

PROYECTO FINAL

Demo1

Leonardo Cortés Marcelo Atencio Federico López Andrés Ruiz

Febrero 2024

Resumen

El siguiente documento presenta el informe de la primer etapa del proyecto final a presentar en el Boot Camp de Data Science de "Soy Henry". En este proyecto se plantea un esquema de trabajo desde la ingesta de datos crudos proporcionados por el cliente hasta el desarrollo de un producto final entregable.

Índice

1. Introducción	2
2. Entendimiento de la situación actual	2
3. Equipo de trabajo	3
4. Objetivos	4
5. Metodología de Trabajo	5
6. Stack Tecnológico	6
7. Alcance	8
8. Key Performance Indicators (KPI)	8
9. Repositorio Git Hub	10
10. Planificación del Proyecto	10
11. Trabajo preliminar	10
11.1. Global Pre ETL (Extract, Transform and Load)	10
11.2. Pre EDA (Exploratory Data Analysis)	11

1. Introducción

ARCOL Data Solutions es una empresa conformada por un equipo de cuatro apasionados expertos en el campo de 'Data' motivados por la constante curiosidad y la dedicación al análisis. Actualmente ARCOL Data Solutions asume el reto de preparar, analizar y modelar una cantidad de datos proporcionados por el cliente con el fin de entregar un análisis exhaustivo junto con un modelo de recomendación que permita al usuario tomar decisiones de inversión en el mercado real del sector turístico en Estados Unidos. Como misión se proponen abordar las preguntas que más inquietan a los usuarios al transformar datos en inteligencia. Para el actual proyecto nuestra principal fuente de información son los Datasets nativos proporcionados por el sitio YELP. Estos conjuntos de datos contienen recomendaciones y calificaciones generadas por usuarios durante sus visitas a diversos lugares en todo Estados Unidos, abarcando desde restaurantes hasta hoteles. Esta información nos ofrece una visión única y detallada de las experiencias de los usuarios en diferentes establecimientos.

Además, complementamos nuestro arsenal de datos con la información suministrada por GOOGLE MAPS. Esta fuente incluye datos como la ubicación precisa de diversos lugares registrados en el sitio, las calificaciones otorgadas por los usuarios, así como los sitios web asociados a estos lugares, entre otros aspectos relevantes. Esta combinación de datos nos permite realizar análisis exhaustivos y proporcionar respuestas informadas a las preguntas que nos planteamos. En ARCOL, estamos comprometidos con la transformación de datos en inteligencia práctica y valiosa para la toma de decisiones.



Figura 1: ARCOL Data Solutions.

2. Entendimiento de la situación actual

1. “El fracaso es una gran oportunidad para empezar otra vez con más inteligencia. (Henry Ford)”
2. “Lo peor que te puede pasar no es no tener una mina de oro, sino tener la mina y no extraer el oro”

Las citas de Henry Ford y la analogía sobre la mina de oro establecen el tono para nuestra propuesta de proyecto, destacando la importancia de aprender de los fracasos y aprovechar

las oportunidades que se presentan. En el entorno digital actual, el éxito empresarial está estrechamente vinculado a la presentación y comunicación efectiva en medios digitales, así como a la capacidad de procesar las recomendaciones y devoluciones de los consumidores. En este contexto, reconocemos que la retroalimentación honesta de los usuarios, facilitada por la privacidad que brinda internet, se ha convertido en una fuente valiosa de información. La capacidad de una empresa para comprender la perspectiva del cliente y mejorar continuamente sus procesos se vuelve crucial. Aquellas organizaciones que no se sumergen en este análisis tienen una alta probabilidad de enfrentar dificultades y fracasos.

Aquí es donde identificamos una clara Oportunidad de Negocio: enfocarnos en proyectos que involucren negocios con tendencia a fracasar, aquellos que enfrentan dificultades o que tienen un potencial de crecimiento no explotado. Nuestra propuesta radica en el uso de herramientas especializadas en el análisis de datos para identificar y comprender patrones, tendencias y áreas de mejora. Al ofrecer a nuestros clientes información precisa y relevante, pretendemos equiparlos con las herramientas necesarias para la toma de decisiones informadas y estratégicas en sus negocios. En resumen, buscamos convertir el fracaso en una oportunidad para el crecimiento empresarial, proporcionando inteligencia a partir de los datos disponibles.

3. Equipo de trabajo

El equipo de ARCOL Data Solutions, está integrado por profesionales con conocimientos integrales en el campo de datos que les permite desarrollar e interactuar en cualquier rol. Para el proyecto específico de «*Modelo de recomendación para inversionistas*», aunque todos los miembros del equipo están involucrados con todas y cada una de las etapas del proyecto, cada integrante lidera un ciclo específico del proyecto, en aras de tener una mejor organización y coordinación.

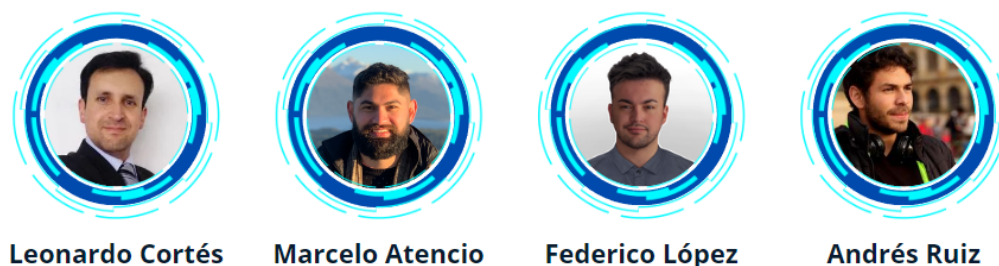


Figura 2: Equipo.

- **Leonardo Cortes.** *Project Manager (PM) y Científico de Datos:*

Leonardo realiza la gestión y coordinación global del proyecto, estableciendo objetivos claros, plazos y entregables para garantizar la oportuna y exitosa disposición del proyecto al cliente. Asigna los recursos, herramientas y tecnologías a usar, para garantizar que el proyecto se complete de manera eficiente. Así mismo, es el encargado de automatizar cada fase del proyecto y ejecutar los modelos de Machine Learning que se implementaran en el producto final entregable.

- **Marcelo Atencio.** *Task Manager e Ingeniero de Datos:*

Marcelo gestiona, coordina y documenta las tareas específicas relacionadas con el proyecto, con base en la gestión direccionada por el PM, asignando labores, realizando el seguimiento de las tareas y gestionando el tiempo y cumplimiento de las mismas, con el fin de cumplir los objetivos planteados por el PM. También es en cargo de liderar la creación e implementación de las funciones de exploración y modelamiento de datos durante el proceso de ingeniería, y dejarlos listos en el data warehouse para los procesos de analítica y machine learning.

- **Federico López** *Analista de Datos:*

La responsabilidad de Federico recae en el análisis, interpretación y presentación de datos para ayudar al cliente a tomar decisiones informadas. En particular, las funciones de Federico son liderar el proceso de analítica de datos, realizando y coordinando la preparación de datos para hacer uso de ellos en el respectivo análisis, la creación de informes y visualizaciones, interpretar los datos hacer recomendaciones con base al análisis realizado.

- **Andrés Ruiz.** *Ingeniero de Datos y de Machine Learning:*

El papel de Andrés dentro del equipo es fundamental pues es el encargado de liderar los procesos de ingeniería de datos y de procesamiento de machine learning. En ingeniería de datos, desde el inicio del proyecto, coordinando y monitoreando el desarrollo del código necesario para la ingesta de datos, la transformación de los mismos y la carga de los conjuntos de datos limpios al data warehouse. Así mismo es el encargado de coordinar el proceso de implementación del análisis exploratorio de los datos y los modelos de aprendizaje, con el fin de establecer con todo el equipo las variables necesarias para el desarrollo de un proyecto exitoso y la entrega de un producto impecable.

4. Objetivos

1. **Objetivo Empresarial a 3-6 Meses:** Encaminar nuestro negocio hacia el éxito, incrementando de manera porcentual y adaptada a las necesidades individuales de cada cliente. Este logro se materializa a través de la implementación de un sólido sistema de recomendación basado en indicadores actualizados de manera constante.
2. **Descripción del Proyecto:** Crear de un sistema de recomendación diseñado especialmente para nuevos inversores. Este sistema se fundamenta en indicadores financieros actualizados en tiempo real, permitiendo a nuestros clientes acceder a información precisa y relevante sobre los proyectos más atractivos en el mercado.
3. **Objetivo del Sistema de Recomendación:** Brindar a nuestros clientes sugerencias informadas y precisas sobre los proyectos más atractivos. Queremos ser la guía confiable en la identificación de oportunidades de negocios prometedoras, permitiendo a los inversores tomar decisiones respaldadas por datos actualizados y relevantes.

4. **Precisión y Confiabilidad:** Proporcionar recomendaciones basadas en datos actualizados constantemente, asegurando la precisión de la información.
5. **Guía Informada:** Ofrecer un sistema que actúe como una brújula para nuevos inversores, ofreciendo una guía informada y adaptada a las necesidades individuales de cada cliente.
6. **Identificación de Oportunidades:** Facilitar la identificación de oportunidades de negocios prometedoras, optimizando el proceso de toma de decisiones para nuestros clientes.
7. **Medición del Éxito:** Medir a través del crecimiento porcentual de la satisfacción de los clientes, el aumento en la rentabilidad de las inversiones y la consolidación de nuestro negocio como referente en el asesoramiento financiero personalizado.

Estamos comprometidos con el éxito de nuestros clientes y estamos seguros de que este sistema de recomendación jugará un papel crucial en su trayectoria hacia el éxito financiero. ¡Juntos, construiremos un futuro próspero!

5. Metodología de Trabajo



Figura 3: Metodología de Trabajo.






Para la ejecución del proyecto, planteamos la implementación de una metodología de Monitoreo Grupal del Progreso Individual, con un responsable visible de cada etapa. Bajo esta metodología, llevamos a cabo tanto tareas colaborativas como asignaciones individuales. En ambas instancias, se realizan comentarios sobre los avances alcanzados, y tras la aprobación unánime del equipo, se integra el nuevo progreso al repositorio o archivo correspondiente. Este enfoque fomenta la transparencia, la colaboración y la eficiencia en el desarrollo del proyecto.

Esta metodología se complementa con roles asignados de manera equitativa entre los cuatro integrantes del proyecto. Esto posibilita el intercambio de conocimientos en cada







etapa del proyecto, promoviendo así la adquisición de versatilidad en el ámbito de Ciencia de Datos.

La designación de responsabilidades para cada tarea y fase implica asignar roles específicos, como encargados de verificar la finalización de las tareas, revisar el contenido y documentar adecuadamente cada paso. Estos roles están claramente especificados en el diagrama de Gantt, proporcionando una estructura organizativa que facilita la ejecución eficiente y colaborativa del proyecto.



6. Stack Tecnológico

-  **Microsoft Fabric:** Es el corazón del proyecto, pues las herramientas integradas permiten que Microsoft Fabric este presente todas las etapas del ciclo de desarrollo del proyecto, abarcando desde la ingesta inicial de datos, pasando por el procesamiento, transformación y análisis hasta la visualización de la información resultante y la automatización de todo el proyecto. También permite integrar los procesos de Machine Learning para realizar los modelos necesarios para entregar el producto final. La elección de esta herramienta se basa en la capacidad que ofrece para proporcionar una experiencia profesional y auténtica en un entorno similar en el que se desenvuelven las empresas especializadas en ingeniería, análisis y ciencia de datos en el sector.
-  **Apache Spark:** Apache Spark es un potente motor de procesamiento distribuido diseñado para manejar grandes volúmenes de datos de manera rápida y eficiente. Una de las características distintivas de Apache Spark es su capacidad para realizar operaciones de procesamiento de datos en memoria, lo que lo hace significativamente más rápido que los sistemas de procesamiento de datos tradicionales. Debido a su gran conjunto de bibliotecas que facilitan de manera versátil el procesamiento de datos en lenguajes python (Pyspark) y la consulta sobre datos estructurados con lenguaje SQL (Spark SQL), Spark es también una herramienta integral que se utilizará durante los ciclos de cada etapa del proyecto.
-  **Jupyter Notebook:** Jupyter Notebook es una aplicación web de código abierto que permite crear y compartir documentos interactivos que contienen código en vivo, visualizaciones, texto explicativo y ecuaciones matemáticas. La característica distintiva de Jupyter Notebook es su capacidad para combinar texto enriquecido, como Markdown o LaTeX, con código ejecutable en varios lenguajes de programación, incluyendo Python. Es una herramienta que se encuentra integrada dentro del entorno de Microsoft Fabric para el procesamiento de código bajo lenguaje Pyspark y SparkSQL, y por lo tanto es una herramienta que se utilizará durante todo el proyecto.
-  **Visual Studio Code:** (VS Code) es un popular editor de código fuente que debido a su interfaz intuitiva, amplia compatibilidad y herramientas integradas, será utilizado para el desarrollo del código Python de las funciones integradas al proyecto, especialmente durante la etapa de ingeniería de datos y desarrollo de Machine Learning.
-  **Python:** Python es un lenguaje de programación de alto nivel, interpretado y multipropósito. Python será utilizado en todas las etapas del proyecto debido a su

simplicidad, versatilidad y a las poderosas bibliotecas que ofrece. En este caso, Python se utilizará como una herramienta fundamental gracias a bibliotecas como NumPy, pandas, Matplotlib, Seaborn, Scikit-Learn, TensorFlow, herramientas que utilizaremos en la primer etapa con la ingesta de datos y análisis primario de datos. Durante la etapa 2 con el desarrollo del EDA y la respectiva selección de datos para almacenar en el warehouse. Y en la etapa final para el desarrollo de los diferentes modelos de Machine Learning que se necesiten realizar.

-  **SQL:** SQL, (Structured Query Language), es un lenguaje de programación diseñado para gestionar y manipular bases de datos relacionales. Al ser el lenguaje estándar de facto para trabajar con bases de datos, será el instrumento que se utilizará particularmente al final del proceso de ingeniería de datos y en el proceso de análisis de datos, al realizar las consultas y modificaciones respectivas para realizar un modelo dimensional y relacionado para luego cargar los datos necesarios a la herramienta que se utilizará para la visualización del análisis.
-  **Power BI:** Power BI es una plataforma de análisis que permite visualizar y compartir datos de manera efectiva para tomar decisiones informadas. Ofrece una amplia gama de herramientas para la preparación de datos, visualización de datos, análisis y colaboración en un solo lugar. Para el el proyecto en particular, se utilizará esta herramienta (que se encuentra implementada dentro del entorno de Microsoft Fabric) durante la etapa que corresponde al análisis de datos, tomando datos ya trabajados y estructurados con anterioridad durante la etapa de ingeniería de datos, para presentar una visualización que permita entender los datos y así permitir que el cliente tome las mejores decisiones con base en ellos.
-  **LaTeX:** LaTeX es un sistema técnico de composición de textos para producir documentos profesionales de alta calidad tipográfica, utilizando un lenguaje de marcado sencillo que describe la estructura lógica del documento, como títulos, secciones, listas, tablas y ecuaciones. Para el caso, el presente documento, así como todos los informes y procesos documentales que se requieran en el proyecto, serán desarrollados con esta herramienta y presentados como un documento PDF de alta calidad y profesionalismo.
-  **Git:** Git es un sistema de control de versiones distribuido que se utilizará durante todos los ciclos del proyecto para realizar seguimiento de avances y trabajo colaborativo entre el equipo.
-  **GitHub:** GitHub es una plataforma de desarrollo colaborativo basada en la web que utiliza el sistema de control de versiones Git. Allí almacenaremos paso a paso las instancias, junto con la documentación y el código realizado para desarrollar el proyecto, por lo cual es una herramienta que estará presente durante cada una de las etapas del proyecto.
-  **Scikit Learn:** Scikit Learn es una biblioteca de aprendizaje automático (machine learning) para el lenguaje de programación Python. Se utiliza para desarrollar y aplicar algoritmos de aprendizaje automático de manera eficiente y sencilla. ES la librería que utilizaremos para trabajar durante la etapa de procesamiento de datos y machine

learning, para realizar la mayoría de los modelos propuestos en nuestros proyectos, debido a la amplia gama de algoritmos de aprendizaje supervisado y no supervisado, así como herramientas para el preprocesamiento de datos.

-  **Fast API:** FastAPI es un marco (framework) web para construir APIs (interfaces de programación de aplicaciones) en Python de manera rápida y eficiente, del que se destacan su alta velocidad, facilidad de uso y su capacidad para generar automáticamente documentación interactiva (Swagger UI y ReDoc) para las APIs que se construyen con él. Por esta razón, es la herramienta que en ARCOL se ha seleccionado para realizar durante la fase de desarrollo de machine learning, los deploy de los modelos y funciones trabajadas previamente durante el procesamiento de datos.
-  **Jira:** Jira es una aplicación de software que se utiliza principalmente para la gestión de proyectos y seguimiento de problemas. EN nuestro caso, Jira será utilizada para planificar, rastrear y gestionar las tareas propuestas por y para el trabajo en equipo.

7. Alcance

- En complemento a los datos proporcionados por nuestro cliente, incorporamos información relevante sobre los datos económicos del mercado real en el sector.
- Uso de herramientas tecnológicas incluidas dentro del servicio que brinda Microsoft Fabric (Data Lake, Data Warehouse, Data Factory, Power BI).
- Desarrollo de un dashboard para la visualización del análisis de la información.
- Ofrecer un sistema de recomendación a través de una API.
- El análisis incluirá negocios comerciales representativos del sector turismo dentro de EEUU, realizando una posterior ampliación del segmento según la necesidad de nuestros clientes.

8. Key Performance Indicators (KPI)

Formulamos un conjunto de indicadores diseñados para proporcionar una evaluación integral de la situación actual de los negocios. Estos indicadores tienen como objetivo principal facilitar a nuestros inversionistas la toma de decisiones informadas, alineándose con la consecución de nuestro objetivo principal. Los indicadores propuestos son:

1. Índice de Satisfacción del Cliente (ISC):

Fórmula: Sean:

x = Promedio de calificaciones de usuarios.

y = Máxima calificación posible

$$ISC = \frac{x}{y} \times 100$$

Objetivo: Evaluar la satisfacción general de los clientes en comparación con el año anterior. Mayor ISC indica mayor satisfacción, lo que puede ser atractivo para inversionistas.

2. Calidad de Servicio (QS) por Ubicación:

Fórmula: Sean:

x = Promedio de calificaciones de usuarios por ubicación.

y = Máxima calificación posible

$$QS = \frac{x}{y} \times 100$$

Objetivo: Evaluar semestralmente la calidad del servicio en ubicaciones específicas. Proporciona información detallada sobre el rendimiento en diferentes áreas geográficas.

3. Índice de Reputación Online (IRO):

Fórmula: Sean:

x = Número total de reviews positivos.

y = Número total de reviews

$$IRO = \frac{x}{y} \times 100$$

Objetivo: Medir la reputación online del negocio cada trimestre. Una alta proporción de reviews positivos puede ser atractiva para inversionistas.

4. Ubicación Estratégica (UE):

Fórmula: Sean:

x = Número de atracciones turísticas cercanas.

y = Distancia promedio a atracciones

$$UE = \frac{x}{y}$$

Objetivo: Evaluar en periodos anuales la ubicación en relación con las atracciones turísticas. Una mayor puntuación indica una ubicación estratégica que puede atraer a más visitantes.

5. Ubicación con Potencial de Crecimiento (UPC):

Fórmula: Sean:

x = Promedio de calificaciones de usuarios en áreas en desarrollo.

y = Promedio de calificaciones de usuarios en áreas consolidadas

$$UPC = \frac{x}{y} \times 100$$

Objetivo: Identificar año a año negocios que están ubicados y/o se expanden a áreas con potencial de crecimiento. Una mayor puntuación indica oportunidades para inversión en expansión.

9. Repositorio Git Hub

En el siguiente link se accede al repositorio del proyecto en GitHub:

[Sistema de Recomendación de Inversiones - Repositorio GitHub](#)

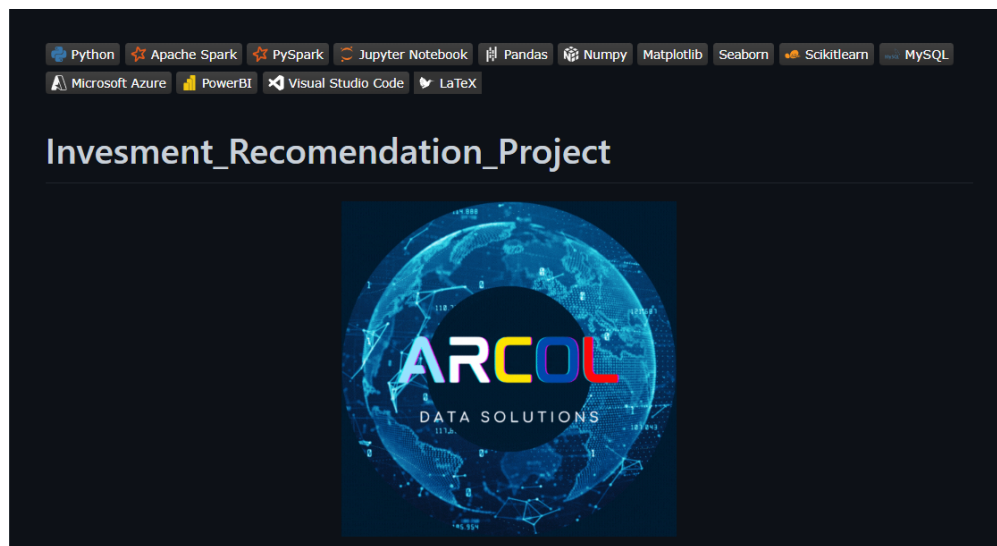


Figura 4: Repositorio GitHub.

10. Planificación del Proyecto

Mediante el uso de un diagrama de Gantt detallamos todas las tareas a ejecutar para alcanzar el objetivo, duración de las mismas y su dependencias.

11. Trabajo preliminar

11.1. Global Pre ETL (Extract, Transform and Load)

En la fase inicial del proyecto, establecimos un Data Lake en Microsoft Fabric y cargamos directamente los conjuntos de datos de Google Maps y Yelp desde un enlace de Google Drive. Desarrollamos un código automatizado para leer archivos en formato .json y .pickle, convirtiéndolos en .parquet para facilitar la manipulación posterior. Destacamos la creación del archivo `utils.py` que contiene funciones esenciales, accedidas desde el código principal.

Optamos por almacenar los archivos en formato parquet desde el principio para reducir el consumo de memoria y agilizar la manipulación de datos en etapas posteriores. Esto se logra gracias a las ventajas de eficiencia y rendimiento que ofrece el formato parquet en comparación con otros. La modularidad del código y la accesibilidad a funciones clave desde `utils.py` contribuyen a la organización y mantenimiento efectivos del proyecto.

11.2. Pre EDA (Exploratory Data Analysis)

1. **Uso de la memoria** Durante este proceso, llevamos a cabo una comparativa entre el consumo de memoria de los archivos originales y su contraparte almacenada en formato .parquet dentro del Data Lake. Inicialmente, los datos nativos representan un consumo aproximado de 36 GB. Sin embargo, después de aplicar un procesamiento posterior, los archivos en formato parquet reducen significativamente este consumo a 10 GB, implicando una reducción del casi 70 % en el uso de recursos. Este eficiente manejo de la memoria no solo optimiza la capacidad de almacenamiento, sino que también mejora la eficiencia global del sistema.
2. **Análisis preliminar de la calidad de datos** Mediante una revisión inicial de los datos, nuestro propósito es visualizar la información proporcionada en cada archivo, comprendiendo el tipo de dato, la cantidad de registros y la presencia de datos nulos. Además, buscamos verificar de manera intuitiva si contamos con la información necesaria para los indicadores propuestos. En este proceso, identificamos cinco aspectos clave que requieren especial atención en relación con la calidad de los datos:
 - a) *Velocidad*: La velocidad de generación constante de datos plantea un desafío significativo para la obtención oportuna de conclusiones e indicadores, dado que estos pueden volverse obsoletos al momento de su creación. En nuestro caso, la base de datos abarca un extenso conjunto de comentarios de clientes recopilados de dos sitios web que incorporan nueva información de manera continua, minuto a minuto. El desafío principal radica en lograr una automatización integral de todo el proceso para adaptarse eficazmente a la expansión constante de la base de datos. Esta automatización no solo garantizará la agilidad en el análisis de los datos existentes, sino que también facilitará la gestión eficiente de la información entrante, asegurando la relevancia y la actualización constante de los indicadores y conclusiones obtenidos. Abordaremos este tema específicamente en el desarrollo del pipeline ETL.
 - b) *Variedad*: La información original está almacenada en archivos con formato .json y .pickle. La transformación de estos archivos a formato .parquet nos brindará la capacidad de optimizar el uso de recursos y mejorar la accesibilidad de los datos.
 - c) *Volumen*: La base de datos cuenta con una cantidad suficiente de registros para llevar a cabo un análisis preciso. En cada carpeta tenemos:
 - Metadata-sitios:
 - 275000 registros
 - 0 registros totalmente nulos
 - 7 columnas que contienen datos nulos; de las cuales 5 tienen mas del 20 % de datos nulos
 - Reviews-estados:
 - 51 archivos los cuales tienen entre 300 mil y 2,8 millones de registros
 - 0 registros nulos

- 3 columnas tienen por arriba del 35 % de datos nulos; las mismas para todos los archivos
- Yelp:
 - 4 archivos con 9,1 millones de registros en su totalidad.
 - 0 registros nulos
 - 1 archivo con 3 columnas que contienen bajo porcentaje de datos nulos
- d) *Valor:* En este contexto, reconocemos la inmensa valía de toda la información proporcionada para alcanzar los objetivos establecidos. Esta valiosa gama de datos incluye calificaciones, comentarios, ubicaciones, precios, descripciones de negocios, información sobre clientes y otros atributos clave. Dado que nuestra máxima prioridad es maximizar la utilidad de estos datos, anticipamos la necesidad de llevar a cabo un exhaustivo proceso de filtrado y estructuración durante la etapa de ETL. Este enfoque nos permitirá otorgar relevancia de manera precisa a la información, alineándose con nuestros propósitos específicos.
- e) *Veracidad:* A pesar de que los datos provienen de dos sitios ampliamente reconocidos en cuanto a reviews y comentarios, resulta imperativo llevar a cabo un análisis exploratorio de datos (EDA) de manera exhaustiva. A través de este análisis, estamos en posición de identificar posibles sesgos, evaluar consistencias, gestionar valores nulos, medir la confiabilidad de la información y abordar cualquier ruido presente en los datos. Este proceso nos capacita para obtener una comprensión profunda de la calidad y las características intrínsecas de los datos, aspecto crucial para llevar a cabo un análisis informado y garantizar la fiabilidad de los resultados obtenidos.