HADOOP COURSE CONTENT

- 1. Hadoop: Overview
  - Move computation not data.
  - Hadoop performance and data scale facts.
  - Hadoop in the context of other data stores.
  - The Apache Hadoop Project.
  - Hadoop – an inside view: MapReduce and HDFS.
  - The Hadoop Ecosystem.
  - What about NoSQL?
- 2. MapReduce Map and Reduce.
  - Java Map Reduce.
  - Running a Distributed Map.
  - Reduce Job Hadoop Streaming: Python
- 3. The Hadoop Distributed Filesystem
  - HDFS Design & Concepts
  - Blocks, Namenodes and Datanodes
  - Hadoop fs The Command-Line Interface
  - Basic Filesystem Operations
  - Reading Data from a Hadoop URL
  - Reading Data Using the FileSystem API
  - Data Flow Anatomy of a File Read
  - Anatomy of a File Write Coherency Model
- 4. How MapReduce Works
  - Anatomy of a MapReduce Job Run
  - Job Submission Job Initialization, Task Assignment, Task Execution
  - Progress and Status Updates
  - Job Completion, Failures
  - Job Scheduling
  - Fair Scheduler

- Shuffle and Sort – Map Side, Reduce Side
- Configuration Tuning
- Task Execution, Speculative Execution, Task JVM Reuse, Skipping Bad Records
- The Task Execution Environment
- Distributed Cache

- 5. Hadoop Administrator
  - Setting Up a Hadoop Cluster
  - Cluster Specification
  - Network Topology
  - Cluster Setup and Installation
  - SSH Configuration
  - Hadoop Configuration
  - Configuration Management
  - Environment Settings
  - Important Hadoop Daemon Properties
  - Hadoop Daemon Addresses and Ports
  - Post Install
  - Benchmarking a Hadoop Cluster: TeraByte Sort on Apache
  - Hadoop on Amazon EC2
  - Monitoring, Logging Routine Administration Procedures
  - Commissioning and Decommissioning Nodes
  - Upgrades
- 6. Pig
  - Installing and Running Pig
  - Execution Types
  - Running Pig Programs
  - User-Defined Functions
- 7. Hive
  - Basic concepts.
  - HiveQL.
  - Serdes
  - Metastore

- 8. HBase
    - Concepts Data Model, Schema Design
    - Test Drive
    - Clients Java
    - REST and Thrift
    - Metrics