Analysis of Yelp Reviews using Big Data Frameworks

Dinesh Narayan Gauda
X17169836
School of Computing
National College of Ireland (NCI)
X17169836@student.ncirl.ie

***Abstract*—Yelp is an online consumer review website which lets consumers post their invaluable comments and feedback. These reviews leverage the wisdom of the crowd and helps the consumer choose from the variety of options available. But, with the amount of data available out there makes it almost impossible for the users to go through all of the reviews and make a decision, the aggregated rating helps but it can be sometimes misleading as well. Another view to these huge amounts of review data with high dimensionality from these websites is that it can provide useful information for the companies which can be used to improve their reputation and thus improve revenue. Thus, this research aims at analysing these large amounts of big data generated from Yelp dataset and find insights such as trends in the market, highest rated business, most popular business, coolest business, and sentiment attached to these businesses. The research makes use of big data frameworks such as MapReduce, Hadoop, Spark, Pig, and hive for the analysis. This project is implemented on Google Cloud Platform. Results of the seven task executed in this project using different big data framework answer the research question.**

***Index Terms*—Big Data Analytics; Yelp Reviews; Google Cloud Platform; MapReduce; PySpark; Hive; Pig; Hadoop**

I. Introduction

In this era of the digital world, how the consumer learns about the quality of the business has changed a lot [1]. Crowd-sourcing websites such as Yelp, Trip Advisor, etc. generate business and product reviews. This gives a consumer lots of information while making a decision and finding the best business/product. The helpfulness of these reviews increases with the number of reviews left. But, with the increase in reviews, it becomes very difficult, tedious and time consuming for the users to analyze all these reviews and extract valuable information from it. This problem requires an analysis that will provide users easy to understand underlying information from these reviews to help them make their decision. There has been a lot of literature on social media analytics that collects data, analyze, and interpret these reviews for multiple causes such as tracking trending topics, popularity, sentiment, opinions about a product/business [2] [3].

A. Motivation

To get a better understanding of different research problems in the field of tourism and hospitality online consumer reviews provided by these websites such as Yelp, TripAdvisor, etc has been studied [4]. The magnitude of the impact of information extracted from reviews not only depends ”on informational content, but also on salience and simplicity of information” [5]. There is evidence from the literature that these reviews have a direct impact on the success of the product/business [6], [7]. It is quite evident from the literature that information generated from these review websites not only have influence on consumer's decision but also the fate of the product/business and thus finding hidden trends and insights will be helpful not only for consumers but also for companies to understand the trending trends in the market, general opinion of consumers about the business/product, negative points about the business/product. A better understanding of these things can help competitors stay ahead in the market and avoid mistakes in the future. Results from the analyses will also ease the process of making decisions for consumers by not making them go through all the reviews manually and provide valuable trends, information, and general opinion through facts and numbers.

The objective of the project is to find meaningful insights from these large amounts of reviews with very dimensionality. Information from the analyses can provide information about what is the general opinion of the customers and what they care which can be helpful in improving there reputation, which will have a direct influence on their revenue. In this paper analyses of data from one of the crowdsourcing websites, Yelp has been done. Yelp is a website where consumers can comment there opinion and feedback about a product/business. Data from Yelp is humongous both in terms of the number of reviews and attributes it provide which makes it impossible for it to be analyzed without the help of big data frameworks. Research question and research objective of this paper are listed below.

B. Research Question

What are the hidden insights in Yelp dataset such as most popular, trending, coolest, and highest rated business in the market and what is the general sentiment attached to it ?

To answer the above research question analyses of the Yelp dataset with the help of Big Data Frameworks such as MapReduce, Hadoop, Spark, Pig, and Hive has been done. This project has been implemented on Google Cloud Platform, and make use of tools provided by it such as DataProc, Google File System (Bucket), Big Query. Tableau has been used for visualizations which have been integrated with Google's Big Query to extract data directly after the analyses.

The remainder of the paper is organized as follows: the next section, Related Work, which reviews related literature on the topic and provides motivation and justification for the proposed study. In Methodology first, the description of the dataset used is listed followed by the data preprocessing done in the project. After that brief information about all the Google Cloud platform tools and big data frameworks used and project setup of the study has been articulated. Then, the results of the study with detailed interpretation have been listed in the section Results. Finally, conclusions are drawn, and limitations and course for future analysis have been discussed.

II. Related Work

Big data analytics has been a profound research paradigm that can analyze a huge amount of data from diverse sources with the help of analytical tools to provide insight and make predictions about the reality [8], [9]. Particularly, with the increase in the big data analytics tools and infrastructure as a service, providers have enabled the researchers to analyze a problem with a lot more ease than ever before. In the study carried out in [7] the authors analyzed the Yelp dataset and concluded that review websites such as Yelp improves the information available about a product/business. The results from there research suggested that rating has a huge impact on revenue of the business and that one-star increase in the rating can increase the revenue by 5-9% and that rating of business impact its revenue 50% more when it has more than 50 reviews. From the viewpoint of companies ratings in these review websites have a direct influence of their revenue, and any improvement in the business can be maximized by a credible user's positive review. thus the companies use these strategies to influence their targeted consumers [10].

The authors in the research [11] made use of a distributed system rather than a centralized system because of its ability to gather large chunks of data from various different sources and process it. The authors in [11] used Amazon's Simple Storage System(S3) along with Apache's MapReduce and Hadoop's HDFS to perform operations on Yelp datasets such as counting average rating, unique tags, and total recommendations. In the research conducted in [12], and [13], processing of Yelp dataset to find out frequent adjectives to describe a business with the help of different interfaces has been done. Visualization of sentiment score of reviews in different colours was shown in these interfaces, but thus these reviews have any effect on ratings of the business/product by comparing sentiment score with ratings was not done which has been done in this research. An interesting fact concluded by the research conducted in [14] in Yelp dataset is a concept termed as 'warm-start bias'. The authors said that the first review that a business receives is highly biased and average first rating is 4.1 whereas the 20th average rating is just 3.69.

Much past research has been done for extracting information from the web [15], opinion mining [16], and review mining [16], [17], [18]. In [19] the authors analyzed the reviews and classified it into positive and negative reviews whereas the authors in [20], and [21] processed review data and determined orientation of the phrases using machine learning algorithms and n-gram techniques. In this digital era, the ever-growing amount of consumer generated textual reviews and the information it contains about the features of product/business and consumer preferences can be extracted using sentiment analysis [22], [16]. Sentiment analysis can be implemented on three different levels review level, sentence level and phrase level. Review [21], [23] and sentence level [24], [25] generalize sentiment of whole review whereas phrase level analysis [26], [27], [28] attempts to extract features of product from the review. In this study sentiment analysis on review level on Yelp dataset has been done along with analysis on the Yelp dataset to extract meaningful information with the help of big data frameworks.

III. Methodology

A. Data Extraction

Yelp dataset was originally put online by the crowdsourcing company Yelp for the students to perform research and analysis on Yelp's data. To perform analysis on data of Yelp website, a subset of Yelp dataset challenge was accessed instead of the original dataset as the original dataset is humongous and subset data met the required minimum criteria of dataset size. Another reason for choosing the subset data was to avoid unnecessary deduction of money for storing and processing of data on Google Cloud Platform. To access the yelp dataset via Yelp API there were limitations of only a few calls per day and bulk download was not allowed, so a subset of a dataset from Kaggle website [29] was downloaded programmatically through the API provided by Kaggle. The dataset downloaded from the website contained data of 5,200,000 user reviews, information on 174,000 businesses for eleven metropolitan areas across four countries and sized around 3.6 GB. For this research four datasets were used as listed below. All the four datasets were downloaded through kaggle API.

1) Yelp Business Dataset: This dataset contained information regarding various businesses. It has around 1,74,568 rows for 13 attributes which included rating, latitude, longitude among other attributes and sized around 130 MB.

2) Yelp Tip Dataset: This dataset contained information regarding tips. It has around 10,98,324 rows for 5 attributes which included business name, date among other attributes and sized around 230 MB.

3) Yelp Review Dataset: This dataset contained information regarding reviews. It has around 52,61,668 rows for 9 attributes which included business name, date, stars among other attributes and sized around 5GB.

4) Yelp User Dataset: This dataset contained information regarding users. It has around 13,26,100 rows for 22 attributes which included average stars, yelping since, name among other attributes and sized around 2.5GB.

B. Preprocessing

The data downloaded through the API pragmatically using R was downloaded as a zip file which was later extracted using R. The data after extraction was in JSON format and so was converted to CSV format using a python script called json_to_csv_converter. Once the data was downloaded and converted into CSV format each file was pre-processed in R. All operations for preprocessing was done in R. Operation performed on each dataset is articulated below:

1) Yelp Business Dataset:

- Double quotes from the record in name address column were removed.
- All empty spaces in each field were replaced by NA and later removed.
- Comma was removed from address records as it was causing problem while inserting data into the hive.
- Neighborhood, and Postal_Code columns were removed as there were not required for analysis.

2) Yelp Review Dataset:

- Records of only business as restaurant and year were filtered, and rest was omitted.
- All special characters along with new line was removed from the text column.
- Double quotes from the record in text column were removed.
- Records matching the business id of the business dataset were only extracted rest were neglected.

3) Yelp User Dataset:

- Empty and rows with NA values were removed.
- Data for the year 2016 was only filtered rest were ignored.
- Double quotes and special characters were removed from elite and name columns.

4) Yelp Tip Dataset:

- Empty and rows with NA values were removed.
- Data for the year 2016 was only filtered rest were ignored.
- Double quotes and special characters were removed from the text.
- A new column named business name was added by joining the business and tip dataset.

Data for the only year 2016 was considered as data size was way too large and would cost more for storing and processing on Google Cloud Platform. A detailed explanation of the project setup, tools and infrastructure used is summarized below.

C. Google Cloud Platform

Google Cloud Platform (GCP), is a cloud computing service provided by Google. Along with infrastructure, it provides a set of management, data storage, computing, data analytics, data visualization and machine learning tools. The account setup is free for a year and comes with an initial balance of 300 dollars which can be used to access services in GCP. GCP was preferred for implementation instead of traditional platforms and local setup for the following advantages of GCP.

- GCP provides all its services as pay as go, which means that the consumers have to only pay for its usage.
- All infrastructure is provided and managed by Google.
- Scaling of infrastructure is possible during development or production. This is also managed by Google and user have nothing to worry about it.
- It provides a list of various tools such as DataProc, Big Query and Google File System.
- Can be accessed from anywhere, and big data tools such as Hadoop, Spark, Pig, and Hive comes pre-installed and configured.
- It provides N worker clusters and range of infrastructure options to choose from as per requirement which makes the processing of Big Data easy.

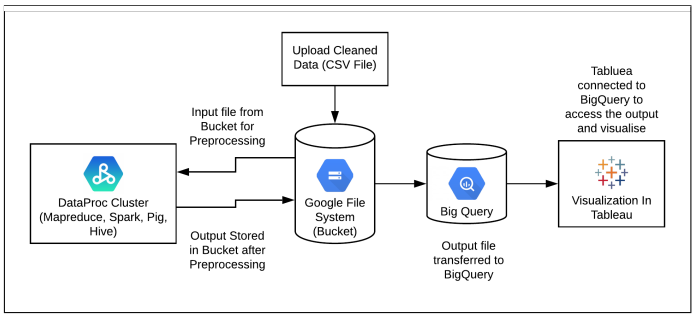

Fig. 1. Project Setup

D. DataProc

DataProc is a cloud service provided by Google for running Big Data frameworks such as Apache Spark and Apache Hadoop clusters. It is a simple and cost-efficient way that can create a cluster in a few minutes that would take days or hours instead traditionally. DataProc can be easily integrated with other GCP services such as BigQuery, and Google File System for data processing and analysis. Dataproc clusters can resize them any time from two to hundreds of nodes in case the load on cluster increase so that the user has more time on focusing on analysing the problem and finding insights rather than setting up the infrastructure. All the above advantages of DataProc over traditional virtual machine made it preferential for implementation over traditional setup. BigData jobs such as a MapReduce job or a Hive job can be run by creating a job on the cluster or directly accessing the cluster as a virtual machine using SSH and executing MapReduce or hive tasks.

E. Google Cloud Storage

Google Cloud Storage is an online file storage web service provided by Google for storing and accessing data on GCP. The services have performance and scalability of Google's cloud and sharing capabilities. It is quite similar to Amazon's S3 online storage service. It is scalable and data can be shared across various services on GCP. This services also come with pay as you go feature and its storage cost are also very cheap which also varies with the location of storage and can be selected per the requirement. The data can be transferred across GCP and even oh Hadoop's HDFS using gsutil. All the above features of Google Cloud Storage made it preferential for storing data rather in a local system.

F. Big Query

BigQuery is a highly scalable enterprise data warehouse that makes data analysis very easy with its inbuilt BI engine. There is no infrastructure to manage and insights can be extracted using SQL with any need of database administrator. Data can be imported in a variety of formats such as Avro, JSON, CSV, etc. The data can be imported from Google Cloud Storage, Google Drive or even can be uploaded from local machines. The output of queries can easily be exported as spreadsheets and reports. All the above features of BiqQuery along with the feature to connect with third-party software such as Tableau gives BiqQuery edge over other traditional databases.

G. Project Setup

The project setup used in this research is as shown in figure 1. The project is implemented using different service provided by Google. First, all data to be processed after cleaning and filtering on R is upload to Google file System (Bucket), then the data is processed on DataProc cluster using different big data frameworks such as MapReduce, Spark, Hive, and Pig here the DataProc cluster created is in one master, two worker node configuration. Once the data is processed the data is again transferred to the Google file system and from there to BigQuery. Once the data is in the table of BigQuery it is accessed through tableau and visualization are done. The files in the Google file system can be accessed by adding a prefix of gs:// to the directory. The files can be transferred from DataProc to Bucket and visa-verse with the help of gsutil, It is better than Scoop which is also used to transfer data from relational databases to Hadoop as it is very fast in terms of processing speed. The Big data jobs can be executed using two ways, by creating a job for respective framework or by directly accessing the cluster as a virtual machine and then executing it. The detailed flow chart that is followed in this project is as shown in figure 2. All BigData jobs in this project have been executed using a virtual machine instance with the help of SSH. Coding for MapReduce design pattern has been done in the local virtual machine then the Jar file has been uploaded on the Google file system for future processing. Coding of PySpark has been using Jupiter on DataProc cluster itself with the help of GCP window's SDK and Web Interface connection to the DataProc cluster using SSH.

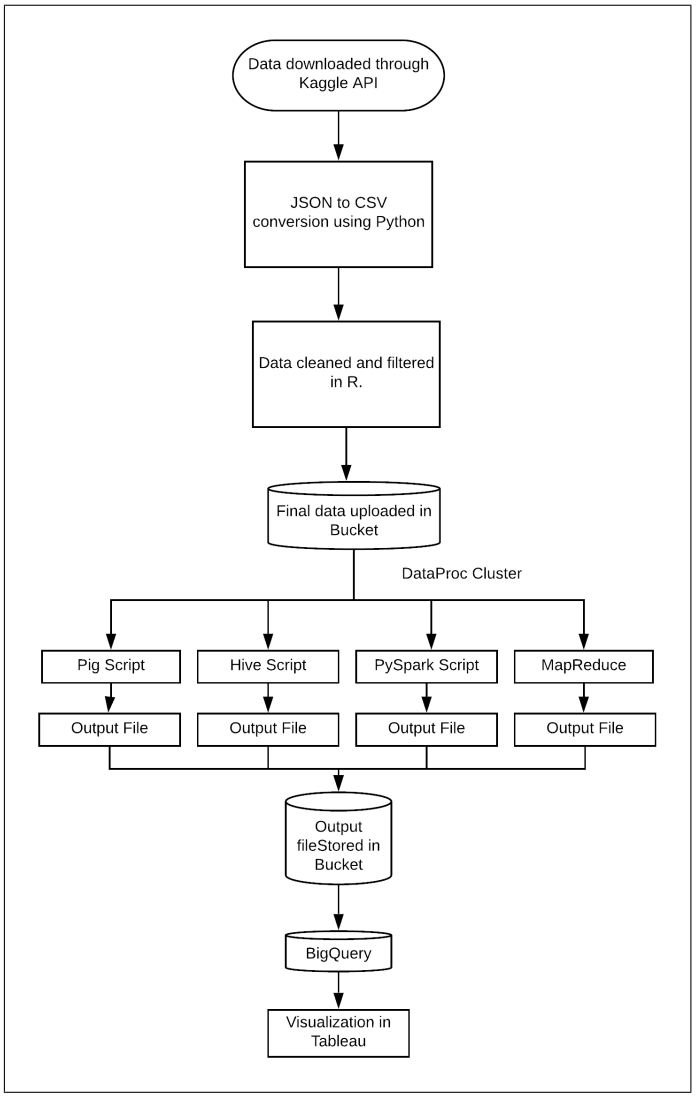

Fig. 2. Project Flow Chart

IV. Results

In total seven tasks are executed using different BigData frameworks in this project, a detailed summary of each is articulated below.

A. Task 1: Most popular dishes in Las Vegas

The first task is to find out the most popular dishes in Las Vegas. To answer the above question, a PySpark code has been written. The output of the results from PySpark code has been visualized as a word cloud wherein, size of the word defines its popularity, i.e. the larger the word size more popular is the dish. As per results are shown in figure 3 Pizza is the most popular dishes in Las Vegas followed by Sushi, lobster, Steak, and others. The results of this query will help the consumers to decide which dishes to try in Las Vegas as per the popularity and answer the research question listed in I-B.

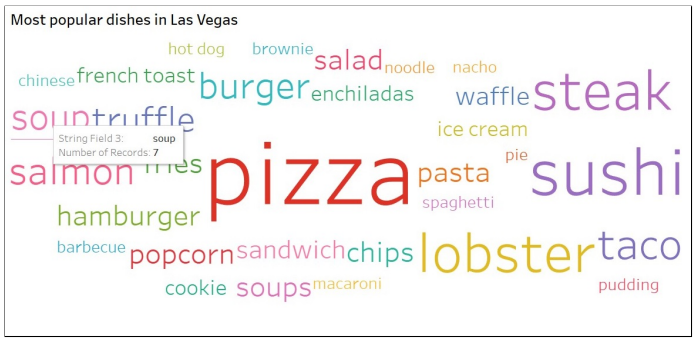

Fig. 3. Most popular dishes in Las Vegas

B. Task 2: Relationship between Sentiment Score and Average Rating

The task here is to find if there is any relationship between sentiment score and an average rating of the reviews across different cities of the world. To answer this question a PySpark code has been written wherein, sentiment score has been calculated of the reviews using positive and negative words in the review per city. The sentiment score ranges from -1 to 1 where, values near -1 means a negative review and values near +1 means a positive review. The average rating is calculated per city and compared with the average sentiment score of the city with the help of a bar chart to find the relationship between these two features. The results as shown in figure 4 shows that there is a multiplex relationship between two features and in some city rating is proportional to sentiment score whereas in some there is no relationship between them. This task helped in understanding the relationship between the sentiment of the consumers and the rating of the business. The results can be used by the company to clear idea about what is consumer's opinion about there business/product in a particular city and try and improve the reputation of the business. The results answer the research question listed in I-B by exploring the relationship between sentiment score and an average rating of a business.

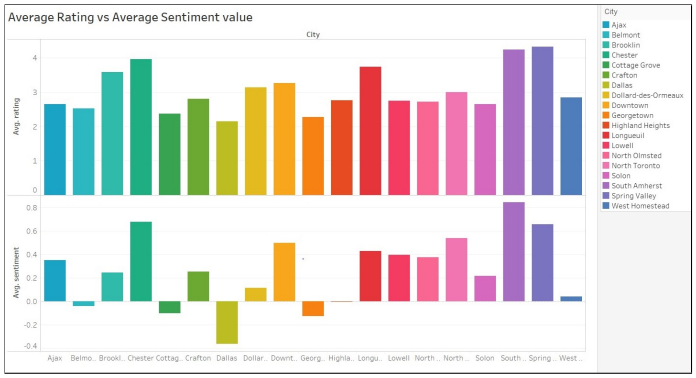

Fig. 4. Sentiment Score vs Average Rating

C. Task 3 : Top 25 Coolest Restaurants

In this task Top 25 coolest restaurants are calculated using a hive script. To answer this question, the addition of cool values of each restaurant is added to get the total coolness value of the restaurant. Filtering of restaurants from all available business is done first then the addition of the cool values. Top 25 coolest restaurants have been plotted using horizontal bars. The larger the bar size more popular the restaurant is, also the colour shown in the visualization fade away gradually as the coolness level decrease. The results as shown in figure 5 shows that the coolest restaurant is Shake Shack followed by Gen Korean BBQ house followed by others. The results of this task will help the consumers in deciding which places to go for dining as per the coolness level and answer the research question listed in I-B.

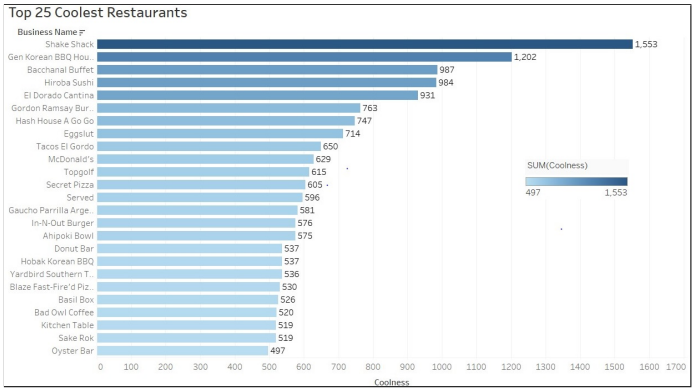

Fig. 5. Top 25 Coolest Restaurants

D. Task 4 : Top 10 Rated Business

Task four is to find the top 10 rated business. To answer this question a MapReduce code in Java has been written using a Top N design pattern. Here top ten rated business was calculated by first calculating the average rating of the business and then sorting it in descending order and finally filtering the results by top 10 records only. The output of the MapReduce job has been visualized in tableau using a treemap wherein the size of the box signifies the average rating of the business. The results as shown in figure 6 show that Abacoo's Steakhouse is the most rated business. The results of this question will aid consumers in finding trending business and will also help companies in keeping a track of the market and their competitors. The results of this task help in answering the research question listed in I-B.

E. Task 5: Average Rating of popular business categories in Las Vegas

In this task, the average rating of popular business categories in Las Vegas is calculated. To answer this question pig script is written which is executed in DataProc cluster. The output of this task has been visualized using packed bubbles in tableau. The size of the bubbles defines the average rating of the business category. The larger the value of average rating the

Fig. 6. Top 10 Rated Business

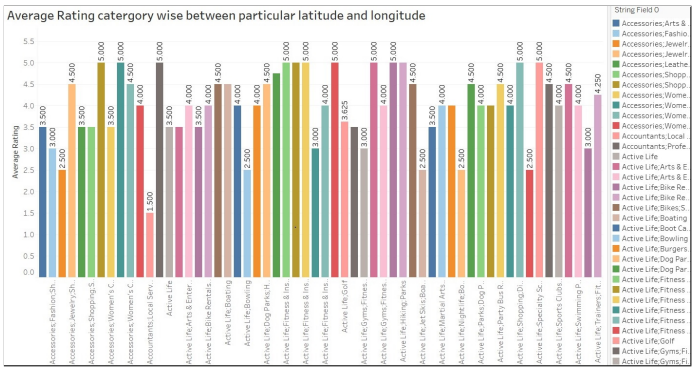

Fig. 8. Average Rating category wise between particular latitude and longitude

larger is the size of the bubbles in the visualization. To answer this question first business categories are grouped then the average rating is calculated of each category a finally sorted in descending order to find an average rating of popular business categories. The answer to this task will help consumers in choosing popular business categories as per the rating. The results of this question also answer the research question listed in I-B.

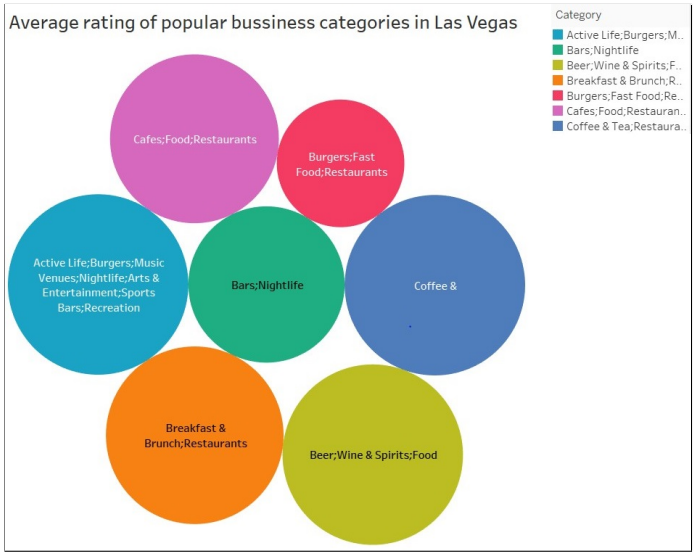

Fig. 7. Average Rating of popular business categories in Las Vegas

F. Task 6: Average Rating category wise between particular latitude and longitude

Task 6 in this project is to find the average rating category wise between following coordinates (43.22145313, -89.21487592) and (42.93172719, -89.61009908). These coordinates are in latitude and longitude format. A pig query has been written to find the answer to this question. The results of this query are plotted using a bar chart in tableau, where each business category is assigned different colours. The answer to this task will help consumers to find the best business category in their area. The results of this question also answer the research question listed in I-B.

G. Task 7: Top 25 Businesses with Most Views

The last task in this project is to find the top 25 businesses with most views. To answer this question a hive query has been written. The output of the result extracted form BigQuery directly into tableau has been visualized using horizontal bars. The larger the bar size more views the restaurant has also, the colour shown in the visualization fade away gradually as the number of views decreases. The answer to this query will help in finding trending businesses which is helpful for both the consumers, as well as companies and the results of the task, also helps in answering the research question listed in I-B.

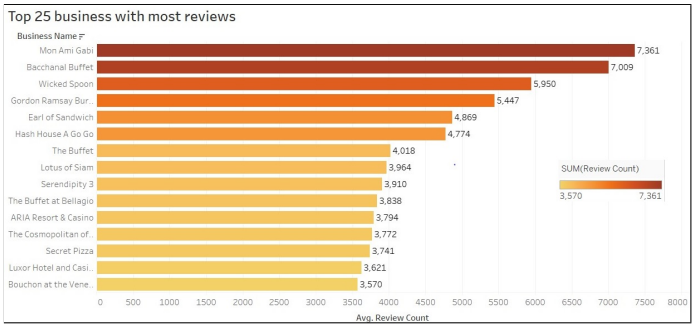

Fig. 9. Top 25 Businesses with Most Views

V. Conclusion

Yelp dataset was explored in this project to get insights and meaningful information reading business and restaurants across different cities in the world using BigData technologies. Four datasets containing different information was used in this project. The project was implemented on Google's Cloud Platform which provides services for analysis such types of BigData problems at a cost-efficient price and high performance which helps in focusing more on analysis and less on infrastructure and setup. Four different BigData technologies namely MapReduce, PySpark, Pig, and Hive were used to perform seven different tasks so as to answer the research question. The results of these task after processing in respective BigData technologies were visualized in tableau to get a better understanding of the outputs. The results of this project

will help consumers in decision making as well as companies in managing their business more effectively.

VI. Future Work

As due to time constraint, few features/attributes of the Yelp dataset were only explored, in future a thorough analysis on the dataset can be done. Also, machine learning methodology can be used in future to predict the fate of business which will be instrumental for the company's management team to stay ahead of their competitors. Real-time data could also be taken instead of historical data to get more accurate results. The analysis can also be extended by considering data from multiple sources to try and get better understanding and results.

References

[1] W. Dai, G. Jin, J. Lee, and M. Luca, "Optimal aggregation of consumer ratings: An application to yelp.com," *SSRN Electronic Journal*, 11 2012.

[2] W. Fan and M. D. Gordon, "The power of social media analytics," *Commun. ACM*, vol. 57, no. 6, pp. 74–81, Jun. 2014. [Online]. Available: http://doi.acm.org/10.1145/2602574

[3] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009. [Online]. Available: https://science.sciencemag.org/content/323/5915/721

[4] M. Schuckert, X. Liu, and R. Law, "Hospitality and tourism online reviews: Recent trends and future directions," *Journal of Travel & Tourism Marketing*, vol. 32, no. 5, pp. 608–621, 2015. [Online]. Available: https://doi.org/10.1080/10548408.2014.933154

[5] D. G. Pope, "Reacting to rankings: Evidence from america's best hospitals," *Journal of Health Economics*, vol. 28, no. 6, pp. 1154 – 1165, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167629609000873

[6] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research*, vol. 43, no. 3, pp. 345–354, 2006. [Online]. Available: https://doi.org/10.1509/jmkr.43.3.345

[7] M. Luca, "Reviews, reputation, and revenue: The case of yelp.com," *SSRN Electronic Journal*, 09 2011.

[8] danah boyd and K. Crawford, "Critical questions for big data," *Information, Communication & Society*, vol. 15, no. 5, pp. 662–679, 2012. [Online]. Available: https://doi.org/10.1080/1369118X.2012.678878

[9] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution that Will Transform how We Live, Work, and Think*, ser. An Eamon Dolan book. Houghton Mifflin Harcourt, 2013. [Online]. Available: https://books.google.ie/books?id=uy4lh-WEhhIC

[10] Z. Zhang, Z. Zhang, and Y. Yang, "The power of expert identity: How website-recognized expert reviews influence travelers' online rating behavior," *Tourism Management*, vol. 55, pp. 15 – 24, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0261517716300048

[11] J. Nandimath, E. Banerjee, A. Patil, P. Kakade, S. Vaidya, and D. Chaturvedi, "Big data analysis using apache hadoop," in *2013 IEEE 14th International Conference on Information Reuse Integration (IRI)*, Aug 2013, pp. 700–703.

[12] K. Yatani, M. Novati, A. Trusty, and K. N. Truong, "Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs," 05 2011, pp. 1541–1550.

[13] J. Huang, O. Etzioni, L. Zettlemoyer, K. Clark, and C. Lee, "Revminer: An extractive interface for navigating reviews on a smartphone," 10 2012, pp. 3–12.

[14] M. Potamias, "The warm-start bias of yelp ratings," 02 2012.

[15] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1411–1428, Oct 2006.

[16] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 12, pp. 1–135, 2008. [Online]. Available: http://dx.doi.org/10.1561/1500000011

[17] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 339–346. [Online]. Available: https://doi.org/10.3115/1220575.1220618

[18] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*, vol. 775152, 10 2003.

[19] S. Eob Lee, D. Kwan Son, and S. Han, "Qtag: tagging as a means of rating, opinion-expressing, sharing and visualizing." 01 2007, pp. 189–195.

[20] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," *Computing Research Repository - CORR*, pp. 417–424, 12 2002.

[21] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," *EMNLP*, vol. 10, 06 2002.

[22] B. Liu and L. Zhang, *A Survey of Opinion Mining and Sentiment Analysis*. Boston, MA: Springer US, 2012, pp. 415–463. [Online]. Available: https://doi.org/10.1007/978-1-4614-3223-4_13

[23] Y. Ainur, Y. Yisong, and C. Claire, "Multi-level structured models for document-level sentiment classification. in," *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1046–1056, 01 2010.

[24] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation (formerly Computers and the Humanities)*, vol. 39, pp. 164–210, 05 2005.

[25] T. Nakagawa, K. Inui, and S. Kurohashi, "Dependency tree-based sentiment classification using crfs with hidden variables." 01 2010, pp. 786–794.

[26] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," 10 2005.

[27] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai, "Automatic construction of a context-aware sentiment lexicon: An optimization approach," 01 2011, pp. 347–356.

[28] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08. New York, NY, USA: ACM, 2008, pp. 231–240. [Online]. Available: http://doi.acm.org/10.1145/1341531.1341561

[29] "Yelp Dataset." [Online]. Available: https://www.kaggle.com/yelp-dataset/yelp-dataset/