Mastering Apache Hadoop Administration

Course Description

This course provides a comprehensive understanding of Hadoop administration, covering cluster setup, configuration, resource management, security, and troubleshooting. Participants will gain hands-on experience in managing Hadoop Distributed File System (HDFS), MapReduce, and related ecosystem components. The training emphasizes practical skills for real-world enterprise deployments.

Duration

5 Days (40 Hours)

Pre-requisites

- Basic knowledge of Linux/Unix operating systems
- Familiarity with Java or scripting languages
- Understanding of distributed systems concepts
- Prior exposure to databases or data warehousing is helpful

Learning Objectives

By the end of this course, participants will be able to:

- Understand Big Data concepts and Hadoop architecture
- Install, configure, and manage Hadoop clusters
- Administer HDFS and MapReduce effectively

- Implement authentication and security using Kerberos
- Manage resources with schedulers and optimize performance
- Perform cluster maintenance, troubleshooting, and recovery

Content Coverage

Module 1: Introduction to Big Data and Hadoop

- What is Big Data?
- Big Data journey and challenges
- Big Data analytics and statistics
- Technologies supported by Big Data
- Introduction to Hadoop and its history
- Breakthroughs and future of Hadoop
- Who is using Hadoop today

Module 2: Hadoop Distributed File System (HDFS)

- Overview of Distributed File Systems
- HDFS architecture and block placement
- NameNode, DataNode, JobTracker, TaskTracker roles
- Secondary NameNode functions
- Typical HDFS workflow
- Data replication and replica placement
- Rack awareness in Hadoop
- Anatomy of file read/write operations

Module 3: MapReduce Framework

- Stages of MapReduce execution
- JobTracker and TaskTracker daemons
- Handling task failures (child, tracker, job, HDFS)
- Fault tolerance mechanisms
- Optimization and tuning strategies

Module 4: Cluster Planning and Design

- Hadoop versions and features
- Hardware selection (Master vs Slave nodes)
- Cluster sizing considerations
- Operating system selection and deployment layout
- Software packages and dependencies
- Disk configuration and mount options
- Network design and typical topologies

Module 5: Installation and Configuration

- Tarball vs Package installation methods
- XML configuration files
- Environment variables setup
- Logging configuration
- HDFS optimization and tuning
- MapReduce optimization and tuning

Module 6: Authentication and Security

- Introduction to Kerberos
- Configuring Hadoop security with Kerberos
- Best practices for secure cluster administration

Module 7: Resource Management

- What is resource management?
- MapReduce scheduling concepts
- Capacity Scheduler configuration
- Fair Scheduler configuration
- Resource optimization techniques

Module 8: Cluster Maintenance

- Managing Hadoop processes (init scripts, manual start/stop)
- Adding and decommissioning DataNodes
- Balancing HDFS block data
- Handling failed disks
- Adding and decommissioning TaskTrackers
- Killing MapReduce jobs and tasks
- Dealing with blacklisted TaskTrackers

Module 9: Troubleshooting

- Common Hadoop failures and problems
- HDFS and MapReduce health checks

- Diagnostic tools and logs
- Recovery strategies

Module 10: Backup and Recovery

- Data backup strategies
- Distributed copy and parallel data ingestion
- NameNode metadata backup and recovery
- Disaster recovery planning