**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

1. **base de datos transaccional:** las bases de datos transaccionales están diseñadas de tal manera de que las operaciones de escritura y lectura se ejecutan respetando un alto nivel de integridad, almacenando sus datos como filas permitiendo consultar todos los datos de una entidad en una misma operación. La principal diferencia con el resto de las bases de datos es que al momento de realizar operaciones analíticas como contar clientes que viven en una ciudad X o bajo cierta condición, estas requieren importar más columnas para hacer una consulta. Ejemplo: ventas de productos en un año
2. **Data warehouse:** Los almacenes de datos son sistemas empresariales que sirven para hacer análisis e informes sobre datos estructurados y semiestructurados de varias fuentes, como transacciones de puntos de venta, automatización de marketing, gestión de relaciones con los clientes (CRM) y mucho más. Los almacenes de datos son perfectos tanto para los análisis ad hoc como para los informes personalizados. Constituyen la ubicación única donde almacenar los datos actuales y el historial de datos. Además, como están diseñados para ofrecer una vista de los datos a lo largo de mucho tiempo, son uno de los componentes principales de la inteligencia empresarial.
3. **Data lake:** Los lagos de datos constituyen una plataforma escalable y segura en la que las empresas pueden hacer varias operaciones: ingerir toda clase de datos de cualquier sistema con independencia de su rapidez, incluso datos procedentes de sistemas on‑premise, en la nube o de informática perimetral; almacenar cualquier tipo o volumen de datos con alta fidelidad; procesar datos en tiempo real o por lotes; y analizar datos usando cualquier tipo de lenguaje, como SQL, Python o R, u otras aplicaciones de datos o de analíticas de terceros.
4. **Lake house:** Los lagos de datos son arquitecturas utilizadas para almacenar grandes cantidades de datos desestructurados y semiestructurados que se recopilan en los distintos sistemas, dispositivos y aplicaciones de la empresa. Los lagos de datos emplean habitualmente infraestructuras de almacenamiento de bajo coste con una interfaz de programa de la aplicación (API) que guarda los datos con formatos de archivos genéricos y abiertos. Esto significa que los lagos de datos son útiles para almacenar datos a escala y ponerlos a disposición de los algoritmos de IA y ML, aunque no se ocupan de la calidad de los datos ni de los requisitos de gobernanza. A medida que a los lagos de datos se van añadiendo datos duplicados, irrelevantes y desorganizados debido a una organización o gestión deficiente, estos lagos se convierten en lo que se conoce como pantanos de datos, lo que hace más difícil extraer información relevante de los datos que contienen.

**Existen similitudes entre bases de datos transaccionales y data warehouse, pues el segundo puede contener más de una base de datos transaccional (tienen la misma lógica). También hay similitudes entre data lake y lake house, pues puede tener en una organización distintos lake según su uso (finanzas, abastecimiento, mantenimiento, etc), pero son independientes entre sí para su procesamiento, lo cual es más productivo cuando se trabaja en grandes volúmenes de datos.**

**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.

**Drivers que hacen necesario la utilización de clúster de computadoras:**

- Requerir rendimiento para disminuir tiempos de procesamiento o acelerar entrenamiento de AI

- Requerir disponibilidad de recursos de manera escalable

- Equilibrar carga de procesamiento o prevenir fallas por uso intensivo de recursos, controlar velocidad de procesamiento y distribuir recursos

- Requerir baja latencia para procesos críticos del negocio

**HPC vs SPARK**

HPC utiliza el bandwidth (ancho de banda) de la red e intensivamente el almacenamiento para procesar los datos, en cambio SPARK utiliza la distribución de datos en RAM en cada nodo disponible, es decir, procesa los datos o sus transformaciones en donde se encuentra la información, lo que minimiza el movimiento y transporte de data.

Ejemplo HPC: simulaciones, predicciones de tormenta

Ejemplo SPARK: sistemas de analítica (cálculo de indicadores, segmentación de clientes)