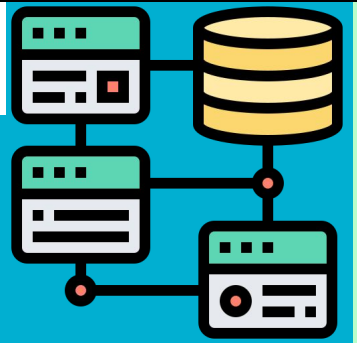
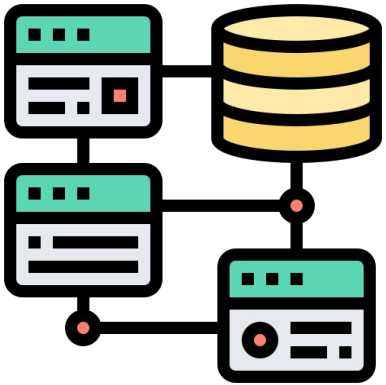
SUPER STORE

ESTRUCTURA DE DATOS

Por: Jessica Cázares

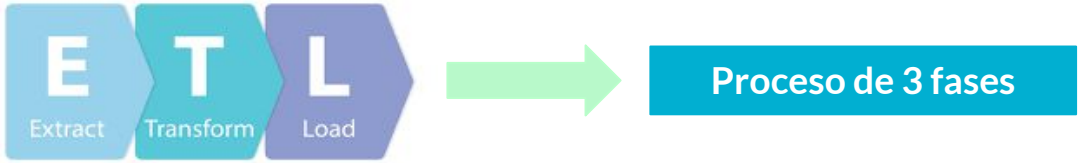


Objetivo



A través del proceso **ETL (Extracción, Transformación y Carga), construir un sistema tabular** que nos permita almacenar datos de manera eficiente y consultar estos datos con mayor facilidad.

Metodología



1. **Extracción (Extraction):** Los datos se extraen desde una o varias fuentes de datos, que pueden ser bases de datos, archivos planos, servicios web u otras fuentes. La extracción implica recopilar la información necesaria para su posterior procesamiento.

2. **Transformación (Transformation):** Los datos extraídos se transforman según los requisitos del sistema de destino. Las transformaciones pueden incluir limpieza de datos, conversión de formatos, combinación de datos de múltiples fuentes, filtrado y otras operaciones que aseguran que los datos sean coherentes y útiles para el análisis.

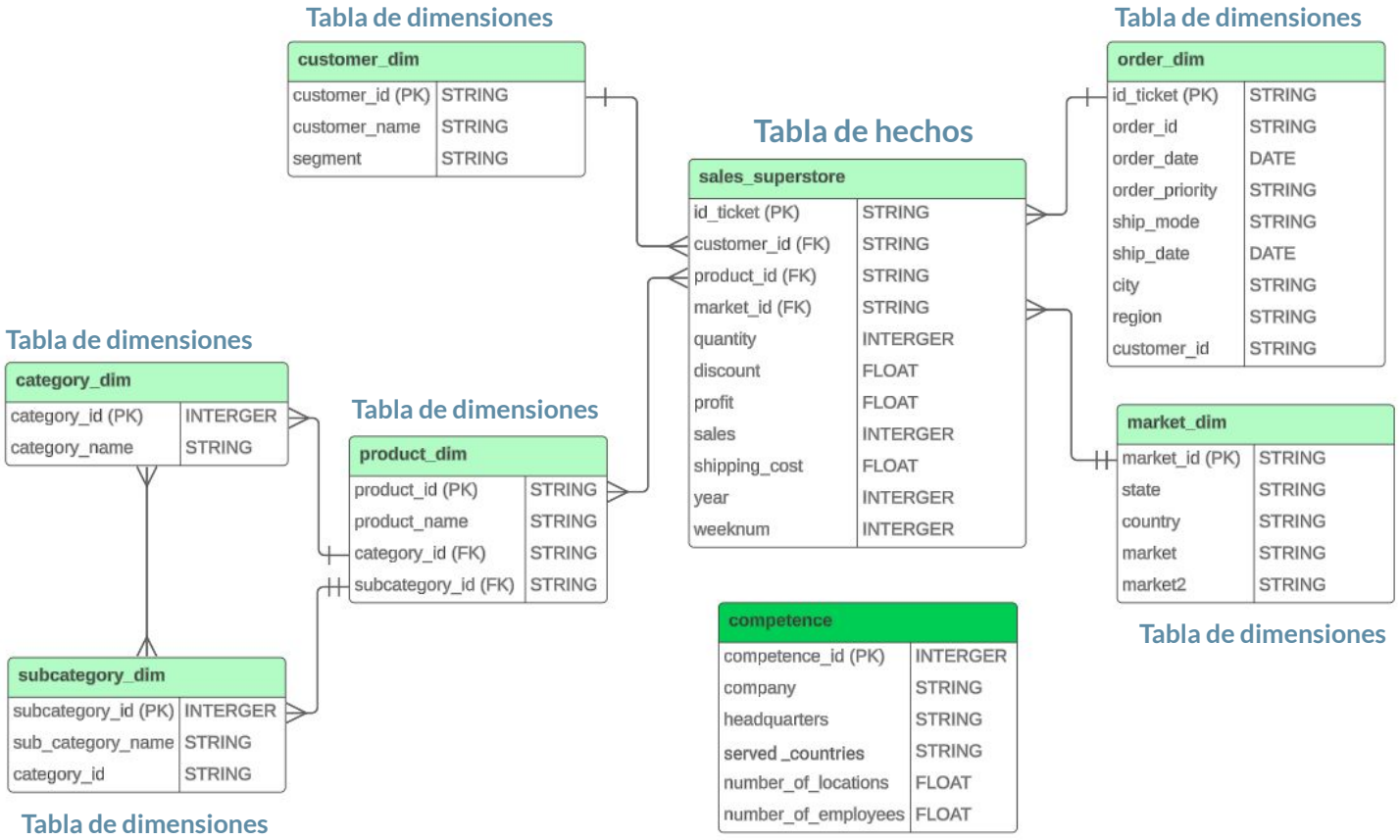
3. **Carga (Load):** La fase final implica cargar los datos transformados en el sistema de destino, que generalmente es un data warehouse o una base de datos diseñada para el análisis de negocios. Los datos ahora están listos para ser consultados y analizados de manera eficiente.

Insumos

- Contamos con un **archivo csv** con la información detallada de las transacciones de **Super store**.

- Para la información de la **competencia** se **extrajo** la información de *wikipidia* usando el paquete Beautiful Soup en Python.

Diseño de estructura de base de datos



Crear estructura de base de datos

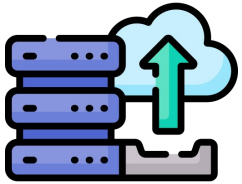
- Creación de tablas de dimensiones y hechos en BigQuery.
- Creación de ID’s y transformación de tipo de datos:
  - **Tabla category_dim**: *category_id*, prefijo “CAT” concatenado con un número incremental de al menos dos dígitos.
  - T**abla subcategory_dim**: *subcategory_id*, prefijo “SUB” concatenando con un número incremental de al menos dos dígitos.
  - **Tabla order_dim:** Se cambió el tipo de dato de las variables *order_date* y *ship_date* de TIMESTAMP a DATE. Se creó un *id_ticket* único concatenando las variables *order_id* y *customer_id*.
  - **Tabla market_dim:** *market_id*, concatenando las variables *state* y *country*.
  - **Tabla compentence_multinational:** *competence_id*, asignando un número incremental a cada *company*.

Diseño de actualización de pipeline

① Extracción

② Transformación

③ Carga



Fuente de Datos

Pre procesamiento

Actualización: Tablas de dimensiones

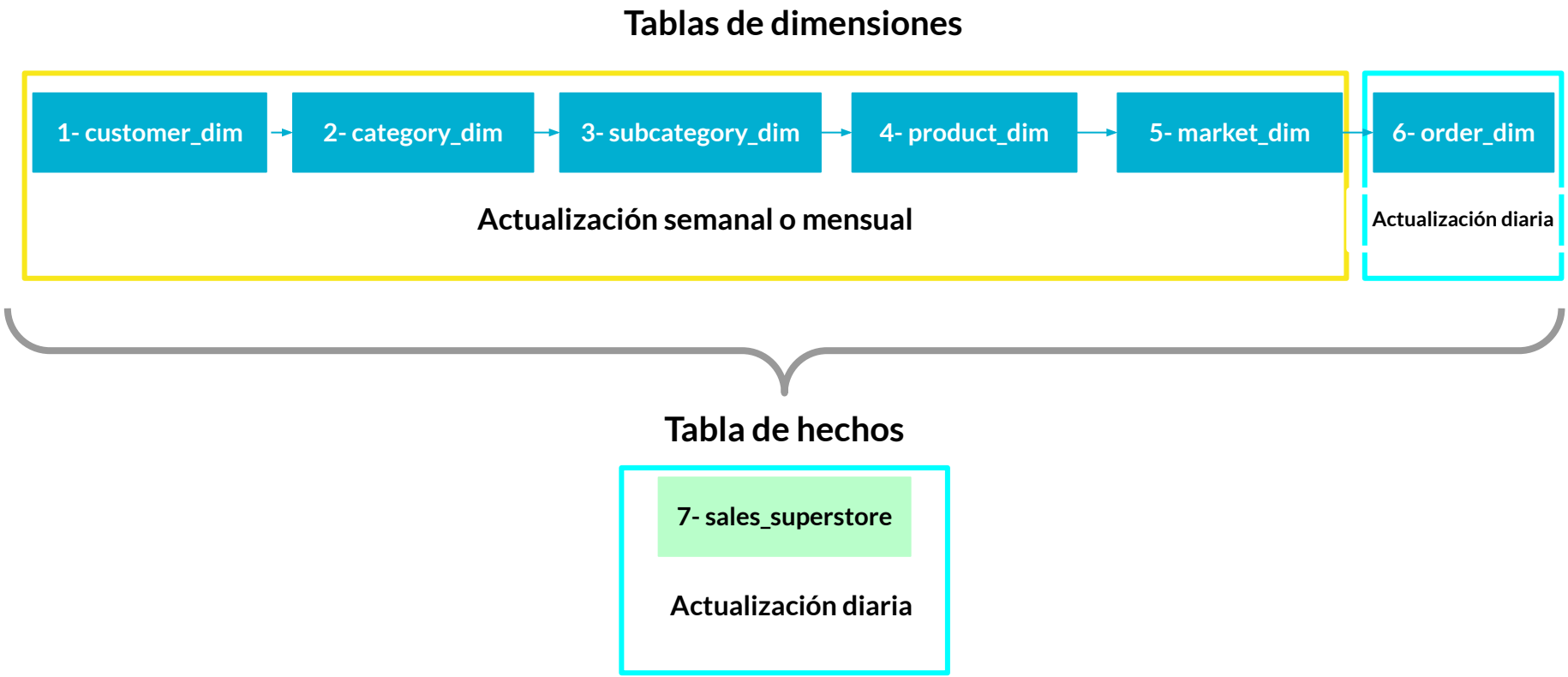
Actualización: Tabla de hechos

Limpieza de datos

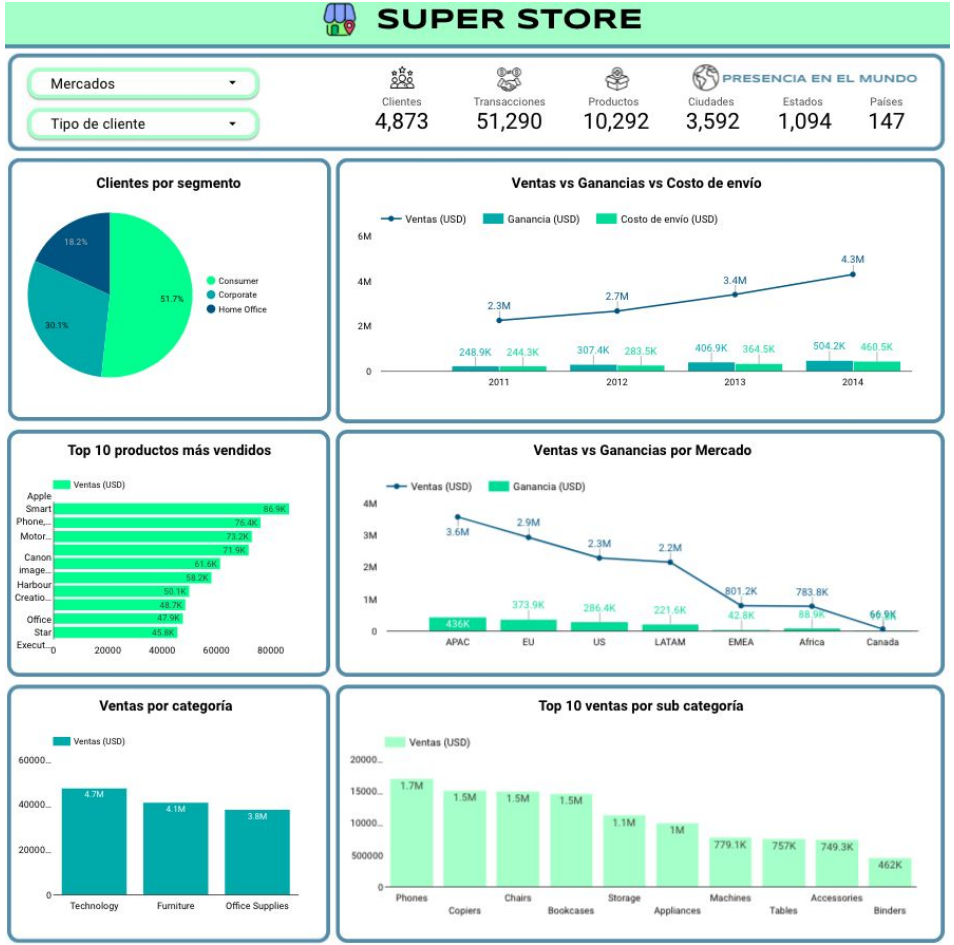
Validación y Transformación

④ Análisis de Datos

Flujo de actualización de tablas



Análisis Exploratorio de Datos



Conclusiones

La **implementación** de un **sistema de tablas de hechos y dimensiones** eficiente, junto con un **pipeline de actualización automatizado**, proporcionará a **Super Store** una estructura sólida para gestionar grandes volúmenes de datos y facilitar el análisis de la información de ventas, productos, clientes y mercados, impulsando así una toma de decisiones más efectiva y basada en datos.

Recomendaciones

- Mantener el **ID único** (***id_ticket***), este campo **se debe generar automáticamente en el proceso de ETL para garantizar que cada transacción sea única**.

- Se deben mantener los identificadores de las tablas de dimensiones (***customer_id***, ***product_id***, ***market_id***, etc.) que se relacionan a la tabla de hechos.

- Dado el volumen de datos y la necesidad de tomar decisiones basadas en información actual, se recomienda que **el pipeline actualice las tablas diariamente para reflejar las ventas más recientes.** Sin embargo, algunas dimensiones como ***products_dim*** o ***customers_dim*** pueden actualizarse semanal o mensualmente si su información no cambia con frecuencia.

- El pipeline debe extraer solo los datos nuevos o modificados desde la última actualización para minimizar el tiempo y los recursos necesarios.

- Se debe incluir un paso de validación para verificar que los datos cargados en las tablas son correctos y completos, evitando inconsistencias.

DASHBOARD:

https://lookerstudio.google.com/reporting/58c871d4-115c-4fec-b35d-976aae6fcd3a

GRACIAS