DATA ANALYTICS
ASSIGNMENT 1
HADOOP

P.Lathika
210701131

INTRODUCTION:

Hadoop is an open-source software framework that is used for storing and processing large amounts of data in a distributed computing environment. It is designed to handle big data and is based on the MapReduce programming model, which allows for the parallel processing of large datasets.

HISTORY:

2003	Google released the paper, Google File System (GFS).
2004	Google released a white paper on Map Reduce.
2006	○ Hadoop introduced. ○ Hadoop 0.1.0 released. ○ Yahoo deploys 300 machines and within this year reaches 600 machines.
2007	○ Yahoo runs 2 clusters of 1000 machines. ○ Hadoop includes HBase.
2008	○ YARN JIRA opened ○ Hadoop becomes the fastest system to sort 1 terabyte of data on a 900 node cluster within 209 seconds. ○ Yahoo clusters loaded with 10 terabytes per day. ○ Cloudera was founded as a Hadoop distributor.
2009	○ Yahoo runs 17 clusters of 24,000 machines. ○ Hadoop becomes capable enough to sort a petabyte. ○ MapReduce and HDFS become separate subproject.
2010	○ Hadoop added the support for Kerberos. ○ Hadoop operates 4,000 nodes with 40 petabytes.

	○ Apache Hive and Pig released.
2011	○ Apache Zookeeper released. ○ Yahoo has 42,000 Hadoop nodes and hundreds of petabytes of storage.
2012	Apache Hadoop 1.0 version released.
2013	Apache Hadoop 2.2 version released.
2014	Apache Hadoop 2.6 version released.
2015	Apache Hadoop 2.7 version released.
2017	Apache Hadoop 3.0 version released.
2018	Apache Hadoop 3.1 version released.

Hardware Requirements

1. **Memory:**
   - At least 8 GB of RAM per machine (16 GB or more is recommended for production environments).
2. **Storage:**
   - At least 500 GB of disk space per machine.
   - Use high-speed disks (SSD) for better performance.
3. **CPU:**
   - Multi-core processors are recommended. At least 4 cores per machine.
4. **Network:**
   - High bandwidth (1 Gbps or higher) network connection between nodes.

Software Requirements

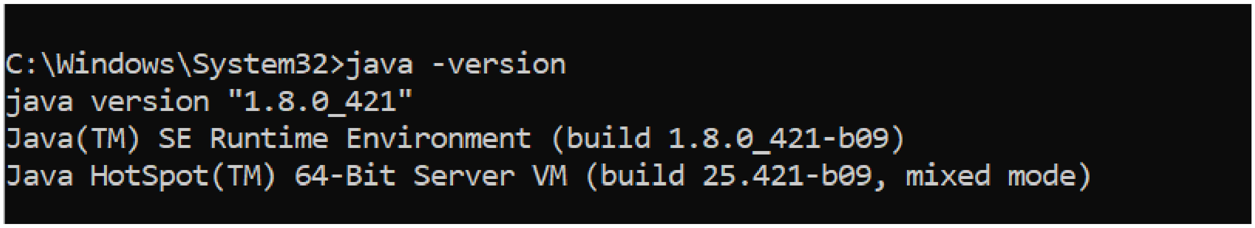
1. **Operating System:**
   - Linux-based OS (e.g., CentOS, Ubuntu, Debian).
   - Some versions of Hadoop support Windows, but Linux is preferred for production.

2. **Java:**
    - Oracle JDK 8 or OpenJDK 8 (Java 8).
    - Some versions of Hadoop may support Java 11, but it's crucial to verify compatibility.
3. **SSH:**
    - Password-less SSH (Secure Shell) setup for communication between nodes.
4. **Hadoop Distribution:**
    - Latest stable version of Hadoop. You can download it from the Apache Hadoop website.
5. **Additional Software:**
    - Python (optional, but recommended for certain Hadoop ecosystem tools).
    - Various Hadoop ecosystem components (e.g., HDFS, YARN, MapReduce, Hive, HBase, etc.) as required by your specific use case.

Configuration Considerations

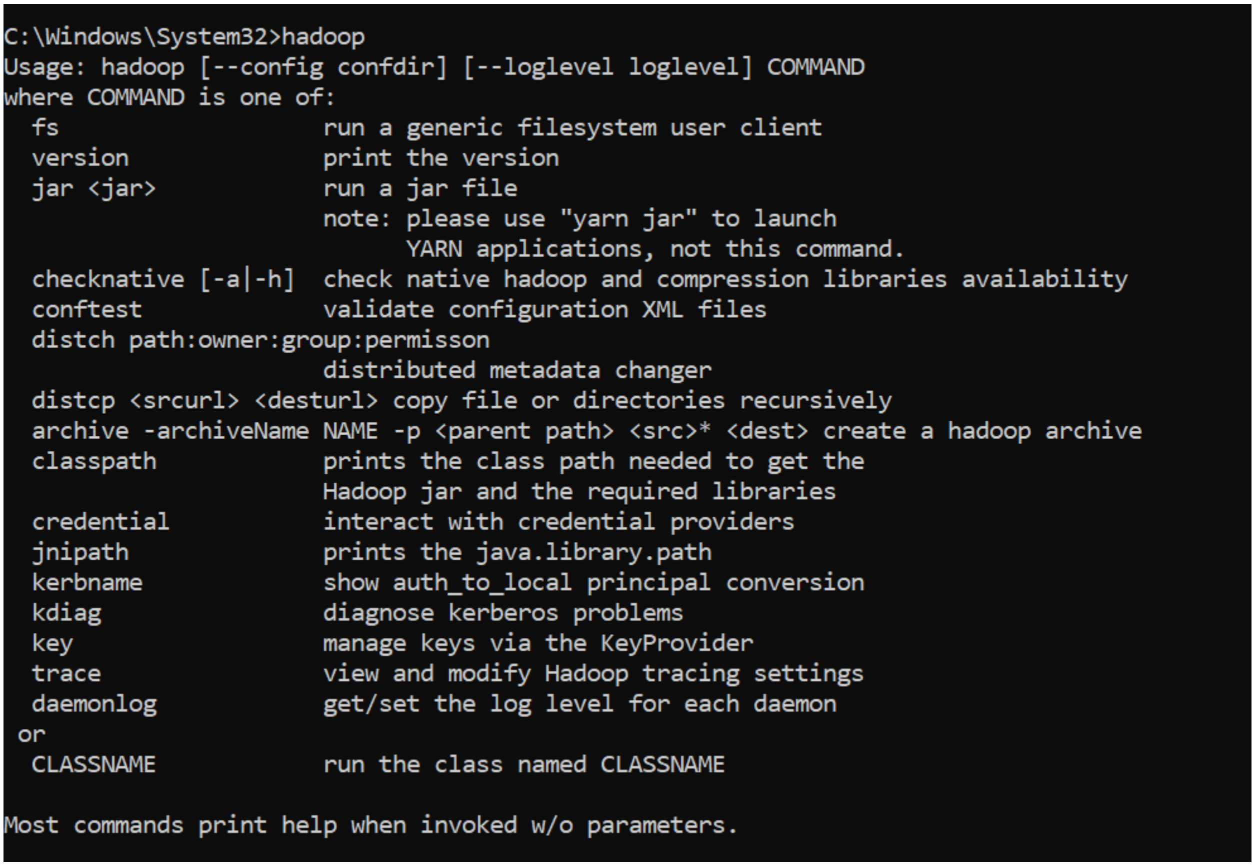
1. **Cluster Management:**
    - Use tools like Apache Ambari, Cloudera Manager, or other cluster management tools for easier setup and maintenance.
2. **Resource Management:**
    - Properly configure YARN for resource allocation.
    - Set appropriate heap sizes for NameNode and DataNode based on available memory.
3. **Replication Factor:**
    - Set the HDFS replication factor based on data redundancy needs (default is 3).
4. **Network Configuration:**
    - Ensure proper network configuration and DNS settings.
    - Optimize network settings for Hadoop traffic.

INSTALLATION STEPS



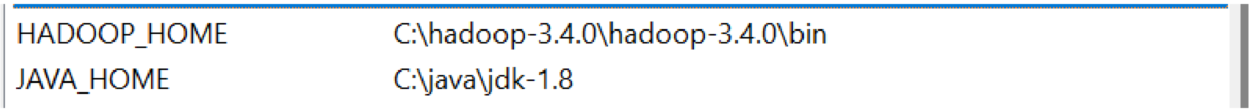
```
C:\Windows\System32>java -version
java version "1.8.0_421"
Java(TM) SE Runtime Environment (build 1.8.0_421-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.421-b09, mixed mode)
```

Check for Hadoop version

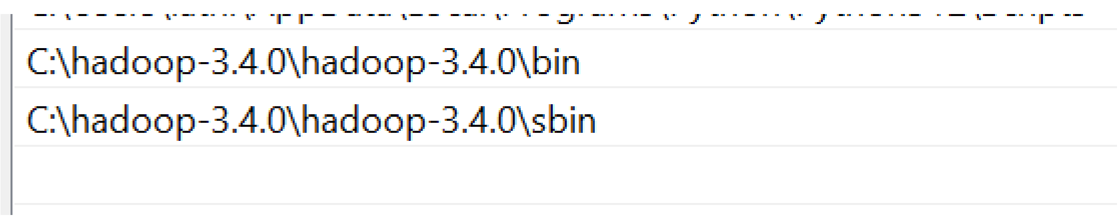


```
C:\Windows\System32>hadoop
Usage: hadoop [--config confdir] [--loglevel loglevel] COMMAND
where COMMAND is one of:
  fs                    run a generic filesystem user client
  version               print the version
  jar <jar>             run a jar file
                        note: please use "yarn jar" to launch
                              YARN applications, not this command.
  checknative [-a|-h]  check native hadoop and compression libraries availability
  conftest              validate configuration XML files
  distch path:owner:group:permisson
                        distributed metadata changer
  distcp <srcurl> <desturl> copy file or directories recursively
  archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
  classpath             prints the class path needed to get the
                        Hadoop jar and the required libraries
  credential            interact with credential providers
  jnipath               prints the java.library.path
  kerbname              show auth_to_local principal conversion
  kdiag                 diagnose kerberos problems
  key                   manage keys via the KeyProvider
  trace                 view and modify Hadoop tracing settings
  daemonlog             get/set the log level for each daemon
 or
  CLASSNAME             run the class named CLASSNAME

Most commands print help when invoked w/o parameters.
```

add path variables for java and Hadoop



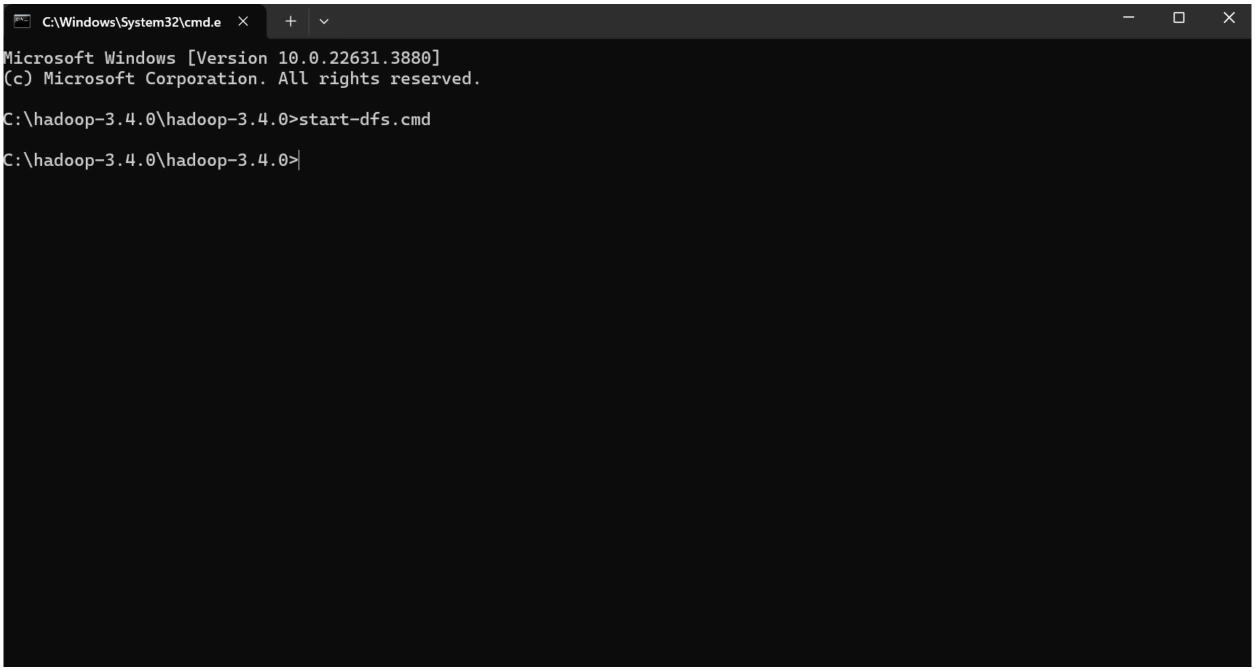
HADOOP_HOME	C:\hadoop-3.4.0\hadoop-3.4.0\bin
JAVA_HOME	C:\java\jdk-1.8



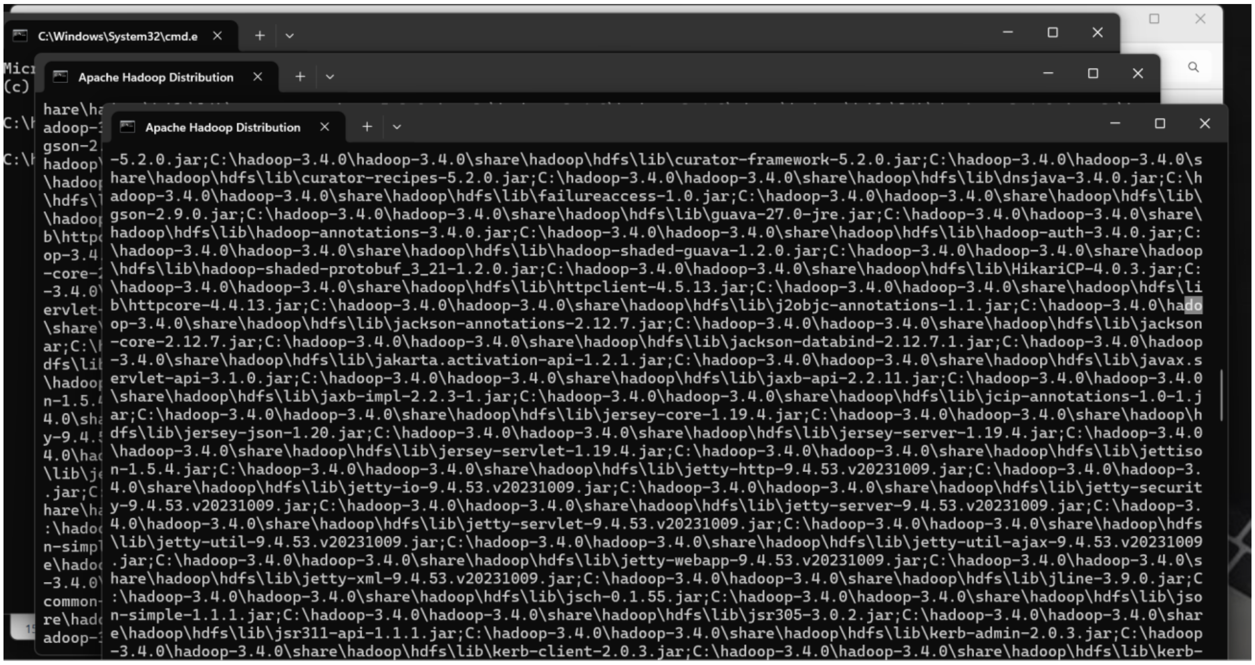
C:\hadoop-3.4.0\hadoop-3.4.0\bin

C:\hadoop-3.4.0\hadoop-3.4.0\sbin

Run start -dfs.cmd



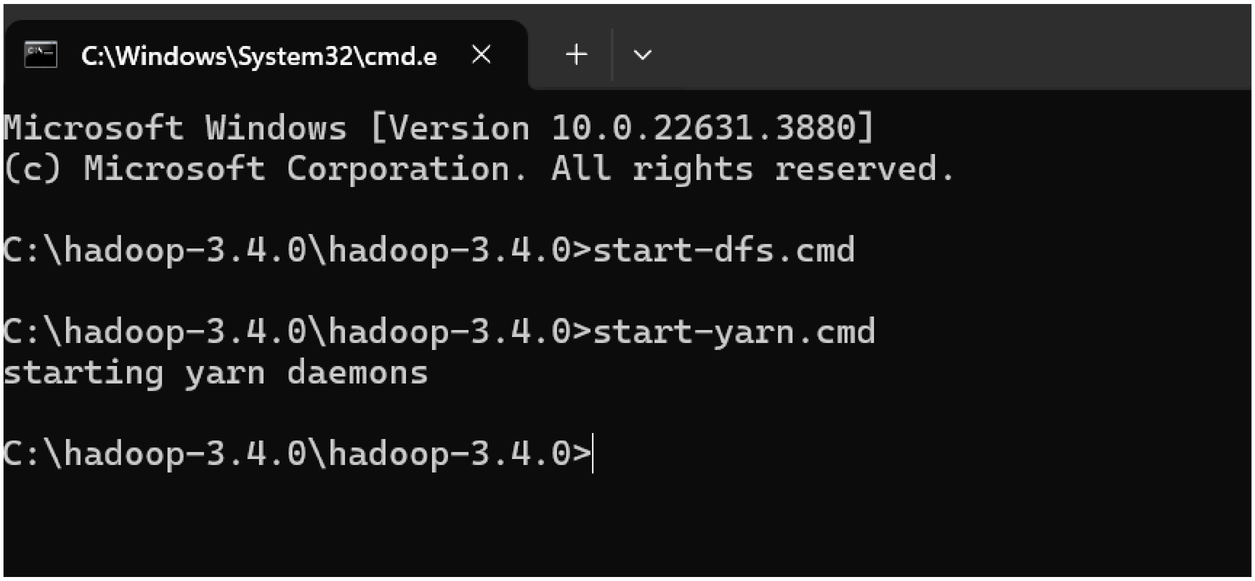
```
Microsoft Windows [Version 10.0.22631.3880]
(c) Microsoft Corporation. All rights reserved.

C:\hadoop-3.4.0\hadoop-3.4.0>start-dfs.cmd

C:\hadoop-3.4.0\hadoop-3.4.0>
```



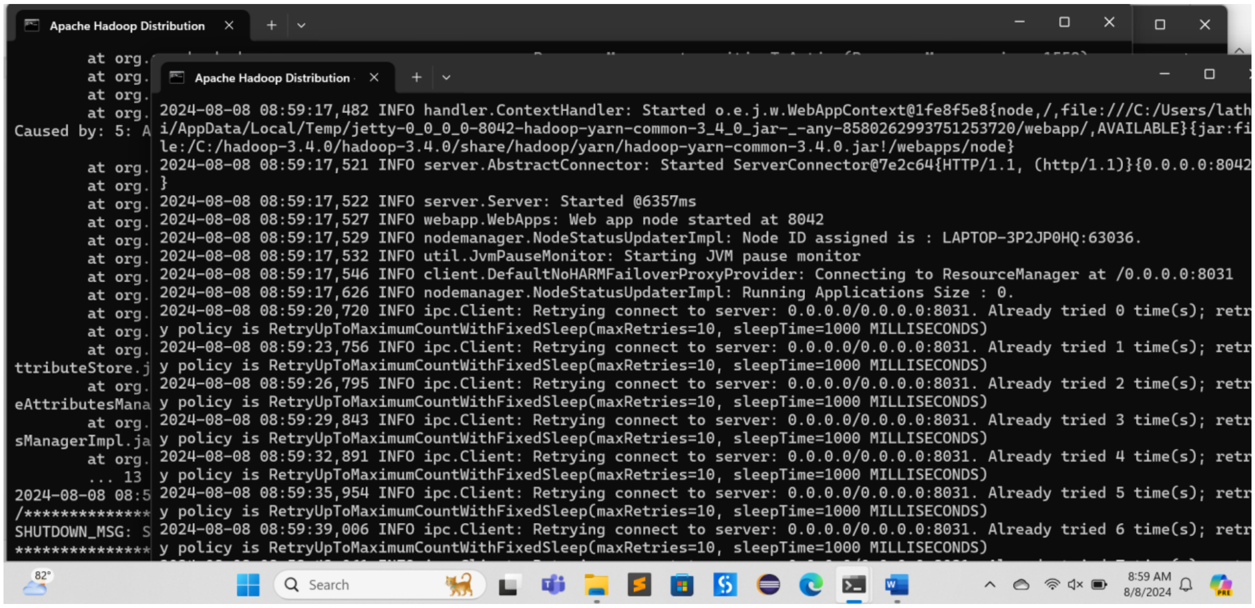
C:\Windows\System32\cmd.e
Apache Hadoop Distribution
Apache Hadoop Distribution

```
-5.2.0.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\curator-framework-5.2.0.jar;C:\hadoop-3.4.0\hadoop-3.4.0\s
hare\hadoop\hdfs\lib\curator-recipes-5.2.0.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\dnsjava-3.4.0.jar;C:\h
adoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\failureaccess-1.0.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\
gson-2.9.0.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\guava-27.0-jre.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\
hadoop\hdfs\lib\hadoop-annotations-3.4.0.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\hadoop-auth-3.4.0.jar;C:
\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\hadoop-shaded-guava-1.2.0.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop
\hdfs\lib\hadoop-shaded-protobuf_3_21-1.2.0.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\HikariCP-4.0.3.jar;C:
\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\httpclient-4.5.13.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\li
b\httpcore-4.4.13.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\j2objc-annotations-1.1.jar;C:\hadoop-3.4.0\hado
op-3.4.0\share\hadoop\hdfs\lib\jackson-annotations-2.12.7.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jackson
-core-2.12.7.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jackson-databind-2.12.7.1.jar;C:\hadoop-3.4.0\hadoop
-3.4.0\share\hadoop\hdfs\lib\jakarta.activation-api-1.2.1.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\javax.s
ervlet-api-3.1.0.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jaxb-api-2.2.11.jar;C:\hadoop-3.4.0\hadoop-3.4.0
\share\hadoop\hdfs\lib\jaxb-impl-2.2.3-1.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jcip-annotations-1.0-1.j
ar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jersey-core-1.19.4.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\h
dfs\lib\jersey-json-1.20.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jersey-server-1.19.4.jar;C:\hadoop-3.4.0
\hadoop-3.4.0\share\hadoop\hdfs\lib\jersey-servlet-1.19.4.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jettiso
n-1.5.4.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jetty-http-9.4.53.v20231009.jar;C:\hadoop-3.4.0\hadoop-3.
4.0\share\hadoop\hdfs\lib\jetty-io-9.4.53.v20231009.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jetty-securit
y-9.4.53.v20231009.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jetty-server-9.4.53.v20231009.jar;C:\hadoop-3.
4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jetty-servlet-9.4.53.v20231009.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs
\lib\jetty-util-9.4.53.v20231009.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jetty-util-ajax-9.4.53.v20231009
.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jetty-webapp-9.4.53.v20231009.jar;C:\hadoop-3.4.0\hadoop-3.4.0\s
hare\hadoop\hdfs\lib\jetty-xml-9.4.53.v20231009.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jline-3.9.0.jar;C
:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jsch-0.1.55.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jso
n-simple-1.1.1.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\jsr305-3.0.2.jar;C:\hadoop-3.4.0\hadoop-3.4.0\shar
e\hadoop\hdfs\lib\jsr311-api-1.1.1.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\kerb-admin-2.0.3.jar;C:\hadoop
-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\kerb-client-2.0.3.jar;C:\hadoop-3.4.0\hadoop-3.4.0\share\hadoop\hdfs\lib\kerb-
```

Run start-yarn.cmd

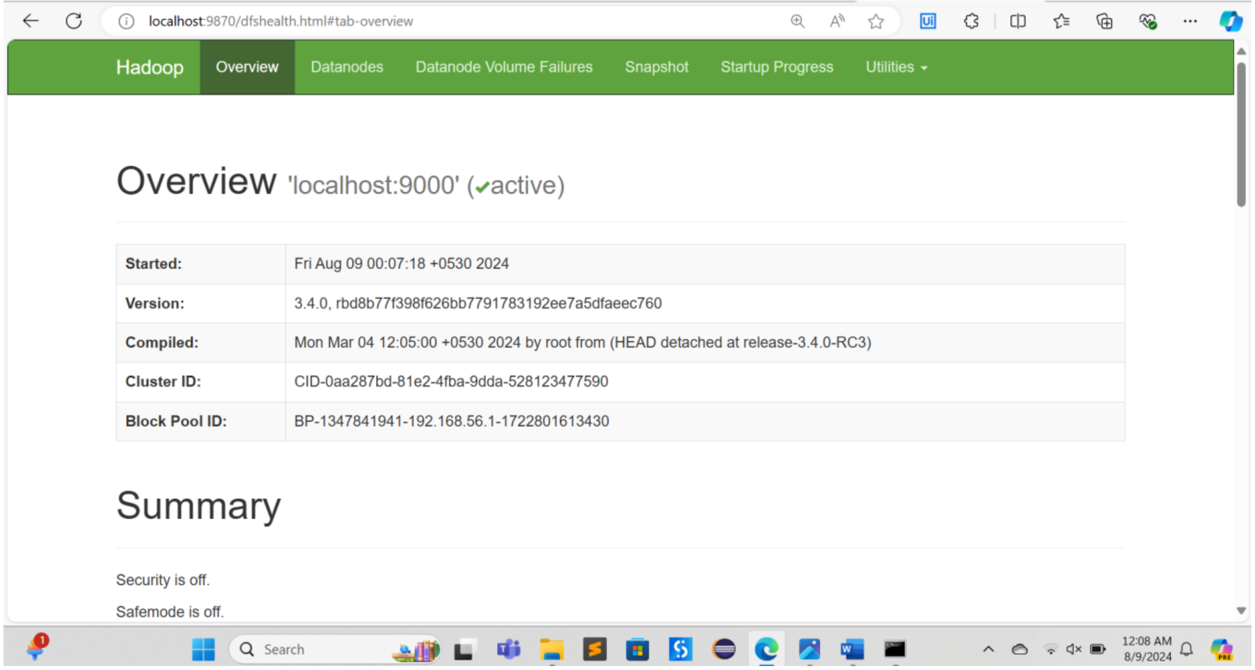


```
Microsoft Windows [Version 10.0.22631.3880]
(c) Microsoft Corporation. All rights reserved.

C:\hadoop-3.4.0\hadoop-3.4.0>start-dfs.cmd

C:\hadoop-3.4.0\hadoop-3.4.0>start-yarn.cmd
starting yarn daemons

C:\hadoop-3.4.0\hadoop-3.4.0>
```



```
        at org.
        at org.
        at org.
        at org. 2024-08-08 08:59:17,482 INFO handler.ContextHandler: Started o.e.j.w.WebAppContext@1fe8f5e8{node,/,file:///C:/Users/lath
Caused by: 5: A i/AppData/Local/Temp/jetty-0_0_0_0-8042-hadoop-yarn-common-3_4_0_jar-_-any-8580262993751253720/webapp/,AVAILABLE}{jar:fi
                le:/C:/hadoop-3.4.0/hadoop-3.4.0/share/hadoop/yarn/hadoop-yarn-common-3.4.0.jar!/webapps/node}
        at org. 2024-08-08 08:59:17,521 INFO server.AbstractConnector: Started ServerConnector@7e2c64{HTTP/1.1, (http/1.1)}{0.0.0.0:8042
        at org. }
        at org. 2024-08-08 08:59:17,522 INFO server.Server: Started @6357ms
        at org. 2024-08-08 08:59:17,527 INFO webapp.WebApps: Web app node started at 8042
        at org. 2024-08-08 08:59:17,529 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : LAPTOP-3P2JP0HQ:63036.
        at org. 2024-08-08 08:59:17,532 INFO util.JvmPauseMonitor: Starting JVM pause monitor
        at org. 2024-08-08 08:59:17,546 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8031
        at org. 2024-08-08 08:59:17,626 INFO nodemanager.NodeStatusUpdaterImpl: Running Applications Size : 0.
        at org. 2024-08-08 08:59:20,720 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:8031. Already tried 0 time(s); retr
        at org. y policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
        at org. 2024-08-08 08:59:23,756 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:8031. Already tried 1 time(s); retr
ttributeStore.j y policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
        at org. 2024-08-08 08:59:26,795 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:8031. Already tried 2 time(s); retr
eAttributesMana y policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
        at org. 2024-08-08 08:59:29,843 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:8031. Already tried 3 time(s); retr
sManagerImpl.ja y policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
        at org. 2024-08-08 08:59:32,891 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:8031. Already tried 4 time(s); retr
        ... 13  y policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-08-08 08:5 2024-08-08 08:59:35,954 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:8031. Already tried 5 time(s); retr
/************** y policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
SHUTDOWN_MSG: S 2024-08-08 08:59:39,006 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:8031. Already tried 6 time(s); retr
*************** y policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
```

Run in the localhost using localhost:9870



Overview 'localhost:9000' (✓active)

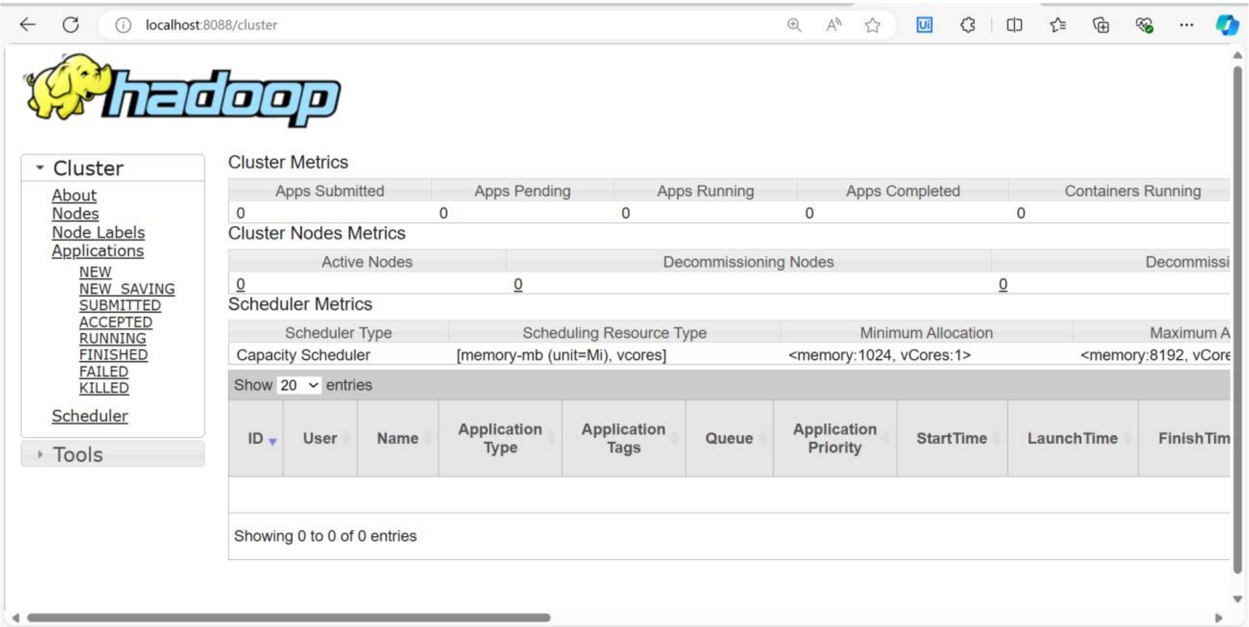
Started:	Fri Aug 09 00:07:18 +0530 2024
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled:	Mon Mar 04 12:05:00 +0530 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-0aa287bd-81e2-4fba-9dda-528123477590
Block Pool ID:	BP-1347841941-192.168.56.1-1722801613430

Summary

Security is off.

Safemode is off.

Run using localhost:8088



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
0	0	0	0	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissi
0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum A
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCore

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTim

Showing 0 to 0 of 0 entries