

Hadoop via Cloudera

Big Data using Hadoop



December 10, 2015

Akolom Gebrezgabhier (ID#108229)

Maged Samaan (ID# 984790)

Content

1. Introduction
2. How to download Cloudera VM
3. Install Virtual Box
4. Using Cloudera VM
5. Pairs Approach Input, Output & Patch file
6. Stipes Approach Input, Output & Patch file
7. Hybrid Approach Input, Output & Patch file
8. Comparison Between Virtual Box and VMware
9. Conclusion

**Introduction Cloudera**

Cloudera is a ready to use virtual machine image that can be used using (VMware, Virtual box or KVM). It has Hadoop platform installed with all its prerequisites, tools, plug-ins, Java and Eclipse to allow you to write MapReduce and run them on HDFS.

To be able to download the image from cloudear.com website, you need to register first and then you will be able to download the image from here (<http://www.cloudera.com/content/www/en-us/downloads/quickstart_vms/5-5.html>).

We are using Cloudera QuickStart VMs v5.5 which have some Prerequisites listed below:

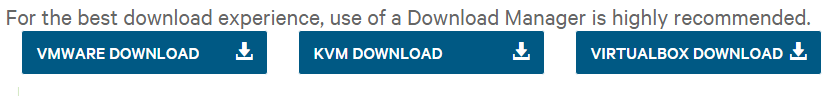
1. These 64-bit VMs require a 64-bit host OS and a virtualization product that can support a 64-bit guest OS.
2. To use a VMware VM, you must use a player compatible with WorkStation 8.x or higher:
   1. Player 4.x or higher
   2. Fusion 4.x or higher
3. Older versions of WorkStation can be used to create a new VM using the same virtual disk (VMDK file), but some features in VMware Tools are not available.
4. The amount of RAM required varies but for this version 4GB is recommended.

**How to download Cloudera VM**

1. Got to the website, (<http://www.cloudera.com/>) and sign up.
2. Go to the current version download page link (<http://www.cloudera.com/content/www/en-us/downloads/quickstart_vms/5-5.html>), or you can change the version from this dropdown list as below, and click the Get it now button.

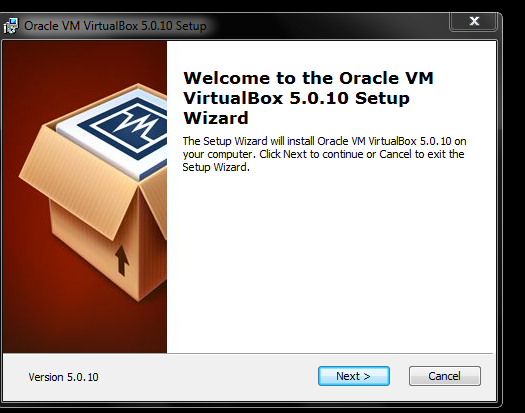


1. Select your virtual image player, you have 3 options Virtual box, VMware player, and KVM as below.

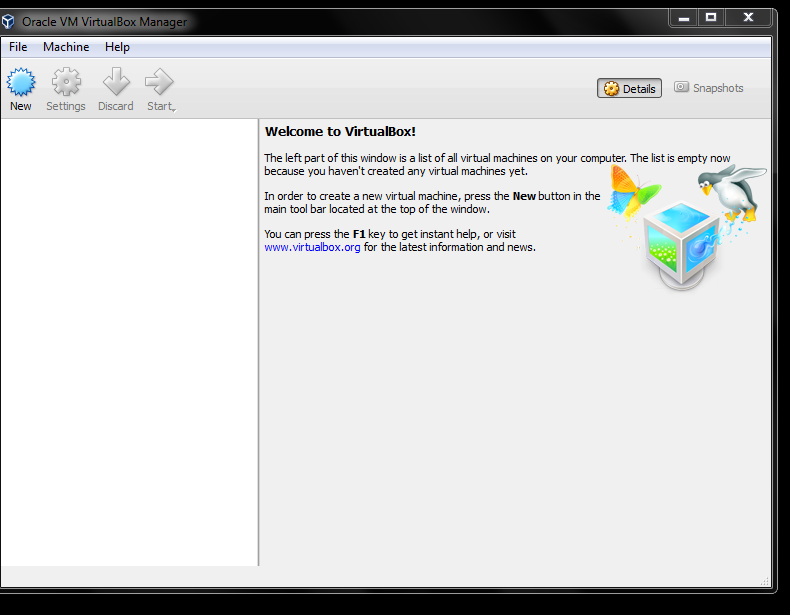


**Install Virtual Box**

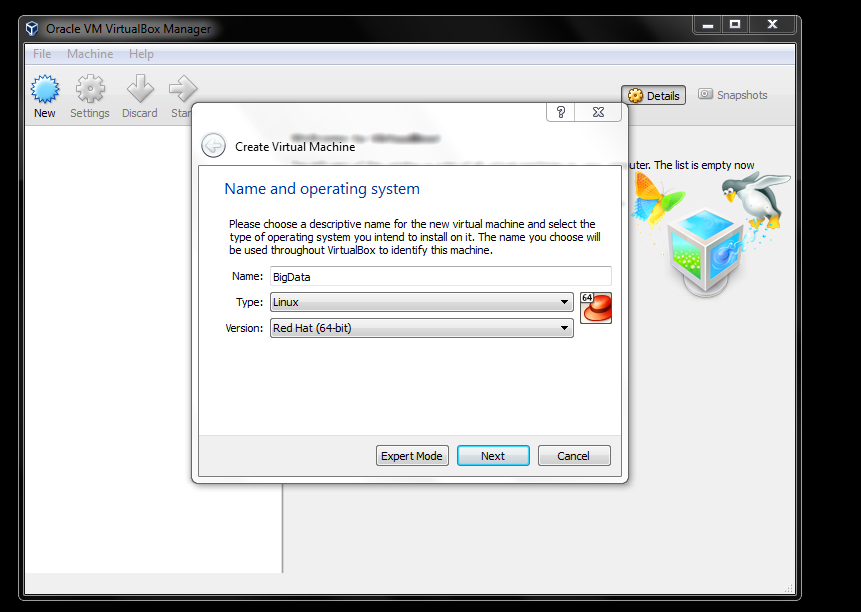
download Virtual box from here (<https://www.virtualbox.org/wiki/Downloads>) and Install it in order to run Virtual Machine Images (VMs) on local machine.



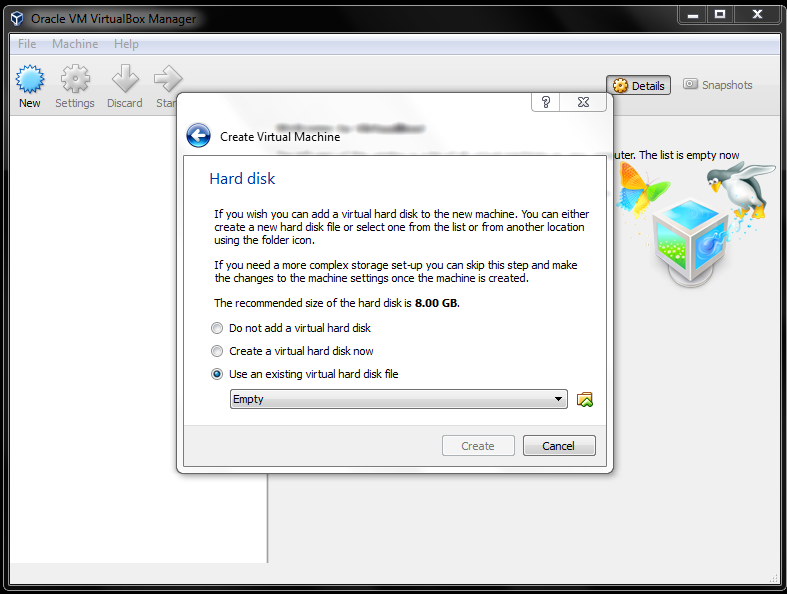
Open the Virtual box and Add the new Image



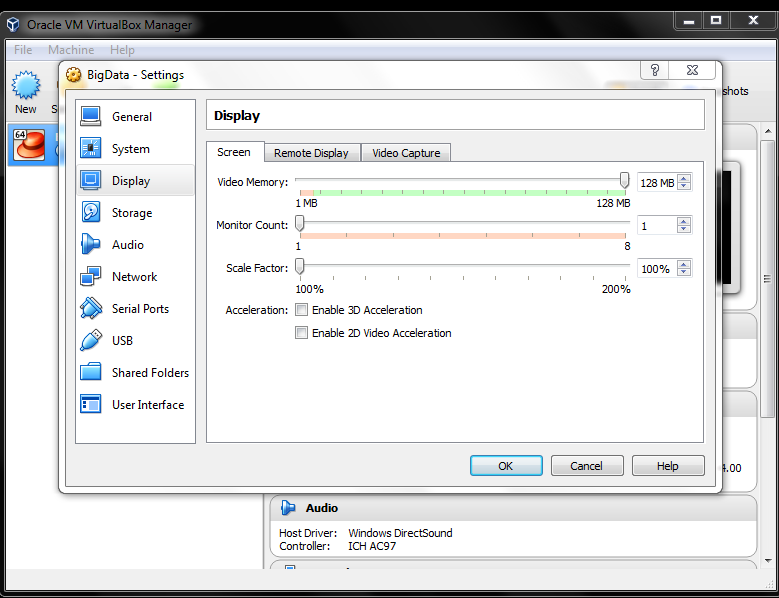
Select Linux from The Type dropdown list, and Red Hat 64 bit from Version Drop down List



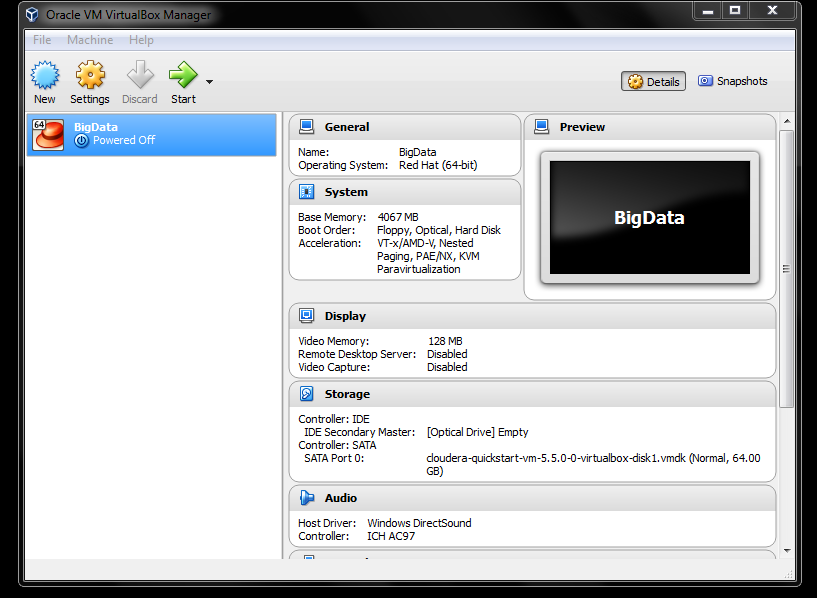
Use Existing file, and select Cloudera VM image you downladed earlier



Set your setting that will be suitable for your environment (RAM size, No of processors, etc…)



Your image now is ready, just hit start and start coding



**Using Cloudera VM**

You now can start writing your MapReduce Program, by using Eclipse IDE (which already installed) or you can install other IDE. By default, a MapReduce template is already presented to us, so if you don’t want to start from scratch, just adjust it and start writing your logic. Also you need to add reference to external Hadoop Client library explicitly, it doesn’t come with default eclipse template.

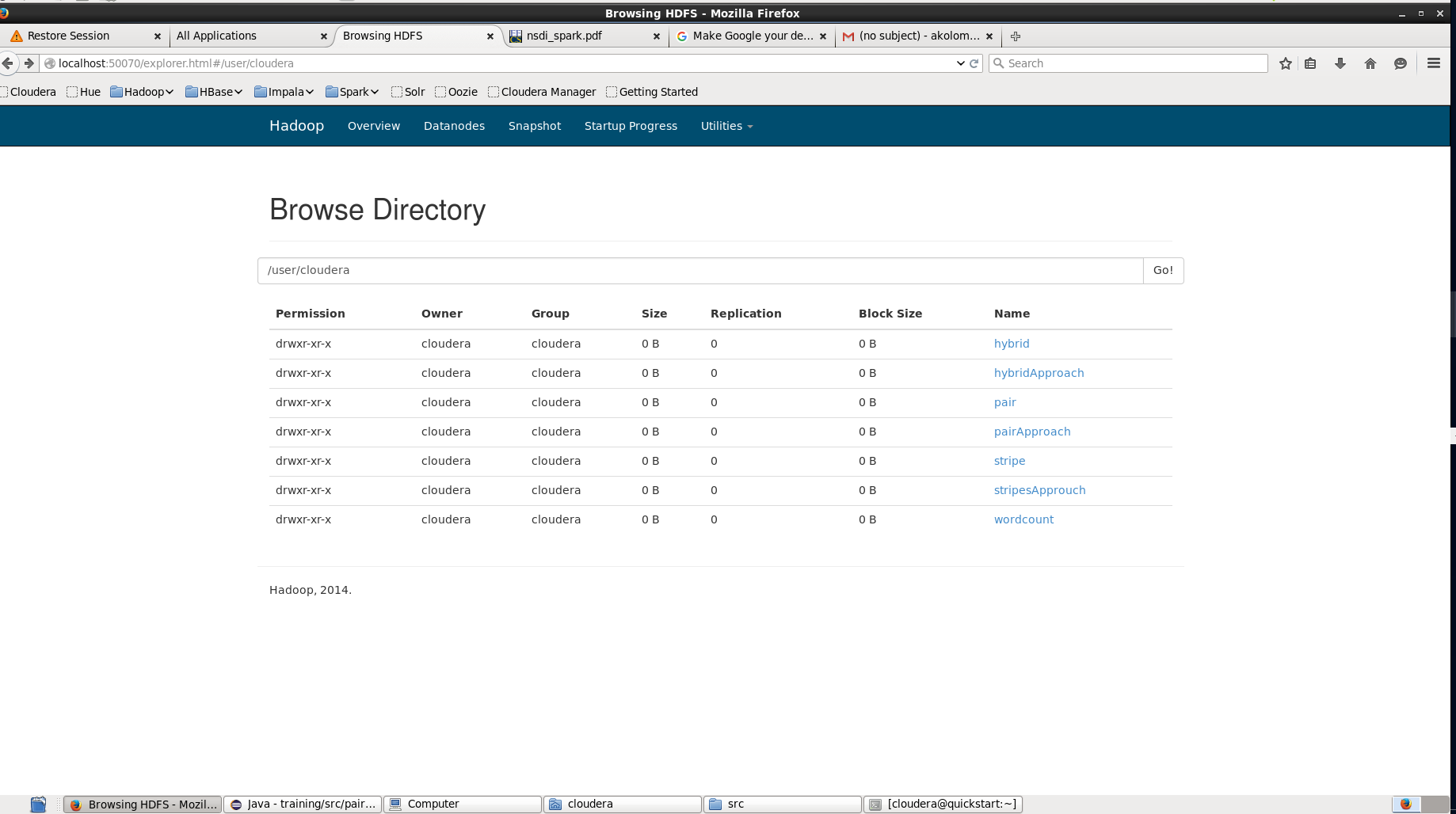
* compile a package (.jar) with mapper and reducer parts (Note: you must include the util package with any approach, as it contains common classes)
* Execute hadoop job by handover jar file, main class name, input location and output location via following terminal commands (for the pairs approach)

hdfs dfs –mkdir pairs/input

hdfs dfs –put input pairs

hadoop jar pairsApproach.jar pairs/input pairs/output

* you can find the output by browsing <http://localhost:50070/> (see the below image)



* you can read the output files using the below command:

hdfs dfs –cat pairs/output/part-r-00000

hdfs dfs –cat pairs/output/part-r-00001

**Pairs Approach Input, Output & Patch file**

* input

425 450 466 471

472 522 525 475

545 440 422 435

545 522 525 475

* output

422,435 1.0

425,450 0.3333333333333333

425,466 0.3333333333333333

425,471 0.3333333333333333

440,422 0.5

440,435 0.5

450,466 0.5

450,471 0.5

466,471 1.0

472,475 0.3333333333333333

472,522 0.3333333333333333

472,525 0.3333333333333333

[cloudera@quickstart ~]$ hdfs dfs -cat pair/output/part-r-00001

522,475 0.5

522,525 0.5

525,475 1.0

545,422 0.16666666666666666

545,435 0.16666666666666666

545,440 0.16666666666666666

545,475 0.16666666666666666

545,522 0.16666666666666666

545,525 0.16666666666666666

* command lines

hdfs dfs -mkdir pairs

hdfs dfs -put /home/cloudera/input pairs/

hadoop jar /home/cloudera/pair.jar pairs/input pairs/output

hdfs dfs -cat pairs/output/part-r-00000

hdfs dfs -cat pairs/output/part-r-00001

**Stipes Approach Input, Output & Patch file**

* input

425 450 466 471

472 522 525 475

545 440 422 435

545 522 525 475

* output

422 [(435,1.0)]

425 [(450,0.3333333333333333)(471,0.3333333333333333)(466,0.3333333333333333)]

435 []

440 [(435,0.5)(422,0.5)]

450 [(471,0.5)(466,0.5)]

466 [(471,1.0)]

471 []

472 [(525,0.3333333333333333)(475,0.3333333333333333)(522,0.3333333333333333)]

475 []

[cloudera@quickstart ~]$ hdfs dfs -cat stripe/output/part-r-00001

522 [(525,0.5)(475,0.5)]

525 [(475,1.0)]

545 [(440,0.16666666666666666)(435,0.16666666666666666)(422,0.16666666666666666)(525,0.16666666666666666)(475,0.16666666666666666)(522,0.16666666666666666)]

* command lines

hdfs dfs -mkdir stripes

hdfs dfs -put /home/cloudera/input stripes/

hadoop jar /home/cloudera/stripes.jar stripes/input stripes/output

hdfs dfs -cat stripes/output/part-r-00000

hdfs dfs -cat stripes/output/part-r-00001

**Hybrid Approach Input, Output & Patch file**

* input

425 450 466 471

472 522 525 475

545 440 422 435

545 522 525 475

* output

[cloudera@quickstart ~]$ hdfs dfs -cat a/output/part-r-00000

422 [(435,1.0)]

425 [(466,0.3333333333333333)(471,0.3333333333333333)(450,0.3333333333333333)]

440 [(435,0.5)(422,0.5)]

450 [(466,0.5)(471,0.5)]

466 [(471,1.0)]

472 [(522,0.3333333333333333)(475,0.3333333333333333)(525,0.3333333333333333)]

[cloudera@quickstart ~]$ hdfs dfs -cat a/output/part-r-00001

522 [(475,0.5)(525,0.5)]

525 [(475,1.0)]

545 [(522,0.16666666666666666)(435,0.16666666666666666)(440,0.16666666666666666)(475,0.16666666666666666)(525,0.16666666666666666)(422,0.16666666666666666)]

* command lines

hdfs dfs -mkdir hybrid

hdfs dfs -put /home/cloudera/input hybrid/

hadoop jar /home/cloudera/hybrid.jar hybrid/input hybrid/output

hdfs dfs -cat hybrid/output/part-r-00000

hdfs dfs -cat hybrid/output/part-r-00001

**Comparison Between Virtual Box and VMware**

|  |  |  |
| --- | --- | --- |
|  | **Virtual Box** | **VMware** |
| **Cost** | Free | Paid for commercial use |
| **Performance** |  | VMware uses different optimizations along with binary translation and hardware assisted virtualization to speed up the whole process which Virtual Box cannot beat |
| **UEFI** | UEFI doesn't support Windows yet. You can still install OS X and Linux though | Workstation's UEFI mode supports Windows which have a hard-core UEFI boot requirement, so you can have UEFI in a VM even if the host is legacy MBR. |
| **Drag & Drop** | N/A | VM ware supports drag and drop from host to client |

**Conclusion**

The easiest way to have your Hadoop Distributed File System (HDFS), installed and working and all the tools and the prerequisites are integrated. You can use Cloudera solution. It will save you a lot of time trying to install and integrate those tools.