

# DATA7201 Practical session

In this session, we are going to learn how to connect to a Hadoop cluster and work out some basic HDFS commands. To work with Hadoop, we need to:

1. Get your zone id.
2. Connect to your zone.

## 1. Get your zone id:

To connect to a cluster node, we need to login into systems/computers that have access to it. These systems/computers are called *zones* (which is a virtual machine hosted in the private UQ cloud).

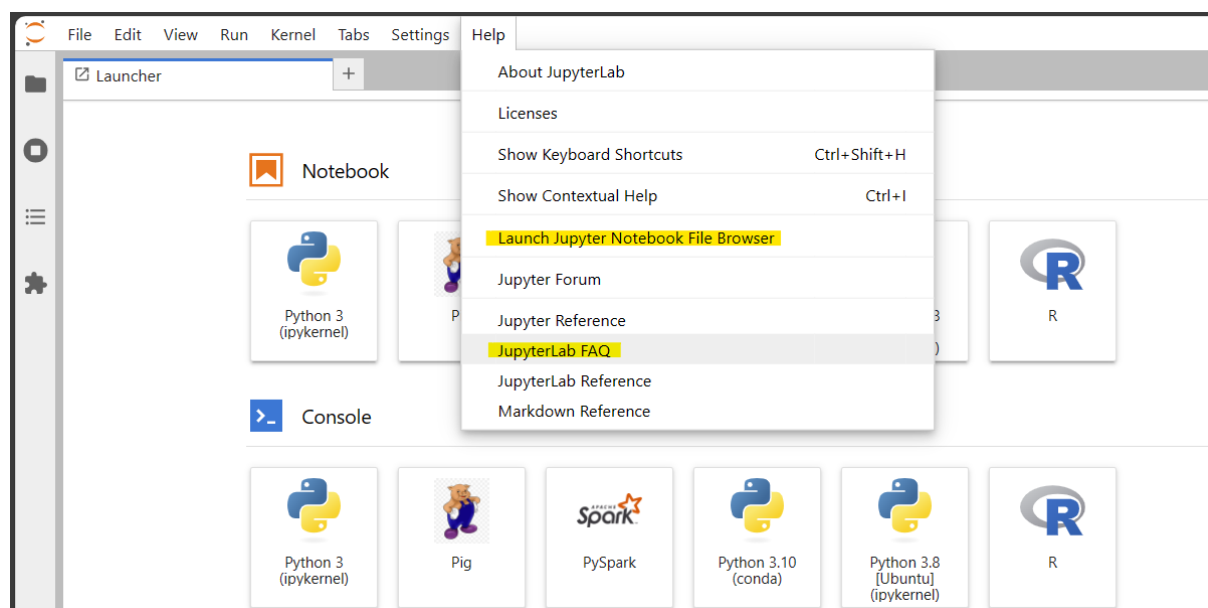
Each student in DATA7201 has his/her own zone to connect to. You will need to find your zone's hostname from the Zone Manager Page (<https://coursemgr.uqcloud.net/data7201>) The examples below demonstrate a zone id "7da3f7d2" and a student id of uqbrogas (this will be your s1234567 student id)

## 2. Connect to your zone:

Open Jupyter Lab by visiting <https://data7201-yourzoneid.uqcloud.net/jupyter/lab> with a web browser. Eg: <https://data7201-7da3f7d2.uqcloud.net/jupyter/lab>

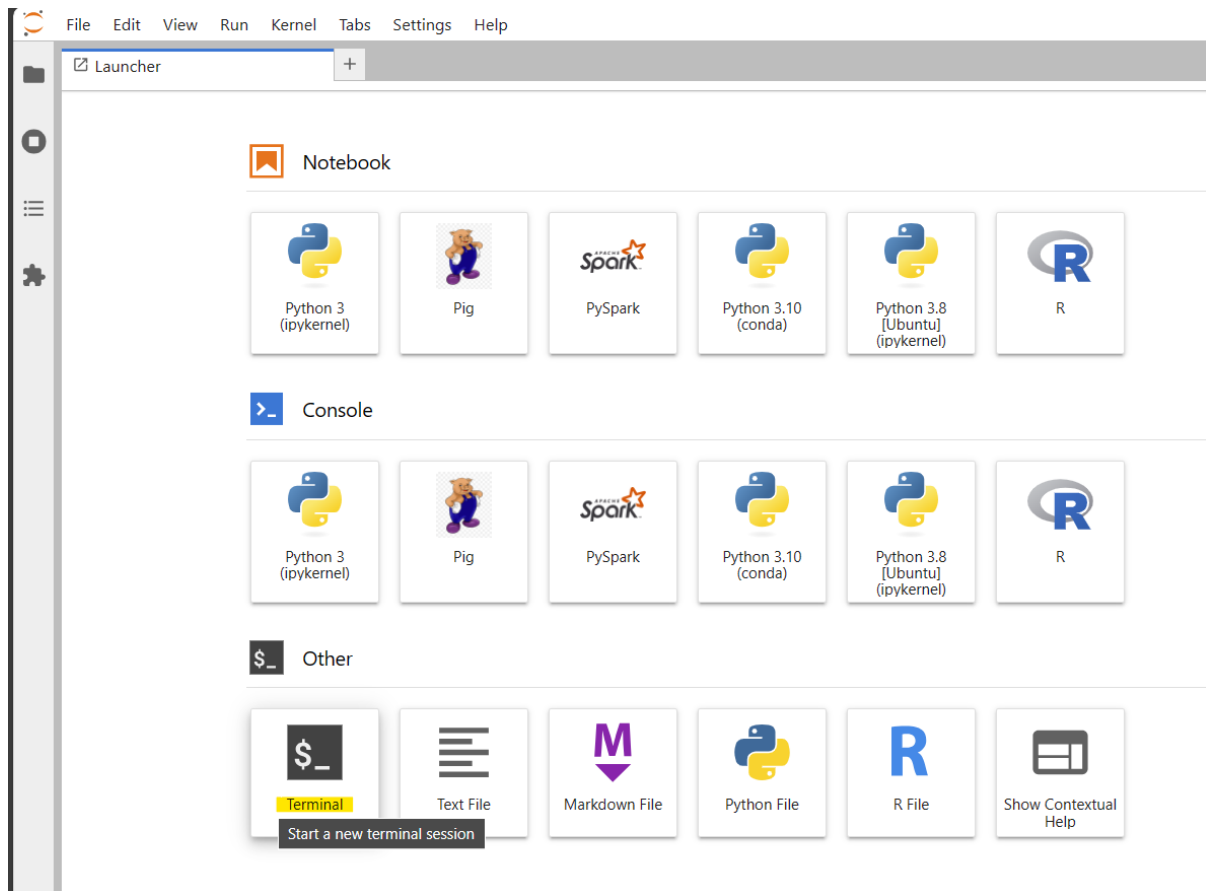
JupyterLab brings the classic notebooks, text editor, terminal, console, and file directory viewer all under one place. It provides a more flexible layout with a drag and drop interface.

If you are new to JupyterLab interface, you can refer to the user guide available under the Help tab -> 'JupyterLab FAQ' section. Or if you like to stick with the older classic Jupyter notebook interface, click on Help tab -> 'Launch Jupyter Notebook File Browser'. For the practical session, we will stick with the JupyterLab interface.



We will work from Terminal to interact with the cluster node.

Choose 'Terminal' from Launcher tab as shown below.



## Interacting with the cluster via command line

Here are some basic commands you can use once logged in to the cluster node.

### Basic Linux terminal commands

**pwd** - tells you about your current location in the local filesystem, e.g., /home/s1

**mkdir** - allows you to create a new directory, e.g., mkdir myfolder

**cd** - allows you to change your current location to another directory e.g., cd myfolder

**ls** - lists the content of the current folder.

**cp** - copies files from a place to another, e.g., cp /home/s1/filename /home/s1/myfolder/filename

**mv** - move files from one place to another, e.g., mv /home/s1/filename /home/s1/myfolder/filename.

**unzip** - decompresses zip files, e.g., unzip file.zip.

**df -h** - shows you the local disk usage.

**wget** - downloads data from a web URL, e.g., wget <http://bbc.com> /home/sxxxxxx

**chmod** - change access control to a file or a folder.

**top** - Shows monitor panel of the server's resources.

**wc -l <filename>** - count lines in a txt file.

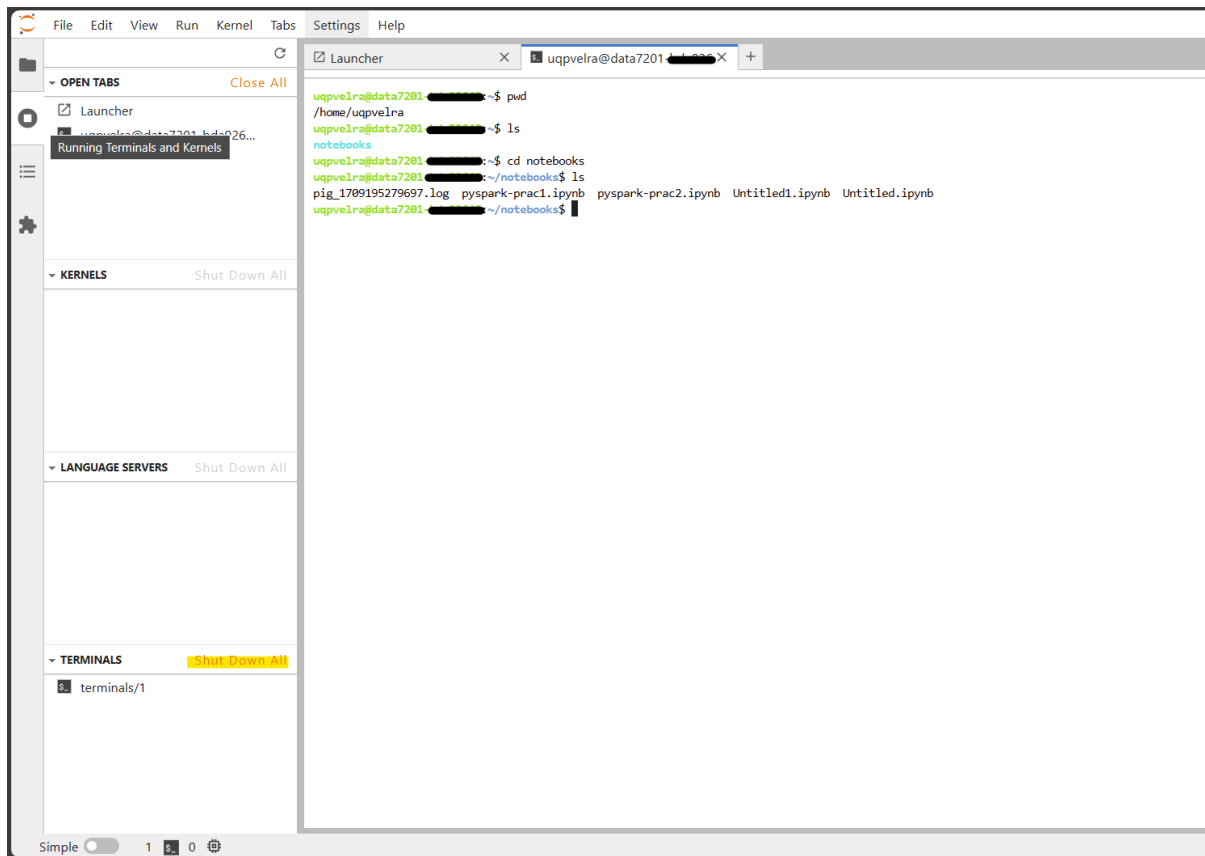
and much more. See, e.g.:

<https://www.hostinger.com/tutorials/ssh/basic-ssh-commands>

<https://www.howtogeek.com/437958/how-to-use-the-chmod-command-on-linux/>

**More help:** Each command has an attached manual explaining its usage and functionalities. It can be accessed by the command **man**, e.g., **man ls**.

**Example screenshot** is shown below. You can work on multiple terminals and Jupyter notebooks concurrently. If terminals or notebooks are not in use, shutdown it by accessing the collapsible Left side bar -> 'Running Terminals and Kernels'.



## 1. Exercise

Download `moby10b.txt` (URL available on blackboard Week3/Practical session) in a new folder called "prac-1" under your home directory using `wget`.

- How many lines does `moby10b.txt` have?
- What permissions does `moby10b.txt` have?

## 2. Exercise

### HDFS Access

Now that you have access to a cluster node, you can put some data onto HDFS (from `data7201-xxxxxx`). Type and execute "`hdfs dfs`" command.

- Which of the basic terminal commands are in `hdfs dfs`?

Also, a full list of available commands can be found here:

<https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-common/FileSystemShell.html>

### 3. Exercise

Create a new folder “prac-1” in HDFS and insert moby10b.txt from your own node. As a hint, you can check the following commands:

**hdfs dfs -ls /** shows the content of the root folder. This is equivalent to “ls /” on the local file system.

**hdfs dfs -mkdir** creates a directory in HDFS, e.g., `hdfs dfs -mkdir /user/sxxxxxx/myfolder`

**hdfs dfs -put** copies files from the local file system to HDFS, e.g., “`hdfs dfs -put localfile /user/sxxxxxx/myfolder`”

**hdfs dfs -rm** deletes files from HDFS.

Another hint, you have assigned a working folder in HDFS at “/user/sxxxxx”

- What permissions does moby10b.txt have?