

BIG DATA CAPSTONE

1.

Search tables...

Databases

- banala
- banalas
- default
- 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

1 select * from moviesdata;

Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED)

Save results...

Logs Results

Filter columns... previous next

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.

hain/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

Search tables...

Databases

- banala
- banalas
- default
- 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

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

Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns... previous next

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.

Ambari - Sandbox - Google Chrome

127.0.0.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

Search tables...

Databases

- banala
- banalas
- default
- foodmart
- shyni
- xademo

Worksheet capstone 5 capstone 3

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

Execute Explain Save as... New Worksheet

100%

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

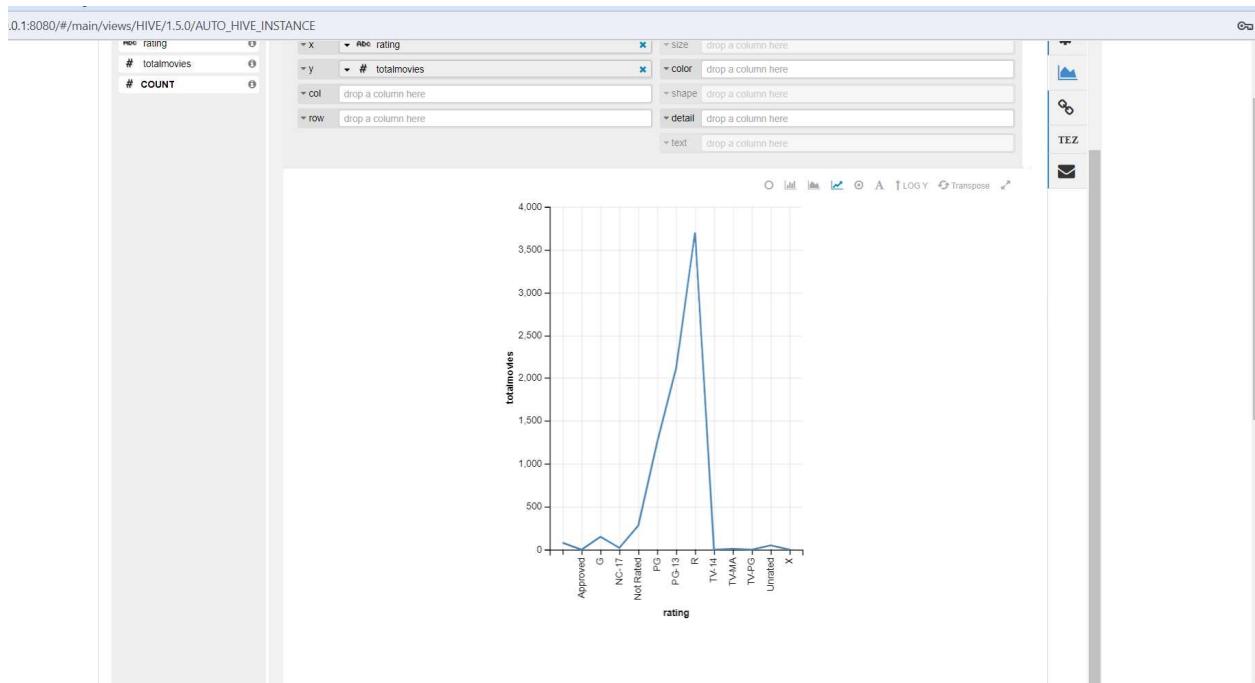
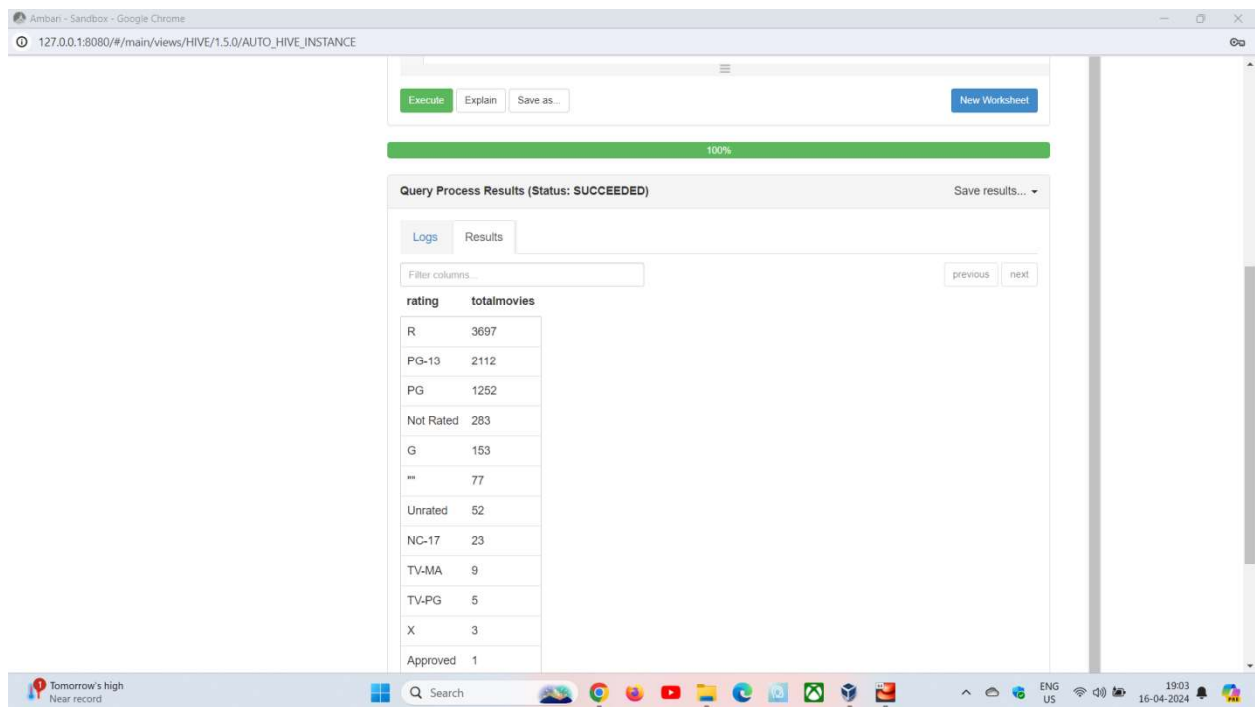
Filter columns... previous next

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

Tomorrow's high
Near record

Search

ENG US 19:03 16-04-2024



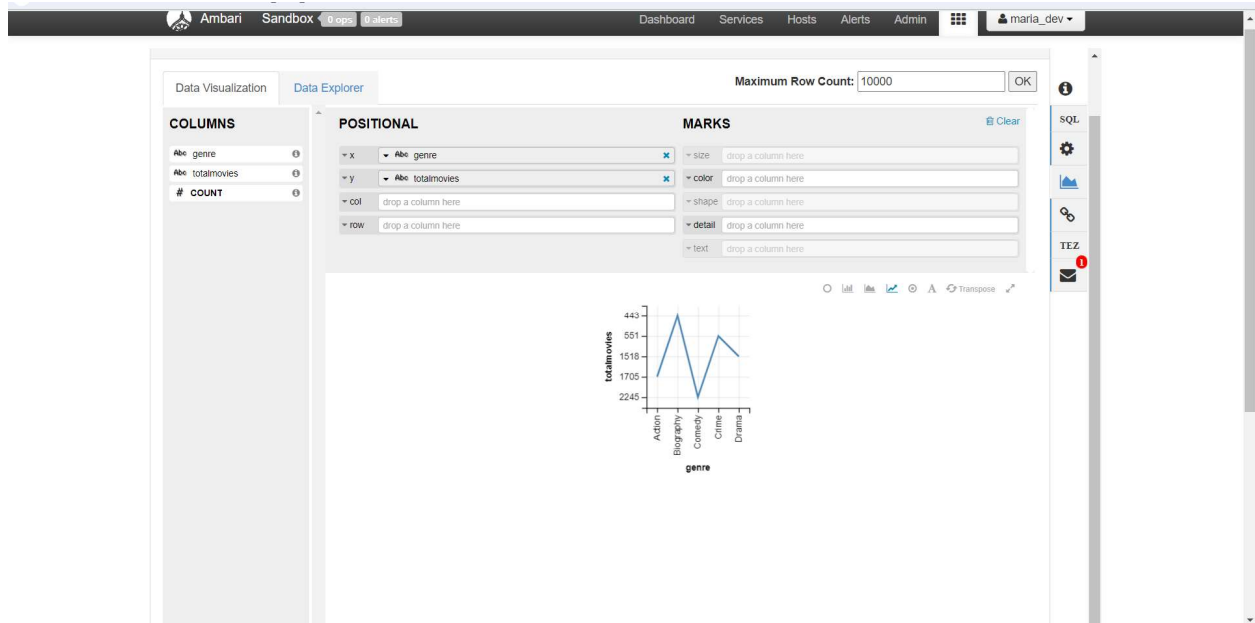
- From the above visualisation I can conclude that top 3 ratings maximum number of movies are
1. R (3697)
 2. PG_13(2112)
 3. PG(1252)

- 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.

The screenshot shows a web interface for executing Hive queries. The top bar displays the URL: 127.0.0.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE. On the left, a sidebar lists databases: vanala, vanalas, default, foodmart, hyni, and xademo. The main area contains a SQL query editor with the following query: `1 select genre,count(*) AS TOTALMOVIES from moviesdata group by genre order by TOTALMOVIES DESC limit 5`. Below the editor are buttons for 'Execute', 'Explain', 'Save as...', and 'New Worksheet'. The 'Query Process Results (Status: SUCCEEDED)' section shows a table with the following data:

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



- 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.

127.0.0.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

Ambari

Sandbox

9 apps

Alerts

Dashboard

Services

Hosts

Alerts

Admin

maria_dev

Hive

Query

Saved Queries

History

UDFs

Upload Table

Database Explorer

default

Search tables...

Databases

vanala

vanalas

default

foodmart

styni

xademo

Query Editor

Worksheet

capstone 3

capstone 4

capstone 13

capstone 5

capstone 12

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

Execute

Explain

Save as...

New Worksheet

100%

Query Process Results (Status: SUCCEEDED)

Save results...

Logs

Results

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

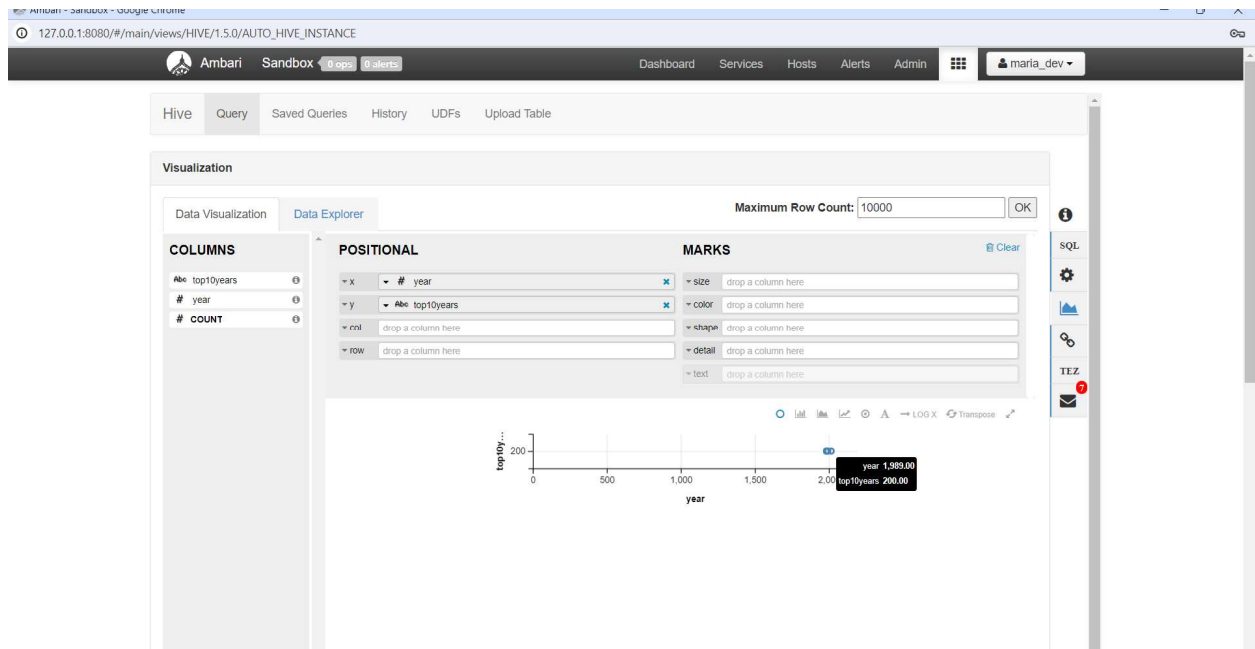
Results

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 .

127.0.0.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

Databases

bookata

bookalas

default

foodmart

olympi

xademo

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

Execute

Explain

Save as...

New Worksheet

100%

Query Process Results (Status: SUCCEEDED)

Save results... ▾

Logs

Results

Filter columns...

previous

next

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.

127.0.0.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

Search tables...

Databases

- analala
- analas
- default
- foodmart
- hynui
- mademo

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

Execute Explain Save as... New Worksheet

100%

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

previous next

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

127.0.0.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

Hive Query Saved Queries History UDFs Upload Table

Visualization

Data Visualization Data Explorer Maximum Row Count: 10000 OK

COLUMNS

- Abe name
- Abe votes
- # COUNT

POSITIONAL

- x Abe name
- y Abe votes
- 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

votes

name

- From the above visualization I can conclude that the total votes polled for each movie were revealed .
- 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.

The screenshot shows a web-based interface for executing Hive queries. On the left, a sidebar lists databases and tables, including 'moviesdata' with columns like name, rating, genre, year, released, score, votes, director, writer, star, country, budget, gross, company, and runtime. The main area displays a SQL query: '1 select director,count(*) AS TOTALMOVIES from moviesdata group by director order by TOTALMOVIES desc;'. Below the query, there are buttons for 'Execute', 'Explain', 'Save as...', and 'New Worksheet'. The 'Query Process Results (Status: SUCCEEDED)' section shows a table of results with columns 'director' and 'totalmovies'.

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

9.

Search tables...

Databases

- banala
- banalas
- default
- 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

```
1 select count(*) AS totalmovies ,writer from moviesdata group by writer order by totalmovies desc limit 5
```

Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

previous next

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

10.

Search tables...

Databases

- banala
- banalas
- default
- 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

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

Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

previous next

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

Search tables...

Databases

banalas

banalas

default

moviesdata

name

rating

genre

year

released

score

votes

director

writer

star

country

budget

gross

company

runtime

sample 07

sample 08

STRING

STRING

STRING

INT

STRING

DOUBLE

INT

STRING

STRING

STRING

STRING

INT

INT

STRING

INT

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

SQL

TEZ

Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED)

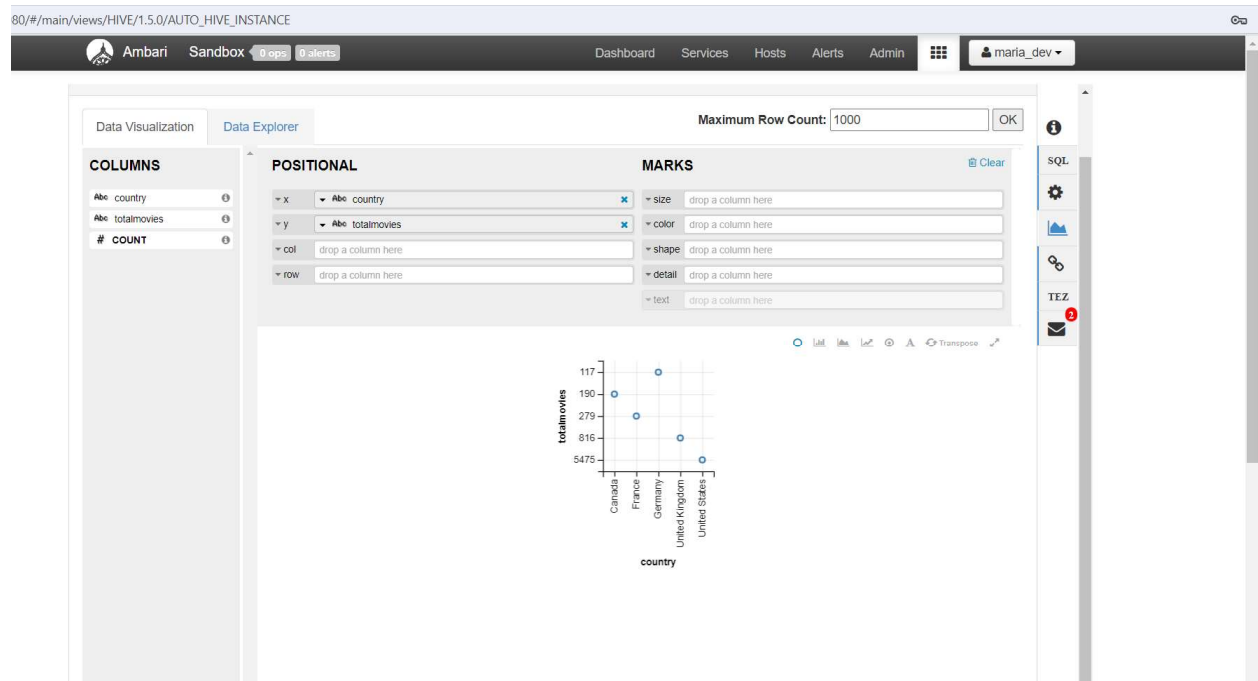
Save results...

Logs Results

Filter columns...

previous next

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



- 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.

127.0.0.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

default

Search tables...

Databases

- banana
- bananas
- default
- foodmart
- hynui
- xademo

Worksheet x capstone 3 x capstone 4 x capstone 13 x capstone 5 x capstone 12 x

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

Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

previous next

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

13.

1.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

default

Search tables...

Databases

hadoop

hadoop

default

moviesdata

name

rating

genre

year

released

score

votes

director

writer

star

country

budget

gross

company

runtime

sample 07

sample 08

STRING

STRING

STRING

INT

STRING

DOUBLE

INT

STRING

STRING

STRING

INT

INT

STRING

INT

capstone 5 x capstone 9 x capstone 10 x capstone 11 x capstone 12 x

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

Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

previous next

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.

30/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

default

Search tables...

Databases

hadoop

hadoop

default

moviesdata

name

rating

genre

year

released

score

votes

director

writer

star

country

budget

gross

company

runtime

sample 07

sample 08

STRING

STRING

STRING

INT

STRING

DOUBLE

INT

STRING

STRING

STRING

INT

INT

STRING

INT

capstone 5 x capstone 9 x capstone 10 x capstone 11 x capstone 13 x Worksheet (14) x

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

Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

previous next

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

15.

1.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

capstone 5 ✕ capstone 9 ✕ capstone 10 ✕ capstone 11 ✕ capstone 13 ✕ capstone 14 ✕

Worksheet (15) ✕

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

Execute Explain Save as... New Worksheet

100%

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

name	star	budget
Fame	Eddie Barth	null

previous next

Databases

Search tables...

- banala
- banalas
- default
- 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

SQL

TEZ 17

16.

Ambari - Sandbox - Google Chrome

127.0.0.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

default

Search tables...

Databases

- vanala
- vanalas
- default
- 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

capstone 5 ✕ capstone 9 ✕ capstone 10 ✕ capstone 11 ✕ capstone 13 ✕ capstone 14 ✕

capstone 15 ✕ Worksheet (17) ✕ Worksheet (18) * ✕

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

Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

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

Ambari - Sandbox - Google Chrome

127.0.0.1:8080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

default

Search tables...

Databases

- vanala
- vanalas
- default
- 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

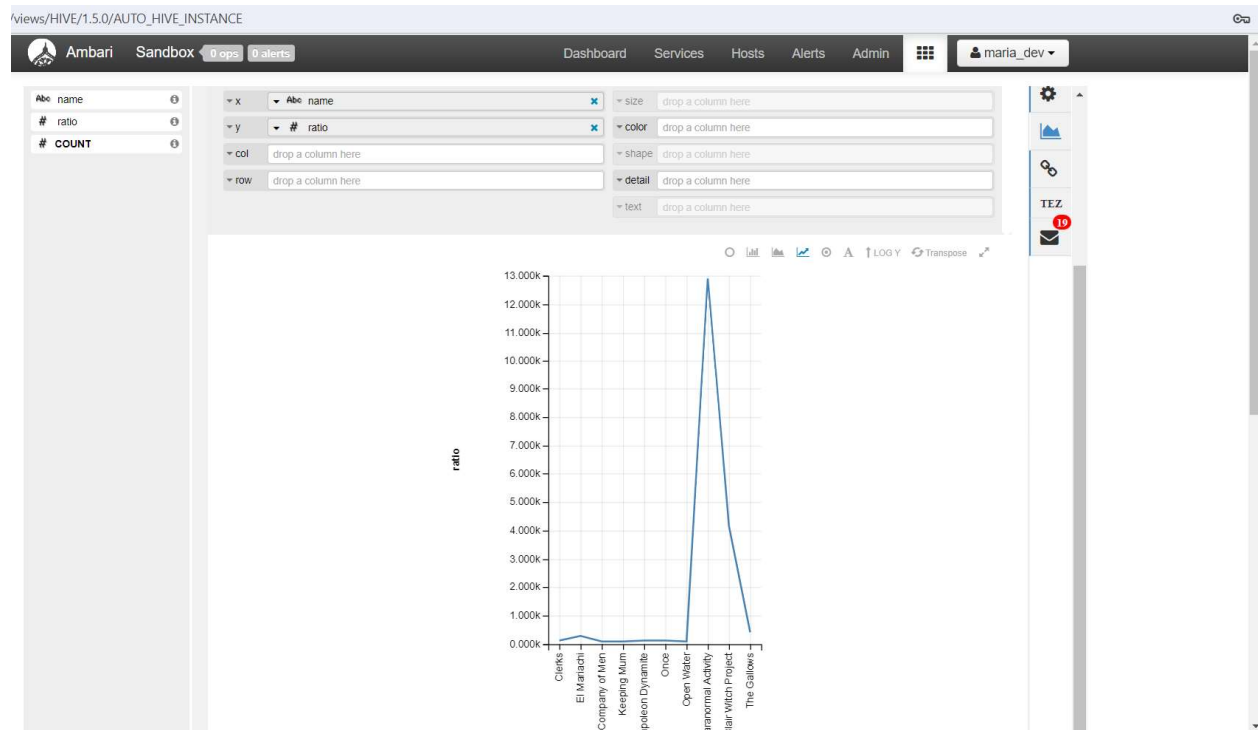
Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

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



- 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)

main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

capstone 5 ✕ capstone 9 ✕ capstone 10 ✕ capstone 11 ✕ capstone 13 ✕ capstone 14 ✕

capstone 15 ✕ capstone 17 ✕ Worksheet (18) ✕

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

Execute Explain Save as... New Worksheet

100%

Query Process Results (Status: SUCCEEDED) Save results... ▾

Logs Results

Filter columns...

previous next

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

main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

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

sample 07 sample 08

Execute Explain Save as... New Worksheet

100%

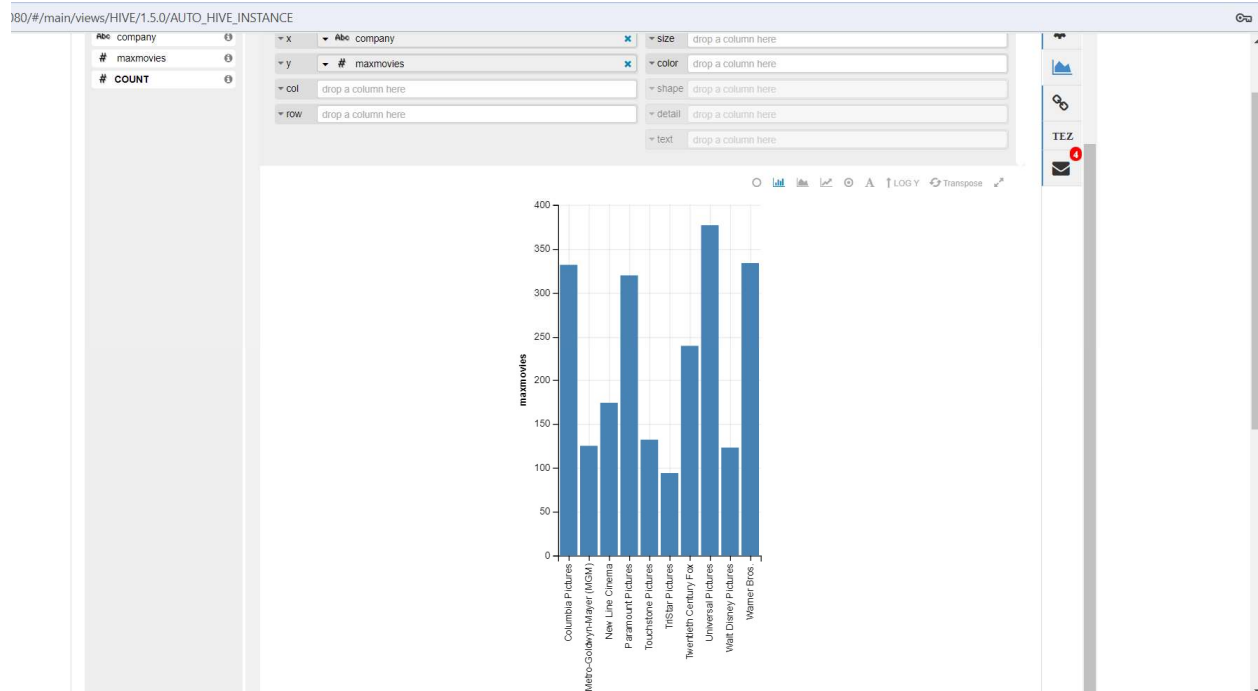
Query Process Results (Status: SUCCEEDED) Save results... ▾

Logs Results

Filter columns...

previous next

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



- 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 revealed .

080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

Search tables...

Databases

rating

genre

year

released

score

votes

director

writer

star

country

budget

gross

company

runtime

sample 07

sample 08

u1

u2

u22

u1

foodmart

STRING

STRING

INT

STRING

DOUBLE

INT

STRING

STRING

STRING

STRING

INT

INT

STRING

INT

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

Execute

Explain

Save as...

New Worksheet

Query Process Results (Status: SUCCEEDED)

Save results...

LogsResults

Filter columns...

previousnext

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

080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

Search tables...

Databases

rating

genre

year

released

score

votes

director

writer

star

country

budget

gross

company

runtime

sample 07

sample 08

u1

u2

u22

u1

foodmart

STRING

STRING

INT

STRING

DOUBLE

INT

STRING

STRING

STRING

STRING

INT

INT

STRING

INT

capstone 15 x capstone 17 x capstone 17 x capstone 18 a x Worksheet (20) x Worksheet (21) x SQL

1 select name, runtime from moviesdata order by runtime;

Execute

Explain

Save as...

New Worksheet

Query Process Results (Status: SUCCEEDED)

Save results...

LogsResults

Filter columns...

previousnext

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.

Default

Search tables...

Databases

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

bu

bu2

bu22

bu

foodmart

capstone 5

capstone 10

capstone 17

capstone 18 a

capstone 18 b

capstone 19 a

capstone 19 b

capstone 19 c

capstone 19 d

```
1 select avg(score) AS AVGSORE from moviesdata;
```

Execute Explain Save as...

New Worksheet

100%

Query Process Results (Status: SUCCEEDED)

Save results...

Logs Results

Filter columns...

previous next

avgscore

6.390410958904098

b.

Default

Search tables...

Databases

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

bu

bu2

bu22

bu

foodmart

capstone 5

capstone 10

capstone 17

capstone 18 a

capstone 18 b

capstone 19 a

capstone 19 b

capstone 19 c

capstone 19 d

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

Execute Explain Save as...

New Worksheet

Query Process Results (Status: SUCCEEDED)

Save results...

Logs Results

Filter columns...

previous next

_c0

3.5589876192650534E7

C.

1:3080/#/main/views/HIVE/1.5.0/AUTO_HIVE_INSTANCE

default

Search tables ...

Databases

rating

genre

year

released

score

votes

director

writer

star

country

budget

gross

company

runtime

sample 07

sample 08

h1

h2

h22

h3

foodmart

STRING

STRING

INT

STRING

DOUBLE

INT

STRING

STRING

STRING

STRING

INT

INT

STRING

INT

INT

INT

INT

INT

INT

INT

capstone 5

capstone 10

capstone 17

capstone 18 a

capstone 18 b

capstone 19 a

capstone 19 b

capstone 19 c

capstone 19 d

SQL

TEZ

1 select avg(gross) AS AVGGROSS from moviesdata;

Execute

Explain

Save as...

New Worksheet

Query Process Results (Status: SUCCEEDED)

Save results...

Logs

Results

Filter columns...

previous

next

avggross

7.748249752233815E7

d.

default

Search tables...

Databases

- banala
- banalas
- default
- 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

capstone 5 ✕ capstone 10 ✕ capstone 17 ✕ capstone 18 a ✕ capstone 18 b ✕ capstone 19 a ✕

capstone 19 b ✕ capstone 19 c ✕ capstone 19 d ✕

1 select avg(runtime) AS AVGRUNTIME from moviesdata;

Execute Explain Save as... New Worksheet

Query Process Results (Status: SUCCEEDED) Save results...

Logs Results

Filter columns...

avgruntime

107.2616127348643

previous next

SQL

TEZ

