

# 基于 Docker 的 Hadoop 和 Spark 实验环境部署

作者：李龙海

## 一. 安装 Docker 引擎

### 1.1 Windows 上安装 Docker 引擎（Docker Desktop）

1. OS 版本要求：在 WIN10 或者 WIN11 上都可以安装 Docker 引擎。
2. 详细安装方法见 Docker 官网的在线说明：Install Docker Desktop on Windows (<https://docs.docker.com/desktop/windows/install/>)
3. 不想看英文说明可以参照如下基本安装方法：
  - a) 下载安装文件，Docker 官网的下载地址为：  
<https://desktop.docker.com/win/main/amd64/Docker%20Desktop%20Installer.exe>
  - b) 下载后双击运行“Docker Desktop Installer.exe”然后根据提示安装。
  - c) 某些 Windows 版本在安装过程中可能会提示安装“WSL2 Linux 内核更新包”（<https://docs.microsoft.com/zh-cn/windows/wsl/install-manual#step-4---download-the-linux-kernel-update-package>），只需要根据“步骤 4 - 下载 Linux 内核更新包”做就可以。
  - d) 安装成功后双击桌面上的 Docker Desktop 图标运行 Docker 引擎，引擎启动后可以在桌面任务栏的右下角看到图标“”。单击该图标可以随时唤醒 Docker 引擎的图形管理界面。
  - e) 在 Docker Desktop 的 Settings 界面的“Docker Engine”栏目中修改配置文件，增加国内的 Docker 镜像仓库地址。例如，本人的配置文件内容如下（红色字体是增加的国内 Docker 镜像仓库地址）。

```
{  
  "builder": {  
    "gc": {  
      "defaultKeepStorage": "20GB",  
      "enabled": true  
    }  
  },  
  "experimental": false,  
  "features": {
```

```
"buildkit": true
},
"registry-mirrors": [
  "https://hub-mirror.c.163.com",
  "https://mirror.baidubce.com"
]
}
```

- f) 在 Windows 命令方式或者 Power Shell 工具中执行如下命令：

```
docker run hello-world
```

如果可以看到自动拉取镜像的过程并且看到提示“Hello from Docker!”，则证明 Docker 引擎安装成功。

## 1.2 Linux 上安装 Docker 引擎

Docker 最初就是专门针对 Linux 系统设计的，因此在 Linux 上安装 Docker 引擎非常方便，请自己在网上搜索安装教程。

## 二. 部署 Hadoop 和 Spark 实验环境

### 2.1 生成所需 Docker 镜像

1. 将教师分发的压缩包 hadoop-sandbox.zip 和 spark-install.zip 分别解压到 xxxxx/hadoop-sandbox 目录和 xxxxx/spark-install 目录

stem (C:) > source > dc22 > hadoop > hadoop-sandbox

<input type="checkbox"/> 名称	修改日期	类型	大小
 conf	2022/6/4 22:51	文件夹	
 data	2022/6/4 22:55	文件夹	
 .gitignore	2022/6/4 22:51	Git Ignore 源文件	1 KB
 docker-compose.yaml	2022/6/5 2:38	Yaml 源文件	3 KB
 README.md	2022/6/4 22:51	Markdown 源文件	6 KB

System (C:) > source > dc22 > hadoop > spark-install

<input type="checkbox"/> 名称	修改日期	类型	大小
 apache-maven-3.8.5-bin.tar.gz	2022/6/4 18:53	GZ 文件	8,470 KB
 Dockerfile	2022/6/5 0:37	文件	1 KB
 profile	2022/6/5 0:46	文件	1 KB
 settings.xml	2022/6/3 21:28	XML 源文件	11 KB
 spark-3.2.1-bin-hadoop3.2.tgz	2022/6/4 12:13	TGZ 文件	293,918 KB

- 在命令终端模式下将当前目录切换到 xxxxx/hadoop-sandbox，然后运行命令：

```
docker-compose up -d
```

首次运行需要保持网络畅通，并耐心等待 Docker 引擎自动下载所需的 Docker 镜像。

所有 Docker 镜像下载完成并成功启动，将看到如下提示：

```
C:\source\dc22\hadoop\hadoop-sandbox>docker-compose up -d
Creating network "hadoop-sandbox_default" with driver "bridge"
Creating hadoop-sandbox_namenode_1 ... done
Creating hadoop-sandbox_datanode_1 ... done
Creating hadoop-sandbox_resourcemanager_1 ... done
Creating hadoop-sandbox_jobhistoryserver_1 ... done
Creating hadoop-sandbox_nodemanager_1 ... done
Creating hadoop-sandbox_clientnode_1 ... done
Creating hadoop-sandbox_front_1 ... done
```

- 在上一步执行成功之后运行如下命令（目的是先关闭这些容器的运行）：

```
docker-compose down
```

```
C:\source\dc22\hadoop\hadoop-sandbox>docker-compose down
Stopping hadoop-sandbox_front_1 ... done
Stopping hadoop-sandbox_clientnode_1 ... done
Stopping hadoop-sandbox_nodemanager_1 ... done
Stopping hadoop-sandbox_jobhistoryserver_1 ... done
Stopping hadoop-sandbox_resourcemanager_1 ... done
Stopping hadoop-sandbox_datanode_1 ... done
Stopping hadoop-sandbox_namenode_1 ... done
Removing hadoop-sandbox_front_1 ... done
Removing hadoop-sandbox_clientnode_1 ... done
Removing hadoop-sandbox_nodemanager_1 ... done
Removing hadoop-sandbox_jobhistoryserver_1 ... done
Removing hadoop-sandbox_resourcemanager_1 ... done
Removing hadoop-sandbox_datanode_1 ... done
Removing hadoop-sandbox_namenode_1 ... done
Removing network hadoop-sandbox_default
```

- 将当前目录切换到 xxxxx/spark-install，然后运行命令（注意最后有个.符号）：

```
docker build -t packet23/hadoop-client:latest .
```

运行该命令时 Docker 引擎将根据 Dockerfile 的指示在名为“**hadoop-client**”的 Docker 镜像中下载并安装 python3.8，安装 spark-3.2.1 和 apache-maven-3.8.5。请保持网络畅通，并耐心等待。

- 上面步骤运行成功之后就生成了实验所需的全部 Docker 镜像。

## 2.2 启动和关闭实验环境

1. 启动实验环境：将当前目录切换到 xxxx/hadoop-sandbox，然后运行命令：

```
docker-compose up -d
```

再运行命令：

```
docker ps
```

可以看到共有 7 个 Docker 容器（虚拟机）在运行：

CONTAINER ID	IMAGE	NAMES	COMMAND	CREATED	STATUS	PORTS
cc26c79e971f	httpd:2.4		"httpd-foreground"	9 seconds ago	Up 5 seconds	127.0.0.1:8842->8842/tcp,
127.0.0.1:8080->8080/tcp,						127.0.0.1:9864->9864/tcp,
127.0.0.1:9870->9870/tcp,						80/tcp, 127.0.0.1:19888->19888/tcp
hadoop-sandbox_frontend_1	packet23/hadoop-client:latest		"/usr/bin/tini -- /d_"	9 seconds ago	Up 6 seconds	127.0.0.1:2222->22/tcp
7ff773bd0dde		hadoop-sandbox_clientnode_1				
a3c019effeed	packet23/hadoop-yarn-nodemanager:latest		"/usr/bin/tini -- /d_"	10 seconds ago	Up 9 seconds	
		hadoop-sandbox_nodemanager_1				
72371f1c476a	packet23/hadoop-mapred-jobhistoryserver:latest		"/usr/bin/tini -- /d_"	10 seconds ago	Up 3 seconds	
		hadoop-sandbox_jobhistoryserver_1				
16ede8b4765d	packet23/hadoop-yarn-resourcemanager:latest		"/usr/bin/tini -- /d_"	12 seconds ago	Up 10 seconds	
		hadoop-sandbox_resourcemanager_1				
fc576b0b00f1	packet23/hadoop-hdfs-datanode:latest		"/usr/bin/tini -- /d_"	12 seconds ago	Up 9 seconds	
		hadoop-sandbox_datanode_1				
7048d69648fa	packet23/hadoop-hdfs-namenode:latest		"/usr/bin/tini -- /d_"	13 seconds ago	Up 11 seconds	
		hadoop-sandbox_namenode_1				

在浏览器中输入网址 <http://localhost:8080/> 可以看到 Hadoop 集群的 5 个节点的运行状态。

### Single node Yarn cluster

- [Resource Manager](#)
- [Name Node](#)
- [Node Manager](#)
- [Data Node](#)
- [Job History Server](#)

另外 2 个节点是 hadoop-client（作为 Hadoop 和 Spark 的客户端系统，用于运行客户端程序），和 httpd（用于提供 Web 监控服务，我们通过 <http://localhost:8080/> 看到内容就是这个节点输出的）。

2. 用如下命令通过 ssh 协议远程登录到 hadoop-client 虚拟机内部：

```
ssh -p 2222 sandbox@localhost
```

如果有交互式提问一律回答 “yes”。登录密码为 “**sandbox**”。

从 ssh 中退出用 “logout” 或 “exit” 命令。

3. 关闭实验环境：将当前目录切换到 xxxx/hadoop-sandbox，然后运行命令：

```
docker-compose down
```

## 三. 关于 HDFS 的实验

1. 启动实验环境
2. 用 ssh 登录到 hadoop-client 虚拟机。以下命令都是在 ssh 登录后运行于 hadoop-client 虚拟机中。

3. 用“**hadoop fs**”或“**hdfs dfs**”命令在分布式文件系统 HDFS 中实现“浏览目录”、“创建子目录”、“删除子目录”、“创建文件”、“拷贝文件”、“移动子目录或文件”、“查看文件内容”、“删除文件”、“上传本地文件”等操作。

<b>hadoop fs -ls</b>	显示当前用户目录下的所有文件和目录
<b>hadoop fs -mkdir test</b>	在当前用户目录下创建子目录 test，创建多级目录 加上 -p
<b>hadoop fs -put .\input.txt test</b>	将本地文件 test.txt 上传到 HDFS 分布式文件系统的 test 目录下
<b>hadoop fs -cat input.txt</b>	查看 HDFS 中文件 input.txt 的内容
<b>hadoop fs -rm input.txt</b>	删除文件
<b>hadoop fs -rm -r test</b>	删除子目录（要加-r 参数）
<b>hadoop fs -cp URI [URI ...] &lt;dest&gt;</b>	cp 复制系统内文件
<b>hadoop fs -get[-ignorecrc] [-crc] &lt;src&gt; &lt;localdst&gt;</b>	下载文件到本地
<b>hadoop fs -mv URI [URI ...] &lt;dest&gt;</b>	将文件从源路径移动到目标路径
<b>hadoop fs -du URI [URI ...]</b>	显示文件大小

4. **特别提醒：**ssh 登录到 hadoop-client 虚拟机后当前目录为（用 pwd 命令可以查看）：  
/home/sandbox。hadoop-client 虚拟机的目录/home/sandbox 和宿主机的目录：  
“**xxxx\hadoop-sandbox\data\clientnode\home**”  
是绑定到一起的，两个目录的内容是完全相同的。因此，你可以利用这两个相互绑定的目录实现宿主机（即你的计算机）和 hadoop-client 虚拟机之间的数据共享。
5. 在浏览器中输入网址 http://localhost:9870/可以看到 HDFS 的 Web 监控服务，在该页面的“utilities”菜单的“Browse Directory”页面中也可以观察到 HDFS 中的目录和文件。

Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities

## Browse Directory

Show  entries
Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	mapred	hadoop	0 B	Jun 04 22:56	0	0 B	<a href="#">mr-history</a>	
<input type="checkbox"/>	drwxrwxrwt	hdfs	supergroup	0 B	Jun 05 01:13	0	0 B	<a href="#">tmp</a>	
<input type="checkbox"/>	drwxr-xr-x	hdfs	supergroup	0 B	Jun 04 22:56	0	0 B	<a href="#">user</a>	

Showing 1 to 3 of 3 entries

Previous
1
Next

6. ssh 登录到 hadoop-client 虚拟机后的用户名为 sandbox，它默认使用的 HDFS 路径为：

“/user/sandbox”

[Hadoop](#) [Overview](#) [Datanodes](#) [Datanode Volume Failures](#) [Snapshot](#) [Startup Progress](#) [Utilities](#)

## Browse Directory

Show  entries

Search:

☐ **Permission**

**Owner**

**Group**

**Size**

**Last Modified**

**Replication**

**Block Size**

**Name**

No data available in table

Showing 0 to 0 of 0 entries

Hadoop, 2022.

## 四. 关于 MapReduce 的实验

1. 启动实验环境
2. 用 ssh 登录到 hadoop-client 虚拟机。以下命令都是在 ssh 登录后运行于 hadoop-client 虚拟机中。
3. 将当前目录切换到“/home/sandbox/mapreduce-demo”，并查看 commands.txt 的内容：

```
sandbox@clientnode:~$ cd /home/sandbox/mapreduce-demo/
sandbox@clientnode:~/mapreduce-demo$ cat commands.txt
mvn clean
mvn package

hadoop fs -mkdir input
hadoop fs -put ./input_file.txt input
hadoop fs -cat input/input_file.txt

hadoop jar ./target/WordCountDemo.jar com.org.xidian.MapReduceWordCountDemo input/input_file.txt output

hadoop fs -ls output
hadoop fs -cat output/part-r-00000
```

4. 依次运行 commands.txt 中的命令。
5. 这些命令的含义如下：

```
#利用Maven编译并打包MapReduce程序
mvn clean
mvn package

#向HDFS上传input_file.txt文件
hadoop fs -mkdir input
hadoop fs -put ./input_file.txt input
hadoop fs -cat input/input_file.txt

#运行WordCount这个MapReduce程序
hadoop jar ./target/WordCountDemo.jar com.org.xidian.MapReduceWordCountDemo input/input_file.txt output

#查看运行结果
hadoop fs -ls output
hadoop fs -cat output/part-r-00000
hadoop fs -rm -r output
```

6. 如果 MapReduce 程序运行成功，则可以看到如下输出（单词计数结果被保存在了 HDFS 的/user/sandbox/output/part-r-00000 文件中。

```
sandbox@clientnode:~/mapreduce-demo$ hadoop fs -cat output/part-r-000000
aaa      4
bbb      3
ccc      3
ddd      4
eee      3
fff      1
ggg      1
www      1
```

7. 也可以利用 HDFS 的 Web 监控服务 <http://localhost:9870/> (“utilities” 菜单的 “Browse Directory”) 观察计算结果。
8. 也可以利用 <http://localhost:19888/jobhistory> 的 MapReduce 任务监控服务观察 MapReduce 程序的运行状态。

## 五. 关于 Spark 的实验

1. 启动实验环境
2. 用 ssh 登录到 hadoop-client 虚拟机。以下命令都是在 ssh 登录后运行于 hadoop-client 虚拟机中。
3. 将当前目录切换到 “/home/sandbox/spark-demo”，并查看 commands 的内容：

```
sandbox@clientnode:~/spark-demo$ pwd
/home/sandbox/spark-demo
sandbox@clientnode:~/spark-demo$ cat commands
mvn clean
mvn package

hadoop fs -put ./input_file.txt
hadoop fs -cat input_file.txt

spark-submit --class org.apache.spark.examples.WordCount --master local ./target/spark-examples-1.0.jar input_file.txt

hadoop fs -ls output
hadoop fs -cat output/part-000000
hadoop fs -rm -r output
```

4. 依次运行 commands 中的命令。
5. 如果上面的 Spark 程序运行成功，则可以看到 output/part-000000 中单词计数的统计结果。
6. 也可以用 Python 语言编写 Spark 程序。/home/sandbox/目录中 wordcount.py 就是用 Python 语言编写的实现单词计数功能的 Spark 程序。其运行方法如下(用 spark-submit)：

```
sandbox@clientnode:~$ pwd
/home/sandbox
sandbox@clientnode:~$ ls
input_file.txt  mapreduce-demo  spark-demo  wordcount.py
sandbox@clientnode:~$ hadoop fs -put ./input_file.txt
sandbox@clientnode:~$ spark-submit wordcount.py
```

7. 查看运行结果的方法如下：

