**Integrantes:**

* **Juan Pablo Vergara**
* **Pablo Rivera**

**Pregunta 1. Desarrollo**

Explique con sus palabras cuáles son las principales diferencias y similitudes entre:

* Una base de datos transaccional
* Un Data Warehouse
* Un Data Lake
* Un Lake House

Proponga un ejemplo y caso de uso para cada una de ellas.

Una base de datos transaccional, como un Data Warehouse predomina el SQL y ambas contienen y almacenan información, sin embargo, una BD transaccional su objetivo es escribir los datos, garantizando su consistencia, por ende, no están procesados, en cambio, Data Warehouse contiene la información ya procesada y su objetivo es más orientado a la reportería.

Una base de datos transaccional puede utilizarse en la gran mayoría de sistemas, por ejemplo, en un ERP que almacene el sistemas de compras de una compañía, desde la solicitud de compra (Cotizaciones), órdenes de compra, recepción de materiales y la factura.

Un Data Warehouse siguiendo el mismo ejemplo anterior, si se procesarán los datos del sistema de compras obteniendo KPI 's, agregaciones, etc. se almacenan en un sistema Data Warehouse para realizar la reportería reactiva como predictiva, llegando a desplegarlo en un BI para la futura toma de decisiones sobre los datos contenidos.

Data Lake con Data Warehouse tienen objetivos en común, ser un gran repositorio de datos, ambos almacenan información en gran escala, sin embargo, de una manera distinta, Data Warehouse almacena información proveniente de sistemas transaccionales que pasan por una normalización y transformación de los datos, por ende, están procesados para un fin, en cambio, Data Lake almacena datos independientemente de la estructura y fuente (Multimedia, datos, etc.) y puede almacenar datos procesados como no procesados.

Un Data Lake puede ser utilizado con el mismo ejemplo anterior, generar reportería, aunque es más factible en empresas que no requieren mucha organización y el origen de los datos no siempre es estructurado, abre el camino a realizar Machine Learning, BI, Reportes, etc.

Data Lake house, combina algunos de los mejores elementos de los Data Lakes y Data Warehouse, en pocas palabras el Data Lake House es una plataforma que contiene todos los datos en un mismo lugar.

Su uso a pesar que es tecnología nueva puede ser factible para diversos usos, ejemplo IA, Machine Learning, BI dado su mix entre las características de un Data Warehouse y Data Lake.

**Pregunta 2. Desarrollo**

Explique con sus palabras cuales son los *drivers* o causas que hacen necesario la utilización de “clústers de computadoras”

Refiérase especialmente al uso de HPC (clústers intensivos en cómputo) vs al uso de clústers para Big Data (tipo Hadoop o Spark)

Proponga un ejemplo o caso de uso para cada uno de ellos.

La necesidad de utilizar clúster de computadoras responde a diversos factores asociados a los datos tales como:

1. Volumen: Las diversas fuentes con las que las empresas recopilan datos, como transacciones comerciales, vídeos, redes sociales, reclamos, etc.

2. Velocidad: Con el crecimiento del Internet de las Cosas, los datos llegan a las empresas a una velocidad sin precedentes. Por ejemplo, los sensores hacen que muchas empresas necesiten administrar estos datos casi en tiempo real.

3. Variedad: Los datos se presentan en todo tipo de formatos: desde datos numéricos estructurados en bases de datos tradicionales hasta documentos de texto no estructurados, correos electrónicos, vídeos, audios y transacciones financieras.

4. Variabilidad: Los flujos de datos tienden a variar mucho más que antes, tanto en velocidad como en volumen, ya sea por ventas peak, por día, por hora, etc.

5. Veracidad: Debido a que los datos provienen de diversas fuentes, resulta complejo verificar su calidad.

Existen distintas tecnologías para procesar grandes volúmenes de datos, entre ellas encontramos HCP (clúster intensivos en cómputo), Hadoop y Spark. Sus principales diferencias radican en su sistema de archivos, arquitectura, modelo de programación y enfoque:

Sistema de archivos: HPC utiliza sistemas de archivos locales o compartidos, mientras que Hadoop utiliza su propio sistema de archivos distribuido, HDFS. Spark también puede utilizar HDFS, pero también puede utilizar otros sistemas de archivos.

Arquitectura: HPC utiliza un conjunto de nodos de computadora de alta potencia que se comunican a través de una red de alta velocidad para realizar cálculos complejos, mientras que Hadoop y Spark utilizan un conjunto de nodos de computadora más simples y menos costosos, que se comunican a través de una red de menor velocidad para procesar grandes conjuntos de datos de manera distribuida.

Modelo de programación: HPC utiliza diferentes modelos de programación según la aplicación específica, como MPI, OpenMP o CUDA. Hadoop utiliza el modelo de programación MapReduce, mientras que Spark utiliza su propio modelo de programación, que es más rápido que MapReduce y utiliza la memoria en lugar de la escritura y lectura en disco.

Enfoque: HPC está diseñado para realizar cálculos intensivos en recursos, como simulaciones científicas, modelos climáticos o análisis numéricos. Hadoop se utiliza comúnmente en aplicaciones de análisis de big data y en el procesamiento de datos en tiempo real, mientras que Spark está diseñado específicamente para procesamiento de datos en memoria y análisis de datos en tiempo real.

Dentro de las aplicaciones HPC podemos encontrar simulaciones científicas, como la simulación del clima, simulaciones de dinámica de fluidos o simulaciones moleculares, ya que requieren un alto rendimiento de procesamiento y cálculo para realizar cálculos intensivos en recursos y modelar sistemas complejos.

Un ejemplo de aplicación de Hadoop es el análisis de registros de servidores web para determinar patrones de uso y comportamiento del usuario, o el análisis de datos de redes sociales para identificar tendencias y patrones de comportamiento.

Spark se utiliza comúnmente en aplicaciones de procesamiento de datos en tiempo real. Por ejemplo, transacciones financieras en tiempo real para detectar fraudes.