

Health Status Prediction

Submitted in partial fulfilment of the Requirements for the Degree of

Master of Computer Applications

A PROJECT REPORT

Submitted by

ARCHANA VARSHNEY

(University Roll No 1900290140009)

NIRMIT BANSAL

(University Roll No 1900290140022)

Batch:2019-2022

Under the Supervision of

VIDUSHI MISHRA

(Assistant Professor)



DEPARTMENT OF COMPUTER APPLICATIONS

DR.APJ ABDUL KALAM TECHNICAL UNIVERSITY

LUCKNOW

(Formerly Uttar Pradesh Technical University, Lucknow)

DECLARATION

I hereby declare that the work presented in this report entitled “health status prediction”, was carried out by me. I have not submitted the matter embodied in this report for the award of any other degree or diploma of any other University or Institute. I have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programs, experiments, results, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources. I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable.

Name :

Roll. No. :

Branch :

CERTIFICATE

Certified that **Archana varshney (University Roll No 1900290140009), Nimit bansal (University Roll No 1900290140022)**, have carried out the project work having “Hospital Management System” for Master of Computer Applications from Dr. A.P.J. Abdul Kalam Technical University (AKTU) (formerly UPTU), Technical University, Lucknow under my supervision. The project report embodies original work, and studies are carried out by the student himself/herself and the contents of the project report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Date:

**Archana varshney
University Roll No. 1900290140009
Nimit bansal
University Roll No. 1900290140022**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:

**Vidushi Mishra
(Assistant Professor)
Department of Computer Applications
KIET Group of Institutions, Ghaziabad**

Signature of External Examiner

Signature of Internal Examiner

ABSTRACT

In healthcare management, a large volume of multi-structured patient data is generated from the clinical reports, doctor's notes, and wearable body sensors. The analysis of healthcare parameters and the prediction of the subsequent future health conditions are still in the informative stage.

A cloud-enabled big data analytic platform is the best way to analyze the structured and unstructured data generated from healthcare management systems. In this paper, a probabilistic data collection mechanism is designed and the correlation analysis of those collected data is performed.

Finally, a stochastic prediction model is designed to foresee the future health condition of the most correlated patients based on their current health status. Performance evaluation of the proposed protocols is realized through extensive simulations in the cloud environment, which gives about 98% accuracy of prediction, and maintains 90% of CPU and bandwidth utilization to reduce the analysis time.

ACKNOWLEDGEMENT

Success in life is never attained single handed. My deepest gratitude goes to my thesis supervisor, **Vidushi Mishra** for his guidance, help and encouragement throughout my research work. Their enlightening ideas, comments, and suggestions.

Words are not enough to express my gratitude to **Dr. Ajay Kumar Shrivastava, Professor and Head, Department of Computer Applications**, for his insightful comments and administrative help at various occasions.

Fortunately, I have many understanding friends, who have helped me a lot on many critical conditions.

Finally, my sincere thanks go to my family members and all those who have directly and indirectly provided me moral support and other kind of help. Without their support, completion of this work would not have been possible in time. They keep my life filled with enjoyment and happiness.

Archana varshney

Nirmit Bansal

Table of Contents

DECLARATION	1
CERTIFICATE	2
ABSTRACT.....	3
ACKNOWLEDGEMENT	4
CHAPTER 1	8
INTRODUCTION	8
1.1 PROJECT DESCRIPTION.....	8
1.2 PROJECT SCOPE.....	10
1.2 TECHNOLOGY USED.....	1
1.3.1 Big Data	1
1.3.2 Hadoop.....	1
1.3.3 JAVA	1
1.4 REQUIREMENTS.....	2
FUNCTIONAL REQUIREMENT.....	2
Check Out of SRS:.....	3
Report Generation of SRS:	3
Non Functional Requirements	4
Security:	4
Performance:	4
Maintainability:	5
Reliability:.....	5
CHAPTER 2	6
LITERATURE REVIEW	6
CHAPTER 3	8
FEASIBILITY STUDY.....	8
Steps Involved in Feasibility Analysis.....	8
3.1 TECHNICAL FEASIBILITY	10
3.2 OPERATIONAL FEASIBILITY.....	11
3.3 BEHAVIORAL FEASIBILITY	12
3.4 Economic Feasibility	13
3.5 TIME FEASIBILITY	13
3.6 FINANCIAL FESIBILITY	14
3.7 Schedule Feasibility	15

CHAPTER 4	16
4.1 HARDWARE REQUIREMENTS	16
4.2 SOFTWARE REQUIREMENTS	16
. ORACLE VM VIRTUALBOX	16
4.3 DATA REQUIREMENTS	17
CHAPTER 5	18
DESIGN	18
5.1 0-LEVEL DATA FLOW DIAGRAM	18
5.2 1-LEVEL DATA FLOW DIAGRAM.....	19
CHAPTER 6	20
IMPLEMENTATION & WORKFLOW	20
6.1 LINUX MACHINE SETUP USING ORACLE VIRTUALBOX AND CENTOS.....	20
6.2 SETTING UP SINGLE NODE CLUSTER	20
In Prerequisites	21
Install Hadoop	22
Step 1: Click here to download the Java 8 Package. Save this file in your home directory.....	22
Step 2: Extract the Java Tar File.	22
Step 3: Download the Hadoop 2.7.3 Package.....	22
Step 4: Extract the Hadoop tar File.	22
Step 5: Add the Hadoop and Java paths in the bash file (.bashrc).	22
Step 6: Edit the Hadoop Configuration files.	24
Step 7: Open <i>core-site.xml</i> and edit the property mentioned below inside configuration tag:...	24
Step 8: Edit <i>hdfs-site.xml</i> and edit the property mentioned below inside configuration tag:	25
Step 9: Edit the <i>mapred-site.xml</i> file and edit the property mentioned below inside configuration tag:.....	26
Step 10: Edit <i>yarn-site.xml</i> and edit the property mentioned below inside configuration tag: ...	27
Step 11: Edit <i>hadoop-env.sh</i> and add the Java Path as mentioned below:	28
Step 12: Go to Hadoop home directory and format the NameNode.	28
Step 13: Once the NameNode is formatted, go to <i>hadoop-2.7.3/sbin</i> directory and start all the daemons.....	29
Start NameNode:	29
Start DataNode:	30
Start ResourceManager:	30
Start NodeManager:	30

Start JobHistoryServer:	31
Step 14: To check that all the Hadoop services are up and running, run the below command...	31
Step 15: Now open the Mozilla browser and go to localhost:50070/dfshealth.html to check the NameNode interface.....	32
CHAPTER 7	33
CONCLUSION AND FUTURE SCOPE	33
REFERENCES	38

CHAPTER 1

INTRODUCTION

1.1 PROJECT DESCRIPTION

Wireless Sensor Networks (WSNs) and mobile networks allow in-hospital and outdoor patients monitoring through Internet of Things (IoT) [1] in which patients are equipped with different smart devices such as in-plant pacemaker, Electrocardiogram (ECG), Electromyography (EMG), Electroencephalography (EEG) and motion sensors, etc. These wearable devices collect health related data such as body temperature, blood pressure and heart rate, which can be applied in physical fitness tracking and medical treatment. Big data in healthcare [2] is an analytic environment to handle the massive volume of structured and unstructured patient data. According to the analysts, the healthcare data volume of USA healthcare system has reached to 150 exabytes in 2011 [3] and has increased to zettabyte scale [4] in the current time. Similarly, the California-based health network Kaiser Permanente has 9 million members and the data [3] collected from Electronic Health Records (EHRs) including doctor notes, clinical reports and pathological images range from 26.5 to 44 petabytes.

The health data are attributed as big data, which is defined by 5Vs in terms of Volume, Velocity, Variety, Value, and Veracity. The collected patient data are of peta or zeta bytes, which describe the volume. The velocity is expressed in terms of data arrival rate from the patients. Variety explains the diversified data sets with respect to the structured, semi-structured and unstructured data sets such as clinical reports, EHRs, and radiological images and veracity explains the truthfulness of the data sets with respect to data availability and authenticity. The collected data are transformed into meaningful insights, which explain the value in 5Vs.

Physiological data of patients are the primary and vital entities in healthcare big data analytic. Hence, valid raw data must be collected with an efficient manner in a medical environment. In advanced healthcare systems, the patient data are collected [5] through wearable devices equipped with different types of sensors. Recently, the advancement in mobile devices [6] such as multi-sensor equipped smart phones are also used as the data collection devices. Hence, colossal amounts of patient data are generated within a hospital network, which needs to be stored and analyzed efficiently. Therefore, a cloud computing [7] enabled distributed storage and processing environment is essential to store and process the healthcare data, which can be accessed anywhere and anytime.

Now-a-days various data intensive applications are emerged, which need some efficient analytic models. Many stochastic approaches [8] are considered by different authors in the recent past for healthcare parameter analysis. Moreover, the similarity [9] between health parameters of a patient is considered by the physicians for better decisions. Big data analytic is applied in healthcare [10] to identify the clusters of patients, diseases and future predictions with the help of various machine learning tools [11]. In a learning healthcare system [12],

data are analyzed and used as insights continuously for patient care. During this process, the patient data are combined with the clinical reports for better suggestions and decisions.

So far limited analyses have accomplished among the patients taking different numbers of health parameters of same or different departments. Even, the existing models cannot support both analysis and processing for the large volume of multi-structured healthcare data. Recently, the high performance of cloud platform provides a scalable and distributable parallel processing framework, i.e. MapReduce [13] for healthcare data processing. MapReduce has the capability to process the large volume of data in parallel on a cloud. The major benefits of MapReduce framework are the scalability and fault-tolerance during massive data processing on a large cloud. Hence, a hybrid model of stochastic and parallel processing framework is planned in a medical environment to process and analyze the huge volume of healthcare data. In our work, MapReduce parallel processing framework [14] is used as a backbone for healthcare big data analysis. Further, the proposed work is extended to a prognosis model for future health condition prediction of the patients.

1.2 PROJECT SCOPE

When patients arrive they make an appointment at the reception to consult a Doctor. These are being recorded in a file. Then again the patients diagnosed symptoms related disease details, ward details and other necessary details are being recorded and those files are being stored in special locations. Calculation of bills and inventory are done manually. As the current system is a file based one, management of the hospital has to put much effort on securing the files. They can be easily damaged by fire, insects and natural disasters.

Also could be misplaced by losing data and information. Limited storage space of the files is another issue that they currently face when the management is manually done. There occurs an issue with the organization of data information and schedules and running the process methodically which leads to the manual system malfunctioning. Hospital Management System . If we want to check a previous record of a patient or other detail. Management will be in a great problem. It's a tough and time taking process to search for a record in a file. Keeping files takes much time and waste much precious man hours. The tendency of making mistakes is high when functioning manually. It is hard to relay on the accuracy of calculation obvious for problems to arise. We plan to overcome the above mentioned problems through a standalone application, to manage the major functions of the Hospital System. The hospital management system we are going to implement will be covering all basic processes done in the hospital. It would handle Employee and Salary management, Patient, Theatre and ward Management, Laboratory management, Transport Management, Pharmacy Management, OPD management and emergency management. In OPD unit, with the OPD and Consultation Management system, the manual doctors channeling details entering process has automated. So the staff does not need to spend time on writing appointment records and updating them in files. And the number issuing process becomes easier and efficient. And keeping the track of patients and medical prescription details allow them to review the details whenever needed. Implementing the Employee & Salary Management system we record Attendance, shifting of employees, their holidays and consulting doctors' schedules.

1.2 TECHNOLOGY USED

1.3.1 Big Data

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size

1.3.2 Hadoop

Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

1.3.3 JAVA

Java is a programming language and computing platform first released by Sun Microsystems in 1995. There are lots of applications and websites that will not work unless you have Java installed, and more are created every day. Java is fast, secure, and reliable. From laptops to data centers, game consoles to scientific supercomputers, cell phones to the Internet, Java is everywhere!

Java is the technology of choice for building applications using managed code that can execute on mobile devices.

What made Java be the technology of choice for mobile development for the Android platform? The Java Programming Language emerged in the mid-1990s; it was created by James Gosling of Sun Microsystems. Incidentally, Sun Microsystems was since bought by Oracle. Java has been widely popular the world over, primarily because of a vast array of features it provides. Java's promise of "Write once and run anywhere" was one of the major factors for the success of Java over the past few decades.

Java even made inroads into embedded processors technology as well; the Java Mobile Edition was built for creating applications that can run on mobile devices. All these, added to Java's meteoric rise, were the prime factors that attributed to the decision of adopting Java as the primary development language for building applications that run on Android. Java programs are secure because they run within a sandbox environment. Programs written in Java are compiled into intermediate code known as bytecode. This bytecode is then executed inside the context of the Java Virtual Machine.

1.4 REQUIREMENTS

FUNCTIONAL REQUIREMENT

Registration Process of SRS (Software Requirements Specification)

- Adding Patients: The Hospital Management enables the staff in the front desk to include new patients to the system.
- Assigning an ID to the patients: The HMS enables the staff in the front desk to provide a unique ID for each patient and then add them to the record sheet of the patient. The patients can utilize the ID throughout their hospital stay.

Check Out of SRS:

- **Deleting Patient ID:** The staff in the administration section of the ward can delete the patient ID from the system when the patient's checkout from the hospital.
- **Adding to beds available list:** The Staff in the administration section of the ward can put the bed empty in the list of beds-available.

Report Generation of SRS:

- **Information of the Patient:** The Hospital Management System generates a report on every patient regarding various information like patients name, Phone number, bed number, the doctor's name whom its assigns, ward name, and more.
- **Availability of the Bed:** The Hospital Management system also helps in generating reports on the availability of the bed regarding the information like bed number unoccupied or occupied, ward name, and more.

Database of SRS:

- **Mandatory Patient Information:** Every patient has some necessary data like phone number, their first and last name, personal health number, postal code, country, address, city, 'patient's ID number, etc.
- **Updating information of the Patient:** The hospital management system enables users to update the information of the patient as described in the mandatory information included.

Non Functional Requirements

There are a lot of software requirements specifications included in the non-functional requirements of the Hospital Management System, which contains various process, namely Security, Performance, Maintainability, and Reliability.

Security:

- Patient Identification: The system needs the patient to recognize herself or himself using the phone.
- Logon ID: Any users who make use of the system need to hold a Logon ID and password.
- Modifications: Any modifications like insert, delete, update, etc. for the database can be synchronized quickly and executed only by the ward administrator.
- Front Desk Staff Rights: The staff in the front desk can view any data in the Hospital Management system, add new patients record to the HMS but they don't have any rights alter any data in it.
- Administrator rights: The administrator can view as well as alter any information in the Hospital Management System.

Performance:

- Response Time: The system provides acknowledgment in just one second once the 'patient's information is checked.
- Capacity: The system needs to support at least 1000 people at once.

- User-Interface: The user interface acknowledges within five seconds.
- Conformity: The system needs to ensure that the guidelines of the Microsoft accessibilities are followed.

Maintainability:

- Back-Up: The system offers the efficiency for data back up.
- Errors: The system will track every mistake as well as keep a log of it.

Reliability:

- Availability: The system is available all the time.

CHAPTER 2

LITERATURE REVIEW

is

The current study performs a systematic literature review (SLR) to synthesise prior research on the applicability of big data analytics (BDA) in healthcare. The SLR examines the outcomes of 41 studies, and presents them in a comprehensive framework. The findings from this study suggest that applications of BDA in healthcare can be observed from five perspectives, namely, health awareness among the general public, interactions among stakeholders in the healthcare ecosystem, hospital management practices, treatment of specific medical conditions, and technology in healthcare service delivery. This SLR recommends actionable future research agendas for scholars and valuable implications for theory and practice. Similar to EHR, an electronic medical record (EMR) stores the standard medical and clinical data gathered from the patients. EHRs, EMRs, personal health record (PHR), medical practice management software (MPM), and many other healthcare data components collectively have the potential to improve the quality, service efficiency, and costs of healthcare along with the reduction of medical errors. The big data in healthcare includes the healthcare payer-provider data (such as EMRs, pharmacy prescription, and insurance records) along with the genomics-driven experiments (such as genotyping, gene expression data) and other data acquired from the smart web of internet of things (IoT) (Fig. [1](#)).

The adoption of EHRs was slow at the beginning of the 21st century however it has grown substantially after 2009 [7, 8]. The management and usage of such healthcare data has been increasingly dependent on information technology. The development and usage of wellness monitoring devices and related software that can generate alerts and share the health related data of a patient with the respective health care providers has gained momentum, especially in establishing a real-time biomedical and health monitoring system. These devices are generating a huge amount of data that can be analyzed to provide real-time clinical or medical care [9]. The use of big data Healthcare industry has not been quick enough to adapt to the big data movement compared to other industries. Therefore, big data usage in the healthcare sector is still in its infancy. For example, healthcare and biomedical big data have not yet converged to enhance healthcare data with molecular pathology. Such convergence can help unravel various mechanisms of action or other aspects of predictive biology. Therefore, to assess an individual's health status, biomolecular and clinical datasets need to be married. One such source of clinical data in healthcare is 'internet of things' (IoT).

In fact, IoT is another big player implemented in a number of other industries including healthcare. Until recently, the objects of common use such as cars, watches, refrigerators and health-monitoring devices, did not usually produce or handle data and lacked internet connectivity. However, furnishing such objects with computer chips and sensors that enable data collection and transmission over internet has opened new avenues. The device technologies such as Radio Frequency IDentification (RFID) tags and readers, and Near Field Communication (NFC) devices, that can not only gather information but interact physically, are being increasingly used as the information and communication systems [3]. This enables objects with RFID or NFC to communicate and function as a web of smart things.

CHAPTER 3

FEASIBILITY STUDY

Feasibility Study can be considered as preliminary investigation that helps the management to take decision about whether study of system should be feasible for development or not.

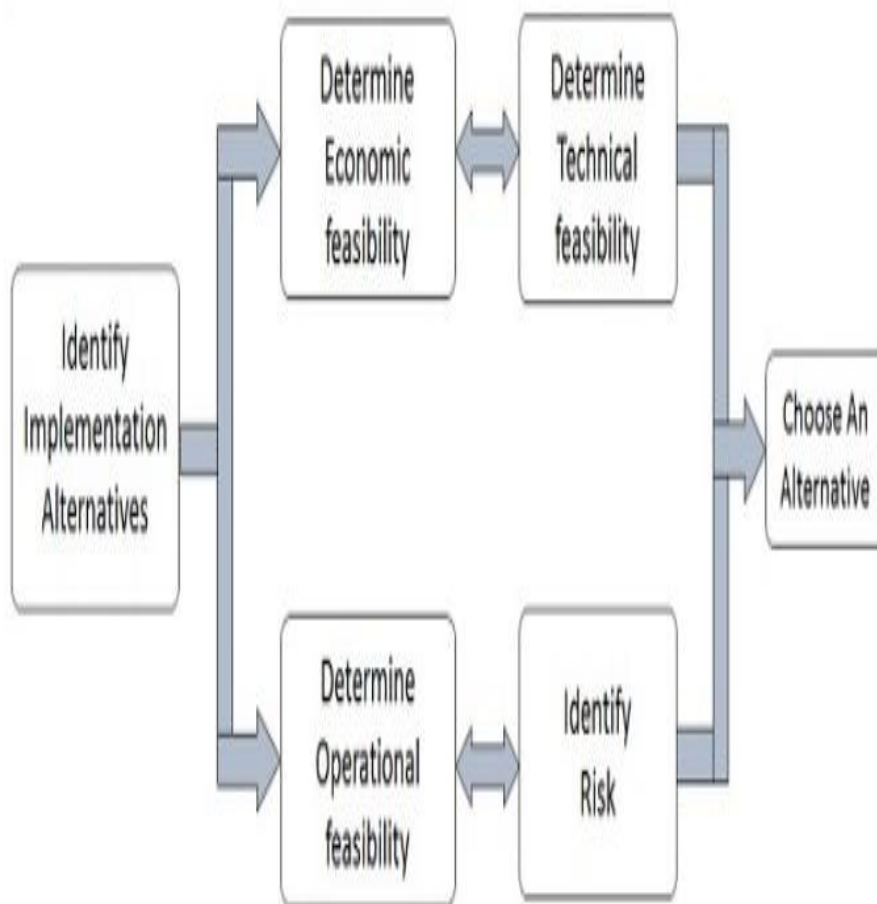
- It identifies the possibility of improving an existing system, developing a new system, and produce refined estimates for further development of system.
- It is used to obtain the outline of the problem and decide whether feasible or appropriate solution exists or not.
- The main objective of a feasibility study is to acquire problem scope instead of solving the problem.
- The output of a feasibility study is a formal system proposal act as decision document which includes the complete nature and scope of the proposed system.

Steps Involved in Feasibility Analysis

The following steps are to be followed while performing feasibility analysis –

- Form a project team and appoint a project leader.
- Develop system flowcharts.
- Identify the deficiencies of current system and set goals.

- Enumerate the alternative solution or potential candidate system to meet goals.
- Determine the feasibility of each alternative such as technical feasibility, operational feasibility, etc.
- Weight the performance and cost effectiveness of each candidate system.
- Rank the other alternatives and select the best candidate system.
- Prepare a system proposal of final project directive to management for approval.



3.1 TECHNICAL FEASIBILITY

This assessment is based on an outline design of system requirements, to determine whether the company has the technical expertise to handle completion of the project when writing a feasibility report, the following should be taken to consideration:

- A brief description of the business to assess more possible factors which could affect the study
- The part of the business being examined
- The human and economic factor
- The possible solutions to the problem

At this level, the concern is whether the proposal is both technically and legally feasible (assuming moderate cost).

The technical feasibility assessment is focused on gaining an understanding of the present technical resources of the organization and their applicability to the expected needs of the proposed system. It is an evaluation of the hardware and software and how it meets the need of the proposed system.

3.2 OPERATIONAL FEASIBILITY

Operational feasibility is the measure of how well a proposed system solves the problems, and takes advantage of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development.

The operational feasibility assessment focuses on the degree to which the proposed development project fits in with the existing business environment and objectives with regard to development schedule, delivery date, corporate culture and existing business processes.

To ensure success, desired operational outcomes must be imparted during design and development. These include such design-dependent parameters as reliability, maintainability, supportability, usability, predictability, disposability, sustainability, affordability and others. These parameters are required to be considered at the early stages of design if desired operational behaviors' are to be realized.

A system design and development requires appropriate and timely application of engineering and management efforts to meet the previously mentioned parameters. A system may serve its intended purpose most effectively when its technical and operating characteristics are engineered into the design.

Therefore, operational feasibility is a critical aspect of systems engineering that needs to be an integral part of the early design phases.

3.3 BEHAVIORAL FEASIBILITY

It evaluates and estimates the user attitude or behavior towards the development of new system.

It helps in determining if the system requires special effort to educate, retrain, transfer, and changes in employee's job status on new ways of conducting business.

People are inherently resistant to change, and computers have been known to facilitate change. An estimate should be made of how strong a reaction the user staff is likely to have toward the development of a computerized system. [t is common knowledge that computer installations have something to do with turnover, transfers, retraining, and changes in employee job status. Therefore, it is understandable that the introduction of a candidate system requires special effort to educate, sell, and train the staff on new ways of conducting business.

3.4 Economic Feasibility

- It is evaluating the effectiveness of candidate system by using cost/benefit analysis method.
- It demonstrates the net benefit from the candidate system in terms of benefits and costs to the organization.
- The main aim of Economic Feasibility Analysis (EFS) is to estimate the economic requirements of candidate system before investments funds are committed to proposal.
- It prefers the alternative which will maximize the net worth of organization by earliest and highest return of funds along with lowest level of risk involved in developing the candidate system.

3.5 TIME FEASIBILITY

A time feasibility study will take into account the period in which the project is going to take up to its completion. A project will fail if it takes too long to be completed before it is useful. Typically this means estimating how long the system will take to develop, and if it can be completed in a given time period using some methods like payback period. Time feasibility is a measure of how reasonable the project timetable is. Given our technical expertise, are the project deadlines reasonable? Some projects are initiated with specific deadlines. It is necessary to determine whether the deadlines are mandatory or desirable.

3.6 FINANCIAL FESIBILITY

In case of a new project, financial viability can be judged on the following parameters:

- Total estimated cost of the project
- Financing of the project in terms of its capital structure, debt to equity ratio and promoter's share of total cost
- Existing investment by the promoter in any other business
- Projected cash flow and profitability The financial viability of a project should provide the following information:[12]
 - Full details of the assets to be financed and how liquid those assets are.
 - Rate of conversion to cash-liquidity (i.e., how easily the various assets can be converted to cash).
 - Project's funding potential and repayment terms.
 - Sensitivity in the repayments capability to the following factors:
 - Mild slowing of sales.
 - Acute reduction/slowing of sales.
 - Small increase in cost.
 - Large increase in cost.
 - Adverse economic condition.

3.7 Schedule Feasibility

It is defined as the probability of a project to be completed within its scheduled time limits, by a planned due date. If a project has a high probability to be completed on-time, then its schedule feasibility is appraised as high. In many cases a project will be unsuccessful if it takes longer than it was estimated: some external environmental conditions may change, hence a project can lose its benefits, expediency and profitability. If a work to be accomplished at a project does not fit the timeframes demanded by its customers, then a schedule is unfeasible (amount of work should be reduced or other schedule compression methods applied).

If the project managers want to see their projects completed before they can lose their utility, they (project managers) need to give proper attention to controlling their schedule feasibility: to calculate and continually reexamine whether it is possible to complete all amount and scope of work lying ahead, utilizing the given amount of resources, within required period of time. Schedule feasibility study includes use of the following matters:

- Project Estimation;
- Gantt and PERT charts;
- CPM (Critical Path Method);
- Change Management;

CHAPTER 4

REQUIREMENT SPECIFICATIONS

4.1 HARDWARE REQUIREMENTS

- . RAM: 8GB**
- . OPERATING SYSTEM: Linux(32bit or 64bit)**

4.2 SOFTWARE REQUIREMENTS

- . ORACLE VM VIRTUALBOX**
- . CentOS 7**
- . Apache NiFi version-1.9.0**
- . Apache Hadoop**
- . Apache hive**

. **Putty**

. **Microsoft Power BI**

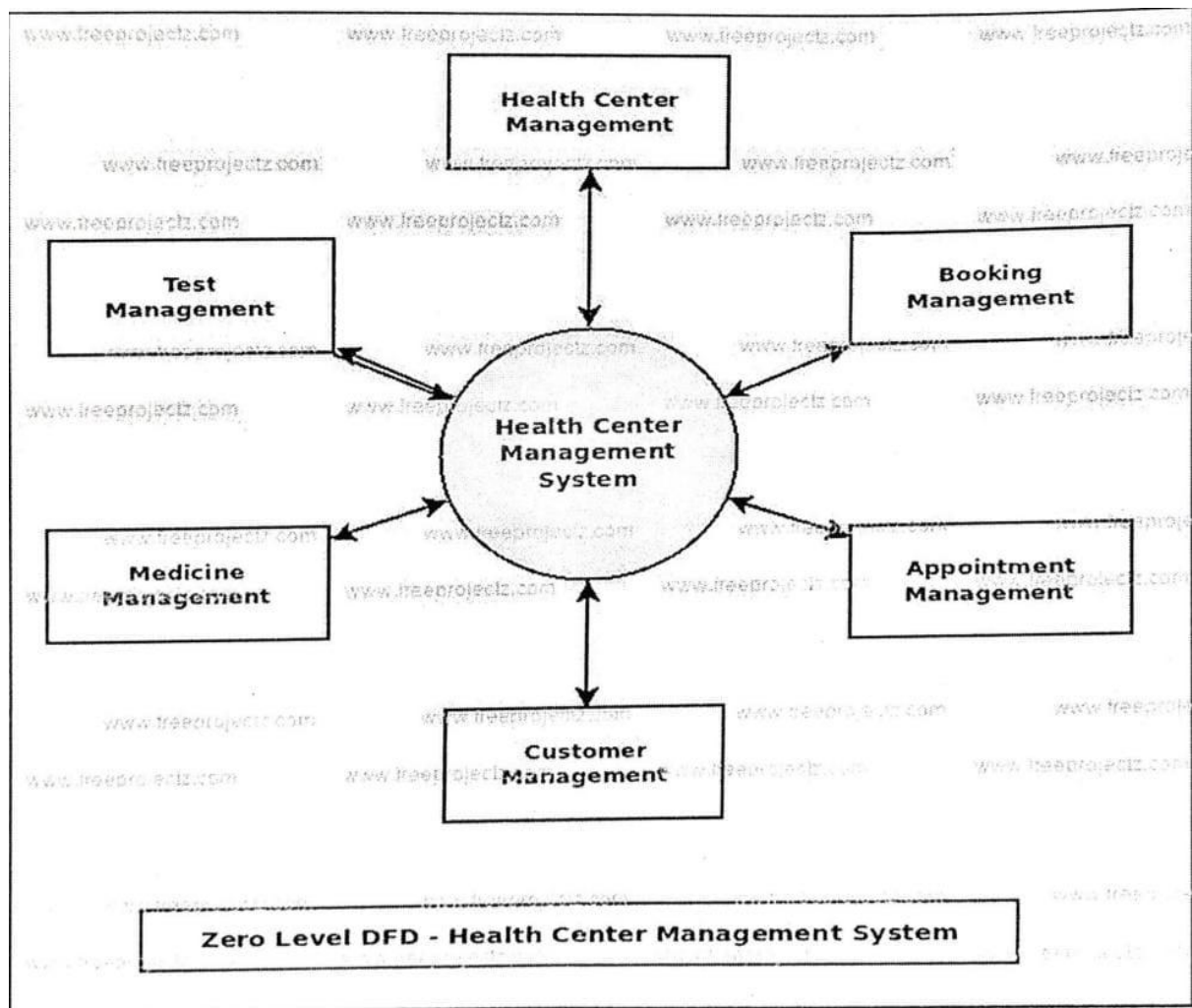
4.3 DATA REQUIREMENTS

For the analysis purpose we needed a large amount of data as we are using Hadoop which is a tool for analysing Big Data. We found twitter dataset on Kaggle website. This data set contains 13 columns. All the entries and Tweet's text in data set are related to Covid19 situation. Kaggle.com is a website that provides dataset for free for its users.

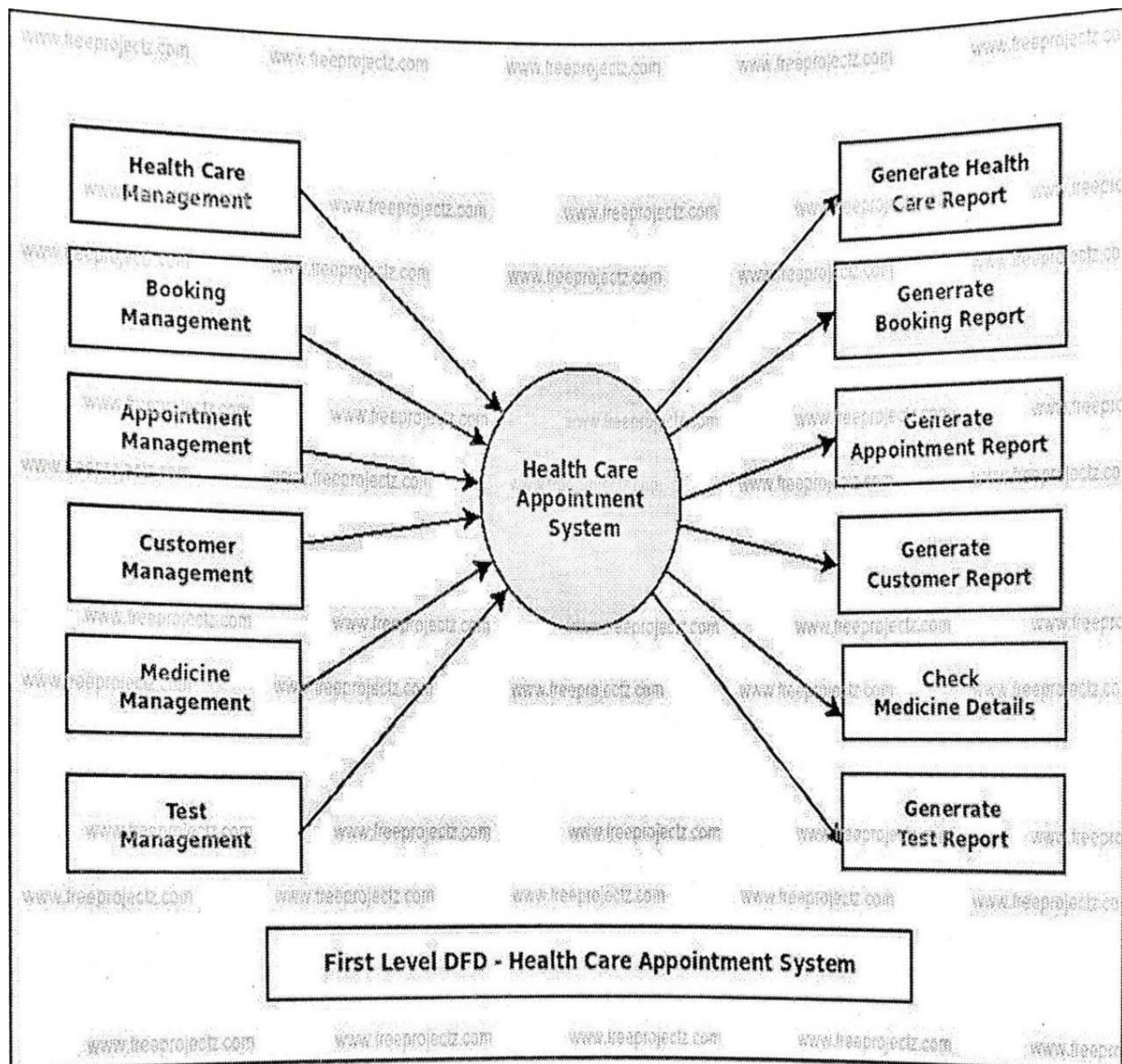
CHAPTER 5

DESIGN

5.1 0-LEVEL DATA FLOW DIAGRAM



5.2 1-LEVEL DATA FLOW DIAGRAM



CHAPTER 6

IMPLEMENTATION & WORKFLOW

6.1 LINUX MACHINE SETUP USING ORACLE VIRTUALBOX AND CENTOS

- . Download Oracle VM Virtualbox for windows and install.
- . Download CentOS 7
- . Setup Linux 64-bit Machine.

6.2 SETTING UP SINGLE NODE CLUSTER

From our previous blogs on [Hadoop Tutorial Series](#), you must have got a theoretical idea about Hadoop, HDFS and its architecture. But to get **Hadoop Certified**, you need good hands-on knowledge. I hope you would have liked our previous blog on [HDFS Architecture](#), now I will take you through the practical knowledge about Hadoop and HDFS. The first step forward is to install Hadoop.

There are two ways to install Hadoop, i.e. **Single node** and **Multi node**.

Single node cluster means only one DataNode running and setting up all the NameNode, DataNode, ResourceManager and NodeManager on a single machine. This is used for studying and testing purposes. For example, let us consider a sample data set inside a healthcare industry. So, for testing whether the Oozie jobs have scheduled all

the processes like collecting, aggregating, storing and processing the data in a proper sequence, we use single node cluster. It can easily and efficiently test the sequential workflow in a smaller environment as compared to large environments which contains terabytes of data distributed across hundreds of machines.

While in a **Multi node cluster**, there are more than one DataNode running and each DataNode is running on different machines. The multi node cluster is practically used in organizations for analyzing Big Data. Considering the above example, in real time when we deal with petabytes of data, it needs to be distributed across hundreds of machines to be processed. Thus, here we use multi node cluster.

In Prerequisites

- *VIRTUAL BOX*: it is used for installing the operating system on it.
- *OPERATING SYSTEM*: You can install Hadoop on Linux based operating systems. Ubuntu and CentOS are very commonly used. In this tutorial, we are using CentOS.
- *JAVA*: You need to install the Java 8 package on your system.
- *HADOOP*: You require Hadoop 2.7.3 package.

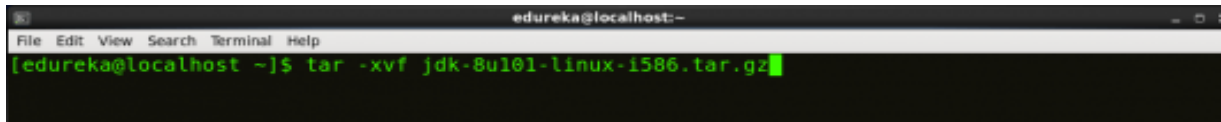
this blog, I will show you how to install Hadoop on a single node cluster.

Install Hadoop

Step 1: [Click here](#) to download the Java 8 Package. Save this file in your home directory.

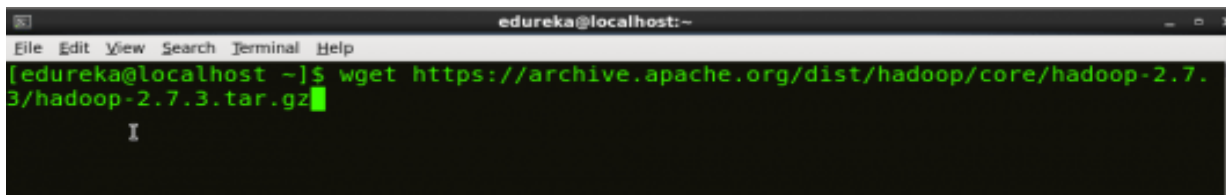
Step 2: Extract the Java Tar File.

Command: `tar -xvf jdk-8u101-linux-i586.tar.gz`

A terminal window titled 'edureka@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The command '[edureka@localhost ~]\$ tar -xvf jdk-8u101-linux-i586.tar.gz' is entered and executed, with a green cursor at the end of the line.

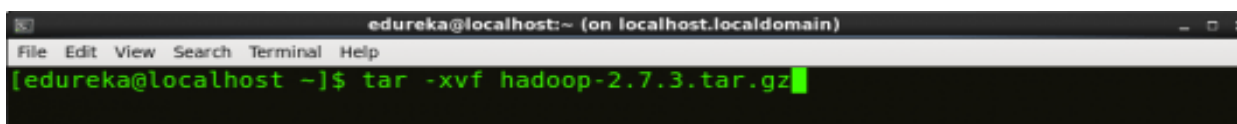
Step 3: Download the Hadoop 2.7.3 Package.

Command: `wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz`

A terminal window titled 'edureka@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The command '[edureka@localhost ~]\$ wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz' is entered and executed, with a green cursor at the end of the line.

Step 4: Extract the Hadoop tar File.

Command: `tar -xvf hadoop-2.7.3.tar.gz`

A terminal window titled 'edureka@localhost:~ (on localhost.localdomain)' with a menu bar (File, Edit, View, Search, Terminal, Help). The command '[edureka@localhost ~]\$ tar -xvf hadoop-2.7.3.tar.gz' is entered and executed, with a green cursor at the end of the line.

Step 5: Add the Hadoop and Java paths in the bash file (.bashrc).

Open. **bashrc** file. Now, add Hadoop and Java Path as shown below.

Command: `vi .bashrc`

```
# User specific aliases and functions

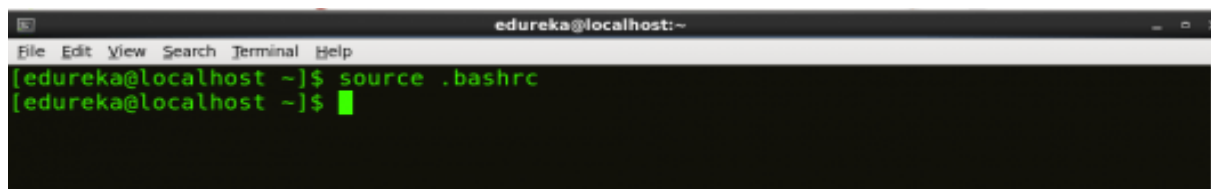
export HADOOP_HOME=$HOME/hadoop-2.7.3
export HADOOP_CONF_DIR=$HOME/hadoop-2.7.3/etc/hadoop
export HADOOP_MAPRED_HOME=$HOME/hadoop-2.7.3
export HADOOP_COMMON_HOME=$HOME/hadoop-2.7.3
export HADOOP_HDFS_HOME=$HOME/hadoop-2.7.3
export YARN_HOME=$HOME/hadoop-2.7.3
export PATH=$PATH:$HOME/hadoop-2.7.3/bin

Set JAVA_HOME
export JAVA_HOME=/home/edureka/jdk1.8.0_101
export PATH=/home/edureka/jdk1.8.0_101/bin:$PATH
```

Then, save the bash file and close it.

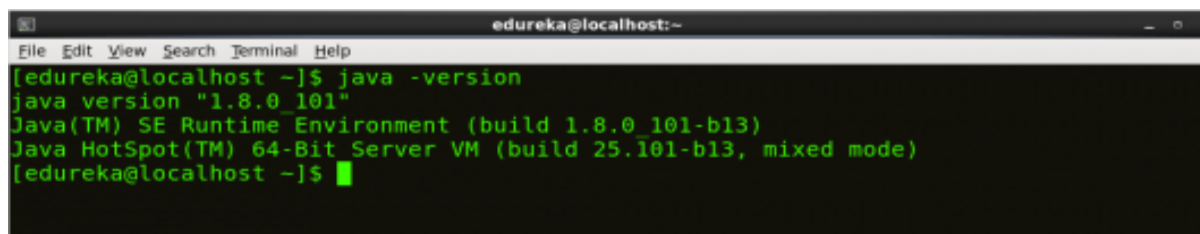
For applying all these changes to the current Terminal, execute the source command.

Command: source .bashrc

A terminal window titled 'edureka@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The prompt is '[edureka@localhost ~]\$'. The user has entered 'source .bashrc' and the prompt has moved to a new line, indicating the command was executed successfully.

To make sure that Java and Hadoop have been properly installed on your system and can be accessed through the Terminal, execute the java -version and hadoop version commands.

Command: java -version

A terminal window titled 'edureka@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The prompt is '[edureka@localhost ~]\$'. The user has entered 'java -version'. The output is: 'java version "1.8.0_101"', 'Java(TM) SE Runtime Environment (build 1.8.0_101-b13)', and 'Java HotSpot(TM) 64-Bit Server VM (build 25.101-b13, mixed mode)'. The prompt has moved to a new line.

Command: hadoop version

```
edureka@localhost:~$ hadoop version
Hadoop 2.7.3
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r baa91f7c6bc9cb92be5982de4719c1c8af91ccff
Compiled by root on 2016-08-18T01:41Z
Compiled with protoc 2.5.0
From source with checksum 2e4ce5f957ea4db193bce3734ff29ff4
This command was run using /home/edureka/hadoop-2.7.3/share/hadoop/common/hadoop-common-2.7.3.jar
[edureka@localhost ~]$
```

Step 6: Edit the [Hadoop Configuration files](#).

Command: `cd hadoop-2.7.3/etc/hadoop/`

Command: `ls`

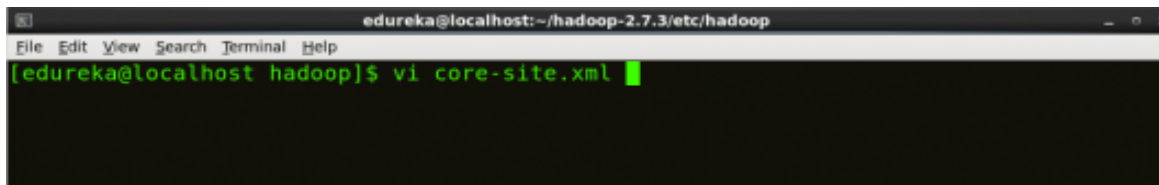
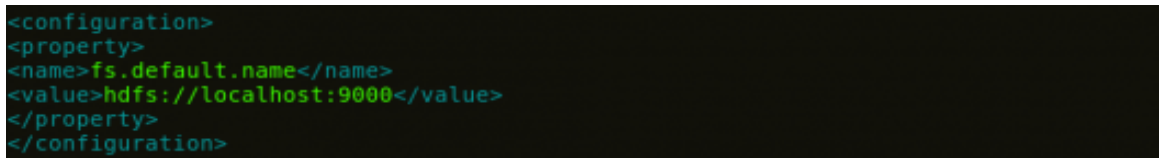
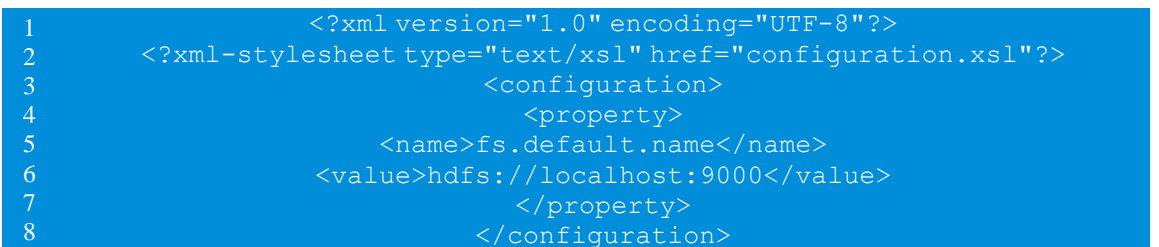
All the Hadoop configuration files are located in **hadoop-2.7.3/etc/hadoop** directory as you can see in the snapshot below:

```
edureka@localhost:~/hadoop-2.7.3/etc/hadoop$ ls
capacity-scheduler.xml      httpfs-env.sh              mapred-env.sh
configuration.xml           httpfs-log4j.properties   mapred-queues.xml.template
container-executor.cfg      httpfs-signature.secret   mapred-site.xml.template
core-site.xml               httpfs-site.xml           slaves
hadoop-env.cmd              kms-acls.xml               ssl-client.xml.example
hadoop-env.sh               kms-env.sh                 ssl-server.xml.example
hadoop-metrics2.properties kms-log4j.properties      yarn-env.cmd
hadoop-metrics.properties  kms-site.xml               yarn-env.sh
hadoop-policy.xml           log4j.properties          yarn-site.xml
hdfs-site.xml               mapred-env.cmd
```

Step 7: Open *core-site.xml* and edit the property mentioned below inside configuration tag:

core-site.xml informs Hadoop daemon where NameNode runs in the cluster. It contains configuration settings of Hadoop core such as I/O settings that are common to HDFS & MapReduce.

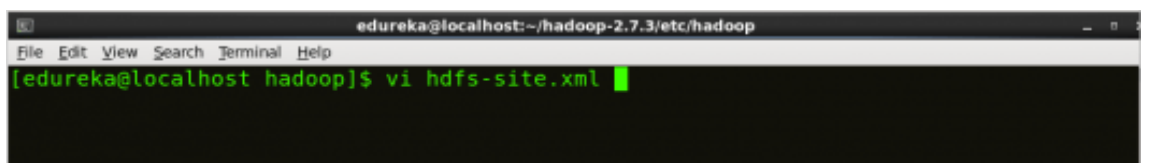
Command: vi core-site.xml

A terminal window titled 'edureka@localhost:~/hadoop-2.7.3/etc/hadoop' with a menu bar (File, Edit, View, Search, Terminal, Help). The command '[edureka@localhost hadoop]\$ vi core-site.xml' is entered and executed, with a green cursor at the end of the line.A terminal window showing the XML configuration for 'fs.default.name'. The text is: <configuration>, <property>, <name>fs.default.name</name>, <value>hdfs://localhost:9000</value>, </property>, </configuration>.A blue code block containing the XML configuration for 'fs.default.name'. The text is: 1 <?xml version="1.0" encoding="UTF-8"?>, 2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>, 3 <configuration>, 4 <property>, 5 <name>fs.default.name</name>, 6 <value>hdfs://localhost:9000</value>, 7 </property>, 8 </configuration>.

Step 8: Edit *hdfs-site.xml* and edit the property mentioned below inside configuration tag:

hdfs-site.xml contains configuration settings of HDFS daemons (i.e. NameNode, DataNode, Secondary NameNode). It also includes the replication factor and block size of HDFS.

Command: vi hdfs-site.xml

A terminal window titled 'edureka@localhost:~/hadoop-2.7.3/etc/hadoop' with a menu bar (File, Edit, View, Search, Terminal, Help). The command '[edureka@localhost hadoop]\$ vi hdfs-site.xml' is entered and executed, with a green cursor at the end of the line.

```
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.permission</name>
<value>>false</value>
</property>
```

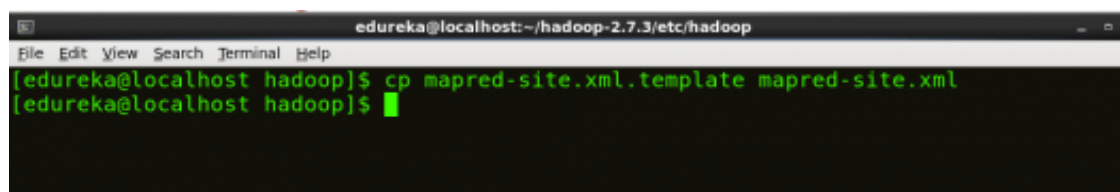
Step 9: Edit the *mapred-site.xml* file and edit the property mentioned below inside configuration tag:

mapred-site.xml contains configuration settings of MapReduce application like number of JVM that can run in parallel, the size of the mapper and the reducer process, CPU cores available for a process, etc.

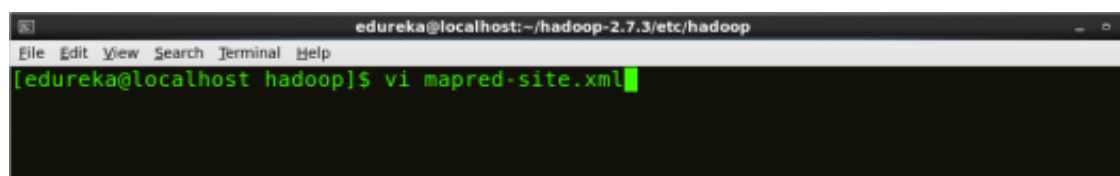
In some cases, *mapred-site.xml* file is not available. So, we have to create the *mapred-site.xml* file using *mapred-site.xml* template.

Command: cp mapred-site.xml.template mapred-site.xml

Command: vi mapred-site.xml.



```
edureka@localhost:~/hadoop-2.7.3/etc/hadoop
File Edit View Search Terminal Help
[edureka@localhost hadoop]$ cp mapred-site.xml.template mapred-site.xml
[edureka@localhost hadoop]$
```



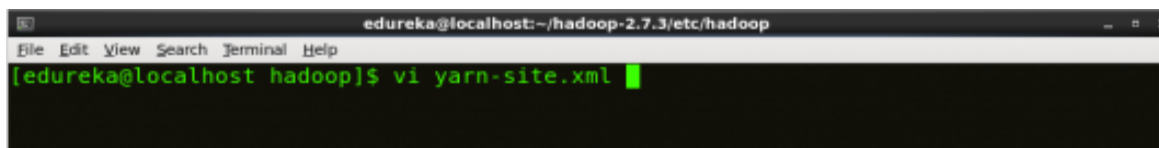
```
edureka@localhost:~/hadoop-2.7.3/etc/hadoop
File Edit View Search Terminal Help
[edureka@localhost hadoop]$ vi mapred-site.xml
```

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

Step 10: Edit *yarn-site.xml* and edit the property mentioned below inside configuration tag:

yarn-site.xml contains configuration settings of ResourceManager and NodeManager like application memory management size, the operation needed on program & algorithm, etc.

Command: vi yarn-site.xml

A terminal window with a dark background. The title bar shows 'edureka@localhost:~/hadoop-2.7.3/etc/hadoop'. The menu bar includes 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The command prompt shows '[edureka@localhost hadoop]\$ vi yarn-site.xml' with a green cursor at the end of the line.

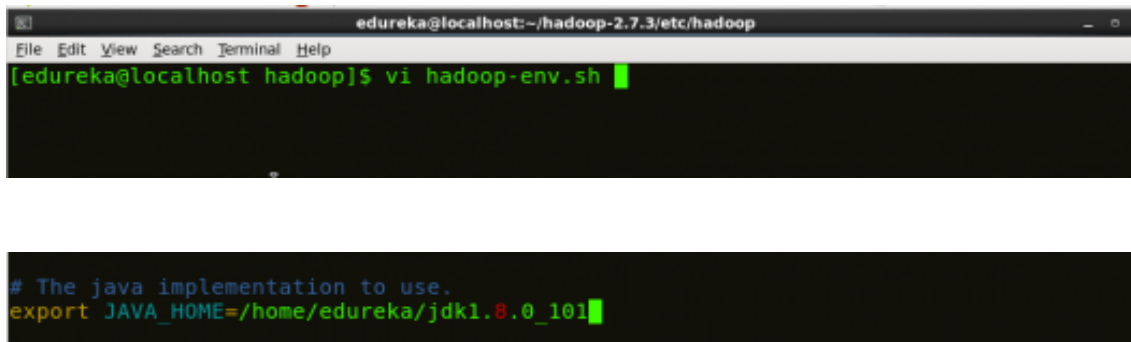
```
edureka@localhost:~/hadoop-2.7.3/etc/hadoop
[edureka@localhost hadoop]$ vi yarn-site.xml
```

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
</configuration>
```

Step 11: Edit *hadoop-env.sh* and add the Java Path as mentioned below:

hadoop-env.sh contains the environment variables that are used in the script to run Hadoop like Java home path, etc.

Command: vi *hadoop-env.sh*

A terminal window titled 'edureka@localhost:~/hadoop-2.7.3/etc/hadoop' with a menu bar (File, Edit, View, Search, Terminal, Help). The prompt is '[edureka@localhost hadoop]\$' and the command 'vi hadoop-env.sh' has been entered. A second terminal window shows the content of the file being edited: '# The java implementation to use.' followed by 'export JAVA_HOME=/home/edureka/jdk1.8.0_101' with a cursor at the end of the line.

```
edureka@localhost:~/hadoop-2.7.3/etc/hadoop
File Edit View Search Terminal Help
[edureka@localhost hadoop]$ vi hadoop-env.sh

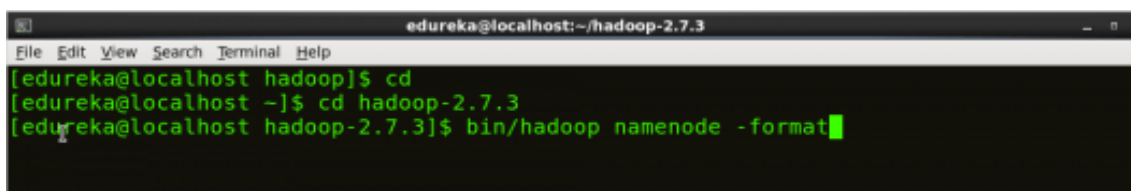
# The java implementation to use.
export JAVA_HOME=/home/edureka/jdk1.8.0_101
```

Step 12: Go to Hadoop home directory and format the NameNode.

Command: cd

Command: cd *hadoop-2.7.3*

Command: *bin/hadoop namenode -format*

A terminal window titled 'edureka@localhost:~/hadoop-2.7.3' with a menu bar (File, Edit, View, Search, Terminal, Help). The prompt is '[edureka@localhost hadoop]\$' and the command 'cd' has been entered. The prompt changes to '[edureka@localhost ~]\$'. Then the command 'cd hadoop-2.7.3' is entered, and the prompt changes to '[edureka@localhost hadoop-2.7.3]\$'. Finally, the command 'bin/hadoop namenode -format' is entered.

```
edureka@localhost:~/hadoop-2.7.3
File Edit View Search Terminal Help
[edureka@localhost hadoop]$ cd
[edureka@localhost ~]$ cd hadoop-2.7.3
[edureka@localhost hadoop-2.7.3]$ bin/hadoop namenode -format
```

This formats the HDFS via NameNode. This command is only executed for the first time.

Formatting the file system means initializing the directory specified by the *dfs.name.dir* variable.

Never format, up and running Hadoop filesystem. You will lose all your data stored in the HDFS.

Step 13: Once the NameNode is formatted, go to `hadoop-2.7.3/sbin` directory and start all the daemons.

Command: `cd hadoop-2.7.3/sbin`

Either you can start all daemons with a single command or do it individually.

Command: `./start-all.sh`

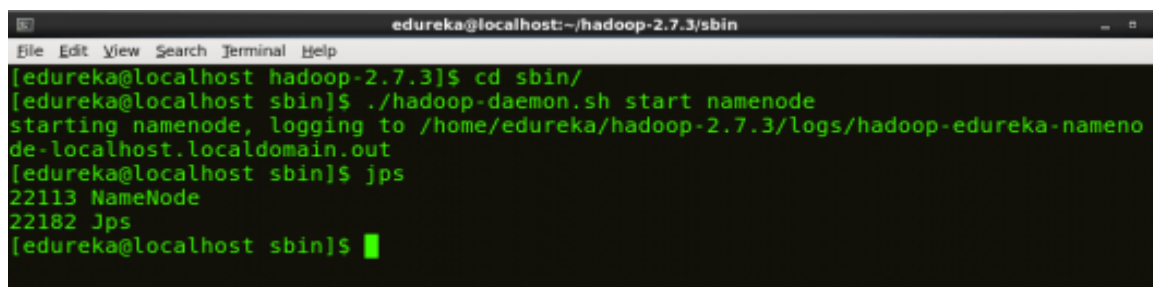
The above command is a combination of *`start-dfs.sh`*, *`start-yarn.sh`* & *`mr-jobhistory-daemon.sh`*

Or you can run all the services individually as below:

Start NameNode:

The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files stored in the HDFS and tracks all the file stored across the cluster.

Command: `./hadoop-daemon.sh start namenode`

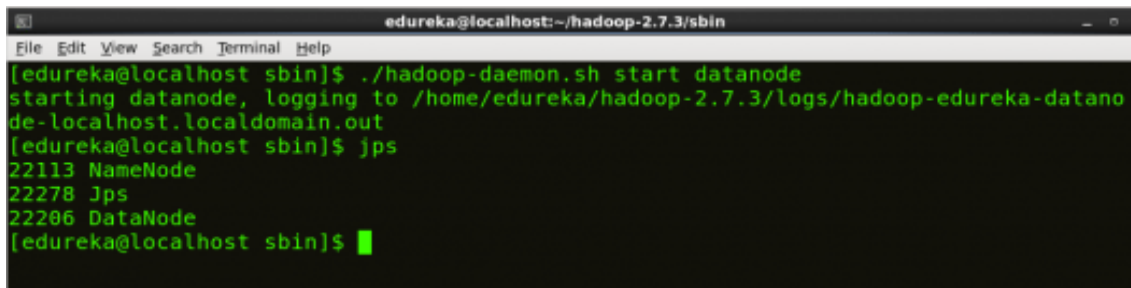
A terminal window titled 'edureka@localhost:~/hadoop-2.7.3/sbin' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the following commands and output:

```
[edureka@localhost hadoop-2.7.3]$ cd sbin/  
[edureka@localhost sbin]$ ./hadoop-daemon.sh start namenode  
starting namenode, logging to /home/edureka/hadoop-2.7.3/logs/hadoop-edureka-namenode-localhost.localdomain.out  
[edureka@localhost sbin]$ jps  
22113 NameNode  
22182 Jps  
[edureka@localhost sbin]$
```

Start DataNode:

On startup, a DataNode connects to the Namenode and it responds to the requests from the Namenode for different operations.

Command: `./hadoop-daemon.sh start datanode`

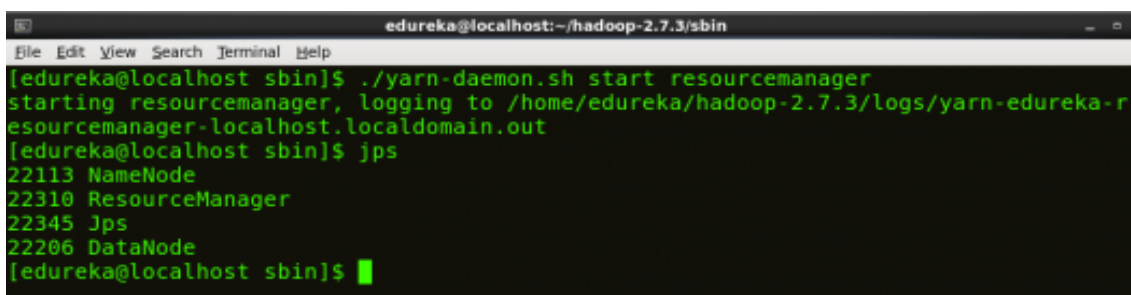


```
edureka@localhost:~/hadoop-2.7.3/sbin
File Edit View Search Terminal Help
[edureka@localhost sbin]$ ./hadoop-daemon.sh start datanode
starting datanode, logging to /home/edureka/hadoop-2.7.3/logs/hadoop-edureka-datano
de-localhost.localdomain.out
[edureka@localhost sbin]$ jps
22113 NameNode
22278 Jps
22206 DataNode
[edureka@localhost sbin]$
```

Start ResourceManager:

ResourceManager is the master that arbitrates all the available cluster resources and thus helps in managing the distributed applications running on the YARN system. Its work is to manage each NodeManagers and the each application's ApplicationMaster.

Command: `./yarn-daemon.sh start resourcemanager`

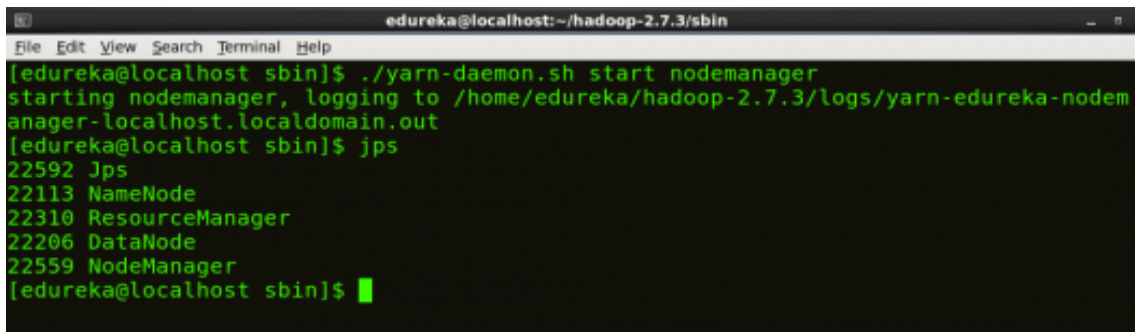


```
edureka@localhost:~/hadoop-2.7.3/sbin
File Edit View Search Terminal Help
[edureka@localhost sbin]$ ./yarn-daemon.sh start resourcemanager
starting resourcemanager, logging to /home/edureka/hadoop-2.7.3/logs/yarn-edureka-r
esourcemanager-localhost.localdomain.out
[edureka@localhost sbin]$ jps
22113 NameNode
22310 ResourceManager
22345 Jps
22206 DataNode
[edureka@localhost sbin]$
```

Start NodeManager:

The NodeManager in each machine framework is the agent which is responsible for managing containers, monitoring their resource usage and reporting the same to the ResourceManager.

Command: `./yarn-daemon.sh start nodemanager`



```
edureka@localhost:~/hadoop-2.7.3/sbin
File Edit View Search Terminal Help
[edureka@localhost sbin]$ ./yarn-daemon.sh start nodemanager
starting nodemanager, logging to /home/edureka/hadoop-2.7.3/logs/yarn-edureka-nodemanager-localhost.localdomain.out
[edureka@localhost sbin]$ jps
22592 Jps
22113 NameNode
22310 ResourceManager
22206 DataNode
22559 NodeManager
[edureka@localhost sbin]$
```

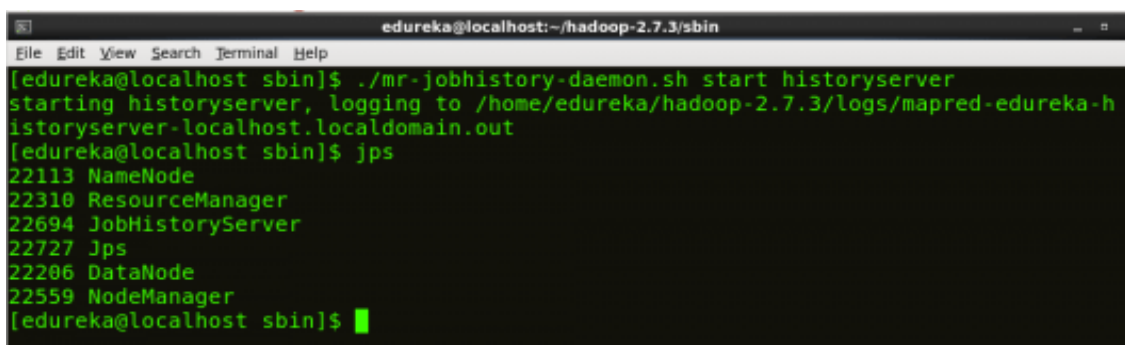
Start JobHistoryServer:

JobHistoryServer is responsible for servicing all job history related requests from client.

Command: `./mr-jobhistory-daemon.sh start historyserver`

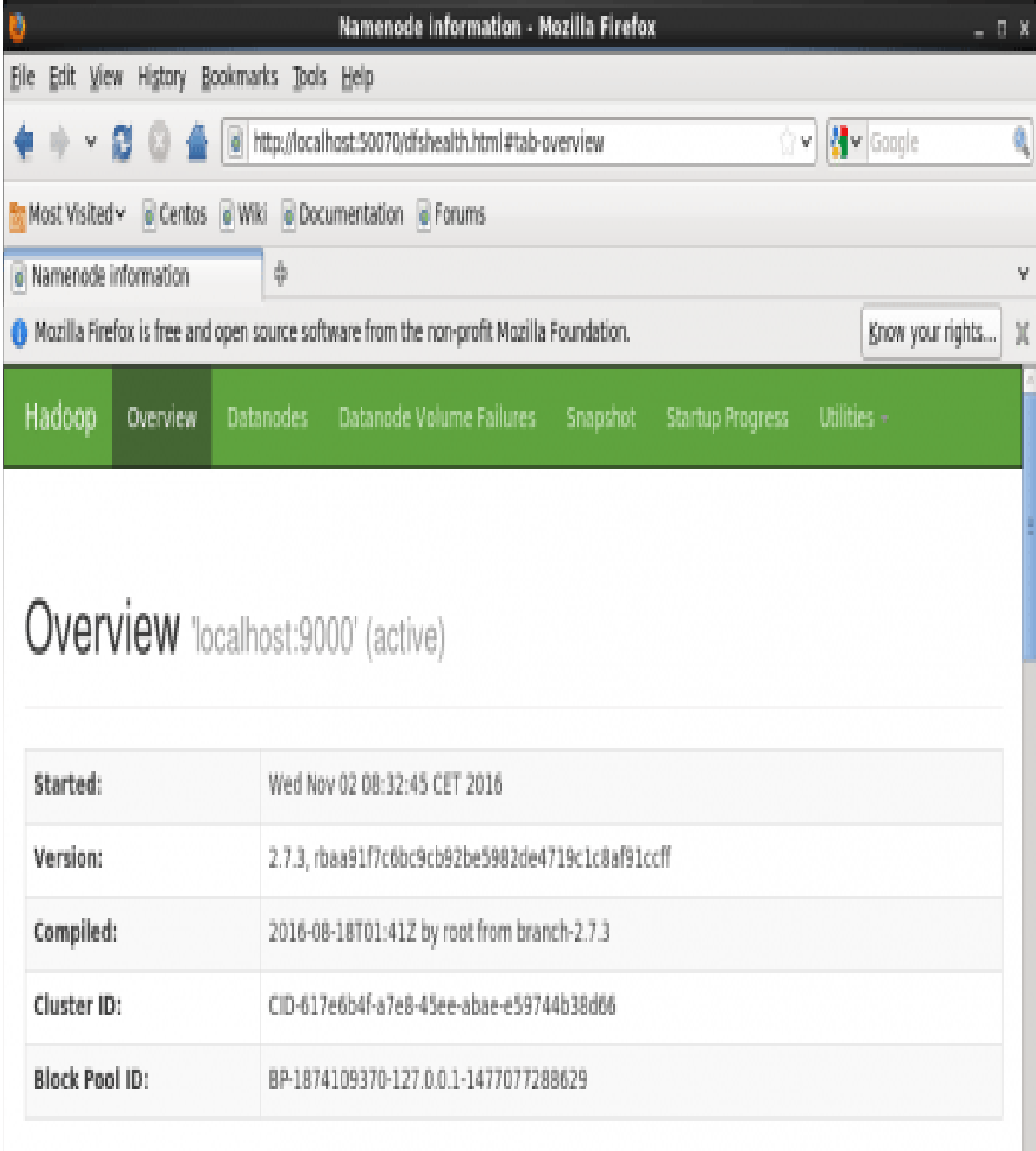
Step 14: To check that all the Hadoop services are up and running, run the below command.

Command: `jps`



```
edureka@localhost:~/hadoop-2.7.3/sbin
File Edit View Search Terminal Help
[edureka@localhost sbin]$ ./mr-jobhistory-daemon.sh start historyserver
starting historyserver, logging to /home/edureka/hadoop-2.7.3/logs/mapred-edureka-historyserver-localhost.localdomain.out
[edureka@localhost sbin]$ jps
22113 NameNode
22310 ResourceManager
22694 JobHistoryServer
22727 Jps
22206 DataNode
22559 NodeManager
[edureka@localhost sbin]$
```

Step 15: Now open the Mozilla browser and go to **localhost:50070/dfshealth.html** to check the NameNode interface.



The screenshot shows a Mozilla Firefox browser window titled "Namenode Information - Mozilla Firefox". The address bar displays the URL `http://localhost:50070/dfshealth.html#tab-overview`. Below the address bar, there are bookmarks for "Centos", "Wiki", "Documentation", and "Forums". A notification bar at the bottom of the browser states "Mozilla Firefox is free and open source software from the non-profit Mozilla Foundation." with a "Know your rights..." link.

The main content area of the browser shows the Hadoop NameNode Information interface. It has a green navigation bar with the following links: "Hadoop", "Overview", "Datanodes", "Datanode Volume Failures", "Snapshot", "Startup Progress", and "Utilities". The "Overview" link is currently selected.

Overview 'localhost:9000' (active)

Started:	Wed Nov 02 08:32:45 CET 2016
Version:	2.7.3, rbaa91f7c6bc9cb92be5982de4719c1c8af91ccff
Compiled:	2016-08-18T01:41Z by root from branch-2.7.3
Cluster ID:	CID-617e6b4f-a7e8-45ee-abae-e59744b38d66
Block Pool ID:	BP-1874109370-127.0.0.1-1477077288629

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

In this paper, a probabilistic data acquisition method is designed for the cloud based healthcare system. Besides ,laCE and leCE algorithms are designed for the intra and inter cluster correlation analysis of the healthcare big data. An FHCP algorithm is designed to predict the future health condition of the patients based on their current health status with the accuracy of 98%. In addition, cloud-based Map Re-duce model is used as the processing framework for our big data analysis. It is observed that our protocol can be used for various applications related to healthcare and patient monitoring such as heart disease prediction or cancer severity classification. Our future work is to implement the proposed data analytic model in the real healthcare domain to analyze the data in real-time data analytic platform such as SPARK.

I have performed analysis on our dataset and our project is completed using these.

. Nifi

. Hadoop

. Map Reduce

. HDFS

.Hive

. Microsoft Power BI

Nowadays, various biomedical and healthcare tools such as genomics, mobile biometric sensors, and smartphone apps generate a big amount of data. Therefore, it is mandatory for us to know about and assess that can be achieved using this data. For example, the analysis of such data can provide further insights in terms of procedural, technical, medical and other types of improvements in healthcare. After a review of these healthcare procedures, it appears that the full potential of patient-specific medical specialty or personalized medicine is under way. The collective big data analysis of EHRs, EMRs and other medical data is continuously helping build a better prognostic framework. The companies providing service for healthcare analytics and clinical transformation are indeed contributing towards better and effective outcome. Common goals of these companies include reducing cost of analytics, developing effective Clinical Decision Support (CDS) systems, providing platforms for better treatment strategies, and identifying and preventing fraud associated with big data. Though, almost all of them face challenges on federal issues like how private data is handled, shared and kept safe. The combined pool of data from healthcare organizations and biomedical researchers have resulted in a better outlook, determination, and treatment of various diseases. This has also helped in building a better and healthier personalized healthcare framework. Modern healthcare fraternity has realized the potential of big data and therefore, have implemented big data analytics in healthcare and clinical practices. Supercomputers to

quantum computers are helping in extracting meaningful information from big data in dramatically reduced time periods. With high hopes of extracting new and actionable knowledge that can improve the present status of healthcare services, researchers are plunging into biomedical big data despite the infrastructure challenges. Clinical trials, analysis of pharmacy and insurance claims together, discovery of biomarkers is a part of a novel and creative way to analyze healthcare big data.

Big data analytics leverage the gap within structured and unstructured data sources. The shift to an integrated data environment is a well-known hurdle to overcome. Interesting enough, the principle of big data heavily relies on the idea of the more the information, the more insights one can gain from this information and can make predictions for future events. It is rightfully projected by various reliable consulting firms and health care companies that the big data healthcare market is poised to grow at an exponential rate. However, in a short span we have witnessed a spectrum of analytics currently in use that have shown significant impacts on the decision making and performance of healthcare industry. The exponential growth of medical data from various domains has forced computational experts to design innovative strategies to analyze and interpret such enormous amount of data within a given timeframe. The integration of computational systems for signal processing from both research and practicing medical professionals has witnessed growth. Thus, developing a detailed model of a human body by combining physiological data and “-omics” techniques can be the next big target. This unique idea can enhance our knowledge of disease conditions and possibly help in the development of novel diagnostic tools. The

continuous rise in available genomic data including inherent hidden errors from experiment and analytical practices need further attention. However, there are opportunities in each step of this extensive process to introduce systemic improvements within the healthcare research.

High volume of medical data collected across heterogeneous platforms has put a challenge to data scientists for careful integration and implementation. It is therefore suggested that revolution in healthcare is further needed to group together bioinformatics, health informatics and analytics to promote personalized and more effective treatments. Furthermore, new strategies and technologies should be developed to understand the nature (structured, semi-structured, unstructured), complexity (dimensions and attributes) and volume of the data to derive meaningful information. The greatest asset of big data lies in its limitless possibilities. The birth and integration of big data within the past few years has brought substantial advancements in the health care sector ranging from medical data management to drug discovery programs for complex human diseases including cancer and neurodegenerative disorders. To quote a simple example supporting the stated idea, since the late 2000's the healthcare market has witnessed advancements in the EHR system in the context of data collection, management and usability. We believe that big data will add-on and bolster the existing pipeline of healthcare advances instead of replacing skilled manpower, subject knowledge experts and intellectuals, a notion argued by many. One can clearly see the transitions of health care market from a wider volume base to personalized or individual specific domain. Therefore, it is essential for

technologists and professionals to understand this evolving situation. In the coming year it can be projected that big data analytics will march towards a predictive system. This would mean prediction of futuristic outcomes in an individual's health state based on current or existing data (such as EHR-based and Omics-based). Similarly, it can also be presumed that structured information obtained from a certain geography might lead to generation of population health information. Taken together, big data will facilitate healthcare by introducing prediction of epidemics (in relation to population health), providing early warnings of disease conditions, and helping in the discovery of novel biomarkers and intelligent therapeutic intervention strategies for an improved quality of life.

REFERENCES

- [1] Laney D. 3D data management: controlling data volume, velocity, and variety, Application delivery strategies. Stamford: META Group Inc; 2001.

- [2] Mauro AD, Greco M, Grimaldi M. A formal definition of big data based on its essential features. *Libr Rev.* 2016;65(3):122–35.

- [3] Gubbi J, et al. Internet of Things (IoT): a vision, architectural elements, and future directions. *Future Gener Comput Syst.* 2013;29(7):1645–60.

- [4] Shvachko K, et al. The hadoop distributed file system. In: *Proceedings of the 2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. New York: IEEE Computer Society; 2010. p. 1–10.

- [5] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM.* 2008;51(1):107–13.

[6] Zaharia M, et al. Apache Spark: a unified engine for big data processing. Commun ACM. 2016;59(11):56–65.

[7] Gopalani S, Arora R. Comparing Apache Spark and Map Reduce with performance analysis using K-means; 2015.

[8] Ahmed H, et al. Performance comparison of spark clusters configured conventionally and a cloud service. Procedia Comput Sci. 2016;82:99–106.

[9] Saouabi M, Ezzati A. A comparative between hadoop mapreduce and apache Spark on HDFS. In: Proceedings of the 1st international conference on internet of things and machine learning.

Liverpool: ACM; 2017. p. 1–4.

[10] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, “The Internet of Things for health care: A comprehensive survey,” IEEE Access, vol. 3, pp. 678–708, 2015.

[11] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, “Big data for health,” IEEE J. Biomed. Health Inform., vol. 19, no. 4, pp. 1193–1208, Jul. 2015.

[12] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Inf. Sci. Syst.*, vol. 2, no. 1, pp. 1–10, 2014.

[13] (Apr. 2014). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. [Online]. Available: <https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf>

[14] K. Lin, F. Xia, W. Wang, D. Tian, and J. Song, "System design for big data application in emotion-aware healthcare," *IEEE Access*, vol. 4, pp. 6901–6909, 2016.

[15] L. A. Tawalbeh, R. Mehmood, E. Benkhelifa, and H. Song, "Mobile cloud computing model and big data analysis for healthcare applications," *IEEE Access*, vol. 4, pp. 6171–6180, 2016.

[16] C. K. Dehury and P. K. Sahoo, "Design and implementation of a novel service management framework for IoT devices in cloud," *J. Syst. Softw.*, vol. 119, pp. 149–161, Sep. 2016.

[17] Z. Yu et al., "Incremental semi-supervised clustering ensemble for high dimensional data clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 701–714, Mar. 2016.

- [18] M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, 2012.
- [19] S. Rallapalli, R. R. Gondkar, and U. P. K. Ketavarapu, "Impact of processing and analyzing healthcare big data on cloud computing environment by implementing hadoop cluster," *Procedia Comput. Sci.*, vol. 85, pp. 16–22, May 2016.
- [20] S. Wang, X. Chang, X. Li, G. Long, L. Yao, and Q. Z. Sheng, "Diagnosis code assignment using sparsity-based disease correlation embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3191–3202, Dec. 2016.
- [21] V. Tresp, J. M. Overhage, M. Bundschuh, S. Rabizadeh, P. A. Fasching, and S. Yu, "Going digital: A survey on digitalization and large-scale data analytics in healthcare," *Proc. IEEE*, vol. 104, no. 11, pp. 2180–2206, Nov. 2016.
- [22] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, "Promises and challenges of big data computing in health sciences," *Big Data Res.*, vol. 2, no. 1, pp. 2–11, 2015.
- [23] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[24] C.-H. Wu and Y.-C. Tseng, "Data compression by temporal and spatial correlations in a body-area sensor network: A case study in Pilates motion recognition," *IEEE Trans. Mobile Comput.*, vol. 10, no. 10, pp. 1459–1472, Oct. 2011.

[25] R. A. Taylor et al., "Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data-driven, machine learning approach," *Acad. Emerg. Med.*, vol. 3, no. 23, pp. 269–278, Mar. 2016.