Introduction to Spark for R Experienced Data Scientists and Analysts

Getting Started with Spark for Experienced Data Scientists Already Working with R

www.triveratech.com

Course Snapshot

- **Course: TTSK7516: Introduction to Spark Programming for R Experienced Data Scientists & Analysts**
- **Duration**: 2 days
- **Skill Level:** Introductory topics for intermediate skilled Data Scientists already fluent in data science techniques in other languages such as SAS and already comfortable with R.
- **Hands-on Learning:** This course is approximately **50% hands-on**, combining expert lecture, real-world demonstrations and group discussions with machine-based practical labs and exercises. Student machines are required.
- **Delivery Options**: This course is available for **onsite private classroom presentation**, **live online virtual presentation**, or can be presented in a **flexible blended learning format** for combined onsite and remote attendees. Please also ask about our **Self-Paced / Video / QuickSkills or Mini-Camp Flex Hours / Short Course** options.
- **Public Schedule**: This course has active dates on our live-online open enrollment **Public Schedule**.
- **Customizable**: This course agenda, topics and labs can be further adjusted to target your specific training skills objectives, tools and learning goals. Please inquire for details.

Overview

Spark is a highly optimized Data Science environment running on Hadoop YARN, with support for Machine Learning through MLib and Mahout, SQL, DataFrames, and Streaming. In this course, Data Scientists dive into the details of practical data science on the Spark platform, including real-world interaction with other systems in modern Data Science environments.

Introduction to Spark for R Experienced Data Scientists & Analysts is intended for existing Data Scientists already fluent in data science techniques in other languages such as SAS and already comfortable with R. This course will be presented in a "rolling lab" approach - a continuous workshop of real-world data exploration involving real-world problems. As such, problems and opportunities will be explored as data suggests and as questions arise. "Lecture" material will be provided only as is necessary to explain the background of the approach being used at the moment. Times and ordering of the material are highly flexible and should be used only as estimates. Student questions and requests will also significantly alter the direction of the workshop.

The objective of the course is to practically transition these data scientists to the R/Spark/Hadoop environment, becoming comfortable with the tools and machine learning libraries and conduct statistical and machine learning analyses they've already been performing in SAS or similar environments.

Learning Objectives

This course is approximately **50% hands-on**, combining expert lecture, real-world demonstrations and group discussions with machine-based practical labs and exercises. The objective of the course is to practically transition these data scientists to the R/Spark/Hadoop environment, becoming comfortable with the tools and machine learning libraries and conduct statistical and machine learning analyses they've already been performing in SAS or similar environments.

Audience & Pre-Requisites

This course is intended for existing Data Scientists already fluent in data science techniques in other languages such as SAS and already comfortable with R.

Take Before: Incoming students should have skills equivalent to the topics in, or should have recently attended, this course as a pre-reqsuite:

- TTDS6682 Introduction to R Programming for Data Scientists

Enhanced Learning Services: Please also ask about our **Pre-Training Class OnRamp & Prep / Primer** offerings, **Skills Gap Assessment Services**, **Case Studies**, **Knowledge Check Quizzes**, **Skills Immersion Programs & Camps**, **Collaborative Mentoring Services** and **Extended Learning Support** & **Post Training** services.

Course Topics / Agenda

Please note that this list of topics is based on our standard course offering, evolved from typical industry uses and trends. We'll work with you to tune this course and level of coverage to target the skills you need most. Topics, agenda and labs may adjust during live delivery based on audience skill-level, needs and participation.

Getting Started - Overview
- Our Data and our problem set
- Accessing the cluster, the data, and the tools
- The Continuous Workshop approach
- "Let's build a model together"
- Focus on analysis, exploration, data munging, algorithms
- Tooling and fundamentals as necessary to get the job done

Spark Overview
- Data Science: The State of the Art
- Hadoop, Yarn, and Spark
- Architectural Overview
- MLib Overview
- HDFS data - Accessing
- Lab Focus
- Working with HDFS data
- Distributed vs. Local Run Modes
- Spark vs. Other tools (when is Spark the right tool for the job?)
- Spark vs. SAS
- Spark Languages (Java, R, Python, and Scala)
- Hello, Spark

Spark Overview
- Spark Core
- Spark SQL
- Spark and Hive
- Lab
- MLib
- Spark Streaming
- Spark API

DataFrames
- DataFrames and Resilient Distributed Datasets (RDDs)
- Partitions
- Adding variables to a DataFrame
- DataFrame Types
- DataFrame Operations
- Dependent vs. Independent variables
- Map/Reduce with DataFrames

Spark SQL
- Spark SQL Overview
- Data stores: HDFS, Cassandra, HBase, Hive, and S3
- Table Definitions
- Queries

Spark MLib
- MLib overview
- MLib Algorithms Overview
- Classification Algorithms
- Regression Algorithms
- Lab Focus
- Brief Comparison to SAS
- Here's your split, how to tune regression
- Decision Trees and forests
- Lab Focus
- Brief Comparison to SAS
- Stepwise approach to Decision Trees
- Working with Exit Criteria
- Recommendation with ALS
- Clustering Algorithms
- Lab Focus
- Key Clustering Algorithms
- Choosing Clustering Algorithms
- Working with key algorithms
- Machine Learning Pipelines
- Linear Algebra (SVD, PCA)
- Statistics in MLib

Spark Streaming
- Streaming overview

Streaming with Kafka
- Kafka overview
- Kafka and Spark Streaming

Data Flow with NiFi
- Apache NiFi overview
- NiFi data flows with Spark/R

Cluster Mode
- Standalone Cluster
- Masters and Workers

Spark - the Big Picture
- Spark in Real-Time and near-Real-Time Decision Support Systems
- Spark in the Enterprise
- Best Practices

Student Materials: Each participant will receive a **Student Guide** with course notes, code samples, software tutorials, step-by-step written lab instructions, diagrams and related reference materials and resource links. Students will also receive the project files (or code, if applicable) and solutions required for the hands-on work.

Hands-On Setup Made Simple! Our dedicated tech team will work with you to ensure our 'easy-access' cloud-based course environment is accessible, fully-tested and verified as ready to go well in advance of the course start date, ensuring a smooth start to class and effective learning experience for all participants. Please inquire for details and options.

For More Information

All courses can be presented **onsite** or **online,** or in a **combined / flex / blended learning format**, tailored to target your specific audience, needs and learning goals. We also offer focused, flexible **short courses, self-paced learning** options, **recorded sessions** and more. We train beginner to advanced skills in all areas we cover, and offer **New Hire / Cohort Training**, **Boot Camps**, **Skills Immersion Programs**, **Reskilling Programs**, **Skills Migration & Transition Programs**, and more. We collaborate with you to ensure all courses are truly targeted to meet your specific needs and learning skills, maximizing your valuable training time, as well as your important budget.

Please also visit our extensive **Public Training Schedule** for training for smaller groups or individuals. Please contact us for course details, **Corporate Rates** and **Special Discount Offers**.

For more information about our dedicated training services, collaborative coaching services, courseware licensing and development services, public course schedule, training management services, partner programs, or to see our complete list of course offerings and special offers please visit us at **www.triveratech.com**, email **Info@triveratech.com** or call us toll free at **844-475-4559.** Our pricing and services are always satisfaction guaranteed.

TRIVERA TECHNOLOGIES ● Collaborative IT Training, Coaching & Skills Development Solutions
www.triveratech.com ● toll free +1-844-475-4559 ● Info@triveratech.com ● Twitter TriveraTech

ONSITE, ONLINE & BLENDED TRAINING SOLUTIONS ● PUBLIC / OPEN ENROLLMENT COURSES ● COURSEWARE LICENSING & DEVELOPMENT
MENTORING ● ASSESSMENTS ● LEARNING PLAN DEVELOPMENT ● SKILLS IMMERSION PROGRAMS / RESKILLING / NEW HIRE / BOOT CAMPS
PARTNER & RESELLER PROGRAMS ● CORPORATE TRAINING MANAGEMENT ● VENDOR MANAGEMENT SERVICES

Trivera Technologies is a Woman-Owned Small-Business Firm

Explore Trivera's Ways to Learn…

TECHNOLOGY MiniCamps
everyCourse EXTRAS
quickSkills FREE COURSES AND WEBINARS
afterCourse COACHING AND SUPPORT