Hadoop Admin

1. The Motivation & Limitation for Hadoop

Motivation of Hadoop
Big data features and challenges
Problems with Traditional Large-Scale Systems
Why Hadoop & Hadoop Fundamental Concepts
Comparison between Hadoop and RDBMS
Is Hadoop replacing RDBMS?
History of Hadoop with Hadoopable problems
Limitation of Hadoop

2. Hadoop Ecosystem & Cluster

Available version Hadoop 1.x & 2
Available Distributions of Hadoop (Cloudera, Hortonworks)
Hadoop Projects & Components
Architecture of Hadoop & Planning for cluster
The Hadoop Distributed File System (HDFS)
Cluster Daemons & Its Functions

- Name Node
- Secondary Node
- Data Nodes
- Application Master and Task Tracker
- Namespace federation

YARN Responsibilities
Deployment of Hadoop Cluster

3. Linux Initials

Installation of Linux (Red Hat)
Basic Linux configurations
Basic Linux commands

- Password less ssh
- IP address and hostname
- Firewall and selinux
- Yum and creating yum repository
- NTP configurations

4. Planning Your Hadoop Cluster

Installation Prerequisites
General Planning Considerations
Choosing the Right Hardware
Network Considerations
Configuring Nodes
Planning for Cluster Management

5. Installation & Deployment of Hadoop

Deployment Types
Setting up Cloudera repository
Installation for Cloudera Manager
Installing Hadoop (Cloudera)
Setting up Cloudera Hadoop environment
Specifying the Hadoop Configuration
Performing Initial HDFS Configuration
Performing Initial YARN and Map Reduce Configuration
Hadoop Logging & Cluster Monitoring

6. 3rd party Vendor Solutions

- Cloudera Manager
- Ambari
- HUE

7. Configuration of services

- Configuring Services
- Configuring HDFS
- Configuring Hadoop Operating System
- (YARN) & Map-Reduce
- Configuring ZooKeeper
- Configuring Hive
- Configuring Pig
- Configuring Schedulers
- Hadoop Logging

8. Advanced Cluster Configuration

- Advanced Configuration Parameters
- Configuring Hadoop Ports
- Explicitly Including and Excluding Hosts
- Rack Awareness and Topology
- Name Node Federation Architecture
- Name Node High-Availability (HA) Architecture

9. Hadoop Security

- Why Hadoop Security Is Important
- Hadoop's Security System Concepts
- What Kerberos is and how it Works
- Securing a Hadoop Cluster with Kerberos

10. Managing and Scheduling Jobs

- Managing Running Jobs
- Scheduling Hadoop Jobs
- Configuring the Fair Scheduler

11. Cluster Maintenance

- Checking HDFS Status
- Copying Data between Clusters
- Adding and Removing Cluster Nodes
- Rebalancing the Cluster
- Cluster Upgrading

12. Sqoop, Flume & HDFS Client

- Sqoop & Flume installation
- Ingesting Data from External (RDBMS) Sources with Sqoop
- Ingesting Data from/to Relational Databases with Sqoop
- Ingesting Data from External Sources with Flume
- Integration of Sqoop and Hbase
- Integration of Flume and Hbase
- Integration of Sqoop and Hive
- Best Practices for Importing Data

13. Conclusion & FAQs

Note:

- Every Topic has practical session
- Hadoop uses different components which discussed in required

Session

- Hue
- Cloudera Manager
- Zookeeper
- Ooozie
- etc.

Prerequisites

This course is best suited to developers and engineers who have some or little bit programming experience. Knowledge of Java is not mandatory, Any programming language can be used with Hadoop and is required to complete the hands-on exercises.

R&D All Hands

1. Building a Search Engine.

One of the awesome characteristics of hadoop courses is how easy it is to offer a search engine wide variety of content. A search engine fundamentally needs distributed search platform which been provided by Hadoop and its related engine. A distributed platform which accepts some sort of search query (elasticsearch, solr) and returns web services that is consumed by web pages. It will open your eyes to new applications that might be built. You will definetly learn alot while building it. "Building a Search Engine" will give you a new perspective on how the Internet works and after you completed this R&D.

Tasks:

- Getting started & basics of elastic search.
- A web crawler (spider) which crawls the data of web pages parsing the HTML and extract the actual content.
- A text based search engine like google to store and index the data from Linux platform.
- Configure the elastic search in distributed environment with hadoop
- Searching data with Kibana web tool for best visual.
- Using Custom Search element Data Rendering, you can write HTML to include rich snippets and other custom data attributes.