Big Data Infrastructure and Technologies for Data Analytics (BDIT4DA)

Practice Guidelines

Hadoop cluster Installation:
Cloudera Quickstart on VirtualBox

Outline

- Hadoop cluster VMs
  - Cloudera Qyuickstart 5.13 – 5.3 GB (discontinued but version 5.13 available)
  - Hortonworks Sandbox – 10 GB
  - Oracle Big Data Lite 4.11 – 22 GB
  - Bitnami Hadoop Stack – 1.5 GB (only CLI and MapReduce)

- Oracle VirtualBox and pre-requisite configuration
- Installing Cloudera Quickstart Hadoop cluster on VirtualBox

Preparation

- Download Cloudera Quickstart 5.13 from below location (not available at Cloudera website since 2020)
  - https://surfdrive.surf.nl/files/index.php/s/2OvUuw5chV42Zwz
  - Directory “vm-software”
- Configure VirtualBox Host-Only Network Adapter

Cloudera Quickstart VM for VirtualBox

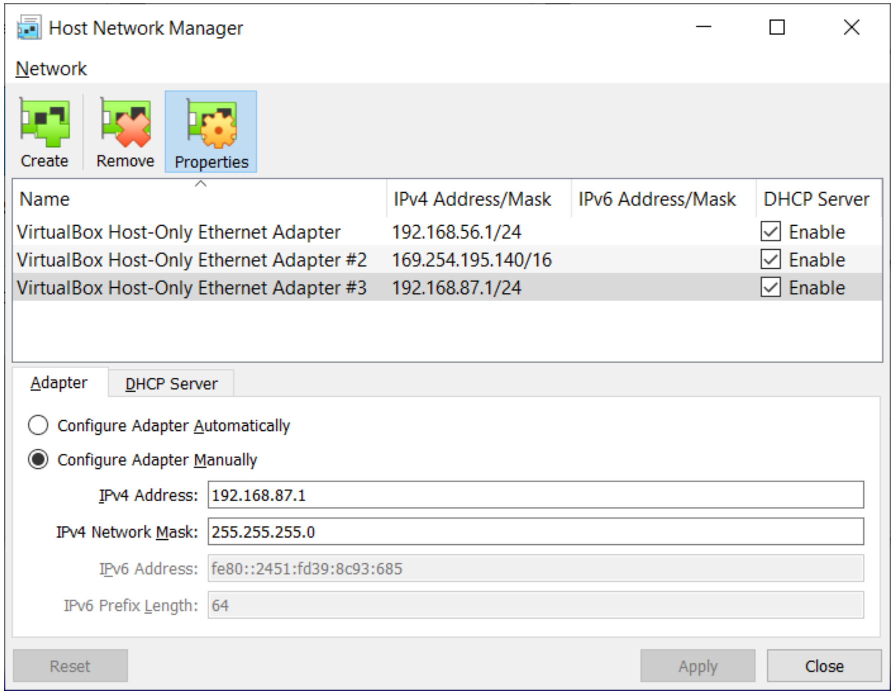
- QuickStart virtual machines (VMs) provide sufficient functionality to try CDH, Cloudera Manager, Impala, and Cloudera Search
  - Minimum required RAM on host machine is 4GB, recommended 8GB
- Cloudera Manager is installed in the VM but is turned off by default.
  - Cloudera strongly recommends that if you use Cloudera Manager, you configure the VM with a minimum of 8 GB RAM and two virtual CPU cores.

VirtualBox Configuration – Host requirements

- Pre-requisite and requirement to host
- Hosting Cloudera Quickstart 5.13 VM require
  - Single node minimum RAM 4GB, recommended 8GB
  - Single node + Cloudera Manager minimum 8GB
- Host machine HardDrive space:
  - Quickstart 5.13 VM for VirtualBox – 5.3 GB
  - VM image installed on the host machine – 9-12 GB
  - Plus additionally for a snapshot 1.2-2.5 GB
- VirtualBox Installation
  - Note: For Windows, you can configure VirtualBox VMs location on drive different from C: (which may be have limited space)

Accessing your VirtualBox Guest from your Host OS: Adding Host-Only network interface

- You need to do is to setup a VirtualBox Host-Only Network and get our guest connected to it.
  - If your guest machine is running, shut it down first.
- New VirtualBox versions
  - Click on File->Host Network Manager in the VirtualBox menu-bar.
  - Use Create/Properties Network Adapter
- Older VirtualBox versions
  - Click on File->Preferences
  - Select the Network option from the side menu and add network adapter
- Configure Adapter Automatically or Manually for better control
  - The default options for the newly-created Host-only network should be fine.
  - If not, you can add the data manually, by clicking on the Edit button in the DHCP Server tab.
  - Save all the settings in Preferences.
- Now open up the settings of your Guest machine and navigate to the Network option from the side menu and click on the Adapter 2 tab.
  - Don't forget to check the Enable Network Adapter option.
  - Save these settings and boot into your Guest machine.
  - After logging in, type ifconfig. Note the new IP, it should be under a new interface like eth0 or vboxnet0. Now you can use this IP to SSH, view the webpages on your machine's Apache Server, etc..



Name	IPv4 Address/Mask	IPv6 Address/Mask	DHCP Server
VirtualBox Host-Only Ethernet Adapter	192.168.56.1/24		☑ Enable
VirtualBox Host-Only Ethernet Adapter #2	169.254.195.140/16		☑ Enable
VirtualBox Host-Only Ethernet Adapter #3	192.168.87.1/24		☑ Enable

○ Configure Adapter Automatically

● Configure Adapter Manually

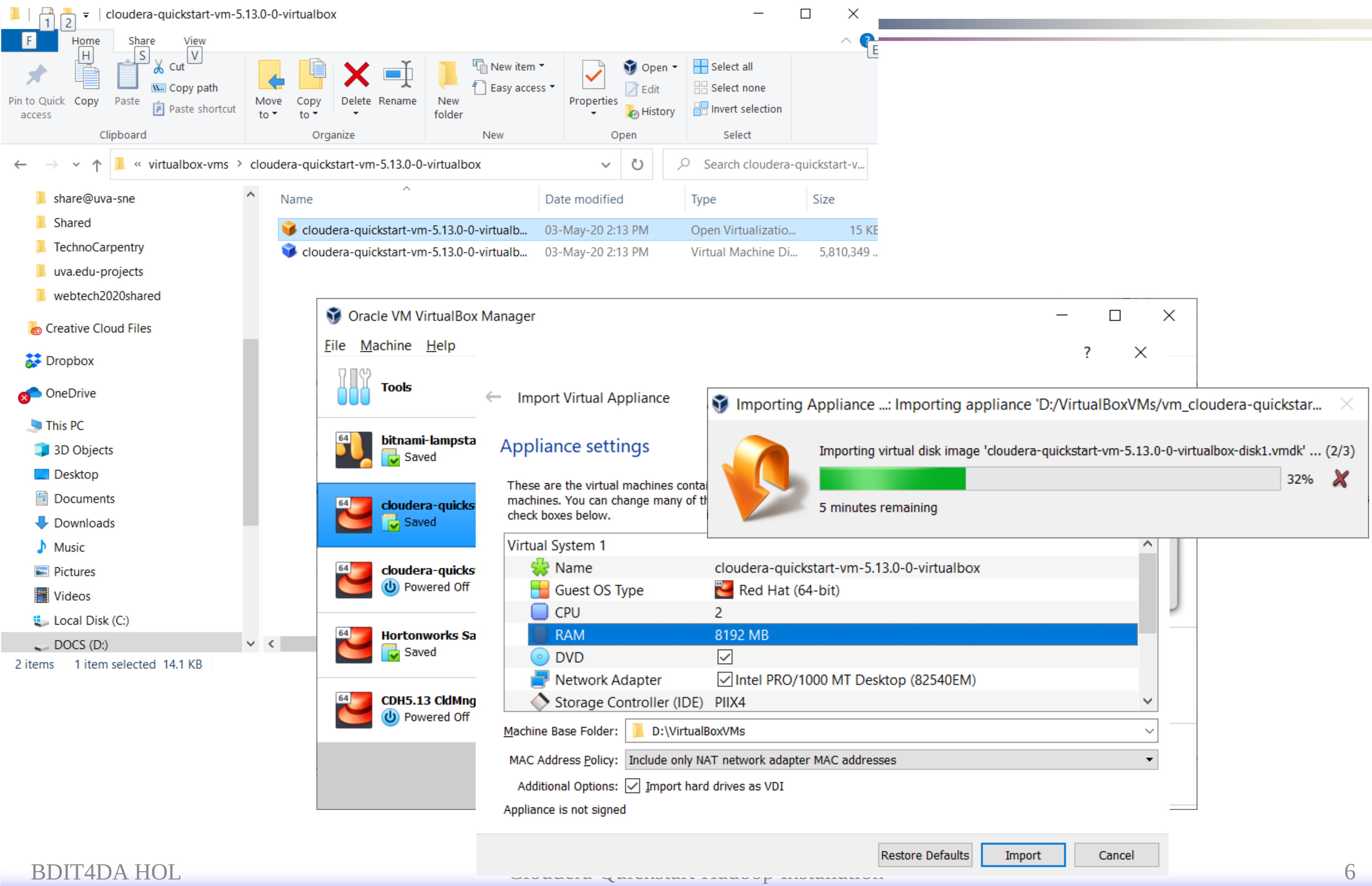
IPv4 Address: 192.168.87.1

IPv4 Network Mask: 255.255.255.0

IPv6 Address: fe80::2451:fd39:8c93:685

IPv6 Prefix Length: 64

Importing VM: Select+Click on .ova File



cloudera-quickstart-vm-5.13.0-0-virtualbox

virtualbox-vms › cloudera-quickstart-vm-5.13.0-0-virtualbox

Name	Date modified	Type	Size
cloudera-quickstart-vm-5.13.0-0-virtualb...	03-May-20 2:13 PM	Open Virtualizatio...	15 KE
cloudera-quickstart-vm-5.13.0-0-virtualb...	03-May-20 2:13 PM	Virtual Machine Di...	5,810,349 ..

2 items 1 item selected 14.1 KB

Oracle VM VirtualBox Manager

bitnami-lampsta	Saved
cloudera-quicks	Saved
cloudera-quicks	Powered Off
Hortonworks Sa	Saved
CDH5.13 CldMng	Powered Off

Import Virtual Appliance

Appliance settings

These are the virtual machines contai machines. You can change many of th check boxes below.

Virtual System 1

Name	cloudera-quickstart-vm-5.13.0-0-virtualbox
Guest OS Type	Red Hat (64-bit)
CPU	2
RAM	8192 MB
DVD	☑
Network Adapter	☑ Intel PRO/1000 MT Desktop (82540EM)
Storage Controller (IDE)	PIIX4

Machine Base Folder: D:\VirtualBoxVMs

MAC Address Policy: Include only NAT network adapter MAC addresses

Additional Options: ☑ Import hard drives as VDI

Appliance is not signed

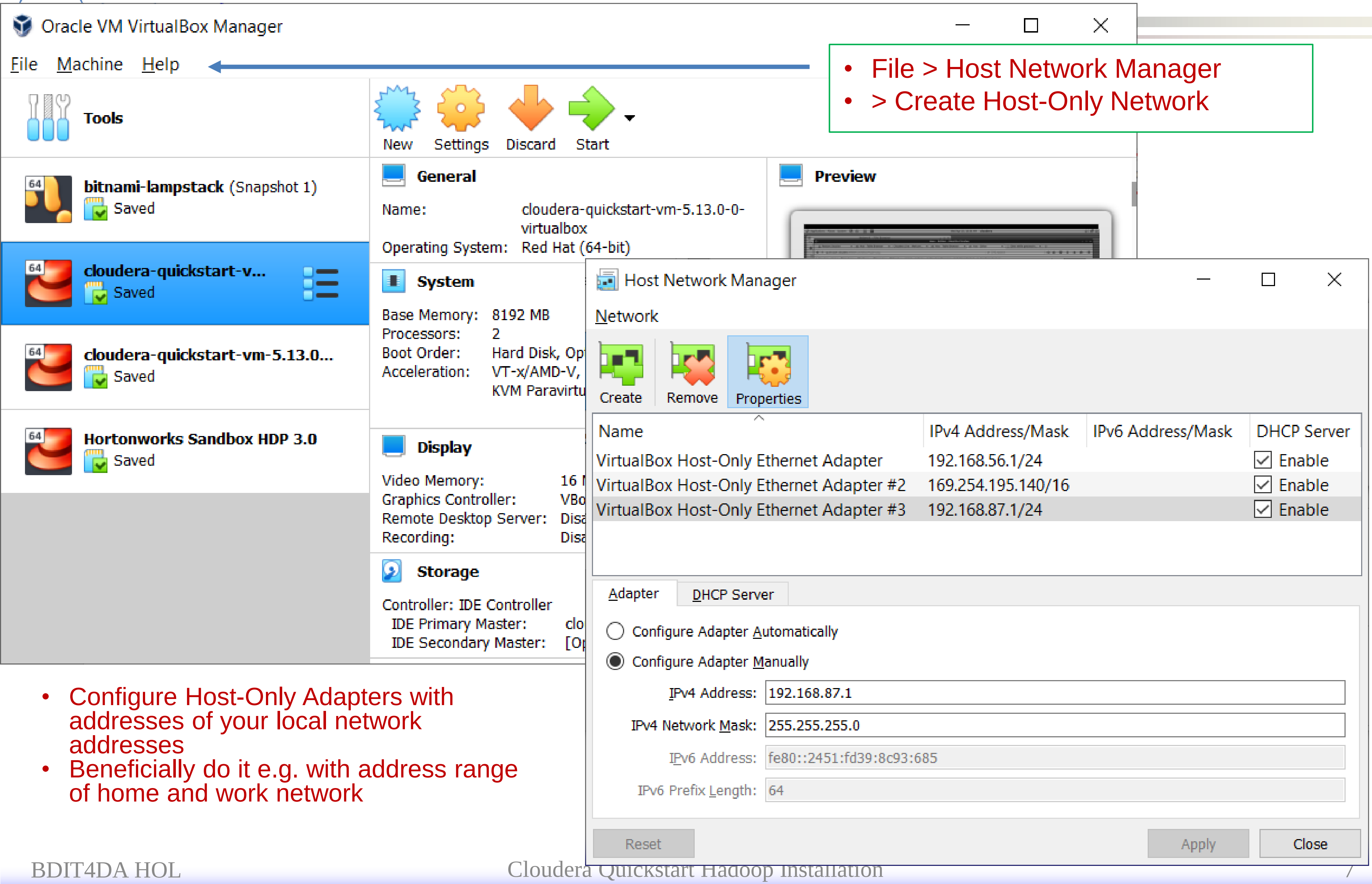
Importing Appliance ...: Importing appliance 'D:/VirtualBoxVMs/vm_cloudera-quickstar...

Importing virtual disk image 'cloudera-quickstart-vm-5.13.0-0-virtualbox-disk1.vmdk' ... (2/3)

32%

5 minutes remaining

VirtualBox Host-Only Network Configuration



Name: cloudera-quickstart-vm-5.13.0-0-virtualbox
Operating System: Red Hat (64-bit)

Base Memory: 8192 MB
Processors: 2
Boot Order: Hard Disk, Op
Acceleration: VT-x/AMD-V, KVM Paravirtu

Video Memory: 16
Graphics Controller: VBo
Remote Desktop Server: Disa
Recording: Disa

Controller: IDE Controller
IDE Primary Master: clo
IDE Secondary Master: [Op

Name	IPv4 Address/Mask	IPv6 Address/Mask	DHCP Server
VirtualBox Host-Only Ethernet Adapter	192.168.56.1/24		☑ Enable
VirtualBox Host-Only Ethernet Adapter #2	169.254.195.140/16		☑ Enable
VirtualBox Host-Only Ethernet Adapter #3	192.168.87.1/24		☑ Enable

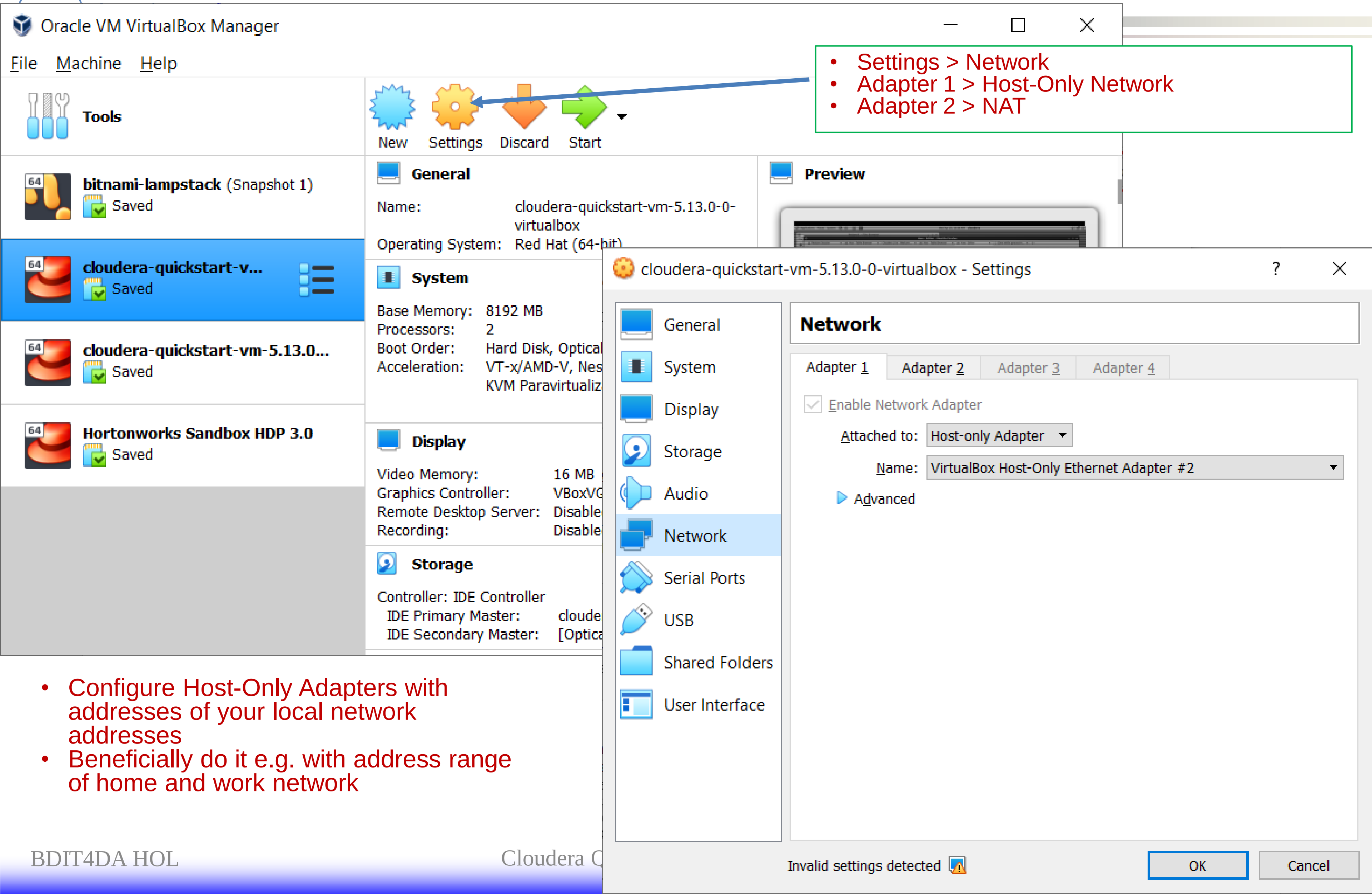
○ Configure Adapter Automatically
● Configure Adapter Manually

IPv4 Address: 192.168.87.1
IPv4 Network Mask: 255.255.255.0
IPv6 Address: fe80::2451:fd39:8c93:685
IPv6 Prefix Length: 64

- File > Host Network Manager
- > Create Host-Only Network

- Configure Host-Only Adapters with addresses of your local network addresses
- Beneficially do it e.g. with address range of home and work network

Guest VM Network Configuration



- Settings > Network
- Adapter 1 > Host-Only Network
- Adapter 2 > NAT

Oracle VM VirtualBox Manager

bitnami-lampstack (Snapshot 1)
Saved

cloudera-quickstart-v...
Saved

cloudera-quickstart-vm-5.13.0...
Saved

Hortonworks Sandbox HDP 3.0
Saved

General

Name: cloudera-quickstart-vm-5.13.0-0-virtualbox
Operating System: Red Hat (64-bit)

System

Base Memory: 8192 MB
Processors: 2
Boot Order: Hard Disk, Optical
Acceleration: VT-x/AMD-V, Nes
KVM Paravirtualiz

Display

Video Memory: 16 MB
Graphics Controller: VBoxVG
Remote Desktop Server: Disable
Recording: Disable

Storage

Controller: IDE Controller
IDE Primary Master: cloude
IDE Secondary Master: [Optica

Preview

cloudera-quickstart-vm-5.13.0-0-virtualbox - Settings

Network

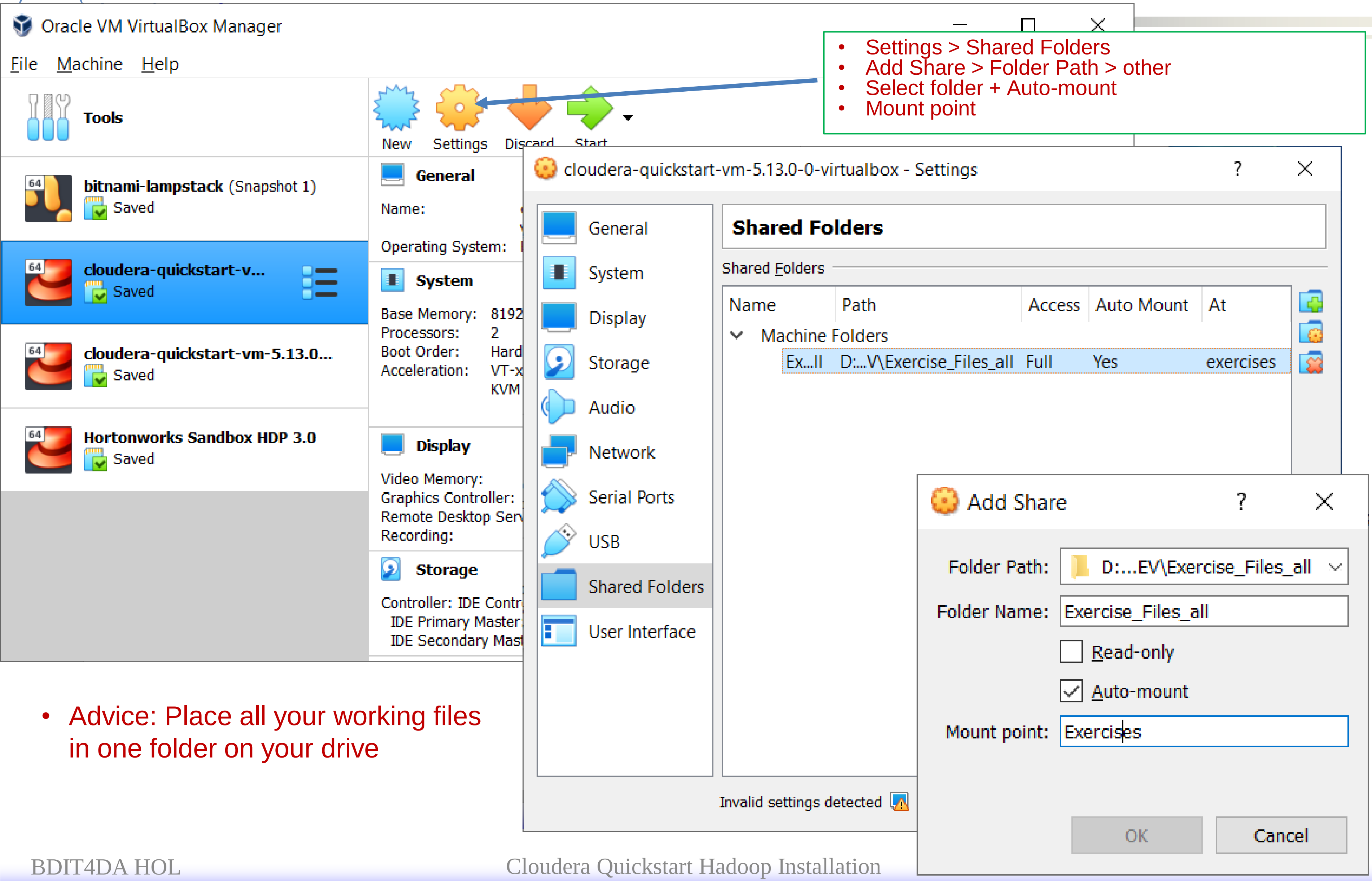
☑ Enable Network Adapter

Attached to: Host-only Adapter
Name: VirtualBox Host-Only Ethernet Adapter #2

Invalid settings detected

- Configure Host-Only Adapters with addresses of your local network addresses
- Beneficially do it e.g. with address range of home and work network

Add Shared Folder: e.g. Exercise_Files_all



- Settings > Shared Folders
- Add Share > Folder Path > other
- Select folder + Auto-mount
- Mount point

Oracle VM VirtualBox Manager

bitnami-lampstack (Snapshot 1)
Saved

cloudera-quickstart-v...
Saved

cloudera-quickstart-vm-5.13.0...
Saved

Hortonworks Sandbox HDP 3.0
Saved

General

Name:

Operating System:

System

Base Memory: 8192
Processors: 2
Boot Order: Hard
Acceleration: VT-x
KVM

Display

Video Memory:
Graphics Controller:
Remote Desktop Serv
Recording:

Storage

Controller: IDE Contr
IDE Primary Master
IDE Secondary Mast

cloudera-quickstart-vm-5.13.0-0-virtualbox - Settings

Shared Folders

Shared Folders

Name	Path	Access	Auto Mount	At
Machine Folders				
Ex...ll	D:...V\Exercise_Files_all	Full	Yes	exercises

Invalid settings detected

Add Share

Folder Path: D:...EV\Exercise_Files_all

Folder Name: Exercise_Files_all

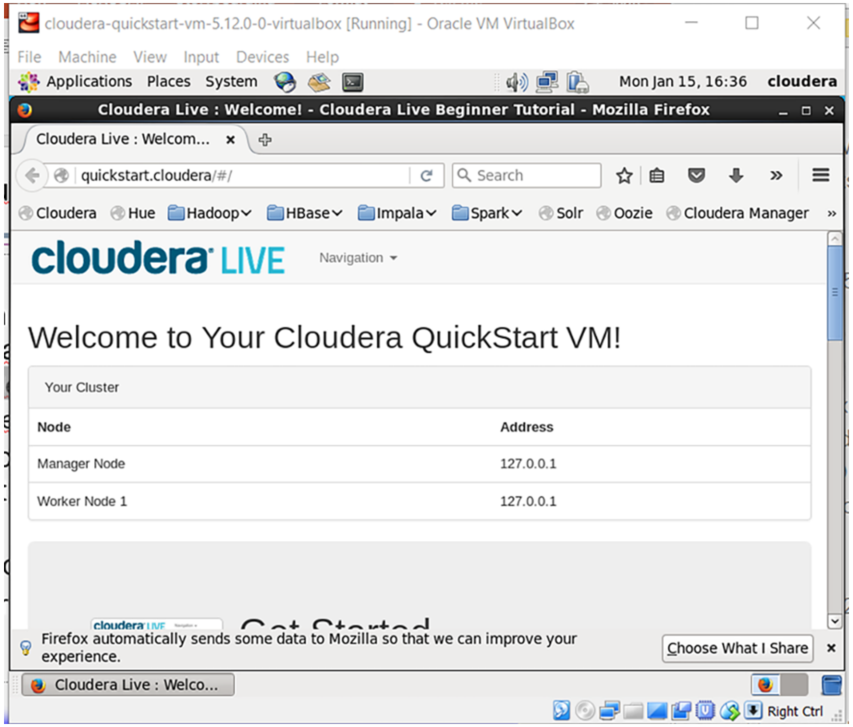
□ Read-only

☑ Auto-mount

Mount point: Exercises

- Advice: Place all your working files in one folder on your drive

Cloudera Quickstart VM for VirtualBox


cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Mon Jan 15, 16:36 cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager
cloudera LIVE
Navigation
Get Started
Choose What I Share
Cloudera Live : Welco...
Right Ctrl

Welcome to Your Cloudera QuickStart VM!

Your Cluster

Node	Address
Manager Node	127.0.0.1
Worker Node 1	127.0.0.1

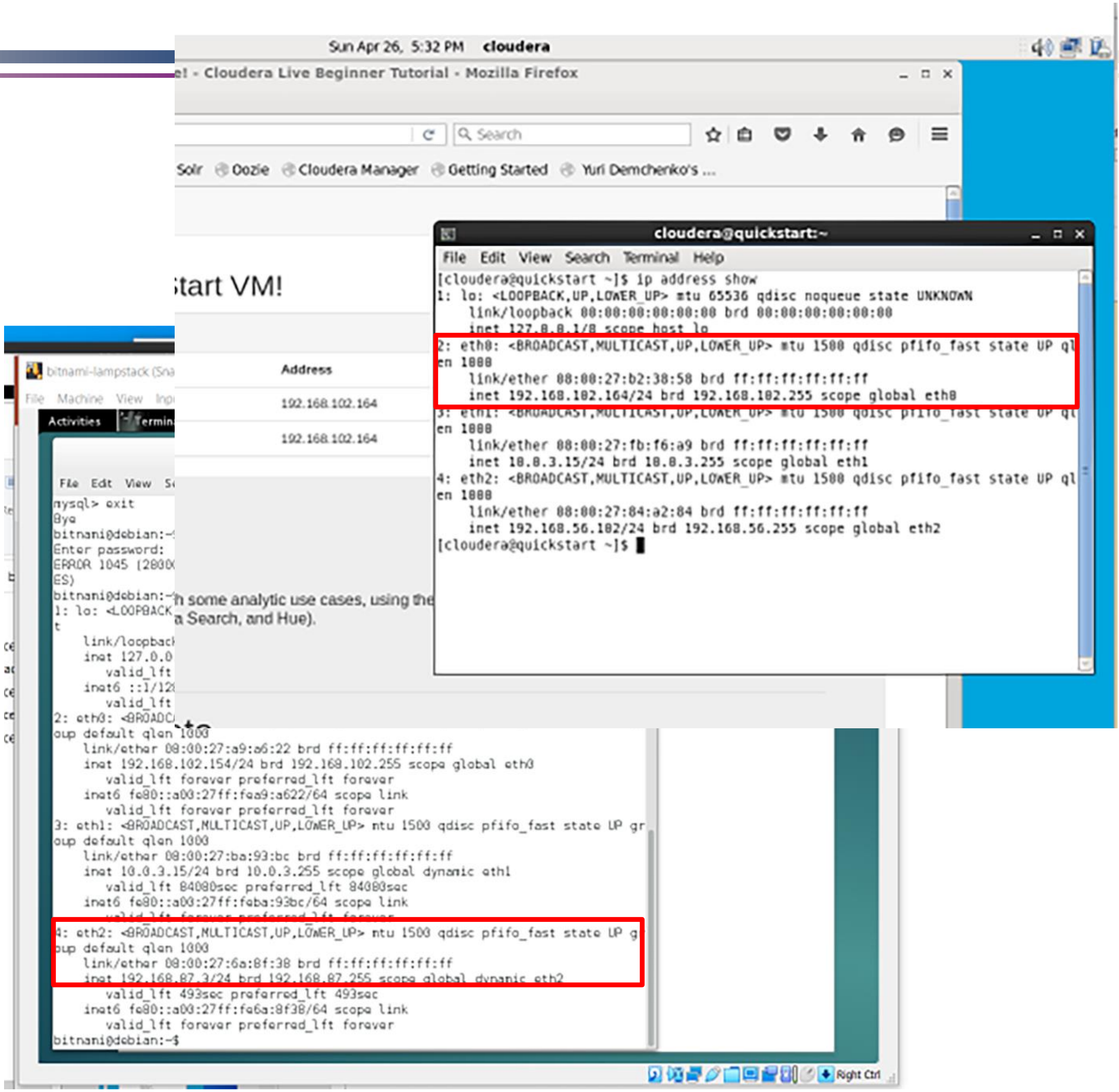
Firefox automatically sends some data to Mozilla so that we can improve your experience.

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.

Find IP address of your VM

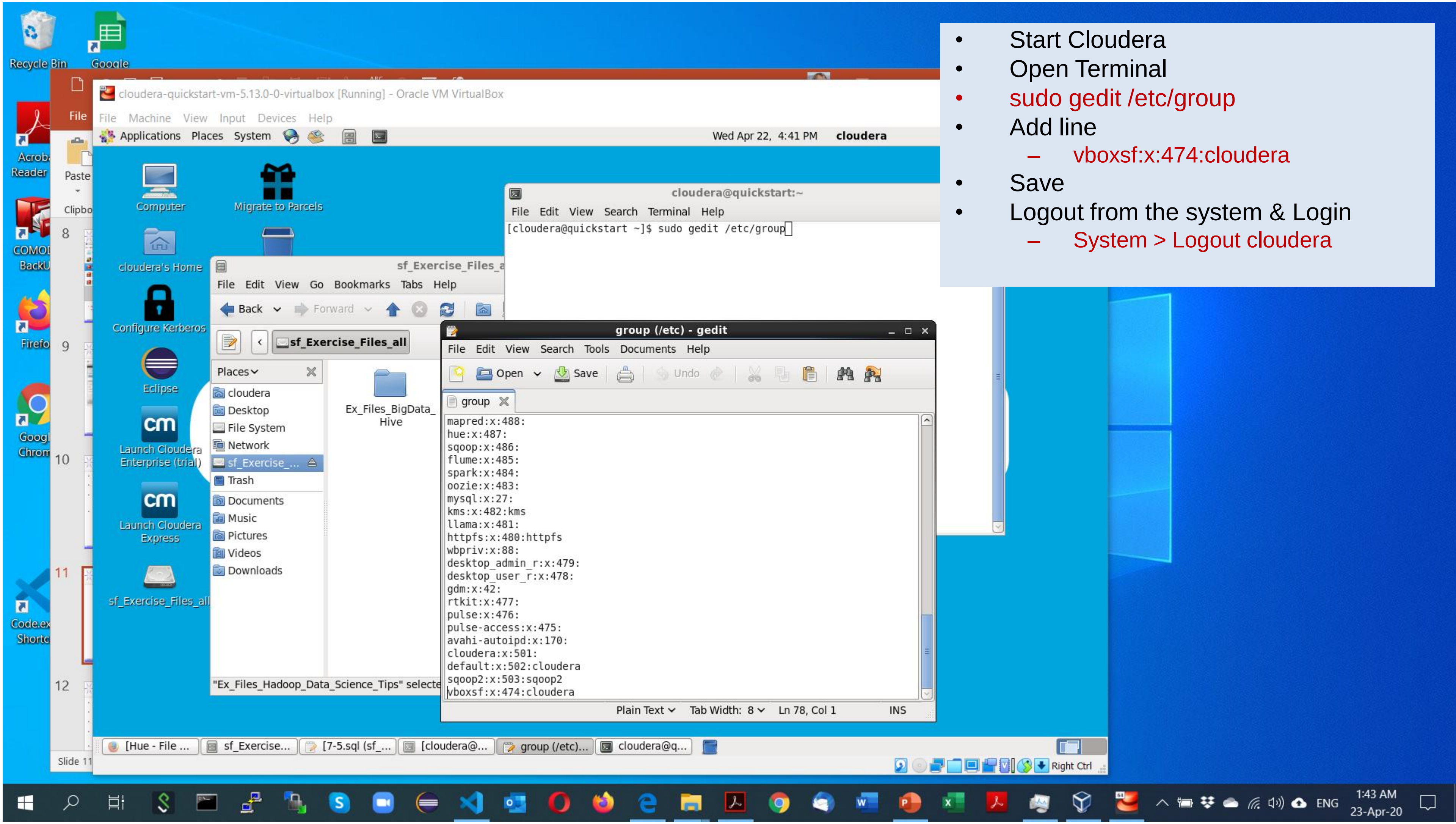
- Debian
  - $ ip address show
- Ubuntu
  - $ ipconfig
- Example:
  - 192.168.102.154 – LAN
  - 192.168.163.3 – VirtualBox Host only
- Use for configuring your SSH or SCP client



```
[cloudera@quickstart ~]$ ip address show
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
2: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP ql
en 1000
    link/ether 08:00:27:b2:38:58 brd ff:ff:ff:ff:ff:ff
    inet 192.168.102.164/24 brd 192.168.102.255 scope global eth0
3: eth1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP ql
en 1000
    link/ether 08:00:27:fb:f6:a9 brd ff:ff:ff:ff:ff:ff
    inet 10.0.3.15/24 brd 10.0.3.255 scope global eth1
4: eth2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP ql
en 1000
    link/ether 08:00:27:84:a2:84 brd ff:ff:ff:ff:ff:ff
    inet 192.168.56.102/24 brd 192.168.56.255 scope global eth2
[cloudera@quickstart ~]$
```

```
mysql> exit
Bye
bitnami@debian:~
Enter password:
ERROR 1045 (2800
ES)
bitnami@debian:~
1: lo: <LOOPBACK
t
    link/loopback
    inet 127.0.0
       valid_lft
    inet6 ::1/12
       valid_lft
2: eth0: <BROADC
oup default qlen 1000
    link/ether 08:00:27:a9:a6:22 brd ff:ff:ff:ff:ff:ff
    inet 192.168.102.154/24 brd 192.168.102.255 scope global eth0
       valid_lft forever preferred_lft forever
    inet6 fe80::a00:27ff:fea9:a622/64 scope link
       valid_lft forever preferred_lft forever
3: eth1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP gr
oup default qlen 1000
    link/ether 08:00:27:ba:93:bc brd ff:ff:ff:ff:ff:ff
    inet 10.0.3.15/24 brd 10.0.3.255 scope global dynamic eth1
       valid_lft 84080sec preferred_lft 84080sec
    inet6 fe80::a00:27ff:feba:93bc/64 scope link
       valid_lft forever preferred_lft forever
4: eth2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP gr
oup default qlen 1000
    link/ether 08:00:27:6a:8f:38 brd ff:ff:ff:ff:ff:ff
    inet 192.168.87.3/24 brd 192.168.87.255 scope global dynamic eth2
       valid_lft 493sec preferred_lft 493sec
    inet6 fe80::a00:27ff:fe6a:8f38/64 scope link
       valid_lft forever preferred_lft forever
bitnami@debian:~$
```

Add Permission for User “cloudera”


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
Applications Places System
Wed Apr 22, 4:41 PM cloudera
cloudera@quickstart:~
sf_Exercise_Files_all
group (/etc) - gedit
Plain Text Tab Width: 8 Ln 78, Col 1 INS

```
[cloudera@quickstart ~]$ sudo gedit /etc/group
```

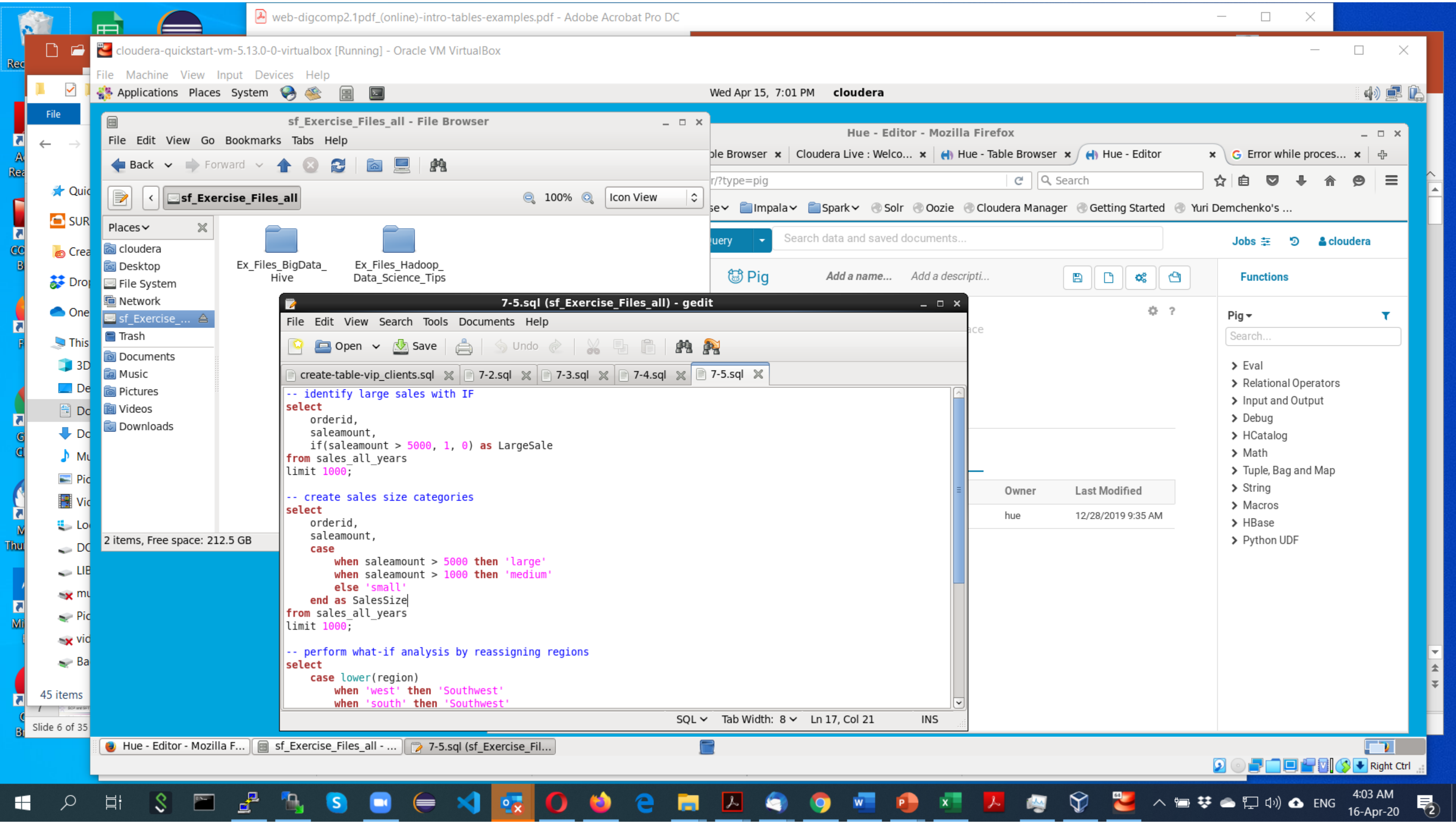
```
mapred:x:488:
hue:x:487:
sqoop:x:486:
flume:x:485:
spark:x:484:
oozie:x:483:
mysql:x:27:
kms:x:482:kms
llama:x:481:
httpfs:x:480:httpfs
wbpriv:x:88:
desktop_admin_r:x:479:
desktop_user_r:x:478:
gdm:x:42:
rtkit:x:477:
pulse:x:476:
pulse-access:x:475:
avahi-autoipd:x:170:
cloudera:x:501:
default:x:502:cloudera
sqoop2:x:503:sqoop2
vboxsf:x:474:cloudera
```

- Start Cloudera
- Open Terminal
- sudo gedit /etc/group
- Add line
  - vboxsf:x:474:cloudera
- Save
- Logout from the system & Login
  - System > Logout cloudera

Working with Cloudera Quickstart

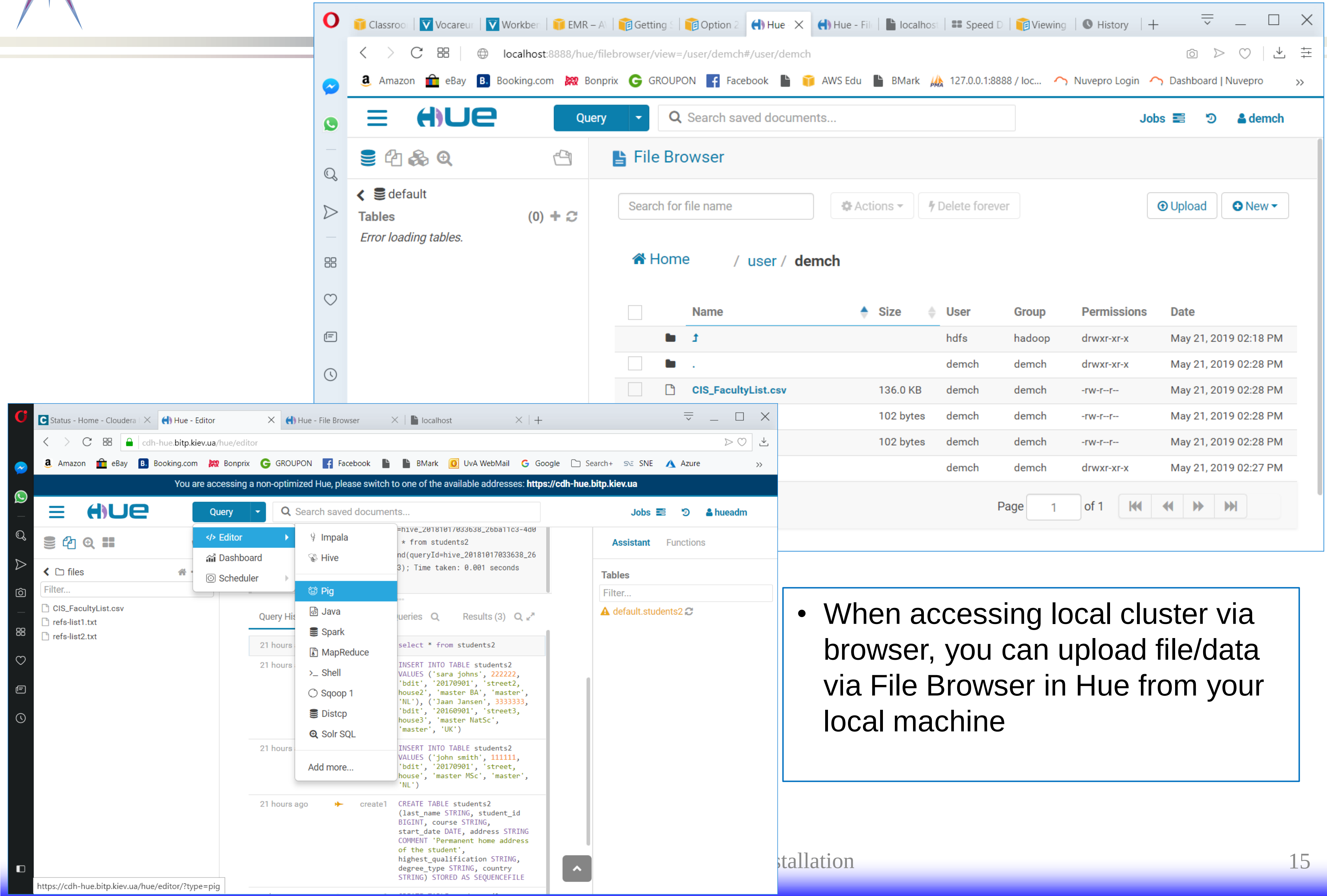
- Mount your local/host directory with exercises to Cloudera HDFS
- Use Hue visual interface
  - Similar to AWS EMR
- Use Query dropdown menu to select Query>Editor Pig, Hive, Java, Scala, etc
- Use Top-left dropdown menu to select:
  - Apps: Editor, Scheduler
  - Browser: Documents, Files, Tables, Indexes, Jobs
- Use Browsers > Files to Upload your datasets or script
  - Select right button Upload > Files
  - When connected with browser via SSH Tunnel, you will have access to local host directories

Exercises local directory mounted to Cloudera FS (typically in the /media share)



```
-- identify large sales with IF
select
    orderid,
    saleamount,
    if(saleamount > 5000, 1, 0) as LargeSale
from sales_all_years
limit 1000;

-- create sales size categories
select
    orderid,
    saleamount,
    case
        when saleamount > 5000 then 'large'
        when saleamount > 1000 then 'medium'
        else 'small'
    end as SalesSize
from sales_all_years
limit 1000;

-- perform what-if analysis by reassigning regions
select
    case lower(region)
        when 'west' then 'Southwest'
        when 'south' then 'Southwest'
```

Hue: File Browser/Directory and Editor view



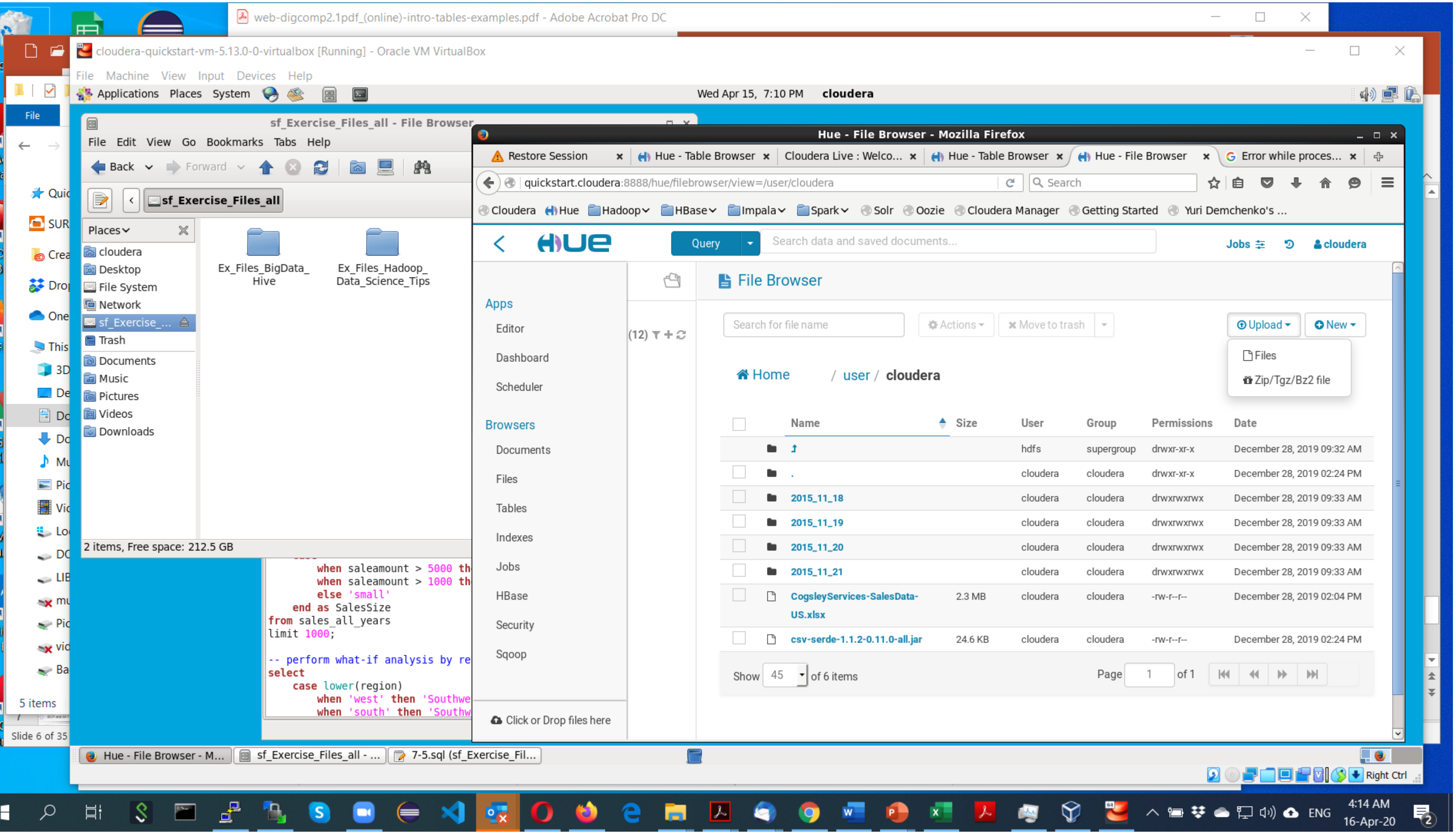
	Name	Size	User	Group	Permissions	Date
	↥		hdfs	hadoop	drwxr-xr-x	May 21, 2019 02:18 PM
□	.		demch	demch	drwxr-xr-x	May 21, 2019 02:28 PM
□	CIS_FacultyList.csv	136.0 KB	demch	demch	-rw-r--r--	May 21, 2019 02:28 PM
		102 bytes	demch	demch	-rw-r--r--	May 21, 2019 02:28 PM
		102 bytes	demch	demch	-rw-r--r--	May 21, 2019 02:28 PM
			demch	demch	drwxr-xr-x	May 21, 2019 02:27 PM

```
=hive_20181017033638_26ba11c3-4d0
 * from students2
nd(queryId=hive_20181017033638_26
3); Time taken: 0.001 seconds
```

```
21 hours          select * from students2

21 hours          INSERT INTO TABLE students2
                  VALUES ('sara johns', 222222,
                  'bdit', '20170901', 'street2,
                  house2', 'master BA', 'master',
                  'NL'), ('Jaan Jansen', 3333333,
                  'bdit', '20160901', 'street3,
                  house3', 'master NatSc',
                  'master', 'UK')

21 hours          INSERT INTO TABLE students2
                  VALUES ('john smith', 111111,
                  'bdit', '20170901', 'street,
                  house', 'master MSc', 'master',
                  'NL')

21 hours ago  create1  CREATE TABLE students2
                  (last_name STRING, student_id
                  BIGINT, course STRING,
                  start_date DATE, address STRING
                  COMMENT 'Permanent home address
                  of the student',
                  highest_qualification STRING,
                  degree_type STRING, country
                  STRING) STORED AS SEQUENCEFILE
```

- When accessing local cluster via browser, you can upload file/data via File Browser in Hue from your local machine

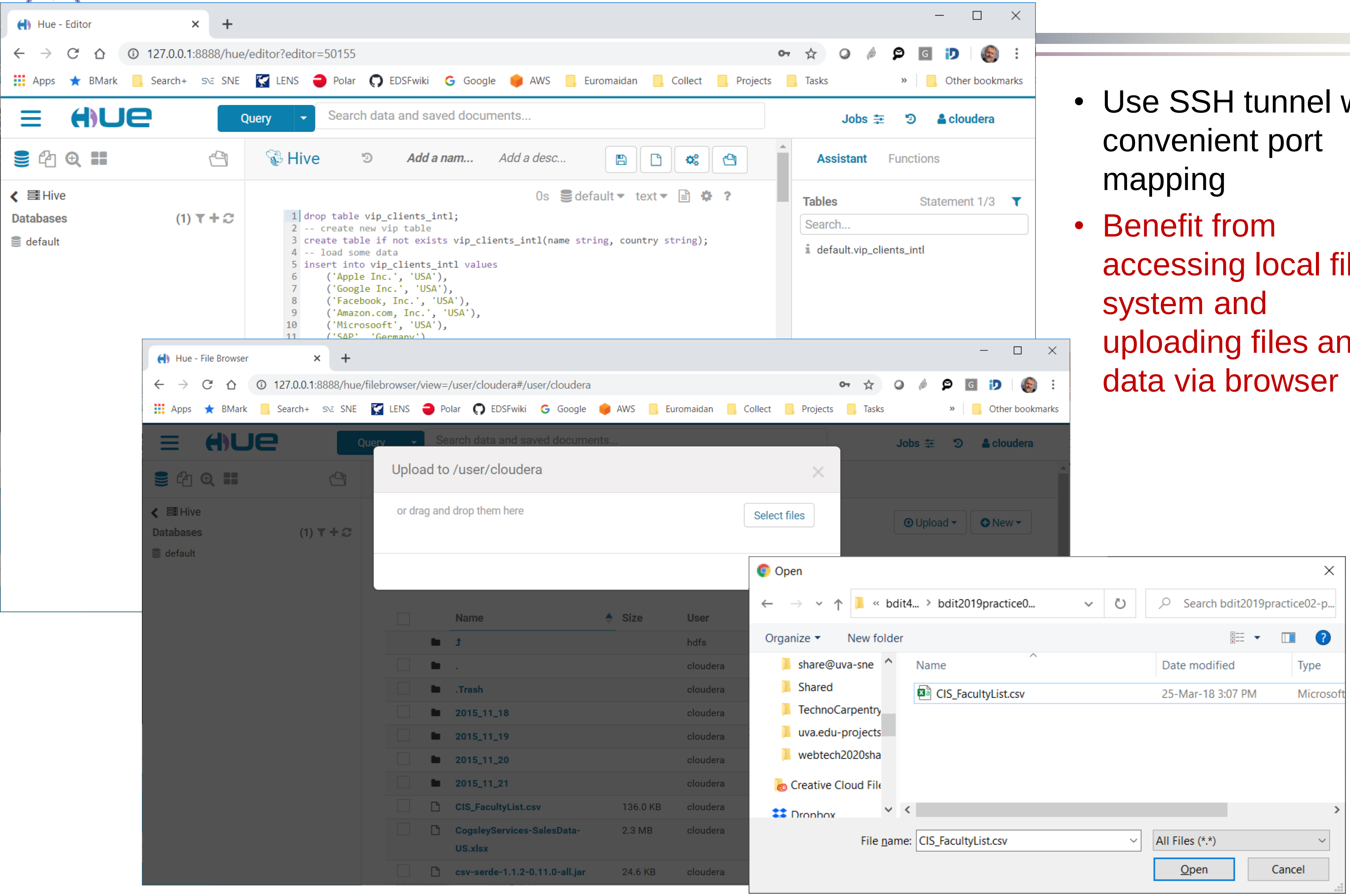
Use File Browser to Upload files


web-digcomp2.1pdf_(online)-intro-tables-examples.pdf - Adobe Acrobat Pro DC
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Wed Apr 15, 7:10 PM cloudera
sf_Exercise_Files_all - File Browser
File Edit View Go Bookmarks Tabs Help
Back Forward
sf_Exercise_Files_all
Places
cloudera
Desktop
File System
Network
sf_Exercise_...
Trash
Documents
Music
Pictures
Videos
Downloads
Ex_Files_BigData_Hive
Ex_Files_Hadoop_Data_Science_Tips
2 items, Free space: 212.5 GB
Hue - File Browser - Mozilla Firefox
Restore Session
Hue - Table Browser
Cloudera Live : Welco...
Hue - Table Browser
Hue - File Browser
Error while proces...
quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started Yuri Demchenko's ...
HUE
Query
Search data and saved documents...
Jobs
cloudera
Apps
Editor
Dashboard
Scheduler
Browsers
Documents
Files
Tables
Indexes
Jobs
HBase
Security
Sqoop
(12)
Click or Drop files here
File Browser
Search for file name
Actions
Move to trash
Upload
New
Files
Zip/Tgz/Bz2 file
Home
/ user / cloudera
Show 45 of 6 items
Page 1 of 1
Hue - File Browser - M...
sf_Exercise_Files_all - ...
7-5.sql (sf_Exercise_Fil...
Right Ctrl
5 items
Slide 6 of 35
4:14 AM
16-Apr-20
ENG

Name	Size	User	Group	Permissions	Date
↥		hdfs	supergroup	drwxr-xr-x	December 28, 2019 09:32 AM
.		cloudera	cloudera	drwxr-xr-x	December 28, 2019 02:24 PM
2015_11_18		cloudera	cloudera	drwxrwxrwx	December 28, 2019 09:33 AM
2015_11_19		cloudera	cloudera	drwxrwxrwx	December 28, 2019 09:33 AM
2015_11_20		cloudera	cloudera	drwxrwxrwx	December 28, 2019 09:33 AM
2015_11_21		cloudera	cloudera	drwxrwxrwx	December 28, 2019 09:33 AM
CogsleyServices-SalesData-US.xlsx	2.3 MB	cloudera	cloudera	-rw-r--r--	December 28, 2019 02:04 PM
csv-serde-1.1.2-0.11.0-all.jar	24.6 KB	cloudera	cloudera	-rw-r--r--	December 28, 2019 02:24 PM

```
      when saleamount > 5000 th
      when saleamount > 1000 th
      else 'small'
   end as SalesSize
from sales_all_years
limit 1000;

-- perform what-if analysis by re
select
   case lower(region)
      when 'west' then 'Southwe
      when 'south' then 'Southw
```

Accessing CDH cluster from your browser



Hue - Editor

127.0.0.1:8888/hue/editor?editor=50155

```
drop table vip_clients_intl;
-- create new vip table
create table if not exists vip_clients_intl(name string, country string);
-- load some data
insert into vip_clients_intl values
    ('Apple Inc.', 'USA'),
    ('Google Inc.', 'USA'),
    ('Facebook, Inc.', 'USA'),
    ('Amazon.com, Inc.', 'USA'),
    ('Microsooft', 'USA'),
    ('SAP', 'Germany')
```

Tables: default.vip_clients_intl

Hue - File Browser

127.0.0.1:8888/hue/filebrowser/view=/user/cloudera#/user/cloudera

Upload to /user/cloudera

or drag and drop them here

Name	Size	User
↑		hdfs
.		cloudera
.Trash		cloudera
2015_11_18		cloudera
2015_11_19		cloudera
2015_11_20		cloudera
2015_11_21		cloudera
CIS_FacultyList.csv	136.0 KB	cloudera
CogsleyServices-SalesData-US.xlsx	2.3 MB	cloudera
csv-serde-1.1.2-0.11.0-all.jar	24.6 KB	cloudera

Open

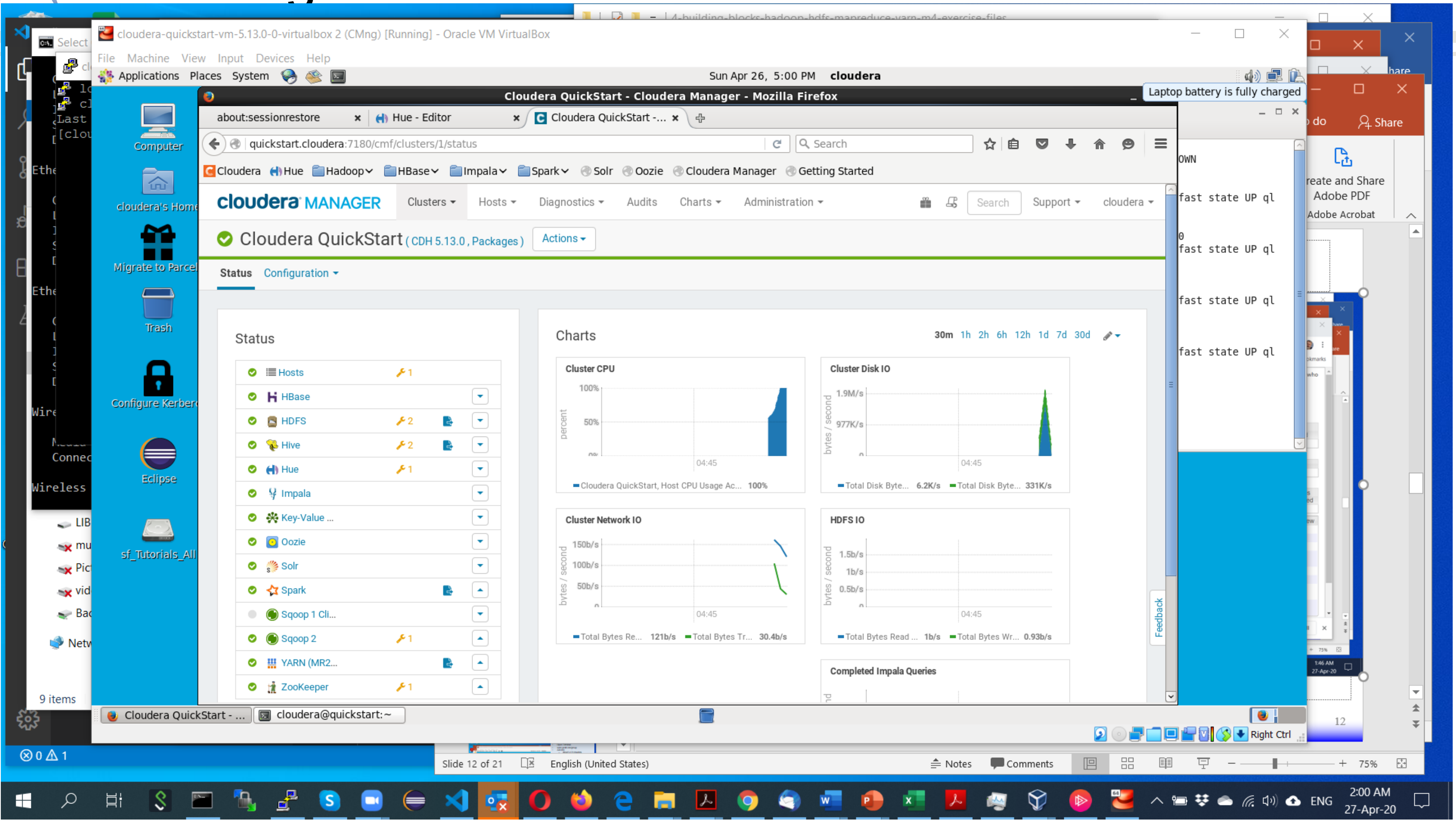
Name	Date modified	Type
CIS_FacultyList.csv	25-Mar-18 3:07 PM	Microsoft

File name: CIS_FacultyList.csv

All Files (*.*)

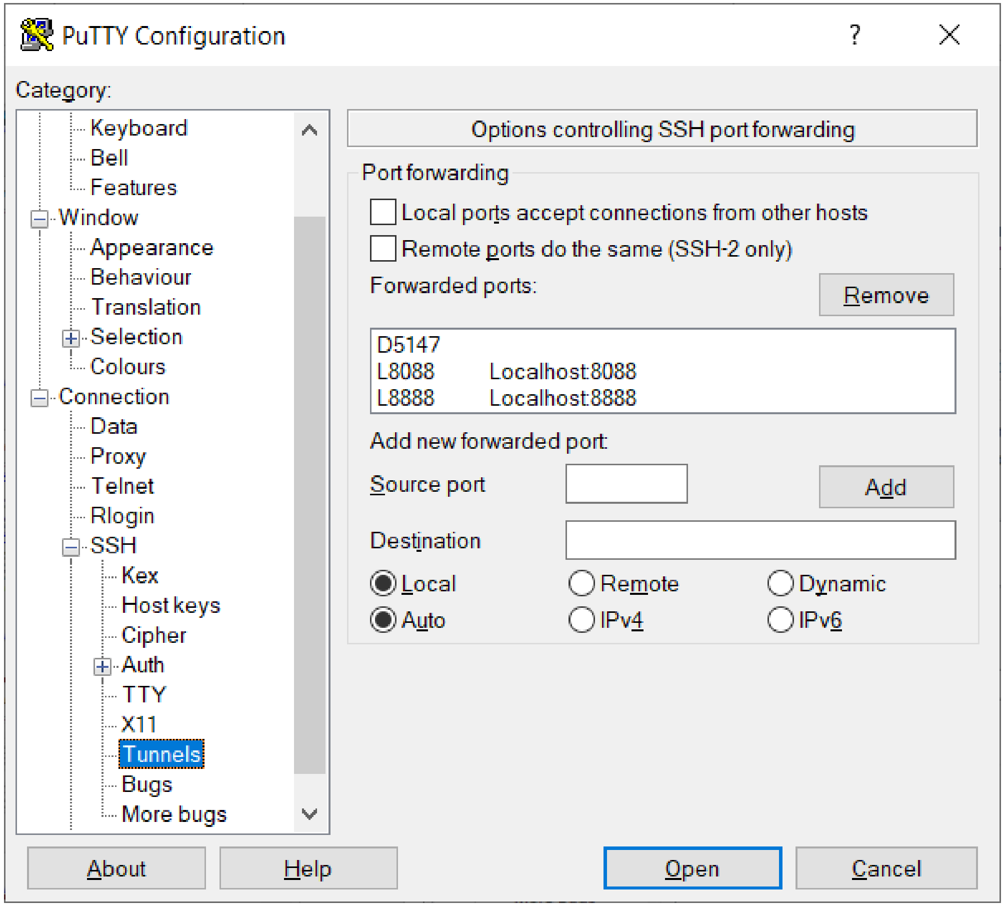
- Use SSH tunnel with convenient port mapping
- Benefit from accessing local file system and uploading files and data via browser

Cloudera Quickstart with enabled Cloudera Manager


cloudera-quickstart-vm-5.13.0-0-virtualbox 2 (CMng) [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Apr 26, 5:00 PM cloudera
Cloudera QuickStart - Cloudera Manager - Mozilla Firefox
Laptop battery is fully charged
about:sessionrestore
Hue - Editor
Cloudera QuickStart -...
quickstart.cloudera:7180/cmf/clusters/1/status
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera MANAGER
Clusters Hosts Diagnostics Audits Charts Administration
Search Support cloudera
Cloudera QuickStart (CDH 5.13.0, Packages)
Actions
Status Configuration
Status
Hosts
HBase
HDFS
Hive
Hue
Impala
Key-Value ...
Oozie
Solr
Spark
Sqoop 1 Cli...
Sqoop 2
YARN (MR2...
ZooKeeper
Charts
30m 1h 2h 6h 12h 1d 7d 30d
Cluster CPU
100%
50%
percent
04:45
Cloudera QuickStart, Host CPU Usage Ac... 100%
Cluster Disk IO
1.9M/s
977K/s
bytes / second
04:45
Total Disk Byte... 6.2K/s
Total Disk Byte... 331K/s
Cluster Network IO
150b/s
100b/s
50b/s
bytes / second
04:45
Total Bytes Re... 121b/s
Total Bytes Tr... 30.4b/s
HDFS IO
1.5b/s
1b/s
0.5b/s
bytes / second
04:45
Total Bytes Read ... 1b/s
Total Bytes Wr... 0.93b/s
Completed Impala Queries
Feedback
Computer
cloudera's Home
Migrate to Parcels
Trash
Configure Kerberos
Eclipse
sf_Tutorials_All
Cloudera QuickStart - ...
cloudera@quickstart:~
Right Ctrl
Create and Share Adobe PDF
Adobe Acrobat
Share
Slide 12 of 21
English (United States)
Notes
Comments
75%
ENG
2:00 AM
27-Apr-20

- To run Cloudera Manager Launch Cloudera Express from Desktop
- Be aware about requirement for configuring at least 2 virtual CPU and minimum memory 8 GB

Configuring Tunnels for known Ports



Options controlling SSH port forwarding

Port forwarding

□ Local ports accept connections from other hosts

□ Remote ports do the same (SSH-2 only)

Forwarded ports:

```
D5147
L8088     Localhost:8088
L8888     Localhost:8888
```

Add new forwarded port:

Source port

Destination

● Local ○ Remote ○ Dynamic

● Auto ○ IPv4 ○ IPv6

- Using PuTTY SSH client
- First, configure dynamic tunnel to EMR cluster
  - Port 8157
  - Using Source port and Dynamic radio button
  - Example:
    - D5147
    - Source L8088
Destination Localhost:8088
    - Source L8888
Destination Localhost:8888
- Configure other local ports
  - See EMR ports table
- This will create a secure tunnel by forwarding a port (the “destination port”) on the remote server to a port (the “source port”) on the local host (127.0.0.1 or localhost).

Configuring SSH tunnel on Windows https://docs.bitnami.com/virtual-machine/faq/get-started/access-phpmyadmin/

- In order to access phpMyAdmin via SSH tunnel, you need an SSH client. In the instructions below we have selected PuTTY, a free SSH client for Windows and UNIX platforms.
  - The first step is to configure PuTTY.
- Once you have your SSH client correctly configured and you have confirmed that you can successfully access your instance using SSH, you need to create an SSH tunnel in order to access phpMyAdmin. Follow these steps:
  - In the “Connection -> SSH -> Tunnels” section, add a new forwarded port by introducing the following values:
    - Source port: 8888
    - Destination: localhost:8888
    - Remember that if you are redirecting HTTP requests to the HTTPS port, you must use destination port 443 instead of 80.
  - This will create a secure tunnel by forwarding a port (the “destination port”) on the remote server to a port (the “source port”) on the local host (127.0.0.1 or localhost).
  - Click the “Add” button to add the secure tunnel configuration to the session. You’ll see the added port in the list of “Forwarded ports”.
- PuTTY configuration
  - In the “Session” section, save your changes by clicking the “Save” button.
  - Click the “Open” button to open an SSH session to the server. The SSH session will now include a secure SSH tunnel between the two specified ports.
- Access the phpMyAdmin console through the secure SSH tunnel you created, by browsing to http://127.0.0.1:8888/phpmyadmin.
- Log in to phpMyAdmin by using the following credentials:
  - Username: root
  - Password: application password. (Refer to our FAQ to learn how to find your application credentials).

Running MapReduce examples on CDH5.13

Follow step, adjust jar name and commands
https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_usage.html#topic_5_2

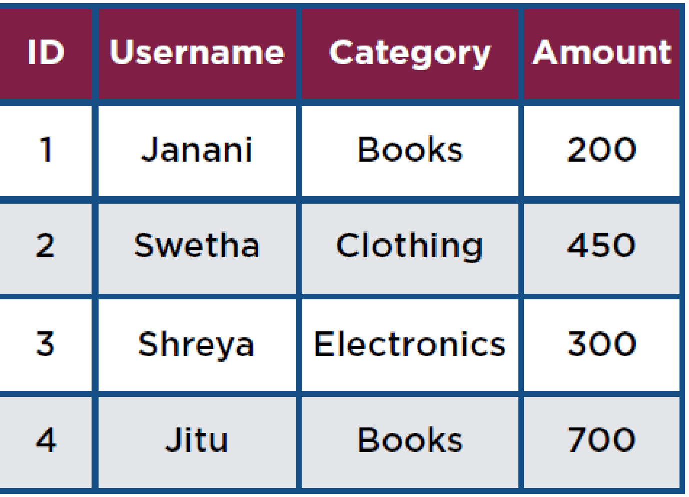
- $ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount /user/cloudera/wordcount/output
- $ hadoop fs –rm –r /user/cloudera/wordcount/output
- $ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 10 10000
- $ hadoop fs –rm –r /user/cloudera/wordcount/output
- $hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordmean /user/cloudera/wordcount/output

Commands/Programs in hadoop-mapreduce-examples.jar

Jobs in the examples JAR file hadoop-mapreduce-examples.jar

- **aggregatewordcount**: An Aggregate-based map/reduce program that counts the words in the input files.
- **aggregatewordhist:** An Aggregate-based map/reduce program that computes the histogram of the words in the input files.
- **bbp:** A map/reduce program that uses Bailey-Borwein-Plouffe to compute the exact digits of pi.
- **dbcount**: An example job that counts the pageview counts from a database.
- **distbbp**: A map/reduce program that uses a BBP-type formula to compute the exact bits of pi.
- **grep:** A map/reduce program that counts the matches to a regex in the input.
- **join:** A job that effects a join over sorted, equally partitioned data sets.
- **multifilewc**: A job that counts words from several files.
- **pentomino:** A map/reduce tile laying program to find solutions to pentomino problems.
- **pi:** A map/reduce program that estimates pi using a quasi-Monte Carlo method.
- **randomtextwriter:** A map/reduce program that writes 10 GB of random textual data per node.
- **randomwriter**: A map/reduce program that writes 10 GB of random data per node.
- **secondarysort**: An example defining a secondary sort to the reduce.
- **sort:** A map/reduce program that sorts the data written by the random writer.
- **sudoku:** A Sudoku solver.
- **teragen:** Generate data for the terasort.
- **terasort:** Run the terasort.
- **teravalidate:** Check the results of the terasort.
- **wordcount:** A map/reduce program that counts the words in the input files.
- **wordmean:** A map/reduce program that counts the average length of the words in the input files.
- **wordmedian:** A map/reduce program that counts the median length of the words in the input files.
- **wordstandarddeviation:** A map/reduce program that counts the standard deviation of the length of the words in the input files.

Do you need Reduce?



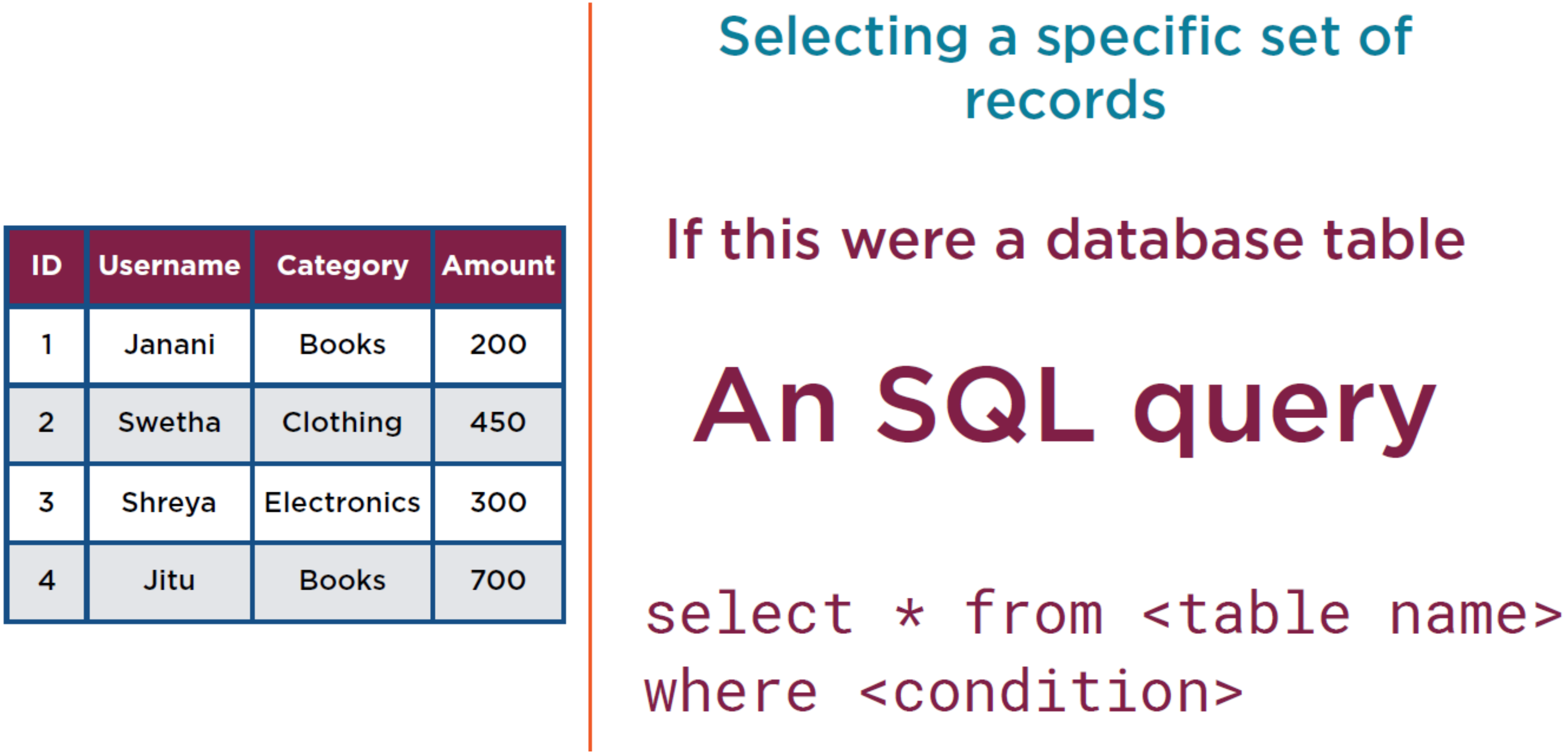
ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700

Which users spent >300?

How many of them bought Books?

Selecting a specific set of records from a dataset

SQL Query can do it



ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700

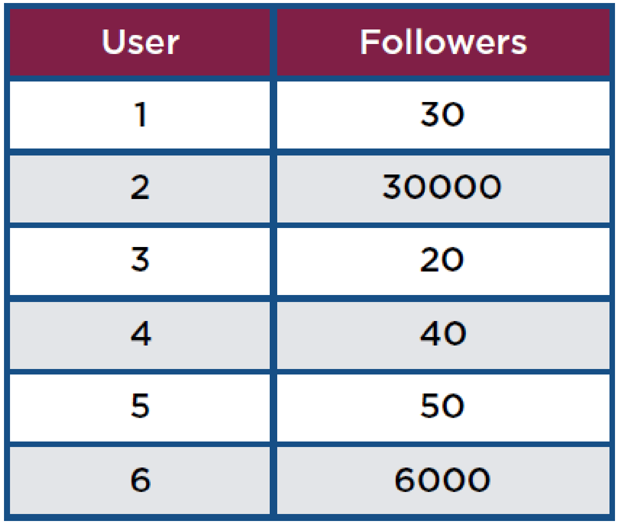
Selecting a specific set of records

If this were a database table

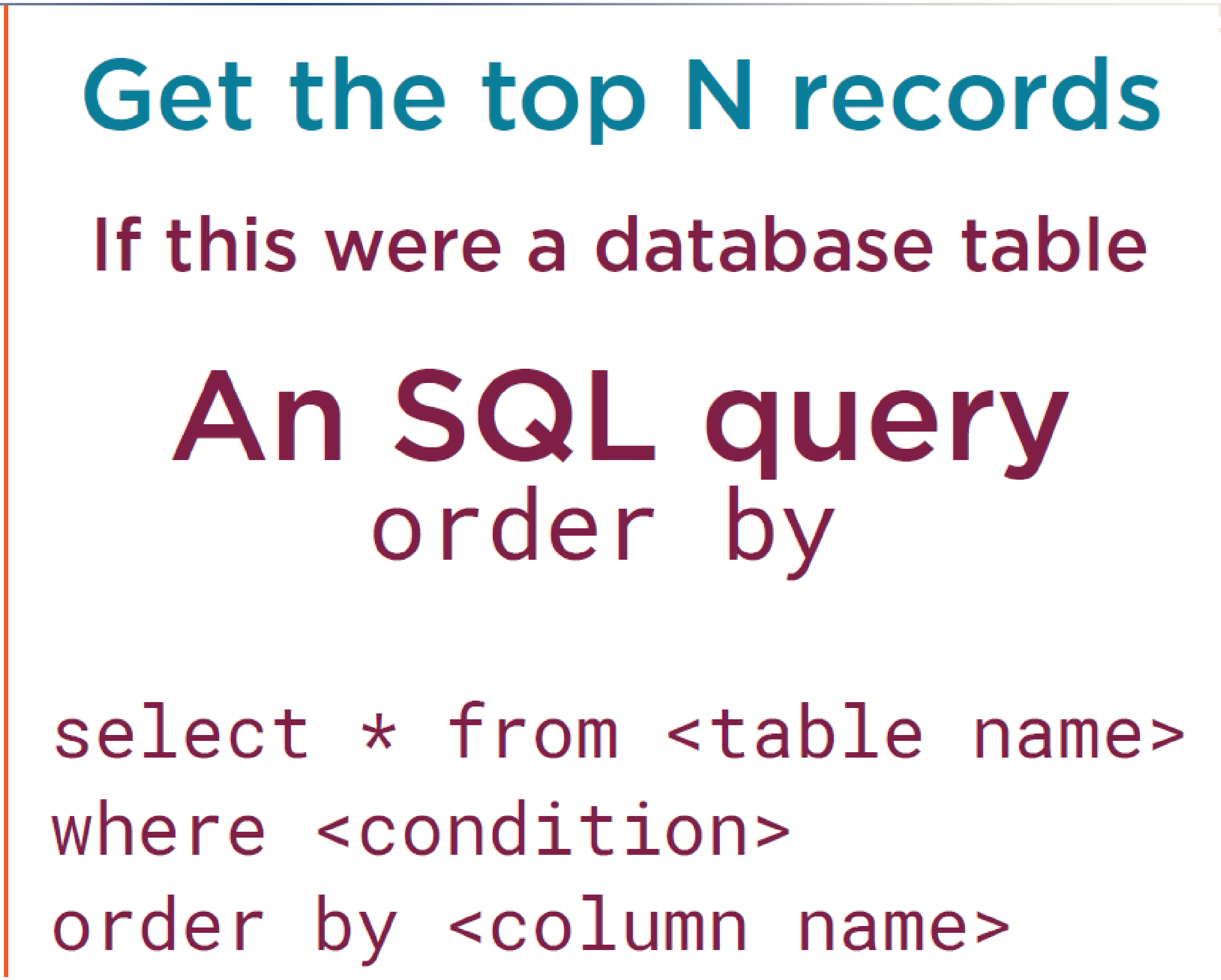
An SQL query

```
select * from <table name>
where <condition>
```

Run MapReduce for Standard Select Operation



User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000



Get the top N records

If this were a database table

An SQL query

`order by`

```
select * from <table name>
where <condition>
order by <column name>
```