

# Supporting Documentation: Task 2 and 3

## Student: Úna McGinn

### CA675 Assignment 1

#### Task 2:

##### Load Data to Pig:

```
Connected, host fingerprint: ssh-rsa 0 95:78:B8:2A:65:9F:43:85: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_mcgin2@cluster-e9cf-m:~$ whoami
una_mcgin2
una_mcgin2@cluster-e9cf-m:~$ hadoop fs -mkdir -p /user/una_mcgin2
una_mcgin2@cluster-e9cf-m:~$ export PATH=${JAVA_HOME}/bin:${PATH}
una_mcgin2@cluster-e9cf-m:~$ export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
una_mcgin2@cluster-e9cf-m:~$ hadoop fs -ls
una_mcgin2@cluster-e9cf-m:~$ env
```

-- Setup new directory

```
una_mcgin2@cluster-e9cf-m:~$ hadoop fs -mkdir -p /user/una_mcgin2
```

```
una_mcgin2@cluster-e9cf-m:~$ hadoop fs -mkdir /PigData
```

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

hadoop fs -put /home/una\_mcgin2/QueryResults1.csv /PigData

hadoop fs -put /home/una\_mcgin2/QueryResults2.csv /PigData

hadoop fs -put /home/una\_mcgin2/QueryResults3.csv /PigData

hadoop fs -put /home/una\_mcgin2/QueryResults4.csv /PigData

hadoop fs -put /home/una\_mcgin2/QueryResults5.csv /PigData

```
una_mcgin2@cluster-e9cf-m:~$ hadoop fs -put /home/una_mcgin2/QueryResults1.csv /PigData
```

```
una_mcgin2@cluster-e9cf-m:~$ hadoop fs -put /home/una_mcgin2/QueryResults2.csv /PigData
```

```
una_mcgin2@cluster-e9cf-m:~$ hadoop fs -put /home/una_mcgin2/QueryResults3.csv /PigData
```

```
una_mcgin2@cluster-e9cf-m:~$ hadoop fs -put /home/una_mcgin2/QueryResults4.csv /PigData
```

```
una_mcgin2@cluster-e9cf-m:~$ hadoop fs -put /home/una_mcgin2/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,ContentLicense: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:chararray,LastEditDate:chararray,LastActivityDate:chararray,Title:chararray,Tags:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:chararray,CommunityOwnedDate:chararray,ContentLicense: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:chararray,DeletionDate:chararray,Score:int,ViewCount:int,Body:chararray,OwnerUserId:int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:chararray,LastEditDate:chararray,LastActivityDate:chararray,Title:chararray,Tags:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:chararray,CommunityOwnedDate:chararray,ContentLicense: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:chararray,DeletionDate:chararray,Score:int,ViewCount:int,Body:chararray,OwnerUserId:int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:chararray,LastEditDate:chararray,LastActivityDate:chararray,Title:chararray,Tags:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:chararray,CommunityOwnedDate:chararray,ContentLicense: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:chararray,DeletionDate:chararray,Score:int,ViewCount:int,Body:chararray,OwnerUserId:int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:chararray,LastEditDate:chararray,LastActivityDate:chararray,Title:chararray,Tags:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:chararray,CommunityOwnedDate:chararray,ContentLicense:chararray);
```

```
grunt> E = LOAD '/PigData/QueryResults5.csv'
' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE') AS (Id:
int,PostTypeId:int,AcceptedAnswerId:int,ParentId:int,CreationDate:chararray,DeletionDate:chararray,Score:int,ViewCount:int,Body:chararray,OwnerUserId:int,OwnerDisplayName:chararray,LastEditorUserId:int,LastEditorDisplayName:chararray,LastEditDate:chararray,LastActivityDate:chararray,Title:chararray,Tags:chararray,AnswerCount:int,CommentCount:int,FavoriteCount:int,ClosedDate:chararray,CommunityOwnedDate:chararray,ContentLicense:chararray);
2021-10-28 19:26:15,421 [main] INFO org.apache.hadoop.conf.Configuration.deprecation
- yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.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(',');

```

```

Files
2021-10-28 19:31:49,723 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at cluster-e9cf-m/10.154.0.3:8032
2021-10-28 19:31:49,724 [main] INFO org.apache.hadoop.yarn.client.AHSPProxy - 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.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-10-28 19:31:49,754 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at cluster-e9cf-m/10.154.0.3:8032
2021-10-28 19:31:49,755 [main] INFO org.apache.hadoop.yarn.client.AHSPProxy - 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.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-10-28 19:31:49,773 [main] INFO org.apache.hadoop.yarn.client.RMPProxy - Connecting to ResourceManager at cluster-e9cf-m/10.154.0.3:8032
2021-10-28 19:31:49,774 [main] INFO org.apache.hadoop.yarn.client.AHSPProxy - 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.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-10-28 19:31:49,795 [main] WARN org.apache.pig.backend.hadoop.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.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>

```

### In Local:

-- Check if file was stored

hadoop fs -ls /HiveDataset

```

una_mcgin2@cluster-e9cf-m:~$ hadoop fs -ls /HiveDataset
Found 6 items
-rw-r--r--  2 una_mcgin2 hadoop      0 2021-10-28 19:31 /HiveDataset/_SUCCESS
-rw-r--r--  2 una_mcgin2 hadoop 62381550 2021-10-28 19:31 /HiveDataset/part-m-00000
-rw-r--r--  2 una_mcgin2 hadoop 57285898 2021-10-28 19:31 /HiveDataset/part-m-00001
-rw-r--r--  2 una_mcgin2 hadoop 55231697 2021-10-28 19:31 /HiveDataset/part-m-00002
-rw-r--r--  2 una_mcgin2 hadoop 50137769 2021-10-28 19:31 /HiveDataset/part-m-00003
-rw-r--r--  2 una_mcgin2 hadoop 10508431 2021-10-28 19:31 /HiveDataset/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 ',';
```



```

OK
Time taken: 0.094 seconds
hive> 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, ContentLicense string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.081 seconds
hive>
> 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, ContentLicense string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.061 seconds
hive>
> 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, ContentLicense string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.071 seconds

```

```

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;
```

```
hive> SELECT id, posttypeid, Score, title from
> F4 SORT by Score DESC limit 10;
Query ID = una_mcginn2_20211028223941_2950648c-ca37-4ec7-bf95-e587bf630cd2
Total jobs = 1
```

```
11227809      1      25933  Why is processing a sorted array faster than processing an unsorted array?
927358      1      23348  How do I undo the most recent local commits in Git?
2003505      1      18514  How do I delete a Git branch locally and remotely?
292357      1      12834  What is the difference between 'git pull' and 'git fetch'?
231767      1      11551  What does the "yield" keyword do?
477816      1      10921  What is the correct JSON content type?
348170      1      10079  How do I undo 'git add' before commit?
5767325      1      9931   How can I remove a specific item from an array?
6591213      1      9792   How do I rename a local Git branch?
1642020      1      9560   What is the "-->" operator in C/C++?
```

-- The top 10 users by post score

```
SELECT owneruserid,
SUM(Score) AS OverallScore from F4
GROUP BY owneruserid
ORDER BY OverallScore DESC LIMIT 10;
```

```
hive> SELECT owneruserid,
> SUM(Score) AS OverallScore from F4
> GROUP BY owneruserid
> ORDER BY OverallScore DESC LIMIT 10;
Query ID = una_mcginn2_20211028195217_0dfbc7a7-5555-4589-9626-bf9306103cd2
Total jobs = 1
```

```
OK
NULL      450416
87234     37672
4883      28817
9951      26799
6068      25944
89904     24024
51816     23719
49153     20203
179736    19530
95592     19479
Time taken: 19.246 seconds, Fetched: 10 row(s)
```

-- 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');
```

```
7653266  
7769163  
7891178  
7905439  
8021951  
8163926  
8170216  
8273474  
8283445  
8351940  
8356264  
8374363  
8383300  
8415957  
8559109  
8564359  
8757603  
8890010  
9056769  
9232623  
9241896  
9306306  
9403545  
9447110  
9623730  
9624577  
9654310  
9707310  
9753241  
9937933  
9964320  
10041534  
10607591  
10684059  
10781988  
11305443  
11520111  
11829057  
12586904  
12787107  
13372486  
Time taken: 22.257 seconds, Fetched: 630 row(s)  
)
```