***Documento de Stack Elegido y Fundamentación***

***Fase 1: Recolección de Datos***

***Fuentes de Datos:*** APIs de Yelp y Google Maps: Ambas fuentes proporcionan datos geolocalizados y de reseñas. Yelp ofrece opiniones y valoraciones de usuarios, mientras que Google Maps aporta información adicional sobre ubicaciones y categorías de negocio. Estas fuentes combinadas ofrecen un conjunto robusto de datos tanto estructurados (puntos de interés, categorías) como no estructurados (comentarios de los usuarios) para un análisis detallado.

***Almacenamiento Inicial:*** AWS S3 / Google Cloud Storage: Estos servicios permiten almacenar grandes volúmenes de datos no estructurados en la nube, proporcionando alta disponibilidad y escalabilidad, lo que es esencial para manejar datos provenientes de diversas fuentes, como las APIs mencionadas.

MySQL: Se utiliza para almacenar datos estructurados de forma organizada, permitiendo consultas rápidas y eficientes, especialmente útiles cuando se requiere un acceso frecuente a los datos procesados y filtrados.

***Fase 2: Procesamiento y Limpieza***

***Herramientas:*** Pandas: Es ideal para la manipulación de datos y la limpieza (eliminación de duplicados, transformación de texto). Pandas facilita el análisis y la preparación de datos mediante operaciones vectorizadas que permiten manejar grandes conjuntos de datos de forma eficiente.

***ETL (Extract, Transform, Load):*** Python, requests, Scrapy: Estos frameworks se utilizan para realizar las tareas de extracción automatizada de datos mediante la integración de diversas APIs (Yelp y Google Maps). Scrapy puede ser utilizado para la recolección de información adicional en caso de que no toda la data sea proporcionada directamente por las APIs.

***Fase 3: Análisis y Modelado***

***Análisis Exploratorio de Datos (EDA):*** Matplotlib, Seaborn y Power BI: Estas herramientas permiten la visualización de datos, siendo fundamentales para obtener una visión general de las tendencias y patrones en los datos. Power BI puede ofrecer dashboards interactivos y gráficos dinámicos útiles para los análisis ejecutivos.

***Análisis de Sentimiento:***TextBlob, VADER y spaCy: Estas bibliotecas se utilizan para analizar el sentimiento de los comentarios de los usuarios. A través de procesamiento de lenguaje natural (NLP), se pueden identificar emociones, opiniones y clasificar las reseñas como positivas, negativas o neutras.

***Machine Learning:*** Scikit-learn: Es la biblioteca seleccionada para implementar modelos predictivos. Los algoritmos de Machine Learning permitirán predecir puntuaciones futuras de los negocios, basándose en datos históricos de reseñas.

***Sistemas de Recomendación:*** TF-IDF + KNN: Estos métodos serán utilizados para crear un sistema de recomendación basado en contenido, donde las reseñas más similares se agrupan. Si se dispone de suficientes datos de usuarios, también se explorará la creación de un sistema colaborativo para recomendaciones personalizadas.

***Fase 4: Almacenamiento y API***

***Bases de Datos:*** MySQL: Se utiliza para almacenar datos organizados y estructurados, optimizados para realizar consultas rápidas.

***Desarrollo de API:***FastAPI: Se emplea para exponer los modelos de análisis y recomendación como endpoints REST, permitiendo la integración de los sistemas en otras aplicaciones y servicios. FastAPI es eficiente y rápido, ideal para la creación de APIs modernas.

Despliegue: Render o AWS Lambda: Estas plataformas se utilizarán para el despliegue de los modelos y la API, permitiendo un acceso eficiente y escalable.

***Fase 5: Orquestación (ETL Automatizado)***

Herramientas: Apache Airflow: Se utilizará para la orquestación de las tareas de ETL, permitiendo la automatización y escalabilidad de los flujos de trabajo. Airflow facilita la programación y ejecución de tareas de forma secuencial y dependiente, mejorando la eficiencia.

Cron + Python Scripts: Alternativa simple para tareas menos complejas o cuando no se necesita una orquestación avanzada.

***Fase 6: Visualización y Presentación***

Herramientas: Power BI, Tableau, Looker Studio: Estas herramientas permitirán la creación de dashboards interactivos y detallados que presenten la evolución del sentimiento por zona, ranking de restaurantes, oportunidades y otros KPIs clave.

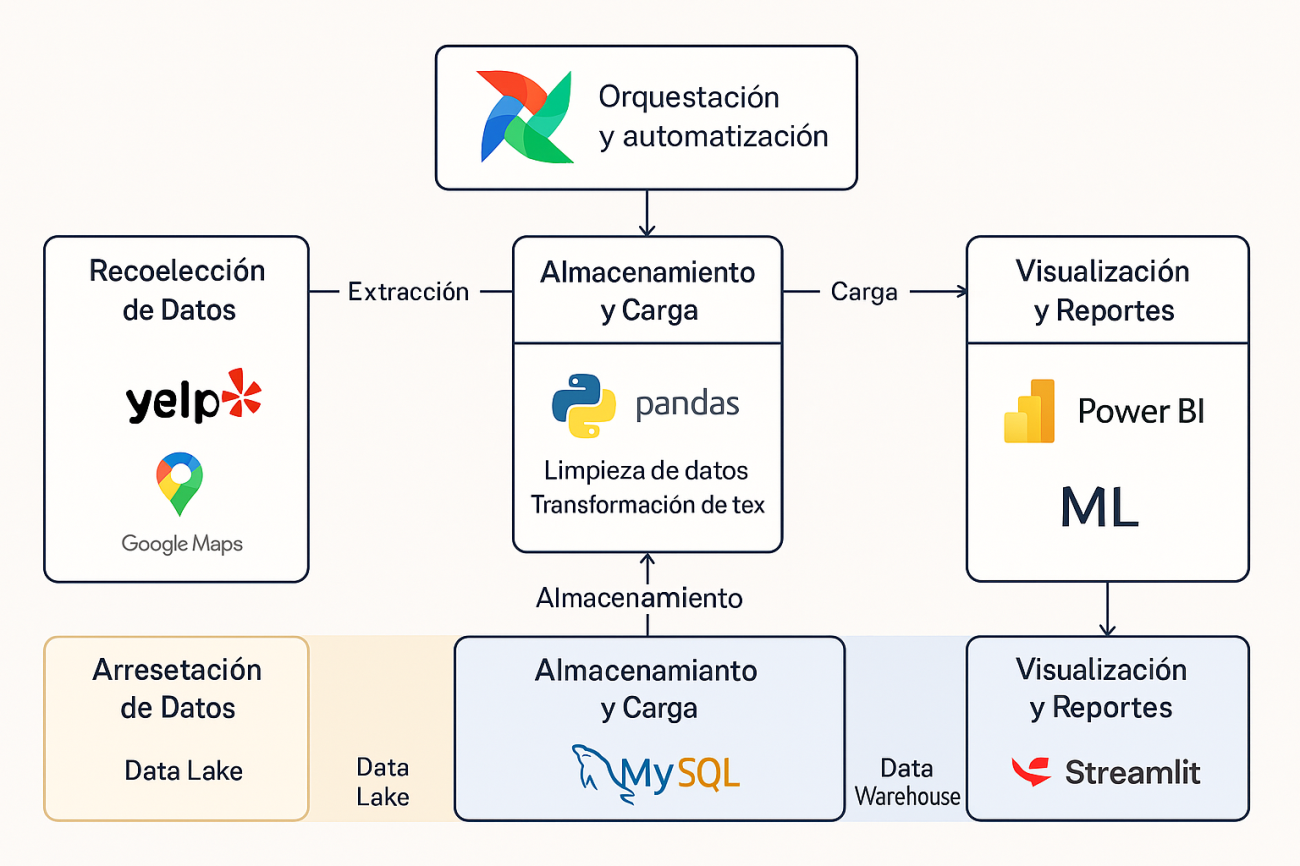
App Interactiva: Streamlit: Se desarrollará una app interactiva que permitirá explorar los datos de forma dinámica, facilitando la visualización de las tendencias y predicciones en tiempo real.

API de Recomendaciones :FastAPI: Se utilizará para exponer el sistema de recomendaciones como un servicio web, permitiendo que los usuarios reciban sugerencias personalizadas basadas en su comportamiento.

***Fase 7: Control de Versiones y Gestión***

Herramientas:GitHub: Se utilizará para gestionar el código fuente, permitiendo un control de versiones eficiente y la colaboración entre los miembros del equipo.

ClickUp: Será utilizado para organizar las tareas y el cronograma del proyecto, asegurando que el equipo mantenga un flujo de trabajo organizado y eficiente.



***Flujo de Trabajo y Ciclo de los Datos***

El ciclo de trabajo y procesamiento de los datos sigue una secuencia lógica que abarca desde la recolección hasta la presentación final de los resultados:

1***. Recolección de Dato***s (Raw Data - Data Lake)

* Ubicación: Data Lake (AWS S3, Google Cloud Storage, CSV temporales).
* Estado: Datos crudos y sin procesar.

***Acciones:***

* Extracción de datos desde las APIs de Yelp y Google Maps.
* Scraping con Scrapy o Selenium si es necesario.
* Almacenamiento en formato JSON o CSV en el Data Lake.

***2. Procesamiento y Limpieza (***Data Transformation - Data Lake / Staging Area)

* Ubicación: Zona de procesamiento o Staging Area.
* Estado: Datos procesados y estructurados.

***Acciones:***

* Eliminación de duplicados, valores nulos, y corrección de formatos.
* Transformación de texto (tokenización, lematización, eliminación de stop words).
* Análisis exploratorio inicial (EDA) con Pandas, Matplotlib, y Seaborn.

***3. Almacenamiento y Carga a Data Warehouse***

* Ubicación: Data Warehouse (MySQL, PostgreSQL, BigQuery).
* Estado: Datos organizados y optimizados para análisis.

***Acciones:***

* Creación de tablas para almacenar datos limpios, análisis de sentimiento y metadatos.
* Optimización de consultas para rapidez y eficiencia.

***4. Análisis y Modelado*** (Machine Learning y Recomendaciones)

* Ubicación: Data Warehouse + procesamiento en memoria (Python, Scikit-learn).
* Estado: Datos listos para modelos predictivos y análisis.

***Acciones:***

* Aplicación de NLP para análisis de sentimiento.
* Implementación de sistemas de recomendación y modelos predictivos.
* Almacenamiento de resultados en el Data Warehouse.

***5. Orquestación y Automatización (ETL Automatizado)***

* Ubicación: Flujos ETL manejados con Apache Airflow o Cron Jobs.
* Estado: Datos actualizados periódicamente.

***Acciones:***

* Programación de tareas recurrentes para extracción, transformación y carga (ETL).
* Automatización de flujos de trabajo para reducir errores manuales.

***6. Visualización y Producto Final (BI & Dashboards)***

* Ubicación: Data Warehouse → Power BI, Tableau, Streamlit.
* Estado: Datos listos para análisis de negocio y toma de decisiones.

***Acciones:***

* Desarrollo de dashboards interactivos en Power BI y Tableau.
* Creación de una aplicación web en Streamlit para exploración dinámica.
* Exposición de recomendaciones a través de FastAPI.

***7. Control de Versiones y Gestión***

* Ubicación: Repositorios de código (GitHub) y herramientas de gestión de tareas (ClickUp).
* Estado: Gestión del código y del equipo de trabajo.

***Acciones:***

* Uso de GitHub para versionado del código.
* Gestión de tareas y cronograma del proyecto en ClickUp.

Este flujo de trabajo permite una gestión eficiente y continua de los datos, asegurando que cada fase del ciclo de datos esté optimizada y automatizada para su análisis final.