ProTech protechtraining.com *"Charting the Course ... ... to Your Success!"*

Comprehensive Programming for Apache Spark 2.3

Course Summary

Description

This three day training course will teach you how to harness Apache Spark 2.3 for large scale data analysis, building big data applications and data processing pipelines. You will learn how to program Spark as efficiently and effectively as possible, by targeting the latest version of the platform (Spark 2.3), 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.

While we're enthusiastic about many of the products in the Big Data ecosystem, the focus of this training course is to make you as proficient and effective as possible with open source Apache Spark, enabling you to apply the fundamental skills gained to whichever products and tools work best for you.

Targeting the latest version of the Spark platform, Apache Spark 2.3, will teach you how to optimize your Spark code to fully leverage the internal changes that make Spark 2.3 faster and more effective. At the same time, this training course will help prepare you for the future of the platform, by teaching you the modern approach to Spark programming required by future releases of the platform.

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 ETL, analytics, machine learning, and streaming operations.
- Understand how Spark should – and shouldn't! – be used within your Big Data application architectures.
- 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

- Intro
- DataFrame/Dataset and SQL Analytics
- Machine Learning Overview
- Streaming Overview
- RDDs and Deep Dive Part 1
- Catalyst/Tungsten and Deep Dive Part 2
- Deployment Overview
- Apache Spark Streaming in Depth
- Machine Learning

Audience

Data analysts, engineers, and scientists who want to conduct analytics with Big Data or build end-to-end applications and data processing pipelines.

Prerequisites

Before taking this course, attendees should have some knowledge of SQL and some background programming in Python, Java, Scala, or R.

Duration

Three days

ProTech "Charting the Course ...

protechtraining.com

... to Your Success!"

Comprehensive Programming for Apache Spark 2.3

Course Outline

I. Intro
- A. Apache Spark Fundamentals and Background
- B. Compute Model
- C. APIs, Use Cases, and Ecosystem
- D. Differences from MapReduce
- E. 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. Hive integration (optional)
- I. Solving analytics problems with DataFrames
- J. DataFrame/Dataset vs. RDDs

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

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

V. RDDs and Deep Dive Part 1
- A. RDD concept, partitioning, APIs
- B. Caching and persistence
- C. DAG and control flow
- D. Job execution: How does Spark use a cluster to run your jobs?
- E. Performance/Troubleshooting: Is my job running well? Improving execution

VI. Catalyst/Tungsten and Deep Dive Part 2
- A. Apache Spark's query optimizer
- B. Encoders and native memory
- C. How Apache Spark converts Dataset operations to RDD/DAG jobs
- D. Understanding the Jobs, Stages, and Tasks of DataFrame/SQL execution
- E. Performance Optimizations (and Gotchas) of Spark
- F. Broadcasts and Broadcast Joins

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

VIII. Apache Spark Streaming in Depth
- A. Streaming processing models: Receiver-based, Receiverless, Structured Streaming (2.x)
- B. APIs for Streaming Logic
- C. Streaming Integration Patterns
- D. Monitoring and Tuning for Streaming
- E. Reliability and Recovery for "Always-On" Apps

IX. Machine Learning
- A. General explanation of predictive analytics / ML (optional)
- B. SparkML feature coverage and integration with other ML toolkits/systems
- C. In-Depth ML Pipelines API examples
- D. Training models
- E. Evaluating models
- F. Tuning models (cross-validation, hyperparameter search)
- G. Deploying models to production

X. Spark Architecture Patterns and Newest Features
- A. A. Inside Spark Streaming Micro-Batch vs. Continuous (Millisecond-Latency) Modes
- B. B. High Performance Python Integration with Arrow
- C. C. Best Practices for Performance in the Latest Spark Versions
- D. D. Spark-as-a-Service
- E. E. Multiuser / Multiapp Spark Deployments