

Apache Hadoop

History Of Hadoop:

Hadoop is an open-source framework that has significantly shaped the world of big data and distributed computing. Here's a concise history of Hadoop:

Early Beginnings

- **2003:** The concept of Hadoop began to take shape when Doug Cutting and Mike Cafarella worked on the Apache Nutch project, which was an open-source web search engine. They were seeking a way to scale out the project's infrastructure.
- **2004:** Doug Cutting and his team at Yahoo! began developing Hadoop, influenced by Google's papers on MapReduce and the Google File System (GFS). These papers provided a blueprint for distributed computing and storage.

Key Milestones

- **2006:** Hadoop became a top-level project of the Apache Software Foundation. It included two main components:
 - **Hadoop Distributed File System (HDFS):** A distributed file system designed to run on commodity hardware.
 - **MapReduce:** A programming model for processing large data sets with a distributed algorithm.
- **2008:** Yahoo! made a major commitment to Hadoop, and the framework began gaining traction in the industry.
- **2009:** Hadoop gained more popularity with the release of version 0.20, which included significant improvements in scalability and performance.
- **2011:** Hadoop 1.0 was released, bringing enhancements like the addition of Hadoop Common and further stability improvements.
- **2012:** Cloudera, a company founded by former Google engineer Christophe Bisciglia and others, played a pivotal role in commercializing Hadoop and supporting its ecosystem. The Hadoop ecosystem began expanding with tools like Hive, Pig, and HBase.
- **2014:** Hadoop 2.0 was released, introducing YARN (Yet Another Resource Negotiator), which allowed Hadoop to run various processing frameworks beyond MapReduce. This version marked a significant architectural change, improving resource management and scalability.
- **2015-2016:** Hadoop became increasingly integrated with other big data tools and platforms. The ecosystem grew to include Apache Spark, Apache Flink, and other data processing and analytics tools.

Recent Developments

- **2018:** Apache Hadoop 3.0 was released, which brought new features such as support for erasure coding and improvements in scalability and resource management.
- **2020s:** The Hadoop ecosystem continued to evolve, incorporating new technologies and improving performance. The rise of cloud-based big data solutions and managed services, such as Amazon EMR, Google BigQuery, and Azure Synapse Analytics, has shifted some focus away from traditional Hadoop deployments.

Legacy and Impact

Hadoop has been a game-changer in the realm of big data. It enabled organizations to process and analyze large volumes of data more efficiently and cost-effectively than ever before. While newer technologies and platforms have emerged, Hadoop's influence remains substantial, particularly in how big data infrastructures are designed and implemented.

Versions of Hadoop:

Hadoop 0.x Series

- **0.1.0 (2006):** The initial release included basic implementations of MapReduce and HDFS.
- **0.17 (2008):** Added significant features and stability improvements, including the introduction of the JobTracker and TaskTracker components.

Hadoop 1.x Series

- **1.0.0 (2012):** Marked the release of Hadoop 1.x, featuring the YARN (Yet Another Resource Negotiator) architecture. This version decoupled MapReduce from resource management, enabling better resource allocation and scalability.

Hadoop 2.x Series

- **2.0.0 (2013):** Official release of Hadoop 2.x with YARN as the default resource manager, improving scalability and flexibility by allowing non-MapReduce applications to run on the Hadoop cluster.
- **2.7.x (2015-2016):** Introduced features such as improved HDFS high availability and enhancements to YARN for better resource management and performance.

Hadoop 3.x Series

- **3.0.0 (2017):** Major release that included support for erasure coding, which improves storage efficiency and fault tolerance. Enhanced YARN for better resource management and the addition of the Hadoop Distributed File System (HDFS) Federation.
- **3.1.x (2018):** Added support for dynamic resource pools and improved support for running multiple versions of YARN.
- **3.2.x (2019):** Included more stability improvements, additional features for performance, and further enhancements to YARN and HDFS.

Hadoop 3.3.x Series

- **3.3.0 (2020):** Focused on continuing improvements, including better scalability, security features, and performance optimizations.

Hadoop 4.x Series

- **4.0.x (2024):** Expected to bring further advancements in scalability, performance, and integration with modern data processing tools and cloud environments. (Note: As of the last update, Hadoop 4.x is anticipated and in development phases; exact features and release details may evolve.)

System Requirements for Hadoop:

Hadoop System Requirements for Windows:

1. Operating System:
 - Windows 10, Windows Server 2016/2019/2022 (Windows Subsystem for Linux or Docker may be used for a more compatible environment).
2. Hardware:
 - **CPU:** 2 GHz multi-core processor (minimum); more cores are recommended for better performance.
 - **RAM:** Minimum 8 GB of RAM (16 GB or more is recommended for larger datasets).
 - **Disk Space:** At least 100 GB of free disk space (more if handling large datasets).
3. Software:
 - **Java:** JDK 8 or JDK 11 (Hadoop 3.x and later versions support Java 11).
 - **SSH:** For a full Hadoop setup, you'll need SSH; however, on Windows, you can use tools like PuTTY or Cygwin for SSH capabilities.

- **Hadoop Distribution:** Download and configure a Hadoop distribution compatible with Windows (e.g., Hadoop binaries or using WSL/Docker).

Installation Steps with Commands:

- Install Java SDK and set the path in environment variables.
- Download Hadoop and set its path in environment variables.

C:\hadoop\bin

- Configure Hadoop core-site.xml file.

```
<configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>/Users/<YOUR_COMPUTER_NAME>/hdfs/tmp/</value>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://127.0.0.1:9000</value>
</property>
</configuration>
```

- Configure Hadoop hdfs-site.xml file.

```
<configuration>
<property>
<name>dfs.data.dir</name>
<value>/Users/<YOUR_COMPUTER_NAME>/hdfs/namenode</value>
</property>
<property>
<name>dfs.data.dir</name>
<value>/Users/<YOUR_COMPUTER_NAME>/hdfs/datanode</value>
</property>
<property><name>dfs.replication</name>
<value>1</value>
</property>
</configuration>
```

- Configure Hadoop mapred-site.xml

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
```

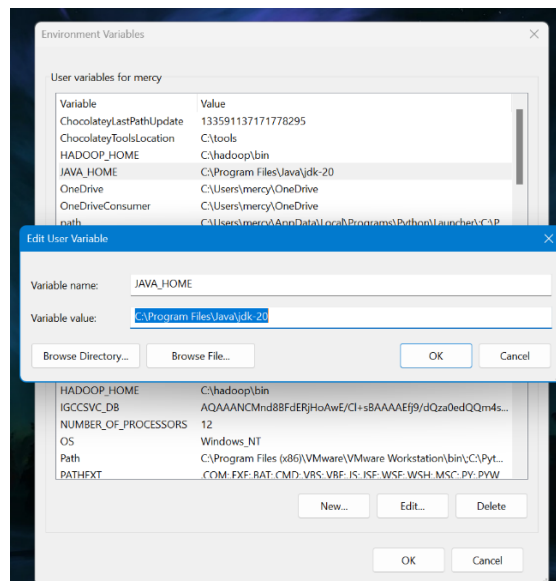
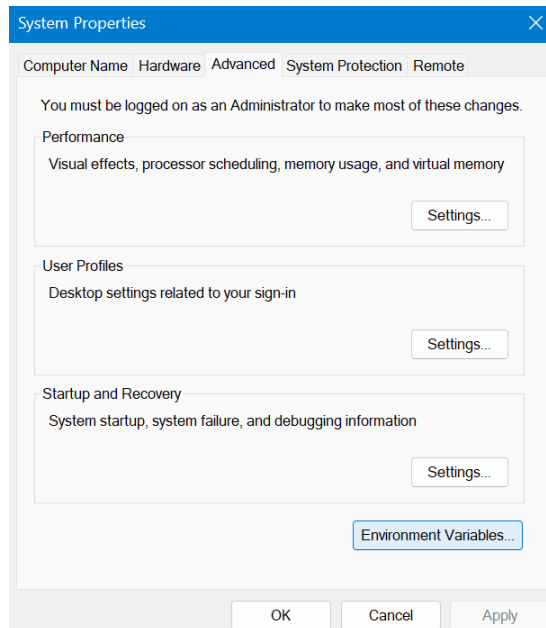
```
<value>yarn</value>
</property>
</configuration>
```

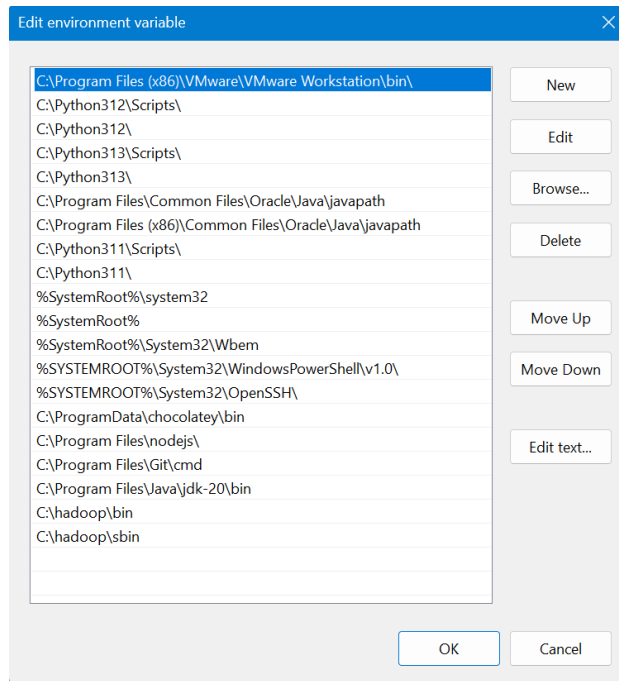
- Configure Hadoop yarn-site.xml

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-
services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
<name>yarn.resourcemanager.hostname</name>
<value>127.0.0.1</value>
</property>
<property>
<name>yarn.acl.enable</name>
<value>0</value>
</property><property>
<name>yarn.nodemanager.env-whitelist</name>
<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOO
P_
CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADO
OP
_MAPRED_HOME</value>
</property>
</configuration>
```

- Start Hadoop
start-all.sh

Installation Screenshots





```
core-site.xml X
C: > hadoop > etc > hadoop > core-site.xml
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20   <property>
21     <name>fs.defaultFS</name>
22     <value>hdfs://localhost:9000</value>
23   </property>
24 </configuration>
25
```

hdfs-site.xml X

C: > hadoop > etc > hadoop > hdfs-site.xml

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3  <!--
4  Licensed under the Apache License, Version 2.0 (the "License");
5  you may not use this file except in compliance with the License.
6  You may obtain a copy of the License at
7
8  http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20   <property>
21     <name>dfs.replication</name>
22     <value>1</value>
23   </property>
24   <property>
25     <name>dfs.namenode.name.dir</name>
26     <value>file:///C:/hadoop/data/namenode</value>
27   </property>
28   <property>
29     <name>dfs.datanode.data.dir</name>
30     <value>file:///C:/hadoop/data/datanode</value>
31   </property>
32 </configuration>
```

mapred-site.xml X

C: > hadoop > etc > hadoop > mapred-site.xml

```
1  <?xml version="1.0"?>
2  <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3  <!--
4  Licensed under the Apache License, Version 2.0 (the "License");
5  you may not use this file except in compliance with the License.
6  You may obtain a copy of the License at
7
8  http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20   <property>
21     <name>mapreduce.framework.name</name>
22     <value>yarn</value>
23   </property>
24 </configuration>
25
```



```
yarn-site.xml X
C: > hadoop > etc > hadoop > yarn-site.xml
1  <?xml version="1.0"?>
2  <!--
3      Licensed under the Apache License, Version 2.0 (the "License");
4      you may not use this file except in compliance with the License.
5      You may obtain a copy of the License at
6
7          http://www.apache.org/licenses/LICENSE-2.0
8
9      Unless required by applicable law or agreed to in writing, software
10     distributed under the License is distributed on an "AS IS" BASIS,
11     WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12     See the License for the specific language governing permissions and
13     limitations under the License. See accompanying LICENSE file.
14 -->
15 <configuration>
16   <property>
17     <name>yarn.nodemanager.aux-services</name>
18     <value>mapreduce_shuffle</value>
19   </property>
20 </configuration>
21
22
```

```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.22621.3958]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>hdfs namenode -format_
```

```

Administrator: Command Prompt
2024-08-03 13:09:16,167 INFO util.GSet: VM type = 64-bit
2024-08-03 13:09:16,167 INFO util.GSet: 1.0% max memory 1000 MB = 10 MB
2024-08-03 13:09:16,168 INFO util.GSet: capacity = 2^20 = 1048576 entries
2024-08-03 13:09:16,172 INFO namenode.FSDirectory: ACLs enabled? true
2024-08-03 13:09:16,172 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2024-08-03 13:09:16,172 INFO namenode.FSDirectory: XAttrs enabled? true
2024-08-03 13:09:16,176 INFO namenode.NameNode: Caching file names occurring more than 10 times
2024-08-03 13:09:16,182 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotFSLimit: 65536, maxSnapshotLimit: 65536
2024-08-03 13:09:16,182 INFO snapshot.SnapshotManager: dfs.namenode.snapshot.deletion.ordered = false
2024-08-03 13:09:16,186 INFO snapshot.SnapshotManager: SkipList is disabled
2024-08-03 13:09:16,190 INFO util.GSet: Computing capacity for map cachedBlocks
2024-08-03 13:09:16,190 INFO util.GSet: VM type = 64-bit
2024-08-03 13:09:16,193 INFO util.GSet: 0.25% max memory 1000 MB = 2.5 MB
2024-08-03 13:09:16,194 INFO util.GSet: capacity = 2^18 = 262144 entries
2024-08-03 13:09:16,203 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2024-08-03 13:09:16,203 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2024-08-03 13:09:16,204 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2024-08-03 13:09:16,211 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2024-08-03 13:09:16,212 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2024-08-03 13:09:16,214 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2024-08-03 13:09:16,214 INFO util.GSet: VM type = 64-bit
2024-08-03 13:09:16,215 INFO util.GSet: 0.029999999329447746% max memory 1000 MB = 307.2 KB
2024-08-03 13:09:16,216 INFO util.GSet: capacity = 2^15 = 32768 entries
Re-format filesystem in Storage Directory root= C:\hadoop\data\namenode; location= null ? (Y or N) Y
2024-08-03 13:09:19,917 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1330978382-192.168.1.5-1722670759882
2024-08-03 13:09:19,918 INFO common.Storage: Will remove files: [C:\hadoop\data\namenode\current\edits_inprogress_00000000000000000001, C:\hadoop\data\namenode\current\fsimage_00000000000000000000, C:\hadoop\data\namenode\current\fsimage_00000000000000000000.md5, C:\hadoop\data\namenode\current\seen_txid, C:\hadoop\data\namenode\current\VERSION]
2024-08-03 13:09:19,975 INFO common.Storage: Storage directory C:\hadoop\data\namenode has been successfully formatted.
2024-08-03 13:09:20,004 INFO namenode.FSImageFormatProtobuf: Saving image file C:\hadoop\data\namenode\current\fsimage.ckpt_00000000000000000000 using no compression
2024-08-03 13:09:20,106 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop\data\namenode\current\fsimage.ckpt_00000000000000000000 of size 400 bytes saved in 0 second
s
2024-08-03 13:09:20,117 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2024-08-03 13:09:20,123 INFO blockmanagement.DatanodeManager: Slow peers collection thread shutdown
2024-08-03 13:09:20,142 INFO namenode.FSNamesystem: Stopping services started for active state
2024-08-03 13:09:20,143 INFO namenode.FSNamesystem: Stopping services started for standby state
2024-08-03 13:09:20,149 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2024-08-03 13:09:20,150 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at Honor/192.168.1.5
*****/

```

```

Administrator: Command Prompt
2024-08-03 13:09:20,143 INFO namenode.FSNamesystem: Stopping services started for standby state
2024-08-03 13:09:20,149 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2024-08-03 13:09:20,150 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at Honor/192.168.1.5
*****/

```

C:\Windows\System32>start-dfs.cmd

```

Administrator: Command Prompt
2024-08-03 13:09:16,194 INFO util.GSet: capacity = 2^18 = 262144 entries
2024-08-03 13:09:16,203 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2024-08-03 13:09:16,203 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2024-08-03 13:09:16,204 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2024-08-03 13:09:16,211 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2024-08-03 13:09:16,212 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2024-08-03 13:09:16,214 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2024-08-03 13:09:16,214 INFO util.GSet: VM type = 64-bit
2024-08-03 13:09:16,215 INFO util.GSet: 0.029999999329447746% max memory 1000 MB = 307.2 KB
2024-08-03 13:09:16,216 INFO util.GSet: capacity = 2^15 = 32768 entries
Re-format filesystem in Storage Directory root= C:\hadoop\data\namenode; location= null ? (Y or N) Y
2024-08-03 13:09:19,917 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1330978382-192.168.1.5-1722670759882
2024-08-03 13:09:19,918 INFO common.Storage: Will remove files: [C:\hadoop\data\namenode\current\edits_inprogress_00000000000000000001, C:\hadoop\data\namenode\current\fsimage_00000000000000000000, C:\hadoop\data\namenode\current\fsimage_00000000000000000000.md5, C:\hadoop\data\namenode\current\seen_txid, C:\hadoop\data\namenode\current\VERSION]
2024-08-03 13:09:19,975 INFO common.Storage: Storage directory C:\hadoop\data\namenode has been successfully formatted.
2024-08-03 13:09:20,004 INFO namenode.FSImageFormatProtobuf: Saving image file C:\hadoop\data\namenode\current\fsimage.ckpt_00000000000000000000 using no compression
2024-08-03 13:09:20,106 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop\data\namenode\current\fsimage.ckpt_00000000000000000000 of size 400 bytes saved in 0 second
s
2024-08-03 13:09:20,117 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2024-08-03 13:09:20,123 INFO blockmanagement.DatanodeManager: Slow peers collection thread shutdown
2024-08-03 13:09:20,142 INFO namenode.FSNamesystem: Stopping services started for active state
2024-08-03 13:09:20,143 INFO namenode.FSNamesystem: Stopping services started for standby state
2024-08-03 13:09:20,149 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2024-08-03 13:09:20,150 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at Honor/192.168.1.5
*****/

Administrator: Apache Hadoop Distribution
2024-08-03 13:10:36,099 INFO datanode.DataNode: Initialization failed for Block pool <registering> (Datanode Uuid 677d10fa-ead0-4630-ba57-3db000662b20) service to localhost/127.0.0.1:9000. Exiting.
java.io.IOException: All specified directories have failed to load.
2024-08-03 13:10:36,099 INFO datanode.DataNode: Removed Block pool <registering> (Datanode Uuid 677d10fa-ead0-4630-ba57-3db000662b20)
2024-08-03 13:10:36,110 WARN datanode.DataNode: Exiting Datanode
2024-08-03 13:10:36,118 INFO datanode.DataNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down DataNode at Honor/192.168.1.5
*****/
C:\Windows\System32>

```

```
Administrator: Command Prompt
Administrator: Apache Hadoop Distribution
2024-08-03 13:11:18,801 INFO nodemanager.NodeManager: SHUTDOWN_MSG:
*****
SHUTDOWN_MSG: Shutting down NodeManager at Honor/192.168.1.5
C:\Windows\System32>
C:\Windows\System32>
2024-08-03 13:11:18,526 INFO resourcemanager.ResourceManager: SHUTDOWN_MSG:
*****
SHUTDOWN_MSG: Shutting down ResourceManager at Honor/192.168.1.5
*****
C:\Windows\System32>
```

localhost

HadoopOverviewDatanodesDatanode Volume FailuresSnapshotStartup ProgressUtilities

Overview 'localhost:9000' (active)

Started:	Sat Aug 03 13:10:35 +0530 2024
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaaec760
Compiled:	Mon Mar 04 12:05:00 +0530 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-b56a8ad0-6722-41a7-9b8e-f649114a138a
Block Pool ID:	BP-1330978382-192.168.1.5-1722670759882

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 60.33 MB of 79 MB Heap Memory. Max Heap Memory is 1000 MB.

Non Heap Memory used 50.52 MB of 53.13 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	0 B
Configured Remote Capacity:	0 B