BIG DATA SPECIALISATION

CloudAge

CLOUDAGE

KONDHWA | PUNE | MAHARASHTRA

WWW.CLOUDAGE.CO.IN

COURSE SYLLABUS

Module 1: What is Big Data?

- Characteristics of Big Data
- What are the V's of Big Data?
- The Impact of Big Data
- Ecosystems of Big Data

Module 2: Introduction to Hadoop

- Understand what Hadoop is
- Components of Hadoop Ecosystem
- Learn about other open source software related to Hadoop
- Understand how Big Data solutions can work on the Cloud

Module 3: The Hortonworks Data Platform

- Hortonworks Data Platform Frameworks
- Apache Ambari
- HDP Installation Using Apache Ambari
- Installing Ambari
- Ambari Web UI
- Managing Hadoop Configuration Properties with Ambari
- Ambari Files View
- Rack Awareness

Module 4: The Cloudera Enterprise Data Hub

- Cloudera Enterprise Data Hub
- CDH Overview
- Cloudera Manager Overview
- Hadoop Administrator Responsibilities
- Installing Cloudera Manager and CDH
- Cluster Installation Overview
- Cloudera Manager Installation
- CDH Installation
- CDH Cluster Services

Module 5: Configuring a Cloudera Cluster

- Overview
- Configuration Settings
- Modifying Service Configurations
- Configuration Files
- Managing Role Instances
- Adding New Services
- Adding and Removing Hosts

Module 6: Cloudera Director: Build Your Cluster

- Cloudera Director: Build Your Cluster
- Add Environment
- Create Cluster Manager Instance
- Add a Cluster
- Cluster Provisioning Status Screens
- Cloudera Director Dashboard

Module 7: Resize your Cluster

- Running an Apache Spark Job
- Add Worker Nodes
- Rerun our Apache Spark Job
- Decommissioning Instances
- Create an Analytics Cluster with Impala
- Terminate a Cluster

Module 8: Hadoop Distributed File System

- Overview
- HDFS Topology and Roles
- Edit Logs and Checkpointing
- HDFS Performance and Fault Tolerance
- HDFS and Hadoop Security Overview
- Web User Interfaces for HDFS
- Using the HDFS Command Line Interface
- Other Command Line Utilities

Module 9: HDFS Data Ingest

- Data Ingest Overview
- File Formats
- Ingesting Data using File Transfer or REST Interfaces
- Importing Data from Relational Databases with Apache Sqoop
- Ingesting Data From External Sources with Apache Flume
- Best Practices for Importing Data

Module 10: Cloudera Director: Build Your Cluster

- Cloudera Director: Build Your Cluster
- Add Environment
- Create Cluster Manager Instance
- Add a Cluster
- Cluster Provisioning Status Screens
- Cloudera Director Dashboard

Module 11: Hive and Impala

- Apache Hive
- Apache Impala

Module 12: YARN and MapReduce

- YARN Overview
- Running Applications on YARN
- Viewing YARN Applications
- YARN Application Logs
- MapReduce Applications
- YARN Memory and CPU Settings

Module 13: Apache Spark

- Apache Spark
- Spark Applications
- How Spark Applications Run on YARN
- Monitoring Spark Applications

Module 14: Planning Your Cluster

- General Planning Considerations
- Choosing the Right Hardware
- Network Considerations
- Virtualization Options
- Cloud Deployment Options
- Configuring Nodes

Module 15: Advanced Cluster Configuration

- Configuring Service Ports
- Tuning HDFS and MapReduce
- Enabling HDFS High Availability

Module 16: Managing Resources

- Configuring groups with Static Service Pools
- The Fair Scheduler
- Configuring Dynamic Resource Pools
- Impala Query Scheduling

Module 17: Impala Query Scheduling

- Checking HDFS Status
- Copying Data Between Clusters
- Rebalancing Data in HDFS
- HDFS Directory Snapshots
- Upgrading a Cluster

Module 18: Monitoring Clusters

- Cloudera Manager Monitoring Features
- Health Tests
- Events and Alerts
- Charts and Reports
- Monitoring Recommendations

Module 19: Cluster Troubleshooting

- Overview
- Troubleshooting Tools
- Misconfiguration Examples
- Essential Points

Module 20: Installing and Managing Hue

- Overview
- Managing and Configuring Hue
- Hue Authentication and Authorization

Module 21: Security Overview

- What Is Security?
- The Need for Security
- Key Principles
- Threat Assessment
- Risk Management

Module 22: Security Architecture

- Scenario Explanation
- Assessing Cluster Security

Module 23: Host Security

- General Server Hardening
- Recommendations
- System Entropy
- Access Control
- Host Firewalls
- Host-Level Monitoring

Module 24: Encrypting Data In Motion

- Encryption Fundamentals
- Certificates
- Configuring Cloudera Manager for TLS

Module 25: Authentication

- Hadoop Authentication Fundamentals
- Kerberos
- Active Directory
- Browser-Based Authentication
- Encrypting Hadoop Data in Motion

Module 26: Authorization

- Authorization Mechanisms
- Cloudera Manager Authorization
- YARN Authorization
- HDFS Authorization
- Apache Sentry

Module 27: Encrypting Data at Rest

- Overview of HDFS Encryption
- Encrypting Data Outside of HDFS
- Hardware Security Modules

Module 28: Additional Considerations

- Auditing
- Data Governance and Lineage Business Continuity and Disaster
- Recovery

Module 29: Apache Kudu

- Kudu Overview
- Architecture
- Installation and Configuration
- Monitoring and Management Tools

Module 30: Apache Kafka

- What Is Apache Kafka?
- Apache Kafka Overview
- Apache Kafka Cluster Architecture
- Apache Kafka Command Line Tools
- Using Kafka with Flume

About Us

A Big Data consulting and solutions provider offering services and training for Big Data Cloud and Machine Learning

We at CloudAge Provide Hadoop Managed Services to Help Traditional Enterprises adopt Apache Hadoop. we Provide Solutions that includes data preparation, data discovery, data availability, and Data Analytics.

Our expertise with Enterprise Distributions deliver a modern platform for analytics data management offerings, in AWS datacenter, Enterprises get one place to store, access, process, secure, and analyse all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Apache Hadoop Open source big data platform is the most widely adopted in the world, and As the leading educator of Hadoop professionals, CloudAge has trained over 1200 individuals worldwide and a seasoned professional services team to help deliver greater time to value.

CloudAge provides best-in-class, technology-managed services and solutions to enterprises that are looking to unlock the potential in their data without the time, cost and complexity associated with traditional big data initiatives. CloudAge delivers an end-to-end solution, so that time is better spent analysing and driving business value from big data. CloudAge provides a full spectrum of services in a private cloud that leverages Hadoop, and helps businesses perform complex analytics and batch-production schedules not possible prior to Hadoop. CloudAge offers a mix of speed, scale, skills, and end-to-end solutions unavailable anywhere else in the big data space.

Insights with Faster Time-to-Value

The analysis of data drives decisions in every business. To gain better business insights, you need to manage the volume, variety, and velocity of data, while applying analytics. With Lenovo-engineered big data validated designs on Lenovo servers, you can harness the power of Apache™ Hadoop® and Apache™ Spark® with Cloudera®, Hortonworks®, IBM® and MapR®. Lenovo servers provide highly reliable and flexible foundations for your business analytics solutions so you can unlock the value of your data and deliver insights faster.

* Outstanding scalability so you can grow as your workloads grow
* Industry-leading transaction processing so you can make better, faster business decisions
* High-throughput capacity that enables you to respond more quickly
* Optimized systems and validated designs for faster time to value