📜 Propuesta de proyecto

En este escrito se deja constancia de la propuesta de trabajo y la planificación de esta:

* Entendimiento de la situación actual
  + Somos un equipo de científicos de datos que ha sido contratado para asesorar sobre una inversión de negocio. Nuestro cliente desea montar un negocio en un estado, de USA. Contamos con dos grades bases de datos: 1- Google Maps: datos sobre locales existentes y características generales; 2-Yelp: con reseñas de usuarios, puntuando los locales y describiendo su experiencia. Se nos permite usar datos adicionales.
* Objetivos

Nuestros objetivos principales se dividen en dos:

* + Responder a nuestro cliente cual es el tipo de negocio más popular, dentro de la categoría restaurantes, y que factores acompañan esta tendencia: lugar, horarios, tipo de comida, entre otros.
  + Desarrollar un sistema de recomendación a usuarios que les sugiera visitar restaurantes que no hayan conocido aun, en base a sus experiencias previas.
* Alcance

La temática planteada es de una inmensidad que supera al recurso tiempo y humano con el que se cuenta. Por ello se ha decidido sectorizar el estudio de mercado. Seleccionando sólo un estado, Wyoming, ente los 50 estados con los que se cuenta; y entre los tipos de negocio, un tipo de negocios, Restaurantes.

El sistema

Consideramos importante el concepto de escalabilidad, ya que el modelo puede servir como una base para abarcar a futuro otros estados, así como otros rubros.

Cabe aclarar que el sistema de recomendación solo será funcional para usuarios de alguna de estas plataformas (Google Maps y Yelp), y que hayan realizado interacciones en alguna de estas aplicaciones.

* Objetivos y KPIs asociados (planteó)

Objetivos específicos:

- 1) Identificar tendencias y patrones de consumo: mediante el análisis de las reseñas y los datos de los locales, identificando variables relacionadas como distribución geográfica, categoría específica, rangos de precios, etc.

- KPI: cantidad de reseñas positivas. Planteamos como meta una tasa promedio de un 1% mensual de aumento en reseñas positivas.

- KPI: Reducir el rating de estrellas faltantes a la mitad, anualmente. Buscando así aumentar la valoración en estrellas de manera asintótica.

- 2) Desarrollar un sistema de recomendaciones para usuarios, sugiriendo nuevos lugares para visitar en función de las elecciones, puntuaciones y reseñas realizadas.

- KPI: La tasa de conversión de recomendaciones en visitas reales a los restaurantes sugeridos por el sistema. Es decir, la proporción de usuarios que siguieron las recomendaciones y visitaron los lugares sugeridos en comparación con el número total de recomendaciones realizadas. Se espera una tasa de concreción del 10 %, del total de recomendaciones.

- KPI: la frecuencia con la que vuelven a utilizar el sistema de recomendación como indicadores de éxito. Se plantea que la misma sea de 1 vez por mes, en promedio.

- 3) (este objetivo sólo se llevará a cabo si se logran terminar en tiempo y forma los anteriores.) Predecir la popularidad de nuevos locales: utilizando los datos de los locales existentes y las reseñas de los usuarios, se puede desarrollar un modelo de aprendizaje automático para predecir la popularidad de los nuevos locales.

- KPI: Medir la popularidad de los locales. Ya que para implementar un modelo que prediga la popularidad de un local hipotético, necesitamos la etiqueta de popular o no popular, lograr o mantener la etiqueta de popularidad positiva.

* Repositorio Github

Armar un repositorio de Github para trabajar colaborativamente con todo el grupo. Debería ser público para que lo pueda ver tanto el mentor como el Product Owner. Van a tener que llevar adelante diferentes branches y controles de versiones de su propio trabajo.

* Solución propuesta:

A continuación, se detallan los objetivos principales a entregar como producto y las tareas específicas para lograr cada uno de los entregables.

Objetivo 1: Identificar tendencias y patrones de consumo, en base al feedback otorgado por clientes:

1. Planificación:

· Definir el alcance del análisis: Qué variables se analizarán, en qué período de tiempo se hará el análisis, etc.

· Identificar las fuentes de datos: Se debe definir cuáles son las fuentes de datos que se utilizarán para el análisis.

· Establecer los objetivos del análisis: Definir los objetivos específicos del análisis, que permitan identificar patrones y tendencias en la información.

1. Limpieza y preprocesamiento:

· Eliminar valores nulos: se deben detectar y eliminar los registros que no tengan información completa o que contengan errores.

· Unificación de datos: si los datos provienen de diferentes fuentes, se deben unificar y homogeneizar para que sean comparables.

· Selección de variables: Seleccionar las variables más relevantes y adecuadas para el análisis.

· Normalización de los datos: se deben normalizar las variables para que todas tengan la misma escala y sean comparables.

1. Análisis exploratorio de datos:

· Visualización de datos: generar gráficos y visualizaciones para comprender mejor los datos y detectar patrones o tendencias.

· Identificación de variables clave: identificar variables clave que puedan influir en los patrones de consumo.

· Análisis de correlaciones: identificar relaciones entre variables y cómo influyen unas en otras.

1. Análisis de patrones de consumo:

· Análisis de series de tiempo: identificar patrones a lo largo del tiempo, cómo se han comportado las variables en diferentes momentos.

· Clustering: agrupar los restaurantes según variables como tipo de comida, ubicación, precio, entre otras.

· Análisis de frecuencias: identificar cuáles son las preferencias de los clientes en cuanto a tipo de comida, ubicación, horarios, entre otros.

1. Presentación de resultados:

· Generación de informes: generar informes claros y precisos que muestren los resultados del análisis.

· Visualización de datos: presentar gráficos y visualizaciones que permitan entender los resultados del análisis.

· Presentación a nuestro cliente: presentar los resultados obtenidos a nuestro cliente, de manera clara y concisa, resaltando los hallazgos más importantes.

1. Disponibilizar la aplicación:

· Desarrollar la aplicación: desarrollar una aplicación que permita a los usuarios interactuar con los resultados del análisis.

· Desplegar la aplicación en un servidor: subir la aplicación a un servidor, para que los usuarios puedan acceder a ella.

· Realizar pruebas de la aplicación: asegurarse de que la aplicación funciona correctamente.

· Poner la aplicación en producción: hacer la aplicación disponible para el público en general.

Objetivo 2: Crear sistema de recomendación.

1. Preprocesamiento de datos:
   * Cargar los datos de las reseñas de Yelp y los datos de locales de Google Maps en un formato adecuado para análisis.
   * Realizar un análisis exploratorio de datos para entender mejor las características y distribución de los datos.
   * Limpiar los datos eliminando registros con información faltante, duplicados y posibles errores.
2. Creación de matriz de valoraciones:
   * Crear una matriz de valoraciones con las puntuaciones de los usuarios a los diferentes locales.
   * Transformar la matriz de valoraciones en una matriz de similitud de usuarios utilizando alguna medida de similitud como la correlación de Pearson o el coseno.
3. Identificación de usuarios similares:
   * Identificar los usuarios similares a partir de la matriz de similitud de usuarios creada en el paso anterior.
   * Definir un umbral para seleccionar a los usuarios más similares.
4. Creación de lista de recomendaciones:
   * Identificar los locales mejor puntuados por los usuarios similares.
   * Filtrar los locales que el usuario ya haya visitado.
   * Ordenar los locales recomendados según su puntuación promedio de los usuarios similares.
5. Evaluación del sistema de recomendación:
   * Evaluar el desempeño del sistema de recomendación utilizando alguna métrica como el error cuadrático medio o la precisión en la recomendación.
   * Ajustar el sistema de recomendación si es necesario.

Una vez creado debemos hacer que este sistema de recomendación sea accesible a nuestros usuarios. Por lo que debemos deployar nuestra aplicación, planificamos hacerlo en Google App Engine. Para lograrlo debemos :}

1. Crear una cuenta en Google Cloud Platform y crear un nuevo proyecto.

· Registrarse en Google Cloud Platform (GCP).

· Crear un nuevo proyecto en la Consola de GCP.

1. Configurar el entorno de desarrollo y preparar la aplicación para su despliegue.

· Descargar e instalar el SDK de Google Cloud.

· Configurar el SDK de Google Cloud para el proyecto.

· Preparar la aplicación para su despliegue en Google App Engine.

1. Desplegar la aplicación en Google App Engine.

· Desplegar la aplicación usando el SDK de Google Cloud.

· Verificar que la aplicación se ha desplegado correctamente.

1. Configurar la escalabilidad y el rendimiento de la aplicación.

· Configurar la escalabilidad y el rendimiento de la aplicación en la Consola de GCP.

· Monitorear el rendimiento y el uso de recursos de la aplicación.

1. Probar la aplicación y solucionar problemas.

· Probar la aplicación en el entorno de producción de Google App Engine.

· Solucionar cualquier problema que surja durante la prueba.

1. Configurar la dirección URL personalizada.

· Configurar una dirección URL personalizada para la aplicación en la Consola de GCP.

Metodología de trabajo

Como equipo hemos decidido, dentro de las metodologías ágiles usar el método Scrum, justificamos el uso de la metodología Scrum en este proyecto por varias razones:

· Colaboración: Scrum fomenta la colaboración entre los miembros del equipo, lo que puede ser muy beneficioso en un proyecto de ciencia de datos, donde se requiere una variedad de habilidades y conocimientos.

· Adaptabilidad: La adaptabilidad es una de las principales características de Scrum, lo que significa que puede ser una buena opción para proyectos de análisis de datos, donde los requisitos y objetivos pueden cambiar a medida que se avanza en el proyecto.

· Entrega iterativa: Scrum se enfoca en la entrega de valor de manera continua y en pequeñas iteraciones, lo que puede ser beneficioso en proyectos de análisis de datos, ya que permite al equipo ajustar el plan según sea necesario y entregar resultados tangibles a los stakeholders del proyecto.

· Gestión del riesgo y la incertidumbre: La metodología Scrum se centra en la gestión efectiva del riesgo y la incertidumbre, lo que puede ser especialmente útil en proyectos de análisis de datos, donde hay una mayor incertidumbre debido a la complejidad de los datos y a los imprevistos que pueden surgir.

Equipo de trabajo - Roles y responsabilidades

El equipo está conformado por 3 integrantes, lo cual implica particularidades a la hora de armar y distribuir roles y responsabilidades. Por lo que considero importante aclarar que la distribución de roles y tareas es a nivel organizativo y didáctico, ya que todos los miembros participarán y aportarán a todas las áreas de trabajo, aunque decidimos dejar responsables centrales en cada tarea, especificados en el siguiente diagrama.

1. Analista de Datos, Eduardo Gamal Gazziro:

· Realizar el análisis exploratorio de los datos para identificar patrones y tendencias de consumo.

· Desarrollar modelos estadísticos para predecir el comportamiento de los usuarios y la popularidad de los restaurantes.

· Realizar la limpieza y transformación de los datos para su uso en los modelos y visualizaciones.

1. Científico de Datos y gestión del proyecto, Martin Rodrigo Morales:

PM:

* Planificación y organización del proyecto.
* Establecimiento de objetivos y plazos.
* Gestión del presupuesto y de los recursos humanos.
* Identificación y gestión de riesgos y problemas.
* Supervisión del progreso del proyecto y ajuste de la planificación según sea necesario.

DC:

* Diseñar e implementar el sistema de recomendación basado en similitud con otros usuarios.

· Desarrollar modelos de aprendizaje automático para mejorar la precisión del sistema de recomendación.

· Evaluar el rendimiento del sistema de recomendación y realizar ajustes según sea necesario.

1. Ingeniero de Datos, Juan José Pérez:

· Implementar y mantener la infraestructura necesaria para el procesamiento y almacenamiento de los datos.

· Desarrollar scripts y herramientas para automatizar el procesamiento de los datos.

· Integrar los diferentes sistemas y herramientas necesarias para el proyecto, como Google Cloud, Spark, BigQuery, Google App Engine y Power BI.

Cuadro de estimación de tareas y stack tecnológico previsto.

A continuación se presenta un cuadro donde se detallan las tareas a nivel general, su estimación temporal, su encargado central y el stack tecnológico planificado para cada actividad.

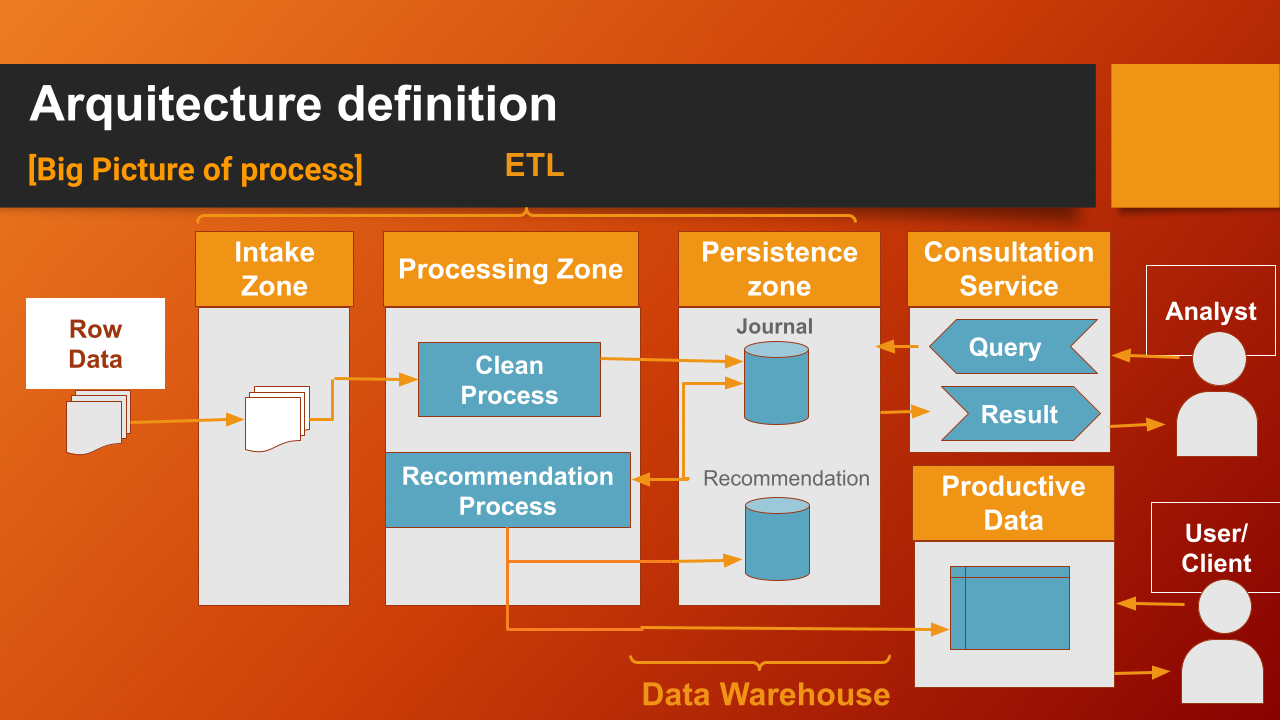
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tarea | Descripción | Dependencias | Responsable | Herramientas | Tiempo estimado |
| Análisis de requerimientos | Identificar los requerimientos del cliente, definir los objetivos del proyecto y las preguntas de investigación | Ninguna | Analista de datos | Google Docs,, Notion | 2 días |
| Selección de herramientas | Investigar y seleccionar herramientas y tecnologías adecuadas para el proyecto, y establecer un entorno de desarrollo | Análisis de requerimientos | Científico de datos | Google Cloud, BigQuery, Spark, Python | 2 días |
| Recopilación y limpieza de datos | Descargar datos de Google Maps y Yelp, fusionarlos en una base de datos y limpiar los datos para su uso | Selección de herramientas | Ingeniero de datos | Python, Spark | 5 días |
| Análisis exploratorio de datos | Identificar patrones y tendencias en los datos, así como cualquier problema en los datos | Recopilación y limpieza de datos | Analista de datos | Python, Spark, Power BI | 3 días |
| Modelado de datos | Desarrollar modelos predictivos para el consumo y la popularidad de los restaurantes, y evaluar su rendimiento | Análisis exploratorio de datos | Científico de datos | Python, Spark | 5 días |
| Integración de modelos | Integrar los modelos en la plataforma de negocios del cliente para proporcionar recomendaciones personalizadas | Modelado de datos | Ingeniero de datos | Google Cloud, BigQuery | 2 días |
| Creación de informes y visualizaciones | Crear informes y visualizaciones para presentar los resultados del análisis de datos y los modelos | Integración de modelos | Analista de datos | Power BI | 3 días |
| Pruebas y validación | Probar y validar el modelo, informes y visualizaciones en diferentes escenarios y en situaciones de datos desconocidas | Integración de modelos, Creación de informes y visualizaciones | Científico de datos | Python, Spark, Power BI | 2 días |
| Documentación y entrega | Documentar el proyecto y entregar los resultados al cliente | Todas las tareas | Equipo completo | Google Docs | 1 día |

**Diagrama de Gantt:**

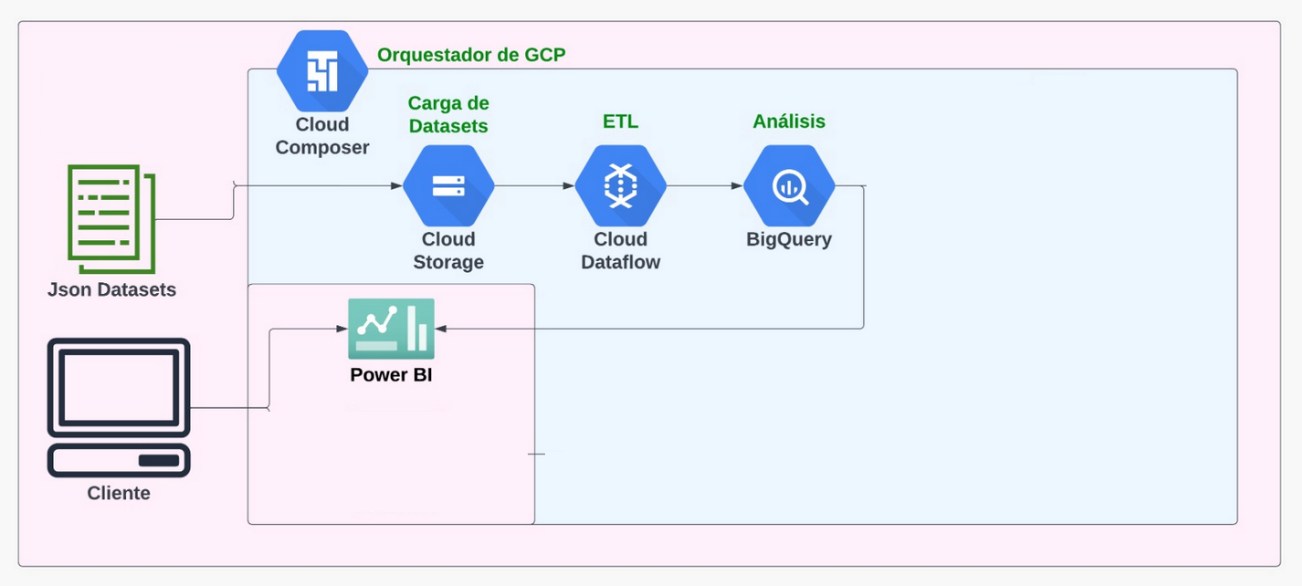
**Aquí se deja el enlace al diagrama de Gantt, el cual contiene las tareas, el tiempo estimado, su estado(por empezar, en progreso, terminada) y las dependencias que entre ellas se estabelce.**

<https://www.notion.so/acb8a036474249bb99b8a35523633658?v=b88dbab532a14580a8e2ba5e4e4547b4>

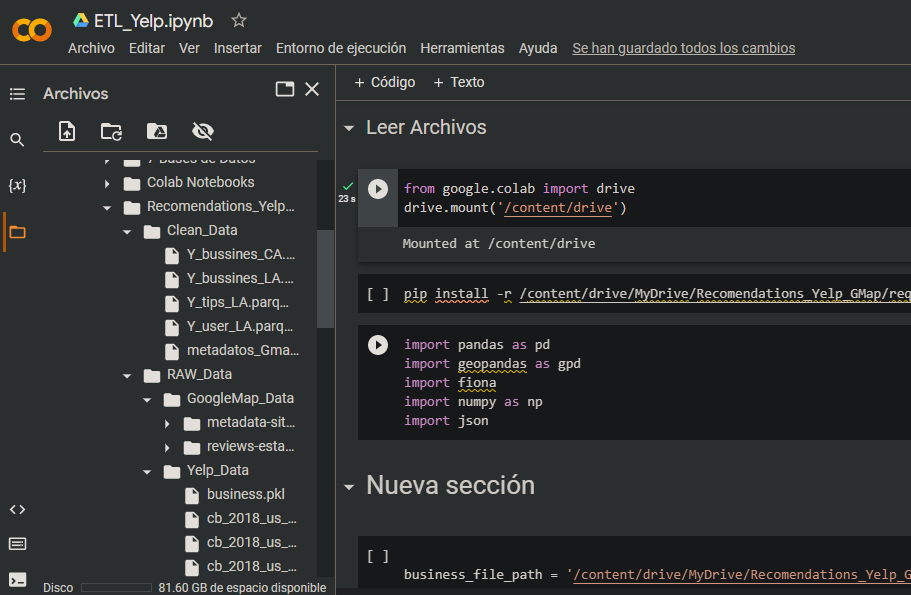
**Data Engineering – Arquitectura de la solución propuesta:**



En principio se propuso la implementación de una arquitectura basada en la nube para el manejo de bigdata (archivos de 5 Gigabytes o más grandes) y automatización del proceso de ETL basado en pipelines. Se estudió las características de los servicios ofrecidos por los principales vendors de servicios en la nube y se eligió en principio Google Cloud por ofrecer un sistema de facturación en el que se paga según las APIs y servicios que se consumen.

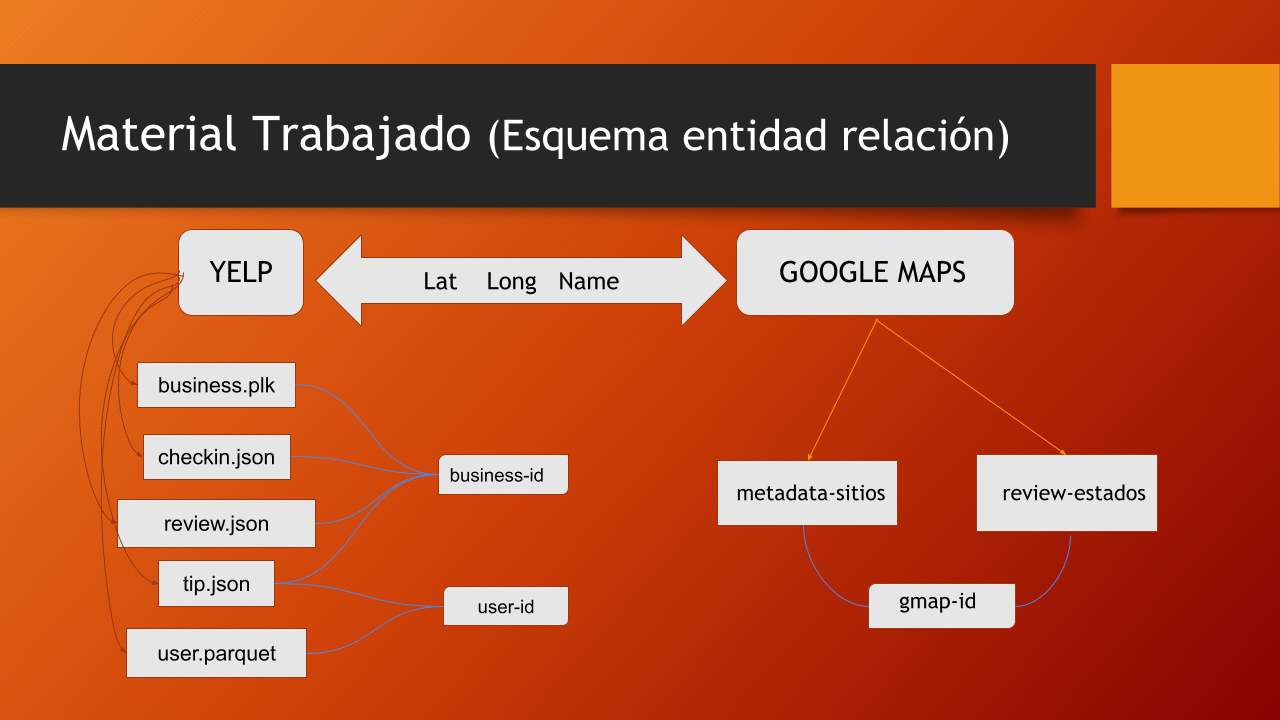


A medida que se avanzo en la implementación se observo que los recursos utilizados junto con los costos de servicios asociados eran mucho más altos que el bajo presupuesto que se tenía para este proyecto así que basándonos en el hecho de que la principal prioridad era la entrega de un Producto Mínimo Viable se desidió hacer un cambio en la estrategia de desarrollo y se optó por un stack tecnológico de bajo costo.



Se eligió Google Colab como plataforma para ejecutar los script de Python en formato notebook de jupiter en los procesos de EDA y ETL y Google Drive como plataforma para almacenar la data cruda y tablas generadas en los procesos de ETL. En esta versión del proyecto no se requirió procesar datos que se generaran en tiempo real sino que se trabajó con data estática, así que no fue necesario la implementación de pipelines, en cambio se diseñó e implementó dos script (ETL\_Yelp y ETL\_GoogleMap) en Python en formato Jupiter que en las primeras etapas del Analisis Exploratorio de Datos (EDA) permitió analizar los datos paso a paso y posteriormente se modificaron para ejecutar el proceso de ETL en una sola ejecución. Finalmente se generaron los archivos con datos estructurados y limpios en formato parquet que posteriormente alimentó a la etapa de Data Analysis y generación de Dashboard e informes.

Esquema Entidad-Relación Hipotético:



Conclusiones:

* Nuestro norte central es entregar un producto mínimo viable.
* Buscamos la calidad en el proceso, mediante la premisa calidad y no cantidad..
* Entendemos que la finalidad última es el aprendizaje de un entorno laboral, con las exigencias y avatares que esto conlleva.
* Creemos que la comunicación, la flexibilidad y el trabajo coordinado en equipo son nuestras herramientas más poderosas para lograr los objetivos planteados.