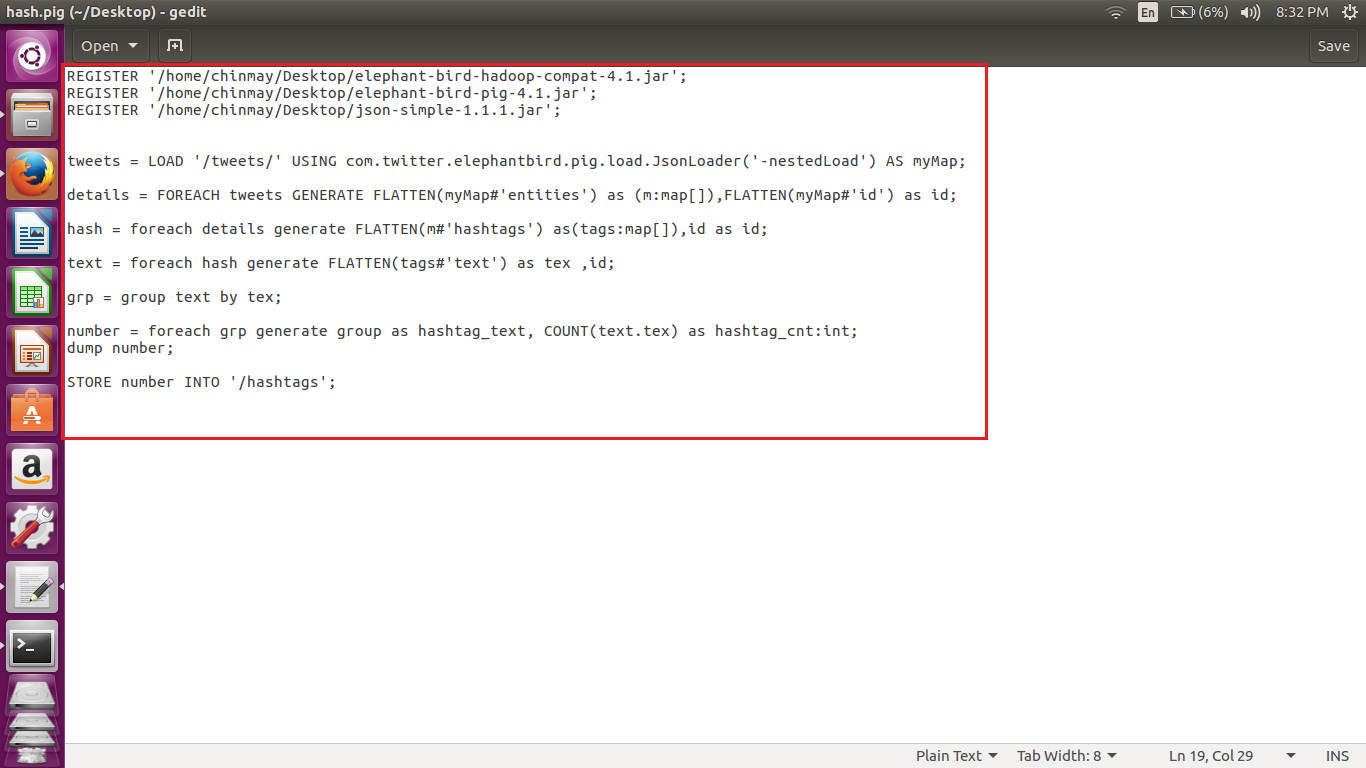
Assignment – 11.4

Write a pig script to find the count of each Hashtag in the dataset. You can download the twitter

data from the below link

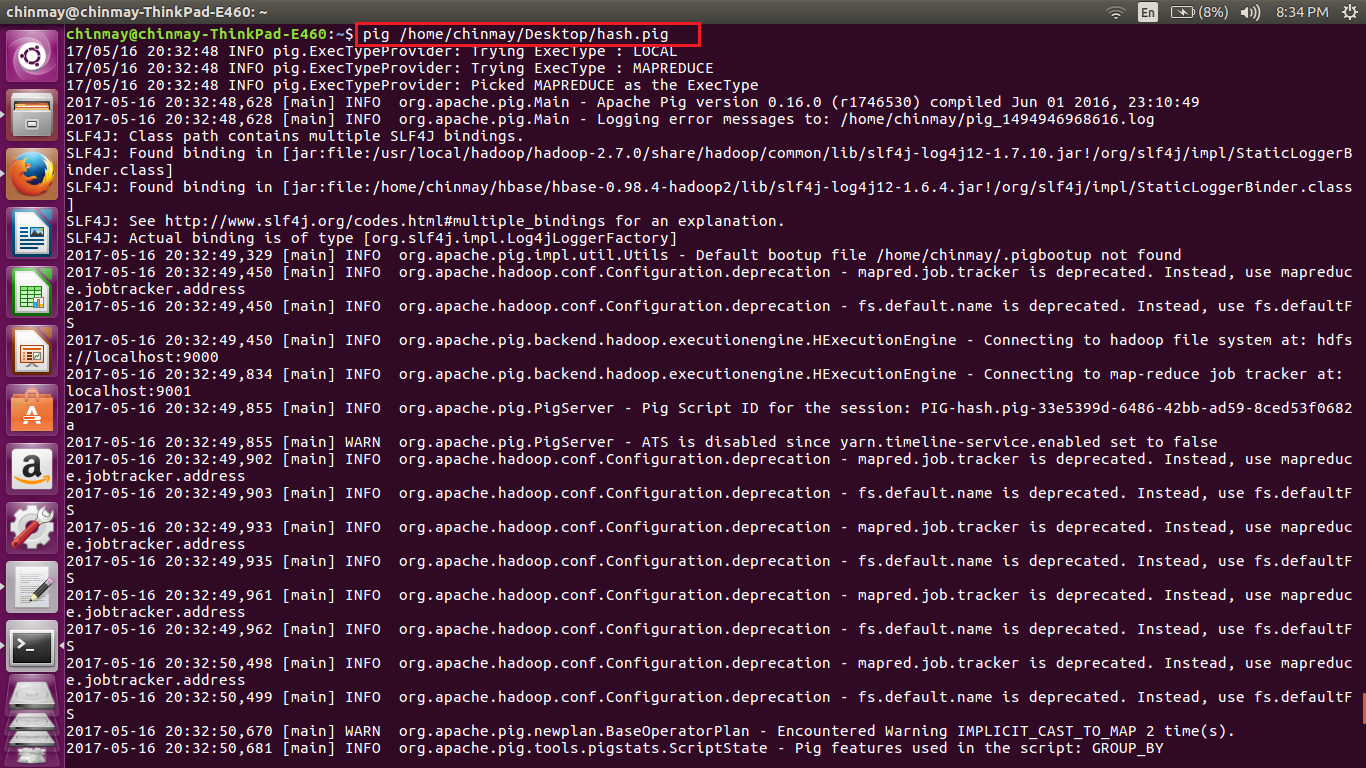


The tweets are in nested Json format and consists of map data types. So they are loaded using the Json loader which supports the map.

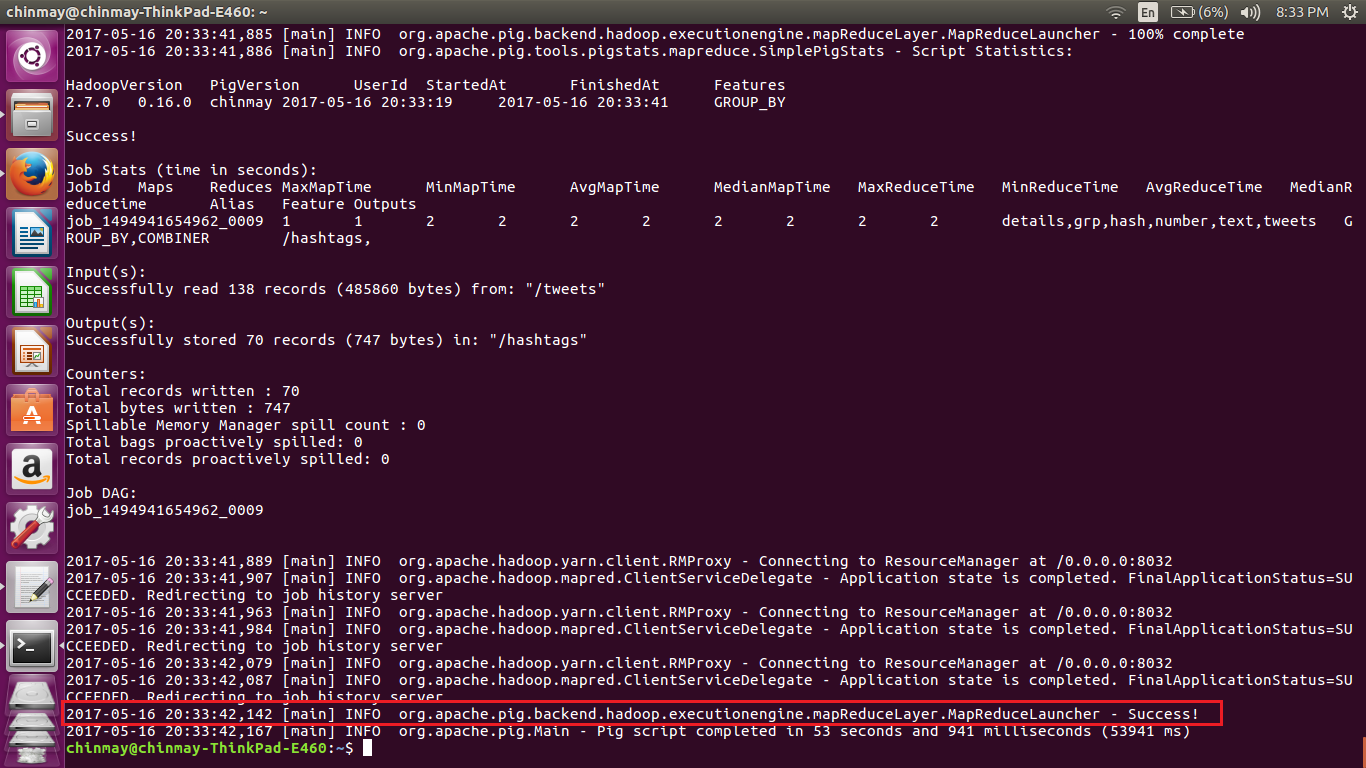
Hence we are using the elephant Json loader.

Steps used

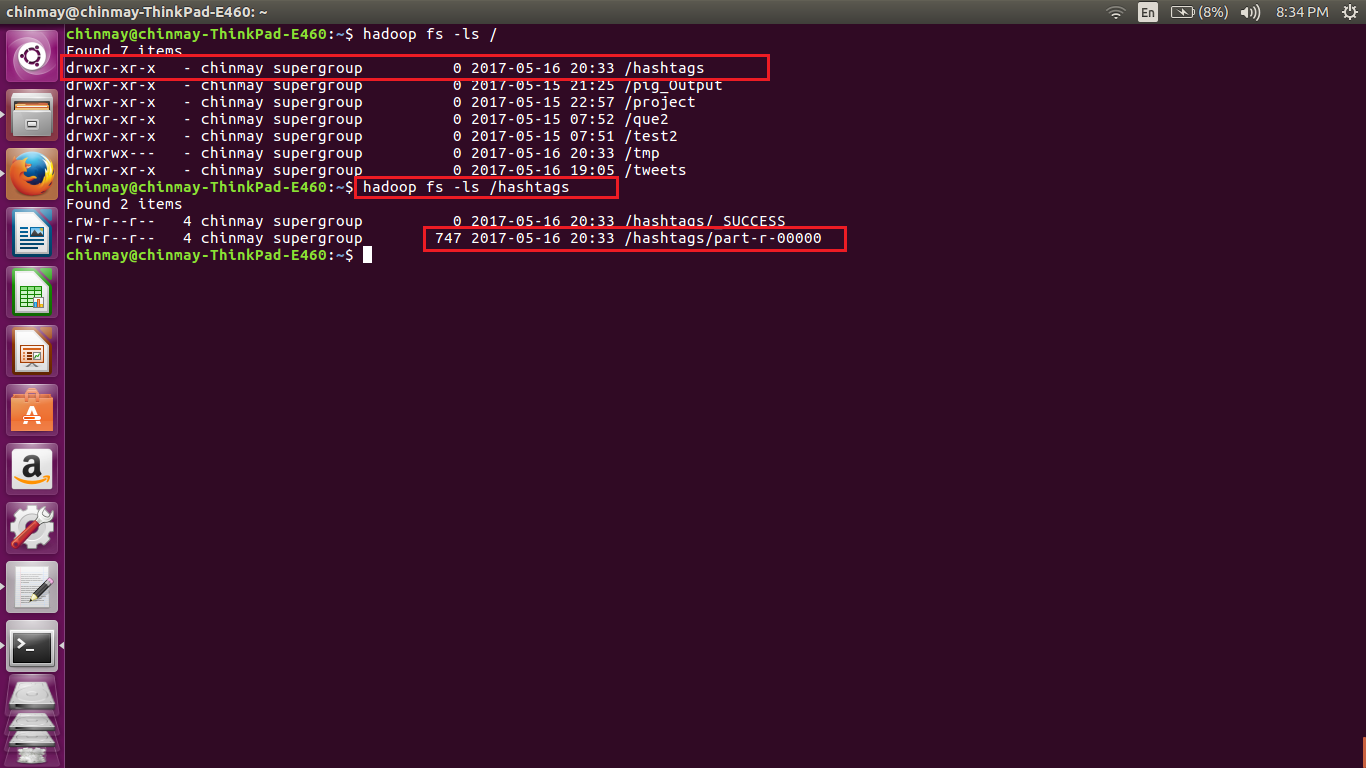
1. Load data using the elephant Json loader
2. Extract entities and the id for each tweet.
3. Now extract the hashtags from the entities which was found in step before.
4. The extracted hashtagsis also a map data type, which can see by describing the relation, from the extracted hashtags, we need to extract textwhich contains the actual hashtag. Hence we have to extract the text which conatins the actual hashtag.
5. Now grouping using the hashtags.
6. And counting the number of hash tags
7. Storing in the hdfs.



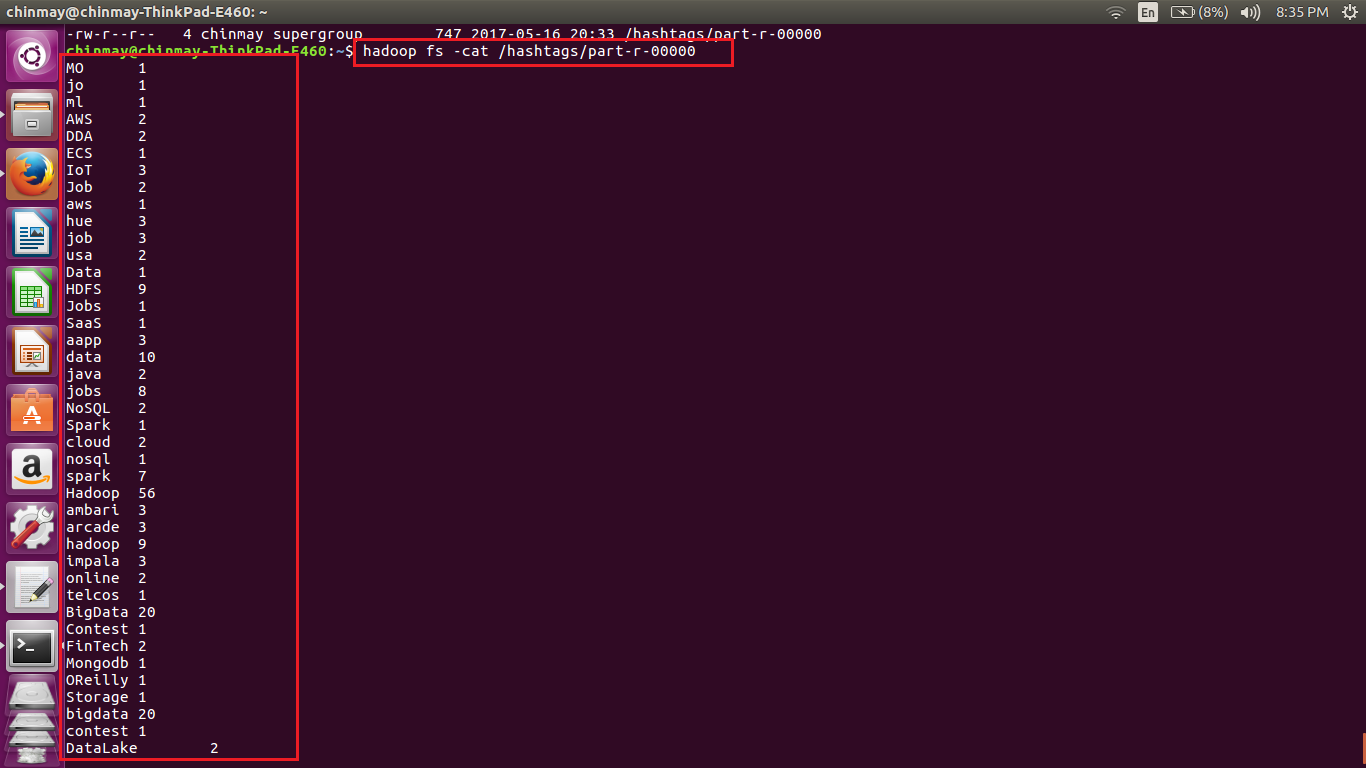
Execution of the Pig script. Here I am using the pig in mapreduce mode.



Storing of the output file is a success.



Here we can see that in the hdfs a file named hastags is formed were the result is stored.





We can see that all the hash tags are listed above.