Mastering Bigdata with HADOOP

Understanding Bigdata and Hadoop

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

YARN

- YARN (Yet another Resource Negotiator) – Next Gen.
- Map Reduce
- What is YARN?
- Difference between Map Reduce & YARN
- YARN Architecture
- Resource Manager Application Master Node Manager.

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
- Single node cluster and Multi node cluster set up Hadoop Administration

Hadoop MapReduce Frameworks

- MapReduce Use Cases

- Why MapReduce
- Hadoop 2.x MapReduce Architecture
- Hadoop 2.x MapReduce Components
- YARN MR Application Execution Flow
- YARN Workflow
- Input Splits
- Relation between Input Splits and HDFS Blocks
- MapReduce: Combiner & Partitioner
- Sequence Input Format
- Xml file Parsing using MapReduce

Pig

- Introduction to Pig
- MapReduce Vs Pig
- Pig Use Cases
- Programming Structure in Pig
- Pig Running Modes
- Pig components
- Pig Execution
- Data Models in Pig
- Pig Data Types
- Shell and Utility Commands
- Pig Latin: Relational Operators
- Joins and COGROUP
- Union
- Diagnostic Operators
- Built In Functions

Hive

- About Hive
- Hive Vs Pig
- Hive Architecture and Components
- 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
- Retail use case in Hive

Advanced Hive and HBase
- Hive QL: Joining Tables
- Dynamic Partitioning
- 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
- Data Loading Techniques
- ZooKeeper Data Model
- Zookeeper Service
- Zookeeper
- Demos on Bulk Loading
- Getting and Inserting Data

- Filters in HBase

Sqoop
- Sqoop Architecture
- Sqoop Installation
- Sqoop Commands(Import, Hive-Import, EVal, Hbase Import, Import All tables,Export)
- Connectors to Existing DBs and DW
- Hands on Exercise

Flume
- Flume Introduction
- Flume Architecture
- Flume Master
- Flume Collector and Flume Agent
- Flume Configurations
- Real Time Use Case using Apache Flume

Spark
- Introduction to Apache Spark
- Role of Spark in Big data
- Who is using Spark
- Installation of SparkShell and StandAlone Cluster
- Configuration
- RDD Operations (Transformations and actions)