knóldus

Deep Dive Into Spark in 5 Days

Developer to Developer Training Series

www.knoldus.com

Course Outline

Apache Spark is the nextgeneration successor to MapReduce. Spark is a powerful, opensource processing engine for data in the Hadoop cluster, optimized for speed, ease of use,and sophisticated analytics. The Spark framework supports streaming data processing and complex, iterative algorithms, enabling applications to run up to 100x faster than traditional Hadoop MapReduce programs.

The 5 day Spark course is aimed at developers who are encountering Spark for the first time and want to understand how to build Big Data Products with Spark. The course would enable participants to build complete, unified Big Data applications combining batch, streaming, and interactive analytics on all their data.

Developers would be able to write sophisticated parallel applications to execute faster decisions, better decisions, and realtime actions, applied to a wide variety of use cases, architectures, and industries.

The course has a practical focus, mixing presentation with indepth handson labs and exercises.

Proposed Structure

Day 1	Day 2
First brush	**RDDs**
◆ Big Data Why and What?	◆ RDD API In Detail
◆ Introduction to Spark	◆ Types of RDD (Pair RDD, Numeric RDD, JDBC RDD, KeyValue etc)
◆ Spark Installation and Modes of Operation	◆ Creating RDD From Different File Formats (Parquet, Avro, JSON, JDBC)
◆ Spark shell	
Programming with Spark	**Parallel Programming**
◆ Spark Fundamentals	◆ Partitions and Data Locality
◆ Role of Spark Context	◆ Executing parallel operations

- Map Reduce in Spark

RDD Fundamentals

- Transformations in RDD
- Actions in RDD

Caching and Persistence

- Caching Overview
- Distributed Persistence

Day 3

Advanced Concepts of RDD

- Accumulators and Broadcast Variables
- RDD Internals
- RDD Lineage

Spark SQL

- Overview
- Role of SQLContext
- Running Spark SQL in Spark shell

DataFrames

- Introduction to Data Frames
- Creating Data Frames
- Transformations and Operations on Data Frames
- Interoperating with RDDs

Day 4

Spark Streaming

- Overview
- Role of StreamingContext
- Receivers
- Streaming Applications

Datasets

- Overview
- Creating Datasets
- Difference between Data Frames and Data Sets.
- Conversion from Data Frame to Dataset and vice versa.

DStreams

- Introduction
- Operations in DStreams
- Sliding Window Operations
- Performance Tuning of DStreams
- Stateful and Stateless Transformations in DStreams

Spark Schedulers

- Overview
- Scheduling Across Applications
- Scheduling Within Application

Spark MLLib

- Data Types
- Basic Statistics
- Classification
- Clustering
- Pipelining

Day 5

Clustering

- Standalone
- Configuration

Monitoring

- Web UI
- REST API

Tuning and Debugging

- Data Serialization
- Memory Management
- Broadcasting Large Variables

Security

- Event Logging
- Encryption
- SSL Configuration
- Standalone mode

Deployment

- Submitting Applications
- Spark Standalone
- Amazon EC2
- Logging

Q&A

Course Prerequisites

To benefit from this course you should have programming experience with Scala or with Python. The language of instruction is Scala. Basic Linux knowledge is expected.

- For more information on the course or a discussion on your custom need, send a mail to info@knoldus.com