Preguntas examen:

¿Cuál es la principal característica de un Data Lake?

Pregunta 1

Respuesta

a.

Es más costoso que un Data Warehouse

b.

Solo admite datos estructurados

c.

Permite almacenar datos en su formato original

d.

Almacena datos procesados exclusivamente

Retroalimentación

La respuesta correcta es: Permite almacenar datos en su formato original

¿Qué tipo de bases de datos utiliza grafos para representar datos?

Pregunta 2

Respuesta

a.

OLAP

b.

Relacionales

c.

Orientadas a grafos

d.

Documentales

Retroalimentación

La respuesta correcta es: Orientadas a grafos

¿Qué técnica de limpieza ETL detecta valores atípicos?

Pregunta 3

Respuesta

a.

Estandarización

b.

Escalado

c.

Discretización

d.

Detección de outliers

Retroalimentación

La respuesta correcta es: Detección de outliers

¿Qué técnica se utiliza para realizar análisis exploratorios de datos?

Pregunta 4

Respuesta

a.

OLTP

b.

Data Lakes

c.

ACID

d.

Data Warehouses

Retroalimentación

La respuesta correcta es: Data Lakes

¿Qué formato es ideal para transmisión de datos en streaming?

Pregunta 5

Respuesta

a.

Parquet

b.

CSV

c.

Avro

d.

JSON

Retroalimentación

La respuesta correcta es: Avro

¿Qué componente gestiona los recursos en un clúster de Hadoop?

Pregunta 6

Respuesta

a.

Namenode

b.

DataNode

c.

YARN

d.

HDFS

Retroalimentación

La respuesta correcta es: YARN

¿Cuál es una ventaja del almacenamiento en la nube?

Pregunta 7

Respuesta

a.

Bajo coste de mantenimiento

b.

Requiere infraestructura física propia

c.

Solo admite datos estructurados

d.

Escalabilidad limitada

Retroalimentación

La respuesta correcta es: Bajo coste de mantenimiento

¿Qué es una arquitectura de tres capas en el contexto ETL?

Pregunta 8

Respuesta

a.

Incluye datos operativos, reconciliados y de presentación

b.

Una arquitectura sin redundancias

c.

Exclusiva para Data Warehouses

d.

Solo almacena datos históricos

Retroalimentación

La respuesta correcta es: Incluye datos operativos, reconciliados y de presentación

¿Qué propiedad implica la replicación en HDFS?

Pregunta 9

Respuesta

a.

Velocidad

b.

Tolerancia a fallos

c.

Capacidad de almacenamiento

d.

Consistencia

Retroalimentación

La respuesta correcta es: Tolerancia a fallos

¿Qué herramienta facilita la ingesta de datos en tiempo real?

Pregunta 10

Respuesta

a.

Amazon Redshift

b.

OLAP

c.

OLTP

d.

Apache Kafka

Retroalimentación

La respuesta correcta es: Apache Kafka

¿Qué tipo de escalabilidad mejora recursos en un nodo existente?

Pregunta 11

Respuesta

a.

Horizontal

b.

Lineal

c.

Vertical

d.

Distribuida

Retroalimentación

La respuesta correcta es: Vertical

¿Qué modelo garantiza la atomicidad, consistencia, aislamiento y durabilidad de transacciones?

Pregunta 12

Respuesta

a.

CAP

b.

OLAP

c.

ACID

d.

BASE

Retroalimentación

La respuesta correcta es: ACID

¿Qué es una transformación ETL típica?

Pregunta 13

Respuesta

a.

Escalado

b.

Estándares de representación

c.

Todas las anteriores

d.

Eliminación de duplicados

Retroalimentación

La respuesta correcta es: Todas las anteriores

¿Qué tipo de escalabilidad implica agregar más nodos a un clúster?

Pregunta 14

Respuesta

a.

Escalabilidad interna

b.

Escalabilidad horizontal

c.

Escalabilidad vertical

d.

Escalabilidad híbrida

Retroalimentación

La respuesta correcta es: Escalabilidad horizontal

¿Cuál es el propósito del procesamiento ETL?

Pregunta 15

Respuesta

a.

Almacenar datos sin procesar

b.

Implementar sistemas de bases de datos

c.

Analizar datos en tiempo real

d.

Extraer, transformar y cargar datos en un almacén

Retroalimentación

La respuesta correcta es: Extraer, transformar y cargar datos en un almacén

¿Qué herramienta de nube ofrece un Data Warehouse escalable?

Pregunta 16

Respuesta

a.

Apache Kafka

b.

Amazon Redshift

c.

Hadoop

d.

Amazon S3

Retroalimentación

La respuesta correcta es: Amazon Redshift

¿Qué técnica de transformación ETL se utiliza para manejar valores perdidos?

Pregunta 17

Respuesta

a.

Imputación

b.

Eliminar registros

c.

Asignar valores fijos

d.

Todas las anteriores

Retroalimentación

La respuesta correcta es: Todas las anteriores

¿Qué sistema es más adecuado para consultas complejas en tiempo real?

Pregunta 18

Respuesta

a.

ETL

b.

OLAP

c.

NoSQL

d.

OLTP

Retroalimentación

La respuesta correcta es: OLAP

¿Cuál es una característica del almacenamiento en bloques?

Pregunta 19

Respuesta

a.

Es menos eficiente que el almacenamiento en archivos

b.

Almacena datos en bloques individuales

c.

Se usa exclusivamente para imágenes

d.

Solo se utiliza en Data Lakes

Retroalimentación

La respuesta correcta es: Almacena datos en bloques individuales

¿Qué propiedad prioriza el modelo BASE en bases de datos distribuidas?

Pregunta 20

Respuesta

a.

Atomicidad

b.

Consistencia

c.

Durabilidad

d.

Disponibilidad

Retroalimentación

La respuesta correcta es: Disponibilidad

Describe cómo la escalabilidad horizontal y vertical impactan el diseño de sistemas distribuidos.

Texto de la respuesta Pregunta 21

La escalabilidad describe la cantidad de carga que puede soportar un sistema distribuido.

La escalabilidad horizontal implica agregar más nodos (más equipos) para distribuir la carga, esto permite que el sistema crezca físicamente y en cuanto a costes de manera flexible y específica. Este tipo puede mejorar la disponibilidad, ya que los datos o la carga de trabajo se distribuyen entre varios nodos y en el caso de fallar hay respaldos que continúan activos. Sin embargo, la coherencia de datos y la sincronización entre nodos puede ser más compleja, lo que requiere soluciones complejas y un poco costosas como particionamiento o replicación de datos.

La escalabilidad horizontal se refiere a mejorar un nodo único aumentando sus capacidades de procesamiento u almacenamiento (memoria, CPU, almacenamiento, etc.), esto supone una solución más simple, pero está limitado por los componentes del servidor y puede volverse costoso a medida que se requieren mejores recursos. Y en el caso de falla no hay

un respaldo que continúe en funcionamiento, es decir, no tiene tanta disponibilidad como en el caso de la horizontal.

Además de todo esto la horizontal cuenta con la ventaja para pequeños proyectos y proyectos de prueba, ya que existe la posibilidad de unir equipos antiguos para realizar trabajos sencillos o de menor capacidad y por un menor costo, como en el caso de una de las actividades propuestas, en la que se unían ordenadores antiguos para formar un clúster.

Analiza los retos y beneficios del uso de almacenamiento en la nube para proyectos de Big Data.

Texto de la respuesta Pregunta 22

El almacenamiento en la nube es una opción eficiente pero que tiene beneficios y retos, estos son los siguientes:

Como beneficios contamos con: escalabilidad ilimitada, ya que tenemos un crecimiento dinámico sin necesidad de infraestructura física y podemos añadir recursos de manera escalable y flexible a nuestro proyecto; menor costo operativo, ya que al no tener hardware propio no necesitamos personal para administrar los servidores o por el mantenimiento físico de estos; disponibilidad y redundancia, debido al servicio contratado podemos tener redundancia geográfica(se almacenan nuestros datos en otros sitios simultáneamente por si se producen pérdidas o fallas) y disponibilidad ya que al no depender de un solo equipo físico, en caso de caía o falla seguiremos teniendo nuestro servicio disponible desde otro servidor, esto hace el sistemas más resiliente; entrada y gestión de datos en tiempo real, los servicios online permiten el procesamiento de datos en tiempo real como en el caso de los dispositivos IOT.

Pero no todo son ventajas, algunas desventajas de el almacenamiento en la nube son los siguientes; costes ocultos, a pesar de reducir los costos iniciales, el aumento de almacenamiento o transferencia de datos puede aumentar la carga de trabajo y necesidades de nuestro sistema contratado lo que resulta en costos adicionales; seguridad y cumplimiento, el almacenamiento en la nube tiene un enorme desafío con la seguridad y el cumplimiento de las normativas establecidas, debido a que las empresas poseen una gran cantidad de datos valiosos, por lo que son responsables de asegurar que sus datos cumplan con regulaciones de privacidad y normativas de protección de datos. Todo esto se dificulta si tenemos en cuenta las normativas específicas de cada país, como pudimos observar en la actividad que investigamos sobre datos estadísticos, en mi caso sobre la tasa de desempleabilidad, donde afectaba la normativa europea y la española sobre la recogida, uso, almacenamiento y tratamiento de los datos; latencia de rendimiento, este es el reto, debido al uso de datos procesados o transferidos a larga distancia del usuario y el servidor accedido puede afectar el rendimiento, especialmente en aplicaciones de Big Data que requieren análisis rápidos y respuestas en tiempo real. Ya que la distancia entre estos trae problemas no solucionables como en el caso de datos más simples de uso habitual como el caso del 5G que permite preprocesar los datos antes de enviarlos debido a la simplicidad y menor carga y volumen de estos.

Explica las diferencias clave entre un Data Warehouse y un Data Lake, incluyendo sus ventajas y limitaciones.

Texto de la respuesta Pregunta 23

Un Data Warehouse y Data Lake son dos tipos de almacenamiento utilizados habitualmente en el entorno Big Dta e Ia.

un Data Warehouse está diseñado para manejar datos estructurados que provienen de diversas fuentes. Estos datos se organizan en tablas y esquemas, y deben pasar por procesos de extracción, transformación y carga (ETL) para garantizar que sean consistentes y fácilmente accesibles para análisis. Mientras que un Data Lake tiene la capacidad de almacenar todos los tipos de datos, incluidos los datos estructurados, semiestructurados (como JSON o XML) y no estructurados (como imágenes o videos). Los datos se almacenan en su formato original, lo que proporciona mayor flexibilidad en el manejo de la información.

El procesamiento de los datos también marca una diferencia importante entre ambos. En un Data Warehouse, los datos se transforman antes de ser cargados en el proceso ETL, lo que implica procesos de limpieza, validación y normalización. Esto asegura que los datos sean mejores para el uso complejo que le queremos dar. Por otro lado, en un Data Lake, los datos se almacenan tal como llegan, sin necesidad de procesamiento previo, lo que permite que los usuarios gestionen y transformen los datos cuando los necesiten. Esta flexibilidad es clave, pero también implica que los datos crudos puedan ser más difíciles de utilizar sin las transformaciones necesarias. por lo que un Dta Warehouse es mejor para generación de informes empresariales y análisis históricos de datos. Por otro lado, los data Lakes son más adecuados para proyectos de Big Data que necesitan almacenar y procesar grandes volúmenes de datos en tiempo real o trabajar con datos no estructurados, como logs, imágenes o videos.

En conclusión los Data Warehouse ofrecen varias ventajas, como la optimización para consultas complejas y el hecho de que los datos se transforman antes de ser almacenados, lo que asegura su calidad y consistencia pero tienen como desventajas su alto costo para escalar y la falta de flexibilidad, mientras que los Data Lakes proporcionan flexibilidad al manejar cualquier tipo de dato, lo que los convierte en una excelente opción para almacenar grandes volúmenes de datos. Sin embargo, también tienen desafíos como la posible mala calidad de los datos debido a la falta de procesamiento previo, lo que puede generar datos desorganizados y difíciles y más costosos de analizar. Además, el riesgo de que el Data Lake se convierta en un "pantano de datos" sin estructura clara es una limitación importante.

Reflexiona sobre la importancia de la limpieza de datos en el proceso ETL y cómo impacta en el análisis de datos.

Texto de la respuesta Pregunta 24

EL proceso ETL (Extracción, transformación y carga) es crucial para garantizar la calidad y precisión de los análisis de datos, durante la extracción y transformación, los datos pueden contener errores, duplicados, valores nulos o inconsistentes que, si no se corrigen, pueden

distorsionar los resultados finales, pero existen algunas técnicas que nos pueden ayudar con esto, como eliminar ese campo, hacer la media de otros campos para ese valor o predecir el valor a través del resto de datos. Unos datos limpios permiten que las herramientas de análisis generen conclusiones más fiables y útiles, facilitando la toma de decisiones empresariales. Si los datos son de mala calidad, los modelos y análisis pueden ser imprecisos o completamente erróneos, lo que afecta la eficacia de las estrategias de negocio.

En resumen, una adecuada limpieza de datos es fundamental para asegurar que el análisis sea basado en información precisa y fiable, lo que maximiza una de las 5V del Big Data, el valor de los datos.

Discute cómo el procesamiento OLAP complementa el OLTP en el análisis empresarial

Texto de la respuesta Pregunta 25

El procesamiento OLAP (procesamiento online vertical) y OLTP (transacción de procesos online) son dos tecnologías complementarias en el análisis empresarial. OLTP se centra en la gestión de transacciones diarias, como la entrada y actualización de datos en tiempo real, y es ideal para operaciones de negocio cotidianas, mientras que OLAP está diseñado para realizar consultas complejas y análisis con varios aspectos, permitiendo a las empresas explorar grandes volúmenes de datos históricos y generar informes detallados.

Mientras OLTP proporciona la información operativa actual, OLAP transforma esa información en informes analíticos útiles para la toma de decisiones estratégicas a nivel histórico y futuro.

Analiza el papel de las herramientas NoSQL en comparación con bases de datos relacionales en aplicaciones modernas.

Texto de la respuesta Pregunta 26

Las herramientas NoSQL a diferencia de las SQL, tienen mayor flexibilidad, ya que permiten almacenar datos semiestructurados o no estructurados como JSON, XML o datos a tiempo real, esto los hace muy útiles en aplicaciones como las redes sociales o el análisis Big Data, sin embargo para datos de gran valor hasta donde he podido observar, se siguen utilizando bases de datos SQL ya que siguen siendo fundamentales para garantizar la consistencia y el control transaccional.

Reflexiona sobre los desafíos éticos y legales que plantea la gestión masiva de datos en la actualidad.

Texto de la respuesta Pregunta 27

La gestión masiva de datos plantea varios desafíos, especialmente en un contexto donde la privacidad, la seguridad y el uso responsable de la información son fundamentales, como es

el caso del Big Data y los datos de usuarios. Uno de los principales desafíos éticos es la protección de la privacidad de los usuarios. A medida que las organizaciones recopilan grandes volúmenes de datos personales, existe el riesgo de que esta información se utilice de manera indebida o se exponga sin el consentimiento adecuado (como el caso de Cambridge analytics). Además, el uso de datos personales sin una clara transparencia puede generar descontento en los usuarios.

Desde la perspectiva legal, las normativas sobre protección de datos como el GDPR que tenemos en Europa han sido implementadas para regular el uso y almacenamiento de datos personales. Sin embargo, el cumplimiento de estas leyes puede resultar complejo, especialmente para las empresas globalizadas, ya que las leyes varían entre los países y regiones, un testimonio de un trabajador de Big Data que vino a darnos una charla nos confesó que en la práctica muchas empresas abusan de vacíos legales en estas leyes para no aplicarlas en sistemas que crean actualmente con el fin de abaratar costes y asegurar una contratación futura para actualizar el sistema, por lo que una mala implementación o una implementación tardía de las leyes también plantea un desafío. Otro desafío legal es la seguridad de los datos, ya que las brechas de seguridad y los ciberataques pueden resultar en la filtración o robo de información confidencial, que en caso de las empresas supone un valor millonario, para ello se han desarrollado técnicas como la confianza cero, que impide filtrar información fuera de la empresa.

Por otro lado, se puede llegar a problemas de discriminación algorítmica, donde los sistemas de inteligencia artificial pueden perpetuar sesgos existentes en los datos. Esto plantea dilemas éticos sobre la justicia y la equidad en el uso de tecnologías que afectan la vida de las personas. Un claro ejemplo de esto es la actividad de análisis de sentimientos, en la que recogían mensajes de la plataforma Twitter que posee un gran sesgo ideológico en cuanto al bando político desde su dueño a muchos de sus usuarios.

En resumen, la gestión masiva de datos implica un equilibrio delicado entre la innovación tecnológica, la protección de derechos y el cumplimiento de normativas, lo que exige a las organizaciones implementar prácticas responsables y transparentes para mitigar los riesgos éticos y legales asociados.