A Brief HandBook to Hadoop Cluster Setup

Contributor: ECE472 23SU Teaching Group

Disclaimer

Please note that this manual is based on our teaching group's experience during the 22SU semester, which had a unique setup due to the pandemic. Some of the methods described in this manual may not be ideal in terms of security. We encourage you to develop your own strategy for set up after having a deeper understanding of Hadoop cluster setup.

Things todo

1. For everyone in one group, work individually to set up your Hadoop cluster in single node (or **single-node in a pseudo-distributed mode** to be exact)
2. Work in a group, properly setup Hadoop with multi node setup.

Before you start

1. Please ensure that you have a Linux system installed on your computer, either directly on the disk or within a virtual machine like VMWare. Please note that our teaching group has limited experience with Mac OS, Windows, or other operating systems, so we may not be able to provide much assistance if you encounter issues with these systems.
2. Kindly note that the version of Hadoop used in our course is **3.2.2**. Some links in the lab manual on canvas may refer to Hadoop 2.x installation, which is not applicable. Also, please be aware that Hadoop has a newer version of 3.3.x, which is not used in our course. Please ensure that you are using the correct version of Hadoop to avoid any compatibility issues.
3. This manual mainly refers to ref.1 and you can treat it as an extended version of ref.1

Single node setup

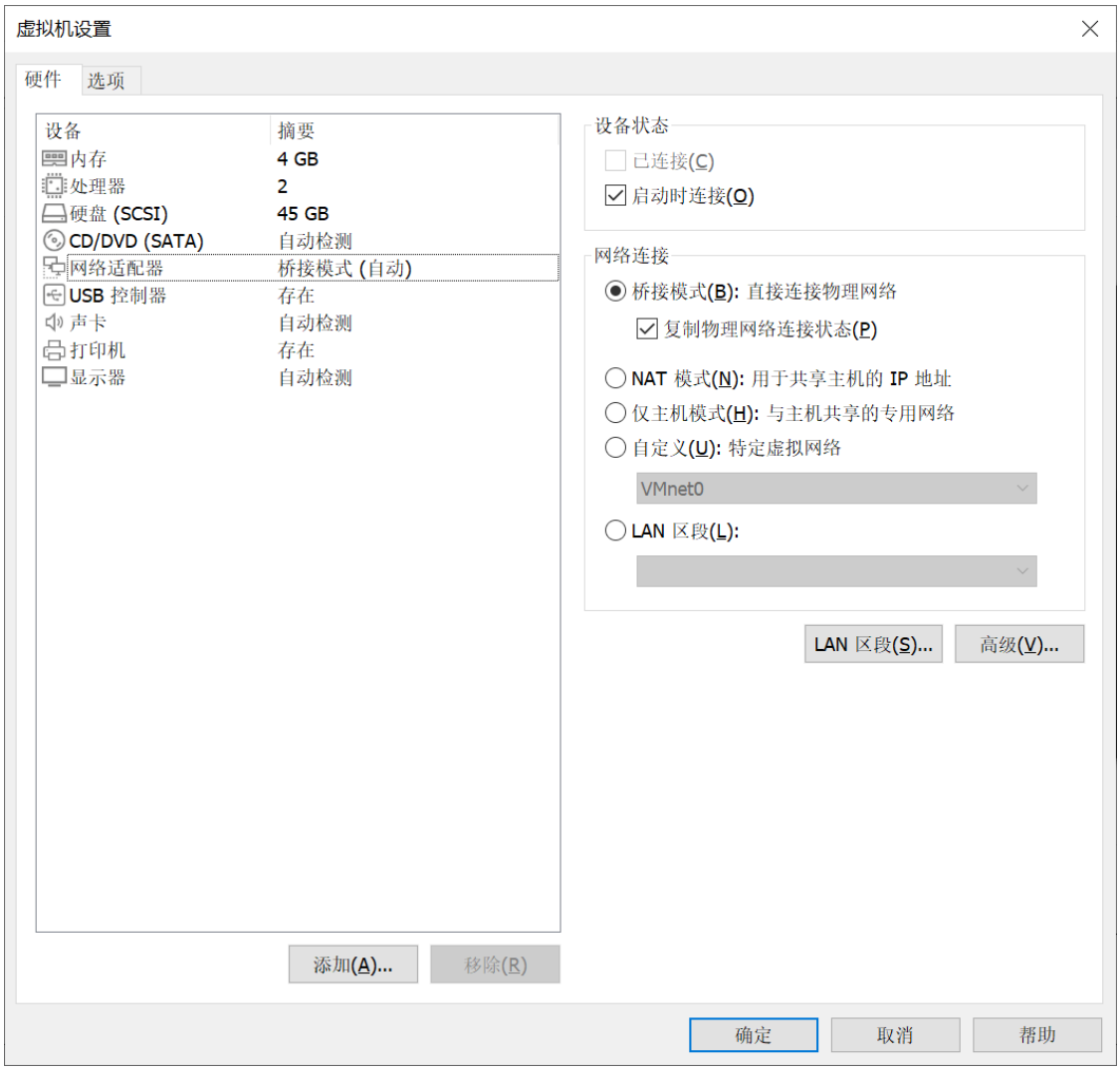
Single Node Setup is quite easy. You just need to follow the instructions from Hadoop official website. Also, it is very easy to switch back to single node after you properly set up the multi-node version.

Multi node setup

1. Set your Virtual Machine network connection to the host machine as **Bridged**. In the reference link, it introduces how to set up such things in Virtual Box.

   Here we introduce how to do that in VMware WorkStation.

   - Right click the **virtual machine** you intended to set network connection for and click **settings（设置)** in the pop-up menu.
   - In the pop-up window, choose **Network Adapter(网络适配器)** and check the bridged mode button.



2. Properly install and setup `ssh` and `pdsh`

   In your virtual machine, install `ssh` and `pdsh` with commands like

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

   add the following command to your shell configuration file (`~/.bashrc` for example):

   ```
   export PDSH_RCMD_TYPE=ssh
   ```

   If you don't know how to use `nano` or `vim` or other tools to edit files in linux, you can ask your teammates or ta for help during lab.

   After that, use `tail ~/.bashrc` to see whether you have this statement in your configuration file.

   At the end of this manual, we will show you the final file of our setup.

3. Generate a ssh key

   We assume all of you have done that. So this part is omitted. However, we recommend you to have a separate ssh key for the Hadoop cluster.

4. Install Java

Hadoop relies compile and run based on java 8.

To do that, use

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

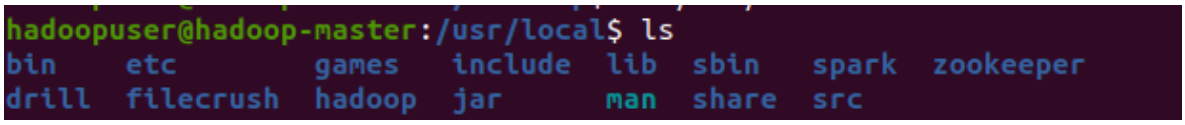
If you happen to have another version of java installed before, you can use following command to specify the default java version:

```
sudo update-alternatives --config java
sudo update-alternatives --config javac #for jdk, if you want
```

5. Download and install Hadoop source code

   You can download source code of Hadoop from Apache Hadoop 3.2.2 Release Page.

   After download, use `tar` command to decompress the file. For the sake of convenience, in this manual, we recommend you to move the extracted folder to `/usr/local` and name as `hadoop`. This will save you lots of time to solve the folder position issue.



6. Set configuration files for Hadoop

   - set Java path for Hadoop

     In `/usr/local/hadoop/etc/hadoop/`, we have a environment file `./hadoop-env.sh`, specify the java home for your hadoop there. For example:

     ```
     export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
     ```

     Search online if you don't know how to find your java home and notice that set it to the java8 home.

   - add hadoop command to your path and Java Home to environment file

     edit your `/etc/environment` file to have following content:

     ```
     PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/local/games:/usr/local/hadoop/bin:/usr/local/hadoop/sbin"
     JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64/jre
     ```

7. Add Hadoop users

   I guess you more or less notice that in previous screenshots, my username is `hadoopuser`. This is what should be done in this step. Create a new hadoopuser to have access to the hadoop folder.

```
sudo adduser hadoopuser
sudo usermod -aG hadoopuser hadoopuser
sudo chown hadoopuser:root -R /usr/local/hadoop/
sudo chmod g+rwx -R /usr/local/hadoop/
sudo adduser hadoopuser sudo
```

8. ip address

   This is the most critical part of this lab. Most of the connection issue of hadoop is due to ip address.

   We recommend you to install and activate sjtu vpn to get a unique ip address within the SJTU network. Alternatively, all of you connect to a shared router will also be fine.

   Then use `ip addr` to get your ip address.

   After that, collect all of your group members ip address and edit `/etc/hosts` file. Following is the example of our file.

```
hadoopuser@hadoop-master:/etc$ cat /etc/hosts
127.0.0.1   localhost hadoop-master
#127.0.1.1  kevinzhang-virtual-machine

# The following lines are desirable for IPv6 capable hosts
::1     ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters


111.186.48.231 hadoop-master #kaiwen
111.186.50.213 hadoop-slave-1 #haoxiang
111.186.46.232 hadoop-slave-2 #yuxuan
111.186.50.212 hadoop-slave-3 #yuxiang
```

   Note that if you are yuxuan in our example, you should have `127.0.0.1  localhost hadoop-slave-2`.

   Also, for every group member, change your hostname to the correspondent name.

```
hadoopuser@hadoop-master:/etc$ cat /etc/hostname
hadoop-master
```

   Then reboot your machine to make the change in effect.

9. share ssh key to enable access without passwd throughout the cluster

   Do the following command:

```
ssh-copy-id hadoopuser@hadoop-master
ssh-copy-id hadoopuser@hadoop-slave-1
ssh-copy-id hadoopuser@hadoop-slave-2
ssh-copy-id hadoopuser@hadoop-slave-3
```

This is a chance to check whether your ip address is correctly set. If success, you can `ssh hadoopuser@hadoop-slave-1` easily and have access to your group members' computer. Please do not do anything bad. :)

10. configuration of Hadoop itself (only master node needs to do it)

    We have several configuration files to set here, we will show you our version here.

    /usr/local/hadoop/etc/hadoop/core-site.xml

```
hadoopuser@hadoop-master:/usr/local/hadoop/etc/hadoop$ cat core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->
<!-- Put site-specific property overrides in this file.
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
    <property>
        <name>hadoop.tmp.dir</name>
        <value>/home/kevin-zhang/hdata</value>
    </property>
</configuration>
-->
<!-- multi -->
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://hadoop-master:9000/</value>
</property>
</configuration>
```

The same as the reference website.

/usr/local/hadoop/etc/hadoop/hdfs-site.xml

```
hadoopuser@hadoop-master:/usr/local/hadoop/etc/hadoop$ cat hdfs-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->
<!-- Put site-specific property overrides in this file.
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
</configuration>
-->
<configuration>
<property>
<name>dfs.namenode.name.dir</name>
<value>/usr/local/hadoop/hadoop-multi-data</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/usr/local/hadoop/hadoop-multi-name</value>
</property>
<property>
<name>dfs.replication</name>
<value>2</value>
</property>
</configuration>
```

Note the **replication** here which is a common question in exam (default replication level) :). And for every node (including master), you should have these folder manually created and content cleaned each time before you boot your cluster.

```
hadoopuser@hadoop-master:/$ ls -al /usr/local/hadoop/
total 232
drwxrwxr-x 12 hadoopuser root         4096 5月  25  2022 .
drwxr-xr-x 16 root       root         4096 6月  21  2022 ..
drwxrwxr-x  2 hadoopuser root         4096 1月   3  2021 bin
drwxrwxr-x  3 hadoopuser root         4096 1月   3  2021 etc
drwxrwxrwx  3 hadoopuser hadoopuser   4096 7月  30  2022 hadoop-multi-data
drwx------  3 hadoopuser hadoopuser   4096 7月  30  2022 hadoop-multi-name
drwxrwxr-x  2 hadoopuser root         4096 1月   3  2021 include
drwxrwxr-x  3 hadoopuser root         4096 1月   3  2021 lib
drwxrwxr-x  4 hadoopuser root         4096 1月   3  2021 libexec
-rw-rwxr--  1 hadoopuser root       150569 12月  5  2020 LICENSE.txt
drwxrwxr-x  3 hadoopuser root        12288 7月  30  2022 logs
-rw-rwxr--  1 hadoopuser root        21943 12月  5  2020 NOTICE.txt
-rw-rwxr--  1 hadoopuser root         1361 12月  5  2020 README.txt
drwxrwxr-x  3 hadoopuser root         4096 1月   3  2021 sbin
drwxrwxr-x  4 hadoopuser root         4096 1月   3  2021 share
```

/usr/local/hadoop/etc/hadoop/workers

For cluster

```
hadoopuser@hadoop-master:/usr/local/hadoop/etc/hadoop$ cat workers
hadoopuser@hadoop-master
hadoopuser@hadoop-slave-1
hadoopuser@hadoop-slave-2
hadoopuser@hadoop-slave-3
#localhost
```

For single node:

```
hadoopuser@hadoop-master:/usr/local/hadoop/etc/hadoop$ cat workers
#hadoopuser@hadoop-master
#hadoopuser@hadoop-slave-1
#hadoopuser@hadoop-slave-2
#hadoopuser@hadoop-slave-3
localhost
```

11. Copy all the configuration files to slave nodes

```
scp /usr/local/hadoop/etc/hadoop/* hadoop-slave-1:/usr/local/hadoop/etc/hadoop/
scp /usr/local/hadoop/etc/hadoop/* hadoop-slave-2:/usr/local/hadoop/etc/hadoop/
scp /usr/local/hadoop/etc/hadoop/* hadoop-slave-3:/usr/local/hadoop/etc/hadoop/
```

12. make the configuration in effect and start hdfs

```
source /etc/environment
hdfs namenode -format
start-dfs.sh
jps
```

Also, you should also follow Step 23 in the reference website.

13. YARN configuration

/usr/local/hadoop/etc/hadoop/yarn-site.xml

```
<property>
<name>yarn.resourcemanager.hostname</name>
<value>hadoop-master</value>
</property>
```

There can also be possible that you don't need this as yarn will automatically set the launch node as host.

Start yarn with `start-yarn.sh`.

/usr/local/hadoop/etc/hadoop/mapred-site.xml

```
hadoopuser@hadoop-master:/usr/local/hadoop/etc/hadoop$ cat mapred-site.xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
    <property>
        <name>yarn.app.mapreduce.am.env</name>
    <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
    </property>
    <property>
```

```
        <name>mapreduce.map.env</name>
    <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
    </property>
    <property>
        <name>mapreduce.reduce.env</name>
    <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
    </property>
</configuration>
```

References

1. Setting up Hadoop 3.2.1 Cluster with Multiple Nodes on Ubuntu Server 20.04 and/or Ubuntu Desktop 20.04 | by Andre Godinho | Analytics Vidhya | Medium
2. Linux 系统交大VPN使用说明-上海交通大学网络信息中心 (sjtu.edu.cn)