PROYECTO FINAL

GOOGLE + YELP – Grupo 8

INFORME DE AVANCE 2 – DATA ENGINEERING

Introducción

De acuerdo a lo establecido en la propuesta inicial, en este informe de avance N°2 se presenta la Ingeniería de Datos del Proyecto.
Este es un pilar fundamental en el desarrollo, ya que se desarrollan las bases para que los procesos sean automáticos y el dataware house pueda ser alimentado constantemente con nuevas bases de datos.

Objetivos

A continuación se establecen los objetivos propuestos para este segundo informe de avance:

a) ETL de Datasets de Google Bussiness y Yelp
b) Diccionario de datos
c) Arquitectura Propuesta y Diagrama
d) Creación del DataWare House
e) Automatizar el DataWare House
f) Desarrollar Proceso Carga Incremental.
g) Utilizar un servicio de Nube

A) ETL de las bases de Google Bussiness y YELP!

Antes de la creación del DataWare House y de comenzar con el proceso de Analytics y Machine Learning, es fundamental y necesario, realizar una correcta Limpieza y Transformación de los datos.

En este sentido, en primer lugar se ha realizado un ETL de cada uno de los datasets con los que se trabajará.

Esto incluyó entre otras acciones, eliminación la estandarización de datos, eliminación de columnas, filtrado a los estados con los que se trabajará, filtrado de datos al rubro bar, nightlife y hoteles, procesamiento de nulls en caso de encontrarlos, creación de nuevas columnas, etc.

B) Diccionario De Datos

Una vez realizado el ETL y con una versión final de cada uno de los datasets, se procedió a elaborar los diccionarios de datos

Los mismos se presentan a continuación:

INFORME DE AVANCE 2 – DATA ENGINEERING

DICCIONARIO DE DATOS YELP

i. Diccionario de Bussines

- business_id: Un identificador único para cada negocio en la base de datos.
- name: El nombre del negocio.
- address: La dirección del negocio.
- city: La ciudad donde se encuentra el negocio.
- state: El estado o provincia donde se encuentra el negocio.
- postal_code: El código postal del área donde se encuentra el negocio.
- latitude: La latitud geográfica del negocio.
- longitude: La longitud geográfica del negocio.
- stars: La calificación promedio del negocio en Yelp, en una escala de 1 a 5 estrellas.
- review_count: El número total de reseñas que ha recibido el negocio en Yelp.
- categories: Las categorías o etiquetas que describen el tipo de negocio, separadas por comas.
- Categoría: se refiere a la categoría objetivo encontrada en ese negocio.

ii. Diccionario de checkin

- business_id: Es un identificador único para cada negocio en Yelp.
- num_visitas: Representa número de veces que se registró un check-ins (visita) en el negocio.

iii. Diccionario de review

- user_id (cadena): Es el identificador único del usuario que ha escrito la reseña.
- Cada usuario tiene un id_usuario único que lo distingue de otros usuarios.
- business_id (cadena): Es el identificador único del negocio que recibió la reseña. Cada negocio tiene una Identificación del negocio única que lo distingue de otros negocios.
- stars (entero): Es la calificación de la reseña en términos de estrellas. Puede ser un valor entero entre 1 y 5, donde 1 es la calificación más baja y 5 es la calificación más alta.
- useful (entero): Representa el número de votos útiles que ha recibido la reseña de otros usuarios. Los usuarios pueden votar si una reseña les resultó útil o no.
- funny (entero): Representa el número de votos de otros usuarios que encontraron la reseña divertida. Los usuarios pueden votar si una reseña les resultó divertida o no.
- cool (entero): Representa el número de votos de otros usuarios que encontraron un cierto factor de fascinación, estilo o innovación que atrae la atención y genera una reacción positiva.
- text (cadena): Es el contenido de la reseña escrita por el usuario. Aquí se encuentra el texto completo de la reseña que describe la experiencia del usuario con el negocio.

- Date(fecha): representa a la fecha en la que se hizo la reseña.
- Sentimiento_score (decimal): contiene valores entre-1 y 1. Estos valores representan el sentimiento a la columna text. Un valor cercano a 1 indica un sentimiento positivo fuerte, mientras que un valor cercano a-1 indica un sentimiento negativo fuerte. Los valores cercanos a 0 sugieren un sentimiento neutro o ambiguo en relación con el texto correspondiente.

iv. Diccionario de tips

- user_id: Es un identificador único para cada usuario en Yelp que ha dejado un consejo (tip) en un negocio.
- business_id: Es un identificador único para el negocio en el que se dejó el consejo.
- text: Es el contenido del consejo que el usuario dejó para el negocio. Es un comentario o sugerencia que el usuario comparte.
- Date(fecha): representa a la fecha en la que se hizo la tips.
- Sentimiento_score (decimal): contiene valores entre-1 y 1. Estos valores representan el sentimiento a la columna text. Un valor cercano a 1 indica un sentimiento positivo fuerte, mientras que un valor cercano a-1 indica un sentimiento negativo fuerte. Los valores cercanos a 0 sugieren un sentimiento neutro o ambiguo en relación con el texto correspondiente.

v. Diccionario de user

- user_id: Es un identificador único para el usuario. Es una cadena de texto que identifica al usuario en Yelp.
- name: Es el nombre del usuario. Es una cadena de texto que representa el nombre del usuario en Yelp.
- review_count: Es la cantidad total de reseñas que el usuario ha escrito en Yelp. Es un número entero que indica cuántas reseñas ha dejado el usuario.
- useful: Representa la cantidad total de votos "útiles" que el usuario ha recibido en sus reseñas por parte de otros usuarios. Es un número entero que indica cuántos usuarios han encontrado útiles las reseñas del usuario.
- funny: Representa la cantidad total de votos "graciosos" que el usuario ha recibido en sus reseñas por parte de otros usuarios. Es un número entero que indica cuántos usuarios han encontrado graciosas las reseñas del usuario.
- cool: Representa la cantidad total de votos "frescos" que el usuario ha recibido en sus reseñas por parte de otros usuarios. Es un número entero que indica cuántos usuarios han encontrado un cierto factor de fascinación, estilo o innovación que atrae la atención y genera una reacción positiva
- average_stars: Representa el numero de estrellas asigna por el usuario en sus reseñas

INFORME DE AVANCE 2 – DATA ENGINEERING

DICCIONARIO DE DATOS GOOGLE BUSSINESS

DICCIONARIO DIM_SITIOS_GOOGLE

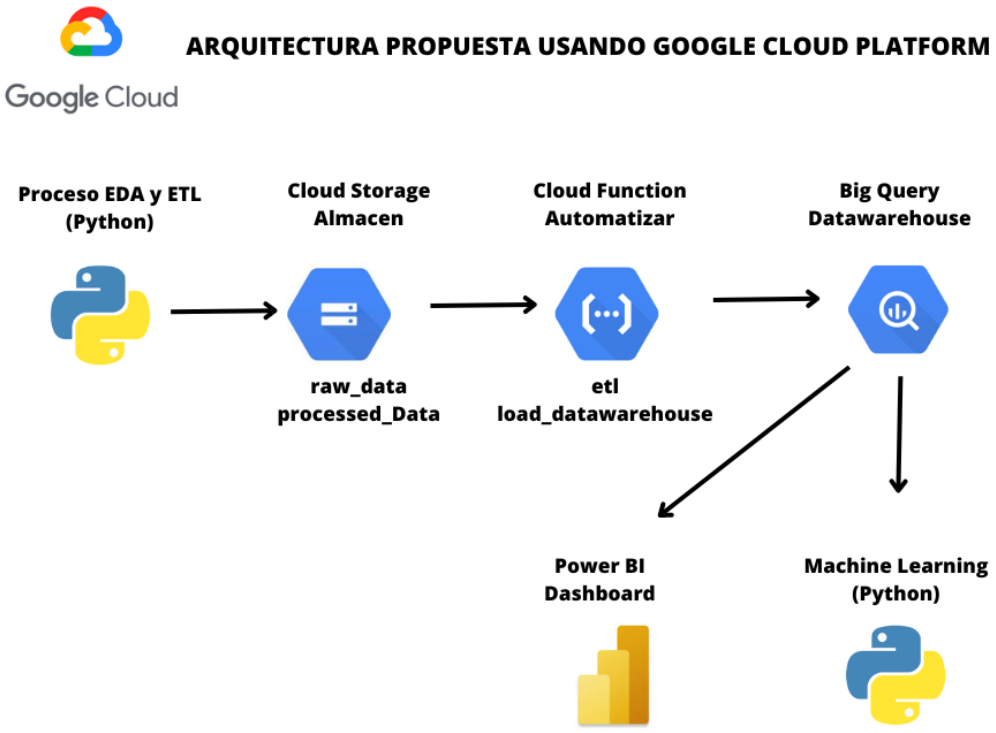
- **name:** nombre del usuario, cadena de texto que representa el nombre del usuario texto
- **address:** La dirección del negocio, texto
- **gmap_id**: código de ubicación global de Google, texto
- **latitude**: ángulo de coordenada geográfica, numero decimal
- **lomgitude**: ángulo de coordenada geográfica, numero decimal
- **categories**: clasificación del establecimiento, texto
- **avg_rating**: promedio de votación del establecimiento, numero decimal
- **num_of_reviews**: número de reseñas del establecimiento, numero entero
- **categoría**: categoría filtrada por Hotel, Bar y Nightlife a la que pertenece el establecimiento, texto

DICCIONARIO REVIEW-GOOGLE

- **user_id**: código del usuario, numero decimal
- **name**: nombre del usuario, texto
- **rating**: votación del usuario al establecimiento, numero entero
- **text**: comentario sobre el establecimiento, texto
- **gmap_id**: código de ubicación global de Google, texto
- **date**: fecha y hora de la reseña, fecha
- **estado**: sigla del estado donde se encuentra el establecimiento, texto
- **sentimiento_score**: contiene valores entre -1 y 1. Estos valores representan el sentimiento a la columna text, numero decimal

INFORME DE AVANCE 2 – DATA ENGINEERING

C) Arquitectura Propuesta y Diagrama



Por la dimensión de los datasets de YELP y GOOGLE BUSINESS que son demasiados grandes, se eligió la arquitectura Big Data.

La tecnología elegida ha sido GOOGLE CLOUD PLATFORM.
Usamos CLOUD STORARE para almacenar los datasets. Tenemos dos buckets:

- RAW_DATA: usado para guardar los datasets originales
- PROCESSED_DATA: usado para guardar los datasets procesado y limpios

Para automatizar usamos CLOUD FUNCTIONS, tenemos dos funciones con lenguaje Python:

- ETL: Realiza el proceso de ETL
- LOAD_DATWAREHOUSE: importa los datasets procesados al DATAWAREHOUSE

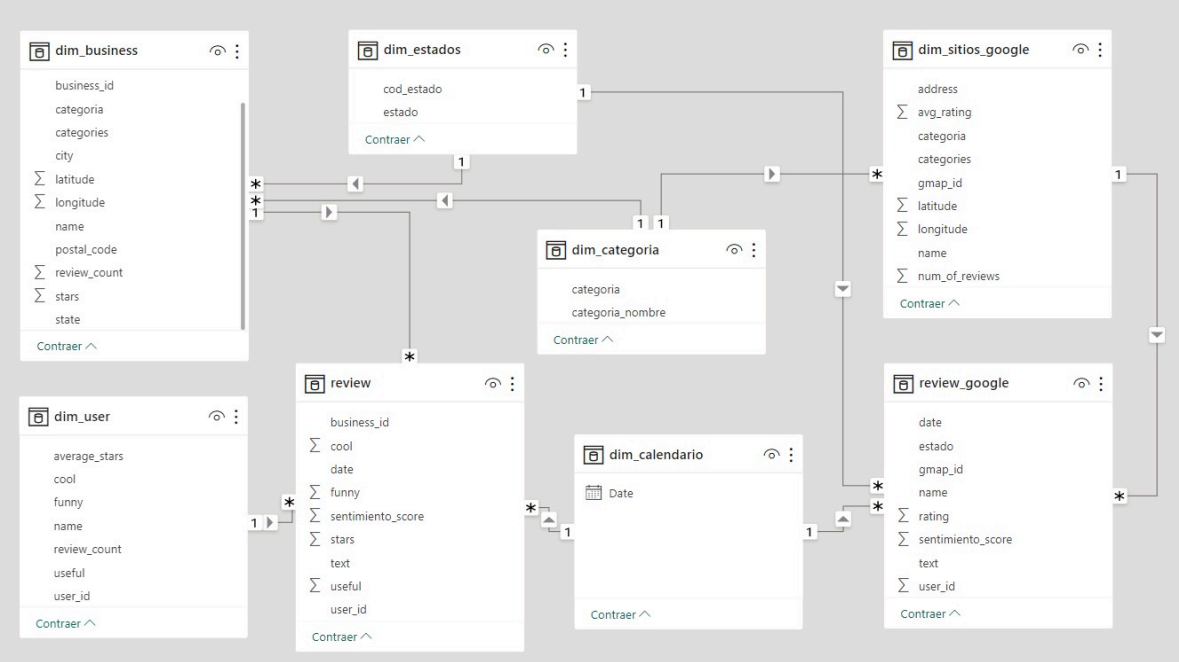
Para crear el DATAWAREHOUSE usamos BIG QUERY, el cual será usado por:

- POWER BI: para diseñar el dashboard
- MACHINE LEARNING: con Python realizamos un sistema de recomendación

INFORME DE AVANCE 2 – DATA ENGINEERING

D) Creación del DataWare House

Se presenta a continuación el Diseño del Data WareHouse.



El Datawarehouse diseñado, posee lo siguiente:

- Tablas Dimensiones:
  - Dim_bussiness
  - Dim_user
  - Dim_estados
  - Dim_categoria
  - Dim_calendario
  - Dim_sitios_google

- Tablas hechos:
  - Reviews
  - Review_google

E) AUTOMATIZAR EL DW

Para automatizar usamos el DW, usamos CLOUD FUNCTIONS, tenemos dos funciones con lenguaje Python:

- ETL: Realiza el proceso de ETL

- LOAD_DATWAREHOUSE: importa los datasets procesados al DATAWAREHOUSE

F) Planificación de Trabajo – Estimación de Esfuerzos

Por último, se deja como recordatorio el cronograma de proyecto.

Tabla 1. Diagrama de Gannt

Tareas	Sprint 1								Sprint 2									Sprint 3									
	Lu.	Ma.	Mi.	Ju.	Vi.	Lu.	Ma.	Mi.	Ju.	Vi.	Lu.	Ma.	Mi.	Ju.	Vi.	Lu.	Ma.	Mi.	Ju.	Vi.	Lu.	Ma.	Mi.	Ju.	Vi.	Lu.	Ma.	Mi.
	31	1	2	3	4	7	8	9	10	11	14	15	16	17	18	21	22	23	24	25	28	29	30	31	1	4	5	6
PUESTA EN MARCHA																												
Kick OFF																												
Asignación de Roles																												
Alcance y establecimiento de los objetivos																												
Creación Repositorio																												
Definición KPIs																												
EDA Preliminar - Analisis Datos																												
Definición Tecnología a utilizar																												
Elaboración del Informe																												
Implementación del Stack Propuesto																												
Reunión Revisión Pre - Entrega																												
Correciones Finales																												
Entrega Informe de Avance 1																												
DATA ENGINEERING																												
Arquitectura Propuesta y Diagrama																												
Diccionario de Datos																												
Creación del DW																												
Automatizar el DW																												
Carga Incremental de Datos																												
Reunión Revisión Pre - Entrega																												
Correciones Finales																												
Entrega Informe de Avance 2																												
ANALYTICS Y MACHINE LEARNING																												
Diseño de Reportes																												
DashBoards																												
Modelo Machine Learning																												
Modelo Machine Learning en Producción																												
Elaboración del Informe																												
Reunión Revisión Pre - Entrega																												
Correciones Finales																												
Entrega Informe Final																												