

Ficha Técnica: Proyecto 5 ETL

Título del Proyecto: Estructura de datos

Objetivo:

A través del proceso ETL(Extracción, Transformación y Carga), construir un sistema tabular para la tienda superstore que nos permita almacenar datos de manera eficiente y consultar estos datos más fácilmente.

Equipo:

Individual.

Herramientas y Tecnologías:

Bigquery, Python, Looker Studio, Google Docs, Google Slices.

Insumos:

CSV Superstore.

CSV Competencia.

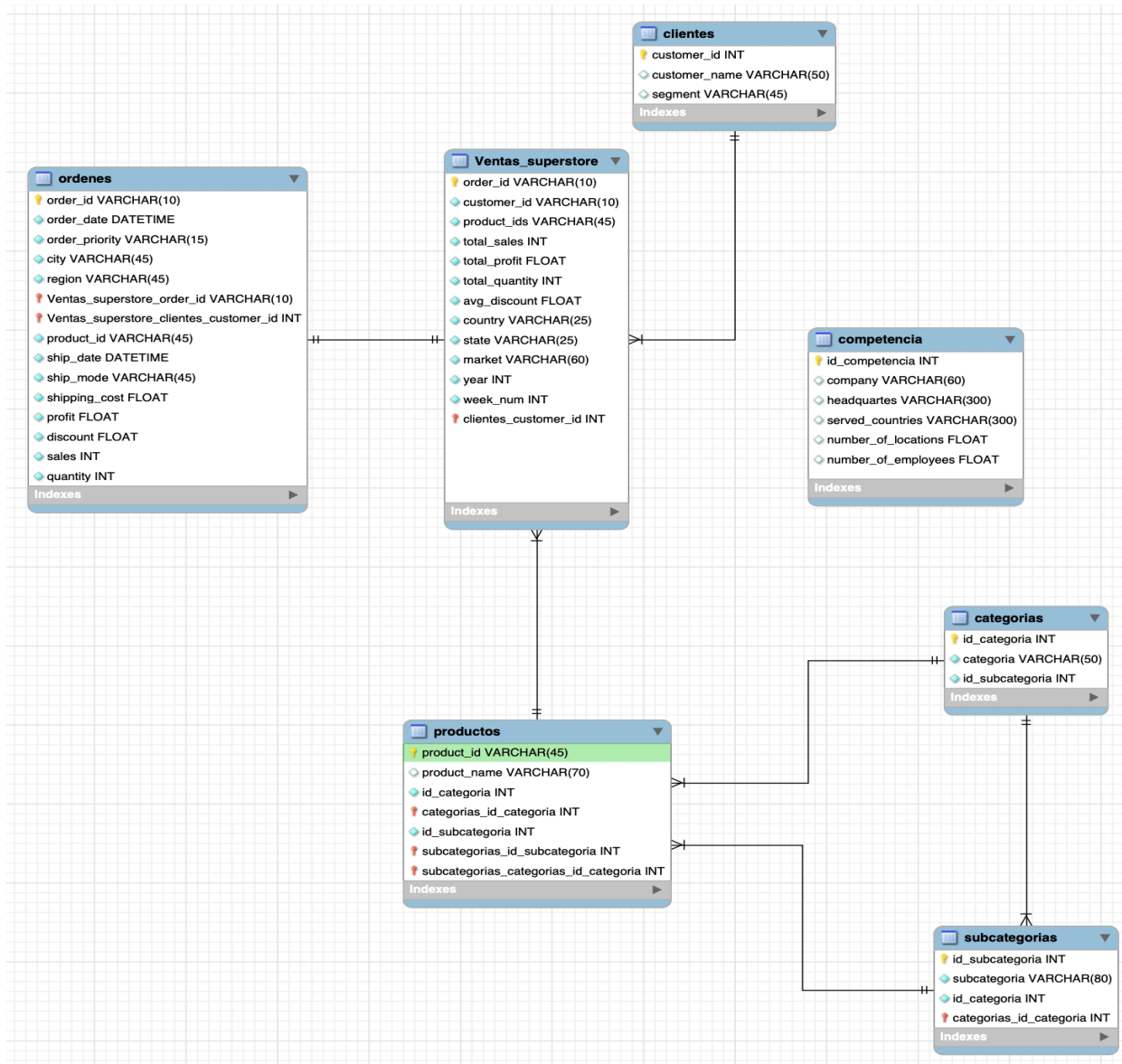
Crear y preparar la Base de Datos:

- Conectar/importar datos a otras herramientas.
- Identificar y manejar valores nulos.
- Identificar y manejar valores duplicados.
- Identificar y manejar datos discrepantes en variables categóricas

- Buscar datos de otras fuentes: Se hizo Web scraping, se utilizó Python para extraer los datos de la tabla "multinacional" de la página https://en.wikipedia.org/wiki/List_of_supermarket_chains, para incluir a los competidores en la estructura de datos de la empresa Super Store. Se eliminaron 59 registros del archivo multinational.csv, que no tenían datos en las columnas de num_empleado, num_locat y served_countries.

Diseñar estructura de la base de datos (tablas de hechos y dimensiones)

Modelo Entidad Relación Superstore



Crear estructura de la base de datos (tablas de hechos y dimensiones)

Se crearon las siguientes tablas con sus respectivos campos.

Clientes: (Dimensión)

Nombre del campo	Tipo
customer_id	STRING
customer_name	STRING
segment	STRING

Categorías: (Dimensión)

Nombre del campo	Tipo
id_categoria	INTEGER
categoria	STRING

Subcategorías: (Dimensión)

Nombre del campo	Tipo
id_subcategoria	INTEGER
subcategoria	STRING

Productos: (Dimensión)

Nombre del campo	Tipo
product_id	STRING
product_name	STRING
id_categoria	INTEGER
id_subcategoria	INTEGER

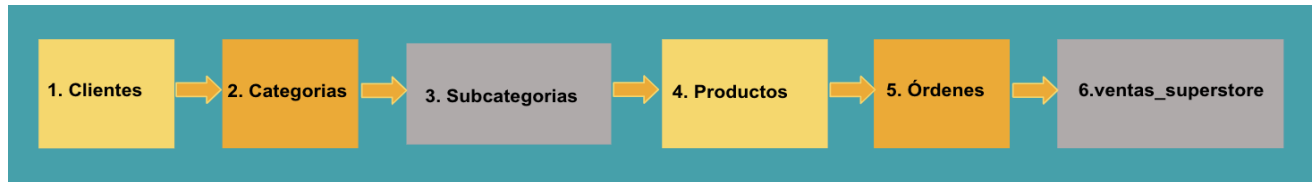
Ordenes : (Dimensión)

Nombre del campo	Tipo
order_id	STRING
order_date	DATE
order_priority	STRING
product_id	STRING
quantity	INTEGER
sales	INTEGER
profit	FLOAT
discount	FLOAT
city	STRING
region	STRING
ship_date	DATE
ship_mode	STRING
shipping_cost	FLOAT

ventas_superstore : (Hechos)

Nombre del campo	Tipo
id_venta	STRING
order_id	STRING
order_date	DATE
customer_id	STRING
product_ids	STRING
total_sales	INTEGER
total_profit	FLOAT
total_quantity	INTEGER
avg_discount	FLOAT
country	STRING
state	STRING
market	STRING
year	INTEGER
weeknum	INTEGER

Flujo de Actualización ETL para base de Datos Superstore.



1. Objetivo del Flujo de Actualización ETL

El objetivo de este flujo de actualización ETL es automatizar la inserción, transformación y carga de datos de diferentes fuentes en una base de datos relacional con tablas de hechos y dimensiones. El flujo asegura que la información de ventas, clientes, productos, competencia y otros datos críticos esté siempre actualizada para análisis y reportes.

2. Descripción de las Fuentes de Datos

Fuente 1: CSV de ventas de superstore

Datos: Ordenes, ventas, clientes, productos.

Frecuencia de actualización: Diaria.

Formato: CSV con columnas: category, city, country, customer_ID, customer_name, discount, market, unknown, order_date, order_id, order_priority, product_id, product_name, profit, quantity, region, row_id, sales, segment, ship_date, ship_mode, shipping_cost, state, sub_category, year, market2, weeknum.

Fuente 2: CSV de Competencia

Datos: Información de empresas competidoras.

Frecuencia de actualización: Mensual.

Formato: CSV con columnas: company, headquarters, number_of_employees, number_of_locations, served_countries.

Fuente 3: API de envíos

Datos: Información sobre el envío de órdenes.

Frecuencia de actualización: Continua.

Formato: JSON con campos : order_id,order_date, ,order_priority, product_id, quantity, sales,profit, discount,city,region,ship_date, ship_mode, shipping_cost.

3. Proceso de Extracción

Fuente 1: CSV de Ventas

El archivo CSV se obtiene desde un sistema de gestión de superstore que exporta datos de ventas a diario.

Herramienta ETL: Se configura una conexión con la carpeta donde se almacenan los archivos CSV diarios.

Fuente 2: CSV de Competencia

La información se extrae una vez al mes desde un archivo CSV proporcionado por el departamento de inteligencia de mercado.

Herramienta ETL: La extracción del CSV se realiza automáticamente una vez al mes.

Fuente 3: API de Envíos

Se conecta a la API de envíos y extrae en tiempo real los datos de las órdenes enviadas.

Herramienta ETL: La herramienta ETL realiza consultas periódicas a la API para obtener actualizaciones.

4. Proceso de Transformación

En esta etapa, los datos crudos se normalizan y preparan para la carga en las tablas de hechos y dimensiones. Se aplican las siguientes transformaciones:

4.1. Normalización de Fechas

Convertir el formato de fecha de las fuentes de ventas y envíos al formato estándar YYYY-MM-DD para asegurar la consistencia en la base de datos.

Origen: order_date, ship_date.

Destino: Tabla de hechos ventas_superstore y tabla ordenes.

4.2. Limpieza de Datos Duplicados

Identificar y eliminar registros duplicados de ventas, clientes y productos. Validar que no haya registros repetidos en el CSV de ventas.

Origen: CSV de Ventas.

Destino: Tabla de hechos ventas_superstore, tabla de dimensión clientes , productos, categorías, subcategorías.

4.3. Cálculo de Nuevas Métricas

Se agrupan todos los productos de una misma orden de la tabla ordenes (Concatenar productos en una sola columna) para obtener la columna product_ids de la tabla ventas_superstore, se suman las ventas por orden de la columna sale de la tabla ordenes y se obtiene la columna total_sale de la tabla ventas_superstore, se suma el profit de la tabla ordenes para obtener el total_profit de la tabla ventas_superstore, se suma quantity de la tabla ordenes para obtener total_quantity de la tabla ventas_superstore.

Origen: CSV de Ventas.

Destino: Tabla de hechos ventas_superstore.

4.4. Mapeo de Relaciones

Asegurar que los campos customer_id, product_id, order_id, id_categoria, id_subcategoria estén correctamente mapeados y no existan valores nulos o no referenciados. Se debe verificar la consistencia de las relaciones entre productos, categorías, subcategorías y clientes.

Origen: CSV de Ventas y CSV de Competencia.

Destino: Tablas de dimensiones (clientes, productos, categorías, subcategorias, competencia).

4.5. Actualización de Categorías y Subcategorías

Verificar que cada producto esté correctamente asignado a una subcategoría y que las subcategorías estén asociadas a la

categoría correspondiente. Si algún producto no tiene una categoría o subcategoría asignada, se debe generar un log de error para su corrección.

Origen: CSV de Productos.

Destino: Tabla de dimensión productos, categorías, y subcategorías.

5. **Proceso de Carga**

Después de las transformaciones, los datos se cargan en las tablas de hechos y dimensiones de la base de datos MySQL.

5.1. Carga de la Tabla de Hechos

Tabla de destino: ventas_superstore

Datos cargados: Datos de ventas, incluyendo order_id, customer_id, product_id, total_sales, total_profit, total_quantity, avg_discount, country, state, year, week_num.

Frecuencia de carga: Diaria.

5.2. Carga de las Dimensiones

- Clientes

Tabla de destino: clientes

Datos cargados: customer_id, customer_name, segment.

Frecuencia de carga: Diaria, tras la limpieza de datos de clientes.

- Productos

Tabla de destino: productos

Datos cargados: product_id, product_name, id_categoria, id_subcategoria.

Frecuencia de carga: Diaria.

- Categorías y Subcategorías

Tabla de destino: categorias, subcategorias

Datos cargados:

Categorías: id_categoria, categoría, id_subcategoria

Subcategorías: id_subcategoria, sub categoria.

Frecuencia de carga: Diaria.

- Competencia

Tabla de destino: competencia

Datos cargados: id_competencia, company, headquarters, number_of_employees, number_of_locations.

Frecuencia de carga: Mensual.

- Ordenes

Tabla de destino: ordenes

Datos cargados: order_id, ship_date, ship_mode, shipping_cost.

Frecuencia de carga: Diaria o continua en tiempo real, dependiendo de la fuente de datos de envío.

6. Proceso de Automatización

Frecuencia de Actualización:

- Ventas y Clientes: Diaria.
- Competencia: Mensual.
- Ordenes: Continua.

Herramientas de Automatización:

Las tareas ETL diarias se automatizan utilizando herramientas como Talend o Airflow. Estas herramientas permiten ejecutar scripts ETL de forma programada y en función de eventos (como la llegada de un nuevo archivo CSV).

Manejo de Errores:


Se configuran alertas para notificar al equipo en caso de fallos en la actualización o transformación de datos. Los errores se registran y almacenan para facilitar su depuración.

7. Monitoreo y Control de Calidad

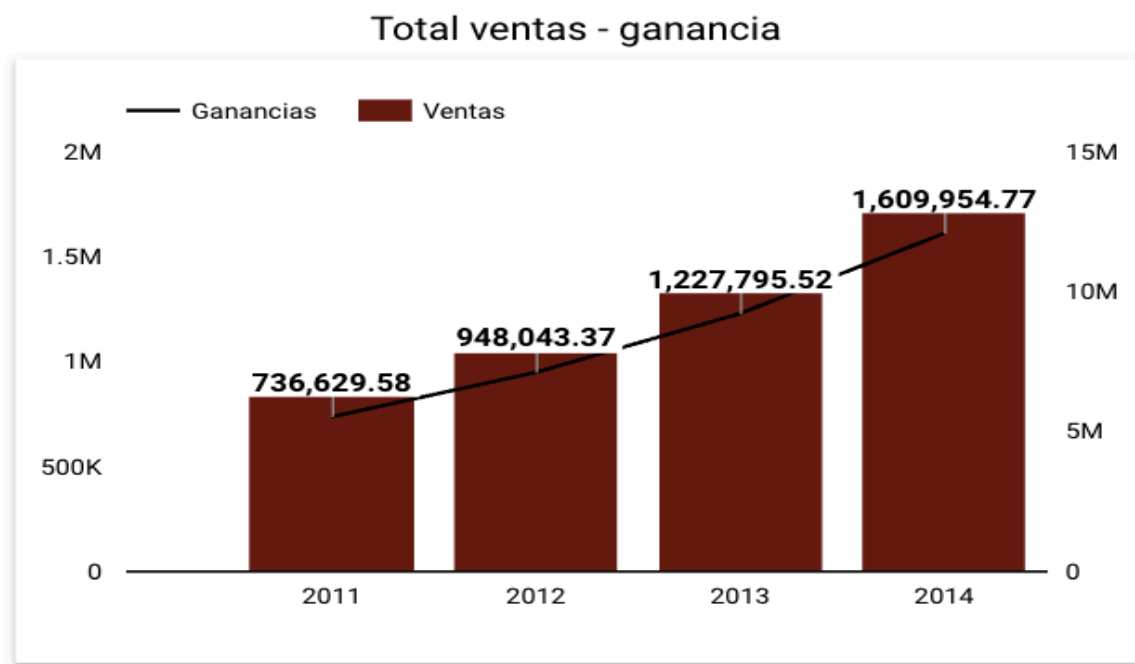
- Se implementan controles de calidad para asegurarse de que los datos estén correctos y actualizados antes de cargarlos.
- Se mantiene un log de las actualizaciones para auditar el proceso de ETL y corregir cualquier error que ocurra durante el proceso.

Análisis Exploratorio

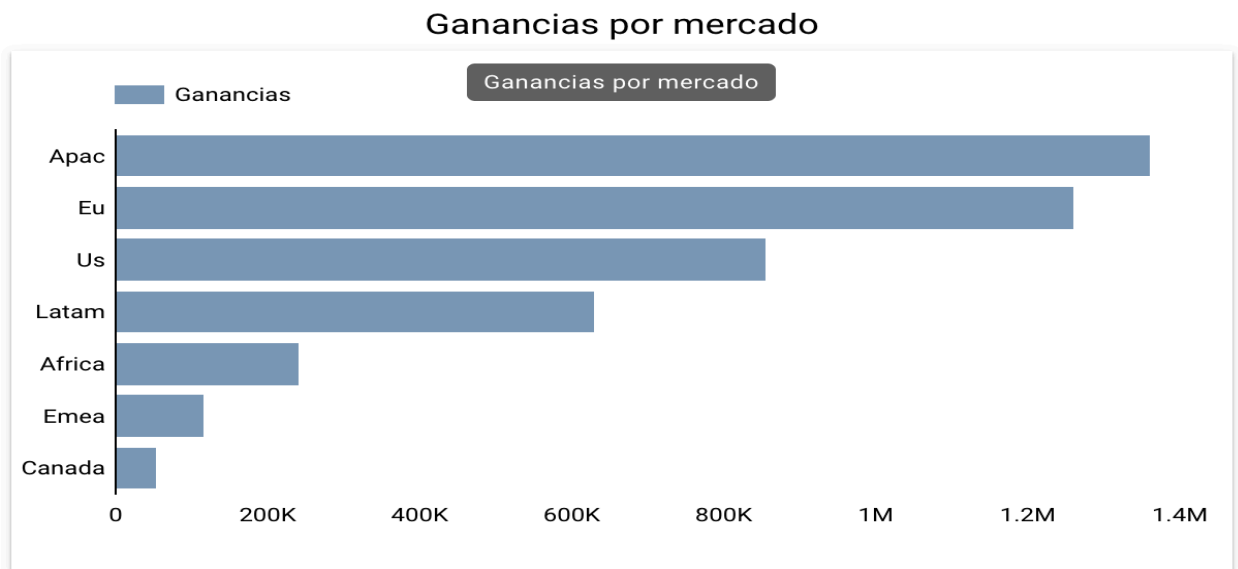
- Obtuvimos los siguientes datos generales de superstore.

							
	Mercados	Total ordenes	Total ventas	Total ganancias	AVG descuentos	Categorías	Subcategorías
	7	25.753	39,4 M	4,5 M	3.641	3	17

- Se puede observar un crecimiento positivo en las ventas y ganancias de Superstore desde 2011 hasta 2014. Las ganancias incrementan año a año, lo que indica una tendencia positiva en este periodo.



- Los mercados APAC, EU y US son los más rentables, es importante mantener e incluso aumentar la inversión en estas regiones. En cambio, África y Canadá continúan siendo los menos rentables, mostrando un bajo rendimiento en términos de ganancias. Superstore debería considerar estrategias para mejorar la presencia en estas regiones.



- Luego de obtener las medidas de tendencia central y dispersión podemos concluir para las métricas de ganancia y ventas:

Medidas de Tendencia central y Dispersión

Mín.Ganancia ▾	Max.Ganancia	AVG Ganancia	STD Ganancia	Min.Ventas	Max.Ventas	AVG Ventas	STD Ventas
-22.746,98	43.811,95	165,58	1.078,17	1	165.620	1.440,84	3.708,79

Ganancias:

Mín. Ganancia: El valor mínimo de ganancia es -22,746.98, lo que indica que algunas transacciones resultaron en pérdidas significativas.

Máx. Ganancia: El valor máximo de ganancia es 43,811.95, lo que refleja una ganancia alta en el mejor escenario.

Promedio (AVG) Ganancia: El promedio de ganancia es 165.58, lo que sugiere que en general las ganancias por transacción son relativamente bajas, comparado con el máximo y la desviación estándar.

Desviación Estándar (STD) Ganancia: La desviación estándar de 1,078.17 muestra una gran dispersión en los datos de ganancia, lo que significa que hay mucha variabilidad entre las transacciones en términos de ganancias y pérdidas.

Existe una alta variabilidad en las ganancias, con algunas transacciones generando ganancias altas y otras ocasionando pérdidas importantes.

El promedio es bajo en comparación con los valores extremos, lo que puede indicar que muchas transacciones tienen ganancias menores o incluso pérdidas.

Ventas:

Mín. Ventas: El valor mínimo de ventas es 1, lo que indica que algunas transacciones tienen volúmenes de ventas extremadamente bajos.

Máx. Ventas: El valor máximo de ventas es 165,620, lo que indica que las mejores transacciones o clientes generan ventas significativamente altas.

Promedio (AVG) Ventas: El promedio de ventas es 1,440.84, lo que sugiere que, en promedio, las ventas son moderadas, aunque mucho más bajas que el valor máximo.

Desviación Estándar (STD) Ventas: La desviación estándar de 3,708.79 indica una alta variabilidad en los volúmenes de ventas, lo que significa que hay muchas transacciones con volúmenes muy diferentes entre sí.

- Las ventas también muestran una alta dispersión, con algunas transacciones generando volúmenes de ventas significativamente altos y otras casi insignificantes. Esto podría indicar una diferencia considerable en el tamaño o tipo de clientes y/o productos.
-
- Tanto las ganancias como las ventas muestran una gran variabilidad en los datos, con altos niveles de dispersión (medidos a través de la desviación estándar).
-
- Los valores promedio son relativamente bajos en comparación con los valores máximos, lo que puede implicar que un pequeño número de transacciones de alto valor está influyendo en el rendimiento total, mientras que muchas otras son de menor escala o incluso resultan en pérdidas.

Conclusiones:

Usar una herramienta ETL o plataforma de integración de datos simplifica y automatiza el flujo de actualización de tu modelo de base de datos. En lugar de manejar todo manualmente, puedes crear pipelines de datos visuales, programar tareas recurrentes y centralizar el monitoreo de tus flujos de actualización. Esto mejora la eficiencia y reduce el riesgo de errores humanos en procesos repetitivos.

Luego del Análisis realizado a los datos de superstore podemos concluir que:

- Mercados como África y Canadá están mostrando un bajo rendimiento en términos de ventas y ganancias. Superstore

debería considerar estrategias para mejorar la presencia en estas regiones, como:

- Mejorar la oferta de productos específicos para estos mercados.
 - Promociones locales o campañas de marketing dirigidas a las necesidades de los clientes locales.
 - Optimizar la logística para reducir costos de envío y mejorar la disponibilidad.
 - Mantener el enfoque en los mercados más rentables:
- Dado que APAC, EU, y US son los mercados más rentables, es importante mantener e incluso aumentar la inversión en estas regiones. Esto puede incluir el lanzamiento de nuevos productos, mejoras en la experiencia del cliente, y la implementación de nuevas tecnologías para mejorar el servicio.
 - Aunque Technology genera ingresos significativos, otras categorías como Furniture y Office Supplies muestran resultados variados. Se recomienda optimizar las subcategorías que generan menos ingresos, como Tables en Furniture o ciertos productos en Office Supplies, ofreciendo promociones o paquetes para mejorar su desempeño.
 - El análisis de competencia muestra una fuerte presencia de competidores en ciertas regiones clave, como Europa y Estados Unidos. Superstore debería vigilar de cerca las estrategias de los competidores en estas regiones para poder diferenciarse mejor. También puede ser una oportunidad expandir su presencia en áreas donde la competencia es menor, como en África o América Latina.

- A pesar de la ligera disminución en las ventas en algunos años, las ganancias han mostrado una tendencia al alza. Superstore debería analizar qué factores están contribuyendo a este aumento de ganancias para replicar estas estrategias en mercados de menor rendimiento.

Recomendaciones:

- Para ciertos análisis críticos, como el monitoreo de rendimiento o seguimiento en tiempo real, es importante contar con un flujo de actualización continua. Asegurarse de que el sistema soporte este tipo de actualización,
- Para análisis menos sensibles al tiempo, se puede programar actualizaciones diarias, semanales o mensuales. La frecuencia de actualización debe depender de las necesidades del análisis y de la frescura de los datos requeridos.
- Para asegurar la trazabilidad y comparar resultados en el tiempo, considerar guardar versiones históricas de los datos. Esto es especialmente útil en análisis como tendencias y proyecciones.
- Usar mecanismos de control de transacciones o procesos de verificación para evitar que existan inconsistencias entre la base de datos original y las actualizaciones que se carguen. Esto es esencial en sistemas de alta concurrencia.
- Implementar procesos de auditoría para revisar el éxito de las cargas de datos. Un buen monitoreo permitirá detectar

problemas (como cargas incompletas o fallas en los datos) antes de que afecten el análisis.

- Documentar el flujo ETL: Documentar las fuentes de datos, las transformaciones realizadas y el flujo de actualización para asegurar que cualquier cambio en el futuro pueda ser realizado de manera controlada.
- Probar el sistema con datos de ejemplo: Antes de implementar el ETL y el flujo de actualización en un entorno de producción, probar con datos de ejemplo para asegurar que el sistema funcione correctamente.
- A medida que los datos aumenten, es crucial optimizar las consultas SQL, el uso de índices y la estructura de la base de datos para mejorar la eficiencia del flujo ETL.
- Asegurarse de que el flujo ETL y de actualización sean escalables. Esto es esencial si se espera que el volumen de datos aumente con el tiempo.
- Proteger los datos sensibles durante todo el proceso, especialmente si se está trabajando con datos personales o confidenciales. Implementar encriptación y controles de acceso.

Limitaciones:

- Falta de datos en la tabla competencia.

Próximos Pasos:

Profundizar en el Análisis, luego de complementar la tabla competencia para hacer comparativos de ventas por mercados.

Enlaces de interés:

Dashboard:

https://lookerstudio.google.com/u/0/reporting/afa01aad-e961-4175-8125-81fb18df782e/page/p_85jj6e3wkd/edit