Proyecto Final - Data 03 - Grupo 11

Equipo:

- Álvarez, Mateo Emmanuel
- Mazzucco, Uriel Agustín
- Pilla, María del Pilar
- Rojas, Martín

Problemática Inicial:

La compañía Amazon quiere mantener su supremacía en los mercados en línea, por lo que contrata a un grupo de especialistas para que analicen los datos de sus reseñas, obtengan insights valiosos para el negocio, e implementen un modelo de recomendación.

Dado el tamaño de los datasets, las capacidades de almacenamiento y cómputo personales no resultan efectivas. Por este motivo utilizaremos herramientas de Cloud Computing. El servicio que utilizaremos es Google Cloud.

Alcances:

Utilizaremos Datasets de reseñas desde mayo de 1996 hasta julio de 2014, y metadata de los productos, facilitados por la empresa. La selección utilizada abarca las reseñas de productos con al menos 5 reseñas. Esto hace un total que supera los 41 millones de reseñas y 9 millones de productos.

Enfoque:

El grupo optó por realizar un análisis a partir de las reseñas de los usuarios, acudiendo a la opinión del consumidor como fuente de información externa sobre los productos, así como información sobre el negocio a partir de la satisfacción de los clientes.

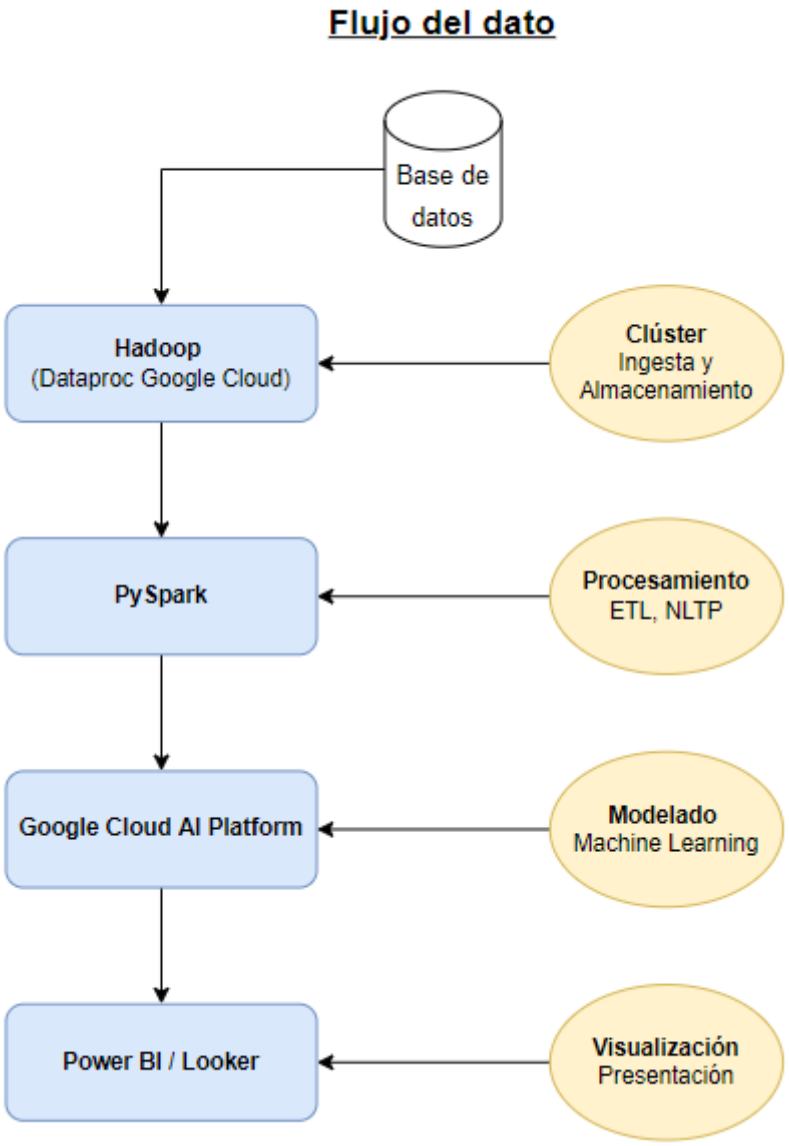
Objetivos:

- Obtener nueva información sobre los productos mediante un procesamiento de texto de las reseñas. Puede tomar la forma de variables y etiquetas.
- Analizar KPI’s surgidas de esta nueva información.
- Generar un algoritmo de recomendación de productos.

Fuera de Alcance:

- Sistema de etiquetas review-based para guiar al usuario.
- Actualización en streaming con toda la información de la empresa.
- Trazado de perfiles para cada usuario de la plataforma.
- Aplicar un filtro regional a las métricas

Proceso:



1. **Ingesta y Almacenamiento: Hadoop.**

   Un clúster en la nube nos permite sortear las limitaciones de almacenamiento local, brinda seguridad y favorece el procesamiento posterior.

   Para ello crearemos un clúster con Dataproc (Google Cloud).

Pipeline: Vía comandos bash obtendremos y descomprimiremos los datasets, para luego ingestarlos y procesarlos a través de Spark.

- **Procesamiento : PySpark**

  Spark nos posibilita implementar soluciones de procesamiento de datos integradas con python, lo cual será de mucha ayuda en el procesamiento de texto. Mediante trabajos.

  I. ETL:

Aunque los datos recibidos ya tienen realizado un proceso de limpieza y selección, realizaremos algunos ajustes de normalización de columnas y nos desharemos de algunas columnas que no aportan utilidad.

  II. Procesamiento de Texto: NLP

Buscaremos palabras claves que aporten información extra sobre el producto desde la percepción del comprador, como la calidad, el precio o el acabado; y la satisfacción del cliente.

  III. Nuevos indicadores:

Con los datos obtenidos del texto crearemos nueva información, en forma de columnas extras.

3. **Modelado: Google Cloud AI, Sklearn**

   Implementaremos un modelo de ML que recomiende productos en base a la información tanto previa como recién generada.

4. **Visualización: PowerBI , Looker, BigQuery**

   Crearemos una presentación analizando los patrones y KPIs que encontremos.

KPIs:

- NPS (Net Promoter Score) (% promotores - %detractores)
- MAU (Monthly Active Users) (n° de personas individuales que accedieron al menos una vez a Amazon en 30 días)
- Ticket promedio (valor total facturado / número de pedidos)
- CSAT (Índice de Satisfacción del Cliente) (número de clientes satisfechos / total clientes)
- Los más comprados en general y de cada categoría.

Herramientas Alternativas:

Otros SQL: No brindan las necesarias soluciones de Cloud Computing.
Docker: No brinda las necesarias soluciones de Cloud Computing.
AWS: Es la primera alternativa a las herramientas de Google Cloud.
Azure: Es la segunda alternativa a las herramientas de Google Cloud.

Diagrama Gantt: