Analysis of Legally Operating Business in New York City using Bigdata Technologies

Programming for Data Analytics

Mathiazhagan Sampath
x18139973
MSc/PGDip Data Analytics 2019/20
Submitted to: Muhammad Iqbal
National College of Ireland, Dublin

***Abstract*—Bigdata is nothing but large amount of data, Bigdata is processed and analyzed to extracted valuable insights using latest technologies. Bigdata is trending technology in market, implementing bigdata in business organization can bring huge profit to business. Legally running business in a city are contributing more to the economy growth of the city on comparing to other sources. In this paper we use Bigdata technologies like Hadoop MapReduce, Sqoop, Pig, Hive and HBase on New York city Legally operating business data to get insights on number of business operating in each district with zip code , type of Legally operating business in New York city, License type of business operating in city with Active and expired license, type of industry operating in city with number of business related to that industry. This project analysis will be helpful for government to monitor the business in New York city and for business people or organization to expand their business in New York and helps in starting a new business based on industry performance.**

***Index Terms*—MapReduce, Business, Sqoop, Hive, Hadoop Distributed File System (HDFS), HBase, Pig, MySQL.**

I. Introduction

New York city is most popularly known as NYC and most popular city in United States of America. NYC contribute an important part to GDP, Business sector in NYC generate greater revenue on comparing to other sectors. There are more than 40 industry sectors in NYC and distributed to different district. Many industries work on License and they need to renew their license accordingly. Legally operating business can be distinguished as Business and Individual and district development depends on number of business in that area.

With bigdata on NYC legally business operation, if a person needs to start a business or to expand their existing business to different place in NYC. Bigdata analysis will be very helpful for that. On the other hand, Government can be benefited by implementing the bigdata technology to find the how fair the business is established and license of each business in an area is renewed or not and future planning of a government can be taken based on the insight. NYC

By using Hadoop MapReduce, Pig and Hive we can analyze the bigdata of NYC legally operating business and the output will be stored to NoSQL database (HBase). File will be moved out from HBase to do visualisation and insights can be used by Government or any other organization.

TABLE I
Details of Technology used in Query

Query	Technology
Query 1	Java MapReduce
Query 2	Java MapReduce
Query 3	Hive
Query 4	Hive
Query 5	Pig

The remaining section of this paper consist of Section 2 which focus on past work related to this field. Methodology is discussed in Section 3 and results are in Section 4 and conclusion is discussed at Section 4. Python will be used to do statistical analysis and Tableau for visualizing the output.

II. Related Work

Business related important decision can be taken thrown data driven approach [1]. For enterprise to make decision bigdata technology are really helpful. To make decision related to enterprise data driven approach is the best and it is split in 4 stages [2], First stage is specifying the source of the data followed by developing scored card based on the appropriate assessment measure and other two steps are technology driven methods by using bigdata technologies like Pig and MapReduce to produce the output. is a research agenda on implementing bigdata analytics in this fast growing digital age to help business model transformation. Digital world and bigdata analytics are together coined as datafication and using MapReduce task to retrieve insights from the data and insights from datafication can be used to change the data model of the business.

Bigdata analytics help in changing the business marketing strategy by collecting the large amount of business data related to marketing and stored on the Hadoop distributed file system (HDFS) and Java MapReduce is used to find the marketing expense type and the expense. From the output of MapReduce, the data is visualized, and important decision has been made to change the strategy of marketing of the company to other approach. In many important decision-making places, bigdata technologies plays an important role and helps in reducing the burden of taking wrong decision by analyzing the data.

Using large dataset of New York Taxi data prediction of the demand using time series forecasting has been done Machine learning techniques but prior to that store the large amount of data for processing has been stored in Hadoop and MapReduce job had ran to classify the data [5], Each trip to the airport from NYC has been sorted using MapReduce and logistic regression has been applied on MapReduce output to predict the demand of the taxi in future.

Bigdata analytics is used in logistics and supply chain management to find the data driven analysis to make important decision in business process [6]. As usual step in bigdata analytics data is used to store in Hadoop distributed file system (HDFS) and data MapReduce job had ran on HDFS data to get insights from the historical bigdata and output has been take to NoSQL database like Hive to carry out future analysis and a change has been made in supply chain systems. There is contradiction on performance of MapReduce on comparing to Spark framework. Through experiments it is found that the performance of MapReduce is low on comaparing to Spark [7,8]. HBase is known for query optimization [9], Social media content are queried stored in HBase and queried, and the performance of the query is very high in HBase. To evaluate the performance of the Bigdata technologies like MapReduce, Pig, Hive [10] used against unstructured social media and concluded on abstraction level. Pig and Hive have high level of abstraction and MapRedcue with low level of abstraction. In [11], movie recommendation is done by analyzing the bigdata on Pig and MapReduce by breaking the traditional approach using rating and other query joins and implemented on upcoming movies.

Overall bigdata plays an important role in many business-related decisions in all sectors/industry. Bigdata driven decision are always leads to the growth of the business.

III. Methodology

New York City legally operating business dataset is maintained by NYC government which consist of different type of industry with zip code and other data. The data is updated on periodic basis and data size is so huge.

A. Dataset Description

Legally operating business of New York City dataset is downloaded from NYC website(https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh) which consist of business-related information in New York city. Data can be downloaded as csv file from website or can be extracted using API connection. Dataset consist of 198,000 rows and 27 columns

B. Data Preprocessing

As the dataset is very large, cleaning the data in Excel is a hectic task. To ease the preprocessing task, Python has been used to clean the data. In Jupyter notebook, dataset is loaded and unwanted column related to our analysis are removed and analysis of the new York leaglly operating business dataset started.

TABLE II
NYC BUSINESS DATASET EXPLANATION

Column Name	Description	Type
DCA License Number	Identification number	Plain Text
License Type	Two license types	Plain Text
License Expiration Date	Exp date of License.	Date & Time
License Status	State of License	Plain Text
License Creation Date	Creation date of License	Date & Time
Industry	Type of industry	Plain Text
Business Name	Name with Secretary	Plain Text
Business Name 2	Trade name.	Plain Text
Address Building	The building number .	Plain Text
Address Street Name	Street name	Plain Text
Secondary Address	The cross-street	Plain Text
Address City	Business is located.	Plain Text
Address State	State-Business is located.	Plain Text
Address ZIP	The zip code	INT
Contact Phone Number	Contact number	Plain Text
Address Borough	The borough	Plain Text
Borough Code	License Category	Plain Text
Community Board	Community Board	Plain Text
Council District	District council	INT
BIN	BIN code	Plain Text
BBL	BBL code	Plain Text
NTA	NTA code	Plain Text
Census Tract	Census Tract of NYC	Plain Text
Detail	Explanation of codes	Plain Text
Longitude	Longitude of business	Plain Text
Latitude	Latitude of business	Plain Text

1) Removal of Unwanted column: BIN, BBL, NTA, Census Tract, Detail, Borough code are columns in raw data removed using python.

2) Converting to Upper case : City name column in downloaded dataset is converted to upper case in python using Jupyter notebook.

3) Exporting the cleaned data : The cleaned and converted data will be stored into nycbusiness.csv

C. Process flow and Architecture

The process flow of this project is represented in Figure 1.

1) MySQL : As the preprocessed data is stored in nycbusiness.csv file, data is stored into MySQL database. To store the data into database, nycbusiness database is created and table nycbusiness is created in database.

2) Data Ingestion Sqoop : Data loaded in MySQL will be imported to Hadoop distributed file system (HDFS) using sqoop. Figure 2 shows the successful import of data from MySQL to HDFS. Sqoop transfer data from database at a faster rate.

Now the data is available in HDFS and MapReduce, Hive will be used to do the analysis of the data in HDFS.

3) MapReduce : MapReduce is a Java programming language used to process the data in HDFS. Eclipse IDE is used to write MapReduce task, Top 10 design pattern of MapReduce is used to execute the MapReduce. Two tasks are executed on HDFS so two driver, mapper and reducer class are created. Using Export function in Eclipse, two executable JAR are created using driver class for the task and stored in hduser home directory. a. Citynyc.jar b. Industrynyc.jar

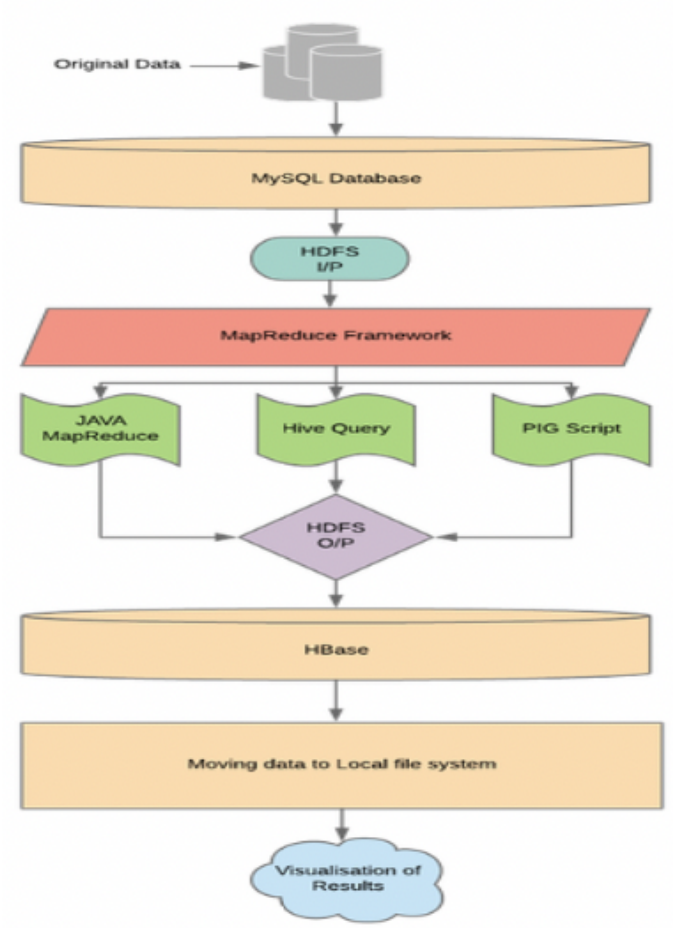

Fig. 1. Process flow of Data Analysis

```
mysql> use nycbusiness;
Database changed
mysql> CREATE TABLE nycbusiness ( DCANumber varchar(255),LicenseType varchar(255
),City varchar(255),Status varchar(255),LicenseCreationDate varchar(255),Industr
y varchar(255),ExpDate varchar(255),
    -> BusinessName varchar(255),Building varchar(255),
    -> Street varchar(255),AddressCity varchar(255),
    -> State varchar(255),Zip INT,AddressBorough varchar(255),CouncilDistrict IN
T,Detail varchar(255),Longitude varchar(255),Lattitude varchar(255));

Query OK, 0 rows affected (0.02 sec)

mysql> load data local infile
    -> '/home/hduser/nycbusiness.csv' into table
    -> nycbusiness fields terminated by ',' lines terminated by '\n';
Query OK, 197463 rows affected, 65535 warnings (1.80 sec)
Records: 197463  Deleted: 0  Skipped: 0  Warnings: 361795

mysql> select count(*) from nycbusiness;
+-----------+
| count(*) |
+-----------+
|   197463 |
+-----------+
1 row in set (0.09 sec)
```

Fig. 2. Data load into MySQL

The file in HDFS is copied as .csv file so that .csv file can be directly imported to other Hive and Pig databases from HDFS. Command to copy the file in HDFS to csv file: **hadoop dfs -cp /nycbusiness/part-m-00000 /nycbusiness/nycbusiness.csv**

MapReduce Task 1 Top 20 Business place in NYC Top 20 Places in NYC with high business numbers are filtered using Hadoop MapReduce. To do that Mapper, Reducer,

```
                FILE: Number of bytes written=207128
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=31532542
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=10403
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=10403
                Total vcore-milliseconds taken by all map tasks=10403
                Total megabyte-milliseconds taken by all map tasks=10652672
        Map-Reduce Framework
                Map input records=197463
                Map output records=197463
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=339
                CPU time spent (ms)=4710
                Physical memory (bytes) snapshot=163401728
                Virtual memory (bytes) snapshot=1914900480
                Total committed heap usage (bytes)=62980096
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=31532542
9/08/08 13:16:06 INFO mapreduce.ImportJobBase: Transferred 30.0718 MB in 27.7546 seconds (1.0835 MB/sec)
9/08/08 13:16:06 INFO mapreduce.ImportJobBase: Retrieved 197463 records.
base) hduser@mathi-VirtualBox:~$
```

Fig. 3. Data Ingestion using Sqoop

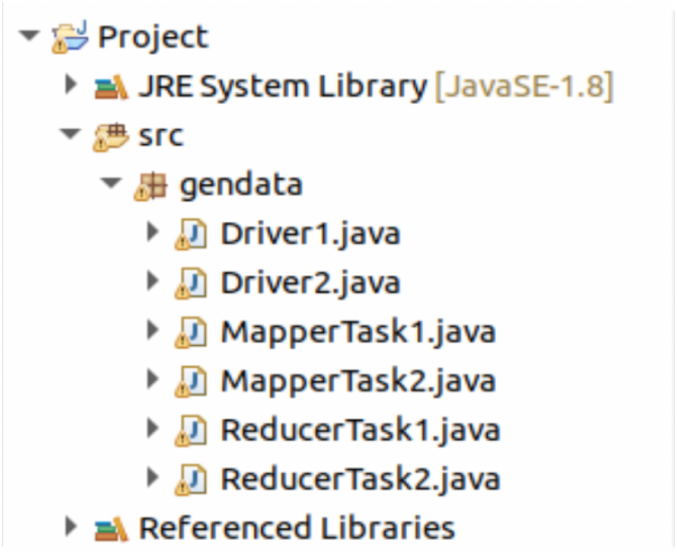

Fig. 4. Eclipse MapReduce Java class

Driver class are created in eclipse and using export function an executable JAR is created. Using JAR, Hadoop MapReduce job had run by passing HDFS file input and output location.

```
(base) hduser@mathi-VirtualBox:~$ hadoop jar citynyc.jar /nycbusiness /nycbusinesscity
In Driver Program
Input: /nycbusiness
Output: /nycbusinesscity
19/08/08 14:05:03 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/08/08 14:05:05 INFO input.FileInputFormat: Total input files to process : 1
19/08/08 14:05:05 INFO mapreduce.JobSubmitter: number of splits:1
19/08/08 14:05:06 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.e
nabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
19/08/08 14:05:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1565254547037_0008
19/08/08 14:05:06 INFO impl.YarnClientImpl: Submitted application application_1565254547037_0008
19/08/08 14:05:06 INFO mapreduce.Job: The url to track the job: http://mathi-VirtualBox:8088/prox
y/application_1565254547037_0008/
19/08/08 14:05:06 INFO mapreduce.Job: Running job: job_1565254547037_0008
19/08/08 14:05:15 INFO mapreduce.Job: Job job_1565254547037_0008 running in uber mode : false
19/08/08 14:05:15 INFO mapreduce.Job:  map 0% reduce 0%
19/08/08 14:05:24 INFO mapreduce.Job:  map 100% reduce 0%
19/08/08 14:05:30 INFO mapreduce.Job:  map 100% reduce 100%
```

Fig. 5. Execution of cityjar

Command to run the JAR file **hadoop jar citynyc.jar /nycbusiness /nycbusinesscity**

The output of citynyc JAR is stored in city folder available under nycoutput in HDFS. The output in HDFS is moved to

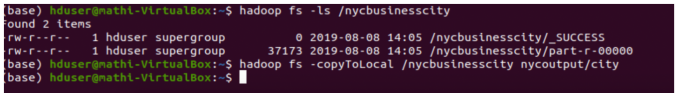

Fig. 6. cityjar execution and output in HDFS location

NoSQL databse(HBase).

MapReduce Task 2 Number of business with respect to Industry in NYC

Similar to above MapReduce, three java class are created in Eclipse and configured as per our logic. After testing the output in Eclipse an executable JAR Industrynyc.jar is created.

Using Hadoop jar command, Industrynyc.jar is executed to run the MapReduce task and output files are stored in HDFS directory. Figure 7 explains the steps involved in successful execution of industry.jar.

Output data is moved to HBase database.

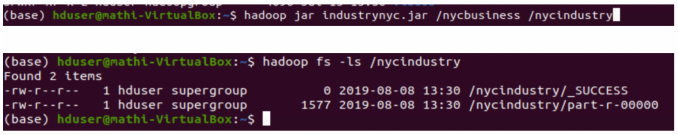

Fig. 7. nycindustry.jar file execution output in filesystem

4) HIVE : To process data in Hadoop, hive is used. It works on top of Hadoop and make query and make process faster. Two tasks are done using hive and explained in detail as follows. Table is created in Hive using create table query to load the data from HDFS, inpath file location is given so that it automatically picks the HDFS location and data will be loaded into Hive table. Figure 8 shows the successful creation and load of data into hive.

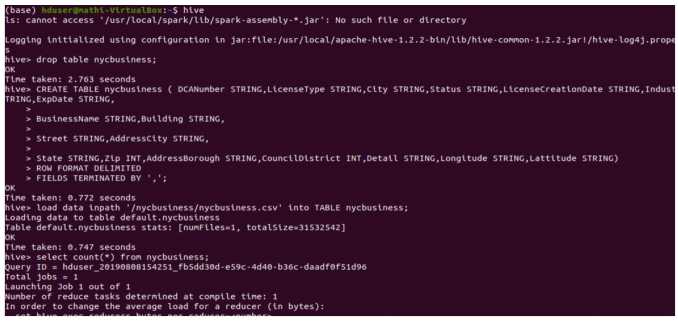

Fig. 8. Load data into HIVE

Hive Task 1 First hive query is to find the number of business in each zip code and find the zipcode with business more than 500. To perform this analysis, data is loaded into hive table from HDFS. Figure 9 shows the execution of first hive task to find the zipcode of NYC with more than 500 business. Output of hive task is stored to /zipcode directory in HDFS.

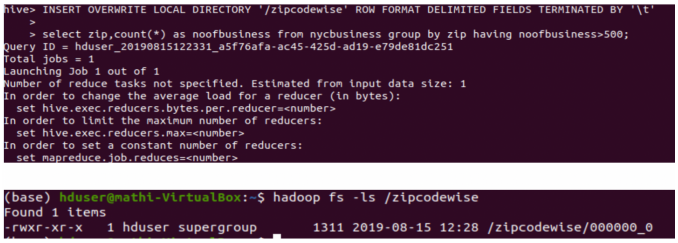

Fig. 9. Execution of First Hive Task

Hive Task 2

Second hive query is to find the business running with Expired license in NYC. To do this analysis Expiration date of License column is chosen and data inside that column is of DD/MM/YYYY format. In hive, the column is chosen and data is split by using / this and year is taken separately. For year value lesser than 2019 are marked as expired and above 2019 as Active Licensed Busines. Figure 10 shows the execution of Hive task 2. Output is stored in HDFS location.

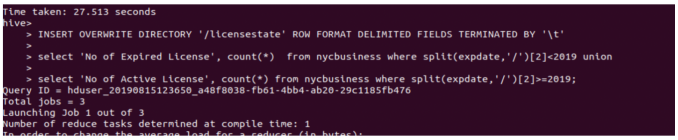

Fig. 10. Execution of Second hive task

5) PIG: Pig is an analytical tool that works on the top of Hadoop. Large dataset available on Hadoop File System are analyze using Pig. To work with pig first data need to be loaded after creation of the table.

Pig Task

There are different type of business running at NYC. To find the number of business for each type in pig, Query is executed to group the business by License type and count is taken for each group. Figure 12 shows the successful execution of pig task. Output of pig task is stored in HDFS location. **Pig Task 1**

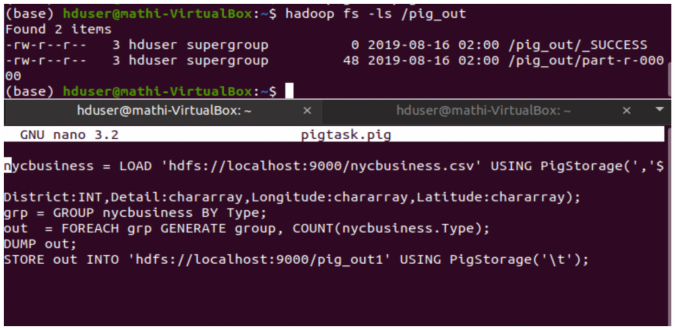

Fig. 11. Execution of pig task

After successful execution of above task, data is moved into HBase. First table is created in HBase and using import command data is moved to Hbase.

bin/hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.columns=HBASE_ROW_KEY,cf1:count hive_task1 zipcodewise000000_0

6) HBase: **Output of MapReduce task 1 is moved to HBase**

```
(base) hduser@pathi-VirtualBox:/usr/local/hbase$ bin/hbase org.apache.hadoop.hbase.mapreduce.ImportTsv  -Dimporttsv.columns=HBASE_ROW_KEY,cf2:count nyccity /nycbusinesscity/part-r-00000
hbase(main):004:0>
hbase(main):005:0* create 'nyccity','cf2'
0 row(s) in 1.2880 seconds

=> Hbase::Table - nyccity
hbase(main):006:0> scan 'nyccity'
ROW                             COLUMN+CELL
 ASTORIA                        column=cf2:count, timestamp=1565865044867, value=4052
 BAYSIDE                        column=cf2:count, timestamp=1565865044867, value=991
 BRONX                          column=cf2:count, timestamp=1565865044867, value=23914
 BROOKLYN                       column=cf2:count, timestamp=1565865044867, value=49015
 CORONA                         column=cf2:count, timestamp=1565865044867, value=2099
 EAST ELMHURST                  column=cf2:count, timestamp=1565865044867, value=907
 ELMHURST                       column=cf2:count, timestamp=1565865044867, value=1578
 FLUSHING                       column=cf2:count, timestamp=1565865044867, value=4597
 FOREST HILLS                   column=cf2:count, timestamp=1565865044867, value=1094
 JACKSON HEIGHTS                column=cf2:count, timestamp=1565865044867, value=1005
 JAMAICA                        column=cf2:count, timestamp=1565865044867, value=4889
 MASPETH                        column=cf2:count, timestamp=1565865044867, value=1085
 NEW YORK                       column=cf2:count, timestamp=1565865044867, value=36743
 OZONE PARK                     column=cf2:count, timestamp=1565865044867, value=1448
 RICHMOND HILL                  column=cf2:count, timestamp=1565865044867, value=1079
 RIDGEWOOD                      column=cf2:count, timestamp=1565865044867, value=1689
 STATEN ISLAND                  column=cf2:count, timestamp=1565865044867, value=10932
 WHITESTONE                     column=cf2:count, timestamp=1565865044867, value=856
 WOODSIDE                       column=cf2:count, timestamp=1565865044867, value=1730
 YONKERS                        column=cf2:count, timestamp=1565865044867, value=1157
20 row(s) in 0.1090 seconds
```

Fig. 12. Moving data from HDFS to NoSQL database

Similar to above task of moving data from HDFS to MapReduce, remaining task data are moved to HBase and Figure 13 shows the list of tables created for Pig, Hive and Mapreduce task.

```
hbase(main):016:0> list
TABLE
hive_task1
hive_task2
newtable
nyccity
nycindustry
pig_nyc
tab3
7 row(s) in 0.0270 seconds

=> ["hive_task1", "hive_task2", "newtable", "nyccity", "nycindustry", "pig_nyc", "tab3"]
hbase(main):017:0>
```

Fig. 13. List of tables in HBase

7) Moving data from HBase to Local: After successful execution of the MapReduce, Pig and Hive task output of those task are moved to HBase and the data in HBase are exported to local directory to visualize the data. Tableau is used to visualize the data.

To automate the process from loading and analysing the data between HDFS, Pig, Hive, Hbase and Local file system nyc.sh shell script is created.

IV. Results and Visualization

This section discuss about the output of the MapReduce, Pig and Hive task stored in HBase and exported to local system later Tableau is used to visualization the output to support the result.

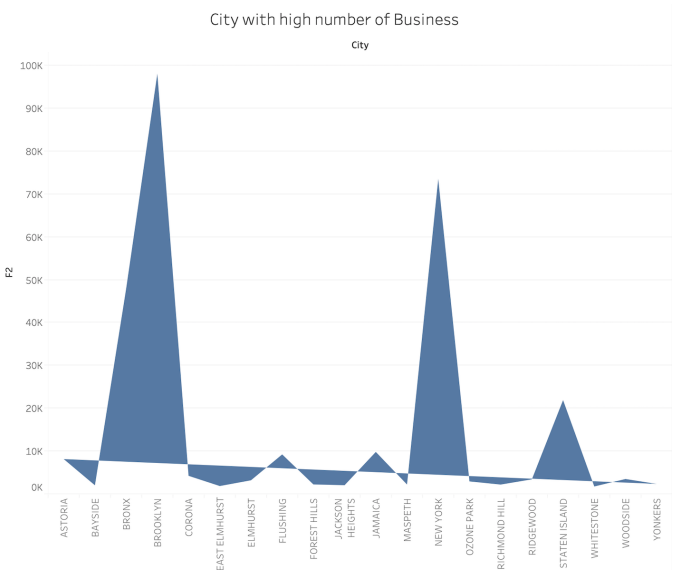

Fig. 14. No of business in city

A. MapReduce Task Output 1

Top 20 city with high number of business is analyzed in MapReduce task 1 and found that more number of business is running in Brookyln followed by New York. Whitestone is having very less number of business.

B. MapReduce Task Output 2

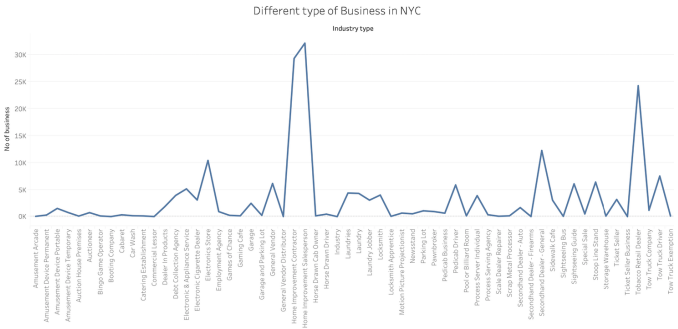

Fig. 15. Different type of business

Figure 15 shows the type of business running in New York city with count. From the graph it is clear that Home Improvement Salesperson job is very high followed by Tobacco retailer dealer. Motion Picture Projectionist business count is very less in NYC.

C. Hive Task Output 1

From the output, Figure 16 is visualised and it shows the number of business running per zipcode. This will be helpful to start a new business in NYC to find in zipcode a new business can be started. 11220 zipcode has very high number of business and 11377 has the lowest number of business.

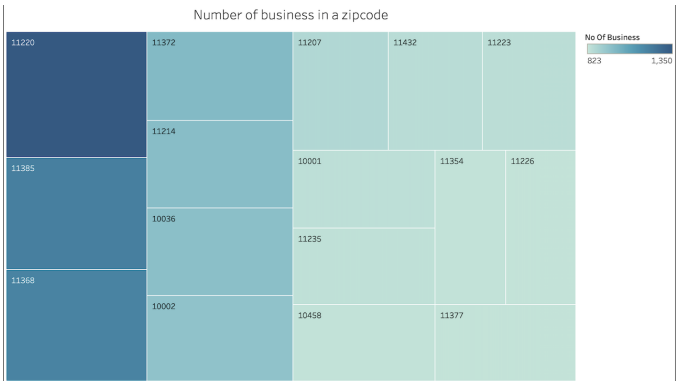

Fig. 16. Number of business with respect to Zipcode

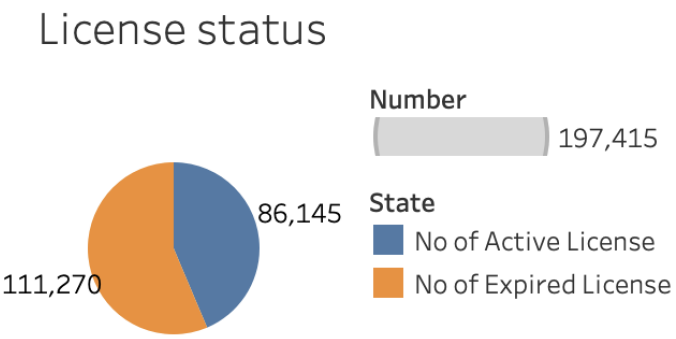

Fig. 17. License Status of Business in NYC

D. Hive Output Task 2

With the output of the Hive task, Figure 17 is visualized and found that number of expired business in NYC are high on comparing to the Active Licensed business. This data can be used by NYC Government to take action on the business running with expired license.

E. Pig Task Output

After executing the pig task, the Figure 18 is generated using the output file and it is clear that Inividual business running in NYC is less on comparing to Business. It can be concluded that Business running by organization is very high on comparing to the Individual ownership in NYC. Business organisation in New York City are high with count more than 120,000 and Individual count of 70,000

V. Conclusion and Future Work

The core of this project is to analyze the New York City Legally operating business dataset to help the Entrepreneur, business organization and Government. As the dataset is too large, Big data tools like MySQL, MapReduce, Hive and Pig, HBase, Python and Tableau for cleaning, analyzing and visualizing the data. Different analyzing task has been performed on dataset like finding the city with high number of business running, Zipcode related business, finding the business in NYC

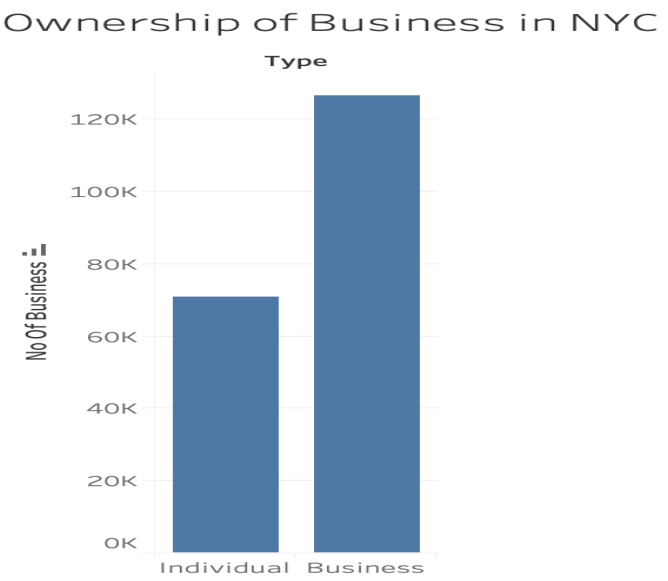

Fig. 18. Business by organisation and Individual in NYC

which has high demand and to get information on business operating in NYC with expired license. The analysis finding will be helpful for Business organization and individual who like to start a new business in NYC can be benefited.
In future, Apache spark can used on this dataset to get high performance on comparing to MapReduce [7].

References

[1] Thomas, L. D. W. and Leiponen, A. (2016) Big data commercialization, IEEE Engineering Management Review.
[2] Koscielniak, H. and Puto, A. (2015) BIG DATA in Decision Making Processes of Enterprises, in Procedia Computer Science.
[3] Loebbecke, C. and Picot, A. (2015) Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda, Journal of Strategic Information Systems.
[4] Erevelles, S., Fukawa, N. and Swayne, L. (2016) Big Data consumer analytics and the transformation of marketing, Journal of Business Research.
[5] Yazici, M. A., Kamga, C. and Singhal, A. (2013) A big data driven model for taxi drivers airport pick-up decisions in New York City, in Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013.
[6] Wang, G. et al. (2016) Big data analytics in logistics and supply chain management: Certain investigations for research and applications, International Journal of Production Economics.
[7] Zaharia, M. et al. (2016) Apache spark: A unified engine for big data processing, Communications of the ACM.
[8] Farook, S., Lakshmi, G. and Tarakeswara, B. (2016) Spark is superior to Map Reduce over Big Data, International Journal of Computer Applications.
[9] Bao, C. and Cao, M., 2019, March. Query Optimization of Massive Social Network Data Based on HBase. In 2019 IEEE

4th International Conference on Big Data Analytics (ICBDA) (pp. 94-97).

[10] Pol, U. R. (2016) Big Data Analysis : Comparision of Hadoop MapReduce , Pig and Hive, International Journal of Innovative Research in Science, Engineering and Technology.

[11] Jain, A. and Bhatnagar, V. (2016) Movie Analytics for Effective Recommendation System using Pig with Hadoop, International Journal of Rough Sets and Data Analysis.