ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

A Comparative Analysis of Apache hive based on MapReduce and Impala based on distributed query engine

Prasad Mitkari[1], Rutuja Banswal [2], Akshata Sirsat[3], Prof. Daivashala Deshmukh[4]

B.Tech Student, Department of Computer Science & Engineering, Dr. BAMU, Maharashtra Institute of Technology, Aurangabad (MS), India

B.Tech Student, Department of Computer Science & Engineering, Dr. BAMU, Maharashtra Institute of Technology, Aurangabad (MS), India

B.Tech Student, Department of Computer Science & Engineering, Dr. BAMU, Maharashtra Institute of Technology, Aurangabad (MS), India

Assistant Professor, Department of Computer Science & Engineering, Dr. BAMU, Maharashtra Institute of Technology, Aurangabad (MS), India

ABSTRACT: Big Data realm moves around 5 Vs- volume, velocity, variety, value and veracity. Analyzing and Storing huge amount of data available in different formats which is increasing with huge velocity to gain values out and it is itself a big deal. Faster and Quick query in the Big Data is very important for getting the valuable information to improve the system performance. To get this goal, number of research institutions and internet companies develop some tools which are respectively Hive (based on MapReduce), Impala (based on Distributed query engine). In this paper, we compare these two-type of query tools and we find the time taken by Impala and Hive to execute the same query. In this paper we contrast the relative merits of two technologies called Apache Hive and Cloudera's Impala. In efficiency and performance, total time taken by both for their execution and their requirements. Basically we focus on performance analysis of both Apache Hive and Impala using Cloudera Distribution Including Apache Hadoop (CDH) and movielens datasets.

KEYWORDS: Hadoop, Hive, Impala, Framework, MapReduce, Analysis

I. INTRODUCTION

Big data is a large amount of drastically increasing real world data. Data is increasing every minute, every second in a day. This large amount of data is used to derive knowledge from the raw data, with this in mind, big data is the computing strategy and technology that are used to handle large datasets. There are five aspects which define the big data such as volume, velocity, variety, veracity, and value. Big Data is usually difficult to store and process using traditional data management systems. Apache Software Foundation introduced processing tools to solve processing challenges, which are used to solve big data related problems and used to derive the patterns and trends from data.

HIVE: Apache Hive is a tool, which built on top of Apache Hadoop framework. At initial phase, it was being developed by Facebook. Later it was given to Apache Software Foundation and developed further as open-source software called as Apache Hive. Hive is mainly used for processing unstructured data. This processing is done in three main steps summarization, query, and analysis. Hive is primarily used for data mining purpose. The interaction with the Hive is done by SQL like query language called HiveQL. Due to its HiveQL, it is very familiar and user-friendly, fast, scalable, and extensible.

DOI: 10.15680/IJIRCCE.2017. 05070101

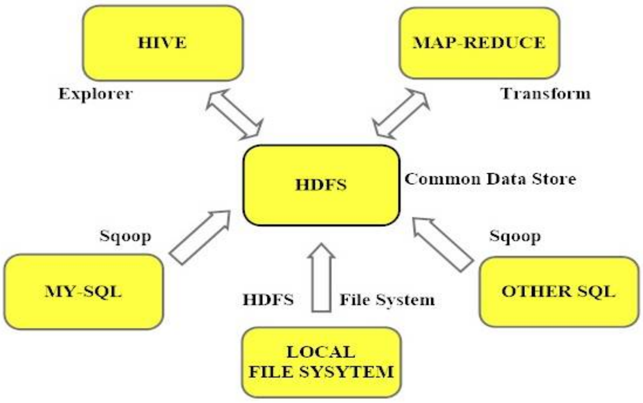


Figure 1: Hive architecture

IMPALA: Impala uses Massive Parallel Processing (MPP) SQL Query Engine and it implements a distributed architecture based on daemon processes. Impala is the best choice when the requirement is a quick result in real time. The intermediate result is stored in In-Memory. Thus, query execution is very fast when compared to other tools.

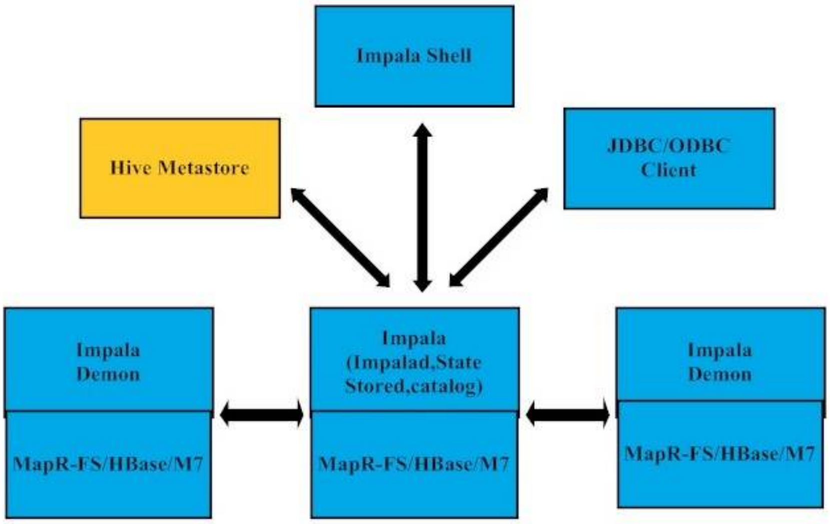


Figure 2: Impala architecture

II. WORKFLOW OF HIVE AND IMPALA

Cloudera’s Impala’s performance is more efficient in terms of execution time as well as the complexity of queries. Apache Hive is Infrastructure developed on Hadoop Framework for analyzing and processing data. Basically, Hive is the front end to parse the SQL queries, designs logical plans that execute in the background by MapReduce and Tez this takes comparatively more time than Impala.

Cloudera’s Impala does not require data to be moved or transformed as it uses Hive's megastore, it can query Hive's tables directly. Impala does not use MapReduce to execute the query. The Impala daemon executes the process on each node to plan queries and, coordinates between them and query execution engine. As Impala uses parallel processing it responds to query very fast using massively parallel processing. Each node accepts query and planner maps these requests to parallel fragment after that the coordinator starts the execution on the name node of the cluster. The Network systems are highly multithreaded thus each node runs efficiently in a cluster.

Apache Hive generates queries at compile-time whereas Impala does runtime code generation for big loops. MapReduce takes more time for running at full capacity. So, in Hive's query suffers cold start problem on the other hand Impala daemon processes are started at boot time. Hive is fault tolerant whereas Impala does not support fault tolerance if query execution fails Impala needs to start again the execution.

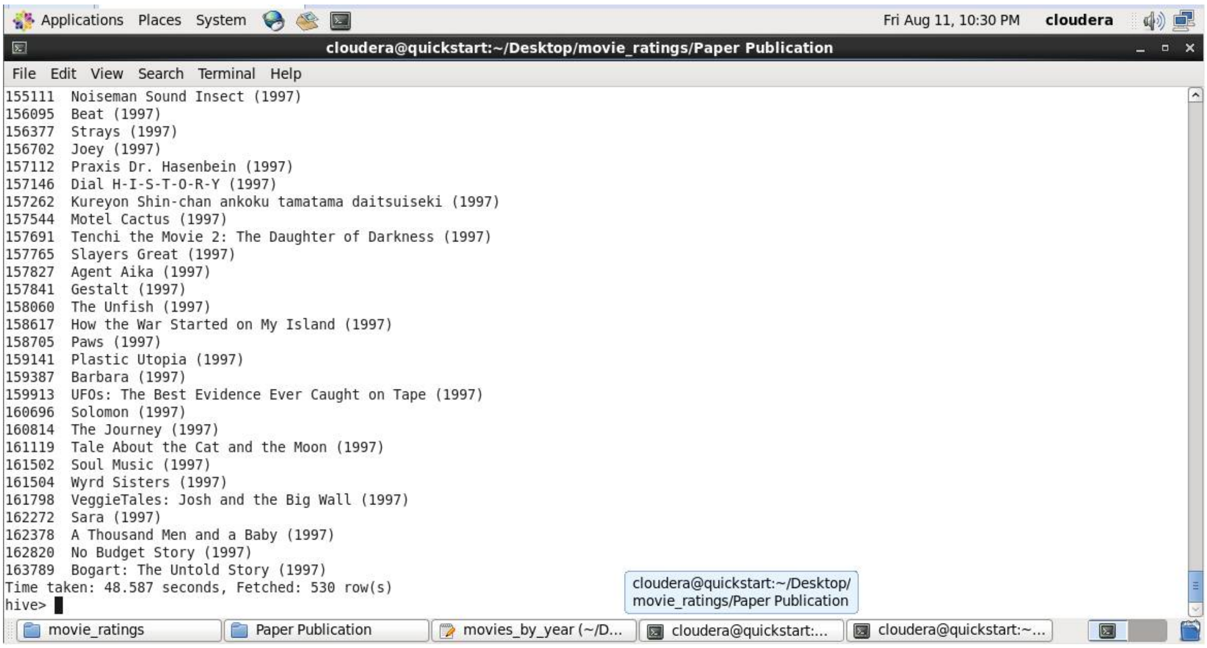
For performance analysis of Apache Hive and Cloudera's Impala, we have used movielens dataset. These are some queries we have executed on movielens dataset. Here we have tables namely ratings and movies.

Attributes List:
movies :- movie_id: int, name: chararray, genres: chararray.
ratings :- user_id: int, movie_id: int, ratings: float, timestamp: int.

Case 1: Movies in Particular (1997) year :	

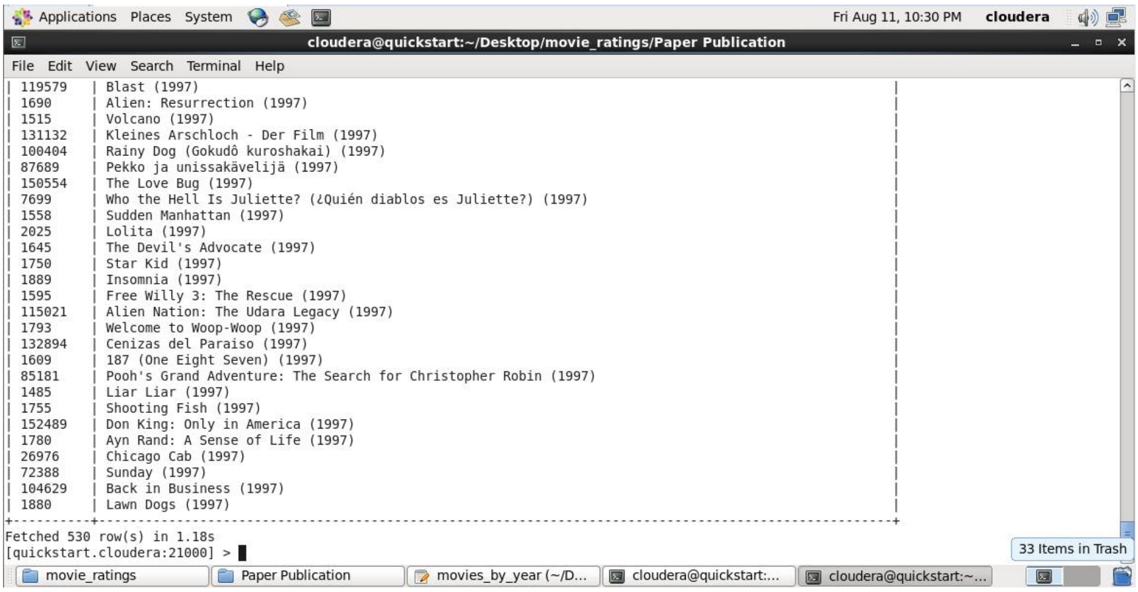
The query is to find movie_id and its name released in 1997 from movies table.	**Query:**
	SELECT DISTINCT movie_id, name
	>FROM movies
	>WHERE name LIKE '%1997%';

HIVE:


Applications Places System
Fri Aug 11, 10:30 PM cloudera
cloudera@quickstart:~/Desktop/movie_ratings/Paper Publication
File Edit View Search Terminal Help
cloudera@quickstart:~/Desktop/
movie_ratings/Paper Publication
movie_ratings
Paper Publication
movies_by_year (~/D...
cloudera@quickstart:...
cloudera@quickstart:~...

```
155111  Noiseman Sound Insect (1997)
156095  Beat (1997)
156377  Strays (1997)
156702  Joey (1997)
157112  Praxis Dr. Hasenbein (1997)
157146  Dial H-I-S-T-O-R-Y (1997)
157262  Kureyon Shin-chan ankoku tamatama daitsuiseki (1997)
157544  Motel Cactus (1997)
157691  Tenchi the Movie 2: The Daughter of Darkness (1997)
157765  Slayers Great (1997)
157827  Agent Aika (1997)
157841  Gestalt (1997)
158060  The Unfish (1997)
158617  How the War Started on My Island (1997)
158705  Paws (1997)
159141  Plastic Utopia (1997)
159387  Barbara (1997)
159913  UFOs: The Best Evidence Ever Caught on Tape (1997)
160696  Solomon (1997)
160814  The Journey (1997)
161119  Tale About the Cat and the Moon (1997)
161502  Soul Music (1997)
161504  Wyrd Sisters (1997)
161798  VeggieTales: Josh and the Big Wall (1997)
162272  Sara (1997)
162378  A Thousand Men and a Baby (1997)
162820  No Budget Story (1997)
163789  Bogart: The Untold Story (1997)
Time taken: 48.587 seconds, Fetched: 530 row(s)
hive>
```

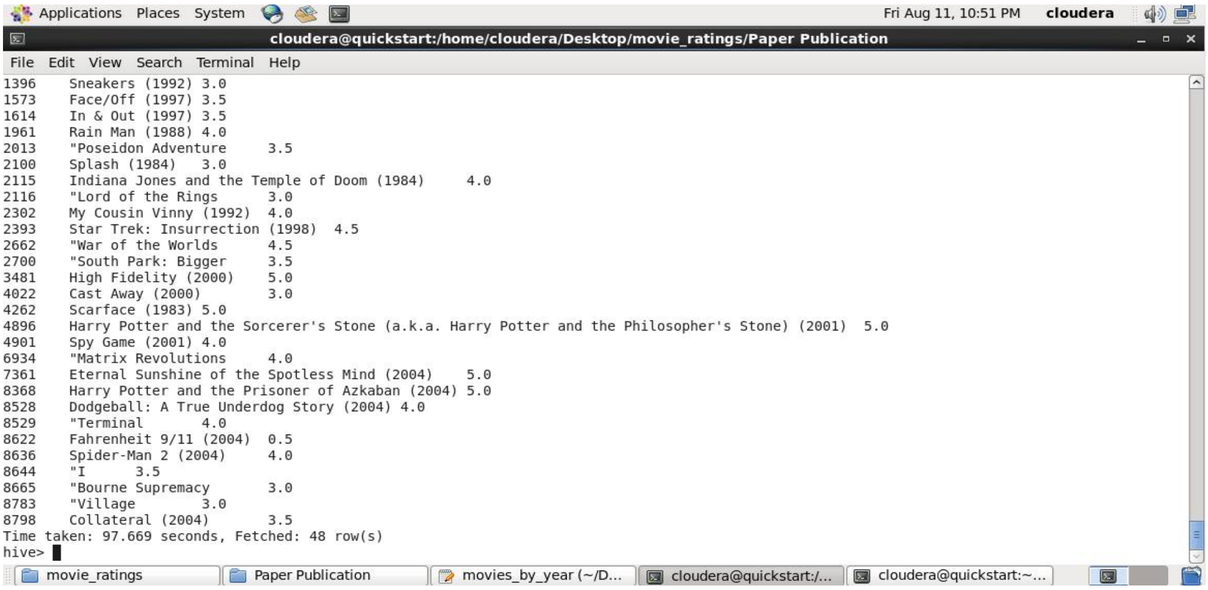
IMPALA:



```
| 119579   | Blast (1997)                                                              |
| 1690     | Alien: Resurrection (1997)                                                |
| 1515     | Volcano (1997)                                                            |
| 131132   | Kleines Arschloch - Der Film (1997)                                       |
| 100404   | Rainy Dog (Gokudô kuroshakai) (1997)                                      |
| 87689    | Pekko ja unissakävelijä (1997)                                            |
| 150554   | The Love Bug (1997)                                                       |
| 7699     | Who the Hell Is Juliette? (¿Quién diablos es Juliette?) (1997)            |
| 1558     | Sudden Manhattan (1997)                                                   |
| 2025     | Lolita (1997)                                                             |
| 1645     | The Devil's Advocate (1997)                                               |
| 1750     | Star Kid (1997)                                                           |
| 1889     | Insomnia (1997)                                                           |
| 1595     | Free Willy 3: The Rescue (1997)                                           |
| 115021   | Alien Nation: The Udara Legacy (1997)                                     |
| 1793     | Welcome to Woop-Woop (1997)                                               |
| 132894   | Cenizas del Paraiso (1997)                                                |
| 1609     | 187 (One Eight Seven) (1997)                                              |
| 85181    | Pooh's Grand Adventure: The Search for Christopher Robin (1997)           |
| 1485     | Liar Liar (1997)                                                          |
| 1755     | Shooting Fish (1997)                                                      |
| 152489   | Don King: Only in America (1997)                                          |
| 1780     | Ayn Rand: A Sense of Life (1997)                                          |
| 26976    | Chicago Cab (1997)                                                        |
| 72388    | Sunday (1997)                                                             |
| 104629   | Back in Business (1997)                                                   |
| 1880     | Lawn Dogs (1997)                                                          |
+----------+---------------------------------------------------------------------------+
Fetched 530 row(s) in 1.18s
[quickstart.cloudera:21000] >
```

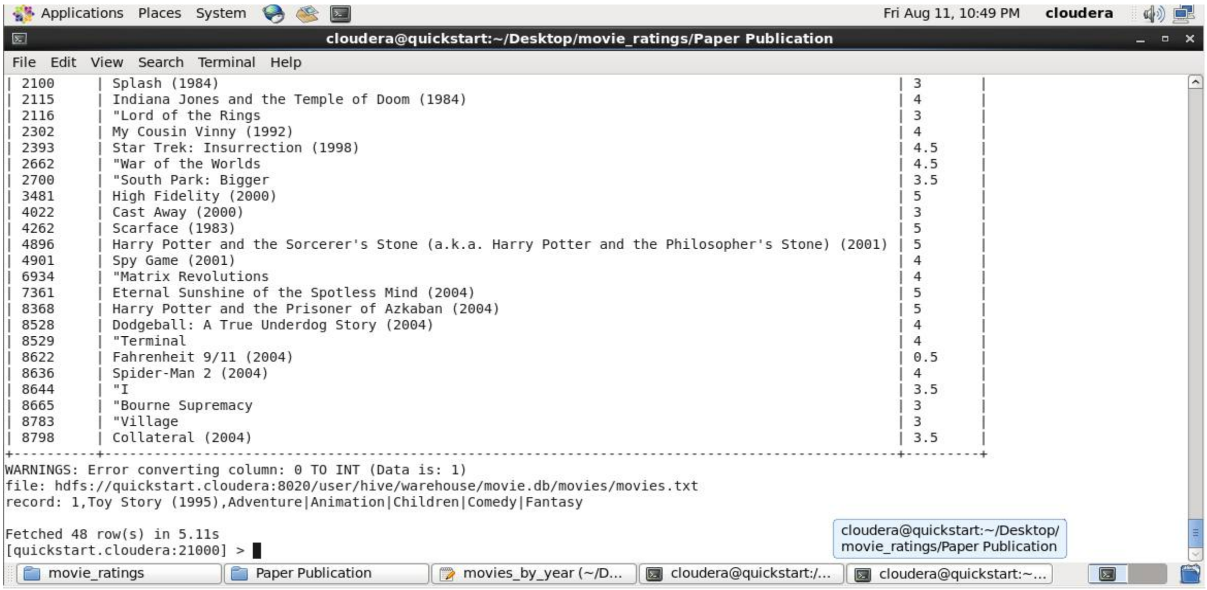
Case 2: Particular User Ratings.	
The query joins two table ratings and movies on the movie_id field to find movie id, name and its rating for a specific user.	**Query:** SELECT r.movie_id, name, ratings > FROM movie.ratings r > JOIN movie.movies m > ON (r.movie_id=m.movie_id) > WHERE user_id=10564;

HIVE:


Applications Places System
Fri Aug 11, 10:51 PM
cloudera
cloudera@quickstart:/home/cloudera/Desktop/movie_ratings/Paper Publication
File Edit View Search Terminal Help
movie_ratings
Paper Publication
movies_by_year (~/D...
cloudera@quickstart:/...
cloudera@quickstart:~...

```
1396    Sneakers (1992) 3.0
1573    Face/Off (1997) 3.5
1614    In & Out (1997) 3.5
1961    Rain Man (1988) 4.0
2013    "Poseidon Adventure     3.5
2100    Splash (1984)   3.0
2115    Indiana Jones and the Temple of Doom (1984)     4.0
2116    "Lord of the Rings      3.0
2302    My Cousin Vinny (1992)  4.0
2393    Star Trek: Insurrection (1998)  4.5
2662    "War of the Worlds      4.5
2700    "South Park: Bigger     3.5
3481    High Fidelity (2000)    5.0
4022    Cast Away (2000)        3.0
4262    Scarface (1983) 5.0
4896    Harry Potter and the Sorcerer's Stone (a.k.a. Harry Potter and the Philosopher's Stone) (2001)  5.0
4901    Spy Game (2001) 4.0
6934    "Matrix Revolutions     4.0
7361    Eternal Sunshine of the Spotless Mind (2004)    5.0
8368    Harry Potter and the Prisoner of Azkaban (2004) 5.0
8528    Dodgeball: A True Underdog Story (2004) 4.0
8529    "Terminal       4.0
8622    Fahrenheit 9/11 (2004)  0.5
8636    Spider-Man 2 (2004)     4.0
8644    "I      3.5
8665    "Bourne Supremacy       3.0
8783    "Village        3.0
8798    Collateral (2004)       3.5
Time taken: 97.669 seconds, Fetched: 48 row(s)
hive>
```

IMPALA:


Applications Places System
Fri Aug 11, 10:49 PM cloudera
cloudera@quickstart:~/Desktop/movie_ratings/Paper Publication
File Edit View Search Terminal Help
cloudera@quickstart:~/Desktop/ movie_ratings/Paper Publication
movie_ratings
Paper Publication
movies_by_year (~/D...
cloudera@quickstart:/...
cloudera@quickstart:~...

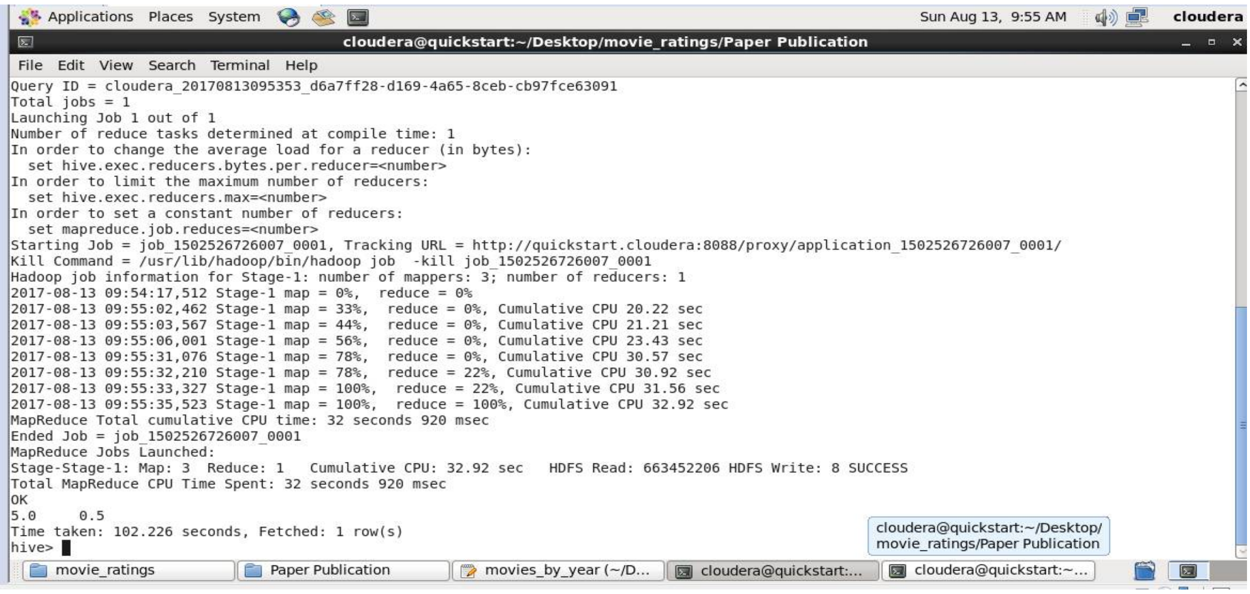
```
| 2100     | Splash (1984)                                                                      | 3       |
| 2115     | Indiana Jones and the Temple of Doom (1984)                                        | 4       |
| 2116     | "Lord of the Rings                                                                 | 3       |
| 2302     | My Cousin Vinny (1992)                                                             | 4       |
| 2393     | Star Trek: Insurrection (1998)                                                     | 4.5     |
| 2662     | "War of the Worlds                                                                 | 4.5     |
| 2700     | "South Park: Bigger                                                                | 3.5     |
| 3481     | High Fidelity (2000)                                                               | 5       |
| 4022     | Cast Away (2000)                                                                   | 3       |
| 4262     | Scarface (1983)                                                                    | 5       |
| 4896     | Harry Potter and the Sorcerer's Stone (a.k.a. Harry Potter and the Philosopher's Stone) (2001) | 5       |
| 4901     | Spy Game (2001)                                                                    | 4       |
| 6934     | "Matrix Revolutions                                                                | 4       |
| 7361     | Eternal Sunshine of the Spotless Mind (2004)                                       | 5       |
| 8368     | Harry Potter and the Prisoner of Azkaban (2004)                                    | 5       |
| 8528     | Dodgeball: A True Underdog Story (2004)                                            | 4       |
| 8529     | "Terminal                                                                          | 4       |
| 8622     | Fahrenheit 9/11 (2004)                                                             | 0.5     |
| 8636     | Spider-Man 2 (2004)                                                                | 4       |
| 8644     | "I                                                                                 | 3.5     |
| 8665     | "Bourne Supremacy                                                                  | 3       |
| 8783     | "Village                                                                           | 3       |
| 8798     | Collateral (2004)                                                                  | 3.5     |
+----------+------------------------------------------------------------------------------------+---------+
WARNINGS: Error converting column: 0 TO INT (Data is: 1)
file: hdfs://quickstart.cloudera:8020/user/hive/warehouse/movie.db/movies/movies.txt
record: 1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy

Fetched 48 row(s) in 5.11s
[quickstart.cloudera:21000] >
```

CASE 3: Finding MAX and MIN ratings from 663452198 records.	

The query finds maximum and minimum rating from entire ratings table which contain 663452198 records.	**Query:**
	SELECT MAX(ratings),MIN(ratings)
	>from ratings;

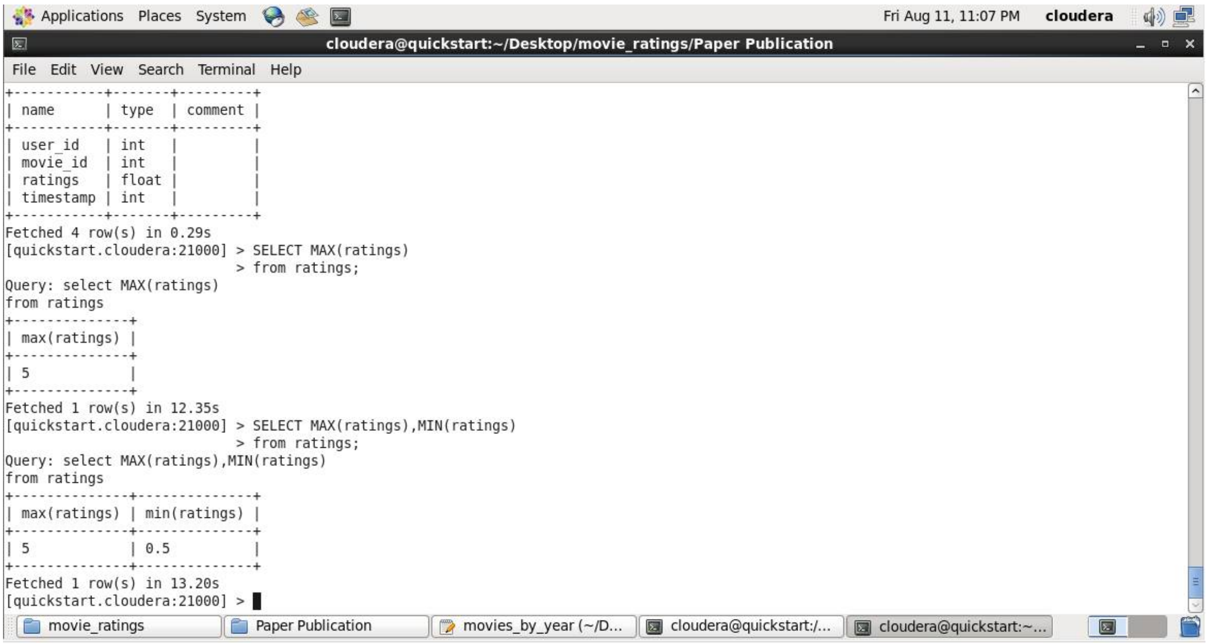
HIVE:


Applications Places System
Sun Aug 13, 9:55 AM cloudera
cloudera@quickstart:~/Desktop/movie_ratings/Paper Publication
File Edit View Search Terminal Help
cloudera@quickstart:~/Desktop/ movie_ratings/Paper Publication
movie_ratings
Paper Publication
movies_by_year (~/D...
cloudera@quickstart:...
cloudera@quickstart:~...

```
Query ID = cloudera_20170813095353_d6a7ff28-d169-4a65-8ceb-cb97fce63091
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1502526726007_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1502526726007_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1502526726007_0001
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 1
2017-08-13 09:54:17,512 Stage-1 map = 0%,  reduce = 0%
2017-08-13 09:55:02,462 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 20.22 sec
2017-08-13 09:55:03,567 Stage-1 map = 44%,  reduce = 0%, Cumulative CPU 21.21 sec
2017-08-13 09:55:06,001 Stage-1 map = 56%,  reduce = 0%, Cumulative CPU 23.43 sec
2017-08-13 09:55:31,076 Stage-1 map = 78%,  reduce = 0%, Cumulative CPU 30.57 sec
2017-08-13 09:55:32,210 Stage-1 map = 78%,  reduce = 22%, Cumulative CPU 30.92 sec
2017-08-13 09:55:33,327 Stage-1 map = 100%,  reduce = 22%, Cumulative CPU 31.56 sec
2017-08-13 09:55:35,523 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 32.92 sec
MapReduce Total cumulative CPU time: 32 seconds 920 msec
Ended Job = job_1502526726007_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 1   Cumulative CPU: 32.92 sec   HDFS Read: 663452206 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 32 seconds 920 msec
OK
5.0     0.5
Time taken: 102.226 seconds, Fetched: 1 row(s)
hive>
```

DOI: 10.15680/IJIRCCE.2017. 05070101

IMPALA:

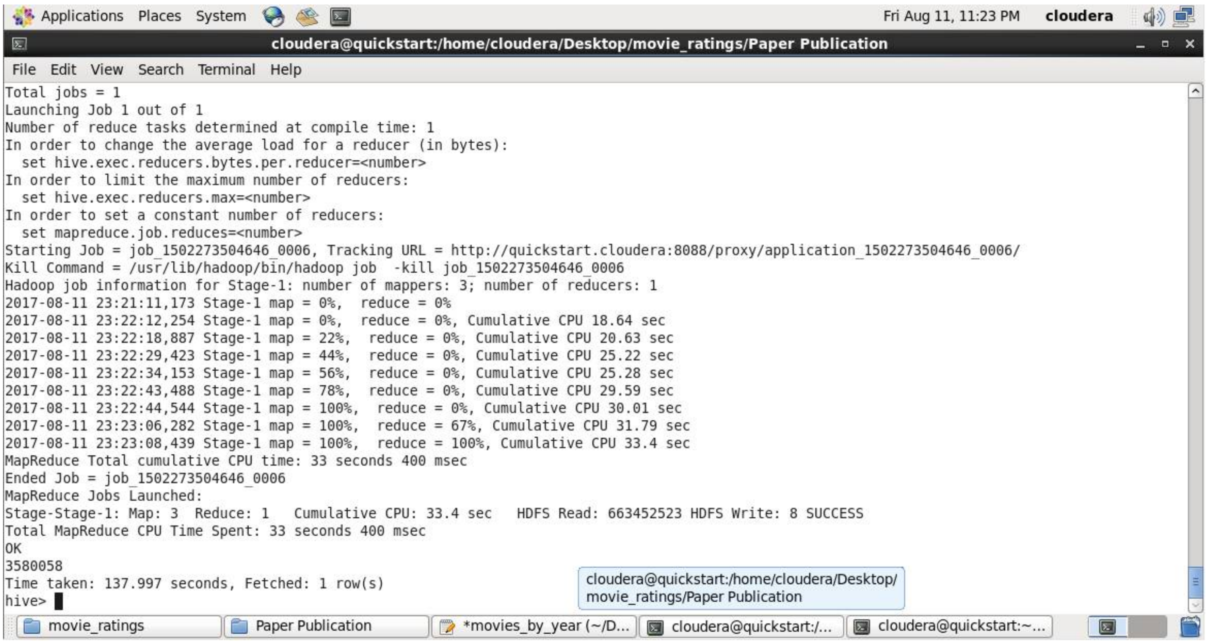


```
+-----------+-------+---------+
| name      | type  | comment |
+-----------+-------+---------+
| user_id   | int   |         |
| movie_id  | int   |         |
| ratings   | float |         |
| timestamp | int   |         |
+-----------+-------+---------+
Fetched 4 row(s) in 0.29s
[quickstart.cloudera:21000] > SELECT MAX(ratings)
                            > from ratings;
Query: select MAX(ratings)
from ratings
+--------------+
| max(ratings) |
+--------------+
| 5            |
+--------------+
Fetched 1 row(s) in 12.35s
[quickstart.cloudera:21000] > SELECT MAX(ratings),MIN(ratings)
                            > from ratings;
Query: select MAX(ratings),MIN(ratings)
from ratings
+--------------+--------------+
| max(ratings) | min(ratings) |
+--------------+--------------+
| 5            | 0.5          |
+--------------+--------------+
Fetched 1 row(s) in 13.20s
[quickstart.cloudera:21000] > █
```

Case 4: Movies by Genres	

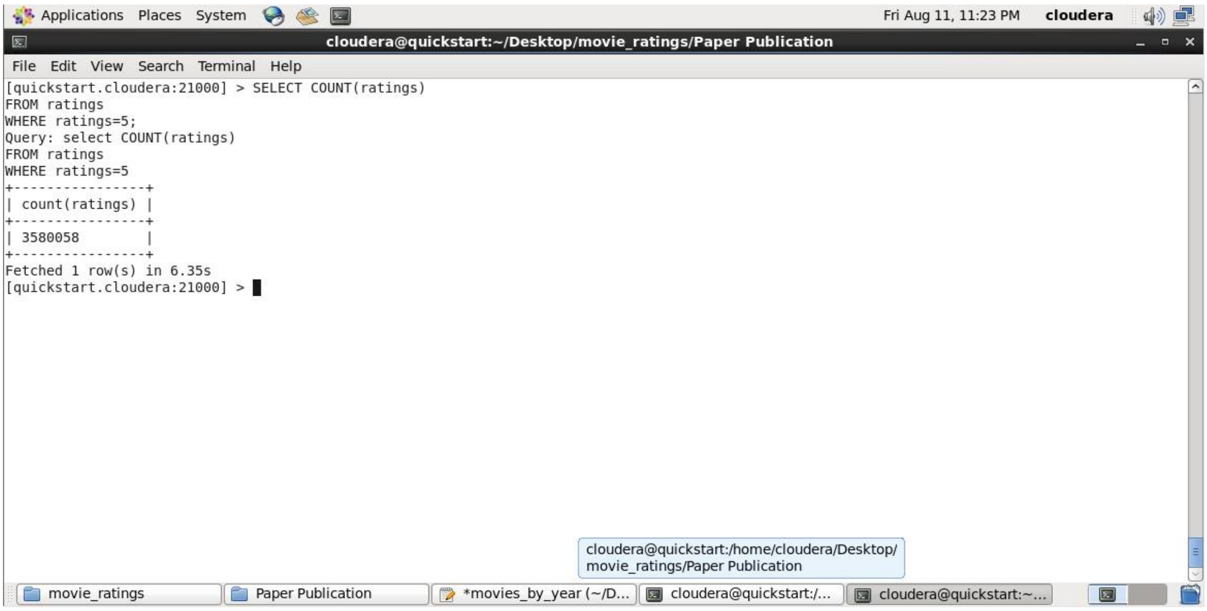
The query is to find the total number of count of the highest rating.	**Query:**
	SELECT COUNT(ratings)
	>FROM ratings
	>WHERE ratings=5;

HIVE:


Applications Places System
Fri Aug 11, 11:23 PM
cloudera
cloudera@quickstart:/home/cloudera/Desktop/movie_ratings/Paper Publication
File Edit View Search Terminal Help
cloudera@quickstart:/home/cloudera/Desktop/
movie_ratings/Paper Publication
movie_ratings
Paper Publication
*movies_by_year (~/D...
cloudera@quickstart:/...
cloudera@quickstart:~...

```
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1502273504646_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1502273504646_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1502273504646_0006
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 1
2017-08-11 23:21:11,173 Stage-1 map = 0%,  reduce = 0%
2017-08-11 23:22:12,254 Stage-1 map = 0%,  reduce = 0%, Cumulative CPU 18.64 sec
2017-08-11 23:22:18,887 Stage-1 map = 22%,  reduce = 0%, Cumulative CPU 20.63 sec
2017-08-11 23:22:29,423 Stage-1 map = 44%,  reduce = 0%, Cumulative CPU 25.22 sec
2017-08-11 23:22:34,153 Stage-1 map = 56%,  reduce = 0%, Cumulative CPU 25.28 sec
2017-08-11 23:22:43,488 Stage-1 map = 78%,  reduce = 0%, Cumulative CPU 29.59 sec
2017-08-11 23:22:44,544 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 30.01 sec
2017-08-11 23:23:06,282 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 31.79 sec
2017-08-11 23:23:08,439 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 33.4 sec
MapReduce Total cumulative CPU time: 33 seconds 400 msec
Ended Job = job_1502273504646_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 1   Cumulative CPU: 33.4 sec   HDFS Read: 663452523 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 33 seconds 400 msec
OK
3580058
Time taken: 137.997 seconds, Fetched: 1 row(s)
hive> █
```

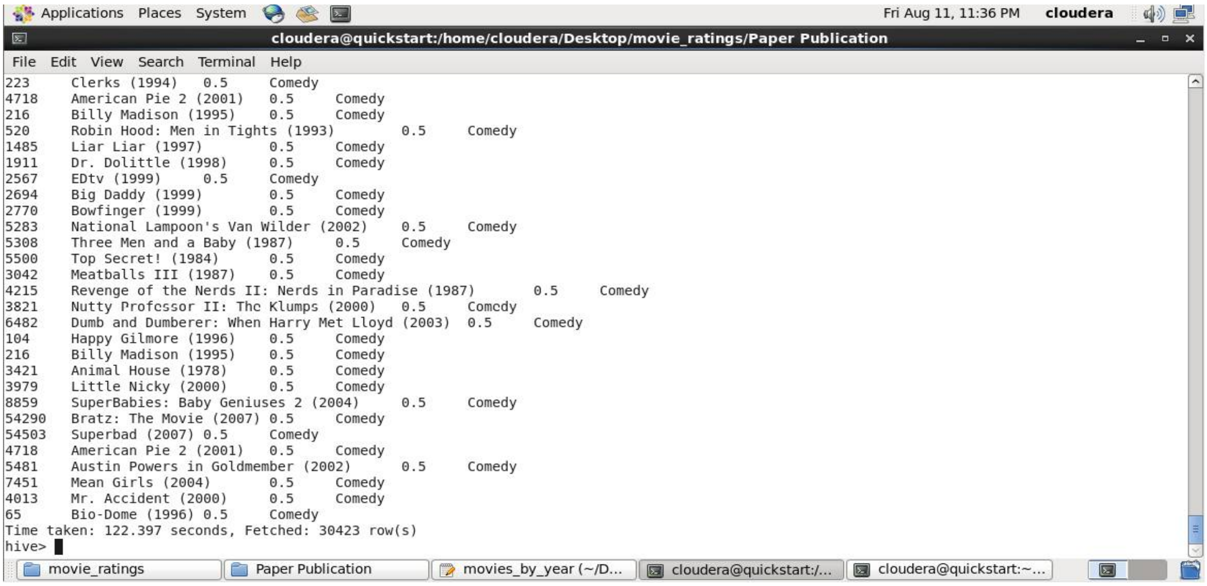
IMPALA:



```
[quickstart.cloudera:21000] > SELECT COUNT(ratings)
FROM ratings
WHERE ratings=5;
Query: select COUNT(ratings)
FROM ratings
WHERE ratings=5
+----------------+
| count(ratings) |
+----------------+
| 3580058        |
+----------------+
Fetched 1 row(s) in 6.35s
[quickstart.cloudera:21000] >
```

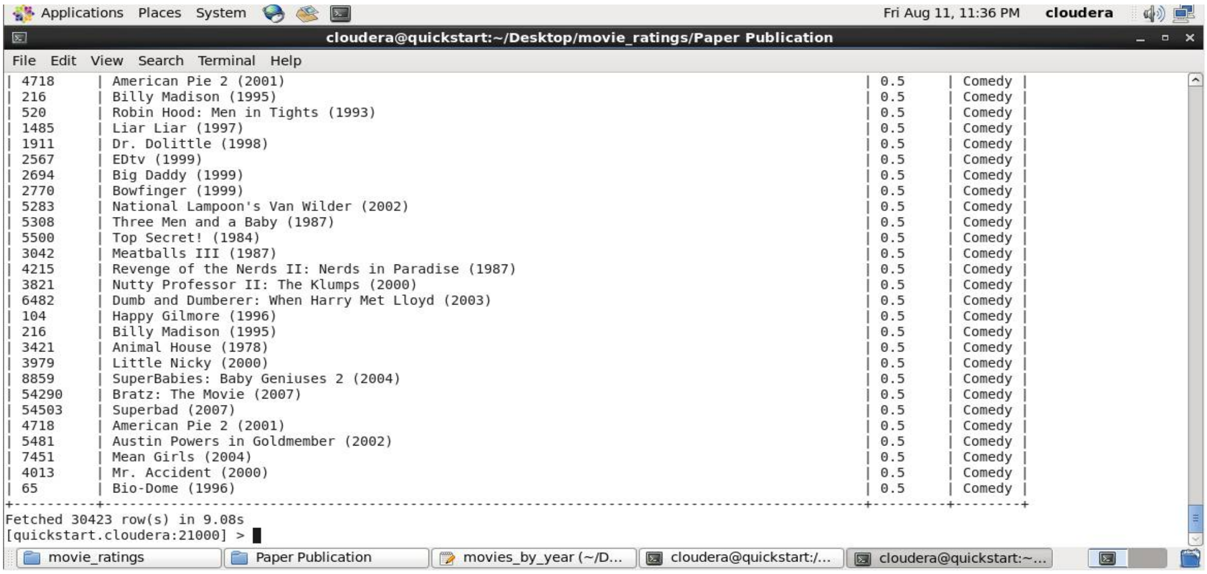
Case 5: Pure Comedy Movies With Lowest Ratings.	
The query is to find movies having genres as ‘Comedy’ with the lowest rating. It joins two tables ratings and movies on the movie_id field that displays movie id, name, ratings and genres.	**Query:** >SELECT r.movie_id, name,ratings,genres >FROM movie.ratings r >JOIN movie.movies m >ON (r.movie_id=m.movie_id) >WHERE genres LIKE 'Comedy' AND ratings=0.5;

HIVE:



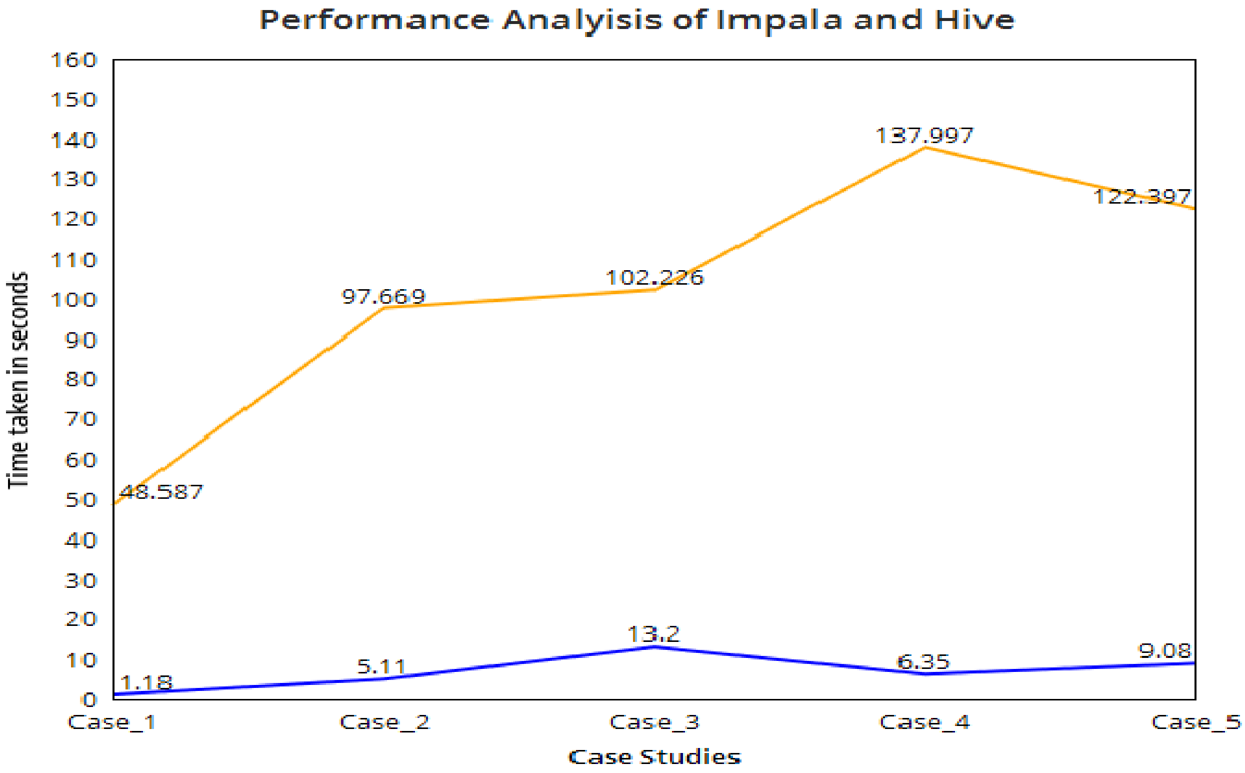
```
223     Clerks (1994)   0.5     Comedy
4718    American Pie 2 (2001)   0.5     Comedy
216     Billy Madison (1995)    0.5     Comedy
520     Robin Hood: Men in Tights (1993)        0.5     Comedy
1485    Liar Liar (1997)        0.5     Comedy
1911    Dr. Dolittle (1998)     0.5     Comedy
2567    EDtv (1999)     0.5     Comedy
2694    Big Daddy (1999)        0.5     Comedy
2770    Bowfinger (1999)        0.5     Comedy
5283    National Lampoon's Van Wilder (2002)    0.5     Comedy
5308    Three Men and a Baby (1987)     0.5     Comedy
5500    Top Secret! (1984)      0.5     Comedy
3042    Meatballs III (1987)    0.5     Comedy
4215    Revenge of the Nerds II: Nerds in Paradise (1987)       0.5     Comedy
3821    Nutty Profcssor II: Thc Klumps (2000)   0.5     Comcdy
6482    Dumb and Dumberer: When Harry Met Lloyd (2003)  0.5     Comedy
104     Happy Gilmore (1996)    0.5     Comedy
216     Billy Madison (1995)    0.5     Comedy
3421    Animal House (1978)     0.5     Comedy
3979    Little Nicky (2000)     0.5     Comedy
8859    SuperBabies: Baby Geniuses 2 (2004)     0.5     Comedy
54290   Bratz: The Movie (2007) 0.5     Comedy
54503   Superbad (2007) 0.5     Comedy
4718    American Pie 2 (2001)   0.5     Comedy
5481    Austin Powers in Goldmember (2002)      0.5     Comedy
7451    Mean Girls (2004)       0.5     Comedy
4013    Mr. Accident (2000)     0.5     Comedy
65      Bio-Dome (1996) 0.5     Comedy
Time taken: 122.397 seconds, Fetched: 30423 row(s)
hive>
```

IMPALA:



```
| 4718    | American Pie 2 (2001)                                                          | 0.5      | Comedy |
| 216     | Billy Madison (1995)                                                           | 0.5      | Comedy |
| 520     | Robin Hood: Men in Tights (1993)                                               | 0.5      | Comedy |
| 1485    | Liar Liar (1997)                                                               | 0.5      | Comedy |
| 1911    | Dr. Dolittle (1998)                                                            | 0.5      | Comedy |
| 2567    | EDtv (1999)                                                                    | 0.5      | Comedy |
| 2694    | Big Daddy (1999)                                                               | 0.5      | Comedy |
| 2770    | Bowfinger (1999)                                                               | 0.5      | Comedy |
| 5283    | National Lampoon's Van Wilder (2002)                                           | 0.5      | Comedy |
| 5308    | Three Men and a Baby (1987)                                                    | 0.5      | Comedy |
| 5500    | Top Secret! (1984)                                                             | 0.5      | Comedy |
| 3042    | Meatballs III (1987)                                                           | 0.5      | Comedy |
| 4215    | Revenge of the Nerds II: Nerds in Paradise (1987)                              | 0.5      | Comedy |
| 3821    | Nutty Professor II: The Klumps (2000)                                          | 0.5      | Comedy |
| 6482    | Dumb and Dumberer: When Harry Met Lloyd (2003)                                 | 0.5      | Comedy |
| 104     | Happy Gilmore (1996)                                                           | 0.5      | Comedy |
| 216     | Billy Madison (1995)                                                           | 0.5      | Comedy |
| 3421    | Animal House (1978)                                                            | 0.5      | Comedy |
| 3979    | Little Nicky (2000)                                                            | 0.5      | Comedy |
| 8859    | SuperBabies: Baby Geniuses 2 (2004)                                            | 0.5      | Comedy |
| 54290   | Bratz: The Movie (2007)                                                        | 0.5      | Comedy |
| 54503   | Superbad (2007)                                                                | 0.5      | Comedy |
| 4718    | American Pie 2 (2001)                                                          | 0.5      | Comedy |
| 5481    | Austin Powers in Goldmember (2002)                                             | 0.5      | Comedy |
| 7451    | Mean Girls (2004)                                                              | 0.5      | Comedy |
| 4013    | Mr. Accident (2000)                                                            | 0.5      | Comedy |
| 65      | Bio-Dome (1996)                                                                | 0.5      | Comedy |
+---------+--------------------------------------------------------------------------------+----------+--------+
Fetched 30423 row(s) in 9.08s
[quickstart.cloudera:21000] > 
```

III. PERFORMANCE ANALYSIS OF HIVE WITH IMPALA FOR ABOVE CASES



Impala	1.18	5.11	13.20	6.35	9.08
Hive	48.587	97.669	102.226	137.997	122.397

IV. CONCLUSION

Cloudera's Impala has many advantages over Apache Hive. Considering the performance of both, Cloudera's Impala is always preferable to developers when it comes to analyzing HDFS or HBase data because it does not require moving this data, Impala uses Hive's metastore. As Cloudera's Impala is written in C/C++ the advantage is it takes less time to execute the query but it has one demerit that it is not suitable for every file format especially for a file written in java. Although Cloudera's Impala is fast when it comes to upgradation project where compatibility is as important as speed, Apache Hive would nudge. Thus we conclude that Cloudera's Impala's performance is better, time efficient than Apache Hive.

REFERENCES

[1]. https://www.cloudera.com/documentation/enterprise/5-8-x/topics/impala_file_formats.html
[2]. https://www.linkedin.com/pulse/20140910142911-22744472-why-is-impala-faster-than-hive
[3]. https://www.slideshare.net/hadooparchbook/impala-architecture-presentation
[4]. https://www.dezyre.com/article/impala-vs-hive-difference-between-sql-on-hadoop-components/180
[5]. https://grouplens.org/datasets/movielens
[6]. Big Data: An Introduction, ARD-IJEET, ISSN- 2320-8821, Volume 5, Issue 1
[7]. https://en.wikipedia.org/wiki/Apache_Hive
[8]. https://impala.apache.org/

BIOGRAPHY

1. Ms. Prasad Mitkari is pursuing his bachelor's degree from Maharashtra Institute of Technology. He is a student of final year, computer science and engineering department. Recently, He completed training on Cloudera's Big Data Course from MIT's Big Data Academy and also doing his final year project in big data area.

2. Ms. Rutuja Banswal is pursuing her bachelor's degree from Maharashtra Institute of Technology. She is a student of final year, computer science and engineering department. Recently, she completed training on Cloudera's Big Data Course from MIT's Big Data Academy and also doing her final year project in big data area.

3. Ms. Akshata Sirsat is pursuing her bachelor's degree from Maharashtra Institute of Technology. She is a student of final year, computer science and engineering department. Recently, she completed training on Cloudera's Big Data Course from MIT's Big Data Academy and also doing her final year project in big data area.

4. Prof. Daivashala Deshmukh is an assistant professor in Maharashtra Institute of Technology. Other than academics she is a coordinator and instructor for Big Data Academy in Computer science and Engineering department, Maharashtra Institute of Technology. She completed training of Cloudera's and HortornWork's Big Data Course. Her current research area of interest is Big Data.