Understanding MapReduce

About this Lab

Objective: To understand how MapReduce works.

During this Lab: Watch as your instructor performs the following steps.

Related lesson: The MapReduce Framework

Lab Files `~/data`

Steps

<!–STEP–>

<h4>1. Put the File into HDFS</h4>

Use more to look at a file named `constitution.txt`. Press `q` to exit when finished.

```
more constitution.txt
```

> *Note: if you don't have the file local, you can get it with a `wget http://www.usconstitution.net/const.txt` command.*

Put the file into HDFS:

```
hdfs dfs -put constitution.txt
```

And to check it:

```
hdfs dfs -ls -R
```

<!–STEP–>

<h4>2. Run the WordCount Job</h4>

The following command runs a wordcount job on the `constitution.txt` and writes the output to `wordcount/output`:

```
yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar wordcount constitution.txt wordcount
```

```
_output
```

> *Note that a MapReduce job gets submitted to the cluster.* ***Wait for the job to complete****. If there is a problem with memory allocation (or other things) see your instructor.*

<!–STEP–>

<h4>3. View the Results</h4>

View the contents of the `wordcount_output` folder:

```
hdfs dfs -ls wordcount_output
```

You should get something like this:

```
Found 2 items
-rw-r--r-- 1    root    root    0           wordcount_out
put/_SUCCESS
-rw-r--r-- 1    root    root    17049       wordcount_out
put/part-r-00000
```


<h4>4. Answer a few questions</h4>

Why is there one `part-r` file in this directory?

> ***Answer:*** *The job only used one reducer.*

What does the "r" in the filename stand for?

> ***Answer:*** *The "r" stands for "reducer."*

View the contents of `part-r-00000`:

```
hdfs dfs -cat wordcount_output/part-r-00000
```

Why are the words sorted alphabetically?

> ***Answer:*** *The key in this MapReduce job is the word, and keys are sorted during the shuffle/sort phase.*

What was the key output by the WordCount reducer?

> ***Answer:*** *The reducer's output key was each word.*

What was the value output by the WordCount reducer?

> ***Answer:*** *The value output by the reducer was the sum of the 1's, which is the number of occurrences of the word in the document.*

Based on the output of the reducer, what do you think the mapper output would be as `key/value` pairs?

> ***Answer:*** *The mapper outputs each word as a key and the number 1 as each value.*

Result

You are finished! Not bad!