Cloudera Professional Services

Saketh Gadde and Jayakrishna Vutukuri

Bank Negara Indonesia - CDP Private Cloud

Version 1.0

9 Nov 2024

Version	Author	Description
1.0	Saketh Gadde	BNI Performance Tuning and Optimization

TABLE OF CONTENTS

IMPORTANT NOTICE

Cloudera, the Cloudera logo, and any other product or service names or slogans contained in this document, except as otherwise disclaimed, are trademarks of Cloudera and its suppliers or licensors, and may not be copied, imitated or used, in whole or in part, without the prior written permission of Cloudera or the applicable trademark holder.

Hadoop and the Hadoop elephant logo are trademarks of the Apache Software Foundation. All other trademarks, registered trademarks, product names and company names or logos mentioned in this document are the property of their respective owners to any products, services, processes or other information, by trade name, trademark, manufacturer, supplier or otherwise does not constitute or imply endorsement, sponsorship or recommendation thereof by us.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Cloudera.

Cloudera may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Cloudera, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

The information in this document is subject to change without notice. Cloudera shall not be liable for any damages resulting from technical errors or omissions which may be present in this document, or from use of this document.

Cloudera, Inc.

395 Page Mill Road
Palo Alto, CA 94306
info@cloudera.com
US: 1-888-789-1488
Intl: 1-650-362-0488
www.cloudera.com

1 Summary

1.1 Overview

This document summarizes the analysis, recommendations for performance and optimization, queries resolved on Bank Negara Indonesia CDP Private Cloud environment remotely undertaken by Saketh Gadde and Jayakrishna Vutukuri (Cloudera Solution Architects) alongside the BNI team.

2 Analysis & Recommendations

This section contains details of the recommendations and Analysis from Cloudera on the BNI CDP Private Cloud on Premises.

2.1 Hive service configuration Tuning

Optimizing Hive service configurations can significantly improve performance by efficiently allocating resources and enhancing the execution of queries. Proper tuning of Hive settings ensures better memory management, smoother data processing, and minimizes the likelihood of job failures due to resource constraints. The following recommendations focus on memory allocation, handling large datasets, and optimizing the execution of tasks in Hive.

Recommendations

*- **hive.tez.container.size** - for better memory allocation to processing containers, reducing the risk of out-of-memory errors and enhancing task efficiency.*

Increased from 2GB to 6GB.

*- **Hive Bytes Per Reducer** - to allow reducers to process larger chunks of data, minimizing the overhead of excessive reducer tasks.*

Increased from 64MB to 128MB

*- **Hive.auto.convert.join.noconditionaltask.size** - for enabling efficient in-memory handling of larger datasets during joins.*

Increased from 50MB to 128MB

2.2 Tez service configuration Tuning

Tez configurations are critical for optimizing resource utilization and improving the performance of Tez tasks. By adjusting these settings, we can reduce latency, improve memory management, and enhance task coordination across the cluster.

Recommendations

*- **Tez sort buffer size** - to enhance the performance of sorting operations within Tez tasks.*

Increased from 1638MB to 2457MB

*- **tez.runtime.unordered.output.buffer.size-mb** - to improve the handling of unordered output data, reducing latency.*

Increased from 100MB to 512MB

- tez.am.resource.memory-mb - *to allocate more memory to the Tez Application Master, enabling better job coordination and resource allocation.*

Increased from 2GB to 6GB

- tez.task.resource.memory-mb - *to provide sufficient memory for Tez tasks, avoiding memory-related failures.*

Increased from 1536MB to 5120MB

- tez.am.launch.cmd-opts - *for better memory management and efficient garbage collection.*

Modified from

"-XX:+PrintGCDetails -verbose:gc -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB" to

"-Xmx4096m -XX:+PrintGCDetails -verbose:gc -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB"

Please refer the following documentation for further details:

Optimizing Hive on Tez Performance - Cloudera Blog

2.3 Improvements observed after 2.1 & 2.2

Performance Gains:
Significant reduction in job execution times, with improved efficiency in handling large-scale data processing tasks. For example, the **transaction_loan query** saw a reduction in execution time from **2 hours to 1 hour**.

Resource Utilization:
Optimized memory and resource usage across Hive and Tez services, reducing the likelihood of job failures due to resource constraints.

Memory Efficiency:
Better memory utilization was observed, particularly for jobs utilizing ACID tables, resulting in fewer **Out-Of-Memory (OOM) errors**.

Scalability:
Enhanced scalability, enabling the cluster to better manage larger datasets and more complex queries, with improved parallelism and memory management.

2.4 System performance check

During the execution of the queries, we monitored the performance of the underlying servers, particularly focusing on how resources are being allocated to the running jobs and queries. Our assessment included evaluating CPU, RAM, and disk utilization to ensure that the system was operating within the expected parameters. The results showed that the resources were being efficiently utilized, with CPU usage remaining within the desired limits, RAM being appropriately allocated to tasks without excessive swapping, and disk I/O operating as anticipated.

Furthermore, we observed that the resource allocation was well-balanced across the nodes, ensuring that no single node was overloaded while others remained underutilized. This confirmed that the current infrastructure setup and configuration were well-suited to handle the workloads, leading to stable performance during query execution. The checks also revealed that there were no signs of resource contention or bottlenecks in the system, which could potentially impact the query performance or job execution times.

Overall, the system was performing optimally, with resources being efficiently distributed and utilized across the cluster, leading to improved overall job and query execution efficiency.

2.5 Yarn service configuration tuning

We conducted a review of the YARN configurations to assess resource management and job scheduling efficiency across the cluster. The analysis revealed that the existing YARN settings were already well-optimized for the current workloads and did not require any adjustments.

Key Findings:

- **Resource Allocation**: The memory and CPU configurations for YARN containers were found to be appropriately allocated for the tasks being executed.
- **Job Scheduling**: The YARN ResourceManager effectively managed resource requests from Spark, Hive, and Impala jobs, ensuring that cluster resources were utilized efficiently.
- **Queue Management**: The queue management policies, including fair resource allocation, were functioning as expected. These settings ensured that resources were evenly distributed across jobs and users, preventing overloading of any individual queue.

Conclusion:

Since no configuration changes were required, it was concluded that the current YARN setup is well-suited to handle the data processing workloads without any performance degradation or resource contention. This confirms that the infrastructure and configuration settings are stable, contributing to the efficient execution of jobs within the cluster.

2.6 Impala service configuration tuning

To optimize **Impala's** performance for large datasets and complex queries, adjustments were made to the default configurations. These changes were based on an evaluation of the system's resource utilization and an identification of memory allocation bottlenecks that were hindering processing efficiency.

Reason

Inadequate memory allocation for various Impala components, including daemons and coordinators, was limiting the system's capacity to process large datasets and manage concurrent workloads effectively.

Recommendations

Once the service is started all the impala components, we have tuned the service with following configurations:

- ***Increased Impala Daemon default group java heap*** - *to support larger data volumes and improve query processing efficiency.*

  Increased from 20GB to 128GB

- ***Increased Impala_Expansi_2024 java heap*** - *to enhance the scalability and performance of data expansion operations.*

  Increased from 20GB to 128GB

- ***Increased Impala Daemon group 1 java heap*** - *to ensure consistent memory availability for critical processes.*

  Increased from 20GB to 128GB

- ***Increased LB-Coordinator*** - *for better load balancing and coordination across nodes.*

  Increased from 32GB to 128GB

- ***Increased Impala-Coordinator*** - *to boost the performance of coordination tasks and metadata management.*

  Increased from 32GB to 128GB

- ***Increased Impala daemon Memory limit*** - *to accommodate larger in-memory operations and reduce disk I/O overhead.*

  Increased from 32GB to 128GB

Improvement

Performance Gains:
The updated configurations led to a significant improvement in **query execution times**, particularly for large and complex workloads.

Memory Efficiency:
Incremental reduction in **memory-related failures** was observed as a result of increased memory allocations for critical Impala components, contributing to more efficient resource utilization.

Scalability:
The system is now better equipped to handle larger datasets and more complex queries, improving Impala's ability to scale efficiently under heavy concurrent workloads.

2.7 Spark service configuration tuning

After analyzing the performance of the **Spark service**, it was determined that the default configurations were not optimized for handling larger datasets and complex workloads. Several adjustments were made to improve resource allocation and execution efficiency, enabling better handling of high-volume data processing tasks within the cluster.

Reason

Spark's default executor memory and core settings were suboptimal for processing large datasets. Insufficient memory allocation caused excessive garbage collection and slow processing, while limited CPU cores hindered parallel task execution, resulting in increased job durations and performance bottlenecks.

Recommendations

- ***spark.executor.memory** - to allocate more memory to Spark executors, allowing them to handle larger datasets without running into memory bottlenecks or out-of-memory errors.*

Increased from 8GB to 24GB

- ***spark.executor.cores** - to provide more CPU cores to each executor, enabling them to process tasks more efficiently and in parallel, reducing job execution time.*

Increased from 2 to 4 cores

Improvement

Following the recommended changes, significant performance improvements were observed across Spark jobs, as confirmed through **Cloudera Observability**. The adjustments in executor memory and cores allowed Spark to process larger datasets more efficiently, resulting in reduced job completion times and increased overall throughput. Additionally, memory-related failures were minimized, providing more stability for Spark jobs. This tuning has led to optimized performance and better resource utilization for big data workloads.

Performance Gains:

- **BNI_CUSTOMER_DEV**: Reduced average runtime from **4 hours 48 minutes** to **15 minutes** for processing approximately **450GB** of data, demonstrating substantial optimization.
- **Mart_ACCOUNT_BNI_DIRECT.PY**: Consistently maintained execution times of **6-8 minutes** for processing **19.4GB** of data.
- **cust_perf_trx.py**: Achieved a **23% reduction** in runtime, improving from **52 minutes** to **40 minutes** while processing **394GB** of data.
- **mart_user_e_channel_mobile.py**: Improved job reliability and optimized data throughput of **up to 3.4TB**, with runtimes averaging **2 hours**, despite occasional spikes in workload.

2.8 Optimizing HDFS Efficiency by Addressing Small File Issues

Small Files in Hive Databases

Identified Small File Issues: During our analysis, we discovered a significant number of small files in both Hive External and Managed Databases, which can lead to inefficiencies in HDFS and increased overhead on the NameNode:

- **Hive External Databases**: 2,016,1914 small files identified.
- **Hive Managed Databases**: 660,2052 small files identified.

These findings prompted a set of recommendations to optimize HDFS storage and improve performance.

Reason

In big data environments, the presence of numerous small files increases the overhead for the HDFS NameNode, which has to manage the metadata for each individual file. This results in higher latency in file operations and contributes to the exhaustion of memory and CPU resources.

Recommendation for Hive to reduce small files

Remove Bucketing for ACID Tables

- *Bucketing in ACID tables can lead to small file creation, especially when partitions remain underutilized.*
- *The following tables are recommended for bucketing removal:*
    - ***`prd_acl_datalake.mart_customer`*** *(8 buckets)*
    - ***`prd_acl_datalake.mart_debitcard`*** *(9 buckets)*
    - ***`prd_acl_datalake.mart_miscrd`*** *(4 buckets)*
    - ***`prd_acl_datalake.mart_miscus`*** *(4 buckets)*

Optimize Non-ACID Tables

- *Non-ACID tables with excessive or inappropriate bucketing configurations contribute to small file generation.*
- *Recommendations:*
    - *Remove bucketing for partitions consistently smaller than **1GB**.*
    - *Reduce the number of buckets for partitions expected to grow beyond **1GB**.*
- *Affected tables include:*
    - ***`prd_acl_datalake.mart_dpk`*** *(8 buckets)*
    - ***`prd_acl_datalake.mart_closed_account`*** *(4 buckets)*
    - ***`prd_acl_datalake.mart_customer_performance_baseline`*** *(4 buckets)*
    - ***`prd_acl_datalake.mart_trx_virtual_account`*** *(8 buckets)*
    - ***`prd_acl_datalake.mart_master_segment_nwow_daily`*** *(4 buckets)*
    - ***`prd_sri_datalake.sri_155_136_jt0004loan_minicore`*** *(5 buckets)*
    - ***`prd_acl_datalake.mart_dpk_dn_daily`*** *(4 buckets)*
    - ***`prd_acl_datalake.mart_cta`*** *(5 buckets)*

Improvement

By implementing these recommendations, the number of small files generated in Hive tables will be significantly reduced. This will lead to more efficient storage utilization and lower overhead on the NameNode. As a result, HDFS will experience improved performance in both file management and query execution.

2.9 CDSW Configurations Review

A detailed review of the **Cloudera Data Science Workbench (CDSW)** node was conducted to analyze performance, focusing on CPU and memory usage across various pods. The current resource configurations for these pods were evaluated for optimization opportunities.

Cause

The review focused on performance metrics, specifically CPU and memory usage, to identify any inefficiencies or bottlenecks in resource allocation.

Based on the current workload, the allocated resources for all pods were deemed adequate, with no evidence of resource contention or excessive usage.

Recommendation

No Changes Required:
The current configuration for CPU and memory requests and limits across all pods is sufficient. The following configurations are recommended to remain unchanged:

- **Web Pod**: CPU: 100m, MEM: 2G (Request: 350Mi, Limit: 2Gi)
- **Livelog**: CPU: 200m, MEM: 500Mi (Request: 200Mi, Limit: 500Mi)
- **TCP Ingress Control**: CPU: 100m, MEM: 2G (Request: 250Mi, Limit: 2Gi)
- **DS Reconciler**: CPU: 100m, MEM: 200Mi (Request: 100Mi, Limit: 200Mi)
- **DS Operator**: CPU: 100m, MEM: 200Mi (Request: 100Mi, Limit: 200Mi)
- **DB Pod**: CPU: 200m, MEM: 2G (Request: 2G, No Limit)

Improvement

Performance Stability:
Maintaining the current configurations ensures stable and predictable performance of CDSW nodes.

Resource Optimization:
With no unnecessary adjustments, the system avoids resource over-provisioning or under-utilization, resulting in efficient resource management.

Scalability:
The existing configuration provides adequate flexibility to accommodate potential workload increases without immediate adjustments.

2.10 Agent client certificate expiration

An expired agent client certificate can cause the Cloudera Manager agents to stop being authenticated by the Cloudera Manager server, which can affect Cloudera Manager's cluster management functionality among other things. Additionally, if the same certificate is in use by services on the host with TLS enabled, their functionality may also cease, with certificate validation errors at dependent services and clients.

It is vital to renew the cluster's host certificates (including that used by the Cloudera Manager agents) well before they expire, for proper continued functioning and availability of the cluster's Cloudera Manager and other TLS-enabled services.

The earliest found date of certificate expiry is on **2025-01-17 23:59:59** (for host **datalakedcdm1.bni.co.id**).

For more details on Cloudera Manager TLS features, please refer:

Document_link:
https://www.cloudera.com/documentation/enterprise/latest/topics/cm_sg_config_tls_security.html

Hosts Affected -
datalakedcdm1.bni.co.id, datalakedcedge1.bni.co.id, datalakedcetl1.bni.co.id, datalakedcm1.bni.co.id, datalakedcm2.bni.co.id, datalakedcm3.bni.co.id, datalakedcw1.bni.co.id, datalakedcw2.bni.co.id, datalakedcw3.bni.co.id, datalakedcw4.bni.co.id, datalakedcw5.bni.co.id, datalakedcw6.bni.co.id, datalakedcw7.bni.co.id, datalakedcw8.bni.co.id, odbdcapi1.bni.co.id, odbdcapi2.bni.co.id, odbdcm3.bni.co.id

Service Name - Cloudera Manager
Role Type - CM_AGENT

2.10 Consistent OS Validator

Inconsistent operating systems are in use across this cluster. This is contrary to our setup guidelines:

Document_link:

- **Setup Guidelines DOC_LINK**
  http://www.cloudera.com/documentation/enterprise/release-notes/topics/rn_consolidated_pcm.html#cdh_cm_supported_os

For advice on upgrading cluster operating systems, see the KB article:

- **KB Article DOC_LINK**
  https://cloudera-portal.force.com/articles/KB_Article/Upgrading-the-Operating-System-on-the-nodes-in-a-CDH-cluster

Hosts Affected :

datalakedcdm1.bni.co.id, datalakedcedge1.bni.co.id, datalakedcetl1.bni.co.id, datalakedcm1.bni.co.id, datalakedcm2.bni.co.id, datalakedcm3.bni.co.id, datalakedcw1.bni.co.id, datalakedcw10.bni.co.id, datalakedcw11.bni.co.id, datalakedcw12.bni.co.id, datalakedcw13.bni.co.id, datalakedcw14.bni.co.id, datalakedcw15.bni.co.id, datalakedcw16.bni.co.id, datalakedcw17.bni.co.id, datalakedcw18.bni.co.id, datalakedcw19.bni.co.id, datalakedcw2.bni.co.id, datalakedcw3.bni.co.id, datalakedcw4.bni.co.id, datalakedcw5.bni.co.id, datalakedcw6.bni.co.id, datalakedcw7.bni.co.id, datalakedcw8.bni.co.id, datalakedcw9.bni.co.id, odbdcapi1.bni.co.id, odbdcapi2.bni.co.id, odbdcm3.bni.co.id

2.11 DataNode-dependent roles running without a DataNode

4 host(s) were found running worker roles but carry no assigned DataNode role. A functioning local DataNode role is necessary for proper performance of the other worker roles.

Worker Roles:

Defined as any of the following: NODEMANAGER, REGIONSERVER, TASKTRACKER, SOLR_SERVER (including SOLR_INFRA).

This warning can be ignored if all the collections of this SOLR instance are on the local filesystem.

Document_link:

https://www.cloudera.com/documentation/enterprise/latest/topics/cm_ig_host_allocations.html

Hosts Affected:

datalakedcm2.bni.co.id, datalakedcm3.bni.co.id, odbdcapi2.bni.co.id, odbdcm3.bni.co.id

Service Name: HDFS

Role Type: DATANODE

Affected Roles: cm_server

2.12 Hbase WAL Provider

HBase currently has an open issue when the Region Servers have the write-ahead-log (WAL) provider set to 'multiwal', as documented in the upstream Jira HBASE-20503. To avoid this issue, Cloudera Support recommends setting both **'hbase.wal.provider'** and **'hbase.wal.meta_provider'** to a single HDFS WAL in Cloudera Manager. Please note that **'hbase.wal.meta_provider'** is not exposed in CM, and you will need to add the property via **'hbase_service_config_safety_valve'** with the value of **'filesystem'**.

Document_link: https://my.cloudera.com/knowledge/HBASE-is-not-responding--ERROR-handlerOpenRegionHandler-?id=291709

2.13 High Availability Role Configured

A majority of Cloudera Data Platform Services support a High Availability Configuration by running multiple instances of roles. The following services did not appear to have HA configured:

- The ATLAS_SERVER role from ATLAS did not have sufficient instance count. Please review this link: https://docs.cloudera.com/cdp-private-cloud-base/7.1.7/atlas-ha/topics/atlas-high-availability-overview.html
- The RANGER_ADMIN role from RANGER did not have sufficient instance count. Please review this link: https://docs.cloudera.com/cdp-private-cloud-base/7.1.7/security-ranger-configuring-advanced/topics/security-ranger-configuring-admin-ha-load-balancer.html
- The OOZIE_SERVER role from OOZIE did not have sufficient instance count. Please review this link: https://docs.cloudera.com/cdp-private-cloud-base/7.1.7/configuring-oozie/topics/oozie-high-availability.html

Document_link: https://docs.cloudera.com/cdp-private-cloud-base/latest/installation/topics/cdpdc-runtime-cluster-hosts-role-assignments.html

Role Type: GATEWAY

2.14 Host Inspector Validator Warning 1

A host is expected to have a single name but resolved itself to a different one (InetAddress.getLocalHost().getHostName()).

Document_link:

https://docs.cloudera.com/documentation/enterprise/5-14-x/topics/cdh_ig_networknames_configure.html

Hosts Affected :

datalakedcedge1.bni.co.id

2.15 Host Inspector Validator Warning 2

All the hosts in the cluster are expected to have one name but resolved (InetAddress.getLocalHost().getHostName()) themselves to another.

Document_link:

https://docs.cloudera.com/documentation/enterprise/5-14-x/topics/cdh_ig_networknames_configure.html

Hosts Affected:

datalakedcdm1.bni.co.id, datalakedcedge1.bni.co.id, datalakedcetl1.bni.co.id, datalakedcm1.bni.co.id, datalakedcm2.bni.co.id, datalakedcm3.bni.co.id, datalakedcw1.bni.co.id, datalakedcw10.bni.co.id, datalakedcw11.bni.co.id, datalakedcw12.bni.co.id, datalakedcw13.bni.co.id, datalakedcw14.bni.co.id, datalakedcw15.bni.co.id, datalakedcw16.bni.co.id, datalakedcw17.bni.co.id, datalakedcw18.bni.co.id, datalakedcw19.bni.co.id, datalakedcw2.bni.co.id, datalakedcw3.bni.co.id, datalakedcw4.bni.co.id, datalakedcw5.bni.co.id, datalakedcw6.bni.co.id, datalakedcw7.bni.co.id, datalakedcw8.bni.co.id, datalakedcw9.bni.co.id, odbdcapi1.bni.co.id, odbdcapi2.bni.co.id, odbdcm3.bni.co.id

2.16 Host Inspector Validator Warning 3

Host expected to have one canonical name but resolved (InetAddress.getLocalHost().getCanonicalHostName()) itself to another.

Document_link:

https://docs.cloudera.com/documentation/enterprise/5-14-x/topics/cdh_ig_networknames_configure.html

Hosts Affected:

datalakedcedge1.bni.co.id

2.17 Host Inspector Validator Warning 4

The hostname datalakedcedge1.bni.co.id is not the first match for address 192.168.104.37 in /etc/hosts. Instead, datalakedcedge01.bni.co.id is the first match. The FQDN must be the first entry in /etc/hosts for the corresponding IP.

Document_link:

https://docs.cloudera.com/documentation/enterprise/5-14-x/topics/cdh_ig_networknames_configure.html

Hosts Affected:

datalakedcedge1.bni.co.id

2.18 Host Inspector Validator Warning 5

Transparent Huge Pages is enabled and can cause significant performance problems. Please see the KB Article for further details. Additionally, the following product documentation exists for version 5.x/6.x. For CDP 7.x, please see the product documentation.

Document_link:

https://cloudera-portal.force.com/articles/KB_Article/Poor-Performance-When-Transparent-Hugepage-Compaction-is-Enabled

https://docs.cloudera.com/documentation/enterprise/latest/topics/cdh_admin_performance.html#cdh_performance__section_hw3_sdf_jq

https://docs.cloudera.com/cdp-private-cloud-base/latest/managing-clusters/topics/cm-disabling-transparent-hugepages.html

Hosts Affected:

odbdcapi1.bni.co.id, odbdcapi2.bni.co.id

2.19 Hue-dependent services (Hbase)

The Hbase service (hbase) could be managed by Hue on this CDP Private Cloud Base cluster, but is not configured for management. There is no requirement to use HUE, but if you wish to use HUE for Hbase, you need to specify which service to manage.

Document_link:

https://docs.cloudera.com/cdp-private-cloud-base/7.1.6/administering-hue/topics/hue-enable-applications-cloudera-manager.html

Service Name:
HUE

2.20 Increased Heap Memory in Reports Manager in Cloudera Manager (Cloudera Customer Advisory 2022-571)

There is increased heap memory usage in Reports Manager in Cloudera Manager 7.4.4 and 7.6.1, which ship, respectively, as part of CDP Private Cloud Base 7.1.7 and 7.1.7 SP1. This issue is documented as Cloudera Customer Advisory 2022-571.

Because of this:

Customers should not place the Reports Manager on the same host as the Service Monitor, as it can cause CPU usage issues.

Cloudera recommends provisioning a host as powerful as possible and dedicating an SSD disk to the Reports Manager.

The Reports Manager heap size should be set to 4 x the fsimage size + additional 2GB at the following location in Cloudera Manager:

Cloudera Management Service > Configuration > headlamp_heapsize.

Additionally, due to the increased heap requirement, Cloudera recommends up to an additional 4x this amount as per the Customer Advisory.

Instead of attempting to increase the Reports Manager heap size as per the above recommendations, Cloudera Support strongly suggests upgrading to the latest version of Cloudera Manager compatible with your version of CDP Private Cloud Base to take advantage of the official fix.

Document_link:

https://docs.cloudera.com/cdp-private-cloud-base/7.1.7/installation/topics/cdpdc-reports-manager.html

https://my.cloudera.com/knowledge/Cloudera-Custom-Advisory-Increased-heap-memory-in-Reports?id=340976

2.21 IP Tables

22 of 28 hosts in this CDP Private Cloud Base cluster have connection tracking enabled in the kernel, even though there aren't firewall rules. This is a needless performance hit.

Review the host command **lsmod** and search the output for **nf_conntrack**.

Hosts Affected:
datalakedcetl1.bni.co.id, datalakedcm1.bni.co.id, datalakedcm2.bni.co.id, datalakedcm3.bni.co.id, datalakedcw1.bni.co.id, datalakedcw10.bni.co.id, datalakedcw11.bni.co.id, datalakedcw12.bni.co.id, datalakedcw13.bni.co.id, datalakedcw14.bni.co.id, datalakedcw15.bni.co.id, datalakedcw16.bni.co.id, datalakedcw17.bni.co.id, datalakedcw18.bni.co.id, datalakedcw19.bni.co.id, datalakedcw2.bni.co.id, datalakedcw3.bni.co.id, datalakedcw5.bni.co.id, datalakedcw6.bni.co.id, datalakedcw7.bni.co.id, datalakedcw8.bni.co.id, datalakedcw9.bni.co.id

Document Link:

https://www.linux.org/threads/what-is-nf_conntrack-used-for.11924/

2.22 Java VM Options Validation

The following role(s) in the cluster appear to be using the JVM GC Log options which are now deprecated and removed starting in JDK 11. If you upgrade the JDK version without removing all the properties on this list:

- -d64
- PrintGCTimeStamps
- PrintGCDateStamps
- PrintTenuringDistribution

- PrintGCCause
- PrintGCDetails
- UseGCLogFileRotation
- NumberOfGCLogFiles
- GCLogFileSize

The role(s) will fail to start with a message similar to:
'Unrecognized VM option 'UseGCLogFileRotation'
Error: Could not create the Java Virtual Machine.
Error: A fatal exception has occurred. The program will exit.

The following role types and their config key carrying one or more of these options:

- Hive_On_Tez
- HIVESERVER2

Document Link:

https://confluence.atlassian.com/confkb/unrecognized-jvm-gc-options-when-using-java-11-1002472841.html

2.23 Kafka: Estimated Heap Size of Broker JVMs

3 Kafka Broker(s) on this cluster were found using a very low heap size (1.00 GiB), well under the minimum recommended value of 4 GiB per Broker.Cloudera strongly recommends specifying a minimum heap size of 4 GiB under Cloudera Manager - Kafka - Configuration - 'Java Heap Size of Broker' for effective use of Kafka in production environments. A very low heap size may result in frequent crashes due to OutOfMemoryErrors and/or slow and intermittent throughput performance due to several Garbage Collection pauses in Kafka's JVM.
Hosts Affected - datalakedcw1.bni.co.id, datalakedcw2.bni.co.id, datalakedcw3.bni.co.id

Service Name - KAFKA

Role Type - KAFKA_BROKER

Document Link:
KB_LINK
https://docs.cloudera.com/cloudera-manager-ibm/7.2.3/installation/topics/cdpdc-kafka.html

2.24 Kerberos configuration

Some hosts on this cluster do not have a KDC timeout set. The default timeout is 30 seconds, which is too long. We advise setting a 3000 msec kdc_timeout in krb5.conf to avoid excessive waits. For further reference please see our documentation.

Hosts Affected :
datalakedcdm1.bni.co.id, datalakedcedge1.bni.co.id, datalakedcetl1.bni.co.id,

datalakedcm1.bni.co.id, datalakedcm2.bni.co.id, datalakedcm3.bni.co.id, datalakedcw1.bni.co.id, datalakedcw10.bni.co.id, datalakedcw11.bni.co.id, datalakedcw12.bni.co.id, datalakedcw13.bni.co.id, datalakedcw14.bni.co.id, datalakedcw15.bni.co.id, datalakedcw16.bni.co.id, datalakedcw17.bni.co.id, datalakedcw18.bni.co.id, datalakedcw19.bni.co.id, datalakedcw2.bni.co.id, datalakedcw3.bni.co.id, datalakedcw4.bni.co.id, datalakedcw5.bni.co.id, datalakedcw6.bni.co.id, datalakedcw7.bni.co.id, datalakedcw8.bni.co.id, datalakedcw9.bni.co.id, odbdcapi1.bni.co.id, odbdcapi2.bni.co.id, odbdcm3.bni.co.id

Document Link:

https://docs.cloudera.com/documentation/enterprise/release-notes/topics/cm_rn_known_issues.html#concept_brv_1q4_15

2.25 Linux - Wider Ephemeral Port Range

1 host(s) carry a reconfigured Linux kernel parameter of 'net.ipv4.ip_local_port_range', wherein the lower-bound of the ephemeral port range is set to lower than the Linux default value of 32768. This can result in service or host restart failures due to transient network client connections squatting on ports required by various Cloudera services. Services may fail with the error 'Address already in use' (EADDRINUSE) when being restarted, which likely disappears when reattempted.

The ports reference for all Cloudera services can be referenced in the Cloudera documentation, wherein the majority of ports lie under the default Linux ephemeral range (32768). The default for Linux is further defined at the Wikipedia link. It is recommended to keep the port range configured at its Linux default value of 32768 through 61000. The current value on the mentioned hosts below can be checked by running the command:
$ sysctl net.ipv4.ip_local_port_range

To reset to the default recommended port range, run the below command in addition to following the guide for persistent configuration:
$ sysctl -w net.ipv4.ip_local_port_range='32768 61000'

As an example, the lowest value observed for lower-bound port configuration was '9000' on host datalakedcm3.bni.co.id.

Hosts Affected:

datalakedcm3.bni.co.id

Document Link:

https://docs.cloudera.com/documentation/enterprise/latest/topics/cm_ig_ports.html ,

https://en.wikipedia.org/wiki/Ephemeral_port#Range , https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_MRG/1.3/html/Realtime_Tuning_Guide/sect-Realtime_Tuning_Guide-General_System_Tuning-Setting_persistent_tuning_parameters.html

2.26 Master/Worker Contention Warning 1

Some hosts are running both Kafka and other Hadoop processes. This is generally unwise, since Kafka can be very resource-intensive. We recommend that Kafka be installed on dedicated hosts that are not used for other cluster roles.

Hosts Affected :

datalakedcw1.bni.co.id, datalakedcw2.bni.co.id, datalakedcw3.bni.co.id

Document Link:
https://docs.cloudera.com/documentation/kafka/latest/topics/kafka_installing.html

2.27 Master/Worker Contention Warning 2

Some hosts in this CDP Private Cloud Base cluster are running Cloudera Manager server (and its management roles) along with other CDP roles and services. CM Server (and its management roles) should be isolated away from the actual cluster, in the event of an outage of that node it would not impact the running cluster.

Hosts Affected:

datalakedcm3.bni.co.id

Document Links:
https://docs.cloudera.com/cdp-private-cloud-base/7.1.6/installation/topics/cdpdc-runtime-cluster-hosts-role-assignments.html ,
https://blog.cloudera.com/deploy-cloudera-edh-clusters-like-a-boss-revamped-part-2/

2.28 Nameservice Caching Daemon

Level: WARN

3 host(s) seems to use both NSCD and SSSD. SSSD is not designed to be used with the NSCD daemon. Even though SSSD does not directly conflict with NSCD, using both services can result in unexpected behavior, especially with how long entries are cached. To avoid this problem, enable caching only for hosts in the /etc/nscd.conf file and rely on the SSSD cache for the passwd, group, services, and netgroup entries.

Hosts Affected :

datalakedcm3.bni.co.id, odbdcapi2.bni.co.id, odbdcm3.bni.co.id

Document Link:
https://docs.cloudera.com/cdp-private-cloud-upgrade/latest/release-guide/topics/cdpdc-os-requirements.html

2.29 Network Parameter Check

28 host(s) have one or more network parameter values out of alignment with the Cloudera recommended value. Below are the 10 network parameters in question along with what we recommend for each one. Please refer to the sysctl command output from the affected hosts for details on the current values of each parameter in question:

- net.core.rmem_default (recommend: 4194304)
- net.core.netdev_max_backlog (recommend: 250000)
- net.ipv4.tcp_low_latency (recommend: 1)
- net.core.rmem_max (recommend: 4194304)
- net.core.wmem_max (recommend: 4194304)
- net.core.optmem_max (recommend: 4194304)
- net.core.somaxconn (recommend: 8192)
- net.core.wmem_default (recommend: 4194304)
- net.ipv4.tcp_wmem (recommend: 4096 65536 4194304)
- net.ipv4.tcp_max_syn_backlog (recommend: 8192)

Hosts Affected :

datalakedcdm1.bni.co.id, datalakedcedge1.bni.co.id, datalakedcetl1.bni.co.id, datalakedcm1.bni.co.id, datalakedcm2.bni.co.id, datalakedcm3.bni.co.id, datalakedcw1.bni.co.id, datalakedcw10.bni.co.id, datalakedcw11.bni.co.id, datalakedcw12.bni.co.id, datalakedcw13.bni.co.id, datalakedcw14.bni.co.id, datalakedcw15.bni.co.id, datalakedcw16.bni.co.id, datalakedcw17.bni.co.id, datalakedcw18.bni.co.id, datalakedcw19.bni.co.id, datalakedcw2.bni.co.id, datalakedcw3.bni.co.id, datalakedcw4.bni.co.id, datalakedcw5.bni.co.id, datalakedcw6.bni.co.id, datalakedcw7.bni.co.id, datalakedcw8.bni.co.id, datalakedcw9.bni.co.id, odbdcapi1.bni.co.id, odbdcapi2.bni.co.id, odbdcm3.bni.co.id

Document Link:
https://docs.cloudera.com/documentation/other/reference-architecture/PDF/cloudera_ref_arch_cdp_dc.pdf

2.30 Performance Degradation Due To Tuned Service

Tuned service was found enabled on 16 host(s). A sample 'systemctl -p status' command output showing tuned service:
Host: datalakedcw18.bni.co.id

```
UNIT FILE STATE
proc-sys-fs-binfmt_misc.automount static
-.mount generated
```

boot-efi.mount generated
tuned.service enabled (...)

Use of tuned daemon service is discouraged on RHEL7 hosts, which comes as the default daemon in RHEL 7 version of Linux. Tuned daemon is introduced to boost the performance of the OS as per needs. However, enabling one profile will have an impact on other profiles. If the tuned configuration is set to performance or other acceptable values, then the warning can be ignored. Though, we still recommend not using tuned for cluster hosts due to varied workloads.

Hosts Affected :

datalakedcdm1.bni.co.id, datalakedcetl1.bni.co.id, datalakedcw10.bni.co.id, datalakedcw11.bni.co.id, datalakedcw12.bni.co.id, datalakedcw13.bni.co.id, datalakedcw14.bni.co.id, datalakedcw15.bni.co.id, datalakedcw16.bni.co.id, datalakedcw17.bni.co.id, datalakedcw18.bni.co.id, datalakedcw19.bni.co.id, datalakedcw9.bni.co.id, odbdcapi1.bni.co.id, odbdcapi2.bni.co.id, odbdcm3.bni.co.id

Service Name:

CDH

Document Link:
https://docs.cloudera.com/documentation/enterprise/5-14-x/topics/cdh_admin_performance.html#cdh_performance__disable-tuned

2.31 Validate Solr Best Practice Configuration

From the SOLR configurations in this CDP Private Cloud Base environment, we identified that slab count is not set according to the best practices. Configuring Apache Solr memory properly is critical for production system stability and performance. We recommend changing the value of slab count according to the direct memory size. Using the formula from this blog (https://blog.cloudera.com/apache-solr-memory-tuning-for-production/): 21474836480 (solr_java_direct_memory_size in bytes) * 0.7 / 134217728 (block size in bytes) = 112 the recommendation for solr_hdfs_blockcache_slab_count. The current value is 1.

Document_link:
https://blog.cloudera.com/apache-solr-memory-tuning-for-production/

2.32 Validate Solr Best Practice Configuration

Default value of archive.max.files in Solr is 100, which is high and typically unnecessary for the processed spool files. Lowering the value will avoid disk space consuming space and requiring manual deletion when there is an urgent requirement for space. This knowledge-based article provides directions to tune to a more reasonable size such as 3.

Document_link:
https://my.cloudera.com/knowledge/Solr-Audit-data-tuning-on-File-spooling-archivemaxfiles-and?id=371469

2.33 YARN's Job History Server UI Access Level Control

The default setting for yarn.webapp.filter-entity-list-by-user is set to 'true', which will block admins from viewing the jobs of other users. Based on your requirements and internal security policies this setting can be adjusted as needed.

Document_link:
https://docs.cloudera.com/runtime/7.2.10/yarn-reference/topics/yarn-config-parameters.html

2.34 Erasure Coding Speed Check

28 host(s) have one or more network adaptors with an insufficient network connection speed. When an adapter has an insufficient network connection speed, Cloudera CDP Private Cloud Base cluster performance can be severely degraded and other unexpected errors can occur. Cloudera Support recommends at least a 10 GB network connection if Erasure Coding is in use. Insufficient network connection speed values found: 2000mb/s, 1000mb/s.

Hosts Affected : datalakedcdm1.bni.co.id, datalakedcedge1.bni.co.id, datalakedcetl1.bni.co.id, datalakedcm1.bni.co.id, datalakedcm2.bni.co.id, datalakedcm3.bni.co.id, datalakedcw1.bni.co.id, datalakedcw10.bni.co.id, datalakedcw11.bni.co.id, datalakedcw12.bni.co.id, datalakedcw13.bni.co.id, datalakedcw14.bni.co.id, datalakedcw15.bni.co.id, datalakedcw16.bni.co.id, datalakedcw17.bni.co.id, datalakedcw18.bni.co.id, datalakedcw19.bni.co.id, datalakedcw2.bni.co.id, datalakedcw3.bni.co.id, datalakedcw4.bni.co.id, datalakedcw5.bni.co.id, datalakedcw6.bni.co.id, datalakedcw7.bni.co.id, datalakedcw8.bni.co.id, datalakedcw9.bni.co.id, odbdcapi1.bni.co.id, odbdcapi2.bni.co.id, odbdcm3.bni.co.id.

Document_link:
https://www.cloudera.com/documentation/enterprise/latest/topics/admin_hdfs_datadurability.html#concept_kmz_mbb_qcb

3 Queries resolved

Following are the issues/queries resolved and discussed.

3.1 Enabling Impala Resource Management and Resource Allocation

Issue: To ensure efficient resource utilization, Impala’s resource management can be enabled and fine-tuned through its configuration settings. In the current setup, YARN's resource management was disabled for Impala, necessitating the adjustment of specific Impala parameters to allocate resources effectively.

Configuration and Resource Allocation: Impala manages its own resource allocation layer, allowing fine-grained control over how resources are assigned to its queries. The key resource management configuration involved tuning the following setting:

- **Impala Daemon Memory Limit:** The **`impalad`** process, which is responsible for handling multiple concurrent queries, is allocated memory through the "Impala daemon memory limit" setting. This defines the overall memory limit for each **`impalad`** process, ensuring sufficient resources for query execution without overloading the node.
    - **Configured Limit:** 128GB

By adjusting this memory setting, the available resources for Impala queries were optimized, ensuring smooth query execution without excessive memory consumption, which could potentially lead to system instability or performance degradation.

Outcome: This configuration ensures that Impala queries run efficiently within the available memory limits, with the Impala Daemon managing the memory allocation for concurrent query executions.

3.2 Issue with Hive Job Execution and Memory Allocation

Request to Rerun Hive Job: We requested the BNI team to rerun a Hive job involving a query on the **`mart_transaction_loan`** ACID table to validate the effectiveness of the previous configurations, excluding any custom settings shared by the BNI team.

Issue Encountered After Rerun: Upon executing the job, the Hive job failed due to YARN terminating the container for exceeding the allocated memory.

Cause

The failure was caused by insufficient memory allocation for Tez tasks and the Application Master. This led to containers exceeding their allocated memory limits, resulting in job termination.

Further Recommendations for Memory Adjustment

To address the memory allocation issue, the following changes were recommended:

- **tez.am.resource.memory-mb**: Increased from 4GB to 6GB to provide more memory to the Tez Application Master.
- **tez.task.resource.memory-mb**: Increased from 4096MB to 5120MB to provide more memory for Tez tasks, reducing the likelihood of out-of-memory errors.
- **tez.am.launch.cmd-opts**: Updated from **`-Xmx4915m`** **to** **`-Xmx4096m`** to fine-tune memory usage during Tez job execution.

Outcome

The BNI team applied the recommended changes to the configurations, resulting in successful job execution with optimized memory allocation.

3.3 Hive Job Issue Analysis in Hue and Resolution Efforts

The BNI team requested an analysis and resolution for a Hive job in Hue that is encountering vertex failures, which is disrupting its normal execution time of 20 minutes.

Analysis

The initial analysis identified potential discrepancies in the metadata or data availability within the Hive environment as the possible cause of the vertex failures. To address these inconsistencies, an MSCK REPAIR operation was performed as an initial corrective measure.

Actions Taken

MSCK Repair Operation:

- We executed the **`MSCK REPAIR TABLE`** command on all relevant databases and tables to rebuild the metadata and synchronize Hive with HDFS.
- Despite performing the MSCK REPAIR operation, the vertex failure issue persisted, indicating that there may be other underlying causes beyond metadata inconsistencies.

Next Steps

- We have requested further information from the BNI team to perform a deeper analysis and determine additional potential causes for the vertex failures.

3.4 Analysis of Non-ACID Table (prd_acl_datalake.mart_loan) with Small File Problem

Issue: The **`prd_acl_datalake.mart_loan`** table was found to generate numerous small files (approximately 200 files, each 2.4 MB in size) on a daily basis. This resulted in inefficiencies in both data processing and storage management.

Action Taken: To address the issue, a duplicate table was created for analysis. The solution involved using the **`df.coalesce(10)`** method, which consolidated the data into 10 partitions of roughly equal size, reducing the number of small files generated.

Outcome: The solution effectively reduced the number of small files generated daily. This reduction in small file overhead helped optimize storage utilization and reduce the strain on the HDFS NameNode, leading to more efficient resource usage. The data is now better distributed, though further steps may be required to improve overall processing performance.

3.5 Ranger Issue with Views in CDSW Spark

Issue: Querying the schema view **`acl_datalake`** resulted in an error related to the underlying physical table schema **`prd_acl_datalake`**. Access to the **`prd_acl_datalake`** schema is restricted, but the error persists when querying the view.

Cause:

- **Spark and Hive Views:** When querying a Hive view through Spark, the Spark engine requires direct access to the underlying tables referenced by the view.
- **Permission Dependency:** Although users may have permissions on the view itself, the query will fail if they lack the necessary privileges on the underlying table, causing access errors.
- **Environment Consistency:** This behavior is inherent to the Spark engine and remains consistent across platforms, including Cloudera Data Science Workbench (CDSW) and base nodes.

Recommendation:

- **Use Impyla for Impala Queries:** In CDSW, configure Impyla as the client to execute Impala queries. This setup allows queries to bypass the need for permissions on the underlying table, ensuring consistent query execution behavior while maintaining access controls.

3.6 Analysis of High CPU and Memory Utilization in CDSW

Issue: The BNI team reported that CPU and memory utilization in Cloudera Data Science Workbench (CDSW) consistently reach near full capacity, resulting in insufficient available resources for job execution and causing delays due to job queuing.

Cause:

- **Limited Node Capacity:** The primary cause of high CPU and memory utilization is that CDSW is operating with only a single node. This configuration restricts the available compute and memory resources, leading to resource contention when multiple jobs are executed concurrently.
- **Resource Exhaustion:** Since there is only one node, the entire workload is constrained by the fixed capacity of that single node. As the resource demand from

multiple simultaneous jobs grows, the CPU and memory quickly become saturated, causing performance degradation and delays in job execution.

Recommendation:

- **Increase Worker Nodes in CDSW:** To resolve this issue, it is recommended to increase the number of worker nodes in the CDSW environment. By adding more nodes, additional compute and memory resources will be made available, which will help distribute the workload more evenly. This approach will not only improve resource availability but also reduce the likelihood of job queuing, leading to faster execution and better overall system performance.
- **Cluster Scaling:** Scaling the CDSW environment by adding more nodes will allow for better parallelization of jobs, enhancing both CPU and memory utilization efficiency. This will also help prevent resource bottlenecks that lead to job queuing, thus ensuring smoother execution of data science workflows in CDSW.

3.7 Spark Job Analysis: Out-of-Memory (OOM) Error

Issue: A Spark job encountered an "Out-of-Memory" (OOM) error during execution, specifically at the point of calling a SQL operation on a file format writer. The error message, "error while calling o79.sql.: org.apache.spark.sparkexception: job aborted.at.org.apache.spark.sql.execution.datasources.file.formatwriter$.write(fileFormatwriter.scala:203) at org.apache.spark.sql.hive.execution.saveas file$class.saveas", was logged, indicating a failure when attempting to write data to a file format.

Upon further investigation in the YARN ResourceManager UI, it was observed that the Spark driver lacked sufficient memory to handle resource-intensive operations, such as executing Hive INSERT statements. This led to a job failure, as the available resources were inadequate to complete the job's processing.

Job Details:

- **Job Name:** First long-running Spark job: **`/home/bnibdteam/bot/ingestion/MART/ACL/MART_CUSTOMER_CREDITCARD/etl_job_development.py`**

Cause: The root cause of the OOM error was identified as **bucketed ACID tables**. These tables, due to their structure and data distribution, require significant memory during processing, particularly during complex operations like INSERTS and joins. The Spark driver was unable to allocate sufficient memory to handle these operations efficiently, causing the job to fail with an OOM error.

Resolution: The issue was resolved by removing the bucketed ACID tables, which alleviated the memory strain on the Spark driver. Without the bucketed structure, the job was able to execute without exhausting memory resources, preventing the OOM error and allowing the job to complete successfully. This adjustment helped distribute the processing load more evenly across available memory resources, ensuring that memory usage was within acceptable limits for the Spark job to execute efficiently.

3.8 Insufficient Memory Due to Limited Node Resources in CDSW

Issue:

"Unschedulable 0/1 nodes are available": This error message indicates that CDSW has only one node available, and that node cannot schedule the requested job due to resource constraints.

"Insufficient memory": The job cannot be executed because the single node's memory is either fully utilized by other processes or the job requires more memory than the node can provide. This suggests that the current configuration is unable to meet the memory demand for the requested job.

Root Cause:

- The issue arises because there is **only one node** available in the **Cloudera Data Science Workbench (CDSW)** environment. With such a limited resource, the system cannot allocate sufficient memory to execute the job. This leads to the node being unable to schedule new tasks.

3.9 Guide for Tuning Static Pools in Your Cluster

For configuring and optimizing static pools for your cluster, no immediate changes are recommended. However, you can refer to the following resources for more details:

https://docs.cloudera.com/cdp-private-cloud-base/7.1.7/managing-clusters/topics/cm-static-service-pools.html ,

https://docs.cloudera.com/cdp-private-cloud-base/7.1.7/impala-manage/topics/impala-static-pool.html