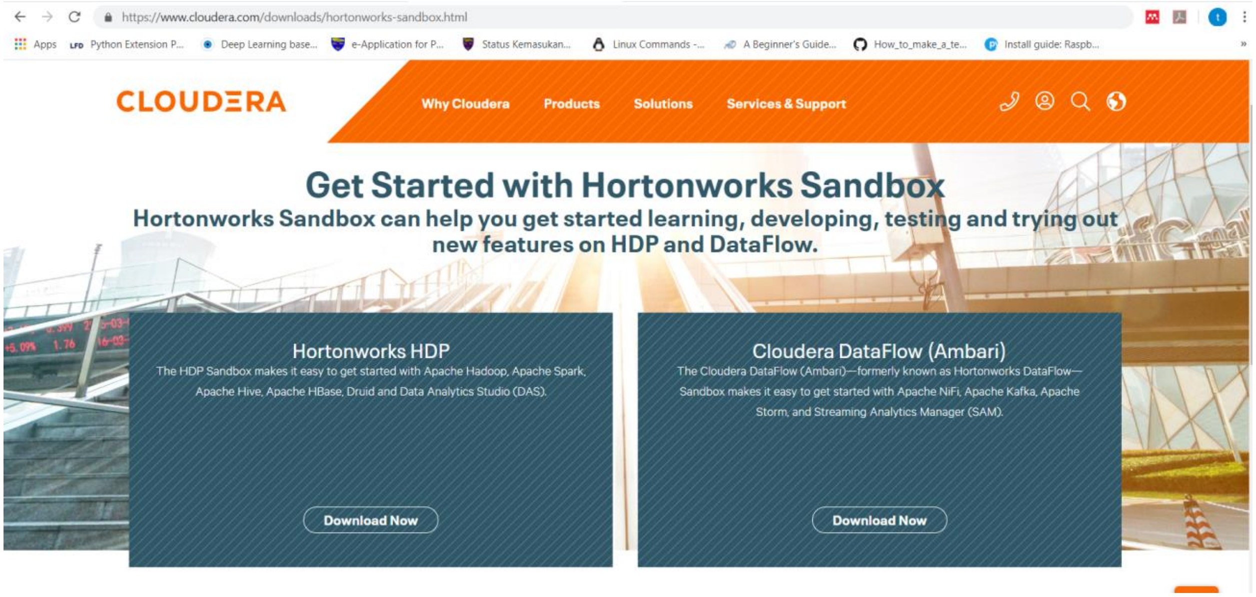
Tan Sia Hong & Tan Chang Jung

Milestone 2: Stone data set into Hive data warehouse

Part 1: Install Hortonworks HDP

1. Download Hortonworks HDP from https://www.cloudera.com/downloads/hortonworks-sandbox.html (20GB)



Get Started with Hortonworks Sandbox

Hortonworks Sandbox can help you get started learning, developing, testing and trying out new features on HDP and DataFlow.

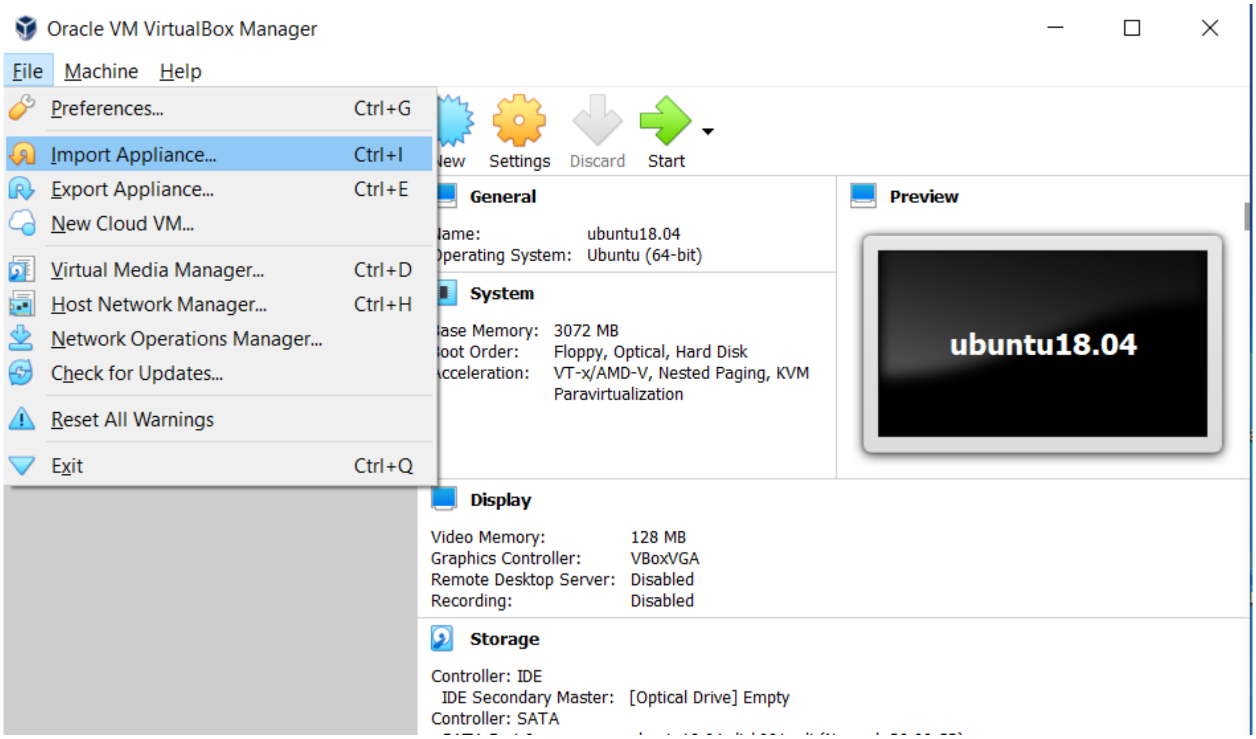
Hortonworks HDP

The HDP Sandbox makes it easy to get started with Apache Hadoop, Apache Spark, Apache Hive, Apache HBase, Druid and Data Analytics Studio (DAS).

Cloudera DataFlow (Ambari)

The Cloudera DataFlow (Ambari)—formerly known as Hortonworks DataFlow—Sandbox makes it easy to get started with Apache NiFi, Apache Kafka, Apache Storm, and Streaming Analytics Manager (SAM).

2. Import Hortonworks.ova into VM VirtualBox.
   VM VirtualBox can download from https://www.virtualbox.org/wiki/Downloads


Oracle VM VirtualBox Manager
File
Machine
Help
Preferences... Ctrl+G
Import Appliance... Ctrl+I
Export Appliance... Ctrl+E
New Cloud VM...
Virtual Media Manager... Ctrl+D
Host Network Manager... Ctrl+H
Network Operations Manager...
Check for Updates...
Reset All Warnings
Exit Ctrl+Q
New
Settings
Discard
Start
Preview
ubuntu18.04

General

```
ame:              ubuntu18.04
perating System:  Ubuntu (64-bit)
```

System

```
ase Memory:   3072 MB
oot Order:    Floppy, Optical, Hard Disk
cceleration:  VT-x/AMD-V, Nested Paging, KVM
              Paravirtualization
```

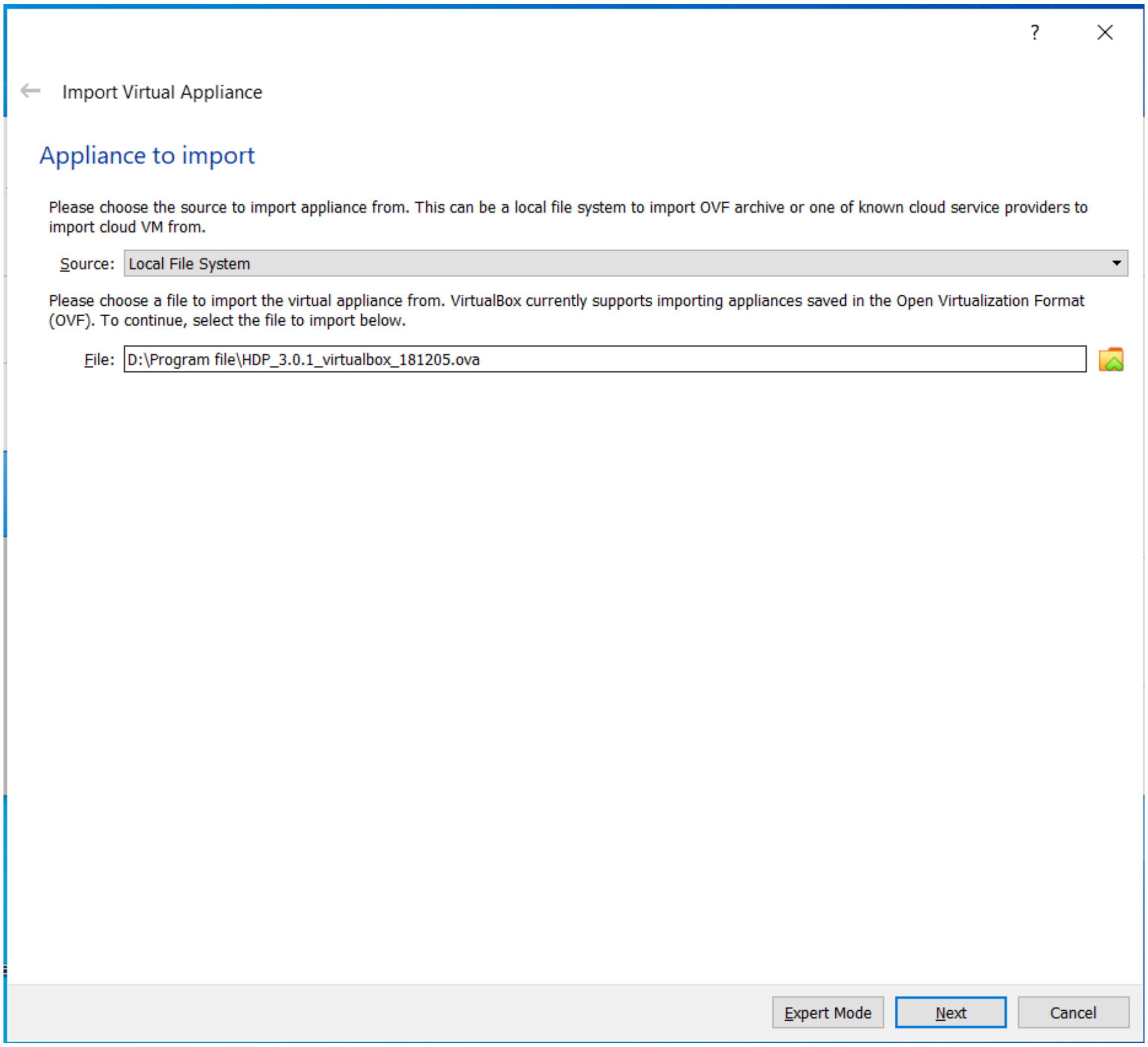
Display

```
Video Memory:           128 MB
Graphics Controller:    VBoxVGA
Remote Desktop Server:  Disabled
Recording:              Disabled
```

Storage

```
Controller: IDE
  IDE Secondary Master:  [Optical Drive] Empty
Controller: SATA
```

Select the hortonworks.ova



Import Virtual Appliance

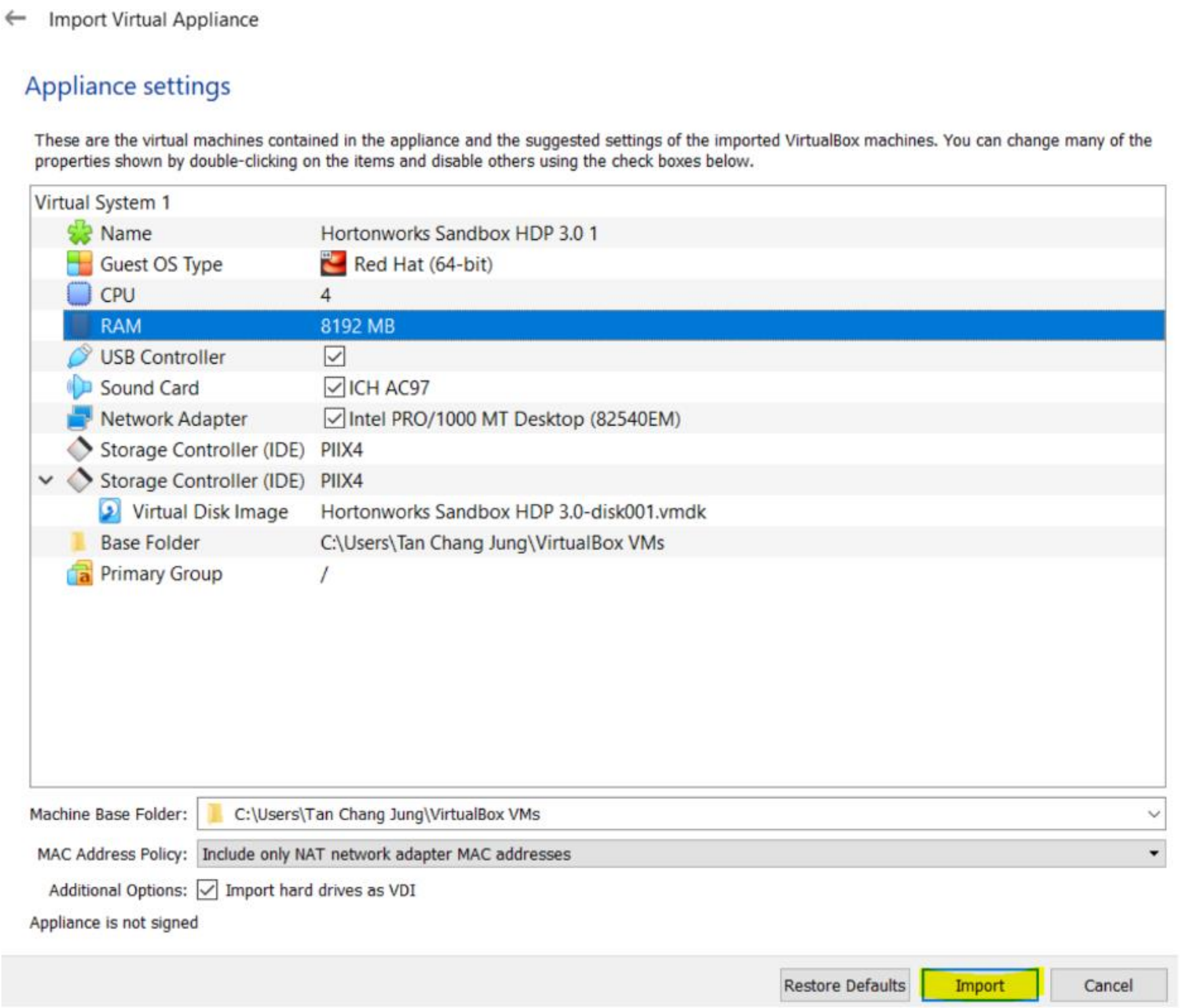
Appliance to import

Please choose the source to import appliance from. This can be a local file system to import OVF archive or one of known cloud service providers to import cloud VM from.

Source: Local File System

Please choose a file to import the virtual appliance from. VirtualBox currently supports importing appliances saved in the Open Virtualization Format (OVF). To continue, select the file to import below.

File: D:\Program file\HDP_3.0.1_virtualbox_181205.ova



← Import Virtual Appliance

Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.

Virtual System 1

Name	Hortonworks Sandbox HDP 3.0 1
Guest OS Type	Red Hat (64-bit)
CPU	4
RAM	8192 MB
USB Controller	☑
Sound Card	☑ ICH AC97
Network Adapter	☑ Intel PRO/1000 MT Desktop (82540EM)
Storage Controller (IDE)	PIIX4
Storage Controller (IDE)	PIIX4
Virtual Disk Image	Hortonworks Sandbox HDP 3.0-disk001.vmdk
Base Folder	C:\Users\Tan Chang Jung\VirtualBox VMs
Primary Group	/

Machine Base Folder: C:\Users\Tan Chang Jung\VirtualBox VMs

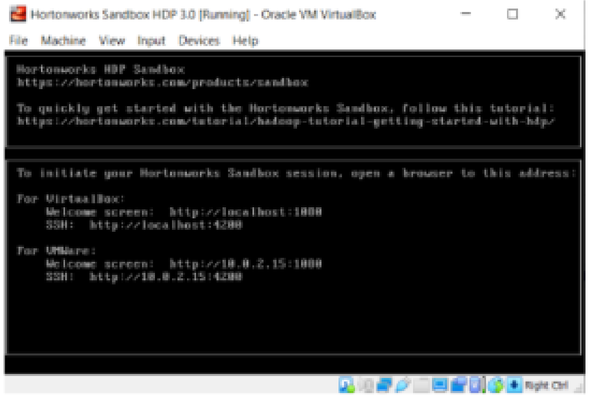
MAC Address Policy: Include only NAT network adapter MAC addresses

Additional Options: ☑ Import hard drives as VDI

Appliance is not signed

*Optimize the number of CPU cores and RAM resources before import.

3. Run the Hortonworks Sandbox HDP 3.0 for extraction and installation. Notice that first time installation will take a few minutes.

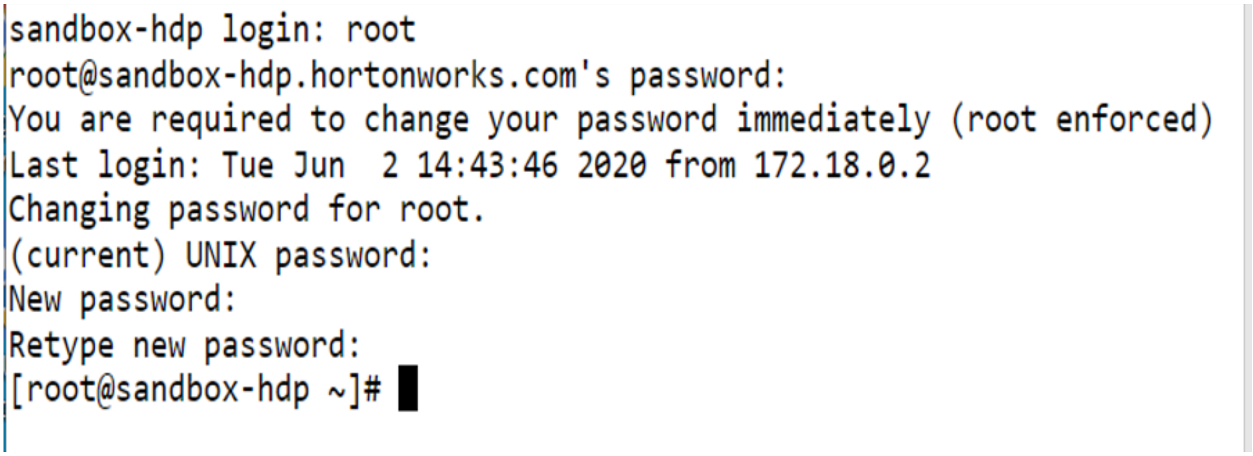


Hortonworks Sandbox HDP 3.0 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Right Ctrl

```
Hortonworks HDP Sandbox
https://hortonworks.com/products/sandbox

To quickly get started with the Hortonworks Sandbox, follow this tutorial:
https://hortonworks.com/tutorial/hadoop-tutorial-getting-started-with-hdp/

To initiate your Hortonworks Sandbox session, open a browser to this address:

For VirtualBox:
    Welcome screen:  http://localhost:1080
    SSH:  http://localhost:4200

For VMWare:
    Welcome screen:  http://10.0.2.15:1080
    SSH:  http://10.0.2.15:4200
```

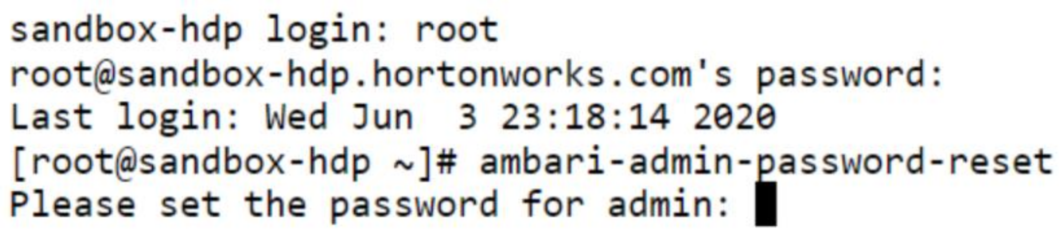
*Installation is completed

4. Use the Google Browser to access http://localhost:4200/ . The default root user login credentials will be:
   User: root
   Password: hadoop
   After logging by default password, you will request to change password. Please change the password for root user.



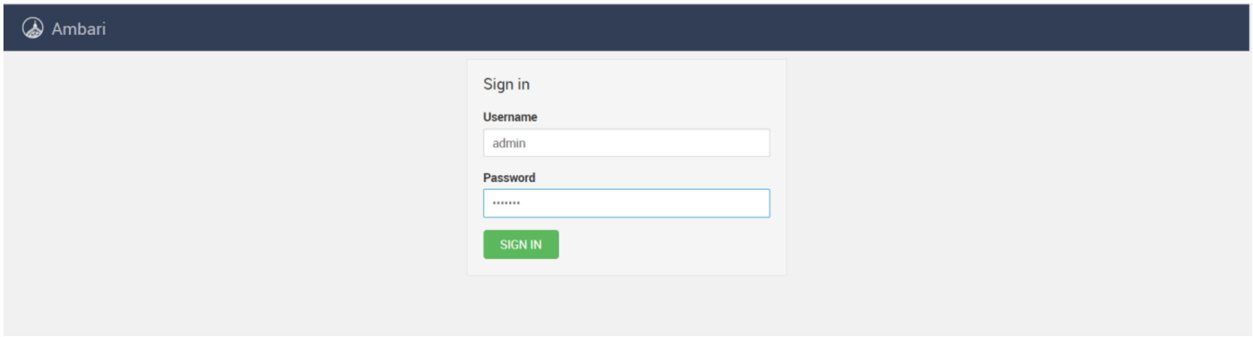
```
sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
You are required to change your password immediately (root enforced)
Last login: Tue Jun  2 14:43:46 2020 from 172.18.0.2
Changing password for root.
(current) UNIX password:
New password:
Retype new password:
[root@sandbox-hdp ~]# █
```

5. Ambari enables system administrators to provision, manage and monitor a Hadoop cluster. Now, Type ‘ambari-admin-password-reset’ to reset the Ambari’ administrator password.



```
sandbox-hdp login: root
root@sandbox-hdp.hortonworks.com's password:
Last login: Wed Jun  3 23:18:14 2020
[root@sandbox-hdp ~]# ambari-admin-password-reset
Please set the password for admin: █
```

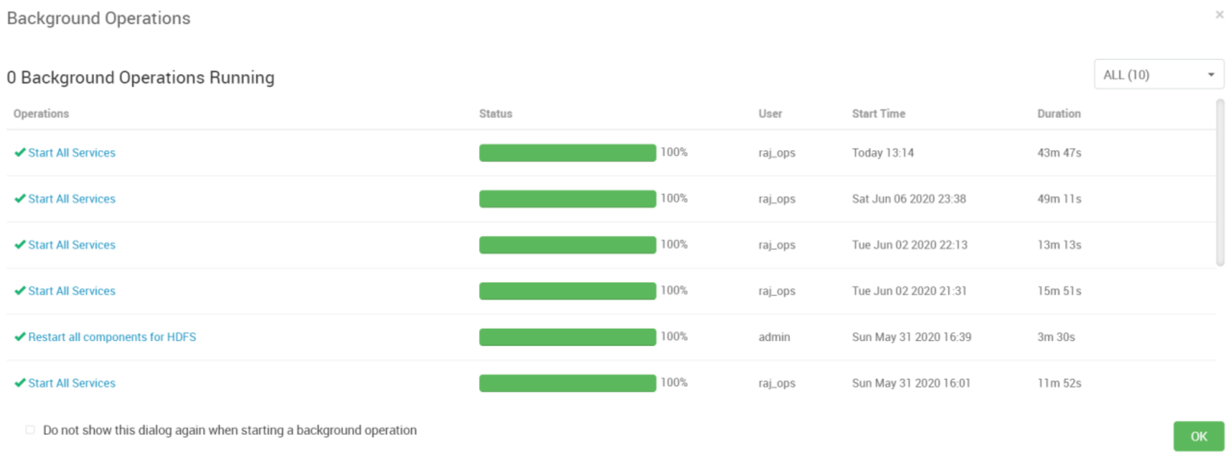
6. Access to http://localhost:8080/ for Ambari login by using Google Brower.



Username: admin

Password: •••••••

7. After logging in, the sandbox is taking some time to starting all the required services that wait time is depend on laptop performance.



Background Operations

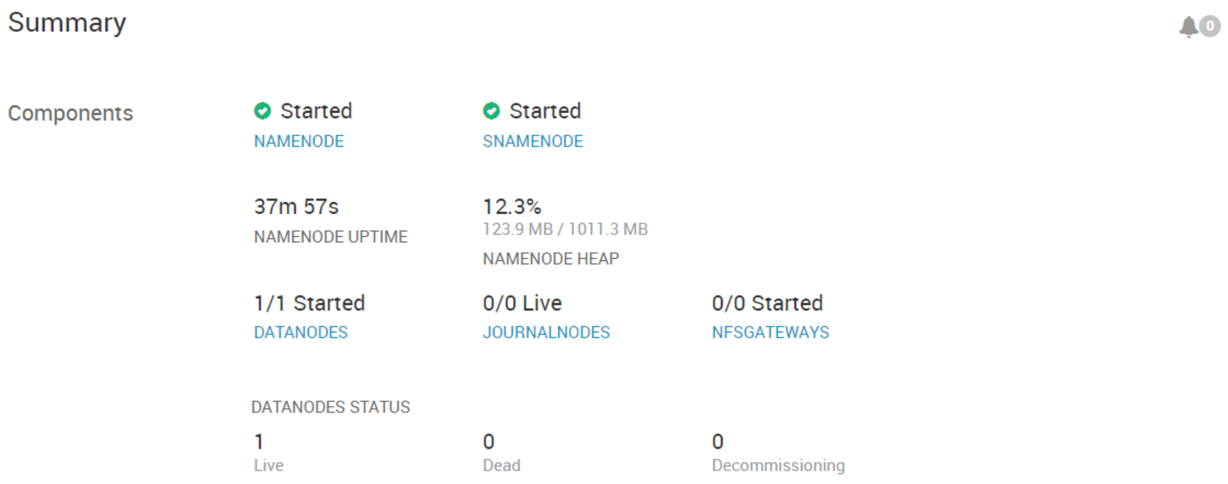
0 Background Operations Running

ALL (10)

Operations	Status	User	Start Time	Duration
✓ Start All Services	100%	raj_ops	Today 13:14	43m 47s
✓ Start All Services	100%	raj_ops	Sat Jun 06 2020 23:38	49m 11s
✓ Start All Services	100%	raj_ops	Tue Jun 02 2020 22:13	13m 13s
✓ Start All Services	100%	raj_ops	Tue Jun 02 2020 21:31	15m 51s
✓ Restart all components for HDFS	100%	admin	Sun May 31 2020 16:39	3m 30s
✓ Start All Services	100%	raj_ops	Sun May 31 2020 16:01	11m 52s

□ Do not show this dialog again when starting a background operation

8. The Hive and HDFS are completely set up and ready to be used.



Summary

Components

Started
NAMENODE

Started
SNAMENODE

37m 57s
NAMENODE UPTIME

12.3%
123.9 MB / 1011.3 MB
NAMENODE HEAP

1/1 Started
DATANODES

0/0 Live
JOURNALNODES

0/0 Started
NFSGATEWAYS

DATANODES STATUS

1
Live

0
Dead

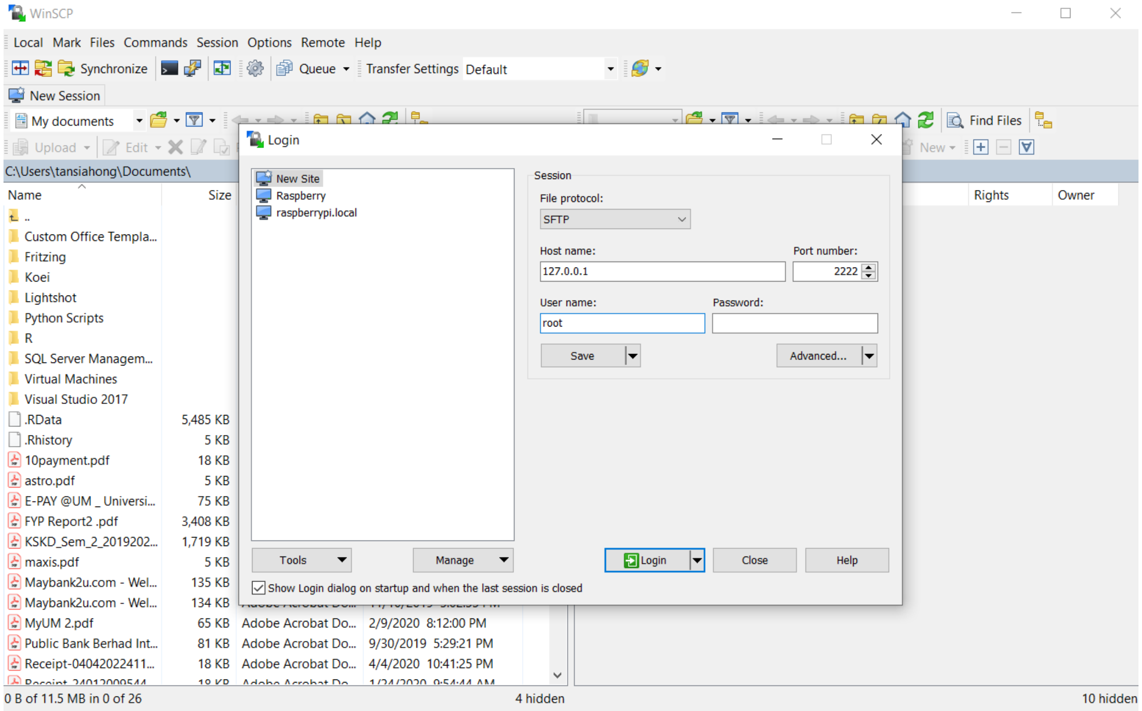
0
Decommissioning

PART 2: Store data into Hive database

1. Download and install WinSCP from https://winscp.net/eng/download.php

WinSCP 5.17 Download

2. Then, logging in the WinSCP by Host name is 127.0.0.1, port number is 2222, user name is root. Note that File protocol must need SCP type.



C:\Users\tansiahong\Documents\

Name	Size		
..			
Custom Office Templa...			
Fritzing			
Koei			
Lightshot			
Python Scripts			
R			
SQL Server Managem...			
Virtual Machines			
Visual Studio 2017			
.RData	5,485 KB		
.Rhistory	5 KB		
10payment.pdf	18 KB		
astro.pdf	5 KB		
E-PAY @UM _ Universi...	75 KB		
FYP Report2 .pdf	3,408 KB		
KSKD_Sem_2_2019202...	1,719 KB		
maxis.pdf	5 KB		
Maybank2u.com - Wel...	135 KB		
Maybank2u.com - Wel...	134 KB		
MyUM 2.pdf	65 KB	Adobe Acrobat Do...	2/9/2020 8:12:00 PM
Public Bank Berhad Int...	81 KB	Adobe Acrobat Do...	9/30/2019 5:29:21 PM
Receipt-04042022411...	18 KB	Adobe Acrobat Do...	4/4/2020 10:41:25 PM

Rights | Owner

Session

File protocol: SFTP

Host name: 127.0.0.1

Port number: 2222

User name: root

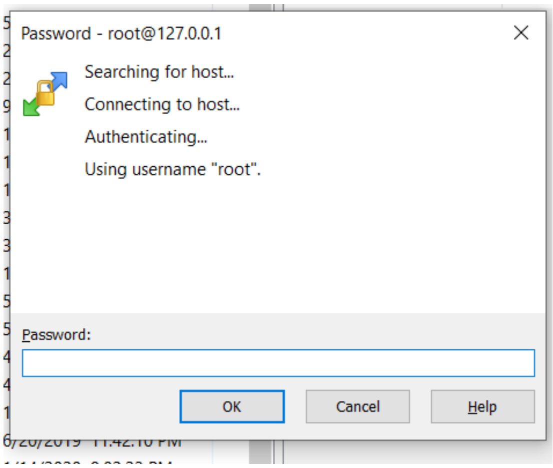
Password:

New Site
Raspberry
raspberrypi.local

☑ Show Login dialog on startup and when the last session is closed

0 B of 11.5 MB in 0 of 26 | 4 hidden | 10 hidden

3. Key in your password of sandbox hadoop.

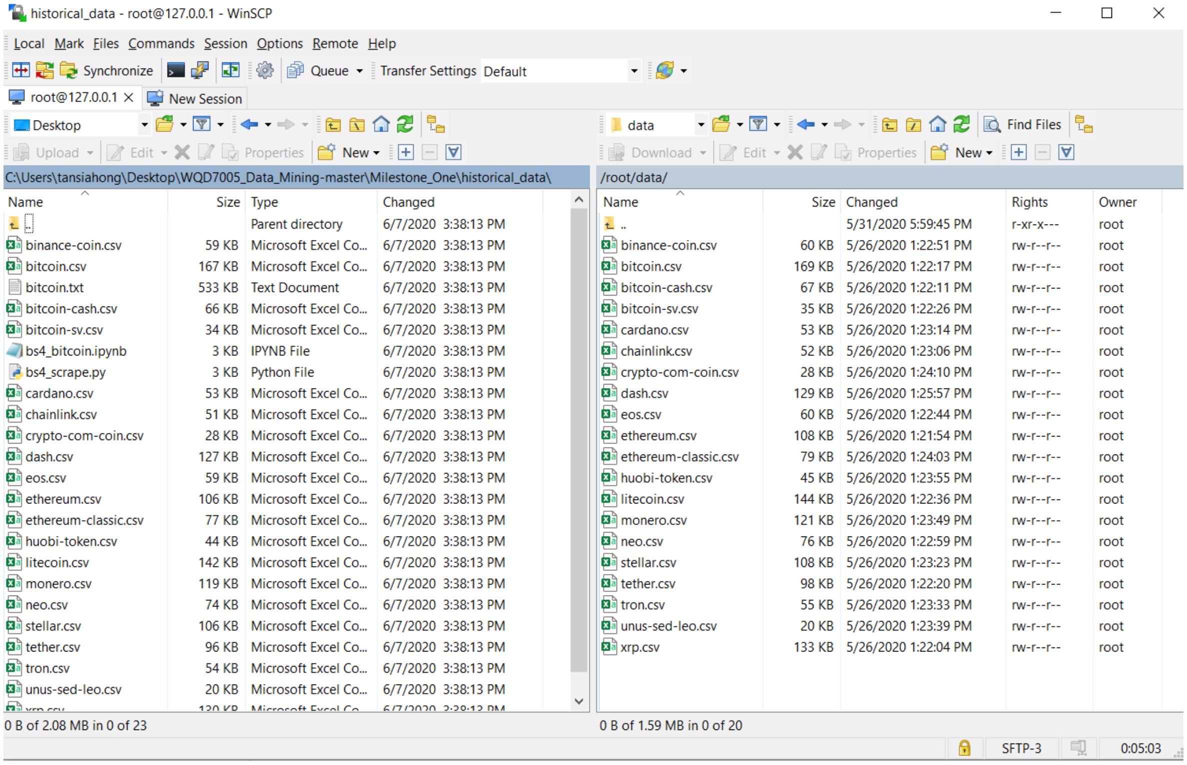


Password - root@127.0.0.1

Searching for host...
Connecting to host...
Authenticating...
Using username "root".

Password:

4. Send local Windows' data into sandbox (VM) by using WinSCP.



C:\Users\tansiahong\Desktop\WQD7005_Data_Mining-master\Milestone_One\historical_data\

Name	Size	Type	Changed
..		Parent directory	6/7/2020 3:38:13 PM
binance-coin.csv	59 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
bitcoin.csv	167 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
bitcoin.txt	533 KB	Text Document	6/7/2020 3:38:13 PM
bitcoin-cash.csv	66 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
bitcoin-sv.csv	34 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
bs4_bitcoin.ipynb	3 KB	IPYNB File	6/7/2020 3:38:13 PM
bs4_scrape.py	3 KB	Python File	6/7/2020 3:38:13 PM
cardano.csv	53 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
chainlink.csv	51 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
crypto-com-coin.csv	28 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
dash.csv	127 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
eos.csv	59 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
ethereum.csv	106 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
ethereum-classic.csv	77 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
huobi-token.csv	44 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
litecoin.csv	142 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
monero.csv	119 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
neo.csv	74 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
stellar.csv	106 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
tether.csv	96 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
tron.csv	54 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
unus-sed-leo.csv	20 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM
xrp.csv	130 KB	Microsoft Excel Co...	6/7/2020 3:38:13 PM

0 B of 2.08 MB in 0 of 23

/root/data/

Name	Size	Changed	Rights	Owner
..		5/31/2020 5:59:45 PM	r-xr-x---	root
binance-coin.csv	60 KB	5/26/2020 1:22:51 PM	rw-r--r--	root
bitcoin.csv	169 KB	5/26/2020 1:22:17 PM	rw-r--r--	root
bitcoin-cash.csv	67 KB	5/26/2020 1:22:11 PM	rw-r--r--	root
bitcoin-sv.csv	35 KB	5/26/2020 1:22:26 PM	rw-r--r--	root
cardano.csv	53 KB	5/26/2020 1:23:14 PM	rw-r--r--	root
chainlink.csv	52 KB	5/26/2020 1:23:06 PM	rw-r--r--	root
crypto-com-coin.csv	28 KB	5/26/2020 1:24:10 PM	rw-r--r--	root
dash.csv	129 KB	5/26/2020 1:25:57 PM	rw-r--r--	root
eos.csv	60 KB	5/26/2020 1:22:44 PM	rw-r--r--	root
ethereum.csv	108 KB	5/26/2020 1:21:54 PM	rw-r--r--	root
ethereum-classic.csv	79 KB	5/26/2020 1:24:03 PM	rw-r--r--	root
huobi-token.csv	45 KB	5/26/2020 1:23:55 PM	rw-r--r--	root
litecoin.csv	144 KB	5/26/2020 1:22:36 PM	rw-r--r--	root
monero.csv	121 KB	5/26/2020 1:23:49 PM	rw-r--r--	root
neo.csv	76 KB	5/26/2020 1:22:59 PM	rw-r--r--	root
stellar.csv	108 KB	5/26/2020 1:23:23 PM	rw-r--r--	root
tether.csv	98 KB	5/26/2020 1:22:20 PM	rw-r--r--	root
tron.csv	55 KB	5/26/2020 1:23:33 PM	rw-r--r--	root
unus-sed-leo.csv	20 KB	5/26/2020 1:23:39 PM	rw-r--r--	root
xrp.csv	133 KB	5/26/2020 1:22:04 PM	rw-r--r--	root

0 B of 1.59 MB in 0 of 20

5. Go to web shell client

```
hdfs fs -put /root/data/*.csv /user/root/datamining/data/
```

This is to copy the file from root directory into another directory HDFS.

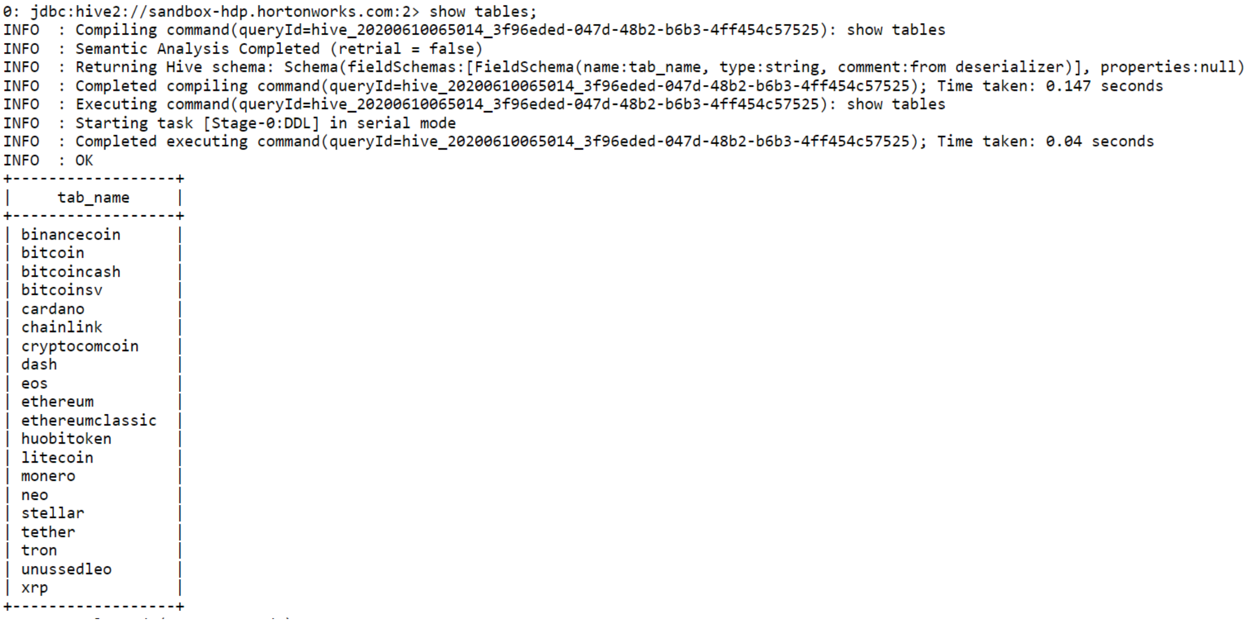
6. Open hive in web shell client, use the database preferred, then create a table under the database.
Eg. Table 'bitcoin'

```
CREATE TABLE bitcoin (MarketDate DATE, Open DOUBLE, High DOUBLE, Low DOUBLE, Close DOUBLE, Volume DOUBLE, MarketCapacity DOUBLE)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

TBLPROPERTIES("skip.header.line.count"="1");
```

Then load the csv data from HDFS into bitcoin table

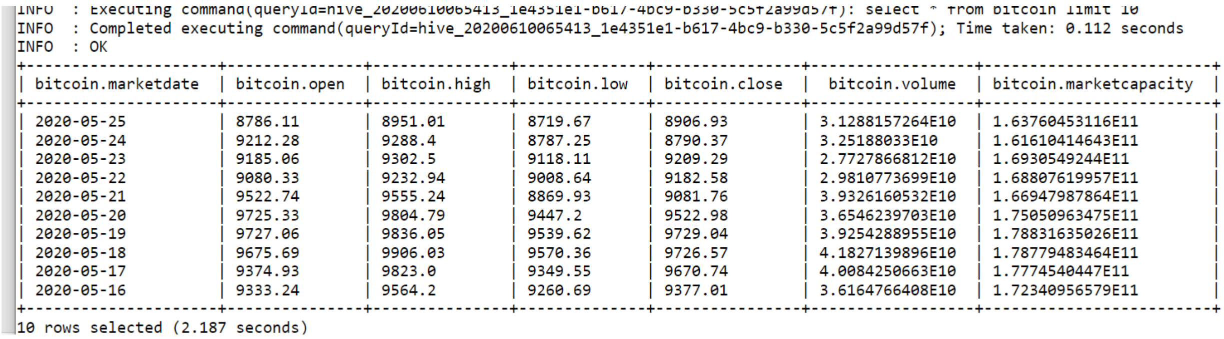
```
LOAD DATA INPATH '/user/root/datamining/data/coin.csv' INTO TABLE bitcoin;
```

7. After load the data into Hive table, you can run query ‘show tables;’ to perform all table as below:



```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> show tables;
INFO  : Compiling command(queryId=hive_20200610065014_3f96eded-047d-48b2-b6b3-4ff454c57525): show tables
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20200610065014_3f96eded-047d-48b2-b6b3-4ff454c57525); Time taken: 0.147 seconds
INFO  : Executing command(queryId=hive_20200610065014_3f96eded-047d-48b2-b6b3-4ff454c57525): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20200610065014_3f96eded-047d-48b2-b6b3-4ff454c57525); Time taken: 0.04 seconds
INFO  : OK
+-------------------+
|     tab_name      |
+-------------------+
| binancecoin       |
| bitcoin           |
| bitcoincash       |
| bitcoinsv         |
| cardano           |
| chainlink         |
| cryptocomcoin     |
| dash              |
| eos               |
| ethereum          |
| ethereumclassic   |
| huobitoken        |
| litecoin          |
| monero            |
| neo               |
| stellar           |
| tether            |
| tron              |
| unussedleo        |
| xrp               |
+-------------------+
```

8. The data show out from Hive table by SQL query ‘Select * from bitcoin limit 10;’



```
INFO  : Executing command(queryId=hive_20200610065413_1e4351e1-b617-4bc9-b330-5c5f2a99d57f): select * from bitcoin limit 10
INFO  : Completed executing command(queryId=hive_20200610065413_1e4351e1-b617-4bc9-b330-5c5f2a99d57f); Time taken: 0.112 seconds
INFO  : OK
+---------------------+----------------+----------------+---------------+-----------------+-------------------+-------------------------+
| bitcoin.marketdate  | bitcoin.open   | bitcoin.high   | bitcoin.low   | bitcoin.close   |  bitcoin.volume   | bitcoin.marketcapacity  |
+---------------------+----------------+----------------+---------------+-----------------+-------------------+-------------------------+
| 2020-05-25          | 8786.11        | 8951.01        | 8719.67       | 8906.93         | 3.1288157264E10   | 1.63760453116E11        |
| 2020-05-24          | 9212.28        | 9288.4         | 8787.25       | 8790.37         | 3.25188033E10     | 1.61610414643E11        |
| 2020-05-23          | 9185.06        | 9302.5         | 9118.11       | 9209.29         | 2.7727866812E10   | 1.6930549244E11         |
| 2020-05-22          | 9080.33        | 9232.94        | 9008.64       | 9182.58         | 2.9810773699E10   | 1.68807619957E11        |
| 2020-05-21          | 9522.74        | 9555.24        | 8869.93       | 9081.76         | 3.9326160532E10   | 1.66947987864E11        |
| 2020-05-20          | 9725.33        | 9804.79        | 9447.2        | 9522.98         | 3.6546239703E10   | 1.75050963475E11        |
| 2020-05-19          | 9727.06        | 9836.05        | 9539.62       | 9729.04         | 3.9254288955E10   | 1.78831635026E11        |
| 2020-05-18          | 9675.69        | 9906.03        | 9570.36       | 9726.57         | 4.1827139896E10   | 1.78779483464E11        |
| 2020-05-17          | 9374.93        | 9823.0         | 9349.55       | 9670.74         | 4.0084250663E10   | 1.7774540447E11         |
| 2020-05-16          | 9333.24        | 9564.2         | 9260.69       | 9377.01         | 3.6164766408E10   | 1.72340956579E11        |
+---------------------+----------------+----------------+---------------+-----------------+-------------------+-------------------------+
10 rows selected (2.187 seconds)
```