**AIT 622 BIG DATA CASE STUDY**

**REAL WORLD MEDICAL ANALYTICS BY CITIUSTECH**

1. **Who**

The medical services industry is one the biggest and the quickest developing field of the general economy. Today the medical services industry is one of the exceptionally created and profoundly tech-prepared industry. All the associations in this industry give experiences, interview and answers for their customers. CitiusTech is one the business driving associations in medical care. This organization offers chances to medical services IT experts to assume a fundamental part to change or improve the manner in which medical services is overseen.

CitiusTech is the biggest provider of medical consideration items and arrangements in business for medical recorded, life science organizations and 2700 and more medical specialists all through the world. The item's and innovative line up of CitiusTech incorporates the product updation and advancement, entomb and intra institutional operability in medical field, client care. It gives top notch administrations and encourages the customer to improve by utilizing best in class innovation. CitiusTech is delivers best in class arrangements in medical consideration all through the world at an entirely reasonable expense. (Citiustech, 2020)

1. **Need**

The customer is a main symptomatic specialist co-op with a large group of patient data that should be dissected to furnish their patients with the best medical arrangements and consultation.

The primary objective for this firm is to change the current arrangement of data warehousing arrangement with the end goal that their customers could get the best assistance as envisioned. The normal to and from data every day is 5 – 10 million messages. The saw needs to improve the perceptions to get more intuitive visualizations of a patient’s indicative data and furthermore not many of the diagnostic inquiries that should have been fathomed to recover data from huge datasets.

The problems that CitiusTech wanted to solve are to have a rigorous and robust data governance. They have been successful in achieving it and have also managed to take care of their data lakes and warehouses and kept them updated from time to time in order to offer the best services to their clients. (Knowledge HUB, 2020)

The foundation confronted a great deal of obstacles that any Diagnostic specialist co-op may have looked now and again, this would be in all honesty a determination cycle of appropriate specialist organization who can help these establishments in arriving at their prerequisites and that issue was illuminated when they met the CitiusTech firm. (Real World clinical analytics, 2020)

This is viewed as a major information issue as it envelops all the V's:

* Variety
* Volume
* Velocity

**VARIETY:**

The diagnostic specialist co-op deals with a scope of data on the grounds that there are various kinds of reports that they produce, for instance, blood test reports, X-beam reports, ECG reports, MMR reports, and so forth.

**VOLUME:**

Diagnostic Service Provider deals with various patient records every day, and when consolidated, these archives add up to petabytes of information. The customer handles 5-10 million messages for each day, which is a ton of information, as expressed above.

**VELOCITY:**

The quantity of patients being treated by the Diagnostic Service Provider keeps on rising step by step and an expansion in the quantity of patients is straightforwardly corresponding to the increment in the quantity of records, which implies there are a fast inflow and outpouring of information demonstrating the attributes of speed.

1. **Challenges**

The primary members would be the specialist co-op and their customers that is The CitiusTech and the diagnostic service co-op including medical care foundation, the patients and all the people associated with the diagnostic service organization.

In order to enhance the service quality the client wanted to replace the existing system with a data lake which could handle immense volume of data which count to petabytes and 5-10 Million messages on a daily basis.

The data is owned by the diagnostics service provider and all the individuals. CitiusTech will handle all the necessary actions related to data access for individuals or firm. Upon request it may allow individuals and firms to alter their own data. They have a very strict rules on the data privacy.

The privacy of data is a huge a deal in organizations so as in CitiusTech they are bound to rules and bound to the individual’s or firm’s data. They do not allow or let the privacy of the data to fall at risk.

Quality of data is a very important part of handling the data. The organizations have few protocols where they can manage and maintain the quality of the data. This is one of the key roles to be played by the data scientist / data manager.

The main challenge for them was to choose proper service provider who could help them in meeting their requirements. This selection process is trickiest part. CitiusTech fits in perfect to solve this problem for the diagnostics service providers.

Handling the huge datasets is a major challenge that can be faced by any organization. Managing these huge datasets and also handling the incoming and outgoing data on a daily basis can be tricky and challenging. CitiusTech helps these medical organizations to serve their customers better.

1. **Stakeholders**

Governing bodies like NCQA, CMS and ONC publish new measures are the stakeholders of CitiusTech. The improvement of clinical, IT, and authoritative cycles to advance execution against these measures has become a center competency that payers and suppliers must overhaul ceaselessly to facilitate and fulfill suppliers and keep accounts sound.

1. **Requirements, resources needed**

CitiusTech was chosen by the customer to build up the framework, given its solid skill across medical services work processes, information mining and normalization, large information preparing and progressed investigation use cases in medical services. CitiusTech set up a group of area specialists and large information experts to do a point by point investigation of the customer's current information climate. In view of its examination, CitiusTech planned the arrangement dependent on Cloudera Hadoop appropriation, MapReduce, Spark Streaming and other Hadoop/Big information advances.

1. **Time**

The time period cannot be defined for this institution as it provides a continuous on-going service to its clients unless and until they want to change the service provider.

1. **Results/Findings**

CitiusTech directed nitty gritty appraisal of the customer's current information climate. The arrangement included structure an adaptable and configurable information model to speak to medical care information and empower progressed examination.

The project is a huge success many institutions have approached and associated with CitiusTech based on the services they provide their clients. The institutions like Amazon, partnering in AWS, Google partnering in Google Cloud Platform, IBM, Microsoft, Automation Anywhere, Tableau, UiPath, Tricentis, Cloudera, Infor, Veeva, Atmosera, ClearData, CloudCheckr, Datica, Talend are active partners with CitiusTech.

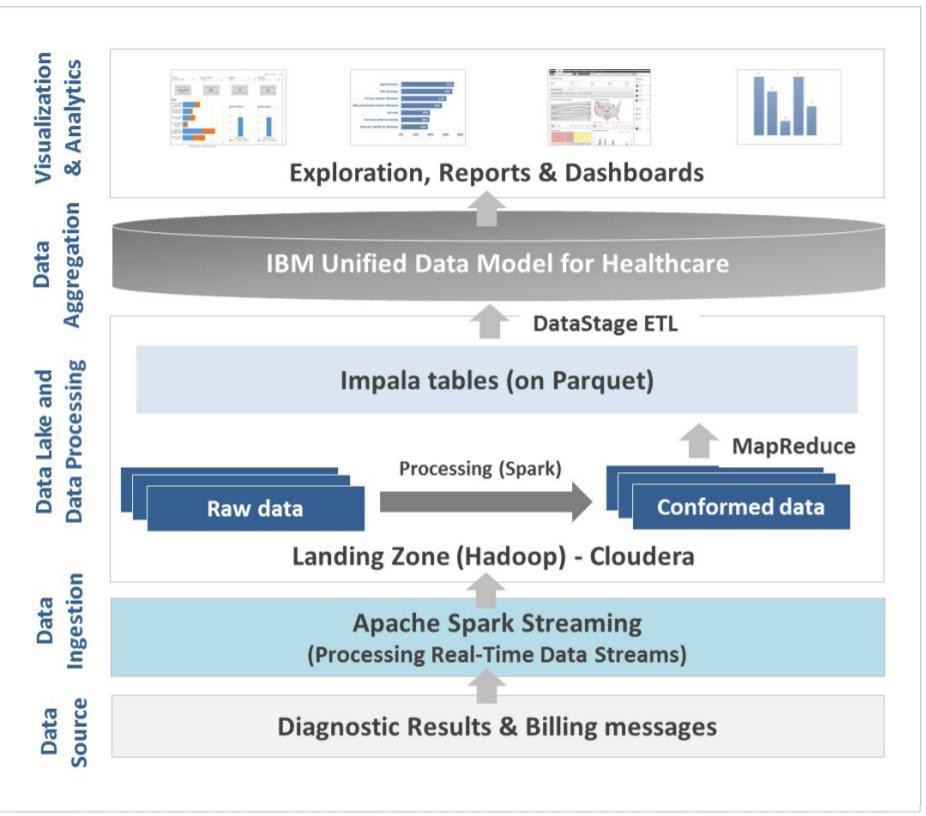


Fig. Sequence of steps followed in the project

* **Data Ingestion :**

Ingestion and storage of HL7 V2.5.1, V3 (Spark Streaming) and BFE messages (billing data flat files) in data lake i.e., HBase + HDFS which is done via Spark and the positions that are capable enough of processing 20,000+ records per second. Using MapReduce to perform Cleaning, Standardizing and de-duplication of data. In standardizing LOINC and SNOMED are done. The HL7 and BFE traverse are stored in HBase tables.

* **Data Processing and Integration:**

MapReduce is used to process the data in the HBase and using Cloudera Impala for converting the messages into parquet format that can be queried. The IBM data stage of ETL is used for exposing the parquet data.

* **Data Aggregation and Storage:**

In this step all the messages are put into the IBM healthcare data model which is on IBM PDA from the Impala tables.

* **Data Presentation:**

Using big data lake as the source of information to downstream analytics applications, providing visualizations using BI/analytics apparatus that leverage multi-dimensional models like Cognos.

As a result of this project CitiusTech was able to associate with many other institutions as described above to expand its services in many places of he world and it has been the best in its sector to lead as one of the most successful medical service provider.

1. **Critique**

The current venture is a good one that addresses all the customer's issues. Whenever requested any progressions or enhancements in the venture I would recommend that we can incorporate some extra highlights into it. A bit of the additional features could take after creation a patient door to give the Diagnostic Report Visualizations genuinely to the patients and subject to the observations conveyed the patients must be given a reality report and suggestions over which pro should be advised, so these are a segment of the modifications that can be made to the current endeavor.

1. **Define Terms**

**HBase**: HBase is an Open-Source non-relational distributed database which has been built on the basis of Google’s Bigtable and is built on Java programing language developed as a part of the Apache Software Foundations. (Wiki, n.d.)

**HDFS**: HDFS is a distributed, scalable and portable file system built on java programing language for Hadoop system. (Wiki, n.d.)

**HL7 V2.5.1**: HL7 is the world's most widely implemented messaging standard for Healthcare data exchange and V2.5.1 stands for version description. (HL7 International, n.d.)

**LOINC, SNOMED**: Loinc is a message code used for HL7. Loinc code is used to define question, Snomed is used to represent the answer. (EHR Intelligence, n.d.)

**Cloudera Impala**: It offers fast and interactive queries to Apache Hadoop and HDFS or HBase. It has similar syntax as SQL and shares the same interface as Apache Hive. (Cloudera, n.d.)

**Parquet**: Apache Parquet is an Open-Source data storage format and originates from the Hadoop ecosystems. (Parquet, n.d.)

# References

(2020). Retrieved from Citiustech: https://www.citiustech.com/aboutus/

*Cloudera*. (n.d.). Retrieved from Cloudera Impala: https://docs.cloudera.com/documentation/enterprise/5-3-x/topics/impala\_intro.html

*EHR Intelligence*. (n.d.). Retrieved from Loinc, Snomed: https://ehrintelligence.com/news/snomed-and-loinc-team-up-ahead-of-icd-10

*HL7 International*. (n.d.). Retrieved from HL7: https://www.hl7.org/implement/standards/product\_brief.cfm?product\_id=94

*Knowledge HUB*. (2020). Retrieved from CitiusTech: https://www.citiustech.com/knowledge-hub/detail?id=tLFf7y1aqx0=

*Parquet*. (n.d.). Retrieved from Apache Parquet: https://parquet.apache.org/

*Real World clinical analytics*. (2020). Retrieved from CitiusTech: https://www.citiustech.com/uploads/knowledgehub/pdf/big-data-for-real-world-

*Wiki*. (n.d.). Retrieved from Apache HBase: https://en.wikipedia.org/wiki/Apache\_HBase

*Wiki*. (n.d.). Retrieved from HDFS: https://en.wikipedia.org/wiki/Apache\_Hadoop#Hadoop\_distributed\_file\_system