Installation and configuration of Hadoop

Rohan Baghel

Indian Institute of Technology Jammu

2021pcs1025@iitjammu.ac.in

July 21, 2025

Overview

1. What is Hadoop?
2. Prerequisites
3. Install Hadoop
4. Configuration in .bashrc
5. Set JAVA_HOME Path
6. Configuration in core-site.xml
7. Configuration in hdfs-site.xml
8. Configuration in mapred-site.xml
9. Configuration in yarn-site.xml
10. Format HDFS & Start Hadoop Cluster

What is Hadoop?

Hadoop is an open source software programming framework for storing a large amount of data and performing the computation. Its framework is based on Java programming with some native code in C and shell scripts.Apache Software Foundation is the developers of Hadoop, and it's co-founders are Doug Cutting and Mike Cafarella.

Prerequisites

- VIRTUAL BOX: it is used for installing the operating system on it. (we don't need if we already have any Linux system.)
- OPERATING SYSTEM: You can install Hadoop on Linux-based operating systems. Ubuntu and Linux-mint are very commonly used. In this tutorial, we are using Linux-mint.
- JAVA: You need to install the Java 8 package on your system.
  TO install java use this command in terminal
  sudo apt install openjdk-8-jdk
  TO check java version cmd is : **java -version**

Prerequisites

- Configure pass-wordless SSH authentication for the local system.

  a. run the following command to generate Public and Private Key Pairs:

  ssh-keygen -t rsa

  b. **cat /.ssh/id_rsa.pub >> /.ssh/authorized_keys**

  c. **chmod 640 /.ssh/authorized_keys**

  d. verify the pass-wordless SSH authentication with the following command: **ssh localhost**

Install Hadoop

- Download the Hadoop 3.3.0 Package.
  Command: wget `https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz`
- Extract the Hadoop tar File. command: tar -xvzf hadoop-3.3.0.tar.gz
- Rename hadoop-3.3.0.tar.gz as hadoop for ease of use.
  command: mv hadoop-3.3.0 hadoop

Configuration in .bashrc

- Add the Hadoop and Java paths in the bash file (.bashrc).Open .bashrc file using command: **vi .bashrc Path** and then add path in .bashrc file as :

```
export JAVA_HOME=/usr/lib/jvm
/ java-1.8.0-openjdk-amd64/
export HADOOP_HOME=/home/username/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
```

Configuration in .bashrc

```
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME
/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME
/bin
export HADOOP_OPTS=”-Djava.library.path=$HADOOP_HOME
/lib/native”
```

- Then, save the bashrc file and close it.
- For applying all these changes to the current Terminal, execute the source command.
  Command: source .bashrc

Set JAVA_HOME Path

- Open the **hadoop-env.sh** file in the nano editor. This file is located in /hadoop/etc/hadoop (Hadoop configuration directory).
  command :nano hadoop-env
- Now, Set JAVA_HOME path:
  export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64/
  NOTE:To save the changes you've made, press **Ctrl+O**. To exit the nano editor, press **Ctrl+X** and then press **'Y'** to exit the editor.
- Now Make two directory using terminal
  1. mkdir -p /hadoopdata/hdfs/namenode
  2. mkdir -p /hadoopdata/hdfs/datanode

Configuration in core-site.xml

- Open the core-site.xml file in the nano editor. This file is also located in the /hadoop/etc/hadoop (Hadoop configuration directory).
  command to open: nano core-site.xml
- Add the following configuration properties:
  $< configuration >$
  $< property >$
  $< name > fs.default.name < /name >< value > hdfs : //localhost : 9000 < /value >$
  $< /property >$
  $< /configuration >$

Configuration in hdfs-site.xml

- Open the hdfs-site.xml file in the nano editor. This file is also located in /hadoop/etc/hadoop (Hadoop configuration directory):
  Command :nano hdfs-site.xml
- Add the following configuration properties and save it:

  $< configuration >$

  $< property >$

  $< name > dfs.replication < /name >< value > 1 < /value ><$
  $/property >< property >< name > dfs.name.dir < /name ><$
  $value > file :$
  $///home/username/hadoop/hadoopdata/hdfs/namenode <$
  $/value >< /property >< property >< name > dfs.data.dir <$
  $/name >< value > file :$
  $///home/username/hadoop/hadoopdata/hdfs/datanode < /value >$
  $< /property >$

  $< /configuration >$

Configuration in mapred-site.xml

- Open the mapred-site.xml file in the nano editor. This file is also located in /hadoop/etc/hadoop (Hadoop configuration directory). Command :nano mapred-site.xml
- Add the following configuration properties and save it:
  $< configuration >$
  $< property >$
  $< name > mapreduce.framework.name < /name >< value > yarn <$
  $/value >< /property >< property >< name >$
  $yarn.app.mapreduce.am.env < /name >< value >$
  $HADOOP\_MAPRED_HOME = \$HADOOP\_HOME < /value ><$
  $/property >< property >< name > mapreduce.map.env <$
  $/name >< value > HADOOP\_MAPRED\_HOME =$
  $\$HADOOP\_HOME < /value >< /property >< property ><$
  $name > mapreduce.reduce.env < /name >< value >$
  $HADOOP_MAPRED_HOME = \$HADOOP\_HOME < /value >$
  $< /property >$
  $< /configuration >$

Configuration in yarn-site.xml

- Open the yarn-site.xml file in the nano editor. This file is also located in /hadoop/etc/hadoop (Hadoop configuration directory).
  Command :nano yarn-site.xml
- Add the following configuration properties and save it:

```
< configuration >
< property >
< name > yarn.nodemanager.aux − services < /name >< value >
mapreduce_shuffle < /value >
< /property >
< /configuration >
```

Format HDFS & Start Hadoop Cluster

- Before starting Hadoop, we need to format HDFS, which can be done using the given command: hdfs namenode -format
- To start the Hadoop cluster we start some services using below command :
  1.Start the HDFS services: start-dfs.sh
  2.Now start the yarn services: start-yarn.sh
  NOTE : The **'jps'** command is used to check whether all the Hadoop processes are running or not.
- Now open the below link for access
  a.`http://localhost:9870`
  b.`http://localhost:9870`
- Congratulations, you have successfully installed a single-node Hadoop cluster.

Thank You!