CS-GY 6513 Big Data Homework 1

Sashank RM

sr6890

HDFS Directory Creation & Listings:

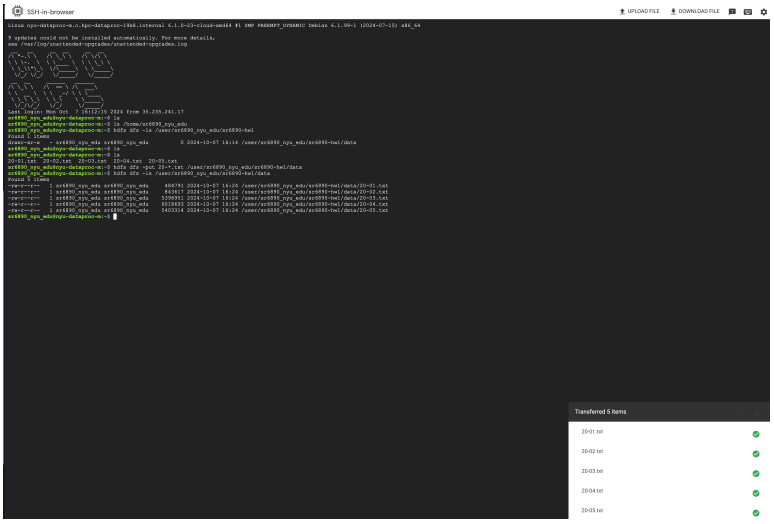
hdfs dfs -mkdir /user/sr6890_nyu_edu/sr6890-hw1

hdfs dfs -mkdir /user/sr6890_nyu_edu/sr6890-hw1/data

```
24  hdfs dfs -mkdir /user/sr6890_nyu_edu/sr6890-hw1
25  hdfs dfs -mkdir /user/sr6890_nyu_edu/sr6890-hw1/data
```

Upload Input Files to HDFS:

hdfs dfs -put 20-*.txt /user/sr6890_nyu_edu/sr6890-hw1/data



Verify the uploaded files:

```
sr6890_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls /user/sr6890_nyu_edu/sr6890-hw1/data
Found 5 items
-rw-r--r--   1 sr6890_nyu_edu sr6890_nyu_edu     484791 2024-10-07 16:24 /user/sr6890_nyu_edu/sr6890-hw1/data/20-01.txt
-rw-r--r--   1 sr6890_nyu_edu sr6890_nyu_edu     843617 2024-10-07 16:24 /user/sr6890_nyu_edu/sr6890-hw1/data/20-02.txt
-rw-r--r--   1 sr6890_nyu_edu sr6890_nyu_edu    5398951 2024-10-07 16:24 /user/sr6890_nyu_edu/sr6890-hw1/data/20-03.txt
-rw-r--r--   1 sr6890_nyu_edu sr6890_nyu_edu    6018693 2024-10-07 16:24 /user/sr6890_nyu_edu/sr6890-hw1/data/20-04.txt
-rw-r--r--   1 sr6890_nyu_edu sr6890_nyu_edu    5403314 2024-10-07 16:24 /user/sr6890_nyu_edu/sr6890-hw1/data/20-05.txt
```

SSH-in-browser UPLOAD FILE DOWNL

```
sr6890_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls /user/sr6890_nyu_edu/
hdfs dfs -ls /user/sr6890_nyu_edu/sr6890-hw1/
Found 1 items
drwxr-xr-x   - sr6890_nyu_edu sr6890_nyu_edu          0 2024-10-08 04:38 /user/sr6890_nyu_edu/sr6890-hw1
Found 5 items
drwxr-xr-x   - sr6890_nyu_edu sr6890_nyu_edu          0 2024-10-07 16:24 /user/sr6890_nyu_edu/sr6890-hw1/data
drwxr-xr-x   - sr6890_nyu_edu sr6890_nyu_edu          0 2024-10-08 04:38 /user/sr6890_nyu_edu/sr6890-hw1/output-id-assign
drwxr-xr-x   - sr6890_nyu_edu sr6890_nyu_edu          0 2024-10-08 04:18 /user/sr6890_nyu_edu/sr6890-hw1/output-job1
drwxr-xr-x   - sr6890_nyu_edu sr6890_nyu_edu          0 2024-10-08 04:19 /user/sr6890_nyu_edu/sr6890-hw1/output-job2
drwxr-xr-x   - sr6890_nyu_edu sr6890_nyu_edu          0 2024-10-08 04:29 /user/sr6890_nyu_edu/sr6890-hw1/output-job3
sr6890_nyu_edu@nyu-dataproc-m:~$
```

Run the First MapReduce Job (Word Count):

a) Part 2.1

cd /home/sr6890_nyu_edu

hadoop com.sun.tools.javac.Main WordCount.java Top10Words.java

jar cf wordcount.jar *.class

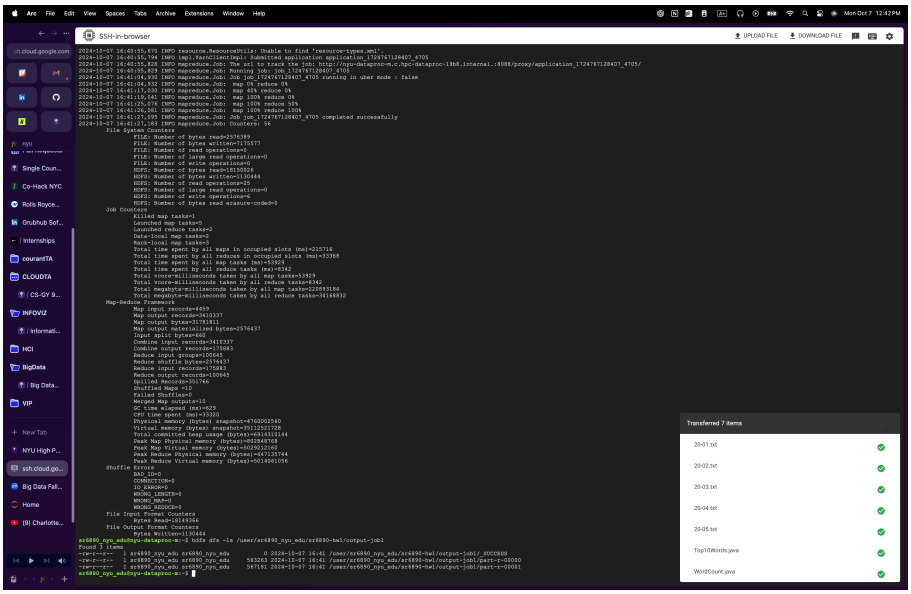
hadoop jar wordcount.jar WordCount /user/sr6890_nyu_edu/sr6890-hw1/data /user/sr6890_nyu_edu/sr6890-hw1/output-job1

hdfs dfs -ls /user/sr6890_nyu_edu/sr6890-hw1/output-job1

hadoop jar wordcount.jar Top10Words /user/sr6890_nyu_edu/sr6890-hw1/output-job1 /user/sr6890_nyu_edu/sr6890-hw1/output-job2

hdfs dfs -cat /user/sr6890_nyu_edu/sr6890-hw1/output-job2/part-r-00000

```
sr6890_nyu_edu@nyu-dataproc-m:~$ ls
20-01.txt  20-02.txt  20-03.txt  20-04.txt  20-05.txt
sr6890_nyu_edu@nyu-dataproc-m:~$ ls
20-01.txt  20-02.txt  20-03.txt  20-04.txt  20-05.txt
sr6890_nyu_edu@nyu-dataproc-m:~$ ls
20-01.txt  20-02.txt  20-03.txt  20-04.txt  20-05.txt
sr6890_nyu_edu@nyu-dataproc-m:~$ ls
20-01.txt  20-02.txt  20-03.txt  20-04.txt  20-05.txt  Top10Words.java  WordCount.java
sr6890_nyu_edu@nyu-dataproc-m:~$ hadoop com.sun.tools.javac.Main WordCount.java Top10Words.java
sr6890_nyu_edu@nyu-dataproc-m:~$ jar cf wordcount.jar *.class
sr6890_nyu_edu@nyu-dataproc-m:~$
```



```
sr6890_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -cat /user/sr6890_nyu_edu/sr6890-hw1/output-job2/part-r-00000
a       48776
in      52034
and     70565
of      74910
<p>     76036
to      87437
.       119369
,       136451
@       141782
the     142860
sr6890_nyu_edu@nyu-dataproc-m:~$
```

b) Part 2.2 (extra credit)