Hadoop & HDFS

Let’s practice!

First Steps [1]

Download and install Virtual Box

Download an Ubuntu release

Create an Ubuntu VM

Pre-Configuration [1]

Pre-requisites for Hadoop:

1. **Oracle Java JDK (version 8)**

2. **Ad-hoc system user for Hadoop**

   (suggested but not needed)

3. **SSH configuration**

Pre-Configuration [2]

1. Oracle Java JDK (version 8)

Install it:

```
sudo add-apt-repository ppa:webupd8team/java

sudo apt-get update

sudo apt-get install oracle-java8-installer
```

Check it:

```
java -version
```

Pre-Configuration [3]

2. Ad-Hoc System User for Hadoop

Create it:

```
sudo addgroup hadoop

sudo adduser --ingroup hadoop hduser
```

Grant it superuser permissions:

```
sudo adduser hduser sudo
```

Pre-Configuration [4]

3. SSH Configuration [1]

Generate the public private keys:

```
su - hduser

ssh-keygen -t rsa -P ""

Generating public/private rsa key pair.
Enter file in which to save the key (/home/hduser/.ssh/id_rsa):
```

Press Enter…

```

Created directory '/home/hduser/.ssh'.
Your identification has been saved in /home/hduser/.ssh/id_rsa.
Your public key has been saved in /home/hduser/.ssh/id_rsa.pub....

```

Pre-Configuration [5]

3. SSH Configuration [2]

Copy the public key to the host you will connect (localhost in our example):

```
cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
```

SSH is required to enable access to the host machine (localhost), so Hadoop does not ask for a password at inconvenient times

Pre-Configuration [5]

3. SSH Configuration [3]

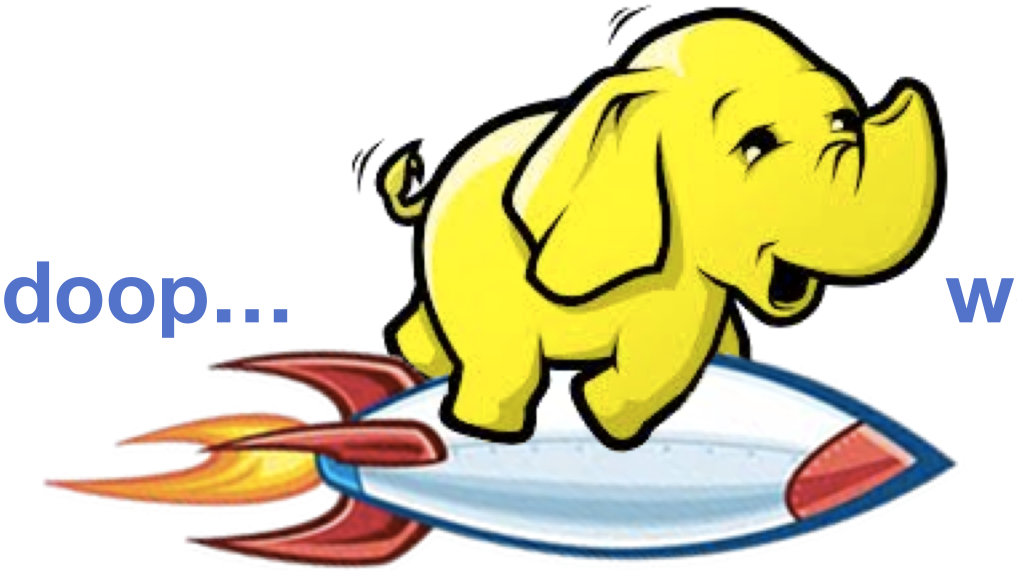
Check configuration:

```
ssh localhost
```

- You should be able to log in without a problem.
- It may be the case to save your local machine’s host key fingerprint to the **hduser** user’s **known_hosts** file

If SSH is not installed, with the root user

```
sudo apt-get install openssh-server
```

Hadoop…  we are coming!

Hadoop Installation [1]

- **Download Hadoop binaries from**

  https://hadoop.apache.org/releases.html

```
wget http://it.apache.contactlab.it/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz
```

- **Or in the downloaded virtual machine do:**

```
sudo mv /home/bigdata/hadoop-2.7.7.tar.gz /home/hduser
sudo chown /home/hduser/hadoop-2.7.7.tar.gz
```

- **Install in /usr/local/**

```
cd
tar -xzf hadoop-2.7.7.tar.gz
sudo mv hadoop-2.7.7 /usr/local
sudo chown -R hduser:hadoop /usr/local/hadoop-2.7.7
sudo ln -s /usr/local/hadoop-2.7.7 /usr/local/hadoop
sudo chown -h hduser:hadoop /usr/local/hadoop
```

Hadoop Installation [2] - Folder content

```
.
├── bin                     % scripts to interact with Hadoop
│   ├── hadoop              % script to interact with all the Hadoop environment
│   └── hdfs                % script to interact with the HDFS part of Hadoop
.
.
.
├── etc
│   └── hadoop              % bash and other scripts
│       ├── hadoop-env.sh   % Env variables used in the scripts to run Hadoop
│       ├── core-site.xml   % I/O settings for Hadoop Core (common to HDFS and MR)
│       ├── hdfs-site.xml   % conf settings for HDFS daemons (namenode and others)
│       ├── mapred-site.xml % conf settings for MR daemons (jobtracker and tasktrackers)
│       ├── core-site.xml   % I/O settings for Hadoop Core (common to HDFS and MR)
│       └── slaves          % contains the adresses to the datanodes
.
.
.
├── sbin                    % scripts to launch Hadoop DFS and Map/Reduce daemons
│   ├── start-dfs.sh        % starts the Hadoop DFS daemons, the namenode and datanodes
│   ├── stop-dfs.sh         % stops the Hadoop DFS daemons
│   ├── start-mapred.sh     % starts the Hadoop Map/Reduce daemon
│   ├── stop-mapred.sh      % stops the Hadoop Map/Reduce daemon
│   ├── start-all.sh        % starts all Hadoop daemons -> deprecated; start first dfs then mapred
│   └── stop-all.sh         % stops all Hadoop daemons  -> deprecated; stop firstdfs then mapred
```

Hadoop Installation [3]

- **Open the configuration file of hduser $HOME/.bashrc (you can use any editor you want, this is just an example):**

```
gedit $HOME/.bashrc
```

- **Add to it the following, and then save:**

```
# Set Hadoop-related environment variables
export HADOOP_HOME='/usr/local/hadoop'

# Set JAVA_HOME
export JAVA_HOME='/usr/lib/jvm/java-8-oracle’
export PATH="$JAVA_HOME/bin:$PATH"

# This is just syntactic sugar, take them as they are
export HADOOP_INSTALL="$HADOOP_HOME"
export HADOOP_MAPRED_HOME="$HADOOP_HOME"
export HADOOP_COMMON_HOME="$HADOOP_HOME"
export HADOOP_HDFS_HOME="$HADOOP_HOME"
export YARN_HOME="$HADOOP_HOME"
export HADOOP_CONF_DIR="${HADOOP_HOME}/etc/hadoop"
```

Hadoop Installation [4]

- **Reload the $HOME/.bashrc file for hduser:**

```
source $HOME/.bashrc
```

- **Edit the file $HADOOP_HOME/etc/hadoop/hadoop-env.sh**

  Replace:

```
export JAVA_HOME=${JAVA_HOME}
```

  With:

```
export JAVA_HOME='/usr/lib/jvm/java-8-oracle'
```

Hadoop Installation [5]

- **Set the directory where data blocks and namenode metadata will be stored:**

```
sudo mkdir -p /usr/local/hadoop-data/namenode
sudo mkdir -p /usr/local/hadoop-data/datanode
sudo chown -R hduser:hadoop /usr/local/hadoop-data
sudo chmod -R 750 /usr/local/hadoop-data
```

- **Edit the file $HADOOP_HOME/etc/hadoop/core-site.xml and add within the <configuration> tags**

```
<property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:8020</value>
    <description>NameNode URI</description>
</property>
```

Hadoop Installation [6]

- **Edit the file $HADOOP_HOME/etc/hadoop/hdfs-site.xml and add the following within the <configuration> tags**

```
<property>
    <name>dfs.datanode.data.dir</name>
    <value>file:///usr/local/hadoop-data/datanode</value>
    <description>DataNode directory</description>
</property>
<property>
    <name>dfs.namenode.name.dir</name>
    <value>file:///usr/local/hadoop-data/namenode</value>
    <description>NameNode directory for namespace and transaction logs storage</description>
</property>
<property>
    <name>dfs.namenode.http-address</name>
    <value>localhost:50070</value>
    <description>Your NameNode hostname for http.</description>
</property>
```

Hadoop Installation [7]

- **In order to create a pseudo-distributed mode, edit the file $HADOOP_HOME/etc/hadoop/hdfs-site.xml and insert the following property inside the <configuration> tags:**

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

- **Please remember that in a real environment the replication factor will be higher than 1**

Hadoop Installation [8]

- **Now, we can format our “distributed” filesystem**

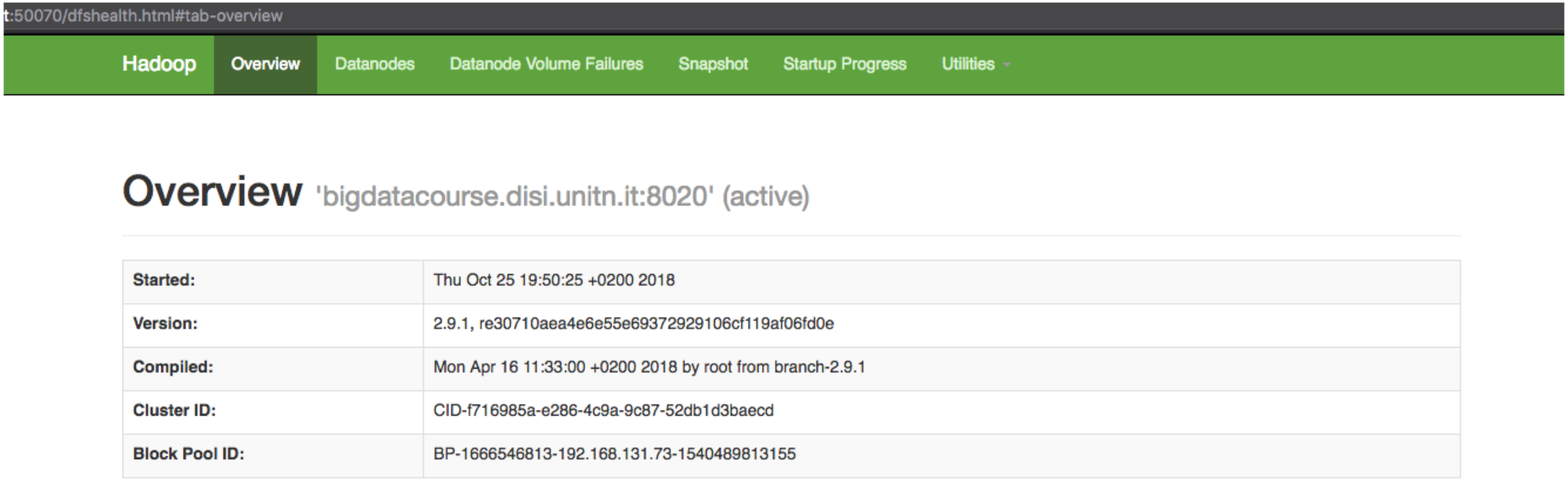
```
$HADOOP_HOME/bin/hdfs namenode -format
```

- **Finally, we can start NameNode and DataNode daemon with only one command:**

```
$HADOOP_HOME/sbin/start-dfs.sh
```

Hadoop Installation [9]

If everything went through, you can see and browse the web interface visiting http://localhost:50070/ from the virtual machine



Overview 'bigdatacourse.disi.unitn.it:8020' (active)

Started:	Thu Oct 25 19:50:25 +0200 2018
Version:	2.9.1, re30710aea4e6e55e69372929106cf119af06fd0e
Compiled:	Mon Apr 16 11:33:00 +0200 2018 by root from branch-2.9.1
Cluster ID:	CID-f716985a-e286-4c9a-9c87-52db1d3baecd
Block Pool ID:	BP-1666546813-192.168.131.73-1540489813155

Explore Hadoop [1]

- **Create the needed directories to execute MR Jobs**

```
$HADOOP_HOME/bin/hadoop fs -mkdir /user
$HADOOP_HOME/bin/hadoop fs -mkdir /user/hduser
```

- **Create a file**

```
echo "write some text here" >> test.txt
```

- **Copy your input files into the distributed filesystem**

```
$HADOOP_HOME/bin/hadoop fs -put test.txt /
```

Explore Hadoop [2]

View the results:

- Via Hadoop filesystem

```
$HADOOP_HOME/bin/hdfs dfs –cat test.txt
```

- Via local filesystem

```
$HADOOP_HOME/bin/hdfs dfs -get test.txt .
cat ./test.txt
```

Explore Hadoop [3]

- **When you are done:**

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

Conclusions

However, think before using it!

Command-line tools can be

235x

faster than your Hadoop cluster

http://aadrake.com/command-line-tools-can-be-235x-faster-than-your-hadoop-cluster.html

References

- Hadoop starting guide

http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html

https://wiki.apache.org/hadoop/GettingStartedWithHadoop

- Hadoop – The Definitive Guide, 4th version, Tom White, O'Reilly 2015

Contacts

For any problem, write me a mail:

daniele.foroni@unitn.it