ORACLE

7 preguntas sobre infraestructura de datos para el éxito de la IA

Al abordar estos puntos clave, los líderes pueden simplificar la adopción de la IA, maximizar el retorno de inversión y potenciar la innovación empresarial

¿Están tus datos listos para la IA?

Por Jeffrey Erickson
Redactor sénior

Imagina los mismos *large language models*, o LLM, que pueden explicar la física cuántica y resumir novelas francesas, adquiriendo un profundo conocimiento sobre los datos operativos y la base de conocimientos únicos de tu empresa. Ahora, la recuperación, combinación y contextualización de esos datos, pasa de ser un pendiente para un analista a una conversación en lenguaje natural entre un usuario empresarial y un agente de IA que puede descubrir información valiosa y luego tomar acciones. Se liberan recursos de TI y tu organización se basa más en los datos.

Parece bueno, ¿verdad? Por eso no es de extrañar que las empresas tecnológicas más ambiciosas, talentosas y bien financiadas del mundo, así como sus competidores startups, estén compitiendo por poner las manos en el poder de la IA generativa. Incluso en medio de todo este revuelo, es difícil exagerar hasta qué punto los modelos de IA generativa pueden multiplicar el valor de tus datos al alterar fundamentalmente la forma en que las personas de tu organización acceden a ellos y los utilizan.

La base de todo esto es una infraestructura adaptada a las necesidades de la IA generativa, con la combinación adecuada de herramientas personalizadas, LLM fundamentales, técnicas y sistemas informáticos para proporcionar respuestas rápidas y dar soporte a casos de uso tan variados como la detección de anomalías y la identificación de objetos. El éxito suele depender de disponer de la potencia suficiente para gestionar operaciones complejas de IA, lo que implica disponer de GPU interconectadas por una red de clústeres capaz de alcanzar un rendimiento ultraelevado y una latencia de microsegundos. También se necesitan bases de datos vectoriales, una forma de aprovechar los agentes de IA para la *retrieval-augmented generation* (RAG), que combina las diversas bases de conocimientos de la empresa con los LLM, e interfaces de agentes de IA fáciles de usar.

> Es difícil exagerar hasta qué punto los modelos de IA generativa pueden multiplicar el valor de tus datos

¿Qué es Agentic AI?

La Agentic AI se refiere a la inteligencia artificial capaz de comprender y responder a la información, así como de perseguir objetivos de forma activa.

Características clave de Agentic AI

Comportamiento proactivo
La IA puede iniciar acciones en lugar de limitarse a reaccionar a estímulos externos.

Adaptación
La IA puede aprender de la experiencia, aceptar comentarios y ajustarse para alcanzar mejor sus objetivos.

Comportamiento orientado a objetivos
La IA tiene objetivos específicos que busca alcanzar y puede trazar los pasos necesarios para llegar a ellos.

Autonomía
La IA puede tomar decisiones y actuar de forma independiente, dentro de unos parámetros determinados.

La promesa de la IA generativa es considerable. También lo son las inversiones en infraestructura necesarias para alcanzarla.

Probablemente hayas oído hablar de las impresionantes estadísticas sobre la expansión de los centros de datos para dar soporte a la IA. Gran parte de ese gasto ha sido realizado por proveedores de nube a hiperescala que buscan ofrecer experiencias de IA excepcionales, y algunas organizaciones con visión de futuro también están preparando sus infraestructuras de datos para dar soporte a lo que está por venir.

Entonces, ¿qué puedes hacer para prepararte ahora? A lo largo de este artículo, presentamos siete preguntas que se plantean muchos arquitectos de TI mientras trabajan para cumplir la promesa de la IA ahora y en el futuro.

1 ¿Cómo beneficiarse de una serie de casos de uso de la IA generativa, incluidos algunos en los que aún nadie ha pensado?

Un puñado de proyectos de IA (como el marketing hiperpersonalizado, los chatbots de atención al cliente realmente buenos o los asistentes de programación que aumentan la productividad) suelen justificar el coste de la infraestructura de apoyo. Y la lista de posibles casos de uso crece día a día, porque, francamente, si puedes imaginarlo, probablemente puedes hacerlo. Esto se debe en gran medida a los avanzados LLM que ahora están integrados en aplicaciones empresariales (como ERP, HCM y SCM), y que se utilizan ampliamente en áreas con mucha actividad en investigación, como la robótica, la sanidad, la genómica y la ingeniería aeroespacial. En otras palabras, en casi todas partes.

Lo fundamental es complementar esos LLM para que tu IA se convierta en el principal experto de tu negocio. Al proporcionar acceso a bibliotecas de datos históricos, operativos, financieros y de documentación, permite que la IA generativa sirva como multiplicador de fuerza de muchas maneras para muchas personas en toda tu empresa.

¿Qué significa eso? Puede significar que los ejecutivos «conversen» con los datos, profundicen en ellos y obtengan nueva inteligencia empresarial en el acto. Puede significar que los agentes de IA se conviertan en socios para el desarrollo de software y las sesiones de intercambio de ideas de I+D, incluso en rápidos creadores de prototipos y maquetas. Puede significar que los vendedores sean capaces de gestionar más clientes potenciales porque la prospección, la comunicación y los detalles del proceso adquieren nuevos niveles de automatización, cortesía de la IA.

Estos son solo algunos ejemplos. Una reunión creativa en la que los equipos de producto, finanzas, recursos humanos y jurídico compartan sus ideas probablemente dará lugar a casos de uso creativos de IA generativa y a un punto de partida sólido para un plan de infraestructura de datos.

¿Quieres más ideas de empresas similares? Ve una serie de casos de uso en el mundo real tan variados como el análisis genético y las retransmisiones deportivas.

Ajuste fino y RAG: dos formas de ayudar a la IA a «entender» tus datos

Tanto el ajuste fino como la RAG ayudan a los modelos generativos de IA a ofrecer respuestas más relevantes contextualmente y adaptadas a tu organización. Aquí tienes en qué se diferencian:

Ajuste fino significa tomar un modelo de propósito general, como Command de Cohere o Llama 3 de Meta, y someterlo a rondas adicionales de entrenamiento en un conjunto de datos más pequeño y específico del dominio. El ajuste ayuda a que el modelo funcione mejor en tareas específicas porque se ha adaptado a los matices y la terminología de un ámbito, como la codificación, las finanzas o la sanidad. Las desventajas son el coste y la necesidad de una infraestructura basada en GPU para soportar el proceso.

RAG es una estructura arquitectónica que ayuda a los modelos de IA de propósito general a ofrecer resultados útiles para organizaciones específicas, proporcionando al modelo de lenguaje (LLM) datos relevantes y específicos mientras formula respuestas a las consultas. El resultado es un sistema de IA que combina la fluidez lingüística de un LLM con la información de tus datos internos para ofrecer respuestas específicas y contextualmente más apropiadas. Contrariamente al ajuste fino de los modelos de IA, el RAG funciona sin modificar el modelo subyacente. Además, el LLM «olvida» los datos que se le han proporcionado una vez finalizadas las consultas, lo que alivia una posible fuente de fugas de datos.

¿Debería utilizar RAG o el ajuste fino?
Aprende cómo elegir

2 ¿Cómo hacer que todos los datos, no estructurados y semiestructurados, estén disponibles para la IA?

El procesamiento de datos no estructurados y semiestructurados para su uso por parte de la IA es un proceso de varios pasos que implica la recopilación e ingestión, el almacenamiento y el procesamiento de los datos: limpieza, extracción de características, normalización, segmentación y, posiblemente, otros pasos. Solo entonces se preparan las imágenes, documentos y archivos de audio y vídeo de tu data lake para el ajuste fino, la vectorización y la RAG.

Datos para el ajuste fino: El ajuste fino de los modelos de IA generativa para tareas específicas puede requerir que se le muestren nuevos datos específicos de la disciplina. Por ejemplo, si estás ajustando un modelo de IA generativa para producir informes médicos, necesitarás un conjunto de datos de informes médicos relevantes y de alta calidad para ayudar al modelo a aprender la terminología y el contexto adecuados.

Datos para resultados continuos de IA: Un sistema de gestión de datos preparado para la IA debe proporcionar a sus LLM acceso a una amplia gama de almacenes de datos, como documentos JSON y datos relacionales y semiestructurados. Es necesario que aplique eficazmente incrustaciones vectoriales a los datos, almacene las incrustaciones como vectores en una base de datos y consulte la base de datos para permitir una búsqueda vectorial eficaz y el procesamiento del lenguaje natural. Esto puede requerir una recopilación continua de datos y un proceso de integración que también incluya una arquitectura RAG para ayudar a mejorar la precisión y relevancia de los resultados de su modelo de IA.

¿Cómo sería si estos sistemas estuvieran en funcionamiento? Un equipo de ventas, por ejemplo, podría guardar el audio de cada llamada y dejar que la IA analizara el sentimiento del cliente mientras crea resúmenes. Se acabaron los garabatos indescifrables cuando los vendedores intentan interactuar y tomar notas al mismo tiempo. Se acabaron los detalles olvidados en una solicitud de presupuesto para el próximo pedido del cliente. Es la misma atención al detalle, pero más precisa y eficiente.

Una clave para el éxito: Tener una base de datos unificada y multimodal que te permita habilitar todos estos procesos sin mover, resincronizar ni volver a proteger los datos a través de sistemas especializados. Un modelo de datos unificado de este tipo puede acelerar las operaciones en varios órdenes de magnitud, ya que permite realizar búsquedas en distintos tipos de datos, incluidos los vectores de IA, sin que el departamento de TI tenga que integrar silos dispares. Esto puede dar lugar a resultados más ricos, ya que los modelos de IA tienen en cuenta las sutiles relaciones entre gran cantidad de información aparentemente no relacionada, pero crítica.

3 ¿Cómo proteger datos sensibles para la IA, preferiblemente con una estrategia única de gobernanza y control de acceso?

Mantener la seguridad y la privacidad de los datos proporcionados a tus herramientas de IA generativa es un principio esencial del diseño de la infraestructura, tanto en los modelos de ajuste fino y RAG como durante el uso diario.

En función de tus necesidades, los datos para tu plataforma de IA generativa pueden anonimizarse, cifrarse o enmascararse. Los datos destinados a la generación de resultados de IA también deben controlarse estrictamente teniendo en cuenta la función del solicitante y empleando procesos de identificación multifactor. Algunas organizaciones optan por ejecutar inferencias en sus propias copias de modelos de IA alojados en infraestructuras dedicadas, o incluso en centros de datos locales, para proteger aún más los datos privados.

Puedes buscar ayuda en tus proveedores tecnológicos y organismos de normalización
Por ejemplo, Oracle es una de las más de 280 organizaciones que colaboran con el National Institute of Standards and Technology (NIST) en su Artificial Intelligence Safety Institute Consortium[1]. El objetivo del AISIC es «desarrollar directrices y normas con base científica y respaldo empírico para la medición y la política de la IA, sentando las bases de la seguridad de la IA en todo el mundo». Los miembros trabajarán para elaborar directrices, herramientas, métodos, protocolos y mejores prácticas que ayuden a la comunidad a desarrollar e implantar la IA de forma segura.

También hay otros esfuerzos similares centrados en la IA.

[1] *NIST Artificial Intelligence Safety Institute Consortium*

Arquitectura de la IA generativa: base de datos convergente frente a múltiples bases de datos de uso único

Con una base de datos multimodal, normalmente no se necesita una ETL y un movimiento de datos extensivos para aprovechar la IA generativa. En su lugar, puede dar soporte a diversos proyectos y cargas de trabajo de IA utilizando una única plataforma de datos, lo que reduce significativamente la complejidad y la sobrecarga de gestión a la vez que minimiza el riesgo que implica el movimiento de datos. Una base de datos multimodal no significa que los datos deban estar en un solo formato o modelo; solo significa que puede elegir su arquitectura de datos para cada tipo de aplicación.

4 ¿Cómo los equipos colaboran en la estrategiade IA generativa manteniendo los estándares establecidos?

En las manos adecuadas, un modelo estándar puede convertirse en una potencia de IA personalizada que solo tu organización podría crear. La gente quiere aprovechar la IA. Darles las herramientas que necesitan para lograrlo hace que sea mucho más probable que los flujos de trabajo de IA generativa se conviertan en parte de la estrategia de tu empresa y no en proyectos de TI puntuales en la sombra.

Aprende a crear un centro de excelencia de IA en tu empresa

Una forma de maximizar tus esfuerzos es un centro de excelencia, que proporcione transparencia y promueva la coherencia entre departamentos. Un centro de excelencia consolida las mejores prácticas, herramientas y técnicas relacionadas con la IA generativa y lidera su uso para resolver problemas empresariales.

Una tecnología central para esta colaboración será una plataforma de ciencia de datos que fomente la colaboración para los pasos técnicos de ajuste fino y la implementación de modelos. Tu proveedor de nube a hiperescala dispondrá de una plataforma de IA generativa en la que los científicos de datos puedan trabajar y mejorar los potentes modelos básicos de Cohere, Meta y otros proveedores. Estas plataformas ofrecen a los científicos de datos y a los equipos de TI un lugar para almacenar, intercambiar y potencialmente reutilizar modelos, conjuntos de datos y etiquetas de datos en todos los servicios.

Los hiperescaladores también ofrecen catálogos de modelos y proporcionan LLM ajustados con los datos y la infraestructura informática para ejecutarlos. Ahora las unidades de negocio pueden aprovechar los éxitos anteriores de la IA en otras partes de la organización.

Con un centro de excelencia y una plataforma líder de ciencia de datos, puedes fomentar una cultura de aprendizaje y mejora continua. Proveedores como Oracle ofrecen servicios totalmente gestionados, como Oracle Cloud Infrastructure (OCI) Generative AI, para integrar a la perfección los LLM en una amplia gama de casos de uso. También puede crear modelos personalizados ajustando los modelos base con tu propio conjunto de datos.

5 ¿Cómo combinar los puntos fuertes de los proveedores de nube para maximizar la disponibilidad de datos para la IA?

Tal vez tu organización esté explorando Google Gemini o Copilot de Microsoft. O podría decidir alojar modelos de código abierto en AWS o modelos de base de Cohere en OCI. Ahora, los nuevos acuerdos entre Oracle y los demás hiperescaladores abren un amplio mundo de posibilidades para las empresas cuyas estrategias de gobernanza de datos corporativos se basan en la tecnología de Oracle Database, en las instalaciones o en la nube. Puede adoptar cualquiera de estos modelos con baja latencia y sobrecarga de gestión, gracias a las innovadoras relaciones multicloud entre Oracle y Azure, Google Cloud y AWS que permiten utilizar los servicios de Oracle Database no solo en OCI, sino dentro de los centros de datos de cada proveedor.

El éxito con esos casos de uso de IA generativa aún no concebidos puede ser más probable cuando se pueden utilizar exactamente los modelos de IA que se deseen, ejecutándose donde se desee, sin comprometer el acceso fácil y seguro a los datos. Los equipos pueden aprovechar los mejores servicios para tareas específicas al tiempo que mantienen la seguridad y resistencia, y reducen los costes.

6 ¿Cómo adquirir, gestionar y financiar los sistemas necesarios para el ajuste fino y la inferencia?

El ajuste fino y el desarrollo de la IA generativa es una tarea que requiere muchos recursos informáticos: cada interacción puede implicar cálculos complejos sobre conjuntos de datos masivos en los que un LLM con miles de millones de parámetros recurre a sistemas informáticos especializados para potenciar su manipulación y análisis de la información. Cuanto más complejo sea el modelo, más cálculos necesitará para procesar los datos y ofrecer resultados relevantes. Necesitará un plan para mantener bajos los costes de computación tanto para el ajuste fino del modelo como para la inferencia, al tiempo que ofrece resultados de alta calidad.

En lugar de asumir los costes y desarrollar los conocimientos necesarios para construir un sistema internamente, muchas organizaciones seleccionan plataformas populares de ciencia de datos, como las que ofrecen los principales proveedores de servicios en la nube, para limpiar los conjuntos de datos y hacerlos pertinentes para una tarea.

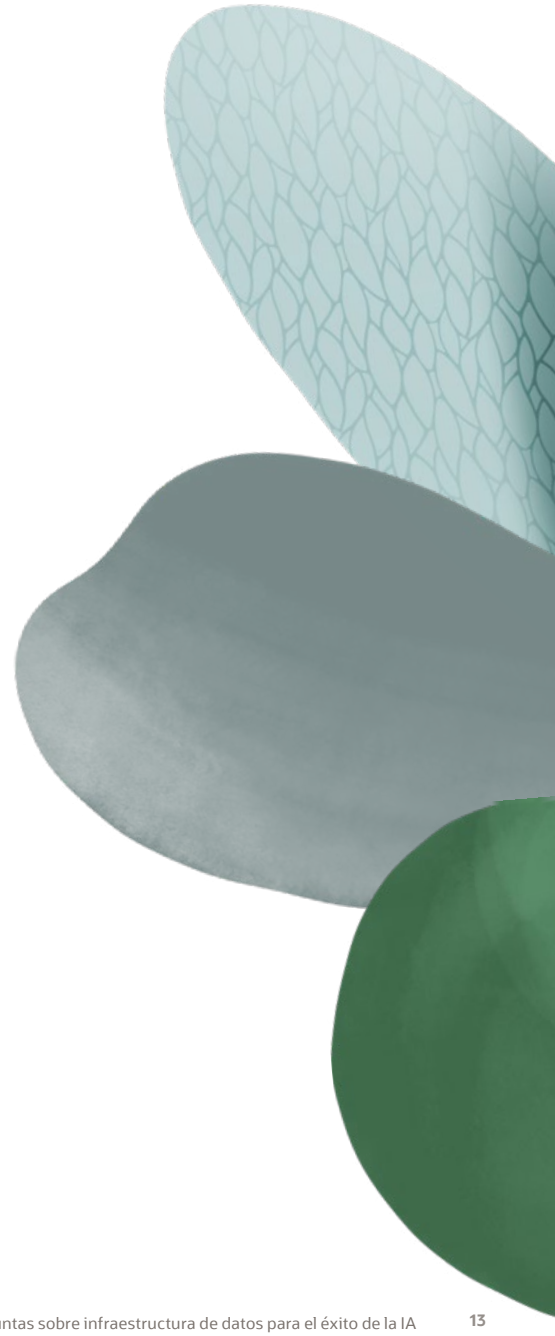
Luego, es importante conocer los servicios de infraestructura de IA de tus proveedores de nube. A menudo puedes obtener acceso o integración con herramientas de código abierto populares, así como con modelos básicos de IA generativa y otras tecnologías necesarias.

Ten en cuenta que la comunidad de código abierto y los hiperescaladores están dispuestos a ayudar a reducir los costes de la IA generativa. En la nube, encontrarás las potentes GPU y las redes de clúster necesarias para acelerar las cargas de trabajo de IA, así como funciones como el escalado automático, las instancias spot, las instancias reservadas, la programación optimizada de trabajos, la distribución eficiente de cargas de trabajo, el multi-tenancy y la optimización de modelos. Estas técnicas maximizan la utilización de los recursos: al adaptar la potencia de cálculo a las demandas exactas de la carga de trabajo en cada momento, las empresas pagan solo por lo que utilizan.

El ajuste fino también está siendo, por así decirlo, ajustado

Un método conocido como T-Few puede reducir la duración del entrenamiento y los recursos informáticos necesarios en comparación con los métodos de ajuste fino convencionales, manteniendo al mismo tiempo una gran precisión.

Por último, querrás asegurarte de que los modelos y la infraestructura funcionan con la máxima eficiencia. Puedes supervisar y optimizar el proceso de inferencia a través del servicio de IA generativa de tu proveedor en la nube o mediante estructuras eficientes de servicio de modelos, como TensorFlow Serving o TorchServe.



7 ¿Se tiene respaldo ejecutivo para el plan de IA? ¿Qué grupos pueden colaborar?

Lo más probable es que tu estrategia de IA comience con tus sistemas empresariales actuales. Los proveedores de CRM, HCM, ERP y otras aplicaciones esenciales están incorporando funciones de IA generativa y agentes de IA en los flujos de trabajo habituales. Piensa en el reconocimiento inteligente de documentos para ayudar a procesar las facturas de los proveedores más rápidamente y con mucho menos trabajo manual, narrativas generativas para aportar una comprensión más profunda de los informes y análisis, y una amplia gama de automatizaciones inteligentes en áreas como el análisis de riesgos y rendimiento. Estos elementos iluminarán el camino, pero necesitarás una estrategia más amplia para llevar la IA a tus flujos de trabajo exclusivos. Esto puede incluir el centro de excelencia mencionado anteriormente.

La incorporación de los servicios de IA generativa y los agentes de IA a las operaciones de tu empresa requerirá la participación y los recursos de toda la organización, incluidos los ejecutivos, los jefes de departamento, los equipos de TI, jurídicos y de cumplimiento normativo y, por supuesto, las personas que utilizarán la nueva tecnología. Si trabajas con las partes interesadas de cada unidad funcional para desarrollar el argumento comercial, centrándose en las ganancias de productividad y las ventajas competitivas, podrán desarrollar un enfoque coordinado.

Una vez establecidos los flujos de trabajo de IA con los que se va a comenzar, es el momento de involucrar a los equipos de TI y de ciencia de datos para comprender qué tipo de arquitectura se necesitará. Las preguntas incluyen: ¿Dónde residirán los LLM? ¿Quién pagará por ellos? ¿Cómo fluirán los datos? ¿Necesitamos nuevas bases de datos vectoriales o arquitecturas RAG? ¿Es necesario ajustar los LLM a un departamento o tarea en particular? ¿De dónde procederán el almacenamiento de datos y la potencia de computación? ¿Nuestro diseño de red y otras decisiones arquitectónicas nos proporcionarán la baja latencia y el alto rendimiento que necesitamos para la inferencia continua de IA?

Prepara una propuesta detallada con plazos, etapas y requisitos de recursos, así como un resumen de los riesgos, incluidas las consideraciones sobre privacidad, seguridad y conformidad de los datos, y los costes. Considera cómo pueden ayudarte tus proveedores de servicios en la nube de confianza.

> Ten en cuenta que ningún modelo de IA será eficaz sin un flujo constante de datos limpios y bien ordenados.

Por lo tanto, considera la implantación de la IA como una extensión de tu estrategia general de datos. Para fomentar la aceptación por parte de los ejecutivos, los equipos querrán confirmar que sus planes de IA están estrechamente alineados con la estrategia empresarial general y el plan de gobernanza de datos de la organización, sin aportar una complejidad innecesaria al mover datos y añadir recursos de gestión.

Soluciones Oracle

Cuando tu infraestructura de datos necesita ofrecer un ajuste fino y una inferencia rápidos y rentables de los modelos de IA generativa más avanzados del mundo, Oracle puede ayudarte.

Oracle Cloud Infrastructure

Oracle AI Infrastructure

OCI ofrece servicios integrales de IA e innovaciones de IA generativa de última generación, todo en una infraestructura de IA de primera clase. Puedes elegir modelos predefinidos o aprovechar un servicio de IA generativa que te permite elegir entre LLM de código abierto o patentados y, después, ajustar los modelos y aumentarlos con los datos de tu propia empresa. Además, tendrás acceso al sistema de gestión de bases de datos líder del sector.

Saber más

Oracle Database 23ai

Oracle Database 23ai es una base de datos multimodal para transacciones y análisis que proporciona una base de datos vectorial integrada junto con datos JSON, relacionales, gráficos y espaciales, así como aprendizaje automático en la base de datos. También puede realizar procesamiento vectorial y ejecutar Oracle Database en plataformas optimizadas de forma exclusiva que solo están disponibles en OCI y como servicios OCI en los centros de datos de nuestros socios multicloud.

Saber más

Servicios de OCI Data Science y de IA generativa

OCI Data Science es un servicio en la nube que permite a los científicos de datos construir, entrenar, desplegar y gestionar modelos de machine learning (ML) de forma colaborativa con tus estructuras de código abierto favoritas, o pueden aprovechar el ML en la base de datos. El servicio de IA generativa de OCI es un servicio totalmente gestionado para integrar los modelos lingüísticos de tu elección en una amplia gama de casos de uso.

Saber más

En OCI, conseguirás un ajuste fino, una inferencia y un procesamiento por lotes más rápidos y rentables. OCI proporciona la escala y el rendimiento necesarios para ejecutar grandes cargas de trabajo de IA con mayor rapidez, sin costes informáticos excesivos.

Cómo puede ayudar Oracle

Las relaciones multicloud únicas pueden facilitar el intercambio de datos.

Organizaciones de todo el mundo utilizan los servicios de Oracle Database que se ejecutan en OCI para crear y ejecutar rápidamente aplicaciones en Oracle Exadata Database Service y aprovechar las funcionalidades de Oracle Autonomous Database. Las relaciones multicloud permiten a las organizaciones que ejecutan aplicaciones en otras nubes de hiperescala obtener los mismos beneficios de Oracle Database con la simplicidad, seguridad y baja latencia de un único ambiente operativo.

Construye y ejecuta aplicaciones rápidamente en AWS, utilizando Exadata Database Service y Autonomous Database con servicios como Amazon Bedrock; en Azure, utilizando Exadata Database Service y Autonomous Database con servicios como Azure OpenAI; y en Google Cloud, utilizando Exadata Database Service y Autonomous Database con servicios como los modelos Vertex AI y Gemini Foundation de Google Cloud.

Saber más

Conéctate con nosotros

Llama al +34 91 631 2174 (España) o visita oracle.com/lad

Fuera de Norteamérica, busca tu oficina local en oracle.com/es/corporate/contact