Lightbend Apache Spark for Scala - Professional

Duration 2 day(s) (APACHE-SPARK-02)

Workshop for Developer

Official Training

Description

This two-day workshop is designed to teach developers how to implement data analytics using Apache Spark for Reactive applications. In this workshop, developers will use hands-on exercises to learn the principles of Apache Spark programming and idioms for specific problems, such as event stream processing, SQL-based analysis on structured data in files, integration with Reactive frameworks like Akka, as well as Hadoop and related tools, and advanced analytics such as machine learning and graph algorithms.

Goals

- Understand how to use the Spark Scala APIs to implement various data analytics algorithms for offline (batchmode) and eventstreaming applications
- Understand Spark internals
- Understand Spark performance considerations
- Understand how to test and deploy Spark applications
- Understand the basics of integrating Spark with Mesos, Hadoop, and Akka

Public

- Developers with basic knowledge of Scala, as covered in “Lightbend Scala Language - Professional”
- Developers with an interest in data science looking to put theory into high-scale practice
- Managers who want to understand how to field applications powered by fast data analytics

Prerequisites

- Experience with Scala, such as completion of Fast Track to Scala course
- Experience with SQL, machine learning, and other Big Data tools will be helpful, but not required.

Structure

50% Theory, 50% Practice

Program

Introduction Why Spark

- How Spark improves on Hadoop MapReduce
- The core abstractions in Spark
- What happens during a Spark job?
- The Spark ecosystem
- Deployment options
- References for more information

Spark's Core API

- Resilient Distributed Datasets (RDD) and how they implement your job
- Using the Spark Shell (interpreter) vs submitting Spark batch jobs
- Using the Spark web console.
- Reading and writing data files
- Working with structured and unstructured data
- Building data transformation pipelines
- Spark under the hood: caching, checkpointing, partitioning, shuffling,etc.
- Mastering the RDD API
- Broadcast variables, accumulators

Spark SQL and DataFrames

- Working with the DataFrame API for structured data
- Working with SQL
- Performance optimizations
- Support for JSON and Parquet formats
- Integration with Hadoop Hive

Processing events with Spark Streaming:

- Working with time slices, “minibatches”, of events
- Working with moving windows of minibatches
- Reuse of code in batchmode and streaming: the Lambda Architecture
- Working with different streaming sources: sockets, file systems, Kafka,etc.
- Resiliency and fault tolerance considerations
- Stateful transformations (e.g., running statistics)

Other Sparkbased Libraries:

- MLlib for machine learning
- Discussion of GraphX for graph algorithms, Tachyon for distributed caching, and BlinkDB for approximate queries

Deploying to clusters:

- Spark’s clustering abstractions: cluster vs. client deployments, coarsegrained and finegrained process management
- Standalone mode
- Mesos
- Hadoop YARN
- EC2
- Cassandra rings

Using Spark with the Lightbend Reactive Platform:

- Akka Streams and Spark Streaming

Conclusions