HADOOP TRAINING

HADOOP/ BIG DATA Course Contents

Big data

- Distributed computing
- Data management – Industry Challenges
- Overview of Big Data
- Characteristics of Big Data
- Types of data
- Sources of Big Data
- Big Data examples
- What is streaming data?
- Batch vs Streaming data processing
- Overview of Analytics
- Big data Hadoop opportunities

Hadoop

- Why we need Hadoop
- Data centres and Hadoop Cluster overview
- Overview of Hadoop Daemons
- Hadoop Cluster and Racks
- Learning Linux required for Hadoop
- Hadoop ecosystem tools overview
- Understanding the Hadoop configurations and Installation.

HDFS (Storage)

- HDFS
- HDFS Daemons – Namenode, Data node, Secondary Namenode
- Hadoop FS and Processing Environment's UIs
- Fault Tolerant
- High Availability
- Block Replication
- How to read and write files
- Hadoop FS shell commands

YARN (Hadoop Processing Framework)

- YARN
- YARN Daemons – Resource Manager, Node Manager etc.
- Job assignment & Execution flow

Apache Hive

- Data warehouse basics
- OLTP vs OLAP Concepts
- Hive
- Hive Architecture
- Metastore DB and Metastore Service
- Hive Query Language (HQL)
- Managed and External Tables
- Partitioning & Bucketing
- Query Optimization
- Hiveserver2 (Thrift server)
- JDBC , ODBC connection to Hive
- Hive Transactions
- Hive UDFs
- Working with Avro Schema and AVRO file format

Apache Pig

- Apache Pig
- Advantage of Pig over Map Reduce
- Pig Latin (Scripting language for Pig)
- Schema and Schema-less data in Pig
- Structured , Semi-Structure data processing in Pig
- Pig UDFs
- HCatalog
- Pig vs Hive Use case

Sqoop

- Sqoop commands
- Sqoop practical implementation

HADOOP TRAINING

- Importing data to HDFS
- Importing data to Hive
- Exporting data to RDBMS
- Sqoop connectors

Flume

- Flume commands
- Configuration of Source, Channel and Sink
- Fan-out flume agents
- How to load data in Hadoop that is coming from web server or other storage
- How to load streaming data from Twitter data in HDFS using Hadoop

Oozie

- Oozie
- Action Node and Control Flow node
- Designing workflow jobs
- How to schedule jobs using Oozie
- How to schedule jobs which are time based
- Oozie Conf file

Scala

- Scala
- Syntax formation, Datatypes , Variables
- Classes and Objects
- Basic Types and Operations
- Functional Objects
- Built-in Control Structures
- Functions and Closures
- Composition and Inheritance
- Scala's Hierarchy
- Traits
- Packages and Imports
- Working with Lists, Collections
- Abstract Members
- Implicit Conversions and Parameters
- For Expressions Revisited
- The Scala Collections API
- Extractors
- Modular Programming Using Objects

Spark

- Spark
- Architecture and Spark APIs
- Spark components
- Spark master
- Driver
- Executor
- Worker
- Significance of Spark context
- Concept of Resilient distributed datasets (RDDs)
- Properties of RDD
- Creating RDDs
- Transformations in RDD
- Actions in RDD
- Saving data through RDD
- Key-value pair RDD
- Invoking Spark shell
- Loading a file in shell
- Performing some basic operations on files in Spark shell
- Spark application overview
- Job scheduling process
- DAG scheduler
- RDD graph and lineage
- Life cycle of spark application
- How to choose between the different persistence levels for caching RDDs
- Submit in cluster mode
- Web UI – application monitoring
- Important spark configuration properties
- Spark SQL overview
- Spark SQL demo
- SchemaRDD and data frames

- Joining, Filtering and Sorting Dataset
- Spark SQL example program demo and code walk through

Why Hadoop?

- Solution for Big Data Problem
- Open Source Technology
- Based on open source platforms
- Contains several tool for entire ETL data processing Framework
- It can process Distributed data and no need to store entire data in centralized storage as it is required for SQL based tools.

Prerequisites for Hadoop Training

- Prerequisites for learning Hadoop include hands-on experience in Core Java and good analytical skills to grasp and apply the concepts in Hadoop. We provide a complimentary Course "Java Essentials for Hadoop" to all the participants who enroll for the Hadoop Training. This course helps you brush up your Java Skills needed to write Map Reduce programs.

Project & Assignments

Course Duration

- 8 Weekends