Task 3: Query data using MapReduce/Pig/Hive

1. In order to load the cleaned data from PIG to HIVE, we create a database and a table in HIVE.

   hive> CREATE DATABASE user_db;
   hive> USE user_db;
   hive> CREATE TABLE user_db.stackdata_analysis (id int, score int, viewcount int, owneruserid int, ownerusername string, title string, tags string, body string);

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> CREATE DATABASE user_db;
OK
Time taken: 0.654 seconds
hive> USE user_db;
OK
Time taken: 0.095 seconds
hive> CREATE TABLE user_db.stackdata_analysis (id int, score int, viewcount int, owneruserid int, title string, tags string,body string);
OK
Time taken: 0.652 seconds
hive>
```

2. Now, we can transfer the cleaned data from PIG into the above created table in HIVE.

   hive> STORE Pig_QueryResults INTO 'user_db.stackdata_analysis' USING org.apache.hive.hcatalog.pig.HCatStorer();

3. In order to verify that the data has been transferred properly, we check the count of the ‘id’ field which returns the value 200,000!

   hive> SELECT COUNT(id) FROM user_db.stackdata_analysis;

4. **Queries for:**
   A. The top 10 posts by score

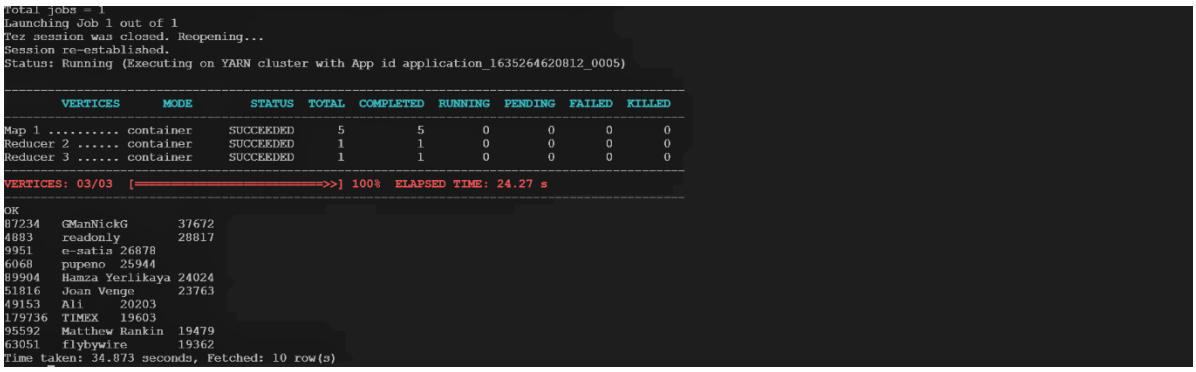
      hive> SELECT id, title, score FROM user_db.stackdata_analysis ORDER BY score DESC LIMIT 10;

```
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1635264620812_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 21.67 s
--------------------------------------------------------------------------------
OK
11227809        Why is processing a sorted array faster than processing an unsorted array?      25933
927358  How do I undo the most recent local commits in Git?     23348
2003505 How do I delete a Git branch locally and remotely?      18514
292357  What is the difference between 'git pull' and 'git fetch'?      12834
231767  What does the "yield" keyword do?       11551
477816  What is the correct JSON content type?  10921
348170  How do I undo 'git add' before commit?  10079
5767325 How can I remove a specific item from an array? 9931
6591213 How do I rename a local Git branch?     9792
1642028 What is the "-->" operator in C/C++?    9560
Time taken: 31.438 seconds, Fetched: 10 row(s)
```

B. The top 10 users by post score

hive> SELECT owneruserid AS USERID, ownerusername, SUM(score) AS SCORE FROM user_db.stackdata_analysis GROUP BY owneruserid having owneruserid is not null SORT BY score DESC LIMIT 10;



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

hive> SELECT COUNT(DISTINCT owneruserid) FROM user_db.stackdata_analysis WHERE UPPER(body) LIKE '%CLOUD%' OR UPPER(title) LIKE '%CLOUD%' OR LOWER(body) LIKE '%cloud%' OR UPPER(tags) LIKE '%CLOUD%' OR LOWER(title) LIKE '%cloud%' OR LOWER(tags) LIKE '%CLOUD%';