DATUM TECH

Data Engineering

PUNTOS A ENTREGAR

- Diagrama ER detallado (tablas, PK, FK y tipo de dato)
- Diccionario de datos (Va aparte)
- Workflow detallando tecnologías

ENTENDIENDO LOS DATOS

Para realizar el diseño del modelo entidad relación, se realizó un análisis exhaustivo del diccionario de los datos empleados para el proyecto. Se identificaron las variables relevantes para el objetivo del proyecto y se definieron sus atributos, tipos de datos y relaciones. A continuación se presentan las variables seleccionadas y sus características.

1. Tabla 1: Sitios_google

Nombre	Descripción
Name_sitio	Atributo que indica el nombre del comercio o establecimiento que se quiere representar como una entidad en el modelo.
Address	Atributo que indica la dirección física del comercio o establecimiento.
Gmap_id	Atributo que indica la identificación única que Google Maps le asigna a la dirección física del comercio o establecimiento.
Latitude	Coordenada Latitud.
Longitude	Coordenada Longitud.
Category	Atributo que indica la categoría del comercio o establecimiento.
Avg_rating	Valoración promedio.

Num_of_review	Número de reseñas que aparecen sobre el comercio o establecimiento, que indican las opiniones o comentarios de los clientes o usuarios que han interactuado con ella.
State	Atributo del nombre del estado o región en donde se encuentra el comercio o establecimiento
Url	URL de Google Maps que direcciona a la zona geográfica donde se localiza dicho comercio o establecimiento.

Fuente: Elaboración propia

La tabla 1 en el modelo entidad-relación es de gran importancia, ya que permite organizar y gestionar de manera estructurada la información de los comercios o establecimientos. A través de sus atributos, como el nombre, la dirección, la categoría, las valoraciones y las reseñas, se facilita la búsqueda, el filtrado y la representación geográfica de los comercios. Además, proporciona datos clave para la toma de decisiones, la evaluación de la calidad y la integración con otras funcionalidades del sistema, mejorando la experiencia global del usuario.

2. Tabla 2: Reviews_google

Nombre	**Descripción**
review_id	Atributo de identificación única.
User_id	Atributo de identificación única del usuario en la plataforma.
Rating	Valoración del usuario sobre el comercio o establecimiento.
Texts	Reseña del usuario sobre el comercio o establecimiento.
Gmap_id	Atributo que indica la identificación única que Google Maps le asigna a la dirección física del comercio o establecimiento.

Fuente: Elaboración propia

La tabla 2 es importante porque representa la información clave relacionada con las interacciones de los usuarios con los comercios o establecimientos en una plataforma. En conjunto, esta tabla proporciona una visión completa de las interacciones de los usuarios y sus opiniones sobre los comercios, lo cual es esencial para la toma de decisiones y la mejora de la plataforma.

3. Tabla 3: state

Nombre	Descripción
state	Atributo de la abreviación del nombre del estado o región en donde se encuentra el comercio o establecimiento.
name	Atributo del nombre del estado o región en donde se encuentra el comercio o establecimiento.

Fuente: Elaboración propia

Está es una tabla dimension que tiene una importancia fundamental en el modelo de entidad relación, ya que permite conectar los dataset de google maps y yelp de manera óptima.

4. Tabla 4: Business_yelp

Nombre	Descripción
Business_id	Atributo identificador único del comercio o establecimiento.
Name_business	Atributo de nombre del comercio o establecimiento.
Address	Atributo que indica la dirección física del comercio o establecimiento.
City	Atributo del nombre de la ciudad en donde se encuentra el comercio o establecimiento.

State	Atributo del nombre del estado o región en donde se encuentra el comercio o establecimiento.
Postal_code	Atributo con el código postal del área donde se encuentra el comercio o establecimiento.
Latitude	Coordenada Latitud.
Longitude	Coordenada Longitud.
Stars	Valoración del usuario sobre el comercio o establecimiento.
Review_count	Número de reseñas que aparecen sobre el comercio o establecimiento, que indican las opiniones o comentarios de los clientes o usuarios que han interactuado con ella.
Attributes	Atributos adicionales asociados al negocio
Categories	Atributo de las categorías o tipos de negocios a los que pertenece

Fuente: Elaboración propia

La tabla 4 permite identificar y relacionar de manera precisa y eficiente los diferentes comercios o establecimientos con sus respectivos atributos y características. Además, al incluir la valoración de los usuarios y el número de reseñas, esta tabla proporciona información relevante para la toma de decisiones de los usuarios, así como para el análisis y la gestión de los negocios.

5. Tabla 5: Tip_yelp

Nombre	**Descripción**
tip_id	Atributo de identificación única.
User_id	Atributo de identificación única del usuario en la plataforma.

Business_id	Atributo identificador único del comercio o establecimiento.
Texts	Reseña del usuario sobre el comercio o establecimiento.
Date	Atributo que indica la fecha en la que el comentario o reseña fue realizada por el usuario.

Fuente: Elaboración propia

La tabla 4 Esta información es valiosa porque permite entender las opiniones y experiencias de los usuarios con respecto a los comercios, lo que ayuda a evaluar la calidad y popularidad de los establecimientos, proporcionando información útil para la toma de decisiones de los usuarios y la gestión de los comercios. Además, esta tabla permite establecer relaciones entre los usuarios, los comercios y las reseñas, lo que facilita el análisis y la extracción de información relevante.

6. Tabla 6: User_yelp

Nombre	**Descripción**
User_id	Atributo de identificación única del usuario en la plataforma.
Avg_stars	Valoración del usuario sobre el comercio o establecimiento.

Fuente: Elaboración propia

Por último, la tabla 6 permite mantener la conexión entre los usuarios y las calificaciones de los comercios o establecimientos.

TRANSFORMANDO LOS DATOS - ETL (Extract Transform Load)

Durante este sprint del proyecto, llevamos a cabo un proceso de Transformación y Carga (ETL) para preparar los datos obtenidos de fuentes de Google Maps y Yelp. A continuación, explicaremos cómo se desarrolló este proceso.

Databricks:

Durante esta tarea utilizamos la tecnología de Databricks. Databricks permitió trabajar con bases de datos y tablas. Creamos una base de datos en Databricks que consistía en una colección de tablas. Estas tablas eran conjuntos de datos estructurados que utilizamos para almacenar y manipular los datos obtenidos de Google Maps y Yelp. Las tablas en Databricks podían ser consultadas y manipuladas utilizando diferentes lenguajes de programación compatibles con Spark.

Databricks y Apache Spark están estrechamente relacionados y son ampliamente utilizados en el análisis de big data debido a sus características y capacidades para procesar y analizar grandes volúmenes de datos de manera eficiente y escalable.

Spark se destaca por su capacidad para procesar datos en memoria, lo que permite un rendimiento muy rápido en comparación con otras tecnologías de procesamiento de datos a gran escala.

En el análisis de big data, Databricks y Spark se utilizan ampliamente debido a las siguientes razones:

1. Procesamiento distribuido:

Spark permite distribuir el procesamiento de datos en clústeres de máquinas, lo que permite escalar horizontalmente y procesar grandes volúmenes de datos de manera eficiente.

2. Velocidad de procesamiento:

Spark realiza operaciones en memoria, lo que brinda un rendimiento rápido para el procesamiento y análisis de datos. Esto es especialmente beneficioso cuando se trabaja con conjuntos de datos grandes o complejos.

3. Capacidad de procesamiento de datos estructurados y no estructurados:

Spark puede manejar una amplia variedad de fuentes de datos, incluyendo datos estructurados (como archivos CSV o bases de datos relacionales) y datos no estructurados (como archivos de texto o datos JSON).

4. Bibliotecas y capacidades de análisis avanzadas:

Spark ofrece una amplia gama de bibliotecas y módulos que facilitan el análisis de datos, como Spark SQL para consultas SQL, Spark Streaming para procesamiento de datos en tiempo real, Spark MLlib para aprendizaje automático, y Spark GraphX para análisis de gráficos.

Utilizamos el componente de Notebooks en Databricks para desarrollar la lógica de transformación de los datos. En los Notebooks, se escribe el código necesario para llevar a cabo las transformaciones requeridas en los datos. Se pudo utilizar diferentes lenguajes de programación, como SQL, Python y Scala, según nuestras necesidades. Los Notebooks nos permitieron ejecutar el código de manera interactiva y realizar visualizaciones relacionadas con el proceso de ETL.

Para mejorar la eficiencia y reutilización de código, se aprovecharon las librerías en Databricks. Este proporciona una amplia gama de librerías incorporadas, y también se fue posible instalar librerías de terceros según las necesidades.

Además, se utilizó la funcionalidad de Jobs en Databricks para automatizar la ejecución del código del Notebook. Se configuraron trabajos automatizados que nos permitieron programar la ejecución del proceso de ETL en momentos específicos, llevando un registro de que tablas ya habían sido transformadas. Creando un proceso de carga incremental. Esto brindó una mayor flexibilidad y control sobre el proceso de ETL, permitiendo ejecutar y supervisar las tareas de manera programada.

Transformación de datos:

Durante este sprint, nos enfocamos en realizar una serie de transformaciones en los datos para asegurarnos de que estuvieran limpios y listos para su análisis. Implementamos diferentes técnicas, como la selección de columnas relevantes, la aplicación de filtros para eliminar datos no deseados y la creación de nuevas columnas basadas en cálculos o manipulaciones de los datos existentes. Estas transformaciones permitieron optimizar la calidad y estructura de los datos, preparándose adecuadamente para las etapas posteriores del análisis.

Carga de datos:

Una vez que los datos fueron transformados, nos enfocamos en cargarlos en una base de datos para su posterior análisis. Durante esta etapa, nos aseguramos de que los datos fueran almacenados de manera eficiente y confiable en la base de datos seleccionada. Organizamos los datos en tablas según su relevancia y utilizamos las capacidades de almacenamiento de la base de datos para realizar una carga eficiente. Este paso fue crucial para garantizar que los datos estuvieran disponibles y accesibles para su posterior análisis.

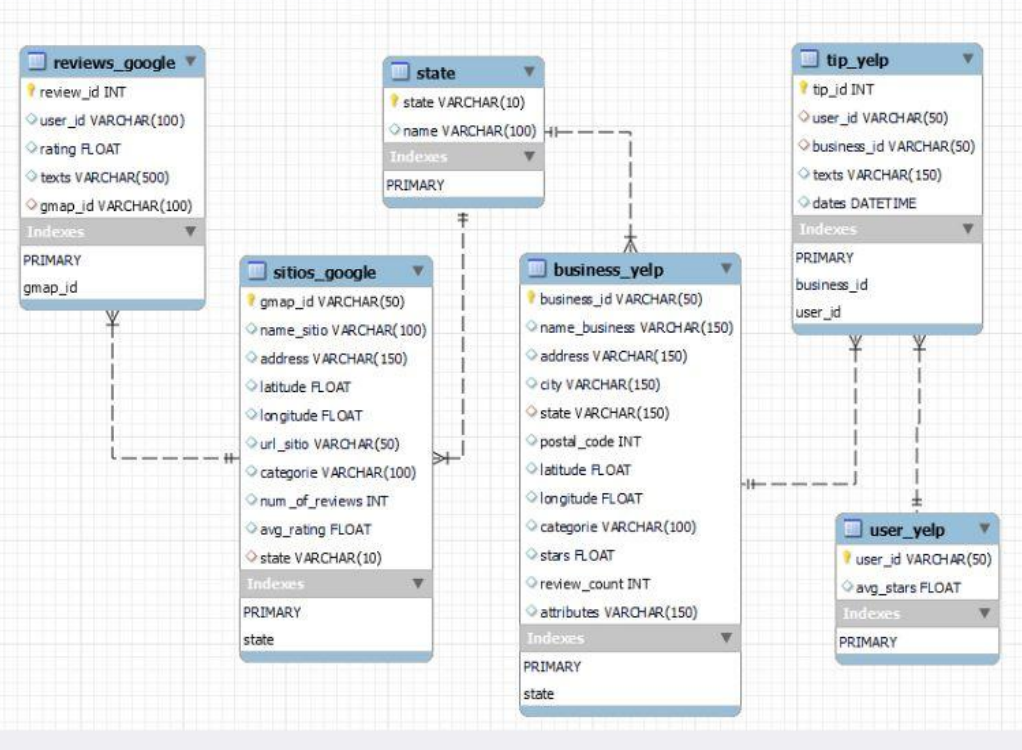
Durante todo el proceso de Transformación y Carga, aplicamos herramientas y técnicas que nos permitieron realizar estas tareas de manera eficiente y escalable. Nos apoyamos en funcionalidades avanzadas de procesamiento de datos para manipular y transformar los datos de manera adecuada, asegurando la integridad y calidad de la información.

Este proceso de ETL nos ha permitido preparar los datos de Google Maps y Yelp de manera efectiva para su posterior análisis. Al realizar las transformaciones necesarias y cargar los datos en una base de datos, hemos creado una base sólida sobre la cual podemos obtener insights valiosos y tomar decisiones fundamentadas.

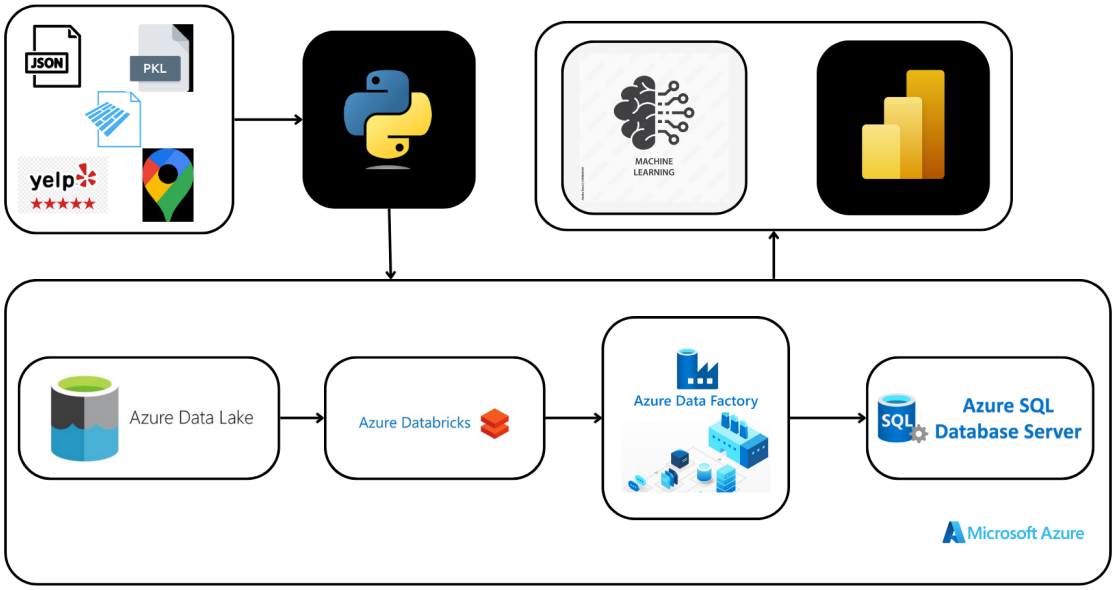
En resumen, durante este sprint del proyecto utilizamos Databricks como nuestra herramienta principal para llevar a cabo el proceso de Transformación y Carga (ETL) de los datos de Google Maps y Yelp. Utilizamos componentes como bases de datos y tablas para almacenar y manipular los datos, Notebooks para desarrollar la lógica de transformación, bibliotecas para mejorar la eficiencia y reutilización de código, y Jobs para automatizar la ejecución del proceso de ETL. Databricks nos proporcionó una plataforma sólida y flexible para llevar a cabo estas tareas de manera eficiente y escalable.

MODELO ENTIDAD RELACIÓN

El modelo entidad relación es una herramienta fundamental para el diseño de bases de datos relacionales, ya que permiten almacenar, organizar y gestionar la información de la organización. Para una empresa dedicada a la gastronomía turística, como lo es Turfood, el modelo entidad relación permite definir las entidades que intervienen en su negocio como, los productos y servicios, las reservas, las facturas, etc. El modelo entidad relación también facilita el identificar los atributos de cada entidad y relación, como por ejemplo, el nombre y dirección del comercio, la valoración y la reseña del servicio, etc. Estos atributos son los que se convertirán en los campos de las tablas de la base de datos y en la base con la que se crearan los kpi, el sistema de reseña y los modelos de machine learning. Además, el modelo entidad relación le permite a la empresa tener una visión clara y estructurada de su información, lo que le ayuda a optimizar sus procesos, mejorar el servicio al cliente, aumentar la rentabilidad y tomar mejores decisiones.



WORKFLOW PROPUESTO INICIALMENTE



Observaciones sobre el WORKFLOW

Cambio de Azure Data Factory por Azure Synapse

Se decidió cambiar el uso de Azure Data Factory por Azure Synapse, debido principalmente a problemas en su implementación. Azure Data Factory es un servicio de integración de datos que permite crear flujos de trabajo para mover y transformar datos desde diversas fuentes. Sin embargo, se encontraron dificultades para configurar y administrar el servicio, así como para integrarlo con otros componentes del entorno de Azure. Por ejemplo, se presentaron problemas para conectar Azure Data Factory con Azure SQL Database, Azure Data Lake Storage y Azure Databricks.

Azure Synapse es una plataforma de análisis de datos que integra los servicios de Azure SQL Database, Azure Data Lake Storage y Azure Databricks en una sola interfaz. Synapse permite crear un sistema ETL automatizado y realizar procesos de carga incremental con mayor facilidad y eficiencia. Al cambiar el uso de Azure Data Factory por Azure Synapse, se logró simplificar la

arquitectura del proyecto, reducir los costos operativos y mejorar el rendimiento y la escalabilidad del sistema. Synapse también facilitó la colaboración entre los diferentes roles involucrados en el proyecto, como los ingenieros, analistas y científicos de datos. Synapse demostró, para este proyecto, ser una solución más completa y robusta para el análisis de datos en la nube.

Empleo de Azure Synapse sobre SQL

Se ha tomado la decisión de reemplazar Azure SQL con Azure Synapse, debido a las ventajas que ofrece este último en términos de gestión de tablas e integración con otros servicios. Azure Synapse es una plataforma analítica que combina capacidades de almacenamiento, procesamiento y visualización de datos, permitiendo crear soluciones escalables y eficientes. Con Azure Synapse, se puede acceder a los datos desde diferentes fuentes, aplicar transformaciones y análisis avanzados, y generar informes y paneles interactivos. Además, Azure Synapse se integra fácilmente con otras herramientas y servicios de Microsoft, como Power BI.

WORKFLOW MODIFICADO

1. Configuración de Azure

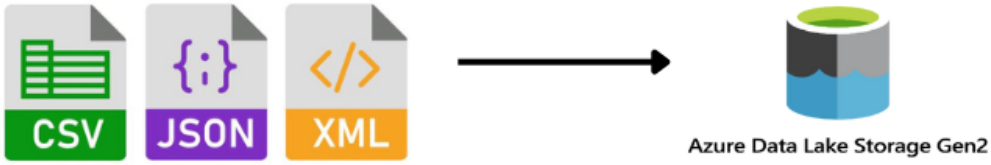
En este punto se desarrollan la creación de los servicios y las configuraciones requeridas.

- Creación de Azure Data Lake con el fin de almacenar los datos.
- Creación del entorno de Azure Databricks para la transformación de los datos.
- Creación del servicio de Azure Synapse para los pipelines y bases de datos.

2. Captura de datos

Se evalúan y emplean las diferentes fuentes de datos propuestas, a fin de tener un conjunto de dataset de utilidad.

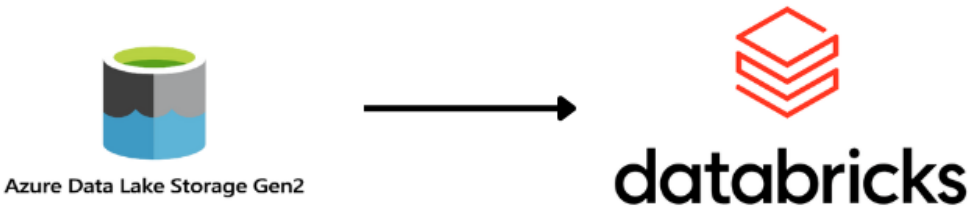
- Captura y carga de datos en Azure Data Lake



3. Extracción y almacenamiento

En conjunto, las tecnologías interactúan entre sí a fin de asegurar un correcto flujo de datos. Los datos serán almacenados en Azure Data lake en bruto de la siguiente manera.

- Los datos se cargan en Azure Data Lake.
- Por medio de Jobs los datos se cargan en Azure Databricks delta Lake.

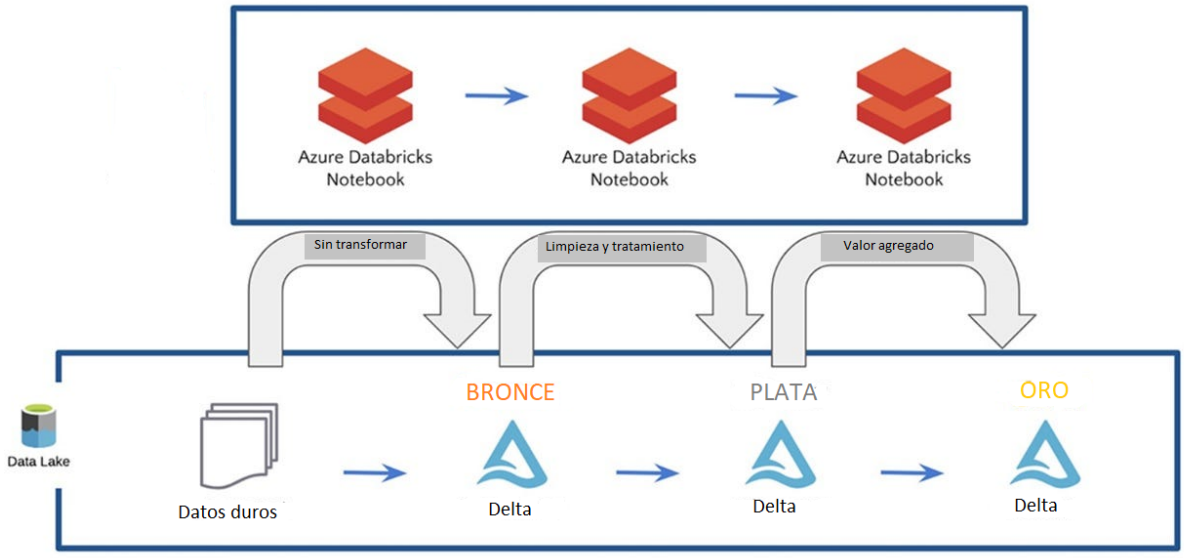


4. Procesamiento en Azure Databricks

Azure Databricks se emplea para el trabajo de procesar los datos en bruto, para ello, se utilizan las capacidades de trabajar con Notebooks compatibles con Python, Scala, R y SQL.

- Creación de procesos de limpieza y transformación de datos, cambiando con ello, los datos en bruto por información de utilidad, esto se hace en tres etapas.

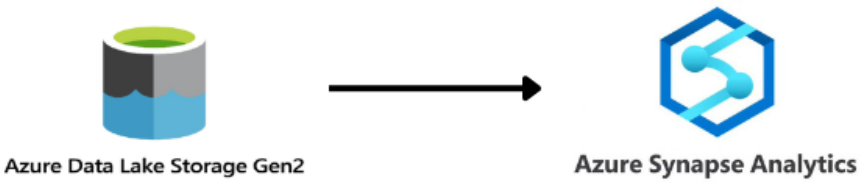
1. Etapa bronce: los datos se guardan en crudo con pocos procesos de transformación. Se mantienen filtros básicos, priorizando la coherencia a la calidad.
2. Etapa silver: se emplean procesos de limpieza de datos, como tratamiento de nulos y vacíos, eliminación de duplicados, corrección del tipo de columna, entre otros. En esta etapa se aplica todo proceso de limpieza y tratamiento de datos necesario.
3. Etapa gold: Aplicación de procesos como agregación de datos, creación de tablas de hechos y dimensiones. Los datos son tratados de maneras más avanzadas, en procesos, que buscan normalizar y enriquecer la información tratada.



5. Carga en Azure Data Lake y Synapse

Cuando los datos se encuentren transformados, se realiza un proceso de copia de la siguiente manera.

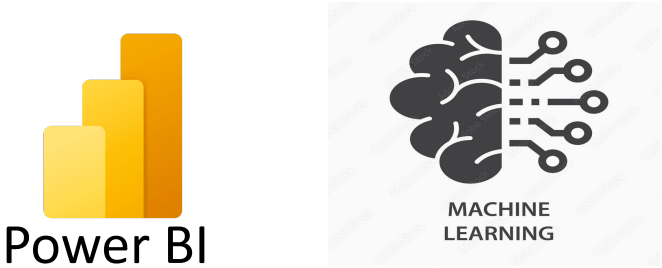
- Los datos de las etapas silver y gold, se carga en Azure Data Lake.
- Los datos finales en la etapa gold se copian desde Data Lake a Synapse.
- Los datos oro se emplean para la creación del modelo Entidad-Relación.
- Define el tipo de dato para cada columna, asegurando la consistencia e integridad de los procesos realizados.
- Validar los datos, por medio de pruebas de verificación.



6. Creación de informes y métodos de machine learning

Por medio de la conexión con Azure Synapse, se emplean los datos para crear informes en PowerBi. Utilizando los diferentes datos almacenados, tanto en Synapse, como en las diferentes etapas en Data Lake (bronze, silver y gold), se entrenan los modelos de machine learning.

- Creación de visualizaciones, empleado PowerBi, usando los datos de Azure Synapse.
- Preparación, entrenamiento y ejecución de los modelos propuestos.

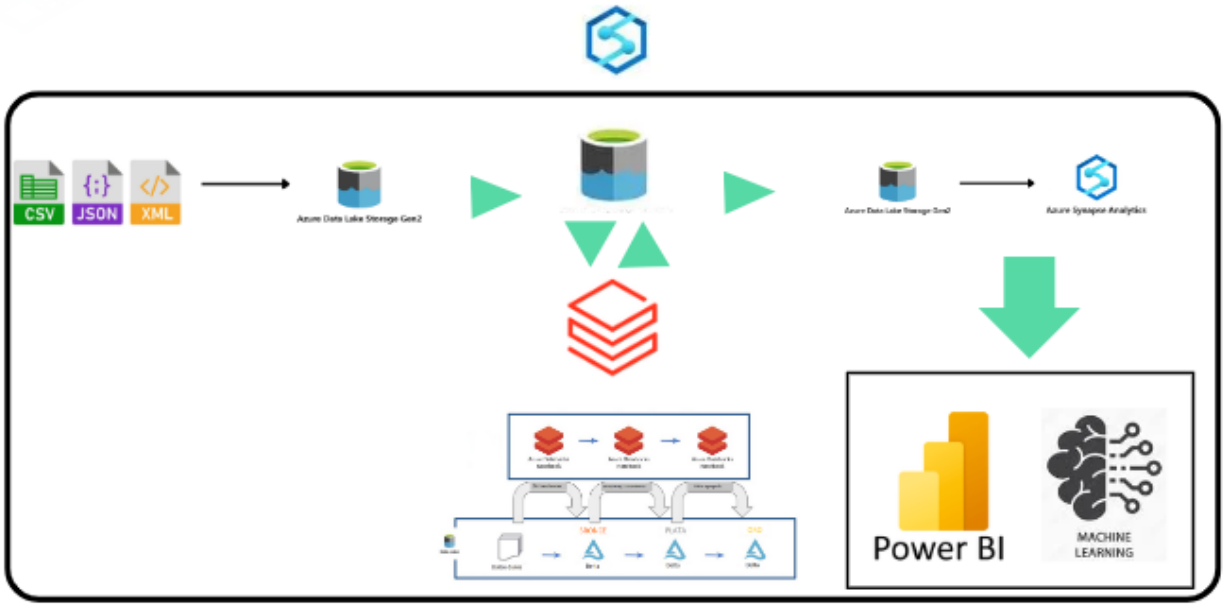


PROPUESTA DE AUTOMATIZACIÓN

A raíz de la modificación propuesta respecto a los recursos que decidimos contemplar en el stack tecnológico vemos la posibilidad de lograr la integración de los mismos mediante la plataforma que nos ofrece Azure, Azure Synapse Analytics.

A través de la conexión que puede establecer dicha plataforma a los recursos utilizados en el stack tecnológico podemos gestionar actividades que involucran tanto al almacenamiento de datos (Azure Data Lake), procesamiento, transformación y carga de datos (Azure Databricks) y a la base de datos SQL en Synapse Analytics. Logrando así integrar desde la ingesta de datos en Azure Data Lake (ya sea desde un entorno local o realizando consultas asociadas a una API), la activación de la ejecución del proceso de transformación de los datos crudos (Azure Databricks), hasta la disponibilización de los datos ya listos para el consumo de los modelos de machine learning y las correspondientes consultas para generar las visualizaciones en Power BI.

De modo que la capacidad de integración que proporciona Azure Synapse Analytics como plataforma unificada para la ingesta, preparación, administración y entrega de datos se observa muy beneficiosa.



CONCLUSIONES

Stack tecnológico optimizado

- Realizamos una cuidadosa selección y ajustes del stack tecnológico para maximizar los recursos.
- Al eliminar Azure Factory y Azure SQL Server, se logró reducir significativamente el proceso de los datos.
- Además, se aprovecha al máximo la capacidad de cada herramienta seleccionada, lo que permite mejorar la eficiencia y la calidad del flujo de trabajo.

Adaptación exitosa

- Durante el desarrollo del proyecto se sortearon desafíos específicos que requieren soluciones flexibles y escalables.
- La adaptación del stack tecnológico permitió abordar estos desafíos de manera efectiva, ajustando las herramientas y componentes según las necesidades.
- Esta capacidad de adaptación brindó mayor agilidad y permitió superar obstáculos de manera eficiente

Avances significativos

- Se lograron importantes avances en varias etapas claves del proyecto.
- En término de procesamientos de datos, se realizó un ETL (extracción, transformación y carga) eficiente, asegurando la calidad y coherencia de los datos para el Dashboard, creación de informes y el entrenamiento de los Modelos de Machine Learning.
- Además, se diseño un sólido modelo entidad-relación que representa de manera precisa y estructurada la información.
-