FIT5202 - Data processing for big data

Activity: Setting up your development environment

In this activity, we will learn how to set up the system and make it ready for processing big data. We will use **Ubuntu 20.04 LTS (Focal Fossa)** as the operating system with the following pre-installed tools and technologies to learn data processing for big data.

1. **Python (3.8)** as a programming language
2. **Jupyter Notebook** as an IDE for python development
3. **Apache Spark (3.0.0)** as a big data processing and analysis tool
4. **Apache Kafka (2.5.0)** as the tool for streaming, and

We will be installing the Ubuntu Image in your own machine using Virtualbox.

Let's get started with the setup.

Using Virtualbox (In personal computer)

Minimum Hardware Requirements for Virtual Machine:

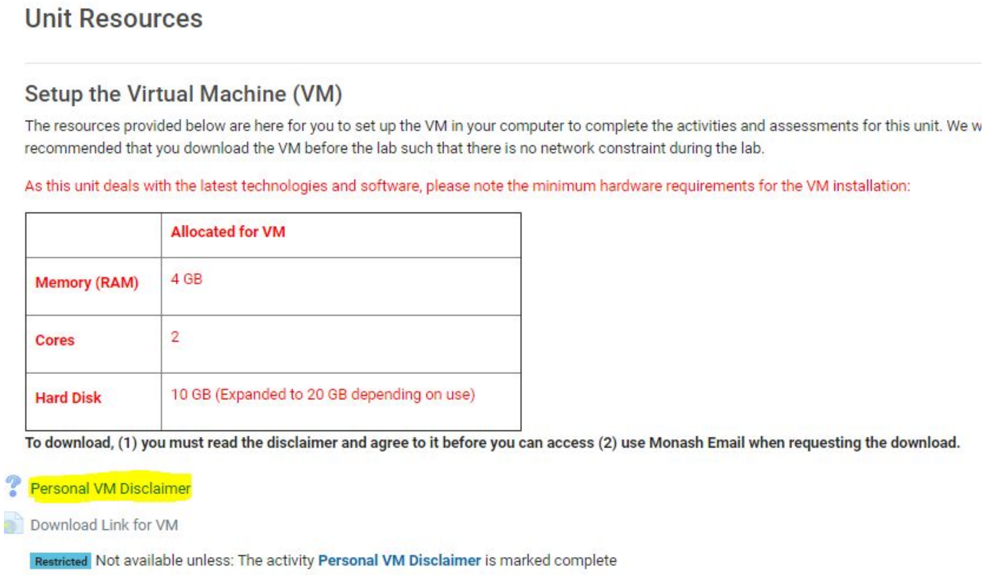
Memory (RAM)	4 GB
Cores	2
Hard Disk	10 GB free space

Step 1: Download and install **Virtualbox 6.1.**

Step 2: Download image file**.**
You need to download the **.ova file** in Moodle, which is a pre-installed version of the latest Ubuntu OS. You can simply import it without having to install it all over again.

Import .OVA File
No installation required (pre-installed)
Download .**ova** file from **Moodle (instructions given below)**
Open Virtualbox
Goto **File -> Import Appliance -> FIT5202 Ubuntu 20 LTS.ova**

In the **Unit Information** Tab in Moodle, please navigate to the bottom of the page. You will find the Unit Resources section with the information and download link to the OVA file as shown in the picture below. **You need to read and accept the VM disclaimer before the link is available for download**.



Unit Resources

Setup the Virtual Machine (VM)

The resources provided below are here for you to set up the VM in your computer to complete the activities and assessments for this unit. We w
recommended that you download the VM before the lab such that there is no network constraint during the lab.

As this unit deals with the latest technologies and software, please note the minimum hardware requirements for the VM installation:

	Allocated for VM
Memory (RAM)	4 GB
Cores	2
Hard Disk	10 GB (Expanded to 20 GB depending on use)

To download, (1) you must read the disclaimer and agree to it before you can access (2) use Monash Email when requesting the download.

Personal VM Disclaimer

Download Link for VM

Restricted Not available unless: The activity **Personal VM Disclaimer** is marked complete

Setting up the Virtual Machine

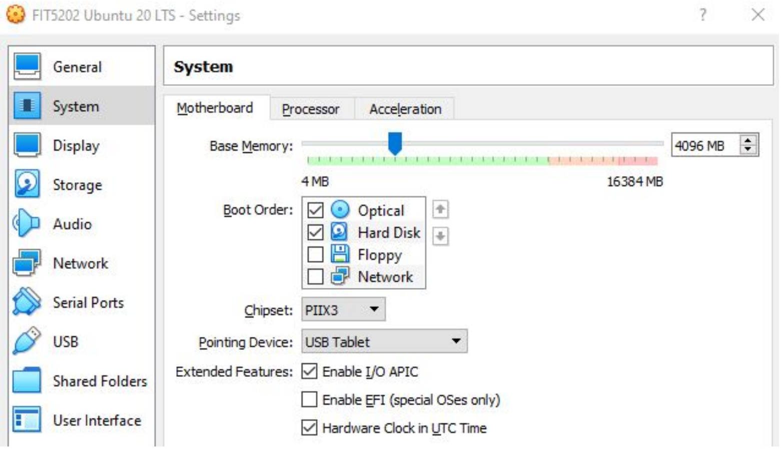
We need to set up the server to use the jupyter notebook in the web browser located in our OS rather than the browser in the virtual machine. We will go through each step to do this. Please note that this should be done after you have imported the .ova image file.

Step 1. Verify System Settings

We first need to verify the resources allocated to the VM.

Note: This should be done without starting the VM.

- Go to Settings -> System -> Motherboard and verify that there is at least 4GB (4096 MB) in Base Memory.



Base Memory: 4096 MB

Boot Order: ☑ Optical ☑ Hard Disk □ Floppy □ Network

Chipset: PIIX3

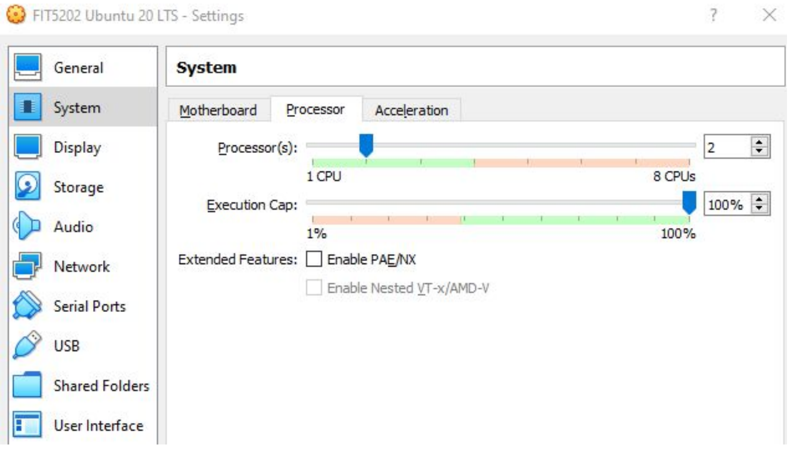
Pointing Device: USB Tablet

Extended Features: ☑ Enable I/O APIC

□ Enable EFI (special OSes only)

☑ Hardware Clock in UTC Time

- Then go to Processor and verify that you have 2 processors allocated for the VM.



Processor(s): 2

Execution Cap: 100%

Extended Features: □ Enable PAE/NX

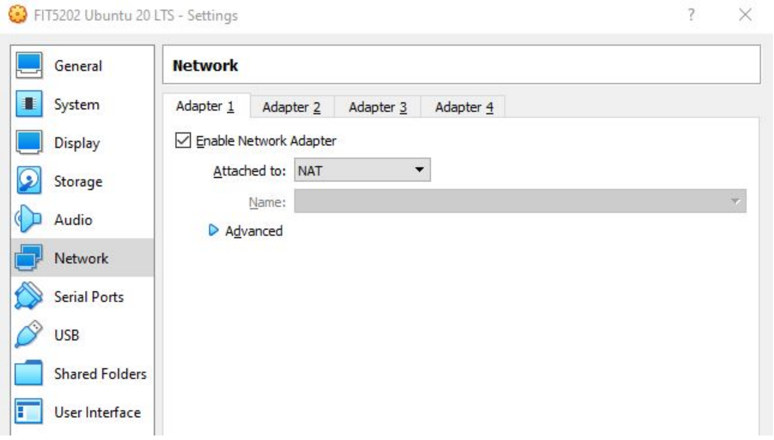
□ Enable Nested VT-x/AMD-V

Step 2: Set Up Additional Network Interface

Note: If you have a Mac, please refer to the last section.

To connect to the jupyter notebook in the VM from your own OS, we need to add a network interface as follows.

- Go to Settings -> Network and it should appear a default adapter in Adapter 1 (NAT) which is enabled.



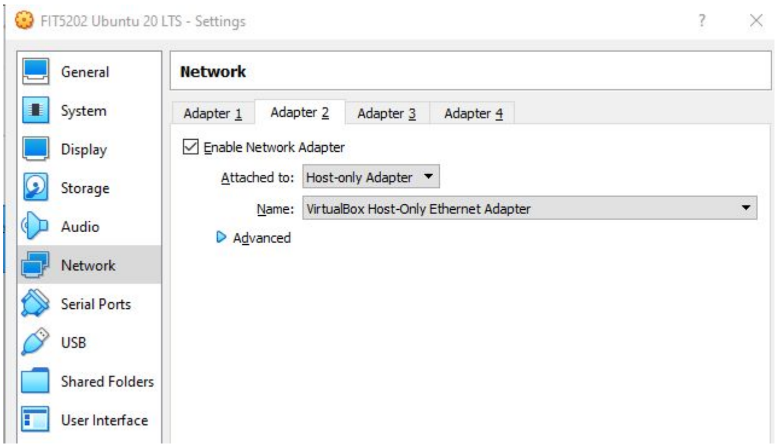
Network

☑ Enable Network Adapter

Attached to: NAT

Name:

- Go to Adapter 2 to add the new interface, select "Enable Network Adapter", choose "Host-only Adapter" and click OK.



Network

☑ Enable Network Adapter

Attached to: Host-only Adapter

Name: VirtualBox Host-Only Ethernet Adapter

Step 3: Get the IP of the New Network Interface

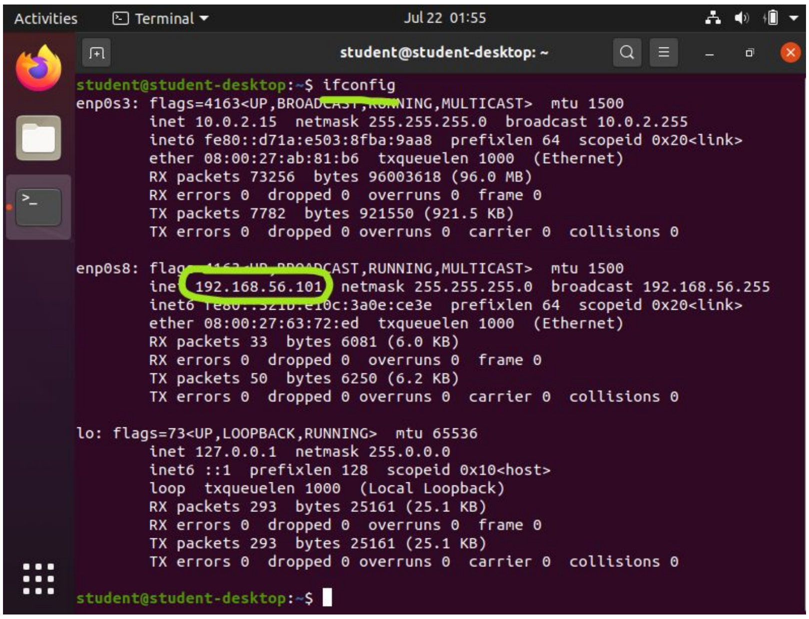
Now that we have a new interface in the VM, we need to find the IP of that interface. For this, start the VM first and access the VM with the following credentials

- Username: student
- Password: student

Then, open the terminal (black button on the left side in Desktop) and execute the following command.

```
ifconfig
```

The picture below shows how it looks in the terminal and the IP needed.



```
student@student-desktop:~$ ifconfig
enp0s3: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 10.0.2.15  netmask 255.255.255.0  broadcast 10.0.2.255
        inet6 fe80::d71a:e503:8fba:9aa8  prefixlen 64  scopeid 0x20<link>
        ether 08:00:27:ab:81:b6  txqueuelen 1000  (Ethernet)
        RX packets 73256  bytes 96003618 (96.0 MB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 7782  bytes 921550 (921.5 KB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

enp0s8: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.56.101  netmask 255.255.255.0  broadcast 192.168.56.255
        inet6 fe80::321b:e10c:3a0e:ce3e  prefixlen 64  scopeid 0x20<link>
        ether 08:00:27:63:72:ed  txqueuelen 1000  (Ethernet)
        RX packets 33  bytes 6081 (6.0 KB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 50  bytes 6250 (6.2 KB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 293  bytes 25161 (25.1 KB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 293  bytes 25161 (25.1 KB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

student@student-desktop:~$
```

The IP needed is the one that is circled. Take note as this will be needed later. For this case, it would be IP: 192.168.56.101.

Step 4: Run Jupyter Notebook in terminal

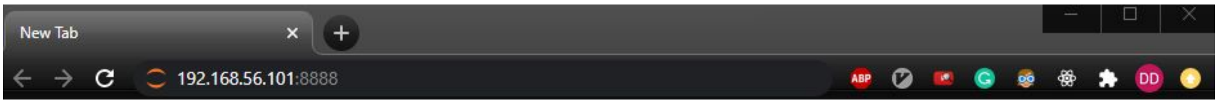
You now have everything you need to run Jupyter Notebook! To run it, execute the following command:

```
jupyter notebook
```

A log of the activities of the Jupyter Notebook will be printed to the terminal. When you run Jupyter Notebook, it runs on a specific port number. The first Notebook you run will usually use port 8888.

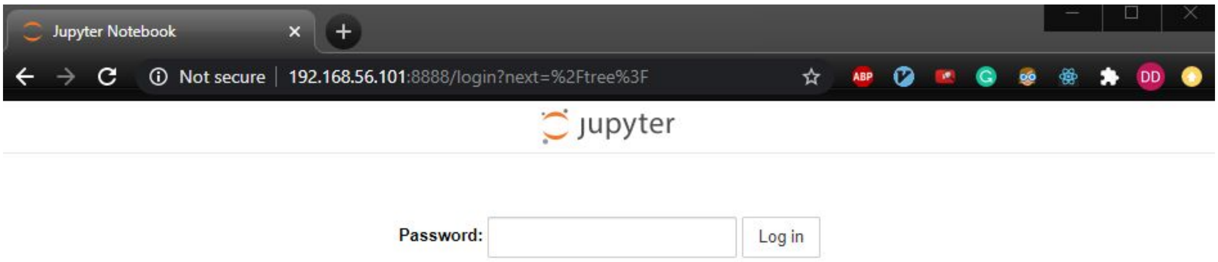
Step 5: Starting Jupyter Notebook in browser

Now that you started jupyter notebook in the VM, go to the browser you use in your own machine (NOT VM browser) and put the following address in the address bar as in the picture. The address will be the IP you obtained in Step 3 with the port (i.e. <yourIP>:8888) REMEMBER that for the example case, the IP was 192.168.56.101.



192.168.56.101:8888

When you access, it will ask you for a password as shown below. The password is also **student**.

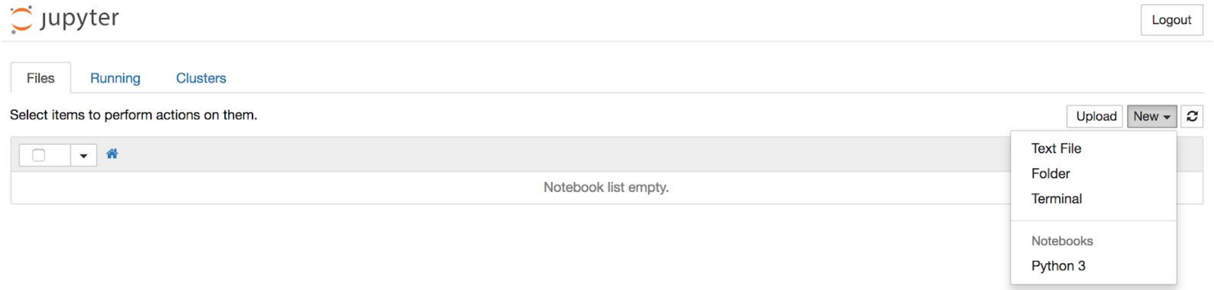


192.168.56.101:8888/login?next=%2Ftree%3F

Step 6: Using Jupyter Notebook

You should now be connected to the jupyter notebook using a web browser. Jupyter Notebook is a very powerful tool with many features. This section will outline a few of the basic features to get you started using the Notebook. Jupyter Notebook will show all of the files and folders in the directory it is run from, so when you're working on a project make sure to start it from the project directory.

To create a new Notebook file, select **New > Python 3** from the top right pull-down menu:



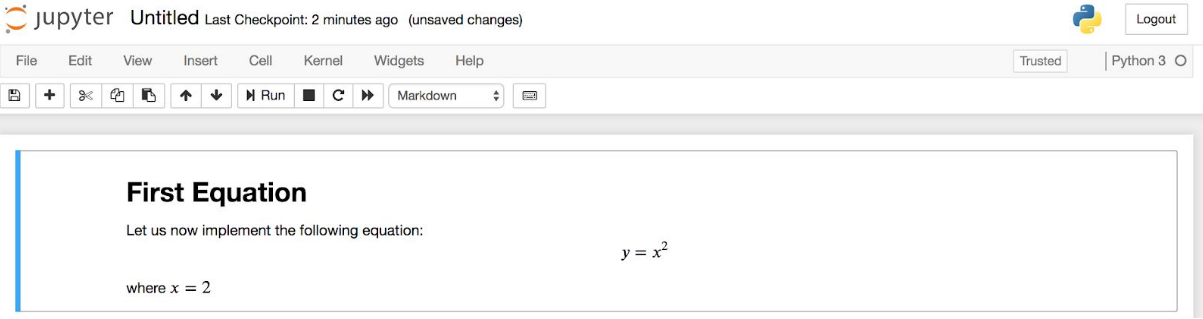
Select items to perform actions on them.

Notebook list empty.

This will open a Notebook. We can now run the Python code in the cell or change the cell to markdown. For example, change the first cell to accept Markdown by clicking **Cell > Cell Type > Markdown** from the top navigation bar. We can now write notes using Markdown and even include equations written in LaTeX by putting them between the $$ symbols. For example, type the following into the cell after changing it to markdown:

```
# First Equation

Let us now implement the following equation:
$$ y = x^2$$

where $x = 2$
```

To turn the markdown into rich text, press CTRL+ENTER, and the following should be the results:



First Equation

Let us now implement the following equation:

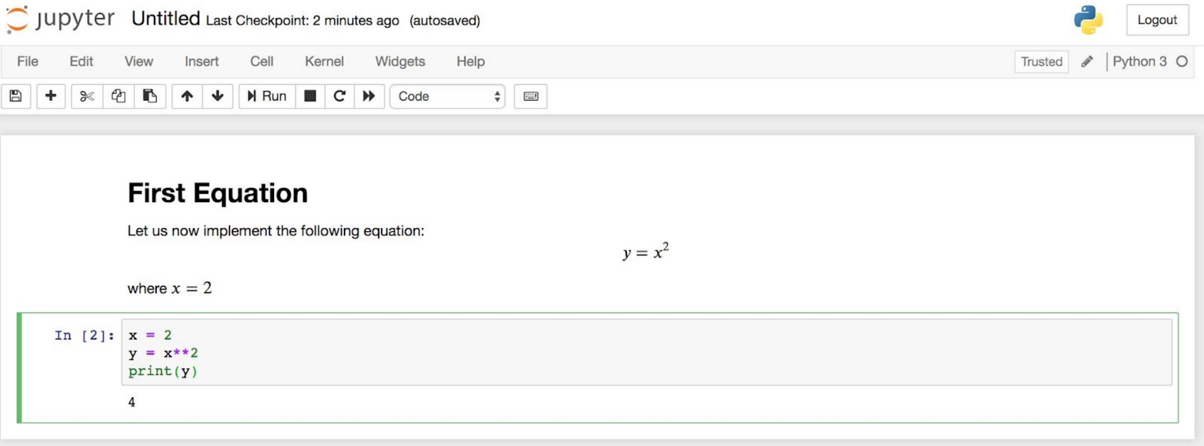
$$y = x^2$$

where $x = 2$

You can use the markdown cells to make notes and document your code. Let's implement that equation and print the result. Click on the top cell, then press ALT+ENTER to add a cell below it. Enter the following code in the new cell.

```
x = 2
y = x**2
print(y)
```

To run the code, press CTRL+ENTER. You'll receive the following results:



First Equation

Let us now implement the following equation:

$$y = x^2$$

where $x = 2$

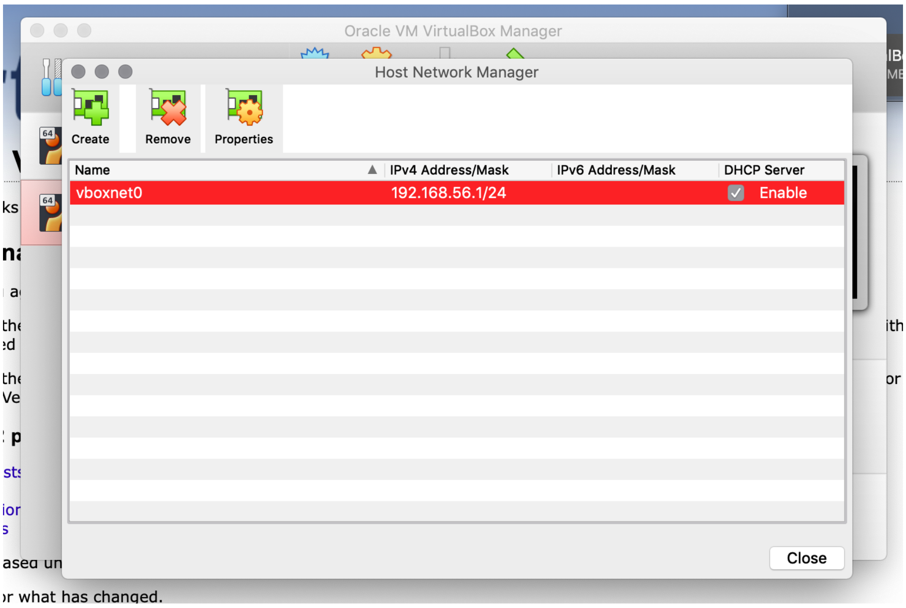
```
In [2]: x = 2
        y = x**2
        print(y)

        4
```

You now have the ability to import modules and use the Notebook as you would with any other Python development environment!

That's it for this activity. Hope you have enjoyed setting up your own machine for processing big data.

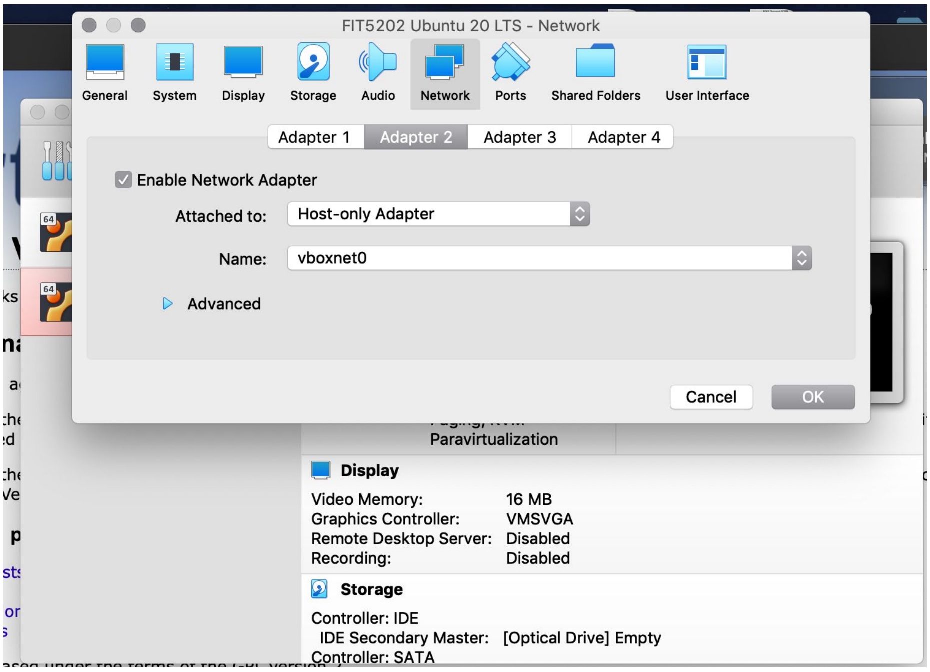
Setting up Network Interface in MAC OS

Navigate to **File -> Host Network Manager**
Click on the “Create” button, this will add a new network called **vboxnet0**, make sure the DHCP Server is enabled.



Name	IPv4 Address/Mask	IPv6 Address/Mask	DHCP Server
vboxnet0	192.168.56.1/24		☑ Enable

Proceed to Settings -> Network
Go to Adapter 2 to add the new interface, select “Enable Network Adapter”, choose “Host-only Adapter”. The newly created network **vboxnet0**, should be automatically selected. Click OK.



☑ Enable Network Adapter

Attached to: Host-only Adapter

Name: vboxnet0

Paravirtualization

Display

Video Memory: 16 MB
Graphics Controller: VMSVGA
Remote Desktop Server: Disabled
Recording: Disabled

Storage

Controller: IDE
IDE Secondary Master: [Optical Drive] Empty
Controller: SATA