Posgrados
CUCEA
El mejor lugar para el talento

Procesamiento de Grandes Bases de Datos

Actividad 1

Profesor: Dr. Osvaldo Guardado

Alumno: Oscar Fernando Flores Garcia
Código: 323019043

Zapopan, Jalisco a 23 de Agosto del 2023

1. Explica cómo se pueden utilizar las redes sociales en el análisis de Big Data y menciona al menos dos tipos de análisis que se pueden realizar.

Las redes sociales generan cantidades masivas de información por sí mismas.
Sobre estos los ejemplos más claros de análisis son análisis de sentimiento por medio de publicaciones usando NLP (Natural Language Processing) o análisis de tendencias por medio de NLP.

2. ¿Qué es el aprendizaje automático (machine learning) y cómo se aplica en el análisis de Big Data?

Es una forma de IA que permite a un sistema aprender datos y actuar con base a lo aprendido en lugar de aprender mediante la programación explícita.

3. ¿Cuáles son las ventajas y desventajas del almacenamiento distribuido en Big Data?

Algunas ventajas del almacenamiento distribuido es la fiabilidad y disponibilidad de los datos, mayor escalabilidad, mayor tolerancia a fallos.

Algunas desventajas son un mayor costo, mayor complejidad de operación y mantenimiento.

4. ¿Por qué la escalabilidad es un aspecto crucial en los sistemas de Big Data?

La complejidad de la escalabilidad está presente en el procesamiento e ingesta de datos, ¿cómo podemos asegurar que la ingesta de datos no se vea detenida por una subida de carga? Muy probablemente tendremos que escalar horizontalmente para mantener un ritmo de procesamiento de ingesta adecuado.

5. Explica la importancia de la visualización de datos en el análisis de Big Data. Menciona al menos dos tipos de visualización.

La visualización de datos es más fácil para algunos datos y análisis inicial por que es más sencillo comprender un gráfico y encontrar tendencias en este que en un csv con miles o millones de filas.

Ejemplos de visualización incluyen histogramas o gráficos de medidor radial.

6. Que es la variabilidad en Big Data y como se puede obtener valor a partir de los datos?

Es distinto a la variedad y se refiere a que la información cambia durante el tiempo, para obtener valor a partir de datos debemos entender e interpretar los significados reales de los datos.

7. Describe dos aplicaciones prácticas de Big Data y cómo puede afectar el análisis de datos?

Una aplicación práctica puede ser un sistema de monitoreo y recomendación por ejemplo, como Amazon analiza millones de datos para recomendar compras similares o determinar gustos similares.

Otra aplicación es como Netflix puede proveer de diversos medios audiovisuales a la población mundial, asignando el lugar más cercano que tiene el medio y como maneja diferentes streams de datos.

8. ¿Qué significa el término “valor” en el contexto de Big Data y cómo se puede obtener valor a partir de datos?

El valor es la meta final con el big data, se refiere a lo que obtenemos después del análisis y modelado de datos, que es lo importante para nosotros como maestrantes o la organización en la que trabajamos el proyecto de Big Data.

9. Cual es el papel de Spark en el procesamiento de Big Data y qué ventajas ofrece en comparación con otras tecnologías?

Es un engine de analíticas unificado para el procesamiento de datos a gran escala con módulos para realizar operaciones SQL, streaming, machine learning y procesamiento de grafos.
Comparado con Hadoop, Spark es una tecnología más avanzada que Hadoop. Apache Hadoop permite agrupar varios equipos para analizar conjuntos de datos enormes en paralelo con mayor rapidez. Apache Spark utiliza el almacenamiento en memoria caché y una ejecución de consultas optimizada para permitir consultas de análisis rápidas en datos de cualquier tamaño.

10. Enumera tres desafíos éticos y legales relacionados con el uso de Big Data en la toma de decisiones.

Privacidad
Es un reto balancear entre privacidad y uso de los datos, a menor privacidad al usuario típicamente existe una mayor usabilidad de los datos y mayor detalle de estos. A una mayor privacidad hay menos detalles en los datos.

Sesgo de IA
Al tener un volumen muy alto de datos no existe el mismo refinamiento en limpieza de estos y no se pueden identificar sesgos en los data sets. La calidad del aprendizaje automático depende de los datos y si en los datos existe un sesgo el modelo padecerá el mismo.

Consentimiento
En el Big Data, la recolección de datos puede estar consensuada, pero el dueño de los datos puede no conocer que sus datos están siendo utilizados como ingesta a un modelo o estadística.

Referencias

Amazon Web Services. (s.f.). *Hadoop y Spark: diferencia entre los marcos de Apache*. AWS. Consultado el 23 de Agosto, 2023, de https://aws.amazon.com/es/compare/the-difference-between-hadoop-vs-spark/

Geeksforgeeks. (2023, March 6). *Difference between Structured, Semi-structured and Unstructured data*. GeeksforGeeks. Consultado el 18 de Agosto ,2023, de https://www.geeksforgeeks.org/difference-between-structured-semi-structured-and-unstructured-data/

GeeksforGeeks. (2023, May 4). *Advantages of Distributed database*. GeeksforGeeks. Consultado el 21 Agosto, 2023, de https://www.geeksforgeeks.org/advantages-of-distributed-database/

Google. (s.f.). *¿Qué es Hadoop?* Google Cloud. Consultado el 21 Agosto, 2023, de https://cloud.google.com/learn/what-is-hadoop?hl=es

Google. (s.f.). *What is Apache Spark?* Google Cloud. Consultado el 23 Agosto, 2023, de https://cloud.google.com/learn/what-is-apache-spark

IBM. (s.f.). *¿Qué es Machine Learning? - México*. IBM. Consultado el 21 Agosto, 2023, de https://www.ibm.com/mx-es/analytics/machine-learning

Oracle. (s.f.). *What Is Big Data?* Oracle. Consultado el 18 Agosto, 2023, de https://www.oracle.com/big-data/what-is-big-data/