Distributed Computing and Big Data Technologies Assignment 1 Project Report

TEAM COMPOSITION:

Ms. Saatchi Nandwani (UB Person Id -50207363), Batch – Blue Code (8 am)

Mr. Akshat Sehgal (UB Person Id – 50198939) – Pink Code (2 pm)

CONTRIBUTIONS:

|  |  |
| --- | --- |
| TASK | CONTRIBUTOR |
| VM & Hadoop DFS Setup | Saatchi/Akshat |
| MapReduce Logic | Akshat |
| Gather inputs | Saatchi |
| Project Report | Saatchi/Akshat |

HADOOP DFS SETUP AND YOUR ENVIRONMENT:

* We installed Oracle VM on our respective Windows machines and used the Hadoop VM image provided by the TA.
* The Linux VM provided to us had Hadoop 2.7.2 and Eclipse 3.1 installed on it.
* We setup Hadoop on the provided VM by providing our JAVA\_HOME and HADOOP\_CLASSPATH paths.
* We modified configuration files like core-site.xml. mapred-site.xml, hdfs-site.xml and yarn-site.xml to setup Hadoop as a single node.

DATA SETS:

We worked with text from some old historic newspaper articles. We obtained the data from the below URL:  
http://chroniclingamerica.loc.gov/.

We worked on 39 articles, decided on the basis of last 2 digits of the person number (50198939).

To concatenate our text files, we used the below web utility:

<http://www.ofoct.com/merge-text-files-online>

DESIGN ISSUES:

While examining the input files, we noticed that a large chunk of the data was corrupted. Hence, we wanted to come up with a mapper program that could uniquely parse proper English words from the text. The idea was to exclude all special characters from the large input text files and then get a unique string token.

For example - 'Tuesday,' ; 'Tuesday.' ; 'Tuesday' should be treated as Tuesday and not as three separate words.  
So we decided on using Java String class’ split() using the regular expression ‘[\\W+](file:///\\W+) ‘ that matches only to words(containing letter and numbers).

PRACTICAL EXPERIENCES:

We faced some blockers during the execution of this project. Below is a summary:

* Namenode did not come up on starting the DFS:  
    
  Frequency of issue - sometimes  
    
  To solve this problem, we had to clear the filesystem and format the namenode before finally starting the dfs again using the below sequence of commands:  
    
  rm -Rf fs.defaultFS  
  hadoop/bin namenode -format  
  sbin/start-dfs.sh
* Datanode did not come up on starting the DFS:  
    
  Frequency of issue – sometimes  
    
  To solve this problem we cleaned the cache of our linux VM, then executed the above set of commands again.