Informe Entrega EE2: dbt Sakila Star

Preparación del entorno de desarrollo

Se han creado tres contenedores $jupyterlab\_container$, $mariadb\_container$ y $clickhouse-server$. Todos están conectados a la misma red ($ee2\_network$), lo que permite que $clickhouse-server$ establezca una conexión con la base de datos $sakila$, alojada en $mariadb\_container$. Para ello, se ha utilizado el siguiente comando:
$CREATE\ DATABASE\ sakila\_proxy\ ENGINE = MySQL('mariadb:3306','sakila','user','userpassword')SETTINGS\ read\_write\_timeout = 10000, connect\_timeout = 100;$

Además, se han configurado los archivos $profiles.yml$ y $dbt\_project.yml$ para que dbt conozca en qué esquema guardar $staging$ y $marts$. Se ha decidido materializar los modelos como tablas en lugar de incrementales, ya que no se ha observado un costo elevado al cargar todos los datos de forma completa. En caso de que posteriormente la cantidad de datos sature el funcionamiento de los modelos, se podría volver a materializar los modelos en incremental.

Desarrollo de la ETL

Una vez establecida la conexión (definida en $profiles.yml$), se extraen las columnas necesarias mediante los scripts SQL ubicados en la carpeta $staging$. Posteriormente, se construyen las dimensiones y la tabla de hechos contenidas en la carpeta $marts$ mediante otros scripts SQL, con el apoyo de los $dim\_XXX.yml$, que permite documentar, verificar nulos, detectar duplicados y definir relaciones entre modelos.

Específicamente para materializar la dimensión $date$ se ha utilizado el siguiente comando en $clickhouse-server$:
$SELECT * FROM\ s3('https://izar.ls.fi.upm.es:30009/sakstar/dim\_date.csv', 'YA9JokyWUb2hFUbKYEEN', '0k2ornkQpVTqUrBbOEsEXOnBEEWgJf4AFQOU4O7Y', 'CSVWithNames')$