UAB
Universitat Autònoma de Barcelona

Antonio Espinosa

Bases de datos analíticas

Sumario

- Analítica de datos
- Análisis de datos en entornos empresariales
- Tipos de bases de datos
- Entornos y usos

Analítica de datos

- Adquirir distintas fuentes de datos no estructurados con distintos formatos
- Gestionar volúmenes de datos más allá de varios Terabytes
- Los tiempos de análisis se acercan al tiempo real
- Es importante la necesidad de escalar la infraestructura de servidores en la nube

¿Por qué construir soluciones a medida?

- Trabajar con nuestros propios datos
- Adaptar las consultas a las propias necesidades
- Integrar múltiples conjuntos de datos: ventas, dirección, I+D
- Aplicar técnicas de *machine learning*
- Adaptar el coste de entornos de análisis existentes

Diferentes tipos de bases de datos

- Proyectos Apache (código abierto)
  - Hadoop, Hbase, Storm, Spark, Impala
- Proveedores en la nube
  - Google BigQuery, Amazon Redshift
- Comerciales
  - Pentaho, Greenplum

Business Analytics

- Práctica de continua exploración de los datos derivados de los procesos de negocio de una compañía
- El objetivo es obtener conocimiento y mejorar la planificación
- Se basa en diversos métodos de análisis de datos aplicados a distintos repositorios de datos

Entornos para procesar datos

- Hadoop
- Impala
- Spark
- BigQuery
- RedShift

Apache Hadoop

- Entorno base para el procesamiento de datos
- Flexible, escalable, programable
- Latencia larga, configuración compleja
- Complejidad de uso, es necesario definir las tareas a un nivel bajo llegando a programarlas en Java
- Ciertos tipos de procesamiento de datos, como el filtrado y el uso de datos con una cierta ordenación natural tienen una fácil implementación en Hadoop
- Hadoop.apache.org

Apache Spark

- Adaptación del modelo **hadoop** para flexibilizar el procesamiento lineal de las fases **map** y **reduce**
- Soporte para la implementación estructuras de procesamiento mas complejas: DAG, bucles y análisis interactivos
- Permite consultas en SQL y el uso de lenguajes alto nivel
- Está eclipsando a **Hadoop** como el entorno de procesamiento analítico más eficiente

Apache Impala

- Diseñado para optimizar la latencia de las consultas
- Arquitectura de sistema orientada al uso intensivo de la memoria
- Permite realizar consultas sobre datos ya almacenados en HDFS o HBase
- Buena solución cuando se necesita realizar un gran número de consultas concurrentes por varios usuarios

Google BigQuery

- Servicio de *Google* para el análisis interactivo de grandes conjuntos de datos
- Gestión de tablas de datos JSON
- Las consultas se realizan en SQL y los resultados son JSON
- Se integra con otros servicios de *Google*: hojas de cálculo
- Se ha presentado un servicio de visualización de datos y analítica: *Analytics data studio*

Amazon RedShift

- Servicio de *Amazon Web Services* específico para conjuntos de datos de analítica
- Los datos se almacenan y gestionan por columnas, permitiendo evitarse la gestión de aquellos atributos que no forman parte de las consultas
- Permite la creación de un cluster de nodos que gestionan las consultas a los datos
- Integración con los otros servicios de *AWS*, como puede ser la importación de datos desde S3

Conclusiones

- Gran ecosistema de sistemas de gestión de datos analíticos
- Elegir las herramientas para el proyecto en función de los requerimientos
  - Volumen datos de entrada
  - Conocimiento técnico del equipo de trabajo
  - Requerimientos de tiempo del análisis
  - Coste de la infraestructura
  - Necesidad de crecer/decrecer rápidamente

Bases de datos analíticas

UAB
Universitat Autònoma
de Barcelona