Sprint 2: Data Engineering Report

Contenidos

Arquitectura y diagrama Entidad-Relación **2**
Diccionario de datos **3**
Data Warehouse **3**
Producto Machine Learning (MVP) - Sistema de Recomendación de Restaurantes **9**

Integrantes, roles y responsabilidades

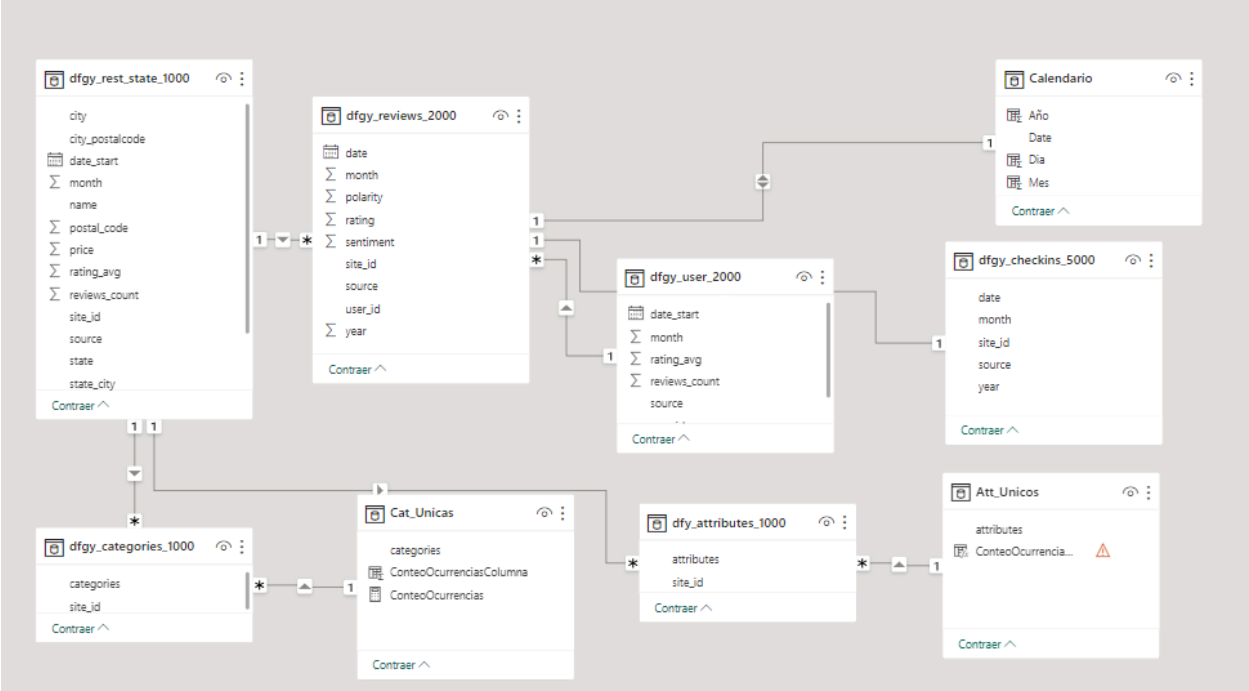
ROL	Name	Email	NameAbr	Github
Machine Learning	Diego Osorio	dosoriofc@gmail.com	DO	dosoriofc
Data Engineer	David Marimón	david.neko26@gmail.com	DM	DaAnMaGi
Data Engineer	Salomón Orozco	Salomonorozcojaramillo@gmail.com	SO	SaloLL
Data Analytics	Marcela Correal	mcorreal@gmail.com	MC	MarceCorreal
Data Analytics	Juan Garate	garatejb@gmail.com	JG	Batxa

Links asociados

Documentos internos

- Repositorio del proyecto https://github.com/Batxa/DS_ProjectFinal.git
- Follow-up de tareas: https://trello.com/b/sH9ofad9/tpfinal
- Cronograma: https://app.teamgantt.com/projects/gantt?ids=3939144

Arquitectura y diagrama Entidad-Relación

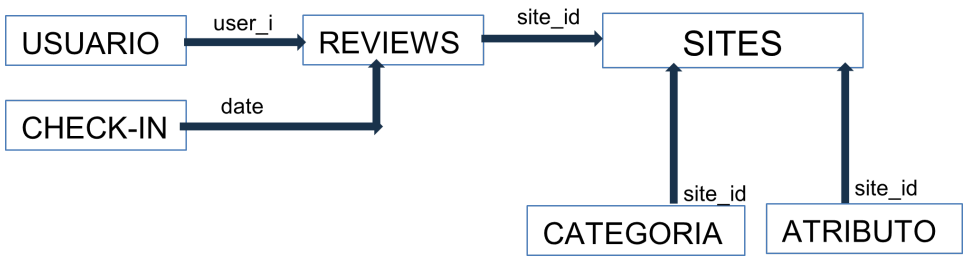


Img. 1

Datasets:

- Sites (locales de restaurantes y bares) – representa a la oferta
- Usuarios – representa a la demanda potencial
- Checkins (Visitas) - demanda real
- Reviews - feedback de los usuarios
- Categorías (Tipo de comida)
- Atributos (Servicios adicionales)

El dato crucial que se encontró, el identificador del establecimiento (site_id), que es el único dato de identificación del local inequívoco, y que relaciona la información de las reseñas de yelp con la información del restaurante en Google.

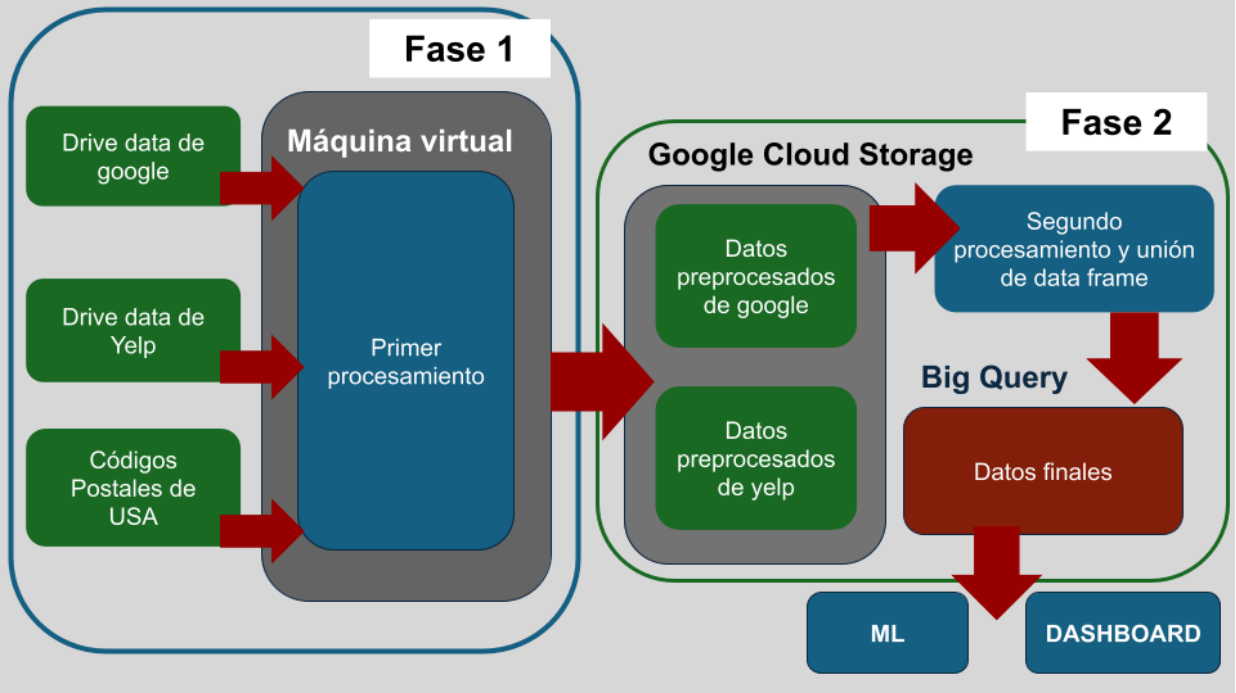


Diccionario de datos

Table name	Field name	Data type	Example	Memo
dfgy_rest	source	object	yelp	Indica la fuente de datos
	site_id	object	0x89c89cfc9b24d1c9:0xe15919fd173da3f8	Identificador de restaurante
	name	object	Pizza Hut	Nombre del restaurante
	state	object	PA	Nombres estados de USA
	city	object	Columbia	Nombres de ciudades de USA
	postal_code	object	17512	Códigos postales de USA
	price	float	2.0	Rango de precios de 1 a 4
	rating_avg	float	4.8	Promedio de rankings recibidos (de 1 a 5)
	reviews_count	integer	67	Cantidad de revisiones recibidas
	date_start	datetime	2016-11-06 16:06:48	Fecha de inicio de actividad ó primer fecha de review
	year	integer	2016	Año
	month	integer	11	Mes
	state_city	object	PA - Columbia	Estado - Ciudad
	city_postalcode	object	Columbia - 17512	Ciudad - Código postal
	state_city_postalcode	object	PA - St. Louis - 17512	Estado - Ciudad - Código postal
dfgy_user	user_id	object	V1HOblSC1bHt5pP33URiqg	Identificador unívoco de usuario
	reviews_count	integer	7	Cantidad de revisiones realizadas
	date_start	datetime	2013-08-28 0:21:08	Primer fecha de review realizado
	rating_avg	float	float	Promedio de rankings realizados (de 1 a 5)
	year	integer	2016	Año
	month	integer	11	Mes
	source	object	google	Indica la fuente de datos
dfgy_checkins	source	object	yelp	Indica la fuente de datos
	site_id	object	0x89c89cfc9b24d1c9:0xe15919fd173da3f8	Identificador de restaurante
	datetime	datetime	2013-08-19 21:26:36	Fecha del checkin
	year	integer	2016	Año
	month	integer	11	Mes
dfgy_reviews	source	object	google	Indica la fuente de datos
	site_id	object	0x89c89cfc9b24d1c9:0xe15919fd173da3f8	Identificador de restaurante
	user_id	object	V1HOblSC1bHt5pP33URiqg	Identificador unívoco de usuario
	datetime	datetime	2016-05-29 19:48:45	Fecha del review
	month	integer	5	Mes
	year	integer	2018	Año
	rating	float	4.0	Rating del review (de 1 a 5)
	polarity	float	0.8271	Valores entre -1 y +1
	sentiment	integer	-1	Valores posibles: -1, 0, +1
dfgy_categories	site_id	object	0x89c89cfc9b24d1c9:0xe15919fd173da3f8	Identificador de restaurante
	categories	object	Sandwich shop	Tipo de restaurante
	source	object	google	Indica la fuente de datos
dfgy_attributes	source	object	google	Indica la fuente de datos
	site_id	object	0x89c89cfc9b24d1c9:0xe15919fd173da3f8	Identificador de restaurante
	attributes	object	BikeParking	Servicios adicionales

Img. 2

Data Warehouse



Img. 3

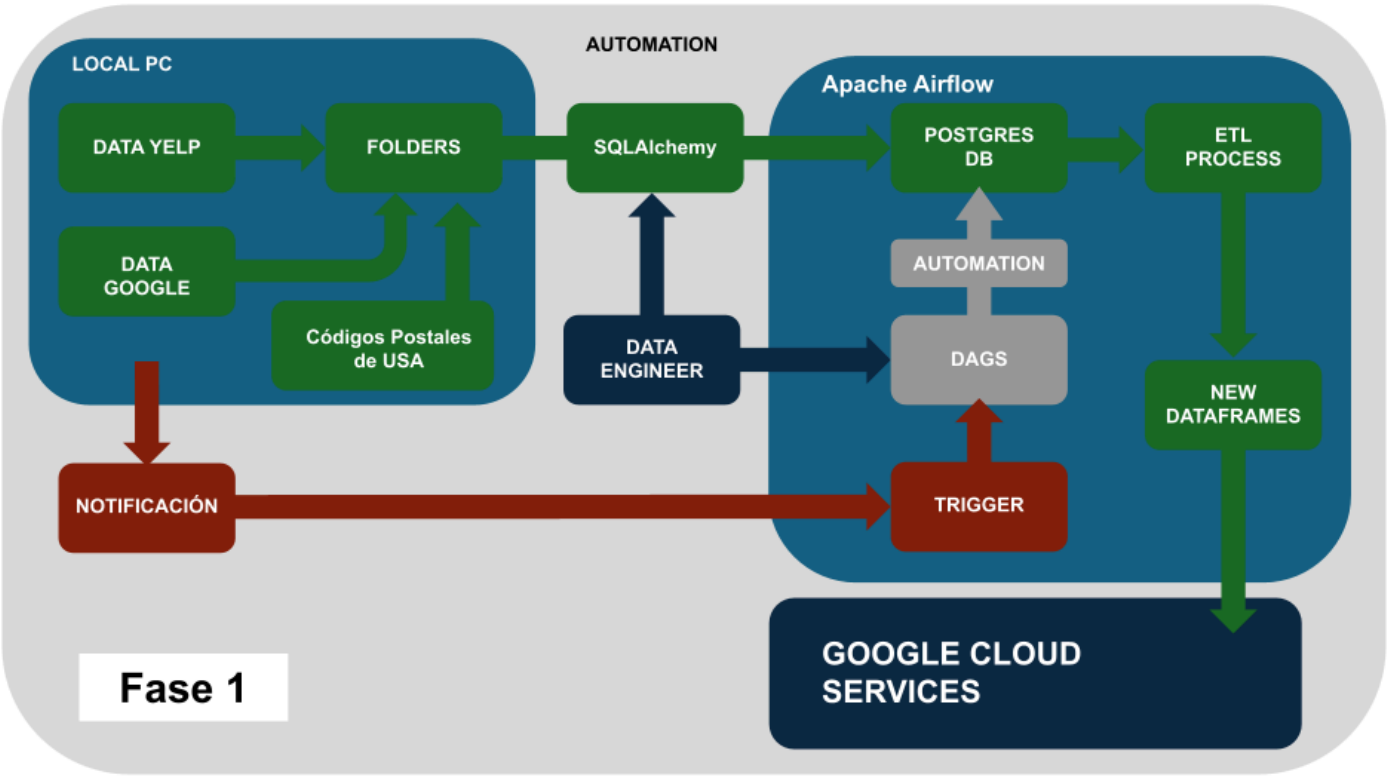
En la imagen 3, podemos observar el proceso general de la infraestructura por la que ocurre la transformación de los datos, el cual fue dividido en dos fases, esto pensando en:

1) La capacidad de procesamiento del dataset completo antes de pasar por las transformaciones.
2) La reducción de costos de infraestructura en la nube (Google Cloud).
3) La capacidad para poder realizar cambios a la transformación de los datos preprocesados, sin la necesidad de que vuelvan a pasar por todo el proceso de limpieza.

Respecto a las herramientas utilizadas principalmente para el proceso, cabe mencionar:

- **Apache AirFlow**: se ha utilizado la imagen de apache Airflow en conjunto con 2 bases de datos de PostgreSQL, para el guardado de el sistema de funcionamiento y configuración de Airflow, DAGS y configuraciones, y la segunda base de datos para el almacenamiento de los datos que resulten de nuestros procesos de ETL.
- **Google Cloud**: Plataforma de servicios en la nube proporcionada por Google. Se decidió trabajar con esta debido a la facilidad que existe para su acceso, además de la amplia documentación existente proporcionada tanto por la misma plataforma, como por la comunidad. Dentro de Google Cloud, se trabajaron los siguientes sistemas:
    - **Cloud Functions**: Herramienta que permite establecer funciones para el procesamiento de código determinadas por una serie de “Triggers” o alarmas iniciadoras. Se escoge por su facilidad de implementación y relativa sencillez.

- **Google Cloud Storage**: Herramienta de almacenamiento de archivos en la nube de Google Cloud. Permite el acceso de los archivos preprocesados desde la máquina virtual y hacia la Cloud Functions establecida para su transformación final.
- **BigQuery**: Almacén de datos en la nube. Se selecciona por la facilidad que tiene de almacenar y analizar grandes volúmenes de datos de forma rápida y económica, permitiendo el uso de consultas SQL estándar, herramientas de aprendizaje automático y la facilidad de escalabilidad de los datos.



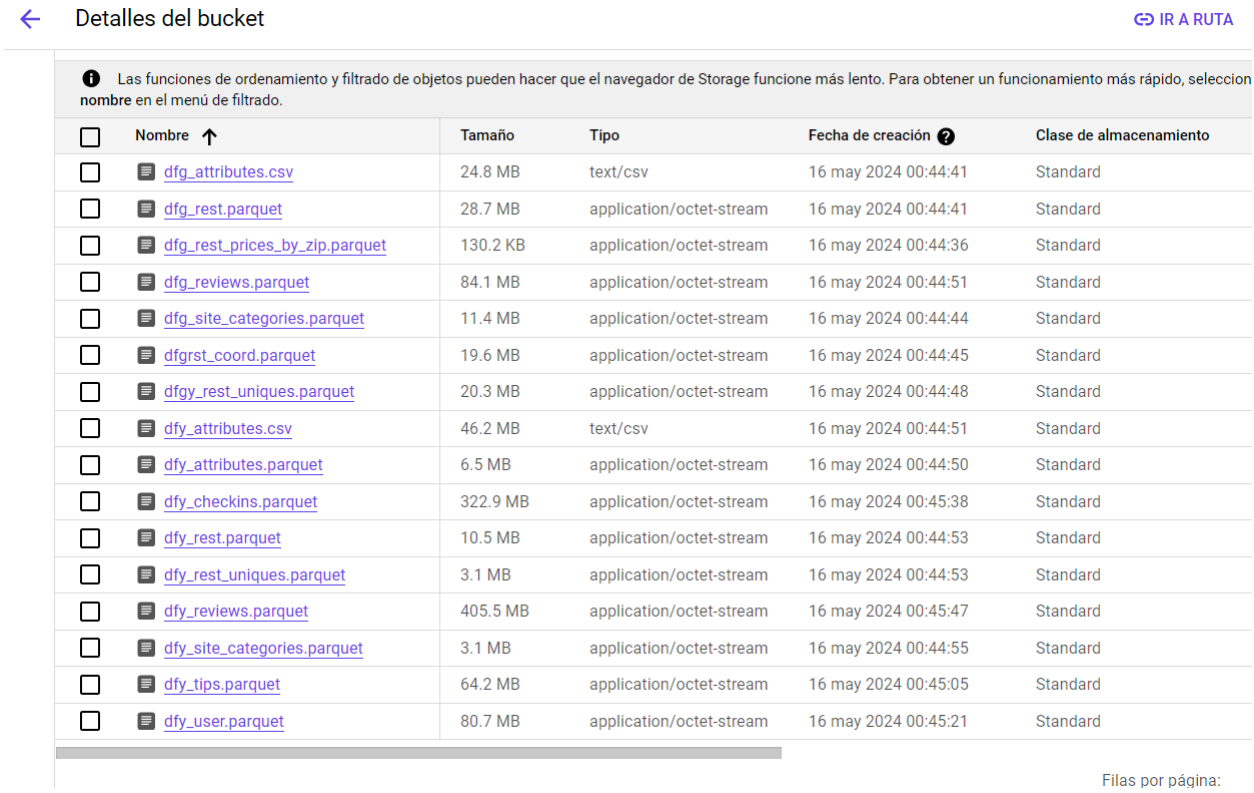
Img. 4

En la imagen 4 podemos observar la primera parte del procesamiento de los datos. Aquí, se descargan las bases de datos originales y son guardadas en un equipo de manera local, serán añadidos a una “BIND MOUNT” que compartirá los datos con nuestro contenedor de Docker, luego un DAG que se encarga de revisar diariamente esta carpeta toma los datos que tenemos, y dependiendo de la carpeta y archivo que sea hará las transformaciones necesarias para que se ajusten a nuestras tablas actuales.

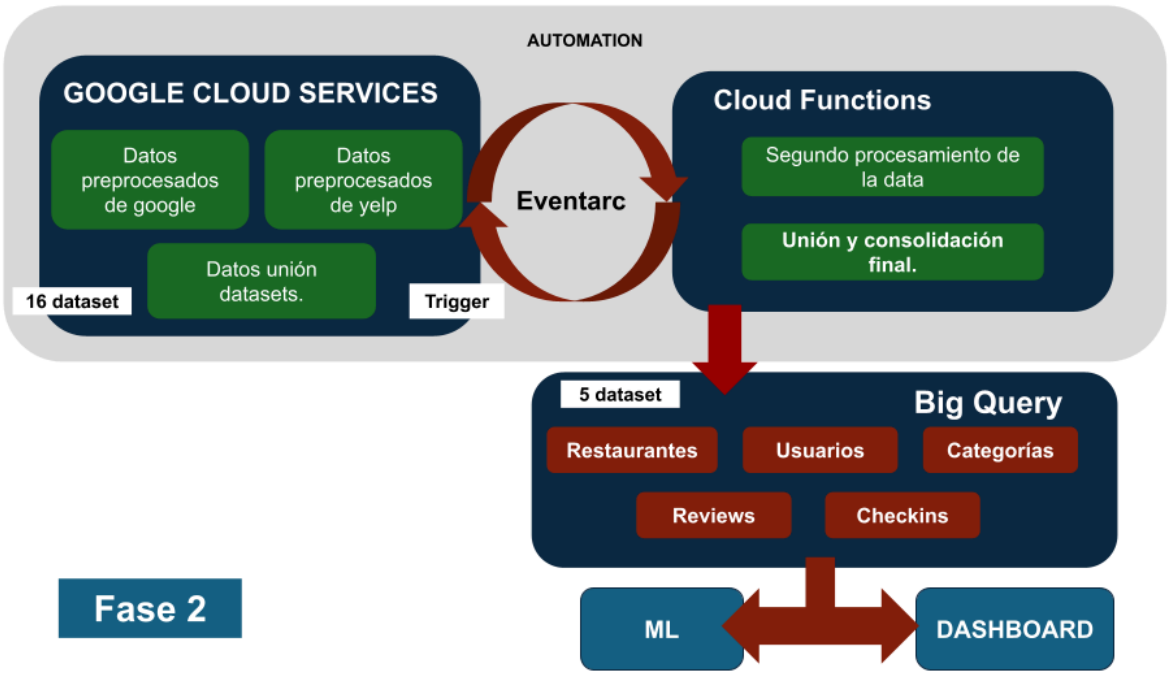
El DAG principal consiste en la limpieza general de la metadata, extrayendo la información correspondiente a locales identificados como Bares y Restaurantes, limpiando de forma general las bases de Google y de Yelp para obtener una estructura similar en los dos datasets, y realizando el cruce y validación de información con la data de los códigos postales de USA.

Después de este proceso se utiliza otro DAG en conjunto con el uso

Esto da como resultado 16 archivos, los cuales son subidos a Google Cloud Storage, como se puede observar en la imagen 5.



Img. 5



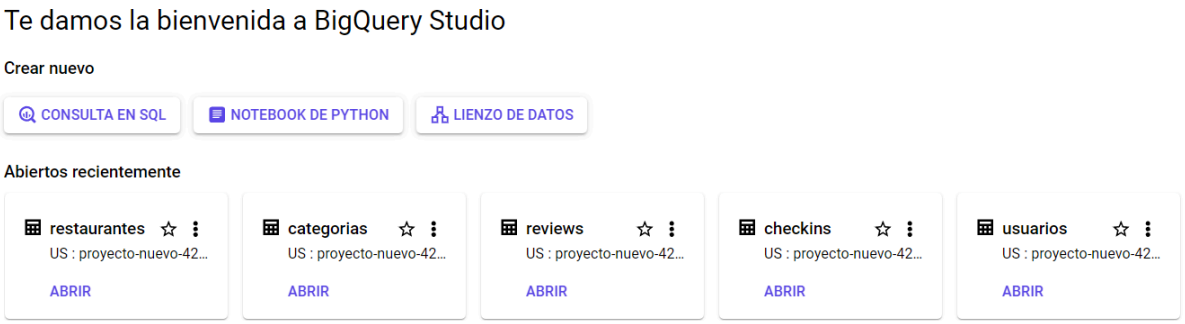
Img. 6

En la imagen 6 observamos la segunda parte del procesamiento de la data, la cual toma los 16 datasets, hace una segunda limpieza más pequeña de los datos, y combina los datos del dataset de Google y el de Yelp en 5 estructuras principales: una con la información de los restaurantes, una con la información de los usuarios, una con la separación de las categorías de los restaurantes, una con la información de las review y una con la información de los checkins.

En este caso, se hace uso de Cloud Functions, quien a través de la funcionalidad ofrecida por Eventarc para el reconocimiento de eventos, reconoce las modificaciones en los archivos subidos a Google Cloud Storage, y ejecuta automáticamente un script de Python que realizará esta segunda limpieza y la unión de los datasets.

Una vez han sido unificados los datos, estos son subidos a 5 tablas en BigQuery, una por cada una de las 5 estructuras principales (restaurantes, reviews, usuarios, checkins y categorías).

En la imagen 7 podemos observar dichas tablas en BigQuery.



Img. 7

Para el proceso mencionado, es posible actualizar la información compartida en BigQuery con nueva data simplemente descargando los nuevos archivos que serán utilizados en la carpeta local, acción que desencadenaría el inicio de todo el proceso mencionado anteriormente.

Producto Machine Learning (MVP) - Sistema de Recomendación de Restaurantes

Se desarrollò un sistema de recomendaciòn de restaurantes que se basa en la comparaciòn de las reseñas de un usuario con las de los otros usuarios registrados en el sistema y, mediante tècnicas de Machine Learning, determina cuales son los usuarios con gustos más parecidos, y en base a esta similitud le recomienda uno (o mas restaurantes) de cualquier categorìa o de sòlo una categorìa especificada por el usuario

Herramientas de desarrollo

El algoritmo de recomendación se desarrolló en Python utilizando la biblioteca open-source de ***Machine Learning Scikit-learn***, y principalmente hace uso de dos de sus funciones *TfidfVectorizer* y *Cosine_similarity*. Adicionalmente, se utilizaron las bibliotecas de Python ***NLTK****/SentimentIntensityAnalyzer* para el análisis de sentimientos de las reseñas y ***fuzzywuzzy*** para la homologaciòn de las categorías, y finalmente ***Streamlit*** para el desarrollo de la la interfaz web interactiva.

1. ***Scikit-learn***
    a. ***TfidfVectorizer*****:** se utilizó para el procesamiento de lenguaje natural (NLP) para transformar el texto de las reseñas en vectores nùmericos que fueron utilizadas en el algoritmo de similitud.
    b. ***Cosine_similarity*****:** se utilizó para calcular la similitud (mediante el algoritmo de la similitud del coseno) entre todos los vectores numéricos que representan las reseñas de los usuarios.
2. ***NLTK (Natural Language Toolkit)***
    a. ***SentimentIntensityAnalyzer:*** se utilizó para evaluar el tono emocional de las reseñas y obtener una puntuación de sentimiento que refleja la positividad, negatividad, neutralidad del sentimiento expresado en las reseñas.
3. ***Fuzzywuzzy:*** se utilizó para comparar las categorías de ambos set de datos, Google y Yelp, y obtener una puntuación de similitud que va del 0% al 100%, donde una puntuación del 100% indica que las cadenas son idénticas; esto se hizo para generar un listado reducido y estandarizado de categorías
4. ***Streamlit*****:** se utilizó para crear interfaz web interactiva que permite el ingreso de los datos y las selecciones de los usuarios y mostrar el resultado del sistema de recomendaciòn

Datos de Entrada:

El sistema permite al usuario ingresar y seleccionar los siguientes paramètros para pedir la recomendación:

1. Identificador ùnico del usuario en la base de datos
2. El número de recomendaciones que desea - Disponible: de 1 al 10
3. El estado donde desea la recomendaciòn - Disponible: todos los estados de Estados Unidos

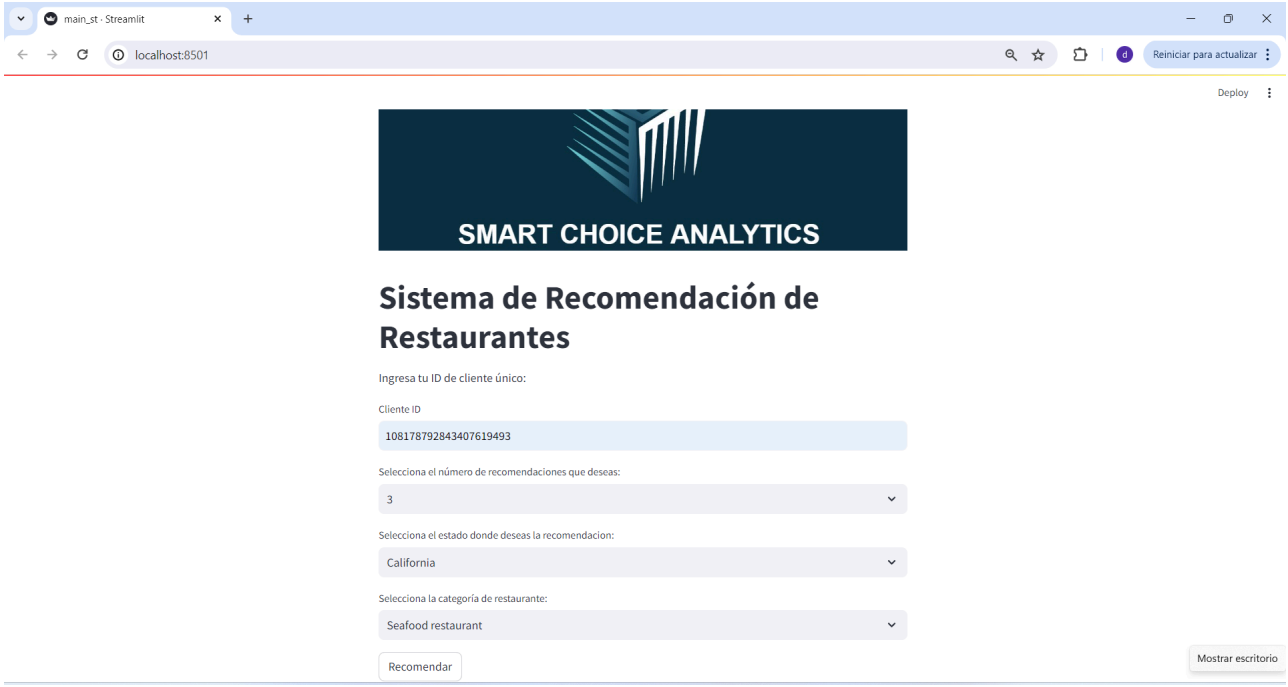
4. La categoría de restaurantes en la que desea la recomendaciòn - Disponible: todas las categorías y la opción All (recomienda sin discriminar la categorìa)

Datos de Salida:

1. Nombre(s) de restaurantes recomendados y para cada uno muestra:
    a. La categoría del restaurante
    b. Una reseña - imprime a manera de muestra la reseña que obtiene el mayor puntaje positivo obtenido con un algoritmo de anàlisis de sentimiento
    c. Rating - el puntaje otorgado al restaurante por el cliente que emitiò la reseña mostrada

Caso de Uso

- Valores de entrada:
- cliente_id = '108178792843407619493'
- número de recomendaciones: 3
- estado: California
- categoria: Seafood restaurant



Img. 8

- Salida obtenida:

Restaurantes recomendados:

1. Mariscos El Picosito:
Categoria: Seafood restaurant

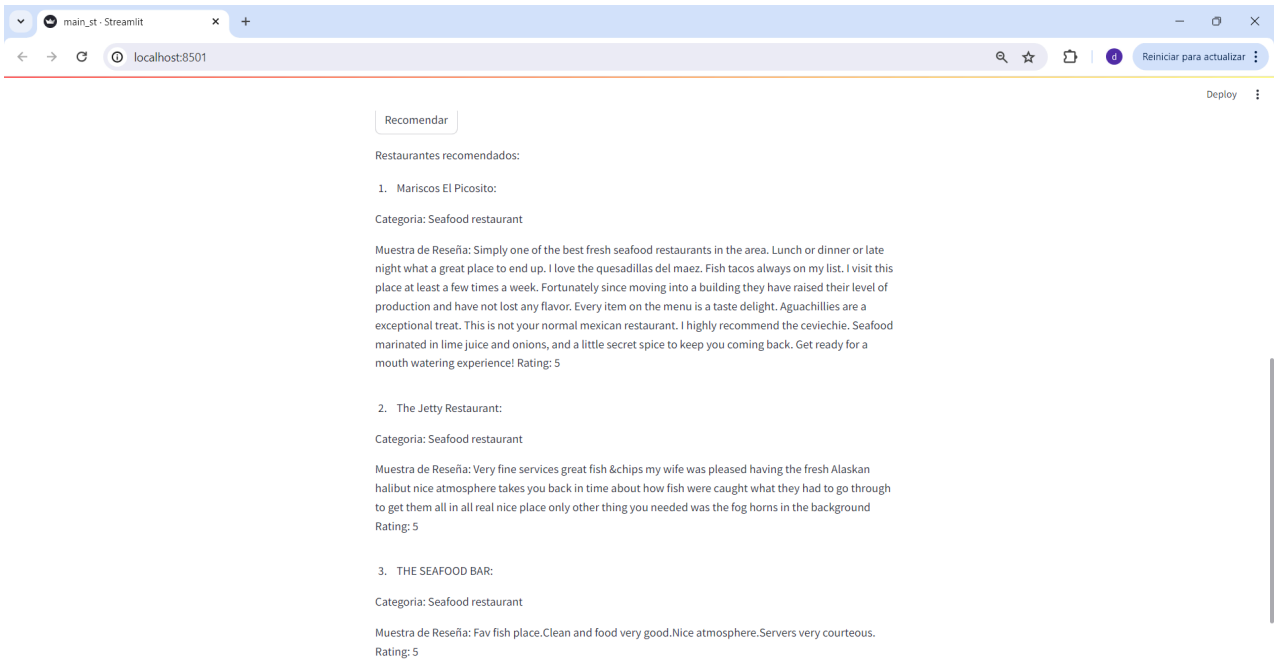
Muestra de Reseña: Simply one of the best fresh seafood restaurants in the area. Lunch or dinner or late night what a great place to end up. I love the quesadillas del maez. Fish tacos always on my list. I visit this place at least a few times a week. Fortunately since moving into a building they have raised their level of production and have not lost any flavor. Every item on the menu is a taste delight. Aguachillies are a exceptional treat. This is not your normal mexican restaurant. I highly recommend the ceviechie. Seafood marinated in lime juice and onions, and a little secret spice to keep you coming back. Get ready for a mouth watering experience! Rating: 5

2. The Jetty Restaurant:
Categoria: Seafood restaurant

Muestra de Reseña: Very fine services great fish &chips my wife was pleased having the fresh Alaskan halibut nice atmosphere takes you back in time about how fish were caught what they had to go through to get them all in all real nice place only other thing you needed was the fog horns in the background Rating: 5

3. THE SEAFOOD BAR:
Categoria: Seafood restaurant

Muestra de Reseña: Fav fish place.Clean and food very good.Nice atmosphere.Servers very courteous. Rating: 5



Img. 9