Instrucciones de instalación de software para Procesamiento Masivo de Datos y Visualización

0. Prerrequisitos y pasos opcionales

Requisitos:

- Docker / Docker Desktop instalado en la máquina
- Mínimo estimado de 12GB de RAM

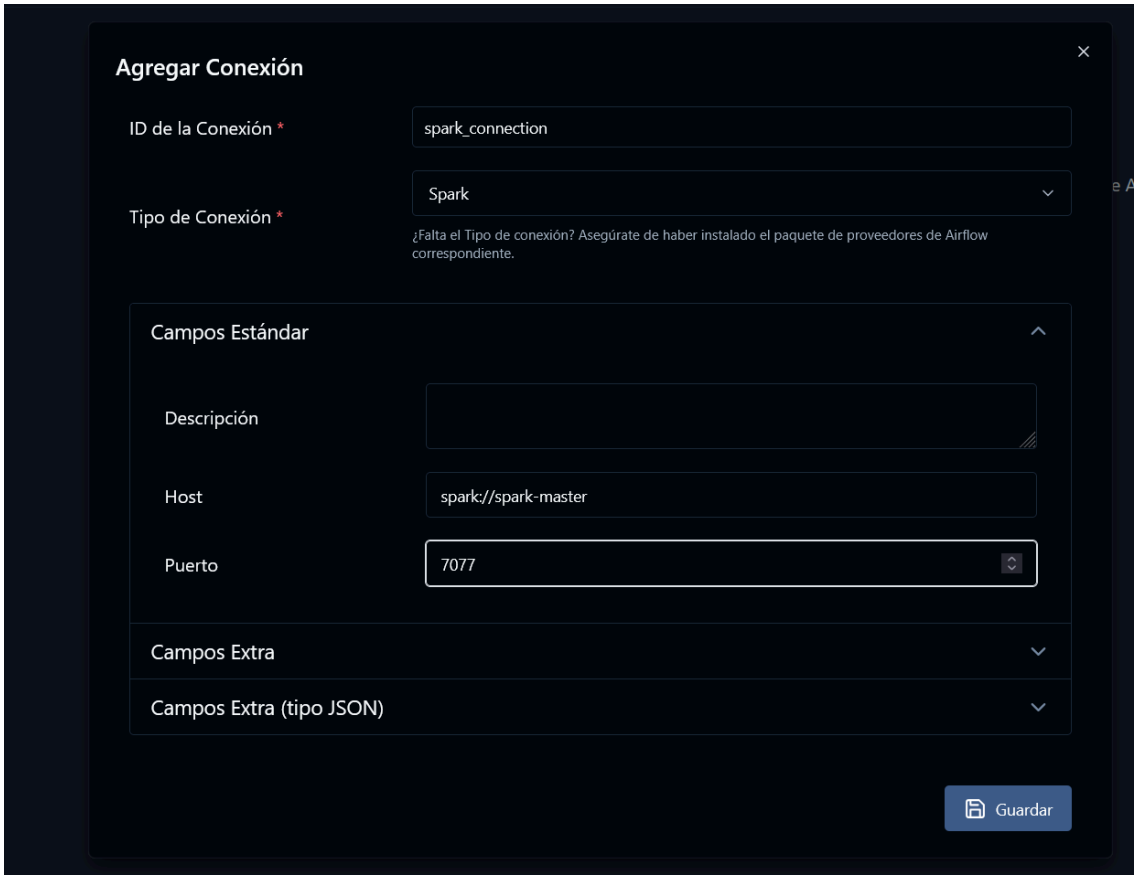
Opcional - Crear una carpeta infra e incluir, al menos:

- Carpeta docker-airflow
- Carpeta docker-jupyterlab
- Carpeta spark
- .env
- requirements.txt
- docker-compose.yml

Si se prefiere, por simplicidad, simplemente descomprimir la carpeta infra tal cual está, entrar en la subcarpeta datalake y pasar al siguiente paso directamente:

1. Ejecutar “docker compose up -d”. Es posible que haya que ejecutarlo 2 veces.
2. Copiar los siguientes archivos desde la carpeta “tests”
    - Copiar “data.csv” en la carpeta “data” que ha aparecido
    - Copiar “spark_direct_submit_demo.py” en la carpeta “jobs”
    - Copiar “demo_spark_direct_submit.py” en la carpeta “airflow/dags”

3. Conectarse a localhost:8080 (airflow/airflow o admin/admin deberían funcionar) y crear una conexión de Spark tal y como se muestra a continuación:



4. Ejecutar el DAG

5. Conectarse a localhost:8081 (Spark-master) y comprobar que el job es recibido por Spark y se ejecuta correctamente