Spark Movie Query Application Tutorial

Student Name	Essa Umar Khan
University	Aston University
Module Code	CS3800
Module Name	Advanced Data Base Systems
Student Number	170077653
Contact Email	khane2@aston.ac.uk

Contents

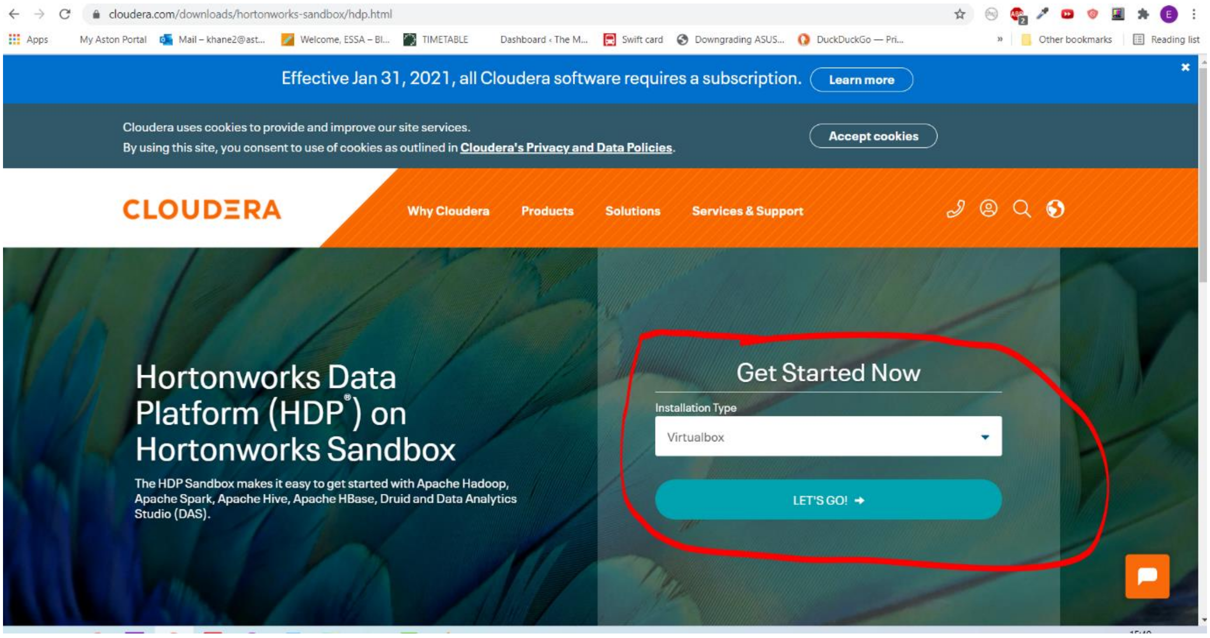
1. Pre-requisites

1) Computer running Windows 10 with 16 GB ram
2) Internet connection
3) Oracle Virtual Box (application was tested on v6.1)
4) Download the entire project source code files from either:
    a. The GitHub repository
    b. The .ZIP submission contains the files from the GitHub repository.

2. Hortonworks Sandbox 2.6.5

2.1 Download

1) Head over to https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html [1] and select 'VirtualBox' as the instillation type.



cloudera.com/downloads/hortonworks-sandbox/hdp.html

Effective Jan 31, 2021, all Cloudera software requires a subscription.

Cloudera uses cookies to provide and improve our site services.
By using this site, you consent to use of cookies as outlined in Cloudera's Privacy and Data Policies.

Hortonworks Data Platform (HDP®) on Hortonworks Sandbox

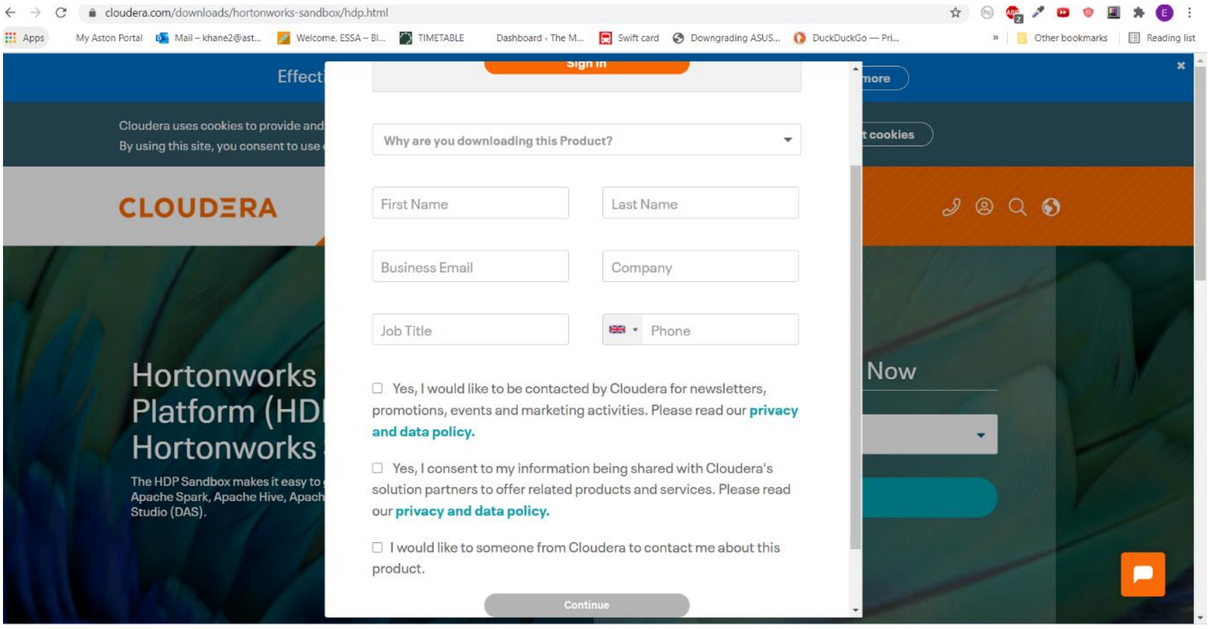
The HDP Sandbox makes it easy to get started with Apache Hadoop, Apache Spark, Apache Hive, Apache HBase, Druid and Data Analytics Studio (DAS).

Get Started Now

Installation Type: Virtualbox

[1] Alternatively, whilst logged into an Aston outlook email account you can download the sandbox from here: https://liveastonac-my.sharepoint.com/:u:/g/personal/khane2_aston_ac_uk/EXMFaW8Clw5LpZ0mMHQN_6YBavbLoT92V4ScCMpU-x4kCg?e=z9bRdH

2) A pop-up will appear and to access the download link you have to complete the form on this pop-up, once you have filled in the information accept all the policies and press submit.



Why are you downloading this Product?

First Name | Last Name

Business Email | Company

Job Title | Phone

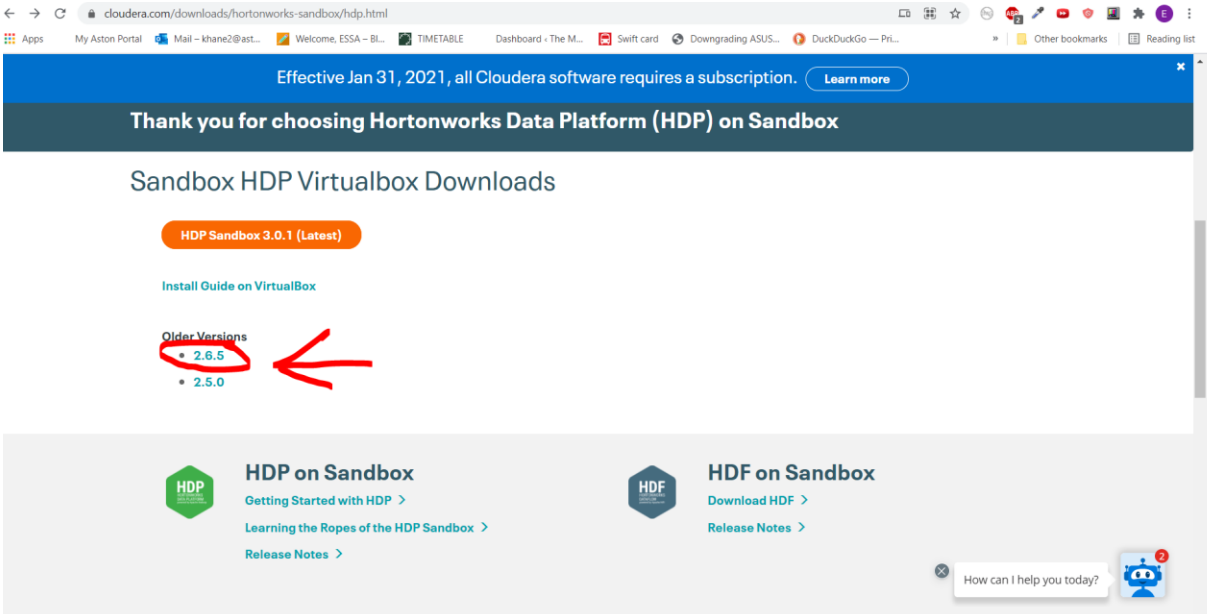
☐ Yes, I would like to be contacted by Cloudera for newsletters, promotions, events and marketing activities. Please read our **privacy and data policy.**

☐ Yes, I consent to my information being shared with Cloudera's solution partners to offer related products and services. Please read our **privacy and data policy.**

☐ I would like to someone from Cloudera to contact me about this product.

Continue

3) Click on 2.6.5 link to start download. It is a very large .ova file so it will take some time to download.



Effective Jan 31, 2021, all Cloudera software requires a subscription.

Thank you for choosing Hortonworks Data Platform (HDP) on Sandbox

Sandbox HDP Virtualbox Downloads

HDP Sandbox 3.0.1 (Latest)

Install Guide on VirtualBox

Older Versions

- 2.6.5
- 2.5.0

HDP on Sandbox

Getting Started with HDP >

Learning the Ropes of the HDP Sandbox >

Release Notes >

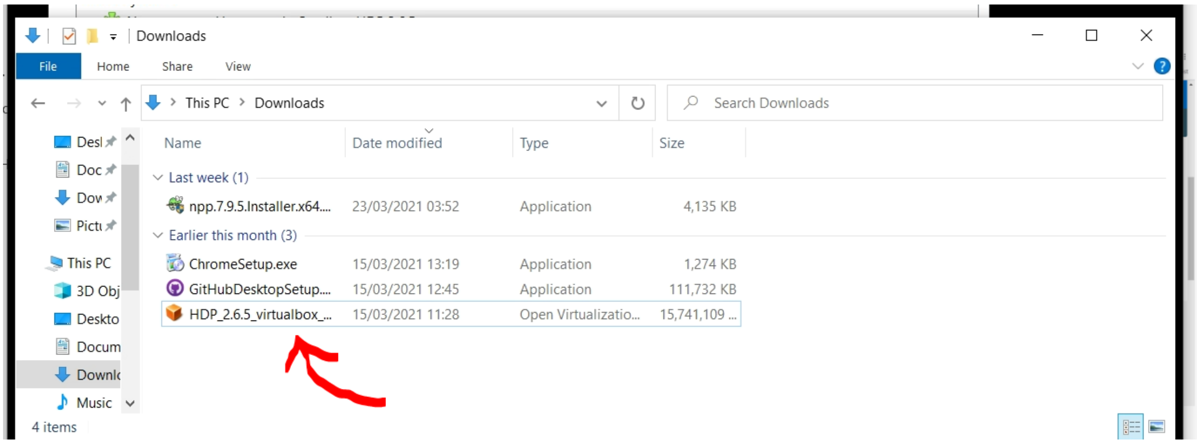
HDF on Sandbox

Download HDF >

Release Notes >

How can I help you today?

4) Once the download is completed, double click the downloaded .ova file:

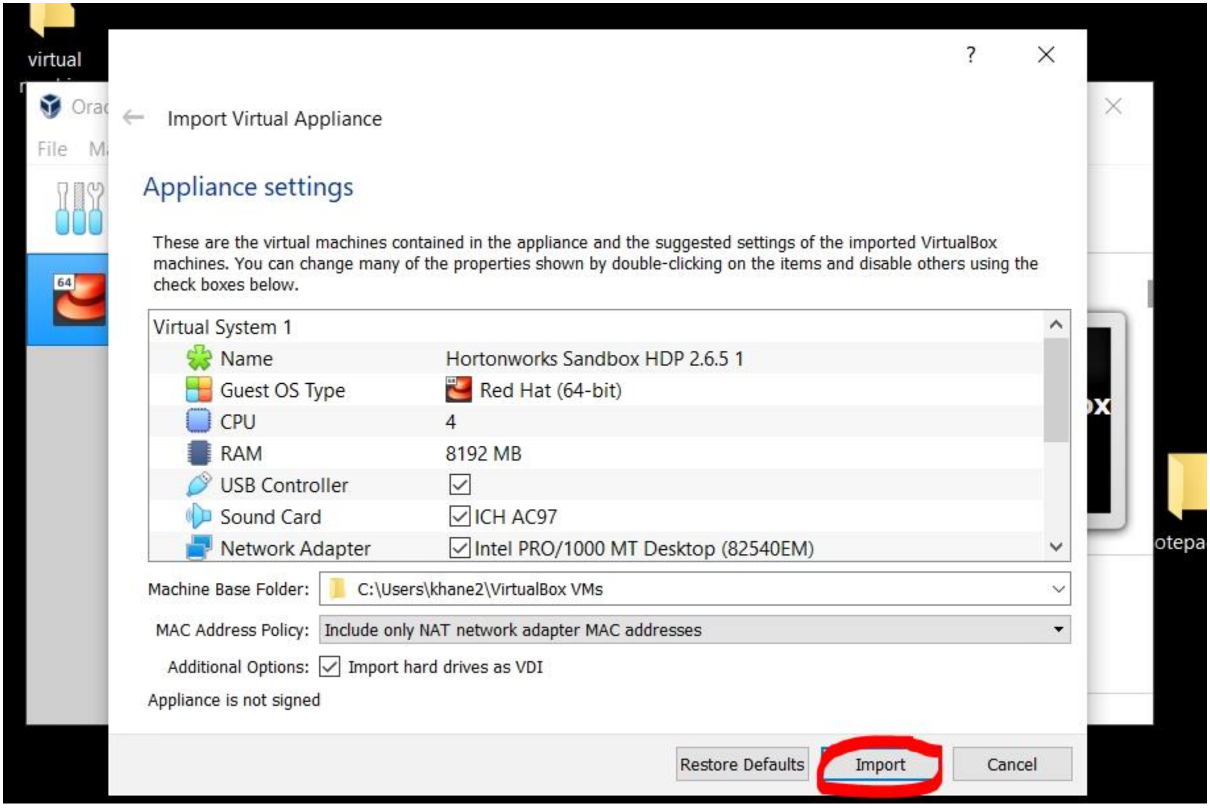


Downloads

Name	Date modified	Type	Size
Last week (1)			
npp.7.9.5.Installer.x64....	23/03/2021 03:52	Application	4,135 KB
Earlier this month (3)			
ChromeSetup.exe	15/03/2021 13:19	Application	1,274 KB
GitHubDesktopSetup....	15/03/2021 12:45	Application	111,732 KB
HDP_2.6.5_virtualbox_...	15/03/2021 11:28	Open Virtualizatio...	15,741,109 ...

4 items

5) *‘Oracle Virtual Box’* software will open with a pop-up like this, just click the import button with the default settings.



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 2.6.5 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)

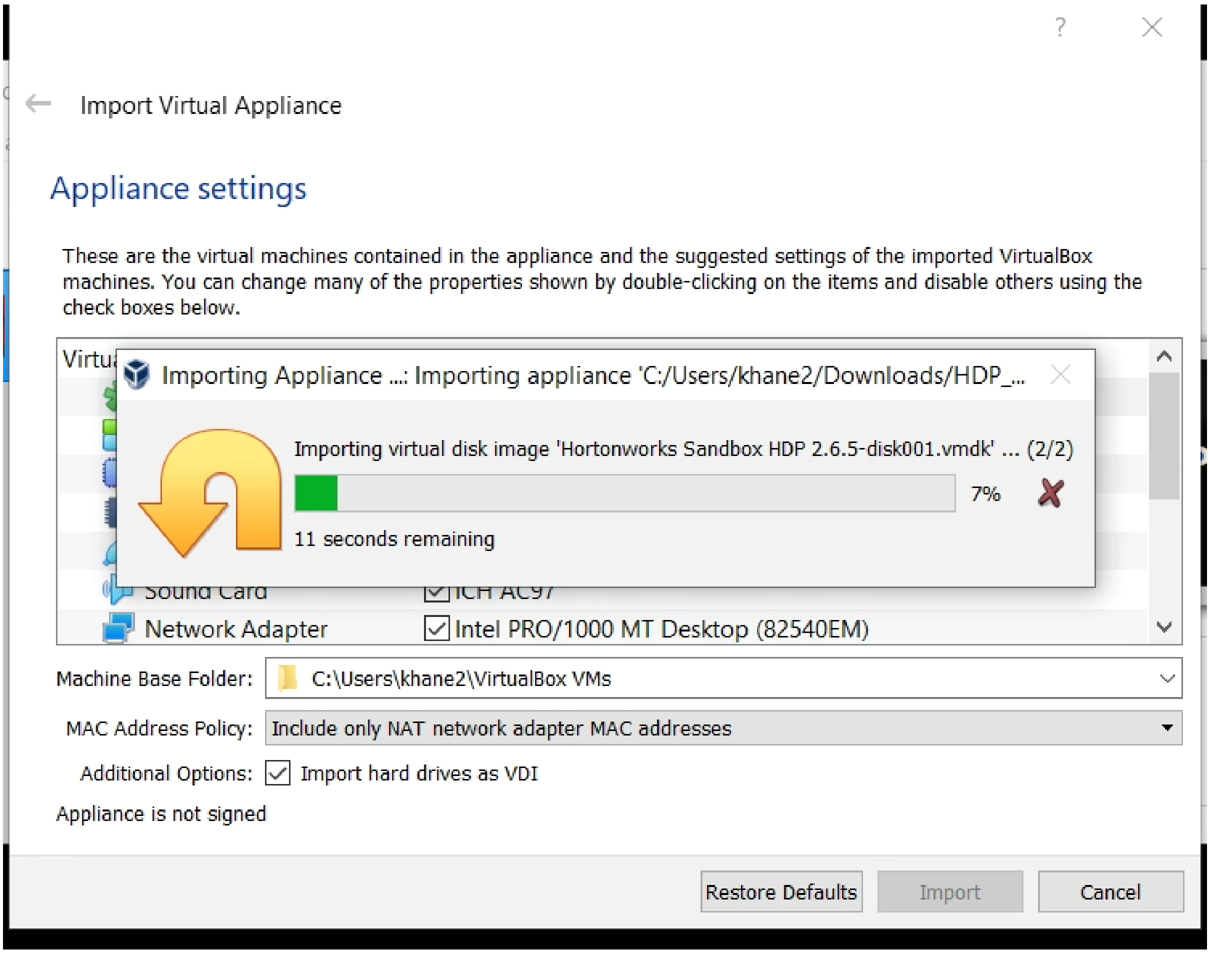
Machine Base Folder: C:\Users\khane2\VirtualBox VMs

MAC Address Policy: Include only NAT network adapter MAC addresses

Additional Options: ☑ Import hard drives as VDI

Appliance is not signed

Once you have pressed import then the image will be imported into virtual box. This will take some time to be copied over:



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.

Importing Appliance ...: Importing appliance 'C:/Users/khane2/Downloads/HDP_...

Importing virtual disk image 'Hortonworks Sandbox HDP 2.6.5-disk001.vmdk' ... (2/2)

7%

11 seconds remaining

Sound Card ☑ ICH AC97

Network Adapter ☑ Intel PRO/1000 MT Desktop (82540EM)

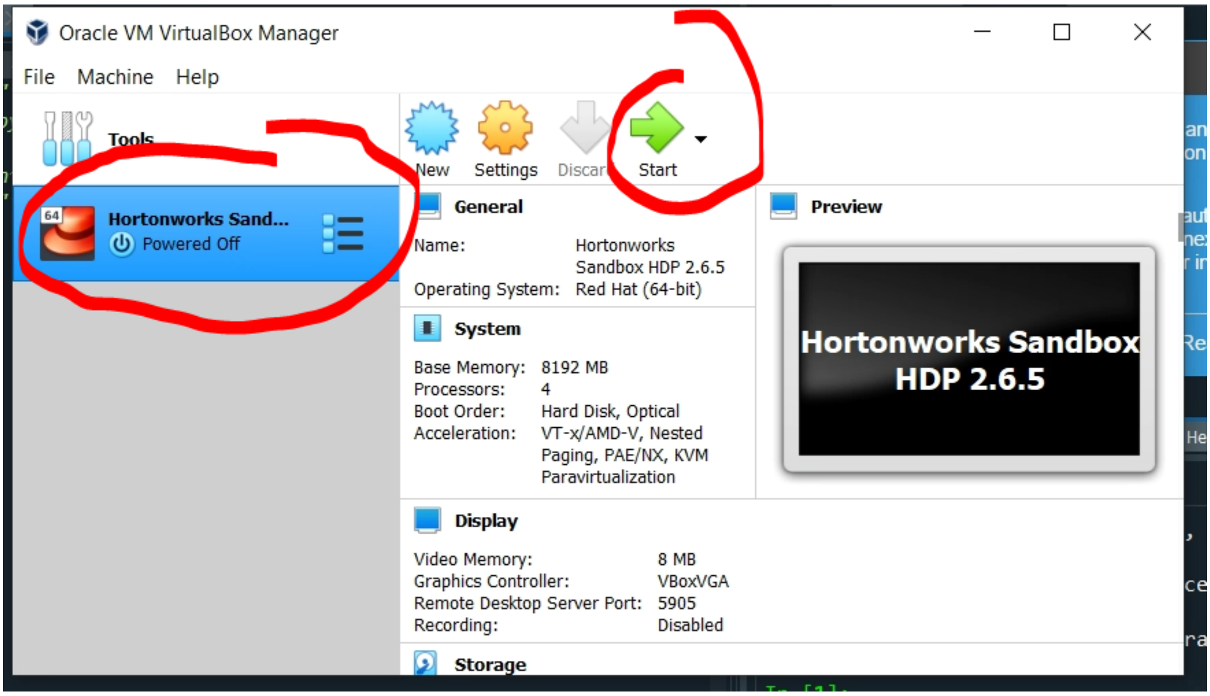
Machine Base Folder: C:\Users\khane2\VirtualBox VMs

MAC Address Policy: Include only NAT network adapter MAC addresses

Additional Options: ☑ Import hard drives as VDI

Appliance is not signed

6) Once the import has completed, select the image and press run virtual box's software main window:



Oracle VM VirtualBox Manager

Hortonworks Sand...
Powered Off

General

Name: Hortonworks Sandbox HDP 2.6.5
Operating System: Red Hat (64-bit)

System

Base Memory: 8192 MB
Processors: 4
Boot Order: Hard Disk, Optical
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX, KVM Paravirtualization

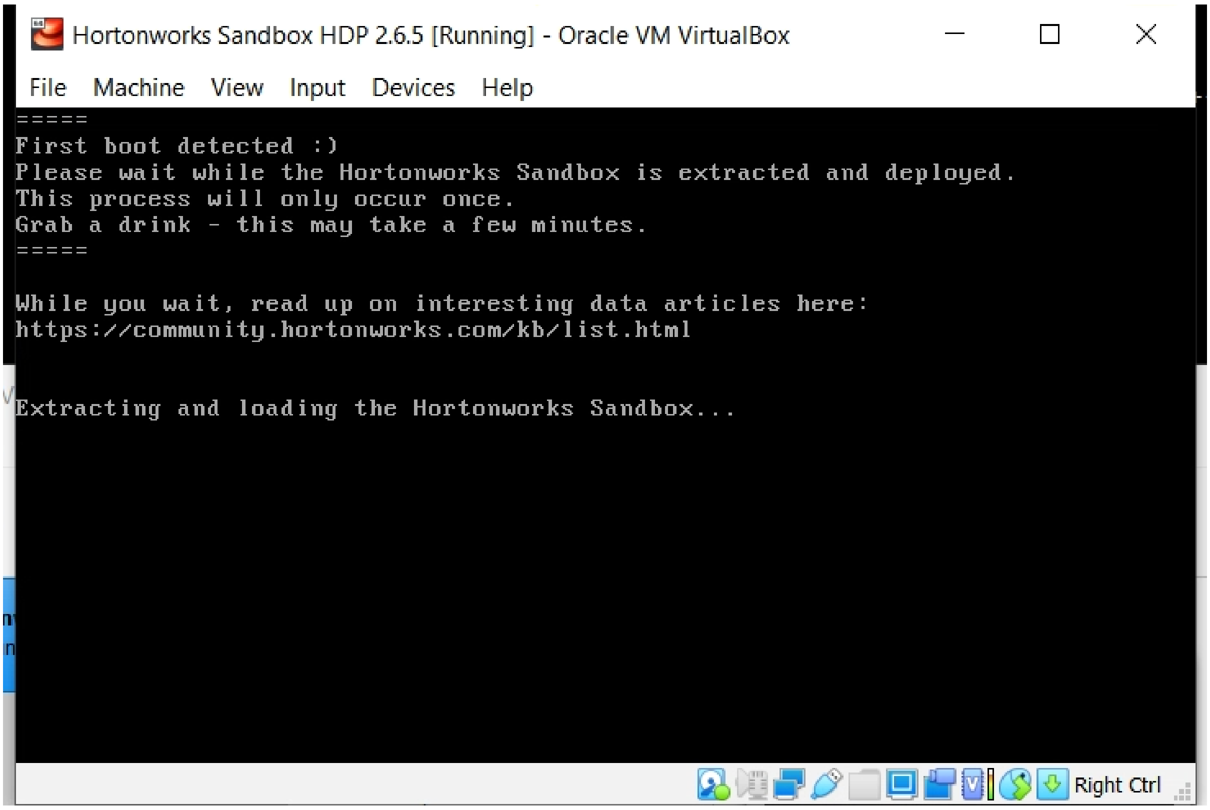
Display

Video Memory: 8 MB
Graphics Controller: VBoxVGA
Remote Desktop Server Port: 5905
Recording: Disabled

Storage

Preview

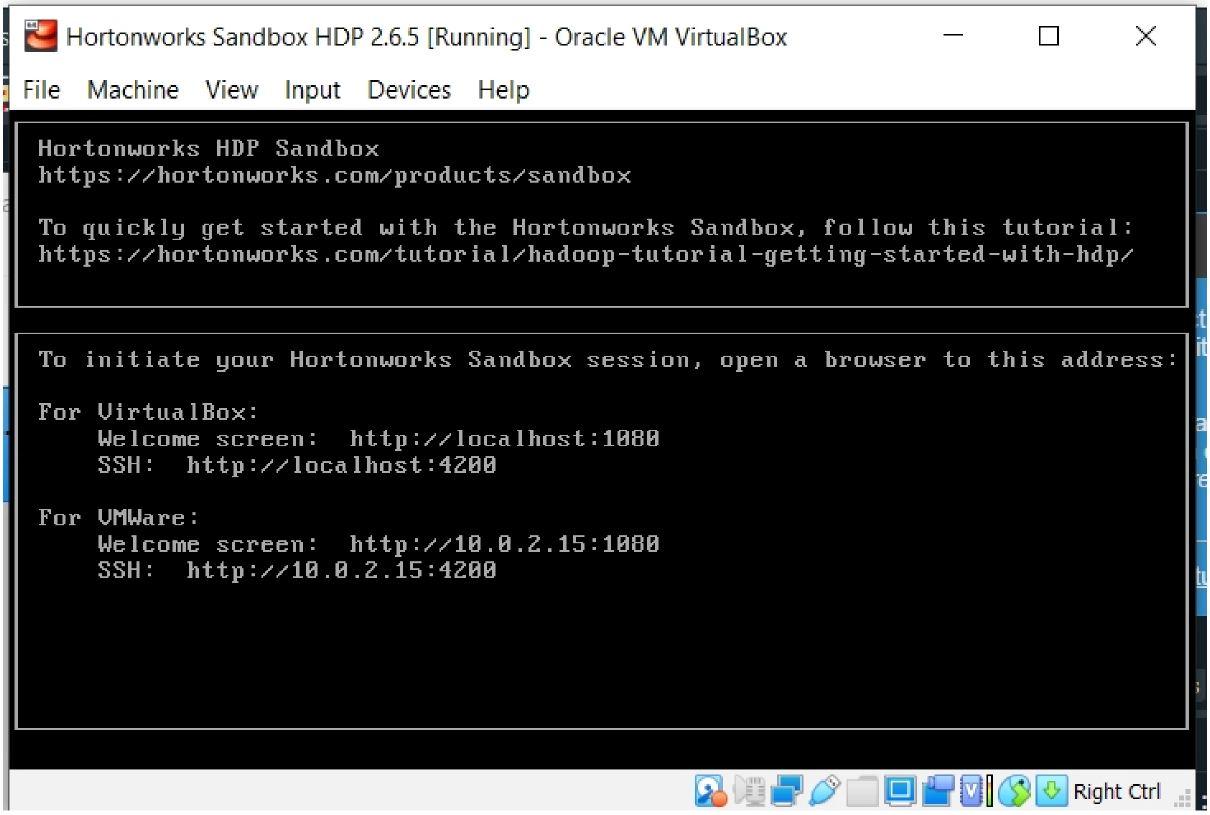
7) There is a ‘first time launch set-up’ boot up process that will now occur, let this complete and do not close the window. You don’t have to do anything but wait. This usually takes around 15 minutes.



Hortonworks Sandbox HDP 2.6.5 [Running] - Oracle VM VirtualBox
File
Machine
View
Input
Devices
Help
Right Ctrl

```
=====
First boot detected :)
Please wait while the Hortonworks Sandbox is extracted and deployed.
This process will only occur once.
Grab a drink - this may take a few minutes.
=====

While you wait, read up on interesting data articles here:
https://community.hortonworks.com/kb/list.html


Extracting and loading the Hortonworks Sandbox...
```

8) Once the set-up in step 7 has completed you will see a screen like the one below.



Hortonworks Sandbox HDP 2.6.5 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Right Ctrl

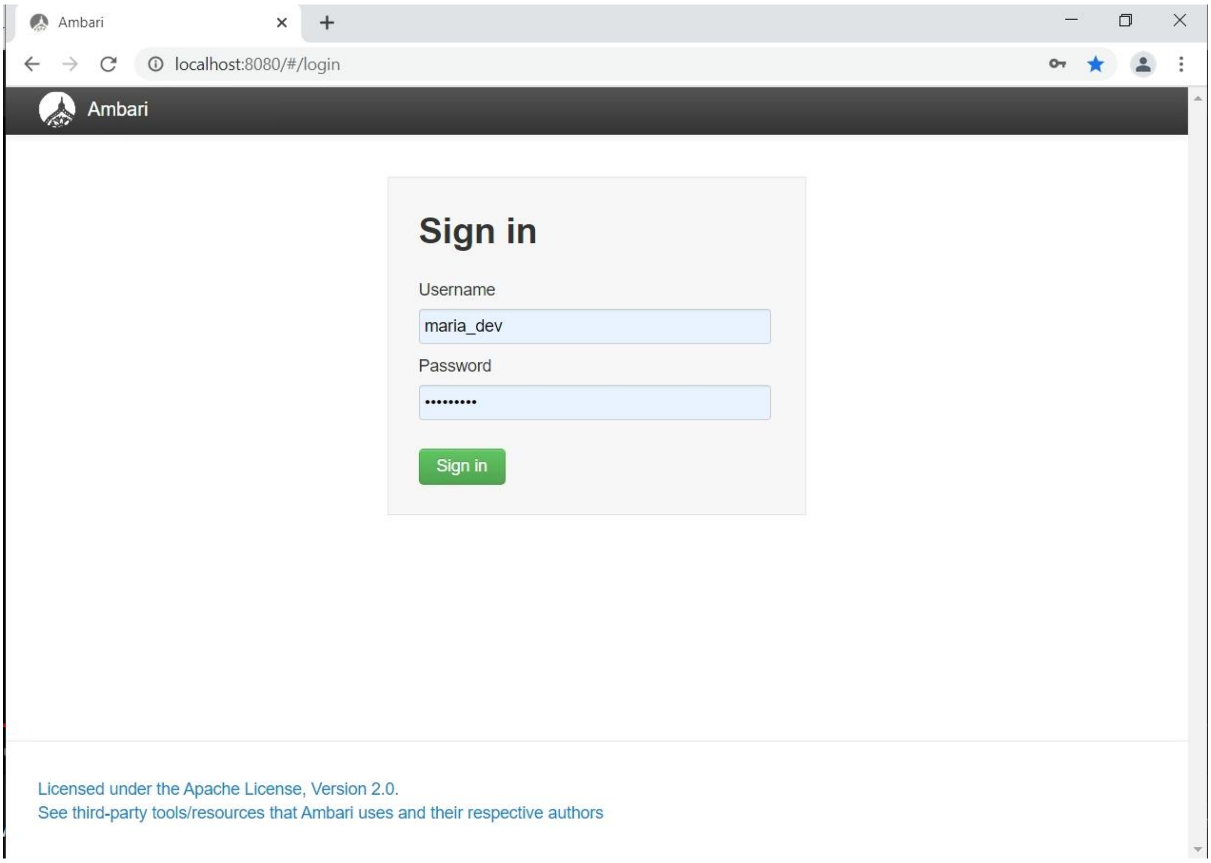
```
Hortonworks HDP Sandbox
https://hortonworks.com/products/sandbox

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
```

As soon as this screen shows all the technologies are loading to be (Hadoop, spark etc.) ready for use – even though it does not indicate this on the above screen. Until this process is completed do not proceed to section 2.3. This process usually takes 15 minutes max to complete, so you can wait for 15 minutes and then proceed to the next section of this tutorial. Alternatively, as this process can finish a lot faster than 15 minutes it is recommended that you visually track the progress of this process please see the forthcoming optional section 2.2: *'Optional: How to visually track apache Ambari services start-up progress.'*

2.2 Optional: How to visually track Apache Ambari services start-up progress.

You have been directed to this optional sub-tutorial, and follows from section 2.1 step 8 of this document.

1) Open an internet browser (chrome is recommended) and go to this link: http://localhost:8080/#/login and login with the credentials:
    a. Username: maria_dev
    b. Password: maria_dev



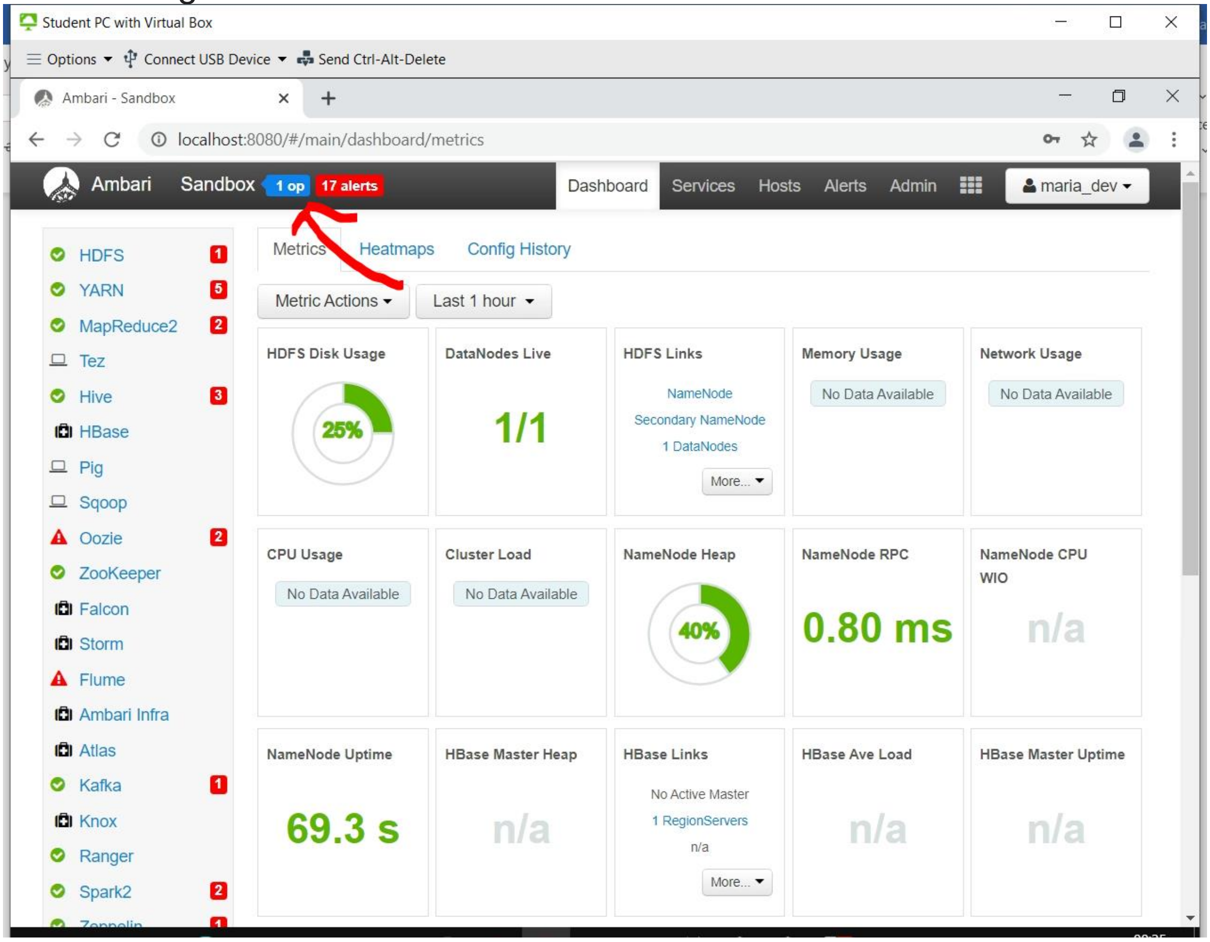
Sign in

Username: maria_dev

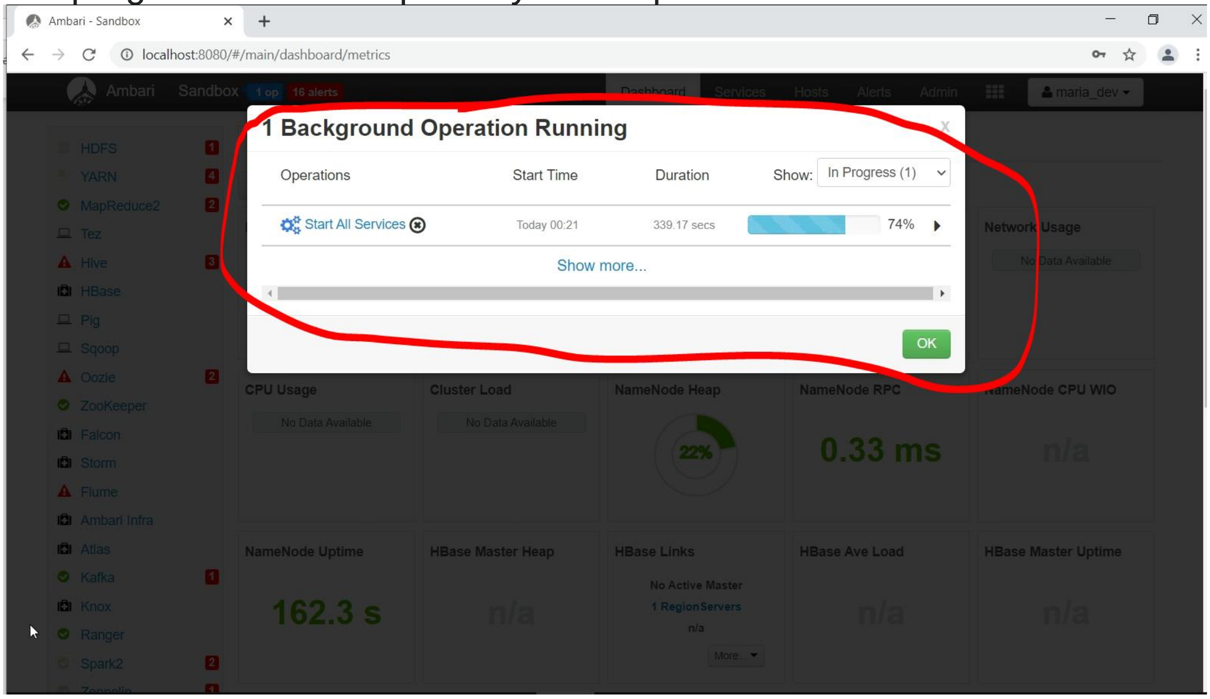
Password: •••••••••

Licensed under the Apache License, Version 2.0.
See third-party tools/resources that Ambari uses and their respective authors

2) Next, you will see something similar to the screen-shot below, click the blue tag:



3) A popup similar as shown in the screenshot below will open. Once, this progress bar is completed you can proceed from section 2.3.



1 Background Operation Running

Operations	Start Time	Duration	Show: In Progress (1)
Start All Services	Today 00:21	339.17 secs	74%

Show more...

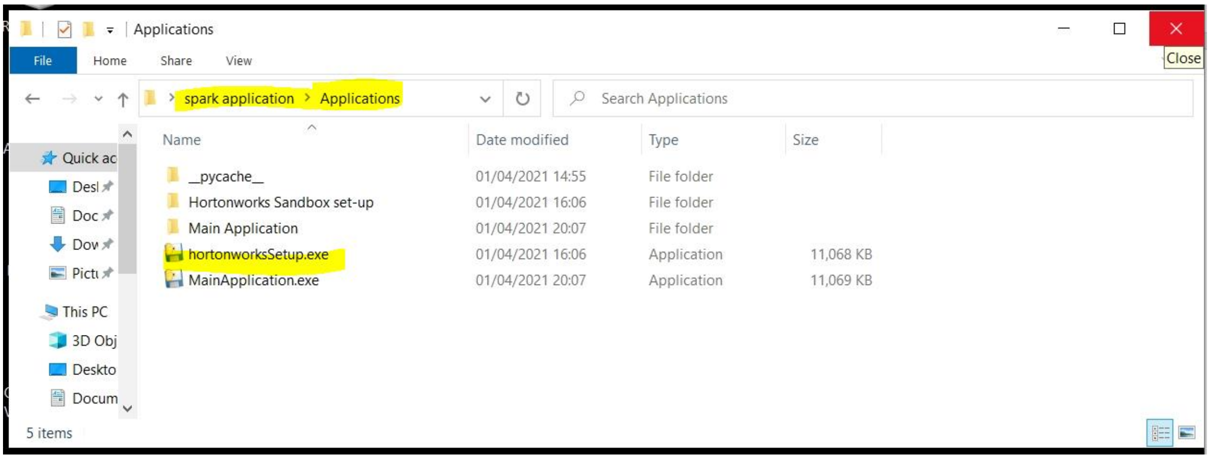
OK

2.3 Set-up

The tutorial from this point onwards can also be followed by watching this youtube video that gives a screen recorded demonstration: https://www.youtube.com/watch?v=Dm_yH_McmbI&ab_channel=EssaKhan

Alternatively you may follow these instructions:

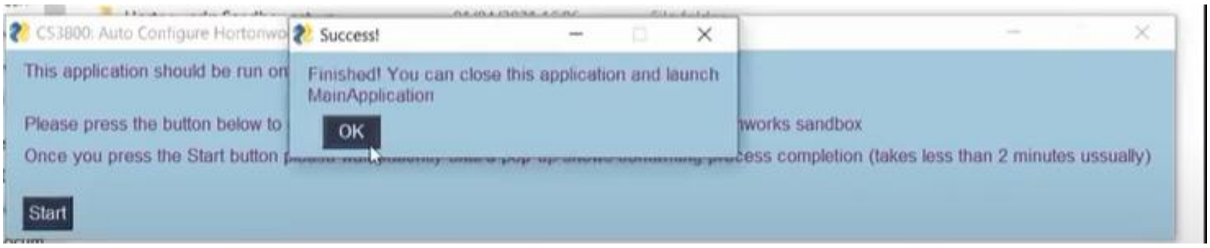
1) Open source-code folder location ‘spark-application/applications’ and launch the application ‘hortonworksSetup.exe’



Name	Date modified	Type	Size
__pycache__	01/04/2021 14:55	File folder	
Hortonworks Sandbox set-up	01/04/2021 16:06	File folder	
Main Application	01/04/2021 20:07	File folder	
hortonworksSetup.exe	01/04/2021 16:06	Application	11,068 KB
MainApplication.exe	01/04/2021 20:07	Application	11,069 KB

2) On the GUI that opens up, follow the on-screen instructions. You must press the ‘start’ button once and let the process complete. Note: hortonworksSetup.exe must only only run once per each Hortonworks sandbox virtual machine image else you will get errors.

   Once process is complete, exit/close this application.



This application should be run on

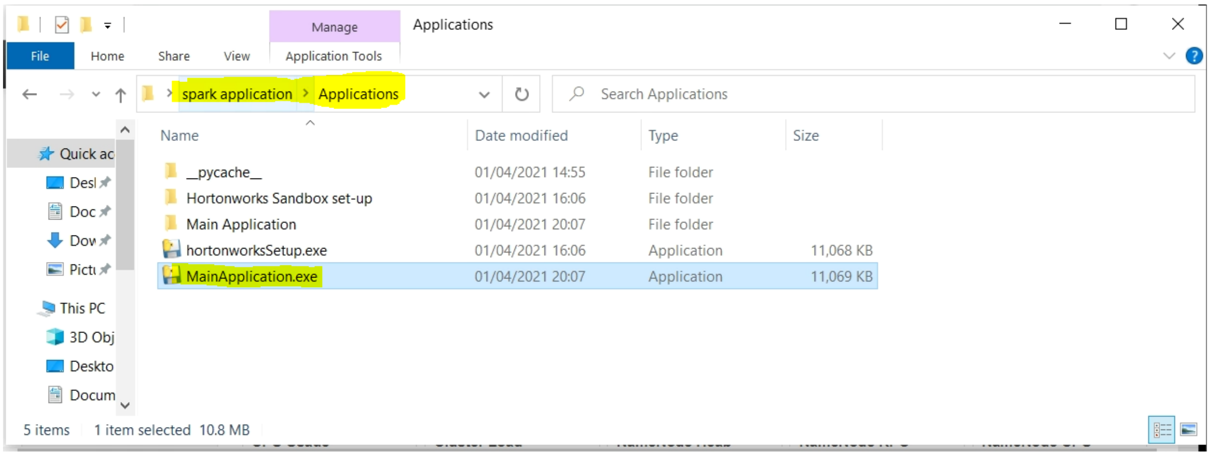
Please press the button below to … works sandbox

Once you press the Start button … ess completion (takes less than 2 minutes usually)

Finished! You can close this application and launch MainApplication

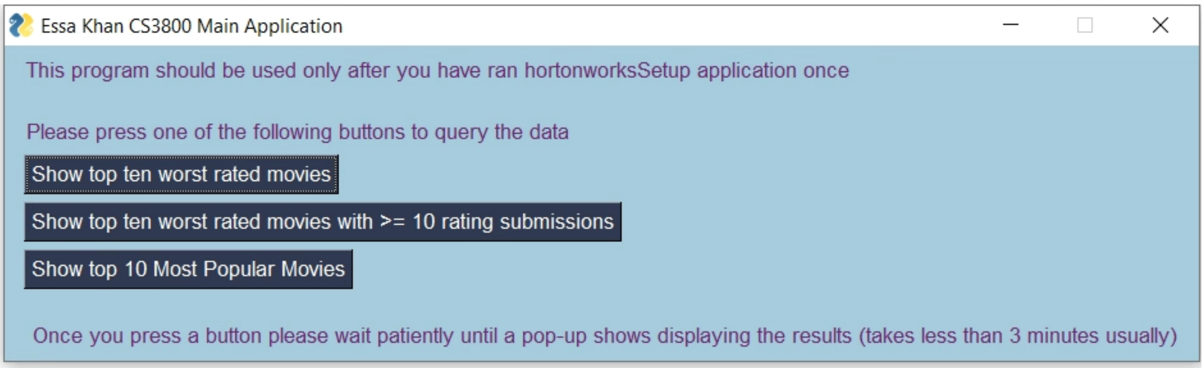
3. Running the Main Application

1) Open source-code folder location ‘spark-application/applications’ and launch the application ‘MainApplication.exe’



Name	Date modified	Type	Size
__pycache__	01/04/2021 14:55	File folder	
Hortonworks Sandbox set-up	01/04/2021 16:06	File folder	
Main Application	01/04/2021 20:07	File folder	
hortonworksSetup.exe	01/04/2021 16:06	Application	11,068 KB
MainApplication.exe	01/04/2021 20:07	Application	11,069 KB

2) A GUI like the one below will open, follow the onscreen instructions to see how to operate the application.


Essa Khan CS3800 Main Application
Show top ten worst rated movies
Show top ten worst rated movies with >= 10 rating submissions
Show top 10 Most Popular Movies

This program should be used only after you have ran hortonworksSetup application once

Please press one of the following buttons to query the data

Once you press a button please wait patiently until a pop-up shows displaying the results (takes less than 3 minutes usually)

You may press each button one by one to query the underlying dataset. The results are shown in a pop-up window. Once one request of results by a button press are output to you then you may proceed to press another button.

4. Optional: Run python .py scripts instead of .exe

This may be helpful alternative if the .exe do not work for some reason. After opening the GUIs this way please follow the respective aforementioned GUIs walkthrough in the previous section.

This has been tested on python 3.7.6

You need to install the relevant packages:

```
pip install pysimplegui
```

```
pip install spur
```

1) To open the GUIs through their scripts directly instead of the provided .exe files:
    a. run the script ‘spark-application\Applications\Hortonworks Sandbox set-up\hortonworksSetup.py’
    b. run the script ‘spark-application\Applications\Main Application\MainApplication.py’