Hadoop Development Training Curriculum

STRUCTURE

www.cromacampus.com | helpdesk@cromacampus.com | +91-120-4155255

Big Data Hadoop Development Training Curriculum

"Hadoop Developer Certification Training to make your career more rewarding and demanding."

Course Objectives:

- Get yourself signed in for our comprehensive, real-world projects led Hadoop Training Program to master the skills and tools in the Hadoop ecosystem.
- Gain all the practical learnings around Hadoop Architecture, Apache PIG, Hive, HBase, YARN, and Programming in MapReduce that will help you in qualifying the competent Hadoop Certifications and make your career a demanding one.
- Prepare yourself for the Hadoop developer global certification exam and get recognition
- Apply for leading MNCs and get hired by top industries across the world.

Course Description:

Croma Campus offers the best Hadoop development Training in Noida with most experienced professionals. Our Instructors are working in Big Data Space and related technologies for years in MNC's.

We aware of industry needs and we are offering Hadoop development Training in more practical way. Our team of Hadoop trainers offers the in-classroom training with best industry practices.

We framed our syllabus to match with the real-world requirements for beginner level to advanced level. Our training will be handled in either weekday or weekends programme depends on participants requirements.

Further, the course content is prepared keeping latest industry trends and certification structure in mind. To know more, you can contact our expert team and get more details.

Course Content:

Module 1: Introduction to Big Data & Hadoop

- Introduction to Big Data
  - Overview of Course
  - What is Big Data?
  - Big Data Analytics
  - Challenges of Traditional System
  - Distributed Systems
- Introduction to Hadoop
  - Components of Hadoop Ecosystem
  - Commercial Hadoop Distributions
  - Why Hadoop?
  - Fundamental Concepts in Hadoop
- Security in Hadoop
  - Why Hadoop Security Is Important?
  - Hadoop's Security System Concepts
  - What Kerberos Is and How it Works?
  - Securing a Hadoop Cluster with Kerberos

- Initial Setup and Configuration
  - Deployment Types
  - Installing Hadoop
  - Specifying the Hadoop Configuration
  - Performing Initial HDFS Configuration
  - Performing Initial YARN and MapReduce Configuration
  - Hadoop Logging

Module 2: HDFS

- What is HDFS?
- Need for HDFS
- Regular File System vs HDFS
- Characteristics of HDFS
- HDFS Architecture and Components
- High Availability Cluster Implementations
- HDFS Component File System Namespace
- Data Block Split
- Data Replication Topology
- HDFS Command Line

Module 3: YARN

- Yarn Introduction
- Yarn Use Case
- Yarn and its Architecture
- Resource Manager
- How Resource Manager Operates?
- Application Master
- How Yarn Runs an Application
- Tools for Yarn Developers

Module 4: Managing and Scheduling Jobs

- Managing Running Jobs
- Scheduling Hadoop Jobs
- Configuring the Fair Scheduler
- Impala Query Scheduling

Module 5: Apache Sqoop

- Apache Sqoop
- Sqoop and Its Uses
- Sqoop Processing

- Sqoop Import Process
- Sqoop Connectors
- Importing and Exporting Data from MySQL to HDFS

Module 6: Apache Flume

- Apache Flume
- Flume Model
- Scalability in Flume
- Components in Flume's Architecture
- Configuring Flume Components
- Ingest Twitter Data

Module 7: Getting Data into HDFS

- Data Ingestion Overview
- Ingesting Data from External Sources with Flume
- Ingesting Data from Relational Databases with Sqoop
- REST Interfaces
- Best Practices for Importing Data

Module 8: Apache Kafka

- Apache Kafka
- Aggregating User Activity Using Kafka
- Kafka Data Model
- Partitions
- Apache Kafka Architecture
- Setup Kafka Cluster
- Producer Side API Example
- Consumer Side API
- Consumer Side API Example
- Kafka Connect

Module 9: Hadoop Clients

- What is a Hadoop Client?
- Installing and Configuring Hadoop Clients
- Installing and Configuring Hue
- Hue Authentication and Authorization

Module 10: Cluster Maintenance

- Checking HDFS Status
- Copying Data between Clusters
- Adding and Removing Cluster Nodes
- Rebalancing the Cluster

- Cluster Upgrading

Module 11: Cloudera Manager

- The Motivation for Cloudera Manager
- Cloudera Manager Features
- Express and Enterprise Versions
- Cloudera Manager Topology
- Installing Cloudera Manager
- Installing Hadoop Using Cloudera Manager
- Performing Basic Administration Tasks using Cloudera Manager

Module 12: Cluster Monitoring and Troubleshooting

- General System Monitoring
- Monitoring Hadoop Clusters
- Common Troubleshooting Hadoop Clusters
- Common Misconfigurations

Module 13: Planning Your Hadoop Cluster

- General Planning Considerations
- Choosing the Right Hardware
- Network Considerations
- Configuring Nodes
- Planning for Cluster Management

Module 14: Advanced Cluster Configuration

- Advanced Configuration Parameters
- Configuring Hadoop Ports
- Explicitly Including and Excluding Hosts
- Configuring HDFS for Rack Awareness
- Configuring HDFS High Availability

Module 15: MapReduce Framework

- What is MapReduce?
- Basic MapReduce Concepts
- Distributed Processing in MapReduce
- Word Count Example
- Map Execution Phases
- Map Execution Distributed Two Node Environment
- MapReduce Jobs
- Hadoop MapReduce Job Work Interaction
- Setting Up the Environment for MapReduce Development
- Set of Classes
- Creating a New Project

- Advanced MapReduce
- Data Types in Hadoop
- Output formats in MapReduce
- Using Distributed Cache
- Joins in MapReduce
- Replicated Join

Module 16: Apache PIG

- Introduction to Pig
- Components of Pig
- Pig Data Model
- Pig Interactive Modes
- Pig Operations
- Various Relations Performed by Developers

Module 17: Apache HIVE

- Introduction to Apache Hive
- Hive SQL over Hadoop MapReduce
- Hive Architecture
- Interfaces to Run Hive Queries
- Running Beeline from Command Line
- Hive Meta Store
- Hive DDL and DML
- Creating New Table
- Data Types
- Validation of Data
- File Format Types
- Data Serialization
- Hive Table and Avro Schema
- Hive Optimization Partitioning Bucketing and Sampling
- Non-Partitioned Table
- Data Insertion
- Dynamic Partitioning in Hive
- Bucketing
- What Do Buckets Do?
- Hive Analytics UDF and UDAF
- Other Functions of Hive

Module 18: No SQL Databases HBase

- NoSQL Databases HBase
- NoSQL Introduction
- HBase Overview
- HBase Architecture
- Data Model

- Connecting to HBase
- HBase Shell

Module 19: Functional Programming using Scala

- Basics of Functional Programming and Scala
- Introduction to Scala
- Scala Installation
- Functional Programming
- Programming with Scala
- Basic Literals and Arithmetic Programming
- Logical Operators
- Type Inference Classes Objects and Functions in Scala
- Type Inference Functions Anonymous Function and Class
- Collections
- Types of Collections
- Operations on List
- Scala REPL
- Features of Scala REPL

Module 20: Apache Spark

- Apache Spark Next-Generation Big Data Framework
- History of Spark
- Limitations of MapReduce in Hadoop
- Introduction to Apache Spark
- Components of Spark
- Application of In-memory Processing
- Hadoop Ecosystem vs Spark
- Advantages of Spark
- Spark Architecture
- Spark Cluster in Real World

Module 21: Hadoop Datawarehouse

- Data warehouse in Hadoop
    - Hadoop and the Data Warehouse
    - Hadoop Differentiators
    - Data Warehouse Differentiators
    - When and Where to Use Which?
- Augmenting Enterprise Data Warehouse
    - Introduction
    - RDBMS Strengths
    - RDBMS Weaknesses
    - Typical RDBMS Scenario
    - OLAP Database Limitations
    - Using Hadoop to Augment Existing Databases

- Benefits of Hadoop
- Hadoop Trade-offs

Advance Programming in Hadoop

Module 22: Writing MapReduce Program

- A Sample MapReduce Program: Introduction
- Map Reduce: List Processing
- MapReduce Data Flow
- The MapReduce Flow: Introduction
- Basic MapReduce API Concepts
- Putting Mapper & Reducer together in MapReduce
- Our MapReduce Program: Word Count
- Getting Data to the Mapper
- Keys and Values are Objects
- What is Writable Comparable?
- Writing MapReduce application in Java
- The Driver
- The Driver: Complete Code
- The Driver: Import Statements
- The Driver: Main Code
- The Driver Class: Main Method
- Sanity Checking the Job's Invocation
- Configuring the Job with Job Conf
- Creating a New Job Conf Object
- Naming the Job
- Specifying Input and Output Directories
- Specifying the Input Format
- Determining Which Files to Read
- Specifying Final Output with Output Format
- Specify the Classes for Mapper and Reducer
- Specify the Intermediate Data Types
- Specify the Final Output Data Types
- Running the Job
- Reprise: Driver Code
- The Mapper
- The Mapper: Complete Code
- The Mapper: import Statements
- The Mapper: Main Code
- The Map Method
- The map Method: Processing the Line
- Reprise: The Map Method
- The Reducer
- The Reducer: Complete Code
- The Reducer: Import Statements
- The Reducer: Main Code

- The reduce Method
- Processing the Values
- Writing the Final Output
- Reprise: The Reduce Method
- Speeding up Hadoop development by using Eclipse
- Integrated Development Environments
- Using Eclipse
- Writing a MapReduce program

Module 23: Introduction to Combiner

- The Combiner
- MapReduce Example: Word Count
- Word Count with Combiner
- Specifying a Combiner
- Demonstration: Writing and Implementing a Combiner

Module 24: Problem-solving with MapReduce

Sorting & Searching large data sets

- Introduction
- Sorting
- Sorting as a Speed Test of Hadoop
- Shuffle and Sort in MapReduce
- Searching

Performing a secondary sort

- Secondary Sort: Motivation
- Implementing the Secondary Sort
- Secondary Sort: Example

Indexing data and inverted Index

- Indexing
- Inverted Index Algorithm
- Inverted Index: Data Flow
- Aside: Word Count

Term Frequency - Inverse Document Frequency (TF- IDF)

- Term Frequency Inverse Document Frequency (TF-IDF)
- TF-IDF: Motivation
- TF-IDF: Data Mining Example
- TF-IDF Formally Defined
- Computing TF-IDF

Calculating Word co- occurrences

- Word Co-Occurrence: Motivation
- Word Co-Occurrence: Algorithm