Mastering Scala with Apache Spark for the Modern Data Enterprise - TTSK7520

Boost your big data expertise with essential skills in Scala, Apache Spark, MLlib, GraphX, and cutting-edge generative AI technologies.

Duration: 5 Days
Skill Level: Intermediate
Available Format: Instructor-Led Online; Instructor-Led, Onsite In Person ; Blended; On Public Schedule

Embark on a journey to master the world of big data with our immersive course on Scala and Spark! **Mastering Scala with Apache Spark for the Modern Data Enterprise** is a five day hands-on course designed to provide you with the essential skills and tools to tackle complex data projects using Scala programming language and Apache Spark, a high-performance data processing engine. Mastering these technologies will enable you to perform a wide range of tasks, from data wrangling and analytics to machine learning and artificial intelligence, across various industries and applications.

What You'll Learn

Overview

Embark on a journey to master the world of big data with our immersive course on Scala and Spark! Mastering Scala with Apache Spark for the Modern Data Enterprise is a five day hands-on course designed to provide you with the essential skills and tools to tackle complex data projects using Scala programming language and Apache Spark, a high-performance data processing engine. Mastering these technologies will enable you to perform a wide range of tasks, from data wrangling and analytics to machine learning and artificial intelligence, across various industries and applications.

Guided by our expert instructor, you'll explore the fundamentals of **Scala programming and Apache Spark** while gaining valuable hands-on experience with Spark programming, RDDs, DataFrames, Spark SQL, and data sources. You'll also explore Spark Streaming, performance optimization techniques, and the integration of popular external libraries, tools, and cloud platforms like AWS, Azure, and GCP. Machine learning enthusiasts will delve into Spark MLlib, covering basics of machine learning algorithms, data preparation, feature extraction, and various techniques such as regression, classification, clustering, and recommendation systems.

You'll also gain experience working with graph processing using Spark GraphX, as well as innovative generative AI technologies, integrating GPT with Spark and Scala for practical applications. Time permitting, you will also be introduced to Spark NLP, covering text preprocessing, classification, and sentiment analysis. With a focus on practical skills and best practices, you'll work on interesting learning objectives and gain hands-on experience with innovative tools in a live, interactive environment.

Upon completing this course, you'll be ready to confidently apply your newly acquired Scala and Apache Spark skills to a wide range of projects. You'll be able to develop efficient and scalable applications, harness the power of machine learning, and analyze large datasets, giving you a competitive edge in the rapidly evolving world of big data and analytics. By integrating these technologies into your daily work, you'll be better prepared to solve complex problems, streamline processes, and ultimately drive value for your organization.

Objectives

Working in a hands-on learning environment led by our expert instructor you'll:

- Develop a basic understanding of Scala and Apache Spark fundamentals, enabling you to confidently create scalable and high-performance applications.
- Learn how to process large datasets efficiently, helping you handle complex data challenges and make data-driven decisions.
- Gain hands-on experience with real-time data streaming, allowing you to manage and analyze data as it flows into your applications.
- Acquire practical knowledge of machine learning algorithms using Spark MLlib, empowering you to create intelligent applications and uncover hidden insights.

- Master graph processing with GraphX, enabling you to analyze and visualize complex relationships in your data.
- Discover generative AI technologies using GPT with Spark and Scala, opening up new possibilities for automating content generation and enhancing data analysis.

If your team requires different topics, additional skills or a custom approach, our team will collaborate with you to adjust the course to focus on your specific learning objectives and goals.

Audience

This **intermediate and beyond** level course is geared for experienced technical professionals in various roles, such as developers, data analysts, data engineers, software engineers, and machine learning engineers who want to leverage Scala and Spark to tackle complex data challenges and develop scalable, high-performance applications across diverse domains. Practical programming experience is required to participate in the hands-on labs.

Pre-Requisites

In order to be successful in this course you should possess:

- Basic understanding of Java programming: Familiarity with Java syntax, data structures, and concepts, such as variables, loops, and conditionals.
- Fundamental knowledge of object-oriented programming (OOP): Experience with OOP principles, such as inheritance, encapsulation, and polymorphism, in any programming language.
- Familiarity with data structures and algorithms: A basic grasp of common data structures, such as arrays, lists, and maps, as well as an understanding of simple algorithms, like sorting and searching.
- Experience with distributed systems: Basic awareness of distributed computing concepts, such as data partitioning, parallel processing, and fault tolerance.
- Basic knowledge of databases: Understanding of database concepts, including data storage, querying, and manipulation using SQL or NoSQL databases.

Next Steps / Follow-on Courses: We offer a wide varietyof follow-on courses for next-level Spark, Scala, programming, AI / Generative AI / GPT, LLMs, machine learning, deep learning, data science skills and more. Please see our **AI & Machine Learning Courses, Learning Journeys & Skills Roadmaps** for options based on your specific role and goals.

TTSCL2104 Fast Track to Scala Programming for OO / Java Developers

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 are subject to change, and may adjust during live delivery based on audience skill level, interests and participation.

Getting Started with Scala and Spark

Introduction to Scala

- Brief history and motivation
- Differences between Scala and Java
- Basic Scala syntax and constructs
- Scala's functional programming features

Introduction to Apache Spark

- Overview and history
- Spark components and architecture
- Spark ecosystem
- Comparing Spark with other big data frameworks
- Lab: Practice basic Scala syntax and functional programming concepts using the REPL.
- Setting up the Development Environment

Basics of Spark Programming SparkContext and SparkSession

- Resilient Distributed Datasets (RDDs)
- Transformations and Actions
- Working with DataFrames

Spark SQL and Data Sources

- Spark SQL library and its advantages
- Structured and semi-structured data sources
- Reading and writing data in various formats (CSV, JSON, Parquet, Avro, etc.)
- Data manipulation using SQL queries
- Lab: Setting up the Environment and Running a Simple Spark Application.
- Lab: Load and query data from different data sources using Spark SQL.

Data Processing and Spark Programming

Basic RDD Operations

- Creating and manipulating RDDs
- Common transformations and actions on RDDs
- Working with key-value data

Basic DataFrame and Dataset Operations

- Creating and manipulating DataFrames and Datasets
- Column operations and functions
- Filtering, sorting, and aggregating data
- Lab: RDD and DataFrame Operations

Introduction to Spark Streaming

- Overview of Spark Streaming
- Discretized Stream (DStream) operations
- Windowed operations and stateful processing

Performance Optimization Basics

- Best practices for efficient Spark code
- Broadcast variables and accumulators

- Monitoring Spark applications

Integrating External Libraries and Tools, Spark Streaming

- Using popular external libraries, such as Hadoop and HBase
- Integrating with cloud platforms: AWS, Azure, GCP
- Connecting to data storage systems: HDFS, S3, Cassandra, etc.
- Lab: Building an End-to-End Spark Application: Create a Spark application to process a large dataset, perform aggregations, and save the results.
- Lab: Implement a simple Spark Streaming application to process real-time data.

Machine Learning Basics with Spark MLlib

Introduction to Machine Learning Basics

- Overview of machine learning
- Supervised and unsupervised learning
- Common algorithms and use cases

Introduction to Spark MLlib

- Overview of Spark MLlib
- MLlib's algorithms and utilities
- Data preparation and feature extraction
- Lab: Data Preparation with Spark MLlib

Linear Regression and Classification

- Linear regression algorithm
- Logistic regression for classification
- Model evaluation and performance metrics

Clustering Algorithms

- Overview of clustering algorithms
- K-means clustering
- Model evaluation and performance metrics

Collaborative Filtering and Recommendation Systems

- Overview of recommendation systems
- Collaborative filtering techniques
- Implementing recommendations with Spark MLlib
- Lab: Basic Machine Learning with Spark MLlib
- Lab: Implementing a Recommendation System

Graph Processing and Generative AI Technologies

Introduction to Graph Processing

- Overview of graph processing
- Use cases and applications of graph processing
- Graph representations and operations
- Introduction to Spark GraphX
- Overview of GraphX
- Creating and transforming graphs
- Graph algorithms in GraphX
- Lab: Graph Processing with GraphX

Big Data Innovation! Using GPT and Generative AI Technologies with Spark and Scala

- Overview of generative AI technologies
- Integrating GPT with Spark and Scala
- Practical applications and use cases

Bonus Topics / Time Permitting

Introduction to Spark NLP

- Overview of Spark NLP
- Preprocessing text data
- Text classification and sentiment analysis
- Lab: Generative AI Technologies and Spark NLP: Integrate GPT for text generation in a Spark application and explore basic NLP tasks using Spark NLP.
- Lab: Text Classification with Spark NLP

Putting It All Together

- Work on a capstone project that integrates multiple aspects of the course, including data processing, machine learning, graph processing, and generative AI technologies.

Related Courses

TTSCL2104	Fast Track to Scala Programming for OO / Java Developers
TTSK7503	Spark Developer | Spark for Big Data, Hadoop & Machine Learning

All applicable course software, digital courseware files or course notes, labs, data sets and solutions, live coaching support channels and rich extended learning and post training resources are provided for you in our “easy access, no install required” high-speed **SkillJourneys™ Learning Experience Platform (LXP),** remote lab and content environment. Course materials, software, resources and post-training platform access periods vary by course.

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.