KOENIG step forward

CLOUDERA Training Partner

KOENIG SOLUTIONS · COURSE OUTLINE

Apache Spark Application Performance Tuning

Learn to tune Apache Spark applications for performance, from architecture and data sources to joins, caching and Spark 3.0 features.

12 MODULES

37 TOPICS

COURSE OVERVIEW

This course examines Spark architecture, including RDDs, DataFrames, Datasets and lazy evaluation, before covering data sources and formats and their impact on performance. It addresses schema inference costs, handling skewed data, and the Catalyst and Tungsten engines that underpin query execution. Further modules cover mitigating shuffles, partitioned and bucketed tables, join performance, Pyspark overhead and UDFs, caching strategies, Workload XM, and new features introduced in Spark 3.0.

WHO THIS COURSE IS FOR

- **Spark Developers**
- **Data Engineers**
- **Pyspark Developers**
- **Performance Engineers**
- **Big Data Analysts**

LEARNING OBJECTIVES

- Explain Spark architecture including RDDs, DataFrames and lazy evaluation
- Compare data sources and formats and their performance impact
- Assess the cost of schema inference and mitigating tactics
- Identify and mitigate skewed data in Spark jobs
- Describe the roles of Catalyst and Tungsten in query execution
- Apply techniques to mitigate shuffles such as broadcast joins
- Configure partitioned and bucketed tables to improve performance
- Analyse join strategies including skewed, bucketed and incremental joins

COURSE SYLLABUS

12 Modules, 37 Topics

01

Spark Architecture

TOPICS

RDDs · DataFrames and Datasets · Lazy Evaluation · Pipelining

02

Data Sources and Formats

TOPICS

Available Formats Overview · Impact on Performance · The Small Files Problem

03

Inferring Schemas

TOPICS

The Cost of Inference · Mitigating Tactics

04

Dealing With Skewed Data

TOPICS

Recognizing Skew · Mitigating Tactics

05

Catalyst and Tungsten Overview

TOPICS

Catalyst Overview · Tungsten Overview

06

Mitigating Spark Shuffles

TOPICS

Denormalization · Broadcast Joins · Map-Side Operations · Sort Merge Joins

07

Partitioned and Bucketed Tables

TOPICS

Partitioned Tables · Bucketed Tables · Impact on Performance

08

Improving Join Performance

TOPICS

Skewed Joins · Bucketed Joins · Incremental Joins

09

Pyspark Overhead and UDFs

TOPICS

Pyspark Overhead · Scalar UDFs · Vector UDFs using Apache Arrow · Scala UDFs

10

Caching Data for Reuse

TOPICS

Caching Options · Impact on Performance · Caching Pitfalls

11

Workload XM (WXM) Introduction

TOPICS

WXM Overview · WXM for Spark Developers

12

What's New in Spark 3.0?

TOPICS

Adaptive Number of Shuffle Partitions · Skew Joins · Convert Sort Merge Joins to Broadcast Joins · Dynamic Partition Pruning · Dynamic Coalesce Shuffle Partitions

WHAT IS INCLUDED

Everything You Need to Succeed

- Koenig course material for every module
- Delivery by certified instructors and subject matter experts
- Hands-on exercises throughout the course
- Course completion certificate from Koenig

ABOUT KOENIG

Who We Are

Established in 1993, Koenig Solutions is a globally recognised training provider with 35+ technology partnerships, 5,000+ training programmes and a network of certified instructors and subject matter experts, supporting individuals and enterprise teams worldwide.

FLEXIBLE DELIVERY

Learn the Way That Suits You

Classroom Training | Live Online Classes (ILO) | Fly-Me-A-Trainer (FMAT) | Flexi (Self-Paced) | 1-on-1 Training

Ready to Get Started?

Talk to a learning advisor about schedules, group training and customised delivery.

Visit koenig-solutions.com | info@koenig-solutions.com