

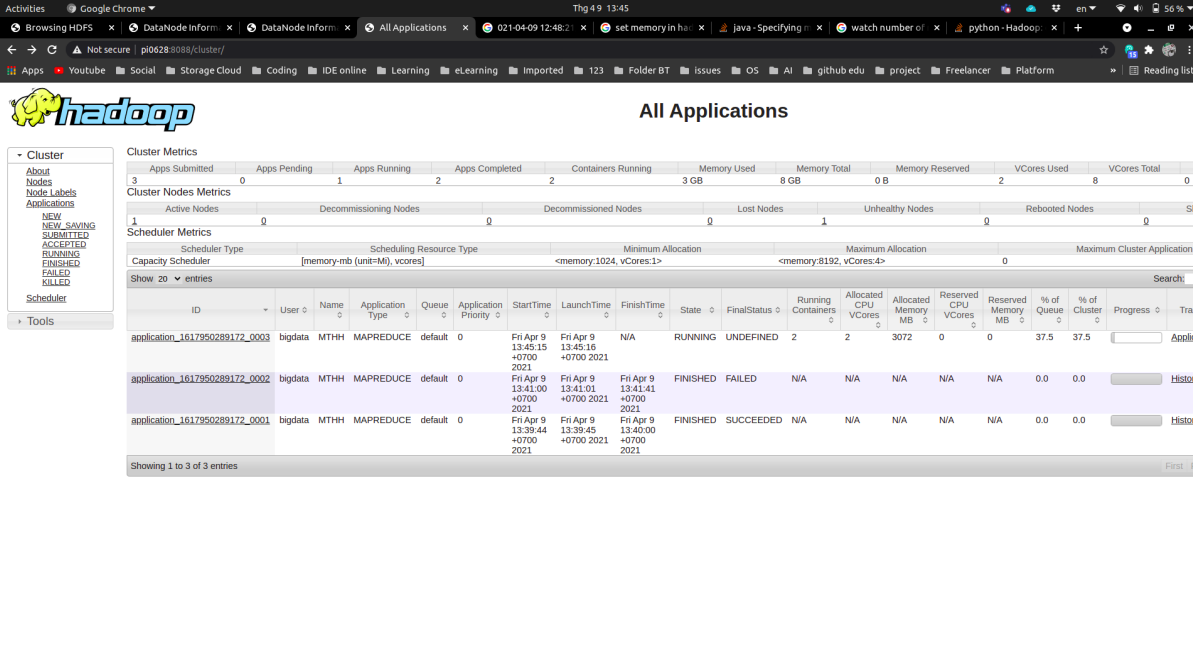
BÁO CÁO LAB2 LƯU TRỮ XỬ LÝ DỮ LIỆU LỚN : CÀI ĐẶT HADOOP VÀ CHẠY MAPREDUCE



Nhóm	MTHH	
Thành viên	Nguyễn Quang Huy	20183554
	Trần Quang Minh	20183594
	Ngô Song Việt Hoàng	20183542
	Nguyễn Văn Thanh	20183632
GVHD	TS. Đào Thành Chung	

1. Hình chụp chạy JOB

Nhóm em đã cấu hình thành công và chạy chương trình WordCount trên 3 máy. Một số hình ảnh nhóm em lưu lại để làm bằng chứng việc cài và chạy thành công (tên JOB : MTHH)



All Applications

Cluster Metrics

Apps Submitted	0	Apps Pending	1	Apps Running	2	Apps Completed	2	Containers Running	3 GB	Memory Used	8 GB	Memory Total	0 B	Memory Reserved	2	VCoers Used	8	VCoers Total	0
----------------	---	--------------	---	--------------	---	----------------	---	--------------------	------	-------------	------	--------------	-----	-----------------	---	-------------	---	--------------	---

Cluster Nodes Metrics

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

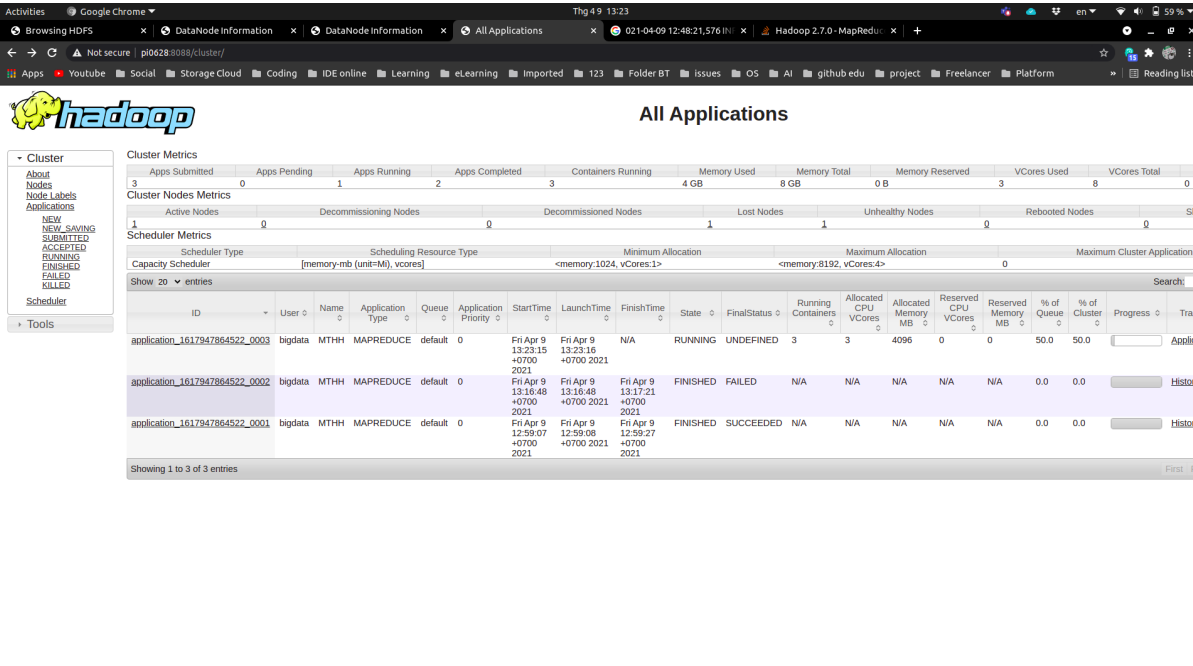
Scheduler Metrics

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

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB	Reserved CPU VCoers	Reserved Memory MB	% of Queue	% of Cluster	Progress	Track
application_1617950289172_0003	bigdata	MTHH	MAPREDUCE	default	0	Fri Apr 9 13:45:15 +0700 2021	Fri Apr 9 13:45:16 +0700 2021	N/A	RUNNING	UNDEFINED	2	2	3072	0	0	37.5	37.5		Applicat
application_1617950289172_0002	bigdata	MTHH	MAPREDUCE	default	0	Fri Apr 9 13:41:00 +0700 2021	Fri Apr 9 13:41:01 +0700 2021	Fri Apr 9 13:41:41 +0700 2021	FINISHED	FAILED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History
application_1617950289172_0001	bigdata	MTHH	MAPREDUCE	default	0	Fri Apr 9 13:39:44 +0700 2021	Fri Apr 9 13:39:45 +0700 2021	Fri Apr 9 13:40:00 +0700 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History

Showing 1 to 3 of 3 entries



All Applications

Cluster Metrics

Apps Submitted	0	Apps Pending	1	Apps Running	2	Apps Completed	3	Containers Running	4 GB	Memory Used	8 GB	Memory Total	0 B	Memory Reserved	3	VCoers Used	8	VCoers Total	0
----------------	---	--------------	---	--------------	---	----------------	---	--------------------	------	-------------	------	--------------	-----	-----------------	---	-------------	---	--------------	---

Cluster Nodes Metrics

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

Scheduler Metrics

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

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB	Reserved CPU VCoers	Reserved Memory MB	% of Queue	% of Cluster	Progress	Track
application_1617947864522_0003	bigdata	MTHH	MAPREDUCE	default	0	Fri Apr 9 13:23:15 +0700 2021	Fri Apr 9 13:23:16 +0700 2021	N/A	RUNNING	UNDEFINED	3	3	4096	0	0	50.0	50.0		Applicat
application_1617947864522_0002	bigdata	MTHH	MAPREDUCE	default	0	Fri Apr 9 13:16:48 +0700 2021	Fri Apr 9 13:16:48 +0700 2021	Fri Apr 9 13:17:21 +0700 2021	FINISHED	FAILED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History
application_1617947864522_0001	bigdata	MTHH	MAPREDUCE	default	0	Fri Apr 9 12:59:07 +0700 2021	Fri Apr 9 12:59:08 +0700 2021	Fri Apr 9 12:59:27 +0700 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History

Showing 1 to 3 of 3 entries

The screenshot shows the Hadoop web interface with the title "All Applications". On the left is a sidebar menu with options like "Cluster", "About Nodes", "Node Labels", "Applications", and "Scheduler". The main content area displays "Cluster Metrics" and a table of application jobs.

Cluster Metrics

Apps Submitted	3	Apps Pending	0	Apps Running	3	Apps Completed	0	Containers Running	0 B	Memory Used	8 GB	Memory Total	0 B	Memory Reserved	0	VCoers Used	8	VCoers Total	0	VCoers
----------------	---	--------------	---	--------------	---	----------------	---	--------------------	-----	-------------	------	--------------	-----	-----------------	---	-------------	---	--------------	---	--------

Cluster Nodes Metrics

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

Scheduler Metrics

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

Applications Table

ID	User	Name	Application Type	Queue	Application Priority	Start Time	Launch Time	Finish Time	State	Final Status	Running Containers	Allocated CPU VCoers	Allocated Memory MB	Reserved CPU VCoers	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI
application_1617950289172_0003	bigdata	MTHH	MAPREDUCE	default	0	Fri Apr 9 13:45:15 +0700 2021	Fri Apr 9 13:45:16 +0700 2021	Fri Apr 9 13:45:33 +0700 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History
application_1617950289172_0002	bigdata	MTHH	MAPREDUCE	default	0	Fri Apr 9 13:41:00 +0700 2021	Fri Apr 9 13:41:01 +0700 2021	Fri Apr 9 13:41:41 +0700 2021	FINISHED	FAILED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History
application_1617950289172_0001	bigdata	MTHH	MAPREDUCE	default	0	Fri Apr 9 13:39:44 +0700 2021	Fri Apr 9 13:39:45 +0700 2021	Fri Apr 9 13:40:00 +0700 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History

Showing 1 to 3 of 3 entries

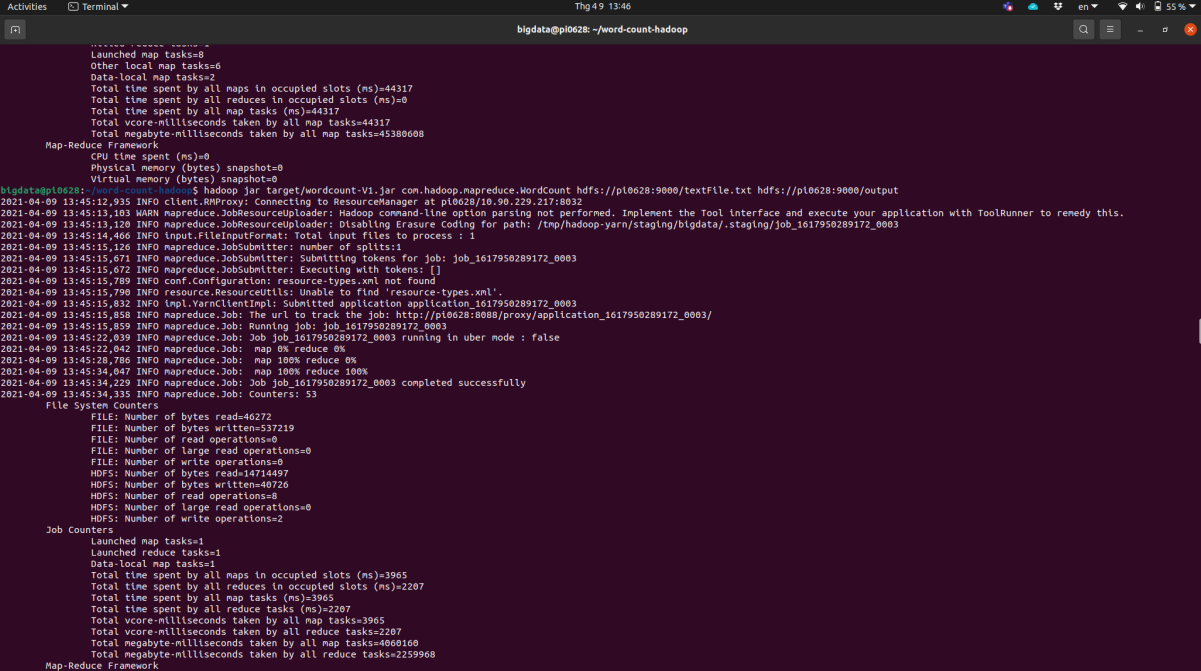
Trong hai hình trên, hình đầu tiên là chụp trong lúc chương trình đang chạy, hình thứ 2 chụp khi đã hoàn tất. (Có 3 chương trình ở đây là bởi vì nhóm em chạy chương trình với 1 số file có dung lượng nhỏ trước để xem chương trình có hoạt động không sau đó mới chạy những file có dung lượng lớn hơn)

2. Hình chụp cách chạy

Chạy file WordCount trên Hadoop sử dụng lệnh :

```
hadoop jar target/wordcount-V1.jar com.hadoop.mapreduce.WordCount  
hdfs://pi0628:9000/textFile.txt hdfs://pi0628:9000/output
```

Hình ảnh minh họa cho việc chạy câu lệnh trên :



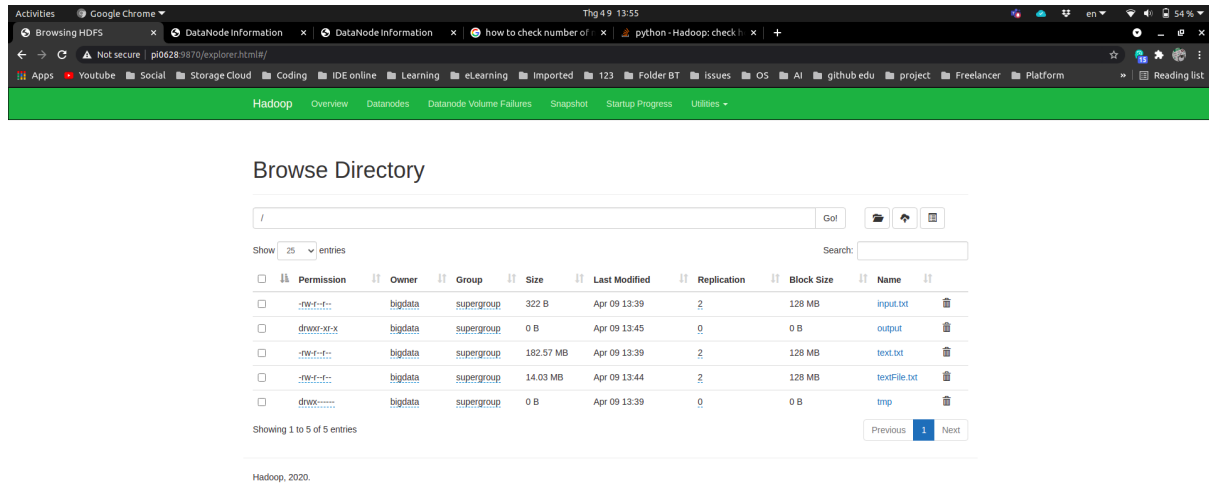
```
Launched map tasks=0  
Other local map tasks=6  
Data-local map tasks=2  
Total time spent by all maps in occupied slots (ms)=44317  
Total time spent by all reduces in occupied slots (ms)=0  
Total time spent by all map tasks (ms)=44317  
Total vcore-millisecseconds taken by all map tasks=44317  
Total megabyte-millisecseconds taken by all map tasks=45380608  
Map-Reduce Framework  
CPU time spent (ms)=0  
Physical memory (bytes) snapshot=0  
Virtual memory (bytes) snapshot=0  
bigdata@pi0628: ~/word-count-hadoop  
bigdata@pi0628:~/word-count-hadoop$ hadoop jar target/wordcount-V1.jar com.hadoop.mapreduce.WordCount hdfs://pi0628:9000/textFile.txt hdfs://pi0628:9000/output  
2021-04-09 13:45:12.935 INFO client.RMProxy: Connecting to ResourceManager at pi0628/10.90.229.217:8032  
2021-04-09 13:45:13.103 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.  
2021-04-09 13:45:13.100 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/bigdata/.staging/job_1617950289172_0003  
2021-04-09 13:45:14.466 INFO input.FileInputFormat: Total input files to process : 1  
2021-04-09 13:45:15.126 INFO mapreduce.JobSubmitter: number of splits:1  
2021-04-09 13:45:15.671 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1617950289172_0003  
2021-04-09 13:45:15.672 INFO mapreduce.JobSubmitter: Executing with tokens: []  
2021-04-09 13:45:15.789 INFO Conf.Configuration: resource-types.xml not found  
2021-04-09 13:45:15.790 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.  
2021-04-09 13:45:15.832 INFO impl.YarnClientImpl: Submitted application application_1617950289172_0003  
2021-04-09 13:45:15.858 INFO mapreduce.Job: The url to track the job: http://pi0628:8088/proxy/application_1617950289172_0003/  
2021-04-09 13:45:15.859 INFO mapreduce.Job: Running job: job_1617950289172_0003  
2021-04-09 13:45:22.039 INFO mapreduce.Job: Job job_1617950289172_0003 running in uber mode : false  
2021-04-09 13:45:22.042 INFO mapreduce.Job: map 0% reduce 0%  
2021-04-09 13:45:28.786 INFO mapreduce.Job: map 100% reduce 0%  
2021-04-09 13:45:34.047 INFO mapreduce.Job: map 100% reduce 100%  
2021-04-09 13:45:34.229 INFO mapreduce.Job: Job job_1617950289172_0003 completed successfully  
2021-04-09 13:45:34.335 INFO mapreduce.Job: Counters: 53  
File System Counters  
FILE: Number of bytes read=46272  
FILE: Number of bytes written=537219  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=14714497  
HDFS: Number of bytes written=40726  
HDFS: Number of read operations=8  
HDFS: Number of large read operations=0  
HDFS: Number of write operations=2  
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)=3965  
Total time spent by all reduces in occupied slots (ms)=2207  
Total time spent by all map tasks (ms)=3965  
Total time spent by all reduce tasks (ms)=2207  
Total vcore-millisecseconds taken by all map tasks=3965  
Total vcore-millisecseconds taken by all reduce tasks=2207  
Total megabyte-millisecseconds taken by all map tasks=4609160  
Total megabyte-millisecseconds taken by all reduce tasks=2259968  
Map-Reduce Framework
```

Trong đó :

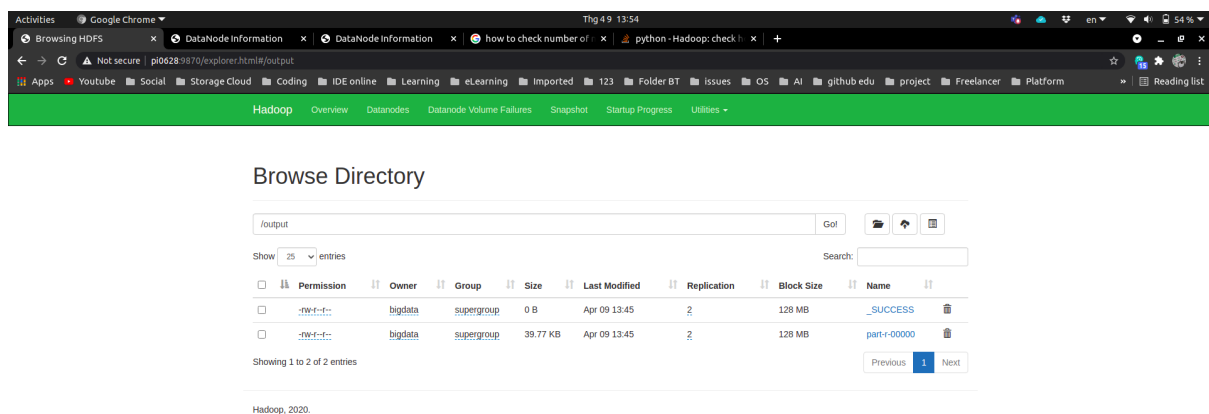
- target/wordcount-V1.jar là đường dẫn tới file jar tại nơi terminal
- com.hadoop.mapreduce.WordCount là đường dẫn tới class chứa hàm main
- hdfs://pi0628:9000/textFile.txt là đường dẫn file đầu vào (pi0628 là tên node master)
- hdfs://pi0628:9000/output là đường dẫn file đầu ra

3. Hình chụp kết quả chạy

Sau khi chạy xong một thư mục output là kết quả được sinh ra trên HDFS như sau :



Bên trong thư mục output sẽ là 2 file, 1 file là part-r-00000 chứa kết quả, 1 file là _SUCCESS không chứa gì cả :



Xem kết quả của chương trình WordCount trong file
part-r-00000 chạy lệnh :

```
hdfs dfs -cat /output/part-r-00000
```

Kết quả sau khi chạy lệnh trên ra là :

```
Activities Terminal Thg 4/9 13:46
bigdata@pi0628: ~/word-count-hadoop

Bytes Read=14714401
File Output Format Counters
Bytes Written=40726
bigdata@pi0628: ~/word-count-hadoop$ hdfs dfs -cat /output/part-r-00000
"AS" 200
"AS" 2200
"AS-Is" 100
"Adaptation" 100
"Copyrights" 100
"Collection" 100
"Collective" 100
"Contribution" 200
"Contributor" 200
"Creative" 100
"Derivative" 200
"Distribute" 100
"French" 200
"JDOM" 200
"JDOM" 100
"Java" 100
"License(s)" 200
"Legal" 100
"License" 100
"License(s)" 200
"Licensed" 100
"Licensors" 300
"Losses" 100
"Notice" 100
"Not" 100
"Object" 100
"Original" 200
"Program" 100
"Publicly" 100
"Recipient" 100
"Reproduce" 100
"Screenplay" 200
"Software(s)" 700
"Source" 100
"The" 200
"This" 100
"Work" 300
"You" 300
"Your" 100
"[]" 100
"Control" 100
"Printed" 100
"Submitted" 100
"Submitted" 100
"Submitted" 100
"Originates" 100
("AGREEMENT"). 100
("Attribution" 100
("CCPL" 200
("Commercial" 100
("Indemnified" 100
"Cover" 100
("Synchronizing" 200
```

```
Activities Terminal Thg 4/9 13:56
bigdata@pi0628: ~/word-count-hadoop

whole, 200
whole. 200
whom 1000
whose 100
will 3100
were 100
wireless 100
with 15100
with. 100
withdraw 400
within 2100
without 7000
words; 100
work 3600
work, 800
work. 300
works; 200
works 1000
works, 400
world-wide, 500
worldwide, 600
would 500
writing 500
writing, 400
writing; 100
written 2100
year 100
you 1600
your 400
252.2277014(a)(1)) 100
$ 100
"AS" 100
"Contribution" 100
"Contributor" 200
"Contributor" 200
"Covered" 200
"Executable" 100
"Executable" 100
"Incompatible" 300
"Initial" 100
"Larger" 200
"Licenseable" 200
"License" 200
"Modifications" 200
"Original" 100
"Participant" 100
"Patent" 200
"Secondary" 100
"Source" 200
"Your" 200
"Your" 400
"as" 100
"commercial" 300
"control" 200
bigdata@pi0628: ~/word-count-hadoop$
```