W7-9

四. 实验过程

Spark 部署

1. 单机集中式部署

1.1 运行 Spark 应用程序

1.1.1 通过 Spark-Shell 运行应用程序I

- 进入 Spark-Shell

```
Spark context Web UI available at http://219.228.135.124:4040
Spark context available as 'sc' (master = local, app id = local-1571552629412).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.4
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_221)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

- 在 scala> 后输入 Scala 代码, 此处执行的是统计 /home/you/spark-2.4.4/RELEASE 文件中的单词数量

```
scala>  sc.textFile("file:///home/syx/spark-2.4.4/RELEASE").flatMap(_.split(" ")
).map((_,1)).reduceByKey(_+_).collect
res0: Array[(String, Int)] = Array((-Psparkr,1), (-B,1), (Spark,1), (-Pkubernete
s,1), (-Pyarn,1), (2.4.4,1), (Build,1), (built,1), (-Pflume,1), (-DzincPort=3036
,1), (flags:,1), (-Phive-thriftserver,1), (-Pmesos,1), (for,1), (-Phive,1), (-Pk
afka-0-8,1), (2.7.3,1), (-Phadoop-2.7,1), (Hadoop,1))
```

1.1.2 通过提交 Jar 包运行应用程序

- `~/spark-2.4.4/bin/spark-submit \` `--master local \` `--class org.apache.spark.examples.SparkPi \` `~/spark-2.4.4/examples/jars/spark-examples_2.11-2.4.4.jar`
- 运行结果如下图所示:

```
Pi is roughly 3.139915699578498
19/10/20 14:29:07 INFO server.AbstractConnector: Stopped Spark@3f2049b6{HTTP/1.1
,[http/1.1]}{0.0.0.0:4040}
19/10/20 14:29:07 INFO ui.SparkUI: Stopped Spark web UI at http://219.228.135.12
4:4040
19/10/20 14:29:07 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMas
terEndpoint stopped!
19/10/20 14:29:07 INFO memory.MemoryStore: MemoryStore cleared
19/10/20 14:29:07 INFO storage.BlockManager: BlockManager stopped
19/10/20 14:29:07 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
19/10/20 14:29:07 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinator
Endpoint: OutputCommitCoordinator stopped!
19/10/20 14:29:07 INFO spark.SparkContext: Successfully stopped SparkContext
19/10/20 14:29:07 INFO util.ShutdownHookManager: Shutdown hook called
19/10/20 14:29:07 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-a
2975307-1598-4195-aad3-05a2e467749c
19/10/20 14:29:07 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-4
3e4f1a0-0865-4bbf-a0e8-fb5969b0e3fb
syx@syx-OptiPlex-7050:~$
```

在运行过程中另起一个终端执行 jps 查看进程. 此时只会出现 SparkSubmit 进程, 应用程序运行结束后该进程消失

```
syx@syx-OptiPlex-7050:~$ jps
25040 NameNode
4386 Jps
30724 RemoteMavenServer36
25497 SecondaryNameNode
27049 NodeManager
26874 ResourceManager
30475 Main
4301 SparkSubmit
349 Launcher
25245 DataNode
```

2.单机伪分布式部署

2.2 修改配置

2.2.1 修改 spark-env.sh 文件

- 在末尾添加

```
export SPARK_MASTER_IP=localhost
export SPARK_MASTER_PORT=7077
export JAVA_HOME=/usr/local/jdk1.8
```

2.2.2 修改 slaves 文件

- `mv ~/spark-2.4.4/conf/slaves.template ~/spark-2.4.4/conf/slaves`

2.2.3修改 spark-defaults.conf 文件

- `mv ~/spark-2.4.4/conf/spark-defaults.conf.template ~/spark-2.4.4/conf/spark-defaults.conf`
- `vi ~/spark-2.4.4/conf/spark-defaults.conf`
- 在末尾添加

```
spark.eventLog.enabled=true
spark.eventLog.dir = hdfs://localhost:9000/tmp/spark_history
spark.history.fs.logDirectory=hdfs://localhost:9000/tmp/spark/spark_history
```

- 并在 HDFS 中建立目录 /tmp/spark_history

  `~/hadoop-2.9.2/bin/hdfs dfs -mkdir -p /tmp/spark_history`

2.3 启动服务

2.3.1 启动 Spark

```
syx@syx-OptiPlex-7050:~$ ~/spark-2.4.4/sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /home/syx/spark-2.4.4
/logs/spark-syx-org.apache.spark.deploy.master.Master-1-syx-OptiPlex-7050.out
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/syx/
spark-2.4.4/logs/spark-syx-org.apache.spark.deploy.worker.Worker-1-syx-OptiPlex-
7050.out
```

2.3.2 启动应用日志服务器

```
syx@syx-OptiPlex-7050:~$ ~/spark-2.4.4/sbin/start-history-server.sh
starting org.apache.spark.deploy.history.HistoryServer, logging to /home/syx/spa
rk-2.4.4/logs/spark-syx-org.apache.spark.deploy.history.HistoryServer-1-syx-Opti
Plex-7050.out
failed to launch: nice -n 0 /home/syx/spark-2.4.4/bin/spark-class org.apache.spa
rk.deploy.history.HistoryServer
        at org.apache.spark.deploy.history.FsHistoryProvider.<init>(FsHistoryPro
vider.scala:207)
        at org.apache.spark.deploy.history.FsHistoryProvider.<init>(FsHistoryPro
vider.scala:86)
        ... 6 more
  Caused by: java.io.FileNotFoundException: File does not exist: hdfs://localhos
t:9000/tmp/spark/spark_history
        at org.apache.hadoop.hdfs.DistributedFileSystem$22.doCall(DistributedFil
eSystem.java:1309)
        at org.apache.hadoop.hdfs.DistributedFileSystem$22.doCall(DistributedFil
eSystem.java:1301)
        at org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkRes
olver.java:81)
        at org.apache.hadoop.hdfs.DistributedFileSystem.getFileStatus(Distribute
dFileSystem.java:1317)
        at org.apache.spark.deploy.history.FsHistoryProvider.org$apache$spark$de
ploy$history$FsHistoryProvider$$startPolling(FsHistoryProvider.scala:257)
        ... 9 more
full log in /home/syx/spark-2.4.4/logs/spark-syx-org.apache.spark.deploy.history
.HistoryServer-1-syx-OptiPlex-7050.out
```

2.4 查看服务信息

- `jps`

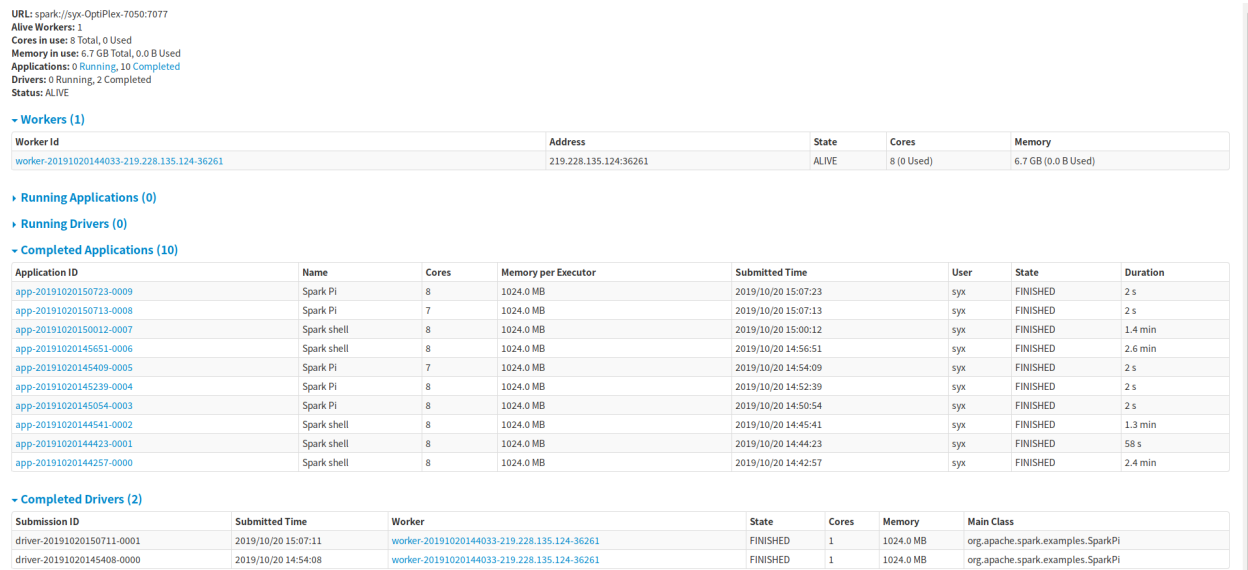
```
syx@syx-OptiPlex-7050:~$ jps
25040 NameNode
30724 RemoteMavenServer36
25497 SecondaryNameNode
27049 NodeManager
26874 ResourceManager
30475 Main
6044 Jps
349 Launcher
25245 DataNode
5597 Master
5758 Worker
```

在单机伪分布式部署模式下, 该节点既充当 Master, 又充当 Worker, 故该节点上会有两个进程: Master 和 Worker

- 查看 Spark 服务日志

```
syx@syx-OptiPlex-7050:~$ ls ~/spark-2.4.4/logs/*.out
/home/syx/spark-2.4.4/logs/spark-syx-org.apache.spark.deploy.history.HistoryServer-1-syx-OptiPlex-7050.out
/home/syx/spark-2.4.4/logs/spark-syx-org.apache.spark.deploy.master.Master-1-syx-OptiPlex-7050.out
/home/syx/spark-2.4.4/logs/spark-syx-org.apache.spark.deploy.worker.Worker-1-syx-OptiPlex-7050.out
```

- 访问 Spark Web 界面, 可看到 Master 和 Worker: http://localhost:8080



URL: spark://syx-OptiPlex-7050:7077
Alive Workers: 1
Cores in use: 8 Total, 0 Used
Memory in use: 6.7 GB Total, 0.0 B Used
Applications: 0 Running, 10 Completed
Drivers: 0 Running, 2 Completed
Status: ALIVE

▾ Workers (1)

Worker Id	Address	State	Cores	Memory
worker-20191020144033-219.228.135.124-36261	219.228.135.124:36261	ALIVE	8 (0 Used)	6.7 GB (0.0 B Used)

▸ Running Applications (0)

▸ Running Drivers (0)

▾ Completed Applications (10)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20191020150723-0009	Spark Pi	8	1024.0 MB	2019/10/20 15:07:23	syx	FINISHED	2 s
app-20191020150713-0008	Spark Pi	7	1024.0 MB	2019/10/20 15:07:13	syx	FINISHED	2 s
app-20191020150012-0007	Spark shell	8	1024.0 MB	2019/10/20 15:00:12	syx	FINISHED	1.4 min
app-20191020145651-0006	Spark shell	8	1024.0 MB	2019/10/20 14:56:51	syx	FINISHED	2.6 min
app-20191020145409-0005	Spark Pi	7	1024.0 MB	2019/10/20 14:54:09	syx	FINISHED	2 s
app-20191020145239-0004	Spark Pi	8	1024.0 MB	2019/10/20 14:52:39	syx	FINISHED	2 s
app-20191020145054-0003	Spark Pi	8	1024.0 MB	2019/10/20 14:50:54	syx	FINISHED	2 s
app-20191020144541-0002	Spark shell	8	1024.0 MB	2019/10/20 14:45:41	syx	FINISHED	1.3 min
app-20191020144423-0001	Spark shell	8	1024.0 MB	2019/10/20 14:44:23	syx	FINISHED	58 s
app-20191020144257-0000	Spark shell	8	1024.0 MB	2019/10/20 14:42:57	syx	FINISHED	2.4 min

▾ Completed Drivers (2)

Submission ID	Submitted Time	Worker	State	Cores	Memory	Main Class
driver-20191020150711-0001	2019/10/20 15:07:11	worker-20191020144033-219.228.135.124-36261	FINISHED	1	1024.0 MB	org.apache.spark.examples.SparkPi
driver-20191020145408-0000	2019/10/20 14:54:08	worker-20191020144033-219.228.135.124-36261	FINISHED	1	1024.0 MB	org.apache.spark.examples.SparkPi

2.5 运行 Spark 应用程序

2.5.1 通过 Spark-Shell 运行应用程序

- 进入 Spark-Shell

```
Spark context Web UI available at http://219.228.135.124:4040
Spark context available as 'sc' (master = spark://127.0.1.1:7077, app id = app-2
0191020144257-0000).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.4
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_221)
Type in expressions to have them evaluated.
Type :help for more information.
```

(注: 如使用 localhost 无法正常启动, 可尝试将 localhost 改为 127.0.1.1)

- 在 scala> 后输入 Scala 代码. 此处执行的是统计 RELEASE 文件中的单词数量

  `sc.textFile("hdfs://localhost:9000/user/you/spark_input/RELEASE").flatMap(_.split(" ")).map((_,1)).reduceByKey(_+_).collect`

  执行后应打印出如下结果

```
scala> sc.textFile("hdfs://localhost:9000/user/syx/spark_input/RELEASE").flatMap(_.split(
" ")).map((_,1)).reduceByKey(_+_).collect
res0: Array[(String, Int)] = Array((-Psparkr,1), (2.4.4,1), (Build,1), (built,1), (-Pflum
e,1), (-Phive-thriftserver,1), (-Pmesos,1), (2.7.3,1), (-Phadoop-2.7,1), (-B,1), (Spark,1
), (-Pkubernetes,1), (-Pyarn,1), (-DzincPort=3036,1), (flags:,1), (for,1), (-Phive,1), (-
Pkafka-0-8,1), (Hadoop,1))
```

2.5.2 通过提交 Jar 包运行应用程序

- (注: 如使用 localhost 无法正常启动, 可尝试将 localhost 改为 127.0.1.1)
- Client 提交模式 (默认), 此模式下 Driver 运行在客户端, 可以在客户端看到应用程序运行过程中的信息

```
syx@syx-OptiPlex-7050:~$ ~/spark-2.4.4/bin/spark-submit \
>   --deploy-mode client \
>  --master spark://127.0.1.1:7077 \
>   --class org.apache.spark.examples.SparkPi \
>   ~/spark-2.4.4/examples/jars/spark-examples_2.11-2.4.4.jar
```

运行结果如下图所示:

```
Pi is roughly 3.1443157215786077
19/10/20 14:50:56 INFO server.AbstractConnector: Stopped Spark@27dc79f7{HTTP/1.1,[http/1.
1]}{0.0.0.0:4040}
19/10/20 14:50:56 INFO ui.SparkUI: Stopped Spark web UI at http://219.228.135.124:4040
19/10/20 14:50:56 INFO cluster.StandaloneSchedulerBackend: Shutting down all executors
19/10/20 14:50:56 INFO cluster.CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each
executor to shut down
19/10/20 14:50:56 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoi
nt stopped!
19/10/20 14:50:56 INFO memory.MemoryStore: MemoryStore cleared
19/10/20 14:50:56 INFO storage.BlockManager: BlockManager stopped
19/10/20 14:50:56 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
19/10/20 14:50:56 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:
 OutputCommitCoordinator stopped!
19/10/20 14:50:56 INFO spark.SparkContext: Successfully stopped SparkContext
19/10/20 14:50:56 INFO util.ShutdownHookManager: Shutdown hook called
19/10/20 14:50:56 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-c3a58d88-6
c59-4b97-8d41-720f5de8dcf8
19/10/20 14:50:56 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-e2131446-e
72a-4c4f-add6-680a91c093d9
syx@syx-OptiPlex-7050:~$
```

在运行过程中另起一个终端执行 jps 查看进程.此时会存在一个 CoarseGrainedExecutorBackend 进程, 负责创建及维护 Executor 对象

```
syx@syx-OptiPlex-7050:~$ jps
23426 NameNode
28114 Jps
28101 CoarseGrainedExecutorBackend
24070 ResourceManager
28008 SparkSubmit
25673 Master
24425 NodeManager
23882 SecondaryNameNode
25835 Worker
23630 DataNode
```

- Cluster 提交模式, 此模式下 Master 会随机选取一个 Worker 节点启动 Driver, 故在客户端看不到应用程序运行过程中的信息

  运行结果如下图所示:

```
syx@syx-OptiPlex-7050:~$ ~/spark-2.4.4/bin/spark-submit   --deploy-mode cluster  --master
 spark://127.0.1.1:7077   --class org.apache.spark.examples.SparkPi   ~/spark-2.4.4/examp
les/jars/spark-examples_2.11-2.4.4.jar
19/10/20 14:54:07 WARN util.Utils: Your hostname, syx-OptiPlex-7050 resolves to a loopbac
k address: 127.0.1.1; using 219.228.135.124 instead (on interface enp0s31f6)
19/10/20 14:54:07 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another addr
ess
19/10/20 14:54:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for yo
ur platform... using builtin-java classes where applicable
```

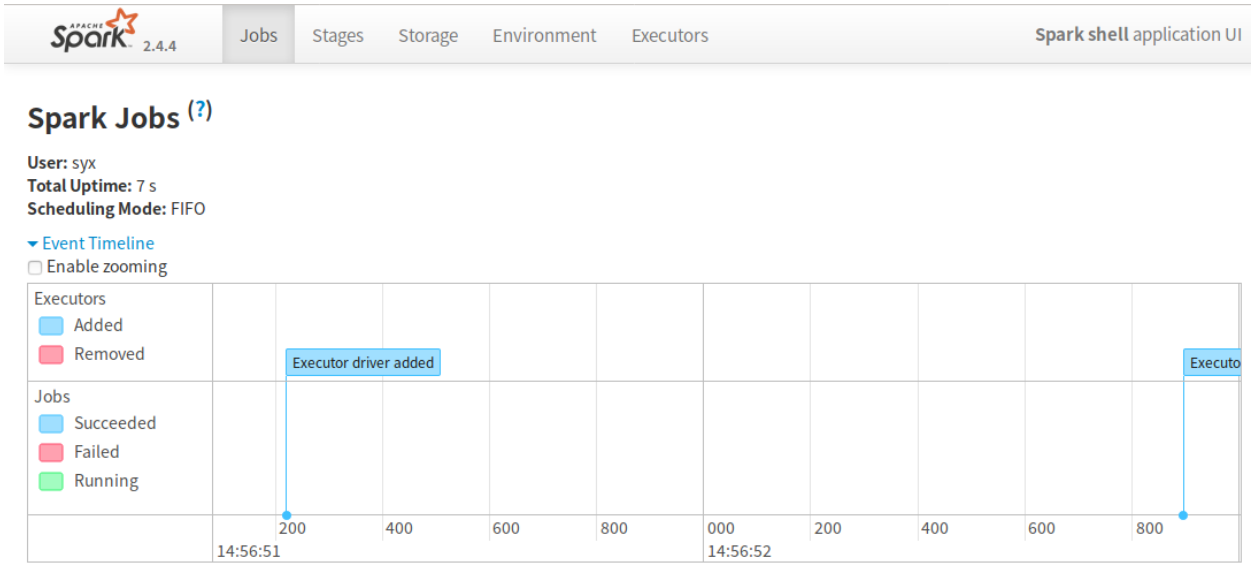
  在运行过程中另起一个终端执行 jps 查看进程.在 Cluster 提交模式下, 还可以看到一个 DriverWrapper 进程

```
syx@syx-OptiPlex-7050:~$ jps
23426 NameNode
28323 SparkSubmit
28469 CoarseGrainedExecutorBackend
24070 ResourceManager
28392 DriverWrapper
25673 Master
28537 Jps
24425 NodeManager
23882 SecondaryNameNode
25835 Worker
23630 DataNode
```

2.6 查看 Spark 程序运行信息

2.6.1 实时查看应用运行情况

- 在应用运行过程中 (如进入 Spark-Shell 之后), 访问 http://localhost:4040



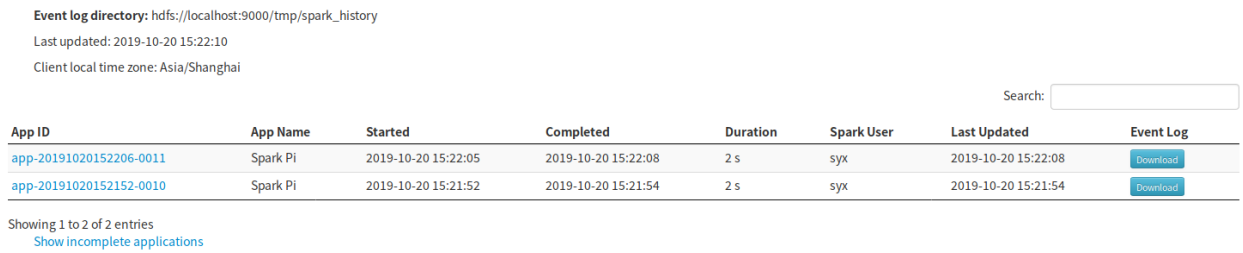
2.6.2 查看 Spark 应用程序日志

- 在提交一个应用程序后，在 ~/spark-2.4.4/work 下会出现应用程序运行日志

```
syx@syx-OptiPlex-7050:~$ ls ~/spark-2.4.4/work
app-20191017184753-0000  app-20191020144257-0000  app-20191020145409-0005
app-20191017185809-0001  app-20191020144423-0001  app-20191020145651-0006
app-20191017185941-0002  app-20191020144541-0002  driver-20191017190133-0000
app-20191017190025-0003  app-20191020145054-0003  driver-20191020145408-0000
app-20191017190134-0004  app-20191020145239-0004
```

2.6.4 查看应用历史记录

- 在应用运行结束后, 访问 http://localhost:18080



2.7 停止服务

- 停止命令

```
syx@syx-OptiPlex-7050:~$ ~/spark-2.4.4/sbin/stop-all.sh
localhost: stopping org.apache.spark.deploy.worker.Worker
stopping org.apache.spark.deploy.master.Master
syx@syx-OptiPlex-7050:~$ ~/spark-2.4.4/sbin/stop-history-server.sh
stopping org.apache.spark.deploy.history.HistoryServer
syx@syx-OptiPlex-7050:~$ jps
577 Jps
23426 NameNode
24070 ResourceManager
24425 NodeManager
23882 SecondaryNameNode
23630 DataNode
```

mistake

1. 修改 `spark-env.sh` 文件

delete the kongge

2. `java_home` is not set

```
syx@syx-OptiPlex-7050:~$ ~/spark-2.4.4/sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /home/syx/spark-2.4.4
/logs/spark-syx-org.apache.spark.deploy.master.Master-1-syx-OptiPlex-7050.out
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/syx/
spark-2.4.4/logs/spark-syx-org.apache.spark.deploy.worker.Worker-1-syx-OptiPlex-
7050.out
localhost: failed to launch: nice -n 0 /home/syx/spark-2.4.4/bin/spark-class org
.apache.spark.deploy.worker.Worker --webui-port 8081 spark://syx-OptiPlex-7050:7
077
localhost:   JAVA_HOME is not set
localhost: full log in /home/syx/spark-2.4.4/logs/spark-syx-org.apache.spark.dep
loy.worker.Worker-1-syx-OptiPlex-7050.out
syx@syx-OptiPlex-7050:~$ vi ~/spark-2.4.4/conf/spark-env.sh
syx@syx-OptiPlex-7050:~$ echo $JAVA_HOME
/usr/local/jdk1.8
```

change `~/spark-2.4.4/conf/spark-env.sh`, add one line:

`export JAVA_HOME=/usr/local/jdk1.8`

```
export SPARK_MASTER_IP=localhost
export SPARK_MASTER_PORT=7077
export JAVA_HOME=/usr/local/jdk1.8
```

3. 单机集中式部署: close the vpn to start the spark-shell by

`~/spark-2.4.4/bin/spark-shell --master local`