

(<https://www.cloudera.com/>)

# Apache Spark Application Performance Tuning

This three-day hands-on training course delivers the key concepts and expertise developers need to improve the performance of their Apache Spark applications. During the course, participants will learn how to identify common sources of poor performance in Spark applications, techniques for avoiding or solving them, and best practices for Spark application monitoring.

*Apache Spark Application Performance Tuning* presents the architecture and concepts behind Apache Spark and underlying data platform, then builds on this foundational understanding by teaching students how to tune Spark application code. The course format emphasizes instructor-led demonstrations illustrate both performance issues and the techniques that address them, followed by hands-on exercises that give students an opportunity to practice what they've learned through an interactive notebook environment. The course applies to Spark 2.4, but also introduces the Spark 3.0 Adaptive Query Execution framework.

## What You Will Learn

Students who successfully complete this course will be able to:

- Understand Apache Spark's architecture, job execution, and how techniques such as lazy execution and pipelining can improve runtime performance

- Evaluate the performance characteristics of core data structures such as RDD and DataFrames

- Select the file formats that will provide the best performance for your application

- Identify and resolve performance problems caused by data skew

- Use partitioning, bucketing, and join optimizations to improve SparkSQL performance

- Understand the performance overhead of Python-based RDDs, DataFrames, and user-defined functions

- Take advantage of caching for better application performance

- Understand how the Catalyst and Tungsten optimizers work

- Understand how Workload XM can help troubleshoot and proactively monitor Spark applications performance

Learn about the new features in Spark 3.0 and specifically how the Adaptive Query Execution engine improves performance

## Book the course

[Register](#) (/about/training/course-

directory.html#t=All&sort=%2540start\_at\_dt%20descending&f:@reference\_code=[Apache%20Spark%20Application%20Performance%20Tuning%20Workshop])

## What to Expect

This course is designed for software developers, engineers, and data scientists who have experience developing Spark applications and want to learn how to improve the performance of their code. This is not an introduction to Spark.

Spark examples and hands-on exercises are presented in Python and the ability to program in this language is required. Basic familiarity with the Linux command line is assumed. Basic knowledge of SQL is helpful.

## Course Topics

### **Spark Architecture**

RDDs

DataFrames and Datasets

Lazy Evaluation

Pipelining

### **Data Sources and Formats**

Available Formats Overview

Impact on Performance

The Small Files Problem

### **Inferring Schemas**

The Cost of Inference

Mitigating Tactics

## **Dealing With Skewed Data**

Recognizing Skew

Mitigating Tactics

## **Catalyst and Tungsten Overview**

Catalyst Overview

Tungsten Overview

## **Mitigating Spark Shuffles**

Denormalization

Broadcast Joins

Map-Side Operations

Sort Merge Joins

## **Partitioned and Bucketed Tables**

Partitioned Tables

Bucketed Tables

Impact on Performance

## **Improving Join Performance**

Skewed Joins

Bucketed Joins

Incremental Joins

## **Pyspark Overhead and UDFs**

Pyspark Overhead

Scalar UDFs

Vector UDFs using Apache Arrow

Scala UDFs

## **Caching Data for Reuse**

Caching Options

Impact on Performance

Caching Pitfalls

## Workload XM (WXM) Introduction

WXM Overview

WXM for Spark Developers

## What's New in Spark 3.0?

Adaptive Number of Shuffle Partitions

Skew Joins

Convert Sort Merge Joins to Broadcast Joins

Dynamic Partition Pruning

Dynamic Coalesce Shuffle Partitions

## Appendix A: Partition Processing

## Appendix B: Broadcasting

## Appendix C: Scheduling

Learn more

## CCA Spark and Hadoop Developer Certification

This course is excellent preparation for the **CCA Spark and Hadoop Developer** (</about/training/certification/cdhhdp-certification/cca-spark.html>) exam. Although we recommend further training and hands-on experience before attempting the exam, this course covers many of the subjects tested.

Certification is a great differentiator. It helps establish you as a leader in the field, providing employers and customers with tangible evidence of your skills and expertise.

**Get certified** (</about/training/certification.html>)


## Advance your career

Big data developers are among the world's most in-demand and highly-compensated technical roles. Check out some of the job opportunities currently listed that match the professional profile, many of which seek CCA qualifications.

## Private training

We also provide private training at your site, at your pace, and tailored to your needs.

**Onsite request** (</about/training/private-training.html>)

**f** (<https://www.facebook.com/cloudera/>)   
**(<https://twitter.com/cloudera>)****in**  
**(<https://www.linkedin.com/company/cloudera>)**

**Partners** (</partners.html>)

**Support** (</about/services-and-support/support-offering.html>)

**Community** (<https://community.cloudera.com>)

**Documentation** (<https://docs.cloudera.com>)

**Careers** (</about/careers.html>)

**Contact Us** (</contact-sales.html>)

US: +1 888 789 1488 (tel:18887891488)

Outside the US: +1 650 362 0488 (tel:16503620488)

 English