Néstor Batista Díaz

Aplicación práctica de tecnologías Big Data

Pregunta 1: Explica qué es un Datawarehouse y un Datamart, y qué limitaciones más importantes tienen.

Un **Datawarehouse** es un sistema de almacenamiento de datos centralizado diseñado para permitir el análisis y la consulta de grandes volúmenes de información procedente de diversas fuentes dentro de una organización. Proporciona una vista unificada de los datos para apoyar la toma de decisiones. Un **Datamart** es una versión más pequeña y específica de un Datawarehouse que se concentra en un área o función particular dentro de una organización, lo que permite un análisis más detallado y relevante para ese departamento en particular.

Las principales limitaciones de un Datawarehouse incluyen su complejidad y costo de implementación, lo que puede ser prohibitivo para pequeñas y medianas empresas. Además, pueden presentar dificultades en la integración de datos de fuentes heterogéneas y en la actualización de los datos a tiempo real. Por otro lado, los Datamarts, al ser más específicos, pueden crear silos de información que dificultan la visión integral de los datos en la organización.

Pregunta 2: Razona si sería posible montar un Datawarehouse utilizando Hadoop como tecnología.

Es completamente posible utilizar Hadoop como tecnología para montar un Datawarehouse. De hecho, Hadoop es una plataforma de software que facilita el procesamiento distribuido de grandes conjuntos de datos a través de clusters de computadoras utilizando modelos de programación sencillos. Puede manejar no solo datos estructurados sino también no estructurados, lo que lo hace ideal para el almacenamiento y análisis de big data. Por lo tanto, puede servir como base para un Datawarehouse que requiera la flexibilidad de manejar grandes volúmenes de información variada, asegurando escalabilidad y confiabilidad en el almacenamiento y procesamiento de datos.

Pregunta 3: Describe qué pasos debería dar una empresa, una vez que ha decidido desplegar una plataforma de Hadoop en la nube, desde su concepción hasta su uso.

- **Planificación y Análisis:** La empresa debe iniciar con una fase de planificación y análisis, definiendo los objetivos, los requerimientos técnicos, y estimando los costos. Esta fase también involucra elegir el proveedor de la nube adecuado que soporte Hadoop.
- **Configuración del entorno en la nube:** Después de seleccionar el proveedor, se procede a configurar el entorno en la nube que aloja la plataforma Hadoop, incluyendo la creación de instancias de servidor, almacenamiento y la configuración de redes y seguridad.
- **Instalación y Configuración de Hadoop:** Una vez establecido el entorno en la nube, se instala y configura Hadoop, adaptándolo a las necesidades específicas de la empresa. Esto incluye la configuración de HDFS (Hadoop Distributed File System), YARN (Yet Another Resource Negotiator) y otros componentes necesarios.

- **Migración de datos:** Antes de poder utilizar Hadoop, la empresa debe migrar los datos existentes al nuevo entorno. Esto puede implicar la traslación de datos desde varios sistemas y formatos a HDFS.
- **Desarrollo y Pruebas:** Con los datos ya en Hadoop, se procede a desarrollar las aplicaciones analíticas y predictivas necesarias, así como a realizar pruebas para asegurar que todo funciona correctamente.
- **Puesta en marcha y monitoreo:** Una vez que las aplicaciones han sido probadas y ajustadas, la plataforma Hadoop se pone en producción. Es crucial establecer un sistema de monitoreo para garantizar el rendimiento y la seguridad del sistema.
- **Capacitación y Soporte:** Finalmente, es esencial capacitar al personal en el uso de Hadoop y establecer un soporte técnico adecuado para resolver cualquier incidencia que pueda surgir.

Pregunta 4: Razona qué es un entorno multitenancy y explica cómo Hadoop consigue ser una tecnología multitenancy.

Un entorno **multitenancy** o de múltiples inquilinos se refiere a una arquitectura de software en la cual una única instancia de una aplicación sirve a múltiples clientes u organizaciones (tenants). Cada inquilino maneja sus propios datos y configuraciones, aunque la aplicación y el sistema operativo subyacentes son los mismos para todos. Este enfoque maximiza la eficiencia de los recursos y reduce costos, ya que los recursos se comparten entre los inquilinos.

Hadoop consigue ser una tecnología multitenancy a través de su capacidad para gestionar de manera eficiente los recursos entre múltiples usuarios y trabajos simultáneamente. Utiliza **YARN (Yet Another Resource Negotiator)**, que es un administrador de recursos que permite que varios procesos de datos puedan correr en el mismo cluster de Hadoop. YARN asigna los recursos del sistema (memoria, CPU) a los diferentes trabajos en ejecución, asegurando que cada inquilino (usuario o trabajo) tenga los recursos necesarios para su ejecución sin interferir con otros inquilinos. Esto facilita el aprovechamiento compartido del cluster Hadoop, manteniendo al mismo tiempo el aislamiento y la seguridad de los datos de cada inquilino.

Pregunta 5: Describe cuáles son las principales diferencias entre Data Lake y Data Mesh.

Las principales diferencias entre un **Data Lake** y un **Data Mesh** radican en su arquitectura y filosofía de gestión de datos.

Un **Data Lake** es un gran repositorio que almacena grandes volúmenes de datos brutos en su formato original. Su foco está en la centralización de los datos para su posterior análisis y procesamiento, sin importar su estructura o formato. Los Data Lakes facilitan el almacenamiento a bajo costo de datos no estructurados y estructurados para su uso en análisis, machine learning y otras aplicaciones analíticas.

Por otro lado, el **Data Mesh** es un enfoque arquitectónico descentralizado para la gestión de datos, que considera los datos como un producto. Prioriza la interconexión de datos distribuidos, con énfasis en la gobernanza, calidad y accesibilidad de los datos. Cada "dominio" dentro de una organización se encarga de gestionar y compartir sus propios datos como si fueran

productos independientes, promoviendo la autonomía y la facilidad de integración entre los diferentes dominios.

Por lo tanto, mientras que un Data Lake se enfoca en la centralización y el almacenamiento masivo de datos, un Data Mesh promueve una estructura organizativa descentralizada, donde la gestión y la calidad de los datos se mantienen cerca de su fuente y se tratan como productos independientes que pueden ser utilizados por toda la organización.

Pregunta 6: Describe cuáles son las principales diferencias entre Data Lake y Datawarehouse.

Las diferencias principales entre un **Data Lake** y un **Datawarehouse** se centran en el formato de los datos, la flexibilidad, y el propósito de cada enfoque.

- **Formato de datos:** Un Data Lake almacena datos en su formato original, sin importar si son estructurados o no estructurados. Esto permite almacenar una amplia variedad de tipos de datos, desde archivos de texto hasta videos. Por otro lado, un Datawarehouse almacena datos que han sido previamente procesados y estructurados de acuerdo con un esquema específico.
- **Flexibilidad:** Los Data Lakes son altamente flexibles, permitiendo almacenar cualquier tipo de dato y modificar el esquema de datos sobre la marcha. En contraste, los Data warehouses requieren un esquema definido previamente, lo que limita la flexibilidad en cuanto a los tipos de datos que pueden ser almacenados y cómo pueden ser utilizados.
- **Propósito:** Mientras que los Data Lakes están diseñados para almacenar grandes volúmenes de datos brutos para exploración y análisis, los Data Warehouses están

enfocados en el almacenamiento de datos estructurados para query y análisis específico en apoyo de la toma de decisiones empresariales.

En resumen, aunque ambos se utilizan para almacenar grandes volúmenes de datos, un Data Lake es más flexible y permite una mayor variedad de tipos de datos y análisis, mientras que un Datawarehouse está más estructurado y se enfoca en el análisis específico de datos para informar la toma de decisiones.