Installation of Apache Hadoop in Standalone Operation

> By default, Hadoop is configured to run in a non-distributed mode, as a single Java process. This is useful for debugging.

We will use the official installation documentation from Apache Hadoop

SSH

SSH into your machine with `ssh student@bdlc-XX.el.eee.intern`, where `XX` is your personal virtual machine number.

Update and Upgrade your Software

Before we start, let's upgrade the software

```
sudo apt update
# answer with Y on: Do you want to continue? [Y/n] Y
sudo apt upgrade
```

Hadoop User

We will create a user `hadoop` and won't user our `student` account. All Hadoop processes will run under the user `hadoop`.

```
# create a new user called 'hadoop'
# give it a new password (and also remember it ;)) and confirm with Y

sudo adduser hadoop
```

The command `sudo adduser hadoop` will look like:

```
# [sudo] password for student:
# Adding user `hadoop' ...
# Adding new group `hadoop' (1004) ...
# Adding new user `hadoop' (1004) with group `hadoop' ...
# Creating home directory `/home/hadoop' ...
# Copying files from `/etc/skel' ...
# New password:
# Retype new password:
# passwd: password updated successfully
# Changing the user information for hadoop
# Enter the new value, or press ENTER for the default
#   Full Name []: Hadoop User
#   Room Number []:
```

```
#   Work Phone []:
#   Home Phone []:
#   Other []:
# Is the information correct? [Y/n] Y
```

Now, we will switch to this user. All following steps should be executed as user `hadoop`.

```
# switch to the user hadoop
su - hadoop

# change to the home directory
cd ~

# verify "who you are" with
whoami
```

Java

I have already preinstalled the correct java version. In other words, you don't have to take care about the java version. For reference, I did:

```
sudo apt install openjdk-11-jdk
```

To verify that we have indeed the version `11.0.13`, run:

```
java --version
```

Following the Apache Hadoop Guide

If not done already, open the official installation guide.

```
# install pdsh
# answer with Y on: Do you want to continue? [Y/n] Y
sudo apt-get install pdsh

# download the latest hadoop distribution
wget https://downloads.apache.org/hadoop/common/hadoop-3.3.1/hadoop-
3.3.1.tar.gz

# check with ls that the file is around
# you should see: hadoop-3.3.1.tar.gz
ls

# extract hadoop-3.3.1.tar.gz
```

```
tar -xvzf hadoop-3.3.1.tar.gz

# check again with ls if you see the folder hadoop-3.3.1
ls

# rename the folder to hadoop
mv hadoop-3.3.1 hadoop
```

We need to specify `JAVA_HOME` in `~/hadoop/etc/hadoop/hadoop-env.sh`. To figure out where java's home is, run:

```
dirname $(dirname $(readlink -f $(which java)))
```

and copy the output (`/usr/lib/jvm/java-11-openjdk-amd64`).

```
change the line from
# export JAVA_HOME=
```

to

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

with

```
nano ~/hadoop/etc/hadoop/hadoop-env.sh
```

Run Apache Hadoop in Standalone Operation

Let's start to run Hadoop. Check if we get the usage help when we type `~/hadoop/bin/hadoop`. Note, it should execute without an error.

```
~/hadoop/bin/hadoop

# should give you the usage help..
#
# Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
#  or    hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
#   where CLASSNAME is a user-provided Java class
# ...
```

✨ congrats ✨ - you finished the standalone operation installation.

Examples

Try to run the example provided in the documentation.

```
mkdir ~/input
cp ~/hadoop/etc/hadoop/*.xml ~/input
ls -al ~/input

~/hadoop/bin/hadoop jar ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-
examples-3.3.1.jar grep ~/input ~/output 'dfs[a-z.]+'
```

Can you explain what the program does?

Homework

Check the Homework

References

- Hadoop - Single Cluster
- Tutorial - Hadoop on Single Node
- Tutorial - Hadoop on Multi Node