PROTECHSKILLLS

COURSE OVERVIEW – BIG DATA: HADOOP

INTRO TO BIG DATA AND HADOOP

- Characteristics of Big Data
- Limitations of traditional large scale systems
- What is Hadoop?
- Evolution of Hadoop
- Comparison of Hadoop with traditional systems
- Where Hadoop is used?
- Understanding Hadoop Architecture
- Components of Hadoop

UNDERSTANDING HDFS

- Design Goals
- Run HDFS commands
- Understanding Blocks
- FS Image and Edit Logs
- Safe mode in HDFS
- Hadoop DFS Admin commands
- Hadoop NameNode and DataNodes Directory Structure
- Name and Space Quota in HDFS
- HDFS Trash Concept

INSTALLING & CONFIGURING HADOOP

- Types of Installation
- Linux VM installation on Windows for Hadoop cluster using Oracle Virtual Box
- Preparing nodes for Hadoop and VM settings
- Basic Linux commands
- Hadoop Deployment – Single node
- Hadoop configuration files
- Running Hadoop services
- Important Web URIs and Logs for Hadoop
- Hadoop Deployment – Clustered mode
- Hands-on

INTRO TO MAP-REDUCE

- What is MapReduce
- MapReduce Architecture
- Understanding the concept of Mappers & Reducers
- Anatomy of MapReduce Program and its phases
- Splits, Blocks and Record Readers
- Concept of Combiner and Partitioner
- Running and Monitoring MapReduce jobs
- Writing your own MapReduce job
- Hands-on

APACHE HIVE

- What is Hive?
- Hive Architecture & Components
- Hive Installation
- Hive Metastore
- Hive DDLs and DMLs
- Hive SQL – Select, Filter, Order By, Group By
- Hive Joins
- Partitioning in Hive
- Bucketing in Hive
- Built-in Functions
- Hive UDFs
- Hands-on

APACHE PIG

- What is Pig?
- Pig installation
- Pig Data Types
- Pig commands
- Pig Relational Operators
- Pig UDFs
- Hands-on

PROTECHSKILLLS

COURSE OVERVIEW – BIG DATA: HADOOP

APACHE SQOOP

- What is Sqoop?
- Installing Sqoop on your cluster.
- Sqoop Architecture
- Importing Data from RDBMS to HDFS
- Importing Data from RDBMS to Hive
- Importing Data using Free Form Query
- Importing All Tables of a Database
- Export Data using Sqoop
- Hands-on

APACHE FLUME

- What is Flume?
- Flume Architecture
- Configuring Flume on your cluster
- Flume Use Cases
- Hands-on

APACHE YARN

- What is YARN?
- Hadoop 1.x Limitations
- Design Goals for YARN
- YARN Architecture
- Components of YARN
- Schedulers and Queues
- Running and Monitoring YARN Applications
- Hands-on

APACHE ZOOKEEPER

- What is Zookeeper?
- Design Goals
- Architecture
- Installation – Standalone / Clustered mode
- Data Model and Hierarchical Namespace
- ZNodes
- Concept of Watches
- Zookeeper command line
- Hands-on

APACHE SPARK

- What is Spark?
- Spark Architecture
- Installation of Scala
- Configuring an Installing Spark
- Resilient Distributed Dataset (RDD)
- Benefits of Unified Platform
- Spark Core Engine
- Spark SQL
- Spark Streaming
- Spark Machine Learning
- Spark GraphX
- Transformations in Spark
- Actions in Spark
- Modes of Execution
- Hands-on