Survey on Big Data Analytics Platforms and Tools Applied in Healthcare System

R.Ramani
Asso Professor, Department of CSE,
P.S.R Engineering College,
Sivakasi, Tamilnadu, India.

Dr.K.Vimala Devi
Professor, Department of CSE,
Velammal Engineering College,
Chennai, Tamilnadu, India.

Abstract *– Gathering data in Healthcare framework for breaking down and overseeing it for the prosperity of individuals' lives, by contributing not exclusively to see new infections yet in addition to foresee results at prior stages and to settle on ongoing choices isn't that simple. These days social insurance information is created in immense sums by cell phones, sensors, patients, emergency clinics, scientists, suppliers and associations. The essential hotspots for gathering patient's information are The Electronic Medical Records (EMRs) and Electronic Health Records (EHRs).Hadoop is utilized for dealing with Big Data like clinic records of patients. Huge Data investigation is utilized for viable basic leadership in social insurance area. Distinctive client can see the outcomes in various organization utilizing perception procedures. The review of this paper portrays how to deal with enormous information and furthermore the methodologies for investigation and expectation.*

Keywords: Big Data Analytics, Health care System, EMRs, EHRs, Hadoop, Visualization techniques.

I. INTRODUCTION

Big Data Analytics has come out with two distinct concepts, i.e., Big Data and Analytics. Together it gives meaningful information from large volume of data. Benefits of Big Data Analytics are Cost Maintenance, Customer Satisfaction and Performance. Healthcare data which is collected by the healthcare industry is huge and practically, it is not so easy to handle manually. Origin of Big Data in healthcare is from the large electronic health datasets. It is very difficult to manage these datasets with the conventional hardware and software. Almost 80% of the healthcare data is unstructured so it brings a challenge for the healthcare industry to make sense of all the unstructured data and leverage it effectively for predicting diseases in the earlier stage. Storing huge volume of data and processing such data is a challenging task in traditional technology. To overcome this problem, Big Data Analytics uses Hadoop framework, as Map Reduce engine and HDFS has the capability to process thousands of petabytes of healthcare data. Hadoop makes use of cheap commodity hardware.

A. Big data Architecture

Big data architecture is the structure how big data will be stored, accessed and managed within big data. How the big data solution will work, the core components (hardware, database, software, storage) used, flow of information is logically defined by Big Data architecture as shown in the figure. This architecture is described in the following sections. We discuss related work in section II; describe the Hadoop ecosystem in section III, present various visualization techniques in section IV.

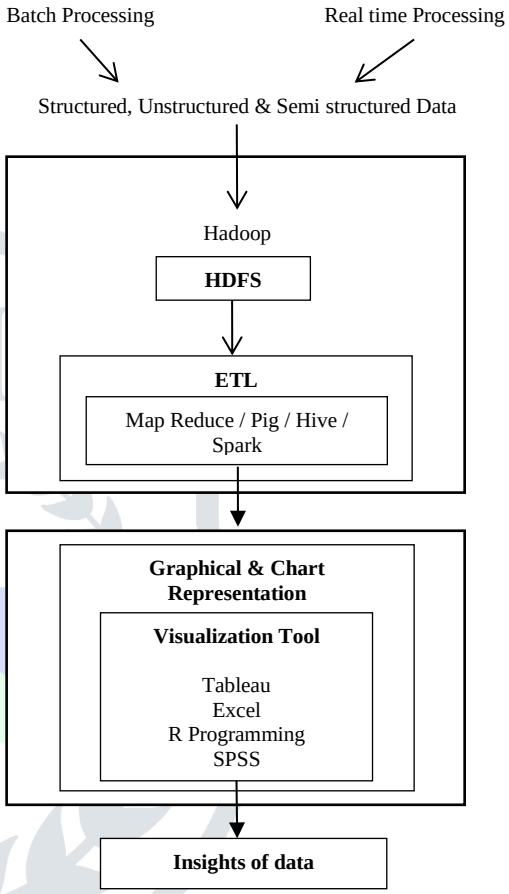

Fig. 1. Big Data Architecture

A.1 Data Source

The one of the most important steps in deciding the architecture is source profiling. It is used to identify the different source systems and categorizing them based on their nature and type.

Key Features:

- The huge amount of data collected from various sources.
- An internal & external source includes Database, File, web service, streams etc.
- Type of data may be structured, semi structured or unstructured.

A.2 ETL

ELT (Extract, Load and Transform) means data ingestion in Hadoop as opposed to ETL (Extract, Transform and Load) in case of traditional warehouses.

Key Features:

- Data extracting from each source.
- If required any data validation or transformation after ingestion is called as Pre-processing.
- Based on mode of ingestion such as batch or real-time processing, segregate the data sources.

A.3 Storage

Any type of data like structured, unstructured and semi-structured can be stored in large amount. The most commonly used storage framework in Big Data world is Hadoop distributed file system. NoSql data stores like MongoDB, HBase, and Cassandra etc are also available for data storage. Capability to scale, self-manage and self-heal are the salient features of Hadoop storage.

It can support for two kinds of analytical requirements:

- Synchronous – Data is analyzed in real-time or near real-time, the storage should be optimized for low latency.
- Asynchronous – Data is captured, recorded and analyzed in batch.

Key Features:

- Type of data may be historical or streaming data.
- Compression requirements.
- Frequency of incoming data.
- Query pattern on the data.
- Consumers of the data.

A.4 Processing

Earlier dynamic RAM is used for storing frequently accessed data .But now a days, it is been stored on multiple disks on a number of machines connected via the network due to the sheer volume. The new way of processing the data is taken closer to data which significantly reduce the network I/O. It can be categorized into Batch, real-time or Hybrid based on the SLA.

- *Batch Processing* – Processing the input in a scheduled way by collecting it for a specified interval of time. Historical data load is a typical batch operation

  Technology Used: MapReduce, Hive, Pig
- *Real-time Processing* – It involves processing whenever data is acquired.

  Technology Used: Impala, Spark, Spark SQL, Tez, and Apache Drill
- *Hybrid Processing* – It's a combination of both batch and real-time processing.

A.5 Consumption & Visualization

Output provided by processing layer is consumed by this layer. Users like administrator, Business users, vendor, partners etc. can consume data in the format they wish to have. Recommendation engine consumes output of analysis and based on the analysis business processes can be triggered.

Different forms of data consumption are:

- *Export Data sets* – As per the requirements of third party, data sets can be generated using hive export or directly from HDFS.
- *Reporting and visualization* – Hadoop can be connected to different reporting and visualization tools using JDBC/ODBC connectivity to hive.
- *Data Exploration* – Sandbox environment is a separate cluster or a separate schema within same cluster that contains subset of actual data and where the Data scientist can build models and perform deep explorations.
- *Adhoc Querying* – Hive, Impala or spark SQL supports Adhoc or Interactive querying.

The type of data and the type of processing should be considered while designing Big Data Architecture. There are multiple technologies offering similar features and claiming to be better than the others.

B. Healthcare Architecture

The clinical information, specialist's composed notes and solutions, restorative pictures, for example, CT and MRI examines results, lab records, drugstore reports, protection documents and other regulatory information, electronic patient records (EPR) information, and internet based life posts, for example, tweets, reports on site pages and so on can be on the whole considered as Healthcare huge information. The totality of administrations offered by all wellbeing disciplines is known as a social insurance framework. This framework is to help the specialists in making convenient and viable determinations by recovering, refreshing, and revealing the patient data productively by encouraging the middle. Diminishing the expenses of treatment, anticipate episodes of pestilences, maintain a strategic distance from preventable illnesses and enhance the personal satisfaction when all is said in done can be adequately done by Healthcare examination. The significant advancements in information the executives, information gathering through electronic medicinal records are profited for Healthcare industry. Information drives a significant number of the choices behind everybody living longer, quickly changing models of treatment conveyance. The drive presently is to comprehend however much about a patient as right on time as could be expected at sufficiently early stage in their life by getting genuine cautioning indications of disease and by treating with far straightforward and more affordable than if it had not been spotted until some other time.

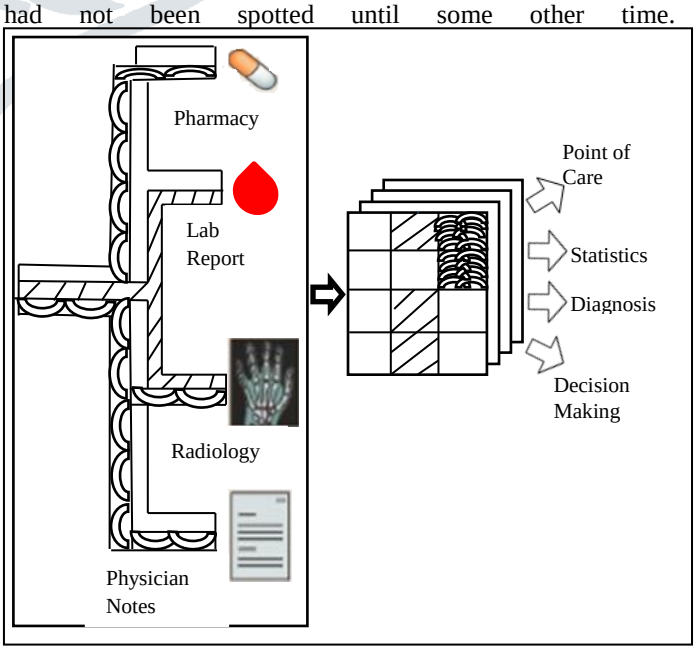

Fig. 2. Healthcare Architecture

II. LITERATURE SURVEY

Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding[1] highlighted data literally concerned about data volumes, here it presented a HACE theorem which suggested that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations. To unleash the full power of the Big Data, High-performance computing platforms are required to support Big Data mining, which imposed systematic designs. In all science and engineering domains, as Big Data is an emerging trend, the need for Big Data mining is arising. By using Big Data technologies, the most relevant and most accurate social sensing feedback to better understand our society at real time can be hopefully achieved.

MimohOjha, Dr. KirtiMathur[2] highlighted the advantages of storing digitalized data in healthcare. They introduced big data technologies to manage those data and to obtain deeper insights for making clinical decisions. A survey at M.Y.hospital, Indore on 150 people had been conducted by them. The relationship between the current working of hospital, current technologies available in the market and how the big data technologies can be used to improve the healthcare facilities at the hospital are examined by them. The parameters like name, age, city, annual income & yearly medical expenses and health insurance are outlined by them. The problems of both patients and doctors are highlighted by the results such as 52 % of people agreed that they wait in a long queue, 51% of people said the facilities offered by the hospital is average and it needs improvement, the majority of the people came from poor background and they cannot afford private hospitals, also only 58% of people have health insurance. In the proposed work, to store patient data digitally and generate a specific unique number for easy retrieval they maintained EHR. The hospital can offer modern facilities to the patients by reducing cost and time by using digitalized data. The use of HDFS to store huge amount of data and Hadoop to analyze the data are required by such EHR. The patient data can be stored in cloud, by the unique number the data can be accessed anytime and anywhere in future is the most valuable and to be followed suggestion given by them.

J.Archenaa and E.A.Mary Anita[3] gave a clear analysis of how Big Data Analytics using Hadoop plays an effective role in performing meaningful analysis on the huge volume of data in healthcare and able to predict the emergency situations before it happens. It concluded that the problem is the lack of information that can be used to support decision-making, planning and strategy but not the data.

H. Gilbert Miller and Peter Mork[4] highlighted a framework that can examine and bring disparate data together to create valuable information that can inform decision making at the enterprise level in an organized way. It accomplished through 1) Data Discovery - Collect and annotate, Prepare and Organize, 2) Data Integration,3) Data Exploitation - Analyze Visualize and Make decisions.

III. PLATFORM & TOOLS FOR HANDLING BIG DATA

A. Hadoop Ecosystem

Hadoop is a combination of HDFS and MapReduce. Hadoop ecosystem consists of different level of layers, storing data, processing stored data, resource allocating and supporting different programming languages to develop various applications are the tasks performed by each layer respectively. The Hadoop ecosystem contains different tools such as Sqoop, Pig, and Hive that are used to help Hadoop modules.

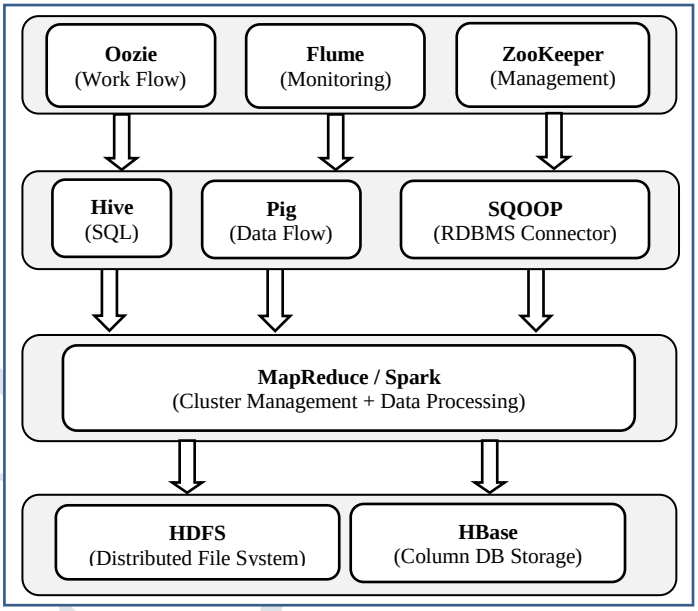

Fig. 3. Hadoop Ecosystem Components

A.1 Hadoop Distributed File System Architecture

HDFS is a file system designed in a way, which can store very large files with streaming data access patterns, running on clusters of commodity hardware. Storing the data in distributed manner in order to compute fast is a technique carried out by HDFS which is a main component of Hadoop. Data in HDFS is stored in blocks of 64MB by default or 128 MB in size which is logical splitting of data in a Datanode in Hadoop cluster. The metadata, which contains all information about data splits in Datanode and is captured in Namenode which is again a part of HDFS. The concept behind storing a file in HDFS is Write Once and Read Many.

HDFS has master/slave architecture. An HDFS cluster consists of a single Namenode, the file system namespace and regulation of access to files by clients is managed by a master server. HDFS file system namespace operations like opening, closing, and renaming files and directories are executed by the Namenode. It also determines the mapping of blocks to Datanodes. The Namenode also manages the list of HDFS files belonging to each block, the current location of the block replicas on the Datanodes, the state of the file, and the access control information which is the metadata for the cluster. The read and write requests from the HDFS file system's clients are served by the Datanodes. They also perform block replica creation, deletion, and replication upon instruction from the Namenode. Each block is replicated some number of times, each replica is allocated on a different Datanode. For HDFS, the default replication factor is three.

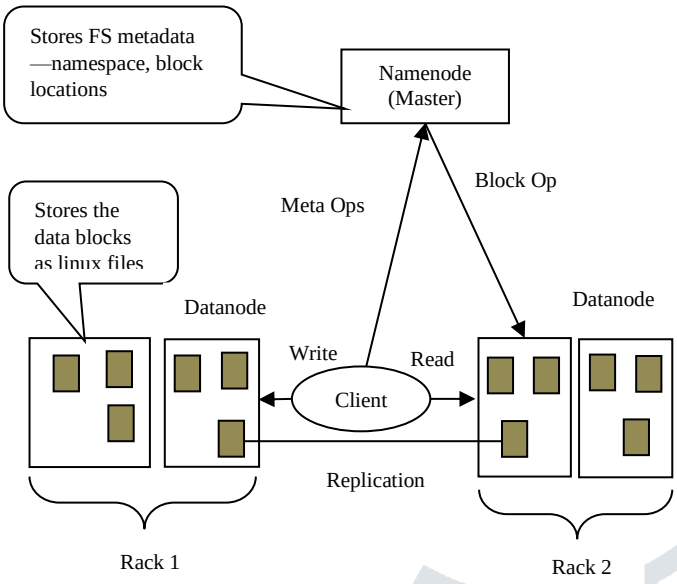

Fig. 4. HDFS Architecture

A.2 MapReduce

The framework used for processing large amount of data on commodity hardware on a huge dataset of cluster ecosystem is called as MapReduce. The two important functions of the MapReduce algorithm are: Map and Reduce.

- *Map Function*

In Map function, individual elements are considered as key/value pairs. It converts a typical dataset into another set of data.

Map (key1, value1) ->List (key2, value2)

- *Reduce Function*

In Reduce function, the output files from a map are considered as an input and then it integrates the data tuples into a smaller set of tuples.

Reduce (key2, List (value2)) ->List (key3, value3)

Any language like JAVA, PYTHON, etc can be used to write MapReduce program. Functionality of MapReduce is mapping of logic into data and once computation is over, the result of Map is collected by the reducer to generate final output result of MapReduce. Whether the data is Structured or Unstructured stored in HDFS, MapReduce Program can be applied.

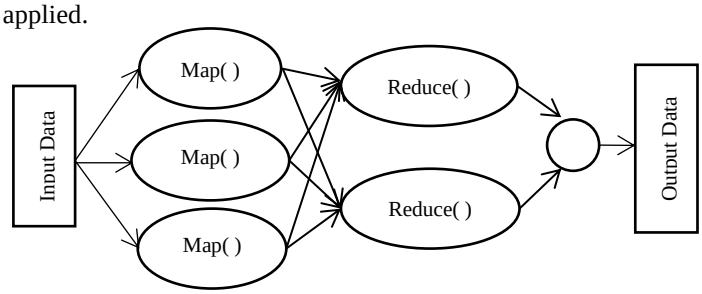

Fig. 5. Map() and Reduce() Function

The MapReduce work flow undergoes different phases and the end result will be stored in HDFS. Client submits the job into Jobtracker and the Jobtracker assign Tasktrackers to coordinate map and reduce phases provide job progress information. All MapReduce jobs that are running on various nodes present in the Hadoop cluster are taken cared by Jobtracker. It plays vital role in scheduling jobs and keeps track of the entire map and reduce jobs. The actual MapReduce process happens in Tasktracker. Intermediate process will take place in between map and reduce stages. Shuffle and sorting of the mapper output data are the operations carried out in the intermediate process. The Intermediate data is going to get stored in local file system.

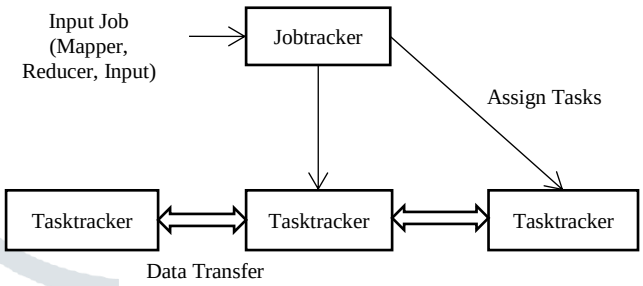

Fig. 6. MapReduce Workflow

A.3 HBase

HBase is a distributed column-oriented database built on top of the Hadoop file system. HBASE was created for large table which have billions of rows and millions of columns with fault tolerance capability and horizontal scalability and based on Google Big Table. HBASE is used for random access of huge data whereas Hadoop can perform only batch processing, and data will be accessed only in a sequential manner. The data can be stored in HDFS either directly or through HBase. Data consumer can reads/accesses the data randomly in HDFS using HBase. HBase sits on top of the Hadoop File System and provides read and write access.

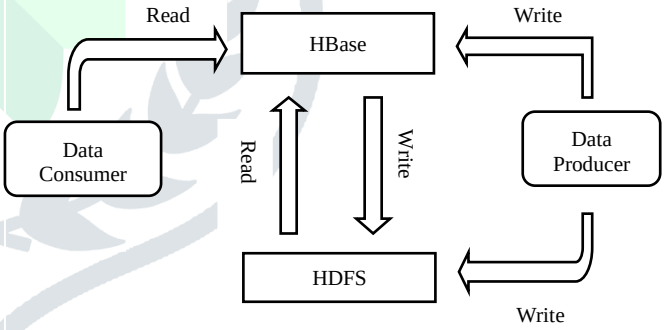

Fig. 7. HBase Architecture

A.4 Spark

Spark is both a programming model and a computing model. A gateway to in-memory computing for Hadoop is provided by it, which is a reason for its popularity and wide adoption. As an alternative to MapReduce that enables workloads to execute in memory, instead of on disk we can go for Spark. Spark accesses data by bypassing the MapReduce processing framework from HDFS, and thus eliminates the resource-intensive disk operations which is required by MapReduce. Spark workloads typically run between 10 and 100 times faster compared to disk execution, by using in-memory computing. Spark can be used independently of Hadoop. However, it is used most commonly with Hadoop as an alternative to MapReduce for data processing. Spark can easily coexist with MapReduce and with other ecosystem components that perform other tasks. Spark supports SQL, which helps

overcome a shortcoming in core Hadoop technology, so it is popular. The Spark programming environment works interactively with Scala, Python, and R shells. It has been used for data extract/transform/load (ETL) operations, stream processing and machine learning development with the Apache GraphX API for graph computation and display. It can run on a variety of Hadoop and non-Hadoop clusters, including Amazon S3.

A.5Sqoop

SQOOP stands for SQL to Hadoop. Sqoop is a tool used for efficiently transferring bulk data between Hadoop and structured data stores such as relational databases. Integrating Sqoop with Oozie, one can schedule and automate import and export tasks. To interact with the traditional business applications, relational database systems are widely used. So, one of the sources that generate Big Data is relational database systems. As we are dealing with Big Data, To achieve benefit of distributed computing and distributed storage, Hadoop stores and processes the Big Data using different processing frameworks like MapReduce, Hive, HBase, Cassandra, Pig etc and storage frameworks like HDFS. Data needed to be transferred between database systems and Hadoop Distributed File System (HDFS), In order to store and analyze the Big Data from relational databases. Sqoop acts like an intermediate layer between Hadoop and relational database systems.

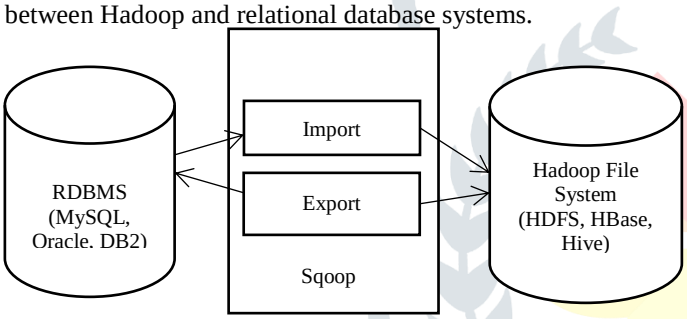

Fig. 8. Sqoop Architecture

A.6 Pig

It is a procedural language platform used to develop a script for MapReduce operations. To provide a rich set of datatypes and operators to perform various operations on the data a high level data processing language called Pig Latin is used. It is used by the Programmers to write a Pig script, and execute them using any of the execution mechanisms like Grunt Shell, UDFs, and Embedded. These scripts will go through a series of transformations applied by the Pig Framework, to produce the desired output after execution. Internally, these scripts are converted into a series of MapReduce jobs, and thus, it makes the programmer's job easy by Apache Pig. It is provided as an alternative to programmer who doesn't want to use Java/Python or SQL and loves scripting to process data.

A.7 Hive

Hive is a platform which is used to develop SQL type scripts to do MapReduce operations. To process structured data in Hadoop a data warehouse infrastructure tool is used, which is nothing but Hive. To summarize Big Data and make querying and analyzing easy, Hive resides on top of Hadoop. Hive is created by Facebook for Many programmers and analysts who are more comfortable with Structured Query Language than Java or any other programming language and later it was donated to Apache foundation. It mainly deals with structured data which is stored in HDFS with a Query Language similar to SQL and known as HQL (Hive Query Language). Hive also run MapReduce program in a backend to process data in HDFS but here programmer has not worried about that backend MapReduce job it will look similar to SQL and result will be displayed on console.

A.8 Oozie

Oozie is a workflow scheduler system to manage Hadoop jobs. It runs both as a server and a client which submits a workflow to the server directly. Running workflow jobs with actions that run Hadoop, MapReduce and Pig jobs is done by specialized Server-based Workflow Engine. It runs in a Java Servlet-Container as a Java Web-Application. Hadoop basically deals with Big Data and when jobs are required to be run in a sequential manner, like output of Job A will be input to Job B and similarly output of job B is input to job C and final output will be output of job C, Oozie is used to automate and execute this workflow sequence.

A.9 Flume

A distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amount of streaming data into the Hadoop Distributed File System is done by Flume. Individual Flume agents collect data generated by data generators such as Facebook, Twitter, which are running on them. Thereafter, the data from the agents which is aggregated and pushed into a centralized store such as HDFS or HBase is collected by a data collector.

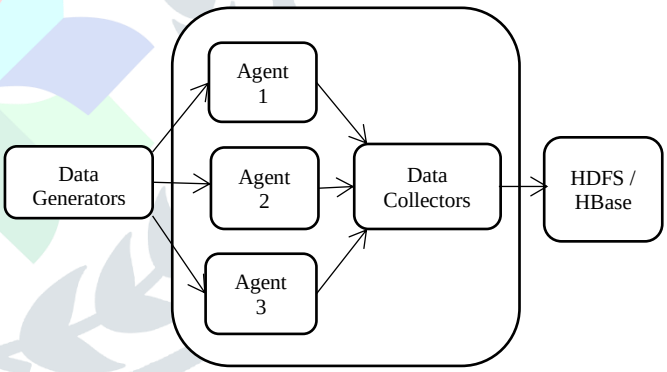

Fig. 9. Flume Architecture

A.10Zookeeper

Zookeeper is based on a simple Client Server model. Maintaining configuration information, naming, providing distributed synchronization, and providing group services is done by a centralized service called Zookeeper. The nodes which request the server for service are the clients and the node which serves the requests is the server. There can be multiple Zookeeper servers which form the ensemble. When services get started, a leader is elected from the group of these servers. Each client is connected to only one server and all the reads are performed by that server only. When a 'write' request comes to a server, it is first sent to the leader, the leader then asks the quorum. Quorum is a strict majority of nodes available in the ensemble that decide on this request. If the quorum responds positively, the 'write' request is considered successful. That's why, a 'write' request takes more time than a 'read' request and when there are fewer 'write' requests and more of 'read'

requests it is suggested that Zookeeper should be implemented in the distributed system. In case of any partial failure, clients can connect to any node and be assured that they will receive the correct, up-to-date information.

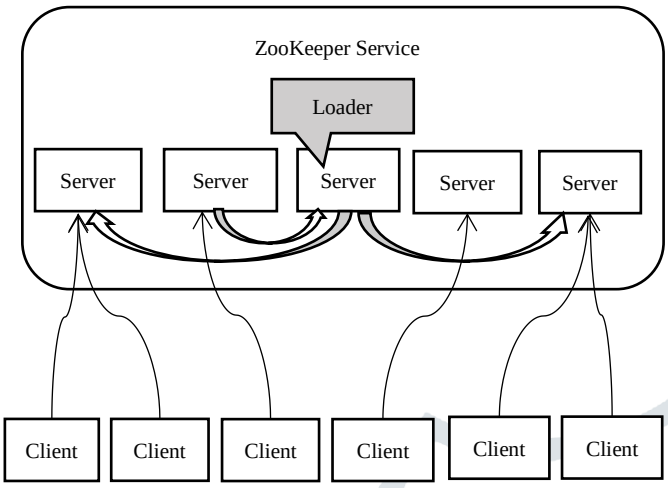

Fig. 10. Zookeeper Architecture

The following table compares the traditional technology with the Hadoop and depicts the overview of big data analytics platform and tools along with its advantages.

TABLE I. OVERVIEW OF BIG DATA ANALYTICS – PLATFORM & TOOLS AND ITS BENEFITS

Concept	Technology			Data Collection (Tools)	Type of Data	Components		Benefits
Big Data Analytics	Hadoop	Version I	Advantages 1.Large volume of data can be stored 2.Less processing time	1.Flume 2.Kafka	Streaming Data	HDFS Hbase	Storage	1.Cost Maintenance 2. Customer Satisfaction 3.Performance
		Version II – YA RN		3.Sqoop	Historical Data	MapReduce Spark	Processing	
	Traditional	-	Limitations 1. To store large volume of data 2. More processing time	-	-	-	-	

IV. DATA VISUALIZATION

Data visualization is the graphical representation of information, with the goal of providing the viewer with a qualitative understanding of the information contents. Regardless of the size of the data, visualizing data is important because it translates information into insight and action. It enables decision makers to see analytics presented visually, so that they can easily grasp difficult concepts or identify new patterns. Business users can understand analytics insights and actually see the reasons why certain recommendations make the most sense with the help of visualizations. As the cost of storing, preparing and querying data is much higher, the approach for visualizing Big Data is especially important. Therefore, organizations must leverage well architected data sources and rigorously apply best practices to allow knowledge workers to query Big Data directly. Few visualization techniques are discussed here to acquire an idea for visualizing data.

A. Tableau

Tableau Software helps to perceive and understand data easily by the user like business users, administrator, vendor etc. Tableau allows you to quickly connect, visualize, and share data with a seamless experience from the PC to the iPad by offering a revolutionary new approach to business intelligence.

Key Features:

- Once online, others can download and manipulate visualizations.
- Desktop application but completed graphics are stored on a public server.
- Store up to 50MB of data with free plan.
- No programming skills required just drag-and-drop interface.

B. R programming

R is highly extensible and provides a wide variety of statistical like linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering and graphical techniques and also provides an Open Source route to participation in such activity.

Key Features:

- Data manipulation, calculation, and graphical display.
- Integrated tools for instant analysis.
- Conditions, loops, user-defined recursive functions, and input/output facilities.
- Define new functions for increased capabilities.

C. Excel

Microsoft Excel is often used to create powerful data visualizations and it is also noted for data manipulation and analysis capabilities. The visualization tools, including recommended charts, quick analysis of the different ways to display the data, and a multitude of control options to change the look and layout of visualizations are wrapped and provided in the latest edition.

Key Features:

- In the same program, perform data analysis and create visualizations.
- Compare various ways to represent the data.
- Change tile, layout and other format options.
- Excel recommends the best visualization for the data.
- Compatible with Microsoft Office products.

D. SPSS

A package used for statistical analysis, Data mining, Text analytics, Data collection, Collaboration & Deployment is offered in Statistical Package for the Social Sciences. SPSS has the capability to handle large amount of data and perform all of the analyses.

Key Features:

- Calculate Descriptive Statistics.
- Compare Means.
- Conduct Cross-Tabulations.
- Recode Data.
- Create Graphs and Charts.

V. IMPLEMENTATION OF BIG DATA IN HEALTHCARE

The overview of implementing big data in healthcare is to analyze the patient data and observing the patients' health. For example, to find the list of patients those who are all likely to be affected by diabetics.

The patient data which is collected from various sources is a structured data and is stored in HDFS. Spark or MapReduce are used to process such data. The normal BP and diabetic level of the humans are maintained in a separate lookup file. During processing, the stored data relates with the lookup file, reports the affected patients list and the necessary information is fetched using query. Using Tableau the generated list is visualized and based on that decision can be taken.

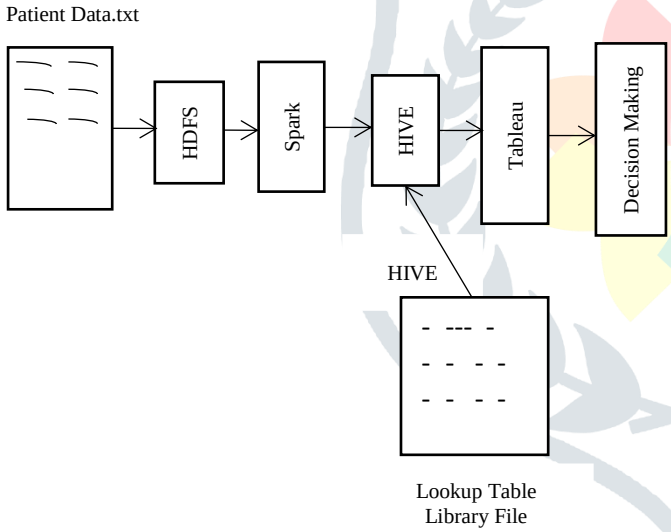

Fig. 11. Implementing Big Data in Healthcare

VI. CONCLUSION

Big data analytics in healthcare is a mixture of clinical innovation and technology altogether. It is almost impossible to manage data over soft or hard copy formats, as the healthcare industry is continuously generating large amount of data in different forms. Every patient has their own digital record which includes demographics, medical history, allergies, laboratory, test results etc. This big data allows for early identification of illness of individual patients. The goal is to help doctors make big data informed decisions within seconds and improve patients' treatment. This is particularly useful in case of patients with complex medical histories, suffering from multiple disorders.

References

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, "Data Mining with Big Data", *IEEE Trans. on Knowledge and Data Engg*, vol. 26, no. 1, pp. 97–107, Jan 2014.

[2] J.Archenaa and E.A.Mary Anita, "A Survey of Big Data Analytics in Healthcare and Government", Procedia Computer Science (Elsevier) 50 (2015) 408 – 413.

[3] H. Gilbert Miller and Peter Mork, "From Data to Decisions: A Value Chain for Big Data", 1520-9202/13/$31.00 pp. 57-59 © 2013 IEEE.

[4] Dharavath Ramesh, PranshuSuraj and Lokendra Saini, "Big data Analytics in Healthcare: A Survey Approach", International Conference on Microelectronics, Computing and Communications (MicroCom), IEEE, ISBN: 978-1-4673-6621-2, 2016.

[5] Madhura A. Chinchmalatpure, Dr. Mahendra P. Dhore, "Review of Big data Challenges in Healthcare Applications" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, PP 06-09, 2016.