CRYSTAL PROJECT SETUP GUIDELINE

September 17th, 2015

Professor:	**Prem Nair**
Student Name:	**Bao Pham**
Student ID:	**984588**

Project Description

In this project you will create a crystal ball to predict events that may happen once a certain event happened.

Example: Amazon will say people who bought "item one" have bought the following items : "item two", "item three", "item four".

For the purpose of this project you can assume that historical customer data is available in the following form:

34 56 29 12 34 56 92 10 34 12 // items bought by a customer, listed in the order she bought it

18 29 12 34 79 18 56 12 34 92 // items bought by another customer, listed in the order she bought it

Let the neighborhood of X, N(X) be set of all term after X and before the next X.

Example: Let Data block be [a b c a d e]

N(a) = {b, c}, N(b) = {c, a, d, e}, N(c) = {a, d, e}, N(a) ={d, e}, N(d) = {e}, N(e) = {}

Environment Setup

Required Softwares

1. Download the Oracle VM Virtual Box of OS X host at https://www.virtualbox.org/wiki/Downloads. The file is VirtualBox-5.0.4-102546-OSX.dmg

3. Download the Cloudera QuickStart for Virtual Box at http://www.cloudera.com/content/cloudera/en/downloads/quickstart_vms/cdh-5-4-x.html. The file is cloudera-quickstart-vm-5.4.2-0-virtualbox.zip

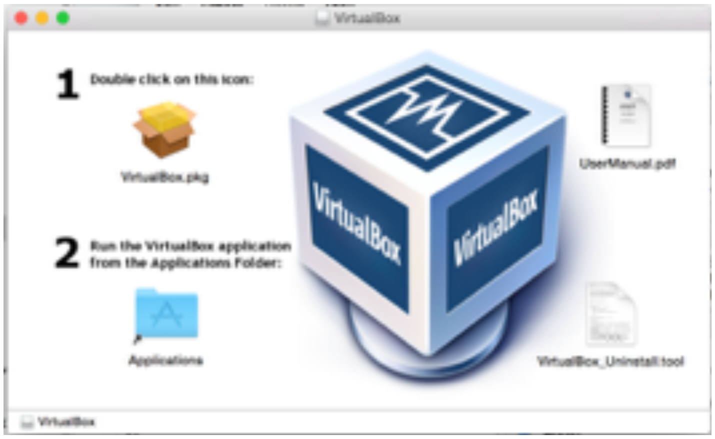
4. Download eclipse IDE for OS X at http://www.eclipse.org/downloads/packages/eclipse-ide-java-ee-developers/marsr

5. Download Apache Hadoop hadoop-2.7.1.tar at http://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.7.1/hadoop-2.7.1.tar.gz

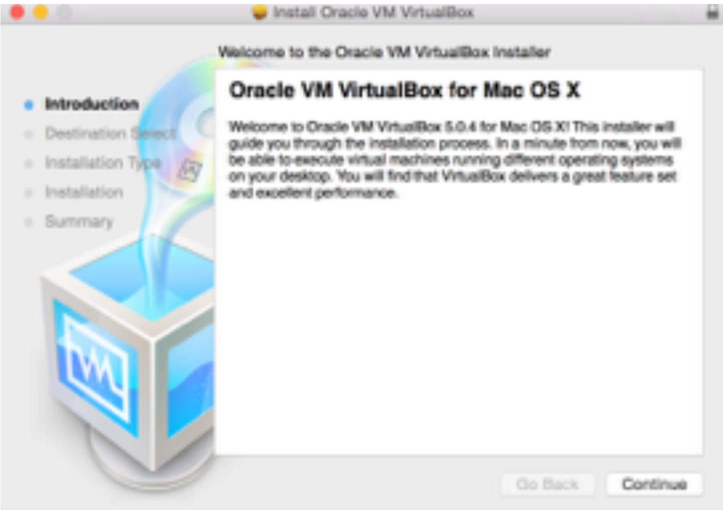
Install Cloudera

Install Virtual Box

1. Double click on VirtualBox-5.0.4-102546-OSX.dmg the following screen will be displayed


VirtualBox
1 Double click on this icon:
VirtualBox.pkg
2 Run the VirtualBox application from the Applications Folder:
Applications
UserManual.pdf
VirtualBox_Uninstall.tool
VirtualBox

2. Select option 1 “Double click on this icon” and select continue

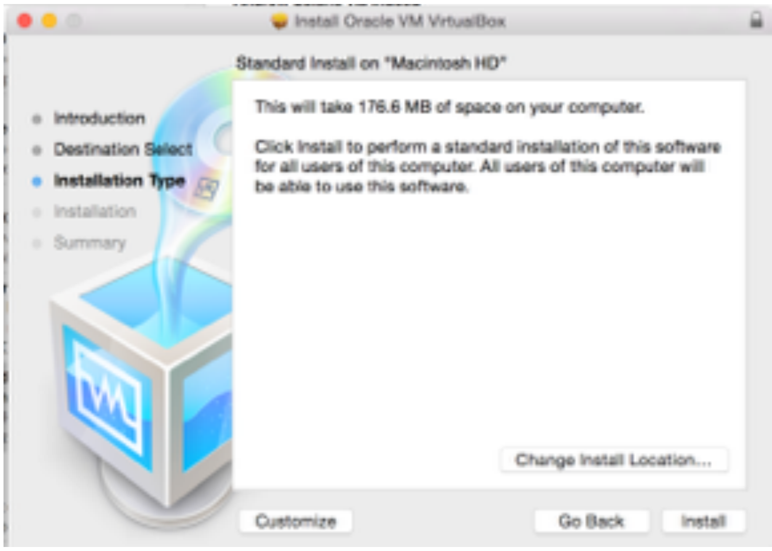


Welcome to the Oracle VM VirtualBox Installer

Oracle VM VirtualBox for Mac OS X

Welcome to Oracle VM VirtualBox 5.0.4 for Mac OS X! This installer will guide you through the installation process. In a minute from now, you will be able to execute virtual machines running different operating systems on your desktop. You will find that VirtualBox delivers a great feature set and excellent performance.

3. Select continue



Standard Install on "Macintosh HD"

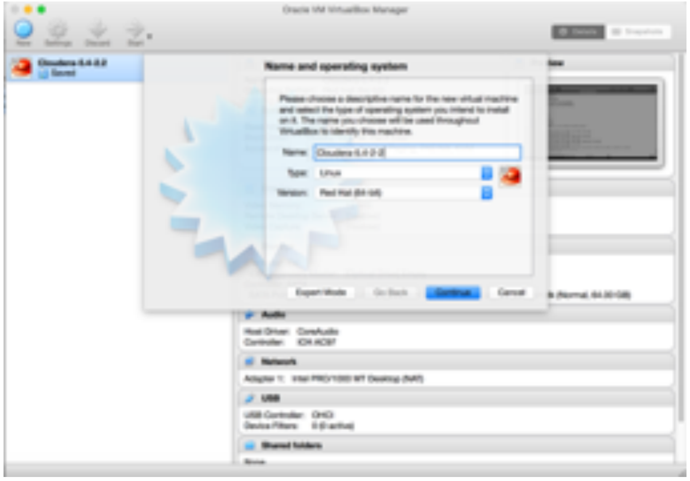
This will take 176.6 MB of space on your computer.

Click Install to perform a standard installation of this software for all users of this computer. All users of this computer will be able to use this software.

4. Select Install, this will install VM VirtualBox on the Mac

Setup Cloudera

1. Open VirtualBox application and then select New button in the menu bar and then enter "Cloudera-5.4-2-2" in the Name field, select "Linux" in the type field and "Red Had (64-bit)" in the Versions field



Name and operating system

Please choose a descriptive name for the new virtual machine and select the type of operating system you intend to install on it. The name you choose will be used throughout VirtualBox to identify this machine.

Name: Cloudera-5.4-2-2

Type: Linux

Version: Red Hat (64-bit)

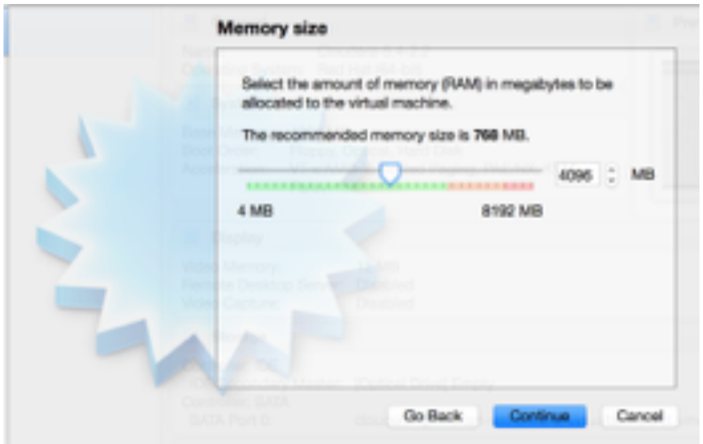
Audio
Host Driver: CoreAudio
Controller: ICH AC97

Network
Adapter 1: Intel PRO/1000 MT Desktop (NAT)

USB
USB Controller: OHCI
Device Filters: 0 (0 active)

Shared folders
None

2. Click Continue, and then select 4096 in the Memory Size screen, then click Continue



Memory size

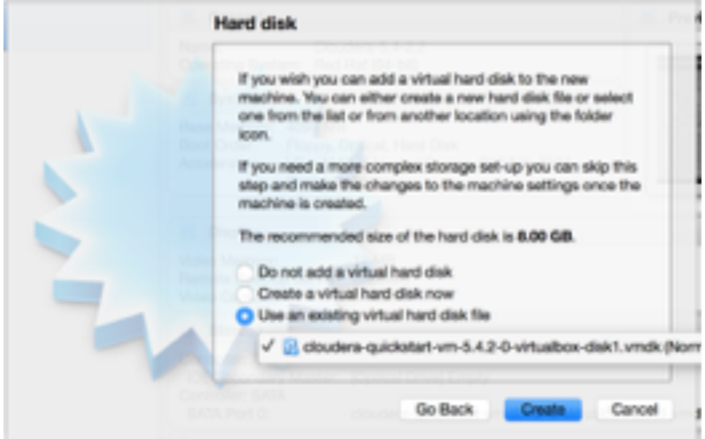
Select the amount of memory (RAM) in megabytes to be allocated to the virtual machine.

The recommended memory size is **768** MB.

4096 MB

4 MB 8192 MB

3. Choose “Use an existing virtual hard disk file” and then select the file from Cloudera QuickStart and select Create



Hard disk

If you wish you can add a virtual hard disk to the new machine. You can either create a new hard disk file or select one from the list or from another location using the folder icon.

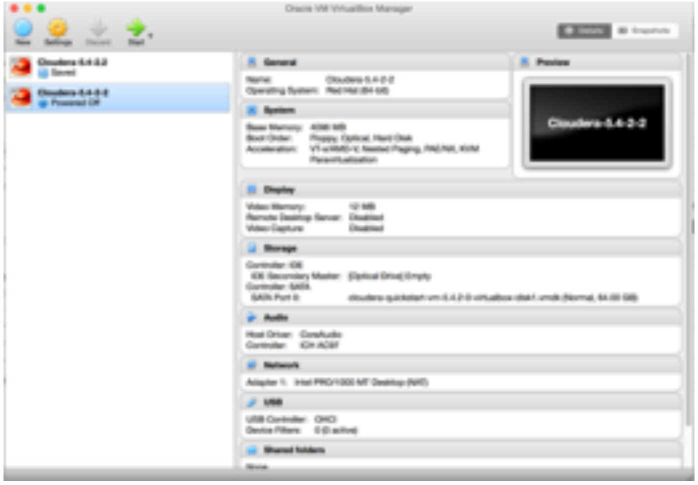
If you need a more complex storage set-up you can skip this step and make the changes to the machine settings once the machine is created.

The recommended size of the hard disk is **8.00 GB**.

○ Do not add a virtual hard disk
○ Create a virtual hard disk now
● Use an existing virtual hard disk file

✓ cloudera-quickstart-vm-5.4.2-0-virtualbox-disk1.vmdk (Norm

4. The screen will be like this



General
Name: Cloudera-5.4-2-2
Operating System: Red Hat (64-bit)

System
Base Memory: 4096 MB
Boot Order: Floppy, Optical, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX, KVM Paravirtualization

Display
Video Memory: 12 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage
Controller: IDE
IDE Secondary Master: [Optical Drive] Empty
Controller: SATA
SATA Port 0: cloudera-quickstart-vm-5.4.2-0-virtualbox-disk1.vmdk (Normal, 64.00 GB)

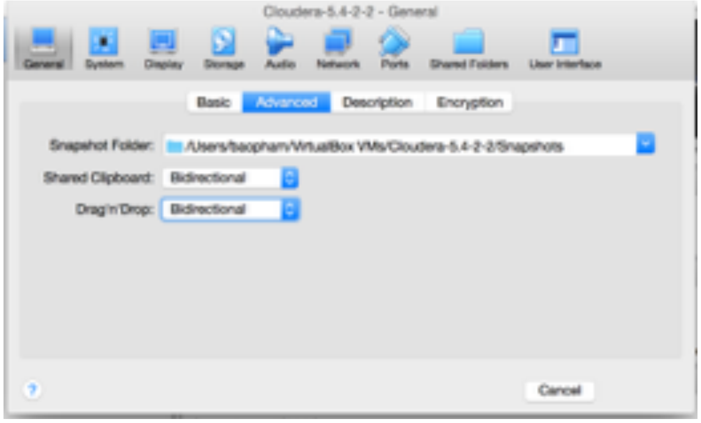
Audio
Host Driver: CoreAudio
Controller: ICH AC97

Network
Adapter 1: Intel PRO/1000 MT Desktop (NAT)

USB
USB Controller: OHCI
Device Filters: 0 (0 active)

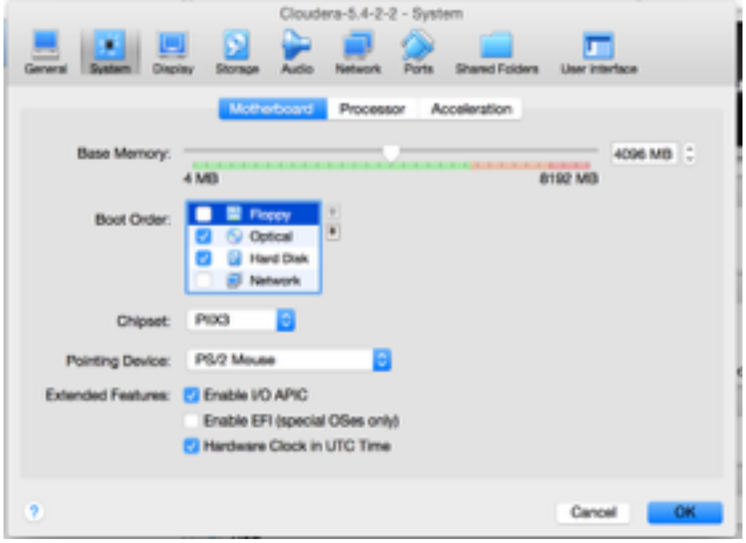
Shared folders
None

5. Select the VM which has just been created and click Settings. In the “Advanced” tab, select “Bidirectional” for Shared Clipboard and Drag ’n’ Drop field



Snapshot Folder: /Users/baopham/VirtualBox VMs/Cloudera-5.4-2-2/Snapshots
Shared Clipboard: Bidirectional
Drag'n'Drop: Bidirectional

6. Select System tab and then uncheck the “Floppy Disk” in boot order



Base Memory: 4096 MB (4 MB – 8192 MB)

Boot Order:
□ Floppy
☑ Optical
☑ Hard Disk
□ Network

Chipset: PIIX3

Pointing Device: PS/2 Mouse

Extended Features:
☑ Enable I/O APIC
□ Enable EFI (special OSes only)
☑ Hardware Clock in UTC Time

Eclipse

Unzip the eclipse and then it’s ready for usage

Apache Hadoop Library

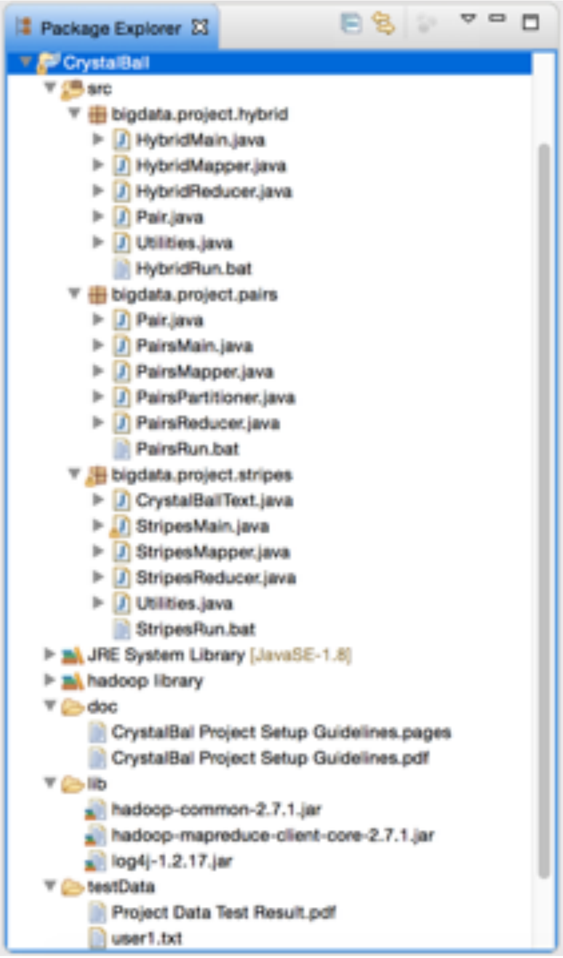
The following are the required libraries for this project. These can be gotten from Apache Hadoop files. Unzip hadoop file to hadoop folder

- hadoop-common-2.7.1.jar at hadoop/share/hadoop/common
- hadoop-mapreduce-client-core-2.7.1.jar at hadoop/share/tools/lib
- log4j-1.2.17.jar

Project Implementation

Project structure

Create a project named CrystalBall with the structure is as below



```
▼ CrystalBall
  ▼ src
    ▼ bigdata.project.hybrid
      ▶ HybridMain.java
      ▶ HybridMapper.java
      ▶ HybridReducer.java
      ▶ Pair.java
      ▶ Utilities.java
        HybridRun.bat
    ▼ bigdata.project.pairs
      ▶ Pair.java
      ▶ PairsMain.java
      ▶ PairsMapper.java
      ▶ PairsPartitioner.java
      ▶ PairsReducer.java
        PairsRun.bat
    ▼ bigdata.project.stripes
      ▶ CrystalBallText.java
      ▶ StripesMain.java
      ▶ StripesMapper.java
      ▶ StripesReducer.java
      ▶ Utilities.java
        StripesRun.bat
  ▶ JRE System Library [JavaSE-1.8]
  ▶ hadoop library
  ▼ doc
      CrystalBal Project Setup Guidelines.pages
      CrystalBal Project Setup Guidelines.pdf
  ▼ lib
      hadoop-common-2.7.1.jar
      hadoop-mapreduce-client-core-2.7.1.jar
      log4j-1.2.17.jar
  ▼ testData
      Project Data Test Result.pdf
      user1.txt
```

- Package bigdata.project.pairs: is used for pairs approach
- Package bigdata.project.stripes: is used for stripes approach
- Package bigdata.project.hybrid: is used for hybrid approach
- Project setup guideline is located at doc folder
- Project libraries are located at lib folder
- Data test and test result are located at testData folder

Run project

In order to run the project, start the Cloudera and then start the terminal. For each approach, using the command lines below to run. These command lines are documented in PairsRun.bat for Pairs approach, StripesRun.bat for Stripes approach and HybridRun.bat for Hybrid approach

1. Pairs approach

```
$ hadoop fs -rm -r /user/cloudera/crystalball
$ hadoop fs -rm -r /user/cloudera/crystalball/input
$ hadoop fs -rm -r /user/cloudera/crystalball/output

$ sudo su hdfs
$ hadoop fs -mkdir /user/cloudera
$ hadoop fs -chown cloudera /user/cloudera
$ exit
$ sudo su cloudera
$ hadoop fs -mkdir /user/cloudera/crystalball /user/cloudera/crystalball/input

$ echo "34 56 29 12 34 56 92 10 34 12" > user1
$ echo "18 29 12 34 79 18 56 12 34 92" > user2
$ echo "26 56 28 39 46 16 56 28 9 46" > user3
$ hadoop fs -put user* /user/cloudera/crystalball/input

$ mkdir -p build
$ javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/* ./build/*.java -d build -
Xlint

$ jar -cvf pairs.jar -C build/ .

$ hadoop jar pairs.jar bigdata.project.pairs.PairsMain /user/cloudera/crystalball/
input /user/cloudera/crystalball/output

$ hadoop fs -cat /user/cloudera/crystalball/output/*
```

2. Stripes approach

```
 $ hadoop fs -rm -r /user/cloudera/crystalball
$ hadoop fs -rm -r /user/cloudera/crystalball/input
$ hadoop fs -rm -r /user/cloudera/crystalball/output

$ sudo su hdfs
$ hadoop fs -mkdir /user/cloudera
$ hadoop fs -chown cloudera /user/cloudera
$ exit
$ sudo su cloudera
$ hadoop fs -mkdir /user/cloudera/crystalball /user/cloudera/crystalball/input

$ echo "34 56 29 12 34 56 92 10 34 12" > user1
$ echo "18 29 12 34 79 18 56 12 34 92" > user2
$ echo "26 56 28 39 46 16 56 28 9 46" > user3
$ hadoop fs -put user* /user/cloudera/crystalball/input

$ mkdir -p build
$ javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/* ./build/*.java -d build -
Xlint

$ jar -cvf stripes.jar -C build/ .

$ hadoop jar stripes.jar bigdata.project.stripes.StripesMain /user/cloudera/
crystalball/input /user/cloudera/crystalball/output


$ hadoop fs -cat /user/cloudera/crystalball/output/*
```

3. Hybrid approach

```
   $ hadoop fs -rm -r /user/cloudera/crystalball
$ hadoop fs -rm -r /user/cloudera/crystalball/input
$ hadoop fs -rm -r /user/cloudera/crystalball/output

$ sudo su hdfs
$ hadoop fs -mkdir /user/cloudera
$ hadoop fs -chown cloudera /user/cloudera
$ exit
$ sudo su cloudera
$ hadoop fs -mkdir /user/cloudera/crystalball /user/cloudera/crystalball/input

$ echo "34 56 29 12 34 56 92 10 34 12" > user1
$ echo "18 29 12 34 79 18 56 12 34 92" > user2
$ echo "26 56 28 39 46 16 56 28 9 46" > user3
$ hadoop fs -put user* /user/cloudera/crystalball/input

$ mkdir -p build
$ javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/* ./build/*.java -d build -Xlint
$ jar -cvf hybrid.jar -C build/ .

$ hadoop jar hybrid.jar bigdata.project.hybrid.HybridMain /user/cloudera/crystalball/
input /user/cloudera/crystalball/output

$ hadoop fs -cat /user/cloudera/crystalball/output/*
```

Test report

The data test and test result is located at doc folder of crystal project. Please refer to it for the detail