COURSE NAME
Hadoop
DURATION
2 Days
COURSE OVERVIEW
Apache Hadoop, the open source data management software that helps organizations analyze massive volumes of structured and unstructured data, is a very hot topic across the tech industry. Employed by such big named websites as eBay, Facebook, and Yahoo, Hadoop can be hosted on cloud environment like **Windows HDinsightService** , where we need to pay for the computing resource we use.
PRE-REQUISITES OF THE PARTICIPANTS
1. Knowledge of Java 2. Basic knowledge of Databases
LAB REQUIREMENTS DETAILS
Hardware/Software requirements: 1. 8 GB RAM windows machine/Mac machine 2. Internet connection for setting up Maven project
COURSE CONTENT

Day 1

Introduction to Hadoop

Hadoop Distributed File System (HDFS)

- HDFS overview and design
- HDFS architecture
- HDFS file storage

Hadoop Distributed File System (HDFS)

- HDFS overview and design
- HDFS architecture
- HDFS file storage
- Component failures and recoveries
- Block placement
- Balancing the Hadoop cluster

Map-Reduce Abstraction

- What MapReduce is and why it is popular

- The Big Picture of the MapReduce
- MapReduce process and terminology
- MapReduce components failures and recoveries
- Working with MapReduce
- Lab: Working with MapReduce

Programming MapReduce Jobs

- Java MapReduce implementation
- Map() and Reduce() methods
- Java MapReduce calling code
- Lab: Programming Word Count

Input/Output Formats and Conversion Between Different Formats

- Default Input and Output formats
- Sequence File structure
- Sequence File Input and Output formats
- Sequence File access via Java API and HDS
- MapFile
- Lab: Input Format
- Lab: Format Conversion

Day 2

MapReduce Features

- Joining Data Sets in MapReduce Jobs
- How to write a Map-Side Join
- How to write a Reduce-Side Join
- MapReduce Counters
- Built-in and user-defined counters
- Retrieving MapReduce counters
- Lab: Map-Side Join

YARN (Hadoop2.0) features: In this class, you will learn what is Yarn and its components**.** We shall how YARN has become the architectural center of Hadoop that allows multiple data processing engines such as **interactive SQL, real-time streaming, data science and batch processing** to handle data stored in a single platform, unlocking an entirely new approach to analytics. We shall look into **Giraph** which is an iterative graph processing system built for high scalability to solve some problem more effectively by processing data as a graph in the Hadoop.

Yarn VsHadoop 1.X: Understand the architecture of YARN and role of different components such as resource manager, Node Manager and App Master. We shall also understand different kind of processing possible
With YARN.

Hive - This class will help you in understanding Hive concepts, Loading and Querying Data in Hive and Hive UDF.

Topics - 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 Outputs, Hive Script, Hive UDF, Hive Demo on Healthcare Data set.

Hands On:

- Understanding the map reduce flow in the Hive-SQL
- Creating Static partition table
- Creating Dynamic partition table
- Loading a unstructured text file into table using Regex serde
- Loading a JSON file into table using Jsonserde
- Creating transaction table
- Creating view and indexes
- Creating ORC, Parquet tables and using compression techniques
- Creating Sequence file table

Writing Java code for UDF
Writing JAVA code to connect with Hive and perform CRUD Operations using JDBC

NoSQL : This class will coverNoSQL in general and HBase in particular. We will see demos on Bulk Loading , Filters. You will also learn what Zookeeper is all about, how it helps in monitoring a cluster, why HBase uses Zookeeper.

Topics - 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.,