28-11-2025

TRABAJO DE ARQUITECTURA DE DATOS

XINYUAN ZHENG

ÍNDICE

1. Introducción

El objetivo principal de este trabajo es diseñar una arquitectura de datos adecuada para las necesidades planteadas por FarmIA, una compañía dedicada a la venta de productos agrícolas.

Se busca definir una solución en la nube que permita gestionar de forma escalable, flexible y eficiente el almacenamiento, procesamiento y análisis de los datos generados por la organización, incluyendo información de ventas en línea, inventarios, sensores instalados en los campos agrícolas y datos de clientes en tiempo real.

Ante la expansión prevista del negocio y el crecimiento del volumen de datos en los próximos 2 años, resulta esencial proponer una arquitectura capaz de centralizar estas fuentes de información en un entorno unificado, garantizando su integración con nuevas aplicaciones, plataformas móviles o canales de venta de terceros. De este modo, se facilitará el desarrollo de capacidades analíticas que mejoren la toma de decisiones y optimicen los procesos logísticos y comerciales de FarmIA.

2. Requisitos

A partir del análisis del contexto de FarmIA y sus necesidades de crecimiento, se identifican los requisitos clave que debe cumplir la arquitectura de datos. Estos se agrupan en tres dimensiones principales: escalabilidad, integración y capacidad analítica, y servirán como base para la selección de tecnologías y el diseño final de la solución.

2.1 Escalabilidad

FarmIA prevé un incremento significativo en su volumen de datos, estimado en más del triple durante los próximos dos años. Aunque este crecimiento será progresivo y limitado al mercado español, la arquitectura debe garantizar:

- Escalado gradual en almacenamiento y cómputo conforme aumenten las necesidades del negocio.
- Mantenimiento del rendimiento ante picos de actividad, especialmente durante campañas o incrementos en los datos IoT.
- Gestión de mayores volúmenes sin requerir infraestructuras sobredimensionadas o propias de entornos con crecimiento internacional acelerado.

2.2 Integración

La compañía trabaja actualmente con una plataforma de ventas online y un sistema local de inventarios, pero incorporará sensores agrícolas, datos de clientes en streaming, aplicaciones móviles y canales de terceros. La arquitectura debe:

- Integrar de forma eficiente fuentes heterogéneas, tanto estructuradas como no estructuradas.
- Soportar ingesta en tiempo real y en procesos batch.
- Conectarse con sistemas internos, servicios cloud, APIs externas y aplicaciones móviles.
- Mantener un diseño modular que permita añadir futuras fuentes sin rediseñar toda la solución.

2.3 Capacidad de análisis

Para mejorar la toma de decisiones, FarmIA necesita capacidades analíticas avanzadas a nivel operativo y estratégico. La arquitectura debe:

- Permitir análisis en tiempo real para monitorizar inventarios, comportamiento de clientes y métricas logísticas.
- Soportar procesos batch para reporting, consolidación y análisis históricos.
- Facilitar el desarrollo futuro de modelos de machine learning basados en datos IoT y comportamiento de usuarios.
- Garantizar un acceso seguro y eficiente a herramientas de visualización y cuadros de mando.

3. Arquitectura seleccionada

Existen múltiples enfoques modernos para el diseño de plataformas de datos, como Data Warehouses, Data Lakes o arquitecturas Lakehouse; que pueden adaptarse a diferentes necesidades empresariales.

Tras analizar las características de cada modelo y compararlas con los requisitos de FarmIA, se selecciona una arquitectura Lakehouse en Azure. Este enfoque combina almacenamiento escalable, integración de datos heterogéneos y capacidades analíticas tanto batch como en tiempo real, lo que lo convierte en una opción idónea para el caso descrito.

A continuación, se describen las capas y componentes principales que conforman la propuesta:

3.1 Capa de Ingesta

La ingesta se adapta a la naturaleza de cada fuente de datos:

- **Azure Data Factory**: cargas batch desde ventas online, inventarios y APIs de terceros.
- **Azure IoT Hub / Event Hubs**: recepción de eventos en tiempo real procedentes de sensores agrícolas y actividad de clientes.
- **Azure API Management**: integración con aplicaciones móviles y futuros partners externos.

3.2 Capa de Almacenamiento:

El repositorio central se implementa sobre **Azure Data Lake Storage Gen2**, organizado mediante el modelo **Medallion Architecture**:

- **Bronze**: datos en bruto procedentes de cargas batch y flujos de eventos.
- **Silver**: datos limpios y normalizados tras procesamiento en Databricks.
- **Gold**: datos preparados para análisis y modelos de negocio.

3.3 Capa de Procesamiento Batch y Streaming

El procesamiento combina cargas periódicas y flujos continuos:

- **Azure Databricks (Delta Lake)**: ejecución de procesos ETL/ELT, transformaciones y modelos analíticos.
- **Delta Live Tables**: definición de pipelines declarativos y control de calidad.

3.4 Capa Analítica

Esta capa permite el acceso a la información procesada:

- **Azure Synapse / Databricks SQL**: consultas analíticas e informes avanzados.
- **Power BI**: visualización y cuadros de mando operativos o estratégicos.
- **Databricks ML**: desarrollo de modelos predictivos basados en datos históricos y en streaming.

3.5 Capa de Servicio y Consumo

Proporciona acceso seguro y gobernado a los datos:

- **Power BI Service**: publicación de dashboards y actualización programada.
- **SQL Endpoints**: consultas ad hoc para equipos analíticos.
- **API Management**: exposición controlada de datos a sistemas y partners.
- **Model Serving (Databricks)**: consumo de modelos de ML en tiempo real o batch.

4. Justificación

La arquitectura Lakehouse propuesta es la opción más adecuada para FarmIA porque responde de manera directa y coherente a los requisitos de escalabilidad, integración y capacidad analítica definidos previamente. A continuación, se detalla cómo cada uno de estos aspectos queda cubierto por los componentes seleccionados.

4.1 Escalabilidad

La arquitectura soporta el crecimiento progresivo previsto por FarmIA mediante el uso de Azure Data Lake Storage Gen2, que ofrece almacenamiento prácticamente ilimitado, y Azure Databricks, que ajusta la capacidad de cómputo de forma elástica en función de la demanda. Además, servicios como Event Hubs permiten absorber picos de datos procedentes de sensores IoT sin necesidad de sobredimensionar la infraestructura. Esto garantiza que la plataforma pueda evolucionar al ritmo del negocio sin requerir rediseños futuros.

4.2 Integración

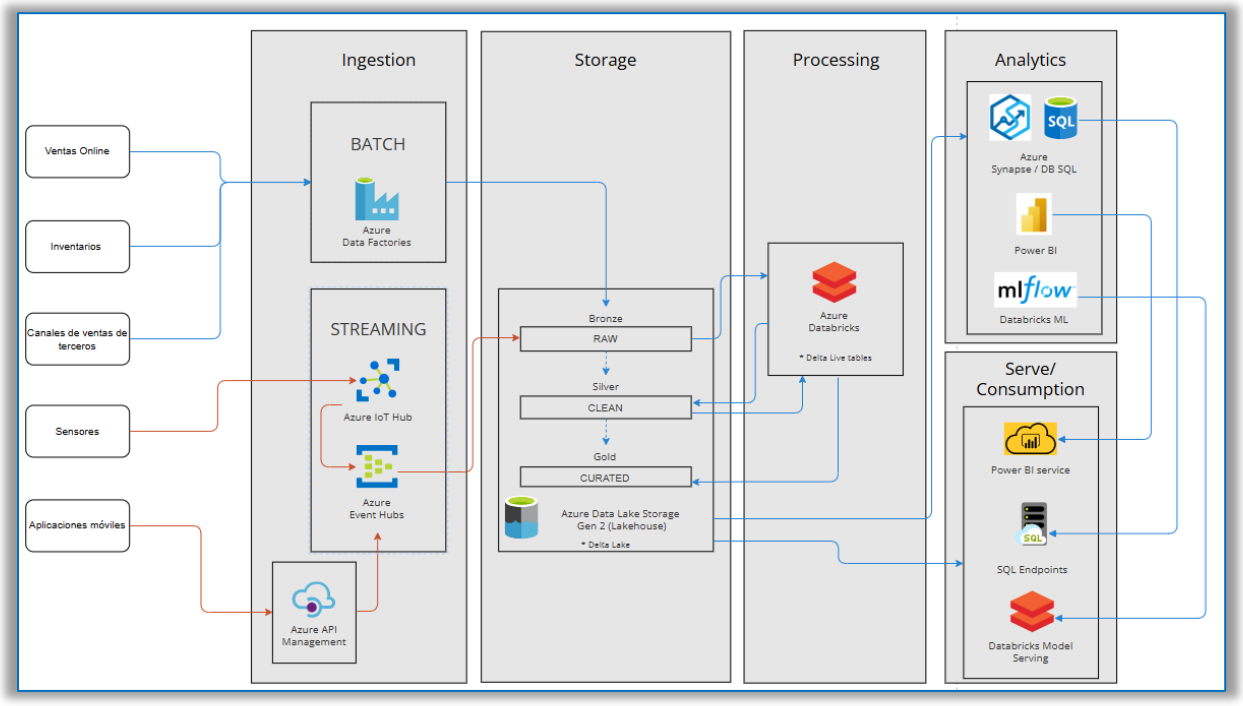
La solución facilita la integración de datos procedentes de múltiples orígenes gracias a la combinación de herramientas específicas para distintos ritmos y formatos: Data Factory para cargas batch, IoT Hub/Event Hubs para ingesta en tiempo real y API Management para aplicaciones móviles o socios externos. El uso de ADLS como repositorio central unifica datos estructurados y no estructurados en un entorno flexible, escalable y fácilmente ampliable conforme evolucionen las necesidades de FarmIA.

4.3 Capacidad Analítica

La arquitectura habilita tanto análisis operativos como estratégicos. Azure Databricks permite procesar datos en tiempo real y en modo batch, generando las capas Silver y Gold del Lakehouse. Los equipos analíticos pueden realizar consultas avanzadas desde Synapse SQL o Databricks SQL y visualizar los resultados a través de Power BI. Además, Databricks incluye capacidades nativas para el desarrollo de modelos predictivos, lo que asegura que la empresa pueda extraer valor inmediato y también evolucionar hacia casos de uso de analítica avanzada y machine learning.

5. Diagrama

El diagrama presentado resume visualmente la arquitectura Lakehouse diseñada para FarmIA, mostrando el flujo completo desde la ingesta de datos hasta su consumo final. Se representan las distintas fuentes de información, los mecanismos de ingesta batch y streaming, y el almacenamiento unificado en Azure Data Lake Storage Gen2 estructurado según el modelo Bronze–Silver–Gold. El procesamiento se centraliza en Azure Databricks mediante Delta Lake, permitiendo transformar, depurar y preparar los datos para su explotación.



Asimismo, se incluyen las capas analítica y de servicio, donde herramientas como Azure Synapse, Power BI y Databricks Model Serving habilitan consultas, visualizaciones y despliegue de modelos predictivos. El esquema refleja de forma clara cómo los diferentes componentes se integran para ofrecer una plataforma escalable, flexible y orientada al análisis.

6. Conclusiones

La arquitectura propuesta ofrece a FarmIA una solución moderna, eficiente y ajustada a sus necesidades reales de crecimiento. Dado que la empresa opera exclusivamente en el mercado español y la previsión de incremento de datos se sitúa en un horizonte de dos a tres años, no resulta necesario adoptar infraestructuras excesivamente complejas o pensadas para escenarios de demanda global. El enfoque Lakehouse basado en Delta Lake permite escalar de forma gradual, absorber nuevas fuentes de datos y acompañar la evolución del negocio sin introducir sobrecostes ni sobredimensionamiento innecesario.

Con ello, FarmIA obtiene una plataforma flexible y preparada para una expansión sostenible, reduciendo costes operativos y reforzando la capacidad de tomar decisiones basadas en datos en el corto y medio plazo.