# Documentación Semana 2

# Introducción

Elaboramos este documento con el objetivo de proporcionar detalles adicionales sobre las actividades y tareas realizadas durante la segunda semana. En dicho período, nuestro enfoque principal fue la ingeniería de datos, que incluyó la planificación, diseño y construcción de la infraestructura del proyecto. Dentro de estas tareas, se dio prioridad a la validación y transformación de dataos junto a la determinación del flujo de trabajo. Asimismo, nos enfocamos en la creación y posterior automatización de los pipelines necesarios para llevar a cabo el proceso de extracción, transformación y carga (ETL), con el propósito de almacenar los datos en un Data Warehouse.

Con el fin de abordar el tema desde lo más general a lo más específico, a continuación, describimos los Pipelines de manera abarcativa y luego WorkFlow para ahondar más en los detalles

Pipelines

Para poder entender mejor de forma visual este proceso, los estructuramos en base las tres instancias del ETL (Extracción, Transformación y Carga)

* Extracción: captura de datos.
* Transformación: limpieza, filtrado de datos, validación y transformación.
* Carga: Almacenamiento de los datos el Data Werehouse.

Diagrama

Descripción generada automáticamente

Definidas las instancias, pasamos a explicar cada paso.

Extracción

Para esta etapa utilizamos como carga inicial los datasets disponibles de ambas plataformas, posteriormente se extraerán dichos datos a través de consultas a Yelp Opern Dataset y Google Places para hacer cargas de datos automáticas.

Dichos Datasets contienen información de las calificaciones y opiniones de los usuarios acerca de los distintos establecimientos, además de los datos geográficos de los mismos. El formato de los datos es principalmente json.

Transformación

Con los datos extraídos procedemos a darles coherencia, consistencia y finalmente los preparamos para la posterior carga. A continuación, los pasos realizados:

1. Limpieza: en este punto corregimos las inconsistencias y eliminamos los registros duplicados o irrelevantes para garantizar la calidad de los datos.
2. Filtrado: En pos de nuestro objetivo el cual se centrar en bares realizamos el filtrado por categoría de establecimiento para obtener solo con los pertinentes.
3. Validación: chequemos y corregimos la estructura y consistencia de los datos para obtener resultados confiables y precisos.
4. Transformación: antes de proceder a la carga les otorgamos a los archivos previamente en formato Json o pkl el formato Parquet para reducir el espacio necesario y optimizar el almacenamiento.

En esta etapa también se realizó el análisis exploratorio de datos en el cual profundizamos en el Workflow

Carga

Con los archivos ya limpios y en formato Parquet procedemos al almacenado de los mismos en GCP (Google Cloud Plataform) La carga se realiza mediante de la siguiente manera:

1. Creamos un proyecto en GCP
2. Seleccionamos un servicio de almacenamiento que en este caso es Cloud Storage (CS)
3. Creamos un bucket en CS, que va a ser el contenedor de los archivos donde almacenarás los datos
4. Finalmente, ingestamos los datos en el servicio de análisis de dicha plataforma BigQuery

Workflow

En esta sección detallamos las distintas etapas y tecnologías utilizadas en la ingeniería de datos. Nuestro objetivo principal se centró en extraer información relevante de los establecimientos de tipo Bar en Estados Unidos junto con sus calificaciones y reseñas. Aplicamos tanto limpieza como transformación a los datos obtenidos, realizamos un análisis exploratorio y utilizamos los servicios en la nube de Google Cloud Platform (GCP) para almacenar, procesar y automatizar el flujo de datos.

En la primera etapa, se emplearon los datasets provenientes de la API de Google Places y la API de Yelp para obtener datos de ubicaciones y reseñas de lugares. La API de Google Places junto con Google Maps nos proporcionan metadatos de los lugares, mientras que la API de Yelp nos brinda información adicional de la reseñas estructurada en distintos archivos.

Posteriormente, se llevó a cabo la limpieza, filtrado, validación y transformación de los datos utilizando Python y bibliotecas como pandas, json, re y os, seaborn, matplotlib, pyrrow y pyarrow.parquet. Se aplicaron filtros por categoría y se eliminaron datos irrelevantes. Los datos limpios se guardaron en formato Parquet, optimizando el espacio de almacenamiento y facilitando la eficiencia en el procesamiento.

A continuación, se realizó un análisis exploratorio detallado de los datos utilizando Python y sus bibliotecas. Se examinaron los puntajes de los lugares en Google Maps y Yelp, identificando patrones y distribuciones significativas. Esto proporcionó una comprensión más profunda de la calidad y popularidad de los lugares analizados.

Por último, se utilizaron los servicios en la nube de Google Cloud Platform (GCP) para almacenar los datos limpios en Cloud Storage en formato Parquet. Además, se emplearon los servicios de BigQuery y DataFlow para crear la base de datos necesaria y su posterior automatización de tareas relacionadas con el procesamiento y análisis de los datos.

El proceso completo abarcó desde la extracción de datos hasta su almacenamiento y análisis en la nube, permitiéndonos obtener información valiosa que utilizaremos más adelante para desarrollar un modelo de machine learning. A continuación, se detallan las diferentes etapas y tecnologías utilizadas en cada una de ellas.

1. Extracción de datos:
   1. Tecnología: Google Places API, Yelp Open Dataset API.
   2. Descripción: Utilizamos para la carga inicial los dataset provenientes de la API de Google Places y la API de Yelp Opren Dataset para extraer datos de ubicaciones y reseñas de lugares específicos en Estados Unidos.
   3. Detalles adicionales:
      1. Utilizamos dichos dataset de Google para extraer datos de reseñas y metadatos de sitios/localizaciones.
      2. Los datos de Google Maps se dividieron en dos partes:
         1. Metadata de sitios: Contiene información sobre los lugares, como nombres, direcciones, categorías, etc.
         2. Reviews por estado: Se crea 51 carpetas, una por cada estado de Estados Unidos, donde se almacenaron las reseñas correspondientes a los bares en cada estado.
      3. Los datos de Yelp se extrajeron de los datasets provenientes de Yelp Open Dataset con datos similares a los de Google pero estructurados de la siguiente manera:
         1. Business.pkl: Contiene información ordenada por establecimientos.
         2. Checkin.json: Información acerca de las fechas de ingreso a los distintos establecimientos.
         3. Review.json: Contiene información de las reseñas realizadas a cada establecimiento agrupadas por el usuario que la emitió, junto con las interacciones.
         4. tip.json: Contiene información más detallada de la reseña de cada usuario.
         5. user.parque: Detalla le información y relación que guardan los usuario entre sí.
2. Limpieza y transformación de datos:
   1. Tecnología: Python, pandas, json, re, os, pyarrow, pyarrow.parquet, Fastparquet.
   2. Descripción: Utilizamos Python junto con las bibliotecas detalladas en el inciso a. para limpiar y transformar los datos extraídos.
   3. Detalles adicionales:
      1. Se filtraron los datos por la categoría "bar" para obtener únicamente los datos relevantes.
      2. Se realizó un filtrado manual adicional para eliminar categorías no relevantes como "barber shop", "eyebrow bar", etc.
      3. Los datos limpios se guardaron en formato Parquet, que ocupa menos espacio y ofrece una mayor eficiencia en el almacenamiento.
3. Análisis exploratorio de datos:
   1. Tecnología: Python, pandas, numpy, matplotlib, seaborn
   2. Descripción: Utilizamos Python junto con las bibliotecas pandas, numpy, matplotlib y seaborn para realizar un análisis exploratorio de los datos.
   3. Detalles adicionales:

Google Maps:

* + 1. Se observa que la mayoría de las puntuaciones en Google Maps para bares se encuentran entre 4 y 5 estrellas, con una inclinación notable hacia las 5 estrellas.
    2. Como referencia para el trabajo, consideraremos puntajes bajos aquellos inferiores a 3.0, donde parece haber un primer pico notable de distinción en la frecuencia. Los puntajes altos serán considerados aquellos por encima de 4.5, debido a la alta frecuencia de las 5 estrellas. Los puntajes moderados estarán en el rango de 3.0 a 4.5.
    3. Cantidad de calificaciones en Google Maps sobre bares: 55,987
    4. Total de reviews de bares: 6,110,842
    5. Puntaje promedio de los bares: 4.235676139103616
    6. Las categorías más numerosas en los datos de Google Maps son las siguientes:
       1. Restaurantes: 25,000
       2. Bares: 8,120 (menos de un tercio de la cantidad de restaurantes)
    7. Principales tipos de bares en Google Maps:
       1. Girl Bars: Los girl bars son lugares donde las mesas son atendidas exclusivamente por mujeres, haciendo del servicio la atracción principal.
       2. Cider Bars: Los cider bars son lugares donde la cidra es la principal atracción.
       3. Cocktail Bars: Los cocktail bars tienen como atracción principal los tragos en general.
       4. Poke Bars: Los poke bars son bares hawaianos.
       5. Eyebrow Bar: Los eyebrow bars son salones de belleza con especialidad en cejas y pestañas, pero no están relacionados con nuestra propuesta, por lo que se filtrarán del conjunto de trabajo.

Yelp

1. Se observa que la mayoría de las puntuaciones en Yelp para bares se encuentran entre 4 y 5 estrellas, con una inclinación notable hacia las 5 estrellas.
2. Cantidad de calificaciones en Yelp sobre bares: 55,987
3. Total de reviews de bares: 1.640.930
4. Puntaje promedio de los bares: 3.65
5. Principales tipos de bares en Yelp
   1. Nightlife Bar: establecimientos que se centran principalmente en ofrecer entretenimiento nocturno y actividades relacionadas con la vida nocturna.
   2. Sushi Bar: establecimientos especializados en la preparación y servicio de sushi
   3. Tapas Bar: establecimientos que se especializan en servir tapas, que son pequeñas porciones de platos tradicionales españoles
6. Servicios en la nube de Google Cloud Platform (GCP):
   1. Tecnología: GCP, Cloud Storage, BigQuery, Dataflow.
   2. Descripción: Utilizamos los servicios de GCP para almacenar, procesar y automatizar el flujo de datos.
   3. Detalles adicionales:
      1. Utilizamos Cloud Storage para almacenar los datos limpios en formato Parquet dentro de un Bucket .
      2. Utilizamos BigQuery para crear la base de datos y las tablas necesarias para el análisis de datos, importado los mismos desde el Bucket
      3. Posteriormente se utilizará Dataflow para la automatización de tareas relacionadas con el procesamiento y análisis de datos.

Modelo De Entidad-Relación (DER)

El modelo de entidad-relación (DER), es una forma de modelar el mundo real en una base de datos de manera que sea fácil de comprender y utilizar para los usuarios.

En nuestro caso, al utilizar Google BigQuery como sistema de gestión de bases de datos, las relaciones del modelo ER tradicional no son necesarias ya que es una base de datos relacional distribuida, no utiliza claves foráneas para establecer relaciones de manera explícita como lo haría un sistema de gestión de bases de datos relacional tradicional.

Si bien las relaciones en un diagrama ER tradicional no se aplican directamente en BigQuery, sigue siendo importante comprender la estructura y la semántica de los datos para poder diseñar consultas eficientes y obtener los resultados deseados. Entender cómo se relacionan las tablas y qué columnas pueden utilizarse para combinar y filtrar los datos de manera efectiva en las consultas.

Por eso generamos un Diagrama de ER para entender las relaciones.

