International Journal of Multidisciplinary Allied Research Review and Practices

IJMARRP ISSN No: 2455-1570 (Volume 3, Issue 7, July 2016)

Analysis and Visualisation of Movie Lens Data Set Using Hive and R

Tripti Mehta
M.Tech Student, GITM Guragon

Sumit Hooda
Asst. Professor GITM, Gurgaon

Abstract— *Big data analytics is the process of examining large sets of data. Analysing Big Data is a challenging task as it is not viable to store such a large amount of data on a traditional data warehouse, making it prohibitively expensive, for that it involves large distributed file systems which should be fault tolerant, flexible and scalable. Hadoop is a popularopen source java based programming frameworkthat implements map-reduce which is being used in companies like Yahoo, Facebook etc. to store and process extremely large data sets in parallel on commodity hardware.. However, the map-reduce programming model is very low level and requires programmers to write bespoke programs which are less flexible, maintain and reuse.this problem was overcome by making use of HiveQL. To execute queries in HiveQL ,a platform is required i.e Hive .It is an open-source data warehouse solution built on top of Hadoop. HiveQL queries are compiled into mapreduce jobs that are executed using Hadoop .In this paper an attempt is made to analyse the movieLen data set and derive some interesting factswith Hive and R which is a popular software used for statistical computing and data visualization.*

Keywords*: - Big Data, HDFS, Hadoop, Hive, MapReduce,R*

I Introduction

The film industry has seen many technical changes over the years, making itself adaptable to its evolving audience's mentality and also contributing to that evolution. Using movies ratings & reviews, both target audiences and public reviews, we have a glimpse into the change process. Any of these data points, once analyzed, may contribute to greater business intelligence.

Analytics has delivered some important insights over the past year Now a days it is easy to obtain readable data from online resources such as the websites Box Office Mojo, BookMyShow.com and many more. if a site like Book My Show has a database of 980GB. It contains all kinds of details about movies like tickets sold, transaction details, information listed on the site. We can use data science and or machine learning to get really good useful information which might be useful for the company and for the entire Film Industry.

Quick and easy access to data has opened up a new world of business questions especially for movie recommendation. If we know the taste of customers, for example which genre of movies customers have booked in past, we can analyze the taste of audience and that too for different segments, different categories of age group, and different regions both at rural & urban level which might be highly useful for mainstream commercial Cinema. The movie makers uses the analytics to study their Fan base The better they understand customers, the more successful they will be giving them the best product.

After analyzing the users ratings ,based on ages, occupations etc the movie makers can have better understanding about the viewer‟s choice expectations which in turn is beneficial for marketing of their movies. This is done by determining the relationship between viewers‟ and their ratings. By making use of effective BigData analysis tools like in this paper Hadoop , R and Hive are made

use of, larger datasets can be analyzed which provides statistically accurate results. These findings provide better understanding about viewers‟ expectations and hence movie choice.

In this paper, we use MovieLens dataset for the analysis purpose .which is an open dataset collected by GroupLen Research Project at the University of Minnesota [8].This dataset is made available for the users on the website to rate the movies. This dataset consists of: 100,000 ratings (1-5) on 1,700 movies from 1000 users.Atleast 20 movies are rated by each user in this dataset. Each user has rated at least 20 movies.,1M ratings on 4000 movies from 6000 users and 72,000 users made 100 thousand ratings on 10,000 movies respectively[2].

II. PROPOSED SYSTEM

A. Data set

The data set used for this study is taken fromhttp://grouplens.org/datasets/movielens/100k/. The system predicts and provides the users with suggestions based on their previous ratings recorded. These findings providebetterunderstanding about viewer expectations and hence movie choice.This data set is analysed using hive and R.

B. Architecture of proposed system

Fig 1 shows the proposed system architecture . The MovieLens data set is given as input to the system, having componentsHive and R.The data set which we have taken is raw ,at first glance it is giving no

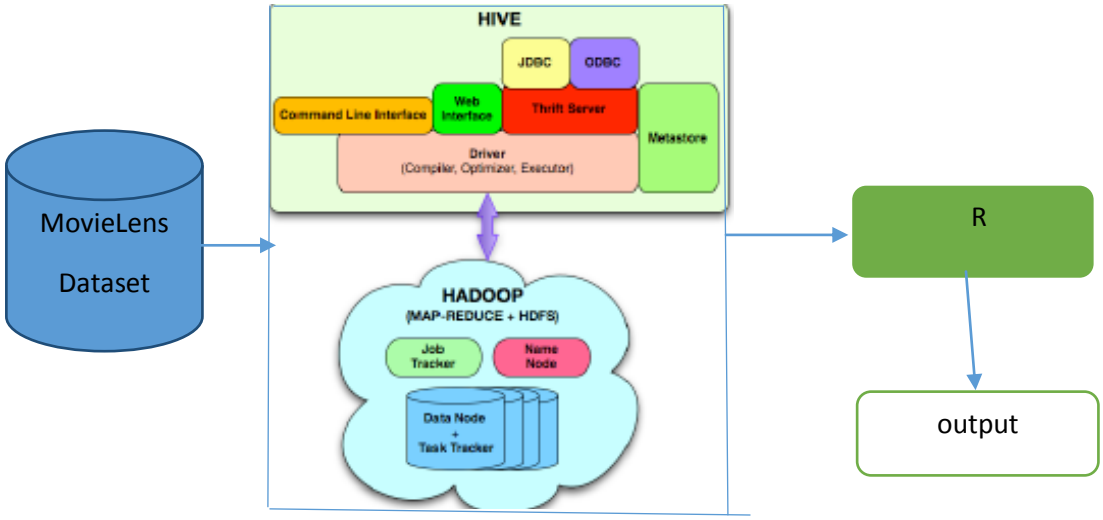

Fig 1: Architecture of a Proposed System

sense. This data is provided as a input to HIVE,then this dataset is analysed and partitioned based on different attribute like genre, occupation, ratings etcthe output that is obtained from hive is well formatted data then proper analysis of this data set will give some interesting facts which is providedas input to R.R is a programming language and software environment for statistical analysis, graphics representation and visualisation[7].We all know that pictures speak more than words, after analysing the data using hive the graphs are generated for each data set using R for visualisation.

C. Hive and R

In this research we have used hive and R for the purpose of analysing the data set.

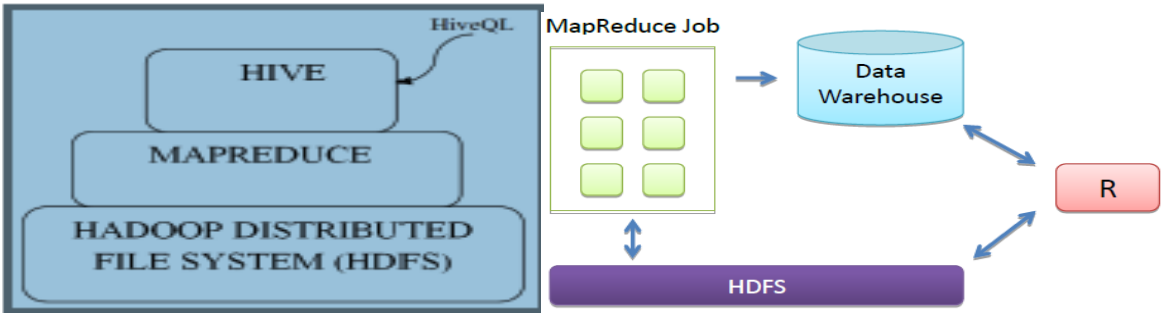

Fig 2: System for Querying and Analysing data using Hive and R

Hive summarises the Big Data makes querying and analysing easy It is a data warehouse infrastructure tool which process structured data and resides on top of the Hadoop to make it possible.HiveQL-the declarative language supported by HIVE is used to express queries which is like SQL. These queriesare thencompiled into Map Reduce jobs and then executed on Hadoop cluster. In addition, HiveQL enables users to plug in custom map-reduce scripts into queries.Hadoop Distributed File system (HDFS) with complete Meta data repository is used for storing flat files in the form of tables. HiveQL is used for querying those tables. Hive keeps the metadata in a relational database to support features like schemas and data partitioning. To know the contents of the HDFS in Hadoop, one needs to write Map Reduce programs. Hive supports partitioning of the table on a particular dimension. For example like we partition the MovieLens dataset on the ratings, occupation etc. This allows to later creating queries on an organized data model.

To the end users who may have no idea about map reduce or no interest in writing Map Reduce programs,Hive is an interesting project because it allows exposing the best parts of hadoop, namely Map reduce and data storage. [4]
AnRprogrammer, being able to read/write files in HDFSas the basic storage mechanism in Hadoop is HDFS (Hadoop Distributed File System).Bounded by the memory constraints of R, this capability allows the analyst to easily work with a data subset and begin some ad hoc analysis it also enables the R programmer to store models or other R objects that can then later be recalled and used in MapReduce jobs. When MapReduce jobs finish executing, they normally write their results to HDFS. Inspection of those results and usage for further analysis in R make this functionality essential [7].

III. EXPERIMENTAL RESULTS

As it was mentioned earlier Hive and R are used for the purpose of analysis. In Hive the data set should be first loaded to it, hence the MovieLens data set is first loaded to Hive. The raw data is just '#' separated which is loaded to Hive.

3.1 MovieLens Dataset schema

For the ease of analysis 100K data set has been chosen from the websitehttp://grouplens.org/datasets/movielens and stored in HDFS. The movies.dat, ratings.dat and the users.dat files have [movieID, tile, gednre], [userID,movieID, rating,

timestamp] and [userID, gender, age, occupation, zipCode] fields respectively;[9] with each field delimited from the other by # symbol.

3.2 Creating tables and loading data

For the above mentioned three files movies, ratings and users tables with the same schema is created. Hive query to create movies table and result for the same is as shown in Fig 3.Similarly, tables have been created for ratings and users files based on their attributes respectively

```
CREATE TABLE movies(movieid INT, title STRING, genres STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '#'
STORED AS TEXTFILE;
```

```
hive> CREATE TABLE movies(
  movieid INT,
  title STRING,
genres STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '#'
STORED AS TEXTFILE;
OK
Time taken: 0.315 seconds
```

Fig 3: Creating movies

The next step is to load the data into the tables after creating allthe three tables. Hive provides us with theutilities to loaddatasets from flat files stored on HDFS using the LOAD DATAcommand.

LOAD DATA LOCAL INPATH <"path_to_flat_file"> OVERWRITE INTO TABLE <table_name>;
The result is as shown below in the Fig 4

```
hive> load data local inpath '/home/training/Desktop/movies.txt' overwrite into
table movies;
Copying data from file:/home/training/Desktop/movies.txt
Copying file: file:/home/training/Desktop/movies.txt
Loading data to table default.movies
Deleted hdfs://localhost/user/hive/warehouse/movies
OK
Time taken: 1.223 seconds
hive>
```

Fig: 4 LOADING DATA INTO MOVIES
The same process is used for loading data into other two tables.

3.3 APPLYING HIVE QUERIES ON DATASETS

1.Effect of occupations on ratings

```
1. SELECT occupations.occupation, count(*)FROM users
JOIN  ratingsON(ratings.userid=users.userid)
JOIN occupations on(users.occupation=occupations.id)
WHERE rating=5
GROUP BY occupations.occupation
2. SELECT occupations.occupation,count(*) FROM users
JOIN occupations ON(occupations.id=users.occupation)
JOIN ratings ON(ratings.userid=users.userid)
WHERE rating=5
```

```
GROUP BYoccupations.occupation,gender;
```

2. Effect of age on ratings

```
1.SELECTusers.age,count(*) FROM ratings
JOIN users ON(ratings.userid=users.userid)
        WHERE rating=5
GROUP BY users.age;
```

Fig 5. and Fig 6 shows the snapshot of the graphs generated using R

The graphical representation can be used to interpret some interesting facts about age groups. Fig 5 shows that Users tend to be mostly in the late teens and mid-thirties, there is an another peak the occurs in the late forties and in Fig 6 there are very few doctors and homemakers, so we can't say anything about these groups with very much confidence

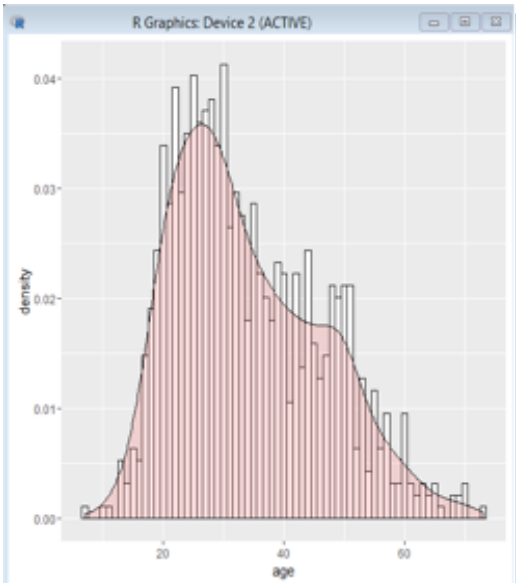

Fig 5:Effects of age

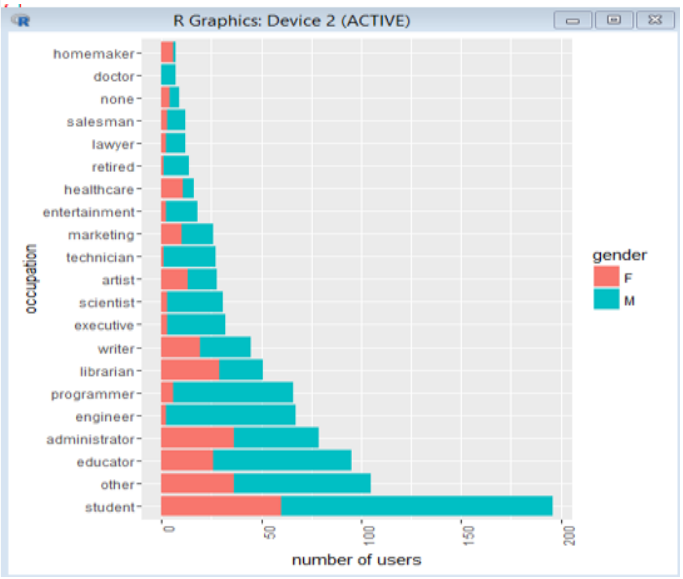

Fig 6:Effect of occupation and gender

IV. CONCLUSIONS AND FUTURE WORK

The conclusion on this topic is that we have executed and reviewed various queries implementation on Hive for large datasets. Mapping and reducing functionalities of Map-reduce and HDFS helped Hive to process larger and unstructured datasets. Executing R code in the context of a MapReduce job elevates the kinds and size of analytics that can be applied to huge datasets, which is not possible by other traditional processing system. Hence Hadoop and R together can be used to solve a variety of analysis problems more efficiently.The facts which were revealed during the process can be used for developing some prediction models.Hive is less optimized as compared to traditional databases like Oracle, MySQL, and PostgreSQLso optimization in Hive has good research possibilities. Also, Hive does not

support Update and Delete functionality yet. So research can be progress in this areaThe facts which were revealed during the process can be used for developing some prediction models.

References:

1. A survey on Big Data Analytics using Map Reduce and Hive on Hadoop Framework Tripti Mehta Dr. Neha Mangla ,IJRAET-Feb2016, Vol-4-I-2(NCRISTM),
2. *Neha Mangla,ShanthiMahesh,ChhayaM, Vidyashree G, Vikas,Atria Institute of Technology, Bangalore.* International Journal of Engineering Research ISSN:2319-6890)(online),2347-5013(print)Volume No.5 Issue: Special 4, pp: 790-991 20 May 2016
3. Deeksha Lakshmi, 2 Iksuk Kim, 3 Jongwook Woo, ARPN Journal of Science and Technology, VOL. 3, NO. 12, December 2013 ISSN 2225-7217
4. Integrating R and Hadoop for Big Data Analysis Bogdan OANCEA (bogdanoancea@univnt.ro) "NicolaeTitulescu" University of Bucharest Raluca Mariana DRAGOESCU (dragoescuraluca@gmail.com) The Bucharest University of Economic Studies
5. Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang,Suresh Antony, HaoLiuandRaghotham Murthy*Facebook Data Infrastructure Team* Hive – A Petabyte Scale Data Warehouse Using Hadoop
6. http://www.oceantara.com/overview-of-hive-for-hadoop/
7. http://www.revolutionanalytics.com
8. http://grouplens.org/datasets/movielens/
9. http://www.ijer.in/ijer/publication/v5si5/04.pdf
10. http://www.iosrjournals.org/iosr-jce/papers/vol18 -issue 2/version-3/n1802038292.pdf
11. https://cwiki.apache.org/confluence/display/Hive/Home;jsesionid=986ED81CB5EB4E18AB21A6BDDE4610EA