Proceso ETL

Extracción

Dentro del flujo principal cargamos todos los archivos denotados en la configuración, ubicado en la dirección base también denotada en la configuración. Después se hace un parse sobre cada archivo csv, cargando cada archivo como un diccionario de datos con 3 campos: nombre, campos y filas.

Transformación

Esta fase se divide en un conjunto de subflujos primarios para cada tabla de hechos designada en la estructura final. Estos reciben la data fuente del proceso de extracción y retornan un diccionario con las tablas de hechos y sus tablas dimensionales ya formateadas, en el orden esperado de inserción. Si encuentra un error crítico, el subflujo termina sutilmente y no entrega tablas adicionales. De esta forma, aun si la transformación de una de las tablas de hecho (y sus tablas dimensionales) falla, no interrumpe el proceso del resto.

Proceso para cada tabla de hechos:

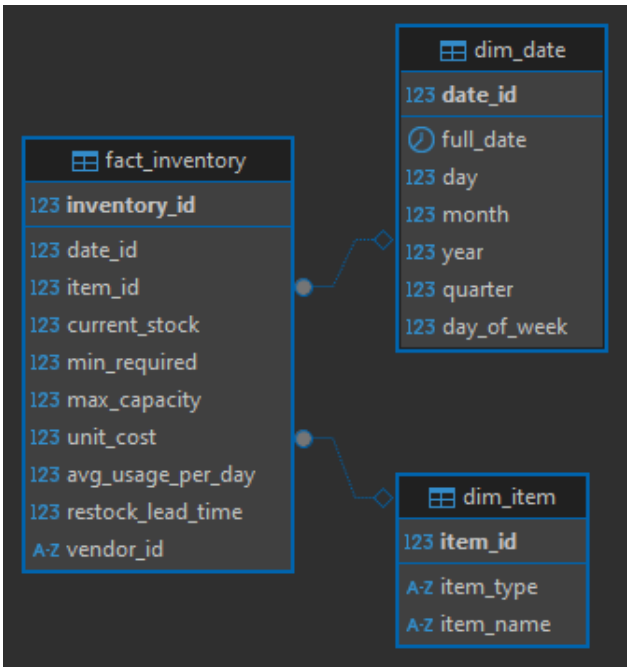
1. Se extra la data csv del diccionario de tablas. Si no existe la tabla esperada, el proceso termina prematuramente.
2. Se formatean los datos de cada campo y cada fila, asegurando que cumplan con su esperado formato y restricciones. De haber un campo que no cumpla el formato esperado, la fila se salta y se continúa el proceso.
3. Simultáneamente, se separan los datos destinados a crear una tabla dimensional. Al haber recorrido la tabla completa, el flujo inicia un nuevo subflujo por cada subtablas para formatearla con los datos extraídos.
4. Al recibir los datos formateados de las subtablas, se hace un proceso de enlazado para campos que dependen de las subtablas, eg.: llaves foráneas.
5. Al finalizar se acumulan todas las tablas creadas en un diccionario anotado con el nombre de la tabla en el orden de dependencia (tablas más dependientes al final).

El proceso es el mismo para cada tabla de forma recursiva. Esto facilita la dispersión de datos entre tablas interdependientes y asegura que todos los datos necesarios para cada tabla de hechos individual este presente independientemente del resto.

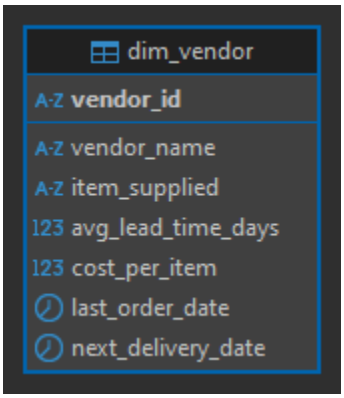
Separación de datos por archivo

Por cada tabla csv, los datos son reestructurados a su esquema esperado. Para cada una la transformación resulta en las siguientes tablas sql:

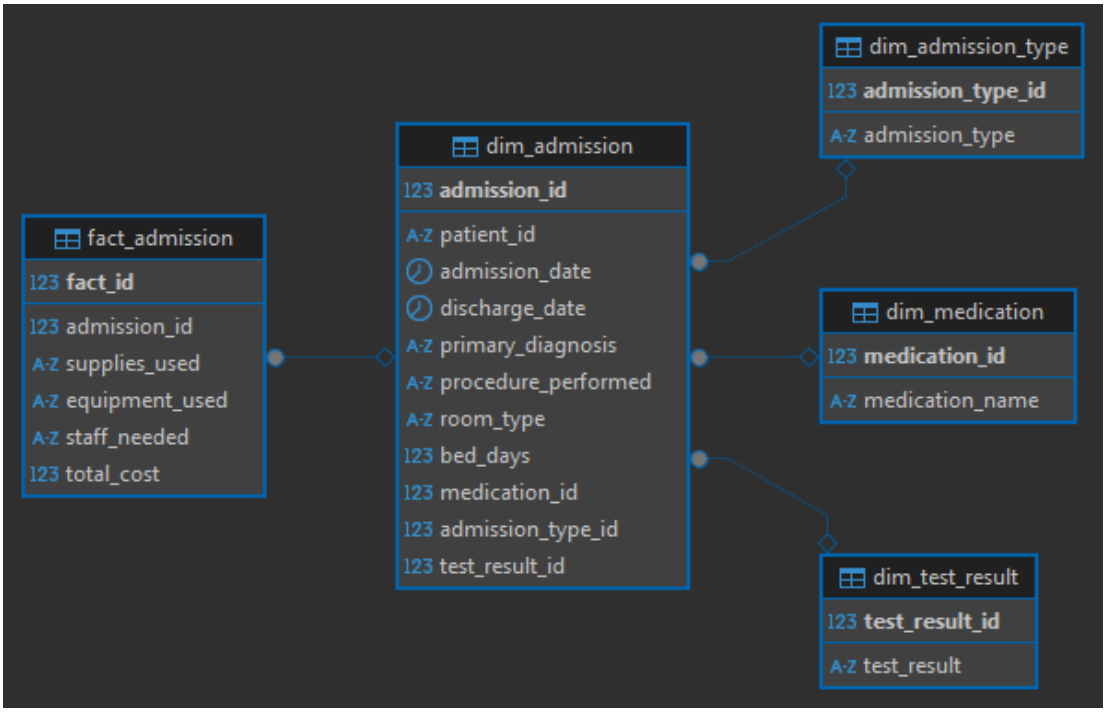
inventory_data



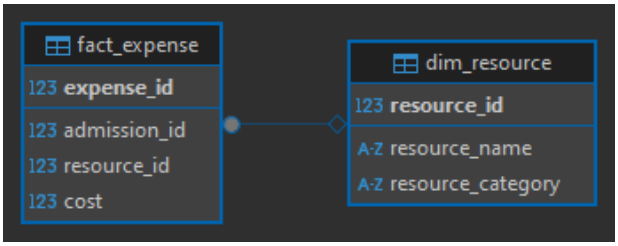
vendor_data



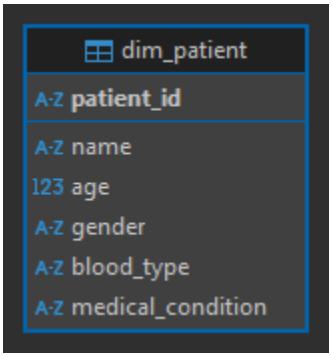
admission_data



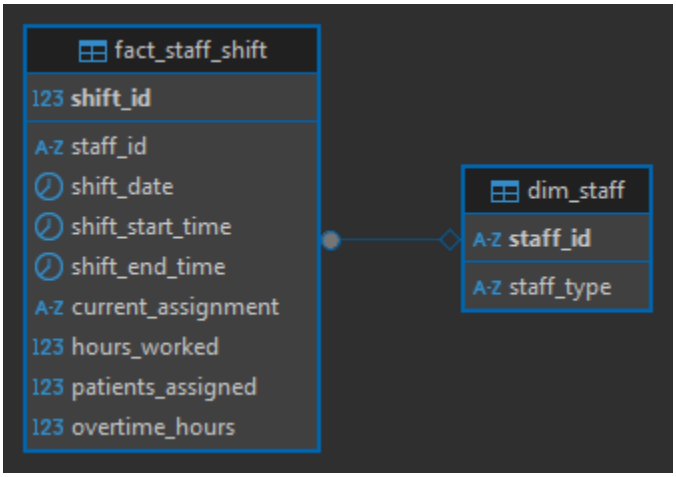
expense_data



patient_data



staff_shift_data



Carga

Finalmente, Prefect hace conexión a la base de datos SQL Server designada en la configuración. Una vez confirme el enlace, el proceso primero vacía tablas relevantes antes de cargar la data en caso de haber datos restantes. El proceso asume que la base de datos ya está creada y con sus tablas debidamente configuradas.