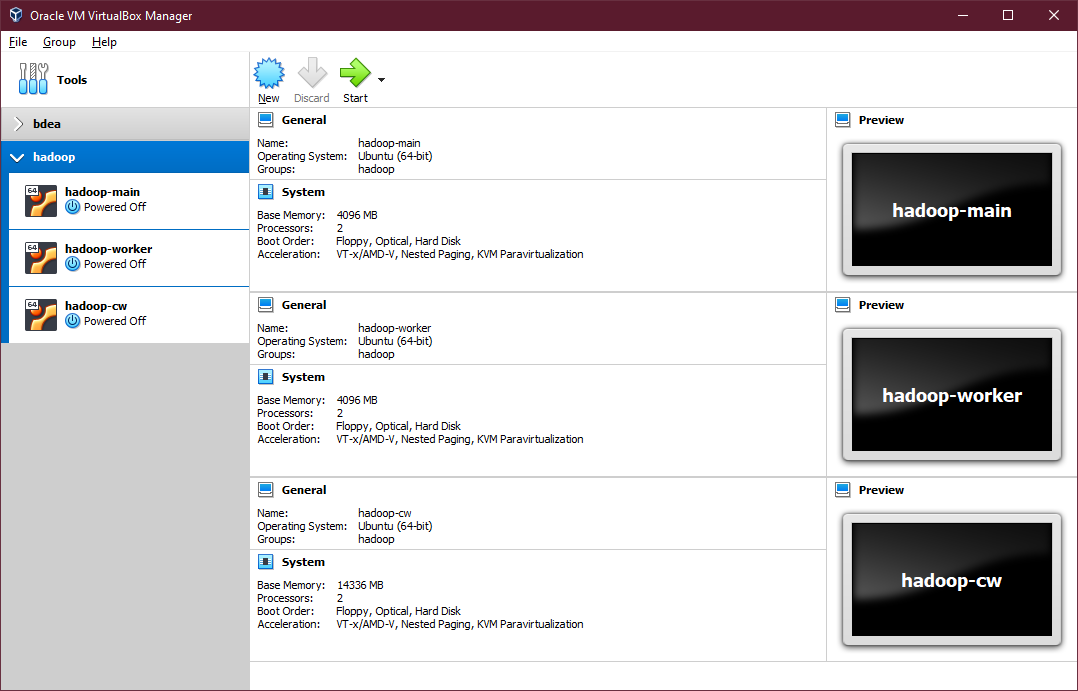
MS4S21 CW-1

Mark Baber – 17076749

# 1 – Experiment 1

This section will detail what was carried out for experiment 1, which was to create a Hadoop cluster using Ubuntu Virtual Machines (VMs) on our local machines. The main task was to replace the existing worker node with a new worker node with the following specifications:



(Figure 1 – Screenshot of the 3 node hadoop cluster)

Seeing as I ran into some issues with my existing cluster from the tutorials, I took a day to re-create the whole cluster again following the notes from the tutorials which I manipulated for myself. The next few steps will be the sequence of events which happened from start to finish, with a working 3 node cluster.

## – Hadoop-Main

After installing Ubuntu Desktop 20.04 Long Term Support ([Instruction Here](https://ubuntu.com/tutorials/install-ubuntu-desktop#1-overview)), I installed the updates and openssh-server so I could ssh into the main node from my computer using [Windows Terminal](#_x.2_-_Windows). This would allow me to copy and paste some of the commands I initially edited during the tutorials, to speed up my process of setting up the VMs.

### 1.1.1 – Set Hostname

I checked the hostname on the main node to make sure it was correctrly assigned as: **hadoop-main**. This was done by editing the file with ***sudo nano /etc/hostname*** and deleting the hostname and changing it to **hadoop-main**.

### 1.1.2 – Set Hosts

Next was to assign the hosts correctly for the main node and the two worker nodes – I already knew I was going to set a static IP as it is good practice, so these were hard coded in a range which I knew was free on my local network. Setting the hosts file was done by typing into a terminal, **sudo nano /etc/hosts** and adding the following:

* 127.0.0.1 localhost
* #127.0.1.1 hadoop-xyz (this would be VirtualMachine hostname)
* 192.168.1.173 hadoop-main
* 192.168.1.174 hadoop-worker
* 192.168.1.175 hadoop-cw

### 1.1.3 – Download and Install Java

To install java on Linux, this can be done with the follow code: **sudo apt install openjdk-8-jdk** which is an open-source version of java freely distributed. Now that java is installed, the path needs to be assigned to the **~/.bashrc** file. This can be done with the following code: **sudo nano ~/.bashrc** and add the following code at the very bottom of the file:

* # JAVA
* export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre

and apply these changes with this: **source ~/.bashrc** – this is all that is currently needed for setting up java on Linux.

### 1.1.4 – Download and Install Hadoop

This next section will cover how to download and install Hadoop which is required for our clusters, this code can be run from everywhere but I usually change directory to downloads when downloading files. This can be done with:

* **cd Downloads/**
* **wget** [**https://archive.apache.org/dist/hadoop/common/hadoop-3.2.2/hadoop-3.2.2.tar.gz**](https://archive.apache.org/dist/hadoop/common/hadoop-3.2.2/hadoop-3.2.2.tar.gz)

Next will be to extract the archive which can be done with **tar -xvf hadoop-3.2.2.tar.gz** and should be moved to /usr/local/hadoop, this can be done with the following code: **sudo mv hadoop-3.2.2 /usr/local/hadoop** – now that we have moved the hadoop files to our local user files, we would need to add the path to **~/.bashrc** again. This can be done with: **sudo nano ~/.bashrc** and adding the 3 lines to the bottom of the file (under the Java entry)

* # HADOOP
* export PATH=$PATH:/usr/local/hadoop/bin:/usr/local/hadoop/sbin
* export CONF=/usr/local/hadoop/etc/hadoop

and applying the changes again with **source ~/.bashrc** – this will conclude this section.

### 1.1.5 – Cloning the VM

## – Hadoop-Worker

## – Hadoop-CW

Software

## x.1 - VirtualBox – Version: 6.1.18 r142142 (Qt5.6.2)

## x.2 - Windows Terminal - Version: 1.7.1033.0