Hadoop/Spark Introduction and Installation

Web Information Retrieval Lab

Office: R1424, Technology Building

Mark Hong

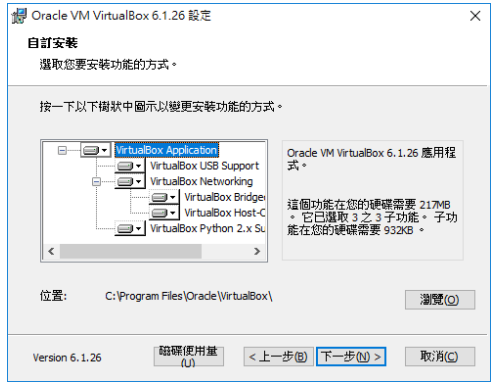
Outline

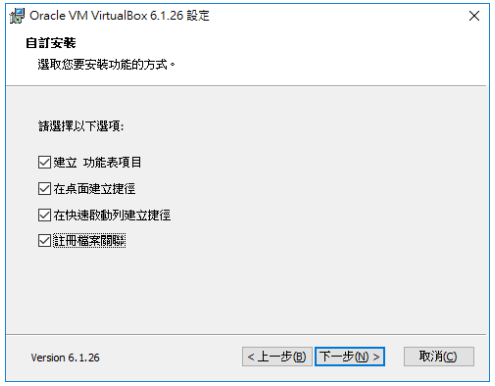
- VirtualBox installation
- Ubuntu introduction
- Linux basic commands
- System configuration
- Hadoop installation
- Spark installation

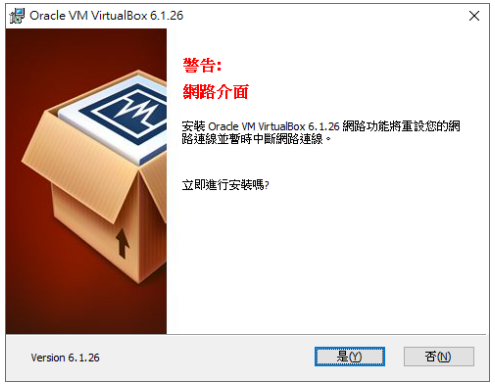
VirtualBox introduction

- Home page: https://www.virtualbox.org/
- Download page: https://www.virtualbox.org/wiki/Downloads

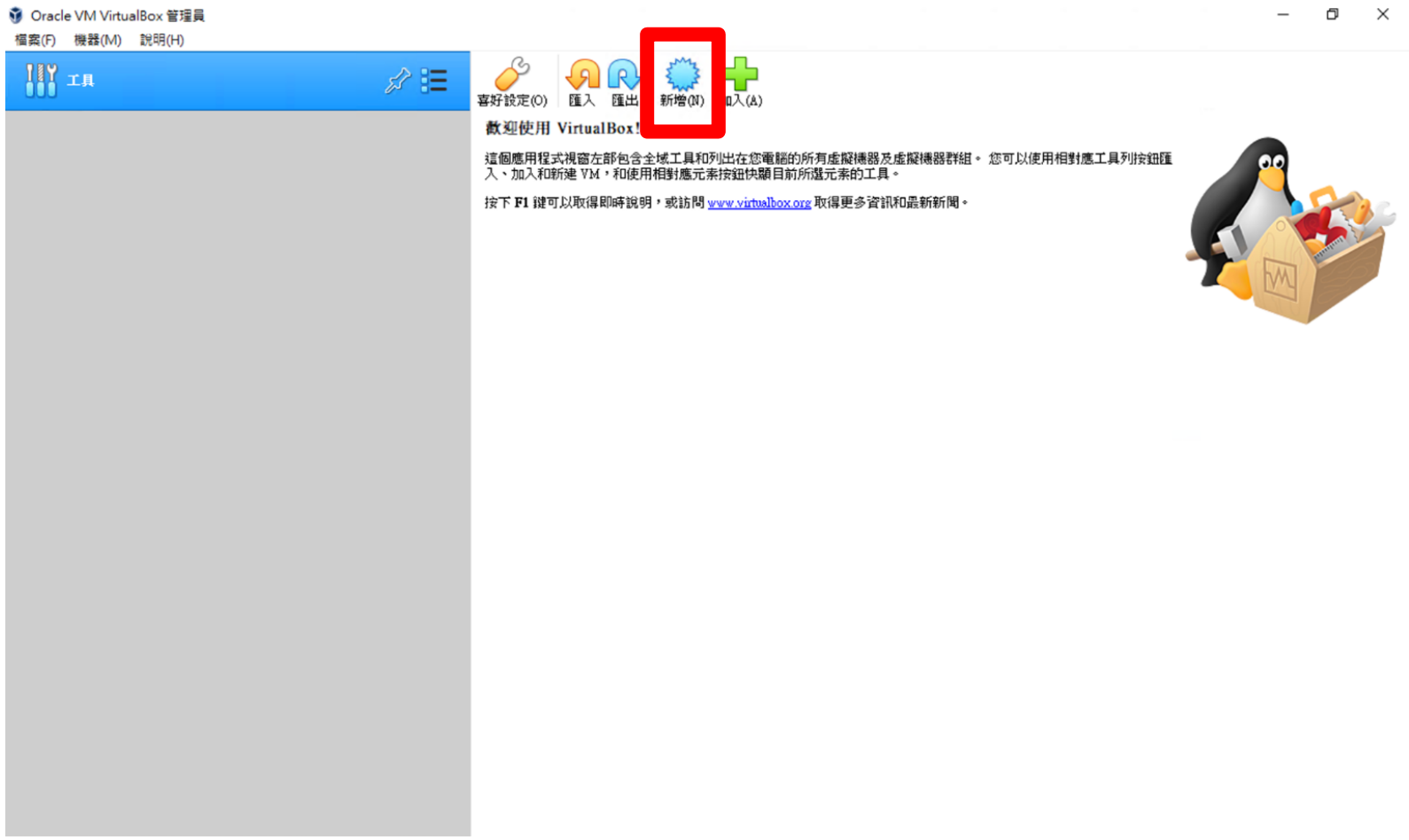
VirtualBox installation steps







Create new virtual machine



Virtual machine specification

Specification	
OS version	Ubuntu 20.04 LTS Desktop
Amount	3
Memory	4 Gb
Disk	50 Gb
User name	bdm

IP	
spark-1	192.168.88.171
spark-2	192.168.88.172
spark-3	192.168.88.173

Ubuntu introduction

- It‘s one of LINUX distributions
- Desktop / Server version
- Newest version is 20.10
- Use 20.04 LTS Desktop to homework
- Download: Click Me

Linux basic commands

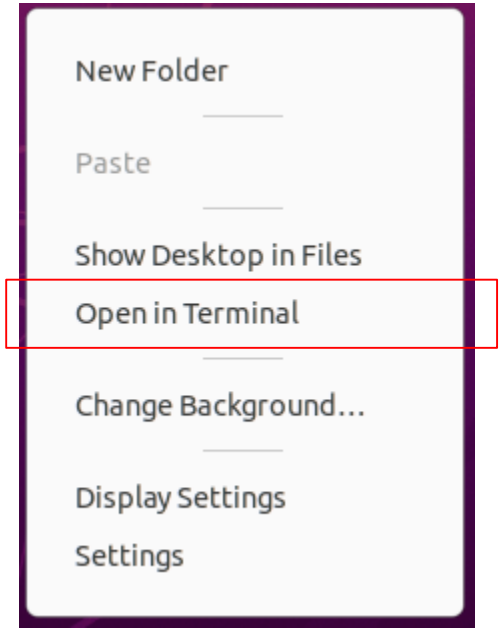
- cd [path]: change directory
- ls [-a –l etc.]: show directory content
- rm [-r –f etc.] [path or filename]: delete directory or file
- mv [source path] [destination path] : move directory or file (rename)
- tar [-zxcvf] [source path] [destination path] : compress or extract file
- man [command] : command’s documentation

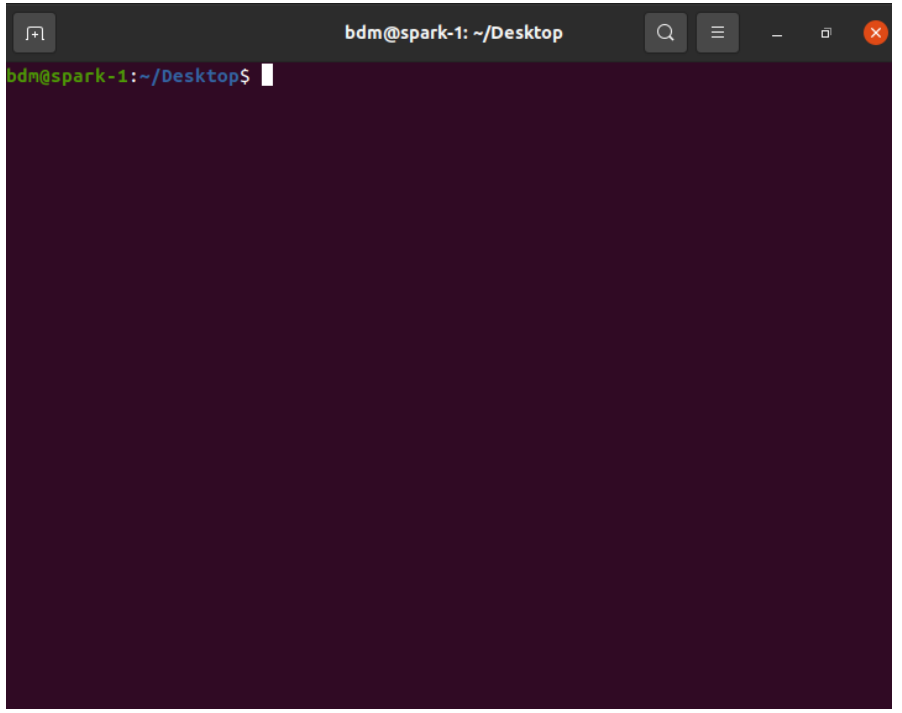
System configuration

Ubuntu basic environment installation

sudo apt update

sudo apt install -y net-tools openssh-server openjdk-8-jdk vim





Generate ssh key and copy to remote host

- ssh-keygen
- ls -al ~/.ssh/
- ssh-copy-id 192.168.88.171
- ssh-copy-id 192.168.88.172
- ssh-copy-id 192.168.88.173

- Check IP: ifconfig

```
bdm@spark-1:~$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/bdm/.ssh/id_rsa):
Created directory '/home/bdm/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/bdm/.ssh/id_rsa
Your public key has been saved in /home/bdm/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:EWOR1mk8yAXvQ+byTT4/j7XKCUpE2z2NEPqZBRFUsDg bdm@spark-1
The key's randomart image is:
+---[RSA 3072]----+
|       .*OoO=.   |
|       .=+O +    |
|       ..E++ .   |
|        .*= * o  |
|        So+=.+ . |
|        .o =  .  |
|         ...+   .|
|         . . oooo.|
|          .   +++.|
+----[SHA256]-----+
```

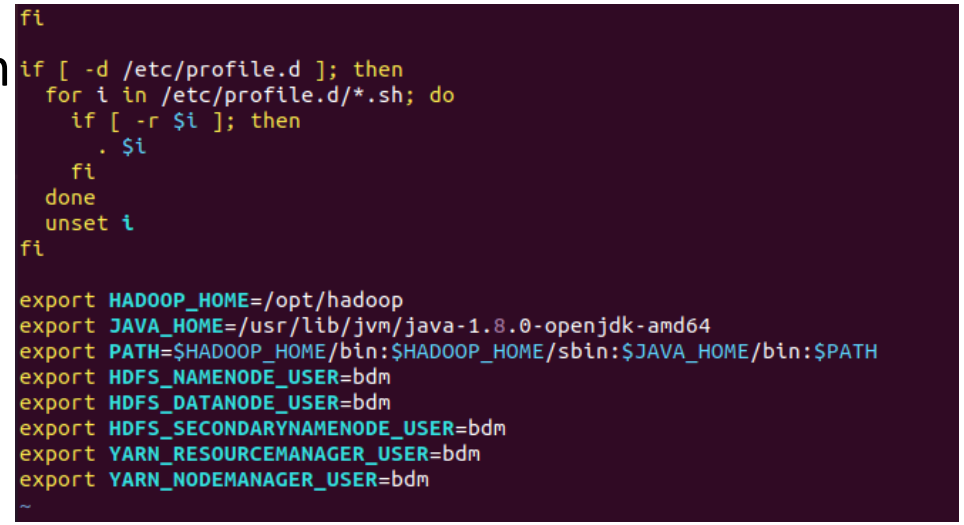
Edit hosts file

sudo vim /etc/hosts

```
127.0.0.1       localhost

192.168.88.171  spark-1
192.168.88.172  spark-2
192.168.88.173  spark-3

# The following lines are desirable for IPv6 capable hosts
::1     ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
```

Edit /etc/profile

- sudo vim /etc/profile
- content:
  - export HADOOP_HOME=/opt/hadoop
  - export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
  - export PATH=$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$JAVA_HOME/bin:$PATH
  - export HDFS_NAMENODE_USER=bdm
  - export HDFS_DATANODE_USER=bdm
  - export HDFS_SECONDARYNAMENODE_USER=bdm
  - export YARN_RESOURCEMANAGER_USER=bdm
  - export YARN_NODEMANAGER_USER=bdm



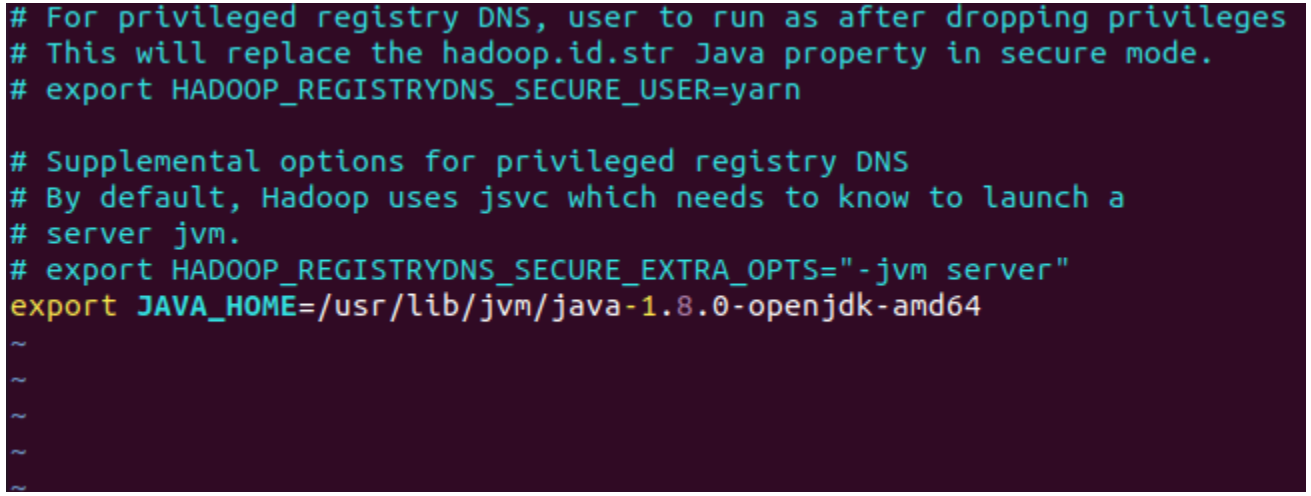
Hadoop installation

Hadoop Installation introduction

- wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz
- tar -zxvf hadoop-3.3.1.tar.gz
- sudo mv hadoop-3.3.1 /opt/hadoop
- cd /opt/hadoop
- Edit files:
  - etc/hadoop/hadoop-env.sh
  - etc/hadoop/core-site.xml
  - etc/hadoop/hdfs-site.xml
  - etc/hadoop/yarn-site.xml
  - etc/hadoop/mapred-site.xml
  - etc/hadoop/workers

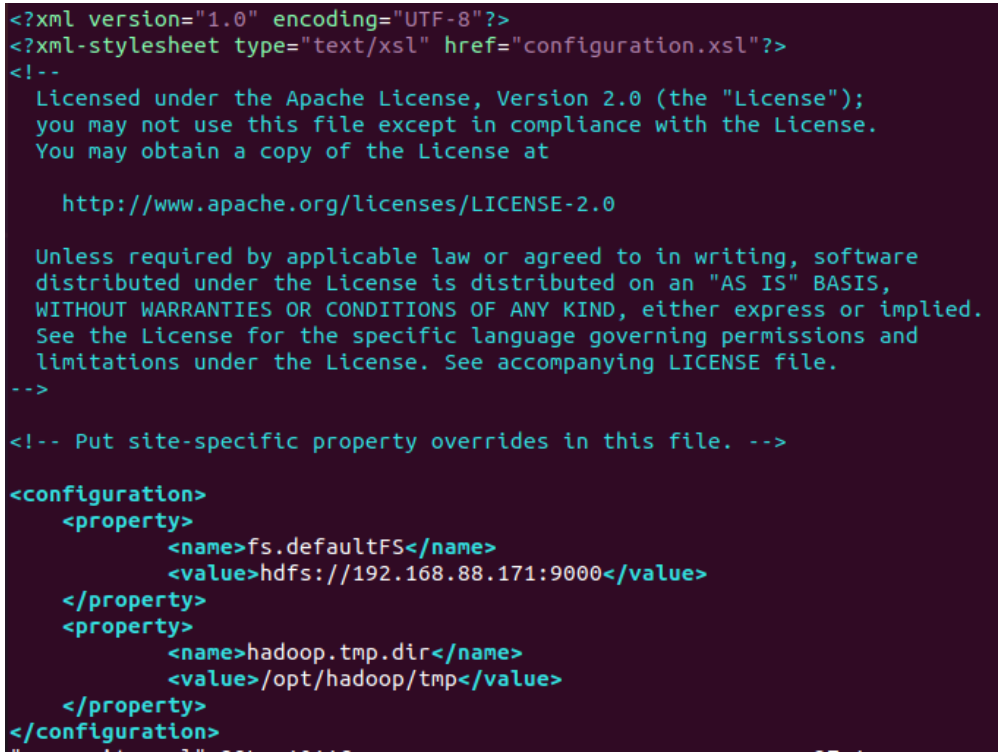
Edit etc/hadoop/hadoop-env.sh

export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64



Edit etc/hadoop/core-site.xml

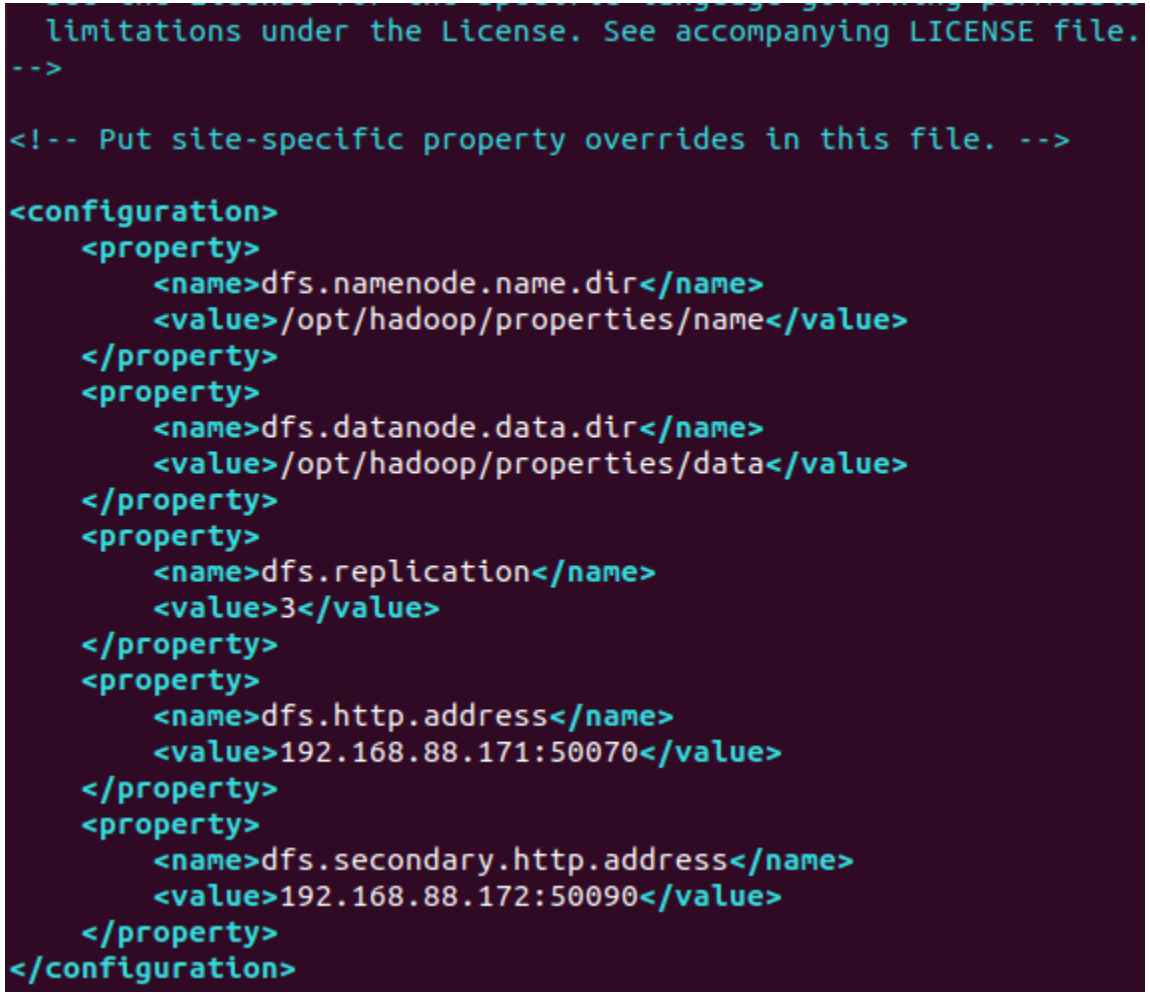
```
<configuration>
   <property>
        <name>fs.defaultFS</name>
        <value>hdfs://192.168.88.171:9000</val
   </property>
   <property>
        <name>hadoop.tmp.dir</name>
        <value>/opt/hadoop/tmp</value>
   </property>
</configuration>
```



```
<?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>fs.defaultFS</name>
            <value>hdfs://192.168.88.171:9000</value>
    </property>
    <property>
            <name>hadoop.tmp.dir</name>
            <value>/opt/hadoop/tmp</value>
    </property>
</configuration>
```

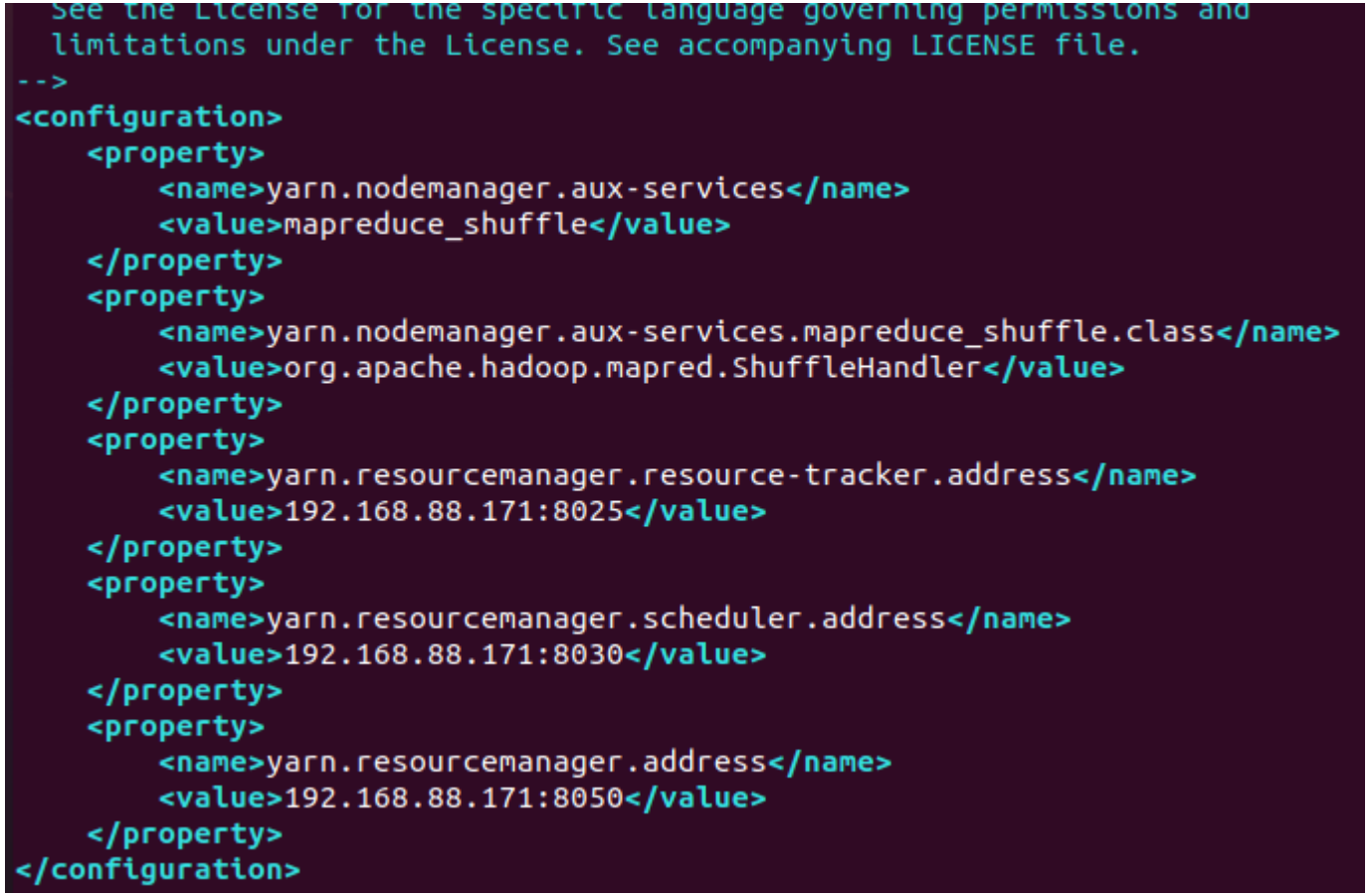
Edit etc/hadoop/hdfs-site.xml

```
<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/opt/hadoop/properties/name</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/opt/hadoop/properties/data</value>
  </property>
  <property>
    <name>dfs.replication</name>
    <value>3</value>
  </property>
  <property>
    <name>dfs.http.address</name>
    <value>192.168.88.171:50070</value>
  </property>
  <property>
    <name>dfs.secondary.http.address</name>
    <value>192.168.88.172:50090</value>
  </property>
</configuration>
```



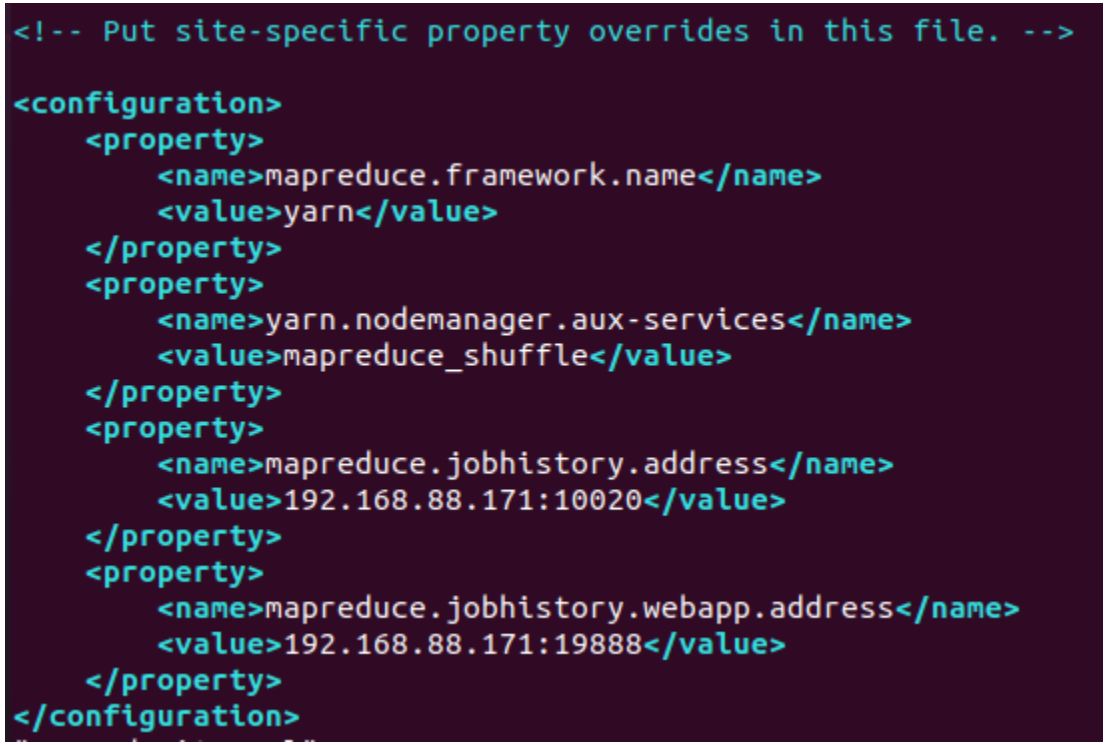
Edit etc/hadoop/yarn-site.xml

```
<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.resource-tracker.address</name>
    <value>192.168.88.171:8025</value>
  </property>
  <property>
    <name>yarn.resourcemanager.scheduler.address</name>
    <value>192.168.88.171:8030</value>
  </property>
  <property>
    <name>yarn.resourcemanager.address</name>
    <value>192.168.88.171:8050</value>
  </property>
</configuration>
```



Edit etc/hadoop/mapred-site.xml

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>mapreduce.jobhistory.address</name>
    <value>192.168.88.171:10020</value>
  </property>
  <property>
    <name>mapreduce.jobhistory.webapp.address</name>
    <value>192.168.88.171:19888</value>
  </property>
</configuration>
```



Edit etc/hadoop/workers

spark-1

spark-2

spark-3

Format hadoop namenode

- cd /opt/hadoop
- bin/hdfs namenode -format

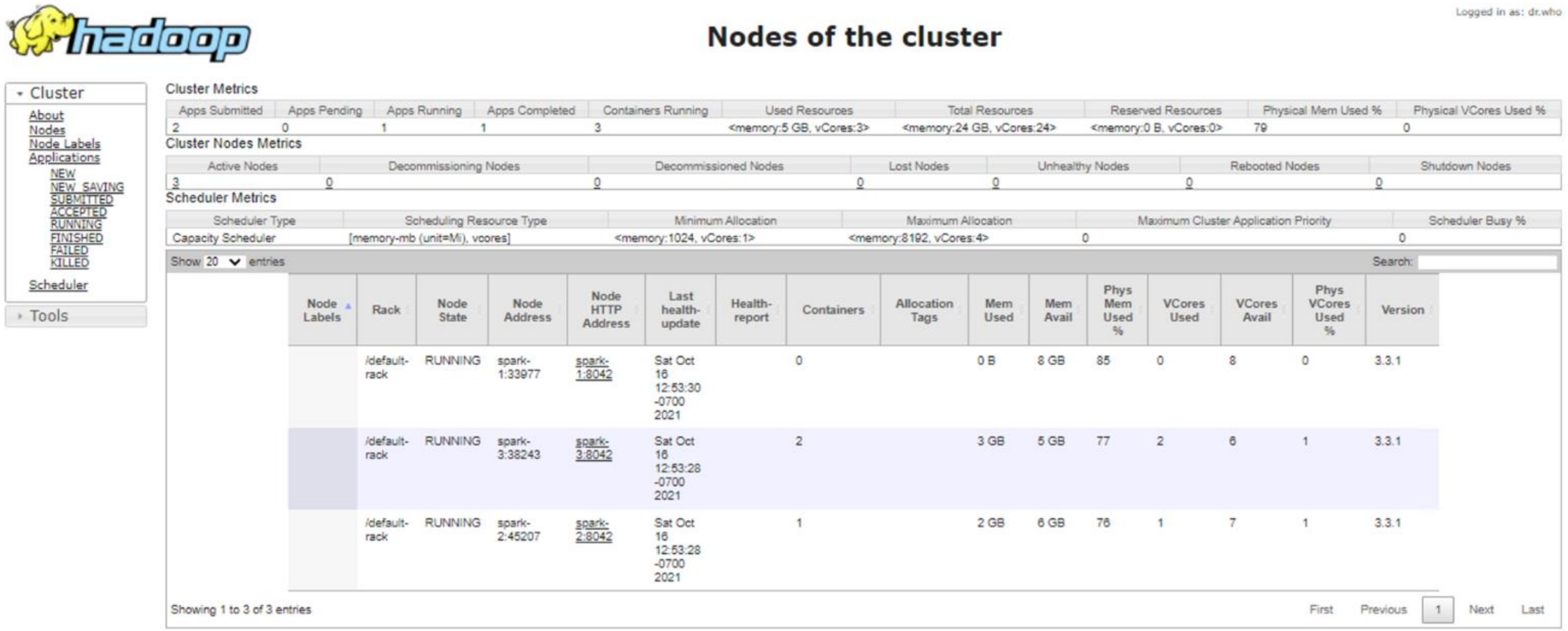
Result: **INFO common.Storage: Storage directory /opt/hadoop/properties/name has been successfully formatted.**

Running Hadoop

- sbin/start-all.sh
- Check hadoop cluster:
  - http://192.168.88.171:8088/cluster/nodes
  - http://192.168.88.171:50070/dfshealth.html#tab-datanode

Node information of the hadoop cluster

http://192.168.88.171:8088/cluster/nodes



Logged in as: dr.who

Nodes of the cluster

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %	Physical VCores Used %
2	0	1	1	3	<memory:5 GB, vCores:3>	<memory:24 GB, vCores:24>	<memory:0 B, vCores:0>	79	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
3	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority	Scheduler Busy %
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0	0

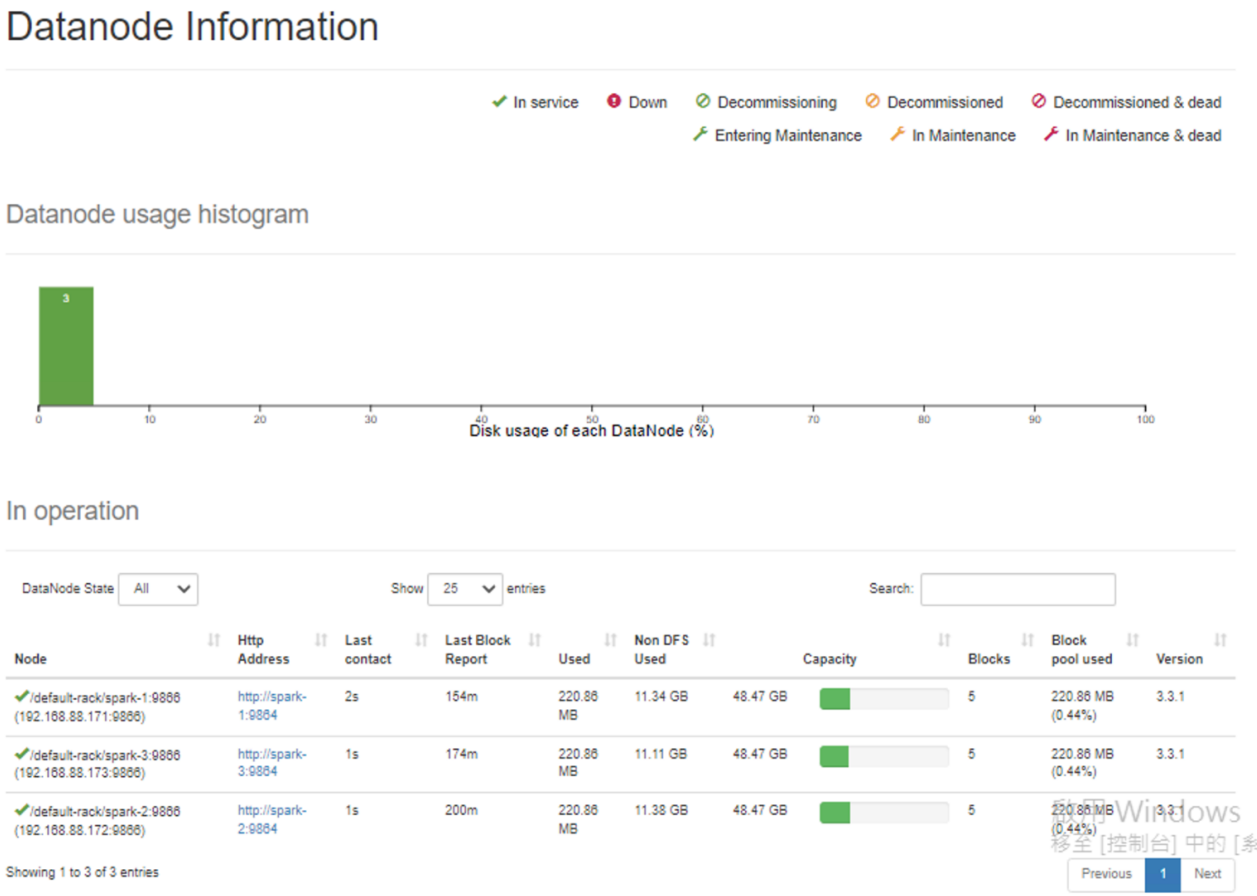
Show 20 entries Search:

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Allocation Tags	Mem Used	Mem Avail	Phys Mem Used %	VCores Used	VCores Avail	Phys VCores Used %	Version
	/default-rack	RUNNING	spark-1:33977	spark-1:8042	Sat Oct 16 12:53:30 -0700 2021		0		0 B	8 GB	85	0	8	0	3.3.1
	/default-rack	RUNNING	spark-3:38243	spark-3:8042	Sat Oct 16 12:53:28 -0700 2021		2		3 GB	5 GB	77	2	6	1	3.3.1
	/default-rack	RUNNING	spark-2:45207	spark-2:8042	Sat Oct 16 12:53:28 -0700 2021		1		2 GB	6 GB	76	1	7	1	3.3.1

Showing 1 to 3 of 3 entries First Previous 1 Next Last

Hadoop HDFS datanode information

http://192.168.88.171:50070/dfshealth.html#tab-datanode



Spark installation

Spark installation introduction

- wget https://www.apache.org/dyn/closer.lua/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz
- tar -zxvf spark-3.1.2-bin-hadoop3.2.tgz
- sudo mv spark-3.1.2-bin-hadoop3.2 /opt/spark
- cd /opt/spark/conf
- cp spark-defaults.conf.template spark-defaults.conf
- cp spark-env.sh.template spark-env.sh
- cp workers.template workers
- Edit files:
  - spark-defaults.conf
  - spark-env.sh
  - workers

Edit conf/spark-defaults.conf

- spark.master spark://192.168.88.171:7077
- spark.serializer org.apache.spark.serializer.KryoSerializer
- spark.ui.enabled true

Edit conf/spark-env.sh

- export HADOOP_HOME=/opt/hadoop
- export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
- export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
- export YARN_CONF_DIR=$HADOOP_HOME/etc/hadoop
- export PYSPARK_PYTHON=/bin/python3
- export PYSPARK_DRIVER_PYTHON=/bin/python3
- export SPARK_MASTER_IP=192.168.88.171
- SPARK_LOCAL_DIRS=/opt/spark

Edit conf/workers

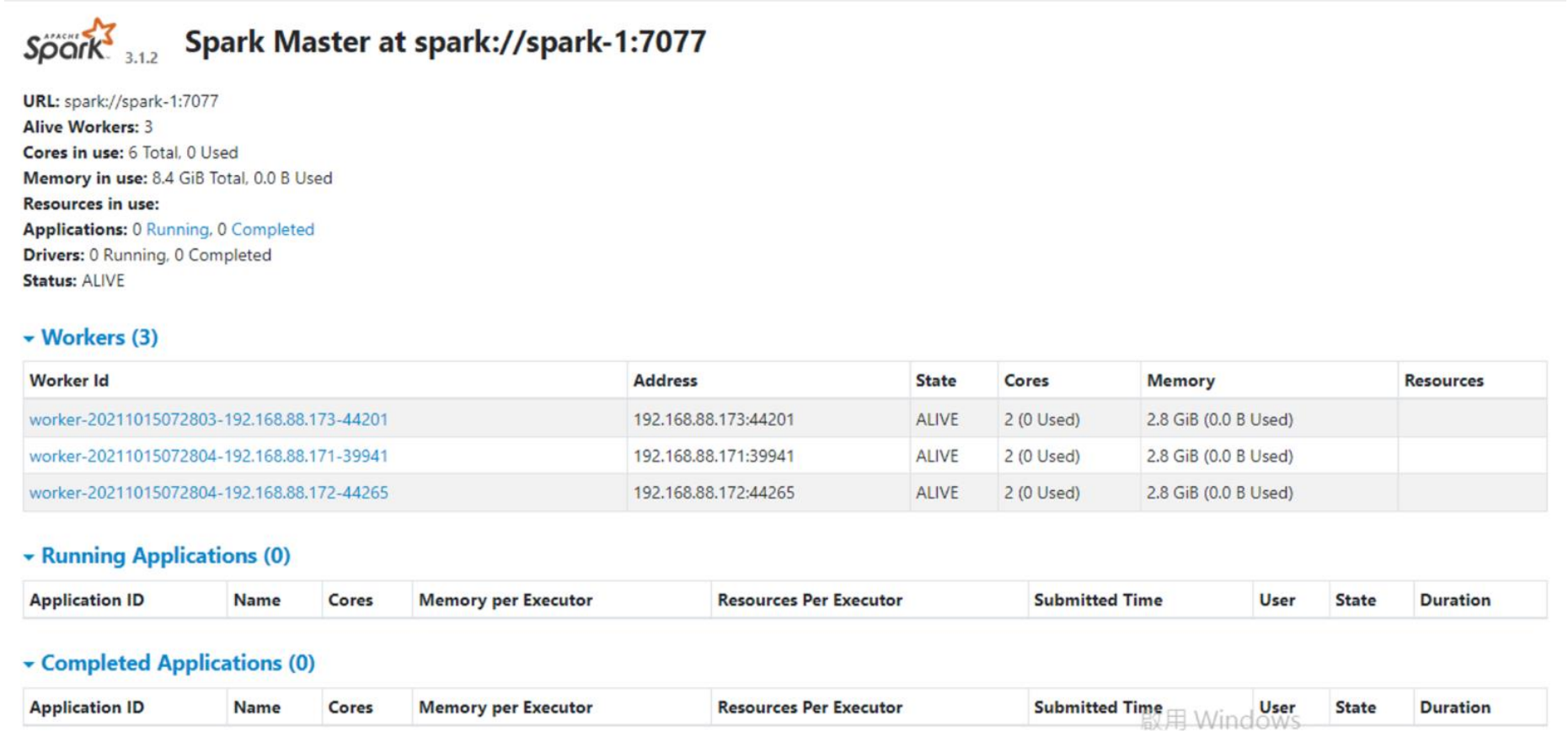
- spark-1
- spark-2
- spark-3

Running Spark

- sbin/start-all.sh
- Check spark cluster:
  - http://192.168.88.171:8080/

Spark web UI

http://192.168.88.171:8080/



Spark 3.1.2 **Spark Master at spark://spark-1:7077**

URL: spark://spark-1:7077
Alive Workers: 3
Cores in use: 6 Total, 0 Used
Memory in use: 8.4 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

▾ **Workers (3)**

Worker Id	Address	State	Cores	Memory	Resources
worker-20211015072803-192.168.88.173-44201	192.168.88.173:44201	ALIVE	2 (0 Used)	2.8 GiB (0.0 B Used)	
worker-20211015072804-192.168.88.171-39941	192.168.88.171:39941	ALIVE	2 (0 Used)	2.8 GiB (0.0 B Used)	
worker-20211015072804-192.168.88.172-44265	192.168.88.172:44265	ALIVE	2 (0 Used)	2.8 GiB (0.0 B Used)	

▾ **Running Applications (0)**

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

▾ **Completed Applications (0)**

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Test spark

bin/pyspark --master yarn --deploy-mode client

```
bdm@spark-1:/opt/spark$ bin/pyspark --master yarn --deploy-mode client
Python 3.8.10 (default, Sep 28 2021, 16:10:42)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
2021-10-15 08:07:19,813 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLev
el(newLevel).
2021-10-15 08:07:22,586 WARN yarn.Client: Neither spark.yarn.jars nor spark.yar
n.archive is set, falling back to uploading libraries under SPARK_HOME.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.1.2
      /_/

Using Python version 3.8.10 (default, Sep 28 2021 16:10:42)
Spark context Web UI available at http://spark-1:4040
Spark context available as 'sc' (master = yarn, app id = application_1634219491
648_0002).
SparkSession available as 'spark'.
>>>
```

Spark application instance on YARN

http://192.168.88.171:8088/cluster/apps