Learn By Example: Hadoop, MapReduce for Big Data problems

Modality: On Demand

Duration: 14 Hours

About this course:

This program is a zoom-in, zoom-out, practical training involving MapReduce, Hadoop, and the art of simultaneous thinking. Let's look at that. Zoom-in, zoom-out: This program is wide as well as deep. It describes Hadoop's components in vivid detail and also provides you a higher-level view of how they communicate. MapReduce, Hadoop, hands-on work out This training should get you to hands-on with Hadoop early on. You will discover how to use both Cloud and Virtual Machines to configure your cluster. Many of MapReduce's main features are covered-including specialized topics such as Secondary Sort and Total Sort. The art of parallel thinking: MapReduce changed the way people thought about analyzing Big Data. It is an art to break down any issue into parallel parts. This program's examples will teach you to "think parallel."

The Data Scientist can earn an average salary of $120,931 per annum.

Course Objective:

· Create a Search Engines Inverted Index: Using MapReduce to simulate the humongous task of constructing an inverted index for a browser

· Enable Hadoop in modes that are pseudo-distributed, standalone, and fully distributed

· Suggest friends on a social networking site: Using a Collaborative filtering algorithm to produce top 10 friend recommendations

· Generate Bigrams from the text: Produce bigrams and measure their frequency distribution in a text corpus

· Using Cloudera Manager to configure a cloud Hadoop cluster on Amazon Web Services

· Tie up several MR jobs

· Configure a cluster of hadoops using Linux Virtual machines

· Total Sort: Filter vast volumes of data globally by filtering input files

· Understanding YARN, MapReduce, and HDFS and how they connect

· Tests unit with MR Unit

· Writing Customized Partitioner

· Secondary sort

· Using Hadoop Streaming Application programming interface to integrate with Python

Audience:

· Engineers who want to create complex distributed data processing applications

· Analysts wishing to harness the power of HDFS where conventional databases no longer cut it

· Data scientists need to add MapReduce to their bag of data processing tricks

Prerequisites:

· You may need some experience in object-oriented programming, in Java ideally. All source code is in Java, and we dive straight into classes, objects, etc

· You will need an IDE that allows you to write Java code, or access the shared source code. Both Eclipse and IntelliJ are superb options

· A bit of access to shells in Unix/Linux will be beneficial but it would not be a blocker

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