Administrative Course
Table of Contents

Pre-requisite: Basic understanding of Linux and RDBMS

Introduction

What is Big Data?

- Hadoop – What, Why & Which?
    - History.
    - Hadoop components.
    - Distributions available – Apache, Horton, Cloudera, Intel, R, Greenplum, etc.
    - Installations – Pseudo and multinode cluster
- Hadoop cluster hardware & network considerations
    - Sizing standards
    - Network recommendations
    - Storage recommendations
- Rack awareness
- Concepts (horizontal scaling, replication, data locality, rack awareness)
- Nodes topologies (NameNode, Secondary NameNode, Standby NameNode, DataNode)
    - NameNode Considerations
    - Secondary vs Standby Namenode
- Hadoop Ecosystem

HDFS & MapReduce

- HDFS basics and high level architecture.
    - What is HDFS and it's basic architecture.
    - How is it different from other Unix OS?
    - FUSE (File system in User SpacE) and architecture
    - Configuration files, logs and directory structure
    - Reading and writing files
    - Copying data into HDFS
        - Using distcp
        - Using Sqoop from RDBMS
        - Using Flume from external sources
    - Checking HDFS with fsck
    - Quotas and Trash
    - Cluster rebalancing
- MapReduce basics and high level architecture.
    - What is MapReduce?
    - Basic architecture, daemons & features
    - MapReduce Version 1 & 2 (YARN)
        - Job tracker & task tracker
        - Resource manager
        - Node manager & application manager
    - Configuration and it's processing power
    - Scheduling and managing running jobs
        - Failover revovery

        - Web UI to monitor jobs
    - Best practices to optimize MR jobs
    - Scheduler types & configuring a Scheduler (FIFO/Fair Scheduler)

Installation

- Pseudo cluster install using vmware
    - Daemons running
    - Web UI to check health of cluster
- Multi-node cluster install
    - Installing a New Node
    - Namenode formatting
    - NameNode recovery options
    - Configuring Hadoop – xml files
    - Hadoop Ports and Web UI
    - Compression codec
    - Creating Users and using Quotas and Trash
    - Log files

Advanced Configuration, Maintenance & Monitoring

- Explicitly Including and Excluding Hosts/Nodes
- Copying Data Between Clusters
- NameNode Metadata Backup
- General System Monitoring
- Common Troubleshooting Issues
- Backup and Recovery
- Create Queues in Capacity Scheduler
- Snapshots (use cases - data backup, protection against user errors and disaster recovery)
- NFS Gateway access to HDFS
- Migration from Hadoop v1 to Hadoop v2

High Availability & Load Balancing

- High Availability
    - Using QJM, Zookeeper & other options
    - Manual and auto failover
    - Role of standby server
    - Fencing options
    - Split Brain Syndrome
- Federation
    - What is Federation?
    - Using Federation for load balancing
- Federation with HA

Security

- Why & what is available?
- Unix accounts with standard permissions, Kerberos & LDAP.

- Securing a Hadoop Cluster with Kerberos.
- Securing a Hadoop Cluster with LDAP.

Ecosystem Highlights

- Hive
- Impala
- Pig
- HBase
- Sqoop
- Oozie
- Flume

Hands-On Exercises

1. Install a pseudo-distributed Hadoop Cluster
2. Using the Job Tracker UI to start and kill jobs
3. Install multi-node Hadoop Cluster
4. Importing data from MySQL or text file
5. Populating HDFS using Sqoop
6. Run MapReduce jobs
7. Using Fair Scheduler
8. Dead nodes and data replication
9. Adding and removing data nodes