Introduction to Large Scale Machine Learning using Scala/Apache Spark MLLIB

Environment SetupInstructions

Step1 - VirtualBox Installation

VirtualBox Installation

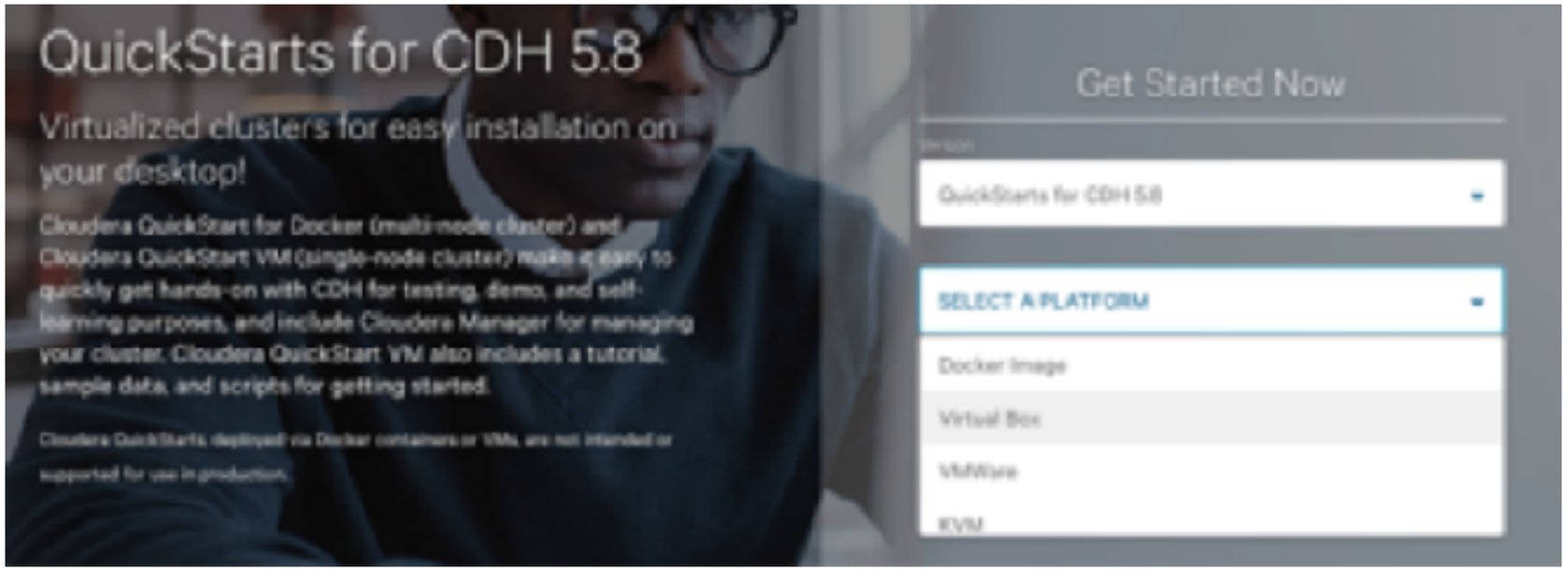
First Download Oracle Virtualbox and install from the below site for your laptop Operating system (Windows / Macbook / any linux laptop)

https://www.virtualbox.org/wiki/Downloads

Step2 - Cloudera QuickStart Virtual machine

Download & Extract the Cloudera QuickStart Virtual Machine

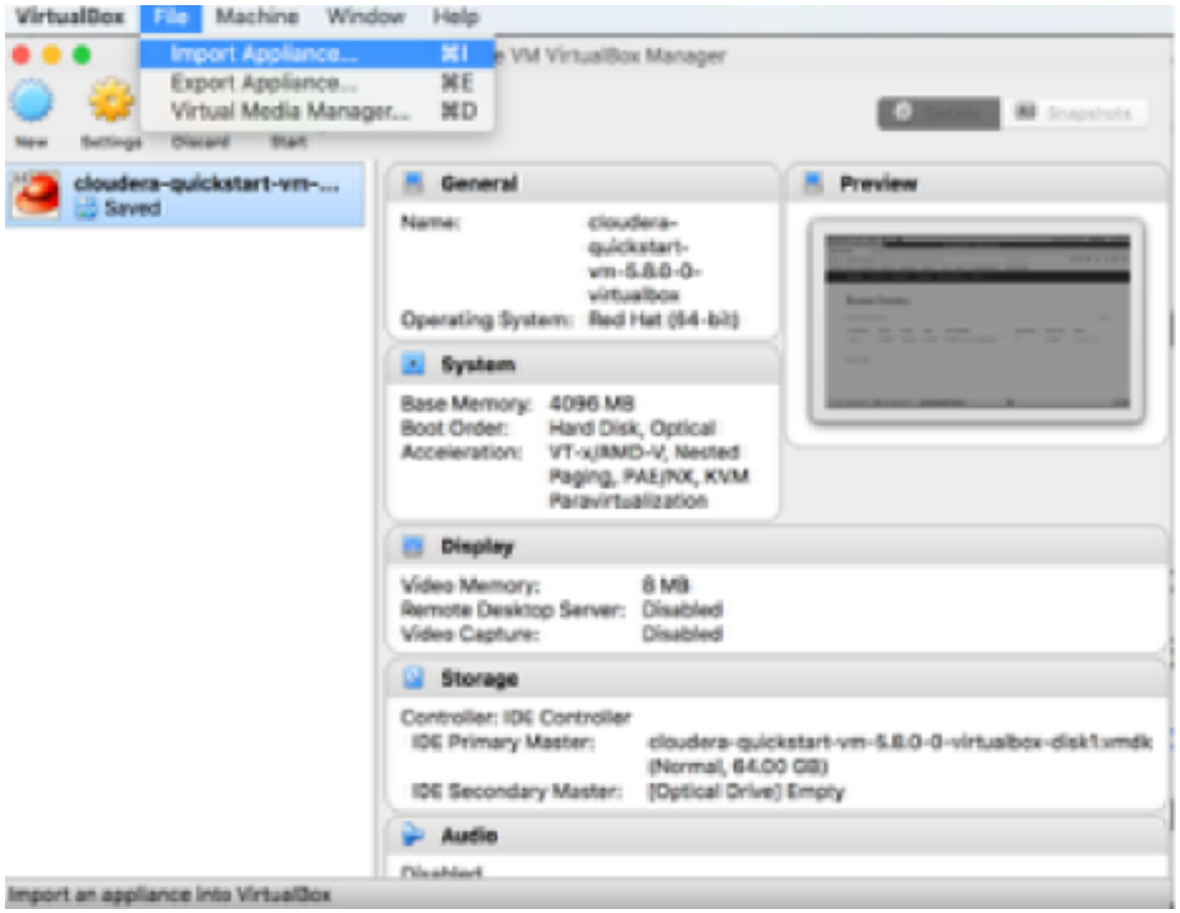
https://www.cloudera.com/downloads/quickstart_vms/5-8.html

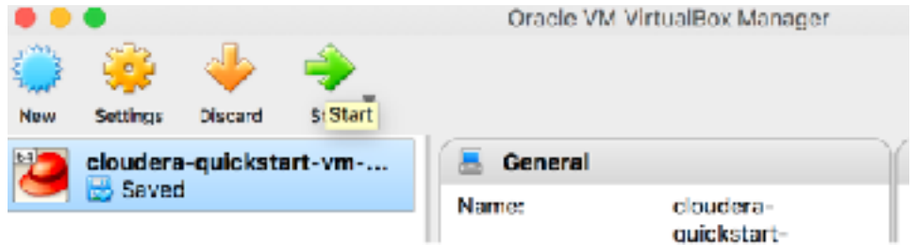


Step3 - Import the Cloudera VM into VirtualBox

Open the Virtualbox software and use “File > import Appliance” and import the downloaded & Extracted Cloudera VM

Next Start the VM clicking green arrow (start) . It takes couple of minutes to get to see the Cloudera Desktop





Step 4 - Open the Eclipse and install Scala 2.10

- On Cloudera Desktop , click the “Eclipse” icon and open the Eclipse java editor.

- use “help > eclipse market place” and search for “Scala IDE” and install

- If asks, please reboot of eclipse ,

Step 5 - Install Apache Kafka & Apache Casandra

- **Kafka installation**
  - open the Unix Terminal by clicking “Terminal” icon on the top bar
  - Create a folder using the unix command “mkdir UAITSummit”
  - Change to the “UAITSummit” directory using the command “cd UAITSummit”
  - wget http://apache.claz.org/kafka/0.10.2.1/kafka_2.10-0.10.2.1.tgz
  - tar -xzf kafka_2.10-0.10.2.1.tgz
- **Cassandra Installation**
  - wget https://archive.apache.org/dist/cassandra/2.2.1/apache-cassandra-2.2.1-bin.tar.gz
  - tar -xvf apache-cassandra-2.2.1-bin.tar.gz
- **Download Data**
  - wget https://s3-us-west-2.amazonaws.com/igebra/sparkData/sfpd.csv
  - wget https://s3-us-west-2.amazonaws.com/igebra/sparkData/trip_data_1.csv