CSCI4180 Tutorial
Introduction to cloud platform 2

XU, Min (xum@cse)
Sep. 26, 2013

Outline

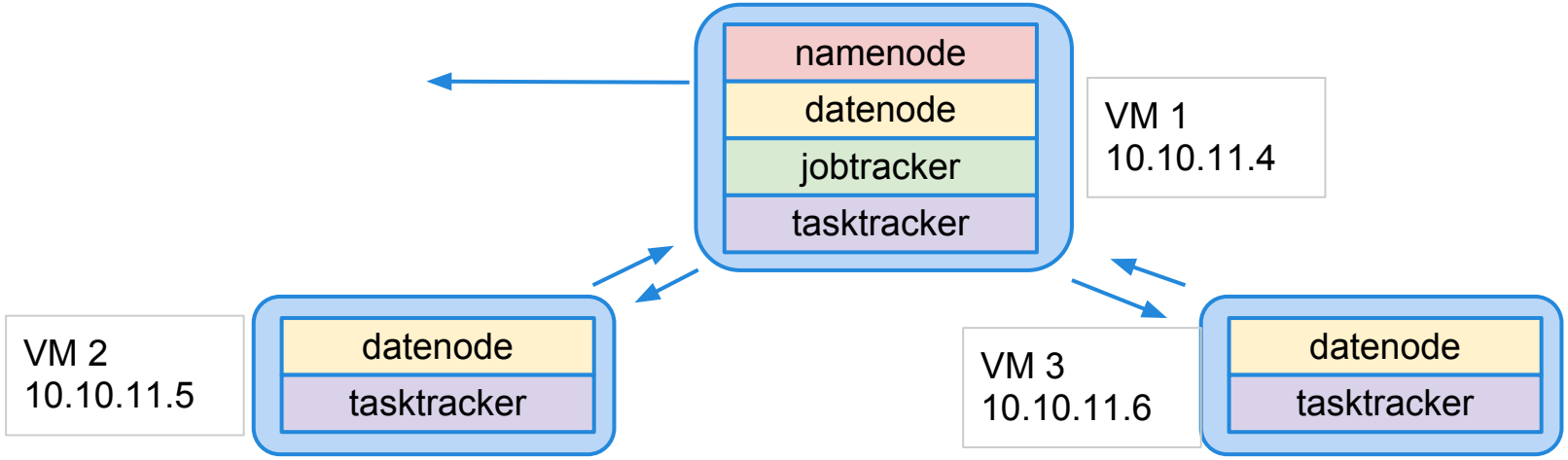
- Setup Hadoop cluster
- Maintain Hadoop cluster
- WordCount Example

Setup Hadoop Cluster

- Last time, we have created several VM instances of our own
- This time, we will set up small-scale Hadoop cluster using our VM instances

Setup Hadoop Cluster

- Architecture
  - 3 instances each group
  - One hosts namenode, datanode, jobtracker and tasktracker
  - The other two host datanode and tasktracker



Setup Hadoop Cluster

- If you launch instances using “Hadoop Lab AMI”, you can skip slides 6 - 12
- Otherwise, follow the instructions from the next slide

Setup Hadoop Cluster

- Make sure you have Internet Access on each node
  - $ export http_proxy='http://proxy.cse.cuhk.edu.hk:8000'
  - $ export https_proxy='http://proxy.cse.cuhk.edu.hk:8000'

Setup Hadoop Cluster

- Install java 1.6 (use your root account) on each node
  - $ apt-get install python-software-properties
  - $ add-apt-repository ppa:webupd8team/java
  - $ apt-get update
  - $ apt-get install oracle-java7-installer

Setup Hadoop Cluster

- Switch to normal user “hadoop”
  $ su hadoop
- If you do not have user “hadoop”
  $ adduser hadoop
    - enter your password when necessary…
  $ su hadoop
  $ cd ~

Setup Hadoop Cluster

- Download Hadoop on each node
  $ wget http://archive.apache.org/dist/hadoop/core/hadoop-0.20.203.0/hadoop-0.20.203.0rc1.tar.gz

Setup Hadoop Cluster

- Place Hadoop (I put it in home directory) on each node

```
$  tar xzf hadoop-0.*.*.tar.gz
$  mv hadoop-0.*.* hadoop
```

Setup Hadoop Cluster

- Set environment variable on each node (I prefer to put them in ~/.bashrc)

```
$ export HADOOP_HOME=~/hadoop
$ export PATH=$PATH:$HADOOP_HOME/bin
```

Setup Hadoop Cluster

- Set hadoop environment on each node

  In ***hadoop/conf/hadoop-env.sh***, add

  ```
  export JAVA_HOME=/usr/lib/jvm/java-7-oracle
  #depends where you put the jvm
  export HADOOP_OPTS=-Djava.net.preferIPv4Stack=true
  ```

Setup Hadoop Cluster

- Set path for HDFS storage on each node (I put it in hadoop/tmp)

  #under HOME directory

  $ mkdir hadoop/tmp

Setup Hadoop Cluster

- Configure SSH on each node

```
$ ssh-keygen -t rsa -P ""
$ cat $HOME/.ssh/id_rsa.pub >> \     $HOME/.
  ssh/authorized_keys
```

Setup Hadoop Cluster

- Configure SSH on namenode only
  - $ ssh-copy-id -i $HOME/.ssh/id_rsa.pub \
    hadoop@10.10.11.4
  - $ ssh-copy-id -i $HOME/.ssh/id_rsa.pub \
    hadoop@10.10.11.5
  - $ ssh-copy-id -i $HOME/.ssh/id_rsa.pub \
    hadoop@10.10.11.6
- Test ssh configuration
  - check whether namenode can ssh all the datanode by key (i.e. no need to type password)

Setup Hadoop Cluster

- Set hadoop core on each node
  - In ***hadoop/conf/core-site.xml*** Add property

```
<property>
        <name>hadoop.tmp.dir</name>
        <value>/home/hadoop/hadoop/tmp</value>
</property>
<property>
        <name>fs.default.name</name>
        <value>hdfs://10.10.11.4:54310</value>
</property>
```

Setup Hadoop Cluster

- Set hadoop mapreduce on each node
  - In ***hadoop/conf/mapred-site.xml*** Add property

```
<property>
   <name>mapred.job.tracker</name>
   <value>10.10.11.4:54311</value>
</property>
```

Setup Hadoop Cluster

- Set hadoop HDFS on each node
  - In ***hadoop/conf/hdfs-site.xml*** Add property

```
<property>
        <name>dfs.replication</name>
        <value>3</value>
</property>
```

Setup Hadoop Cluster

- Set hadoop masters on namenode
  - In ***hadoop/conf/masters*** Add hostname which is supposed to run JobTracker and NameNode

    10.10.11.4

Setup Hadoop Cluster

- Set hadoop slaves on namenode
  - In ***hadoop/conf/slaves*** Add hostname which is supposed to run JobTracker and NameNode

    10.10.11.4

    10.10.11.5

    10.10.11.6

Setup Hadoop Cluster

- Format namenode on namenode

```
$ hadoop namenode -format
```

Setup Hadoop Cluster

- Start hadoop on namenode
  - $ start-dfs.sh
  - $ start-mapred.sh

  # you can type “jps” to see whether the startup is successful

  # it looks like the follow, but I don’t have the datanode and tasktracker

Setup Hadoop Cluster

- Stop hadoop on namenode

  $ stop-mapred.sh

  $ stop-dfs.sh

Setup Hadoop Cluster

- Some operations related to HDFS
  - From Local to HDFS

    $ hadoop dfs -copyFromLocal <local dir/file> <hdfs URI> (for user home URI: /user/username)
  - List files in HDFS

    $ hadoop dfs -ls <hdfs URI>
  - Cat files in HDFS

    $ hadoop dfs -cat <hdfs URI>
  - From HDFS to local

    $ hadoop dfs -copyToLocal <local dir/file> <hdfs URI>

Maintain Hadoop cluster

- Add one more instance into cluster
  - Stop Hadoop services on namenode
  - For the new instance, repeat steps from slide 6 to slide 17
  - Add ip of new instance in ***hadoop/conf/slaves*** on namenode
  - Format namenode and start Hadoop
- Remove one instance from cluster
  - Stop Hadoop services on namenode
  - Remove ip of the instance from ***hadoop/conf/slaves***
  - Format namenode and start Hadoop

Maintain Hadoop cluster

- Change namenode to another instance
  - Stop Hadoop on old namenode
  - Do instructions from slide 6 to 19 on the new namenode
  - Modify configure files on each datanode (slides 15-17)
  - Format namenode and start Hadoop

Maintain Hadoop cluster

- Snapshot your instance in case of disasters!
  - Create Snapshot: login ⇒ Instances & Volumes ⇒ select an instance, and click Snapshot. (may take a few minutes to be done)
  - View Snapshot: login ⇒ Images & Snapshots ⇒ #Instance Snapshots
  - Launch instance from saved point: login ⇒ Images & Snapshots ⇒ select snapshot, and launch it.
- Try not to create too many snapshots to waste your hard disk space!
  - Every time you create a new snapshot, you are welcome to delete the older version

WordCount Example

- Download the java source code from course website, say, WordCount.java, to your namenode, home directory
- Compile and run the program

```
$ mkdir wordcount
$ javac -classpath $HADOOP_HOME/hadoop-core-0.20.203.0.jar WordCount.java -d wordcount
$ jar -cvf wordcount.jar -C wordcount/ .
$ hadoop jar wordcount.jar /path/to/input/file /path/to/output/
```

Email me if any problem related to cloud platform