**M S Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Data Analytics on Medicare Payments

Submitted by

Sriram Anupindi 1MS12CS111

Suhail T N 1MS12CS115

Mayank Sharma 1MS11CS062

Ashresh Prasad 1MS11CS034

*In partial fulfillment for the award of the degree of*

# *Bachelor of Engineering in Computer Science & Engineering*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu/), **May 2015**

# Abstract

The objective of this project is to derive a relationship between the various features of the data set chosen. The relationship will help identify the trends and patterns among the various features of the data set. The data set chosen will provide information regarding Medicare [1] payments during a certain fiscal year covering a certain demographic for a list of popular procedures. Due to the vast size and complexity of the data set a parallel, distributed algorithm on clusters will be utilized to simplify the data analytics operation. The simplified data will be parsed in order to extract the required features and values . Machine learning and Data Mining techniques are used to extract a relationship between the extracted features. Visualization techniques are used to provide a comprehensive and graphical representation of the results. At the end of the project we wish to gain a better understanding how Medicare Payments are dependent on procedures and vary from demographic to demographic.

#### Contents

***Abstract i***

***List of Figures***

***List of Tables***

1 **INTRODUCTION**

* 1. General Introduction……………….
  2. Statement of the Problem…………..
  3. Objectives of the project……………
  4. Project deliverables……………
  5. Current Scope………………………
  6. Future Scope……………………….

1. **PROJECT ORGANIZATION**
   1. Software Process Models
   2. Roles and Responsibilities
2. **LITERATURE SURVEY**

3.1….Introduction

3.2…Main Body

3.3 Conclusion of Survey

1. **SOFTWARE REQUIREMENT SPECIFICATIONS**

5.1 External Interface Requirements

5.2.1 User Interfaces

5.2.2 Hardware Interfaces

5.2.3 Software Interfaces

5.2.4 Communication Interfaces

5.2 Functional Requirements

5.3.1 Functional Requirement 1.1

5.3.1 Functional Requirement 1.2

5.3.1 Functional Requirement 1.3

5.3.1 Functional Requirement 1.4

5.3.1 Functional Requirement 1.5

5.3 Software System Attributes

5.4.1 Reliability

5.4.2 Availability

5.4.3 Security

5.4.4 Portability

5.4.5 Maintainability

5.4.6 Performance

5.4 Performance Requirements

5.5 Design Constraints

5.6 Other Requirements

1. **DESIGN**
   1. Introduction
   2. Architecture Design
   3. Data flow diagram
2. **IMPLEMENTATION**
   1. Tools Introduction
   2. Technology Introduction
   3. Overall view of the project in terms of implementation
   4. Explanation of Algorithm and how it is been implemented
3. **TESTING**

**7.1** Results and Snapshots

1. **CONCLUSION & SCOPE FOR FUTURE WORK**
2. **REFERENCES**

**INTRODUCTION**

General Introduction

Health Insurance is one of the most sought after form of Insurance, accounts for 19.3 percent profit margin as of 2008. In most countries medical insurance is provided by employers for their employees, often many employees take additional insurance in order to cover all members of their family. In many medical situations patients end up paying additional charges in spite of shelling out huge sums of premiums yearly. This practice seems quite skeptical, in order to shed some light onto how much surcharge a patient pays we decided to investigate Medicare.

Medicare is a national social insurance program administrated by the U.S federal government. This system is implemented by using a 30 private insurance companies spread across the United States. Medicare provides its services for Americans aged 65 and older who have worked and paid into the system, where the system refers to the U.S economy. Medicare also covers persons with disabilities as well as those who permanent kidney damage. In the year 2011 Medicare was the primary payer for an estimated 15.3 million inpatient procedures, which represents 47.2 % of the total aggregate inpatient hospital costs in the United States. Medicare consists of two parts[2]:

1. Hospital Insurance- Covers hospital procedures, discharges, and medical tests which a standard part of Medicare.
2. Medical Insurance- Is not included in the standard Medicare as it involves paying a premium.[3]

Statement of the Problem

To identify relationships between the features of the data set pertaining to the provision of medical services (Hospitals) with regard to the 100 most frequent discharges spread across a vast number of geographical locations in the United States of America. The lists of relationships that have been identified by us are:

1. State wise count of the no of cases for a particular procedure.
2. Average cost for a particular procedure across all hospitals in a particular state.
3. Classification of hospitals as high, medium and low for a particular state based on the number of discharges.

Objectives of the project

Through this project we will aim to achieve basic understating of the data analytics on large data sets. During the course of the project we wish to arrive at a meaningful conclusion from the raw data selected. Since the data is of huge size, we require a robust storage and computational framework which leads us to Hadoop. During the utilization of Hadoop we get a better understating of Cluster based computing

With the due course of time we will be utilizing many tools and algorithms such as:

1. Pig script [4]
2. Octave [5]
3. Clustering Algorithms [6]
4. Scripting Languages
5. Matrix Manipulations
6. Data Driven Documents using JS [7]

Project deliverables

1. A collection of simplified documents from the Original data set selected for the Project.
2. An interactive web interface, which will give a comprehensive and concise view of results, acquired.
3. To visually show the distribution of number of cases for a particular procedure in the USA.
4. Utilizing data maps for java script we provide a holistic view of the average cost distribution in the United States for a particular procedure.
5. Clustering techniques will be implemented to provide a relative grading of hospitals in a particular state.

We will provide a holistic, complete and simplified view of the original data so as to arrive at a meaningful conclusion.

Current Scope

In the US currently, Medicare is being implemented by using many private insurance companies, which are spread across the United States. Medicare provides its services for Americans aged 65 and older who have worked and paid into the system. This analysis is designed keeping in mind the US demographic. It proves to be beneficial for US lawmakers in making better judgment in making the Healthcare policies. This analysis will help in determining which area needs immediate attention on this front. Not only the lawmakers but also the citizens of the United States could use this analysis for a better understating and transparency of the Insurance system.

Future Scope

Patients end up paying additional charges in spite of shelling out huge sums of premiums /taxes yearly. This practice seems quite skeptical. In India central government organizations such as CGHS and respective state government health schemes like Jyothi Sanjeevini scheme, Rajiv Arogya Bhagya can be supported with hospital insurance and medical insurances so that many people aged above 65 who mainly are retired government servants can be benefited. A similar analysis can be used for the Indian demographic, benefiting citizens and policy makers.

1. **PROJECT ORGANIZATION**

Software Process Models

A software process model is an abstract representation of a process. It presents a description of a process. When we describe and discuss processes, we usually talk about the activities in these processes such as specifying a data model, designing a user interface, etc. and the ordering of these activities.

The waterfall model is a sequential design process, used in software development processes, in which progress is seen as flowing steadily downwards (like a waterfall) through the phases of conception, initiation, analysis, design, construction, testing, production/implementation and maintenance. We chose Waterfall model for this Project as it is easy to implement.

The model can be split into the following five phases:

Data acquisition

Data sets used for our project are huge in size. We acquired the data from the United States government website. It holds information regarding procedures for a particular fiscal year per state.

Filtering the data

The data sets being large in size and also contain unwanted data that needs to be filtered. Hence, using Hadoop and Pig script does filtering.

Extracting key features and parsing

Not all the fields are of the desired formats and specification. Hence, the required features need to be extracted by using parsing techniques and feature extraction. This process simplifies various algorithmic implementations.

Algorithms

Due to large complexity we are going to reduce the original data set to a simpler data set by using distributed processing. Popular clustering algorithms will be implemented to help find similarities between data points.

Results with visualization techniques

The resulting analysis will be stored in distributed files. Popular visualization techniques will be applied for a better understating of result acquired.

Roles and Responsibilities

Since the project requires multiple skill sets, each and every individual was assigned equal responsibility.

The project can be divided into the following five major responsibilities:

1. Data acquisition and filtering: To justice this area of the project the person must be sound with Hadoop[8] and Pigscript.
2. Storage management and parsing: The person must be well versed in the HDFS storage architecture. They should be efficient in python for efficiently parsing the required data.

1. Front end web development: The person must have good HTML, CSS and web development skills.
2. Visualization development: The person must be comfortable with using d3.js, handling of file input and outputs.

1. Documentation: The person should be a proficient typist and must have extraordinary vocabulary and grammar skills.

All the members of the team should be co-operative, organized and well aware of the project goals, aspirations, deadlines and outcomes. Each and every individual has a key role in making the project a success.

1. **LITERATURE SURVEY**

Introduction

The overall goals of this project were firstly to establish the significance of the general field of study, then identify a place where a new contribution could be made. The bulk of the chapter was on critically evaluating the different methodologies used in data analytics so as to identify the appropriate approach and implement it.

Main Body

For the fast and successful completion of the project it was divided in to different phases

Data acquisition, filtering and feature extractions, algorithmic implementations and visual

representation of data.

Data acquisition:

The first task of our project was to find a database, which contains more than 1 lakhs of data. After extensive searching we realized that the medical insurance in USA is very prominent and holds lots of data. Then we started searching for the database related to it and finally we found it in www.usa.gov . The data is about different medical procedures in different hospitals throughout USA.

Filtering and feature extractions:

Since the data acquired is huge and contains unwanted details it was difficult to perform different tasks using it. In order to overcome this we decided to filter data using distributed cluster computing. Multiple options such as Hadoop, Spark[9], etc were available which satisfy the above criteria. Since Hadoop was compatible with our system, we chose that over spark.

Since there were anomalies in the data it made it difficult while extracting key features from the datasets. These anomalies were removed using python programming language. We chose python[10] over every other programming languages due to its simplicity, robustness and reliability.

Algorithmic implementations:

We read more about our problem statements and came up with lots of algorithms that can help during the course of our project. After further research on the algorithms we shortlisted them based on its simplicity, time complexity and compatibility with our system. The algorithms, which we used in our project are-

1. Map Reduce[11] on a HDFS[12] System: This Data Mining technique helps to reduce the number of features and records which is to be extracted from the original data set.

2. Clustering using K-means[13]: This Data Mining technique is utilized to help classify hospitals as high, medium & low intake for a particular procedure based on the number of discharges.

Visual representation of data:

In order to represent our project results in a more striking manner we decided to do a front end to our project. There are many traditional ways such as bar charts, histograms etc to represent data. In order to represent our data in more interactive and dynamic manner we decided to explore various java scripts options like polymaps.js, d3.js, springy.js [14] etc. We chose d3.js because many libraries such as nvd3 and data maps. These libraries allow us to represent data in various geometric figures as well as geographical map representations.

Conclusion of Survey

During the course of the project there were so many obstacles that we had to overcome but with the help thorough literature survey as clearly mentioned above we could meet our deadlines and successfully complete our project.

**Software Requirement Specification**

1. External Interface Requirements
   1. User Interfaces:

Simple, elegant and comprehensive website which represent data in a graphical manner. Graphical representations include pie charts, maps and plots.

* 1. Hardware Interfaces:

The application will run on a laptop with min 4GB RAM, 64 bit configuration, should have NIC, and runs on Linux configuration. No Further hardware devices or interfaces will be required for this analysis.

* 1. Software Interfaces:

1. Hadoop: Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware

2. Java 1.7: Java is used to work with Apache Hadoop

3. Octave: Octave is software featuring a high-level programming language, primarily intended for numerical computations.

4. Python: Python is a widely used general-purpose, high-level programming language and it is used for passing and editing the dataset.

5. D3.js: D3.js is a JavaScript library for producing dynamic, interactive data visualizations in web browsers.

* 1. Communication Interfaces:

Internet connection and a web browser are required in order to make use of several functions and to be executed such as searching, viewing and downloading.

1. Functional Requirements
   1. Retrieving Input :

The software will receive 2 inputs

● Medical Procedure

● State Name

* 1. Map reducing:

Map reducing is done in Hadoop with the help of in-built java libraries and it’s functions are written in pig script.

* 1. Passing data:

Passing data is done with the help of Python. Using python ‘$’ signs in the data are removed so that the data can be treated as a floating value.

* 1. Matrix Operation:

Data is entered to the octave which converts it in the form of a matrix and then K-means algorithm is applied to it.

* 1. Output:

Output is graphically represented in the web page. Graphical representation includes maps, pie-charts and plots. These are implemented using D3.

1. Software System Attributes
   1. Reliability:

The software will meet all of the functional requirements without any fail and graphically display the appropriate data.

* 1. Availability:

Software is available anytime depending on the internet connection.

* 1. Security:

Since the software is not taking any personal information from the user it will never have security issues.

* 1. Portability:

This software is designed to run on any operating system.

* 1. Maintainability:

The software should be written clearly and well documented. Special care is taken in developing front end to make it user friendly and easy to use.

1. Performance Requirements:

All the operations should run in the background without affecting the normal process of system.

1. Design Constraints:

Data taken is only for one year.

Hadoop is slow compared to Apache Spark.

Hadoop is heavy resource consumer.

1. Other Requirements:

Knowledge on different medical procedures, hospitals and states of USA

**DESIGN**

1. Introduction
   1. **Number of Modules**

1. Hadoop

2. Java

3. Octave

4. Python

5. D3

* 1. **Modules description**

**Hadoop:**

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.

**Java:**

Java is used to work with Apache Hadoop

**Octave:**

Octave is software featuring a high-level programming language, primarily intended for numerical computations.

**Python:**

Python is a widely used general-purpose, high-level programming language.

**D3:**

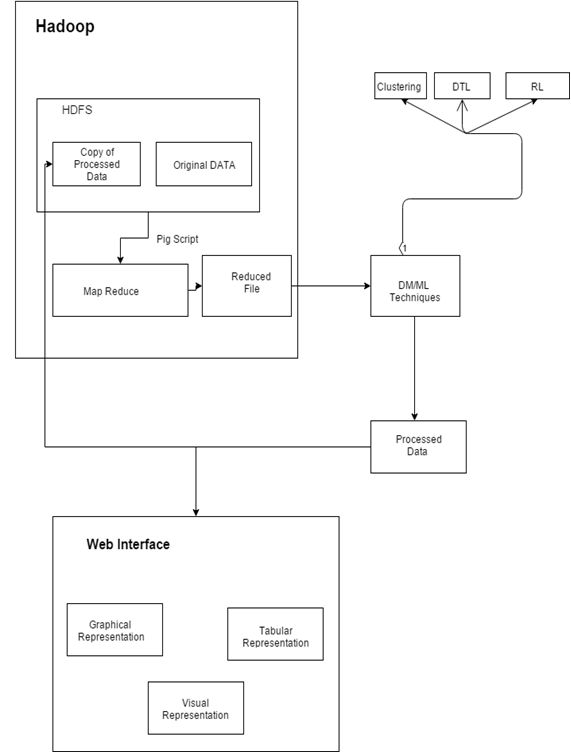
D3.js is a JavaScript library for producing dynamic, interactive data visualizations in web browsers.

* 1. **Algorithm design**

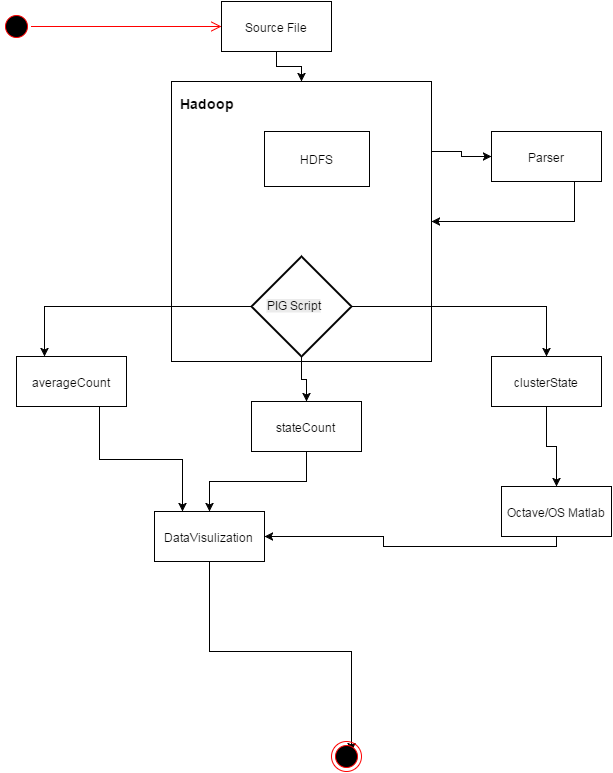
**1.** **Map Reduce on HDFS**: This Data Mining technique helps reduce the number of features and records which is to be extracted from the original data set.

**2. Clustering using K-means or HAG:** This Data Mining technique is utilized to help classify hospitals as high, medium & low intake for a particular procedure based on the number of discharges.

1. Architecture Design



1. Data flow diagram



**IMPLEMENTATION**

Tools Introduction

The tools used for the implementation of this project are many, varying from storage and computation of data to visualization of data. Below paragraphs will highlight the importance of each tool that is used in this project.

The following tools are:

* Pig Script: **Apache Pig**is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.
* Octave: GNU Octave is a high-level interpreted language, primarily intended for numerical computations. It provides capabilities for the numerical solution of linear and nonlinear problems, and for performing other numerical experiments. It also provides extensive graphics capabilities for data visualization and manipulation.
* Python: Python is a widely used [general-purpose](https://en.wikipedia.org/wiki/General-purpose_programming_language), high-level programming language.[[20]](https://en.wikipedia.org/wiki/Python_(programming_language)#cite_note-AutoNT-34-20)[[21]](https://en.wikipedia.org/wiki/Python_(programming_language)#cite_note-21) Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer [lines of code](https://en.wikipedia.org/wiki/Lines_of_code) than would be possible in languages such as C++ or Java.[[22]](https://en.wikipedia.org/wiki/Python_(programming_language)#cite_note-Summerfield-22)[[23]](https://en.wikipedia.org/wiki/Python_(programming_language)#cite_note-23) The language provides constructs intended to enable clear programs on both a small and large scale.
* D3: D3 allows you to bind arbitrary data to a Document Object Model (DOM), and then apply data-driven transformations to the document. For example, you can use D3 to generate an HTML table from an array of numbers. Or, use the same data to create an interactive SVG bar chart with smooth transitions and interaction.

Technology Introduction

Due to large size of the data we need to opt for more cluster based approach. The following two technologies are utilized:

* Hadoop: Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.
* HDFS: HDFS is a Java-based file system that provides scalable and reliable data storage, and it was designed to span large clusters of commodity servers. HDFS has demonstrated production scalability of up to 200 PB of storage and a single cluster of 4500 servers, supporting close to a billion files and blocks.

Overall view of the project in terms of implementation

In order for the Hadoop to work a java ecosystem is required on the workstation. Following the installation of Hadoop involving the setting up of data nodes and name nodes, the workstation is ready for analytics. In order to simplify the process of writing map reduce algorithms, Apache pig binary libraries are included onto Hadoop. The source file is stored on HDFS for later processing which is achieved through running Pig scripts. The results of the map reduce are also stored in HDFS and later downloaded to the local file system.

The downloaded data contains quite some anomalies which need to be rectified in order for a successful implementation of the algorithm. This is achieved by utilizing languages like Python, which parse the required data. The parsed data is restored on HDFS for further map reduce processing.

Using Matlab functions, which run on the octave platform, perform the clustering of the data. The results of which are stored in the web application front end folder so as to be visually displayed by web browser. This visualization of data is implemented by using d3 java script libraries. The similar implementation flow is maintained for achieving the other objects of the project.

Explanation of Algorithm and how it is been implemented

In this section of the report we will be exploring the major algorithms implemented

in the project.

Map reduce algorithm: The MapReduce algorithm contains two important tasks, namely Map and Reduce.

* The map task is done by means of Mapper Class
* The reduce task is done by means of Reducer Class.

Mapper class takes the input, tokenizes it, maps and sorts it. The output of Mapper class is used as input by Reducer class, which in turn searches matching pairs and reduces them.

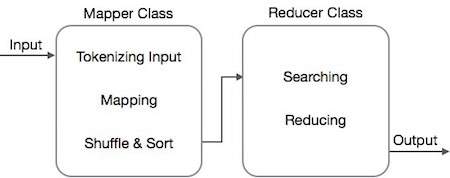


Fig1:Workflow of Map Reduce

MapReduce implements various mathematical algorithms to divide a task into small parts and assign them to multiple systems. In technical terms, MapReduce algorithm helps in sending the Map & Reduce tasks to appropriate servers in a cluster.

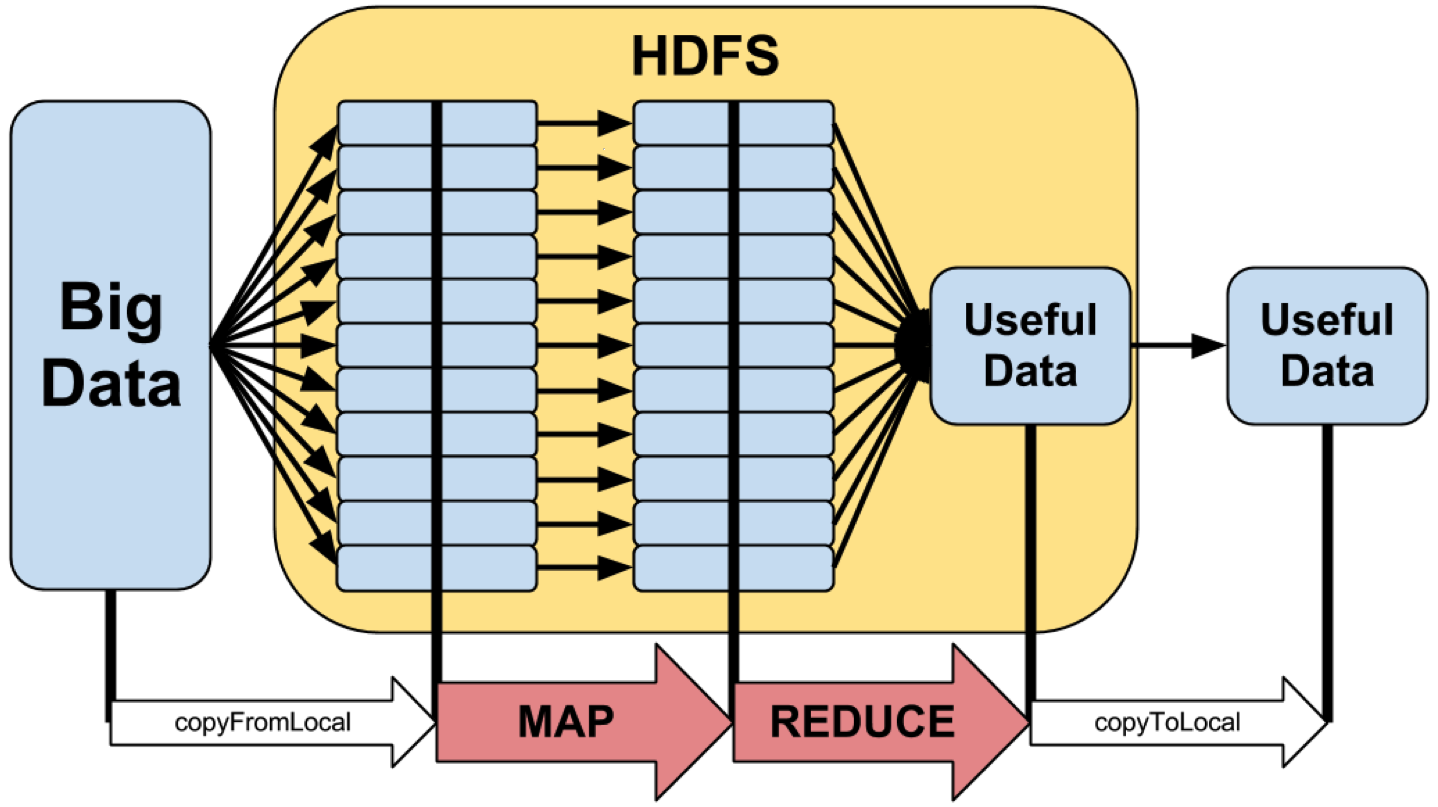


Fig2: HDFS & Map Reduce

In our project we are implementing map reduce algorithms with the help of Pig script. Below are the examples of the Pig script used in our project.

Pig Script to calculate average cost for a procedure per state:

L1 = LOAD '/user/hduser/PBL/avgcost.csv' USING PigStorage(',');

L2 = FILTER L1 BY($0==' ');

L3 = FOREACH L2 GENERATE $1,$2;

L4 = GROUP L3 BY $0;

L5 = FOREACH L4 GENERATE $0,SUM($1.$1),COUNT($1);

L6 = FOREACH L5 GENERATE $0,(float)$1/$2;

STORE L6 INTO '/user/hduser/PBLOUT2' USING PigStorage(',');

Pig Script to calculate number of cases handled by a hospital in a particular state:

S1 = LOAD '/user/hduser/PBL/source.csv' USING PigStorage(',');

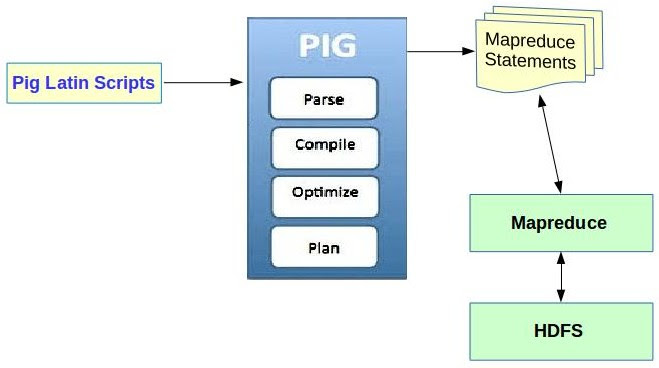
S2 = FILTER S1 BY $5==' ';

S3 = FOREACH S2 GENERATE $1,$8;

S4 = GROUP S3 BY $0;

S5 = FOREACH S4 GENERATE $0,SUM($1.$1);

STORE S5 INTO '/user/hduser/PBLOUT2' USING PigStorage(',');



1. Fig3:Pig Script integration with HDFS

Clustering algorithm: Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for [statistical](https://en.wikipedia.org/wiki/Statistics) [data analysis](https://en.wikipedia.org/wiki/Data_analysis), used in many fields, including [machine learning](https://en.wikipedia.org/wiki/Machine_learning), [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), [image analysis](https://en.wikipedia.org/wiki/Image_analysis), [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval), and [bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics).

There are many clustering algorithms available in the market, to name a few

Centroid based clustering, Distribution based clustering and Density based clustering. Now based on our data we decided to use Centroid based clustering also commonly known has K-means algorithm. We use K-mean because here we need to classify the data into three clusters High, Medium and Low. K-means is implemented using Matlab function; the first step in the function was to extract the data from the source CSV file following which the data was stored in a matrix format. The matrix is then manipulated and passed with a K- means function. The result is a vector, which is concatenated with the original data and stored.

K-means algorithm: The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the *k*-means algorithm; it is also referred to as Lloyd’s Algorithm, particularly in the computer science community.

Given an initial set of *k* means *m*1(1),…,*mk*(1) (see below), the algorithm proceeds by alternating between two steps:

**Assignment step**: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means).

S_i^{(t)} = \big \{ x_p : \big \| x_p - m^{(t)}_i \big \|^2 \le \big \| x_p - m^{(t)}_j \big \|^2 \ \forall j, 1 \le j \le k \big\},

Where each x_p is assigned to exactly one S^{(t)}, even if it could be assigned to two or more of them.

**Update step**: Calculate the new means to be the centroids of the observations in the new clusters.

m^{(t+1)}_i = \frac{1}{|S^{(t)}_i|} \sum_{x_j \in S^{(t)}_i} x_j 

The below Matlab code illustrates this implementation.

%Read contents of file and perform K-means algorithm

%fid input file and fout is output file

%number of rows is to be determined

fid=fopen("WY.csv",'r');

fout=fopen("WYout.csv","w");

kmean=[zeros(11,1)];

tline=fgets(fid);

cnt=1;

while ischar(tline)

tmp=strsplit(tline,',');

val=str2double(tmp{1});

val2=str2double(tmp{2});

%disp(val);

kmean(cnt,1)=val;

kmean(cnt,2)=val2;

tline=fgets(fid);

cnt=cnt+1;

end

%disp(kmean);

X=kmean(:,1);

Y=kmean(:,2);

disp(size(X));

disp(size(Y));

plot(Y,'\*');

%plot(kmean,'\*');

resul=kmeans(kmean,3);

%disp(resul);

C=[kmean,resul];

disp(C);

%fprintf(fout,"%d,%d,%d",C);

dlmwrite(fout,C,',');

fclose(fout);

**TESTING**

Results and Snapshots

The result of the above analysis is generation of a website where users can view data obtained with regard to the originally listed program objectives.

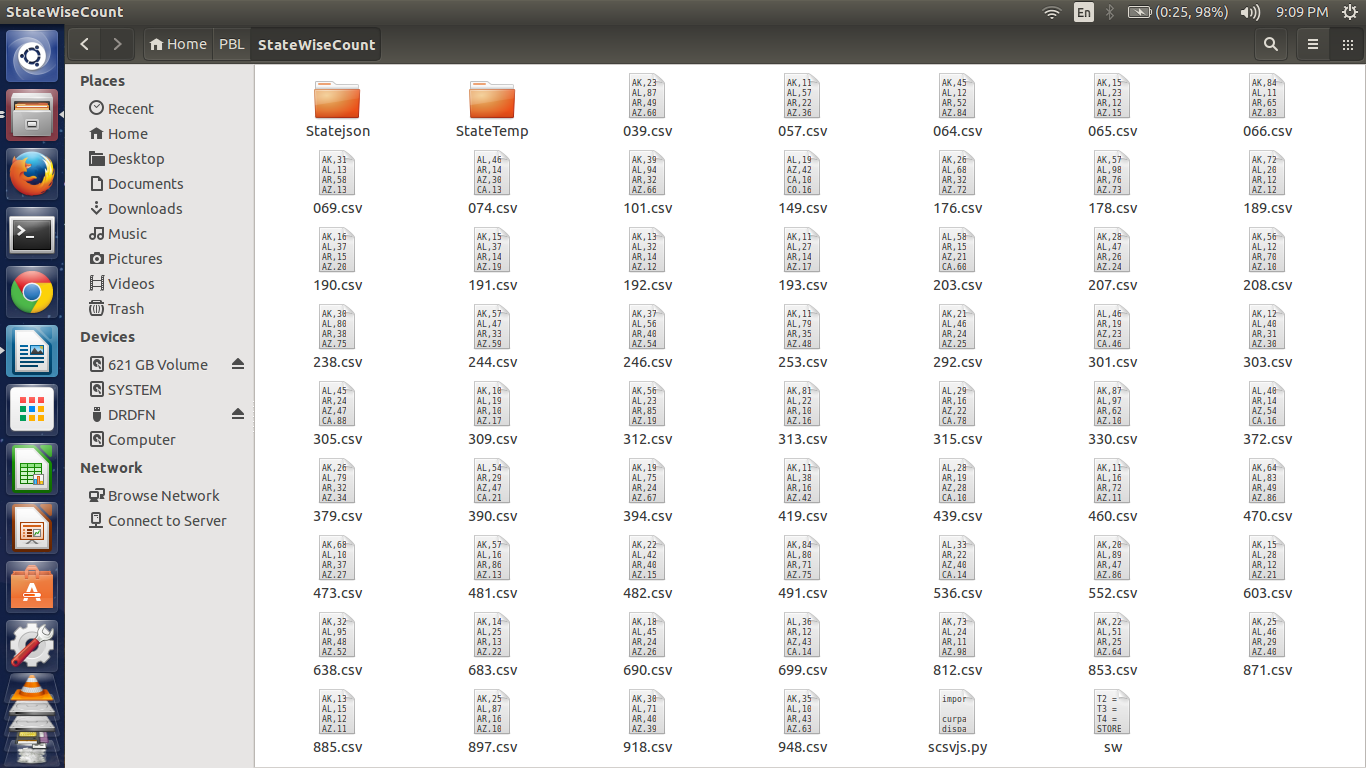


Fig 4:StateWise Count for a collection of Procedures

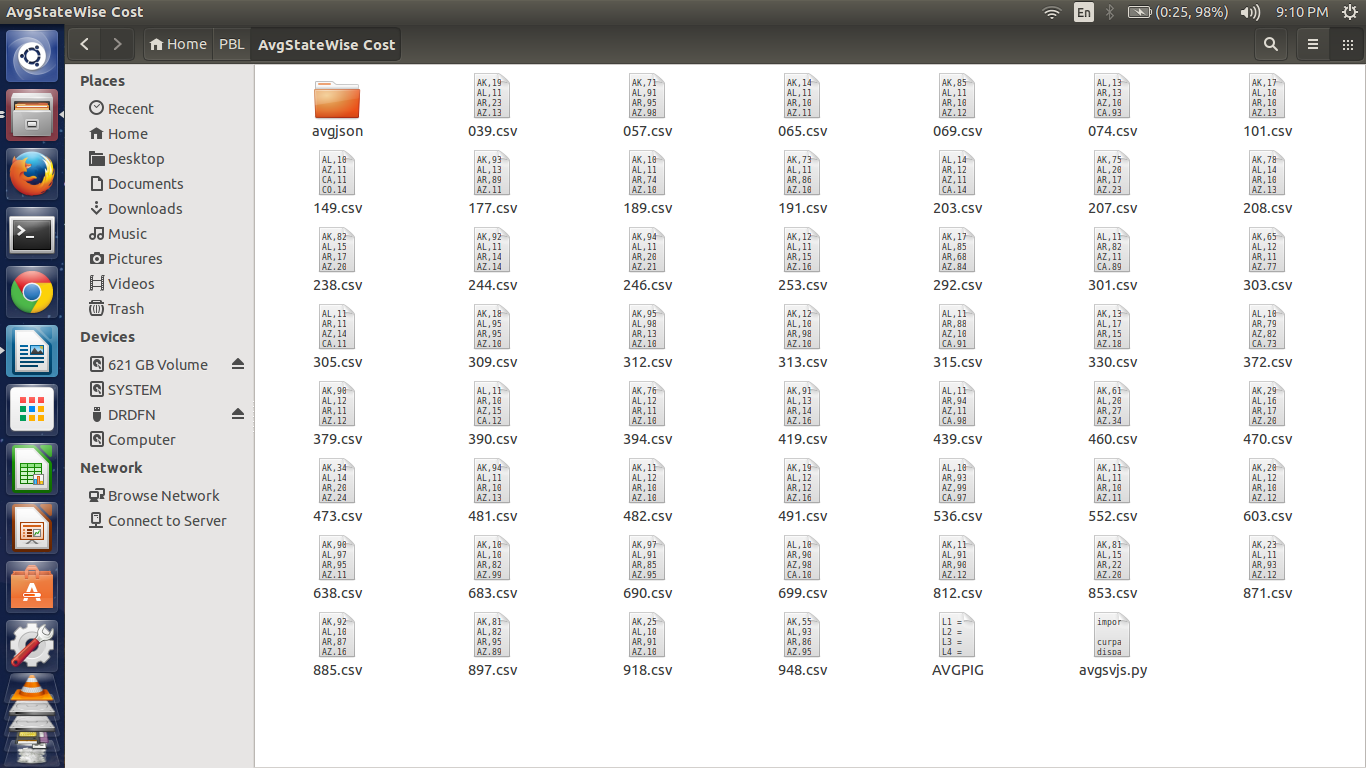


Fig 5:Average Cost per State for a collection of Procedures

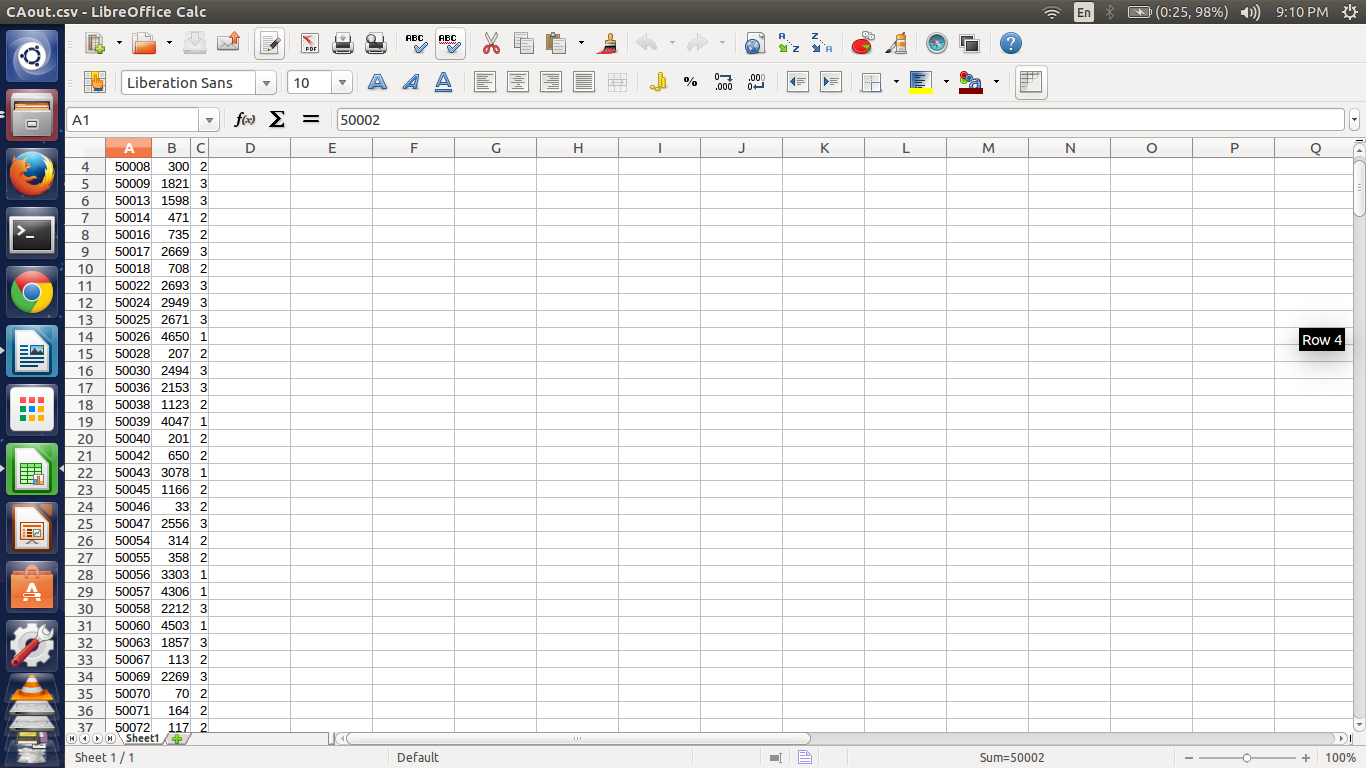


Fig 6:Sample output of hospitals clustered based on a State

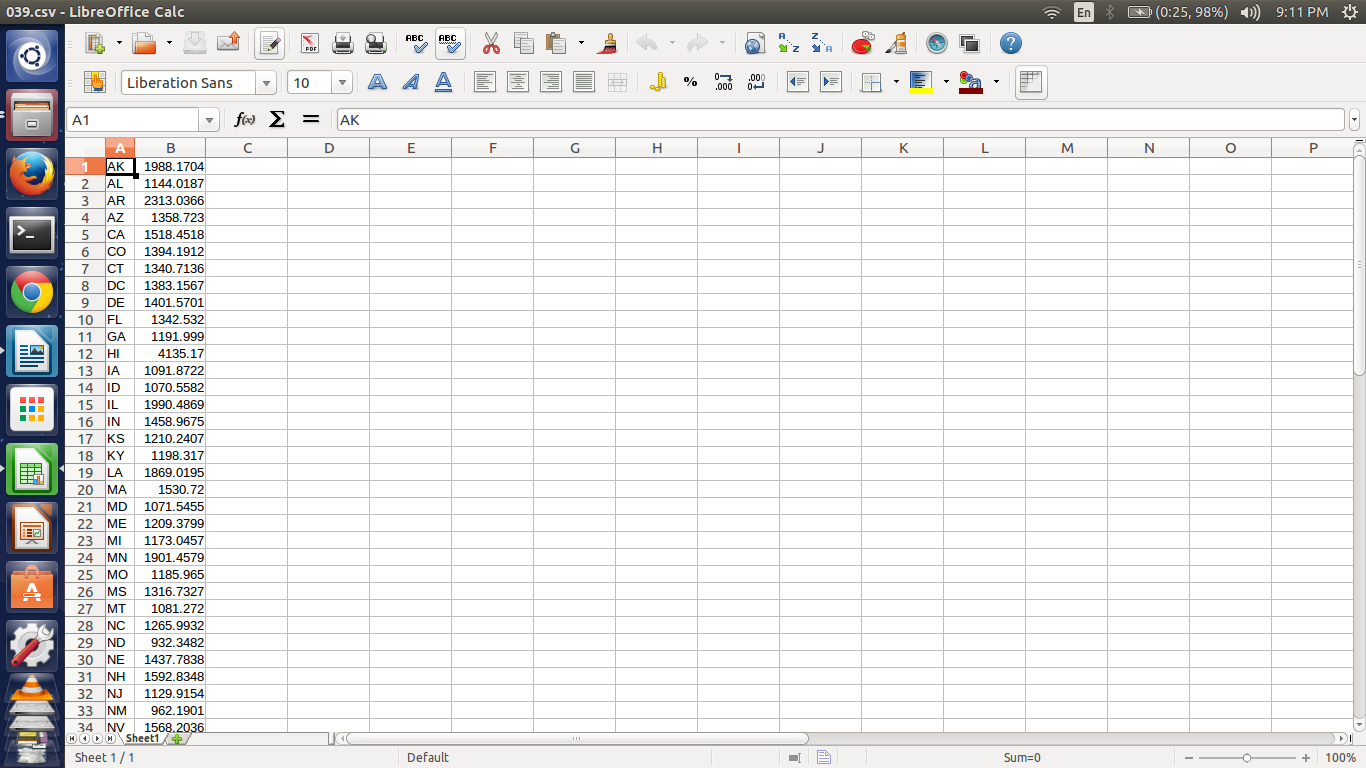


Fig 7:Sample output of average cost per state for a procedure



Fig 8:Sample output of number of cases per state for a procedure



Fig 8:Home Screen of Website

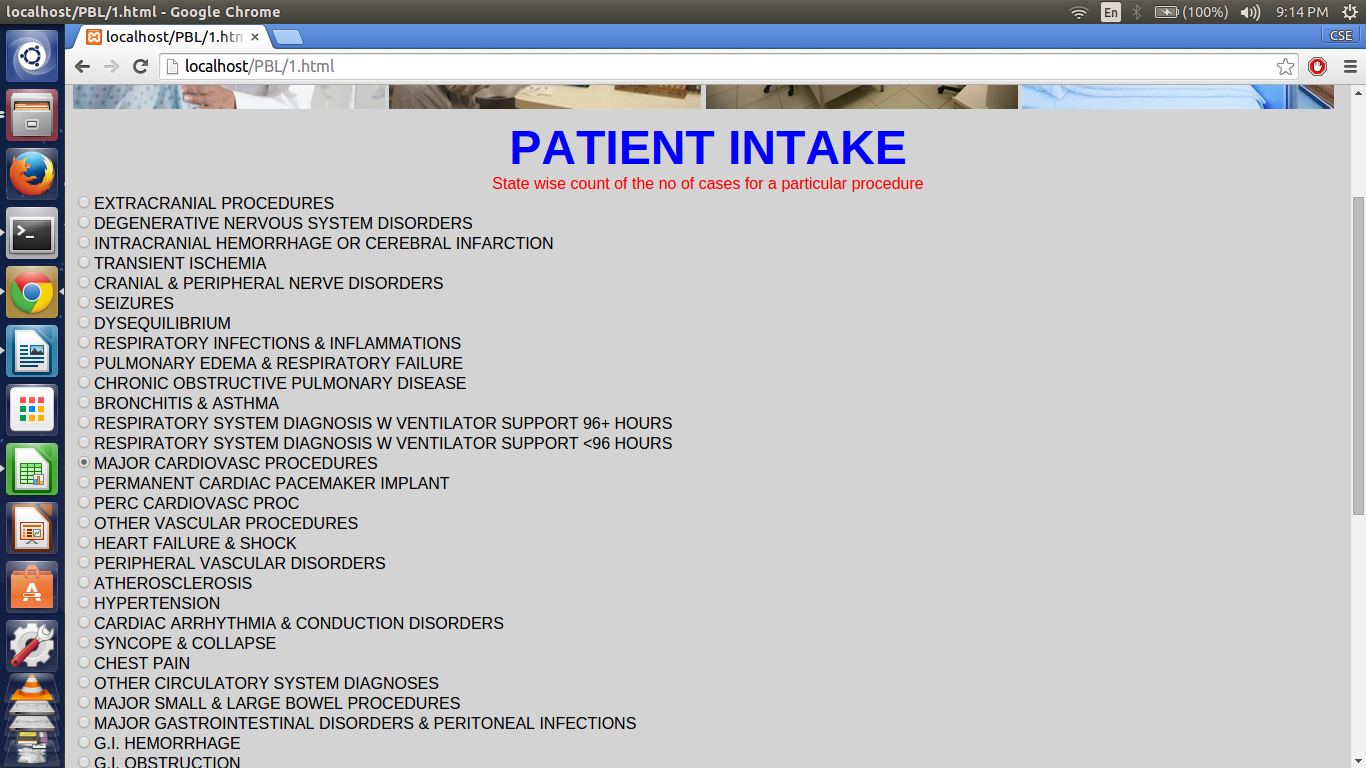


Fig 9:Patient Intake Landing Screen

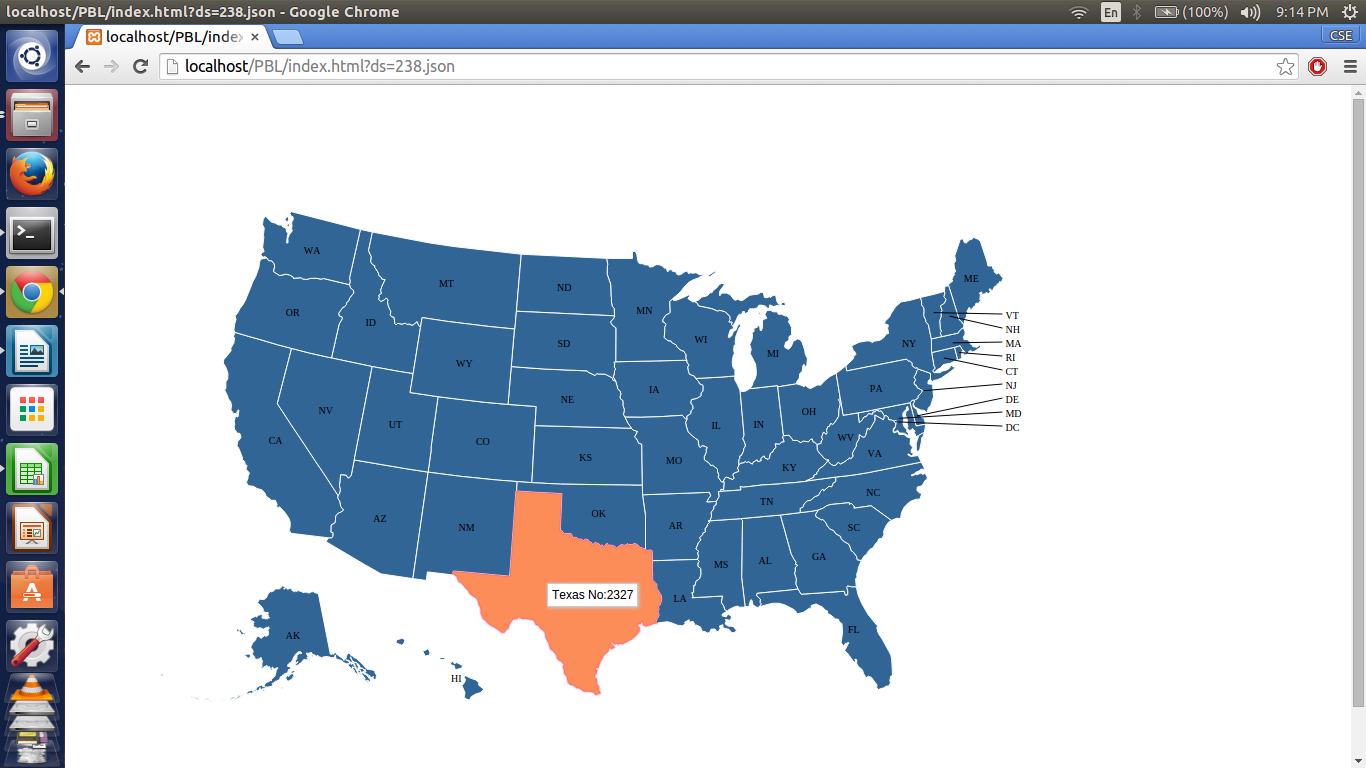


Fig 9:Patient Intake Result



Fig 9:Average Payment Landing Screen



Fig 9:Average Payment Result



Fig 10:Hospital Popularity Landing Page

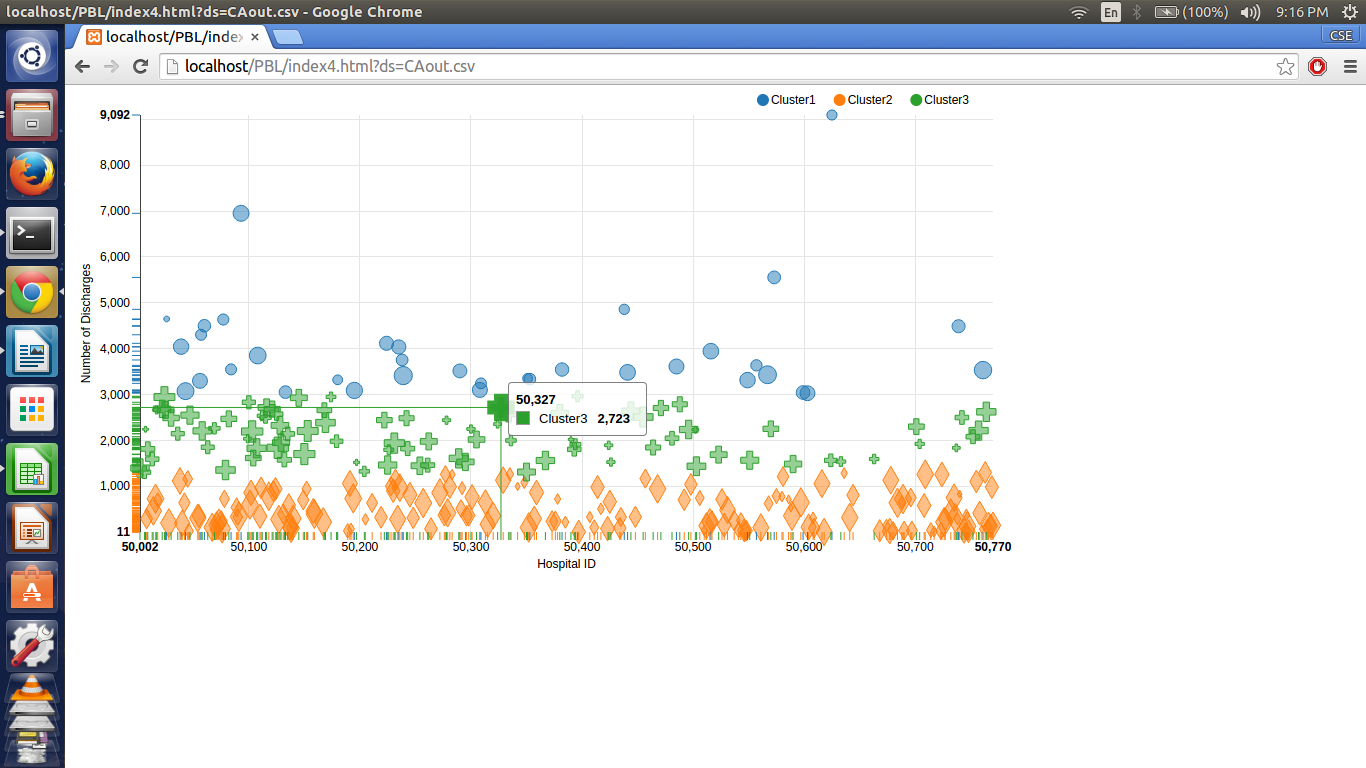


Fig 10:Hospital Popularity Results

1. CONCLUSION & SCOPE FOR FUTURE WORK

After successfully implementing the above project we arrive at a conclusion, data analytics on Medicare payments is a fruitful analytical endeavor. During the analysis we observed noticeable variations in the number of cases as well as average payment made by patients for a particular procedure across different states. These variations benefit policy makers and citizens and improve the quality of their decisions. The clustering of Hospitals in a particular state based on the number of discharges helps officials into classifying them as high, medium and low priority hospitals.

The future scope of this project is to include more data sets covering fiscal years post 2011; this will expand the analytical reach of the data as predictive algorithms can be applied. Another major area for growth is to implement the analytical model in the Indian demographic to investigate central and state government health schemes such as CGHS, Rajiv Arogya Bhagya respectively.

References

1. Medicare : https://en.wikipedia.org/wiki/Medicare\_(United\_States)
2. Profit of Health Care: http://marginalrevolution.com/marginalrevolution/2009/09/how-profitable-are-health-insurance-companies.html
3. Social Security :https://www.ssa.gov/medicare/
4. Pig: https://pig.apache.org/docs/r0.7.0/tutorial.html
5. Octave: https://en.wikibooks.org/wiki/Octave\_Programming\_Tutorial
6. Cluster Analysis: https://en.wikipedia.org/wiki/Cluster\_analysis
7. D3: https://github.com/mbostock/d3/wiki/Tutorials
8. Hadoop: http://www.sas.com/en\_us/insights/big-data/hadoop.html
9. Spark: https://en.wikipedia.org/wiki/Apache\_Spark
10. Python: http://www.tutorialspoint.com/python/
11. MapReduce: https://en.wikipedia.org/wiki/MapReduce
12. HDFS: http://hortonworks.com/hadoop/hdfs/
13. K Means: https://en.wikipedia.org/wiki/K-means\_clustering
14. JS Visualizations: http://www.sitepoint.com/twelve-javascript-libraries-data-visualization/