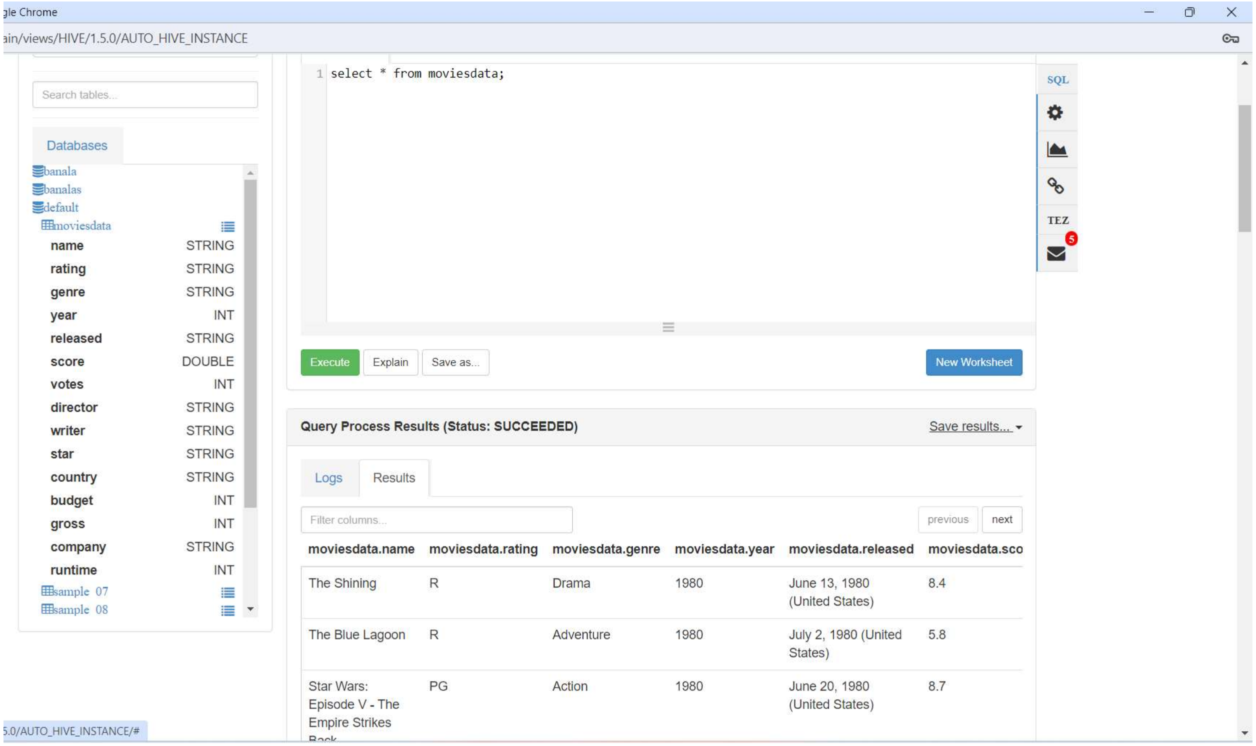
BIG DATA CAPSTONE

1.



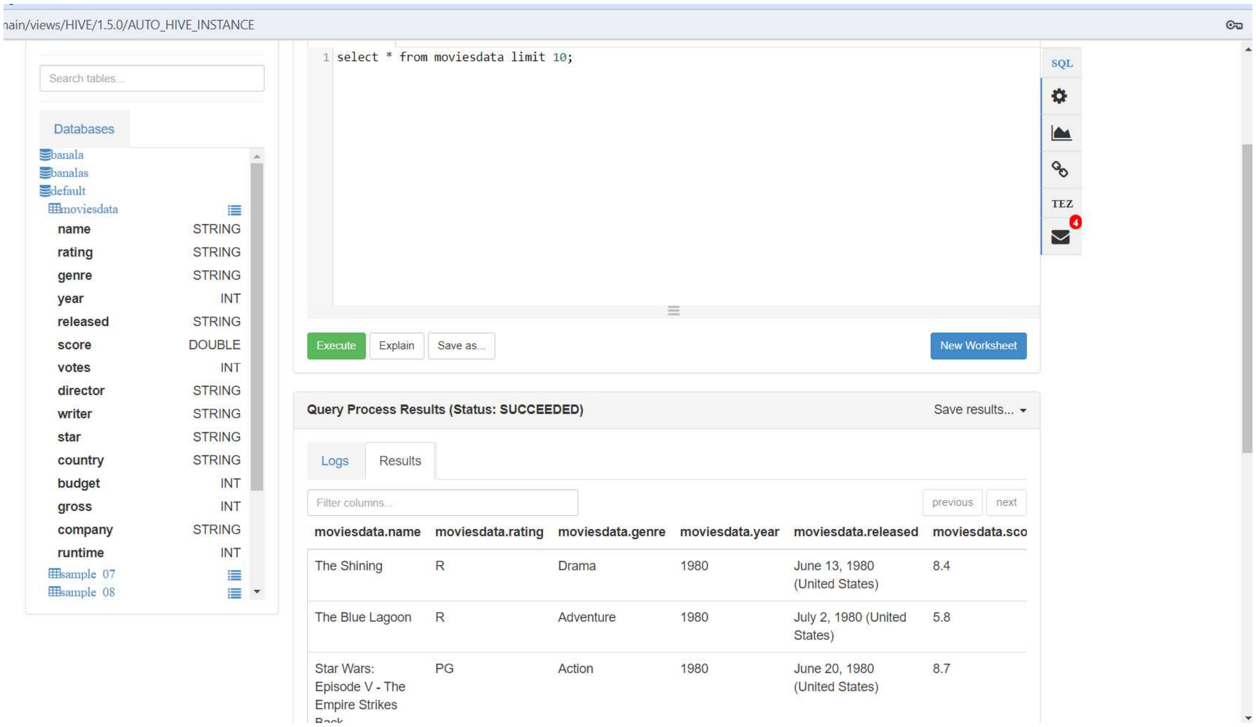
name	STRING
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

```
select * from moviesdata;
```

Query Process Results (Status: SUCCEEDED)

moviesdata.name	moviesdata.rating	moviesdata.genre	moviesdata.year	moviesdata.released	moviesdata.sco
The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4
The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8
Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7

2.



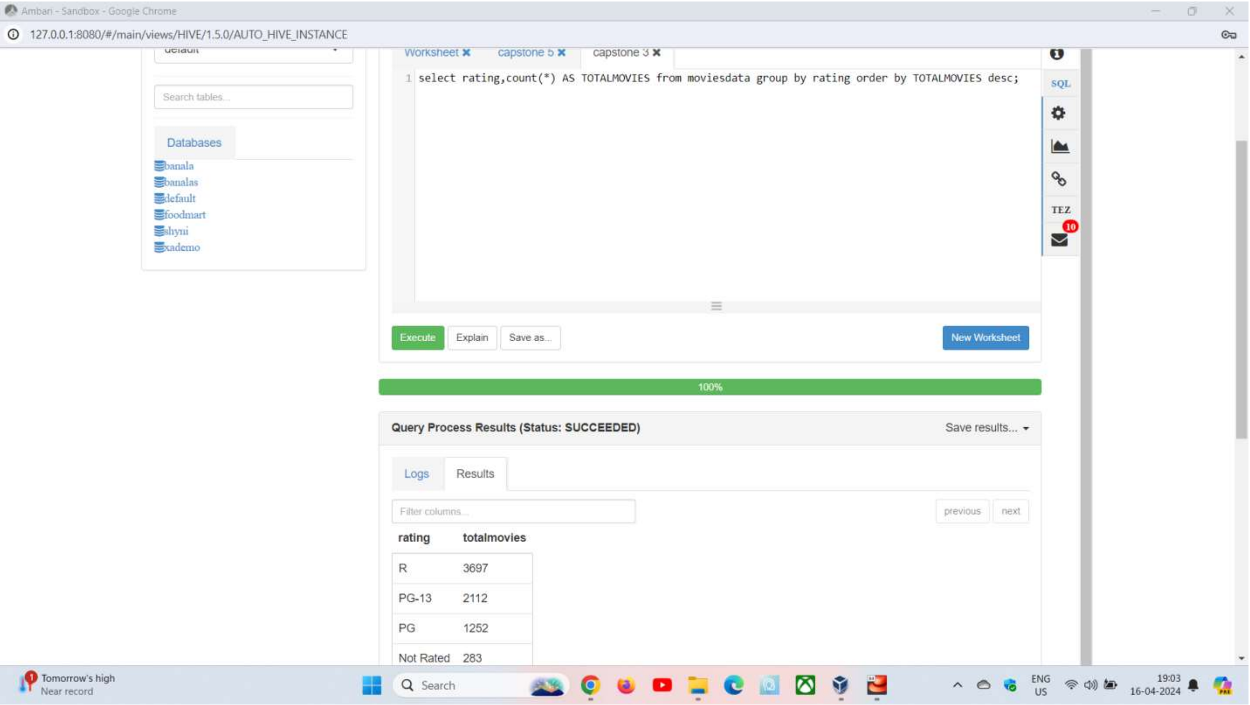
name	STRING
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

```
select * from moviesdata limit 10;
```

Query Process Results (Status: SUCCEEDED)

moviesdata.name	moviesdata.rating	moviesdata.genre	moviesdata.year	moviesdata.released	moviesdata.sco
The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4
The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8
Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7

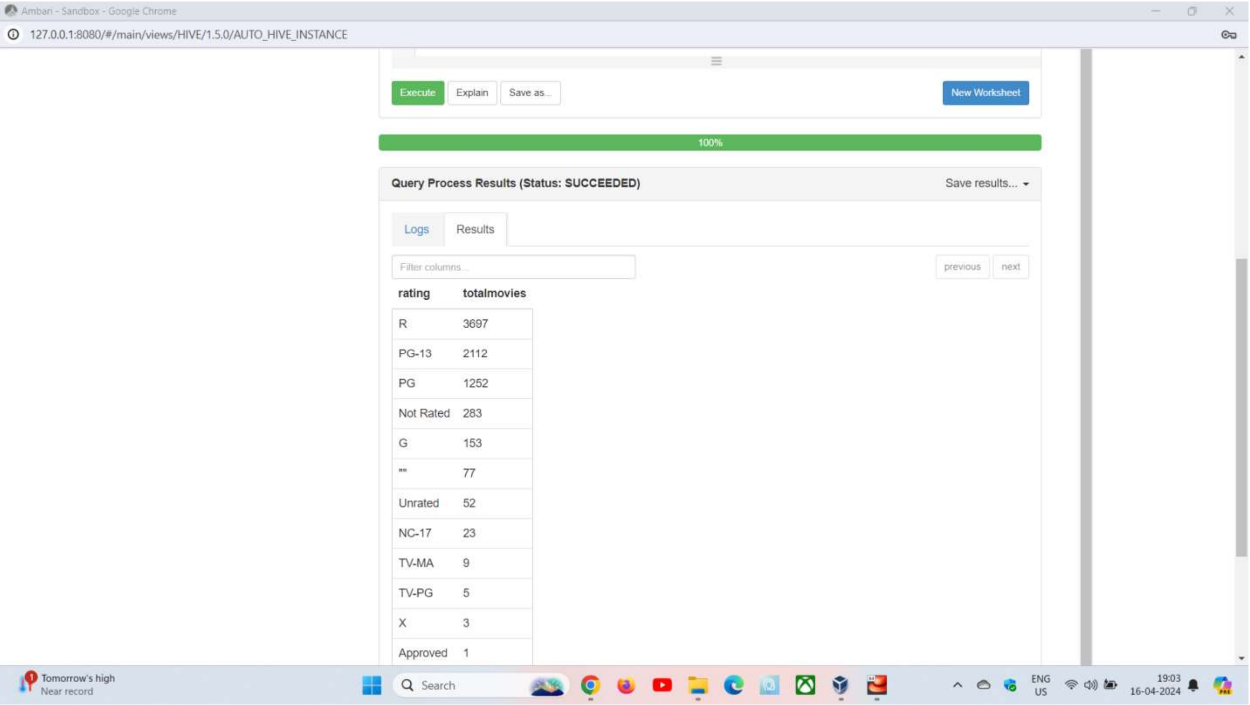
3.



```
select rating,count(*) AS TOTALMOVIES from moviesdata group by rating order by TOTALMOVIES desc;
```

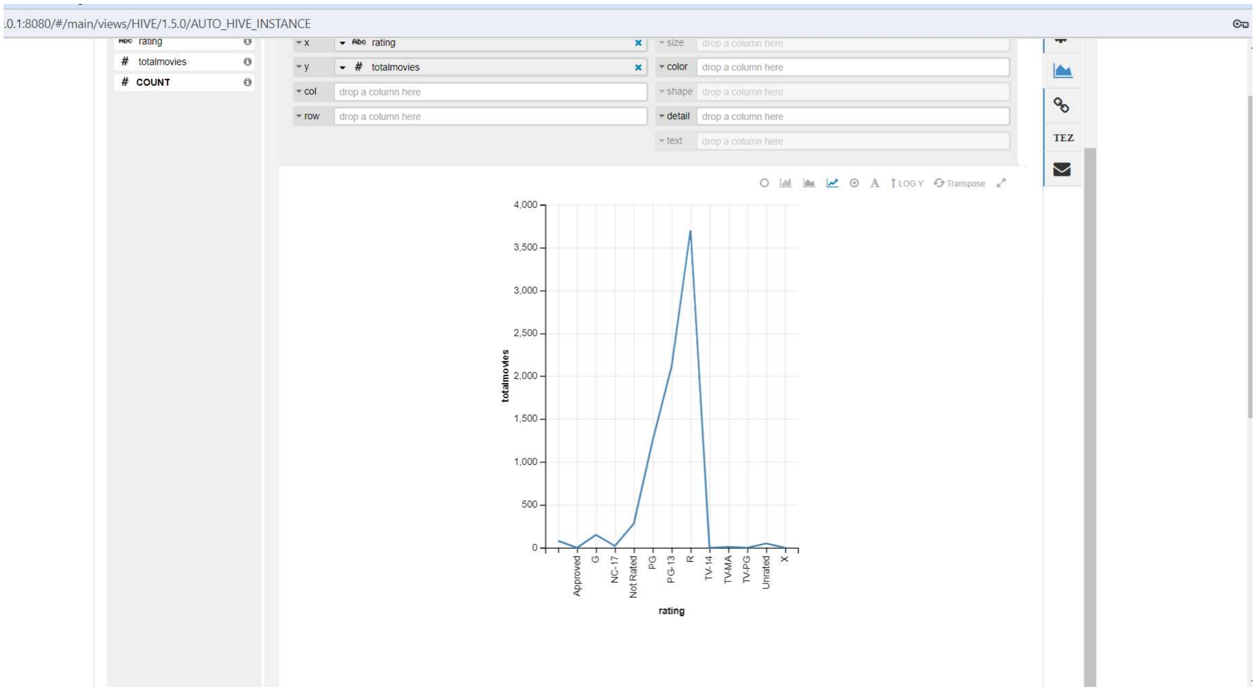
Query Process Results (Status: SUCCEEDED)

rating	totalmovies
R	3697
PG-13	2112
PG	1252
Not Rated	283



Query Process Results (Status: SUCCEEDED)

rating	totalmovies
R	3697
PG-13	2112
PG	1252
Not Rated	283
G	153
""	77
Unrated	52
NC-17	23
TV-MA	9
TV-PG	5
X	3
Approved	1

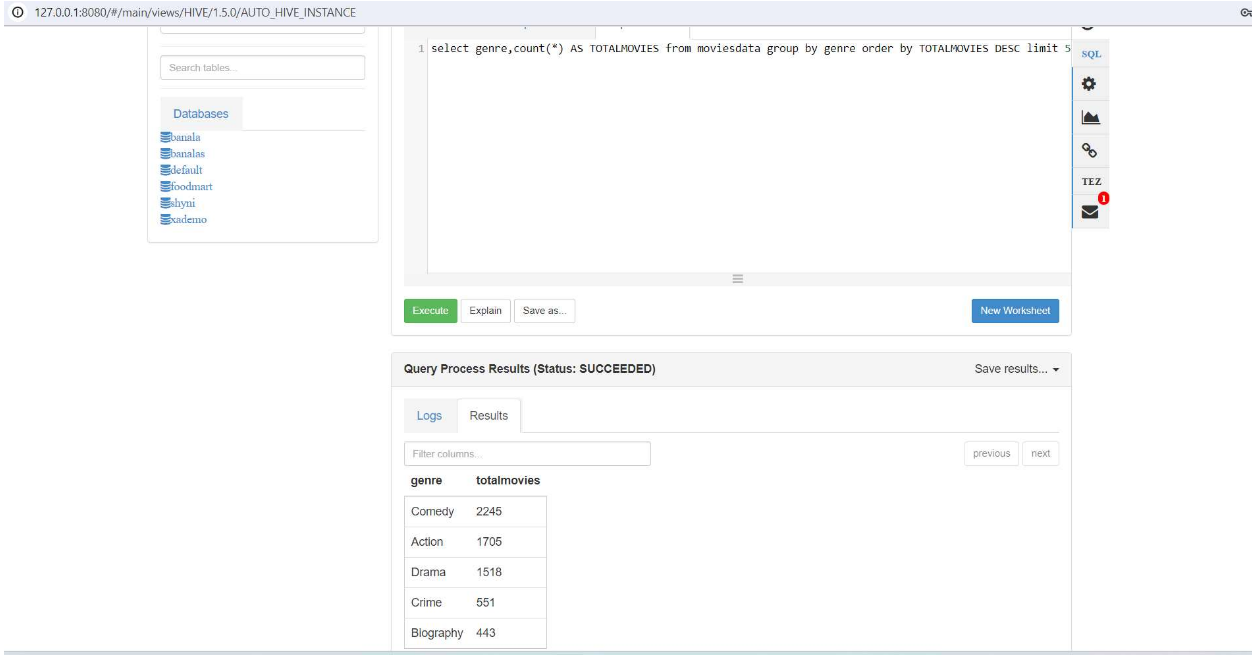


- ➢ From the above visualisation I can conclude that top 3 ratings maximum number of movies are

  1. R (3697)
  2. PG_13(2112)
  3. PG(1252)

- That means maximum number of people from each rating who watched top 3 movies are r,pg_13,pg
- From this top most watched movies are 3697 .

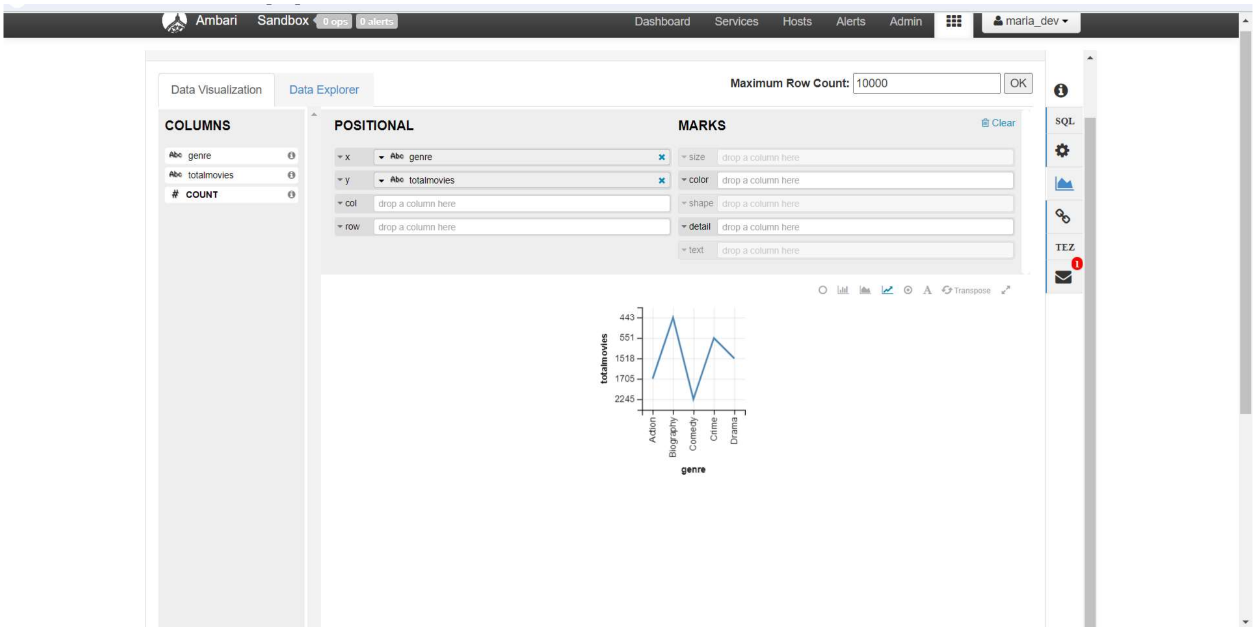
4.



```
select genre,count(*) AS TOTALMOVIES from moviesdata group by genre order by TOTALMOVIES DESC limit 5
```

Query Process Results (Status: SUCCEEDED)

genre	totalmovies
Comedy	2245
Action	1705
Drama	1518
Crime	551
Biography	443



Maximum Row Count: 10000

COLUMNS
- genre
- totalmovies
- # COUNT

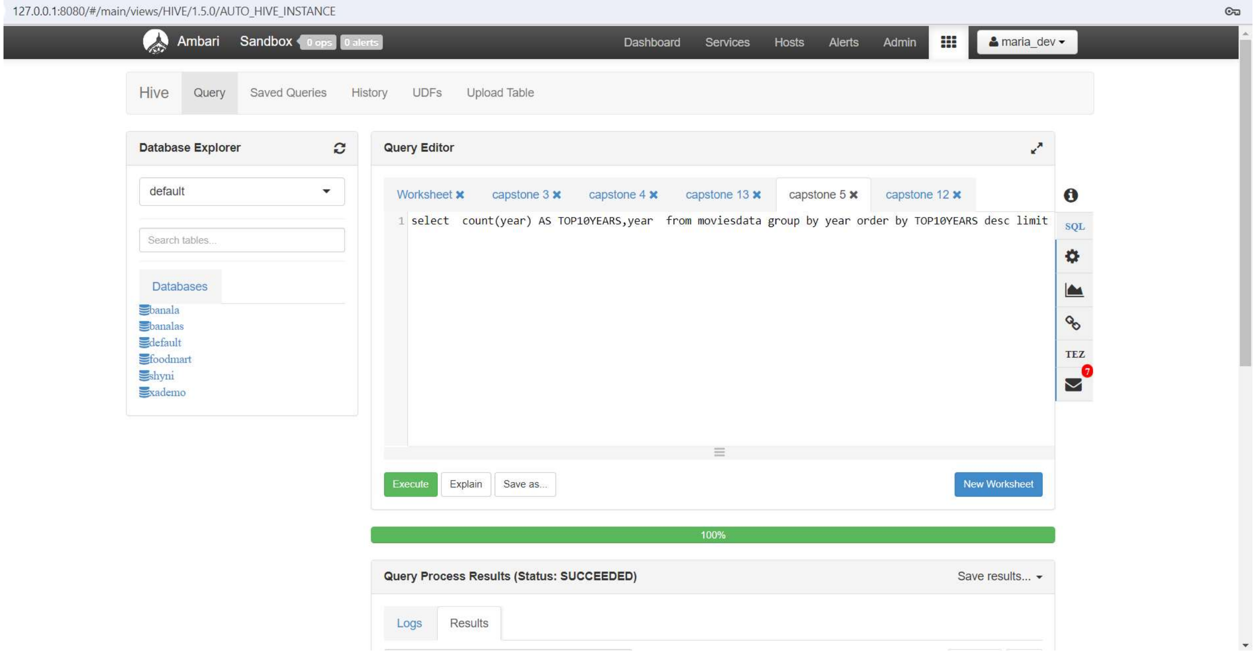
POSITIONAL
- x: genre
- y: totalmovies
- col: drop a column here
- row: drop a column here

MARKS
- size: drop a column here
- color: drop a column here
- shape: drop a column here
- detail: drop a column here
- text: drop a column here

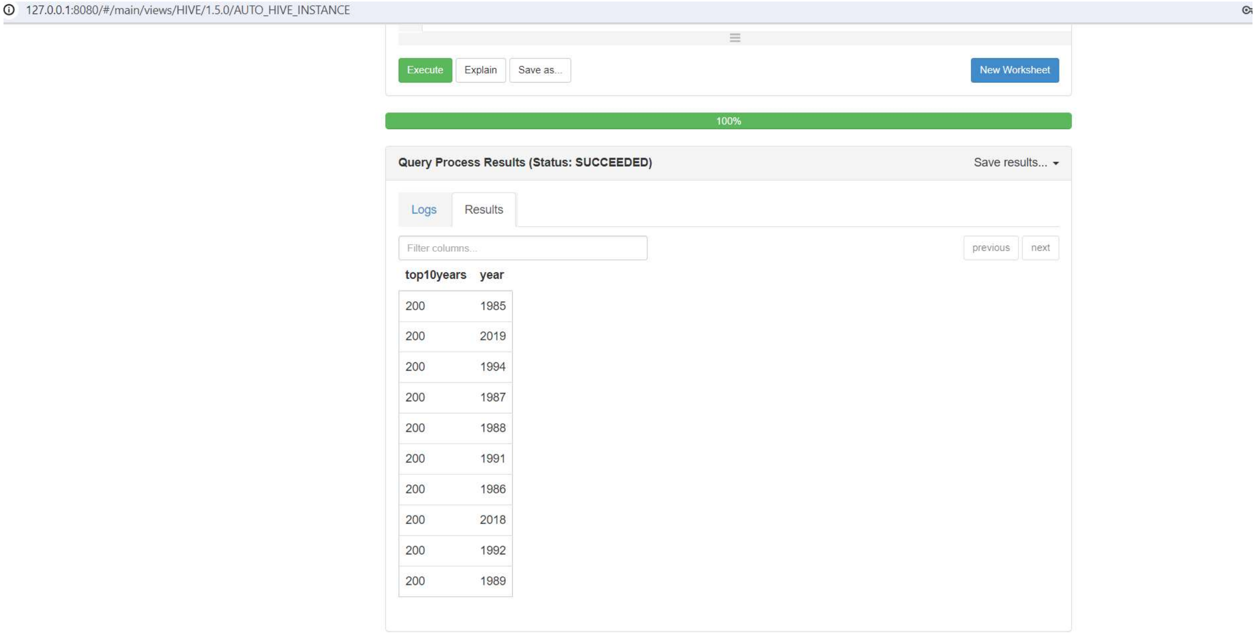
- From the above visualization I can conclude that from each genre top 3 movies are

  1.Comedy (2245)
  2.Action (1705)
  3.Drama (1518)
- From this most number of people watched the comedy movies .
- Most number of directors liked to make comedy movies.

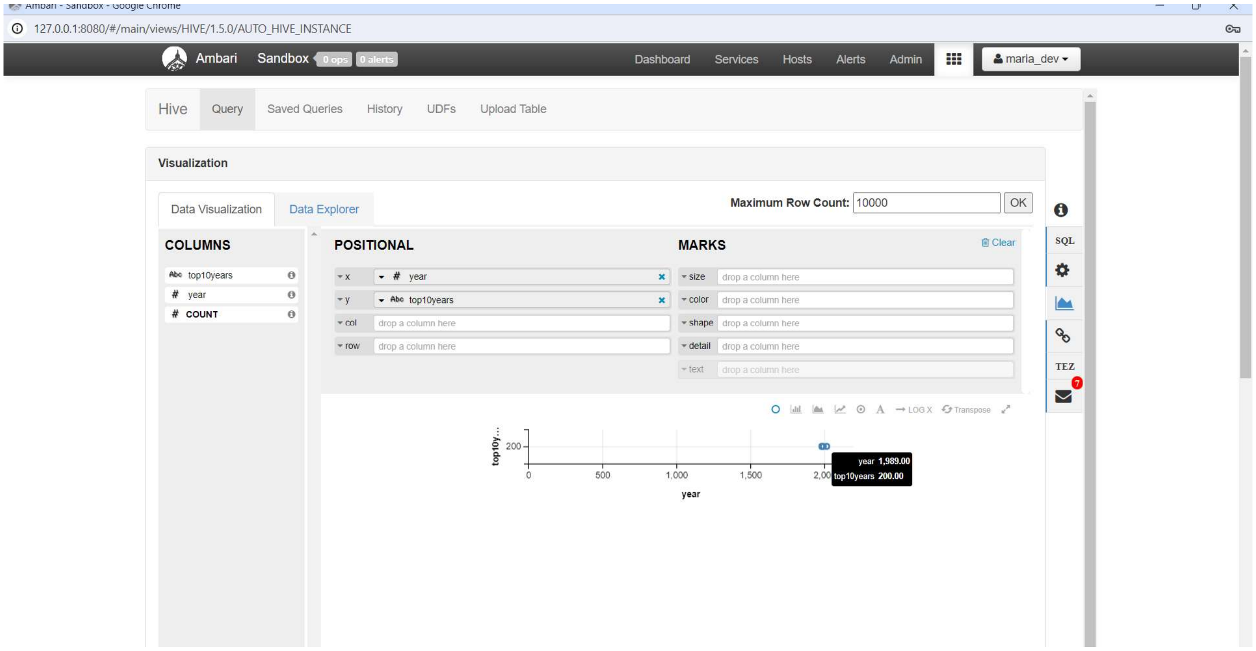
5.



```
select  count(year) AS TOP10YEARS,year  from moviesdata group by year order by TOP10YEARS desc limit
```

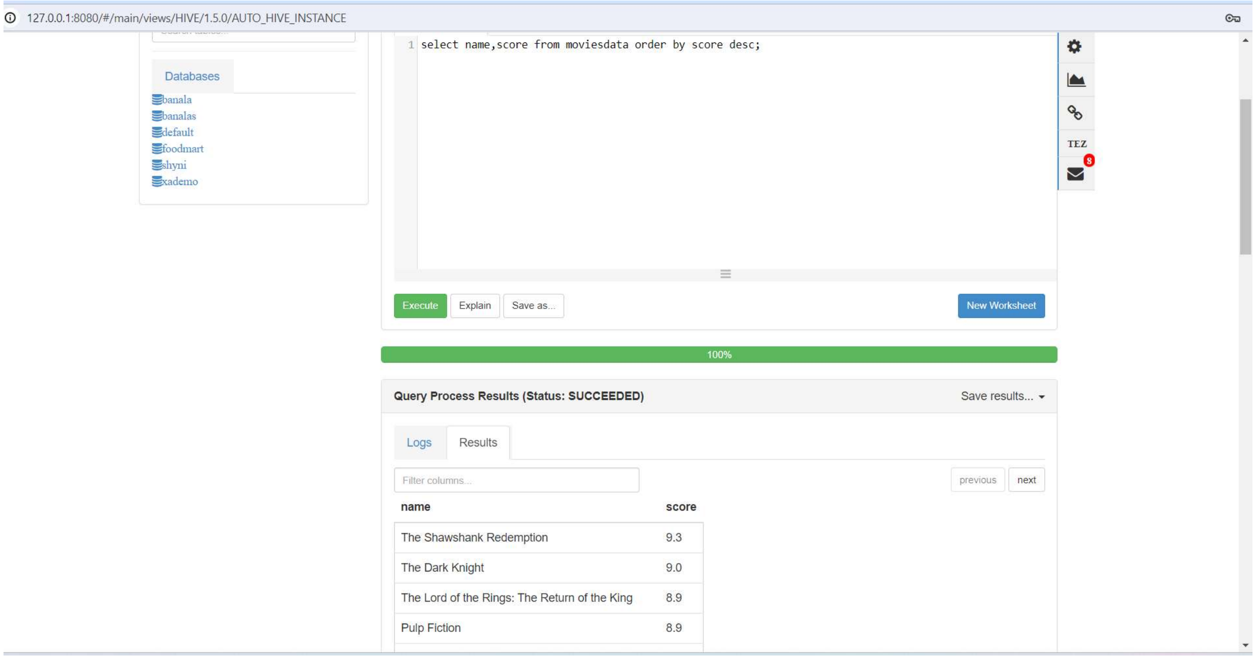

127.0.0.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE
Execute
Explain
Save as...
New Worksheet
100%
Query Process Results (Status: SUCCEEDED)
Save results...
Logs
Results
Filter columns...
previous
next

top10years	year
200	1985
200	2019
200	1994
200	1987
200	1988
200	1991
200	1986
200	2018
200	1992
200	1989



- From the above visualisation I can conclude that in the year of 1985 200 movies were released .
- In the years of 2019,1994 ,1987,1988,1991,1986, 2018,1992,1989 total 200 movies were released .
- For top 10 years all the movies released were same .

6.

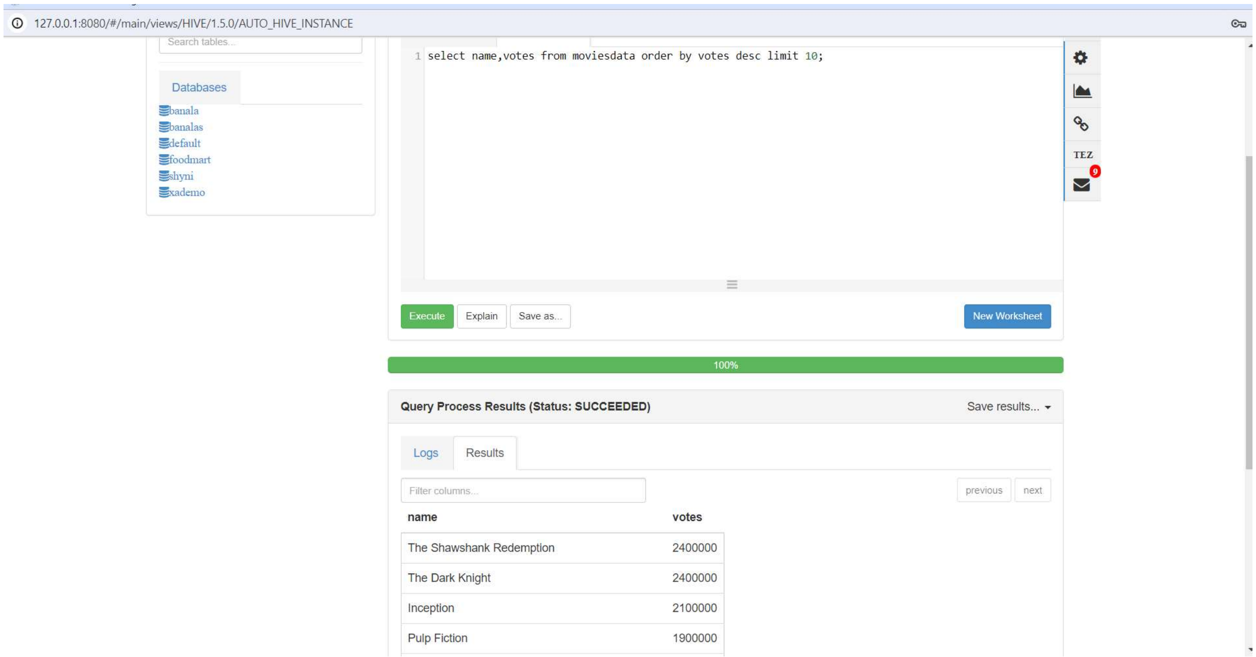


```
select name,score from moviesdata order by score desc;
```

Query Process Results (Status: SUCCEEDED)

name	score
The Shawshank Redemption	9.3
The Dark Knight	9.0
The Lord of the Rings: The Return of the King	8.9
Pulp Fiction	8.9

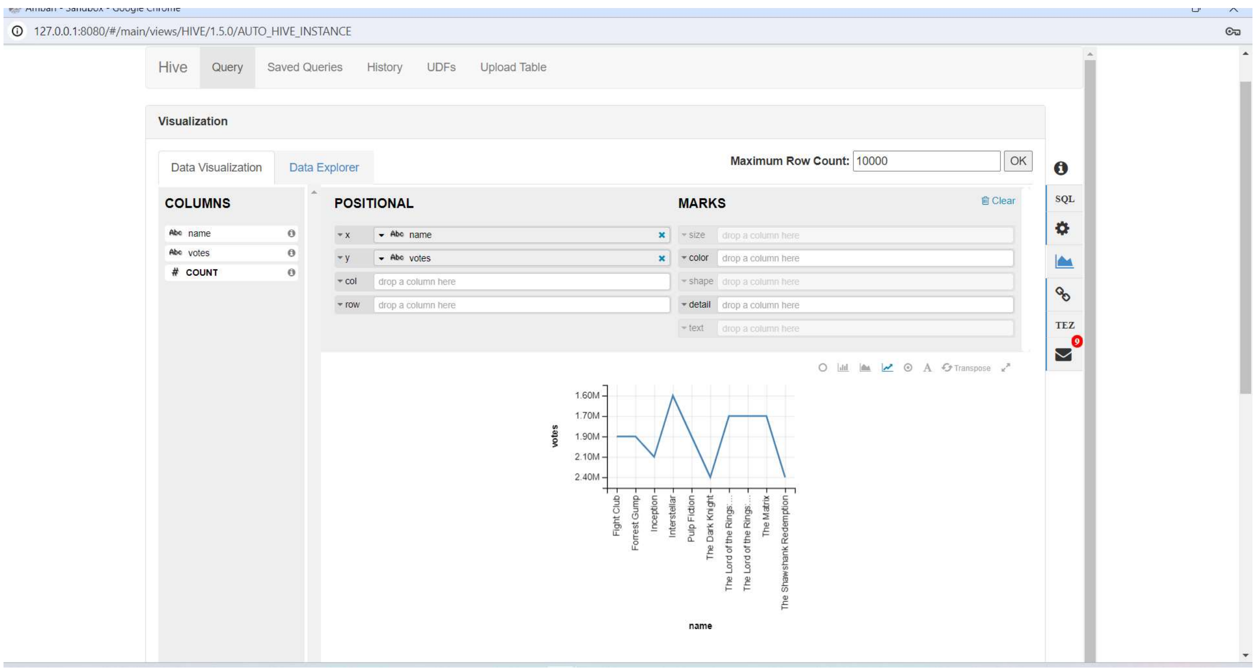
7.



```
select name,votes from moviesdata order by votes desc limit 10;
```

Query Process Results (Status: SUCCEEDED)

name	votes
The Shawshank Redemption	2400000
The Dark Knight	2400000
Inception	2100000
Pulp Fiction	1900000



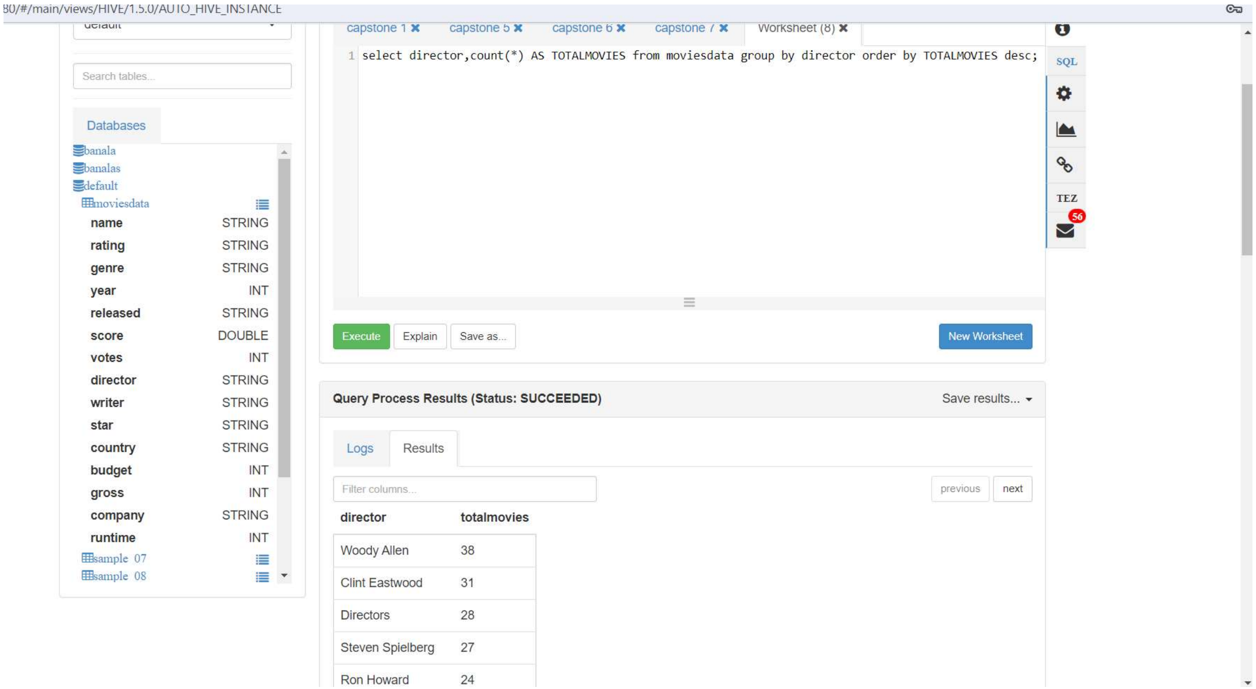
Maximum Row Count: 10000

x: name

y: votes

- ➢ From the above visualization I can conclude that the total votes polled for each movie were revelaed .
- ➢ Here the top 3 votes polled for the movies are 1.2400000 (The shawshank redemption)
  2.2400000 (The dark knight)
  3.2100000(Inception)
- ➢ Total votes polled means 2400000 people watched the 2 different movies .

8.



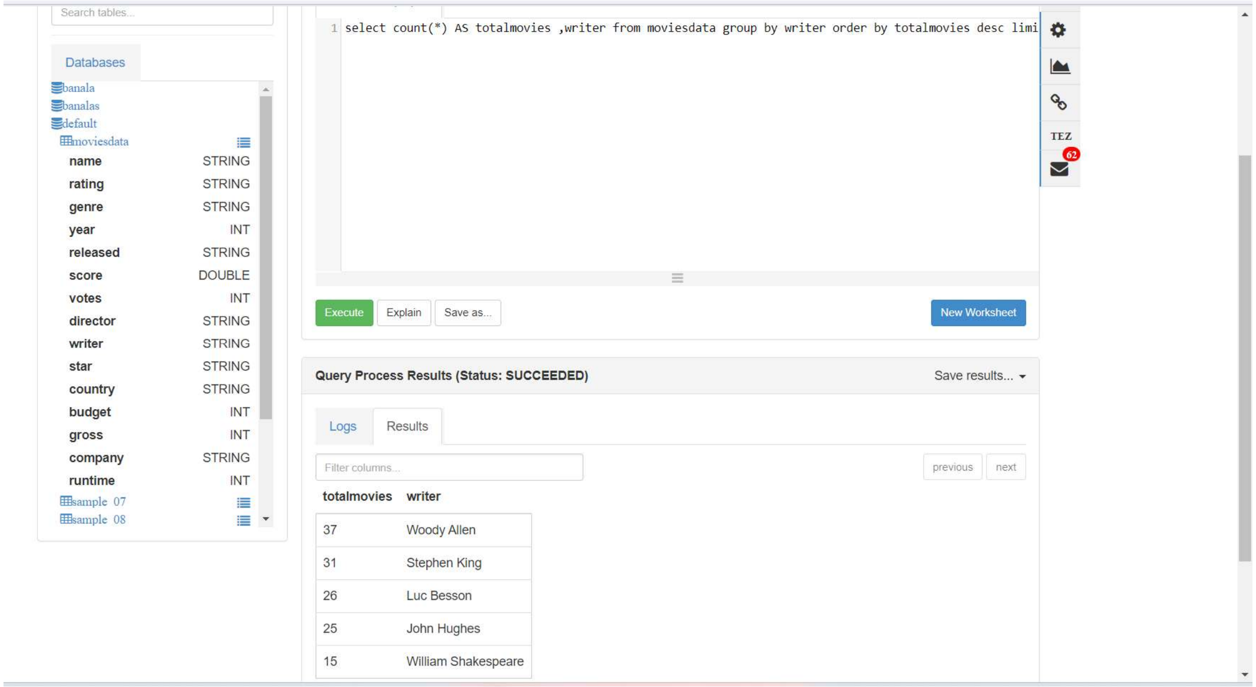
name	STRING
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

```
select director,count(*) AS TOTALMOVIES from moviesdata group by director order by TOTALMOVIES desc;
```

Query Process Results (Status: SUCCEEDED)

director	totalmovies
Woody Allen	38
Clint Eastwood	31
Directors	28
Steven Spielberg	27
Ron Howard	24

9.

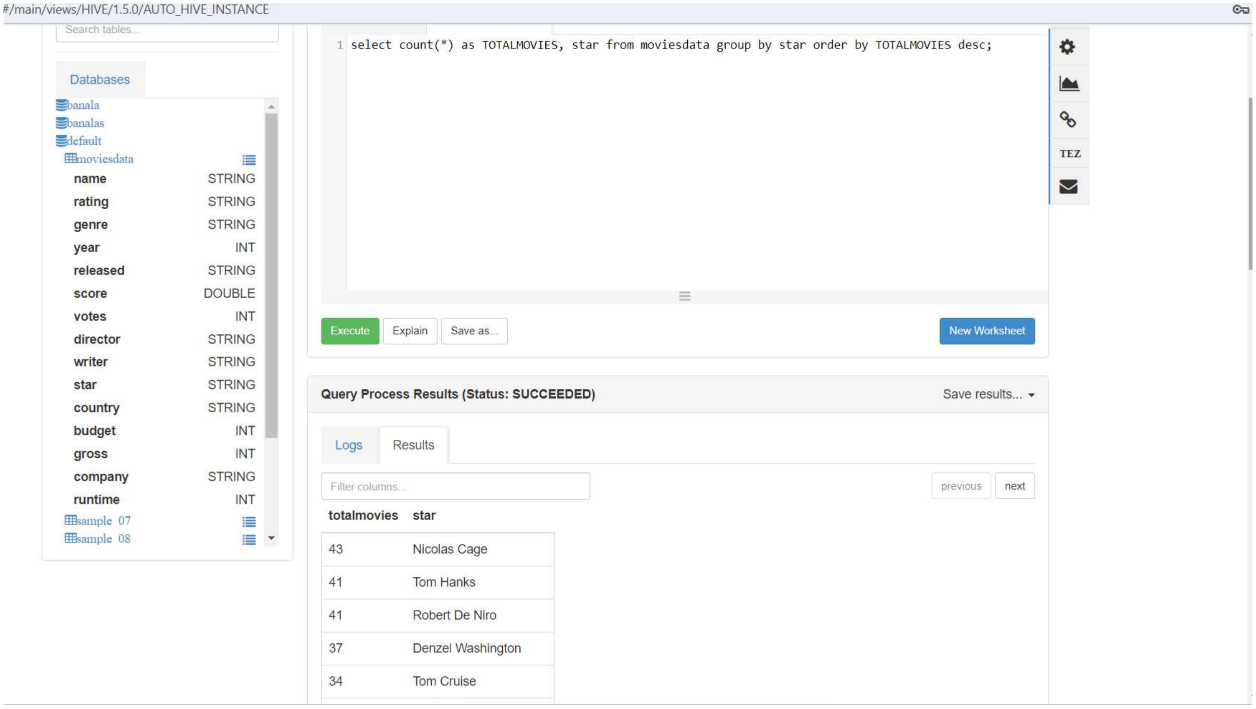


```
select count(*) AS totalmovies ,writer from moviesdata group by writer order by totalmovies desc limi
```

Query Process Results (Status: SUCCEEDED)

totalmovies	writer
37	Woody Allen
31	Stephen King
26	Luc Besson
25	John Hughes
15	William Shakespeare

10.

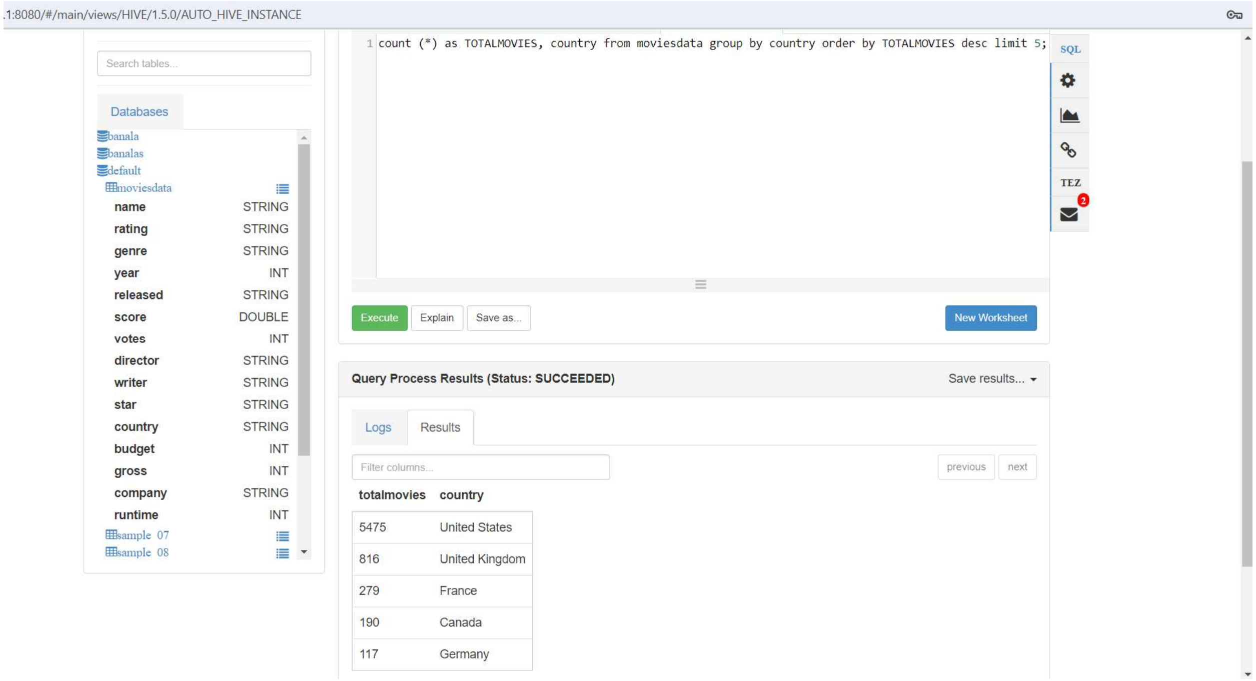


```
select count(*) as TOTALMOVIES, star from moviesdata group by star order by TOTALMOVIES desc;
```

Query Process Results (Status: SUCCEEDED)

totalmovies	star
43	Nicolas Cage
41	Tom Hanks
41	Robert De Niro
37	Denzel Washington
34	Tom Cruise

11.



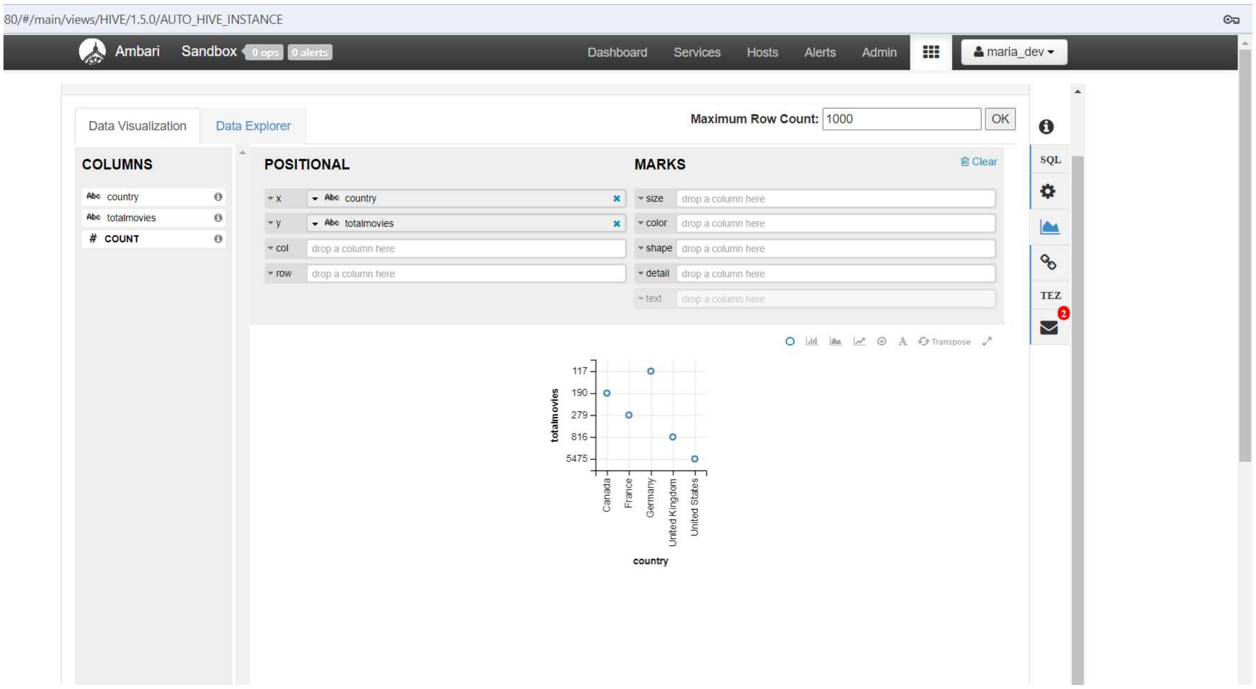
.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

name	STRING
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

```
count (*) as TOTALMOVIES, country from moviesdata group by country order by TOTALMOVIES desc limit 5;
```

Query Process Results (Status: SUCCEEDED)

totalmovies	country
5475	United States
816	United Kingdom
279	France
190	Canada
117	Germany



80/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

Maximum Row Count: 1000

COLUMNS

- Abc country
- Abc totalmovies
- # COUNT

POSITIONAL

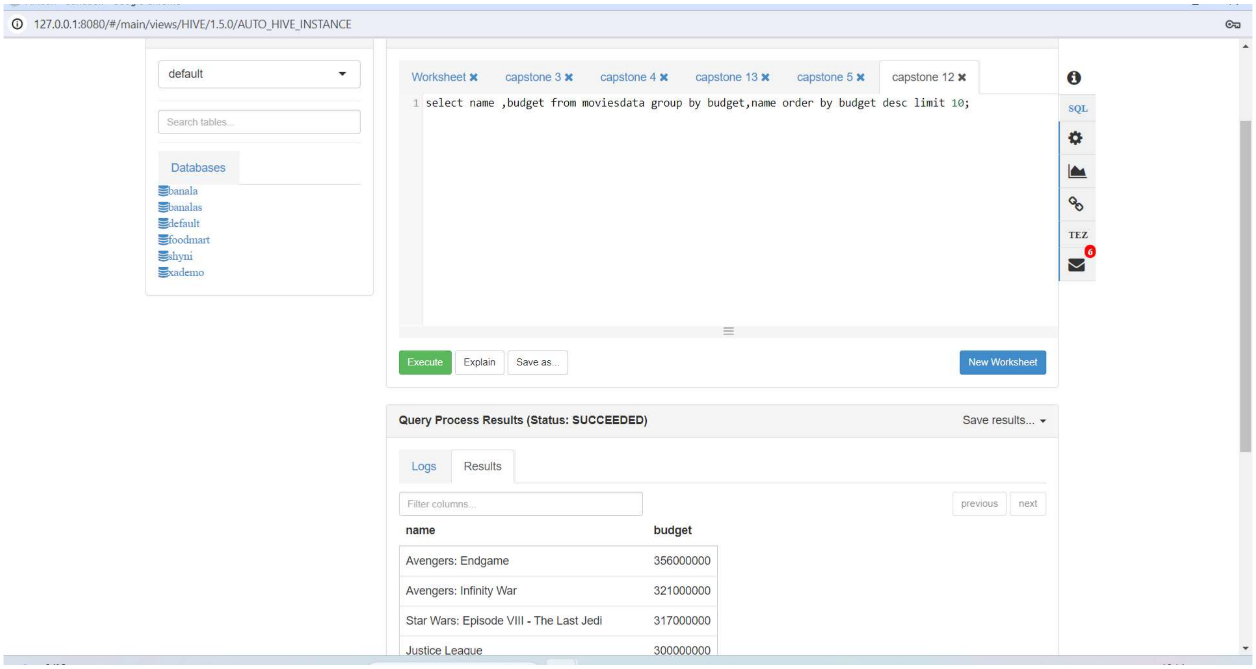
- x: Abc country
- y: Abc totalmovies
- col: drop a column here
- row: drop a column here

MARKS

- size: drop a column here
- color: drop a column here
- shape: drop a column here
- detail: drop a column here
- text: drop a column here

- From the above visualisation I can conclude that each country how many movies were watched will be revealed .
- Here the top 3 countries total watched movies are 1. United states (5475)
  2. United kingdom (816)
  3.france (279)
- That means from united states 5475 movies generated the top most revenue.

12.

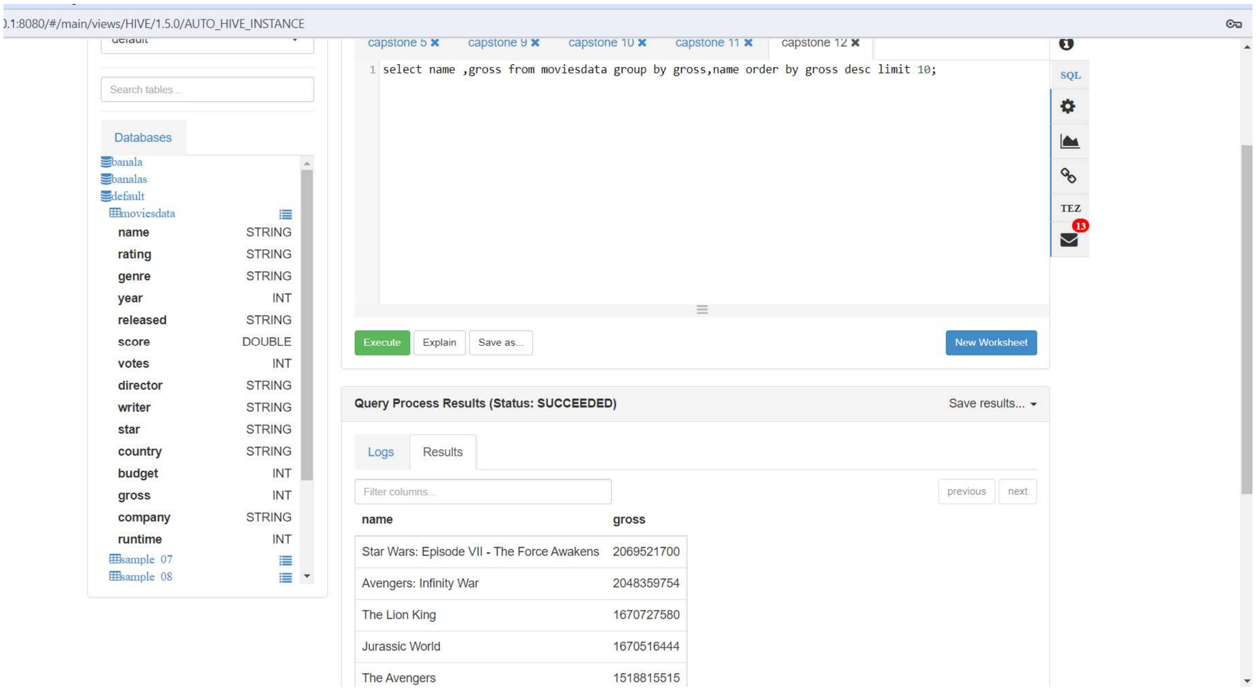


```
select name ,budget from moviesdata group by budget,name order by budget desc limit 10;
```

Query Process Results (Status: SUCCEEDED)

name	budget
Avengers: Endgame	356000000
Avengers: Infinity War	321000000
Star Wars: Episode VIII - The Last Jedi	317000000
Justice League	300000000

13.



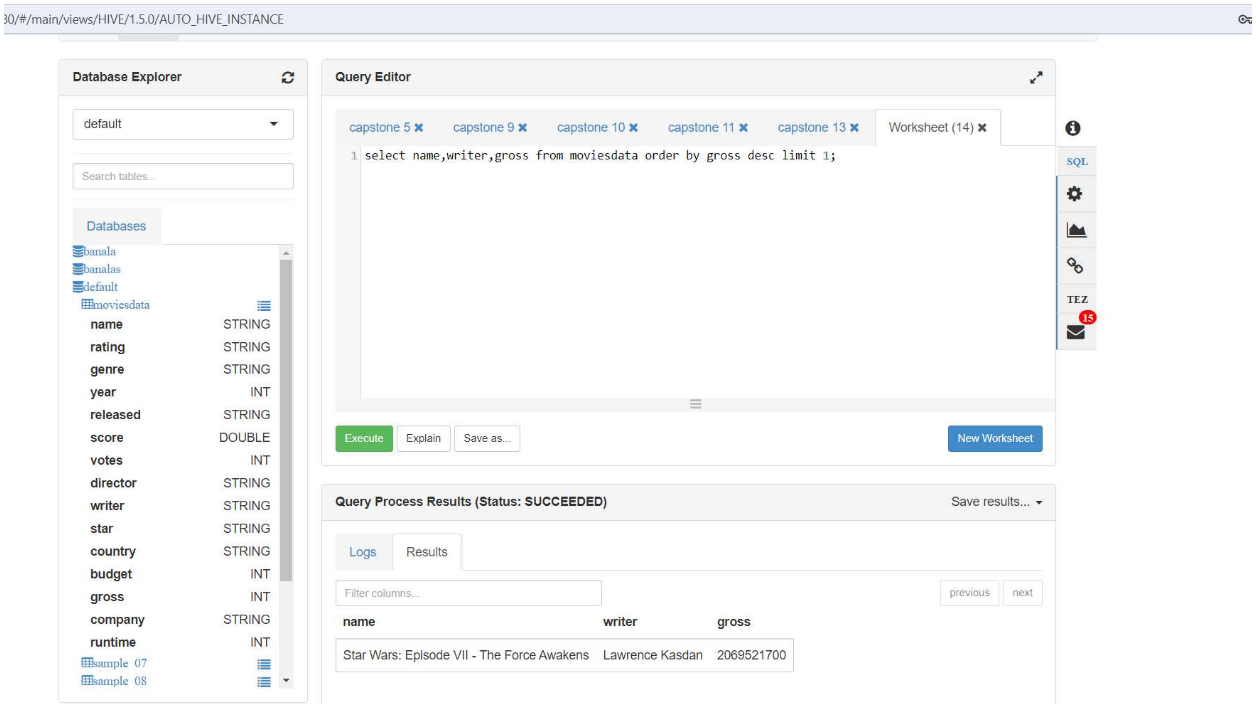
```
select name ,gross from moviesdata group by gross,name order by gross desc limit 10;
```

name	STRING
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

Query Process Results (Status: SUCCEEDED)

name	gross
Star Wars: Episode VII - The Force Awakens	2069521700
Avengers: Infinity War	2048359754
The Lion King	1670727580
Jurassic World	1670516444
The Avengers	1518815515

14.



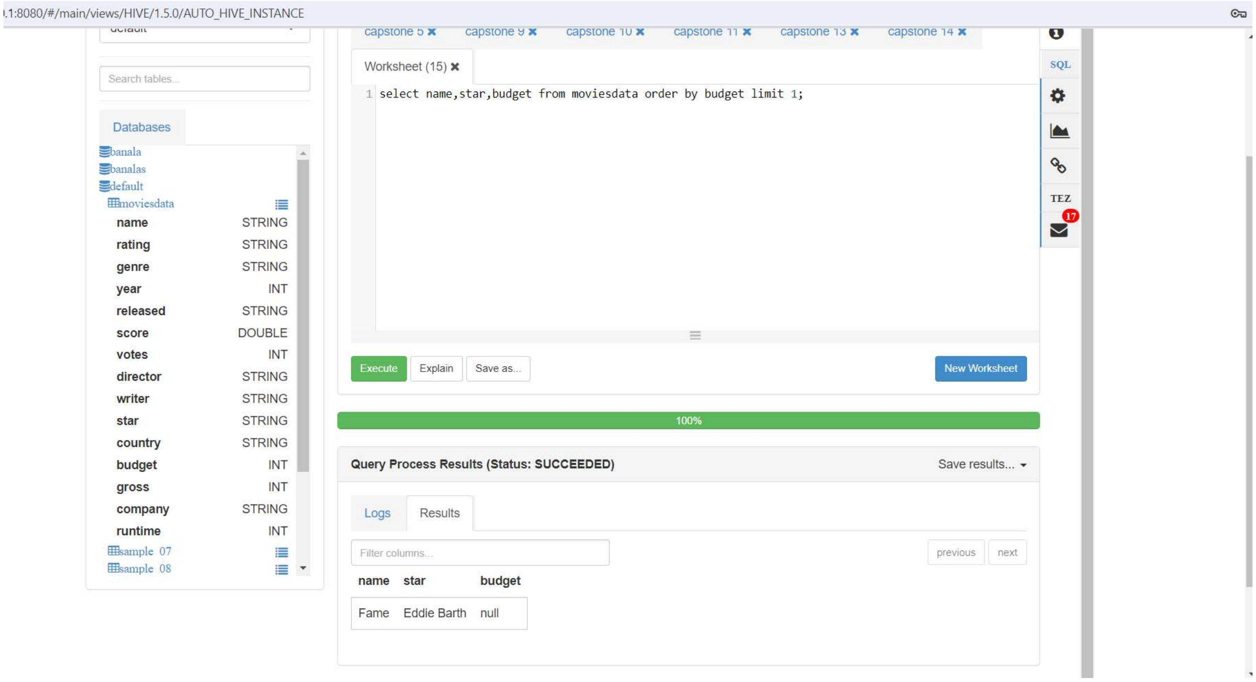
```
select name,writer,gross from moviesdata order by gross desc limit 1;
```

name	STRING
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

Query Process Results (Status: SUCCEEDED)

name	writer	gross
Star Wars: Episode VII - The Force Awakens	Lawrence Kasdan	2069521700

15.



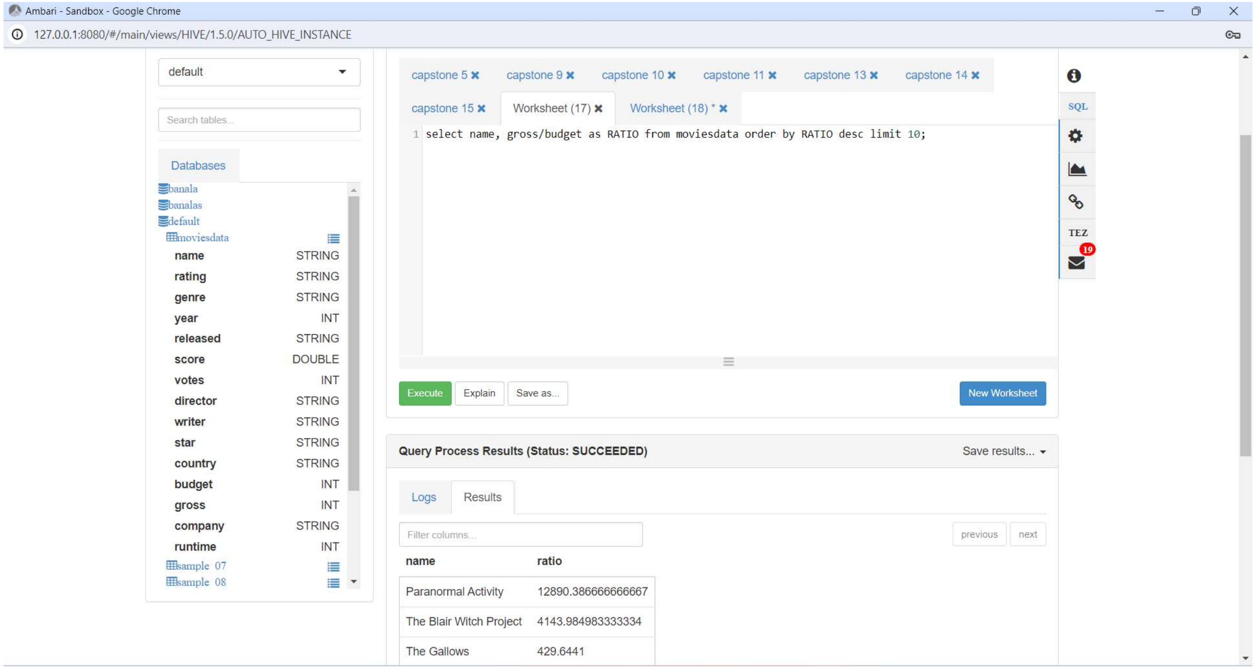
name	STRING
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

```
select name,star,budget from moviesdata order by budget limit 1;
```

Query Process Results (Status: SUCCEEDED)

```
name   star         budget
Fame   Eddie Barth  null
```

16.

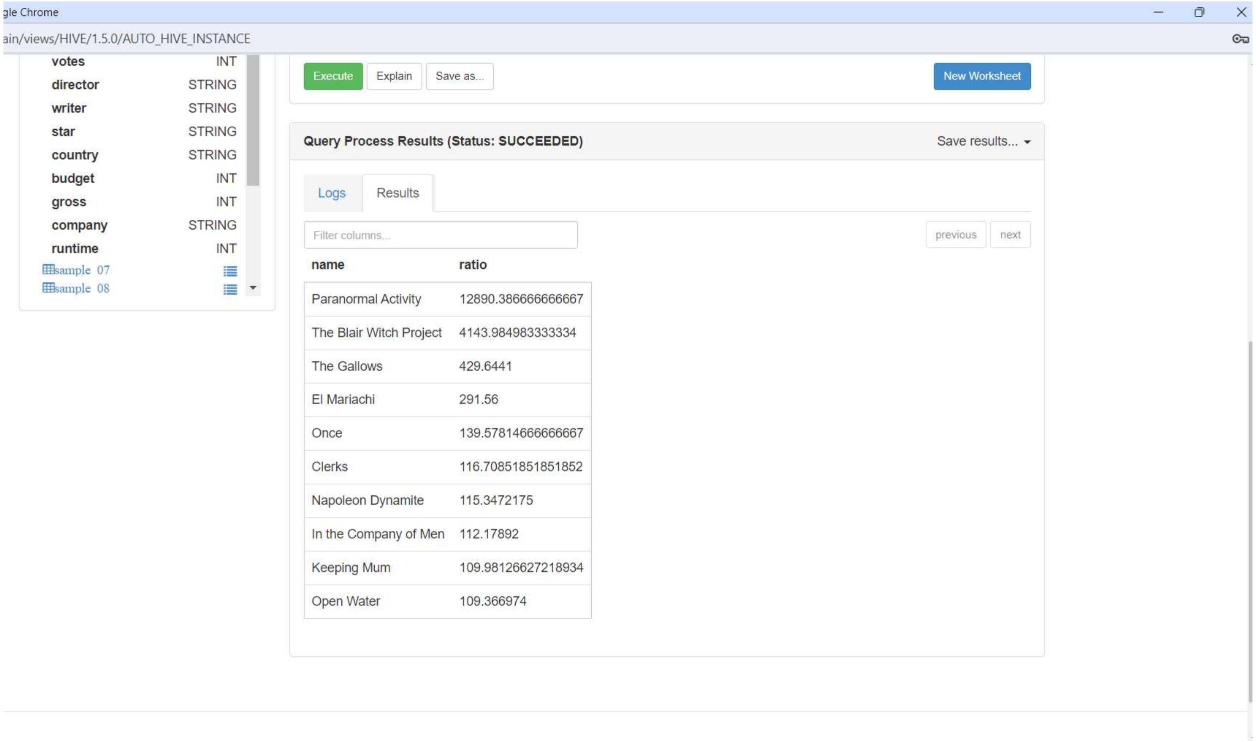


name	STRING
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

```
select name, gross/budget as RATIO from moviesdata order by RATIO desc limit 10;
```

Query Process Results (Status: SUCCEEDED)

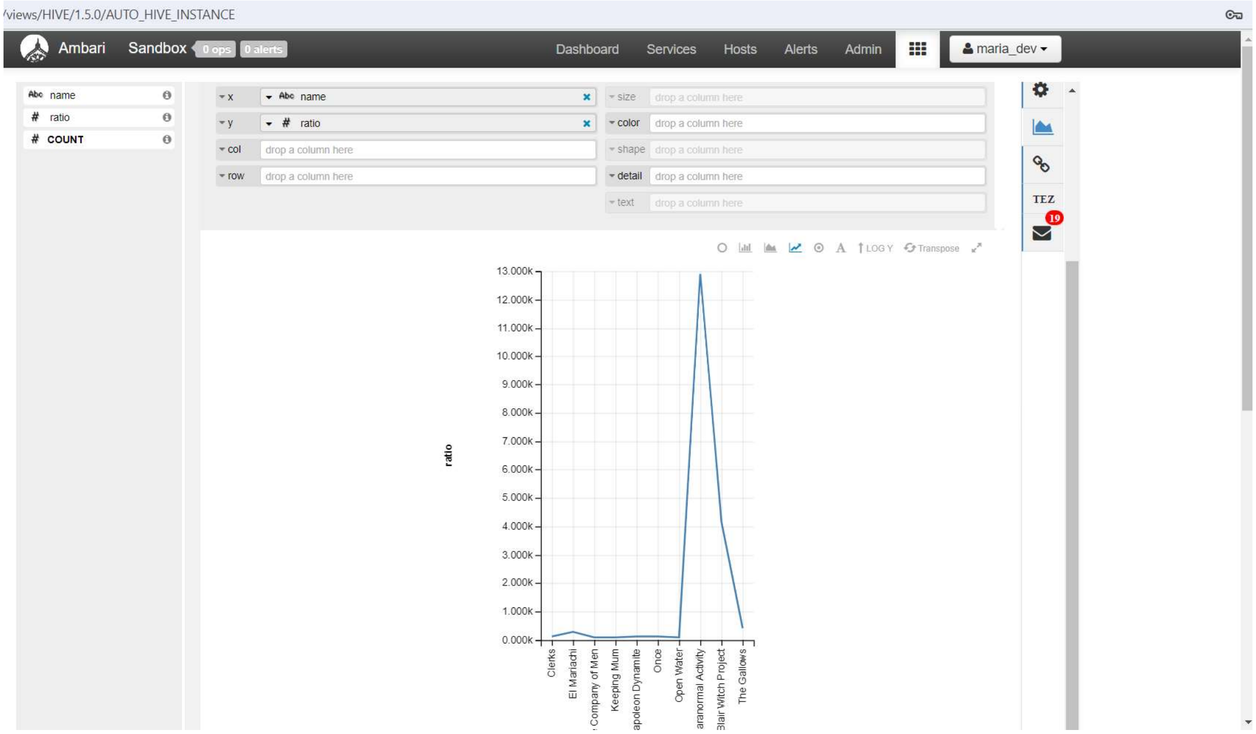
name	ratio
Paranormal Activity	12890.386666666667
The Blair Witch Project	4143.984983333334
The Gallows	429.6441



votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

Query Process Results (Status: SUCCEEDED)

name	ratio
Paranormal Activity	12890.386666666667
The Blair Witch Project	4143.984983333334
The Gallows	429.6441
El Mariachi	291.56
Once	139.57814666666667
Clerks	116.70851851851852
Napoleon Dynamite	115.3472175
In the Company of Men	112.17892
Keeping Mum	109.98126627218934
Open Water	109.366974

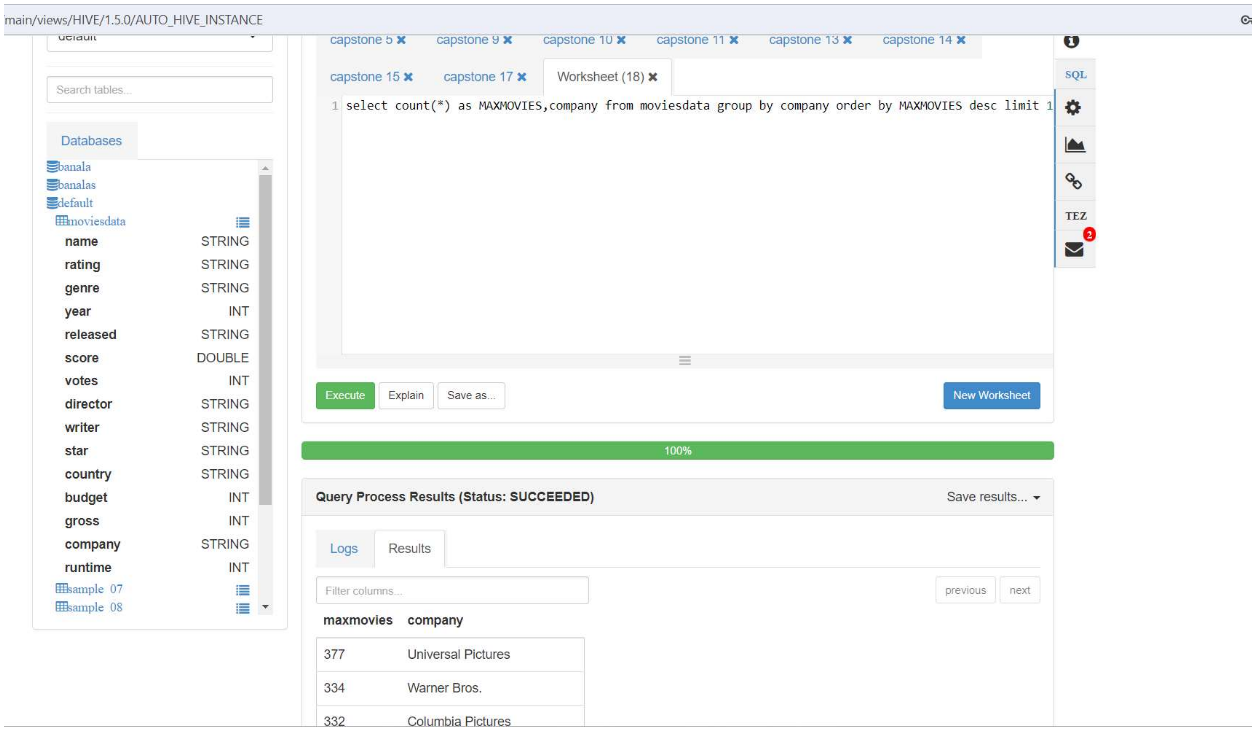


x: name

y: ratio

- From the above visualization I can conclude that gross to budget ratio means the profit of each movie is revealed here
- That means top 3 highest profitable movies are 1.Paranormal activity (12890.38)
  2.The blair witch project (4143.98)
  3. The gallows (429.64)

17.



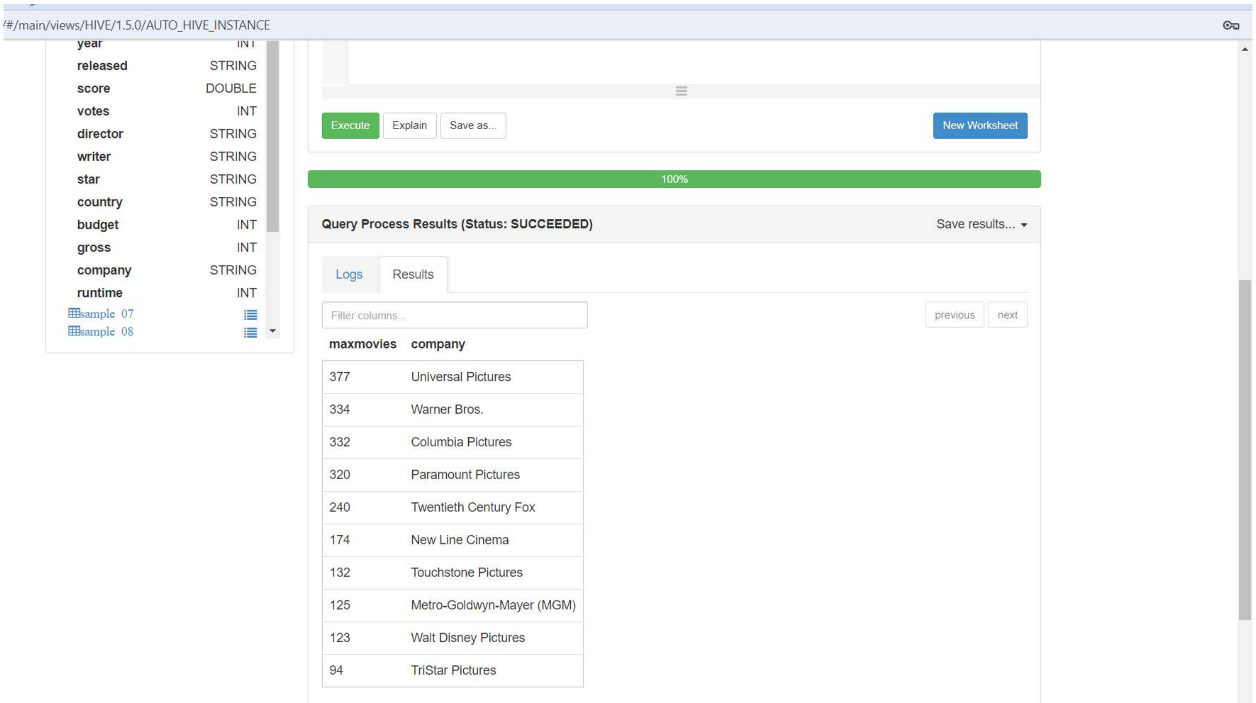
```
select count(*) as MAXMOVIES,company from moviesdata group by company order by MAXMOVIES desc limit 1
```

moviesdata	
name	STRING
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

sample_07
sample_08

Query Process Results (Status: SUCCEEDED)

maxmovies	company
377	Universal Pictures
334	Warner Bros.
332	Columbia Pictures

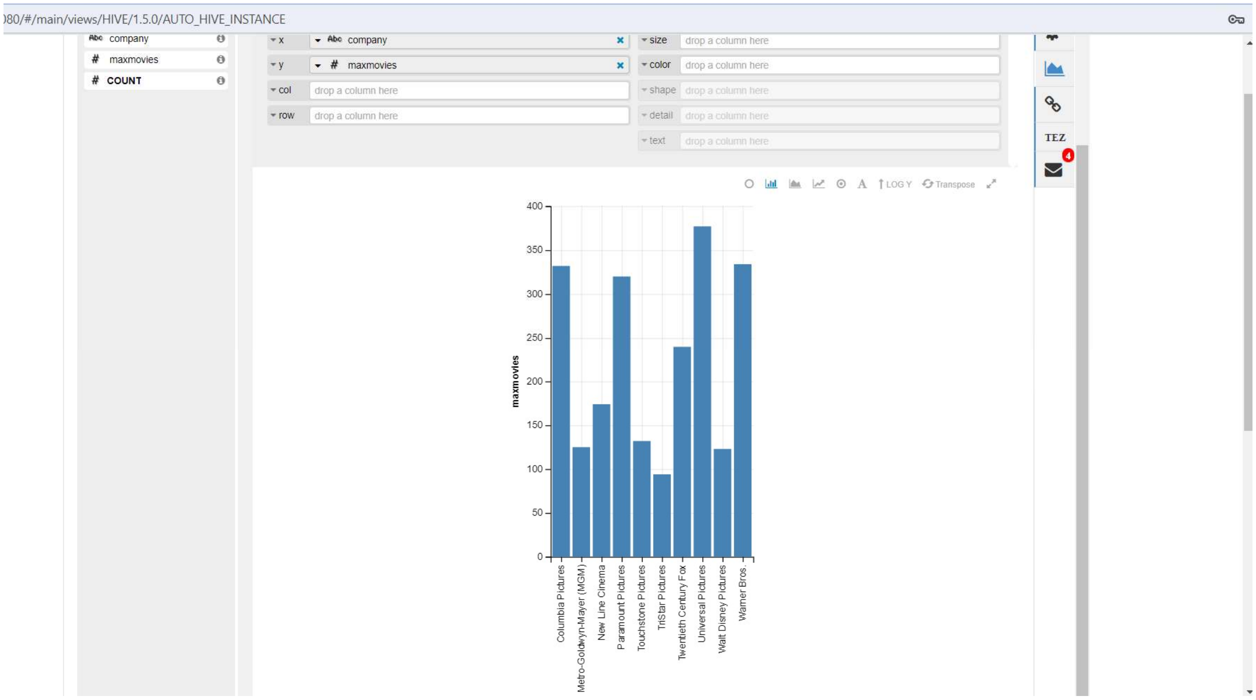


year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

sample_07
sample_08

Query Process Results (Status: SUCCEEDED)

maxmovies	company
377	Universal Pictures
334	Warner Bros.
332	Columbia Pictures
320	Paramount Pictures
240	Twentieth Century Fox
174	New Line Cinema
132	Touchstone Pictures
125	Metro-Goldwyn-Mayer (MGM)
123	Walt Disney Pictures
94	TriStar Pictures



)80/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

Abc company
# maxmovies
COUNT

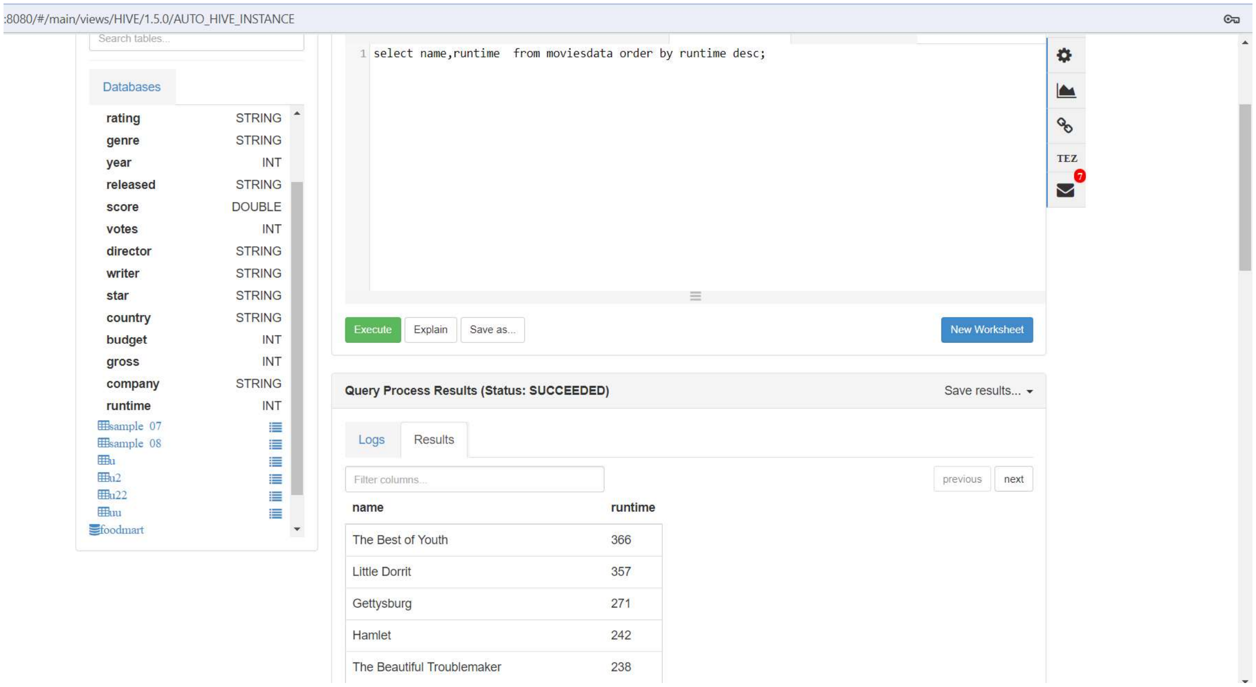
x: Abc company
y: # maxmovies
col: drop a column here
row: drop a column here
size: drop a column here
color: drop a column here
shape: drop a column here
detail: drop a column here
text: drop a column here

- From the above visualisation I can conclude that which company make the highest number of movies were revealed here.
- Here the top 3 companies make the movies are 1.universal pictures (377)
  2.warner bros (334)
  3.columbia pictures (332)

- From the above each company total movies production is revelaed .

18.

a.



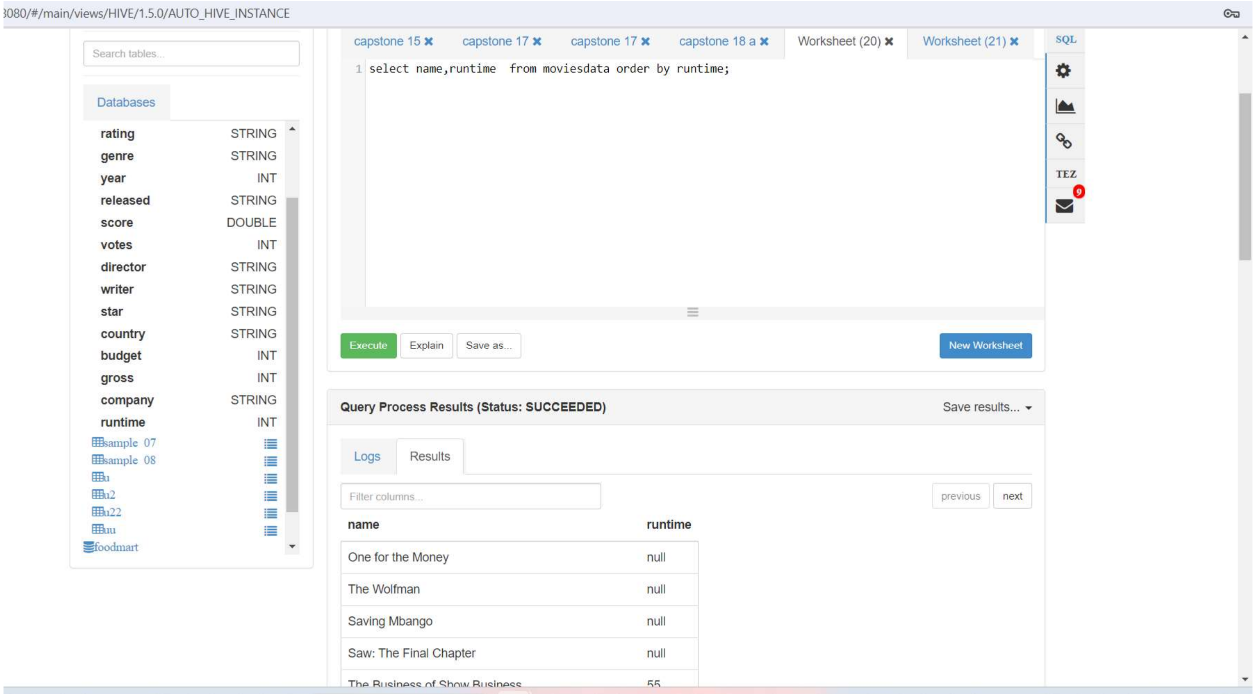
:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

```
select name,runtime  from moviesdata order by runtime desc;
```

Query Process Results (Status: SUCCEEDED)

name	runtime
The Best of Youth	366
Little Dorrit	357
Gettysburg	271
Hamlet	242
The Beautiful Troublemaker	238



3080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

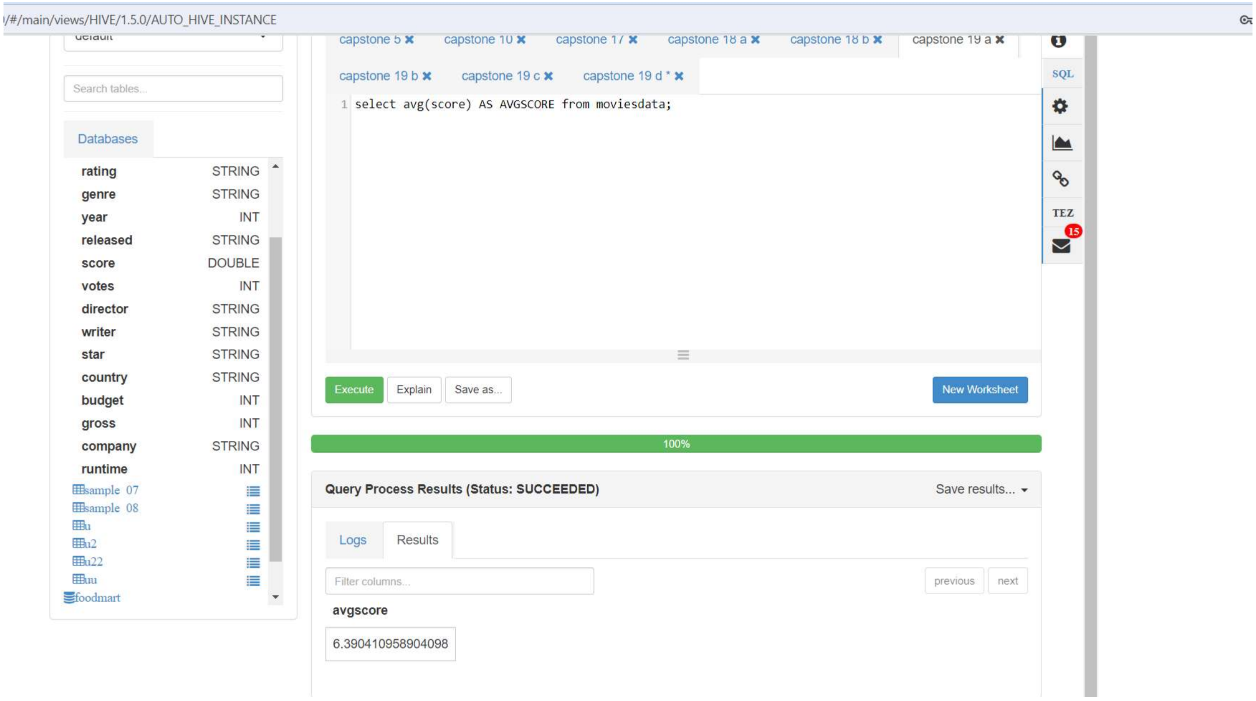
```
select name,runtime  from moviesdata order by runtime;
```

Query Process Results (Status: SUCCEEDED)

name	runtime
One for the Money	null
The Wolfman	null
Saving Mbango	null
Saw: The Final Chapter	null
The Business of Show Business	55

19.

a.



```
select avg(score) AS AVGSCORE from moviesdata;
```

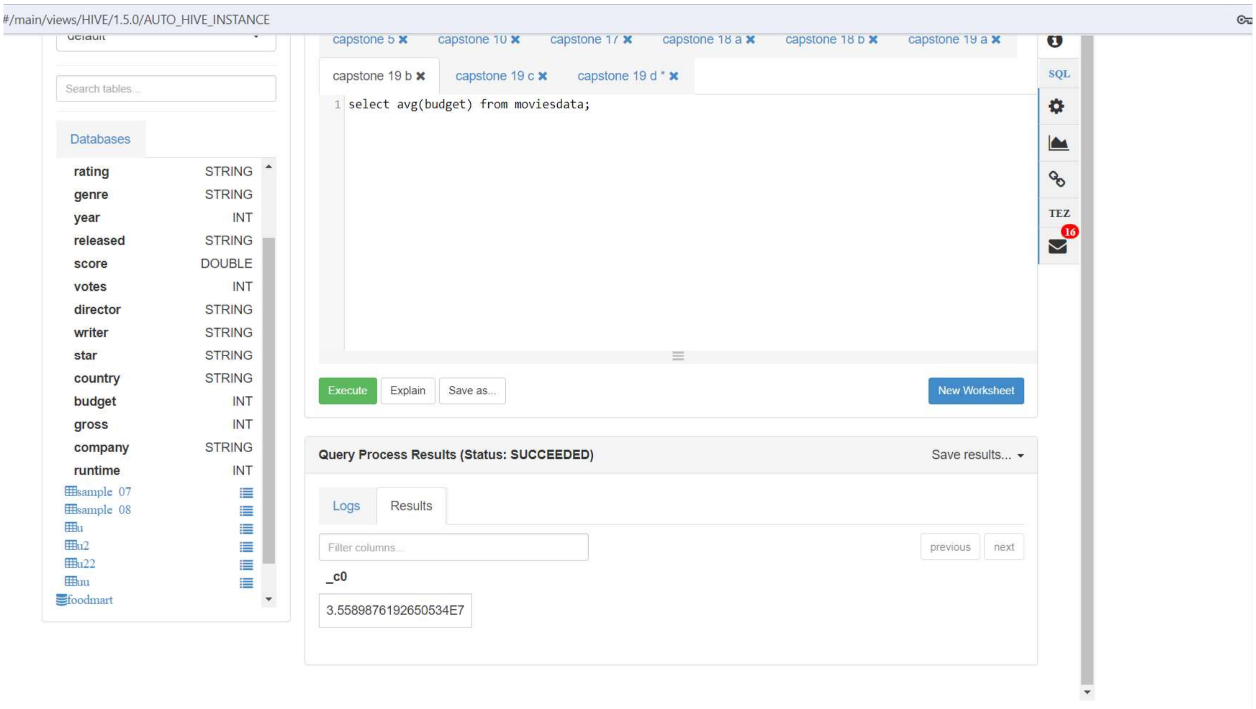
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

sample_07
sample_08
u
u2
u22
uu
foodmart

Query Process Results (Status: SUCCEEDED)

avgscore
6.390410958904098

b.



```
select avg(budget) from moviesdata;
```

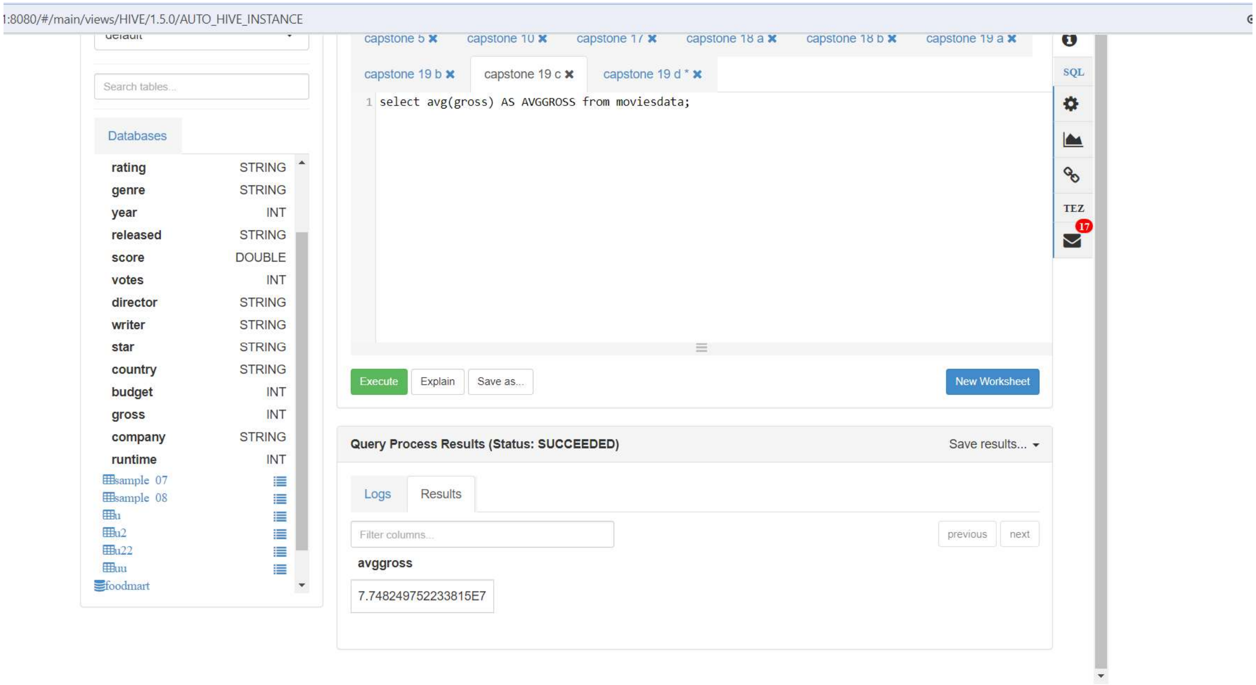
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

sample_07
sample_08
u
u2
u22
uu
foodmart

Query Process Results (Status: SUCCEEDED)

_c0
3.5589876192650534E7

c.



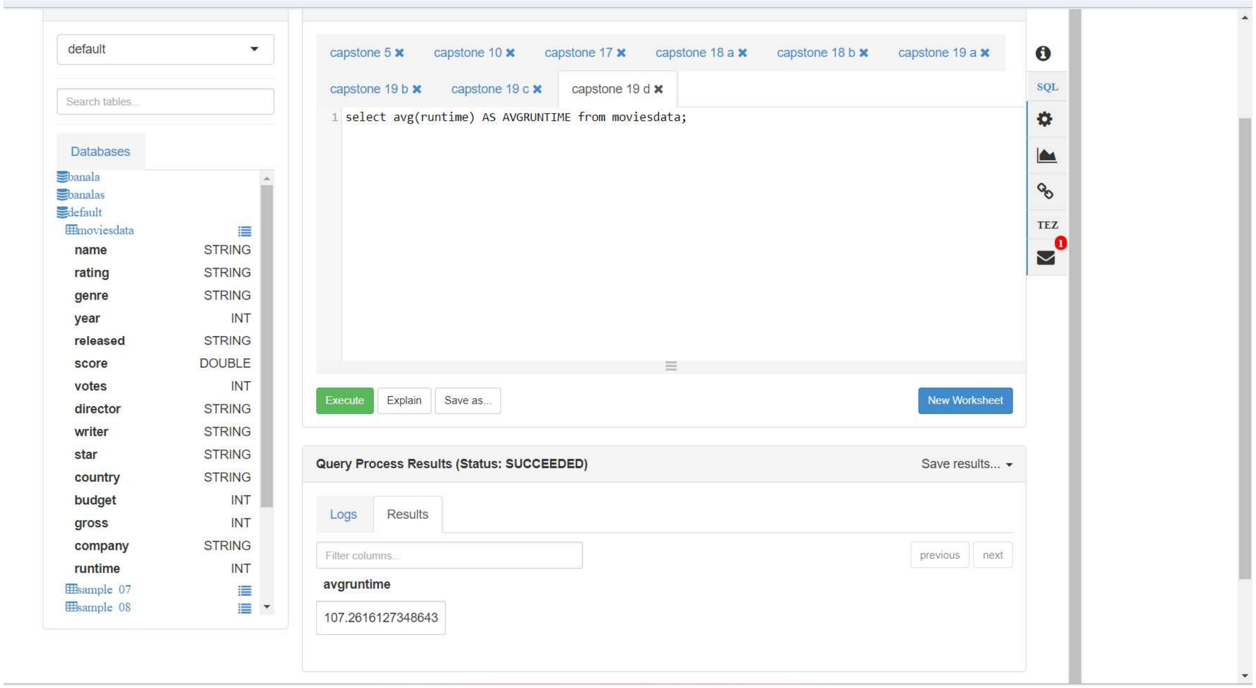
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

```
select avg(gross) AS AVGGROSS from moviesdata;
```

Query Process Results (Status: SUCCEEDED)

```
avggross
7.748249752233815E7
```

d.



name	STRING
rating	STRING
genre	STRING
year	INT
released	STRING
score	DOUBLE
votes	INT
director	STRING
writer	STRING
star	STRING
country	STRING
budget	INT
gross	INT
company	STRING
runtime	INT

```
select avg(runtime) AS AVGRUNTIME from moviesdata;
```

Query Process Results (Status: SUCCEEDED)

```
avgruntime
107.2616127348643
```