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.

```
ca675@ca675-VB:/usr/local/hadoop/hive$ hdfs dfs -copyFromLocal '/home/ca675/Downloads/union.csv' /stackexchange
```

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.

1	Id	Score	ViewCount	Body	OwnerUserId	Title	Tags
2	20928023	7	24631	<p>I'm	3161430	How to use QFileDialog options and retrieve saveFileName?	<pyqt><exec><pyqt4><qfiledialog>

2	20928023	7	24631	I'm trying to	3161430	1033581	How to use C	pyqt exec pyqt4 qfiledialog

Loading CSV into hive server

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

```
hive> CREATE EXTERNAL TABLE new  (
    >         Id STRING,
    >         PostTypeId STRING,
    >         AcceptedAnswerId STRING,
    >         CreationDate STRING,
    >         Score INT,
    >         ViewCount INT,
    >         Body STRING,
    >         OwnerUserId STRING,
    >         LastEditorUserId STRING,
    >         LastEditDate STRING,
    >         LastActivityDate STRING,
    >         Title STRING,
    >         Tags STRING,
    >         AnswerCount INT,
    >         CommentCount INT,
    >         FavoriteCount INT
    >     )
    >     ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    >     LOCATION '/user/root/stackexchange-input';
OK
Time taken: 1.792 seconds
hive> show tables
    > ;
OK
```

Query 1: Top 10 posts by Score

Input:-

```
hive> select id, title, score from posts order by score desc limit 10;
```

output:-

```
OK
11227809        Why is it faster to process a sorted array than an unsorted array?      17267
927358  How to undo last commit(s) in Git?      13359
2003505 How to delete a Git branch both locally and remotely?   9787
179123  How to modify existing  7687
477816  What is the correct JSON content type?  7671
292357  What is the difference between 'git pull' and 'git fetch'?      7480
111102  How do JavaScript closures work?        7246
1642028 What is the ""-->"" operator in C++?    6313
503093  How to redirect to another webpage in JavaScript/jQuery?       6213
231767  What does the ""yield"" keyword do in Python?   6120
```

Query 2: Top 10 users by post Score

Input:-

```
hive>  select owneruserid, sum(viewcount) as viewc from posts group by owneruserid order by viewc desc limit 10;
```

Output:-

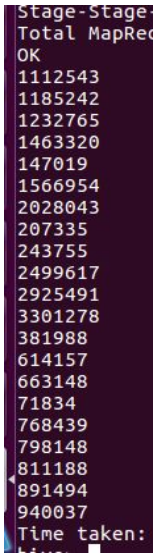
```
Total MapReduce CPU Time Spent: 0 msec
OK
49153   33483594
51816   30713568
4653    25306899
39677   24459705
104015  23453700
4872    21078102
6068    20067921
48523   19505625
4883    18667788
46646   18568620
Time taken: 13.267 seconds, Fetched: 10 row(s)
```

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

Input:-

```
hive> select distinct owneruserid from posts where instr(body, 'hadoop')!=0;
```

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.