Data Analytics on Medicare Payments

Sriram Anupindi,Suhail T N **,** Mayank Sharma,Ashresh Prasad

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..

Keywords: Medicare, Hadoop, Map-Reduce, Apache Pig, D3.js

1. **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 damages. 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]

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.
4. **Design**

The design of the system is best described in the form of its architecture diagram.

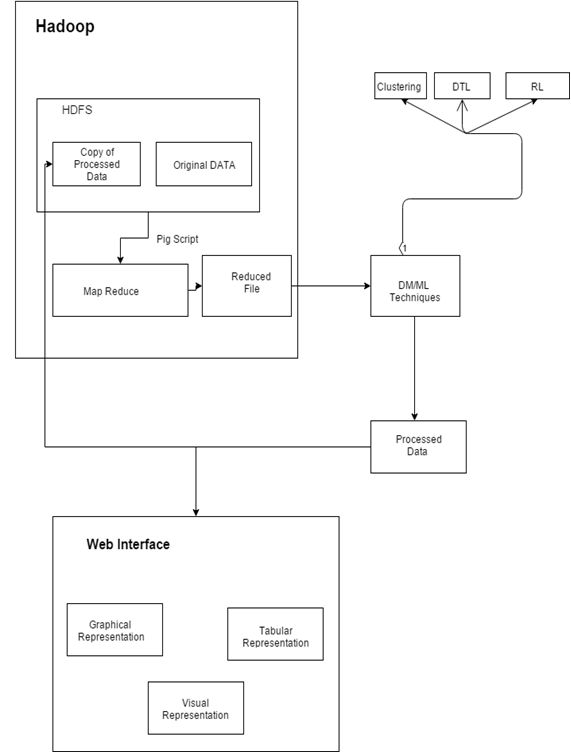


Fig1.Architecture Diagram

The system can be divided into various modules, where each module has a specific purpose, objective and task to accomplish. The modules in the system are:

**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.

Following the design of the system we need the design the algorithms to be utilized. The algorithms that we plan to utilize are:

**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.

The following diagram will highlight the flow of the data during the execution of the project.

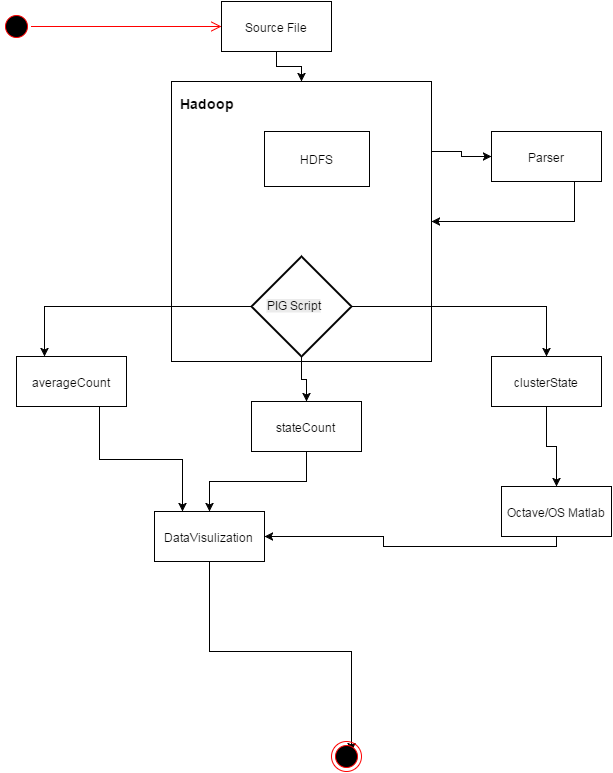


Fig2: Data Flow Diagram

1. **Implementation**

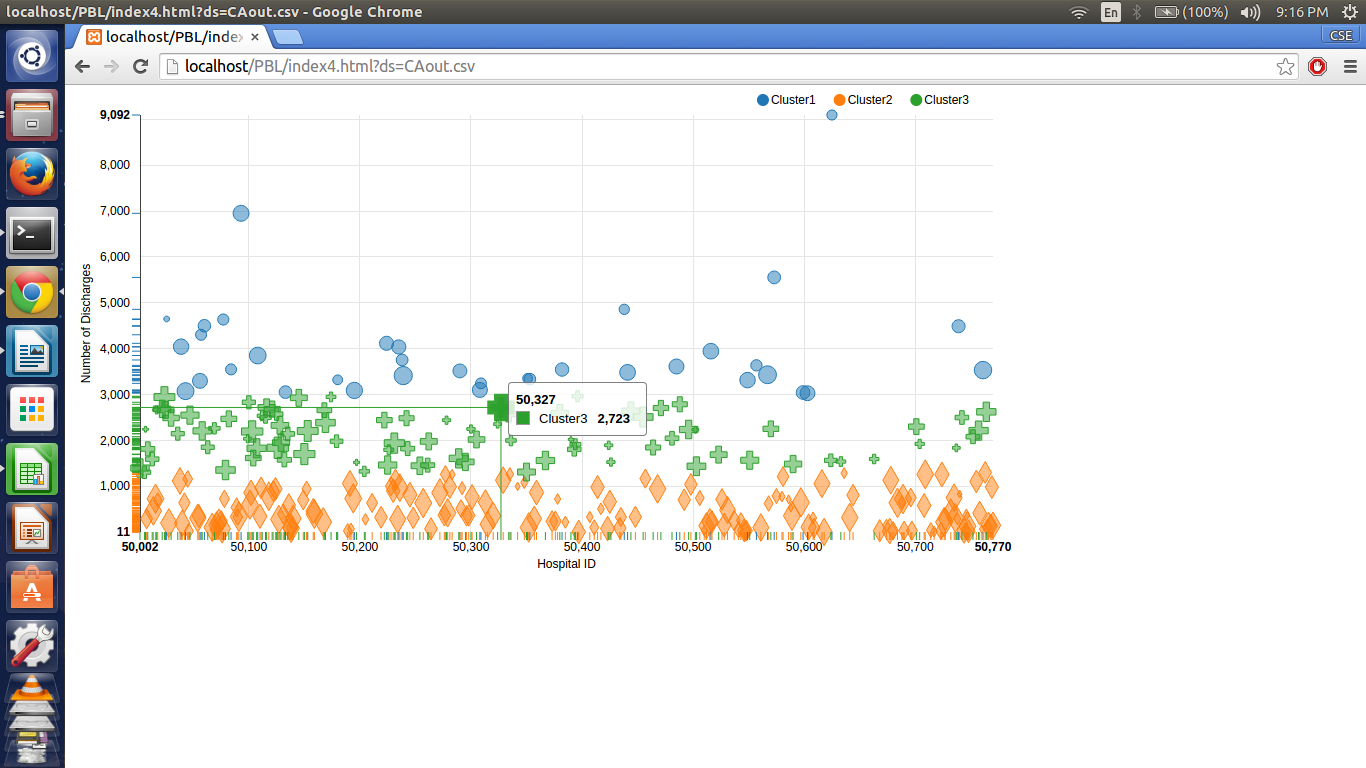
The tools used for the implementation of this project are many, varying from storage and computation of data to visualization of data. Below paragraphs will highlight the importance of each tool that is used in this project.

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, high-level programming language.[20][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.

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

1. **Results**

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. The results obtained on analysing the data are:

* State Wise comparison of Number of Cases for a particular procedure [Figure 3].
* State Wise comparison of Average Payment for a particular procedure [Figure 4].
* Clustering of Hospitals based on Discharges for a given state [Figure 5].

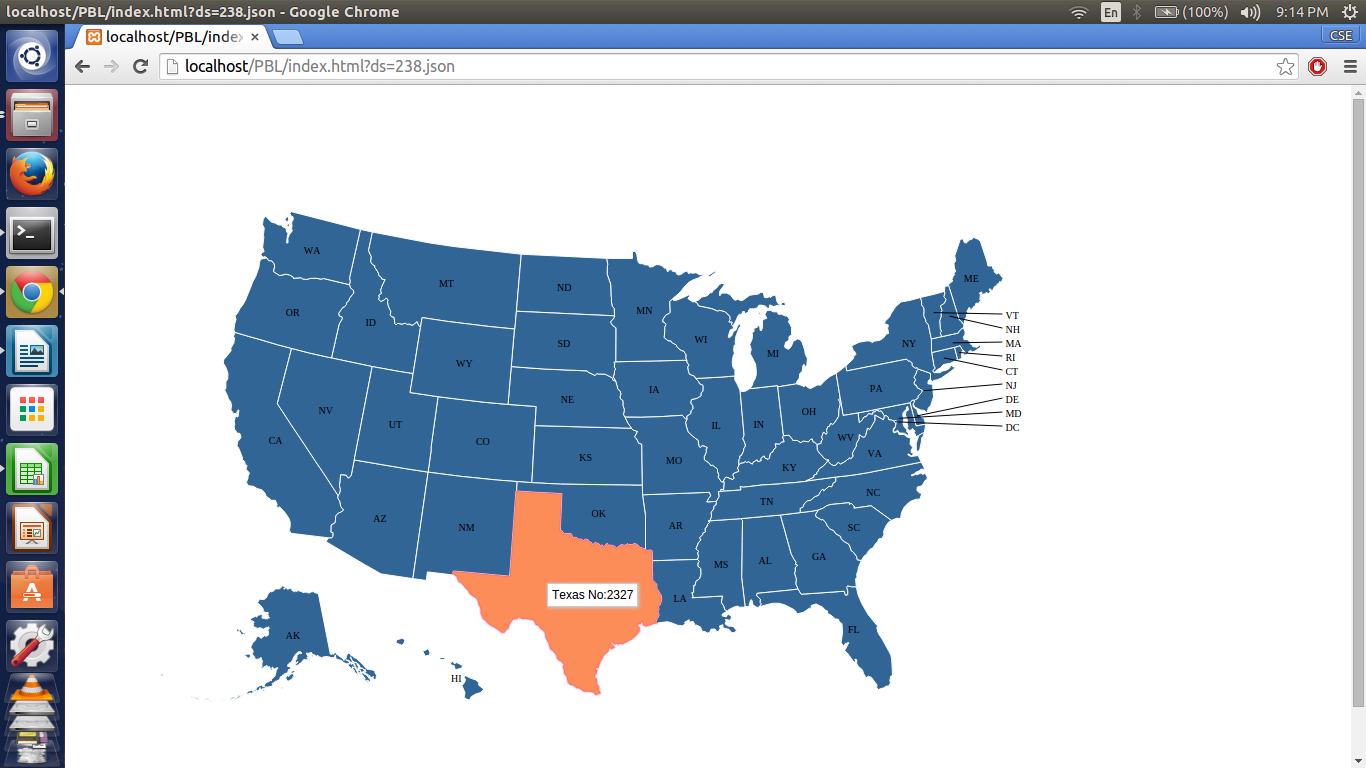


Fig 3:State Wise Count

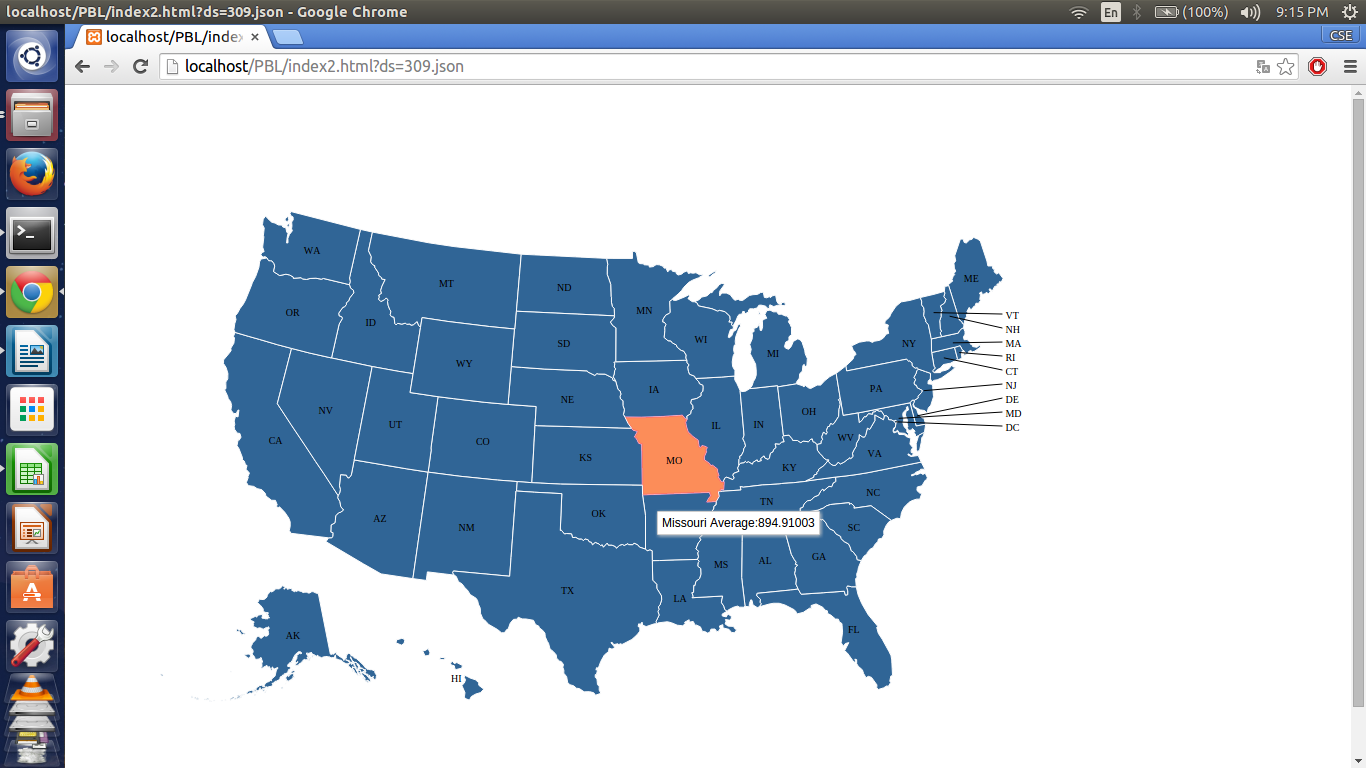


Fig4: State Wise Avg Payment

Fig5:Cluster for a State

1. Scope and Future work

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.

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.

1. Conclusion

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

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