

KOENIG SOLUTIONS • COURSE OUTLINE

Cloudera Data Analyst Training for Apache Hadoop

Learn to store, query and analyse large datasets in Hadoop using Apache Hive and Impala, from basic queries to performance optimisation.

12

MODULES

55

TOPICS

COURSE OVERVIEW

This course covers the fundamentals of Apache Hadoop, including HDFS, YARN, MapReduce, Spark, Pig, Hive, Impala and Sqoop, before introducing Hive and Impala for data processing and analysis. It then addresses querying with Hive and Impala, data management, data storage and performance through partitioning and file formats, and working with multiple datasets using joins and unions. Later modules cover analytic functions and windowing, text analysis, optimisation of Hive and Impala, extending Hive and Impala with custom code and user-defined functions, and comparing Hive, Impala and relational databases to choose the appropriate tool.

WHO THIS COURSE IS FOR

Data Analysts

Business Analysts

Data Scientists

Database Administrators

Hadoop Developers

LEARNING OBJECTIVES

- Explain the motivation and overview of the Hadoop ecosystem
- Describe how Hive and Impala store and query data
- Apply Hive and Impala query language syntax to datasets
- Implement data management tasks such as loading and altering tables
- Apply partitioning and file formats to improve data storage performance
- Analyse multiple datasets using joins and unions
- Apply analytic functions and windowing to query results
- Compare Hive, Impala and relational databases for a given task

12 Modules, 55 Topics

01

Apache Hadoop Fundamentals

TOPICS

- The Motivation for Hadoop
- Hadoop Overview
- Data Storage: HDFS
- **Distributed Data Processing:** YARN, MapReduce, and Spark
- **Data Processing and Analysis:** Pig, Hive, and Impala
- Database Integration: Sqoop
- Other Hadoop Data Tools
- Exercise Scenario Explanation

02

Introduction to Apache Hive and Impala

TOPICS

- What Is Hive?
- What Is Impala?
- Why Use Hive and Impala?
- Schema and Data Storage
- Comparing Hive and Impala to Traditional Databases
- Use Cases

03

Querying with Apache Hive and Impala

TOPICS

Databases and Tables · Basic Hive and Impala Query Language Syntax · Data Types · Using Hue to Execute Queries · Using Beeline (Hive's Shell) · Using the Impala Shell

COMMON OPERATORS AND BUILT-IN FUNCTIONS

Operators · Scalar Functions · Aggregate Functions

04

Data Management

TOPICS

Data Storage · Creating Databases and Tables · Loading Data · Altering Databases and Tables · Simplifying Queries with Views · Storing Query Results

05

Data Storage and Performance

TOPICS

Partitioning Tables · Loading Data into Partitioned Tables · When to Use Partitioning · Choosing a File Format · Using Avro and Parquet File Formats

06

Working with Multiple Datasets

TOPICS

UNION and Joins · Handling NULL Values in Joins · Advanced Joins

07

Analytic Functions and Windowing

TOPICS

Using Common Analytic Functions · Other Analytic Functions · Sliding Windows

COMPLEX DATA

Complex Data with Hive · Complex Data with Impala

08

Analyzing Text

TOPICS

- Using Regular Expressions with Hive and Impala
- Processing Text Data with SerDes in Hive
- Sentiment Analysis and n-grams

09

Apache Hive Optimization

TOPICS

Understanding Query Performance · Bucketing · Hive on Spark

10

Apache Impala Optimization

TOPICS

How Impala Executes Queries · Improving Impala Performance

11

Extending Apache Hive and Impala

TOPICS

- Custom SerDes and File Formats in Hive
- Data Transformation with Custom Scripts in Hive
- User-Defined Functions
- Parameterized Queries

12

Choosing the Best Tool for the Job

TOPICS

- Comparing Hive, Impala, and Relational Databases

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