

Ozone Performance Tuning

Date published: 2022-08-30

Date modified: 2024-07-19



Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

Configuring Ozone services.....	4
Performance tuning for Ozone.....	6
Node maintenance.....	7

Configuring Ozone services

This section helps you to understand the purpose of the parameters you need to set and the disk type you require for services like DataNode, OM, Recon, and SCM.



Note:

- Cloudera recommends you to configure the correct values during the initial configuration of the Ozone service for the below properties. Changing this configuration later can cause service disruption.

For example, Cloudera does not recommend moving or changing the data directory once Ozone service is in use. Changing or moving the data directory later is disruptive and brings down the Ozone service.

- You must configure the Ozone Manager Metadata directory and Ozone Manager Data directory in the same filesystem. If you configure the Ozone Manager Metadata directory and Ozone Manager Data directory on different filesystems on the host server, the hard link will fail causing the snapshot installation job to fail.

DataNode

Setting name	Default value	Purpose	Disktype	Mount point	Minimum size (GB)	Notes
ozone.metadata.dirs	/var/lib/hadoop-ozone/datanode/ozone-metadata	One or more directories used for storing Ozone metadata. OzoneManager, SCM, and Datanode will write the metadata to this path.	HDD	/var/app/lib/	100	
dfs.container.ratis.data.dir	/var/lib/hadoop-ozone/datanode/ratis/data	One or more directories used for storing Datanode Ratis metadata.	NVMe (Preferred)	/var/app/lib/hadoop-ozone/datanode/ratis	1024	RAID 1 NVMe or SSDNVMe/SSD is not required but strongly recommended.
hdds.datanode.dir	None	List of directories where user data is stored on each DataNode.	HDD	/data/{1..N}	Size and Number varies	<p>The mount points are numeric values for each data drive 1 - n</p> <p>These disks must not be shared with HDFS or another storage system.</p> <p>If not configured this falls back to dfs.datanode.data.dir. However it is strongly recommended to configure this explicitly.</p>

Ozone Manager

Setting name	Default value	Purpose	Disktype	Mount point	Minimum size (GB)	Notes
ozone.metadata.dirs	/var/lib/hadoop-ozone/om/ozone-metadata	One or more directories used for storing Ozone metadata. OzoneManager, SCM, and Datanode will write the metadata to this path.	NVMe (required)	/var/app/lib/hadoop-ozone/om/	4096	RAID 1 NVMe required
ozone.om.ratis.storage.dir	/var/lib/hadoop-ozone/om/ratis	This directory stores Ozone Manager's Ratis metadata like logs.	NVMe (required)	/var/app/lib/hadoop-ozone/om/		
ozone.om.db.dirs	/var/lib/hadoop-ozone/om/data	Directory where the Ozone Manager stores its metadata.	NVMe (required)	/var/app/lib/hadoop-ozone/om/		
ozone.om.snapshot.diff.db.dir		Directory for Ozone snapshot diff scratch space.	HDD or SSD	/data/1	256 (One HDD with 256GB free space will be sufficient to store snapshot diffs.)	<p>This is not available in Cloudera Manager. You must configure this through the safety valve.</p> <p>By default, if no value is specified here then the same locations as specified in ozone.metadata.dirs will be used.</p>

Recon

Setting name	Default value	Purpose	Disktype	Mount point	Minimum size (GB)	Notes
ozone.metadata.dirs	/var/lib/hadoop-ozone/recon/ozone-metadata	One or more directories used for storing Ozone metadata. OzoneManager, SCM, and Datanode will write the metadata to this path.	HDD	/var/app/lib/	4096	These 3 settings can share the same NVMe dedicated to Recon
ozone.recon.db.dir	/var/lib/hadoop-ozone/recon/data	Directory where the Recon Server stores its metadata.	NVMe (required)	/var/app/lib/hadoop-ozone/recon		
ozone.recon.om.db.dir	/var/lib/hadoop-ozone/recon/om/data	Where recon keeps OM snapshot DB.	NVMe (required)	/var/app/lib/hadoop-ozone/recon		
ozone.recon.scn.db.dir	/var/lib/hadoop-ozone/recon/scm/data	Directory where the Recon Server stores StorageContainerManager's metadata	NVMe (required)	/var/app/lib/hadoop-ozone/recon		

Storage Container Manager

Setting name	Default value	Purpose	Disktype	Mount point	Minimum size (GB)	Notes
ozone.metadata.dirs	/var/lib/hadoop-ozone/scm/ozone-metadata	One or more directories used for storing Ozone metadata. OzoneManager, SCM, and Datanode will write the metadata to this path.	NVMe or SSD	/var/app/lib/hadoop-ozone/scm	4096	These 3 settings can share the same NVMe dedicated to the SCM. RAID 1 NVMe or SSD required.
ozone.scm.db.dirs	/var/lib/hadoop-ozone/scm/data	Directory where the Storage Container Manager stores its metadata	NVMe or SSD	/var/app/lib/hadoop-ozone/scm		
ozone.scm.ha.ratis.storage.dir	/var/lib/hadoop-ozone/scm/ratis	Storage directory used by SCM to write Ratis logs.	NVMe or SSD	/var/app/lib/hadoop-ozone/scm		



Note: If NVMe is not available, then a SAS SSD can be used instead.

Performance tuning for Ozone

Learn how to use the Ozone configuration properties to tune Ozone to work optimally on your cluster.

For performance optimization and to overcome known issues, Cloudera recommends the following additional configurations:

Configuration Property	Value	Remarks
Ozone Configuration		
Maximum Process File Descriptors	100,000	
Java Heap Size of Ozone Manager	31GB	
Java Heap Size of Storage Container Manager	31GB	
Java Heap Size of Recon	31GB	Can be increased to 64GB for higher load.
Java Heap Size of S3 Gateway	31GB	
Java Heap Size of DataNode	31GB	
Ozone Service Advanced Configuration Snippet (Safety Valve) for ozone-conf/ozone-site.xml	<pre><property> <name>hdds.prometheus.endpoint.token</name> <value>*****</value> <description>Prometheus Token</description> </property></pre>	<p>Prometheus is an optional role and this configuration can be ignored if Prometheus is not being used.</p> <p>Disables Prometheus SPNEGO and uses Token Based Authentication.</p>
Ozone Configuration Ozone DataNode Advanced Configuration Snippet (Safety Valve) for ozone-conf/ozone-site.xml		
ozone.container.cache.size	8192	
ozone.container.cache.lock.stripes	8192	
hdds.datanode.du.factory.classname	org.apache.hadoop.hdds.fs.DedicatedDiskSpaceUsageFactory	

Configuration Property	Value	Remarks
dfs.container.ratis.leader.pending.bytes.limit	2GB	
Ozone Configuration Storage Container Manager Advanced Configuration Snippet (Safety Valve) for ozone-conf/ozone-site.xml		
ozone.scm.datanode.pipeline.limit	10	If DataNodes have different number of disks, pick the highest number.
ozone.scm.pipeline.owner.container.count	10	
ozone.scm.pipeline.creation.auto.factor.one	FALSE	
ozone.scm.container.placement.impl	org.apache.hadoop.hdds.scm.container.placement.algorithms.SCMContainerPlacementCapacity	EdgortileSSCMContainerPlacementCapacity on capacity available instead of random allocation.
Ozone Configuration Ozone Recon Advanced Configuration Snippet (Safety Valve) for ozone-conf/ozone-site.xml		
ozone.recon.task.pipelinesync.interval	120s	
ozone.recon.task.missingcontainer.interval	3600s	
Ozone Configuration Ozone Recon Advanced Configuration Snippet (Safety Valve) for ozone-conf/ozone-site.xml		
hdds.datanode.replication.work.dir	/tmp (is the default value)	The /tmp directory is used for staging transient files as part of data replication. Cloudera recommends a minimum of 50 GB size for the Datanode Replication Working Directory. Cloudera recommends you to use SSD for better performance during data replication. It can use the same SSD as datanode metadata.
hdds.datanode.replication.work.dir	</path/to/custom/replication/work/dir>	To use a non-default directory as Datanode Replication Working Directory
ozone.om.enable.filesystem.paths	true	
scm.container.client.max.size	<number of active pipelines in the cluster>	
scm.container.client.idle.threshold	120s	

Node maintenance

You must provide the value for the `hdds.scm.replication.maintenance.remaining.redundancy` parameter to put the cluster's nodes under maintenance.

The `hdds.scm.replication.maintenance.remaining.redundancy` parameter defines the number of redundant containers in a group which must be available for a node to enter maintenance.

If you are putting a node into maintenance mode and reduces the redundancy below the value defined, the node will remain in the ENTERING_MAINTENANCE state until a new replica is created.

Parameter	Default value	Description
hdds.scm.replication.maintenance.remaining.redundancy	Unset	The value is taken from the Ozone default configuration.
	1	The EC container has at least dataNum + 1 online allowing the loss of 1 more replica before data becomes unavailable.
hdds.scm.replication.maintenance.replica.minimum	1	This is for Ratis containers only.



Note: For EC containers, if nodes are under maintenance, reconstruction reads might be required if some of the data replicas are offline. This is seamless to the client but affects the read performance.