

Acquiring and Manipulating Data in Hadoop

Week 1

If you want to really do some serious playing with Hadoop you probably need this set up on your own PC or Laptop. Instructions for this are given in a separate document in Blackboard.

On a Lab PC

Go to the VMStore and download a copy of the VMImage from the Common area. The Image is called HDP_2.4 SHU. You will need to put it in the D:drive as this is the only place you have write access to. This means you will lose the work you do at the end of the session unless you copy it to a pen-drive or to your own VMStore. Take the following steps:

- From the VMStore select Q:\ACESStudent\VMStore\COMMON\BDDS\HDP_2.4 SHU and copy this to the D:\ drive)
- Open VMware Workstation from the desktop
- Once VMWare is loaded go to File and Select "Open..."
- Go to the D: drive where you copied the (VM) Virtual Machine image and select the file called HDP_2.4 SHU.vmx to load this VM into the VMware Workstation environment.
- Once the VM image is loaded into the VMware Workstation, then select the option "Power on this Virtual Machine" to start up the VM. (If you are using VMPlayer on your own PC this may say Play Virtual Machine instead)
- When you are asked, select "**I copied it**"
- If asked about software updates, decline to update software by clicking **Remind Me Later**
- The VM will then go through the process of starting the Linux machine and then starting Hadoop. This may take a few minutes. Wait until the screen looks something like the figure 1 below.

Later we will add some other tools, but just to get us started, this will do for now. You are probably itching to do something with Hadoop! For the time being we will not be logging in directly, so ignore the footer of Figure 1 and we will use a browser.

What happens when you start your Sandbox?

First of all the VM boots up. It is running CentOS, which is a free Linux variant derived from Red Hat Enterprise Linux. Once the operating system is running Hadoop and many supporting processes are started. At the end of the sequence we have, in effect, a Hadoop server that we will be connecting to and using. It has the Hadoop Distributed File System running as a single-node Hadoop cluster. You will need to look at other learning resources to gain an understanding of how HDFS works in principle; and to begin to work on production scale Hadoop which is likely to be running in multi-node mode. Here we are going to get on with using the tools within the Sandbox on the grounds that use often assists with learning.

We can connect to the Sandbox in a number of ways. Later we will be using PuTTY to connect to it and issue command line instructions, for example. But the easiest way to explore the environment is with the browser-based tools that come for free when you use the Hortonworks VM.

Exploring the Browser interface to Hadoop

“Hello World” is often the first task tutorials on programming languages cover and is so famous it has its own Wikipedia page: http://en.wikipedia.org/wiki/%22Hello,_world!%22_program ! With Big Data the first task tends to be a word count. It is a relatively simple concept to get used to the idea of querying data – how many instances of each unique word are there in a particular collection of words? In the case of Hadoop it also allows you to get an early exposure to another key building block: MapReduce. We will come across the MapReduce algorithm often as we use our tools simply because it is such a key building block in the Hadoop environment. However we are not going to explain the MapReduce algorithm here. We cover it in more detail later, but just do a Google search on MapReduce if you want a quick overview before starting, or read the Wikipedia pages: <http://en.wikipedia.org/wiki/MapReduce>.

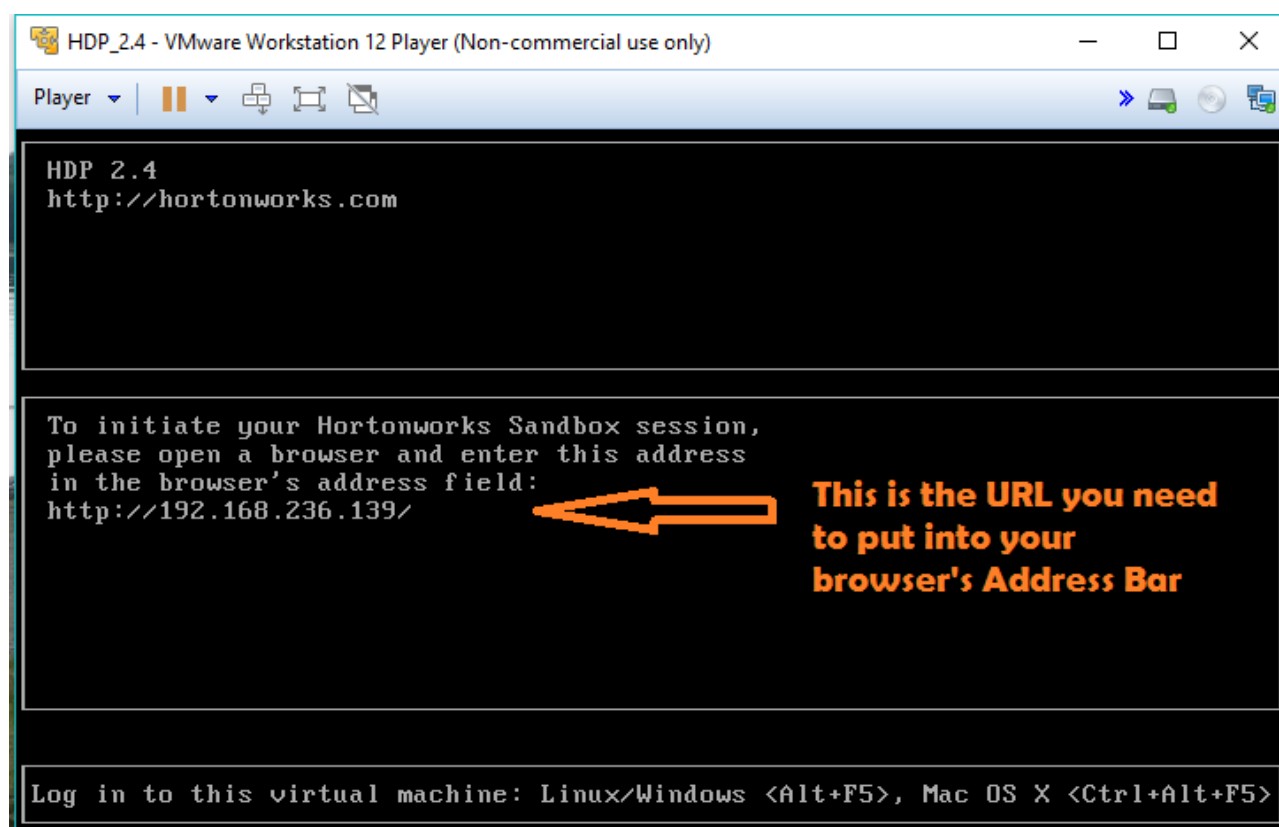


Figure 1

WARNING: we are in new and fast moving territory here. Don't expect things to work as you would hope they would. I know, from personal experience, Microsoft Edge Browser (Window 10+) does not work. For the examples I use Avant (<http://avantbrowser.com/>) but only because that has been on my home server for years! Firefox works... you will just have to “suck it and see!” if you are using different browsers. IE in the labs seemed to work in July.

When the VM is fully started and has provided you with a URL (as per figure 1 above), type that address into your browser's address bar. Your screen should look something like figure 2. BUT, this URL will probably be different on your PC to that one in the figure. Use your URL!

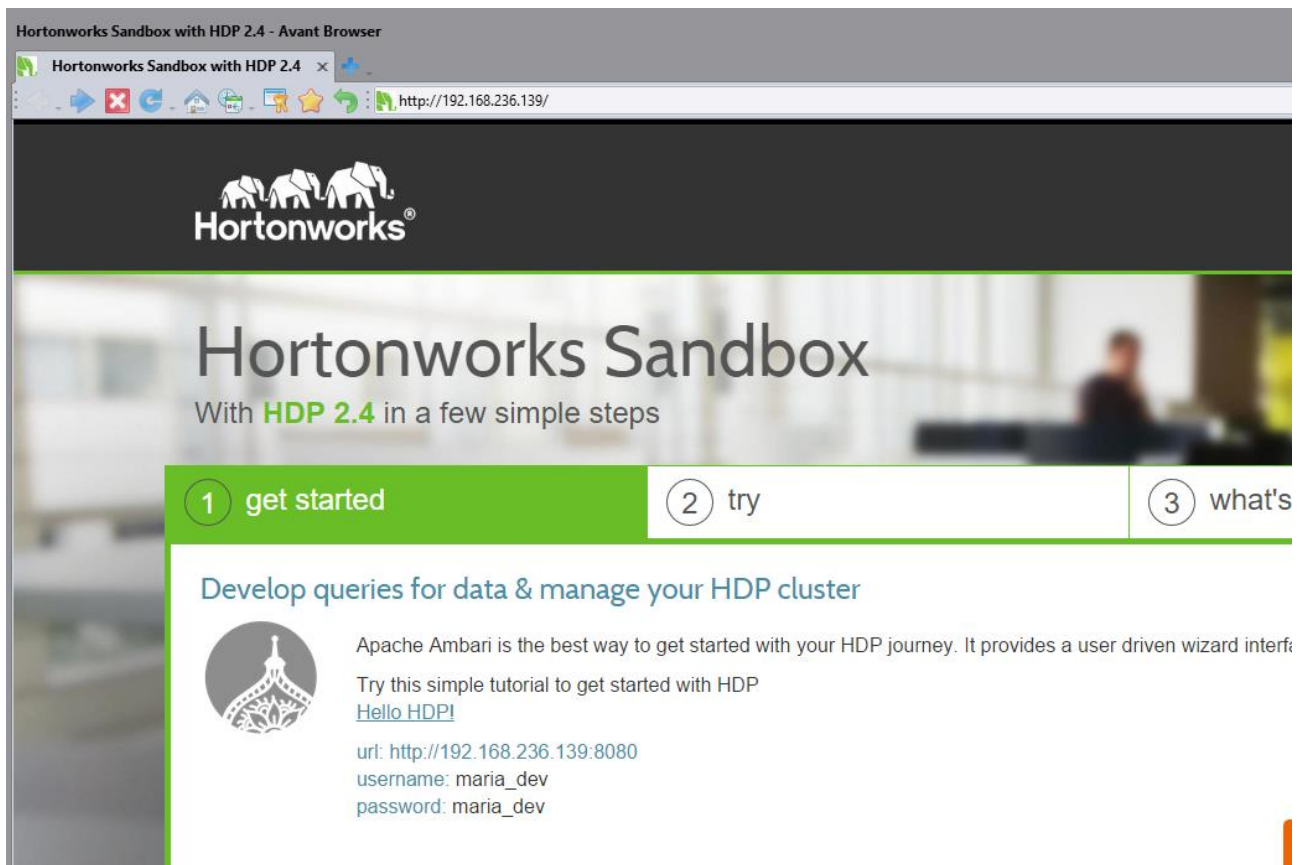


Figure 2

You will see there are many tutorials provided on the website you land on. Later, do explore them to add to your understanding of Hadoop and remember that we are only here focusing on a subset of the functionality built into the Hadoop environment.

From now on, to get straight to the browser-based menu we will be using you should add the port number **:8080** to the address you used. In my case this is how I open the browser interface to the Sandbox (see figure 3): <http://192.168.236.139:8080/>

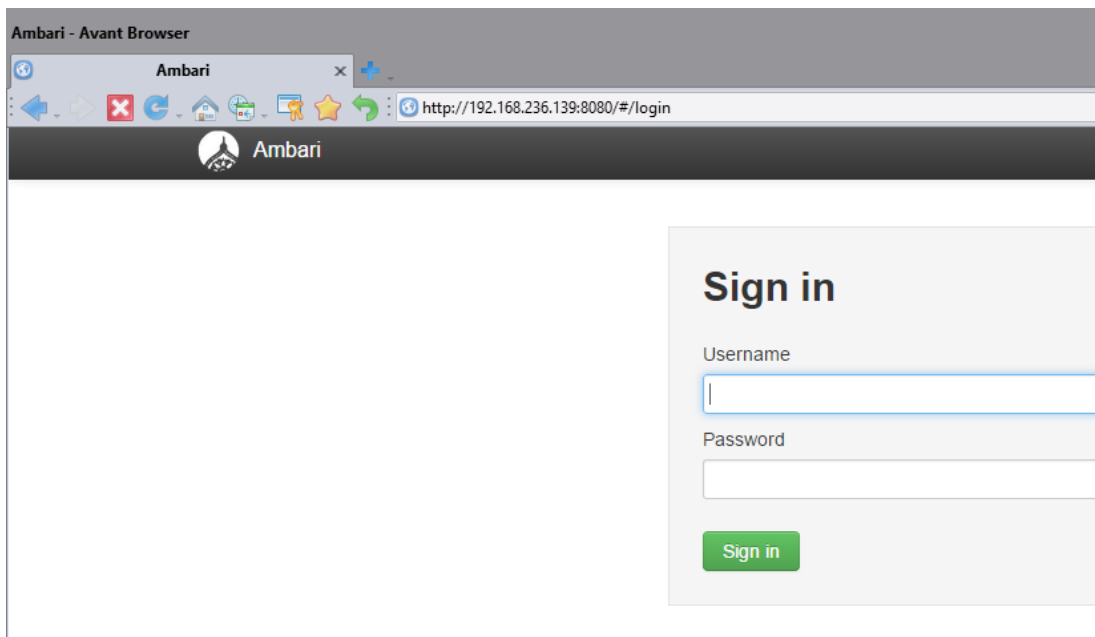


Figure 3

To get us started we will use the account provided by Hortonworks, called `maria_dev` and the password is also `maria_dev`.

The opening page is called the **Dashboard**. We will see how some of these bits work through the weeks. For the time being, do not worry about what is being reported on the dashboard. We are using this as our menu. Your browser should now look something like figure 4.

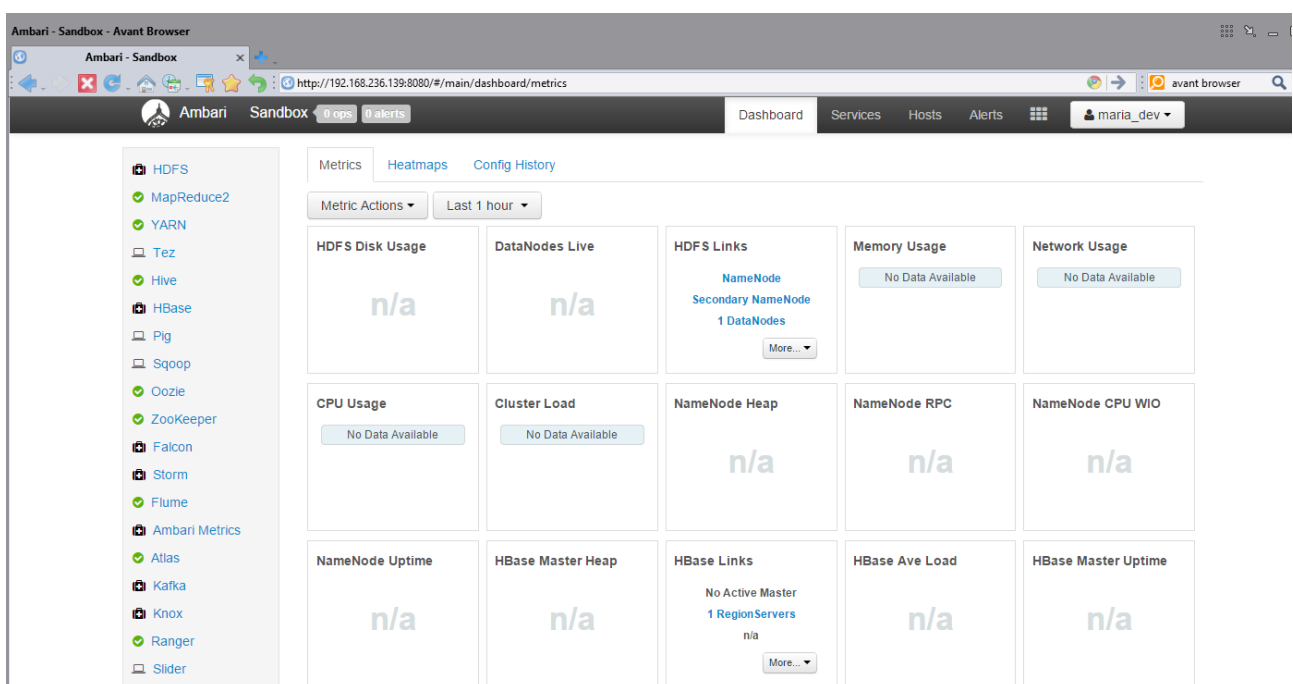


Figure 4

Working with HDFS

Hadoop Distributed File System (HDFS) is the open source distributed file system that underpins the Hadoop infrastructure. As a file system its job is to control how data is stored and retrieved. (If you haven't come across this term before, try this: https://en.wikipedia.org/wiki/File_system.) HDFS is at the core of all we will cover in our Hadoop tutorials. You can think of it as a bit like DOS, or the Windows file system. You will certainly see familiar terms like “directory” in this file system. It is, however, important to remember that HDFS does do things differently. We explore some of those differences through this module.

For the time being, take a closer look at the icons on the top bar (figure 5). You will see a 3x3 grid, which is the menu selection item. If you hover your mouse over this you will see you get some options.

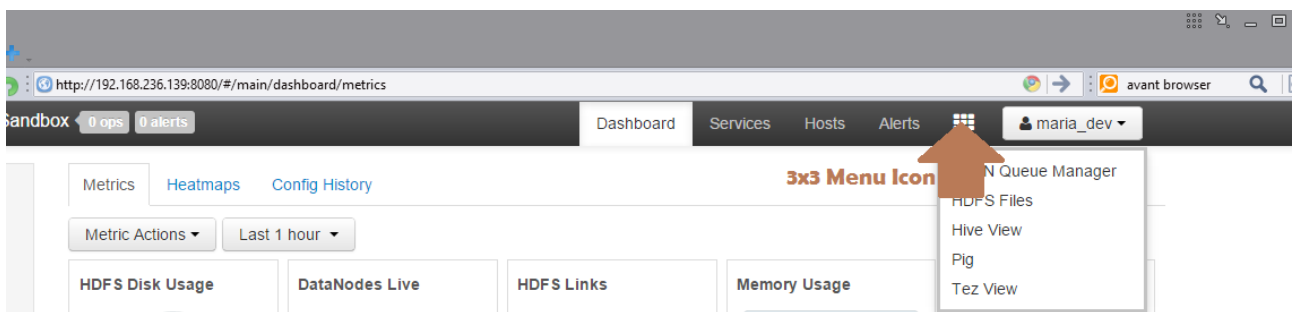


Figure 5

From that menu, select HDFS Files. Your screen should now look something like Figure 6.

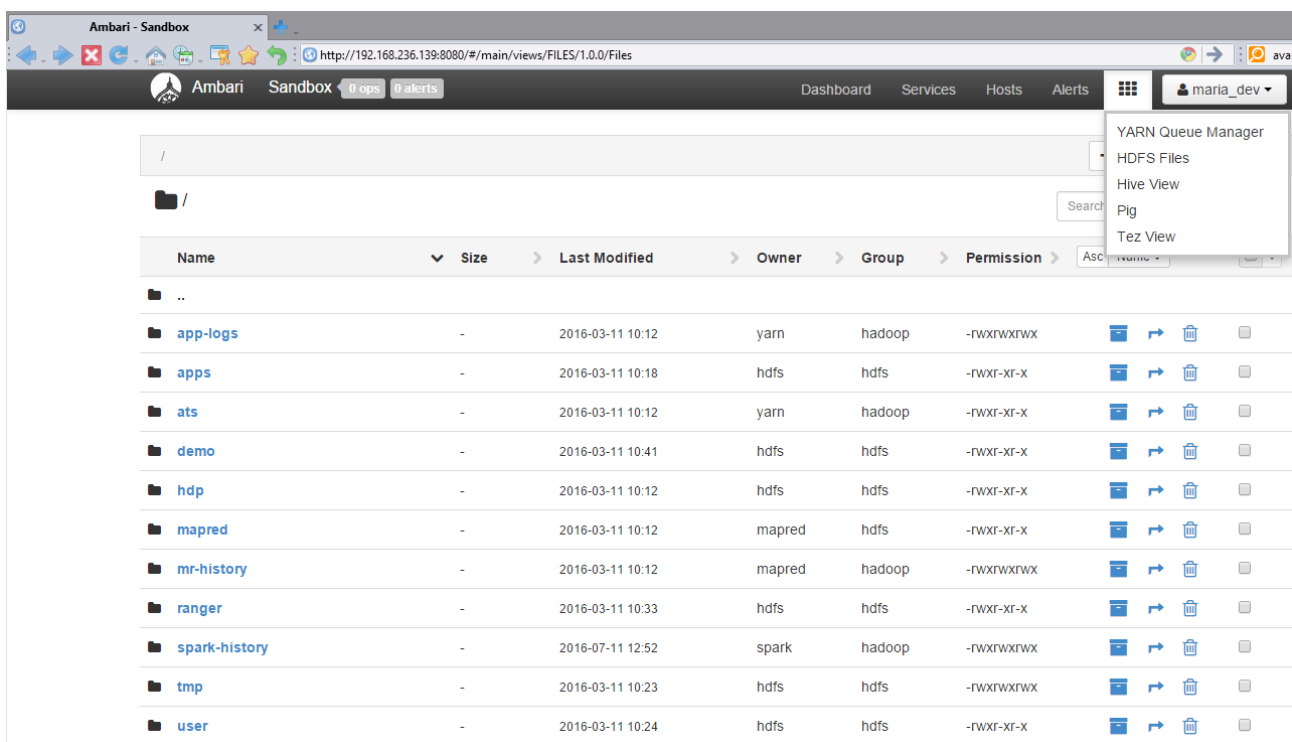


Figure 6

What is happening here?

Before we get started it is really important for us to understand what we are looking at here. This is a typical file browser, not too dissimilar to what you get with Windows File Explorer, except it is in a Web Browser interface. **But what is listed are the directories and files WITHIN Hadoop's HDFS**, not those of either the Host Windows PC or the Guest CentOS VM.

Across the top we get some buttons to do file management-type tasks, like create new directories. To the left of that we see path-based breadcrumbs indicating where we are within the structure of HDFS. You can use these to short-cut your navigation.

Going down in rows are lists of files and directories. Next to each is listed their properties and links to allow management tasks such as move or rename.

To start with, we are in a directory (folder if you are a Windows user) called root, shown as a “ / “.

Take some time to look around and then click on the link to the **user** directory. We have logged in as maria_dev, so now click into that directory. Your screen should now look like figure 7.

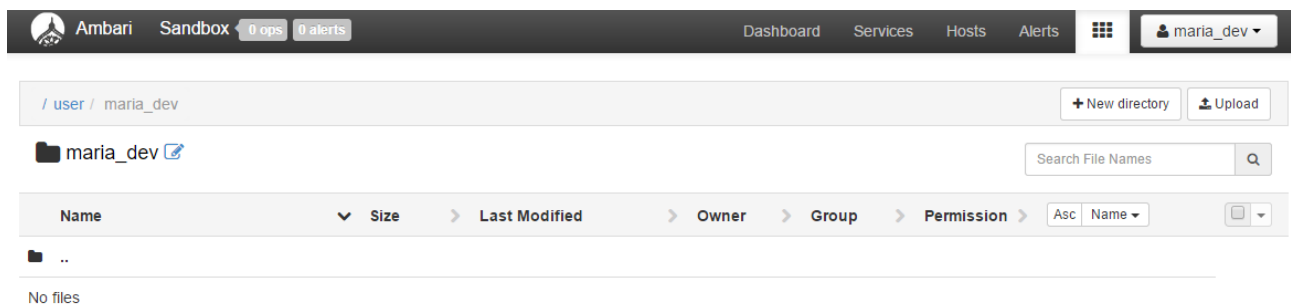


Figure 7

We are now going to create a directory into which we will place some text in readiness for carrying out a word count. Click on the **New Directory** button on the right of the screen and then create a new Directory called something like **tutorials**. Now move into that directory and see that it is empty.

Keeping your Ambari Sandbox session active, in a different Browser session, we need to find a nicely sized file with lots of words in it. Project Gutenberg (<http://www.gutenberg.org/>) is a good place to look as it has over 46000 books stored for free use. James Joyce's Ulysses will do nicely as an example. You can find it here: <http://www.gutenberg.org/files/4300/4300.txt>. Download to your PC, but remember where you put it! You might like to rename it to something more meaningful, like Joyce.txt.

Now return to the HDFS File Manager and click on the upload button (see Figure 8). Select the file you saved to your PC.

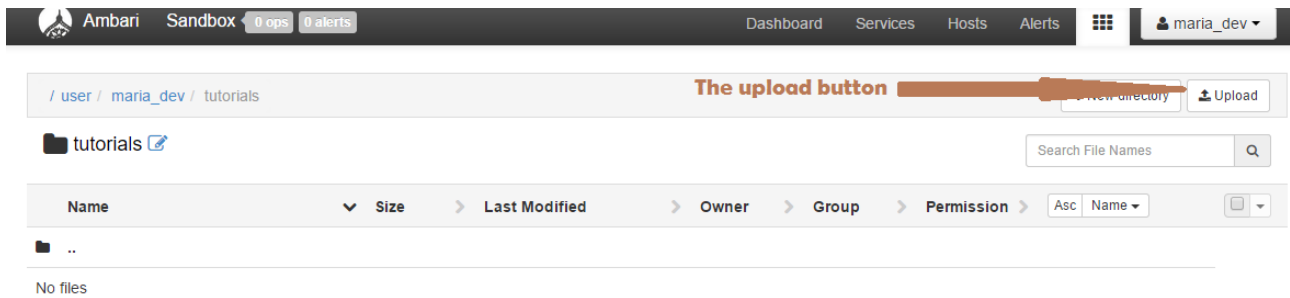


Figure 8

You should then see the document in your HDFS File Manager (see figure 9). Try clicking the link to see the content.

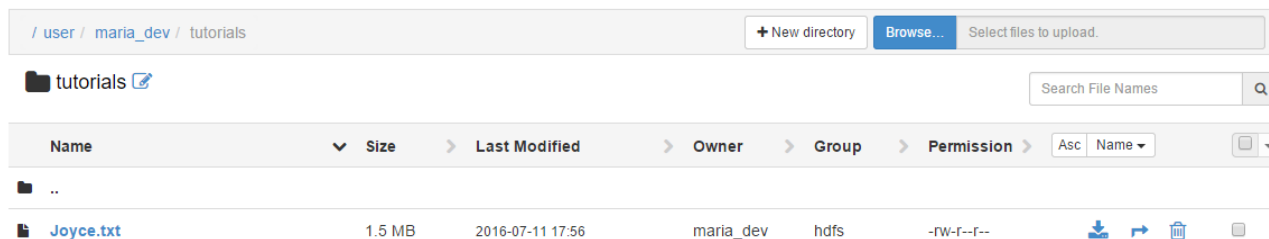


Figure 9

Just to be clear, what we have done is fetched a file from an external source. Saved that onto the local file system of the host PC, and then uploaded that file into HDFS, saving it in a directory we have created especially for our use.

There are many more things we can do with the File Browser and we will be coming back to it in the future.

Downloading and using PuTTY to interact with the Sandbox

To run the wordcount programme we will need to provide the Sandbox with some command line input. One of the ways to log on to a multi-user system from another computer, is with a secure shell (SSH) session and PuTTY allows us to do this. Remember that your Sandbox, although it is actually on the same PC as your other software, is running as if it was a separate server. This means we will use PuTTY running in your Windows host system to talk to the Hadoop system running on the Centos VM.

If you are in a lab, simply type Putty in the Start button since it is installed on our network.

If you are running this on your own PC you will first need to use a browser and connect to the PuTTY site and download the executable: <http://the.earth.li/~sgtatham/putty/latest/x86/putty.exe> . It does not need installing. Just put it somewhere where you will remember, for example in a folder called putty in your windows folder. You may like to put a link on your desktop as you may be using this quite often. Make sure your Sandbox is running and that you know the connection URL.

Start PuTTY and enter the address that the Sandbox indicates and then click open, as seen in figure 10. Save your session by giving it a name and clicking **Save** to save having to put your address in every time you use PuTTY.

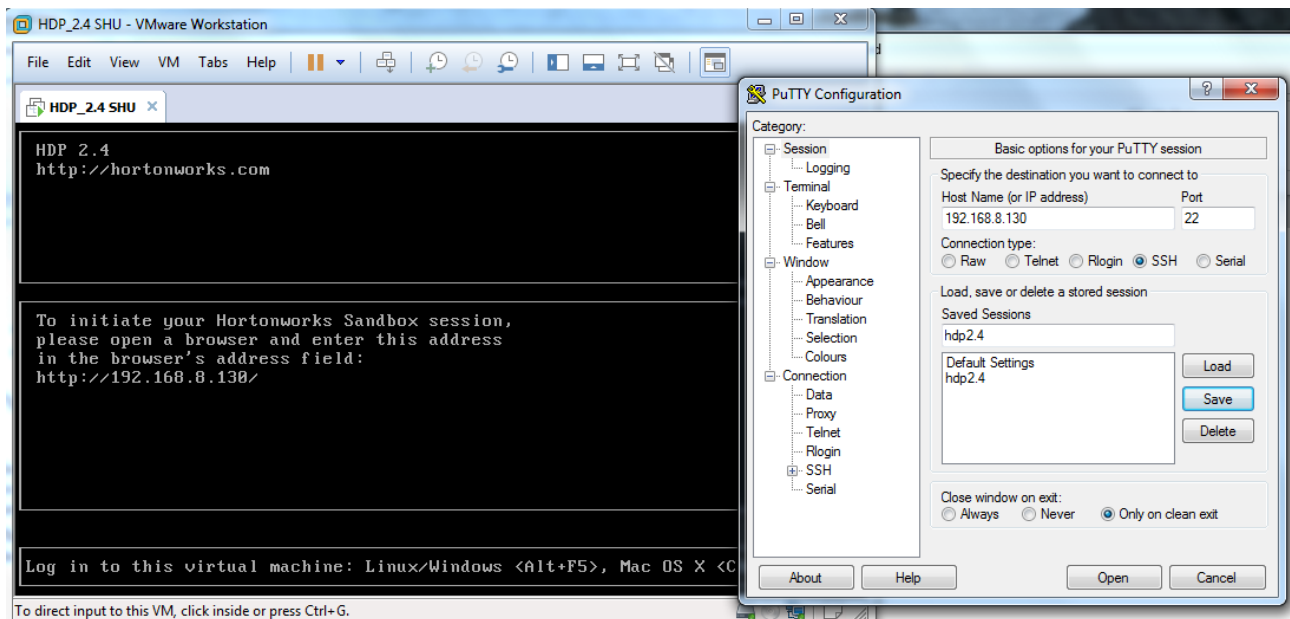


Figure 10

Click on **Open**.

You will then be asked to log in. We will connect as user **root** with a password of **hadoop**. The first time you log in you will be asked to change the password (see figure 11). Make it something you will remember! But it is quite picky in terms of not allowing simple passwords. For example, I used Rovers and it didn't like it, and the tried DoncasterRovers and it liked it.

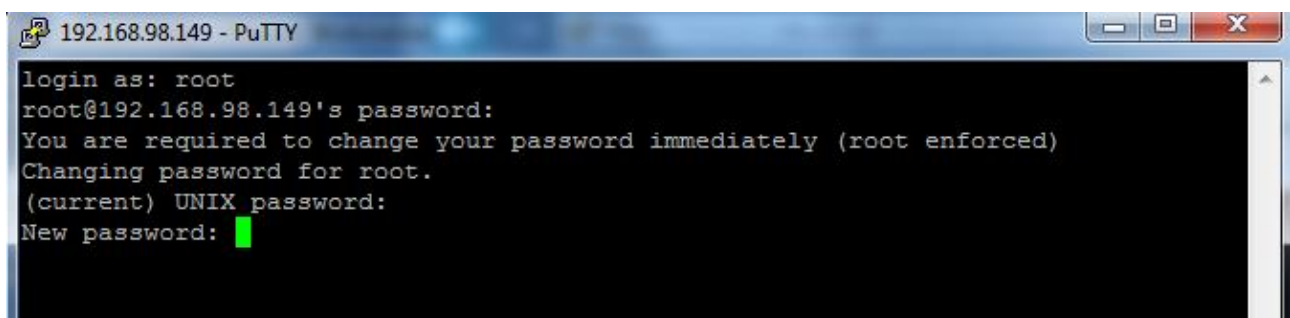


Figure 11

We should then have a command line with a prompt, awaiting our input, which should look something like this: **[root@sandbox ~]#**

If you are new to Linux and command line environments, for the time being, just copy the steps and try not to worry too much about the operating system commands. We suggest you get some Linux experience separately - there are tutorial links in Blackboard.

Warning: Root is a privileged account that has access to all the commands and files on a Linux system, so be careful!

Don't do this now, but to close a PuTTY session just type exit at the command prompt.

Finally, for now, we are going to run the wordcount programme. Or to be more precise we are going to run Hadoop and pass it some parameters that tell it to do the wordcount. We will have to type the command very precisely. The first word is hadoop, to call the hadoop executable. The parameters that follow in this case are:

- jar – to tell hadoop to run a jar file
- the location and filename of the jar file that hadoop should run
- the class to call from within the jar file (wordcount in this case)
- the location and filename of the document to be counted (which must be stored in HDFS)
- the location that output should be stored in – again in HDFS. (This should not already exist)

For interest, in the version of the Sandbox we are using, the hadoop executable is stored in:
/usr/hdp/2.4.0.0-169/hadoop/bin

Move to this directory before running the code example by issuing this command:

cd /usr/hdp/2.4.0.0-169/hadoop/bin

The wordcount function can be found in this jar:

/usr/hdp/2.4.0.0-169/hadoop-mapreduce/hadoop-mapreduce-examples.jar

So the full command we need to type into the PuTTY session is:

**hadoop jar /usr/hdp/2.4.0.0-169/hadoop-mapreduce/hadoop-mapreduce-examples.jar
wordcount /user/maria_dev/tutorials/Joyce.txt /user/maria_dev/tutorials/JoyceOutput**

Once you press enter you will begin to see messages appear in the PuTTY terminal session. Just wait until you get the input prompt back again before doing anything else – this may take a few minutes. Then refresh your HDFS File Manager browser session to view the output created. It should look like figure 12.




/ user / maria_dev / tutorials / JoyceOutput			
JoyceOutput 			
Name	Size	Last Modified	
..			
 _SUCCESS	0.1 kB	2016-07-13 14:43	
 part-r-00000	508.0 kB	2016-07-13 14:43	

Figure 12

Look at the contents of that directory and then look at the wordcount output in the file unhelpfully called **part-r-00000**. We now know how many of each individual word there were in this book!

We have now used Hadoop to store some data and analysed that data to discover what its constituent words are. This is a good example to get us used to our new environment but there is much more to data analysis than just counting words!

What Next?

You could try to see if you have understood the concepts described above by selecting your own book, or other text source, to carry out a wordcount on. We will cover much more in subsequent weeks but for this week you should just improve your familiarity with the environment we are working in, including the Linux operating system.

Make sure you close the PuTTY. Close the VM by clicking the **Player** drop-down and then click **Power** followed by **Shut Down Guest**. Say **Yes**, you are sure you want to close down.