

Tutorial: Hadoop VM Installation

ITCS 3190: Cloud Computing for Data Analysis

Prerequisite

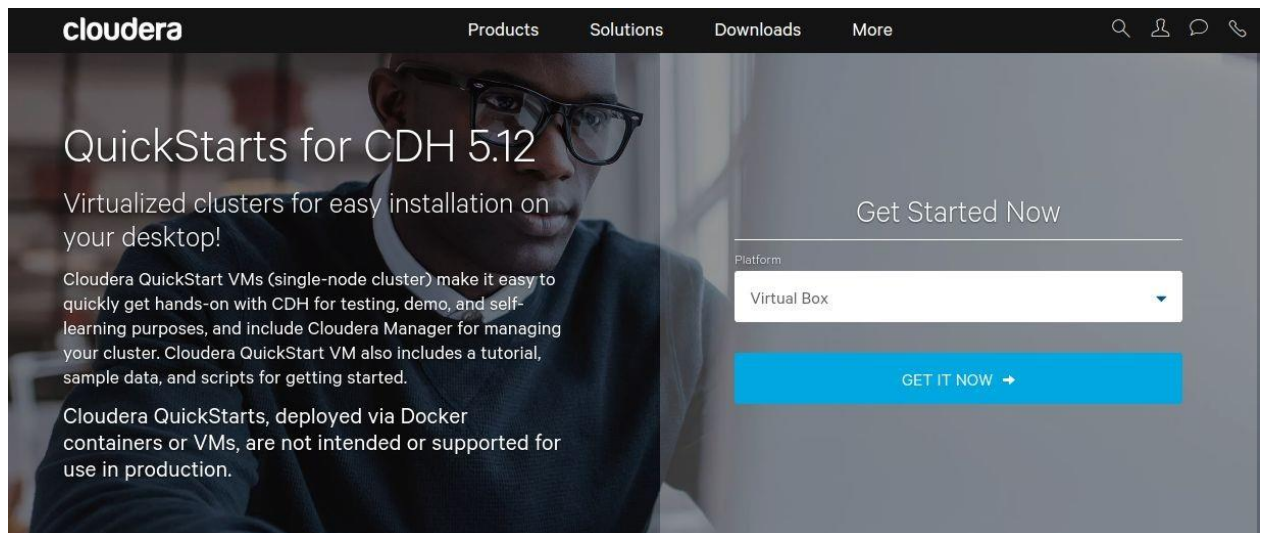
Any 64-bit Windows, Mac, or Linux machine, 4GB RAM minimum.

1. Download Hadoop Virtual Machine

1.1 Download Cloudera QuickStart Virtual Machine (VM)

Go to https://www.cloudera.com/downloads/quickstart_vms/5-12.html

1.2 Choose “Virtual Box” as platform



After the zip file is fully downloaded, unzip it.

2. Install Virtual Box

2.1 Download Virtual Box

Go to <https://www.virtualbox.org/wiki/Downloads>, and download the Virtual Box for your OS.

- **VirtualBox 5.1.26 platform packages.** The binaries are released under the terms of the GPL version 2.
 - [Windows hosts](#)
 - [OS X hosts](#)
 - [Linux distributions](#)
 - [Solaris hosts](#)

Once it is installed, you can use it to run Cloudera QuickStart VM.

3. Start Cloudera QuickStart VM

3.1 Import Cloudera QuickStart VM

The unzipped folder from step 1.2 will contain .ovf file, double clicking will open Virtual Box . click Import.



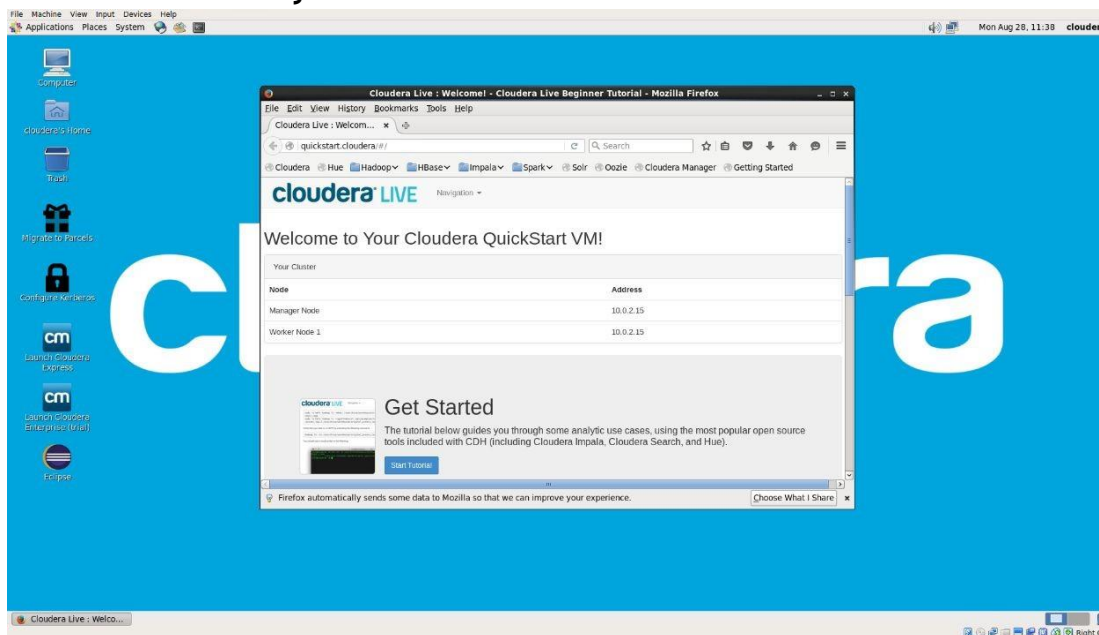
While importing

1. If your system has 8GB or more RAM , select RAM as 4096 MB
2. Else if your system has 4GB RAM, select RAM as 2048 MB.

3.2 After import is completed, select the Cloudera QuickStart Virtual Machine and then click “Start” to start it.



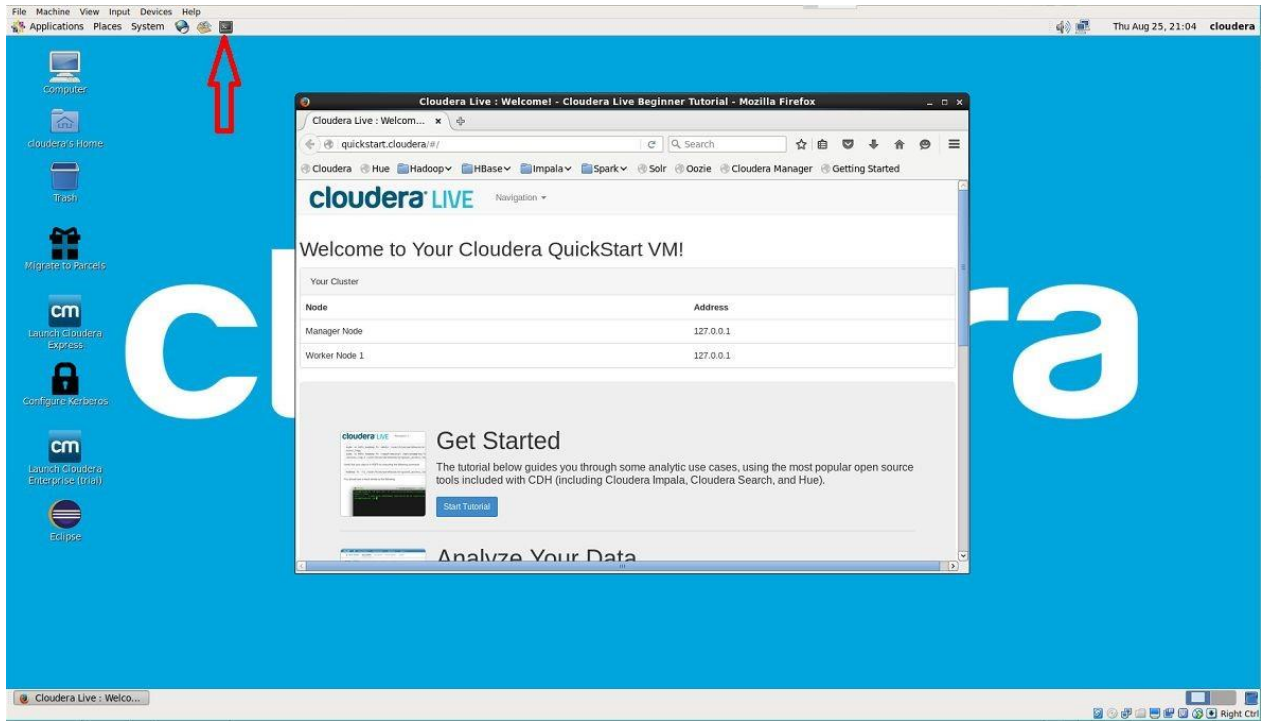
3.3 Be patient as it will take minutes until the VM is started. Once started, the browser will be automatically launched as below.



4. Basic Linux usage

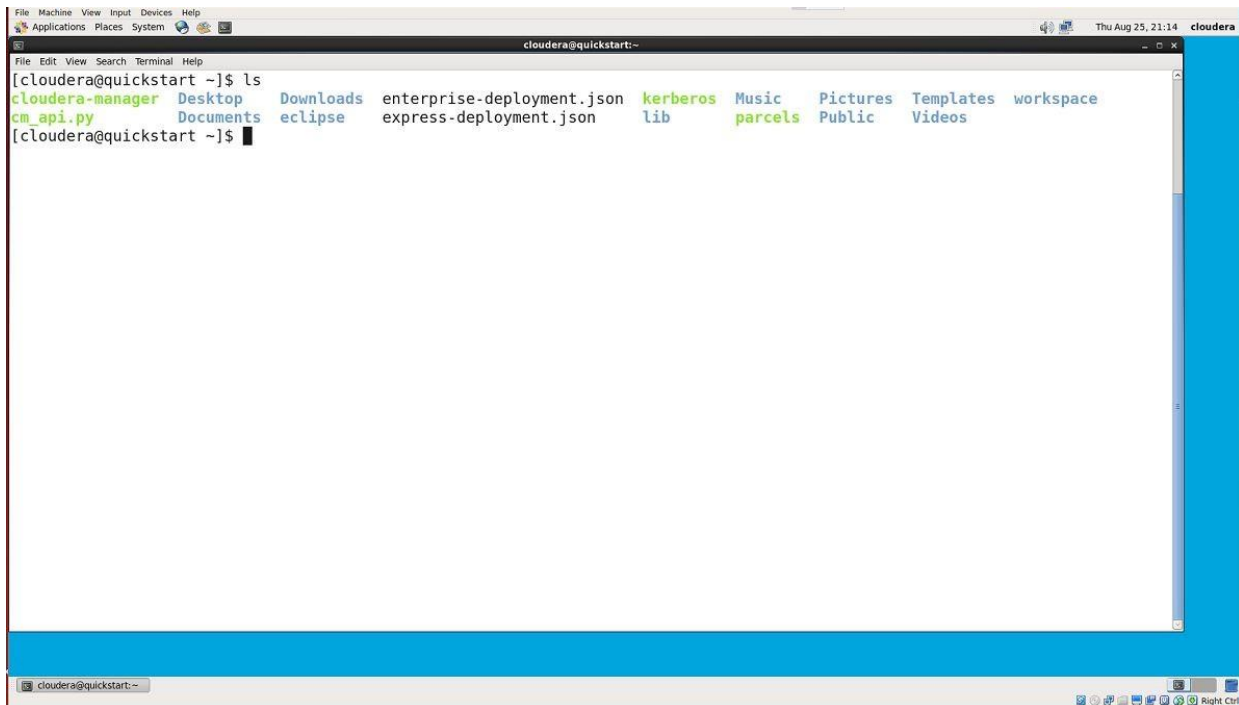
For those of you who do not have any experience using Linux, don't panic. In this particular tutorial, all you have to know is that Linux is a command line based operating system, and most operations are done in the terminal. From now on, all the operations are done inside the VM.

4.1 Open a terminal



4.2 Run a command

Let us type "ls", which mean list all files and directories in current directory.



```
cloudera@quickstart:~$ ls
cloudera-manager  Desktop  Downloads  enterprise-deployment.json  kerberos  Music  Pictures  Templates  workspace
cm_api.py         Documents eclipse     express-deployment.json    lib        parcels  Public    Videos
```

5. Run Hadoop Example

Follow the tutorials to compile and run the Word Count example.

https://www.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_usage.html

The WordCount.java source file will be provided along with this assignment.
It can be copied and pasted.

N.B.: if you get below exception, just ignore it:

WARN hdfs.DFSClient: Caught exception

java.lang.InterruptedExcep

For explanation of the source code, please refer to the link below.

https://www.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount1_source.html

This is the more detailed explanation of the WordCount example. Please read and understand it.

https://www.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_walk_through.html