Yelp Restaurants Big Data Analysis

Maria Boldina, Meena Muthiah, Prasanna Natuthurai, Julia Stachurska | Dr. Jongwook Woo | California State University, Los Angeles

Problem/ Question

Since its inception in 2004, Yelp has collected a staggering amount of data about local businesses including over 142 million reviews from users.

The goal of this project is to take advantage of the Hadoop Big Data analytics tools to capture some valuable insights about the restaurant industry from the 2017 Yelp challenge data set.

Hypothesis

- Fast food is the most popular category on Yelp
- Bars usually receive favorable reviews
- Majority of Yelp users leave positive reviews about restaurants
- Elite users might have a bias in their ratings

Project Overview

- Our objective is to use Hadoop Big Data analysis tools in order to provide insightful analytics which can help restaurants owners make important decisions regarding CRM, marketing and to have a competitive edge.
- We started by downloading and extracting data from the 2017 Yelp challenge dataset. The extracted JSON files have been uploaded to IBM Bluemix BigInsights cluster (HDFS) with the help of AMBARI file browser. Since JSON files in the dataset contain numerous nested attributes we chose to use Rcongiu JSON SerDe. This library enables Apache Hive to read and write in JSON format. Tables necessary for our analysis were created using HiveQL.
- Finally, we used Tableau software for visualization of the results.

Variables / Research

Data set URL: https://www.yelp.com/dataset

Data set size:

- 1 TAR file - 2.28 GB compressed
- 6 JSON files - 5.79 GB uncompressed
- Categories: business, reviews, user, check-in, tip and photos.

This set includes information about local businesses across 4 countries (USA, Canada, UK, Germany).

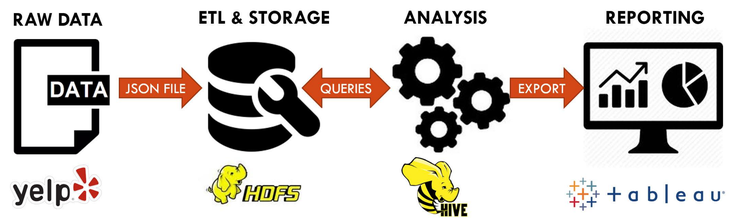
- 4.7M reviews
- 1.1M users
- 156K businesses
- 12 metropolitan areas

Technical Specifications

BigInsights for Apache Hadoop

- IBM Bluemix BigInsights cluster
- 1 Data Node | vCPU = 4 (24 GB RAM)
- 1 Management Node | vCPU = 12 (48 GB RAM)
- Data Disk - 1 TB SATA
- CPU Speed - 2.30 GHz
- Version - IOP 4.2
- Operating System - CentOS 6.6
- Data Centre - Washington DC

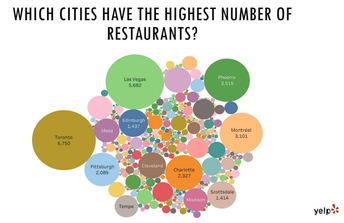
Procedure

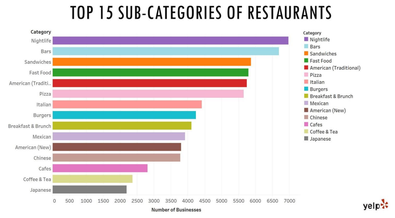


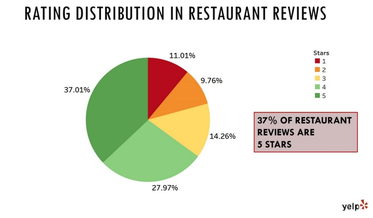
Observations

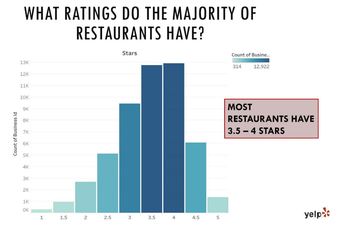
- The results of the analysis have shown that not all our assumptions in the hypothesis section were correct
- Most popular categories of restaurants on Yelp are nightlife, bars, sandwiches and fast food
- Majority of restaurants on Yelp have ratings b/w 3.5 and 4 stars
- 37% of restaurant reviews are 5 stars
- Almost 28% of bad (1-2 star) reviews are for fast food restaurants
- Restaurants with the most reviews are located in Las Vegas
- Only 3% of Yelp users are Elite members (30K out of 1.1M)
- Elite and non-elite users give restaurants very similar star ratings

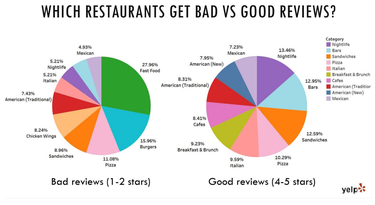
Visualizations

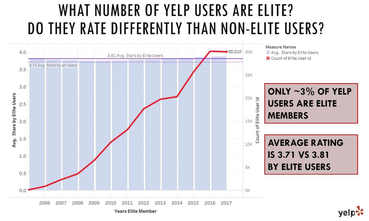












Conclusion

In this paper, we have successfully used Hadoop Big Data analytics tools to capture some valuable insights about the restaurant industry from the 2017 Yelp challenge data set.

From our analysis we had a better understanding of various aspects of the restaurants on Yelp such as which cities have the highest number of restaurants, top 15 sub-categories of restaurants, distribution of ratings in restaurant reviews, the number of Elite users on Yelp and how their ratings compare to non-elite users.

Works Cited

- Data set URL https://www.yelp.com/dataset
- Serde source https://github.com/rcongiu/Hive-JSON-Serde
- IBM Bluemix: https://console.bluemix.net/data/bic/
- Tableau https://www.tableau.com
- JSON with Tableau https://onlinehelp.tableau.com/current/pro/desktop/en-us/examples_json.html