BIDATING

EMPRESA CONTRATANTE:
Wyndham Hotels & Resorts

OBJETO DEL CONTRATO:
ASESORAMIENTO INTEGRAL EN PLAN DE EXPANSIÓN DE LÍNEA HOTELERA PREMIUM A LO LARGO DEL TERRITORIO ESTADOUNIDENSE

Marzo del 2024

ÍNDICE

1.INTRODUCCIÓN

En el marco de la consultoría contratada por Wyndham Hotels & Resorts, se presenta la primera entrega del proyecto donde se realiza un planteamiento general del proyecto, a donde se desea ly que pasos y metodología se va a seguir para la ejecución del mismo.

Bidating como aliado estratégico de Wyndham Hotels & Resorts, y asesor en su plan de expansión para el periodo 2024 - 2026, pone a disposición su equipo de profesionales para brindar primero las herramientas para la toma de decisiones en cuanto a la expansión de Hoteles de la línea premium de la cadena y en segundo lugar brindar herramientas de control y análisis de esta inversión tomada de forma informata y data driven.

2.OBJETIVOS

2.1. Objetivo General

Ayudar a nuestro cliente Wyndham Hotels & Resorts a ejecutar un plan de expansión a nivel nacional de su línea Premium Hotelera con el objetivo de aperturar 100 hoteles a nivel nacional entre el presente año 2026. Lo anterior entregando al cliente los análisis y herramientas para tomar una desición impulsada por datos.

2.2. Objetivos Específicos

- Realizar un análisis de los datos de reseñas de restaurantes de Google y Yelp para establecer las tendencias del consumidor por ubicación geográfica.
- Por medio de fuentes de datos a parte de las reseñas de como atracciones turísticas, museos y poblacionales establecer indicadores y correlaciones de qué datos monitorear para establecer en qué lugar geográfico de los estados unidos hay más probabilidad de éxito para un hotel de línea premium.
- Realizar un dashboard interactivo que le permita a nuestro cliente visualizar los KPI e indicadores definidos que servirán para realizar un seguimiento y control al plan de expansión.

3.FUNDAMENTACIÓN

Respecto a la industria de Wyndham Hotels & Resorts

La industria hotelera juega un papel crucial en la economía de los Estados Unidos y se espera que para el año 2025 esta industria genera cerca de 2.5 billones de dólares teniendo una participación proyectada del PIB nacional de cerca del 2.5%.

Respecto al proyecto realizado

Al momento de tomar una decisión de inversión a gran escala es de vital importancia tomar las precauciones para asegurar un éxito de dicha inversión, un elemento importante es tomar decisiones basadas en datos y análisis de los mismos para identificar oportunidades, retos y establecer objetivos a cumplir para un seguimiento continuo de el desarrollo del proyecto y por ende del éxito de la inversión.

4.KEY PERFORMANCE INDICATORS

4.1. **Impacto de proyecto (IP):** medido en el porcentaje de variación semanal del número de reseñas de 4 o 5 estrellas dejadas en los hoteles aperturados.

Meta semanal : 5%

Mes	Semana	Reseñas positivas	%Variación
Abril	s1	24	
Abril	s2	27	**11%**
Abril	s3	22	**-23%**

$$IP\ (\%) = (1 - \frac{Numero\ reseñas\ positivas\ semana\ anterior}{Numero\ reseñas\ positivas\ semana\ en\ curso})\ x\ 100$$

4.2. **Relevancia de Competencia (RC):** Medición de las reseñas nuevas de los hoteles competidores sobre el número de hoteles presentes en el mismo condado comparado a las reseñas nuevas de los hoteles aperturados, semanalmente.

Lo anterior para Hoteles de la competencia que estén dentro de la zona de influencia del hotel aperturado.

Meta mensual : ≤ 50%

$$RC\ (\%)\ =\ (1\ -\ \frac{\frac{nHCD}{nRNHCD}}{\frac{nHA}{nRHA}})\ x\ 100$$

nHCD = Nùmero Hoteles competencia Directa.
nRNHCD = Número de reseñas nuevas de Hoteles competencia Directa.
nHA = Número Hoteles aperturados.
nRHA = Número de reseñas nuevas de Hoteles Aperturados.

4.3. Reputación Online (RO): Reputación online del Hotel aperturado medida anualmente calculando el número de reseñas positivas sobre el número de reseñas totales de ese Hotel. Entiéndase por reseñas positivas 4 y 5 estrellas. Mediciòn mensual.

Meta mensual : ≥ 65%

$$RO\ (\%)\ =\ (\frac{Nùmero\ de\ reseñas\ positivas}{Nùmero\ de\ reseñas\ totales})\ *\ 100$$

5. MÉTRICAS Y ANÁLISIS

- **Densidad de mercado:** El número de hoteles identificados como posible competencia directa por kilómetro cuadrado a nivel de estado y/o a nivel de condado.

- **POV del usuario:** por medio de análisis de reseñas e información complementaria determinar que hace que un hotel sea relevante en Google y Yelp y que aspectos resaltan los usuarios más, tanto positivos como negativos.

6.METODOLOGÍA DE TRABAJO

Como marco de trabajo se sigue la metodología ágil en específico el scrum. Se ha establecido un plazo de 45 días (mes y medio) dividido en 3 ciclos de desarrollo cortos denominados sprints. Cada sprint tiene definido objetivos e hitos entregables en cada uno de ellos. Se detallan a continuación:

6.1. Sprint 1: puesta en marcha del proyecto y trabajo con datos

- Definición del problema, objetivos, alcance y fundamentación del proyecto
- Establecimiento de los KPI's
- ETL de los datos. (preliminar)
- EDA de los datos (preliminar)

- Repositorio en Github.
- Implementación stack tecnológico
- Metodología de trabajo
- Equipo de trabajo - Roles y responsabilidades
- Cronograma general - Gantt
- Análisis preliminar de calidad de datos

6.2. Sprint 2: data engineering

- ETL completo
- EDA completo
- Estructura de datos implementada (DW, DL, etc).
- Pipeline ETL automatizado
- Diseño del Modelo ER
- Pipelines para alimentar el DW
- Data Warehouse
- Automatización
- Validación de datos
- Documentación
- Diagrama ER detallado (tablas, PK, FK y tipo de dato)
- Diccionario de datos
- Workflow detallando tecnologías
- Análisis de datos de muestra
- MVP/ Proof of Concept de producto de ML ó MVP/ Proof of Concept de Dashboard

6.3. Sprint 3: data analytics + ML

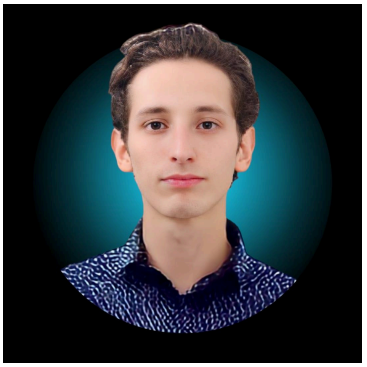
- Diseño de Reportes/Dashboards
- KPIs
- Modelos de ML
- Modelo de ML en producción
- Documentación

7.EQUIPO DE TRABAJO

Bidating pone a disposición un equipo multidisciplinario y experto para ofrecer soluciones a su medida.

Marco Data Analyst

Paulo Data Engineer



Kevin Data Engineer

Edgar Data Scientist

Vcitor Project Manager

Fernanda Data Analyst

- **Marco Caro. Data Analyst – Data Architect:**

La responsabilidad de Marco recae en el análisis, interpretación y presentación de datos para ayudar al cliente a tomar decisiones informadas. Además, como data architect se encarga de diseñar la arquitectura de datos y garantizar su integridad y eficiencia.

- **Paulo Lara. Data Engineer – Task Manager:**

Paulo es fundamental para garantizar que los datos se recolecten, almacenen, procesen y transformen de manera eficiente y efectiva, se asegurará de que los datos sean precisos, completos y confiables mediante la implementación de controles de calidad de datos, limpieza y normalización de datos. También como task manager se encarga de monitorear el progreso del equipo en relación a las tareas asignadas, elaborando informes diarios del estado de las asignaciones, los hitos alcanzados y cualquier problema o riesgo identificado.

- **Kevin Davison. Data Engineer – Data Strategist:**

El papel de Kevin es muy importante como diseñador y constructor de la infraestructura necesaria para almacenar y procesar grandes volúmenes de datos de manera eficiente. Esto incluye el diseño de base de datos, almacenes de datos, data lakes y pipelines de datos. Desde su rol como data strategist colabora con los líderes de la organización para

definir la estrategia global de datos, identificando objetivos, prioridades y áreas de enfoque para la gestión y utilización de datos.

- **Edgar Esteban. Data Scientist – Technology Manager:**

Edgar es responsable de convertir datos en insights accionables que impulsen el valor y la innovación en un proyecto de datos, utilizando una combinación de habilidades matemáticas y tecnológicas. En su papel como Technology Manager es el encargado de garantizar que las tecnologías y herramientas utilizadas en el proyecto sean adecuadas, eficientes y seguras, y que se utilicen de manera efectiva para lograr los objetivos del proyecto.

- **Víctor Orestes. Data Scientist- Project Manager:**

Víctor es el encargado de extraer conocimientos y generar valor a partir de los datos disponibles, desarrollar modelos predictivos utilizando técnicas de machine learning y análisis predictivo, para predecir eventos futuros o comportamientos basados en datos históricos. En su función como Project Manager lidera y coordina todas las actividades del proyecto, desde la planificación inicial hasta la implementación y el cierre, asegurando que se alcancen los objetivos de manera exitosa y dentro de los límites establecidos.

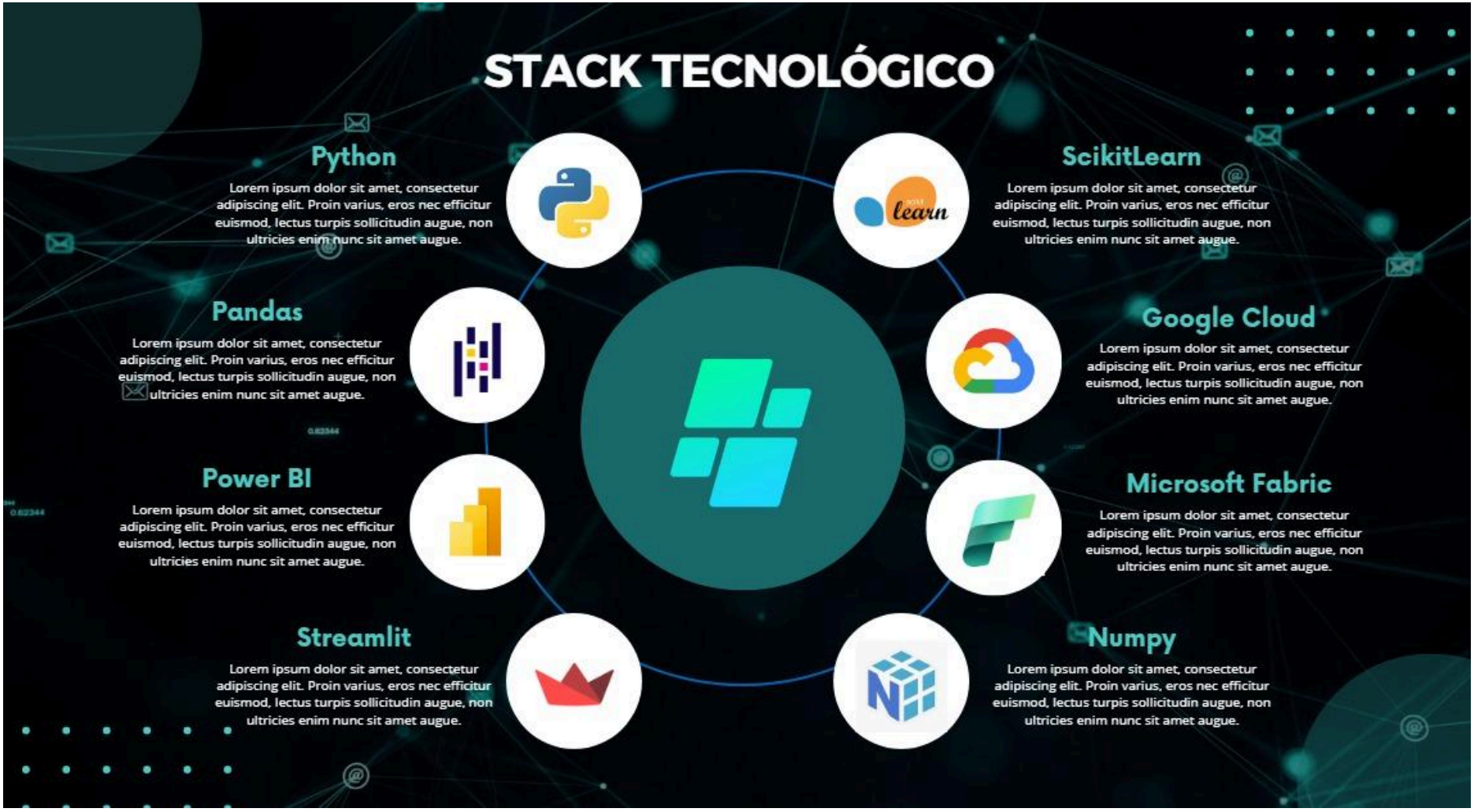
- **María Helguero. Data Analyst – Data Storyteller:**

La función de María como analista de datos es recopilar datos de diversas fuentes, como bases de datos, archivos, APIs o herramientas de terceros. Luego, limpiar y procesar los datos para eliminar valores atípicos, datos faltantes o errores. Utiliza herramientas y técnicas de análisis de datos para explorar y comprender la estructura. En su rol como data storyteller nos aportará en la conversión de los insights obtenidos a partir de los datos en historias comprensibles y persuasivas, utilizando técnicas de narración y visualización de datos.

8.STACK TECNOLÓGICO

El stack tecnológico empleado estará dentro de la plataforma integrada Microsoft Fabric, el cual ofrece toda la gama de herramientas para desarrollar el ciclo completo de nuestro proyecto. Facilita la ingesta y transformación de los datos, de forma que se podrá trabajar desde el data lakehouse que aloja los datos sin procesar, pasando por los entornos para escribir y ejecutar código hasta la generación de la base de datos con datos ya transformados e implementación de machine learning e inteligencia de negocios.

En la imágen se verán las herramientas, programas, lenguajes de programación y librerías contemplados para llevar a cabo cada una de las etapas de la ejecución del proyecto.


STACK TECNOLÓGICO
Python
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin varius, eros nec efficitur euismod, lectus turpis sollicitudin augue, non ultricies enim nunc sit amet augue.
ScikitLearn
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin varius, eros nec efficitur euismod, lectus turpis sollicitudin augue, non ultricies enim nunc sit amet augue.
Pandas
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin varius, eros nec efficitur euismod, lectus turpis sollicitudin augue, non ultricies enim nunc sit amet augue.
Google Cloud
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin varius, eros nec efficitur euismod, lectus turpis sollicitudin augue, non ultricies enim nunc sit amet augue.
Power BI
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin varius, eros nec efficitur euismod, lectus turpis sollicitudin augue, non ultricies enim nunc sit amet augue.
Microsoft Fabric
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin varius, eros nec efficitur euismod, lectus turpis sollicitudin augue, non ultricies enim nunc sit amet augue.
Streamlit
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin varius, eros nec efficitur euismod, lectus turpis sollicitudin augue, non ultricies enim nunc sit amet augue.
Numpy
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin varius, eros nec efficitur euismod, lectus turpis sollicitudin augue, non ultricies enim nunc sit amet augue.

Citaremos algunas de las tecnologías, para entrar en el contexto de su funcionamiento.

- **Microsoft Fabric:** Es el corazón del proyecto, pues las herramientas integradas permiten que Microsoft Fabric esté presente en 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 del proyecto.

- **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.

- **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. Dentro de ella trabajaremos con NumPy, pandas, Matplotlib, Seaborn, Scikit-Learn, Tensor Flow.

- **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.

- **Streamlit**: Es una biblioteca de python que permite crear aplicaciones web interactivas para el análisis de datos y la visualización de manera rápida y sencilla. Ofrece una amplia gama de widgets para la entrada de datos, gráficos interactivos y capacidades de visualización, lo que la convierte en una herramienta poderosa para la creación rápida de prototipos y la implementación de aplicaciones de análisis y ciencia de datos.

8.1. El papel de microsoft Fabric

Esta es una plataforma integrada en Azure basada en Apache Spark que permite la ejecución y administración de ingeniería y Ciencias de datos. Combina componentes de orígenes internos y código abierto proporcionado por los clientes una solución completa en una misma plataforma. Apache es una potente biblioteca informática distribuida de código abierto que permite realizar tareas de análisis y procesamiento de datos a gran escala y proporciona información, un alto rendimiento para la experiencia en la ingeniería y ciencia de datos.

El Data Lake es una capa de almacenamiento de código abierto que aporta transacciones y otras características de confiabilidad integrado a Fabric, la mejora, las capacidades de procesamiento garantizan la coherencia de los datos en múltiples operaciones simultáneas. Cuenta con paquetes por defecto de Java, Scala, Python o R, paquetes compatibles como ya sabemos con diversos lenguajes y entornos de programación cuya instalación y configuración automática lo hace muy sencillo de usar. Tiene compatibilidad con varias rutinas se basa en un sólido sistema operativo de código abierto el cual lo permite que utilizar y ser compatible con varias configuraciones de hardware y equipos de sistemas. Fabric ofrece una solución completa de experiencia de ingeniería y ciencia de datos con integraciones nativas de azur y power bi. Solución todo en 1 que abarca el movimiento de datos la ciencia de datos el análisis en tiempo real la inteligencia empresarial y ofrece un conjunto completo de servicios e incluyendo un lago de datos ingeniería e integración de datos todo en un solo lugar vamos a explicar un poco más en profundidad que es. Data Integration es una parte fundamental porque consolida los datos desde diversas fuentes en un único lugar es decir podemos transformar limpiar y enriquecer antes de cargarlo al lago de datos.

El funcionamiento el flujo se utiliza se realiza en realidad con Azure Data Factory se ejecuta en Apache Spark el cual facilita la extracción transformación y carga de datos desde diferentes orígenes. Data Factory es una plataforma de orquestación de datos que permite crear programas y administrar flujos de trabajo de datos. ¿Cómo funciona? define pipeline que mueven datos entre diferentes ubicaciones como base de datos almacenes de datos y servicios en la nube los flujos de trabajo se pueden programar y automatizar según las necesidades del negocio.

Synapse Data Engineering es una experiencia central que permite a los ingenieros de datos transformar los datos a gran escala utilizando Apache Spark. ¿Cómo funciona? Los ingenieros de datos pueden crear flujos de trabajo de transformación democratización de datos y construir una arquitectura robusta de Lake House este combina lo mejor del data Lake y el data warehouse facilitando la ingesta transformación y comparación de datos organizativos.

Synapse Data Science permite a los científicos de datos realizar análisis avanzado y construir modelos predictivos bueno acá utilizaremos Júpiter notebook Apache Spark para explorar datos crear modelo y obtener insights la integración con power bi y asus machine learning permiten llevar los resultados a la producción. Synapse Data Warehouse es el almacenamiento y consulta eficiente de grandes volúmenes de datos estructurados funciona con asus sinnaps analytics anterior conocido como assur SQL data warehouse es parte de fabric y proporciona un entorno escalable para

consultas analítica permite la creación de almacenes de datos para análisis empresarial.

Synapse Data Time Analytics lo que permite en este caso es brindar información en tiempo real. La integración con Power BI y Azure permite que el lago de datos pueda ser visualizado y analizado directamente en power bi además la integración con azure machine learning permite llevar modelo de datos a la producción.

Data Activation data activador es una función que lo que permite es que se active una alerta cuando en tiempo real vamos a consumir determinados datos este sistema nos va a ir actualizando sí en tiempo en forma remota en tiempo real y forma remota los nuevos datos que se vayan cargando. Es importante tener en cuenta que nosotros cuando trabajemos con olela like es como se como si fuera nuestro onedrive de nuestra computadora en una misma nube vamos a tener cargados todos los datos de tal manera que van a estar disponibles para trabajar dentro de la organización desde cualquier lugar

Por otro lado sí en la orquestación de datos desde el Lake housing creo relaciones de tabla un diagrama de la que hausen podremos ver cómo se pueden relacionar las diferentes tablas o los dataset y como dije anteriormente One like guardaremos todo en un repositorio como si fuera onedrive de datos. Podríamos decir que este software es un diseño preliminar y detallado donde se codifique y depura, se realizan testing y pruebas previas para la generación y mantenimiento en la ingeniería de datos en lo que hace a científico de datos y analista de negocios como a los usuarios de negocios es decir en una misma plataforma pueden trabajar todas las partes consumir la información que necesitan los analistas de negocios y nosotros.

Ahora vamos a hablar de nuestro trabajo en específico en esta plataforma. Vamos a instalar la librería missigno en Python, cuya utilidad para visualizar datos faltantes o nulos en conjunto de datos el cual nos sirve para proporcionar datos y visualizaciones que permiten identificar fácilmente patrones de datos faltantes en un conjunto de datos a lo que puede ser crucial en el proceso de limpieza y preparación de datos antes del análisis.

A lo que hace la importación de librerías vamos a importar: os: el cual proporciona funciones para interactuar con el sistema operativo como manipular archivos y directorios. pandas: esta librería es muy poderosa porque permite realizar análisis en Python proporciona estructura de datos flexibles y herramientas para trabajar con conjunto de datos de manera eficiente. numpy: es una librería que realiza computaciones numéricas en Python ofrece un soporte para matrices multidimensionales y funciones matemáticas de alto rendimiento. pyarrow: esta librería procesa y almacena

datos en formato columnar es eficiente para trabajar con grandes volúmenes de datos especialmente en entornos distribuidos. matplotlib:

Esta librería permite la visualización de datos y ofrece una amplia variedad de funciones para crear gráficos de alta calidad y personalización. warnings: bueno acá lo que vamos a hacer en nuestro libro de Python es manejar las advertencias de Python controlaremos cómo se manejan y muestran las advertencias durante la ejecución del código el cual lo configuraremos para que ignore las advertencias. pickle: con esta librería lo que haremos es serializar y de serializar objetos de pyc es decir la serialización sería convertir si un objeto en una secuencia de bytes mientras que la DC serialización es el proceso inverso es decir convertir una secuencia de bytes de vuelta a un objeto Python en memoria por qué la vamos a usar porque esto es útil para guardar objetos complejos en archivos o enviarlos a través de una red por ejemplo esta librería es especialmente útil cuando se trabaja con objeto que no se pueden guardar fácilmente en un archivo de texto como instancias de clases personalizadas listas diccionarios entre otros.

9.Exploratory Data Analysis

En la etapa de transformación de los datos se abordarán los siguientes aspectos básicos para realizar el tratamiento previo y limpieza de datos para poder llevar a cabo la carga de datos al data wharehouse

- Verificación De tipo de dato de columna
- Dimensionalidad de los datos
- Valores nulos y en creo en caso de las columnas numéricas
- Verificación gráfica de outliers
- Indagación de consistencia de los datos, máximos, mínimos.

10.Datos Google reviews y Yelp: ETL / EDA

Inicialmente se tuvieron diferentes datos provenientes de dos plataformas de reseñas de negocios, google reviews y Yelp, estos datos vienen en diferentes formatos, y archivos. En las imágenes a continuación veremos los esquemas de datos y los formatos de las tablas.

10.1. Informe ETL / EDA Yelp Reviews

Las reseñas de Yelp vienen separadas en 5 archivos cada uno con formato e informaciòn diferente pero con columnas relacionales entre sì, se describirà lo observado en cada uno de estos archivos y asimismo se enumeraràn las acciones tomadas para proceder.

10.1.1. **Review:**

De este dataframe podemos observar que respeta el formato y estructura asignado a cada tipo de dato, por lo que el tratamiento o normalizacion en un principio no se ve necesario.En cuanto a datos nulos, no se observan como tampoco registros duplicados.

1. división de la columna "date" en dos columnas: "fecha" y "hora".
2. realizar análisis de sentimiento de la columna "text", creando una nueva columna llamada "ANALISIS_SENTIMIENTO".
3. eliminación de las columnas "useful", "funy", "cool" y "text", porque son innecesarias, no aportan información relevante para el propósito del proyecto, nos basta con mantener la columna "star" y el análisis de sentimiento realizado sobre la columna text que finalmente se volcó en la columna "ANALISIS_SENTIMIENTO".

10.1.2. User:

Podemos observar que los datos en las columnas se comportan adecuadamente tanto en su tipo como en la sintaxis
presente en ellos. No se observan casos en los que se deban formatear o normalizar en principio, siempre teniendo en cuenta
que esto esta condicionado al alcance, desarrollo y rumbo que llevemos en el ETL y tratamiento de estos datos
Como observacion adicional, las columnas compliment se comportan como valoracion de sentimiento, por lo que el tratamiento en conjunto de ellas no se descarta en un futuro. Esta columna no contiene datos nulos. Esta columna no contiene registros duplicados.

1. eliminación de las columnas *"yelping_since", "useful", "funny", "cool", "elite", "fans", "average_stars", "compliment_hot", "compliment_more", "compliment_profile", "compliment_cute", "compliment_list", "compliment_note", "compliment_plain", "compliment_cool", "compliment_funny", "compliment_writer", "compliment_photos"* porque algunas de ellas se vinculan con columnas que fueron eliminadas del data set "REVIEW", las otras resultan innecesarias ya que no aportan información de relevancia para el proyecto.
2. eliminación de duplicados borrar duplicados de columna user_id.

10.1.3. Tip:

La columna tips no presenta mayores problemas en cuanto al tipo de dato en cada columna ni en su estructura o disposición, ya que conservan su integridad y su sintaxis. En el desarrollo del ETL podremos valorar si es el formato adecuado o si son necesarios cambios o normalizaciones. No se observan valores nulos.
No se observan registros duplicados en esta columna.

1. división de la columna "date" en dos columnas: "fecha" y "hora".
2. realizar análisis de sentimiento de la columna "text", creando una nueva columna llamada *"ANALISIS_SENTIMIENTO".*
3. eliminación de columnas "text", *"compliment_count"*, porque basta con quedarnos con la columna *"ANALISIS_SENTIMIENTO"* y la segunda es innecesaria si de la tabla *"USER"* se eliminaron todas las columnas que hacían referencia a "compliment".

10.1.4. Business:

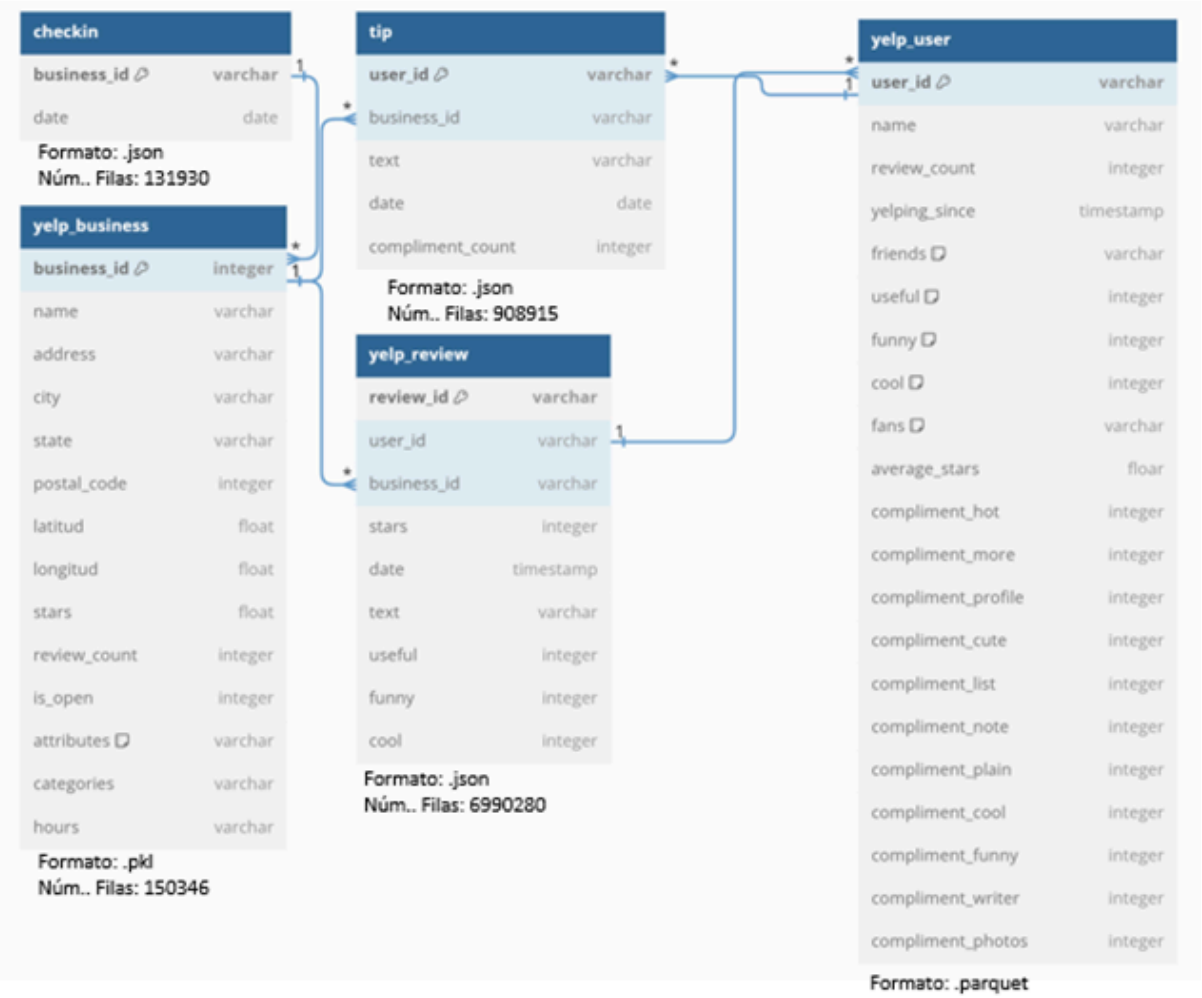
Este dataframe cuenta con varias columnas con datos anidados (por ejemplo 'attributes' o 'hours'), con datos en formato string pero con estructura de lista o tupla (como la columna 'categories'), entre otros. Esto nos dificulta el manejo y analisis preliminar de la calidad de estos

datos. Podemos señalar también que las columnas 'attributes' y 'hours' cuentan con una cantidad moderada de datos nulos,
que debemos consensuar si condicionan su valor en el análisis. No se observan registros duplicados

10.1.5. Checkin:

Este dataframe no presenta registros duplicados, aunque si debemos remarcar que los datos de la columna date se encuentran en un formato de cadena (string) pero contienen la estructura de dato datetime, por lo que le tratamiento de los datos de esta columna debe ser enfocado respetando esta estructura. No se observan registros duplicados ni datos nulos.

10.2. Esquema inicial de datos Yelp



10.3. Informe ETL / EDA Google

De Google existen gran cantidad de archivos pero pertenecientes a los mismos datasets, hay 2 grandes datasets que son “reviews” y “metadata”, el principal reto de esto fue unir esa gran cantidad de archivos en 2 datasets independientes ya que había archivos separados.

10.3.1. Metadata:

Se presentan en las columnas 'hours','MISC' y 'state' tipos de datos string que tienen sintaxis de estructuras de datos, como por ejemplo array, que debemos manejar de manera de respetar la estructura y segregar o reconstruir los datos individuales y/o necesarios para nuestro análisis,

Las columnas 'price' y 'description' cuentan con más del 90% de datos nulos, lo que imposibilita realizar un análisis con esos datos, no asi en el caso de las columnas 'hours','MISC' y 'state' que a pesar de contar

con mas del 20%/25% de datos nulos, debemos establecer un criterio de utilizacion de sus datos ya que no amerita su eliminacion por falta de ellos. Se observan pocos registros duplicados.

Acciones tomadas:

1. se eliminaron columnas que no eran relevantes o no sumaban valor para el estudio correspondiente, estas fueron: 'price' con un 90% de datos nulos, 'descrption' con una presencia de nulos de casi el 92%, 'state' solo nos dice hora de cierre y apertura, 'misc' es un diccionario que posee datos de servicios opcionales o accesibilidad al establecimiento, 'relative_results' nos da información de otros lugares parecidos al establecimiento calificado y por último 'hours' nos indica los horarios en que se maneja dicho negocio, pero falta información ya que no nos dice los días completos de la semana.

2. se creó una nueva columna partiendo de 'category', y se asignó por las palabras claves usadas en el dataset "BUSINESS" las categorías de forma resumida y se por medio de la palabra 'otro', se desestimaron aquellas categorías que no serán relevantes para el estudio, a esta nueva columna se le llamó 'category summary'.

10.3.2. Reviews:

Las columnas 'text', 'pics' y 'resp' cuentan con un porcentaje pronunciado de datos nulos, principalmente las dos últimas, lo que las hace inutilizables para nuestro análisis.

Con respecto a la columna 'text' debemos establecer el criterio de utilización y manejo de valores faltantes, en caso de una cantidad grande de estos.

También debemos establecer el valor que le daremos a la data incluida en esta última columna para nuestro análisis posterior, y si vale la pena tenerla.

Las columnas restantes del dataframe no presentan mayores complicaciones ni contaminación o data anidada. Presenta un bajo nivel de registros duplicados.

1.se seleccionaron tres columnas que no aplicaban para nuestro análisis, dichas columnas son: 'resp' y 'pics' ambas poseían una gran cantidad de nulos, inclusive 'resp' poseía información que podíamos encontrarla en otra columna, la columna 'time' no poseía un valor significativo para nuestro estudio.

2. en razon del grán cumulo de datos, y teniendo en cuenta las dificultades que se presentaron al momento de hacer análisis de

sentimientos con tablas más reducidas, analizamos la columna ‘estado’ para luego concluir cuáles tomar para la nueva cadena hotelera de nuestro cliente, nos enfocamos en aquellos que se situaban cercano a la coste este y oeste de estados unidos, seleccionando a: 'California', 'Florida', 'Nevada', 'NewYork', 'Massachusetts',
'Mississippi','Oregon','illinois','Louisiana','Alabama','Texas','Washington','Georgia','NorthCarolina','SouthCarolina','Virginia','Maryland','Delaware','New jersey', así fue como se procedió a filtrar nuestra data mediante los ya mencionados.

3. sobre la columna ’text’, aquella que nos suministra información sobre las recomendaciones dadas, verificamos la presencia nulos los cuales se sustituyeron por (N/A), para así proceder a instalar e importar todos los paquetes necesarios para aplicar ‘Análisis de Sentimientos’ a dicha columna.

Nota aclaratoria:

Para realizar el procedimiento de análisis de sentimiento se utilizó la biblioteca "NLTK" y su ejecución se llevó a cabo en modo local **(Fernanda)** y por medio de un notebook en el entorno de fabric **(Paulo)**, tarea que se desdobló en razón que dentro de la nube, insumió un término mayor a 4 horas, situación que me llevó a buscar formas de optimizar la función, luego de probar distintas alternativas de optimización, terminé ejecutando la tarea con un código optimizado con el uso de la librería de "joblib" y la importación de la función "Parallel", que me permitió ejecutar la acción con la implementación de 10 núcleos de CPU.

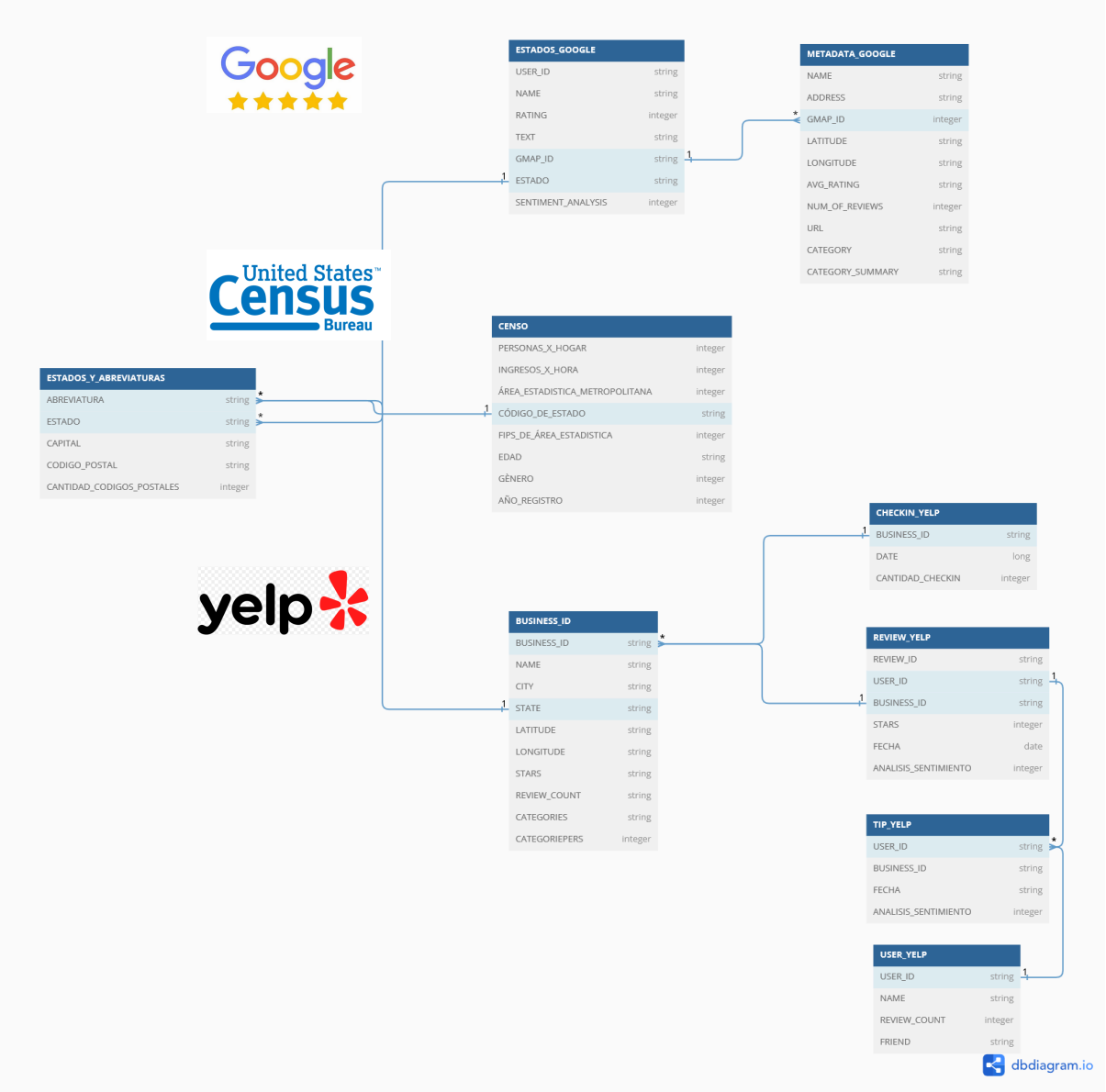
Por otra parte, las reseñas fueron clasificadas del siguiente modo:

0 = Malo
2 = Positivo
1 = Neutral

Cabe aclarar que, este procesamiento se realizó de modo previo a la realización de todo el proceso de transformación de datos dentro de los respectivos dataflows.

10.4. Esquema inicial de datos Google

10.5. Diagrama ER luego de procesamiento de datos ETL



10.6. Conclusiones ETL / EDA

Los datos obtenidos de las fuentes anteriormente citadas (Google Maps y Yelp!), presentan congruencia e integridad general, señalando su coherencia en cuanto a formato de organización y tipo y estructura presente en cada uno de los ficheros y dentro de estos en cada una de las features y columnas,nos facilita y habilita el manejo y posterior carga en una base de datos a construir, para el desarrollo de nuestro proyecto.

En un análisis más específico, la correlación entre los datos de las distintas fuentes es alta, lo que favorece una posible conexión y unificación coherente que construya la relación necesaria para una manipulación y análisis exitoso y óptimo

Señalamos que se cuenta con una baja proporción de datos nulos en general, y los encontrados no interfieren con el valor de la data extraída o no son relevantes en el análisis para nuestro proyecto

También es casi inexistente la presencia de registros duplicados y la estructura y tipo de dato en la cual se encuentran registrados los datos, no dificulta en mayor medida su manejo (solo en casos excepcionales y particulares en algunas columnas de algunos dataframes), y en todo caso, la presencia de datos anidados, por ejemplo, no conlleva una gran dificultad de normalización, ya que se respeta la sintaxis para su manejo adecuado y su posterior desempaquetamiento (de ser necesario).

En conclusión, vemos que los datos a nuestra disposición son utilizables y manejables, no presentan grandes obstáculos y nos brindaran valor a nuestro análisis.
Como observación final, el que contemos con datos óptimos, no quita el hecho de que para un análisis más completo y profundo en el desarrollo de nuestro proyecto debamos obtener nueva data complementaria o data de nuevas fuentes que nos brinde una perspectiva más amplia y agregue mayorvalor a nuestro análisis.

10.7. Datos de la oficina del censo de los Estados Unidos

A través de la API del Bureau de Censo de Estados Unidos, se recopilaron datos de encuestas poblacionales desde el año 2010 hasta el 2023 con el objetivo de obtener información sobre las áreas de las ciudades y los ingresos promedio por ciudad. Estos datos se agruparon por condado, lo que dio como resultado una lista que abarcaba todos los registros de rangos de ingresos. Posteriormente, se calcularon los promedios de ingresos por condado.

Además, se accedió a la API para facilitar la asignación de identificadores únicos a los registros y así poder mapear la información con cadenas de texto que identificaran los datos de manera más comprensible. Este proceso ayudó a mejorar la claridad y utilidad de la información recopilada, permitiendo un análisis más efectivo de las tendencias y diferencias en los ingresos a nivel de condado en los Estados Unidos.

Para más detalle referirse al archivo trabajado ETL_CENSO y la transformación de estos datos. El dataset obtenido tiene 1’712307 filas y 8 columnas a partir del año 2011.

10.8. Trabajo con Web Scraping con el Entorno: Microsoft Azure

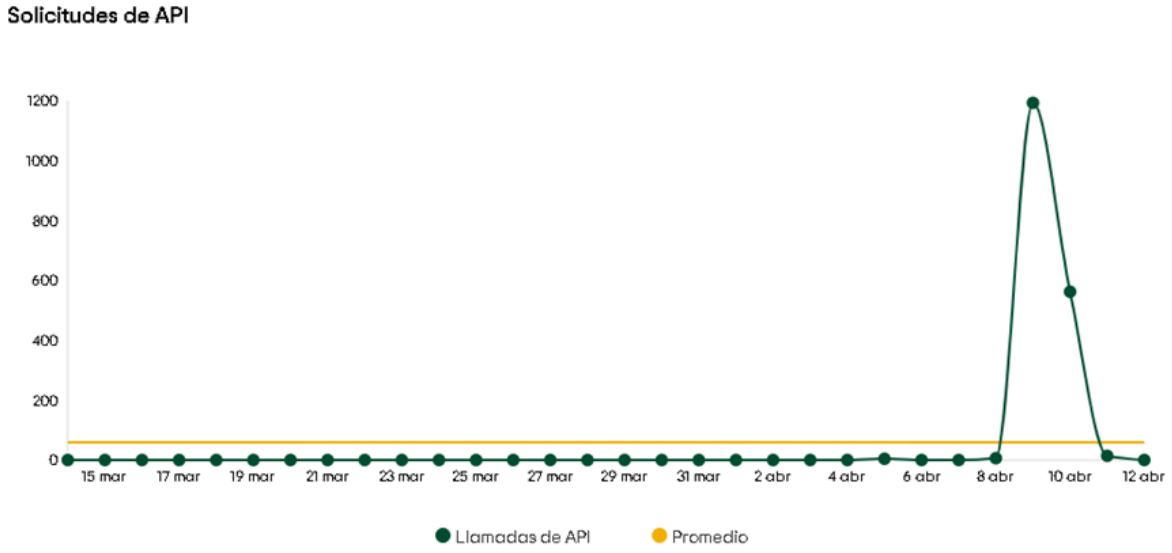
Para este trabajo, se utilizará la herramienta de Microsoft Fabric: Synapse Data Science en Azure. Se ingresará el código o se codeará en el flujo de trabajo de ciencia de datos en Synapse Data Science. Para ello, se utilizará la plataforma integrada Azure Synapse Analytics donde se ha creado la cuenta de almacenamiento: Microsoft.Azure.SynapseAnalytics-20240403150313.

Posteriormente, se ingresará a la pestaña Desarrollar y a Synapse Studio para terminar de configurar los pasos indicados. Posteriormente, se accederá a: https://web.azuresynapse.net para ingresar al área de trabajo, el cual se llama: PF-HENRY, donde se creará un nuevo notebook para comenzar a trabajar.

En el notebook se creará una lista con el Excel que se había generado a partir de todas las sucursales de la página oficial de Wyndham y, a su vez, este listado se vincularán todos los "locationId" desde la empresa Tridadvisor.

Con la clave API generada de TripAdvisor Content API se generarán las consultas viendo el gráfico del consumo realizado en el siguiente gráfico:

API	Solicitudes	Errores
Detalles de la ubicación	2	0
Opiniones de la ubicación	1777	1334



10.8.1. Requerimientos de la API de Tripadvisor

TripAdvisor establece límites de llamadas diarias y por segundo en cada clave utilizada para acceder a la API de contenido. El límite de velocidad se mide en una ventana móvil de 24 horas. La ventana comienza con la primera llamada realizada y se restablece 24 horas después. No hay una hora fija del día en la que la ventana se reinicie. Si la ventana se restablece y no se realiza ninguna llamada para una clave determinada, no se abre una nueva ventana hasta que se realice la siguiente llamada para una clave determinada. Las llamadas por segundo funcionan de manera similar.

Límite de llamadas diarias para detalles de ubicación, fotos y reseñas API: determinado por tu presupuesto gratuito de hasta 5000 consultas de locationId cuyo **Límite de llamadas diarias para las API de búsqueda: 10000 Consultas por segundo (QPS): 50**

El código que tiene el sistema es:

python

Copy code

```
import

"https://api.content.tripadvisor.com/api/v1/location/nearby_search?langu
age=en"

          "accept"  "application/json"



print
```

El código generado en Data Factory Lakehouse Notebook TripAdvisor con lenguaje PySpark (Python) tuvo que hacer varios ensayos para adaptar el consumo de la misma de forma masiva y no individualmente.

Una vez ejecutado, se pudo ver el resultado del DataFrame finalizado en Data Wrangler sin salir del entorno donde se trabajó y se pudo verificar satisfactoriamente que se pudieron consumir los "locationId".

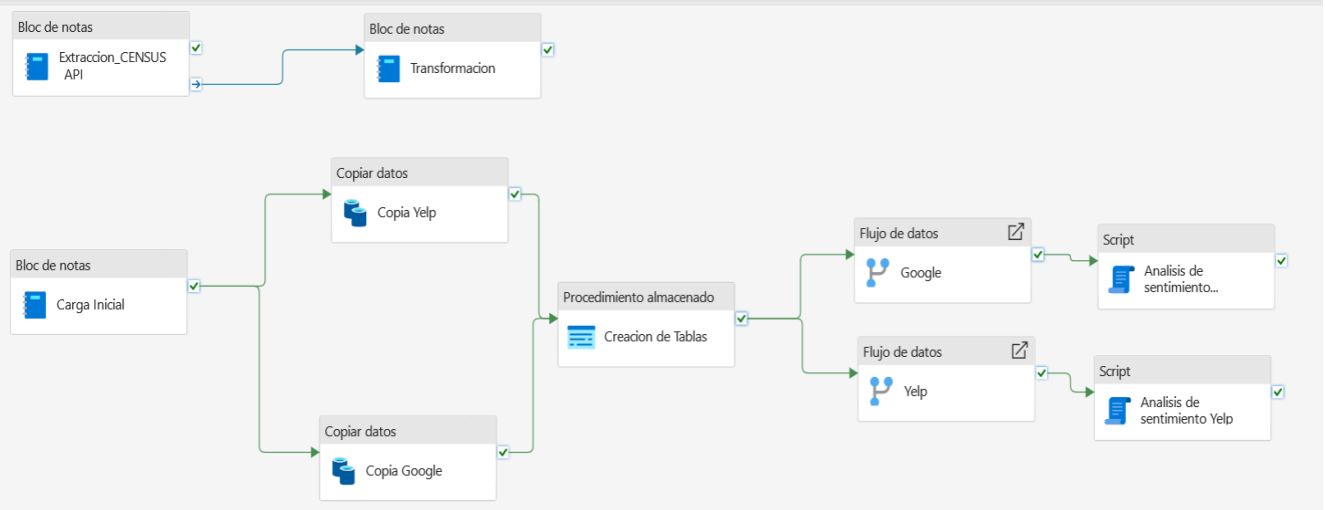
Verificado, se procedió a guardar en el mismo entorno de Microsoft Fabric el archivo generado en formato Parquet.

11.Pipeline y ciclo de vida del dato

Para realizar la secuencia de datos y procesos requeridos para lograr tener los datos para ser analizados y presentados (pipeline), se eligió Data Factory de Microsoft Fabric.

En el marco del presente proyecto de análisis de datos para asesorar a nuestro cliente en su plan de expansión dentro del entorno de Microsoft Fabric, la implementación de un pipeline en Data Factory ofrece una serie de ventajas significativas. Data Factory proporciona una plataforma robusta y escalable para orquestar y automatizar el flujo de datos desde diversas fuentes hacia un repositorio centralizado, permitiendo un procesamiento eficiente y en tiempo real. Además, su integración nativa con otras herramientas y servicios de Microsoft, como Azure Blob Storage, Azure SQL Database y Power BI, simplifica la creación de flujos de trabajo complejos y la generación de insights relevantes para la toma de decisiones estratégicas. Con características de monitorización y gestión avanzadas, Data Factory garantiza la fiabilidad y la calidad de los datos, optimizando así el rendimiento y la eficiencia de todo el proceso de análisis.

A continuación se presenta el esquema establecido del Pipeline:



Dentro del Pipeline se ejecutan los procesos paralelos de ETL y análisis de los datos del CENSO descargados y los datos de las reseñas de Google Yelp. Como se puede observar tenemos diferentes elementos dentro de dicho pipeline.

11.1. Carga Inicial

Notebook que realiza la unificación y transformación de formato de los múltiples archivos que vienen de Google y de Yelp.

11.2. Procedimiento de Copia de archivos

Se realiza una copia de estos archivos cargados inicialmente para tener una copia de seguridad siempre ya que esta carga inicial puede tardar y consume recursos por lo que es recomendable siempre tener una data lista para ser transformada original y trabajar sobre la copia.

11.3. Procedimiento guardado

Procedimiento para realizar la creación de tablas que serán consumidas por los dataflow de Yelp y de Google.

11.4. Dataflows o flujo de datos

El papel de un dataflow en un proceso de ETL en Microsoft Fabric es proporcionar una representación visual de cómo los datos son extraídos, transformados y cargados a través de una serie de operaciones lógicas. Los dataflows en Microsoft Fabric permiten a los usuarios definir estas operaciones utilizando una interfaz gráfica intuitiva, lo que facilita el diseño y la implementación de flujos de datos complejos.

11.5. Scripts de análisis de datos

Notebooks que procesan los datos luego del ETL para obtener insigths y contribuir al desarrollo del objetivo principal.

12. Diccionario de datos final luego de ETL

12.1. Procedimiento guardado

Contiene las reseñas completas, incluyendo el user_id que escribió el review y el business_id por el cual se escribe la reseña

NOMBRE DE COLUMNA	TIPO DE DATO	CONTENIDO

| REVIEW_ID | String | "zdSx_SD6obEhz9VrW9uAWA" |

| USER_ID | String | "Ha3iJu77CxlrFm-vQRs_8g" |

| BUSINESS_ID | String | "tnhfDv5Il8EaGSXZGiuQGg" |

| STARS | Entero | 4 |

| FECHA | Date | "2016-03-09" |

| HORA | Date | 00:00 |

|ANALISIS_SENTIMIENTO | Entero | 1 |

12.2. User Yelp

Data del usuario incluyendo referencias a otros usuarios amigos y a toda la metadata asociada al usuario.

| NOMBRE DE COLUMNA | TIPO DE DATO | CONTENIDO |

|---|---|---|

| USER_ID | String | "Ha3iJu77CxlrFm-vQRs_8g" |

| NAME | String | "Sebastien" |

| REVIEW COUNT | Entero | 56 |

| FRIENDS | Lista | ["wqoXYLWmpkEH0YvTmHBsJQ", "KUXLLiJGrjtSsapmxmpvTA", "6e9rJKQC3n0RSKyHLViL-Q"] |

12.3. User Yelp

Tips (consejos) escritos por el usuario. Los tips son más cortas que las reseñas y tienden a dar sugerencias rápidas.

NOMBRE DE COLUMNA	TIPO DE DATO	CONTENIDO
USER_ID	String	"49JhAJh8vSQ-vM4Aourl0g"
BUSINES_ID	String	"tnhfDv5Il8EaGSXZGiuQGg"
FECHA	String	"2013-09-20"
ANALISIS_SENTIMIENTO	Entero	0

12.4. Business Yelp

Contiene información del comercio, incluyendo localización, atributos y categorías.

NOMBRE DE COLUMNA	TIPO DE DATO	CONTENIDO
BUSINESS_ID	String	"tnhfDv5Il8EaGSXZGiuQGg"
NAME	String	"Garaje"
CITY	String	"San Francisco"
STATE	String	"CA"
LATITUDE	Float	37.7817529521
LONGITUDE	Float	-122.39612197
STARS	Float	4.5

| REVIEW_COUNT | Entero | 1198 |

| CATEGORIE | Lista | ["Mexican", "Burgers", "..."] |

| CATEGORIEPERS | Objeto | Hotel |

12.5. Checkin Yelp

Contiene información del comercio, incluyendo localización, atributos y categorías.

Registros en el negocio.

| NOMBRE DE COLUMNA | TIPO DE DATO | CONTENIDO |

|---|---|---|

| BUSINESS_ID | String | "tnhfDv5Il8EaGSXZGiuQGg" |

| DATE | String | "2016-04-26 19:49:16, …" |

| CANTIDAD_CHECKIN | Entero | 3 |

12.6. Metadata Google

En el archivo se dispone la metadata referente a información del comercio, incluyendo localización, atributos y categorías.

| NOMBRE DE COLUMNA | TIPO DE DATO | CONTENIDO |

|---|---|---|

| NAME | String | 'Walgreens Pharmacy' |

| ADRESS | String | 'Walgreens Pharmacy, 124 E North St, Kendallville, IN 46755' |

| GMAP_ID | String | '0x881614ce7c13acbb:0x5c7b18bbf6ec4f7e' |

| LATITUDE | Decimal | 41.451859999999996 |

| LONGITUDE | Decimal | -85.2666757 |

| CATEGORY | String | ['Pharmacy'] |

| AVG_RATING | Decimal | 4.2 |

| NUM_OF_REVIEWS | Entero | 5 |

| URL | String | 'https://www.google.com/maps/place//data=!4m2!3m1!1s0x881614ce7c13acb b:0x5c7b18bbf6ec4f7e?authuser=-1&hl=en&gl=us' |

| CATEGORY SUMMARY | String | Health |

12.7. Estados Google

El archivo donde se disponibiliza las reviews de los usuarios de los estados de USA. Se conforma de la siguiente manera

| NOMBRE DE COLUMNA | TIPO DE DATO | CONTENIDO |

|---|---|---|

| USER_ID | Entero/Decimal | '101463350189962023774' |

| NAME | String | 'Jordan Adams' |

| RATING | Entero | 5 |

| TEXT | String | 'Cool place, great people, awesome dentist!' |

| GMAP_ID | String | '0x87ec2394c2cd9d2d:0xd1119cfbee0da6f3' |

| ESTADO | String | 'arizona' |

| SENTIMENT_ANALYSIS | Entero | '2' |

12.8. Estados y Abreviaturas

El archivo contiene información de identificación de los Estados que integran los Estados Unidos.

NOMBRE DE COLUMNA	TIPO DE DATO	CONTENIDO
ABREVIATURA	String	'AL'
ESTADO	String	'Alabama'
CAPITAL	String	‘Montgomery’
CÓDIGO POSTAL	Entero	'35004 to 36975'
CANTIDAD CODIGO POSTALES	Entero	'2655'

12.9. CENSO

Contiene datos demográficos y estadísticas relacionadas con la población de Estados Unidos.

NOMBRE DE COLUMNA	TIPO DE DATO	CONTENIDO
PERSONAS_X_HOGAR	Entero	'1'
INGRESOS_X_HORA	Entero	'80'
ÁREA_ESTADISTICA_METROPOLITANA	Entero	‘0’
CÓDIGO_DE_ESTADO	String	'NH'
FIPS_DE_ÁREA_ESTADISTICA	Entero	'715'
EDAD	Entero	'53'
GÉNERO	Entero	'1'
AÑO_REGISTRO	Entero	'2011'

12.10. Trip Advisor

1. se eliminó la columna 'Helpful Votes' por ser irrelevante para el proyecto.

TRIPADVISOR.

Contiene las reseñas completas del sitio web
https://www.tripadvisor.com.ar/developers

NOMBRE DE COLUMNA	TIPO DE DATO	CONTENIDO
ID	String	"945868640"
LOCATION_ID	String	"7184808"
PUBLISH_DATE	String	"2024-04-08T14:18:25Z"
RATING	Entero	4
TEXT	String	"It is a wetlands, and it does have birds for an overall experience. I would much rather drive out to Blackpoint Or Viera wetlands."
TITLE	String	'Worth the visit if in the area'
TRIP TIPE	String	'Family'
TRAVEL_DATE	Date	2023-08-31
USERNAME	String	'tinkerbsb'
ANALISIS_SENTIMIENTO	Entero	1