MODULO 07 - EJERCICIO 01-A

ALEXIS YURI M.

Analiza un caso de uso propuesto (por ejemplo: redes sociales, salud digital, transporte urbano o e-commerce) e identifica cómo se manifiestan las 5V's de Big Data. Argumenta con ejemplos y propone tecnologías del ecosistema Big Data que podrían aplicarse.

Se toma como caso el análisis de datos de una plataforma de redes sociales como X (ex Twitter).

Identificación de las 5V's

- Volumen: X procesa un inmenso volumen de datos. Cada día se generan cientos de millones de tweets, retweets, likes, comentarios, y mensajes directos. Este volumen no cabe en una base de datos "tradicional".

- Velocidad: La velocidad de generación de datos es extremadamente alta. Los tweets se publican en tiempo real, con peaks significativos durante eventos de gran repercusión (noticias de última hora, eventos deportivos, conciertos). Los sistemas de Big Data deben procesar estos datos a medida que llegan.

- Variedad: La variedad de datos es alta e incluye:

Datos estructurados: metadatos de usuario (nombre, ID), la fecha y hora del tweet.

Datos semi-estructurados: el contenido del tweet en formato JSON.

Datos no estructurados: imágenes, videos, GIFs adjuntos, y el texto de cada tweet.

- Veracidad: La veracidad de los datos en redes sociales es un gran desafío. Los tweets pueden contener fake news, opiniones sesgadas, spam, y bots. Un análisis de Big Data debe incluir métodos para identificar y filtrar esta información no confiable para obtener resultados válidos.

- Valor: El valor se extrae al analizar los datos para obtener insights. Por ejemplo, las empresas pueden analizar la opinión de los clientes sobre sus productos, los gobiernos pueden monitorear el sentimiento público ante ciertas políticas, y los analistas de marketing pueden identificar tendencias o influencers.

Herramientas para cada V

- Volumen: Para manejar el almacenamiento de grandes volúmenes, se puede usar un sistema de archivos distribuido como HDFS (Hadoop Distributed File System).

- Velocidad: Para procesar datos en tiempo real (por ejemplo streaming), se pueden usar herramientas como Apache Kafka para la ingesta y Apache Spark Streaming para el procesamiento.

- Variedad: Para almacenar y procesar datos con esquemas flexibles, se pueden utilizar bases de datos NoSQL como MongoDB (para documentos) o Cassandra (para grandes volúmenes de datos distribuidos).

- Veracidad: Se pueden aplicar algoritmos de Machine Learning sobre plataformas como Spark para la detección de spam, análisis de sentimiento, o clasificación de usuarios (bots o humanos).

- Valor: Para la visualización y análisis de los datos procesados, se pueden usar herramientas de Business Intelligence (BI) como Tableau o Power BI. Los modelos de machine learning también son clave para extraer valor, por ejemplo, prediciendo tendencias o segmentando a la audiencia.

Esquema de la solución.

Tabla resumen del análisis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **V de Big Data** | **Caso de uso plataforma X** | **Herramientas aplicables** |  |
|  | Volumen | Cientos de millones de tweets diarios. | HDFS  (Hadoop Distributed File System) |  |
|  | Velocidad | Peaks de tweets durante eventos en vivo. | Apache Kafka, Spark Streaming |  |
|  | Variedad | Texto, imágenes, videos, metadatos de usuario. | MongoDB, Cassandra |  |
|  | Veracidad | Fake news, spam, bots. | Apache Spark  (para Machine Learning) |  |
|  | Valor | Análisis de sentimiento, tendencias, influencers. | Tableau, Power BI |  |
|  |  |  |  |  |