HOW TO APPROACH BUILDING A HEALTHCARE DATA LAKE

SUMMARY BY DR. ALVIN ANG

CONTENTS

INTRODUCTION

- This manuscript is a summary of two sources:
    - How to Approach Building a Healthcare Data Lake Roadmap (O'Dowd 2018)
    - Making Growing Healthcare Data Lakes Actionable for Providers (O'Dowd 2018)

PROBLEM STATEMENT

- More connected digital devices contribute data
    - → Increasing amounts of health data
    - → Healthcare Data Lakes are flooding.
- Healthcare Organizations want to make better use of the data stored in Data Lakes.
- Organizations need to
    - Understand the nature of the data being collected
    - How to store and access that data
    - How to make the data actionable
    - Create a roadmap to leverage data in order to improve workflow and patient care
- These takes significant IT infrastructure planning.
- It's about creating a learning health system.

STRUCTURED DATA

- Examples of Structured Data:
    - Patients' Demographic Information
    - Electronic Health Record (EHR)
        - Patient's medical history
        - Diagnoses
        - Medications
        - Treatment plans
        - Immunization dates
        - Allergies
        - Radiology images
        - Laboratory and test results
    - Diagnosis and Procedure codes,
    - Medication Codes
- Structured Data is stored in Traditional Data Warehouses as Files.
- Structured Data is easier to analyze and store because it has straightforward boundaries and stored in standardized format.

UNSTRUCTURED DATA

- Examples of Unstructured Data:
    - Images
    - Numbers
    - Complex Data Sets
- Unstructured Data is not organized.
- Unstructured Data is stored in the Data Lake
- It is often too vast to be retrieved conveniently or used for analytics.
- But organizations need tools to make their data (Structured / Unstructured) actionable.

- Utilizing tools such as
    - Hadoop
    - Object Storage
    - Blockchain
- Can turn data lakes from storage dumps to active tools.

HADOOP

- Hadoop can handle both Structured and Unstructured Data.
- Hadoop can turn Data Lakes from storage dumps to active tools.
- Hadoop is ***not*** a data warehouse
- Hadoop is an open-source software
- Hadoop distributes large amounts of data to different processing nodes, which later then combines the collected results.
- Hadoop separates unstructured data into nodes that are individual parts of a larger data structure.
- The nodes are linked together to combine the stored data stored.
- This approach allows data to be processed faster, since the system is working with smaller batches of localized data instead of the entire warehouse.
- Hadoop Distributed File System (HDFS) is the primary distributed storage used by Hadoop applications
- HDFS is not a physical database, but it collects data and stores it in clusters until an organization is ready to use it.

OBJECT STORAGE

- Distributed Object Storage = Software + Hardware solutions that are based on shared architecture and supports scale-out file technology.
- Object Storage treats Data as Objects, NOT Files / Blocks.
- Object Storage uses Unique Identifiers.
- Objects Storage does not have a Hierarchical Structure.
- Object Storage should not be used to run Analytics because it's not the fastest storage solution.
- But Object Storage provides fastest accessibility → Unlike tape where you have to know the serial number, track the tape, and physically retrieve it.

BLOCKCHAIN

- The most significant challenge of the data lake is knowing how to trace the data back to where it originated - to give it context.

- Blockchain creates an unchangeable ledger that can trace permission, access, and transmission of data to create a controlled data lake.

- It's also possible to have patient ownership pools.

- Blockchain is a potential remedy for stagnant data lakes.

CLOUD MODELS FOR DATA LAKE

<table>
<tr><th colspan="2">Cloud</th><th>On Premise</th></tr>
<tr><td colspan="2">Flexibility

Able to purchase more storage space when needed.</td><td>Inflexible

Need to add more hardware servers when more storage space is needed.</td></tr>
<tr><td colspan="2">Well Connected / Easily Scalable

More applications are moved to the cloud & more compute cycles are accessed for analytics.

More control over IT infrastructure because dan deploy Hadoop</td><td>Hard to Scale

Must purchase, manage and maintain On Premise hardware servers = must employ IT staff.

Cannot Deploy Hadoop</td></tr>
<tr><th>Public Cloud</th><th>Private Cloud</th><th></th></tr>
<tr><td>The most scalable data storage solution.

Storage space can be added or dropped as the size of an organization changes.

Popular for temporary projects as well as data migration.

Less expensive option for health systems that have a lot of unstructured data.</td><td>Gives health IT staff direct control over the contents stored.

Health Organizations can keep a close eye on Protected Health Information (PHI).</td><td></td></tr>
<tr><td colspan="2">How to decide which one?

• Budget

• Staff

• Amount of data that needs to be stored.

• No matter what a hospital's budget is, the more data that is produced the more expensive it will be to store it.</td><td></td></tr>
</table>

REFERENCES

O'Dowd, E. (2018). "How to Approach Building a Healthcare Data Lake Roadmap." from https://hitinfrastructure.com/news/how-to-approach-building-a-healthcare-data-lake-roadmap.

O'Dowd, E. (2018). "Making Growing Healthcare Data Lakes Actionable for Providers." from https://hitinfrastructure.com/news/making-growing-healthcare-data-lakes-actionable-for-providers.

ABOUT THE AUTHOR

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.