Admin Assignments [Hadoop]

Pre-Reqs Tasks:

1. Disable IPV6
2. Setup NTP
3. Understand SSH, SCP and RSync
4. Configuring FQDN for server
5. Understand Reverse DNS lookup's and hosts file
6. IPTables
7. SELinux
8. Starting/Stopping services
9. Enable services at specified run time levels

Tasks

1. Setup Pseudo Distributed Hadoop cluster
2. Distribute single node cluster to a multi node deployment using MR1
3. Restrict running an additional SNN on the cluster
4. Replace MR1 on the existing cluster with YARN (MR2)
5. Run a simple mapreduce program in local mode and distributed mode
6. Insert data into HDFS with specified replication factor and change the replication factor for a specified dir/file in HDFS
7. Benchmarking cluster
    1. using TeraGen, TeraSort, TeraValidate
    2. Explore HiBench
8. Fine tune cluster from following perspectives
    1. OS level tuning: Disable Swapping, Open file limits for a process
    2. Utilize cluster resources (CPU, RAM)
    3. JBOD Configuration for Worker Nodes
    4. RAID Configuration for Master Nodes or NFS mount points for Master nodes
9. Setup Rack Topology for an existing cluster
10. Disabling/Enabling trash for HDFS
11. Commission/Decommission of Worker Nodes
12. Performing a minor Upgrading on the cluster
13. Performing a major upgrade on the cluster
14. Setup High Availability for NameNode using NFS
15. Setup High Availability for NameNode using Journal Quorum
16. Setup High Availability for MapReduce Master
17. Setup kerberos and configure hadoop cluster to use kerberos as authentication back end
18. Back & Disaster Recovery
    1. Write a script to back up metadata every hour to a remote NFS mount point as CRON Job
    2. Performing backup's to S3 using DistCP
19. Ops scripting
    1. Script to delete log files older than 7 days as CRON Jobs
    2. Integrate Ganglia and Nagios for monitoring a cluster
20. Alerting Scripts for Nagios
    1. Check DFS status based on number of datanodes alive
    2. Check Namenode status
    3. Check SNN status
    4. Check JT/RM status
    5. Check TT/NM status
    6. Check JN status
    7. Check ZK status
21. Integrate Hue with existing cluster
22. Deploy cluster with Cloudera Manager & explore CM

23. Deploy cluster with Apache Ambari & explore Ambari
24. Deploy a hadoop cluster on Amazon VPC using public and private subnet. Where your gateway should be located in the public subnet and rest of the hadoop nodes should be located in the private subnet only accessible from the gateway and should be able to connect to the internet for downloading packages using NAT.
    1. Side task: When do you want to use public subnet deployment vs private subnet deployment?
25. Explore Amazon EMR
26. Explore Ankus