ASM Educational Center (ASM)

25 YEARS Service Excellence

Hadoop for Administrators

Course Outline:

- ## Introduction
  - Hadoop history, concepts
  - Ecosystem
  - Distributions
  - High level architecture
  - Hadoop myths
  - Hadoop challenges (hardware / software)
  - Labs: discuss your Big Data projects and problems

- ## Planning and installation
  - Selecting software, Hadoop distributions
  - Sizing the cluster, planning for growth
  - Selecting hardware and network
  - Rack topology
  - Installation
  - Multi-tenancy
  - Directory structure, logs
  - Benchmarking
  - Labs: cluster install, run performance benchmarks

- ## HDFS operations
  - Concepts (horizontal scaling, replication, data locality, rack awareness)
  - Nodes and daemons (NameNode, Secondary NameNode, HA Standby NameNode, DataNode)
  - Health monitoring
  - Command-line and browser-based administration
  - Adding storage, replacing defective drives
  - Labs: getting familiar with HDFS command lines
- ## Data ingestion
  - Flume for logs and other data ingestion into HDFS
  - Sqoop for importing from SQL databases to HDFS, as well as exporting back to SQL
  - Hadoop data warehousing with Hive
  - Copying data between clusters (distcp)
  - Using S3 as complementary to HDFS
  - Data ingestion best practices and architectures
  - Labs: setting up and using Flume, the same for Sqoop
- ## MapReduce operations and administration
  - Parallel computing before mapreduce: compare HPC vs Hadoop administration
  - MapReduce cluster loads
  - Nodes and Daemons (JobTracker, TaskTracker)
  - MapReduce UI walk through
  - Mapreduce configuration
  - Job config

  - Optimizing MapReduce
  - Fool-proofing MR: what to tell your programmers
  - Labs: running MapReduce examples

- ## YARN: new architecture and new capabilities
  - YARN design goals and implementation architecture
  - New actors: ResourceManager, NodeManager, Application Master
  - Installing YARN
  - Job scheduling under YARN
  - Labs: investigate job scheduling

- ## Advanced topics
  - Hardware monitoring
  - Cluster monitoring
  - Adding and removing servers, upgrading Hadoop
  - Backup, recovery and business continuity planning
  - Oozie job workflows
  - Hadoop high availability (HA)
  - Hadoop Federation
  - Securing your cluster with Kerberos
  - Labs: set up monitoring

- ## Optional tracks
  - Cloudera Manager for cluster administration, monitoring, and routine tasks; installation, use. In this track, all exercises and labs are performed within the Cloudera distribution environment (CDH5)

- Ambari for cluster administration, monitoring, and routine tasks; installation, use. In this track, all exercises and labs are performed within the Ambari cluster manager and Hortonworks Data Platform (HDP 2.0)