CESTE

Escuela Internacional de Negocios

Zaragoza (España)

¿Quién soy?

- Pablo Sanz Caperote
- Doble Grado en Ingeniería Informática – Matemáticas
- Profesional con más de 3 años de experiencia entorno al mundo de los datos.
- Casi 2 años trabajando con Databricks
- Varias certificaciones en Clouds (AWS, Azure, GCP)

www.linkedin.com/in/pablosanzcaperote

CESTE

Índice del curso

1. Ingeniería de Datos y Pipeline ETL con Delta Lake
    - ETL
    - Delta Tables
    - Pipelines

2. Manejo de Permisos, Workflows y Jobs en entornos productivos
    - Seguridad y permisos: Unity Catalog
    - Herramientas Orquestación

3. Machine Learning en Producción con MLflow y Databrick
    - Construction de modelos
    - Modelos en Databricks con MLflow
    - Despliegue

Recapitulación | Nociones básicas Databrick Bloque I

1. dbutils

2. Catalog > Schema (Database) > Table

3. Spark:
    1. read_csv
    2. withColumn()
    3. filter()
    4. groupBy()

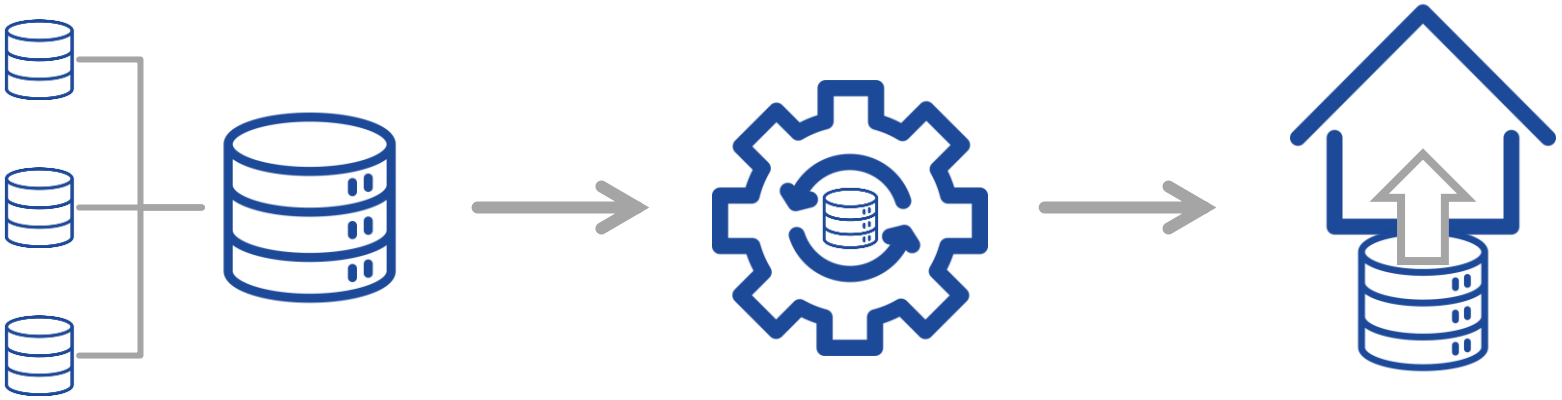
1. Ingeniería de Datos y Pipeline ETL con Delta Lake

Índice | Ingeniería de Datos y Pipeline ETL con Delta Lake

Objetivos | Construir ETLs basadas en Delta Tables

1. Comprender a construir ETL en Databricks
2. Aprender a trabajar con Delta Tables
3. Aplicar buenas prácticas de ingeniería de datos para entornos de trabajo escalables y colaborativos

ETL | Extract Transform Load



Extract

Recolecta y valida los datos de diferentes fuentes

Transform

Procesa y organiza los datos extraídos para convertirlos en útiles

Load

Mueve los datos transformados a un repositorio

ETL | Tipos de Extracciones

Carga Total

- Carga inicial
- Tablas fijas
- Migraciones, datasets históricos, …

Carga incremental

- Datos nuevos desde la última carga
- Columna de partición
- Cargas diarias, semanales, …

Carga en tiempo real

- Los datos se procesan y almacenan en cuanto están disponibles
- Herramientas específicas

En la parte práctica veremos los diferentes orígenes que existen (csv, json, bases de datos)

ETL | Tipos de Transformaciones

Limpieza de Datos

- Eliminación valores nulos
- Eliminación duplicados

Filtrado

- Filtrado por reglas
- Filtrado por tiempo

Unificación

- Conversión de tipos
- Mapeo de columnas

Operaciones

- Agregaciones
- Normalización
- Estandarización

Anonimización

- Protección de datos sensibles

ETL | Tipos de Cargas (Load)

Base datos relacionales
- Datos estructurados
- MySQL, PostgreSQL

Base datos NoSQL
- Datos semi-estructurados o no estructurados
- Clave-valor (Redis)
- Documentos (MongoDB)
- Grafos (Neo4j)

Data Lakes
- Grandes volúmenes de datos en forma “cruda”
- AWS S3, Azure Data Lake

Data Warehouse
- Grandes volúmenes de datos listos para consulta
- Snowflake, Google BigQuery

Almacenamiento distribuido
- Sistemas a gran escala
- HDFS

Delta Lakes | Una solución perfecta para ETL

Capa de almacenamiento open-source nativa en Databricks sobre Apache Spark basado en Parquet y logs de transacciones utilizando Delta Tables

Transacciones ACID

Schema enforcement

Time Travel

Optimización de datos

CESTE

Transacciones ACID | La base de la integridad en el Data Lake

Atomicity

Consistency

Isolation

Durability

- Las operaciones se completan totalmente o fallan sin dejar cambios
- Los datos siempre cumplen con el esquema definido
- Escrituras concurrentes no interfieren entre sí
- Los datos persisten tras fallos del sistema

Schema Enforcement | Carga datos correctos o no los cargues

Enforcement

Si cargas datos que no cumplen el esquema esperado, el proceso falla (protección automática)

Evolution

Se puede permitir que el esquema evolucione automáticamente si queremos mediante (*mergeSchema*)

Time Travel | Versionado inteligente para tus Delta Tebles

Cada Delta Table guarda un log de transacciones que registra todos los cambios permitiendo consultar versiones anteriores de una tabla

Auditar cambios

Recuperar datos eliminados

Comparar evolución de datos

CESTE

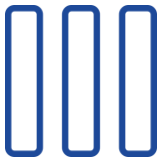
Optimización de datos | La clave para consultas ultrarrápidas

Optimizer

Compacta pequeños archivos Parquet en archivos grandes

mejora el rendimiento de lectura.

Z - Ordering



Ordena los datos físicamente por una o más columnas

mejora búsquedas selectivas.

Delta Tables | La clave para tratar con los datos en Databricks

databricks

CESTE

www.ceste.es