Course Syllabus

After completing this course, you will be able to solve any data engineering and data science problem using Apache Spark. With 30.000+ commits, 1000+ contributors, nearly 1.000.000 lines of code and 200+ man years of effort, Apache Spark is the most active Apache Software Foundation project and one of the largest open source projects ever. In IBM alone, 3500 researchers and developers are working with Apache Spark and IBM calls it "potentially the most significant open source project of the next decade." You'll master Apache Spark, for BigData but also for SmallData problems. You'll be able to go beyond CPU, main memory and storage limitations by making use of large scale compute clusters (IBM provides a free Apache Spark cluster for you during the course which you can continue to use afterwards, all free of charge). You'll understand how parallel code is written, capable of running on thousands of CPUs. You'll be able to use simple SQL statements on Petabytes of data using Apache SparkSQL and the Apache Spark DataFrame API. You'll be able to explain how Tungsten and Catalyst are transforming SQL queries into cost based optimized dynamic execution graphs. A significant advantage Spark has over other state-of-the-art frameworks like TensorFlow. You'll be able to apply machine learning algorithms on Petabytes of data using Apache SparkML Pipelines. Join us to learn one of the de-facto standards in data science, successfully applied by companies like Alibaba, Apple, Amazon, Baidu, eBay, IBM, NASA, Samsung, SAP, TripAdvisor, Yahoo! and Zalando. Prerequisites: - basic python programming - basic machine learning (optional introduction videos are provided in this course as well) - basic SQL skills for optional content The following courses are recommended taking before taking this class (unless think you have the skills already) https://www.coursera.org/learn/python-for-applied-data-science or similar https://www.coursera.org/learn/machine-learning-with-python or similar https://www.coursera.org/learn/sql-data-science or similar for optional lectures

Every week has an assessment quiz. If you pass all quizzes, you'll pass the course. The quiz in week 4 is the most complex which needs you to answer concepts from the complete course. There are also a couple of practice quizzes which are not graded.

Syllabus:

- Week 1: Introduction

-- Understanding how Apache Spark works

--- What is Big Data?

--- Data storage solutions

--- Parallel data processing strategies of Apache Spark

--- Functional programming basics

--- Resilient Distributed Dataset and DataFrames - ApacheSparkSQL

- Week 2: Scaling Math for Statistics on Apache Spark

-- Experience parallel programming on Apache Spark

--- Averages

--- Standard deviation

--- Skewness

--- Kurtosis

--- Covariance, Covariance matrices, correlation

- Week 3: Introduction to Apache SparkML

-- Introduction to Apache SparkML

--- How ML Pipelines work

--- Introduction to SparkML

--- Extract - Transform - Load

-- Unsupervised Learning with Apache SparkML

--- Introduction to Clustering: k-Means

--- Using K-Means in Apache SparkML

- Week 4: Supervised and Unsupervised learning with SparkML

-- Supervised Learning with Apache SparkML

--- Linear Regression

--- LinearRegression with Apache SparkML

--- Logistic Regression

--- LogisticRegression with Apache SparkML

-- Course Project

Mark as completed