Data Lake Value Proposition

Medical Data Processing Company

Srijana Thapa

Agenda

- What is a Data Lake
- Components of a Data Lake
- Data Lake vs Data Warehouse
- Business Value of Data Lake Solution
- Proposed Data Lake Architecture for Medical Data Processing system

What is a Data Lake

- A data lake is a centralized storage system that holds large volume of raw data in its native format.
    - Structure data
    - semi- Structure data
    - Unstructured data
- Enables real-time and bath processing
- Schema on read
- Scalable & cost- effective

Components of Data Lake

- Ingestion – Tools to ingest data
    - Tools like- Apache Kafka, Apache sqoop
- Processing- Tools to process the Data in Data Lake
    - Tools like- Hadoop MapReduce, Hive, Pig, Spark
- Storage- Stored Data at Scale in a Data Lake
    - Hadoop HDFS
- Serving- Presenting the Data or Results
    - Apache HBase, PostgreSQL

Data Warehouse

- Structured data
- Schema-on-write
- Batch Processing
- Higher Cost

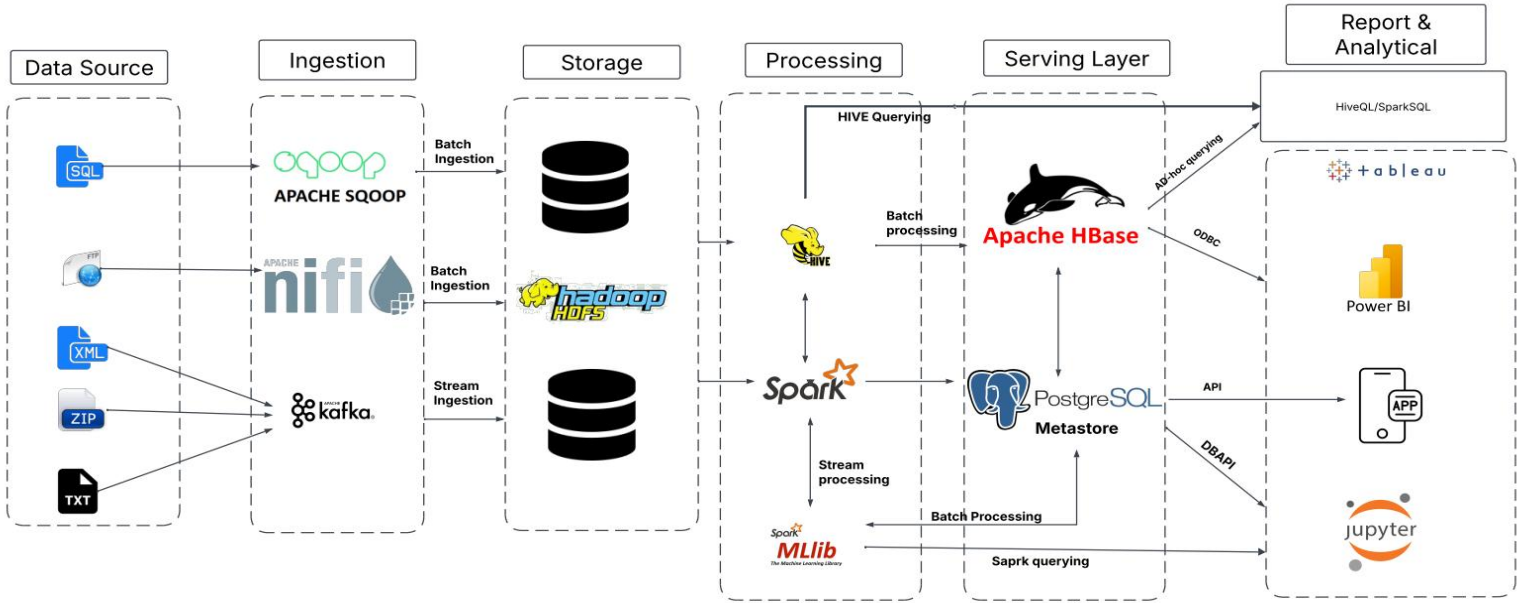
Data Lake

- Store raw data
- Schema-on-read
- Batch and real-time data Processing
- low Cost

Business Value of Data Lake

- Centralized Data Storage
  - Store all type of medical data in one place
- Real –time Analytics
  - Support real-time data processing
- Scalability and flexinility
- Cost effective
  - Reduces costs compared to traditional storage solutions.

Data Lake Architecture



THANK YOU