Yelp Data Analysis

Bhagyashree Bhagwat
Manali Joshi
Priyanka Purushu
GROUP E
CAL STATE LA.

Abstract: This paper aims at performing data analysis and providing insights on Yelp's Dataset using HIVE & PIG and presenting visualization with Tableau, Powerview & 3D Maps. From the available data we did best to find insights like sentiment analysis, best suited restaurants for tourists, forecast on number of potential users' to join Yelp in coming years , who are the most active users and who's reviews are more popular among the users and lastly the city having the most no of closed business's.
Data-set-url : https://www.yelp.com/dataset_challenge
Data Size : 2.62 GB

1. Introduction

The analysis is done on Yelp Dataset. This set includes information about local businesses in 10 cities across 4 countries dated from January 1, 2004 to July, 2016.
1) U.K.: Edinburgh
2) Germany: Karlsruhe
3) Canada: Montreal and Waterloo
4) U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison.

The dataset includes five datafiles:
1) Business
2) Review
3) User
4) Check-in
5) Tips.
The analysis is primarily done on Business, Review and User data and is focused only on the six U.S cities.

2. System Specifications

To began work on this dataset we used different services offered on IBM Bluemix cloud computing paltform. We extensively worked on Hadoop system components like Hive and Pig.
Our Bluemix cluster details are as follows:

- Cluster Type - Hadoop IBM Big Insights
- No. of Data Nodes – 1 node | vCPU = 4(24 GB RAM)
- No. of Management Nodes – 1 node | vCPU = 12 (48 GB RAM)
- Data Disk - 1 TB SATA | Data Storage- 244 GB
- CPU Speed - 2.30 GHz
- Version - IOP 4.2 [IBM Open Version Platform]
- Operating System - CentOS 6.6 [Linux]

3. Tools Used

We used Hive and Pig to extract the dataset files from Json to csv format. Also, we ran the queries in Hive and Pig. Secondly, for the visualization of the analysis output, we used Tableau software and tools such as 3D Maps & Powerview on Microsoft Excel.

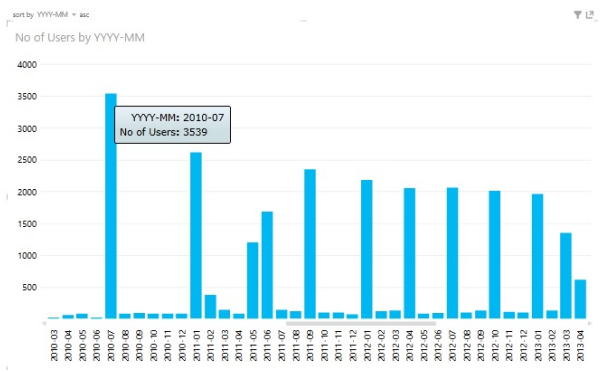
4. Workflow

The following is the step to step workflow carried out for this analysis.
Step 1 - Extract Dataset from Json to CSV using Pig & Hive
Step 2 - Data cleaning in Pig
Step 3 - Upload Dataset to HDFS
Step 4 - Analyzing data and building queries
Step 5 - Run commands in Hive and Pig
Step 6 - Visualize results using Tableau, Excel Power View & 3D Maps.
Step 7 - Summarize the desired output.

5. Data Analysis

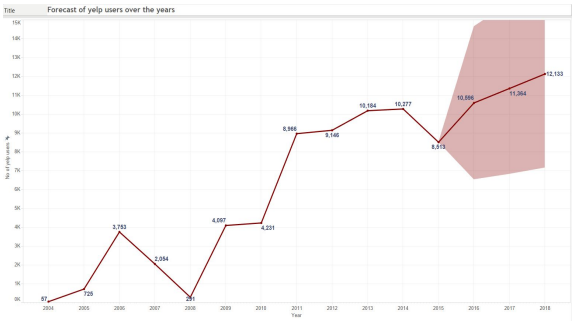
Post data extraction and cleaning, we ran queries to get the following insights.

5.1 In which year Yelp has got a maximum number of users?



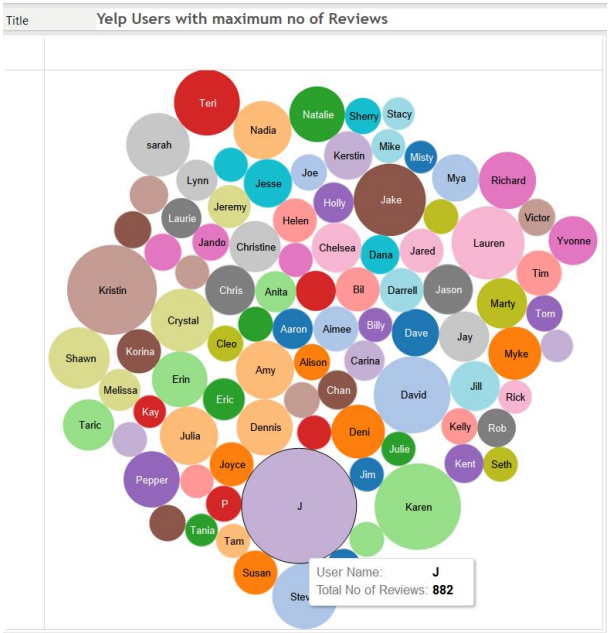
With the help of stacked columns, we have visualized in which year & month yelp got the maximum number of users. Additionally, in which year the maximum number of user had joined yelp. From the analyzed & visualized data we can see that Yelp got the max no of users in the month of July 2010.

5.2 Forecast of users joining Yelp in coming years.



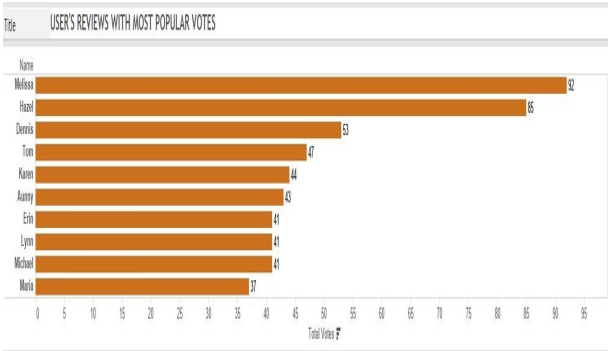
On the basis of previous insight, we have forecasted how many new users will be joining yelp in upcoming years. By the end of 2018, 12133 new users will be joining yelp.

5.3 The most active Yelp users.



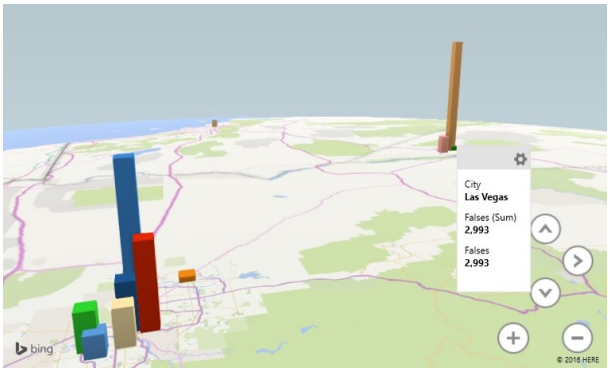
From the bubble chart, we have found who are the most active users. Most active users have been found on the basis of the number of reviews they've written. A user named J has written maximum 882 reviews.

5.4 Most popular users' review based on votes.



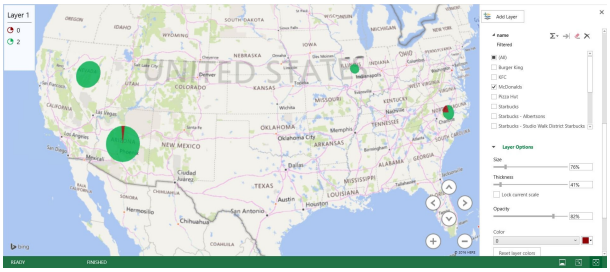
Top 10 users whose reviews got more popular on Yelp. User Melissa's review is the most popular with 92 votes.

5.5 Which city has the maximum no of closed business?



On the 3D Map, we have visualised that Las Vegas city has the maximum 2993 number of closed businesses. We got a useful insight from this analysis that, whoever wants to open the new business can avoid Las Vegas as it has shown the maximum number of closed businesses.

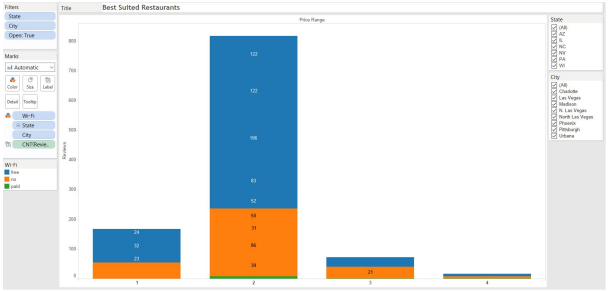
5.6 Sentiment Analysis on Top 8 food chains in the USA based on user reviews.



We have found which are the Top 8 food chains in the USA and we have done sentiment analysis based on

them. Sentiment analysis provides negative, neutral & positive reviews by the user for those food chains on 3D Maps.

5.7 Best suited restaurants for tourists.



Based on facilities like City, State, Parking, Wi-Fi, Price Range & User Reviews. On tableau screenshot, we can see that restaurants which fall under Price Range $$ (2) are the most suitable for tourists.

6. Summary of Insights

- During our analysis we learned that, Yelp got the max no of users in the month of July 2010.
- Further, we displayed a forecast of new Yelp users joining in the coming years .
- We displayed Top 100 most active users in the Bubble chart.
- Top 10 user reviews which got more popular on Yelp.
- On the 3D Map, we visualised that Las Vegas city has the max no of closed businesses.
- Sentiment analysis which includes positive, neutral & negative reviews about top 5
- restaurants based on geo location. We only displayed for McDonalds.
- List of Best suited restaurants for tourists in terms of Price range, City, State, Wi-fi
- Connection and Parking

7. Github URL

https://github.com/mjoshi1110/cali/

8. References

https://calstatela.edu/jwoo5/classes/2016/fall/cis5200/

https://www.yelp.com/dataset_challenge

https://console.ng.bluemix.net/catalog/services/biginsights-for-apache-hadoop

https://www.tableau.com