# Spark Developer (DM500)

24 Hours

## Outline

Apache Spark is the next-generation successor to MapReduce. Spark is a powerful, open-source 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 course enables participants to build complete, unified big data applications combining batch, streaming, and interactive analytics on all their data. With Spark, developers can write sophisticated parallel applications to execute faster decisions, better decisions, and real-time actions, applied to a wide variety of use cases, architectures, and industries

## Prerequisites

- Familiarity with basic concepts of the Hadoop Eco-system
- Experience with one of the modern programming languages

## Contents

## Module 4: Writing and Deploying Spark Applications

- Spark Applications vs. Spark Shell
- Creating the SparkContext
- Building a Spark Application (Scala and Java)
- Running a Spark Application
- The Spark Application Web U
- Hands-On Exercise: Write and Run  a Spark Application
- Configuring Spark Properties
- Logging
- Exercise

## Module 5: Parallel Processing

- Review: Spark on a Cluster
- RDD Partitions
- Partitioning of File-based RDDs
- HDFS and Data Locality
- Executing Parallel Operations
- Stages and Tasks
- Exercise

## Module 6: Spark RDD Persistence

- RDD Lineage
- RDD Persistence Overview
- Distributed Persistence

## Module 7: Basic Spark Streaming

- Spark Streaming Overview
- Example: Streaming Request Count
- DStreams
- Developing Spark Streaming Applications
- Exercise

## Module 8: Advanced Spark Streaming

- Multi-Batch Operations
- State Operations
- Sliding Window Operations
- Advanced Data Sources

## Module 9: Common Patterns in Spark Data Processing

- Common Spark Use Cases
- Iterative Algorithms in Spark
- Graph Processing and Analysis
- Machine Learning
- Example: k-mean

## Module 10: Improving Spark Performance

- Shared Variables: Broadcast Variables
- Shared Variables: Accumulators
- Common Performance Issues
- Diagnosing Performance Problems

## Module 11: Spark SQL and DataFrames

- Spark SQL and the SQL Context
- Creating DataFrames
- Transforming and Querying DataFrames
- Saving DataFrames
- DataFrames and RDDs
- Comparing Spark SQL, Impala and Hive-on-Spark
- Exercise