Foundations for Big Data Systems and Programming

Pejman Rasti

Email: prasti@esaip.org
pejman.rasti@univ-angers.fr

Course Website: Access from your "Moodle" portal

1

Hadoop implementation steps

- Installation of VM
- Network Configuration
- Define a model
  - Hadoop with two nodes
- Hadoop installation
- Copy file in nodes and apply a map reduce query

2

Installation of VM

- Oracle VmBox
  - Installation of ubuntu
    - 2 nodes (set vmdk)
    - Assign manual IP
      - *sudo su*
      - *nano /etc/hosts*
    - Ping the another node
    - disable firewall
      - *ufw disable*
    - Check Java version
      - java -version

3

Installation of VM

- Continue
  - Install wget
    - *apt-get install wget*
  - Install ssh server
    - *apt-get install openssh-srver*
  - download oracle jdk 1.8
  - Download Hadoop 2.6.5
  - Give the same privilege of root to the user
    - *Visudo*
      - *Under root line*
        - *pej ALL=(ALL:ALL) ALL*

4

Installation of VM

- Continue
  - Make folder for JDK and unzip the file
    - *cd /usr/lib*
    - *mkdir jvm*
    - *cd jvm*
    - *tar -xvf /home/pej/Downloads/jdk…*
    - *ls -all*
    - *ls jdk1…/bin*
    - *cd jdk…*
  - Copy the path of JDK in bash file
    - *cd*
    - *nano .bashrc*
    - Add these lines to the file
      - *export JAVA_HOME=/usr/lib/jvm/jdk…*
      - *export PATH=$PATH:$JAVA_HOME/bin*
    - *source .bashrc*

5

Installation of VM

- Continue
  - Copy the path of JDK in user bash file as well
    - *su – pej* (*"pej" is a user name)*
    - *nano .bashrc*
    - Add these lines to the file
      - *export JAVA_HOME=/usr/lib/jvm/jdk…*
      - *export PATH=$PATH:$JAVA_HOME/bin*
    - *source .bashrc*
  - Extraction of Hadoop file
    - *cd /usr/local*
    - *sudo tar -xvf /home/pej/Downloads/hadoop…*
    - *ls -all*
  - Changing the name from hadoop-2.6.5 to Hadoop
    - *sudo ln -s hadoop-2.6.5 hadoop*

6

Installation of VM

- Continue
  - Change the ownership of the folder
    - sudo chown -R pej:pej Hadoop
    - ls –all
  - Add the path of Hadoop to the bash file
    - cd
    - nano .bashrc
    - Copy these lines
      - export HADOOP_INSTALL=/usr/local/hadoop
      - export PATH=$PATH:$HADOOP_INSTALL/bin
      - export PATH=$PATH:$HADOOP_INSTALL/sbin
    - source .bashrc
  - Check if the Hadoop installed properly
    - hadoop version

7

Installation of VM

- Continue
  - Make ssh paswordless for both nodes
    - On the first node
      - ssh-keygen -t rsa -P “”
      - ls -all .ssh
    - On the second node
      - ssh-keygen -t rsa -P “”
      - ls -all .ssh
  - Copy ssh public key of node 1 to 2 and vice versa
    - On the first node
      - ssh-copy-id -i $HOME/.ssh/id_rsa.pub pej@m2
    - On the second node
      - ssh-copy-id -i $HOME/.ssh/id_rsa.pub pej@m1

8

Installation of VM

- Continue
  - Copy the ssh in authorized key on both nodes
    - *cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys*
  - *Copy JDK and Hadoop zip files into the second node*
    - scp /home/pej/Downloads/jdk… pej@m2:/tmp
    - scp /home/pej/Downloads/hoodop… pej@m2:/tmp
  - *Make the extractions of files in the second node*
    - sudo mkdir */usr/lib/jvm*
    - *cd /usr/lib/jvm*
    - *sudo tar -xvf /tmp/jdk….*
    - *sudo chown -R root:root jdk1….*
    - Cd /usr/local
    - sudo tar -xvf /tmp/ha..
    - sudo ln -s hadoop-2.6.5 hadoop
    - sudo chown -R pej:pej hadoop*

9

Installation of VM

- Continue
  - Copy the bash file from node 1 to node 2
    - scp .bashrc pej@m2:/home/pej
  - On the second node
    - source .bashrc
    - java -version
    - hadoop version

10

Hadoop Configuration

- Hadoop Configuration
  - Local files can be found
    - ls /usr/local/hadoop/etc/hadoop
  - We will make our model like
    - Machine 1
      - NameNode, DataNode, ResourceManager, NodeManager
    - Machine 2
      - DataNode, NodeManager, SecondNameNode

11

Hadoop Configuration

- Continue
  - On Machine 1
    - cd /usr/local/hadoop/etc/Hadoop
    - nano core-site.xml
    - Copy following lines in the file

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://m1:9000</value>
</property>
<property>
<name>dfs.permissions</name>
<value>false</value>
</property>
</configuration>
```

12

Hadoop Configuration

- Continue
  - The second file is for defining processing unit (YARN)
    - First we rename the file
      - mv mapred-site.xml.template mapred-site.xml
      - nano mapred-site.xml
    - Copy following lines in the file

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

13

Hadoop Configuration

- Continue
  - The third file is hdfs-site.xml to define replication factors, where my name node store the meta data
    - nano hdfs-site.xml
    - Copy following lines in the file

```
<property>
<name>dfs.replication</name>
<value>2</value>
</property>
```

As we have 2 data node replication should be 2.

```
<property>
<name>dfs.namenode.name.dir</name>
<value>/abc/name</value>
</property>
```

For saving the meta data we need a path - *later we should make abc directory*

14

Hadoop Configuration

- Continue

```
<property>
<name>dfs.datanode.data.dir</name>
<value>/abc/data1</value>
<final>true</final>
</property>
```

For saving blocks of datanode, we should give a path

```
<property>
<name>dfs.namenode.http-address</name>
<value>m1:50070</value>
</property>
```

The port that name node

```
<property>
<name>dfs.namenode.secondary.http-address</name>
<value>m2:50090</value>
</property>
```

Defining the location of Secondary name node

15

Hadoop Configuration

- Continue
  - Next file will be yarn-site.xml which is about resource manager, resource tracker …
    - Nano yarn-site.xml
    - Copy following lines in the file

```
<property>
<name>yarn.resourcemanager.address</name>
<value>m1:9001</value>
</property>
```

The resource manager should be on m1

```
<property>
<name>yarn.resourcemanager.resource-tracker.address</name>
<value>m1:8031</value>
</property>
```

The resource tracker is on m1 as well

16

Hadoop Configuration

- Continue

```
<property>
<name>yarn.nodemanager.aux-services.mapreduce_shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>

<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>

<property>
<name>yarn.resourcemanager.nodes.include-path</name>
<value>/usr/local/hadoop/etc/hadoop/include</value>
</property>
```

17

Hadoop Configuration

- Continue

```
<property>
<name>yarn.resourcemanager.nodes.exclude-path</name>
<value>/usr/local/hadoop/etc/hadoop/exclude</value>
</property>

<property>
<name>yarn.log-aggregation-enable</name>
<value>true</value>
</property>

<property>
<name>yarn.nodemanager.remote-app-log-dir</name>
<value>/tmp</value>
</property>
```

18

Hadoop Configuration

- Continue
  - Next file is the file slaves to define machines
    - nano slaves
    - Write in the file
      - m1
      - m2

19

Hadoop Configuration

- Continue
  - Now, copy the files into m2 machine.
    - scp core-site.xml pej@m2:/usr/local/hadoop/etc/Hadoop
    - scp mapred-site.xml pej@m2:/usr/local/hadoop/etc/Hadoop
    - scp yarn-site.xml pej@m2:/usr/local/hadoop/etc/Hadoop
    - scp hdfs-site.xml pej@m2:/usr/local/hadoop/etc/Hadoop
    - scp slaves pej@m2:/usr/local/hadoop/etc/hadoop
  - Switching to the machine 2
    - ssh m2
    - cd /usr/local/hadoop/etc/hadoop/
    - Remove the mapred template file
      - rm -rf mapred-site.xml.template

20

Hadoop Configuration

- Continue
  - we modify hdfs-site.xml
    - We just need to remove property of namenode

```
<property>
<name>dfs.namenode.name.dir</name>
<value>/abc/name</value>
</property>
```

Must be removed

    - Then path of datanode /abc/data2

```
<property>
<name>dfs.datanode.data.dir</name>
<value>/abc/data2</value>
<final>true</final>
</property>
```

21

Hadoop Configuration

- Continue
  - we should add check point to see how frequently name node check the point

```
<property>
<name>dfs.namenode.checkpoint.period</name>
<value>600</value>
</property>
```

22

Hadoop Configuration

- Continue
  - Make the directory of abc in both machines
    - cd
    - sudo mkdir /abc
    - sudo chown -R pej:pej /abc
    - ssh m1
    - cd
    - sudo mkdir /abc
    - sudo chown -R pej:pej /abc

23

Hadoop Configuration

- Continue
  - The last step is to formatting Hadoop (must be done once)
    - hdfs namenode -format
  - we can check the cluster ids
    - cat /abc/name/current/VERSION
  - Let's start our cluster
    - start-all.sh

This run two code of start-dfs.sh and start-yarn.sh

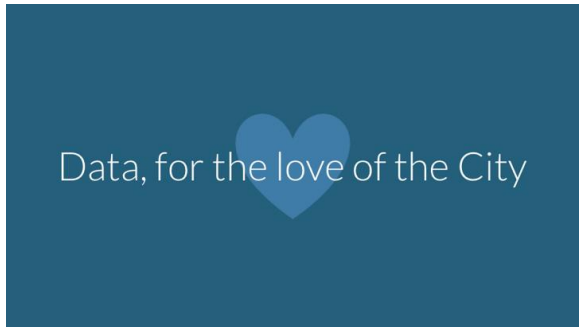
  - To check how cluster is configured
    - jps
    - hdfs dfsadmin -report

24

Hadoop Configuration

- Continue
  - We can check it on web browser by
    - http://m1:50070
  - And also (resource manager)
    - http://m1.8088

25



THANK YOU

Assoc. Prof. Pejman Rasti

26