

N words count using MapReduce

We are introducing how to get top N words count from different articles and sort them accordingly using hadoop MapReduce paradigm.

MapReduce Problem Statement:

We have N number of articles in text format and we are interested in finding word frequency and also want to sort them accordingly so that we can find which words are most occurring among all those files.

I have tested the code in following environment

- Java: 1.7.0_75
- Hadoop: 1.0.4
- Sample Input:

we do have N number of files in text format. I have used 20 big text files for performing this test.

- Data Preparation:

Once we have collected all the input files, we have to upload them in HDFS.

I have created /input/articles directory and put all those files in that directory.

- Solution :

We will use 2 steps to perform this task.

- 1.Using core MapReduce

We will use 1 mapper for parsing the files and count the single word occurrence of a particular word.

We will use 1 reducer for the total count of the word frequency.

Once the mapper and reducer task is completed, we will have a partition file in our HDFS.

- 2. We will sort the data using sort utility based on frequency count data.

I will give a detailed explanation of this program and how to run it at the end of this document.