BIG DATA HADOOP TRAINING

Introduction to Big Data

- What is Big Data?
- What are the challenges for processing big data?
- What technologies support big data?
- 3V's of Big Data and Growing.
- Problems with traditional large-scale systems.

Introduction to Hadoop

- An Overview of Hadoop
- History of Hadoop
- Hadoop Core
    - The Hadoop Distributed File System
    - MapReduce Programming model
- Hadoop Ecosystem
- Real Life Use Cases

Hadoop Cluster Setup

- Setup & Configuration details
- Local mode
- Pseudo distributed mode
- Distributed mode
- Using Cloudera CDH.

Hadoop Distributed File System (HDFS)

- HDFS Design & Concepts
- Building Blocks of Hadoop
    - o Name Node (NN) and its functionality
    - o Data Node(DN) and its functionality
    - o Secondary Name Node(SNN) and its functionality

- Replica and Block placement
- HDFS user and admin commands.
- Basic File System Operations
- HDFS Java Client API
- Read and Write flow
- Safe mode
- dist CP - Data loading into HDFS parallel
- Hadoop Data Archives
- Data Integrity and Compression

Map Reduce

- Components of MapReduce
- JobTracker and its functionality
- Task Track and its functionality
- Job execution flow
- MapReduce Programming Model
- Mapper
- Reducer
- Writable and Writable Comparator
- Map Reduce old and new API's.
- Input Formatters and its associated Record Readers
- Input Splits
- Output Formatters and its associated Record Writers
- Configuration and Writing MR jobs in Eclipse.
- Running MR Job on Local Mode.
- Running MR Job on Cluster/Distributed Mode
- Shuffle Sort
- Combiner
- Partitioner
- Job submission flow
- Speculative Execution
- Raw Comparator
- Different File Formats (Sequence File, Map File, Other File Formats)
- Hands-on MapReduce Program Examples

Advance Map Reduce Programming

- Custom Writable
- Custom Partitioner

- Custom Combiner
- Custom Input and output Formatters
- Custom Sorting (Secondary Sorting)
- Distributed Cache
- Counters & Reporter
- Compression techniques
- Joins
- Chaining of MR Jobs
- Adding third party libraries to MR Jobs

Programming Practices

- Writing MapReduce Programs with Eclipse IDE
- Setup Maven Project for writing MapReduce Jobs.
- Web UI for monitoring cluster
- Side Data Distribution Techniques
- Sending Job specific parameters
- Using Distributed Cache
- Performance tuning
- Partitioning MR Job output into multiple output files.

Apache PIG

- Introduction to Apache Pig
- Setup & Configurations
- Pig Latin through Grunt Shell
- Data types
- Relational Operators
- Expressions and Functions
- Working with Pig Script
- Writing reusable script by parameter substitution
- Writing UDF's
- Pig Joins
- Load and Processing Complex Data with Pig
- Hands-on writing Pig Script
- Data Fu/Piggy Bank

Apache Hive

- Introduction to Apache Hive
- Hive vs SQL
- Setup & Configuration
- Hive Architecture
- Meta Store
- Different Data Types
- Hive CLI
- Hive QL
- DDL and DML Operations
- Hive build in operators and functions
- Create Partitioned tables
- Create User Defined Functions
- Bucketing
- Working with different File Formats
- Perform a join of two datasets with Hive
- Tuning

Apache H Base

- H Base introduction
- When Should I Use H Base
- H Base Vs HDFS
- Setup & Configurations
- Key Design
- Column families
- H Base shell commands
- Basic CRUD operations
- Web Based UI
- H Base Architecture
- H Base Components
- Zookeeper
- Compaction
- H Base Hands-on
- Map reduce integration
- Pig Integration
- Hive Integration
- H Base Clients