

# Hadoop

## 安装Hadoop

安装环境: MacOS

```
$ brew install hadoop
```

遇到error:

```
Error: Cannot install hadoop because conflicting formulae are installed.  
yarn: because both install `yarn` binaries
```

```
Please `brew unlink yarn` before continuing.
```

```
Unlinking removes a formula's symlinks from /usr/local. You can  
link the formula again after the install finishes. You can --force this  
install, but the build may fail or cause obscure side effects in the  
resulting software.
```

尝试unlink yarn:

```
$ brew unlink yarn
```

```
Unlinking /usr/local/Cellar/yarn/1.22.4... 2 symlinks removed
```

再次安装hadoop:

```
$ brew install hadoop
```

```
==> Summary
```

```
🍺 /usr/local/Cellar/openjdk/13.0.2+8_2: 631 files, 314.6MB
```

```
==> Installing hadoop
```

```
🍺 /usr/local/Cellar/hadoop/3.2.1_1: 22,397 files, 815.6MB, built in 55  
seconds
```

```
==> Caveats
```

```
==> openjdk
```

```
For the system Java wrappers to find this JDK, symlink it with  
sudo ln -sf /usr/local/opt/openjdk/libexec/openjdk.jdk  
/Library/Java/JavaVirtualMachines/openjdk.jdk
```

```
openjdk is keg-only, which means it was not symlinked into /usr/local,  
because it shadows the macOS `java` wrapper.
```

If you need to have openjdk first in your PATH run:

```
echo 'export PATH="/usr/local/opt/openjdk/bin:$PATH"' >>  
/Users/qq/.bash_profile
```

For compilers to find openjdk you may need to set:

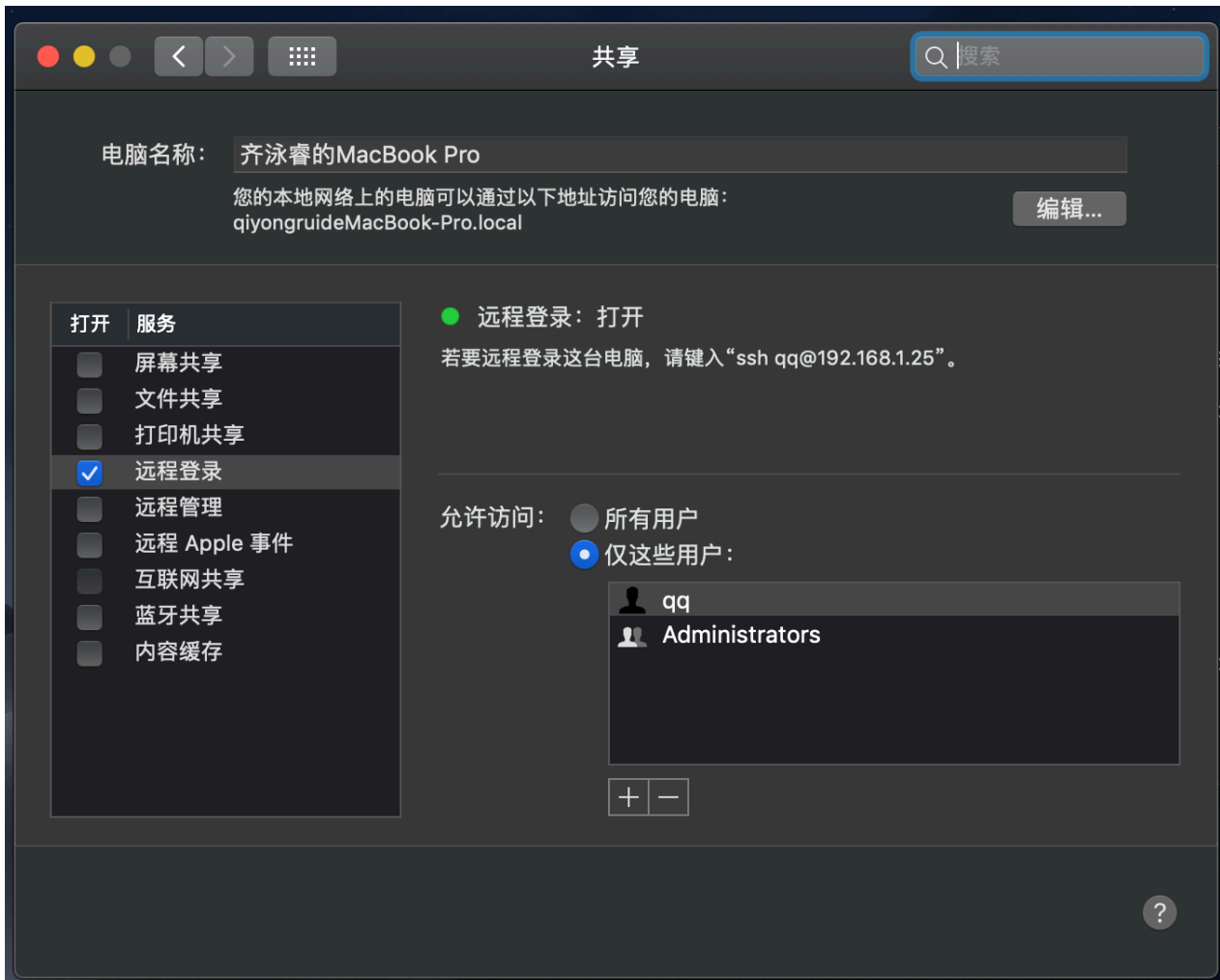
```
export CPPFLAGS="-I/usr/local/opt/openjdk/include"
```

查看是否安装成功:

```
$ hadoop version  
Hadoop 3.2.1  
Source code repository https://gitbox.apache.org/repos/asf/hadoop.git -r  
b3cbbb467e22ea829b3808f4b7b01d07e0bf3842  
Compiled by rohithsharmaks on 2019-09-10T15:56Z  
Compiled with protoc 2.5.0  
From source with checksum 776eaf9eee9c0ffc370bcbcl888737  
This command was run using  
/usr/local/Cellar/hadoop/3.2.1_1/libexec/share/hadoop/common/hadoop-common-  
3.2.1.jar
```

## 配置SSH

## 设置中打开共享权限



Mac本身带有ssh功能，Hadoop需要通过SSH来启动Slave列表中各台主机的守护进程。此处配置使得系统运行中免密码登录和访问节点。

密钥对保存在`/.ssh/id_rsa`文件中，进入`/.ssh`目录复制生成文件：

```
$ cp id_rsa.pub authorized_keys
```

最后输入 `ssh localhost` 测试

## 配置文件

进入 `/usr/local/Cellar/hadoop/3.2.1_1/etc/hadoop`

### hadoop-env.sh

首先进入 `/usr/libexec/java_home` 查看Java路径：

```
/Library/Java/JavaVirtualMachines/jdk1.8.0_201.jdk/Contents/Home
```

在hadoop-env.sh中删除 `#export JAVA_HOME=` 前的注释，并在其后添加Java路径：

```
hadoop-env.sh — Edited
hadoop-env.sh > No Selection
39 # Therefore, the vast majority (BUT NOT ALL!) of these defaults
40 # are configured for substitution and not append. If append
41 # is preferable, modify this file accordingly.
42
43 ###
44 # Generic settings for HADOOP
45 ###
46
47 # Technically, the only required environment variable is JAVA_HOME.
48 # All others are optional. However, the defaults are probably not
49 # preferred. Many sites configure these options outside of Hadoop,
50 # such as in /etc/profile.d
51
52 # The java implementation to use. By default, this environment
53 # variable is REQUIRED on ALL platforms except OS X!
54 export JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_201.jdk/Contents/Home
55
56 # Location of Hadoop. By default, Hadoop will attempt to determine
57 # this location based upon its execution path.
58 # export HADOOP_HOME=
59
60 # Location of Hadoop's configuration information. i.e., where this
61 # file is living. If this is not defined, Hadoop will attempt to
62 # locate it based upon its execution path.
63 #
64 # NOTE: It is recommend that this variable not be set here but in
65 # /etc/profile.d or equivalent. Some options (such as
66 # --config) may react strangely otherwise.
67 #
68 # export HADOOP_CONF_DIR=${HADOOP_HOME}/etc/hadoop
69
```

## core-site.xml

修改core-site.xml的文件参数，配置NameNode的主机名和端口号：

```
19 <configuration>
20   <property>
21     <name>hadoop.tmp.dir</name>
22     <value>/usr/local/Cellar/hadoop/hdfs/tmp</value>
23     <description>A base for other temporary directories</description>
24   </property>
25   <property>
26     <name>fs.default.name</name>
27     <value>hdfs://localhost:8020</value>
28   </property>
29 </configuration>
```

## hdfs-site.xml

设置HDFS的默认备份方式。

变量dfs.replication指定了每个HDFS数据库的复制次数。通常为3, 由于我们只有一台主机和一个伪分布式模式的DataNode，将此值修改为1：

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

## mapred-site.xml

配置MapReduce中的地址和端口号：

```
<configuration>
  <property>
    <name>mapred.job.tracker</name>
    <value>localhost:8021</value>
  </property>
</configuration>
```

# 格式化

初次安装和使用Hadoop之前格式化hdfs

```
$ /usr/local/Cellar/hadoop/3.2.1_1/bin/hadoop namenode -format
```

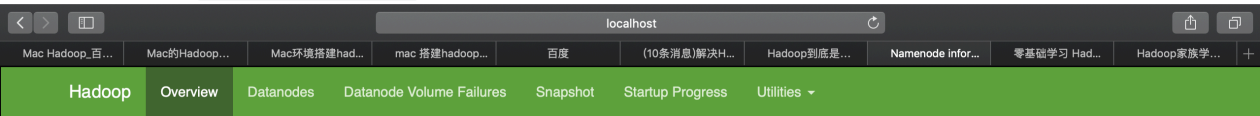
# 启动Hadoop

先启动NameNode和DataNode

```
$ cd /usr/local/Cellar/hadoop/3.2.1_1/sbin
$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [qiyongruideMacBook-Pro.local]
2020-06-23 21:02:32,845 WARN util.NativeCodeLoader: Unable to load native-
hadoop library for your platform... using builtin-java classes where
applicable
```

(WARN暂时不会有影响.....就先放在这里叭.....

web浏览框输入 localhost:9870



## Overview 'localhost:8020' (active)

Started:	Tue Jun 23 21:02:24 +0800 2020
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 23:56:00 +0800 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-3b6a04ce-8d12-4177-b7a2-abb7709a7ccf
Block Pool ID:	BP-1402840440-192.168.1.25-1592910989435

## 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 90.78 MB of 251 MB Heap Memory. Max Heap Memory is 1.78 GB.  
Non Heap Memory used 48.81 MB of 49.94 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	112.8 GB
Configured Remote Capacity:	0 B
DFS Used:	4.15 GB (4%)

启动Hadoop守护进程

```
start-all.sh
```

web地址栏输入 localhost:8088



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	Memory Used	Memory Total	Memory Reserved
0	0	0	0	0	0 B	8 GB	0 B

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes
1	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=M), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Re Mi N
No data available in table															

Showing 0 to 0 of 0 entries