Participante

Edward Neftalí Liriano Gómez 2022-0437

Profesor

Francis Ramírez

Asignatura

Electiva 1 (Big Data)

Tema

Resumen de los conceptos básicos

Índice

Introducción a la ingeniería de datos

La ingeniería de datos es un campo fundamental en el manejo y procesamiento de datos a gran escala. Consiste en diseñar, construir y mantener sistemas que permiten recopilar, almacenar y analizar grandes volúmenes de información. Un ingeniero de datos es el responsable principal de integrar, transformar y consolidar datos de varios sistemas de datos estructurados y no estructurados en estructuras adecuadas para crear soluciones de análisis.

Hay tres tipos principales de datos con los que un ingeniero de datos trabajará:

1. **Estructurados:** El elemento principal de un archivo estructurado es que las filas y columnas se alinean de forma coherente en todo el archivo. Formato común(CSV).

2. **Semiestructurados:** Los datos semiestructurados son datos como archivos de notación de objetos JavaScript (JSON), que pueden requerir acoplamiento antes de cargarlos en el sistema de origen.

3. **Datos no estructurados**: Los datos no estructurados incluyen datos almacenados como pares clave-valor que no cumplen los modelos relacionales estándar.

Algunas de las tareas principales que realizará un ingeniero de datos incluyen integración de datos, transformación de datos y consolidación de datos.

- La integración de datos implica establecer vínculos entre los servicios operativos y analíticos y los orígenes de datos para permitir el acceso seguro y confiable a los datos en varios sistemas.
- Los datos operativos normalmente deben transformarse en una estructura y formato adecuados para el análisis, a menudo como parte de un proceso de extracción, transformación y carga
- La consolidación de datos es el proceso de combinar datos extraídos de varios orígenes de datos en una estructura coherente, normalmente para admitir análisis e informes.

Exploración de los conceptos de los datos principales

Los datos pueden recopilarse de manera más fácil y almacenarse de forma más barata, lo que permite que casi todas las empresas puedan tener acceso a ellos. Las soluciones de datos incluyen tecnologías de software y plataformas que pueden facilitar la recopilación, el análisis y el almacenamiento de información valiosa.

El formato de archivo específico que se usa para almacenar datos depende de una serie de factores, entre los que se incluyen los siguientes:

1. El tipo de datos que se almacenan (estructurados, semiestructurados o no estructurados).
2. Las aplicaciones y los servicios que tendrán que leer, escribir y procesar los datos.
3. La necesidad de que los archivos de datos sean legibles para los usuarios o estén optimizados para un almacenamiento y procesamiento eficientes.

Tipos de formatos:

Archivos de texto delimitado: Los datos se almacenan como texto sin formato con delimitadores de campo y terminadores de fila específicos. El formato más común para los datos delimitados son los valores separados por comas (CSV), en los que los campos están separados por comas y las filas finalizan con un retorno de carro o una nueva línea.

Notación de objetos JavaScript: JSON es un formato omnipresente en el que se usa un esquema de documento jerárquico para definir entidades de datos (objetos) que tienen varios atributos. Cada atributo puede ser un objeto (o una colección de objetos), lo que hace de JSON un formato flexible adecuado tanto para datos estructurados como semiestructurados.

Objeto binario grande: Todos los archivos se almacenan como datos binarios (1 y 0), pero en los formatos legibles que se describen anteriormente, los bytes de datos binarios se asignan a caracteres imprimibles (normalmente a través de un esquema de codificación de caracteres como ASCII o Unicode). Aun así, algunos formatos de archivo, especialmente para los datos no estructurados, almacenan los datos como datos binarios sin formato que las aplicaciones deben interpretar y representar.

Procesamiento de datos transaccionales para base de datos:

Los sistemas transaccionales suelen ser de gran volumen; a veces, controlan muchos millones de transacciones en un solo día. Se debe poder acceder a los datos que se procesan con mucha rapidez. El trabajo que realizan los sistemas transaccionales a menudo se conoce como procesamiento de transacciones en línea (OLTP).

Estas operaciones se aplican transaccionalmente, de una forma que garantiza la integridad de los datos almacenados en la base de datos. Para ello, los sistemas OLTP aplican transacciones que admiten la denominada semántica ACID:

- **Atomicidad**: cada transacción se trata como una unidad única, la cual se completa correctamente o produce un error general. Por ejemplo, una transacción que conlleve el adeudo de fondos de una cuenta y el abono de la misma cantidad en otra debe completar ambas acciones. Si alguna de las acciones no se puede completar, se debe producir un error en la otra.
- **Coherencia**: las transacciones solo pueden pasar los datos de la base de datos de un estado válido a otro. Para continuar con el ejemplo anterior del adeudo y el abono, el estado completado de la transacción debe reflejar la transferencia de fondos de una cuenta a la otra.
- **Aislamiento:** las transacciones simultáneas no pueden interferir entre sí y deben dar lugar a un estado coherente de la base de datos. Por ejemplo, mientras la transacción para transferir fondos de una cuenta a otra está en proceso, otra transacción que comprueba el saldo de las cuentas debe devolver resultados coherentes. Es decir, la transacción de comprobación del saldo no

puede recuperar un valor para una cuenta que refleje el saldo *antes* de la transferencia y un valor para la otra cuenta que refleje el saldo *después* de la transferencia.

- **Durabilidad**: cuando se ha confirmado una transacción, permanece confirmada. Una vez que la transacción de transferencia de la cuenta se ha completado, los saldos revisados de las cuentas se conservan, de modo que, incluso si el sistema de base de datos se desactiva, la transacción confirmada se refleje cuando se vuelva a activar.

Procesamiento de datos analíticos

El procesamiento de datos analíticos usa sistemas de solo lectura (o *principalmente* de lectura) que almacenan grandes volúmenes de datos históricos o métricas empresariales. Los análisis pueden basarse en una instantánea de los datos en un momento concreto o en una serie de instantáneas.

Los detalles específicos de un sistema de procesamiento analítico pueden variar según la solución, pero una arquitectura común para el análisis a escala empresarial tiene el siguiente aspecto:

1. Los datos operativos se extraen, transforman y cargan (ETL) en un lago de datos para su análisis.
2. Los datos se cargan en un esquema de tablas: normalmente en un *almacén de lago de datos* basado en Spark con abstracciones tabulares en los archivos del lago de datos o en un *almacenamiento de datos* con un motor SQL totalmente relacional.
3. Los datos del almacenamiento de datos se pueden agregar y cargar en un modelo de procesamiento analítico en línea (OLAP) o un *cubo*. Los valores numéricos agregados (*medidas*) de las tablas de hechos se calculan para intersecciones de *dimensiones* a partir de tablas de dimensiones. Por ejemplo, los ingresos de ventas podrían sumarse por fecha, cliente y producto.
4. Los datos del lago de datos, el almacenamiento de datos y el modelo analítico se pueden consultar para generar informes, visualizaciones y paneles.

Exploración de roles y servicios de datos

Hay una amplia variedad de roles implicados en la administración, el control y el uso de datos. Algunos roles están orientados a los negocios, mientras que otros implican más ingeniería. La organización puede definir roles de maneras distintas o asignarles nombres diferentes, pero los que se describen en esta unidad resumen la clasificación más habitual de las tareas y las responsabilidades.

La ingeniería de datos implica diversos roles y servicios esenciales:

- Ingeniero de datos: Diseña y optimiza pipelines de datos.
- Científico de datos: Analiza datos para obtener conocimiento.
- Analista de datos: Interpreta datos para generar informes.

Identificación de los servicios de datos

Microsoft Azure es una plataforma de nube que usan las aplicaciones y la infraestructura de TI de algunas de las organizaciones más grandes del mundo. Incluye numerosos servicios para admitir soluciones en la nube, incluidas cargas de trabajo de datos transaccionales y analíticos.

Entre los servicios claves se encuentran:

- Extracción, Transformación y Carga (ETL): Procesos para limpiar y preparar datos.
- Infraestructura en la nube: Herramientas como AWS, Azure y Google Cloud Platform para almacenar y procesar datos.
- Seguridad de Datos: Protección contra accesos no autorizados y cumplimiento de regulaciones.

Conclusión

La ingeniería de datos juega un papel esencial en el manejo del volumen y complejidad de los datos modernos. Comprender sus conceptos básicos, roles y servicios permite aprovechar los datos de manera eficiente y garantizar que las empresas puedan tomar decisiones informadas. Este campo sigue creciendo rápidamente, marcando la pauta para el futuro de la tecnología.

Además, la evolución de herramientas y técnicas dentro de este campo promete facilitar aún más el manejo de datos a gran escala. Innovaciones como la inteligencia artificial y el aprendizaje automático están cada vez más integradas en los procesos de ingeniería de datos, automatizando tareas y proporcionando análisis más profundos. Las organizaciones que invierten en este campo no solo optimizan sus operaciones, sino que también ganan una ventaja competitiva significativa al utilizar datos para predecir tendencias, mejorar la experiencia del cliente y responder de manera proactiva a los cambios del mercado.