Healthcare Analysis Using Big Data

Archanendra Srivastava
Inderprastha Engineering College
Dr. A.P.J Abdul Kalam University
Ghaziabad (U.P.), India

Chahat Sharma
Asst. Professor,
Inderprastha Engineering College,
Ghaziabad, India.

***Abstract*—Healthcare Analytics is a term used to describe the healthcare analysis activities that can be undertaken as a result of data collected from areas within healthcare that includes claims and cost data, pharmaceutical, research and development (R&D) data, clinical data (collected from electronic medical records (EHRs)), and patient behavior and sentiment data (patient behaviors and preferences) and other types of health related data too. Big Data in healthcare is being used to predict epidemics, cure diseases and to avoid preventable deaths. The main aim of this paper is to describe the nascent field of big data analytics in healthcare and to perform real-time analytics on healthcare using Hadoop and RStudio, and help executives and research professionals from hospitals, health systems and other healthcare provider organizations to identify and understand big data analytics for innovative uses of data assets that can enable them to reduce costs and provide more accessible care.**

***Index Terms*— Big data, healthcare, healthcare analytics**

1. INTRODUCTION

A. Healthcare Analytics

Healthcare Analytics is used to analyze patient's medical records and provide them better and cost-effective care. Today, population of our country is increasing rapidly and so the medical data, but the treatment provided to them is not up to the mark. Therefore Healthcare Analytics is a term used to provide a treatment based on analysis instead of theory-based treatment, so that patient can get better care.

In current scenario, Healthcare providers face significant obstacles in implementing analytical decisions, business intelligence tools and data warehousing because the health data is diverse, comprising structured, semi-structured and unstructured information in a range of formats. At the heart of many healthcare industries, debate is what to do about data, how to realize its value for quality care and how to use it to bend the cost curve [1].

The global healthcare industry is experiencing fundamental transformation as it moves from a volume-based business to a value and quality based business. With increasing demands from consumers for enhanced healthcare quality and increased value, healthcare providers and payers are under pressure to deliver better outcomes [2].Challenges in Healthcare Analytics includes scalable computation, analytical platform and collaboration across different domains [3].Healthcare industries must identify and establish proven strategies and best practices to manage this diverse and distributed data.

B. Big Data

Big Data, as the name suggests, Big and Data, where Big stands for huge and enormous amount of data and Data here signifies the three types of data: Structured, semi-structured and unstructured data .Structured data can be defined as the data that is used to query and report against predetermined data types. It includes relational databases, flat files in form of records etc. Unstructured data might or might not have logical or repeating pattern and consist typically of metadata in different formats such as e-mails, videos, text and images. Semi-structured data includes data exchange formats such as JSON data, and web data in the form of cookies.

In 2012, Dr.Yan Mo won the Nobel Prize in literature and this was probably the most controversial Nobel Prize of this category, as Mo speaks Chinese, lives in a socialist country, and had the Chinese government's support. If we search on Google with "Yan Mo Nobel Prize" ,we get 1,050,000 web pointers on the internet. "For all types of praises and criticisms" said Mo," I am grateful" .What types of praises and criticism has Mo actually received over his 31-year writing career? Can we summarize all types of opinions in real-time fashion, including updated, cross-referenced discussion by critics. This type of summarization program is an excellent example for Big-Data processing, as the information comes from multiple, heterogeneous, autonomous sources with complex and evolving relationships, and keep growing [4].

If we talk about Flickr, a public picture sharing site, which has a traffic of 1.8 million photos per day. If we assume that the size of each photo is 2Mb, then the result is 3.6Tb storage, every single day.

As "a picture is worth a thousand words ",the billions of picture on Flickr are a treasure tank for us to explore the human society ,social events , public affairs ,disasters etc. only if we have the powers to harness the enormous amount of data. Along with above examples, the era of Big-Data has arrived (Nature Editorial 2008; Mervis J.2012; Labrandish and jagadish 2012).

Think of the following: Every second, there are around 8.22 tweets on twitter, and in every minute, nearly 510 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded on Facebook. Every hour, Walmart, a global discount departmental store chain, handles more than 1 million customer transactions and every day, consumers make around 11.5 million payments using PayPal. According to IBM, "Every day, we create 2.5 quintillions bytes of data, so much that 90% of the data in the world today has been created in the last two years alone". This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals.

Big Data is slowly becoming ubiquitous and every arena of business, health or general living standards can now implement Big Data [5]. We can also visualize Big Data technically by dividing its functionality into number of classes (Fig 1.2.1).

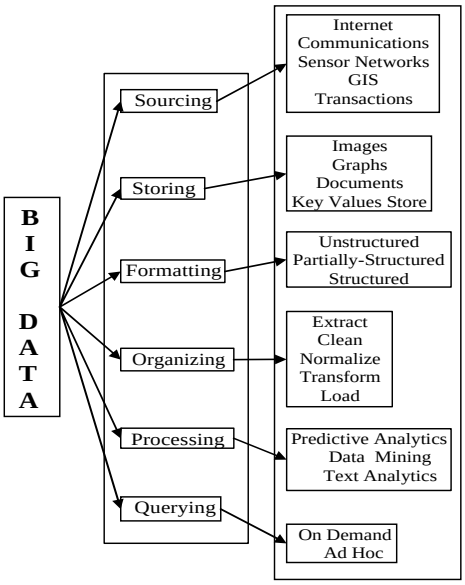


Fig 1.2.1 [6] Technical View of Big Data

C. Big Data in Healthcare

Data in Healthcare can be characterized as follows:

Genomic Data: It refers to genotyping, gene expression and DNA sequence [7].

Clinical Data and Clinical notes: About 80% of this data are unstructured documents, images and clinical or transcribed notes (Yang.et.al, 2014).

Behavioral Data and Patient Sentiment Data: Web and Social media data, mobility sensor data or streamed data.

Health publications and Clinical Representative data: Text based publications (journal articles, clinical research) and clinical text-based reference practice guidelines and wealth product data [8].

Big data in Healthcare is described by four primary characteristics: volume, velocity, variety and veracity. Over time, health-related data will be created and accumulated continuously, resulting in an incredible volume of data. The already daunting volume of existing healthcare data includes personal medical records, radiology images, clinical trial data FDA(Food and Drug Administration) submissions, human genetics and population data genomic sequences, etc. Newer forms of big data, such as 3D imaging, genomics and biometric sensor readings, are also fueling this exponential growth [9].

Most healthcare data has been traditionally static—paper files, x-ray films, and scripts. Velocity of mounting data increases with data that represents regular monitoring, such as multiple daily diabetic glucose measurements (or more continuous control by insulin pumps), blood pressure readings, and EKGs(electrocardiogram). Meanwhile, in many medical situations, constant real-time data (trauma monitoring for blood pressure, operating room monitors for anesthesia, bedside heart monitors, etc.) can mean the difference between life and death.

As the nature of health data has evolved, so too have analytics techniques scaled up to the complex and sophisticated analytics necessary to accommodate volume, velocity and variety. Gone are the days of data collected exclusively in electronic health records and other structured formats. Increasingly, the data is in multimedia format and unstructured. The enormous variety of data—structured, unstructured and semi-structured—is a dimension that makes healthcare data both interesting and challenging.

Structured data is data that can be easily stored, queried, recalled, analyzed and manipulated by machine. Historically, in healthcare, structured and semi-structured data includes instrument readings and data generated by the ongoing conversion of paper records to electronic health and medical records. Historically, the point of care generated unstructured data: office medical records, handwritten nurse and doctor notes, hospital admission and discharge records, paper prescriptions, radiograph films, MRI, CT and other images.

Veracity assumes the simultaneous scaling up in granularity and performance of the architectures and platforms, algorithms, methodologies and tools to match the demands of big data.

Veracity in healthcare data faces many of the same issues as in financial data, especially on the payer side such as is this the correct patient/hospital/payer/reimbursement code/dollar amount? Other veracity issues are unique to healthcare: Are diagnoses/treatments/prescriptions/procedures/outcomes captured correctly?

The '4Vs' are an appropriate starting point for a discussion about big data analytics in healthcare. But there are other issues to consider, such as the number of architectures and platforms, and the dominance of the open source paradigm in the availability of tools and the challenge of developing methodologies and the need for user-friendly interfaces. While the overall cost of hardware and software is declining, these issues have to be addressed to harness and maximize the potential of big data analytics in healthcare. A 2011 McKinsey report estimated that the healthcare industry can potentially realize $300 billion in annual value by leveraging big data.

Today, Information Technology functions are developing, not only as a technology services provider, but also as a strategic provider that can develop and integrate industries infrastructure to facilitate and ensure quality of service [10].

Big Data in healthcare helps us in reducing waste and inefficiency in the areas like clinical operations, genomic analytics, patient profile analytics etc.

D. Big Data challenges and strategies to leverage Big Data

The healthcare data is rarely standardized, often fragmented, or generated in legacy IT systems with incompatible formats. The diversity in healthcare data and the variety of data in healthcare, from various medical records and other sources results in healthcare providers facing challenges in implementing business intelligence (B.I.) tools, and data warehousing , as well as an overarching general reluctance among organizations to share their data .

- According to a report in 2011, only 34% analysts were able to capture data from EHR (Electronic Health Record) to help patients and 43% were unable to collect sufficient data to improve care. It will take significant efforts to shift attitudes and educate providers about available and emerging technologies [11].
- Most clinical data is in unstructured form, especially the Electronic Health Record data, making it difficult to access for effective analysis. Therefore, data usability has been identified as major issue especially with CDS (Clinical Decision Support).
- Maximum amount of data in healthcare sector is fragmented among labs, hospital systems, financial IT systems and Electronic Health Records, which is another obstacle in leveraging big data in healthcare. Smaller organizations with multiple systems and taxonomies always feel difficulty in extracting useful information for data mapping.
- Analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additional layer of complexity.
- Key to overcome these fragmentation obstacles is "normalizing" the data and shifting to a "culture of best practices", best experiences and using data from various components of health IT to improve care and lower costs in holistic way.

- We can leverage big data by implementing a data governance framework, a framework for enterprise wide data governance is necessary to ensure the success of any effort to leverage big data for healthcare.
- All organizations need to be able to make decisions regarding management of data and realizing value from data, how to minimize cost, complexity and risk, and how to ensure compliance with ever growing legal, regulatory, and other requirements. The physicians and nurses of hospitals and other healthcare provider organizations must be given training in analytics to understand how big data can add value to overall healthcare performance.
- Don't scale up, scale out: Some organizations may be prone to lean towards replacing their older servers with bigger and more powerful servers. Today's trend is to scale out; to improve performance and scalability of a system by adding nodes for processing and data storage. The approach may be worth considering because it can make systems easier to manage and to expand to accommodate big data solutions.

.

2. LITERATURE REVIEW

- Using analytics, healthcare companies can go beyond analyzing the significance of their historical data .They can explore 'what-if' scenarios without causing inconvenience or driving up costs. They can test what might happen if certain situations occurred, without actually performing physical experiments or simply relying on trial-and-error approach. Analytics can be seen as the key for successfully managing the organization.
- In their research work, Bhosale and Gadekar describes Big Data as innovative techniques and technologies to capture, store, distribute, manage and analyze pentabytes or larger sized data sets with high velocity and different structure. Big Data is a data whose scale, diversity and complexity require new architecture, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it [12].
- As described by Shvacko et al (2010), Hadoop HDFS is designed to store large datasets reliably, and to stream those data sets at high bandwidth to user applications. In their research, they describe the architecture of HDFS and their report on experience using HDFS to manage 25 PB of enterprise data at Yahoo[13].
- According to Cortade, Gordan and Leniban, Healthcare Organizations around the world are challenged by pressure to reduce costs, improve coordination and outcomes, provide more with less and be more patient centric. Building analytic competency can help these organizations harness big data to create actionable insights, set their future visions, improve outcomes and reduce time to value.
- Kellerman and Jones states that the potential of health IT to transform healthcare can and will only be fully realized by a comprehensive three-pronged effort by government, private vendors and providers.
- Joseph.M.Woodside has presented that in-efficient vendors can be identified having poor lifestyle decisions and must be compliance with preventive care programs. For individuals, incentives can be given, such as cash, gift cards, which are considered as big change in healthcare system [14].
- However, Bhattacherjee and Hikmet (2007) and Castrol (2007) proposed that "It is hard to implement information in small clinics and organization with high costs due to reduced efficiencies of scale" [15].
- In their paper, Dr. Kumar Sarvana et al proposed that Predictive Analysis can help healthcare providers accurately expect and respond to the patient needs. Big data Analytics in Hadoop's implementation provides systematic way in achieving better outcomes like availability and affordability of healthcare especially in rural areas [16].
- Alkhatib et al proposed some data analytical tools and techniques that have been used to improve healthcare analytics in many areas such as medical operations, reports, decision making, and prediction and prevention system. Moreover they had discussed weakness, disadvantages, problems and gaps of traditional healthcare data analytic technique in order to manage healthcare big data. They proposed a technique that promises to leverage large amount of healthcare data properly, since doctors and nurses would be able to determine diseases and risk easily like some certain types of cancer, diabetes and blood pressure, as well as provide needed treatment in right time.
- According to Borana Mukesh et al, the healthcare data should be properly analyzed so that we can deduce that in which group or gender, diseases attack the most [17].
- Raghupati and Kudyb proposed that the healthcare industry historically has generated large amounts of data, driven by record keeping, compliance and regulatory requirements and patient care. While most of the data is in pen-paper format, the current scenario is to digitalize the large amount of data [18].

3. RESEARCH WORK

A. TOOLS USED

Tools used in this process are Hadoop and RStudio.

Hadoop Ecosystem can be defined as a compendious collection of tools and technologies that can be implemented to provide effective and efficient Big Data solutions in a cost-effective manner [19]. Hadoop MapReduce and Hadoop Distributed File System (HDFS) and YARN (Yet Another Resource Negotiator) are the core components of the Hadoop Ecosystem. Along with these two, the Hadoop Ecosystem provides a collection of different elements to deal with Big Data. It includes Oozie(workflow management),Chukwa(monitoring),Zookeeper(manage --ment), Mahout(machine learning) ,Sqoop(SQL- to-hadoop) , hive(data warehousing) etc.

In the hadoop MapReduce model, a compute "job" is decomposed into smaller "tasks", which corresponds to separate Java Virtual Machine (JVM) processes in hadoop implementation. It primarily supports two operations: Map and Reduce. These operations execute in parallel on a set of slave nodes(as shown in Fig 3.5.2(a) and 3.5.2(b)). MapReduce works on master-slave approach in which master control entire activities like collection, segregation and delegation of data among slaves [20].

. Hadoop Distributed File System (HDFS), is a primary storage system used by hadoop applications, that provides a fault-tolerant, scalable and distributed

approach for storing and managing huge volumes of data[21]. It works on the "write once-read many times" approach that makes it capable of handling huge volumes of data with the least possibilities of error.

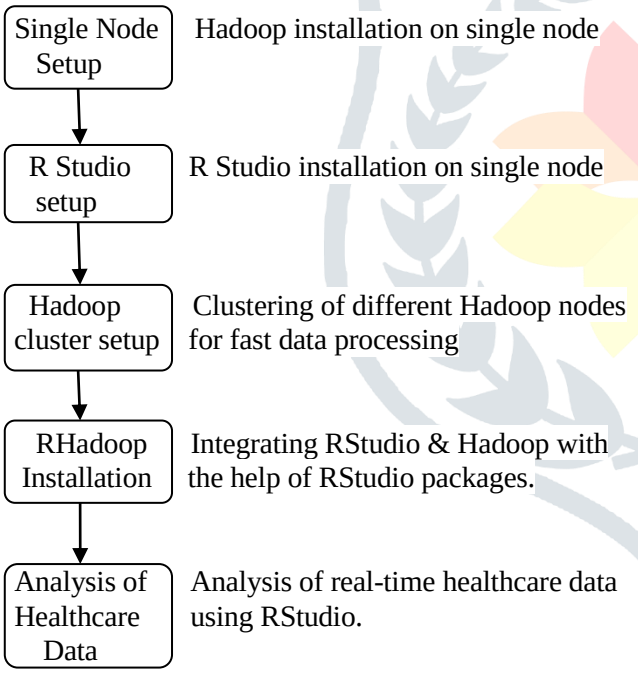
Hadoop YARN is a core hadoop service that supports major services like global resource management (Resource Manager and Node Manager) and pre-application management (Application Master)[22].

The Resource Manager is a master service that manages the Node Manager in each node of the cluster and also has a scheduler that allocates system resources to running applications.

R Studio is an IDE of R, a programming language for statistical computing and graphics. To some people, R is just the 18th letter of alphabet, but for the data analysts, it is a programming language that enables them to analyze and visualize data effectively and efficiently. RStudio includes a console, syntax highlighting editor that contain tools for plotting history, debugging and workspace management [23].

Analytical and visualization power of R
+ Big Data storage and processing of Hadoop
= Ideal solution for Big Data Analytics

B. METHODOLOGY



C. SINGLE NODE SETUP

1. Hadoop requires a working Java 1.5+ installation. We have used jdk 1.8.0.
2. For that, install java from oracle website.
3. For checking the existence of JVM
   command: /usr/local/jvm/
4. Next step is configuring SSH.
   command: ssh –keygen –t rsa –P " ".
5. The command will create a RSA key pair with an empty password. We have to unlock the key without user interaction.
6. Enabling SSH access to our local machine with this newly created key.
7. Hadoop installation:- Download Hadoop from Apache download mirrors and extract the contents of Hadoop package to a location of your choice. We picked /home/archi/hadoop-2.7.1
8. For extracting the Hadoop package
   command: sudo tar xzf hadoop-2.7.1.tar.gz
9. Update /.bashrc file:- For that , firstly move to the location of /.bashrc file and then edit the file by the editor of your choice[24].(shown in Fig 3.3.1)

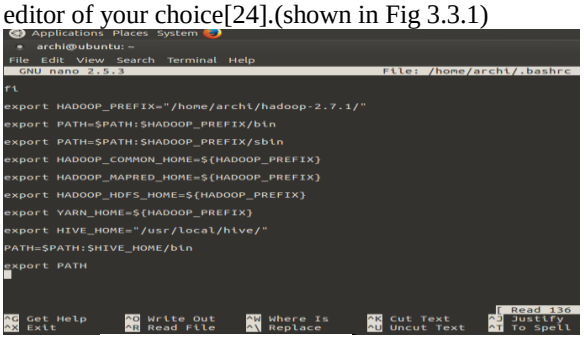


Fig 3.3.1 /.bashrc file

10. The only required environment variable we have to configure for Hadoop is JAVA_HOME. Open conf/hadoop-env.sh in the editor of your choice
11. export JAVA_HOME=/usr/lib/jvm/java-1.8.

15 Start your single node cluster by typing start-all.sh command on your command line terminal

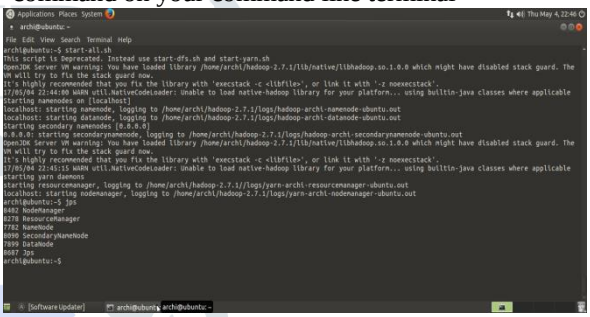


Fig 3.3.2 Starting hadoop on single node

D. R Studio SETUP

- For RStudio installation, we have to perform two steps:
- Firstly install R from CRAN mirror.(latest version of R for linux is 3.2.3)
- Install RStudio Desktop from official website of RStudio(www.rstudio.org).
- After installation, R Studio will appear same as shown in Fig 3.4.1.

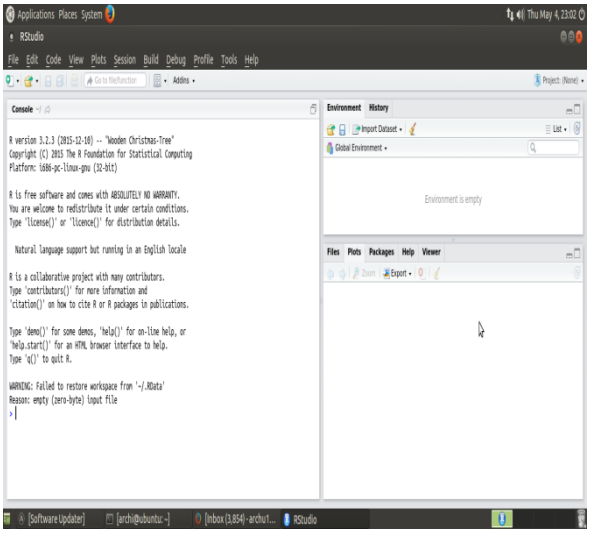


Fig 3.4.1 RStudio installation on single node

E. HADOOP CLUSTER SETUP

- Firstly, assure that Hadoop is installed on all the nodes. You can check it by typing hadoop –version on your command line.
- For connectivity, we have to point out that both machines must be able to reach each other over the network. The easiest is to put both machines in the same network with regard to hardware and software configuration, for example connect both machines via a single hub or switch and configure the network interfaces to use a common network such as 192.168.0.x/24. Update /etc/hosts file by the editor of your choice.
- Next step is SSH access. So , for connecting from master to master type ssh master and for master to slave type ssh slave on command line of master node.
- Typically, one machine in the cluster is designated as NAME NODE and another machines as DATA NODE.
- The primary NameNode and the JobTracker will always be the machines on which you run the bin/start-dfs.sh and bin/start-mapred.sh scripts.
- On master, update the conf/masters file by typing master in it.
- On master, update the conf/slave file. We have to include one master and all the slaves in use.
- We must change the configuration files conf/core-site.xml, conf/mapred-site.xml and conf/hdfs-site.xml on all the machines.
- Run the command start-all.sh on the master node.
- We can also view information of Data Nodes by typing localhost:50070 in our web browser(shown in Fig 3.5.1(a) and 3.5.1(b))

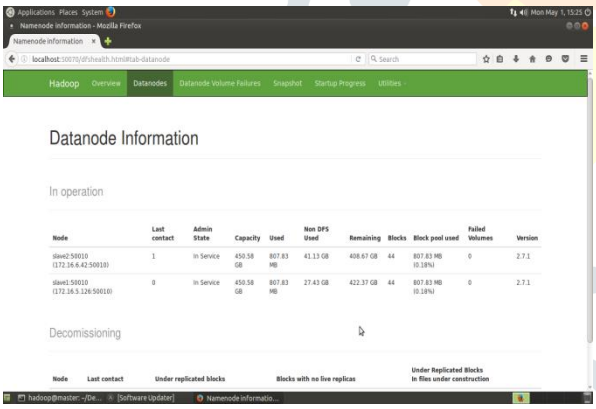


Fig 3.5.1(a) Data nodes and their information

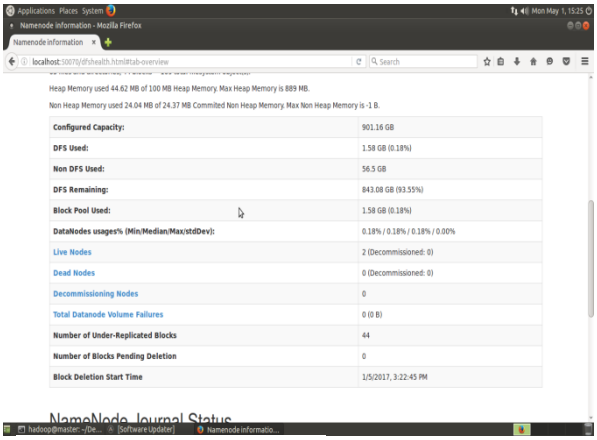


Fig 3.5.1(b)

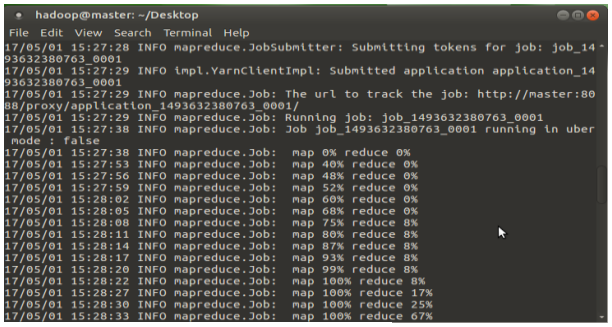


Fig 3.5.2(a) Map Reduce job

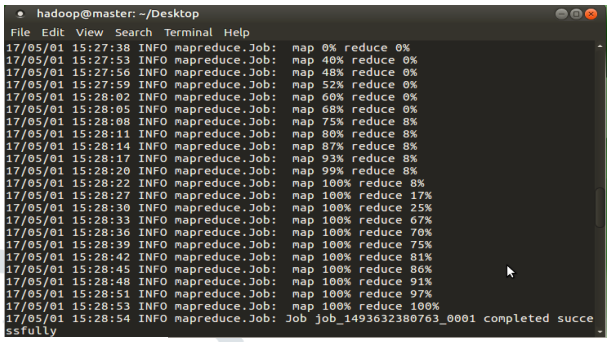


Fig 3.5.2(b)

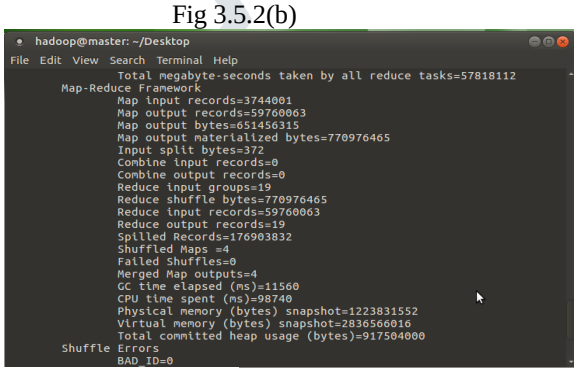


Fig 3.5.3 CPU time spent

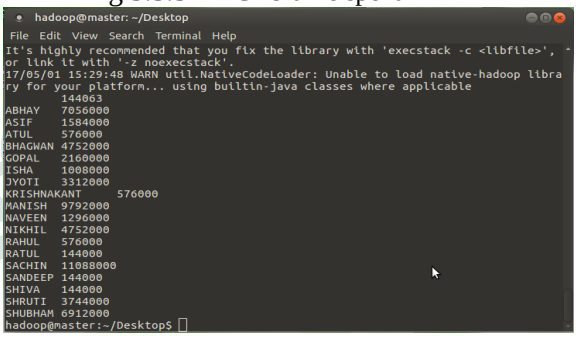


Fig 3.5.4 Running Word Count program

E. RHADOOP INSTALLATION-INTEGRATING HADOOP AND RStudio

RHadoop is a collection of 3 packages for providing large data operations wthin R environment.These p packages are :

1. Rhdfs 2.Rmr 3.Rhbase

Rhdfs:It is a package of R that provides the basic connectivity to the HDFS. R programmers can browse,read,write and modify files stored in HDFS from R.

Rmr: It allows developers to perform Ststistical Analysis in R via Hadoop's MapReduce functionality on a Hadoop cluster

Rhbase: is an interface for operating the Hadoop's HBase data source, stored at the distributed network via a thrift server[25].

Other packages includes rJava, RJSonio, itertools, digest, Rcpp, httr, functional, devtools, plyr, reshape2.

For using Rhdfs, we have to set configuration paths[26](as shown in Fig 3.6.1)

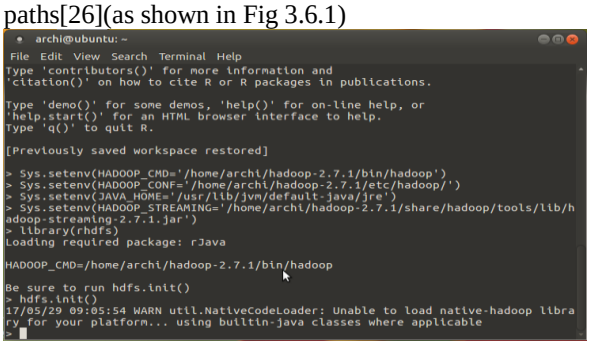


Fig 3.6.1 Setting configuration paths

After proper installation of Rhdfs, we can directly access HDFS files from RStudio.

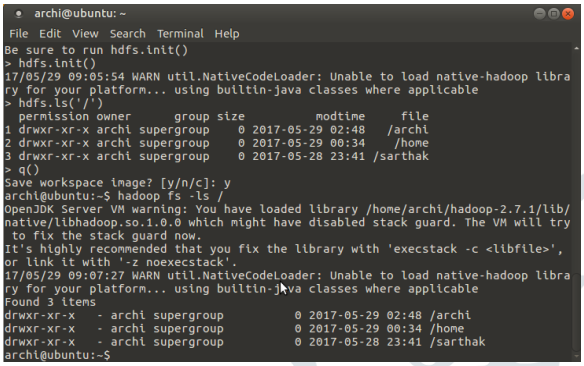


Fig 3.6.2 Accessing HDFS files from RStudio

F. ANALYSIS OF HEALTHCARE DATA

Analysis of Healthcare data using RStudio

1) Datasets needs to be imported first.
2) In our research work, we have taken four different tables of Doctor, patients, Disease, Medicines.
3) We can now apply various commands for analyzing different things.(as shown in Fig 3.7.1)

For Ex: view (Disease)

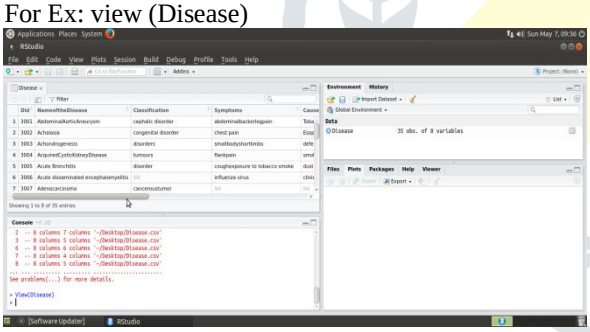


Fig 3.7.1 Disease table

For viewing the summary of column Effective% of Medicine table.(shown in Fig 3.7.2).

Command: summary (Medicine$Effective)

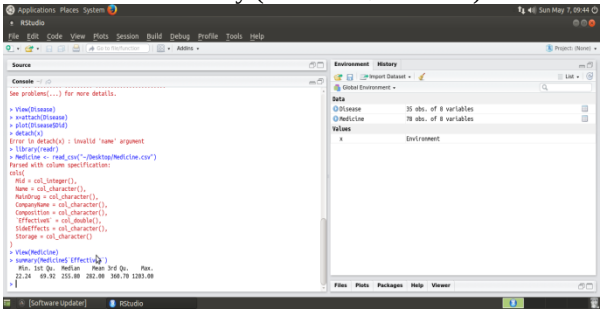


Fig 3.7.2 Effective % of Medicine table

For plotting the columns of table Disease(shown in Fig 3.7.3)

Command: plot (Disease$Id)

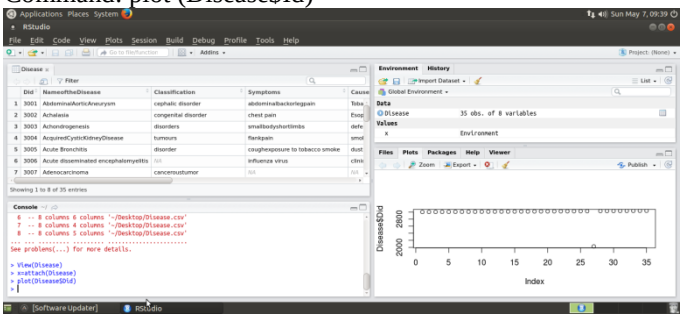


Fig 3.7.3 Disease table column

For viewing the histogram (shown in Fig 3.7.4)

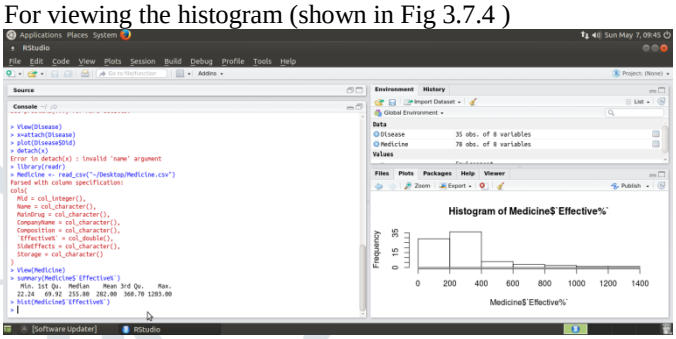


Fig 3.7.4 Histogram

With the help of Histogram (as shown in Fig 3.7.4), we can visualize our data pictorially and can analyze. data more effectively. In Fig 3.7.4, we have drawn a histogram showing the effective % of Medicine table. Similarly, histogram of various other data can also be drawn. Therefore, with the help of R Studio, data can be analyzed pictorially and graphically. Other commands of RStudio are available on the official website of RStudio (www.rstudio.org).

4. CONCLUSIONS AND FUTURE SCOPE

At the end of the paper, we can conclude that we have entered an era of Big Data, era where large volumes of healthcare data are available and effective and efficient analysis of that data can help us in reducing waste and inefficiency in the healthcare areas like clinical operations, genomic analytics, patient profile analytics etc.

In our research work, real-time healthcare data is processed in clustered environment along with analysis using RStudio, integrated with hadoop so that larger sets of data can be processed. Histograms of specific data sets were plotted to visualize data in pictorial manner. With our framework, research professionals and analysts from healthcare organizations can directly analyze huge volume of data in less time and without loss of data, resulting in analytics-based care, instead of theory-based care. This model will provide a great help in rural areas where doctors are still using pen-paper based treatment.

REFERENCES

[1] Cottle M, Hover W, Kohn M, Trelster W N, Institute for Health Technology transformation, New York,2013

[2] Cortada W J, Gordon D and Lenihan B," The value of analytics in healthcare", IBM Global Business Services Executive Report, 2011.

[3] Sun J and .Reddy K C, "Big data analytics for healthcare", IBM.

[4] Xindong W, Xingquan Z, Gong-Quing W, Ding W, "Data mining with big data", 2012.

[5] Mukherjee S and Shaw R, "Big data-concepts, applications, challenges and future scope", International Journal of Advanced Research in Computer and Communication Engineering, Vol.5,Issue 2,Kolkata,India,2016.
[6] Samuel J S, RVP K, Sasidhar K and Bharathi R C, "A survey of big data and its research", APRN Journal of Engineering and Applied Sciences, Chennai ,India, Vol. 10, No 8,2015, pp3343-3347.
[7] Ketal P, Kulennavar N, "A survey on big data analytics in healthcare", International Journal of Computer Science and Information Technology, 2014, Vol.5 (4), pp5865-5868.
[8] Weng L, Alexander A C, "Big data in medical applications and healthcare", 2015.
[9] Raghupati W and Raghupati V, "Big data analytics in healthcare: promise and potential", 2014, pp1-10.
[10] Ronge L, Mantzana C and Wilson V,"Healthcare information system research-revelation and visions", European Journal of IS, Vol. 16, pp669-671.
[11] "Big data: the next frontier for innovation, competition and productivity", McKinsey Global Institute, McKinsey & Company, 2011.
[12] Bhosale S H, Gadekar P, "A review paper on Big data and Hadoop", 2014.
[13] Shvacko K, Kuang H, Radia S, Sunnyvale C R," The Hadoop Distributed File System", IEEE, California, USA, 2010.
[14] Woodside M J, "virtual health management", 11th International Conference on Information Technology: New Generation 978-1-4799-3187-3/14, 2014.
[15] Alkhatib A M , Khoei T A, Ghapanchi H A, "Analysis of research in healthcare data analytics", Australasian Conference on Information System, Sydney, 2015.
[16] Kumar S, Eswari T, Sampath P, S Lavanya, "Predictive methodology for diabetic data analysis in big data",2nd International Symposium on Big Data and Cloud Computing(ISBCC'15), India, 2015,pp203-208
[17] Borana M, Giri M , Kamble S , Deshpande K, Edake S, "Healthcare analysis using Hadoop" ,International Research Journal of Engineering and Technology (IRJET) ,Maharashtra, India, 2015,pp583-586.
[18] Raghupati W and Kudyb S , "data mining in healthcare improving efficiency and productivity", Healthcare Informatics ,India,2010, pp2231-2234.
[19] Dhavapriya M and Yasodha N,"Big data analytics: challenges and solutions using Hadoop, MapReduce and Big Table", International Journal of Computer Science and Technology (IJCST), 2016, Vol.4, pp5-14.
[20] Dean J and Ghemawat S, "MapReduce: A flexible data processing tool", CACM, 2010, pp72-77.
[21] Borkar D S, Surtakar S C, "A Review Paper on the Hadoop Distributed File System", International Journal for Research in Science and Engineering, India,Vol.1, pp211-216.
[22] Vavilapalli V, Murthy A, Douglas C, Agarwal S, Kovar M, Evans R, Graves T, Lowe J, "Apache Hadoop YARN: Yet Another Resource Negotiator", SoCC'13, Santa Clara , California, USA. 1-3 Oct.2013,
[23] https://www.rstudio.com/products/RStudio/
[24]http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/
[25]https://www.dezyre.com/article/r-hadoop-a-perfect-match-for-big-data/292
[26] https://acadgild.com/blog/integration-r-hadoop/