华东师范大学数据科学与工程学院实验报告

课程名称：分布式模型与编程	**年级：2017**	**上机实践成绩：**
指导教师：徐辰	**姓名：熊双宇**	**学号：10174102103**
上机实践名称：Hadoop部署与编程		**上机实践日期：2019.9.20-2019.10.17**
上机实践编号：实验1	**组号：11**	**上机实践时间：18:00-19:30**

一. 实验目的

1. 学习Hadoop v1和Hadoop v2的部署，理解单机集中式、单机伪分布式的区别;
2. 学会通过系统日志查找部署和编程中遇到的错误;
3. 通过系统部署理解Hadoop的体系架构，以及Hadoop v1和Hadoop v2之间的差异，初步体会Yarn的作用;
4. 学习基于Hadoop v2 API的编程，包括HDFS和MapReduce;
5. 了解Hadoop Streaming编程

二. 实验任务

1. HDFS 1.0部署【第3周】: 单机伪分布式（在个人用户下独立完成）、分布式（多位同学新建一个相同的用户，例如ecnu，协作完成）
2. MapReduce 1.0部署【第3周】: 单机集中式、单机伪分布式（在个人用户下独立完成）、分布式（多位同学新建一个相同的用户，例如ecnu，协作完成）
3. HDFS 2.0部署【第4周】: 单机伪分布式（在个人用户下独立完成）、分布式（多位同学新建一个相同的用户，例如ecnu，协作完成）
4. MapReduce 2.0部署【第4周】: 单机集中式、单机伪分布式（在个人用户下独立完成）、分布式（多位同学新建一个相同的用户，例如ecnu，协作完成）
5. HDFS编程【第5周】
6. MapReduce编程【第6周】
7. Hadoop Streaming编程【第6周】： 该内容选做

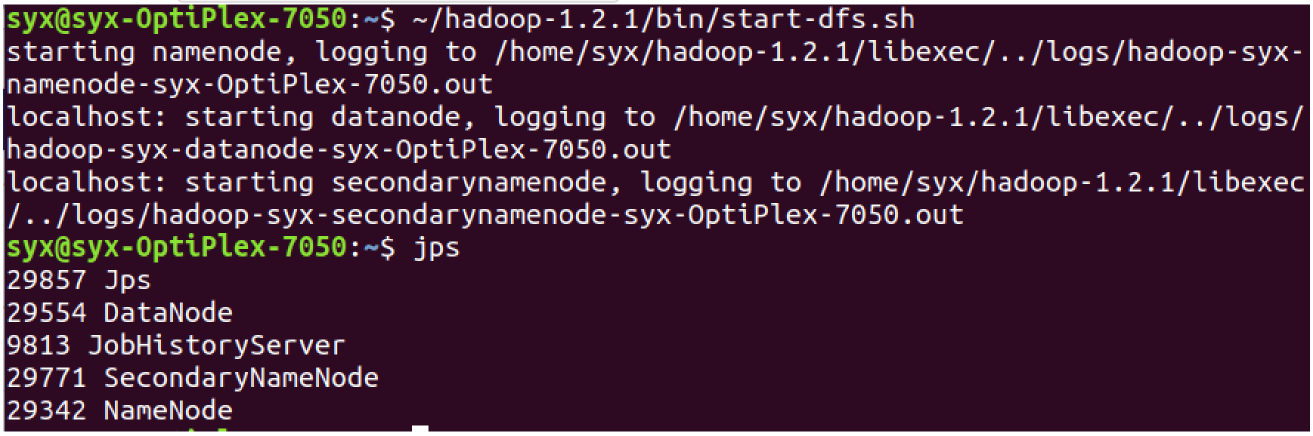
三. 使用环境

1. Ubuntu18.04
2. hadoop-1.2.1
3. hadoop-2.9.2

四. 实验过程

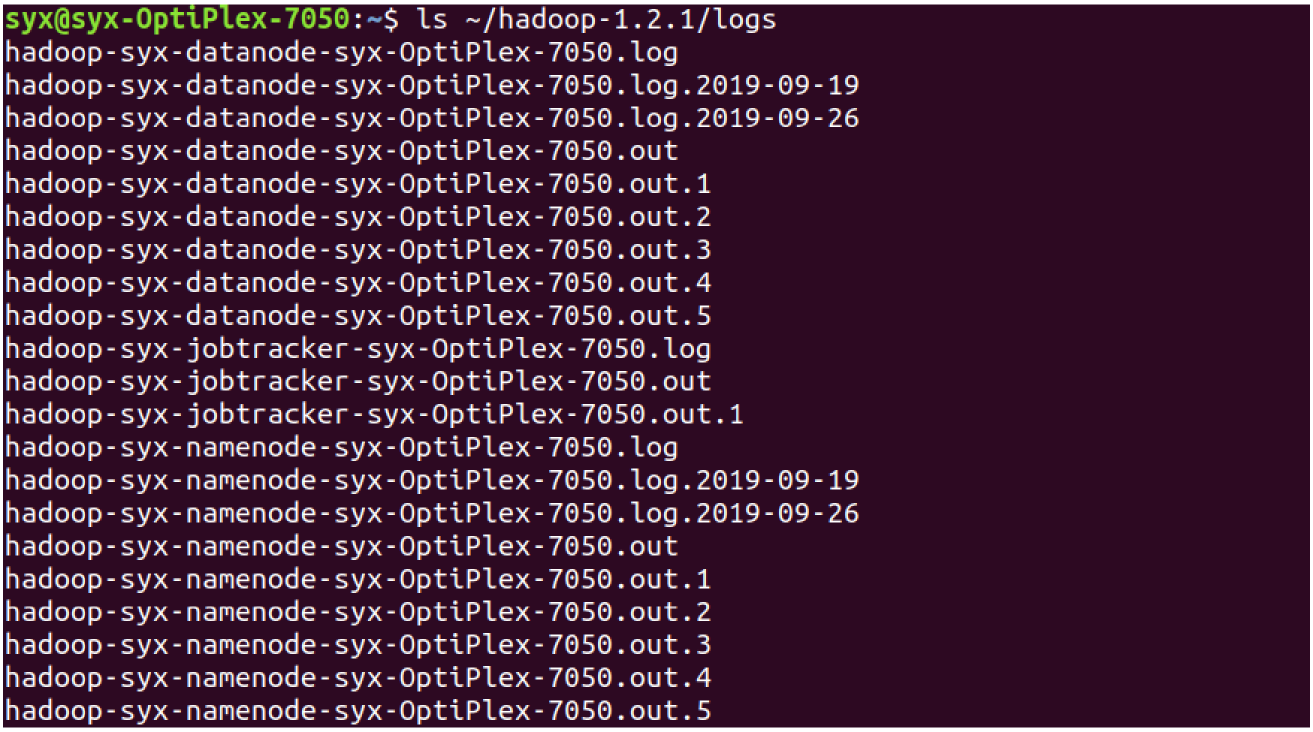
1. HDFS v1部署

1.1 启动HDFS: `~/hadoop-1.2.1/bin/start-dfs.sh`, 使用jps查看启动后的进程



```
syx@syx-OptiPlex-7050:~$ ~/hadoop-1.2.1/bin/start-dfs.sh
starting namenode, logging to /home/syx/hadoop-1.2.1/libexec/../logs/hadoop-syx-
namenode-syx-OptiPlex-7050.out
localhost: starting datanode, logging to /home/syx/hadoop-1.2.1/libexec/../logs/
hadoop-syx-datanode-syx-OptiPlex-7050.out
localhost: starting secondarynamenode, logging to /home/syx/hadoop-1.2.1/libexec
/../logs/hadoop-syx-secondarynamenode-syx-OptiPlex-7050.out
syx@syx-OptiPlex-7050:~$ jps
29857 Jps
29554 DataNode
9813 JobHistoryServer
29771 SecondaryNameNode
29342 NameNode
```

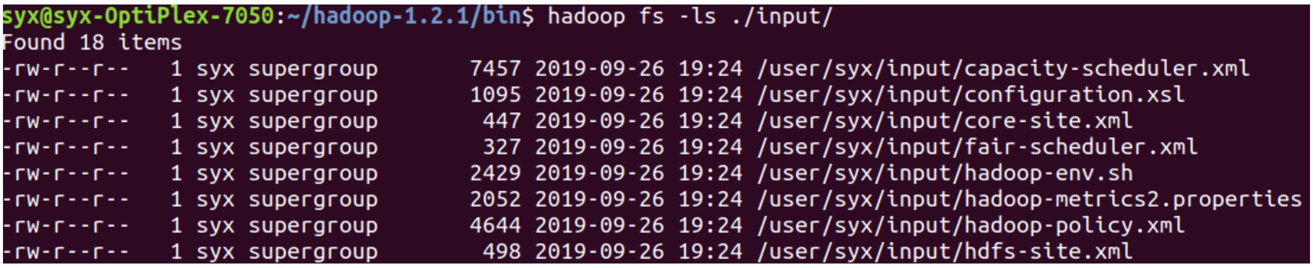
1.2 查看HDFS服务信息



```
syx@syx-OptiPlex-7050:~$ ls ~/hadoop-1.2.1/logs
hadoop-syx-datanode-syx-OptiPlex-7050.log
hadoop-syx-datanode-syx-OptiPlex-7050.log.2019-09-19
hadoop-syx-datanode-syx-OptiPlex-7050.log.2019-09-26
hadoop-syx-datanode-syx-OptiPlex-7050.out
hadoop-syx-datanode-syx-OptiPlex-7050.out.1
hadoop-syx-datanode-syx-OptiPlex-7050.out.2
hadoop-syx-datanode-syx-OptiPlex-7050.out.3
hadoop-syx-datanode-syx-OptiPlex-7050.out.4
hadoop-syx-datanode-syx-OptiPlex-7050.out.5
hadoop-syx-jobtracker-syx-OptiPlex-7050.log
hadoop-syx-jobtracker-syx-OptiPlex-7050.out
hadoop-syx-jobtracker-syx-OptiPlex-7050.out.1
hadoop-syx-namenode-syx-OptiPlex-7050.log
hadoop-syx-namenode-syx-OptiPlex-7050.log.2019-09-19
hadoop-syx-namenode-syx-OptiPlex-7050.log.2019-09-26
hadoop-syx-namenode-syx-OptiPlex-7050.out
hadoop-syx-namenode-syx-OptiPlex-7050.out.1
hadoop-syx-namenode-syx-OptiPlex-7050.out.2
hadoop-syx-namenode-syx-OptiPlex-7050.out.3
hadoop-syx-namenode-syx-OptiPlex-7050.out.4
hadoop-syx-namenode-syx-OptiPlex-7050.out.5
```

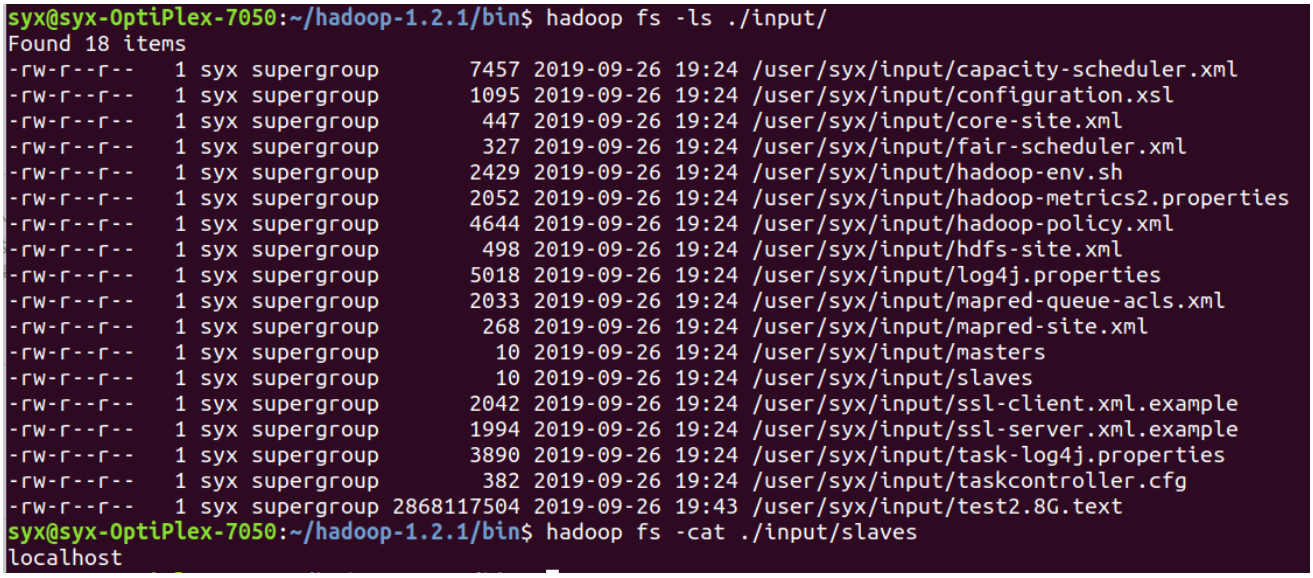
1.3 常用的HDFS Shell命令

1.3.1 directory



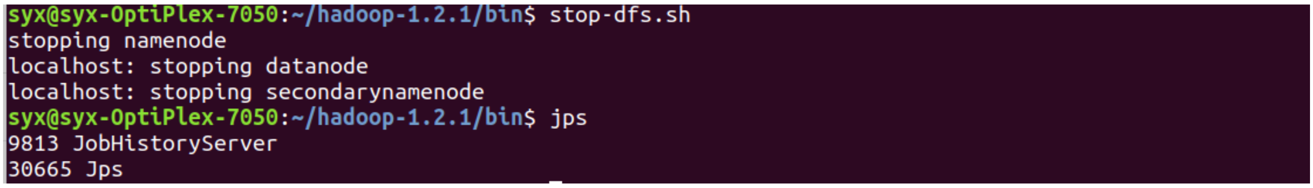
```
syx@syx-OptiPlex-7050:~/hadoop-1.2.1/bin$ hadoop fs -ls ./input/
Found 18 items
-rw-r--r--   1 syx supergroup       7457 2019-09-26 19:24 /user/syx/input/capacity-scheduler.xml
-rw-r--r--   1 syx supergroup       1095 2019-09-26 19:24 /user/syx/input/configuration.xsl
-rw-r--r--   1 syx supergroup        447 2019-09-26 19:24 /user/syx/input/core-site.xml
-rw-r--r--   1 syx supergroup        327 2019-09-26 19:24 /user/syx/input/fair-scheduler.xml
-rw-r--r--   1 syx supergroup       2429 2019-09-26 19:24 /user/syx/input/hadoop-env.sh
-rw-r--r--   1 syx supergroup       2052 2019-09-26 19:24 /user/syx/input/hadoop-metrics2.properties
-rw-r--r--   1 syx supergroup       4644 2019-09-26 19:24 /user/syx/input/hadoop-policy.xml
-rw-r--r--   1 syx supergroup        498 2019-09-26 19:24 /user/syx/input/hdfs-site.xml
```

1.3.2 file



```
syx@syx-OptiPlex-7050:~/hadoop-1.2.1/bin$ hadoop fs -ls ./input/
Found 18 items
-rw-r--r--   1 syx supergroup       7457 2019-09-26 19:24 /user/syx/input/capacity-scheduler.xml
-rw-r--r--   1 syx supergroup       1095 2019-09-26 19:24 /user/syx/input/configuration.xsl
-rw-r--r--   1 syx supergroup        447 2019-09-26 19:24 /user/syx/input/core-site.xml
-rw-r--r--   1 syx supergroup        327 2019-09-26 19:24 /user/syx/input/fair-scheduler.xml
-rw-r--r--   1 syx supergroup       2429 2019-09-26 19:24 /user/syx/input/hadoop-env.sh
-rw-r--r--   1 syx supergroup       2052 2019-09-26 19:24 /user/syx/input/hadoop-metrics2.properties
-rw-r--r--   1 syx supergroup       4644 2019-09-26 19:24 /user/syx/input/hadoop-policy.xml
-rw-r--r--   1 syx supergroup        498 2019-09-26 19:24 /user/syx/input/hdfs-site.xml
-rw-r--r--   1 syx supergroup       5018 2019-09-26 19:24 /user/syx/input/log4j.properties
-rw-r--r--   1 syx supergroup       2033 2019-09-26 19:24 /user/syx/input/mapred-queue-acls.xml
-rw-r--r--   1 syx supergroup        268 2019-09-26 19:24 /user/syx/input/mapred-site.xml
-rw-r--r--   1 syx supergroup         10 2019-09-26 19:24 /user/syx/input/masters
-rw-r--r--   1 syx supergroup         10 2019-09-26 19:24 /user/syx/input/slaves
-rw-r--r--   1 syx supergroup       2042 2019-09-26 19:24 /user/syx/input/ssl-client.xml.example
-rw-r--r--   1 syx supergroup       1994 2019-09-26 19:24 /user/syx/input/ssl-server.xml.example
-rw-r--r--   1 syx supergroup       3890 2019-09-26 19:24 /user/syx/input/task-log4j.properties
-rw-r--r--   1 syx supergroup        382 2019-09-26 19:24 /user/syx/input/taskcontroller.cfg
-rw-r--r--   1 syx supergroup 2868117504 2019-09-26 19:43 /user/syx/input/test2.8G.text
syx@syx-OptiPlex-7050:~/hadoop-1.2.1/bin$ hadoop fs -cat ./input/slaves
localhost
```

1.4 停止HDFS服务



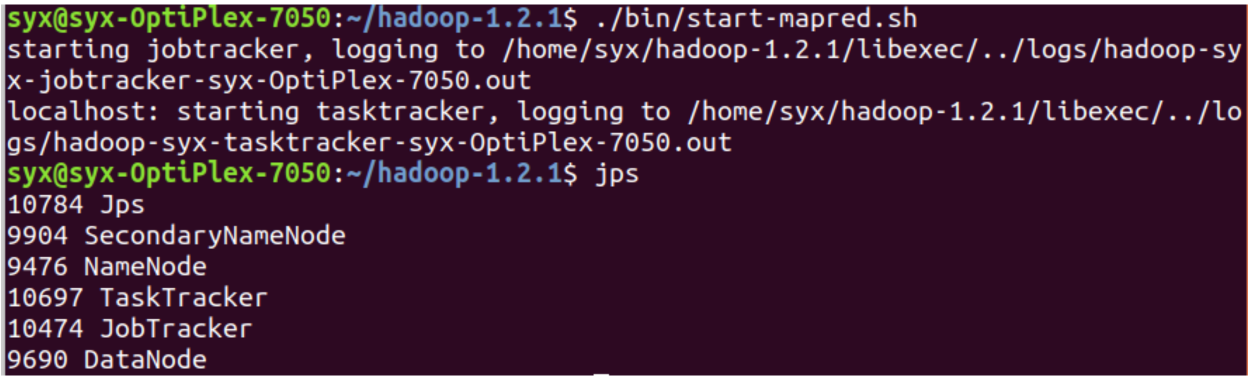
```
syx@syx-OptiPlex-7050:~/hadoop-1.2.1/bin$ stop-dfs.sh
stopping namenode
localhost: stopping datanode
localhost: stopping secondarynamenode
syx@syx-OptiPlex-7050:~/hadoop-1.2.1/bin$ jps
9813 JobHistoryServer
30665 Jps
```

2. MapReduce v1部署

2.1单机集中式部署

2.1.1 启动MapReduce服务

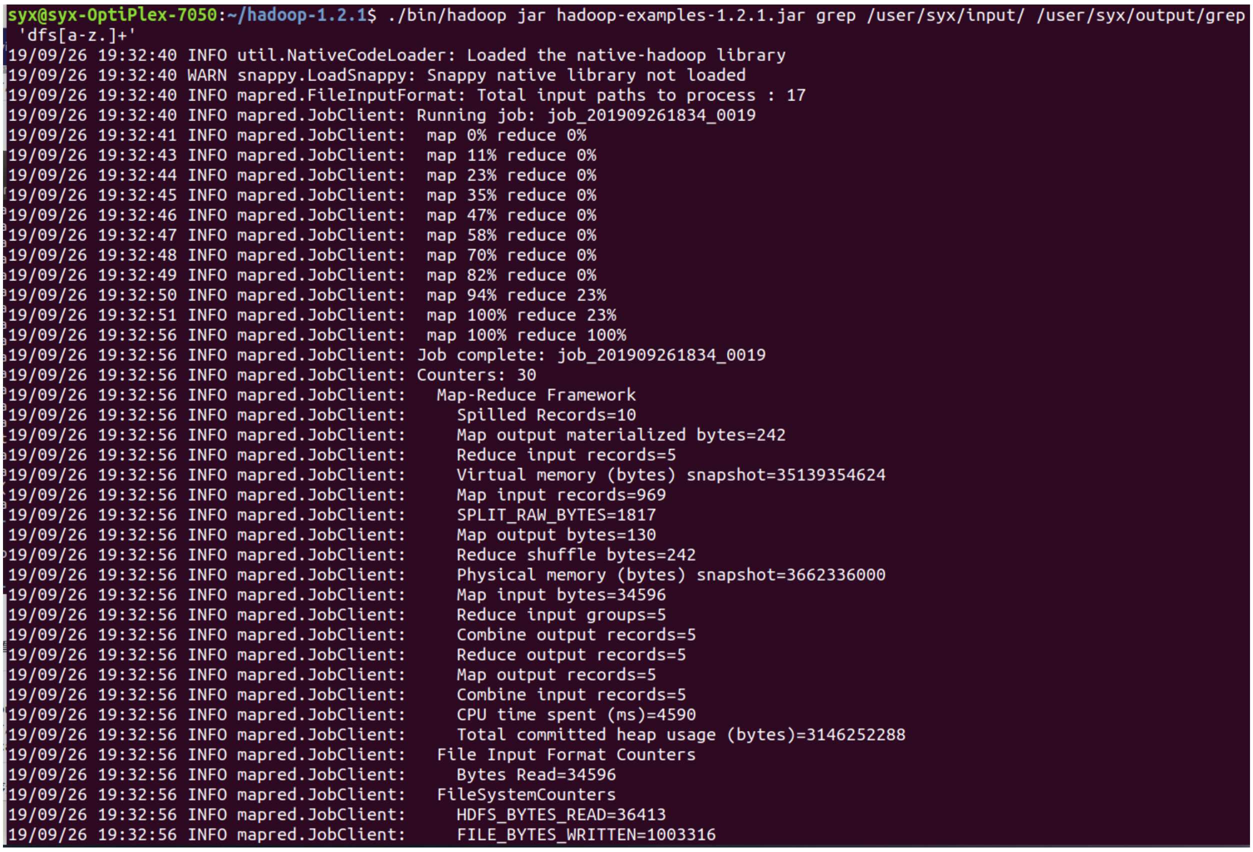
- 查看进程，验证是否成功启动服务



```
syx@syx-OptiPlex-7050:~/hadoop-1.2.1$ ./bin/start-mapred.sh
starting jobtracker, logging to /home/syx/hadoop-1.2.1/libexec/../logs/hadoop-sy
x-jobtracker-syx-OptiPlex-7050.out
localhost: starting tasktracker, logging to /home/syx/hadoop-1.2.1/libexec/../lo
gs/hadoop-syx-tasktracker-syx-OptiPlex-7050.out
syx@syx-OptiPlex-7050:~/hadoop-1.2.1$ jps
10784 Jps
9904 SecondaryNameNode
9476 NameNode
10697 TaskTracker
10474 JobTracker
9690 DataNode
```

2.1.2 提交MapReduce应用程序

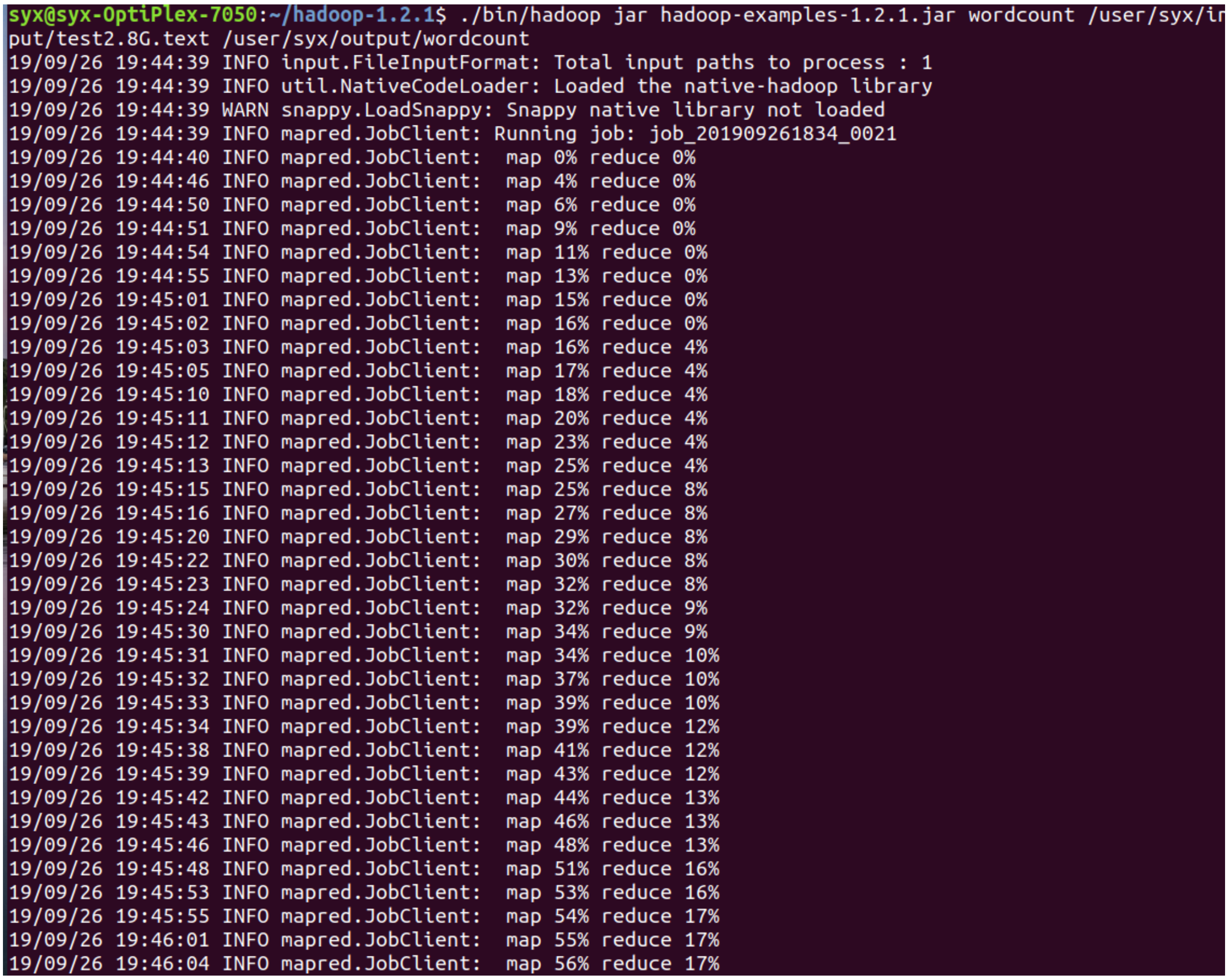
- 提交jar命令并查看运行结果, 运行grep示例:



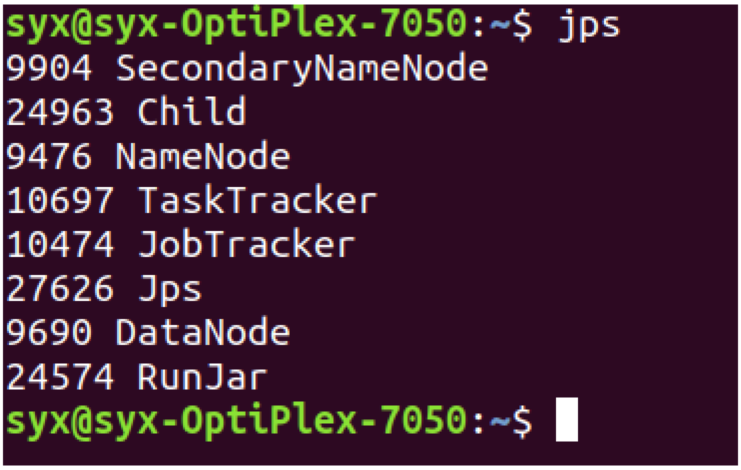
```
syx@syx-OptiPlex-7050:~/hadoop-1.2.1$ ./bin/hadoop jar hadoop-examples-1.2.1.jar grep /user/syx/input/ /user/syx/output/grep
 'dfs[a-z.]+'
19/09/26 19:32:40 INFO util.NativeCodeLoader: Loaded the native-hadoop library
19/09/26 19:32:40 WARN snappy.LoadSnappy: Snappy native library not loaded
19/09/26 19:32:40 INFO mapred.FileInputFormat: Total input paths to process : 17
19/09/26 19:32:40 INFO mapred.JobClient: Running job: job_201909261834_0019
19/09/26 19:32:41 INFO mapred.JobClient:  map 0% reduce 0%
19/09/26 19:32:43 INFO mapred.JobClient:  map 11% reduce 0%
19/09/26 19:32:44 INFO mapred.JobClient:  map 23% reduce 0%
19/09/26 19:32:45 INFO mapred.JobClient:  map 35% reduce 0%
19/09/26 19:32:46 INFO mapred.JobClient:  map 47% reduce 0%
19/09/26 19:32:47 INFO mapred.JobClient:  map 58% reduce 0%
19/09/26 19:32:48 INFO mapred.JobClient:  map 70% reduce 0%
19/09/26 19:32:49 INFO mapred.JobClient:  map 82% reduce 0%
19/09/26 19:32:50 INFO mapred.JobClient:  map 94% reduce 23%
19/09/26 19:32:51 INFO mapred.JobClient:  map 100% reduce 23%
19/09/26 19:32:56 INFO mapred.JobClient:  map 100% reduce 100%
19/09/26 19:32:56 INFO mapred.JobClient: Job complete: job_201909261834_0019
19/09/26 19:32:56 INFO mapred.JobClient: Counters: 30
19/09/26 19:32:56 INFO mapred.JobClient:   Map-Reduce Framework
19/09/26 19:32:56 INFO mapred.JobClient:     Spilled Records=10
19/09/26 19:32:56 INFO mapred.JobClient:     Map output materialized bytes=242
19/09/26 19:32:56 INFO mapred.JobClient:     Reduce input records=5
19/09/26 19:32:56 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=35139354624
19/09/26 19:32:56 INFO mapred.JobClient:     Map input records=969
19/09/26 19:32:56 INFO mapred.JobClient:     SPLIT_RAW_BYTES=1817
19/09/26 19:32:56 INFO mapred.JobClient:     Map output bytes=130
19/09/26 19:32:56 INFO mapred.JobClient:     Reduce shuffle bytes=242
19/09/26 19:32:56 INFO mapred.JobClient:     Physical memory (bytes) snapshot=3662336000
19/09/26 19:32:56 INFO mapred.JobClient:     Map input bytes=34596
19/09/26 19:32:56 INFO mapred.JobClient:     Reduce input groups=5
19/09/26 19:32:56 INFO mapred.JobClient:     Combine output records=5
19/09/26 19:32:56 INFO mapred.JobClient:     Reduce output records=5
19/09/26 19:32:56 INFO mapred.JobClient:     Map output records=5
19/09/26 19:32:56 INFO mapred.JobClient:     Combine input records=5
19/09/26 19:32:56 INFO mapred.JobClient:     CPU time spent (ms)=4590
19/09/26 19:32:56 INFO mapred.JobClient:     Total committed heap usage (bytes)=3146252288
19/09/26 19:32:56 INFO mapred.JobClient:   File Input Format Counters
19/09/26 19:32:56 INFO mapred.JobClient:     Bytes Read=34596
19/09/26 19:32:56 INFO mapred.JobClient:   FileSystemCounters
19/09/26 19:32:56 INFO mapred.JobClient:     HDFS_BYTES_READ=36413
19/09/26 19:32:56 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=1003316
```

2.1.3 查看运行过程中的进程

- 运行 wordcount 示例，并且查看系统执行该任务过程中启动的进程：



```
syx@syx-OptiPlex-7050:~/hadoop-1.2.1$ ./bin/hadoop jar hadoop-examples-1.2.1.jar wordcount /user/syx/in
put/test2.8G.text /user/syx/output/wordcount
19/09/26 19:44:39 INFO input.FileInputFormat: Total input paths to process : 1
19/09/26 19:44:39 INFO util.NativeCodeLoader: Loaded the native-hadoop library
19/09/26 19:44:39 WARN snappy.LoadSnappy: Snappy native library not loaded
19/09/26 19:44:39 INFO mapred.JobClient: Running job: job_201909261834_0021
19/09/26 19:44:40 INFO mapred.JobClient:  map 0% reduce 0%
19/09/26 19:44:46 INFO mapred.JobClient:  map 4% reduce 0%
19/09/26 19:44:50 INFO mapred.JobClient:  map 6% reduce 0%
19/09/26 19:44:51 INFO mapred.JobClient:  map 9% reduce 0%
19/09/26 19:44:54 INFO mapred.JobClient:  map 11% reduce 0%
19/09/26 19:44:55 INFO mapred.JobClient:  map 13% reduce 0%
19/09/26 19:45:01 INFO mapred.JobClient:  map 15% reduce 0%
19/09/26 19:45:02 INFO mapred.JobClient:  map 16% reduce 0%
19/09/26 19:45:03 INFO mapred.JobClient:  map 16% reduce 4%
19/09/26 19:45:05 INFO mapred.JobClient:  map 17% reduce 4%
19/09/26 19:45:10 INFO mapred.JobClient:  map 18% reduce 4%
19/09/26 19:45:11 INFO mapred.JobClient:  map 20% reduce 4%
19/09/26 19:45:12 INFO mapred.JobClient:  map 23% reduce 4%
19/09/26 19:45:13 INFO mapred.JobClient:  map 25% reduce 4%
19/09/26 19:45:15 INFO mapred.JobClient:  map 25% reduce 8%
19/09/26 19:45:16 INFO mapred.JobClient:  map 27% reduce 8%
19/09/26 19:45:20 INFO mapred.JobClient:  map 29% reduce 8%
19/09/26 19:45:22 INFO mapred.JobClient:  map 30% reduce 8%
19/09/26 19:45:23 INFO mapred.JobClient:  map 32% reduce 8%
19/09/26 19:45:24 INFO mapred.JobClient:  map 32% reduce 9%
19/09/26 19:45:30 INFO mapred.JobClient:  map 34% reduce 9%
19/09/26 19:45:31 INFO mapred.JobClient:  map 34% reduce 10%
19/09/26 19:45:32 INFO mapred.JobClient:  map 37% reduce 10%
19/09/26 19:45:33 INFO mapred.JobClient:  map 39% reduce 10%
19/09/26 19:45:34 INFO mapred.JobClient:  map 39% reduce 12%
19/09/26 19:45:38 INFO mapred.JobClient:  map 41% reduce 12%
19/09/26 19:45:39 INFO mapred.JobClient:  map 43% reduce 12%
19/09/26 19:45:42 INFO mapred.JobClient:  map 44% reduce 13%
19/09/26 19:45:43 INFO mapred.JobClient:  map 46% reduce 13%
19/09/26 19:45:46 INFO mapred.JobClient:  map 48% reduce 13%
19/09/26 19:45:48 INFO mapred.JobClient:  map 51% reduce 16%
19/09/26 19:45:53 INFO mapred.JobClient:  map 53% reduce 16%
19/09/26 19:45:55 INFO mapred.JobClient:  map 54% reduce 17%
19/09/26 19:46:01 INFO mapred.JobClient:  map 55% reduce 17%
19/09/26 19:46:04 INFO mapred.JobClient:  map 56% reduce 17%
```

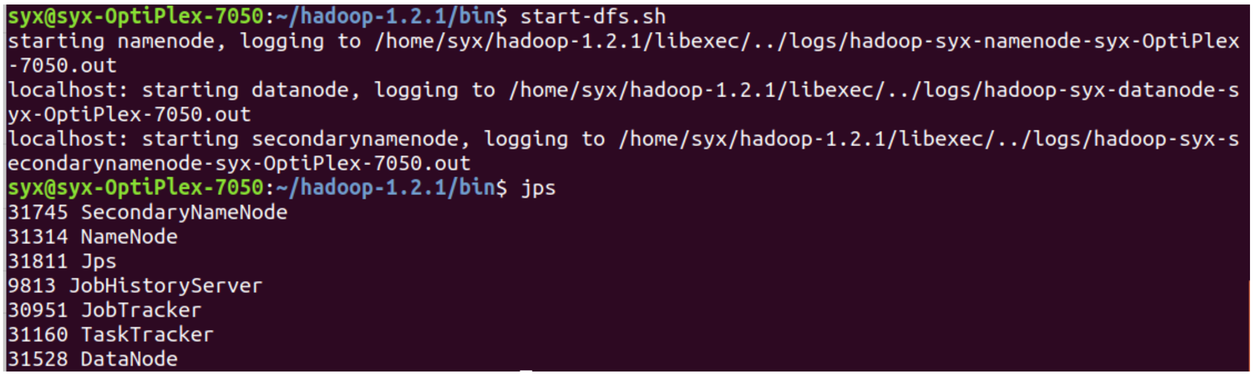


```
syx@syx-OptiPlex-7050:~$ jps
9904 SecondaryNameNode
24963 Child
9476 NameNode
10697 TaskTracker
10474 JobTracker
27626 Jps
9690 DataNode
24574 RunJar
syx@syx-OptiPlex-7050:~$ 
```

2.2 单机伪分布式部署

2.2.1 启动MapReduce服务

- `~/hadoop-1.2.1/bin/start-mapred.sh`
- `~/hadoop-1.2.1/bin/start-dfs.sh`

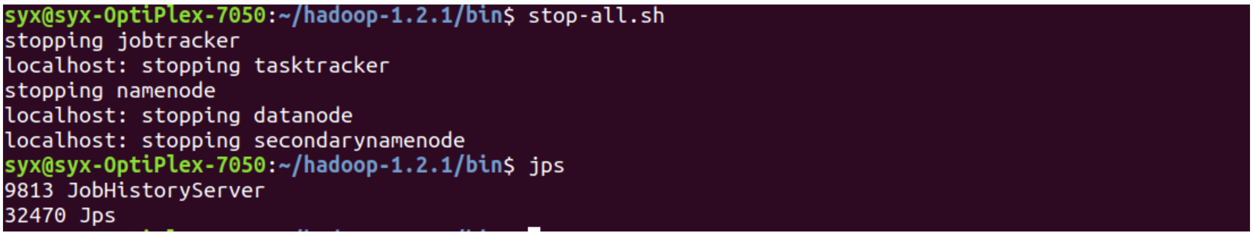


```
syx@syx-OptiPlex-7050:~/hadoop-1.2.1/bin$ start-dfs.sh
starting namenode, logging to /home/syx/hadoop-1.2.1/libexec/../logs/hadoop-syx-namenode-syx-OptiPlex
-7050.out
localhost: starting datanode, logging to /home/syx/hadoop-1.2.1/libexec/../logs/hadoop-syx-datanode-s
yx-OptiPlex-7050.out
localhost: starting secondarynamenode, logging to /home/syx/hadoop-1.2.1/libexec/../logs/hadoop-syx-s
econdarynamenode-syx-OptiPlex-7050.out
syx@syx-OptiPlex-7050:~/hadoop-1.2.1/bin$ jps
31745 SecondaryNameNode
31314 NameNode
31811 Jps
9813 JobHistoryServer
30951 JobTracker
31160 TaskTracker
31528 DataNode
```

2.3 停止MapReduce服务

2.3.1 停止命令

- 使用jps查看进程，不再出现NameNode、SecondaryNameNode、DataNode、JobTracker、TaskTracker等进程则服务停止



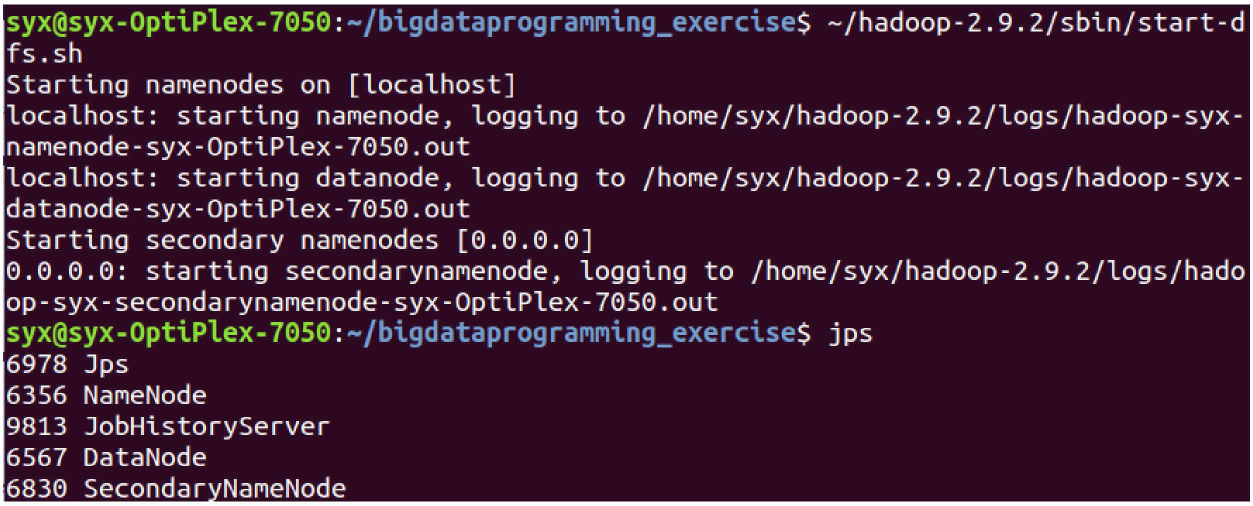
```
syx@syx-OptiPlex-7050:~/hadoop-1.2.1/bin$ stop-all.sh
stopping jobtracker
localhost: stopping tasktracker
stopping namenode
localhost: stopping datanode
localhost: stopping secondarynamenode
syx@syx-OptiPlex-7050:~/hadoop-1.2.1/bin$ jps
9813 JobHistoryServer
32470 Jps
```

3. HDFS v2部署

3.1 单机伪分布式部署

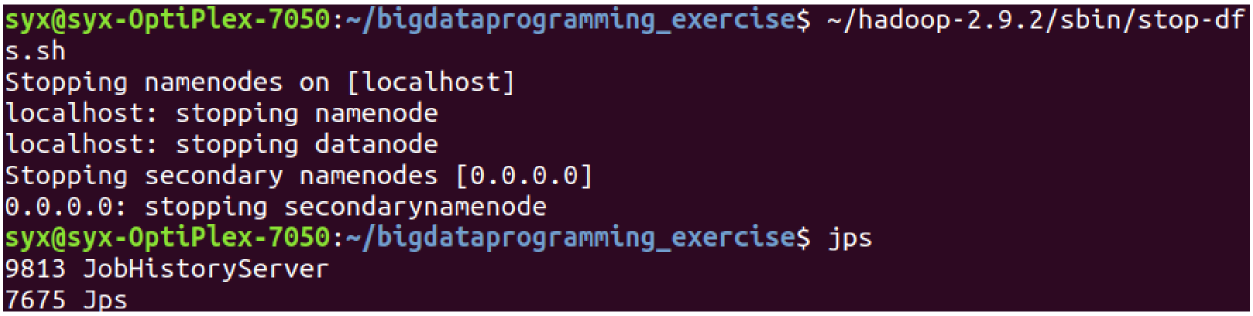
3.1.1 HDFS 服务

- 启动, `jps` 查看 HDFS 进程. 若出现 NameNode, DataNode, SecondaryNameNode, 则表示启动成功



```
syx@syx-OptiPlex-7050:~/bigdataprogramming_exercise$ ~/hadoop-2.9.2/sbin/start-d
fs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/syx/hadoop-2.9.2/logs/hadoop-syx-
namenode-syx-OptiPlex-7050.out
localhost: starting datanode, logging to /home/syx/hadoop-2.9.2/logs/hadoop-syx-
datanode-syx-OptiPlex-7050.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/syx/hadoop-2.9.2/logs/hado
op-syx-secondarynamenode-syx-OptiPlex-7050.out
syx@syx-OptiPlex-7050:~/bigdataprogramming_exercise$ jps
6978 Jps
6356 NameNode
9813 JobHistoryServer
6567 DataNode
6830 SecondaryNameNode
```

- 停止, 使用 jps 查看进程, 不再出现 NameNode, DataNode, SecondaryNameNode 则表示服务停止



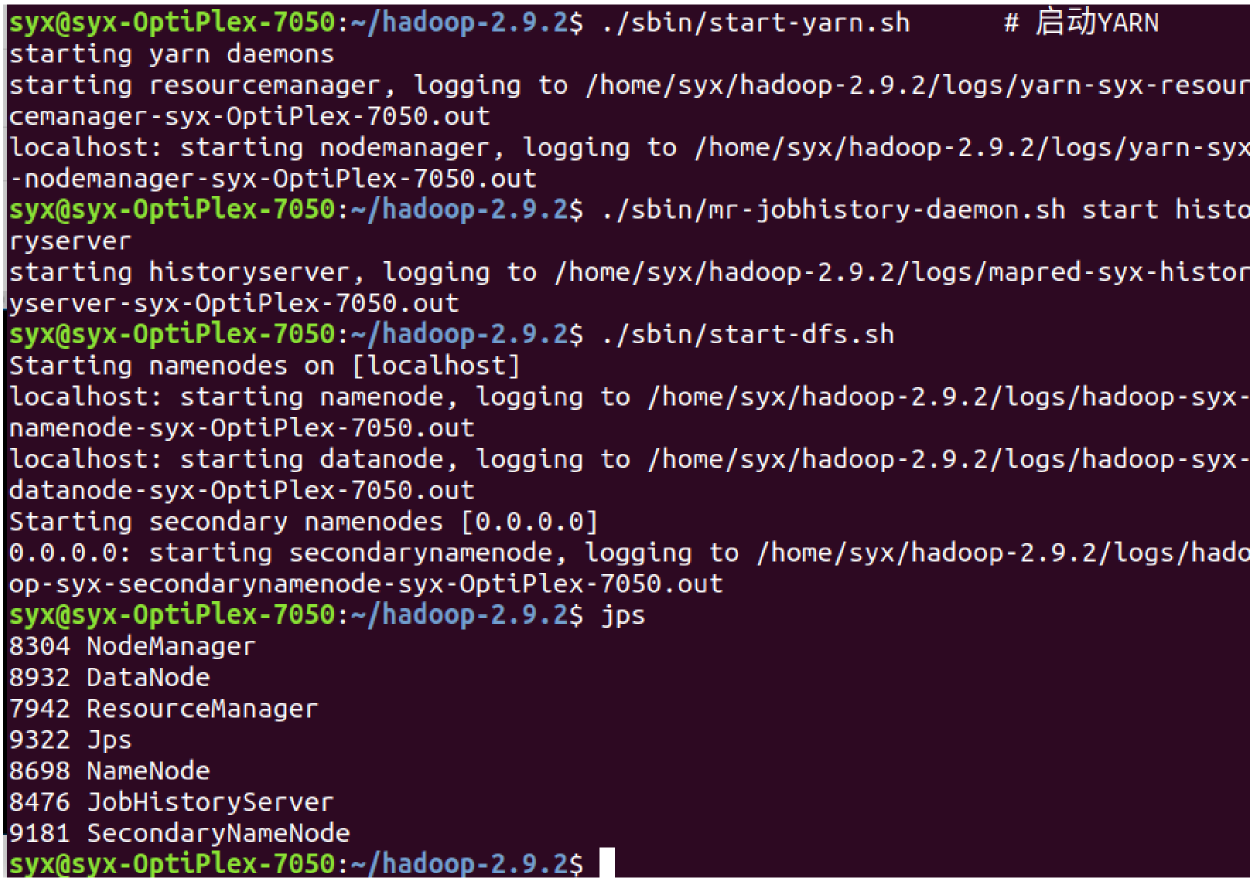
```
syx@syx-OptiPlex-7050:~/bigdataprogramming_exercise$ ~/hadoop-2.9.2/sbin/stop-df
s.sh
Stopping namenodes on [localhost]
localhost: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
syx@syx-OptiPlex-7050:~/bigdataprogramming_exercise$ jps
9813 JobHistoryServer
7675 Jps
```

4. MapReduce v2 部署

4.1 单机集中式部署

4.1.1 启动MapReduce服务

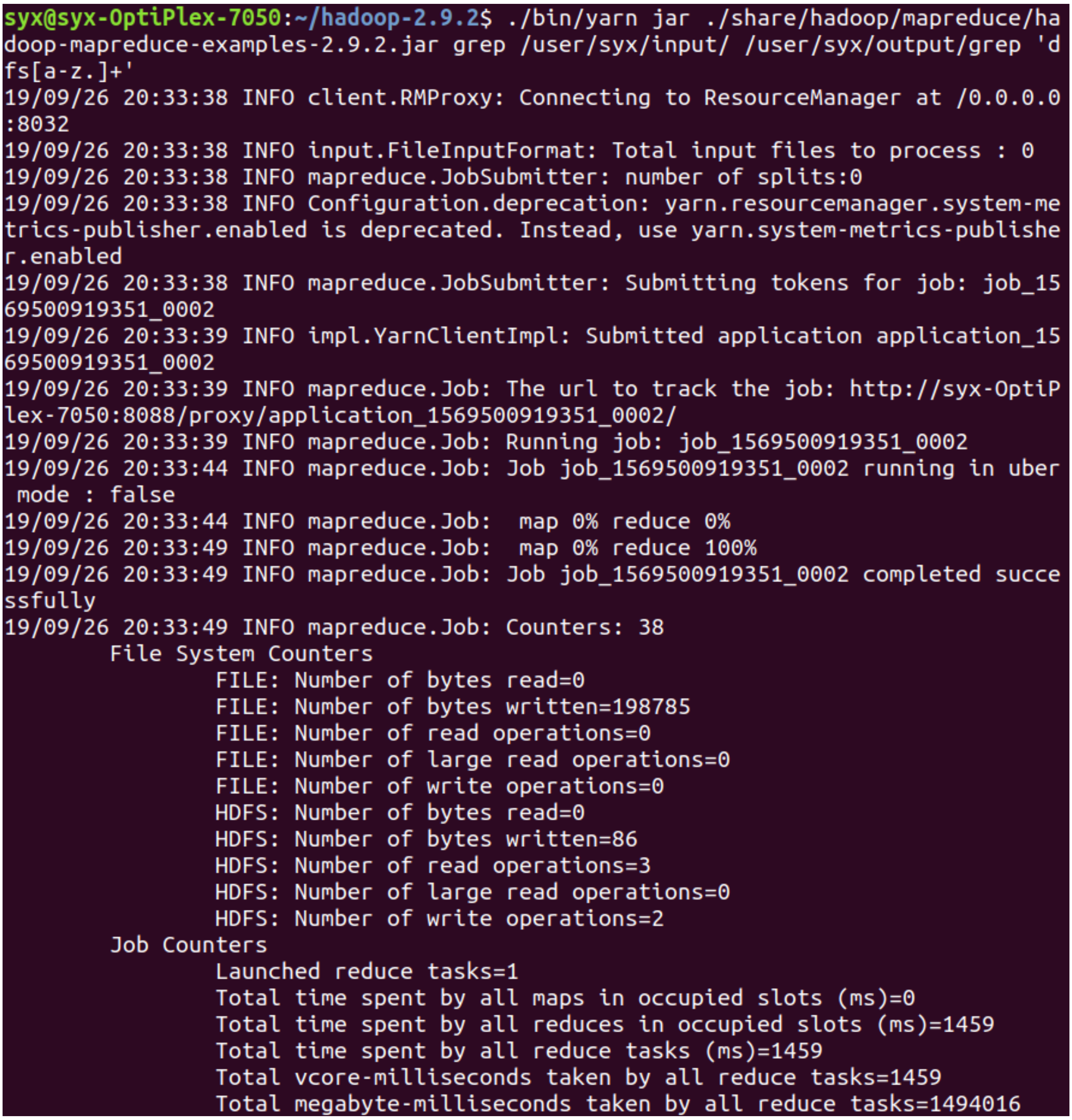
- 启动YARN命令
- 开启历史服务器
- 启动HDFS服务
- `jps`



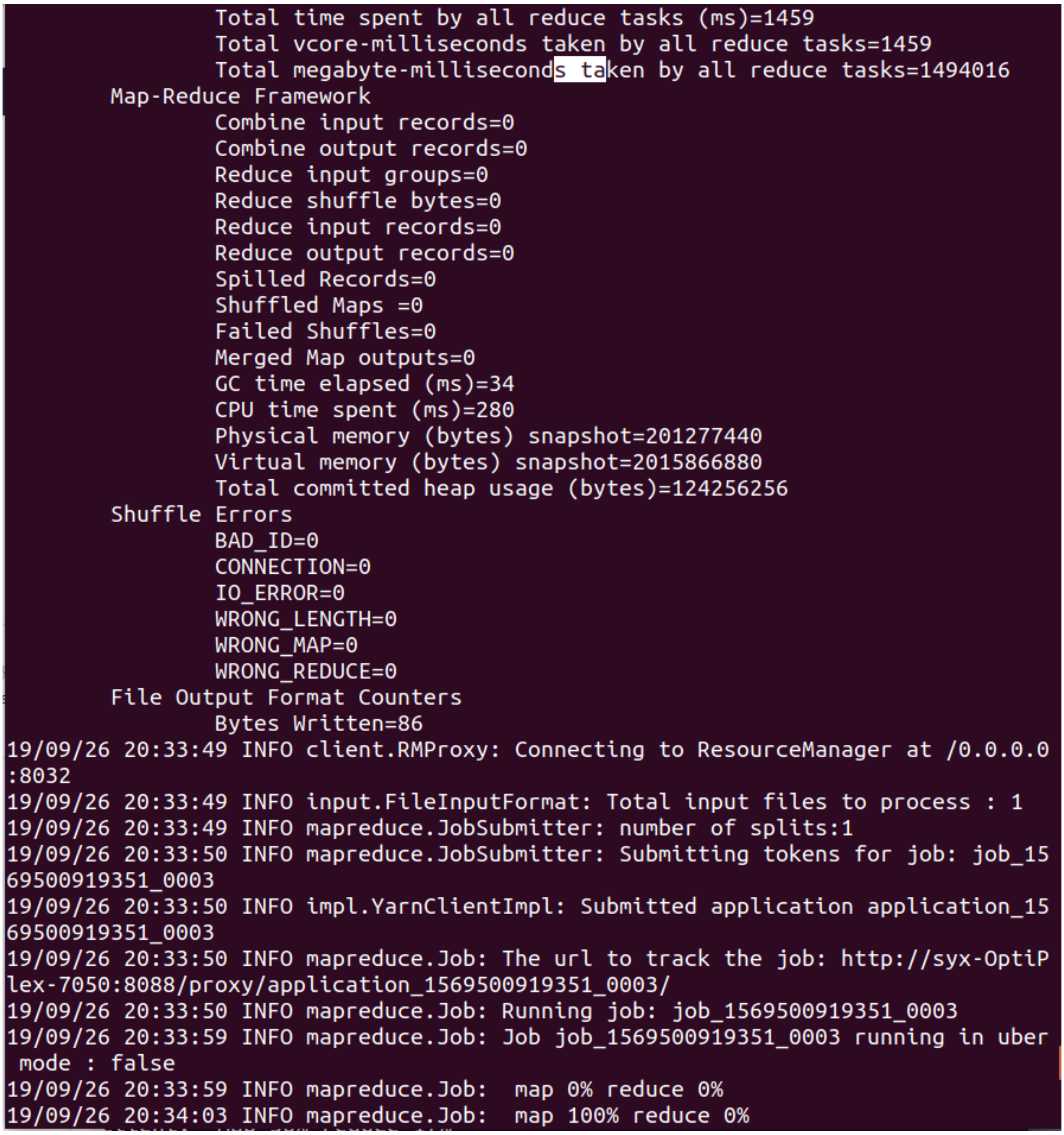
```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./sbin/start-yarn.sh      # 启动YARN
starting yarn daemons
starting resourcemanager, logging to /home/syx/hadoop-2.9.2/logs/yarn-syx-resour
cemanager-syx-OptiPlex-7050.out
localhost: starting nodemanager, logging to /home/syx/hadoop-2.9.2/logs/yarn-syx
-nodemanager-syx-OptiPlex-7050.out
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./sbin/mr-jobhistory-daemon.sh start histo
ryserver
starting historyserver, logging to /home/syx/hadoop-2.9.2/logs/mapred-syx-histor
yserver-syx-OptiPlex-7050.out
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/syx/hadoop-2.9.2/logs/hadoop-syx-
namenode-syx-OptiPlex-7050.out
localhost: starting datanode, logging to /home/syx/hadoop-2.9.2/logs/hadoop-syx-
datanode-syx-OptiPlex-7050.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/syx/hadoop-2.9.2/logs/hado
op-syx-secondarynamenode-syx-OptiPlex-7050.out
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ jps
8304 NodeManager
8932 DataNode
7942 ResourceManager
9322 Jps
8698 NameNode
8476 JobHistoryServer
9181 SecondaryNameNode
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ 
```

4.1.2 运行MapReduce应用程序

- 提交jar命令并查看运行结果, 运行grep示例, 结果如下:

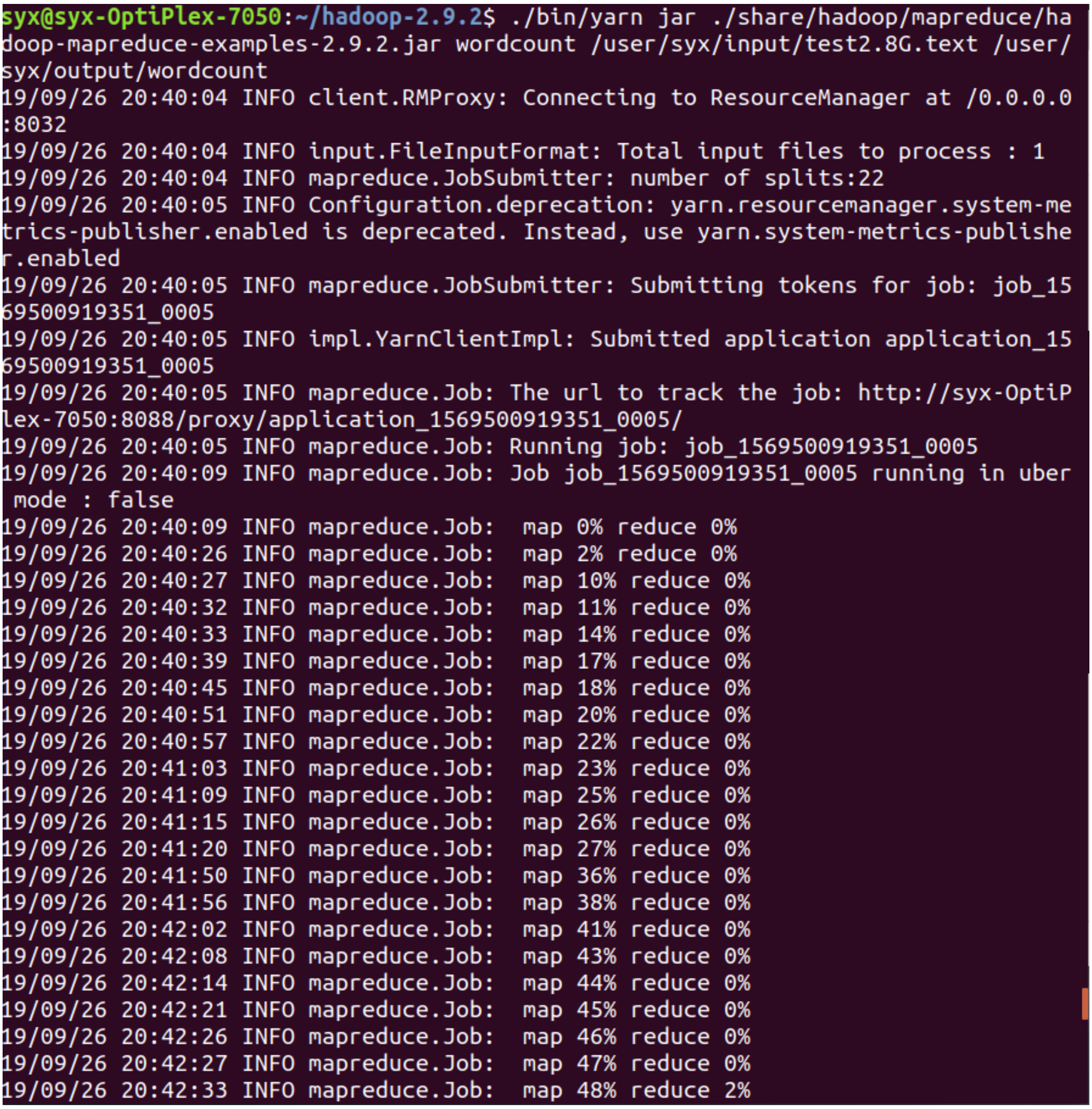


```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./bin/yarn jar ./share/hadoop/mapreduce/ha
doop-mapreduce-examples-2.9.2.jar grep /user/syx/input/ /user/syx/output/grep 'd
fs[a-z.]+'
19/09/26 20:33:38 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
19/09/26 20:33:38 INFO input.FileInputFormat: Total input files to process : 0
19/09/26 20:33:38 INFO mapreduce.JobSubmitter: number of splits:0
19/09/26 20:33:38 INFO Configuration.deprecation: yarn.resourcemanager.system-me
trics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publishe
r.enabled
19/09/26 20:33:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
69500919351_0002
19/09/26 20:33:39 INFO impl.YarnClientImpl: Submitted application application_15
69500919351_0002
19/09/26 20:33:39 INFO mapreduce.Job: The url to track the job: http://syx-OptiP
lex-7050:8088/proxy/application_1569500919351_0002/
19/09/26 20:33:39 INFO mapreduce.Job: Running job: job_1569500919351_0002
19/09/26 20:33:44 INFO mapreduce.Job: Job job_1569500919351_0002 running in uber
 mode : false
19/09/26 20:33:44 INFO mapreduce.Job:  map 0% reduce 0%
19/09/26 20:33:49 INFO mapreduce.Job:  map 0% reduce 100%
19/09/26 20:33:49 INFO mapreduce.Job: Job job_1569500919351_0002 completed succe
ssfully
19/09/26 20:33:49 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=198785
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=0
                HDFS: Number of bytes written=86
                HDFS: Number of read operations=3
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched reduce tasks=1
                Total time spent by all maps in occupied slots (ms)=0
                Total time spent by all reduces in occupied slots (ms)=1459
                Total time spent by all reduce tasks (ms)=1459
                Total vcore-milliseconds taken by all reduce tasks=1459
                Total megabyte-milliseconds taken by all reduce tasks=1494016
```



```
                Total time spent by all reduce tasks (ms)=1459
                Total vcore-milliseconds taken by all reduce tasks=1459
                Total megabyte-milliseconds taken by all reduce tasks=1494016
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=0
                Reduce shuffle bytes=0
                Reduce input records=0
                Reduce output records=0
                Spilled Records=0
                Shuffled Maps =0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=34
                CPU time spent (ms)=280
                Physical memory (bytes) snapshot=201277440
                Virtual memory (bytes) snapshot=2015866880
                Total committed heap usage (bytes)=124256256
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=86
19/09/26 20:33:49 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
19/09/26 20:33:49 INFO input.FileInputFormat: Total input files to process : 1
19/09/26 20:33:49 INFO mapreduce.JobSubmitter: number of splits:1
19/09/26 20:33:50 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
69500919351_0003
19/09/26 20:33:50 INFO impl.YarnClientImpl: Submitted application application_15
69500919351_0003
19/09/26 20:33:50 INFO mapreduce.Job: The url to track the job: http://syx-OptiP
lex-7050:8088/proxy/application_1569500919351_0003/
19/09/26 20:33:50 INFO mapreduce.Job: Running job: job_1569500919351_0003
19/09/26 20:33:59 INFO mapreduce.Job: Job job_1569500919351_0003 running in uber
 mode : false
19/09/26 20:33:59 INFO mapreduce.Job:  map 0% reduce 0%
19/09/26 20:34:03 INFO mapreduce.Job:  map 100% reduce 0%
```

- 运行 wordcount 示例

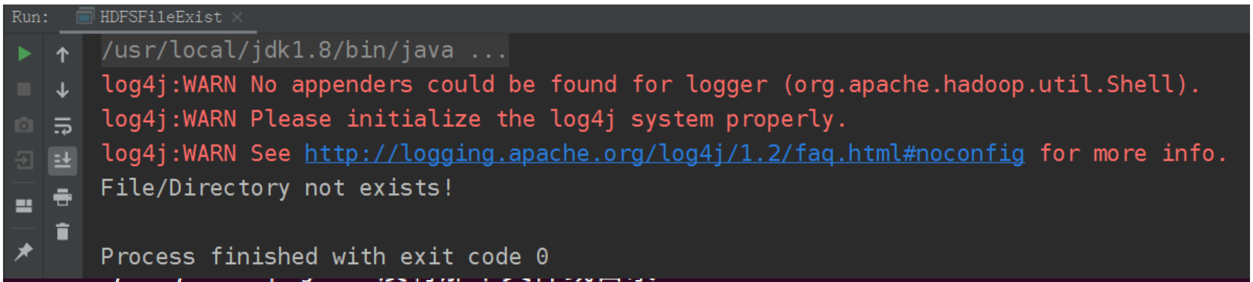


```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./bin/yarn jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar wordcount /user/syx/input/test2.8G.text /user/syx/output/wordcount
19/09/26 20:40:04 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/09/26 20:40:04 INFO input.FileInputFormat: Total input files to process : 1
19/09/26 20:40:04 INFO mapreduce.JobSubmitter: number of splits:22
19/09/26 20:40:05 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
19/09/26 20:40:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1569500919351_0005
19/09/26 20:40:05 INFO impl.YarnClientImpl: Submitted application application_1569500919351_0005
19/09/26 20:40:05 INFO mapreduce.Job: The url to track the job: http://syx-OptiPlex-7050:8088/proxy/application_1569500919351_0005/
19/09/26 20:40:05 INFO mapreduce.Job: Running job: job_1569500919351_0005
19/09/26 20:40:09 INFO mapreduce.Job: Job job_1569500919351_0005 running in uber mode : false
19/09/26 20:40:09 INFO mapreduce.Job:  map 0% reduce 0%
19/09/26 20:40:26 INFO mapreduce.Job:  map 2% reduce 0%
19/09/26 20:40:27 INFO mapreduce.Job:  map 10% reduce 0%
19/09/26 20:40:32 INFO mapreduce.Job:  map 11% reduce 0%
19/09/26 20:40:33 INFO mapreduce.Job:  map 14% reduce 0%
19/09/26 20:40:39 INFO mapreduce.Job:  map 17% reduce 0%
19/09/26 20:40:45 INFO mapreduce.Job:  map 18% reduce 0%
19/09/26 20:40:51 INFO mapreduce.Job:  map 20% reduce 0%
19/09/26 20:40:57 INFO mapreduce.Job:  map 22% reduce 0%
19/09/26 20:41:03 INFO mapreduce.Job:  map 23% reduce 0%
19/09/26 20:41:09 INFO mapreduce.Job:  map 25% reduce 0%
19/09/26 20:41:15 INFO mapreduce.Job:  map 26% reduce 0%
19/09/26 20:41:20 INFO mapreduce.Job:  map 27% reduce 0%
19/09/26 20:41:50 INFO mapreduce.Job:  map 36% reduce 0%
19/09/26 20:41:56 INFO mapreduce.Job:  map 38% reduce 0%
19/09/26 20:42:02 INFO mapreduce.Job:  map 41% reduce 0%
19/09/26 20:42:08 INFO mapreduce.Job:  map 43% reduce 0%
19/09/26 20:42:14 INFO mapreduce.Job:  map 44% reduce 0%
19/09/26 20:42:21 INFO mapreduce.Job:  map 45% reduce 0%
19/09/26 20:42:26 INFO mapreduce.Job:  map 46% reduce 0%
19/09/26 20:42:27 INFO mapreduce.Job:  map 47% reduce 0%
19/09/26 20:42:33 INFO mapreduce.Job:  map 48% reduce 2%
```

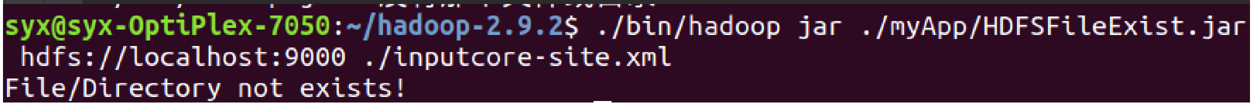
5.HDFS 应用编程实践

5.1 使用IntelliJ IDEA编写 测试HDFS中是否存在一个文件 应用程序

5.1.1运行 `测试HDFS中是否存在一个文件` 应用程序 HDFSFileExist:

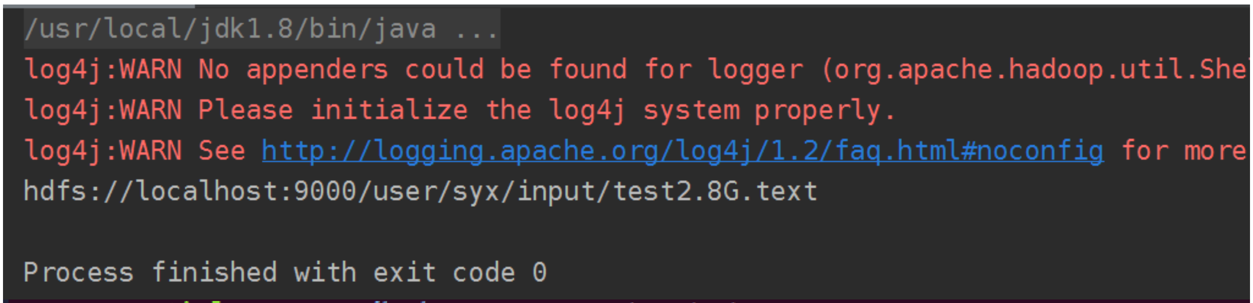
- 

```
/usr/local/jdk1.8/bin/java ...
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
File/Directory not exists!

Process finished with exit code 0
```

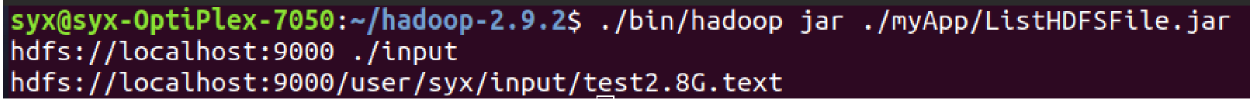
- 

```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./bin/hadoop jar ./myApp/HDFSFileExist.jar hdfs://localhost:9000 ./inputcore-site.xml
File/Directory not exists!
```

5.2 列出目录下所有文件 ListHDFSFile:

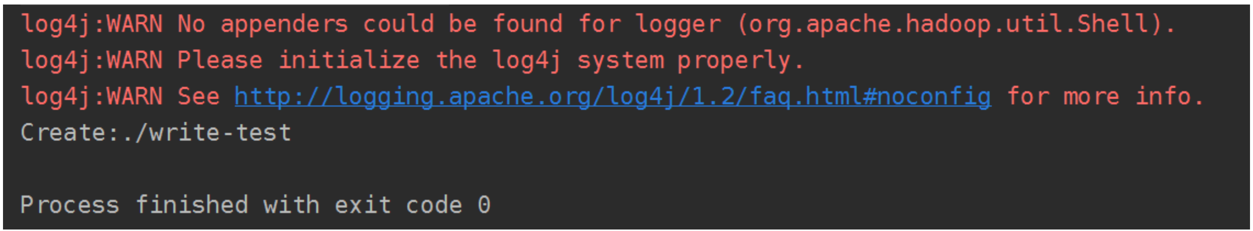
- 

```
/usr/local/jdk1.8/bin/java ...
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.She
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more
hdfs://localhost:9000/user/syx/input/test2.8G.text

Process finished with exit code 0
```

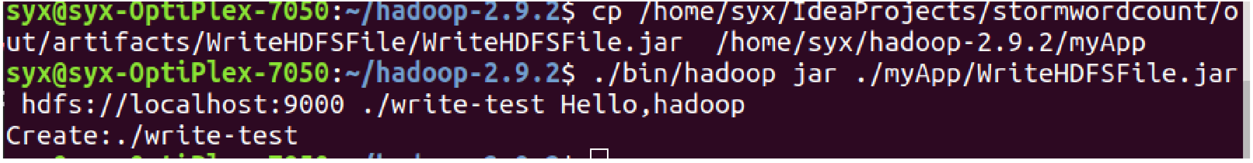
- 

```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./bin/hadoop jar ./myApp/ListHDFSFile.jar
hdfs://localhost:9000 ./input
hdfs://localhost:9000/user/syx/input/test2.8G.text
```

5.3 写入文件 WriteHDFSFIle

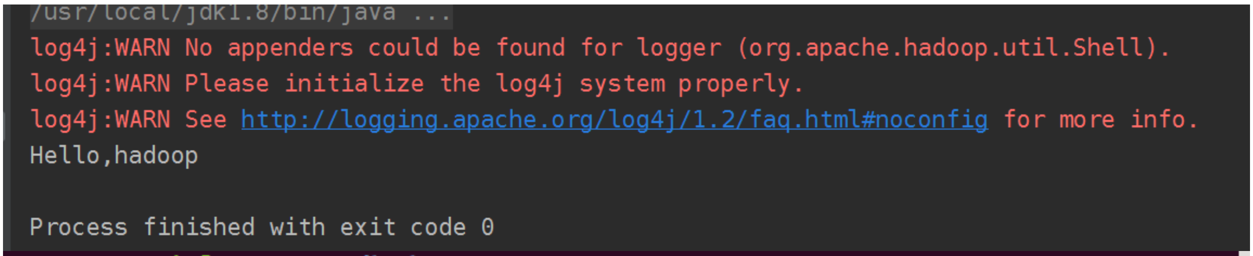
- 

```
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Create:./write-test

Process finished with exit code 0
```

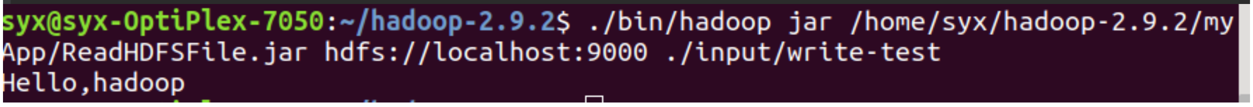
- 

```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ cp /home/syx/IdeaProjects/stormwordcount/o
ut/artifacts/WriteHDFSFile/WriteHDFSFile.jar  /home/syx/hadoop-2.9.2/myApp
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./bin/hadoop jar ./myApp/WriteHDFSFile.jar
 hdfs://localhost:9000 ./write-test Hello,hadoop
Create:./write-test
```

5.4 读取文件 ReadHDFSFile

- 

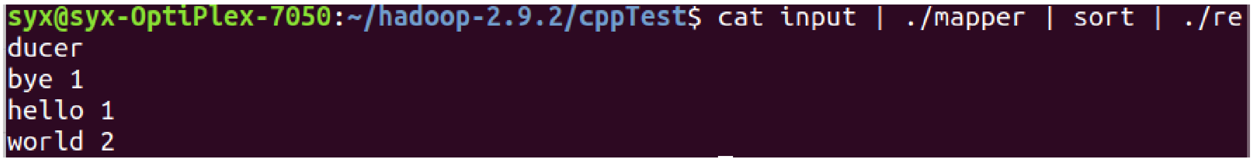
```
/usr/local/jdk1.8/bin/java ...
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Hello,hadoop

Process finished with exit code 0
```

- 

```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./bin/hadoop jar /home/syx/hadoop-2.9.2/my
App/ReadHDFSFile.jar hdfs://localhost:9000 ./input/write-test
Hello,hadoop
```

6. Hadoop Straming 介绍与实践

6.1使用 C++ 编写 Mapper/Reducer 源文件, 脚本测试

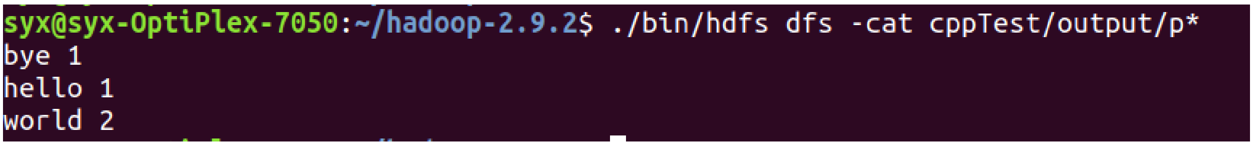
- 

```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2/cppTest$ cat input | ./mapper | sort | ./re
ducer
bye 1
hello 1
world 2
```

6.2 使用 Hadoop Streaming 运行

6.2.1 伪分布式或分布式部署时

- 输出结果和脚本测试结果相同



```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./bin/hdfs dfs -cat cppTest/output/p*
bye 1
hello 1
world 2
```

6.3 Hadoop Streaming Shell 示例

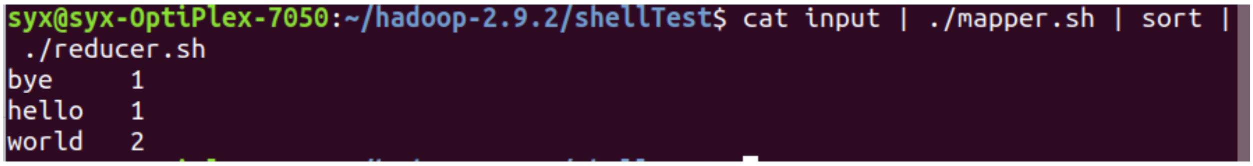
6.3.1 使用 Shell 编写 Mapper/Reducer 源文件, 脚本测试

- 使用 Shell 编写 Mapper 和 Reducer, 实现 wordcount
- 编写输入文件 vi ~/hadoop-2.9.2/shellTest/input

  `hello world`

  `bye world`

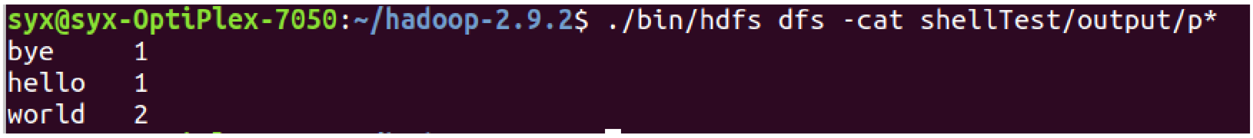
- 使用脚本测试



```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2/shellTest$ cat input | ./mapper.sh | sort |
 ./reducer.sh
bye     1
hello   1
world   2
```

6.3.2 使用 Hadoop Streaming 运行

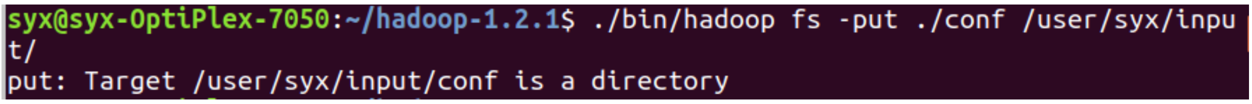
- 伪分布式或分布式部署时



```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./bin/hdfs dfs -cat shellTest/output/p*
bye     1
hello   1
world   2
```

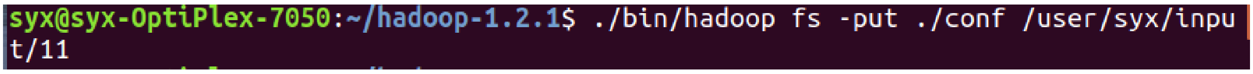
五. 总结

1. FS



```
syx@syx-OptiPlex-7050:~/hadoop-1.2.1$ ./bin/hadoop fs -put ./conf /user/syx/inpu
t/
put: Target /user/syx/input/conf is a directory
```

solu:



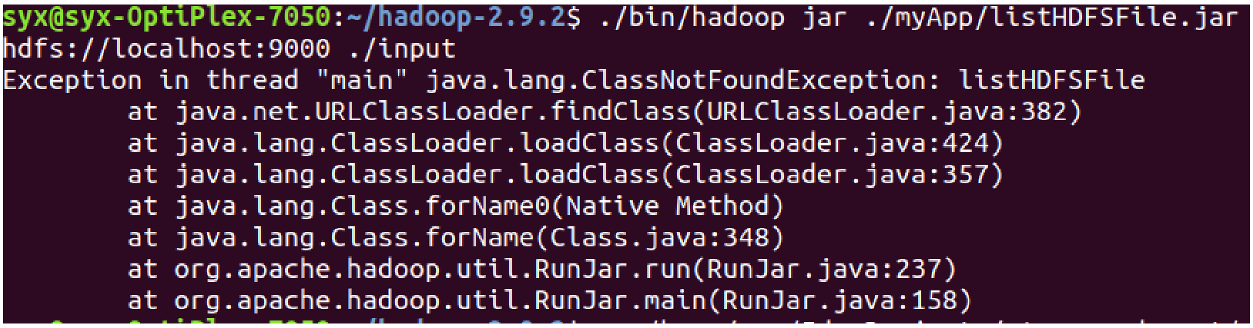
```
syx@syx-OptiPlex-7050:~/hadoop-1.2.1$ ./bin/hadoop fs -put ./conf /user/syx/inpu
t/11
```

hadoop's FS is different from linux's FS. I have tried

```
sudo rm -rf /user/syx/input
```

but this "input" is not the one in the hadoop FS.

2. Ignore the upper-case of the file name



```
syx@syx-OptiPlex-7050:~/hadoop-2.9.2$ ./bin/hadoop jar ./myApp/listHDFSFile.jar
hdfs://localhost:9000 ./input
Exception in thread "main" java.lang.ClassNotFoundException: listHDFSFile
        at java.net.URLClassLoader.findClass(URLClassLoader.java:382)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:424)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:357)
        at java.lang.Class.forName0(Native Method)
        at java.lang.Class.forName(Class.java:348)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:237)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:158)
```

the correct name of class is : ListHDFSFile

3. 按照实验步骤做时，要理解每一步的含义，注意warn提示，方便debug;