****

**Hadoop Cluster Health Check**

University of Maryland | Production Cluster 1 | November 2017

# 

# 

# 

|  |  |  |  |
| --- | --- | --- | --- |
| **Revision History** | | | |
| **Version** | **Author** | **Description** | **Date** |
| V1 | Michael Arnold | Initial Version | 10 Nov 2017 |

# 

# Table of Contents

[**Table of Contents**](#_30j0zll) **2**

[**Section 1.0 Overview**](#_1fob9te) **3**

[**Section 2.0 Evaluation Script Analysis**](#_wygny74pbucj) **3**

[**Section 3.0 Cloudera Manager Analysis**](#_9g6wd4sabojb) **4**

[**Section 4.0 YARN Tuning**](#_l5yaoobqhukw) **5**

# Section 1.0 Overview

The Clairvoyant team performs an initial review of the Hadoop cluster to verify various configurations are following best practices and to scan for existence of known issues. The aim for the initial review process is as follows:

1. Verifying the configuration of operating system, network, and hardware.
2. Review the Hadoop deployment strategy.
3. Review cluster configuration and provide any recommendations for any changes.
4. Review and provide recommendations for security on the cluster. (if applicable)

This cluster review document describes our initial system architecture findings and related recommendations.

# Section 2.0 Evaluation Script Analysis

This section provides an analysis from evaluation script output. These items are **not** ranked in any particular order.

1. RAM sizes are different between nodes of the same type. Four Dell PowerEdge R630s have 251 GiB and two have 235 GiB as seen from the OS. There may be firmware configuration or OS kernel boot parameters that are inconsistent.
2. Firmware levels are different between nodes. Four Dell PowerEdge R630s have version 2.0.2, one has version 2.4.3, and one has version 2.5.5. This was discovered when looking into item 2.0.1 above.
3. Multiple Java Development Kits (JDKs) are installed. Recommendation is to uninstall all except a single Oracle JDK 8.
4. The $JAVA\_HOME environment variable is not set at the OS layer across all machines in the cluster. Recommended to set $JAVA\_HOME to the desired JDK.
5. Multiple IP addresses are assigned to each node in the cluster. Cloudera does not [support multihoming](https://www.cloudera.com/documentation/enterprise/release-notes/topics/rn_consolidated_pcm.html#cdh_cm_network_security). This may result in intermittent inaccessibility of services to certain clients when DNS names resolve to interfaces on networks that are unreachable from the client.
6. DNS entries for host A and PTR records are not configured. In general we recommend the use of DNS over /etc/hosts as this allows all clients in the environment to access the cluster. It also becomes more important when enabling security features like Kerberos and TLS in Hadoop.
7. The Name Service Caching Daemon (NSCD) service is not in use. Recommend to install NSCD and cache DNS lookups for improved cluster performance and lower DNS server utilization.
8. Transparent Huge Pages (THP) is not completely disabled. Recommendation is to disable THP across all nodes.
9. The Kernel parameter vm.swappiness is inconsistently set across machines in the cluster. Recommended value is 1.
10. Swap space is configured on all nodes and is fairly large in some cases. Recommendation is to disable or minimize swap size as it is undesirable for Hadoop processes to swap.
11. Available Entropy is below 500 and it is recommended to be above 1000. Installation of the rng-tools package and starting the rngd process can achieve this.
12. Network Maximum Transmission Unit (MTU) is set to 1500. Increased network performance can be achieved if configured to use Jumbo frames (9000). This is highly dependent upon network switch configuration and should not be blindly set without coordination with Network Administration staff.

# Section 3.0 Cloudera Manager Analysis

This section provides an analysis of Cloudera Manager configuration. These items are ranked in order of importance.

1. HDFS “Journal Node Edits Directory” is set to “/data1/dfs/jn”. HDFS “NN Data Directory” is set to “/data1/dfs/nn”. These services are latency sensitive and should be using separate, dedicated filesystems/disks in order to not have their disk IO impacted by each other.
2. The Cloudera Manager server is using JDK 7. JDK 8 is recommended and is already installed. This corresponds to issue 2.0.10.
3. HDFS “Block Count Threshold” is set to “2,000,000” and has been reached by some nodes. This indicates a [small](https://blog.cloudera.com/blog/2009/02/the-small-files-problem/) [files](https://community.hortonworks.com/articles/15104/small-files-in-hadoop.html) [problem](https://snowplowanalytics.com/blog/2013/05/30/dealing-with-hadoops-small-files-problem/). Recommend to determine if there are too many small files in HDFS and combine them into fewer, larger files.
4. Recommend to implement Static Service Pools. This will allow YARN and HBase to coexist and not impact each other’s performance.
5. Recommend to configure dynamic resource pools for YARN with task preemption. This will allow jobs to be submitted to separate YARN queues and for large, long-running jobs to not fully consume the cluster resources, blocking smaller jobs.
6. The Host configuration item “Java Home Directory” is set to “/usr/java/jdk\_1.8.0\_92”. Recommendation is to leave this property unset. This corresponds to issue 2.0.10.
7. YARN tuning is needed for the cluster as the given parameters are inconsistent.
8. HBase tuning may be needed, depending upon use as the given parameters are unoptimized.
9. ZooKeeper client connection limit is still the default of 60. Should be 300 or more.
10. Recommend to upgrade to latest CM and CDH versions to avoid security issues with CM 5.7.1 and CDH 5.7.1.
11. Recommend to implement YARN HA to complement HDFS HA.
12. Recommend to implement HBase HA to complement HDFS HA.
13. Remove the YARN safety valve properties “yarn.resourcemanager.hostname” and “yarn.nodemanager.aux-services”. These properties should not be required and will conflict with YARN HA.
14. Impala is installed but the service is stopped. Recommend to remove the service if it is not in use.
15. Hue is using the SQLite3 database. Recommend to migrate from SQLite3 to MySQL for better concurrent (multi user) operation.
16. The Cloudera Management Service (CMS) Activity Monitor service and it’s database can be removed. This service is only required if running MapReduce version 1 (ie JobTracker and TaskTracker processes).
17. All empty “\* Group1” Role Group configurations in HDFS can be removed.
18. YARN “MapReduce submit replication” is set to “3”. It should be returned to the default of 10. This will allow better parallel operation of YARN jobs.
19. “Suppress Health Test: Agent Status” is “enabled”. It should be disabled so that the agent health test alerts are sent.

# Section 4.0 YARN Tuning

This is the Clairvoyant recommended initial tuning for YARN. This does not take into account actual jobs running on the cluster.

Worksheet:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Host Components | | | Quantity | Description | | |
| RAM | | | 128 | Gigabytes | | |
| CPU | | | 48 | 2 CPUs \* 8 cores + HT | | |
| HDD (Hard Disk Drive) | | | 24 | 12x3TB SATA III Hard Drives in JBOD Configuration | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Service | | | Category | CPU (cores) | Memory (MB) | Notes |
| Operating System | | | Overhead | 1 | 8192 | Most operating systems use 4-8GB minimum. |
| Task overhead | | | Overhead | 0 | 8192 | Allow additional memory overhead for task buffers such as the HDFS Sort I/O buffer, JVM overheads, etc. |
| Cloudera Manager agent | | | Overhead | 1 | 1024 | Allocate 1GB for Cloudera Manager agents, which track resource usage on a host. |
| Other services | | | Overhead | 1 | 1024 | Enter the required cores or memory for services not listed above. |
| HDFS DataNode | | | CDH | 1 | 4096 | Allocate 1GB for the HDFS DataNode. |
| Impala daemon | | | CDH | 0 | 0 | (Optional Service) Suggestion: Allocate at least 16GB memory when using Impala. |
| Hbase RegionServer | | | CDH | 4 | 10240 | (Optional Service) Suggestion: Allocate no more than 12-16GB memory when using HBase Region Servers. |
| Solr Server | | | CDH | 0 | 0 | (Optional Service) Suggestion: Minimum 1GB for Solr server. More will be necessary depending on index sizes. |
| YARN NodeManager | | | CDH | 1 | 1024 | Allocate 1GB for the YARN NodeManager. |
| Available Resources | | |  | 39 | 97280 |  |
| Physical Cores to Vcores Multiplier | | |  | 1 |  | Set this ratio based on the expected number of concurrent threads per core. Use 1 for CPU intensive tasks up to 4 for standard I/O bound tasks. |
| YARN Available Vcores | | |  | 39 |  | This value will be used in STEP 4 for YARN Configuration |
| YARN Available Memory | | |  |  | 97280 | This value will be used in STEP 4 for YARN Configuration |

Results:

These are the values that should be set for the YARN configuration.

|  |  |  |
| --- | --- | --- |
| YARN Configuration Property | Value |  |
| yarn.nodemanager.resource.cpu-vcores | 39 | Copied from STEP 2 "Available Resources" |
| yarn.nodemanager.resource.memory-mb | 97280 | Copied from STEP 2 "Available Resources" |

|  |  |  |
| --- | --- | --- |
| YARN Container Configuration Property (Vcores) | Value | Description |
| yarn.scheduler.minimum-allocation-vcores | 1 | Minimum vcore reservation for a container |
| yarn-scheduler.maximum-allocation-vcores | 20 | Maximum vcore reservation for a container |
| yarn.scheduler.increment-allocation-vcores | 1 | Vcore allocations must be a multiple of this value |
| YARN Container Configuration Property (Memory) | Value | Description |
| yarn.scheduler.minimum-allocation-mb | 1024 | Minimum memory reservation for a container |
| yarn.scheduler.maximum-allocation-mb | 80000 | Maximum memory reservation for a container |
| yarn.scheduler.increment-allocation-mb | 512 | Memory allocations must be a multiple of this value |

|  |  |  |
| --- | --- | --- |
| Property | Value | Description |
| yarn.app.mapreduce.am.resource.cpu-vcores | 1 | AM container vcore reservation |
| yarn.app.mapreduce.am.resource.mb | 1024 | AM container memory reservation |
| ApplicationMaster Java Maximum Heap Size (available in CM) | 768 | AM Java heap size |
| mapreduce.map.cpu.vcores | 1 | Map task vcore reservation |
| mapreduce.map.memory.mb | 12000 | Map task memory reservation |
| mapreduce.map.java.opts.max.heap | 8000 | Map task Java heap size |
| mapreduce.reduce.cpu.vcores | 1 | Reduce task vcore reservation |
| mapreduce.reduce.memory.mb | 12000 | Reduce task memory reservation |
| mapreduce.reduce.java.opts.max.heap | 8000 | Reduce Task Java heap size |
| mapreduce.task.io.sort.mb | 256 | Spill/Sort memory reservation |