UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

Setting-up a HDFS cluster

Lecturers: Petar Jovanovic and Sergi Nadal

October 28th, 2016

In this document, we are going to see how to start up a small HDFS with three machines. More precisely, such cluster will be composed of a master machine running as NameNode and two other slaves running as DataNodes.

We are mainly going to work with three different technologies:

1. Hadoop main stack (i.e., HDFS, YARN and MapReduce)
2. HBase (to see in future lectures).
3. JRE 7

We accordingly will focus only on the one containing HDFS for this practical session. Note that all this software could be installed from packages, but by working with them this way we will just be able to freely delete and recreate things as many times as we want, which will improve the didactic output of this session. So, do not worry about breaking things!

Importantly, though, you need to know your environment is composed of three virtual machines. The first one is supposed to act as master in every single technology we use, while the others will be the slaves. The hostnames are *master*, *slave1* and *slave2*, respectively.

1 Configure Hadoop

Log into the master machine and configure the HDFS for this session. The first thing you need to do is to obtain the files in the home folder:

```
cp tarballs/hadoop-2.5.1.tar.gz ~/.
tar xf hadoop-2.5.1.tar.gz

cp tarballs/jre-7u75-linux-x64.gz ~/.
tar xf jre-7u75-linux-x64.tar.gz
```

Now we start working on the real setting-up:

1. The file *hadoop-2.5.1/etc/hadoop/slaves* should include the list of the machine hostnames that will work as DataNodes. Following the same example in this document, the content should be:

```
slave1
slave2
```

2. Next step is to configure file *hadoop-2.5.1/etc/hadoop/core-site.xml*. The first property defines the HDFS URL where the NameNode is going to be listening on. The second property defines the path where HDFS data is going to be physically stored.

```
<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://master:27000</value>
        </property>
        <property>
                <name>hadoop.tmp.dir</name>
                <value>/home/bdma**/data</value>
        </property>
</configuration>
```

3. File *hadoop-2.5.1/etc/hadoop/hdfs-site.xml* needs to be updated as following. The first property disables the minimum block size a user can define for a given file, and the second property disables the safemode threshold for start-up. Note these two properties are not really essential for the setting-up of a Hadoop cluster, but they will allow going through the exercises more comfortably.

```
<configuration>
        <property>
                <name>dfs.namenode.fs-limits.min-block-size</name>
                <value>0</value>
        </property>
        <property>
                <name>dfs.namenode.safemode.threshold-pct</name>
                <value>0</value>
        </property>
</configuration>
```

4. Remember Hadoop is Java-based which means we need a Java Runtime Environment below. Here, instead of installing Java by packages, we are going to use a JRE instance in our home directory. Thus, in order to link it with Hadoop, add the following lines in *hadoop-2.5.1/etc/hadoop/hadoop-env.sh*.

```
export JAVA_HOME="/home/bdma**/jre1.7.0_75"
export HADOOP_HEAPSIZE=2000
```

5. Finally, we need to propagate on all these files and configurations to the other nodes. So we therefore type:

```
scp -r hadoop-2.5.1 jre1.7.0_75 bdma**@slave1:.
scp -r hadoop-2.5.1 jre1.7.0_75 bdma**@slave2:.
```

2 Starting up the cluster

Now the cluster should be perfectly configured and, hence, it should be ready to run. The start-up process is very simple and just needs of two steps:

1. Format the NameNode. Type, in the NameNode:

   ```
   hadoop-2.5.1/bin/hdfs namenode -format
   ```

2. Start it up!

   ```
   hadoop-2.5.1/sbin/start-dfs.sh
   ```

After running those two commands, the HDFS web UI should be available at port 50070. Thus, you should check your email to see what public address opens to your master:50070. Finally, note that if we try to explore a little bit on our HDFS directory and we run the next command we are going to obtain the error below since no user folder has been yet defined:

```
hadoop-2.5.1/bin/hdfs dfs -ls
ls: ‘.’: No such file or directory
```

In order to solve this, we type:

```
hadoop-2.5.1/bin/hdfs dfs -mkdir /user
hadoop-2.5.1/bin/hdfs dfs -mkdir /user/bdma**
```

3 Stopping the cluster

To shut the Hadoop cluster down, run:

```
hadoop-2.5.1/sbin/stop-dfs.sh
```