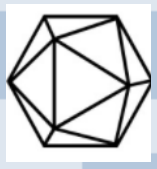


Mejora de Estrategias de Mercadotecnia

Data Engineering

DeLorean Data Consulting

Marzo 2023

Mejora de Estrategias de Mercadotecnia

Data Engineering

Abstracto

El presente documento establece la arquitectura y el modelo de los datos que usará nuestra aplicación final.

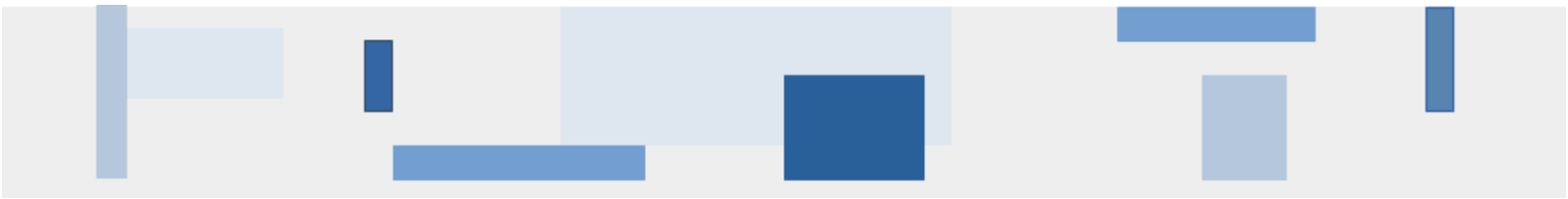
Se ha decidido por la plataforma de Google Cloud debido a los servicios especializados para inteligencia artificial, la encriptación de los datos y al rápido crecimiento de la plataforma en la cuota de mercado, lo que brinda una ventaja competitiva en el desarrollo final de la aplicación.

En primer lugar se describe la plataforma tecnológica seleccionada, así como el flujo de los datos desde el repositorio original hasta su preparación en la solución propuesta.

En segundo lugar se describen los procesos de automatización que dan soporte al ciclo de vida de los datos.

Finalmente se detalla la estructura de los datos y las relaciones entre los diferentes datos.

Tecnologías



El sector hotelero y gastronómico en Estados Unidos, es un sector en constante crecimiento y altamente competitivo. Según cifras del World Travel & Tourism Council en sus reportes de impacto económico, en 2021, la industria turística generó ingresos por valor de $771.8 mil millones de dólares, de los cuales $40.3 mil millones de dólares correspondieron a visitantes extranjeros , lo que representa un aumento del 21.7% respecto a 2020. Además, Estados Unidos ha sido durante mucho tiempo uno de los destinos turísticos más populares, lo que hace que la elección de la ubicación adecuada para un negocio sea aún más importante.

En este contexto, el uso de datos de Google Maps se ha vuelto cada vez más importante para las empresas que buscan comprender mejor la ubicación geográfica de sus clientes y la percepción que tienen de sus productos o servicios. Los datos de Google Maps proporcionan información valiosa sobre la cantidad de negocios cercanos, la cantidad de clientes potenciales y la calidad de los servicios que se ofrecen en la zona. Estos datos pueden ayudar a identificar tendencias y patrones en la demanda de servicios de hotelería y gastronomía en diferentes áreas geográficas.

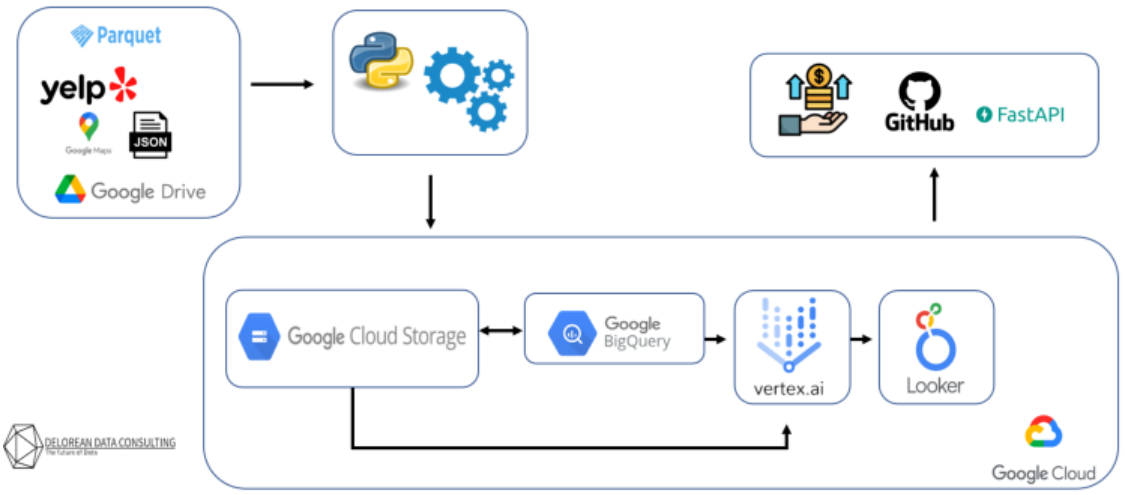
En cuanto al contexto económico de la zona, Estados Unidos ha experimentado un fuerte crecimiento en los últimos años y ha sido durante mucho tiempo uno de los principales motores de la economía. La contribución del turismo al PIB nacional para el año 2021 fue de 6.1%, según se indica en los reportes de impacto económico antes mencionados. Además, Estados Unidos posee estados que

tienen ubicaciones estratégicas que las convierten en una importante puerta de entrada a América Latina y el Caribe, siendo un destino atractivo para los inversores extranjeros. En resumen, Estados Unidos es un mercado dinámico y en constante evolución que ofrece numerosas oportunidades para las empresas del sector hotelero y gastronómico.

Flujo de Trabajo

A efecto de integrar un esquema de los datos que fuera útil para el desarrollo de la aplicación se desarrolló un repositorio (Datalake) en Google Storage para los datos crudos importados desde Drive.

Posteriormente se implementó un datawarehouse en una instancia Google BigQuery, que permite el tratamiento de datos ordenados por lo que se tuvo que aplicar procesos de limpieza a los datos crudos.



Posterior a realizar una extracción, transformación y carga adecuada a toda la data correspondiente, lo que se busca es realizar un modelo adecuado de Machine Learning, con el objetivo de llevar a cabo las predicciones buscadas para lograr cumplir los objetivos del proyecto. Google Cloud, nos ofrece diferentes herramientas para realizar modelos de ML, herramientas como: AutoML, Previsión de Vertex AI y BigQuery ML. Para la continuidad del proyecto y facilidad del mismo, se usará la herramienta de Vertex AI, la cual permite ejecutar propias rutinas de entrenamiento personalizadas e implementar modelos de cualquier tipo en una arquitectura sin servidores, además ofrece servicios adicionales, como ajuste y supervisión de hiperparámetros, para facilitar el desarrollo de un modelo.

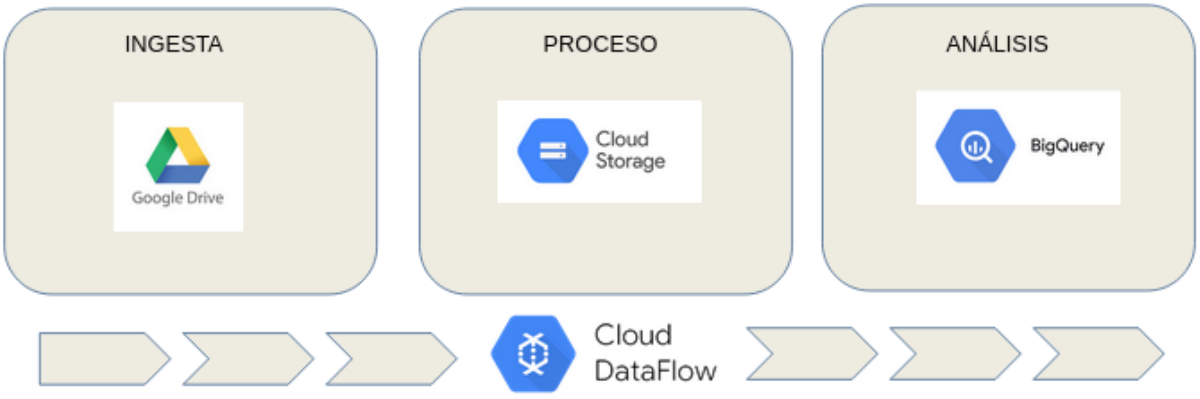
Finalmente, para la realización de presentaciones, reportes y dashboards se utilizará la herramienta Looker Studio, la cual se encuentra integrada a los demás servicios que ofrece Google Cloud, lo que permitirá tener todo el trabajo realizado en un mismo sistema integrado.

Automatización

Descripción

La ingesta de los datos no es un proceso que pueda realizarse de manera manual, primero, debido al tamaño de los mismos y segundo por la naturaleza cambiante del origen de los datos.

Por tal motivo se desarrollaron procesos automáticos (canalizaciones) para realizar éstas tareas.



Estas canalizaciones se realizan en dos pasos, la primera, recupera los datos del repositorio original, está considerado, en la etapa de producción realizar una carga incremental en este punto a efecto de mantener actualizada la información. Estos datos se almacenan en una instancia de Google Storage. que definimos como “datalake” ya que almacena los datos crudos, con apenas algún proceso de transformación.

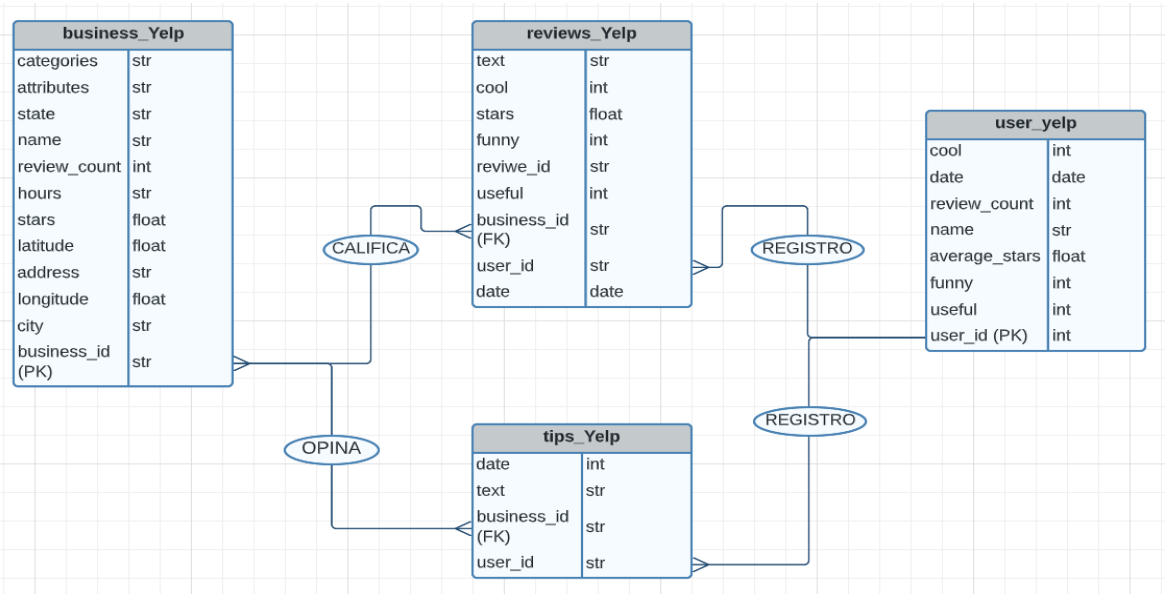
Un segundo proceso de “canalización” se realiza para integrar los datos a una segunda instancia de almacenamiento, en la herramienta google bigQuery.

Datawarehouse

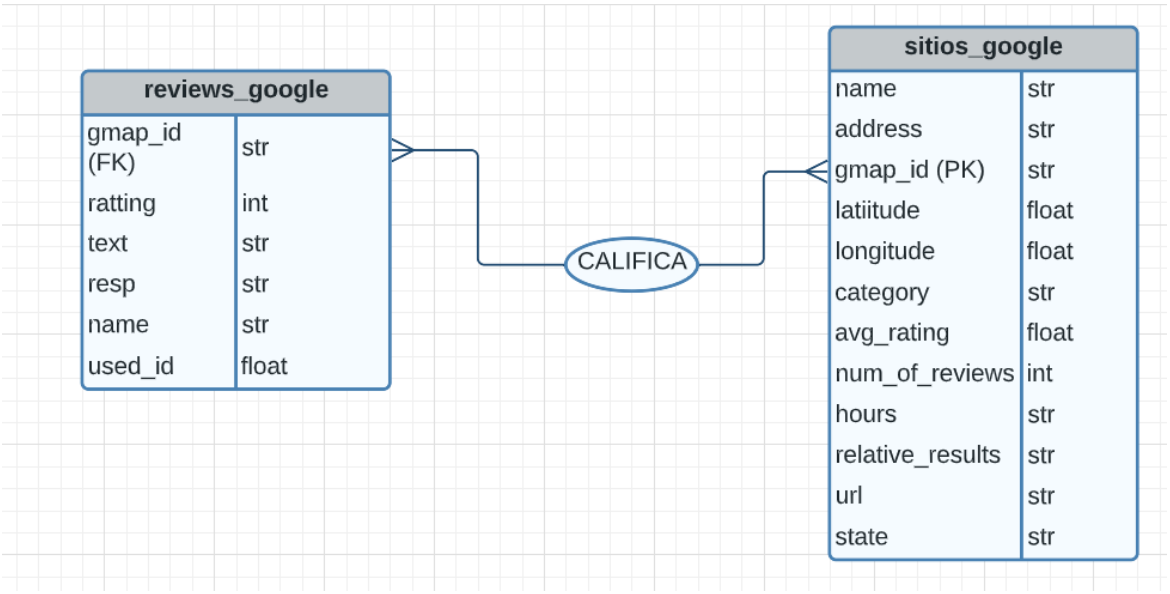
Diagrama entidad Relación

Los datos, en la instancia de DatawareHouse, se encuentran organizados de acuerdo al siguiente diagrama

Yelps



Google Reviews



Diccionario de Datos

Business (Yelp)

Num	Nombre	Tipo	Tamaño	Observaciones
1	business_id	str		Pk
2	categories	str		
3	attributes	str		
4	state	str		
5	name	str		
6	review_count	int		
7	hours	str		
8	stars	float		
9	latitude	float		
10	address	str		
11	longitude	float		
12	city	str		

Reviews_yelp

Num	Nombre	Tipo	Tamaño	Observaciones
1	text	str		
2	cool	int		
3	stars	float		

4	funny	int		
5	review_id	str		PK
6	useful	str		
7	business_id	str		FK
8	user_id	str		FK

Tips

Num	Nombre	Tipo	Tamaño	Observaciones
1	date	date		
2	cool	int		
3	review_count	int		
4	name	str		
5	funny	int		
6	useful	int		
7	user_id	int		FK
8	busibusiness_id	str		FK

Google_reviews

Num	Nombre	Tipo	Tamaño	Observaciones
1	gmap_id	str		FK
2	rating	int		

3	text	str		
4	resp	str		
5	name	str		
6	user_id	float		PK

Google_sites

Num	Nombre	Tipo	Tamaño	Observaciones
1	name	str		
2	address	str		
3	gmap_id	str		PK
4	latitude	float		
5	category	str		
6	Longitud	float		
7	category	str		
8	avg_rating	str		
9	num_of_reviews	int		
10	hours	str		
11	relativeresults	str		
12	url	str		
13	state	str		