Faysal Aziz

Hadoop and Big Data

HDFS, MapReduce, YARN, ecosystem tools and cluster operations

FA-0788 · 5 days

Data & Analytics · DevOps, Cloud & Infrastructure · Software Development

---

Introduction

Explore the Hadoop ecosystem through guided examples of distributed storage, processing and cluster operations. Learn HDFS, MapReduce and YARN concepts and compare associated data tools. Installation and administration exercises use a prepared compatible stack; the course does not promise competence for every production scenario.

Learning outcomes

- Explain Hadoop architecture, HDFS, MapReduce and YARN roles.
- Perform guided HDFS operations and examine replication and rack awareness.
- Explore MapReduce workflows, formats, debugging and processing patterns.
- Use introductory Pig/HBase/Hive examples and compare ingestion and Spark roles.
- Discuss cluster planning, monitoring, administration, backup and security.
- Distinguish HDFS high availability and federation.

Prerequisites

Basic command-line and data-handling familiarity is helpful. Use the prepared Linux/Java-compatible Hadoop environment and selected component versions; programming examples are guided.

Training outline

1. Day 1 — Hadoop introduction and HDFS

- Big Data Case Studies
- Brief History of Hadoop
- Need for Hadoop
- Hadoop Architecture
- RDBMS vs Hadoop
- Vendor Comparison
- Hardware Recommendations
- Hadoop Installation

Distributed storage

- HDFS Basics
- HDFS Architecture
- Data Read and Write Process
- HDFS Permissions
- Data Replication
- HDFS Accessibility
- HDFS Filesystem Operations
- HDFS Interfaces
- Heartbeats
- Rack Awareness
- distcp

2. Day 2 — MapReduce and YARN

- MapReduce Basics
- MapReduce Workflow
- MapReduce Framework
- Hadoop Data Types
- MapReduce Internals
- Job Formats
- Debugging and Profiling

- Distributed Cache
- Combiner Functions
- MapReduce Streaming
- MapReduce Counters, Sorting and Joins

Resource management

- YARN Infrastructure
- YARN ResourceManager
- YARN ApplicationMaster
- YARN NodeManager
- YARN Container

A YARN container is an allocation/execution context, not automatically a Docker container.

3. Day 3 — Pig and HBase

- Pig Architecture
- Pig Installation and Modes
- Grunt and Pig Script
- Pig Latin Commands
- UDF and Data Processing Operator

- HBase Architecture
- HBase Installation
- HBase Configuration
- HBase Schema Design
- HBase Commands
- MapReduce Integration
- HBase Security

Select compatible releases and a prepared environment for component demonstrations.

4. Day 4 — Query, ingestion and workflow tools

Hive Architecture

- Hive shell
- Hive Data Types
- HiveQL

- Hadoop Workflow
- Sqoop as historical bulk-transfer context: retired, not required for new installation.
- Flume log-ingestion concepts; current Apache rework is not a mandatory new deployment dependency.
- Spark's role in the broader Hadoop processing ecosystem.

5. Day 5 — Cluster management, administration and security

Cluster Planning

- Cluster Installation and Configuration
- Cluster Testing
- Cluster Benchmarking
- Cluster Monitoring

Hadoop Administration

- dfsadmin, fsck and balancer
- Hadoop Logging
- Hadoop Data Backup
- Addition and removal of nodes

Hadoop Security

- Authentication
- Data Confidentiality
- Configuration

Availability and federation

- HDFS HA
- HDFS Federation

Authentication and confidentiality require distinct controls. High availability and replication do not replace an appropriate backup strategy.

© 2026 Faysal Aziz · Kuala Lumpur