iNeuron

**Problem statement:**

**In this case study, we are giving a Movies metadata example of how to use HIVE on top of the HADOOP for different exploratory data analysis. In here, we have a predefined** dataset (movies_metadata.csv, credits.csv, ratings.csv,links.csv.

**Queries for Hive Case study**

**Tasks:**

**Create table for movies_metadata.csv using complex data type struct and map**

**Hive > Create tables for all datasets**

```
create table movies_metadata

(

adult boolean,

budget bigint,

genre array<string>,

id bigint,

imdb_id bigint,

original string,

popular double,

production_house struct<production_companies:string,production_countries:string>,

release_date date,

revenue bigint,

runtime int,

language map<string,string>,

status string,

title string,

video boolean,

vote_average double,

vote_count int
)
row format delimited
fields terminated by ','
collection items terminated by '|'
map keys terminated by ':'
TBLPROPERTIES ("skip.header.line.count"="1");
```

Load data into table from local file system movies_metadata.csv

```
Hive > load data local inpath'/home/cloudera/sidd/Challenge/Mini_project_4/movies_metadata.csv' into table movies_metadata;
```

```
Hive > select distinct(year(release_date) from movies_metadata;
```

Create Table for ratings.csv

```
Hive > create table ratings
(
userId int,
movieId bigint,
rating double,
timestamp timestamp
)
row format delimited
fields terminated by ','
TBLPROPERTIES ("skip.header.line.count"="1");
```

Load data into table from Local file system ratings.csv

```
load data local inpath'/home/cloudera/sidd/Challenge/Mini_project_4/ratings.csv' into table ratings;
```

**Create table fpr credits.csv using complex data type struct**

```
Hive > create table credits

(

Id int,

casts struct<cast_id:int,character:string,credit_id:string,gender:int,id:int,name:string>,

crew struct<credit_id:string,department:string,gender:int,id:bigint,job:string,name:string>

)

row format delimited

fields terminated by ','

collection items terminated by '|'

TBLPROPERTIES ("skip.header.line.count"="1");
```

**Load data into Table from local file system credits.csv**

```
Hive > load data local inpath'/home/cloudera/sidd/Challenge/Mini_project_4/credits.csv' into table credits;
```

**Create Table for  links**

```
Hive > create table links

(

movieId int,

imdbId bigint,

tmdbId bigint

)

row format delimited

fields terminated by ','

TBLPROPERTIES ("skip.header.line.count"="1");
```

**Tables are recreated with partitioning and bucketing**

```
create table movies_metadata_partitioned
(
adult boolean,
budget bigint,
genre array<string>,
id bigint,
imdb_id bigint,
original string,
popular double,
production_house struct<production_companies:string,production_countries:string>,
release_date date,
revenue bigint,
runtime int,
language map<string,string>,
title string,
video boolean,
vote_average double,
vote_count int
)
partitioned by (status string)
clustered by (vote_average) INTO 10 BUCKETS;
```

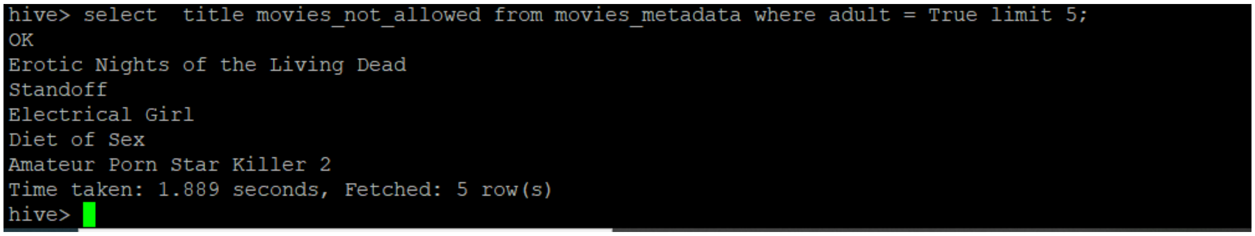
```
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;

set hive.enforce.bucketing = true;



insert into table movies_metadata_partitioned partition(status) select
adult,budget,genre,id,imdb_id,original,popular,production_house,release_date,revenue,
runtime,language,title,video,vote_average,vote_count,status from movies_metadata2;



create table links_partitioned
(
movieId int,
imdbId bigint,
tmdbId bigint
)
clustered by (movieId) sorted by (movieId) INTO 5 BUCKETS;

insert overwrite  table links_partitioned select * from links;



create table credits_partitioned
(
Id int,
casts struct<cast_id:int,character:string,credit_id:string,gender:int,id:int,name:string>,
crew struct<credit_id:string,department:string,gender:int,id:bigint,job:string,name:string>
)
clustered by (Id) sorted by (Id) INTO 5 BUCKETS;
```

```
insert overwrite table credits_partitioned select * from credits;

create table ratings_partitioned
(
userId int,
movieId bigint,
rating double,
timestamp timestamp
)
clustered by (rating) sorted by (rating) INTO 10 BUCKETS;

insert overwrite table ratings_partitioned select * from ratings;
```

**Analytical queries:**

**find the movies list which are not allowed for childs**

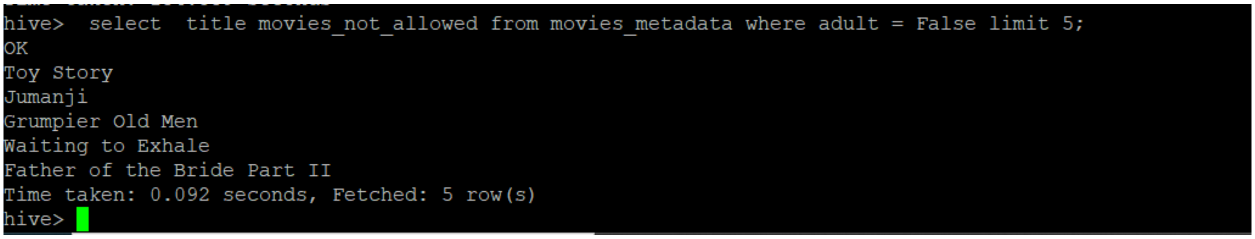
**Hive > select  title from movies_metadata where adult = True;**



```
hive> select  title movies_not_allowed from movies_metadata where adult = True limit 5;
OK
Erotic Nights of the Living Dead
Standoff
Electrical Girl
Diet of Sex
Amateur Porn Star Killer 2
Time taken: 1.889 seconds, Fetched: 5 row(s)
hive>
```

**find the movies list which are  allowed for childs**

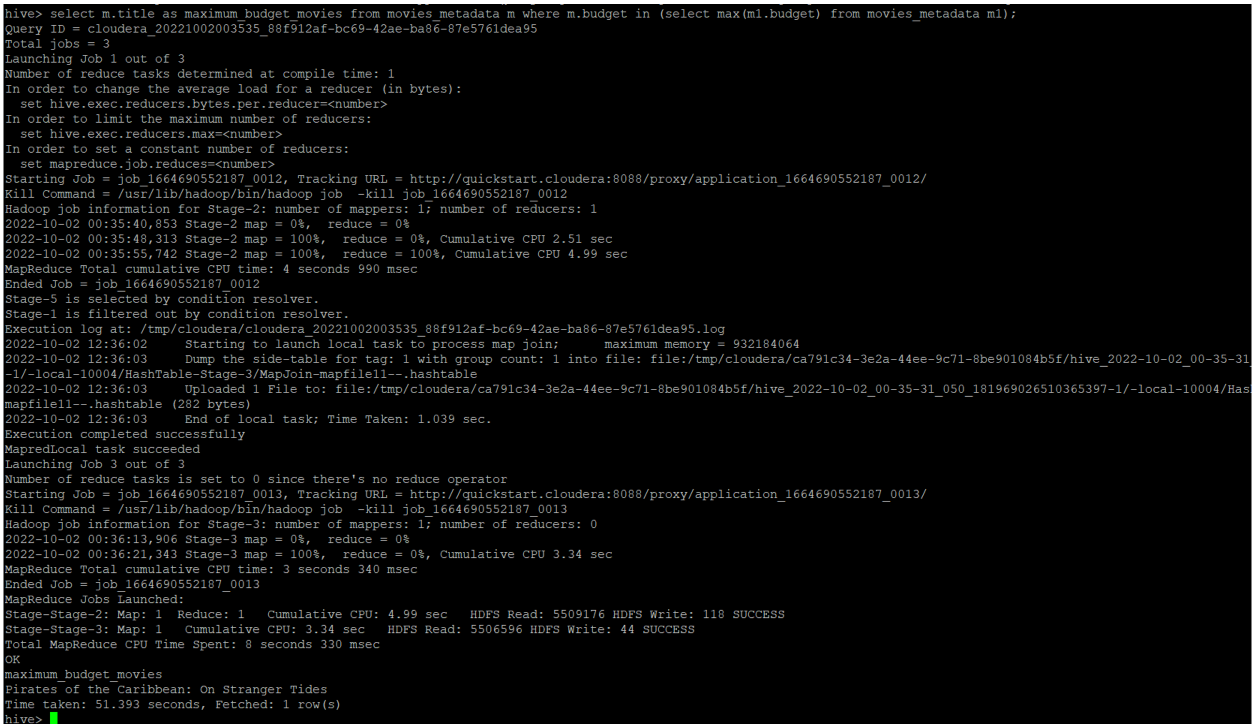
**Hive > select  title from movies_metadata where adult = False;**



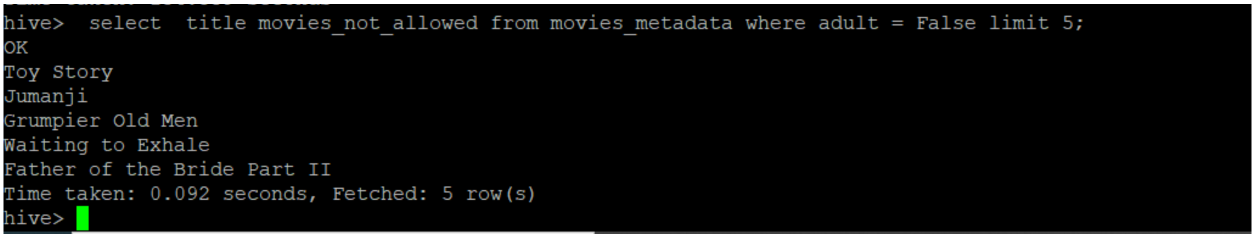
```
hive>  select  title movies_not_allowed from movies_metadata where adult = False limit 5;
OK
Toy Story
Jumanji
Grumpier Old Men
Waiting to Exhale
Father of the Bride Part II
Time taken: 0.092 seconds, Fetched: 5 row(s)
hive>
```

**Find out the movie which  has highest budget**

**Hive > select m.title as maximum_budget_movies from movies_metadata m where m.budget in (select max(m1.budget) from movies_metadata m1);**



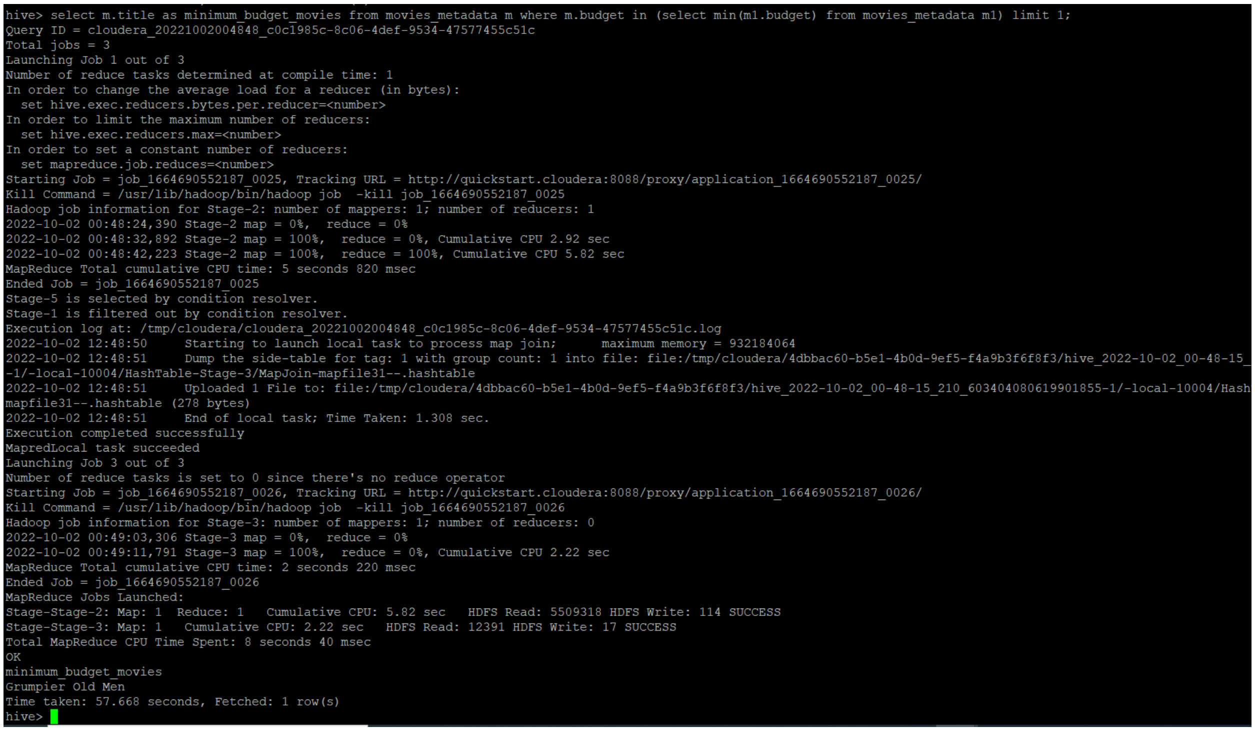
```
hive> select m.title as maximum_budget_movies from movies_metadata m where m.budget in (select max(m1.budget) from movies_metadata m1);
Query ID = cloudera_20221002003535_88f912af-bc69-42ae-ba86-87e5761dea95
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 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_1664690552187_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0012
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-02 00:35:40,853 Stage-2 map = 0%,  reduce = 0%
2022-10-02 00:35:48,313 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.51 sec
2022-10-02 00:35:55,742 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.99 sec
MapReduce Total cumulative CPU time: 4 seconds 990 msec
Ended Job = job_1664690552187_0012
Stage-5 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution log at: /tmp/cloudera/cloudera_20221002003535_88f912af-bc69-42ae-ba86-87e5761dea95.log
2022-10-02 12:36:02     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 12:36:03     Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_00-35-31_
-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile11--.hashtable
2022-10-02 12:36:03     Uploaded 1 File to: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_00-35-31_050_181969026510365397-1/-local-10004/Has
mapfile11--.hashtable (282 bytes)
2022-10-02 12:36:03     End of local task; Time Taken: 1.039 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0013
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-10-02 00:36:13,906 Stage-3 map = 0%,  reduce = 0%
2022-10-02 00:36:21,343 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 3.34 sec
MapReduce Total cumulative CPU time: 3 seconds 340 msec
Ended Job = job_1664690552187_0013
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.99 sec   HDFS Read: 5509176 HDFS Write: 118 SUCCESS
Stage-Stage-3: Map: 1   Cumulative CPU: 3.34 sec   HDFS Read: 5506596 HDFS Write: 44 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 330 msec
OK
maximum_budget_movies
Pirates of the Caribbean: On Stranger Tides
Time taken: 51.393 seconds, Fetched: 1 row(s)
hive>
```



```
hive>  select  title movies_not_allowed from movies_metadata where adult = False limit 5;
OK
Toy Story
Jumanji
Grumpier Old Men
Waiting to Exhale
Father of the Bride Part II
Time taken: 0.092 seconds, Fetched: 5 row(s)
hive>
```

**Find out the movie which  has highest budget**

**Hive > select m.title as minimum_budget_movies from movies_metadata m where m.budget in (select min(m1.budget) from movies_metadata m1) limit 1;**

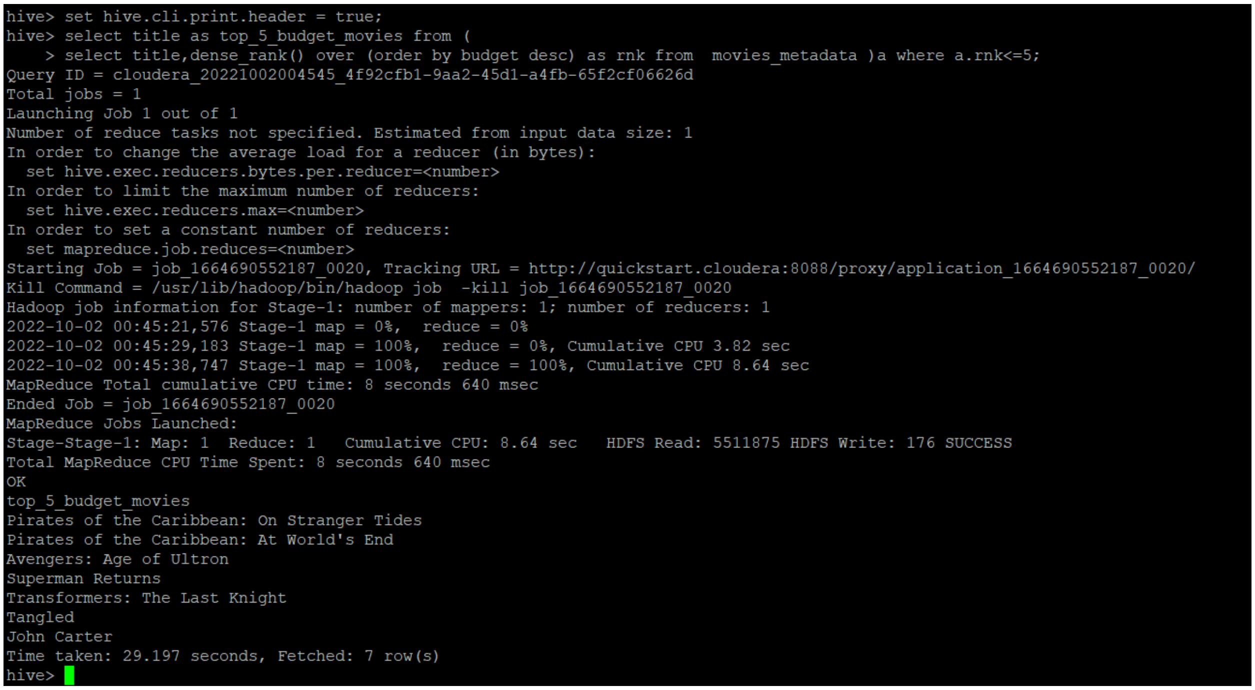


```
hive> select m.title as minimum_budget_movies from movies_metadata m where m.budget in (select min(m1.budget) from movies_metadata m1) limit 1;
Query ID = cloudera_20221002004848_c0c1985c-8c06-4def-9534-47577455c51c
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile 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_1664690552187_0025, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0025/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0025
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-02 00:48:24,390 Stage-2 map = 0%,  reduce = 0%
2022-10-02 00:48:32,892 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.92 sec
2022-10-02 00:48:42,223 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.82 sec
MapReduce Total cumulative CPU time: 5 seconds 820 msec
Ended Job = job_1664690552187_0025
Stage-5 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution log at: /tmp/cloudera/cloudera_20221002004848_c0c1985c-8c06-4def-9534-47577455c51c.log
2022-10-02 12:48:50     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 12:48:51     Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_00-48-15_
-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile31--.hashtable
2022-10-02 12:48:51     Uploaded 1 File to: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_00-48-15_210_603404080619901855-1/-local-10004/Hash
mapfile31--.hashtable (278 bytes)
2022-10-02 12:48:51     End of local task; Time Taken: 1.308 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0026, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0026/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0026
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-10-02 00:49:03,306 Stage-3 map = 0%,  reduce = 0%
2022-10-02 00:49:11,791 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.22 sec
MapReduce Total cumulative CPU time: 2 seconds 220 msec
Ended Job = job_1664690552187_0026
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.82 sec   HDFS Read: 5509318 HDFS Write: 114 SUCCESS
Stage-Stage-3: Map: 1   Cumulative CPU: 2.22 sec   HDFS Read: 12391 HDFS Write: 17 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 40 msec
OK
minimum_budget_movies
Grumpier Old Men
Time taken: 57.668 seconds, Fetched: 1 row(s)
hive> 
```

**which movie has highest budget top 5 using dense_rank function**

**Hive > select * from (**

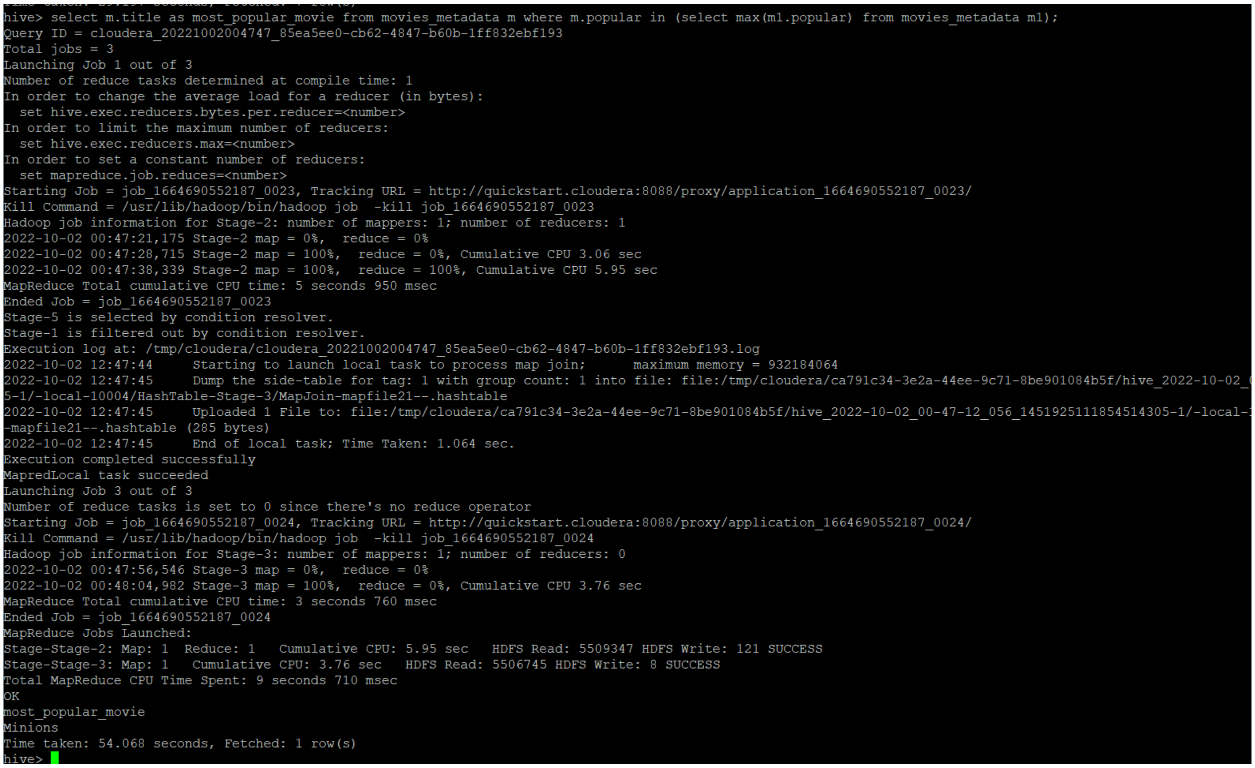
**select m.title,dense_rank(order by budget desc) as rnk from  movies_metadata m)a where a.rnk<=5;**



```
hive> set hive.cli.print.header = true;
hive> select title as top_5_budget_movies from (
    > select title,dense_rank() over (order by budget desc) as rnk from  movies_metadata )a where a.rnk<=5;
Query ID = cloudera_20221002004545_4f92cfb1-9aa2-45d1-a4fb-65f2cf06626d
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_1664690552187_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-02 00:45:21,576 Stage-1 map = 0%,  reduce = 0%
2022-10-02 00:45:29,183 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.82 sec
2022-10-02 00:45:38,747 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.64 sec
MapReduce Total cumulative CPU time: 8 seconds 640 msec
Ended Job = job_1664690552187_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.64 sec   HDFS Read: 5511875 HDFS Write: 176 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 640 msec
OK
top_5_budget_movies
Pirates of the Caribbean: On Stranger Tides
Pirates of the Caribbean: At World's End
Avengers: Age of Ultron
Superman Returns
Transformers: The Last Knight
Tangled
John Carter
Time taken: 29.197 seconds, Fetched: 7 row(s)
hive> █
```

**which is most popular movie**

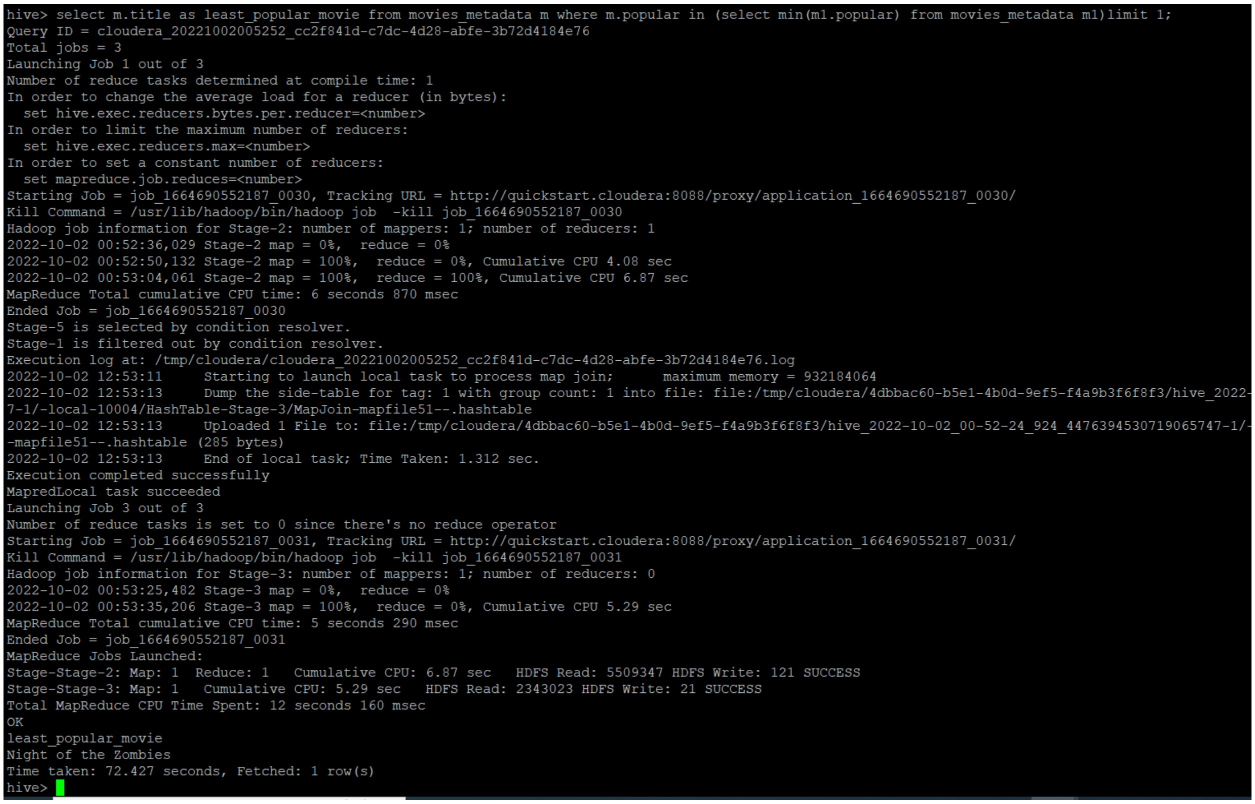
**Hive > select m.title as most_popular_movie from movies_metadata m where m.popular in (select max(m1.popular) from movies_metadata m1);**



```
hive> select m.title as most_popular_movie from movies_metadata m where m.popular in (select max(m1.popular) from movies_metadata m1);
Query ID = cloudera_20221002004747_85ea5ee0-cb62-4847-b60b-1ff832ebf193
Total jobs = 3
Launching Job 1 out of 3
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_1664690552187_0023, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0023/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0023
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-02 00:47:21,175 Stage-2 map = 0%,  reduce = 0%
2022-10-02 00:47:28,715 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.06 sec
2022-10-02 00:47:38,339 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.95 sec
MapReduce Total cumulative CPU time: 5 seconds 950 msec
Ended Job = job_1664690552187_0023
Stage-5 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution log at: /tmp/cloudera/cloudera_20221002004747_85ea5ee0-cb62-4847-b60b-1ff832ebf193.log
2022-10-02 12:47:44     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 12:47:45     Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_
5-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile21--.hashtable
2022-10-02 12:47:45     Uploaded 1 File to: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_00-47-12_056_1451925111854514305-1/-local-
-mapfile21--.hashtable (285 bytes)
2022-10-02 12:47:45     End of local task; Time Taken: 1.064 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0024, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0024/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0024
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-10-02 00:47:56,546 Stage-3 map = 0%,  reduce = 0%
2022-10-02 00:48:04,982 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 3.76 sec
MapReduce Total cumulative CPU time: 3 seconds 760 msec
Ended Job = job_1664690552187_0024
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.95 sec   HDFS Read: 5509347 HDFS Write: 121 SUCCESS
Stage-Stage-3: Map: 1   Cumulative CPU: 3.76 sec   HDFS Read: 5506745 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 710 msec
OK
most_popular_movie
Minions
Time taken: 54.068 seconds, Fetched: 1 row(s)
hive>
```

**which is least popular movie**

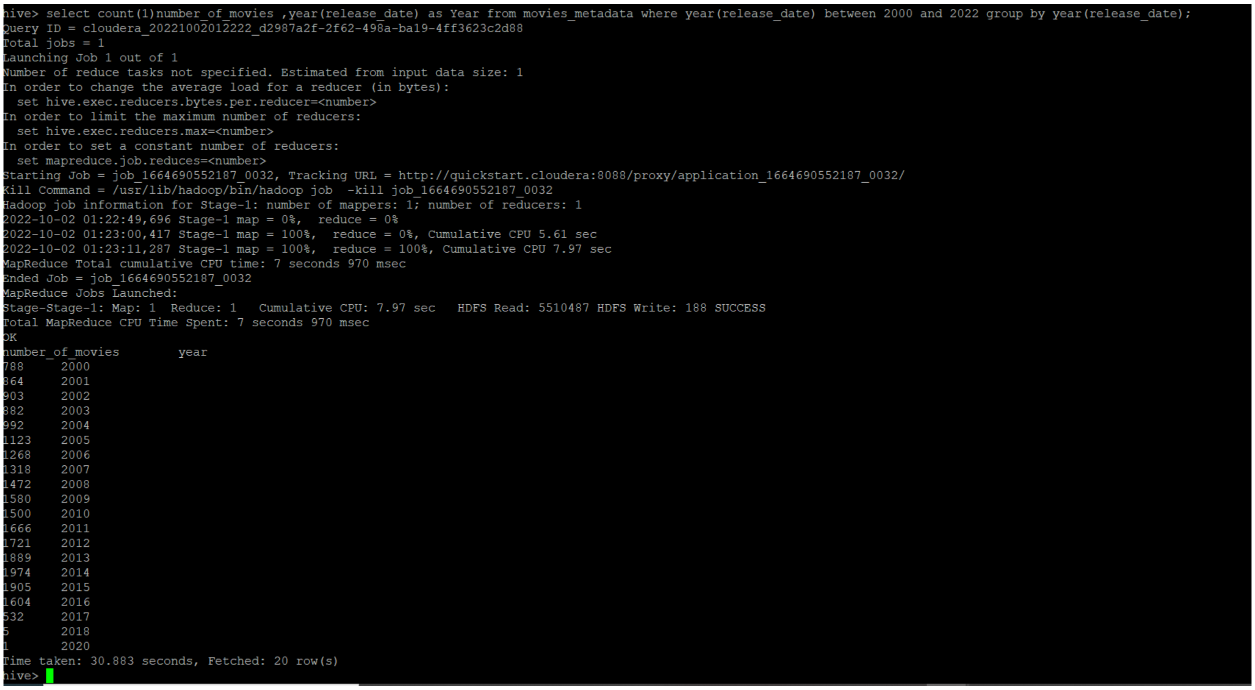
**Hive > select m.title as least_popular_movie from movies_metadata m where m.popular in (select min(m1.popular) from movies_metadata m1);**



```
hive> select m.title as least_popular_movie from movies_metadata m where m.popular in (select min(m1.popular) from movies_metadata m1)limit 1;
Query ID = cloudera_20221002005252_cc2f841d-c7dc-4d28-abfe-3b72d4184e76
Total jobs = 3
Launching Job 1 out of 3
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_1664690552187_0030, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0030/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0030
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-02 00:52:36,029 Stage-2 map = 0%,  reduce = 0%
2022-10-02 00:52:50,132 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 4.08 sec
2022-10-02 00:53:04,061 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 6.87 sec
MapReduce Total cumulative CPU time: 6 seconds 870 msec
Ended Job = job_1664690552187_0030
Stage-5 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution log at: /tmp/cloudera/cloudera_20221002005252_cc2f841d-c7dc-4d28-abfe-3b72d4184e76.log
2022-10-02 12:53:11     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 12:53:13     Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-
7-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile51--.hashtable
2022-10-02 12:53:13     Uploaded 1 File to: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_00-52-24_924_4476394530719065747-1/-
-mapfile51--.hashtable (285 bytes)
2022-10-02 12:53:13     End of local task; Time Taken: 1.312 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0031, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0031/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0031
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-10-02 00:53:25,482 Stage-3 map = 0%,  reduce = 0%
2022-10-02 00:53:35,206 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 5.29 sec
MapReduce Total cumulative CPU time: 5 seconds 290 msec
Ended Job = job_1664690552187_0031
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 6.87 sec   HDFS Read: 5509347 HDFS Write: 121 SUCCESS
Stage-Stage-3: Map: 1   Cumulative CPU: 5.29 sec   HDFS Read: 2343023 HDFS Write: 21 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 160 msec
OK
least_popular_movie
Night of the Zombies
Time taken: 72.427 seconds, Fetched: 1 row(s)
hive>
```

**number of movies releases per year**

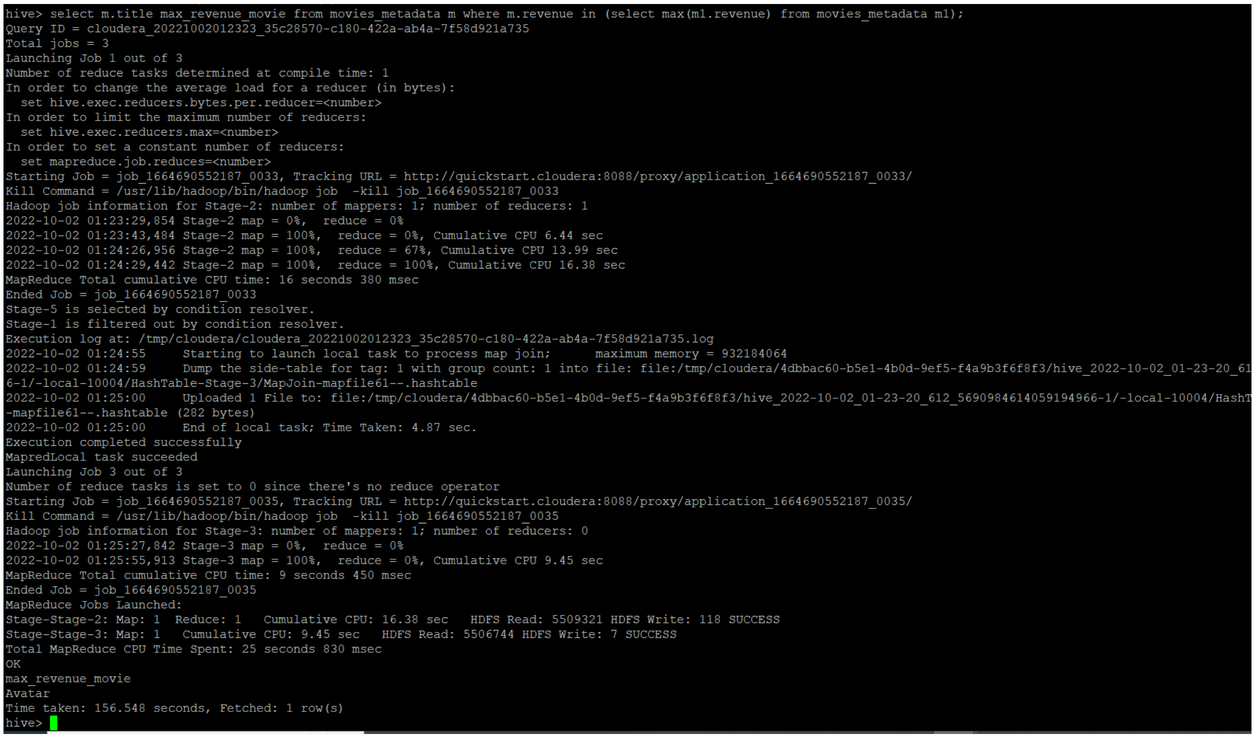
**Hive > select count(1),year(release_date) as Year from movies_metadata group by year(release_date);**



```
hive> select count(1)number_of_movies ,year(release_date) as Year from movies_metadata where year(release_date) between 2000 and 2022 group by year(release_date);
Query ID = cloudera_20221002012222_d2987a2f-2f62-498a-ba19-4ff3623c2d88
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_1664690552187_0032, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0032/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0032
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-02 01:22:49,696 Stage-1 map = 0%,  reduce = 0%
2022-10-02 01:23:00,417 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.61 sec
2022-10-02 01:23:11,287 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.97 sec
MapReduce Total cumulative CPU time: 7 seconds 970 msec
Ended Job = job_1664690552187_0032
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.97 sec   HDFS Read: 5510487 HDFS Write: 188 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 970 msec
OK
number_of_movies        year
788     2000
864     2001
903     2002
882     2003
992     2004
1123    2005
1268    2006
1318    2007
1472    2008
1580    2009
1500    2010
1666    2011
1721    2012
1889    2013
1974    2014
1905    2015
1604    2016
532     2017
5       2018
1       2020
Time taken: 30.883 seconds, Fetched: 20 row(s)
hive>
```

**which movie has hight revenue**

**Hive > select m.title max_revenue_movie from movies_metadata m where revenue = (select max(m1.revenue) from movies_metadata m1);**



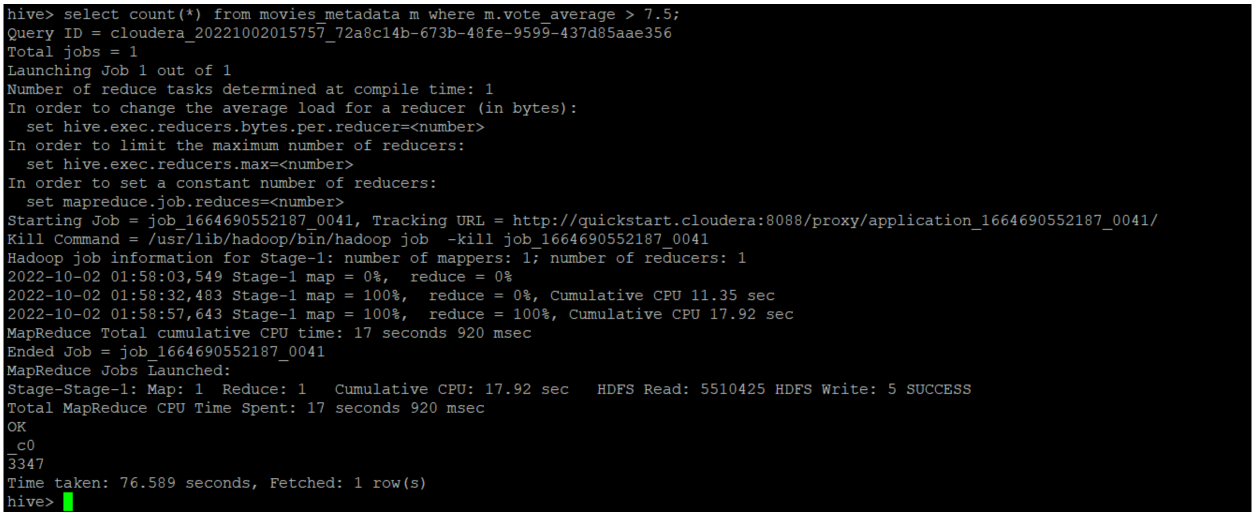
```
hive> select m.title max_revenue_movie from movies_metadata m where m.revenue in (select max(m1.revenue) from movies_metadata m1);
Query ID = cloudera_20221002012323_35c28570-c180-422a-ab4a-7f58d921a735
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 1
In 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_1664690552187_0033, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0033/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0033
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-02 01:23:29,854 Stage-2 map = 0%,  reduce = 0%
2022-10-02 01:23:43,484 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 6.44 sec
2022-10-02 01:24:26,956 Stage-2 map = 100%,  reduce = 67%, Cumulative CPU 13.99 sec
2022-10-02 01:24:29,442 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 16.38 sec
MapReduce Total cumulative CPU time: 16 seconds 380 msec
Ended Job = job_1664690552187_0033
Stage-5 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution log at: /tmp/cloudera/cloudera_20221002012323_35c28570-c180-422a-ab4a-7f58d921a735.log
2022-10-02 01:24:55     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 01:24:59     Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_01-23-20_61
6-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile61--.hashtable
2022-10-02 01:25:00     Uploaded 1 File to: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_01-23-20_612_5690984614059194966-1/-local-10004/HashT
-mapfile61--.hashtable (282 bytes)
2022-10-02 01:25:00     End of local task; Time Taken: 4.87 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0035, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0035/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0035
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-10-02 01:25:27,842 Stage-3 map = 0%,  reduce = 0%
2022-10-02 01:25:55,913 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 9.45 sec
MapReduce Total cumulative CPU time: 9 seconds 450 msec
Ended Job = job_1664690552187_0035
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 16.38 sec   HDFS Read: 5509321 HDFS Write: 118 SUCCESS
Stage-Stage-3: Map: 1   Cumulative CPU: 9.45 sec   HDFS Read: 5506744 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 25 seconds 830 msec
OK
max_revenue_movie
Avatar
Time taken: 156.548 seconds, Fetched: 1 row(s)
hive>
```

**which movie has lowest revenue**

**Hive > select m.title min_revenue_movie from movies_metadata m where revenue = (select min(m1.revenue) from movies_metadata m1);**

find the number of movies which has vote_average is greater than 7.5

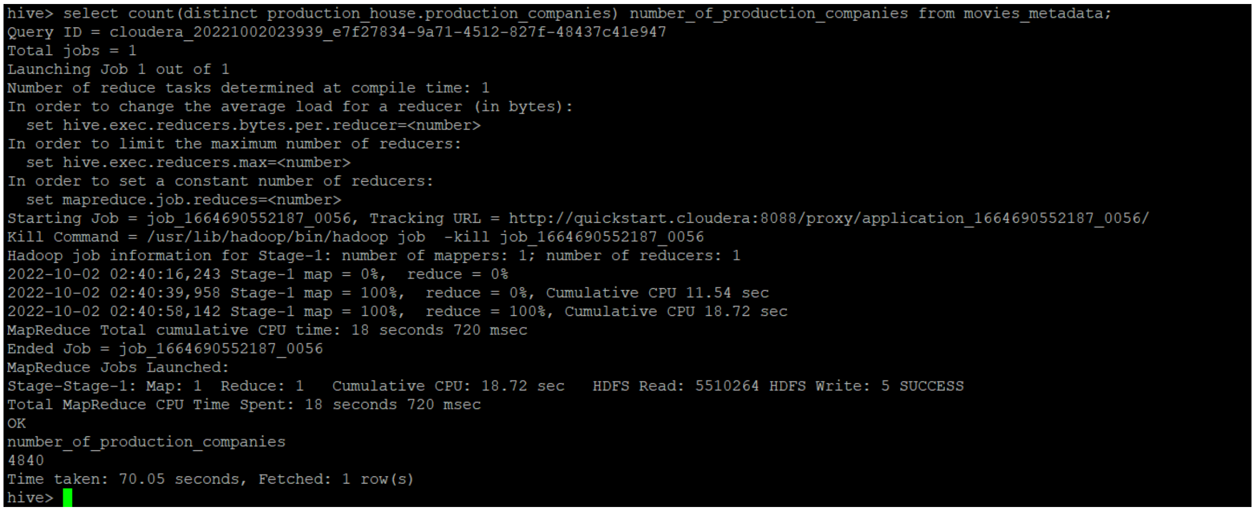
selecct count(*) from movies_metadata where vote_average > 7.5;



```
hive> select count(*) from movies_metadata m where m.vote_average > 7.5;
Query ID = cloudera_20221002015757_72a8c14b-673b-48fe-9599-437d85aae356
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_1664690552187_0041, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0041/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0041
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-02 01:58:03,549 Stage-1 map = 0%,  reduce = 0%
2022-10-02 01:58:32,483 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 11.35 sec
2022-10-02 01:58:57,643 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 17.92 sec
MapReduce Total cumulative CPU time: 17 seconds 920 msec
Ended Job = job_1664690552187_0041
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 17.92 sec   HDFS Read: 5510425 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 920 msec
OK
_c0
3347
Time taken: 76.589 seconds, Fetched: 1 row(s)
hive>
```

**Find out total number of production_companies**

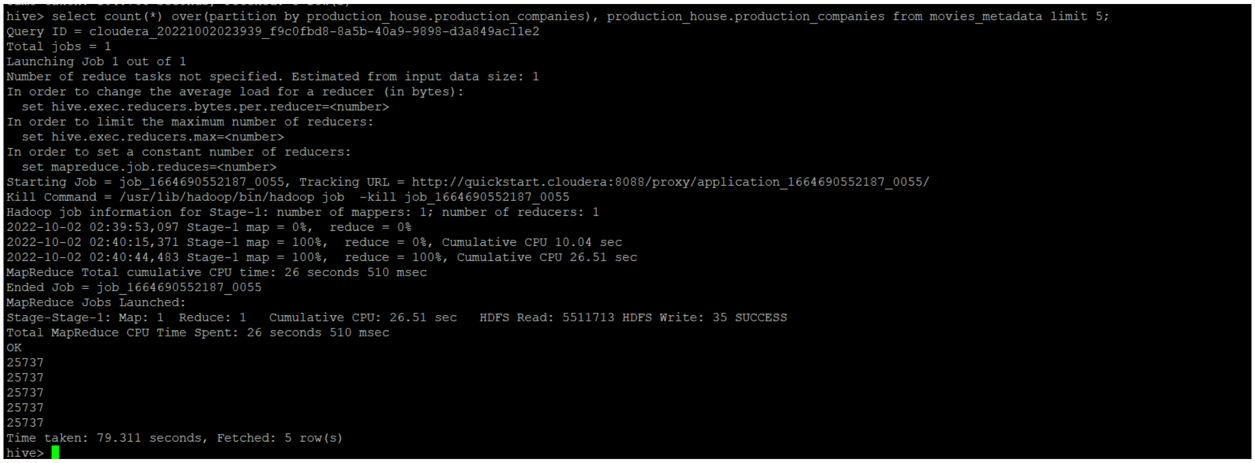
**Hive > select count(distinct production_house.production_companies) number_of_production_companies from movies_metadata;**



```
hive> select count(distinct production_house.production_companies) number_of_production_companies from movies_metadata;
Query ID = cloudera_20221002023939_e7f27834-9a71-4512-827f-48437c41e947
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_1664690552187_0056, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0056/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0056
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-02 02:40:16,243 Stage-1 map = 0%,  reduce = 0%
2022-10-02 02:40:39,958 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 11.54 sec
2022-10-02 02:40:58,142 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 18.72 sec
MapReduce Total cumulative CPU time: 18 seconds 720 msec
Ended Job = job_1664690552187_0056
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 18.72 sec   HDFS Read: 5510264 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 720 msec
OK
number_of_production_companies
4840
Time taken: 70.05 seconds, Fetched: 1 row(s)
hive>
```

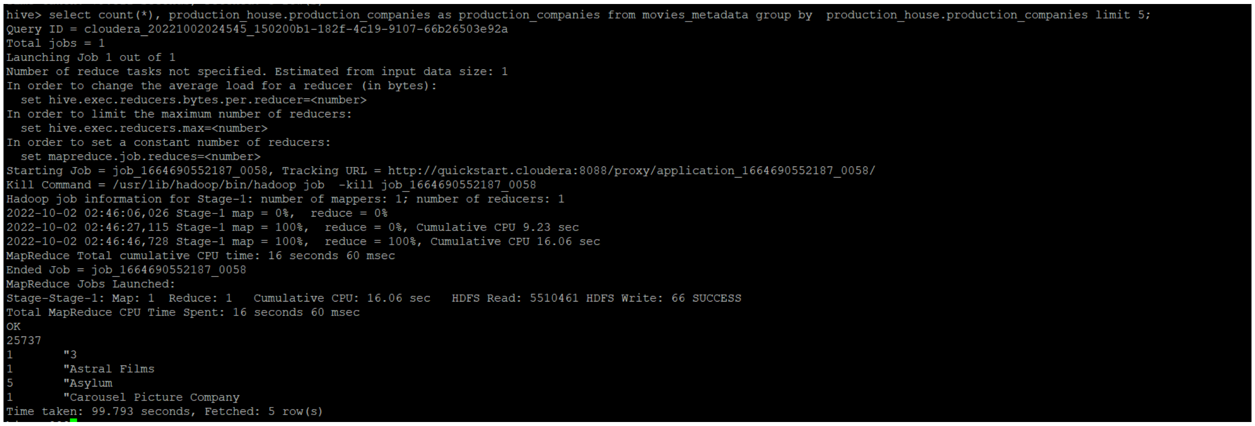
**Find out number of movies produced of production companies**

**Hive > select count(*) over(partitioned by production_house.production_companies) from movies_metadata;**



```
hive> select count(*) over(partition by production_house.production_companies), production_house.production_companies from movies_metadata limit 5;
Query ID = cloudera_20221002023939_f9c0fbd8-8a5b-40a9-9898-d3a849ac11e2
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_1664690552187_0055, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0055/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0055
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-02 02:39:53,097 Stage-1 map = 0%,  reduce = 0%
2022-10-02 02:40:15,371 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 10.04 sec
2022-10-02 02:40:44,483 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 26.51 sec
MapReduce Total cumulative CPU time: 26 seconds 510 msec
Ended Job = job_1664690552187_0055
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 26.51 sec   HDFS Read: 5511713 HDFS Write: 35 SUCCESS
Total MapReduce CPU Time Spent: 26 seconds 510 msec
OK
25737
25737
25737
25737
25737
Time taken: 79.311 seconds, Fetched: 5 row(s)
hive>
```

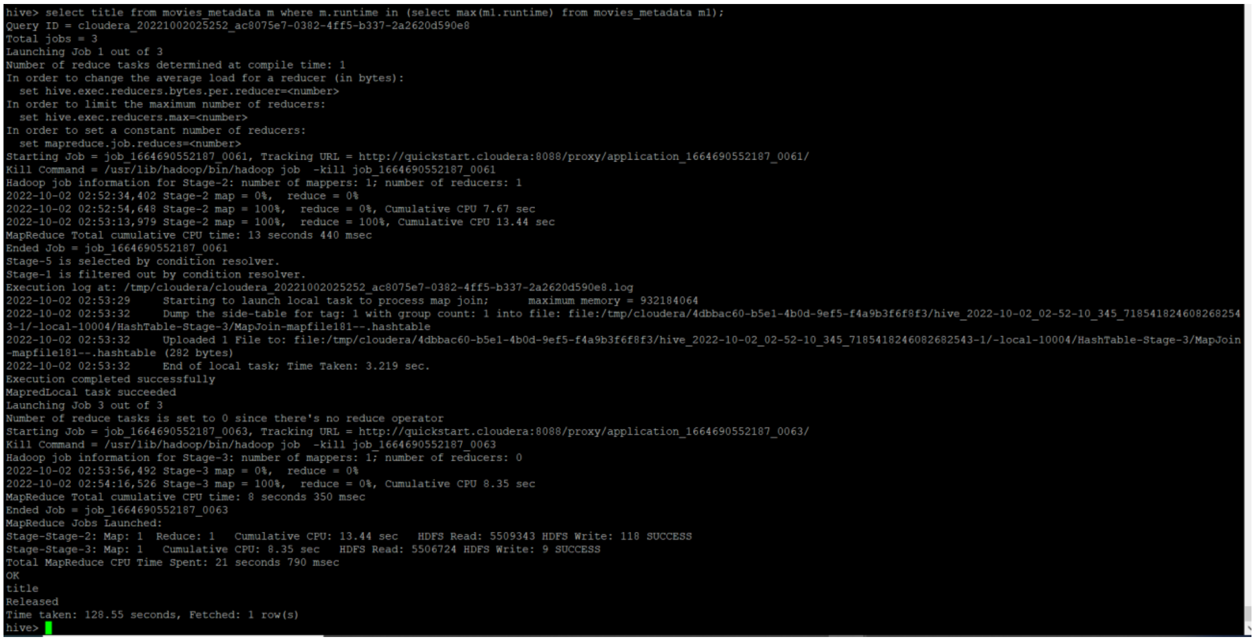
**Hive > select count(*), production_house.production_companies as production_companies from movies_metadata group by  production_house.production_companies limit 5;**



```
hive> select count(*), production_house.production_companies as production_companies from movies_metadata group by  production_house.production_companies limit 5;
Query ID = cloudera_20221002024545_150200b1-182f-4c19-9107-66b26503e92a
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_1664690552187_0058, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0058/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0058
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-02 02:46:06,026 Stage-1 map = 0%,  reduce = 0%
2022-10-02 02:46:27,115 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.23 sec
2022-10-02 02:46:46,728 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 16.06 sec
MapReduce Total cumulative CPU time: 16 seconds 60 msec
Ended Job = job_1664690552187_0058
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 16.06 sec   HDFS Read: 5510461 HDFS Write: 66 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 60 msec
OK
25737
1       "3
1       "Astral Films
5       "Asylum
1       "Carousel Picture Company
Time taken: 99.793 seconds, Fetched: 5 row(s)
```

**Find the name of movie which is longest duration**

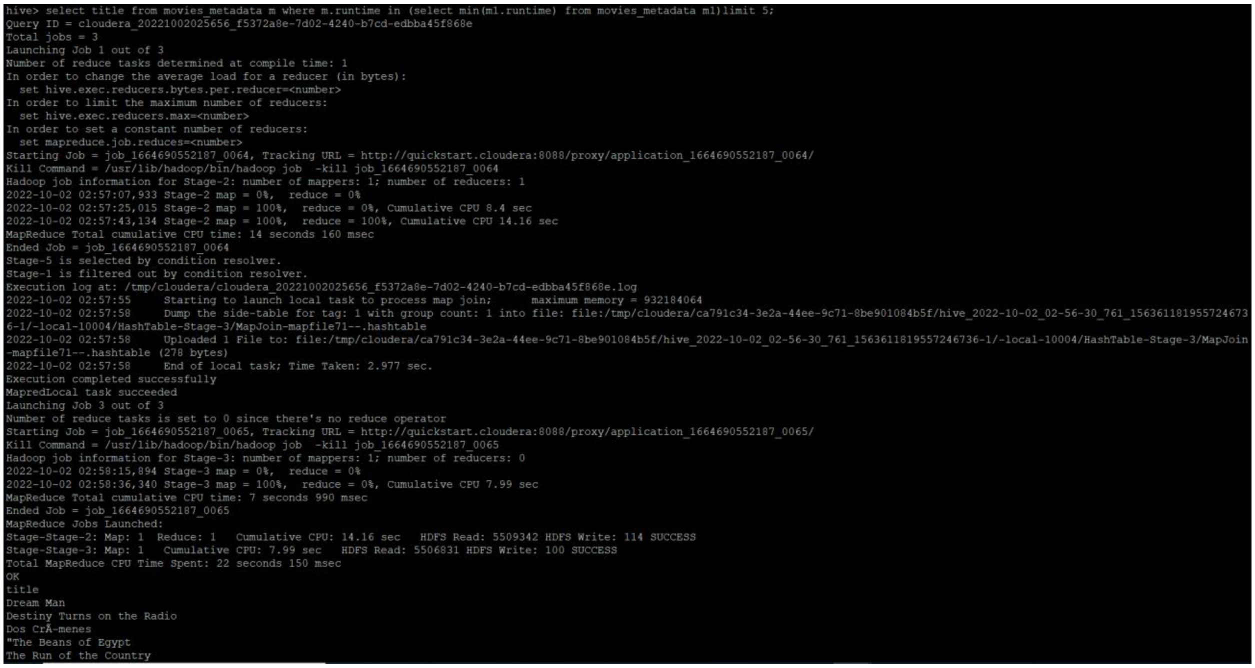
**Hive > select title from movies_metadata m where m.runtime in (select max(m1.runtime) from movies_metadata m1);**



```
hive> select title from movies_metadata m where m.runtime in (select max(m1.runtime) from movies_metadata m1);
Query ID = cloudera_20221002025252_ac8075e7-0382-4ff5-b337-2a2620d590e8
Total jobs = 3
Launching Job 1 out of 3
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_1664690552187_0061, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0061/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0061
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-02 02:52:34,402 Stage-2 map = 0%,  reduce = 0%
2022-10-02 02:52:54,648 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 7.67 sec
2022-10-02 02:53:13,979 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 13.44 sec
MapReduce Total cumulative CPU time: 13 seconds 440 msec
Ended Job = job_1664690552187_0061
Stage-5 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution log at: /tmp/cloudera/cloudera_20221002025252_ac8075e7-0382-4ff5-b337-2a2620d590e8.log
2022-10-02 02:53:29     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 02:53:32     Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-52-10_345_718541824608268254
3-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile181--.hashtable
2022-10-02 02:53:32     Uploaded 1 File to: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-52-10_345_7185418246082682543-1/-local-10004/HashTable-Stage-3/MapJoin
-mapfile181--.hashtable (282 bytes)
2022-10-02 02:53:32     End of local task; Time Taken: 3.219 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0063, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0063/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0063
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-10-02 02:53:56,492 Stage-3 map = 0%,  reduce = 0%
2022-10-02 02:54:16,526 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 8.35 sec
MapReduce Total cumulative CPU time: 8 seconds 350 msec
Ended Job = job_1664690552187_0063
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 13.44 sec   HDFS Read: 5509343 HDFS Write: 118 SUCCESS
Stage-Stage-3: Map: 1   Cumulative CPU: 8.35 sec   HDFS Read: 5506724 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 21 seconds 790 msec
OK
title
Released
Time taken: 128.55 seconds, Fetched: 1 row(s)
hive>
```

**Find the name of movie which is very short in duration**

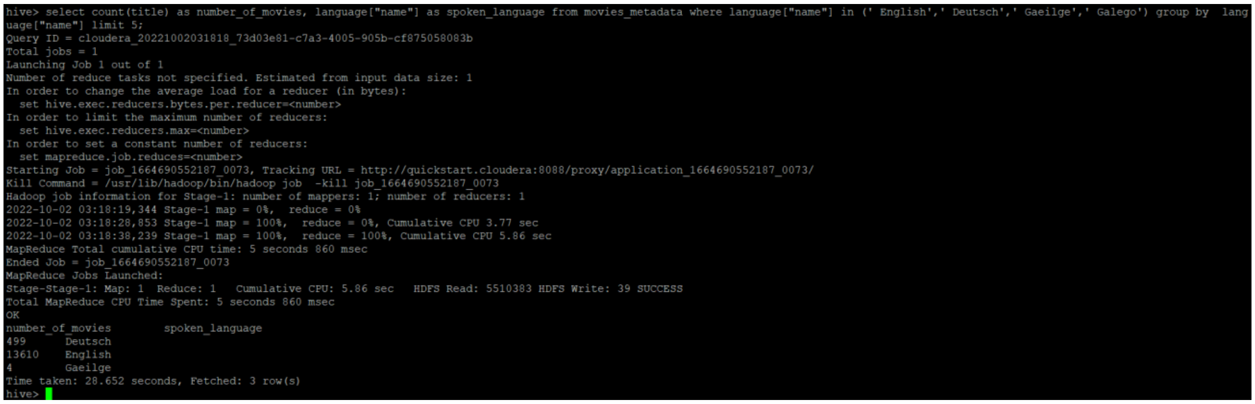
**Hive > select title from movies_metadata m where m.runtime in (select min(m1.runtime) from movies_metadata m1);**



```
hive> select title from movies_metadata m where m.runtime in (select min(m1.runtime) from movies_metadata m1)limit 5;
Query ID = cloudera_20221002025656_f5372a8e-7d02-4240-b7cd-edbba45f868e
Total jobs = 3
Launching Job 1 out of 3
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_1664690552187_0064, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0064/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0064
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-02 02:57:07,933 Stage-2 map = 0%,  reduce = 0%
2022-10-02 02:57:25,015 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 8.4 sec
2022-10-02 02:57:43,134 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 14.16 sec
MapReduce Total cumulative CPU time: 14 seconds 160 msec
Ended Job = job_1664690552187_0064
Stage-5 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution log at: /tmp/cloudera/cloudera_20221002025656_f5372a8e-7d02-4240-b7cd-edbba45f868e.log
2022-10-02 02:57:55     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 02:57:58     Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_02-56-30_761_156361181955724673
6-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile71--.hashtable
2022-10-02 02:57:58     Uploaded 1 File to: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_02-56-30_761_1563611819557246736-1/-local-10004/HashTable-Stage-3/MapJoin
-mapfile71--.hashtable (278 bytes)
2022-10-02 02:57:58     End of local task; Time Taken: 2.977 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0065, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0065/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0065
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-10-02 02:58:15,894 Stage-3 map = 0%,  reduce = 0%
2022-10-02 02:58:36,340 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 7.99 sec
MapReduce Total cumulative CPU time: 7 seconds 990 msec
Ended Job = job_1664690552187_0065
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 14.16 sec   HDFS Read: 5509342 HDFS Write: 114 SUCCESS
Stage-Stage-3: Map: 1   Cumulative CPU: 7.99 sec   HDFS Read: 5506831 HDFS Write: 100 SUCCESS
Total MapReduce CPU Time Spent: 22 seconds 150 msec
OK
title
Dream Man
Destiny Turns on the Radio
Dos CrÃ-menes
"The Beans of Egypt
The Run of the Country
```

**Find the number of movies according to spoken language (English,Deutsch,Gaeilge,Galego)**

**Hive > select count(title) as number_of_movies, language["name"] as spoken_language from movies_metadata where language["name"] in (' English',' Deutsch',' Gaeilge',' Galego') group by  language["name"] limit 5;**



```
hive> select count(title) as number_of_movies, language["name"] as spoken_language from movies_metadata where language["name"] in (' English',' Deutsch',' Gaeilge',' Galego') group by  lang
uage["name"] limit 5;
Query ID = cloudera_20221002031818_73d03e81-c7a3-4005-905b-cf875058083b
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_1664690552187_0073, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0073/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0073
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-02 03:18:19,344 Stage-1 map = 0%,  reduce = 0%
2022-10-02 03:18:28,853 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.77 sec
2022-10-02 03:18:38,239 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.86 sec
MapReduce Total cumulative CPU time: 5 seconds 860 msec
Ended Job = job_1664690552187_0073
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.86 sec   HDFS Read: 5510383 HDFS Write: 39 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 860 msec
OK
number_of_movies        spoken_language
499     Deutsch
13610   English
4       Gaeilge
Time taken: 28.652 seconds, Fetched: 3 row(s)
hive> 
```

## Join Operation

1) **Find the name of movie in which cast is Woody(voice)**

**Hive > select m.title from movies_metadata m join**

**credits_partitioned c on (m.id = c.id)**

**where casts.name = 'TomHanks';**

**To improve the performance we can use either**

**Hive > select /*+ MAPJOIN(m) */m.title as name_of_movie from credits_partitioned c join**

**movies_metadata m on (c.id = m.id)**

**where casts.name = 'TomHanks';**

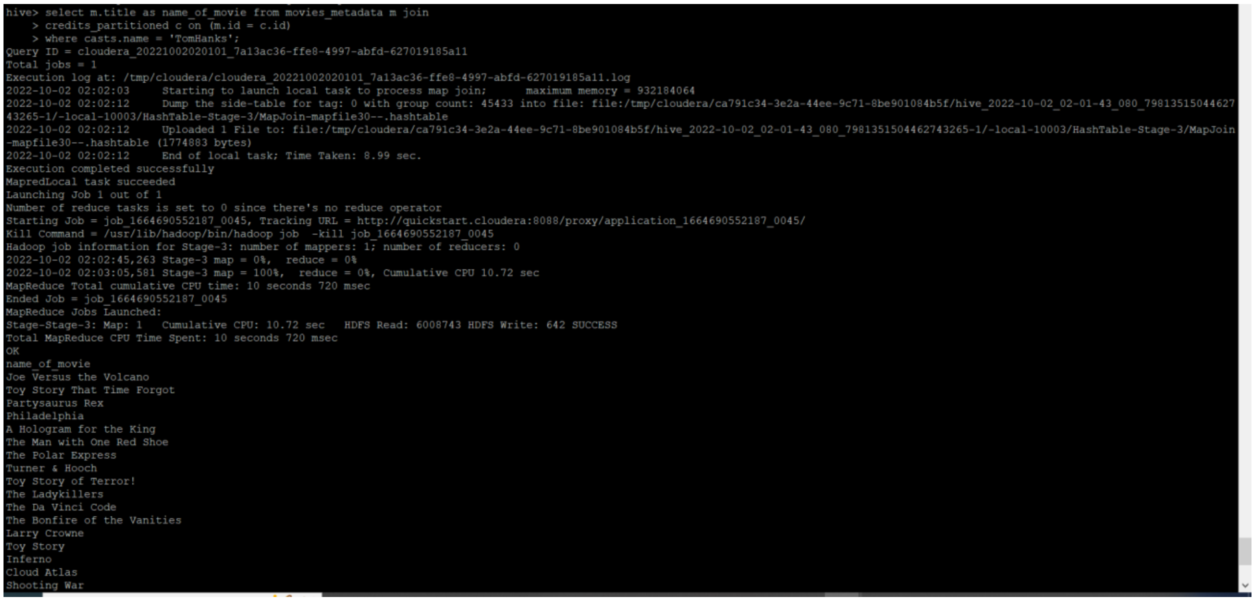
**Or**

**set hive.auto.convert.join=true;**

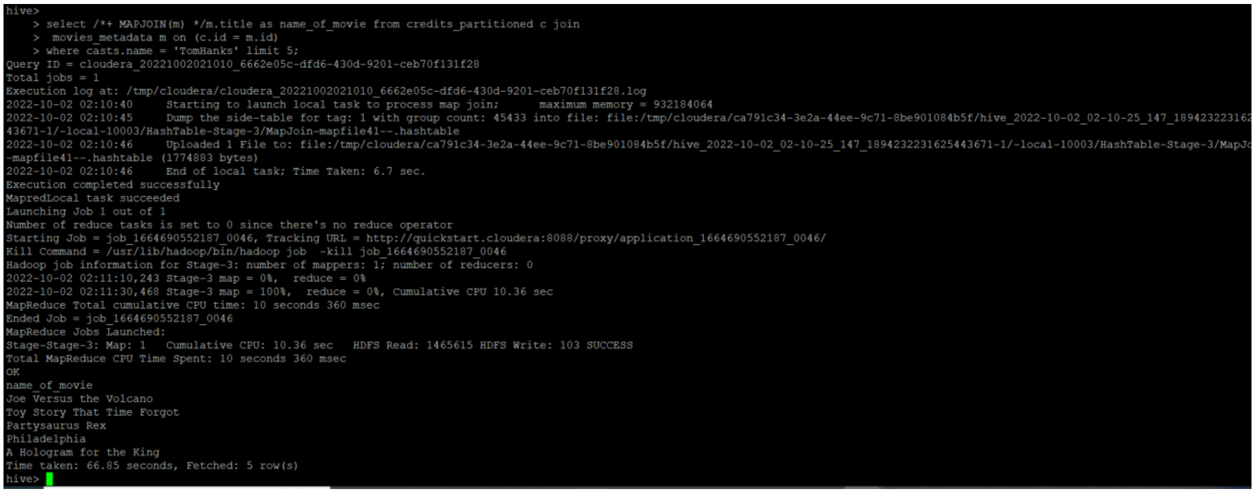
**Optimize Auto Join Conversion**

**set hive.auto.convert.join.noconditionaltask = true;**

**set hive.auto.convert.join.noconditionaltask.size = 10000;**



```
hive> select m.title as name_of_movie from movies_metadata m join
    > credits_partitioned c on (m.id = c.id)
    > where casts.name = 'TomHanks';
Query ID = cloudera_20221002020101_7a13ac36-ffe8-4997-abfd-627019185a11
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20221002020101_7a13ac36-ffe8-4997-abfd-627019185a11.log
2022-10-02 02:02:03     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 02:02:12     Dump the side-table for tag: 0 with group count: 45433 into file: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_02-01-43_080_79813515044627
43265-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile30--.hashtable
2022-10-02 02:02:12     Uploaded 1 File to: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_02-01-43_080_7981351504462743265-1/-local-10003/HashTable-Stage-3/MapJoin
-mapfile30--.hashtable (1774883 bytes)
2022-10-02 02:02:12     End of local task; Time Taken: 8.99 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0045, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0045/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0045
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-10-02 02:02:45,263 Stage-3 map = 0%,  reduce = 0%
2022-10-02 02:03:05,581 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 10.72 sec
MapReduce Total cumulative CPU time: 10 seconds 720 msec
Ended Job = job_1664690552187_0045
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 10.72 sec   HDFS Read: 6008743 HDFS Write: 642 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 720 msec
OK
name_of_movie
Joe Versus the Volcano
Toy Story That Time Forgot
Partysaurus Rex
Philadelphia
A Hologram for the King
The Man with One Red Shoe
The Polar Express
Turner & Hooch
Toy Story of Terror!
The Ladykillers
The Da Vinci Code
The Bonfire of the Vanities
Larry Crowne
Toy Story
Inferno
Cloud Atlas
Shooting War
```



```
hive>
    > select /*+ MAPJOIN(m) */m.title as name_of_movie from credits_partitioned c join
    >  movies_metadata m on (c.id = m.id)
    > where casts.name = 'TomHanks' limit 5;
Query ID = cloudera_20221002021010_6662e05c-dfd6-430d-9201-ceb70f131f28
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20221002021010_6662e05c-dfd6-430d-9201-ceb70f131f28.log
2022-10-02 02:10:40     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 02:10:45     Dump the side-table for tag: 1 with group count: 45433 into file: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_02-10-25_147_189423223162
43671-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile41--.hashtable
2022-10-02 02:10:46     Uploaded 1 File to: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_02-10-25_147_1894232231625443671-1/-local-10003/HashTable-Stage-3/MapJo
-mapfile41--.hashtable (1774883 bytes)
2022-10-02 02:10:46     End of local task; Time Taken: 6.7 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0046, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0046/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0046
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-10-02 02:11:10,243 Stage-3 map = 0%,  reduce = 0%
2022-10-02 02:11:30,468 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 10.36 sec
MapReduce Total cumulative CPU time: 10 seconds 360 msec
Ended Job = job_1664690552187_0046
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 10.36 sec   HDFS Read: 1465615 HDFS Write: 103 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 360 msec
OK
name_of_movie
Joe Versus the Volcano
Toy Story That Time Forgot
Partysaurus Rex
Philadelphia
A Hologram for the King
Time taken: 66.85 seconds, Fetched: 5 row(s)
hive>
```

2) **Find the name of movie directed by director HowardDeutch which has rating between 3 and 5**

**We can use /*+ STREAMTABLE(large_table_name) */ to stream large table data into reducer side that increase the performance in query execution**

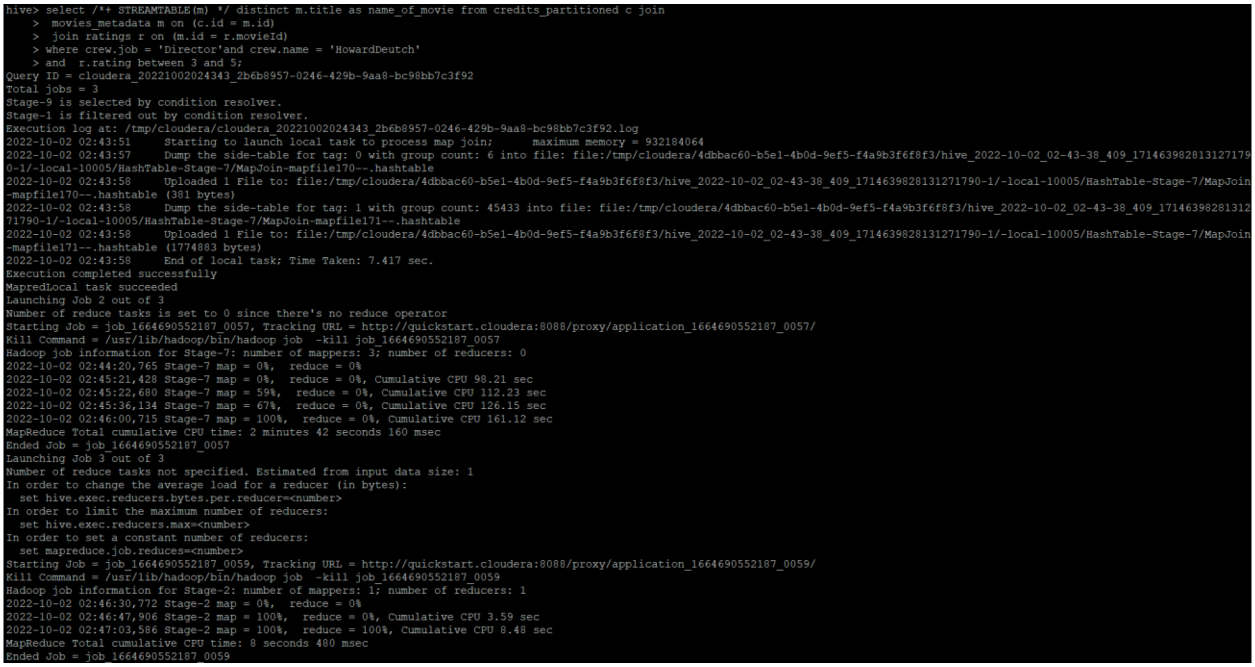
**Hive > select /*+ STREAMTABLE(m) */ distinct m.title as name_of_movie from credits_partitioned c join**

**movies_metadata m on (c.id = m.id)**

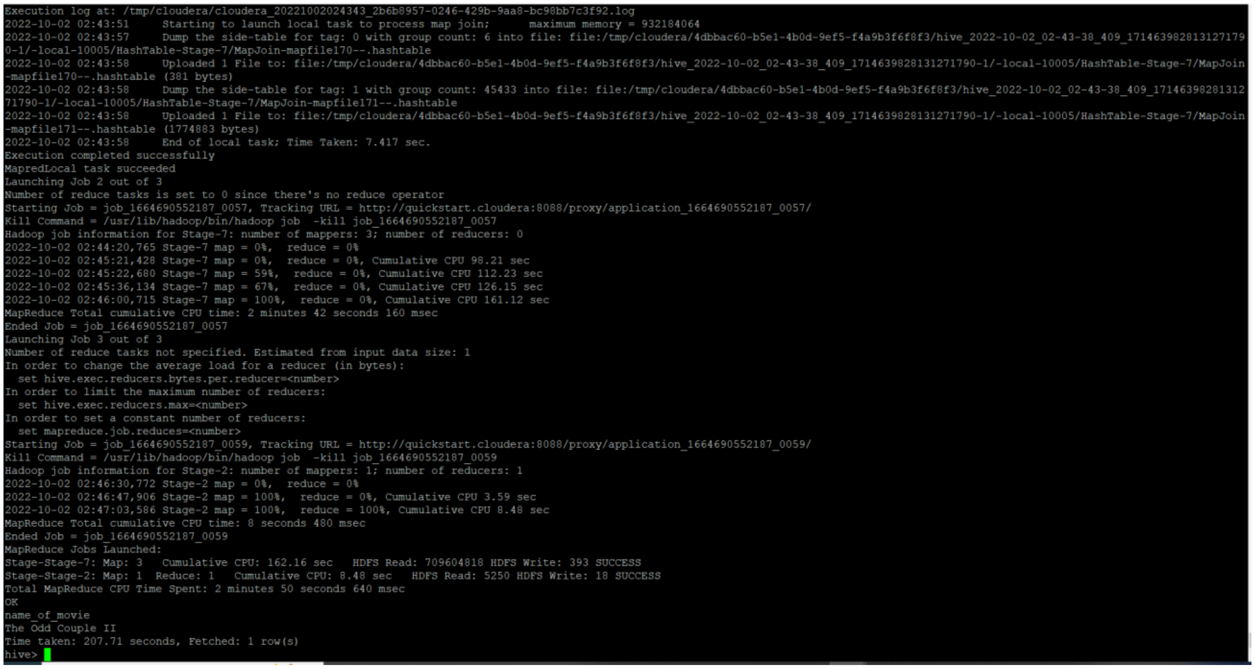
**join ratings r on (m.id = r.movieId)**

**where crew.job = 'Director'and crew.name = 'HowardDeutch'**

**and r.rating between 3 and 5;**



```
hive> select /*+ STREAMTABLE(m) */ distinct m.title as name_of_movie from credits_partitioned c join
    >  movies_metadata m on (c.id = m.id)
    >  join ratings r on (m.id = r.movieId)
    > where crew.job = 'Director'and crew.name = 'HowardDeutch'
    > and  r.rating between 3 and 5;
Query ID = cloudera_20221002024343_2b6b8957-0246-429b-9aa8-bc98bb7c3f92
Total jobs = 3
Stage-9 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution log at: /tmp/cloudera/cloudera_20221002024343_2b6b8957-0246-429b-9aa8-bc98bb7c3f92.log
2022-10-02 02:43:51     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 02:43:57     Dump the side-table for tag: 0 with group count: 6 into file: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-43-38_409_171463982813127179
0-1/-local-10005/HashTable-Stage-7/MapJoin-mapfile170--.hashtable
2022-10-02 02:43:58     Uploaded 1 File to: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-43-38_409_1714639828131271790-1/-local-10005/HashTable-Stage-7/MapJoin
-mapfile170--.hashtable (381 bytes)
2022-10-02 02:43:58     Dump the side-table for tag: 1 with group count: 45433 into file: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-43-38_409_17146398281312
71790-1/-local-10005/HashTable-Stage-7/MapJoin-mapfile171--.hashtable
2022-10-02 02:43:58     Uploaded 1 File to: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-43-38_409_1714639828131271790-1/-local-10005/HashTable-Stage-7/MapJoin
-mapfile171--.hashtable (1774883 bytes)
2022-10-02 02:43:58     End of local task; Time Taken: 7.417 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0057, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0057/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0057
Hadoop job information for Stage-7: number of mappers: 3; number of reducers: 0
2022-10-02 02:44:20,765 Stage-7 map = 0%,  reduce = 0%
2022-10-02 02:45:21,428 Stage-7 map = 0%,  reduce = 0%, Cumulative CPU 98.21 sec
2022-10-02 02:45:22,680 Stage-7 map = 59%,  reduce = 0%, Cumulative CPU 112.23 sec
2022-10-02 02:45:36,134 Stage-7 map = 67%,  reduce = 0%, Cumulative CPU 126.15 sec
2022-10-02 02:46:00,715 Stage-7 map = 100%,  reduce = 0%, Cumulative CPU 161.12 sec
MapReduce Total cumulative CPU time: 2 minutes 42 seconds 160 msec
Ended Job = job_1664690552187_0057
Launching Job 3 out of 3
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_1664690552187_0059, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0059/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0059
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-02 02:46:30,772 Stage-2 map = 0%,  reduce = 0%
2022-10-02 02:46:47,906 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.59 sec
2022-10-02 02:47:03,586 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 8.48 sec
MapReduce Total cumulative CPU time: 8 seconds 480 msec
Ended Job = job_1664690552187_0059
```



```
Execution log at: /tmp/cloudera/cloudera_20221002024343_2b6b8957-0246-429b-9aa8-bc98bb7c3f92.log
2022-10-02 02:43:51     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 02:43:57     Dump the side-table for tag: 0 with group count: 6 into file: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-43-38_409_171463982813127179
0-1/-local-10005/HashTable-Stage-7/MapJoin-mapfile170--.hashtable
2022-10-02 02:43:58     Uploaded 1 File to: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-43-38_409_1714639828131271790-1/-local-10005/HashTable-Stage-7/MapJoin
-mapfile170--.hashtable (381 bytes)
2022-10-02 02:43:58     Dump the side-table for tag: 1 with group count: 45433 into file: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-43-38_409_17146398281312
71790-1/-local-10005/HashTable-Stage-7/MapJoin-mapfile171--.hashtable
2022-10-02 02:43:58     Uploaded 1 File to: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-43-38_409_1714639828131271790-1/-local-10005/HashTable-Stage-7/MapJoin
-mapfile171--.hashtable (1774883 bytes)
2022-10-02 02:43:58     End of local task; Time Taken: 7.417 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1664690552187_0057, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0057/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0057
Hadoop job information for Stage-7: number of mappers: 3; number of reducers: 0
2022-10-02 02:44:20,765 Stage-7 map = 0%,  reduce = 0%
2022-10-02 02:45:21,428 Stage-7 map = 0%,  reduce = 0%, Cumulative CPU 98.21 sec
2022-10-02 02:45:22,680 Stage-7 map = 59%,  reduce = 0%, Cumulative CPU 112.23 sec
2022-10-02 02:45:36,134 Stage-7 map = 67%,  reduce = 0%, Cumulative CPU 126.15 sec
2022-10-02 02:46:00,715 Stage-7 map = 100%,  reduce = 0%, Cumulative CPU 161.12 sec
MapReduce Total cumulative CPU time: 2 minutes 42 seconds 160 msec
Ended Job = job_1664690552187_0057
Launching Job 3 out of 3
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_1664690552187_0059, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0059/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0059
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-02 02:46:30,772 Stage-2 map = 0%,  reduce = 0%
2022-10-02 02:46:47,906 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.59 sec
2022-10-02 02:47:03,586 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 8.48 sec
MapReduce Total cumulative CPU time: 8 seconds 480 msec
Ended Job = job_1664690552187_0059
```

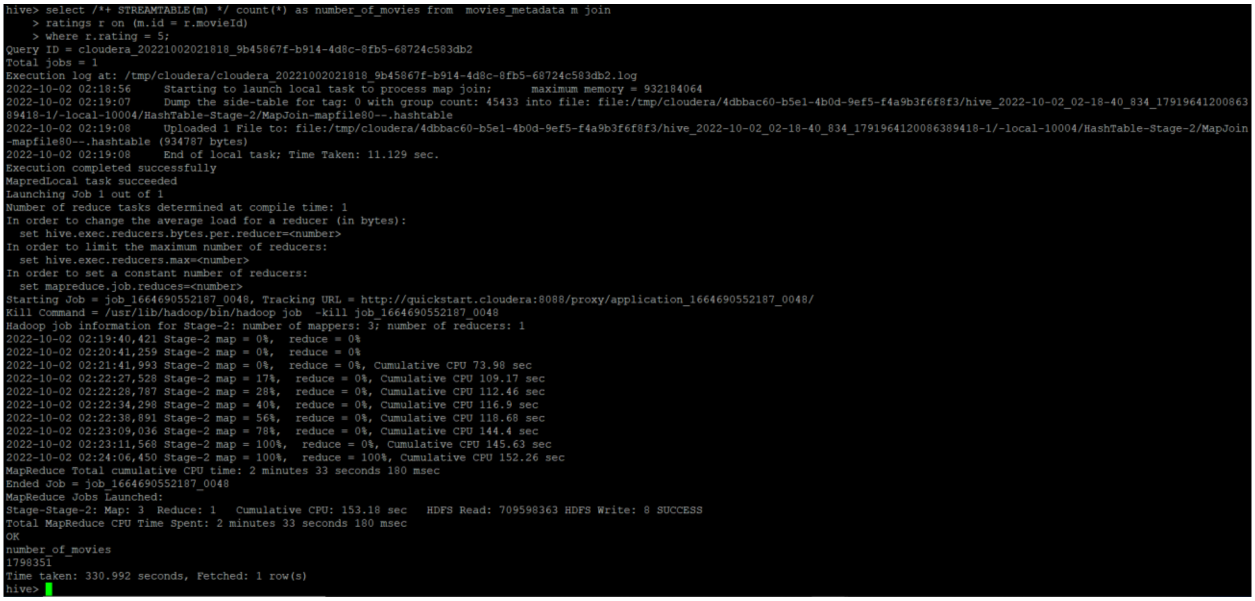
```
MapReduce Jobs Launched:
Stage-Stage-7: Map: 3   Cumulative CPU: 162.16 sec   HDFS Read: 709604818 HDFS Write: 393 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 8.48 sec   HDFS Read: 5250 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 50 seconds 640 msec
OK
name_of_movie
The Odd Couple II
Time taken: 207.71 seconds, Fetched: 1 row(s)
hive>
```

### 3) Find the number of movies which has rating 5

**Hive > select /*+ STREAMTABLE(m) */ count(*) as number_of_movies from movies_metadata m join**

**ratings r on (m.id = r.movieId)**

**where r.rating = 5;**



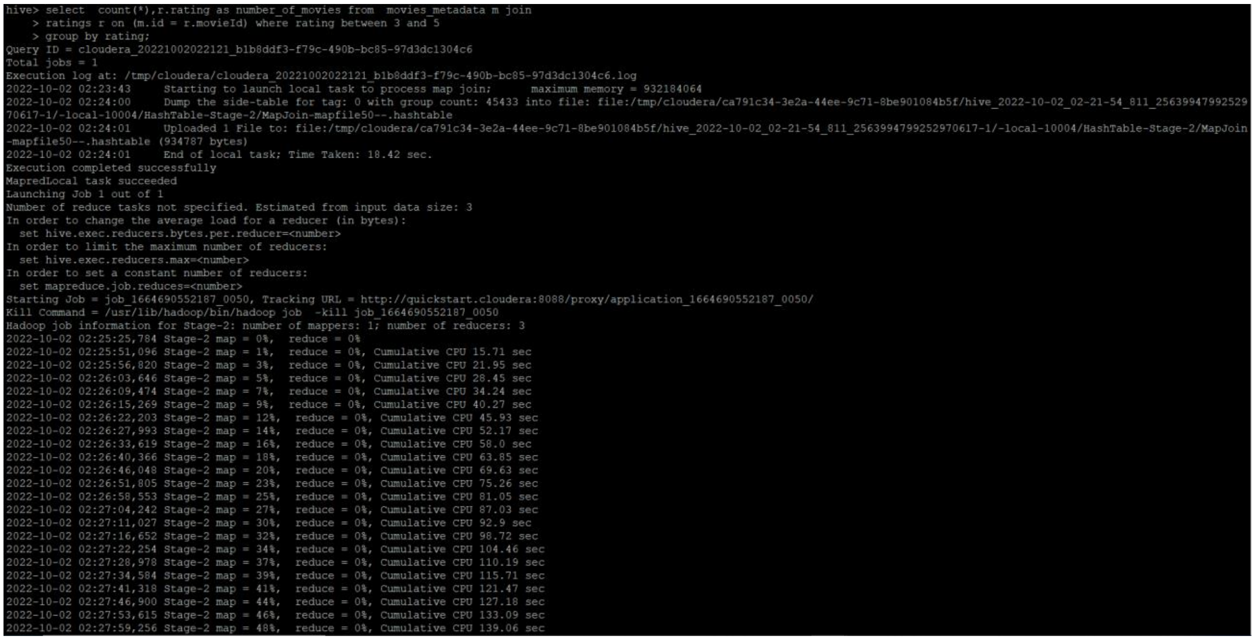
```
hive> select /*+ STREAMTABLE(m) */ count(*) as number_of_movies from  movies_metadata m join
    > ratings r on (m.id = r.movieId)
    > where r.rating = 5;
Query ID = cloudera_20221002021818_9b45867f-b914-4d8c-8fb5-68724c583db2
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20221002021818_9b45867f-b914-4d8c-8fb5-68724c583db2.log
2022-10-02 02:18:56     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 02:19:07     Dump the side-table for tag: 0 with group count: 45433 into file: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-18-40_834_17919641200863
89418-1/-local-10004/HashTable-Stage-2/MapJoin-mapfile80--.hashtable
2022-10-02 02:19:08     Uploaded 1 File to: file:/tmp/cloudera/4dbbac60-b5e1-4b0d-9ef5-f4a9b3f6f8f3/hive_2022-10-02_02-18-40_834_1791964120086389418-1/-local-10004/HashTable-Stage-2/MapJoin
-mapfile80--.hashtable (934787 bytes)
2022-10-02 02:19:08     End of local task; Time Taken: 11.129 sec.
Execution completed successfully
MapredLocal task succeeded
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_1664690552187_0048, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0048/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0048
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 1
2022-10-02 02:19:40,421 Stage-2 map = 0%,  reduce = 0%
2022-10-02 02:20:41,259 Stage-2 map = 0%,  reduce = 0%
2022-10-02 02:21:41,993 Stage-2 map = 0%,  reduce = 0%, Cumulative CPU 73.98 sec
2022-10-02 02:22:27,528 Stage-2 map = 17%,  reduce = 0%, Cumulative CPU 109.17 sec
2022-10-02 02:22:28,787 Stage-2 map = 28%,  reduce = 0%, Cumulative CPU 112.46 sec
2022-10-02 02:22:34,298 Stage-2 map = 40%,  reduce = 0%, Cumulative CPU 116.9 sec
2022-10-02 02:22:38,891 Stage-2 map = 56%,  reduce = 0%, Cumulative CPU 118.68 sec
2022-10-02 02:23:09,036 Stage-2 map = 78%,  reduce = 0%, Cumulative CPU 144.4 sec
2022-10-02 02:23:11,568 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 145.63 sec
2022-10-02 02:24:06,450 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 152.26 sec
MapReduce Total cumulative CPU time: 2 minutes 33 seconds 180 msec
Ended Job = job_1664690552187_0048
MapReduce Jobs Launched:
Stage-Stage-2: Map: 3  Reduce: 1   Cumulative CPU: 153.18 sec   HDFS Read: 709598363 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 33 seconds 180 msec
OK
number_of_movies
1798351
Time taken: 330.992 seconds, Fetched: 1 row(s)
hive>
```

### 4) Find the number of movies according to rating

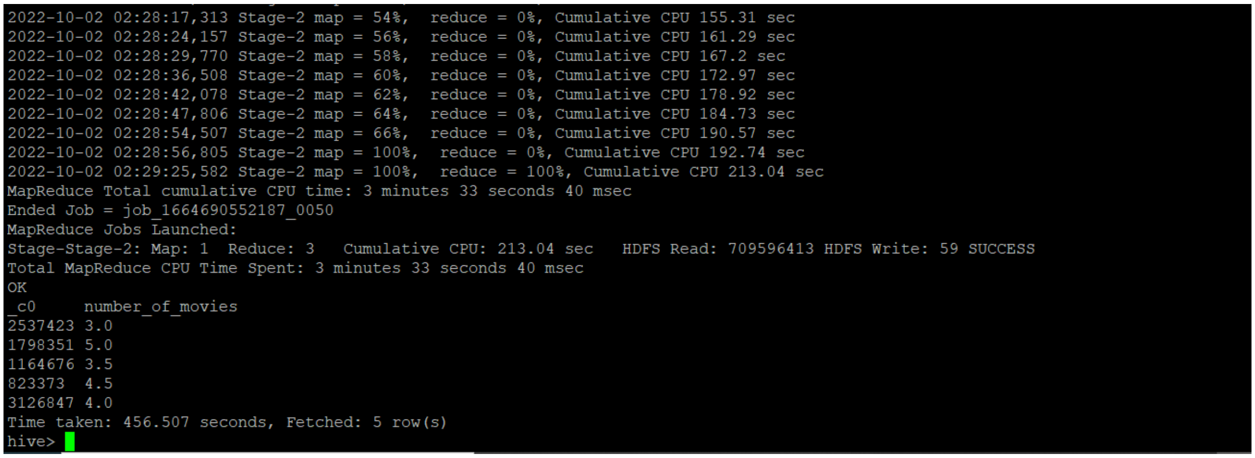
**Hive > select  count(*),r.rating as number_of_movies from  movies_metadata m join**

**ratings r on (m.id = r.movieId) where rating between 3 and 5**

**group by rating;**



```
hive> select  count(*),r.rating as number_of_movies from  movies_metadata m join
    > ratings r on (m.id = r.movieId) where rating between 3 and 5
    > group by rating;
Query ID = cloudera_20221002022121_b1b8ddf3-f79c-490b-bc85-97d3dc1304c6
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20221002022121_b1b8ddf3-f79c-490b-bc85-97d3dc1304c6.log
2022-10-02 02:23:43     Starting to launch local task to process map join;      maximum memory = 932184064
2022-10-02 02:24:00     Dump the side-table for tag: 0 with group count: 45433 into file: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_02-21-54_811_25639947992529
70617-1/-local-10004/HashTable-Stage-2/MapJoin-mapfile50--.hashtable
2022-10-02 02:24:01     Uploaded 1 File to: file:/tmp/cloudera/ca791c34-3e2a-44ee-9c71-8be901084b5f/hive_2022-10-02_02-21-54_811_2563994799252970617-1/-local-10004/HashTable-Stage-2/MapJoin
-mapfile50--.hashtable (934787 bytes)
2022-10-02 02:24:01     End of local task; Time Taken: 18.42 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 3
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_1664690552187_0050, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664690552187_0050/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664690552187_0050
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 3
2022-10-02 02:25:25,784 Stage-2 map = 0%,  reduce = 0%
2022-10-02 02:25:51,096 Stage-2 map = 1%,  reduce = 0%, Cumulative CPU 15.71 sec
2022-10-02 02:25:56,820 Stage-2 map = 3%,  reduce = 0%, Cumulative CPU 21.95 sec
2022-10-02 02:26:03,646 Stage-2 map = 5%,  reduce = 0%, Cumulative CPU 28.45 sec
2022-10-02 02:26:09,474 Stage-2 map = 7%,  reduce = 0%, Cumulative CPU 34.24 sec
2022-10-02 02:26:15,269 Stage-2 map = 9%,  reduce = 0%, Cumulative CPU 40.27 sec
2022-10-02 02:26:22,203 Stage-2 map = 12%,  reduce = 0%, Cumulative CPU 45.93 sec
2022-10-02 02:26:27,993 Stage-2 map = 14%,  reduce = 0%, Cumulative CPU 52.17 sec
2022-10-02 02:26:33,619 Stage-2 map = 16%,  reduce = 0%, Cumulative CPU 58.0 sec
2022-10-02 02:26:40,366 Stage-2 map = 18%,  reduce = 0%, Cumulative CPU 63.85 sec
2022-10-02 02:26:46,048 Stage-2 map = 20%,  reduce = 0%, Cumulative CPU 69.63 sec
2022-10-02 02:26:51,805 Stage-2 map = 23%,  reduce = 0%, Cumulative CPU 75.26 sec
2022-10-02 02:26:58,553 Stage-2 map = 25%,  reduce = 0%, Cumulative CPU 81.05 sec
2022-10-02 02:27:04,242 Stage-2 map = 27%,  reduce = 0%, Cumulative CPU 87.03 sec
2022-10-02 02:27:11,027 Stage-2 map = 30%,  reduce = 0%, Cumulative CPU 92.9 sec
2022-10-02 02:27:16,652 Stage-2 map = 32%,  reduce = 0%, Cumulative CPU 98.72 sec
2022-10-02 02:27:22,254 Stage-2 map = 34%,  reduce = 0%, Cumulative CPU 104.46 sec
2022-10-02 02:27:28,978 Stage-2 map = 37%,  reduce = 0%, Cumulative CPU 110.19 sec
2022-10-02 02:27:34,584 Stage-2 map = 39%,  reduce = 0%, Cumulative CPU 115.71 sec
2022-10-02 02:27:41,318 Stage-2 map = 41%,  reduce = 0%, Cumulative CPU 121.47 sec
2022-10-02 02:27:46,900 Stage-2 map = 44%,  reduce = 0%, Cumulative CPU 127.18 sec
2022-10-02 02:27:53,615 Stage-2 map = 46%,  reduce = 0%, Cumulative CPU 133.09 sec
2022-10-02 02:27:59,256 Stage-2 map = 48%,  reduce = 0%, Cumulative CPU 139.06 sec
```



```
2022-10-02 02:28:17,313 Stage-2 map = 54%,  reduce = 0%, Cumulative CPU 155.31 sec
2022-10-02 02:28:24,157 Stage-2 map = 56%,  reduce = 0%, Cumulative CPU 161.29 sec
2022-10-02 02:28:29,770 Stage-2 map = 58%,  reduce = 0%, Cumulative CPU 167.2 sec
2022-10-02 02:28:36,508 Stage-2 map = 60%,  reduce = 0%, Cumulative CPU 172.97 sec
2022-10-02 02:28:42,078 Stage-2 map = 62%,  reduce = 0%, Cumulative CPU 178.92 sec
2022-10-02 02:28:47,806 Stage-2 map = 64%,  reduce = 0%, Cumulative CPU 184.73 sec
2022-10-02 02:28:54,507 Stage-2 map = 66%,  reduce = 0%, Cumulative CPU 190.57 sec
2022-10-02 02:28:56,805 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 192.74 sec
2022-10-02 02:29:25,582 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 213.04 sec
MapReduce Total cumulative CPU time: 3 minutes 33 seconds 40 msec
Ended Job = job_1664690552187_0050
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 3   Cumulative CPU: 213.04 sec   HDFS Read: 709596413 HDFS Write: 59 SUCCESS
Total MapReduce CPU Time Spent: 3 minutes 33 seconds 40 msec
OK
_c0     number_of_movies
2537423 3.0
1798351 5.0
1164676 3.5
823373  4.5
3126847 4.0
Time taken: 456.507 seconds, Fetched: 5 row(s)
hive>
```