Big Data Analytics Final Report

Authors Name/s - Dhruv Motwani (dgm2138), Harshit Saxena (hs2873) , Abhinav Iyappan (avi2111)

I. Introduction

The Healthcare and Medical industry is one of the biggest industries in the US and even the world, with estimated net transactions of $3.24 trillion. American consumers spend hundreds of thousands of dollars in insurance and medical costs. In such a huge industry, even small percentage increases in efficiency can results in million of dollars in increased revenue and higher market share. It is no surprise that companies form collaborations and have huge data analytic departments to maximize their profits. From a marketing point of view, Pharmaceuticals usually give some sort of monetary benefits to healthcare providers like hospitals and doctors so that these providers can use the services of these companies. For example, the maker of a particular brand of medicinal pills may pay a doctor to use and vouch for its medicines, rather than some generic or competing brand. Benefits may also be in the form of travel reimbursements, speaking fees, research grants, or even ownership interests in the company. Such payments by drug and device companies are so commonplace, that the payments total 7.52 Billion dollars, from 12 million payments per year. We try to analyse some of this data to capture and see trends in payments due to geographic or demographic factors.

II. System Overview

The dataset that we are using is from openpayments. Open Payments is a federal program, required by the Affordable Care Act, that collects information about the payments drug and device companies make to physicians and teaching hospitals for things like travel, research, gifts, speaking fees, and meals. It also includes ownership interests that physicians or their immediate family members have in these companies. This data is then made available to the public each year on this website

. **For patients, consumers, and the public,** Open Payments can be used to learn about the relationships between physicians and applicable manufacturers and GPOs. We encourage patients to discuss these relationships with their healthcare providers.

For physicians and teaching hospital representatives, reviewing the data reported about you in the Open Payments system can ensure that this information is accurate. You can also:

- Use the information reported about you to plan for questions from patients

- Healthcare comprises 17.5% of US GDP.
- 19 major drug companies had atleast one board member who also held leadership role at atleast one academic medical center.

III. Algorithm

MLlib fits into Spark's APIs and interoperates with NumPy in Python (as of Spark 0.9) and R libraries (as of Spark 1.5). You can use any Hadoop data source (e.g. HDFS, HBase, or local files), making it easy to plug into Hadoop workflows. ML algorithms include:

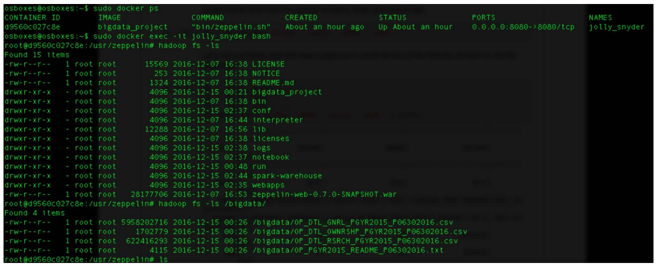
- Classification: logistic regression, naive Bayes,...
- Regression: generalized linear regression, survival regression,...
- Decision trees, random forests, and gradient-boosted trees
- Recommendation: alternating least squares (ALS)
- Clustering: K-means, Gaussian mixtures (GMMs),...
- Topic modeling: latent Dirichlet allocation (LDA)
- Frequent itemsets, association rules, and sequential pattern mining

IV. Software Package Description

Spark is a Data processing engine/framework developed by the Apache foundation. It is an improvement upon Hadoop because of its ability to process data in memory, rather than copying intermediate results to HDFS like Hadoop does. Programs can potentially be run upto 100x faster in Spark when compared to Haoop. It allows easy use by being compatible with common programming languages like R, Python, Scala and Java. It has loads of libraries that allow us to perform operations in SQL and Machine Learning among others using MLlib.

MLlib: MLlib is Spark's Machine learning library. It is extremely scalable, sophisticated and versatile. It provides functions for various operations like Classification, Regressions and Clustering using multiple methods like Decision trees, Gaussian Mixtures, and Latent Dirichlet Allocations.

2. Docker: Docker is a software containerization platform, much like Virtual Machine hosts like VMware or VirtualBox. It allows for wrapping software in a complete filesystem inclusive of system tools and libraries. The end result is that Docker appears to users as its own virtual machine, when it is in fact a container running on part of a remote server.

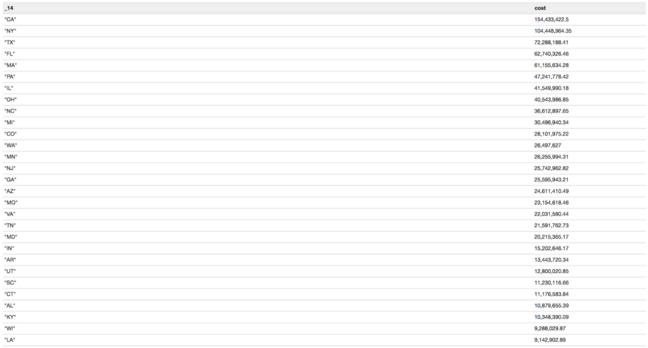


```
osboxes@osboxes:~$ sudo docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS              PORTS                    NAMES
d9560c027c8e        bigdata_project     "bin/zeppelin.sh"   About an hour ago   Up About an hour    0.0.0.0:8080->8080/tcp   jolly_snyder
osboxes@osboxes:~$ sudo docker exec -it jolly_snyder bash
root@d9560c027c8e:/usr/zeppelin# hadoop fs -ls
Found 15 items
-rw-r--r--   1 root root      15569 2016-12-07 16:38 LICENSE
-rw-r--r--   1 root root        253 2016-12-07 16:38 NOTICE
-rw-r--r--   1 root root       1324 2016-12-07 16:38 README.md
drwxr-xr-x   - root root       4096 2016-12-15 00:21 bigdata_project
drwxr-xr-x   - root root       4096 2016-12-07 16:38 bin
drwxr-xr-x   - root root       4096 2016-12-15 02:37 conf
drwxr-xr-x   - root root       4096 2016-12-07 16:44 interpreter
drwxr-xr-x   - root root      12288 2016-12-07 16:56 lib
drwxr-xr-x   - root root       4096 2016-12-07 16:38 licenses
drwxr-xr-x   - root root       4096 2016-12-15 02:38 logs
drwxr-xr-x   - root root       4096 2016-12-15 02:37 notebook
drwxr-xr-x   - root root       4096 2016-12-15 00:48 run
drwxr-xr-x   - root root       4096 2016-12-15 02:44 spark-warehouse
drwxr-xr-x   - root root       4096 2016-12-15 02:35 webapps
-rw-r--r--   1 root root   28177706 2016-12-07 16:53 zeppelin-web-0.7.0-SNAPSHOT.war
root@d9560c027c8e:/usr/zeppelin# hadoop fs -ls /bigdata/
Found 4 items
-rw-r--r--   1 root root 5958202716 2016-12-15 00:26 /bigdata/OP_DTL_GNRL_PGYR2015_P06302016.csv
-rw-r--r--   1 root root    1702779 2016-12-15 00:26 /bigdata/OP_DTL_OWNRSHP_PGYR2015_P06302016.csv
-rw-r--r--   1 root root  622416293 2016-12-15 00:26 /bigdata/OP_DTL_RSRCH_PGYR2015_P06302016.csv
-rw-r--r--   1 root root       4115 2016-12-15 00:26 /bigdata/OP_PGYR2015_README_P06302016.txt
root@d9560c027c8e:/usr/zeppelin# ls
```

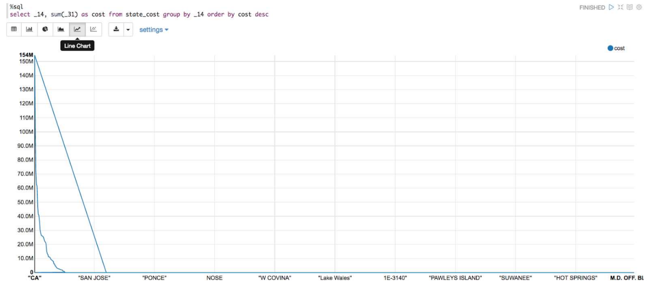
3. Zeppelin: Zeppelin is a programming notebook much like Sagemath or Jupyter. However, there are some key differences. The main difference is that Zeppelin being an Apache project supports Big Data languages like Hadoop, Pig etc. It also allows for Spark accessibility using API calls in python etc. Another useful feature is the ability to code in different languages in different sections, which is not possible in Jupyter. Having a notebook allows us to both code and visualize results on the fly in the same software bundle, rather than having to print, take screenshots, save and collate among other things.

V. Experiment Results

We first did a count of the transactions, grouping by state. We saw clear signs of higher number of transactions in more populous state by demographic, and not by geographical area. For example California and New York are the highest transacted states.

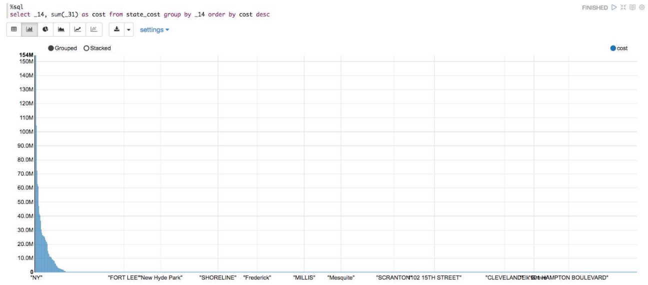


_14	cost
"CA"	154,433,422.5
"NY"	104,448,964.35
"TX"	72,288,188.41
"FL"	62,740,326.46
"MA"	61,155,634.28
"PA"	47,241,778.42
"IL"	41,549,990.18
"OH"	40,543,986.85
"NC"	36,612,897.65
"MI"	30,496,940.34
"CO"	28,101,975.22
"WA"	26,497,627
"MN"	26,255,994.31
"NJ"	25,742,962.82
"GA"	25,595,943.21
"AZ"	24,611,410.49
"MO"	23,154,618.46
"VA"	22,031,580.44
"TN"	21,591,762.73
"MD"	20,215,365.17
"IN"	15,202,646.17
"AR"	13,443,720.34
"UT"	12,800,020.85
"SC"	11,230,116.66
"CT"	11,176,583.84
"AL"	10,879,655.39
"KY"	10,348,390.09
"WI"	9,288,029.87
"LA"	9,142,902.89

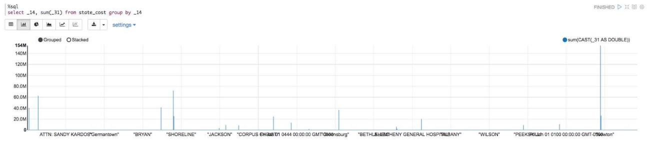


```
%sql
select _14, sum(_31) as cost from state_cost group by _14 order by cost desc
```

We can Also see that not only that are the poular states leading in terms of pure nuber of transactions, they also have the most expensive transactions too. This can be attributed to the fact that California and New York have financial and technological hubs like New York City and San Francisco among others which contain headquarters of the biggest pharmaceutical companies on earth.



```
%sql
select _14, sum(_31) as cost from state_cost group by _14 order by cost desc
```



```
%sql
select _14, sum(_31) from state_cost group by _14
```

of your algorithm. Show
how did you evaluate the performance of your algorithm.

VI. Conclusion

Below is a rough summary of the work done as part of the project.
Initial Literature review and plan formulation.
Initial testing of Spark Mllib-cleaning and data organizing done in Python.
Data writing to HDFS using hadoop.
Analysis of data in Spark Mllib
Visualization using matplotlib in Zeppelin.

Acknowledgment

The authors would like to thank first and foremost the developers of the open-source projects that made development and testing of this project possible, specifically: MLLib in Spark, Zeppelin, . We also thank and acknowledge Professor Ching-Yung Lin for his time and efforts as professor in this Big Data Analytics course, and Eric and other TAs that supported him.

Appendix

References

[1] https://openpaymentsdata.cms.gov/

[2] https://zeppelin.apache.org/

[3] http://spark.apache.org/.