This Course:

Introduction to Big Data

By the end of this course you will be able to...

Describe the Big Data landscape including examples of real world big data problems and approaches.

Identify the high level components in the data science lifecycle and associated data flow.

Explain the V’s of Big Data and why each impacts the collection, monitoring, storage, analysis and reporting, including their impact in the presence of multiple V’s.

Identify big data problems and be able to recast problems as data science questions.

Summarize the features and significance of the HDFS file system and the MapReduce programming model and how they relate to working with Big Data.

**A “Small” Definition of Big Data**

The term ‘big data’ seems to be popping up everywhere these days. And there seems to be as many uses of this term as there are contexts in which you find it: ‘big data’ is often used to refer to any dataset that is difficult to manage using traditional database systems; it is also used as a catch-all term for any collection of data that is too large to process on a single server; yet others use the term to simply mean “a lot of data”; sometimes it turns out it doesn’t even have to be large. So what exactly is big data?

A precise specification of ‘big’ is elusive. What is considered big for one organization may be small for another. What is large-scale today will likely seem small-scale in the near future; petabyte is the new terabyte. Thus, size alone cannot specify big data. The complexity of the data is an important factor that must also be considered.

Most now agree with the characterization of big data using the 3 V’s coined by Doug Laney of Gartner:

· Volume: This refers to the vast amounts of data that is generated every second/minute/hour/day in our digitized world.

· Velocity: This refers to the speed at which data is being generated and the pace at which data moves from one point to the next.

· Variety: This refers to the ever-increasing different forms that data can come in, e.g., text, images, voice, geospatial.

A fourth V is now also sometimes added:

· Veracity: This refers to the quality of the data, which can vary greatly.

There are many other V's that gets added to these depending on the context. For our specialization, we will add:

· Valence: This refers to how big data can bond with each other, forming connections between otherwise disparate datasets.

The above V’s are the dimensions that characterize big data, and also embody its challenges: We have huge amounts of data, in different formats and varying quality, that must be processed quickly.

It is important to note that the goal of processing big data is to gain insight to support decision-making. It is not sufficient to just be able to capture and store the data. The point of collecting and processing volumes of complex data is to understand trends, uncover hidden patterns, detect anomalies, etc. so that you have a better understanding of the problem being analyzed and can make more informed, data-driven decisions. In fact, many consider value as the sixth V of big data:

· Value: Processing big data must bring about value from insights gained.

To address the challenges of big data, innovative technologies are needed. Parallel, distributed computing paradigms, scalable machine learning algorithms, and real-time querying are key to analysis of big data. Distributed file systems, computing clusters, cloud computing, and data stores supporting data variety and agility are also necessary to provide the infrastructure for processing of big data. [Workflows](http://words.sdsc.edu/words-data-science/workflows) provide an intuitive, reusable, scalable and reproducible way to process big data to gain verifiable value from it in and enable application of same methods to different datasets.

With all the data generated from social media, smart sensors, satellites, surveillance cameras, the Internet, and countless other devices, big data is all around us. The endeavor to make sense out of that data brings about exciting opportunities indeed!

Source: <http://words.sdsc.edu/words-data-science/big-data>

Data Science is about extracting knowledge from data. At the WorDS Center (words.sdsc.edu), we define data science as a multidisciplinary craft that combines people, process, computational and Big Data platforms, application-specific purpose and programmability. Publications and provenance of the data products leading to these publications are also important for data science, but we start by defining 5 P's that take significant part in the data science activities.

* **Purpose:**The purpose refers to the challenge or set of challenges defined by your big data strategy. The purpose can be related to a scientific analysis with a hypothesis or a business metric that needs to be analyzed based often on Big Data.
* **People**: The data scientists are often seen as people who possess skills on a variety of topics including: science or business domain knowledge; analysis using statistics, machine learning and mathematical knowledge; data management, programming and computing. In practice, this is generally a group of researchers comprised of people with complementary skills.
* **Process**: Since there is a predefined team with a purpose, a great place for this team to start with is a process they could iterate on. We can simply say, People with Purpose will define a Process to collaborate and communicate around! The process of data science includes techniques for statistics, machine learning, programming, computing and data management. A process is conceptual in the beginning and defines the course set of steps and how everyone can contribute to it. Note that similar reusable processes can be applicable to many applications with different purposes when employed within different workflows. Data science workflows combine such steps in executable graphs. We believe that process-oriented thinking is a transformative way of conducting data science to connect people and techniques to applications. Execution of such a data science process requires access to many datasets, Big and small, bringing new opportunities and challenges to Data Science. There are many Data Science steps or tasks, such as Data Collection, Data Cleaning, Data Processing/Analysis, Result Visualization, resulting in a Data Science Workflow. Data Science Processes may need user interaction and other manual operations, or be fully automated.Challenges for the data science process include 1) how to easily integrate all needed tasks to build such a process; 2) how to find the best computing resources and efficiently schedule process executions to the resources based on process definition, parameter settings, and user preferences.
* **Platforms**: Based on the needs of an application-driven purpose and the amount of data and computing required to perform this application, different computing and data platforms can be used as a part of the data science process. This scalability should be made part of any data science solution architecture.
* **Programmability**: Capturing a scalable data science process requires aid from programming languages, e.g., R, and patterns, e.g., MapReduce. Tools that provide access to such programming techniques are key to making the data science process programmable on a variety of platforms.

To summarize, data science can be defined as a craft of using the five pieces identified above. Having a process between the more business driven P’s people and purpose and the more technical driven P’s platforms and programmability leads to a streamlined approach that starts and ends with a defined business value, team accountability and collaboration in mind.

# ****MapReduce****

Map-Reduce is a scalable programming model that simplifies distributed processing of data. Map-Reduce consists of three main steps: Mapping, Shuffling and Reducing. An easy way to think about a Map-Reduce job is to compare it with act of ‘delegating’ a large task to a group of people, and then combining the result of each person’s effort, to produce the final outcome.

Let’s take an example to bring the point across. You just heard about this great news at your office, and are throwing a party for all your colleagues! You decide to cook Pasta for the dinner. Four of your friends, who like cooking, also volunteer to join you in preparation. The task of preparing Pasta broadly involves chopping the vegetables, cooking, and garnishing.

Let’s take the job of chopping the vegetables and see how it is analogous to a map-reduce task. Here the raw vegetables are symbolic of the input data, your friends are equivalent to compute nodes, and final chopped vegetables are analogous to desired outcome. Each friend is allotted onions, tomatoes and peppers to chop and weigh.

You would also like to know how much of each vegetable types you have in the kitchen. You would also like to chop these vegetables while this calculation is occurring. In the end, the onions should be in one large bowl with a label that displays its weight in pounds, tomatoes in a separate one, and so on.

## MAP

To start with, you assign each of your four friends a random mix of different types of vegetables.



They are required to use their ‘compute’ powers to chop them and measure the weight of each type of veggie. They need to ensure not to mix different types of veggies. So each friend will generate a mapping of <key, value> pairs that looks like:

**Friend X:**

<tomatoes, 5 lbs>

<onions, 10 lbs>

<garlic, 2 lbs>

**Friend Y:**

<onions, 22 lbs>

<green peppers, 3 lbs>

…

Seems like you are having a really big party! Now that your friends have chopped the vegetables, and labeled each bowl with the weight and type of vegetable, we move to the next stage: Shuffling.

SHUFFLE: This stage is also called Grouping. Here you want to group the veggies by their types. You assign different parts of your kitchen to each type of veggie, and your friends are supposed to group the bowls, so that like items are placed together:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/W2o7c8AwEeWYyAqU4MStYQ_c4c5370b022450dc31126280db02e70c_tomato.png?expiry=1553472000000&hmac=Vt7YbWhCM450ClAggINosGC9HiAgiIKAImbOl32EDfI

**North End of Kitchen:**

<tomatoes, 5 lbs>

<tomatoes, 11 lbs>

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/SFKwksAwEeWcpBJ24z85uQ_98203f10b39000550a6af9bb7ee35e03_onion.png?expiry=1553472000000&hmac=W-Wu1nUmXRYLSHAulsyR1D1xVj_bspu79cKFZWzzxks

**West End of Kitchen:**

<onions, 10 lbs>

<onions, 22 lbs>

<onions, 1.4 lbs>

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/qcogxsAvEeWQhhJguuWa9w_db6903659b9a1199dcbd7ae9dbf90dce_pepper.png?expiry=1553472000000&hmac=LP2k7S-dTE1ZghHTfuzje3tbAf__SGg1c87wBGbX44Q

**East End of Kitchen:**

<green peppers, 3 lbs>

<green peppers, 10 lbs>

The party starts in a couple of hours, but you are impressed by what your friends have accomplished by Mapping and Grouping so far! The kitchen looks much more organized now and the raw material is chopped. The final stage of this task is to measure how much of each veggie you actually have. This brings us to the Reduce stage.

## REDUCE

In this stage, you ask each of your friend to collect items of same type, put them in a large bowl, and label this large bowl with sum of individual bowl weights. Your friends cannot wait for the party to start, and immediately start ‘reducing’ small bowls. In the end, you have nice large bowls, with total weight of each vegetable labeled on it.



<tomatoes, 16 lbs>



<green peppers, 13lbs>



<onions, 33.4 lbs>

The number represents total weight of that vegetable after reducing from smaller bowls

Your friends (‘compute nodes’) just performed a Map-Reduce task to help you get started with cooking the Pasta. Since you were coordinating the entire exercise, you are “The Master” node of this Map-Reduce task. Each of your friends took roles of Mappers, Groupers and Reducers at different times. This example demonstrates the power of this technique.

This simple and powerful technique can be scaled very easily if more of your friends decide to join you. In future, we will continue to add more articles on different open source tools that will help you easily implement Map-Reduce to solve your computational problems.

**Source:**<http://words.sdsc.edu/words-data-science/mapreduce>

Installation on Mac

## Learning Goals

In this activity, you will:

* Download and Install VirtualBox.
* Download and Install Cloudera Virtual Machine (VM) Image.
* Launch the Cloudera VM.

Hardware Requirements: (A) Quad Core Processor (VT-x or AMD-V support recommended), 64-bit; (B) 8 GB RAM; (C) 20 GB disk free. How to find your hardware information: Open Overview by clicking on the Apple menu and clicking “About This Mac.” Most computers with 8 GB RAM purchased in the last 3 years will meet the minimum requirements. You will need a high speed internet connection because you will be downloading files up to 4 Gb in size.

Instructions

Please use the following instructions to download and install the Cloudera Quickstart VM with VirutalBox before proceeding to the Getting Started with the Cloudera VM Environment video. The screenshots are from a Mac but the instructions should be the same for Windows. Please see the discussion boards if you have any issues.

1. **Install VirtualBox.** Go to <https://www.virtualbox.org/wiki/Downloads> to download and install VirtualBox for your computer. The course uses Virtualbox 5.1.X, so we recommend clicking [VirtualBox 5.1 builds](https://www.virtualbox.org/wiki/Download_Old_Builds_5_1) on that page and downloading the older package for ease of following instructions and screenshots. However, it shouldn't be too different if you choose to use or upgrade to VirtualBox 5.2.X.

2. **Download the Cloudera VM.** Download the Cloudera VM from <https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.4.2-0-virtualbox.zip>. The VM is over 4GB, so will take some time to download.

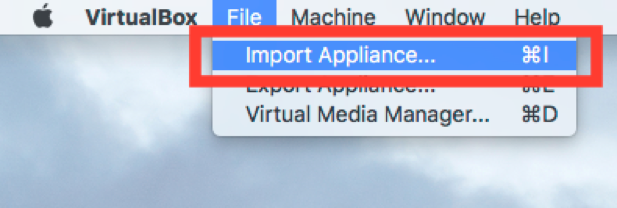
3. **Unzip the Cloudera VM**:

On Mac: Double click cloudera-quickstart-vm-5.4.2-0-virtualbox.zip

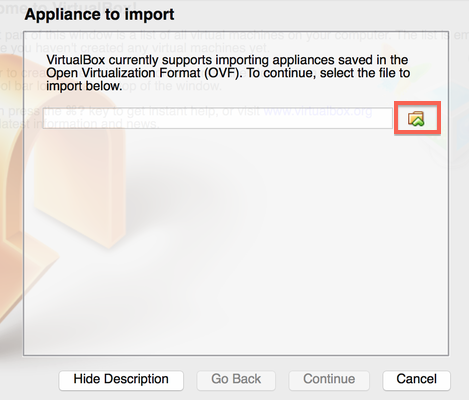
On Windows: Right-click cloudera-quickstart-vm-5.4.2-0-virtualbox.zip and select “Extract All…”

4. **Start VirtualBox**.

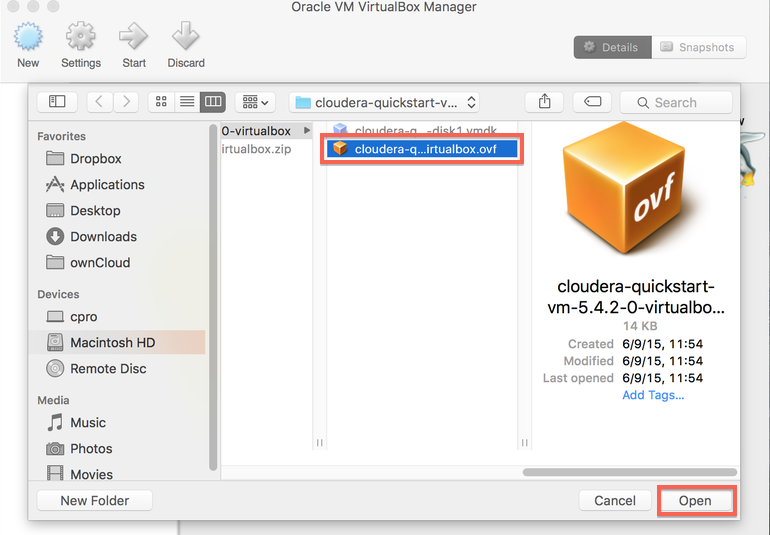
5. **Begin importing**. Import the VM by going to File -> Import Appliance



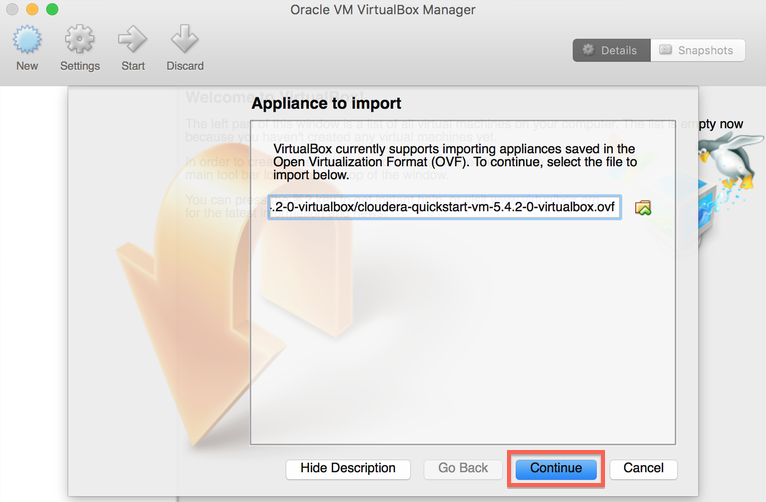
6. **Click the Folder icon.**



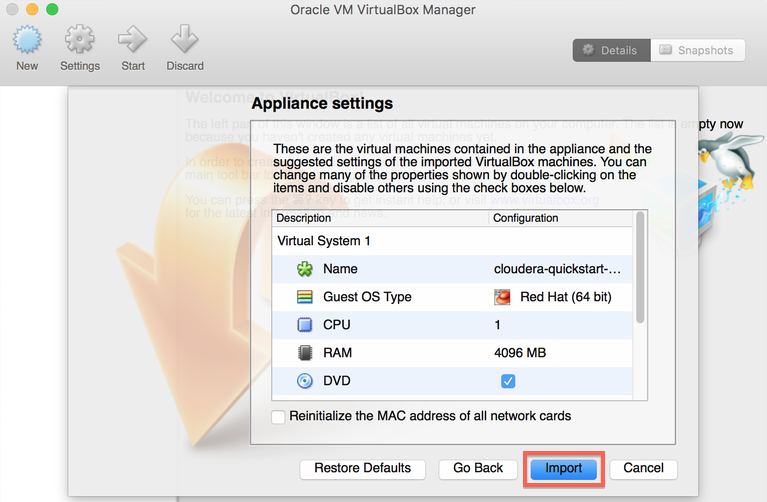
7. **Select the cloudera-quickstart-vm-5.4.2-0-virtualbox.ovf** from the Folder where you unzipped the VirtualBox VM and click Open.



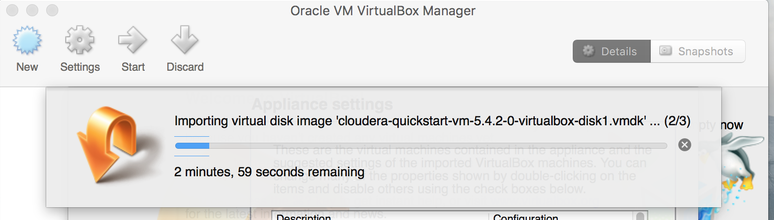
8. **Click Continue** to proceed.



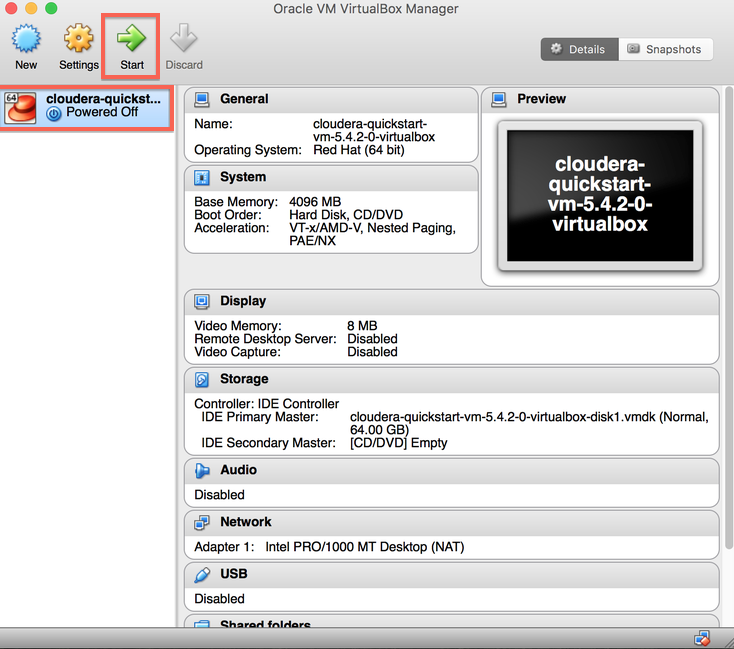
9. **Click Import.**



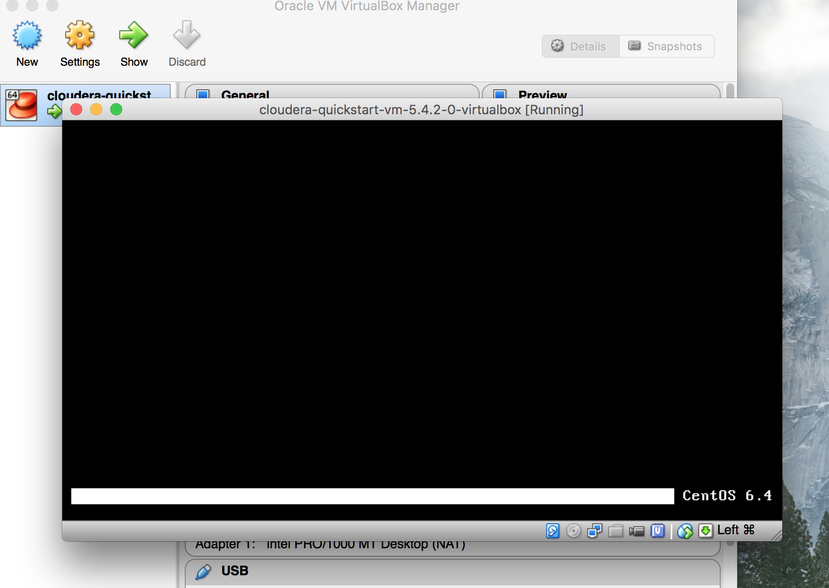
10. **The virtual machine image will be imported.** This can take several minutes.



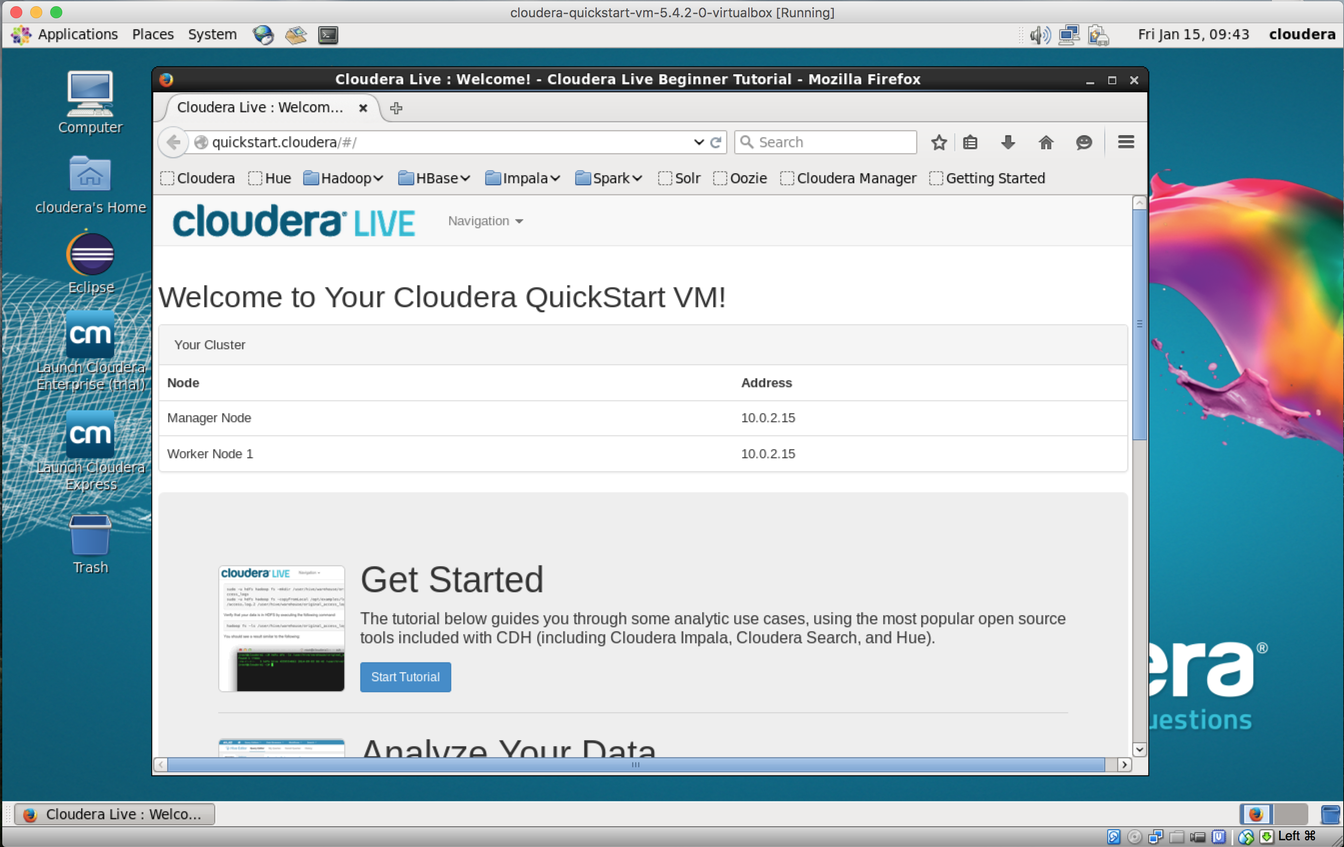
11. **Launch Cloudera VM.** When the importing is finished, the quickstart-vm-5.4.2-0 VM will appear on the left in the VirtualBox window. Select it and click the Start button to launch the VM.



12. **Cloudera VM booting.**It will take several minutes for the Virtual Machine to start. The booting process takes a long time since many Hadoop tools are started.



13. **The Cloudera VM desktop.**Once the booting process is complete, the desktop will appear with a browser.



Installation on Windows

## Learning Goals

In this activity, you will:

* Download and Install VirtualBox.
* Download and Install Cloudera Virtual Machine (VM) Image.
* Launch the Cloudera VM.

Hardware Requirements: (A) Quad Core Processor (VT-x or AMD-V support recommended), 64-bit; (B) 8 GB RAM; (C) 20 GB disk free. How to find your hardware information: Open System by clicking the Start button, right-clicking Computer, and then clicking Properties. Most computers with 8 GB RAM purchased in the last 3 years will meet the minimum requirements.You will need a high speed internet connection because you will be downloading files up to 4 Gb in size.

Instructions

Please use the following instructions to download and install the Cloudera Quickstart VM with VirutalBox before proceeding to the Getting Started with the Cloudera VM Environment video. The screenshots are from a Mac but the instructions should be the same for Windows. Please see the discussion boards if you have any issues.

1. **Install VirtualBox.** Go to <https://www.virtualbox.org/wiki/Downloads> to download and install VirtualBox for your computer. The course uses Virtualbox 5.1.X, so we recommend clicking [VirtualBox 5.1 builds](https://www.virtualbox.org/wiki/Download_Old_Builds_5_1) on that page and downloading the older package for ease of following instructions and screenshots. However, it shouldn't be too different if you choose to use or upgrade to VirtualBox 5.2.X. For Windows, select the link "**VirtualBox 5.1.X for Windows hosts x86/amd64**" where 'X' is the latest version.

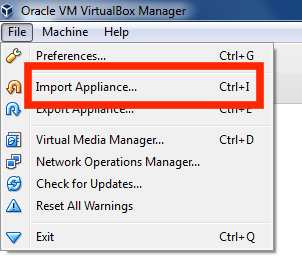
2. **Download the Cloudera VM.** Download the Cloudera VM from<https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.4.2-0-virtualbox.zip>. The VM is over 4GB, so will take some time to download.

3. **Unzip the Cloudera VM**:

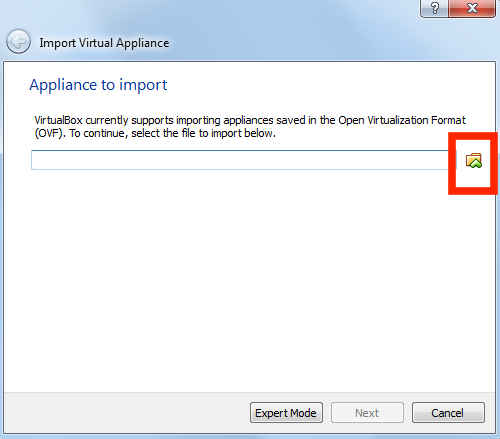
Right-click cloudera-quickstart-vm-5.4.2-0-virtualbox.zip and select “Extract All…”

4. **Start VirtualBox**.

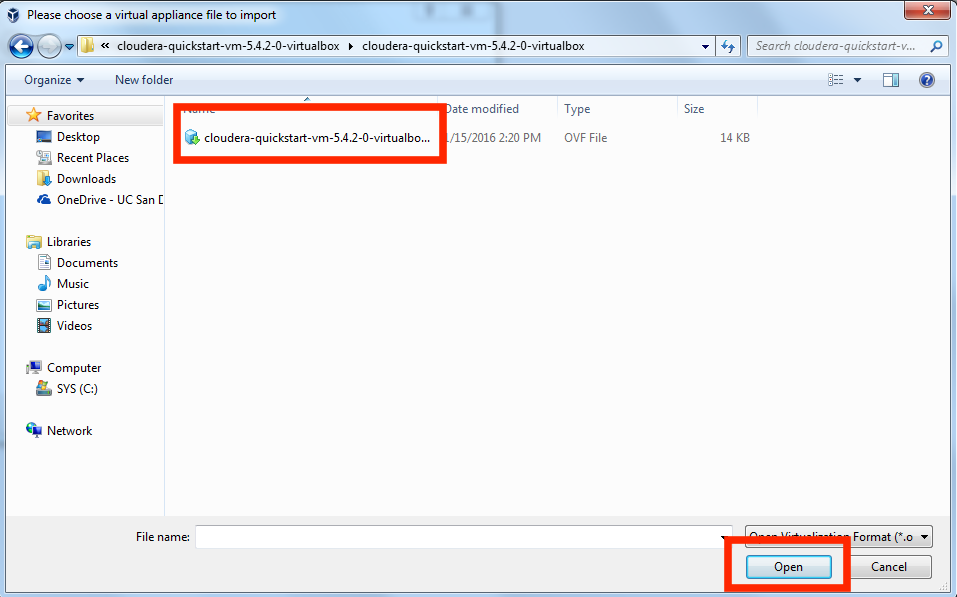
5. **Begin importing**. Import the VM by going to File -> Import Appliance



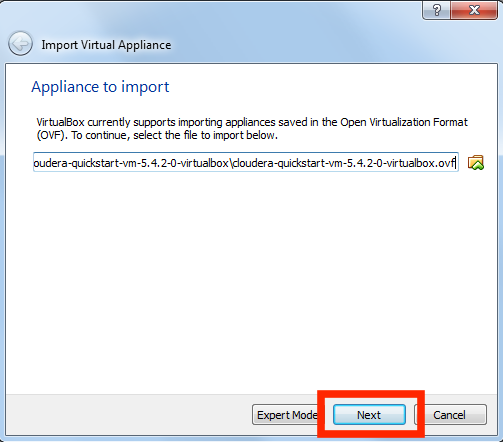
6. **Click the Folder icon.**



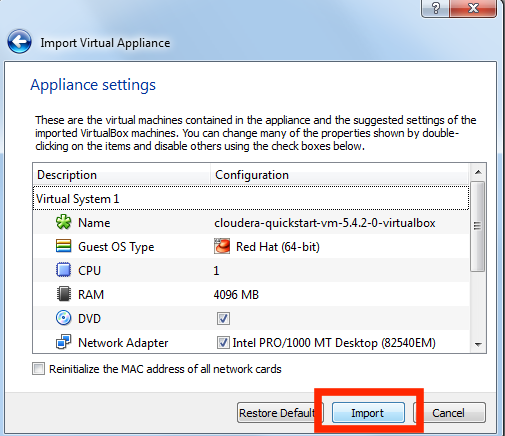
7. **Select the cloudera-quickstart-vm-5.4.2-0-virtualbox.ovf** from the Folder where you unzipped the VirtualBox VM and click Open.



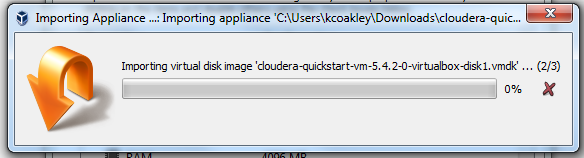
8. **Click Next** to proceed.



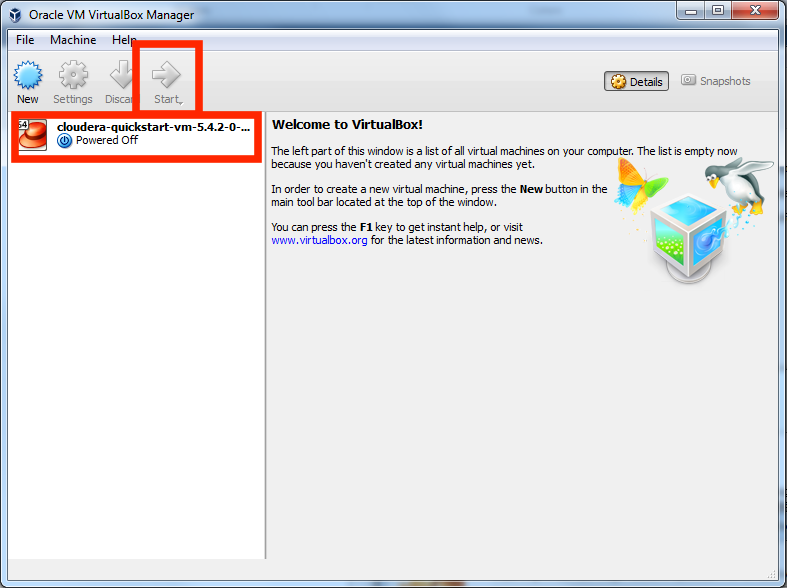
9. **Click Import.**



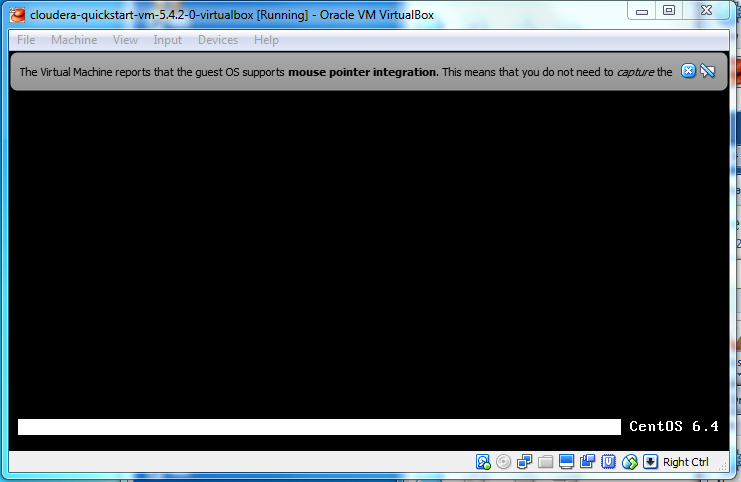
10. **The virtual machine image will be imported.** This can take several minutes.



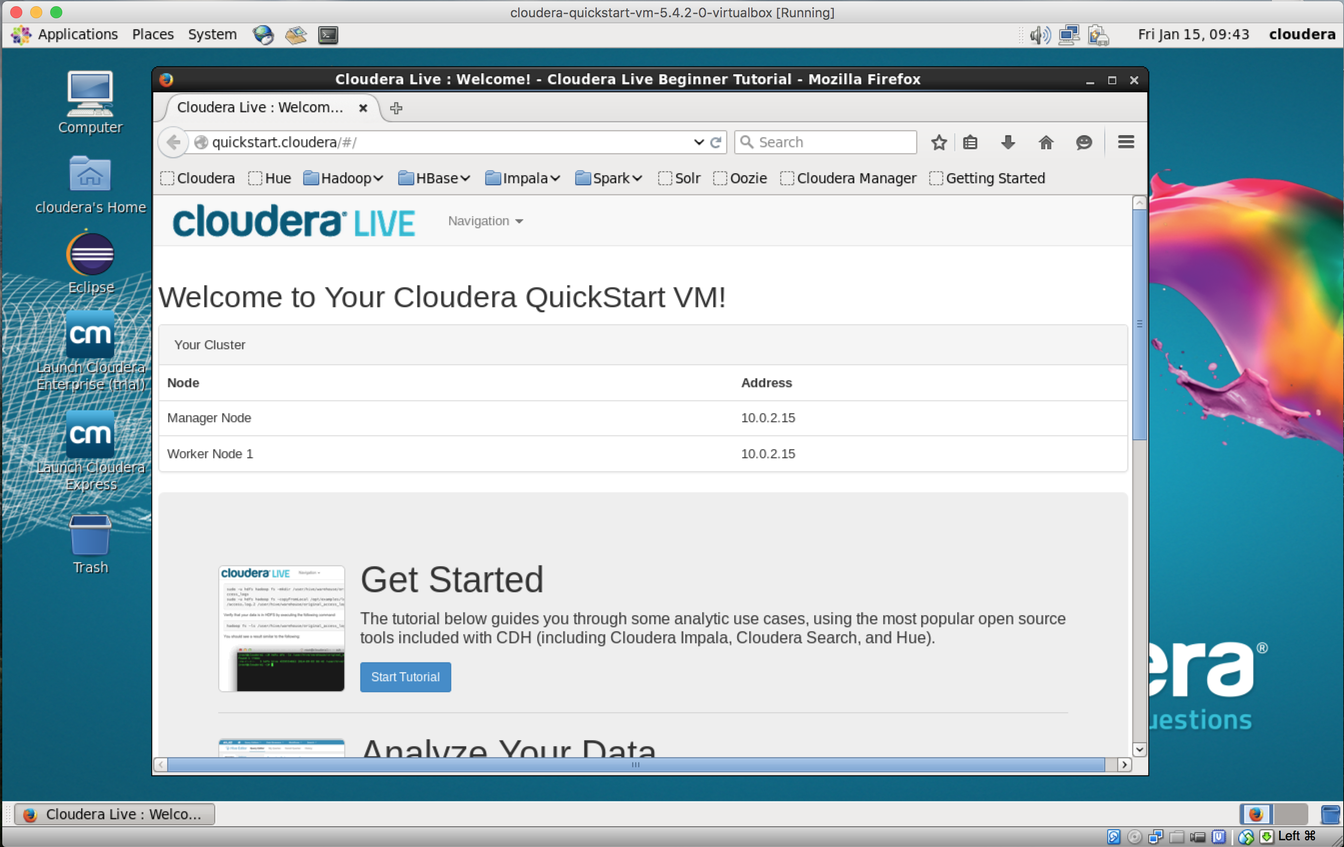
11. **Launch Cloudera VM.** When the importing is finished, the quickstart-vm-5.4.2-0 VM will appear on the left in the VirtualBox window. Select it and click the Start button to launch the VM.



12. **Cloudera VM booting.**It will take several minutes for the Virtual Machine to start. The booting process takes a long time since many Hadoop tools are started.



13. **The Cloudera VM desktop.**Once the booting process is complete, the desktop will appear with a browser.



* **I am getting a failed to import error when I try to import Cloudera into Virtual Box?**

The PC BIOS needs to be set to allow virtual technology.

* **My screen keeps freezing even though virtual machine is in 'running state'?**

Often times the virtual machine takes time to finish booting. If it takes more than five minutes try restarting the virtual machine.

* **A message keeps saying that name mode is in safe mode so the file cannot be created?**

After opening the Virtual Box, right click on the VM and look for "Normal Start."