Big Data with Hadoop

Lab Instructions Part 2: Cluster Management and Cluster Administration

Introduction

This document helps you navigate through your first meeting with Hadoop. The document also contains the questions to address in your lab report[1].

The exploration of Hadoop is the lab; there are no particular goals to be achieved and reported. The questions to answer in your report are here to force you to stop and reflect on your work.

Feedback to These Instructions

If you find any problems, faults, omissions, etc. in these instructions, or have suggestions for improvements, please notify the teacher.

[1] See the *Lab Instructions Part 0 – General Instructions* for guidance on, and the requirements of, the report.

Table of Contents

A Few General Hints

Time quota: *You have a limited time quota; it should be more than sufficient, but make sure to not waste it. See* Lab Instructions Part 0: General Instructions*.*

Patience needed: *Do not rush. See* Lab Instructions Part 1: Connecting to the Azure Labs Virtual Machine.

Services starting order: *Starting these services in a specific order is important to ensure they function properly. Here is the recommended order.*

1. ***ZooKeeper****: Start ZooKeeper first as it is a coordination service for distributed applications.*
2. ***HDFS****: Next, start the Hadoop Distributed File System (HDFS). This includes starting the NameNode, Secondary NameNode, and DataNodes.*
3. ***YARN****: After HDFS, start the Yet Another Resource Negotiator (YARN) services, including the ResourceManager and NodeManagers.*
4. ***MapReduce2****: Finally, start the MapReduce2 services, which depend on YARN and HDFS being up and running*

Don't be afraid: *The environment is here for you to experiment with, stretch your knowledge, and make mistakes. Your* Azure Labs VM *is completely your own; nothing you do can affect anyone else.*

Possibility to reset VM: *If the lab environment seems to feel unwell, your VM can quickly be reset to the state where you first opened it;* *see page 22.*

Ambari

Log In

Navigate to *the Ambari Dashboard:*

- ***Click*** *the "Ambari" icon on the Favorites bar.*

OR

- ***Navigate*** *to localhost:8080 or [127.0.0.1:8080](127.0.0.1:8080)*

⇒ *This will take you to the* ***Ambari*** *sign in page; see Figure 1.*

Log in using:

Username: ***admin***

Password: ***Gik2q3_VT***

⇒ *This will take you to the Ambari Dashboard; see Figure 2.*

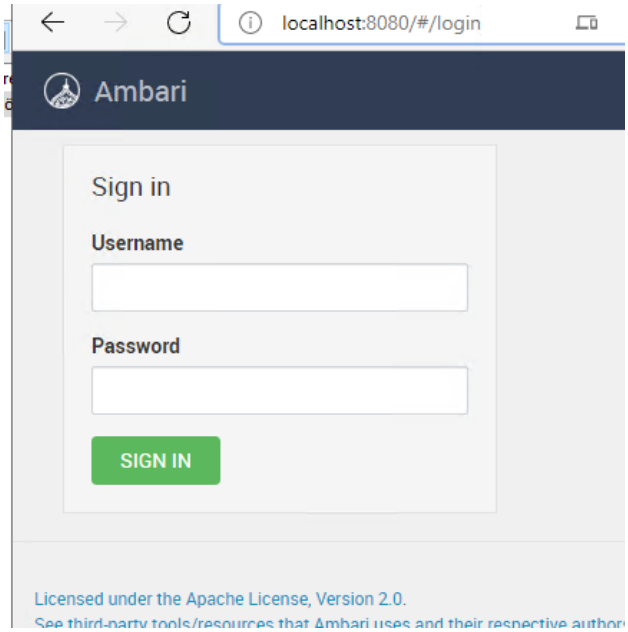

Figure 1: Ambari sign in page.

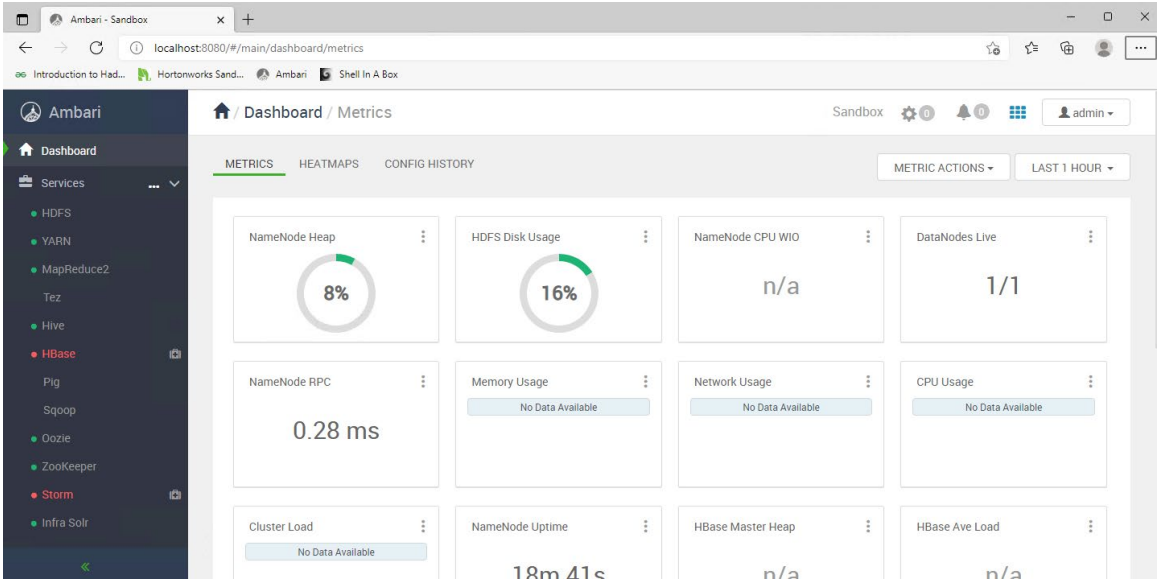

Figure 2: Ambari Dashboard.

May be you will see the left menu on top but the right manu down to the web page. HDP is not the part of cloudera now so there is GUI problem.

The Ambari Dashboard

Explore *the Ambari Dashboard on your own, with the help of Internet resources.*

Answer/discuss *the question(s) below in your lab report.*

Question Set A: The Ambari Dashboard

The Ambari Dashboard is intended to provide an overview of the Hadoop cluster. In your opinion:

1. *Describe three ways in which the Ambari Dashboard provides you with useful information.*
2. *Mention other information you would like to have access to on the Dashboard, and when/how you would have use of it.*

Hadoop Services

To the left in Ambari we see the list of "services", that together make up Hadoop. For example, HDFS (see Figure 3), Hive (Figure 4), and MapReduce (Figure 5).

***Explore** the information about each service.*

- *Take help from Internet resources to better understand what you see.*
- *If you studied and presented a specific Hadoop tool/service in the seminar/PM, have an extra look on that service.*

***Explore** what commands are available on the "Action" button; see Figure 6.*

***Answer/discuss** the question(s) below in your lab report.*

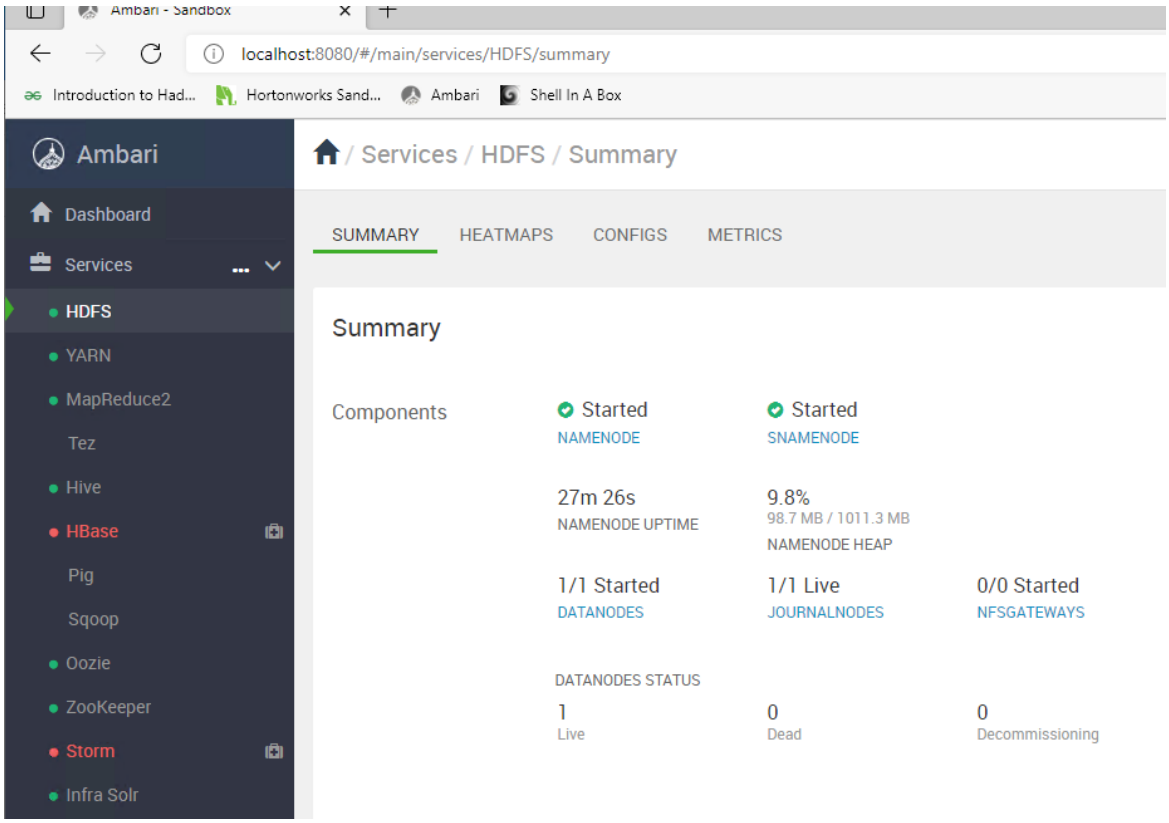

Figure 3: The HDFS service.

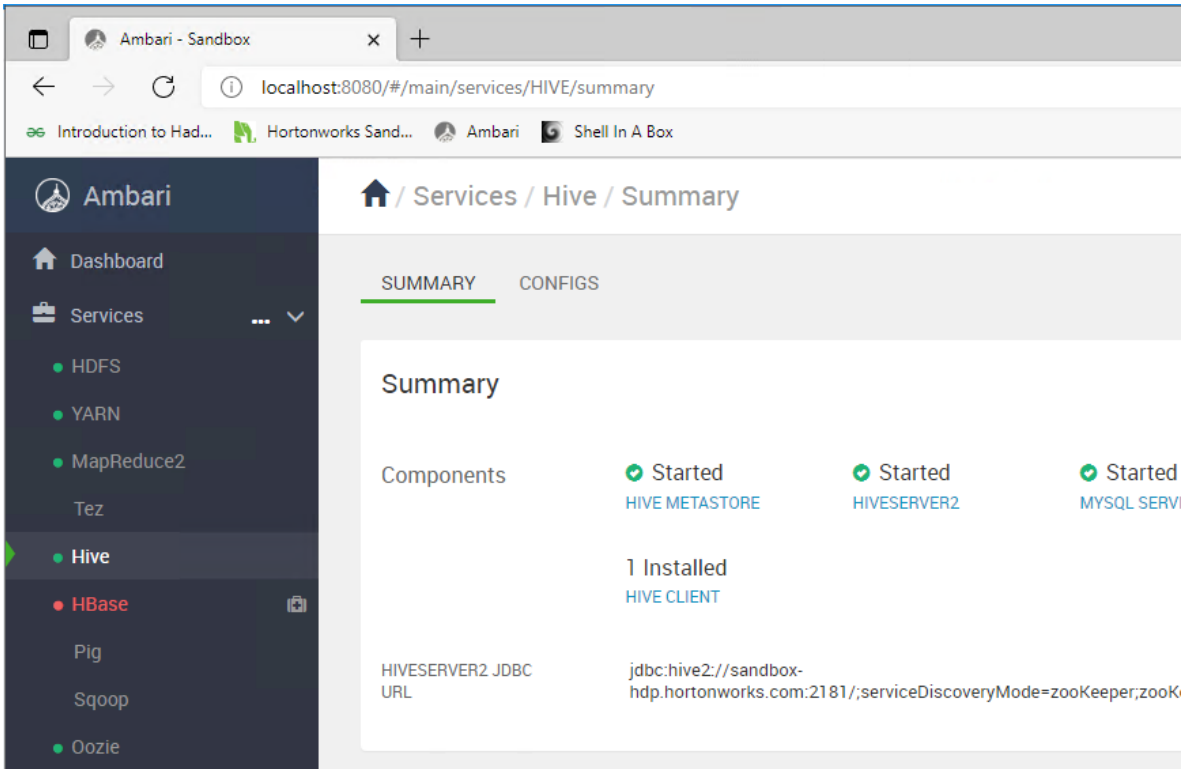

Figure 4: The Hive service.

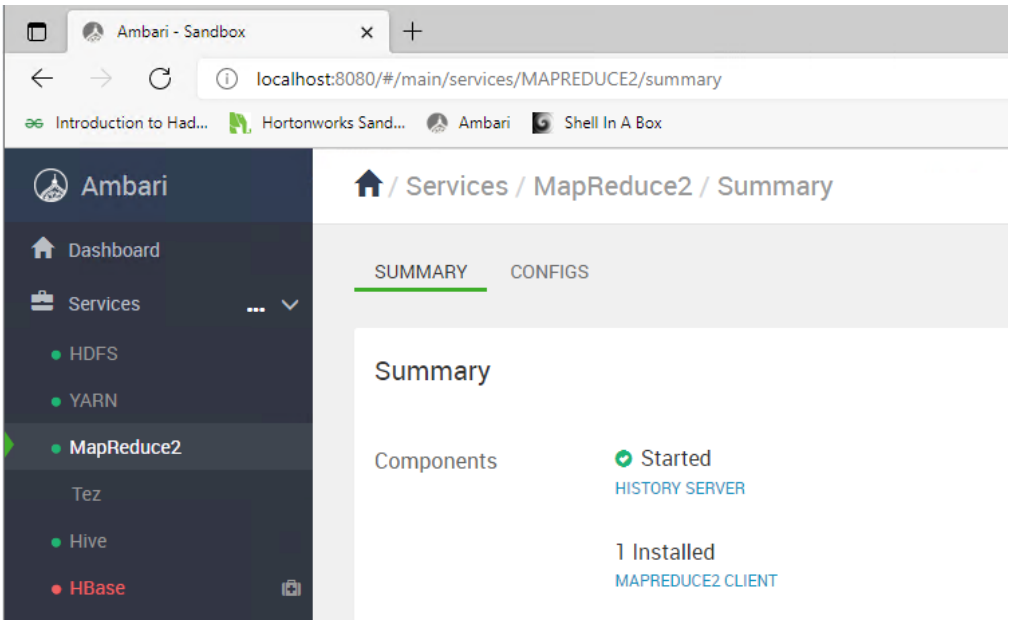

Figure 5: The MapReduce service.

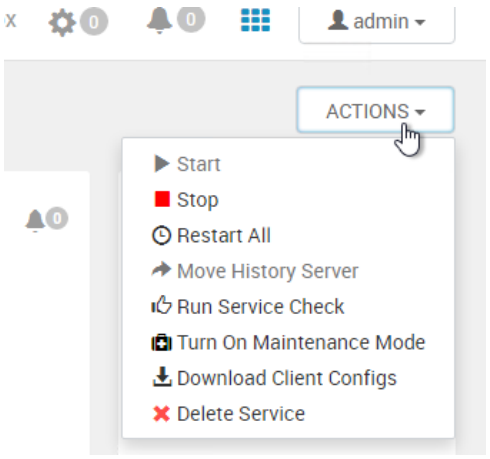

Figure 6: The "Action" menu for a service (in this case MapReduce).

Question Set B: Services

1. *What is a "service"?*
2. *What services (names only) are available in this cluster?*
3. *What are the services used for? (Choose three services; search for information in the lecture slides and on the Internet, and relate to what you see in Ambari in two sentences for each service.)*
4. *What information is available for several, most, or all, services?*
5. *Is there some information for some service that you would like to have available already on the Dashboard?*

Question Set C: Service Actions

1. *What do actions "Start" and "Stop" do?*
2. *What do actions "Turn On Maintenance Mode" and "Turn Off Maintenance Mode" do?*
3. *What does the action "Run Service Check" do?*

Hadoop Hosts

Click *"Hosts" in the menu to the left.*

Study *the information.*

Add *a new host. (You will not succeed – but try, and see how far you will get!)*

Answer/discuss *the question(s) below in your lab report.*

Question Set D: Host Management

1. *What is a "host"?*
2. *How many hosts are there in this cluster?*
3. *What information can you see about each host?*
4. *What does "Rack" mean? Why is that important information; when would you use this information?*
5. *Describe briefly what happens when you (try to) add a host. Speculate what would happen if you actually had another computer connected and added this.*

HDFS

Select *"HDFS" in the left menu.*

Explore *each of the tabs "Summary", "Heatmaps", "Configs" and "Metrics".*

Change *HDFS configuration.*

- ***Read*** *all the warnings you will get – in order to understand, to learn.*
- ***Apply*** *the changes anyway!*

Answer/discuss *the question(s) below in your lab report.*

If needed, ask your teacher to ***reset the Azure Labs Virtual Machine****.*

Question Set E: HDFS

1. *Describe HDFS briefly in your own words; in what way is HDFS a cornerstone in Hadoop? (Two to three sentences.)*
2. *Does HDFS seem to be running well on this cluster? Why do you think so?*
3. *What does NAMENODE and SNAMENODE mean? Are both started?*
4. *Are any DATANODEs running?*
5. *Why are there only one of each kind of NODE (NAMENODE, SNAMENODE, DATANODE)?*
6. *Describe briefly how you can configure each kind of NODE.*

If HDFS seems messed up now, no harm is done; you may consider resetting your Azure Labs VM*; see page 22.*

Ambari Alerts

Click *"Alerts" in the menu to the left.*

Sort *and* ***filter*** *the alerts in various ways.*

Study *the following alerts in a little more detail:*

- *Those starting with "NameNode".*
- *Those starting with "HDFS".*

Answer/discuss *the question(s) below in your lab report.*

Question Set F: Alerts

1. *What is an "alert"?*
2. *What alerts have been issued during your own lab time? Explain these (very briefly).*
3. *Why are some alerts marked with status "None"? (Why would you have to be alerted that there is no alert?)*
4. *Explain briefly how serious the "NameNode" and "HDFS" alerts would be, if the status was anything else than "None". (Select three alerts and write a sentence each.)*

Cluster Administration

Expand *"Cluster Admin" the menu to the left in Ambari.*

Click *"Stack and Versions".*

Click *the button "Manage versions" (under the tab "Versions") and* ***confirm*** *that you want to continue.*

Explore *the new Ambari interface that appears.*

Answer/discuss *the question(s) below in your lab report.*

Question Set G: Cluster Administration

1. *Characterize the difference between "Cluster Management" (i.e. all the previous sections) and "Cluster Administration".*

Ambari does its best to protect you from making mistakes; read all warnings etc. carefully, and do <u>your</u> best to apply changes, register new versions, change base URLs etc. The important thing is not to actually register new versions, etc.; the goal is to learn.

Cluster Information

Click *"Cluster Information" in the menu to the left (under "Cluster Management"). Please note that even though this is under "Cluster Management", it is an administrative task. Definitionally, the lines are sometimes blurred between Cluster Information and Cluster Management.*

Answer/discuss *the question(s) below in your lab report.*

Question Set H: Cluster Information

1. *Characterize in one or two sentences what you see.*

Hadoop Versions

Click *"Versions" in the menu to the left (under "Cluster Management").*

Explore *this part of the cluster management interface.*

Register *a Hadoop version.*

Add *an operating system.*

Answer/discuss *the question(s) below in your lab report.*

Question Set I: Hadoop Versions

1. *Describe in one to three sentences the type of information a cluster manager must understand.*

Remote Clusters

Click *"Remote Clusters" in the menu to the left (under "Cluster Management").*

Register *a remote cluster.*

Answer/discuss *the question(s) below in your lab report.*

Question Set J: Remote Clusters

1. *What do you think a remote cluster is?*

Ambari does its best to protect you from making mistakes. You should read all warnings etc. carefully, and do <u>your</u> best to apply changes, register new versions, change base URLs etc. The important thing is not to actually register new versions, etc.; the goal is to learn.

However, don't be afraid to apply changes. This is a sandbox, i.e. a safe place for you to play and experiment.

After managing your cluster for a while, it is not unlikely that you want to, once again, reset your Azure Labs VM; see page 22.

Authorization and User Accounts

***Go to** the Ambari Administration interface (i.e. http://localhost:8080)*

Service Accounts

***Click** "Service Accounts" in the menu to the left.*

***Answer/discuss** the question(s) below in your lab report.*

Question Set K: Service Accounts

1. *What is it that we see?*

Ambari User Accounts

So far, we have been logged in as the Ambari "**admin**" user. The Hortonworks Sandbox comes with some other configured user accounts, which we will look at now.

Navigate *to the Cluster Management view; see the procedure in "Cluster Management" on page 12.*

Click *"Users" in the menu to the left. Make a mental or physical note of the users that are displayed.*

Add *some new users. Explore the various options and settings available.*

Sign out *from Ambari; see Figure 7.*

Log in *with the other user accounts that you created, and/or those that you noted three steps ago. (The passwords are listed in Table 1 on page 19.)*

Explore *the Ambari user interface as the new user.*

Answer/discuss *the question(s) below in your lab report.*

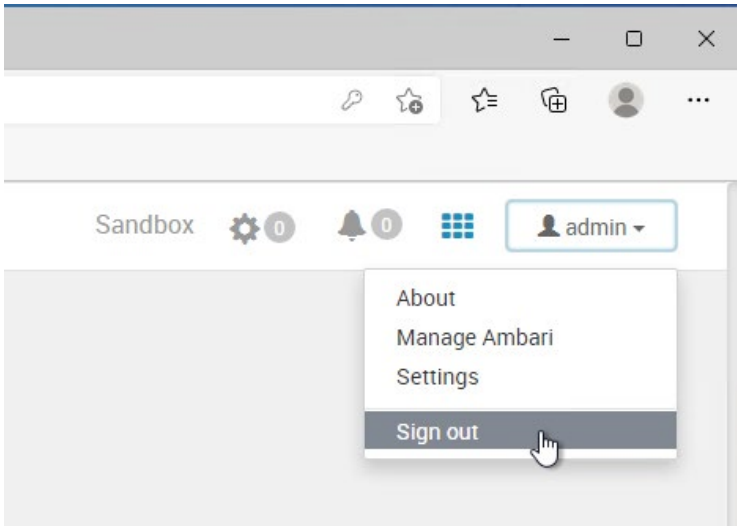

Figure 7: Signing out from Ambari.

Question Set L: Authorization and Users

1. *What information and actions are available in the Ambari user interface, depending on the "User Access" setting? (List three things per user, that are different from some other user.)*
2. *What kind of work role do you think this user is intended for?*

Cluster Management In Linux

In this section, we step outside Ambari and log in with a Linux user account directly to the "sandbox-hdp" *Docker container*.

***Navigate to** the Shell Web Client:*

- ***Click** the "Shell In a Box" icon on the Favorites bar.*

OR

- ***Navigate** to localhost:4200.*

Log in using:

Username: ***root***

Password: ***Gik2q3_VT***

See Figure 8.

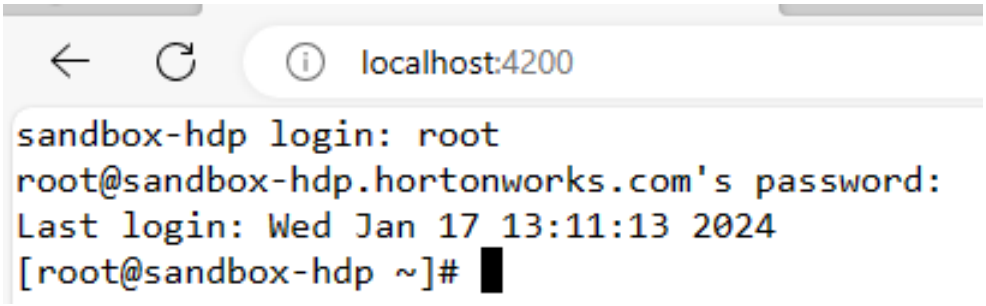

Figure 8: The Shell Web Client.

If you are new to Linux/Unix, you may want to have "Appendix B: Linux Fundamentals" (page 20) within reach.

Go to *the* `/sbin` *folder and list the contents:*

```
cd /sbin

ls
```

Go to *the* `/home` *folder and list the contents.*

Go to *the* `/hadoop` *folder, list the contents, and* ***explore*** *it further.*

Change *the password of the Ambari user "admin" with the command:*

```
ambari-admin-password-reset
```

Try *some of the Hadoop-related commands that you find at your own leisure.*

Answer/discuss *the question(s) below in your lab report.*

Question Set M: Cluster Management In Linux

1. *Describe how terms and concepts, Hadoop services, Ambari users, etc. that you have met earlier in these instructions, seem to relate to folders in the file structure.*
2. *Why, do you think, is there a specific command to reset the password for the Ambari user "admin"?*

Appendix A: User Accounts

Table 1 provides the fundamental information about user accounts to use in the lab.

Username	***Password***	***Special Information***
root	*Gik2q3_VT*	*Linux root user; see "Appendix B: Linux Fundamentals", page 20*
admin	*Gik2q3_VT*	*Ambari main administrator*
raj_ops	*raj_ops*	
maria_dev	*maria_dev*	
holger_gov	*holger_gov*	
amy_ds	*amy_ds*	

Table 1: User accounts.

Note: *It is not necessarily possible to use all of these user accounts in all parts of the lab. It is part of the lab to explore and learn the differences between different types of accounts.*

Appendix B: Linux Fundamentals

In the section "Cluster Management In Linux" on page 17, you log in to the "sandbox-hdp" *Docker container* running Linux and are met by a command prompt.

This section may be useful if you are entirely new to Linux/Unix, as it describes some basic features and commands.

The `root` User

The user account "root" is the superuser on a Linux/Unix system, i.e. this account has all privileges.

File Structure

The following things are worth to note:

- Directories are separated with forward slash "/" (division sign), not backslash "\" as in Windows.

Folder	***Description***
`/`	*The root of the file tree. (There is no drive letter as in Windows.)*
`/bin`	*Contains binary files, i.e. programs and commands.*
`/sbin`	*Contains "special binaries", i.e. more programs and commands.*
`/home`	*Contains home directories for all users.*
`~`	*Shortcut to the logged in user's home directory, e.g.* `/home/root`

Table 2: Important folders.

Commands

Table 3 lists some of the most useful commands in this lab series.

Command	***Description***
`cd`	*Change directory*
`cat`	*Print contents of a file*
`cd ..`	*Change directory: go up one level*
`cd ~`	*Change directory: go to the current user's home directory*
`help`	*Prints general help*
`help` *commandname*	*Prints help on a specific command*
`ls`	*List the contents of a folder*
`man` *commandname*	*Prints the manual for a specific command*
`pwd`	*Print working directory*

Table 3: Linux/Unix commands.

Appendix C: Troubleshooting

For problems outside Hadoop, i.e. with the *Azure Labs VM*, *Remote Desktop Connection*, and *Docker*, see the Troubleshooting section in *Lab Instructions Part 1*.

Also, the Troubleshooting section in *Lab Instructions Part 1* mentions the virtue of patience, especially while Hadoop is starting.

Reset Azure Labs VM

If the Hadoop Sandbox does not seem to feel well, consider **resetting the Azure Labs VM** to the exact same template image you had at the beginning of *Lab Instructions Part 1*.

***Back up** any custom data from your Azure Labs VM, such as MapReduce programs.*

***Ask** the teacher to reset your VM; see Figure 9.*

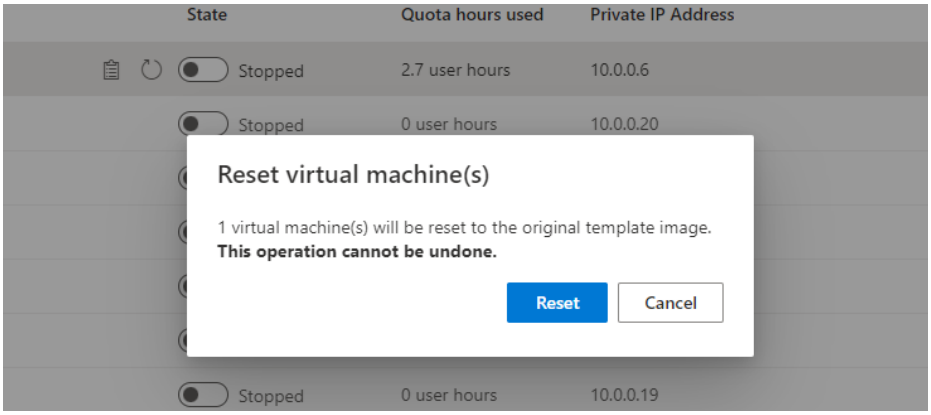

Figure 9: Resetting the Azure Labs Virtual Machine.

Unfortunately, it seems as only the teacher has access to the reset functionality. Contact your teacher and he/she will reset it asap.