RECOMMENDATION SYSTEM: FAST FOOD FRANCHISES

MINDFUL DATA

Indice

Introducción

De acuerdo a la propuesta inicial presentada la semana pasada, en este informe de avance N°2 se presenta la parte de ingeniería de datos del proyecto. Este es un pilar fundamental en el desarrollo, que permite la automatización del proceso y la carga de nuevos datos.

Etapas de trabajo.

Tal y como se presentó en el informe anterior, nos encontramos en la segunda semana de trabajo correspondiente a la etapa de Data Engineering.

Durante esta etapa se crea, se implementa y se automatiza el datawarehouse. Además, se plantean los MVPs del dashboard y de los modelos de Machine Learning preliminares.

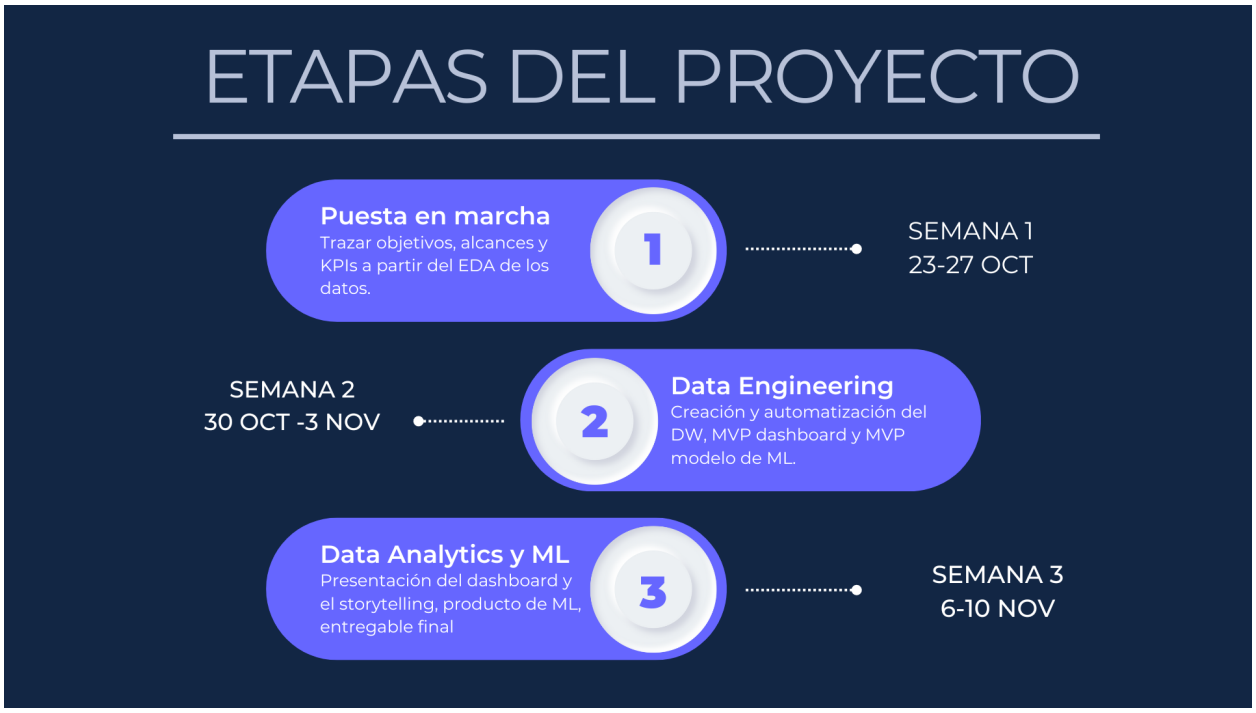
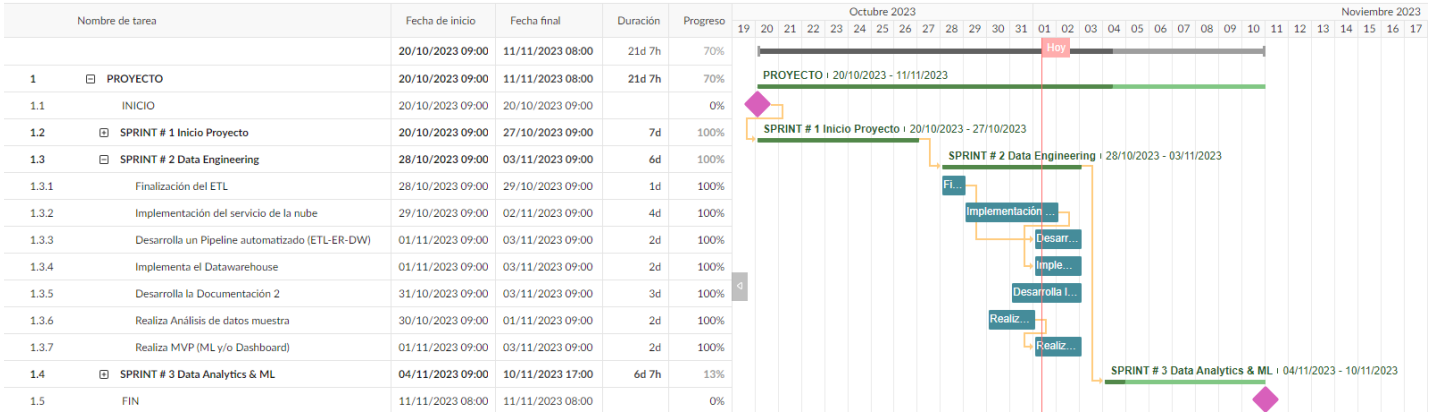


Diagrama de Gantt - Semana 2

A continuación se presenta el diagrama de Gantt con las actividades planificadas para la segunda semana de trabajo.



	Nombre de tarea	Fecha de inicio	Fecha final	Duración	Progreso
		20/10/2023 09:00	**11/11/2023 08:00**	21d 7h	70%
1	**PROYECTO**	**20/10/2023 09:00**	**11/11/2023 08:00**	**21d 7h**	70%
1.1	INICIO	20/10/2023 09:00	20/10/2023 09:00		0%
1.2	**SPRINT # 1 Inicio Proyecto**	**20/10/2023 09:00**	**27/10/2023 09:00**	**7d**	100%
1.3	**SPRINT # 2 Data Engineering**	**28/10/2023 09:00**	**03/11/2023 09:00**	**6d**	100%
1.3.1	Finalización del ETL	28/10/2023 09:00	29/10/2023 09:00	1d	100%
1.3.2	Implementación del servicio de la nube	29/10/2023 09:00	02/11/2023 09:00	4d	100%
1.3.3	Desarrolla un Pipeline automatizado (ETL-ER-DW)	01/11/2023 09:00	03/11/2023 09:00	2d	100%
1.3.4	Implementa el Datawarehouse	01/11/2023 09:00	03/11/2023 09:00	2d	100%
1.3.5	Desarrolla la Documentación 2	31/10/2023 09:00	03/11/2023 09:00	3d	100%
1.3.6	Realiza Análisis de datos muestra	30/10/2023 09:00	01/11/2023 09:00	2d	100%
1.3.7	Realiza MVP (ML y/o Dashboard)	01/11/2023 09:00	03/11/2023 09:00	2d	100%
1.4	**SPRINT # 3 Data Analytics & ML**	**04/11/2023 09:00**	**10/11/2023 17:00**	**6d 7h**	13%
1.5	FIN	11/11/2023 08:00	11/11/2023 08:00		0%

Objetivos

Como mencionamos anteriormente este proyecto tiene como objetivo recomendar una franquicia del rubro fast food, en base a un análisis de los negocios, las reviews y los usuarios de Yelp y Google Maps del estado de Florida, en los últimos 5 años.

Específicamente en esta semana los objetivos a cumplir son:

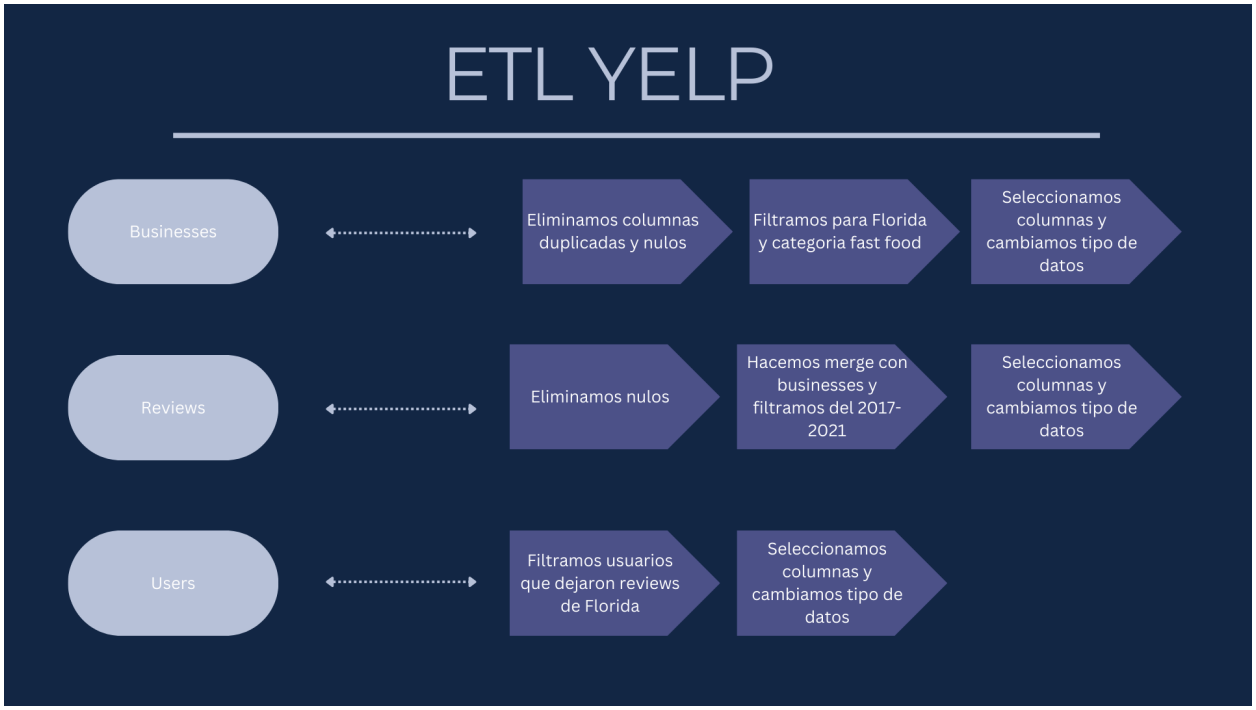
- Realizar el ETL de datasets de Google Bussiness y Yelp.
- Generar el diccionario de datos
- Presentar la arquitectura propuesta.
- Creación del datawarehouse y el proceso de automatización del mismo.
- Desarrollar un proceso de carga incremental
- Validar los datos ingresados.

ETL completo

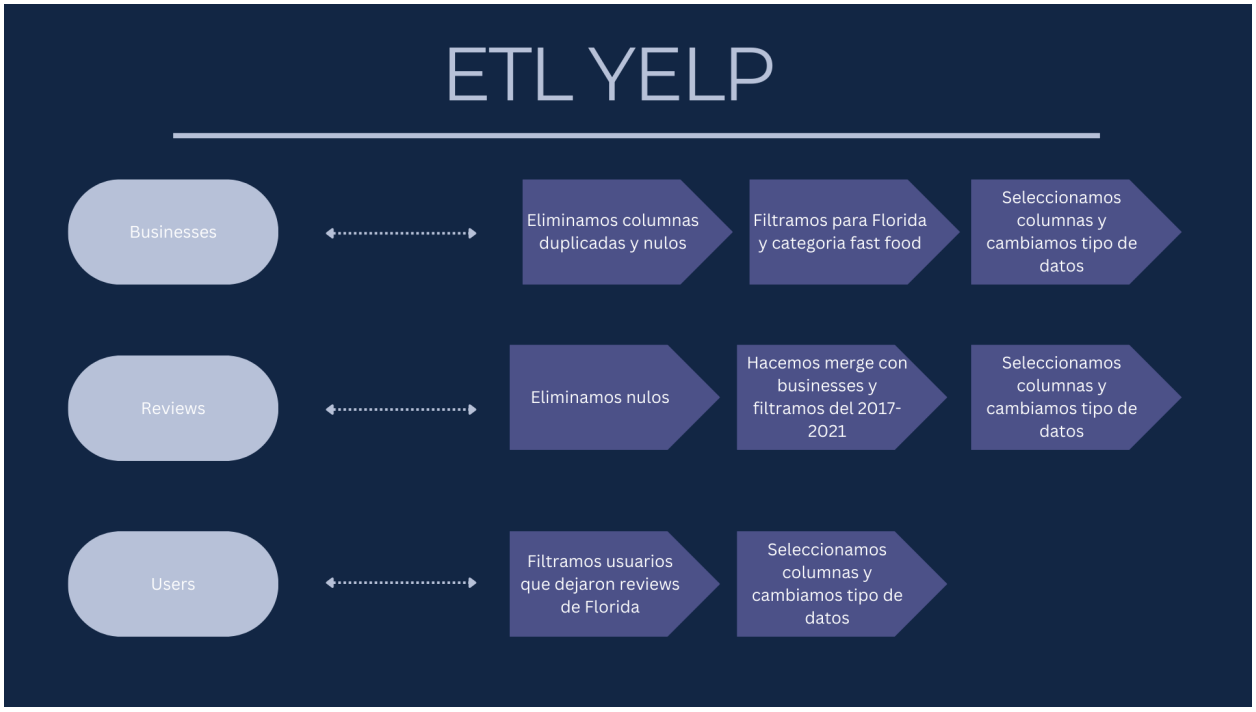
Antes de establecer un Data Warehouse y dar inicio al proceso de Analytics y Machine Learning, resulta esencial llevar a cabo una adecuada depuración y transformación de los datos. En este contexto, en primer lugar, se efectuó un proceso ETL para cada uno de los conjuntos de datos con los que se va a trabajar.

El ETL se realiza en primera instancia de manera local, para crear las funciones que nos van a permitir automatizar el proceso.

Para la información proveniente de Yelp, el ETL para cada dataset lleva los siguientes pasos:



Para los archivos json provenientes de Google Maps, se realizaron los siguientes pasos:



Diccionario de datos

1. Diccionario de Business de Yelp

Columna	Descripción
Business_id (str)	Identificador único para cada negocio en la base de datos.
Name (str)	Nombre del negocio.
City (str)	Ciudad donde se encuentra el negocio.
Latitude (float)	Latitud geográfica del negocio.
Longitude (float)	Longitud geográfica del negocio.
Stars (float)	Calificación promedio del negocio en Yelp, en una escala de 1 a 5 estrellas.
Review_count (int)	Número total de reseñas que ha recibido el negocio en Yelp.

2. Diccionario de Reviews de Yelp

Columna	Descripción
Review_id (str)	Identificador único para cada review.
User_id (str)	Identificador único del usuario que ha escrito la reseña.
Business_id (str)	Identificador único del negocio que recibió la reseña.
Stars (float)	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.
Text (str)	Es el contenido de la reseña escrita por el usuario.
Date (date)	Representa a la fecha en la que se hizo la reseña.
City (str)	Ciudad donde se encuentra el negocio.
Sentiment_analysis (int)	Estos valores representan el sentimiento de la columna text. 0 es negativo, 1 es neutro y 2 es positivo.

3. Diccionario Users de Yelp

Columna	Descripción
User_id (str)	Identificador único del usuario.
Name (str)	Nombre del usuario.
Review_count (int)	Número total de reseñas que ha dejado el usuario.
Average_stars (float)	Promedio de stars que ha dejado el usuario.

4. Diccionario de business de Google Maps

Columna	Descripción
Gmap_id (str)	Código de ubicación global de Google Maps.
Name (str) (str)	Nombre del negocio.
City (str)	Ciudad donde se encuentra el negocio.

Postal_code (float)	Código postal de la ciudad donde se encuentra el negocio.
Latitude (float)	Latitud geográfica del negocio.
Longitude (float)	Longitud geográfica del negocio.
Category (str)	Categorías a las que pertenece ese negocio.
Avg_rating (float)	Representa el promedio de puntuaciones del negocio.
Num_of_reviews (int)	Número de reseñas del negocio.

5. Diccionario de Reviews de Google Maps

Columna	Descripción
Review_id (str)	Identificación única de la review.
User_id (float)	Identificación única del usuario.
Name (str)	Nombre del usuario.
Rating (int)	Representa las puntuaciones del usuario al negocio.
Text (str)	Es el contenido de la reseña escrita por el usuario.
Gmap_id (str)	Código de ubicación global de Google Maps.
Date (date)	Fecha y hora de la reseña.
Sentiment_analysis (int)	Estos valores representan el sentimiento de la columna text.

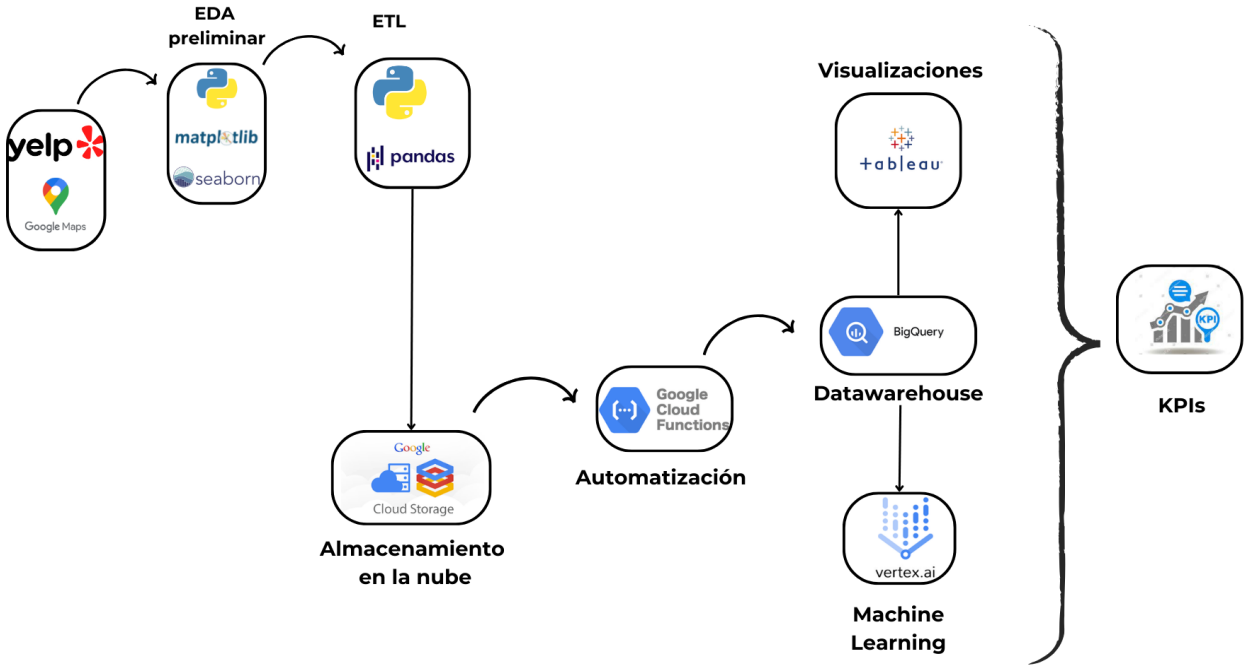
6. Diccionario de Calendar

Columna	Descripción
Fecha (date)	Fecha del calendario.
Año (int)	Año del calendario.
Trimester (int)	Trimestre del calendario.
Month (int)	Mes del calendario.
Day (número) (int)	Número del día de la semana para el calendario.
Day (nombre) (str)	Nombre del día de la semana para el calendario.

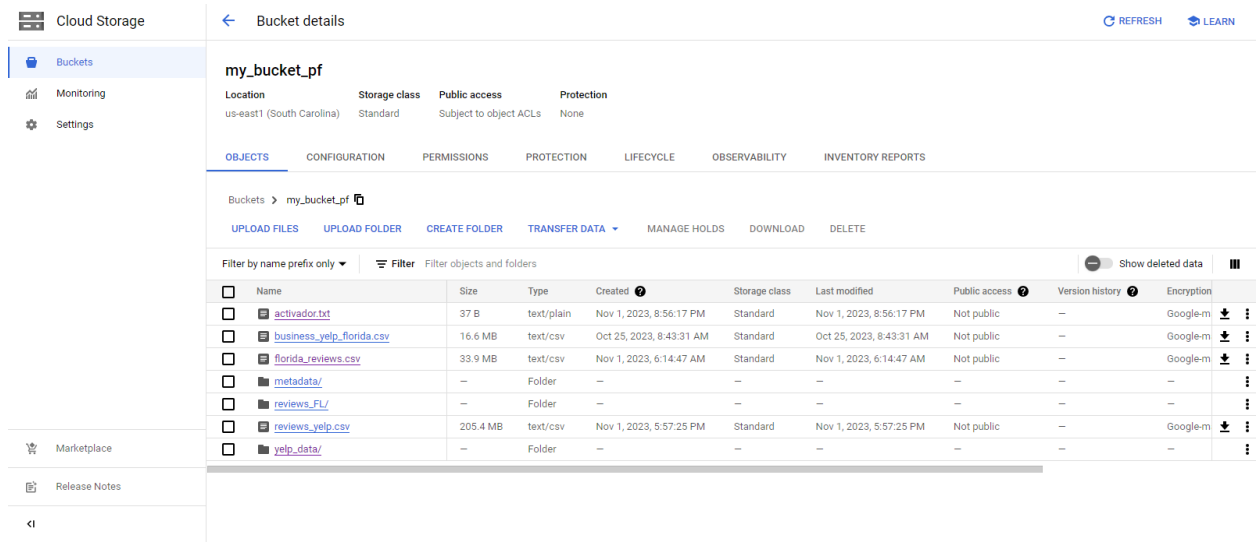
7. Diccionario de Ciudades

Columna	Descripción
City_id	Identificación única de la ciudad.
City_name	Nombre de la ciudad.
Latitude (float)	Latitud geográfica de la ciudad.
Longitude (float)	Longitud geográfica de la ciudad.

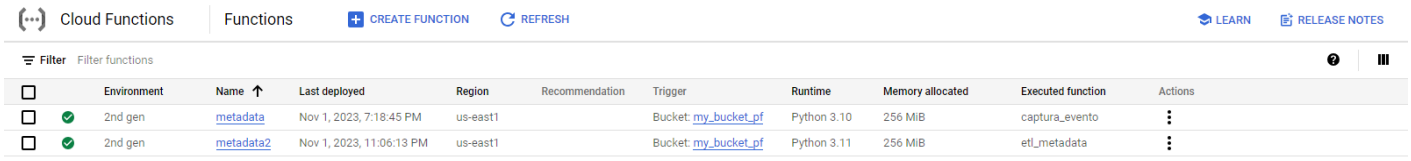
Arquitectura planteada



- Data Lake:
  Disponibilizamos los archivos crudos en el Google Cloud Storage. El mismo que se utiliza para describir la arquitectura de almacenamiento de datos en formato json que permite almacenar grandes volúmenes de datos en su formato original, sin estructuración previa.
  En el data lake, cargamos los archivos RAW o crudos.

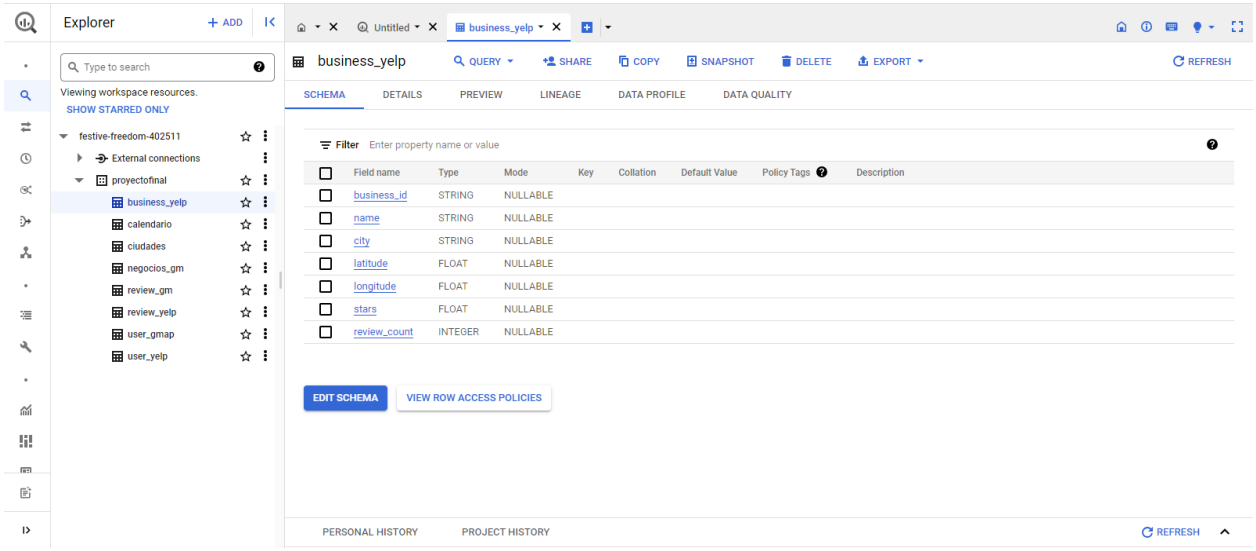


- Cloud Functions:
  Cloud Functions es un servicio de computación sin servidor que nos permite ejecutar código en respuesta a eventos.



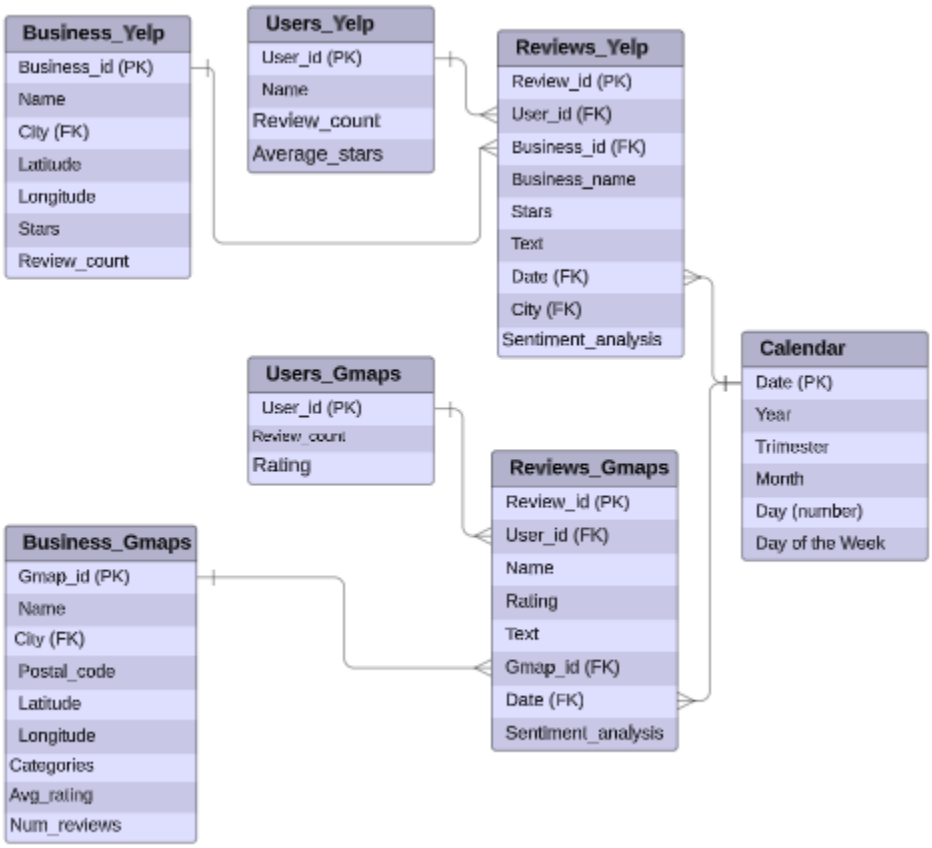
- Data Warehouse:
  Después de realizar las transformaciones y limpieza de los datos, el data frame de pandas lo guarda como un archivo CSV temporal y se carga en una tabla de BigQuery. Este actúa como un almacén de datos, también conocido como data warehouse.

  En nuestro Data Warehouse almacenamos los datos estructurados de manera eficiente para su posterior análisis y consulta.



Data Warehouse

Se presenta a continuación el diagrama entidad-relación de las tablas que componen el datawarehouse Big Query.



El Datawarehouse diseñado, posee la siguiente información:

- Tablas Dimensiones:
    - Bussiness_yelp
    - Business_gmaps
    - Users_yelp
    - Users_gmaps
    - Calendario

- Tablas hechos:
    - Reviews_yelp
    - Reviews_gmaps

Automatización

En esta instancia, aún nos encontramos trabajando en la automatización de las funciones del Cloud Function de manera que el proceso funcione de manera automática a partir de un activador.
Hemos tenido inconvenientes con la implementación de la automatización del ETL, por lo que la carga incremental estará definida y presentada para la etapa final del proyecto.

MVP del dashboard

Para la creación del dashboard se retoman los KPis presentados la semana anterior:

- *Crecimiento de Reseñas Positivas:* Este KPI se centra en el aumento porcentual en el número de reseñas positivas en comparación con el año anterior
  Objetivo: Aumento del 5% en el número de reseñas positivas de Yelp y Google para los negocios de Fast food en comparación al año anterior.

  Fórmula: (Cantidad de reseñas positivas año actual - Cantidad de reseñas positivas año anterior) /(reseñas anterior)

- *Disminución de Reseñas Negativas:* Este KPI se centra en la disminución porcentual en el número de reseñas positivas en comparación con el año anterior.
  Objetivo: Disminución del número de reseñas negativas del 5% de Yelp y Google para los negocios de Fast food es menor en comparación al año anterior .

  Fórmula: (Cantidad de reseñas negativas año anterior - Cantidad de reseñas negativas año actual) /(reseñas anterior)

- *Índice de satisfacción de usuarios de los principales negocios de fast food (según Yelp y Google)* para recomendar una cadena de fast food: Mide la satisfacción global de los usuarios con un negocio, combinando las calificaciones y ponderaciones de las reseñas de los usuarios en las plataformas Yelp y Google.
  Objetivo: Aumentar el Índice de Satisfacción del Usuario en un 10% durante el próximo semestre.
  Fórmula:
  Índice de Satisfacción del Usuario = (Ponderación_Yelp * Rating_Yelp + Ponderación_Google * Rating_Google) / (Ponderación_Yelp + Ponderación_Google)

  - Ponderación_Yelp: Peso en proporción al total (1- total google)
  - Rating_Yelp: Puntuación promedio de Yelp para un restaurante o categoría de restaurantes.
  - Ponderación_Google: Peso en proporción del total (1- total yelp)
  - Rating_Google: Puntuación promedio de Google para un restaurante o categoría de restaurantes.
- *Aumento la cantidad de reseñas por restaurante fast-food:* Mide la interacción en el restaurante (o restaurantes, si tiene varias sucursales). Suma el total de reseñas que se obtuvieron.
  Objetivo: Elevar la cantidad de reseñas por restaurante por lo menos un 15% en comparación con el año anterior.
  Fórmula: (Cantidad de reseñas año anterior - Cantidad de reseñas año actual) /(reseñas anterior)

- *Aumento de la tasa anual de retención de usuarios:* Mide la tasa de usuarios que escriben reseñas año a año.
  Objetivo: Aumentar la tasa anual de usuarios que dejan reseñas en un 5% en comparación al año anterior.
  Fórmula: (Usuarios que dejan reseñas en el año actual - usuarios que dejan reseñas en el año anterior) / usuarios que dejan reseñas en el año anterior.

KPI	Descripción	Fórmula	Objetivo
Crecimiento de reseñas positivas	Este KPI se centra en el aumento porcentual en el número de reseñas positivas en comparación con el año anterior	[(Cantidad de Reseñas Positivas en el Año Actual - Cantidad de Reseñas Positivas en el Año Anterior) / Cantidad de Reseñas Positivas en el Año Anterior] x 100	Aumento del 5% en el número de reseñas positivas para los negocios de Fast food en comparación al año anterior.
Disminución de reseñas negativas	Este KPI se centra en la disminución porcentual en el número de reseñas positivas en comparación con el año anterior	[(Cantidad de Reseñas negativas en el Año Anterior - Cantidad de Reseñas negativas en el Año Actual) / Cantidad de Reseñas Negativas en el Año Anterior] x 100	Disminución del 5% en el número de reseñas negativas de los negocios de Fast food es menor en comparación al año anterior .
Satisfacción de Clientes en Fast Food	Mide la satisfacción global de los usuarios un negocio, combinando las calificacione ponderaciones de las reseñas de los usuarios en las plataformas Yelp y Google.	Ponderación_Google * Rating_Google) / (Ponderación_Yelp + Ponderación_Google)	**Aumentar el Índice de Satisfacción del Usuario en un 10% durante el próximo semestre**
Aumentar la cantidad de reseñas por restaurante	Mide la interacción en el restaurante (o restaurantes, si tiene varias sucursales). Suma el total de reseñas que se obtuvieron.	(Cantidad de reseñas año anterior - Cantidad de reseñas año actual) /(reseñas anterior)	**Elevar la cantidad de reseñas por restaurante por lo menos un 15% en comparación con el año anterior.**
Aumento de la tasa anual de retención de usuarios	Mide la tasa de usuarios que escriben reseñas año a año.	(Usuarios que dejan reseñas en el año actual - usuarios que dejan reseñas en el año anterior) / usuarios que dejan reseñas en el año anterior	**Aumentar la tasa anual de usuarios que dejan reseñas en un 5% en comparación al año anterior.**

Cabe mencionar que en esta instancia, se presenta una versión simplificada del dashboard que se planea implementar, incluida la conexión con DW. Se incluyen algunas de las visualizaciones preliminares y datos de muestra.