

CÀI ĐẶT HADOOP

1. Cập nhật và cài java 11:

- *sudo apt update*

```
ubuntu@ubuntu-bigdata:~$ sudo apt update
[sudo] password for ubuntu:
Hit:1 http://vn.archive.ubuntu.com/ubuntu jammy InRelease
Hit:2 http://vn.archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:3 http://vn.archive.ubuntu.com/ubuntu jammy-backports InRelease
Get:4 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Get:5 http://security.ubuntu.com/ubuntu jammy-security/main amd64 DEP-11 Metadata [43,2 kB]
Get:6 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 DEP-11 Metadata [40,1 kB]
Fetched 194 kB in 2s (89,1 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
425 packages can be upgraded. Run 'apt list --upgradable' to see them.
```

- *sudo apt install openjdk-11-jdk*

```
Running hooks in /etc/ca-certificates/update.d...
done.
done.
Processing triggers for sgml-base (1.30) ...
Setting up x11proto-dev (2021.5-1) ...
Setting up libxau-dev:amd64 (1:1.0.9-1build5) ...
Setting up libice-dev:amd64 (2:1.0.10-1build2) ...
Setting up libsm-dev:amd64 (2:1.2.3-1build2) ...
Setting up libxdmcp-dev:amd64 (1:1.1.3-0ubuntu5) ...
Setting up libxcb1-dev:amd64 (1.14-3ubuntu3) ...
Setting up libx11-dev:amd64 (2:1.7.5-1ubuntu0.2) ...
Setting up libxt-dev:amd64 (1:1.2.1-1) ...
```

2. Kiểm tra phiên bản đã cài: *java -version*

```
ubuntu@ubuntu-bigdata:~$ java -version
openjdk version "11.0.20" 2023-07-18
OpenJDK Runtime Environment (build 11.0.20+8-post-Ubuntu-1ubuntu122.04)
OpenJDK 64-Bit Server VM (build 11.0.20+8-post-Ubuntu-1ubuntu122.04, mixed mode, sharing)
```

3. Kiểm tra đường dẫn thư mục JAVA_HOME:

dirname \$(dirname \$(readlink -f \$(which java)))

```
ubuntu@ubuntu-bigdata:~$ dirname $(dirname $(readlink -f $(which java)))
/usr/lib/jvm/java-11-openjdk-amd64
```

4. Tạo người dùng Hadoop: *sudo adduser hadoop*

```
ubuntu@ubuntu-bigdata:~$ sudo adduser hadoop
Adding user 'hadoop' ...
Adding new group 'hadoop' (1001) ...
Adding new user 'hadoop' (1001) with group 'hadoop' ...
Creating home directory '/home/hadoop' ...
Copying files from '/etc/skel' ...
New password:
BAD PASSWORD: The password is shorter than 8 characters
Retype new password:
passwd: password updated successfully
Changing the user information for hadoop
Enter the new value, or press ENTER for the default
  Full Name []:
   Room Number []:
    Work Phone []:
    Home Phone []:
       Other []:
Is the information correct? [Y/n] y
```

- Chuyển tài khoản qua người dùng Hadoop để thao tác: `su - hadoop`
- Định cấu hình SSH (lưu ý: không tạo passphrase) cho người dùng Hadoop:

- Tạo cặp khóa SSH: `ssh-keygen -t rsa`

```
hadoop@ubuntu-bigdata:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:tyn1il553PlfR0lr0fdKt+44XsaPouudnxqfH4YbjCg hadoop@ubuntu-bigdata
The key's randomart image is:
+---[RSA 3072]-----+
|
|   .
|  S o o o
|   o.*+o* o
|  E..+.*oOo|
|   .+ +oBB=B|
|   .o.=oB%O=|
+---[SHA256]-----+
```

- Sao chép khóa chung đã tạo và đặt quyền:
`cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys`
`chmod 640 ~/.ssh/authorized_keys`

```
hadoop@ubuntu-bigdata:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hadoop@ubuntu-bigdata:~$ chmod 640 ~/.ssh/authorized_keys
```

- Truy cập SSH đến localhost: `ssh localhost`

```
hadoop@ubuntu-bigdata:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:tA3pQh44eZvHK3wCqQ84qmwhaIH6T/PbCXc0CGHqAI.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
Welcome to Ubuntu 22.04.1 LTS (GNU/Linux 6.2.0-26-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

425 updates can be applied immediately.
259 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.
```

**Nếu không thể ssh thì hãy kiểm tra ssh đã được install chưa?*

- Download Hadoop 3.3.4:
`wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz`
(quá trình xảy ra lâu, tùy thuộc kết nối mạng)

```
hadoop@ubuntu-bigdata:~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz
--2023-08-25 21:45:11-- https://dlcdn.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 695457782 (663M) [application/x-gzip]
Saving to: 'hadoop-3.3.4.tar.gz'

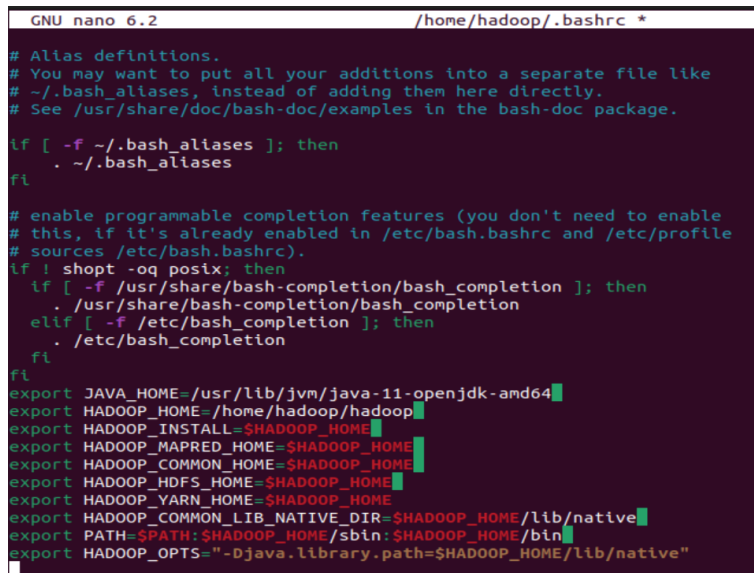
hadoop-3.3.4.tar.gz 100%[=====] 663,24M 2,39MB/s in 4m 39s

2023-08-25 21:50:24 (2,38 MB/s) - 'hadoop-3.3.4.tar.gz' saved [695457782/695457782]
```

- Giải nén file: `tar xzf hadoop-3.3.4.tar.gz`

10. Đổi tên file đã giải nén (nếu muốn): `mv hadoop-3.3.4 hadoop`
11. Cấu hình Hadoop và biến môi trường của Java, mở file: `nano ~/.bashrc`
`export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64`
`export HADOOP_HOME=/home/hadoop/hadoop`
`export HADOOP_INSTALL=$HADOOP_HOME`
`export HADOOP_MAPRED_HOME=$HADOOP_HOME`
`export HADOOP_COMMON_HOME=$HADOOP_HOME`
`export HADOOP_HDFS_HOME=$HADOOP_HOME`
`export HADOOP_YARN_HOME=$HADOOP_HOME`
`export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native`
`export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin`
`export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"`

**Thêm 10 dòng export vào cuối file!*



```
GNU nano 6.2 /home/hadoop/.bashrc *
# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

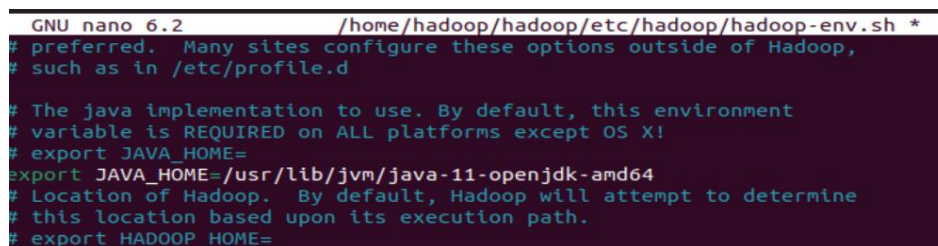
if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
    if [ -f /usr/share/bash-completion/bash_completion ]; then
        . /usr/share/bash-completion/bash_completion
    elif [ -f /etc/bash_completion ]; then
        . /etc/bash_completion
    fi
fi

export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

(màu xanh chuỗi vuông chỉ là khoảng trắng thừa, không ảnh hưởng)

12. Load lại cấu hình vừa thêm: `source ~/.bashrc`
13. Cấu hình cho biến môi trường của JAVA_HOME:
`nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh`
và tìm ngay dòng có `#export JAVA_HOME=` và cấu hình giá trị như đã tìm thấy ở bước 3.



```
GNU nano 6.2 /home/hadoop/hadoop/etc/hadoop/hadoop-env.sh *
# preferred. Many sites configure these options outside of Hadoop,
# such as in /etc/profile.d

# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
# export JAVA_HOME=
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
# Location of Hadoop. By default, Hadoop will attempt to determine
# this location based upon its execution path.
# export HADOOP_HOME=
```

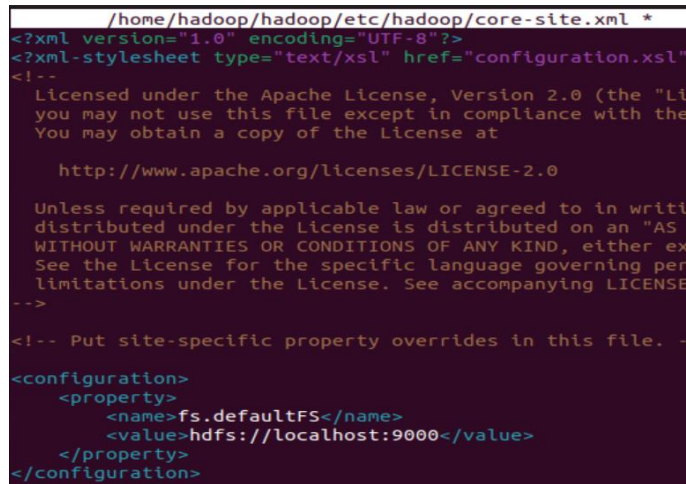
14. Tạo 2 thư mục trong thư mục home của người dùng Hadoop:

```
mkdir -p ~/hadoopdata/hdfs/{namenode,datanode}
```

15. Sửa tên file trong **core-site.xml**, bằng cách thêm các dòng lệnh sau:

```
nano $HADOOP_HOME/etc/hadoop/core-site.xml
```

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```



```
/home/hadoop/hadoop/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 for details.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

16. Sửa đường dẫn thư mục NameNode và DataNode trong **hdfs-site.xml**, bằng cách thêm các dòng lệnh sau:

```
nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
```

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.name.dir</name>
    <value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>
  </property>
  <property>
```



```

<name>dfs.data.dir</name>
<value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>

</property>

</configuration>

```

```

/home/hadoop/hadoop/etc/hadoop/hdfs-site.xml
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>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.name.dir</name>
    <value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>
  </property>
  <property>
    <name>dfs.data.dir</name>
    <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
  </property>
</configuration>

```

17. Sửa trong **mapred-site.xml**, bằng cách thêm các dòng lệnh sau:
[nano \\$HADOOP_HOME/etc/hadoop/mapred-site.xml](nano $HADOOP_HOME/etc/hadoop/mapred-site.xml)

```

<configuration>

  <property>

    <name>mapreduce.framework.name</name>

    <value>yarn</value>

  </property>

</configuration>

```

```

/home/hadoop/hadoop/etc/hadoop/mapred-site.xml
?xml version="1.0"?>
?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>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>

```

18. Sửa trong **yarn-site.xml**, bằng cách thêm các dòng lệnh sau:
nano \$HADOOP_HOME/etc/hadoop/yarn-site.xml

```
<configuration>

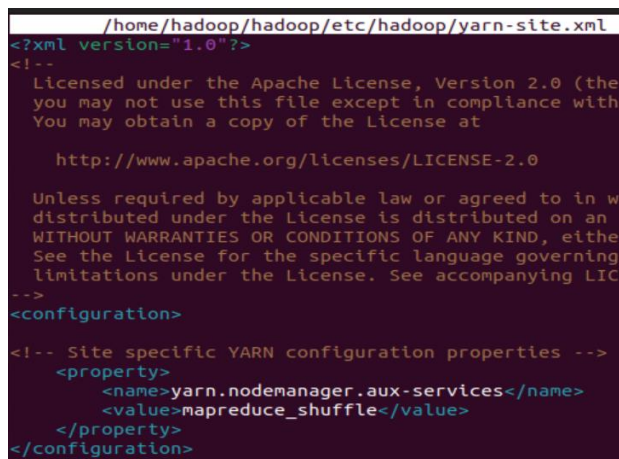
    <property>

        <name>yarn.nodemanager.aux-services</name>

        <value>mapreduce_shuffle</value>

    </property>

</configuration>
```

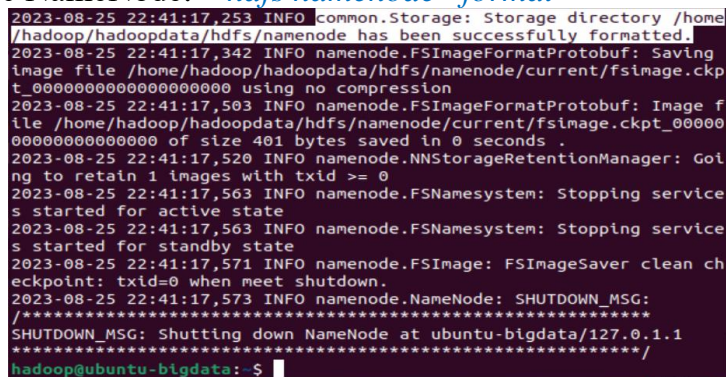


```
/home/hadoop/hadoop/etc/hadoop/yarn-site.xml
<?xml version="1.0"?>
<!--
 Licensed under the Apache License, Version 2.0 (the
 you may not use this file except in compliance with
 You may obtain a copy of the License at

 http://www.apache.org/licenses/LICENSE-2.0

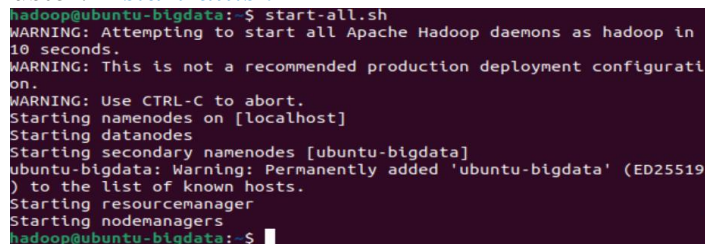
 Unless required by applicable law or agreed to in w
 distributed under the License is distributed on an
 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, eithe
 See the License for the specific language governing
 limitations under the License. See accompanying LIC
 -->
<configuration>
<!-- Site specific YARN configuration properties -->
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
</configuration>
```

19. Format thư mục NameNode: *hdfs namenode -format*



```
2023-08-25 22:41:17,253 INFO common.Storage: Storage directory /home
/hadoop/hadoopdata/hdfs/namenode has been successfully formatted.
2023-08-25 22:41:17,342 INFO namenode.FSImageFormatProtobuf: Saving
image file /home/hadoop/hadoopdata/hdfs/namenode/current/fsimage.ckpt
t_000000000000000000 using no compression
2023-08-25 22:41:17,503 INFO namenode.FSImageFormatProtobuf: Image f
ile /home/hadoop/hadoopdata/hdfs/namenode/current/fsimage.ckpt_00000
0000000000000 of size 401 bytes saved in 0 seconds .
2023-08-25 22:41:17,520 INFO namenode.NNStorageRetentionManager: Goi
ng to retain 1 images with txid >= 0
2023-08-25 22:41:17,563 INFO namenode.FSNamesystem: Stopping service
s started for active state
2023-08-25 22:41:17,563 INFO namenode.FSNamesystem: Stopping service
s started for standby state
2023-08-25 22:41:17,571 INFO namenode.FSImage: FSImageSaver clean ch
eckpoint: txid=0 when meet shutdown.
2023-08-25 22:41:17,573 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at ubuntu-bigdata/127.0.1.1
*****/
hadoop@ubuntu-bigdata:~$
```

20. Start Hadoop cluster: *start-all.sh*



```
hadoop@ubuntu-bigdata:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in
10 seconds.
WARNING: This is not a recommended production deployment configurati
on.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu-bigdata]
ubuntu-bigdata: Warning: Permanently added 'ubuntu-bigdata' (ED25519
) to the list of known hosts.
Starting resource manager
Starting nodemanagers
hadoop@ubuntu-bigdata:~$
```

(nếu đặt passphrase ở bước 6, thì không thể start Hadoop cluster như hình!)

21. Truy cập Hadoop từ trình duyệt qua đường dẫn <http://localhost:9870> or <http://localhost:8088>

Course: Dữ liệu lớn (GV: ...)

Buoi1 - Google Drive

Namenode information

+

localhost:9870/dfshealth.html#tab-overview

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Overview 'localhost:9000' (active)

Started:	Fri Aug 25 22:45:29 +0700 2023
Version:	3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb
Compiled:	Fri Jul 29 19:32:00 +0700 2022 by stevel from branch-3.3.4
Cluster ID:	CID-7e89c981-8c5a-4353-9b9b-2983f9ac256a
Block Pool ID:	BP-361549562-127.0.1.1-1692978077164

Summary


Course: Dữ liệu lớn (GV: ...)

Buoi1 - Google Drive

All Applications

+

localhost:8088/cluster



Cluster

About

Nodes

Node Labels

Applications

NEW

SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Cluster Metrics

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

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decon
0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Mi
Capacity Scheduler	[memory-mb (unit=Mb), vcores]	<memory:1024, vCore

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	Finis
Showing 0 to 0 of 0 entries									