

Preparing Dataset.

Step 1:

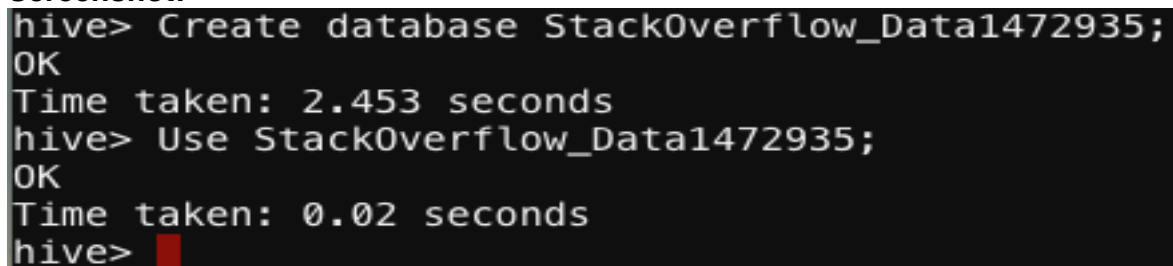
Create Database.

Command:

```
Create database StackOverflow_Data1472935;
```

```
Use StackOverflow_Data1472935;
```

Screenshot:



```
hive> Create database StackOverflow_Data1472935;  
OK  
Time taken: 2.453 seconds  
hive> Use StackOverflow_Data1472935;  
OK  
Time taken: 0.02 seconds  
hive> █
```

Step 2:

Create four tables.

Command:

```
create table comments1472935 (id int, userid int) row format delimited fields terminated by  
' ';
```

```
create table posts1472935 (id int, post_type smallint, creationdate timestamp, score int,  
viewcount int, owneruserid int, title string, answercount int, commentcount int) row format  
delimited fields terminated by ' ';
```

```
create table posttypes1472935 (id int, name string) row format delimited fields terminated  
by ' ';
```

```
create table users1472935 (id int, reputation int, displayname string, loc string, age tinyint)  
row format delimited fields terminated by ' ';
```

Screenshot:

```
hive> create table comments1472935 (id int, userid int) row format delimited fields terminated by ',';
OK
Time taken: 0.03 seconds
hive>
hive> create table posts1472935 (id int, post_type smallint, creationdate timestamp, score int, viewcount int, owneruserid int, title string, answercount int, commentcount int) row format delimited fields terminated by ',';
OK
Time taken: 0.052 seconds
hive>
hive> create table posttypes1472935 (id int, name string) row format delimited fields terminated by ',';
OK
Time taken: 0.034 seconds
hive>
hive> create table users1472935 (id int, reputation int, displayname string, loc string, age tinyint) row format delimited fields terminated by ',';
OK
Time taken: 0.037 seconds
hive>
```

Step 3:

Load dataset into created tables.

Command:

```
load data local inpath
"/mnt/home/edureka_1472935/Stack_Overflow_Dataset/comments.csv" overwrite into table
comments1472935;
```

```
load data local inpath "/mnt/home/edureka_1472935/Stack_Overflow_Dataset/posts.csv"
overwrite into table posts1472935;
```

```
load data local inpath
"/mnt/home/edureka_1472935/Stack_Overflow_Dataset/posttypes.csv" overwrite into table
posttypes1472935;
```

```
load data local inpath "/mnt/home/edureka_1472935/Stack_Overflow_Dataset/users.csv"
overwrite into table users1472935;
```

Screenshots:

```
hive> load data local inpath "/mnt/home/edureka_1472935/Stack_Overflow_Da
taset/comments.csv" overwrite into table comments1472935;
Loading data to table stackoverflow_data1472935.comments1472935
Table stackoverflow_data1472935.comments1472935 stats: [numFiles=1, numRo
ws=0, totalSize=804550, rawDataSize=0]
OK
Time taken: 0.263 seconds
hive>
hive> load data local inpath "/mnt/home/edureka_1472935/Stack_Overflow_Da
taset/posts.csv" overwrite into table posts1472935;
Loading data to table stackoverflow_data1472935.posts1472935
Table stackoverflow_data1472935.posts1472935 stats: [numFiles=1, numRows=
0, totalSize=2499773, rawDataSize=0]
OK
Time taken: 0.245 seconds
hive>
hive> load data local inpath "/mnt/home/edureka_1472935/Stack_Overflow_Da
taset/posttypes.csv" overwrite into table posttypes1472935;
Loading data to table stackoverflow_data1472935.posttypes1472935
Table stackoverflow_data1472935.posttypes1472935 stats: [numFiles=1, numR
ows=0, totalSize=116, rawDataSize=0]
OK
Time taken: 0.256 seconds
hive>
hive> load data local inpath "/mnt/home/edureka_1472935/Stack_Overflow_Da
taset/users.csv" overwrite into table users1472935;
Loading data to table stackoverflow_data1472935.users1472935
Table stackoverflow_data1472935.users1472935 stats: [numFiles=1, numRows=
0, totalSize=1559871, rawDataSize=0]
OK
Time taken: 0.242 seconds
hive>
```

Solutions:

- A. Find the display name and no. of posts created by the user who has got maximum reputation.**

Command:

```
select displayname, reputation from users1472935 group by displayname, reputation order
by reputation desc limit 1;
```

Screenshots:

```
hive> select displayname, reputation from users1472935 group by displayname,
  reputation order by reputation desc limit 1;
Query ID = edureka_1472935_20210427155151_1222f2c2-b2ff-42ae-89e8-0c1d89522c68
Total jobs = 2
Launching Job 1 out of 2
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_1616262377730_6896, Tracking URL = http://ip-20-0-21-161.
ec2.internal:8088/proxy/application_1616262377730_6896/
Kill Command = /opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/
bin/hadoop job -kill job_1616262377730_6896
```

Result: Jon Skeet with 736381 posts.

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.38 sec HDFS Read: 156
6238 HDFS Write: 1442879 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 6.59 sec HDFS Read: 144
7696 HDFS Write: 17 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 970 msec
OK
Jon Skeet 736381 ←
Time taken: 52.476 seconds, Fetched: 1 row(s)
hive> █
```

B. Find the average age of users on the Stack Overflow site.

Command:

```
select avg(age) from users1472935;
```

Screenshots:

```
hive> select avg(age) from users1472935;
Query ID = edureka_1472935_20210427155454_5225c7a8-10fe-4834-b1df-ffdd9d7ca2dd
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_1616262377730_6900, Tracking URL = http://ip-20-0-21-161.ec2.internal:8088/proxy/application_1616262377730_6900/
Kill Command = /opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/bin/hadoop job -kill job_1616262377730_6900
```

Results: Average age is 35.

```
MapReduce Total cumulative CPU time: 5 seconds 860 msec
Ended Job = job_1616262377730_6900
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.86 sec HDFS Read: 1567308 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 860 msec
OK
35.263352397712275 ←
Time taken: 36.129 seconds, Fetched: 1 row(s)
hive> 
```

C. Find the display name of user who posted the oldest post on Stack Overflow (in terms of date).

Command:

```
select u.displayname, p.creationdate from users1472935 u join posts1472935 p on (u.id = p.owneruserid) order by p.creationdate limit 1;
```

Screenshots:

```
hive> select u.displayname, p.creationdate from users1472935 u join posts1472935 p on (u.id = p.owneruserid) order by p.creationdate limit 1;
Query ID = edureka_1472935_20210427160000_a7775116-d639-454c-ba6c-d19f97fa8361
Total jobs = 1
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
```

Result: Eggs McLaren posted the post on 2008-7-31-21:42:52.

```

MapReduce Total cumulative CPU time: 8 seconds 490 msec
Ended Job = job_1616262377730_6903
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 8.49 sec HDFS Read: 251
0610 HDFS Write: 33 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 490 msec
OK
Eggs McLaren 2008-07-31 21:42:52 ←
Time taken: 39.758 seconds, Fetched: 1 row(s)
hive> █

```

D. Find the display name and no. of comments done by the user who has got maximum reputation.

Command:

```

select u.displayname, p.commentcount, max(u.reputation) as repu from users1472935 u
join posts1472935 p on u.id = p.owneruserid join comments1472935 c on c.userid =
p.owneruserid group by u.displayname, p.commentcount order by repu desc limit 1;

```

Screenshots:

```

hive> select u.displayname, p.commentcount, max(u.reputation) as repu from u
sers1472935 u join posts1472935 p on u.id = p.owneruserid join comments14729
35 c on c.userid = p.owneruserid group by u.displayname, p.commentcount orde
r by repu desc limit 1;
Query ID = edureka_1472935_20210427161515_ae20f47f-e5eb-485a-9c4a-3e88544e47
1b
Total jobs = 2
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M;
support was removed in 8.0
Execution log at: /tmp/edureka_1472935/edureka_1472935_20210427161515_ae20f4
7f-e5eb-485a-9c4a-3e88544e471b.log

```

Result:

```

MapReduce Total cumulative CPU time: 5 seconds 90 msec
Ended Job = job_1616262377730_6917
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 9.13 sec HDFS Read: 251
3449 HDFS Write: 96862 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 5.09 sec HDFS Read: 102
080 HDFS Write: 20 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 220 msec
OK
Jon Skeet NULL 736381 ←
Time taken: 44.557 seconds, Fetched: 1 row(s)
hive> █

```

E. 1) Find the display name of user who has created maximum no. of posts on Stack Overflow.

Command:


```
select u.displayname, count(*) as count from users1472935 u join posts1472935 p on
p.owneruserid = u.id group by u.displayname, p.owneruserid order by count desc limit 1;
```

Screenshots:

```
hive> select u.displayname, count(*) as count from users1472935 u join posts
1472935 p on p.owneruserid = u.id group by u.displayname, p.owneruserid orde
r by count desc limit 1;
Query ID = edureka_1472935_20210427163636_147a31d5-c596-4687-9ad0-a01253e209
2d
Total jobs = 2
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M;
support was removed in 8.0
```

Result:

```
MapReduce Total cumulative CPU time: 4 seconds 820 msec
Ended Job = job_1616262377730_6928
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 8.04 sec HDFS Read: 251
0863 HDFS Write: 177493 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 4.82 sec HDFS Read: 182
188 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 860 msec
OK
aku 274
Time taken: 71.514 seconds, Fetched: 1 row(s)
hive>
```

2) Find the display name of user who has commented maximum no. of posts on Stack Overflow.

Command:

```
select u.displayname, p.commentcount from users1472935 u join posts1472935 p on
p.owneruserid = u.id group by u.displayname, p.commentcount order by p.commentcount
desc limit 1;
```

Screenshots:

```
hive> select u.displayname, p.commentcount from users1472935 u join posts147
2935 p on p.owneruserid = u.id group by u.displayname, p.commentcount order
by p.commentcount desc limit 1;
Query ID = edureka_1472935_20210427163737_b9a5e6cc-871f-4e00-8329-3a95596f6e
8a
Total jobs = 2
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M;
support was removed in 8.0
```

Result:

```

MapReduce Total cumulative CPU time: 5 seconds 180 msec
Ended Job = job_1616262377730_6931
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 7.97 sec HDFS Read: 251
0062 HDFS Write: 363696 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 5.18 sec HDFS Read: 368
558 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 150 msec
OK
Justin Standard 57
Time taken: 40.712 seconds, Fetched: 1 row(s)
hive>

```

F. Find the owner name and id of user whose post has got maximum no. of view counts so far.

Command:

```
select u.displayname, u.id, p.viewcount from users1472935 u join posts1472935 p on u.id =
p.owneruserid order by p.viewcount desc limit 1 ;
```

Screenshots:

```

hive> select u.displayname, u.id, p.viewcount from users1472935 u join posts
1472935 p on u.id = p.owneruserid order by p.viewcount desc limit 1 ;
Query ID = edureka_1472935_20210427164141_7c30fd79-0ac1-4f1e-ab9d-2a044e4a30
9a
Total jobs = 1
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M;
support was removed in 8.0

```

Result: Owner name: Shadow_x99, ID: 244, viewcounts: 758492

```

MapReduce Total cumulative CPU time: 7 seconds 470 msec
Ended Job = job_1616262377730_6933
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 7.47 sec HDFS Read: 251
0889 HDFS Write: 22 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 470 msec
OK
Shadow_x99      244      758492
Time taken: 37.889 seconds, Fetched: 1 row(s)
hive>

```

G. 1) Find the title and owner name of post who has got maximum no. of Answer count.

Command:

```
select u.displayname, p.title, p.answercount from users1472935 u join posts1472935 p on
u.id = p.owneruserid order by p.answercount desc limit 1;
```


Screenshots:

```
hive> select u.displayname, p.title, p.answercount from users1472935 u join
posts1472935 p on u.id = p.owneruserid order by p.answercount desc limit 1;

Query ID = edureka_1472935_20210427164545_d34fc30e-b094-4814-b637-3186d1642b
a2
Total jobs = 1
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M;
support was removed in 8.0
```

Result:

```
MapReduce Total cumulative CPU time: 7 seconds 860 msec
Ended Job = job_1616262377730_6934
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 7.86 sec HDFS Read: 251
1018 HDFS Write: 70 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 860 msec
OK
Charles Roper What non-programming books should programmers read? 316
Time taken: 44.142 seconds, Fetched: 1 row(s)
hive>
```

2) Find the title and owner name of post who has got maximum no. of
Comment count.

Command:

```
select p.title, u.displayname, p.commentcount from users1472935 u join posts1472935 p on u.id
= p.owneruserid where p.title != '' order by p.commentcount desc limit 1;
```

Screenshots:

```
hive> select p.title, u.displayname, p.commentcount from users1472935 u join
posts1472935 p on u.id = p.owneruserid where p.title != '' order by p.comm
entcount desc limit 1;
Query ID = edureka_1472935_20210427164949_ca6e35fb-55fa-40c4-bf70-eea589f4eb
3a
Total jobs = 1
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M;
support was removed in 8.0
```

Result:

```
MapReduce Total cumulative CPU time: 7 seconds 230 msec
Ended Job = job_1616262377730_6938
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 7.23 sec HDFS Read: 251
1491 HDFS Write: 84 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 230 msec
OK
What is the single most influential book every programmer should read? NotM
yself 35
Time taken: 42.22 seconds, Fetched: 1 row(s)
hive>
```

H. Find the location which has maximum no of Stack Overflow users.

Command:

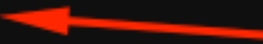
```
select loc, count(*) as count from users1472935 group by loc order by count desc limit 10;
```

Screenshots:

```
hive> select loc, count(*) as count from users1472935 group by loc order by
count desc limit 10;
Query ID = edureka_1472935_20210427170606_8fbdda2f-3f28-4581-8979-e4eb5e1cbc
77
Total jobs = 2
Launching Job 1 out of 2
```

Result:

```
20208
United States 2452
United Kingdom 1170
London United Kingdom 826
San Francisco CA 497
Seattle WA 430
Australia 424
New York NY 412
Germany 405
California 405
Time taken: 51.577 seconds, Fetched: 10 row(s)
hive> 
```



I. Find the total no. of answers, posts, comments created by Indian users.

Command:

```
select count(*) from posts1472935 p join users1472935 u on u.id = p.owneruserid where
p.post_type = 2 and u.loc == 'India';
```

```
select count(p.id) from posts1472935 p join users1472935 u on u.id = p.owneruserid where
u.loc == 'India';
```

```
select count(*) from comments1472935 c join users1472935 u on u.id = c.userid where u.loc ==
'India';
```

Screenshots:

```
hive> select count(*) from posts1472935 p join users1472935 u on u.id = p.owneruserid where p.post_type = 2 and u.loc == 'India';
Query ID = edureka_1472935_20210427171414_776e764a-0eaa-449c-abb3-710d9066b191
Total jobs = 1
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
```

```
hive> select count(p.id) from posts1472935 p join users1472935 u on u.id = p.owneruserid where u.loc == 'India';
Query ID = edureka_1472935_20210427171515_9be236c6-8ac9-40a8-948e-2bc37fe991a3
Total jobs = 1
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
```

```
hive> select count(*) from comments1472935 c join users1472935 u on u.id = c.userid where u.loc == 'India';
Query ID = edureka_1472935_20210427171818_79c6d5d3-2da4-41f3-8611-ca1f2792f698
Total jobs = 1
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
```

Results:

```
MapReduce Total cumulative CPU time: 6 seconds 910 msec
Ended Job = job_1616262377730_6946
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 6.91 sec HDFS Read: 2512192 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 910 msec
OK
31 ← Answer
Time taken: 40.927 seconds, Fetched: 1 row(s)
hive> █
```

```
MapReduce Total cumulative CPU time: 6 seconds 400 msec
Ended Job = job_1616262377730_6949
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 6.4 sec HDFS Read: 2510571 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 400 msec
OK
62 ← no. of posts
Time taken: 36.283 seconds, Fetched: 1 row(s)
hive> █
```

```
MapReduce Total cumulative CPU time: 6 seconds 0 msec
Ended Job = job_1616262377730_6951
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 6.0 sec HDFS Read: 1570
465 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 0 msec
OK
150 ←
Time taken: 26.29 seconds, Fetched: 1 row(s)
hive> 
```

no. of comments