PREGUNTAS DE REPASO PARTES 1 Y 2

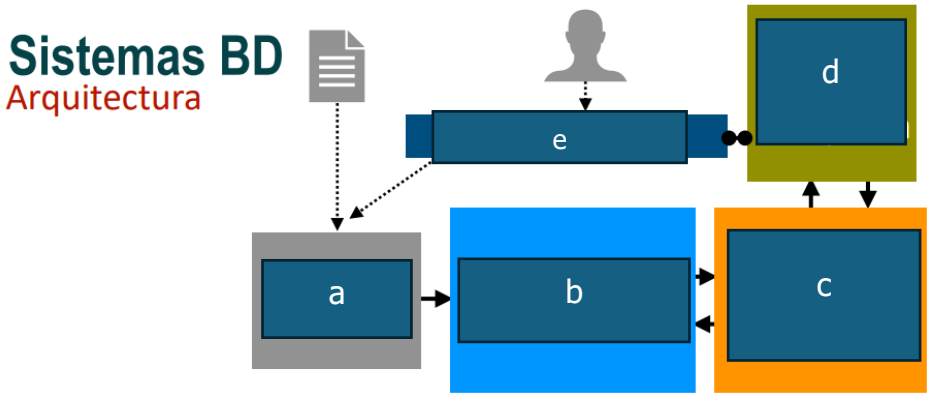
1. **Completa:**

2. **Indica 4 lugares de los que se puede extraer información:**

3. **Indica dos tipos de bases de datos en las que se pueden almacenar la información extraída:**

4. **Especifica distintos tipos de análisis que pueden llevarse e cabo con los datos almacenados en función de la urgencia con la que deben procesarse:**

5. **Una arquitectura prototípica está formada por una serie de elementos. A partir de las funciones de cada uno, indica el nombre del elemento:**



a. **Es el punto de entrada de la información, donde se recopilan y almacenan los datos.**
b. **Toma los datos ingresados y realiza transformaciones o análisis básicos necesarios antes de pasar la información a otros componentes.**
c. **Usa los datos procesados para entrenar modelos y extraer patrones o información valiosa.**
d. **Permite realizar consultas sobre los datos procesados, proporcionando respuestas en tiempo real para facilitar el análisis.**
e. **Es la interfaz de usuario que permite la interacción con el sistema, proporcionando acceso y visualización de los datos y resultados de consultas y análisis.**

6. **Relaciona cada característica con el tipo de procesamiento adecuado (por lotes, streaming o tiempo real):**
    a. **Permite manejar grandes volúmenes de datos de manera eficiente**
    b. **Este método es ideal para aplicaciones que requieren respuestas instantáneas**
    c. **Los datos se procesan en pequeños lotes de manera continua y secuencial a medida que llegan**
    d. **Todos los datos deben ser recopilados y almacenados en la etapa de ingestión de datos antes de iniciar el análisis**
    e. **A medida que los datos se ingieren, se envían a la aplicación de procesamiento, que los analiza al instante**
    f. **El procesamiento se realiza sin la intervención activa del usuario**
    g. **La capa de procesamiento toma todos los datos acumulados y los analiza en una sola ejecución**
    h. **Su análisis es inmediato**
    i. **Se realiza en tiempo cercano al real**
    j. **El motor de búsqueda interviene después de que el procesamiento ha finalizado y los datos han sido analizados**
    k. **Los datos se recopilan de forma continua y se envían a la capa de procesamiento en micro-lotes**
    l. **Su función principal en este contexto es permitir que el usuario consulte y recupere información específica de los resultados ya procesados, facilitando la exploración y el análisis posterior**

7. **Esta arquitectura suele almacenarse en clústeres en la nube. Consecuentemente, surgen algunos problemas. Identifícalos a partir de las siguientes definiciones:**
    a. **Este problema se refiere a la posibilidad de que usuarios no autorizados accedan, copien o visualicen información confidencial. Esto representa un gran riesgo a nivel empresarial y gubernamental, y, por ello, se invierten grandes sumas de dinero en medidas de seguridad para proteger los datos.**
    b. **Ocurre porque la distribución de datos en múltiples servidores o colecciones de ordenadores plantea riesgos de protección si los protocolos de seguridad son deficientes o si el sistema se utiliza incorrectamente. Esto puede resultar en la exposición de información sensible, lo que no solo representa un riesgo de**

seguridad, sino que también puede dañar la reputación de la empresa.

c. Ocurren cuando un componente del sistema no puede manejar el volumen de solicitudes o tareas que recibe, lo cual ralentiza el rendimiento general del sistema. Esto puede suceder, por ejemplo, si muchas tareas se están procesando simultáneamente en un mismo servidor. Para mitigar este problema, los administradores de la infraestructura pueden dimensionar adecuadamente los recursos, replicar hardware y balancear la carga entre diferentes servidores.
d. Surge cuando aumenta la demanda de recursos y el sistema debe adaptarse para manejarla sin pérdida de rendimiento. Puede implicar la necesidad de añadir más recursos en la nube, como almacenamiento y potencia de procesamiento, para asegurar que el sistema crezca de manera eficiente.
e. Se da cuando múltiples tareas se ejecutan simultáneamente y comparten los mismos recursos. Requiere estrategias de sincronización y coordinación eficientes, como bloqueos o algoritmos de control de acceso, para asegurar que los recursos compartidos se usen de manera óptima.
f. En un sistema distribuido, múltiples nodos trabajan juntos para procesar las tareas. Si uno de estos nodos falla, el sistema debe ser capaz de continuar operando sin pérdida de datos. Esto implica implementar mecanismos de redundancia, recuperación automática y liberación adecuada de recursos para asegurar que el fallo de un nodo no afecte el procesamiento general.
g. Implica garantizar que los servidores no se caigan o que ciertos servicios no queden inoperativos.

8. **¿Para qué tipo de procesamiento es óptimo el MapReduce?**

9. **¿Dónde, cuándo y por qué se originó MapReduce?**

10. **¿Cómo se llamó la BD y el sistema de archivos?**

11. **Basándose en MapReduce, ¿qué empresa creó Hadoop MapReduce que, más tarde, se conoció como YARN MapReduce?**

12. **¿Cómo se llamaban ahora la BD y el sistema de archivos?**

13. **¿Cómo se llamaron finalmente estos dos proyectos creados por Apache y qué características fundamentales tienen?**

14. **¿Qué 3 etapas conoces del MapReduce y en qué formato están los datos producidos por cada una?**

15. **¿Qué 2 funciones son las que debe definir el usuario?**