

Clase 10. Bootcamp Bases de Datos en la Nube

Exploración de analítica de datos en tiempo real

Luis Beltrán
Microsoft MVP



Hola soy...

¡Hola a todos! Antes de sumergirnos en esta presentación, quiero tomarme un momento para hablar un poco sobre mí. [Soy Luis Beltrán](#), y he trabajado en algunos proyectos tales como desarrollo de aplicaciones móviles, soluciones informáticas de Inteligencia Artificial alojadas en la nube, bases de datos, consultoría de TI y capacitaciones oficiales de certificación de Microsoft.

Mi pasión por las bases de datos es contagiosa, así que prepárense para unirnos en un emocionante bootcamp juntos.



Objetivos de la clase/Intro

Hoy daremos un repaso sobre las opciones de Analítica de datos en Azure y posteriormente procederemos a conocer en qué consiste el análisis de datos en tiempo real con Apache Spark y Synapse. Además, estudiaremos un poco de Fabric.





Tabla de contenido

Clase #10

/01

Repaso: Análisis de datos en Azure

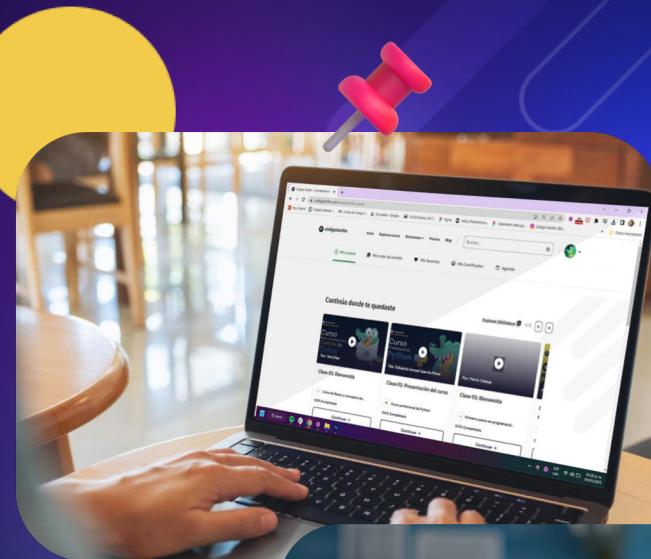
Opciones para analítica de datos

/02

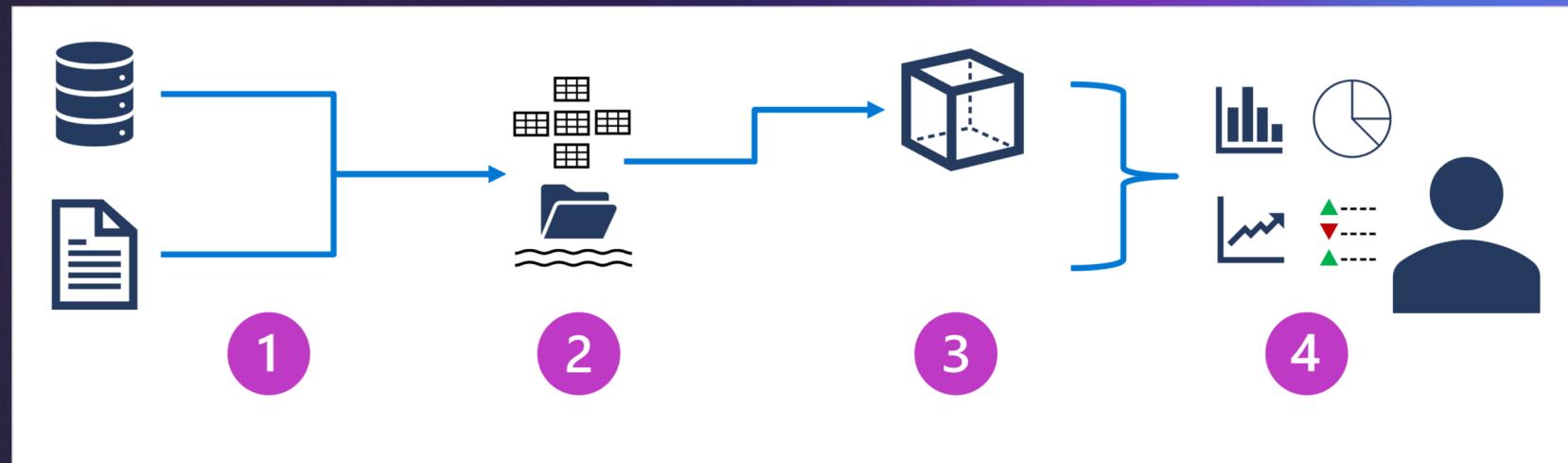
Streaming y análisis en tiempo real

Con Azure Stream Analytics,
Spark, Synapse y Fabric

/01 Repaso: Análisis de datos en Azure



Elementos de una solución de análisis de datos a gran escala



Ingestión y
procesamiento de
datos

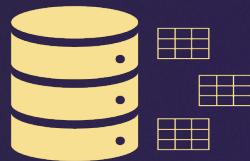
Almacén de datos
analíticos

Modelo de datos
analíticos

Visualización de
datos

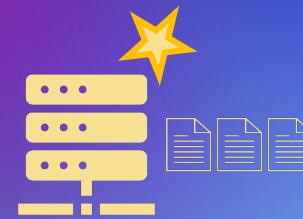


Procesamiento de datos en análisis a gran escala



Base de datos relacional

- Modelo bien establecido para el almacenamiento y el procesamiento de datos relacionales.
- Compatibilidad total con el lenguaje SQL para consultas y la manipulación de datos.



Apache Spark

- Plataforma de código abierto para el procesamiento de datos escalable y distribuido.
- Código de procesamiento de datos multilenguaje (Python, Scala, Java, SQL...).

Arquitecturas de almacén de datos analíticos



Data Warehouse

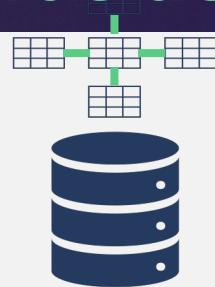
- Modelo bien establecido para el almacenamiento y el procesamiento de datos relacionales.
- Compatibilidad total con el lenguaje SQL para consultas y la manipulación de datos.



Almacén de lago de datos

- Plataforma de código abierto para el procesamiento de datos escalable y distribuido.
- Código de procesamiento de datos multilenguaje (Python, Scala, Java, SQL...).

Arquitecturas de almacén de datos analíticos



Data Warehouse

- Los datos se almacenan en una base de datos relacional y se consultan mediante un motor de consultas SQL.
- Los datos se *desnormalizan* para optimizar las consultas
 - Normalmente como un esquema de estrella o copo de nieve de *hechos* numéricos que se pueden agregar mediante *dimensiones*.



Arquitecturas de almacén de datos analíticos



Almacén de lago de datos

- Los archivos de datos se almacenan en un sistema de archivos distribuido (un *lago de datos*) y se suelen procesar con Apache Spark.
- Los metadatos se usan para definir tablas que proporcionan una interfaz SQL relacional a los datos del archivo.
 - Normalmente se usa un formato *delta lake* para proporcionar funcionalidad de base de datos transaccional.

Servicios de análisis de datos PaaS



Azure Synapse Analytics

- Solución unificada para el almacenamiento de datos relacionales y el análisis del lago de datos
- Procesamiento escalable y consulta mediante varios entornos de ejecución de análisis
 - SQL de Synapse
 - Apache Spark
 - Synapse Data Explorer
- Experiencia interactiva en Azure Synapse Studio
- Integración de canalización integrada para la ingestión y el procesamiento de datos

Úselo para una única solución analítica unificada a gran escala en Azure

Servicios de análisis de datos PaaS



Azure Databricks

- Implementación en Azure de la plataforma de análisis en la nube Databricks
- Spark y consulta de SQL escalables para el análisis del lago de datos
- Experiencia interactiva en el área de trabajo de Azure Databricks
- Use Azure Data Factory para implementar canalizaciones de ingestión y procesamiento de datos

Úsalo para aprovechar las aptitudes de Databricks y para la portabilidad en la nube

Servicios de análisis de datos PaaS



Azure HDInsight

- Implementación en Azure de frameworks de "macrodatos" comunes de Apache basados en un lago de datos
 - Hadoop: consulta de archivos de lago de datos mediante tablas de Hive
 - Spark: use las API de Spark para consultar datos y abstraer el almacenamiento de archivos subyacente como tablas
 - Kafka: procesamiento de eventos en tiempo real
 - Storm: procesamiento de flujos
 - HBase: almacén de datos NoSQL

Úselo cuando necesite admitir varias plataformas de código abierto



Microsoft Fabric



Datos
Integración
Data Factory



Datos
Ingeniería
Synapse



Datos
Almacén
Synapse



Datos
Ciencia
Synapse



Real-Time
Analytics
Synapse



Business
Intelligence
Power BI



Observabilidad
aplicada
Data Activator



Base de datos unificada
OneLake

Unificado

Experiencia en el
producto SaaS

Seguridad y
gobernanza

Proceso

Almacenamiento

Empresas
modelo



códigofacilito

Exploración del análisis de datos en Microsoft Fabric



1

¿Qué servicios de Azure puede usar para crear una canalización para la ingestión y el procesamiento de datos?

- Azure SQL Database y Azure Cosmos DB
- Azure Synapse Analytics y Azure Data Factory
- Azure HDInsight y Azure Databricks

2

¿Qué debe definir para implementar una canalización que lea datos de Azure Blob Storage?

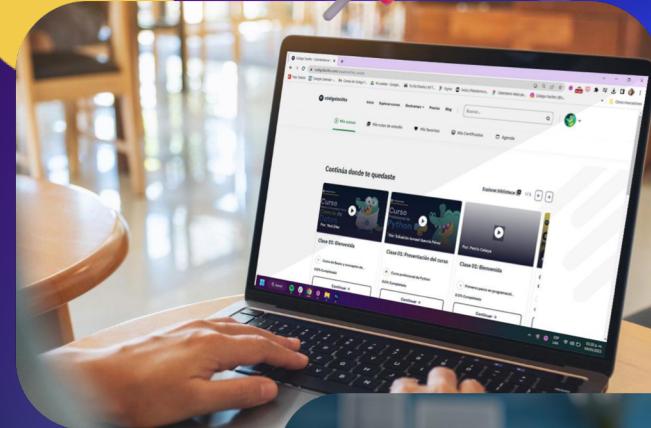
- Un servicio vinculado para la cuenta de Azure Blob Storage
- Un grupo de SQL dedicado en el área de trabajo de Azure Synapse Analytics
- Un clúster de Azure HDInsight en la suscripción

3

¿Qué motor de procesamiento distribuido de código abierto incluye Azure Synapse Analytics?

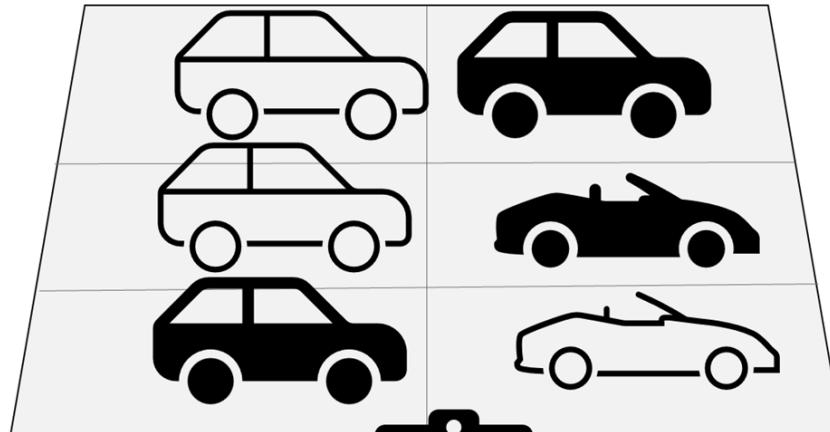
- Apache Hadoop
- Apache Spark
- Apache Storm

/02 Streaming y análisis en tiempo real



Procesamiento por lotes

Procesamiento por lotes

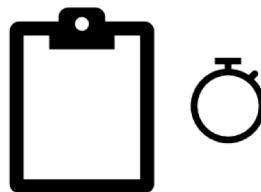


Los datos se recopilan
y procesan a intervalos
regulares

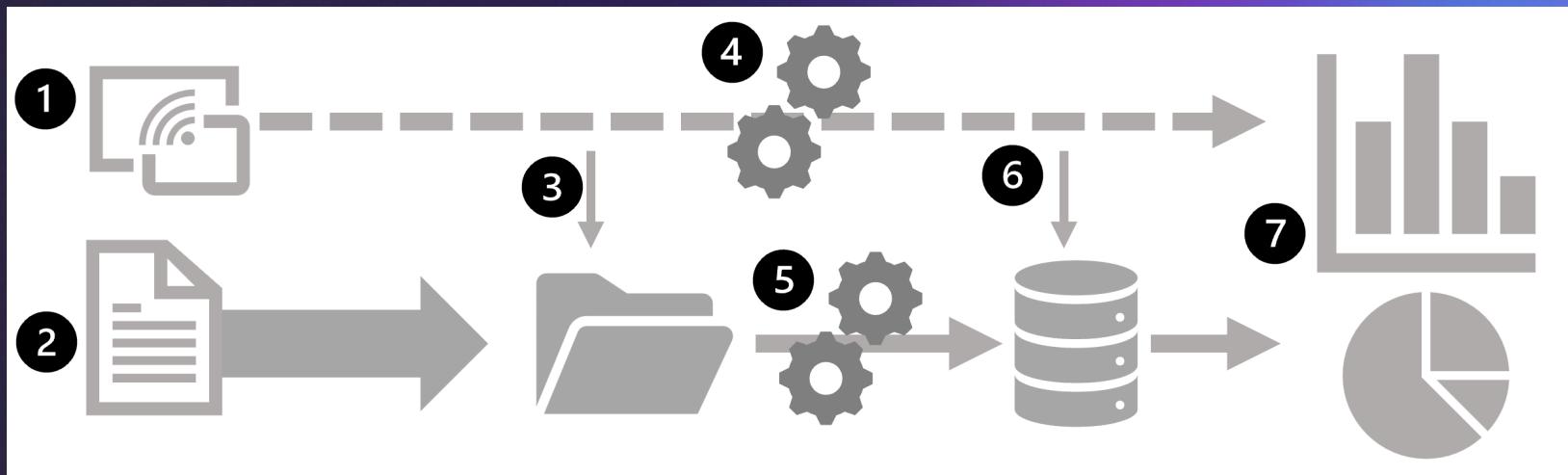
Procesamiento de flujos (streams)

Procesamiento de flujos

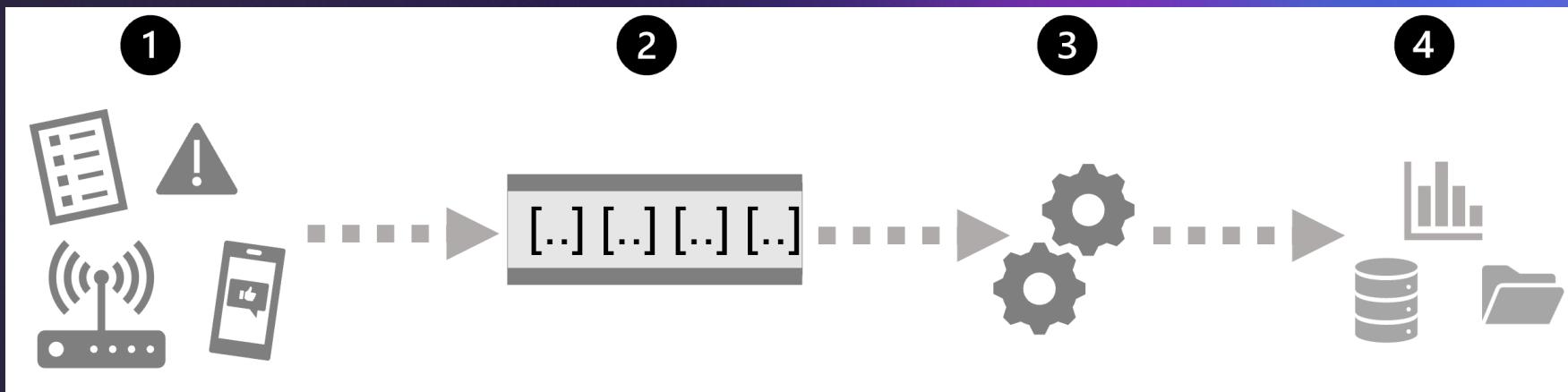
Los datos se
procesan en
tiempo casi real
conforme llegan



Combinación



Una arquitectura general para el procesamiento de flujos



Servicios de análisis en tiempo real

- Azure Stream Analytics
- Spark Structured Streaming
- Microsoft Fabric

Orígenes para el procesamiento de flujos

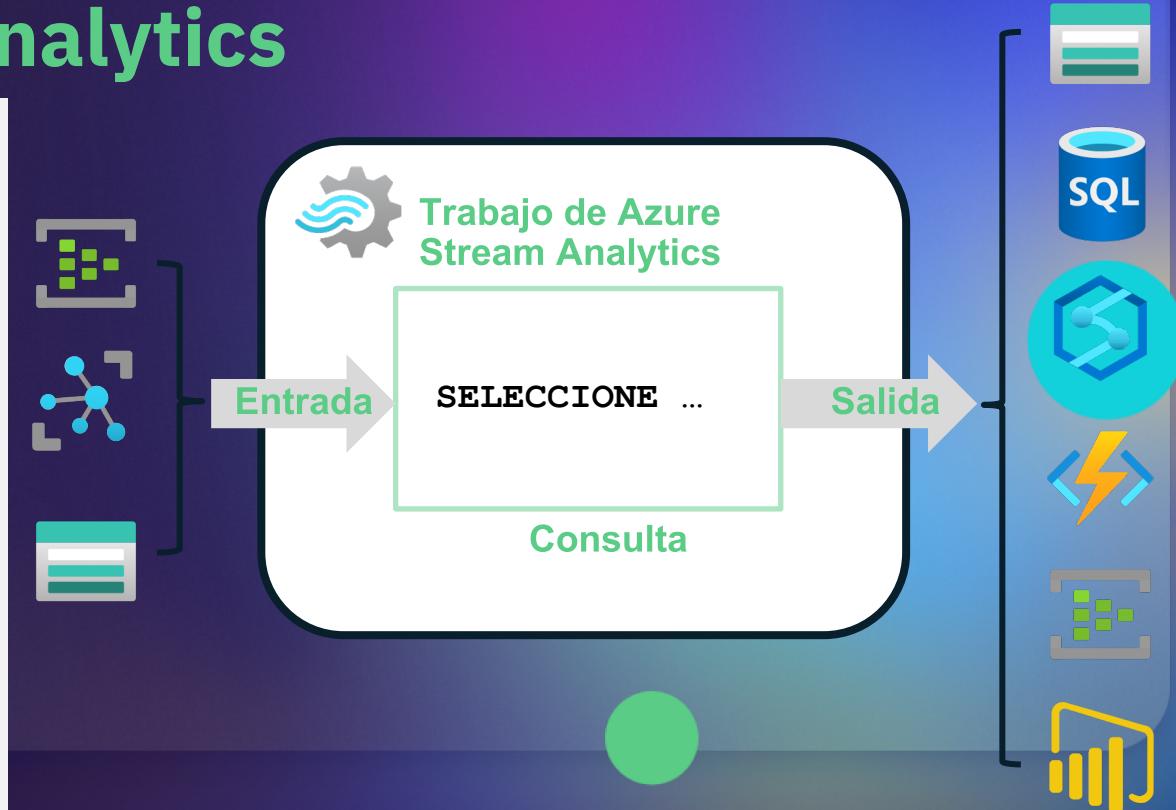
- Azure Event Hubs
- Azure IoT Hub
- Azure Data Lake Store Gen 2
- Apache Kafka

Receptores para el procesamiento de flujos

- Azure Event Hubs
- Azure Data Lake Store Gen 2, Microsoft OneLake o Azure blob Storage
- Azure SQL Database, Azure Databricks o Microsoft Fabric
- Microsoft Power BI

Procesamiento de datos en tiempo real con Azure Stream Analytics

- Cree un *trabajo* o un *clúster* de Azure Stream Analytics.
- Ingiera datos de una *entrada* (Azure Event Hubs, Azure IoT Hub, Azure Blob Storage container)
- Procese datos con una *consulta perpetua*.
- Envíe los resultados a una *salida* (Azure Blob Storage, Azure SQL Database, Azure Synapse Analytics, Azure Functions, Azure Event Hubs, Power BI).



Análisis de telemetría y registro en tiempo real con Azure Data Explorer

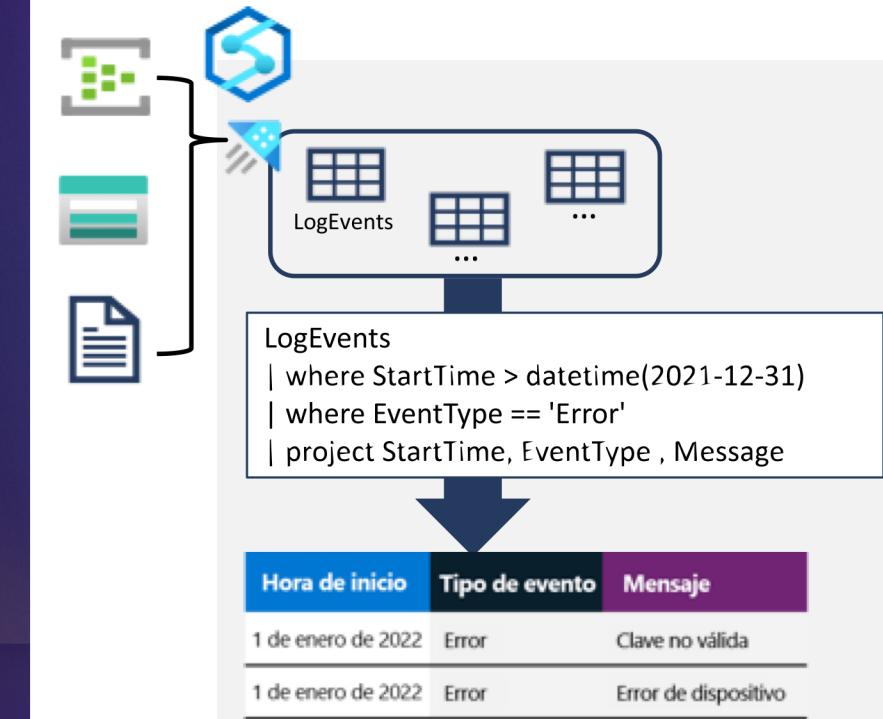
Alto rendimiento, servicio escalable para datos por lotes y de streaming

- **Servicio** Azure Data Explorer dedicado
- **Entorno de ejecución de** Azure Synapse Data Explorer en Azure Synapse Analytics

Los datos se ingieren desde el streaming y los orígenes por lotes en tablas de una base de datos

Las tablas se pueden consultar con Kusto Query Language (KQL):

- Sintaxis intuitiva para consultas de solo lectura. Optimizado para datos de serie temporal y telemetría sin procesar



Análisis con streaming estructurado de Apache Spark (Spark Structured Streaming)



Para procesar los datos de flujos en Spark, puede usar la biblioteca de **Spark Structured Streaming**.

- Proporciona una API para ingerir, procesar y generar resultados de flujos de datos perpetuos.

Spark Structured Streaming se compila en una estructura denominada **dataframe**.

Lectura de datos de un origen de datos en tiempo real, como un centro de Kafka, un almacén de archivos o un puerto de red, a un dataframe "sin límite" que se rellena continuamente con nuevos datos del flujo.

Defina una consulta en el dataframe que selecciona, proyecta o suma los datos.

Los resultados de la consulta generan otro dataframe.

Delta Lake

Delta Lake es una capa de almacenamiento de código abierto que agrega compatibilidad con la coherencia transaccional, el cumplimiento del esquema y otras características de almacenamiento de datos a Data Lake Storage.

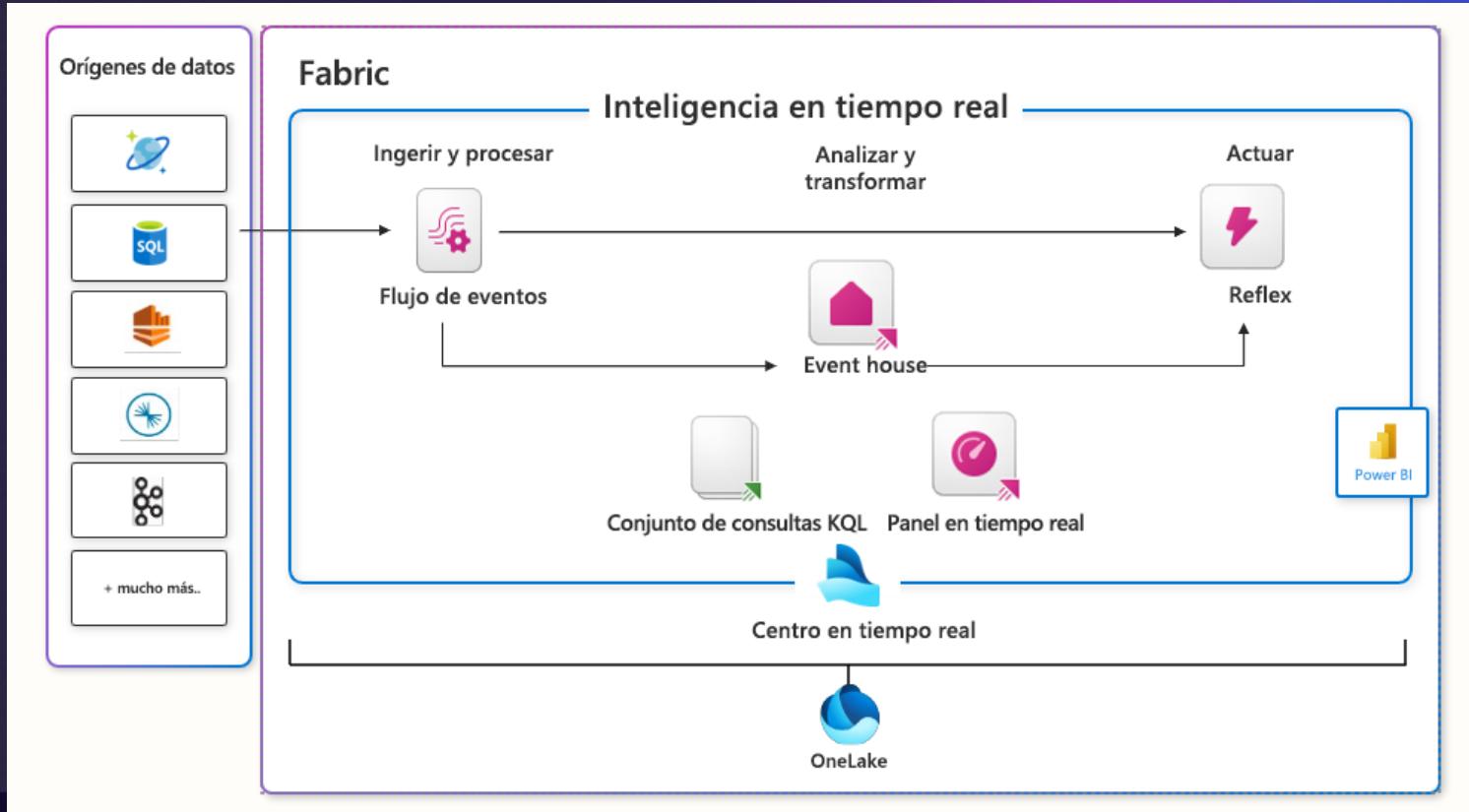
Unifica el almacenamiento para datos por lotes y de flujos.

Se puede usar en Spark para definir tablas relacionales para el procesamiento por lotes y de flujos.

Cuando se usa para el procesamiento de flujos, una tabla de Delta Lake se puede usar como un origen de flujos para las consultas en datos en tiempo real o como un receptor en el que se escribe un flujo de datos.

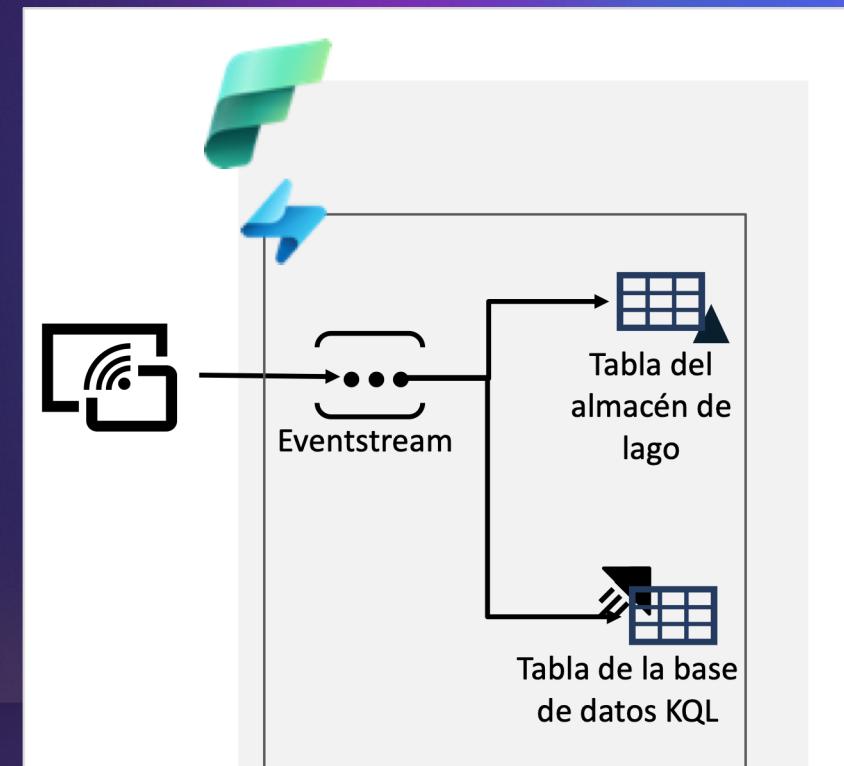
Los runtimes de Spark en Microsoft Fabric y Azure Databricks incluyen compatibilidad con Delta Lake.

Delta Lake + Structured Streaming de Spark = solución óptima cuando es necesario abstraer los datos procesados por lotes y flujos en un lago de datos detrás de un esquema relacional para realizar consultas y análisis basados en SQL.



Análisis en tiempo real en Microsoft Fabric

- Compatibilidad con la ingestión continua de datos de varios orígenes.
- Captura de datos de streaming en un **Eventstream**.
- Escritura de datos en tiempo real en una tabla en un almacén de lago o una base de datos KQL.
- Consulta de datos en tiempo real con SQL o KQL.
- Creación de visualizaciones en tiempo real.





códigofacilito

Exploración del análisis en tiempo real en Microsoft Fabric



1 ¿Qué definición de *procesamiento de flujos* es correcta?

- Los datos se procesan continuamente a medida que llegan nuevos registros de datos
- Los datos se recopilan en un almacén temporal y todos los registros se procesan de forma conjunta como un lote
- Los datos están incompletos y no se pueden analizar

2 ¿Qué servicio usaría para capturar datos continuamente de una instancia de IoT Hub, agregarlos por períodos temporales y almacenar los resultados en Azure SQL Database?

- Azure Cosmos DB
- Azure Stream Analytics
- Azure Storage

3

¿Qué lenguaje usaría para consultar datos de registro en tiempo real en Azure Synapse Data Explorer?

- SQL
- Python
- KQL

¿Qué has aprendido hoy?

El procesamiento en tiempo real es un elemento común de las soluciones de análisis de datos empresariales. Microsoft Azure ofrece una variedad de servicios que puede usar para implementar el procesamiento de flujos y el análisis en tiempo real.



En este módulo, has aprendido sobre:

- Comparación del procesamiento por lotes y por flujos
- Descripción de elementos comunes de las soluciones de datos de flujos
- Descripción de las características y funciones de Azure Stream Analytics
- Descripción de las características y funcionalidades de la inteligencia en tiempo real de Microsoft Fabric
- Descripción de las características y funciones de Spark Structured Streaming en Azure



Directorio de links

Exploración de los aspectos básicos del análisis a gran escala
<https://learn.microsoft.com/es-mx/training/modules/examine-components-of-modern-data-warehouse/>

Exploración de los aspectos básicos del análisis en tiempo real
<https://learn.microsoft.com/es-mx/training/modules/explore-fundamentals-stream-processing/>





Directorio de links



Ingest realtime data with Azure Stream Analytics and Azure Synapse Analytics

<https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/18-Ingest-stream-synapse.html>

Create a realtime report with Azure Stream Analytics and Microsoft Power BI

<https://microsoftlearning.github.io/dp-203-azure-data-engineer/Instructions/Labs/19-Stream-Power-BI.html>

¡Gracias por tu atención!

Luis Beltrán
Microsoft MVP
about.me/luis-beltran

