Taller Evaluado 01 – Ingesta y Validación de Datos en Azure Databricks

Sesiones relacionadas:

- Sesión 2: Mecanismos de ingesta en Azure Databricks
- Sesión 3: Validación de esquemas y manejo de errores

Duración estimada: 3 horas (ejercicio práctico evaluado)

Objetivo del taller

Que los estudiantes pongan en práctica y refuercen lo aprendido en las sesiones 2 y 3, aplicando técnicas de ingesta de datos, validación de esquemas, manejo de errores y registro de auditoría, utilizando un dataset de origen real o simulado.

Al finalizar, el estudiante será capaz de:

1. Seleccionar e implementar un mecanismo de ingesta adecuado (Auto Loader) para un dataset dado.
2. Aplicar un esquema explícito (StructType) para controlar la estructura de los datos.
3. Detectar y separar registros válidos e inválidos usando _rescued_data o badRecordsPath.
4. Registrar los registros inconsistentes en una **tabla Delta de auditoría** usando saveAsTable().
5. Documentar el proceso y los hallazgos en un notebook bien estructurado.

Descripción de la actividad

Cada estudiante deberá:

1. Escoger un dataset de origen (puede ser un archivo CSV, JSON o Parquet disponible en un volumen de Unity Catalog o cargado manualmente).
2. Diseñar y ejecutar un flujo de ingesta hacia Databricks utilizando **Auto Loader** según el formato y el volumen del archivo.
3. Definir un esquema explícito con StructType y aplicarlo durante la carga.
4. Configurar un badRecordsPath y capturar información de registros inválidos.
5. Separar datos válidos e inválidos en tablas Delta diferentes.

6. Agregar una columna de auditoría con la ruta de archivo de origen (_metadata.file_path).
7. Guardar los registros inválidos en una **tabla Delta de auditoría** usando saveAsTable().
8. Documentar las decisiones tomadas, las dificultades encontradas y los resultados obtenidos.

Requisitos técnicos

- Workspace de Databricks free edition activo y acceso a Unity Catalog.
- Dataset de origen accesible desde el entorno de Databricks.
- Permisos para crear tablas en el esquema de trabajo.

Entregable

- Un notebook en Databricks con:
    - Código ejecutado y validado.
    - Resultados de consultas a las tablas válidas e inválidas.
    - Creación de la tabla Delta de auditoría.
    - Imágenes incrustadas mostrando las tablas Delta creadas y su contenido.
    - Explicación escrita de los pasos realizados y conclusiones.
- Versión exportada del notebook en formato **HTML**.
- Tablas Delta creadas en el esquema asignado.

💡 **Recomendaciones**

- Probar el flujo con un subconjunto de datos antes de la ejecución final.
- Usar rutas controladas para badRecordsPath y carpetas de checkpoint.
- Verificar el conteo de registros entre origen y destino.
- Validar que la tabla de auditoría contenga exclusivamente registros inconsistentes.

- Realizar capturas de pantalla o exportar vistas de Spark UI si es relevante para la explicación.