquilinos a cada agente.

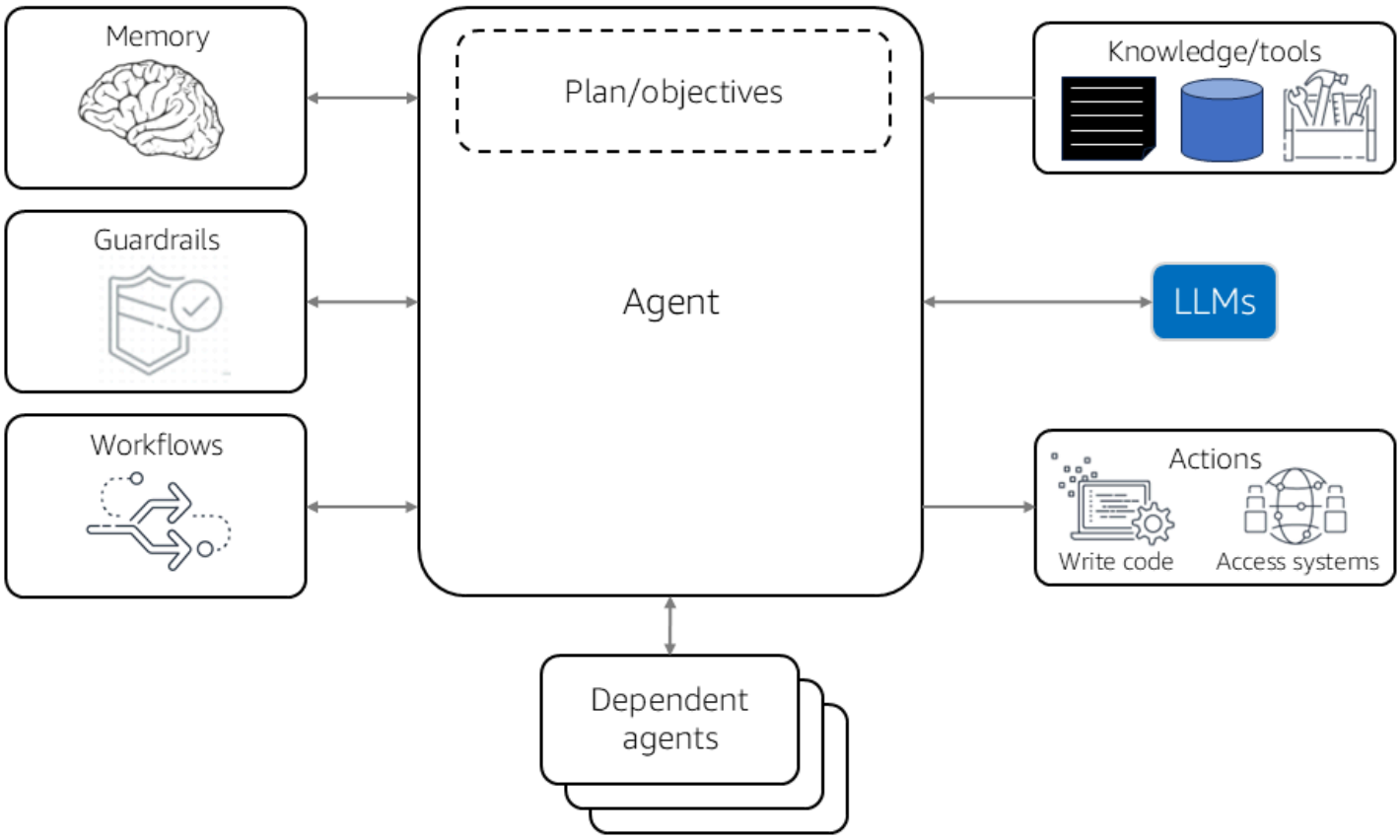
El modelo del lado derecho, que puede ser más común, representa un sistema de agentes que abarca varias entidades. Los agentes se crean, administran y operan de forma independiente, por lo que cada uno tiene sus propios esquemas de autenticación y autorización. En este caso, el desafío es que necesitamos una forma universal de resolver y compartir el contexto de los inquilinos entre estos agentes. Esto se basa en un modelo más distribuido en el que cada agente debe poder autenticar los sistemas o los usuarios y entregarlos a un inquilino de acuerdo con los mecanismos aplicados.

Crear agentes conscientes de los inquilinos

La multitenencia influye en la forma en que implementamos a los agentes individuales. A medida que un agente procesa las solicitudes, considere cómo el contexto del inquilino afecta a la forma en que un agente accede a los datos, toma decisiones e invoca acciones. Para entender mejor cómo y dónde afecta la multitenencia al perfil de su agente, primero determine cómo los constructos pueden formar parte de cualquier agente.

El desafío es que el alcance, la naturaleza y el diseño de los agentes no son nada concretos, ya que los proveedores toman sus propias decisiones sobre el diseño de la experiencia de los agentes. En última instancia, el objetivo de un agente es que es un servicio de aprendizaje autónomo que puede acceder a una variedad de herramientas, fuentes de datos y memoria para determinar la mejor manera de resolver una tarea.

Es menos importante saber exactamente qué estrategias y patrones utiliza un agente. En un modelo multiusuario, es más importante identificar cómo se configuran, se accede y se aplican las distintas partes de un agente. Pensemos en un posible entorno de agentes que dependa de una serie de recursos y mecanismos para alcanzar sus objetivos. En el siguiente diagrama se muestra un ejemplo de un agente de este tipo.

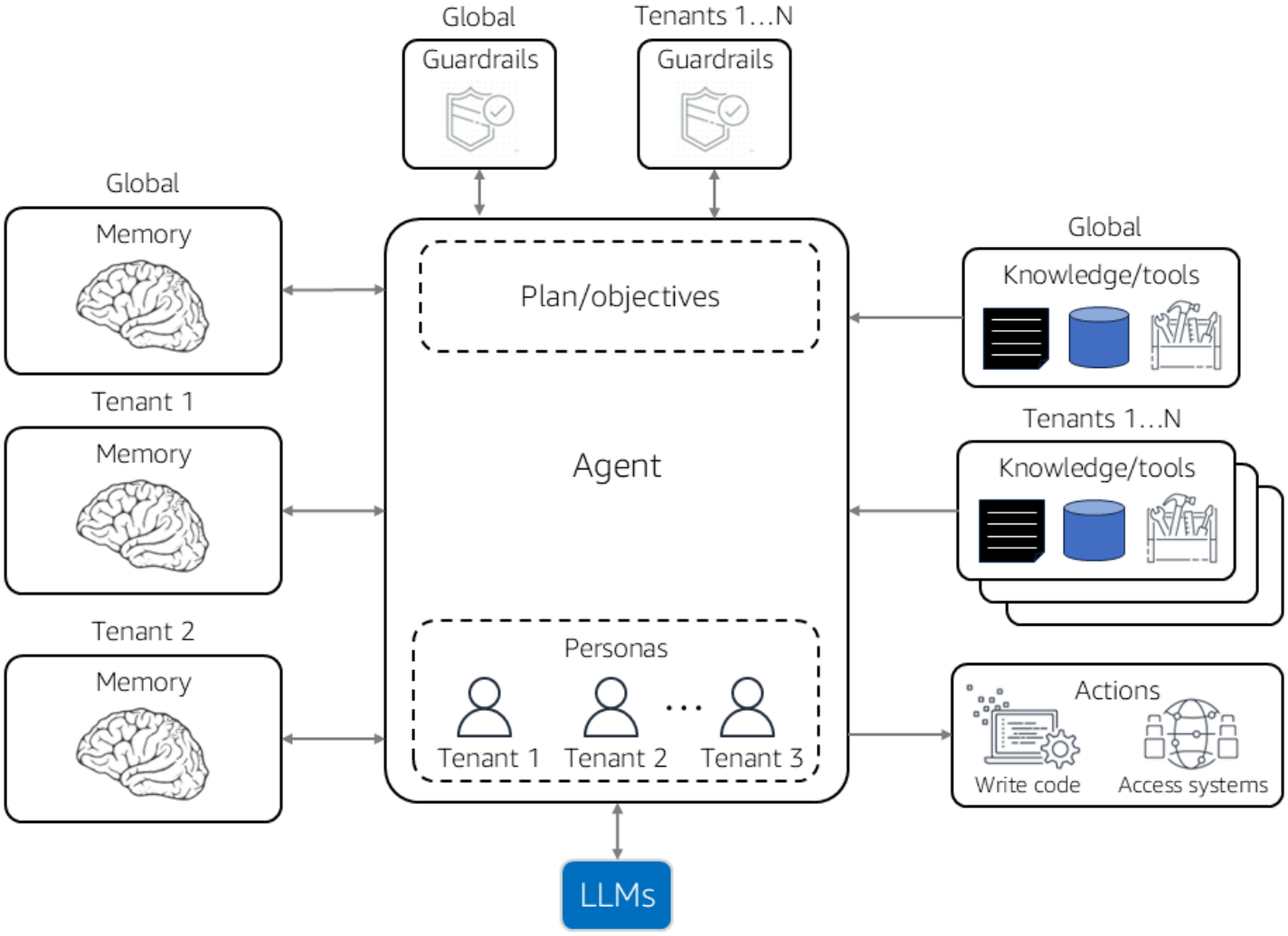


Este diagrama representa una amplia gama de posibilidades de la agencia y muestra varias herramientas y mecanismos que podrían combinarse para lograr un objetivo. En la parte izquierda del diagrama, observe cómo un agente depende de la memoria como parte de su contexto, de las barreras para definir las políticas que guían sus actividades y de los flujos de trabajo que se dirigen a tareas específicas. Algunos podrían argumentar que los flujos de trabajo no deberían incluirse en este contexto, pero puede haber situaciones en las que los flujos de trabajo sean parte integral de la experiencia de un agente.

La parte derecha del diagrama muestra cómo las aportaciones, como el conocimiento y las herramientas, pueden proporcionar información y contexto adicionales que mejoran las capacidades del agente. Luego, el agente genera acciones, como escribir código o acceder a los sistemas. En la parte inferior del diagrama se muestra cómo los agentes dependen de uno o más agentes internos o de terceros que pueden organizarse como parte de un sistema más amplio.

Ahora podemos pensar en lo que significa introducir la multitenencia. La tenencia nos obliga a considerar cómo y dónde un agente introduce las estrategias y los mecanismos que dictan los comportamientos y las acciones. Esto añade otra dimensión a la forma en que pensamos sobre los agentes en términos de sus conocimientos, aprendizaje, herramientas y memoria.

Consideremos ahora cómo modificar este modelo para que sea compatible con la multitenencia. El siguiente diagrama muestra un ejemplo de un modelo multiagente.



En este diagrama, presentamos los personajes de los inquilinos que pretenden moldear la forma en que un agente integra el contexto del inquilino. Por ejemplo, en la parte izquierda del diagrama, la memoria del agente está alterada para admitir la memoria específica del inquilino. Lo mismo ocurre en la parte derecha del diagrama, donde el agente apoya los conocimientos y las herramientas específicos del inquilino. El mismo soporte se aplica también a las barandillas.

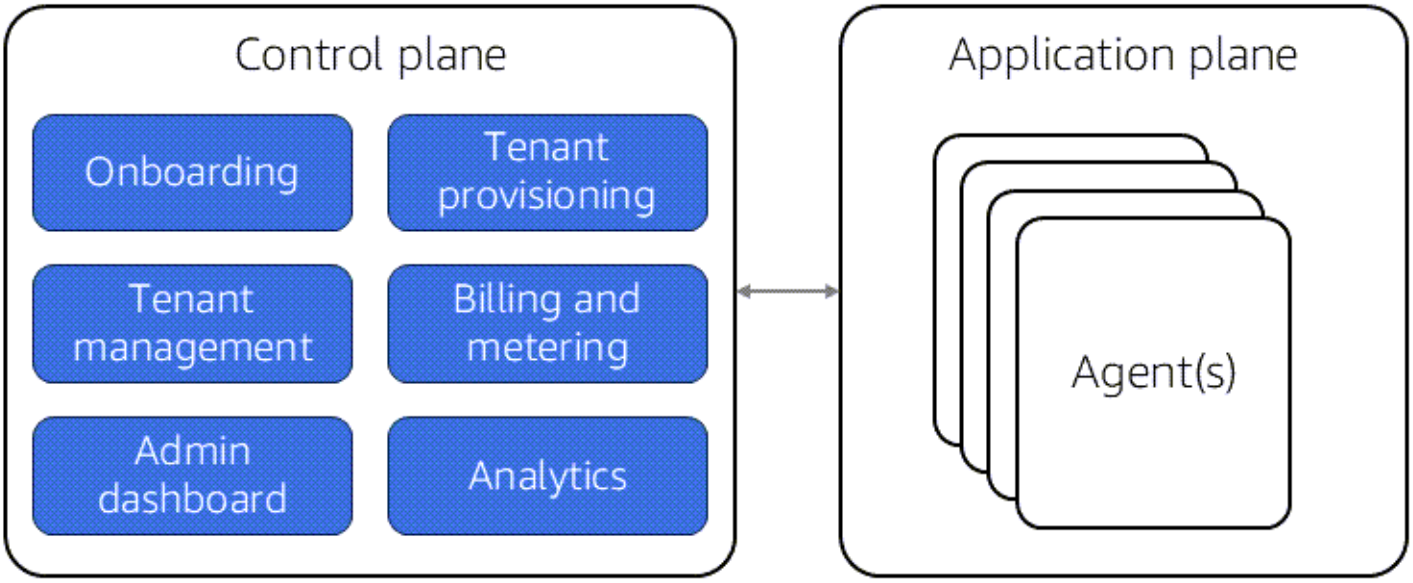
Este puede ser un ejemplo extremo, ya que no todos los aspectos de un agente multiusuario requieren recursos por arrendatario. El punto es que debes considerar cómo adaptar tu agente a inquilinos específicos puede mejorar su eficacia. Este enfoque le permite a su agente aumentar su impacto y valor, proporcionar un contexto más relevante en sus respuestas y desarrollar capacidades especializadas. De este modo, el agente podrá aprender, adaptarse y realizar tareas que se adapten exclusivamente a diferentes personas.

La idea principal es que el contexto del inquilino afecta directamente a la forma en que se crean los agentes. También puede moldear las interacciones de los inquilinos con entidades externas, incluidos otros agentes. La creación de un agente con múltiples inquilinos presenta desafíos tradicionales, como los vecinos ruidosos, el aislamiento de los inquilinos, la organización por niveles, las limitaciones y la administración de costos. El diseño y la arquitectura de su agente deben abordar estos conceptos fundamentales de múltiples inquilinos, que analizaremos en la siguiente sección.

Empleo de planos de control en entornos de agentes

Las mejores prácticas para varios usuarios suelen dividir las implementaciones en dos partes distintas: un plano de control y un plano de aplicaciones. El plano de control proporciona un único panel de control para acceder a los mecanismos operativos, de administración y de organización que abarcan a los usuarios del entorno. El plano de la aplicación es donde residen la lógica empresarial, las características y las capacidades funcionales.

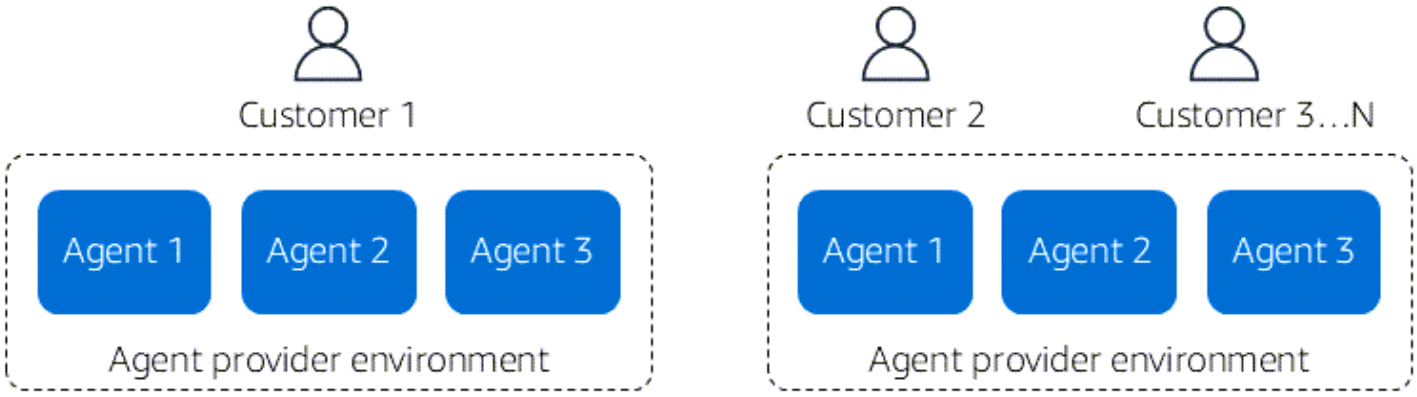
Esta división de responsabilidades también se aplica a los modelos de agencia. Un agente multiusuario requiere cierto grado de administración, operación y conocimientos centralizados, y tiene sentido abordar estas necesidades de forma continua a través de un plano de control. El siguiente diagrama muestra una vista conceptual de cómo se dividen estos planos en un entorno de agente como servicio (AAAs).



Este diagrama muestra la separación tradicional de los planos de control y de aplicación. La novedad es que el plano de control ahora administra los agentes que componen un entorno AAA. El plano de control interactúa con todos los agentes porque suponemos que los crea, administra e implementa un solo proveedor.

Este modelo introduce niveles adicionales de complejidad, especialmente en lo que respecta al ciclo de vida de los agentes y la coordinación con terceros, pero mantiene la separación fundamental de las preocupaciones. El plano de control sigue proporcionando las mismas funciones básicas, ya que organiza la configuración de los agentes, permite observar a los inquilinos y los agentes, recopila datos de consumo y medición para la facturación y gestiona las políticas de los inquilinos.

Este escenario se vuelve más complejo si se considera un sistema multiagente que incorpora agentes de varios proveedores. En el siguiente diagrama se muestra un ejemplo de un modelo de este tipo.



Este diagrama muestra cuatro agentes de diferentes proveedores que forman parte de un sistema multiagente. Los proveedores externos siguen operando e implementando cada agente, que está configurado para permitir el acceso autorizado desde uno o más proveedores. Sin embargo, los agentes permanecen bajo el control del proveedor, por lo que cada agente mantiene su propio plano de control.

Básicamente, estos agentes multiusuario se comportan como servicios de terceros que se integran con otros agentes. Por lo tanto, deben tener su propio plano de control para proporcionar la operación, la configuración y la administración centralizadas de las capacidades de un agente.

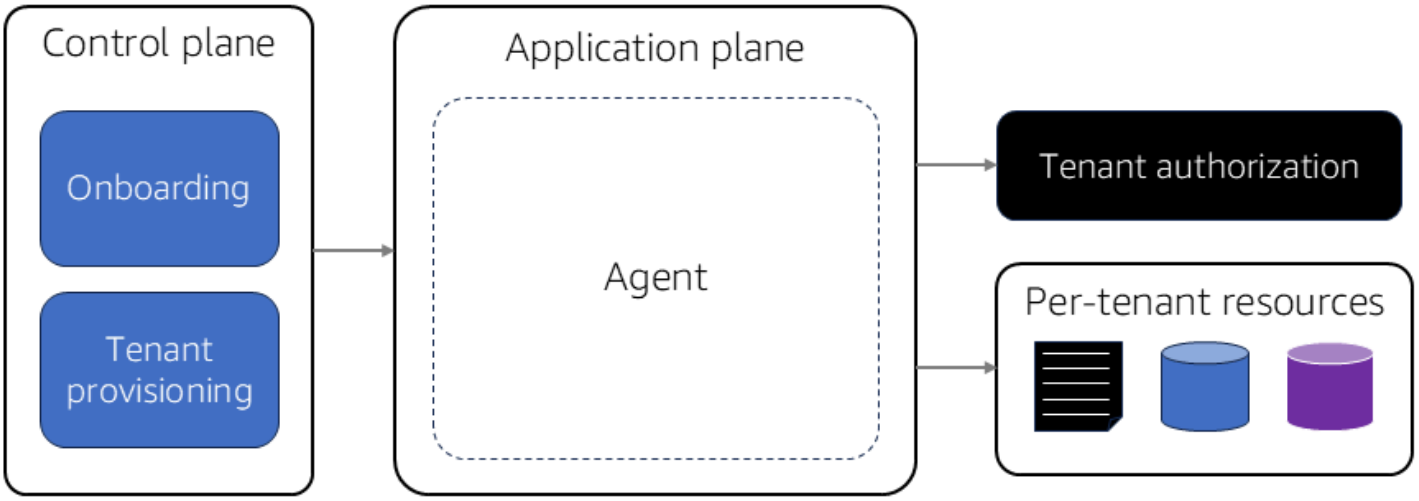
Suponemos que los agentes son servicios independientes que se ejecutan en una experiencia alojada por el proveedor. Sin embargo, esto puede no resultar claro en un escenario en el que un agente consumidor impone más restricciones sobre cómo y dónde alojar a un agente.

Incorporación de inquilinos a agentes

La incorporación suele ser una parte vital de cualquier entorno de AAA. La forma de crear, configurar y aprovisionar los inquilinos suele implicar muchas partes móviles, integraciones y herramientas. La experiencia de incorporación de agentes puede requerir los mismos servicios que se encuentran en un plano de control de la AAA, que incluyen la identidad de los inquilinos, la organización por niveles, el aprovisionamiento de los recursos por inquilino y la configuración de las políticas de los inquilinos.

Su enfoque de la incorporación de agentes depende del tamaño y el modelo de arrendamiento del entorno de su agencia. Los agentes agrupados o aislados tienen sus propios matices, y la elección de utilizar un solo agente o varios agentes también afecta al proceso de incorporación. El siguiente

diagrama muestra una visión conceptual de cómo la incorporación afecta a la configuración de un agente.



Cada vez que incorpore a un agente, el plano de control debe tomar las medidas necesarias para que el inquilino pueda acceder al agente. La forma de presentar a los inquilinos varía según el modelo de autorización del agente, pero suponga que va a crear una identidad de inquilino que asocie las solicitudes de los agentes a las de los inquilinos individuales. Este contexto de inquilino dicta la experiencia del agente al aplicarla a las rutas, los ámbitos y el control de acceso.

La incorporación también puede requerir que configure los recursos por inquilino que utilice un agente. Es aquí donde el servicio de aprovisionamiento de inquilinos del plano de control conecta al agente con los datos y recursos específicos del inquilino que el agente consulta.

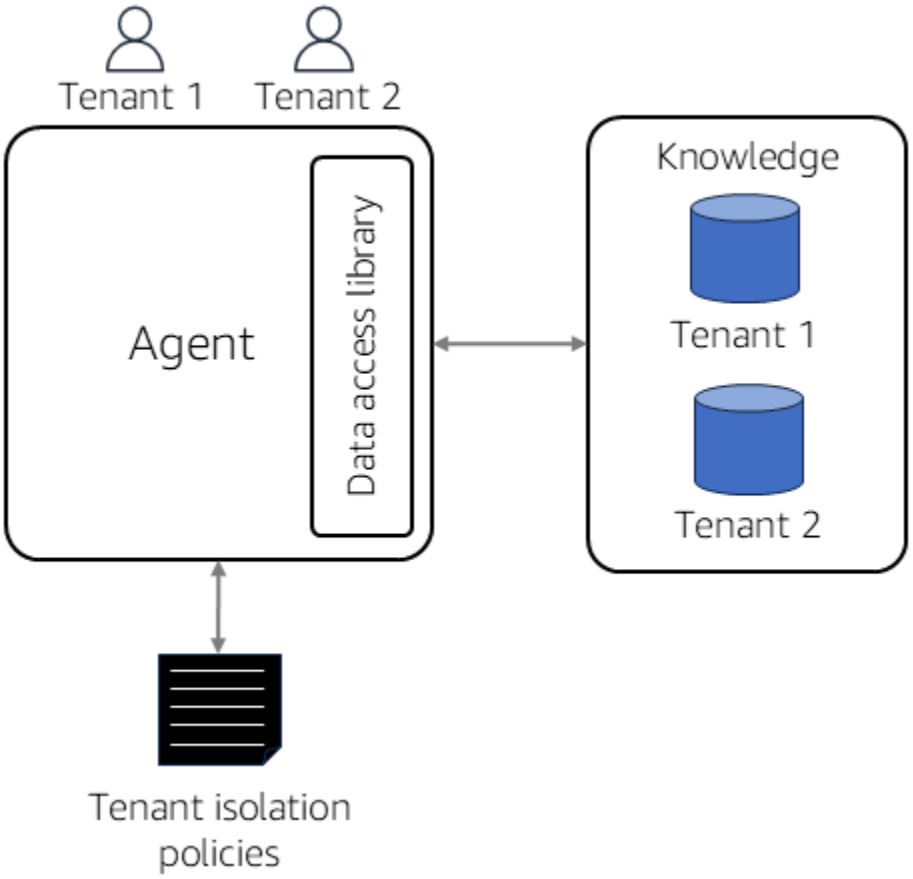
Si su sistema se basa en la integración de agentes externos, también debe abordar las necesidades de esos agentes durante el proceso de incorporación. El funcionamiento de esto depende de los mecanismos de seguridad e integración para autorizar el acceso entre los agentes. Lo ideal es que los pasos necesarios para organizar y configurar la agent-to-agent autenticación y la autorización se aborden mediante la incorporación automática.

Hacer cumplir el aislamiento de los inquilinos

El aislamiento de inquilinos es un concepto que se aplica a todos los entornos con varios inquilinos. Significa que sus políticas y estrategias garantizan que un inquilino no pueda acceder a los recursos de otros inquilinos. En el caso de los agentes con varios inquilinos, es posible que tengas que introducir conceptos y mecanismos que ayuden a hacer cumplir los requisitos de aislamiento de los inquilinos y a los agentes.

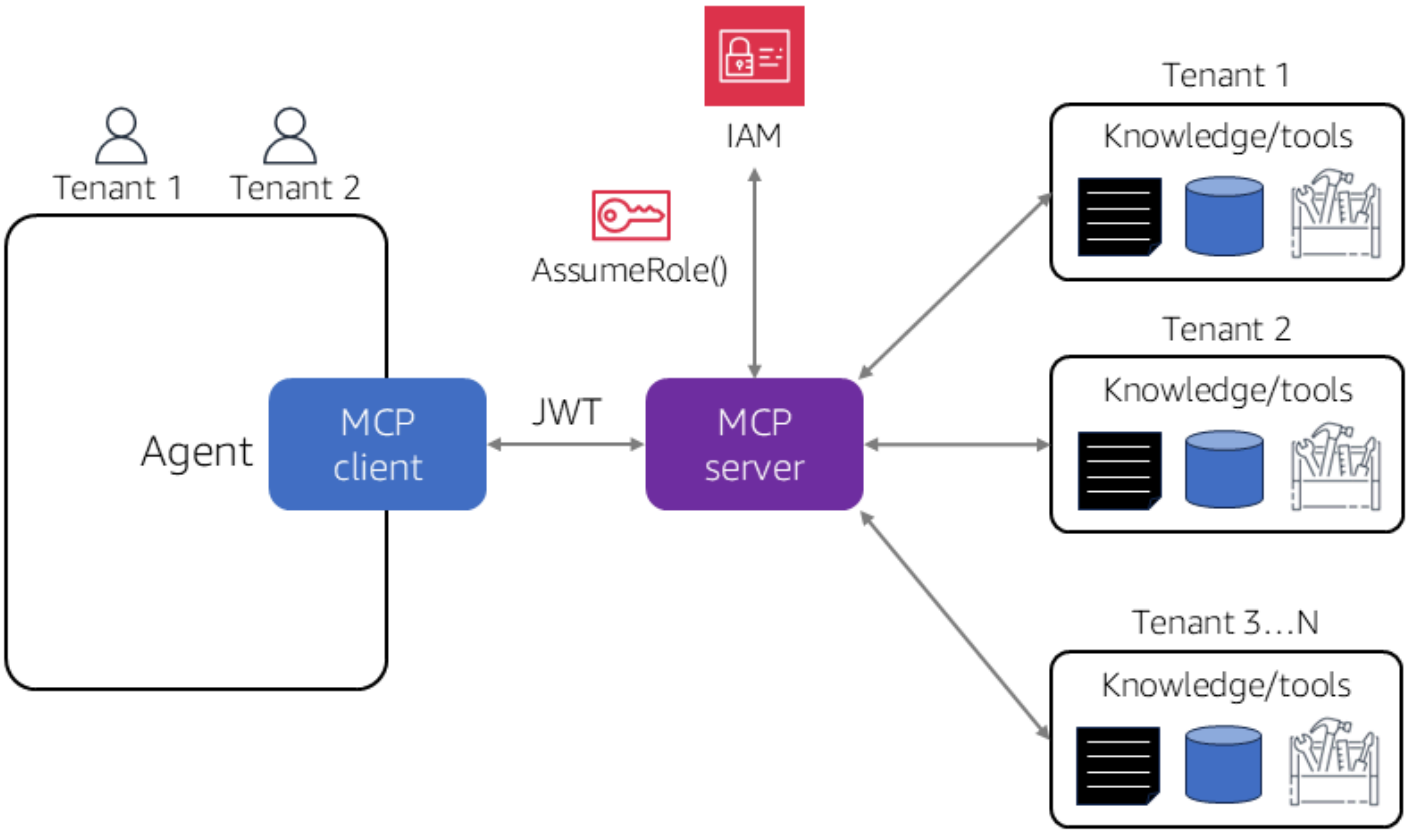
Aplicar el aislamiento de inquilinos es similar a otras estrategias que utilizan los sistemas multiarrendatarios tradicionales. Por lo general, cuando diseñe una arquitectura AAA, identifique cualquier área del sistema en la que una solicitud o acción pueda acceder a los recursos para determinar si la solicitud sobrepasa los límites de algún inquilino. Por ejemplo, los microservicios pueden depender de tablas de Amazon DynamoDB dedicadas por inquilino. Esto requiere que introduzca políticas que garanticen que otro inquilino no pueda acceder a la tabla de un inquilino.

En este caso, considere el aislamiento de los inquilinos desde la perspectiva de un agente y sus interacciones con cualquiera de sus recursos por inquilino. El siguiente diagrama muestra un ejemplo conceptual de cómo los agentes aplican las políticas de aislamiento de inquilinos para controlar el acceso a los recursos de los inquilinos.



En la parte derecha de este diagrama, el agente tiene información sobre cada inquilino almacenada en bases de datos vectoriales independientes. A medida que el agente procesa una solicitud, examina el contexto del inquilino que la presenta. En base a esto, el agente aplica una política de aislamiento adecuada para garantizar que los inquilinos no puedan acceder a los datos o recursos fuera de los límites designados.

Si su agente utiliza un protocolo de contexto modelo (MCP), también puede implementar su modelo de aislamiento de inquilinos. El siguiente diagrama muestra un ejemplo de cómo introducir el MCP y aplicar políticas de aislamiento.



El MCP es un protocolo estandarizado que un agente utiliza para integrarse con cualquier herramienta, dato y recurso. En este ejemplo, un cliente MCP y un servidor MCP interactúan con los conocimientos y las herramientas específicos del inquilino que se muestran en la parte derecha del diagrama. El contexto del inquilino fluye del cliente al servidor, y el servidor utiliza este contexto para adquirir las credenciales del servicio (IAM) relacionadas con el inquilino. AWS Identity and Access Management Las credenciales controlan el acceso a los recursos de cada inquilino, lo que garantiza que un inquilino pueda acceder a los recursos de otro inquilino.

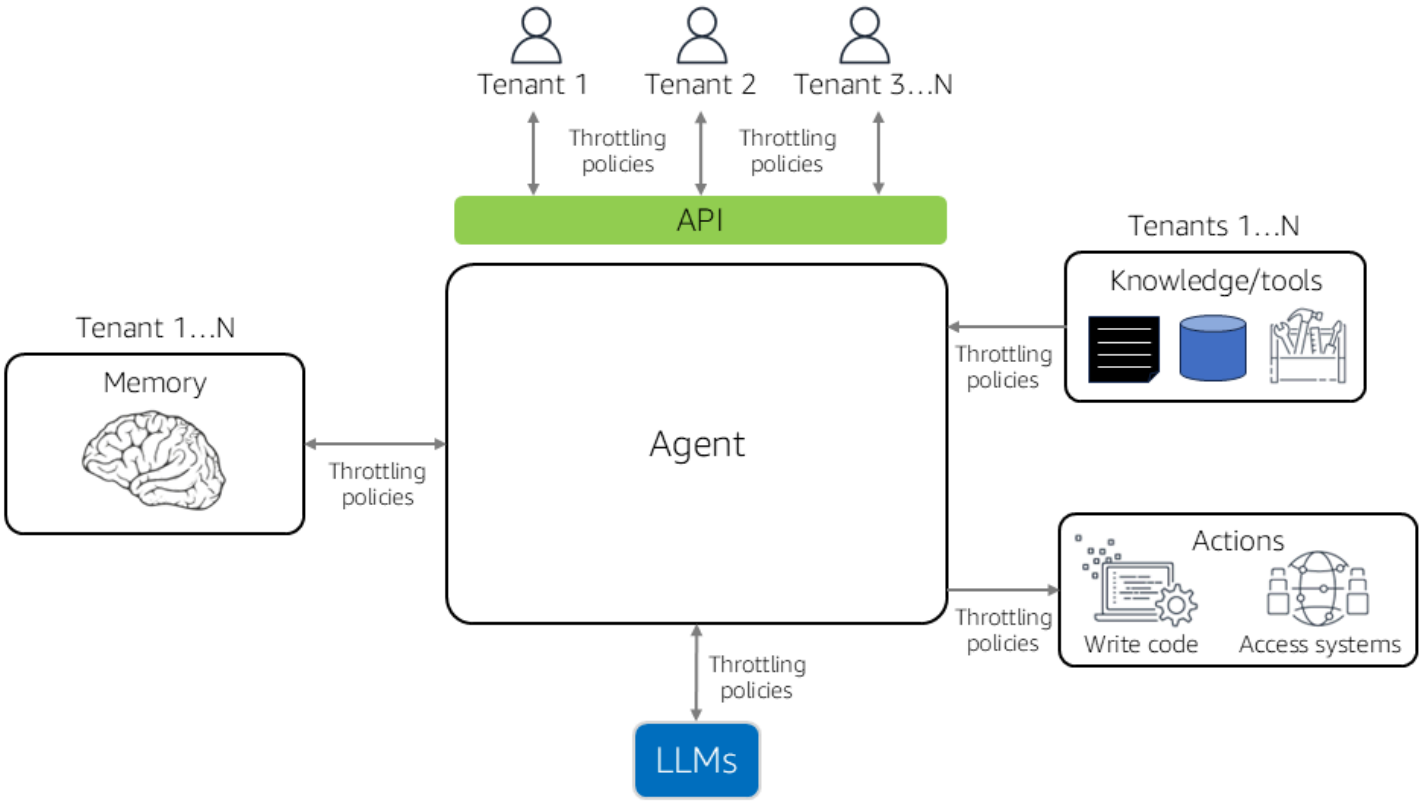
A medida que los agentes incorporan la opción de arrendamiento múltiple, deben introducir mecanismos que apliquen políticas de aislamiento de inquilinos a medida que procesan las solicitudes. En algunos casos, la IAM puede ayudar a limitar el acceso a los recursos de los inquilinos. En otros casos, es posible que deba introducir otras herramientas o marcos para aplicar las políticas de aislamiento de los inquilinos.

Vecinos y agentes ruidosos

En un entorno AAA con varios inquilinos en el que varios inquilinos comparten un agente, piense dónde y cómo introducir políticas que eviten que los vecinos sean ruidosos. Las políticas pueden

introducir restricciones de uso general que se apliquen a todos los tipos de consumo, o bien pueden tener políticas basadas en inquilinos o niveles que impongan restricciones en función de una persona determinada. Es posible que impongas mayores restricciones de consumo a los inquilinos del nivel básico que a los inquilinos del nivel premium.

Esta noción de limitación se puede aplicar en varios puntos de la arquitectura. El siguiente diagrama muestra un ejemplo de algunas áreas en las que es posible introducir políticas de vecindad ruidosa.



En nuestra revisión previa de la implementación de varios agentes, examinamos los diferentes recursos que su agente puede utilizar y destacamos el potencial de recursos por inquilino dentro de un agente. Cada punto de contacto es un área potencial para introducir políticas de limitación, lo que ayuda a garantizar que los inquilinos no superen los límites de consumo de su sistema o las políticas de estratificación de un inquilino.

Los mejores lugares para introducir protecciones para los vecinos ruidosos son los puntos de la arquitectura en los que los inquilinos comparten recursos. Estos componentes compartidos o agrupados, como el procesamiento, la memoria y los modelos de lenguaje de gran tamaño, son los más susceptibles a la degradación del rendimiento si un solo inquilino consume de forma desproporcionada. APIs

Un lugar natural para aplicar la regulación es en el punto de entrada del agente, a veces denominado «borde exterior». Aquí puedes introducir límites globales o de tenant-tier-based tarifas antes de que el agente comience a procesar la solicitud. La limitación también se puede aplicar en una parte más profunda de la ruta de ejecución, por ejemplo, cuando el agente llama a un LLM, accede a la memoria o invoca herramientas compartidas.

Estas políticas lo ayudan a hacer cumplir el uso justo, a mantener la resiliencia de los agentes ante la presión y a preservar una experiencia coherente entre los inquilinos. En función de tus objetivos, puedes centrarte en la protección general del sistema (resiliencia) o en gestionar de forma pormenorizada la experiencia de los inquilinos (por ejemplo, con prestaciones por niveles).

Datos, operaciones y pruebas

Los agentes y la propiedad de los datos

Al revisar la implementación de los agentes, se destacan los escenarios en los que un agente se basa en los datos de un inquilino determinado. En este caso, considere el ciclo de vida de los datos y, lo que es más importante, dónde se almacenan. Esto es especialmente importante para los sectores y los casos de uso en los que la naturaleza de los datos influye en la forma en que un agente accede a ellos.

Los proveedores de AAA deben evaluar cómo resolver los problemas de datos en un entorno multiusuario, que pueden afectar a la incorporación, el aislamiento y las operaciones de un agente. Los matices y las estrategias aplicables varían según las herramientas, las tecnologías y los datos que se consuman. Puede abordarlo de muchas maneras, algo que debe tener en cuenta al crear cualquier oferta de AAA.

Operaciones con agentes multiarrendatarios

A medida que cree entornos de agentes, piense en cómo operar y administrar sus agentes. Como proveedor, necesita métricas, datos, información y registros que le permitan monitorear el estado, la escala y la actividad de un agente. Esto es más pronunciado en un entorno de agencia con varios inquilinos, en el que querrá entender cómo los inquilinos individuales consumen los recursos de los agentes.

Esto es aún más importante en entornos con varios agentes cuando se necesita información sobre las interacciones entre los agentes. Ser capaz de perfilar las actividades entre los agentes y realizar un seguimiento de ellas puede ser esencial para solucionar los problemas que afectan a la escala, la precisión y la eficacia del sistema.

Los equipos de operaciones también pueden elaborar perfiles de las interacciones de la LLM para hacerse una mejor idea de las cargas que soportan los LLMs agentes. Estos datos son esenciales para refinar la implementación de los agentes. También puede ofrecer a los equipos operativos una visión de cómo los agentes y el arrendamiento afectan al perfil de costos general de un sistema.

Capacitación y pruebas de agentes con múltiples inquilinos

Uno de los desafíos asociados a los agentes inmobiliarios es que se espera que aprendan y evolucionen. También significa que debemos probar nuestro agente, perfeccionarlo y mejorar su precisión antes de ponerlo en producción. Hay muchas áreas en las que puede inspeccionar y evaluar si su agente está evaluando y categorizando correctamente la intención o si está eligiendo e invocando las herramientas y acciones adecuadas. La lista de variables es extensa, pero se trata, en última instancia, de garantizar que su agente encuentre los resultados que le permitan alcanzar sus objetivos.

El examen de todas las partes móviles y los principios relacionados con las pruebas de los agentes va más allá del alcance de este documento, pero tenga en cuenta que las estrategias de prueba añaden complejidad a los entornos de AAA con varios usuarios. Por ejemplo, si un agente tiene datos, memoria y otras estructuras que se aplican contextualmente a cada inquilino, los resultados del agente pueden determinarse en función de los recursos de cada inquilino.

Si utiliza un agente para simular un escenario, es posible que necesite ampliar la simulación para adaptarla a casos de uso específicos de cada inquilino. En consecuencia, debe refinar los procedimientos de validación para tener en cuenta los casos en los que los criterios de validación difieran para cada inquilino.

Consideraciones y discusión

¿Dónde encaja el SaaS?

Los expertos de la industria debaten activamente sobre cómo los agentes influyen en el panorama del software como servicio (SaaS). Si bien es cierto que los agentes están cambiando el software de muchos sistemas, es exagerado sugerir que los agentes hacen que los modelos de entrega queden obsoletos. Es probable que algunos proveedores de SaaS se vean perturbados por la adopción de agentes, y otros podrían replantearse por completo su propuesta de valor al optar por un modelo de agente como servicio (AAAs). Otros pueden lograr un equilibrio mediante la introducción selectiva de agentes para abordar necesidades específicas.

Este tema es interesante porque adoptar los mejores principios de SaaS puede representar la próxima evolución del SaaS. Esto podría significar que el SaaS está avanzando, o puede significar que los principios fundamentales del SaaS se están empaquetando y realizando en un modelo basado en agentes. Probablemente sea menos importante decidir dónde termina la terminología, pero parece poco probable que el SaaS como concepto desaparezca. Es más probable que los agentes den forma a la presencia de SaaS.

En última instancia, debemos decidir qué estrategias se pueden aplicar a la tecnología AAA, es decir, permitir que las organizaciones adopten arquitecturas y estrategias empresariales basadas en los agentes para que los proveedores puedan maximizar la eficiencia, el valor y el impacto de sus sistemas de agentes. Los agentes no son cajas negras. Los agentes consumen recursos, escalan las operaciones, dependen de los datos y generan costos, todos factores que los proveedores deben abordar. Los proveedores de agentes deben evaluar cómo los principios de múltiples inquilinos pueden moldear las ofertas de servicios y optimizar los modelos operativos.

Explicación

El panorama de las agencias sigue evolucionando y los diseños varían según los dominios, los casos de uso previstos y los sectores objetivo. Parte de esta evolución incluye perfeccionar aún más nuestra visión de las estrategias, los patrones y las ventajas y desventajas que los arquitectos tienen en cuenta al diseñar y construir agentes.

Una estrategia integral de agentes debe alinearse con los objetivos comerciales y técnicos. Esto incluye definir los mercados y las personas objetivo, establecer estrategias de gestión de

precios y recursos y determinar cómo se adaptan los agentes a los sistemas más grandes. Estas consideraciones son particularmente importantes a la hora de ofrecer servicios AAA, en los que la escala, la rentabilidad y la innovación son los objetivos principales.

Las capacidades operativas son igualmente importantes. El entorno debe permitir la supervisión de la actividad de los agentes, las métricas de estado y los patrones de uso. Esto se vuelve más complejo en los sistemas con varios agentes, donde las operaciones deben coordinarse entre agentes independientes.

En general, este análisis de los agentes solo es una muestra superficial de las diversas consideraciones arquitectónicas que podrían formar parte de los sistemas de agentes. Más allá de seleccionar las herramientas, los marcos y los marcos adecuados LLMs, el éxito depende de la creación de una arquitectura que cumpla con los requisitos empresariales en materia de escalabilidad, eficiencia, implementación y tenencia múltiple.

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	—	14 de julio de 2025

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

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 unos 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 CSPM 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 Operational Readiness Reviews (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 del 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.

Region

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 Consola de administración de AWS 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 el 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.