**[ Health Insurance Marketplace]**

Fall 2019

[10/20/2019]

Name Ashish Gir

UBID 1029647

Name Shalyrajsinh Jadeja

Table of Contents

1. Introduction 3

2. Problem Definition 4

3. Data Set 5

4. Modeling MapReduce Programming 6

5. Evaluation Plan 7

6. Project Plan 8

7. Conclusions 9

# Introduction

# *The Health Insurance Marketplace Public Use Files contain data on health and dental plans offered to individuals and small businesses through the US Health Insurance Marketplace.*

*This data was originally prepared and released by the*[*Centers for Medicare & Medicaid Services (CMS)*](https://www.cms.gov/cciio/resources/data-resources/marketplace-puf.html)*. Please read the*[*CMS Disclaimer-User Agreement*](https://www.cms.gov/CCIIO/Resources/Data-Resources/Downloads/Data-Disclaimer-User-Agreement.pdf)*before using this data.*

*The Centers for Medicare & Medicaid Services (CMS) Center for Consumer Information and Insurance Oversight (CCIIO) is committed to increasing transparency in the Health Insurance Exchange. While health plan information including benefits, copayments, premiums, and geographic coverage is publically available on Healthcare.gov, CMS also publishes downloadable public use files (PUFs) so that researchers and other stakeholders can more easily access Exchange data.*

*The Health Insurance Exchange Public Use Files (Exchange PUFs) are available for plan years 2014 to 2020 to support timely benefit and rate analysis. It is important to note that the 2020 Exchange PUFs will be updated regularly to reflect the plan data that consumers will see when shopping for an Exchange Qualified Health Plan (QHP). Data for the 2020 Exchange PUFs were imported to CMS systems by October 1, 2019.*

# 2. Problem Definition

*To get started, here are some data exploration ideas:*

* *How do plan rates and benefits vary across states?*
* *How do plan benefits relate to plan rates?*
* *How do plan rates vary by age?*
* *How do plans vary across insurance network providers?*
* *How do plan rates vary across state?*
* *How do plan benefits vary across state?*
* *How did plan rates change from 2014 to 2016?*
* *How do plan benefits relate to plan rates?*
* *How do plan rates vary by age?*
* *How do plan rates vary with additional family?*
* *How do plan rates vary between smokers and non-smokers?*
* *How do plans vary across insurance network providers?*
* *How can you analyze one source of variation in this data while controlling for the others?*

# Data Set

*This data was originally prepared and released by the*[*Centers for Medicare & Medicaid Services (CMS)*](https://www.cms.gov/cciio/resources/data-resources/marketplace-puf.html)*. Please read the*[*CMS Disclaimer-User Agreement*](https://www.cms.gov/CCIIO/Resources/Data-Resources/Downloads/Data-Disclaimer-User-Agreement.pdf)*before using this data.*

*Here, we've processed the data to facilitate analytics. This processed version has three components:*

### *1. Original versions of the data*

*The original versions of the 2014, 2015, 2016 data are available in the "raw" directory of the download and "../input/raw" on Kaggle Scripts. Search for "dictionaries" on*[*this page*](https://www.cms.gov/cciio/resources/data-resources/marketplace-puf.html)*to find the data dictionaries describing the individual raw files.*

### *Combined CSV files that contain In the top level directory of the download ("../input" on Kaggle Scripts), there are six CSV files that contain the combined at across all years:*

* ***BenefitsCostSharing.csv***
* ***BusinessRules.csv***
* ***Network.csv***
* ***PlanAttributes.csv***
* ***Rate.csv***
* ***ServiceArea.csv***

# A picture containing screenshot Description automatically generated4. Methods

# *Median Monthly Premiums by state :*

# *Filter Data to 2016*

# *Remove odd values from the IndividualRate column (those >=9000).*

# *Calculate the median “IndividualRate” across all plans offered in each state*

# *Append state names onto the dataset*

# *Plans and Carriers by state :*

# *Create subset of 2015 dataset*

# *Remove odd values from the IndividualRate column (those >=9000).*

# *By state distribution: We mainly want to see the number of insurance carriers that are available per state.*

# *Also the number of plans they offer by state. It shows the difference of accessibility by state.*

*On this paper we worked with Map-Reduce Programming and Spark. The Hive Execution was also done by importing the data into hive tables. Through Map Reduce and Spark the Median Monthly Premiums by state was done and obtained the results. The data was first separated into individual two columns in order to make the execution possible.*

*Each of the methodology is described in two of the following phases:*

*Map-Reduce:*

*Using the Hadoop map-reduce we can analyze the nature of the Insurance statistics.*

*Hive:*

*All of the columns from the data set were imported to the hive tables in order to execute the queries in sql format.*

*Spark:*

*With the spark we can separate a full data set into various tuples using spark context. We can grab the data of the name of the Insurance, Id of the data and the data from bias column to find out the nature of the insurance with further steps.*

*After defining the column in terms of tuples we can loop over all the tuples to get the data of individual in terms of columns specified in the tuple. Now we can map the values that are distinct ie Median Monthly Premiums by state, Plans and Carriers by state*

## Total Size of Files (.csv) = >10GB Method using AWS (Steps as given below)

## Create a File Data Set and an Analysis

Complete the following procedure to create a data set and an analysis:

1. Check [Data Source Limits](https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html) to make sure your target file doesn't exceed data source limits.
2. On the Amazon QuickSight start page, choose **Manage data**.
3. On the **Your Data Sets** page, choose **New data set**.
4. In the **FROM NEW DATA SOURCES** section of the **Create a Data Set** page, choose **Upload a file**.
5. In the **Open** dialog box, browse to a text file, select it, and then choose **Open**.

A file must be 1 GB or less to be uploaded to Amazon QuickSight.

1. Choose **Next**.
2. Choose **Visualize**.

# Adding a Spark Step

You can use Amazon EMR steps to submit work to the Spark framework installed on an EMR cluster. For more information, see [Steps](https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-overview.html#emr-overview-data-processing) in the Amazon EMR Management Guide. In the console and CLI, you do this using a Spark application step, which runs the spark-submit script as a step on your behalf. With the API, you use a step to invoke spark-submit using command-runner.jar.

**To submit a Spark step using the console**

1. Open the Amazon EMR console at <https://console.aws.amazon.com/elasticmapreduce/>.
2. In the **Cluster List**, choose the name of your cluster.
3. Scroll to the **Steps** section and expand it, then choose **Add step**.
4. In the **Add Step** dialog box:
   * For **Step type**, choose **Spark application**.
   * For **Name**, accept the default name (Spark application) or type a new name.
   * For **Deploy mode**, choose **Cluster** or **Client** mode. Cluster mode launches your driver program on the cluster (for JVM-based programs, this is main()), while client mode launches the driver program locally. For more information, see [Cluster Mode Overview](https://spark.apache.org/docs/latest/cluster-overview.html) in the Apache Spark documentation.

**Note**

Cluster mode allows you to submit work using S3 URIs. Client mode requires that you put the application in the local file system on the cluster master node.

* + Specify the desired **Spark-submit options**. For more information about spark-submit options, see [Launching Applications with spark-submit](https://spark.apache.org/docs/latest/submitting-applications.html#launching-applications-with-spark-submit).
  + For **Application location**, specify the local or S3 URI path of the application.
  + For **Arguments**, leave the field blank.
  + For **Action on failure**, accept the default option (**Continue**).

1. Choose **Add**. The step appears in the console with a status of Pending.
2. The status of the step changes from **Pending** to **Running** to **Completed** as the step runs. To update the status, choose the **Refresh** icon above the **Actions** column.
3. The results of the step are located in the Amazon EMR console Cluster Details page next to your step under **Log Files** if you have logging configured. You can optionally find step information in the log bucket you configured when you launched the cluster.

# 5. Evaluation Plan

*Data visualization is a method for displaying vast amounts of data in a comprehensible way. It’s considered to be modern equivalent of visual communication and involves both the creation and study of the visual representation of information. Essentially, complex data is abstracted to a relevant schematic form, including attributes or variables for specific groups or units of information. The visual representation allows for greater discernment of patterns, relationships, trends, and anomalies.*

*Using data visualization software from AWS Marketplace can help you easily analyze the vast amounts of data your business collects and generates. Additional benefits include.*

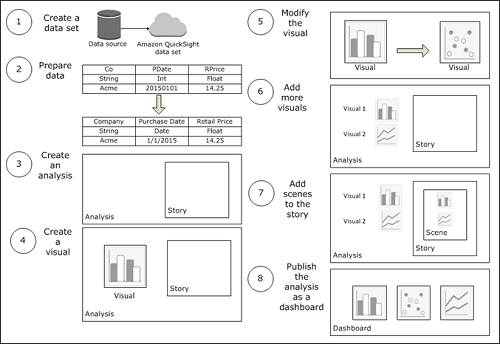
* *Being able to process information visually makes it easier to grasp information than sifting through spreadsheets and reports.*
* *Generating analytics without custom programming. Data visualization tools allows you to experiment with different scenarios by adjusting the selections.*
* *Collating data for analysis. Depending on what your data visualization tool can do, either you process the data first and assemble it in a single location, or your tool pulls data from its existing locations.*

# 6. Project Plan

*When you create an analysis, the typical workflow is as follows:*

1. *Connect to a data source, and then create a new dataset or choose an existing dataset.*
2. *(Optional) If you created a new dataset, prepare the data (for example, by changing field names or data types).*
3. *Create a new analysis.*
4. *Add a visual to the analysis by choosing the fields to visualize. Choose a specific visual type, or use AutoGraph and let Amazon QuickSight choose the most appropriate visual type, based on the number and data types of the fields that you select.*
5. *(Optional) Modify the visual to meet your requirements (for example, by adding a filter or changing the visual type).*
6. *(Optional) Add more visuals to the analysis.*
7. *(Optional) Add scenes to the default story to provide a narrative about some aspect of the analysis data.*
8. *(Optional) Publish the analysis as a dashboard to share insights with other users.*

*The following graphic illustrates a typical Amazon QuickSight workflow.*



# 7. Conclusions

The data set is fairly complex to run the analysis. But if we do it through the map reduce and spark methods we can categorize the data in various level and run the analysis to get the results as expected. This would need the setup of whole Hadoop environment and data has to be pushed to HDFS to run the analysis.

The Health Insurance Marketplace Public Use Files contain data on health and dental plans offered to individuals and small businesses through the US Health Insurance Marketplace and this Big Data Analysis helps us to find Median Monthly Premiums by state and Plans and Carriers by state. We credit the extensive data gathering on dataset with rich abundant and precise data to thank for, We conclude that we have understood to read and process data and to generate a valuable information.