Laureano Mandrile

Big Data-

1. Cuando hablamos de proyecto de Big Data, ¿qué características tienen la cantidad de datos?
2. ¿Enumere las características del Big Data?}
3. ¿A qué hace referencia el tipo de error semántico que nos podemos encontrar en los datos?
4. ¿Por qué muchos de los datos de un proyecto de Big Data no son útiles?
5. ¿Fases del proceso ETL?
6. ¿Con qué tipos de datos podemos trabajar en un proyecto de Big data?
7. ¿Cuál es el proceso que se utiliza para obtener datos en el proyecto Big data?
8. ¿Por qué viene dada la variedad en un proyecto de Big data?
9. ¿Qué tipo de proyecto no son capaces de ser procesados por los sistemas informáticos tradicionales?
10. ¿Qué significan las siglas ETL?
11. ¿Cuántos datos se genera por hora, minuto y segundo en un proyecto de Big Data?
12. Los datos en sí mismos carecen de una cualidad muy importante. ¿Cuál es?
13. ¿A quién afecta la velocidad con la que se generan los datos en el proceso de toma de decisiones?
14. ¿Qué tipo de datos generado por un teléfono inteligente puede ser utilizado en Big data?
15. ¿Cuál es el término empleado para referirse a toda aquella cantidad ingente de datos que no pueden ser siempre procesados por los sistemas informáticos actuales?
16. ¿En qué tipo de base de datos generalmente se usa le proceso de minería de datos?
17. ¿Qué es un almacén de datos?
18. ¿Con que tipo de información se trabaja en minería de datos?
19. Las siglas KDD son referidas a ¿?
20. ¿Qué es la minería de datos?
21. ¿Cuál de las siguientes son algoritmos utilizados en minería de datos?
22. ¿Qué es el proceso de extracción del conocimiento?
23. ¿Qué problema se resuelve al cargar los datos en un almacén de datos corporativos?
24. ¿Cuál es el objetivo de las técnicas de almacenamiento de datos?
25. ¿A qué tema está orientado un data mart?
26. ¿Con que objetivo se emplea el almacén de datos?
27. ¿Qué es un Data Mart?
28. ¿Qué significa que el almacén de datos está integrado?
29. ¿Qué se debe hacer una vez que los datos se han extraído de las fuentes de datos y se han transformado y limpiado?
30. ¿Cómo se llama la sistematización del proceso de identificar el valor en los datos?
31. Suelen caracterizarse por su gran volumen, alta velocidad de generación y amplia variedad de tipos y fuentes.
32. Las características que tienen son:

* **Velocidad:** se refiera a los datos en movimiento, rapidez en la generación, acceso y análisis de datos en su entorno de explotación.
* **Variedad:** tiene que ver con gestionar la complejidad de múltiples fuentes y formatos de datos.
* **Volumen:** esta dimensión se refiere a la cantidad de datos, grandes colecciones de información creadas para diferentes usos y propósitos.
* **Valor:** uso eficiente para sacar el máximo valor a los datos, algoritmos de análisis de gran rendimiento y herramientas para la visualización.
* **Veracidad:** para que las fuentes de datos representen la verdad y no estén alteradas.

1. Se refiere a errores en el significado o interpretación de los datos, que pueden ser causados por datos incorrectos, mal etiquetados, o mal interpretados.
2. Muchos datos pueden ser no útiles debido a su irrelevancia para la pregunta de investigación, baja calidad, la volatilidad de los mismos, o porque son redundantes.
3. ETL significa **Extracción**, **Transformación** y **Carga**. Las fases son:

* Extracción de datos de fuentes diversas.
* Transformación de datos para corregir inconsistencias y formatearlos.
* Carga de datos en el destino final para su análisis.

Hay quien divide las fases de los procesos ETL en cinco pasos, que serían: Extraer, Limpiar (para confirmar la calidad de los datos), Transformar, Cargar y Analizar.

1. Tenemos datos de tipo:

* **Estructurados:** Cualquier dato que se pueda almacenar, acceder y procesar en formato fijo recibe el nombre de dato «estructurado». Durante este periodo de tiempo, el talento en la informática ha logrado mejores resultados en el desarrollo de técnicas de trabajo con ese tipo de datos (donde ya se conoce el formato) y se ha derivado valor. Ej.: bases de datos SQL.
* **No estructurados:** Son cualquier dato de forma desconocida o cuya estructura se clasifica como un dato no estructurado. Además, de ser enorme en tamaño, los datos no estructurados plantean múltiples desafíos con respecto a su procesamiento para derivar valor de ellos. Ej.: textos, imágenes y videos.
* **Semiestructurados:** Los datos semiestructurados pueden contener ambos tipos de datos. Suelen tener un formato que se puede definir, pero el usuario no lo puede comprender fácilmente y requiere el uso de reglas complejas que ayuden a determinar cómo leer cada pieza de la información. Ej.: XML, JSON.

1. El Big Data se genera a través de muchas de las actividades que realizamos a diario. Por ello, las fuentes de datos son verdaderamente diversas: dispositivos GPS, sensores de reconocimiento facial o emails son solo algunos ejemplos.

Es decir, los servicios en línea, sitios web, aplicaciones y muchos otros dispositivos analizan datos constantemente para conseguir que sus servicios sean más eficaces y desarrollar nuevos productos.

1. La variedad viene dada por las múltiples fuentes de datos disponibles hoy en día, como redes sociales, sensores, transacciones comerciales, registros médicos, entre otros.

El Big Data no posee un formato único ni proviene de una sola fuente, sino que puede tratarse de imágenes, correos electrónicos, vídeos. La diversidad en los datos es muy importante para cubrir un rango amplio. Esta diversidad es característica y, a la vez, un reto, puesto que condiciona las metodologías con las que se analizaran los datos.

1. Los proyectos de Big Data que no pueden ser procesados por sistemas informáticos tradicionales a menudo involucran conjuntos de datos extremadamente grandes o complejos, que requieren tecnologías y enfoques más avanzados. Ej.:

* **Protección de especies en peligro de extinción:** El Snow Leopard Trust y la National Conservation Foundation utilizan Big Data para prevenir la caza ilegal de tigres en la India, analizando avistamientos y ubicaciones de caza para identificar zonas de alta actividad ilegal.

1. ETL significa **Extracción**, **Transformación** y **Carga.**
2. La cantidad de datos depende del proyecto, pero estas son estimaciones y promedios encontrados:

* Para el año 2025, se espera que se generen más de 463 exabytes de datos cada día.
* En todas las industrias, las empresas generan aproximadamente 2 zettabytes (2,000,000,000,000,000,000 bytes) de datos al día.
* En 2020, se estimó que cada persona creaba alrededor de 1.7 megabytes de datos cada segundo.

1. La cualidad que a menudo carecen los datos por sí solos es el contexto. El contexto es esencial para comprender el negocio y para interpretar y evaluar los resultados de cualquier proceso de minería de datos. Al inicio de un proyecto, es crucial plantear las preguntas correctas y buscar datos relevantes o determinar qué preguntas pueden responderse con los datos existentes. Al final del proceso, el contexto nos ayuda a interpretar y evaluar los resultados correctamente​​. Además, el contexto de los datos también es clave en la toma de decisiones de negocio, y es tan importante como los modelos y algoritmos utilizados en el análisis de los datos​​.
2. Afecta a líderes y analistas de negocio, ya que una rápida generación de datos puede requerir decisiones más rápidas y adaptativas.
3. Los datos de ubicación, uso de aplicaciones, datos de sensores, registros de llamadas y mensajes, entre otros.
4. Se refiere a menudo como "Big Data" o "datos a gran escala".
5. La minería de datos moderna se basa en la computación virtual y en la nube, así como en las bases de datos in-memory, para gestionar datos de muchas fuentes de manera rentable y escalar según la demanda.
6. Un almacén de datos es una gran base de datos centralizada diseñada para el análisis y la generación de informes.
7. Se trabaja con información no trivial, que es implícita, previamente desconocida y potencialmente útil.
8. Se refiere a Knowledge Discovery in Databases (Descubrimiento de Conocimiento en Bases de Datos).
9. La minería de datos es una técnica asistida por computadora que se utiliza en los análisis para procesar y explorar grandes conjuntos de datos. Gracias a las herramientas y métodos de minería de datos, las organizaciones pueden descubrir patrones y relaciones ocultas en sus datos. La minería de datos transforma datos en bruto en conocimiento práctico. Las compañías utilizan dicho conocimiento para resolver problemas, analizar las consecuencias en el futuro de decisiones empresariales y aumentar sus márgenes de beneficio.
10. Los algoritmos que se usan en minería de datos son:

* **Limpieza de datos:** La información muchas veces procede de distintas fuentes y muchas de ellas poco fiables. Por lo tanto, este tipo de metodología tiene como único objetivo limpiar la información.
* **Feature engineering:** Una vez la información ha sido limpiada es necesario que sea transformada para encajar con el tipo de algoritmo que vamos a emplear.
* **Árboles de decisión:** Los árboles de decisión son un tipo de algoritmo supervisado que genera diferentes decisiones de manera jerárquica dividiendo los datos en subconjuntos según sus características. Este tipo de modelo de machine learning puede ser usado tanto para problemas de clasificación y de regresión.
* **Random forest:** El random forest o bosque aleatorio es un algoritmo que pertenece a los métodos ensambladores (ensamble methods), más concretamente al método de bagging.
* **Máquinas de vectores de soporte o support vector machines:** Las máquinas de vector de soporte son un algoritmo muy usado en problemas de clasificación (aunque también sirven para regresión) dentro del machine learning y de la minería de datos.
* **Técnicas de clusterización:** Las técnicas de clusterización pertenecen al grupo de modelos de aprendizaje autónomo no supervisados ya que no necesitan datos de entrenamiento. Las técnicas de clusterización o agrupación consisten en juntar puntos del dataset en grupos donde sus características sean similares.
* **Naive Bayes:** El método de data mining Naive Bayes basa sus predicciones en el famoso teorema de Bayes. Es un clasificador que supone la independencia o no correlación de las características.
* **Redes Neuronales:** Las redes neuronales son actualmente los algoritmos más populares en inteligencia artificial e ingeniería de datos. Estos modelos hacen uso de una red de neuronas y conexiones que imitan la funcionalidad de las neuronas de nuestro sistema nervioso.

1. Es un proceso iterativo y sistemático utilizado para descubrir conocimientos útiles de grandes volúmenes de datos.
2. Se resuelven inconsistencias y se unifican los datos provenientes de diferentes fuentes para facilitar el análisis.
3. Proporcionar un repositorio centralizado de información histórica para análisis y toma de decisiones.
4. Un Data Mart está orientado a un tema o departamento específico dentro de una organización.
5. Apoyar el proceso de toma de decisiones proporcionando datos integrados y consolidados para el análisis.
6. Es una versión más pequeña y enfocada de un almacén de datos, usualmente dedicada a una sola área de negocio.
7. Significa que los datos de diferentes fuentes han sido combinados y homogeneizados.
8. Los datos deben ser cargados en el almacén de datos para su uso y análisis futuros.
9. Este proceso se conoce como el proceso de extracción del conocimiento o KDD.