HDFS AND MAP REDUCE - HADOOP ADMINISTRATION

[1]P.Srilakshmi, [2]N.Siva Rama Krishna Prasad, [3]Dr.M.I.Thariq Hussan

[1]Associate Professor, [2]Assistant Professor, [3]Professor and Head
Department of Information Technology,
Guru Nanak Institutions Technical Campus, Hyderabad, India

Abstract: **HDFS administrator is one who administers and manages Hadoop clusters and all other resources in the entire Hadoop ecosystem. A Hadoop admin's job is not visible to all other groups or end users. The role of a Hadoop admin is mainly associated with tasks that involve installing and monitoring Hadoop cluster, HDFS file structure, locations, and the updated files. Map Reduce administration includes monitoring the list of applications, configuration of nodes, application status, etc. HDFS (Hadoop Distributed File System) contains the user directories, input files, and output files. The Map Reduce commands such as put and get used for storing and retrieving. After starting the Hadoop framework (daemons) by passing the command "startall.sh" on "/$HADOOP_HOME/sbin", pass the following URL to the browser "http://localhost:50070". It describes the file structure of HDFS and data node information in a cluster. A map reduce application is a collection of jobs such as map job, combiner, partitioner and reduce job. It is mandatory to monitor and maintain the following configuration of data node where the application is suitable. Each application consists of the number of data nodes and resources. To monitor all these things, it is imperative that we should have a user interface. After starting the Hadoop framework by passing the command "start-ll.sh" on "/$HADOOP_HOME/sbin", pass the following URL to the browser "http://localhost:8080". It describes the details of the particular application and its local host address of the running node.**

Index Terms **- HDFS Monitoring, Map Reduce Monitoring**

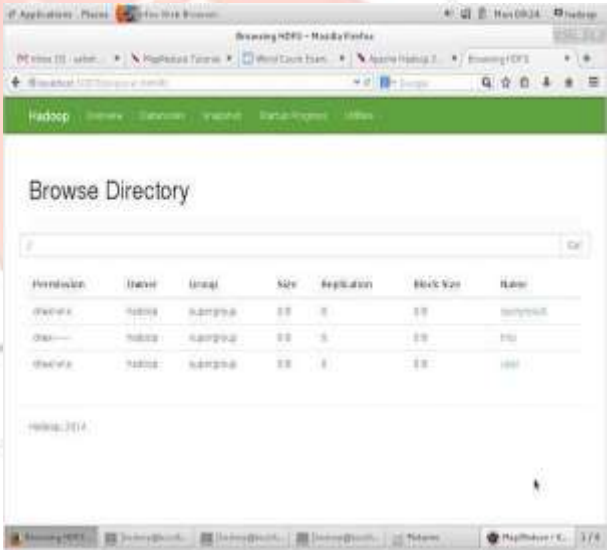
I. INTRODUCTION

Hadoop administration includes both HDFS and map reduce administration. HDFS administration includes monitoring the HDFS file structure, locations, and the updated files. Map reduce administration includes monitoring the list of applications, configuration of nodes, application status, etc.

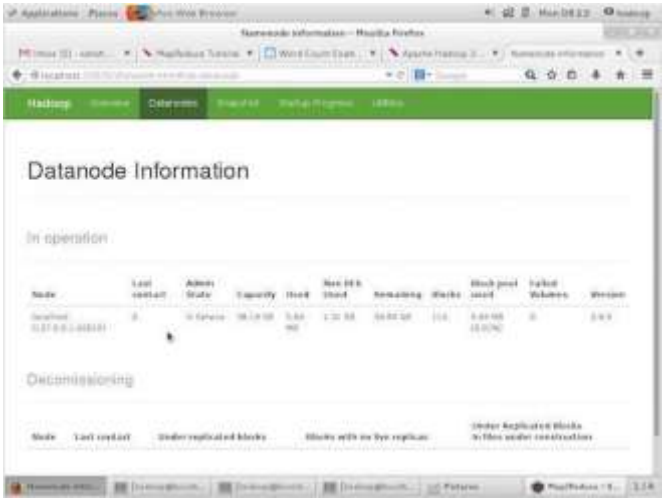
1.1 HDFS Monitoring

HDFS (Hadoop Distributed File System) contains the user directories, input files, and output files. Use the Map Reduce commands, **put** and **get,** for storing and retrieving. After starting the Hadoop framework (daemons) by passing the command "startall.sh" on "/$HADOOP_HOME/sbin", pass the following URL to the browser "http://localhost:50070". You should see the following screen on your browser.

The following screenshot shows how to browse the HDFS. It shows the files in the "/user/hadoop" directory.



The following screenshot shows the data node information in a cluster with its configurations and capacities.



1.2 Map Reduce Job Monitoring

A Map Reduce application is a collection of jobs (Map job, Combiner, Practitioner, and Reduce job). It is mandatory to monitor and maintain the following configuration of data node where the application is suitable. Each application consists of the number of data nodes and resources. To monitor all these things, it is imperative that we should have a user interface. After starting the Hadoop framework by passing the command "start-all.sh" on "/$HADOOP_HOME/sbin", pass the following URL to the browser "http://localhost:8080".

You should see the following screen on your browser. In the above screenshot, the hand pointer is on the application ID.
It describes the following information:

- The application name
- Type of that application
- Current status, Final status

The following screenshot shows the details of a particular application.

The following screenshot describes the currently running nodes information. Here, the screenshot contains only one node. A hand pointer shows the local host address of the running node.

II. HADOOP ADMINISTRATOR

In the Hadoop world, a system administrator is called a Hadoop Administrator. Hadoop admin roles and responsibilities include setting up Hadoop clusters. Other duties involve backup, recovery and maintenance. Hadoop administration requires good knowledge of hardware systems and excellent understanding of Hadoop architecture.

2.1 Roles and Responsibilities

Responsible for implementing and support of the Enterprise Hadoop environment involves designing, capacity arrangement, cluster set up, performance fine-tuning, monitoring, structure planning, scaling and administration.

The administrator consultant works closely with infrastructure, network, database, business intelligence and application teams to ensure that business applications are highly available and performing within agreed on service levels.

1. Need to implement concepts of Hadoop eco system such as YARN, Map Reduce, HDFS, HBase, Zookeeper, Pig and Hive.
2. In charge of installing, administering, and supporting Windows and Linux operating systems in an enterprise environment.
3. Accountable for storage, performance tuning and volume management of Hadoop clusters and Map Reduce routines.
4. In command of setup, configuration and security for Hadoop clusters using Kerberos.
5. Monitor Hadoop cluster connectivity and performance.

- Manage and analyze Hadoop log files.
- File system management and monitoring.
- Develop and document best practices
- HDFS support and maintenance.
- Setting up new Hadoop users.
- Responsible for the new and existing administration of Hadoop infrastructure.
- Include DBA Responsibilities like data modeling, design and implementation, software installation and configuration, database backup and recovery, database connectivity and security.

Hadoop admin roles and responsibilities on popular job portals like Dice.com, Glassdoor.com, and Monster.com are as follows:

- Hadoop administration includes ongoing administration of the Hadoop infrastructure.

- Keeps track of Hadoop Cluster connectivity and security
- Capacity planning and screening of Hadoop cluster job performances.
- HDFS maintenance and support
- Setting up new Hadoop users.

2.2 Hadoop Administration Skills

- Strong scripting skills in Linux environment
- Hands on experience in Oozie, HCatalog, Hive
- Knowledge of HBase for efficient Hadoop administration

III. HOW TO SECURE HADOOP

Current versions of Hadoop are more secure than earlier versions. However, by default they are not secure out-of-the-box, therefore you need to get your hands on the configuration files and make sure that security relevant options are actually enabled.

Also, it is important to keep in mind that the only way of making your cluster secure is to protect it at different layers, from the lower (that is OS-level security) up to application-level and network-level security.

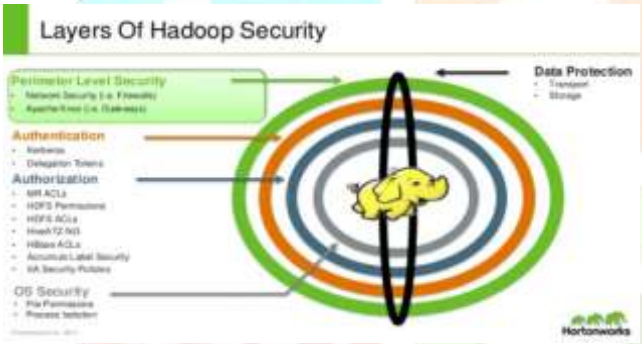


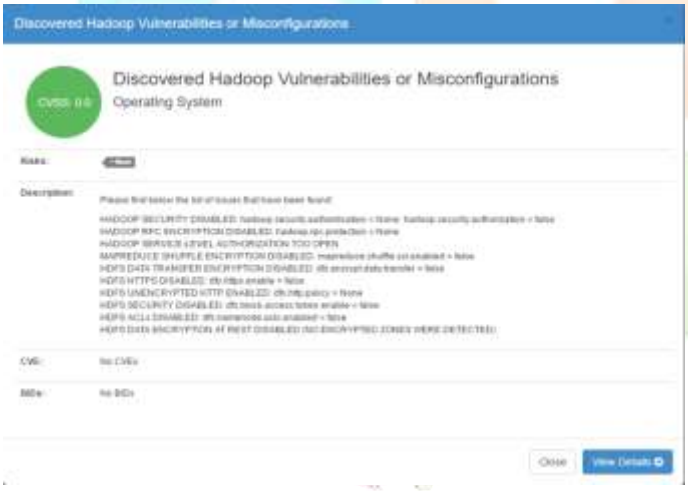
Fig1: Architecture of Layers of Hadoop Security

Here is a minimum set of options we strongly recommend to enable in order to secure Hadoop clusters:

1. Enable HDFS extended ACLs by adding the following properties to *hdfs-site.xml*
   dfs.namenode.acls.enabled true
2. Enable Hadoop security module and strong authentication (Kerberos) by adding the following properties to *core-site.xml*
   hadoop.security.authentication
   kerberos
   Hadoop.security.authorization
   true
3. Secure HDFS by adding the following properties to *hdfs-site.xml* (in particular, enable HTTPS and advanced authorization)
   dfs.block.access.token.enable
   true
   dfs.namenode.keytab.file
   /etc/hadoop/conf/hdfs.keytab
   dfs.namenode.kerberos.principal
   hdfs/_HOST@YOUR-REALM.COM
   dfs.namenode.kerberos.internal.spnego.principal
   HTTP/_HOST@YOUR-REALM.COM
   dfs.secondary.namenode.keytab.file
   /etc/hadoop/conf/hdfs.keytab
   dfs.secondary.namenode.kerberos.principal
   hdfs/_HOST@YOUR-REALM.COM
   dfs.secondary.namenode.kerberos.internal.spnego.principal.
   HTTP/_HOST@YOUR-REALM.COM
   dfs.datanode.data.dir.perm
   700
   dfs.datanode.address
   0.0.0.0:1004
   dfs.datanode.http.address
   0.0.0.0:1006
   dfs.datanode.keytab.file
   /etc/hadoop/conf/hdfs.keytab
   dfs.datanode.kerberos.principal
   dfs/_HOST@YOUR-REALM.COM
   dfs.web.authentication.kerberos.principal
   HTTP/_HOST@YOUR_REALM
   dfs.http.policy HTTPS_ONLY
4. Use Web HDFS (REST API for HDFS), make sure Kerberos authentication is on by adding the following properties to *hdfs-site.xml*
   dfs.web.authentication.kerberos.principal
   HTTP/_HOST@YOUR-REALM.COM
   dfs.web.authentication.kerberos.keytab
   /etc/hadoop/conf/HTTP.keytab
5. Enable transparent data encryption and configure the Key Provider (which will take of generating and providing encryption keys). You can use the *hdfs crypto* to test your configuration.
6. Finally, as a general security warning, make sure that firewall rules are correctly set and limit the access from the Internet only to necessary services. Indeed, by default Hadoop will expose two web interfaces, one for the Resource Manager (on port 8088) and one for Name Node (on port 50070), which are not protected by authentication, therefore they can potentially leak sensitive and critical information

3.1 Continuous Security and Vulnerability Assessment

Monitoring a Hadoop cluster becomes a headache for system administrators and developers, therefore an automated tool can be of great help. That's why we embedded these Hadoop security analysis into our product Elastic Workload Protector in order to help you continuously monitor the security of your Hadoop cluster(s) and be notified as soon as there is a potential issue or mis-configuration.



IV. HADOOP CLUSTER INSTALLATION

Installing a Hadoop cluster typically involves unpacking the software on all the machines in the cluster or installing it via a packaging system as appropriate for your operating system. It is important to divide up the hardware into functions.

Typically one machine in the cluster is designated as the Name Node and another machine as the Resource Manager, exclusively. These are the masters. Other services (such as Web App Proxy Server and Map Reduce Job History server) are usually run either on dedicated hardware or on shared infrastructure, depending upon the load.

The rest of the machines in the cluster act as both Data Node and Node Manager.

4.1 Configuring Hadoop in Non-Secure Mode

Hadoop's Java configuration is driven by two types of important configuration files:

- Read-only default configuration -core-default.xml, hdfs-default.xml, yarn-default.xml and mapred-default.xml.
- Site-specific configuration -etc/hadoop/core-site.xml,etc/hadoop/hdfs-site.xml, etc/hadoop/yarn-site.xml and etc/hadoop/mapred-site.xml.

Additionally, you can control the Hadoop scripts found in the bin/ directory of the distribution, by setting site-specific values via the etc/hadoop/hadoop-env.sh and etc/hadoop/yarn-env.sh.

To configure the Hadoop cluster you will need to configure the environment in which the Hadoop daemons execute as well as the configuration parameters for the Hadoop daemons.

HDFS daemons are Name Node, Secondary Name Node, and Data Node. YARN daemons are Resource Manager, Node Manager, and Web App Proxy. If Map Reduce is to be used, then the Map Reduce Job History Server will also be running. For large installations, these are generally running on separate hosts.

4.2 Day-to-Day Hadoop Cluster Care

The lifetime of a Hadoop Administrator revolves around making, managing and watching the Hadoop Cluster. However, cluster administration isn't an identical activity practiced through and thru by directors from round the globe. The most variable during this case is that the "Distribution of Hadoop" or in easy words a 'cluster' based mostly wherever you decide on the cluster watching tools. The various distributions of Hadoop are Cloudera, Horton works, Apache and MapR. Apache distribution is in fact the Open supply Hadoop distribution.

As associate administrator, if I need to setup Hadoop clusters on the Horton works/Cloudera distribution, my job are easy as a result of all the configurations files are gift on startup. However, within the case of the open supply Apache distribution of Hadoop, we've got to manually setup all the configurations like Core-Site, HDFS-Site, YARN-Site and Map Red-Site.

Once we've got created the cluster, we've got to make sure that the Cluster is active and accessible the least bit times. For this, all the nodes within the cluster need to be setup. They're Name Nodes, Data Node, Active & Standby Name Nodes, Resource Manager and also the Node Manager.

Name Nodes is that the Heart of the cluster. It consists of information that helps the cluster to acknowledge the info and coordinate all the activities. Since lots depends on the Name Nodes, we've got to make sure 100 percent dependableness and for this, we've got one thing known as the Standby Name Nodes that acts because the backup for the Active Name Nodes. Name Nodes stores the information, whereas the particular information is hold on within the information Node within the type of Blocks. The Resource Manager takes care of the cluster's CPU and memory resources the least bit times for all the roles whereas the applying Master manages the particular jobs.

If all the above services are running and are active at all times, then your Hadoop Cluster is ready for use.

When setting up the Hadoop Cluster, the administrator will also need to decide the cluster size based on the amount of data that is to be stored in the HDFS. Since the replication factor of HDFS is 3, 15 TB of free space is required to store 5 TB of data in the Hadoop cluster. The replication factor is set at 3 in order to increase the Redundancy and Reliability. Cluster growth based on storage capacity is a very effective technique that is implemented in the clusters. We can add new systems to the existing cluster and thereby increase the storage space to any number of times.

Another important activity we have to perform as a Hadoop Administrator is that we have to monitor the cluster on a regular basis. While monitor the Cluster it is to ensure that it is up and running at all times and to keep track of the performance. Clusters can be monitored using the various cluster monitoring tools. We choose the appropriate cluster monitoring tools based on the distribution of Hadoop that you are using.

4.3 Monitoring tools for distribution of Hadoop

Open Source Hadoop/Apache Hadoop à Nagios/Ganglia/Ambari
Shell scripting/Python Scripting.

Cloud era Hadoop à Cloud era Manager + Open Source Hadoop tools

Horton works à Apache Ambari + Open Source Hadoop tools.

Ganglia is used for monitoring Compute Grids i.e a bunch of servers working on the same task to achieve a common goal. It is like a cluster of clusters.

Ganglia is also used to monitor the various statistics of the cluster. Nagios is used for monitoring the different servers, the different services running in the servers, switches, and network bandwidth via SNMP etc.

Nagios and Ganglia are open source which is why both are slightly difficult to manage when compared to Ambari and Cloud era Manager. The former is the monitoring tool used by Horton works distribution while Cloud era uses the latter.
Apache Ambari and Cloud era Manager are more popular tools because they come along with the Hadoop Distributions providing you with around 10,000 statistics to monitor. But the drawback is that they are not open source.

4.4 Key skills For Hadoop admin

- Mater of UNIX commands
- Sound knowledge UNIX based File System
- Excellent hold on shell scripting
- Deep understanding and knowledge of Operating System, Scheduling, Process Management.
- Command on Hadoop cluster setup Single Node, Pseudo Distributed and fully distributed Mode.
- Knowledge of Networking.
- What does a Hadoop Admin do?
- Installation and Configuration
- Cluster Maintenance
- Resource Management
- Security Management
- Troubleshooting
- Cluster Monitoring
- Backup and Recovery Task

If working with open source Apache Distribution then hadoop admin's has to manually setup all the configurations- Core-Site, HDFS-Site, YARN-Site and Map Red-Site. However, when working with popular hadoop distribution like Horton works, Cloud era or MapR the configuration files are setup on startup and the hadoop admin need not configure them manually.

4.5 Backup and recovery tasks

Many people don't consider backups since Hadoop has 3X replication by default. Also, Hadoop is often a repository for data that resides in existing data warehouses or transactional systems, so the data can be reloaded. That is not the only case anymore! Social media data, ML models, logs, third-party feeds, open APIs, IoT data, and other sources of data may not be reloadable, easily available, or in the enterprise at all. So, this is not critical single-source data that must be backed up and stored forever.

There are a lot of tools in the open-source space that allow you to handle most of your backup, recovery, replication, and disaster recovery needs. There are also some other enterprise hardware and software options.

V. CONCLUSION

Hadoop distributed file systems provides a high throughput access to data of an application and is suitable for applications that need to work with large data sets. It is designed to hold terabytes or petabytes of data and provides higher throughput access to this data.

Files containing data are stored redundantly across number of machines for higher availability and durability to failure. Moving computation is faster than the moving data. By default, the Map Reduce framework gets input data from the Hadoop Distributed File System (HDFS). The reduce phase uses results from map tasks as input to a set of parallel reduce tasks. The reduce tasks consolidate the data into final results.

By default, the Map Reduce framework stores results in HDFS.

REFERENCES

[1] https://www.dezyre.com/article/what-are-the-job-responsibilities-of-a-hadoop administrator/306
[2] https://elasticsecurity.com/2016/08/10/big-data-security-how-to-secure-your-hadoop-cluster/
[3] https://hadoop.apache.org/docs/r2.7.2/hadoopprojectdist/hadoophdfs/TransparentEncryption.html# Configuration
[4] https://www.slideshare.net/amalgjose89/deployment-and-management-of-hadoop-clusters
[5] https://www.dezyre.com/article/hadoop-cluster-overview-what-it-is-and-howtosetup-one/356
[6] https://www.slideshare.net/amalgjose89/deployment-and-management-of-hadoopclusters
[7] https://www.slideshare.net/amalgjose89/deployment-and-management-of-hadoopclusters

[8] http://www.xoomtrainings.com/blog/what-day-to-day-activities-does-a-hadoop-admin-do
[9] https://cfs22.simplicdn.net/ice9/free_resources_article_thumb/Applications_of_big_data_infog raphic.png
[10] https://mapr.com/hadoop-security-and-big-data-governance-mapr/