Big Data Analysis of Local Business and Reviews

Ruchi Singh
California State University, LA
+14696185151
rsingh26@calstatela.edu

Yashaswi Ananth
California State University, LA
+19176404175
yananth@calstatela.edu

Dr. Jongwook Woo
California State University, LA
+13233432916
jwoo5@calstatela.edu

ABSTRACT
In this paper, we have analyzed the local business data and reviews to get insights on the popularity of a business and factors responsible for it. We have also analyzed the sentiments of the customer reviews for better understanding of business popularity. The total size of the dataset is 180 MB and we have adopted cloud computing service like Big Data Hadoop using HiveQL and Pig for query. Our preferred choice of framework for this project was Big Data Hadoop primarily because it is an open source software and its scalability and flexibility best suited our requirements.

This project has helped us in understanding various aspects of the Local Businesses; factors driving their popularity, customer review patterns and regions that favor certain businesses the most, are a few of the many aspects to name. Analyzing the customer sentiments based on their reviews has helped us in realizing the importance of customer satisfaction, also it is possible to derive action plans to improve business performance to keep up with the competition.

CCS Concepts
• CCS → Information systems → Information retrieval → Retrieval tasks and goals → Information extraction
• CCS → Information systems → Information retrieval → Retrieval tasks and goals → Sentiment analysis
• CCS → Information systems → Information retrieval → Retrieval tasks and goals → Business intelligence

Keywords
Local Business; Data Analysis; Hadoop; Big Data; Reviews; Visualization; IBM Bluemix; Azure; Hive.

1. INTRODUCTION
Yelp and Google Local engages a large percentage of customers for reviews about local business. It can be great source of information for business owners to improve their services and users to choose best services available. More than 3 Exabyte of data is created every day on the internet, making it a challenge to store this huge amount of data and retire useful information from it.

Our objective is to provide insightful analytics on existing business registered on these platforms, help future business owners to make important decisions regarding services or business expansion and find business trends.

We have analyzed a data set from Yelp and Google Local that spans a variety of businesses such as restaurants, shopping, nightlife, medical, education, entertainment, common services, etc. in various cities of the world, using Big Data, a non-expensive framework that can store a huge variety of data set and process it in parallel [1][2]. We have worked on Hadoop ecosystem to perform the Big data analytics and have built the project on IBMs Biginsight as well as Azure HDinsight.

Biginsight is an open source service provided by IBM Open Platform, and includes all the Hadoop ecosystem components integrated with Apache Spark for a good processing power. We also built the project on Microsoft's ready to roll-out service called "HDInsight" using Hortonworks Hadoop on the Azure platform. The Hive ODBC driver from Microsoft and Hortonworks provides more options to Excel and PowerPivot users to query data within Hadoop running the cloud computing service.

The technical specifications for the platforms used for the project are as follows:

<table>
<tr><td colspan="2">Azure Specifications</td><td colspan="3">IBM Bluemix Specifications</td></tr>
<tr><td>Worker Node</td><td>4</td><td rowspan="2">Management Nodes: 1</td><td>vCPU</td><td>12</td></tr>
<tr><td>Header Node</td><td>4</td><td>RAM</td><td>48 GB</td></tr>
<tr><td>Number of Cores</td><td>24</td><td rowspan="3">Data Nodes: 1</td><td>vCPU</td><td>4</td></tr>
<tr><td>RAM</td><td>14 GB</td><td>RAM</td><td>24 GB</td></tr>
<tr><td>Disk Size</td><td>200 GB</td><td>Data Disk</td><td>1 TB SATA</td></tr>
<tr><td>Operation System</td><td>LINUX</td><td colspan="2">CPU Speed</td><td>2.4 GHz Intel Xeon ES-2673</td></tr>
</table>

Table 1: Technical Specifications

HiveQL and Pig are the querying tools built on top of Hadoop that are used to query data within HDFS. They inherit all of Hadoop's fault tolerance features and are scalable for Big Data. The Hive language resembles SQL, which makes it useful for creating reports by Data Analysts whereas Pig is a Procedural Data Flow language used for programming by researchers and programmers.

Visualizations for this project are generated in Tableau and Excel power view that create multi-faceted views of data and help communicate complex ideas simply.

2. RELATED WORK
Published work related to local business analysis is focused on specific categories of business in general. Some of them have performed a detail analysis of business of a location, city or country unlike this paper which studies countries like US and UK.

For instance, Gwo-Hshiung Tzeng [3] concentrates on the criteria for a good restaurant location in Taipei. Whereas Tsung-Yu Chou [4] evaluates the importance of infrastructure cost and environmental factors responsible for setting up a hotel business. Predicting Usefulness of Yelp Reviews by Xinyue et al. uses MATLAB to perform language processing techniques.

Our work has a more holistic and complete approach to the data analysis taking in account every aspect of local business. The goal of this paper is to provide all the possible insights from the data set, like defining a method that analyzes the factors of success of all kinds of businesses, instead of focusing on a single category of business, finding business trends, customer response, popularity of every business etc. Furthermore, using Hadoop Big Data and Cloud Computing services to store and process massive business data set.

3. DATA SPECIFICATIONS

The Local Business dataset collected from Yelp and Google APIs dated between 2005 and 2016, is rich in information about the local businesses in various cities in 14 different states in U.S and 4 other cities that include: U.K: Edinburgh, Germany: Karlsruhe, Canada: Montreal and Waterloo. The yelp data set is of size 90MB in CSV file format with 334,335 rows and 108 columns. The Google Local reviews dataset is of size 85MB in JASON file format with 117,486 and 10 columns. To be able to use this data in Hadoop, the JSON file was converted to a CSV file using SerDe (Serializer/Deserializer) in Hive. Alternatively, JSON to CSV file format conversion can also be done in PIG using the built-in function JSONLoader().

The business data contained a lot of junk data which made it difficult to import into Hive tables for analysis. To clean the data, we removed the duplicate rows, eliminated complete NULL valued columns. We also had to format the date column to yyyy-mm-dd time columns to timestamps.

4. ANALYSIS OF DATASET

The clean data was imported into HDFS, exploratory data analysis was performed on the dataset using LOAD and DUMP in Pig to gather some initial facts about our data. For example, the local business data included 2,903,217 entries having attributes such as business id, address, name, longitude, latitude, etc. The reviews data included 1,174,850 hive entries having attributes that included business id, user id, review id, stars, review date, and the user comments. Separate tables were created for business and reviews in HDFS using Hive QL. Reviews table creation is as follows:

"CREATE EXTERNAL TABLE IF NOT EXISTS reviews(funny int, useful int, cool int, userid string, reviewid string, stars int, reviewday date, type string, businessid string, comment string)ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/rsingh26/LocalBussinessReviews/' TBLPROPERTIES ('skip.header.line.count'='1');"

A master table was created by joining the business and reviews table on common business id in each table. There are 452 categories of business, we have grouped them broadly into 6 categories listed below. For each category, separate tables were created to further analyze the business grouped in them: - Education, Entertainment, Food, Medical, Services and Shopping.

4.1 INFORMATION EXTRACTION

The data has been analyzed in the form of answers to the following questions related to business data.

4.1.1 What is the review count for each category of business to understand which business category gets most number of reviews by the customers?

The Hive QL to get the sum of the count for each category is as follows and its visualization is in Figure 1.

```
select sum(review_count) from Education;
select sum(review_count) from Entertainment;
select sum(review_count) from Food;
select sum(review_count) from Medical;
select sum(review_count) from Services;
select sum(review_count) from Shopping;
```

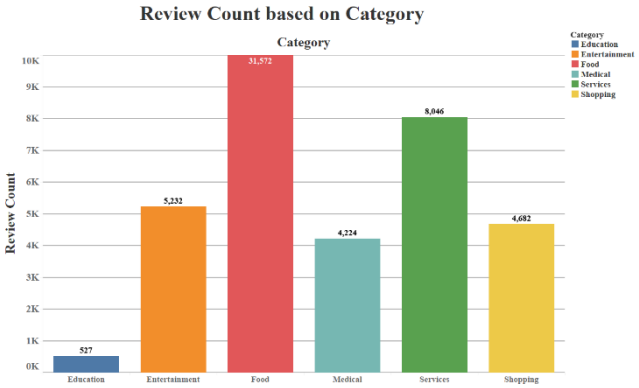

Figure 1. Review count category-wise

Inference: Users tend to review the food businesses more compared to the other categories of business like education, medical, shopping and services. Although categories like Education and Medical are equally vital categories as Food for day to day life. This insight can be used to drive medical and educational organizations to encourage customers to rate the services in-turn having an impact on business and service quality.

4.1.2 Which sub categories are most popular in different states in US?

The business data was collected from 14 states in US as shown in Figure 2. and Figure 3. It has been visualized using Excel power-view.

State	Category
Alabama	Food
Arizona	Home Services
Arkansas	Home Services
California	Entertainment
Illinois	Food
Minnesota	Local Services
Nevada	Food
New Mexico	Steak Houses
North Carolina	Active Life
Oregon	Hotels and Travel
Pennsylvania	Food
South Carolina	Food
Texas	Banks and Credit Union
Wisconsin	Food

Figure 2. Business Categories and States

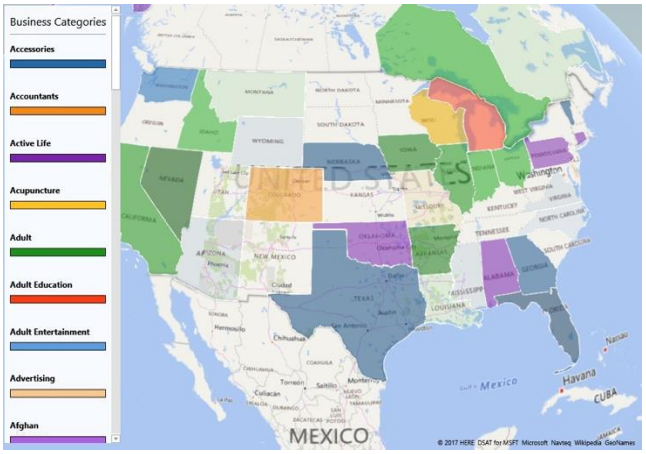

Figure 3. Sub categories grouped by state

4.1.3 Which is the most popular city for each business category?

The Tree map representation in Tableau is used to visualize the business categories grouped by city. The size of the rectangle represents the review count for a city and every business category is represented with a different color, refer Figure 4.

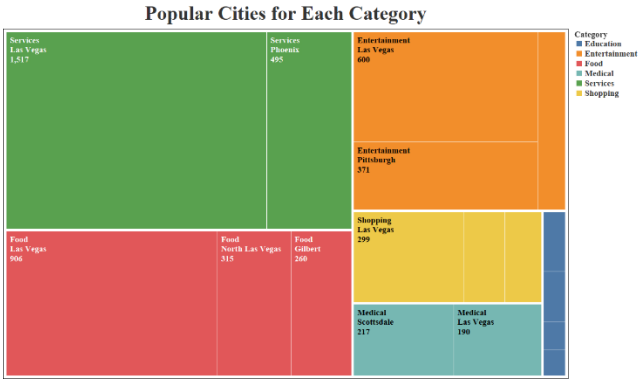

Figure 4. Business categories grouped by city

Inference: In every category of business maximum number of review count comes from Las Vegas. The possible reason for this could be the large number of tourists that visit this city every year, use the services and leave feedbacks. As another possibility, the business in this city encourage customers to review their business to improve the customer satisfaction and increase popularity.

4.1.4 What is the count of reviews for the sub-categories of Shopping?

Bubble chart in Tableau is used to visualize the shopping sub-categories grouped by city in Figure 5. The distinct color of the bubble represents different sub-category and size of the bubble represents the review count. Bubbles with city names are the most prominent sub category of Shopping in the respective city and the number below the city name is the review count.

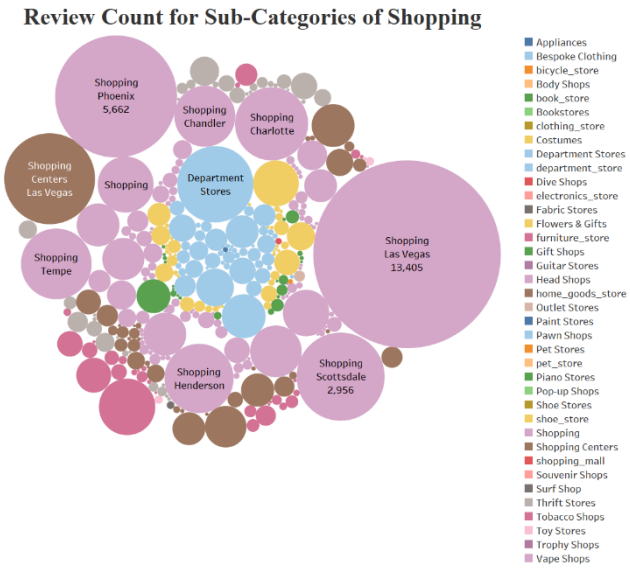

Figure 5. Review count for the sub categories of Shopping

4.1.5 What is the sentiment analysis the customer reviews for all the Local Businesses in a city?

Customer's sentiment is analyzed by the sum of the positive, negative and neutrals words, used in the review, in accordance to the English dictionary. Cumulative sentiment is calculated for every review and summed all the business in a city and represented in pie chart for each city. The size of the pie graph is the number of people who voted the review useful.

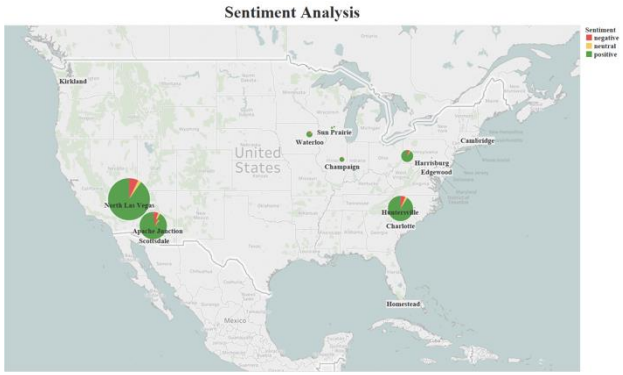

Figure 6. Sentiment Analysis based on customer reviews

Inference: More than 60% people in a city write positive reviews for the business and services they use. Las Vegas has the highest number of useful review votes.

4.1.6 What is the percentage of positive, neutral and negative reviews by the customers for the sub category of Local Businesses?

Sentiment analysis for the sub categories of Services is done in Figure 7. Similar analysis can be done for Education, Entertainment, Food, Medical and Shopping as well.

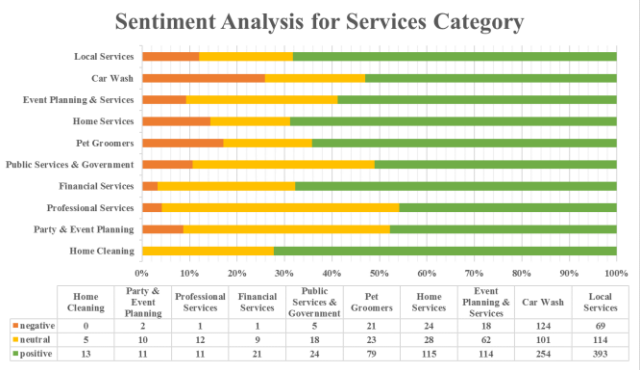

Figure 7. Percentage of positive, neutral and negative customer reviews for the Services category

Inference: Services businesses mostly received positive reviews from the customers and the Home cleaning services did not receive any negative feedback.

4.1.7 What is the maximum number of business reviewed by individual users?

The following Hive QL query was used to find the maximum number of business reviewed by a single reviewer and Figure. 8 visualizes it in Excel charts.

```
select userid, count(businessid) from review_analysis group by userid limit 10;
```

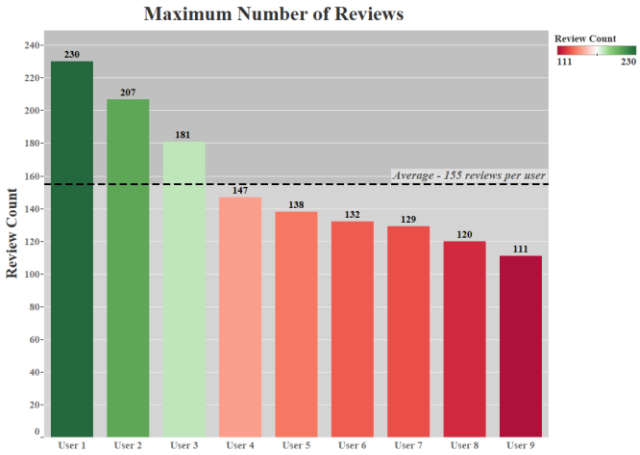

Figure 8. Maximum count of reviews made by individual users

Inference: One of the user has written reviews for almost 250 different business over a span of 10 years. Further text analysis of the reviews could help us in understanding the authenticity of these reviews.

4.1.8 What are the factors responsible for the popularity of the Local Businesses?

The top and bottom rated Food business are listed below in Table. 1 and Table.2 respectively. Various attributes like happy hour, reservation, parking, vegan etc. were compared and analyzed. It led to the conclusion that the attributes listed in Table. 3 are hugely responsible for the good reviews and ratings of a food business. Similar analysis can be done with other categories of business as well.

Top Rated Food Businesses			
City	Name	Reviews	Stars
Las Vegas	Art of Flavours	359	5
Las Vegas	Poke Express	315	5
Las Vegas	Brew Tea Bar	306	5
Gilbert	Frost Gelato	260	5
Las Vegas	Dutch Bros. Coffee	241	5
Phoenix	Handcrafted American	232	5
Las Vegas	Karaoke Bar	192	5
Montreal	Kem CoBa	156	5
Mesa	Gelato Dolce Vita	149	5
Las Vegas	Tast Crepes	148	5

Table. 2 Top rated food businesses

Bottom Rated Food Businesses			
City	Name	Reviews	Stars
Las Vegas	KFC	19	1
Las Vegas	MCDonald's	17	1
Charlotte	Pizza Hut	15	1
Chandler	Dairy Queen	14	1
Maricopa	Dairy Queen	14	1
Scottsdale	Food Truck Festival	12	1
Pittsburgh	Pizza Hut	11	1
Carnegie	Walmart	10	1
Queens	Burger King	9	1
Surprise	Church's Chicken	9	1

Table. 3 Bottom rated food businesses

Reservation		Ambience		Wheelchair		TV		WiFi	
Top	Low	Top	Low	Top	Low	Top	Low	Top	Low
No	No	Yes	No	Yes	No	Yes	No	Free	No
Yes	No	No	Yes	Yes	No	No	No	Free	No
Yes	No	Yes	No	No	No	No	No	Free	No
No	No	No	No	Yes	No	No	No	No	No
Yes	No	Yes	No	No	No	Yes	Yes	Free	No
Yes	No	No	Yes	Yes	No	Yes	No	Free	No
Yes	Yes	Yes	Yes	Yes	No	Yes	No	Free	Free
Yes	No	Yes	No	Yes	Yes	Yes	Yes	No	Free
Yes	Yes	Yes	Yes	No	No	Yes	Yes	Free	No
Yes	No	Yes	No	Yes	No	No	No	Free	Free
80%	**20%**	**70%**	**40%**	**70%**	**10%**	**60%**	**30%**	**80%**	**30%**

Table. 4 Factors influencing the popularity of the food businesses

Inference: More than 70% of the top 10 food business have reservation, ambience, wheelchair facility, TV and free Wi-Fi whereas less than 30% of bottom 10 food business have these facilities.

5. CONCLUSION

Using Hadoop, Hive, Pig, and Tableau enhanced the exploratory possibilities and analytical capability, to store and process Big Data in parallel computing. The findings from the research can be concluded as: food is the most popular category of Local Business based on the review count. Las Vegas is the most popular city for local business in every category of business. Reservation, ambience, Wi-Fi are the main factors responsible for the popularity of food businesses. More than 60% of people in a city write positive reviews for local business. And on record the maximum number of business reviewed by a single reviewer is 250.

6. ACKNOWLEDGMENTS

We are grateful to Mahsa Tayer Farahani and Hemamalini Madhanguru for their help and support in this project.

7. REFERENCES

[1] Woo, Jongwook., and Xu, Yuhang., 2011. "*Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing*" The 2011 international Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2011), Las Vegas (July 18-21, 2011).

[2] Woo, Jongwook., 2013. "*Market Basket Analysis Algorithms with MapReduce*" DMKD-00150, Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, Oct 28 2013, Volume 3, Issue 6, pp. 445-452, ISSN 1942-4795.

[3] Chou, T. Y., Hsu, C. L., and Chen, M. C., 2008. "*A Fuzzy Multi-Criteria Decision Model for International Tourist Hotels Location Selection*" International Journal of Hospitality Management, Vol. 27, No. 2, pp. 293-301, 2008.

[4] Tzeng, G. H., Teng, M. H., Chen, J. J., and Opricovic, S., 2002. "*Multicriteria selection for a restaurant location in Taipei*" International Journal of Hospitality Management, Vol. 21, No. 2, pp. 171-187.