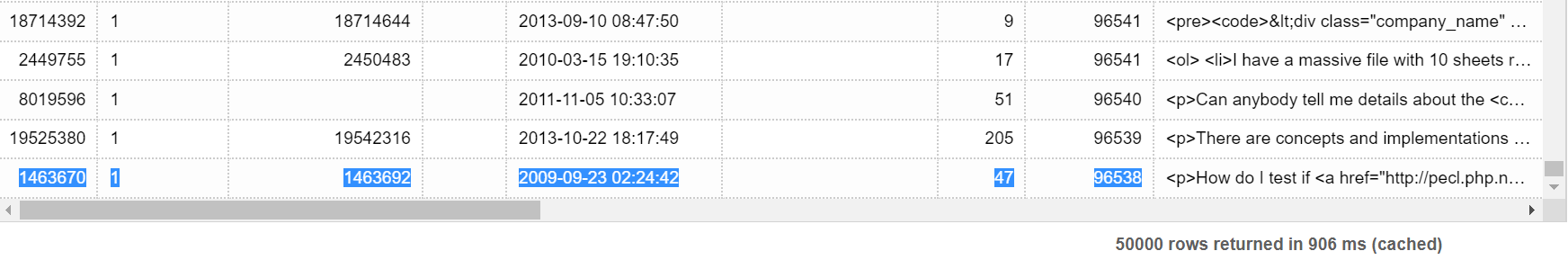
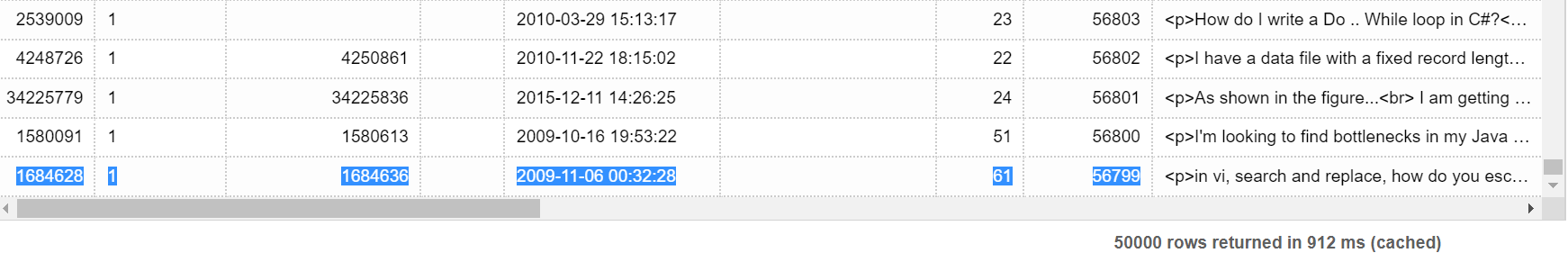
**1.Acquire the top 200,000 posts by viewcount (see notes on Data Acquisition)**

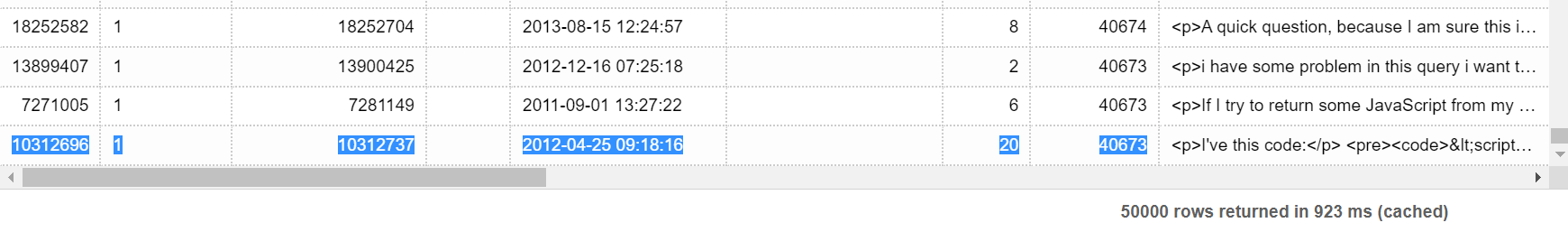
SELECT top 50000 \* FROM posts WHERE posts.ViewCount < 100000000 ORDER BY posts.ViewCount DESC;



SELECT top 50000 \* FROM posts WHERE posts.ViewCount <= 96538 and posts.Id <> 1463670 ORDER BY posts.ViewCount DESC;

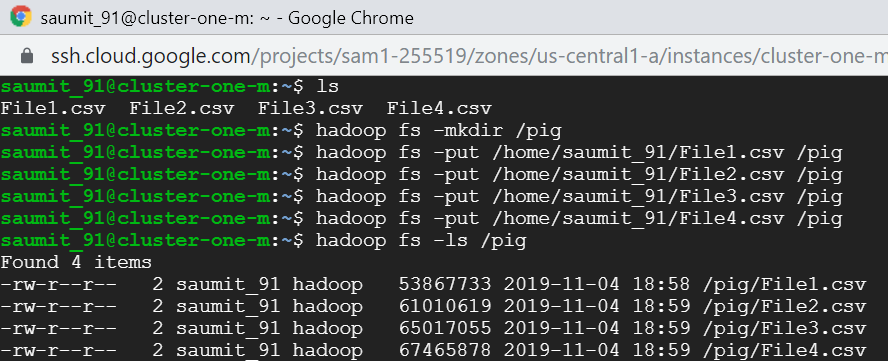


SELECT top 50000 \* FROM posts WHERE posts.ViewCount <= 56799 and posts.Id <> 1684628 ORDER BY posts.ViewCount DESC;

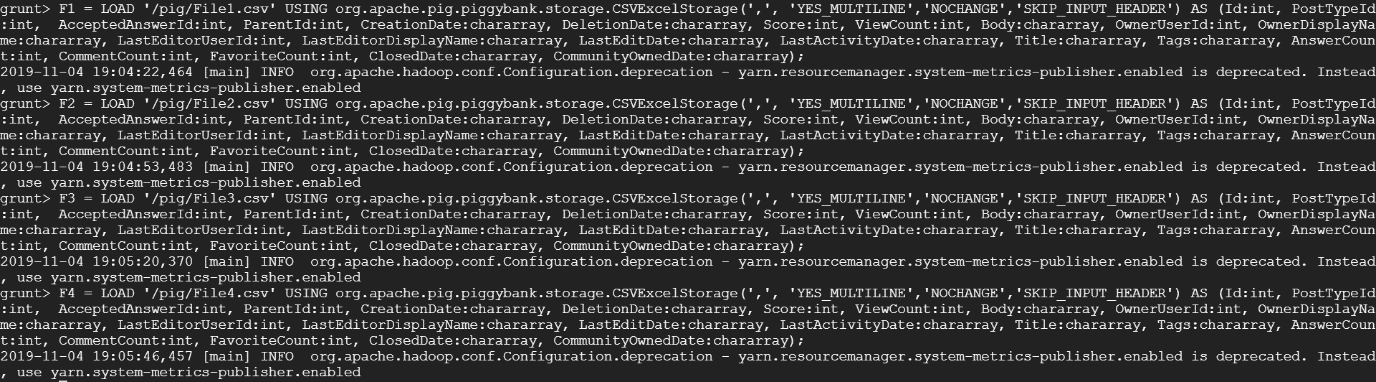


SELECT top 50000 \* FROM posts WHERE posts.ViewCount <= 40673 and posts.Id <> 13899407 and posts.Id <> 7271005 and posts.Id <> 10312696 ORDER BY posts.ViewCount DESC;

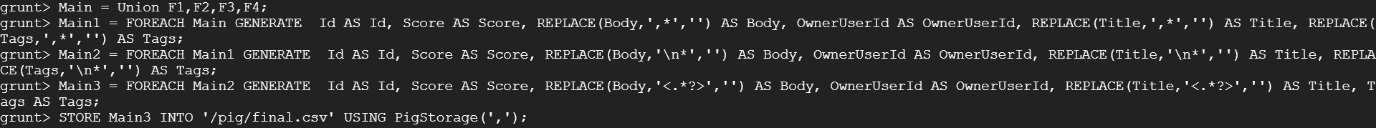
**2.Using Pig or MapReduce, extract, transform and load the data as applicable**



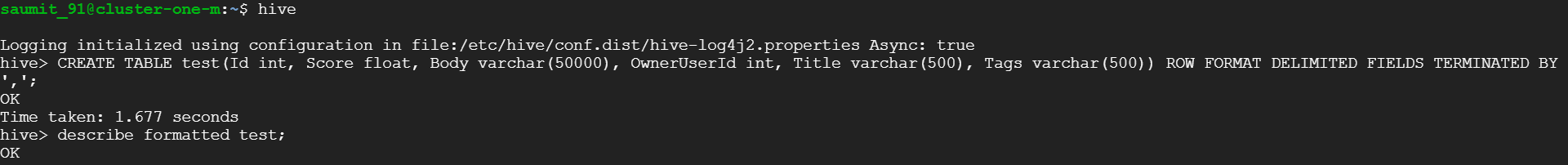
*Extracting*

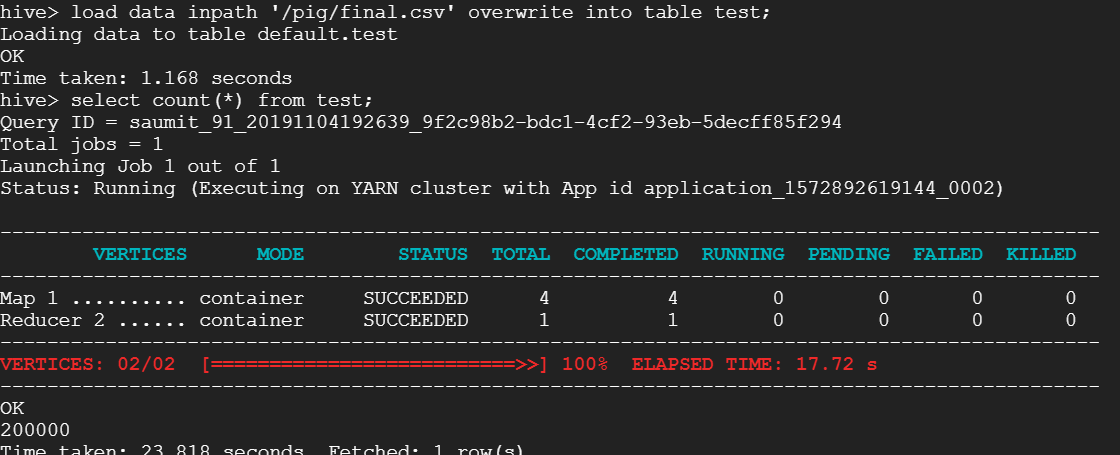


*Transforming (cleaning)*



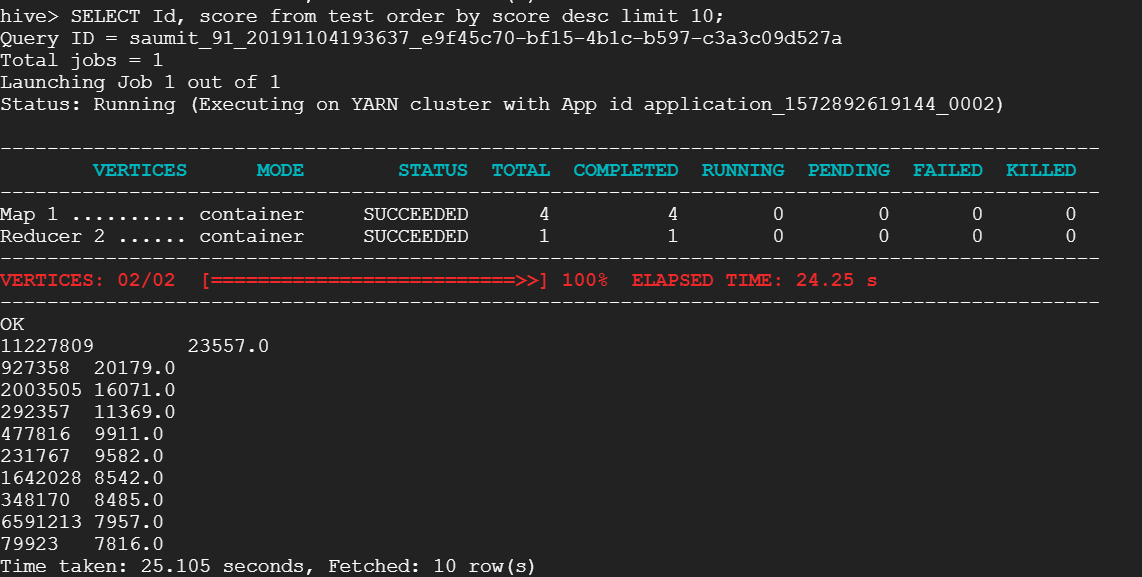
*Loading*



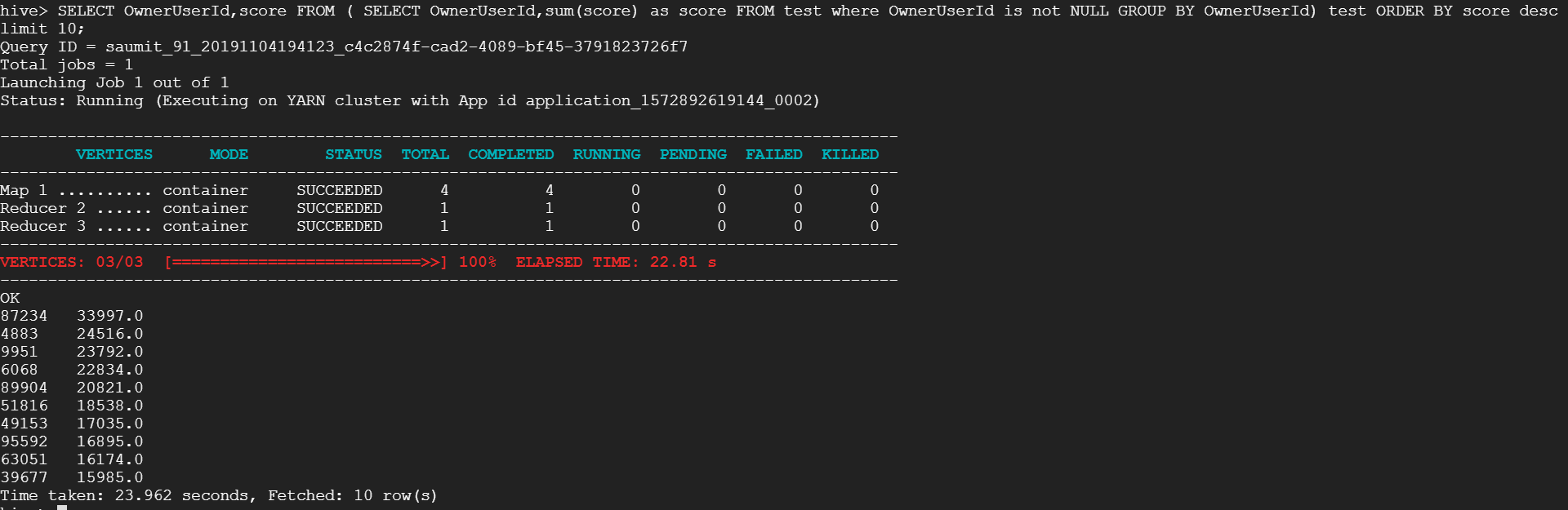


**3.Using Hive and/or MapReduce, get:**

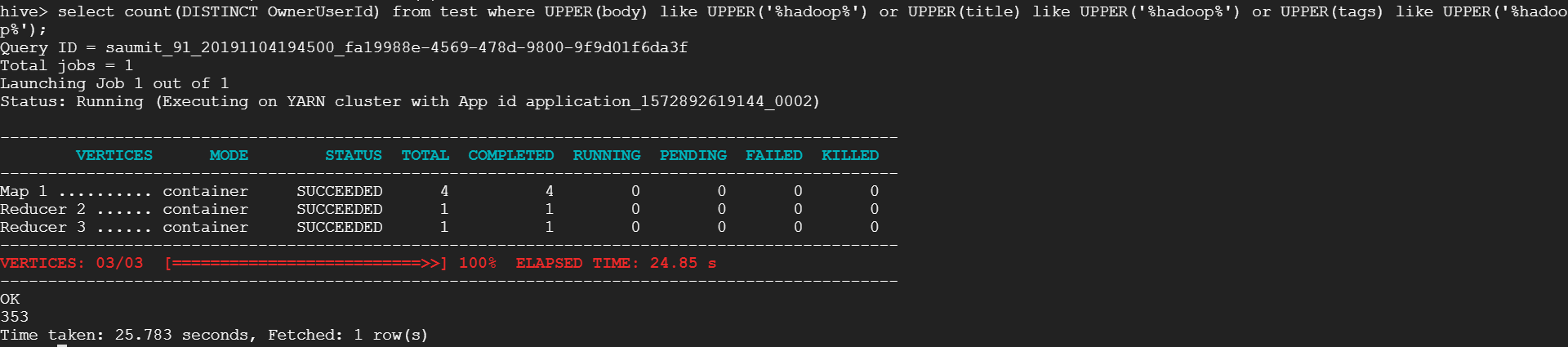
**I. The top 10 posts by score**



**II. The top 10 users by post score**



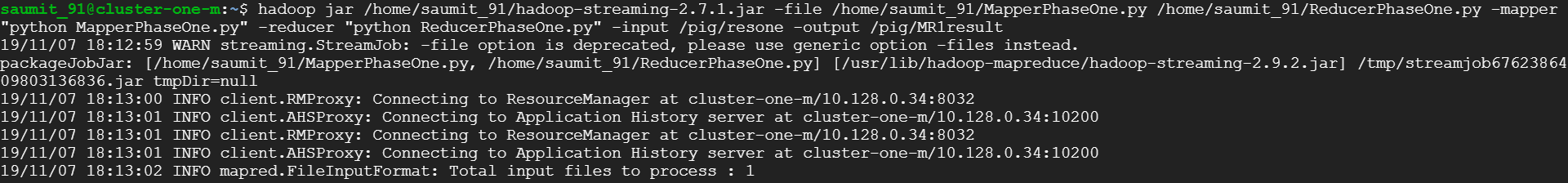
**III. The number of distinct users, who used the word “Hadoop” in one of their posts**

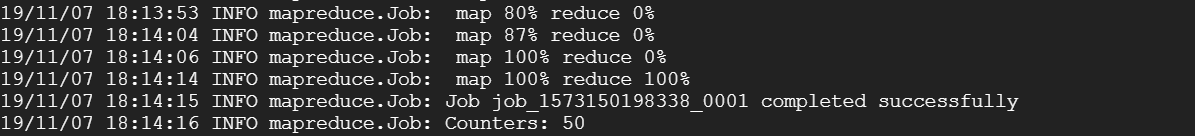


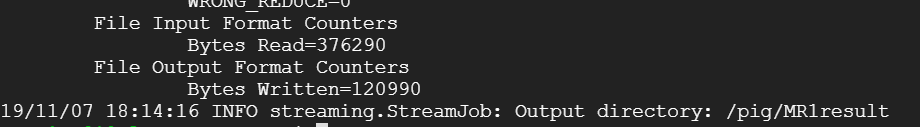
**4.Using Mapreducecalculate the per-user TF-IDF (just submit the top 10 terms for each of the top 10 users from Query 3.II)**

Stored the required data from table into pig directory (code provided in source code file).

Ran the below script three time for each mapreduce phase (one script run screenshot is provided)







*Results of 3 users out of 10 provided below*

