Cloud Application Development

Week -8

WEEK-8: DATA INTENSIVE PROGRAMMING Install Hadoop single node cluster and run simple applications like word count

You must have got a theoretical idea about Hadoop, HDFS and its architecture. But to get Hadoop Certified you need good hands-on knowledge. I hope you would have liked our previous blog on ***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**.

A 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 the 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 a single node cluster. It can easily and efficiently test the sequential workflow in a smaller environment as compared to large environments which contain 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 a multi-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 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

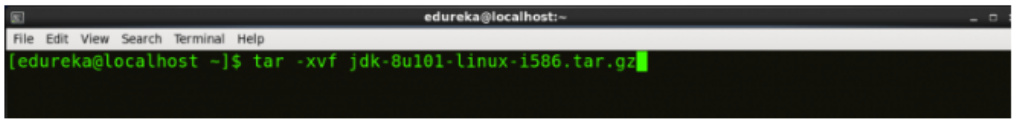

Fig: Hadoop Installation – Extracting Java Files

Step 3: Download the Hadoop 2.7.3 Package.

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

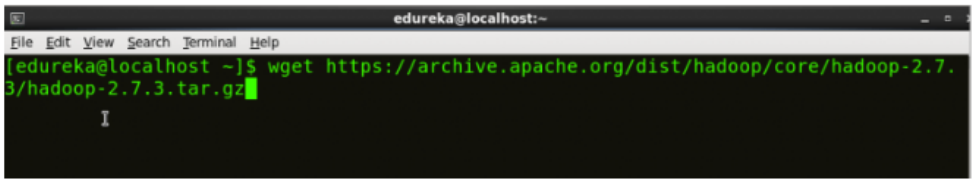

Fig: Hadoop Installation – Downloading Hadoop

Step 4: Extract the Hadoop tar File.

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

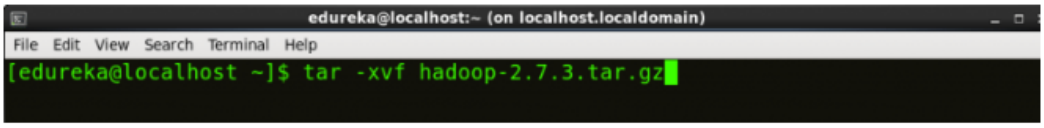

Fig: Hadoop Installation – Extracting Hadoop Files

Step 4: Extract the Hadoop tar File.

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

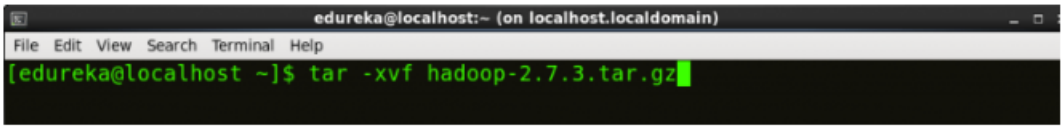

Fig: Hadoop Installation – Extracting Hadoop Files

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.

Learn more about the Hadoop Ecosystem and its tools with the Hadoop Certification.

***Command*:** vi .bashrc

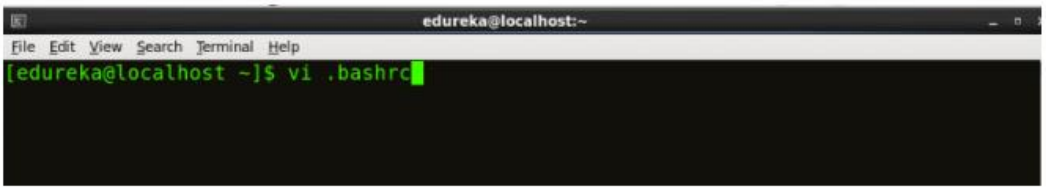

Fig: Hadoop Installation – Setting Environment Variable

Then, save the bash file and close it.

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

***Command*:** source .bashrc

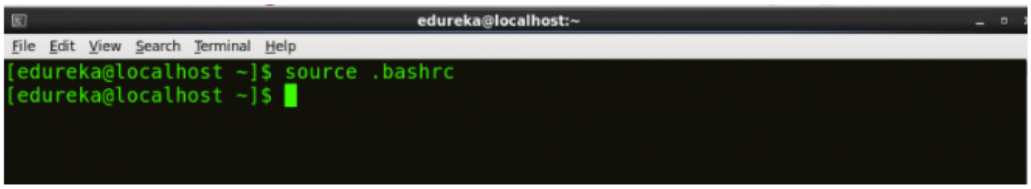

Fig: Hadoop Installation – Refreshing environment variables

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

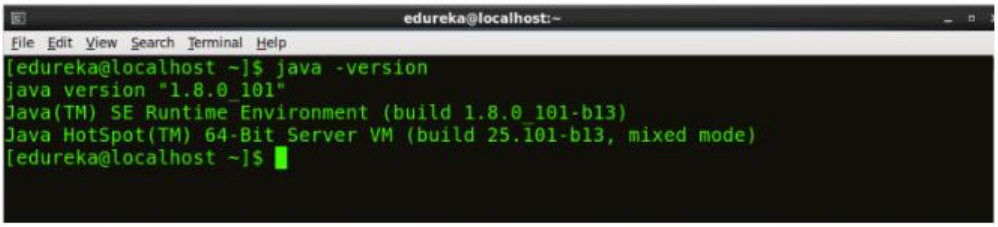

Fig: Hadoop Installation – Checking Java Version

***Command*:** hadoop version

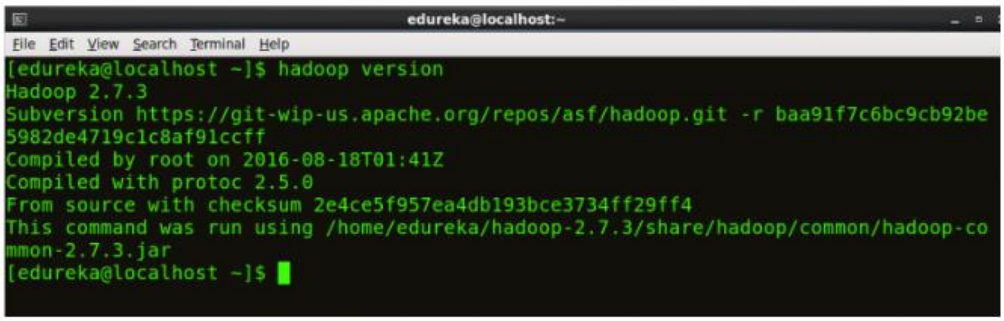

Fig: Hadoop Installation – Checking Hadoop Version

Step 6: Edit the **Hadoop Configuration files**.

Command: cd hadoop-2.7.3/etc/hadoop/

Command: ls