Significance of Big Data Frameworks and Speculative Approaches in Healthcare Systems

G.P. Hegde
Department of Information Science Engineering, SDM Institute of Technology, Ujire, Mangalore
Email: gphegde123@gmail.com
Nagaratna Hegde
Department of Computer Science and Engineering, VCE, Hyderabad
Email: nagaratnaph@gmail.com

---**ABSTRACT**--

Due to rapid generation of large numbers of integrated medical data from various communities of the world, health care systems need to be stored and streamed properly. Heterogeneous data has been scattered from different healthcare medical records has various attributes and primarily they are not structured using data frameworks. In this paper we emphasized the various frameworks of big data and its significance in healthcare systems. This paper also focuses on the significance of speculative approaches and the streaming process of data frameworks. This would be helpful for researchers to analyse and evaluate the characteristics of frameworks with respect to network throughput and latency. Selection of nodes for different stages of healthcare is also a challenging issue while selecting data frameworks. State of art approaches show the role of big data frameworks in other sectors of applications.

Keywords – **Big data, pig, spark, flume, frameworks, Hadoop.**

Date of Submission: May 16, 2021 Date of Acceptance: Jun 27, 2021

I. INTRODUCTION

Today, billions of people are accessing and extracting large amounts of data through the internet, social media and social networks. Collection of large amounts of healthcare data and processing those data is a challenging task in various fields. The data are stored and manipulated based on the data streaming process. Systematic analysis of data and its visualization makes the explosion of hidden data. Data analytics makes the explosion of data obtained from big data Hadoop frameworks during its evaluation. This evaluated data gives significant information about healthcare business intelligence. The volume of data dictation and its performance analysis is essential for new technologies development and innovations. Generation of heterogeneous monthly data or half yearly data from various companies, hospitals, institutes and forests sectors are stored in data nodes of data warehouses, in order to improve the performance of data accessing and its evaluation [1]. The integration of Internet of Things and big data analytics gives powerful strength for customers to solve their needs and requirements in various aspects. Dynamic processing and big data analysis process of healthcare data has been carried out in a systematic way by big data analytics frameworks [2]. Millions of data transactions have been carried out daily due to fast increasing IoT usages in various smart cities. Velocity factor is adopted by the author for analysis of big data [3]. Smart city applications are flooded over today and giving a lot of information for real time data. The IoT technologies are widely used with Big Data and IoT equipment are placed in various fields of cities. In the education system, traffic control system, and home automations system smart city-based services are used [4]. Integration of embedded devices with Metadata depends on data accessing performance of Hadoop frameworks. Today, the requirement of big data analysis is showing a tremendous job in connection with the Internet of Things (IoT). The valid requirements can be boosted by various types of resources along with big data and IoT. In this digital world to process the real time data in smart cities most of the industries have to be maintaining their data in big data bases. Classification of data can be made more effective using Hadoop cluster technology [5]. Author in [6] presented the taxonomy of streaming analytics frameworks, architecture of a data stream processing system. They have addressed challenges of data frameworks development and its infrastructure with a selection of suitable streaming frameworks. Data can be extracted from various media such as sensors, Realtime Data Database Management System (RDBMS), social communication systems, social networks, weblogs, log transactions [7][8]. Figure 1 shows 6V's of big data analytics and streaming of big data nature.

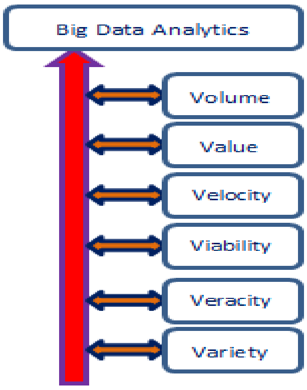

Figure 1 Different views of big data analytics

Figure 1 shows the different parameters or views of big data analytics in real time system. In big data analytics value parameter is applicable in clinically appropriate data longitudinal studies. Volume parameter describes high throughput technologies continuous monitoring of vital signs. Velocity parameter illustrates about high speed processing for fast clinical decision support. Increasing data generation rate by the health infrastructure. Variety parameter related to heterogeneous and unstructured medical data sources. It is differences in frequencies and taxonomies. Veracity factors discuss unreliable medical data quality. Here data coming from uncontrolled environments. Variability parameter illustrates serial health effects and disease evolution Non deterministic models of illness and health.

II. Speculative Approaches

A number of speculative approaches can be employed to recognize irregularities in vast amounts of data from different datasets. The frameworks available for the analysis of healthcare data are as follows:

Predictive Analytics in Healthcare: It is identified as a business intelligence approach, since from the past three years. The prediction algorithms have been properly utilized in machine learning. These methods are supported in arranging larger healthcare records using big data analytics by including multimedia analytics. In the medical field patients may get disturbed by any risk factors of unhealthiness.

Machine Learning in Healthcare: The concept of machine learning is very similar to that of data mining, both of which scan data to identify patterns. Rather than extracting data based on human understanding, as in data mining applications, machine learning uses that data to improve the program's understanding. Detection of breast cancer survivability using decision trees and remedies for depression [9][10]. Heart disease diagnostics using k-nearest neighbors [11]. Classification of genes and detection of diabetes mellitus using support vector machine [12][13] Machine learning identifies data patterns and then alters the program function accordingly.

Electronic Health Records: EHR represents the most widespread health application of big data in healthcare. Each patient has his/her own medical records, with details that include their medical history, allergies diagnosis, symptoms, and lab test results. Patient records are shared in both public and private sectors with healthcare providers via a secure information system. These files are modifiable, in that doctors can make changes over time and add new medical test results, without the need for paper work or duplication of data.

III. Big Data Frameworks

Author in paper [14] has made discussion on Megabyte to Gigabyte and Gigabyte to Exabyte concepts with transmission of big data with batch processing. They have also discussed Hadoop frameworks and presented big data challenges. Apache Hadoop is a tool or data framework consisting of clusters with data nodes and name nodes called Hadoop distributed file systems. MapReduce is lying above the Hadoop cluster for transmission of mapped information into reduced form of processed data. The large volume data has been processed using HDFS. Approximately billions of people access the internet, this can be supported by Hadoop clusters simultaneously. Hadoop cluster uses write once /read many concepts. Hence HDFS does not support random writing of data into its clustering by the clients. HDFS manages streaming of large volumes of data extracted from the hard disk. The size of the HDFS block is 64MB or 128MB and does not support a direct caching mechanism. Data locality is one of the main features of HDFS. Each cluster of HDFS consists of NameNode and multiple nodes. Name node maintains the metadata required to store and retrieve the actual data from the DataNodes. NameNode never stores actual data and is used for transmission. Various big data analytics tools are available for handling healthcare systems and conceptual architecture illustrates data gathering [14]. A medical data processing using ETL tasks with MapReduce Java based platform and solutions were addressed in [15]. Hadoop with SQL platform is placed above the HDFS and it influences the recovery from fault tolerance system for RDBMS by providing medical data information with flexibility and scalability issues as mentioned in [16]. Medical big data analysis has been made in [17] and explained about chronical diseases monitoring with analytics and visualization tools. Figure 2 shows the storage platform frameworks used in healthcare data processing.

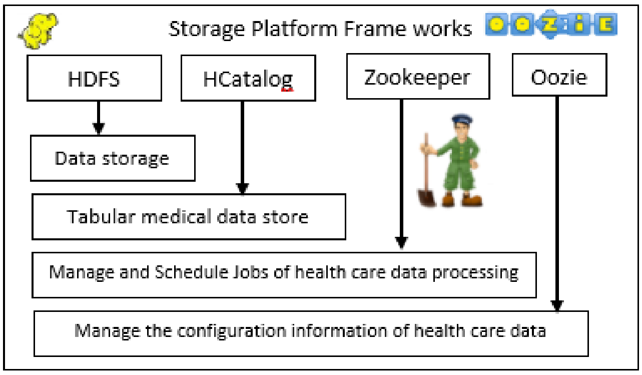
Storage Platform Frame works
OOZIE
HDFS
HCatalog
Zookeeper
Oozie
Data storage
Tabular medical data store
Manage and Schedule Jobs of health care data processing
Manage the configuration information of health care data

Figure 2: Data Storage of health care data records

Apache Pig: Apache Pig is one of the available open-source platforms being used to better analyze big data. Pig is an alternative to the MapReduce programming too. First developed by the Yahoo web service provider as a research project, Pig allows users to develop their own user-defined functions and supports many traditional data operations such as join, sort, filter, etc.

Apache Sqoop: It is a simple command line interface application. Sqoop is located between Hadoop and relational databases. It is used to import the data from RDBMS to HDFS of Hadoop and export data from HDFS of Hadoop. Sqoop is tested with Microsoft SQL server ie

Postgres SQL, MySQL and Oracle. There are two versions of Sqoop version 1 and version 2. Sqoop tool can be used with Java database connectivity (JDBC). In version 1 Sqoop extraction of data has been carried out using connectors written for specific databases. But version-2 does not support connectors. In version 1 direct transformation of data would carry out between RDBMS to Hive and vice versa. Figure 3 shows general purpose executable engine.

Apache Storm: It is one of the open source big data essential Hadoop tools used as data framework in dynamic systems. Storm makes unbounded streams of data in serial fashion. It consists of two clusters namely master node and slave node. It is a fast queuing streaming agency. Apache storm processes the data stream wise and divides the streams if it is necessary for every stage of transmission. In storm master nodes are called nimbus and slave nodes are called supervisor nodes. For unexpected unrecoverable failure, the storm vanishes until it reacts. The supervisor nodes find the number of regions it can provide which is applicable to clusters. Each supervisor node supervises the data process in Hadoop data frameworks, based on the number of supervisors allocated regions. Apache storm is not enough to make control cluster present state. Zoopkeeper solves the communication complexity between master node and supervisor node. Figure 2 shows the Apache storm architecture. The cardiological ambulance control is an emergency communication medical system. This real time platform-based stream fault tolerance is managed by Apache Storm in big data analytics, this concept was described in [18].

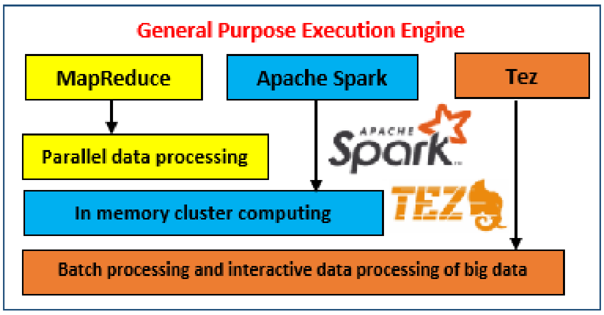
General Purpose Execution Engine
MapReduce
Apache Spark
Tez
Parallel data processing
APACHE Spark
In memory cluster computing
TEZ
Batch processing and interactive data processing of big data

Figure 3: Execution engine platform-based frame works

Apache Flink: is a big data essential tool used in distributed systems. It supports bounded and unbounded streams of messages. It works for all common cluster's areas. It is also an open source tool that runs real time dynamic data pipelines with billions of values within a second by using fault tolerance technique. It never runs micro batches. User sends various job tasks to the job manager then it is split into individual tasks. These tasks are transferred to slave nodes. Slave calculates statistical steps and the same thing is computed. So that distributed computation can be carried out. The fault tolerance in Flink maintains the status of the job based on results. These snaps consist of checkpoints that can return faults status if any communication failure takes place. Apache Flink is used in batch and stream processing. In distributed systems it works for less amount of data latency with more fault tolerance. It supports data stream application programming interface and dataset interface. The Flink works on batch data. Task slot is one of the major parts of Flink that acts as an execution resource. These slots are allocated in task manager. All the parallel talks are run through these slots only. Flink executes parallel tasks. Figure 3 shows the architecture of Apache Flink.

Apache Spark: For the purpose of Big Data processing Apache Spark utilized to influence the uniform execution graphs. Spark SQL tool is supported by this Apache frame Spark. The Big Data analytics and IoT based smart city pplications can be viewed by Spark. This is an essential tool in memory processing. If scalable primary memory is full then Spark can accommodate data in the hard disk. It consists of a driver node, a worker node and a cluster manager. One of the main parts in Spark is Spark context is an entry point. The integration takes place in worker nodes. The cluster provides higher bandwidth to pass the data packets from driver to worker nodes. The performance can be increased by disabling the security features of clusters. Figure 4 shows the basic architecture of Apache Spark. The author [19] presented Spark based MRI scanned data processing applications to store the data. Using high speed data stream processing with batch interactive strategies. The HDFS based system was implemented in order to prevent disk writing.

Apache Flume: This framework extracts the Big Data for further evaluation in distributed systems. It integrates all the data in the Meta database. It supports data frame streaming. Apache Flume is used to collect log data present in log files from web servers. It collects the data in batch and gives highly available services. Collected data is integrated and sent to the central server. Flume collects data from various web servers as data generators and passes them to channels and stored in the sink. Apache Flume supports context routing [20]. Figure 5 shows the basic architecture of Apache Flume. The heart disorder patients risk prediction and classification can be recommended by machine learning techniques with Hadoop batch libraries supported by data frame work like Apache Mahout as described in[21]. Big data bases like Asthma Exacerbations prediction of health care systems can be achieved by the Hive database platform of HDFS supported systems [22].

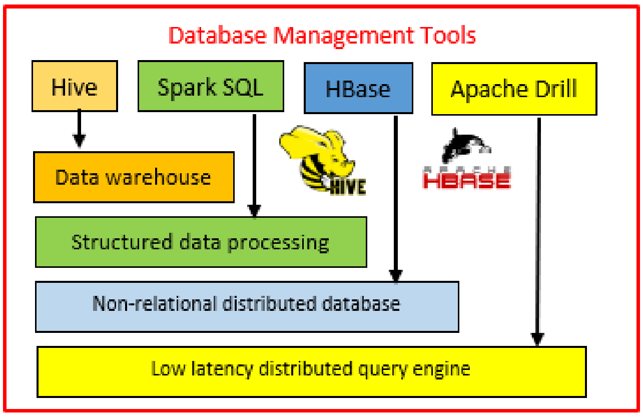
Database Management Tools
Hive
Spark SQL
HBase
Apache Drill
Data warehouse
HIVE
APACHE HBASE
Structured data processing
Non-relational distributed database
Low latency distributed query engine

Figure 4 Database platform management frameworks

IV. Big Data Health Care Databases

There are various discriminated health care data such as social media data, Genomic data sets, Electronic Medical Records, medical ontologies data, RNA, DNA data and public health datasets. Traditional database system cannot handle storage management of large volume of big data of health care system. The storage cost of health care data can be reduced by utilizing various big data frameworks data storage tools. Huge data has heterogeneous nature. EMR also called EHR is defined in the beginning health analytical data. MRI scanned data base need systematic analysis of discrimination of medical information due to large volume real time data. Table 1 shows medical analytics using big data analytics benefits. Healthcare data uses different streaming of data during loading, extraction and transmission. Network parameters like processing time, processor utilization time, Latency, throughput, execution time, sustainable input rate, execution time, task performance, scalability, and fault tolerance. Comparative statement of different frameworks has been presented in Table 2. From our experiment we can say that Flink work best compared to other frameworks. Spark also having good performance, among above network parameters. The main novelty of this study is to understand the data frameworks capabilities, this would be helping the researchers to select specific frameworks for health care data analysis. Flink yields efficient during measuring of processing time, compared to Storm and Spark.

Table 1. Medical analytics using Big data techniques

BDA for healthcare	Description of data storage	Literature
Medical Analysis	Pre identification of diseases based on symptoms and taking necessary treatment with primary investigation	[23]
Community healthcare	Prevention of diseases should be carried out in order to reduce risk factors.	[24]
Medical care monitoring	The present situation of Hospital monitoring has been carried out with respect to records of healthcare	[25][26]
Patient care system	Fast relief treatment can be given to prevent multiple times admitting to hospital	[28]

V. Conclusion

In medical analytics and healthcare systems managing huge amounts of data is a challenging task and handling critical decisions on the patients is also a difficult job. The large set of medical data is accessed in the Hadoop system by HDFS and map-reducer which makes accurate classification of data.

Table 2. Comparison between performance of frameworks

Parameter	Health care for	Hadoop	Spark	Flink	Storm
Processing time	Big data	Less fast	High speed	Low speed	Low speed
	Small dataset	Slow	Fast	Slow	Slow
CPU utilization time	Batch mode	High	High	Less	Less
	Store mode	Less	High	Less	High
Latency	RAM 3 S framework	Low	High	Low	Low
	Different group of databases	Low	Low	Low	High
Throughput	RAM 3 S framework	Low	Low	High	High
	Different group of databases	Low	Low	High	High
Fault tolerance	Fault and errors	Low	Low	High	Low

This paper addresses the speculative approaches of the medical healthcare system. Apache Storm supports real time processing of medical data with high scalability by handling continuous messages until it is ended by users. State of art mentioned in this paper highlights the various benefits of big data analytics. This review work compares the different Hadoop tools in terms of medical analytics. Real time health monitoring is influenced by big data framework stream processing such as Apache Flink. This work helps for various researches for finding enormous and enhancement of healthcare data analytics in real time life.

References

[1] S. Agarwal state of fast data and streaming applications survey, 2017.

[2] M. Mohammadi and A. Al-Fuqaha, “Enabling Cognitive Smart Cities Using Big Data and Machine Learning: Approaches and Challenges,” IEEE Communications Magazine, vol. 56, no. 2, pp. 94–101, 2018.

[3] Gani, A., Siddiqa, A., Shamshirband, S., &Hanum, F.: A survey on indexing techniques for big data: taxonomy and performance evaluation. Knowledge and Information Systems, 46(2), 241–284, 2016.

[4] Gubbi, J., Buyya, R., Marusic, S., &Palaniswami, M.: Internet of Things (IoT): a vision, architectural elements, and future directions. Future Generation Computer Systems, 29(7), 1645– 1660, 2013.

[5] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, “Internet of things for smart cities,” IEEE

Internet of Things Journal, vol. 1, no. 1, pp. 22–32, 2014.

[6] H. Isah, T. Abughofa, S. Mahfuz, D. Ajerla, F. Zulkernine and S. Khan, "A Survey of Distributed Data Stream Processing Frameworks," in *IEEE Access*, vol. 7, pp. 154300-154316, 2019.

[7] Manovich L, Trending: the promises and the challenges of big social data. In: Gold MK (ed) Debates in the digital humanities. University of Minessota Press, Minneapolis, pp 460–475, 2012.

[8] Burgess J, Bruns A Twitter archives and the challenges of "Big Social Data" for media and communication research. M/C J 15(5), 1–7, 2012.

[9] U. Khan, J. P. Choi, H. Shin, and M. Kim, "Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare," in Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, 2008, pp. 5148–5151.

[10] A.Andreescu et al., "Empirically derived decision trees for the treatment of late-life depression," American Journal of Psychiatry, vol. 165, no. 7, pp. 855–862, 2008.

[11] M. Shouman, T. Turner, and R. Stocker, "Applying k-nearest neighbour in diagnosing heart disease patients," International Journal of Information and Education Technology, vol. 2, no. 3, pp. 220–223, 2012.

[12] M. P. Brown et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," Proceedings of the National Academy of Sciences, vol. 97, no. 1, pp. 262–267, 2000.

[13] Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," IEEE transactions on information technology in biomedicine, vol. 14, no. 4, pp. 1114–1120, 2010.

[14] T. J. Hannan, "Electronic medical records," Health informatics: An overview, vol. 133, 1996.

[15] X. Fei, X. Li, and C. Shen, "Parallelized text classification algorithm for processing large-scale TCM clinical data with MapReduce," in Information and Automation, 2015 IEEE International Conference on, 2015, pp. 1983–1986.

[16] M. Bittorf et al., "Impala: A modern, open-source SQL engine for Hadoop," in Proceedings of the 7th Biennial Conference on Innovative Data Systems Research, 2015.

[17] R. Lin, Z. Ye, H. Wang, and B. Wu, "Chronic Diseases and Health Monitoring Big Data: A Survey," IEEE Reviews in Biomedical Engineering, 2018.

[18] T. Jones, "Process real-time big data with Twitter Storm," IBM Technical Library, 2013.

[19] M. Zaharia et al., "Apache spark: a unified engine for big data processing," Communications of the ACM, vol. 59, no. 11, pp. 56–65, 2016.

[20] R. Lin, Z. Ye, H. Wang, and B. Wu, "Chronic Diseases and Health Monitoring Big Data: A Survey," IEEE Reviews in Biomedical Engineering, 2018.

[21] D. Lyubimov and A. Palumbo, Apache Mahout: Beyond MapReduce. CreateSpace Independent Publishing Platform, 2016.

[22] A.Thusoo et al., "Hive: a warehousing solution over a map-reduce framework," Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1626–1629, 2009.

[23] Marco Viceconti, Peter Hunter and Rod Hose, "Big Data Big Knowledge: Big Data for Personalized Healthcare", *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1209-1215, 2015.

[24] Naoual El aboudi and Laila Benhlima, "Big Data Management for Healthcare Systems: Architecture Requirements and Implementation", *Advances in Bioinformatics*, pp. 1-10, 2018.

[25] Yichuan Wanga, LeeAnn Kungb, William Yu Chung Wangc and Casey G. Cegielskid, "An Integrated Big Data Analytics-Enabled Transformation Model: Application to Health care", *Elsevier Journal of Information & Management*, vol. 55, pp. 64-79, 2018.

[26] Sari I. Lakkis, Maher Elshakankiri, "IoT based emergency and operational services in medical care systems", *Internet of Things Business Models Users and Networks 2017*, pp. 1-5, 2017.

[27] B.Thillaieswari, "Comparative Study on Tools and Techniques of Big Data Analysis", International Journal of Advanced Networking & Applications (IJANA) Volume: 08, Issue: 05 Pages: 61-66, 2017.

[28] B.Prasanna, A.Prema, K.Chelladurai , "E-Health for Security and Privacy in Health Care System Using Hadoop Map Reduce", International Journal of Advanced Networking & Applications (IJANA) Volume: 08, Issue: 05 Pages: 101-104 , 2017.

AUTHORS BIOGRAPHY

First-Author received M. Tech. and PhD degree in Computer Science and Engineering from Visvesvaraya Technical University (VTU) Belgaum. India, in 2009 and 2018 respectively. From 1994 to 2009, he worked as Lecturer, and became Assistant Professor in 2010 in department of Computer Science and Engineering at the SDM Institute of Technology, Ujire, and Mangalore, India. Affiliated to VTU Belgaum, India. From 2018 he is working as Associate Professor in Information Science and Engineering SDMIT, Ujire. First-Author is the author of over 26 technical publications, proceedings. His research interests include Image processing, Computer Networks, Software engineering, Software architecture and Big data analytics

Second-Author received MTech. degree in Computer Science Engineering from the NITK Surthkal India, and Ph.D. degree in Computer Science and Engineering from the Jawaharlal Technological University (JNTU) Hyderabad, India in 1999 and 2007, respectively. From 1999 to 2005, she worked as Assistant Professor in Computer Science Engineering department of SDM Engineering College Dharwad. She is currently a Professor in the Department of Computer Science and Engineering of Vasvi College of Engineering, affiliated to Osmania University, Hyderabad, India. Second-Author is the author of over 41 technical publications. Her research interests include Image Processing, Pattern Recognition, Artificial Intelligence Neural Networks, Data communication and Interfacing, Data mining and Computer Organization. Recognized as supervisor from Osmania University, Reviewer for International journals of Information Fusion. And ACM journal