Big Data Hadoop Course Content

Course Description:

Big data Hadoop framework has many applications where the data is analyzed in big volumes. Big Data with Hadoop enables multiple types of analytic workloads to execute on similar data, at the same time, at a huge scale on industry-standard hardware. Hadoop is written in Java programming language. As the future scope of Big data Hadoop is high, it is used in social networking platforms like Facebook, Yahoo, Google, Linkedin, Twitter and etc. Hadoop is one of the most in-demand skills, so master the skills through industry experts training to become a Hadoop Developer expert. Our Big data Hadoop online course is designed to give you a competitive edge in the ever-evolving IT job market.

Hachion's Big Data Hadoop tutorial is well prepared by skilled trainers for the beginners, middle-level, and professionals. Basic and advanced topics are also included in the course syllabus to enhance your professional skills. This course provides in-depth knowledge of the Big Data framework using Hadoop and its ecosystem, and explores various applications and tools to process and analyze large volumes of data. Our Big Data Hadoop training will give practical knowledge on HDFS, MapReduce, Hbase, Hive, Pig, Yarn, Oozie, Flume and Sqoop concepts using real-time applications like retail, social media, aviation, tourism, finance domain. Upon completing the course learners will gain expert knowledge in Big Data Hadoop and its ecosystem.

Course Content:

Understanding Big Data and Hadoop

- Big Data
- Limitations and Solutions of existing Data Analytics Architecture
- Hadoop
- Hadoop Features
- Hadoop Ecosystem
- Hadoop 2.x core components
- Hadoop Storage: HDFS
- Hadoop Processing: MapReduce Framework
- Hadoop Different Distributions

Hadoop Architecture and HDFS

- Hadoop 2.x Cluster Architecture - Federation and High Availability
- A Typical Production Hadoop Cluster
- Hadoop Cluster Modes
- Common Hadoop Shell Commands
- Hadoop 2.x Configuration Files
- Single node cluster and Multi node cluster set up Hadoop Administration

Hadoop MapReduce Framework

- MapReduce Use Cases
- Traditional way Vs MapReduce way
- Why MapReduce
- Hadoop 2.x MapReduce Architecture
- Hadoop 2.x MapReduce Components
- YARN MR Application Execution Flow
- YARN Workflow
- Anatomy of MapReduce Program
- Demo on MapReduce
- Input Splits
- Relation between Input Splits and HDFS Blocks
- MapReduce: Combiner & Partitioner
- Demo on de-identifying Health Care Data set
- Demoon Weather Data set

Advanced MapReduce

- Counters
- Distributed Cache
- MRunit
- Reduce Join
- Custom Input Format
- Sequence Input Format
- Xml file Parsing using MapReduce

Pig

- About Pig
- MapReduce Vs Pig
- Pig Use Cases
- Programming Structure in Pig
- Pig Running Modes
- Pig components
- Pig Execution
- Pig Latin Program
- Data Models in Pig
- Pig Data Types
- Shell and Utility Commands
- Pig Latin : Relational Operators
- File Loaders, Group Operator
- COGROUP Operator
- Joins and COGROUP

- Union
- Diagnostic Operators
- Specialized joins in Pig
- Built In Functions (Eval Function, Load and Store Functions, Math function, String Function, Date Function, Pig UDF, Piggybank, Parameter Substitution (PIG macros and Pig Parameter substitution)
- Pig Streaming
- Testing Pig scripts with Punit
- Aviation use case in PIG
- Pig Demo on Healthcare Data set

Hive

- Hive Background
- Hive Use Case
- About Hive
- Hive Vs Pig
- Hive Architecture and Components
- Metastore in Hive
- Limitations of Hive
- Comparison with Traditional Database
- Hive Data Types and Data Models
- Partitions and Buckets
- Hive Tables (Managed Tables and External Tables)
- Importing Data
- Querying Data
- Managing Output
- Hive Script
- Hive UDF
- Retail use case in Hive
- Hive Demo on Healthcare Data set
- Advanced Hive and HBase
- Hive QL: Joining Tables
- Dynamic Partitioning
- Custom Map/Reduce Scripts
- Hive Indexes and views Hive query optimizers
- Hive : Thrift Server, User Defined Functions
- HBase: Introduction to NoSQL Databases and HBase
- HBase v/s RDBMS
- HBase Components
- HBase Architecture
- Run Modes & Configuration
- HBase Cluster Deployment

- Advanced HBase
- HBase Data Model
- HBase Shell
- HBase Client API
- Data Loading Techniques
- ZooKeeper Data Model
- Zookeeper Service
- Zookeeper
- Demos on Bulk Loading
- Getting and Inserting Data
- Filters in HBase

Spark

- Spark Introduction
- Spark Architecture
- Spark RDD's
- Spark Components(Streaming and Spark SQL)
- Programming in Spark + Spark Streaming

Scoop

- Sqoop Installation
- Import Data.(Full table, Only Subset, Target Directory, protecting Password, file format other than CSV, Compressing, Control Parallelism, All tables Import)
- Incremental Import(Import only New data, Last Imported data, storing Password in Metastore, Sharing Metastore between Sqoop Clients)
- Free Form Query Import
- Export data to RDBMS,HIVE and HBASE
- Hands on Exercises

Processing Distributed Data with Apache Spark

- What is Apache Spark
- Spark Ecosystem
- Spark Components
- History of Spark and Spark Versions/Releases
- Spark a Polyglot
- What is Scala?
- Why Scala?
- SparkContext
- RDD

Oozie and Hadoop Project

- Flume and Sqoop Demo

- Oozie
- Oozie Components
- Oozie Workflow
- Scheduling with Oozie
- Demo on Oozie Workflow
- Oozie Co-ordinator
- Oozie Commands
- Oozie Web Console
- Oozie for MapReduce
- PIG
- Hive, and Sqoop
- Combine flow of MR
- PIG
- Hive in Oozie
- Hadoop Project Demo
- Hadoop Integration with Talend