

Running MapReduce Programs On Single Node Hadoop Cluster - Word Count/Word Frequency

Expt No: 3

March 06, 2019

Author: Subalakshmi Shanthosi S (186001008)

Aim

Implementation of MapReduce in Hadoop single node cluster.

Description

- Apache Hadoop
 - Large Scale, Open Source Software Framework.
 - Supports Three Projects:
 - * Hadoop Common.
 - * HDFS : Hadoop Distributed File System.
 - * MapReduce.
- Hadoop MapReduce
 - Hadoop Programming Model and Software Framework.
 - Computational Processing:
 - * Unstructured Data : File system
 - * Structured Data : Database
 - MapReduce Layer has job and task tracker nodes.
 - Cluster nodes:
 - * Single JobTracker per master.
 - * Single TaskTracker per slave.
 - Fundamental Steps:
 - * Map Step:
 - Master node slices problem input into several subproblems input.
 - Distributes data slices to worker nodes.
 - Worker nodes processes and hands over the control to master.
 - * Reduce Step:
 - Master node takes the answers to the sub problems and combines them in a predefined way to get the output/answer to original problem.

Software's Used

- Ubuntu 16.04 LTS
- Hadoop 1.0.3

Hadoop MapReduce Architecture

- MapReduce Architecture:

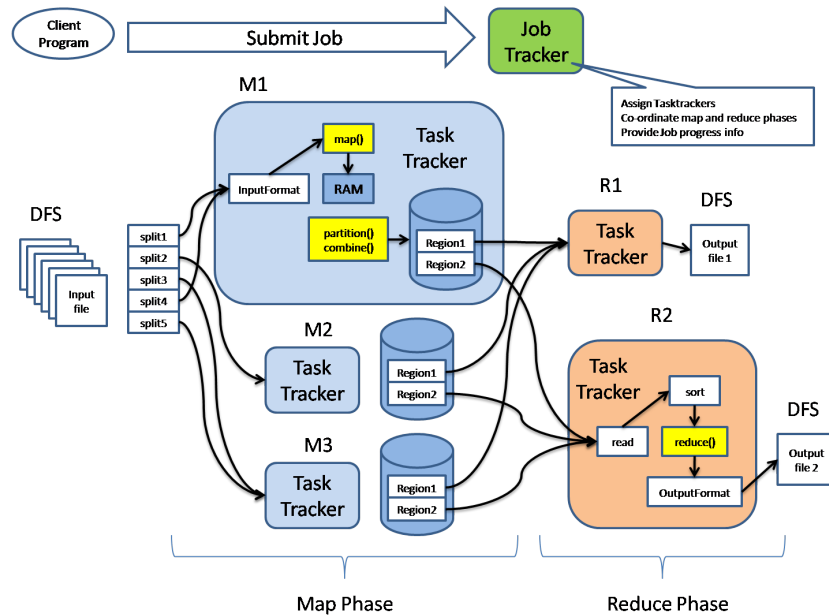


Figure 1: Hadoop MapReduce Architecture Diagram.

Procedure

1. Launch Ubuntu 16.04 LTS in virtual environment.
2. Login to the OS with sudo permission and install the following packages using apt-get command
 - openssh-server
 - openssh-client
 - java jdk 8
 - javac compiler
 - hadoop 1.0.3

3. Install and configure appropriate environment variables for Hadoop 1.0.3.
4. Start Hadoop DFS by invoking script start-all.sh in bin directory.
5. Examine the running Hadoop Job Process using jps command.
6. Download and place the input files in appropriate directory.
7. Copy the input Files from Local File System to HDFS using command attribute CopyFromLocal of dfs command.
8. Run the MapReduce Job.
9. Retrieve the Job Result from HDFS.
10. View stats from Web Interface for the following information listed below.
 - JobTracker Web Interface - <http://localhost:50030/>
 - TaskTracker Web Interface - <http://localhost:50060/>
 - NameNode Web Interface - <http://localhost:50070/>

Output

```
hduser@client-VirtualBox: /usr/local/hadoop/bin$ ./start-all.sh
Warning: $HADOOP_HOME is deprecated.

starting namenode, logging to /usr/local/hadoop/libexec/../logs/hadoop-hduser-na
menode-client-VirtualBox.out
localhost: starting datanode, logging to /usr/local/hadoop/libexec/../logs/hadoo
p-hduser-datanode-client-VirtualBox.out
localhost: starting secondarynamenode, logging to /usr/local/hadoop/libexec/../l
ogs/hadoop-hduser-secondarynamenode-client-VirtualBox.out
starting jobtracker, logging to /usr/local/hadoop/libexec/../logs/hadoop-hduser-
jobtracker-client-VirtualBox.out
localhost: starting tasktracker, logging to /usr/local/hadoop/libexec/../logs/ha
doo-hduser-tasktracker-client-VirtualBox.out
```

Figure 2: Starting Hadoop DFS.

```
hduser@client-VirtualBox: /usr/local/hadoop/bin$ jps
2960 Jps
2548 DataNode
2677 SecondaryNameNode
2758 JobTracker
2893 TaskTracker
hduser@client-VirtualBox: /usr/local/hadoop/bin$
```

Figure 3: Examining Running Hadoop Process.

```
hduser@client-VirtualBox: /tmp/gutenberg$ ls -l /tmp/gutenberg/
total 3608
-rw-r--r-- 1 hduser hadoop 1586396 Apr 12 09:19 4300-0.txt
-rw-r--r-- 1 hduser hadoop 1428841 Apr 12 09:19 5000-8.txt
-rw-r--r-- 1 hduser hadoop 674570 Apr 12 09:19 pg20417.txt
hduser@client-VirtualBox: /tmp/gutenberg$
```

Figure 4: Placing the inputFiles in appropriate location.

```
hduser@client-VirtualBox: /usr/local/hadoop/bin
hduser@client-VirtualBox: /usr/local/hadoop/bin$ ./hadoop dfs -copyFromLocal /tmp
/gutenberg /user/hduser/gutenberg/
Warning: SHADOOP_HOME is deprecated.

hduser@client-VirtualBox: /usr/local/hadoop/bin$ ./hadoop dfs -ls /user/hduser/gu
tutenberg/
Warning: SHADOOP_HOME is deprecated.

Found 4 items
-rw-r--r-- 1 hduser supergroup 1586396 2019-04-12 12:12 /user/hduser/gutenb
erg/4380-0.txt
-rw-r--r-- 1 hduser supergroup 1428841 2019-04-12 12:12 /user/hduser/gutenb
erg/5080-8.txt
drwxr-xr-x - hduser supergroup 0 2019-04-12 12:13 /user/hduser/gutenb
erg/gutenberg
-rw-r--r-- 1 hduser supergroup 674570 2019-04-12 12:12 /user/hduser/gutenb
erg/pg28417.txt
hduser@client-VirtualBox: /usr/local/hadoop/bin$
```

Figure 5: Copy Files from Local to HDFS.

```
hduser@client-VirtualBox: /usr/local/hadoop$ bin/hadoop jar hadoop-examples-1.0.3
.jar wordcount /user/hduser/gutenberg /user/hduser/gutenberg-output
Warning: SHADOOP_HOME is deprecated.

19/04/12 09:41:42 INFO Input.FileInputFormat: Total input paths to process : 3
19/04/12 09:41:42 INFO util.NativeCodeLoader: Loaded the native-hadoop library
19/04/12 09:41:42 WARN snappy.LoadSnappy: Snappy native library not loaded
19/04/12 09:41:43 INFO mapred.JobClient: Running job: job_201904120938_0002
19/04/12 09:41:44 INFO mapred.JobClient: map 0% reduce 0%

19/04/12 09:42:15 INFO mapred.JobClient: map 66% reduce 0%
19/04/12 09:42:29 INFO mapred.JobClient: map 100% reduce 0%
19/04/12 09:42:42 INFO mapred.JobClient: map 100% reduce 100%
19/04/12 09:42:47 INFO mapred.JobClient: Job complete: job_201904120938_0002
19/04/12 09:42:47 INFO mapred.JobClient: Counters: 29
19/04/12 09:42:47 INFO mapred.JobClient: Map-Reduce Framework
19/04/12 09:42:47 INFO mapred.JobClient: Spilled Records=256111
19/04/12 09:42:47 INFO mapred.JobClient: Map output materialized bytes=14888
53
19/04/12 09:42:47 INFO mapred.JobClient: Reduce input records=102468
19/04/12 09:42:47 INFO mapred.JobClient: Virtual memory (bytes) snapshot=896
1723904
19/04/12 09:42:47 INFO mapred.JobClient: Map input records=78710
19/04/12 09:42:47 INFO mapred.JobClient: SPLIT_RAW_BYTES=358
```

Figure 6: Run MapReduce Task for Input Files.

```

hduser@client-VirtualBox:/usr/local/hadoop$ bin/hadoop dfs -ls /user/hduser
Warning: $HADOOP_HOME is deprecated.

Found 3 items
drwxr-xr-x - hduser supergroup      0 2019-04-12 09:40 /user/hduser/gutenb
urg
drwxr-xr-x - hduser supergroup      0 2019-04-12 09:42 /user/hduser/gutenb
urg-output

```

Figure 7: Retrieve Hadoop MapReduce Output from appropriate folder.

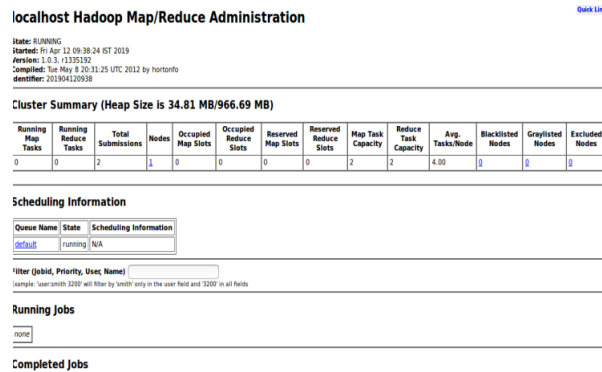


Figure 8: JobTracker of HDFS Web Interface.

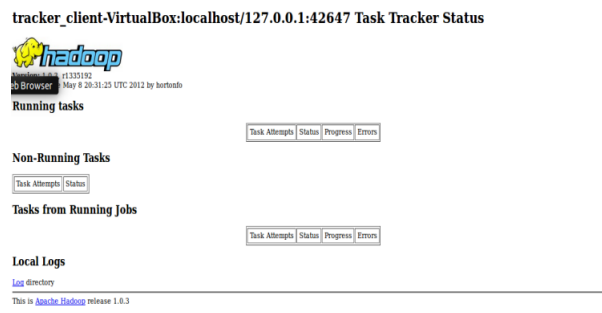


Figure 9: Task Tracker of HDFS Web Interface.

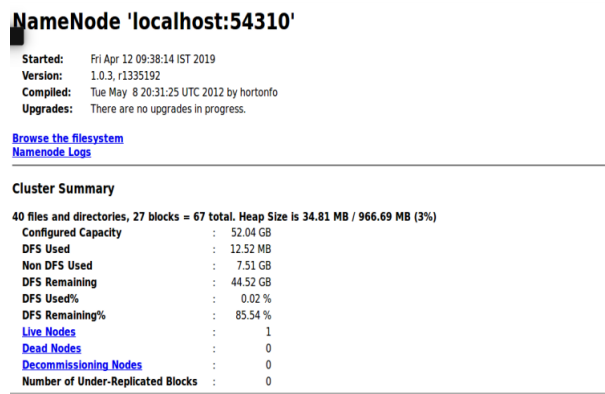


Figure 10: Web Interface of HDFS Namenode.

Result

Thus the hadoop MapReduce program for finding word count and frequency was successfully executed and its results are obtained for further processing.