Examen sobre Introducción a Big Data

Duración: 60 minutos

Puntuación máxima: 100 puntos

Pregunta 1: Define qué es Big Data y menciona al menos tres características clave.

Respuesta: Big Data se refiere a conjuntos de datos extremadamente grandes y complejos que son difíciles de procesar y analizar utilizando métodos tradicionales. Sus características clave incluyen volumen, variedad, velocidad, veracidad y valor.

Pregunta 2: ¿Cuáles son las tres "V" originales de Big Data? Explícalas brevemente.

Respuesta: Las "3V" originales de Big Data son:

Volumen: Hace referencia a la cantidad masiva de datos que se generan y recopilan.

Variedad: Se refiere a la diversidad de tipos y formatos de datos que existen.

Velocidad: Hace referencia a la rapidez con la que los datos son generados y procesados.

Pregunta 3: Enumera al menos cinco fuentes de datos que puedan contribuir al volumen de datos en un entorno de Big Data.

Respuesta: Algunas fuentes de datos que pueden contribuir al volumen de datos en Big Data son: redes sociales, sensores de IoT, registros de aplicaciones, transacciones financieras, sistemas de vigilancia, entre otros.

Pregunta 4: Explica la diferencia entre datos estructurados, no estructurados y semi-estructurados en el contexto de Big Data.

Respuesta:

Datos estructurados: Son datos organizados en tablas con filas y columnas, como datos almacenados en bases de datos relacionales.

Datos no estructurados: Son datos que no siguen un formato fijo, como texto libre, imágenes y videos.

Datos semi-estructurados: Son datos que tienen alguna estructura, pero no se ajustan a un esquema fijo, como datos en formato XML o JSON.

Pregunta 5: ¿Por qué el procesamiento en tiempo real es un desafío importante en Big Data? Proporciona un ejemplo.

Respuesta: El procesamiento en tiempo real es un desafío en Big Data porque requiere que los datos sean analizados y utilizados casi de inmediato para tomar decisiones rápidas. Un ejemplo sería el análisis en tiempo real de datos de sensores en un coche autónomo para tomar decisiones de conducción segura en tiempo real.

Pregunta 6: ¿Qué son los modelos sintéticos en Big Data y para qué se utilizan?

Respuesta: Los modelos sintéticos en Big Data son conjuntos de datos generados artificialmente para representar situaciones específicas o ampliar la diversidad de datos disponibles. Se utilizan para probar algoritmos y modelos, realizar análisis exploratorios y realizar pruebas de rendimiento.

Pregunta 7: Describe brevemente la importancia de la variedad de datos en Big Data y cómo puede afectar el análisis de datos.

Respuesta: La variedad de datos en Big Data es importante porque los datos provienen de diversas fuentes y están en diferentes formatos. Esto puede afectar el análisis de datos al requerir herramientas y técnicas específicas para procesar y extraer conocimientos de cada tipo de dato.

Pregunta 8: ¿Cuál es el papel de Hadoop en el procesamiento y almacenamiento de Big Data? Menciona al menos dos de sus componentes principales.

Respuesta: Hadoop es un framework de código abierto ampliamente utilizado en Big Data. Su papel es proporcionar una infraestructura para el procesamiento distribuido y el almacenamiento escalable de grandes volúmenes de datos. Algunos de sus componentes principales son Hadoop Distributed File System (HDFS) para almacenamiento distribuido y MapReduce para el procesamiento paralelo.

Pregunta 9: ¿Qué significa el término "veracidad" en el contexto de Big Data y por qué es importante?

Respuesta: La veracidad en el contexto de Big Data se refiere a la calidad y confiabilidad de los datos. Es importante asegurarse de que los datos utilizados para análisis y toma de decisiones sean precisos y representativos de la realidad para evitar resultados erróneos o sesgados.

Pregunta 10: ¿Cuáles son algunos de los desafíos de seguridad y privacidad que se enfrentan en el manejo de Big Data? Proporciona ejemplos.

Respuesta: Algunos desafíos de seguridad y privacidad en Big Data incluyen la protección de datos sensibles, la identificación y prevención de brechas de seguridad y el cumplimiento de regulaciones de privacidad. Un ejemplo es la filtración de información personal en bases de datos de empresas, lo que podría resultar en el robo de identidad o el uso indebido de datos.

Pregunta 11: 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.

Respuesta: Las redes sociales pueden ser una fuente rica de datos para el análisis de Big Data. Se pueden utilizar para realizar análisis de sentimiento, donde se evalúa la opinión de los usuarios sobre un producto o tema. También se pueden emplear para identificar tendencias y patrones en el comportamiento de los usuarios.

Pregunta 12: ¿Qué es el aprendizaje automático (machine learning) y cómo se aplica en el análisis de Big Data?

Respuesta: El aprendizaje automático es una rama de la inteligencia artificial que se enfoca en desarrollar algoritmos que permiten a las máquinas aprender y mejorar su rendimiento a partir de datos sin una programación explícita. En el análisis de Big Data, el aprendizaje automático se utiliza para identificar patrones, realizar predicciones y clasificar datos en grandes conjuntos.

Pregunta 13: ¿Cuáles son las ventajas y desventajas del almacenamiento distribuido en Big Data?

Respuesta: Las ventajas del almacenamiento distribuido en Big Data incluyen una mayor capacidad de almacenamiento, escalabilidad y redundancia, lo que brinda mayor confiabilidad y disponibilidad de datos. Sin embargo, las desventajas pueden incluir una mayor complejidad en la administración y la necesidad de abordar problemas de consistencia en datos distribuidos.

Pregunta 14: ¿Por qué la escalabilidad es un aspecto crucial en los sistemas de Big Data?

Respuesta: La escalabilidad es crucial en los sistemas de Big Data porque estos deben ser capaces de manejar volúmenes masivos de datos que pueden crecer rápidamente con el tiempo. Los sistemas escalables pueden adaptarse a la demanda y continuar proporcionando un rendimiento óptimo a medida que aumenta la carga de trabajo.

Pregunta 15: Explica la importancia de la visualización de datos en el análisis de Big Data. Menciona al menos dos tipos de visualización.

Respuesta: La visualización de datos es importante en el análisis de Big Data porque permite representar la información de manera gráfica y comprensible, facilitando la identificación de patrones y tendencias. Algunos tipos de visualización incluyen gráficos de barras, diagramas de dispersión, mapas de calor y gráficos de líneas.

Pregunta 16: ¿Qué es la variabilidad en Big Data y cómo puede afectar el análisis de datos?

Respuesta: La variabilidad en Big Data se refiere a la inconsistencia y la falta de estructura en los datos. Esta variabilidad puede dificultar el análisis, ya que diferentes fuentes de datos pueden tener formatos distintos y requerir procesamiento y transformación adicionales antes de poder ser analizados.

Pregunta 17: Describe dos aplicaciones prácticas de Big Data en la industria o la investigación.

Respuesta: Dos aplicaciones prácticas de Big Data son:

Análisis de datos en tiempo real en la industria financiera para detectar fraudes en transacciones y prevenir actividades sospechosas.

Análisis de datos de salud para identificar patrones y tendencias que puedan conducir a la prevención temprana de enfermedades.

Pregunta 18: ¿Qué significa el término "valor" en el contexto de Big Data y cómo se puede obtener valor a partir de los datos?

Respuesta: En el contexto de Big Data, el término "valor" se refiere a la capacidad de los datos para proporcionar información significativa y útil que pueda ser utilizada para tomar decisiones informadas. Para obtener valor a partir de los datos, es necesario realizar análisis y aplicar técnicas de procesamiento y minería de datos para extraer conocimientos y patrones valiosos.

Pregunta 19: ¿Cuál es el papel de Spark en el procesamiento de Big Data y qué ventajas ofrece en comparación con otras tecnologías?

Respuesta: Spark es un framework de procesamiento de datos en tiempo real y análisis de Big Data. Su papel es permitir el procesamiento paralelo y distribuido de grandes conjuntos de datos de manera eficiente y rápida. Spark ofrece ventajas como velocidad de procesamiento, soporte para múltiples fuentes de datos y una API fácil de usar para desarrolladores.

Pregunta 20: Enumera tres desafíos éticos y legales relacionados con el uso de Big Data en la toma de decisiones.

Respuesta: Tres desafíos éticos y legales relacionados con el uso de Big Data son:

Privacidad y protección de datos personales.

Sesgos y discriminación en los algoritmos de aprendizaje automático.

Transparencia y explicabilidad en la toma de decisiones automatizadas basadas en Big Data.