Tarea Sistemas de Big Data 01 David Carlón

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.*

Aumento de capacidad del clúster.

Para poder aumentar la capacidad del clúster, lo mas interesante seria optar por un modelo de escalado horizontal, que permite añadir mas nodos al clúster y aumentar su capacidad.

Capas de arquitectura Big Data necesarias.

Capa de ingestión: Dado que hay diferentes fuentes externas de datos, esta capa seria indispensable, teniendo que adaptarse a las fuentes y empleando los protocolos necesarios para interpretar los datos recibidos.

Capa de colección: Teniendo en cuenta que son de diferentes fuentes, deberíamos tener esta capa para unificarlos, representarlos y posteriormente usarlos.

Capa de almacenamiento: Necesaria para almacenar los 500TB de datos que hay inicialmente, y poder aumentarla 100TB anualmente. Podremos optar por un sistema HDFS y con el escalado horizontal nos aseguramos de que sea escalable y pueda crecer fácilmente. También puede optarse por una solución de almacenamiento en la nube como AWS o Google Cloud.

Capa de Procesamiento: Indispensable para poder proveer de infraestructura a la siguiente capa (la de consulta y analítica) necesaria también en este caso, podemos utilizar Apache Spark para un procesamiento distribuido y paralelo.

Capa de consulta y analítica: Esta capa es necesaria para hacer que los datos sean accesibles, interpretables y útiles para la junta directiva.

Capa de visualización: Con todas las capas anteriores en marcha, es necesario esta capa para ver de manera clara la información obtenida, podemos implementar un sistema de visualización como Tableau para crear cuadros de mando interactivos.

Capa de seguridad y monitorización: Para garantizar la integridad, confidencialidad y disponibilidad de los datos son necesarias estas dos capas.

ACID

La capa de almacenamiento, especialmente para datos internos y transaccionales, podría requerir características ACID. Se podría considerar una base de datos relacional tradicional o bases de datos NoSQL que admitan transacciones.

OLTP

Si, será necesario un sistema de OLTP para gestionar transacciones diarias con clientes y para manejar datos en tiempo real generados por fuentes internas.

OLAP

Sí, para analíticas descriptivas y la creación de informes interactivos que requiere la junta directiva, se necesita un subsistema OLAP. Herramientas como Apache Kylin o Apache Druid pueden ser útiles para realizar análisis OLAP en grandes conjuntos de datos**.**

Almacén de datos.

El sistema debe incluir un almacén de datos (data warehouse), para almacenar datos históricos. Estos datos se utilizaran generar modelos predictivos y decidir a que clientes ofrecer ofertas en función de su comportamiento. Pudiendo utilizar soluciones como Amazon Redshift, Google Bigquery o Snowflake.

Estrategia de procesamiento para creara el cuadro de mandos

La estrategia de procesamiento que propongo sería realizar procesos semanales de captura y organización de datos, guardar esta "foto" de la información en un almacén analítico como Amazon Redshift o Google BigQuery, y luego conectar la herramienta de visualización, por ejemplo, Tableau, directamente a estos datos actualizados cada semana. Esto nos permitirá crear un cuadro de mandos interactivo con filtros que los directivos pueden personalizar antes de sus reuniones mensuales. El énfasis aquí está en el procesamiento por lotes, ya que no necesitamos datos en tiempo real. Podemos usar herramientas como Apache Spark para manejar eficientemente el procesamiento de grandes cantidades de datos y garantizar que el cuadro de mandos siempre muestre información actualizada y relevante.

Creación de modelos predictivos.

Considerando las necesidades de Construcciones D8, sería beneficioso la implementación de modelos predictivos mediante herramientas como TensorFlow y scikit-learn. Estas plataformas brindan capacidades avanzadas para construir algoritmos predictivos, siendo especialmente útiles para tomar decisiones informadas sobre ofertas personalizadas a clientes basadas en su comportamiento pasado. La aplicación de estos modelos podría mejorar la eficiencia operativa y optimizar las transacciones diarias, al tiempo que permitiría cumplir con los objetivos estratégicos de la empresa. La utilización de estas técnicas avanzadas, como parte de la estrategia de análisis de datos, podría potenciar la capacidad de ofrecer soluciones más personalizadas y estratégicas, alineándose con los objetivos de la empresa y mejorando la toma de decisiones a nivel directivo. No obstante, es esencial evaluar cuidadosamente la viabilidad y el valor añadido de los modelos predictivos en relación con los objetivos y recursos disponibles.