Introduction to Apache Hadoop

Course ID: ITG-BIG-100

Introduction to Apache Hadoop ITG-BIG-100

Course Overview:

This 5 day course introduces the student to the Hadoop architecture, file system, and the Hadoop Ecosystem. This course will cover the basic concepts of processing unstructured files. This course will also cover MapReduce, Data Types and Formats, HDFS and the Hadoop Ecosystem; Pig, Hive, HBase, and ZooKeeper.

Prerequisites:

Basic Linux Knowledge and some knowledge of traditional database systems.

Who Should Attend this course?

Those who have an interest in Big Data both developers and administrators.

• Introduction to Apache Hadoop	• Apache Hadoop Installation
• Hadoop File System (HDFS)	• Managing HDFS
• MapReduce	• MapReduce Types & Formats
• Hadoop Administration	• Pig & Pig Installation
• Working with Pig	• Hive & Hive Installation
• Working with HBase	• Sqoop Installation

Day 01 Modules: ITG-BIG-100

Module 01: Introduction to Apache Hadoop

- What is Hadoop?
- Current State of Data
- Comparing Hadoop with Traditional RDBMS
- Hadoop History
- Apache Hadoop Ecosystem

Module 02: Preparing your Environment for Installation

- Hadoop Releases
- Hadoop Installation Modes
- Stand Alone Mode
- Pseudo-Distributed Mode
- Fully-Distributed Mode
- Prerequisites to install Apache Hadoop

Module 03: HDFS

- What is HDFS?
- HDFS Architecture
- HDFS Block Placement
- Writing Data to HDFS
- Reading Data from HDFS

Module 04: Managing HDFS

- Data Compressions in HDFS
- Hadoop File System implementations
- Inserting data into HDFS
- HDFS Federation
- High Availability

Day 02 Modules: ITG-BIG-100

Module 05: MapReduce

- MapReduce Architecture
- Mapper Class
- Reducer Class
- Driver Class
- Packing Jar and Running MapReduce Job

Module 06: MapReduce Types & Formats

- MapReduce – Key and Value Pairs
- Serialization
- Hadoop Data Types
- File Input Format
- File Output Format

Module 07: Advanced MapReduce Concepts

- Classic MapReduce Framework
- Failure Cases in Classic MapReduce
- Yarn
- Yarn Components
- MapReduce Job Scheduling
- Counters
- Sorting with Partitioner
- Joins

Module 08: Hadoop Administration

- Troubleshooting Hadoop
- Optimizing Hadoop
- Benchmarking Hadoop
- Administering Hadoop

Day 03 Modules: ITG-BIG-100

Module 09: Pig

- Pig Overview
- Pig Architecture
- Execution Modes
- Interfaces for running Pig Scripts
- Pig Latin vs SQL
- Pig Latin structure
- Pig Latin data types

Module 10: Hands on Pig

- Pig Built in Functions
  - Eval
  - Load and Store
  - Math Functions
  - Tuple, Bag, Map Functions
- User Defined Functions
- Pig Diagnostic Operators

Module 11: Hive

- Hive Architecture
- Hive QL Data Types
- Comparing Hive with RDBMS

Module 12: Hands on Hive

- HiveQL
- Writing UDF's
- Creating Tables
- Querying tables using HiveQL

Day 04 Modules: ITG-BIG-100

Module 13: HBase

- HBase Overview
- Comparing Column-Oriented & Row-Oriented Databases
- HBase vrs RDBMS
- HBase Architecture
- Data Model in HBase

Module 14: Hands on HBase

- HBase Internals
- HBase Shell Commands
- HBase Client

Module 15: Sqoop

- What is Sqoop
- Importing with Sqoop
- Sqoop Connectors
- Downloading Apache Sqoop
- Installing Apache Sqoop
- Configuring Apache Sqoop
- Importing Data from RDBMS to HDFS and Hive
- Exporting data from HDFS to RDBMS

Module 16: Zookeeper

- Zookeeper Overview
- ZooKeeper Data Consistencies
- ZooKeeper Architecture
- ZooKeeper Data Model
- Watches
- Downloading Apache Zookeeper
- Installing Apache Zookeeper
- Configuring Apache Zookeeper
- Using the Command Line Interface in Zookeeper

Day 05 Modules: ITG-BIG-100

Module 17: Case Studies

- Search Engines
- Social Media
- Retail
- Government