Daftar Isi

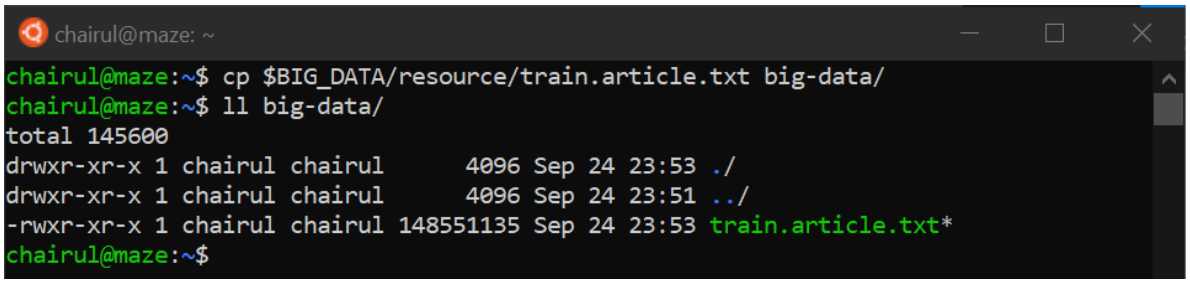
Laporan HDFS dan MapReduce

- Chairul Imam 1608107010040
- Nani Fiddini 1508107010060

Unduh file teks

File teks yang akan digunakan dapat diunduh pada URL berikut train.article.txt. Setelah diunduh, duplikasi file tersebut ke WSL Ubuntu dengan perintah

```
$ cp $BIG_DATA/resource/train.article.txt big-data/
```

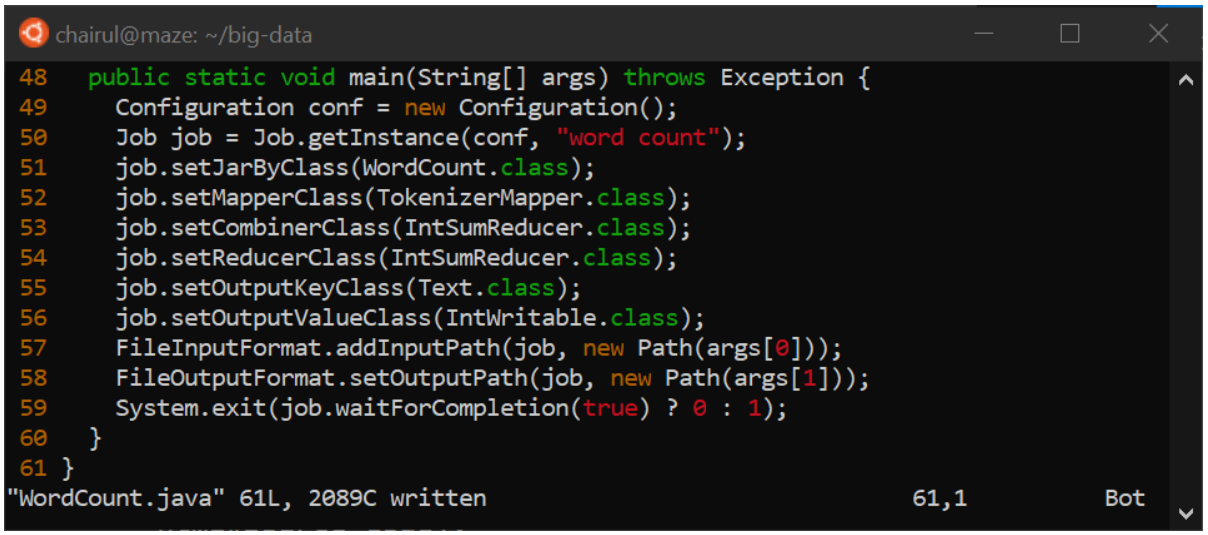


> **Info**
>
> command **cp** pada contoh diatas menerima 2 parameter, **SOURCE** dan **DEST.** Sesuaikan dengan lokasi file dan folder pada sistem anda.

Salin source code

Salin dan simpan *source code* WordCount v1.0 ke file `WordCount.java`. *Source code* dapat diakses pada URL berikut WordCount v1.0.

```
$ vi big-data/WordCount.java
```



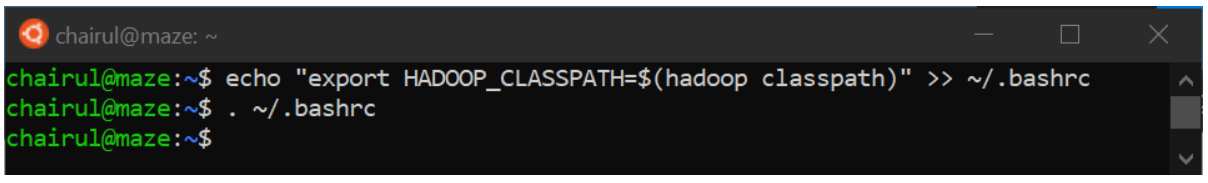
Tambah environment variable HADOOP_CLASSPATH

Pastikan bahwa variabel lain seperti **JAVA_HOME** telah ditambahkan pada file **.bashrc**. Jalankan perintah berikut untuk menambahkan variabel **$HADOOP_CLASSPATH**

```
$ echo "export HADOOP_CLASSPATH=$(hadoop classpath)" >> ~/.bashrc
```

Kemudian *reload environment* dengan perintah

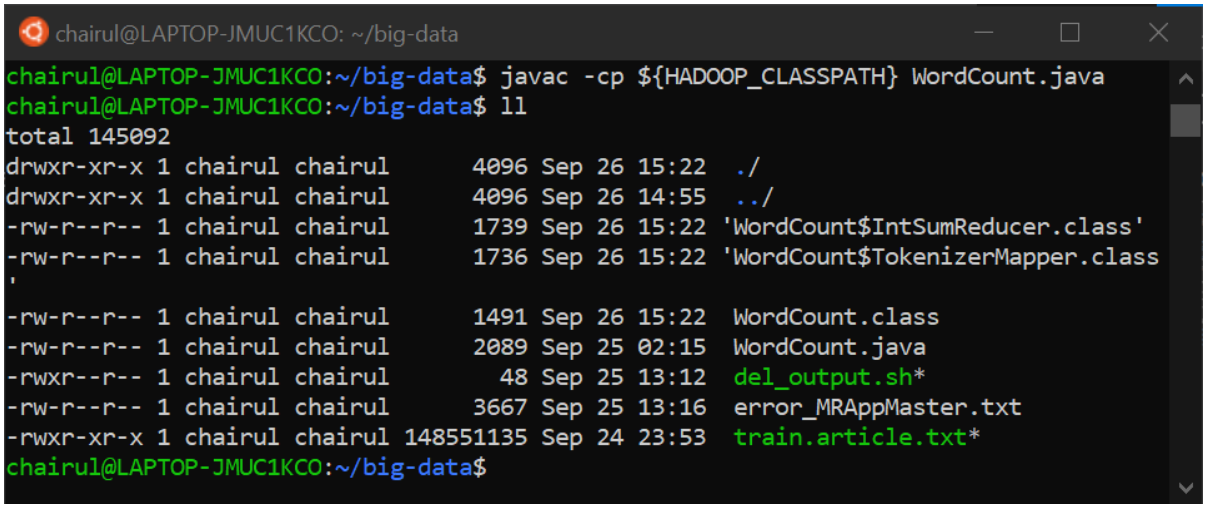
```
$ source ~/.bashrc

atau

$ . ~/.bashrc
```



Compile WordCount.Java dan buat Jar

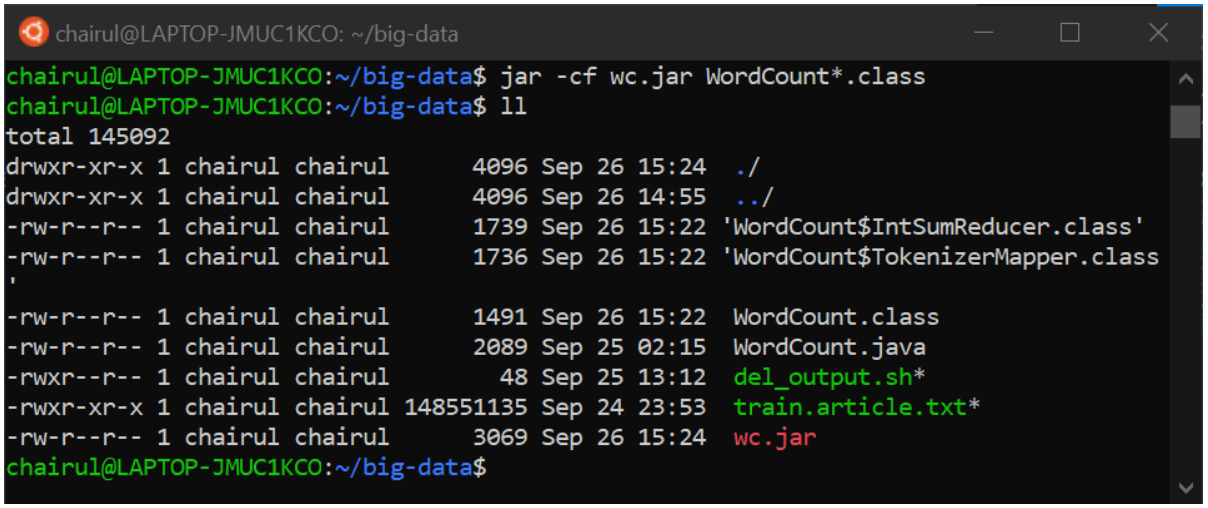
File **WordCount.java** dapat dikompilasi dengan perintah

```
$ javac -cp ${HADOOP_CLASSPATH} WordCount.java -d WordCount/
```



Kemudian dari file-file **.class** tersebut akan dibuat sebuah file **Jar**. Lakukan dengan perintah

```
$ jar -cf wc.jar WordCount*.class
```

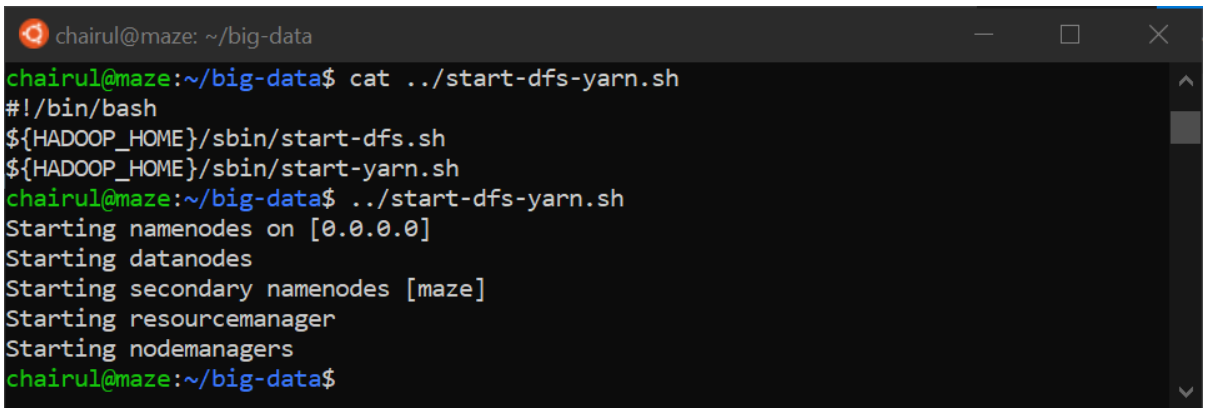


> **Info**
>
> Opsi **-cf** pada perintah **jar** adalah opsi untuk membuat file jar dengan nama yang dilewatkan pada parameter/argumen.

Jalankan DFS dan YARN

Setelah selesai membuat file jar, dilanjutkan dengan menjalankan DFS dan YARN dengan perintah

```
$HADOOP_HOME/sbin/start-dfs.sh
$HADOOP_HOME/sbin/start-yarn.sh
```



Tip

Lelah mengetik perintah yang sama berulang kali ?

Coba pakai **Bash Script**.

Pada contoh diatas saya menggabungkan dua perintah kedalam 1 file **.sh**. Untuk dapat mengeksekusi file tersebut lakukan

```
$ chmod 744 start-dfs-yarn.sh
```

Kemudian file dapat dieksekusi dengan

```
$ ./start-dfs-yarn.sh
```

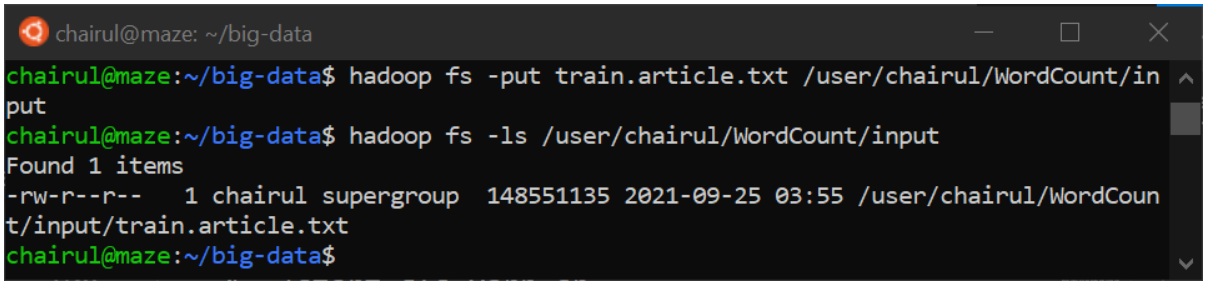
Membuat folder input dan memindahkan file teks

Folder input ini digunakan sebagai folder penyimpanan file teks **train.article.txt**. Lakukan perintah berikut untuk membuat folder input

```
$ hadoop fs -mkdir -p /user/chairul/WordCount/input
```

Kemudian lanjutkan dengan memindahkan file **train.article.txt** ke folder **/input**

```
$ hadoop fs -put train.article.txt /user/chairul/WordCount/input
```

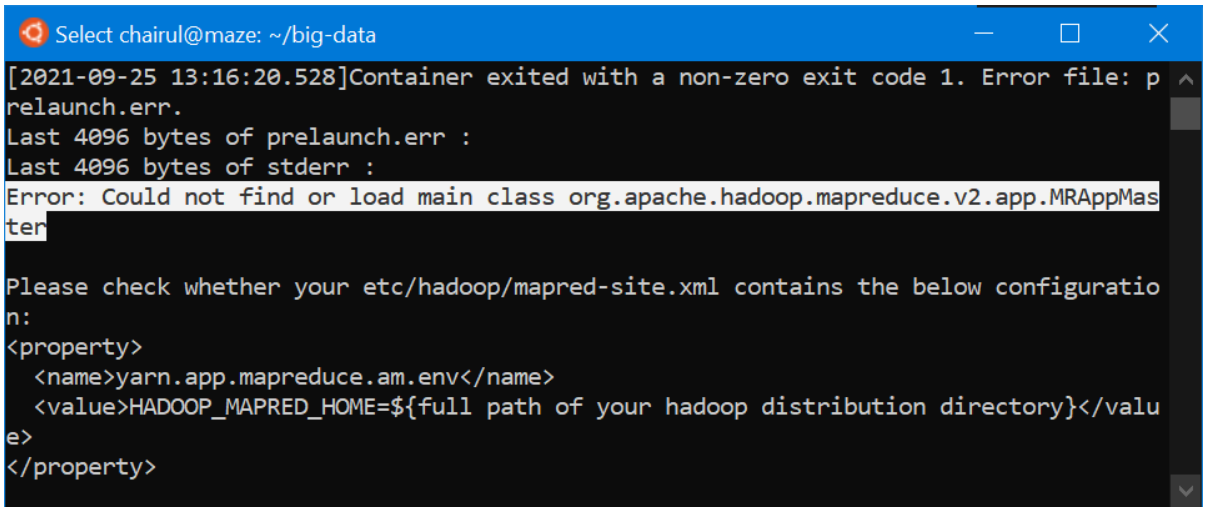


Jalankan Jar

Kemudian eksekusi perintah berikut untuk menjalankan program MapReduce yang menghitung frekuensi kata,

```
$ hadoop jar wc.jar WordCount /user/chairul/WordCount/input
  /user/chairul/WordCount/output
```

Sesaat setelah dijalankan, terminal menampilkan **Error: Could not find or load main class org.apache.hadoop.mapreduce.v2.app.MRAppMaster**.

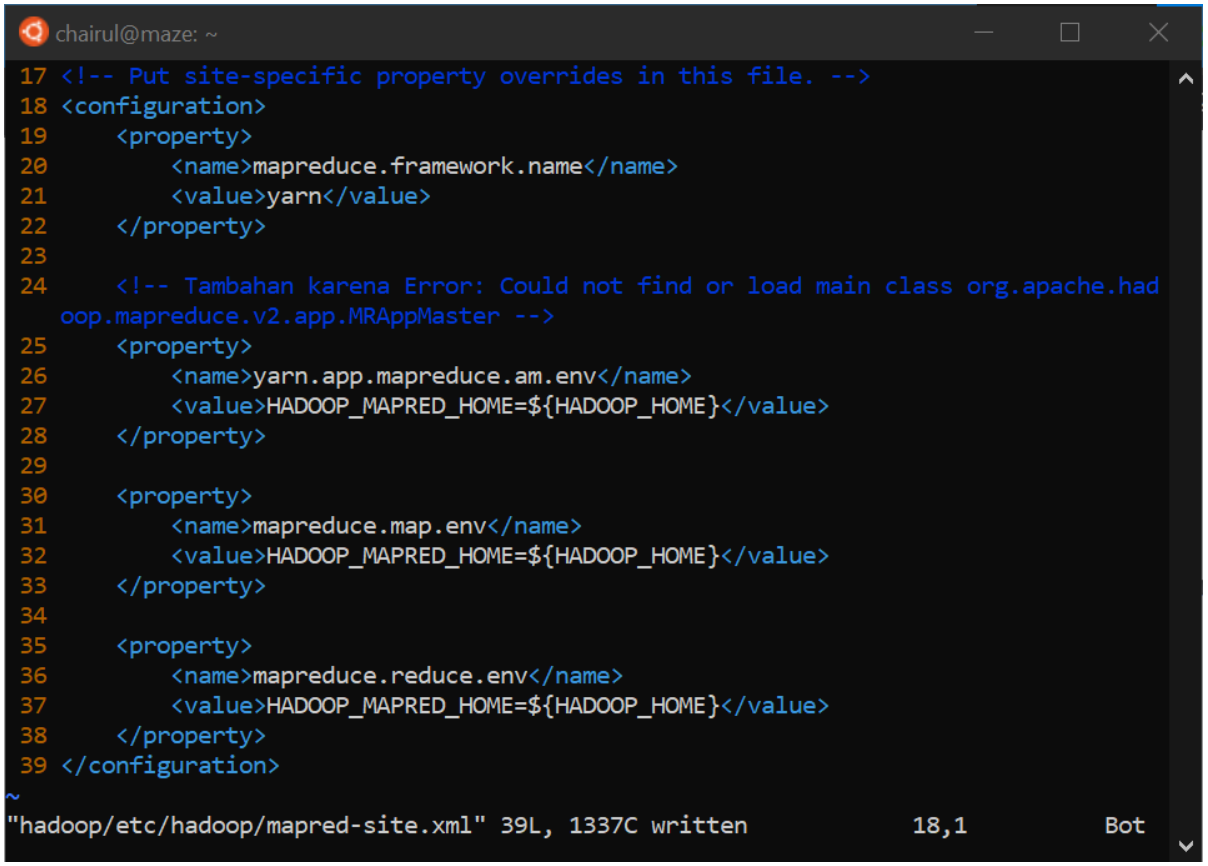


Troubleshoot Error MRAppMaster

Error ini terjadi karena *main class* **org.apache.hadoop.mapreduce.v2.app.MRAppMaster** tidak ditemukan atau dimuat. Detail error dapat diakses pada Error MRAppMaster. Untuk memperbaiki error ini, tambahkan beberapa konfigurasi berikut pada file **mapred-site.xml**.

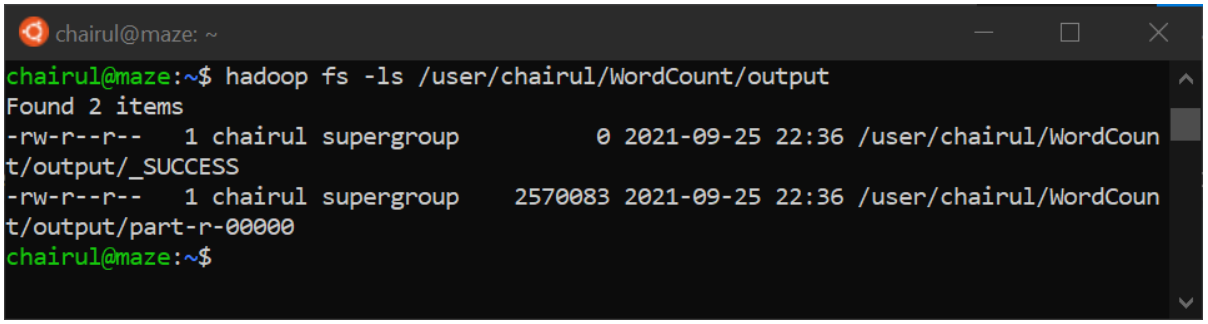
```
<!-- Tambahan karena Error: Could not find or load main class org.apache.hac
<property>
    <name>yarn.app.mapreduce.am.env</name>
    <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
</property>
<property>
    <name>mapreduce.map.env</name>
    <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
</property>
<property>
    <name>mapreduce.reduce.env</name>
    <value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
</property>
```

Sehingga file **mapred-site.xml** terlihat seperti berikut



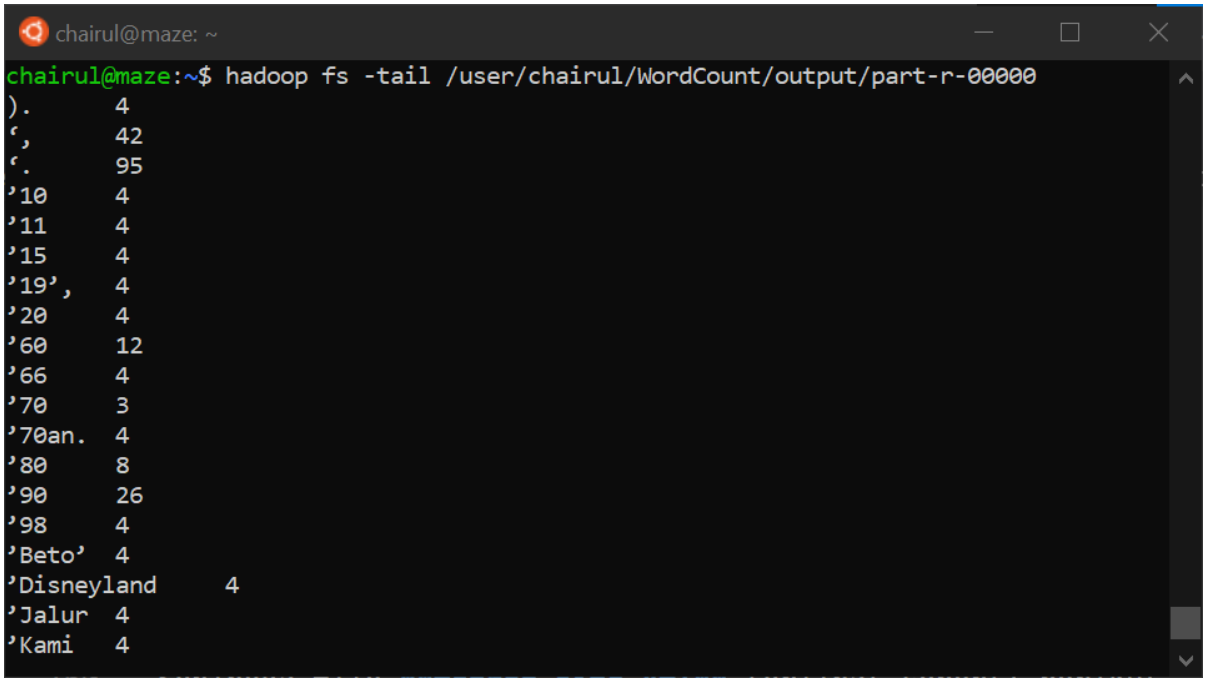
Kemudian coba jalankan kembali program MapReduce tersebut. Program berhasil berjalan dan hasil frekuensi kata dapat diakses pada folder **/output**.

```
$ hadoop fs -ls /user/chairul/WordCount/output
```



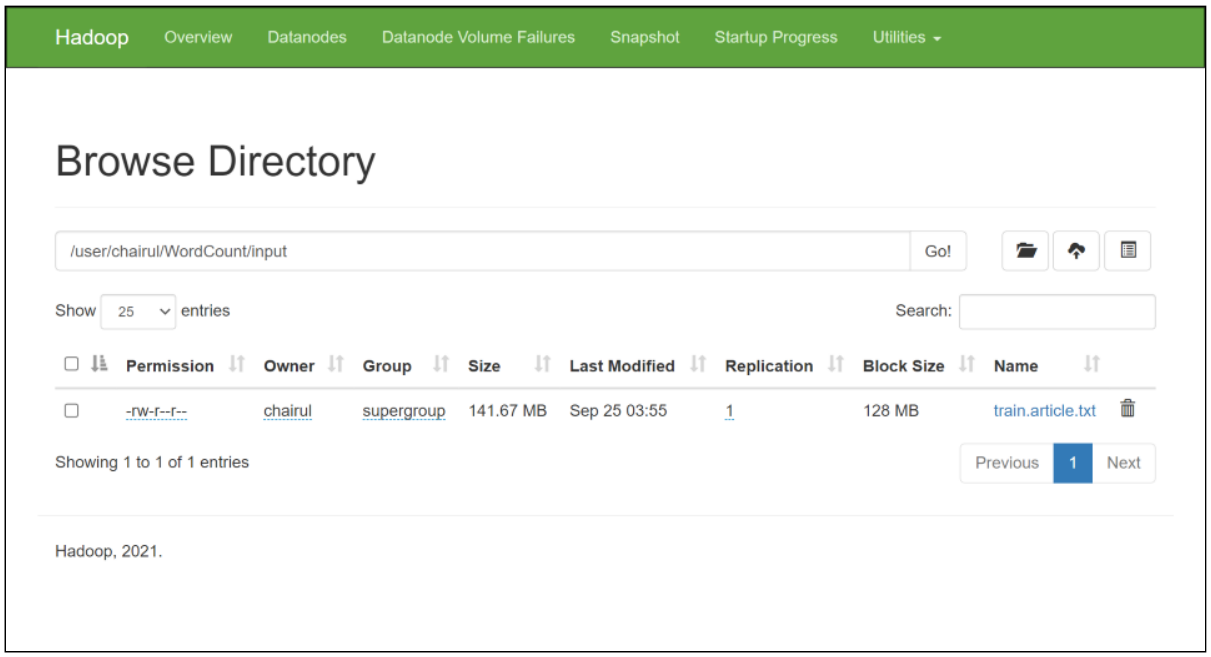
Sedangkan untuk menampilkan beberapa frekuensi kata terakhir dapat dilakukan dengan

```
$ hadoop fs -tail /user/chairul/WordCount/output/part-r-00000
```



Tampilan dashboard Hadoop

Folder /input dapat diakses dengan menginput **/user/chairul/Wordcount/input** pada form kemudian klik **Go** atau **Enter**.



Sedangkan untuk mengakes folder /output, ganti nilai form menjadi **/user/chairul/Wordcount/output**.