

RECOMMENDATION SYSTEM: FAST FOOD FRANCHISES



MINDFUL DATA

Indice

Indice.....	2
Introducción.....	3
Etapas de trabajo.....	3
Diagrama de Gantt - Semana 2.....	3
Objetivos.....	4
ETL completo.....	4
Diccionario de datos.....	6
Arquitectura planteada.....	9
Data Warehouse.....	11
Automatización.....	12
MVP del dashboard.....	12
MVP Sistema de Recomendación.....	15

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.

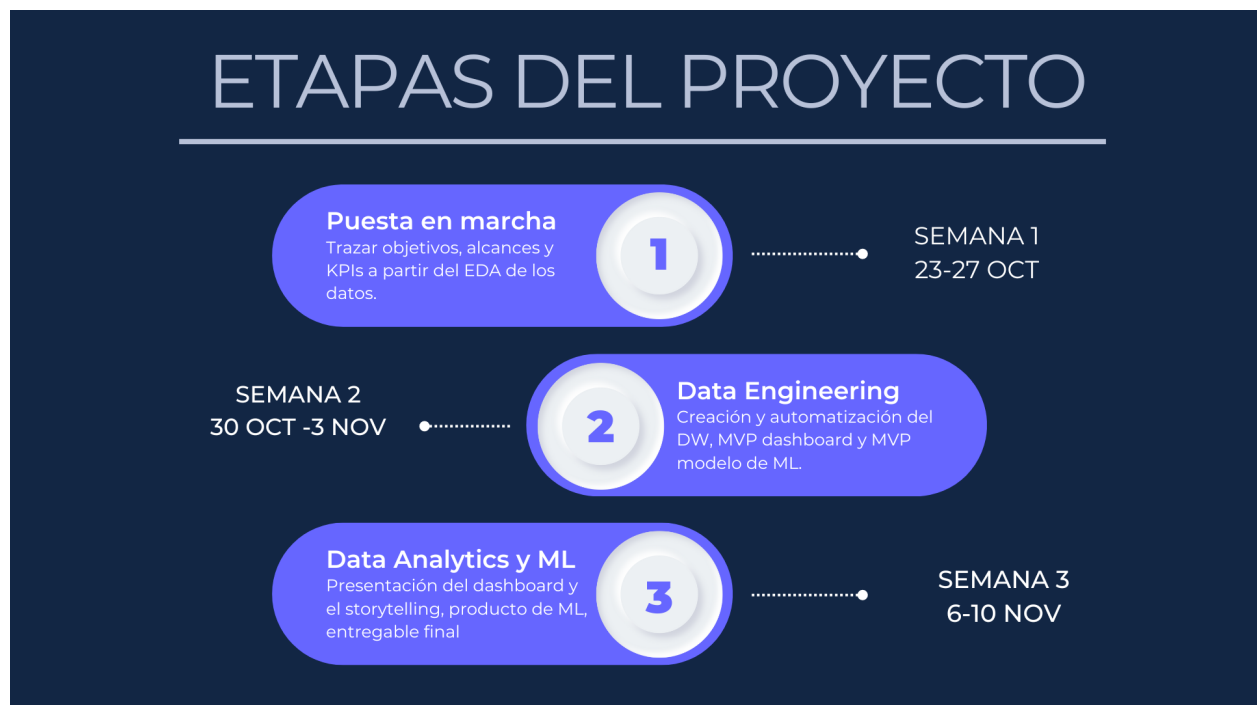
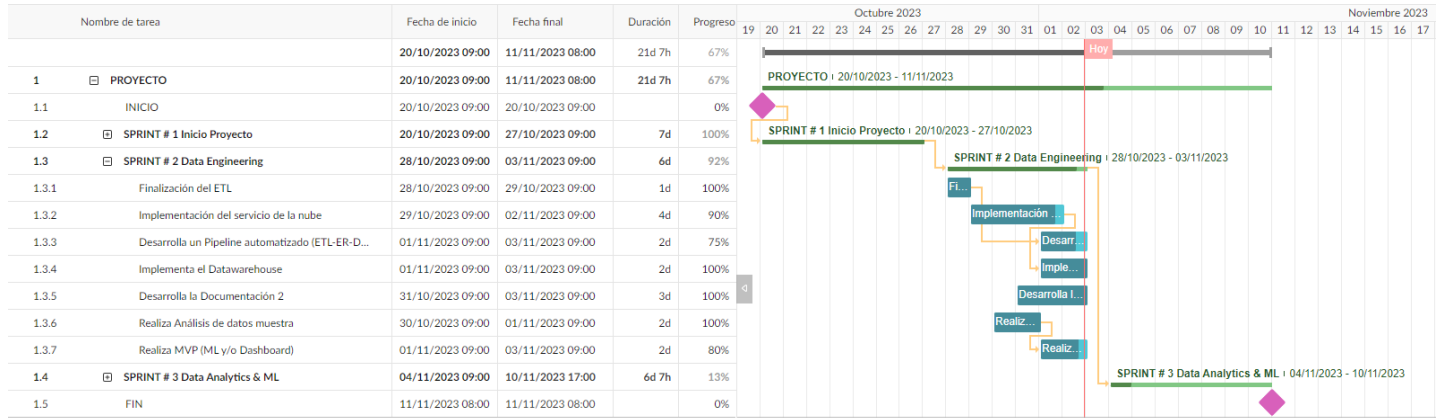


Diagrama de Gantt - Semana 2

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

INFORME DE AVANCE: PROYECTO FINAL - SEMANA 2



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:

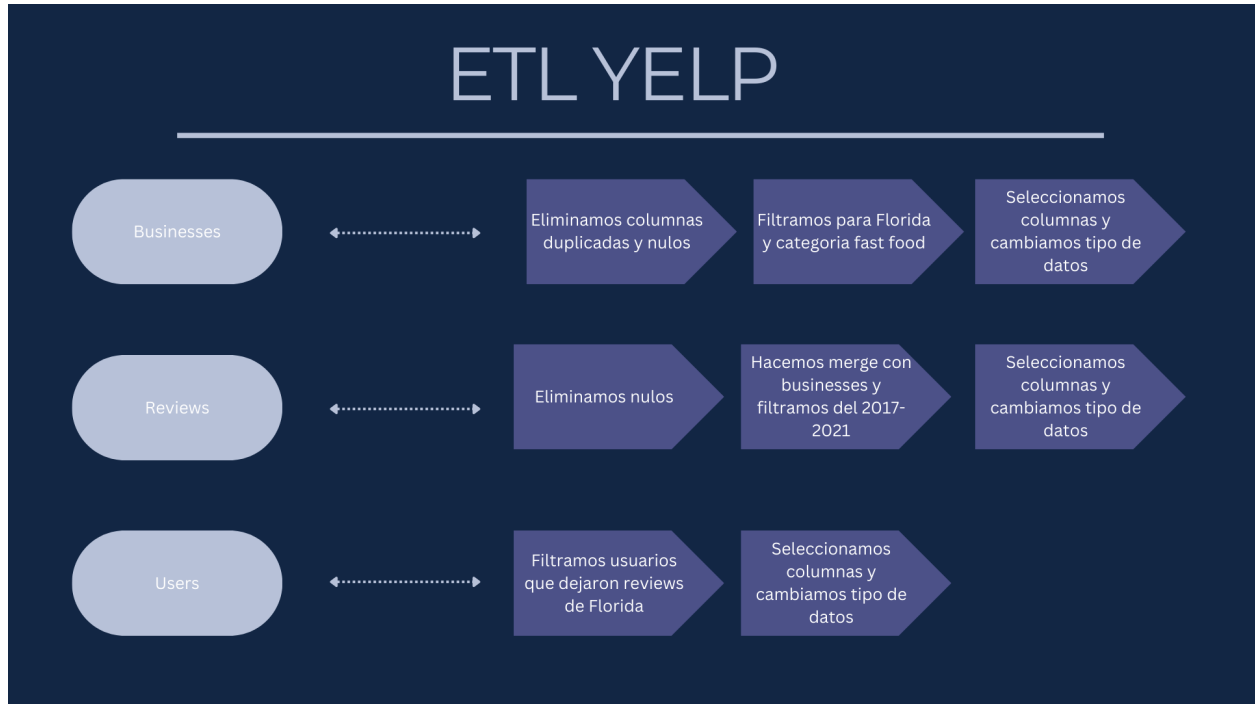
- Realizar el ETL de datasets de Google Bussiness y Yelp.
- Generar el diccionario de datos
- Presentar la arquitectura propuesta.
- Creación del datawarehouse y el proceso de automatización del mismo.
- Desarrollar un proceso de carga incremental
- Validar los datos ingresados.

ETL completo

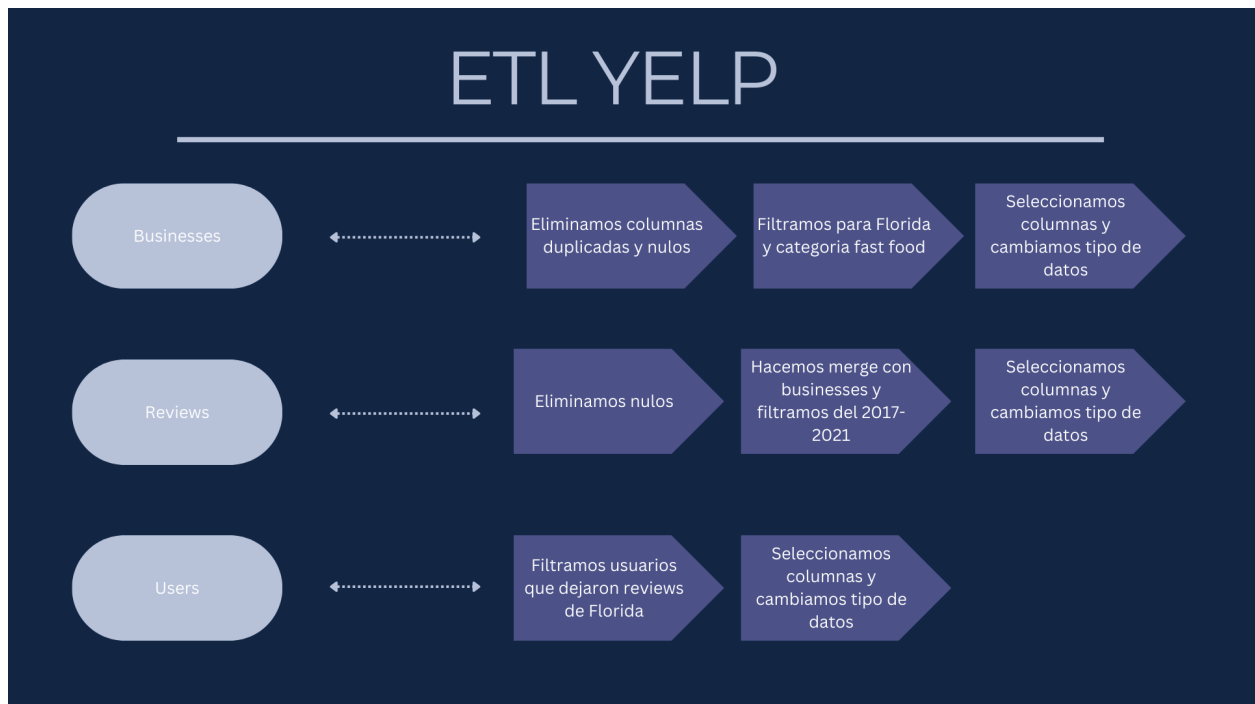
Antes de establecer un Data Warehouse y dar inicio al proceso de Analytics y Machine Learning, resulta esencial llevar a cabo una adecuada depuración y transformación de los datos. En este contexto, en primer lugar, se efectuó un proceso ETL para cada uno de los conjuntos de datos con los que se va a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

INFORME DE AVANCE: PROYECTO FINAL - SEMANA 2

	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.

INFORME DE AVANCE: PROYECTO FINAL - SEMANA 2

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.

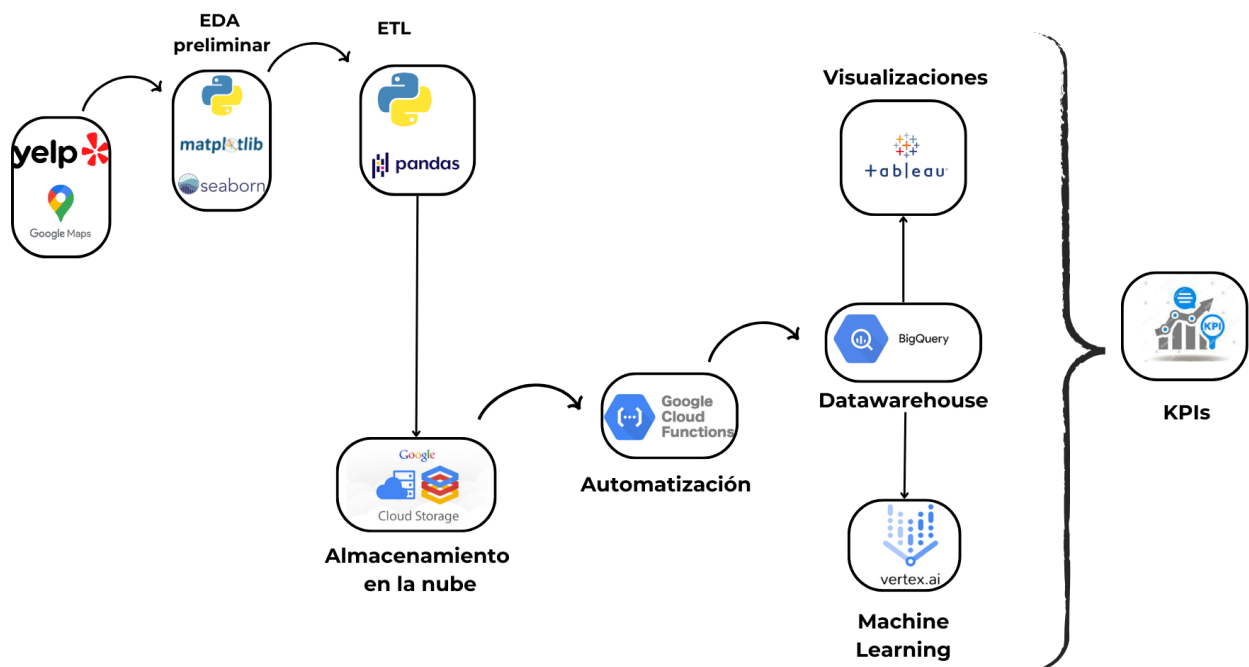
5. Diccionario de Reviews de Google Maps

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.

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.

INFORME DE AVANCE: PROYECTO FINAL - SEMANA 2

Cloud Storage

Bucket details

my_bucket_pf

Location: us-east1 (South Carolina) | Storage class: Standard | Public access: Subject to object ACLs | Protection: None

OBJECTS | CONFIGURATION | PERMISSIONS | PROTECTION | LIFECYCLE | OBSERVABILITY | INVENTORY REPORTS

Upload FILES | Upload FOLDER | Create FOLDER | Transfer DATA | Manage Holds | Download | Delete

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption
activador.txt	37 B	text/plain	Nov 1, 2023, 8:56:17 PM	Standard	Nov 1, 2023, 8:56:17 PM	Not public	—	Google-m
business_yelp_florida.csv	16.6 MB	text/csv	Oct 25, 2023, 8:43:31 AM	Standard	Oct 25, 2023, 8:43:31 AM	Not public	—	Google-m
florida_reviews.csv	33.9 MB	text/csv	Nov 1, 2023, 6:14:47 AM	Standard	Nov 1, 2023, 6:14:47 AM	Not public	—	Google-m
metadata/	—	Folder	—	—	—	—	—	—
reviews_FL/	—	Folder	—	—	—	—	—	—
reviews_yelp.csv	205.4 MB	text/csv	Nov 1, 2023, 5:57:25 PM	Standard	Nov 1, 2023, 5:57:25 PM	Not public	—	Google-m
yelp_data/	—	Folder	—	—	—	—	—	—

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

Cloud Functions

Functions

CREATE FUNCTION | REFRESH

LEARN | RELEASE NOTES

Filter functions

Environment	Name	Last deployed	Region	Recommendation	Trigger	Runtime	Memory allocated	Executed function	Actions
2nd gen	metadata	Nov 1, 2023, 7:18:45 PM	us-east1		Bucket: my_bucket_pf	Python 3.10	256 MiB	captura_evento	
2nd gen	metadata2	Nov 1, 2023, 11:06:13 PM	us-east1		Bucket: my_bucket_pf	Python 3.11	256 MiB	etl_metadata	

- 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.

INFORME DE AVANCE: PROYECTO FINAL - SEMANA 2

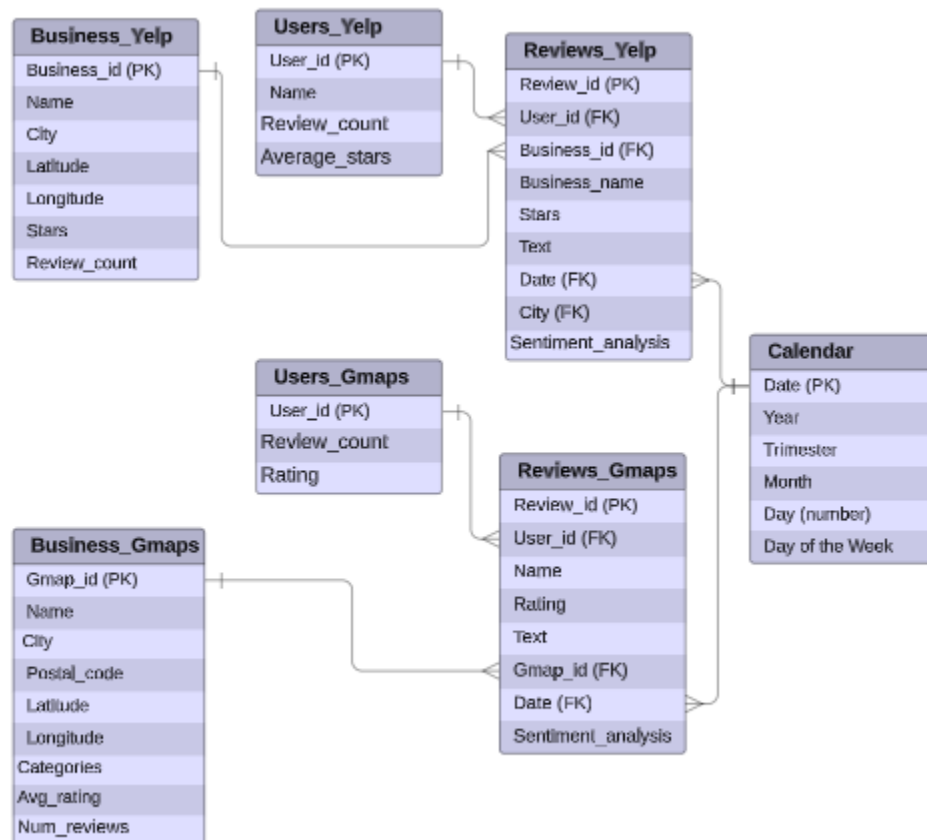
The screenshot shows the Google Cloud BigQuery Explorer interface. On the left, the 'Explorer' pane displays a project hierarchy: 'festive-freedom-402511' > 'projectofinal' > 'business_yelp'. The main pane shows the 'business_yelp' table schema with the following fields:

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
business_id	STRING	NULLABLE					
name	STRING	NULLABLE					
city	STRING	NULLABLE					
latitude	FLOAT	NULLABLE					
longitude	FLOAT	NULLABLE					
stars	FLOAT	NULLABLE					
review_count	INTEGER	NULLABLE					

Buttons for 'EDIT SCHEMA' and 'VIEW ROW ACCESS POLICIES' are visible below the schema table. The bottom of the interface shows 'PERSONAL HISTORY' and 'PROJECT HISTORY' tabs.

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: $(\text{Cantidad de reseñas positivas año actual} - \text{Cantidad de reseñas positivas año anterior}) / (\text{reseñas anterior})$

- ❖ 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: $(\text{Cantidad de reseñas negativas año anterior} - \text{Cantidad de reseñas negativas año actual}) / (\text{reseñas anterior})$

- ❖ 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 3% durante el próximo semestre.

Fórmula:

$$\text{Índice de Satisfacción del Usuario} = \frac{(\text{Ponderación_Yelp} * \text{Rating_Yelp} + \text{Ponderación_Google} * \text{Rating_Google})}{(\text{Ponderación_Yelp} + \text{Ponderación_Google})}$$

- 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 de 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: $(\text{Cantidad de reseñas año anterior} - \text{Cantidad de reseñas año actual}) / (\text{reseñas anterior})$


- ❖ 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.

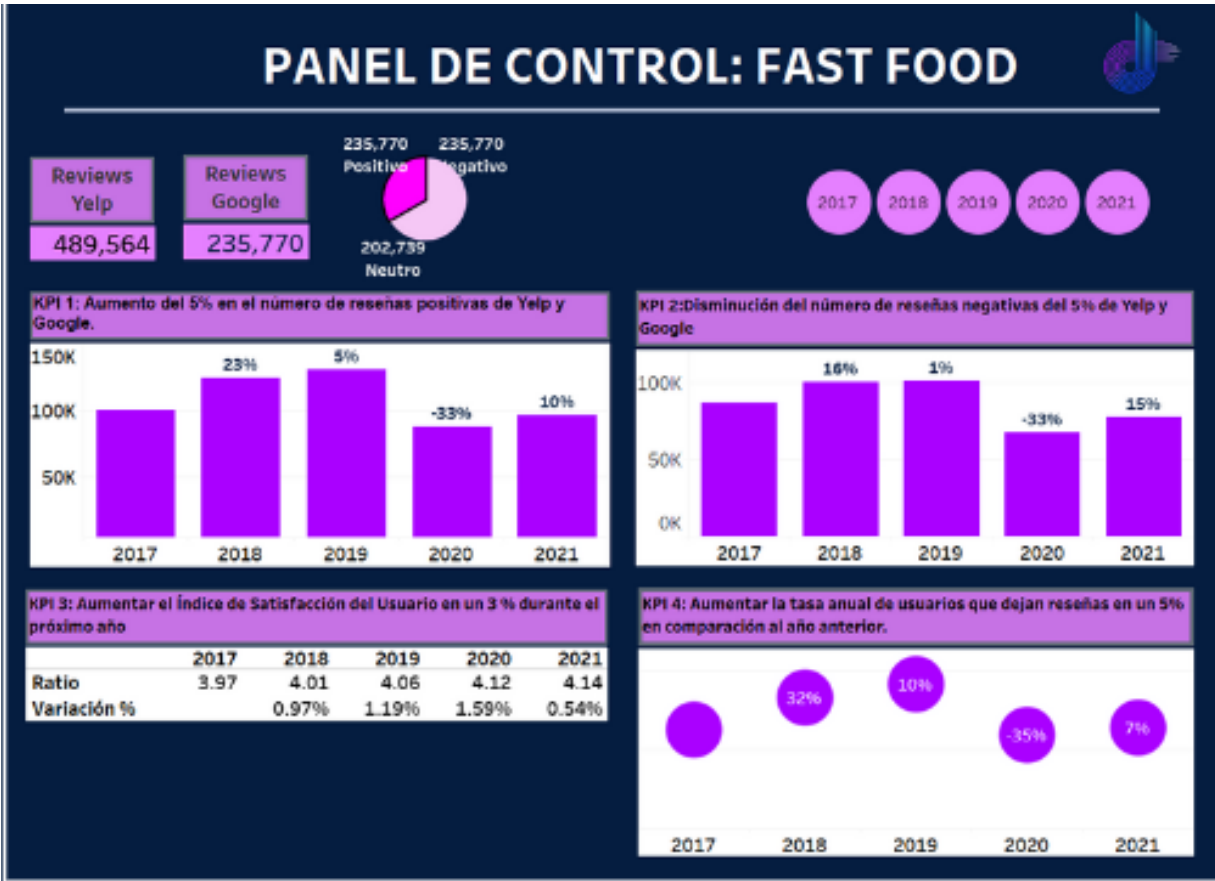
Fórmula: $(\text{Usuarios que dejan reseñas en el año actual} - \text{usuarios que dejan reseñas en el año anterior}) / \text{usuarios que dejan reseñas en el año anterior}$

INFORME DE AVANCE: PROYECTO FINAL - SEMANA 2

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	$\frac{[(\text{Cantidad de Reseñas Positivas en el Año Actual} - \text{Cantidad de Reseñas Positivas en el Año Anterior}) / \text{Cantidad de Reseñas Positivas en el Año Anterior}] \times 100}{}$	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	$\frac{[(\text{Cantidad de Reseñas negativas en el Año Anterior} - \text{Cantidad de Reseñas negativas en el Año Actual}) / \text{Cantidad de Reseñas Negativas en el Año Anterior}] \times 100}{}$	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 con un negocio, combinando las calificaciones y ponderaciones de las reseñas de los usuarios en las plataformas Yelp y Google.	 $\text{Índice de Satisfacción Usuario} = \frac{(\text{Ponderación_Yelp} * \text{Rating_Yelp} + \text{Ponderación_Google} * \text{Rating_Google})}{(\text{Ponderación_Yelp} + \text{Ponderación_Google})}$	Aumentar el Índice de Satisfacción del Usuario en un 3% 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.	$\frac{(\text{Cantidad de reseñas año anterior} - \text{Cantidad de reseñas año actual})}{(\text{reseñas anterior})}$	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.	$\frac{(\text{Usuarios que dejan reseñas en el año actual} - \text{usuarios que dejan reseñas en el año anterior})}{\text{usuarios que dejan reseñas en el año anterior}}$	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.



MVP Sistema de Recomendación

Una vez extraída, cargada y transformada la data lista para consumir por los departamentos de Analytics y Machine Learning, nuestro EDA permitirá entender bien los datos que entrenaremos en el modelo de machine learning para armar el Sistema de Recomendación de restaurantes y brindar recomendaciones a los usuarios de ambas plataformas la posibilidad de conocer nuevos sabores basados en sus experiencias previas.

- El proyecto se divide en los siguientes procesos:
- Análisis exploratorio (EDA)
- Análisis de sentimientos.
- Extracción, carga y transformación de datos.
- Modelo de aprendizaje automático y desarrollo del sistema de recomendación.

Primera parte, hemos realizado un análisis exploratorio de los datos de Yelp, que consisten en cinco tablas: negocios, reseñas, usuarios, tips y ranking.

- Extracción las categorías relevantes del archivo 'business_yelp_florida.csv', que contiene información sobre los negocios de Florida, como el nombre, la dirección, la puntuación, el número de reseñas y las características.
- Ffiltrado las 30 ciudades con más negocios y hemos creado un gráfico del top de ciudades. También hemos creado columnas categorías para cada negocio, como restaurantes, hoteles, bares, etc.
- Filtrado las categorías más relevantes para nuestro cliente, que es parte de un conglomerado de empresas de restaurantes y afines.
- Filtrado de categorías al archivo 'yelp_reviews_florida.csv', que contiene las reseñas de los usuarios sobre los negocios.

Segunda parte, hemos realizado un análisis de sentimientos de las reseñas de los usuarios, tanto de Yelp como de Google Map con el objetivo de **conocer la opinión de los usuarios sobre los negocios de nuestro interés y cruzarlos con los datos de Google Maps**, que contienen información sobre hoteles, restaurantes y otros negocios afines al turismo y ocio.

El análisis de sentimientos nos permite clasificar las reseñas en positivas, negativas o neutras, según el tono y el contenido de las mismas. Hemos utilizado una nube de palabras para visualizar las palabras más frecuentes y más relevantes en cada tipo de reseña. Así, podemos identificar los aspectos que más valoran o critican los usuarios sobre los negocios.

Tercera parte, hemos desarrollado un modelo de aprendizaje automático que utiliza los datos procesados y analizados para **predecir cuáles serán los rubros de los negocios que más crecerán o decaerán**, y dónde es conveniente reemplazar los nuevos locales de restaurantes y afines.

El modelo se basará en técnicas de regresión y clasificación, que nos permiten estimar variables numéricas o categóricas, respectivamente. El modelo tendrá en cuenta factores como la ubicación, la puntuación, el número de reseñas, el tipo de negocio, el análisis de sentimientos y la competencia.

Cuarta parte, hemos creado un sistema de recomendación de restaurantes y afines que brinda a los usuarios de ambas plataformas la posibilidad de conocer nuevos sabores basados en sus experiencias previas. El sistema se basa en la similitud del coseno, que es una medida que calcula el grado de parecido entre dos vectores. En este caso, los vectores representan a los usuarios y a los negocios, y la similitud se basa en las preferencias y el comportamiento de los usuarios.

El sistema toma como input un usuario y devuelve como output una lista de negocios recomendados, que son los que frecuentan los usuarios similares y que no son frecuentados por el usuario. Para saber si el negocio es frecuentado por el usuario, se analiza dónde ha dejado comentarios y dónde se ha hecho checkin.

INFORME DE AVANCE: PROYECTO FINAL - SEMANA 2

Estas son las principales etapas y resultados de nuestro proyecto de data science, que esperamos que les haya gustado y que les haya aportado valor.

- Queremos sumar para la presentación final:
- Visualización de nube de palabras de la data de gmaps.
- Conexión de una data externa, ya sea API o data sintética para mejorar los modelos predictivos.
- Procesamiento de los archivos de checkin y tips de Yelp, para mejorar el sistema de recomendación.
- Implementación del sistema de recomendación en la nube.
- Comparativa de rendimiento con otro producto externo.