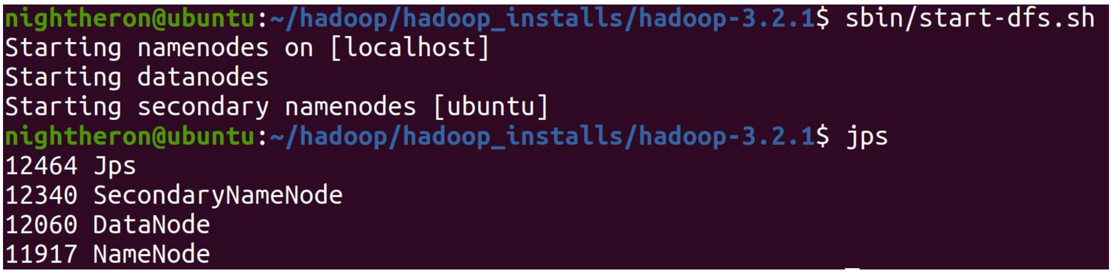
实验 1：Hadoop 系统安装、开源大数据系统实验工具链使用

陈德丹 221220159

任务 1：Hadoop 伪分布式环境安装与配置，HDFS 基本操作

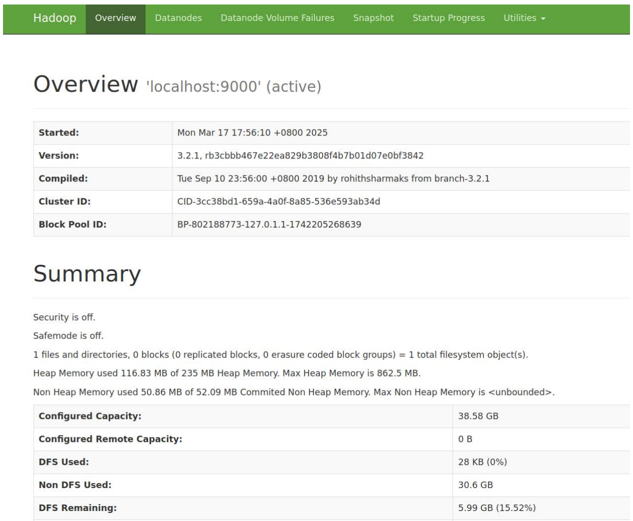
○ 按照教程安装并配置 hadoop 和 java

○ 启动 NameNode daemon 和 DataNode daemon



```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu]
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ jps
12464 Jps
12340 SecondaryNameNode
12060 DataNode
11917 NameNode
```

○ 浏览 HDFS NameNode 的 Web 接口



Overview 'localhost:9000' (active)

Started:	Mon Mar 17 17:56:10 +0800 2025
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 23:56:00 +0800 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-3cc38bd1-659a-4a0f-8a85-536e593ab34d
Block Pool ID:	BP-802188773-127.0.1.1-1742205268639

Summary

Security is off.

Safemode is off.

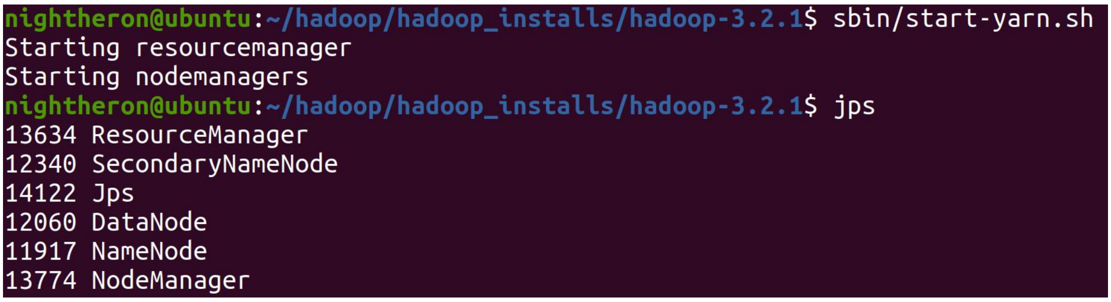
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 116.83 MB of 235 MB Heap Memory. Max Heap Memory is 862.5 MB.

Non Heap Memory used 50.86 MB of 52.09 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

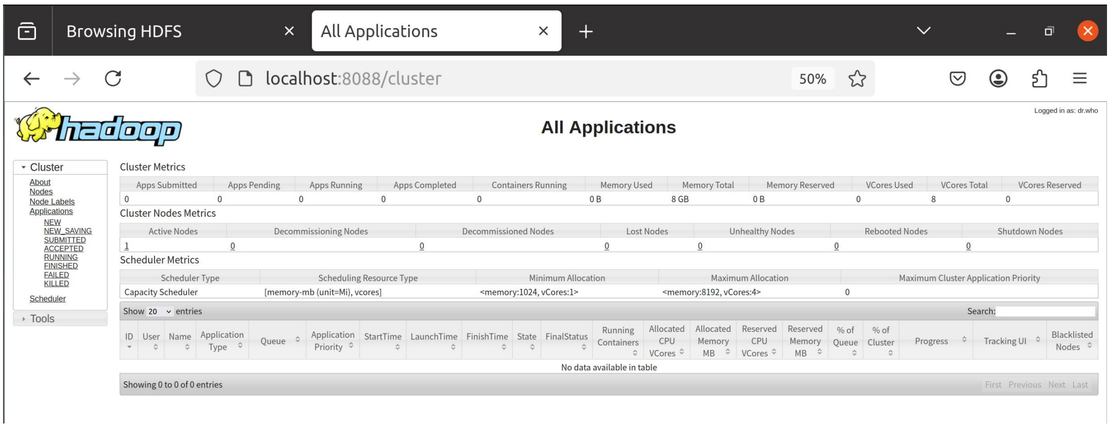
Configured Capacity:	38.58 GB
Configured Remote Capacity:	0 B
DFS Used:	28 KB (0%)
Non DFS Used:	30.6 GB
DFS Remaining:	5.99 GB (15.52%)

○ 启动 ResourceManage daemon 和 NodeManage daemon



```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ sbin/start-yarn.sh
Starting resourcemanager
Starting nodemanagers
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ jps
13634 ResourceManager
12340 SecondaryNameNode
14122 Jps
12060 DataNode
11917 NameNode
13774 NodeManager
```

○ 浏览 Resource Manager 的 Web 接口



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	8 GB	0 B	0	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

Scheduler Metrics

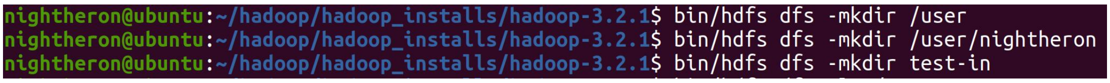
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																				

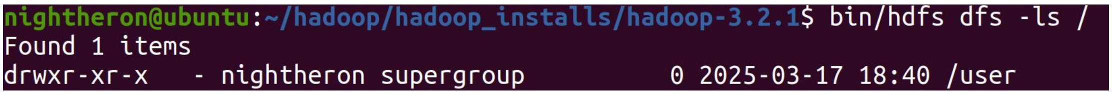
Showing 0 to 0 of 0 entries First Previous Next Last

○ 创建 HDFS 目录



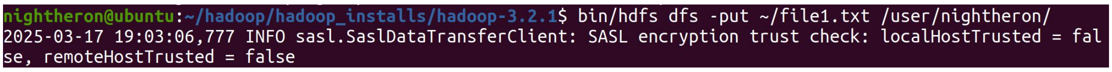
```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hdfs dfs -mkdir /user
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hdfs dfs -mkdir /user/nightheron
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hdfs dfs -mkdir test-in
```

○ 查看 HDFS 根目录



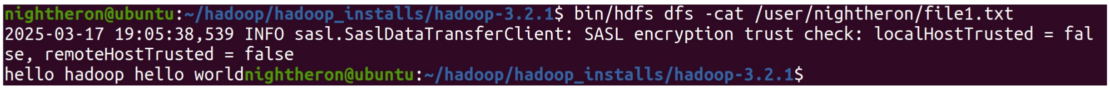
```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - nightheron supergroup          0 2025-03-17 18:40 /user
```

○ 向 HDFS 上传一个文件 file1.txt: hello hadoop hello world



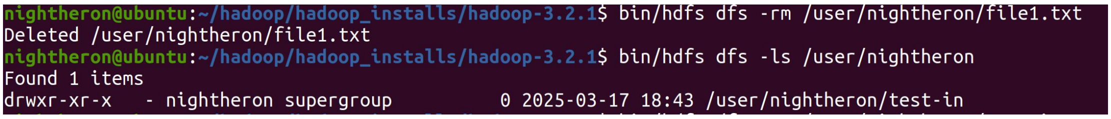
```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hdfs dfs -put ~/file1.txt /user/nightheron/
2025-03-17 19:03:06,777 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = fal
se, remoteHostTrusted = false
```

○ 显示上传的文件的内容



```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hdfs dfs -cat /user/nightheron/file1.txt
2025-03-17 19:05:38,539 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = fal
se, remoteHostTrusted = false
hello hadoop hello worldnightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$
```

○ 删除上传的文件

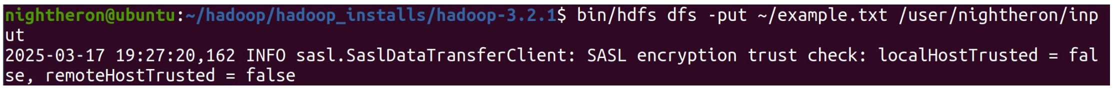


```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hdfs dfs -rm /user/nightheron/file1.txt
Deleted /user/nightheron/file1.txt
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hdfs dfs -ls /user/nightheron
Found 1 items
drwxr-xr-x   - nightheron supergroup          0 2025-03-17 18:43 /user/nightheron/test-in
```

可以看到已成功删除

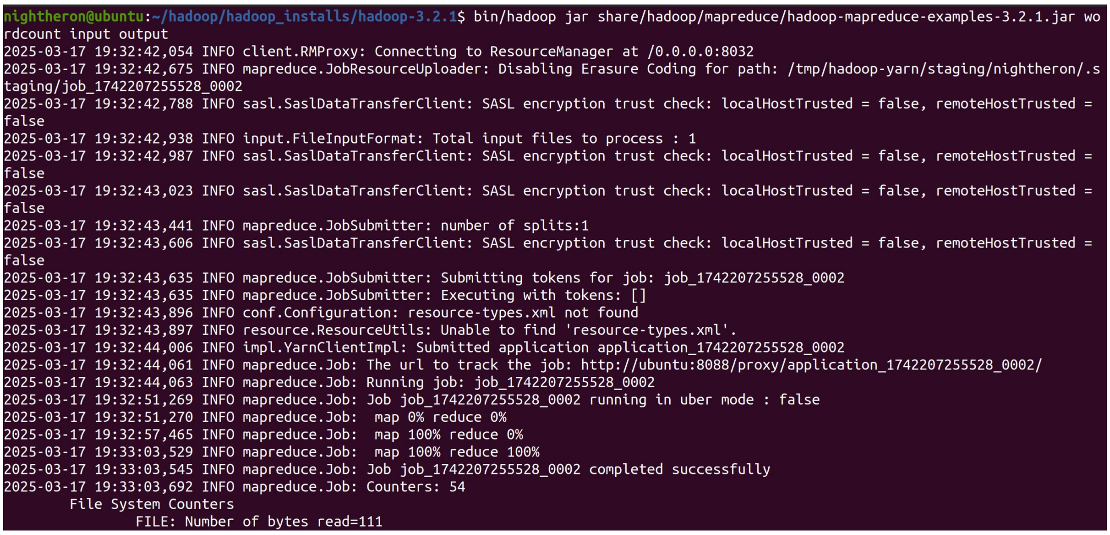
任务 2：执行 Hadoop 官方示例程序 wordcount

○ 上传 example.txt 到 HDFS

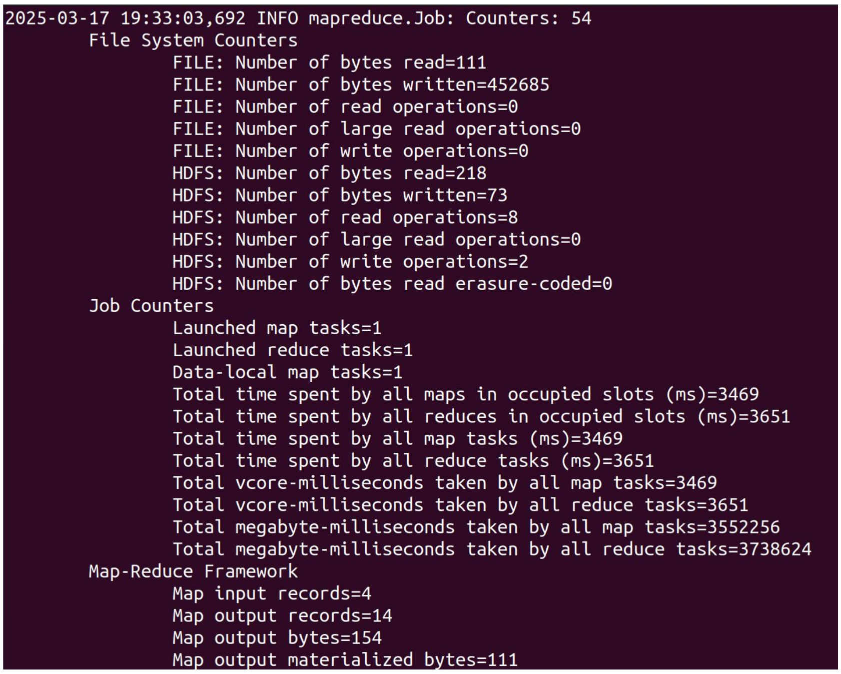


```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hdfs dfs -put ~/example.txt /user/nightheron/inp
ut
2025-03-17 19:27:20,162 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = fal
se, remoteHostTrusted = false
```

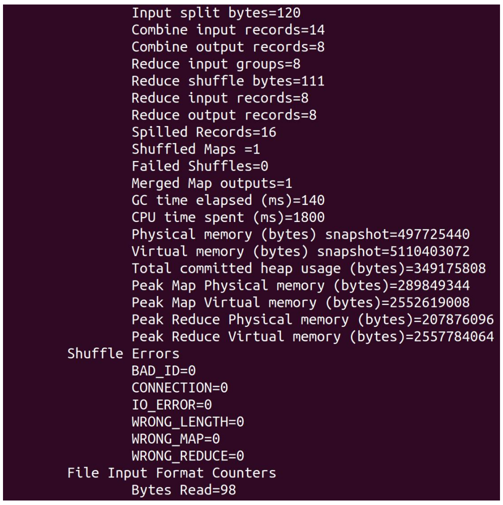
○ 执行 wordcount



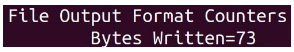
```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar wo
rdcount input output
2025-03-17 19:32:42,054 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2025-03-17 19:32:42,675 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/nightheron/.s
taging/job_1742207255528_0002
2025-03-17 19:32:42,788 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
2025-03-17 19:32:42,938 INFO input.FileInputFormat: Total input files to process : 1
2025-03-17 19:32:42,987 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
2025-03-17 19:32:43,023 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
2025-03-17 19:32:43,441 INFO mapreduce.JobSubmitter: number of splits:1
2025-03-17 19:32:43,606 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
2025-03-17 19:32:43,635 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1742207255528_0002
2025-03-17 19:32:43,635 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-03-17 19:32:43,896 INFO conf.Configuration: resource-types.xml not found
2025-03-17 19:32:43,897 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-03-17 19:32:44,006 INFO impl.YarnClientImpl: Submitted application application_1742207255528_0002
2025-03-17 19:32:44,061 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1742207255528_0002/
2025-03-17 19:32:44,063 INFO mapreduce.Job: Running job: job_1742207255528_0002
2025-03-17 19:32:51,269 INFO mapreduce.Job: Job job_1742207255528_0002 running in uber mode : false
2025-03-17 19:32:51,270 INFO mapreduce.Job:  map 0% reduce 0%
2025-03-17 19:32:57,465 INFO mapreduce.Job:  map 100% reduce 0%
2025-03-17 19:33:03,529 INFO mapreduce.Job:  map 100% reduce 100%
2025-03-17 19:33:03,545 INFO mapreduce.Job: Job job_1742207255528_0002 completed successfully
2025-03-17 19:33:03,692 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=111
```



```
2025-03-17 19:33:03,692 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=111
                FILE: Number of bytes written=452685
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=218
                HDFS: Number of bytes written=73
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=3469
                Total time spent by all reduces in occupied slots (ms)=3651
                Total time spent by all map tasks (ms)=3469
                Total time spent by all reduce tasks (ms)=3651
                Total vcore-milliseconds taken by all map tasks=3469
                Total vcore-milliseconds taken by all reduce tasks=3651
                Total megabyte-milliseconds taken by all map tasks=3552256
                Total megabyte-milliseconds taken by all reduce tasks=3738624
        Map-Reduce Framework
                Map input records=4
                Map output records=14
                Map output bytes=154
                Map output materialized bytes=111
```

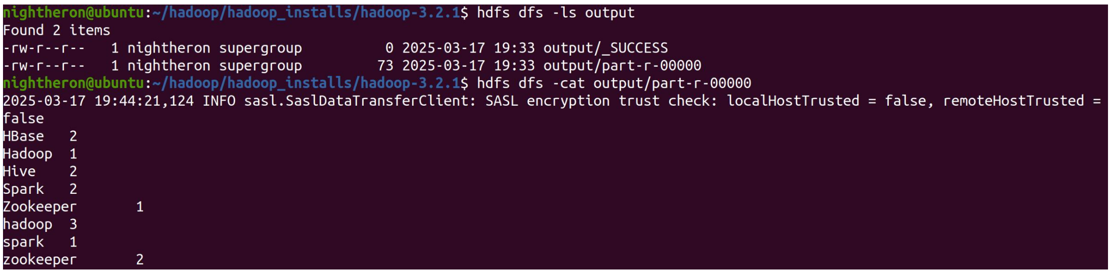


```
                Input split bytes=120
                Combine input records=14
                Combine output records=8
                Reduce input groups=8
                Reduce shuffle bytes=111
                Reduce input records=8
                Reduce output records=8
                Spilled Records=16
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=140
                CPU time spent (ms)=1800
                Physical memory (bytes) snapshot=497725440
                Virtual memory (bytes) snapshot=5110403072
                Total committed heap usage (bytes)=349175808
                Peak Map Physical memory (bytes)=289849344
                Peak Map Virtual memory (bytes)=2552619008
                Peak Reduce Physical memory (bytes)=207876096
                Peak Reduce Virtual memory (bytes)=2557784064
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=98
```



```
File Output Format Counters
        Bytes Written=73
```

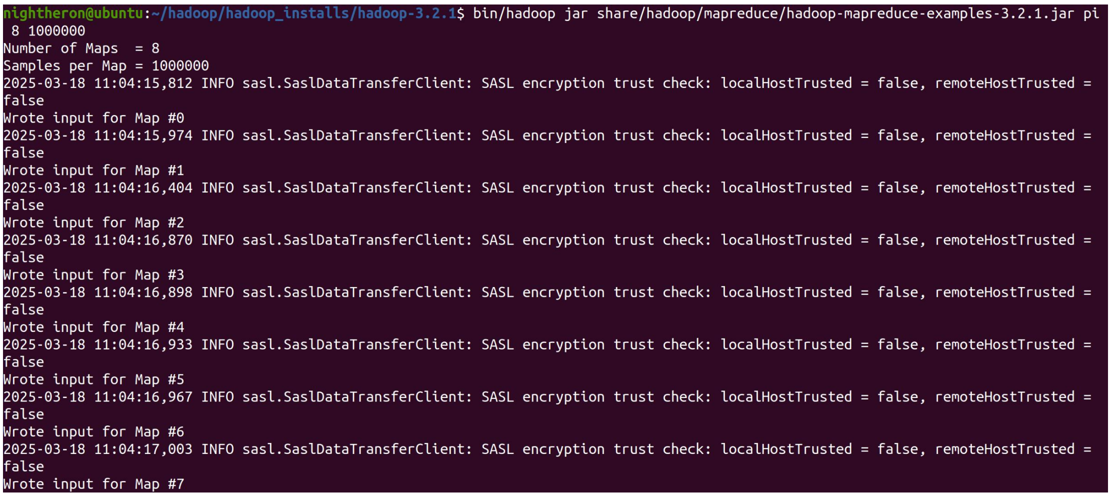
○ 查看执行结果:



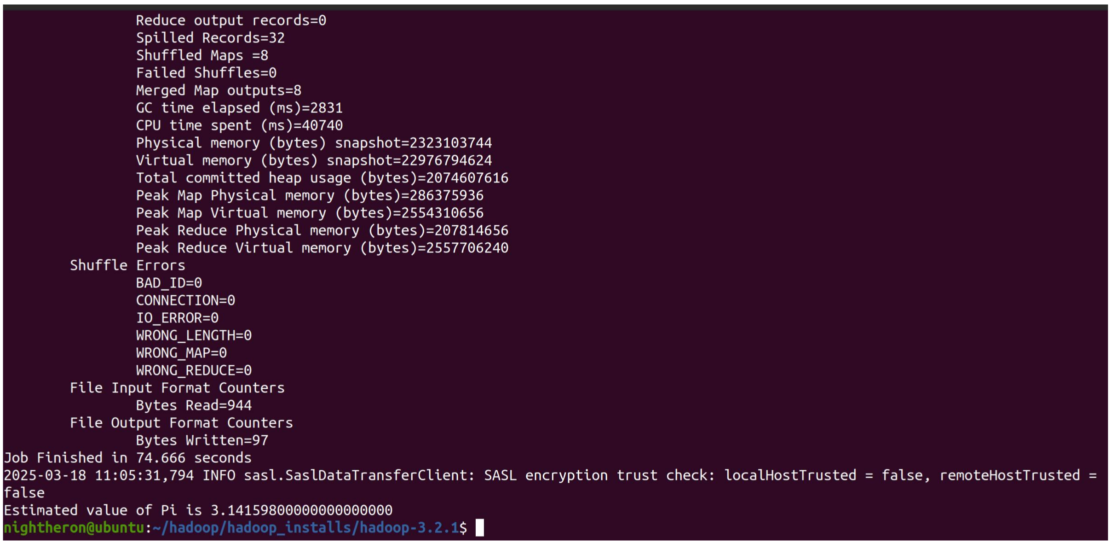
```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ hdfs dfs -ls output
Found 2 items
-rw-r--r--   1 nightheron supergroup          0 2025-03-17 19:33 output/_SUCCESS
-rw-r--r--   1 nightheron supergroup         73 2025-03-17 19:33 output/part-r-00000
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ hdfs dfs -cat output/part-r-00000
2025-03-17 19:44:21,124 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
HBase   2
Hadoop  1
Hive    2
Spark   2
Zookeeper       1
hadoop  3
spark   1
zookeeper       2
```

附加实验(可选)：执行更多的 Hadoop 官方示例程序

1.本地运行 pi 示例程序，观察 MapReduce 任务执行过程(8 个 map 任务，每个任务 1,000,000 个样本)

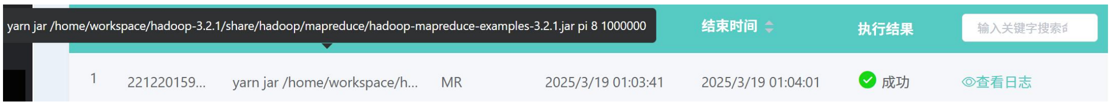


```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar pi
 8 1000000
Number of Maps  = 8
Samples per Map = 1000000
2025-03-18 11:04:15,812 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
Wrote input for Map #0
2025-03-18 11:04:15,974 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
Wrote input for Map #1
2025-03-18 11:04:16,404 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
Wrote input for Map #2
2025-03-18 11:04:16,870 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
Wrote input for Map #3
2025-03-18 11:04:16,898 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
Wrote input for Map #4
2025-03-18 11:04:16,933 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
Wrote input for Map #5
2025-03-18 11:04:16,967 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
Wrote input for Map #6
2025-03-18 11:04:17,003 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
Wrote input for Map #7
```



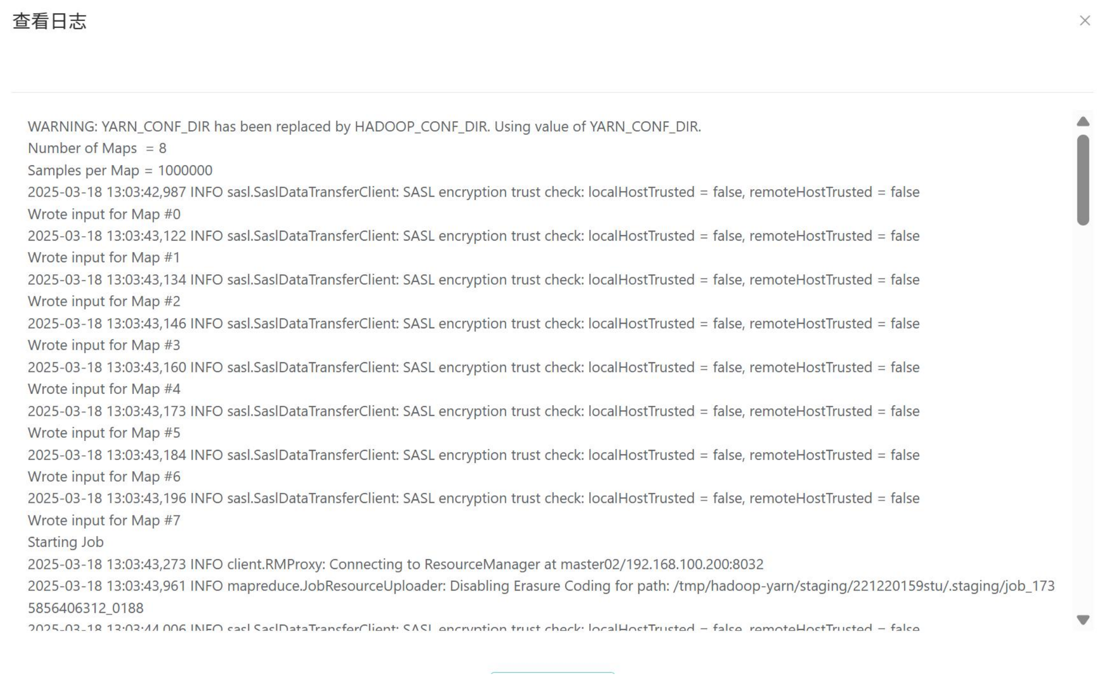
```
                Reduce output records=0
                Spilled Records=32
                Shuffled Maps =8
                Failed Shuffles=0
                Merged Map outputs=8
                GC time elapsed (ms)=2831
                CPU time spent (ms)=40740
                Physical memory (bytes) snapshot=2323103744
                Virtual memory (bytes) snapshot=22976794624
                Total committed heap usage (bytes)=2074607616
                Peak Map Physical memory (bytes)=286375936
                Peak Map Virtual memory (bytes)=2554310656
                Peak Reduce Physical memory (bytes)=207814656
                Peak Reduce Virtual memory (bytes)=2557706240
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=944
        File Output Format Counters
                Bytes Written=97
Job Finished in 74.666 seconds
2025-03-18 11:05:31,794 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
Estimated value of Pi is 3.14159800000000000000
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$
```

2.实验平台上运行 pi 示例程序，观察 MapReduce 任务执行过程(8 个 map 任务，每个任务 1,000,000 个样本)

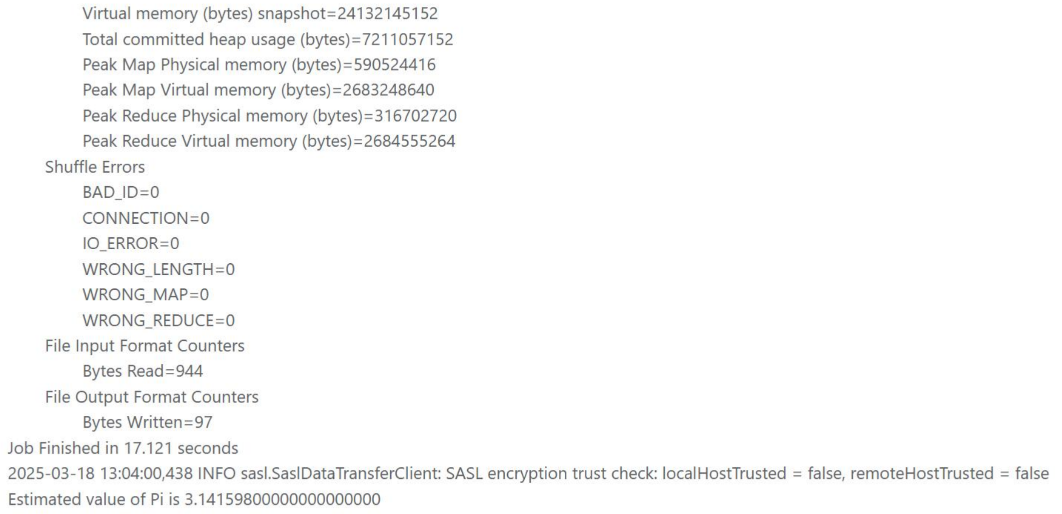


```
yarn jar /home/workspace/hadoop-3.2.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar pi 8 1000000
```

					结束时间	执行结果	
1	221220159...	yarn jar /home/workspace/h...	MR	2025/3/19 01:03:41	2025/3/19 01:04:01	✓ 成功	查看日志



```
WARNING: YARN_CONF_DIR has been replaced by HADOOP_CONF_DIR. Using value of YARN_CONF_DIR.
Number of Maps  = 8
Samples per Map = 1000000
2025-03-18 13:03:42,987 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Wrote input for Map #0
2025-03-18 13:03:43,122 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Wrote input for Map #1
2025-03-18 13:03:43,134 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Wrote input for Map #2
2025-03-18 13:03:43,146 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Wrote input for Map #3
2025-03-18 13:03:43,160 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Wrote input for Map #4
2025-03-18 13:03:43,173 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Wrote input for Map #5
2025-03-18 13:03:43,184 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Wrote input for Map #6
2025-03-18 13:03:43,196 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Wrote input for Map #7
Starting Job
2025-03-18 13:03:43,273 INFO client.RMProxy: Connecting to ResourceManager at master02/192.168.100.200:8032
2025-03-18 13:03:43,961 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/221220159stu/.staging/job_1735856406312_0188
2025-03-18 13:03:44,006 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
```

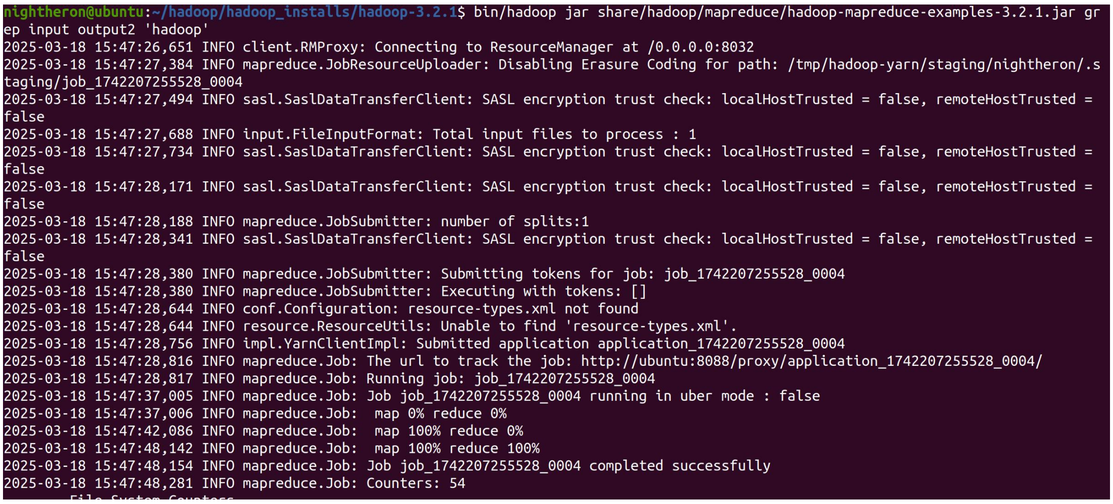


```
		Virtual memory (bytes) snapshot=24132145152
		Total committed heap usage (bytes)=7211057152
		Peak Map Physical memory (bytes)=590524416
		Peak Map Virtual memory (bytes)=2683248640
		Peak Reduce Physical memory (bytes)=316702720
		Peak Reduce Virtual memory (bytes)=2684555264
	Shuffle Errors
		BAD_ID=0
		CONNECTION=0
		IO_ERROR=0
		WRONG_LENGTH=0
		WRONG_MAP=0
		WRONG_REDUCE=0
	File Input Format Counters
		Bytes Read=944
	File Output Format Counters
		Bytes Written=97
Job Finished in 17.121 seconds
2025-03-18 13:04:00,438 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Estimated value of Pi is 3.14159800000000000000
```

3.比较 1 和 2 的执行时间，说明你的发现

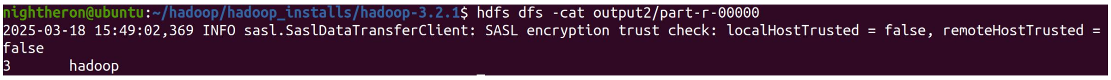
本地执行时间为 1 分 16 秒，实验平台执行时间为 16 秒，实验平台执行速度远快于本地，可见由于在集群模式下，任务会被分配到不同的节点进行并行处理，集群模式的并行处理能够显著提高执行速度。

4.本地或实验平台上运行 grep 示例程序，找出给定的 example.txt 中"hadoop"这个单词出现的次数



```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar gr
ep input output2 'hadoop'
2025-03-18 15:47:26,651 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2025-03-18 15:47:27,384 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/nightheron/.s
taging/job_1742207255528_0004
2025-03-18 15:47:27,494 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
2025-03-18 15:47:27,688 INFO input.FileInputFormat: Total input files to process : 1
2025-03-18 15:47:27,734 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
2025-03-18 15:47:28,171 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
2025-03-18 15:47:28,188 INFO mapreduce.JobSubmitter: number of splits:1
2025-03-18 15:47:28,341 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
2025-03-18 15:47:28,380 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1742207255528_0004
2025-03-18 15:47:28,380 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-03-18 15:47:28,644 INFO conf.Configuration: resource-types.xml not found
2025-03-18 15:47:28,644 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-03-18 15:47:28,756 INFO impl.YarnClientImpl: Submitted application application_1742207255528_0004
2025-03-18 15:47:28,816 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1742207255528_0004/
2025-03-18 15:47:28,817 INFO mapreduce.Job: Running job: job_1742207255528_0004
2025-03-18 15:47:37,005 INFO mapreduce.Job: Job job_1742207255528_0004 running in uber mode : false
2025-03-18 15:47:37,006 INFO mapreduce.Job:  map 0% reduce 0%
2025-03-18 15:47:42,086 INFO mapreduce.Job:  map 100% reduce 0%
2025-03-18 15:47:48,142 INFO mapreduce.Job:  map 100% reduce 100%
2025-03-18 15:47:48,154 INFO mapreduce.Job: Job job_1742207255528_0004 completed successfully
2025-03-18 15:47:48,281 INFO mapreduce.Job: Counters: 54
        File System Counters
```

hadoop 共出现 3 次



```
nightheron@ubuntu:~/hadoop/hadoop_installs/hadoop-3.2.1$ hdfs dfs -cat output2/part-r-00000
2025-03-18 15:49:02,369 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted =
false
3       hadoop
```