

El proceso ETL desarrollado integra la información transaccional de Wide World Importers y la transforma en un modelo multidimensional compuesto por tres dimensiones principales (Dates, StockItem y PackageType) y una tabla de hechos enriquecida (FactTableOrderLines). El flujo inicia con la extracción de los archivos CSV suministrados por la empresa y cargados previamente en Google Cloud Storage. A partir de estos archivos, se construyen de manera independiente las dimensiones, cada una con sus propias transformaciones y flujo hacia BigQuery, y posteriormente se consolida la tabla de hechos con las medidas e identificadores necesarios para el análisis requerido.

La dimensión DimDates se construye a partir del archivo de órdenes, del cual se deriva el campo OrderDate. Este valor es transformado en el nodo Dates para generar un identificador único DateID en formato YYYYMMDD, acompañado de los componentes de fecha (día, mes, año, trimestre). Luego, un proceso de eliminación de duplicados garantiza una única fila por fecha. Finalmente, la dimensión es cargada a BigQuery como una tabla limpia y estructurada que permitirá relacionar los hechos mediante llaves foráneas basadas en fecha.

La dimensión DimStockItem se basa en el archivo StockItems. En primera instancia, se aplica un Wrangler (StockItemsClean) donde se normalizan tipos de datos y se seleccionan únicamente los atributos relevantes para el análisis. Esta transformación asegura que los valores numéricos y de texto tengan un formato consistente y compatible con BigQuery. Una vez limpiada, la información se almacena directamente en la tabla DimStockItem, conformando así la dimensión de productos necesarios para enriquecer la tabla de hechos con atributos como precio, tasa de impuesto o empaque.

La nueva dimensión DimPackageType fue incorporada al ETL como extensión del proceso original. A partir del archivo PackageTypes, se realiza una primera limpieza y tipificación en el nodo SCD2. Allí se agregan las columnas propias de un manejo de historia tipo 2 simplificado: StartDate, EndDate e IsCurrent, asignando valores por defecto que permiten simular la trazabilidad temporal de los tipos de empaque. Posteriormente, mediante el nodo GenerateID se genera un surrogate key (PackageTypeUuid) utilizando UUID para asegurar unicidad en la dimensión. Finalmente, la dimensión es cargada a BigQuery como una tabla independiente que podrá relacionarse con la tabla de hechos a través del campo PackageTypeID presente en OrderLines.

Una vez construidas las dimensiones, el ETL procede con la consolidación de la tabla de hechos FactTableOrderLines. En la primera etapa, OrderLines y Orders son combinadas mediante un Join para incorporar información del encabezado de las órdenes. Luego, se aplica un Wrangler adicional (FormatDateID) que introduce el DateID para conectar con la dimensión de fechas. En paralelo, la tabla OrderLines es unida con StockItemsClean mediante un segundo Join (Joiner2), que permite anexar el precio unitario y otros atributos del producto. Posteriormente, en el nodo CreateTotalPrice se genera la medida nueva requerida por el laboratorio: TotalPrice, calculada como Quantity por UnitPrice. El flujo resultante contiene todas las llaves foráneas (StockItemID, PackageTypeID y DateID), así como las medidas necesarias, y finalmente se carga a la tabla FactTableOrderLines en BigQuery en modo append.

En conjunto, este proceso ETL final integra todas las dimensiones requeridas, incorpora la nueva dimensión PackageType con control de historia simplificado y ajusta la tabla de hechos agregando la nueva medida TotalPrice y el identificador PackageTypeID. El resultado es un modelo multidimensional consistente, completo y apto para análisis posteriores mediante consultas SQL o herramientas de visualización como Power BI y Tableau.