Getting Started with R & Hadoop

From Local VM to the Cloud

TDWI World Boston

Pre-Conference Workshop

Hynes Convention Center
Boston, MA

Saturday, September 15, 2012

by **Jeffrey Breen**

email: jeffrey@jeffreybreen.com
http://jeffreybreen.wordpress.com
Twitter: @JeffreyBreen

http://bit.ly/tdwibos

Part 1: Setting up the Local VM

Code & more on github:

http://bit.ly/tdwibos
(https://github.com/jeffreybreen/tutorial-201209-TDWI-big-data)

Overview

- Download and install a virtual machine containing a configured and working version of Hadoop
- Install R, RStudio, and RHadoop packages
- Test our installation by running a simple Hadoop job written in R

Thank you, Cloudera

- Cloudera's Hadoop Demo VM provides everything you need to run small jobs in a virtual environment
- Hadoop 0.20 + Flume, HBase, Hive, Hue, Mahout, Oozie, Pig, Sqoop, Whirr, Zookeeper
- Based on CentOS 5.8 & available for VMware, KVM and VirtualBox:

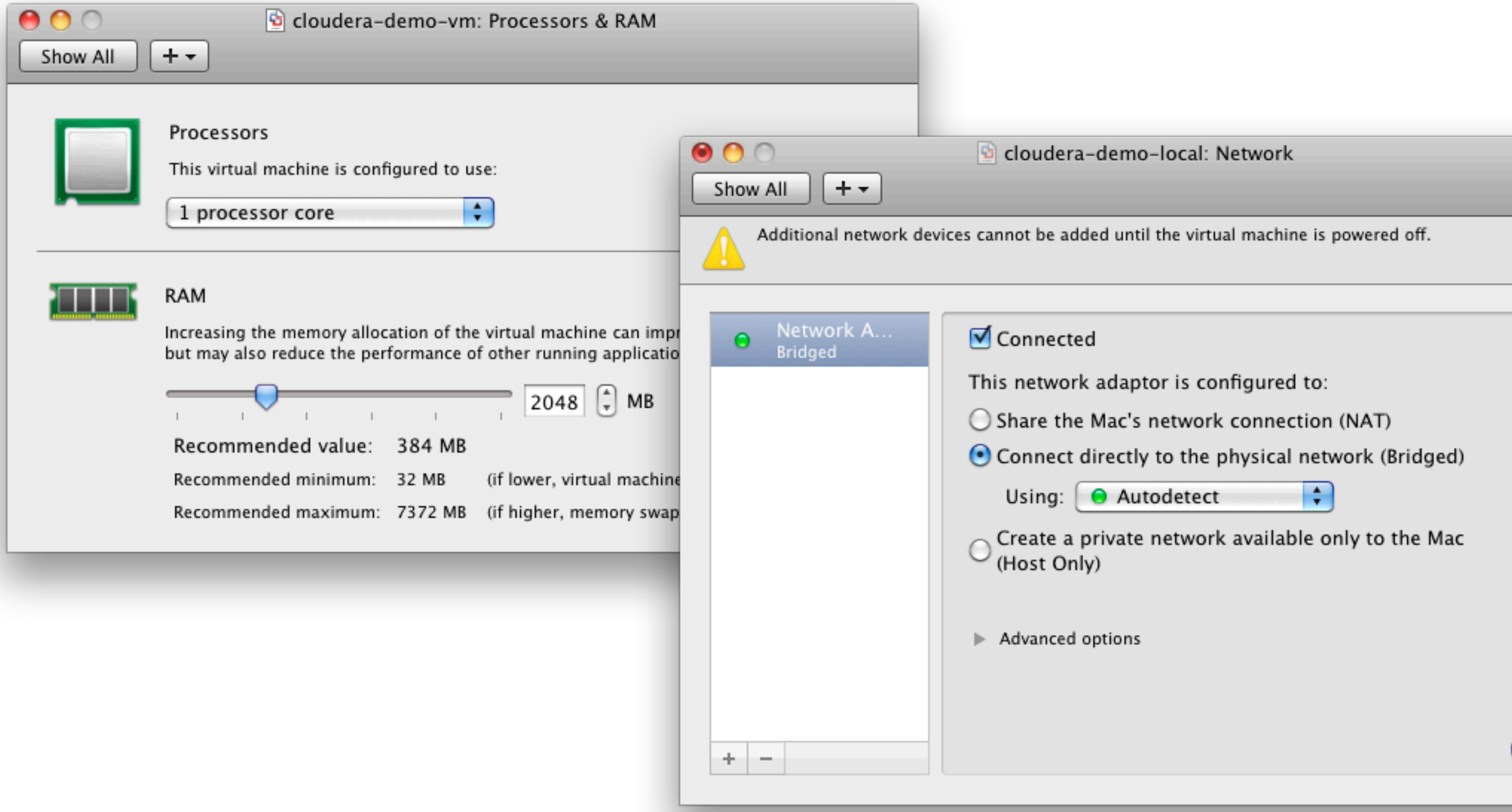
  https://ccp.cloudera.com/display/SUPPORT/Cloudera%27s+Hadoop+Demo+VM
- Provides a common base which we will use for our later cluster, etc. work
- Older version was known as the 'training VM' and came with tutorials and data still available on github:

  https://github.com/cloudera/cloudera-training

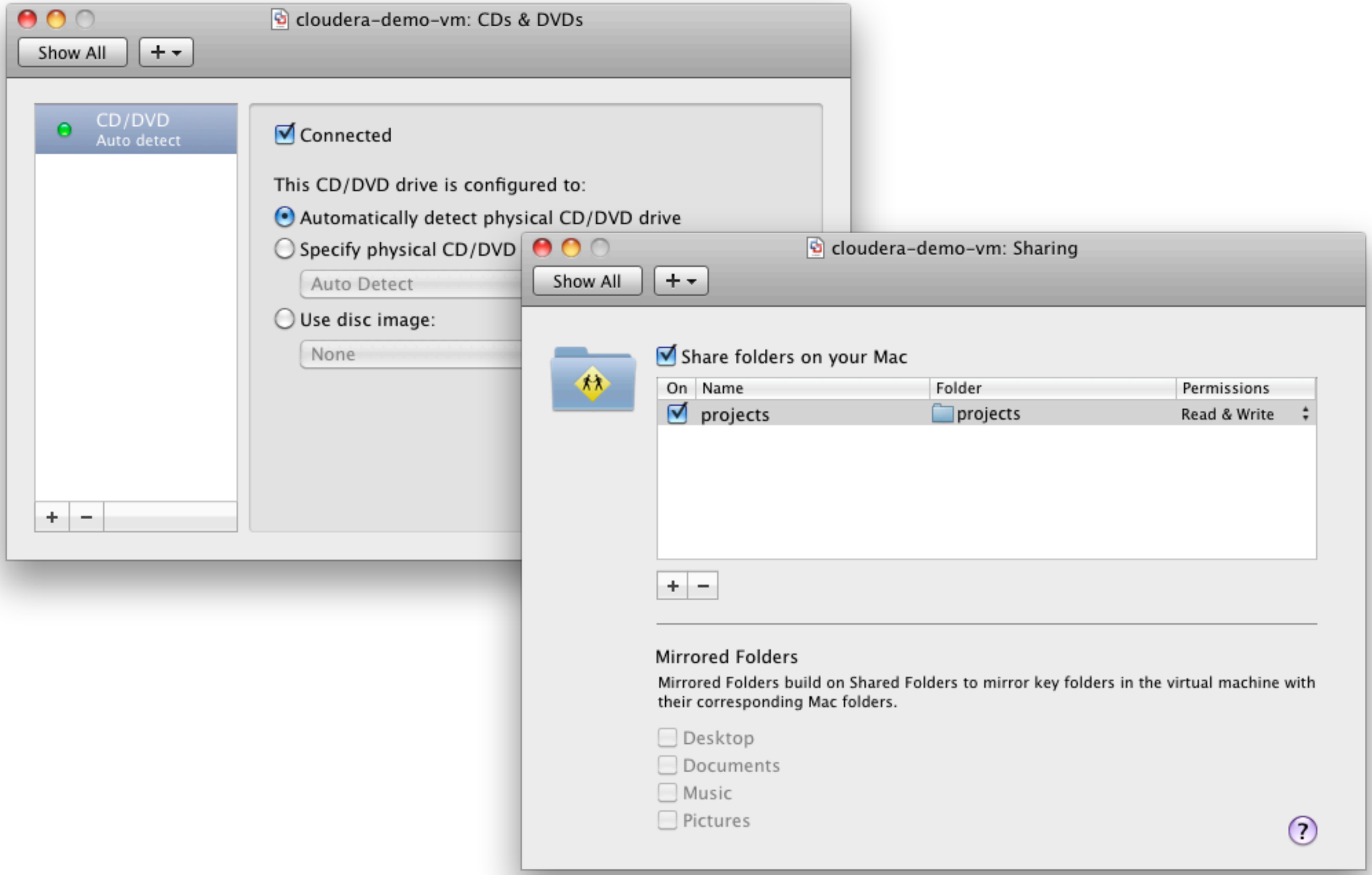
A couple of tweaks

- Give it more RAM
  - CDH3 uses 1GB by default, CDH4 starts with 3+GB
  - not configured with a swap file
- Use Bridged networking vs. NAT or Host-only
  - Virtual machine will get its own IP address on your network
  - Experienced DNS errors with whirr while sharing an IP
- Extras: Set up shared folders & add a CD-ROM
  - Shared folders make it easy to share data & code between your computer and the VM
  - Add a CD-ROM drive if you want to install VMware tools or any ISO file

Important


cloudera-demo-vm: Processors & RAM
Show All
Processors
This virtual machine is configured to use:
1 processor core
RAM
Increasing the memory allocation of the virtual machine can impr
but may also reduce the performance of other running applicatio
2048
MB
Recommended value: 384 MB
Recommended minimum: 32 MB (if lower, virtual machine
Recommended maximum: 7372 MB (if higher, memory swap
cloudera-demo-local: Network
Show All
Additional network devices cannot be added until the virtual machine is powered off.
Network A...
Bridged
Connected
This network adaptor is configured to:
Share the Mac's network connection (NAT)
Connect directly to the physical network (Bridged)
Using: Autodetect
Create a private network available only to the Mac (Host Only)
Advanced options

Nice to have


cloudera-demo-vm: CDs & DVDs
Show All
CD/DVD
Auto detect
Connected
This CD/DVD drive is configured to:
Automatically detect physical CD/DVD drive
Specify physical CD/DVD
Auto Detect
Use disc image:
None
cloudera-demo-vm: Sharing
Show All
Share folders on your Mac
On
Name
Folder
Permissions
projects
projects
Read & Write
Mirrored Folders
Mirrored Folders build on Shared Folders to mirror key folders in the virtual machine with their corresponding Mac folders.
Desktop
Documents
Music
Pictures

Yes, it's that easy

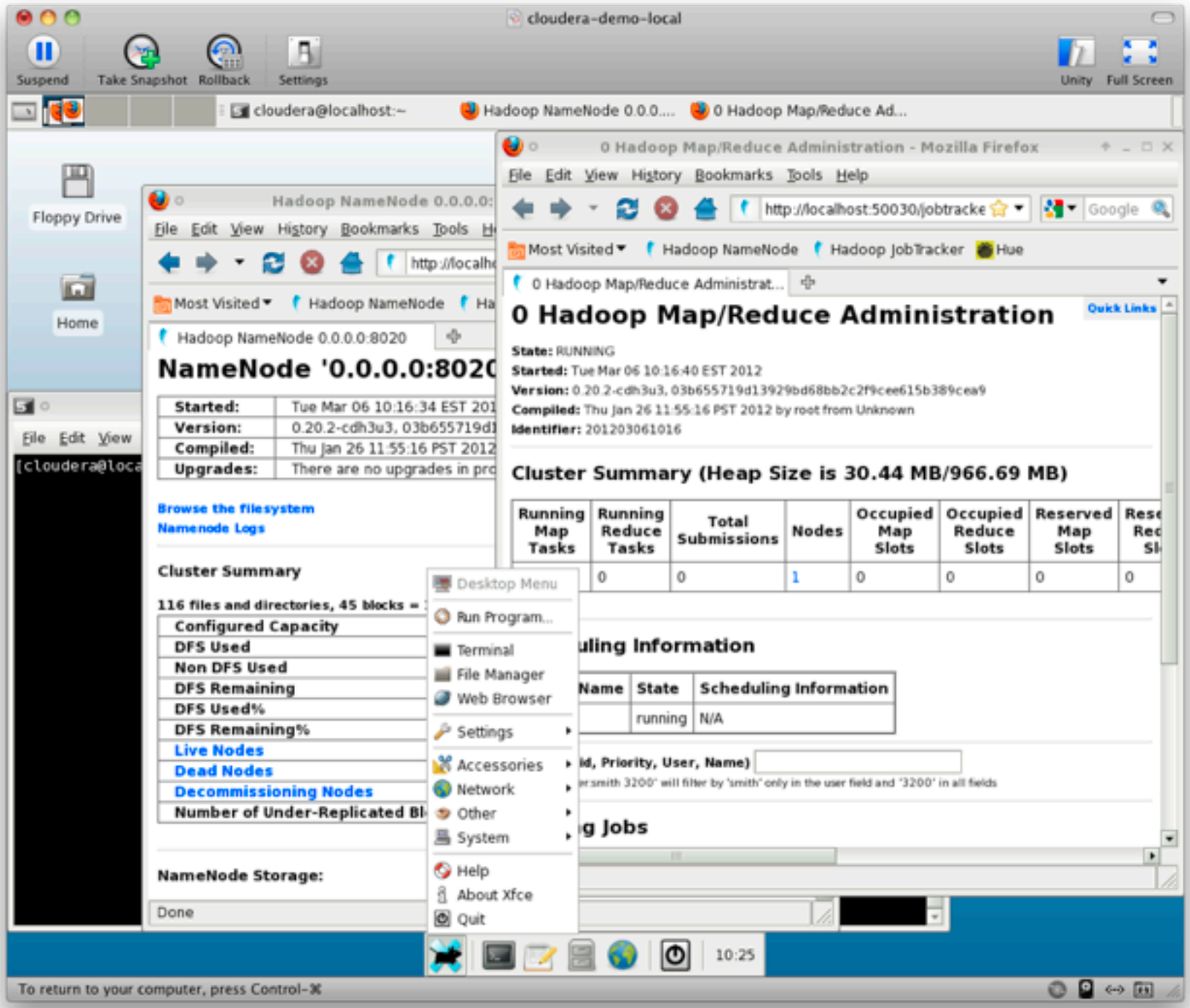
Boot VM and log in as "cloudera". (Password = "cloudera" too)

Execute as root with "`sudo`"

"`sudo su -`" for root shell

Hadoop already running

Firefox contains bookmarks to admin pages



Well, almost.

- Install VMware tools and link to shared folder on host PC

```
$ sudo mkdir /mnt/vmware
$ sudo mount /dev/hda /mnt/vmware
$ tar zxf /mnt/vmware/VMwareTools-8.4.7-683826.tar.gz
$ cd vmware-tools-distrib/
$ sudo ./vmware-install.pl
$ ln -s /mnt/hgfs/projects/tutorial-201209-TDWI-big-data/ ~/.
```

- First add the EPEL repository then install git, wget, and R

```
$ sudo rpm -Uvh http://dl.fedoraproject.org/pub/epel/5/x86_64/epel-release-5-4.noarch.rpm
$ sudo yum -y install git wget R
```

- Set Hadoop environment variables so R can find them too

CDH3:

```
$ sudo ln -s /etc/default/hadoop-0.20 /etc/profile.d/hadoop.sh
$ cat /etc/profile.d/hadoop.sh | sed 's/export //g' > ~/.Renviron
```

CDH4:

```
$ sudo ln -s /etc/default/hadoop-0.20-mapreduce /etc/profile.d/hadoop.sh
$ cat /etc/profile.d/hadoop.sh | sed 's/export //g' > ~/.Renviron
```

Warning: Pages of fast-scrolling gibberish to follow

But it’s all going to be OK

```
[cloudera@localhost ~]$ sudo mkdir /mnt/vmware
[cloudera@localhost ~]$ sudo mount /dev/hda /mnt/vmware
mount: block device /dev/hda is write-protected, mounting read-only
[cloudera@localhost ~]$ tar zxf /mnt/vmware/VMwareTools-8.4.7-416484.tar.gz
[cloudera@localhost ~]$ cd vmware-tools-distrib/
[cloudera@localhost vmware-tools-distrib]$ sudo ./vmware-install.pl
Creating a new VMware Tools installer database using the tar4 format.

Installing VMware Tools.

In which directory do you want to install the binary files?
[/usr/bin]

What is the directory that contains the init directories (rc0.d/ to rc6.d/)?
[/etc/rc.d]

What is the directory that contains the init scripts?
[/etc/rc.d/init.d]

In which directory do you want to install the daemon files?
[/usr/sbin]

In which directory do you want to install the library files?
[/usr/lib/vmware-tools]

The path "/usr/lib/vmware-tools" does not exist currently. This program is
going to create it, including needed parent directories. Is this what you want?
[yes]

In which directory do you want to install the documentation files?
[/usr/share/doc/vmware-tools]

The path "/usr/share/doc/vmware-tools" does not exist currently. This program
is going to create it, including needed parent directories. Is this what you
want? [yes]

The installation of VMware Tools 8.4.7 build-416484 for Linux completed
successfully. You can decide to remove this software from your system at any
time by invoking the following command: "/usr/bin/vmware-uninstall-tools.pl".

Before running VMware Tools for the first time, you need to configure it by
invoking the following command: "/usr/bin/vmware-config-tools.pl". Do you want
this program to invoke the command for you now? [yes]

Initializing...

Making sure services for VMware Tools are stopped.

Stopping VMware Tools services in the virtual machine:
   Guest operating system daemon:                          [  OK  ]
   Virtual Printing daemon:                                [  OK  ]
   Unmounting HGFS shares:                                 [  OK  ]
   Guest filesystem driver:                                [  OK  ]


Found a compatible pre-built module for vmmemctl.  Installing it...


Found a compatible pre-built module for vmhgfs.  Installing it...
```

```
[cloudera@localhost ~]$ sudo yum -y install wget git
Loaded plugins: fastestmirror
Loading mirror speeds from cached hostfile
 * base: mirror.symnds.com
 * epel: mirror.symnds.com
 * extras: mirrors.einstein.yu.edu
 * updates: mirror.symnds.com
epel                                                                                                                | 3.4 kB     00:00
epel/primary_db                                                                                                     | 3.7 MB     00:01
Setting up Install Process
Resolving Dependencies
There are unfinished transactions remaining. You might consider running yum-complete-transaction first to finish them.
The program yum-complete-transaction is found in the yum-utils package.
--> Running transaction check
---> Package git.x86_64 0:1.7.4.1-1.el5 set to be updated
--> Processing Dependency: perl-Git = 1.7.4.1-1.el5 for package: git
--> Processing Dependency: perl(Error) for package: git
--> Processing Dependency: perl(Git) for package: git
---> Package wget.x86_64 0:1.11.4-2.el5_4.1 set to be updated
--> Running transaction check
---> Package perl-Error.noarch 1:0.17010-1.el5 set to be updated
---> Package perl-Git.x86_64 0:1.7.4.1-1.el5 set to be updated
--> Finished Dependency Resolution

Dependencies Resolved

================================================================================================================================
 Package                         Arch                         Version                              Repository              Size
================================================================================================================================
Installing:
 git                             x86_64                       1.7.4.1-1.el5                        epel                   4.5 M
 wget                            x86_64                       1.11.4-2.el5_4.1                     base                   582 k
Installing for dependencies:
 perl-Error                      noarch                       1:0.17010-1.el5                      epel                    26 k
 perl-Git                        x86_64                       1.7.4.1-1.el5                        epel                    28 k

Transaction Summary
================================================================================================================================
Install       4 Package(s)
Upgrade       0 Package(s)

Total download size: 5.1 M
Downloading Packages:
(1/4): perl-Error-0.17010-1.el5.noarch.rpm                                                                          |  26 kB     00:00
(2/4): perl-Git-1.7.4.1-1.el5.x86_64.rpm                                                                            |  28 kB     00:00
(3/4): wget-1.11.4-2.el5_4.1.x86_64.rpm                                                                             | 582 kB     00:00
(4/4): git-1.7.4.1-1.el5.x86_64.rpm                                                                                 | 4.5 MB     00:01
--------------------------------------------------------------------------------------------------------------------------------
Total                                                                                       2.6 MB/s | 5.1 MB     00:02
warning: rpmts_HdrFromFdno: Header V3 DSA signature: NOKEY, key ID 217521f6
epel/gpgkey                                                                                                         | 1.7 kB     00:00
Importing GPG key 0x217521F6 "Fedora EPEL <epel@fedoraproject.org>" from /etc/pki/rpm-gpg/RPM-GPG-KEY-EPEL
Running rpm_check_debug
Running Transaction Test
Finished Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing     : wget                                                                                                1/4
  Installing     : perl-Error                                                                                          2/4
  Installing     : git                                                                                                 3/4
  Installing     : perl-Git                                                                                            4/4

Installed:
  git.x86_64 0:1.7.4.1-1.el5                                    wget.x86_64 0:1.11.4-2.el5_4.1

Dependency Installed:
  perl-Error.noarch 1:0.17010-1.el5                                perl-Git.x86_64 0:1.7.4.1-1.el5

Complete!
```

```
[cloudera@localhost ~]$ sudo rpm -Uvh http://dl.fedoraproject.org/pub/epel/5/x86_64/epel-release-5-4.noarch.rpm
Retrieving http://dl.fedoraproject.org/pub/epel/5/x86_64/epel-release-5-4.noarch.rpm
warning: /var/tmp/rpm-xfer.CPJMIi: Header V3 DSA signature: NOKEY, key ID 217521f6
Preparing...                ########################################### [100%]
   1:epel-release           ########################################### [100%]
[cloudera@localhost ~]$ sudo yum -y install R
Loaded plugins: fastestmirror
Loading mirror speeds from cached hostfile
 * base: mirror.symnds.com
 * epel: mirrors.einstein.yu.edu
 * extras: mirrors.einstein.yu.edu
 * updates: mirror.symnds.com
Setting up Install Process
Resolving Dependencies
There are unfinished transactions remaining. You might consider running yum-complete-transaction first to finish them.
The program yum-complete-transaction is found in the yum-utils package.
--> Running transaction check
---> Package R.x86_64 0:2.14.1-1.el5 set to be updated
--> Processing Dependency: libRmath-devel = 2.14.1-1.el5 for package: R
--> Processing Dependency: R-devel = 2.14.1-1.el5 for package: R
--> Running transaction check
---> Package R-devel.x86_64 0:2.14.1-1.el5 set to be updated
--> Processing Dependency: R-core = 2.14.1-1.el5 for package: R-devel
--> Processing Dependency: zlib-devel for package: R-devel
--> Processing Dependency: tk-devel for package: R-devel
--> Processing Dependency: texinfo-tex for package: R-devel
--> Processing Dependency: tetex-latex for package: R-devel
--> Processing Dependency: tcl-devel for package: R-devel
--> Processing Dependency: pcre-devel for package: R-devel
--> Processing Dependency: libX11-devel for package: R-devel
--> Processing Dependency: gcc-gfortran for package: R-devel
--> Processing Dependency: gcc-c++ for package: R-devel
--> Processing Dependency: bzip2-devel for package: R-devel
---> Package libRmath-devel.x86_64 0:2.14.1-1.el5 set to be updated
--> Processing Dependency: libRmath = 2.14.1-1.el5 for package: libRmath-devel
--> Running transaction check
---> Package R-core.x86_64 0:2.14.1-1.el5 set to be updated
--> Processing Dependency: xdg-utils for package: R-core
--> Processing Dependency: cups for package: R-core
--> Processing Dependency: libgfortran.so.1()(64bit) for package: R-core
---> Package bzip2-devel.x86_64 0:1.0.3-6.el5_5 set to be updated
---> Package gcc-c++.x86_64 0:4.1.2-51.el5 set to be updated
--> Processing Dependency: gcc = 4.1.2-51.el5 for package: gcc-c++
--> Processing Dependency: libstdc++-devel = 4.1.2-51.el5 for package: gcc-c++
---> Package gcc-gfortran.x86_64 0:4.1.2-51.el5 set to be updated
--> Processing Dependency: libgmp.so.3()(64bit) for package: gcc-gfortran
---> Package libRmath.x86_64 0:2.14.1-1.el5 set to be updated
---> Package libX11-devel.x86_64 0:1.0.3-11.el5_7.1 set to be updated
--> Processing Dependency: xorg-x11-proto-devel >= 7.1-2 for package: libX11-devel
--> Processing Dependency: libXau-devel for package: libX11-devel
--> Processing Dependency: libXdmcp-devel for package: libX11-devel
---> Package pcre-devel.x86_64 0:6.6-6.el5_6.1 set to be updated
---> Package tcl-devel.x86_64 0:8.4.13-4.el5 set to be updated
---> Package tetex-latex.x86_64 0:3.0-33.13.el5 set to be updated
--> Processing Dependency: tetex-dvips = 3.0 for package: tetex-latex
--> Processing Dependency: tetex = 3.0 for package: tetex-latex
--> Processing Dependency: netpbm-progs for package: tetex-latex
---> Package texinfo-tex.x86_64 0:4.8-14.el5 set to be updated
--> Processing Dependency: texinfo = 4.8-14.el5 for package: texinfo-tex
---> Package tk-devel.x86_64 0:8.4.13-5.el5_1.1 set to be updated
---> Package zlib-devel.x86_64 0:1.2.3-4.el5 set to be updated
--> Running transaction check
```

Pretty impressive for cut-and-pasting a few commands, eh?

Checking on Hadoop

- Get VM's IP address with 'ifconfig'

  ```
  $ ifconfig
  ```

- Connect to web admin interface
  - Job tracker on port 50030

    http://192.168.1.132:50030
  - Name Node (HDFS) on 50070

    http://192.168.1.132:50070/
  - Full list

http://www.cloudera.com/blog/2009/08/hadoop-default-ports-quick-reference/

0 Hadoop Map/Reduce Administration

http://192.168.1.132:50030/jobtracker.jsp

Google

0 Hadoop Map/Reduce Administration

Quick Links

State: RUNNING
Started: Wed Sep 12 15:45:48 EDT 2012
Version: 2.0.0-mr1-cdh4.0.0, Unknown
Compiled: Mon Jun 4 17:31:58 PDT 2012 by jenkins from Unknown
Identifier: 201209121545

Cluster Summary (Heap Size is 30.44 MB/

Running Map Tasks	Running Reduce Tasks	Total Submissions	Nodes	Occupied Map Slots	Oc R
0	0	0	1	0	0

Scheduling Information

Queue Name	State	Scheduling Information
default	running	N/A

Filter (Jobid, Priority, User, Name)

Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in

Running Jobs

none

Retired Jobs

none

Hadoop NameNode 0.0.0.0:8020

http://192.168.1.132:50070/dfshealth.jsp

Google

NameNode '0.0.0.0:8020' (active)

Started:	Wed Sep 12 15:44:14 EDT 2012
Version:	2.0.0-cdh4.0.0, 5d678f6bb1f2bc49e2287dd69ac41d7232fc9cdc
Compiled:	Mon Jun 4 16:52:42 PDT 2012 by jenkins from Unknown
Upgrades:	There are no upgrades in progress.
Cluster ID:	CID-7a1e2271-73b7-4fec-ba84-45c53fb98bff
Block Pool ID:	BP-1361448423-127.0.0.1-1340247796653

Browse the filesystem
NameNode Logs

Cluster Summary

Security is *OFF*
24 files and directories, 1 blocks = 25 total.
Heap Memory used 20.07 MB is 61% of Commited Heap Memory 32.88 MB. Max Heap Memory is 966.69 MB.
Non Heap Memory used 33.48 MB is 95% of Commited Non Heap Memory 35.06 MB. Max Non Heap Memory is 130 MB.

Configured Capacity	: 9.39 GB
DFS Used	: 32 KB
Non DFS Used	: 4.53 GB
DFS Remaining	: 4.86 GB
DFS Used%	: 0 %
DFS Remaining%	: 51.77 %

Let's Install RStudio

- Current download link and instructions at **http://rstudio.org/download/server**

```
$ wget http://download2.rstudio.org/rstudio-server-0.96.331-x86_64.rpm
```

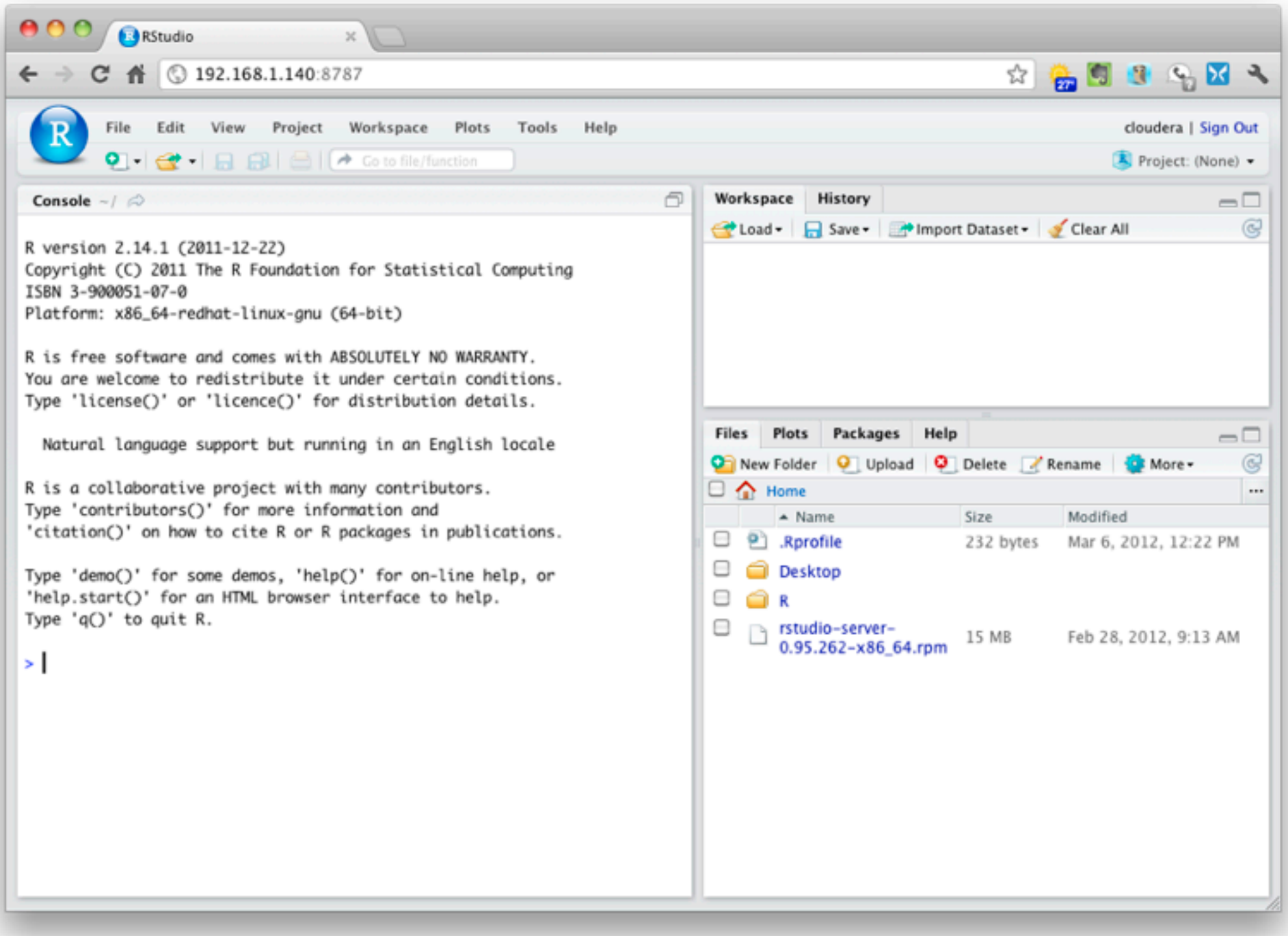
- Install from RPM

```
$ sudo rpm -Uvh rstudio-server-0.96.331-x86_64.rpm
```

- Access from browser via port 8787
  - e.g., **http://192.168.1.132:8787/**

```
[cloudera@localhost ~]$ wget http://download2.rstudio.org/rstudio-server-0.95.262-x86_64.rpm
--2012-03-06 12:14:24--  http://download2.rstudio.org/rstudio-server-0.95.262-x86_64.rpm
Resolving download2.rstudio.org... 216.137.39.181, 216.137.39.217, 216.137.39.222, ...
Connecting to download2.rstudio.org|216.137.39.181|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 15748959 (15M) [application/x-redhat-package-manager]
Saving to: `rstudio-server-0.95.262-x86_64.rpm'

100%[======================================================================================================>] 15,748,959  1.83M/s   in 7.2s

2012-03-06 12:14:31 (2.09 MB/s) - `rstudio-server-0.95.262-x86_64.rpm' saved [15748959/15748959]

[cloudera@localhost ~]$ sudo rpm -Uvh rstudio-server-0.95.262-x86_64.rpm
Preparing...                ########################################### [100%]
   1:rstudio-server         ########################################### [100%]
rsession: no process killed
Starting rstudio-server: [  OK  ]
[cloudera@localhost ~]$ ifconfig
eth0      Link encap:Ethernet  HWaddr 00:0C:29:4B:77:1D
          inet addr:192.168.1.140  Bcast:192.168.1.255  Mask:255.255.255.0
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:75039 errors:0 dropped:0 overruns:0 frame:0
          TX packets:36742 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:104953280 (100.0 MiB)  TX bytes:3061577 (2.9 MiB)
          Interrupt:59 Base address:0x2000

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          UP LOOPBACK RUNNING  MTU:16436  Metric:1
          RX packets:78954 errors:0 dropped:0 overruns:0 frame:0
          TX packets:78954 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:14608044 (13.9 MiB)  TX bytes:14608044 (13.9 MiB)
```

RStudio Success



Install RHadoop's **rmr** package

- RHadoop is an open source project sponsored by Revolution Analytics and is one of several available to make it easier to work with R and Hadoop
  - The **rmr** package contains all the mapreduce-related functions, including generating Hadoop streaming jobs and basic data exchange with HDFS
- First install prerequisite packages (run R as root to install system-wide)

```
$ sudo R
> install.packages( c('RJSONIO', 'itertools', 'digest', 'Rcpp'),
      repos='http://cran.revolutionanalytics.com')
```

- Download the latest stable release (1.3.1) from github

```
$ wget --no-check-certificate https://github.com/downloads/RevolutionAnalytics/RHadoop/rmr_1.3.1.tar.gz
```

- Install the package from the tar file

```
$ sudo R CMD INSTALL rmr_1.3.1.tar.gz
```

- Test that it loads

```
$ R
> library(rmr)
Loading required package: Rcpp
Loading required package: RJSONIO
Loading required package: itertools
Loading required package: iterators
Loading required package: digest
```

Let's test it out

Let's harness the power of Hadoop... to square a few numbers

```
library(rmr)

small.ints = 1:1000

small.int.path = to.dfs(1:1000)

out = mapreduce(input = small.int.path,
      map = function(k,v) keyval(v, v^2)
      )

df = as.data.frame(
        from.dfs( out, structured=T ) )
```

see R/0-square-integers.R
and https://github.com/RevolutionAnalytics/RHadoop/wiki/Tutorial

RStudio

192.168.1.130:8787

File Edit Code View Project Workspace Plots Tools Help

cloudera | Sign Out

Go to file/function

Project: (None)

test-square-integers.R

Source on Save

Run

Source

```
1  library(rmr)
2
3  small.ints = 1:1000
4  small.int.path = to.dfs(1:1000)
5  out = mapreduce(input = small.int.path,
6                  map = function(k,v) keyval(v, v^2) )
7  df = as.data.frame( from.dfs( out, structured=TRUE) )
8  colnames(df) = c('n', 'n2')
9
```

9:1 (Top Level) R Script

Workspace History

Load Save Import Dataset Clear A

Data	
df	1000 obs. of 2 variables
Values	
small.ints	integer[1000]
Functions	
out()	
small.int.path()	

Console ~/

```
> out = mapreduce(input = small.int.path,
+                 map = function(k,v) keyval(v, v^2) )
packageJobJar: [/tmp/RtmpzGA7L2/rmr-local-env, /tmp/RtmpzGA7L2/rmr-global-env,
/tmp/RtmpzGA7L2/rhstr.map28c25255bc95, /var/lib/hadoop-0.20/cache/cloudera/hadoop-
unjar8485390741174873780/] [] /tmp/streamjob4769379180629343125.jar tmpDir=null
12/09/12 23:17:48 INFO mapred.FileInputFormat: Total input paths to process : 1
12/09/12 23:17:49 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-
0.20/cache/cloudera/mapred/local]
12/09/12 23:17:49 INFO streaming.StreamJob: Running job: job_201209122232_0003
12/09/12 23:17:49 INFO streaming.StreamJob: To kill this job, run:
12/09/12 23:17:49 INFO streaming.StreamJob: /usr/lib/hadoop-0.20/bin/hadoop job  -
Dmapred.job.tracker=0.0.0.0:8021 -kill job_201209122232_0003
12/09/12 23:17:49 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdetails.jsp?
jobid=job_201209122232_0003
12/09/12 23:17:50 INFO streaming.StreamJob:  map 0%  reduce 0%
12/09/12 23:18:03 INFO streaming.StreamJob:  map 50%  reduce 0%
12/09/12 23:18:04 INFO streaming.StreamJob:  map 100%  reduce 0%
12/09/12 23:18:14 INFO streaming.StreamJob:  map 100%  reduce 33%
12/09/12 23:18:16 INFO streaming.StreamJob:  map 100%  reduce 100%
12/09/12 23:18:19 INFO streaming.StreamJob: Job complete: job_201209122232_0003
12/09/12 23:18:19 INFO streaming.StreamJob: Output: /tmp/RtmpzGA7L2/file28c242f7a00
> df = as.data.frame( from.dfs( out, structured=TRUE) )
> colnames(df) = c('n', 'n2')
> head(df)
  n n2
1 1  1
2 2  4
3 3  9
4 4 16
```

Files Plots Packages Help

R: R and Hadoop Streaming Connector

Find in Topic

R and Hadoop Streaming Connector

Documentation for package ‘rmr’ version 1.3.1

- DESCRIPTION file.

Help Pages

dfs.empty	Check if a dfs file is empty
equijoin	Equijoins using map reduce
from.dfs	Read or write R objects from or to the file system

Next up:

Getting Started with R & Hadoop