Hadoop Architecture Fundamentals

Course Number 3531 – 8 Hours

Overview

Apache Hadoop is an open-source distributed fault-tolerant system that leverages commodity hardware to achieve large-scale agile data storage and processing. Hadoop empowers applications to work with thousands of nodes and petabytes of data without exposing the complexity of clustering to the end user.

This course discusses the design principles behind Apache Hadoop and explains the architecture of its core sub-systems: HDFS and MapReduce

On Completion, Delegates will be able to

Understand the main Hadoop components and other open source software related to Hadoop. Understand how HDFS works and the concepts of map and reduce operations.

Who Should Attend

This course is intended for developers, architects and technical managers who wish to understand Hadoop's architecture.

Prerequisites

This course assumes no prior knowledge of Hadoop. Participants should be comfortable with Java code and familiar with DWH concepts

Course Contents

Module 1: Big Data Brief Overview

- Big Data Characteristics and Use-Cases
- New Data Categories
- Big Data vs. Traditional Enterprise Relational Data

Module 2: Introduction to Hadoop

- Hadoop vs. Traditional Large-Scale Data Storage and Processing
- Introduction to Hadoop Ecosystem: HDFS, MapReduce, Pig, Hive, HBase
- Hadoop Distributors: Cloudera, Hortonworks, MapR
- Available Java Runtime Environments for Hadoop

Module 3: MapReduce

- MapReduce Motivation and Core Concepts
- How Does MapReduce Work?
- Implementing Common MapReduce Patterns
- Core Hadoop API Interfaces and Classes

Module 4: The Hadoop Distributed File System (HDFS)

- How HDFS works?
- Writing a File to HDFS
- HDFS Command Line
- Hadoop Cluster Architecture overview

Module 5: Hadoop Related Projects

- Data Warehousing with Hive
- Parallel Processing with Pig
- Data Storage with HBase
- Common Utilities: Sqoop, Flume, Zookeeper, etc.