**Project Phase 1 Submission**

**Team members:**

Hari Manasa Ganganagunta

Sree Naga Sayana

Akhila Naga Venkata Atluri

**Source Code :**

https://github.com/harimanasag/BigDataManagement/tree/master/project%20phase%201/Source

**Goal:**

To collect Tweets using Twitter’s Streaming APIs

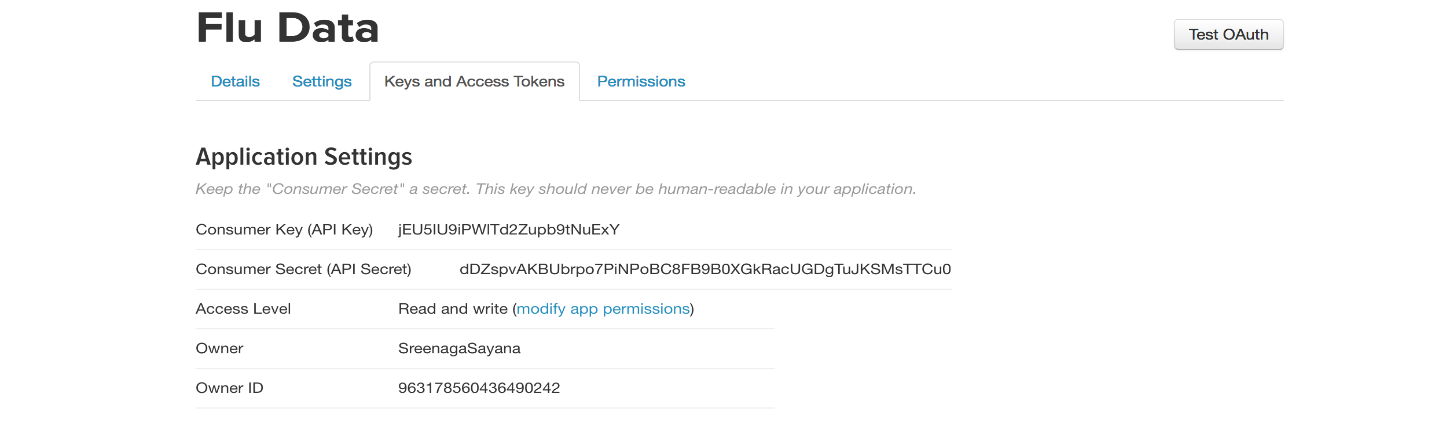
Extracted all the hashtags and URLs in the tweets

Performed Word Count example in Apache Hadoop and Apache Spark on the extracted hashtags/URLs and collect the output and log files from Hadoop.

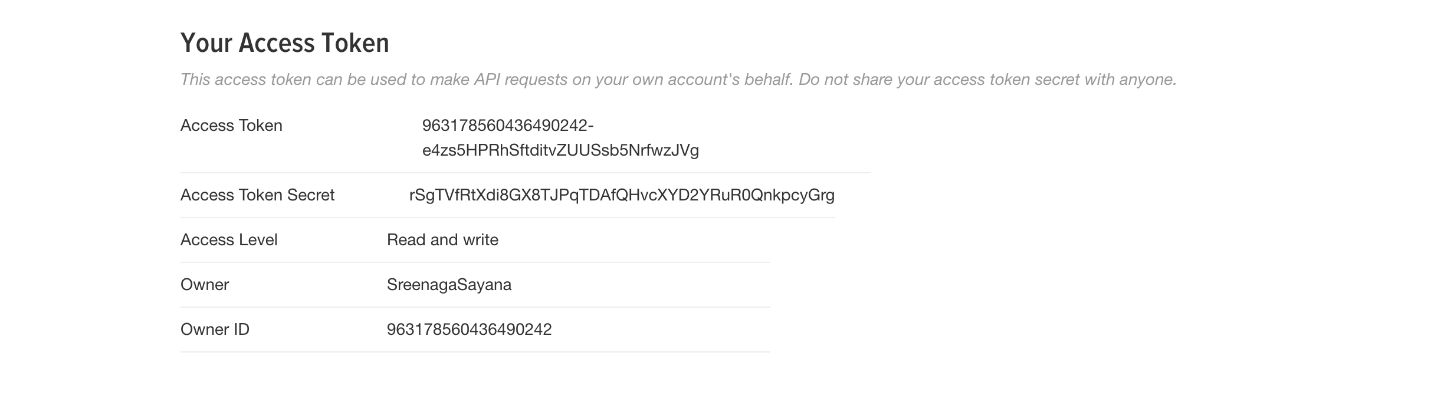
Add a README file.

**Approach:**

We have created a new application which enabled the generation of consumer key and consumer secret.



Requested for Access token generation, and the access token and access token secret were generated.



**Extraction of twitter data using python code**:

We have extracted the twitter data of size 1.4 GB, we have added filter naming SERIES for the extraction of data from twitter.

Source code is available in the github link provided in the first page.

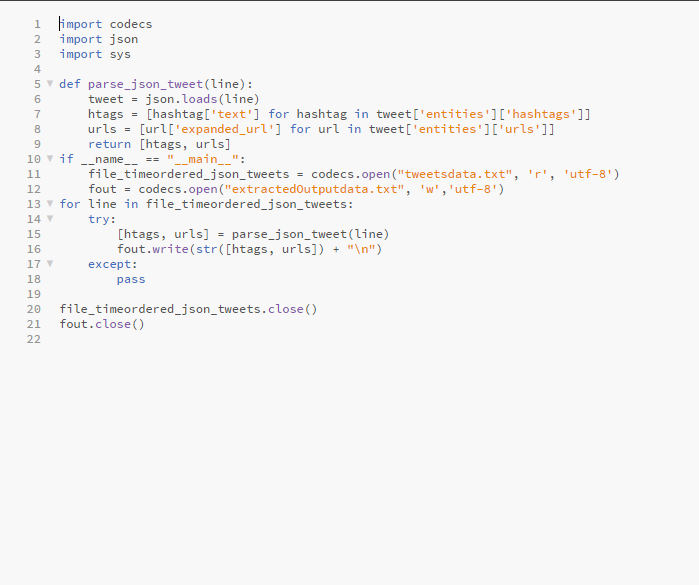
Screenshot of the code is passed below:



Extracted twitter data is consolidated to only hashtags and URLS using python code.

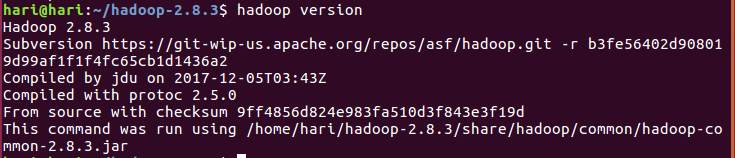
Source code is available in the github link provided in the first page.

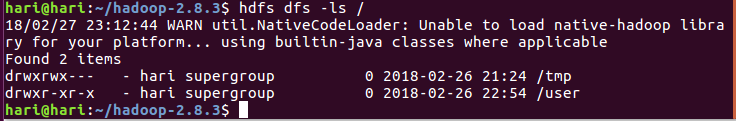
Screenshot of the code is placed below:



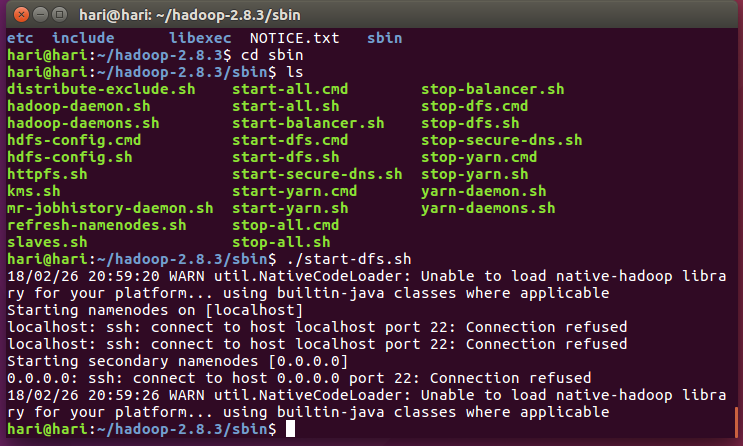
**Performed Word count example in Apache Hadoop**

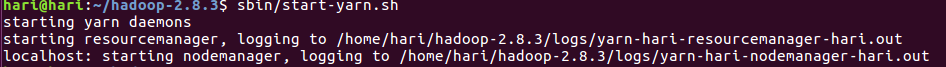
Checking for the Hadoop version:



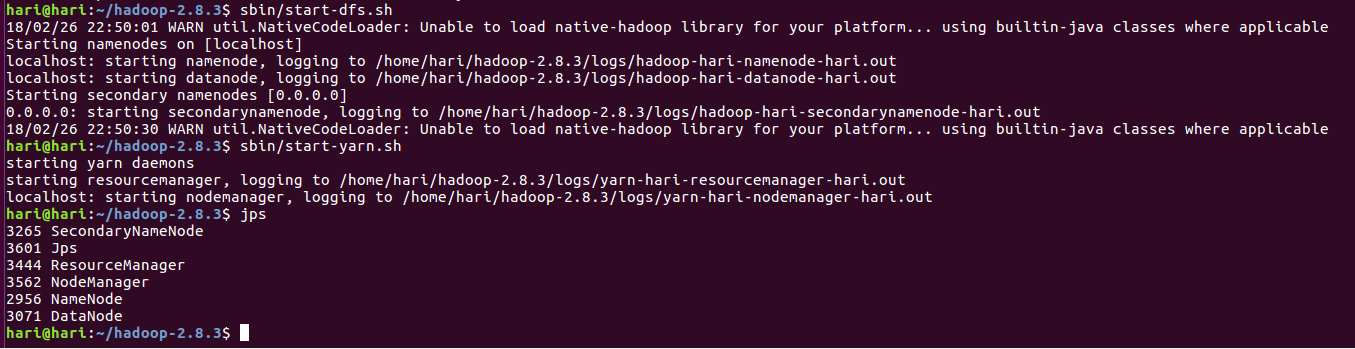


Starting dfs.sh and yarn.sh:





JPS Command:



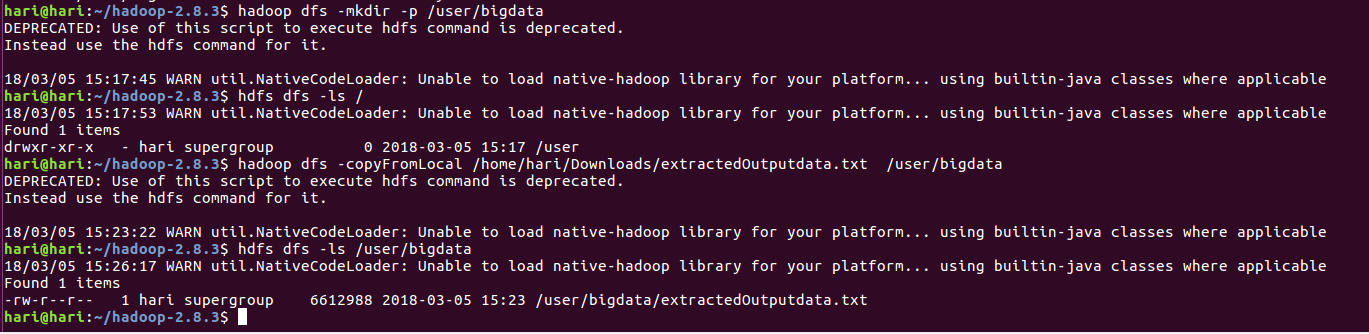
**Creation of directory:**

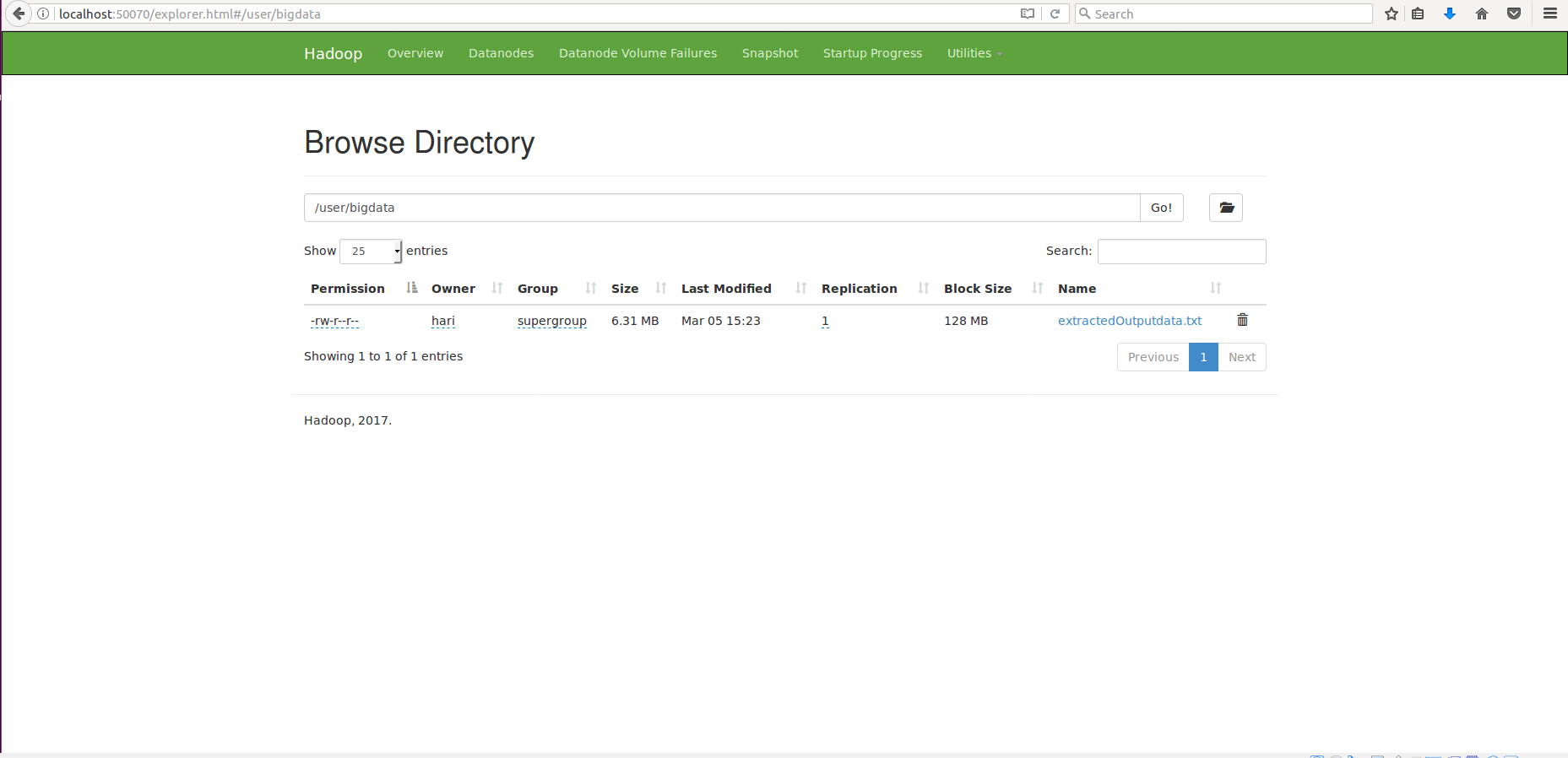
We have created a directory in the localhost:50070, under utilities > browse directory

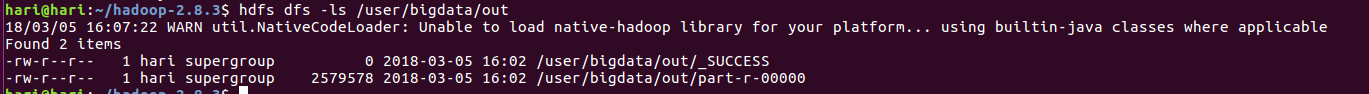
Named the directory as /user/bigdata.

And then copied the output file from the local directory to dfs

Please find the screenshot below:



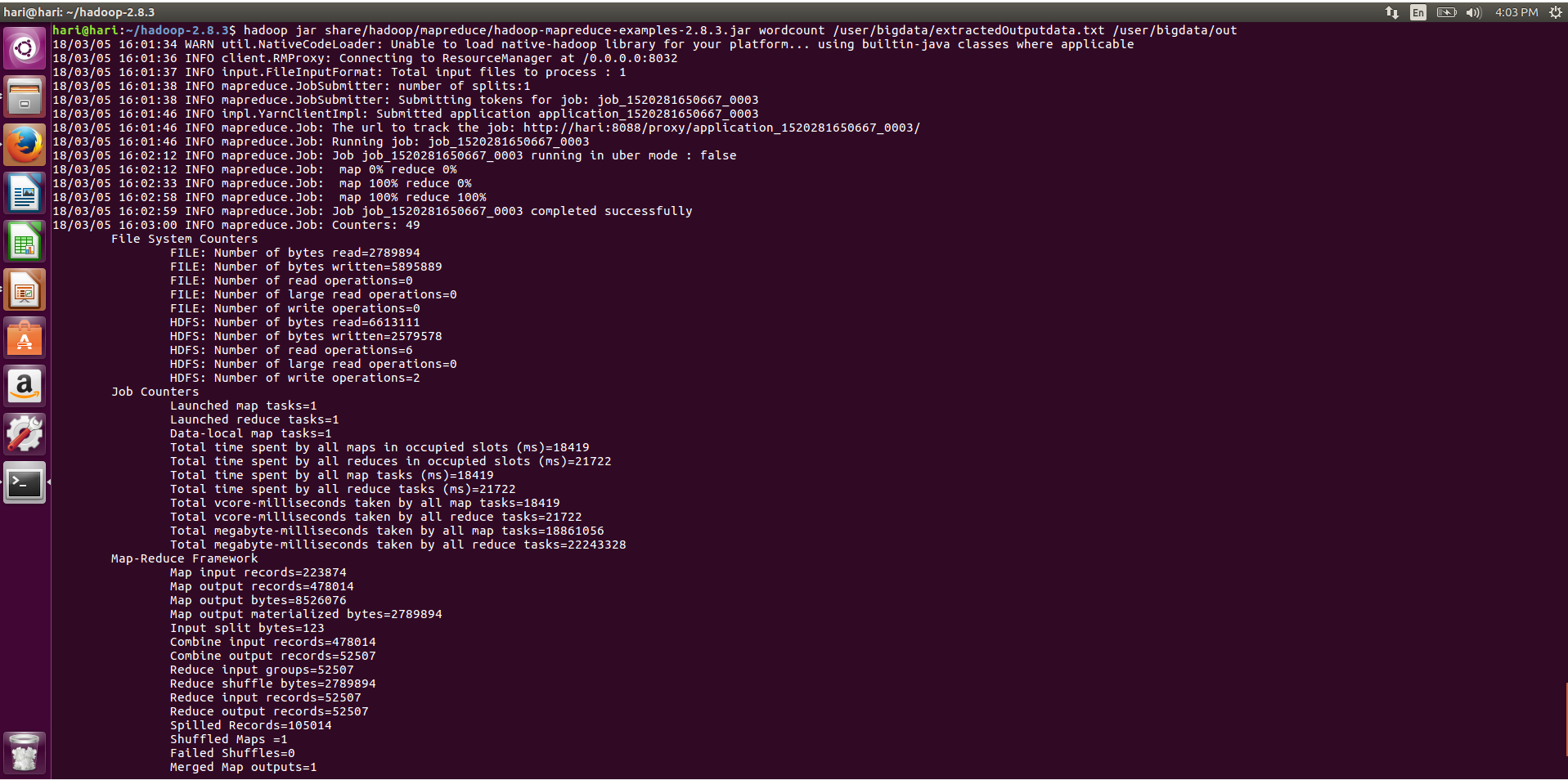




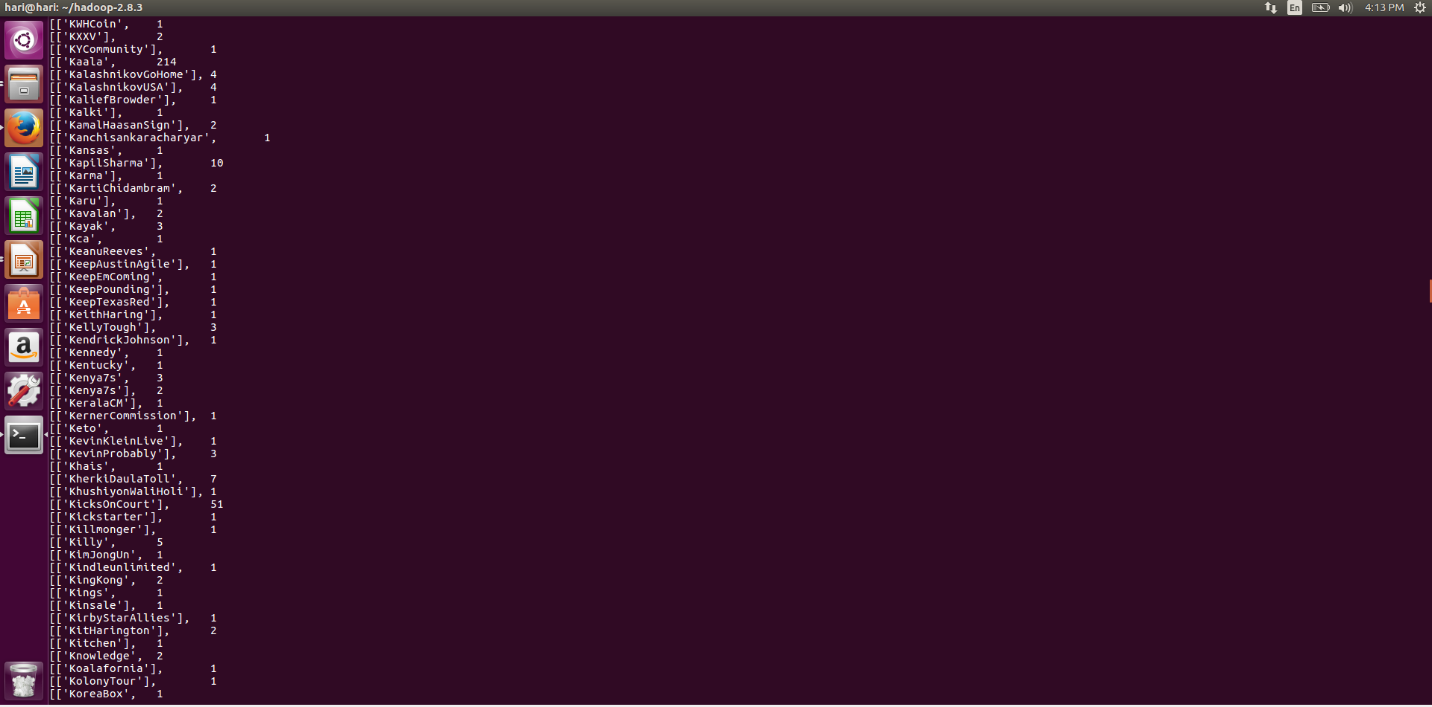
**Performing Word count:**

We have run the word count commands presented in the screenshots and obtained respective output.

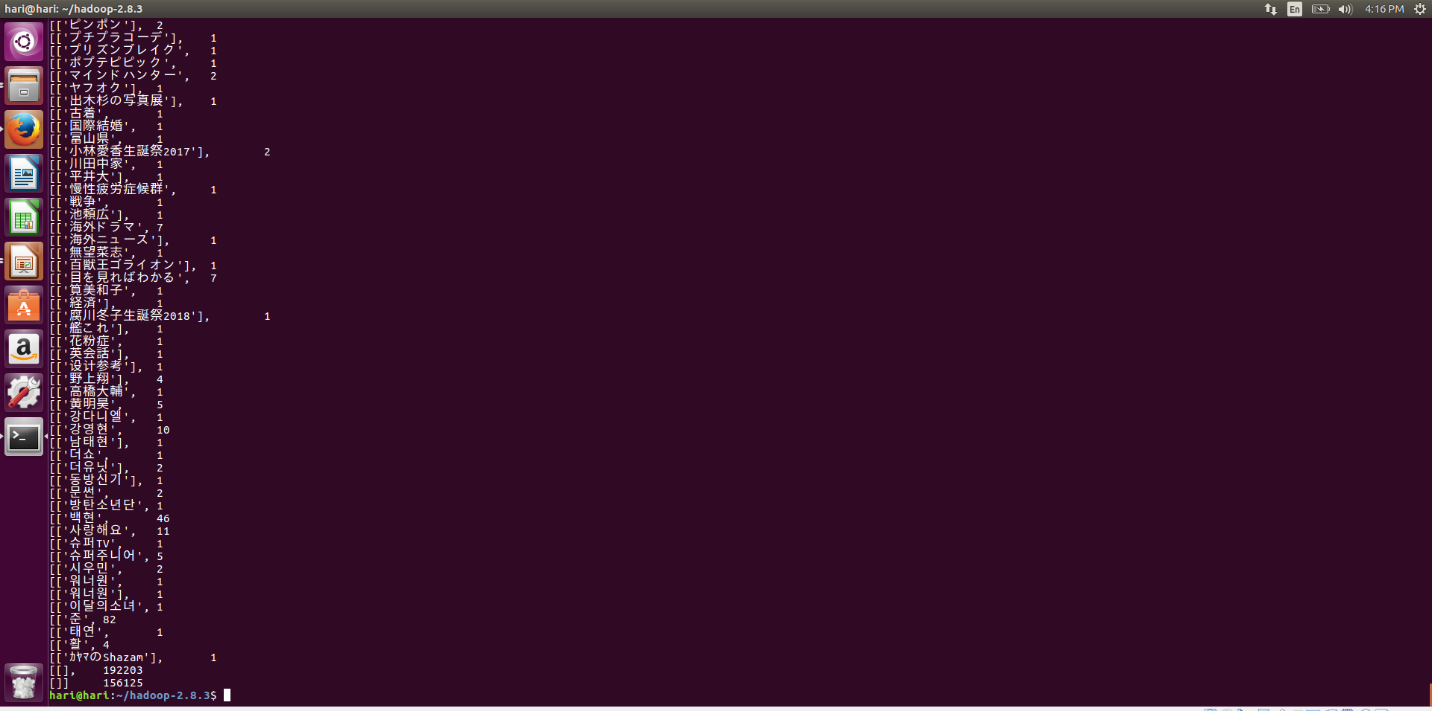
Please find the screenshots below:



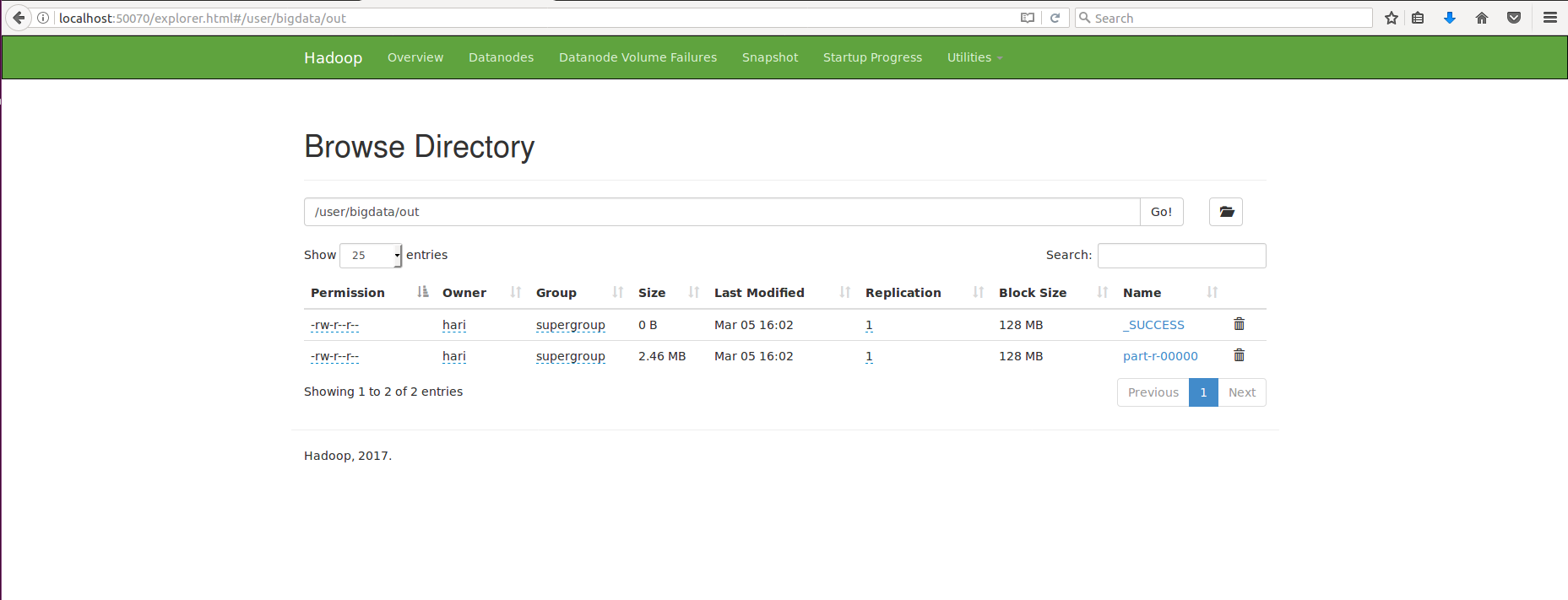




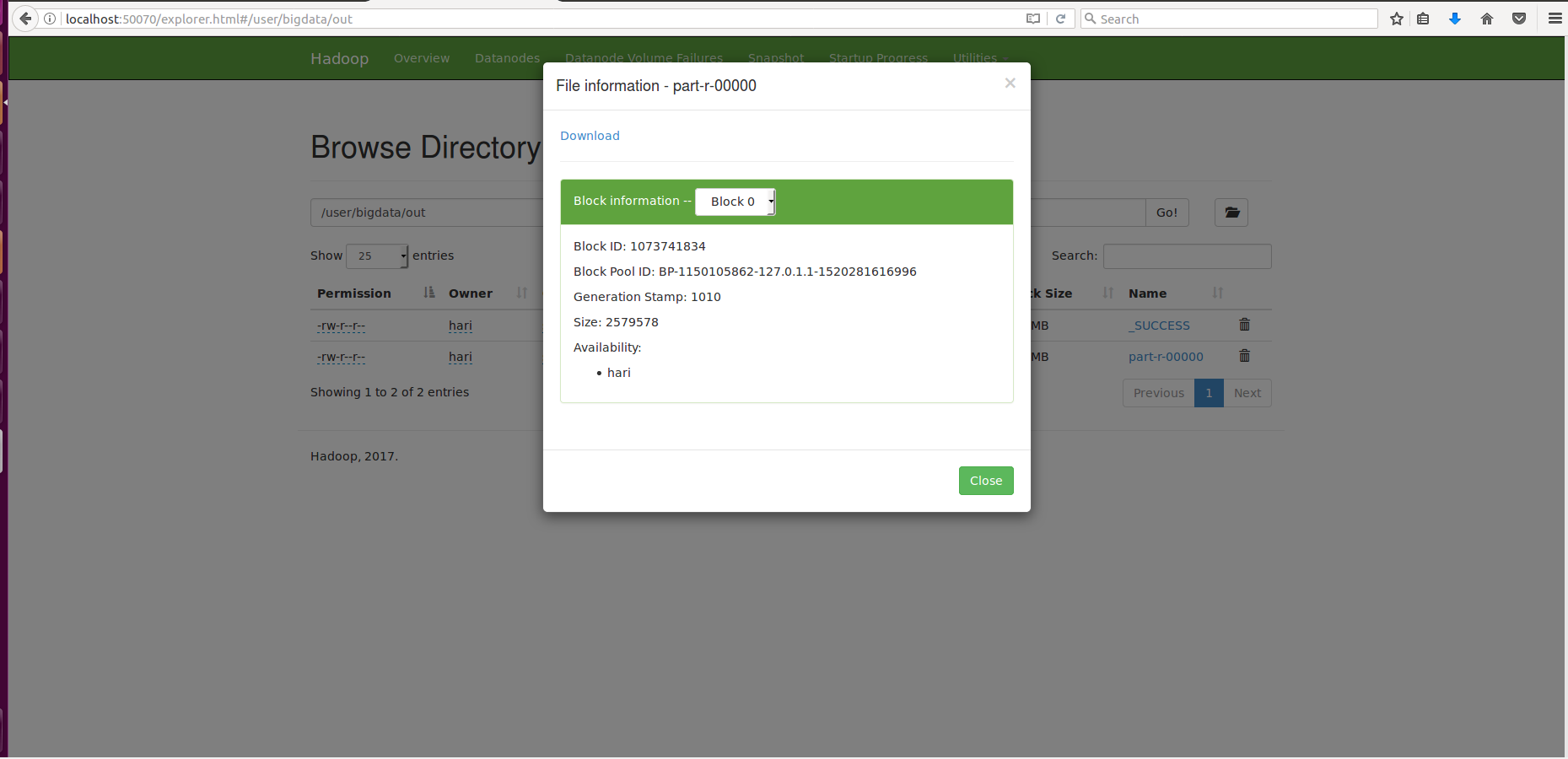




Logfile for the wordcount example execution on Hadoop:



Downloaded the output file from the directory.

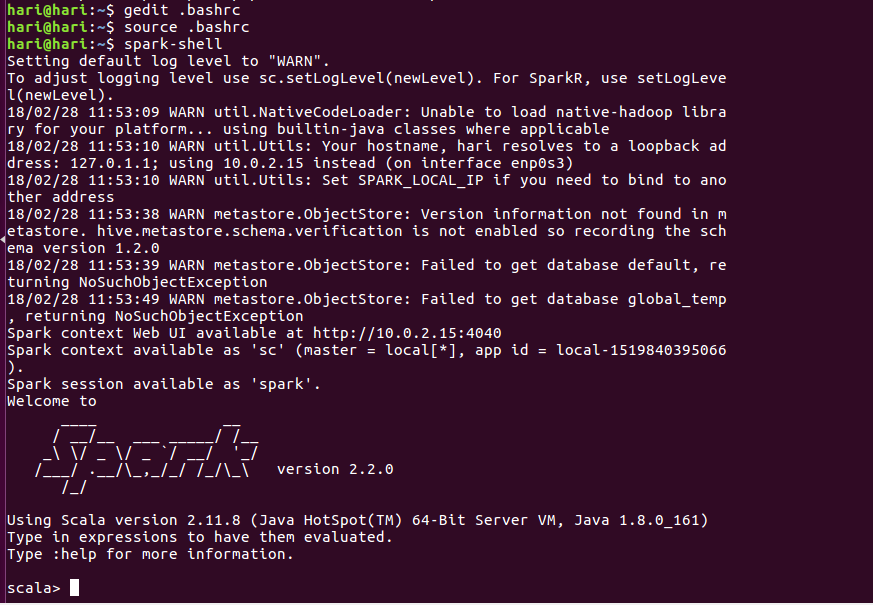


Output file output file naming **part-r-00000**

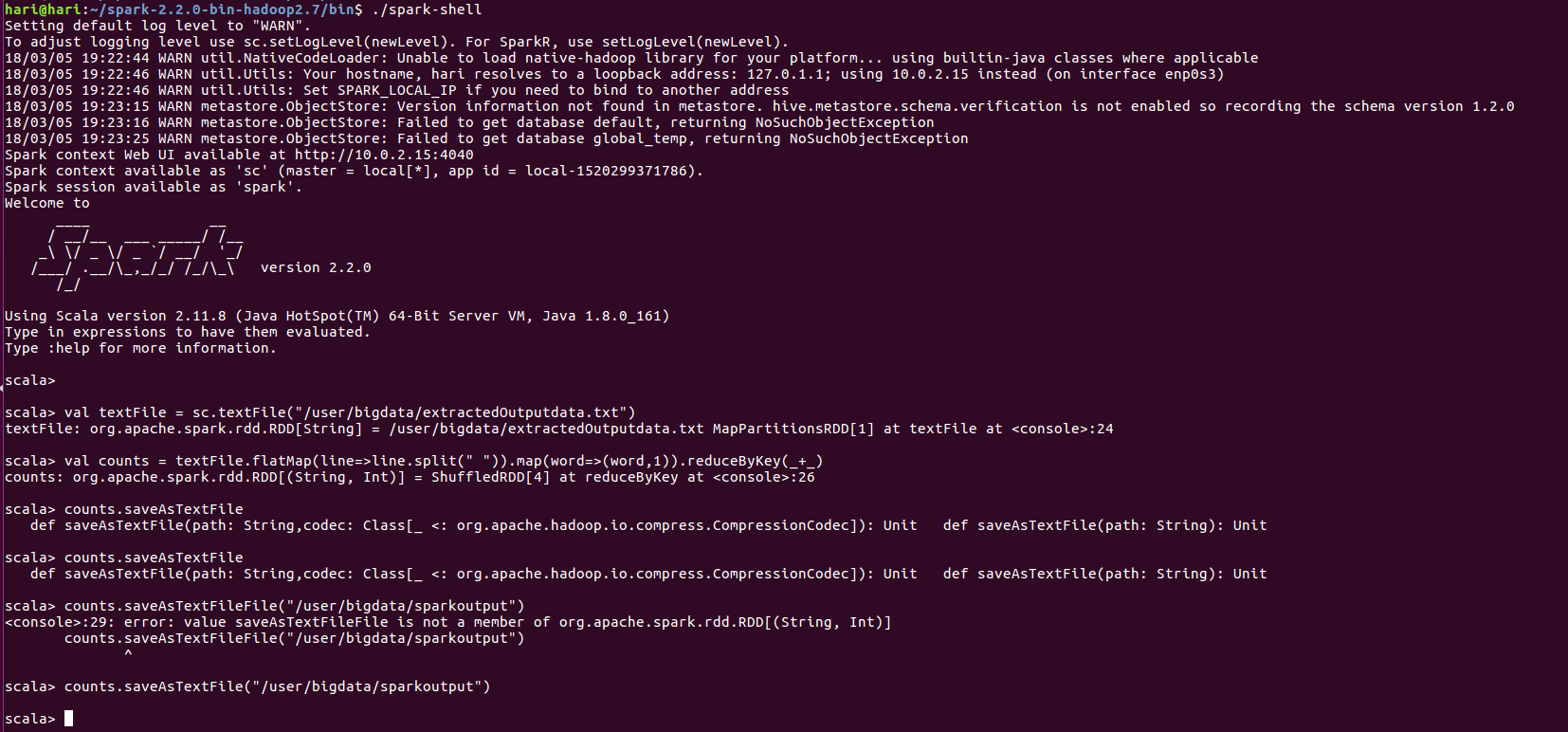
Please find the output file in the source code provided in the link above

**Performed Word count example on Apache Spark:**

Getting started with spark..



**Performed word count using spark:**



Collected the data output from the directory.

Collected the output and logfiles from Hadoop.

Each and every source code and log files are available in github repository mentioned below: