ENGINEERING BIG-DATA SYSTEMS

PROJECT REPORT
"Yelp Dataset Analysis"
April 2016

By:
Dishu Jindal
Pawan Matta
Rimple Talati

OVERVIEW

We have selected yelp academic data set. We have chosen Yelp data set because it is large in size, which allows us to do research, and analysis in details. In addition, it helps the travelers to explore the places and thus we want to help them by analyzing it in detail.

We have used many technologies/ programming languages like Virtualization, AWS Elastic Map Reduce, Machine Learning- Using Mahout Library, Pig, Hive, HBase, Python, Java, R Studio and Tableau

Virtualization is a software that separates physical infrastructures to create various dedicated resources. It is the fundamental technology that powers cloud computing. Virtualization software makes it possible to run multiple operating systems and multiple applications on the same server at the same time. It enables businesses to reduce IT costs while increasing the efficiency, utilization and flexibility of their existing computer hardware.

For our project, we have selected Elastic map reduce services of AWS. We have created a cluster, which includes one master and two slaves. While creating a cluster we have selected applications, which we need while working over on analysis like Hadoop, Pig, Hive, HBase and Mahout.

ANALYSIS

We have done following analysis for our project:

- Sentimental Analysis of user's reviews based on business
- Business Recommendations for Users
- Compare the morning, afternoon and evening check-in counts of users based on the business and cities

FIRST ANALYSIS - Sentimental Analysis

In this, we are going to analyze the reviews given by users for a particular business, which helps us to find top 20 and worst 20 business reviews.

We have used two Yelp's JSON files: - Business and Review

We have used Dictionary (tsv file) which consists of total 8000 words of positive and negative.

We have uploaded these JSON files to HBase tables using hive. For uploading data, we are storing data into staging table so that we can fetch required fields from these staging tables and upload into the HBase Tables. After uploading the data to HBase tables, we are going to retrieve these data using pig script so that we execute sentimental analysis script to compute top 20 and worst 20 business according to reviews given by the user. After getting the output in csv file, we get that file from Hadoop dfs, copy it on our local system, and send this csv file as input to R script for displaying the results.

SECOND ANALYSIS – Business Recommendations for users

In this, we are going to recommend businesses to users using their similarity percentage via mahout library.

We have used following JSON files: - Business, Review and Users

Initially, we have user id and business id as string which we can't use as input in mahout library so we created a mapping file for business id's and user id's. After creating business mapping and user mapping csv files, we uploaded these csv files and review JSON file to Hive staging tables. After executing these scripts, we will get the input file, which we can provide, to mahout library, which will have user id, business id and review rating. We will execute the mahout command to compute recommendations for all users. For displaying this output in Tableau, we wrote a java code, which will parse this file as input and convert the data into csv file, which we gave as input file to Tableau. In Tableau, we can see the recommended business for a particular user and see users recommended for a particular business.

THIRD ANALYSIS – Comparing Morning, Afternoon and Evening Check-in by users

In this, we are going to do analysis on the check-in times by users based on the business and cities.

We are using following JSON files for this analysis: - Business and Check-in

Before uploading check-in file to Hive, we cleaned the data using java code because this JSON file has nested json objects and dynamic keys. After cleaning the data, we uploaded these JSON files to staging tables of hive using Hive scripts. After executing these commands, we got all the required data into hive Tables and then we used join statements of hive scripts to get the desired outputs. After getting csv file as output, we pass that file as input to tableau and get the following output.

CONCLUSION

Thus, we conclude that Yelp dataset was used to perform mainly three analysis, which are sentimental analysis, check-in and recommendations to users. We found that user reviews create a lot of impact on the business. To understand user requirements and to provide them with recommendations was the goal which we achieved after implementing the project.