Introduction: The analysis of the stack exchange dataset is done in the assignment using big data technology.

Acquiring the 200000 posts using ViewCount-

* Select top 50000 \* from posts where posts.ViewCount>19350 ORDER BY posts.ViewCount
* Select top 50000 \* from posts where posts.ViewCount>24630 ORDER BY posts.ViewCount
* Select top 50000 \* from posts where posts.ViewCount>34081 ORDER BY posts.ViewCount
* Select top 50000 \* from posts where posts.ViewCount>57481 ORDER BY posts.ViewCount

The files are downloaded as CSV files and merged using pig which contains 200,000 posts. The report contains the input and output of the tasks.

Task 1:Extract Transform Load:

The data source is hadoop distributed file system(HDFS) and data format is CSV. This section shows the cleaning and formatting of the dataset. The dataset is first loaded into HDFS.

Screen Shot 2017-03-28 at 11.43.02.png

Reading the dataset using hadoop and ETL using mapreduce

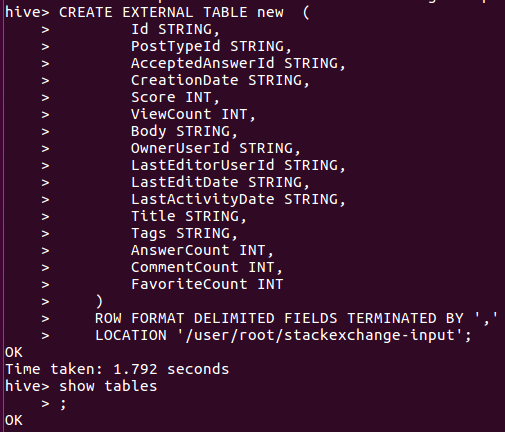
The difficulty in reading the dataset in hadoop is to identify the multiline row. The column body contains multiple line which needs to be cleaned. The dataset is cleaned using pig. All the unwanted words and characters are cleaned and removed.The figure 1 and 2 show the data cleaning for the same reading.

Screen Shot 2017-03-28 at 01.29.09.png

Screen Shot 2017-03-28 at 01.36.30.png

Loading CSV into hive server

The new table is created and the dataset is loaded using the Hive query.

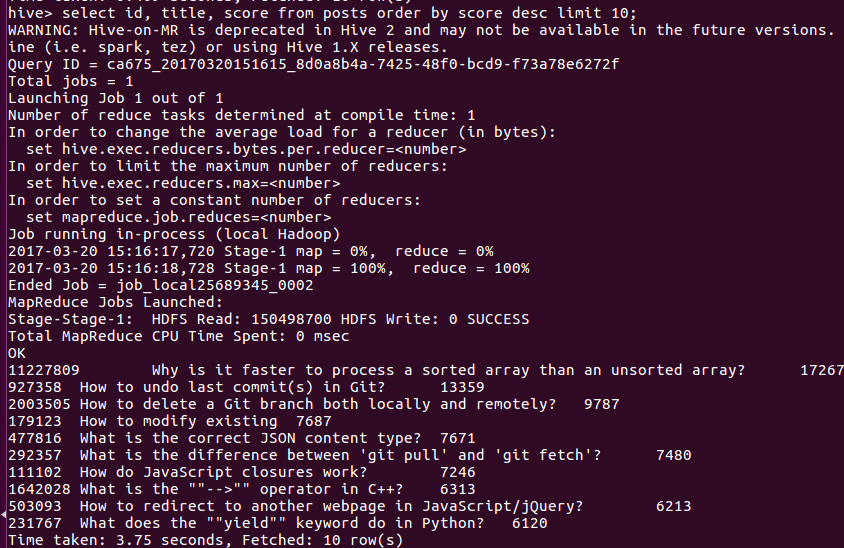


**Query 1**: Top 10 posts by Score

**Input:**-

Screen Shot 2017-03-27 at 23.21.27.png

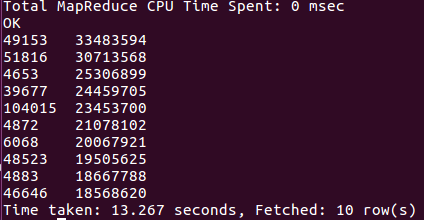
**output:-**



**Query 2**: Top 10 users by post Score

**Input:-**Screen Shot 2017-03-27 at 23.33.04.png

**Output:-**

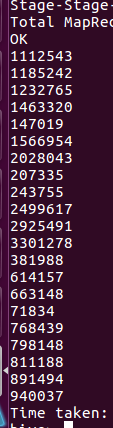


**Query 3**: The number of distinct users, who used the word ‘hadoop’ in one of their posts:

**Input:-**

Screen Shot 2017-03-27 at 23.37.44.png

**Output:-**



**TF/IDF:**

The TF/IDF is done using hive. Please check the screenshot folder for the programming**.**

**Conclusion:** Hive is used to do the queries of the assignment.

* Pig is used to clean the dataset and all the datasets were merged.
* As i had no knowledge of java so unfortunately i could not do the TF/IDF using mapreduce. I tried to do it using hive and i failed in that.
* Some values in the OwnerUserId were missing which were converted to zero.
* Used pig commands for neglecting newlines in the body column of posts.
* Learnt to work with pig and hive.