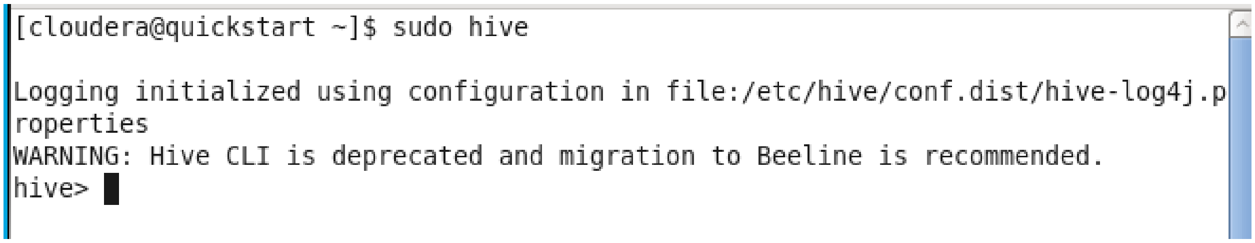
EXPERIMENT 5

Name	Shreya Shetty
UID	2019141059
Batch	A
Class	TE IT
Subject	BDA

AIM: Extract facts in real world dataset using Hive.

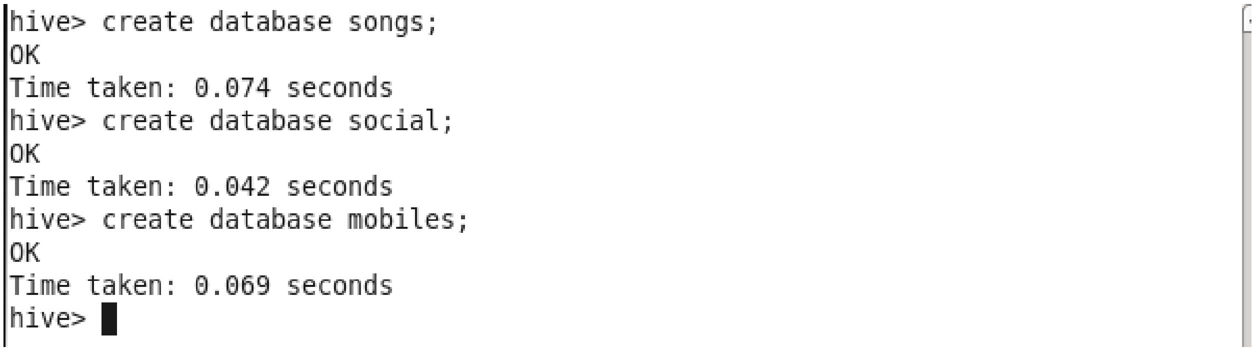
COMMANDS:

1. Starting Hive on Cloudera
   $sudo hive;



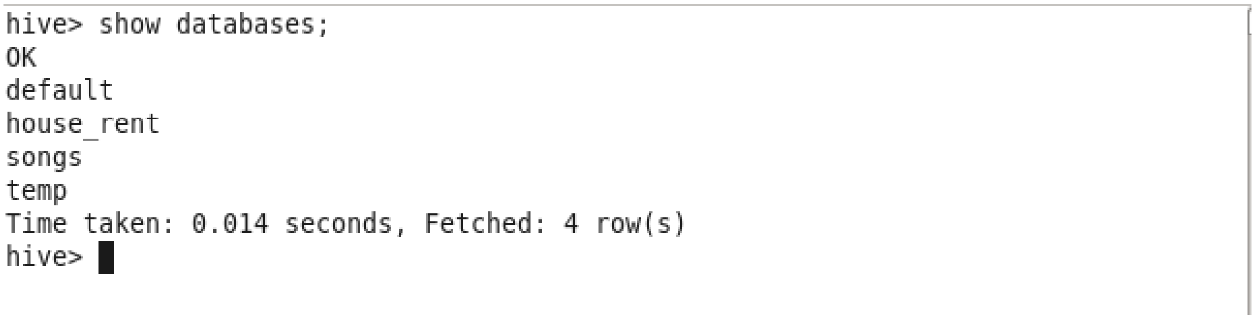
```
[cloudera@quickstart ~]$ sudo hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> 
```

2. Creating a Database ‘songs’
   $create database songs;



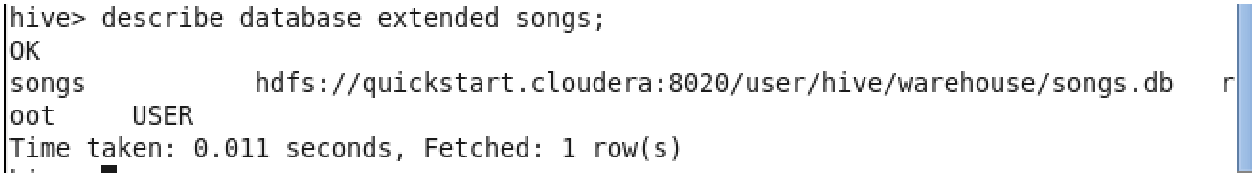
```
hive> create database songs;
OK
Time taken: 0.074 seconds
hive> create database social;
OK
Time taken: 0.042 seconds
hive> create database mobiles;
OK
Time taken: 0.069 seconds
hive> 
```

3. To show all the Databases present
   $show databases;



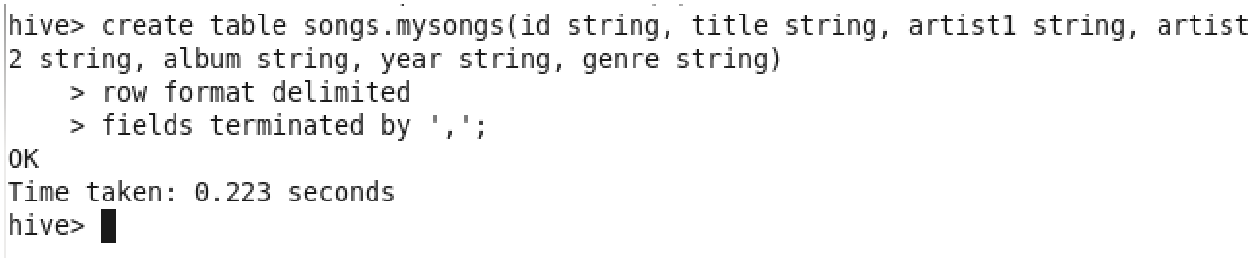
```
hive> show databases;
OK
default
house_rent
songs
temp
Time taken: 0.014 seconds, Fetched: 4 row(s)
hive> █
```

4. Describing database i.e. the format of the database
   $describe database extended songs;



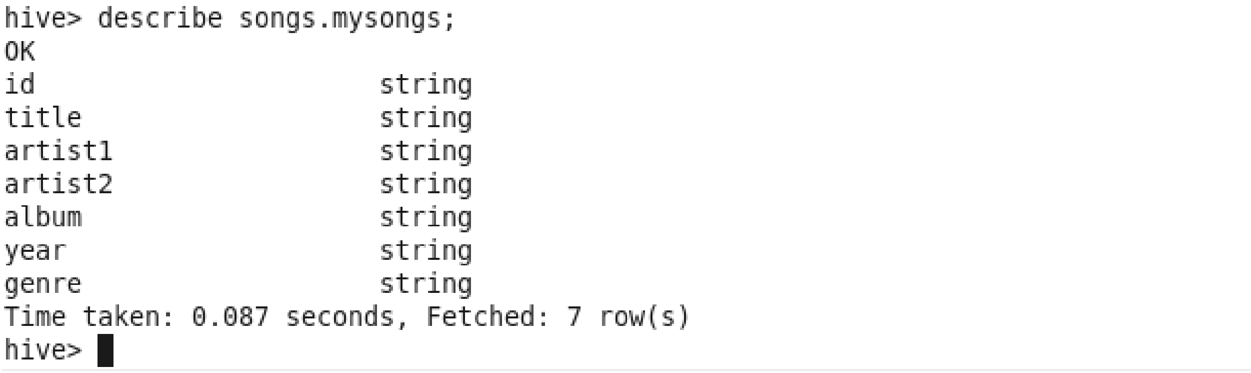
```
hive> describe database extended songs;
OK
songs           hdfs://quickstart.cloudera:8020/user/hive/warehouse/songs.db   r
oot     USER
Time taken: 0.011 seconds, Fetched: 1 row(s)
```

5. Creating Table ‘mysongs’ in database ‘songs’
   $create table songs.mysongs(id string, title string, artist1 string, artist2 string, album string, year string, genre string)
   >row format delimited
   >fields terminated by ‘,’;



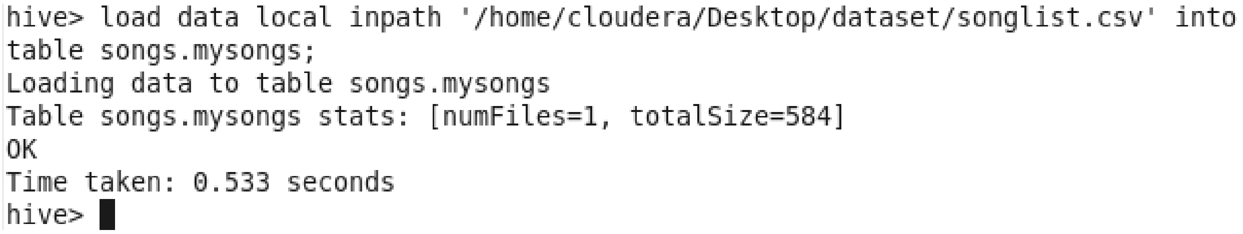
```
hive> create table songs.mysongs(id string, title string, artist1 string, artist
2 string, album string, year string, genre string)
    > row format delimited
    > fields terminated by ',';
OK
Time taken: 0.223 seconds
hive> █
```

6. Describing Table ‘mysongs’ i.e. the format of the table
   $describe songs.mysongs;



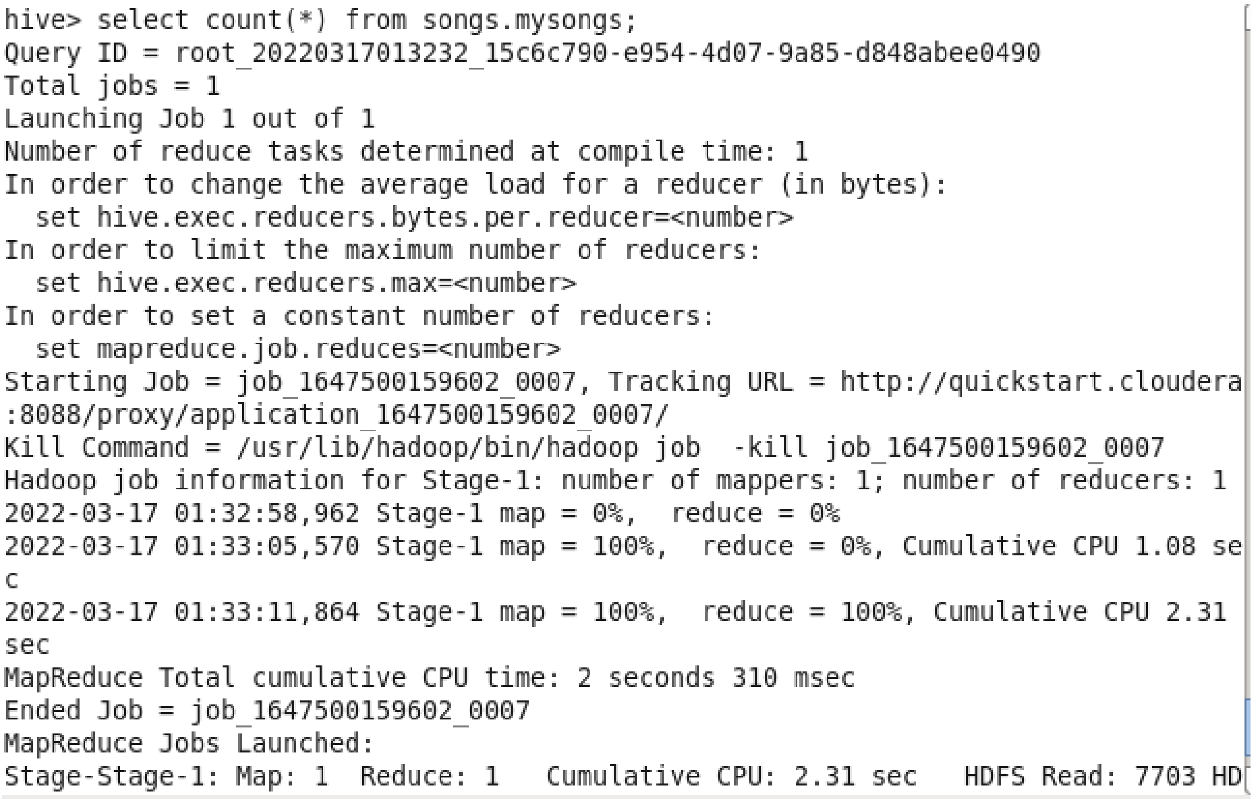
```
hive> describe songs.mysongs;
OK
id                      string
title                   string
artist1                 string
artist2                 string
album                   string
year                    string
genre                   string
Time taken: 0.087 seconds, Fetched: 7 row(s)
hive> █
```

7. Loading Data from csv file into table
   $load data inpath ‘/home/cloudera/Desktop/dataset/songlist.csv’ into table songs.mysongs;

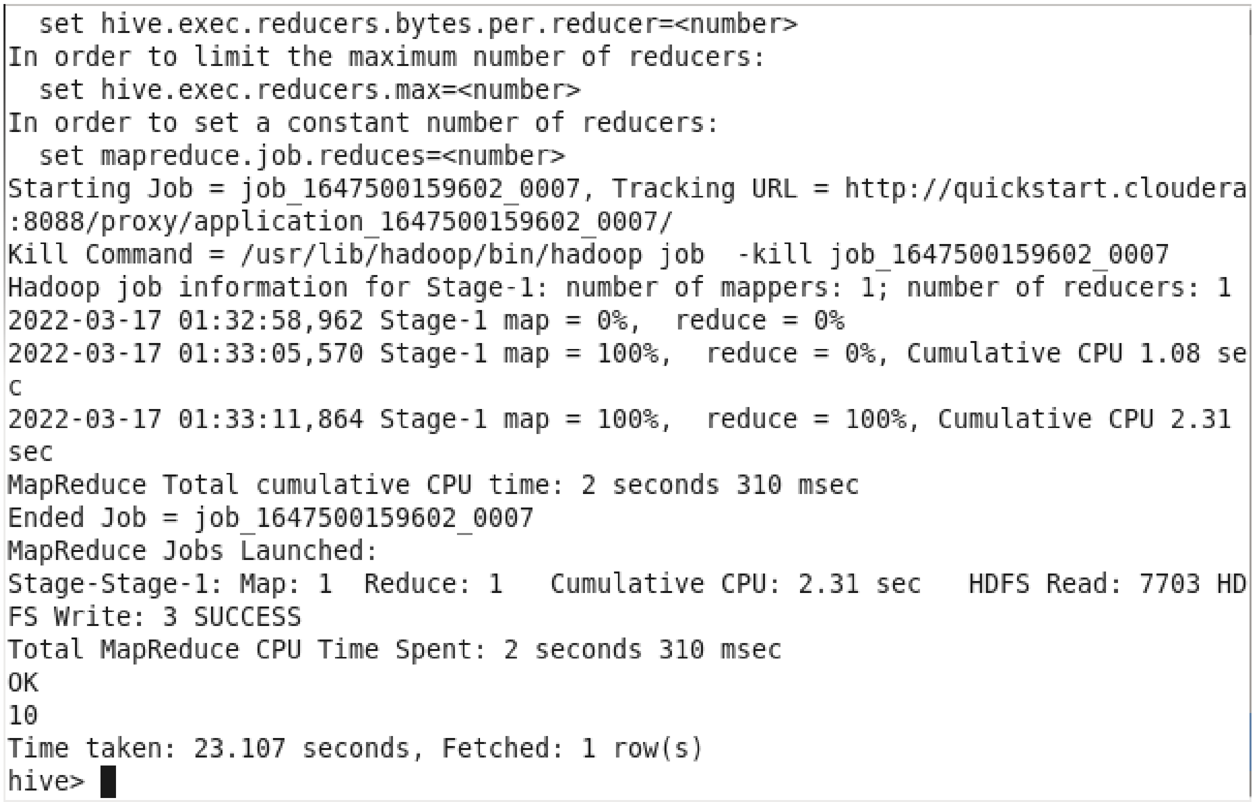


```
hive> load data local inpath '/home/cloudera/Desktop/dataset/songlist.csv' into
table songs.mysongs;
Loading data to table songs.mysongs
Table songs.mysongs stats: [numFiles=1, totalSize=584]
OK
Time taken: 0.533 seconds
hive> █
```

8. Selecting Count of all columns from Table ‘mysongs, to get the total number of rows
   $select count(*) from songs.mysongs;

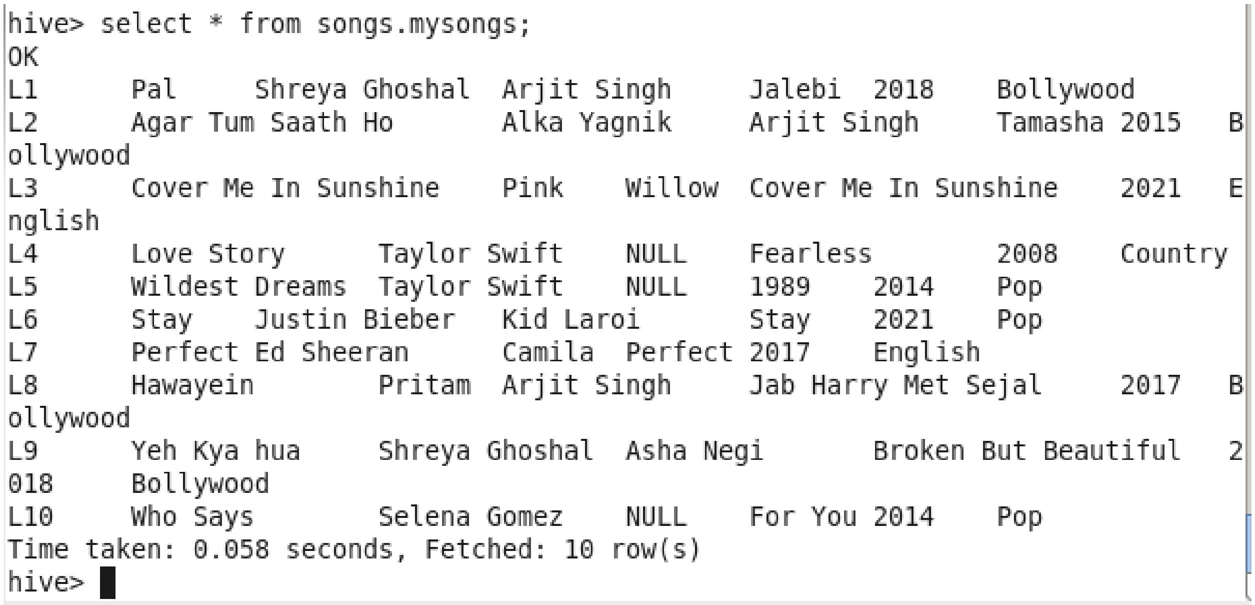


```
hive> select count(*) from songs.mysongs;
Query ID = root_20220317013232_15c6c790-e954-4d07-9a85-d848abee0490
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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_1647500159602_0007, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1647500159602_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1647500159602_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-17 01:32:58,962 Stage-1 map = 0%,  reduce = 0%
2022-03-17 01:33:05,570 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.08 se
c
2022-03-17 01:33:11,864 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.31
sec
MapReduce Total cumulative CPU time: 2 seconds 310 msec
Ended Job = job_1647500159602_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.31 sec   HDFS Read: 7703 HD
```



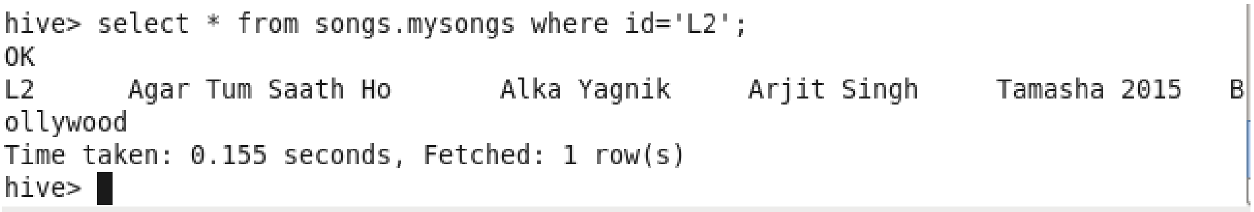
```
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1647500159602_0007, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1647500159602_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1647500159602_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-17 01:32:58,962 Stage-1 map = 0%,  reduce = 0%
2022-03-17 01:33:05,570 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.08 se
c
2022-03-17 01:33:11,864 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.31
sec
MapReduce Total cumulative CPU time: 2 seconds 310 msec
Ended Job = job_1647500159602_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.31 sec   HDFS Read: 7703 HD
FS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 310 msec
OK
10
Time taken: 23.107 seconds, Fetched: 1 row(s)
hive> █
```

9. Selecting all rows from table ‘mysongs’
   $select * from songs.mysongs;



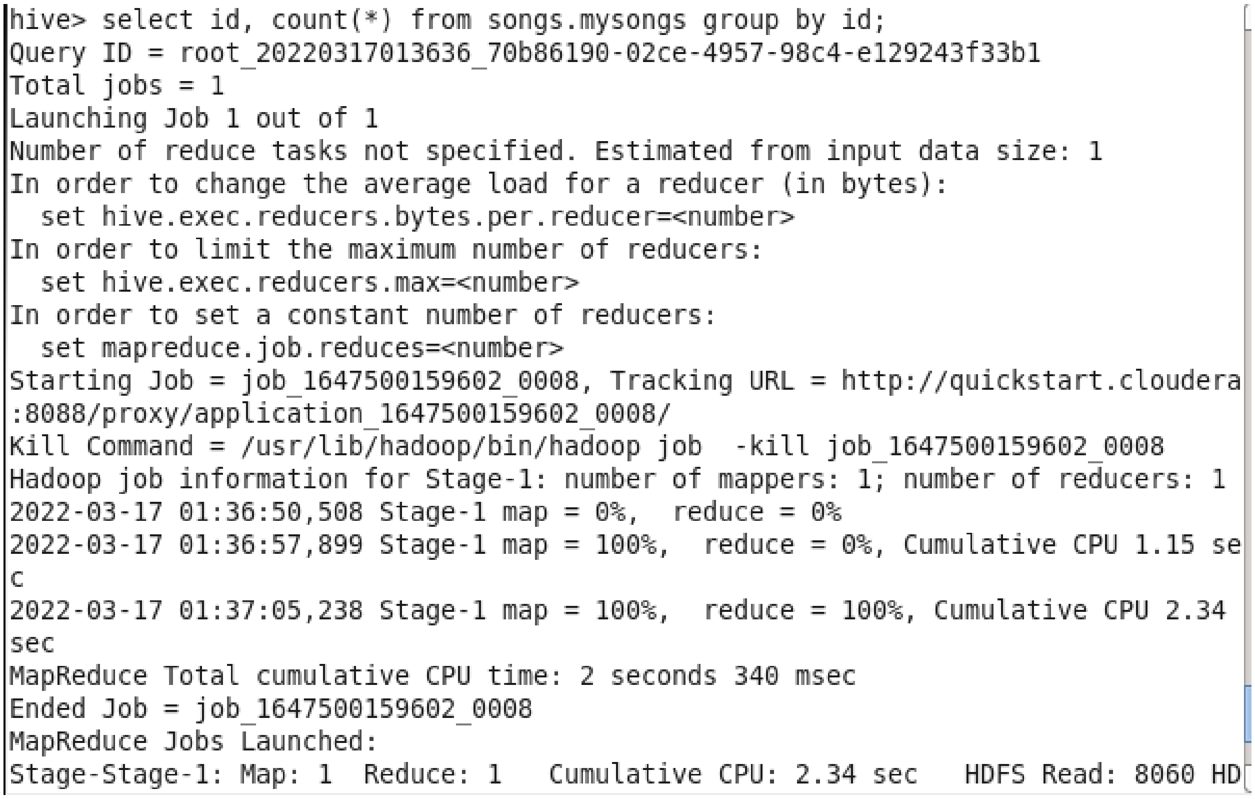
```
hive> select * from songs.mysongs;
OK
L1      Pal     Shreya Ghoshal  Arjit Singh     Jalebi  2018    Bollywood
L2      Agar Tum Saath Ho       Alka Yagnik     Arjit Singh     Tamasha 2015    B
ollywood
L3      Cover Me In Sunshine    Pink    Willow  Cover Me In Sunshine    2021    E
nglish
L4      Love Story      Taylor Swift    NULL    Fearless        2008    Country
L5      Wildest Dreams  Taylor Swift    NULL    1989    2014    Pop
L6      Stay    Justin Bieber   Kid Laroi       Stay    2021    Pop
L7      Perfect Ed Sheeran      Camila  Perfect 2017    English
L8      Hawayein        Pritam  Arjit Singh     Jab Harry Met Sejal     2017    B
ollywood
L9      Yeh Kya hua     Shreya Ghoshal  Asha Negi       Broken But Beautiful    2
018     Bollywood
L10     Who Says        Selena Gomez    NULL    For You 2014    Pop
Time taken: 0.058 seconds, Fetched: 10 row(s)
hive> █
```

10. Selecting a particular row from table where id is ‘L2’
    $select * from songs.mysongs where id=‘L2’;

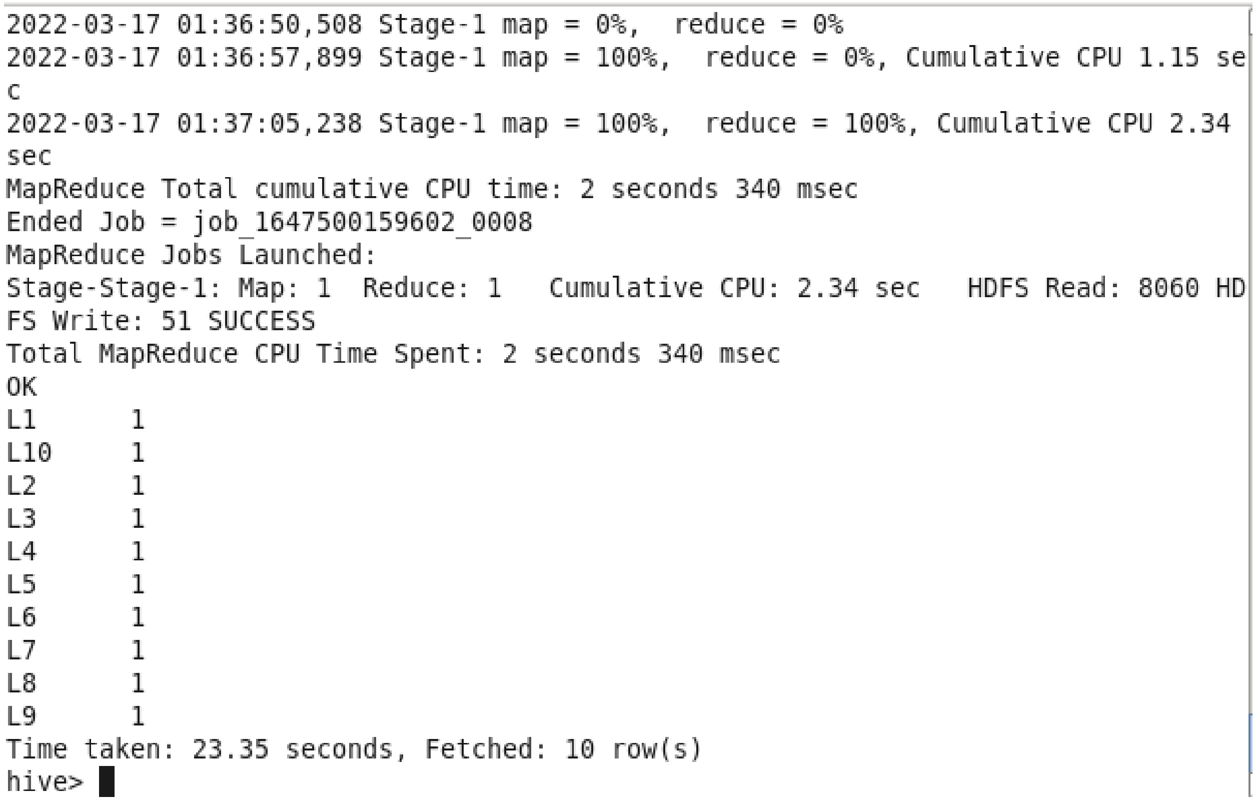


```
hive> select * from songs.mysongs where id='L2';
OK
L2      Agar Tum Saath Ho       Alka Yagnik     Arjit Singh     Tamasha 2015   B
ollywood
Time taken: 0.155 seconds, Fetched: 1 row(s)
hive> 
```

11. Selecting Count of rows from table grouped by Id
    $select id, count(*) from songs.mysongs group by id;



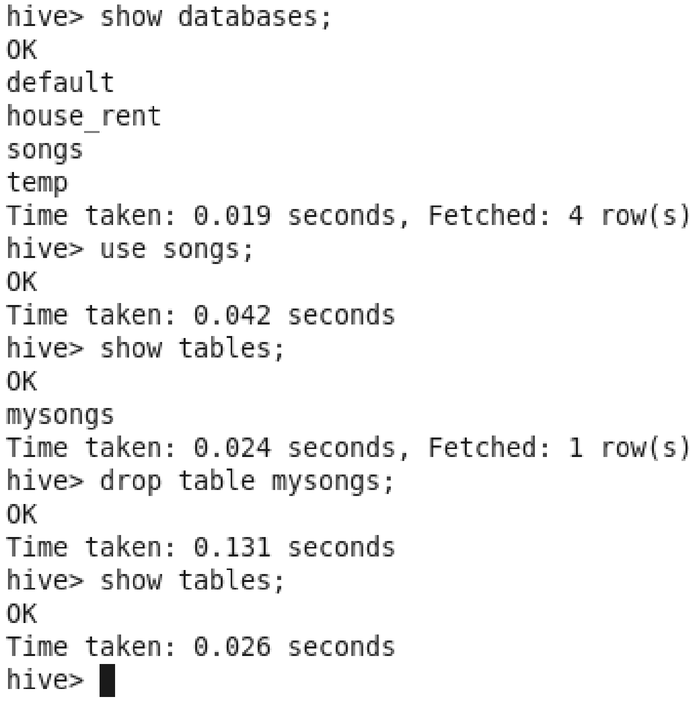
```
hive> select id, count(*) from songs.mysongs group by id;
Query ID = root_20220317013636_70b86190-02ce-4957-98c4-e129243f33b1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  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_1647500159602_0008, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1647500159602_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1647500159602_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-17 01:36:50,508 Stage-1 map = 0%,  reduce = 0%
2022-03-17 01:36:57,899 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.15 se
c
2022-03-17 01:37:05,238 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.34
sec
MapReduce Total cumulative CPU time: 2 seconds 340 msec
Ended Job = job_1647500159602_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.34 sec   HDFS Read: 8060 HD
```



```
2022-03-17 01:36:50,508 Stage-1 map = 0%,  reduce = 0%
2022-03-17 01:36:57,899 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.15 se
c
2022-03-17 01:37:05,238 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.34
sec
MapReduce Total cumulative CPU time: 2 seconds 340 msec
Ended Job = job_1647500159602_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.34 sec   HDFS Read: 8060 HD
FS Write: 51 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 340 msec
OK
L1      1
L10     1
L2      1
L3      1
L4      1
L5      1
L6      1
L7      1
L8      1
L9      1
Time taken: 23.35 seconds, Fetched: 10 row(s)
hive> █
```

12. Deleting Table ‘mysongs’

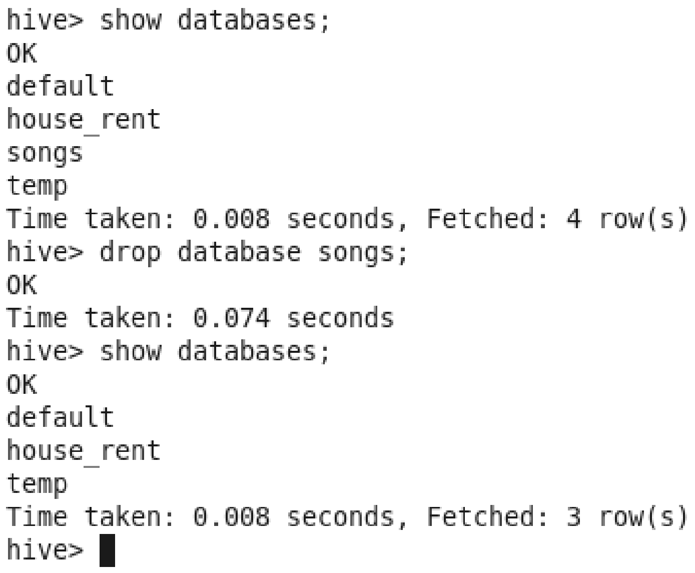
$drop table mysongs;



```
hive> show databases;
OK
default
house_rent
songs
temp
Time taken: 0.019 seconds, Fetched: 4 row(s)
hive> use songs;
OK
Time taken: 0.042 seconds
hive> show tables;
OK
mysongs
Time taken: 0.024 seconds, Fetched: 1 row(s)
hive> drop table mysongs;
OK
Time taken: 0.131 seconds
hive> show tables;
OK
Time taken: 0.026 seconds
hive> █
```

13. Deleting Database ‘songs’

    $drop database songs;



```
hive> show databases;
OK
default
house_rent
songs
temp
Time taken: 0.008 seconds, Fetched: 4 row(s)
hive> drop database songs;
OK
Time taken: 0.074 seconds
hive> show databases;
OK
default
house_rent
temp
Time taken: 0.008 seconds, Fetched: 3 row(s)
hive> █
```

Conclusion:

In this experiment, I learnt to use and run commands on Apache Hive. Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.