Data Lake Value Proposition

Medical Data Processing Company

Francis Odo

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

Executive summary

- A data Lake is a central data repository for structured, unstructured and semi-structured data for machine learning, as well as analytics as needed.
- Data Lake can be scaled to any size.
- Data lake takes in data in its original raw format from multiple different sources.
- Data lake is a cost-effective method of storing data for later processing

Components of Data Lake

- Data ingestion
- Data Storage
- Data Security
- Data Analytics
  .
- Data Governance

Data Warehouse

- Store structured data in specific format only.
- Meant for specific use(fixed) within the organization.
- Restrictions for format and schemas on ingestion or WRITE.
- Usually meant for business professionals
- Optimized for reporting

Data Lake

- Can store structured, unstructured and semi-structured data
- Intended use or purpose of data may be known/ unknown or future use
- Restrictions for format and schemas on READ
- Difficult to navigate for non-professionals- For Data Scientists
- Optimized for cost efficiency

Business Value of Data Lake

- Flexibility - Storing raw data from different sources in a common repository with no restrictions on ingestion. Data can be shaped as need through processing

- Advanced Analytics - The capability to take in raw data from multiple and varying sources for advanced data analytics and visualization in real-time with high availability

- Reduced cost and scalability - Compared to Data Warehouse

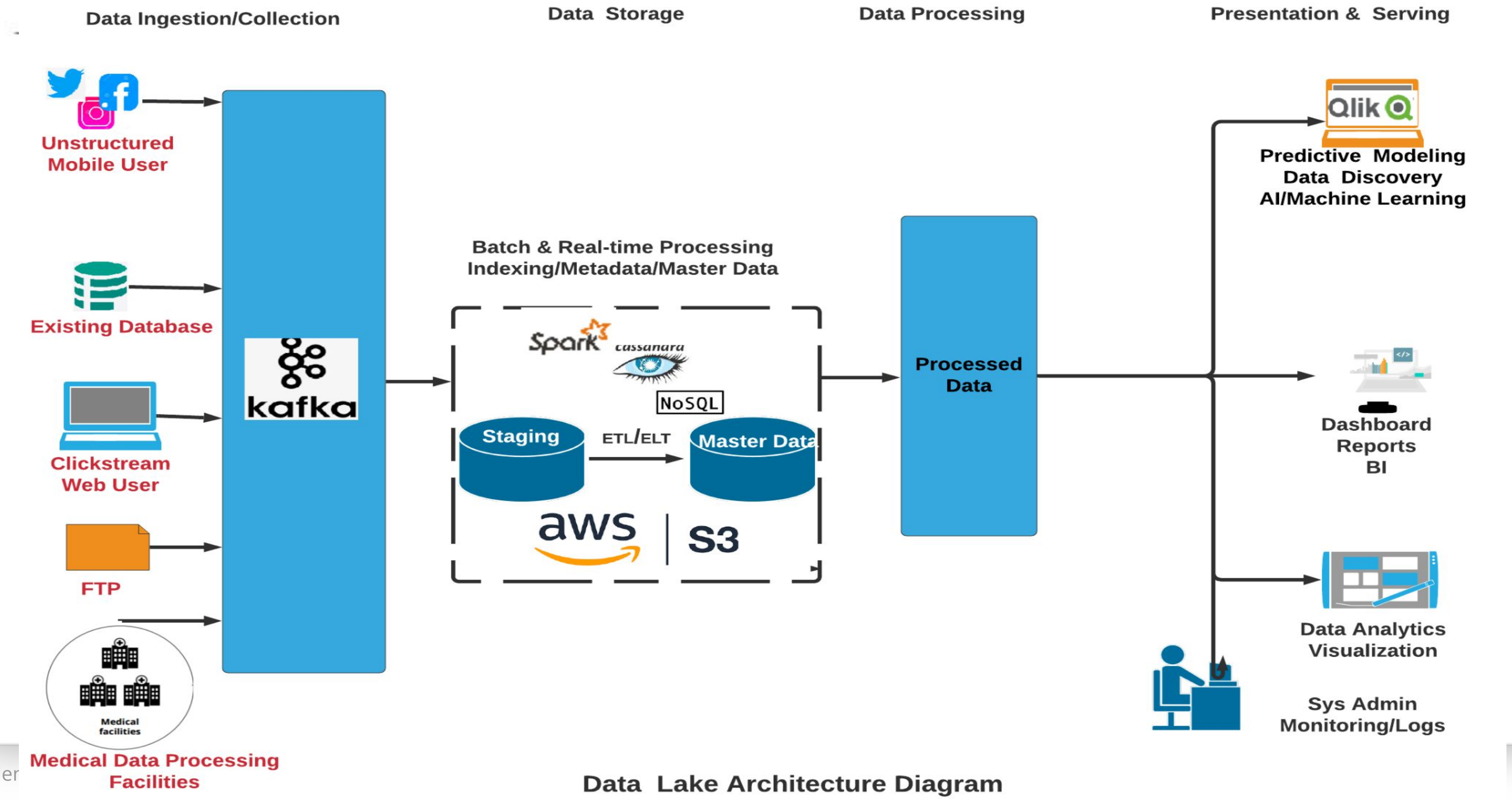
- A 360 degrees view of the customer

Why not Data Warehouse?

A Data Warehouse will not fully address the needs and requirements at hand because:

- Data Warehouse must have structured data, but requirements include varying unstructured/semi-structured data
- Data Ware Warehouse has limitations and restrictions with respect to data size
- Data Warehouse is not as cost effective compared to Data Lake
- A Data Warehouse can be an ingesting component to a Data Lake.

Data Lake Architecture



Data Lake Architecture Diagram

Data Lake Architecture

Data Ingestion Layer

- Data is ingested from multiple varying sources and data types(Structured, Unstructured, Semi-Structured) Electronic Medical Records, Clickstream and other types from mobile and we users

Data Collection and Storage

- Data is collected in its original raw format with little or no restrictions. The data will need to be processed (Tagged, Indexed, Metadata Managed) at the staging area. Using the ETL/ELT process and data governance, it is then stored as Master Data repository. Amazon S3 Bucket is the storage. Computing Platform is Amazon EC2

Data Processing

- The main tool for processing is the Apache Spark. The database engine is Cassandra NoSQL, purposely for handling Big Data. Apache Spark offers a significantly fast processing power due to its in-memory processing capability.

Data Presentation and Data Serving

- After the data is cleaned and processed, it is now ready for Analytics, AI/Machine Learning and any decision-support use

UDACITY

THANK YOU