Training on Big Data and Hadoop

Module 1. What is Big Data & Why Hadoop?

- What is Big Data?
- Traditional data management systems and their limitations
- What is Hadoop?
- Why is Hadoop used?
- The Hadoop eco-system
- Big data/Hadoop use cases

Module 2. HDFS (Hadoop Distributed File System) and installing Hadoop on single node

- HDFS Architecture
- HDFS internals and use cases
- HDFS Daemons
- Files and blocks
- Namenode memory concerns
- Secondary namenode
- HDFS access options
- Installing and configuring Hadoop
- Hadoop daemons
- Basic Hadoop commands
- Hands-on exercise

Module 3. Advanced HDFS concepts

- HDFS workshop
- HDFS API
- How to use configuration class
- Using HDFS in MapReduce and programmatically
- HDFS permission and security
- Additional HDFS tasks
- HDFS web-interface
- Hands-on exercise

Module 4. Cloud computing overview and installing Hadoop

- Cloud computing overview
- SaaS/PaaS/IaaS
- Characteristics of cloud computingSaaS/PaaS/IaaS
- Cluster configurationsSaaS/PaaS/IaaS
- Configuring Masters and Slaves
- Module 5.Introduction to MapReduce
- MapReduce basics
- Functional programming concepts
- List processing
- Mapping and reducing lists
- Putting them together in MapReduce
- Word Count example application
- Understanding the driver, mapper and reducer

- Closer look at MapReduce data flow
- Additional MapReduce functionality
- Fault tolerance
- Hands-on exercises

Module 6. MapReduce workshop

- Hands-on work on MapReduce

Module 7. Advanced MapReduce concepts

- Understand combiners & partitioners
- Understand input and output formats
- Distributed cache
- Understanding counters
- Chaining, listing and killing jobs
- Hands-On Exercise
- Module 8. Using Pig and Hive for data analysis
- Pig program structure and execution process
- Joins & filtering using Pig
- Group & co-group
- Schema merging and redefining functions
- Pig functions
- Understanding Hive
- Using Hive command line interface
- Data types and file formats

- Basic DDL operations
- Schema design
- Hands-on examples

Module 9. Introduction to HBase, Zookeeper & Sqoop

- HBase overview, architecture & installation
- HBase admin: test
- HBase data access
- Overview of Zookeeper
- Sqoop overview and installation
- Importing and exporting data in Sqoop
- Hands-on exercise

Module 10. Introduction to Oozie, Flume and advanced Hadoop concepts

- Overview of Oozie and Flume
- Oozie features and challenges
- How does Flume work
- Connecting Flume with HDFS
- YARN
- HDFS Federation
- Authentication and high availability in Hadoop

Module 11. Building a web-log analysis POC using MapReduce

- Designing structures for POC
- Developing MapReduce code

- Push data using Flume into HDFS
- Run MapReduce code
- Analyse the output

System Configuration: Minimum 4GB RAM is needed in the Machines.