**Principles of Big Data – Project Phase 1**

**Team Members:**

Sravya Para

Hemasree Koganti

Sai Priyanka Gurajada

**Source code link:** <https://github.com/sravyapara/bigdata_project/tree/master/phase1>

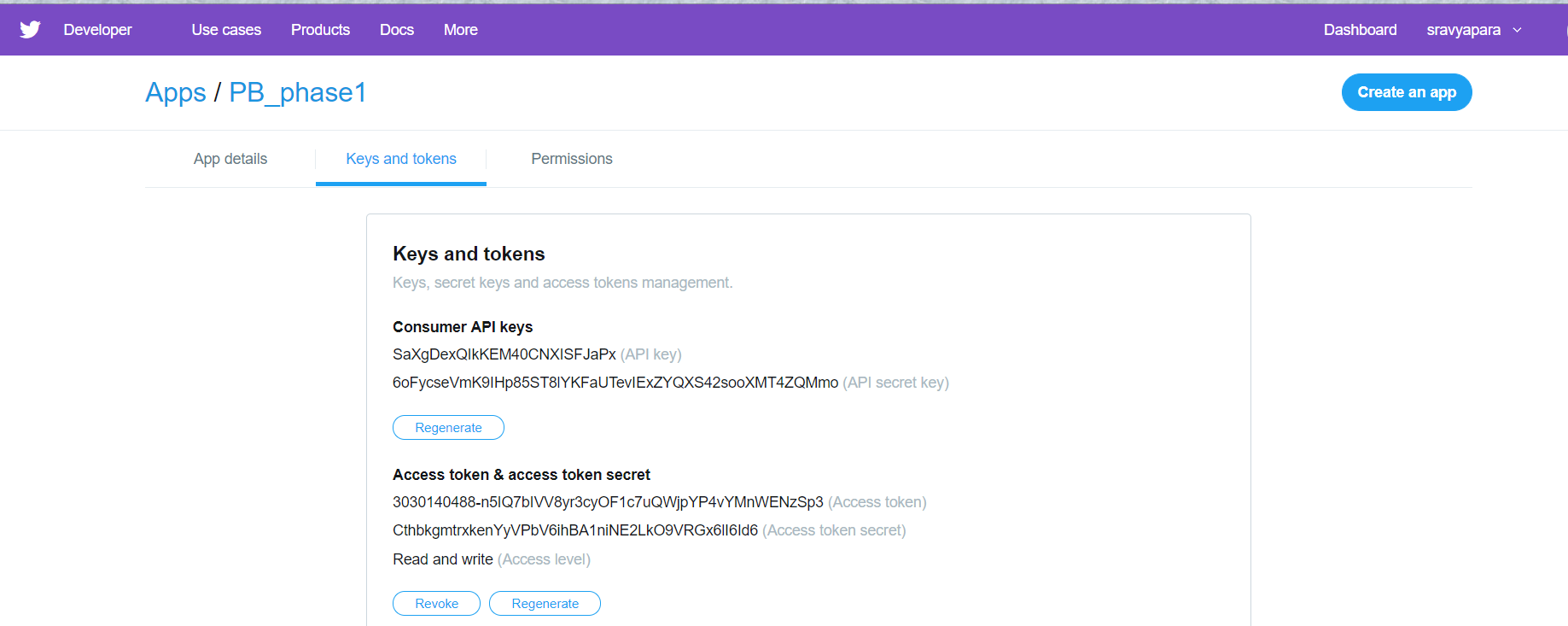
In this phase we collect the tweets from twitter and extract the hashtags and URL’s and run the wordcount program in Apache Hadoop and Apache Spark

**Tasks in Phase 1:**

* Generating the Access tokens and consumer keys from twitter API.
* Collecting the tweets using python tweepy module in a json file and then extracting the URL’s and hashtags into a txt file.
* Running the word count on the extracted URLs and hash tags file in Apache Hadoop and Apache Spark.

**Generating the tokens in twitter:**

Generated the twitter tokens from developer account



**Collecting the twitter data:**

Collected data from twitter using tweepy module based on mostly trending topics like Netflix, marvel.

Below is the code to collect data from twitter



**Extraction of the hash tags and urls in twitter data:**

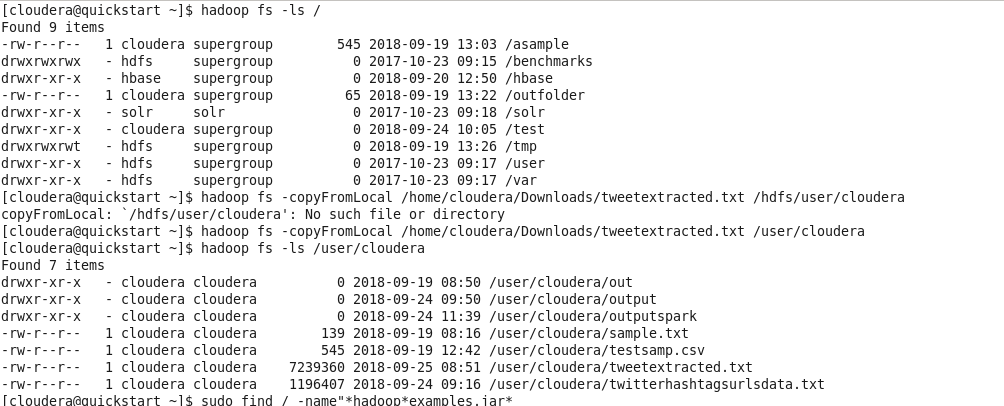
The hashtags and URLs are extracted from the json file .

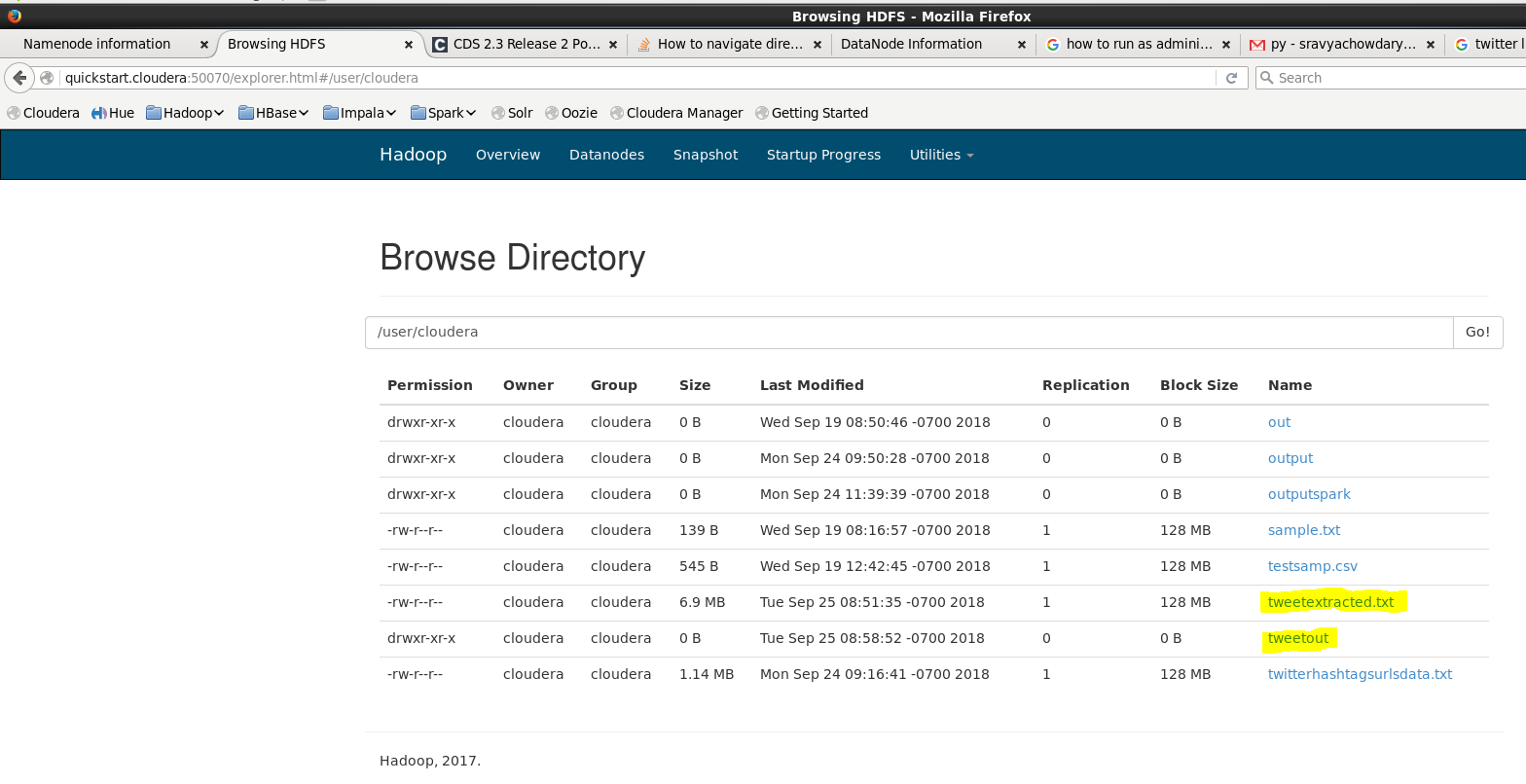
The code snippet as below:



**Word count in Apache Hadoop**

The extracted file is placed in the HDFS location.

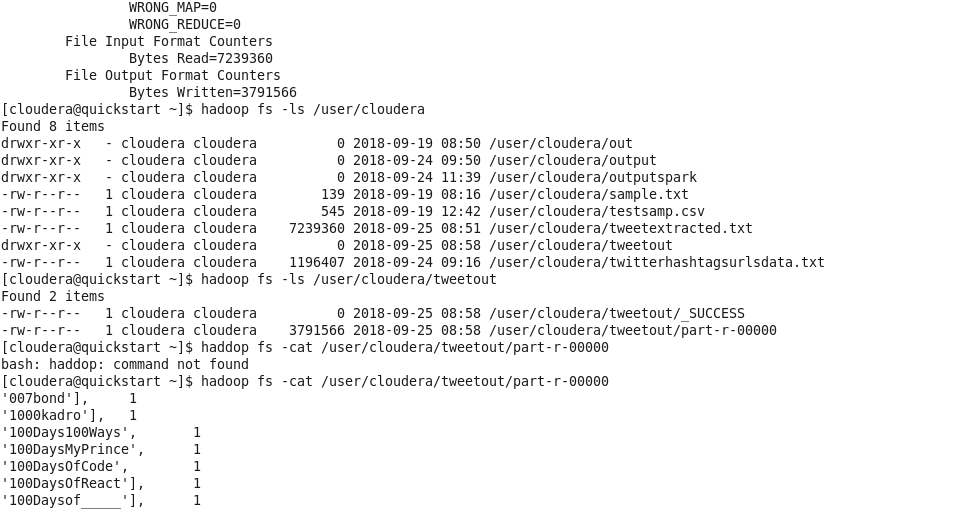




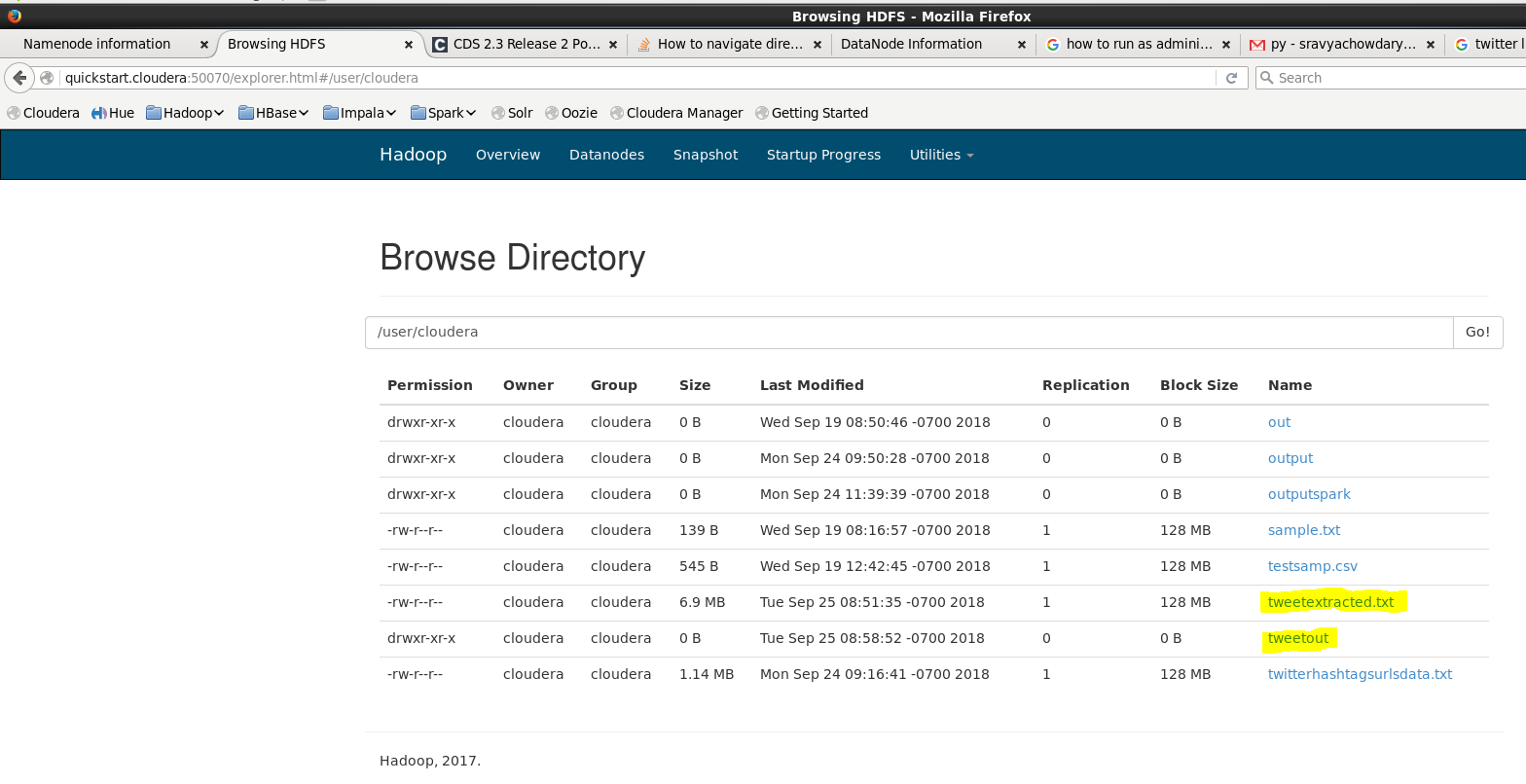
We ran the wordcount using a jar file . BY running a sudo find / -name “\*hadoop\*examples\*.jar” , we could get a hadoop jar file for running the word count.

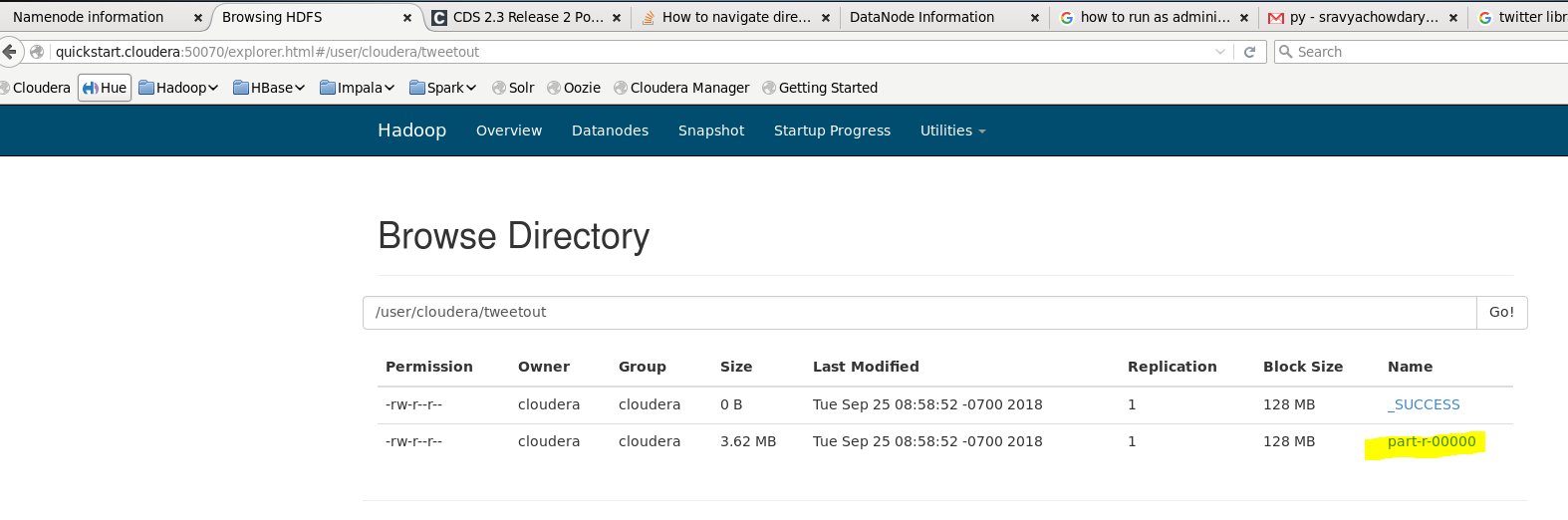


Ouput:



Ouput files being saved in the hdfs location:





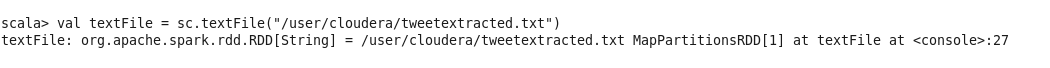
**Word Count in Apache Spark:**

The same input file in the hdfs location is made used.

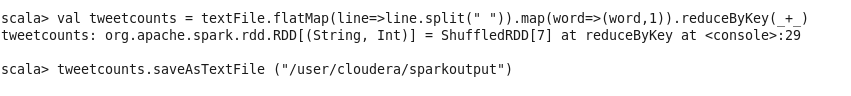
The following screenshots shows how we ran the wordcount in spark

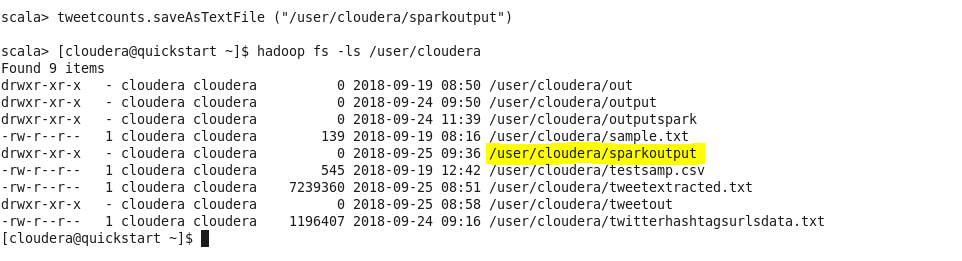


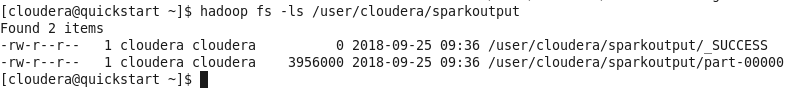
Giving the input file path



For the word count:

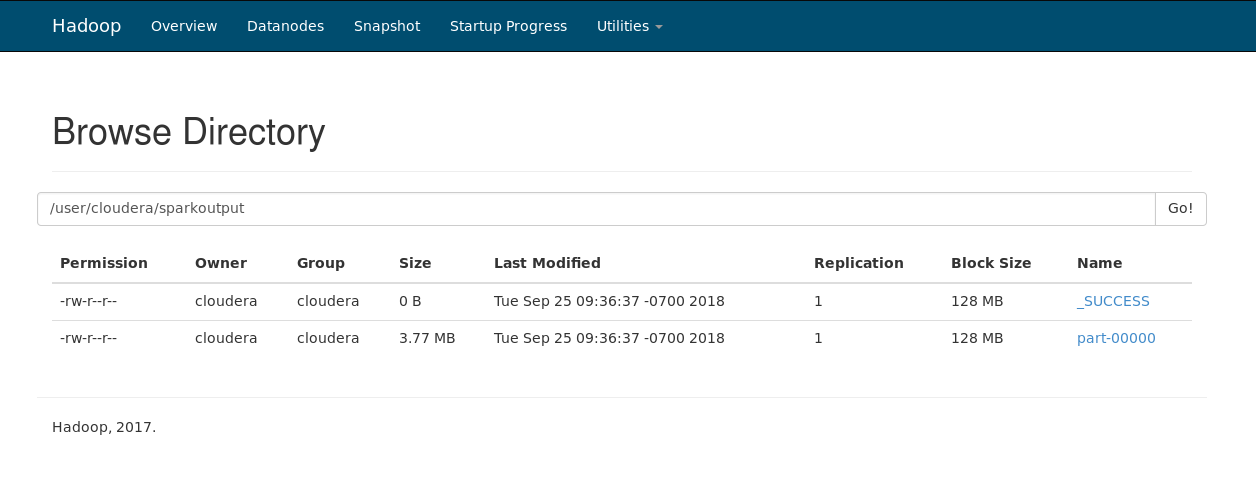




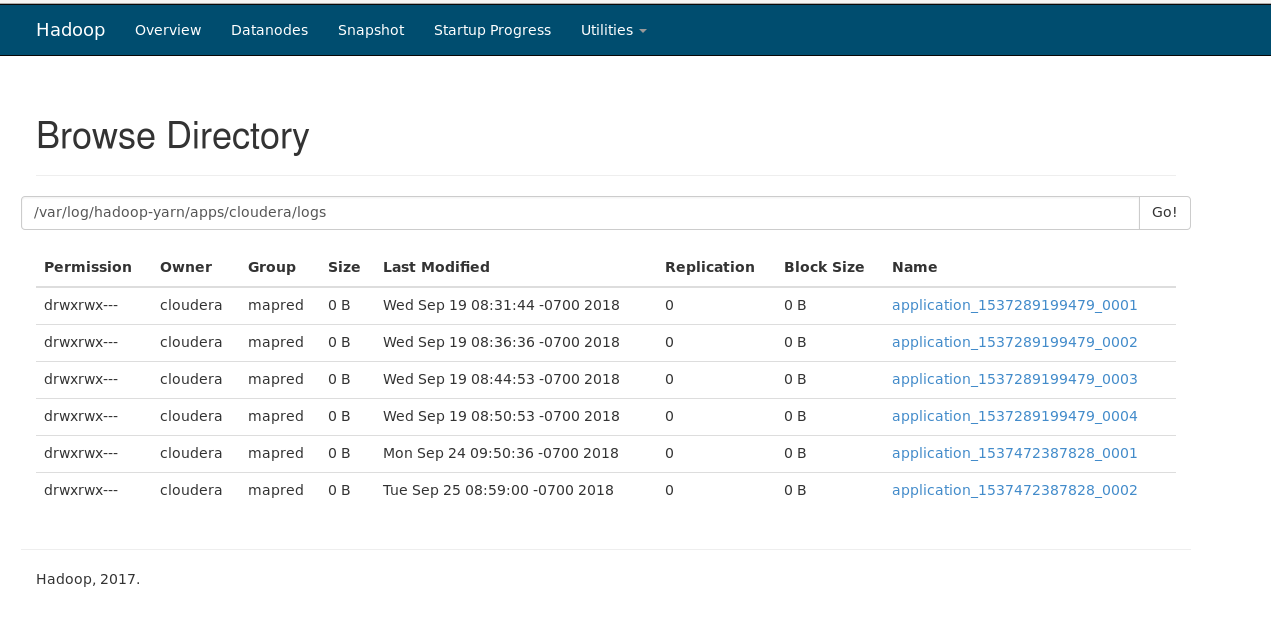


Output:





Collecting the log files:



The output files and log files can be found in below GitHub directory:

<https://github.com/sravyapara/bigdata_project/tree/master/phase1>