BA-BEAD Big Data Engineering for Analytics

CDH Workshop Series

1. Lab Setup Instructions

In Academic Partnership with:

cloudera®

ACADEMIC PARTNER

NUS National University of Singapore | ISS INSTITUTE OF SYSTEMS SCIENCE

Contents

Homework

This Setup instruction document provides 'General Notes' for the further labs regarding use of a Virtual Machine running the CentOS Linux distribution. The distributed Virtual Machine (VM in short) has CDH (Cloudera's Distribution, including Apache Hadoop) is installed in Pseudo-Distributed mode. Pseudo-Distributed mode is a method of running Hadoop whereby all Hadoop daemons run on the same machine. It is, essentially, a cluster consisting of a single machine. It works just like a larger Hadoop cluster, the only difference (apart from speed, of course!) being that the block replication factor is set to 1, since there is only a single DataNode available.

Prerequisite

This course is designed for developers and engineers who have programming experience. Apache Spark examples and homework labs are presented in Scala or Python, therefore, the ability to program in one of those languages is required. Basic familiarity with the Linux command line is assumed. Basic knowledge of SQL is helpful; prior knowledge of Hadoop is not required.

Software Essentials

This courses will use a VMware Virtual Machine (VM), which is configured with everything required for the class. Each student should have a Windows or Macintosh computer on which to run the VM. Below are the additional computer requirements:

- Minimum RAM required: 8GB
- Minimum Free Disk Space: 25GB
- VMware Player or Virtual Box Player
- Student machines must support a 64-bit VMware guest image. Please check the relevant machine configuration requirements with your machine before proceeding with installation.
- Student machines must have VT-x virtualization support enabled in the BIOS.

VMware / VBox Player

To download VMware Player:

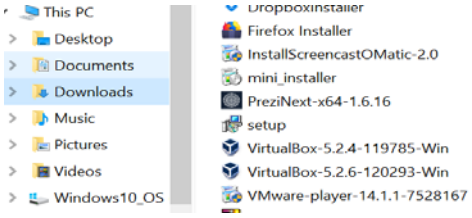
- Navigate to the VMware or Virtual Box Download Center.
- Locate the appropriate Player.
- Select the installer from the list according to your host operating system.
- Click Download.

Notes:

You can only have one version of Player installed at once. You must uninstall any previous version of Player before installing a new version.

Example to install VirtualBox Player on a Windows host:

Download site: https://www.virtualbox.org/wiki/Downloads



- Log in to the Windows host.
- Open the folder where the Vbox Player installer was downloaded. The default location is the Downloads folder for the user account on the Windows host.
- Right-click the installer executable file and click Run as Administrator.
- Follow the on-screen instructions to finish the installation.
- Restart the host machine.

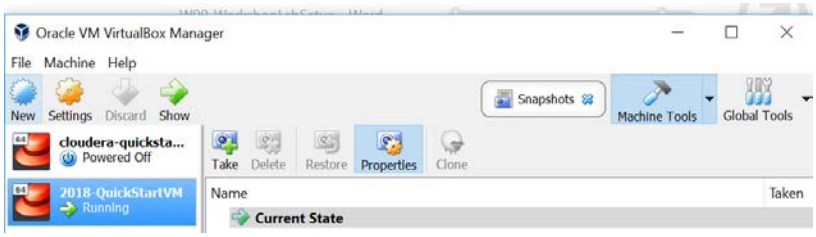
You can refer to the additional support video at: https://youtu.be/hr6JZfusDLQ

Obtaining the Quick VM from Cloudera

This is the link for download: https://www.cloudera.com/downloads/quickstart_vms/5-12.html

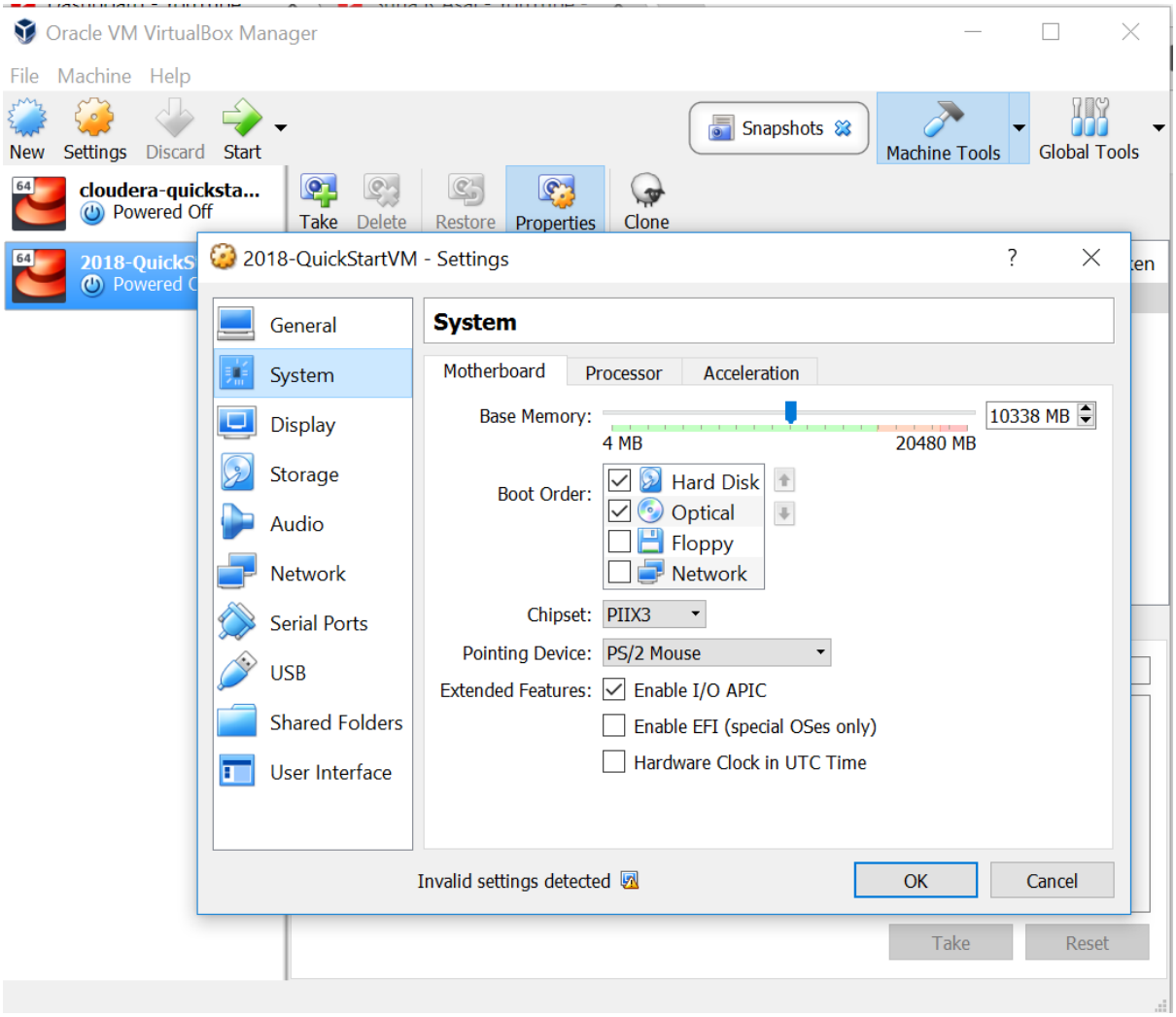
Cloudera QuickStart VMs (single-node cluster) make it easy to quickly get hands-on with CDH for testing, demo, and self-learning purposes, and include Cloudera Manager for managing your cluster. Cloudera QuickStart VM also includes a tutorial, sample data, and scripts for getting started.

The official documentation is available from:
https://www.cloudera.com/documentation/enterprise/latest/topics/introduction.html



Getting Started

In most cases, the QuickStart VM requires no administration beyond managing the installed products and services.



Accounts

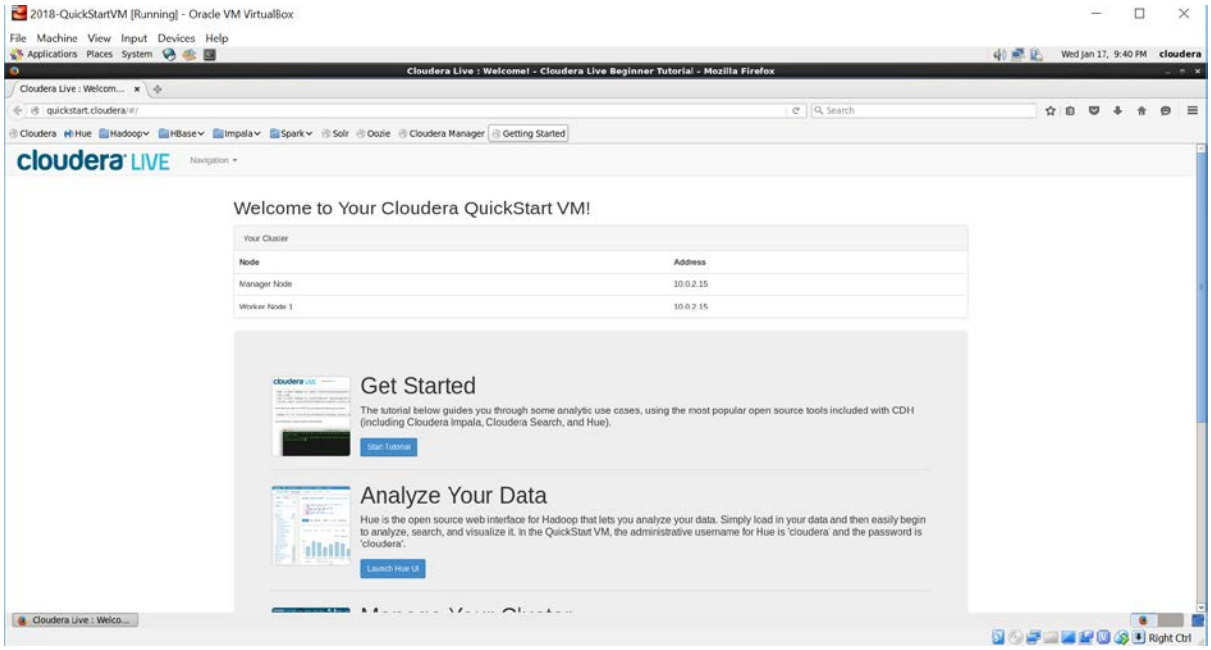
Once you launch the VM, you are automatically logged in as the cloudera user. The account details are:

- username: cloudera
- password: cloudera

The cloudera account has sudo privileges in the VM. The root account password is cloudera.

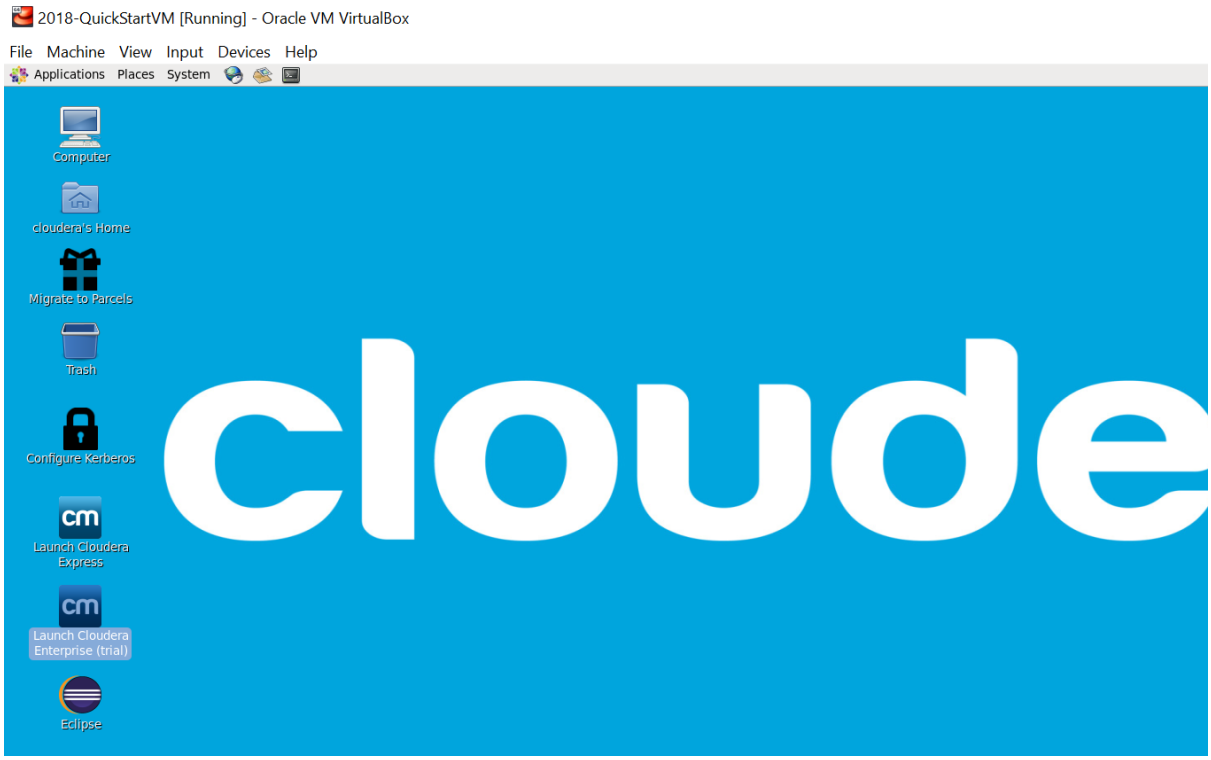
The root MySQL password (and the password for other MySQL user accounts) is also cloudera.

Hue and Cloudera Manager use the same credentials.



Start-up Services

Before starting the homework assignments, run the course express or enterprise trail edition script in the quick VM desktop window:

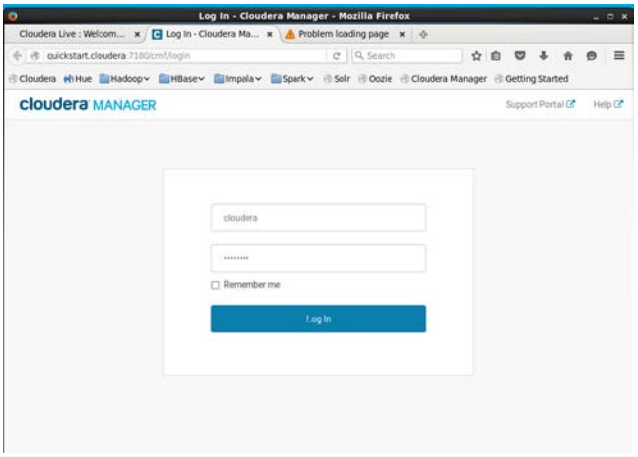


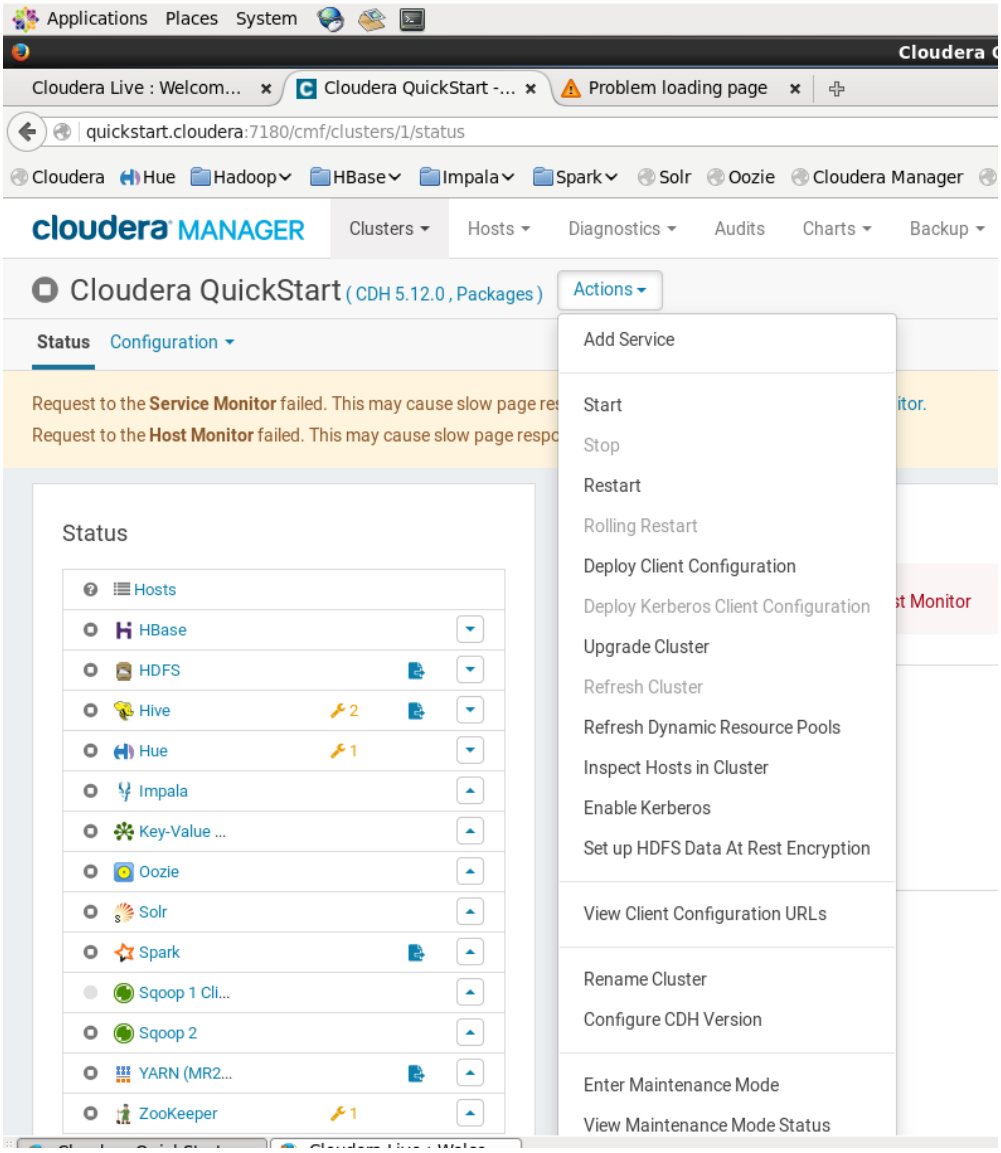
This script will enable services and set up any data required for the course.

Working with the Virtual Machine

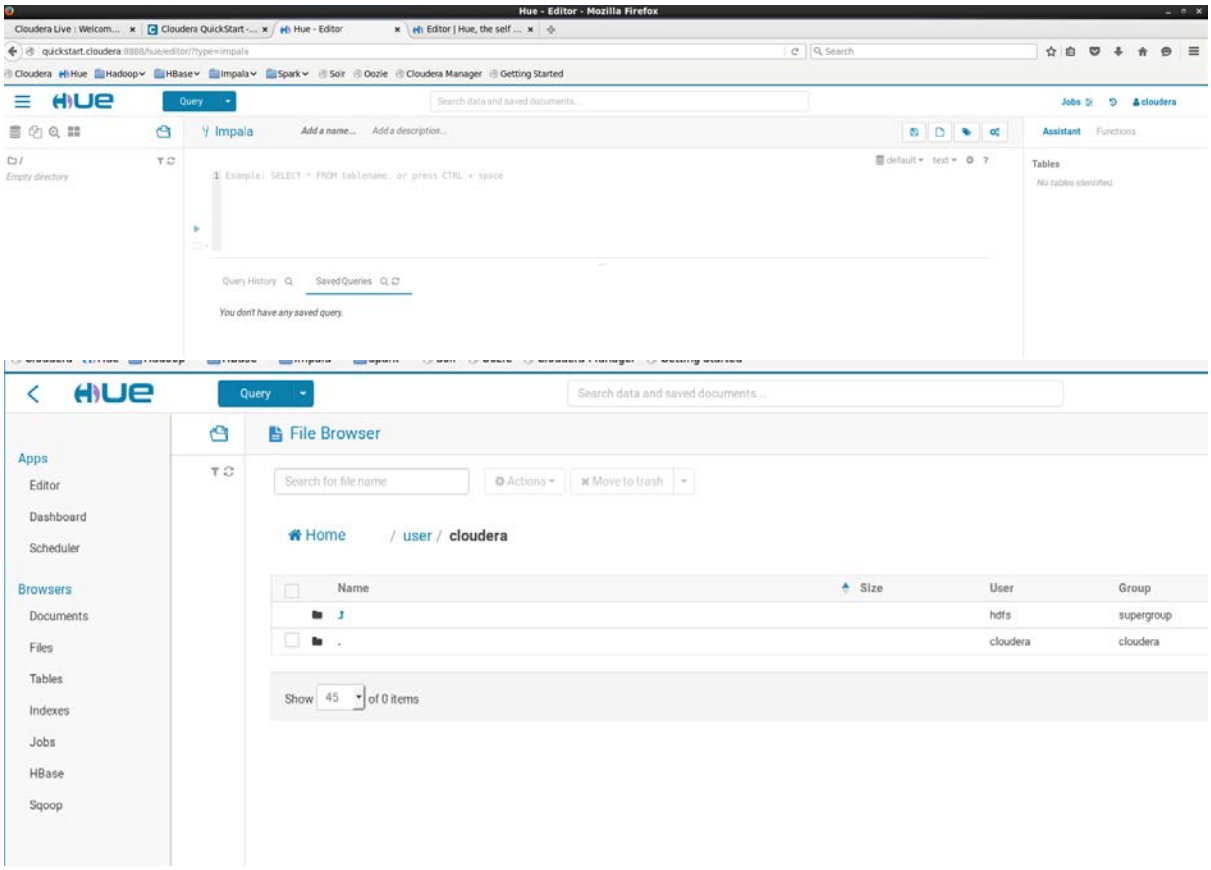
1. The VM is set to automatically log in as the user `cloudera`. Should you log out at any time, you can log back in as the user `cloudera` with the password training.

2. Once the enterprise trial scripts are activated, the cloudera cluster can be well managed using cloudera manager page, Cluster's Start Services - that looks similar to the screen shot below:





3. Should you need it, the root password is `cloudera`. You may be prompted for this if, for example, you want to change the keyboard layout. In general, you should not need this password since the `cloudera` user has unlimited sudo privileges. (For example Hue Log In)



4. In some command-line steps in the `homework`, you will see lines like this:

```
$ hdfs dfs -put shakespeare \
/user/cloudera/shakespeare
```

The dollar sign (**$**) at the beginning of each line indicates the Linux shell prompt. The actual prompt will include additional information (e.g., `[cloudera@localhost workspace]$`) but this is omitted from these instructions for brevity.

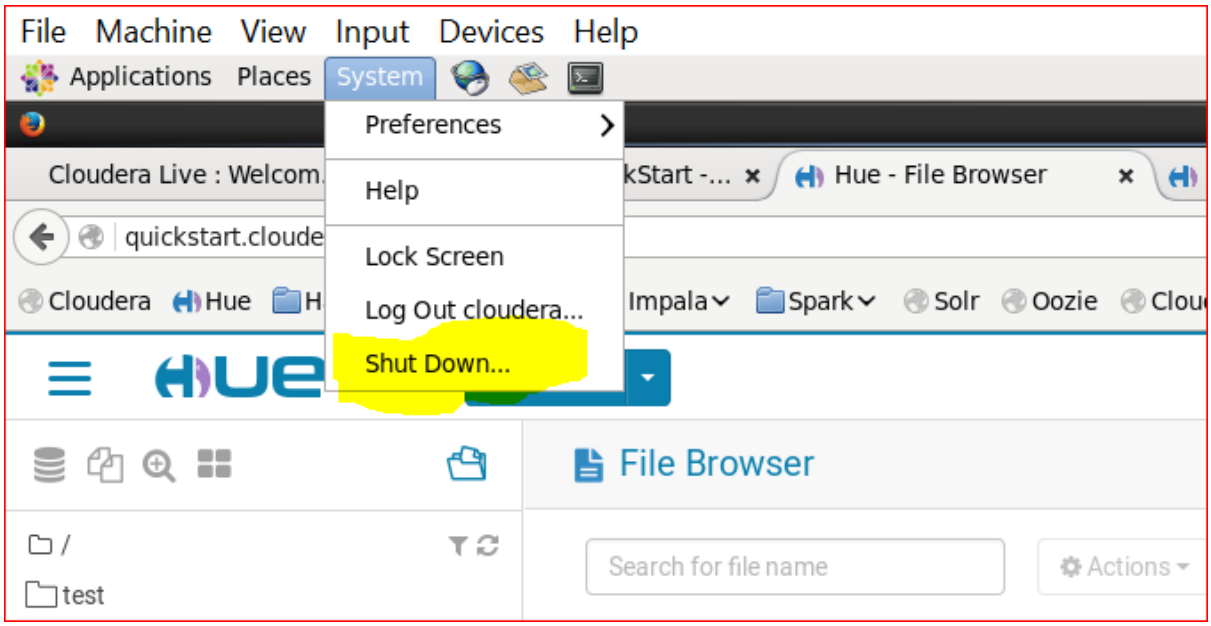
The backslash (\) at the end of the first line signifies that the command is not completed, and continues on the next line. You can enter the code exactly as shown (on two lines), or you can enter it on a single line. If you do the latter, you should *not* type in the backslash.

Points to note during the homework

As the exercises progress, and you gain more familiarity with Hadoop and Spark, we provide fewer step-by-step instructions; as in the real world, we merely give you a requirement and it's up to you to solve the problem!

Switching off the Virtual Image.

When you have finished working with the image you can properly close the machine by choosing the Shut Down option under the System Menu of the Cent OS Linux Virtual Image.



---------END OF DOCUMENT----------