IT BROTHERS – Consultora especialista en ciencia de datos

INTEGRANTES DEL STAFF:

JIMÉNEZ, JOSE
ALONSO, LUCILA
ARGUMEDO, HÉCTOR
TALAVERA, RICARDO
UMBERT, NORBERTO

ÍNDICE

INTRODUCCIÓN

A continuación, se desarrollará el trabajo realizado durante la segunda semana de proyecto, la cual consistía en la normalización de los datos, creación del pipeline del flujo de trabajo y diseño del Data Warehouse.

Como equipo de trabajo, nos organizamos para llevar a cabo el trabajo cubriendo las necesidades del cliente.

OBJETIVO

El objetivo de esta etapa se basa en automatizar el pipeline de trabajo, tener los datos procesados y almacenados en un Data Warehouse con la estructura adecuada para llevar a cabo su análisis. Es decir, se trata de hacer una limpieza correcta de los datos, la cual nos va a permitir el posterior análisis en una plataforma de visualización adecuada, para realizar los reportes necesarios.

TAREAS REALIZADAS

- Data Lake.

Para implementar el Data Warehouse, nos encargamos de investigar sobre diversas plataformas en la nube y llegamos entre todos a la conclusión de que la herramienta más adecuada sería Google Cloud Platform, la cual es un conjunto de recursos y servicios de computación en la nube pública de Google. Esta plataforma nos ofrece la posibilidad de trabajar con Google Cloud Storage, para almacenar los datos no estructurados, Big Query para realizar consultas SQL, además de que la transmisión de estos están en un formato completamente encriptado, proporciona servicios de copia de seguridad y posee un amplio volumen de memoria.

Los datos fueron provistos por el cliente a través de un bucket dentro de la nube de google cloud desde la cual nuestro equipo tiene el acceso permitido y pudimos llevarlos a nuestro propio bucket, a través de Google Cloud Functions, creando así el Data Lake, es decir, un repositorio de almacenamiento que contiene los datos en bruto y se mantienen allí hasta que sea necesario.

Luego, utilizando nuevamente la herramienta Google Cloud Functions, generamos una carga delta de los datos, lo que genera que sólo se guarden los

datos que no han sido almacenados en nuestro sistema, en caso de que se actualice alguna de las tablas por parte del cliente. Esto proporciona mucha utilidad para evitar sobrecargas en el sistema fuente. Además, en otro bucket propio, se va generando un archivo con el historial de cargas, detallando el archivo con la fecha en que el cliente lo subió, discriminando si es un archivo nuevo para su proceso, o descartando si es un archivo repetido.

A su vez, utilizando nuevamente GCF, se realizó la extracción, transformación y carga, correspondiente a cada dataframe. Este proceso es automático, en la función ya están declaradas las transformaciones necesarias, para que al tomar los nuevos datos se apliquen y se trasladen al bucket donde quedarían los datos listos para stage. Luego, utilizamos Big Query para formar un conjunto de datos, es decir el Data Warehouse, creando tablas desde el Google Cloud Storage con nuestros datos de 'stage' y realizarles consultas en lenguaje SQL, cada vez que sea necesario.

- ETL

Con respecto al ETL, se decidió que se va a trabajar con 9 (nueve) de los datasets contenidos en el análisis exploratorio de los datos, ya que las tablas "Closed_deals" cuenta con una gran cantidad de valores nulos, y la tabla "Marketing_qualified_leads" no contiene información relevante para nuestra propuesta de análisis.

En cuanto a los demás dataframes, se les hicieron las correcciones adecuadas depende de lo que le faltaba, o le sobraba, a cada uno. Es decir, se cambiaron los nombres de las columnas para un mejor entendimiento de las mismas, se agregó una columna con la traducción de sus datos al español (en la tabla Product_category_name), se eliminaron algunos valores duplicados, se separaron algunos datos por región (como en el caso de la tabla "Geolocation"), se realizó el cambio de algunos caracteres (ya que "Sao Pablo", también figuraba como "São Pablo") e incluso se eliminaron emojis, ya que impide la correcta lectura de los datos.

WORKFLOW

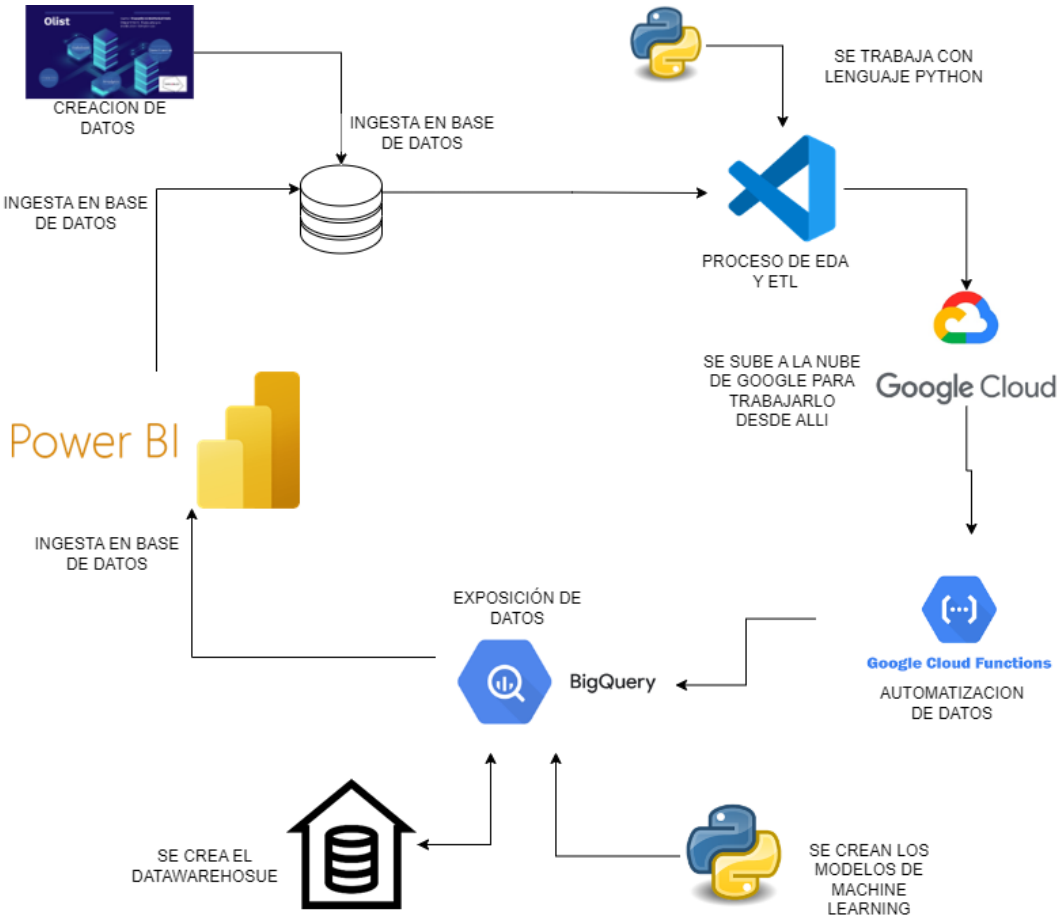
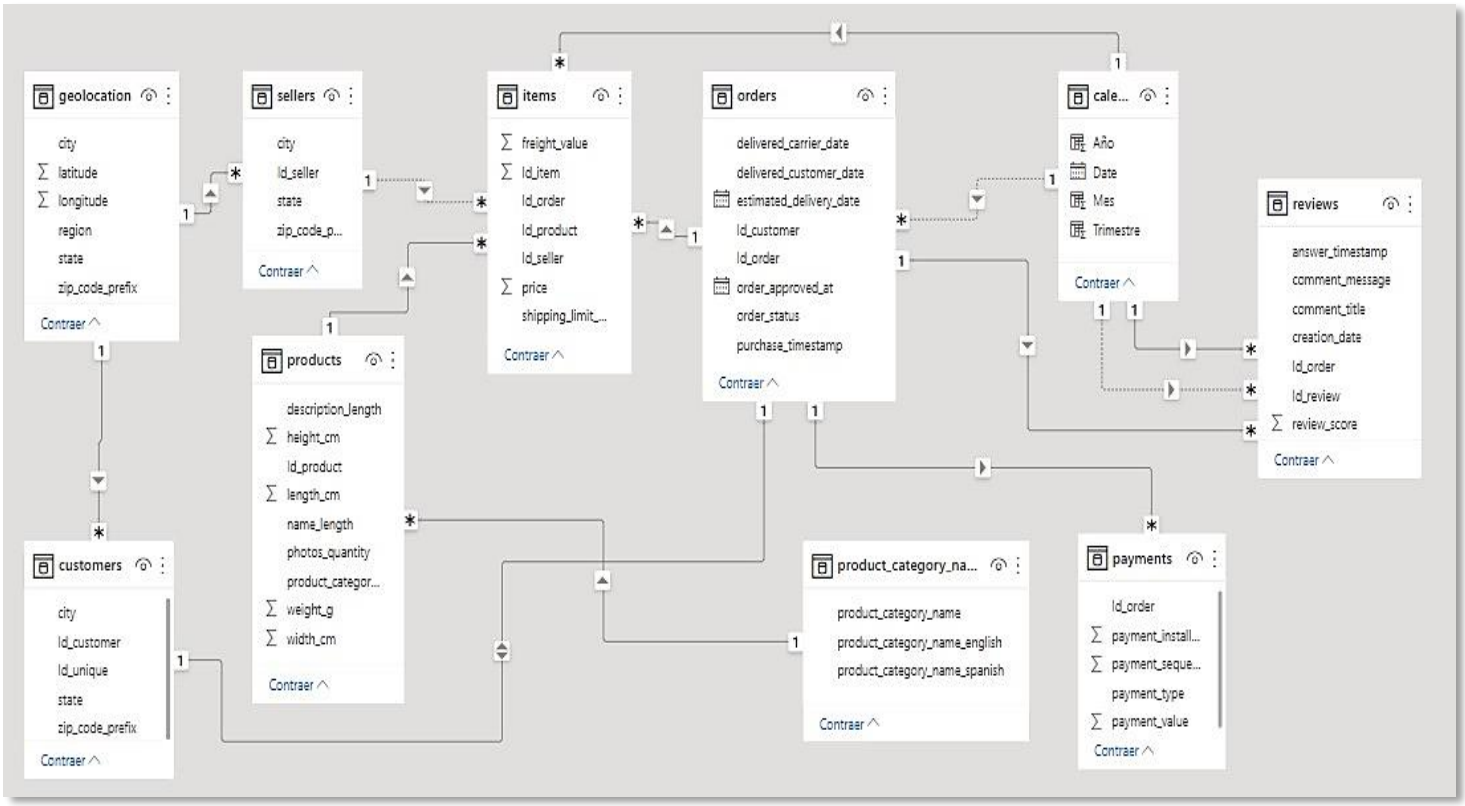


DIAGRAMA ENTIDAD – RELACIÓN

Se puede observar en el siguiente diagrama 3 tablas de hechos (orders, payments y reviews) y 6 tablas de dimensiones (geolocation, sellers, products, products_category_name, customers e items). Además, se agregó una tabla calendario, ya que se va a realizar un análisis basado en fechas.



CRONOGRAMA DE LA SEMANA

Tarea Macro	Microtarea	Dias de elaboracion	hs trabajada	Progreso	Recursos	16-ene	17-ene	18-ene	19-ene	20-ene
SEGUNDA SEMANA	Diseño adecuado del Modelo ER	1	1,5	100,00%	L N R					
	Pipelines para alimentar el DW	1	3	100,00%	N R					
	Data Warehouse	1	5	100,00%	H J					
	Automatización	1	28	100,00%	L H J N R					
	Validación de datos	1	26	100,00%	L H					
	Documentación	3	3	100,00%	L H J N R					
	Diagrama ER detallado (tablas, PK, FK y tipo de dato)	1	1	100,00%	N R					
Data Engineering	Diccionario de datos	1	2	100,00%	N R					
	Workflow detallando tecnologías	1	4	100,00%	H J					
	Elaboración de entregables 2da semana	1	4	100,00%	L H J N R					

ACLARACIONES

Ante lo ya expuesto de las tareas a realizar con el equipo, nos vemos en la determinación de llevar adelante ciertas observaciones en cuanto a la entrega de dichos datos.

La primera observación a hacer es en el formato de los datos, se acuerda con el cliente que tanto la empresa como nuestro equipo entregarán siempre con el mismo formato los archivos, por lo que nos permitiría leerlos, entenderlos y modificarlos si hiciera falta, de una manera fácil y eficiente.

La segunda observación que hacemos es en cuanto al uso de las herramientas a implementar para poder trabajar los datos en la nube. Nuestro equipo considera que resulta más óptimo utilizar las herramientas nativas del entorno de la nube, en este caso GCP, (utilizando Google Cloud Functions) desestimando Airflow para no trabajar con los archivos de manera local, ya que así estamos realizando un flujo de trabajo con mayor practicidad, minimizando costos e integrando con más facilidad.