Analysis of New York State Data Using Big Data Technologies

Shreeya Namboori[1]
MSc in Data Analytics
National College of Ireland
x18128947@student.ncirl.ie

***Abstract*— As Government in each countries move towards digitization they look forward to adopting new technologies that can help them in their task. Government databases contains huge amounts of datasets that can be used for analysis to help government officials solve or realize an issue that exists within their country or state, this is where Big Data technologies come in. New York State data has been used for gaining some insights or recognizing some patterns using datasets of employment, crime, consumer price index and substance abuse . Since the datasets are quite huge Big Data technologies like Hadoop MapReduce, Apache Hive and Apache Pig that are efficient in storing, processing and operating on these datasets are used. The output obtained from each of these technologies is then used for visualization using Tableau and PowerBI softwares. Analysis of data using visualization brings out some interesting information about the state and it's different counties.**

Keywords- Apache Hive, Hadoop MapReduce, HBase, MySQL , Apache Pig

I. INTRODUCTION

The advancement in technology and the introduction of internet has given rise to huge amounts of data that is being generated at a rate of 2.5 quintilion bytes of data every day[1], which is a huge amount of data. All the data that is being accumulated is an invaluable resource that helps in gaining some insights, recognizing pattern or finding potential causes to a problem that is currently being dealt with. Therefore Big data technologies have gained popularity because unlike the traditional systems they are able to efficiently process large amount of data which can be used for gaining information. Hadoop MapReduce, Apache Hive, Apache Pig and Apache Spark are some of the famous Big data programming platforms due to their ability of efficiently processing large datasets. Many government websites have now made the datasets contained by them public so that everyone can have the right to access the information about their country to maintain transparency. The size of some of these datasets are very large and might contain general information like employment, unemployment or information on social or economic issues. Analysis of these datasets can help in finding some interesting information that can be used by the government to make better policies accordingly if any issue is realized from the analysis. In this paper datasets of New York State are used for Big data analysis using Hadoop MapReduce, Hive and Pig.

Background and Motivation

New York State is located in the north eastern part of the Unites States (U.S) and it is the fourth most populated state in US with an estimated population of 19.54 million residents in the year 2018[2] . One of the most popular city, New York City is present in this state having approximately 40 percent of the states population. New York State consists of 62 counties. Since this is one of the most populous state in US its dataset can be analyzed for finding some useful information about the state that could be used by government officials in resolving an existing issue in the sate by creating policies. It has been noticed that gradually the population in the states has been decreasing therefore analysis of dataset can help in determining the potential causes or the effect this decline can had on the state of New York.

Research Question

What are the hidden insights or pattern in the New York State datasets about employment, crimes and substance abuse?

II. RELATED WORK

The analysis of criminal activities or crimes to get knowledge about different types of crimes happening in a city has been realized as important as getting some information from the analysis could target at curbing these crooked activities [1]. The crime analysis has been done using Hadoop MapReduce because of its ability to deal and getting information out of a very large dataset efficiently. The result analysis show the frequently occurring crimes and the number of crimes that occurred in a city in a span of 10 years. MapReduce is a complied language having lower level of abstraction [2]. For mapreduce a code is very lengthy and more development efforts are required but the code efficiency is high as compared to hive using wordcount program as an example the performance of Hive and MapReduce were compared. It was found that although mapreduce has a low level of abstraction and lengthy code as compared to

[1]https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

[2]https://en.wikipedia.org/wiki/New-York-City

hive but the efficiency of its code is greater than that of hive.

The comparison of Hive , Pig and MapReduce is done in another research and it is found that Hive and Pig comprising of higher level of abstraction and comparatively shorter code have less code efficiency then the MapReduce programs [3]. Pig is realized as more advantageous then Hive as it offers more optimization and control on data flow.Another research analyzed crime data in Chennai using Hive and MapReduce [4]. The dataset having 100K rows of data of crime in Chennai region for the year 2005 to 2015 was used by Hive and Pig to get some relevant information. Hive was used to get information about the numbers of different crimes in an area and the total number of crime that happened in a period of time. MapReduce was used to get the data for average and total number of crime per year, get the areas with the maximum number of crimes and number of crimes over the years. R tool and correlation matrix with heat map was used for visualizing the output of Hive and MapReduce.

A Big data analytics framework has been suggested for e-government in India as it realized that Big data analytics has started playing major role in many countries [5]. Australian Government information Management Office had adopted the strategy of Big Data analysis to provide better services to its people while Sweden is using Big Data analysis to find pattern in traffic .Similarly many more countries are adopting Big Data analytics to help them get insights in the problems they are facing so they could efficiently deal with it. Introduction of social media platforms like Facebook, Twitter have led to the accumulation of large amount of data therefore Hadoop is proposed for storing and processing this data [6]. Enterprises are implementing Big Data analytics readily as it helps in improving their operations, efficiency and sales. It is stated that Big data analytics has made huge impact in sectors such as banking, healthcare and campaign departments.

III. Methodology

Introduction

This section discusses the dataset used and the architecture of the implementation.

Data Collection

For this research project four datasets have been used and three of them are from the New York State government website. The first dataset contains the information about employment from 1990 to 2019 by industry. This dataset has 390K

The second dataset contains information about crimes in the state by county and by the agency the crimes were reported to from 1990 to 2019. This dataset has 15 columns and 19.3k rows. The third dataset contains information about the number of people that were admitted for treatment for alcohol and substance abuse in each county for different programs from the year 2007 to 2018. This dataset has 79.4 K rows and 7 columns. The fourth dataset contains information about consumer price index that is extracted from the us inflation calculator website (after getting the site owners permission) from the year 1913 to 2019 using extract-areas function of R. This data has 16 columns and 105 rows.

Dataset cleaning

1. Employment dataset [3]
   - The dataset was checked for NA values which were omitted from the dataset.
   - All the unnecessary characters like commas, hyphen and round brackets were removed from the dataset using lapply.function of R.
   - The finally cleaned dataset was then saved as a csv.
2. Substance abuse treatment dataset[4]
   - The dataset was checked for NA values which were omitted from the dataset.
   - All the unnecessary characters like commas, hyphen and round brackets were removed from the dataset using lapply function of R.
   - A unique id was created for each row using tibble.
   - The finally cleaned dataset was then saved as a csv
3. State Crime dataset [5]
   - The dataset was checked for NA values which were omitted from the dataset.
   - All the unnecessary characters like commas, hyphen and round brackets were removed from the dataset using lapply function of R.
   - A unique id was created for each row using tibble.
   - All the unnecessary columns from the dataset were removed and the remaining columns in this dataset were 6. This file was then saved as csv.
4. Consumer price index (CPI) dataset [6]
   - After extracting the data from the website all the columns except for year, annual average cpi and percentage change in cpi columns were removed from the dataset as they were not required.
   - As the three columns did not had a proper name they were renamed properly
   - NA values in the dataset was checked and removed and the file was the saved as csv.

BigData Processing

The cleaned datasets were loaded into MySQL database tables(Figure 2) . After all the four tables in MySQL were successfully loaded with the data of the four datasets, the data from these tables was then imported to HDFS using sqoop. Data from HDFS was then used by MapReduce, Hive

[3] https://data.ny.gov/Economic-Development/Current-Employment-Statistics-Beginning-1990/6k74-dgkb

[4] https://data.ny.gov/Human-Services/Chemical-Dependence-Treatment-Program-Admissions-B/ngbt-9rwf

[5] https://data.ny.gov/Public-Safety/Index-Crimes-by-County-and-Agency-Beginning-1990/ca8h-8gjq

[6] https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/

```
library(tidyverse)
library(XML)
library(stringr)
library(dplyr)
library(htmltab)
library(readr)
url = "https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/"
cpi <- htmltab(doc=url, which=1)
head(cpi,100)
  names(cpi)
cpi<-data.frame(cpi)
cpi<- cpi[-c(2,4:13)]
cpi<- cpi[-c(2,4)]
names(cpi)[names(cpi) == "V13"] <- "Year"
names(cpi)[names(cpi) == "Percent.Change.1"] <- "Percent_change"
names(cpi)[names(cpi) == "Annual"] <- "Annual_Avg"
cpi<- cpi[-c(1), ]

cpi$'Percent_change' <- as.double(as.character(cpi$'Percent_change'))
cpi$'Annual_Avg' <- as.double(as.character(cpi$'Annual_Avg'))
colSums(is.na(cpi))
cpi<-na.omit(cpi)
write.csv(cpi,file="uscpi.csv",row.names=FALSE)
```

Fig. 1. CPI table extracted using R

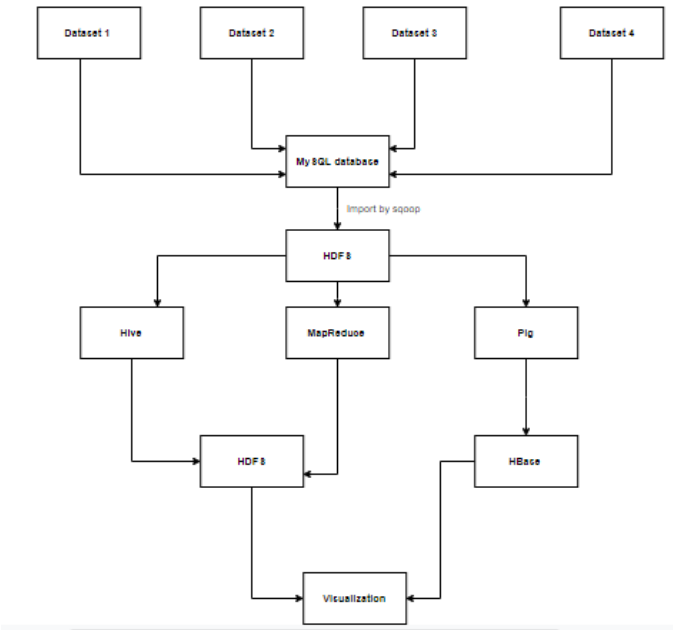

Fig. 2. Big Data Process

and Pig for processing and getting the required information. The output results of MapReduce and Hive were stored in HDFS again while the output of Pig was stored in HBase nosql database These results were then visualized using Tableau and PowerBI software to gain some insights about the information obtained.

MySQL

MySQL relational database was used for saving all the four csv files after they were cleaned. Inside the Test database of MySQL four tables named employ (for employment dataset), cpi (for cpi datset), substanceabuse (for substance abuse treatment dataset) and crimereport (for state crime dataset) were created. These tables were then loaded with the csv data using the command LOAD DATA LOCAL INFILE. Figure 3 ,4 ,5 and 6 shows the four tables in SQL and the number of entries (rows) it has inside them.

```
mysql> select COUNT(*) FROM substanceabuse;
+----------+
| COUNT(*) |
+----------+
|    79411 |
+----------+
1 row in set (0.33 sec)
```

Fig. 3. Substance abuse table in MySQL

```
mysql> select COUNT(*) FROM cpi;
+----------+
| COUNT(*) |
+----------+
|      105 |
+----------+
1 row in set (0.03 sec)
```

Fig. 4. CPI table in MySQL

sql.PNG

```
mysql> select COUNT(*) FROM employ;
+----------+
| COUNT(*) |
+----------+
|   389985 |
+----------+
1 row in set (4.70 sec)
```

Fig. 5. Employment table in MySQL

```
mysql> select COUNT(*) FROM crimereport;
+----------+
| COUNT(*) |
+----------+
|    19324 |
+----------+
1 row in set (0.24 sec)
```

Fig. 6. Crime table in MySQL

Sqoop

After the data from csv file was successfully loaded into the MySQL database, Sqoop was used for importing the data from each of these tables into HDFS (Hadoop Distributed File System) using sqoop import command.

IV. Results

The data present in HDFS was processed by MapReduce, Hive and Pig to get some information from the large datasets and their outputs were visualized using Tableau and PowerBI.

1. Hive Task 1 :

Data from HDFS was processed by Hive for finding total number of crimes (reported by different agencies from all the counties) in New York State from 1990 to 2017.

```
hive> select sum(totalcrime),year from crimeHive group by year order by year de
sc;
```

Fig. 7. Hive query 1

```
602884  2015
635149  2014
671001  2013
708977  2012
706152  2011
712756  2010
711583  2009
730963  2008
921990  2007
1228079 2006
1251065 2005
1275044 2004
1321013 2003
1366099 2002
1408230 2001
1467524 2000
1509190 1999
1623925 1998
1771085 1997
1869331 1996
2039612 1995
2226215 1994
2430768 1993
2547943 1992
2698494 1991
2701049 1990
Time taken: 85.984 seconds, Fetched: 28 row(s)
hive>
```

Fig. 8. Hive query 1 output

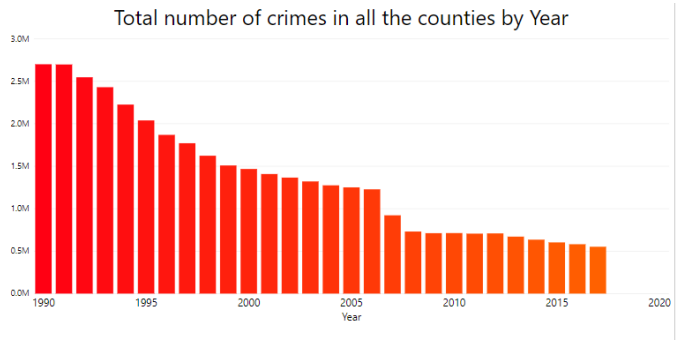

Fig. 9. Hive query 1 visualization

Hive was used for getting total number of crimes reported by different agencies in different counties on a yearly basis (1990 to 2017) and the output of hive was used for visualization in PowerBI. It can be seen from the visualization that the number of crimes keeps on decreasing after 1990 and the number of crimes in 2017 are comparatively less than the crimes in 1990.

2. Hive Task 2 :

The annual average CPI value and total number of crimes are taken after the year 2010 in descending order of the year using Hive that requires joining of two tables one of cpi and the other of crime.

```
hive> select sum(c.totalcrime),b.annualavg as cb from crimeHive c join cpiHive
b on (b.year=c.year) where b.year>2010 group by b.annualavg;
```

Fig. 10. Hive query 2

```
 HDFS Write: 111 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 970 msec
OK
1412304 224.939
1417954 229.594
1342002 232.957
1270298 236.736
1205768 237.017
1162616 240.007
1103544 245.12
Time taken: 56.64 seconds, Fetched: 7 row(s)
hive>
```

Fig. 11. Hive query 2 output

From the visualization (using PowerBI)of the hive output it can be seen that the years having high annual average cpi have less number of crimes. Since the values are in descending order of the year it can be noted that as the cpi keeps on increasing the total number of crime reported keeps on decreasing.

Total number of crimes	cpi
1103544	245.12
1162616	240.01
1205768	237.02
1270298	236.74
1342002	232.96
1412304	224.94
1417954	229.59

Fig. 12. Hive query 2 visualization

3. Pig Task :

The data for total number of people that were admitted for treatment of alcohol and substance abuse (cocaine, marijuana, heroin,opinods etc) in different counties yearly was taken using Pig.

```
user_record = LOAD 'hdfs://localhost:50070/sqoop/substanceabuse/part-m-00000'
USING PigStorage(',')
AS(id:INT,year:INT,county:chararray,program:chararray,service:chararray,age:chS
DUMP user_record;
state_record = GROUP user_record BY year;
result = FOREACH state_record GENERATE group,
SUM(user_record.num);
STORE result INTO 'pigca3.txt' USING PigStorage (',');
DUMP result;
```

Fig. 13. Pig Query

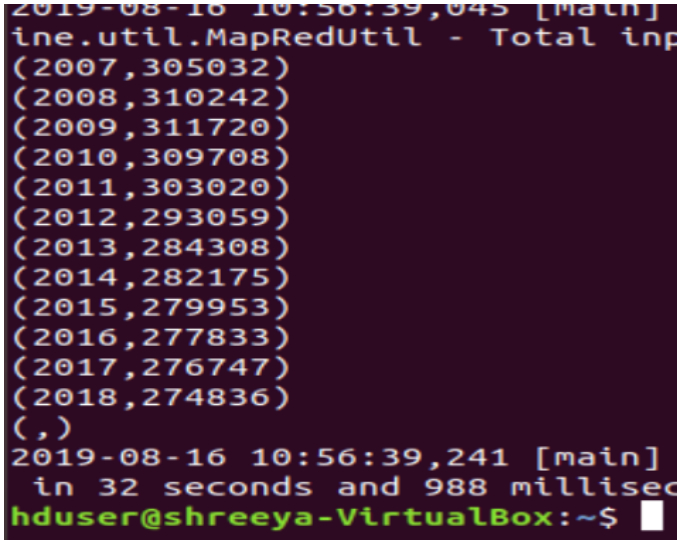

Fig. 14. Pig query result

From the Pig query output visualization (using Tableau) it can be seen that as the year keeps on increasing the number of people that were admitted for treatment of alcohol and substance abuse keeps on decreasing gradually.

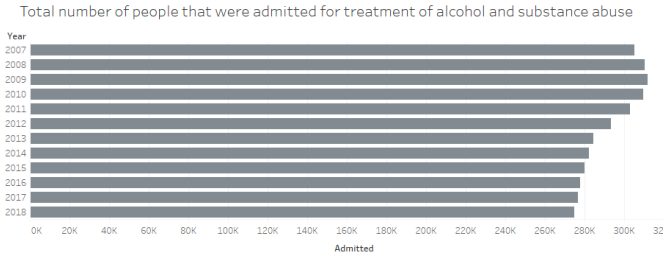

Fig. 15. Pig output visualization

4. MapReduce Task 1 :

The data for total number of people that were employed by government in all counties from 2011 to 2018 was obtained using MapReduce.

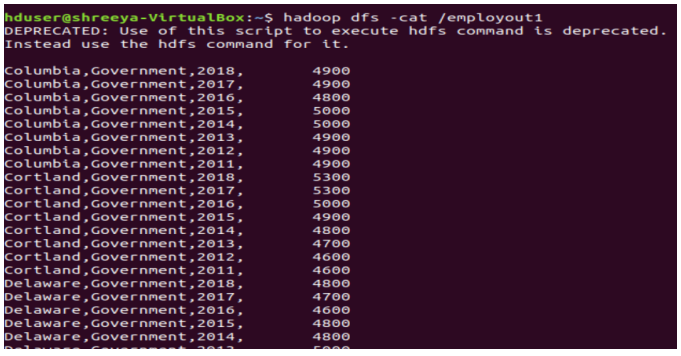

Fig. 16. Mapreduce Task 1 result

From the Tableau visualization it can be seen that Nassau county has the highest number of government jobs for all the years between 2011 and 2018. Albany county ranks second in having total number of government jobs .

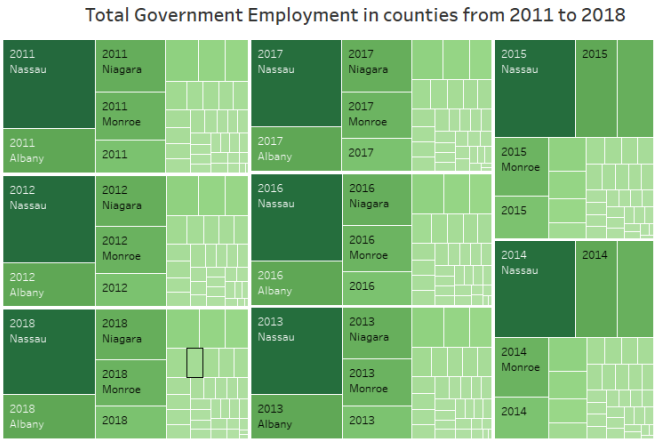

Fig. 17. Mapreduce Task 1 visualization

5. MapReduce Task 2 :

The data for total number of people who were under 18 and were admitted for the treatment of alcohol and substance abuse by county from 2011 to 2018 was taken using MapReduce.

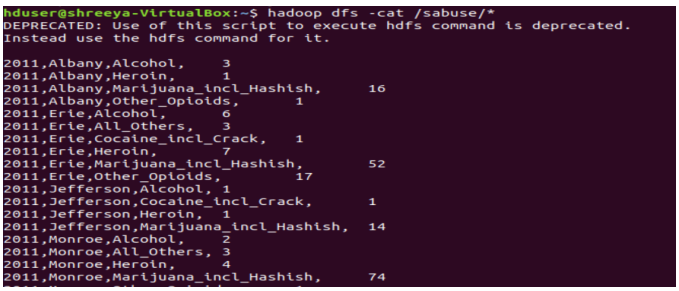

Fig. 18. Mapreduce Task 2 result

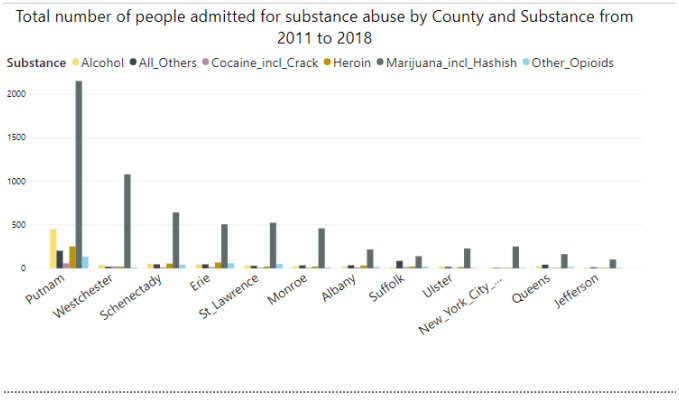

Fig. 19. Mapreduce Task 2 visualization

From the Tableau visualization it can be seen that the state of Putnam had the highest number of teenagers that were admitted for the treatment of alcohol and substance abuse.Westcheter county has the second most number of

teenagers that were admitted for substance abuse from the period of 2011 to 2018.

6. MapReduce Task 3 :

The data for the Top 20 counties along with industries by (number of) employment for the year 2018 was taken using MapReduce.

```
hduser@shreeya-VirtualBox:~$ hadoop dfs -cat /top20/*
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Albany,PrivateServiceProviding,2018,    330900
Albany,TotalPrivate,2018,       377200
Monroe,PrivateServiceProviding,2018,    382600
Niagara,PrivateServiceProviding,2018,   406800
Albany,ServiceProviding,2018,   433800
Monroe,TotalPrivate,2018,       460100
Monroe,ServiceProviding,2018,   465900
Niagara,TotalPrivate,2018,      479200
Albany,TotalNonfarm,2018,       480100
Niagara,ServiceProviding,2018,  498300
Monroe,TotalNonfarm,2018,       543400
OrangeRocklandWestchesterMetroArea,PrivateServiceProviding,2018,        546500
Niagara,TotalNonfarm,2018,      570700
OrangeRocklandWestchesterMetroArea,TotalPrivate,2018,   619900
OrangeRocklandWestchesterMetroArea,ServiceProviding,2018,       656600
OrangeRocklandWestchesterMetroArea,TotalNonfarm,2018,   730000
Nassau,PrivateServiceProviding,2018,    1001700
Nassau,TotalPrivate,2018,       1157200
Nassau,ServiceProviding,2018,   1202600
Nassau,TotalNonfarm,2018,       1358100
hduser@shreeya-VirtualBox:~$
```

Fig. 20. Mapreduce Task 3 result

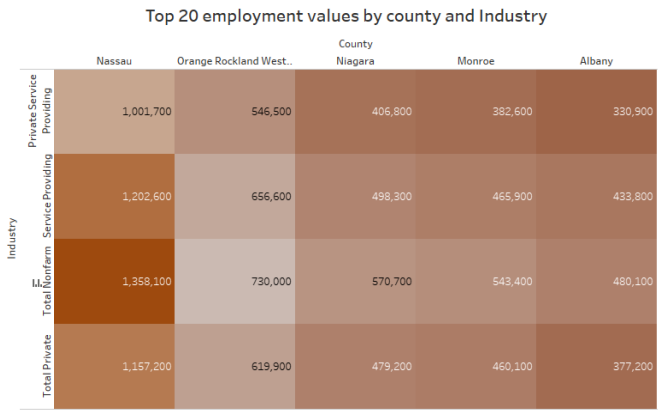

Fig. 21. Mapreduce Task 3 visualization

From the Tableau visualization it can be seen that only five counties for four industry made it to the list of Top 20 by the employment number. Nassau county had the highest employment from Total Nonfarm industry. Nassau actually takes the Top 4 position in the Top 20 for all the four industries which means Nassau county has a higher employment for these industrial sectors than any other county.

V. Conclusion

Hadoop MapReduce, Apache Hive and Apache Pig were used in this research for processing the large New York State datasets and some interesting information were observed after visualizing the processed data. Nassau county had the highest number of employment for four sectors, Putnam county had highest number of teenagers that were admitted for treatment of alcohol and substance abuse, Nassau county had highest employment from 2011 to 2018 , with each passing year as cpi increases the total number of crimes becomes less, total number of crime keeps on decreasing with each passing year and the number of people who were admitted for substance abuse have become comparatively less in 2018 compared to 2011. These findings can help out the state government in making or implementing new rules and regulations.

References

[1] A. Kumar, V. S, and R. Kayalvizhi, "Using Mapreduce Techniques to Predict and Examine Crime Pattern," *International Journal of Engineering & Technology*, vol. 7, no. 3.12, p. 43, 2018.

[2] T. Mehta and N. Mangla, "A Survey Paper on Big Data Analytics using Map Reduce and Hive on Hadoop Framework," *National Conference on Recent Innovations in Science, Technology & Management (NCRISTM)*, no. February, pp. 112–118, 2016.

[3] U. R. Pol, "Big Data Analysis : Comparision of Hadoop MapReduce , Pig and Hive," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 5, no. 6, pp. 9687–9693, 2016.

[4] P. B. Minajagi and P. R. Nadagoudar, "Crime Analysis and Prediction Using Big Data," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 2, no. 4, pp. 477–482, 2017.

[5] M. R. Rajagopalan and S. Vellaipandiyan, "Big data framework for national E-governance plan," *International Conference on ICT and Knowledge Engineering*, pp. 1–5, 2013.

[6] M. Sogodekar, S. Pandey, I. Tupkari, and A. Manekar, "Big data analytics: Hadoop and tools," *IEEE Bombay Section Symposium 2016: Frontiers of Technology: Fuelling Prosperity of Planet and People, IBSS 2016*, pp. 1–6, 2017.