Learn By Example: Hadoop, MapReduce for Big Data problems

Modality: Self-Paced Learning
Duration: 14 Hours
SUBSCRIPTION: Learn, Master, Master Plus

About this course:

This course is a zoom-in, zoom-out, hands-on workout involving Hadoop, MapReduce and the art of thinking parallel. Let’s parse that. Zoom-in, Zoom-Out: This course is both broad and deep. It covers the individual components of Hadoop in great detail, and also gives you a higher level picture of how they interact with each other. Hands-on workout involving Hadoop, MapReduce: This course will get you hands-on with Hadoop very early on. You’ll learn how to set up your own cluster using both VMs and the Cloud. All the major features of MapReduce are covered - including advanced topics like Total Sort and Secondary Sort. The art of thinking parallel: MapReduce completely changed the way people thought about processing Big Data. Breaking down any problem into parallelizable units is an art. The examples in this course will train you to "think parallel".

The average salary for Data Scientist is $120,931 per year.

Course Objective:

After completing this course, students will be able to:

- Recommend friends in a Social Networking site: Generate Top 10 friend recommendations using a Collaborative filtering algorithm.
- Build an Inverted Index for Search Engines: Use MapReduce to parallelize the humongous task of building an inverted index for a search engine.
- Generate Bigrams from text: Generate bigrams and compute their frequency distribution in a corpus of text.
- Install Hadoop in Standalone, Pseudo-Distributed and Fully Distributed modes
- Set up a hadoop cluster using Linux VMs.
- Set up a cloud Hadoop cluster on AWS with Cloudera Manager.
- Understand HDFS, MapReduce and YARN and their interaction
- Chain multiple MR jobs together
- Write your own Customized Partitioner
- Total Sort: Globally sort a large amount of data by sampling input files
- Secondary sorting
- Unit tests with MR Unit
- Integrate with Python using the Hadoop Streaming API

Audience:

This course is intended for:
Analysts who want to leverage the power of HDFS where traditional databases don't cut it anymore
Engineers who want to develop complex distributed computing applications to process lot's of data
Data Scientists who want to add MapReduce to their bag of tricks for processing data

Prerequisites:

- You'll need an IDE where you can write Java code or open the source code that's shared. IntelliJ and Eclipse are both great options.
- You'll need some background in Object-Oriented Programming, preferably in Java. All the source code is in Java and we dive right in without going into Objects, Classes etc
- A bit of exposure to Linux/Unix shells would be helpful, but it won't be a blocker

Suggested prerequisites courses:

- Object Oriented Programming in Java

Course Outline:

- Introduction
- Why is Big Data a Big Deal
- Installing Hadoop in a Local Environment
- The MapReduce "Hello World"
- Run a MapReduce Job
- Juicing your MapReduce - Combiners, Shuffle and Sort and The Streaming API
- HDFS and Yarn
- MapReduce Customizations For Finer Grained Control
- The Inverted Index, Custom Data Types for Keys, Bigram Counts and Unit Tests!
- Input and Output Formats and Customized Partitioning
- Recommendation Systems using Collaborative Filtering
- Hadoop as a Database
- K-Means Clustering
- Setting up a Hadoop Cluster
- Appendix