Hadoop Developer Foundation: Explore Hadoop, HDFS, Hive, Yarn, Spark & More

Learn the Modern Skills & Tools Required to Process Large Data Streams in the Hadoop Ecosystem

www.triveratech.com

Course Snapshot

- **Course: TTDS6509 Hadoop Developer Foundation: Explore Hadoop, HDFS, Hive, Yarn, Spark and More**
- **Duration**: 4 days
- **Skill Level:** Intermediate
- **Audience**: This course is geared for experienced Developers, Managers and Architects (with development experience) who seek to be proficient in modern skills working with Hadoop, Hive and Spark in an enterprise data environment.
- **Hands-on Learning**: This course is approximately 50% hands-on lab to 50% lecture ratio, combining engaging lecture, demos, group activities and discussions with machine-based student labs and exercises. Student machines are required.
- **Delivery Options**: This course is available for **onsite private classroom presentation**, **live online virtual presentation**, or can be presented in a **flexible blended learning format** for combined onsite and remote attendees. Please also ask about our **Self-Paced** / **Video** / **QuickSkills or Mini-Camp Flex Hours / Short Course** options.
- **Customizable**: This course agenda, topics and labs can be further adjusted to target your specific training skills objectives, tools and learning goals. Please inquire for details.

Overview

Apache Hadoop is the classical framework for processing Big Data, and Spark is a new in-memory processing engine.

Hadoop Developer Foundation | Working with Hadoop, HDFS, Hive, Yarn, Spark and More is a lab-intensive hands-on Hadoop course that explores processing large data streams in the Hadoop Ecosystem. Working in a hands-on learning environment, students will learn techniques and tools for ingesting, transforming, and exporting data to and from the Hadoop Ecosystem for processing, as well as processing data using Map/Reduce, and other critical tools including Hive and Pig. Towards the end of the course, we'll introduce other useful tools such as Spark and Oozie and discuss essential security in the ecosystem.

NOTE: This course agenda can be adjusted to add review and discussion of pending desired exam and Certifications as needed. We'll collaborate with your organization to tune the agenda as needed to accommodate additional prep topics and review.

Learning Objectives

This "skills-centric" course is about **50% hands-on lab and 50% lecture,** designed to train attendees in core big data/ Spark development and use skills, coupling the most current, effective techniques with the soundest industry practices. Throughout the course students will be led through a series of progressively advanced topics, where each topic consists of lecture, group discussion, comprehensive hands-on lab exercises, and lab review.

Working in a hands-on learning environment led by our expert Hadoop team, students will explore:

- Introduction to Hadoop
- HDFS
- YARN
- Data Ingestion
- HBase
- Oozie
- Working with Hive
- Hive (Advanced)
- Hive in Cloudera
- Working with Spark
- Spark Basics
- Spark Shell
- RDDs (Condensed coverage)
- Spark Dataframes & Datasets
- Spark SQL
- Spark API programming
- Spark and Hadoop
- Machine Learning (ML / MLlib)
- GraphX
- Spark Streaming

Need different skills or topics? If your team requires different topics or tools, additional skills or custom approach, this course may be further adjusted to accommodate. We offer additional Big Data / Data Science, Hadoop, development, programming, analytics, Python/R, Spark, and other related topics that may be blended with this course for a track that best suits your needs.

Audience & Pre-Requisites

This in an **intermediate-level** course is geared for experienced developers seeking to be proficient in Hadoop, Spark tools & related technologies. Attendees should be experienced Python developers who are comfortable with programming languages. Students should also be able to navigate Linux command line, and who have basic knowledge of Linux editors (such as VI / nano) for editing code.

In order to gain the most from this course, attending students should be:

- Familiar with basic Python programming
- Comfortable in Linux environment (be able to navigate Linux command line, edit files using vi or nano)

Next-Steps / Follow on Training: We offer a wide variety of Big Data, Analytics, Data Science, Hadoop, AI / Machine Learning, Python / R / Scala / Java Programming and other related courses that can advance your skills to the next level after this course. Please see our **Big Data Training Suite** course list for details, or please inquire for next step recommendations for follow on courses or Learning Plan options based on your role and goal.

Enhanced Learning Services: Please also ask about our **Pre-Training Class OnRamp & Prep / Primer** offerings, **Skills Gap Assessment Services**, **Case Studies**, **Knowledge Check Quizzes**, **Skills Immersion Programs & Camps**, **Collaborative Mentoring & Coaching Services** and **Extended Learning Support** & **Post Training** services.

Course Topics / Agenda

Please note that this list of topics is based on our standard course offering, evolved from typical industry uses and trends. We will work with you to tune this course and level of coverage to target the skills you need most. Course topics, agenda and labs are subject to change, and may adjust during live delivery based on audience skill-level, interests and participation.

Day One

Introduction to Hadoop

- Hadoop history, concepts
- Ecosystem
- Distributions
- High-level architecture
- Hadoop myths
- Hadoop challenges
- Hardware and software
- Lab: first look at Hadoop

HDFS

- Design and architecture
- Concepts (horizontal scaling, replication, data locality, rack awareness)
- Daemons: Namenode, Secondary Namenode, Datanode
- Communications and heart-beats
- Data integrity
- Read and write path
- Namenode High Availability (HA), Federation
- Labs: Interacting with HDFS

Day Two

YARN

- YARN Concepts and architecture
- Evolution from MapReduce to YARN
- Labs: Running a sample YARN program

Data Ingestion

- **Flume** for logs and other data ingestion into HDFS
- **Sqoop** for importing from SQL databases to HDFS, as well as exporting back to SQL
- Copying data between clusters (distcp)
- Using S3 as complementary to HDFS
- Data ingestion best practices and architectures
- **Oozie** for scheduling events on Hadoop
- Labs: setting up and using Flume, the same for Sqoop

HBase

- (Covered in brief)
- Concepts and architecture
- HBase vs RDBMS vs Cassandra
- HBase Java API
- Time series data on HBase
- Schema design
- Labs: Interacting with HBase using shell; programming in HBase Java API ; Schema design exercise

Oozie

- Introduction to Oozie
- Features of Oozie
- Oozie Workflow
- Creating a MapReduce Workflow
- Start, End, and Error Nodes
- Parallel Fork and Join Nodes
- Workflow Jobs Lifecycle
- Workflow Notifications
- Workflow Manager
- Creating and Running a Workflow

- Exercise: Create an Oozie Workflow from Terminal
- Exercise: Create an Oozie Workflow Using Java API
- Oozie Coordinator Sub-groups
- Oozie Coordinator Components, Variables, and Parameters
- Exercise: Create an Oozie Workflow from HUE

Day Three

Working with Hive
- Architecture and design
- Data types
- SQL support in Hive
- Creating Hive tables and querying
- Partitions
- Joins
- Text processing
- Labs: various labs on processing data with Hive

Hive (Advanced)
- Transformation, Aggregation
- Working with Dates, Timestamps, and Arrays
- Converting Strings to Date, Time, and Numbers
- Create new Attributes, Mathematical Calculations, Windowing Functions
- Use Character and String Functions
- Binning and Smoothing
- Processing JSON Data
- Execution Engines (Tez, MR, Spark)
- Many labs

Day Four

Hive in Cloudera (or tools of choice)

Working with Spark

Spark Basics
- Big Data, Hadoop, Spark
- What's new in Spark v2
- Spark concepts and architecture
- Spark ecosystem (core, spark sql, mlib, streaming)
- Labs: Installing and running Spark

Spark Shell
- Spark web UIs
- Analyzing dataset – part 1
- Labs: Spark shell exploration

RDDs (Condensed coverage)
- RDDs concepts
- RDD Operations / transformations
- Labs : Unstructured data analytics using RDDs
- Data model concepts
- Partitions
- Distributed processing
- Failure handling
- Caching and persistence
- Lab on the above

Spark Dataframes & Datasets
- Intro to Dataframe / Dataset
- Programming in Dataframe / Dataset API
- Loading structured data using Dataframes
- Labs: Dataframes, Datasets, Caching

Spark SQL
- Spark SQL concepts and overview
- Defining tables and importing datasets
- Querying data using SQL
- Handling various storage formats : JSON / Parquet / ORC
- Labs: querying structured data using SQL; evaluating data formats

Spark API programming (Scala and Python)
- Introduction to Spark API
- Submitting the first program to Spark
- Debugging / logging
- Configuration properties
- Labs : Programming in Spark API, Submitting jobs

Spark and Hadoop
- Hadoop Primer: HDFS / YARN
- Hadoop + Spark architecture
- Running Spark on YARN
- Processing HDFS files using Spark
- Spark & Hive
- Lab

Capstone project
- Team design workshop
- The class will be broken into teams
- The teams will get a name and a task
- They will architect a complete solution to a specific useful problem, present it, and defend the architecture based on the best practices they have learned in class

Optional Additional Topics – Please Inquire for Details

Machine Learning (ML / MLlib)
- Machine Learning primer
- Machine Learning in Spark: MLlib / ML
- Spark ML overview (newer Spark2 version)
- Algorithms: Clustering, Classifications, Recommendations
- Labs: Writing ML applications in Spark

GraphX
- GraphX library overview
- GraphX APIs
- Labs: Processing graph data using Spark

Spark Streaming
- Streaming concepts
- Evaluating Streaming platforms
- Spark streaming library overview
- Streaming operations
- Sliding window operations
- Structured Streaming
- Continuous streaming
- Spark & Kafka streaming
- Labs: Writing spark streaming applications

Student Materials: Each participant will receive a **Student Guide** with course notes, code samples, software tutorials, step-by-step written lab instructions, diagrams and related reference materials and resource links. Students will also receive the project files (or code, if applicable) and solutions required for the hands-on work.

Hands-On Setup Made Simple! Our dedicated tech team will work with you to ensure our 'easy-access' cloud-based course environment is accessible, fully-tested and verified as ready to go well in advance of the course start date, ensuring a smooth start to class and effective learning experience for all participants. Please inquire for details and options.

For More Information

All courses can be presented **onsite** or **online,** or in a **combined / flex / blended learning format**, tailored to target your specific audience, needs and learning goals. We also offer focused, flexible **short courses, self-paced learning** options, **recorded sessions** and more. We train beginner to advanced skills in all areas we cover, and offer **New Hire / Cohort Training**, **Boot Camps**, **Skills Immersion Programs**, **Reskilling Programs**, **Skills Migration & Transition Programs**, and more. We collaborate with you to ensure all courses are truly targeted to meet your specific needs and learning skills, maximizing your valuable training time, as well as your important budget.

Please also visit our extensive **Public Training Schedule** for training for smaller groups or individuals. Please contact us for course details, **Corporate Rates** and **Special Discount Offers**.

For more information about our dedicated training services, collaborative coaching services, courseware licensing and development services, public course schedule, training management services, partner programs, or to see our complete list of course offerings and special offers please visit us at **www.triveratech.com**, email **Info@triveratech.com** or call us toll free at **844-475-4559.** Our pricing and services are always satisfaction guaranteed.

TRIVERA TECHNOLOGIES • Collaborative IT Training, Coaching & Skills Development Solutions
www.triveratech.com • toll free +1-844-475-4559 • Info@triveratech.com • Twitter TriveraTech

ONSITE, ONLINE & BLENDED TRAINING SOLUTIONS • PUBLIC / OPEN ENROLLMENT COURSES • COURSEWARE LICENSING & DEVELOPMENT
MENTORING • ASSESSMENTS • LEARNING PLAN DEVELOPMENT • SKILLS IMMERSION PROGRAMS / RESKILLING / NEW HIRE / BOOT CAMPS
PARTNER & RESELLER PROGRAMS • CORPORATE TRAINING MANAGEMENT • VENDOR MANAGEMENT SERVICES

Trivera Technologies is a Woman-Owned Small-Business Firm

Explore Trivera's Ways to Learn…

TECHNOLOGY MiniCamps · everyCourse EXTRAS · quickSkills FREE COURSES AND WEBINARS · afterCourse COACHING AND SUPPORT