Apartado 1: Prediseño un sistema para Big Data

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.

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

Para aumentar capacidad del clúster se deberá hacer un *escalado horizontal* (scale-out), esto se consigue añadiendo más nodos a un clúster, consiguiendo mejor rendimiento del mismo, aunque como contraposición esto conlleva un gran trabajo de diseño y reimplantación.

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

Revisando las múltiples capas de Big Data podemos decir que las capas necesarias como mínimo serian:

- **Capa de ingestión**: Los datos se obtienen desde múltiples fuentes con las que es necesario conectarse. Es necesario que esta capa se adapte a las fuentes y no a la inversa.
- **Capa de colección**: Una vez obtenidos los datos hay que darles una estructura. Hay que unificarlo para representarlo como un único conjunto de datos.
- **Capa de almacenamiento**: Tenemos que almacenar esos datos. Para ello utilizaremos sistemas de almacenamiento distribuido.
- **Capa de procesamiento**: Esta capa provee de infraestructura a la siguiente, necesaria para analizar grandes volúmenes de datos. Se recomendaría en tiempo real ya que se generan datos de forma continua.
- **Capa de consulta y analítica**: Se realiza la estadística, algoritmia o análisis que se considere basándose en la capa previa.
- **Capa de visualización**: Interacciona con el usuario final donde puede consultar los datos o acceder a cuadros de mando interactivos.

Bajo mi punto de vista y ciñéndome a la pregunta estás capas serían lo mínimo, pero es cierto que sería mas que recomendable utilizar las siguientes.

- **Capa de seguridad**: Protección de los datos.
- **Capa de monitorización**: Da soporte al sistema supervisando el estado de los datos…

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

Ya que vamos a estar realizando transacciones con clientes en el día a día, la respuesta es **sí**, estas transacciones deberán cumplir las propiedades **ACID** en la capa de consulta y analítica:

- Atomicidad (**A**tomicity).
- Consistencia (**C**onsistency).
- Aislamiento (**I**solation).
- Durabilidad (**D**urability).

4. **Indicar si será necesario un subsistema OLTP.**

Según la definición de OLTP y el caso práctico la respuesta es **sí**, ya que *OnLine Transaction Processing* esta orientado a transacciones con datos operacionales día a día. Normalmente contra bases de datos relacionales, incluyendo tareas como (insertar, eliminar…)

5. **Indicar si será necesario un subsistema OLAP.**

Ya que la junta directiva quiere poder acceder a un cuadro de mandos para ver analíticas que empleen todos los datos disponibles una semana antes de reunirse y que además deberán ser interactivas… La respuesta es **sí**, esto ayudará a responder consultas complejas en poco tiempo.

6. **Indicar si habrá un almacén de datos.**

La respuesta es **sí**, ya que necesitan tener almacenados todos los datos de modo que no se pierdan y accesibles en todo momento dado que realizan transacciones muy a menudo.

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

Sin lugar a dudas sería estrategia en **tiempo real** ya que se utiliza para analíticas interactivas en las que la respuesta deba ser rápida. Además, para realizarla con un gran volumen de datos se suele emplear con subsistemas OLAP y dado que ya necesitamos este subsistema para el cuadro de mandos para la junta directiva sin duda esa sería la mejor estrategia.

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

La respuesta es **sí**, dado que necesitan poder decidir a que clientes ofrecer ciertas ofertas en función de su comportamiento pasado.