

PERSONALIZATION FOR FEATURE SELECTION IN CLINICAL FORM AUTO FILLING

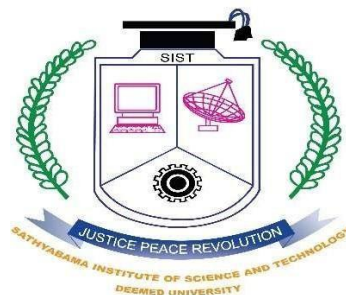
Submitted in Partial Fulfilment of The Requirements

For The Award Of

Bachelor of Engineering Degree in Computer Science and

Engineering By

VEMPATI SRINADH(37110823)



**DEPARTMENT OF COMPUTER SCIENCE AND
ENGINEERING SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND
TECHNOLOGY (DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by

NAAC JEPPIAARNAGAR, RAJIV GANDHI SALAI, CHENNAI

I-

600119

MARCH-2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NACC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai –
600119 www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this project report is the bonafide work of **V E M P A T I S R I N A D H (3 7 1 1 0 8 2 3)** who carried out the project entitled "**PERSONALIZATION FOR FEATURE SELECTION IN CLINICAL FORM AUTO FILLING**" under my supervision from **November-2020 to April -2021**.

Internal Guide

Mr. Saravanan. M. E.,

Head of the Department

Submitted for viva voce Examination held on _____.

Internal Examiner

External Examiner

DECLARATION

I, VEMPATI SRINADH(37110823)herebydeclare thatthe projectreportentitled “
**PERSONALIZATIONFOR FEATURE SELECTIONIN CLINICAL FORM AUTO
FILLING**”

donebymeundertheguidanceof**Mr.Saravanan.,M.E.**,issubmittedinpartialfulfilmentoft
herequirementsfortheawardofBachelorofEngineeringdegreeinComputerScienceAnd
Engineering.

DATE:

PLACE:CHENNAI.

SIGNATUREOFTHECANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr.T. Sasikala M.E., Ph.D., Dean**, School of Computing, **Dr. S. Vigneshwari M.E., Ph.D.** and **Dr. L. Lakshmanan M.E., Ph.D., Heads** of the Department of computer science and engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Mr. Saravanan M.E.**, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Computer Science and Engineering who were helpful in many ways for the completion of the project.

ABSTRACT

Feature learning and selection have been widely applied in many research areas because of their good performance and lower complexity. Traditional methods usually treat all terms with same feature sets, such that performance can be damaged when noisy information is brought by wrong features for a given term. Here, we propose a term-based personalization approach to finding the best features for each term. First, features are given as the input so that we focus on selection strategies. Second, we present a feature searching method to generate feature candidate subsets for each term. Here, the big data only deals with the analysis of the patient details, problems, reports and hospitals specialties. Finally, we obtain the personalized feature set for each term as a subset of all features so they can be used to integrate onto the searching method.

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	v
	LIST OF FIGURES	viii
1	INTRODUCTION	1
	1.1 OVERVIEW	1
	1.2 BIG DATA	1
	1.3 TYPES OF BIG DATA	2
	1.4 CHALLENGES OF BIG DATA	3
	1.5 APPLICATIONS OF BIG DATA	3
	1.6 FEATURE SELECTION	5
	1.6.1 CHALLENGES OF FEATURE SELECTION	5
	1.6.2 OBJECTIVE OF FEATURE SELECTION	5
2	LITERATURE SURVEY	7
	2.1 CLASSIFY SELF-REPORTED HEALTH STATUS	7
	2.2 DYNAMIC PREDICTION IN CLINICAL SURVIVAL ANALYSIS	7
	2.3 GUEST-EDITORIAL BIOMEDICAL INFORMATICS IN CLINICAL ENVIRONMENTS	8
	2.4 LEARNING MULTI-LABEL SCENE CLASSIFICATION	8
	2.5 PROGRESSION OF HYPERTENSIVE HEART DISEASE	9
3	AIM AND SCOPE OF THE PROJECT	10
	3.1 AIM	10
	3.2 SCOPE OF THE PROJECT	10
	3.3 EXISTING SYSTEM	10
	3.3.1 DRAWBACKS OF EXISTING SYSTEM	11
	3.4 PROPOSED SYSTEM	11
	3.5 ADVANTAGES	13
4	METHODOLOGY	15
	4.1 MODULES	15

	4.1.1 PRE-PROCESSINGCLINICALDATABASE	15
	4.1.2 STORAGE	16
	4.1.3 ANALYZE QUERY	19
	4.1.4 ANALYSISLATINSCRIPT (PIG)	23
	4.1.5 PROCESSING (MAPREDUCE)	25
	4.2 ALGORITHMS	28
	4.3 REQUIREMENTANALYSIS	29
	4.3.1 HARDWARE ENVIRONMENT	29
	4.3.2 SOFTWARE ENVIRONMENT	30
5	RESULT	31
	OUTPUT	31
6	CONCLUSIONANDFUTUREWORK	33
	REFERENCES	34
	APPENDIX	35
	A. SOURCECODE	35
	B. SCREENSHOTS	42
	C. PUBLICATION	45

LISTOFFIGURES

FIGURENO	FIGURENAME
1.1	BIGDATA
3.1	MULTINODECLUSTER
4.1	PRE-PROCESSING
4.2	STORAGE
4.3	HDFSARCHITECTURE
4.4	ANALYSING
4.5	HIVE
4.6	MAPREDUCE
4.7	ANALYZEQUERY
4.8	PIG
4.9	ANALYSISOFPIG
4.10	MAPREDUCE
4.11	MAPREDUCEARCHITECTURE

CHAPTER

1 INTRODUCTION

1.1 OVERVIEW

N

It is broadly acknowledged that well-being data devices and methods can be misused effectively to help specialists in diagnosing and treating their patients all the more productively. Utilizing their experience and information, the doctors order patients and analyze their infections, however in doing as such, it is plausible that they submit a few errors, especially when they need sufficient experience or when their staff of judgment is poor. In such circumstances, Clinical Decision Support (CDS) frameworks, including frameworks that give finding, customized restorative estimation, treatment and important learning, would be useful to the doctors by method for giving them particular learning, patients' data and keen applications, which can enhance the proficiency of their basic leadership form CDS frameworks center on separating attributes of patients, in light of which they order patients and give relating clinical recommendations to the doctors.

Traditional methods usually treat all terms with same feature sets, such that performance can be damaged when noisy information is brought by wrong features for a given term. Here, we propose a term-based personalization approach to finding the best features for each term. First, features are given as the input so that we focus on selection strategies. Second, we present a feature searching method to generate feature candidate subsets for each term. The Big data deals with the analysis of the patient details, problems, reports and hospital specialties. Finally, we obtain the personalized feature set for each term as a subset of all features.

1.2 BIG DATA

The data which is beyond the storage capacity and beyond the processing power such a data is called Big Data. Big data means really a big data; it is a collection of large data sets that cannot be processed using traditional computing techniques. Big data is not merely a data; rather it has become a complete subject, which involves various tools, techniques and frameworks. Data which are very large in size is called Big Data. Normally we work on data of size MB (Wordbook, Excel) or maximum GB (Movies, Codes) but data in Petabytes i.e. 10^{15} -bytes size is called Big Data. It is stated that almost 90% of today's data has

being generated in the past 6 years. However, there are certain basic tenets of Big Data that will make it even simpler to answer what is Big Data.

- It refers to a massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.

The same concept applies on Big Data. Big Data says, till today, we were okay with storing the data into our servers because the volume of the data was pretty limited, and the amount of time to process this data was also okay. But now in this current technological world, the data is growing too fast and people are relying on the data a lot of times. Also the speed at which the data is growing, it is becoming impossible to store the data into any server.

1.3 TYPES OF BIG DATA

1.3.1 Structured

By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc., will be presented in an organized manner.

1.3.2 Unstructured

Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data.

1.3.3 Semi-Structured

Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.

1.4 CHALLENGES WITH BIG DATA

- *Data Quality* : The problem here is the 4th V i.e. Veracity. The data here is very messy, inconsistent and incomplete. Dirty data cost \$600 billion to the companies every year in the United States.
- *Discovery*: Finding insights on Big Data is like finding a needle in a haystack. Analyzing petabytes of data using extremely powerful algorithms to find patterns and insights are very difficult.
- *Storage*: The more data an organization has, the more complex the problems of managing it can become. The question that arises here is "Where to store it?". We need a storage system which can easily scale up or down on-demand.
- *Analytics*: In the case of Big Data, most of the time we are unaware of the kind of data we are dealing with, so analyzing that data is even more difficult.
- *Security*: Since the data is huge in size, keeping it secure is another challenge. It includes user authentication, restricting access based on user, recording data access histories, proper use of data encryption etc.
- *Lack of Talent*: There are a lot of Big Data projects in major organizations, but a sophisticated team of developers, data scientists and analysts who also have sufficient amount of domain knowledge is still a challenge.

1.5 APPLICATIONS OF BIG DATA

Cannot talk about data without talking about the people, people who are getting benefited by Big Data applications. Almost all the industries today are leveraging Big Data applications in one or the other way.



Fig1.1BigData

- *SmarterHealthcare:* Making use of the petabytes of patient's data, the organization can extract meaningful information and then build applications that can predict the patient's deteriorating condition in advance.
- *Telecom:* Telecom sectors collect information, analyze it and provide solutions to different problems. By using Big Data applications, telecom companies have been able to significantly reduce data packet loss, which occurs when networks are overloaded, and thus, providing a seamless connection to their customers.
- *Retail:* Retail has some of the tightest margins, and is one of the greatest beneficiaries of big data. The beauty of using big data in retail is to understand consumer behavior. Amazon's recommendation engine provides suggestions based on the browsing history of the consumer.
- *TrafficControl:* Traffic congestion is a major challenge for many cities globally. Effective use of data and sensors will be key to managing traffic better as cities become increasingly densely populated.
- *Manufacturing:* Analyzing big data in the manufacturing industry can reduce component defects, improve product quality, increase efficiency, and save time and money.
- *Search Quality:* Every time we are extracting information from Google, we are simultaneously generating data for it. Google stores this data and uses it to improve its search quality.

1.6 FEATURESELECTION

Feature selection has been an important research area in data mining, which chooses a subset of relevant features for use in the model building. This paper aims to provide an overview of feature selection methods for big data mining. First, it discusses the current challenges and difficulties faced when mining valuable information from big data. A comprehensive review of existing feature selection methods in big data is then presented. Herein, we approach the review from two aspects: methods specific to a particular kind of big data with certain characteristics and applications of methods in classification analysis, which are significantly different from the existing review work. This paper also highlights the current issues of feature selection in big data and suggests the future research directions.

1.6.1 Challenges Of Feature Selection

Compared to traditional data, some influential points need to be highlighted on extracting valuable information from big data. Taking the 3V characteristics into consideration, traditional feature selection methods face the following

threefold challenges with respect to the case of big data. Traditional feature selection methods usually require large amounts of learning time, so it is hard for processing speed to catch up with the change of big data. Generally, big data not only

include an immense amount of irrelevant and/or redundant features, but also have possible noises of different degrees and different types, which greatly increases the difficulty of selecting feature. Some data are unreliable/forged, due to different means of acquisition, or even loss, which further enhances the complexity of feature selection.

1.6.2 Objective Of Feature Selection

Feature learning and selection have been widely applied in many research areas because of their good performance and lower complexity.

- To propose a term-based personalization approach to finding the best features for each term.
- First, features are given as the input so that we focus on selection strategies.

- Second, we present a feature searching method to generate feature candidate subsets for each term.
- Here, the big data only deals with the analysis of the patient details, problems, reports and hospital specialties.
- Finally, we obtain the personalized feature set for each term as a subset of all features so they can be used to integrate into the searching method.

CHAPTER 2 LITERATURE SURVEY

2.1 TITLE: A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients with Heart Disease using Activity Tracker Data.

2.1.1 Author: Yiwen Meng, William Speier, Member, Chrisandra Shufelt, Sandy Joung, Jennifer E Van Eyk, C. Noel Bairey Merz, Mayra Lopez, Brennan Spiegel and Corey W. Arnold

2.1.2 Description: Constructing statistical models using personal sensor data could allow for tracking health status over time, thereby enabling the possibility of early intervention. The goal of this study was to use machine learning algorithms to classify patient-reported outcomes (PROs) using activity tracker data in a cohort of patients with stable ischemic heart disease (SIHD). A population of 182 patients with SIHD was monitored over a period of 12 weeks. Each subject received a Fitbit Charge 2 device to record daily activity data, and each subject completed eight Patient-Reported Outcomes Measurement Information Systems (PROMIS®) shortform at the end of each week as a self-assessment of their health status. Two models were built to classify PRO scores using activity tracker data. The first model treated each week independently, while the second used a Hidden Markov model (HMM) to take advantage of correlations between successive weeks. Retrospective analysis compared the classification accuracy of the two models and the importance of each feature. In the independent model, a random forest classifier achieved a mean area under curve (AUC) of 0.76 for classifying the Physical Function PRO.

2.2 TITLE: Dynamic Prediction in Clinical Survival Analysis using Temporal Convolutional Networks

2.2.1 Author: Daniel Jarrett, Jinsung Yoon, Member, IEEE, and Mihaela van der Schaar, Fellow, IEEE

2.2.2 Description: Accurate prediction of disease trajectories is critical for early identification and timely treatment of patients at risk. Conventional methods

in survival analysis are often constrained by strong parametric assumptions and

limited in their ability to learn from high-dimensional data. This paper develops a novel convolutional approach that addresses the drawbacks of both traditional statistical approaches as well as recent neural network models for survival. We present MATCH-Net: A Missing Awareness Temporal Convolutional Hitting-

time Network, designed to capture temporal dependencies and heterogeneous interactions in covariate trajectories and patterns of missingness. To the best of our knowledge, this is the first investigation of temporal convolutions in the context of dynamic prediction for personalized risk prognosis

2.3 TITLE: Guest-Editorial Biomedical Informatics in Clinical Environments

2.3.1 Author: M. AKAY and D. I. FOTIADIS

2.3.2 Description: The point of this uncommon segment is to give an outline of the rising biomedical informatics advancements and their application in inquire about and clinical conditions. Later advancements in biomedical informatics have made strategies, strategies and apparatuses, which depend on the investigation of heterogeneous information, information mining, choice emotionally supportive networks, multiscale demonstrating, and so on. These separation from the improvement of such frameworks what's more, the genuine clinical situation is still sufficiently long, and just some of them have been utilized in a clinical scale. Physiological signals, such as Electrocardiograms and Electroencephalograms, are used for the diagnosis of the patient status and detection of events. In, an unsupervised, robust, and computationally fast algorithm that uses Modified Multiscale Sample Entropy and Kurtosis to automatically identify the independent eye blink components is presented.

2.4 TITLE: Learning multi-label scene classification

2.4.1 Author: Matthew R. Boutella, Jiebo Luo

2.4.2 Description: In great example acknowledgment issues, classes are fundamentally unrelated by definition. Classification mistakes happen when the classes cover in the component space. We look at a different circumstance, happening when the classes are, by definition, not totally unrelated. Such issues emerge in semantic scene and archive classification and in restorative analysis. We present a system to deal with such issues and apply it to the issue of semantic

scene classification where a characteristic scene may contain various protests to such an extent that the scene can be depicted by different class names (e.g., a scene with a mountain in the foundation). Such an issue presents difficulties to the exemplary example acknowledgment worldview and requests a different treatment. We talk about methodologies for preparing and testing in this situation and present new measurements for assessing singular precedents, class review and exactness, and by and large precision. Investigations demonstrate that our strategies are appropriate for scene classification; besides, our work seems to sum up to other classification issues of a similar sort.

2.5 TITLE: The Progression of Hypertensive Heart Disease

2.5.1 Author: Mark H. Drazner

2.5.2 Description: Hypertension remains a noteworthy general medical issue related with significant grimness and mortality. Hypertensive coronary illness is a group of stars of variations from the norm that incorporates left ventricular hypertrophy (LVH), systolic and diastolic brokenness, and their clinical appearances including arrhythmias and symptomatic heart disappointment. The work of art worldview of hypertensive coronary illness is that the left ventricular (LV) divider thickens in light of hoisted blood weight as a compensatory instrument to limit divider push. In this way, after a progression of ineffectively portrayed occasions ("change to disappointment"), the left ventricle expands, and the LV discharge portion (EF) decreases (characterized thus as "widened heart failure").¹ The reason for this survey is to centre around the key strides in the movement of hypertensive coronary illness, featuring ongoing advances too as some uncertain discussions.

CHAPTER 3

AIM AND SCOPE OF THE PROJECT

3.1 AIM

To propose a term-based personalization approach to finding the best features for each term. The big data only deals with the analysis of the patient details, problems, reports and hospital specialties. Finally, we obtain the personalized feature set for each term as a subset of all features so they can be used to integrate onto searching method. Analyzing clinical data by using Hadoop tool along with some Hadoop ecosystems like HDFS, MapReduce, Sqoop, Hive and Pig. By using these tools we can process no limitation of data, no data lost problem, we can get high throughput, maintenance cost also very less and it is an open source software, it is compatible on all the platforms since it is Java based. In clinical

data details, this is based on whenever patient enters their ID in hospital database within not too much time their data will present.

3.2 SCOPE OF THE PROJECT

Analyzing Clinical Data by using Hadoop tool along with some Hadoop ecosystems like HDFS, MapReduce, Sqoop, Hive and Pig. By using these tools, we can process no limitation of data, no data lost problem, we can get high throughput, maintenance cost also very less and it is an open source software, it is compatible on all the platforms since it is Java based. In clinical, we will know which hospital is curing better for particular health problems.

3.3 EXISTING SYSTEM

Existing concept deals with providing backend by using MySQL which contains lot of drawbacks i.e. data limitation is that processing time is high when the data is huge and once data is lost, we cannot recover so thus we propose a concept by using Hadoop tool. MySQL is a fast, easy to use RDBMS being used for many small and big businesses. MySQL is developed, marketed and supported by MySQL AB, which is a Swedish company. MySQL is becoming so popular because of many good reasons.

- MySQL is released under an open-source license. So you have nothing to pay to use it.
- MySQL is a very powerful program in its own right. It handles a large subset of the functionality of the most expensive and powerful database packages.
- MySQL uses a standard form of the well-known SQL data language.
- MySQL works on many operating systems and with many languages including PHP, PERL, C, C++, JAVA, etc.
- MySQL works very quickly and works well even with large datasets.
- MySQL is very friendly to PHP, the most appreciated language for web development.
- MySQL supports large databases, up to 50 million rows or more in a table. The default file size limit for a table is 4GB, but you can increase this (if your operating system can handle it) to a theoretical limit of 8 million terabytes (TB).
- MySQL is customizable. The open-source GPL license allows programmers to modify the MySQL software to fit their own specific environments.

3.3.1 Drawbacks of Existing System

- We can process limitation of data.
- We get results with take more time and maintenance cost is very high.

3.4 PROPOSED SYSTEM

Proposed concept deals with providing database by using Hadoop tool we can analyze no limitation of data and simply add number of machines to the cluster and we get results with less time, high throughput and maintenance cost is very less and we are using joins, partitions and bucketing techniques in Hadoop.

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop consists of the Hadoop Common package, which provides filesystem and operating system level abstractions, a MapReduce either MapReduce and the Hadoop Distributed File System (HDFS). The Hadoop Common package contains files and scripts needed to start Hadoop.

For effective scheduling of work, every Hadoop-compatible filesystem should provide location awareness, which is the name of the rack, specifically the network switch where a worker node is. Hadoop applications can use this information to execute code on the node where the data is, and, failing that, on the same rack/switch to reduce backbone traffic. HDFS uses this method when replicating data for data redundancy across multiple racks. This approach reduces the impact of a rack power outage or switch failure if any of these hardware failures occurs, the data will remain available.

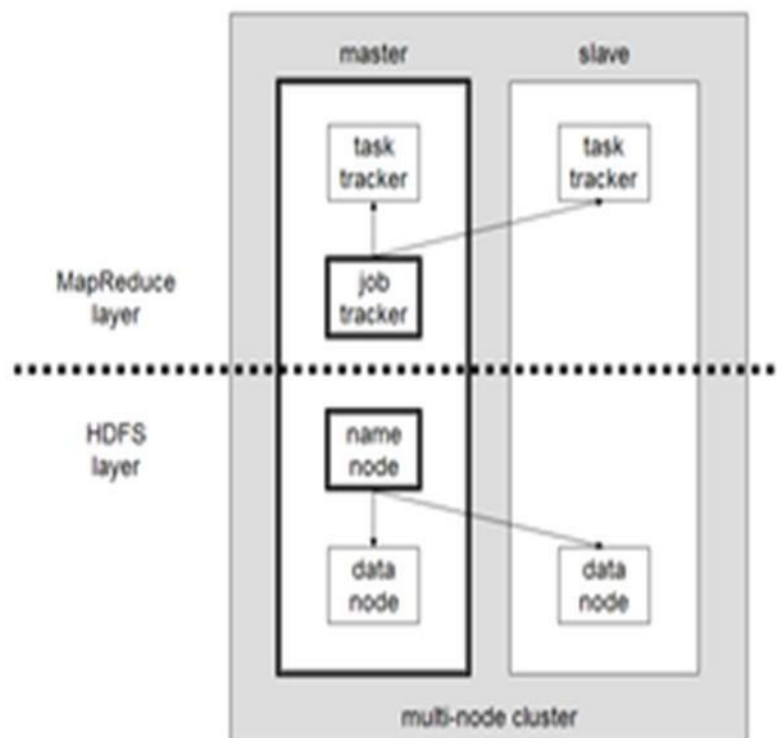


Fig3.1MultiNodeCluster

A small Hadoop cluster includes a single master and multiple worker nodes. The master node consists of a Job Tracker, Task Tracker, Name Node, and DataNode. A slave acts as both a Data Node and Task Tracker, though it is possible to have data-only and compute-only worker nodes. These are normally used only in nonstandard applications.

Hadoop requires Java Runtime Environment (JRE) 1.6 or higher. The standard startup and shutdown scripts require that Secure Shell (SSH) be set up between nodes in the cluster.

In a larger cluster, HDFS nodes are managed through a dedicated NameNode server to host the file system index, and a secondary Name Node that can generate snapshots of the name node memory structures, thereby preventing file-system corruption and loss of data. Similarly, a standalone Job Tracker

server can manage job scheduling across nodes. When Hadoop MapReduce is used with an alternate file system, the Name Node, secondary Name Node, and DataNode architecture of HDFS are replaced by the file-system-specific equivalents.

Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware. HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput. Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines.

Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop performs.

- Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M).
- These files are then distributed across various cluster nodes for further processing.
- HDFS, being on top of the local file system, supervises the processing.
- Blocks are replicated for handling hardware failure.

- Checking that the code was executed successfully.
- Performing the sort that takes place between the map and reduce stages.
- Sending the sorted data to a certain computer.
- Writing the debugging logs for each job.

3.4.1 Advantages

- No data loss problem.
- Less time consumption.
- Accuracy is improved.
- Efficient data processing.
- Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatically distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
- Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

CHAPTER 4 METHODOLOGY

Analyzing clinical data by using Hadoop framework along with some Hadoop ecosystems like hdfs, MapReduce, Sqoop, hive and pig. By using these tools we can process huge amount of data, no data lost problem, we can get high throughput, maintenance cost also very less and it is an open source software it is compatible on all the platforms since it is Java based.

4.1 MODULES

- 1) Pre-processing Clinical Database
- 2) Storage
- 3) Analyse Query
- 4) Analysis Latin Script (Pig)
- 5) Processing (MapReduce)

4.1.1 Pre-processing Clinical Database

In this module, Analyzing the data with different kinds of fields in Microsoft Excel then it is converted into comma delimited format which is said to be csv (comma separator value) file and moved to MySQL backup through Database.

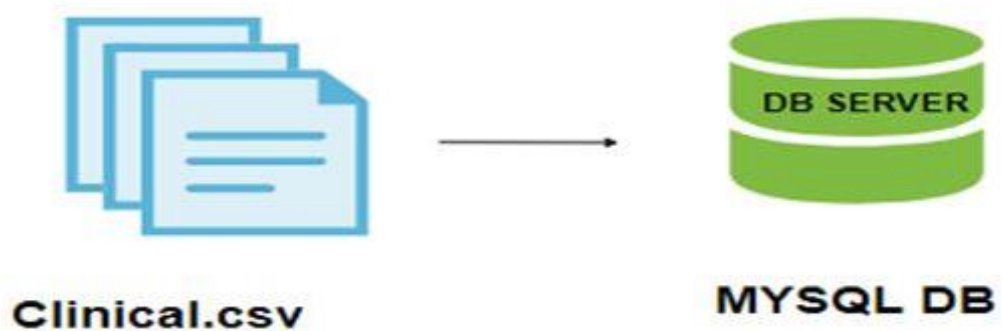


Fig4.1 Pre-processing.

Here by getting historical data we have to convert those historical batch processing data from (.xlsc) format to (.csv) format and by taking backup of all those data in MYSQL Database to avoid loss of data.

Pre-processing is a very important step in data analysis and can heavily influence the modelling results. In the following, we will show the main problems associated with clinical databases. According to Brause et al (2001), a large amount

untofmissingdataandunevensamplingtimesaretypicalformedicaldata

and should be taken into account in all approaches for medical data diagnosis. To get a good clean data from a medical database, a systematic pre-processing procedure is proposed and is described in more detail.

- *Define Features:* Through a systematic approach, identifying all clinical concepts contained in EHR data and defining features conveying each concept, including its type (numerical or categorical) and recording mechanism to determine feature value; features must account for different data.
- *Process Data:* Manipulating the feature set to improve homogeneity and avoid data dispersion by mitigating redundancy (concepts represented with different designations) and granularity (clinical concept is expressed with different levels of detail), which are tackled by combining different features referring to same clinical concept into a single feature.
- *Assess feature values:* Determining the value of each clinical feature (variable) for each dataset instance, by querying the extracted database entries according to the feature types and recording mechanisms.

4.1.2 STORAGE

In this module we are getting all those backup data which we have stored in MySQL and importing all those data by use of Sqoop command to HDFS (Hadoop Distributed File System). Now all the data are stored in HDFS where it is ready to get processed by use of Hive. By storing the huge stock data into HDFS, the solution provided is much more robust, reliable, economical, and scalable. Whenever data size is increasing, you can just add some more nodes, configure into Hadoop and that's all. If sometime any node is down, then even other nodes are ready to handle the responsibility due to data replication. By managing the Hive schema into embedded database or any other standard SQL database, we are able to utilize the power of SQL as well.

Sqoop (SQL-to-Hadoop) is a big data tool that offers the capability to extract data from non-Hadoop data stores, transform the data into a form usable by Hadoop, and then load the data into HDFS. This process is called ETL, for Extract, Transform, and Load.

While getting data into Hadoop is critical for processing using MapReduce, it is also critical to get data out of Hadoop and into an external data source for use in other kinds of application. Sqoop is able to do this as well.

While it is sometimes necessary to move the data in real time, it is most often necessary to load or unload data in bulk. Like Pig, Sqoop is a command-line interpreter. You type Sqoop commands into the interpreter and they are executed one at a time. Four key features are found in Sqoop.

- *Bulk Import:* Sqoop can import individual tables or entire databases into HDFS. The data is stored in the native directories and files in the HDFS filesystem.
- *Direct Input:* Sqoop can import and map SQL (relational) databases directly into Hive and HBase.
- *Data Interaction:* Sqoop can generate Java classes so that you can interact with the data programmatically.
- *Data Export:* Sqoop can export data directly from HDFS into a relational database using a target table definition based on the specifics of the target database.

Sqoop works by looking at the database you want to import and selecting an appropriate import function for the source data. After it recognizes the input, it then reads the metadata for the table (or database) and creates a class definition of your input requirements.

You can force Sqoop to be very selective so that you get just the columns you are looking for before input rather than doing an entire input and then looking for your data. This can save considerable time. The actual import from the external database to HDFS is performed by a MapReduce job created behind the scenes by Sqoop.

Sqoop is an effective tool for non-programmers. The other important item to note is the reliance on underlying technologies like HDFS and MapReduce. You see this repeatedly throughout the elements of the Hadoop ecosystem.

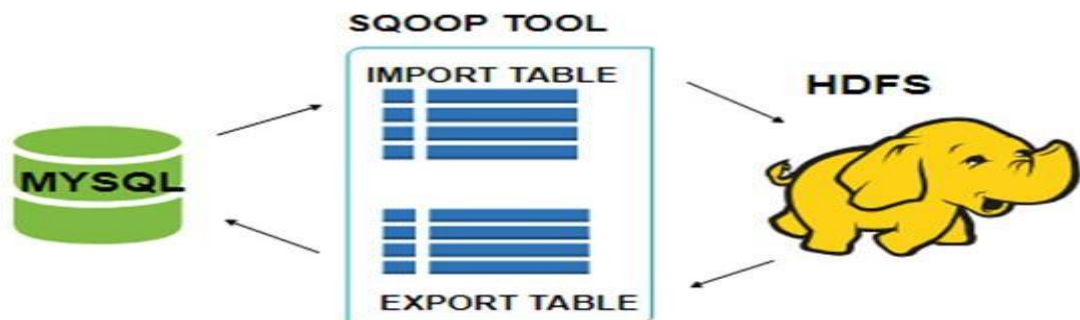


Fig4.2Storage.

All the existing Database Management Systems are designed with SQL standard in mind. However, each DBMS differs with respect to dialect to some extent. So, this difference poses challenges when it comes to data transfer across the systems. Sqoop Connectors are components which help overcome these challenges.

Data transfer between Sqoop and external storage system is made possible with the help of Sqoop's connectors.

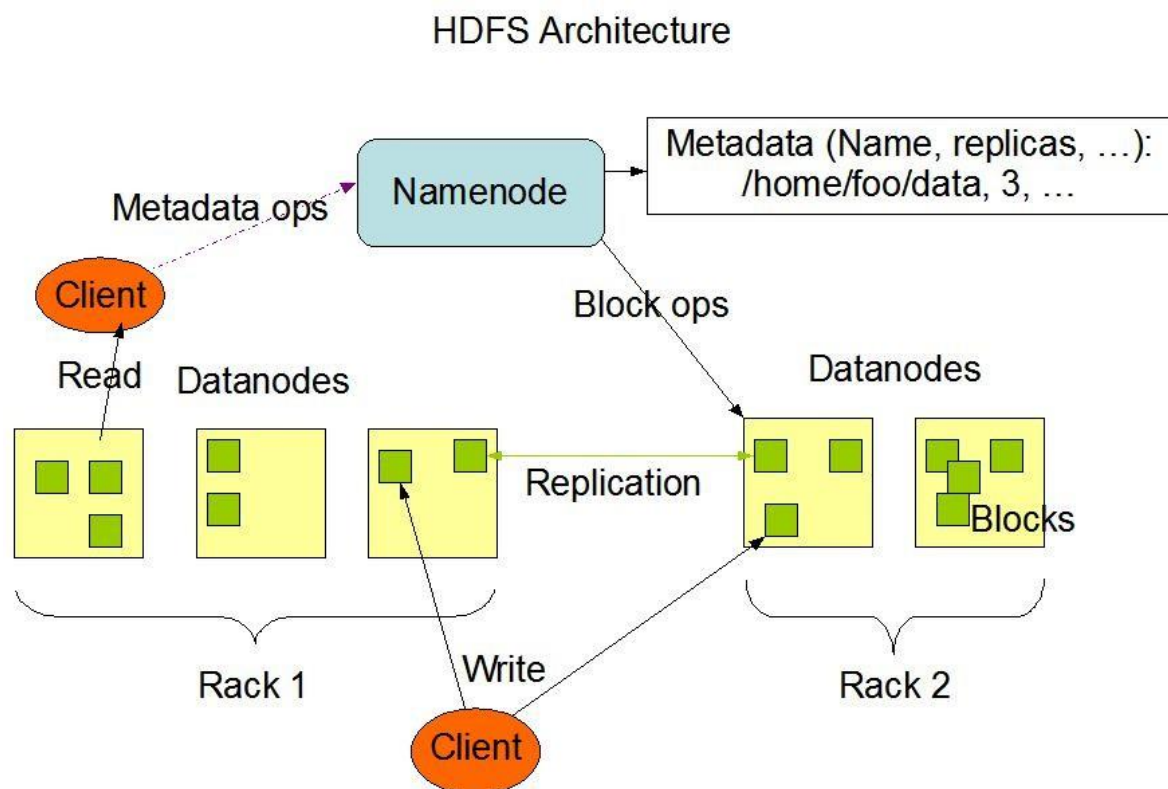


Fig4.3 HDFS Architecture

Sqoop has connectors for working with a range of popular relational databases, including MySQL, PostgreSQL, Oracle, SQL Server, and DB2. Each of these connectors knows how to interact with its associated DBMS. There is also a generic JDBC connector for connecting to any database that supports Java's JDBC protocol. In addition, Sqoop provides optimized MySQL and PostgreSQL connectors that use database-specific APIs to perform bulk transfer efficiently.

Analytical processing using Hadoop requires loading of huge amounts of data from diverse sources into Hadoop clusters. This process of bulk data

load into Hadoop, from heterogeneous sources and then processing it, comes with a

certain set of challenges. Maintaining and ensuring data consistency and ensuring efficient utilization of resources, are some factors to consider before selecting the right approach for data load.

- *Data Load Using Scripts:* The traditional approach of using scripts to load data is not suitable for bulk data load into Hadoop; this approach is inefficient and very time-consuming.
- *Direct Access To External Data Via Map-Reduce Application:* Providing direct access to the data residing at external systems (without loading into Hadoop) for map-reduce applications complicates these applications. So, this approach is not feasible. In addition to having the ability to work with enormous data, Hadoop can work with data in several different forms. So, to load such heterogeneous data into Hadoop, different tools have been developed.

4.1.3 ANALYZE QUERY

In this module we are getting all those data from HDFS to HIVE by use of Sqoop import command where HIVE is ready to analyze. Here in HIVE we can process only structured data to analyze. By extracting only the meaningful data and neglecting unclenched data we can analyze the data in more effective manner by use of HIVE. The queries are executed by MapReduce, Apache Tez or Apache Spark. HIVE queries run on MapReduce take long to run because of batch processing. Spark provides a way to run low latency queries. Spark provides better performance than MapReduce without requiring any changes in queries. HIVE is able to access data stored in HDFS. HIVE is a data warehouse system for data summarization and analysis and for querying of large data systems in the open-source Hadoop platform. It converts SQL-like queries into MapReduce jobs for easy execution and processing of extremely large volumes of data.



Fig4.4Analysing.

In this module we are getting those data from HDFS to HIVE by use of Sqoop import request. Where hive is set up to separate. Here in HIVE we can process in a general sense regulated data to look at. By restricting only, the gigantic data and blamable unclenched data we can separate the data in a magnificently reasonable course by utilization of hive to process the information that begins from the mapper. In the wake of setting it up, passes on another system of yield, which will be affirmed. The three important functionalities for which Hive is deployed are data summarization, data analysis, and data query. The query language, exclusively supported by Hive, is HiveQL. This language translates SQL-like queries into MapReduce jobs for deploying them on Hadoop. HiveQL also supports MapReduce scripts that can be plugged into the queries. Hive increases schema design flexibility and also data serialization and deserialization.

Hive is best suited for batch jobs, rather than working with web log data and append-only data. It cannot work for online transaction processing (OLTP) systems since it does not provide real-time querying for row-level update. Hive services such as Meta store, Filesystem, and Job Client in turn communicate with Hive storage and perform the following actions.

- Metadata information of tables created in Hive is stored in Hive "Metastore database".
- Query results and data loaded in the tables are going to be stored in Hadoop cluster or nHDFS.

4.1.3.1 Job Execution Flow

The data flow in Hive behaves in the following pattern.

- Executing Query from the UI (User Interface).
- The driver is interacting with Compiler for getting the plan. (Here plan refers to query execution) process and its related metadata information gathering.
- The compiler creates the plan for a job to be executed. Compiler communicating with Metastore for getting metadata request.
- Metastore sends metadata information back to compiler.

- Compiler communicating with Driver with the proposed plan to execute the query.
- Driver Sending execution plan to Execution engine.
- Execution Engine (EE) acts as a bridge between Hive and Hadoop to process the query. For DFS operations.
- EE should first contact Name Node and then to Data nodes to get the values stored in tables.
- EE is going to fetch desired records from Data Nodes. The actual data of tables resides in data node only. While from Name Node it only fetches the metadata information for the query.
- It collects actual data from data nodes related to mentioned query
- Execution Engine (EE) communicates bi-directionally with Metastore present in Hive to perform DDL (Data Definition Language) operations. Here DDL operations like CREATE, DROP and ALTERING tables and databases are done. Metastore will store information about database name, table names and column names only. It will fetch data related to query mentioned.
- Execution Engine (EE) in turn communicates with Hadoop daemon such as Name node, Data nodes, and job tracker to execute the query on top of Hadoop file system.

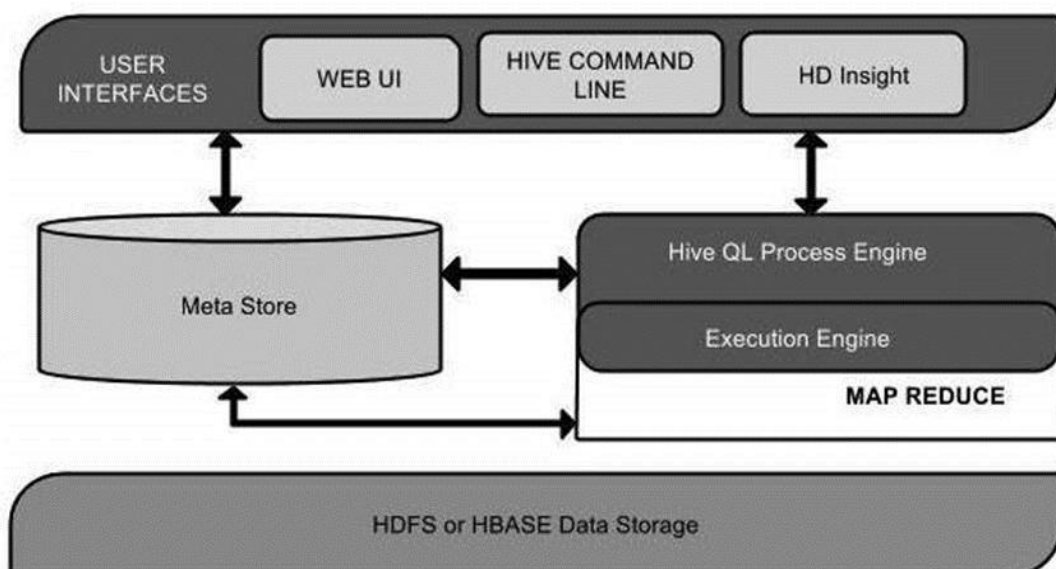


Fig4.5HIVE.

4.1.3.2 *Different Modes Of Hive*

Hive can operate in two modes depending on the size of data nodes in Hadoop. These modes are Local mode and Map reduce mode

When To Use Local Mode

- If the Hadoop is installed under pseudo mode with having one data node we use Hive in this mode
- If the data size is smaller in terms of limited to single local machine, we can use this mode
- Processing will be very fast on smaller datasets present in the local machine

When To Use Map Reduce Mode

- If Hadoop is having multiple data nodes and data is distributed across different nodes we use Hive in this mode.
- It will perform on large amount of datasets and query going to execute in parallel way.
- Processing of large datasets with better performance can be achieved through this mode

In Hive, we can set this property to mention which mode Hive can work? By default, it works on Map Reduce mode and for local mode you can have the following setting. From the Hive version 0.7 it supports a mode to run map reduce jobs in local mode automatically data which is available in hdfs while processing the data first internally it will going to convert into the (key, value) pair that input will goes to mapper then sorting and shuffling of data will happened and this intermediate data will going to pass to the reducer and internally combiner will combine the data of key and value data and finally it will going to pass to the hdfs for storage purpose.

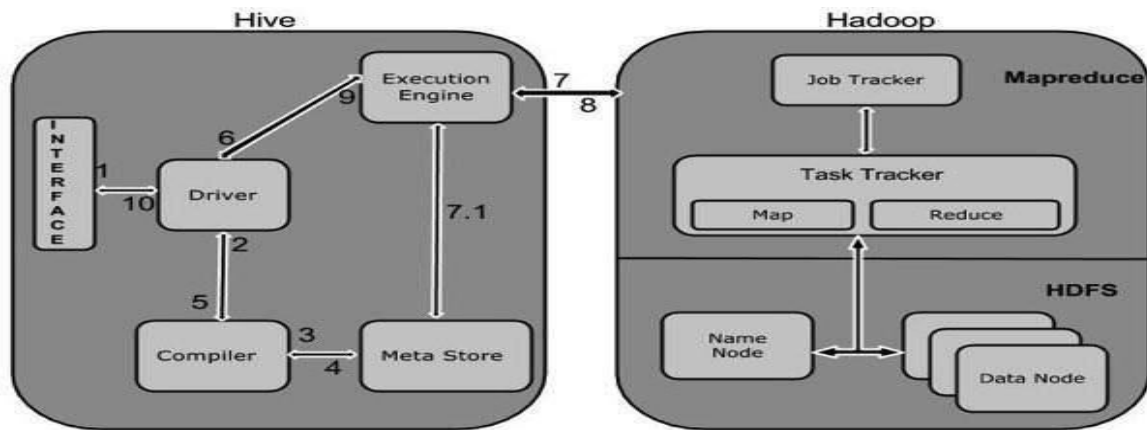


Fig4.6MapReduce.

- It stores schema in a database and processed data into HDFS.
- It is designed for OLAP.
- It provides SQL type language for querying called HiveQL or HQL.
- It is familiar, fast, scalable, and extensible.

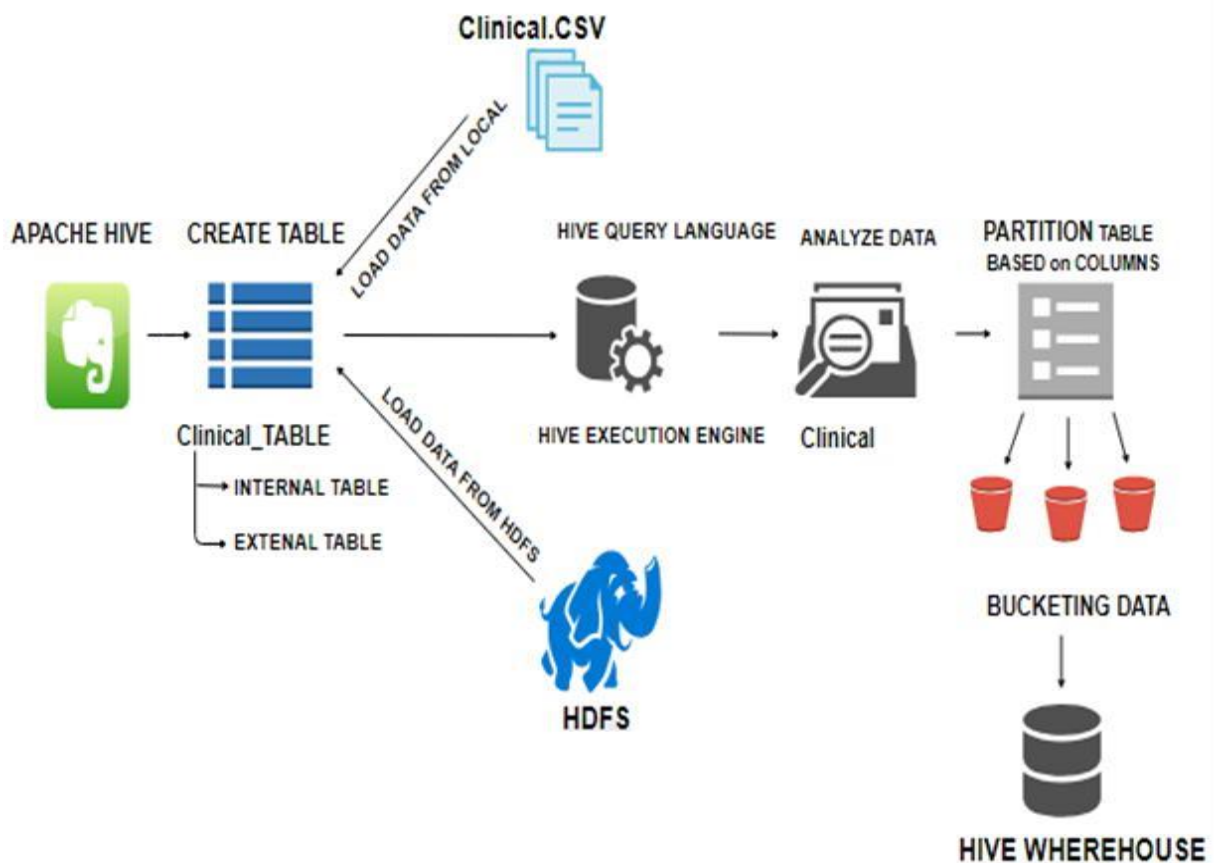


Fig4.7AnalyzingQuery

4.1.4 ANALYSIS IN SCRIPT (PIG)

To analyze Clinical using Pig, programmers need to write scripts using Pig Latin language and execute them in interactive mode using the Grunt shell. All these scripts are internally converted to Map and Reduce tasks. After invoking the Grunt shell, you can run your Pig scripts in the shell. Except LOAD and STORE, while performing all other operations, Pig Latin statements take a relation as input and produce another relation as output. As soon as you enter a Load statement

in the Grunt shell, its semantic checking will be carried out. To see the contents of the schema, you need to use the Dump operator. Only after performing the dump operation, the Map Reduce job for loading the data into the file system will be carried out. Pig provides many built-in operators to support data operations like grouping, filters, ordering, etc.

The major benefit of PIG is that it works with data that are obtained from various sources and store the results into HDFS (Hadoop Data File System). The programmers have to write the scripts in Pig Latin language which are then converted into Map and reduce tasks with the Pig Engine component (Apache Pig has a component called Pig Engine. It usually accepts the Pig Latin scripts to convert them into Map Reduce jobs). Generally, the Apache Pig gives an abstraction to reduce the complexity of developing MapReduce Programming for the developers. The common reason for using the Pig is that it gives hand to write short programs.

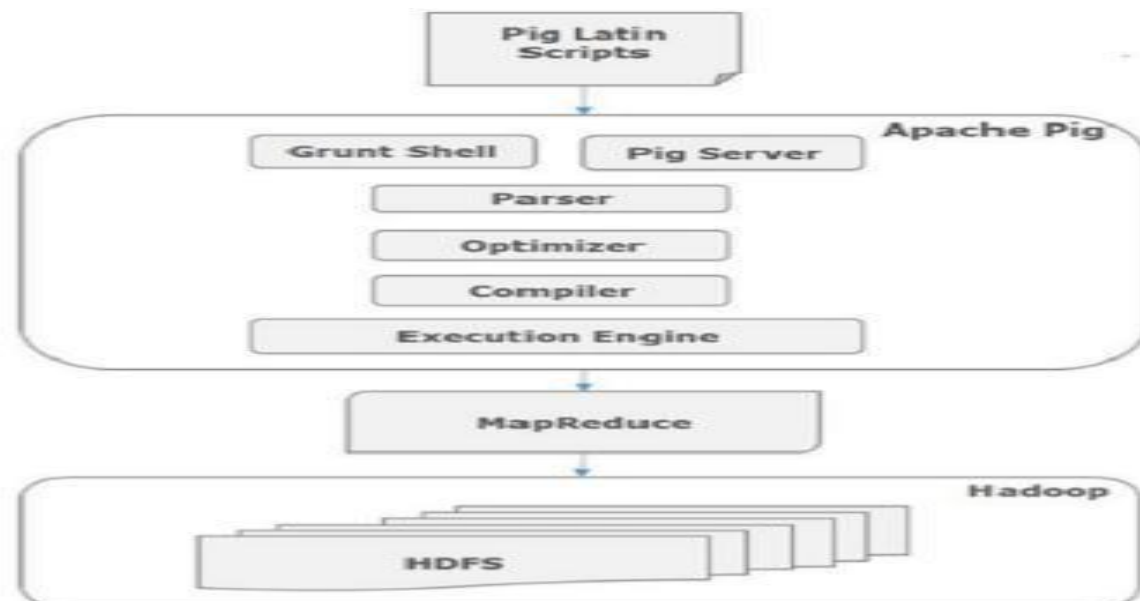


Fig4.8Pig

The advanced features of Apache Pig enable the programmers to do more work than other frameworks. This also eases the life of a data engineer in maintaining various ad hoc queries on the data sets. In fact, Apache Pig is a boon for all the programmers and so it is most recommended to use in data management.

The Pig Latin Compiler is defined as the runtime environment which converts the Pig source code into executable code. Generally, most of the executable code exists in the form of MapReduce form. First of all, the programmers have to write Pig scripts and analyze them. These pig scripts are processed with the help of Apache Pig components such as a parser, optimizer, compiler and finally to the execution engine. Now, you will be able to get the executable code of Pig Latin which is to be converted into MapReduce tasks.

The MapReduce tasks are then stored into the Hadoop Distributed File System (HDFS).

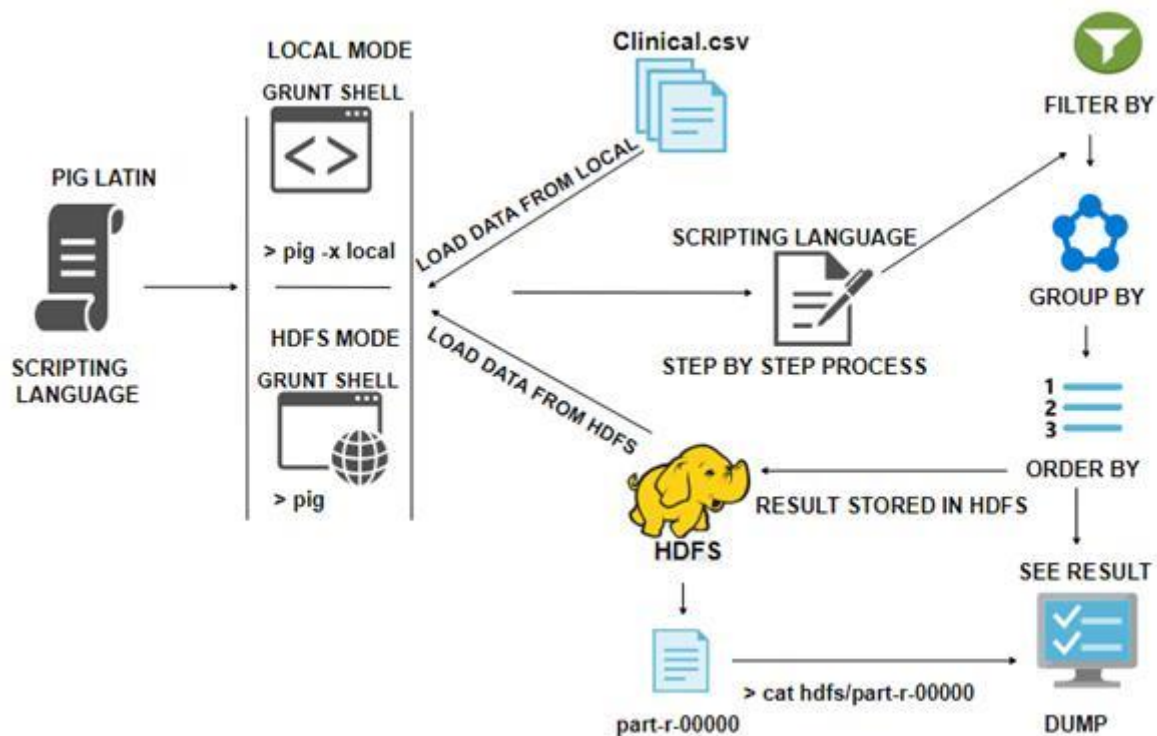


Fig4.9 Analysis Of PIG

4.1.5 PROCESSING(MAPREDUCE):

MapReduce is a framework using which we can write applications to process huge amounts of Clinical, in parallel, on large clusters of commodity hardware in a reliable manner. Map Reduce is a processing technique and a program model for distributed computing based on java. The MapReduce

algorithm contains two important tasks, namely Map and Reduce. Map Reduce program executes in three stages, namely map stage, shuffle stage, and reduce stage. The map or mapper's job is to process the input data. Generally, the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data. This stage is the combination of the Shuffle stage and the Reduce stage.

The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS. Since Hive schema is created on a standard SQL database, you get the advantage of running SQL queries on the huge dataset also and are able to process GBs or TBs of data with simple SQL queries. Since actual data resides in HDFS, so these Hive SQL queries are being converted into MapReduce jobs and these parallelized map reduce jobs process these huge volumes of data and achieve scalable and fault-tolerant solutions.

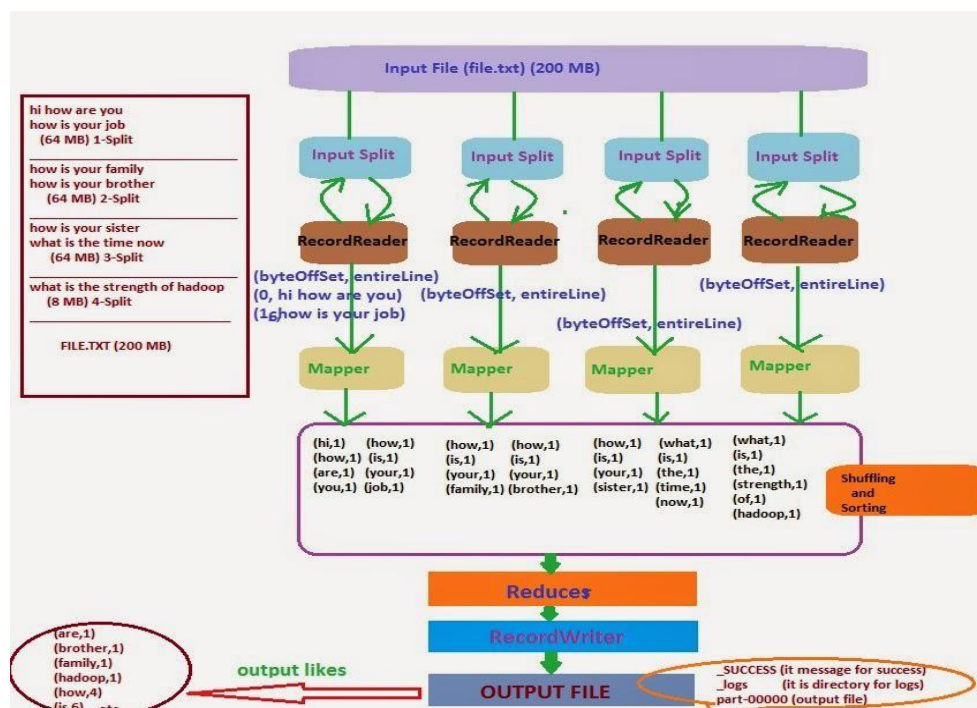


Fig4.10 MapReduce Process.

One map task is created for each split which then executes map function for each record in the split.

- It is always beneficial to have multiple splits because the time taken to process a split is small as compared to the time taken for processing of the

whole input. When the splits are smaller, the processing is better to load balanced since we are processing the splits in parallel.

- However, it is also not desirable to have splits too small in size. When splits are too small, the overload of managing the splits and map task creation begins to dominate the total job execution time.
- For most jobs, it is better to make a split size equal to the size of an HDFS block (which is 64MB, by default).
- Execution of map tasks results into writing output to a local disk on the respective node and not to HDFS.
- Reason for choosing local disk over HDFS is, to avoid replication which takes place in case of HDFS store operation.
- Map output is intermediate output which is processed by reduce tasks to produce the final output.
- Once the job is complete, the map output can be thrown away. So, storing it in HDFS with replication becomes overkill.
- In the event of node failure, before the map output is consumed by the reduce task, Hadoop reruns the map task on another node and re-creates the map output.
- Reduce task doesn't work on the concept of data locality. An output of every map task is fed to the reduce task. Map output is transferred to the machine where the reduce task is running.
- On this machine, the output is merged and then passed to the user-defined reduce function.
- Unlike the map output, reduce output is stored in HDFS (the first replica is stored on the local node and other replicas are stored on off-rack nodes). So, writing the reduce output.

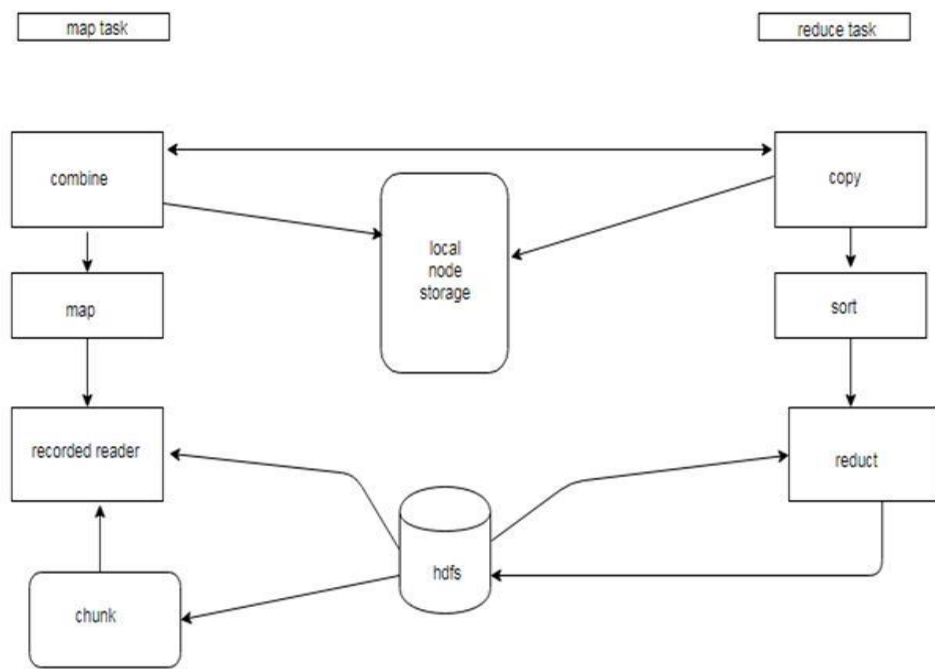


Fig4.11MapReduceArchitecture

4.2 ALGORITHMS

Introduces the important concept of local aggregation in Map Reduce and strategies for designing efficient algorithms that minimize the amount of partial results that need to be copied across the network. The proper use of combiners is discussed in detail, as well as the “in-mapper combining” design pattern.

Generally MapReduce paradigm is based on sending the computer to where the data resides. MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

- **Map Stage:** The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
- **Reduce Stage:** This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

4.2.1 WordCount

The mapper emits an intermediate key-value pair for each word in a document. The reducer sums up all counts for each word.

```

1: class Mapper
2:   method Map (doc id a, doc
   d)
3:   forall term t ∈ doc d do
4:     Emit(term t, count1)
1: class Reducer
2:   method Reduce (term t, counts [c1, c2, . .
   .])
3:   sum ← 0
4:   for all count c ∈ counts [c1, c2, . . .]
   do
5:     sum ← sum + c
6:   Emit(term t, countsum)

```

4.2.2

```

1: class Mapper
2:   method Map (doc id a, doc
   d)
3:   H ← new AssociativeArray
4:   forall term t ∈ doc d do
5:     H{t} ← H{t} + 1. Tally counts for entire
   document
6:   forall term t ∈ H do
7:     Emit(term t, countH{t})

```

4.2.3

```

1: class Mapper
2:   method Initialize
3:   H ← new AssociativeArray
4:   method Map (doc id a, doc
   d)
5:   forall term t ∈ doc d do
6:     H{t} ← H{t} + 1. Tally counts across
   documents
7:   method Close
8:   forall term t ∈ H do
9:     Emit(term t, countH{t})

```

4.3 REQUIREMENT ANALYSIS

Requirement analysis determines the requirements of a new system. This project analyses on product and resource requirement, which is required for this successful system. The product requirement includes input and output requirements it gives the wants in terms of input to produce the required output.

The resource requirements give in brief about the software and hardware that are needed to achieve the required functionality.

4.3.1 Hardware Environment

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design. It shows what the systems do and not how it should be implemented.

- Harddisk : 40GB
- Monitor : 15' colour with VGA card support
- Ram : 4GB DDRam
- Processor : Pentium IV and above (or) equivalent
- Processor speed : Minimum 500MHZ

4.3.2 Software Environment

The software requirements are the specification of the system. It should include both a definition and a specification of requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the team's and tracking the team's progress throughout the development activity.

- Operating system : VMWare
- Frontend : MySQL
- Backend : Hadoop

CHAPTER 5 RESULT

To analyze no limitation of data and simply add number of machines to the cluster and we get results with less time, high throughput and maintenance cost is very less and we are using joins, partitions and bucketing techniques in Hadoop.

Big Data are radically changing research. The unprecedented advances in automated collection of large-scale molecular and clinical data pose major challenges to data analysis and interpretation, calling for the development of new computational approaches. The creation of powerful systems for the effective use of Big Data in Personalized will require significant scientific and technical developments, including in infrastructure, engineering, project and financial management. We review here how the evolution of data-driven methods offers the possibility to address many of these problems, guiding the formulation of hypotheses on systems functioning and the generation of mechanistic models, and facilitating the design of clinical procedures.

OUTPUT

```
hive> select HOSPITAL_NAM,CIT,PATIENT_NAM,count(PROBLE),PROBLE FROM hos1
where PROBLE = 'FEVER' GROUP BY HOSPITAL_NAM,CIT,PATIENT_NAM,PROBLE LIMIT
20;
```

```
OK
hospital_nam      cit      patient_nam      _c3      proble
A.G.HOSPITAL      CHENNAI  VIKASNI 1        FEVER
A.G.HOSPITAL      CHENNAI  VISHETA 1        FEVER
AASHIANA HOSPITAL CHENNAI  SENDHEN 1        FEVER
AGARWALLS EYE INSTITUTE PRIVATE LIMITED CHENNAI SANITHI 1        FEVER
AGARWALLS EYE INSTITUTE PRIVATE LIMITED CHENNAI VIRANGANA 1
FEVER
AGARWALLS EYE INSTITUTE PRIVATE LIMITED CHENNAI YASHNA 1        FEVER
AMAR HOSPITAL      CHENNAI  ADHIYAMAAN 1        FEVER
AMAR HOSPITAL      CHENNAI  KRISHNA-DWAIPAYANA 1        FEVER
AMAR HOSPITAL      CHENNAI  KUPPURAM 1        FEVER
AMAR HOSPITAL      CHENNAI  NALLAMUTHU 1        FEVER
AMAR HOSPITAL      CHENNAI  NANJAIYA 1        FEVER
AMAR HOSPITAL      CHENNAI  SAMPREETI 1        FEVER
AMAR HOSPITAL      CHENNAI  SEVANDHI 1        FEVER
AMAR HOSPITAL      CHENNAI  VRITTI 1        FEVER
APOLLO CANCER HOSPITAL CHENNAI  ACHYUTAM 1        FEVER
APOLLO CANCER HOSPITAL CHENNAI  ADHIYAMAAN 1        FEVER
APOLLO CANCER HOSPITAL CHENNAI  NAMBI 1        FEVER
APOLLO CANCER HOSPITAL CHENNAI  PARUSHI 1        FEVER
APOLLO CANCER HOSPITAL CHENNAI  RIDHUSHNI 1        FEVER
APOLLO CANCER HOSPITAL CHENNAI  SANJUSHREE 1        FEVER
Time taken: 11.48 seconds
```

```
hive> select * from hos1 where PROBLE = 'MELANOMA' LIMIT 5;
```

```
Ended Job = job_201809242128_0002
```

```
OK
```

```
1011 APOLLO MEDICAL CENTRE CHENNAI 2/3/2016 3266 ADITHAN 49 M
MELANOMA 97 98 84 126 CHECKS THE PATIENTS BICARBONATE (CARBON
DIOXIDE) SENT TO DOCTOR TREATS PATIENT BY SEEING REPORTS DESCRIBED TO
PATIENTS 15821

1073 JAYADEV POLYCLINIC HOSPITAL CHENNAI 6/25/2015 5641 TOKKA 54 F
MELANOMA 99 108 93 130 CHECKS THE PATIENTS BICARBONATE (CARBON
DIOXIDE) SENT TO DOCTOR TREATS PATIENT BY SEEING REPORTS DESCRIBED TO
PATIENTS 6494

1083 LAKSHMI MATERNITY HOSPITAL CHENNAI 6/26/2016 4654 VIKASNI 20 F
MELANOMA 99 116 95 140 CHECKS THE PATIENTS BICARBONATE (CARBON
DIOXIDE) SENT TO DOCTOR TREATS PATIENT BY SEEING REPORTS DESCRIBED TO
PATIENTS 2832

1059 HANDE HOSPITAL CHENNAI 1/12/2017 4978 ROSHITA 91 F MELANOMA
97 97 93 117 CHECKS THE PATIENTS BICARBONATE (CARBON DIOXIDE) SENT
TO DOCTOR TREATS PATIENT BY SEEING REPORTS DESCRIBED TO PATIENTS 20104

1027 CHINNAMMAL ENT HOSPITAL CHENNAI 11/21/2015 2315 NALLAPERUMAL 34
M MELANOMA 97 106 85 142 CHECKS THE PATIENTS BICARBONATE
(CARBON DIOXIDE) SENT TO DOCTOR TREATS PATIENT BY SEEING REPORTS DESCRIBED
TO PATIENTS 21552
```

```
*****
```

```
hive> SELECT HOSPITAL_NAM,CIT,PATIENT_NAM,TEMPERATURE_MI FROM hos1 where
TEMPERATURE_MI >= 90 or TEMPERATURE_MI < 90 limit 5;
```

```
Ended Job = job_201809242222_0041
```

```
OK
```

```
APOLLO SPECIALITY HOSPITALS CHENNAI ABHINAVAN 97
ASIAN HOSPITAL CHENNAI NALLANBAN 99
A.G.HOSPITAL CHENNAI ABHISEKA 97
A.G.M.HOSPITAL CHENNAI ABIJITH 98
A.V.M.MEDICALS ENT RESEARCH FOUNDATION PVT LT CHENNAI ACHYUTAM 97

Time taken: 3.781 seconds
```

CHAPTER

6 CONCLUSION AND FUTURE WORK

CONCLUSION

This model presented a study on Clinicalis to help the doctors by getting the idea of what treatment has been given to that particular patient in the past. A patient can get an idea to see which hospital is treating better for particular problems.

Analyze the Clinical data using Hadoop ecosystem where Hive and Pig used for analysis and MapReduce for processing.

FUTURE ENHANCEMENT

Apache Spark is an open source processing engine built around speed, ease of use, and analytics. If you have large amounts of data that requires low latency processing that a typical MapReduce program cannot provide, Spark is the alternative. Spark provides in-memory cluster computing for lightning fast speed and supports Java, Scala, and Python APIs for ease of development.

REFERENCES

- [1] Bhatti U A, Huang M, Zhang Y, et al. Research on the Smartphone Based Health Systems for Strengthening Healthcare Organization [M]//Smart Health. 2017.
- [2] Black A D, Car J, Pagliari C, et al. "The impact of eHealth on the quality and safety of health care: a systematic overview", *PLoS Medicine*, vol. 8, no. 1, 2011.
- [3] Cao P, Xian Medical University. "The Impact of Big Data on the Development of Chinese Health Service Industry", *Journal of Electronic Test*, 2014
- [4] Daniel Jarrett, Jinsung Yoon, Member, IEEE, and Mihaela van der Schaar, Fellow, IEEE "Dynamic Prediction in Clinical Survival Analysis using Temporal Convolutional Networks".
- [5] Hermann M, Pentek T, Otto B. Design Principles for Industrie 4.0 Scenarios, pp. 3928-3937, 2016.
- [6] Horsky J, Schiff GD, Johnston D, et al. "Methodological Review: Interface design principles for usable decision support: A targeted review of best practices for clinical prescribing interventions", *Journal of Biomedical Informatics*, vol. 45, no. 6, pp. 1202-1216, 2012.
- [7] L. Lakshmanan, k. Vijaykumar, Parthasarathy, S. Ramani, "Cluster based distributed architecture for prediction of students performance in higher education" *Cluster Computing*, vol. 16, pp. 5-7, 2019.
- [8] M. Akay, D. I. Fotiadis, K. S. Nikita and R. W. Williams, "Guest Editorial: Biomedical Informatics in Clinical Environments," *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 1, pp. 149-150, 2015.
- [9] Matthew R. Boutella, Jiebo Luob "Learning multi-label scene classification", 2015.
- [10] Mark H. Drazner "The Progression of Hypertensive Heart Disease", 2016.
- [11] Thuemmler C, Bai C. Health 4.0: How Virtualization and Big Data are Revolutionizing Healthcare, Springer International Publishing, 2017.
- [12] Vancampfort D, Mugisha J, Hallgren M, et al. "The prevalence of diabetes mellitus type 2 in people with alcohol use disorders: a systematic review and large-scale meta-analysis", *Psychiatry Research*, vol. 246, pp. 394-400, 2016.
- [13] Yu W D, Prakash C, Swati S, et al. "A Modelling Approach to Big Data Based Recommendation Engine in Modern Health Care Environment", *Computer Software and Applications Conference, IEEE Computer Society*, pp. 75-86, 2015.

APPENDIX

A. SAMPLECODE

MySQL>select*fromtreat limit5;

```

-----
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
|HOSPITAL_I|HOSPITAL_NAM          |CIT   |DAT
|PATIENT_I|PATIENT_NA
M|AG|GENDE|PROBLE                |TEMPERATURE_MI
|TEMPERATURE_MA|BP_MIN|BP_MA|INTERACT_WITH_DOCTO|
LAB_TES          |REPORT    |TREATMEN          |
PRESCRIPTIO      |BIL|
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
|   1000|APOLLOSPECIALITYHOSPITALS          | CHENNAI
|10/29/2015|1243   |ABHINAVAN  |26| M    |FEVER    |   99|
100|   87|100|CHECKSTHEPATIENTS|ACTHSUPPRESSION
| SENT TO DOCTOR | TREATS PATIENT BY SEEING REPORTS |
DESCRIBEDTOPATIENTS| 23004|
|   1001|ASIANHOSPITAL          |CHENNAI|12/31/2017|1565
|NALLANBAN|   15|M    |BACKPAIN  |   100|   96|   94|  108
|CHECKS THE PATIENTS | ADRENOCORTICOTROPIC HORMONE (ACTH)
|SENT TO DOCTOR | TREATS PATIENT BY SEEING REPORTS |
DESCRIBEDTOPATIENTS| 4059|
|   1002|A.G.HOSPITAL          |CHENNAI|11/27/2015|5564
|ABHISEKA  |45| M    |NEURALPAIN|   97|   109|   91|  127
|CHECKS THE PATIENTS | ALANINE AMINOTRANSFERASE (ALT) | SENT
TODOCTOR | TREATS PATIENT BY SEEING REPORTS | DESCRIBED
TOPATIENTS| 18508|

```

```
| 1003|A.G.M.HOSPITAL |CHENNAI|1/19/2015|1147
|ABIJITH | 68|M |APPENDIX | 97| 96| 94 |136
|CHECKSTHEPATIENTS|ALBUMIN | SENT TO DOCTOR
|TREATSPATIENTBYSEEINGREPORTS|DESCRIBEDTOPATIENTS|20115
```

```
| 1004 | A.V.M. MEDICALS ENT RESEARCH FOUNDATION PVT LT
|CHENNAI|4/11/2017|2013 |ACHYUTAM |10| M |CANCER
|97| 107| 80| 120|CHECKSTHEPATIENTS|ALKALINE
PHOSPHATASE | SENT TO DOCTOR | TREATS PATIENT BY
SEEINGREPORTS|DESCRIBED TOPATIENTS|22761|
```

```
+ + + + +
+ + + + +
+ + + + +
+ + + + +
```

5rowsinset(0.01sec)

```
=====
=====
=====
```

SQOOP:-

```
[traininglocalhost~] $Sqoop import --connect jdbc: mysql://localhost/movie lens --
username training--password training--table treat--m1--target-dir.clinic;
```

```
-----
-----
```

Eval:-

```
[traininglocalhost~] $Sqoopeval --connect jdbc: mysql://localhost/movielens --
username training--password training--query"select*fromtreatlimit5"
```

```
+++++
+++++
+
```

```

+_____+_____+_____+_____+_____+_____
|_+___+___+___+___+___+___+___+___+___+___+___+___
|_+_____+_____+_____+_____+_____+_____+_____
|_____+___+___+

```

```

| 1000|APOLLOSPECIALITYHOSPITALS | CHENNAI
|10/29/2015|1243 |ABHINAVAN |26| M |FEVER | 99|
100| 87|100|CHECKSTHEPATIENTS|ACTHSUPPRESSION
| SENT TO DOCTOR | TREATS PATIENT BY SEEING REPORTS |
DESCRIBEDTOPATIENTS| 23004|
| 1001|ASIANHOSPITAL |CHENNAI|12/31/2017|1565
|NALLANBAN| 15|M |BACKPAIN | 100| 96| 94| 108
|CHECKS THE PATIENTS | ADRENOCORTICOTROPIC HORMONE (ACTH)
|SENT TO DOCTOR | TREATS PATIENT BY SEEING REPORTS |
DESCRIBEDTOPATIENTS| 4059|
| 1002|A.G.HOSPITAL |CHENNAI|11/27/2015|5564
|ABHISEKA |45| M |NEURALPAIN| 97| 109| 91| 127
|CHECKS THE PATIENTS | ALANINE AMINOTRANSFERASE (ALT) | SENT
TODOCTOR | TREATS PATIENT BY SEEING REPORTS | DESCRIBED
TOPATIENTS| 18508|
| 1003|A.G.M.HOSPITAL |CHENNAI|1/19/2015|1147
|ABIJITH | 68|M |APPENDIX | 97| 96| 94 |136
|CHECKSTHEPATIENTS|ALBUMIN | SENT TO DOCTOR
|TREATSPATIENTBYSEEINGREPORTS|DESCRIBEDTOPATIENTS|20115
|
| 1004 | A.V.M. MEDICALS ENT RESEARCH FOUNDATION PVT LT
|CHENNAI|4/11/2017|2013 |ACHYUTAM |10| M |CANCER
|97| 107| 80| 120|CHECKSTHEPATIENTS|ALKALINE
PHOSPHATASE | SENT TO DOCTOR | TREATS PATIENT BY
SEEINGREPORTS|DESCRIBED TOPATIENTS|22761|

```

5rowsinset(0.01sec)

listingdatabases:

[traininglocalhost~] \$Sqoop list-databases --connect jdbc:

mysql://localhost/movielens--username training--password training

18/09/11 01:10:07 WARN tool.BaseSqoopTool: Setting your password on thecommand-lineis insecure. Considerusing-Pinstead.

18/09/11 01:10:07 INFO manager.MySQLManager: Preparing to use a MySQLstreamingresultset.

information

schemamovie lens

training

HIVE:-

[traininglocalhost~] \$

hiveHivehistory

file=/tmp/training/hive_job_log_training_201809242142_713934136.txt

hive>

hive> show

databases;OK

default

Timetaken:1.618seconds

=====

=====

=====

hive> create database

hospital;OK

Timetaken:0.239seconds

=====

=====

=====

hive>showdatabases;

OK

default

hospital

Timetaken:0.067seconds

```
=====
=====
=====
```

hive> use

hospital;OK

Timetaken:0.067seconds

```
=====
=====
=====
```

CreateHiveTable:-

```
hive> create table hos1(HOSPITAL_I int,HOSPITAL_NAMstring, ITstring,
ATstring,PATIENT_lint,PATIENT_NAMstring,AGstring,GENDEstring,PROBLEstri
ng,TEMPERATURE_Mlint,TEMPERATURE_MAint,BP_MINint,BP_MAint,INTER
ACT_WITH_DOCTOstring,LAB_TESstring,REPORTstring,TREATMENstring,PR
ESCRPTIOstring,BILint) row format delimited fieldsterminatedby',';
```

OK

Time taken: 0.334

secondsLoadingDataintoTheTa

ble:-

```
hive>loaddatainpath'/user/training/clin/part-m-00000'intotablehos1;
```

```
-----
```

Loading data to table

hospital.hos1OK

Time taken: 0.399

secondsHIVEANALYSIS:-

Howtoapply conditionsandseetheresult?

```
+++++
```

```
hive>select*fromhos1wherePROBLE='MELANOMA'LIMIT5;
```

OK

```
hospital    hospital_namecitdat    patient_i    patient_nam    ag
            gendeproble temperature mi
temperature_mabp_minbp_maint interact_with_docto lab_tesreporttreatmenpre
scriptio    bil
1011  APOLLOMEDICALCENTRECHENNAI12/24/2015    3266  ADITHAN
49    M    MELANOMA    100    104    97    131    CHECKS
THEPATIENTS    BICARBONATE (CARBON DIOXIDE) SENT
TO DOCTOR TREATSPATIENTBYSEEINGREPORTS
DESCRIBEDTOPATIENTS9454
1073  JAYADEVPOLYCLINICHOSPITAL    CHENNAI6/16/2016
5641TOKKA54    F    MELANOMA    98    116
100    102    CHECKS THEPATIENTS    BICARBONATE (CARBON
DIOXIDE) SENT TO DOCTOR TREATSPATIENTBYSEEINGREPORTS
DESCRIBEDTOPATIENTS9669
1083  LAKSHMIMATERNITYHOSPITAL    CHENNAI10/14/2015
4654VIKASNI20    F    MELANOMA    98    95    86
101    CHECKS THEPATIENTS BICARBONATE (CARBON
DIOXIDE)SENT TO DOCTOR TREATSPATIENTBYSEEINGREPORTS
DESCRIBED TOPATIENTS 241781059
HANDEHOSPITALCHENNAI3/21/2015    4978  ROSHITA91    F
MELANOMA    99    102    86    109    CHECKS THE
PATIENTSBICARBONATE (CARBON DIOXIDE)SENT TO DOCTOR TREATS
PATIENT BYSEEINGREPORTS    DESCRIBEDTOPATIENTS12559
1027  CHINNAMMALENTHOSPITALCHENNAI8/1/2017    2315
NALLAPERUMAL 34    M    MELANOMA    98    103    81    123
CHECKS THE PATIENTSBICARBONATE (CARBON DIOXIDE)SENT
TODOCTOR TREATS PATIENT BY SEEING REPORTSDESCRIBED
TOPATIENTS17890
```

Timetaken:4.452seconds

Howtoapply conditionsontheclinicaldata?

+++++

```
hive>selectCOUNT(*) fromhos1wherePROBLE='MELANOMA';
```

```
OK
```

792

Timetaken:14.559seconds

```
hive> select HOSPITAL_NAM, CIT, PATIENT_NAM, count (PROBLE),  
PROBLEFROM hos1 where PROBLE = 'FEVER' GROUP BY HOSPITAL_NAM,  
CIT,PATIENT_NAM,PROBLELIMIT 20;
```

OK

hospital cit patient _c3 proble

A.G.HOSPITAL CHENNAI VIKASNI1 FEVER

A.G.HOSPITAL CHENNAI VISHETA1

FEVERAASHIANAHOSPITAL

CHENNAISENDHEN1 FEVER

AGARWALLS EYE INSTITUTE PRIVATE LIMITED CHENNAI SANITHI
1FEVER

AGARWALLSEYEINSTITUTEPRIVATELIMITEDCHENNAIVIRANGANA

1

FEVER

AGARWALLS EYE INSTITUTE PRIVATE LIMITED CHENNAI

YASHNA1FEVER

AMARHOSPITAL CHENNAIADHIYAMAAN 1 FEVER

AMARHOSPITAL CHENNAIKRISHNA-DWAIPAYANA 1

FEVERAMARHOSPITAL CHENNAIKUPPURAM 1

FEVER

AMARHOSPITAL CHENNAINALLAMUTHU 1

FEVERAMARHOSPITAL

CHENNAINANJAIYA 1

FEVERAMARHOSPITAL

CHENNAISAMPREETI 1

FEVERAMARHOSPITAL

CHENNAISEVANDHI 1

FEVERAMARHOSPITAL CHENNAI

VRITTI1 FEVER

APOLLOCANCERHOSPITALCHENNAIACHYUTAM 1 FEVER

APOLLOCANCERHOSPITALCHENNAIADHIYAMAAN	1	FEVER
APOLLOCANCERHOSPITALCHENNAINAMBI1		
		FEVERAPOLL
OCANCERHOSPITALCHENNAIPARUSHI1		
		FEVERAPOLL
OCANCERHOSPITALCHENNAIRIDHUSHNI	1	FEVER
APOLLOCANCERHOSPITALCHENNAISANJUSHREE	1	FEVER
Timetaken:11.48seconds		

B. SCREENSHOTS

datasets - Microsoft Excel

BP MIN																
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	HOSPITAL ID	HOSPITAL NAME	CITY	DATE	PATIENT ID	PATIENT NAME	AGE	GENDER	PROBLEM	TEMP	BP MIN	BP MAX	INTERACT WITH DOC	LAB TEST	REPORTS	TREATMENT
1	1000	APOLLO SPECIALITY HC	CHENNAI	10/10/16	1243	ADHIVAN	26	M	FEVER	98	118	98	125	CHECKS THE PATIENT ACTH SUPPRESSION	SENT TO DOCT	TREATS PATIENT
2	1001	ASIAN HOSPITAL	CHENNAI	10/10/17	1905	NALLANBAN	15	M	BACKPAIN	99	104	89	134	CHECKS THE PATIENT ADRENOCORTICOTROPIC HORMONE (ACTH)	SENT TO DOCT	TREATS PATIENT
3	1002	A.G.HOSPITAL	CHENNAI	6/6/17	5564	ADHSEKA	45	M	NEURAL PAIN	100	95	81	140	CHECKS THE PATIENT ALANINE AMINO TRANSFERASE (ALT)	SENT TO DOCT	TREATS PATIENT
4	1003	A.G.M.HOSPITAL	CHENNAI	10/11/17	1747	ABUTH	68	M	APPENDIX	99	103	96	110	CHECKS THE PATIENT ALBUMIN	SENT TO DOCT	TREATS PATIENT
5	1004	A.V.M.MEDICALS	CHENNAI	10/16/17	2013	ADHUTAM	10	M	CANCER	98	100	90	110	CHECKS THE PATIENT ALKALINE PHOSPHATASE	SENT TO DOCT	TREATS PATIENT
6	1005	AASHANA HOSPITAL	CHENNAI	6/28/17	2586	ADALARACHAN	5	M	SUGAR	98	110	92	102	CHECKS THE PATIENT ALLERGY TESTS	SENT TO DOCT	TREATS PATIENT
7	1006	AGARWALLS EYEINSTIT	CHENNAI	10/11/17	3645	ADALARASEN	7	M	LEG FRACTURE	97	96	81	139	CHECKS THE PATIENT ALPHA-FETOPROTEIN (AFP)	SENT TO DOCT	TREATS PATIENT
8	1007	AMARI HOSPITAL	CHENNAI	2/4/17	2574	ADALERU	8	M	HEADACHE	97	95	91	145	CHECKS THE PATIENT AMYLASE	SENT TO DOCT	TREATS PATIENT
9	1008	APOLLO CANCER HOSP	CHENNAI	6/23/15	5566	ADHIVAN	2	M	MALARIA	99	96	92	148	CHECKS THE PATIENT ANTIBODY / TESTS (COOMBS TEST)	SENT TO DOCT	TREATS PATIENT
10	1009	APOLLO HOSPITALS	CHENNAI	5/23/15	3165	ADHIEU	19	M	TYFOD	97	95	84	102	CHECKS THE PATIENT ANTINUCLEAR ANTIBODIES (ANA)	SENT TO DOCT	TREATS PATIENT
11	1010	APOLLO HOSPITALS	CHENNAI	7/11/17	1448	ADHIVAMAN	43	M	BASAL-CELL CARCIN	99	108	95	132	CHECKS THE PATIENT ASPARTATE AMINO TRANSFERASE (AST)	SENT TO DOCT	TREATS PATIENT
12	1011	APOLLO MEDICAL CENT	CHENNAI	10/16/16	3296	ADITHAN	49	M	MELANOMA	99	99	90	125	CHECKS THE PATIENT BICARBONATE (CARBON DIOXIDE)	SENT TO DOCT	TREATS PATIENT
13	1012	B.M.HOSPITAL	CHENNAI	10/23/15	3151	JAYASLAN	21	F	SKIN CANCER (NON-ME	98	109	96	146	CHECKS THE PATIENT BILIRUBIN	SENT TO DOCT	TREATS PATIENT
14	1013	B.R.S.HOSPITALS	CHENNAI	4/6/15	2156	KAALAPRAN	83	M	PRIMARY EFFUSION LY	99	116	87	133	CHECKS THE PATIENT BLOOD CULTURE	SENT TO DOCT	TREATS PATIENT
15	1014	BALAJI HOSPITAL	CHENNAI	11/5/17	3156	KOVLKANI	45	F	ACUTE BRONCHITIS	97	102	84	102	CHECKS THE PATIENT BLOOD GLUCOSE	SENT TO DOCT	TREATS PATIENT
16	1015	BALAJI HOSPITAL	CHENNAI	7/23/15	3152	KRISHNA-CHAPRA	23	M	ACUTE EOSINOPHILIC	99	102	81	116	CHECKS THE PATIENT BLOOD TYPE	SENT TO DOCT	TREATS PATIENT
17	1016	BEGUM SHAFIA HOSPIT	CHENNAI	6/22/16	2161	KUPPURAM	7	M	ACUTE LYMPHOBLAST	100	110	97	103	CHECKS THE PATIENT BLOOD UREA NITROGEN (BUN)	SENT TO DOCT	TREATS PATIENT
18	1017	BEST HOSPITAL PRIVAT	CHENNAI	10/11/15	2175	MEYAPPAN	30	M	ACUTE MYELOID LEUKA	98	98	81	149	CHECKS THE PATIENT BREAST CANCER (BRCA) GENE	SENT TO DOCT	TREATS PATIENT
19	1018	BHARATH HOSPITAL	CHENNAI	3/18/17	5456	MOOPRAN	42	M	ACUTE MYELOID OEDI	97	118	89	142	CHECKS THE PATIENT C-REACTIVE PROTEIN (CRP)	SENT TO DOCT	TREATS PATIENT
20	1019	BHARATH HOSPITALS	CHENNAI	6/21/17	1102	MOORTHY	64	M	AIDS-RELATED LYMPH	98	110	91	111	CHECKS THE PATIENT CALCIUM (CA)	SENT TO DOCT	TREATS PATIENT
21	1020	BHARATH RAJAA HOSP	CHENNAI	3/23/15	2163	MOOVENDAN	12	M	ANAPLASTIC LARGE CI	100	103	86	147	CHECKS THE PATIENT CARDIAC ENZYME STUDIES	SENT TO DOCT	TREATS PATIENT
22	1021	BILIRUTHI HEALTH CENT	CHENNAI	3/21/17	5454	NALLADURAI	34	M	ANGIOIMMUNOBLASTIC	98	95	81	140	CHECKS THE PATIENT CD4+ COUNT	SENT TO DOCT	TREATS PATIENT
23	1022	BONRAVE HOSPITAL	CHENNAI	2/10/16	2551	NALLAYAN	64	M	B-CELL PROLYMPHOC	98	117	97	104	CHECKS THE PATIENT CHEMISTRY / SCREEN	SENT TO DOCT	TREATS PATIENT
24	1023	BONRAVE HOSPITAL	CHENNAI	11/15/15	6113	NALLAGAN	21	M	BURKITT'S LYMPHOMA	100	111	91	110	CHECKS THE PATIENT CHLAMYDIA TESTS	SENT TO DOCT	TREATS PATIENT
25	1024	C.M.HOSPITAL	CHENNAI	10/11/15	1212	NALLAPUTHU	73	M	CHRONIC LYMPHOCT	100	102	82	131	CHECKS THE PATIENT CHLORIDE (CL)	SENT TO DOCT	TREATS PATIENT
26	1025	C.S.I.KALYAN MULTI SP	CHENNAI	3/16/17	2132	NALLANBAN	19	M	CHRONIC MYELOIDIC	98	97	94	139	CHECKS THE PATIENT CHOLESTEROL AND TRIGLYCERIDES	SENT TO DOCT	TREATS PATIENT
27	1026	CHILD TRUST HOSPITA	CHENNAI	10/23/17	1213	NALLANBAN	21	M	CUTANEOUS T-CELL L	97	105	87	141	CHECKS THE PATIENT COBALAMIN	SENT TO DOCT	TREATS PATIENT
28	1027	CHINNAMMAL ENT HOSP	CHENNAI	3/26/16	2135	NALLAPERUMAL	34	M	DIFFUSE LARGE B-CEL	99	95	98	136	CHECKS THE PATIENT COMPLETE BLOOD COUNT (CBC)	SENT TO DOCT	TREATS PATIENT
29	1028	CHRISTUDAS HOSPITAL	CHENNAI	5/15/17	5213	NALLAPPA	39	M	FOLLICULAR LYMPHO	100	112	88	141	CHECKS THE PATIENT COOMBS TEST	SENT TO DOCT	TREATS PATIENT
30	1029	CIHOSPITAL	CHENNAI	6/10/16	2132	NALLAPRAN	46	M	HARRY CELL LEUKEMIA	98	108	92	145	CHECKS THE PATIENT CREATININE AND CREATININE CLEARANCE	SENT TO DOCT	TREATS PATIENT
31	1030	CITY TOWER HOSPITAL	CHENNAI	6/21/16	5846	NALLARASAN	10	M	HEPATOSPLENIC T-CE	98	101	93	121	CHECKS THE PATIENT GLOBULIN	SENT TO DOCT	TREATS PATIENT
32	1031	COSMETIC SURGERIES	CHENNAI	6/11/16	4561	NALLASUNAN	81	M	HODGKINS LYMPHOMA	98	110	89	134	CHECKS THE PATIENT GLUCOSE	SENT TO DOCT	TREATS PATIENT
33	1032	CRESCENT HOSPITAL	CHENNAI	10/4/17	5813	NALLATHAMBI	15	M	HARRY CELL LEUKEMIA	99	111	90	147	CHECKS THE PATIENT GLYCOHEMOGLOBIN (HBA1C, A1C)	SENT TO DOCT	TREATS PATIENT
34	1033	C.S.I.RAJY HOSPITAL	CHENNAI	3/5/17	5146	NALLAYAN	56	M	INTRAVASCULAR LAR	99	102	92	143	CHECKS THE PATIENT GONORRHEA	SENT TO DOCT	TREATS PATIENT
35	1034	CANCER INSTITUTE	CHENNAI	3/15/15	7546	NALLANAN	39	M	LARGE GRANULAR LYT	98	106	100	129	CHECKS THE PATIENT GROWTH HORMONE	SENT TO DOCT	TREATS PATIENT
36	1035	CHENNAI KALAPPA HOS	CHENNAI	3/8/17	8744	NALLASUN	104	M	LYMPHOPLASMACYTIC	97	117	87	146	CHECKS THE PATIENT HDL CHOLESTEROL	SENT TO DOCT	TREATS PATIENT

Ready Sheet1 - Sheet2 Sheet3 70% 5:30 PM 28/10/2018

NameNode 'localhost.localdomain:8020'

Started:	Wed Nov 01 23:37:21 PDT 2017
Version:	0.20.2-cdh3u2, 95a824e400b2a94fe1c11f1ef9db4c672ba43cb
Compiled:	Thu Oct 13 21:51:41 PDT 2011 by root from Unknown
Upgrades:	There are no upgrades in progress

[Browse the filesystem](#)
[NameNode Logs](#)

Cluster Summary

232 files and directories, 136 blocks = 368 total, Heap Size is 15.31 MB / 193.38 MB (7%)

Configured Capacity	17.33 GB
DFS Used	6.8.94 MB
Non DFS Used	5.66 GB
DFS Remaining	11.61 GB
DFS Used%	0.39 %
DFS Remaining%	66.97 %
Live Nodes	1
Dead Nodes	0
Decommissioning Nodes	0
Number of Under-Replicated Blocks	81

NameNode Storage:

Storage Directory	Type	State
/var/lib/hadoop-0.20/cache/hadoop/dfs/name	IMAGE_AND_EDITS	Active

HDFS/user/training - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://localhost.localdomain:50075/browseDirectory.jsp?dir=%2Fuser%2Ftraining&namenodeip=50070

HDFS/user/training

Contents of directory /user/training

Goto /user/training go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
clin	dir				2018-10-01 23:39	rwxr-xr-x	training	supergroup
clin	dir				2018-09-24 23:39	rwxr-xr-x	training	supergroup
clin	dir				2018-09-24 22:03	rwxr-xr-x	training	supergroup
clin11	dir				2018-10-30 03:57	rwxr-xr-x	training	supergroup
qpl_gg	dir				2018-10-02 02:50	rwxr-xr-x	training	supergroup
qpl_pg	dir				2018-10-02 02:45	rwxr-xr-x	training	supergroup
nan	file	10.58 KB	1	64 KB	2018-10-01 03:31	rwxr-xr-x	training	supergroup

Go back to DFS home

Local logs

Done

Cloudera_training_VM_1.7 - VMware Workstation

File Edit View VM Tabs Help

Cloudera_training_VM_1.7 x Cloudera_training_VM_1.7 x

Applications Places System 3:51 AM

HDFS/user/training/clin11/part-m-00000 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://localhost.localdomain:50075/browseBlock.jsp?blockid=1226394596809336040&blockSize=110206016&genstamp=221

HDFS/user/training/clin11/part-m-00000

Goto /user/training/clin11 go

Go back to dir listing

Advanced view/download options

View next chunk

```

1000, APOLLO SPECIALITY HOSPITALS, CHENNAI, 9/5/2016, 1243, ABHINAVAN, 20, M, FEVER, 97, 95, 80, 128, CHECKS THE PATIENTS, ACTH SUPPRESSION, SENT TO DOCTOR, TREATS PATIENT BY SEEING REPORTS, DESCRIBED TO PATIENTS, 7318
1001, ASIAN HOSPITAL, CHENNAI, 9/12/2005, 1565, BALLANRAN, 15, M, BACKPAIN, 99, 97, 80, 148, CHECKS THE PATIENTS, ADRENOCORTICOTROPIC HORMONE (ACTH), SENT TO DOCTOR, TREATS PATIENT BY SEEING REPORTS, DESCRIBED TO PATIENTS, 16595
1002, A. G. HOSPITAL, CHENNAI, 10/12/2017, 5564, ABHISEKA, 45, M, NEURAL PAIN, 97, 111, 94, 108, CHECKS THE PATIENTS, ALANINE AMINOTRANSFERASE (ALT), SENT TO DOCTOR, TREATS PATIENT BY SEEING REPORTS, DESCRIBED TO PATIENTS, 17918
1003, A. G. H. HOSPITAL, CHENNAI, 2/22/2005, 1147, ABIZITH, 60, M, APPENDIX, 98, 117, 92, 102, CHECKS THE PATIENTS, ALBUMEN, SENT TO DOCTOR, TREATS PATIENT BY SEEING REPORTS, DESCRIBED TO PATIENTS, 16614
1004, A. V. MEDICALS ENT RESEARCH FOUNDATION PVT LT, CHENNAI, 7/6/2005, 2033, ACHYUTAN, 10, M, CANCER, 97, 96, 80, 131, CHECKS THE PATIENTS, ALKALINE PHOSPHATASE, SENT TO DOCTOR, TREATS PATIENT BY SEEING REPORTS, DESCRIBED TO PATIENTS, 21002
1005, AASHIANA HOSPITAL, CHENNAI, 2/1/2017, 2596, ADALARACHAN, 5, M, SUGAR, 99, 115, 89, 102, CHECKS THE PATIENTS, ALLERGY TESTS, SENT TO DOCTOR, TREATS PATIENT BY SEEING REPORTS, DESCRIBED TO PATIENTS, 7183
1006, AGARWALLS EYE INSTITUTE PRIVATE LIMITED, CHENNAI, 3/18/2017, 9045, ADALARASEN, 7, M, LEG FRACTURE, 99, 97, 94, 103, CHECKS THE PATIENTS, ALPHA-FETOPROTEIN (AFP), SENT TO DOCTOR, TREATS PATIENT BY SEEING REPORTS, DESCRIBED TO PATIENTS, 5620
1007, AMAR HOSPITAL, CHENNAI, 8/17/2016, 2574, ADALERU, 8, M, HEADACHE, 99, 110, 82, 132, CHECKS THE PATIENTS, AMYLASE, SENT TO DOCTOR, TREATS PATIENT BY SEEING REPORTS, DESCRIBED TO PATIENTS, 21024
1008, APOLLO CANCER HOSPITAL, CHENNAI, 2/12/2015, 5566, ADHIBHAN, 2, M, MALARIA, 98, 104, 80, 148, CHECKS THE PATIENTS, ANTIBODY TESTS (COOMBS TEST), SENT TO DOCTOR, TREATS PATIENT BY SEEING

```

Done reference (~\Desktop) - gedit

[training@localhost ~] reference (~\Desktop) ... HDFS user/training/clin...

To direct input to this VM, move the mouse pointer inside or press Ctrl-G.

```
mysql> select * from treat limit 5;
```

HOSPITAL_I	HOSPITAL_NAM	CIT	DAT	PATIENT_I	PATIENT_NAM	AG	GENDE	PROBLE	TEMPERATURE_MI	TEMPERATURE_MA	BP_MIN	BP_MA	INTERACT_WITH_DOCTO	LAB_TES	REPORT	TREATMEN	PRESSCRIPTIO	BIL
1000	APOLLO SPECIALITY HOSPITALS	CHENNAI	10/29/2015	1243	ABHINAVAN	26	M	FEVER	99	100	87	100	CHECKS THE PATIENTS	ACTH SUPPRESSION	SENT TO DOCTOR	TREATS PATIENT BY SEEING REPORTS	DESCRIBED TO PATIENTS	23004
1001	ASIAN HOSPITAL	CHENNAI	12/31/2017	1565	NALLANBAN	15	M	BACKPAIN	100	96	94	108	CHECKS THE PATIENTS	ADRENOCORTICOTROPIC HORMONE (ACTH)	SENT TO DOCTOR	TREATS PATIENT BY SEEING REPORTS	DESCRIBED TO PATIENTS	4059
1002	A.G.HOSPITAL	CHENNAI	11/27/2015	5564	ABHISEKA	45	M	NEURAL PAIN	97	109	91	127	CHECKS THE PATIENTS	ALANINE AMINOTRANSFERASE (ALT)	SENT TO DOCTOR	TREATS PATIENT BY SEEING REPORTS	DESCRIBED TO PATIENTS	18508
1003	A.G.M.HOSPITAL	CHENNAI	1/19/2015	1147	ABIJITH	68	M	APPENDIX	97	96	94	136	CHECKS THE PATIENTS	ALBUMIN	SENT TO DOCTOR	TREATS PATIENT BY SEEING REPORTS	DESCRIBED TO PATIENTS	20115
1004	A.V.M.MEDICALS ENT RESEARCH FOUNDATION PVT LT	CHENNAI	4/11/2017	2013	ACHYUTAM	10	M	CANCER	97	107	80	120	CHECKS THE PATIENTS	ALKALINE PHOSPHATASE	SENT TO DOCTOR	TREATS PATIENT BY SEEING REPORTS	DESCRIBED TO PATIENTS	22761

```
hive > insert overwrite table buck select * from hos1;
```

Ended Job = job_201809262149_0007

Ended Job = 848722547, job is filtered out (removed at runtime).

Moving data to: hdfs://localhost/tmp/hive-training/hive_2018-09-26_23-25-39_693_5255021456270881882/-ext-10000

Loading data to table hospital.gende_buckets

Deleted hdfs://localhost/user/hive/warehouse/hospital.db/gende_buckets

Table hospital.gende_buckets stats: (num_partitions: 0, num_files: 1, num_rows: 0, total_size: 11000670)

49999 Rows loaded to gende_buckets

OK

Time taken: 5.908 seconds

C. PUBLICATIONWITHPLAGIARISMREPORT

TurnitinOriginality Report

Processed on: 21-Feb-
202010:56AMCSTID:1261533112
WordCount: 1913
Submitted:1

SimilarityIndex
13%

SimilaritybySource

InternetSources: 9%
Publications: 11%
StudentPapers: 9%

Personalization For
FeatureSelection In Clinical
FormAuto Filling By K.

Tejasri[
Lakshmanan, Dr.L.

7%match(Internetfrom05-May-
2019)<https://es.scribd.com/document/341486263/MapReduce-Algorithms>

3%match(publications)

[L. Ramanathan, G. Parthasarathy, K. Vijayakumar, L. Lakshmanan, S. Ramani."Cluster-based distributed architecture for prediction of student's performance in higher education", Cluster Computing, 2018](#)

1%match(Internetfrom29-Oct-
2019)<https://ijsiet.org/vol32018/11.pdf>

1%match(studentpapersfrom04-Jun-2016)

[Submitted to ABV- Indian Institute of Information Technology and Management Gwalior on 2016-06-04](#)

<1%match(studentpapersfrom25-Apr-2016)

[Submitted to Savitribai Phule Pune University on 2016-04-25](#)

Personalization For Feature Selection In Clinical Form Auto Filling
K. Tejasri[1], L. Sreevaishnavi[2], DR. L. Lakshmanan[3]. [1], [2] UG [Student, Department of Computer Science Engineering, Sathyabama Institute of Science and Technology](#). [3] [Professor, Department of Computer science, Sathyabama Institute of Science and Technology](#) Abstract:--
Feature learning and selection will be fully used in many technical areas for the reason because of their advanced implementation and less delicacy. Usually established models are customarily treating altogether based on terms with the even feature sets, such that should be able to strident and that gives
Accompanied information can give some distinct attributes for the given model. Here, just we introduced one terminology i.e., personalization accessible to discuss the properties based on each and every form. Each point will be given as the input so that methods will give much importance to these selections scheme. Second, it describes property of searching pattern to induce feature aspirant subsets for every phrase. Here, towering facts of individual deals with the investigation of these serene details, problems, Intelligence and hospital specialties. Finally, it attains the personalized phrase set for every term to relate for all phrases so they can be used to integrate onto searching method.
Keywords—Big Data, HDFS, MapReduce, CSV.I. INTRODUCTION Curing is superior for scrupulous health problem. Feature selection is a searching pattern of a incisive skill for proposing the new bring out subsets, next to with evaluation measures which contains different feature subsets. By [HDFS] spread

out to limit storage uses and can raise algorithm speed. The term "Big data" is the gathering of huge chunks of information that can be used by old techniques. i.e., Social media, it consists of the structured, unstructured, semi-structured data. It can transfer the data from Hadoop with the Sqoop and possess the data using Hive. It can analyze the data using Pig. It is analyzing clinical data with the use of Hadoop appliance with the Hadoop terms like HDFS, MapReduce, Sqoop, Hive and Pig. While using all these set of tools, it will afford to run unbounded data. By this we will get high output, and Continuation value also very low and the software that can be used here is everyone can modify and share easily. It executed on all the objectives and it is based upon Java. From this

clinical form feature selection will know which hospital problem.

Here using the [CSV] comma delimited which load the data set in this we are not giving the data entries. II. RELATED WORK Daniel Jarrett [4] exact prediction of diseases are critical and it is essential for premature recognition and identify suitable cure of patients at risk. Methods in survival psychoanalysis are frequently followed by highly flavoured parametrical assumptions and inadequate in their ability to be trained from high-dimensional data. This

paper develops a novel come close

to that address the corner of cooperation time honoured arithmetical approaches as splendidly as recent setup network models for

survival. We match MATCH-Net: a game strategy Awareness Sequential Hitting-time Network, deliberate for capture activist dependencies. L. LAKSHMANAN

[7] Primarily contributes towards the EDM by introducing the cluster based distributed architecture for predicting the student performance for the helpful academics. The future cluster-based two-tier architecture for Predicts the student's accomplishment by collecting the academic performance of the students in earlier records. The anticipated architecture has two joining layer, and two distributive layers for performing various operations, such as clustering, feature extraction

and the Prediction. M. AKAY [8] point of this uncommon segment is to give for an outline of the increasing biomedical informatics advancements and their applications in inquire about and clinical conditions. After advancements in biomedical informatics have made

strategies, strategies and apparatuses, which depend on the investigation of heterogeneous information, information mining, choice emotionally supportive networks, multi-scale demonstrating, and soon. These separation from the enhancements of such frameworks what's

more, the genuine clinical situation is still fully long, and

just some of them have been used in a clinical scale. Physiological signals. Matthew R. Boutella [9] In a great example acknowledgment things classes are fundamental unrelated by definition and Classification mistakes can happen when the classes cover in the component space. will look at a different situation, happening when the classes are, by definition, not totally unrelated. Such issues emerge in significant scene and hide classification and in restorative analysis. Presenting a system to deal with such issues and apply it to the issues of significant scene classification, where each characteristic scene may contain different protest to such an extent that the scene can be depicted by different class names (e.g., an "old scene with a mountain in the foundation"). Mark H. Drazner [10] Hypertension remains an noteworthy general medical issue related with correct grimness and mortality. Hypertensive coronary illness is a group of star of variations from the norm of hypertensive coronary illness is that the left ventricular (LV) divider thickens in light of high blood

weight as a compensatory instrument to limit divider push. In this way, after a progression of ineffectively portrayed occasions ("change to disappointment"), the left ventricle expands, and the LV discharge portion (EF) decreases (characterized thus as "widened heart failure"). III. PROPOSED SYSTEM: Anticipated

system deals with benevolent file by using Hadoop tool. it will be able to dissect a barrier of facts and plainly tally up to machinery to the cluster and contract domain effect with reduced amount of time, important throughput and maintenance rate is enormous, not as much and by using partitions and bucketing terminologies in System.

ADVANTAGES IN PROPOSED SYSTEM: a. No information

shortfall problem. b. Systemized helpful data is processed. Fig 3.1

Proposed System. IV. ALGORITHMS: Algorithm 4.1: Wordcount Mapper emit

san intermediate key-value two of a kind for both word in a document. The reducers sum up entirely counts for all word. 1: class status Mapper2: process

plot(codcrendtialsa, codd) 3: together period $\in$ codd look

after 4: Emit(term t, calculate 1) 1: class Reducer 2: Approach Reduce(term t, counts) 3: the lot $\leftarrow$ 0 4: for entirely consider c $\in$ counts[count] do organize 5: sum

m = sum + c 6: Emit(term t, count) Algorithm 4.2: Wordcount mapper with associative arrays 1: wordcount Mapper 2: method Map terms(credentials id a, codd) 3: for all term contains H

$\leftarrow$ to the new Associative Array 4: for all the term t consists $\in$ codd do 5:

H{t} $\leftarrow$ H{t} + 1. Tally calculates for total document 6: count t $\in$ H do 7: Emit the term contains (term t, count H{t}) Algorithm 4.3: Word

count mapper will use the "map-reduce" 1: class

Mapper(wordcount) 2: procedure map term t do initialize 3: H $\leftarrow$ new Associative Array 4: procedure Map(credentials id a, codd) 5: for all the term t $\in$ codd do 6:

H{t} $\leftarrow$ H{t} + 1. Tally calculates for all documents 7: method will stop. 8:

for all term t consists of t $\in$ H do 9: Emit the (term t,

count H{t}) V. METHODOLOGIES Analysing the information with an assortment various kinds of properties in Microsoft Excel do extremely well configure it will follow

that it will be formed as a comma delimited form i.e., (comma separator value) bind with the MySQL encouragement such a Database. 1.) Pre-

processing: By getting paid chronological facts it translates individuals will convert those

chronological batch handing out files from (.xls) layout to (.csv) setup and by retrieving everyone of the files MySQL catalog to prevent be a victim of data. Fig 5.

1 pre-

processing. 2.) Processing (Map reduce): It is a framework which can be capable to inscribe applications to deal with

colossal chunk of Clinical, with both spaces on corpulent forms and garnish the stack in a effortless way. It is a giving out skill and a agenda prototypical for scatterer computed based upon java. plan demote algorithm which hold two imperative ways, specifically Map and Reduce. Which demote the program executes in three phases, namely map phase, shuffle-

phase, and reduce phase. This job is to the method of key data. Fig 5.2 processing. 3.) Analyse Query: Takes every part of the figures from HDFS within the HIVE by

depletion of sqoop import decree. where hive is available to ring to go for analysis sort out point in HIVE it will deal with no more than the structured fact to analyze. By retrieving some meaningful data and avoiding indistinguishable data in further victorious method by consume of hive. Fig 5.3 Analyse Query. 4.) Analysis

is Latin Script (Pig): Inspect Clinical using Pig, developer to create their scripts via the language i.e., Pig Latin and effect those parameters for the use of means mode while using the Grunt shell. Every part of this scripts are interlink that should be converted to Map and Reduce ways. After connecting it will be present in the Grunt

shell, Presses and run the Pig scripts in the shell. Except Insert and Cache, time liability everyone other operations, ease charge lots the figures in file system be capable of move out. Pig provides lot of built-in operators to be familiar with

numbers of operations. Fig 5.4 Analysis Latin Script. 1 V. RESULTS AND DISCUSSION RESULTS: CONCLUSION It presents a reading on Clinical to be of assistance to

doctors by getting the intent of what treatment has been agreed to that not worthy tolerant in the past. A patient cannot tolerate be

able to develop an awareness to look at which infirmaries are treating improved for a stidious problems. Analyze the clinical data via Hadoop for where Hive and Pig used for analysis and MapReduce for processing.

REFERENCES [1] Bhatti UA, Huang M, Zhang Y, et al. explore on the Smartphone Based Health Systems for Strengthening Healthcare Organization [M] // Smart Health. 2017. [2] Black AD, CarJ, Pagliari C, et al. "The contact of Health on the trait and shelter of healthcare: a systematic overview", *PLoS Medicine*, vol. 8, no. 1, 2011.

[3] Cao P, Xian Medical University. "The effect of great figures on the Expansion of Chinese Health Service overhaul Industry", *Journal of Electronic Test*, 2014 [4] Daniel Jarrett, Jinsung Yoon, Member, IEEE, and Mihaela Fronder-Schaar, Fellow, IEEE "Dynamic Prediction in Clinical Survival Analysis using by earthy Convolutional Networks".

[5] Hermann M, Pentek T, Otto B. Plan ideology for Industrie 4.0 Scenarios, pp. 3928–3937, 2016. [6] Horsky J, Schiff G D, Johnston D, et al. "Methodological Review: Interface design principles for usable decision support: A targeted review of best practices for clinical prescribing interventions", *Journal of Biomedical Informatics*, vol. 45, no. 6, pp. 1202–1216, 2012. [7] L. Lakshmanan, k. Vijayakumar, G. Parthasarthy, S. Ramani, "Cluster based distributed architecture for prediction of students performance in higher education". *Cluster Computing*, vol. 16, pp. 5–7 2019. [8] M. Akay, D. I. Fotiadis, K. S. Nikita and R. W. Williams, "Guest Editorial: Biomedical Informatics in Clinical Environments", *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 1, pp. 149–150, 2015. [9] Matthew R. Boutella, Jiebo Luob "Learning multi-label scene classification", 2015. [10] Mark H. Drazner "The Progression of Hypertensive Heart Disease", 2016. [11] Thummler C, Bai C. Health 4.0: How Virtualization and Big Data are Revolutionizing Healthcare, Springer International Publishing, 2017.

[12] Vancampfort D, Mugisha J, Hallgren M, et al. "The prevalence of diabetes mellitus type 2 in people with alcohol use disorders: a systematic review and large scale meta-analysis", *Psychiatry Research*, vol. 246, pp. 394–400, 2016. [13] Yu W D, Pratiksha C, Swati S, et al. "A Modeling Approach to Big Data Based Recommendation Engine in Modern Health Care Environment", *Computer Software and Applications Conference, IEEE Computer Society*, pp. 75–86, 2015.