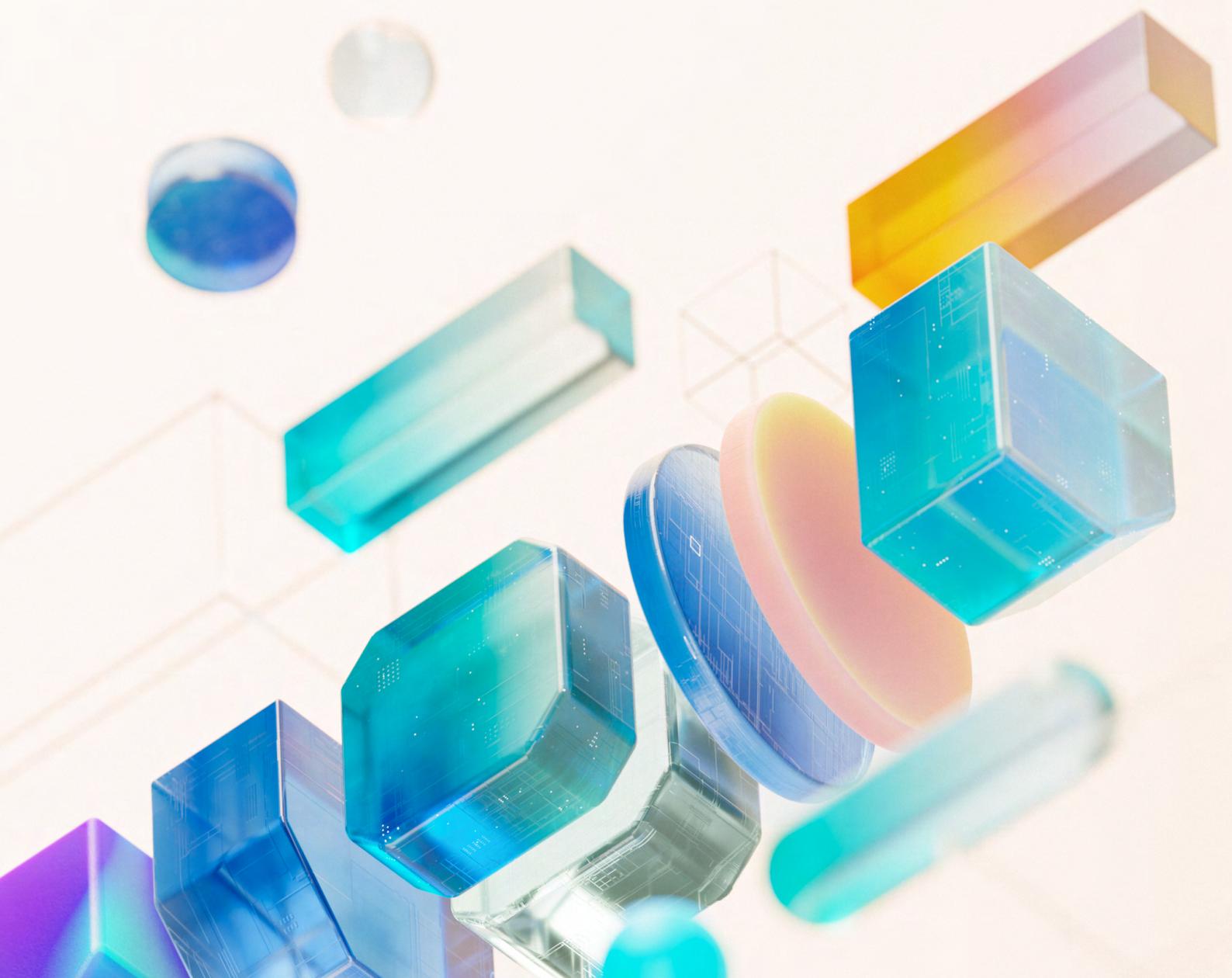




Sinergia en los análisis: unificar Azure Databricks y Microsoft Fabric



Sinergia en los análisis: unificar Azure Databricks y Microsoft Fabric

3 /

Potencia análisis de datos modernos con Azure Databricks y Microsoft Fabric

26 /

Actividad de Data Factory y Azure Databricks en Microsoft Fabric

4 /

Simplifica las cargas de trabajo de análisis con Azure Databricks y Microsoft Fabric

28 /

Mejora las funcionalidades organizativas con IA generativa

10 /

Arquitectura de medallas en Azure Databricks y Microsoft Fabric

33 /

Explora casos de uso reales con ejemplos prácticos

15 /

Utiliza datos de almacén de lago con Azure Databricks y Microsoft Fabric

42 /

Logra la excelencia con Azure Databricks y Microsoft Fabric

22 /

Mejor juntos: Azure Databricks, Catalog, Unity Catalog y Microsoft Fabric Purview

43 /

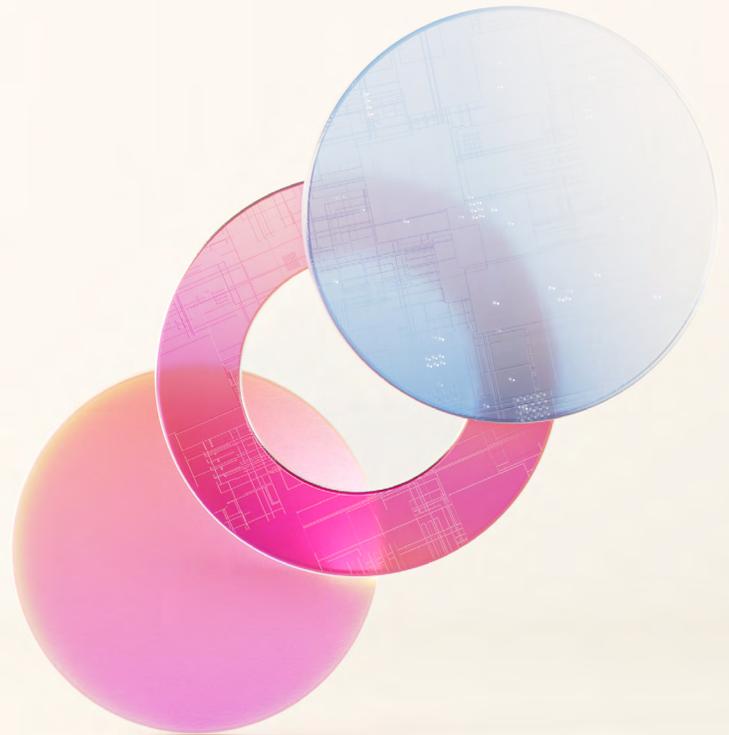
Siguientes pasos

Potencia análisis de datos modernos con Azure Databricks y Microsoft Fabric

En la actualidad, las organizaciones necesitan entender cómo administrar la creciente avalancha de datos de una manera eficiente y perspicaz. Un almacén de lago de datos combina el gran almacenamiento de un lago de datos con el procesamiento estructurado de diversos servicios de datos. Admite un almacenamiento de datos amplio y análisis complejos sin realizar concesiones. Es algo más que una solución de almacenamiento: mejora la inteligencia de los datos y admite análisis avanzados, para abordar el desafío de convertir datos abundantes en conocimientos útiles.

Los entornos de cloud ofrecen grandes recursos computacionales y escalabilidad bajo demanda. A medida que aumentan los datos de una organización, su infraestructura puede crecer junto a ellos de forma fluida y rentable. Esta sinergia entre las plataformas de cloud y la arquitectura de lago de datos es crucial, ya que proporciona una base resiliente y adaptable para cualquier empresa que quiera prosperar en la economía basada en datos. La administración eficaz y unas medidas de seguridad sólidas son esenciales en el panorama de los datos basados en el cloud para proteger este valioso activo.

Tanto Azure Databricks como Microsoft Fabric son soluciones de análisis integrales. Fabric incluye herramientas más fáciles de usar para la empresa y Azure Databricks tiene una plataforma de IA integrada, pero como ambos dependen de la misma capa de datos, se pueden usar juntos como un todo más potente. Azure Databricks, Fabric y OneLake permiten a las organizaciones agilizar su arquitectura de datos, simplificando las cargas de trabajo de análisis y permitiendo una administración y un análisis de datos eficientes en una plataforma unificada.



Simplifica las cargas de trabajo de análisis con Azure Databricks y Microsoft Fabric

Con la arquitectura moderna de lagos de datos, las empresas pueden sacar partido de las sinergias entre Azure Databricks y Microsoft Fabric. Tanto Azure Databricks como Fabric ofrecen un conjunto unificado y completo de herramientas para una amplia gama de casos de análisis avanzados y funcionan juntos para proporcionar un abanico completo de soluciones para trabajar con un almacén de lago de datos. Con elementos de ingeniería de datos, ciencia de datos, almacenamiento de datos y Power BI, ofrecen características de análisis de gran alcance, una experiencia cohesiva para los usuarios y un único repositorio de datos accesible para diversas herramientas de análisis. Azure Databricks también proporciona un seguimiento exhaustivo de la gestión y el linaje de los datos y los activos de IA en una sola experiencia unificada.

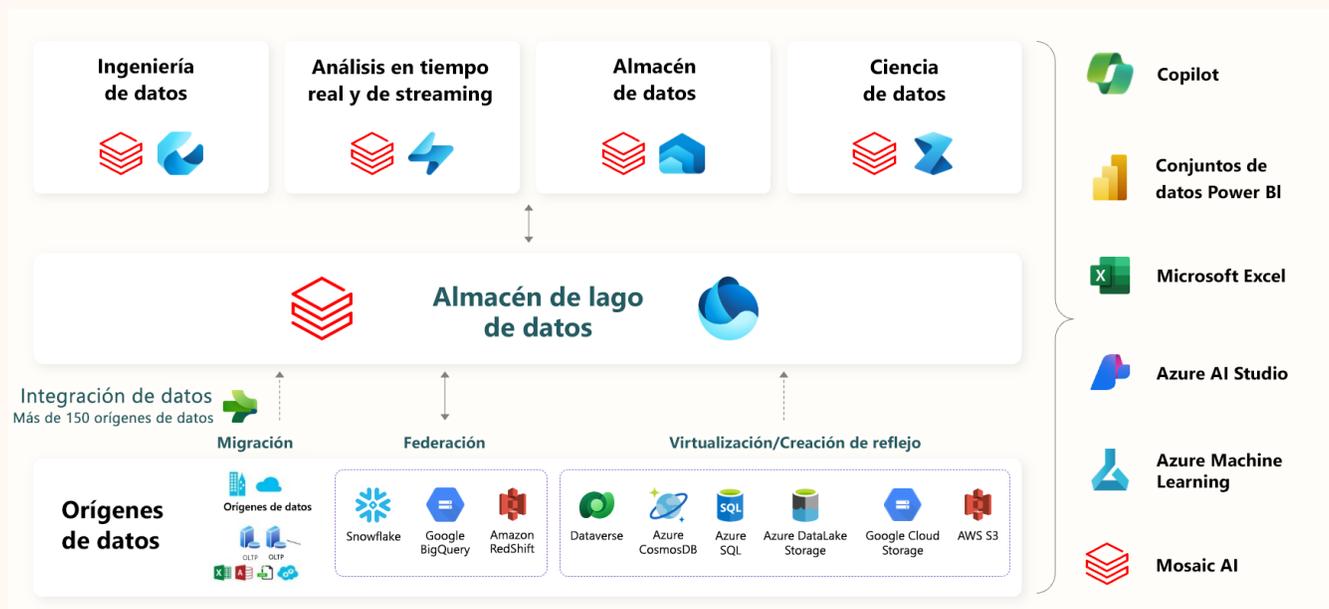


Figura 1: Integración de Azure Databricks y Microsoft Fabric en una arquitectura de almacén de lago

Maximiza el potencial de los datos con Azure Databricks y Microsoft Fabric

Gracias a la integración de Azure Databricks y Fabric, los usuarios pueden cambiar sin problemas de una plataforma a otra, lo que aporta a los clientes una solución cohesionada y potente para la administración y el análisis de datos, y hace posible proyectos de IA y machine learning con facilidad y eficiencia.

Administración de datos

OneLake centraliza los datos de diversas fuentes. La integración de Azure Databricks con Fabric no solo revoluciona la administración de datos, la escalabilidad y el procesamiento de datos, sino que además centraliza los datos de una amplia gama de orígenes a través de OneLake. Este enfoque integral garantiza que Azure Databricks pueda conectarse sin problemas con los datos almacenados en Azure Data Lake Storage (ADLS), distintas bases de datos y el propio OneLake. Esto simplifica la administración de grandes volúmenes de datos, mejora la capacidad de escalar proyectos de datos y agiliza la canalización del procesamiento de los mismos.

- **Almacenamiento centralizado:** el uso de OneLake dentro de Fabric permite la administración centralizada de datos, lo que simplifica el acceso y la gestión de los datos y garantiza que Azure Databricks utilice directamente los datos para los procesos analíticos.
- **Integración perfecta:** la integración perfecta entre Azure Databricks y Data Factory en Fabric hace posible flujos de trabajo optimizados, desde la ingesta y validación de los datos hasta la transformación. Esta integración permite una estrategia de administración de datos cohesiva que contribuye al análisis de datos, la ciencia de datos y los proyectos de IA.
- **Mayor seguridad y accesibilidad:** las áreas de trabajo premium de Azure Databricks admiten credenciales diferidas, lo que refuerza la seguridad y la facilidad de acceso a los recursos de OneLake. Esta característica garantiza un acceso seguro y sencillo a los datos centralizados para su posterior procesamiento y análisis.

Escalabilidad y reproducibilidad

Azure Databricks y Fabric admiten flujos de trabajo de datos escalables, proyectos reproducibles de IA y análisis y capacidades de procesamiento dinámico de datos. Esta integración permite a las organizaciones administrar grandes volúmenes de datos de manera eficiente, garantizar resultados coherentes en sus entornos de datos y adaptar la potencia de procesamiento para satisfacer diferentes requisitos, lo que impulsa operaciones de datos fiables y escalables.

- **Flujos de trabajo de datos escalables:** las actividades de Azure Databricks dentro de Data Factory se han diseñado para admitir flujos de trabajo de datos escalables. Las organizaciones pueden gestionar de forma eficiente grandes cantidades de datos y escalar sus operaciones de procesamiento y análisis de datos según sea necesario, sin renunciar al rendimiento ni a la fiabilidad.
- **Proyectos de IA reproducibles:** la integración garantiza que los proyectos de IA y análisis sean reproducibles, con lo que se benefician de características como el control de versiones de datos y el seguimiento de linaje. Estas características, disponibles de forma nativa en ambas plataformas, mejoran la fiabilidad de los proyectos de IA y garantizan la coherencia en todos los entornos de datos.

- **Funciones de procesamiento**

dinámico dinámicos: Azure Databricks ofrece funciones de procesamiento dinámico de datos que se adaptan a los distintos volúmenes de datos y a los diversos requisitos de procesamiento. Esta flexibilidad es crucial para que las organizaciones escalen con eficiencia sus operaciones de análisis de datos.

Procesamiento de datos

Al combinar las sólidas capacidades de transformación de datos de Azure Databricks con la sofisticada orquestación de las canalizaciones de Data Factory, estas tecnologías crean sinergias que respaldan una amplia gama de tareas de procesamiento de datos.

- **Transformación eficiente de los datos:** Azure Databricks destaca en la transformación de los datos almacenados en OneLake y otros orígenes. Juntos, Azure Databricks y Fabric admiten una amplia gama de tareas de procesamiento de datos, incluida la exploración, limpieza y preparación de datos, cruciales para preparar conjuntos de datos para IA y machine learning.

- **Orquestación de flujos de trabajo complejos:** las canalizaciones de Data Factory que incluyen actividades de Azure Databricks permiten la orquestación de flujos de trabajo de transformación de datos complejos. Estas canalizaciones pueden validar orígenes de datos, copiar datos en el almacenamiento designado y ejecutar cuadernos para la transformación de datos, lo que proporciona una solución integral para el procesamiento de datos.

La sinergia entre Azure Databricks y Fabric, sobre todo a través de Data Factory, mejora la administración de datos, garantiza la escalabilidad y reproducibilidad y facilita un procesamiento de datos eficiente. Esta integración es fundamental para que las organizaciones utilicen sus datos para realizar análisis útiles y tomar decisiones basadas en IA.

Ventajas de la arquitectura de almacén de lago

La arquitectura de almacén de lago se basa en el formato de almacenamiento de código abierto Delta Lake. Además de sus capacidades técnicas, como la coherencia de transacciones ACID, mejora la eficacia general de toda la plataforma. Además, permite el uso mediante varios motores de procesamiento, ya que utiliza formatos abiertos y permite que herramientas como Azure Databricks y Fabric funcionen con la misma copia de los datos al mismo tiempo.

Las empresas no tienen que depender de una sola herramienta para procesar sus datos; en lugar de ello, pueden seleccionar la mejor herramienta para cada proyecto. La arquitectura de almacén de lago revoluciona la forma en que las empresas administran, escalan y procesan sus datos. Este enfoque innovador garantiza la integridad y la coherencia de los datos mediante el soporte transaccional, lo que fomenta una plataforma de administración de datos más eficaz. Por otro lado, la capacidad de la arquitectura de combinar diferentes formatos de almacenamiento bajo un mismo techo simplifica el complejo panorama de los patrimonios de datos. Facilita el ajuste dinámico de los recursos informáticos en consonancia con la demanda en tiempo real, eliminando el derroche de aprovisionamiento, mejorando la rentabilidad y el uso de recursos.

En el núcleo de esta arquitectura se encuentra la integración de amplios lagos de datos y almacenes de datos estructurados, con lo que se crea un entorno óptimo para fomentar la innovación en IA y machine learning. Esto garantiza el acceso a la potencia computacional y a los datos, acelerando la innovación y agilizando la administración de sistemas de datos diversos.

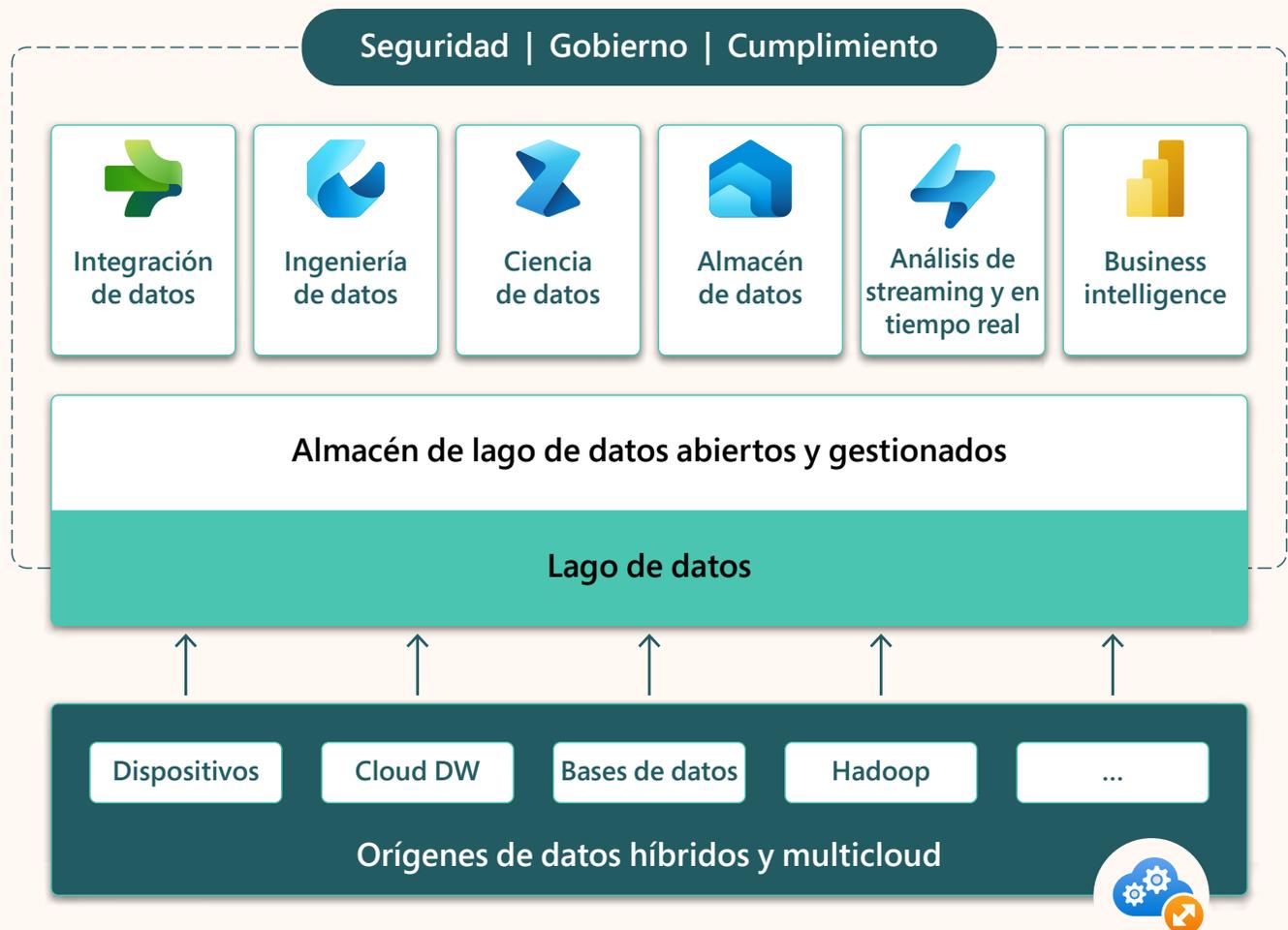


Figura 2: Análisis en almacén de lago

Con la combinación única de la amplia capacidad de almacenamiento de un lago de datos y el entorno estructurado y optimizado para consultas de un almacén de datos, el almacén de lago moderno surge como la plataforma ideal para desarrollar e implementar algoritmos de IA. Esta doble capacidad garantiza que los proyectos de IA puedan utilizar la potencia computacional y la accesibilidad de los datos necesarios, acelerando la innovación y reduciendo los costes asociados con la administración de sistemas de datos separados.

Al simplificar la arquitectura de datos y reducir la complejidad de la infraestructura, las empresas pueden centrarse en generar valor a través de la IA, en lugar de enfrentarse a los desafíos de administración de datos de las siguientes maneras:

- La arquitectura del almacén de datos guarda archivos de Delta Lake en una cuenta de ADLS. Este servicio de almacenamiento en el cloud es extremadamente rentable y el formato Delta Lake permite el almacenamiento de datos estructurados y no estructurados.
- La creación de modelos de IA requiere grandes cantidades de potencia de computación tanto de CPU tradicionales como de GPU avanzadas. Dado que la arquitectura del almacén de lago de datos permite el uso de múltiples motores de computación, incluidos Azure Databricks y Fabric, las empresas pueden aportar el tipo idóneo de potencia de procesamiento a sus tareas de exploración y modelado de datos.

Las empresas pueden utilizar las capacidades avanzadas de machine learning e IA de Azure Databricks y Fabric en todo su patrimonio de datos almacenado en un almacén de lago de datos. Estas herramientas incluyen la administración de experimentos de extremo a extremo y conjuntos de herramientas automatizados de machine learning que pueden impulsar los proyectos de IA.



Arquitectura de medallas en Azure Databricks y Microsoft Fabric

La arquitectura de medallas es un enfoque sofisticado dentro del concepto más amplio de la arquitectura de almacén de lago, diseñado para agilizar los flujos de trabajo de datos, desde la ingesta hasta el conocimiento. Fundamentalmente, consta de tres capas: bronce, plata y oro, cada una con una finalidad diferenciada en el ciclo de vida de los datos.

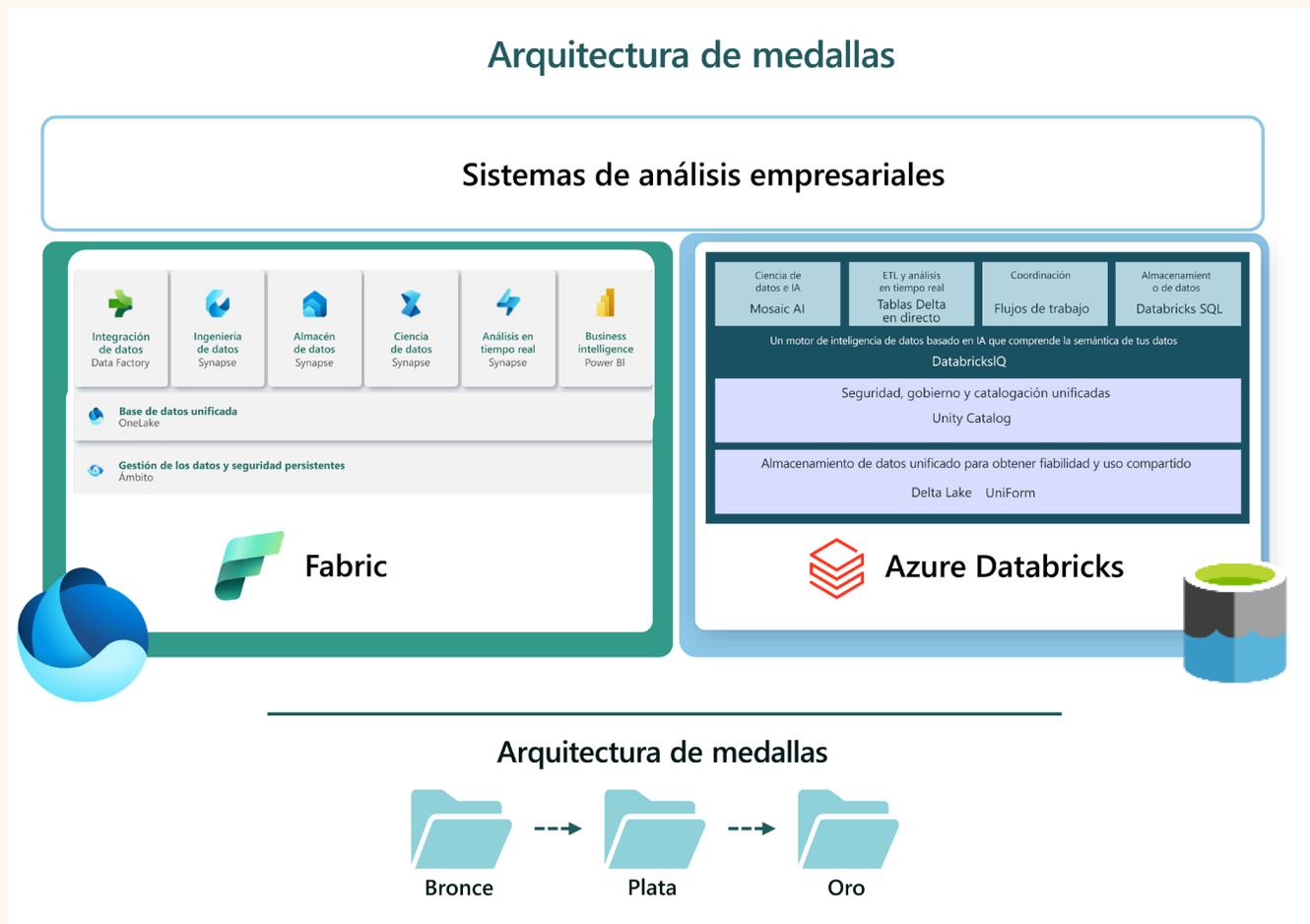


Figura 3: Arquitectura de medallas

Las tres capas de la arquitectura de medallas son:

1. **Bronce (sin procesar)**: en esta capa, los datos sin procesar se ingieren inicialmente, conservando su forma original. Actúa como una zona provisional y es crucial para captar toda la granularidad de los datos sin pérdida de fidelidad.
2. **Plata (validado)**: en esta capa, los datos de fuentes dispares se combinan, fusionan y conforman, de modo que estén listos para tareas analíticas más complejas. La capa de plata está diseñada para proporcionar una vista empresarial de las principales entidades empresariales y es fundamental para sustentar los análisis de autoservicio y las necesidades de almacenamiento de datos intermedio.
3. **Oro (enriquecido)**: en esta capa, los datos se optimizan aún más para necesidades empresariales concretas y a menudo se estructuran en formatos desnormalizados y optimizados para lectura que son aptos para consultas e informes de alto rendimiento. La capa de oro suele albergar modelos de datos que se utilizan directamente en aplicaciones de business intelligence y sistemas de toma de decisiones. Los datos se convierten en un verdadero activo empresarial en la capa de oro y aportan conocimientos valiosos y útiles.

Azure Databricks y Fabric utilizan esta arquitectura para mejorar sus ofertas analíticas y de administración de datos. En conjunto, crean un entorno robusto en el que los datos no solo fluyen sin problemas a través de cada etapa de la arquitectura de medallas, sino que también se enriquecen y se hacen más accesibles.

Integración de almacén de lago con Azure Databricks y Microsoft Fabric

El patrón de almacén de lago primero representa un enfoque transformador de la administración y el análisis de datos. Este enfoque se basa en un sistema de almacenamiento de datos por niveles, organizando los datos en capas bronce, plata y oro. Este flujo estructurado de datos facilita aplicaciones de procesamiento, análisis y machine learning más eficientes para transformar los datos sin procesar en datos optimizados para las empresas.

Azure Databricks destaca en el procesamiento de grandes volúmenes de datos con su motor de análisis basado en Spark, al gestionar eficazmente las transformaciones de datos necesarias para la transición de los datos de la capa bronce a la capa plata.

Fabric proporciona una plataforma de análisis cohesionada que se integra profundamente con Azure Databricks. Ofrece herramientas sofisticadas de administración de datos y ayuda a conectar e ingerir datos de forma fluida de varias fuentes a través de su amplio ecosistema de conectores.

Esta integración garantiza que los datos se trasladen libremente por cada capa de la arquitectura de medallas, manteniendo la integridad y la coherencia, al mismo tiempo que se minimiza la complejidad y la sobrecarga.

Base de formatos de almacenamiento de código abierto

La adopción de Apache Parquet y Delta Lake permite a OneLake y Azure Databricks optimizar los motores de Fabric y mejorar la interoperabilidad entre sus plataformas. Esta estrategia garantiza la gestión robusta de grandes conjuntos de datos, facilita el acceso sin problema a los datos a través de la arquitectura de almacén de lago y reduce las complejidades que suelen asociarse con la administración de arquitecturas de datos a gran escala:

- **Estandarización en Apache Parquet y Delta Lake:** OneLake adopta estos formatos para manejar grandes conjuntos de datos y compatibilidad con funciones transaccionales (propiedades ACID). Esta estandarización garantiza que todos los datos de todos los motores de Fabric se optimicen tanto

para el rendimiento como para la compatibilidad, con lo que se obtienen flujos de trabajo de procesamiento de datos más eficientes.

- **Optimización de motores de Fabric para Apache Parquet y Delta Lake:** al rediseñar motores de procesamiento de datos para optimizarse para estos formatos, el sistema garantiza operaciones de datos de alto rendimiento, que son fundamentales para procesar grandes volúmenes de datos de manera eficiente.
- **Interoperabilidad en todo el sistema:** la capacidad de Azure Databricks de leer cualquier artefacto de Fabric en OneLake destaca la interoperabilidad de estas tecnologías y garantiza que se pueda acceder a los datos y utilizarlos sin problemas en diferentes partes de la arquitectura del almacén de lago de datos.

El uso de formatos de almacenamiento de código abierto dentro de una arquitectura de almacén de lago maximiza la utilidad de los datos, aumenta la eficiencia operativa y reduce la complejidad que tradicionalmente se asocia a la administración de arquitecturas de datos a gran escala. Apache Parquet y Delta Lake facilitan esta tarea al garantizar que los datos se almacenen en un formato robusto y ampliamente compatible, lo que facilita a las organizaciones la integración y el análisis de datos en diversos sistemas y plataformas.

Integración de formatos de código abierto con arquitectura de medallas

El uso de formatos de código abierto como Apache Parquet y Delta Lake estandariza el almacenamiento de datos y el acceso dentro de este sistema integrado, lo que permite funciones avanzadas de administración de datos como transacciones ACID y evolución de esquemas. La potente combinación de las capacidades de procesamiento de Azure Databricks y las herramientas de administración de Fabric dentro de la arquitectura de medallas permite a las empresas abordar la arquitectura de datos de una manera escalable, eficiente y muy favorable, lo que facilita la generación de conocimientos transformadores.

Azure Databricks destaca en procesamiento y análisis de datos, utilizando Apache Spark para realizar transformaciones y análisis de datos sólidos a escala. Su integración con ADLS Gen2 permite a Azure Databricks manejar enormes conjuntos de datos de manera eficiente, con el fin de prepararlos para un posterior procesamiento analítico. Fabric amplía las capacidades de Azure Databricks al ofrecer herramientas adicionales para la administración de datos, como acceso sencillo a orígenes de datos a través de más de 200 conectores nativos y mecanismos simplificados de ingesta de datos. Esto permite a las empresas implementar una estrategia de datos integral que abarca todo, desde la ingesta hasta el análisis útil.

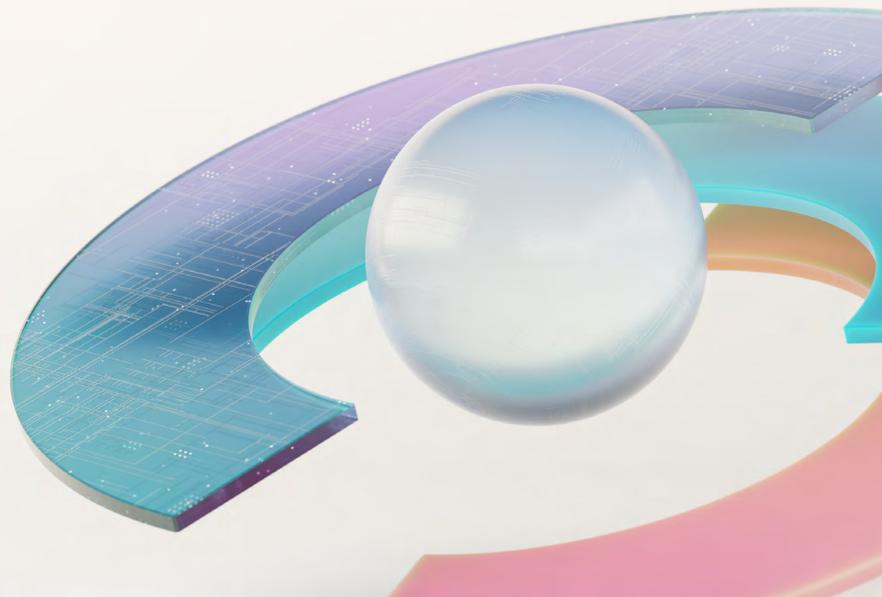
Dentro de esta arquitectura, OneLake ayuda a centralizar la administración de datos sin necesidad de trasladar físicamente los datos. Se puede acceder a datos almacenados en varias ubicaciones y analizarlos como si estuvieran en un único repositorio. Las capacidades de federación de Azure Databricks complementan aún más esto al permitir consultas en diferentes almacenes de datos, con lo que mejora la flexibilidad y el ámbito del análisis de datos.

La sinergia entre Azure Databricks y Fabric proporciona una base sólida para crear arquitecturas avanzadas de almacenes de lagos. Esta combinación simplifica la administración de datos en orígenes de datos dispares y mejora las capacidades analíticas de las organizaciones, con lo que pueden obtener conocimientos útiles de manera más eficiente y con mayor precisión. Esto permite a las empresas maximizar el valor de sus activos de datos de una manera segura y escalable.

Casos de patrones de almacén de lago primero

Al adoptar la arquitectura de medallas con Azure Databricks y Fabric, integrar los datos de almacenes de lagos con las cargas de trabajo de Fabric y utilizar los datos de OneLake con Lakehouse Federation, las organizaciones pueden habilitar un modelo de almacenamiento sofisticado y en niveles dentro de un entorno de almacén de lago. Este enfoque unificado facilita la administración y el procesamiento eficientes de los datos. Fomenta respuestas ágiles a las necesidades operativas y los conocimientos basados en datos, al mismo tiempo que mejora la escalabilidad y la flexibilidad del análisis de datos en varios sistemas de almacenamiento:

- **Arquitectura de medallas con Azure Databricks y Fabric:** este enfoque utiliza un modelo de almacenamiento por niveles dentro de un entorno de almacén de lago de datos, lo que facilita una administración y un procesamiento de datos eficientes. Mediante el uso de Azure Databricks junto con Fabric, las organizaciones pueden administrar canalizaciones de análisis de datos a gran escala de manera más eficaz.
- **Integración de datos de almacén de lago de datos de Azure Databricks con las cargas de trabajo de Fabric:** en este caso, los datos almacenados y administrados en un almacén de lago de datos de Azure Databricks se pueden utilizar directamente con las herramientas de análisis en Fabric. Esta integración permite obtener una respuesta más ágil a los conocimientos basados en datos y las necesidades operativas.
- **Uso de datos de OneLake en Azure Databricks con Lakehouse Federation:** esta configuración permite usar datos en varios sistemas de almacenamiento dentro de Azure Databricks. Al federar orígenes de datos, los usuarios pueden consultar los datos de estas fuentes como si fueran una sola entidad, lo que mejora la flexibilidad y la escalabilidad de las operaciones de análisis de datos.



Utiliza datos de almacén de lago con Azure Databricks y Microsoft Fabric

La sinergia entre Azure Databricks y Microsoft Fabric ofrece a las organizaciones una forma eficaz y potente de manejar sus cargas de trabajo de datos. Desde la ingesta y el almacenamiento hasta el análisis y la elaboración de informes, las empresas se benefician de un marco de trabajo seguro y gobernado. Esta flexibilidad permite a los equipos seleccionar la plataforma que mejor se adapte a las necesidades de sus proyectos, lo que garantiza una integración perfecta dentro del ecosistema empresarial más amplio.

Por ejemplo, un equipo de ciencia de datos que trabaje principalmente en entornos de programación de cuadernos valorará las características avanzadas de la interfaz de usuario de Azure Databricks y la flexibilidad para administrar bibliotecas de Spark avanzadas en los clústeres, y los ingenieros de IA agradecerán la capacidad nativa de ajustar los modelos en sus datos. Los analistas empresariales pueden preferir la facilidad de uso de los flujos de datos con poco código en Fabric para crear rápidamente canalizaciones que transformen los datos y creen nuevos conjuntos de datos en la capa oro del almacén de lago. Ambos equipos pueden utilizar sus herramientas preferidas para

trabajar en los mismos conjuntos de datos sin necesidad de que ninguno de los equipos realice copias de los datos en su propio entorno.

Interactuar con los datos de almacenes de lago

Los datos del almacén de lago dentro del ecosistema de Azure se suelen almacenar en ubicaciones en el cloud, que se pueden clasificar en dos tipos principales:

1. **Cuentas ADLS:** ADLS es un sistema de almacenamiento en el cloud optimizado para cargas de trabajo de análisis. Las empresas pueden crear y administrar cuentas ADLS para responder a sus necesidades de administración de datos.
2. **OneLake:** OneLake es también una cuenta de ADLS, pero a diferencia de otras cuentas, los clientes de Azure no la administran directamente. En su lugar, se crea como parte de Fabric y este mismo lo administra. No aparece en el portal de Azure y, aunque los clientes pueden interactuar con los datos que contiene, no tienen mucho control sobre la propia cuenta.

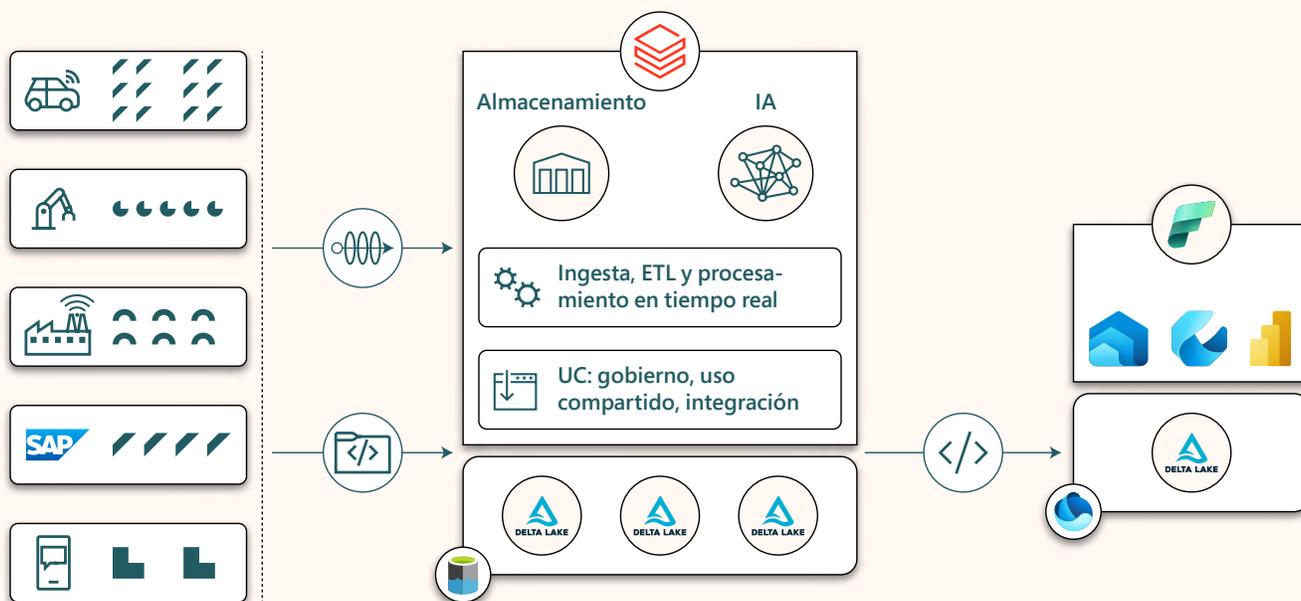


Figura 4: Enlace de cuentas ADLS a OneLake con métodos abreviados

OneLake presenta métodos abreviados para ayudar a los profesionales de datos a acceder a los datos en OneLake y varias cuentas ADLS. Los métodos abreviados permiten a los profesionales de datos enlazar los datos de Unity Catalog (con métodos abreviados de Azure Databricks) o cuentas ADLS externas a OneLake, lo que hace que aparezcan unificadas. Los usuarios pueden acceder a los datos de estas cuentas de forma fluida, sin darse cuenta de que son de diferentes orígenes. Los métodos abreviados ayudan a administrar los datos de manera eficiente virtualizando el acceso a orígenes de datos externos sin duplicación innecesaria, con lo que se obtienen unos procesos de formación e implementación de modelos de IA escalables y eficientes.

Integra Azure Databricks con Power BI para mejorar la visualización de datos

Para escenarios avanzados de visualización y paneles con un almacén de lago de datos de Azure, la mayoría de las empresas eligen Microsoft Power BI como herramienta de preferencia. Esta potente herramienta de visualización y análisis ahora se ofrece como parte de Fabric, lo que permite a las empresas integrar completamente la administración y facturación de Power BI con otros recursos de Fabric.

Azure Databricks se integra a la perfección con Power BI. Los almacenes de Databricks SQL y Unity Catalog ofrecen una solución flexible y escalable para Power BI en el almacén de lago. Los datos procesados por Azure Databricks se pueden utilizar de tres maneras:

1. **Azure Databricks Direct Publish para Power BI:** Databricks ahora puede sincronizar automáticamente las tablas, incluidas las relaciones, con los modelos semánticos de Power BI con un solo clic. Esto ayuda a los analistas a crear informes y paneles más rápido que nunca.
2. **Conector de Azure Databricks en Power BI Desktop:** Azure Databricks permite al cliente de Power BI conectarse a un clúster de Azure Databricks, que puede consultar y procesar los datos del almacén de lago de datos automáticamente y enviar los resultados a Power BI para su visualización.
3. **Modo Power BI Direct Lake:** Power BI puede usar su nuevo modo Direct Lake para leer directamente los datos de Delta Lake que se han escrito en una ubicación de almacenamiento de Azure. Pueden ser datos escritos por Azure Databricks o por Fabric, y la ubicación de almacenamiento puede ser la cuenta de OneLake o cualquier otra cuenta de ADLS.

En la sección anterior se detalló cómo utilizar Azure Databricks para procesar datos sin procesar, prepararlos para la generación de informes y luego escribirlos en el almacén del lago.

Modo Direct Lake en Power BI para leer y visualizar datos de Azure Databricks

Con el almacenamiento OneLake, los archivos se almacenan en el formato eficiente de Delta Lake. El motor VertiPaq ha optimizado estas tablas de Delta Lake, por lo que son altamente eficientes para su consumo por parte de Power BI. Esto permite a Power BI interactuar directamente con las tablas de Delta Lake almacenadas en OneLake sin necesidad de una capa de almacenamiento en caché intermedia, como Azure Analysis Services o conjuntos de datos de Power BI. Este nuevo modo de acceso, denominado modo Direct Lake, proporciona acceso a los datos en tiempo real sin necesidad de actualizar los modelos en Power BI.

Publicar directamente conjuntos de datos en áreas de trabajo de Power BI:

- Publicar desde la interfaz de usuario de Azure, sin Power BI Desktop
- Publicar esquemas enteros con relaciones de tablas (PK/FK)

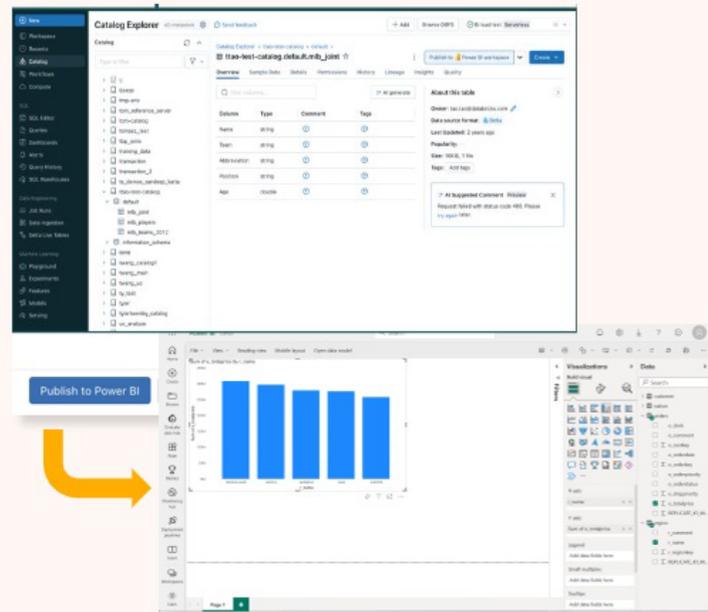


Figura 5: Integración profunda de Power BI

El conjunto de datos predeterminado incluye todas las tablas del almacén de lago, lo que permite a los usuarios establecer relaciones y aplicar diversos cambios de modelado. Estos conjuntos de datos de Unity Catalog se pueden publicar directamente en Power BI. Los usuarios pueden acceder y editar un modelo semántico publicado con el editor de modelado web al que se puede acceder a través de Power BI.

En la vista del modelo en el editor de modelado web, puedes ver si hay una conexión de Direct Lake colocando el cursor sobre los encabezados de tabla. Direct Lake también permite crear nuevos conjuntos de datos de Power BI directamente a través de la web. Este proceso garantiza el uso de Direct Lake para la conexión. Para obtener más información sobre el uso del editor web para modelos semánticos, el siguiente documento te ayudará a empezar: [Edita modelos de datos en el servicio Power BI \(preview\) - Power BI | Microsoft Learn](#).

Integra Azure Databricks con Power BI para mejorar los flujos de trabajo de datos

La integración de Azure Databricks con Power BI aporta importantes ventajas para la administración y visualización de datos, lo que mejora tanto la seguridad como el rendimiento en los flujos de trabajo de análisis de datos:

1. En primer lugar, la integración permite experiencias de visualización de datos más seguras e interactivas directamente desde el lago de datos, lo que evita la latencia y los costes asociados con los flujos de trabajo tradicionales de procesamiento de datos. Utiliza Microsoft Entra ID para la autenticación, lo que simplifica la experiencia del usuario, aumenta la seguridad y elimina la necesidad de tokens de acceso personal. Esta integración garantiza que los controles de seguridad en el nivel de lago de datos se aplican dentro de Power BI, manteniendo políticas de seguridad coherentes en todas las plataformas.¹
2. En segundo lugar, la arquitectura del almacén de lago semántica agiliza la ingesta y el almacenamiento de datos. Proporciona una capa de almacenamiento unificada que admite una amplia gama de formatos y estructuras de datos, lo que aumenta en gran medida la eficiencia del procesamiento y la transformación

de los datos. Esta configuración no solo simplifica la pila de análisis, sino que también mejora la calidad de los datos y la accesibilidad de las herramientas de BI, lo que permite crear modelos y análisis de datos más sofisticados directamente en conjuntos de datos grandes.²

3. Por último, la integración admite escenarios de análisis avanzados para simplificar la administración y el análisis de grandes volúmenes de datos. En este caso, la opción DirectQuery de Power BI desempeña una función crucial, ya que permite a los usuarios realizar análisis en tiempo real sin extraer los datos del almacén del lago. Esta capacidad es fundamental para mantener la precisión exacta en informes y paneles, proporcionando a las empresas conocimientos profundos y útiles de forma inmediata.³

Estas características convierten colectivamente a Azure Databricks y Power BI en una combinación sólida que ofrece a las empresas herramientas avanzadas para aprovechar sus datos de forma eficaz y segura. Para obtener más información sobre la integración de Power BI y almacén de lago de Azure, consulta el siguiente documento: <https://learn.microsoft.com/fabric/get-started/directlake-overview>

¹ [Potencia tu BI con Microsoft Power BI y Lakehouse en Azure Databricks: parte 1 - Fundamentos - Microsoft Community Hub](#)

² [Impulso con Power BI y Lakehouse en Azure Databricks: parte 3 - Ajuste de Azure Databricks SQL - Microsoft Community Hub](#)

³ <https://docs.databricks.com/partners/bi/power-bi.html>

Utiliza Data Activator para alertar sobre los cambios en los datos de Azure Databricks a través de Power BI

Un caso común en las empresas es que quieren recibir alertas si determinadas métricas superan ciertos umbrales. Por ejemplo, es posible que quieran saber si se produce un aumento repentino e inesperado de las ventas de un artículo determinado, o bien si el volumen de las transacciones se ha desplomado por debajo del rango normal, lo que indica un posible problema en la canalización de transacciones.

Estos casos se pueden gestionar con una nueva característica de Fabric denominada Data Activator. Esta herramienta sin código supervisa los datos de un informe de Power BI y adopta automáticamente acciones si los datos coinciden con determinados patrones o alcanzan los umbrales especificados. Cuando se producen estos eventos, Data Activator puede emprender una acción como alertar a un usuario o iniciar un flujo de trabajo de Power Automate.

Para habilitar Data Activator, sigue la documentación oficial aquí: <https://learn.microsoft.com/fabric/data-activator/>

Para crear una alerta con Data Activator cuando la temperatura de un congelador descienda por debajo de 30 °F en un informe de Power BI, sigue estos pasos para supervisar las temperaturas del congelador en un área de trabajo de Fabric:

1. Confirma que tu informe de Power BI, que incluye datos de temperatura del congelador, se publique online en un área de trabajo de Fabric equipada con capacidad Premium.
2. Elige el elemento visual de temperatura:
 - a. **Accede al informe:** abre el informe específico de Power BI que realiza un seguimiento de las temperaturas del congelador.
 - b. **Selecciona el elemento visual correspondiente:** encuentra el elemento visual que muestra las temperaturas del congelador.
3. Haz clic en los puntos suspensivos (...) en la esquina superior derecha del elemento visual de temperatura y selecciona **Set Alert** (Ajustar alerta) o utiliza el botón **Set Alert** (Ajustar alerta) que se encuentra en la barra de herramientas de Power BI.
4. En el panel **Set Alert** (Ajustar alerta), especifica cómo deseas recibir alertas (por correo electrónico o por Teams). Si el elemento visual incluye varios congeladores (dimensiones), utiliza la lista desplegable **For each** (Para cada uno) para seleccionar la dimensión concreta (congelador) que desea supervisar.
5. Define la condición de alerta, por ejemplo, cuando la temperatura descienda por debajo de 30 °F. Data Activator supervisará la temperatura y te notificará cuando se cumpla esta condición.
6. Decide dónde guardar el desencadenador de Data Activator en Power BI. Puedes agregarlo a un elemento de desencadenante existente o crear uno nuevo.

7. Haz clic en **Create alert** (Crear alerta) para finalizar el desencadenador de Data Activator. Opcionalmente, puedes eliminar la selección de **Start my alert** (Iniciar mi alerta) si prefieres editar el desencadenador en Data Activator antes de activarlo.

Siguiendo estos pasos, has configurado correctamente una alerta en Data Activator para que te notifique cuando la temperatura de un congelador supervisado descienda por debajo de 30 °F, lo que te permite adoptar medidas inmediatas, si es necesario. Una vez completadas estas actualizaciones de datos, debes recibir la alerta de Data Activator que se ha configurado.

Utiliza Lakehouse Monitoring with Alerts para alertar sobre cambios en Azure Databricks

Con frecuencia, las empresas necesitan alertas cuando las métricas de calidad de los datos superan ciertos umbrales. Por ejemplo, es posible que quieran saber si se produce un aumento repentino e inesperado en el número de valores que faltan en un campo determinado, lo que indica un posible problema en la canalización de transacciones, o bien si la calidad de las predicciones de un modelo de machine learning ha disminuido, lo que indica la necesidad de volver a entrenar el modelo con datos más recientes.

Estos casos se pueden gestionar con una característica de Azure Databricks denominada Lakehouse Monitoring with Alerts. Esta herramienta sin código

supervisa la calidad de los datos en Unity Catalog y adopta automáticamente acciones si los datos cumplen ciertas condiciones o superan los umbrales. Cuando se producen estos eventos, las alertas emprenden una acción especificada, como enviar una notificación por correo electrónico, Slack o Teams. La alerta también puede llamar a una acción de webhook, lo que permite a los usuarios crear flujos de trabajo ampliables y personalizados basados en los cambios en los datos.

Un monitor es un proceso que se ejecuta según una programación específica para comprobar la calidad de los datos de una tabla determinada. Cuando un usuario crea un monitor, calcula las métricas de calidad de los datos de la tabla y almacena los valores actuales en una tabla del sistema independiente. Cada vez que se ejecuta el monitor, vuelve a calcular las métricas de calidad y las compara con los valores originales. Si la calidad ha descendido, se emitirá una alerta. Para obtener información detallada sobre cómo se puede crear un monitor, consulta el siguiente documento: <https://docs.databricks.com/lakehouse-monitoring/create-monitor-ui.html>

Si un monitor detecta que la calidad de los datos de la tabla ha descendido, emitirá la alerta especificada. Esto se puede utilizar para enviar una notificación a los equipos de ingeniería de datos para que puedan investigar más. Para obtener información detallada sobre cómo se pueden configurar estas alertas, consulta el siguiente documento: <https://docs.databricks.com/lakehouse-monitoring/monitor-alerts.html>

Mejor juntos: Azure Databricks, Unity Catalog y Microsoft Fabric Purview

A medida que aumenta la demanda de análisis y las plataformas de datos evolucionan para convertirse en sistemas más complejos, es fundamental la gestión de la plataforma (administración de la disponibilidad, usabilidad, integridad y seguridad de los datos). En una arquitectura de almacén de lago, la gestión de los datos ayuda a garantizar que los datos se clasifiquen, cataloguen y administren correctamente. Mediante la implementación de una gestión de los datos eficaz, las organizaciones pueden administrar sus datos correctamente y utilizarlos para impulsar el valor empresarial.

La gestión eficaz de los datos en una arquitectura de almacén de lago requiere la implementación de políticas, procedimientos y normas para la administración de datos. Esto incluye definir la propiedad y dirección de los datos, establecer estándares de calidad de los datos e implementar medidas de seguridad y cumplimiento de los datos. Para proporcionar estas capacidades cruciales de gestión de los datos, tanto Azure Databricks como Microsoft Fabric ofrecen características modernas y eficaces.

Unity Catalog en Azure Databricks

Azure Databricks incluye Unity Catalog, que proporciona un control de acceso preciso detallado y centralizado para las ubicaciones de almacenamiento de los datos de una organización, auditoría del acceso a los datos y seguimiento del linaje desde la ingesta a todas las cargas de trabajo de datos, y Azure Databricks proporciona controles de acceso a nivel de columna y de fila y herramientas de detección de datos. Ahora también incluye tablas del sistema, que proporcionan una manera directa de consultar datos de auditoría, datos de facturación y linaje. Además, Unity Catalog está respaldado por funcionalidades de IA para documentar automáticamente las tablas y columnas, facilitar la búsqueda semántica y ayudar a mostrar productos de datos relacionados.

Microsoft Purview para controlar Microsoft Fabric

Fabric se integra con Microsoft Purview para la gestión de los datos, la protección de la información y la prevención de la pérdida de datos. Las características de protección de la información permiten a las empresas detectar, clasificar y proteger los datos almacenados en el almacén de lago y aplicar etiquetas de confidencialidad. La prevención de la pérdida de datos utiliza políticas para detectar cuándo se cargan datos confidenciales en modelos semánticos de Power BI u otros activos de Fabric admitidos. También puede ayudar a detectar datos confidenciales habituales. Fabric también incluye herramientas para descubrir el linaje de los datos, de manera que se pueda realizar un seguimiento de los datos a través del proceso analítico a medida que se trasladan desde su origen, por las diversas transformaciones y hasta los diversos modelos de informes.

Microsoft Purview y Unity Catalog para optimizar la gestión de los datos

Microsoft Purview y Unity Catalog son dos potentes herramientas diseñadas para mejorar la gestión y la administración de los datos en entornos de cloud, sobre todo para los usuarios de los amplios servicios en el cloud que ofrece por Microsoft.

Las amplias funcionalidades de gestión de Microsoft Purview pueden extenderse al entorno de Azure Databricks, donde Unity Catalog aplica medidas concretas de gestión y seguridad a las áreas de trabajo de Azure Databricks. Esta integración permite a las organizaciones mantener una estrategia de gestión coherente en todas las plataformas, lo que mejora la seguridad y la eficiencia operativa. Las organizaciones pueden garantizar que las políticas de datos se apliquen de manera uniforme, que el linaje de datos sea claro y auditable, y que se cumplan todos los requisitos de cumplimiento normativo en todo su patrimonio de datos.

Unity Catalog ofrece una solución de gestión sofisticada y centralizada para administrar una variedad de activos de datos dentro de la plataforma de almacén de lago de Azure Databricks. Se integra perfectamente con Azure para proporcionar funcionalidades de gestión detalladas, como el control de accesos, la auditoría y el linaje de datos. Unity Catalog simplifica la administración de datos en varias áreas de trabajo de Azure Databricks, lo que permite a las organizaciones aplicar políticas coherentes de seguridad y cumplimiento en sus activos de datos, ya sean archivos, tablas o modelos de machine learning.

Unity Catalog proporciona un único punto de control para las políticas de acceso a los datos, que se aplican de manera uniforme en todas las áreas de trabajo. Esto garantiza que la gestión de los datos no solo esté centralizada, sino que también se integre profundamente en el entorno de Azure Databricks, lo que mejora la seguridad y la gestión. Además, Unity Catalog admite la detección integral de los datos, lo que facilita a los usuarios la búsqueda y el acceso a los datos que necesitan mientras cumplen con los controles y las políticas de acceso definidos. Este enfoque unificado ayuda a agilizar las operaciones y reduce la complejidad que suele asociarse a la administración de entornos de datos grandes y diversos.

Microsoft Purview permite a las empresas mantener el control sobre sus datos a través de Fabric, con lo que se logra una integración y administración perfectas de los datos de varios orígenes hasta informes detallados. Junto con un conjunto de herramientas para proteger los datos confidenciales en distintos entornos, Microsoft Purview proporciona funcionalidades como la detección, clasificación y protección de datos confidenciales mediante etiquetas de confidencialidad. También facilita estrategias integrales de auditoría y prevención de pérdida de datos adaptadas específicamente a entornos complejos como los modelos semánticos de Power BI.

Prácticas recomendadas

La integración de Azure Databricks con Microsoft Purview se centra en maximizar la gestión y la seguridad de los datos en los entornos de Azure Databricks. Entre las principales prácticas recomendadas para esta integración se incluyen las siguientes:

- **Acceso seguro a datos críticos:** Microsoft Purview se puede utilizar para detectar y clasificar automáticamente datos dentro de Azure Databricks, visualizar el linaje de los datos y administrar los controles de acceso de forma eficaz. Esto garantiza que solo el personal autorizado pueda acceder a datos confidenciales o críticos y que todas las políticas de datos se apliquen de manera coherente en todos los servicios de Azure.
- **Utiliza dos conectores independientes para administrar metadatos:** Microsoft Purview ofrece dos conectores independientes para Azure Databricks. La mayoría de las empresas utilizarán el conector Azure Databricks Unity Catalog porque Unity Catalog aporta muchas de las características modernas de Azure Databricks. Sin embargo, para los clientes que aún no han migrado a Unity Catalog y siguen usando Hive para administrar sus metadatos, Microsoft Purview también tiene un conector de Azure Databricks Hive Metastore que se puede utilizar.

- **Utiliza conjuntos de reglas personalizados:** las empresas pueden usar Microsoft Purview para analizar catálogos, esquemas, tablas y vistas. Como práctica recomendada, las empresas deben utilizar conjuntos de reglas personalizadas, además de los conjuntos de reglas proporcionados por Microsoft Purview. La creación de un conjunto de reglas personalizadas para diferentes regiones del mundo puede acelerar el proceso de análisis utilizando solo las reglas de clasificación necesarias en una región concreta.
- **Indica la confidencialidad de los datos con herramientas de etiquetado:** las herramientas de etiquetado de Microsoft Purview se pueden utilizar en los datos de Unity Catalog para indicar la confidencialidad de los archivos y las columnas de datos. Estas etiquetas viajan con los datos y las pueden utilizar otras herramientas del ecosistema de datos de Microsoft, como SharePoint y Power BI, para aplicar automáticamente políticas de gestión de datos.

Las funcionalidades combinadas de Microsoft Purview, la seguridad de Azure en OneLake y Unity Catalog respaldan una estrategia de gestión de datos resiliente y ágil, para que las empresas puedan utilizar sus activos de datos eficazmente en un panorama digital.

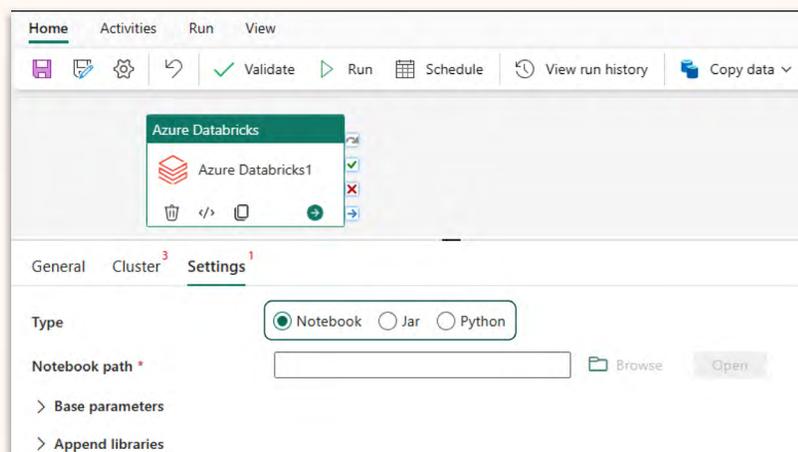


Actividad de Data Factory y Azure Databricks en Microsoft Fabric

La actividad de Azure Databricks en Microsoft Fabric representa una evolución importante del procesamiento de datos en los entornos de cloud, integrando las amplias funcionalidades de Azure Data Factory en un marco más unificado y robusto. Con la nueva actividad de Azure Databricks, los usuarios pueden crear y administrar fácilmente canalizaciones de datos en Fabric, incorporando tareas de análisis y procesamiento sofisticadas directamente en sus flujos de trabajo.

Los usuarios pueden configurar clústeres de Azure Databricks que se utilizan para el procesamiento de datos directamente en Fabric, como la funcionalidad que se ofrece en Azure Data Factory. Esto incluye la capacidad de configurar instancias puntuales de Azure para acceder a la capacidad de computación de Azure no utilizada a costes reducidos y especificar políticas de clúster para garantizar que las configuraciones del clúster cumplan los estándares y requisitos de la organización.

Una actividad que abarca los tres tipos de trabajo: Notebook, Jar, Python



Compatibilidad con Unity Catalog e integración con Policy ID

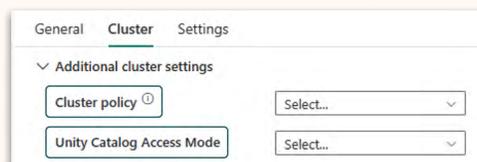


Figura 6: integración perfecta entre Azure Databricks y Data Factory con Microsoft Fabric

Además, la integración mejorada de Azure Databricks presenta varias características nuevas. Estas son:

- **Configura el modo de acceso a Unity Catalog:** los usuarios pueden configurar el modo de acceso de Unity Catalog, lo que mejora la gestión y la seguridad al administrar los permisos de forma más meticulosa.
- **Ejecuta varias tareas en una sola actividad de Databricks:** los usuarios pueden ejecutar varias tareas, como cuadernos, JARs y scripts de Python, dentro de una sola actividad de Databricks, con lo que se agiliza el proceso y se reduce la complejidad previamente asociada a la administración de varios tipos de trabajos de datos.
- **Supervisa ejecuciones de cuadernos en tiempo real:** durante la operación, los usuarios pueden iniciar una canalización de datos y supervisar inmediatamente su ejecución. El sistema que proporciona enlaces directos a los resultados en la instancia de Azure Databricks. Esta capacidad de supervisión en tiempo real permite a los usuarios realizar un seguimiento de los detalles de la ejecución del cuaderno, incluido el rendimiento del clúster y la eficiencia computacional, directamente desde Fabric.

Esta integración no solo incorpora las funcionalidades existentes de Azure Databricks desde Azure Data Factory en Fabric, sino que también presenta nuevas funciones, como la política de clústeres y la compatibilidad con Unity Catalog, lo que mejora la experiencia general de administración de datos y análisis.



Mejora las funcionalidades organizativas con IA generativa

Los modelos avanzados de IA han cambiado el panorama tecnológico. Las empresas están intentando sacar el máximo partido de sus datos y utilizar la IA para ampliar sus funcionalidades empresariales. Esto incluye el uso de la IA generativa para crear su almacén de lago de datos en Azure.

En la siguiente sección se detallará cómo describir la arquitectura que deseas en ChatGPT y hacer que genere código que se pueda implementar en Azure.

Genera arquitecturas y código con Azure OpenAI

Mediante la ingeniería de instrucciones, puedes generar planes de arquitectura y código utilizando el mismo proceso que se utilizaría para resumir un artículo técnico.

Introducción de solicitudes

```
Quiero que actúes como ingeniero de datos. Te proporcionaré detalles sobre un caso en el que un cliente necesita utilizar los datos almacenados en ADLS Gen2. Azure Databricks y Microsoft Fabric accederán a ellos para cargar datos para modelos de machine learning. Tu trabajo consistirá en crear una arquitectura que responda a esta necesidad.
```

Respuesta de Azure OpenAI

```
Para abordar el caso en el que un cliente necesita utilizar los datos almacenados en Azure Data Lake Storage Gen2 (ADLS Gen2), a los que tienen acceso Azure Databricks y Microsoft Fabric para cargar datos en modelos de machine learning, diseñaremos una arquitectura que aproveche eficazmente estos servicios. Esta arquitectura tiene como objetivo optimizar el procesamiento de datos, mejorar la gestión de los datos y garantizar operaciones de machine learning escalables.
```

Información general de la arquitectura

- Capa de almacenamiento de datos: Azure Data Lake Storage Gen2: actúa como el repositorio central para almacenar datos sin procesar, datos procesados y artefactos de machine learning. ADLS Gen2 está optimizado para casos de análisis a gran escala y admite espacios de nombres jerárquicos, lo que simplifica la administración de datos.
- Capa de procesamiento y análisis de datos: Azure Databricks: sirve como la plataforma principal para el procesamiento, análisis y desarrollo de modelos de machine learning. Azure Databricks proporciona un entorno de colaboración con compatibilidad con varios lenguajes de programación y se integra perfectamente con ADLS Gen2 para obtener acceso directo a los datos.
- Capa de gestión y administración de datos: Microsoft Fabric: se utiliza para administrar y gestionar los datos en todo el entorno de datos. Microsoft Fabric permite crear un modelo de gestión de los datos unificado que abarca los datos de ADLS Gen2 y los datos procesados o generados por Azure Databricks.
- Operativización de machine learning: Azure Machine Learning: se integra con Azure Databricks para implementar y administrar modelos de machine learning. Este servicio proporciona herramientas para el control de versiones, la supervisión y el servicio de modelos, lo que facilita la operativización de los modelos de machine learning.

Flujo de trabajo detallado

1. Ingesta de datos: los datos se ingieren en ADLS Gen2 desde diversos orígenes, incluidas bases de datos estructuradas, dispositivos de IoT, archivos de registro, etc. Estos datos se almacenan en formato sin procesar dentro de una estructura jerárquica del sistema de archivos.

Nota

El resto del contenido generado por Azure OpenAI se puede encontrar en el *apéndice*.

Es posible que las indicaciones y respuestas indicadas aquí no reflejen la experiencia exacta.

Genera y corrige código con Databricks Assistant

Databricks Assistant es una herramienta basada en IA integrada dentro del entorno de Azure Databricks, diseñada para mejorar la productividad de programación, que ayuda a los usuarios con la generación de código, la resolución de errores y la documentación directamente en cuadernos de Databricks. Utiliza las funcionalidades de los servicios de IA de Azure, mejorando la forma en que los desarrolladores interactúan con los datos y el código dentro de la plataforma.

Databricks Assistant actúa como una herramienta sólida que ayuda a los desarrolladores automatizando tareas rutinarias, optimizando el código, explicando funcionalidades y solucionando problemas, todo dentro del área de trabajo de Azure Databricks. Esto no solo acelera el proceso de desarrollo, sino que también ayuda a mantener un alto estándar de calidad del código y la documentación, lo que lo convierte en un activo en la ingeniería de datos y los flujos de trabajo de análisis.

Databricks Assistant mejora la productividad en la ciencia e ingeniería de datos mediante el uso de IA para ayudar con la generación de código, la resolución de errores y la documentación dentro del entorno de Azure Databricks de las siguientes maneras:

- **Generación de código:** Databricks Assistant simplifica la programación al permitir a los usuarios introducir sus requisitos en lenguaje natural. Puede generar consultas SQL ejecutables o transformar código de un lenguaje a otro, como convertir código pandas de Python en PySpark. Esta característica acelera el desarrollo y reduce los errores de codificación manuales.
- **Resolución de errores:** Databricks Assistant puede identificar y aclarar rápidamente los errores de codificación y ofrecer soluciones mediante la generación de fragmentos de código correctivo. Esto resulta de gran utilidad tanto para los programadores principiantes como para los más experimentados, ya que proporciona soluciones inmediatas a problemas comunes de sintaxis y tiempo de ejecución, con lo que se minimiza el tiempo de inactividad.
- **Documentación:** ayuda a documentar el código generando automáticamente comentarios que explican la funcionalidad de los bloques de código, respaldando el mantenimiento de bases de código limpios y comprensibles, fundamentales para la sostenibilidad de los proyectos a largo plazo y la colaboración en equipo.

- **Ayuda y aprendizaje contextuales:** Databricks Assistant ayuda en el contexto al comprender el entorno del usuario, incluidas tablas, esquemas y consultas anteriores que se utilizan con frecuencia. Utiliza este contexto para proporcionar respuestas precisas y fragmentos de código adaptados, lo que mejora el soporte específico del proyecto.
- **Visualizaciones y paneles:** en entornos de visualización como Lakeview, Databricks Assistant puede generar visualizaciones de datos basadas en indicaciones del usuario, con lo que se puede crear e iterar rápidamente pantallas visuales sin necesidad de una experiencia técnica profunda.
- **Procesos de ingesta de datos y ETL:** Databricks Assistant optimiza la ingesta de datos y las tareas ETL automatizando la generación y optimización de código, acelerando la configuración y la ejecución de canalizaciones de datos.
- **Seguridad y cumplimiento:** Azure Databricks garantiza que las interacciones con Databricks Assistant sean seguras, respetando los permisos de los usuarios y las políticas de gestión de los datos, lo que lo hace apto para su uso en entornos sensibles y regulados.
- **Integración y accesibilidad:** se puede acceder a Databricks Assistant desde cuadernos, editores de SQL y editores de archivos, lo que lo convierte en una herramienta versátil para distintas tareas de datos.
- **Comentarios e iteración:** los usuarios pueden proporcionar comentarios directamente a través de la plataforma, lo que ayuda a mejorar la precisión

y la funcionalidad de Databricks Assistant. Esto garantiza que la herramienta evolucione en respuesta a las necesidades y los desafíos de los usuarios.

Las áreas de trabajo de Azure Databricks se pueden aprovisionar con una prueba gratuita de 14 días que incluye acceso a Assistant, para que los usuarios potenciales puedan evaluar sus funcionalidades e integrarlas en sus flujos de trabajo sin una inversión inicial.

A continuación, se muestran algunos ejemplos de cómo Databricks Assistant se puede utilizar en diferentes casos para mejorar la productividad y simplificar las tareas:

1. Generación de código

Escenario: un usuario necesita extraer y resumir los datos de ventas por región de un DataFrame.

Entrada del usuario:

```
Genera una consulta SQL para
sumar todas las ventas por
región desde la tabla sales_
data.
```

Respuesta:

```
sqlCopy code

SELECT region, SUM(sales) AS
total_sales FROM sales_data
GROUP BY region;
```

Utilidad: ayuda al usuario a obtener rápidamente el código que necesita sin tener que escribir manualmente la consulta SQL.

2. Corrección de errores

Escenario: un usuario ha escrito código de PySpark que falla debido a un error de sintaxis.

Entrada del usuario:

```
Este código no funciona. ¿Me puedes ayudar a corregirlo?
```

Diagnóstico y corrección: Databricks Assistant identifica que falta una coma en una operación de DataFrame y sugiere el código corregido con un cambio resaltado.

Utilidad: el usuario recibe sugerencias inmediatas de corrección y comentarios, lo que acelera la solución de problemas y reduce la frustración.

3. Documentación del código

Escenario: un desarrollador quiere agregar comentarios a una función compleja para mejorar la legibilidad del código.

Entrada del usuario:

```
¿Puedes documentar esta función explicando cada paso?
```

Respuesta: Databricks Assistant añade comentarios antes de cada línea o bloque de código importante, explicando lo que hace, como, por ejemplo, inicializar variables, gestionar errores y flujos lógicos.

Utilidad: garantiza que el código sea comprensible para referencias futuras o para otros miembros del equipo, lo que mejora el mantenimiento.

Estos ejemplos ilustran las ventajas prácticas de Databricks Assistant en entornos de desarrollo reales, con lo que se agiliza el proceso de programación, se simplifica la resolución de errores y se garantiza una documentación exhaustiva.

Nota

Es posible que las indicaciones y respuestas indicadas aquí no reflejen la experiencia exacta.



Explora casos de uso reales con ejemplos prácticos

En ejemplos anteriores, utilizaste código de Python para leer los datos y agregarlos para responder a algunas preguntas empresariales. En esta sección se examina una alternativa al código de Python para leer datos y cómo se puede usar la IA para que los usuarios empresariales puedan consultar datos del almacén de lago usando el inglés en lugar de un lenguaje de consulta.

Usa el SDK en inglés para Spark para escribir consultas en Azure Databricks y Fabric

Para utilizar el SDK en inglés para Apache Spark, se deben cumplir los siguientes requisitos:

Nota

Azure Databricks recomienda usar GPT-4.

1. **Instala el paquete de SDK en inglés:** comienza añadiendo el SDK a tu entorno. Utiliza el comando `%pip install pyspark-ai --upgrade` en tu cuaderno para asegurarte de que tienes la versión más reciente.
2. **Reinicia el kernel de Python:** después de la instalación, debes reiniciar el kernel de Python para aplicar las actualizaciones. Ejecuta `dbutils.library.restartPython()` en una nueva celda para restablecer el entorno.
3. **Establece la clave de API de OpenAI:** tu clave de API de OpenAI es necesaria para la autenticación. Implementala estableciendo una variable de entorno con el código de Python

```
os.environ['OPENAI_API_KEY'] = '<tu-clave-de-api-de-openai>',
```

 sustituyendo `<tu-clave-de-api-de-openai>` por tu clave de API.
4. **Activa el SDK:** para usar el SDK, actívalo dentro de tu cuaderno. Esto implica inicializar el SDK con tu modelo de lenguaje preferido (como GPT-4) y, a continuación, activarlo para empezar a interpretar consultas en inglés.
5. **Crea un DataFrame:** utiliza consultas SQL dentro del cuaderno para recuperar datos desde el área de trabajo de Azure Databricks y guardarlos como DataFrame. Este DataFrame será la base para tus consultas en inglés.

6. **Consulta en inglés:** por último, consulta el DataFrame haciendo preguntas en inglés sencillo. El SDK interpreta estas preguntas y ejecuta las consultas SQL correspondientes, devolviendo los resultados directamente a tu cuaderno.

Un ejemplo de consulta en inglés con el SDK en inglés para Apache Spark podría ser algo así:

```
¿Cuál fue la distancia media de viaje por día durante el mes de enero de 2016? Imprime las medias a la décima más cercana.
```

Esta consulta demuestra cómo se puede utilizar el idioma normal para realizar actividades de análisis de datos, como calcular los promedios de un conjunto de datos con el SDK en inglés, lo que permite a Apache Spark interpretar y ejecutar instrucciones en inglés.

Otro ejemplo de consulta con inglés para el SDK en inglés para Apache Spark podría ser el siguiente:

```
Muéstrame los ingresos totales de cada categoría de productos en el último trimestre.
```

Este tipo de consulta ilustra cómo los usuarios pueden solicitar métricas financieras concretas, como los ingresos

totales, desglosadas por categorías durante un periodo definido, como el último trimestre, utilizando un lenguaje natural. Este enfoque simplifica las tareas complejas de análisis de datos en preguntas sencillas en inglés.

Creación de un cuaderno en Microsoft Fabric

Los cuadernos de Fabric son una herramienta clave para crear trabajos de Apache Spark y realizar experimentos de machine learning. Gracias a su compatibilidad con visualizaciones avanzadas e integración de texto Markdown, ofrece una plataforma interactiva basada en web muy popular entre científicos de datos e ingenieros para la codificación. Los científicos de datos se basan en estos cuadernos para desarrollar e implementar modelos de machine learning, incluidas las fases de experimentación, seguimiento de modelos e implementación. Los cuadernos de Fabric ofrecen:

- Usabilidad inmediata sin necesidad de configuración
- Una interfaz intuitiva con poco código para la exploración y el procesamiento de datos
- Mayor seguridad de los datos mediante características integradas de nivel empresarial
- La capacidad de analizar datos en varios formatos (incluidos CSV, TXT, JSON, Parquet y Delta Lake) con las sólidas capacidades de Spark

Creación de cuadernos

Al crear un cuaderno, los usuarios tienen dos opciones: crear uno nuevo o importar uno existente. Las organizaciones pueden crear un nuevo cuaderno siguiendo el flujo de trabajo de creación de elementos de Fabric que ya conocen:

1. Inicia un nuevo cuaderno directamente desde **Data Engineering** de Fabric o la página de inicio de **Data Science**, o bien a través de la opción **New** (Nuevo) del área de trabajo.
2. Elige **Import Notebook** (Importar cuaderno) en la misma ventana para importar el cuaderno existente, como un archivo de cuaderno de Azure Databricks.
3. Una vez que tienes un cuaderno abierto, puedes agregarle código para escribir datos en OneLake.

Trabajar con datos en OneLake es sencillo y no implica una configuración compleja para acceder a los datos.

Carga de datos en OneLake mediante un cuaderno de ingeniería de datos de Microsoft Fabric

```
from pyspark.sql import SparkSession

# Initialize Spark session (assuming it's not already initialized)

spark = SparkSession.builder.appName("ParkDataImport").getOrCreate()

# URL to the CSV file

data_url = "https://www.dropbox.com/s/268uogek0mcypn9/park-data.csv?raw=1"

# Read the CSV data directly into a Spark DataFrame

df = spark.read.option("header", "true").csv(data_url)

# Assuming csv_table_name, parquet_table_name, and delta_table_name are
defined elsewhere in your code

csv_table_name = "park_data_csv"

parquet_table_name = "park_data_parquet"

delta_table_name = "park_data_delta"
```

```
# Save dataframe as CSV files to Files section of the default Lakehouse

df.write.mode("overwrite").format("csv").save("Files/" + csv_table_
name)

# Save dataframe as Parquet files to Files section of the default
Lakehouse

df.write.mode("overwrite").format("parquet").save("Files/" + parquet_
table_name)

# Save dataframe as a delta lake, parquet table to Tables section of
the default Lakehouse

df.write.mode("overwrite").format("delta").saveAsTable(delta_table_
name)

# Save the dataframe as a delta lake, appending the data to an existing
table

# Make sure the table exists and the schema matches to avoid errors

df.write.mode("append").format("delta").saveAsTable(delta_table_name)
```

Lectura y análisis de datos

Una vez que se hayan cargado correctamente los datos, intenta leerlos y analizarlos:

```
# Basic Data Analysis

# Count of animal sightings by type (excluding squirrels)

animal_sightings = spark.sql("""

SELECT Animal_Type, COUNT(*) as Total_Sightings

FROM park_data_view

WHERE Animal_Type != 'Squirrel'

GROUP BY Animal_Type

ORDER BY Total_Sightings DESC
```

```
""")
animal_sightings.show()

# Average temperature and most common weather conditions
avg_temp = spark.sql("""
SELECT AVG(Temperature) as Average_Temperature
FROM park_data_view
""")
avg_temp.show()

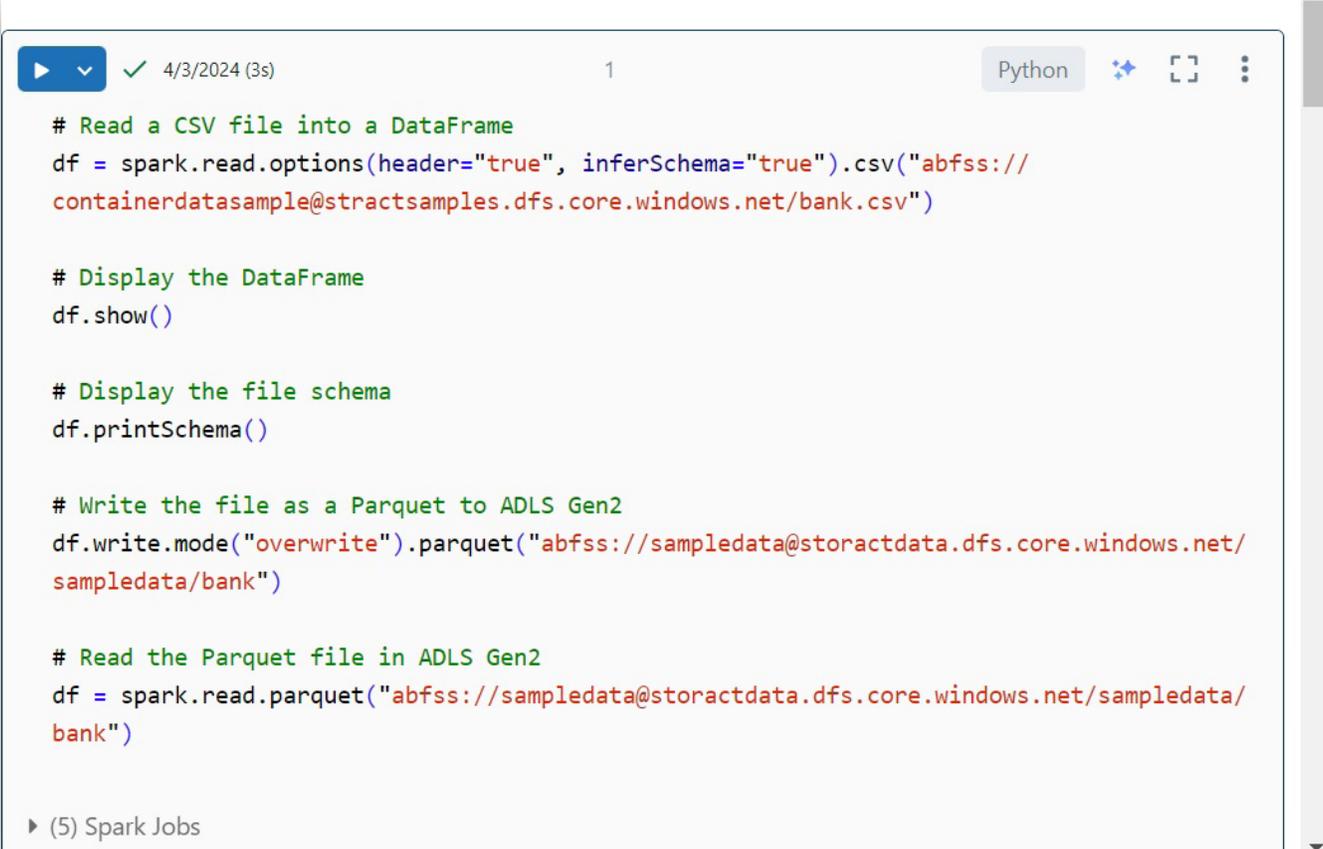
common_weather = spark.sql("""
SELECT Weather, COUNT(*) as Frequency
FROM park_data_view
GROUP BY Weather
ORDER BY Frequency DESC
LIMIT 5
""")
common_weather.show()

# Total count of squirrel sightings
squirrel_sightings = spark.sql("""
SELECT COUNT(*) as Total_Squirrel_Sightings
FROM park_data_view
WHERE Animal_Type = 'Squirrel'
""")
squirrel_sightings.show()
```

Creación y modificación de una tabla Delta de Parquet en Azure Databricks con cambios reflejados en fabric

Azure Databricks y Fabric proporcionan un entorno de almacén de lago de datos que permite a las empresas acceder y analizar sus datos simultáneamente, utilizando diferentes herramientas. Esto permite una amplia gama de actividades de procesamiento de datos en el mismo conjunto de datos, lo que facilita a las organizaciones la administración y la derivación de conocimientos de su información de manera eficiente.

1. Abre tu área de trabajo de Azure Databricks con el navegador que prefieras e inicia un nuevo cuaderno de Azure Databricks.



```
▶ 4/3/2024 (3s) 1 Python
```

```
# Read a CSV file into a DataFrame
df = spark.read.options(header="true", inferSchema="true").csv("abfss://
containerdatasample@stractsamples.dfs.core.windows.net/bank.csv")

# Display the DataFrame
df.show()

# Display the file schema
df.printSchema()

# Write the file as a Parquet to ADLS Gen2
df.write.mode("overwrite").parquet("abfss://sampledata@storactdata.dfs.core.windows.net/
sampledata/bank")

# Read the Parquet file in ADLS Gen2
df = spark.read.parquet("abfss://sampledata@storactdata.dfs.core.windows.net/sampledata/
bank")
```

▶ (5) Spark Jobs

Figura 7: Cuaderno de ejemplo

2. Copia y pega el siguiente script en tu nuevo cuaderno. A continuación, ejecuta el siguiente script de Python en tu cuaderno para crear una tabla Delta dentro de tu cuenta de ADLS Gen2. Este script lee algunos datos de Parquet de ejemplo y luego los escribe como una tabla Delta en tu cuenta ADLS:

```
#python

# Adjust the file path to point to your sample parquet data using the
following format:

"abfss://<storage name>@<container name>.dfs.core.windows.
net/<filepath>"

# The line below reads Parquet files from your ADLS account

df = spark.read.format('Parquet').load("abfss://datasetsv1@olsdemo.dfs.
core.windows.net/demo/full/dimension_city/")

#This line writes the read data as Delta tables back into your ADLS
account

df.write.mode("overwrite").format("delta").save("abfss://datasetsv1@
olsdemo.dfs.core.windows.net/demo/adb_dim_city_delta/")
```

Y, por supuesto, Azure Databricks también puede leer los datos en la cuenta de ADLS.

3. Azure Databricks también puede modificar los mismos conjuntos de datos que se crearon al principio anteriormente con Fabric. Para ver esto en acción, anexa algunas filas nuevas a las tablas de Delta Lake que creaste en OneLake:

```
# Import the necessary libraries

from pyspark.sql import SparkSession

from pyspark.sql.functions import lit

# Initialize a Spark session

spark = SparkSession.builder.appName("AppendToDeltaTable").
getOrCreate()
```

```
# Define the path to your Delta Lake table in OneLake

# Replace '<your-delta-table-path>' with the actual path to your Delta
Lake table

delta_table_path = "abfss://<container-name>@<storage-account-name>.
dfs.core.windows.net/<your-delta-table-path>"

# Create a DataFrame with the new rows you want to append

# Replace the column names and values with those relevant to your table

new_rows = [

    ("NewValue1", 10),

    ("NewValue2", 20)

    # Add as many rows as needed

]

# Define the schema based on your Delta Lake table structure

# This is an example schema; adjust it to match your table's columns
and data types

schema = ["ColumnName1", "ColumnName2"]

# Create a DataFrame with the new data

new_data_df = spark.createDataFrame(new_rows, schema)

# Append the new data to the Delta Lake table

# Ensure the table format is set to 'delta' for Delta Lake
compatibility

new_data_df.write.format("delta").mode("append").save(delta_table_path)

# Verify by reading back the data from the Delta Lake table

df = spark.read.format("delta").load(delta_table_path)

df.show()
```

Como ilustran los ejemplos, un almacén de lago de datos, creado sobre cualquier plataforma con las ventajas de las plataformas abiertas, permite a las empresas utilizar varios motores para trabajar en la misma copia de los datos al mismo tiempo.

Conector de Azure Databricks en Power BI

El conector de Power BI para Azure Databricks proporciona una integración perfecta entre Power BI y Azure Databricks, lo que permite a las organizaciones conectar, analizar y visualizar los datos almacenados en Azure Databricks con facilidad. Esta integración admite la autenticación Microsoft Entra ID, lo que elimina la necesidad de que los administradores generen tokens de acceso personal para la conexión. Se ha diseñado para mejorar la conectividad de datos y las experiencias de análisis, lo que permite una visualización de datos eficiente y segura directamente desde el lago de datos.

1. Obtén el nombre de host del servidor de Azure Databricks y la ruta HTTP para configurar la conexión en Power BI.
2. Inicia Power BI Desktop.
3. Elige **Get Data** (Obtener datos) en la página de inicio o navega hasta **File > Get Data** (Archivo > Obtener datos).
4. Busca **Azure Databricks**.
5. Selecciona **Azure Databricks connector** (Conector de Azure Databricks) y a continuación, haz clic en **Connect** (Conectar).
6. Introduce el nombre de host del servidor y la ruta HTTP que has obtenido antes.
7. Decide entre los modos **Import** (Importar) y **DirectQuery** para la conectividad de datos. Para obtener más información sobre estas opciones, plantéate la posibilidad de leer acerca del uso de [DirectQuery en Power BI Desktop](#).
8. Selecciona tu método de autenticación de preferencia:
 - a. **Token de acceso personal:** introduce el token de acceso personal de Azure Databricks.
 - b. **Microsoft Entra ID:** elige **Sign in (Iniciar sesión)** y sigue las indicaciones.
 - c. **Nombre de usuario/contraseña:** normalmente esta opción no se aplica.
9. Después de la autenticación, Power BI te presentará la ventana **Navigator**. Aquí puedes seleccionar los datos de Azure Databricks que quieras consultar. Si tu área de trabajo tiene Unity Catalog habilitado, primero seleccionarás un catálogo, seguido de un esquema y una tabla.

Para las cargas de trabajo que necesitan la potencia de procesamiento y la flexibilidad que ofrece Azure Databricks, las empresas pueden utilizar las funcionalidades de visualización avanzadas de Power BI junto con Azure Databricks.

Logra la excelencia con Azure Databricks y Microsoft Fabric

La integración de Azure Databricks y Microsoft Fabric representa un enfoque transformador para administrar y analizar datos dentro de entornos de cloud modernos. Azure Databricks proporciona una plataforma de alto rendimiento para el procesamiento de datos y análisis basados en IA, mientras que Fabric mejora estas capacidades con herramientas de administración de datos robustas. Esta combinación permite a las organizaciones aprovechar los análisis avanzados y las soluciones de IA de manera más eficaz.

En conjunto, Azure Databricks y Fabric optimizan las cargas de trabajo de análisis al proporcionar un acceso a los datos sin problemas sin necesidad de copias de datos redundantes. Esta integración admite consultas directas desde Power BI, lo que mejora el rendimiento y una arquitectura de datos simplificada.

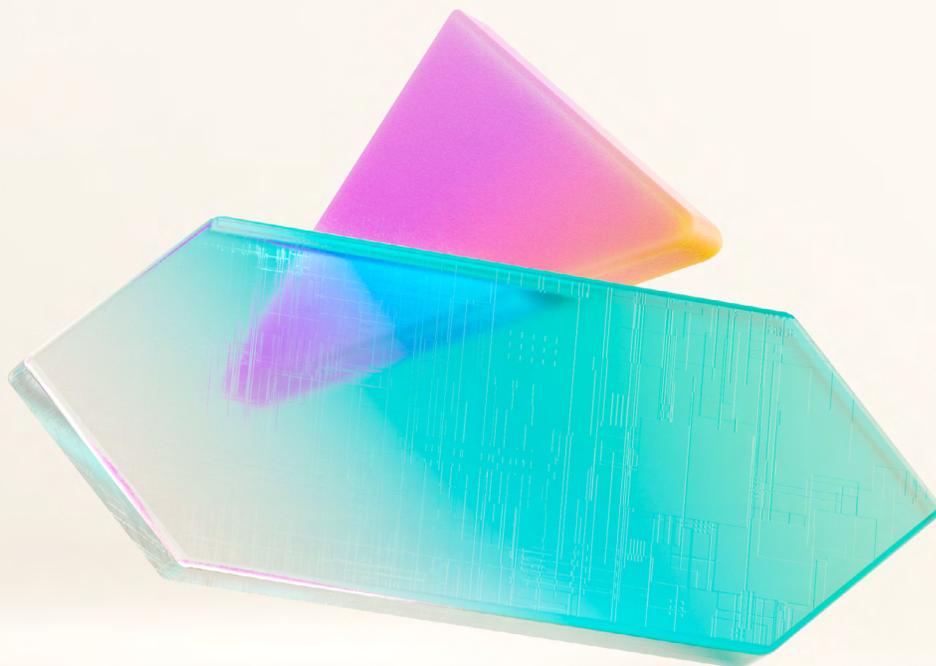
La sinergia entre Azure Databricks y Fabric libera todo el potencial de la arquitectura de los almacenes de lagos, lo que mejora la capacidad de administrar, procesar y analizar grandes conjuntos de datos de manera más eficaz. Esto permite a las organizaciones implementar soluciones de IA con facilidad y abordar desafíos analíticos complejos.

Además, las capacidades combinadas de Azure Databricks y Fabric abordan todo el ciclo de vida de los proyectos de IA y machine learning, desde la ingesta y el almacenamiento de datos hasta el análisis y la implementación. Esta integración ayuda a las organizaciones a mantener el control sobre sus procesos de datos y a garantizar que las iniciativas de IA sean escalables y cumplan las normativas del sector.

Las organizaciones pueden mejorar su gestión de los datos, impulsar la eficiencia operativa y fomentar la innovación en IA y el análisis con la eficaz integración de Azure Databricks y Fabric, con el fin de estar bien equipadas ante una economía basada en datos y alcanzar sus objetivos estratégicos.

Siguientes pasos

- Habla con un [especialista en ventas](#).
- Obtén más valor de tus datos con [Microsoft Fabric](#).
- Obtén conocimientos de tus datos y crea soluciones de IA con [Azure Databricks](#).
- [Mira el seminario web](#) para obtener más información sobre cómo Azure Databricks y Microsoft Fabric funcionan mejor juntos.



Apéndice

Introducción de solicitudes

Quiero que actúes como ingeniero de datos. Te proporcionaré detalles sobre un caso en el que un cliente necesita utilizar los datos almacenados en ADLS Gen2. Azure Databricks y Microsoft Fabric accederán a ellos para cargar datos para modelos de machine learning. Tu trabajo consistirá en crear una arquitectura que responda a esta necesidad.

Respuesta de Azure OpenAI

Para abordar el caso en el que un cliente necesita utilizar los datos almacenados en Azure Data Lake Storage Gen2 (ADLS Gen2), a los que tienen acceso Azure Databricks y Microsoft Fabric para cargar datos en modelos de machine learning, diseñaremos una arquitectura que aproveche eficazmente estos servicios. Esta arquitectura tiene como objetivo optimizar el procesamiento de datos, mejorar la gestión de los datos y garantizar operaciones de machine learning escalables.

Información general de la arquitectura

- Capa de almacenamiento de datos: Azure Data Lake Storage Gen2: actúa como el repositorio central para almacenar datos sin procesar, datos procesados y artefactos de machine learning. ADLS Gen2 está optimizado para casos de análisis a gran escala y admite espacios de nombres jerárquicos, lo que simplifica la administración de datos.
- Capa de procesamiento y análisis de datos: Azure Databricks: sirve como la plataforma principal para el procesamiento, análisis y desarrollo de modelos de machine learning. Azure Databricks proporciona un entorno de colaboración con compatibilidad con varios lenguajes de programación y se integra perfectamente con ADLS Gen2 para obtener acceso directo a los datos.
- Capa de gestión y administración de datos: Microsoft Fabric: se utiliza para administrar y gestionar los datos en todo el entorno de datos. Microsoft Fabric permite crear un modelo de gestión de los datos unificado que abarca los datos de ADLS Gen2 y los datos procesados o generados por Azure Databricks.

- Operativización de machine learning: Azure Machine Learning: se integra con Azure Databricks para implementar y administrar modelos de machine learning. Este servicio proporciona herramientas para el control de versiones, la supervisión y el servicio de modelos, lo que facilita la operativización de los modelos de machine learning.

Flujo de trabajo detallado

1. Ingesta de datos: los datos se ingieren en ADLS Gen2 desde diversos orígenes, incluidas bases de datos estructuradas, dispositivos de IoT, archivos de registro, etc. Estos datos se almacenan en formato sin procesar dentro de una estructura jerárquica del sistema de archivos.

2. Procesamiento y exploración de datos:

- Los cuadernos de Azure Databricks se utilizan para acceder a los datos sin procesar almacenados en ADLS Gen2. Los ingenieros y científicos de datos procesan y exploran estos datos mediante Spark, realizando tareas como limpieza de datos, la transformación e ingeniería de características.
- Los datos procesados se pueden volver a almacenar en ADLS Gen2 en un formato optimizado (por ejemplo, Delta Lake) para su posterior análisis o machine learning.

3. Linaje y gestión de datos:

- Microsoft Fabric está configurado para catalogar los activos de datos en ADLS Gen2 y los productos de datos generados por Azure Databricks. Proporciona una vista unificada de los datos, aplicando políticas, reglas de calidad de los datos y medidas de seguridad.
- Microsoft Fabric permite configurar el linaje y el seguimiento de los datos para mantener la visibilidad de las transformaciones y el uso de los datos durante todo el ciclo de vida de los datos.

4. Entrenamiento e implementación de modelos de machine learning:

- Los científicos de datos utilizan Azure Databricks para desarrollar y entrenar modelos de machine learning, aprovechando el MLflow integrado para el seguimiento de experimentos y la administración de modelos.

- Los modelos entrenados se registran en Azure Machine Learning, que se utiliza para implementar estos modelos en producción como servicios web o contenedores, lo que permite puntuar en tiempo real o por lotes.

5. Supervisión y administración:

- Azure Machine Learning proporciona capacidades de supervisión de modelos, que rastrean el rendimiento de modelos y el desfase de los datos a lo largo del tiempo. Facilita los flujos de trabajo de reentrenamiento cuando los modelos se vuelven obsoletos.
- Microsoft Fabric supervisa el aspecto de la gestión y el cumplimiento, lo que garantiza que el acceso a los datos y las implementaciones de modelos cumplan los estándares organizativos y reglamentarios.

Esta arquitectura proporciona una solución completa para aprovechar los datos de ADLS Gen2 en Azure Databricks y Microsoft Fabric, lo que facilita el desarrollo, la gestión y la operativización de modelos de machine learning eficientes. Garantiza la escalabilidad, la seguridad y el cumplimiento al responder a las necesidades dinámicas de las organizaciones modernas basadas en datos.

Nota

Es posible que las indicaciones y respuestas indicadas aquí no reflejen la experiencia exacta.

Introducción de solicitudes

En función de lo anterior, ¿puedes proporcionar instrucciones de compilación detalladas con código?

Respuesta de Azure OpenAI

La creación de una guía de compilación detallada con código para integrar Azure Data Lake Storage Gen2 (ADLS Gen2), Azure Databricks y Microsoft Fabric en una arquitectura cohesionada para el machine learning requiere varios pasos. A continuación, describiremos fases clave en este proceso, incluida la configuración, el procesamiento de datos y la administración del ciclo de vida del modelo de machine learning. En esta guía se presupone una comprensión básica de los servicios de Azure y una suscripción donde puedas implementar estos servicios.

1. Ajuste y configuración

Azure Data Lake Storage Gen2

1. Crea una cuenta de almacenamiento de ADLS Gen2:

- Navega hasta el portal de Azure.
- Crea una nueva cuenta de almacenamiento, seleccionando tu suscripción, un grupo de recursos y el tipo de cuenta "StorageV2".
- Habilita el espacio de nombres jerárquico para utilizar características de ADLS Gen2.

2. Configura el acceso a la cuenta de almacenamiento:

- En la cuenta de almacenamiento, ve a la sección "Access Keys" (Claves de acceso) y anota la cadena de conexión para usarla posteriormente.
- Como alternativa, configura la autenticación de Azure Active Directory (Azure AD) para lograr un acceso más seguro.

Azure Databricks

1. Crea un área de trabajo de Azure Databricks:

- En el portal de Azure, crea un nuevo servicio de Azure Databricks.
- Una vez completada la implementación, dirígete al área de trabajo de Azure Databricks e inicia el área de trabajo.

2. Crea un clúster en Azure Databricks:

- Dentro del área de trabajo de Azure Databricks, ve a "Compute" (Computación).
- Crea un nuevo clúster y asegúrate de que sea compatible con las bibliotecas que tienes previsto usar.

3. Monta ADLS Gen2 en Azure Databricks (para obtener acceso directo a los datos):

- En un cuaderno de Azure Databricks, utiliza el siguiente código para montar el almacenamiento ADLS Gen2 y tener acceso directo:

Python

```
configs = { "fs.azure.account.auth.type": "OAuth", "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider", "fs.azure.account.oauth2.client.id": "<Application-ID>", "fs.azure.account.oauth2.client.secret": dbutils.secrets.get(scope="<scope-name>", key="<secret-key>"), "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/<Directory-ID>/oauth2/token" } dbutils.fs.mount( source="abfss://<filesystem>@<storage-account-name>.dfs.core.windows.net/", mount_point="/mnt/<mount-name>", extra_configs=configs)
```

- Sustituye los marcadores de posición con los detalles de registro reales de tu aplicación de Azure AD y los detalles de almacenamiento de ADLS Gen2.

Microsoft Fabric

1. Configuración de Microsoft Fabric:

- Actualmente, Microsoft Fabric es un punto de integración conceptual y puede que no tenga instrucciones de configuración directas orientadas al cliente disponibles públicamente. Normalmente, el ajuste implicaría configurar el entorno para catalogar y administrar activos de datos en Azure Databricks y ADLS Gen2.
- Con fines de demostración, vamos a continuar con los pasos siguientes suponiendo que Microsoft Fabric se utiliza para gestionar y proteger el acceso a los datos, garantizando el cumplimiento y la calidad de los datos.

2. Procesamiento de datos con Azure Databricks

1. Carga datos desde ADLS Gen2:

- Utiliza la ruta montada para leer los datos en un DataFrame de Azure Databricks:

```
pythonCopy code
```

```
df = spark.read.csv("/mnt/<mount-name>/path/to/data.csv")
```

2. Transformación de datos:

- Realizar las transformaciones de datos necesarias mediante las operaciones DataFrame de Spark.

3. Guarda los datos procesados:

- Vuelve a guardar los datos procesados en ADLS Gen2 en un formato optimizado, como Delta Lake:

pythonCopy code

```
df.write.format("delta").save("/mnt/<mount-name>/path/to/processed/data/")
```

3. Entrenamiento e implementación de modelos de machine learning

1. Entrenamiento de modelos:

- Utiliza Azure Databricks para entrenar un modelo de machine learning con los datos procesados.
- Ejemplo usando Spark MLlib para la regresión logística:

pythonCopy code

```
from pyspark.ml.classification import LogisticRegression # Assuming df is a DataFrame prepared for training # Train a model lr = LogisticRegression(featuresCol='features', labelCol='label') model = lr.fit(df) # Save the model model.write().overwrite().save("/mnt/<mount-name>/path/to/model")
```

2. Implementación del modelo:

- Integra con Azure Machine Learning para la implementación, supervisión y administración de modelos. Esta parte es compleja y normalmente implica registrar el modelo, crear un script de puntuación e implementar el modelo como un servicio web.

Nota

Es posible que las indicaciones y respuestas indicadas aquí no reflejen la experiencia exacta.

