Abstract—This paper explores the advantages of utilizing the Hadoop ecosystem for performing Big Data analytics. One of the main advantages Hadoop has over traditional technologies includes high speed replication and availability. Data within the Hadoop ecosystem is replicated to maximize fault tolerance in the system. The Hadoop ecosystem includes the Hadoop File System (HDFS) and MapReduce. HDFS is a distributed file system that provides high throughput access to application data. MapReduce, on the other hand, is a programming model that allows for parallel and distributed processing. Using a distributed storage approach also allows for data to be scaled-out horizontally, which is more cost effective than traditional storage designs which scale vertically.

To demonstrate how Hadoop can be used to analyze Big Data, Hive, PIG, and Tableau will be used to perform initial exploratory data analysis. The dataset chosen for this analysis comes from the popular web site (http://www.yelp.com/dataset_challenge) called Yelp. Yelp specializes in reviews of various businesses such as restaurants, shops, and services. A few examples of services include dentists, hair stylists, and mechanics. The data set that is currently available to the public is part of a yearly challenge that is open to the public and made for research purposes. The latest data includes information about local business in 10 cities across 4 countries. The cities include:
- U.K; Edinburgh
- Germany: Karlsruhe
- Canada: Montreal and Waterloo
- U.S: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, and Madison.

I. INTRODUCTION

Websites of consumer reviews are on the rise. Popular food sites like Chowhound.com and Phood.com are some examples of the many places on the internet where you can find reviews about food. For travel advice, you can consult tripadvisor.com or fodors.com. What makes Yelp unique from the other websites, however, is that the reviews on their site are written exclusively by the user community. There are no professional reviews. The reviews also span a variety of businesses such as restaurants, shopping, nightlife, entertainment, and common services. Because Yelp is an aggregate of user reviews on a variety of businesses, it contains valuable data on consumer and business trends. Yelp is also available through a smartphone application. Therefore, real-time data is recorded whenever a user “checks in” or writes a review or quick tip.

The goal of this paper is to use the Hadoop ecosystem to help glean insights on the latest Yelp dataset. We intend to gather information about the most popular businesses in a specific town and perhaps learn more about certain demographic trends.

II. DATA PREPARATION

The original dataset provided is a single Tape Archive (TAR) file, 588MBs in size. Within the tar file are five JavaScript Object Notation (JSON) files representing five different categories. The categories include: Business, Check-in, Reviews, Tip, and User Information. Due to the magnitude of data, only the Business and Check-In data was analyzed for this project.

To prepare the data for use in Hadoop the JSON files were extracted from the TAR file and converted to a CSV format. The JSON to CSV conversion was performed using a python script available through github.com. [1] The CSV files were then converted to a tab delimited text file before it was imported into the Hadoop Distributed File System (HDFS).

The Business data contained a lot of noise, which made it difficult to import into Hadoop. To clean the data we took the liberty of removing the columns which we determined had little amount of useful data. For example, attribute columns with less than 5% of data were removed from the data set. Carriage returns from the “full address” column were also removed. All clean-up tasks were performed in Excel.

III. EXPLORATORY DATA ANALYSIS

After the data was imported into HDFS, exploratory data analysis was performed on the dataset using Hive. Hive is a querying tool built on top of Hadoop that is used to query files within HDFS. Hive inherits all of Hadoop’s fault tolerance features and is scalable for Big Data. The Hive language also resembles SQL, which makes it an efficient tool for ad hoc queries.

Using Hive we can gather some initial facts about our data. For example, the Business data table includes 61,184 entries and 71 columns of attribute data such as name, longitude, latitude, etc. The Tips data table includes 505,858 hive entries and 6 attributes. A snapshot of this output is shown in Figures 1-4 below.
To look for trends, we ask the following questions about the data.

1) Which states are the most active on Yelp?

Active is defined as having the most number of reviews. It is apparent from the query below that Arizona, followed by Nevada, has the most number of reviews.

```
select state, count(*) from yelpdata group by state order by count(*) desc
```

The top 3 reviewed states are AZ, NV, and NC. We can see that the majority of reviews are for restaurants and a few airports.

2) What kinds of businesses receive the most reviews in each state?

The top 3 reviewed states are AZ, NV, and NC. We can see that the majority of reviews are for restaurants and a few airports.
3) What kinds of businesses receive the most reviews in each state?

Most tips come from restaurants such as Starbucks and McDonalds. However there are some services like Bank of America, Firestone Complete Auto Care, and Banfield Pet Hospital which receive a disproportionately high amount of tips. The reason for high tip counts may warrant further investigation.

PIG is a procedural data-flow language within the Hadoop ecosystem that can also be used to retrieve information from the dataset. Since it is a procedural language it can be used for controlling more complicated queries, step-by-step. PIG scripts are ideal for running reports that could be performed periodically.

Since Arizona has been determined to have the most reviews out of all the states (Figure 5), we may want to ask the following questions with PIG:

1) Which states are the most active on Yelp?

An example of the code in PIG is shown below.

```pig
--10 Most Reviewed Restaurants in AZ
```
The 10 most reviewed restaurants in AZ include:

- **Pizzeria Bianco**, Phoenix – 1453 reviews
- **Four Peaks Brewing Co**, Tempe - 1241 reviews
- **Cibo**, Phoenix, 1202 reviews
- **FEZ**, Phoenix, 1117 reviews
- **Cornish Pasty Company**, Tempe, 1033 reviews
- **Postino Arcadia**, Phoenix, 995 reviews
- **Lux**, Phoenix, 956 reviews
- **Gallo Blanco**, Phoenix, 924 reviews
- **The Mission**, Scottsdale, 908 reviews
- **Citizen Public House**, Scottsdale, 886 reviews

2) *What times do most people dine at the most reviewed restaurant?*

The check-in data set was mined for the distribution of check-in times in PIG.

```pig
B = filter A by categories matches '.*Restaurants.*';
C = filter B by state == 'AZ';
D = foreach C generate name, city, review_count, business_id;
E = order D by review_count desc;
F = limit E 10;
dump F;
```
The original dataset for check-in times were provided as 168 separate attribute columns. Each column represents an hourly time slot for a 24 hour period, spanning across 7 days of the week. The outputs were a string of check-in counts. To help visualize the results, the outputs from PIG were imported into Tableau.

It is apparent from the Tableau output that Saturday is the busiest day of the week for this restaurant.

IV. CONCLUSION

Using Hadoop can greatly enhance our analytics capability with Big Data. In our analysis of Yelp data, Hive, PIG, and Tableau can greatly enhance exploratory data analysis. We discovered the majority of Yelp activity residing in the United States. Reviews in foreign cities in Canada (QC) and in the UK (EDH) were relatively low compared to the US, which could represent growing business areas for Yelp. Within the Yelp business and check-in data sets, we discovered Arizona and Nevada have the most activity on Yelp. We also drilled down to the most popular Pizza business in AZ and used check-ins to analyze its business activity.

V. FUTURE WORK

Although the business and check-in datasets had timestamps, the specific dates of the check-ins and reviews were not disclosed. Future studies could involve investigating the date information to help us answer questions regarding business growth within each city.

ACKNOWLEDGMENT

The analysis derived in this paper uses Open Source components. You can find the source code of his open source projects along with license information below. We acknowledge and are grateful to the developer for his contribution to open source.

References


BIOGRAPHY

Chih-Hao Dai received his B.S. degree in Applied Mathematics in 2001 from the University of California at Los Angeles. His background is in cost engineering and affordability. He is a senior systems engineer at Northrop Grumman and is currently pursuing his M.S. degree in Predictive Analytics at Northwestern University.

Sunil Kakade Sunil has over 18 years' experience in various technical and leadership roles focused on IT transformation initiatives, analytics driven software development and Big Data technologies. Sunil is a Big Data and Data Science evangelist with hands-on expertise in implementing Hadoop and NoSQL technologies. Sunil leads the architecture and delivery of business systems for a fortune 100 company. His areas of interests include Machine Learning, Data Science, Legacy systems modernization and Open Source technologies. Sunil has a Master of Science in Information Technology and Management from the Illinois Institute of Technology at Chicago and obtained a Master of Science in Mechanical Engineering from the Indian Institute of Technology, Madras, researching applications of Artificial Intelligence to manufacturing processes. His research papers have been presented at multiple international conferences and published in technical journals. He has patents pending in areas of IT and retail processes.