Spark Developer | Spark for Big Data, Hadoop & Machine Learning - TTSK7503

Explore Spark Essentials | Ecosystem, Data Structures, Spark SQL, APIs, RRDs, Hadoop, Streaming and More

Duration: 3 Days
Skill Level: Introductory
Available Format: Instructor-Led Online; Instructor-Led, Onsite In Person ; Blended; On Public Schedule

Apache Spark, a significant component in the Hadoop Ecosystem, is a cluster computing engine used in Big Data. Building on top of the Hadoop YARN and HDFS ecosystem, it offers order-of-magnitude faster processing for many in-memory computing tasks compared to Map/Reduce.

What You'll Learn

Overview

Apache Spark, a significant component in the Hadoop Ecosystem, is a cluster computing engine used in Big Data. Building on top of the Hadoop YARN and HDFS ecosystem, it offers order-of-magnitude faster processing for many in-memory computing tasks compared to Map/Reduce. It can be programmed in Java, Scala, Python, and R - the favorite languages of Data Scientists - along with SQL-based front ends. With advanced libraries like Mahout and MLib for Machine Learning, GraphX or Neo4J for rich data graph processing as well as access to other NOSQL data stores, Rule engines and other Enterprise components, Spark is a lynchpin in modern Big Data and Data Science computing.

Geared for experienced developers**, Spark Developer | Introduction to Spark for Big Data, Hadoop & Machine Learning** provides students with a comprehensive, hands-

on exploration of enterprise-grade Spark programming, interacting with the significant components mentioned above to craft complete data science solutions. Students will leave this course armed with the skills they require to begin working with Spark in a practical, real world environment.

This course is offered in support of the Python programming language but can also be offered for R or Java with advance notice and planning. Our team will work with you to coordinate the languages, tools and environment that will work best for your organization and needs. Please inquire for details.

Objectives

This "skills-centric" course is about **50% hands-on lab and 50% lecture,** designed to train attendees in core big data/ Spark development and use skills, coupling the most current, effective techniques with the soundest industry practices. Throughout the course students will be led through a series of progressively advanced topics, where each topic consists of lecture, group discussion, comprehensive hands-on lab exercises, and lab review.

This course provides indoctrination in the practical use of the umbrella of technologies that are on the leading edge of data science development focused on Spark and related tools. Working in a hands-on learning environment, students will explore:

- Spark Ecosystem
- Spark Shell
- Spark Data structures (RDD, DataFrame, Dataset)
- Spark SQL
- Modern data formats and Spark
- Spark API
- Spark & Hadoop & Hive
- Spark ML overview
- GraphX
- Time-permitting: Spark Streaming
- Time-permitting: Optional Capstone Workshop (Time-Permitting)

Need different skills or topics? If your team requires different topics or tools, additional skills or custom approach, this course may be further adjusted to accommodate. We offer additional programming, Python / R, Spark, AI/Machine Learning/Deep Learning, data science, analytics and other related topics that may be blended with this course for a track that best suits your needs. Our team will collaborate with you to understand your needs and will target the course to focus on your specific learning objectives and goals.

Audience

This foundation-level course is geared for intermediate skilled, experienced Developers and Architects (with basic Python experience) who seek to be proficient in advanced, modern development skills working with Apache Spark in an enterprise data environment.

Pre-Requisites

This foundation-level course is geared for intermediate skilled, experienced Developers and Architects (with basic Python experience) who seek to be proficient in advanced, modern development skills working with Apache Spark in an enterprise data environment.

Take Before: Students should have attended the course(s) below, or should have basic skills in these areas:

- **TTPS4800** Introduction to Python Programming

TTPS4800	Introduction to Python Programming Basics

Agenda

Please note that this list of topics is based on our standard course offering, evolved from typical industry uses and trends. We will work with you to tune this course and level of coverage to target the skills you need most. Course agenda, topics and labs are subject to adjust during live delivery in response to student skill level, interests and participation.

Spark Introduction

- Big data, Hadoop, Spark
- Spark concepts and architecture
- Spark components overview
- Labs: installing and running Spark

The first look at Spark

- Spark shell
- Spark web UIs
- Analyzing dataset – part 1
- Labs: Spark shell exploration

Spark Data structures

- Partitions
- Distributed execution
- Operations: transformations and actions
- Labs: Unstructured data analytics using RDDs

Caching

- Caching overview
- Various caching mechanisms available in Spark
- In memory file systems
- Caching use cases and best practices
- Labs: Benchmark of caching performance

DataFrames and Datasets

- DataFrames Intro
- Loading structured data (JSON, CSV) using DataFrames
- Using schema
- Specifying schema for DataFrames

- Labs: DataFrames, Datasets, Schema

Spark SQL

- Spark SQL concepts and overview
- Defining tables and importing datasets
- Querying data using SQL
- Handling various storage formats: JSON, Parquet, ORC
- Labs: querying structured data using SQL; evaluating data formats

Spark and Hadoop

- Hadoop Primer: HDFS, YARN
- Hadoop + Spark architecture
- Running Spark on Hadoop YARN
- Processing HDFS files using Spark
- Spark & Hive

Spark API

- Overview of Spark APIs in Scala / Python
- The lifecycle of a Spark application
- Spark APIs
- Deploying Spark applications on YARN
- Labs: Developing and deploying a Spark application

Spark ML Overview

- Machine Learning primer
- Machine Learning in Spark: MLib / ML
- Spark ML overview (newer Spark2 version)
- Algorithms overview: Clustering, Classifications, Recommendations
- Labs: Writing ML applications in Spark

GraphX

- GraphX library overview
- GraphX APIs
- Create a Graph and navigating it
- Shortest distance
- Pregel API
- Labs: Processing graph data using Spark

Time Permitting Topics

Spark Streaming

- Streaming concepts
- Evaluating Streaming platforms
- Spark streaming library overview
- Streaming operations
- Sliding window operations
- Structured Streaming
- Continuous streaming
- Spark & Kafka streaming
- Labs: Writing spark streaming applications

Workshop

- Attendees will work on solving real-world data analysis problems using Spark

Related Courses

TTSK7520	Mastering Scala with Apache Spark for the Modern Data Enterprise

TTSK7503	Spark Developer | Spark for Big Data, Hadoop & Machine Learning

All course software (limited versions, for course use only), digital courseware files or course notes, labs / data sets and solutions (as applicable) are provided for you in our "easy access / no install required" high-speed remote lab environment. Our tech team works with every student to ensure everyone is set up with working access and ready to go prior to every course start date, ensuring a smooth delivery and great hands-on experience. Please ask for details.

For More Information

Please contact us or call 844-475-4559 toll free for more information about our training services (instructor-led, self-paced or blended), coaching and mentoring services, public course enrollment or questions, partner programs, courseware licensing options and more.