Course: MMDS

Tutorial 01

Set up Hadoop

Prerequisites

- Operation system: Ubuntu 18.04 LTS (highly recommended)
- Chipset: x86 (optional to the expected version of Hadoop)
- If you do not have Ubuntu, consider virtual machines
    - VMWare Fusion (MacOS)
    - Oracle Virtual Box (Windows)
    - Docker (Windows)

Installation

Install Hadoop

1. Install Java 8 (highly recommended)

   ```
   sudo apt install openjdk-8-jre-headless
   sudo apt install openjdk-8-jdk-headless
   ```

2. Install ssh and pdsh

   ```
   sudo apt install ssh
   sudo apt install pdsh
   ```

3. Setup passphrase for ssh

   ```
   ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
   cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
   chmod 0600 ~/.ssh/authorized_keys
   ```
   # ensure the file `~/.ssh/authorized_keys` exists

   ```
   # check whether you can ssh to localhost
   ssh localhost
   ```

4. Configure rcmd to ssh as default

```
sudo nano /etc/pdsh/rcmd_default
# add “ssh” to the file
# save & quit

# sudo echo “ssh” > /etc/pdsh/rmcd_default
```

5. Download Hadoop 3.2.1 (highly recommended)
   https://hadoop.apache.org/release/3.2.1.html

```
cd Desktop
wget
https://archive.apache.org/dist/hadoop/common/hadoop-3.2.1/hado
op-3.2.1.tar.gz
tar -xvf hadoop-3.2.1.tar.gz
```

6. Declare JAVA_HOME for Hadoop

```
# cd to the extracted folder of Hadoop
nano etc/hadoop/hadoop-env.sh

# add this line to the end of the file
# check your own Java path if different
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
export PATH=${JAVA_HOME}/bin:${PATH}
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
# save & quit
```

7. Verify installation

```
# cd to the extracted folder of Hadoop
bin/hadoop
```

Set up Pseudo-Distributed Mode

Configuration

Edit these following files

etc/hadoop/core-site.xml

```
<configuration>
    <property>
```

```
            <name>fs.defaultFS</name>
            <value>hdfs://localhost:9000</value>
        </property>
    </configuration>
```

etc/hadoop/hdfs-site.xml

```
    <configuration>
        <property>
            <name>dfs.replication</name>
            <value>1</value>
        </property>
    </configuration>
```

Run a MapReduce job locally

1. Format the filesystem

   ```
   bin/hdfs namenode -format
   ```

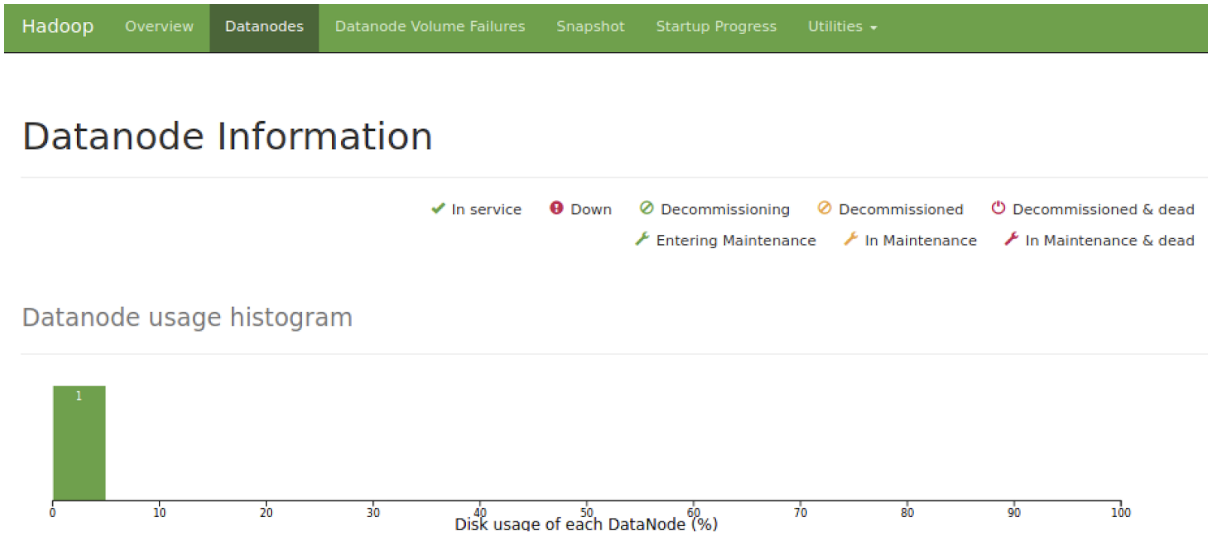
2. Start NameNode daemon and DataNode daemon

   ```
   sbin/start-dfs.sh
   ```
   # check log output in .../logs as needed

   # if you fail to start a datanode, then run sbin/stop-all.sh and sudo rm -rf /tmp/*

3. Browse the web interface for the NameNode; by default it is available at

   http://localhost:9870/



4. Make the HDFS directories required to execute MapReduce jobs:

```
bin/hdfs dfs -mkdir /user
bin/hdfs dfs -mkdir /user/<username>
```

5. Copy the input files into the distributed filesystem:

```
bin/hdfs dfs -mkdir input
bin/hdfs dfs -put etc/hadoop/*.xml input
```

```
ntan@ubuntu:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -put etc/hadoop/*.xml input
2021-05-05 13:24:27,376 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:24:28,078 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:24:28,118 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:24:28,556 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:24:28,589 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:24:29,031 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:24:29,081 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:24:29,114 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:24:29,147 INFO sasl.SaslDataTransferClient: SASL encryption trust che
ck: localHostTrusted = false, remoteHostTrusted = false
```

6. Run some of the examples provided:

```
bin/hadoop jar
share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar grep input
output 'dfs[a-z.]+'
```

ensure the corresponding jar file exists in the folder mapreduce/

```
ntan@ubuntu:~/Desktop/hadoop-3.2.1$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar grep input output 'dfs[a-z.]+'
2021-05-05 13:30:45,054 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-05-05 13:30:45,475 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ntan/.staging/job_1620196238613_0001
2021-05-05 13:30:45,603 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:30:45,774 INFO input.FileInputFormat: Total input files to process : 9
2021-05-05 13:30:45,840 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:30:46,284 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:30:46,699 INFO mapreduce.JobSubmitter: number of splits:9
2021-05-05 13:30:46,858 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-05 13:30:46,900 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1620196238613_0001
2021-05-05 13:30:46,900 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-05-05 13:30:47,175 INFO conf.Configuration: resource-types.xml not found
2021-05-05 13:30:47,175 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-05-05 13:30:47,546 INFO impl.YarnClientImpl: Submitted application application_1620196238613_0001
2021-05-05 13:30:47,646 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1620196238613_0001/
2021-05-05 13:30:47,647 INFO mapreduce.Job: Running job: job_1620196238613_0001
2021-05-05 13:30:55,872 INFO mapreduce.Job: Job job_1620196238613_0001 running in uber mode : false
2021-05-05 13:30:55,878 INFO mapreduce.Job:  map 0% reduce 0%
2021-05-05 13:31:15,196 INFO mapreduce.Job:  map 67% reduce 0%
```

Note: after setting up YARN, this step requires [Connecting to ResourceManager at /0.0.0.0:8032]

7. Examine the output files: Copy the output files from the distributed filesystem to the local filesystem and examine them:

   ```
   bin/hdfs dfs -get output output
   cat output/*
   ```

   or
   View the output files on the distributed filesystem:

   ```
   bin/hdfs dfs -cat output/*
   ```

8. When you're done, stop the daemons with:

   ```
   sbin/stop-dfs.sh
   ```

Execute job on YARN

The following instructions assume that 1. ~ 4. steps of the above instructions are already executed.

5. Configure parameters as follows:
   etc/hadoop/mapred-site.xml

   ```
   <configuration>
       <property>
           <name>mapreduce.framework.name</name>
           <value>yarn</value>
       </property>
       <property>
   ```

```
            <name>mapreduce.application.classpath</name>

<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOO
P_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>
    </property>
</configuration>
```

etc/hadoop/yarn-site.xml

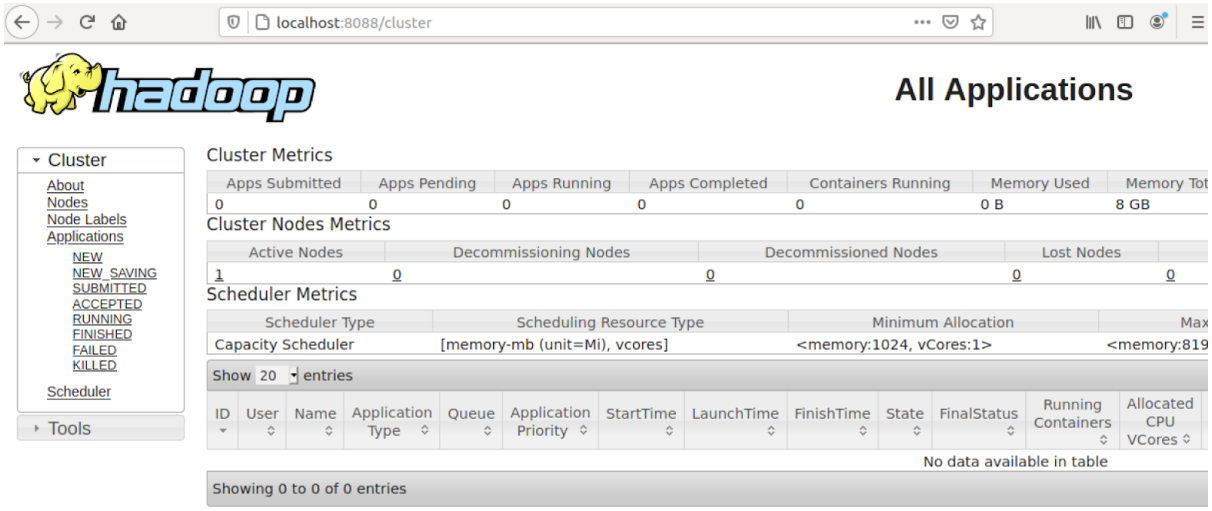
```
<configuration>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
    <property>
        <name>yarn.nodemanager.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOO
P_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HA
DOOP_MAPRED_HOME</value>
    </property>
</configuration>
```

6. Start ResourceManager daemon and NodeManager daemon:

   ```
   sbin/start-yarn.sh
   ```

7. Browse the web interface for the ResourceManager; by default it is available at:

   http://localhost:8088/



8. Run a MapReduce job.
9. When you're done, stop the daemons with:

```
sbin/stop-yarn.sh
```

Read more about modes of Hadoop here

- Local (Standalone) Mode
- Pseudo-Distributed Mode
- Fully-Distributed Mode

References

- https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html
- https://www.programmersought.com/article/93394144266/
- clean up Hadoop: https://stackoverflow.com/questions/26545524/there-are-0-datanodes-running-and-no-nodes-are-excluded-in-this-operation
- Turn off safemode: https://stackoverflow.com/questions/15803266/name-node-is-in-safe-mode-not-able-to-leave