**Analysis of Health Insurance Marketplace using Hive and Pig**

**Kulwinder Kaur, Jalashri, Moath**

**Prof. Jongwook Woo**

**Grad student Department of Computer Information System**

**California state University, Los Angeles**

[kkaur7@calstatela.edu](mailto:kkaur7@calstatela.edu)**,** [jraval@calstatela.edu](mailto:jraval@calstatela.edu)**,** [maldarr](mailto:maldarr@calstatela.edu)@calstate.edu

**Abstract:** A large scale data set facilitates the transparency of analysis and visualization. In this term paper, we delve into the Health Insurance Marketplace data using Hive and Hive Ql codes to better understand the average rates of Health plans, average rates of dental plans only in particular state, comparing the rates and comparing the tobacco rates for different age groups for different years.

1. **Introduction**

Health Insurance Industry is one of the biggest and all time booming industry across the word. But, there are many ambiguities and dilemmas for common people in understanding which plan is most suitable and how much average premium should be paid, which criteria should be considered for which age-group. The Insurance is expensive in US, also, it is not easy for the researchers and stake-holders to get the access to marketplace. This project is to bring an idea and show the right directions to stake-holders, analysts and common people to choose QHP (Quality Health Plan). CMS (centre For Medicare & Medicaid Services) and CCIIO (Consumer Information & Insurance Oversight) are committed to increase transparency in Health Insurance marketplace. The Centres for Medicare & Medicaid Services, CMS, is a part of the Department of Health and Human Services (HHS). They have published PUFs (Public used Files) which are available for all and provides the authentic information for rates and various network plans. The data considered under research are for years: 2014, 2015 and 2016. The research paper only includes the data from state partnership, marketplaces & state based marketplace that rely on the federal information technology .

The huge variety of data availibity led us to dive into this data and analyze it. In this research paper the main analysis is based on Average rate between various network and insurance premium for specific age group, average rate of different states for different ages using hive commands, average rate for 2016 in various states, Tobacco v/s non-tobacco rates analysis using Tableau, different network plans and average premium rates for different states, for particular age-group in specific year and also analysis for stand alone dental plans and the average premium paid for it which shows the cheapest and also most expensive plan in a particular year for particular age-group and for particular state. This analysis will make it easy for the customers to select cheapest and quality health plan and it will also be advantageous for stake-holders. To make this research valuable for the public, stakeholders and businessmen we did three analysis.

Big Data is the data available due to rapid increase in data available on websites and social networking sites. Big data is part of our lives now. People are utilizing this data to make better understanding of the problems. It is expensive to store and process big data. One of the cheapest solution is Hadoop. Hive works on Hadoop framework in parallel. Hive QL is easy to use.

1. Specifications

2.1 Data

There are seven PUFs available for Rates, Plan attribute, Benefits cost sharing, Service area, Plan ID crosswalk data, Quality data, Network data, business rules and many more. We have used Rate data dictionary and Network Data Dictionaries only for years 2014,2015 and 2016. The data on which the research is done is near 750 MB.

**2.2 Technical**

Each file is 701MB and the total size of this data is more than 3.5GB. We are using Bluemix Biginsight platform which is an open standards platform for building, running and managing apps. There are two nodes, Management node and Data node where the RAM is 48GB and 24GB respectievely and the data disk is 1TB. The file format for the data we used is CSV (comma separated values). The research has been done with the help of Hive and Pig. For the visualization we have used Excel, Powerview and tableau. Apache Hive is a data warehoude built using Hadoop for providing data summarization, query and analysis. The advantage behind using Hive is; data analyzed is stored in HDFS which provides scalibility, redundancy, etc., and easy to use while being similar to SQL. This research has also used Pig, a high level language that runs on Apache Hadoop. Pig can execute its Hadoop jobs in Mapreduce and Apache Spark. We have used Powerview, Tablue and excel for our visualizations. Powerview is a feature of Excel used for interactive data exploration, visualization and presentations, while, Tablue being a sofware used for visualization mainly focused on Business Intelligence. The FTP is Winscp, a free open source File transfer protocol for windows, transfers files between local computer and remote computer securely. For secure transfers, it uses secure shell (SSH) and supports SCP protocol in addition to SFTP.

1. **Implementation**

This research focused on the Average of the rates paid by the people in a particular age group of that state in year 2014 and 2015. This research supports the comparison of rates of healthplans provided by a particular state in year 2014 and 2015. The insights consists of average rate of a year, comparison of average in year 2014 and 2015, Average rate of dental only plans of a state, average rates of health plans in 2016, average rates of tobacco and non-tobacco in year 2016. As a result of this anlysis, it broadens the oppurtunity of choosing the health plans by customers, stakeholders and vendors. To make the files usable we converted these files to csv files and then upload. There was a file that contained unknown values like 99999, 99, 0 , for removing those values we used filter.

* 1. **Analysis of Averate Rate Analysis of 2014 and 2015 and comparison of average rates**

In order to do average rate analysis, we needed Rate\_PUF file from 2014, we used wget to upload the files and this file we uploaded to HDFS using put command Same procedure is adapted for uploading and storing other files. All files are stored into one projectfile folder in which we create different directories.

* 1. **Features of Health Insurance Database:**

The Health Insurance Database has following tables:

rate2014, rate2015, r2014, r2015, net2014, n2014, avgout2014, avgout2015, pro2014, proout2014, n2014\_dent, pro2014\_dental and d2014. Each of these tables have specific set of columns which reflects the necessary information. Each rate2014 and 2015 tables have near 25 columns, net2014 has 14 columns, r2014 and r2015 have 8 columns each, avgout2014 and avgout2015 have 3 columns, n2014 has 4 columns, pro2014 and proout2014 have 2 columns each, n2014\_dent has 5 columns, pro2014\_dental has 2 columns and d2014 had 2 columns. Research is done by using HDFS and HIVE and one of the visualization is expplained in detail as follows:

* 1. **Features of Hive Database:**

As data is stored in the HDFS of the system, all tables are created using HIVEQL:

CREATE EXTERNAL TABLE IF NOT EXISTS rate2014(

BusinessYear string,

StateCode string,

IssuerId string,

SourceName string,

VersionNum string,

ImportDate string,

IssuerId2 string,

FederalTIN string,

RateEffectiveDate string,

RateExpirationDate string,

PlanId string,

RatingAreaId string,

Tobacco string,

Age string,

IndividualRate int,

IndividualTobaccoRate int,

Couple int,

PrimarySubscriberAndOneDependent int,

PrimarySubscriberAndTwoDependents int,

PrimarySubscriberAndThreeOrMoreDependents int,

coupleAndOneDependent int,

CoupleAndTwoDependents int,

CoupleAndThreeOrMoreDependents int,

RowNumber string

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE LOCATION '/user/sskk04/projectfile/2014/rate/Rate\_puf2014'

TBLPROPERTIES ('skip.header.line.count'='2');

After this a table r2014 is created using rate2014 with the necessary fields and eliminating fields with unknown values:

CREATE TABLE r2014 As

select BusinessYear,

StateCode, IssuerId,

PlanId,

Tobacco,

Age,

IndividualRate,

IndividualTobaccoRate

from rate2014

where IndividualRate != 999999 and IndividualRate != 99 and IndividualRate != 0;

We have created table for 2015 in the same way described above.

Then the average of the rates is calculated with avg() function in hive of particular age group. The code for creating those tables is as follows:

* 1. **Query 1- Averagerate of 2014 & 2015 for all age groups:**

Create table avgout2014 As

Select Statecode, age, avg(IndividualRate) FROM r2014 GROUP BY Statecode,age;

Create table avgout2015 As

Select Statecode, age, avg(IndividualRate) FROM r2015 GROUP BY Statecode,age;

* 1. **Put the output in a location in under the username for downloading file in excel:**

INSERT OVERWRITE LOCAL DIRECTORY '/home/sskk04/temp1' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT \* FROM avgout2014;

According to the data needed, the 9 more tables are created using create. For selecting the columns from the tables we have used select sattement. The join is also used for r2014 and n2014 to create a relationaship between these two tables and using unique key IssuerId. The result of these two tables resulted in Healthplan rates provided by different providers. We also concluded the averagerates of Dental only plans also by different providers.

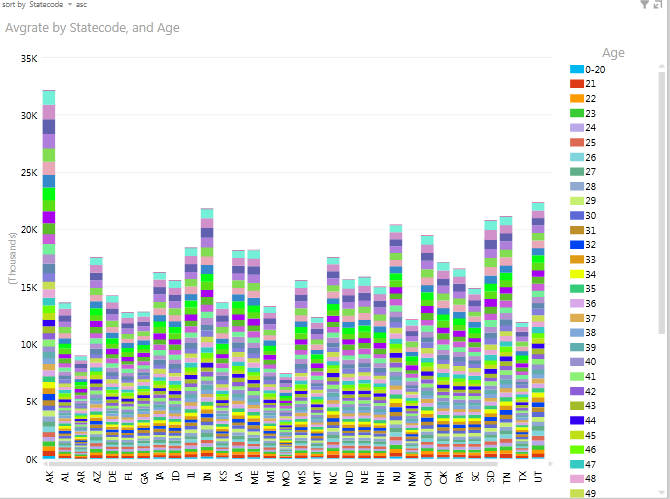
1. **Visualization :**

For 2014 After creating the tables and running the queries, In excel the file is retrieved by first downloading from Winscp and saved as excel file.

In excel file is opened for visualization. For the visualization, open file in Powerview and the output is as follows:

* 1. **Average rates of different age groups in different states in year 2014**

**Figure1**

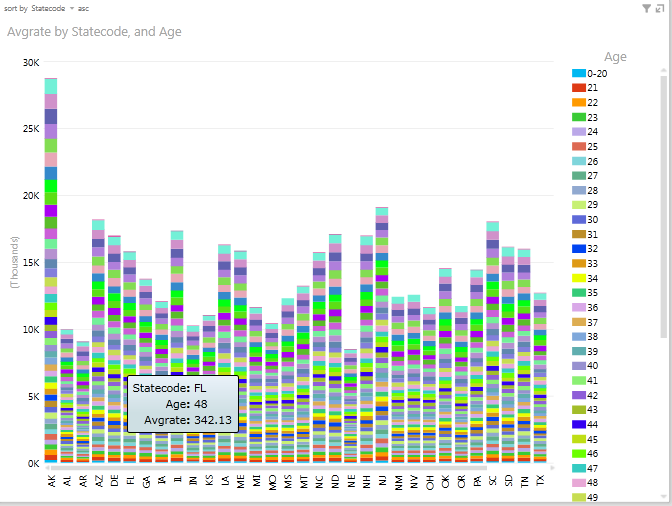


This visualization shows the average rates of different age groups in year 2014. The average rates each age group can be viewed by moving the cursor.

With the availibility of this graph, the summarization of the rates of different ages in different states will be easy for the customers, stakeholders, vendors as well as providers.

**4.2 Average rates of different age groups in different states in year 2015**

**Figure2**

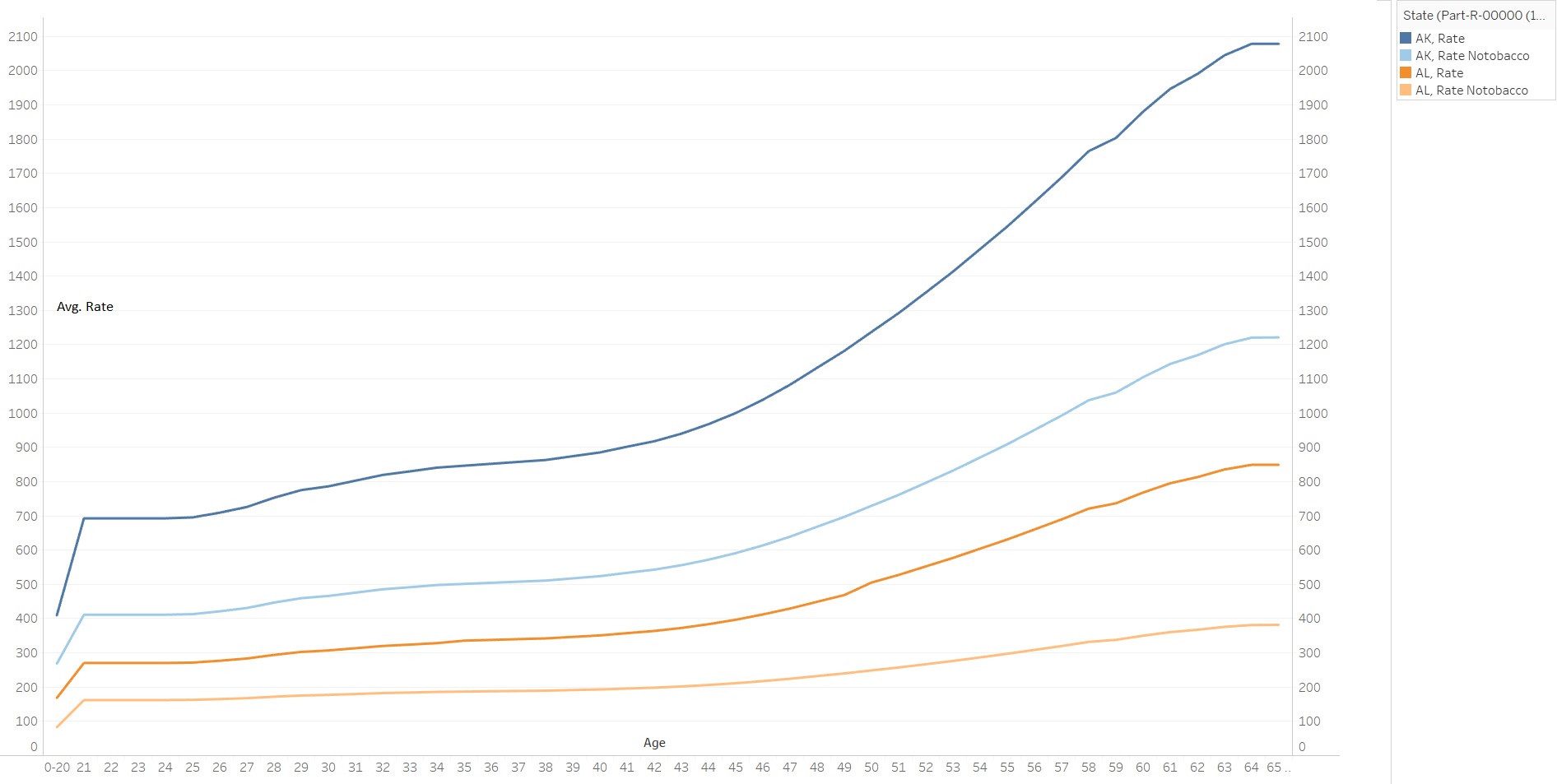


In the above, the visualization of averaage rates of different age groups in different states in year 2015 is shown. For Florida in age 48, its showing average rate $342.13. With these two figures, Figure 1 & Figure 2 , the average rates of 2014 and 2015 can be compared.

For age group 0-20, the rates in Alaska and Missouri changed very little. There are some states where the rates changed which we can visualize using this graph. Alaska has the highest rate $249.94 and Missourie has lowest rate $65.84 in 2014. For age group 0-20, Alaska has highest average rate $249.94 and Wisconsin has the second highest  $174.74. Missouri has the lowest average rate $65.84 and Arkansas has the second lowest $71.76 in 2015.

* 1. **Analysis of Average rates for 2016**

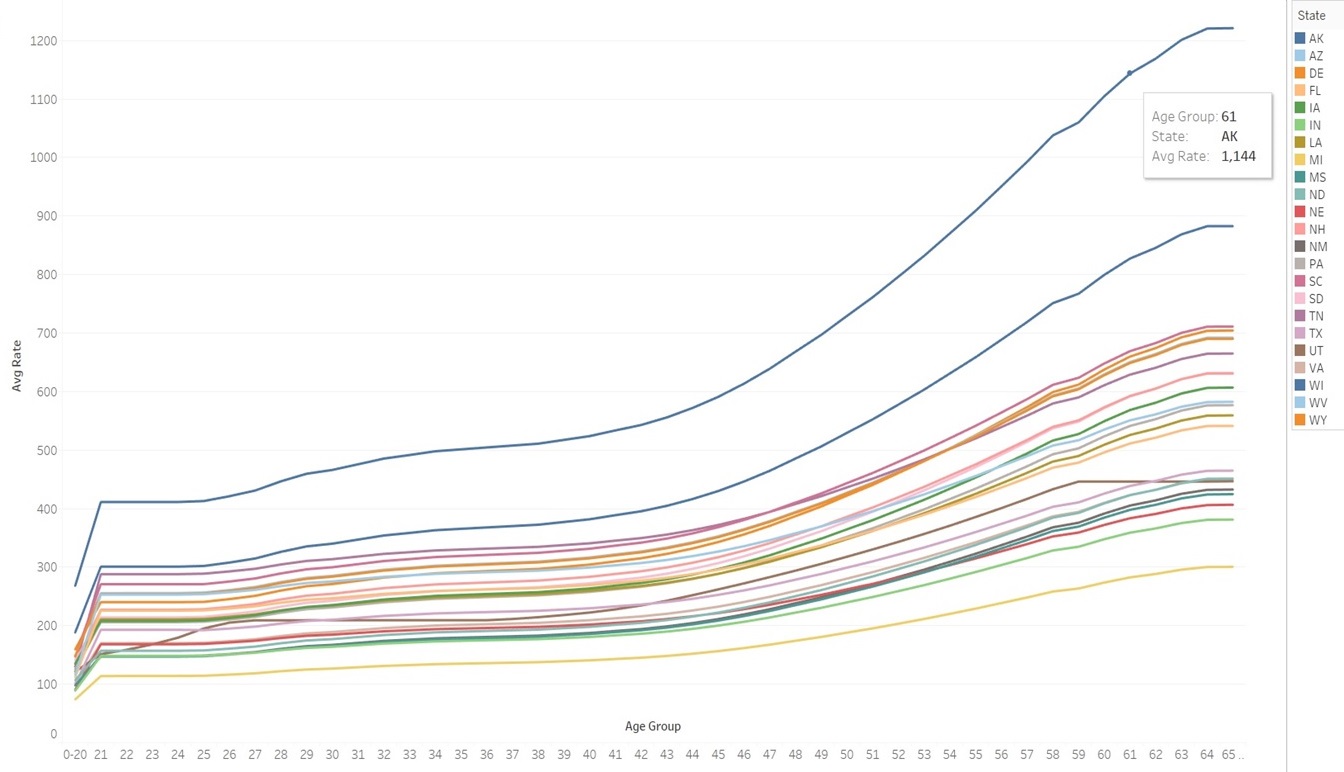
For 2016 rates analysis, our team decided to use a different tool, for educational purposes. Thus, we used Pig Latin language to processes the data that we have, and Tableau to visualize the results. After importing the 2016 data from the csv file, we started our analysis. So, we created a relation that contains all the field and the entries that were provided in the CMS website. Then, we started selecting certain columns that are necessary for conducting the analysis. After some mathematical calculations and grouping steps, we ended up with some interesting results. To make these



results easier to read and compare, we exported the relation as an excel file as a preparation for the visualization part. In Tableau, we imported the results from the excel file as a

a Data source ending up with a clear graph showing the price trends in different states grouped by age group.

**Figure 3 Average Insurance Rates 2016**



As shown in (Figure3), we can see that Alaska holds the highest monthly rate over all the age groups. On the other hand, Michigan has the lowest rate for the same age groups. It is important to notice that there are some crucial points in this graph where the average rate line intersects with the other state lines. These points must be covered by researchers and financial counselors to see the reasons that might cause this unpredicted fluctuation. For example, from the figure below we can see one of these fluctuation points at age group 25 between Texas and Utah where the rate in Texas before this point was higher than Utah, $193 and $180 respectively.

**4.4 Tobacco effect on the average rates in 2016**

After looking at the data that we have, we noticed a major factor that could affect the monthly health rate dramatically; it was the Tobacco usage. To see the damages on the overall price, we needed to pass the data to Pig again[[1]](#endnote-1) and then to Tableau for a thoughtful analysis. The results were shocking; around 72% increase in the price on average for all states. In (Figure 4) we picked only two states Alaska with the blue color and Alabama with Yellow color; The darker line represents the price with tobacco use while the light one shows the price without tobacco use. And it is clear from the graph the impact of the Tobacco use on the price.

1. **Conclusion:**

This research of health insurance marketplace led us to the conclusion that large availibility of data and interactive visualization of analysis is significant an helpful for the people who are seeking for the quality and affordable healthplans. The research is also beneficial for the stakeholders and vendors also. The transperancy in this data and analysis is high. The level transperancy in this data and analysis is high.

1. **References**:

[1] "Health Insurance Marketplace Public Use Files," in *The Centers for Medicare & Medicaid Services*, 2016. [Online].

Available: https://www.cms.gov/CCIIO/Resources/Data-Resources/marketplace-puf.html. Accessed: Dec. 9, 2016.

[2] The Pig code is available for the public access in the github project folder <https://github.com/kkaur7/healthinsurancedata>

1. [↑](#endnote-ref-1)