DISEÑO DE UN DATA MART PARA EL DATASET ECOMMERCEFAKEDATA

SISTEMAS DE BASES DE DATOS AVANZADAS

INTEGRANTES DEL GRUPO

BAJAÑA FERNANDEZ LUIS ANDRES

CAÑARTE ORMAZA JOHN EDUARDO

COTERA MENDOZA WASHINGTON GENARO

MITE GUILLEN BRIAN ANDREI

Proceso de ETL, Diseño de Data Mart y Visualización en Power BI

# **1. Introducción**

En este documento se presenta el proceso completo de construcción de un Data Mart de comercio electrónico a partir de datos transaccionales. El trabajo incluye la extracción, transformación y carga (ETL) de los datos en un modelo dimensional, su explotación mediante Power BI y la generación de indicadores claves de negocio (KPIs).

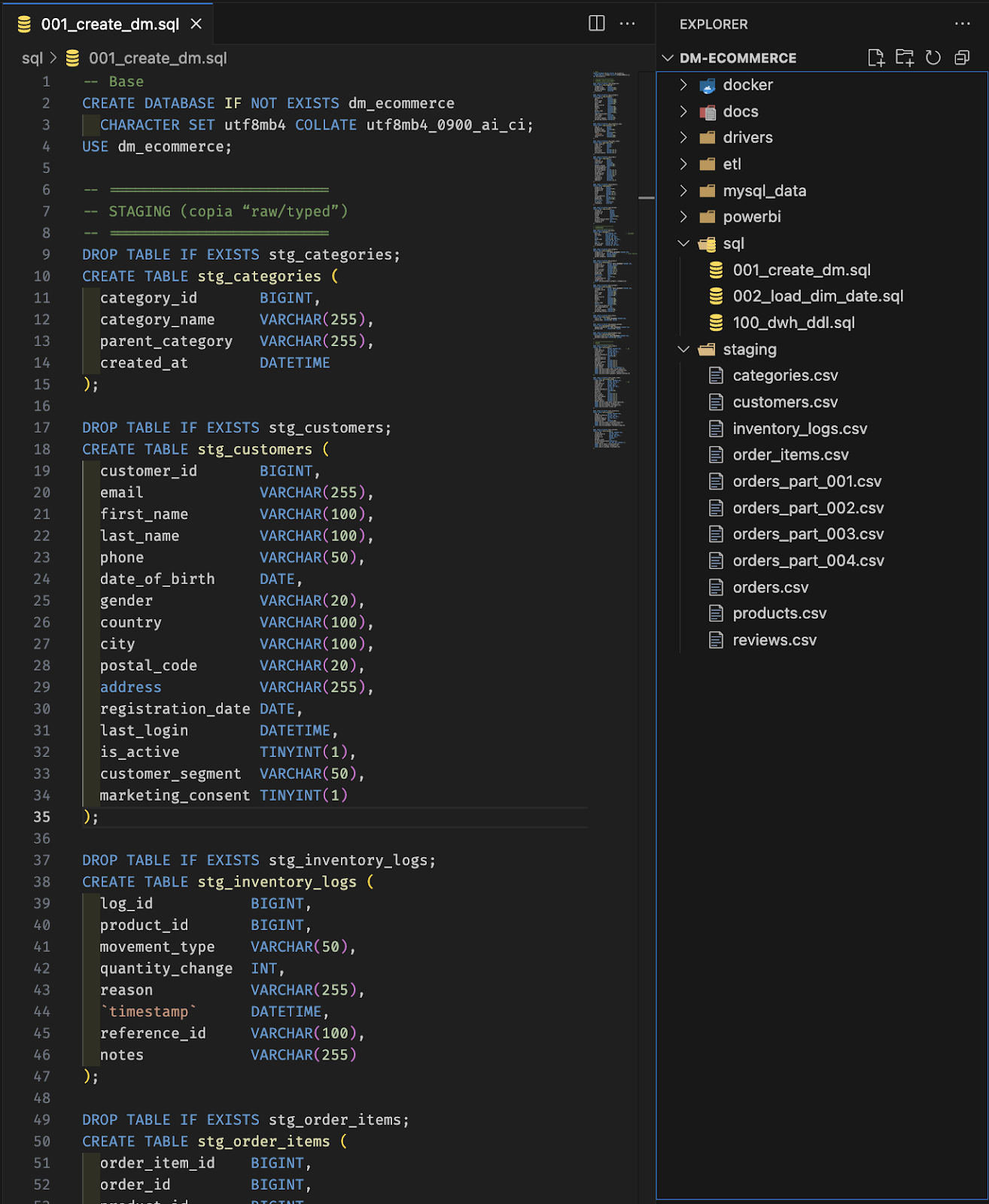
# **2. Dataset seleccionado.**

El dataset proviene de Kaggle – [ECommerceFakeData](https://www.kaggle.com/datasets/sachingupta26/ecommercefakedata).

Incluye información clave de un ecosistema de comercio electrónico:

* **categories.csv** → categorías de productos, con jerarquías (Electronics → Smartphones).
* **products.csv** → catálogo de productos con atributos de marca, precio, costo, peso, dimensiones, descripción.
* **customers.csv** → clientes con datos demográficos, geográficos, segmentación de marketing, fecha de registro y última conexión.
* **orders.csv y orders\_part\_00x.csv** → pedidos realizados por los clientes, con fechas, estado de la orden, método de pago, direcciones de envío/facturación, descuentos, impuestos, costos de envío y totales.
* **order\_items.csv** → detalle de ítems dentro de cada pedido (qué producto, cuántas unidades, precio unitario, descuentos).
* **reviews.csv** → reseñas de clientes con calificaciones, comentarios y votos de utilidad.
* **inventory\_logs.csv** → movimientos de inventario (entrada, salida, devoluciones, ajustes) con cantidades, razones y fechas.

Este dataset fue elegido porque:

* Contiene información transaccional, de clientes y de productos típica de un e-commerce real.
* Tiene un volumen considerable (>500MB) que permite probar procesos ETL y modelado dimensional a escala.
* Permite analizar no solo ventas, sino también satisfacción del cliente (reviews) y logística (inventario), ampliando el alcance del Data Mart.  
  

# **3. Diseño del Data Mart.**

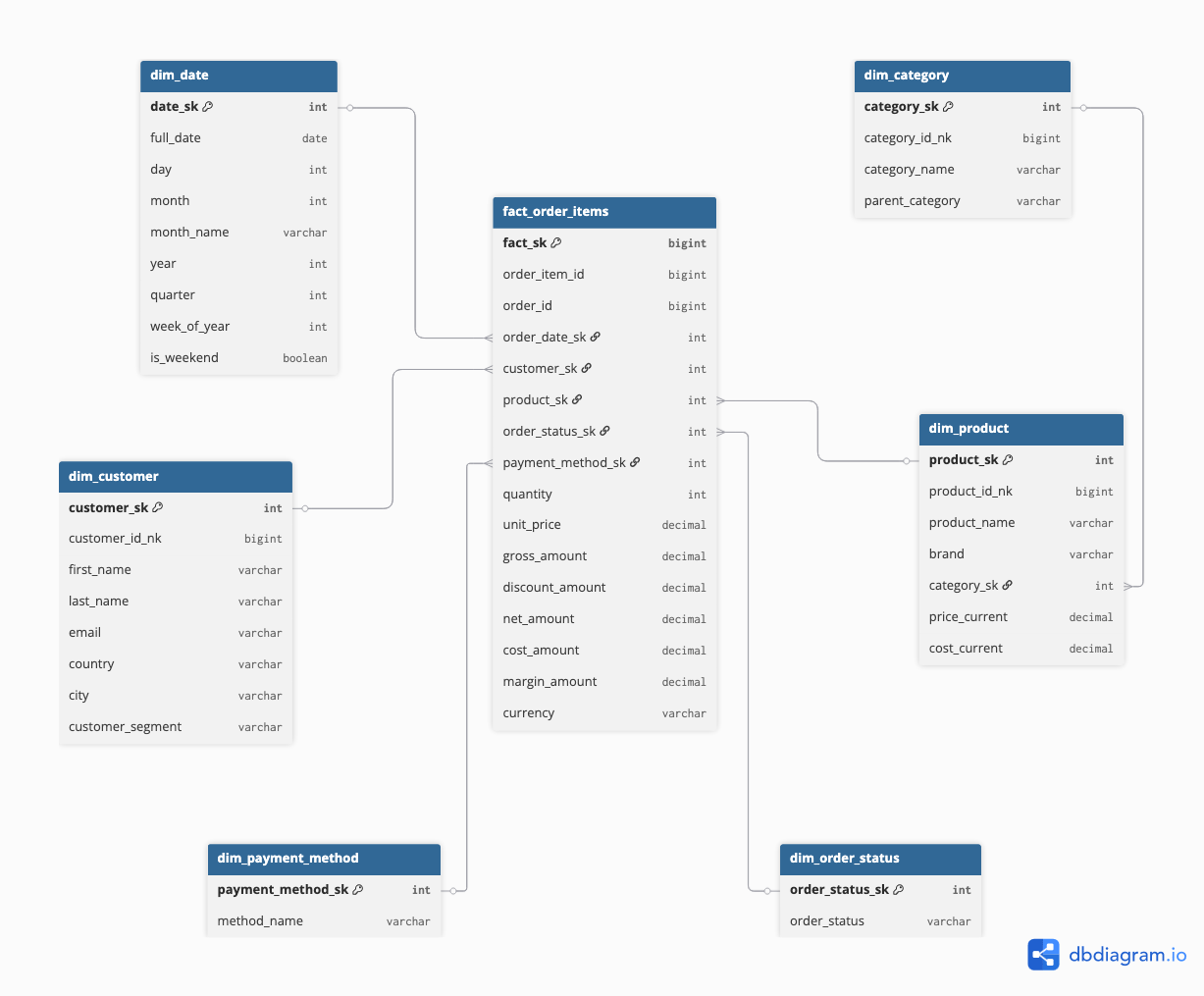
Se diseñó un Data Mart orientado a ventas, siguiendo un modelo dimensional en estrella, donde la tabla de hechos principal es fact\_order\_items (una fila por ítem de pedido).

La razones de este diseño son:

* Centralizar métricas de ventas en una única tabla de hechos.
* Permitir análisis por múltiples perspectivas (producto, categoría, cliente, tiempo, estado del pedido, método de pago).
* Definir la granularidad al nivel más bajo (detalle de ítem de pedido) para mantener flexibilidad analítica.

# **4. Diagramas del modelo dimensional**

La granularidad definida es una fila por ítem de pedido (order\_item\_id).



# **5. Proceso ETL (Pentaho)**

## **5.1 ¿Qué es un ETL?**

Un ETL (Extract, Transform, Load) es un proceso que:

* Extrae datos desde diversas fuentes (archivos CSV, bases transaccionales, APIs).
* Transforma los datos para limpiarlos, enriquecerlos y adaptarlos al modelo destino.
* Carga los datos en un Data Warehouse o Data Mart, optimizados para análisis.

## **5.2 ETL en este proyecto**

Se utilizó Pentaho Data Integration (PDI):

* Transformaciones de staging para normalizar y limpiar los datos.
* Uso de pasos cómo Text File Input, Replace in String, Strings Cut para corregir formatos de fechas.
* Database Lookup para asignar llaves sustitutas (SK) desde las dimensiones.
* Table Output para cargar los hechos en el Data Mart.

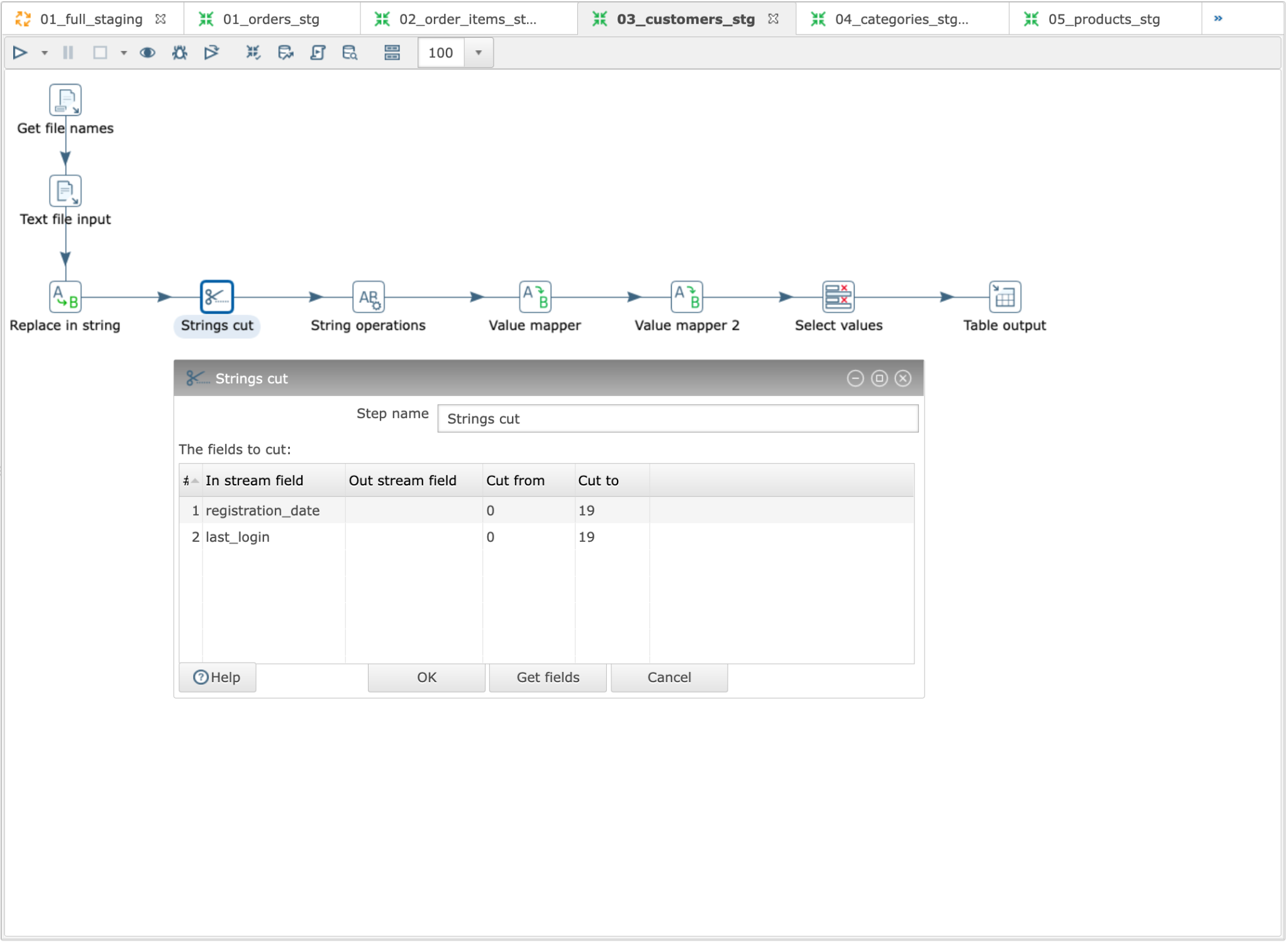
El proceso ETL fue implementado en Pentaho Data Integration (PDI - Spoon). Consta de tres fases principales:

* Carga a staging: ingestión de archivos CSV originales a tablas temporales en la base de datos.
* Construcción de dimensiones: normalización y carga de las tablas dim\_customer, dim\_product, dim\_category, dim\_order\_status, dim\_payment\_method y dim\_date.
* Carga de hechos: integración de los datos de staging y dimensiones para poblar fact\_order\_items.

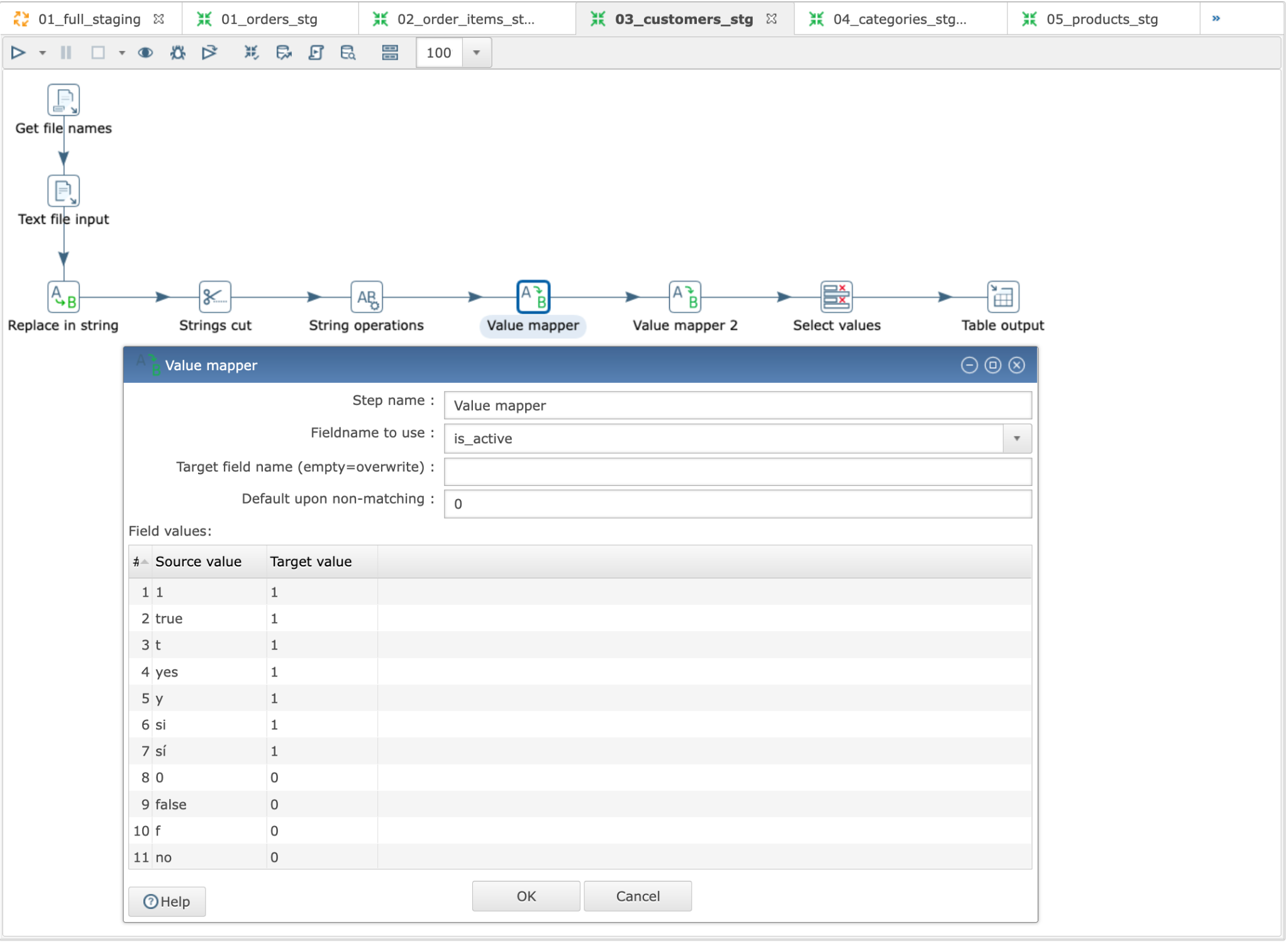
Cada transformación .ktr y job .kjb fue versionado en la carpeta ETL del proyecto. Se implementaron pasos de limpieza como la normalización de fechas (YYYYMMDD) para asegurar consistencia entre la tabla de hechos y la dimensión fecha.

Algunas capturas del proceso en pentaho.

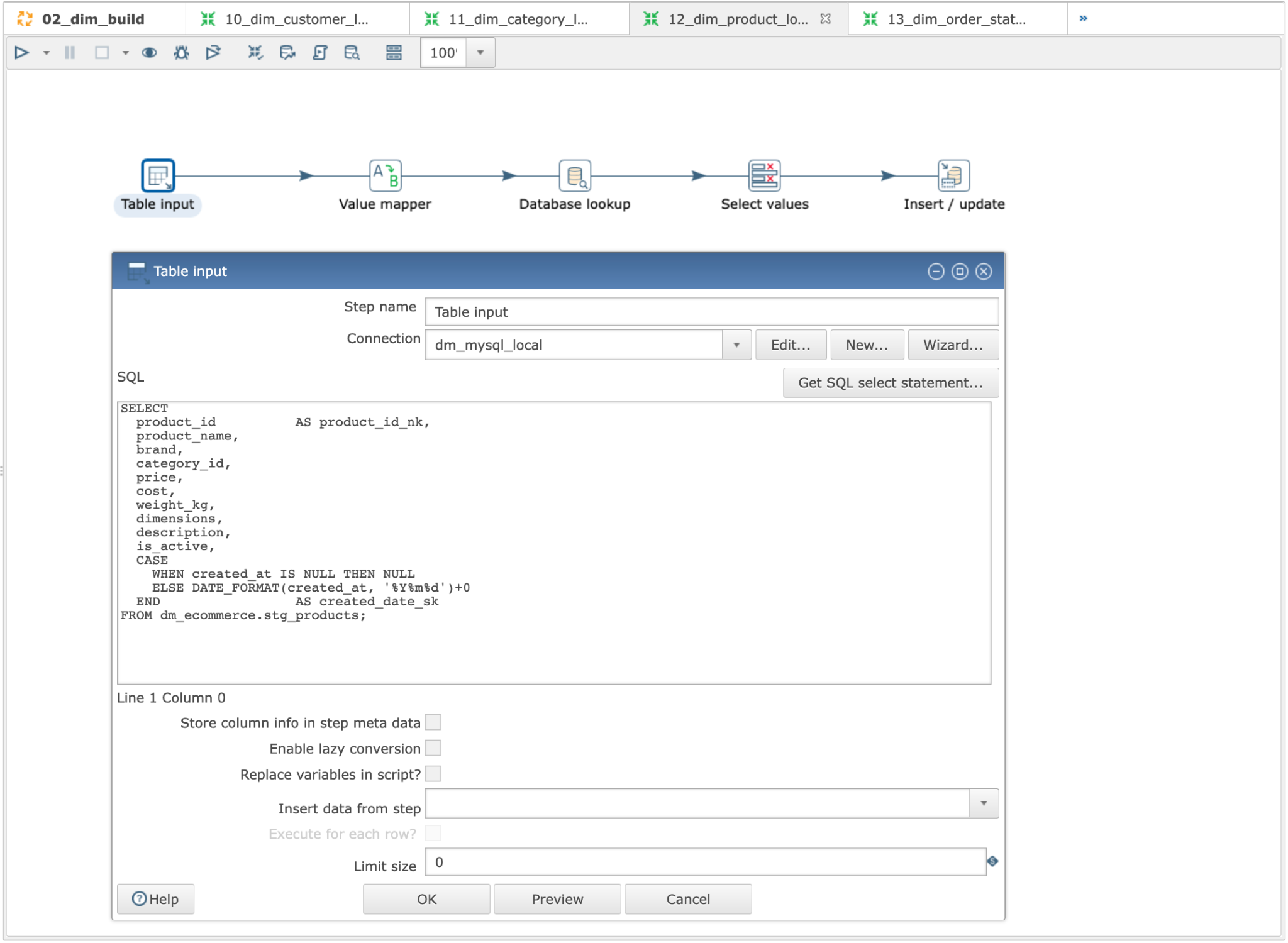
* Por la estructura de las fechas en el dataset fue necesario usar “Replace in string” y “String cut” en cada para cada tabla que las tuviera.



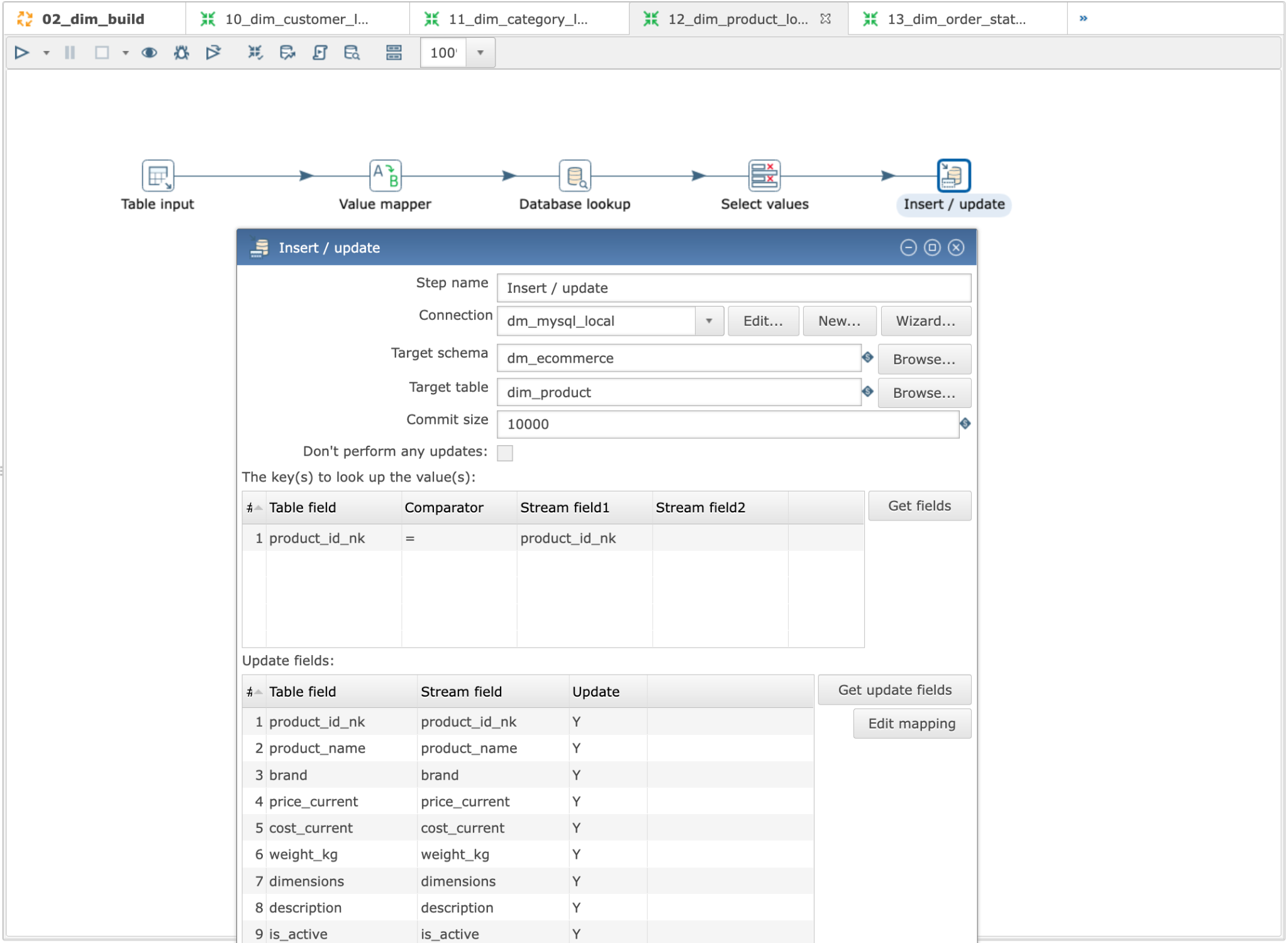
* Por seguridad se utilizó “Value mapper” en los casos donde se trabajó con datos booleanos.



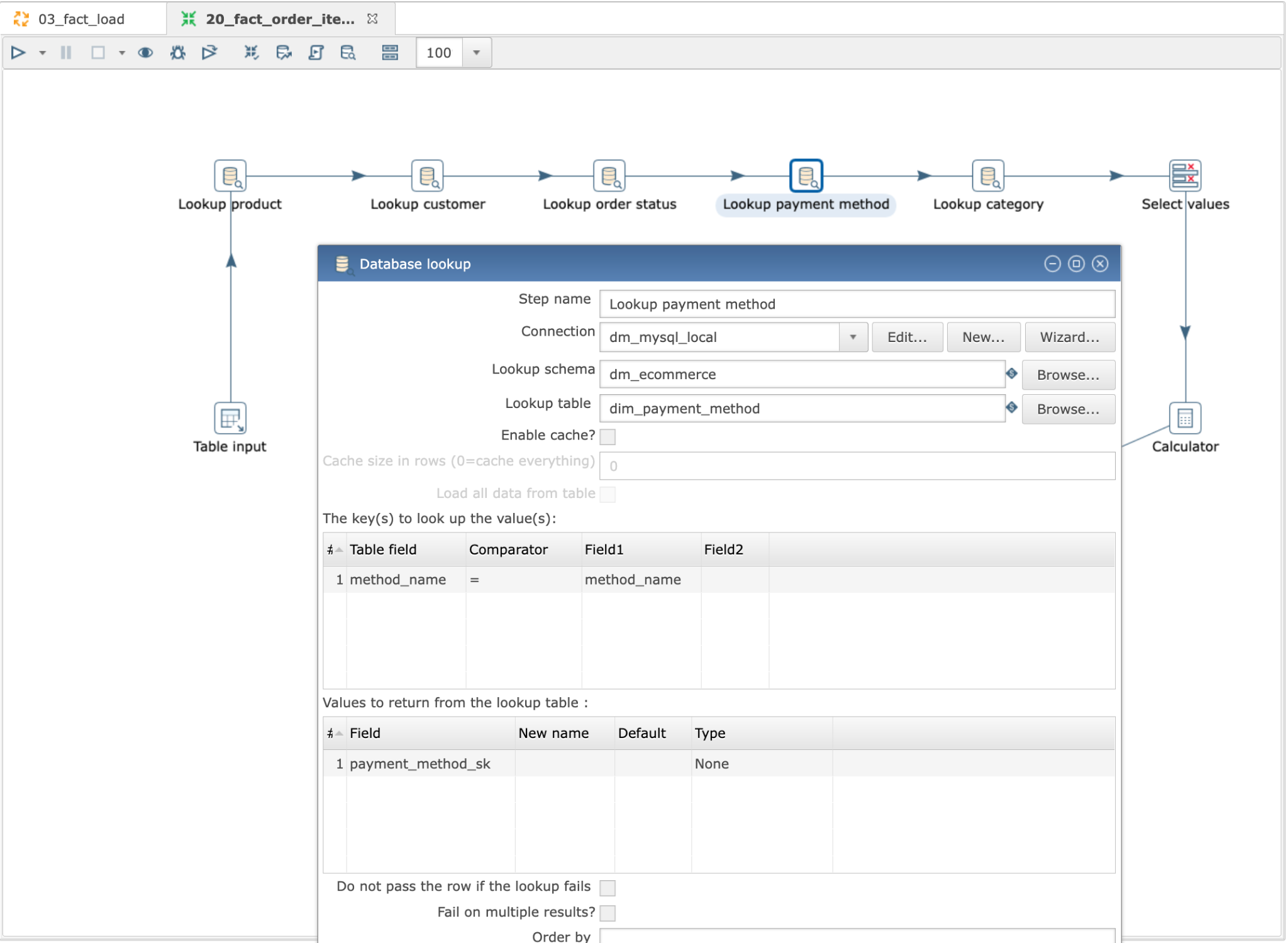
* En la construcción de las dimensiones se usó “Table input” para cargar los datos de las tablas creadas en staging.



* Para especificar donde se guardaran los datos de las dimensiones se usó “Insert/Update”.



* Para cargar columnas de otras tablas se usó “Database Lookup”, la cual permite especificar qué campos necesitamos indicando las columnas por las que se relacionan.



# **6. Power BI**

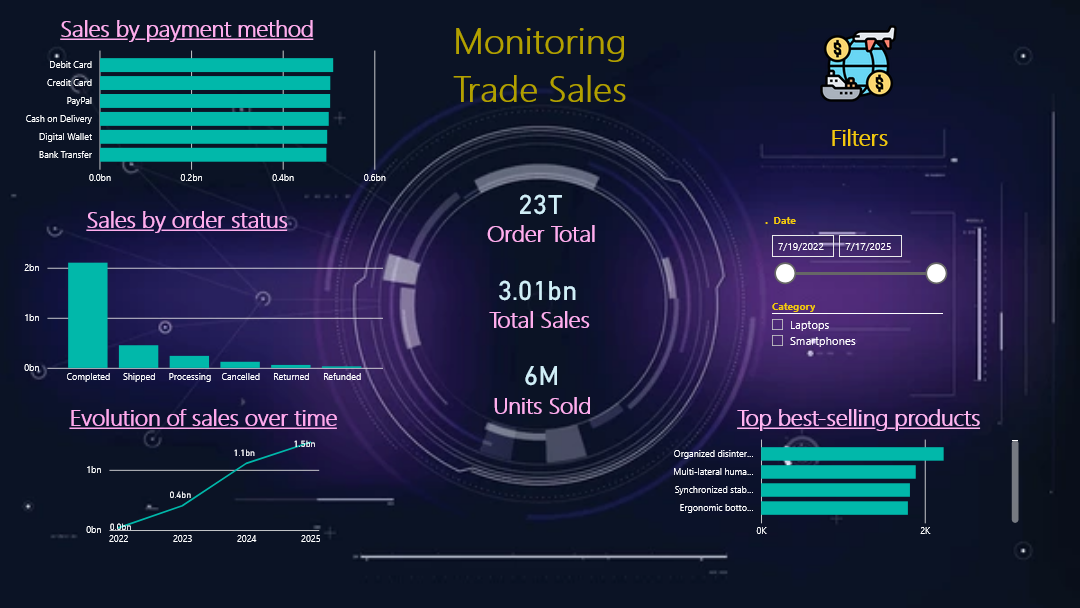
## **6.1 ¿Qué es Power BI?**

Power BI es una herramienta de Microsoft para el análisis de datos y la construcción de dashboards interactivos. Permite a los usuarios finales explorar la información del Data Mart mediante visualizaciones gráficas dinámicas.

## **6.2 Dashboards creados**

Se generaron reportes que permiten responder preguntas clave de negocio:

* Tendencia de ventas en el tiempo (gráfico de líneas).
* Ventas por categoría de producto (gráfico de barras).
* Estado de ventas realizadas(gráfico de barras).
* Métodos de pago más utilizados (gráfico circular).



# Los indicadores se integraron en un único dashboard que resume la información clave y se puede filtrar de forma interactiva con los gráficos implementados y las opciones de filtros incluidas.

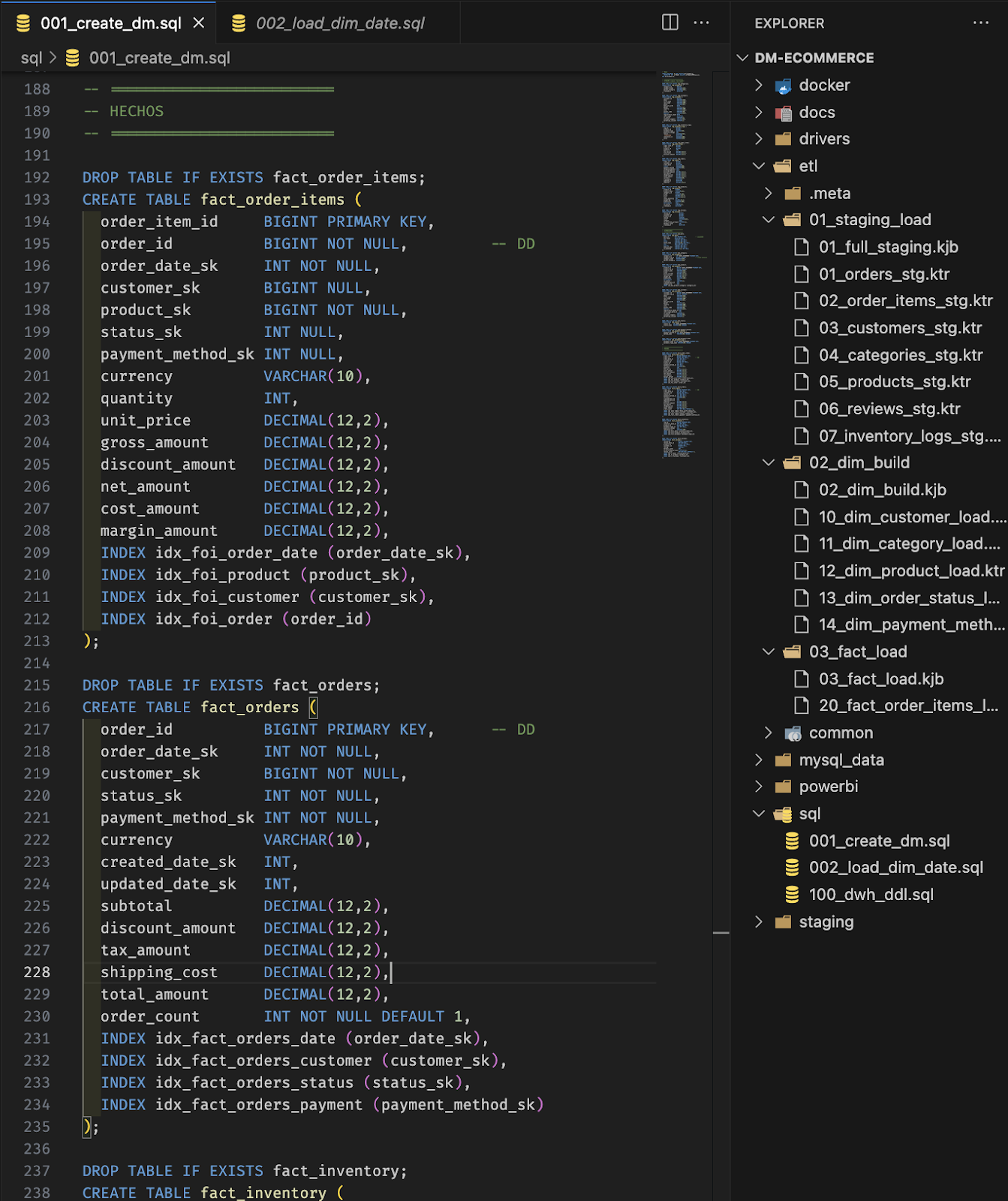
# **7. Tabla de requisitos del negocio**

| **Indicador** | **Variables** | **Características** | **Proceso** | **Método** | **Visualización** |
| --- | --- | --- | --- | --- | --- |
| **Ventas netas** | fact\_order\_items.quantity, fact\_order\_items.unit\_price, fact\_order\_items.discount\_amount, fact\_order\_items.net\_amount | Cantidad monetaria (USD) | Proceso de venta | Ventas brutas = Σ(quantity \* unit\_price)  Ventas netas = Σ(net\_amount) | Tarjeta KPI + tendencia (línea) |
| **Ticket promedio (AOV)** | fact\_order\_items.net\_amount, fact\_order\_items.order\_id | Ratio monetario | Proceso de venta | AOV = Σ(net\_amount) / #órdenes | Tarjeta + línea temporal |
| **Margen bruto %** | fact\_order\_items.net\_amount, fact\_order\_items.cost\_amount, fact\_order\_items.margin\_amount | Porcentaje | Proceso de venta | Margen% = (Σ(margin\_amount) / Σ(net\_amount)) \* 100 | Semáforo / barra por categoría |
| **Tasa de recompra** | fact\_order\_items.customer\_sk, dim\_customer.customer\_id | Porcentaje de clientes | Proceso de venta | Clientes con ≥2 pedidos / Clientes totales (periodo) | Barra / línea |
| **Rating promedio** | fact\_reviews.rating, fact\_reviews.product\_sk | Promedio 1–5 | Postventa | Avg(rating) por producto/categoría | Gráfico de barras |
| **Ítems con stock crítico** | fact\_inventory.stock\_quantity, dim\_product.product\_name, dim\_category.category\_name | Conteo / % | Inventario | % de productos con stock\_quantity < umbral | Tarjeta + barra |

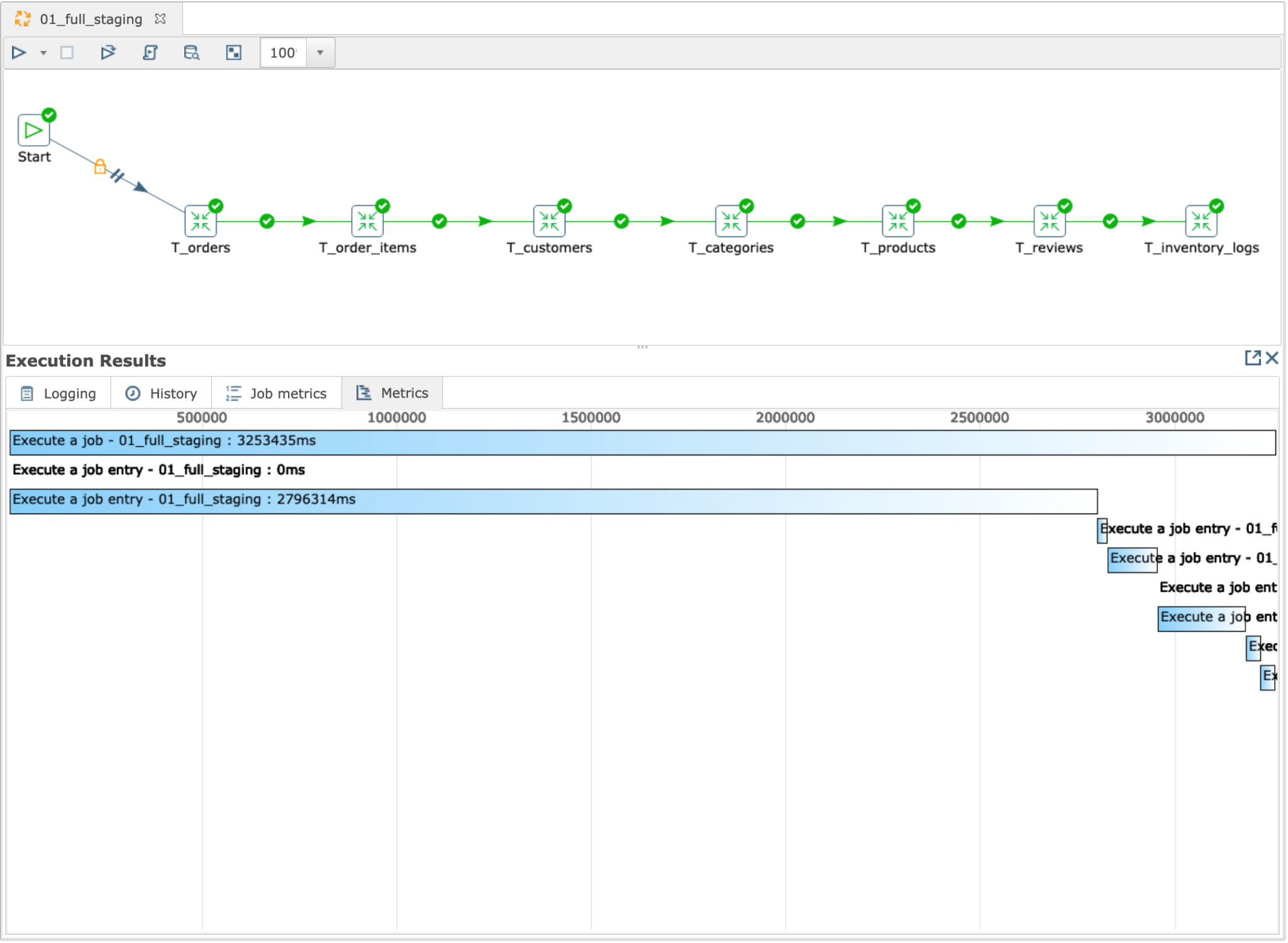
# **8. Código y flujo del proceso ETL**

Se incluyen ejemplos del código SQL y configuraciones en Pentaho para la creación de la tabla de hechos:

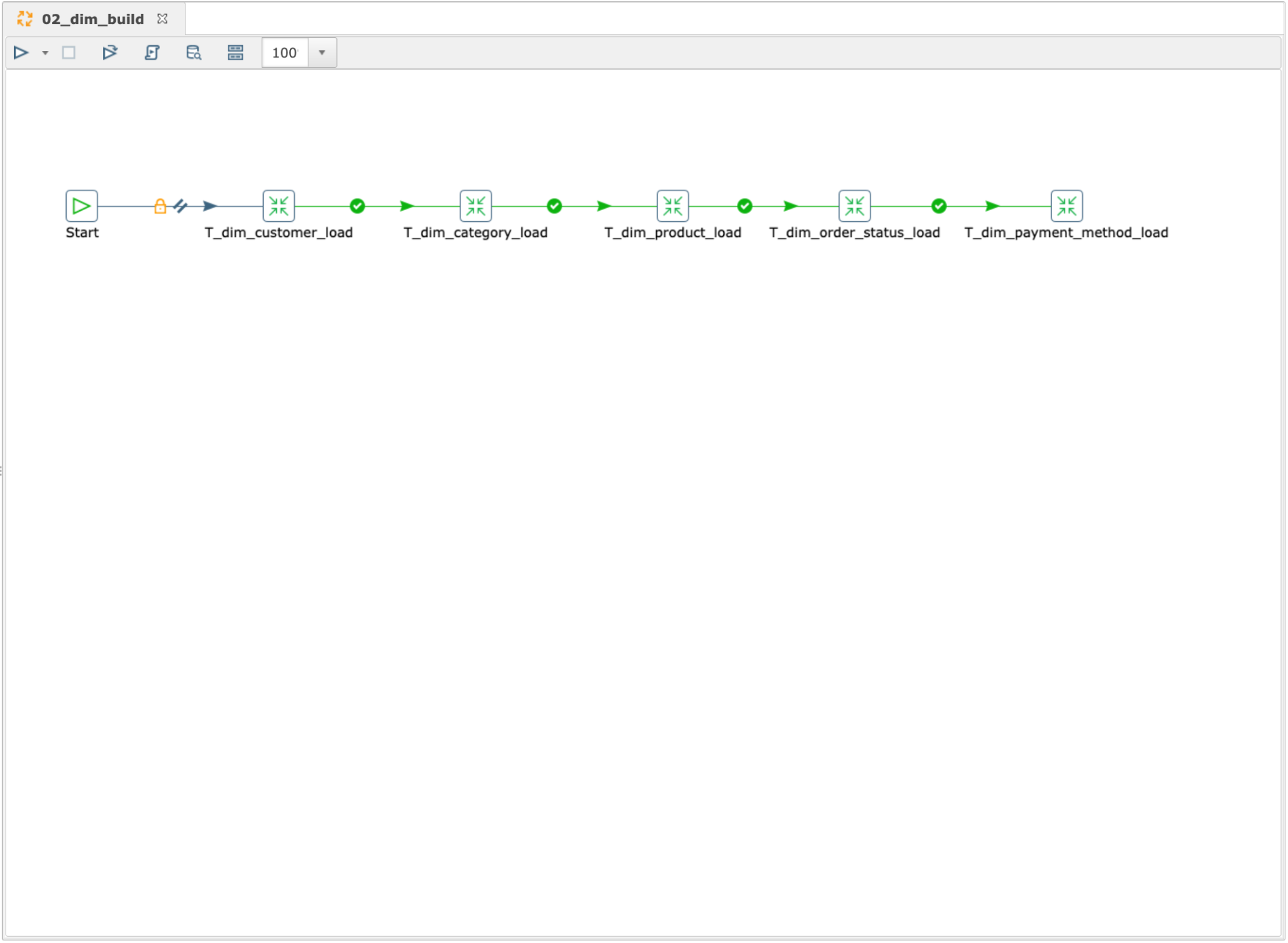
* **Creación de tablas de dimensiones y hechos.**



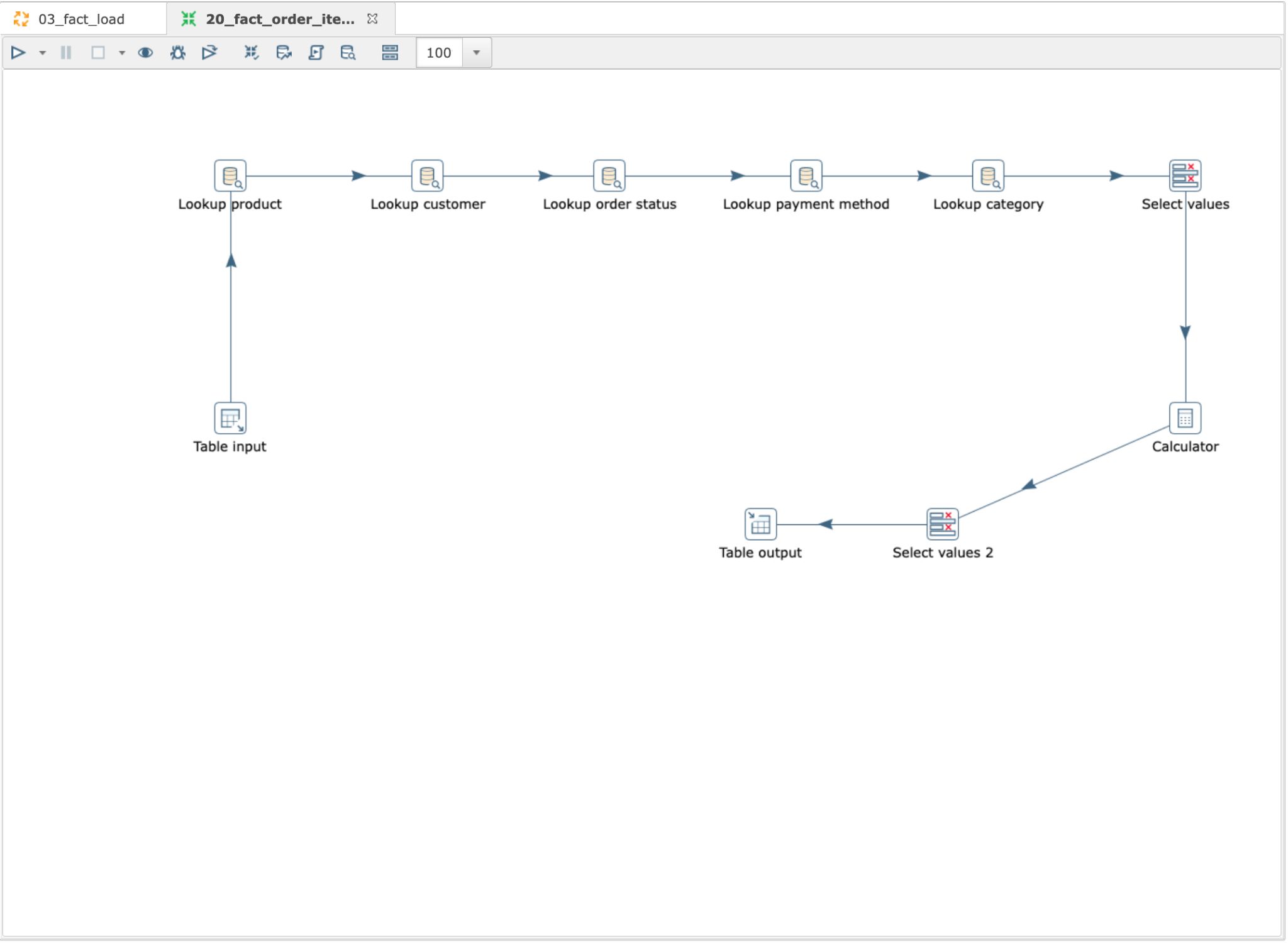
* **Carga a staging.**



* **Construcción de dimensiones.**



* **Carga de hechos.**



El código SQL (creación de Data Mart, carga de dimensiones, vistas finales) y las transformaciones Pentaho (.ktr, .kjb) se organizaron en carpetas según su etapa:

* /etl/01\_staging\_load: Cargas a staging desde CSV.
* /etl/02\_dim\_build: Construcción de dimensiones.
* /etl/03\_fact\_load: Carga de hechos.
* /sql: scripts SQL de creación, carga y vistas.

# **9. Conclusiones**

* Se construyó un Data Mart con granularidad a nivel de ítem de pedido.
* El proceso ETL con Pentaho permitió limpiar y transformar los datos de staging a un esquema en estrella.
* Con Power BI se generaron dashboards interactivos que cumplen con los requisitos de indicadores del negocio.
* El modelo permite escalabilidad para añadir nuevos hechos como inventario o reseñas.