**Apache Hadoop MapReduce**

**Objectives**:

* Understand the MapReduce programming model.
* Setting up Hadoop on a single node and a cluster of nodes.

**Overview**:

* It is required to install Hadoop on both single node cluster and multiple node cluster.
* Next, you will practice running few HDFS commands and executing Hadoop jobs.

**Procedure**:

1. Initiate all the Hadoop Daemon processes by the following command

**start-all.sh**

1. Create a sample directory in HDFS using the below command in the picture (Create 3 directories named input, Input, sample\_directory)



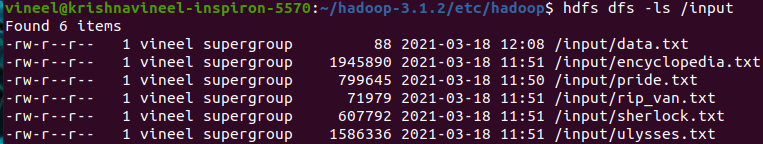
1. Download some Plain Text files of UTF-8 format related to some of the books as mentioned in the assignment sheet. (From Gutenberg Project) use the following command to download the text files

**wget -O <filename.txt> <URL link>**

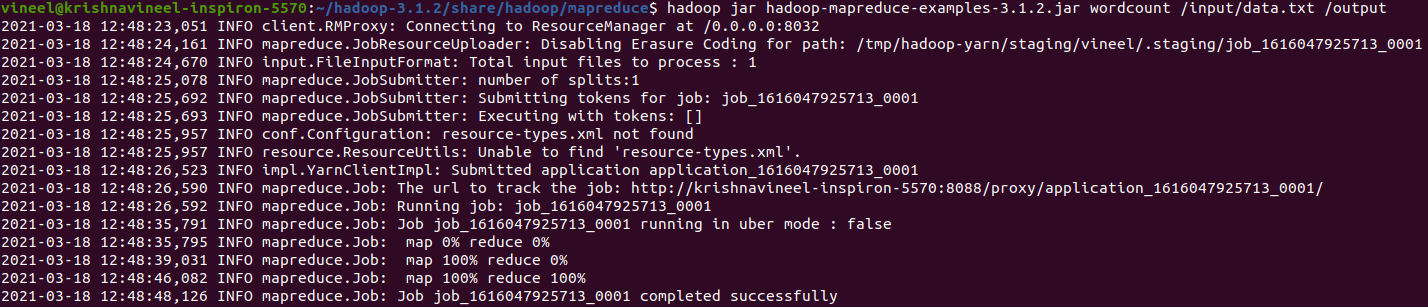
1. Store the .txt files into a directory called “Guten” and prepare a special data.txt file which has 4 lines of data - <college\_name> <roll\_number, year of join>. Create an “intro.txt” which has a few basic lines of data.
2. Now using the Terminal in Ubuntu, copy these .txt files into HDFS directory named “input” and “Input”.
3. Copy the data of all books in addition to “data.txt” into the HDFS directory “input” and copy the “intro.txt” into HDFS “Input” directory.

-copyFromLocal.png

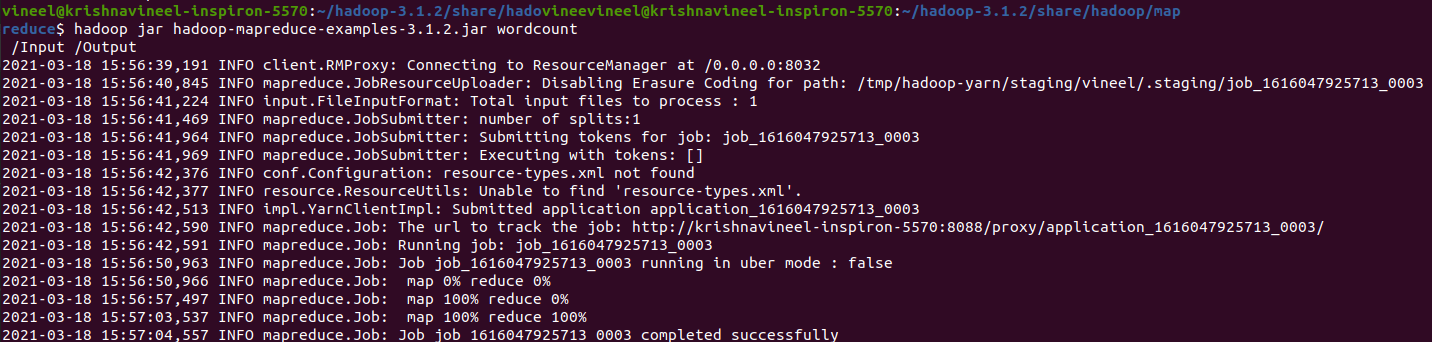
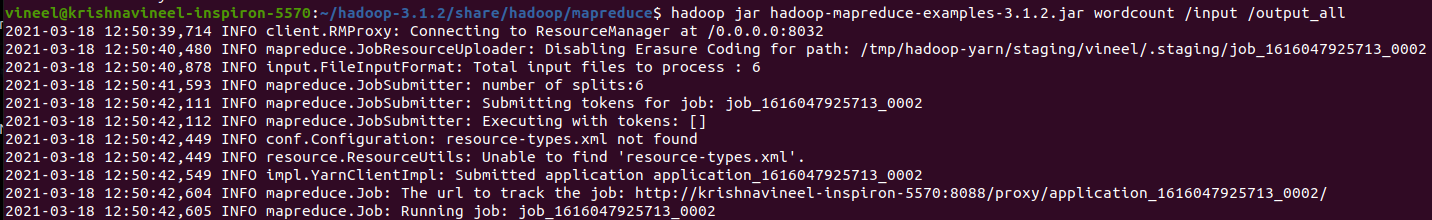
1. Using the Terminal, check whether the files got copied successfully or not.



1. Now run the Hadoop-MapReduce word count program on the “data.txt” file and store the result in the “output” directory.



1. In the “output” directory, it can be observed that there are count values or the <key, value>: <word, count> pairs.
2. Similarly, run the word count program for all the 5 text files that are taken from the Gutenberg project.
3. The output for all the text files in the “input” directory is stored in the “output\_all” directory, whereas the output for “intro.txt” in the “Input” directory is stored in the “Output” directory.



1. Here 3 jobs were done, namely

* Running word count on data.txt 🡪JOB1
* Running word count on text files taken from Gutenberg project 🡪JOB2
* Running word count on intor.txt 🡪JOB3

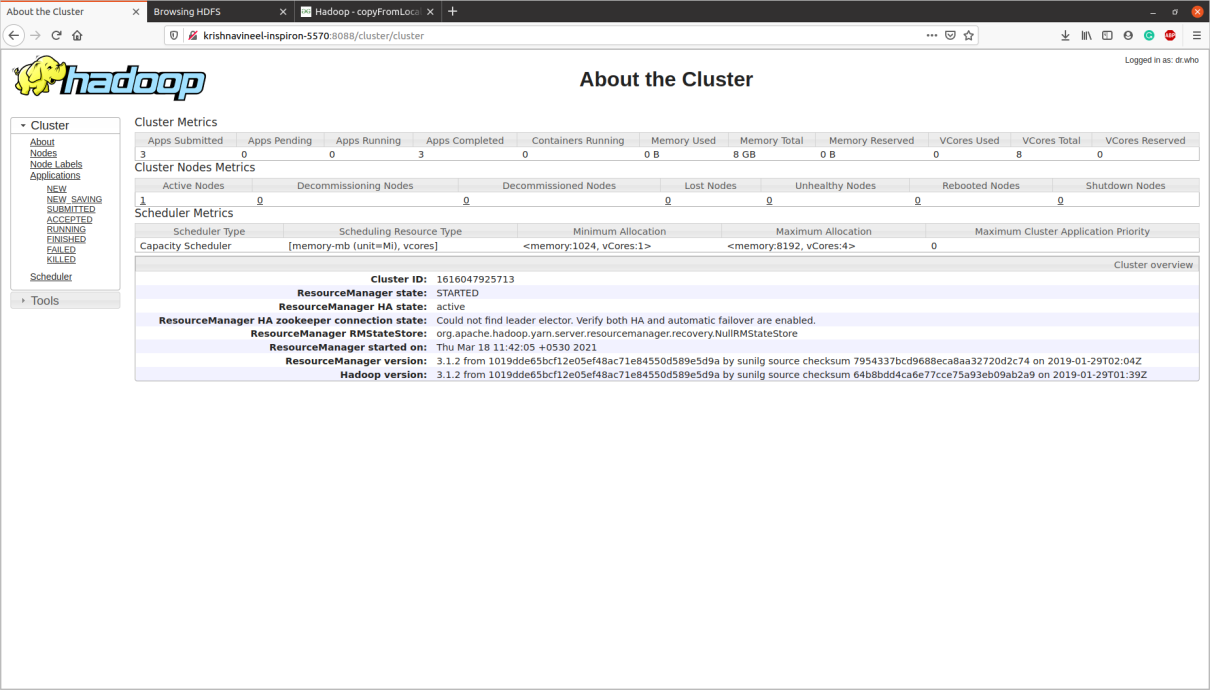
1. Now store the output data into the local, machine using the following command.

copyToLocal.png

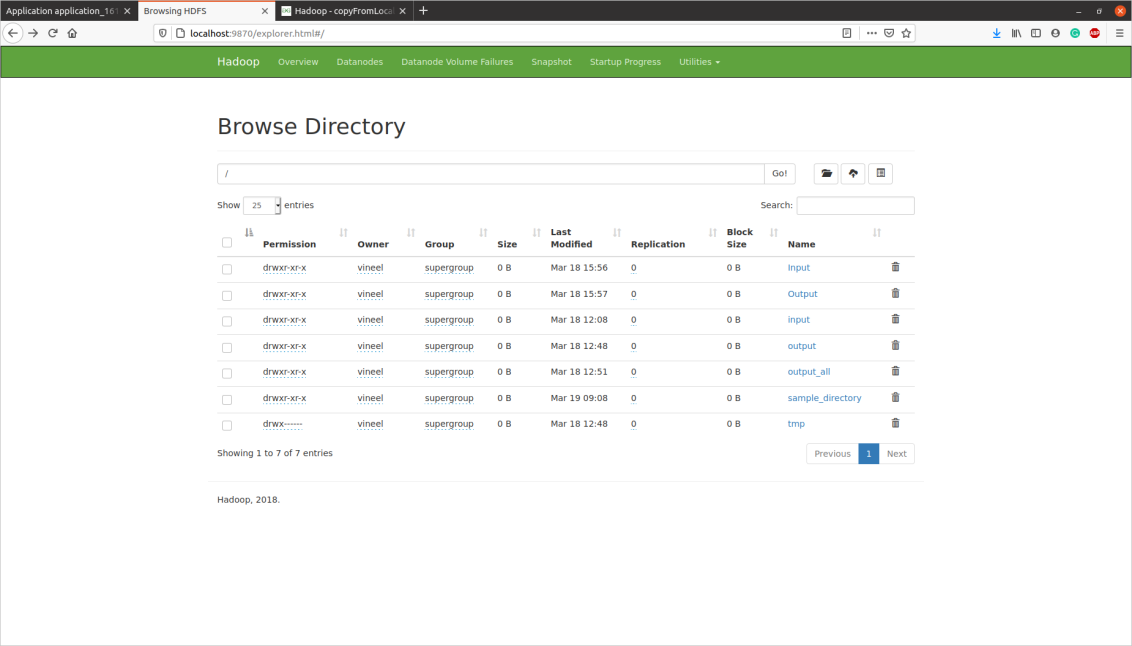
1. All jobs were accomplished and find the attached images below.
2. Stop all the daemon processes by using the below command

**stop-all.sh**

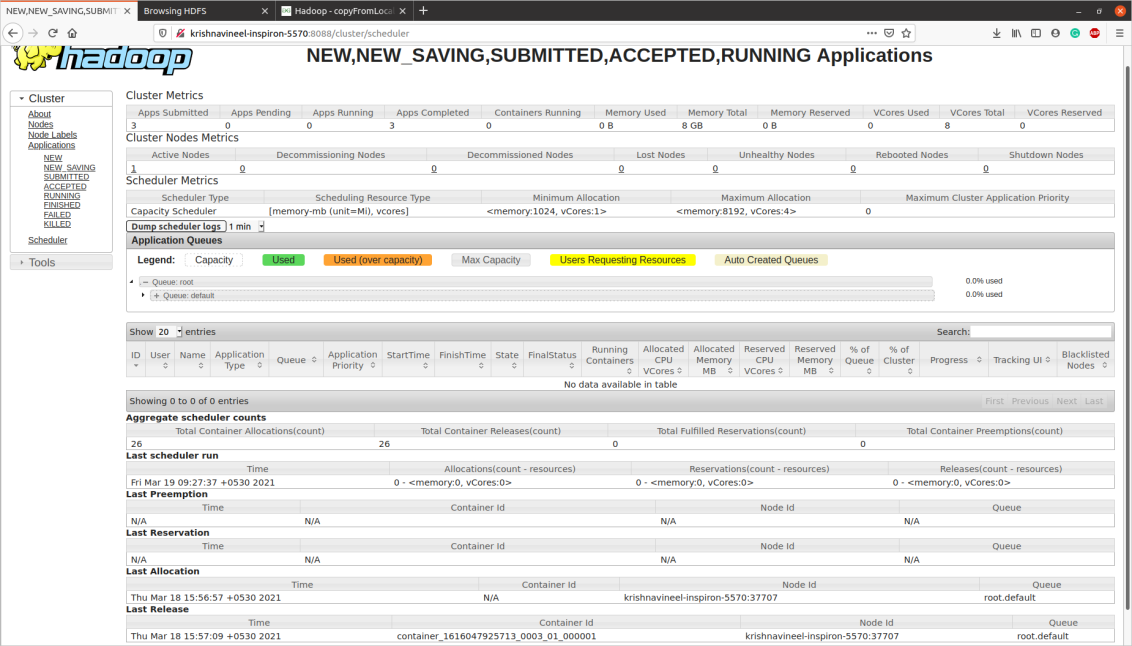
Here are some of the images that describe the details of the jobs and node.



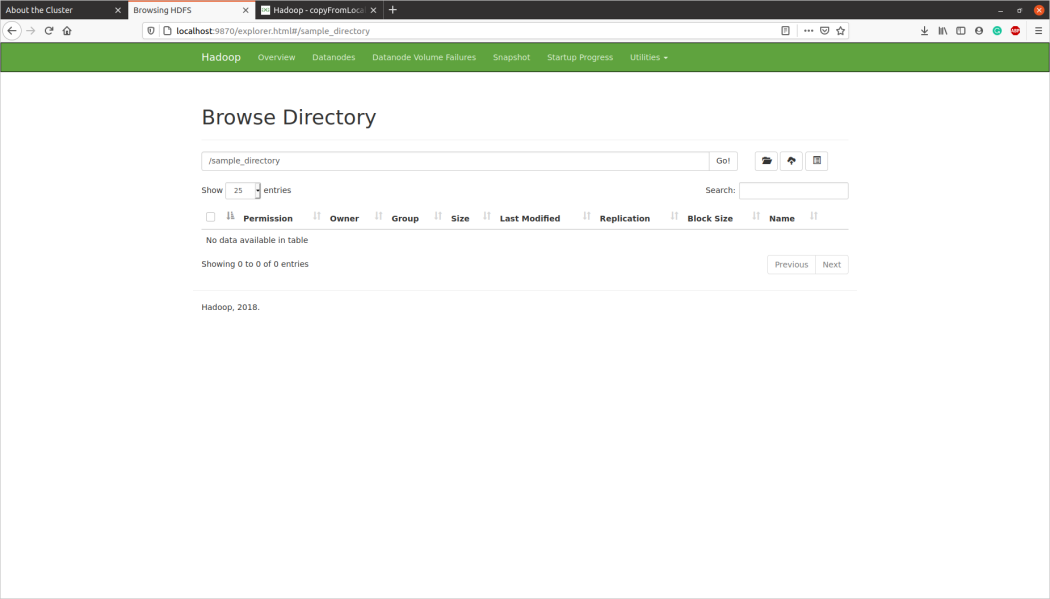
About the Cluster



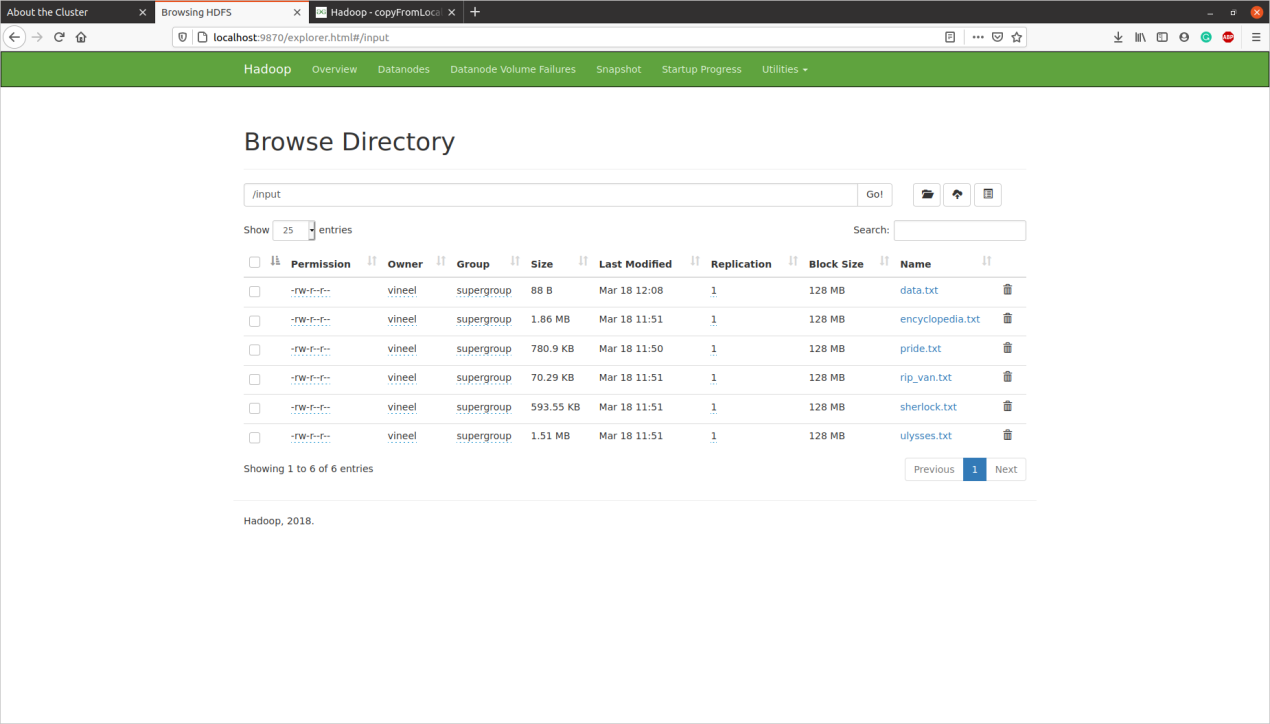
Browse Directory



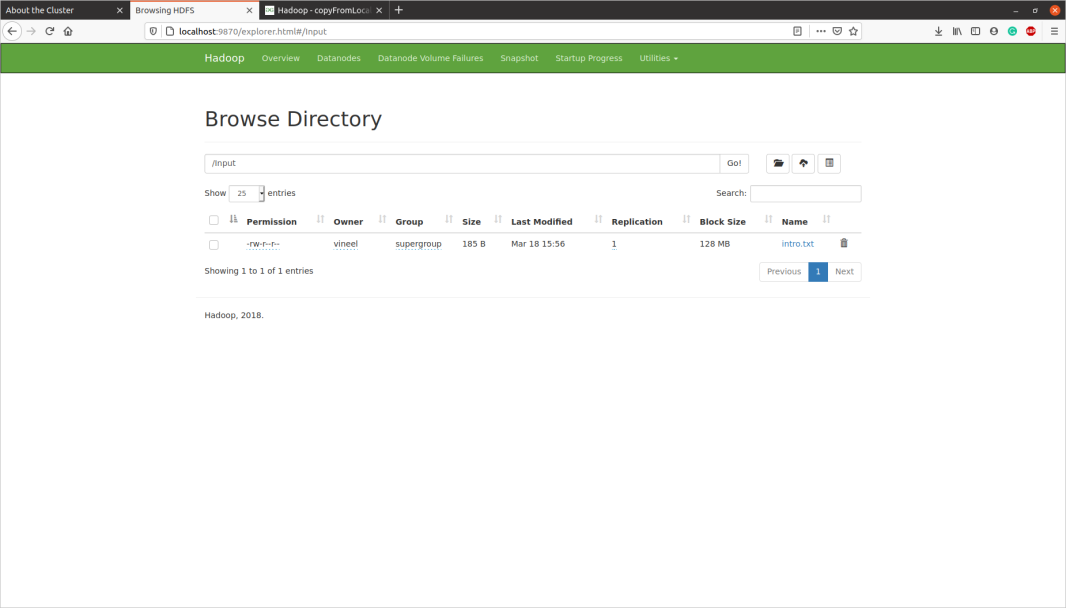
Job Scheduler Details

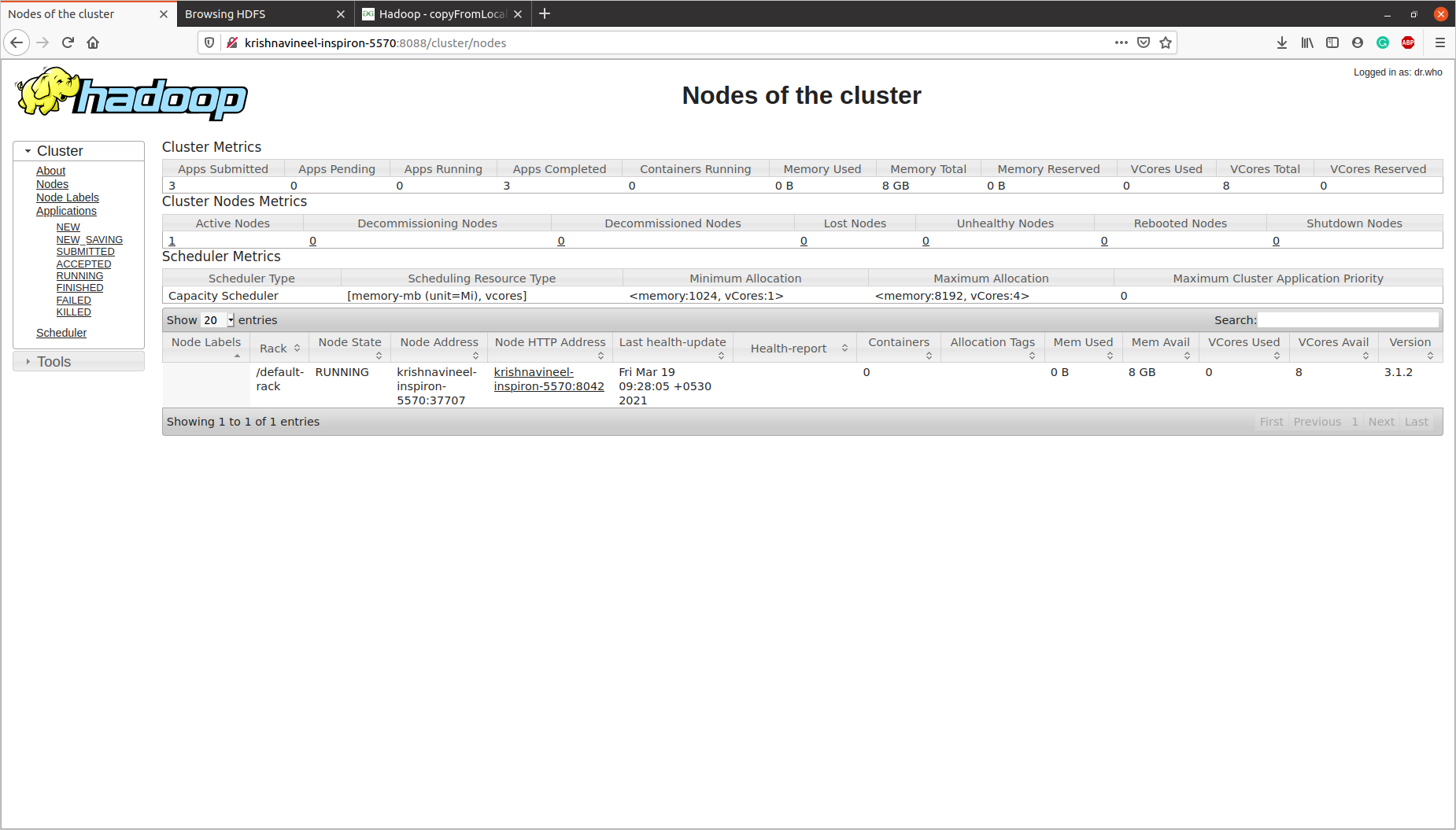


Sample\_directory

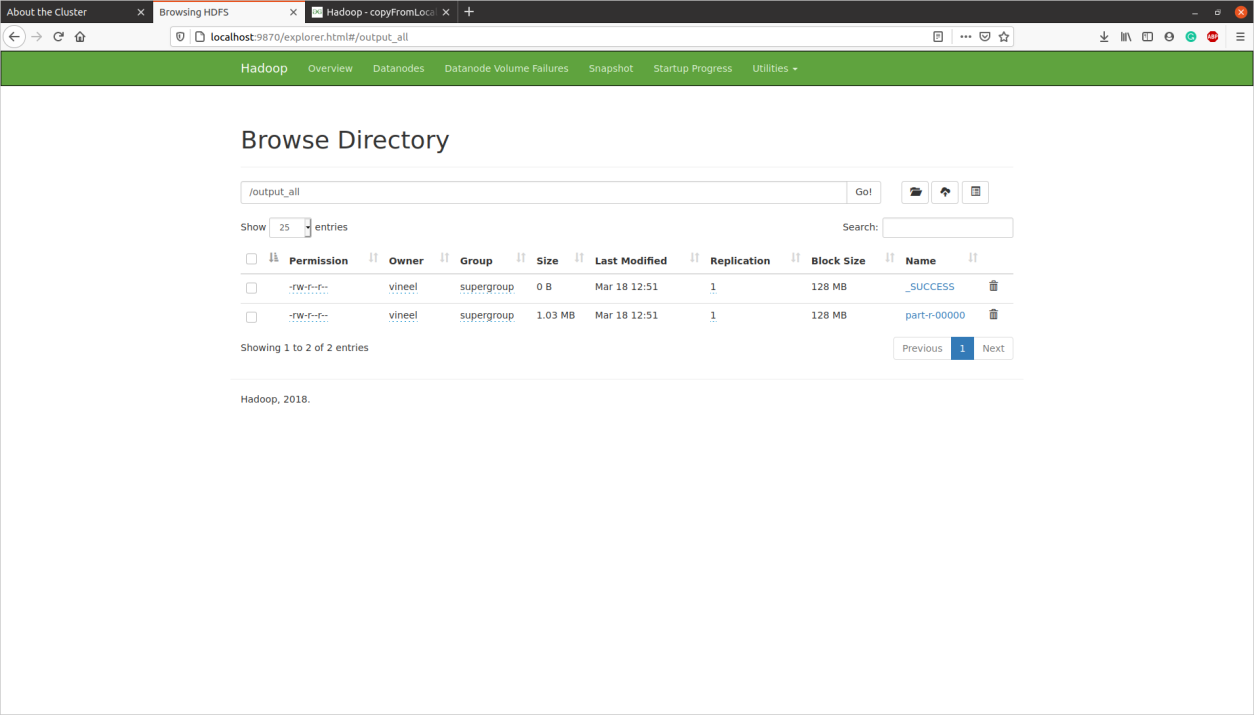


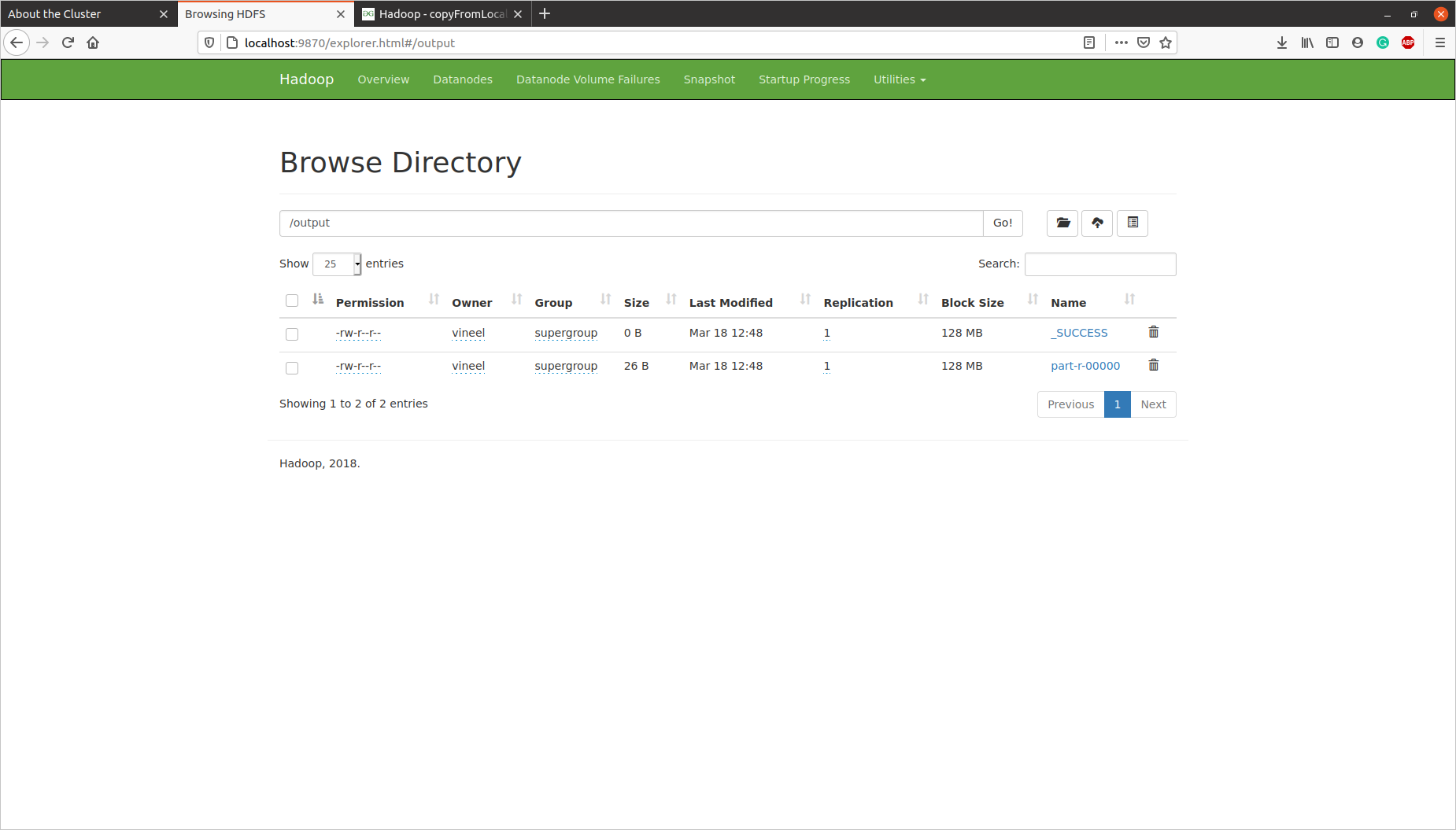
input directory



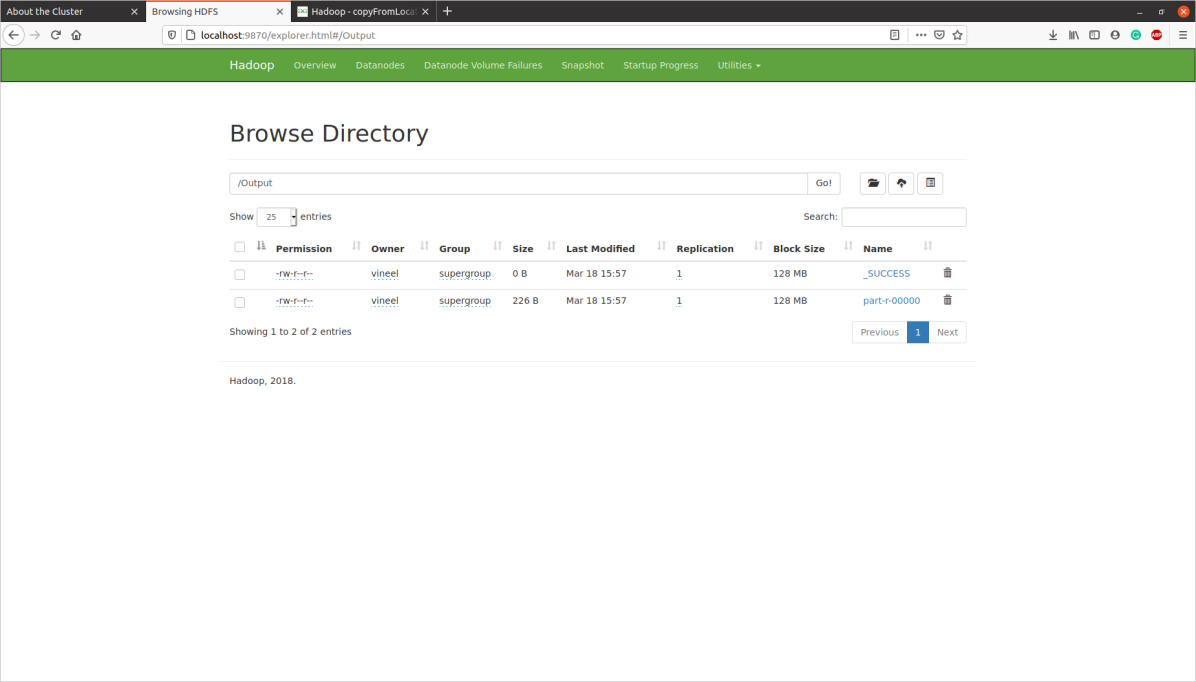
Input directory

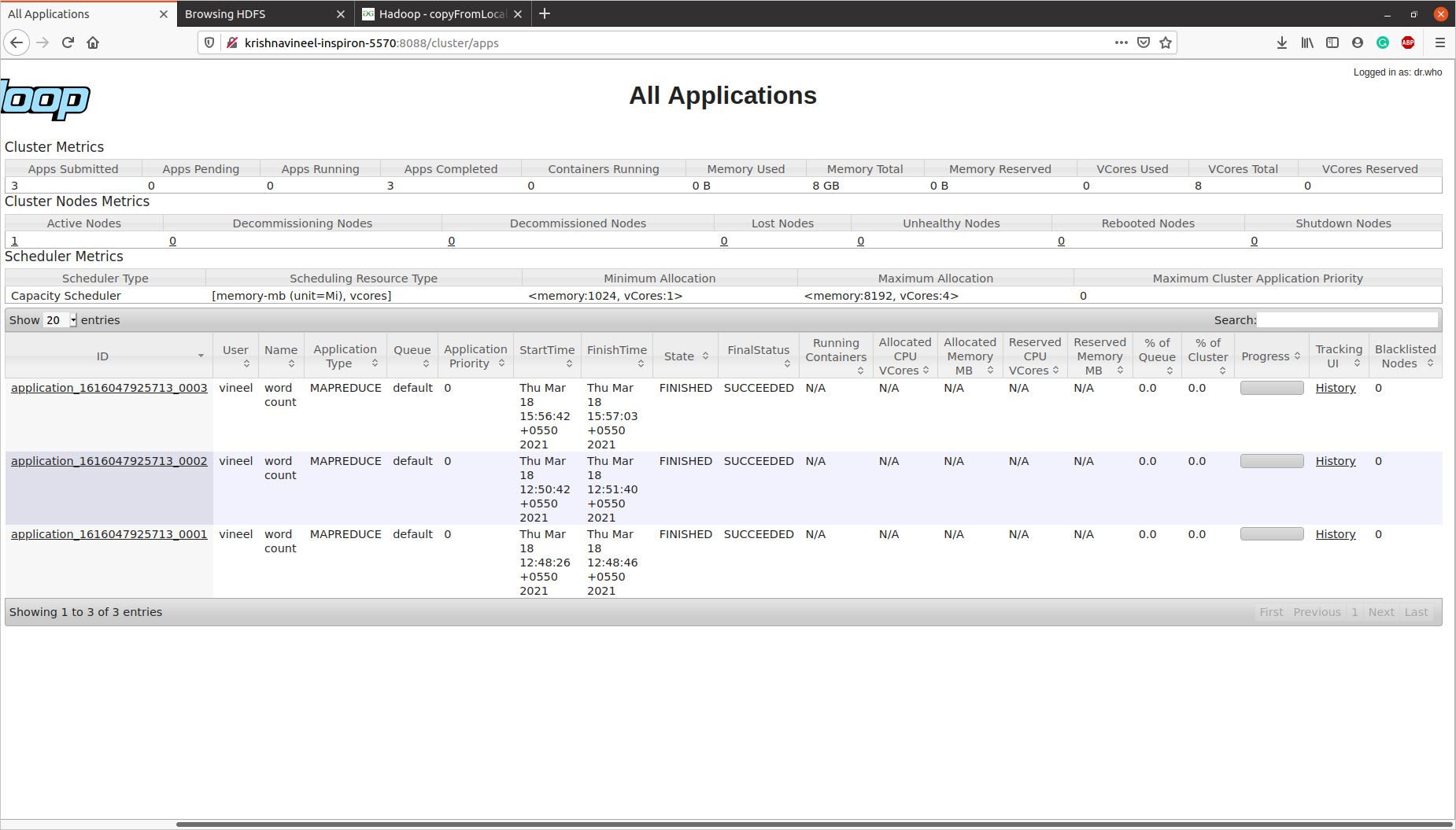
Nodes of the cluster



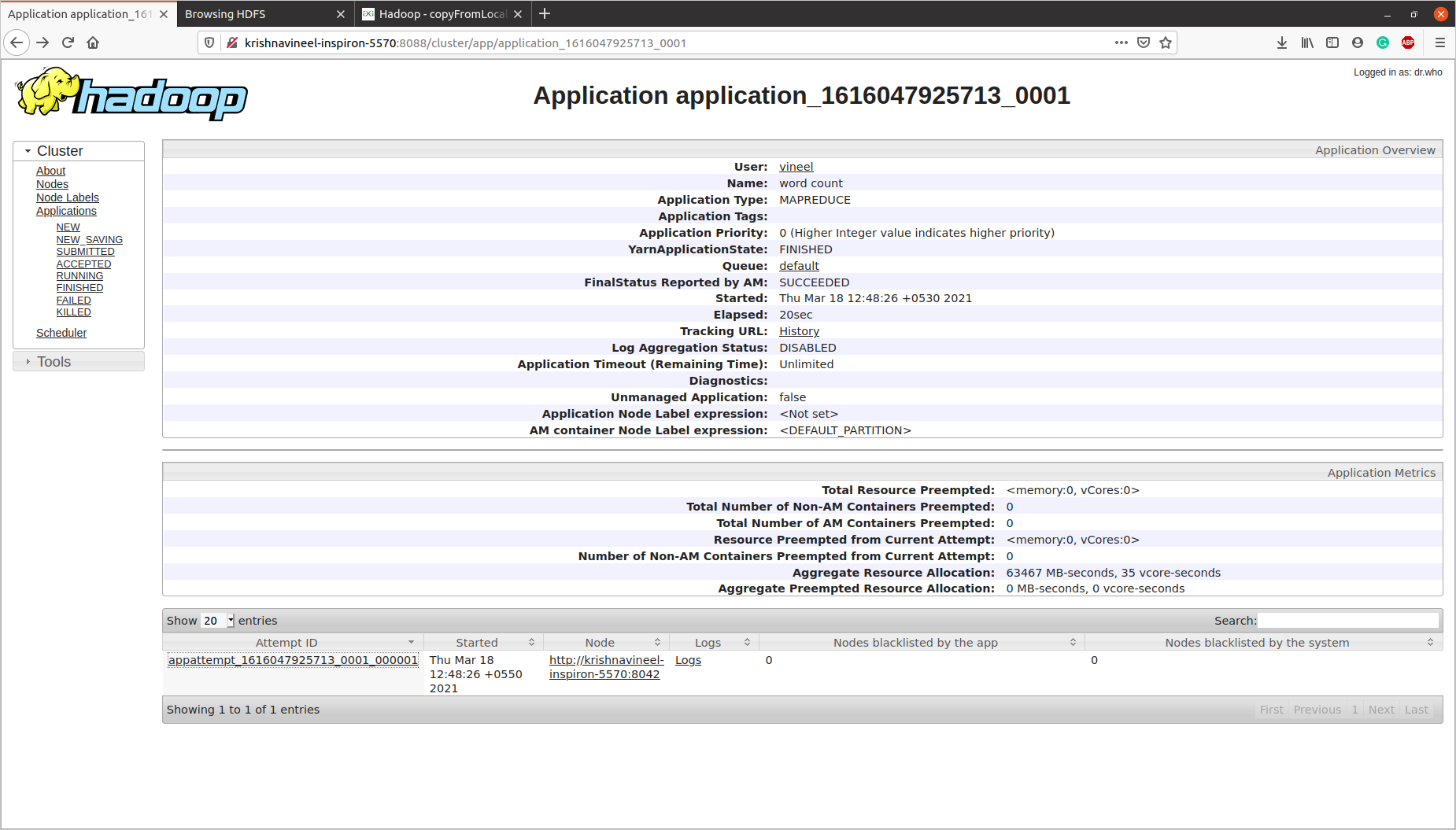
output\_all directory

output directory

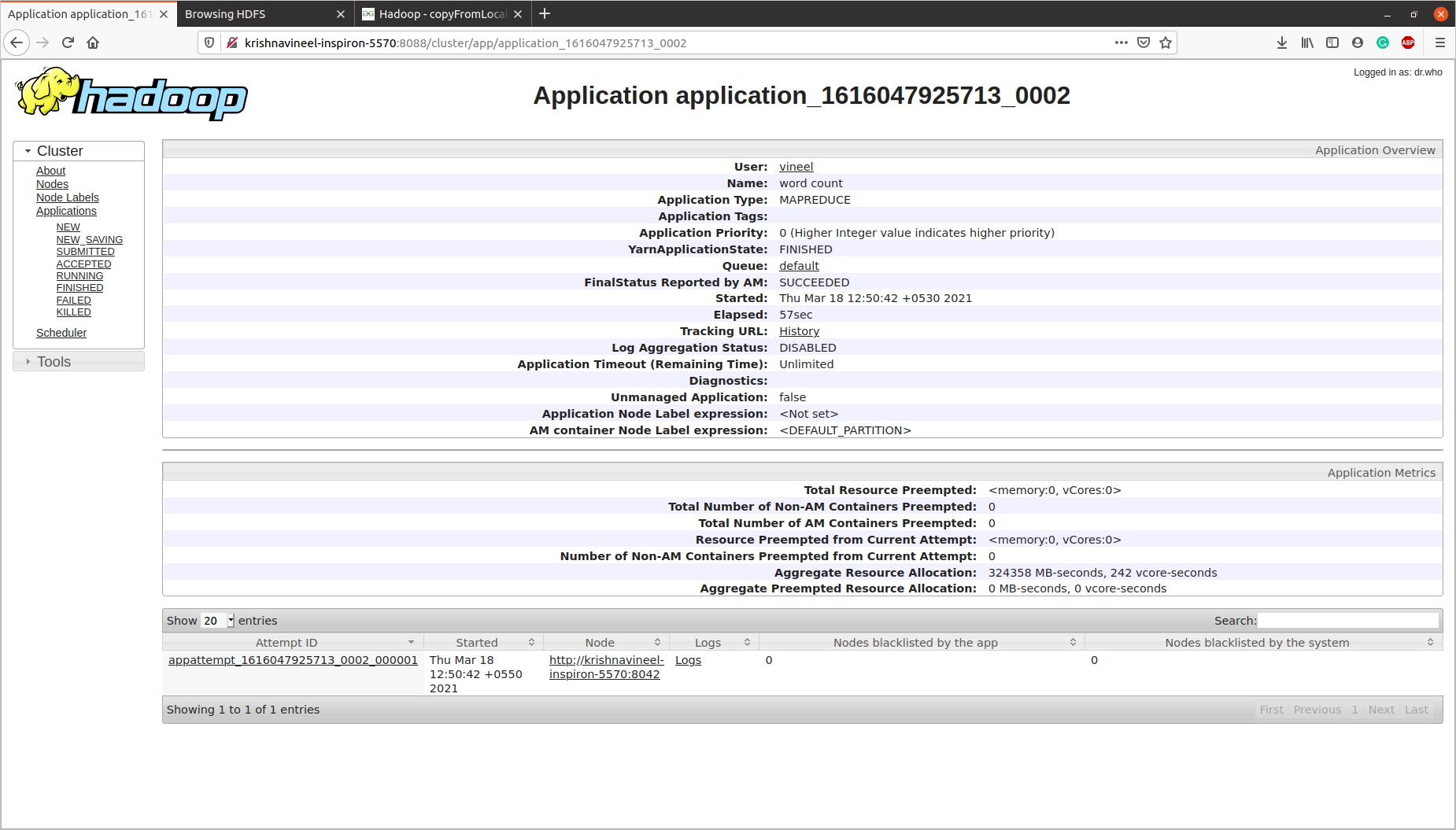


Output directory

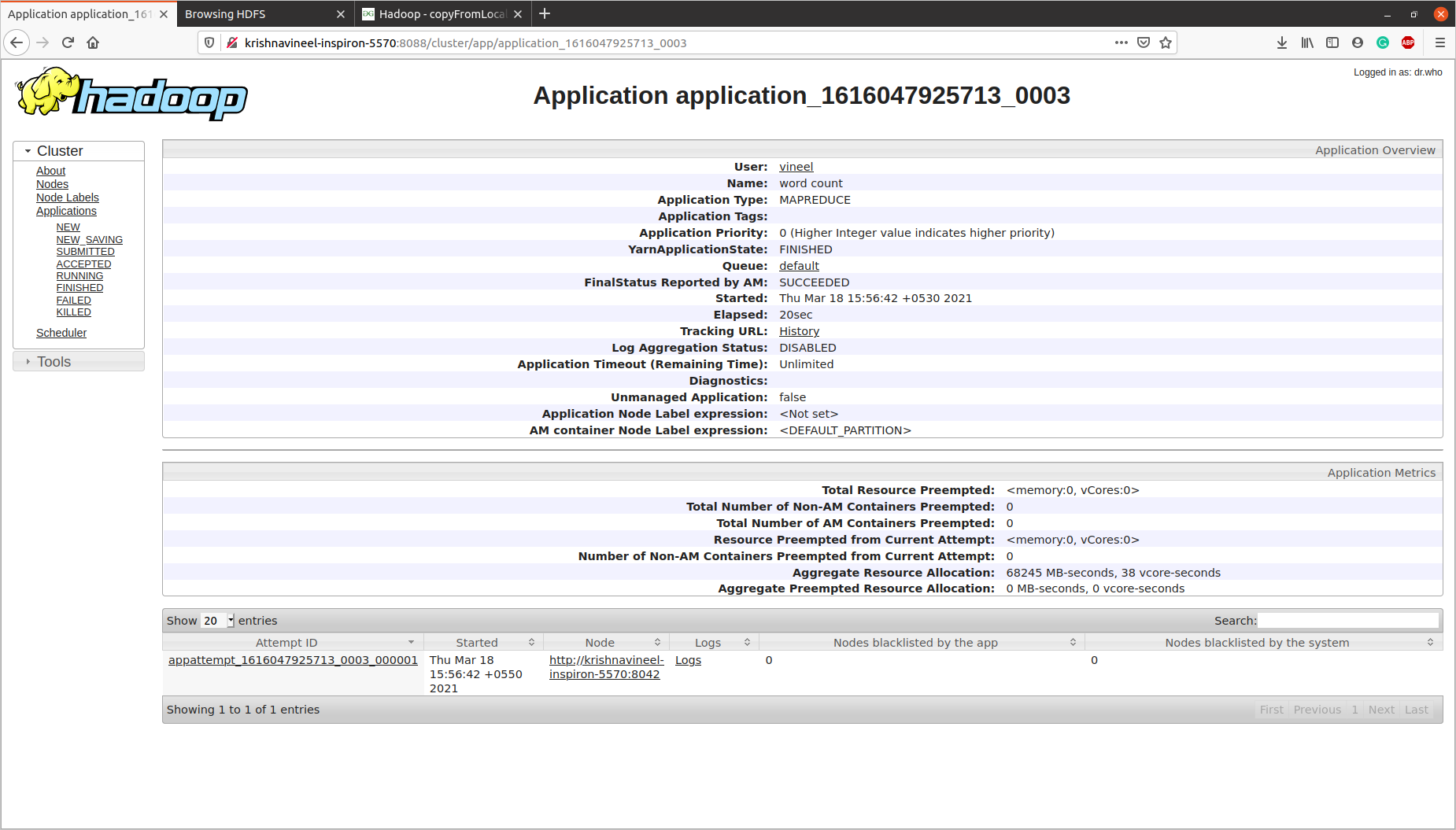
Details of All Applications



JOB – 1 Web Interface



JOB – 2 Web Interface



JOB – 3 Web Interface

**REFERENCE LINKS:**

* <https://drive.google.com/drive/folders/1DfFzXTgnkqZdo6y4SxUX326aHgEWDw52?usp=sharing>
* <https://drive.google.com/drive/folders/13yFAfYe7XI1u7FZZAq6aWteNC6dCVd8m?usp=sharing>

The above provided links contains the screenshots of the outputs and all the text files with their output files as well. It also contains the <key, value>: <word, count> pairs for all the words present in the text files.

THE END