cloudera®

> "Cloudera University was by far the most well-executed technical training I have attended. I feel confident that I can build my own big data application with an enterprise data hub, and I look forward to using the tools I learned in the classroom."
>
> PricewaterhouseCoopers

Cloudera Developer Training for Apache Spark

Take Your Knowledge to the Next Level and Solve Real-World Problems with Training for Hadoop and the Enterprise Data Hub

Cloudera University's three-day training course for Apache Spark 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.

Advance Your Ecosystem Expertise

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.

Hands-On Hadoop

Through instructor-led discussion and interactive, hands-on exercises, participants will navigate the Hadoop ecosystem, learning topics such as:

- Using the Spark shell for interactive data analysis
- The features of Spark's Resilient Distributed Datasets
- How Spark runs on a cluster
- Parallel programming with Spark
- Writing Spark applications
- Processing streaming data with Spark

Audience & Prerequisites

This course is best suited to developers and engineers. Course examples and exercises are presented in Python and Scala, so knowledge of one of these programming languages is required. Basic knowledge of Linux is assumed. Prior knowledge of Hadoop is not required.

Course Outline: Cloudera Developer Training for Apache Spark

Introduction

Why Spark?

- Problems with Traditional Large-Scale Systems
- Introducing Spark

Spark Basics

- What is Apache Spark?
- Using the Spark Shell
- Resilient Distributed Datasets (RDDs)
- Functional Programming with Spark

Working with RDDs

- RDD Operations
- Key-Value Pair RDDs
- MapReduce and Pair RDD Operations

The Hadoop Distributed File System

- Why HDFS?
- HDFS Architecture
- Using HDFS

Running Spark on a Cluster

- Overview
- A Spark Standalone Cluster
- The Spark Standalone Web UI

Parallel Programming with Spark

- RDD Partitions and HDFS Data Locality
- Working With Partitions
- Executing Parallel Operations

Caching and Persistence

- RDD Lineage
- Caching Overview
- Distributed Persistence

Writing Spark Applications

- Spark Applications vs. Spark Shell
- Creating the SparkContext
- Configuring Spark Properties
- Building and Running a Spark Application
- Logging

Spark, Hadoop, and the Enterprise Data Center

- Overview
- Spark and the Hadoop Ecosystem
- Spark and MapReduce

Spark Streaming

- Spark Streaming Overview
- Example: Streaming Word Count
- Other Streaming Operations
- Sliding Window Operations
- Developing Spark Streaming Applications

Common Spark Algorithms

- Iterative Algorithms
- Graph Analysis
- Machine Learning

Improving Spark Performance

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

Conclusion

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc., 1001 Page Mill Road, Palo Alto, CA 94304, USA