

# HADOOP

Hadoop is an open source framework overseen by Apache Software Foundation which is written in Java for storing and processing of huge datasets with the cluster of commodity hardware. There are mainly two problems with the big data. First one is to store such a huge amount of data and the second one is to process that stored data. The traditional approach like RDBMS is not sufficient due to the heterogeneity of the data. So Hadoop comes as the solution to the problem of big data i.e. storing and processing the big data with some extra capabilities.

## History of Hadoop-

Hadoop was started with Doug Cutting and Mike Cafarella in the year 2002 when they both started to work on Apache Nutch project. Hadoop originated in the early 2000s from the need to efficiently process massive amounts of data, inspired by Google's distributed file system (GFS) and MapReduce model. Created by Doug Cutting and Mike Cafarella as part of the Nutch project and later developed at Yahoo, Hadoop became an Apache Software Foundation top-level project in 2008. It quickly gained traction among major tech companies like Facebook, LinkedIn, and Twitter for its ability to handle large-scale data. Over time, Hadoop's ecosystem expanded with projects like Pig, Hive, and HBase, enhancing its capabilities and solidifying its role as a cornerstone of the big data movement, driving data-driven business models and influencing newer big data technologies.

## Versions of Hadoop-

Hadoop has undergone several versions since its inception, each introducing new features, improvements, and optimizations.

1. Hadoop 0.1.0 to 0.20.x - 2006 to 2009
2. Hadoop 0.21.x to 1.x - 2010 to 2012
3. Hadoop 2.x - 2012 to 2017
4. Hadoop 3.x - 2017 till present day

Each version of Hadoop has progressively enhanced its scalability, reliability, and flexibility, making it a robust solution for big data processing and analytics.

## System requirements for Hadoop all os-

Hadoop can run on various operating systems, but it is primarily designed for Unix-based systems like Linux and macOS. It is also possible to run Hadoop on Windows. Here are the general system requirements for running Hadoop on these operating systems:

- **Java:** Java 8 or higher (Java 11 recommended).
- **Memory:** Minimum of 8GB RAM
- **Storage:** At least 10GB of free disk space for basic operation
- **CPU:** Multi-core processors are recommended for better performance.
- **Network:** Reliable network connectivity for cluster operations.

## Hadoop Installation Steps for macOS -

To install hadoop in mac os , Open Terminal and follow the steps below.

- brew install hadoop

```
(base) madhumithak@madhumithas-MacBook-Air ~ % cd /opt/homebrew/Cellar/hadoop/3.4.0/libexec/etc/hadoop
(base) madhumithak@madhumithas-MacBook-Air hadoop % code hadoop-env.sh
(base) madhumithak@madhumithas-MacBook-Air hadoop % code core-site.xml
(base) madhumithak@madhumithas-MacBook-Air hadoop % code hdfs-site.xml
(base) madhumithak@madhumithas-MacBook-Air hadoop % code mapred-site.xml
(base) madhumithak@madhumithas-MacBook-Air hadoop % code yarn-site.xml
(base) madhumithak@madhumithas-MacBook-Air hadoop %
```

- hadoop-env.sh file

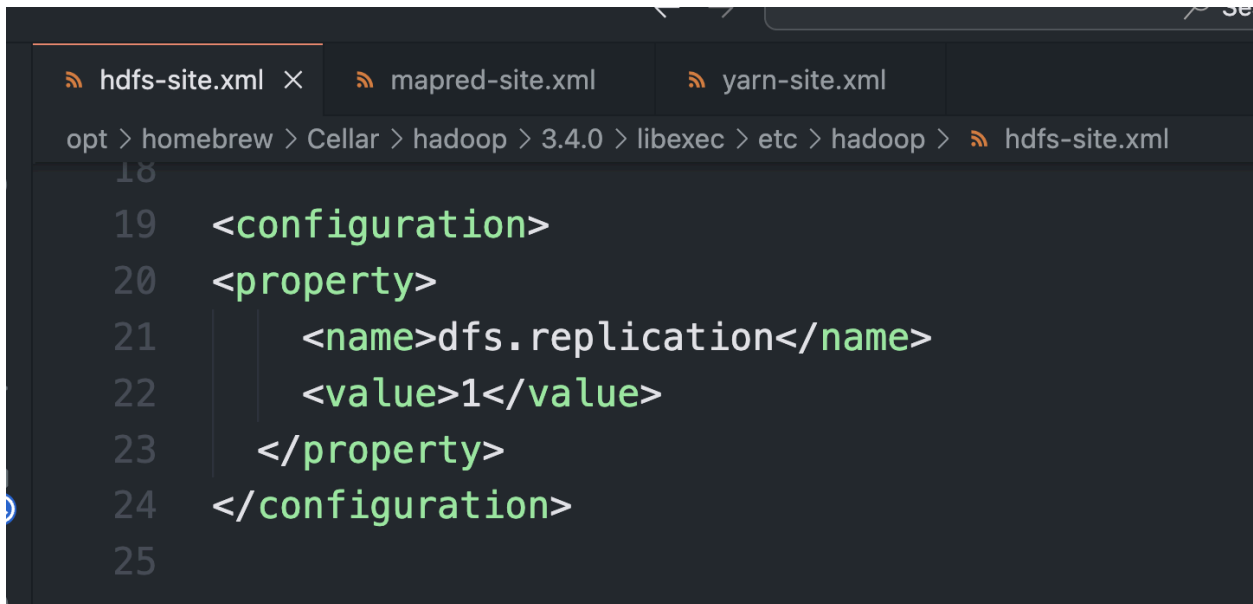
```
# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk-18.0.2.1.jdk/Contents/Home

# The language environment in which Hadoop runs. Use the English
# environment to ensure that logs are printed as expected.
export LANG=en_US.UTF-8
```

- Core-site.xml file

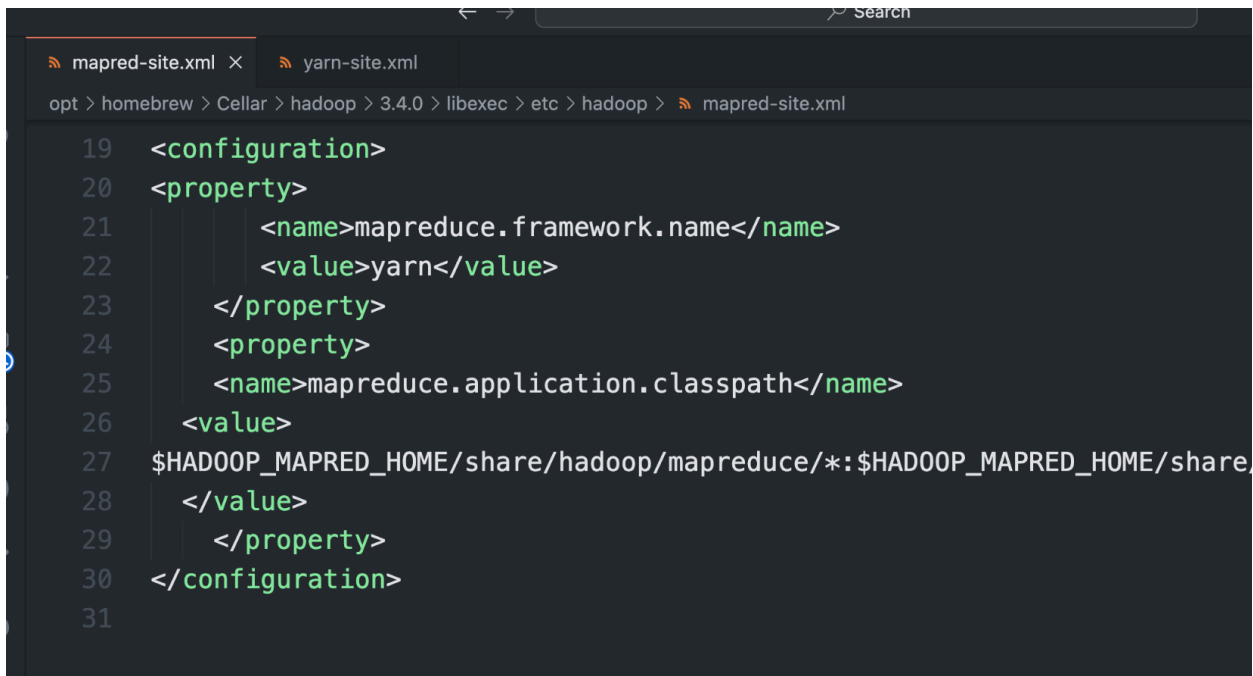
```
core-site.xml × hdfs-site.xml mapred-site.xml yarn-site.xml
opt > homebrew > Cellar > hadoop > 3.4.0 > libexec > etc > hadoop > core-site.xml
19 <configuration>
20 <property>
21   <name>fs.defaultFS</name>
22   <value>hdfs://localhost:9000</value>
23 </property>
24 </configuration>
25
```

- Hdfs-site.xml file



```
18
19 <configuration>
20 <property>
21   <name>dfs.replication</name>
22   <value>1</value>
23 </property>
24 </configuration>
25
```

- Mapred-site.xml file



```
19 <configuration>
20 <property>
21   <name>mapreduce.framework.name</name>
22   <value>yarn</value>
23 </property>
24 <property>
25   <name>mapreduce.application.classpath</name>
26   <value>
27 $HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share,
28   </value>
29 </property>
30 </configuration>
31
```

- Yarn-site.xml file

```
yarn-site.xml x
opt > homebrew > Cellar > hadoop > 3.4.0 > libexec > etc > hadoop > yarn-site.xml
14  -->
15  <configuration>
16
17  <!-- Site specific YARN configuration properties -->
18  <property>
19      <name>yarn.nodemanager.aux-services</name>
20      <value>mapreduce_shuffle</value>
21  </property>
22  <property>
23      <name>yarn.nodemanager.env-whitelist</name>
24      <value>
25  JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASS
26  </value>
27  </property>
28  </configuration>
29  |
```

- RSA key

```
(base) madhumithak@madhumithas-MacBook-Air hadoop % ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
/Users/madhumithak/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /Users/madhumithak/.ssh/id_rsa
Your public key has been saved in /Users/madhumithak/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:wUti4Zinq8j8wcXllwclvEYEWQt91TrvQIBWTjfD0Yo madhumithak@madhumithas-MacBook-Air.local
The key's randomart image is:
+---[RSA 3072]-----+
|      +oB=oB+      |
|      = Bo=+.oo    |
|      o B.B.+ o    |
|      .+o+ Eo=     |
|      .o .So..o    |
|      . . . . .    |
|      o.          o |
|o. ..            . |
|.oo.            |
+-----[SHA256]-----+
(base) madhumithak@madhumithas-MacBook-Air hadoop %
```

```
+---[SHA256]---+
(base) madhumithak@madhumithas-MacBook-Air ~ % cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
(base) madhumithak@madhumithas-MacBook-Air ~ % hadoop namenode -format
WARNING: Use of this script to execute namenode is deprecated.
WARNING: Attempting to execute replacement "hdfs namenode" instead.

2024-08-08 08:17:20,214 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = madhumithas-MacBook-Air.local/127.0.0.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.4.0
*****/
```

- Start the server

```
(base) madhumithak@madhumithas-MacBook-Air ~ % start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as madhumithak in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [madhumithas-MacBook-Air.local]
2024-08-08 08:18:15,499 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting resourcemanager
Starting nodemanagers
(base) madhumithak@madhumithas-MacBook-Air ~ %
```

- Open localhost

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities ▾

Overview

'localhost:9000' (✓active)

Started:	Thu Aug 08 08:18:08 +0530 2024
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaee760
Compiled:	Mon Mar 04 12:05:00 +0530 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-1273d0fe-e152-45ca-8a48-72e83f9d8eb3
Block Pool ID:	BP-2046741561-127.0.0.1-1723085241244

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 58.1 MB of 102 MB Heap Memory. Max Heap Memory is 2 GB.

Non Heap Memory used 50.63 MB of 53.38 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	228.27 GB
Configured Remote Capacity:	0 B

Non Heap Memory used 50.63 MB of 53.38 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	228.27 GB
Configured Remote Capacity:	0 B
DFS Used:	4 KB (0%)
Non DFS Used:	199.75 GB
DFS Remaining:	28.52 GB (12.49%)
Block Pool Used:	4 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Thu Aug 08 08:18:08 +0530 2024
Last Checkpoint Time	Thu Aug 08 08:17:21 +0530 2024
Last HA Transition Time	Never
Enabled Erasure Coding Policies	RS-6-3-1024k

## NameNode Journal Status

Current transaction ID: 1

Journal Manager	State
FileJournalManager(root=/tmp/hadoop-madhumithak/dfs/name)	EditLogFileOutputStream(/tmp/hadoop-madhumithak/dfs/name/current/edits_inprogress_00000000000000000001)

## NameNode Storage

Storage Directory	Type	State
/tmp/hadoop-madhumithak/dfs/name	IMAGE_AND_EDITS	Active

## DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	228.27 GB	4 KB (0%)	28.52 GB (12.49%)	4 KB	1

Hadoop, 2024.

- Stopping server

```
(base) madhumithak@madhumithas-MacBook-Air:~$ hadoop % stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as madhumithak in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [madhumithas-MacBook-Air.local]
2024-08-08 08:20:30,194 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Stopping nodemanagers
Stopping resourcemanager
(base) madhumithak@madhumithas-MacBook-Air:~$ hadoop %
```