Hadoop config

lucic@inf.ethz.ch

April 17, 2013

This is the recommended Hadoop setup for the **pseudo-distributed** (local) mode. We refer to the Hadoop installation folder as **HADOOP**. Configuration files are located in **HADOOP/conf** folder. Before running any of the changes make sure you **backup the configuration folder**. If you make any changes to these files you need to **restart Hadoop** (**HADOOP/bin/stop-all.sh** followed by **HADOOP/bin/start-all.sh**)

1 hadoop-env.sh

Make sure that you set the correct JAVA_HOME in hadoop-env.sh (otherwise Hadoop won't start):

```
export JAVA_HOME="/path/to/java/home"
```

Other settings in this file are not important at this point.

2 core-site.xml

This file contains configuration information that overrides the default values for core Hadoop properties. Here we want to set the IP and the port of the namenode. Since you are running in pseudo-distributed mode you can use *localhost* as IP. **WARNING**: if the namenode is starting normally on your machine there is no need to update your configuration file to match this setting.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

3 hdfs-site.xml

This file contains site-specific HDFS configuration. **WARNING**: if your HDFS is working properly there is no need to change these properties.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <!-- These are local directories used to hold distributed filesystem data on
       the master node and slave nodes respectively. -->
  <property>
    <name>dfs.name.dir</name>
    <value>/local/hadoop/hadoopdata/hdfsname</value>
```

```
    <final>true</final>
  </property>
  <property>
    <name>dfs.data.dir</name>
    <value>/local/hadoop/hadoopdata/hdfsdata</value>
    <final>true</final>
  </property>
  <!-- A base for other temporary directories. Default is /tmp/hadoop-${user.name} -->
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/var/hadoopdata/temp</value>
  </property>
  <!-- Default block replication. The actual number of replications can be
       specified when the file is created. -->
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

4 mapred-site.xml

This file contains the configuration information that overrides the default values for MapReduce configuration properties. **WARNING**: Update properties that are missing from your configuration or those that exist but are set to a different value. The critical property to set is **mapred.reduce.tasks** which enable multiple simultaneous reduce tasks. If you already have the **mapred.job.tracker** property set, you don't have to change it.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>mapred.job.tracker</name>
    <value>localhost:9001</value>
  </property>
  <property>
    <name>mapred.tasktracker.map.tasks.maximum</name>
    <value>NUMBER_OF_CORES_YOUR_PROCESSOR_HAS</value>
  </property>
  <property>
    <name>mapred.tasktracker.reduce.tasks.maximum</name>
    <value>NUMBER_OF_CORES_YOUR_PROCESSOR_HAS</value>
  </property>
  <!-- The default number of reduce tasks per job. Typically set to 99% of the
       cluster's reduce capacity, so that if a node fails the reduces can still be
       executed in a single wave. Set this to the same value as the one above.
       This enables multiple simultaneous Reduce tasks. -->
  <property>
    <name>mapred.reduce.tasks</name>
    <value>NUMBER_OF_CORES_YOUR_PROCESSOR_HAS</value>
  </property>
</configuration>
```