ANALYSIS OF LONDON HOUSING DATA

In Hadoop Using MapReduce Technique

Aswin Sivam Ravikumar
x16134621
National College of Ireland
Masters in data analytics

Abstract— **This electronic document is about the London housing data by many measures London is one of the greatest countries on the earth. during the years between 1997 to 2012 the London's economy had doubled in size growing from 150 billion to 350 billion so here we can see that there is a rapid increase in London population ranging between 8.6 million to 9 million so this rapid growth result says that London has more youthful population. But this success and positives has also some challenges the fall of affordable homes and apartments so the shortage of housing makes the markets price higher so to know the prices I have analysed the London housing data for the years between 2013 and 2014 it contains vast amount of data so I have used Hadoop distributed framework first the data is stored to the SQL from excel and they are loaded to the HDFS on the Linux machine to do the parallel processing after that map reducing technique is processed and the output will be analysed using the qlik sense and using the Python libraries like pandas, script.**

Keywords-hdfs, map reducing, SQL, Linux, Qlik sense, Python libraries

1. INTRODUCTION

What's the future real estate in London? This may be an interesting question but this question can be answered globally around the world here if we look out the primal real-estate we will find the most of the residential properties are bought by the outsiders of the London the majority of the peoples will already buy an apartment or houses in the London or some of them are already planning to buy it's not a social media thing between them but it has become a trend, it's a global trend looking at a places like London as a places to buy property the demand has always increased here because due to the seismic earthquake caused and also the political changes in some countries. So, the investors need a better place not only to invest their money they need some basic aspects like if there is a city that has a low crime, has a language that they can easily understand, have a police force that they trust, have a democracy and stable currency value it can be considered as the place in part of the world that they can operate their business easily so these places have more demands in the housing.

In London the housing prices data, their location will be registered so the government will gain large amount of data when the vast amount of data is generated the data can be explored deeply and by analysing them the people can easily find the affordable houses to their budget limits so these analyzation and visualisation will helps them know the current market values so according to the market values people can make a plan where London's year 2013-2014 dataset is the best-suited dataset for this project as it contains the no of attributes that can be used in the Hadoop distributed environment the dataset is collected from the Kaggle it's an open data set repository so the data can be used for the exploration as well as the analysis process.

To explore the London real estate data, the real estate data will be first stored in the MySQL and then it's exported to the Hadoop in HDFS environment and then mined using map reduce technique so it needs to answer the following business requirements and the business queries.

1.) what is the average price to buy an apartment in each region?

2.) which is the demanding region in London?

3.) which is the most expensive and least expensive region in London?

4.) what is the percentage of houses sold on the town in yearly basis?

The following section will give the details of the related work by discovering the London's data set and how I have effectively used the data set in the Hadoop distributed environment using the map reducing technique. The below passages will show the methods I have used in the project to analyze and visualize the dataset and their obtained results and conclusions are mentioned.

2. RELATED WORKS

London’s real estate housing data is an open data available in many data repositories it’s a public data set used by many researchers so by analyzing them many useful hidden insights can be derived and used in the business developments. The first was done by the Aaron in Imperial college London they predicted the price of houses using machine learning they have created a mobile application so that it can generate the housing prices but it’s a prediction various regression methods Gaussian processes is one of the methods it’s one of the flexible and probabilistic models. The author states that the problem is more successful and it has the ability to produce the predicted values more accurately when compared to the other prediction models. This method has been separated in two ways which mean the application developed by them separated in two ways one is a client and the server side so the heavy computations will be done on the server side (analyzing) and the visualization will be done on the client side. It also has defects more factors also have chances to affect the prediction, for prediction, the information should be accurate but mostly it will have the lack of information.

Steven c. Bourassa, Eva Cantoni, Martin Hoesli written a paper called spatial dependence, housing sub markets, and house price prediction so this paper contains some of the alternative methods here they have used the spatial dependence, hedonic price models, geostatistical models, lattice models, mass appraisal so this approach will allow the sub markets which vary from the house to house so by having this model the author states it will have accuracy.

Byeonghwa Park and Jae known bae have used the machine learning algorithms for the house price prediction here they have used the naïve Bayesian and AdaBoost and they have compared both the algorithm’s based upon the accuracy performance so what they say is this model will assist the house seller and the real estate agent to make their decisions better and they have also used another algorithm named RIPPER so depends upon the accuracy the house prediction will vary

The above-mentioned users only have used the machine learning algorithm’s and some statistical models as they failed to use the Hadoop environment. In my perspective Hadoop environment have created a revolution in big data industry it helps to analyse the big data so easily how its handle its separate the data accordingly and sends the data to the different nodes, so I have chosen map reducing technique to analyse the London housing data, map reducing is one of the most efficient techniques used all over the big data world.

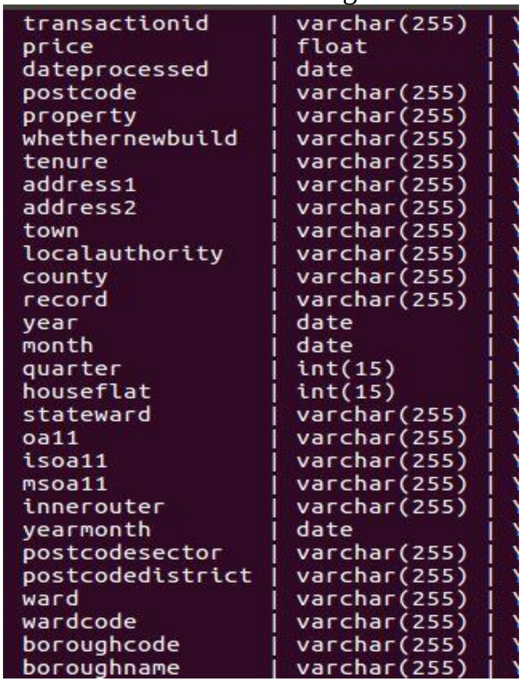
3.METHODOLOGY

Description of dataset

The dataset consists of 29 attributes and 2 million instances. Some of the main attributes used is price, whether the property is new build or an old one, their post code, year so these would be more helpful to know about the house prices and at which location they are present here they have given the post code using the post code we can derive the latitude and longitude so by using the obtained latitude and longitude we can derive the location.

These attributes are sufficient to explore the data and helps to find out the hidden information present in it, the price attribute is one of the most important with the help of that we can find out the average price for the house present in the London by using the real estate we can find out the approximate price per region so the attribute price and the region plays a key role in this dataset.

The schema of the dataset is given below.



```
transactionid    | varchar(255) |
price            | float        |
dateprocessed    | date         |
postcode         | varchar(255) |
property         | varchar(255) |
whethernewbuild  | varchar(255) |
tenure           | varchar(255) |
address1         | varchar(255) |
address2         | varchar(255) |
town             | varchar(255) |
localauthority   | varchar(255) |
county           | varchar(255) |
record           | varchar(255) |
year             | date         |
month            | date         |
quarter          | int(15)      |
houseflat        | int(15)      |
stateward        | varchar(255) |
oa11             | varchar(255) |
isoa11           | varchar(255) |
msoa11           | varchar(255) |
innerouter       | varchar(255) |
yearmonth        | date         |
postcodesector   | varchar(255) |
postcodedistrict | varchar(255) |
ward             | varchar(255) |
wardcode         | varchar(255) |
boroughcode      | varchar(255) |
boroughname      | varchar(255) |
```

Data processing

The London housing data 2013-2014 is downloaded from a Kaggle as we all know that Kaggle is an open data set repository I have downloaded the data and stored on a local disk. After storing the data on my machine by using the MySQL I have created the database and table called housing dB, housing TBL and by using the script I have loaded the data from local disk into an SQL database. Before that, I have used the PIG script this script would allow you to remove the duplicate and null values so the process would be same as the ETL and then loaded to the into SQL table which I manually created.

The second process is storing the data from pig to the MySQL under the particular circumstances using the appropriate schemas.

Why choosing MySQL because it helps to store the data easily it has the speed, reliability, data integrity and scalability and queries are simple, easily executable.

Why Hadoop because it's a distributed framework and more powerful in the terms of computation the parallel processing can be done easily.

By combing the MySQL and Hadoop we will have a high scalability so more powerful system is formed with the tremendous speeds so it would be more helpful by choosing this type of environments.

The dataset from MySQL is loaded to the HDFS, it's a distributed file system it consists of the single name node and a master spaces which manages every process happening over the entire system. How the data will be stored in HDFS

The data from MySQL is loaded into HDFS using the sqoop, is the environment used to transfer the data from the relational databases into the Hadoop environment.

After storing the programming part is done using the eclipse the map reduce is done here the map reducing has three steps one is the driver, mapper and then reduced in reducer.

Map reducing programs are performed and executed on the java eclipse environment in the driver part the input and the output paths are given.

Driver

- INPUT- DATA TO BE LOADED - the path of the file will be given.
- OUTPUT- whether the output file will be stored on the local disk.

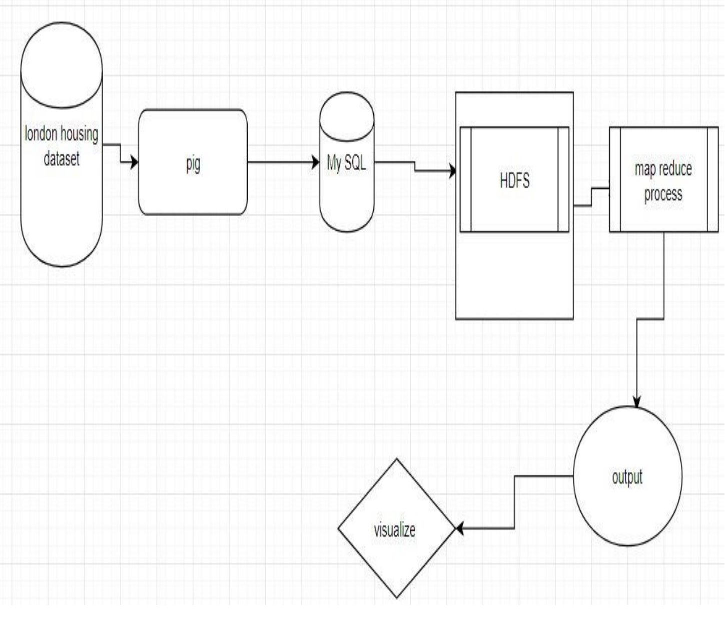
Mapper

- Mapping the values or whatever we need the column number will be specified and the program will fetch it accordingly.
- The main aim mapping is it takes the data in one format and it converts into another format so that would-be breakdown into tuples.
- The process will be taken place in different multiple nodes.

Reducer

- The data comes from the mapper to the reducer so it combines the key values and the values will be reduced the output will be reduced into a new form and the output data will be stored into the hdfs.

Data processing steps



The above diagram shows the data processing steps and the overall process what happened this project.

Map Reduce

Mapper 1
Input- price, county
Output
Key- county
Value- price

Mapper 2
Input -price, town
Output
Key-town
Value-price

Mapper 3
Input- county
Output
Key- county
Value- county (no of times occurred)

The data which is stored in the HDFS will be accessed through the java environment which we use here is eclipse where the map reducing task takes place. The main function what happens here is the file present in the Hadoop is split into key values so that the MapReduce process will take place how the hdfs will separate is it will separate the files into 64mb blocks.

The eclipse is an environment where we can perform the MapReduce for that we need to export the jars from the Hadoop to the eclipse environment and start your coding. I have already mentioned the process of MapReduce in the data processing part.

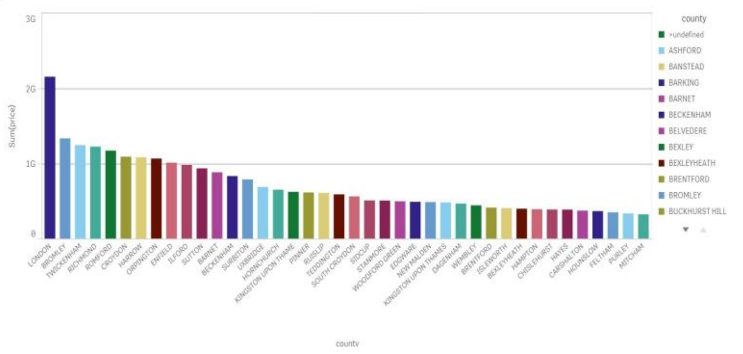
In MapReduce 1 the input key attribute is county and the price so the key values are passed into the reducer and the prices are summed to find out the average price in the county.

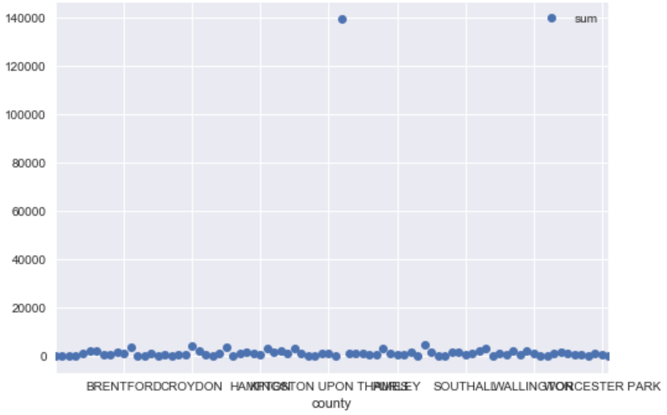
In MapReduce 2 I have used the word count so the mapper will map and send to the receiver where the attribute I have used is the county by having the word count we can find which place has sold the more houses.

The output from all the MapReduce programs will be stored in the HDFS after than we need to convert into the normal text files so the visualization process will be made easily using the qlik sense and the python libraries.

The output of the map reduces 1 gives the county and the average price for each county here the below image shows that the x axis consists of county and the y axis consists of the average price that each county has.
So here is that clear London have more housing price when compared to other countries Barnet and the UK bridge seems to have the average price of the housing and the lowest average price seems to be in the Mitcham. So depends on this the user can decide that where they need to buy a housing its according to their personal budgets.

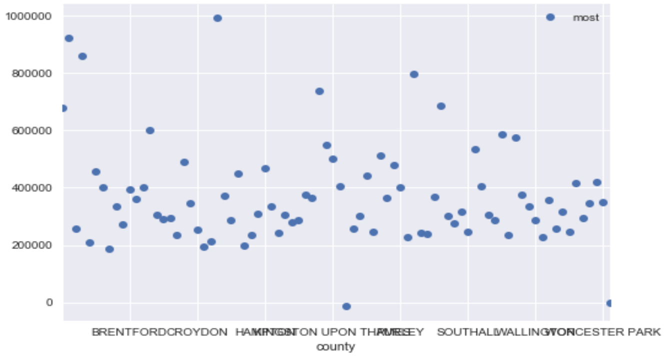
VISUALISATION

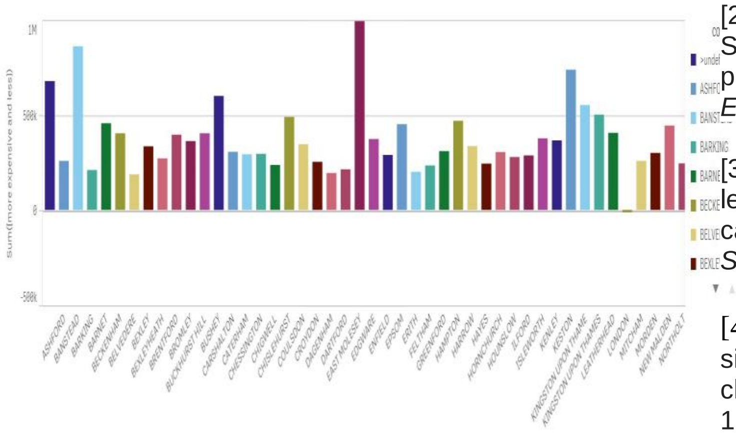




Output 2

The output 2 gives which plays have sold most of the houses so that I have used the word count program so that we can find that which place has sold the more houses so I have taken the town as an attribute and performed the MapReduce.
In mapper, it maps all values and sent to the reducers and its forms a new form of output and its visualized in the qlik sense.





The output shows that which places the houses are sold so from the graph it’s clear that the place named east Molesey have the more houses so the investors have the more options to buy the houses there.so by this graph, the users will have a brief idea where are the more houses left out.

4.CONCLUSION

With the help of the map reducing and Hadoop the London’s housing data have been analysed, and their respected outputs are visualised using the qlik sense and the Python libraries by analysing and the visualisation we have found that which places have the highest and the lowest houses, and what’s the average price to buy a house in the London, and which is the most expensive and which is the least expensive place. At last the visualized graphs shows the answers for the business question that we have raised.

5.RELATED WORKS

The housing data will be updated for every one year or two years as the data set and the data size increases it would be more helpful for the users to predict the house prices not only to predict the house prices the Hadoop have developed more when compared to the MapReduce many high-level techniques have increased like the Apache spark and Apache kudu when we this type of techniques the data can be more effectively managed and more efficiently analysed.

6.REFERENCES

[1]. Ng, A. and Deisenroth, M., 2015. *Machine learning for a London housing price prediction mobile application*. Technical Report, June 2015, Imperial College, London, UK.

[2]. Bourassa, S.C., Cantoni, E. and Hoesli, M., 2007. Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics*, *35*(2), pp.143-160.

[3]. Park, B. and Bae, J.K., 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), pp.2928-2934.

[4]. Dean, J. and Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), pp.107-113.