Curso Big Data (25h)

Diego Sevilla Ruiz — Universidad de Murcia

CFP Carlos III

6 sesiones × 4 h + 1 h trabajo autónomo

Objetivo: Dejar un mini-lakehouse reproducible con ETL batch + streaming, consultas SQL y dashboard, orquestado con un DAG simple. Mostrará un proceso End-to-End (E2E) de Big Data moderno: ingesta → almacenamiento → procesamiento → SQL → dashboard → orquestación

Agenda del curso

1. Arranque del entorno y datos (4 h)
2. Spark batch I (4 h)
3. Spark batch II + Catálogo/Trino (4 h)
4. Ingesta en streaming (Kafka + Structured Streaming) (4 h)
5. SQL interactivo + Lakehouse ligero + BI (4 h)
6. Orquestación (4 h)
7. Trabajo autónomo del alumnado (1 h)

Trabajo Previo (previo a la Sesión 1)

- Tener disponible una instalación de Ubuntu reciente (preferiblemente) o Windows con WSL2 (si maneja bien docker y linux)
- Instalar **Docker** y **Docker Compose** y conocer el uso básico
- Conocer el lenguaje **Python** y tener conocimientos de **SQL**
- Conocer **git** y ser capaz de clonar repositorios
- Dataset de ejemplo: **ventas** (CSV)

Sesión 1 — Arranque del entorno y datos (4 h)

Objetivos

- Levantar stack local (Spark, HDFS, MinIO, Trino, Hive Metastore, UIs)
- Entender formatos columnares y particionado
- Entender almacenamiento HDFS, Hadoop vs S3 (MinIO)

Contenidos

- Docker/Compose: instalación, arranque — **40 min**
- Intro al pipeline y UIs (Spark/HDFS/MinIO/Trino) — **40 min**
- Formatos: CSV/JSON vs **Parquet**, particiones, uso con Pandas y Spark — **80 min**
- Almacenamiento: **HDFS** vs **MinIO**; esquemas de rutas `hdfs://` y `s3a://` — **40 min**

Práctica

- Convertir **CSV → Parquet** y **particionar por fecha**

Sesión 2 — Spark batch I (4 h)

Objetivos

- Cargar, transformar y persistir con **DataFrames**

Contenidos y tiempos

- Lecturas/escrituras, `select`, filtros — **120 min**
- Joins y agregaciones — **60 min**
- Buenas prácticas: `repartition/coalesce`, shuffles (intro) — **60 min**

Práctica

- **Mini-ETL “ventas” (bronze → silver):** limpieza, tipificación, Parquet particionado

Sesión 3 — Spark batch II + Catálogo/Trino (4 h)

Objetivos

- Ejecutar el ETL sobre HDFS/MinIO y exponer datasets como tablas

Contenidos y tiempos

- Rendimiento: particionado por año/mes, conteo de archivos, tamaños — **60 min**
- Comparativa IO HDFS vs MinIO — **40 min**
- Catálogo: **Hive Metastore** y tablas externas — **60 min**
- **Trino SQL**: filtros por partición, agregaciones, joins — **80 min**

Práctica

- Correr ETL en HDFS **y** MinIO; **registrar tablas** (external)
- Consultar desde Trino y validar pushdown por particiones

Sesión 4 — Ingesta en streaming (4 h)

Objetivos

- Ingerir eventos con **Kafka** y persistir Parquet mediante **Structured Streaming**

Contenidos y tiempos

- Kafka: topics, particiones, retención — **60 min**
- Spark Structured Streaming: Kafka → Parquet — **120 min**
- Validación de datos en llegada — **60 min**

Práctica

- Productor simulado → **consumo streaming** a “bronze-stream”
- Visualizar llegada en Trino / inspección de archivos

Sesión 5 — SQL + Lakehouse ligero + BI (4 h)

Objetivos

- Explorar con SQL, versionado mínimo ACID y generación de KPIs

Contenidos y tiempos

- SQL interactivo en Trino sobre silver/gold — **90 min**
- Lakehouse (demo) (Iceberg): ACID, MERGE/UPSERT, time-travel — **60 min**
- BI (Superset/Metabase): modelo bronze/silver/gold y datos para dashboard — **90 min**

Práctica

- **Dashboard**: ventas por año/categoría; 2–3 KPIs

Sesión 6 — Orquestación (4 h)

Objetivos

- Encadenar el pipeline y dejar un E2E reproducible

Contenidos y tiempos

- **Airflow**: DAG simple (ingesta → ETL → refresco tabla → dashboard) — **90 min**
- Taller de integración E2E — **120 min**

Práctica

- **DAG funcional** que dispara productor Kafka, ejecuta ETL Spark, refresca tablas y genera un screenshot del dashboard

Sesión 7 — Trabajo autónomo (1 h)

Objetivo

- Afinar y entregar el mini-proyecto

Qué hacer

- Checklist de entrega, rúbrica, dudas puntuales

Producto final esperado

- Repo con `docker-compose`, ETL batch/streaming, tablas registradas y dashboard con 2–3 KPIs

Qué entra

- Docker/Compose + UIs
- Parquet y particiones
- Spark batch (ETL “ventas”)
- HDFS y MinIO
- Kafka + Structured Streaming (pipeline mínimo)
- Metastore + Trino
- Dashboard (sólo generación, no visualización)
- DAG simple de orquestación

Qué se recorta para 25h

- Spark **internals** avanzados y tuning profundo → solo *buenas prácticas* esenciales
- **Delta/Iceberg** → *demo ligera* (ACID, MERGE, rollback) sin profundizar
- Operaciones/costes/logs → pinceladas integradas en el DAG
- Visualización y generación on-line del dashboard → solo generación automática