Assignment 5

MapReduce Word Count

* Harshil Patel (40163431)

Overview :

The goal of the assignment is to implement a Hadoop Daemon on a single node and run a map reduce program to count number of words in multiple files. In addition to this, a stopwords file must be provided to skip all the stop words from the input file before counting the words in them.

Implementation:

The detailed steps to setup Hadoop as a single node on a local machine and run the program have been provided in the other file (Hadoop Setup.docx).

The WordCount.java contains the following methods:

run() : To configure MapReduce job arguments like path to input and output directories, set . up Map and Reduce Classes.

Setup(): To setup the input test files which contains text information and call the . . parseStopWordsFile() method

parseStopWordsFile(): To read stopwords.txt from Hadoop File System and store all the . stop words into a set.

map(): To map all the words from all test files into a list after comparing whether they exist . in stop word list or not

reduce(): To count the occurrences of each words from the mapped list.

Other files : test1.txt, test2.txt, stopwords.txt

After setting up Hadoop, create Input and Output directories inside HDFS using

hadoop fs -mkdir <path>

Upload the test files and stopwords file to HDFS using

hadoop fs -put <src> <dest>

Compile the WordCount.java using

bin/hadoop com.sun.tools.javac.Main <path>/WordCount.java

Create and Execute Jar file using

jar -cvf wordcount.jar -C /home/patel/Downloads/data/ .

hadoop jar wordcount.jar WordCount /WordCount/Input /WordCount/Output /WordCount/stopwords.txt

Check output on terminal using:

hadoop dfs -cat /WordCount/Output/\*