Microsoft Fabric

Microsoft

Análisis de lakehouse

con Microsoft Fabric y Azure Databricks

Análisis de lakehouse

con Microsoft Fabric y Azure Databricks

Desbloquea el valor de los datos

Hoy en día, los datos se comparan a menudo con el oro, por lo que empresas de todo el mundo compiten para extraer, refinar y capitalizar sus inmensos activos de datos. Esta ola transformadora está impulsada por una infraestructura de cloud público que proporciona a las empresas una enorme capacidad de computación y almacenamiento para procesar sus crecientes patrimonios de datos. Microsoft Azure, la plataforma en el cloud líder de hoy en día, proporciona servicios que permiten a las organizaciones hacer frente a estos modernos desafíos de datos. A la vanguardia está Microsoft Fabric, un nuevo motor de análisis que está remodelando el ámbito de la administración de datos. En combinación con Azure Databricks, las empresas disponen de un amplio conjunto de herramientas para aprovechar el poder del cloud y abordar sus iniciativas de datos más ambiciosas.

Desde ingenieros de datos hasta responsables de la toma de decisiones, Fabric proporciona a todos los usuarios las herramientas y los conocimientos necesarios para dirigir a las empresas hacia el futuro. Al presentar una plataforma cohesionada que combina diversas funcionalidades de análisis, Fabric garantiza que las empresas puedan desbloquear el potencial de sus datos sin la complejidad que supone administrar varias herramientas. Con OneLake, una plataforma de datos pionera, Fabric permite realizar análisis integrales proporcionando una capa de almacenamiento seguro y administrado para crear y mantener un data lakehouse moderno.

¿Qué es un lakehouse?

La arquitectura de lakehouse es el enfoque moderno para crear una plataforma de análisis escalable para el patrimonio de datos en crecimiento de una empresa. Combina la precisión de un almacén de datos tradicional con la enorme escala y flexibilidad de un data lake. Mediante el uso de herramientas modernas como Microsoft Fabric, Microsoft Power BI y Azure Databricks, las empresas pueden crear un lakehouse que satisfaga las necesidades de los ingenieros de datos, analistas del negocio y científicos de datos, compartiendo una única copia de los datos almacenada en un formato abierto y controlada por un catálogo unificado.

La arquitectura de lakehouse de Azure adopta en su totalidad la potencia de Apache Spark, un motor de análisis escalable para las empresas. Con los asistentes de IA de Fabric y Azure Databricks, los equipos pueden trabajar más eficazmente para compartir datos, encontrar nuevos conocimientos y crear modelos de IA avanzados.

La arquitectura tradicional de los data lakes y los almacenes

Los data lakes o lagos de datos son grandes depósitos que almacenan una gran variedad de tipos de datos brutos sin procesar. Como repositorios centralizados, los data lakes están diseñados para ingerir y almacenar grandes volúmenes de datos en su forma más auténtica. Esta arquitectura se caracteriza por su flexibilidad y permite a las empresas almacenar datos estructurados, semiestructurados y no estructurados sin las restricciones de un esquema predefinido.

Los almacenes de datos son repositorios estructurados, diseñados para realizar consultas y análisis eficientes. Un almacén de datos almacena datos que se han limpiado, organizado, procesado y transformado para que estén listos para las aplicaciones de business intelligence (BI).

Área	Data Lake	Almacén de datos
Almacén de datos	Puede capturar y retener datos no estructurados, semiestructurados y estructurados en su formato sin procesar. Un data lake almacena todos los tipos de datos, independientemente del origen y la estructura.	Puede capturar y retener únicamente datos estructurados. Un almacén de datos almacena datos en métricas cuantitativas con sus atributos. Los datos se transforman y limpian.
Definición del esquema	Normalmente, el esquema se define después de almacenar los datos. Esto ofrece una alta agilidad y permite capturar datos con bastante facilidad, pero requiere trabajo al final del proceso (esquema de lectura).	Normalmente, el esquema se define antes de almacenar los datos. Requiere trabajo al principio del proceso, pero ofrece rendimiento, seguridad e integración (esquema de escritura).
Calidad de los datos	Cualquier dato, saneado o no saneado (como datos sin procesar).	Datos cuidadosamente seleccionados que sirven como la versión central de la realidad.
Usuarios	Un data lake es ideal para los usuarios a los que se les permite realizar análisis profundos, como científicos de datos, ingenieros de datos y analistas de datos.	Un almacén de datos es ideal para usuarios operativos como analistas del negocio, ya que está bien estructurado y es fácil de usar y entender.
Precio y rendimiento	El coste de almacenamiento es relativamente bajo en comparación con un almacén de datos y los resultados de las consultas son más rápidos.	El coste de almacenamiento es elevado y la consulta de los resultados lleva mucho tiempo.
Accesibilidad	Un data lake tiene pocas restricciones y es fácilmente accesible. Los datos se pueden cambiar y actualizar rápidamente.	Un almacén de datos está estructurado por diseño, lo que dificulta el acceso y la manipulación.

Tabla 1: Comparación de los data lakes y los almacenes de datos

Administración y análisis de datos con Microsoft Fabric y Azure Databricks

Microsoft Fabric es una plataforma unificada de análisis que reúne todas las herramientas de datos y análisis que las organizaciones necesitan. Azure Databricks es un servicio de Azure totalmente administrado que permite el análisis y la IA en un lakehouse de Azure.

Fabric y Azure Databricks proporcionan soluciones integrales que permiten a los ingenieros de datos, científicos de datos, administradores de datos, analistas de datos y consumidores de datos trabajar juntos para encontrar conocimientos valiosos. Dado que ambas plataformas se basan en el almacenamiento de Delta Lake, los dos servicios pueden funcionar juntos, compartiendo las mismas copias de los datos. Cuando se combinan, Fabric y Azure Databricks ofrecen una potente sinergia que mejora las soluciones de análisis de datos, procesamiento e IA. A continuación se incluyen cinco ejemplos concretos de cómo Fabric funciona mejor con Azure Databricks:

- **Canalizaciones de datos eficientes**: Azure Databricks permite a los ingenieros de datos aprovechar la potencia de Apache Spark con la aceleración de Photon para crear canalizaciones de datos eficientes y escalables. Estas canalizaciones pueden proporcionar datos a OneLake, lo que garantiza que los datos estén disponibles para el análisis y otras operaciones. Esta integración garantiza que los ingenieros y científicos de datos disfruten de una experiencia ininterrumpida al trabajar con datos en ambas plataformas.

- **Almacenamiento de datos unificado con OneLake**: Azure Databricks puede interactuar directamente con los datos almacenados en OneLake. Tanto si los datos se originan en sistemas on-premises como si se ingieren de fuentes como Azure Databricks, OneLake proporciona las herramientas para consolidar estos datos. Esta arquitectura minimiza las copias de datos, ofrece un gobierno consolidado y permite a los usuarios utilizar sus aplicaciones preferidas como Azure Databricks para las consultas y la ciencia de datos.

- **Sinergia de IA y machine learning**: Azure Databricks ayuda a crear modelos complejos de IA y machine learning. Cuando se combina con las características de Copilot basadas en IA de Fabric, las empresas pueden obtener conocimientos de sus datos de manera más eficiente. Esta sinergia garantiza que los científicos de datos puedan crear e implementar modelos en Azure Databricks y usar después Fabric para el análisis de BI, utilizando las funcionalidades de IA de ambas plataformas.

- **Flexibilidad con el almacenamiento de datos**: hay dos enfoques posibles para cargar datos en OneLake mediante Azure Databricks:

  - **Lakehouse basado en ADLS**: los datos se pueden almacenar en tablas de Delta Lake en una cuenta de Azure Data Lake Storage (ADLS). Después se puede crear un acceso directo a este almacenamiento en la base de datos de Fabric Lakehouse para que se pueda acceder a los datos de forma sencilla y segura dentro de Fabric.
  - **Lakehouse basado en OneLake**: los datos se pueden almacenar directamente en la ubicación de ADLS de OneLake. Este enfoque requiere identificar la ubicación de almacenamiento predeterminada para Fabric Lakehouse. Una vez almacenados los datos en esta ubicación, se puede acceder a ellos y usarse en Fabric, lo que garantiza una integración perfecta entre las dos plataformas.

- **Visualizaciones de datos mejoradas**: una vez almacenados y procesados los datos, la creación de visualizaciones se convierte en un paso crucial. Con la integración de Azure Databricks y Fabric, los usuarios pueden crear fácilmente conjuntos de datos de Power BI a partir de los datos procesados. Esta integración garantiza que los científicos y analistas de datos puedan visualizar sus datos, obtener conocimientos y tomar decisiones fundamentadas basadas en el análisis proporcionado por ambas plataformas.

La integración entre Fabric y Azure Databricks ofrece una solución integral para las necesidades de análisis de datos de una organización. Con el lanzamiento de la preview pública de OneLake, estos dos eficaces almacenes de datos tienen el potencial de simplificar las tareas de análisis.

OneLake: el núcleo de Microsoft Fabric

OneLake, también conocido como Microsoft Fabric Lake, es el elemento fundamental de todos los servicios de Fabric. Proporciona un centro de almacenamiento unificado para los datos de la organización, basado en el robusto ADLS Gen2. OneLake atiende a una base de usuarios diversa, desde profesionales experimentados hasta desarrolladores en ciernes, con el objetivo principal de descomponer los silos de datos. Facilita la detección y el intercambio de datos, y garantiza un cumplimiento de la seguridad centralizado.

Unificar los datos con Microsoft Fabric

Microsoft Fabric va más allá de las herramientas de análisis tradicionales. Proporciona una plataforma unificada que simplifica el proceso de análisis, desde la integración de los datos hasta la obtención de conocimientos en tiempo real. Ofrece una solución completa de análisis para las empresas, con lo que se elimina la necesidad de servicios específicos de varios proveedores. Fabric reúne componentes nuevos y existentes de Power BI, Azure Synapse y Azure Data Factory en una única plataforma de software como servicio (SaaS) para garantizar una experiencia de usuario cohesionada.

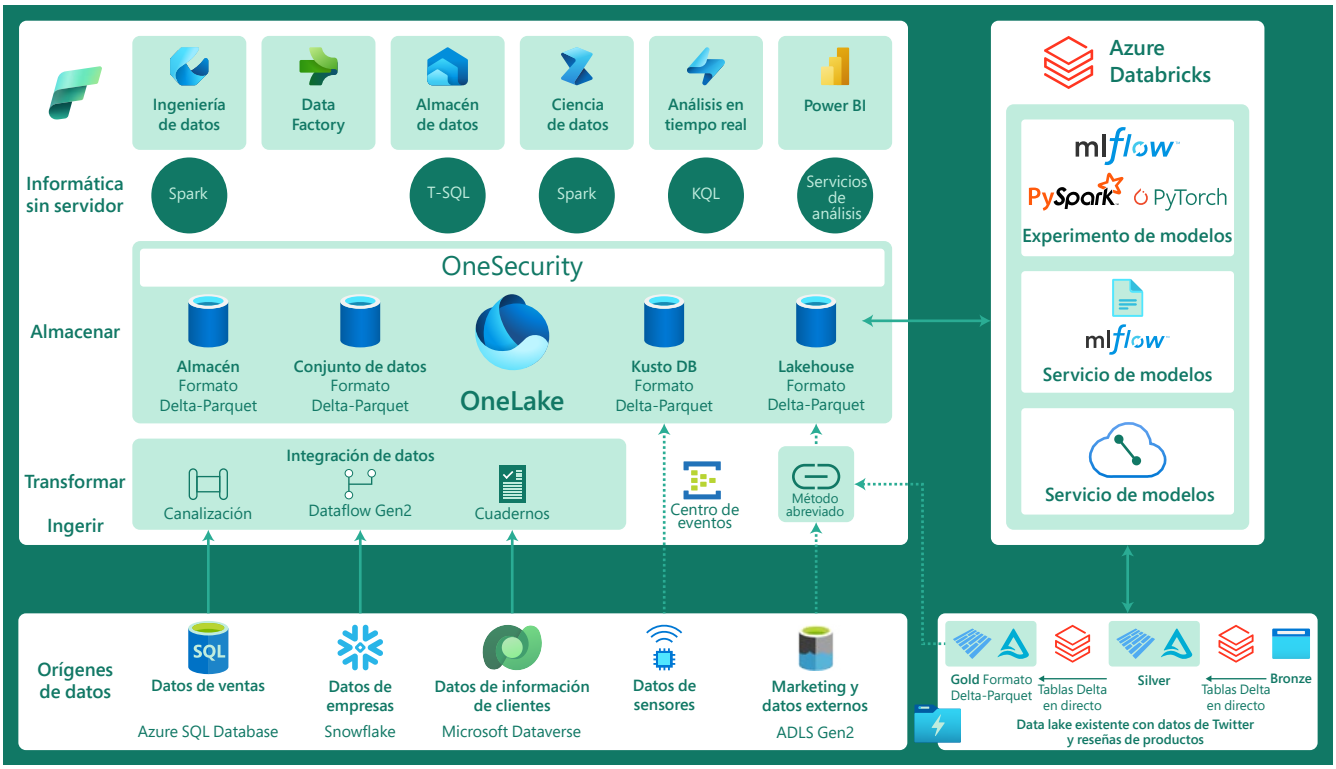

Figura 1: Arquitectura de Microsoft Fabric y Azure Databricks

Componentes de Microsoft Fabric

Microsoft Fabric ofrece soluciones de análisis integrales. Conecta todos los servicios de orígenes de datos y análisis a través de experiencias líderes del sector en una amplia variedad de categorías.

- **Synapse Data Engineering**: Synapse Data Engineering, una experiencia básica de Microsoft Fabric, dota a los ingenieros de datos de una plataforma Spark de primer nivel y facilita la transformación de los datos a gran escala dentro del paradigma del lakehouse. Simplifica el proceso de ingesta, transformación e intercambio de los datos de la organización en un formato abierto.

- **Data Factory**: Data Factory se ha diseñado para proyectos complejos híbridos de extracción, transformación y carga (ETL) e integración de datos. Ayuda a transformar los datos sin procesar desorganizados en conocimientos útiles. Data Factory permite la creación de flujos de trabajo basados en datos, conocidos como canalizaciones, para ingerir datos de diversos almacenes de datos. Estas canalizaciones pueden utilizar flujos de datos o servicios de computación como Azure HDInsight Hadoop, Azure Databricks y Azure SQL Database para transformar los datos.

- **Synapse Data Science**: Synapse Data Science desbloquea el valor de los datos en el flujo de trabajo de análisis de una organización. Proporciona herramientas y funciones que permiten a los científicos de datos crear, implementar y administrar fácilmente modelos de machine learning mediante integraciones de Azure Machine Learning.

- **Synapse Data Warehouse**: integrado con Power BI, Fabric ofrece funciones de almacenamiento de datos de última generación. Es compatible con un formato de datos abierto, lo que garantiza que los ingenieros de datos y los equipos de TI puedan trabajar fácilmente almacenando los datos en el formato abierto de Delta Lake para disfrutar de un rendimiento de SQL de primer nivel.

- **Synapse Real-Time Analytics**: Synapse Real-Time Analytics se ha diseñado para manejar datos observacionales de varios orígenes y procesar datos semiestructurados en grandes volúmenes. Proporciona información en tiempo real para que las empresas puedan tomar decisiones fundamentadas basadas en los datos más recientes.

- **Power BI**: Power BI es una completa colección de servicios de software, aplicaciones y conectores diseñados para transformar los orígenes de datos no relacionados en conocimientos coherentes, visualmente inmersivos e interactivos. Tanto si los datos residen en una hoja de cálculo de Excel como en una combinación de almacenes de datos on-premises y basados en el cloud, Power BI facilita la conexión, la visualización y el intercambio de conocimientos. Su compatibilidad con soluciones on-premises y en el cloud garantiza que las empresas puedan elegir dónde desean almacenar sus informes en función de sus necesidades específicas.

- **Data Activator**: Data Activator ofrece una experiencia sin código que permite a los analistas del negocio realizar operaciones automáticas a partir de los datos. Esto simplifica el proceso de activación de datos para que los usuarios puedan obtener conocimientos útiles a partir de sus datos.

Ventajas de Microsoft Fabric

Las empresas no paran de buscar herramientas que puedan integrar sin problemas grandes cantidades de datos que ofrezcan conocimientos que impulsen la toma de decisiones y la innovación. Si bien muchas soluciones prometen funciones de análisis completas, Microsoft Fabric ofrece una experiencia verdaderamente integrada adaptada a las necesidades diversas de las empresas modernas. Se ha diseñado para aumentar la simplicidad y la integración, ofreciendo un conjunto completo de servicios de análisis. En comparación con los data lakes y almacenes tradicionales, Fabric ofrece varias ventajas:

- **Conjunto completo de análisis**: Fabric actúa como una solución de análisis integral que abarca el movimiento de datos, la ciencia de datos, el análisis en tiempo real y la BI, lo que simplifica el proceso de análisis.
- **Ecosistema integrado**: Fabric proporciona un ecosistema integrado de servicios, incluidos data lake, ingeniería de datos e integración de datos. Esto garantiza un flujo de datos fluido en toda la canalización de análisis, lo que reduce los silos de datos y aumenta la eficiencia.
- **Implementación y administración simplificadas**: al ser una plataforma SaaS, Fabric administra la configuración y el mantenimiento de la infraestructura, lo que permite a las organizaciones centrarse en el análisis en lugar de en la administración de la infraestructura.
- **Escalabilidad y flexibilidad**: Fabric, diseñado para escalarse con facilidad, puede gestionar distintas cargas de trabajo, lo que garantiza un procesamiento eficiente de los datos sin problemas de rendimiento.
- **Análisis en tiempo real**: Fabric admite análisis en tiempo real, lo que permite a las organizaciones obtener conocimientos de los flujos de datos en tiempo real y tomar decisiones fundamentadas rápidamente.
- **Funciones avanzadas de ciencia de datos**: la plataforma ofrece herramientas para que los científicos de datos realicen análisis de datos complejos, desarrollen modelos de machine learning y obtengan conocimientos predictivos.
- **Solución integral**: Fabric proporciona una solución de análisis integral que simplifica todo el flujo de trabajo, desde la ingesta de datos hasta la visualización.
- **Seguridad y cumplimiento**: al formar parte del ecosistema de Microsoft, Fabric garantiza medidas de seguridad robustas y certificaciones de cumplimiento que protegen los datos confidenciales.
- **Interfaz fácil de usar**: diseñada para ser accesible, la interfaz de la plataforma atiende las necesidades de usuarios técnicos y no técnicos, lo que democratiza el análisis de datos en todas las organizaciones.
- **Rentabilidad**: al integrar un conjunto de servicios, Fabric puede ahorrar costes en comparación con la administración de varias soluciones de análisis independientes.

Al ofrecer estos servicios y características, Fabric garantiza que las empresas dispongan de un conjunto completo de herramientas para abordar todas sus necesidades de análisis, desde la integración de datos hasta la obtención de conocimiento en tiempo real.

Integración de IA

Azure OpenAI Service se basa en la infraestructura optimizada de IA diseñada específicamente para Azure. Varios productos de Microsoft se basan en esta infraestructura. La integración de IA es la piedra angular de Microsoft Fabric, ya que es lo que la diferencia de otras plataformas de análisis:

- **Plataforma unificada impulsada por IA**: Fabric no es solo una plataforma de análisis; es una plataforma de análisis basada en IA. Esto significa que cada capa de Fabric está integrada con funciones de IA, lo que hace que el análisis de datos sea más intuitivo, eficiente y potente. Esto permite a las organizaciones aprovechar todo el potencial de sus datos.

- **Modelos lingüísticos y de IA generativa**: Fabric utiliza servicios de modelos lingüísticos y de IA generativa, como Azure OpenAI Service. Estos servicios permiten la creación de experiencias de IA cotidianas que redefinen la forma en que los empleados utilizan su tiempo. Por ejemplo, para poder disfrutar de experiencias de IA específicas de cada organización se requiere un suministro constante de datos limpios desde un sistema de análisis bien administrado. Fabric garantiza que estos datos estén disponibles y se procesen de manera eficiente.

- **Copilot de Microsoft Fabric**: una de las características más destacadas de Fabric es Copilot. Con Copilot integrado en todas las experiencias de datos de Fabric, los usuarios pueden emplear un lenguaje conversacional para realizar una serie de tareas. Esto incluye la creación de flujos de datos y canalizaciones, la generación de código, la creación de modelos de machine learning y la visualización de los resultados. Este enfoque de IA conversacional simplifica las tareas complejas y hace que la plataforma sea más fácil de usar. Además, Copilot cumple los compromisos adquiridos por Microsoft en términos de seguridad y privacidad de los datos, garantizando que los datos de la organización permanezcan protegidos.

- **Conocimientos basados en IA**: la integración de Fabric y Power BI permite realizar análisis basados en IA para que los analistas y los usuarios de la empresa puedan obtener conocimientos valiosos a partir de los datos. La experiencia de Power BI dentro de Fabric también está profundamente integrada en Microsoft 365, lo que garantiza que los conocimientos estén disponibles fácilmente donde trabajan los usuarios de la empresa.

- **Data Activator**: Data Activator in Fabric, una característica que estará disponible próximamente, ofrece detección y supervisión de datos en tiempo real. Activa notificaciones y acciones basadas en patrones especificados en los datos, todo ello dentro de un entorno sin código. Esta función basada en IA mejora aún más las funcionalidades de la plataforma para proporcionar conocimientos pertinentes en el momento oportuno.

Enfoque de IA responsable

Microsoft reconoce la importancia de una IA responsable, especialmente con tecnologías potentes como los modelos generativos. Microsoft ha adoptado un enfoque iterativo para los modelos de gran tamaño, trabajando estrechamente con OpenAI y sus clientes para evaluar casos de uso y abordar riesgos potenciales.

Azure OpenAI Service ha implementado medidas de seguridad acordes con los principios de IA responsable de Microsoft. Antes de acceder al servicio se exige a los desarrolladores que describan su caso de uso o aplicación previstos. Existen filtros de contenido para supervisar tanto la entrada proporcionada al servicio como el contenido generado para asegurar el cumplimiento de las directrices de las políticas.

Microsoft ha identificado los seis principios siguientes para guiar el desarrollo y uso de la IA:

- **Imparcialidad**: Microsoft hace hincapié en la importancia de la imparcialidad en los sistemas de IA. La empresa quiere garantizar que los sistemas de IA asignen oportunidades, recursos o información de manera equitativa para todos los usuarios.

- **Fiabilidad y protección**: Microsoft da prioridad a la fiabilidad y la seguridad de los sistemas de IA, garantizando que funcionen de manera coherente y prevengan cualquier posible daño.

- **Privacidad y seguridad**: proteger los datos de los usuarios es primordial. Microsoft garantiza que los sistemas de IA cumplan los estándares más altos de privacidad y seguridad, protegiendo la información de los usuarios del acceso no autorizado y de las infracciones.

- **Inclusividad**: Microsoft cree en la creación de sistemas de IA que atiendan las necesidades de todo el mundo, garantizando que todos los usuarios, independientemente de su experiencia o aptitudes, puedan beneficiarse de la tecnología.

- **Transparencia**: Microsoft está comprometido con la transparencia de los sistemas de IA, asegurándose de que los usuarios entiendan cómo estos sistemas funcionan y toman decisiones.

- **Responsabilidad**: Microsoft se hace responsable de los sistemas de IA que desarrolla, garantizando que existan mecanismos para abordar cualquier problema o preocupación relacionado con la tecnología.

En esencia, estos principios para guiar el desarrollo de una IA responsable priorizan la imparcialidad, la seguridad, la privacidad, la inclusividad, la transparencia y la responsabilidad, garantizando que las tecnologías de IA sean fiables y beneficiosas para todos los usuarios.

Modelo de precios simplificado

Microsoft Fabric ha desarrollado un modelo de precios optimizado, diseñado para optimizar la eficiencia y acorde con los objetivos financieros de la empresa. Para esta estrategia, es fundamental la adquisición de un único conjunto de créditos computacionales diseñados para potenciar todas las operaciones dentro del marco de Fabric.

El modelo de precios se basa en un sistema unificado de créditos computacionales. Ya se trate de ingeniería de datos, integración o análisis, Fabric emplea un grupo de créditos consolidado. Este enfoque ofrece flexibilidad operativa y un posible ahorro de costes. Aunque existen herramientas especializadas que satisfacen necesidades específicas, Fabric integra varias funciones en una plataforma económica, lo que facilita la planificación financiera. Sin embargo, pueden surgir problemas cuando los diferentes departamentos, como los ingenieros de datos y los analistas del negocio, tienen presupuestos separados. La toma de decisiones colaborativa resulta esencial para una asignación eficiente de los créditos.

El modelo de precios de Microsoft Fabric ofrece:

- **Rentabilidad**: la combinación de diversas tareas de análisis en una única estructura de precios puede producir un ahorro considerable. Elimina las molestias de realizar un seguimiento de los costes de los distintos servicios, lo que garantiza gastos más predecibles.

- **Adaptabilidad**: el sistema de créditos compartidos es versátil. Los créditos se pueden emplear allí donde más se necesiten, en consonancia con los requisitos empresariales en constante cambio.

- **Opciones sencillas**: un modelo de precios unificado elimina la complejidad de los múltiples costes de servicio, lo que simplifica las decisiones y reduce las tareas administrativas.

- **Aumento del uso de la plataforma**: Fabric ofrece precios sencillos y soluciones completas que pueden ser un imán para las empresas.

- **Preparación para el futuro**: para abordar la expansión de las empresas y sus necesidades de datos variables, este modelo de precios ayuda a las organizaciones a prever el crecimiento y conocer la demanda variable de análisis.

Programa más rápidamente con GitHub Copilot, Visual Studio Code y Azure Databricks

Herramientas como GitHub Copilot, con tecnología de OpenAI Codex, están cambiando la forma en que las organizaciones diseñan y utilizan los sistemas. Cuando se combinan con las características de la aplicación de Visual Studio Code, Copilot ofrece una experiencia de primer nivel para los desarrolladores. Pero para crear sistemas de IA como GitHub Copilot, no basta con disponer de un sencillo editor de código. Aquí es donde resulta útil la extensión Databricks para Visual Studio Code, que permite a los desarrolladores trabajar en Visual Studio Code, pero usar la potencia de Azure Databricks para implementar sus proyectos. Se trata de combinar un desarrollo sencillo y una ejecución eficaz con los consejos de programación de Copilot.

Crea un ayudante de programación de IA para Azure Databricks

Este tutorial te mostrará cómo puedes programar más rápidamente con Visual Studio Code, GitHub Copilot y la extensión Databricks para Visual Studio Code. Con la extensión Databricks para Visual Studio Code, cualquier código creado en Visual Studio Code puede ejecutarse sin problemas en Azure Databricks. También ayuda en el registro de modelos, lo que facilita la implementación.

Para seguir este ejemplo, necesitas:

- Tener Visual Studio Code 1.69.1 o superior instalado y configurado para programar en Python.
- Garantizar que Visual Studio Code se esté ejecutando y tenga un proyecto local abierto.
- Generar un token de acceso personal de Azure Databricks para el área de trabajo de Azure Databricks de destino.
- Añadir tu token de acceso personal de Azure Databricks como un campo de token junto con la URL de la instancia del área de trabajo (e.g., `https://dbc-a7h3345c-d6e7.cloud.databricks.com`) al perfil de configuración `DEFAULT` en tu archivo local `.databrickscfg`.
- Asegurarte de que Visual Studio Code esté configurado para la programación en Python. Normalmente, esto implica la instalación de la extensión de Python desde el marketplace de Visual Studio Code y la configuración del intérprete de Python.

- Generar un token de acceso personal de Azure Databricks para el área de trabajo de Azure Databricks de destino. Este token se usará para autenticar tu código de Visual Studio con Azure Databricks.
- Tener acceso a GitHub Copilot. Consulta la página siguiente para obtener más información sobre una versión de evaluación gratuita: https://github.com/features/copilot.

Instalar la extensión Databricks para Visual Studio Code

Para instalar la extensión Databricks para Visual Studio Code:

1. Abre Visual Studio Code.
2. Ve a la vista `Extensions` pulsando `Ctrl+Shift+X` o haciendo clic en el icono cuadrado de la barra lateral.
3. En la barra de búsqueda, escribe `Databricks extension for Visual Studio Code.`
4. En los resultados de búsqueda, busca la extensión y haz clic en el botón verde `Install`.
5. Espera a que finalice la instalación.

Configurar la extensión Databricks para Visual Studio Code

Para configurar la extensión Databricks para Visual Studio Code, sigue estos pasos:

1. Haz clic en el logotipo de Azure Databricks en la barra lateral de Visual Studio Code.
2. En el panel `Configuration` que aparece, haz clic en `Configure Databricks.`
3. Aparecerá una paleta de comandos en la que se te pedirá que selecciones el `host de Databricks`. Introduce la URL del área de trabajo de Azure Databricks y pulsa `Enter`.
4. A continuación, se te pedirá que te autentiques. Haz clic en `DEFAULT: Authenticate using the DEFAULT profile`.
5. Para configurar tu clúster de Azure Databricks, vuelve al panel `Configuration`, haz clic en `Cluster` y, a continuación, haz clic en el icono de engranaje que aparece. Esto te permitirá seleccionar y configurar el clúster que deseas utilizar.
6. Si el clúster que has elegido no se está ejecutando, haz clic en el icono `Start cluster` situado junto a `Cluster`.
7. Ahora configura el destino de sincronización. En el panel `Configuration`, haz clic en `Sync Destination` y, a continuación, haz clic en el icono de engranaje. Aparecerá una paleta de comandos con un nombre de destino de sincronización generado aleatoriamente. Haz clic en ella para seleccionar el destino.

Instalar la extensión GitHub Copilot

Para instalar la extensión GitHub Copilot, sigue estos pasos:

1. Abre Visual Studio Code:
2. Pulsa `Ctrl+Shift+X` o haz clic en el icono cuadrado de la barra lateral para abrir la vista Extensiones.
3. Escribe `GitHub Copilot` en la barra de búsqueda y localiza la extensión GitHub Copilot en los resultados de la búsqueda.
4. Haz clic en el botón verde `Install` situado junto a la extensión GitHub Copilot y espera a que finalice la instalación.

Una vez instalada, la extensión GitHub Copilot te pedirá que te autentiques con tus credenciales de GitHub. Sigue las instrucciones de la pantalla para iniciar sesión y conceder los permisos necesarios. Con ambas extensiones instaladas y configuradas, estás todo listo para desarrollar código en Visual Studio Code utilizando las funciones de Azure Databricks y las sugerencias basadas en IA de GitHub Copilot.

Escribir un script de Python con sugerencias basadas en IA

Sigue estos pasos para generar un script de Python con sugerencias basadas en IA:

1. Crea un nuevo archivo de Python.
2. En Visual Studio Code, haz clic en `File > New File.`
3. Guarda el archivo con una extensión `.py`: por ejemplo, `script.py` o `codehelping.py`.
4. Empieza a escribir tu código de Python. Cuando empieces a escribir, GitHub Copilot proporcionará automáticamente sugerencias de código en tiempo real.
5. Por ejemplo, escribe `How to create a Azure databricks dataframe in python.`

6. Una vez enviada la instrucción, verás un resultado similar al de la *figura 2*.

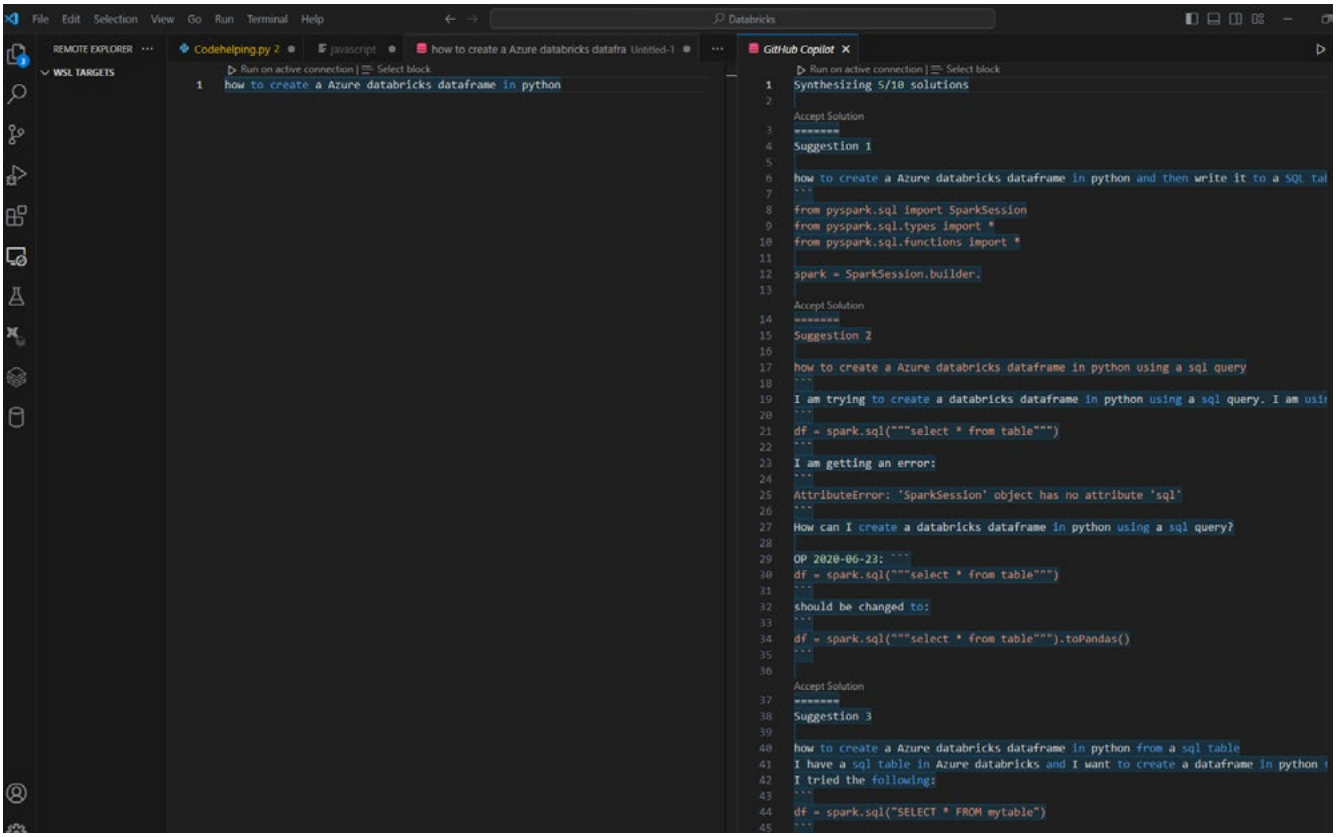

Figura 2: Script de Python generado por IA

Esto demuestra el poder de Azure Databricks como una plataforma líder para los datos y la IA cuando se combina con herramientas para desarrolladores como Visual Studio Code y GitHub Copilot.

Integración de Azure Databricks con OneLake

OneLake, dentro de Microsoft Fabric, actúa como una ubicación unificada para almacenar los datos de una organización. El poder de la arquitectura de lakehouse proviene de permitir que varios motores de análisis procesen los datos almacenados en OneLake. Estos motores podrían ser los que se incluyen en Fabric, como Synapse Data Warehouse. Sin embargo, otros servicios de Azure, como Azure Databricks, también pueden procesar los datos almacenados en OneLake. Esto permite a las organizaciones utilizar toda la potencia de Azure dentro del patrimonio de datos de Fabric.

El siguiente tutorial muestra cómo conectarse a OneLake a través de Azure Databricks. La conexión con OneLake permite leer y escribir datos en Microsoft Fabric Lakehouse desde un área de trabajo de Azure Databricks.

Para seguir este ejemplo, necesitarás:

- Un área de trabajo establecida de Fabric y Fabric Lakehouse
- Un área de trabajo premium de Azure Databricks

> **Nota:**
> Solo las áreas de trabajo premium de Azure Databricks admiten credenciales diferidas de Microsoft de Azure Active Directory, que son esenciales para esta integración.

Guía paso a paso

Sigue estos pasos para integrar Azure Databricks con OneLake:

1. Para empezar, deberás configurar tus clústeres de Azure Databricks para que usen las `credential passthrough`. Esto permite a Azure Databricks usar tu identidad de usuario al conectarte a una cuenta de Azure Storage, en este caso, tu cuenta de OneLake.

2. Abre tu área de trabajo de Azure Databricks. En el panel de la izquierda, selecciona `Compute`. A continuación, haz clic en el `botón Create compute`.

3. Sigue el procedimiento habitual para crear un nuevo clúster de Azure Databricks. Sin embargo, asegúrate de expandir la sección `Advanced options` y activar la opción `Enable credential passthrough for user-level data access`. Esto te permitirá utilizar la identidad de usuario de Microsoft para conectarte a OneLake.

4. Cuando hayas establecido todas las opciones de configuración deseadas, haz clic en `Create compute`.

5. Cuando el nuevo clúster esté listo, haz clic en `Workspace` en el panel del lado izquierdo de la pantalla. Haz clic en el botón `Add` en la esquina superior derecha de la pantalla y selecciona `Notebook` en el menú desplegable. Esto creará un nuevo cuaderno de Azure Databricks.

6. Cuando se abra el cuaderno, utiliza el menú desplegable de la parte superior del cuaderno para asegurarte de que estás conectado al nuevo clúster que acabas de crear.

7. Para continuar con el tutorial, necesitarás obtener los detalles de tu cuenta de almacenamiento de OneLake. Para encontrar esta información, dirígete a Fabric Lakehouse y localiza la ruta de Azure Blob Filesystem (ABFS). Esta ruta se puede encontrar en el panel `Properties` de la página del área de trabajo.

> **Nota:**
> Azure Databricks admite exclusivamente el controlador ABFS para las interacciones con ADLS Gen2 y OneLake. Una ruta de ejemplo podría ser la siguiente: `abfss://myWorkspace@onelake.dfs.fabric.microsoft.com/`.

8. Ahora estás listo para usar Azure Databricks para cargar datos en tu cuenta de almacenamiento de OneLake. Pega el siguiente código en una celda del cuaderno. Este código creará un nuevo conjunto de datos en Microsoft Fabric basado en los datos de ejemplo de Azure Databricks:

```
oneLakePath = 'abfss://myWorkspace@onelake.dfs.fabric.microsoft.
com/myLakehouse.lakehouse/Files/'

# Load data from a Databricks public dataset into a dataframe.
#Alternatively, you can source data from elsewhere in Fabric or
#from another ADLS Gen2 account you own. For example:

yellowTaxiDF = spark.read.format("csv").option("header", "true").
option("inferSchema", "true").load("/databricks-datasets/nyctaxi/
tripdata/yellow/yellow_tripdata_2019-12.csv.gz")

#Process your data as needed. This could involve filtering,
#joining with other datasets, or other transformations. For
#instance:

filteredTaxiDF = yellowTaxiDF.where(yellowTaxiDF.fare_amount<4).
where(yellowTaxiDF.passenger_count==4)
display(filteredTaxiDF)

#Write the processed dataframe to your Fabric Lakehouse using the
#OneLake path you saved earlier:

filteredTaxiDF.write.format("csv").option("header", "true").
mode("overwrite").csv(oneLakePath)

#To ensure your data was written successfully, read the newly
#loaded file and display a sample:

lakehouseRead = spark.read.format('csv').option("header", "true").
load(oneLakePath) display(lakehouseRead.limit(10))
```

Con este tutorial, has integrado correctamente Azure Databricks con OneLake en Fabric. Ahora puedes leer y escribir datos fácilmente en Fabric con Azure Databricks.

Conclusión

La administración y el análisis de datos eficaces son primordiales para que las empresas obtengan una ventaja competitiva. Microsoft Fabric es una solución crucial que ofrece una plataforma unificada, escalable y segura diseñada pensando en las necesidades de las empresas modernas. Esta plataforma no solo garantiza una integración y administración de datos fluidas, sino que también favorece la colaboración, refuerza la seguridad e impulsa la innovación.

Pasos siguientes

- **Obtén más información sobre Microsoft Fabric y comienza con una versión de evaluación gratuita.**
- **Permite casos de uso de datos, análisis e IA con Azure Databricks.**
- **Aprende a crear un lakehouse con OneLake en Fabric.**