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

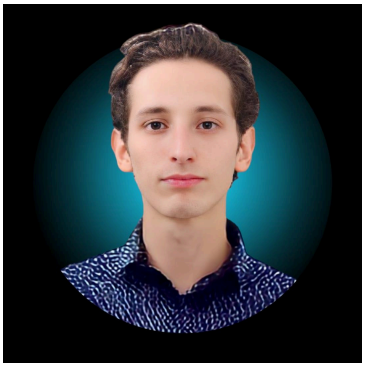
- Diseño de Reportes/Dashboards
- KPIs
- Modelos de ML
- Modelo de ML en producción
- Documentación

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.

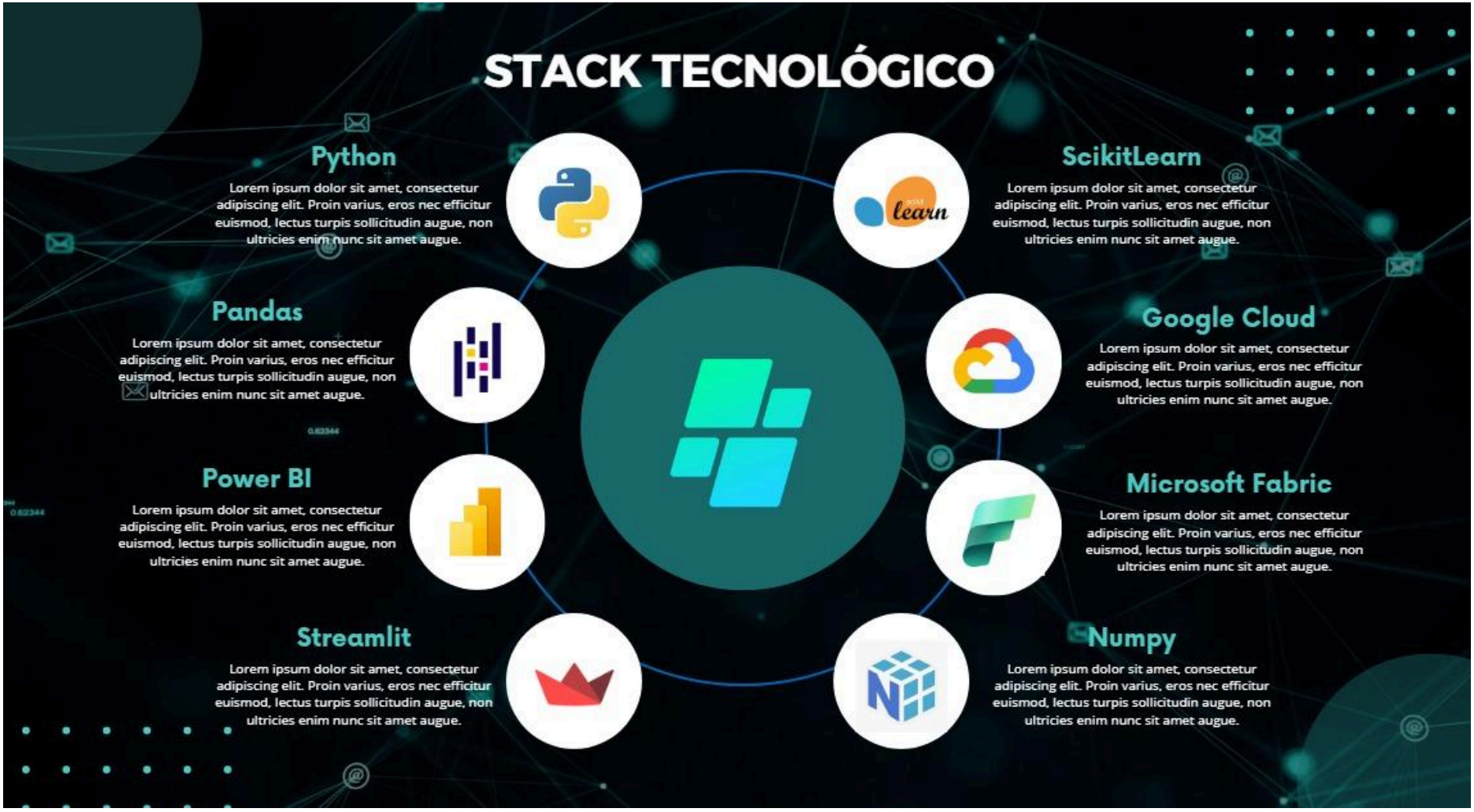
- **María Helguero. Data Analyst – Data Storyteller:**

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.

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.

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.

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.

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.

10.1.4. 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.

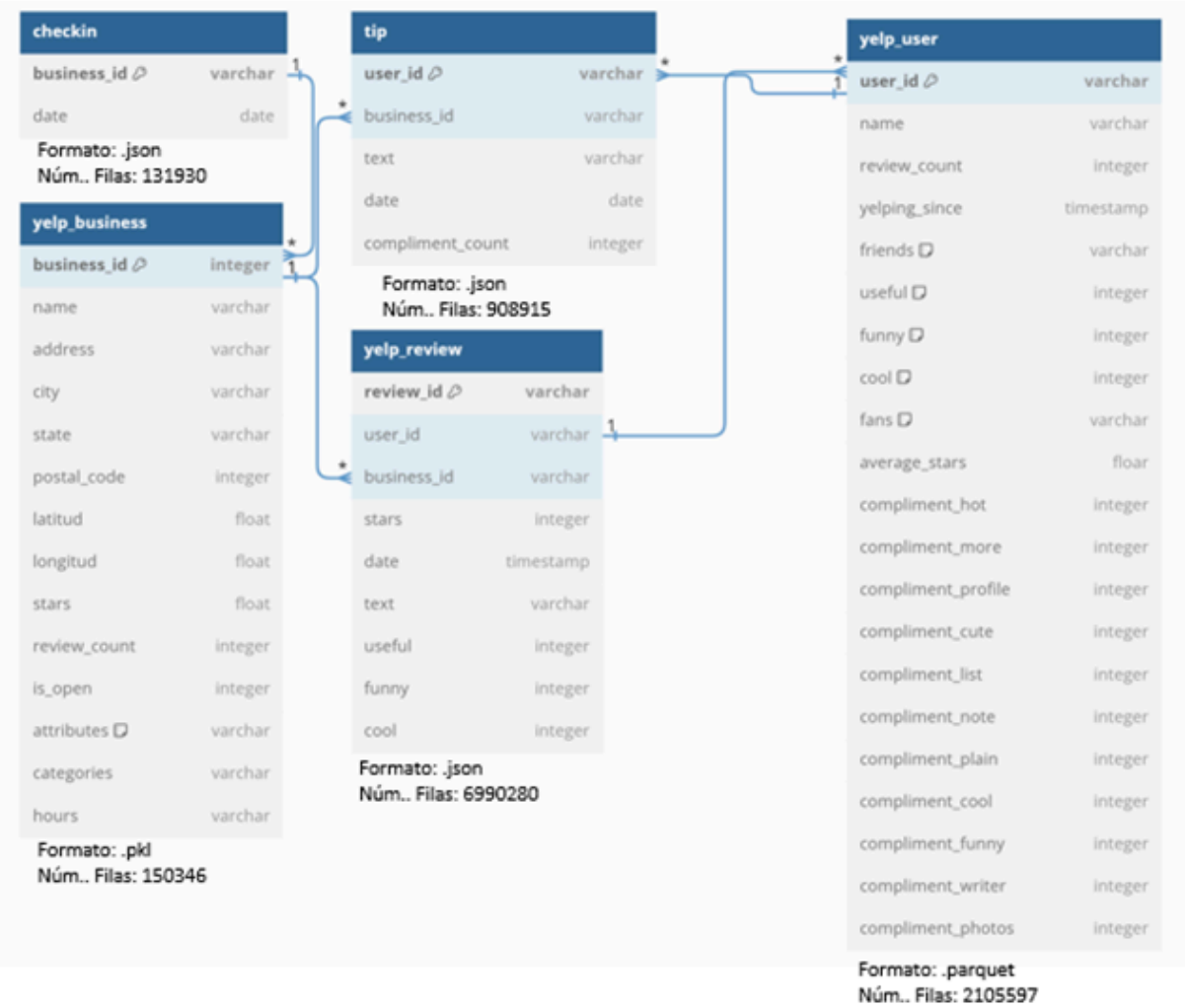
10.1.5. Business:

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

10.1.6. 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 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.

10.3.2. Reviews:

Las columnas 'text', 'pics' y 'resp' cuentan con un porcentaje pronunciado de datos nulos, principalmente las dos ultimas, lo que las hace inutilizables para nuestro analisis.
Con respecto a la columna 'text' debemos establecer el criterio de utilizacion y manejo de valores faltantes, en caso de una cantidad grande de estos.
Tambien debemos establecer el valor que le daremos a la data incluida en esta ultima columna para nuestro analisis 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.

10.4. Esquema de datos Google

10.5. Conclusiones ETL / EDA preliminar

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.