HDP Administration Training Contents

Day1

- Introduction and Motivation of Hadoop
    - What is BigData?
    - Challenges in BigData
    - Challenges in traditional application
    - Finding new requirements
    - What is Hadoop?
    - Features of Hadoop
    - Hadoop v/s RDBMS
    - Challenges in Hadoop
    - Overview of HDP ecosystem

- Pseudo mode / Single Node set up of Apache Hadoop (2.7.0)
    - Prerequisites
    - Why password less ssh keys?
    - Important configuration
    - Formatting Namenode
    - Starting and stopping the Hadoop process / daemons
    - Overview of NameNode and ResourceManager Web UI

- Setting Ambari (2.1.0) using internet connectivity
    - Preparing machines
    - Setting up Ambari server
    - Creating one node HDP cluster
        - Adding hosts
        - Choosing services
        - Configuration
    - Overview of NameNode and ResourceManager Web UI
    - Overview of Configuration via Ambari

- Important Configuration
    - fs.defaultFS
    - dfs.block.size
    - dfs.replication
    - dfs.datanode.data.dir
    - dfs.namenode.name.dir
    - Incompatible cluster IDs
    - Understanding safemode concept

Day2

- HDFS Shell commands
    - What is HDFS?
    - Various restriction
    - What is append only in HDFS?
    - Navigating HDFS from command line and UI
    - Understanding and creating home directory for a user
    - mkdir
    - put, copyFromLocal
    - get, copyToLocal
    - ls, ls –R
    - cat
    - fsck
    - chmod, chown

- Local Repository Setup
    - Motivation of local repository set up
    - Setting local HTTP server and hosting Ambari, HDP and HDP utils repository
    - Installing Ambari Server and setting up
    - Installing Ambari Agent and setting up

Day3

- Hadoop Architecture
    - Overview of daemon process
    - HDFS design considerations
        - Horizontal scaling
        - Small files v/s big files
        - What is Batch mode?
        - Writeonce and read many times
        - Fault tolerance w.r.t data and job
    - Roles and responsibilities of
        - NameNode
            - What is Metadata?
            - Importance of fsimage and edits files
            - Scalability limit of Hadoop
        - DataNode
            - Block reports
            - Under replication , over replication scenarios
        - Secondary NameNode
            - Understanding the merging of edits file into fsimage file
            - Effect on Safemode
        - ResourceManager

                - What was the problem of JobTracker?
                - How yarn is capable of running the legacy applications in distributed fashion
            - NodeManager
                - Understanding the concept of container
                - Understanding the concept of application master
        - How does read (get) and write (put) operations happens on HDFS
            - How block corruptions are handled in HDFS

- Commission and Decommissioning
    - How to add new host via Ambari and commission DN and NM
    - How to decommission DN and NM and remove host via Ambari

- HDFS Quotas
    - Types of Quotas
    - Understanding the quota accounting on HDFS

Day 4

- Upgradation Ambari and HDP stack
    - Why do we require upgrades?
    - What is rolling and express upgrade
    - Upgrading Ambari to 2.2.0
    - Upgrading HDP from 2.3 to 2.4

- High Availability
    - NameNode HA
        - Motivation
        - Role of journal nodes
        - Role of Zookeeper and ZKFC
        - Design consideration
        - Who does the role of Secondary NameNode
        - Setting NameNode HA via Ambari
    - Resource Manager HA
        - Motivation
        - Setting Resource Manage HA via Ambari

- Race Awareness
    - Understanding Rack
    - What is rack awareness
    - What is block placement policy?
    - Assigning hosts to racks via Ambari

- Configuring Alerts and Notifications
    - How to create our own group and set up the notification?

Day 5

- Backup and Recovery
    - How to create snapshot of a directory?
    - How to recover a snapshot?
    - How to delete a snapshot?
    - How to rename a snapshot?

- Integrating Ambari with the LDAP
    - Motivation
    - Integrating openLDAP server with Ambari

- Hadoop Scheduler
    - Overview of various Scheduler
    - Understanding capacity scheduler in details
    - Understanding how to achieve multi tenancy via capacity scheduler
    - Configuring queues via Ambari

Day 6

- MapReduce programing model
    - Basics of MR job
    - Understanding basic terminologies
        - Task attempts
        - Speculative execution
        - Fail and killed tasks
        - What is mapper and reducer?
    - Execution of job on yarn framework
        - Uber jobs
        - How failures are handled?

- Data Ingestion Mechanisms
    - Sqoop
        - Importing and exporting the data
        - Incremental report
        - Code gen
        - Options file
    - Flume
        - Basics

        - Source, channel and sink
    - Overview of available source, channels and sinks
    - Configuring a flume agent to push the data to HDFS

Day 7

- Introduction to Hive
    - Motivation
    - Installing and configuring Hive metastore and hive server2
    - Configuring Hive server 2 HA
    - Understanding hive views
    - Running Hive queries on Tez

- Introduction to HBase
    - Motivation
    - CAP theorem
    - HBase architecture
    - HBase daemons
    - How to use HBase shell to perfom put, get scan

- Spark Administration
    - Adding spark services via ambari
    - Overview of architecture
    - Overview of RDD
    - Administering spark on yarn
    - Introduction to spark sql and spark streaming

Day 8

- General planning considerations
- Hadoop Security
    - Understanding the prinicples of security
    - Configuring and installing Kerberos
    - Installing and using Ranger
        - Setting up policies
    - Installing and using Knox
        - Using webHDFS
        -
- Data governance tools (optional)
    - Configuring Falcon

    - Configuring Atlas
- Using Slider (optional)