Apache Hadoop

1.1 History of Hadoop

Hadoop has its origins in the early era of the World Wide Web. As the Web grew to millions and then billions of pages, the task of searching and returning search results became one of the most prominent challenges. Startups like Google, Yahoo, and AltaVista began building frameworks to automate search results. One project called Nutch was built by computer scientists Doug Cutting and Mike Cafarella based on Google's early work on MapReduce (more on that later) and Google File System. Nutch was eventually moved to the Apache open source software foundation and was split between Nutch and Hadoop. Yahoo, where Cutting began working in 2006, open sourced Hadoop in 2008.

While Hadoop is sometimes referred to as an acronym for High Availability Distributed Object Oriented Platform, it was originally named after Cutting's son's toy elephant.

1.2 Versions of Hadoop

Hadoop version 1 (Hadoop 1.x or Hadoop v1):

This was the initial release of Hadoop, also known as Apache Hadoop Core.

1. It consisted of two main components: Hadoop Distributed File System (HDFS) and MapReduce.

2. Hadoop 1.x had limitations such as a single point of failure (NameNode) and a lack of support for running multiple workloads simultaneously.

3. It was suitable for batch processing applications but had scalability and performance limitations for certain use cases.

Hadoop version 2 (Hadoop 2.x or Hadoop v2):

- Hadoop 2.x introduced significant architectural changes and improvements over Hadoop 1.x.

- The major enhancement was the introduction of YARN (Yet Another Resource Negotiator), which decoupled the resource management and job scheduling functionalities from MapReduce. This allowed Hadoop to support multiple processing frameworks beyond MapReduce.

- Hadoop 2.x addressed scalability and multi-tenancy issues by providing better resource utilization and support for running multiple applications simultaneously.

- With YARN, Hadoop 2.x became more versatile, supporting various processing models such as batch processing, interactive querying, real-time processing, etc.

- Hadoop 2.x also introduced various other improvements and features, including HDFS High Availability (HA) and federation, performance optimizations, and enhancements to MapReduce.

Hadoop version 3 (Hadoop 3.x or Hadoop v3):

- Hadoop 3.x builds upon the improvements made in Hadoop 2.x and introduces several new features and enhancements.

- One significant enhancement in Hadoop 3.x is support for erasure coding in HDFS, which provides more efficient data storage compared to traditional replication methods, reducing storage overhead.

- Hadoop 3.x also includes performance improvements, optimizations, and various updates to its components, including HDFS, YARN, and MapReduce.

- Another important addition in Hadoop 3.x is support for resource types in YARN, which allows users to define custom resources beyond CPU and memory, enabling better resource management for diverse workloads.

- Additionally, Hadoop 3.x continues to improve scalability, reliability, and security features compared to previous versions.

1.3 System Requirements for Hadoop

Hadoop System Requirements for macOS:

1. **Operating System:**
   - macOS (Catalina, Big Sur, Monterey, Ventura)
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:** Built-in SSH capabilities on macOS can be used.
   - **Homebrew:** Using Homebrew for installing dependencies and Hadoop itself can simplify setup.
4. **Additional:**
   - **File System:** Ensure sufficient file system space and performance for Hadoop operations.
   - **Network Configuration:** Proper network configuration and connectivity are required for distributed processing.

1.4 Installation Steps with Commands

- Install Java JDK and JAVA Home

```
/usr/libexec/java_home
echo $JAVA_HOME
```

- Enable SSH for localhost

```
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
chmod 0600 ~/.ssh/id_rsa.pub
ssh localhost
```

- Download Hadoop and modify the zprofile

  ```
  source ~/.zprofile
  ```
- Configure Hadoop

  ```
  sudo code $HADOOP_HOME/etc/hadoop/hadoop-env.sh
  /usr/libexec/java_home
  export
  JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_333.jdk
  /Contents/Home
  ```

- Edit core-site.xml

  ```
  <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>
  ```

- Edit hdfs-site.xml

  ```
  <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>
```

- Edit mapred-site.xml

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

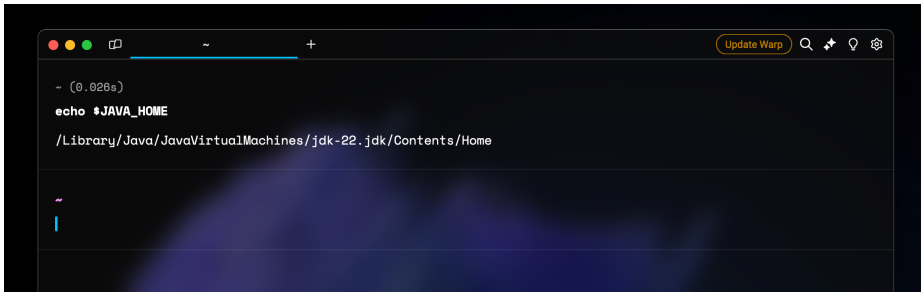
- Edit yarn-site.xml

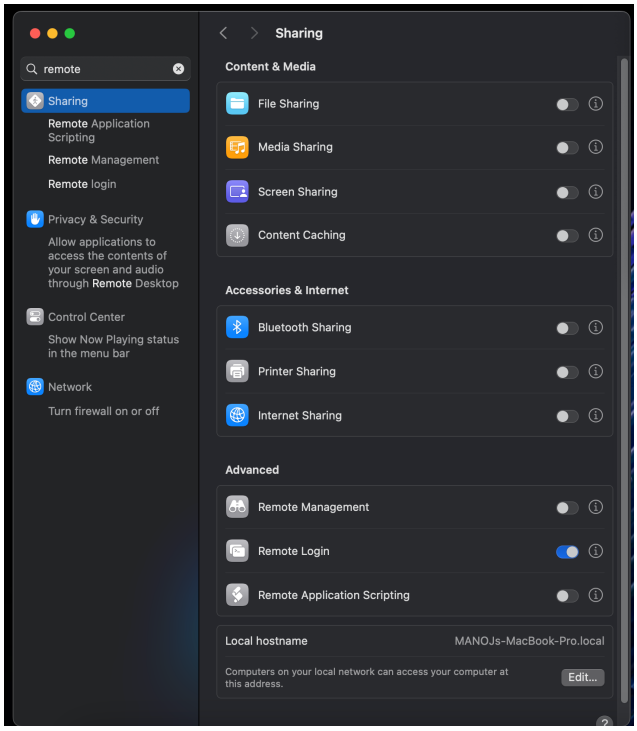
```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-
services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
<name>yarn.resourcemanager.hostname</name>
<value>127.0.0.1</value>
</property>
<property>
<name>yarn.acl.enable</name>
<value>0</value>
</property>
```

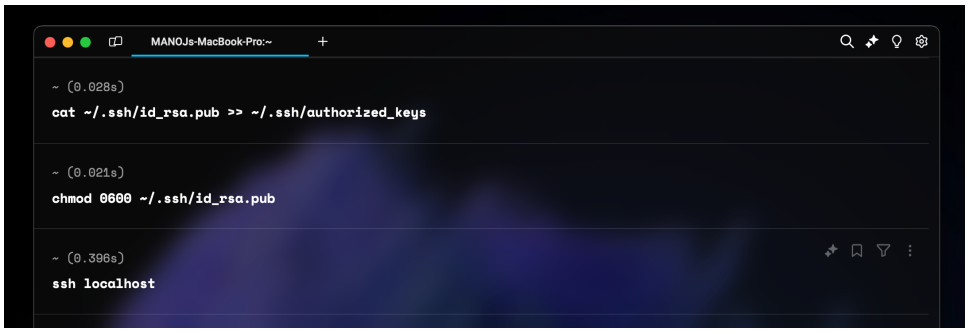
```
<property>
<name>yarn.nodemanager.env-whitelist</name>
<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_
CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP
_MAPRED_HOME</value>
</property>
</configuration>
```

- Start Hadoop
  `start-all.sh`

1.5 Installation Screenshots



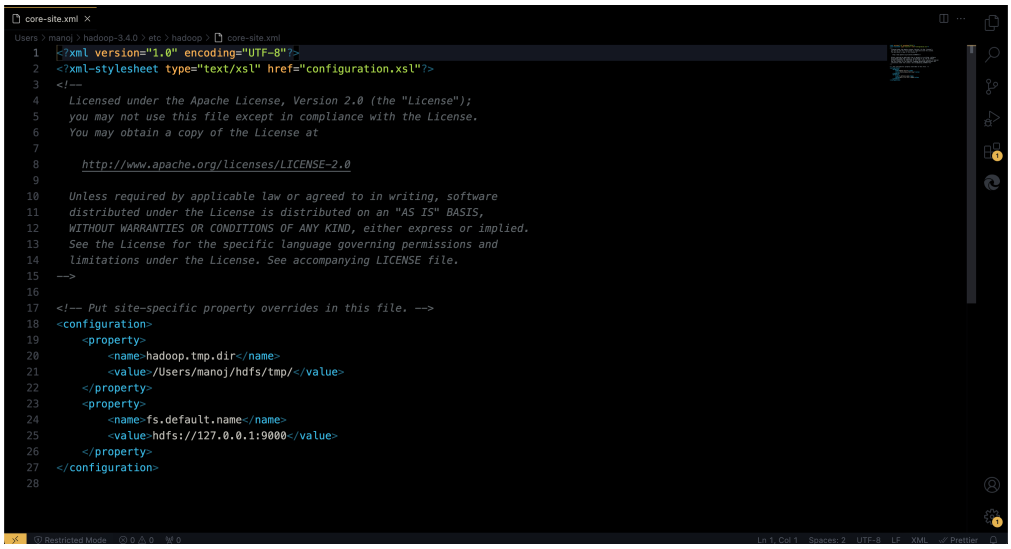


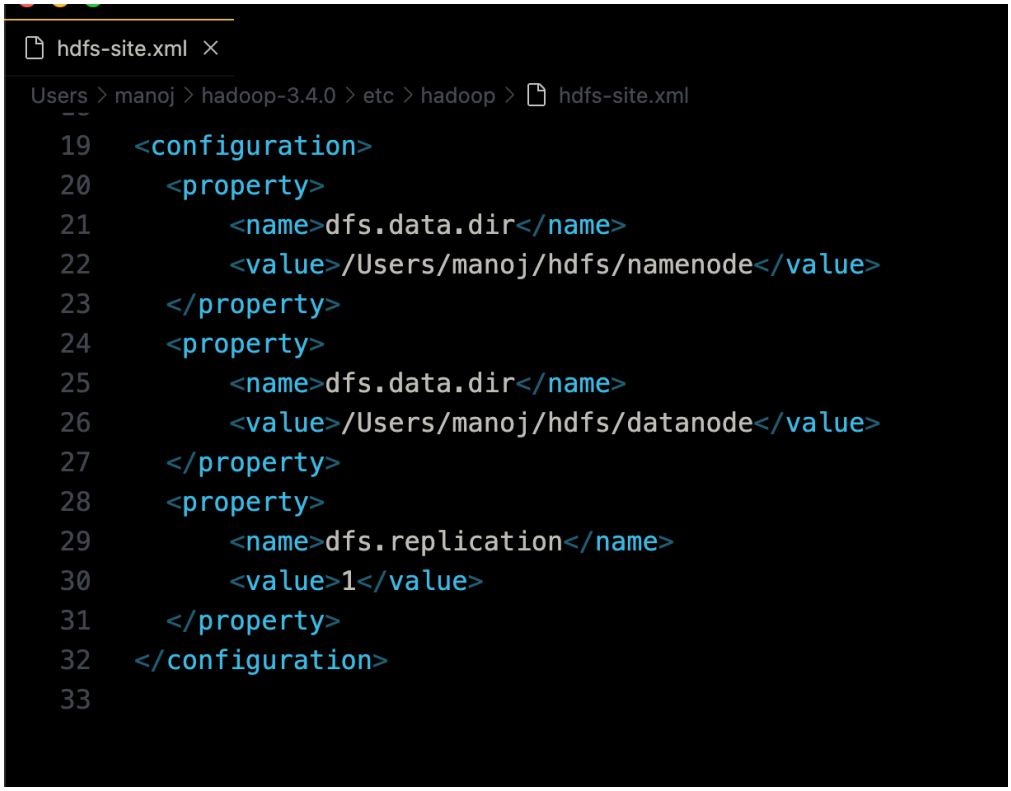


Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-512.

Version	Release date	Source download	Binary download	Release notes
3.4.0	2024 Mar 17	source (checksum signature)	binary (checksum signature) binary-aarch64 (checksum signature)	Announcement
3.3.6	2023 Jun 23	source (checksum signature)	binary (checksum signature) binary-aarch64 (checksum signature)	Announcement
2.10.2	2022 May 31	source (checksum signature)	binary (checksum signature)	Announcement


core-site.xml
Users > manoj > hadoop-3.4.0 > etc > hadoop > core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at
http://www.apache.org/licenses/LICENSE-2.0
Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>/Users/manoj/hdfs/tmp/</value>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://127.0.0.1:9000</value>
</property>
</configuration>
Restricted Mode
Ln 1, Col 1
Spaces: 2
UTF-8
LF
XML
Prettier


hdfs-site.xml
Users > manoj > hadoop-3.4.0 > etc > hadoop > hdfs-site.xml
<configuration>
<property>
<name>dfs.data.dir</name>
<value>/Users/manoj/hdfs/namenode</value>
</property>
<property>
<name>dfs.data.dir</name>
<value>/Users/manoj/hdfs/datanode</value>
</property>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
</configuration>
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33

mapred-site.xml

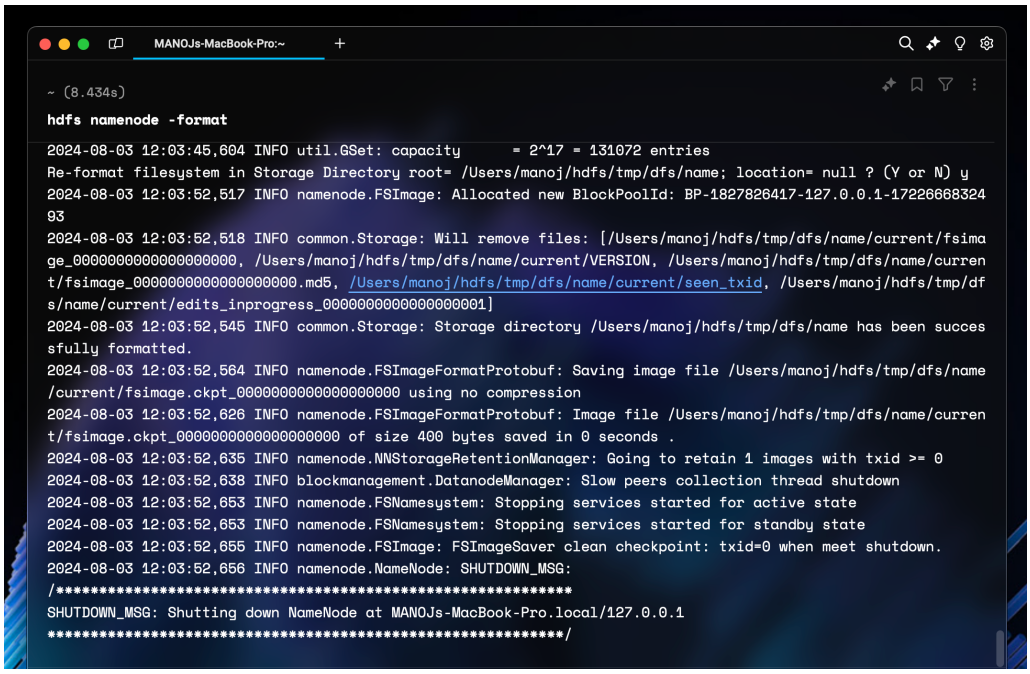
Users > manoj > hadoop-3.4.0 > etc > hadoop > mapred-site.xml

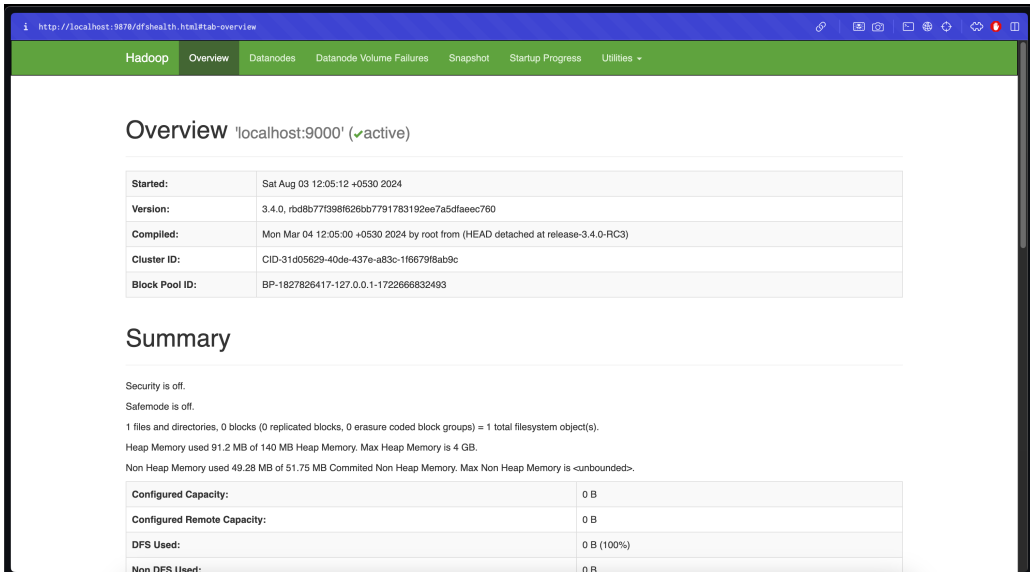
```
16
17  <!-- Put site-specific property overrides in this file. -->
18  <configuration>
19    <property>
20      <name>mapreduce.framework.name</name>
21      <value>yarn</value>
22    </property>
23  </configuration>
24
```

yarn-site.xml

Users > manoj > hadoop-3.4.0 > etc > hadoop > yarn-site.xml

```
15
16  <configuration>
17    <property>
18      <name>yarn.nodemanager.aux-services</name>
19      <value>mapreduce_shuffle</value>
20    </property>
21    <property>
22      <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
23      <value>org.apache.hadoop.mapred.ShuffleHandler</value>
24    </property>
25    <property>
26      <name>yarn.resourcemanager.hostname</name>
27      <value>127.0.0.1</value>
28    </property>
29    <property>
30      <name>yarn.acl.enable</name>
31      <value>0</value>
32    </property>
33    <property>
34      <name>yarn.nodemanager.env-whitelist</name>
35      <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRE
36    </property>
37  </configuration>
38
```


MANOJs-MacBook-Pro:~
~ (8.434s)
hdfs namenode -format
2024-08-03 12:03:45,604 INFO util.GSet: capacity = 2^17 = 131072 entries
Re-format filesystem in Storage Directory root= /Users/manoj/hdfs/tmp/dfs/name; location= null ? (Y or N) y
2024-08-03 12:03:52,517 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1827826417-127.0.0.1-1722666832493
2024-08-03 12:03:52,518 INFO common.Storage: Will remove files: [/Users/manoj/hdfs/tmp/dfs/name/current/fsimage_0000000000000000000, /Users/manoj/hdfs/tmp/dfs/name/current/VERSION, /Users/manoj/hdfs/tmp/dfs/name/current/fsimage_0000000000000000000.md5, /Users/manoj/hdfs/tmp/dfs/name/current/seen_txid, /Users/manoj/hdfs/tmp/dfs/name/current/edits_inprogress_0000000000000000001]
2024-08-03 12:03:52,545 INFO common.Storage: Storage directory /Users/manoj/hdfs/tmp/dfs/name has been successfully formatted.
2024-08-03 12:03:52,564 INFO namenode.FSImageFormatProtobuf: Saving image file /Users/manoj/hdfs/tmp/dfs/name/current/fsimage.ckpt_0000000000000000000 using no compression
2024-08-03 12:03:52,626 INFO namenode.FSImageFormatProtobuf: Image file /Users/manoj/hdfs/tmp/dfs/name/current/fsimage.ckpt_0000000000000000000 of size 400 bytes saved in 0 seconds .
2024-08-03 12:03:52,635 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2024-08-03 12:03:52,638 INFO blockmanagement.DatanodeManager: Slow peers collection thread shutdown
2024-08-03 12:03:52,653 INFO namenode.FSNamesystem: Stopping services started for active state
2024-08-03 12:03:52,653 INFO namenode.FSNamesystem: Stopping services started for standby state
2024-08-03 12:03:52,655 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2024-08-03 12:03:52,656 INFO namenode.NameNode: SHUTDOWN_MSG:
/**
SHUTDOWN_MSG: Shutting down NameNode at MANOJs-MacBook-Pro.local/127.0.0.1
**/


http://localhost:9870/dfshealth.html#tab-overview
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Overview 'localhost:9000' (active)
Started: Sat Aug 03 12:05:12 +0530 2024
Version: 3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled: Mon Mar 04 12:05:00 +0530 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID: CID-31d05629-40de-437e-a83c-1f6679f8ab9c
Block Pool ID: BP-1827826417-127.0.0.1-1722666832493
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 91.2 MB of 140 MB Heap Memory. Max Heap Memory is 4 GB.
Non Heap Memory used 49.28 MB of 51.75 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.
Configured Capacity: 0 B
Configured Remote Capacity: 0 B
DFS Used: 0 B (100%)