Table of Contents

[1. Purpose 2](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438404)

[2. Tools Used 2](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438405)

[3. Apache Hadoop Single Node setup 2](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438406)

[3.1 Prerequisites 2](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438407)

[3.1.1 Supported Platforms 2](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438408)

[3.1.2 Required Software 2](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438409)

[3.1.3 Installing Software 2](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438410)

[3.1.4 Download 3](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438411)

[3.2 Prepare to Start the Hadoop Cluster 3](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438412)

[3.2.1 Pseudo-Distributed Operation 3](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438413)

[3.2.2 Configuration 4](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438414)

[3.2.3 Setup passphraseless ssh 5](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438415)

[3.2.4 Execution 5](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438416)

[4. Installing CDH 5 with YARN on a Single Linux Node in Pseudo-distributed mode 6](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438417)

[5. Templates and References 6](file:///C:\Users\huynh\Desktop\Cloudera%20Hadoop%20-%20Single%20Node%20Setup.docx#_Toc448438418)

# Purpose

This document describes how to set up and configure a single-node Hadoop installation so that you can quickly perform simple operations using Hadoop MapReduce and the Hadoop Distributed File System (HDFS).

# Tools Used

|  |  |
| --- | --- |
| * Cloudera | Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data: The Enterprise Data Hub. Cloudera offers enterprises one place to store, process, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. |

# Apache Hadoop Single Node setup

## Prerequisites

### Supported Platforms

GNU/Linux is supported as a development and production platform. Hadoop has been demonstrated on GNU/Linux clusters with 2000 nodes.

Win32 is supported as a development platform. Distributed operation has not been well tested on Win32, so it is not supported as a production platform.

### Required Software

Required software for Linux and Windows include:

* JavaTM 1.6.x, preferably from Sun, must be installed.
* ssh must be installed and sshd must be running to use the Hadoop scripts that manage remote Hadoop daemons.

Additional requirements for Windows include:

* Cygwin - Required for shell support in addition to the required software above.

### Installing Software

If your cluster doesn't have the requisite software you will need to install it.

For example on Ubuntu Linux:

$ sudo apt-get install ssh

$ sudo apt-get install rsync

On Windows, if you did not install the required software when you installed cygwin, start the cygwin installer and select the packages:

* openssh - the Net category

### Download

To get a Hadoop distribution, download a [recent stable](http://hadoop.apache.org/core/releases.html) release from one of the Apache Download Mirrors

## Prepare to Start the Hadoop Cluster

Unpack the downloaded Hadoop distribution. In the distribution, edit the file conf/hadoop-env.sh to define at least JAVA\_HOME to be the root of your Java installation.

Try the following command:

$ bin/hadoop

This will display the usage documentation for the hadoop script.

Now you are ready to start your Hadoop cluster in one of the three supported modes:

* Local (Standalone) Mode
* Pseudo-Distributed Mode
* Fully-Distributed Mode

### Pseudo-Distributed Operation

Hadoop can also be run on a single-node in a pseudo-distributed mode where each Hadoop daemon runs in a separate Java process.

### Configuration

Use the following:

conf/core-site.xml:

|  |
| --- |
| <configuration>  <property>  <name>fs.default.name</name>  <value>hdfs://localhost:9000</value>  </property>  </configuration> |

conf/hdfs-site.xml:

|  |
| --- |
| <configuration>  <property>  <name>dfs.replication</name>  <value>1</value>  </property>  </configuration> |

conf/mapred-site.xml:

|  |
| --- |
| <configuration>  <property>  <name>mapred.job.tracker</name>  <value>localhost:9001</value>  </property>  </configuration> |

### Setup passphraseless ssh

Now check that you can ssh to the localhost without a passphrase:

$ ssh localhost

If you cannot ssh to localhost without a passphrase, execute the following commands:

$ ssh-keygen -t dsa -P '' -f ~/.ssh/id\_dsa

$ cat ~/.ssh/id\_dsa.pub >> ~/.ssh/authorized\_keys

### Execution

Format a new distributed-filesystem:

$ bin/hadoop namenode -format

Start the hadoop daemons:

$ bin/start-all.sh

The hadoop daemon log output is written to the ${HADOOP\_LOG\_DIR} directory (defaults to ${HADOOP\_HOME}/logs).

Browse the web interface for the NameNode and the JobTracker; by default they are available at:

* NameNode - http://localhost:50070/
* JobTracker - http://localhost:50030/

Copy the input files into the distributed filesystem:

$ bin/hadoop fs -put conf input

Run some of the examples provided:

$ bin/hadoop jar hadoop-examples-\*.jar grep input output 'dfs[a-z.]+'

Examine the output files:

Copy the output files from the distributed filesystem to the local filesytem and examine them:

$ bin/hadoop fs -get output output

$ cat output/\*

or

View the output files on the distributed filesystem:

$ bin/hadoop fs -cat output/\*

When you're done, stop the daemons with:

$ bin/stop-all.sh

# Installing CDH 5 with YARN on a Single Linux Node in Pseudo-distributed mode

Please refer to [this URL](http://www.cloudera.com/documentation/enterprise/5-3-x/topics/cdh_qs_yarn_pseudo.html) for more detail

# Templates and References

* https://hadoop.apache.org/docs/r1.2.1/single\_node\_setup.html
* http://www.cloudera.com/documentation/enterprise/5-3-x/topics/cdh\_qs\_yarn\_pseudo.html