华东师范大学数据科学与工程学院实验报告

课程名称：分布式编程模型与系统　　**年级：**2021　　**上机实践成绩：**

指导教师：徐辰　　**姓名：**彭一珅　　**学号：**10215501412

上机实践名称： Spark 部署　　**上机实践日期：**2024.4.25

一、实验目的

学习 Spark 的部署，简单使用 Spark-Shell

查看 Spark 的运行日志，体会与 MapReduce 运行过程中日志的区别

通过系统部署理解体系架构，体会 Spark 与 MapReduce 之间的区别

二、实验任务

完成 Spark 的单机集中式部署，单机伪分布式部署以及分布式部署

在单机伪分布式和分布式部署方式下，分别以 Client 和 Cluster 提交方式来运行示例程序

三、实验环境

Ubuntu 操作系统

JDK 版本：1.8

Spark 版本：2.4.7

Hadoop 版本：2.10.1

四、实验过程

单机集中式部署

下载并配置 Spark 后，运行 Spark 应用程序

```
ubuntu@10-23-3-42:~$ mv ~/spark-2.4.7/conf/spark-env.sh.template ~/spark-2.4.7/conf/spark-env.
sh
ubuntu@10-23-3-42:~$ vim spark-env.sh
ubuntu@10-23-3-42:~$ vim ~/spark-2.4.7/conf/spark-env.sh
ubuntu@10-23-3-42:~$ ~/spark-2.4.7/bin/spark-shell --master local
24/04/25 19:04:17 WARN util.Utils: Your hostname, 10-23-3-42 resolves to a loopback address: 1
27.0.1.1; using 10.23.3.42 instead (on interface eth0)
24/04/25 19:04:17 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
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://10.23.3.42:4040
Spark context available as 'sc' (master = local, app id = local-1714043070771).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.7
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_171)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

输入 scala 代码，统计 RELEASE 文件中的单词数量，执行后打印出如下结果

```
scala> sc.textFile("file:///home/ubuntu/spark-2.4.7/RELEASE").flatMap(_.split(" ")).map((_,1))
.reduceByKey(_ + _).collect
res0: Array[(String, Int)] = Array((-Psparkr,1), (-B,1), (Spark,1), (-Pkubernetes,1), (-Pyarn,
1), (revision,1), (Build,1), (built,1), (-DzincPort=3038,1), (-Pflume,1), ((git,1), (2.6.5,1),
 (flags:,1), (-Pmesos,1), (for,1), (-Pkafka-0-8,1), (-Phadoop-provided,1), (14211a1),1), (2.4.
7,1), (Hadoop,1))

scala>
```

通过提交 jar 包运行程序，输出 pi 的近似值 3.135……

```
ms on localhost (executor driver) (2/2)
24/04/25 19:11:51 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all
completed, from pool
24/04/25 19:11:51 INFO scheduler.DAGScheduler: ResultStage 0 (reduce at SparkPi.scala:38) fini
shed in 1.072 s
24/04/25 19:11:51 INFO scheduler.DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38, too
k 1.282561 s
Pi is roughly 3.135635678178391
24/04/25 19:11:51 INFO server.AbstractConnector: Stopped Spark@3153ddfc{HTTP/1.1,[http/1.1]}{0
.0.0.0:4040}
24/04/25 19:11:51 INFO ui.SparkUI: Stopped Spark web UI at http://10.23.3.42:4040
24/04/25 19:11:51 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint st
opped!
24/04/25 19:11:51 INFO memory.MemoryStore: MemoryStore cleared
24/04/25 19:11:51 INFO storage.BlockManager: BlockManager stopped
24/04/25 19:11:51 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
24/04/25 19:11:51 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: Outp
utCommitCoordinator stopped!
24/04/25 19:11:51 INFO spark.SparkContext: Successfully stopped SparkContext
24/04/25 19:11:51 INFO util.ShutdownHookManager: Shutdown hook called
24/04/25 19:11:51 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-4825a1a0-c552-4
819-a964-0c04aa435980
24/04/25 19:11:51 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-7c47acd3-2285-4
c11-9985-5e77c42b1830
ubuntu@10-23-3-42:~$
```

在运行过程中另启动一个终端执行 jps 查看进程，已出现 SparkSubmit 进程

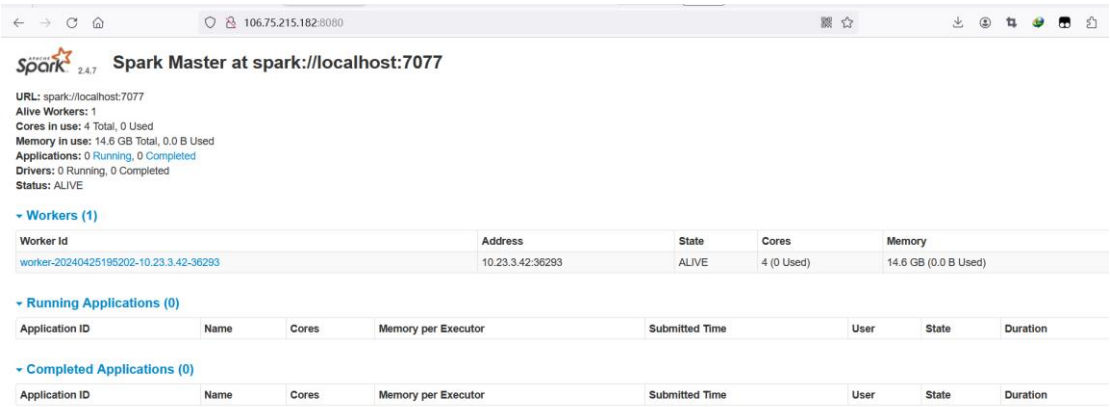
```
ubuntu@10-23-3-42:~$ jps
1422782 Jps
ubuntu@10-23-3-42:~$ jps
1422809 SparkSubmit
1422859 Jps
ubuntu@10-23-3-42:~$
```

单机伪分布式部署：

启动 spark 和 hadoopHDFS 单机伪分布式服务，查看进程

```
ubuntu@10-23-3-42:~/hadoop-2.10.1/etc/hadoop$ ~/spark-2.4.7/sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /home/ubuntu/spark-2.4.7/logs/spar
k-ubuntu-org.apache.spark.deploy.master.Master-1-10-23-3-42.out
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/ubuntu/spark-2.4.
7/logs/spark-ubuntu-org.apache.spark.deploy.worker.Worker-1-10-23-3-42.out
ubuntu@10-23-3-42:~/hadoop-2.10.1/etc/hadoop$ ~/spark-2.4.7/sbin/start-history-server.sh
starting org.apache.spark.deploy.history.HistoryServer, logging to /home/ubuntu/spark-2.4.7/l
ogs/spark-ubuntu-org.apache.spark.deploy.history.HistoryServer-1-10-23-3-42.out
ubuntu@10-23-3-42:~/hadoop-2.10.1/etc/hadoop$ jps
1425685 HistoryServer
1425132 SecondaryNameNode
1425737 Jps
1425609 Worker
1425418 Master
1424922 DataNode
1424746 NameNode
ubuntu@10-23-3-42:~/hadoop-2.10.1/etc/hadoop$
```

访问 Spark Web 页面，可以看到 Master 和 Worker



运行 Spark 应用程序，统计 RELEASE 文件中单词数量，打印出如图结果：

```
res0: Array[(String, Int)] = Array((-Psparkr,1), (-B,1), (Spark,1), (-Pkubernetes,1), (-Pyarn,
1), (revision,1), (Build,1), (built,1), (-DzincPort=3038,1), (-Pflume,1), ((git,1), (2.6.5,1),
 (flags:,1), (-Pmesos,1), (for,1), (-Pkafka-0-8,1), (-Phadoop-provided,1), (14211a1),1), (2.4.
7,1), (Hadoop,1))
```

Client 提交方式提交 jar 包，运行结果如图所示，得到 pi 的近似值 3.140……

```
24/04/25 20:08:31 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all
 completed, from pool
24/04/25 20:08:31 INFO scheduler.DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38, to
ok 2.573842 s
Pi is roughly 3.140755703778519
24/04/25 20:08:31 INFO server.AbstractConnector: Stopped Spark@2b27cc70{HTTP/1.1,[http/1.1]}{
0.0.0.0:4040}
24/04/25 20:08:31 INFO ui.SparkUI: Stopped Spark web UI at http://10.23.3.42:4040
24/04/25 20:08:31 INFO cluster.StandaloneSchedulerBackend: Shutting down all executors
24/04/25 20:08:31 INFO cluster.CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each exec
utor to shut down
24/04/25 20:08:31 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint s
topped!
24/04/25 20:08:31 INFO memory.MemoryStore: MemoryStore cleared
24/04/25 20:08:31 INFO storage.BlockManager: BlockManager stopped
24/04/25 20:08:31 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
24/04/25 20:08:31 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: Out
putCommitCoordinator stopped!
24/04/25 20:08:31 INFO spark.SparkContext: Successfully stopped SparkContext
24/04/25 20:08:31 INFO util.ShutdownHookManager: Shutdown hook called
24/04/25 20:08:31 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-a703e8ef-825c-
48aa-a2b6-0165c03530e0
24/04/25 20:08:31 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-b6e42355-6fb7-
4615-a1c3-9346ed22890c
ubuntu@10-23-3-42:~/hadoop-2.10.1/etc/hadoop$
```

出现 CoarseGrainedExecutorBackend 进程，负责创建及维护 Executor 对象

```
ubuntu@10-23-3-42:~$ jps
1425685 HistoryServer
1426820 Jps
1426710 SparkSubmit
1425132 SecondaryNameNode
1426799 CoarseGrainedExecutorBackend
1425609 Worker
1425418 Master
1424922 DataNode
1424746 NameNode
```

Cluster 提交方式，Master 随机选取一个 Worker 节点启动 Driver，因此看不到运行过程

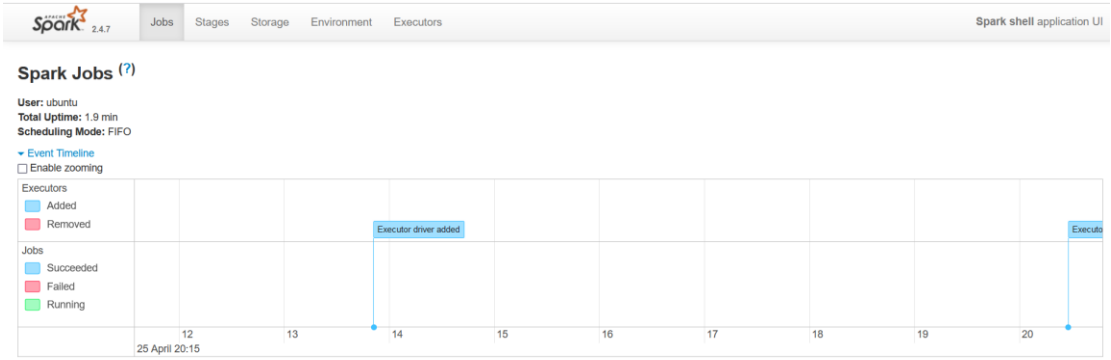
的信息

```
ubuntu@10-23-3-42:~$ ~/spark-2.4.7/bin/spark-submit --deploy-mode cluster --master spark://lo
calhost:7077 --class org.apache.spark.examples.SparkPi ~/spark-2.4.7/examples/jars/spark-exam
ples_2.11-2.4.7.jar
24/04/25 20:13:03 WARN util.Utils: Your hostname, 10-23-3-42 resolves to a loopback address:
127.0.1.1; using 10.23.3.42 instead (on interface eth0)
24/04/25 20:13:03 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
```

使用 jps 查看进程，可以看到 DriverWrapper 进程

```
ubuntu@10-23-3-42:~$ jps
1425685 HistoryServer
1427253 SparkSubmit
1427331 DriverWrapper
1425132 SecondaryNameNode
1425609 Worker
1427355 Jps
1425418 Master
1424922 DataNode
1424746 NameNode
```

查看 Spark 应用程序日志：



查看运行历史记录：

Completed Applications (4)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20240425201309-0003	Spark Pi	3	1024.0 MB	2024/04/25 20:13:09	ubuntu	FINISHED	7 s
app-20240425200824-0002	Spark Pi	4	1024.0 MB	2024/04/25 20:08:24	ubuntu	FINISHED	7 s
app-20240425195733-0001	Spark shell	4	1024.0 MB	2024/04/25 19:57:33	ubuntu	FINISHED	9.5 min
app-20240425195444-0000	Spark shell	4	1024.0 MB	2024/04/25 19:54:44	ubuntu	FINISHED	2.5 min

五、实验总结

在本次实验中学习了 Spark 的部署，并运行了示例程序，对 Spark 的体系架构有了初步的理解。

华东师范大学数据科学与工程学院实验报告

课程名称： 分布式编程模型与系统	**年级：** 2021	**上机实践成绩：**
指导教师： 徐辰	**姓名：** 彭一珅	**学号：** 10215501412
上机实践名称： Spark 编程	**上机实践日期：**	2024.5.9

一、实验目的

学习编写简单的基于 RDD API 的 Spark 程序

掌握在 IDEA 中调试 Spark 相关程序，以及在单机伪分布式、分布式部署方式下运行 Spark 相关程序的方法

二、实验任务

完成 WordCount 示例程序的编写

在单机伪分布式和分布式部署方式下运行 WordCount 示例程序

三、实验环境

操作系统：ubuntu20.04

Jdk 版本：1.8

Hadoop 版本：2.10.1

Spark 版本：2.4.7

Scala 版本：2.11.12

四、实验过程

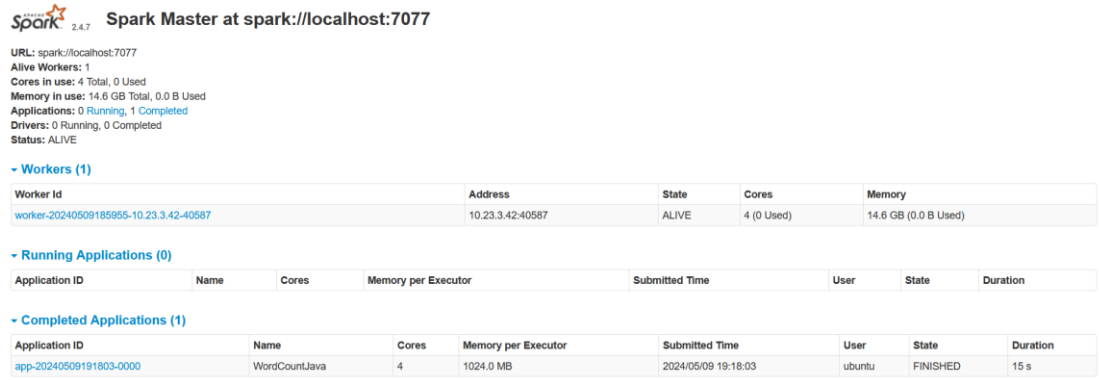
从实验网站下载 SparkDemo 项目资源，打包为 jar 包并在伪分布式系统下运行

```
ubuntu@10-23-3-42:~$ ~/hadoop-2.10.1/bin/hdfs dfs -cp ./input/pd.test ./spark_input
cp: `./input/pd.test': No such file or directory
ubuntu@10-23-3-42:~$ ~/hadoop-2.10.1/bin/hdfs dfs -mkdir input
ubuntu@10-23-3-42:~$ ~/hadoop-2.10.1/bin/hdfs dfs -put pd.test input/
ubuntu@10-23-3-42:~$ ~/hadoop-2.10.1/bin/hdfs dfs -cp ./input/pd.test ./spark_input
ubuntu@10-23-3-42:~$ ~/spark-2.4.7/bin/spark-submit --master spark://localhost:7077 --class c
n.edu.ecnu.spark.example.java.wordcount.WordCount /home/ubuntu/spark-2.4.7/myApp/RddWordCount
Java.jar hdfs://localhost:9000/user/ubuntu/spark_input hdfs://localhost:9000/user/ubuntu/spar
k_output
24/05/09 19:17:59 WARN util.Utils: Your hostname, 10-23-3-42 resolves to a loopback address:
127.0.1.1; using 10.23.3.42 instead (on interface eth0)
24/05/09 19:17:59 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/05/09 19:18:00 INFO spark.SparkContext: Running Spark version 2.4.7
24/05/09 19:18:00 INFO spark.SparkContext: Submitted application: WordCountJava
24/05/09 19:18:00 INFO spark.SecurityManager: Changing view acls to: ubuntu
24/05/09 19:18:00 INFO spark.SecurityManager: Changing modify acls to: ubuntu
24/05/09 19:18:00 INFO spark.SecurityManager: Changing view acls groups to:
```

运行结束输出

```
24/05/09 19:18:18 INFO scheduler.DAGScheduler: ResultStage 1 (runJob at SparkHadoopWriter.sca
la:78) finished in 3.522 s
24/05/09 19:18:18 INFO scheduler.DAGScheduler: Job 0 finished: runJob at SparkHadoopWriter.sc
ala:78, took 8.861479 s
24/05/09 19:18:18 INFO io.SparkHadoopWriter: Job job_20240509191809_0007 committed.
24/05/09 19:18:18 INFO server.AbstractConnector: Stopped Spark@79f227a9{HTTP/1.1,[http/1.1]}{
0.0.0.0:4040}
24/05/09 19:18:18 INFO ui.SparkUI: Stopped Spark web UI at http://10.23.3.42:4040
24/05/09 19:18:18 INFO cluster.StandaloneSchedulerBackend: Shutting down all executors
24/05/09 19:18:18 INFO cluster.CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each exec
utor to shut down
24/05/09 19:18:18 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint s
topped!
24/05/09 19:18:18 INFO memory.MemoryStore: MemoryStore cleared
24/05/09 19:18:18 INFO storage.BlockManager: BlockManager stopped
24/05/09 19:18:18 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
24/05/09 19:18:18 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: Out
putCommitCoordinator stopped!
24/05/09 19:18:18 INFO spark.SparkContext: Successfully stopped SparkContext
24/05/09 19:18:18 INFO util.ShutdownHookManager: Shutdown hook called
24/05/09 19:18:18 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-bb55c17c-e5ec-
4666-82c6-01c19629ddb0
24/05/09 19:18:18 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-60ceb611-89f9-
42ca-bcde-17932e7b5a32
ubuntu@10-23-3-42:~$
```

查看运行记录：



五、实验总结

在本次实验中，进行了伪分布式的 wordcount 示例程序运行，在 IDEA 中调试 Spark 程序，并在虚拟机上进行运行。

华东师范大学数据科学与工程学院实验报告

课程名称：分布式编程模型与系统　**年级：**2021　**上机实践成绩：**

指导教师：徐辰　**姓名：**彭一珅　**学号：**10215501412

上机实践名称：基于 Yarn 部署 Spark　**上机实践日期：**5.16

一、实验目的

通过基于 Yarn 部署 Spark，深入理解 Yarn 的作用，体会“一个平台，多个框架”

二、实验任务

完成 Spark2.4.7on Yarn 的单机伪分布式以及分布式部署

在两种部署方式下以不同的提交方式运行词频统计及 pi 近似值计算两个示例程序

三、实验环境

操作系统：ubuntu20.04

Jdk 版本：1.8

Hadoop 版本：2.10.1

Spark 版本：2.4.7

四、实验过程

单机伪分布式部署

运行 hadoop 和 yarn 之后，启动如下进程

```
ubuntu@10-23-3-42:~$ jps
2050589 NameNode
2042221 ResourceManager
2051002 SecondaryNameNode
2045945 JobHistoryServer
2050744 DataNode
2051607 Jps
2042358 NodeManager
2051425 HistoryServer
ubuntu@10-23-3-42:~$
```

在 spark-shell 中运行应用程序，统计 RELEASE 文件中单词数量

```
scala> sc.textFile("spark_input/RELEASE").flatMap(_.split(" ")).map((_,1)).reduceByKey(_ + _).colle
ct
res1: Array[(String, Int)] = Array((-Psparkr,1), (Build,1), (built,1), (-Pflume,1), ((git,1), (-Pme
sos,1), (-Phadoop-provided,1), (14211a1),1), (-B,1), (Spark,1), (-Pkubernetes,1), (-Pyarn,1), (revi
sion,1), (-DzincPort=3038,1), (2.6.5,1), (flags:,1), (for,1), (-Pkafka-0-8,1), (2.4.7,1), (Hadoop,1
))
```

通过 client 方式提交 jar 包，出现一个 ExecutorLauncher 进程和若干个 CoarseGrainedExecutorBackend 进程

```
ubuntu@10-23-3-42:~$ jps
2050589 NameNode
2042221 ResourceManager
2053548 Jps
2051002 SecondaryNameNode
2045945 JobHistoryServer
2050744 DataNode
2042358 NodeManager
2053350 ExecutorLauncher
2053539 CoarseGrainedExecutorBackend
2053186 SparkSubmit
2051425 HistoryServer
```

Cluster 提交方式

```
24/05/16 19:52:01 INFO yarn.Client:
         client token: N/A
         diagnostics: AM container is launched, waiting for AM container to Register with RM
         ApplicationMaster host: N/A
         ApplicationMaster RPC port: -1
         queue: default
         start time: 1715860320314
         final status: UNDEFINED
         tracking URL: http://10-23-142-76:8088/proxy/application_1715859333061_0002/
         user: ubuntu
```

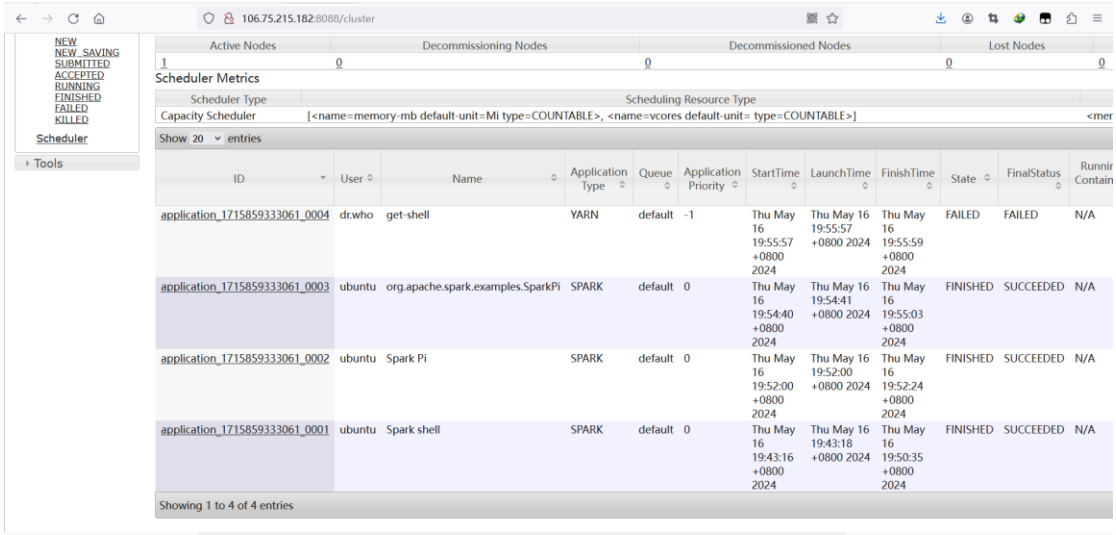
存在一个 ApplicationMaster 进程

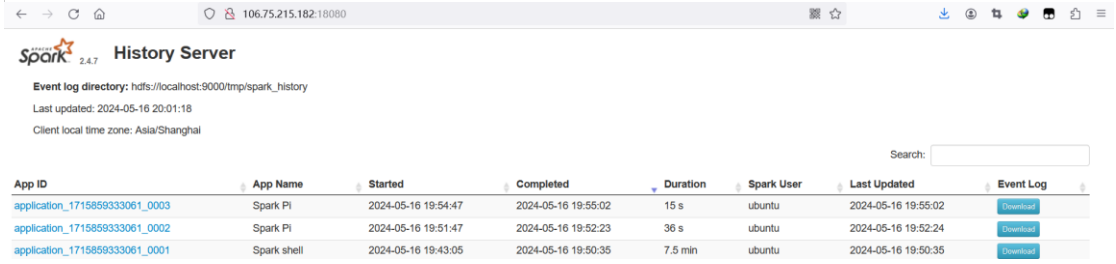
```
ubuntu@10-23-3-42:~$ jps
2054094 Jps
2050589 NameNode
2042221 ResourceManager
2051002 SecondaryNameNode
2045945 JobHistoryServer
2050744 DataNode
2054071 ApplicationMaster
2042358 NodeManager
2053921 SparkSubmit
2051425 HistoryServer
```

运行结果如下

```
24/05/16 19:55:04 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: 10.23.3.42
         ApplicationMaster RPC port: 45165
         queue: default
         start time: 1715860480848
         final status: SUCCEEDED
         tracking URL: http://10-23-142-76:8088/proxy/application_1715859333061_0003/
         user: ubuntu
24/05/16 19:55:04 INFO util.ShutdownHookManager: Shutdown hook called
24/05/16 19:55:04 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-a0057df3-7816-4bd9-9
62b-0fbc956ab5b8
24/05/16 19:55:04 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-2d264b60-4a01-46ad-b
9a6-48bb6a1905eb
```

查看 Spark 程序运行信息





五、实验总结

在本次实验中，在单机集中式、单机伪分布式进行了 Yarn 部署 Spark 过程，深入理解了 Yarn 的作用，并在两种部署方式下以不同的提交方式运行了词频统计和 Pi 近似值计算的程序。