

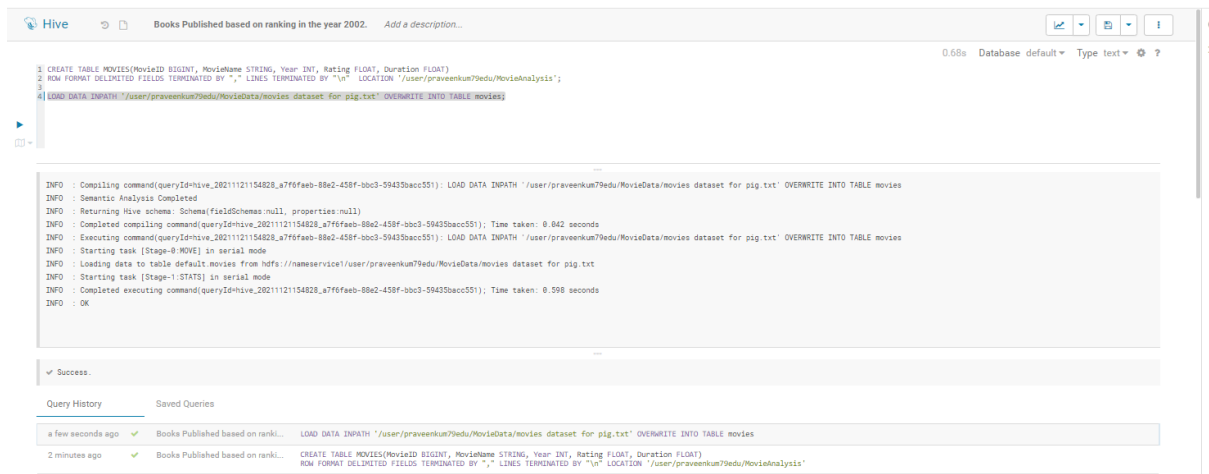
Movie Data Analysis

Creating tables and loading data into it

Create movies table and load data into it

```
CREATE TABLE MOVIES(MovieID BIGINT, MovieName STRING, Year INT, Rating FLOAT, Duration FLOAT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY "," LINES TERMINATED BY "\n" LOCATION
'/user/praveenkum79edu/MovieAnalysis';
```

```
LOAD DATA INPATH '/user/praveenkum79edu/MovieData/movies dataset for pig.txt' OVERWRITE INTO TABLE movies;
```



The screenshot shows the Hive console interface. The top bar displays the title "Books Published based on ranking in the year 2002." and the Hive logo. The main area contains the SQL commands entered:

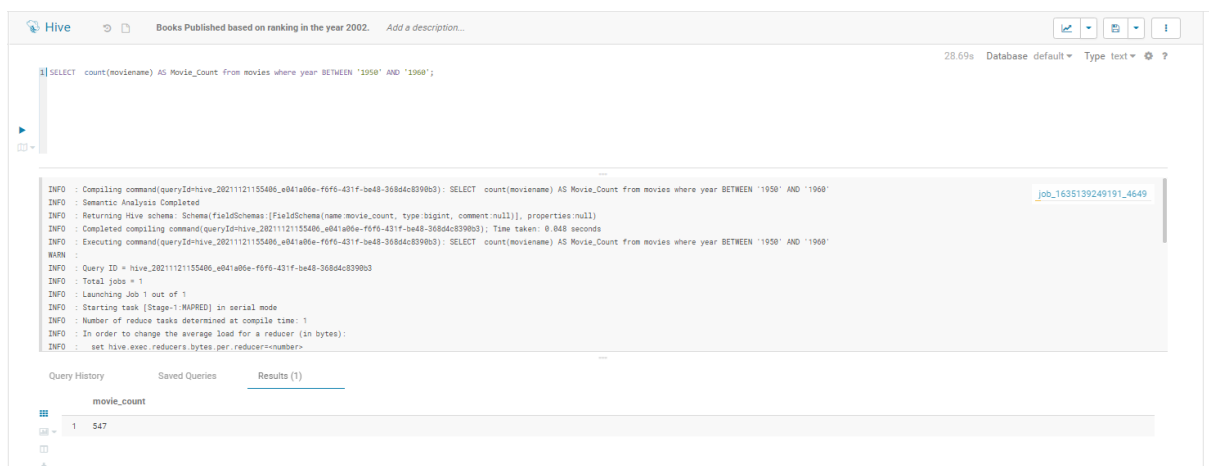
```
1 CREATE TABLE MOVIES(MovieID BIGINT, MovieName STRING, Year INT, Rating FLOAT, Duration FLOAT)
2 ROW FORMAT DELIMITED FIELDS TERMINATED BY "," LINES TERMINATED BY "\n" LOCATION '/user/praveenkum79edu/MovieAnalysis';
3
4 LOAD DATA INPATH '/user/praveenkum79edu/MovieData/movies dataset for pig.txt' OVERWRITE INTO TABLE movies;
```

 The output shows the execution details, including the compilation of the command, the semantic analysis, and the loading of data into the table. The status at the bottom is "Success."

Problem Statement

Find the number of movies released between 1950 and 1960

```
SELECT count(moviename) AS Movie_Count from movies where year BETWEEN '1950' AND '1960';
```



The screenshot shows the Hive console interface. The top bar displays the title "Books Published based on ranking in the year 2002." and the Hive logo. The main area contains the SQL query entered:

```
1 SELECT count(moviename) AS Movie_Count from movies where year BETWEEN '1950' AND '1960';
```

 The output shows the execution details, including the compilation of the command, the semantic analysis, and the execution of the query. The status at the bottom is "Success." The results are displayed in a table with the following data:

movie_count
1 547



Find the number of movies having rating more than 4

select count(moviename) As Movie_greater_4_rating from movies where rating > 4;

Hive interface showing the execution of the query: `select count(moviename) As Movie_greater_4_rating from movies where rating > 4;`

The interface displays the query execution details, including the command, schema, and the results. The results show a single row with the value 897 for the column `movie_greater_4_rating`.

```
INFO : Compiling command(queryId=hive_20211121155848_e4538658-5612-468f-a057-e09d0c1eeab0):
select count(moviename) As Movie_greater_4_rating from movies where rating > 4
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name=movie_greater_4_rating, type=bigint, comment=null)], properties=null)
INFO : Completed compiling command(queryId=hive_20211121155848_e4538658-5612-468f-a057-e09d0c1eeab0); Time taken: 0.054 seconds
INFO : Executing command(queryId=hive_20211121155848_e4538658-5612-468f-a057-e09d0c1eeab0):
select count(moviename) As Movie_greater_4_rating from movies where rating > 4
WARN :
INFO : Query ID = hive_20211121155848_e4538658-5612-468f-a057-e09d0c1eeab0
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Number of reduce tasks determined at compile time: 1
```

movie_greater_4_rating
897



Find the movies whose rating are between 3 and 4

select moviename As MovieName, rating from movies where rating between 3 AND 4;

Hive interface showing the execution of the query: `select moviename As MovieName, rating from movies where rating between 3 AND 4;`

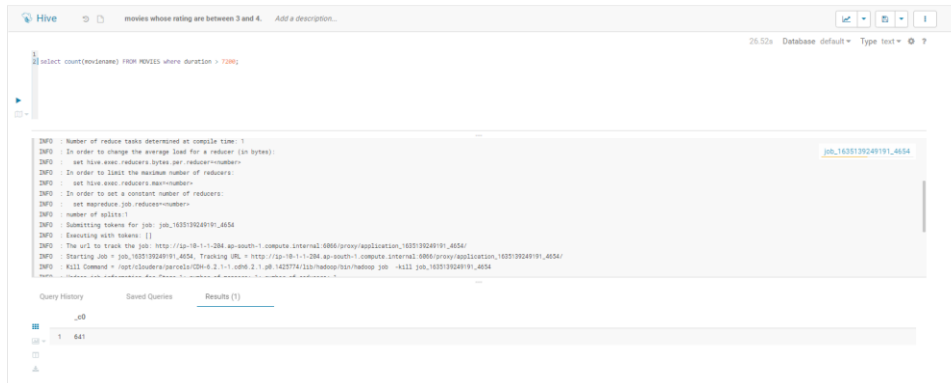
The interface displays the query execution details, including the command, schema, and the results. The results show a list of movies with their ratings, sorted by rating in descending order.

```
INFO : Compiling command(queryId=hive_20211121161519_ef702f34-988a-4f45-bc5d-13bd6eb13853):
select moviename As MovieName, rating from movies where rating between 3 AND 4
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name=moviename, type=string, comment=null), FieldSchema(name=rating, type=float, comment=null)], properties=null)
INFO : Completed compiling command(queryId=hive_20211121161519_ef702f34-988a-4f45-bc5d-13bd6eb13853); Time taken: 0.002 seconds
INFO : Executing command(queryId=hive_20211121161519_ef702f34-988a-4f45-bc5d-13bd6eb13853):
select moviename As MovieName, rating from movies where rating between 3 AND 4
WARN :
INFO : Query ID = hive_20211121161519_ef702f34-988a-4f45-bc5d-13bd6eb13853
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Number of reduce tasks is set to 0 since there's no reduce operator
```

moviename	rating
1 The Nightmare Before Christmas	3.9
2 The Mummy	3.5
3 Orphans of the Storm	3.2
4 One Magic Christmas	3.8
5 Murie's Wedding	3.5
6 Mother's Boys	3.4
7 Nosferatu: Original Version	3.5
8 Nick of Time	3.4
9 Broken Blossoms	3.3
10 Big Night	3.6
11 The Boys from Brazil	3.6
12 The Breakfast Club	4
13 The Bride of Frankenstein	3.7

Find the number of movies with duration more than 2 hours (7200 second)

select count(moviename) FROM MOVIES where duration > 7200;



The screenshot shows the Hive web interface. The query entered is `select count(moviename) FROM MOVIES where duration > 7200;`. The execution time is 26.52s. The results section shows a single row with the value 641.

```
1
2 select count(moviename) FROM MOVIES where duration > 7200;
```

26.52s Database: default Type: text

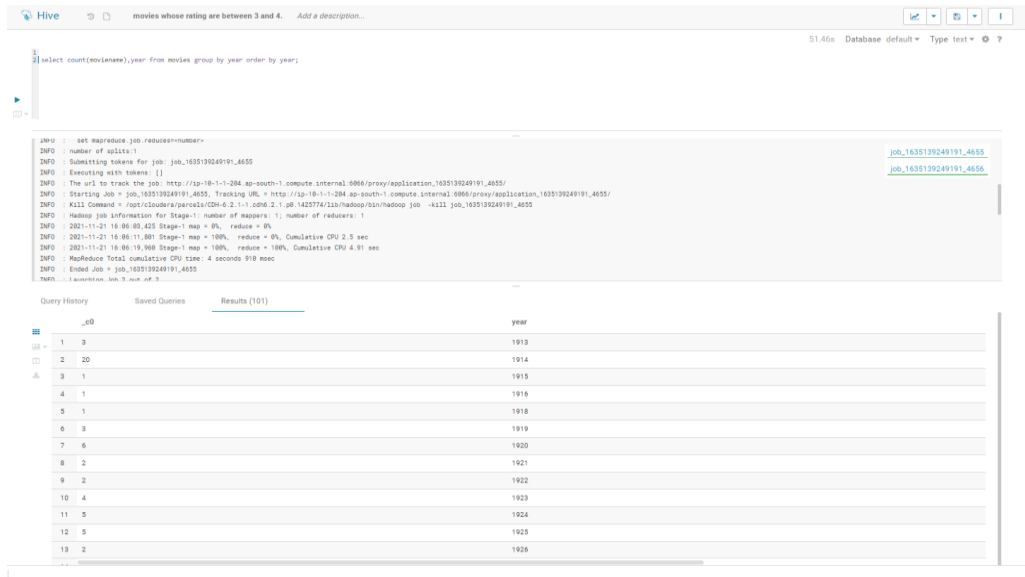
INFO : Number of reduce tasks detected at compile time: 1
INFO : In order to change the average load for a reducer (in bytes):
INFO : set hive.exec.reducers.bytes.per.reducer=number=
INFO : In order to limit the maximum number of reducers:
INFO : set hive.exec.reducers.max=number=
INFO : In order to set a constant number of reducers:
INFO : set mapreduce.job.reducers=number=
INFO : number of splits: 1
INFO : Submitting tokens for job: job_1635139249191_4654
INFO : Executing with tokens: []
INFO : The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:8086/proxy/application,1635139249191_4654/
INFO : Starting job = job_1635139249191_4654, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:8086/proxy/application,1635139249191_4654/
INFO : Kill Command = /opt/clooudera/parcels/CDH-6.2.1-1.cdh6.2.1.jb.1425774/lib/hadoop/bin/hadoop job -kill job_1635139249191_4654
WARN : Reduce job information for Stage-1: number of mappers: 1
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO : 2021-11-21 18:06:00.625 Stage-1 map = 0%, reduce = 0%, Cumulative CPU 2.5 sec
INFO : 2021-11-21 18:06:19.960 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.91 sec
INFO : MapReduce Total cumulative CPU time: 4 seconds 910 msec
INFO : Ended Job = job_1635139249191_4654
WARN : LaunchTask Job 3: succ ret 5

Query History Saved Queries Results (1)

	count
1	641

Find the list of years and number of movies released each year

select count(moviename),year from movies group by year order by year;



The screenshot shows the Hive web interface. The query entered is `select count(moviename),year from movies group by year order by year;`. The execution time is 51.46s. The results section shows a table with 101 rows, each containing a year and its corresponding count.

```
1
2 select count(moviename),year from movies group by year order by year;
```

51.46s Database: default Type: text

INFO : set mapreduce.job.reducers=number=
INFO : number of splits: 1
INFO : Submitting tokens for job: job_1635139249191_4655
INFO : Executing with tokens: []
INFO : The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:8086/proxy/application,1635139249191_4655/
INFO : Starting job = job_1635139249191_4655, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:8086/proxy/application,1635139249191_4655/
INFO : Kill Command = /opt/clooudera/parcels/CDH-6.2.1-1.cdh6.2.1.jb.1425774/lib/hadoop/bin/hadoop job -kill job_1635139249191_4655
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO : 2021-11-21 18:06:00.625 Stage-1 map = 0%, reduce = 0%, Cumulative CPU 2.5 sec
INFO : 2021-11-21 18:06:19.960 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.91 sec
INFO : MapReduce Total cumulative CPU time: 4 seconds 910 msec
INFO : Ended Job = job_1635139249191_4655
WARN : LaunchTask Job 3: succ ret 5

Query History Saved Queries Results (101)

	count	year
1	3	1913
2	20	1914
3	1	1915
4	1	1916
5	1	1918
6	3	1919
7	6	1920
8	2	1921
9	2	1922
10	4	1923
11	5	1924
12	5	1925
13	2	1926



Find the total number of movies in the dataset

```
select count(moviename)from movies;
```

Hive

movies whose rating are between 3 and 4. Add a description...

27.41s Database default Type text ?

```
1 select count(moviename)from movies;
```

INFO : In order to limit the maximum number of reducers:
INFO : set hive.exec.reducers.max=<number>
INFO : In order to set a constant number of reducers:
INFO : set mapreduce.job.reducers=<number>
INFO : number of splits:1
INFO : Submitting tokens for job: job_1635139249191_4657
INFO : Executing with tokens: []
INFO : The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1635139249191_4657/
INFO : Starting Job = job_1635139249191_4657, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1635139249191_4657/
INFO : Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1-cdh6.2.1.jar:425774/lib/hadoop/bin/hadoop job -kill job_1635139249191_4657
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO : 2021-11-21 16:08:04.041 Stage-1 map = 0%, reduce = 0%
INFO : 2021-11-21 16:08:11.185 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.37 sec

Query History

Saved Queries

Results (1)

	_c0
1	49590