Proyecto de Inversión Gastronómica y Afines en Mercado De Estados Unidos

INTEGRANTES:

Víctor Vargas
Guillermo del Rio
Michael Martinez Chinchilla
Julian Scarpeccio
Benjamin Zelaya

Tabla De Contenido

1. INTRODUCCIÓN

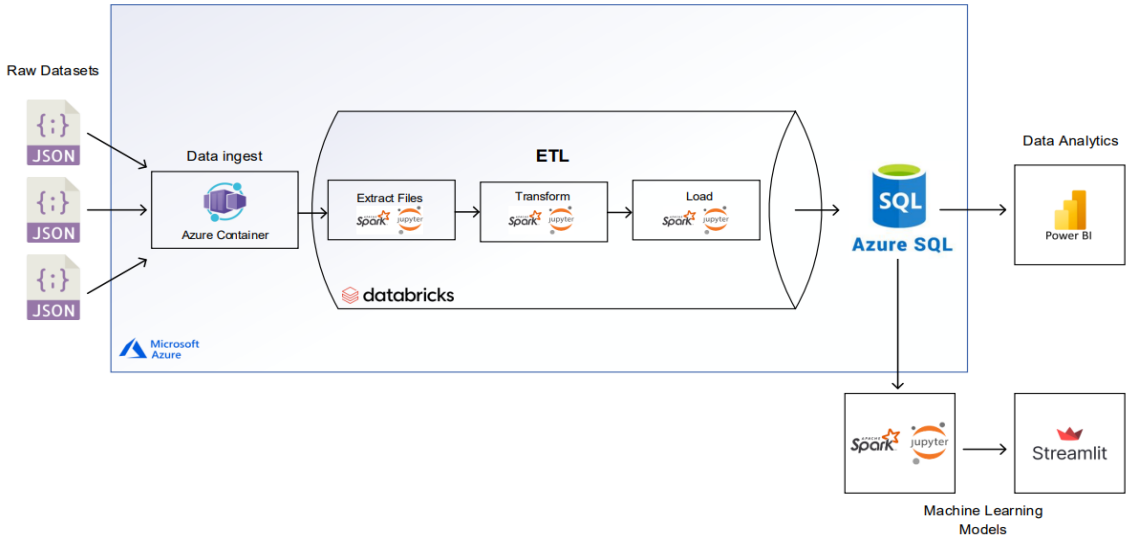
Podemos observar el trabajo realizado en el segundo sprint, el cual consiste en la creación y optimización del pipeline de datos (ETL), Asimismo, se pondrá en marcha la implementación de un datawarehouse automatizado, acompañado de la carga incremental.

El equipo de trabajo ha demostrado una sólida capacidad de organización y colaboración para abordar las demandas de nuestra propuesta de inversión para nuestro cliente.

2. OBJETIVOS

El objetivo principal del segundo sprint radica en la automatización completa del pipeline de datos. Esta automatización conlleva la transformación de los datos en formatos adecuados, seguida de su almacenamiento en un datawarehouse diseñado con precisión para facilitar análisis posteriores.

3. Arquitectura



Ingesta de Datos:
Los archivos de datos en los formatos CSV, Parquet y DBF están almacenados localmente . Estos archivos contienen información esencial sobre la movilidad urbana y son la base de nuestro análisis.

Orquestación con Apache Airflow en Docker:
Utilizamos Apache Airflow para orquestar y programar el flujo de trabajo de ingesta, procesamiento y carga de los datos. Airflow asegura que las tareas se ejecuten automáticamente y en el orden correcto, optimizando la eficiencia del sistema.

Proceso ETL Automatizado:
El proceso de Extracción, Transformación y Carga (ETL) se realiza a través del lenguaje de programación Python, además, hemos implementado con airflow el mecanismo de carga incremental para actualizar los datos de manera eficiente.

Almacenamiento en MySQL:
Los datos procesados se almacenan en una base de datos MySQL.

Análisis de Datos en Power BI:
Una vez que los datos se encuentran en la base de datos, se pueden analizar y visualizar utilizando Power BI. Esto permite identificar tendencias, patrones y obtener información valiosa para la toma de decisiones informadas.

Modelo de Machine Learning y Streamlit:
Se creará un modelo de Machine Learning en Python. Este modelo se implementará en una aplicación interactiva utilizando Streamlit, lo que permitirá a los usuarios utilizar el modelo y obtener recomendaciones en tiempo real.

3.2 KPI

Los siguientes KPIs se utilizarán para evaluar el éxito de nuestras soluciones y la mejor opción de inversión de franquicia para nuestro cliente:

KPI: Tasa de Satisfacción del Cliente.

Descripción: Este KPI mide el porcentaje de clientes satisfechos en función del rating de opiniones recopiladas en plataformas como Yelp y Google Maps. Calificación por encima de un umbral específico de puntuación de Rating = 4 .
Tasa de Satisfacción del Cliente = (Número de Rating/ Total de Rating) x 100

KPI: Cantidad de Sucursales por Conglomerado de Estados.
Descripción: Permite identificar cuantas sucursales tendremos por conglomerado de estados.

KPI: Porcentaje de Restaurantes con Alta Calificación.

Descripción: Este KPI muestra el porcentaje de restaurantes que tienen una calificación por encima de un umbral específico (por ejemplo, 4 estrellas). La fórmula es:
Porcentaje de Restaurantes con Alta Calificación = (Número de Restaurantes con Rating >= Umbral) / Total de Restaurantes x 100

KPI: Top 5 Franquicias por Conglomerado de Estados:
Identificar las mejores franquicias por conglomerado de estados.

Después de haber compartido los indicadores clave de rendimiento (KPIs) que describen la situación actual, ahora introducimos dos KPIs que serán fundamentales para evaluar el éxito de la inversión una vez que haya transcurrido un tiempo desde su ejecución:

KPI: Comparación de Satisfacción del Cliente Promedio por Estado con el Nivel de Satisfacción de la Franquicia Elegida.
Descripción: Este indicador se establece con el propósito de evaluar el nivel de satisfacción del cliente en la franquicia elegida por el inversor y compararlo con la satisfacción promedio de los clientes dentro del estado en el que se realizó la inversión.

Si el nivel Promedio de satisfacción de nuestra Franquicia está por encima del promedio del estado significará que estamos por encima de la media y que mantenemos la calidad y servicio a nuestros clientes, lo cual nos indica un buen desempeño.

KPI: Comparación del Nivel de Satisfacción del Cliente en Franquicias Seleccionadas frente al Promedio Nacional.
Descripción: Este indicador de desempeño tiene como objetivo evaluar la satisfacción de los clientes en la franquicia que se ha seleccionado por el inversor, y contrastar con la satisfacción promedio de los usuarios en dicha franquicia a nivel nacional.

4) ETL

De los 3 dataset entregados, se realizaron transformaciones como limpieza de nulos y se redujo a utilizar el 30% de los datos ya que son los prodigios para nuestros análisis y modelo de machine learning.

Los dataset dados por la empresa y extraídos se encuentran el la carpeta llamada Datasets de nuestro repositorio y los datasets ya limpios en el cual se utilizaran para nuestro análisis y modelo de m.l se encuentran dentro de sprint_2 llamada la carpeta dataset_limpios.

5. Diccionario de datos

5.1.

Este conjunto de datos contiene…

.

- Id_estado : Indica
- Nombre_Estado: Indica el
- Nombre_Local : Sonido del motor
- Latitud: Señales de alerta (bocina)
- Longitud : Total de sonidos
- Categoria : Nombre del barrio
- Promedio_rating:
- Cantidad_Reviews:
- Comentario:

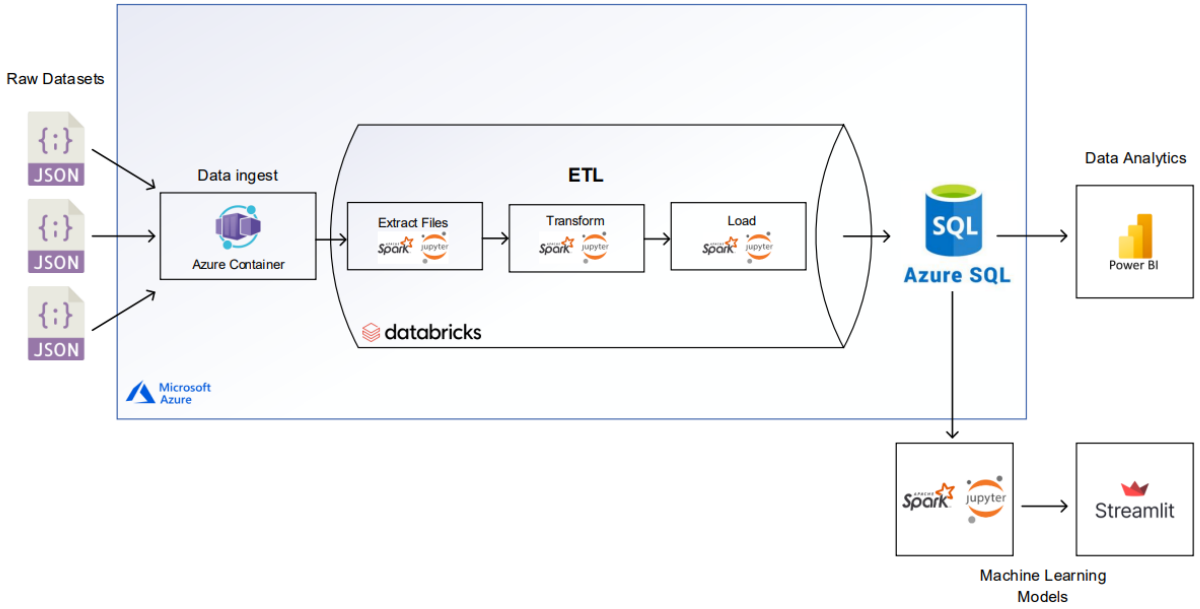
5.2.

Dataset.

6. Análisis Exploratorio

En el proceso de realizar el Análisis Exploratorio de Datos (EDA)

7. Creación del Data Warehouse (Diagrama Entidad Relación)(Aquí va el de Power Bi)



8. Tareas por realizar

8.1 Automatizar el Data Warehouse

Para lograr la automatización del data warehouse, implementaremos el uso de Airflow. En el cuaderno, configuraremos un flujo de trabajo que facilitará la ejecución de nuestro proceso ETL, asegurando así la transformación y carga eficiente de los datos depurados en MySQL.

8.2 Carga incremental del Data Warehouse

Airflow realiza una carga incremental al emplear un enfoque inteligente para identificar y procesar solo los nuevos datos o los que han cambiado desde la última ejecución. Utilizando marcas de tiempo y metadatos, Airflow compara los registros existentes con los nuevos datos, permitiendo una actualización eficiente de la base de datos sin tener que procesar nuevamente todo el conjunto de datos. Esto agiliza el proceso y optimiza el uso de recursos al reducir la carga de trabajo y el tiempo requerido para mantener la integridad y actualidad de la información en el sistema.