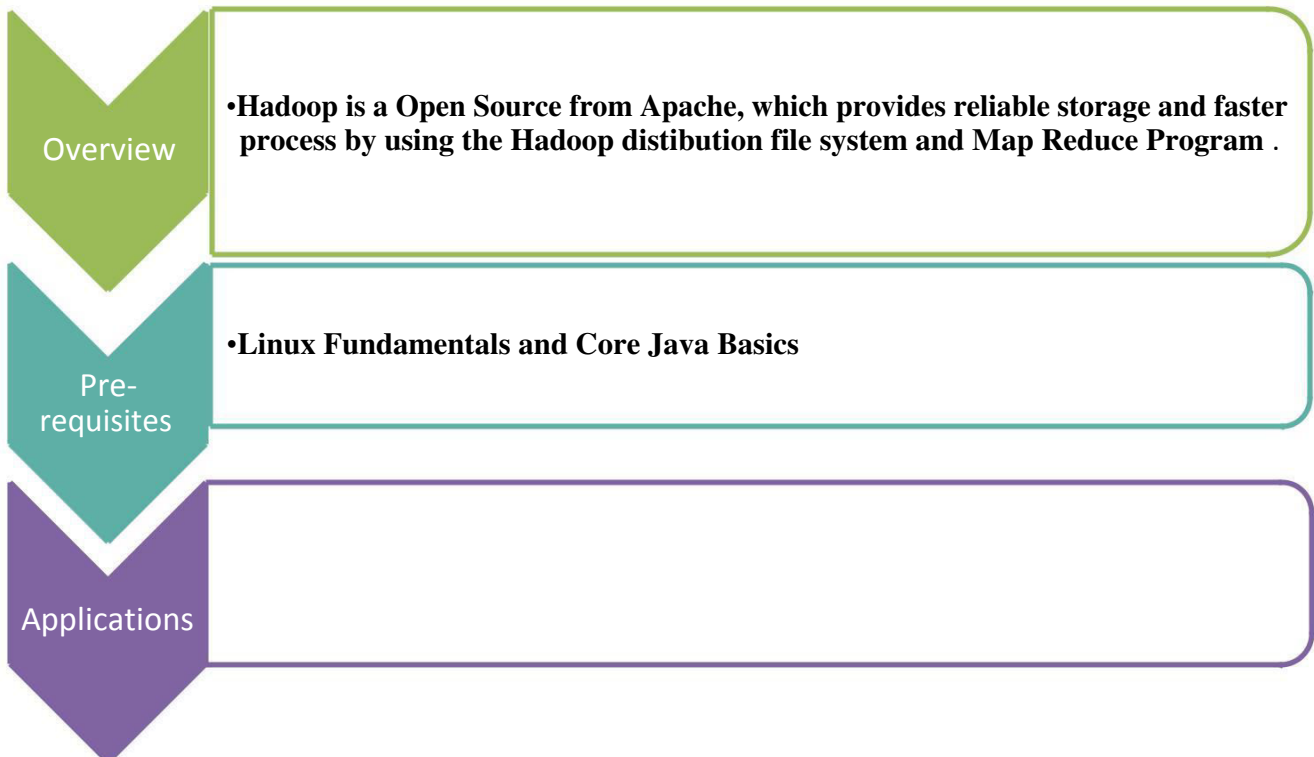


Course Brochure



COURSE CONTENTS

Introduction

- ❖ What is Cloud Computing
- ❖ What is Grid Computing
- ❖ What is Virtualization
- ❖ How above three are inter-related to each other
- ❖ What is Big Data
- ❖ Introduction to Analytics and the need for big data analytics
- ❖ Hadoop Solutions - Big Picture
- ❖ Hadoop distributions
- ❖ Comparing Hadoop Vs. Traditional systems
- ❖ Volunteer Computing
- ❖ Data Retrieval - Random Access Vs. Sequential Access
- ❖ NoSQL Databases

The Motivation for Hadoop

- ❖ Problems with traditional large-scale systems
- ❖ Data Storage literature survey
- ❖ Data Processing literature Survey
- ❖ Network Constraints
- ❖ Requirements for a new Approach

Hadoop: Basic Concepts

- ❖ What is Hadoop?
- ❖ The Hadoop Distributed File System
- ❖ How MapReduce Works
- ❖ Anatomy of a Hadoop Cluster

Hadoop demons

- ❖ Master Daemons
 - ❖ Name node
 - ❖ Job Tracker
 - ❖ Secondary name node
 - ❖ Slave Daemons
 - ❖ Job tracker
 - ❖ Task tracker

HDFS (Hadoop Distributed File System)

- ❖ Blocks and Splits
 - ❖ Input Splits
 - ❖ HDFS Splits
 - ❖ Data Replication
 - ❖ Hadoop Rack Aware
 - ❖ Data high availability
 - ❖ Data Integrity
 - ❖ Cluster architecture and block placement
 - ❖ Accessing HDFS
 - ❖ JAVA Approach
 - ❖ CLI Approach

Programming Practices & Performance Tuning

Developing MapReduce Programs in Local Mode

Running without HDFS and Mapreduce

Pseudo-distributed Mode

Running all daemons in a single node

Fully distributed mode

Running daemons on dedicated Nodes

Hadoop Administrative Tasks

Setup Hadoop cluster of Apache, Cloudera and HortonWorks

- ❖ Install and configure Apache Hadoop
- ❖ Make a fully distributed Hadoop cluster on a single laptop/desktop (Pseudo Mode)
- ❖ Install and configure Cloudera Hadoop distribution in fully distributed mode

COURSE CONTENTS

- ❖ Install and configure **HortonWorks Hadoop** Distribution in fully distributed mode
- ❖ Monitoring the cluster
- ❖ Getting used to management console of Cloudera and Horton Works
- ❖ Name Node in Safe mode
- ❖ Meta Data Backup
- ❖ Integrating Kerberos security in hadoop
- ❖ Ganglia and Nagios Cluster monitoring
- ❖ Benchmarking the Cluster
- ❖ Commissioning/Decommissioning Nodes

Hadoop Developer Tasks

Writing a MapReduce Program

- ❖ Examining a Sample MapReduce Program
- ❖ With Several Examples
- ❖ Basic API Concepts
- ❖ The Driver Code
- ❖ The Mapper
- ❖ The Reducer
- ❖ Hadoop's Streaming API

Performing several Hadoop Jobs

- ❖ The configure and close Methods
- ❖ Sequence Files
- ❖ Record Reader
- ❖ Record Writer
- ❖ Role of Reporter
- ❖ Output Collector
- ❖ Processing video files and audio files
- ❖ Processing image files
- ❖ Processing XML files

- ❖ Processing Zip files
- ❖ Counters
- ❖ Directly Accessing HDFS
- ❖ ToolRunner
- ❖ Using The Distributed Cache

Common MapReduce Algorithms

- ❖ Sorting and Searching
- ❖ Indexing
- ❖ Classification/Machine Learning
- ❖ Term Frequency – Inverse Document Frequency
- ❖ Word Co-Occurrence
- ❖ Hands-On Exercise: Creating an Inverted Index
- ❖ Identify Mapper
- ❖ Identify Reducer
- ❖ Exploring well known problems using MapReduce applications

Debugging MapReduce Programs

- ❖ Testing with MRUnit
- ❖ Logging
- ❖ Other Debugging Strategies

Advanced MapReduce Programming

- ❖ A Recap of the MapReduce Flow
- ❖ Custom Writables and WritableComparables
- ❖ The Secondary Sort
- ❖ Creating InputFormats and OutputFormats
- ❖ Pipelining Jobs With Oozie
- ❖ Map-Side Joins
- ❖ Reduce-Side Joins

COURSE CONTENTS

Monitoring and debugging on a Production Cluster

- ❖ Counters
- ❖ Skipping Bad Records
- ❖ Rerunning failed tasks with Isolation Runner

Tuning for Performance

- ❖ Reducing network traffic with combiner
- ❖ Reducing the amount of input data
- ❖ Using Compression
- ❖ Running with speculative execution
- ❖ Refactoring code and rewriting algorithms
- ❖ Parameters affecting Performance
- ❖ Other Performance Aspects

Hadoop Ecosystem

Hive

- ❖ Hive concepts
- ❖ Hive architecture
- ❖ Install and configure hive on cluster
- ❖ Create database, access it console
- ❖ Buckets,Partitions
- ❖ Joins in Hive
- ❖ Inner joins
- ❖ Outer joins
- ❖ Hive UDF
- ❖ Hive UDAF
- ❖ Hive UDTF
- ❖ Develop and run sample applications in Java to access hive
- ❖ Load Data into Hive and process it using Hive

PIG

- ❖ Pig basics
- ❖ Install and configure PIG on a cluster
- ❖ PIG Vs MapReduce and SQL
- ❖ PIG Vs Hive
- ❖ Write sample Pig Latin scripts
- ❖ Modes of running PIG
- ❖ Running in Grunt shell
- ❖ Programming in Eclipse
- ❖ Running as Java program
- ❖ PIG UDFs
- ❖ PIG Macros
- ❖ Load data into Pig and process it using Pig

Sqoop

- ❖ Install and configure Sqoop on cluster
- ❖ Connecting to RDBMS
- ❖ Installing Mysql
- ❖ Import data from Oracle/Mysql to hive
- ❖ Export data to Oracle/Mysql
- ❖ Internal mechanism of import/export
- ❖ Import millions of records into HDFS from RDBMS using Sqoop

HBase

- ❖ HBase concepts
- ❖ HBase architecture
- ❖ Region server architecture
- ❖ File storage architecture
- ❖ HBase basics
- ❖ Cloumn access
- ❖ Scans
- ❖ HBase Use Cases
- ❖ Install and configure HBase on cluster

COURSE CONTENTS

- ❖ Create database, Develop and run sample applications
- ❖ Access data stored in HBase using clients like Java
- ❖ Map Reduce client to access the HBase data
- ❖ HBase and Hive Integration
- ❖ HBase admin tasks
- ❖ Defining Schema and basic operation

Cassandra

- ❖ Cassandra core concepts
- ❖ Install and configure Cassandra on cluster
- ❖ Create database, tables and access it console
- ❖ Developing applications to access data in Cassandra through Java
- ❖ Install and Configure OpsCenter to access Cassandra data using browser

Oozie

- ❖ Oozie architecture
- ❖ XML file specifications
- ❖ Install and configure Oozie on Cluster
- ❖ Specifying Work flow
- ❖ Action nodes

- ❖ Control nodes
- ❖ Oozie job coordinator
- ❖ Accessing Oozie jobs command line and using web console
- ❖ Create a sample workflows in oozie and run them on cluster
- ❖ **Zookeeper, Flume, Chukwa,Avro, Scribe, Thrift, HCatalog**
- ❖ Flume and Chukwa Concepts
- ❖ Use cases of Thrift ,Avro and scribe
- ❖ Install and Configure flume on cluster
- ❖ Create a sample application to capture logs from Apache using flume
- ❖ **Analytics Basics**
- ❖ Analytics and big data analytics
- ❖ Commonly used analytics algorithms
- ❖ Analytics tools like R and Weka
- ❖ R language basics
- ❖ Mahout
- ❖ **CDH4 Enhance ments**
- ❖ Name Node High Availability
- ❖ Name Node federation
- ❖ Fencing
- ❖ YARN