Install Hadoop: Setting up a Single Node Hadoop Cluster

Shubham Sinha (https://www.edureka.co/blog/author/shubham-sinha/) | Jul 16,2018 |

Install Hadoop: Setting up a Single Node Hadoop Cluster

From our previous blogs on ***Hadoop Tutorial Series (https://www.edureka.co/blog/hadoop-tutorial/)***, you must have got a theoretical idea about Hadoop, HDFS and its architecture. I hope you would have liked our previous blog on ***HDFS Architecture (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/)***, now I will take you through the practical knowledge about Hadoop and HDFS. The first step forward is to install Hadoop.

There are two ways to install Hadoop, i.e. **Single node** and **Multi node**.

Single node cluster means only one DataNode running and setting up all the NameNode, DataNode, ResourceManager and NodeManager on a single machine. This is used for studying and testing purposes. For example, let us consider a sample data set inside a healthcare industry. So, for testing whether the Oozie jobs have scheduled all the processes like collecting, aggregating, storing and processing the data in a proper sequence, we use single node cluster. It can easily and efficiently test the sequential workflow in a smaller environment as compared to large environments which contains terabytes of data distributed across hundreds of machines.

While in a **Multi node cluster**, there are more than one DataNode running and each DataNode is running on different machines. The multi node cluster is practically used in organizations for analyzing Big Data. Considering the above example, in real time when we deal with petabytes of data, it needs to be distributed across hundreds of machines to be processed. Thus, here we use multi node cluster.

In this blog, I will show you how to install Hadoop on a single node cluster.

Prerequisites

- *VIRTUAL BOX*: it is used for installing the operating system on it.
- *OPERATING SYSTEM*: You can install Hadoop on Linux based operating systems. Ubuntu and CentOS are very commonly used. In this tutorial, we are using CentOS.
- *JAVA*: You need to install the Java 8 package on your system.
- *HADOOP*: You require Hadoop 2.7.3 package.

Install Hadoop

Step 1: Click here (https://goo.gl/ipdJJa) to download the Java 8 Package. Save this file in your home directory.

Step 2: Extract the Java Tar File.

***Command*:** tar -xvf jdk-8u101-linux-i586.tar.gz

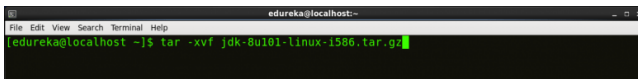


Fig: Hadoop Installation – Extracting Java Files

Step 3: Download the Hadoop 2.7.3 Package.

***Command*:** wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz

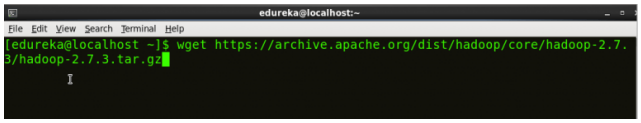


Fig: Hadoop Installation – Downloading Hadoop

Step 4: Extract the Hadoop tar File.

***Command*:** tar -xvf hadoop-2.7.3.tar.gz

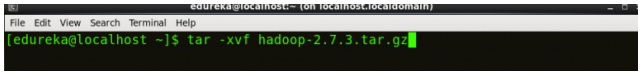


Fig: Hadoop Installation – Extracting Hadoop Files

Step 5: Add the Hadoop and Java paths in the bash file (.bashrc).

Open**. bashrc** file. Now, add Hadoop and Java Path as shown below.

***Command*:** vi .bashrc

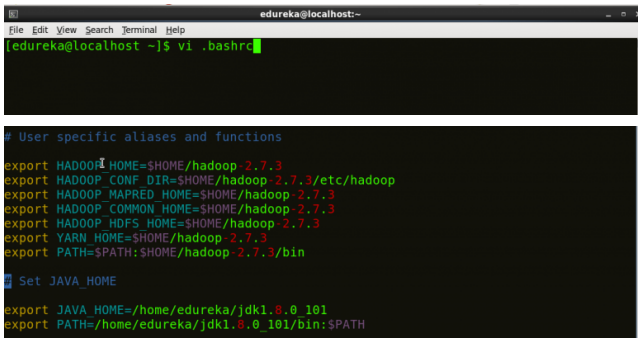


Fig: Hadoop Installation – Setting Environment Variable

Then, save the bash file and close it.

For applying all these changes to the current Terminal, execute the source command.

***Command*:** source .bashrc

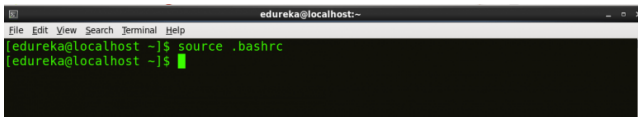


Fig: Hadoop Installation – Refreshing environment variables

To make sure that Java and Hadoop have been properly installed on your system and can be accessed through the Terminal, execute the java -version and hadoop version commands.

***Command*:** java -version

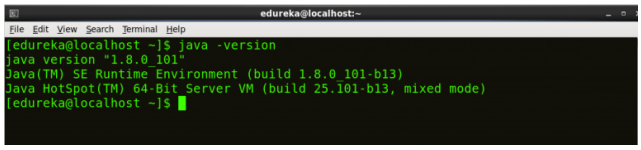


Fig: Hadoop Installation – Checking Java Version

***Command*:** hadoop version

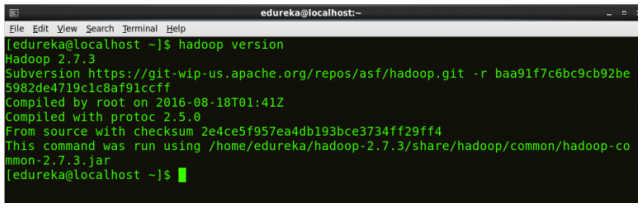


Fig: Hadoop Installation – Checking Hadoop Version

Step 6: Edit the **Hadoop Configuration files** (https://www.edureka.co/blog/explaining-hadoop-configuration/).

Command: cd hadoop-2.7.3/etc/hadoop/

Command: ls

All the Hadoop configuration files are located in **hadoop-2.7.3/etc/hadoop** directory as you can see in the snapshot below:

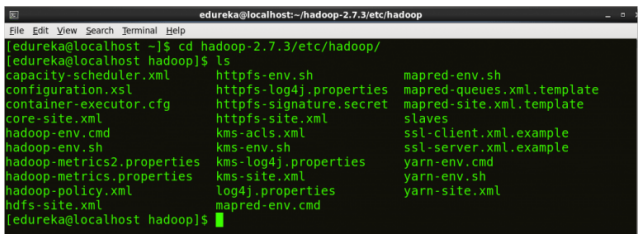


Fig: Hadoop Installation – Hadoop Configuration Files

Step 7: Open *core-site.xml* and edit the property mentioned below inside configuration tag:

core-site.xml informs Hadoop daemon where NameNode runs in the cluster. It contains configuration settings of Hadoop core such as I/O settings that are common to HDFS & MapReduce.

***Command*:** vi core-site.xml

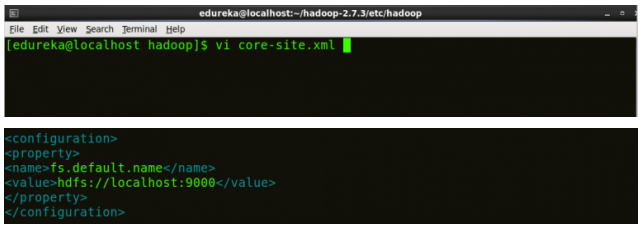


Fig: Hadoop Installation – Configuring core-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000 (hdfs://localhost:9000)</value>
</property>
</configuration>
```

Step 8: Edit *hdfs-site.xml* and edit the property mentioned below inside configuration tag:

hdfs-site.xml contains configuration settings of HDFS daemons (i.e. NameNode, DataNode, Secondary NameNode). It also includes the replication factor and block size of HDFS.

***Command*:** vi hdfs-site.xml

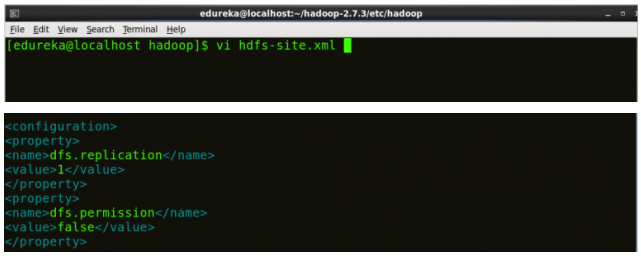


Fig: Hadoop Installation – Configuring hdfs-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.permission</name>
<value>false</value>
</property>
</configuration>
```

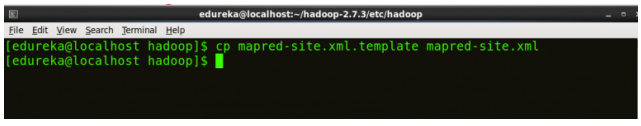
Step 9: Edit the *mapred-site.xml* file and edit the property mentioned below inside configuration tag:

mapred-site.xml contains configuration settings of MapReduce application like number of JVM that can run in parallel, the size of the mapper and the reducer process, CPU cores available for a process, etc.

In some cases, mapred-site.xml file is not available. So, we have to create the mapred-site.xml file using mapred-site.xml template.

***Command*:** cp mapred-site.xml.template mapred-site.xml

***Command*:** vi mapred-site.xml.



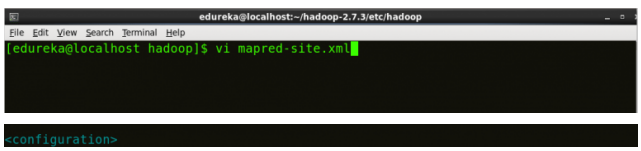


Fig: Hadoop Installation – Configuring mapred-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

Step 10: Edit *yarn-site.xml* and edit the property mentioned below inside configuration tag:

yarn-site.xml contains configuration settings of ResourceManager and NodeManager like application memory management size, the operation needed on program & algorithm, etc.

***Command*:** vi yarn-site.xml

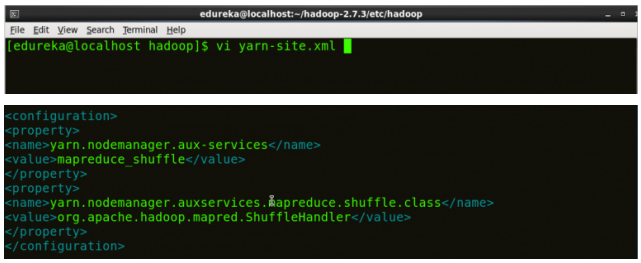


Fig: Hadoop Installation – Configuring yarn-site.xml

```
<?xml version="1.0">
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
</configuration>
```

Step 11: Edit *hadoop-env.sh* and add the Java Path as mentioned below:

hadoop-env.sh contains the environment variables that are used in the script to run Hadoop like Java home path, etc.

***Command*:** vi hadoop–env.sh

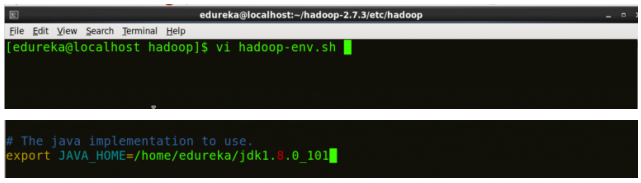


Fig: Hadoop Installation – Configuring hadoop-env.sh

Step 12: Go to Hadoop home directory and format the NameNode.

***Command*:** cd

***Command*:** cd hadoop-2.7.3

***Command*:** bin/hadoop namenode -format

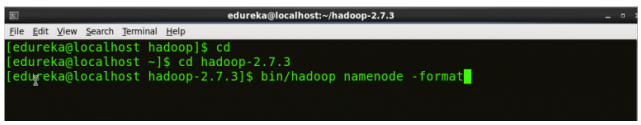


Fig: Hadoop Installation – Formatting NameNode

This formats the HDFS via NameNode. This command is only executed for the first time. Formatting the file system means initializing the directory specified by the dfs.name.dir variable.

Never format, up and running Hadoop filesystem. You will lose all your data stored in the HDFS.

Step 13: Once the NameNode is formatted, go to hadoop-2.7.3/sbin directory and start all the daemons.

Command: cd hadoop-2.7.3/sbin

Either you can start all daemons with a single command or do it individually.

Command: ./start-all.sh

The above command is a combination of ***start-dfs.sh, start-yarn.sh*** & ***mr-jobhistory-daemon.sh***

Or you can run all the services individually as below:

Start NameNode:

The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files stored in the HDFS and tracks all the file stored across the cluster.

Command: ./hadoop-daemon.sh start namenode

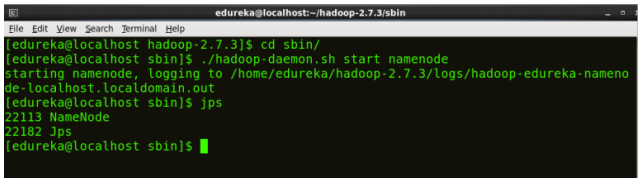


Fig: Hadoop Installation – Starting NameNode

Start DataNode:

On startup, a DataNode connects to the Namenode and it responds to the requests from the Namenode for different operations.

Command: ./hadoop-daemon.sh start datanode

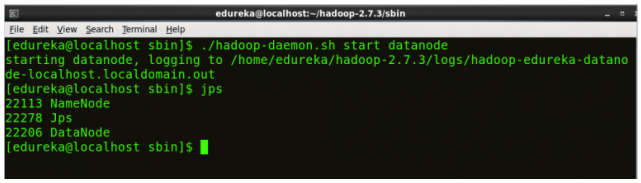


Fig: Hadoop Installation – Starting DataNode

Start ResourceManager:

ResourceManager is the master that arbitrates all the available cluster resources and thus helps in managing the distributed applications running on the YARN system. Its work is to manage each NodeManagers and the each application's ApplicationMaster.

Command: ./yarn-daemon.sh start resourcemanager

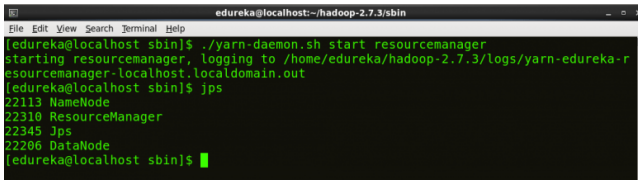


Fig: Hadoop Installation – Starting ResourceManager

Start NodeManager:

The NodeManager in each machine framework is the agent which is responsible for managing containers, monitoring their resource usage and reporting the same to the ResourceManager.

Command: ./yarn-daemon.sh start nodemanager

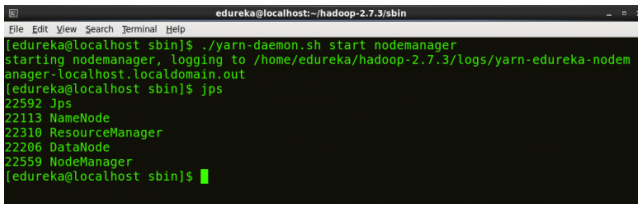


Fig: Hadoop Installation – Starting NodeManager

Start JobHistoryServer:

JobHistoryServer is responsible for servicing all job history related requests from client.

***Command*:** ./mr-jobhistory-daemon.sh start historyserver

Step 14: To check that all the Hadoop services are up and running, run the below command.

Command: jps

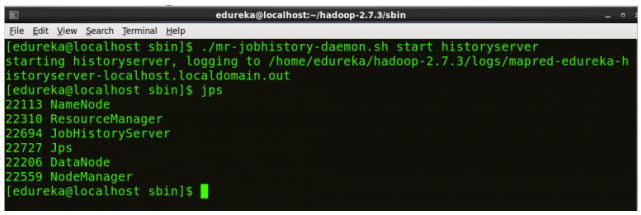


Fig: Hadoop Installation – Checking Daemons

Step 15: Now open the Mozilla browser and go to **localhost**:**50070/dfshealth.html** to check the NameNode interface.

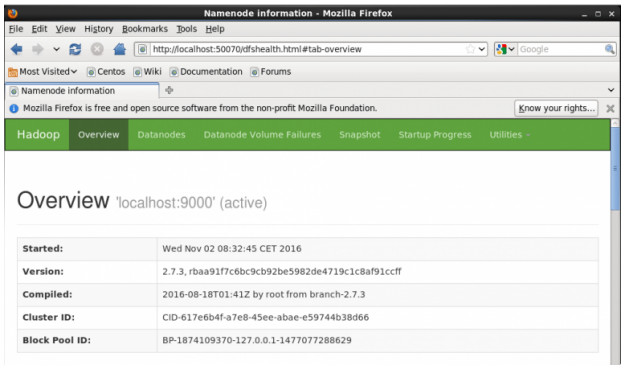


Fig: Hadoop Installation – Starting WebUI

Congratulations, you have successfully installed a single node Hadoop cluster in one go. In our next blog of ***Hadoop Tutorial Series (https://www.edureka.co/blog/hadoop-tutorial/)***, we will be covering how to install Hadoop on a multi node cluster as well.

*Now that you have understood how to install Hadoop, check out the **Hadoop training (https://www.edureka.co/big-data-and-hadoop/)** by Edureka, a trusted online learning company with a network of more than 250,000 satisfied learners spread across the globe. The Edureka Big Data Hadoop Certification Training course helps learners become expert in HDFS, Yarn, MapReduce, Pig, Hive, HBase, Oozie, Flume and Sqoop using real-time use cases on Retail, Social Media, Aviation, Tourism, Finance domain.*

Got a question for us? Please mention it in the comments section and we will get back to you.

About Shubham Sinha (25 Posts (https://www.edureka.co/blog/author/shubham-sinha/))

Shubham Sinha is a Big Data and Hadoop expert working as a Research Analyst at Edureka. He is keen to work with Big Data related technologies such as Hadoop, Spark, Flink and Storm and web development technologies including Angular, Node.js & PHP.

PREVIOUS NEXT

Related Posts

Big Data Tutorial: All You Need To Know About Big Data!
37.7K
(https://www.edureka.co/blog/big-data-tutorial)

Hadoop Tutorial: All you need to know about Hadoop!
87.9K
(https://www.edureka.co/blog/hadoop-tutorial/)

HDFS Tutorial: Introduction to HDFS & its Features
35.2K
(https://www.edureka.co/blog/hadoop-tutorial)

Hadoop Ecosystem: Hadoop Tools for Crunching Big Data
44.8K
(https://www.edureka.co/blog/hadoop-ecosystem)

Comments

9 Comments

10 Comments https://www.edureka.co/blog/ Rajiv Chaudhuri

Recommend Share Sort by Best

Join the discussion…

Ritesh • 2 months ago
Can be install hadoop on Win10 machine and use python as base languagae.
1 • Reply • Share ›

Carlos Arbeláez • 8 months ago
Hi everyone, I have some troubles running "vi .bashrc" command. When I use it the console does not show as it looks in the image. I'm newbie with linux so I do not know how to proceed, i'll be really gratefull for help
Reply • Share ›

eniolorunmo • 10 months ago
Thanks Mr Sinha, I have installed a single node cluster, but my job get stuck saying "Accepted, waiting for AM container to allocate, launched and registered with RM", I have checked and edited configuration files yet I dont seem to find a way around this problem. Please help
Reply • Share ›

bhavya pruthi • 10 months ago
where are we typing this
Command: tar -xvf jdk-8u101-linux-i586.tar.gz
Reply • Share ›

Shubham Sinha ➚ bhavya pruthi • 10 months ago
We are typing this command in the directory where the Java file has been downloaded.
Reply • Share ›

Kiran Edu • a year ago
hi,

I am on Win 10. Plz suggest what is the best option for Virtual Box?
Reply • Share ›

Shubham Sinha ➚ Kiran Edu • 10 months ago
I would suggest you to use Oracle Virtual Box.
Reply • Share ›

bhavya pruthi ➚ Kiran Edu • 10 months ago
i am facing the same problem
plz help
Reply • Share ›

deepak • a year ago
I am facing problem when trying to run ./start-all.sh. I am getting permisson denied error.

[edureka@localhost sbin]$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
17/04/07 10:11:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
edureka@localhost's password:
localhost: starting namenode, logging to /usr/lib/hadoop-2.2.0/logs/hadoop-edureka-namenode-localhost.localdomain.out
localhost: /usr/lib/hadoop-2.2.0/sbin/hadoop-daemon.sh: line 157: /usr/lib/hadoop-2.2.0/hadoop2_data/hdfs/pid/hadoop-edureka-namenode.pid: Permission denied
edureka@localhost's password:
localhost: starting datanode, logging to /usr/lib/hadoop-2.2.0/logs/hadoop-edureka-datanode-localhost.localdomain.out
localhost: /usr/lib/hadoop-2.2.0/sbin/hadoop-daemon.sh: line 157: /usr/lib/hadoop-2.2.0/hadoop2_data/hdfs/pid/hadoop-edureka-datanode.pid: Permission denied
Starting secondary namenodes [0.0.0.0]
edureka@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/lib/hadoop-2.2.0/logs/hadoop-edureka-secondarynamenode-localhost.localdomain.out

0.0.0.0: /usr/lib/hadoop-2.2.0/sbin/hadoop-daemon.sh: line 157: /usr/lib/hadoop-2.2.0/hadoop2_data/hdfs/pid/hadoop-edureka-secondarynam...: Permission denied
17/04/07 10:12:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons

see more

• Reply • Share ›

EdurekaSupport Mod ➔ deepak • a year ago

Hey Deepak, thanks for checking out our blog. Here's what you need to do:
Use : sudo ./start-all.sh Then Use : sudo jps

If you are using edureka VM the follow the below steps.

==>Open the terminal and fire reboot.

==> Now when the VM start fire sudo service hadoop-master restart

==> sudo jps.

And if you still want to start from sbin the first got to root

=> su (fire this command in the terminal and when it prompts for password give edureka as password.)
Hope this helps. Cheers!

• Reply • Share ›

Related Blogs


(https://www.
data-tutorial)

Big Data Tutorial: All You Need To Know About Big... (https://www.edureka.co/blog/big-data-tutorial)


(https://www.
tutorial/)

Hadoop Tutorial: All you need to know about Hadoop! (https://www.edureka.co/blog/hadoop-tutorial/)


(https://www.
tutorial)

HDFS Tutorial: Introduction to HDFS & its Features (https://www.edureka.co/blog/hdfs-tutorial)


(https://www.
ecosystem)

Hadoop Ecosystem: Hadoop Tools for Crunching Big Data (https://www.edureka.co/blog/hadoop-ecosystem)