Hadoop Introduction - Part II

Goal: In this tutorial you will run a Hadoop MapReduce WordCount example job, which reads text files and counts how often words occur.

Before you begin

- Set up a single-node Hadoop cluster as described in Part I of the tutorial.

Download example input data

- Download the following ebooks from Project Gutenberg as text files in "**Plain Text UTF-8**" encoding:
    - Metamorphosis by Franz Kafka.
    - The Adventures of Sherlock Holmes by Arthur Conan Doyle.
    - The Adventures of Tom Sawyer by Mark Twain.

  Create a directory named "**gutenberg**" in your home directory and store these files.

```
[user@vcl-host ~]$ mkdir gutenberg
```

- Restart the Hadoop single-node cluster.

```
[user@vcl-host ~]$ cd $HOME/hadoop/hadoop-1.2.1
[user@vcl-host hadoop-1.2.1]$ bin/start-all.sh
```

- Copy the files to the HDFS.

```
[user@vcl-host hadoop-1.2.1]$ bin/hadoop dfs -copyFromLocal
$HOME/gutenberg /gutenberg
[user@vcl-host hadoop-1.2.1]$ bin/hadoop dfs -ls /gutenberg
Found 3 items
-rw-r--r--  1 user supergroup  594933 […] /gutenberg/pg1661.txt
-rw-r--r--  1 user supergroup  141419 […] /gutenberg/pg5200.txt
-rw-r--r--  1 user supergroup  421884 […] /gutenberg/pg74.txt
```

Run WordCount job

- Run the WordCount example job.

```
[user@vcl-host hadoop-1.2.1]$ bin/hadoop jar
hadoop*examples*.jar wordcount /gutenberg /gutenberg-output
[…] INFO input.FileInputFormat: Total input paths to process : 3
[…] INFO util.NativeCodeLoader: Loaded the native-hadoop library
[…] WARN snappy.LoadSnappy: Snappy native library not loaded
[…] INFO mapred.JobClient: Running job: job_xxx
[…] INFO mapred.JobClient:  map 0% reduce 0%
```

```
[…] INFO mapred.JobClient:  map 33% reduce 0%
[…] INFO mapred.JobClient:  map 66% reduce 0%
[…] INFO mapred.JobClient:  map 100% reduce 0%
[…] INFO mapred.JobClient:  map 100% reduce 100%
[…] INFO mapred.JobClient: Job complete: job_xxx
[…]
```

- Copy the output file of the WordCount example job from the HDFS to your local file system.

```
[user@vcl-host hadoop-1.2.1]$ bin/hadoop dfs -getmerge
/gutenberg-output $HOME/gutenberg-output
```

Open file "**gutenberg-output**" in "**~/gutenberg-output**." Output file must look as follows:

```
"'A             1
"'About         1
"'Absolute      1
"'Ah!'          2
"'Ah,           2
[…]
```

- Submit file "**gutenberg-output**" through Moodle (rename the file "**hadoop-single-node-user**," where **user** is your Unity ID).
- Stop your single-node Hadoop cluster.

```
[user@vcl-host hadoop-1.2.1]$ bin/stop-all.sh
stopping jobtracker
152.xxx.xxx.xxx: stopping tasktracker
stopping namenode
152.xxx.xxx.xxx: stopping datanode
152.xxx.xxx.xxx: stopping secondarynamenode
```

References

- http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/