Supporting Documentation: Task 2 and 3
Student: Úna McGinn
CA675 Assignment 1

Task 2:

Load Data to Pig:

```
Connected, host fingerprint: ssh-rsa 0 95:70:B0:2A:65:9F:43:05:C3:22:8E:46:D4:07
:45:A4:EF:CF:5F:B4:A8:B8:1B:D7:50:AE:72:AD:FA:4D:8C:2D
Linux cluster-e9cf-m 5.10.0-0.bpo.8-amd64 #1 SMP Debian 5.10.46-4~bpo10+1 (2021-
08-07) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
una_mcginn2@cluster-e9cf-m:~$ whoami
una_mcginn2
una_mcginn2@cluster-e9cf-m:~$ hadoop fs -mkdir -p /user/una_mcginn2
una_mcginn2@cluster-e9cf-m:~$ export PATH=${JAVA_HOME}/bin:${PATH}
una_mcginn2@cluster-e9cf-m:~$ export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
una_mcginn2@cluster-e9cf-m:~$ hadoop fs -ls
una_mcginn2@cluster-e9cf-m:~$ env
```

-- Setup new directory

```
una_mcginn2@cluster-e9cf-m:~$ hadoop fs -mkdir -p /user/una_mcginn2
```

```
una_mcginn2@cluster-e9cf-m:~$ hadoop fs -mkdir /PigData
```

-- The data was loaded to the Pig Data folder.

hadoop fs -put /home/una_mcginn2/QueryResults1.csv /PigData
hadoop fs -put /home/una_mcginn2/QueryResults2.csv /PigData
hadoop fs -put /home/una_mcginn2/QueryResults3.csv /PigData
hadoop fs -put /home/una_mcginn2/QueryResults4.csv /PigData
hadoop fs -put /home/una_mcginn2/QueryResults5.csv /PigData

```
una_mcginn2@cluster-e9cf-m:~$ hadoop fs -put /home/una_mcginn2/QueryResults1.csv /PigData
```

```
una_mcginn2@cluster-e9cf-m:~$ hadoop fs -put /home/una_mcginn2/QueryResults2.csv /PigData
una_mcginn2@cluster-e9cf-m:~$ hadoop fs -put /home/una_mcginn2/QueryResults3.csv /PigData
una_mcginn2@cluster-e9cf-m:~$ hadoop fs -put /home/una_mcginn2/QueryResults4.csv /PigData
una_mcginn2@cluster-e9cf-m:~$ hadoop fs -put /home/una_mcginn2/QueryResults5.csv /PigData
```

```
REGISTER /usr/lib/pig/piggybank.jar;
```

Source: https://stackoverflow.com/questions/31460632/reading-a-csv-file-in-pig

Code in Pig:

```
A = LOAD '/PigData/QueryResults1.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE') AS
(Id:int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:chararr
ay,DeletionDate:chararray,Score:int,ViewCount:int,Body:chararray,OwnerUserId:
int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:cha
rarray,LastEditDate:chararray,LastActivityDate:chararray,Title:chararray,Tags
:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:char
array,CommunityOwnedDate:chararray,ContendLicense:chararray);

B = LOAD '/PigData/QueryResults2.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE') AS
(Id:int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:chararr
ay,DeletionDate:chararray,Score:int,ViewCount:int,Body:chararray,OwnerUserId:
```

```
int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:cha
rarray,LastEditDate:chararray,LastActivityDate:chararray,Title:chararray,Tags
:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:char
array,CommunityOwnedDate:chararray,ContendLicense:chararray);

C = LOAD '/PigData/QueryResults3.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE') AS
(Id:int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:chararr
ay,DeletionDate:chararray,Score:int,ViewCount:int,Body:chararray,OwnerUserId:
int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:cha
rarray,LastEditDate:chararray,LastActivityDate:chararray,Title:chararray,Tags
:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:char
array,CommunityOwnedDate:chararray,ContendLicense:chararray);

D = LOAD '/PigData/QueryResults4.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE') AS
(Id:int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:chararr
ay,DeletionDate:chararray,Score:int,ViewCount:int,Body:chararray,OwnerUserId:
int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:cha
rarray,LastEditDate:chararray,LastActivityDate:chararray,Title:chararray,Tags
:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:char
array,CommunityOwnedDate:chararray,ContendLicense:chararray);

E = LOAD '/PigData/QueryResults5.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE') AS
(Id:int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:chararr
ay,DeletionDate:chararray,Score:int,ViewCount:int,Body:chararray,OwnerUserId:
int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:cha
rarray,LastEditDate:chararray,LastActivityDate:chararray,Title:chararray,Tags
:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:char
array,CommunityOwnedDate:chararray,ContendLicense:chararray);
```

```
grunt> E = LOAD '/PigData/QueryResults5.csv
' USING org.apache.pig.piggybank.storage.CS
VExcelStorage(',', 'YES_MULTILINE') AS (Id:
int,PostTypeId:int,AcceptedAnswerId:int,Par
entId:int,CreationDate:chararray,DeletionDa
te:chararray,Score:int,ViewCount:int,Body:c
hararray,OwnerUserId:int,OwnerDisplayName:c
hararray,LastEditorUserId:int,LastEditorDis
playName:chararray,LastEditDate:chararray,L
astActivityDate:chararray,Title:chararray,T
ags:chararray,AnswerCount:int,CommentCount:
int,FavoriteCount:int,ClosedDate:chararray,
CommunityOwnedDate:chararray,ContendLicense
:chararray);
2021-10-28 19:26:15,421 [main] INFO  org.ap
ache.hadoop.conf.Configuration.deprecation
- yarn.resourcemanager.system-metrics-publi
sher.enabled is deprecated. Instead, use ya
rn.system-metrics-publisher.enabled
```

Data Cleaning:

-- The data is cleaned, accounting for line breaks, new lines, carriage returns and commas.

```
A1 = FOREACH A GENERATE Id, PostTypeId, AcceptedAnswerId, ParentId,
CreationDate, DeletionDate, Score, ViewCount,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Body,'\\n',''),'\\r',''),'\\r\\n','')
,'<br>',''),',',' ') AS Body, OwnerUserId,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(OwnerDisplayName,'\\n',''),'\\r',''),
'\\r\\n',''),'<br>',''),',',' ') AS OwnerDisplayName, LastEditorUserId,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(LastEditorDisplayName,'\\n',''),'\\r'
,''),'\\r\\n',''),'<br>',''),',',' ') AS LastEditorDisplayName, LastEditDate,
LastActivityDate,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Title,'\\n',''),'\\r',''),'\\r\\n',''
),'<br>',''),',',' ') AS Title,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Tags,'\\n',''),'\\r',''),'\\r\\n','')
,'<br>',''),',',' ') AS Tags, AnswerCount, CommentCount, FavoriteCount,
ClosedDate, CommunityOwnedDate, ContendLicense;

B2 = FOREACH B GENERATE Id, PostTypeId, AcceptedAnswerId, ParentId,
CreationDate, DeletionDate, Score, ViewCount,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Body,'\\n',''),'\\r',''),'\\r\\n','')
,'<br>',''),',',' ') AS Body, OwnerUserId,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(OwnerDisplayName,'\\n',''),'\\r',''),
'\\r\\n',''),'<br>',''),',',' ') AS OwnerDisplayName, LastEditorUserId,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(LastEditorDisplayName,'\\n',''),'\\r'
,''),'\\r\\n',''),'<br>',''),',',' ') AS LastEditorDisplayName, LastEditDate,
LastActivityDate,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Title,'\\n',''),'\\r',''),'\\r\\n',''
),'<br>',''),',',' ') AS Title,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Tags,'\\n',''),'\\r',''),'\\r\\n','')
,'<br>',''),',',' ') AS Tags, AnswerCount, CommentCount, FavoriteCount,
ClosedDate, CommunityOwnedDate, ContendLicense;

C3 = FOREACH C GENERATE Id, PostTypeId, AcceptedAnswerId, ParentId,
CreationDate, DeletionDate, Score, ViewCount,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Body,'\\n',''),'\\r',''),'\\r\\n','')
,'<br>',''),',',' ') AS Body, OwnerUserId,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(OwnerDisplayName,'\\n',''),'\\r',''),
'\\r\\n',''),'<br>',''),',',' ') AS OwnerDisplayName, LastEditorUserId,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(LastEditorDisplayName,'\\n',''),'\\r'
,''),'\\r\\n',''),'<br>',''),',',' ') AS LastEditorDisplayName, LastEditDate,
LastActivityDate,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Title,'\\n',''),'\\r',''),'\\r\\n',''
),'<br>',''),',',' ') AS Title,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Tags,'\\n',''),'\\r',''),'\\r\\n','')
,'<br>',''),',',' ') AS Tags, AnswerCount, CommentCount, FavoriteCount,
ClosedDate, CommunityOwnedDate, ContendLicense;

D4 = FOREACH D GENERATE Id, PostTypeId, AcceptedAnswerId, ParentId,
CreationDate, DeletionDate, Score, ViewCount,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Body,'\\n',''),'\\r',''),'\\r\\n','')
,'<br>',''),',',' ') AS Body, OwnerUserId,
```

```
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(OwnerDisplayName,'\\n',''),'\\r',''),
'\\r\\n',''),'<br>',''),',',' ') AS OwnerDisplayName, LastEditorUserId,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(LastEditorDisplayName,'\\n',''),'\\r'
,''),'\\r\\n',''),'<br>',''),',',' ') AS LastEditorDisplayName, LastEditDate,
LastActivityDate,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Title,'\\n',''),'\\r',''),'\\r\\n',''
),'<br>',''),',',' ') AS Title,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Tags,'\\n',''),'\\r',''),'\\r\\n','')
,'<br>',''),',',' ') AS Tags, AnswerCount, CommentCount, FavoriteCount,
ClosedDate, CommunityOwnedDate, ContendLicense;

E5= FOREACH E GENERATE Id, PostTypeId, AcceptedAnswerId, ParentId,
CreationDate, DeletionDate, Score, ViewCount,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Body,'\\n',''),'\\r',''),'\\r\\n','')
,'<br>',''),',',' ') AS Body, OwnerUserId,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(OwnerDisplayName,'\\n',''),'\\r',''),
'\\r\\n',''),'<br>',''),',',' ') AS OwnerDisplayName, LastEditorUserId,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(LastEditorDisplayName,'\\n',''),'\\r'
,''),'\\r\\n',''),'<br>',''),',',' ') AS LastEditorDisplayName, LastEditDate,
LastActivityDate,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Title,'\\n',''),'\\r',''),'\\r\\n',''
),'<br>',''),',',' ') AS Title,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Tags,'\\n',''),'\\r',''),'\\r\\n','')
,'<br>',''),',',' ') AS Tags, AnswerCount, CommentCount, FavoriteCount,
ClosedDate, CommunityOwnedDate, ContendLicense;
```

Source: https://www.simplilearn.com/tutorials/hadoop-tutorial/pig

https://www.educba.com/pig-commands/

```
-- Union the subsets to make one dataset F

F = UNION A1, B2, C3, D4, E5;

G = FOREACH F GENERATE Id, Title, PostTypeId, Score;

Dump G;

-- Store the clean data from F to Hive directory

Store F INTO '/HiveDataset' USING PigStorage(',');
```

```
2021-10-28 19:31:49,723 [main] INFO  org.apache.hadoop.yarn.clien
t.RMProxy - Connecting to ResourceManager at cluster-e9cf-m/10.15
4.0.3:8032
2021-10-28 19:31:49,724 [main] INFO  org.apache.hadoop.yarn.clien
t.AHSProxy - Connecting to Application History server at cluster-
e9cf-m/10.154.0.3:10200
2021-10-28 19:31:49,726 [main] INFO  org.apache.hadoop.mapred.Cli
entServiceDelegate - Application state is completed. FinalApplica
tionStatus=SUCCEEDED. Redirecting to job history server
2021-10-28 19:31:49,754 [main] INFO  org.apache.hadoop.yarn.clien
t.RMProxy - Connecting to ResourceManager at cluster-e9cf-m/10.15
4.0.3:8032
2021-10-28 19:31:49,755 [main] INFO  org.apache.hadoop.yarn.clien
t.AHSProxy - Connecting to Application History server at cluster-
e9cf-m/10.154.0.3:10200
2021-10-28 19:31:49,758 [main] INFO  org.apache.hadoop.mapred.Cli
entServiceDelegate - Application state is completed. FinalApplica
tionStatus=SUCCEEDED. Redirecting to job history server
2021-10-28 19:31:49,773 [main] INFO  org.apache.hadoop.yarn.clien
t.RMProxy - Connecting to ResourceManager at cluster-e9cf-m/10.15
4.0.3:8032
2021-10-28 19:31:49,774 [main] INFO  org.apache.hadoop.yarn.clien
t.AHSProxy - Connecting to Application History server at cluster-
e9cf-m/10.154.0.3:10200
2021-10-28 19:31:49,777 [main] INFO  org.apache.hadoop.mapred.Cli
entServiceDelegate - Application state is completed. FinalApplica
tionStatus=SUCCEEDED. Redirecting to job history server
2021-10-28 19:31:49,795 [main] WARN  org.apache.pig.backend.hadoo
p.executionengine.mapReduceLayer.MapReduceLauncher - Encountered
Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 55 time(s).
2021-10-28 19:31:49,795 [main] INFO  org.apache.pig.backend.hadoo
p.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

In Local:

```
-- Check if file was stored
hadoop fs -ls /HiveDataset
```

```
una_mcginn2@cluster-e9cf-m:~$ hadoop fs -ls /HiveDataset
Found 6 items
-rw-r--r--   2 una_mcginn2 hadoop          0 2021-10-28 19:31 /Hi
veDataset/_SUCCESS
-rw-r--r--   2 una_mcginn2 hadoop   62381550 2021-10-28 19:31 /Hi
veDataset/part-m-00000
-rw-r--r--   2 una_mcginn2 hadoop   57285898 2021-10-28 19:31 /Hi
veDataset/part-m-00001
-rw-r--r--   2 una_mcginn2 hadoop   55231697 2021-10-28 19:31 /Hi
veDataset/part-m-00002
-rw-r--r--   2 una_mcginn2 hadoop   50137769 2021-10-28 19:31 /Hi
veDataset/part-m-00003
-rw-r--r--   2 una_mcginn2 hadoop   10508431 2021-10-28 19:31 /Hi
veDataset/part-m-00004
```

Source: https://spark.apache.org/docs/3.1.1/sql-ref-syntax-ddl-create-table-hiveformat.html

Go to Hive:

Code:

```
CREATE TABLE A4 (Id int, PostTypeId int, AcceptedAnswerId int, ParentId int,
CreationDate string, DeletionDate string, Score int, ViewCount int, Body
string, OwnerUserId int, OwnerDisplayName string, LastEditorUserId int,
LastEditorDisplayName string, LastEditDate string, LastActivityDate string,
```

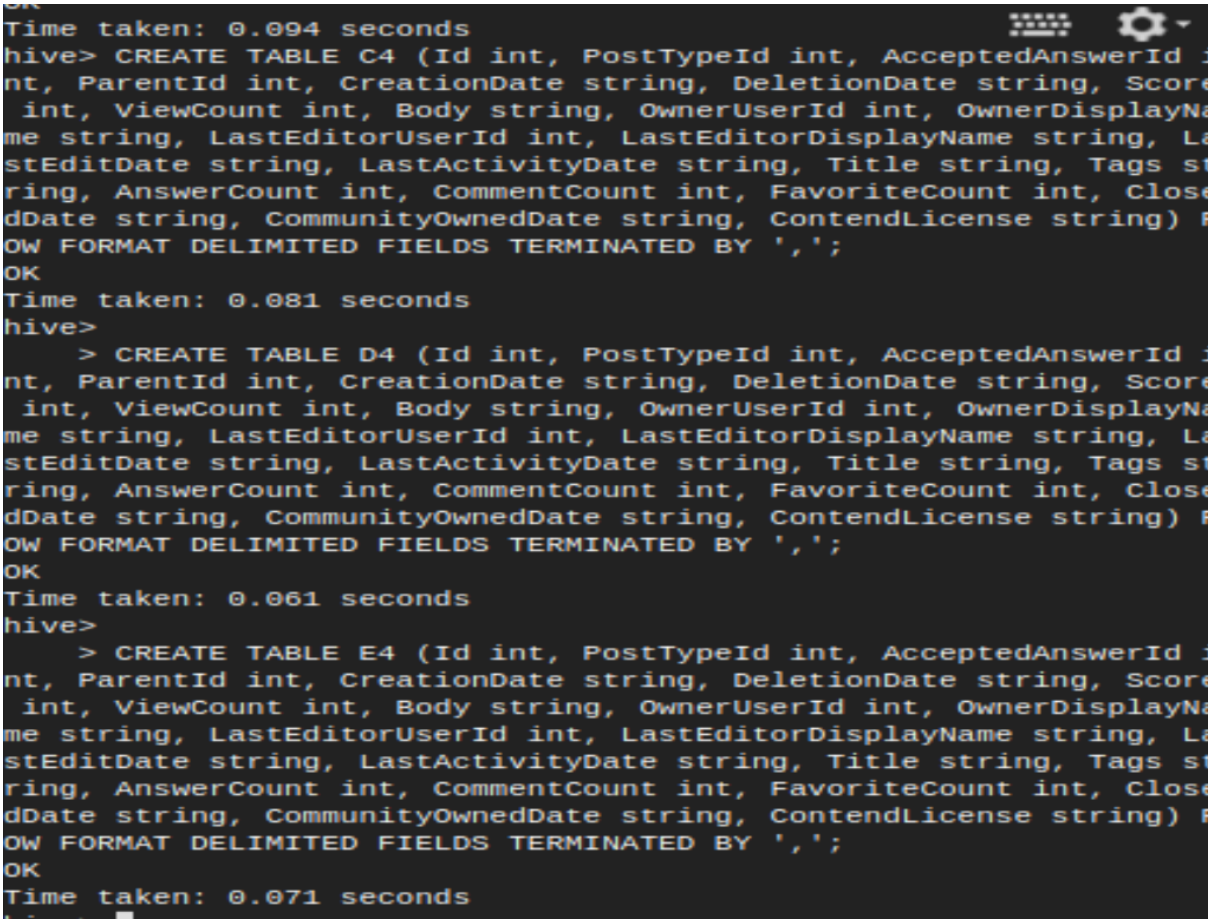
```
Title string, Tags string, AnswerCount int, CommentCount int, FavoriteCount
int, ClosedDate string, CommunityOwnedDate string, ContendLicense string) ROW
FORMAT DELIMITED FIELDS TERMINATED BY ','
TBLPROPERTIES('skip.header.line.count'='1');
```

```
CREATE TABLE B4 (Id int, PostTypeId int, AcceptedAnswerId int, ParentId int,
CreationDate string, DeletionDate string, Score int, ViewCount int, Body
string, OwnerUserId int, OwnerDisplayName string, LastEditorUserId int,
LastEditorDisplayName string, LastEditDate string, LastActivityDate string,
Title string, Tags string, AnswerCount int, CommentCount int, FavoriteCount
int, ClosedDate string, CommunityOwnedDate string, ContendLicense string) ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

```
CREATE TABLE C4 (Id int, PostTypeId int, AcceptedAnswerId int, ParentId int,
CreationDate string, DeletionDate string, Score int, ViewCount int, Body
string, OwnerUserId int, OwnerDisplayName string, LastEditorUserId int,
LastEditorDisplayName string, LastEditDate string, LastActivityDate string,
Title string, Tags string, AnswerCount int, CommentCount int, FavoriteCount
int, ClosedDate string, CommunityOwnedDate string, ContendLicense string) ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

```
CREATE TABLE D4 (Id int, PostTypeId int, AcceptedAnswerId int, ParentId int,
CreationDate string, DeletionDate string, Score int, ViewCount int, Body
string, OwnerUserId int, OwnerDisplayName string, LastEditorUserId int,
LastEditorDisplayName string, LastEditDate string, LastActivityDate string,
Title string, Tags string, AnswerCount int, CommentCount int, FavoriteCount
int, ClosedDate string, CommunityOwnedDate string, ContendLicense string) ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

```
CREATE TABLE E4 (Id int, PostTypeId int, AcceptedAnswerId int, ParentId int,
CreationDate string, DeletionDate string, Score int, ViewCount int, Body
string, OwnerUserId int, OwnerDisplayName string, LastEditorUserId int,
LastEditorDisplayName string, LastEditDate string, LastActivityDate string,
Title string, Tags string, AnswerCount int, CommentCount int, FavoriteCount
int, ClosedDate string, CommunityOwnedDate string, ContendLicense string) ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',';
```




```
LOAD DATA INPATH '/HiveDataset/part-m-00000' INTO TABLE A4;

LOAD DATA INPATH '/HiveDataset/part-m-00001' INTO TABLE B4;

LOAD DATA INPATH '/HiveDataset/part-m-00002' INTO TABLE C4;

LOAD DATA INPATH '/HiveDataset/part-m-00003' INTO TABLE D4;

LOAD DATA INPATH '/HiveDataset/part-m-00004' INTO TABLE E4;
```

```
Loading data to table default.a4
OK
Time taken: 0.342 seconds
hive> LOAD DATA INPATH '/HiveDataset/part-m-00001' INTO TABLE B4;

Loading data to table default.b4
OK
Time taken: 0.235 seconds
hive> LOAD DATA INPATH '/HiveDataset/part-m-00002' INTO TABLE C4;

Loading data to table default.c4
OK
Time taken: 0.199 seconds
hive> LOAD DATA INPATH '/HiveDataset/part-m-00003' INTO TABLE D4;

Loading data to table default.d4
OK
Time taken: 0.203 seconds
hive> LOAD DATA INPATH '/HiveDataset/part-m-00004' INTO TABLE E4;

Loading data to table default.e4
OK
Time taken: 0.192 seconds
hive>
```

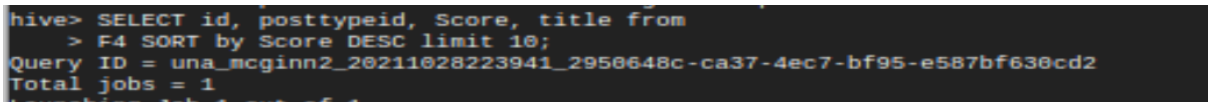
```
-- Create table F4 as resultant table of Union of A4, B4, C4, D4 and E4

create table F4 as
select * From A4 union all
select * From B4 union all
select * From C4 union all
select * From D4 union all
select * From E4;

-- Remove records where Body is not equal to 'Body'

insert overwrite table F4

select * from F4

where Body <> 'Body';


SELECT COUNT(*) FROM F4;
```

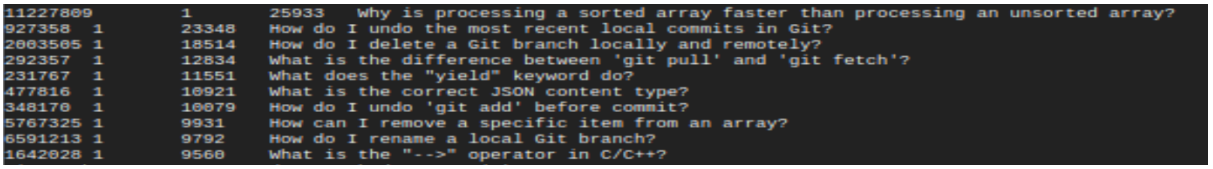
```
Loading data to table default.f4
OK
Time taken: 19.383 seconds
hive> SELECT COUNT(*) FROM F4;
OK
199807
Time taken: 0.138 seconds, Fetched: 1 row(s)
hive>
```

Task 3:

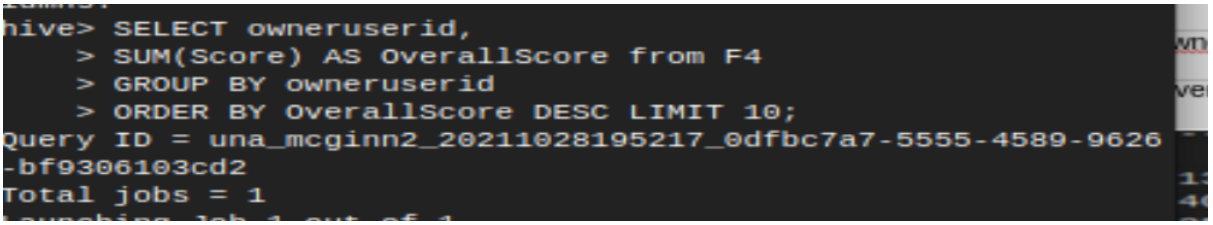
```
-- The top 10 posts by score
```

```
SELECT id, posttypeid, Score, title from
F4 SORT by Score DESC limit 10;
```






```
--  The top 10 users by post score

SELECT owneruserid,

SUM(Score) AS OverallScore from F4

GROUP BY owneruserid

ORDER BY OverallScore DESC LIMIT 10;
```




```
-- The number of distinct users, who used the word “cloud” in one
of their posts

SELECT DISTINCT(OwnerUserId) From F4
```

```
WHERE (Body REGEXP 'cloud') OR (Title REGEXP 'cloud');
```

```
hive> SELECT DISTINCT(OwnerUserId) From F4
    > WHERE (Body REGEXP 'cloud') OR (Title REGEXP 'cloud');
```