

Week Three Assignment

Objective: To learn about MapReduce.

Instruction: This assignment you need to run a word count program to calculate the words in the War and Peace novel authored by Leo Tolstoy.

The source code is from the [Hadoop: Example: WordCount v1.0 website](#).

Steps:

- 1) **Create a directory names wordcount_classes**
`mkdir wordcount_classes`
- 2) **Compile the WordCount.java program, use the 'hadoop classpath' command as defined by your specific Hadoop installation**
`javac -cp `hadoop classpath` -d wordcount_classes WordCount.java`
- 3) **Create a java archive for distribution**
`jar -cvf wordcount.jar -C wordcount_classes/`
- 4) **Create a directory and move the file into HDFS**
`hdfs dfs -mkdir /user/arman/war-and-peace`
`cp ../../Data/war-and-peace.txt`
`hdfs dfs -put war-and-peace.txt war-and-peace`
- 5) **run work count, but first**
`hadoop jar wordcount.jar WordCount war-and-peace war-and-peace-output`
- 6) **check for output form Hadoop job**
`hdfs dfs -ls war-and-peace-output`
- 7) **move it back to working directory (example of "hadoop dfs -get")**
`hdfs dfs -get war-and-peace-output/part-00000.`

**# Note: If you run program again it won't work because /war-and-peace-output exists.
Hadoop will not overwrite files!**