Trabajo Fin de Máster :
Pipeline de Datos Ferroviarios en Azure

Opción 2: Preparación de un pipeline de datos.

Tecnologías utilizadas: Preparación y Explotación de Datos con Spark, Cosmos DBy FastAPI

Autor: Judith Castillo Martínez

Introducción

Desarrollo de un pipeline de datos ferroviarios en Azure.

Objetivo: preparar y disponibilizar datos.

Tecnologías: Spark (PySpark), Databricks, Azure, Cosmos DB, FastAPI.

Configuración inicial

- Configuración del espacio en Azure
- Data Lake
- Clústeres de Databricks
- Cosmos DB (API MongoDB)
- Entorno seguro, escalable y listo para ejecutar el pipeline.

Estructura del pipeline

El pipeline se compone de 4 grupos de notebooks:

- 1. Autodescarga
- 2. Ingesta (RAW)
- 3. Limpieza (CLEAN)
- 4. Publicación (Cosmos DB).

1. Autodescarga

SCRIPT PARA DESCARGAR AUTOMÁTICAMENTE LOS FICHEROS GTFS.

SE ALMACENAN EN EL DATA LAKE SIN INTERVENCIÓN MANUAL.

2. Ingesta en RAW

Los datos se guardan en la capa RAW en formato Delta.

Se conserva el formato original con versionado y trazabilidad.

3. Limpieza en CLEAN

- 12 notebooks de transformación.
- Reglas de calidad
- Eliminación de duplicados
- Validación de horarios
- Estandarización de nombres.
- Resultado: datos coherentes y listos para análisis.

4. Publicación en Cosmos DB

Carga de datos limpios en Cosmos DB (API MongoDB).

Se generan identificadores únicos.

Las tablas se exponen como colecciones listas para consulta.

Job en Databricks (Orquestación)

Se creó un **Job en Databricks** que encadena las etapas en orden: autodescarga → ingesta → limpieza → carga.

Este Job asegura la **ejecución secuencial, trazabilidad y control de errores**, facilitando la operación y la repetibilidad del pipeline.

Explotación con FastAPI

- API desarrollada en FastAPI sobre las colecciones de Cosmos DB.
- Endpoints disponibles: rutas, viajes, paradas, tiempos de paso.
- Despliegue en Azure App Service.
- Consumo desde aplicaciones o herramientas de BI.

Conclusión

Pipeline en Azure con ciclo completo de datos.

Desde ingesta automática hasta exposición en API moderna.

Flujo orquestado, escalable y alineado con el negocio ferroviario.