INFORME DE AVANCE - SPRINT 2

Aylén Sol Guzman, Edith Priscila Muniz Cantu, Maria Marcela Diaz, Jorge Andres Jola Hernandez, Franco Dylan Damian Luna Pedroso.

Introducción

En concordancia con los lineamientos establecidos en la propuesta inicial, el presente Informe de Avance N°2 aborda la Ingeniería de Datos del Proyecto. Este componente se erige como un pilar fundamental en el desarrollo, dado que sienta las bases para la automatización de procesos y permite la alimentación continua del data warehouse con nuevas bases de datos. La implementación de la Ingeniería de Datos resulta crucial para garantizar la eficiencia y la actualización constante de la infraestructura, contribuyendo así al éxito y la evolución continua del proyecto.

En este segundo sprint se llevaron a cabo las siguientes acciones:

1. Utilizar un servicio de Nube
2. Llevar a cabo ETL de Datasets de Google y Yelp
3. Desarrollar Proceso Carga Incremental
4. Automatizar y crear el DataWareHouse
5. Construir diccionario de datos
6. Construcción versión inicial del Dashboard
7. Inicio de Desarrollo de productos de ML

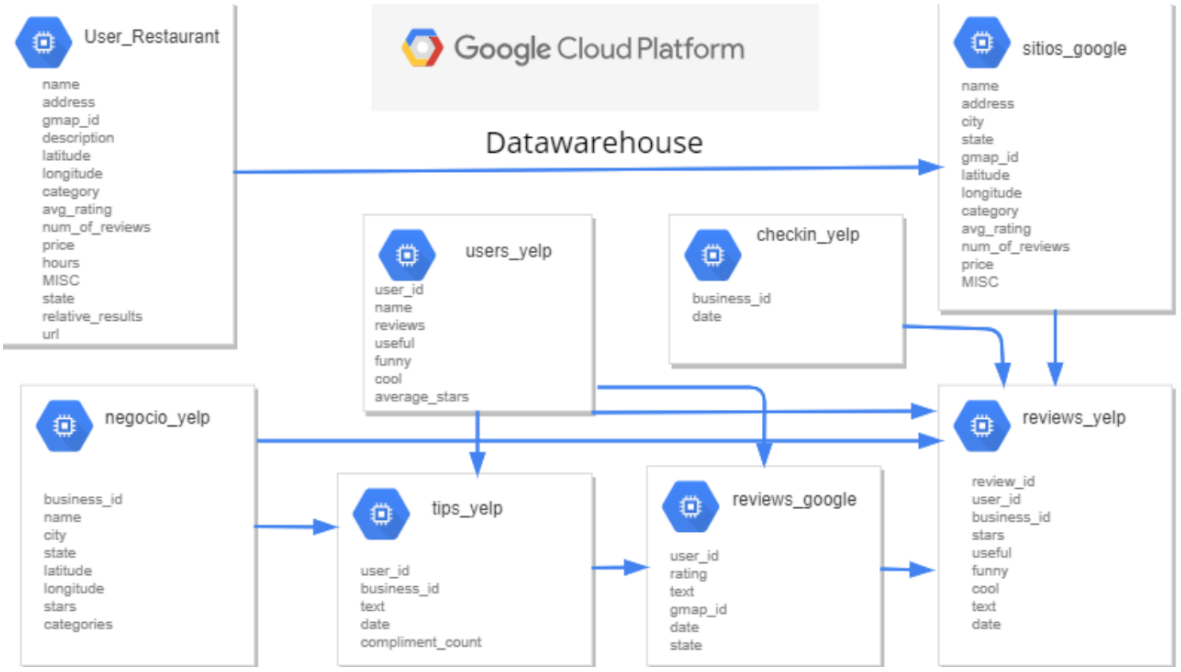
Servicio de Nube

Para la implementación de un servicio de nube, optamos por Google Cloud debido a sus notables ventajas y eficiencias. Google Cloud ofrece una infraestructura sólida y confiable, respaldada por su capacidad para escalar según las necesidades del proyecto. Su integración con diversas herramientas y servicios facilita la gestión y el despliegue eficiente de aplicaciones y recursos. Además, la seguridad avanzada y las opciones flexibles de almacenamiento y procesamiento de datos hacen de Google Cloud una elección idónea para satisfacer nuestras necesidades de forma integral y efectiva.

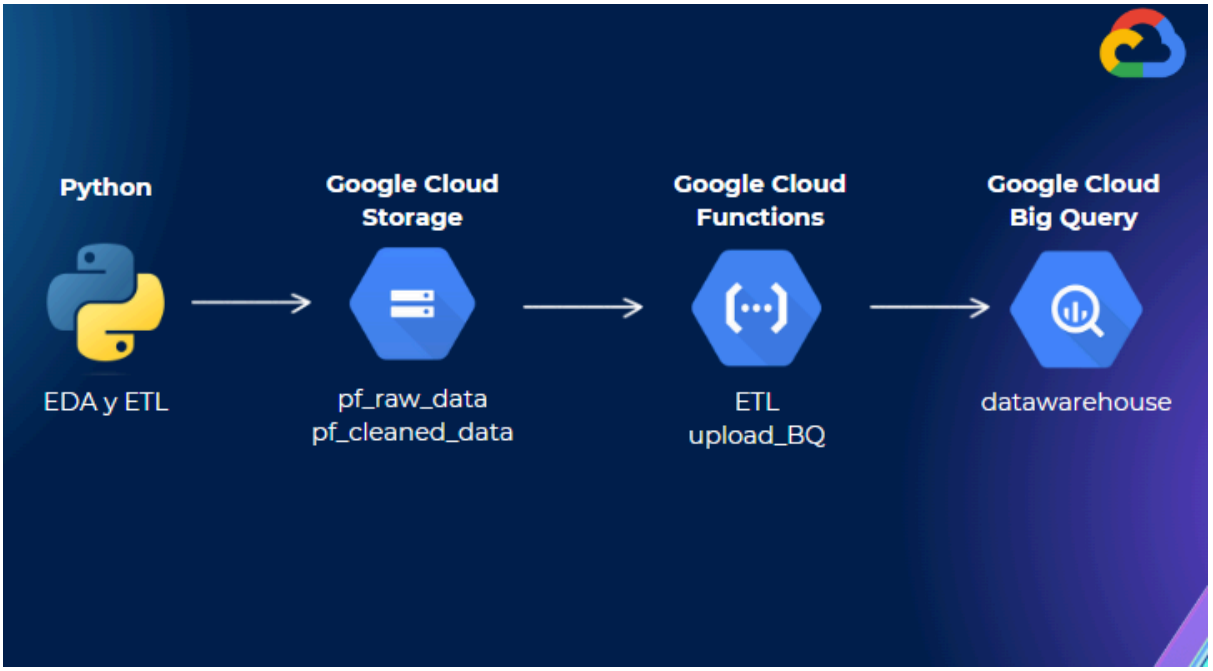
ETL

Antes de embarcarnos en la creación del DataWareHouse y adentrarnos en el análisis de datos y Machine Learning, resulta imperativo llevar a cabo una meticulosa limpieza y transformación de los datos. En este contexto, se ha realizado inicialmente un proceso de Extract, Transform, Load (ETL) para cada uno de los conjuntos de datos que serán utilizados. Este procedimiento abarcó diversas acciones, como la estandarización de datos, eliminación de columnas innecesarias, filtrado según los estados de interés, focalización en el rubro de restaurantes, tratamiento de valores nulos, entre otras medidas.

DataWareHouse



Proceso de Carga Incremental



Dada la magnitud de los conjuntos de datos de Yelp y Google, que se caracterizan por ser extraordinariamente extensos, se optó por implementar una arquitectura Big Data. En esta estrategia, utilizamos Google Cloud Storage como plataforma para almacenar los datasets, organizando la información en dos buckets específicos: "pf_raw_data", destinado a guardar los datasets originales, y "pf_cleaned_data", destinado a almacenar los datasets procesados y depurados.

Para llevar a cabo la automatización de procesos, empleamos Google Cloud functions, donde se implementaron dos funciones escritas en Python:
1. ETL: Encargada de ejecutar el proceso de Extract, Transform, Load (ETL).
2. Upload_BQ: Responsable de importar los datasets procesados al Datawarehouse.

En la creación del Datawarehouse se recurrió a Big Query, una herramienta que será aprovechada por diversas instancias como Looker y Python (Machine Learning).

Diccionarios

DICCIONARIO DE DATOS YELP

1. Diccionario de Business
- **business_id:** Identificador único para cada negocio en la base de datos.
- **name:** Nombre del negocio.
- **address:** Dirección del negocio.
- **city:** Ciudad donde se encuentra el negocio.
- **state:** Estado o provincia donde se encuentra el negocio.
- **postal_code:** Código postal del área donde se encuentra el negocio.
- **latitude:** Latitud geográfica del negocio.
- **longitude:** Longitud geográfica del negocio.
- **stars:** Calificación promedio del negocio en Yelp, en una escala de 1 a 5 estrellas.
- **review_count:** Número total de reseñas que ha recibido el negocio en Yelp.
- **categories:** Categorías o etiquetas que describen el tipo de negocio, separadas por comas.
- **Categoría:** Se refiere a la categoría objetivo encontrada en ese negocio.

2. Diccionario de Check-in
- **business_id:** Identificador único para cada negocio en Yelp.
- **num_visitas:** Representa el número de veces que se registró un check-in (visita) en el negocio.

3. Diccionario de Reviews
- **user_id (cadena):** Identificador único del usuario que ha escrito la reseña.
- **business_id (cadena):** Identificador único del negocio que recibió la reseña.
- **stars (entero):** Calificación de la reseña en términos de estrellas (1 a 5).
- **useful (entero):** Número de votos útiles que ha recibido la reseña.
- **funny (entero):** Número de votos de otros usuarios que encontraron la reseña divertida**.**
- **cool (entero):** Número de votos de otros usuarios que encontraron un factor de fascinación en la reseña.
- **text (cadena):** Contenido completo de la reseña escrita por el usuario.
- **Date (fecha):** Fecha en la que se hizo la reseña.
- **Sentimiento_score (decimal):** Valores entre -1 y 1 que representan el sentimiento hacia el texto de la reseña.

4. Diccionario de Consejos
- **user_id:** Identificador único para cada usuario en Yelp que ha dejado un consejo en un negocio.
- **business_id:** Identificador único para el negocio en el que se dejó el consejo.
- **text:** Contenido del consejo que el usuario dejó para el negocio.

- **Date (fecha):** Fecha en la que se hizo el consejo.
- **Sentimiento_score (decimal):** Valores entre -1 y 1 que representan el sentimiento hacia el texto del consejo.

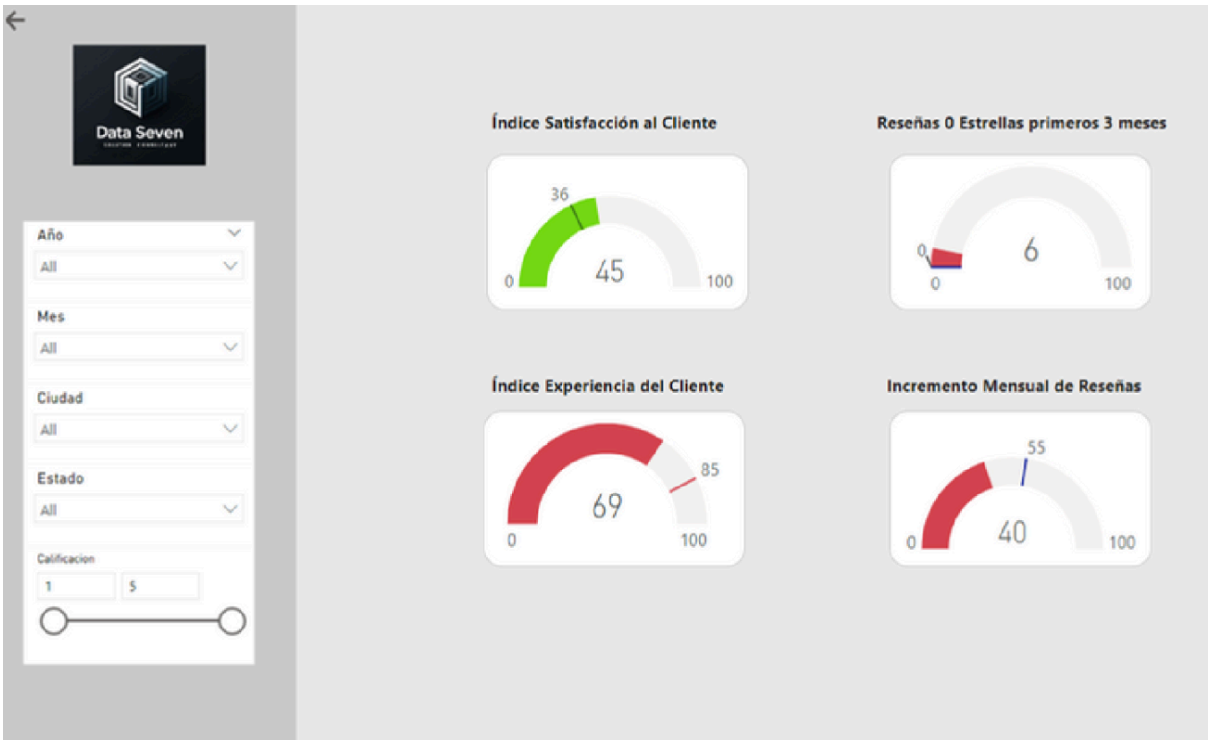
5. Diccionario de Users
- **user_id:** Identificador único para el usuario en Yelp.
- **name:** Nombre del usuario en Yelp.
- **review_count:** Cantidad total de reseñas que el usuario ha escrito en Yelp.
- **useful:** Cantidad total de votos "útiles" recibidos en las reseñas del usuario.
- **funny:** Cantidad total de votos "graciosos" recibidos en las reseñas del usuario.
- **cool:** Cantidad total de votos "frescos" recibidos en las reseñas del usuario.
- **average_stars:** Número de estrellas asignado por el usuario en sus reseñas.

DICCIONARIO DE DATOS GOOGLE

1. Diccionario de sitios_google
- **name:** Denominación del usuario, secuencia de caracteres que refleja el nombre del usuario.
- **address:** Ubicación del negocio, secuencia de caracteres.
- **gmap_id:** Código global de ubicación de Google, secuencia de caracteres.
- **latitude:** Ángulo de coordenada geográfica, número decimal.
- **longitude:** Ángulo de coordenada geográfica, número decimal.
- **categories:** Clasificación del establecimiento, secuencia de caracteres.
- **avg_rating:** Calificación promedio del establecimiento, número decimal.
- **num_of_reviews:** Cantidad de reseñas del establecimiento, número entero.
- **category:** Categoría clasificada como Hotel, Bar y Nightlife a la que pertenece el establecimiento, secuencia de caracteres.

2. Diccionario de review_google
- **user_id:** Código de identificación del usuario, número decimal.
- **name:** Nombre del usuario, secuencia de caracteres.
- **rating:** Calificación del usuario al establecimiento, número entero.
- **text:** Comentario acerca del establecimiento, secuencia de caracteres.
- **gmap_id:** Código global de ubicación de Google, secuencia de caracteres.
- **date:** Fecha y hora de la reseña, fecha.
- **state:** Sigla del estado donde se ubica el establecimiento, secuencia de caracteres.
- **sentiment_score:** Define valores entre -1 y 1 que representan el sentimiento hacia la columna text, número decimal.

Versión inicial del Dashboard


Data Seven
Año
All
Mes
All
Ciudad
All
Estado
All
Calificacion
1
5
Índice Satisfacción al Cliente
36
45
0
100
Reseñas 0 Estrellas primeros 3 meses
6
0
100
Índice Experiencia del Cliente
85
69
0
100
Incremento Mensual de Reseñas
55
40
0
100


Data Seven
Año
All
Mes
All
Ciudad
All
Estado
All
Calificacion
1
5
Distribución de Calificaciones
Sum of Calificacion
400
200
0
1
2
3
4
5
Calificacion
Calificación Promedio
3.04
Calificaciones por sentimiento
298 (19....)
318
(20.93%)
903 (59....)
Pos
Neu
Neg
Qty Calificaciones por Plataforma
243
(48.6%)
257
(51.4%)
Plataforma
Yelp
Google M...
Calificaciones por Año
Count of Calificacion
Average of Calificacion
Qty Reviews
40
20
0
3.2
3.0
2.8
Calificación Promedio
2005
2010
2015
2020
Año

Producto de Machine Learning

El producto final es un modelo de recomendación desplegado en Google Cloud Vertex AI, diseñado para proporcionar recomendaciones de restaurantes basadas en el análisis de sentimientos y similitud de contenido de las reseñas. Este modelo utiliza técnicas avanzadas de procesamiento de lenguaje natural, como el análisis de sentimientos mediante NLTK, tokenización y lematización, así como la vectorización TF-IDF para representar las características de las reseñas. Además, se emplea el algoritmo de vecinos más cercanos (KNN) con métrica de similitud de coseno para encontrar restaurantes similares en función de las preferencias del usuario. El producto incluye una función que toma como entrada una ciudad específica y devuelve recomendaciones personalizadas. Este enfoque permite una adaptabilidad significativa, y se menciona una mejora adicional para incorporar la cantidad mínima de estrellas como entrada para ajustar aún más las sugerencias según las preferencias

individuales del usuario. El modelo se almacena en Google Cloud Storage y se despliega en Vertex AI Prediction, ofreciendo así una solución integral y escalable para el desarrollo de sistemas de recomendación basados en contenido.

			SPRINT 1					SPRINT 2					SPRINT 3				
SPRINT	TAREAS	RESPONSABLE	L	M	X	J	V	L	M	X	J	V	L	M	X	J	V
PUESTA EN MARCHA DEL PROYECTO Y TRABAJO CON DATOS	ORGANIZACION SCRUM	Marcela															
	IDENTIFICAR ROLES DE LOS PARTICIPANTES	Todos															
	CREACION DEL REPOSITORIO DE GITHUB	Jorge Jola															
	PREVISUALIZACION DE LOS DATOS	Jorge Jola / Marcela															
	PLANTEAR OBJETIVOS	Aylen															
	PLANTEAR ALCANCE	Todos															
	DEFINICION DE KPIS	Aylen															
	DESCRIBIR PRODUCTOS ESPERADOS	Aylen															
	ENTENDIMIENTO DE LA SITUACION DEL NEGOCIO	Todos															
	IDENTIFICAR PROBLEMAS POTENCIALES Y SUS SOLUCIONES	Todos															
	EDA PRELIMINAR DE LOS DATOS	Priscila															
	ANALISIS PRELIMINAR DECALIDAD DE DATOS	Priscila															
	DOCUMENTACION DEL SPRINT 1	Priscila															
	STACK TECNOLOGICO	Priscila															
	DEMO 1																
DATA ENGINEERING	EXTRAER DATOS DE YELP Y GOOGLE MAPS.	Jorge / Marcela															
	CONFIGURAR PIPELINES PARA CARGA INCREMENTAL EN UN DATA WAREHOUSE.	Jorge / Marcela															
	DISEÑAR UN MODELO ENTIDAD-RELACIÓN DETALLADO.	Jorge / Aylen															
	DETALLAR TABLAS, RELACIONES Y TIPOS DE DATOS PARA EL MODELO ER.	Jorge															
	IMPLEMENTAR SCRIPTS PARA LA AUTOMATIZACIÓN DEL ETL Y ACTUALIZACIÓN DE DATOS.	Franco															
	VERSIÓN INICIAL DEL DASHBOARD CON DATOS DE MUESTRA.	Priscila															
	COMENZAR EL DESARROLLO DE PRODUCTOS DE ML.	Aylen															
	DOCUMENTACIÓN Y REPORTE DETALLADO.	Priscila / Marcela															
	DICCIONARIO DE DATOS.	Priscila / Jorge															
	DEMO 2																
DATA ANALYTICS + ML	DASHBOARD FINAL CON TODAS LAS VISUALIZACIONES	Priscila / Marcela															
	IMPLEMENTAR MODELOS DE ANÁLISIS DE SENTIMIENTOS Y PREDICCIÓN	Aylen															
	SELECCIONAR MODELOS FINALES Y REALIZAR FEATURE ENGINEERING SI ES NECESARIO	Aylen / Priscila															
	PREPARAR DOCUMENTACIÓN FINAL CON DETALLES DEL ANÁLISIS, PRODUCTOS Y LAS PREDICCIONES	Jorge															
	VIDEO DEMOSTRATIVO DEL PROYECTO	Franco															
	PREPARAR UNA PRESENTACIÓN FINAL DESTACANDO LOGROS Y RESULTADOS CLAVE	Todos															
	DEMO 3																

Figura x . Diagrama de Gantt para el proyecto