Analysing the Insights of Visitors for Hotels in the European Countries

Piyush Narkhede
MSc Cloud Computing
National College of Ireland
Dublin, Ireland
x17151538@student.ncirl.ie

***Abstract*—This paper gives an idea about hotel business insights by using data analytics techniques. The dataset for the project is taken from Kaggle and it is used to check the hotel reviews, people visiting for the country and some future prediction for tourism management. Our dataset is for the hotels in Europe contains hotel reviews and number of customers visited and reviewed. We used this data to analyze insights based on hotels. We are going to show hotels in which people gave negative reviews so that visitor will know which hotel they should select for a stay. This report will be helpful for the hotels to improve their customer segments and for the government who is planning for more visitors in their country. For that, we are using Python and Microsoft Excel for cleaning, Hadoop (Apache Hive, MariaDB, Sqoop) for deriving insights and Tableau for visualizing that data into graphical images. In future, if we get data from each county in the world then we will make analysis and will show outcomes whether which hotels are best in which country based on its type by using big data analytics.**

***Index Terms*—business insights, data analytics, hotel reviews, Kaggle, predictions, analyze, Python, Microsoft Excel, Hadoop, big data**

I. Introduction

Recent years, people are traveling around the world for various reasons such as business, picnic, holiday, family tour so many reasons. In Europe, people are mostly coming either for holiday or business. In 2017, approx. 87M tourists visited in France, that was a breaking record; 58.3M went to Italy; also the Netherlands received 17.9M visitors. // ref so, there is a need to see about any stay options at the destination country. If any visitor is going first time to visit a specific country then that visitor dont know about hotels and other options. This report will at least help visitors to find which hotels are best and which are worst based on historical customers reviews. Also, the government also get an idea of how many visitors coming to stay in their country with their nationalities. So, that they can plan about future services for nationalitys visitors.

We are using Hadoop for MapReduce and hive for query processing. With this output, it is important to show that output in a meaningful way, so we used Tableau [3] and MS Excel to show visualization and graphs for showing results. Proper analysis has to be done on an existing dataset so that we can show proper results for visualization. We have taken dataset from Kaggle [1]. Kaggle is the platform where all types of data are present. It is a mine of data and a key factor for data scientists. We got our dataset from Kaggle which is Hotel reviews in Europe. All details about this dataset and processing we will see in further sections.

1) Objective and motivation: Objective for this topic is to work with big data environment and come with some outcomes which will be useful for local or international tourists. We can predict whether which hotel should they choose for stay.

Business Question for this paper is : Would be the business insights if any data analytics technique (Map-Reduce) is applying on large number of dataset(Hotel Reviews) and generated output, useful to take decisions?

Question states that whether some business insight by using this project, will be possible or not? We tried to show all possibilities by using Map-Reduce and Hadoop in this report.

In further distribution of paper are as follows: Section II contains related work done with same topic; Section III is explaining implementation in details such as data preprocessing, implementation with architecture flow diagram; section IV is explaining results and outcome got from implementation;and last section concludes paper with giving some future direction for this topic.

II. Related Work

Hadoop [5] is fastest computation platform which we can use to process and store data into storage. It is one of the best project of Apache. Hadoop uses mapreduce technique which we can use to process with big data. [7] implemented hotel recommendation system based on hotel reviews. By using type of traveler, location and visitor amenities, author trying for recommendation. For this system, author used text mining and semi-supervised learning algorithm to group features. So that recommendation will be accurate. Author doing data preprocessing, after that operation using data and recommendation based on result of algorithm.

[6] this research is based on customer satisfaction. On the basis of online reviews, their system detects some low features using text mining and giving that notification to hotel administration. So, that hotel administration staff will be taking some decisions to improve their facilities.

[4] developed Hadoop framework based recommendation system for hotels. Author implemented recommendation system for users, sentiment analysis for hotel administration

staff whether get to know about services based on sentiment analysis. Architecture of their proposed system is shown in figure. Author started with cleaning of dataset, after that they extracted all keywords from all the data. They took some rating and gave it to processing unit. By using some extraction techniques, they are showing outputs as recommended hotels.

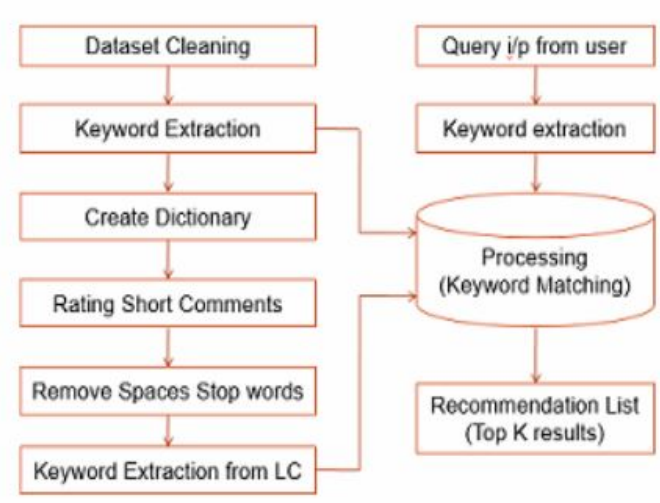

Fig. 1. Proposed system by [4]

[8] proposed a system for evaluating historical reviews based on hotel from multiple sources like blogs and websites. They were analyzing data and creating graphs on basis of data because always graphical representation is better than other representations.They used intelligent learning techniques like neural networks, k-means and web crawlers to analyze graph and it can be customize output as per users needs.

III. Methodology

A. Technical Specification:

For processing and working on data, we use some data analytics and technologies as follows:

- **Programming Language:** Python
- **Data Cleaning Tool:** Anaconda:Spyder, MS Excel
- **Framework:** Map-Reduce (Hadoop)
- **Database:** MariaDB(MySQL), Hive
- **Data integration:** Sqoop
- **Scripting:** Shell Script and HQL
- **Visualization Tool:** Tableau, MS Excel

We used MS Excel and Spyder [2] in which we run python program for data cleaning. There are many tools available for cleaning but these are mostly used and recommended tools for data cleaning. For storing data we used MariaDB which is a similar tool like MySQL and for using that data for processing, we used sqoop to transfer the dataset to the hive which we are using for query processing. After getting that output, we are visualizing that output in Tableau and MS Excel so that we can show those outputs to readers. After final output we use HQL and shell script to store all commands and queries in a script file.

B. Data preprocessing:

For processing data, it is important to clear that data by using any tool. We use Anaconda: Spyder tool to run python code for a dataset. It is a tool to process data, advanced editing, and analysis with debugging and profiling features for data. It is a development tool with the data exploration and interactive execution with deep inspection of data. We wrote python code in Spyder and started to clean null values and removing spaces from our dataset. Then we put that dataset into the MS Excel which is a global tool used for analyzing and processing data developed by Microsoft. We distributed column into different columns to separate values. We removed some columns which we dont want and removed all complexities and cleared that data. In python, we used ”pandas” package which provides features for cleaning such as ’dropna()’, strip(), isnull(), etc.

The dataset contains approximately 3lacs rows and 16 columns with data their meanings as follows:

Attribute	Description
id	it is primary key for table (integer)
hotel_country	country where hotel is located (stirng)
review_month	month of review(string)
review_year	year of review (integer)
average_score	average of score which hotel that got(integer)
hotel_name	name of the hotel (string)
reviewer_nationality	nationality of reviewer (string)
negative_words	negative wordcount which reviewer gave (integer)
no_of_reviews	total number of reviews (integer)
reviewers_score	score given by reviewer out of 10 (integer)
days_since_review	days from review (integer)
trip_type	reviewer’s trip type (string)
customer_type	type of customer (string)
night_stay	stay of reviewer (integer)

We used this data because this contains negative and positive review count so that it was easy to understand whether that hotel is worst or best. Big challenge we faced that was, only one column in dataset was in JSON format and my whole data in comma separated value(CSV) format so we used MS Excel to divide whole data by using text to column using separator as ’,’ and we used find and replace to remove ’[’ and’]’ brackets. After this process we again clean that data by python so that we removed all null values and spaces from columns.

From our dataset, we are using hotel name and country for results and with those parameters some more parameters like customer type, nationality, trip type and year when they visited so that we can predict for future visitor or change services as per result.

C. Map-reduce using Hadoop:

Map-Reduce is framework which deals with big data. It is used to precessing and generating big data. Map reduce is mapping data in first stage. Then is shuffles whole data and reducing to another stage. So that phase we are getting key

value pairs and then it is easy to use those values for next processing.

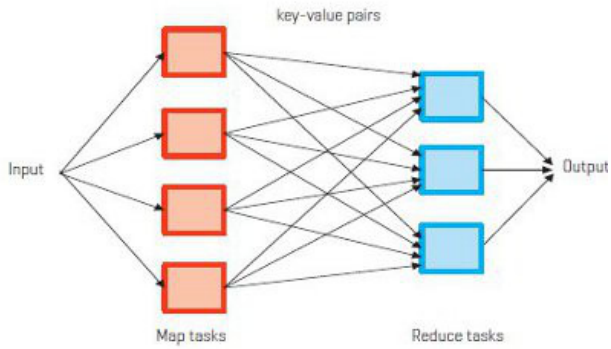

Fig. 2. Working of MapReduce

After cleaning all dataset, we have to process on that data. So, we used MariaDB which is platform like MySQL. It is database as a service for storing and processing data. So, we imported our dataset into MariaDB. We planned to work with mapreduce so now our task is to use data from MariaDB to hive.

So, we used sqoop for integrating dataset from MariaDB to hive. Sqoop is a tool present in hadoop which we installed under hadoop group. Sqoop took those dataset and passing it to hive by using tunneling process. Sqoop is taking those data into hadoop (hdfs/dfs) and then importing those data into hive using –import parameter in command. Hive got dataset from hadoop by using sqoop. Now we started work with hive. Apache Hive is data warehouse software which provides data analysis, query processing and data summarization. We are using hive because our dataset contains more that 3lacs of data and hive is best software to deal with big data than MySQL //ref. After processing all data we are storing those data into output file which we want for visualization.

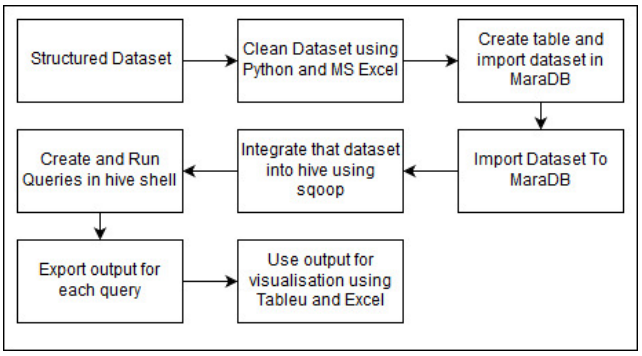

Fig. 3. Work flow diagram for this project

IV. Result

For visualization, we used Tableau tool which is mostly using for showing data results into graphical format. Explaining result iin graphical format is easy and better than other techniques. We can see our outcomes came from out project:

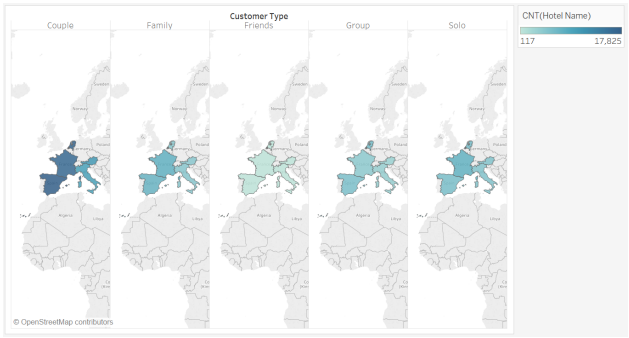

Fig. 4. Map of type of customer visiting to given countries(Visualization 1)

1) Visualization 1:: As we can see in map, most of the couples are visiting France, Spain and Netherlands than family and solo visitors visiting to that countries. If we see about overall ranking for friends then it is very less than other types. So, it means visitors coming with friends are very less for these countries. Group of visitors are nearly same for all countries so we can't predict that much information than other types. For Italy, there are very less visitors than other country has.

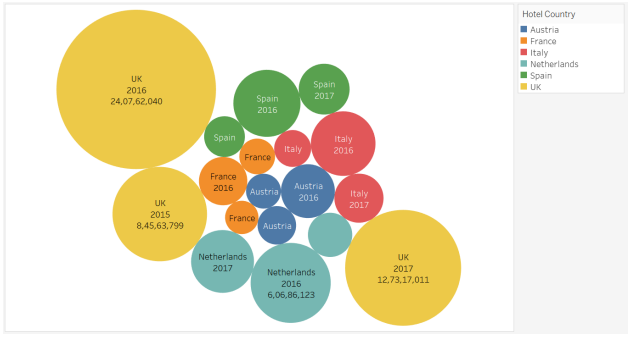

Fig. 5. Visitors visited for country divided by years(Visualization 2)

2) Visualization 2:: If we see about year by year, then their is twice growth in UK visitors for 2016 to 2017. In 2015, there were very less visitors to all countries than 2016. For Italy, there are more visitors than France in 2016 and Netherlands is holding second position for 2016 user reviews or visitors. From this diagram we can easily predict how many visitors are visiting those countries and what plans that countries should do in future for visitors and many more things.

3) Visualization 3:: By seeing that graph for visitors with their visit types, we can see that visitors for leisure trip are most as compare to business trip. For Spain, there are more than 30k visitors are vising for leisure and Netherlands and France are on nearly equal position as compare to Italy and Austria. For France, most of the poeple visiting for business

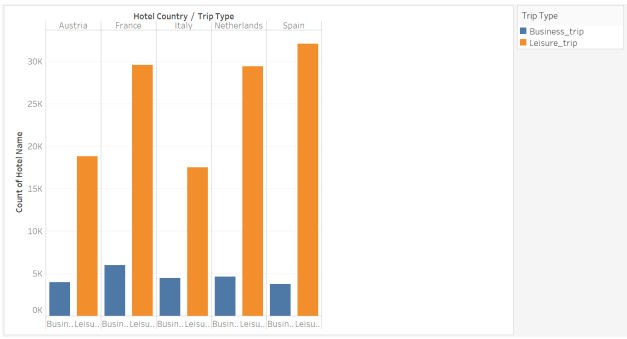

Fig. 6. Visitors who visits countries by their trip type(Visualization 3)

than other countries but still for all countries, visitors for business trips are same. People are coming to visit for Leisure, so that hotel should take decision that some services regarding leisure for visitor should be more than regular facilities which they have.

For all these visualizations, most of the time we are not taking UK as comparator because our dataset has some issue about UK, there are more entries for UK than other countries so that we remove UK from those two visualization. For visitor ranking by year, we took UK because, we were comparing those values inside UK so we can use UK for comparison.

V. Conclusion and Future work

The meaningful factors regarding hotel reviews and customer nationality based on historical dataset has been showed successfully in this report. This report will be helping to take decision whether which hotels are better to stay for what reasons and some positive/negative factors about hotels. We showed all result which calculated by Map-reduce and visualizing techniques.

Another important concept we have seen in this process that if we are using MariaDB(MySQL) or MS Excel for dealing with data, then it takes time to process any operation. But if we are using python language or Apache hive, then it is performing that operation in very less time. Hence, we performed our queries to see output time taking by both methods and we conclude that efficiency for python and Apache hive with Map-reduce is better than MS Excel and MariaDB.

We are already using hive and hadoop(map-reduce) technique for data processing. So, in future, if we get more data like hotel reviews from all over world, we can easily work by just following this report. Also, by this report, it is easy to take decision for visitor, either business or leisure, which hotel is better to stay or for administration, for making changes in their hotel if they are getting negative reviews.

Cloud is again best option to use with map reduce. In this report, we used CentOS and physical system for dealing with our dataset. But in future, if we are getting million rows of dataset then it will not be easy to deal with big data. So, instead of using physical processing units, we will advice to use cloud platform like AWS and Azure which are providing high processing computing(HPC) unit for dealing with big data. Cloud is easy to access and retrieve data with fault tolerance and reliability features so it is better to use cloud instead of physical machine for big data.

References

[1] Kaggle: Your Home for Data Science. [Online]. Available: https://www.kaggle.com/. [Accessed: 17-Aug-2018].

[2] Spyder:: Anaconda Cloud. [Online]. Available: https://anaconda.org/anaconda/spyder. [Accessed: 17-Aug-2018].

[3] "Tableau Software, Tableau Software. [Online]. Available: https://www.tableau.com/. [Accessed: 17-Aug-2018].

[4] Cherapanukorn, V. and Charoenkwan, P. (2017). Word cloud of online hotel reviews in chiang mai for customer satisfaction analysis, Digital Arts, Media and Technology (ICDAMT), International Confer- ence on, IEEE, pp. 146-151.

[5] Europe Made Billions from Tourists. Now Its Turning Them Away (n.d.). Time .URL: http://time.com/5349533/europe-against-tourists/

[6] Jalan, K. and Gawande, K. (2017). Context-aware hotel recommendation system based on hybrid ap-proach to mitigate cold-start-problem, 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), IEEE, pp. 2364-2370. 1

[7] Jian, M.-S., Fang, Y.-C., Wang, Y.-K. and Cheng, C. (2017). Big data analysis in hotel customer response and evaluation based on cloud, Advanced Communication Technology (ICACT), 2017 19th International Conference on, IEEE, pp. 791-795.

[8] Shrote, K. R. and Deorankar, A. V. (n.d.). Hotel recommendation system using hadoop and mapreduce for big data.