DATUM TECH

Data Analytics y Machine Learning

PUNTOS A ENTREGAR

- Dashboard (diseño)
- Dashboard (conexión a DB)
- ML en producción

CAMBIOS Y AMPLIACIÓN DEL STACK TECNOLÓGICO

Azure Synapse por Databricks

Synapse es una plataforma de análisis de datos que integra servicios de Azure como Azure Data Lake Storage gen 2, Azure SQL Database, Azure Cosmos DB, entre otros. En el desarrollo del proyecto y en consecuencia a la evolución natural del mismo, se decide cambiar el uso de Azure Databricks por Synapse. Razones para esta decisión son:

1. Se presenta una experiencia unificada para el análisis de datos, que permite trabajar con lenguajes como SQL, Python, Scala y R en un solo entorno.
2. Se tiene una integración nativa con Power BI, que facilita la visualización y el consumo de los datos analizados.
3. Una mayor velocidad y rendimiento, gracias a la optimización de los motores de procesamiento y almacenamiento de datos.
4. Sistema de gestión de los servicios de azure integrado, lo que permite desde Synapse enviar y recibir datos con gran facilidad.
5. Synapse gracias a su motor apache spark permite manejar grandes volúmenes de datos, sin perder flexibilidad, escalabilidad y eficiencia.

La integración de Synapse con SQLPool y el motor de Apache Spark permite transferir de forma eficiente grandes volúmenes de datos entre el entorno de ejecución de Spark y el SQLPool dedicado. El conector de SQLPool dedicado para Apache Spark se implementa usando Scala y admite Scala y Python ofreciendo las siguientes caracteristicas:

- Leer desde SQLPool dedicado de Synapse: leer grandes conjuntos de datos desde tablas (internas y externas) y vistas de SQLPool dedicado de Synapse.

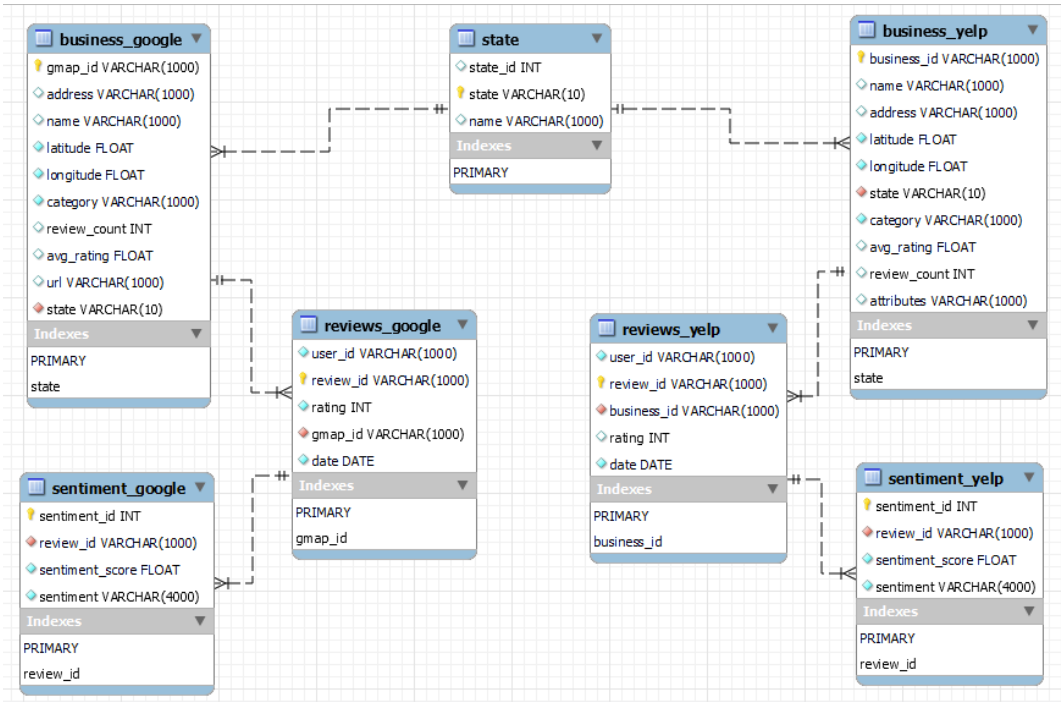
- Soporte para la propagación de consultas.
- Escribir en SQLPool dedicado de Synapse: ingerir grandes volúmenes de datos en tipos de tabla internos y externos.

Por estos motivos se tomó la decisión de migrar las operaciones realizadas en los notebooks de Databricks, a Synapse que gracias a su motor de Apache los mismos notebooks se ejecutan, incluso ya deja de ser necesaria la conexión con el Data Lake para la extracción de datos.

Github + Render

GitHub es una plataforma de desarrollo colaborativo que permite alojar y gestionar proyectos de código abierto. Render es un servicio de computación en la nube que ofrece una forma fácil de desplegar aplicaciones web, funciones sin servidor, bases de datos y otros recursos. Ambos servicios se integran entre sí y permiten crear flujos de trabajo automatizados para el desarrollo y la implementación de modelos. Con este método, se tiene un modelo de recomendación funcional y en línea, sin necesidad de emplear servidores ni infraestructura compleja. Además, se puede aprovechar la potencia de GitHub y Render para actualizar el modelo, escalar y monitorizar de forma sencilla.

MODELO DE ENTIDAD-RELACIÓN DEFINITIVO



DICCIONARIO DE DATOS

Tabla 1: state

Nombre	Descripción
state_id	Atributo de identificación única.
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.

Tabla 2: business_google

Nombre	Descripción
gmap_id	Atributo que indica la identificación única que Google Maps le asigna a la dirección física del comercio o establecimiento.
address	Atributo que indica la dirección física del comercio o establecimiento.
name	Atributo que indica el nombre del comercio o establecimiento que se quiere representar como una entidad en el modelo.
latitude	Coordenada de Latitud
longitude	Coordenada de Longitud
category	Atributo que indica la categoría del 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.

avg_rating	Valoración promedio.
url	URL de Google Maps que direcciona a la zona geográfica donde se localiza dicho comercio o establecimiento.
state	Atributo del nombre del estado o región en donde se encuentra el comercio o establecimiento

Fuente: Elaboración propia

Esta tabla 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.

Tabla 3: business_yelp

Nombre	Descripción
business_id	Atributo identificador único del comercio o establecimiento, que asigna la plataforma Yelp.
name	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.
latitude	Coordenada de Latitud
longitude	Coordenada de Longitud
state	Atributo del nombre del estado o región en donde se encuentra el comercio o establecimiento
category	Atributo que indica la categoría del comercio o establecimiento.
avg_rating	Valoración promedio.

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.

Fuente: Elaboración propia

La tabla 3 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

Tabla 4: reviews_google

Nombre	Descripción
user_id	Atributo de identificación única del usuario en la plataforma.
review_id	Atributo de identificación única de identificación de reseña.
rating	Valoración 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.
date	Atributo que indica la fecha en la que el comentario o reseña fue realizada por el usuario.

Fuente: Elaboración propia

La tabla 4 es importante porque representa la información clave relacionada con las interacciones de los usuarios con los comercios o establecimientos en la plataforma Google Maps. 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.

Tabla 5: reviews_yelp

Nombre	Descripción
user_id	Atributo de identificación única del usuario en la plataforma.
review_id	Atributo de identificación única de identificación de reseña.
business_id	Atributo identificador único del comercio o establecimiento, que asigna la plataforma Yelp.
rating	Valoración 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 5 es de importancia ya que presenta información clave relacionada con las interacciones de los usuarios con los comercios o establecimientos en la plataforma Yelp. 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.

Tabla 6: sentiment_google

Nombre	Descripción
sentiment_id	Atributo de identificación única del sentimiento en la plataforma.
review_id	Atributo de identificación única de identificación de reseña de la plataforma Google Maps..
sentiment_score	Valoración otorgada por el modelo de sentimientos a la reseña.
sentiment	Atributo de clasificación del sentimiento.

Fuente: Elaboración propia

La tabla 6 es de importancia ya que presenta información clave relacionada con las opiniones de los usuarios respecto de los comercios o establecimientos en la plataforma Google Maps. Proporcionando los diferentes tipos de sentimientos asociados a las diferentes reseñas .

Tabla 7: sentiment_yelp

Nombre	Descripción
sentiment_id	Atributo de identificación única del sentimiento en la plataforma.
review_id	Atributo de identificación única de identificación de reseña de la plataforma Yelp.
sentiment_score	Valoración otorgada por el modelo de sentimientos a la reseña.
sentiment	Atributo de clasificación del sentimiento.

Fuente: Elaboración propia

La tabla 7 es de importancia ya que presenta información clave relacionada con las opiniones de los usuarios respecto de los comercios o establecimientos en la plataforma Yelp. Proporcionando los diferentes tipos de sentimientos asociados a las diferentes reseñas .

MODELO DE RECOMENDACIONES

Un modelo de recomendaciones es una herramienta que ayuda a los usuarios a encontrar los lugares que más les gusten y se acomoden a sus necesidades, esto según sus preferencias o características. Los modelos de recomendaciones son muy útiles e importantes para ofrecer un servicio personalizado y mejorar la experiencia del usuario, así como para aumentar las ventas o la fidelización de los clientes. Con lo anterior en mente, Datum Tech desarrolló un modelo de recomendaciones amigable para el usuario y potente en sus resultados. A partir de ello se crearon dos endpoints con modelos diferentes.

Similitud del coseno

La similitud del coseno es una medida de la similitud entre dos vectores, funciona calculando el coseno del ángulo entre los dos vectores, que es igual al producto escalar de los vectores dividido por el producto de sus normas. El resultado es un valor entre -1 y 1, donde -1 indica una similitud nula, 0 indica una similitud media y 1 indica una similitud máxima. La similitud del coseno es de utilidad e importante en el modelo de recomendación porque permite calcular la afinidad entre los perfiles de los usuarios y las características que se quieren recomendar.

El modelo sigue los siguientes pasos para encontrar los mejores sitios adaptados al gusto del cliente:

1. Filtra los datos según el estado y la categoría introducidos por el usuario.
2. Selecciona la información relevante para mostrar las recomendaciones (nombre, dirección, rating, categoría y atributos del sitio).
3. Calcula la matriz de similitud del coseno para encontrar el sitio que mejor se adapte a los parámetros introducidos del usuario.
4. Ordena los índices de los sitios según su similitud con la información que se toma de referencia, de mayor a menor.
5. Devuelve los primeros 5 lugares con mayor similitud a las características deseadas.

De esta manera, se genera una lista de 5 sitios similares, según su ubicación geográfica, su rating y su número de reseñas. Estos sitios pueden ser considerados como recomendaciones para el usuario que busca lugares para visitar en su estado y categoría de interés.

A continuación, un ejemplo en el estado de California con preferencias de un pub, restaurante:

name	address	avg_rating	categories	attributes
Ball & Chain	Ball & Chain, 1643 N Cahuenga Blvd, Los Angele...	4.2	Gastropub, Restaurant	https://www.google.com/maps/place//data=!4m2!3...
Porter's House	Porter's House, 20209 Rinaldi St, Porter Ranch...	4.4	Pub, Restaurant	https://www.google.com/maps/place//data=!4m2!3...
Pappy McGregor's Pub & Grill	Pappy McGregor's Pub & Grill, 1865 Monterey St...	4.1	Irish pub, Bar & grill, European restaurant, G...	https://www.google.com/maps/place//data=!4m2!3...
Trifecta Tavern	Trifecta Tavern, 2600 Via De La Valle Ste 100,...	4.4	Bar, Pub, Restaurant	https://www.google.com/maps/place//data=!4m2!3...
Whistling Duck Tavern	Whistling Duck Tavern, 1040 University Ave, Sa...	4.6	Gastropub, Asian fusion restaurant, Bar, Chine...	https://www.google.com/maps/place//data=!4m2!3...

K-vecinos cercanos

El modelo de k-vecinos cercanos es un método de aprendizaje automático supervisado que se basa en la idea de que los objetos similares tienden a estar cerca unos de otros en el espacio de características. Este método consiste en asignar a un nuevo objeto el valor más frecuente entre sus k-vecinos más cercanos, donde k es un número entero positivo que se elige previamente. Para medir la distancia entre los objetos, se puede utilizar cualquier métrica adecuada, como la distancia euclidiana o la distancia de Manhattan. El modelo de k-vecinos cercanos es simple de implementar y puede adaptarse a diferentes tipos de problemas de clasificación o regresión, por ello se emplea un segundo modelo que utiliza el método de k-vecinos cercanos en conjunto con twitter-xlm-roberta-base-sentiment.

El uso de modelo k-vecinos con cardiffnlp/twitter-xlm-roberta-base-sentiment permite evaluar no solo la cercanía de los locales según la distancia y las características, también permite evaluar el sentimiento de las reseñas respecto al local, creando así un filtro extra que asegura que la experiencia sea lo más positiva posible. La combinación de ambos modelos permite aprovechar el conocimiento transferido del modelo pre-entrenado y adaptarlo a los datos específicos del problema de análisis de sentimiento sobre las reseñas.

A continuación, un ejemplo en el estado de Arizona con preferencias de un local de estilo italiano.

name	address	avg_rating	categories	attributes	Sentimiento
Fresco Italian Cafe on the Canal	340 W Michigan St	4.5	Italian, Beer Bar, Bars, American (New), Night...	BikeParking, BusinessAcceptsCreditCards, Cater...	positivo
Diavola Pizza	1134 E 54th St, Ste I	4.5	Italian, Restaurants, Pizza, Bars, Nightlife, ...	BikeParking, BusinessAcceptsCreditCards, Cater...	positivo
Buona Pizza	7407 Shadeland Ave	4.0	Restaurants, Italian, Pizza	BusinessAcceptsCreditCards, GoodForKids, HasTV	positivo
Chris' Pizza Village Pleasant View	244 Village Sq Pleasent, Ste 100	4.0	Chicken Wings, Italian, Salad, Pizza, Restaurants	BikeParking, BusinessAcceptsCreditCards, GoodF...	positivo
Mirko Pasta	2037 N Mt Juliet Rd	4.0	Food, Italian, Pasta Shops, Specialty Food, Fa...	BusinessAcceptsCreditCards, Caters, GoodForKid...	positivo

Análisis de sentimientos

El análisis de sentimientos es una tarea específica del procesamiento del lenguaje natural que consiste en identificar y extraer las opiniones, emociones y actitudes expresadas en un texto. Esta tarea tiene múltiples aplicaciones prácticas, como el monitoreo de la reputación de una marca, la detección de la satisfacción del cliente o la clasificación de reseñas de productos. Sin embargo, el análisis de sentimientos también plantea varios desafíos, como la ambigüedad, la ironía o el sarcasmo, que pueden dificultar la interpretación correcta de los textos. Para lograr un idóneo análisis sobre los sentimientos de las reseñas se emplearon las siguientes tecnologías:

1. NLTK (Natural Language Toolkit)

La librería NLTK de Python es una herramienta versátil y poderosa para el análisis de sentimientos. NLTK identifica y extrae la opinión o 'sentimiento' de un texto respecto a un tema o entidad. El análisis de sentimientos proporcionado por NLTK puede aplicarse a diversos campos, como los negocios, la política, la educación, la salud, entre otros. NLTK es una de las principales librerías de Python para el procesamiento del lenguaje natural (NLP por sus siglas en ingles) dada su eficacia y facilidad de uso, para tareas como, la tokenización, la etiquetación gramatical, el análisis sintáctico o la generación de texto. Entre sus funcionalidades, se encuentra el módulo nltk.sentiment, que contiene clases y métodos para realizar el análisis de sentimientos con diferentes enfoques y técnicas.

Uno de los componentes más destacados de este módulo es el analizador VADER (Valence Aware Dictionary and sEntiment Reasoner), el cual detecta el sentimiento de un texto con base en aspectos como la intensidad, la negación, el énfasis o el contexto. VADER utiliza un diccionario léxico que asigna valores numéricos a las palabras según su polaridad y su intensidad (positiva(1), negativa(-1) o neutra(0)). Además, incorpora reglas gramaticales y sintácticas para ajustar estos valores en función de la estructura del texto.

A continuación se muestra un ejemplo del resultado el análisis de sentimientos con NLTK usando Python:

text	Puntaje de sentimiento	Sentimiento
mixed feelings place. pros- great location. really heart everything. close attractions french quarter. history. many famous people stayed here. price great! sheets soft. saltwater pool amazing spoiled normal pools. staff friendly. cons- wifi bathroom light worked portion time. everything older outdated (carpet, tile, balcony, bedspread, elevator). ice box lower level. coffee makers refrigerators rooms.	0.9523	positivo
tied best indian philly. delivery mess (containers broken open, food cold) may restaurant's fault. tough say. solid 3-star now.	-0.4588	negativo
manager friendly. staff seems run well there. food fresh, fries, potato cakes hot. food takes minutes asks us seat someone brings food table. never issue here. tend hit grocery shopping buy lot junk store.	0.765	positivo
must try!! order take rate servers always positive experience. place usually busy go food always hot pick up, someone always greets soon walk place clean welcoming. personally love drunken noodle dish get shrimp beef good portion size make 2 meals plus! husband gets chicken satay coconut rice loves it. definitely recommend place!!	0.9798	positivo
first glance restaurant seems italian restaurant. careful. menu lot items relationship italian cuisine. white table cloths also lend certain ethnic identity. restaurant food menu ole ole. quick note continue, tommy's warehouse district anyone else. setting gorgeous maintains characteristics buildings area. design decoration help relaxing step above. food truly star. never bad dish specific parts menu truly unique. always looked louisiana restaurant first. ordered turtle soup beet salad along linguini. food service fantastic. third time eaten perfect time.	0.9809	positivo

2. BERT (Bidirectional Encoder Representations from Transformer)

BERT es un modelo de procesamiento del lenguaje natural, el cual por medio de redes neuronales (deep learning) logra entender el contexto y significado de un texto dado.
El modelo codifica las palabras a partir de los transformers, los que permiten procesar secuencias largas de datos. Una de las características importantes de BERT es su bi-direccionalidad, ésta permite analizar los textos, tanto de derecha a izquierda, como de izquierda a derecha. BERT se especializa en el procesamiento de lenguaje natural (NLP), permitiendo abarcar tareas como: compresión de lectura, generación de texto, traducción automática, análisis de sentimientos, entre otros.

El modelo cardiffnlp/twitter-xlm-roberta-base-sentiment es una herramienta muy útil y potente para el análisis de sentimientos, ya que, consisten en un modelo BERT afinado para el análisis de sentimientos en tweets, además, emplea una versión multilingüe de BERT llamada XLM-RoBERTa que puede procesar textos en 100 idiomas diferentes. cardiffnlp/twitter-xlm-roberta-base-sentiment ha sido entrenado con más de 198 millones de tweets etiquetados con sentimientos positivos, negativos o neutros y ha demostrado tener un alto rendimiento y precisión en esta tarea. El modelo puede asignar una etiqueta de sentimiento (positivo, neutral o negativo) a un texto, como lo puede ser una reseña de un restaurante, un hotel u otro servicio. El modelo tiene una alta precisión y puede manejar el lenguaje informal y los emojis que se usan comúnmente en las redes sociales.

A continuación, se muestra un ejemplo del resultado del análisis de sentimientos:

text	Sentimiento
first, review market, dining room, order go. advice love place, place go order here. eating pretty regularly never tried order take recent visit. one people group needed get home quickly planning maynard's figured stick it. never considered market carryout certainly trying difficult. mistake trying jerk asking service offered, really know. guess market seemed like carryout friendly menu. anyway, issue way lady behind counter handled situation. handled polite, carryout. however, really rude clearly irritated. seriously need that. went menu intended carryout proper take containers. ok, get carryout. really, brought home leftovers plenty times consider proper containers necessary make reasons annoyed asked carryout. tell something offer come back next time dine in. vibe place usually fun; sad ordering listening people order. pretty short guy ordered us too. dining questions menu; possibly still irritated order. good experience. love guessing back. lesson learned, dine only. see person behind counter heading across street congress ended afternoon.	Negativo
place amazing.. relaxing daughter loved made friends enjoyed view... good food.	Positivo
prosciutto burger more. told good sale. fried mushroom delicious fried pickles salty. however, agree burgers taste average high dollar. $13 probably worth $7-8. much better burgers around lower price. worst thing added tips bill. two parties 7 people together. added 20% ask additional tip. normally tip, (easier calculate 18%), like forced appreciate service! felt slightly insulted!	Negativo
great fusion food! would't necessarily categorize place modern vietnamese bistro bistro asian twist. beef steak eggs good much beef make meal. chicken green waffles good waffles heavy. chicken good sauce waffles made dish. pho good great good. kimchi burger great fries fantastic. worth effort come dinner.	Positivo
yes, pancakes good! still, never stand line breakfast food. mean "break fast," super hungry. hungry, want food right away whine starving. boyfriend reminds me, "you're starving. people rwanda starving. hungry." true. still, warned someone born raised nashville must get 7:30 a.m., did. three party four vacation one working. turn seated, working one take call went outside yet rest us walked table. none crappy policy waiting entire party ready seated -- policy common busy dc restaurants. sat drooled menu friend finished business outside. phone call, mean. pooping outside. waited, waited, server, kept checking us topping coffee patiently waited friend join us ordering. imagine that, rushed getting 'tude busy staff! food came -- sweet potato, cornmeal, caribbean, regular potato pancakes shared enjoyed table. sweet potato pancakes came sort creamy cinnamon syrup, drank dispenser. pancakes heavenly, light, fluffy, good were, served vehicle wonderful sauce/syrup! hype. really good, get early!	Negativo

CONCLUSIÓN

En el desarrollo del proyecto se optó por utilizar Synapse como una alternativa a Azure Databricks, ya que permite una integración más profunda con los servicios de Azure, una menor complejidad de gestión y una mayor escalabilidad y rendimiento. Además, se implementaron dos modelos de recomendación que se adaptan a las preferencias de los usuarios y ofrecen resultados personalizados. Por último, se aplicó técnicas de procesamiento de lenguaje natural para extraer los sentimientos de los usuarios a partir de sus reseñas, utilizando modelos pre-entrenados, con estos métodos se puso clasificar los comentarios en positivos, neutros o negativos, y obtener indicadores de sentimiento de los usuarios.