UDACITY

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

<Provide high-level executive summary - no more than 3-4 sentences of what is a data lake.

- 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. Schema is on enforced when reading the data.
- Data lake is a cost-effective method of storing data for later processing

Components of Data Lake

< **Think About:** What elements are needed to make a data lake? What different layers and/or tools involved? Identify at least 3 in your high level list here. Elaborate more on each component in video.>

- Data ingestion - the data ingestion layer takes data from websites, mobile apps, social media, IoT devices, and existing Data Management systems
- Data Storage - The storage component store and process raw data and support encryption and compression while remaining cost-effective.
- Data Security - It should be highly secure from the use of multi-factor authentication, authorization, role-based access, data protection, etc.
- Data Analytics - After data is ingested, it should be quickly and efficiently analyzed using data analytics and machine learning tools to derive valuable insights and move vetted data into a data warehouse.
- Data Governance - Maintain proper enterprise-level data quality through every aspect of the processes from ingestion to Analytics.

Data Lake vs Data Warehouse

< **Think about:** Research if a Data Warehouse can still be a viable solution instead of a Data Lake? What is the difference between both? >

< You will complete this information on the next slide. Please provide at least 3 items for each. No need to add any content on this slide >
< Video tip: While presenting the differences, elaborate why a Data Lake solution makes more sense for Medical Data Processing Company over a Data Warehouse approach? How the Data Lake / Big Data characteristics different from Data warehouse>

Data Warehouse

- Stores data in files or folders of a specific structure that is processed and refined. Structured data only.
- Meant for specific use within the organization. Purpose is fixed
- restrictions for format and schemas on ingestion or WRITE.
- Usually meant for or handled by business professionals
- Can store structured data only
- Optimized for reporting

Data Lake

- Holds data in its original raw format. Unprocessed and Ungoverned data
- Intended use or purpose of data may be known and unknown or stored for the future
- Restrictions for format and schemas on READ
- Difficult to navigate for those that are unfamiliar with unprocessed/unstructured data. Usually handled by Data Scientists.
- Can store structured, unstructured and semi-structured data
- Optimized for cost efficiency

Business Value of Data Lake

~~**< Think about**~~**:** Provide an executive level summary based on your analysis of problem statement. How would implementing a Data Lake solution help Medical Data Processing Company? Please identify at least 3 business outcomes of building a Data Lake?>
<Video tip: Explain **“Why”** Data Lake solution makes sense for Medical Data Processing Company. **>**

Problem Statement- Medical data processing requires the capability of being able to handle Petabyte-size varying types of data as imaging, binary, chat, messaging, and others from EMR, Insurance Service providers etc. Faced with needs and challenges to join, merge, correlate and analyze the data.

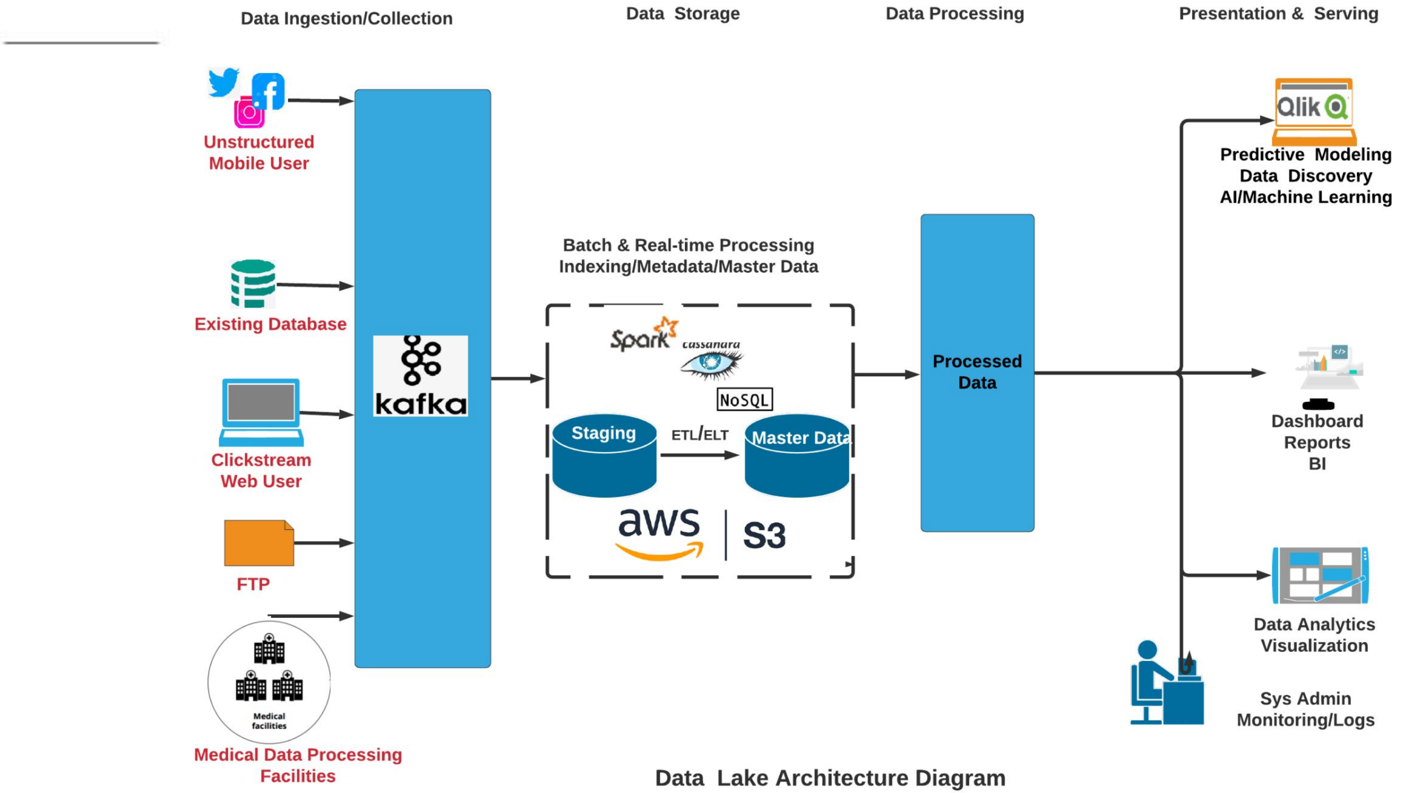
- Flexibility - Storing raw data from different sources in a common repository with little or no restrictions on ingestion. Data can be shaped as need through processing
- Advanced Analytics - The capability to take in raw data from multiple and varying sources for advanced data analytics and visualization in real-time with high availability
- Reduced cost and scalability - A data lake is significantly cost effective to implement compared to a data warehouse mainly due to its flexibility
- A 360 degrees view of the customer

Why not Data Warehouse?

A Data Warehouse will not fully address the needs and requirements of the Medical Processing Company for the following reasons:

- Data Warehouse must have structured data, but requirements include varying unstructured data
- Data Ware Warehouse has limitations and restrictions with respect to data size, which is better handled in a Data Lake
- 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