Movie Dataset Analysis using Hadoop-Hive

Ashwitha T A
Department of Computer Science and Engineering
NMAM Institute of Technology, Nitte
ashwithapoojary213@gmail.com

Anisha P Rodrigues
Department of Computer Science and Engineering
NMAM Institute of Technology, Nitte
anishapr@nitte.edu.in

Niranjan N Chiplunkar
Department of Computer Science and Engineering
NMAM Institute of Technology, Nitte
nchiplunkar@nitte.edu.in

***Abstract*—In today's world there is a huge growth in data. This data is generated from variety of sources like social media, industry, transaction records, cell phone, GPS signals etc. It is difficult and challenging to store such a huge amount data in traditional data warehouse. Big Data is the dataset with 3 V's that are Volume, Variety and Velocity and difficult to store and process using traditional database management systems. Big Data Analytics is the way of processing the large amount of data. Hadoop is a popular open source software which is very useful in analyzing the larger data. Hadoop provides several tools for this purpose like Hive, Pig, Hbase, Cassandra etc. In this paper, we have used Hadoop framework. For the analysis of movie dataset Hive tool is used with Hadoop framework. We have got significant improvement in processing time for analyzing dataset compared to traditional system**

Keywords—Big Data, Hadoop, Hive,Map-Reduce,HDFS.

I. INTRODUCTION

Today the data is growing in a very high speed. These data can be produced by any sources like industries, social media, cell-phones, scientific source etc. We can refer this large data as Big Data. Till now there is no particular measure is defined on size of the Big Data [1]. In the beginning, Big Data was adopted by Facebook, LinkedIn, Google etc. The reason might be the rapid change of the data [3]. The three characteristics of Big Data are volume, velocity and variety. Volume refers to the size of the data. It is growing bigger day by day. According to the expert's analysis, few years later the data can cross 25 Zettabytes. Velocity refers to how fast the data is being processed. For this we can consider the examples like posting comment or image on Facebook and watching video on YouTube etc. Variety is referring to different types of data and different sources which produce these data. The data can be structured, semi-structured or unstructured. It can be in any formats like image, csv files, text files, audio, video etc. In addition to these characteristics, two more have been defined: veracity and value. Veracity refers to the trustworthiness of the data and value refers to extracting meaningful information from datasets [2].

As the data are produced, we need to store and analyse it. The traditional data processing techniques are failed to analyse the Big Data. Today about 80% of data are unstructured and these unstructured data are impossible to analyse using relational database systems. So Big Data analytics is introduced for processing these large amount of structured, unstructured and semi-structured data with the help of existing software tools and with very less amount of time [2]. Big data analytics makes use of Hadoop framework for analyzing the larger datasets. Hadoop is an open-source, reliable, scalable and shared computing framework developed especially to process the huge amount of data. It was initially developed by Google in 2004 and at present it is maintained by Apache. The main two components Hadoop are Hadoop Distributed File System (HDFS) and MapReduce. HDFS stores the data in the form of blocks. MapReduce is used for processing, sharing and clustering [3].

Hadoop ecosystem provides several tools for Big Data analytics. Some of the tools are Hive, Pig, Hbase, Cassandra, Mahout, Flume, Avro etc.

Pig is an analysis tool of Hadoop ecosystem. Pig has its own programming language called Pig Latin. To covert the Pig Latin scripts into MapReduce job Pig Runtime is used. Hive is called as data warehousing software. It has its own query language called HiveQL. It is used for processing the larger datasets stored in data warehouse. Hbase is used to store the large amount of data especially structured data. It stores data in the form of tables and Pig or Hive queries can be used to analyse these data. Like Hbase, Cassandra is also a database which stores the structured data. In Casandra, the data is replicated in several nodes so even if the one node fails it doesn't make any difference. Its data will be available in other nodes. Therefore Cassandra is called as fault-tolerant system. Hbase and Cassndra are NoSQL databases and they are column oriented. Mahout is a framework developed by Apache. The main purpose of this is to implement the data mining algorithms. Flume is another data analyzing tool provided by Hadoop eco-system. It is used especially to analyse the log data. Its architecture is very simple and it is robust. Avro is schema-dependent. It is mainly used to serialise the data so that it can analysed easily. It declares different data structures with the help of JSON format. Java, C, C++, Python, C# and Ruby are the languages supported by Avro [3].

In this paper we have used Hive with Hadoop framework for analysing movie dataset. Hive is built on the top of Hadoop and it has its own query language HiveQL which is similar to SQL. Hive will internally convert the queries into MapReduce job [3]. The reason for why Hive is better than MySQL is that, Hive is most suitable for larger datasets and MySQL is suitable for smaller datasets.

978-1-5386-2044-1/17/$31.00 ©2017 IEEE

In this paper section II contains the related works. Section III describes the architecture of proposed system. The analysis and result of the experiment is described in section IV and V.

II. RELATED WORKS

Ammar Fuad et al. proposed a method to analyze the performance MySQL cluster, Hive and Pig [7]. For the experiment three different sizes of movie lens datasets are considered. The result showed that MySQL cluster processing time increases as the data size increases. Because of step-by-step execution nature of the Pig, its processing time exceeded the processing time of the Hive.

Karan Sachdeva et al. compared the performance of Map-Reduce, Hive and Pig by considering unstructured, semi-structured and structured dataset [8]. From the result it's been proved that to process structured data Hive is the efficient tool. For processing semi-structured and unstructured data Pig and Map-Reduce respectively are efficient.

Analysis of Meteorological and Oceanographic data is difficult using MySQL because it consists of several number of small files and each file might contain 20 to 300 columns. Ali Usman Abdullahi et al. have analysed the Meteorological and Oceanographic data for indexed and non-indexed table using Hive [9]. There are three different types of queries have been executed on the tables. From the result it is shown that type 1 and type 3 queries have shown better response time for indexed table. The type 2 showed different processing time for indexed and non-indexed table depending upon the size of the data. It is possible to reduce the number of mappers so that response time can be increased.

Aditya Bhardwaj et al. have analysed Twitter data using Hive [10]. Analysis is performed to predict the Map-Reduce time and total job completion time for different cluster size. From the result it is proved that as the cluster size increases the Map-Reduce time also increases and total job completion time decreases.

It is possible to increase the performance of Hadoop/Hive with help of Multi Query Optimization technique and distributed Hive. Varun Garg [11] considered 11 queries from TPC-H and different sizes of datasets are used. From the experiments the author has shown that performance of distributed Hive is greater than the conventional Hive.

Xiaoyu Wang et al. [12] performed analysis on Internet traffic data using Hadoop and Hive based traffic analysis system. The libpcap files are pre-processed with less amount of time. They have shown that the system proposed by them is error free and increases execution speed.

Taoying Liu et al. [13] presented the implementation Standard Science DBMS which is a benchmark of distributed scientific data on Hive. Hive queries are compared with SciDB queries. The amount of time taken to load the data by both SciDB and Hive is same. For the smaller input data the performance of Hive is slower when compared to SciDB.

S K Pushpa et al. [14] analyzed airport data using Hive and found out that it is more efficient and faster when compared to traditional approach. Dharaben Patel et al. have analyzed the huge amount of network traffic data using Hive [15]. Hive queries are written to find-out different types of security attacks. To visualize the result Apache Zeppelin tool is used. As the part of future work using this method more number of security attacks can be found.

Hive performance time can be predicted by determining the Map-Reduce job execution time [16]. Amit Sangroya et al. have proposed a linear regression model. Hive processing time decreases with increase in the size of data. The proposed method helps in predicting the Hive performance time with reduced error rate.

The Connected Vehicles can exchange information about location and security. Large amount of data are produced by Connected Vehicle. Weija Xu et al. [17] analyzed these data using Hive and compared result with PostgreSQL. The experiment conducted showed that Hive query performance time is lesser than PostgreSQL.

The Earth science data are always in NetCDF format. This format is not supported in HDFS, therefore we cannot analyse this data using Haddop tools. Shujia Zhou et al. [19] proposed a system that will convert the NetCDF format to CSV format making it easy to visualize and to analyze by Haddop Tools like Spark, Hive.

III. PROPOSED SYSTEM

DATASET

The datasets are taken from IMDb website.The Movie dataset consists of information about the year of release, title, language, imdb ratings, FB likes, genre etc. In this paper we have considered four datastes for analysis: Movie dataset, Genre dataset, Rating dataset, User dataset. Movie dataset consists of 10650 rows, Genre dataset consists of 5044 rows, Rating dataset consists of 6-7 lakh rows and User dataset consists of 6040 rows. The description about the datasets is as follows.

MOVIE TITLE	Title of the movie.
LANGUAGE	Language used in movie.
YEAR	Movie release year.
IMDB RATINGS	IMDB ratings for a particular movie.
FB LIKES	Facebook likes obtained for a particular movie.

Table 1: Movie Dataset

MOVIE ID	Unique ID for a particular movie.
TITLE	Name of the movie.
GENRE	Describes the category to which a movie belongs.

Table 2: Genre Dataset

USER ID	Unique ID for a particular user.
MOVIE ID	Unique ID for a particular movie.
RATINGS	The rating given by the user for a particular movie.

Table 3: Rating Dataset

USER ID	Unique ID for a particular user.
GENDER	Describes whether user is male or female.

Table 4: User Dataset

ARCHITECTURE

Hadoop is open-source software maintained by Apache. Hadoop stands for High-availability distributed object oriented platform. This framework is used for shared processing of huge datasets across multiple clusters with the help of simple programming models. Hadoop has two different components for storing and processing of data. The storing component is Hadoop Distributed File System (HDFS) and processing component is MapReduce [5]. HDFS is a file storing system of Hadoop and it was invented by Doug Cutting. HDFS divides the input into number of blocks. These blocks will be stored in datanodes. The datanodes here refers to the machines in the cluster. In order to access the blocks Name nodes are used. HDFS is made fault tolerant by replicating the blocks. The default replication factor is 3 [3]. MapReduce has two phase: Map phase and Reduce phase. The input and output of both the phases are key-value pair. Mapper divides the input into several subproblems and Reduce combines it to produce solution. MapReduce consists of Job tracker and Task tracker. The job is assigned to the task tracker by job tracker [5].

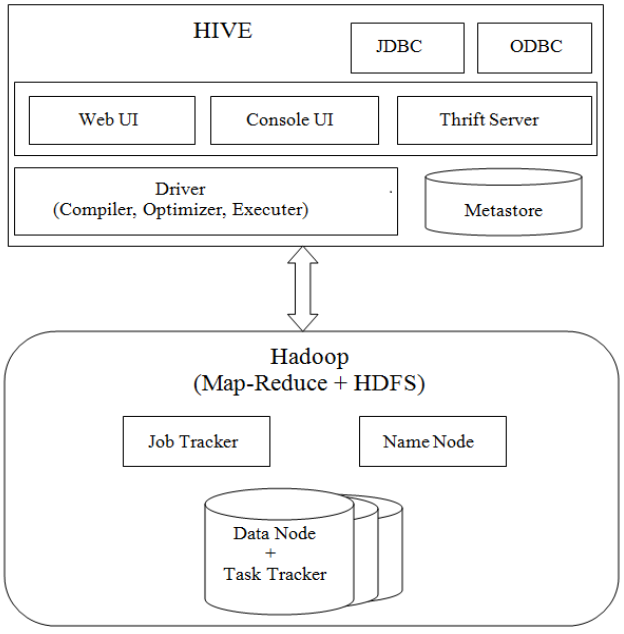


Figure 1: Hive Architecture

Hive was initially popular within facebook users. It is built on the top of the Hadoop. As shown in the figure 1, Hive consists of metastore, driver, thrift server, JDBC/ODBC driver, query compiler [6]. Metastore contains the information about the table like its schema, location etc. Driver is like a controller. It is used to execute Hive queries. The external user can interact with the system through thrift server. Query compiler is used to compile HiveQL into directed acyclic graph. The tables, partitions, and buckets are used to organize the data. Tables are stored in HDFS. Partition can be stored in the directory of table as a sub directory. Bucket is stored as a file either within partition's directory or in table's directory. Hive will convert the queries into MapReduce job and these are executed on interpreter. For processing static data, for analyzing large dataset and when there is necessity to use queries instead of scripts Hive is well suited [4].

IV. ANALYSIS OF DATA

The analysis involves following steps.

1. Rating of action movies on user gender.
2. Rating of adventure movies on user gender.
3. Highest rated movie in a year.
4. Maximum FB likes in a year.

For analyzing dataset, we have created the views of each table. The purpose of creating the view is to minimize the execution time. For example if there is a table with 1000 columns and we want to access 10 columns from that. In this case a view of those 10 columns is created. Instead of scanning the entire table for 10 columns, query can be executed on the view to minimize the execution time. In our analysis these views are then joined on the condition to get the result.

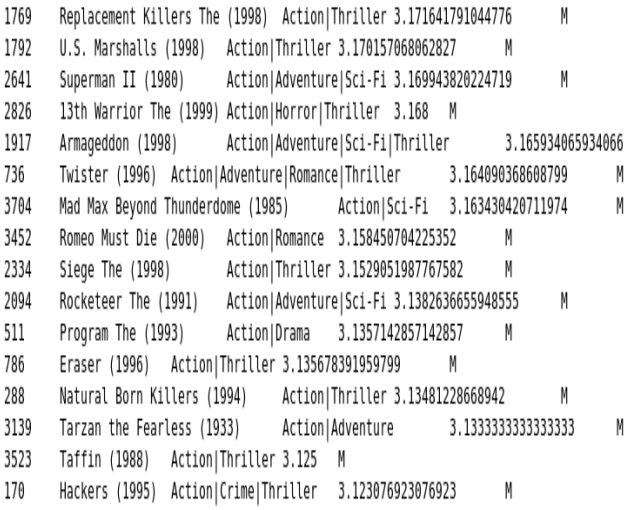
The figure 2 shows the result of average rating given by the female gender for each action category movie. We have used three views: Movie_Action which has the columns movie_id, title, genre as Action from Genre dataset. Female_view has the colums user_id and gender as F from user dataset. Rat_F has movie_id, ratings, gender.

```
3654    Guns of Navarone The (1961)     Action|Drama|War        4.061224489795919       F
1221    Godfather: Part II The (1974)   Action|Crime|Drama      4.04093567251462        F
2194    Untouchables The (1987) Action|Crime|Drama      4.021164021164021       F
110     Braveheart (1995)       Action|Drama|War        4.016483516483516       F
1910    I Went Down (1997)      Action|Comedy|Crime     4.0     F
139     Target (1995)   Action|Drama    4.0     F
3137    Sea Wolves The (1980)   Action|War      4.0     F
2924    Drunken Master (Zui quan) (1979)        Action|Comedy   4.0     F
251     Hunted The (1995)       Action  4.0     F
2823    Spiders The (Die Spinnen 1. Teil: Der Goldene See) (1919)       Action|Drama    4.0     F
2756    Wanted: Dead or Alive (1987)    Action  4.0     F
2737    Assassination (1987)    Action  4.0     F
390     Faster Pussycat! Kill! Kill! (1965)     Action|Comedy|Drama     4.0     F
1832    Heaven's Burning (1997) Action|Drama    4.0     F
1434    Stranger The (1994)     Action  4.0     F
624     Condition Red (1995)    Action|Drama|Thriller   4.0     F
1287    Ben-Hur (1959)  Action|Adventure|Drama  3.9765625       F
1277    Cyrano de Bergerac (1990)       Action|Drama|Romance    3.948905109489051       F
1387    Jaws (1975)     Action|Horror   3.946875        F
```

Figure 2: Rating of action movies by female gender.

978-1-5386-2044-1/17/$31.00 ©2017 IEEE

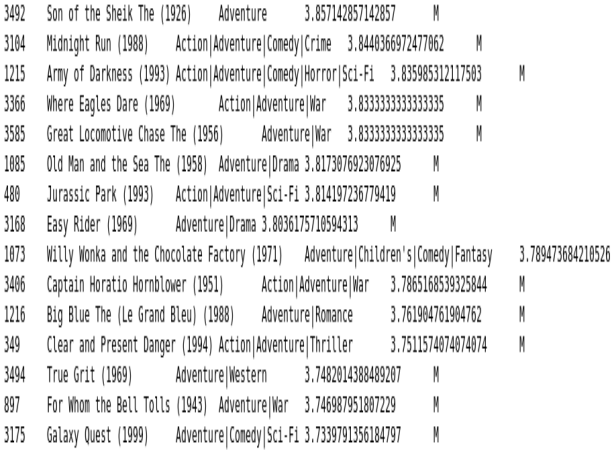
The figure 3 shows the result of average rating given by the male gender for each action category movie. Three views have been created: Movie_Action view, Male_view which has the colums user_id and gender as M from user dataset. Rat_M has movie_id, ratings, gender.



1769	Replacement Killers The (1998)	Action|Thriller	3.171641791044776	M
1792	U.S. Marshalls (1998)	Action|Thriller	3.170157068062827	M
2641	Superman II (1980)	Action|Adventure|Sci-Fi	3.169943820224719	M
2826	13th Warrior The (1999)	Action|Horror|Thriller	3.168	M
1917	Armageddon (1998)	Action|Adventure|Sci-Fi|Thriller	3.1659340659340666	M
736	Twister (1996)	Action|Adventure|Romance|Thriller	3.164090368608799	M
3704	Mad Max Beyond Thunderdome (1985)	Action|Sci-Fi	3.163430420711974	M
3452	Romeo Must Die (2000)	Action|Romance	3.158450704225352	M
2334	Siege The (1998)	Action|Thriller	3.1529051987767582	M
2094	Rocketeer The (1991)	Action|Adventure|Sci-Fi	3.1382636655948555	M
511	Program The (1993)	Action|Drama	3.1357142857142857	M
786	Eraser (1996)	Action|Thriller	3.135678391959799	M
288	Natural Born Killers (1994)	Action|Thriller	3.13481228668942	M
3139	Tarzan the Fearless (1933)	Action|Adventure	3.1333333333333333	M
3523	Taffin (1988)	Action|Thriller	3.125	M
170	Hackers (1995)	Action|Crime|Thriller	3.123076923076923	M

Figure 3: Rating of action movies by male gender.

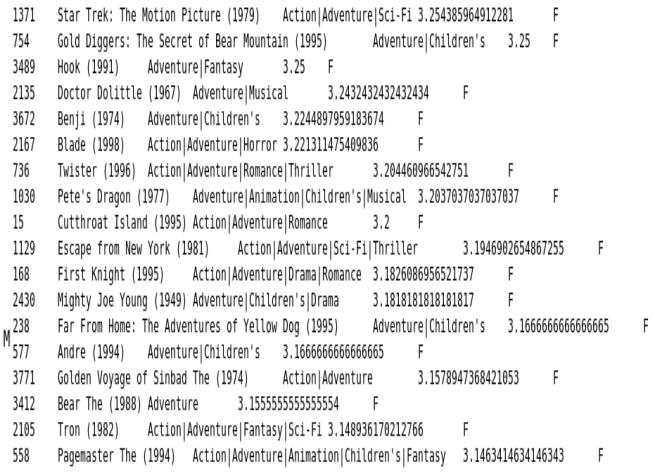
The figure 4 shows the result of average rating given by the female gender for each adventure category movie. Three views are: Movie_Adventure which has the columns movie_id, title, genre as Adventure from Genre dataset. Female_view and Rat_F view.



3492	Son of the Sheik The (1926)	Adventure	3.857142857142857	M
3104	Midnight Run (1988)	Action|Adventure|Comedy|Crime	3.844036972477062	M
1215	Army of Darkness (1993)	Action|Adventure|Comedy|Horror|Sci-Fi	3.835985312117503	M
3366	Where Eagles Dare (1969)	Action|Adventure|War	3.8333333333333335	M
3585	Great Locomotive Chase The (1956)	Adventure|War	3.8333333333333335	M
1085	Old Man and the Sea The (1958)	Adventure|Drama	3.8173076923076925	M
480	Jurassic Park (1993)	Action|Adventure|Sci-Fi	3.814197236779419	M
3168	Easy Rider (1969)	Adventure|Drama	3.8036175710594313	M
1073	Willy Wonka and the Chocolate Factory (1971)	Adventure|Children's|Comedy|Fantasy	3.7894736842105263	M
3406	Captain Horatio Hornblower (1951)	Action|Adventure|War	3.7865168539325844	M
1216	Big Blue The (Le Grand Bleu) (1988)	Adventure|Romance	3.761904761904762	M
349	Clear and Present Danger (1994)	Action|Adventure|Thriller	3.7511574074074074	M
3494	True Grit (1969)	Adventure|Western	3.7482014388489207	M
897	For Whom the Bell Tolls (1943)	Adventure|War	3.7469879518072289	M
3175	Galaxy Quest (1999)	Adventure|Comedy|Sci-Fi	3.7339791356184797	M

Figure 4: Rating of adve nture movies by male gender.

The figure 5 shows the result of average rating given by the male gender for each adventure category movie. Views used are: Movie_Adventure , Male_view and Rat_M view.



1371	Star Trek: The Motion Picture (1979)	Action|Adventure|Sci-Fi	3.254385964912281	F
754	Gold Diggers: The Secret of Bear Mountain (1995)	Adventure|Children's	3.25	F
3489	Hook (1991)	Adventure|Fantasy	3.25	F
2135	Doctor Dolittle (1967)	Adventure|Musical	3.2432432432432434	F
3672	Benji (1974)	Adventure|Children's	3.2244897959183674	F
2167	Blade (1998)	Action|Adventure|Horror	3.221311475409836	F
736	Twister (1996)	Action|Adventure|Romance|Thriller	3.204460966542751	F
1030	Pete's Dragon (1977)	Adventure|Animation|Children's|Musical	3.2037037037037037	F
15	Cutthroat Island (1995)	Action|Adventure|Romance	3.2	F
1129	Escape from New York (1981)	Action|Adventure|Sci-Fi|Thriller	3.1946902654867255	F
168	First Knight (1995)	Action|Adventure|Drama|Romance	3.1826086956521737	F
2430	Mighty Joe Young (1949)	Adventure|Children's|Drama	3.1818181818181817	F
238	Far From Home: The Adventures of Yellow Dog (1995)	Adventure|Children's	3.1666666666666665	F
577	Andre (1994)	Adventure|Children's	3.1666666666666665	F
3771	Golden Voyage of Sinbad The (1974)	Action|Adventure	3.1578947368421053	F
3412	Bear The (1988)	Adventure	3.1555555555555554	F
2105	Tron (1982)	Action|Adventure|Fantasy|Sci-Fi	3.148936170212766	F
558	Pagemaster The (1994)	Action|Adventure|Animation|Children's|Fantasy	3.146341463146343	F

Figure 5: Rating of adventure movies by female gender.

The figure 6 shows the result of highest rated movie in a particular year. For this two views have been created. Movie_Rat with year, title and imdb_rating from Movie dataset. Movie_Rat_Max with year and maximum imdb_rating for that year.

1938	8.0	You Can't Take It with You
1939	8.2	Gone with the Wind
1939	8.2	Mr. Smith Goes to Washington
1940	8.2	Rebecca
1941	7.8	How Green Was My Valley
1942	8.6	Casablanca
1943	7.0	A Guy Named Joe
1944	6.5	Bathing Beauty
1945	8.0	The Lost Weekend
1946	8.6	It's a Wonderful Life
1947	7.7	The Lady from Shanghai
1948	7.8	Red River
1949	7.4	She Wore a Yellow Ribbon
1950	7.0	Annie Get Your Gun
1951	8.0	A Streetcar Named Desire
1952	8.3	Singin' in the Rain
1953	7.8	From Here to Eternity
1954	8.7	Seven Samurai
1955	8.1	Ordet
1956	7.4	Moby Dick
1957	8.9	12 Angry Men
1958	8.1	Cat on a Hot Tin Roof
1959	8.3	Some Like It Hot
1960	8.5	Psycho
1961	8.3	Judgment at Nuremberg
1962	8.4	Lawrence of Arabia
1962	8.4	To Kill a Mockingbird

Figure 6: Highest rated movie in a year.

The figure 7 shows the result of which movie has got maximum Facebook likes in a particular year. The views created here are Movie_FB with year, title and fb_likes from Movie dataset. Movie_FB_Max with year and maximum fb_likes for that year.

```
1967    3000    Point Blank
1968    24000   2001: A Space Odyssey
1969    548     Sweet Charity
1970    690     Waterloo
1971    819     Escape from the Planet of the Apes
1972    43000   The Godfather
1973    18000   The Exorcist
1974    14000   Young Frankenstein
1974    14000   The Godfather: Part II
1975    32000   One Flew Over the Cuckoo's Nest
1976    35000   Taxi Driver
1977    33000   Star Wars: Episode IV - A New Hope
1978    13000   Grease
1979    23000   Alien
1980    37000   The Shining
1981    16000   Raiders of the Lost Ark
1982    34000   Blade Runner
1982    34000   E.T. the Extra-Terrestrial
1983    19000   Scarface
```

Figure 7: Maximum FB likes in a year.

V. PERFORMANCE ANALYSIS

In this paper Movie dataset is analysed using HiveQL. We executed Hive queries and compared its response time with SQL response time. MySQL is well suited for smaller datasets. As the dataset size increases, it might take longer time to process it and it may also require more memory for processing. But in the case of Hive, it is best suited for larger dataset. From the experiment we observed that the response time is more in MySQL. Therefore it is more efficient to use Hive for large dataset. We have also observed that by creating views of the table it is possible to reduce the execution time The result of the experiment is as shown in the table 5.

Description	Hive Response time (in secs)	SQL response time (in secs)
Rating of action movies by female gender	165.455	209.743
Rating of action movies by male gender	150.987	200.7778
Rating of adventure movies by female gender	111.382	203.4589
Rating of adventure movies by male gender	103.018	198.894
Highest rated movie	38.245	112.563
Maximum fb likes in a year	37.985	109.549

Table 5: Analysis Result

The figure 8 represent the result in bar graph where it clearly shows the execution time difference between Hive and MySQL.

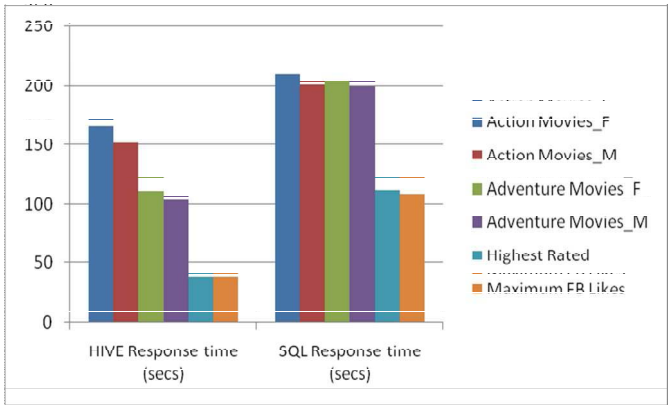


Figure 8: Response time analysis

VI. CONCLUSION

Hive is mainly used to process large amount of data. It is faster than SQL on low-cost machine. Hive performance is poor on smaller dataset but as the data size increases its processing time decreases. SQL is efficient and more robust for smaller data.

In order to find the relationship between year, movie title, imdb rating facebook likes, genre the dataset is analysed using HiveQL and its performance is compared with SQL performance. The result states that Hive performs better than SQL for larger dataset. For the future work we can consider the director, actor to predict which director has directed the best rated movie and which actor has got role in that. User ratings can be determined for other category.

VII. REFERENCES

[1] Yojna Arora, Dr Dinesh Goyal, "Big Data: A Review of Analytics Methods & Techniques", IEEE, 2016 2nd International Conference on Contemporary Computing and Informmatics.

[2] Rahul Kumar Chawda, Dr Ghanshyam Thankur, "Big Data and Advanced Analytics Tools", IEEE, 2016 Symposium on Colossal Data Analysis and Networking.

[3] Brijesh Dhyani, Anurag Barthwal, "Big Data Analytics using Hadoop", International Journal of Computer Applications (0975-8887) Volume 108-No 12, December 2014.

[4] Ashish Thusoo et al. "Hive- A Petabyte Scale Data Warehouse Using Hadoop", 978-1-4244-5446-4, 2010 IEEE.

[5] http://hadoop.apache.org/

[6] https://hive.apache.org/

[7] Ammar Fuad, Alva Erwin, Henru Purnomo Ipung, " Processing Performance on Apache Pig, Apache Hive and MySQL Cluster", IEEE, 2014 International Conference on Information, Communication Technology and System.

[8] Karan Sachdeva et al., "Comparison of Data Processing Tools in Hadoop", IEEE, 2016 International Conference on Electrical, Electronics,

978-1-5386-2044-1/17/$31.00 ©2017 IEEE

Communication, Computer and Optimization Techniques.

[9] Ali Usman Abdullahi, Rohiza Ahmad, Nordin M Zakaria, "Big Data: Performance Profiling of Meteorological and Oceanographic Data on Hive", IEEE, 2016 3rd International Conference On Computer And Information Sciences.

[10] Aditya Bhardwaj et al., "Big Data Emerging Technologies: A CAseStudy with Analyzing Twitter Data using Apache Hive", IEEE, 2015 RAECS UIET Panjab University Chandigarh.

[11] Varun Garg, "Optimization of Multiple Queries for Big Data with Apache Hadoop/Hive", IEEE, 2015 International Conference on Computational Intelligence and Communication Networks.

[12] Abdeltawab M. Hendawi et al., "Hobbits: Hadoop and Hive Based Internet Traffic Analysis", 2016 IEEE International Conference on Big Data.

[13] Taoying Liu, Jing Liu, Hong Liu, Wei Li, "A Performance Evaluation of Hive for Scientific Data Management", 2013 IEEE International Conference on Big Data.

[14] S K Pushpa, Manjunath T N, Srividhya, "Analysis of Airport Data using Hadoop-Hive: A Case Study", International Journal of Computer Applications (0975 – 8887) National Conference on "Recent Trends in Information Technology" (NCRTIT-2016).

[15] Dharaben Patel, Xiaohong Yuan, Kaushik Roy, Aakiel Abernathy, "Analyzing Network Traffic Data Using Hive Queries", 978-1-5386-1539-3/17/2017 IEEE.

[16] Amit Sangroya, Reha Singhal, "Performance Assurance Model for HiveQL on Large Data Volume", 2015 IEEE 22nd International Conference on High Performance Computing Workshops.

[17] Weijia Xu et al., "Supporting Large Scale Connected Vehicle Data Analysis using Hive", 2016 IEEE International Conference on Big Data.

[18] Alexander C.Shulyak, Lizy K. John, "Identifying Performance Bottlenecks in Hive: Use of Processor Counters", 978-1-4673-9005-7/16/2016 IEEE International Conference on Big Data.

[19] Shujia Zhou et al., "Visualization and Diagnosis of Earth Science Data through Hadoop and Spark", 978-1-4673-9005-7/16/2016 IEEE International Conference on Big Data.