Lab #3 – Hadoop

Introduction

Welcome to Lab #3.

Background

This lab introduces Hadoop ecosystem, HDFS (Hadoop Distributed File System), and MapReduce program.

Goals of Lab

- Know the basic knowledge of HDFS, and write simple MapReduce programs in Python.

Pre-requisites

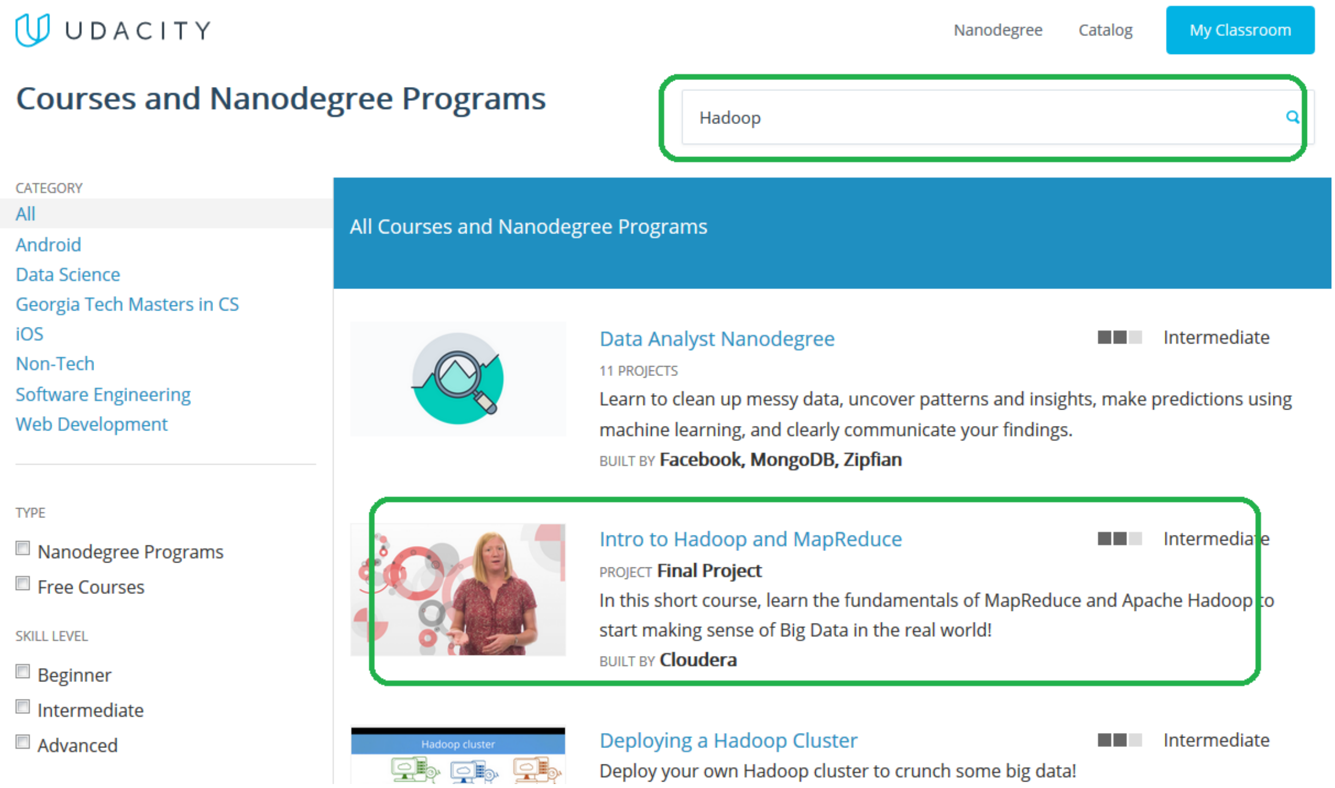
- VirtualBox

Section 1 – Introduction to Hadoop ecosystem, HDFS, and MapReduce

1. Register for the website:

https://www.udacity.com/

2. Go to Course Catalog and select "Intro to Hadoop and MapReduce"



Courses and Nanodegree Programs

Hadoop

CATEGORY

- All
- Android
- Data Science
- Georgia Tech Masters in CS
- iOS
- Non-Tech
- Software Engineering
- Web Development

TYPE

□ Nanodegree Programs

□ Free Courses

SKILL LEVEL

□ Beginner

□ Intermediate

□ Advanced

All Courses and Nanodegree Programs

Data Analyst Nanodegree ■■□ Intermediate

11 PROJECTS

Learn to clean up messy data, uncover patterns and insights, make predictions using machine learning, and clearly communicate your findings.

BUILT BY **Facebook, MongoDB, Zipfian**

Intro to Hadoop and MapReduce ■■□ Intermediate

PROJECT **Final Project**

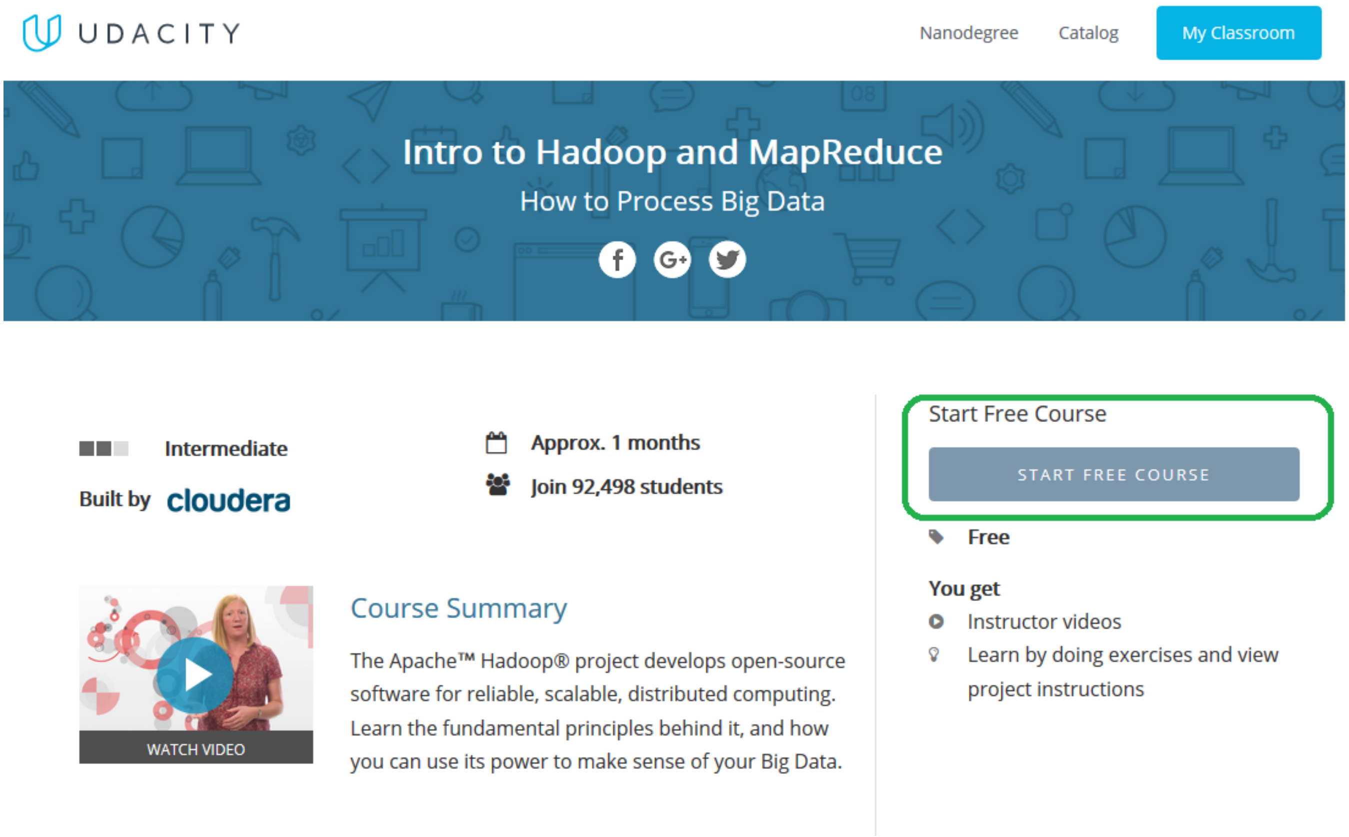
In this short course, learn the fundamentals of MapReduce and Apache Hadoop to start making sense of Big Data in the real world!

BUILT BY **Cloudera**

Deploying a Hadoop Cluster ■■□ Intermediate

Deploy your own Hadoop cluster to crunch some big data!

3. Select "Start free courseware"



UDACITY

Nanodegree Catalog My Classroom

Intro to Hadoop and MapReduce

How to Process Big Data

Intermediate

Built by cloudera

Approx. 1 months

Join 92,498 students

Start Free Course

START FREE COURSE

Free

You get

Instructor videos

Learn by doing exercises and view project instructions

WATCH VIDEO

Course Summary

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing. Learn the fundamental principles behind it, and how you can use its power to make sense of your Big Data.

4. Watch the video for **Lesson 1 Big data**, **Lesson 2 HDFS and MapReduce**, **Lesson 3 MapReduce code**
   - Make sure you understand the concept of HDFS and how data is processed with MapReduce
   - Make sure you know how to load data set from local disk to HDFS and run the MapReduce code

Section 2 – Download virtual machine and install in VirtualBox

1. Download the virtual machine package which has CDH pre-installed from: https://docs.google.com/document/d/1v0zGBZ6EHap-Smsr3x3sGGpDW-54m82kDpPKC2M6uiY/pub
2. Follow the instruction to install the VM in VirtualBox

3. Run sample code of calculating the total sales per store in VM

Homework

1. Write Python program for the following three questions of project part 1.
   1) Find a sales breakdown by product category across all of our stores.
   2) Find the monetary value for the highest individual sale for each separate store.
   3) Find the total sales value across all the stores, and the total number of sales. (Assume there is only one reducer.)

2. Take a screen shot after running MapReduce code for question 1. Copy and paste the mapper and reducer code for question 1. Copy and paste the result for question 1.

3. Take a screen shot after running MapReduce code for question 2. Copy and paste the mapper and reducer code for question 2. What are the values for the following store:

   Anchorage
   Bakersfield
   Colorado Springs

4. Take a screen shot after running MapReduce code for question 3. Copy and paste the mapper and reducer code for question 3. What is the total number of sales and the total sales value from all the stores?

Deliverables

- Create a {Microsoft Word| PDF } document containing the answers to the test section.
- Name the file <Last Name>_<First Name>_Lab03.{docx|pdf}
- Send the file to me via slack private message