Fast Track: Getting Started with the Cloudera QuickStart VM

Xavier Morera

PASSIONATE ABOUT TEACHING

@xmorera www.xaviermorera.com

Getting Started with Big Data

BIG
DATA

Ready to get started

And you need to practice and learn… now

Cloudera QuickStart VM is for you

- Learn Hadoop
- Try new ideas
- Test Big Data jobs
- Demo your application

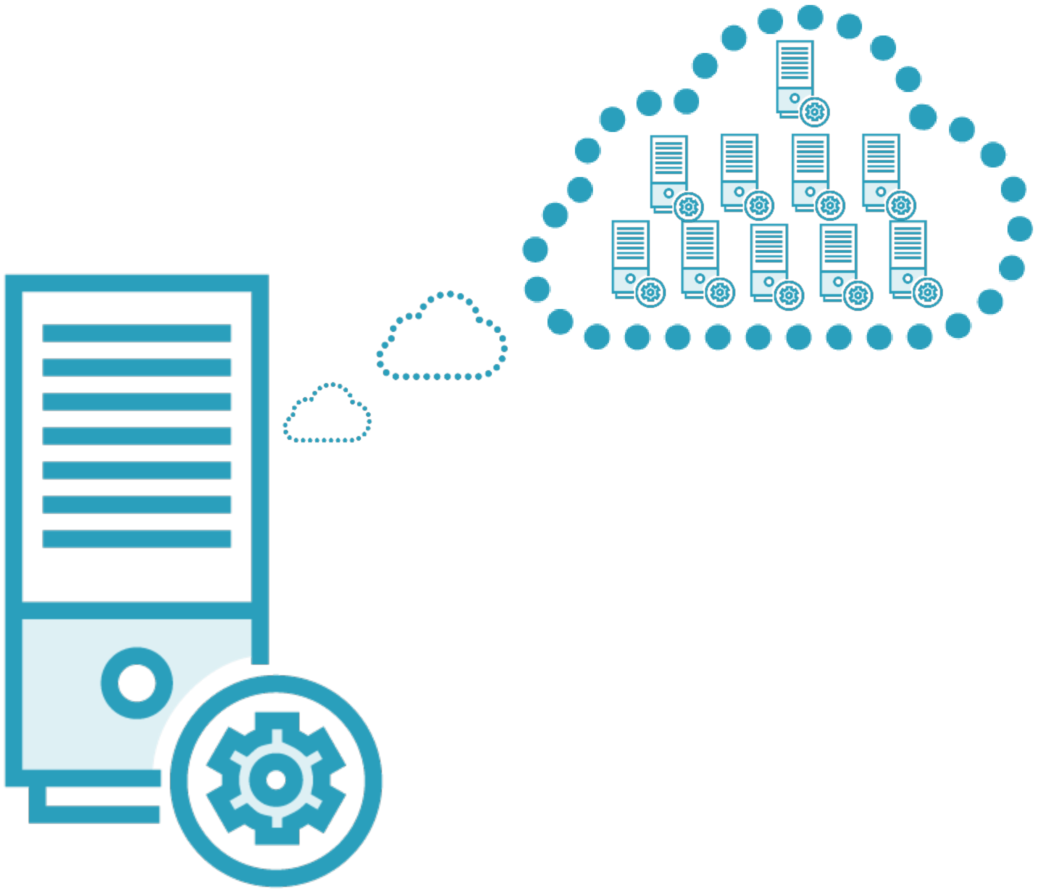
All from a single virtual machine

Pseudo-cluster

What Is the Cloudera QuickStart VM?

Cluster

Pseudo-Cluster



Fast Track to Big Data with Cloudera

Hosting Your Virtual Machine

Docker

KVM

VMWare

Virtual Box
(Oracle)

VirtualBox

My choice for this training

Multiplatform

Solid choice

- Good performance
- Stable

Active Community

- Possible to contribute

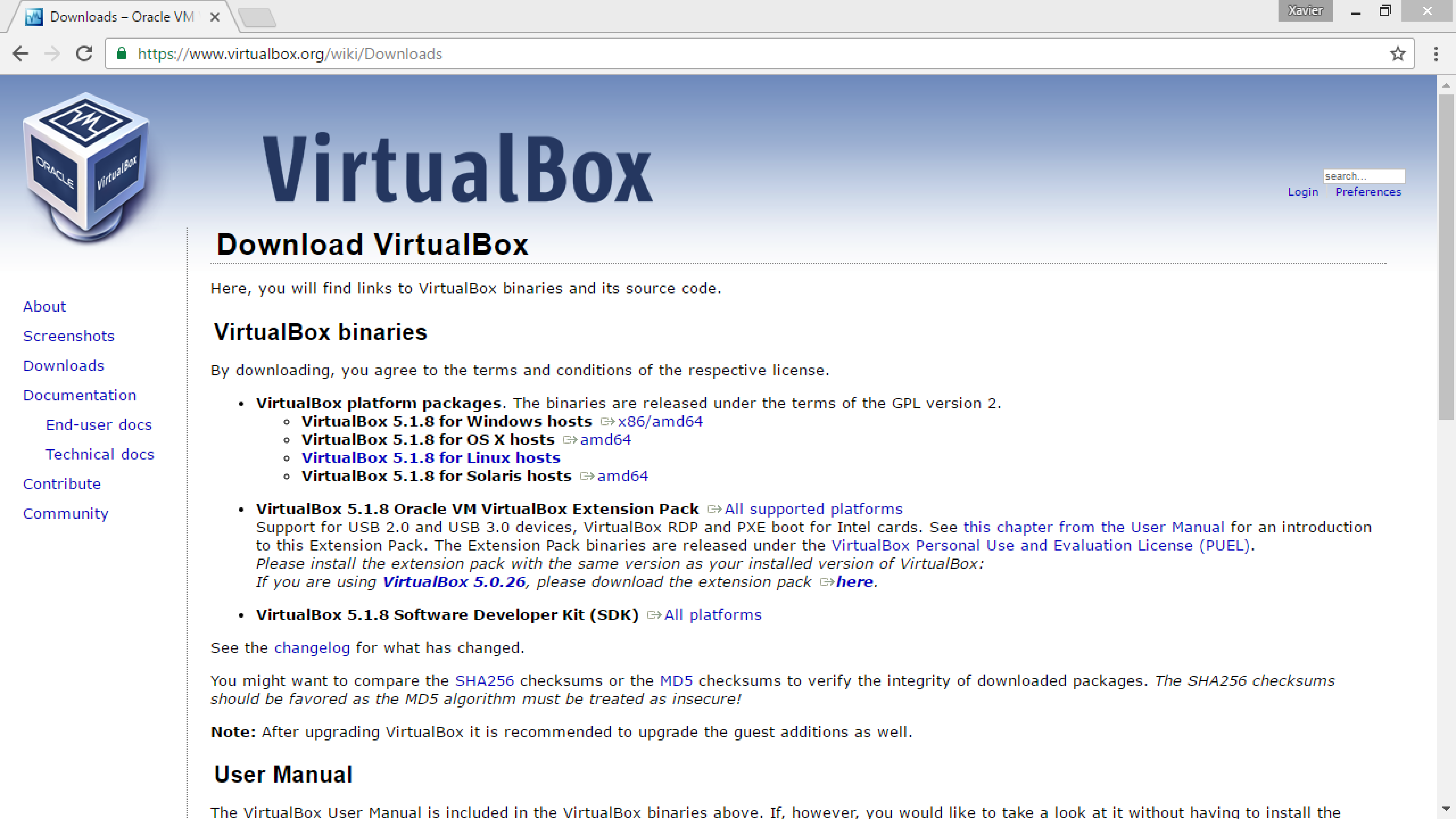
Free (really free)

- Even with multiple virtual machines

Demo

Exercise 1

Getting a Virtualization Environment Up & Running with 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 platform packages**. The binaries are released under the terms of the GPL version 2.
  - **VirtualBox 5.1.8 for Windows hosts** ⇨x86/amd64
  - **VirtualBox 5.1.8 for OS X hosts** ⇨amd64
  - **VirtualBox 5.1.8 for Linux hosts**
  - **VirtualBox 5.1.8 for Solaris hosts** ⇨amd64
- **VirtualBox 5.1.8 Oracle VM VirtualBox Extension Pack** ⇨All supported platforms
  Support for USB 2.0 and USB 3.0 devices, VirtualBox RDP and PXE boot for Intel cards. See this chapter from the User Manual for an introduction to this Extension Pack. The Extension Pack binaries are released under the VirtualBox Personal Use and Evaluation License (PUEL).
  Please install the extension pack with the same version as your installed version of VirtualBox:
  *If you are using **VirtualBox 5.0.26**, please download the extension pack ⇨**here**.*
- **VirtualBox 5.1.8 Software Developer Kit (SDK)** ⇨All platforms

See the changelog for what has changed.

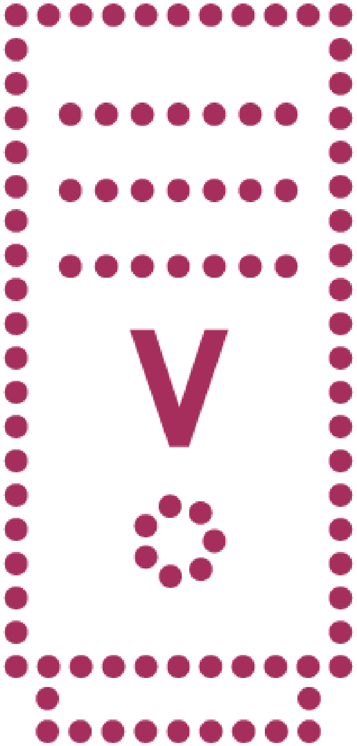
You might want to compare the SHA256 checksums or the MD5 checksums to verify the integrity of downloaded packages. *The SHA256 checksums should be favored as the MD5 algorithm must be treated as insecure!*

Note: After upgrading VirtualBox it is recommended to upgrade the guest additions as well.

User Manual

The VirtualBox User Manual is included in the VirtualBox binaries above. If, however, you would like to take a look at it without having to install the

Getting the Cloudera QuickStart VM



Getting started is easy

Download QuickStart VM

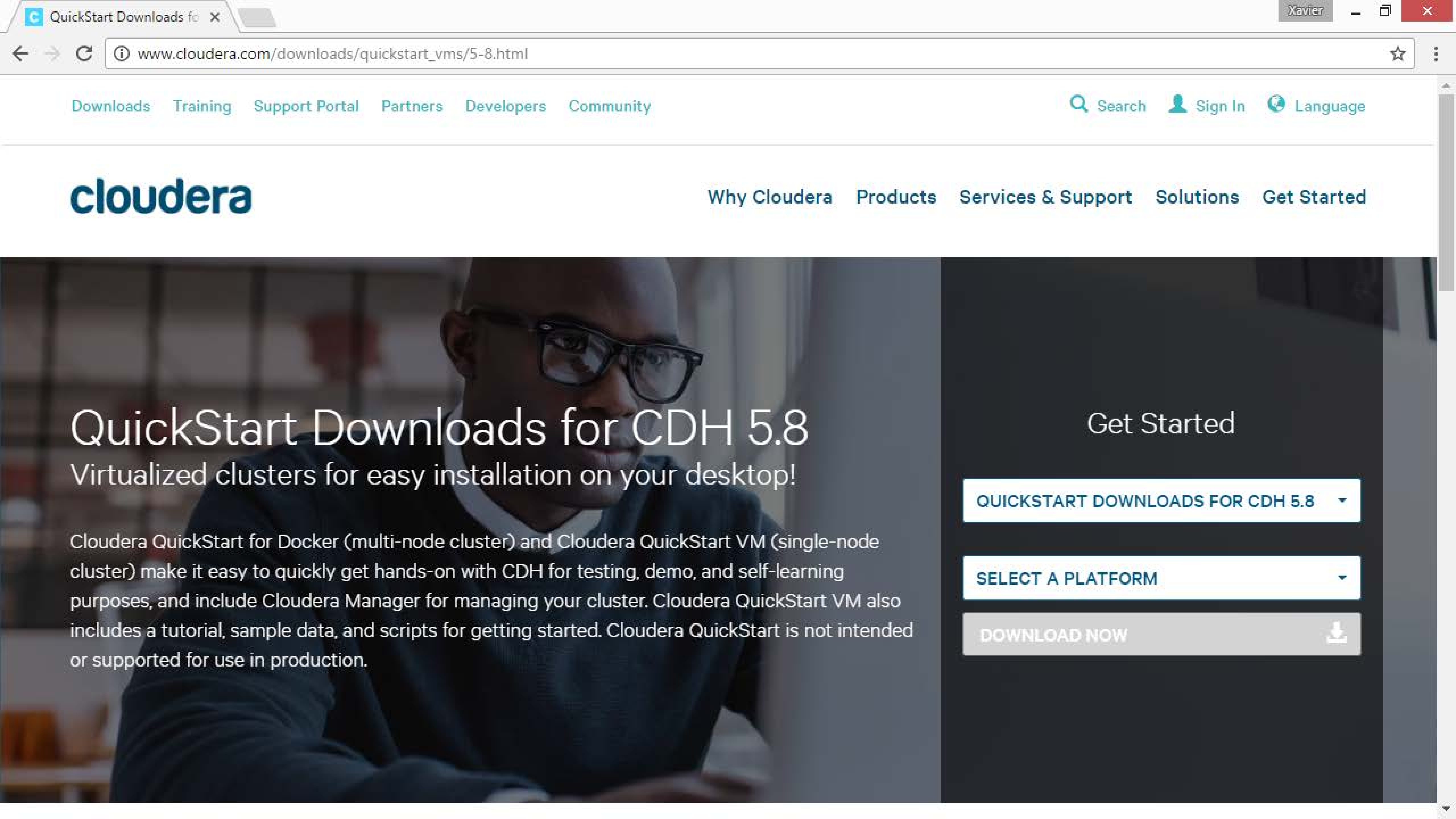
- From Cloudera.com

Import Appliance in VirtualBox

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

Demo

Exercise 2
Getting The Cloudera QuickStart VM Up & Running



QuickStart Downloads 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 QuickStart is not intended or supported for use in production.

Cloudera Manager

End to end application for managing CDH clusters

- Yes, plural

Intuitive and complete web UI

Versions

- Free
- Enterprise

Recommended values

- Like having an expert right next to you

Cloudera Manager

Features

- Automated deployment & configuration
- Monitoring, reporting & log management
- Zero downtime with rolling upgrades
- Extensible
- Integrated with Cloudera Support

Making Hadoop Easy with Cloudera Manager

Simple Administration for Apache Hadoop

Cloudera Manager is the industry's trusted tool for managing Hadoop in production

Starting Cloudera Manager

Turned off by default in QuickStart VM

Minimum configuration:

- 8 GiB RAM
- 2 virtual CPU cores

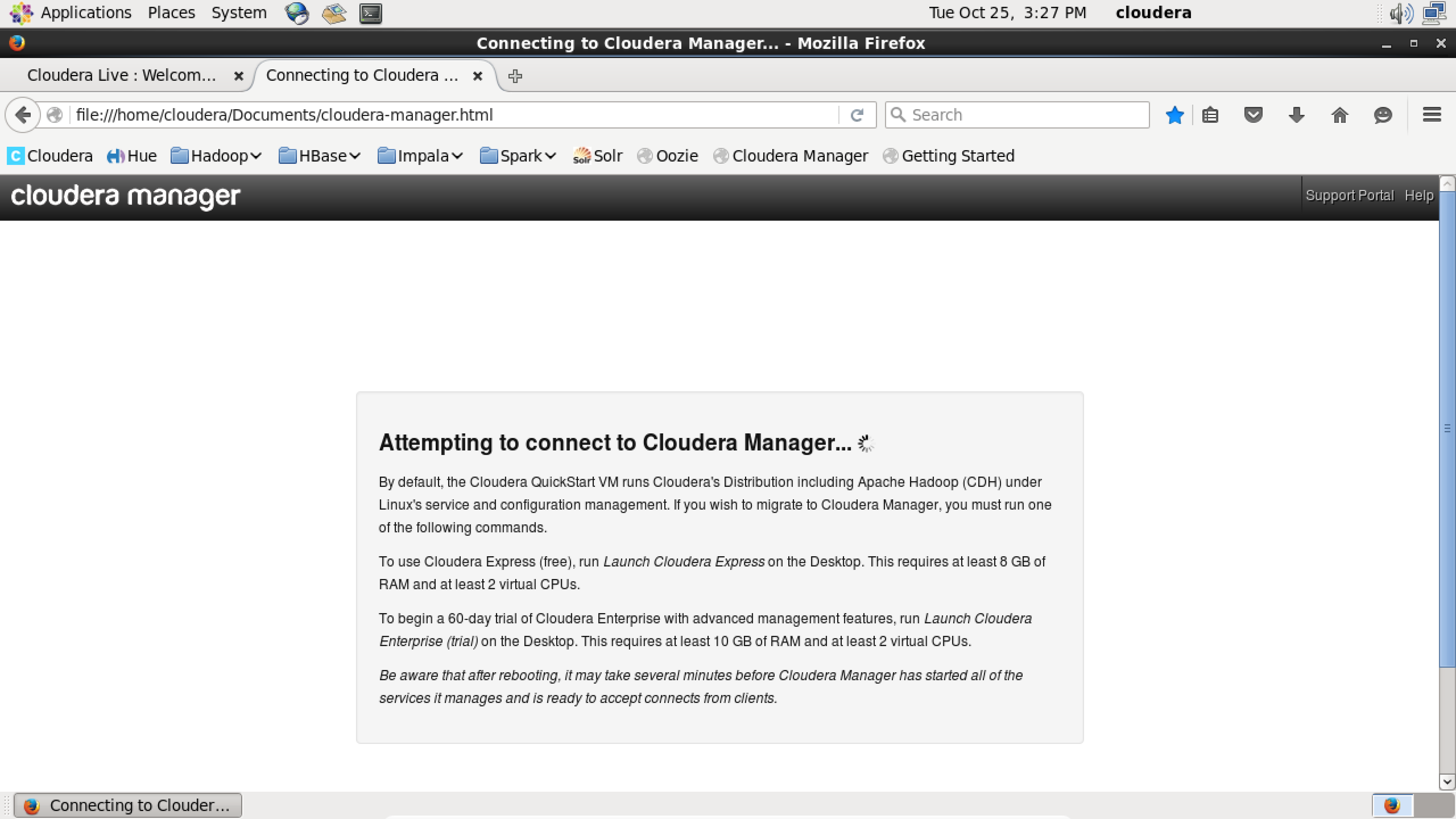
“Launch Cloudera Manager” icon in desktop

- Select desired version

http://quickstart.cloudera:7180

Demo

Exercise 3
Starting Cloudera Manager in the QuickStart VM



Applications Places System
Tue Oct 25, 3:27 PM cloudera
Connecting to Cloudera Manager... - Mozilla Firefox
Cloudera Live : Welcom...
Connecting to Cloudera ...
file:///home/cloudera/Documents/cloudera-manager.html
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera manager
Support Portal Help
Connecting to Clouder...

Attempting to connect to Cloudera Manager...

By default, the Cloudera QuickStart VM runs Cloudera's Distribution including Apache Hadoop (CDH) under Linux's service and configuration management. If you wish to migrate to Cloudera Manager, you must run one of the following commands.

To use Cloudera Express (free), run *Launch Cloudera Express* on the Desktop. This requires at least 8 GB of RAM and at least 2 virtual CPUs.

To begin a 60-day trial of Cloudera Enterprise with advanced management features, run *Launch Cloudera Enterprise (trial)* on the Desktop. This requires at least 10 GB of RAM and at least 2 virtual CPUs.

Be aware that after rebooting, it may take several minutes before Cloudera Manager has started all of the services it manages and is ready to accept connects from clients.

Let’s See Some Big Data in Action

Scenario you can relate to

Picture this

- Friend that asks many questions
- i.e. “should I learn Java or C#?”
- i.e. “tips on vi”

How can we help him?

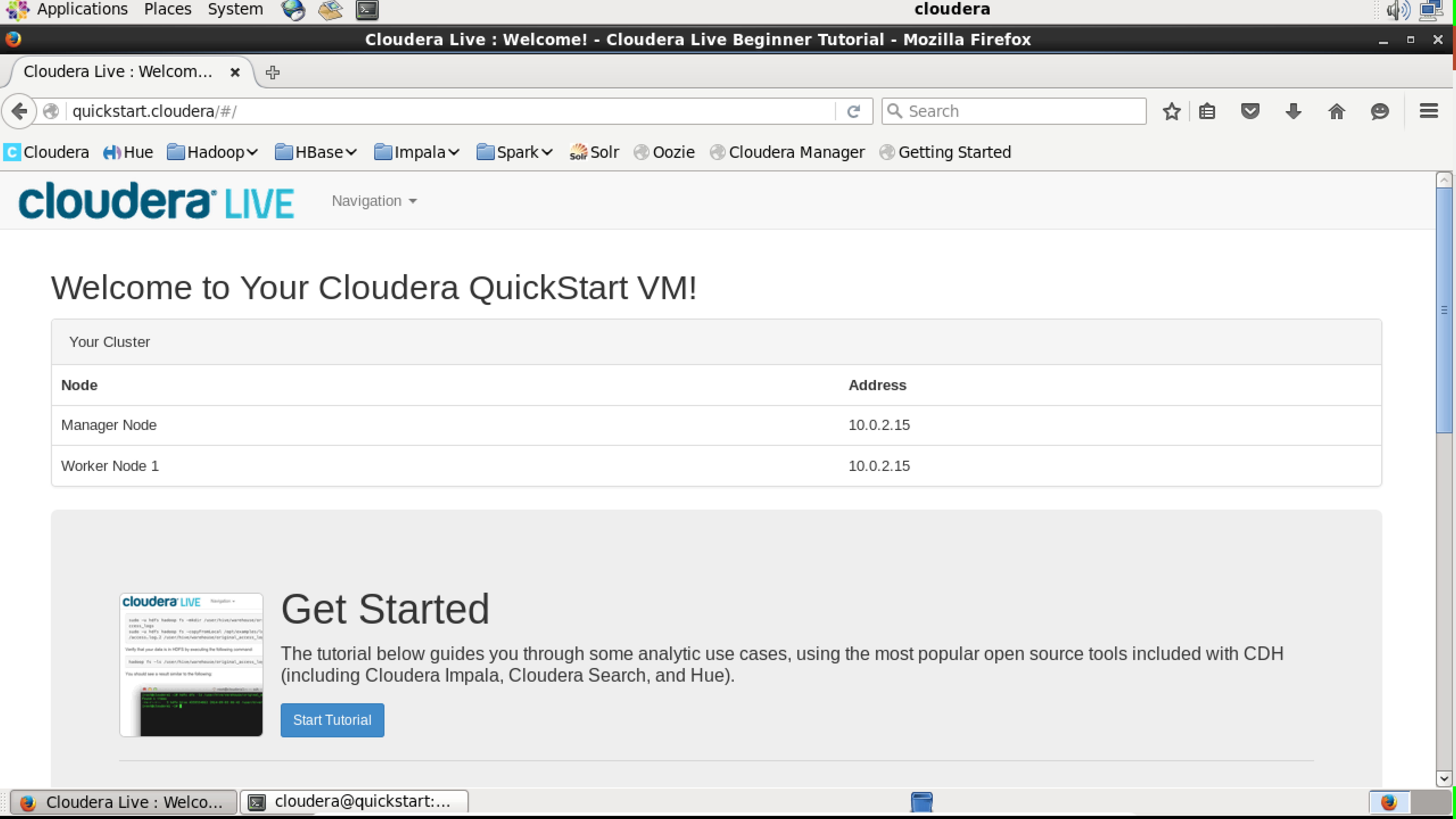
You think and get an idea using Hadoop

Let’s use StackExchange and get answers

Demo

Exercise 4
Loading a StackExchange Site into HDFS and Asking a Few Questions

key-bindings vimscript vimrc command-line regular-expression syntax-highlighting
search cursor-movement cut-copy-paste indentation more tags



Welcome to Your Cloudera QuickStart VM!

Your Cluster

Node	Address
Manager Node	10.0.2.15
Worker Node 1	10.0.2.15

Get Started

The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).

What's Next? A Real Cluster

We have a Big Data playground

- Ran a few jobs
- POCs, demos, learning
- Not a cluster.... It is a pseudo-cluster

Still a single machine

- Limitations

Time to scale

What are my options?

Where Can I Set up My Cluster?

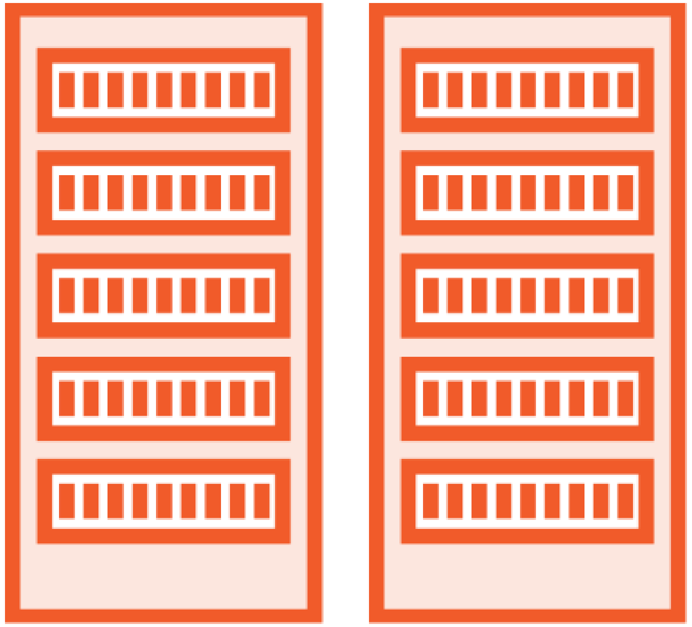
On-Premises

Amazon Web Services (AWS)

Microsoft Azure

Google Cloud

On-Premises



On-premises can mean physical or virtual

Starting point

Why?

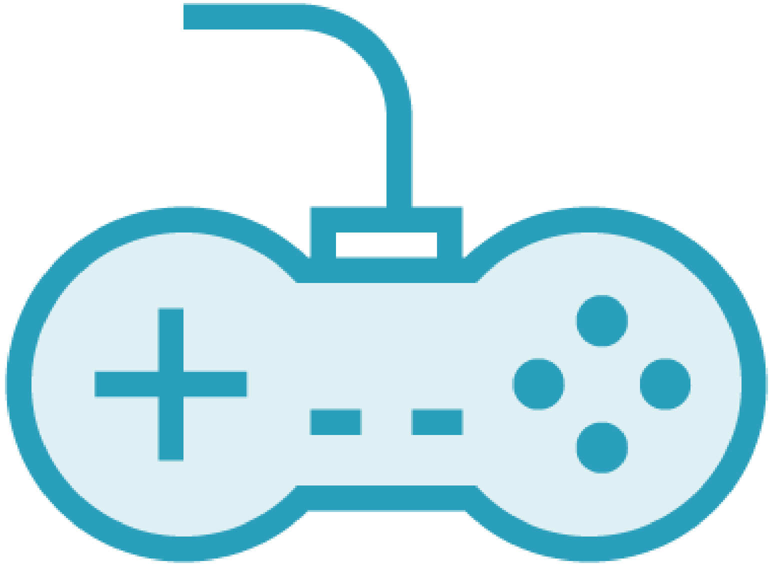
Focus on Cloudera

- Learn the installation process
- Avoid cloud specific details (for now)

But if you go for cloud

- Consider Cloudera Director

Cloudera Director



Makes it easy to deploy and manage production ready clusters in the cloud

Part of Cloudera Enterprise

Centralized administration

- Integrated with Cloudera Manager

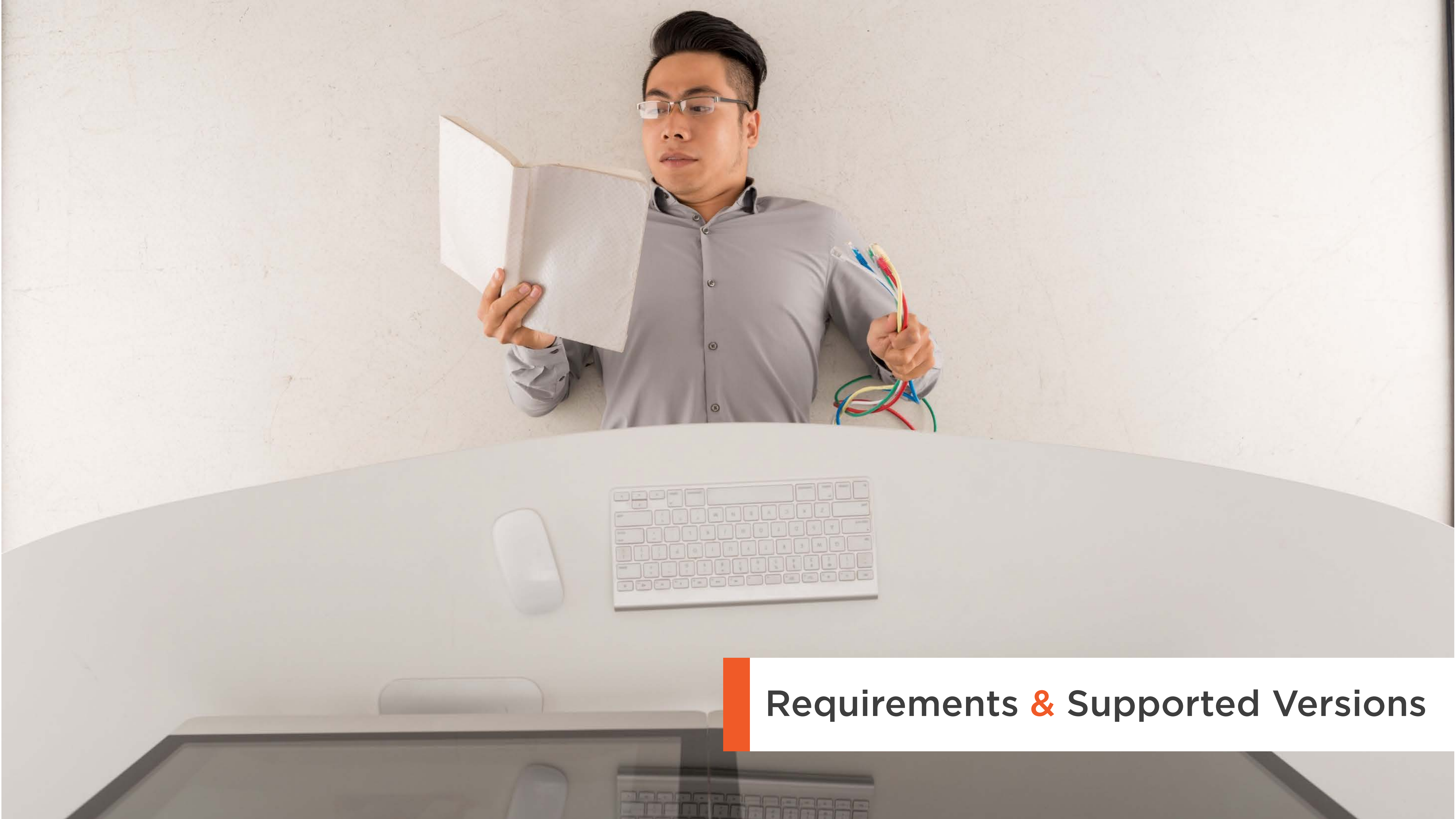
Automates many processes

- Adding an environment

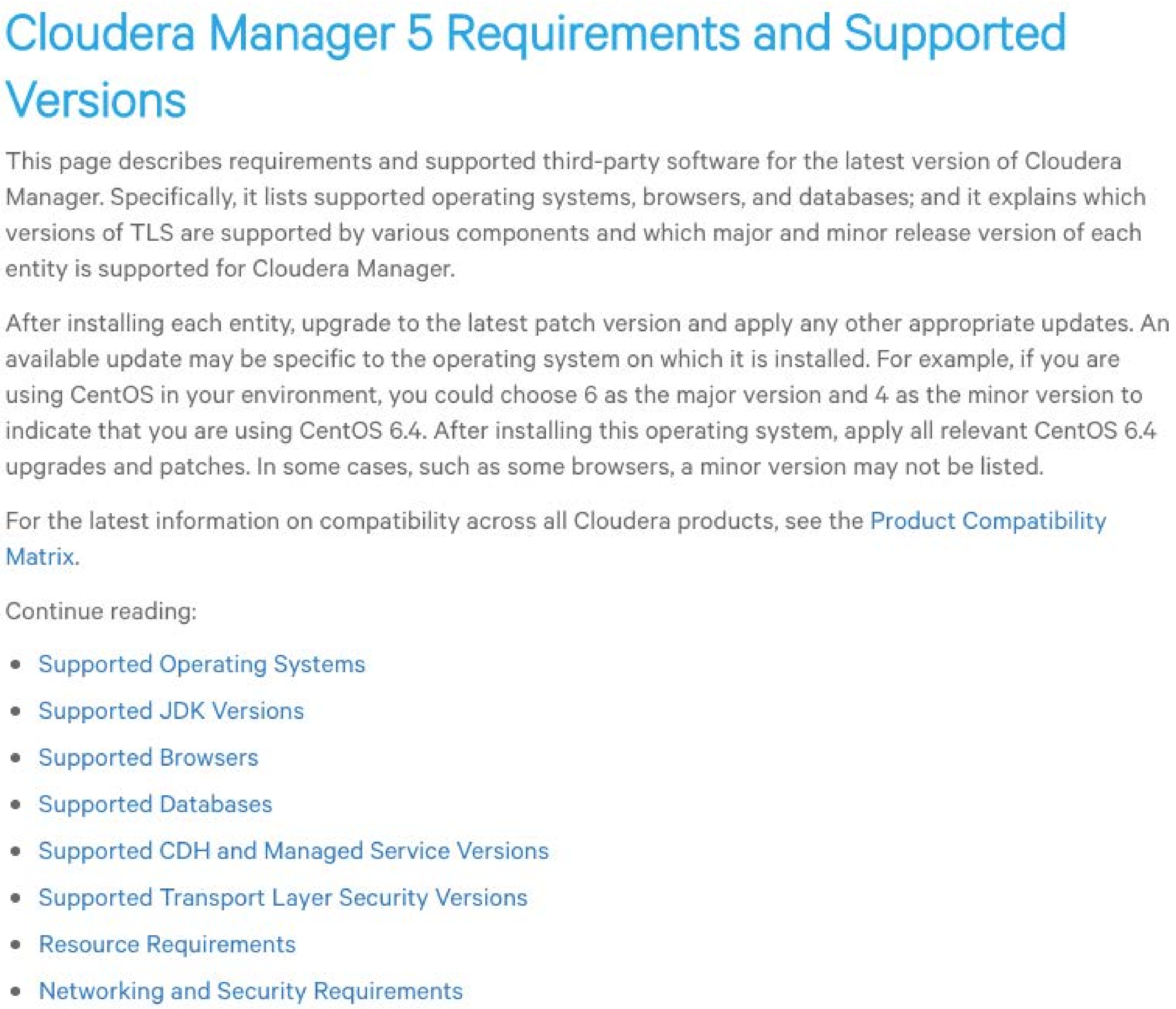
Works with AWS, Azure & Google Cloud

- Other integrations on the way

Deploy and Manage the Lifecycle of Cloudera Enterprise in the Cloud



Requirements & Supported Versions



Cloudera Manager 5 Requirements and Supported Versions

This page describes requirements and supported third-party software for the latest version of Cloudera Manager. Specifically, it lists supported operating systems, browsers, and databases; and it explains which versions of TLS are supported by various components and which major and minor release version of each entity is supported for Cloudera Manager.

After installing each entity, upgrade to the latest patch version and apply any other appropriate updates. An available update may be specific to the operating system on which it is installed. For example, if you are using CentOS in your environment, you could choose 6 as the major version and 4 as the minor version to indicate that you are using CentOS 6.4. After installing this operating system, apply all relevant CentOS 6.4 upgrades and patches. In some cases, such as some browsers, a minor version may not be listed.

For the latest information on compatibility across all Cloudera products, see the Product Compatibility Matrix.

Continue reading:

- Supported Operating Systems
- Supported JDK Versions
- Supported Browsers
- Supported Databases
- Supported CDH and Managed Service Versions
- Supported Transport Layer Security Versions
- Resource Requirements
- Networking and Security Requirements

Supported Versions: Operating Systems & JDK

Operating System

- Linux
- Red Hat Enterprise Linux (compatible), CentOS, SUSE Enterprise, Ubuntu, Debian
- Same version across cluster for support
- Check exact versions in docs

JDK

- Most CDH versions support 7 & 8
- Deploy on same major version
- Patches & updates

Supported Databases & Browsers

Databases

- Extremely important in your cluster
    - Hold all your cluster’s information
    - Remember backups!
- Embedded PostgreSQL in non prod
- Production
    - MariaDB, MySQL, SQLite, PostgreSQL, Oracle, Derby*

Browsers

- Hue & Cloudera Manager
- Firefox, Chrome, Internet Explorer, Safari

Additional Details

Supported CDH & Managed Service Versions

Permissions

- Single user mode

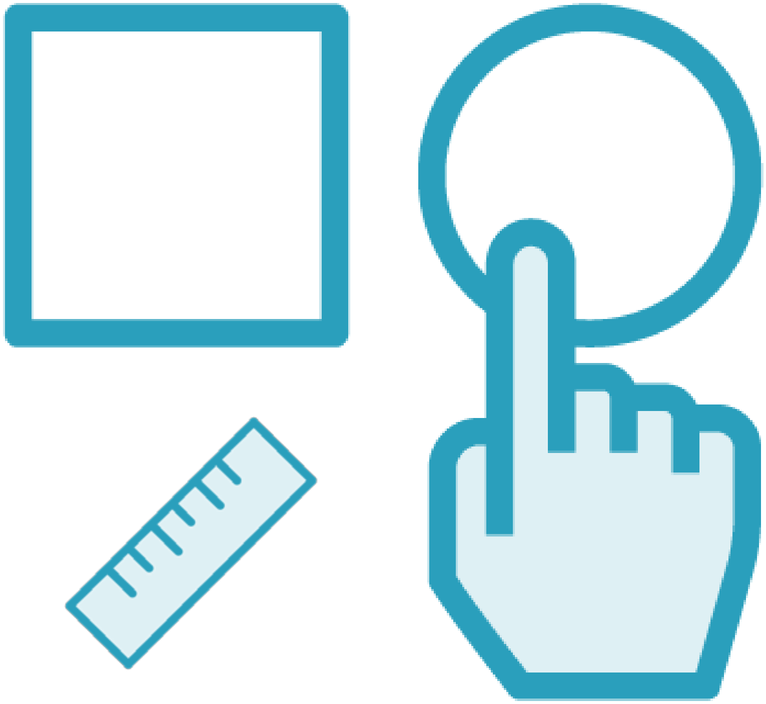
Transport Layer Security Versions

Resource Requirements

Networking and Security Requirements

- IPV4 / IPV6 not supported

Selecting Hardware for your Cluster



Hadoop is different

- Take computation to the data

Not a One-Size-Fits-All situation

Vary by role

“It depends”

- Your type of application
- Chicken and the egg problem

Sizing exercise (benchmark)

- Easier in the cloud

Workload Type Matters

IO Bound

Read from disk or network

Sorting

Indexing

Grouping

Data importing and exporting

Data movement and transformation

CPU Bound

Complex operations on input data

Classification

Clustering

Complex text mining

Natural-language processing

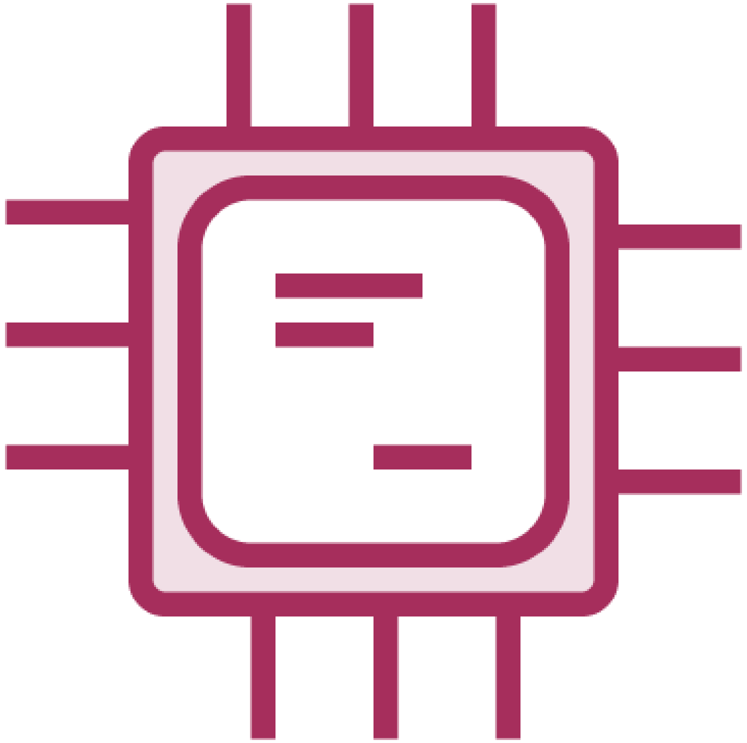
Feature extraction

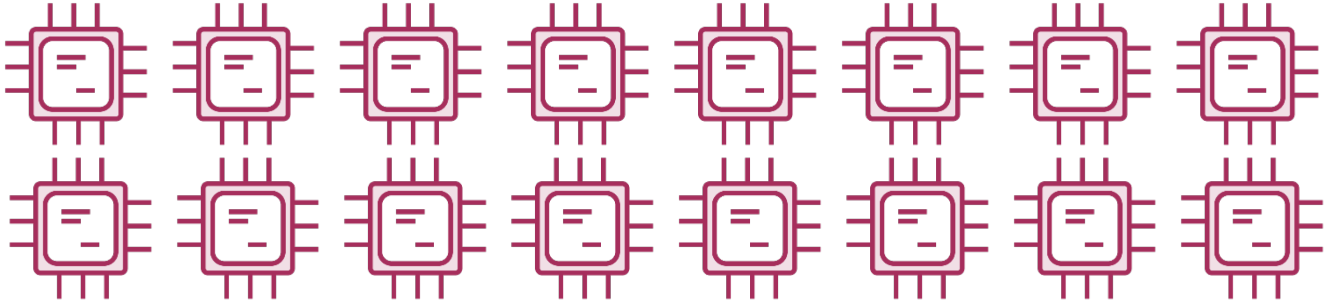
Minimum Memory Requirements

CDH	4+ GiB
Cloudera Express	8+ GiB
Cloudera Enterprise	10+ GiB

Common to see 64 GiB to 128 GiB

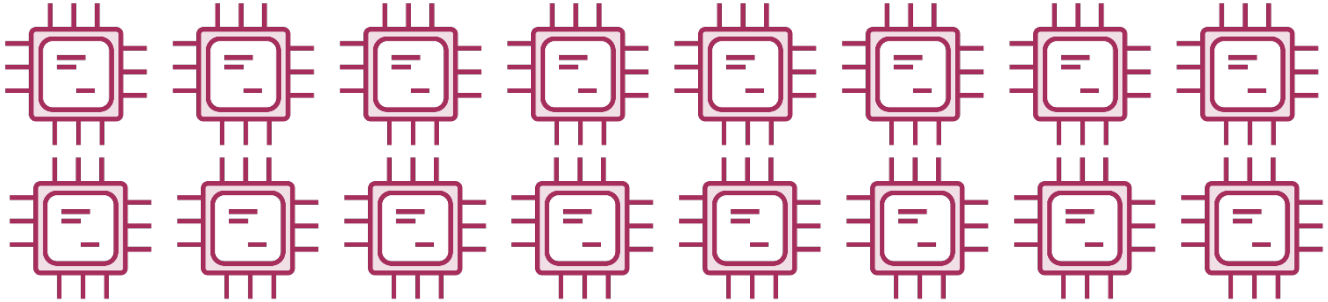
Processor





Always at least 2, but the more the merrier

i.e. 32 cores not far fetched



Takeaway

Cloudera QuickStart VM

FastTrack to Big Data

One download away

Pseudo-cluster

Everything we need

- POCs
- Demos
- Learning
- Testing

Requirements

- Recommended to read documentation
- The more the merrier