

Spark-2.3.2-bin-hadoop2.6 镜像构建

一、 实现目的

基于容器技术，搭建 Spark-Hadoop 开发环境，以供开发人员快速掌握并应用到实际开发中，有关 Spark 的优点等知识这里不再赘述。暂且把优点总结如下：

- 1) 技术积累
- 2) 省时省力省钱

二、 环境准备

1) sequenceiq/docker-spark 源码

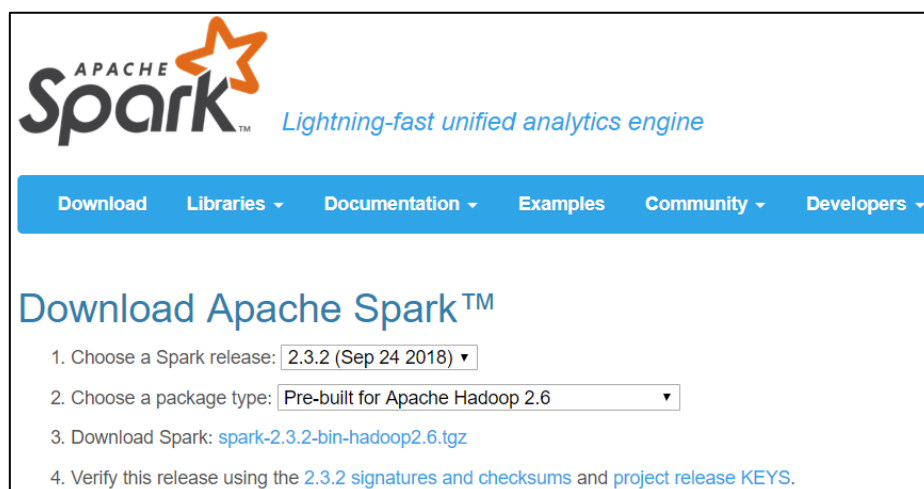
下载 sequenceiq/ docker-spark 镜像构建源码，下载命令如下：

```
git clone https://github.com/sequenceiq/docker-spark
```

2) sequenceiq/hadoop-docker:2.6.0 官网镜像

```
docker pull sequenceiq/hadoop-docker:2.6.0
```

3) spark-2.3.2-bin-hadoop2.6.tgz



4) jdk-8u25-linux-x64.tar.gz

请自行下载。

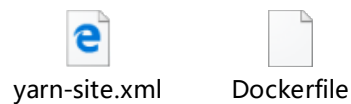
三、 镜像构建

1) 目录结构

基础目录结构如下图所示：

```
[root@localhost docker-spark]# tree
.
├── bootstrap.sh
├── Dockerfile
├── jdk-8u25-linux-x64.tar.gz
├── LICENSE
├── README.md
├── spark-2.3.2-bin-hadoop2.6.tgz
├── yarn-remote-client
│   ├── core-site.xml
│   └── yarn-site.xml
└── yarn-site.xml
```

其中，docker-spark 为 sequenceiq/docker-spark 源码目录，Dockerfile 及 yarn-site.xml 如下附件。



2) 构建命令

```
docker build --rm -t kngines/spark:2.3.2 .
```

3) 结果输出

```
Complete!
Removing intermediate container 4304bb84242e
--> b64f1cc9f4b6
Step 20/20 : ENTRYPOINT ["/etc/bootstrap.sh"]
--> Running in 8b5ad162196e
Removing intermediate container 8b5ad162196e
--> 45ad0468703a
Successfully built 45ad0468703a
Successfully tagged kngines/spark:2.3.2
[root@localhost docker-spark]# docker images
```

REPOSITORY	TAG	IMAGE ID	CREATED	SIZE
kngines/spark	2.3.2	45ad0468703a	10 seconds ago	3.05GB

4) 创建容器

```
docker run -it -p 8088:8088 -p 8042:8042 -p 4040:4040 -h sandbox  
kngines/spark:2.3.2 bash
```

5) 运行模式

1. Local 模式命令

```
spark-shell --master local --driver-memory 1g --executor-memory 1g --executor-cores 1
```

2. Yarn 模式命令

```
spark-shell --master yarn --deploy-mode client --driver-memory 512m --executor-memory 512m --executor-cores 1
```

6) 运行效果

```
[root@localhost docker-spark]# docker run -it -p 8088:8088 -p 8042:8042 -p 4040:4040 -h sandbox kngines/spark:2.3.2 bash
/
Starting sshd:
[ OK ]
Starting namenodes on [sandbox]
sandbox: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-sandbox.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-sandbox.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-sandbox.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-sandbox.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-sandbox.out
bash-4.1# java -version
java version "1.5.0"
gij (GNU libgcj) version 4.4.7 20120313 (Red Hat 4.4.7-23)

Copyright (C) 2007 Free Software Foundation, Inc.
This is free software; see the source for copying conditions. There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
bash-4.1# java -version
gij: unrecognized option -- '-version'
Try 'gij --help' for more information.
bash-4.1# spark-shell --master local --driver-memory 1g --executor-memory 1g --executor-cores 1
2019-01-22 00:17:02 WARN NativeCodeLoader:62 - Unable to load native-hadoop 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 setLogLevel(newLevel).
Spark context Web UI available at http://sandbox:4040
Spark context available as 'sc' (master = local, app id = local-1548134228625).
Spark session available as 'spark'.
Welcome to

  ____      _
 / ___|    / \
| |  | |  / _ \
| |  | | / ___ \
| |  | |/_/   \_\
| |  | |
|_|  |_|

 version 2.3.2

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_25)
Type in expressions to have them evaluated.
Type :help for more information.
```

```
hadoop fs -mkdir -p /spark
hadoop fs -put hello.txt /spark
hadoop fs -cat /spark/out/p*
```

五、 Web UI

3) Spark Web UI

← → ↻ 🔒 不安全 | 192.168.136.129:8088/proxy/application_1548134203925_0002/jobs/

SPARK

2.3.2

Jobs

Stages

Storage

Environment

Executors

Spark shell applica

Spark Jobs (?)

User: root
Total Uptime: 2.4 h
Scheduling Mode: FIFO
Completed Jobs: 2
[Event Timeline](#)

Completed Jobs (2)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	count at <console>:25 count at <console>:25	2019/01/22 02:18:03	0.1 s	1/1	2/2
0	runJob at SparkHadoopWriter.scala:78 runJob at SparkHadoopWriter.scala:78	2019/01/22 00:50:52	4 s	2/2	4/4

← → ↻ 🔒 不安全 | 192.168.136.129:8088/proxy/application_1548134203925_0002/executors/

SPARK

2.3.2

Jobs

Stages

Storage

Environment

Executors

Spark shell applica

Executors

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Blacklisted
Active(3)	0	0.0 B / 293.5 MB	0.0 B	2	0	0	6	6	8 s (0.5 s)	56 B	203 B	203 B	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B	0
Total(3)	0	0.0 B / 293.5 MB	0.0 B	2	0	0	6	6	8 s (0.5 s)	56 B	203 B	203 B	0

Executors

Show 20 entries

Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs	Thread Dump
driver	sandbox:34472	Active	0	0.0 B / 97.8 MB	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B		Thread Dump
1	sandbox:41680	Active	0	0.0 B / 97.8 MB	0.0 B	1	0	0	3	3	4 s (0.4 s)	19 B	100 B	98 B	stdout stderr	Thread Dump
2	sandbox:35823	Active	0	0.0 B / 97.8 MB	0.0 B	1	0	0	3	3	4 s (0.2 s)	37 B	103 B	105 B	stdout stderr	Thread Dump

Showing 1 to 3 of 3 entries

[Previous](#) 1 [Next](#)