Big Data Hadoop—Real Time Project— Insurance

Jayateertha M Tatti

joy.tat@gmail.com

26-Oct-2016

Analyze health reports across years for the US market

A US-based insurance provider has decided to launch a new medical insurance program targeting various customers. To help this customer understand the current realities and the market better. We have to perform a series of data analytics tasks using Hadoop. The customer has provided the data set.

Project goals:

1) Calculate the average % of people, aged between 18 and 64, who have obtained insurance from private players from 2001 to 2011
2) Calculate the average percentage of people, aged 65 years or more, who are solely covered by public insurance from the year 2001 to 2011

Data set in xls format from the URL :
http://www.census.gov/hhes/www/hlthins/data/utilization/tables.html
Note: I have downloaded from SimpliLearn download section.

Hadoop Architecture

- Vmware
- Linux Ubuntu 16.04 Lts
- Hadoop 2.7.2

```
admin@poorvi-HP-Pavilion-dv4-Notebook-PC:~$ hadoop version
Hadoop 2.7.2
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r b165c4fe8a74265c792ce23f546c64604acf0e41
Compiled by jenkins on 2016-01-26T00:08Z
Compiled with protoc 2.5.0
From source with checksum d0fda26633fa762bff87ec759ebe689c
This command was run using /home/admin/hadoop/share/hadoop/common/hadoop-common-2.7.2.jar
admin@poorvi-HP-Pavilion-dv4-Notebook-PC:~$ jps
4040 NodeManager
4410 Jps
3579 DataNode
3917 ResourceManager
3759 SecondaryNameNode
admin@poorvi-HP-Pavilion-dv4-Notebook-PC:~$
```

Approach

- Referred lesson 3 to load the dataset in HDFS

- Referred lesson 6 Pig to arrive at the below results detailed in following pages

Conversion xls to csv files

For conversion linux gnumeric was used

Input files are listed below

```
admin@poorvi-HP-Pavilion-dv4-Notebook-PC:~/Documents/SimpliLearn/Hadoop/Downlaod/projects/Insurance_Data_Set/data/back/excel$ ls
2001.xls  2002.xls  2003.xls  2004.xls  2005.xls  2009.xls  2010.xls  2011.xls
```

ssconvert -S 2001.xls 2001.csv

ssconvert created one csv per sheet. Hence 0 to 5 csv files for each xls files.
Refer the image below

```
admin@poorvi-HP-Pavilion-dv4-Notebook-PC:~/Documents/SimpliLearn/Hadoop/Downlaod/projects/Insurance_Data_Set/data$ ls
2001.csv.0  2001.csv.5  2002.csv.4  2003.csv.3  2004.csv.2  2005.csv.1  2009.csv.0  2009.csv.5  2010.csv.4  2011.csv.3
2001.csv.1  2002.csv.0  2002.csv.5  2003.csv.4  2004.csv.3  2005.csv.2  2009.csv.1  2010.csv.0  2010.csv.5  2011.csv.4
2001.csv.2  2002.csv.1  2003.csv.0  2003.csv.5  2004.csv.4  2005.csv.3  2009.csv.2  2010.csv.1  2011.csv.0  2011.csv.5
2001.csv.3  2002.csv.2  2003.csv.1  2004.csv.0  2004.csv.5  2005.csv.4  2009.csv.3  2010.csv.2  2011.csv.1  back
2001.csv.4  2002.csv.3  2003.csv.2  2004.csv.1  2005.csv.0  2005.csv.5  2009.csv.4  2010.csv.3  2011.csv.2
```

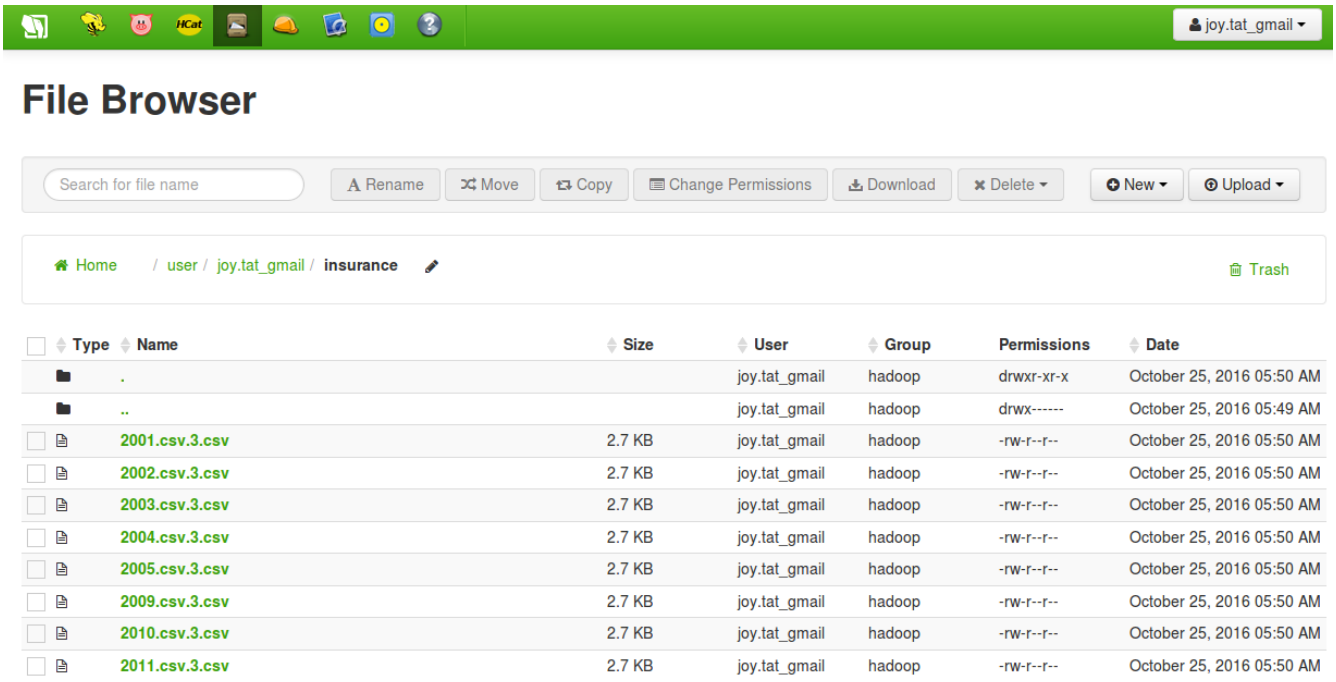
Data cleaning performed:

The original file has hierarchical data stored in csv format. The Tab4 (after split <*year*>.csv.3) contains the data to be processed. For a given requirement there multiple records are present. We need to fetch the specific parent child relation records. Hence I took data cleaning approach to analyze the data set. Read the parent value append it to child value (*logic listed in python code*). So that data becomes normalized. On this data set DML operations can be performed. Refer the image of insurance_clean_data.py to clean data. Once the data is cleaned upload to CloudLab for further processing using Pig.

```
import xlrd
import csv
import glob
import pandas as pd

def List_All_CSV_Files_From_Folder(filePath, fileType):
        allFiles = glob.glob(filePath + fileType)
        list_ = []
        for file_ in allFiles:
                #Call Append_Year_To_CSV method
                #Append_Year_To_CSV(file_, file_)
                clean_data(file_)
#Add the root data to all child data for processing
def clean_data(inputFileName):
        content = []
        total_per = []
        dataframe = pd.read_csv(inputFileName, usecols=[0,1], header=None)
        privious_data=''
        for index, row in dataframe.iterrows():
                data, per, i = row[0], row[1], index
                if data[0] == ' ':
                        data = privious_data + data.strip()
                else :
                        privious_data = data
                content.append(data)
                total_per.append(per)
        #Create DataFrame with required columns only
        CleanedDataFrame = pd.DataFrame({'Insurance_Description':content, 'Total_Per': total_per})
        #Create the csv file using required columns
        #print 'outputFileName : ' + outputFileName
        CleanedDataFrame.to_csv(inputFileName+'.csv', index=None)

def Process_File():
        #print 'Started Processing files..'
        filePath =r'/home/admin/Documents/SimpliLearn/Hadoop/Downlaod/projects/Insurance_Data_Set/data' # use your path
        fileType = '/*.csv.3'
        List_All_CSV_Files_From_Folder(filePath, fileType)

Process_File()
```

- Above python code converts *.3 files *.3.csv files
- Upload the *.csv.3.csv files to cloudlab in insurance folder
- Execute pig scripts on these files

Refer Uploaded CloudLab file image



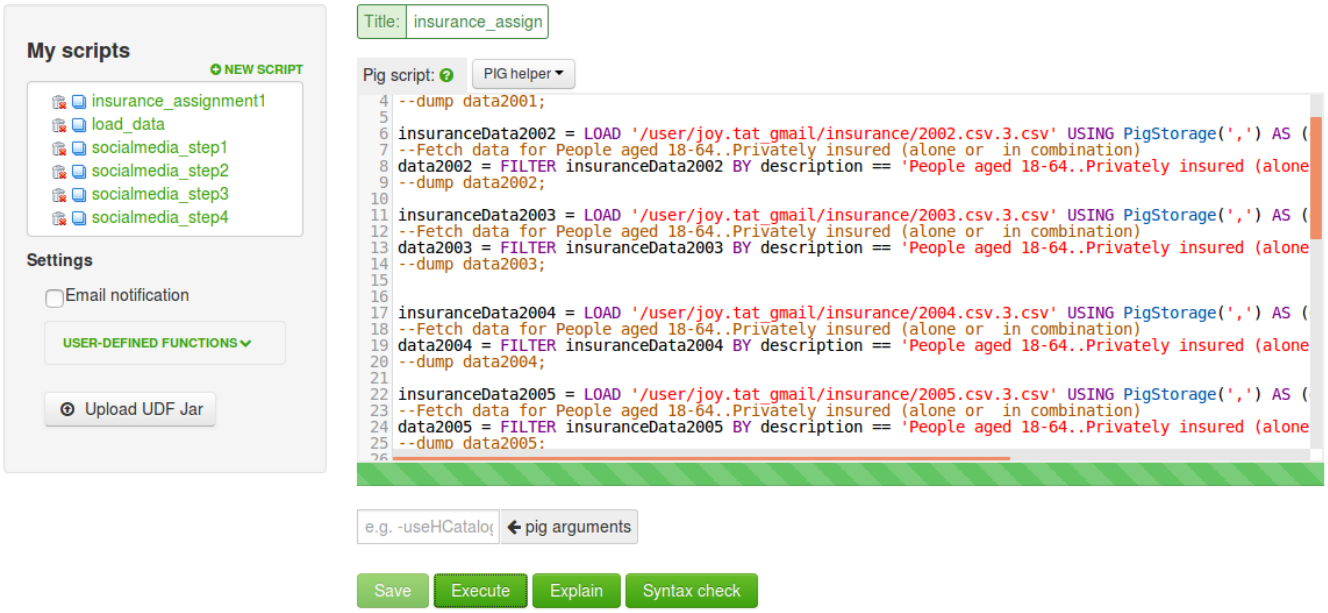
Analysis 1:

- Calculate the average % of people, aged between 18 and 64, who have obtained insurance from private players from 2001 to 2011
- Input the data in HDFS
- Use Pig to arrive at the conclusions
- criteria is : `'People aged 18-64..Privately insured (alone or  in combination)'`

Query is :

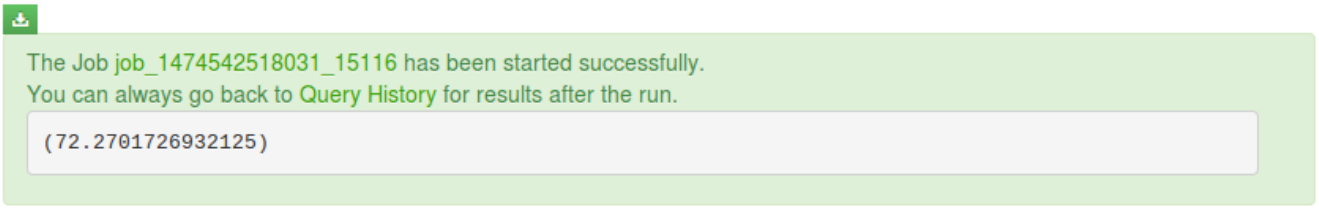
```
insuranceData2001 = LOAD '/user/joy.tat_gmail/insurance/2001.csv.3.csv' USING PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2001 = FILTER insuranceData2001 BY description == 'People aged 18-64..Privately insured (alone or  in combination)';
--dump data2001;

insuranceData2002 = LOAD '/user/joy.tat_gmail/insurance/2002.csv.3.csv' USING PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2002 = FILTER insuranceData2002 BY description == 'People aged 18-64..Privately insured (alone or  in combination)';
--dump data2002;

insuranceData2003 = LOAD '/user/joy.tat_gmail/insurance/2003.csv.3.csv' USING PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2003 = FILTER insuranceData2003 BY description == 'People aged 18-64..Privately insured (alone or  in combination)';
--dump data2003;


insuranceData2004 = LOAD '/user/joy.tat_gmail/insurance/2004.csv.3.csv' USING PigStorage(',') AS (description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2004 = FILTER insuranceData2004 BY description == 'People aged 18-64..Privately insured (alone or  in combination)';
```

```
--dump data2004;

insuranceData2005 = LOAD '/user/joy.tat_gmail/insurance/2005.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2005 = FILTER insuranceData2005 BY description == 'People aged 18-64..Privately insured (alone
or  in combination)';
--dump data2005;

insuranceData2009 = LOAD '/user/joy.tat_gmail/insurance/2009.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2009 = FILTER insuranceData2009 BY description == 'People aged 18-64..Privately insured (alone
or  in combination)';
--dump data2009;

insuranceData2010 = LOAD '/user/joy.tat_gmail/insurance/2010.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2010 = FILTER insuranceData2010 BY description == 'People aged 18-64..Privately insured (alone
or  in combination)';
--dump data2010;

insuranceData2011 = LOAD '/user/joy.tat_gmail/insurance/2011.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 18-64..Privately insured (alone or  in combination)
data2011 = FILTER insuranceData2011 BY description == 'People aged 18-64..Privately insured (alone
or  in combination)';
--dump data2011;

sub_data2001 = FOREACH data2001 GENERATE total;
sub_data2002 = FOREACH data2002 GENERATE total;
sub_data2003 = FOREACH data2003 GENERATE total;
sub_data2004 = FOREACH data2004 GENERATE total;
sub_data2005 = FOREACH data2005 GENERATE total;
sub_data2009 = FOREACH data2009 GENERATE total;
sub_data2010 = FOREACH data2010 GENERATE total;
sub_data2011 = FOREACH data2011 GENERATE total;

result = UNION data2011, data2010, data2009, data2005, data2004, data2003, data2002, data2001;
grp= Group result all;

insurance_average = FOREACH grp GENERATE (double)AVG(result.total) AS Total_Average;
dump insurance_average;
```



Results

People aged 18-64 opted Privately insured (alone or in combination) is 72.27017269%



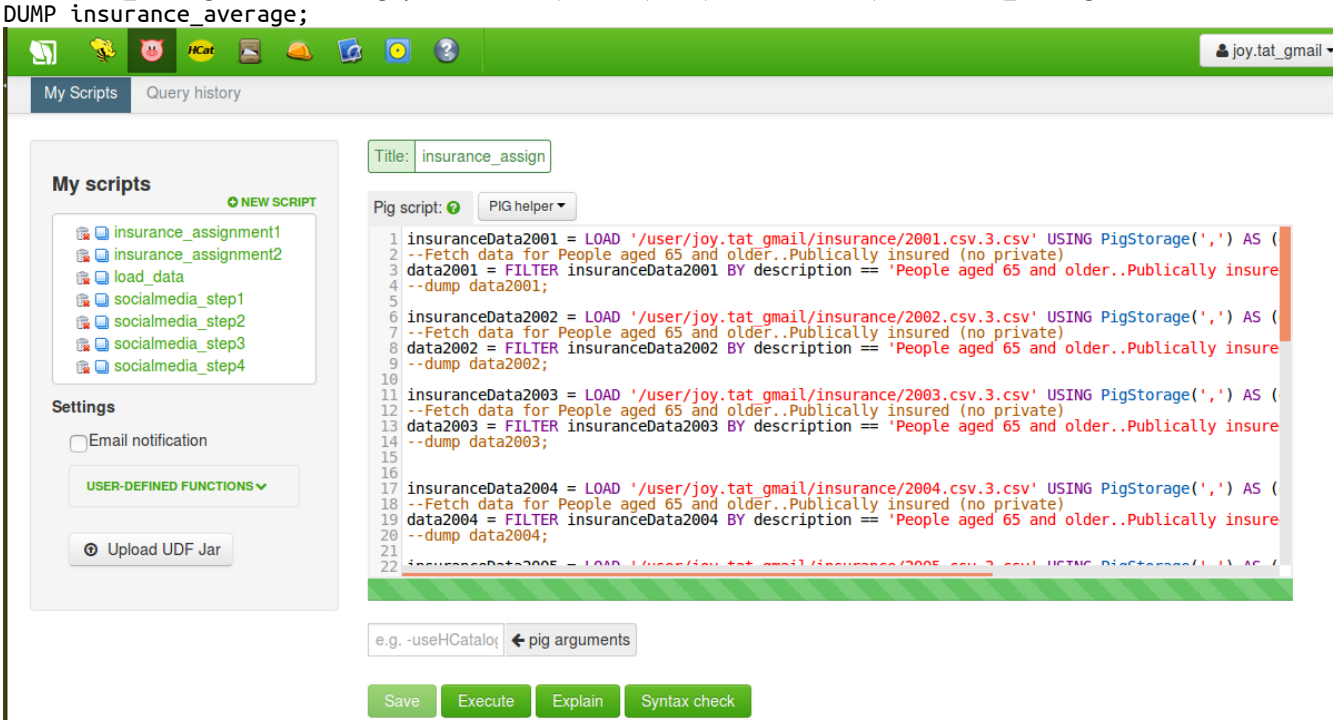
Analysis 2:

- Calculate the average percentage of people, aged 65 years or more, who are solely covered by public insurance from the year 2001 to 2011
- Input the data in HDFS
- Use Pig to arrive at the conclusions
- criteria is : `'People aged 65 and older..Publically insured (no private)'`

Query is:

```
insuranceData2001 = LOAD '/user/joy.tat_gmail/insurance/2001.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2001 = FILTER insuranceData2001 BY description == 'People aged 65 and older..Publically insured
(no private)';
--dump data2001;

insuranceData2002 = LOAD '/user/joy.tat_gmail/insurance/2002.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2002 = FILTER insuranceData2002 BY description == 'People aged 65 and older..Publically insured
(no private)';
--dump data2002;

insuranceData2003 = LOAD '/user/joy.tat_gmail/insurance/2003.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2003 = FILTER insuranceData2003 BY description == 'People aged 65 and older..Publically insured
(no private)';
--dump data2003;


insuranceData2004 = LOAD '/user/joy.tat_gmail/insurance/2004.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2004 = FILTER insuranceData2004 BY description == 'People aged 65 and older..Publically insured
(no private)';
--dump data2004;

insuranceData2005 = LOAD '/user/joy.tat_gmail/insurance/2005.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2005 = FILTER insuranceData2005 BY description == 'People aged 65 and older..Publically insured
(no private)';
--dump data2005;

insuranceData2009 = LOAD '/user/joy.tat_gmail/insurance/2009.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2009 = FILTER insuranceData2009 BY description == 'People aged 65 and older..Publically insured
(no private)';
--dump data2009;

insuranceData2010 = LOAD '/user/joy.tat_gmail/insurance/2010.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2010 = FILTER insuranceData2010 BY description == 'People aged 65 and older..Publically insured
(no private)';
--dump data2010;

insuranceData2011 = LOAD '/user/joy.tat_gmail/insurance/2011.csv.3.csv' USING PigStorage(',') AS
(description:chararray, total: double);
--Fetch data for People aged 65 and older..Publically insured (no private)
data2011 = FILTER insuranceData2011 BY description == 'People aged 65 and older..Publically insured
(no private)';
--dump data2011;

sub_data2001 = FOREACH data2001 GENERATE total;
sub_data2002 = FOREACH data2002 GENERATE total;
sub_data2003 = FOREACH data2003 GENERATE total;
sub_data2004 = FOREACH data2004 GENERATE total;
sub_data2005 = FOREACH data2005 GENERATE total;https://www.google.co.in/?gfe_rd=cr&ei=eWwPWIyaG-
mq8wee4L-ADw
sub_data2009 = FOREACH data2009 GENERATE total;
```

```
sub_data2010 = FOREACH data2010 GENERATE total;
sub_data2011 = FOREACH data2011 GENERATE total;

result = UNION data2011, data2010, data2009, data2005, data2004, data2003, data2002, data2001;
grp= GROUP result ALL;

insurance_average = FOREACH grp GENERATE (double)AVG(result.total) AS Total_Average;
DUMP insurance_average;
```



Results:
People aged 65 and older opted Publically insured (no private) is 24.528774111%