Zomato Restaurant Analysis Using Pig and Hadoop

Abstract

Food sector is one of the largest areas in industries across the globe. There is urge for carrying out analysis for most of the population across the globe who are always excited to try different cuisines in different parts of the country and with good quality of food. This research analysis aims to examine the worth for money restaurants across the various parts of the country based on various factors such as value for money, rating, quality index, frequency of visits and types of cuisines. The examine of restaurant industry is done on the urge of the populations who are motivated to try the best cuisines across the different part of the globe and which city serves the best foods in the market with large number of restaurants. For the analysis of food industry, Zomato API dataset is used which is publicly available on Kaggle.

Keywords: Hadoop, MapReduce, PIG, Hive, HDFS, HBase, MySQL, SQOOP, MS Excel.

Introduction

The restaurant food industry is one of the universal and biggest sectors across the globe. This industry is reaching to every household in one way or another with an extensive variety of products and services. Due to the internet, reviews of many products and services are accessible online which includes restaurants too. Based on the literature study, researchers demonstrate the impact of reviews on the businesses. So, the restaurant also begun to use the universal element of electronic communications through social media and different food application platform, which not only help their product in marketing but also it helps the consumer to make decisions. There are many food web platforms available such as YELP, Zomato, Swish, Dine, food panda etc. through which consumer can find the best restaurant reviews based on food quality index, price range, cuisines, rating and frequency of customers.

Over the past few years, tourist's business has significantly growth. Due to the inadequate amount of time tourists are not able to strategize their trip efficiently i.e which places to visit, where to dine and which cuisines especially when the user is new to the place. With the help of online food platforms user can narrow down their choices and helps in decision making. Most of the population across the globe are excited about food quality and which restaurant is best in the business.

For the analysis of restaurant, Zomato is the best platform as it has highest number of restaurants registered around the world. The data is available at www.kaggle.com, which is publicly available for researches. The data is collected through Zomato API, the raw data collected was in JSON format later it was saved in comma separated values 'CSV'. For the analysis, we make use of HBase as database to store data from the CSV file. The table is exported from HBase to HDFS server through commands. Then the table hosted on HDFS server is fetched and loaded into PIG to perform MapReduce task using appropriate queries. The result of the MapReduce task is stored into local file system using -copyToLocal command and then visualizing it into MS Excel.

This paper is divided into following sections (II) Related Work. (III) Methodology. (IV) Evaluation of Results. (V) Conclusions.

Related Work

This paper demonstrates the analysis of online reviews of restaurant based on 360degree experience and factors such as food, quality and service. For this analysis, Michelin starred restaurant's dataset located Portugal and Spain was used. The overall ratings of the restaurant ranges from 0 to 5 and it is marked based on food, facility, worth and atmosphere. The analysis was done on 202 restaurants with the count of ratings approx. 82,000. Regression analysis was performed on the dataset i.e. means, standard deviation and correlations. The analysis outcomes demonstrate that the frequency of customer reviews have significant effect on restaurant reviews rating. The researchers also experimented sentiment analysis of text reviews the restaurant received. The analysis result was quite precise, but it cannot show the value for money is directly proportional to the ratings, quality and services.

The experiment is used to analyze food safety, health and hygiene of restaurants located across New York City. The dataset used for this study was collected via restaurant inspection done from December 2011 to end of 2016. Nearly about 24K restaurants were inspected and analyze per year across the city. Th dataset consists of half a million rows and it is publicly available on NYC open data website. The restaurants were letter Grade rating Grade A: 0-13, Grade B: 14 - 27, Grade C: 28 or more based on their safety, health and hygiene. The grade points indicate the area of doubt. The analysis was done based on number of restaurants fall in each type of cuisines and their aggregate rating. The analysis was done using Hadoop, HDFS, HBase as database, Hive to perform MapReduce task and MySQL to store output. The data was analyzed based on food specialize restaurants, but they lack in location of the restaurant area or zip code wise so that it would be easy to identify more deeper which region doesn't have standard food quality.

Methodology

Apache Hadoop: Hadoop is an evolutionary tool developed by the Apache Software Foundation (ASF). Hadoop is the open-sourced and most commonly used framework for storing and processing applications with huge amount of data. It provides distributed storage and distributed processing of clusters. This has been an evolutionary tool for processing big data, also for all analytical applications such as making predictions, data mining techniques and machine learning. Hadoop is good in dealing structured and unstructured data, which makes it as preferable tool for not only analyzing and processing huge data but also, for crafting it to user's requirement.

The hardware failures generally occur in the system and to control it automatically, every element of the Hadoop was designed based on fault tolerance assumptions. Due to fault tolerance, application continues to perform computational task in the case of node failure. In the recent past days, new upgraded Apache Hadoop 3.0 was released integrated with the current core elements such as HDFS, and MapReduce it was advanced with YARN cluster feature. Apache Hadoop YARN is responsible for managing the resources and scheduling of tasks.

HDFS: HDFS implies Hadoop Distributed File System, one of the vital core elements of the

Apache Hadoop Model. HDFS is extensible and dependable file system.

HDFS Features:

- High throughput time.
- Java-based file system.
- Adaptive to fault tolerance.
- Robustness.
- Smart recovery of loss data.
- Extensively scalable and reliable data storing.

Due to the above-mentioned features HDFS is effective in managing, accessing and forming required space for processing and storing big data.

HBase: It is an open source database management system, which provides real time read and write access to huge datasets. HBase is column-oriented DBMS. It runs on top of Hadoop Distributed File System (HDFS). It is commonly used for sparse datasets. HBase works reliably on the top of HDFS, along with more YARN data access engines.

Dataset Description

The dataset of restaurant was carried out by the researchers based on Zomato registered restaurant through Zomato API and it is publicly available on “www.kaggle.com”. The dataset has multiple different variety of columns which are used to analyze and identify which city has highest number of good restaurants based on ratings, votes and analyzing pattern of expensive restaurant with quality of food.

Restaurant ID (unique):

- City
- Address
- Location
- Cuisines (types)
- Average Cost for Two
- Currency (representing Country)
- Has Table booking (Y or N)
- Has Online Delivery (Y or N)
- Switch to order menu (Y or N)
- Price range (0-5)
- Aggregate rating (0-5)
- Rating Color (Red-orange-yellow-green)
- Rating Text (Excellent-Good-fair -poor)
- Votes

Apache PIG: It is a tool created by Apache software Foundations for analyzing and processing extensively large amount of data. It is an Open source software.

PIG features:

- Parallelization of tasks.
- Reading and Writing data from HDFS becomes simple.
- Simple to write queries in PIG.
- It can process any type of data i.e independent of data type.

Setting Up Hadoop Environment

For setting up Hadoop environment in our system, editing bash profile is extremely important. Editing bash profiles includes overall path of the Hadoop Distributed File System from reading and writing into database and executing PIG and MapReduce scripts. For successful implementation, we need to configure bash profile and all xml files as per our requirements. Below is our bashrc profile:

After successfully configuring and setting up the system, the very first step is to load the data from “CSV” file into HBase table. But before loading data into HBase make sure to run all Hadoop services by executing ‘sbin/start-dfs.sh’, ‘sbin/start-yarn.sh’, ‘bin/start-hbase.sh’ in terminal. In HBase, the cluster distributive property is set at True, by doing so it will store the data across range of clusters.

Initializing database and MapReduce environment.

Once the whole setup is complete and making sure all the services are running through command “jps”. Then we will load data from “Zomato.CSV” file into HBase table. For creating table in HBase, table name ‘project_restro’ with under namespace ‘pda_restro’ running script ‘createtable.sh’. After creating table, loading the data into HBase table through executing ‘load.sh’ in HBase home path ‘/usr/local/hbase’. But, while loading data we ensure that first column key in HBase table is unique. After loading the data into table ‘project_pda’ with under namespace ‘pda_restro’ script in HBase, we will observe that HBase stores the data across multiple clusters.

After successfully loading data into HBase table, the next step is to load the HBase table into HDFS. This is done by the inbuilt export function of hbase mapreduce class which exports the data from HBase to HDFS. The next step is most important step, load the data from HDFS into PIG. This is done through PIG grunt shell. In grunt, create table by running ‘pigloaddata.pig’ script which injected by code. Then comes executing queries in grunt shell to perform MapReduce task.

For running MapReduce Algorithm in Hive, we need to create a table database to load using ‘Create Table database’.

MapReduce Design Patterns

Design pattern is a critical technique to efficiently deal with frequently occurring data related issues. Under design pattern, MapReduce is an environment which has led the foundation for top level analysis tools such as PIG and Hive. There are number of classes in MapReduce patterns which are as follows filter.

MapReduce Algorithms

The first algorithm was used to carry out range of price per city has. The price range was from 0 - 5. As the range goes increasing from 0 to 5 the city is getting expensive.

1. A = GROUP by City;
2. B = Generate AVG Price range for EACH A;
3. DUMP B;
4. STORE Y USING PigStorage(‘,’) INTO HDFS;

The second algorithm was used to carry out average rating of zomato restaurant as per city across globe. The ratings are from 0 – 5.

1. X = GROUP BY (Currency, City);
2. Y =Generate AVG Aggregate rating for EACH X;
3. DUMP Y;
4. STORE Y USING PigStorage(‘,’) INTO HDFS;

The third algorithm was used to find out which cuisines is the best and most popular among population across the country.

1. A = GROUP BY Cusines;
2. B =Generate COUNT Votes for EACH A;
3. DUMP Y;
4. STORE A USING PigStorage(‘,’) INTO HDFS;

The fourth algorithm was used to carry out the total number of restaurants falls in which rating text category.

1. A = GROUP BY Rating_text;
2. B =Generate COUNT Restaurant_ID for EACH A;
3. DUMP B;
4. STORE B USING PigStorage(‘,’) INTO HDFS;

Experimented Results

The most expensive city for cuisines across the globe based on Zomato registered restaurants was carried out by running first algorithm. But we have visualized top 20 expensive Cities as the dataset was quite big to visualize. It can be interpreted from the graph that most of the cities around the globe lies in the budget, as there are almost nil cities in the range greater than 4.