ProTech "Charting the Course ...

protechtraining.com

... to Your Success!"

Comprehensive Apache Spark 2.3 for Machine Learning and Data Science

Course Summary

Description

This three day training course for Data Scientists and Analysts will teach you how to harness Apache Spark 2.3 for large scale data analysis, predictive modeling, and machine learning tasks. You will learn how to program Spark as efficiently and effectively as possible, by targeting the latest version of the platform, and learning the modern approach necessary to fully leverage the advantages it offers.

The entirety of the course is taught hands-on, using real code and interactive examples. In addition, longer labs allow attendees to work together to apply their growing Spark knowledge to solve common challenges faced by organizations running complex Big Data applications in production.

Both lectures and lab activities use real-world datasets, so that you can practice getting Apache Spark to work well in-spite of real-world challenges. You'll also gain hands-on experience with performance tuning and troubleshooting.

Apache Spark 2 brings a suite of new features and speed improvements – but it also works differently under the hood, and requires a slightly different approach to programming in-oder to get the most out of it.

This course focuses entirely on Spark 2 and will teach you how to program for the latest version of Spark (currently Spark 2.3) in the most performant, most effective, and easiest way possible.

Objectives

By the end of this course, students will be able to:

- Program Apache Spark in the most performant, easy, modern, and effective ways possible to perform data wrangling, feature selection, model building, validation, tuning, and serving, as well as extending Spark ML to add your feature processing tools and new parallel ML algorithms.
- Apache Spark has strengths and limitations, like anything else – learn exactly what those are, so you can get the most out of Spark together with other tools.
- Learn how Apache Spark processes your jobs so that you can troubleshoot, analyze, and improve performance if they don't run well.
- See important patterns, tricks, tips, and gotchas so that you don't have to learn them the hard way.

Topics

- Introduction
- DataFrame/Dataset and SQL Analytics
- Machine Learning Overview
- Streaming Overview
- Using Apache Spark with the ML / Predictive Analytics Process
- Understanding Apache Spark Job Performance
- Additional Spark ML Algorithms and Features
- Integrating Apache Spark with Other Machine Learning Systems
- Extending Spark ML
- Apache Spark Model Deployment Patterns
- Apache Spark Cluster Deployment Overview (Optional)

Comprehensive Apache Spark 2.3 for Machine Learning and Data Science

Course Summary (cont'd)

Audience

This course is designed for data scientists or analysts involved in predictive modeling, who want to explore machine learning where data is too large for single-machine tools.

Prerequisites

There are no prerequisites for this course.

Duration

Three days

ProTech "Charting the Course ...

protechtraining.com

... to Your Success!"

Comprehensive Apache Spark 2.3 for Machine Learning and Data Science

Course Outline

I. Introduction
- A. Apache Spark Fundamentals and Background
- B. Compute Model
- C. APIs, Use Cases, and Ecosystem
- D. Core Architecture

II. DataFrame/Dataset and SQL Analytics
- A. Overview of concepts and APIs
- B. Using SQL with Spark
- C. Queries
- D. Data import/export, formats
- E. Parallelism and UI basics
- F. DataFrame operators, columns
- G. Caching
- H. Solving analytics problems with DataFrames
- I. DataFrame/Dataset vs. RDDs

III. Machine Learning Overview
- A. Understanding Spark's API Patterns
- B. Basics of Transformers, Estimators, and Pipelines
- C. Simple Linear Regression

IV. Streaming Overview
- A. Spark Structured Streaming with DataFrames
- B. Patterns and I/O Considerations

V. Using Apache Spark with the ML / Predictive Analytics Process
- A. Data Manipulation, Exploration
- B. Cleaning and Preprocessing Data
- C. Feature Selection and Transformation
- D. Modeling
- E. Evaluation
- F. Tuning and Cross-validation

VI. Understanding Apache Spark Job Performance
- A. Job execution: How does Spark use a cluster to run your jobs?
- B. Apache Spark GUI
- C. Improving execution
- D. Patterns and Anti-Patterns

VII. Additional Spark ML Algorithms and Features
- A. Classification
- B. Clustering
- C. Frequent Pattern Mining
- D. Natural Language Processing

VIII. Integrating Apache Spark with Other Machine Learning Systems
- A. Common algorithms/approaches not in Spark ML
- B. Model parallelism with scikit-learn
- C. Deep learning with external libraries such as H2O or TensorFlow

IX. Extending Spark ML
- A. Leveraging Spark with Parallel Algorithms
- B. Implementing Transformers
- C. Implementing Estimators
- D. Implementing a distributed ML algorithm

X. Apache Spark Model Deployment Patterns
- A. Persisting Models
- B. Serving models with Spark
- C. Building models with Spark and serving them outside of Spark
- D. Online / incremental learning and prediction

XI. Apache Spark Cluster Deployment Overview (Optional)
- A. Cluster manager options, pros/cons
- B. Patterns for submitting jobs to Spark or exposing Spark as a service
- C. Spark standalone clustering, YARN, and Mesos deployment
- D. Beyond standalone analytics or ETL: Integrating Spark services into your architecture