Big Data Project on Healthcare Domain

Hitesh Bhoir[1] Vivek Gupta[2] Linomon Kuriakose[3] Iqbal Shaikh[4]
[1,2,3]UG Student [4]Professor
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]TCOE, Mumbai University, Maharashtra, India

Abstract— Big data is an enormous amount of information that can do wonders. It has become an area of special interest for the past two decades due to a great potential that is hidden in it. Various industries in the public and private sectors generate, store and analyze data Big data aimed at improving the services provided. In the healthcare industry Various sources of big data include hospital records, patient medical records, Health exams and device results that are part of the Internet of Things. Proper management and analysis is required to derive this data. Meaningful information. Otherwise, quickly analyze big data to find a solution It's comparable to looking for a needle in a haystack. There are various challenges Associated with every step in processing big data that can only be exceeded by usage High-end computing solution for big data analysis. Make it relevant for this reason Public health solutions need to be perfect for healthcare providers Has the right infrastructure to systematically generate and analyze Big data. Efficient management, analysis and interpretation of big data is subject to change A game by paving the way for modern healthcare. So different Industries, including healthcare, are taking proactive steps to transform this. Potential for better service and financial benefits. Strong integration of biomedicine Health data can revolutionize modern healthcare organizations Medical therapy and personalized medicine.

Keywords: Healthcare, HDFS, YARN

I. INTRODUCTION

Information was the key to a better organization and new development. more The information we have can be more optimally organized to provide the best result. For this reason, data collection is an important part of any organization. we You can also use this data to predict the current trends for a particular parameter. Future event. Production started amid growing awareness Collect more of almost all data through the adoption of technology Development in this direction. Today we are facing a flood of situations Social activities, science, work, in a sense, you can compare the current situation with a large amount of data. where it has become unmanageable with currently available technologies. This has led to the creation of the term 'big data' to describe data that is large and unmanageable. In order to meet our present and future social needs, we need to develop new strategies to organize this data and derive meaningful information. One such special social need is healthcare. Like every other industry, healthcare organizations are producing data at a tremendous rate that presents many advantages and challenges at the same time. In this review, we discuss about the basics of big data including its management, analysis and future prospects especially in healthcare sector

II. CHALLENGES

To make all the data available at one place & make sense of it. Data Security (Authenticate & Authorized). Data processing in real-time. Data should be accessible, traceable, & audited at any time when needed.

III. PROPOSED SYSTEM

To make all your data available in one place and use it wisely. Data security (authentication and authorization). Real-time data processing. Data should be accessible, traceable, and auditable whenever needed. We recommended a Big Data Platform powered by Cloudera EDH with network partner Amazon Web Services with AWS, Cloudera Version 5.16.2, HDFS, Kafka, HIVE, Apache Zookeeper, Sentry, HUE, Sparks, Apache Sqoop.

IV. WHAT IS HADOOP

Apache™ Hadoop® is an open source software project that enables the distributed processing of large data sets across clusters of servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer.

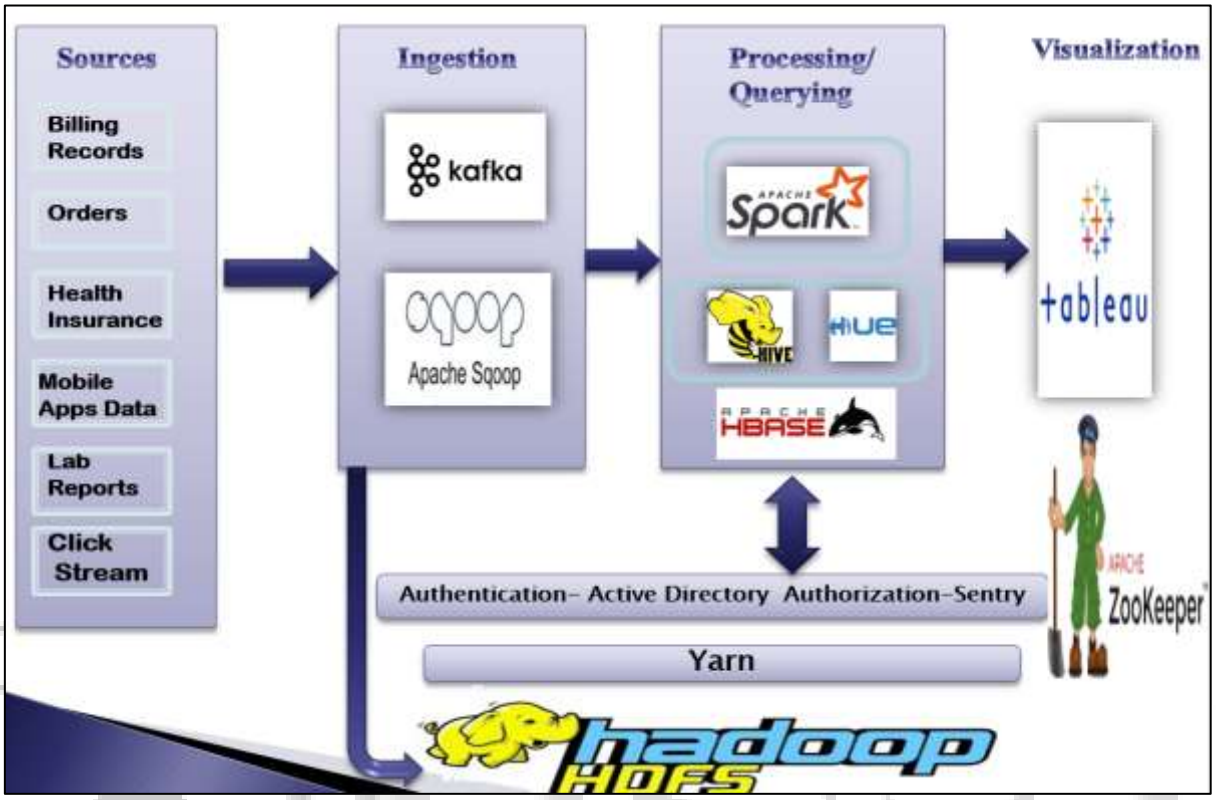
Apache Hadoop has two pillars:

1) YARN - Yet Another Resource Negotiator (YARN) assigns CPU, memory, and storage to applications running on a Hadoop cluster. The first generation of Hadoop could only run MapReduce applications. YARN enables other application frameworks (like Spark) to run on Hadoop as well, which opens up a wealth of possibilities.
2) HDFS - Hadoop Distributed File System (HDFS) is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together the file systems on many local nodes to make them into one big file system. Hadoop is supplemented by an system of Apache projects, such as Pig, Hive and Zookeeper, that extend the value of Hadoop and improves its usability. Hadoop changes the economics and the dynamics of large scale computing. Its impact can be bubbled down to four salient characteristics.
3) Scalable– New nodes can be added as needed, and added without needing to change data formats, how data is loaded, how jobs are written, or the applications on top
4) Cost effective– Hadoop brings massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all your data.
5) Flexible – Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources. Data from multiple sources can be joined and

aggregated in arbitrary ways enabling deeper analyses than any one system can provide.

6) Fault tolerant– When you lose a node, the system redirects work to another location of the data and continues processing without missing a fright beat.

V. DATA FLOW



VI. CONCLUSION & FUTURE WORKS

By implementing Enterprise Data Hub and advanced analytics, integration of entire Health Ecosystem became possible. Which enables to bring data from variety of sources. The Cloudera EDH met the core requirements such as big data lake, including a very robust security and audit framework. Hence, the agility on delivering new products & new insights that help improve patient care and support healthier lifestyles. Also, achieved much better outcomes, both patient-related and financial-related.

REFERENCES

[1] Pol, International Journal of Advanced Research in Computer Science and Software Engineering 4(11), November - 2014, pp. 1028-1034
[2] Apache Hadoop HDFS Architecture Guide.url:https://hadoop.apache. org/docs/r1.2.1/hdfs_design.html.
[3] Cloudera. Machine Learning, Analytics, Cloud - Cloudera.url:https://www.cloudera.com/.
[4] J. Archenaa and M. Anita, "Health recommender system using big data analytics," Journal of Management Science and Business Intelligence, vol. 86, no. 5, pp. 17–24, 2017.
[5] L. R. Nair and S. D. Shetty, "Applying spark-based machine learning model on streaming big data for health status prediction," Computers & Electrical Engineering, vol. 65, no. 1, pp. 393–399, 2018.

