Hadoop Administration Fundamentals

Introduction

- Hadoop history and concepts
- Ecosystem
- Distributions
- High level architecture
- Hadoop myths
- Hadoop challenges (hardware/software)

Planning and installation

- Selecting software and Hadoop distributions
- Sizing the cluster and planning for growth
- Selecting hardware and network
- Rack topology
- Installation
- Multi-tenancy
- Directory structure and logs
- Benchmarking

HDFS operations

- Concepts (horizontal scaling, replication, data locality, rack awareness)
- Nodes and daemons (NameNode, Secondary NameNode, HA Standby NameNode, and DataNode)
- Health monitoring
- Command-line and browser-based administration
- Adding storage and replacing defective drives

MapReduce operations

- Parallel computing before MapReduce: compare HPC versus Hadoop administration
- MapReduce cluster loads
- Nodes and Daemons (JobTracker and TaskTracker)
- MapReduce UI walk through
- MapReduce configuration

- Job config
- Job schedulers
- Administrator view of MapReduce best practices
- Optimizing MapReduce
- Fool proofing MR: what to tell your programmers
- YARN: architecture and use

Advanced topics

- Hardware monitoring
- System software monitoring
- Hadoop cluster monitoring
- Adding and removing servers and upgrading Hadoop
- Backup, recovery, and business continuity planning
- Cluster configuration tweaks
- Hardware maintenance schedule
- Oozie scheduling for administrators
- Securing your cluster with Kerberos
- The future of Hadoop