VOL. 3, NO. 12, December 2013 ISSN 2225-7217
ARPN Journal of Science and Technology

http://www.ejournalofscience.org

Analysis of Movie Lens Data Set using Hive

[1] Deeksha Lakshmi, [2] Iksuk Kim, [3] Jongwook Woo
[1] Grad Student, Department of Computer Information Systems
[2] Assoc. Prof., Department of Marketing
[3] Assoc. Prof., Department of Computer Information Systems
California State University Los Angeles

[1] Deeksha.Lakshmi@calstatela.edu, [2] ikim@calstatela.edu, [3] jwoo5@calstatela.edu

ABSTRACT

Large scale data set provides the better opportunity to find out much better data relationship in the area of business intelligence. In the paper, we implement our systems using Hadoop that has been popular to store and compute Big Data. However, it is not easy to write Hadoop Map Reduce code. Therefore, we use Hive and Hive QL codes to understand the relationships between ratings and the users' profiles for the different movies in the Movie Lens data set.

Keywords: *Hadoop, Big Data, Hive, Data Analysis*

1. INTRODUCTION

Data grows exponentially past several years because of social network, sensor networks, bioinformatics, and smartphones and it is called Big Data. It is expensive to store and process Big Data with the legacy approach, many solutions has risen. However, Hadoop framework has received most popular highlights [4]. Hive has been used for the data analyst as its syntax is similar to SQL [2] and it runs on Hadoop framework in parallel. Thus, even though analyst does not know programming languages such as Java and Perl, s/he can easily write codes in Hive QL (Hive Query Language) [3].

All major companies realize the need for better marketing to reach their desired target audience. Many new methods have been tried, tested and successfully implemented by the industry. The latest in this category is the use and analysis of Big Data to understand the audience and target them accordingly. In this study we use a Movie data set which is [1]movie lens data set [1] and comprises of a number of users of different age groups from 18 to 60 and a number of genres of movies from many years. The user's ratings for the movies range from 1 (bad) to 5 (excellent). We analyze the database using My SQL, Sqoop, Hive to find patterns between the user's demographics and the movie ratings.

Section 2 summaries the movie data set. Section 3 describes the data analysis view. In Section 4, we describe the components of our data analysis systems and Hive codes. Section 5 discusses the results of our experimental evaluation. We conclude the research in Section 6.

2. HIVE and MOVIE DATA

This section briefly introduces Hive and Movie lens data set. Hive is a declarative programming language that runs on Hadoop framework. Movie Lens is a open data set provided by University of Minnesota.

2.1 Hive

While Yahoo built Hadoop Map Reduce systems to analyze Big Data, Yahoo realized that it is too expensive to write Map Reduce codes in high level language such as Java, Ruby, and Perl. Thus, Yahoo designed data flow language called Pig, that is relative easy to write a data analysis code with the simple syntax, for example, business logic in 100 lines of Java code can be built in 5 – 10 lines of Pig code.

Similarly, Face book implemented Hive QL that is a declarative language with the similar syntax of SQL. Even though it was developed for data warehousing in Face book, now it is an Apache open-source project. Hive runs on the client machine and the queries are submitted to the Hadoop clusters. In the experiments of the paper, Hive is used.

2.2 Movie Lens

Group Lens research by University of Minnesota has collected movie data set and made the web site available for the users to rate the movies [1]. Movie Lens data set is composed of 100K, 1M, 10M data sets, which have 100 thousands ratings from 1,000 users on 1,700 movies, 1 million ratings from 6,000 users on 4,000 movies, and 10 millions ratings with 100 thousands tags from 72,000 users on 10,000 movies respectively.

The paper uses 1M data set that is composed of RATINGS, USERS, MOVIES data sets with the following fields:

MOVIERATINGS: [User ID, Movie ID, Rating, Timestamp]
USERS: [User ID, Gender, Age, Occupation, Zip-code]
MOVIES: [Movie ID, Title, Genres]

3. IMPLICATIONS OF DATA ANALYSIS

This research provides an opportunity to investigate the relationship between movie audience and movie ratings. Like any other area in the field of marketing, understanding audience profile plays a major role in movie marketing.

[1] We acknowledge the use of the Movie Lens data set by Group Lens Lab at the Department of Computer Science and Engineering, University of Minnesota, Twin Cities.

This study supports the generalization of the relationship between movie ratings and audience profile found in the Big Data set. From a theoretical perspective, the findings of this research lend further support to audience profile and movie rating, showing that Big Data analysis is a useful tool for investigating audience profile. Consequently, this research provides a foundation for marketing researchers who want to offer further explanation for theories of consumer behaviour in the movie marketing environment. In other words, this research does not exhaust in any way what we know from past research about the previous the relationship between audience and consumer movie choice behaviour. The area of Big Data and consumer behaviour can now be fully exploited as it relates to and potentially explains issues in movie audience behavior.

Second, this research broadens the power of outcomes by the size of the dataset. Marketing research is typically burdened by the data set dilemma: small data sets are best for efficient data collection, but generating results with better statistical durability would require a large data set. However, through this research, we empirically demonstrate that statistically durable and efficient data sets can converge. This clearly suggests that the Big Data method can be extended to current areas of marketing research.

Managerially, this research provides marketers with the importance of the audience profile in the movie business. For movie marketers, awareness of audience attributes is important because the audiences' profile and ratings become a crucial factor in movie choice. In response to the specific findings of this research, if movie marketers desire to improve ratings, they must attempt two things. First and foremost, movie marketers must realize that the audience profile is not a neutral factor to all genres of movies, but one that a priori is seen by audiences as more diversified. Since certain audience profiles increase the probability of a positive movie rating, marketers must understand the nature of the profile of each and every movie. Finally, the results highlight the importance of constructing the Big Data set. The lack of a large dataset feeds the uncertainty of the audience profile and impairs the success of approaching the target audience.

4. MOVIE DATA ANALYSIS USING HIVE

In order to collect Movie Lens data set, we built the systems as shown in Fig 1. Movie Lens data set is migrated to and stored at My SQL database in the beginning. And, using Sqoop [5], data from the database is transferred to HDFS (Hadoop Distributed File Systems) in the process of Map Reduce in parallel. Sqoop is a Hadoop ecosystem to transfer data of database to Hadoop HDFS systems in parallel using Map Reduce.

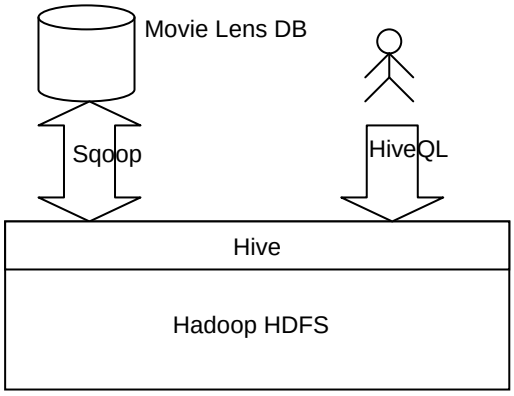

Fig 1: Data Analysis Systems

4.1 Characteristics of the My SQL Database

The Movie Lens database has the following tables: MOVIE, GENRE, MOVIEGENRE, MOVIERATING, OCCUPATION and USER. Each of these tables has a specific set of columns which indicate necessary information.

Here below is a list of some of the general numbers pertaining to these tables. Total Number of RATINGS 1-5 all included is 1,000,205. Total Number of 5 ratings in the MOVIERATING table is 226,309. Total Number of USERS is 6,040.

4.2 Characteristics of the Hive Data Set

As data is stored at HDFS of the systems describe above, MOVIE, MOVIERATING, USER, and OCCUPATION tables are created at HDFS using Hive QL, which is SQL-like as follows:

```
CREATE EXTERNAL TABLE movie
(Id INT, name STRING, year INT)
ROW FORMAT DELIMITED FIELDS TERMINATED by '\t'
LOCATION '/user/kodicalk/movie'

CREATE EXTERNAL TABLE movie rating
(use rid INT, movie id INT, rating INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
LOCATION '/user/kodicalk/movie rating'

CREATE EXTERNAL TABLE user
(id INT, gender STRING, age INT, occupation id INT, zip INT)
Row format delimited fields terminated by '\t'
Location '/user/kodicalk/user'

CREATE EXTERNAL TABLE occupation
(occupation id INT, name STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
LOCATION '/user/kodicalk/occupation'
```

5. EXPERIMENTAL RESULT

On Amazon AWS, *AWS EC2 m1.small* instance type nodes are launched as a Hadoop cluster. Those are run on small instances of *AWS EC2*, where each instance as *m1.small* is $0.06/hr. and is composed of 1 core (1 *EC2* compute unit), 1.7GB memory and 160GB storage on 32 bits platform with *Ubunt-10.02 OS* and Hive is running on the cluster in order to observe the effect of user demographics on the ratings as follows:

5.1 Effect of Gender on the Ratings

The effect of the user's gender is analyzed on the ratings that the users give to the movies. The Hive QL to get the effect of the user's gender with the rating 5 is as follows and its result is shown at Table 1.

```
SELECT u.gender, COUNT(*)
FROM movie rating r, user u
 WHERE r.userid = u.id  AND r.rating = 5
GROUP BY gender;
```

Table 1: Rating of 5 split by gender

GENDER	COUNT
F	58,546
M	167,763

Table 1 shows that more numbers of males rate a movies as '5' than females.

5.2 Effect of Occupation on Ratings

The following Hive QL is to select data with the rating of 5 split by different occupation categories and Table 2 shows the result.

```
SELECT o.name, count(*)
FROM movie rating r, occupation o, user u
WHERE r.userid = u.id
AND o.id = u.occupationid
AND r.rating = 5
GROUP BY o.name;
```

Table 2: Rating of 5 split by different occupation categories

OCCUPATION	COUNT
college/grad student	30,272
other or not specified	28,178
executive/managerial	23,044
academic/educator	18,603
technician/engineer	16,208

The following Hive QL is to select data with the rating of 5 split by different occupation and gender categories and Table 3 shows the result Effect of occupation and gender on ratings

```
SELECT o.name, count(*)
FROM movie rating r, occupation o, user u
WHERE r.userid = u.id
AND o.id = u.occupationid
AND r.rating = 5
GROUP BY o.name, gender;
```

Table 3: Rating of 5 split by different occupation and gender categories

OCCUPATION	GENDER	COUNT
college/grad student	M	22,927
other or not specified	M	19,698
executive/managerial	M	18,892
technician/engineer	M	14,346
programmer	M	11,773
academic/educator	M	11,476

In all the occupations, Table 2 shows that the college students give the rating of 5 in the most. And, Table 3 presents that the number of males giving a rating of 5 is higher than the number of females giving a rating of 5 except in the homemakers section.

6. EFFECT OF AGE ON RATINGS

The following Hive QL is to select data with the range of ages on rating of 5 and Table 4 shows the result.

```
SELECT u.age, count(*)
FROM movie rating r, user u
WHERE r.userid = u.id
AND r.rating = 5
GROUP BY u.age;
```

Table 4: Effect of age on rating of 5

AGE	COUNT(*)
1	6,802
18	40,558
25	85,730
35	44,710
45	19,142
50	18,599
56	10,768

Table 4 illustrates that the users in the age group of 25 give the highest rating of 5 followed by age group 35 and age group 18.

The following Hive QL is to select data with the range of ages and the occupations on rating of 5 and Table 4 shows the result.

```
SELECT o.name, u.age, COUNT(*)
FROM user u, occupation o, movie rating r
WHERE r.userid = u.id
AND o.id = u.occupationid
AND r.rating = 5
GROUP BY o.name, u.age;
```

Table 5: Effect of Age and Occupation on Ratings

OCCUPATION	AGE	COUNT
college/grad student	18	20,273
other or not specified	25	13,381
college/grad student	25	8,962
executive/managerial	25	8,952
executive/managerial	35	7,051

Table 5 shows that college/grad students in the ages of 18 give the rating of 5 in the most, Then, others in the ages of 25 give the rating of 5 at the next.

7. EFFECT OF OCCUPATION AND GENDER ON RATINGS

The following Hive QL is to select data with the genders and the occupations on rating of 5 and Table 6 shows the result.

```
SELECT o.name, gender, COUNT(*)
FROM user u, occupation o, movie rating r
WHERE r.userid = u.id
AND o.id = u.occupationid
AND r.rating = 5
GROUP BY o.name, gender;
```

Table 6: Effect of occupation and gender on ratings

AGE	OCCUPATION	GENDER	COUNT
18	college/grad student	M	15,239
25	other or not specified	M	9,509
25	college/grad student	M	6,934
25	executive/managerial	M	6,928
25	technician/engineer	M	6,104
25	programmer	M	5,631

Table 6 shows that the number of Male users with a rating of 5 is higher than the number of Female users in any age category, especially in the group of college/graduate students.

8. CONCLUSION

Large scale data set gives us the better opportunity to find out much better investigation in the area of the business intelligence.

The paper shows how Big Data can be analyzed easily using Hive on Hadoop frameworks in parallel. The paper presents the systems that transfer data from My SQL DB to HDFS using Sqoop. Using Hive, Hive QL codes, SQL-like codes, are processed in parallel on Amazon AWS.

In order to show the relationships between ratings and the users' profiles, Movie Lens data is analyzed to find out the ratings of 5 by gender, age, occupation, and both occupation and gender. The results illustrate that the male users around ages 25 as college/graduate students give the ratings of 5.

REFERENCES

[1] Movie Lens Data Set, Group Lens Lab, Department of Computer Science and Engineering, University of Minnesota, Twin city, http://license.umn.edu/technologies/z05173_movielens-database

[2] Apache Hive, http://hive.apache.org/

[3] Apache Hive Query Language Manual, https://cwiki.apache.org/confluence/display/Hive/LanguageManual

[4] Apache Hadoop Project, http://hadoop.apache.org/

[5] Apache Sqoop, http://sqoop.apache.org/

[6] "Market Basket Analysis Algorithms with Map Reduce", Jongwook Woo, DMKD-00150, Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, Oct 28 2013, Volume 3, Issue 6, pp445-452, ISSN 1942-4795

[7] "Market Basket Analysis Algorithm on Map/Reduce in AWS EC2", in International Journal of Advanced Science and Technology (IJAST), Jongwook Woo, Science & Engineering Research Support society (SERSC), Sept 2012, Volume 46, No 3, pp25-38, ISSN 2005-4238

[8] "Market Basket Analysis Algorithm with No SQL DB HBase and Hadoop", Jongwook Woo, Siddharth Basopia, Yuhang Xu, Seon Ho Kim, The Third International Conference on Emerging Databases (EDB 2011), Songdo Park Hotel, Incheon, Korea, Aug. 25-27, 2011

[9] "Apriori-Map/Reduce Algorithm", Jongwook Woo, The 2012 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2012), Las Vegas (July 16-19, 2012)

AUTHOR PROFILES

Deeksha Lakshmi is currently pursuing Masters in Information systems at California State University Los Angeles. She completed her BS in Instrumentation Engineering from Visveswariah Technological University, India in 2006. Her interests include pattern analysis in Big Data using Hadoop Map Reduce and Hive.

Iksuk Kim is currently an Associate Professor at Department of Marketing, California State University Los Angeles. He received the MS and Ph. D , from Purdue University in 1997 and 2001, respectively. His research interests are Big Data application in Marketing, and Entertainment Marketing.

Jongwook Woo is currently an Associate Professor at Computer Information Systems at California State University, Los Angeles. He received the BS and the MS degree, both in Electronic Engineering from Yonsei University in 1989 and 1991, respectively. He obtained his second MS degree in Computer Science and received the PhD degree in Computer Engineering, both from University of Southern California in 1998 and 2001, respectively. His research interests are Information Retrieval /Integration /Sharing on Big Data, Map/Reduce algorithm on Hadoop Parallel/Distributed/Cloud Computing, and n-Tier Architecture application in e-Business, smartphone, social networking and bioinformatics applications. He has published more than 40 peer reviewed conference and journal papers. He also has consulted many entertainment companies in Hollywood.