HADOOP – BIG DATA ANALYSIS FRAMEWORK
Duration: 32 Hrs

Course Outline:

About this tutorial

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.
This brief tutorial provides a quick introduction to Big Data, Map Reduce algorithm, and Hadoop Distributed File System.

Audience

This tutorial has been prepared for professionals aspiring to learn the basics of Big Data Analytics using Hadoop Framework and become a Hadoop Developer. Software Professionals, Analytics Professionals, and ETL developers are the key beneficiaries of this course.

Prerequisites

Before you start proceeding with this tutorial, we assume that you have prior exposure to Core Java, database concepts, and any of the Linux operating system flavours.

BIG DATA OVERVIEW

- What is Big Data?
- What Comes Under Big Data?
- Benefits of Big Data
- Big Data Technologies
- Operational vs Analytical Systems
- Big Data Challenges

BIG DATA SOLUTIONS

- Traditional Enterprise Approach
- Google's Solution
- Hadoop

INTRODUCTION TO HADOOP

- Hadoop Architecture
- Map Reduce
- Hadoop Distributed File System
- How Does Hadoop Work?
- Advantages of Hadoop

ENVIORNMENT SETUP

- Pre-installation Setup
- Installing Java
- Downloading Hadoop
- Hadoop Operation Modes
- Installing Hadoop in Standalone Mode
- Installing Hadoop in Pseudo Distributed Mode
- Verifying Hadoop Installation

HDFS OVERVIEW
- Features of HDFS
- HDFS Architecture
- Goals of HDFS

HDFS OPERATIONS
- Starting HDFS
- Listing Files in HDFS
- Inserting Data into HDFS
- Retrieving Data from HDFS
- Shutting Down the HDFS

COMMAND REFERENCE
- HDFS Command Reference

MAPREDUCE
- What is MapReduce?
- The Algorithm
- Inputs and Outputs (Java Perspective)
- Terminology
- Example Scenario
- Compilation and Execution of Process Units Program
- Important Commands
- How to Interact with MapReduce Jobs

STREAMING
- Example using Python
- How Streaming Works
- Important Commands

MULTI-NODE CLUSTER
- Installing Java
- Creating User Account
- Mapping the nodes
- Configuring Key Based Login
- Installing Hadoop
- Configuring Hadoop
- Installing Hadoop on Slave Servers
- Configuring Hadoop on Master Server
- Starting Hadoop Services
- Adding a New DataNode in the Hadoop Cluster
- Adding a User and SSH Access
- Set Hostname of New Node
- Start the DataNode on New Node
- Removing a DataNode from the Hadoop Cluster