Tarea para SBD01.

Título de la tarea: Prediseño de un sistema Big Data

Curso de Especialización: Inteligencia Artificial y Big Data Módulo profesional: Sistemas de Big Data

1.- Descripción de la tarea.

Caso práctico

La empresa **Construcciones D8** se ha puesto en contacto con la empresa consultora en la que trabajas para que les realicéis un prediseño de lo que sería un sistema Big Data para resolver las siguientes necesidades.

- ✓ Hay distintas fuentes externas a su empresa que producen datos interesantes para ellos y les interesaría poder conectarse a ellas para obtenerlos.
- ✓ Esas fuentes tienen conjuntos de datos estáticos o que se actualizan anualmente.
- ✓ Además hay fuentes internas de la propia empresa que generan datos de forma continua y hay que irlos obteniendo sobre la marcha.
- ✓ La cantidad de datos actualmente es de aproximadamente 500TB, y calculan que se producen otros 100TB nuevos cada año.
- ✓ Quieren poder mantener almacenados todos esos datos de modo no se pierdan y además accesibles en todo momento.
- ✓ Se realizan transacciones debido a la interacción con clientes en el día a día.
- ✓ La junta directiva se reúne una vez al mes y quiere poder acceder a un cuadro de mandos para ver analíticas descriptivas que empleen todos los datos que estuviesen disponibles una semana antes de reunirse. Tales analíticas deben ser interactivas, siendo los directivos capaces de realizar filtrados de información de modo que las gráficas mostradas se actualicen según la información seleccionada.
- ✓ Quieren poder decidir a qué clientes ofrecerles ciertas ofertas en función de lo que se sabe de su comportamiento pasado.

¿Qué te pedimos que hagas?

- **Apartado 1: Prediseño un sistema para Big Data**

  Crea un documento en el que explicas cómo sería el sistema a emplear para resolver las necesidades Big Data del supuesto práctico. Deberás:

  - Indicar qué habrá que hacer para ir aumentando la capacidad del clúster según se reciben nuevos datos.

    Cuando llegue el momento, se debe realizar un escalado horizontal del clúster, añadiendo nuevos nodos (nuevos servidores con su almacenamiento y memoria propios) para poder optimizar las tareas de flujos de datos y su persistencia.

  - Indicar qué capas de la arquitectura Big Data necesitarán estar presentes como mínimo en el sistema a crear.

    - La capa de ingestión, para adaptarse a las diversas fuentes desde las que obtendremos los datos, usando su protocolo y pudiendo interpretarlos.
    - La capa de colección, para unificar los formatos de información recibida.

- La capa de almacenamiento, inevitable para crear el sistema.
- La capa de procesamiento, para elegir la estructura (en tiempo real para este caso).
- La capa de consulta y analítica para obtener datos y analizarlos.
- La capa de visualización para interactuar con el usuario

Adicionalmente, serían convenientes las capas de seguridad y de monitorización para aportar capas de protección, gestión y control.

- Indicar si alguna parte del sistema necesitará cumplir con las características ACID.

Claro, vamos a tener que trabajar con datos de forma transaccional, por lo que el principio ACID estará presente en los procesos de consulta y analítica de datos.

- Indicar si será necesario un subsistema OLTP.

Si. Necesitamos de un subsistema OLTP, pues está orientado a transacciones de datos contra bases de datos relacionales, que según el enunciado, se realizan a diario. Esto nos permitirá tiempos de respuesta de menos de un segundo que puedan ser mal percibidos por el cliente.

- Indicar si será necesario un subsistema OLAP.

Volviendo al enunciado, la Junta Directiva debe poder hacer consultas analíticas, por lo que es necesario que convivan los dos tipos de subsistema, el OLTP para los procesos de inmediatez con las transferencias de datos con los clientes, y el OLAP para las consultas de analítica a través de bases de datos multidimensionales, optimizadas para responder a consultas complejas (relacionales o transaccionales) en un tiempo muy corto.

- Indicar si habrá un almacén de datos.

Creo que es necesario mantener un almacén de datos como repositorio centralizado de toda la información guardada, y así poder realizar consultas analíticas (aprovechando que hemos mantenido para ello tablas bajo un subsistema OLAP.

- Indicar qué estrategia de procesamiento habrá que emplear para poder crear el cuadro de mandos que quiere la junta directiva.

Una estrategia de procesamiento “en tiempo real”, es la más indicada para este supuesto, pues permite mostrarnos resultados de forma rápida en consultas analíticas interactivas.

Una vez más, los datos almacenados en memoria en subsistemas OLAP, permitirán responder en corto espacio de tiempo.

- Indicar si será necesario crear modelos predictivos a partir de los datos.

Si. El último punto del enunciado, nos indica que quieren decidir sobre los clientes a los que ofrecer distintas ofertas. Para ello, es necesario crear modelos predictivos entre los datos más actuales, mediante técnicas de minería de datos.