WHITEPAPER

CALIDAD Y GOBERNANZA DE DATOS

Los seis principios de los datos preparados para la IA

Cómo establecer un conjunto de datos básicos para la IA

Índice

Resumen ejecutivo

- En este documento se describen seis principios para garantizar que los datos estén preparados para usarlos con la inteligencia artificial (IA).
- Dichos principios son los siguientes: los datos deben ser diversos, oportunos, precisos, seguros, reconocibles y las máquinas deben poder consumirlos con facilidad.
- En el documento también se describe la puntuación de confianza de la IA, que ayuda a evaluar hasta qué punto sus datos cumplen estos principios.

Introducción

Se espera que la inteligencia artificial (IA) mejore en gran medida sectores como la asistencia sanitaria, la fabricación y la atención al cliente, lo que generará experiencias de mayor calidad tanto para clientes como para empleados. De hecho, existen tecnologías de IA como el aprendizaje automático (ML) que ya han ayudado a responsables de la gestión de datos a producir predicciones matemáticas, generar conocimientos y mejorar la toma de decisiones. Además, tecnologías de IA emergentes como la IA generativa (GenAI) pueden crear contenido sorprendentemente realista que tiene el potencial de mejorar la productividad en prácticamente todos los aspectos del ámbito empresarial.

Según Gartner, en 2025, la IA generativa formará parte de la plantilla en el 90 % de empresas de todo el mundo, mientras que en 2026 más del 80 % de las empresas habrá implantado aplicaciones con capacidad de IA generativa en su entorno de producción.

Aun así, una IA de éxito debe basarse en datos de calidad, y en este documento se describen seis principios para garantizar que sus datos estén preparados para la IA.

[1] «Analysts to Discuss Generative AI Trends and Technologies» (Los analistas debaten las tendencias y tecnologías de la IA generativa), Gartner, octubre de 2023.

Los seis principios de los datos preparados para la IA

Sería ingenuo creer que simplemente añadiendo datos a diversas iniciativas de IA se puede esperar un resultado mágico, pero eso es lo que hacen muchos responsables de la gestión de datos. Aunque pueda parecer que este enfoque funciona en los primeros proyectos de IA, los científicos de datos cada vez pasan más tiempo corrigiendo y preparando los datos a medida que los proyectos maduran.

Además, los datos empleados para la IA deben ser de gran calidad y estar preparados con gran precisión para estas aplicaciones inteligentes. Para ello, hay que dedicar muchas horas a depurar y optimizar manualmente los datos para garantizar su precisión e integridad, además de organizarlos de una manera que las máquinas puedan entender con facilidad. Asimismo, estos datos suelen necesitar información adicional, como definiciones y etiquetas, para enriquecer el significado semántico del aprendizaje automático y ayudar a la IA a realizar tareas de forma más efectiva.

Por eso, cuanto antes se puedan preparar los datos para los procesos de IA posteriores, mayor será el beneficio. Emplear datos preparados previamente para la IA es como darle a un cocinero las verduras ya lavadas y cortadas, en vez de un saco lleno de comestibles: ahorra tiempo y esfuerzo, y ayuda a garantizar que el plato elaborado llegue al comensal rápidamente. En el diagrama siguiente se definen seis principios esenciales para garantizar la «preparación» de los datos y su facilidad de uso con la IA.

En el resto de apartados de este documento se describe detalladamente cada principio.

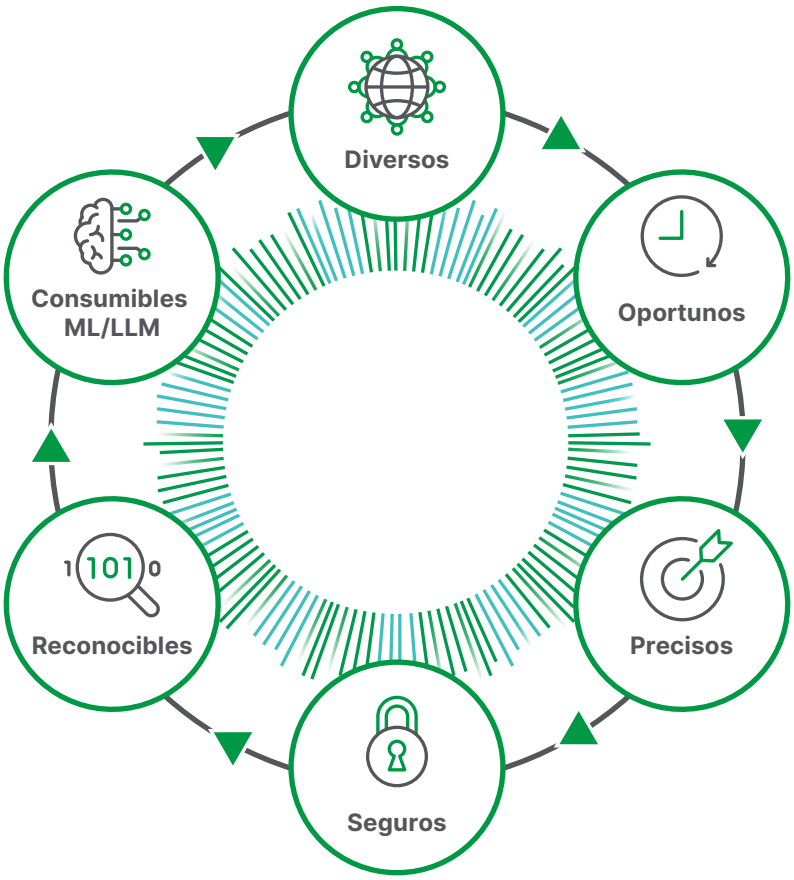

Figura 1. Seis principios de los datos preparados para la IA

1 Los datos deben ser diversos.

Los sesgos en los sistemas de IA, también conocidos como sesgos de algoritmo o de aprendizaje automático, aparecen cuando las aplicaciones de IA generan resultados que reflejan sesgos humanos, como la desigualdad social. Es algo que puede suceder cuando el proceso de desarrollo del algoritmo incluye prejuicios en sus supuestos o, lo que es más habitual, cuando los datos de entrenamiento presentan sesgos. Por ejemplo, un algoritmo de puntuación de crédito puede denegar un préstamo si emplea constantemente un grupo restringido de atributos financieros.

Por eso, nuestro primer principio se centra en proporcionar una amplia variedad de datos a los modelos de IA, lo que aumenta la diversidad de los datos y reduce el sesgo para ayudar a garantizar una menor probabilidad de que las aplicaciones de IA tomen decisiones injustas.

Un conjunto de datos diverso implica que los modelos de IA no se basen en conjuntos de datos restringidos y aislados. De hecho, los datos deben extraerse de una amplia variedad de orígenes de datos que abarquen distintos patrones, perspectivas, variaciones y situaciones pertinentes para la tarea en cuestión. Es posible que estos datos estén bien estructurados y se alojen en la nube o in situ en las instalaciones. También podrían existir en sistemas mainframe, de bases de datos o SAP, o bien en una aplicación de software como servicio (SaaS). Asimismo, también es posible que los datos de origen no estén estructurados y figuren como archivos o documentos en una unidad empresarial.

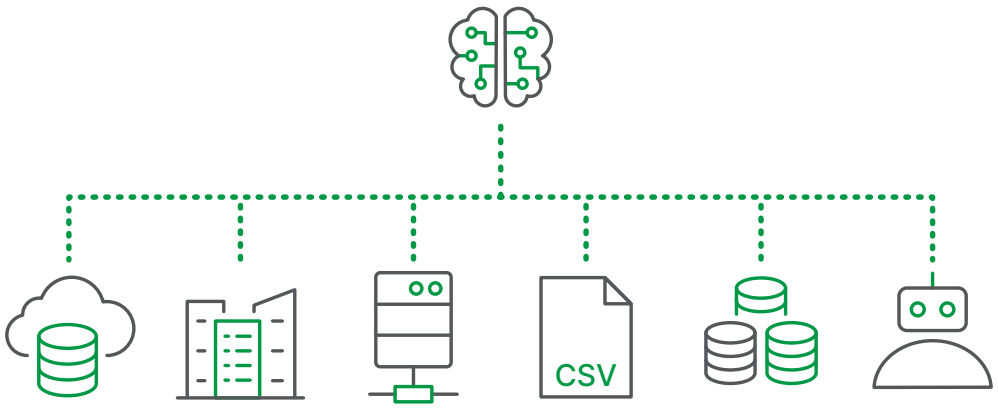

Figura 2. Diversidad de los datos

Es fundamental adquirir datos variados de diversas formas para la integración en sus aplicaciones de ML e IA generativa.

2 Los datos deben ser oportunos.

Si bien es cierto que las aplicaciones de ML e IA generativa avanzan en su trabajo gracias a la diversidad de los datos, el grado de actualidad de dichos datos también es crucial. Del mismo modo que una previsión meteorológica basada en las condiciones de ayer no es adecuada para planificar un viaje hoy, los modelos de IA entrenados con información obsoleta pueden generar resultados inexactos o irrelevantes. Además, los datos actualizados permiten a los modelos de IA seguir el ritmo de las tendencias, adaptarse a las circunstancias cambiantes y ofrecer los mejores resultados posibles. Por tanto, el segundo principio de los datos preparados para la AI es su carácter oportuno.

Es fundamental que desarrolle e implemente canales de datos en tiempo real de baja latencia para que sus iniciativas de IA garanticen datos oportunos. A menudo se utiliza la **captura de datos modificados (CDC)** para recabar datos oportunos de sistemas de bases de datos relacionales; por su parte, la **captura de datos de transmisión** se emplea con los datos que se originan en dispositivos del IoT que requieren un procesamiento de baja latencia. Una vez que se capturan los datos, se actualizan los repositorios de destino y **se aplican cambios continuamente** en tiempo casi real para garantizar la máxima actualidad de los datos.

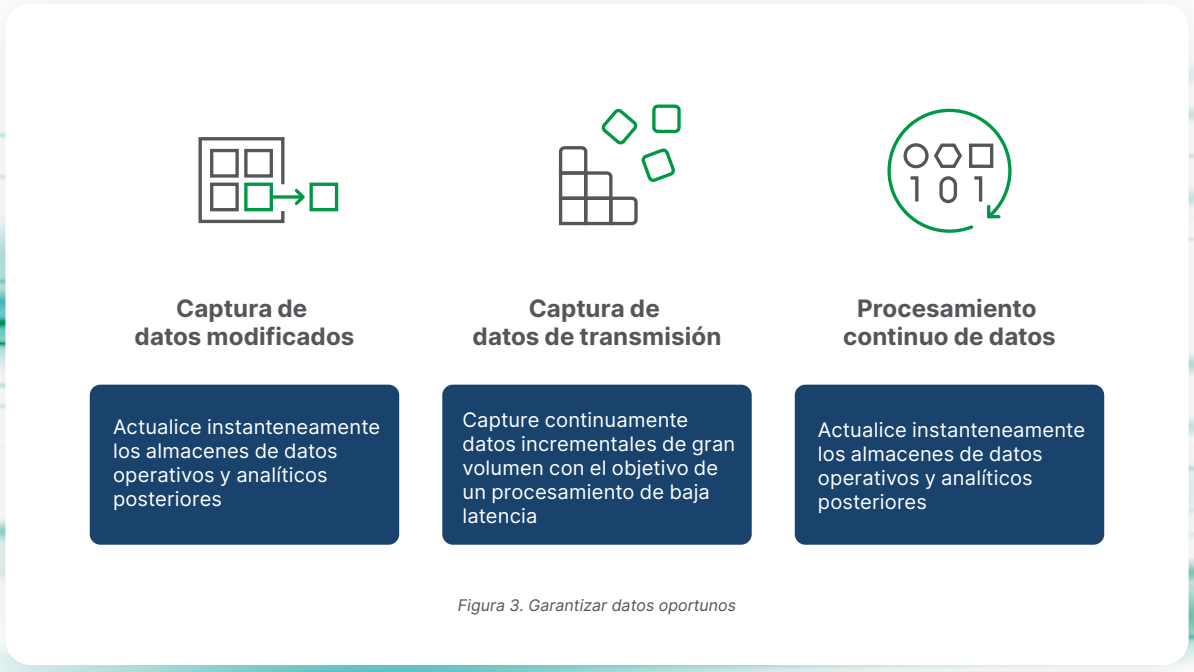

Figura 3. Garantizar datos oportunos

Recuerde que unos datos oportunos permiten predicciones más precisas y con mejor información.

3 Los datos deben ser precisos.

El éxito de cualquier iniciativa de ML o IA generativa depende de un ingrediente clave: datos correctos. El motivo es que los modelos de IA actúan como si fueran sofisticadas esponjas que absorben información para aprender y realizar tareas. Si la información no es precisa, es como si la esponja estuviera absorbiendo agua sucia, lo que provoca resultados sesgados, creaciones que no tienen sentido y, en última instancia, fallos de funcionamiento del sistema de IA. Por tanto, la precisión de los datos es el tercer principio y un precepto fundamental para desarrollar aplicaciones de IA fiables y de confianza.

La precisión de los datos incluye tres aspectos. El primero consiste en **perfilar los datos de origen** para comprender sus características, integridad, distribución, redundancia y forma. El perfilado también se suele denominar análisis exploratorio de datos (o EDA).

El segundo aspecto es la **puesta en práctica de estrategias de saneamiento** creando, implementando y supervisando continuamente la eficacia de las reglas de calidad de los datos. Es posible que su administrador de datos deba participar en esta fase para ayudar con la desduplicación y combinación de datos. La IA también puede ayudar a automatizar y acelerar el proceso mediante sugerencias sobre la calidad de los datos.

El aspecto final consiste en habilitar el linaje de datos yel análisis del impacto. Se realizará con herramientas para ingenieros y científicos de datos que resalten el efecto de los posibles cambios de datos y rastreen el origen de los datos para evitar la modificación accidental de los datos utilizados por los modelos de IA.

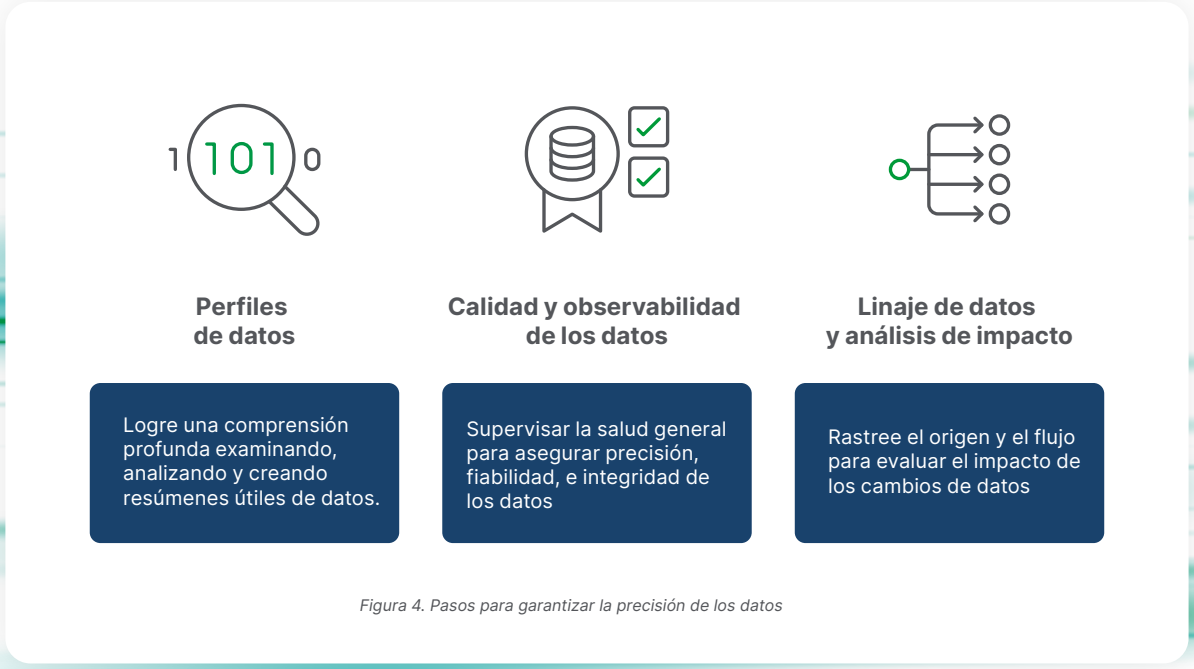

Figura 4. Pasos para garantizar la precisión de los datos

Unos datos precisos y de alta calidad garantizan que los modelos puedan identificar patrones y relaciones pertinentes, lo que derivará en decisiones, generación y predicciones más precisas.

4 Los datos deben estar seguros.

Los sistemas de IA suelen utilizar datos confidenciales, incluida información personal (PII), registros financieros e información empresarial privilegiada, por lo que el uso de estos datos exige responsabilidad. Dejar los datos desprotegidos en aplicaciones de IA es como dejar abierta la puerta de una caja fuerte. Los agentes maliciosos podrían robar información confidencial, manipular los datos de entrenamiento para sesgar los resultados o incluso interrumpir por completo los sistemas de IA generativa. Proteger los datos es fundamental para salvaguardar la privacidad, mantener la integridad de los modelos y garantizar el desarrollo responsable de aplicaciones de IA potentes. Por tanto, la seguridad de los datos es el cuarto principio de los datos preparados para la IA.

Una vez más, existen tres tácticas que pueden servir de ayuda para automatizar la seguridad de los datos a escala, ya que es casi imposible hacerlo manualmente. La **clasificación de datos** detecta, clasifica y etiqueta los datos que alimentan la siguiente etapa. La **protección de datos** define políticas como la ocultación, la tokenización y el cifrado para enmascarar los datos. Por último, la **seguridad de datos** define las políticas que describen el control de acceso, es decir, quién puede acceder a los datos. Los tres conceptos se combinan de la siguiente manera: primero, deben definirse niveles de privacidad y deben etiquetarse datos con una designación de seguridad de *sensible*, *confidencial* o *restringido*. En segundo lugar, se debe aplicar una política de protección para ocultar los datos restringidos. Por último, se debe utilizar una política de control de acceso para limitar los derechos de acceso.

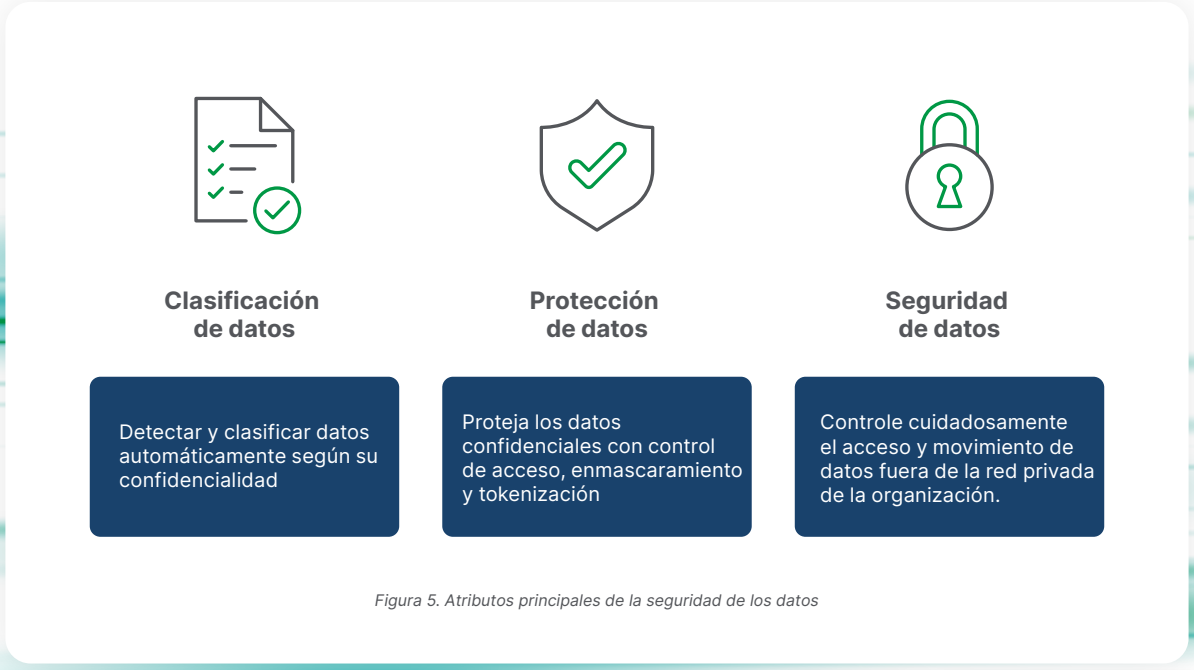

Figura 5. Atributos principales de la seguridad de los datos

Estas tres tácticas protegen sus datos y son cruciales para mejorar la confianza general en su sistema de IA y salvaguardar la reputación de esta tecnología.

5 Los datos deben ser reconocibles.

Los principios que hemos comentado hasta ahora se han centrado principalmente en suministrar con rapidez los datos pertinentes, en el formato correcto y a las personas, sistemas o aplicaciones de IA adecuados. Pero la acumulación de datos no es suficiente. Los datos preparados para la IA deben ser reconocibles y se debe poder acceder a ellos con facilidad en el propio sistema. Imagine una biblioteca en la que todos los libros están guardados bajo llave: el conocimiento está ahí, pero no se puede utilizar. Los datos reconocibles desbloquean el verdadero potencial de las tecnologías de ML e IA generativa, ya que permiten que estas cargas de trabajo encuentren la información que necesitan para aprender, adaptarse y generar resultados innovadores. Por tanto, la capacidad de reconocimiento es el quinto principio de los datos preparados para la IA.

Como era de esperar, en el centro de la capacidad de reconocimiento se encuentran unas buenas prácticas de metadatos. Además de los metadatos técnicos asociados con los conjuntos de datos de IA, también se deben definir los metadatos empresariales y la tipificación semántica. La **tipificación semántica** proporciona un significado adicional para los sistemas automatizados, mientras que los términos empresariales adicionales ofrecen un mayor contexto para ayudar con la comprensión humana. Una de las mejores prácticas consiste en crear un **glosario empresarial** que asigne términos empresariales a elementos técnicos en los conjuntos de datos, para garantizar una comprensión común de los conceptos. La mejora asistida por IA también se puede utilizar para generar documentación automáticamente e incorporar semántica empresarial a partir del glosario. Por último, todos los metadatos están indexados y permiten realizar búsquedas mediante un **catálogo de metadatos**.

Figura 6. Atributos principales de los datos reconocibles

Este enfoque garantiza que los datos se pueden reconocer y aplicar directamente, y que sean prácticos y significativos para la tarea de IA que se está realizando.

6 El aprendizaje automático (ML) y los modelos de lenguaje de gran tamaño (LLM) deben ser capaces de consumir los datos con facilidad.

Ya hemos mencionado que las aplicaciones de ML e IA generativa son herramientas potentes, pero su potencial se basa en la capacidad de consumir datos fácilmente. A diferencia de los humanos, que pueden descifrar notas manuscritas o buscar información en hojas de cálculo desordenadas, estas tecnologías necesitan que la información se represente en formatos específicos. Imagine dar de comer a una persona quisquillosa: si no se come lo que le sirve, pasará hambre. De igual forma, las iniciativas de IA no tendrán éxito si los datos no tienen el formato correcto para los experimentos de ML o las aplicaciones de LLM. Facilitar el consumo de los datos ayuda a desbloquear el potencial de estos sistemas de IA, lo que les permite ingerir información sin problemas y convertirla en acciones inteligentes para obtener resultados creativos. Por tanto, facilitar el consumo de los datos es el principio final de los datos preparados para la IA.

Cómo facilitar el consumo de los datos para el aprendizaje automático

La transformación de datos es el héroe anónimo de los datos que los sistemas de ML pueden consumir. Aunque algoritmos como los de regresión lineal se llevan todo el protagonismo, la calidad y la forma de los datos que se utilizan para entrenarlos son igual de importantes. Además, el esfuerzo invertido en depurar, organizar y facilitar que los modelos de ML consuman los datos genera importantes recompensas. Los datos preparados permiten que los modelos aprendan de manera efectiva, lo que genera predicciones precisas, resultados fiables y, en última instancia, el éxito de todo el proyecto de ML.

Sin embargo, los formatos de los datos de entrenamiento dependen en gran medida de la infraestructura de ML subyacente. Los sistemas de ML tradicionales se basan en discos, y gran parte del flujo de trabajo de los científicos de datos se centra en establecer las mejores prácticas y procedimientos de codificación manual para gestionar grandes volúmenes de archivos. En los últimos tiempos, los sistemas de ML basados en lakehouse han utilizado un almacén de funciones de tipo base de datos y el flujo de trabajo de los científicos de datos ha pasado de lenguaje SQL a un lenguaje de primera clase. El resultado es que las estructuras de datos tabulados de gran calidad y con el formato adecuado son el formato de datos más consumible y práctico para los sistemas de ML.

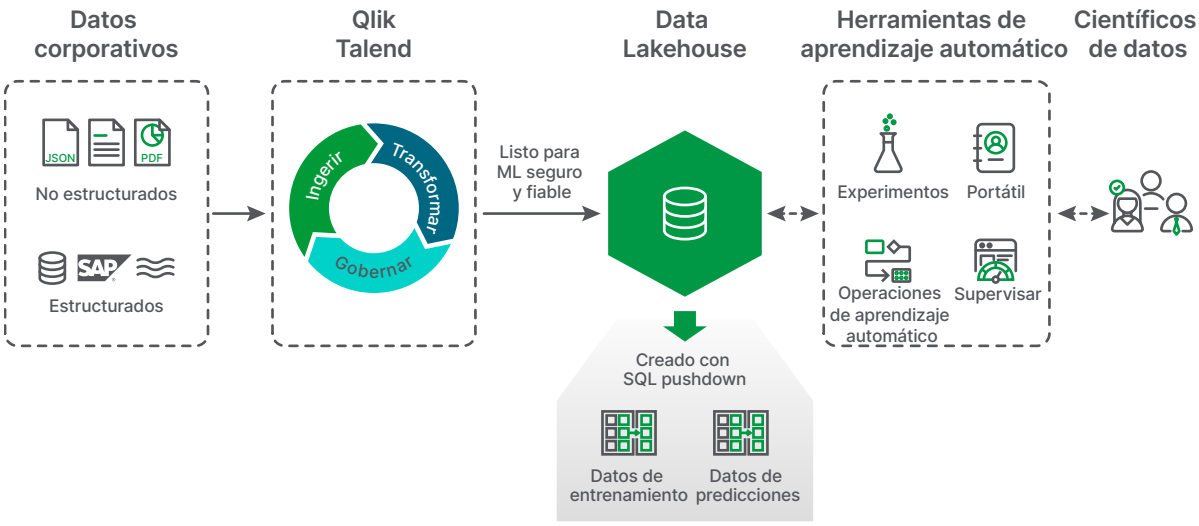

Figura 7. Facilitar el consumo de los datos para ML

Cómo facilitar el consumo de los datos para la IA generativa

Los modelos de lenguaje de gran tamaño (LLM) como GPT-4 de OpenAI, Claude de Anthropic y LaMDA y Gemini de Google AI, han sido entrenados previamente con grandes cantidades de datos de texto y son el corazón de la IA generativa. Se calcula que el modelo GPT-3 de OpenAI se ha entrenado con aproximadamente 45 TB de datos, es decir, más de 300 mil millones de tokens. A pesar de esta gran cantidad de información introducida, los LLM no pueden responder a preguntas específicas sobre su empresa porque no tienen acceso a los datos de la empresa. La solución es mejorar estos modelos con su propia información para generar aplicaciones de IA más correctas, pertinentes y fiables.

El método para integrar sus datos corporativos en una aplicación basada en LLM se denomina generación mejorada por recuperación (o RAG). Esta técnica suele emplear información de texto derivada de fuentes no estructuradas basadas en archivos, como presentaciones, correos archivados, documentos de texto, archivos PDF, transcripciones, etc. A continuación, el texto se divide en fragmentos que se pueden gestionar y se convierte en una representación numérica que el LLM utiliza en un proceso conocido como integración. Posteriormente, estas integraciones se almacenan en una base de datos vectorial como Chroma, Pinecone o Weviate. Conviene señalar que muchos proveedores de bases de datos tradicionales, como PostgreSQL, Redis y SingleStoreDB, también admiten vectores. Además, algunas plataformas en la nube como Databricks, Snowflake y Google BigQuery también han incorporado recientemente compatibilidad con vectores.

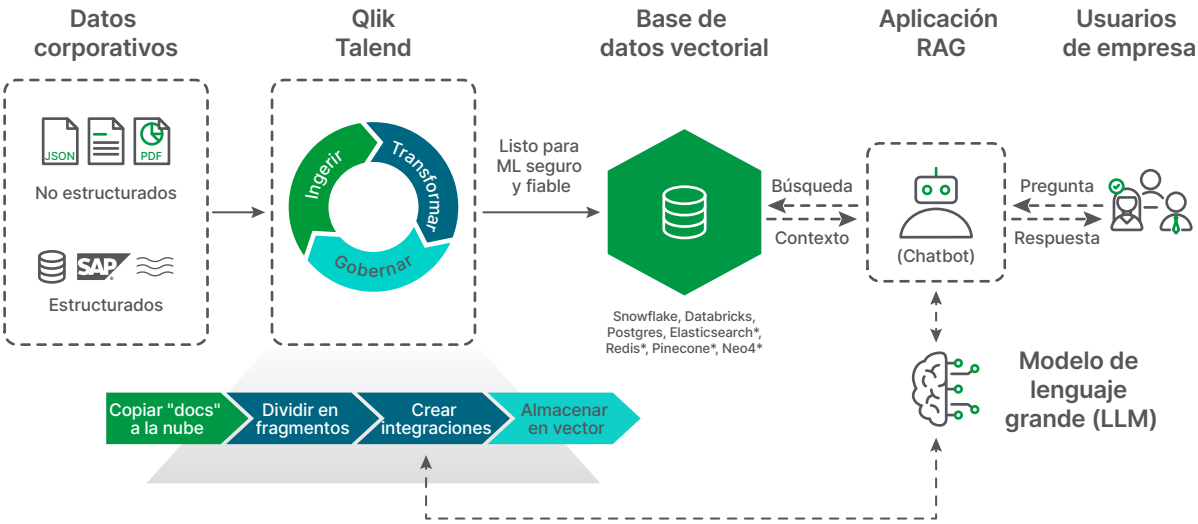

Figura 8. Facilitar el consumo de los datos para la IA generativa

Independientemente de si sus datos de origen están o no están estructurados, el enfoque de Qlik garantiza que sus aplicaciones de IA generativa, RAG o basadas en LLM puedan consumir datos de calidad con facilidad.

La puntuación de confianza en la IA

Una vez definidos los seis principios básicos de preparación e idoneidad de los datos, aún quedan preguntas por responder: ¿se pueden codificar y convertir los principios con facilidad para el uso diario? ¿Y cómo se puede distinguir con rapidez si los datos están preparados para la IA o no? Una posibilidad es utilizar la puntuación de confianza en la IA de Qlik como un indicador de preparación global y comprensible.

La puntuación de confianza en la IA asigna una dimensión independiente a cada principio y luego agrega cada valor para crear una puntuación compuesta, un atajo rápido y fiable para evaluar hasta qué punto los datos están preparados para la IA. Además, debido a que los datos empresariales cambian constantemente, la puntuación de confianza se comprueba periódicamente y se recalibra con frecuencia para rastrear las tendencias de preparación de los datos.

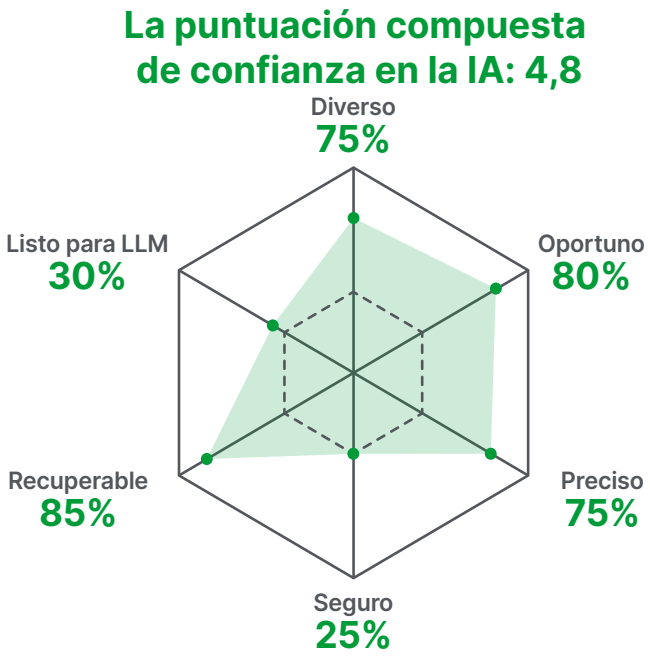

Figura 9. La puntuación de confianza de preparación de la IA

La puntuación de confianza de la IA agrega varias métricas a una puntuación de preparación única y fácil de comprender.

Conjunto de datos básicos de Qlik Talend para la IA

Nunca ha existido una necesidad tan grande de datos de alta calidad en tiempo real que lleven a decisiones más reflexivas, eficiencia operativa e innovación empresarial. Por eso, las organizaciones de éxito quieren trabajar con las soluciones de calidad e integración de datos líderes de Qlik Talend para proporcionar de manera eficiente datos de confianza a warehouses, repositorios y otras plataformas de datos empresariales. Nuestras excelentes y completas soluciones utilizan canales automatizados, transformaciones inteligentes y calidad de conjuntos de datos fiable para ofrecer la agilidad a la que aspiran los profesionales de los datos respecto a la gobernanza y el cumplimiento normativo que esperan las empresas.

Por eso, no importa si crea warehouses o repositorios para obtener análisis esclarecedores, si moderniza infraestructuras de datos operativos para la eficiencia empresarial o si utiliza de datos de varias nubes para iniciativas de inteligencia artificial, Qlik Talend puede guiarle.

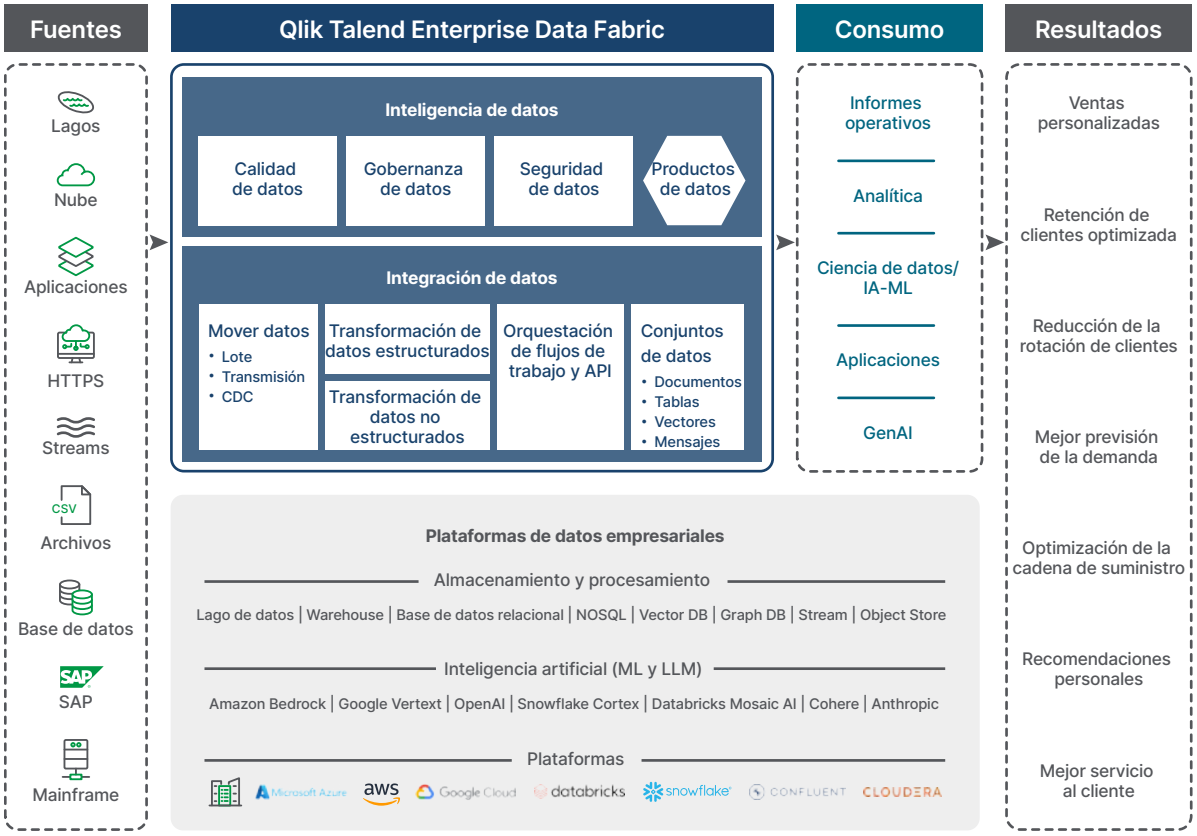

Figura 10. Qlik Talend Enterprise Data Fabric para IA y análisis

Conclusión

A pesar del poder transformador del aprendizaje automático y el potencial de crecimiento explosivo de la IA generativa, la preparación de los datos sigue siendo la piedra angular de cualquier implementación de IA de éxito. En este documento se han descrito seis principios clave para establecer conjuntos de datos básicos sólidos y fiables que se combinen para ayudar a su organización a desarrollar todo el potencial de la IA.

¿Quiere desarrollar todo el potencial de IA de su organización?

Empezar aquí

Acerca de Qlik

Qlik convierte entornos de datos complejos en información útil, generando resultados empresariales estratégicos. Con más de 40.000 clientes en todo el mundo, nuestra cartera de productos integra capacidades avanzadas de IA y ML empresarial, mejorando de manera constante la calidad de los datos. Destacamos en integración y gobernanza de datos, ofreciendo soluciones integrales que operan con fuentes de datos dispares. El análisis intuitivo en tiempo real de Qlik revela patrones ocultos, permitiendo a los equipos abordar desafíos complejos y aprovechar nuevas oportunidades. Nuestras herramientas de IA/ML, prácticas y escalables, permiten tomar mejores decisiones con mayor rapidez. Como socio estratégico, nuestra tecnología y experiencia hacen que nuestros clientes sean más competitivos, independientemente de la plataforma que utilicen.

qlik.com

24-DQ-0001-01-WHM