Sistemas Big data

Introducción

03/10/2023

/\* ------------------------------------------------------------------------------------------------- \*/

06/10/2023

El profesor especialista se llama Ricardo

Conceptos:

Pipeline: conjunto de procesos y herramientas. similar al concepto de ETL (Extract, Transform and Load). Su objetivo principal es analizar grandes volumenes de datos de manera eficiente. Está compuesto de los siguientes procesos:

1. Ingesta de Datos: los datos pueden obtenerse de diferentes fuentes (CSV, BD, HTTP, API) Capturar de manera eficiente los datos. Herramientas de captura como Kafka, SISS
2. Procesamiento de Datos: **El profesor prefiere hacer un preanálisis (validación de datos) como comprobar que un numero es un número, una fecha es una fecha.** Después viene el proceso de transformación y limpieza (seleccionar solamente aquellos datos necesarios).
3. Almacenamiento de Datos:
4. Procesamiento y Análisis Avanzado: Aquí puede entrar la Ia a través del Machine Learning, para hacer predicción de compras, etc…
5. Entrega de resultados: realizar reportes o análisis de las conclusiones sacadas de los datos

Dibujo en pizarra de lo que seria un pipeline (flujo de datos)

1. Datos (sources): CSV, DB, HTTP, Rest
2. Captura Kafka, Python, Pentaho, SISS
3. Preanálisis (validación de datos)
4. Transformación y limpieza: seleccionar solo lo que nos interesa
5. Machine Learning
6. Reporte Resultados

La línea entre Big Data y Bussniess Inteligence es fina. Solo se diferencia en el volumen de datos (a partir de 10 TB de datos)

Hay dos arquitecturas: funcional y técnica