10/11/2023

Unit 05 – Storage

Tipos de archivos

* Estructurados: como CSV o SQL
* Semiestrucurados: JSON
* No estructurados: imagenes, texot sin formato
* Logs: ELK

Formatos de archivos communes en Big Data

* Avro: comunicación entre componenetes big datas, para la serialización
* Parquet: formato columnar. Se almacena por columnas
* ORC: formato columnar. Se almacena por columnas
* Sequence: Secuencias binarias

Almacenamiento en filas vs columnas

* Descripcion:
* Eficiente: para actualizaciones individuales es mas eficiente almacenamiento pro filas. El almacenamiento por columnas es mas eficiente para operaciones y analisis de datos
* No eficiente: al reves de lo de arriba
* Adeuacdo: almacenamiento por filas es adecuado para transacciones. Por columnas para data warehouse
* Ejemplo: filas es SQL. Columnas es Cassandra o Apache HBase

Repositorios

Almacenamiento centralizado y gestionado para almcenar gran cantidad de datos. Caracteristicas:

* Escalabilidad
* Diversidad de datos: estructurados, no estructurados, etc…
* Eficiente:
* Procesamiento Distribuido: en clusters
* Metadatos y catalogación: saber como estan organizados los datos
* Seguridad y control de acceso:
* Almacenamiento en bruto: nuna perder los datos originales aunque estos se transformen a posteriori
* Integración en herramientas de análisis:
* Tolerancia a fallos: trabajar con cluster por si uno falla, hay otro

Herramientas

* Apache HDFS (Hadoop distributed File System)
* Cassandra
* S3
* Azure Data Lake
* Azure Databrick

Cubos OLAP

Para realizar análisis multidimensionales. Varias tablas de dimensiones aportan datos a una tabla de hecho donde se realiza la agregación

Tipos de cubos

* HOLAP

Dimensiones y medidas

Herramientas OLAP

Los sistemas OLAP mas populares son:

* Mondrian

Data Warehouse

Ssitema de almacenamiento. Datos esctrurados. Se puede resumir en datos preparados para hacer un análisis. Características clave:

* Estructura
* Orientado a BI

Componentes Warehouse:

* Origenes de datos: diversas fuentes
* ETL: Extraer, transformar y cargar los datos
* Alamcen de datos (Data warehouse): Se gurdan los datos ya consolidades y tranformados.
* Esquema dimensional:
* Herramientas de consultas y reporting: como power BI
* Herramientas de visualización: PowerBI o se puede utilizar Python
* Metadatos
* Seguridad
* Backup
* Rendimiento

Beneficios

* Eficiencia de análisis
* Alta calidad de datos
* Apoyo toma decisiones
* Conssitencia decisiones empresariales

Herramientas:

* Redshift
* Google bigquery

Data lake

Como dataware pero para datos semiestrucvutrados

16/11/2023

Data lake

La diferencia de los datalake es el origen de los datos, son datos de todo tipo (estructurado, no estrucurado, semi-estrucutrado). A través del data lake se pueden transformar para llevarlos a un data warehouse, machine learning o data science.

Características Clave:

* Almacenamiento bruto:
* Escalable
* Variedad de datos:
* Flexibilidad:

Componentes de un datalake

* Almacenamiento de datos: preferiblemente un sistema distribuido
* Ingesta de datos: recoger los datos, ya sea en tiempo real, por lotes. Siempre datos en bruto.

Datos de un data lake:

* Estructurados:
* Semiestructurados
* No estructurados

Data LakeHouse

Intenta coger lo mejor de data lake y data warehouse

Características:

* Almacenamiento en bruto
* Procesamiento integrado
* Escalabilidad
* …

Modelo de datos:

* Almacenamiento en bruto
* Modelo de datos estructurados

Herramientas:

* Hive
* Delta Lake

Gobernanza de datos

Políticas y estándares para garantizar la seguridad y integridad de los datos. Se encarga de que cumpla las leyes y políticas que se aplican a los datos.

Data Owners y data swewars

Los propietarios de los datos son la empresa. El data steward es el responsable e de los datos y se encarga de que cumplan los estándares y políticas establecidas

Calidad de datos

* Evaluar la calidad de los datos

Seguridad y privacidad de datos

Cumplimiento Normativo