



Introduction to Large Scale Machine Learning using Scala/Spark MLLIB

Environment Setup Instructions



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>

The screenshot shows the 'VirtualBox' website's download page. At the top, the word 'VirtualBox' is written in a large, bold, blue font. Below it, the heading 'Download VirtualBox' is underlined. A paragraph of text states: 'Here, you will find links to VirtualBox binaries and its source code.' This is followed by the sub-heading 'VirtualBox binaries'. Another paragraph reads: 'By downloading, you agree to the terms and conditions of the respective license.' Below this is a bulleted list starting with 'VirtualBox 5.1.18 platform packages. The binaries are released under the terms of the GPL version 2.' The list includes four items: 'Windows hosts', 'OS X hosts', 'Linux distributions', and 'Solaris hosts', each preceded by a small blue icon.

VirtualBox

Download VirtualBox

Here, you will find links to VirtualBox binaries and its source code.

VirtualBox binaries

By downloading, you agree to the terms and conditions of the respective license.

- **VirtualBox 5.1.18 platform packages.** The binaries are released under the terms of the GPL version 2.
 - [Windows hosts](#)
 - [OS X hosts](#)
 - [Linux distributions](#)
 - [Solaris hosts](#)



Step2 - Cloudera QuickStart Virtual machine

Download & Extract the Cloudera QuickStart Virtual Machine

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

QuickStarts for CDH 5.8

Virtualized clusters for easy installation on your desktop!

Cloudera QuickStart for Docker (multi-node cluster) and Cloudera QuickStart VM (single-node cluster) make it easy to quickly get hands-on with CDH for testing, demo, and self-learning purposes, and include Cloudera Manager for managing your cluster. Cloudera QuickStart VM also includes a tutorial, sample data, and scripts for getting started.

Cloudera QuickStarts, deployed via Docker containers or VMs, are not intended or supported for use in production.

Get Started Now

Version

QuickStarts for CDH 5.8

SELECT A PLATFORM

Docker Image

Virtual Box

VMWare

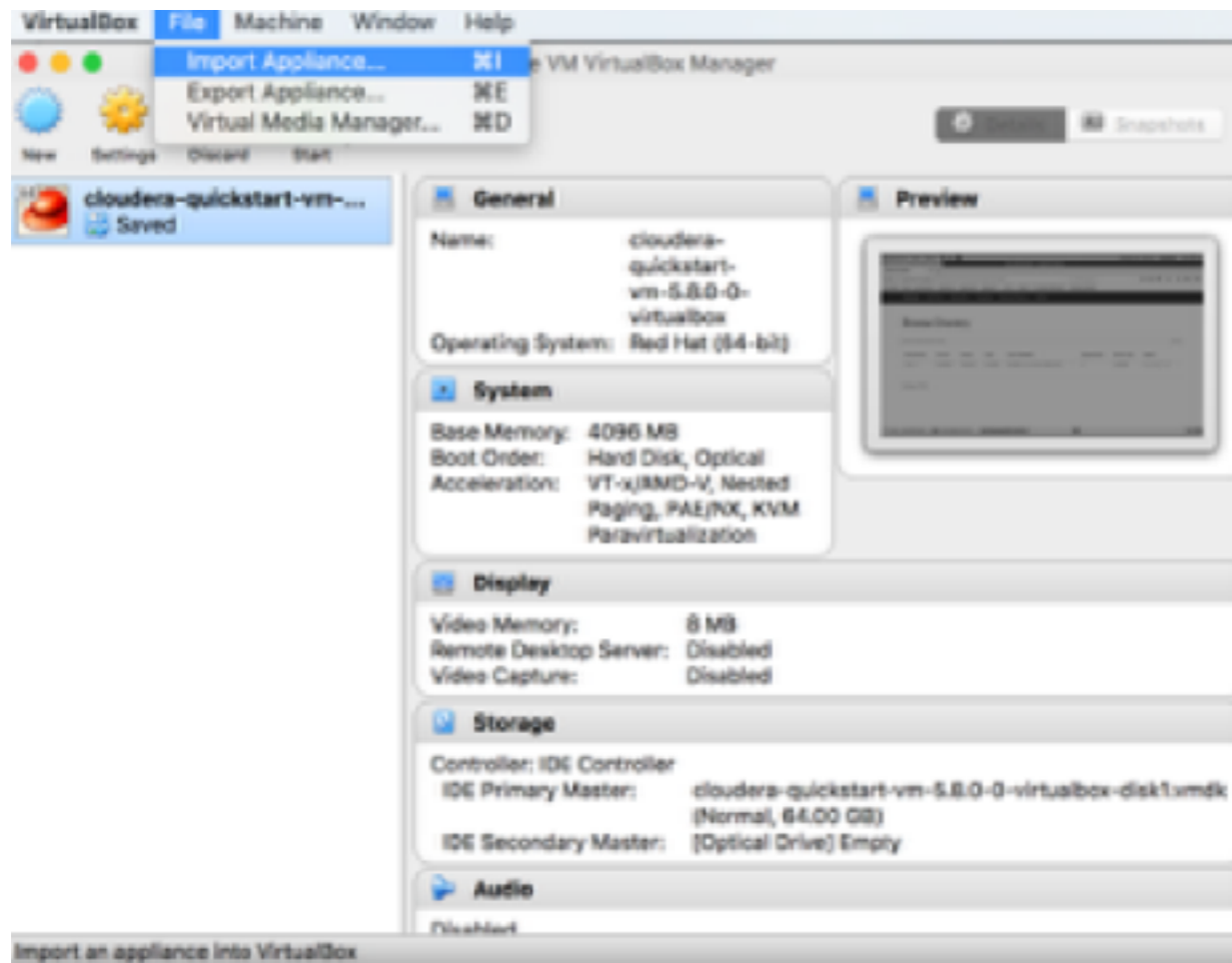
KVM



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 Cassandra

- **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`