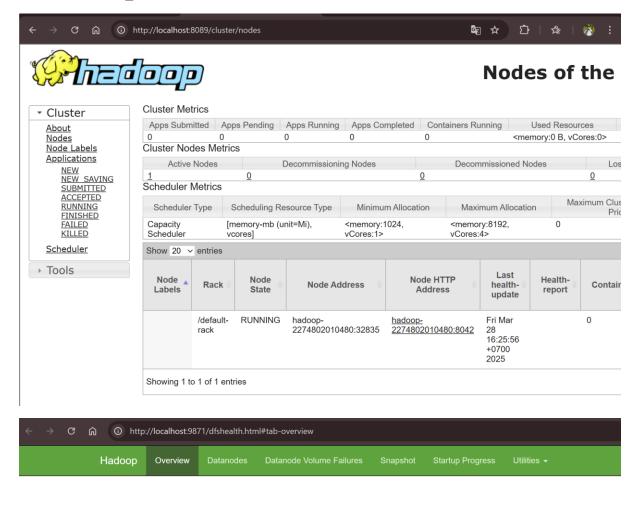
2274802010480_Tran Khanh Linh



Overview 'localhost:9000' (ractive)

Started:	Fri Mar 28 16:38:00 +0700 2025
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled:	Mon Mar 04 13:35:00 +0700 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-f3690cda-0a60-4603-b957-5bf5affd7357
Block Pool ID:	BP-608357847-172.23.0.76-1743154673555

LAB1:HDFS

Copy_to_local:

```
from snakebite.client import Client

client = Client('localhost', 9000)
for f in client.copyToLocal(['/input/input.txt'], '/tmp'):
    print f
```

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$ python2 copy_to_local.py {'path': '/tmp/input.txt', 'source_path': '/input/input.txt', 'result': True, 'error': ''} hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$ cat /tmp/input.txt jack be nimble jack be quick jack jumped over the candlestickhadoop@hadoop-2274802010480:~/ih/HadoopWithPython/python/HDFS$ h
```

Delete.py

```
from snakebite.client import Client

client = Client('localhost', 9000)
for p in client.delete(['/foo', '/input'], recurse=True):
    print p
```

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$ python2 delete.py {'path': '/foo', 'result': True} {'path': '/input', 'result': True} hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$
```

Get-pip.py

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc han
h/HadoopWithPython/python/HDFS$ python2 get-pip.py
DEPRECATION: Python 2.7 reached the end of its life on January
1st, 2020. Please upgrade your Python as Python 2.7 is no longe
r maintained. pip 21.0 will drop support for Python 2.7 in Janu
ary 2021. More details about Python 2 support in pip can be fou
nd at https://pip.pypa.io/en/latest/development/release-process
/#python-2-support pip 21.0 will remove support for this functi
onality.
Defaulting to user installation because normal site-packages is
 not writeable
Collecting pip<21.0
  Using cached pip-20.3.4-py2.py3-none-any.whl (1.5 MB)
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 20.3.4
    Uninstalling pip-20.3.4:
      Successfully uninstalled pip-20.3.4
Successfully installed pip-20.3.4
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc han
h/HadoopWithPython/python/HDFS$
```

List directory.py

```
from snakebite.client import Client

client = Client('localhost', 9000)
for x in client.ls(['/']):
    print x
```

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$ python2 list_directory.py {'group': u'supergroup', 'permission': 493, 'file_type': 'd', 'access_time': 0L, 'block_replication': 0, 'modification_time': 1743155609645L, 'length': 0L, 'blocksize': 0L, 'owner': u'hadoop', 'path': '/input'} hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$
```

List file.py:

```
from snakebite.client import Client

client = Client('localhost', 9000)
for x in client.ls(['/input.txt']):
    print x
```

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$ python2 list_file.py {'group': u'supergroup', 'permission': 420, 'file_type': 'f', 'access_time': 1743156439018L, 'block_replication': 1, 'modifica tion_time': 1743156439233L, 'length': 61L, 'blocksize': 1342177 28L, 'owner': u'hadoop', 'path': '/input.txt'} hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$
```

Mkdir.py:

```
from snakebite.client import Client

client = Client('localhost', 9000)
for p in client.mkdir(['/foo/bar', '/input'], create_parent=True):
    print p
```

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$ python2 mkdir.py {'path': '/foo/bar', 'result': True} {'path': '/input', 'result': False, 'error': "mkdir: `/input': File exists"} hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$
```

Text.py:

```
from snakebite.client import Client

client = Client('localhost', 9000)
for l in client.text(['/input/input.txt']):
    print l
```

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$ python2 text.py jack be nimble jack be quick jack jumped over the candlestick hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_han h/HadoopWithPython/python/HDFS$
```

LAB2 Luigui:

Wordcount.py:

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc han
h/HadoopWithPython/python/Luigi$ python2 wordcount.py WordCount
--input-file input.txt --output-file wordcount output.txt --lo
cal-scheduler
DEBUG: Checking if WordCount(input_file=input.txt, output_file=
wordcount output.txt) is complete
DEBUG: Checking if InputFile(input_file=input.txt) is complete
INFO: Informed scheduler that task
                                     WordCount input txt wordco
unt output 0a6198752c
                        has status
                                     PENDING
INFO: Informed scheduler that task
                                     InputFile_input_txt_fd27c1
       has status
2e07
                    DONE
INFO: Done scheduling tasks
INFO: Running Worker with 1 processes
```

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc han
h/HadoopWithPython/python/Luigi$ cat wordcount output.txt
be
jumped
        1
over
        1
candlestick
                1
nimble 1
jack
        3
quick
        1
the
        1
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc han
h/HadoopWithPython/python/Luigi$
```

Luigi_pig.py:

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/Luigi$ python2 -m luigi --module luig i_pig WordCount --input-file /user/hadoop/input.txt --output-file /user/hadoop/output --scheduler-host localhost --scheduler-p ort 8084

DEBUG: Checking if WordCount(input_file=/user/hadoop/input.txt, output_file=/user/hadoop/output, script_path=/home/hadoop/iDra gonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/pig/wordcount.pig) is complete

DEBUG: Running file existence check: hadoop fs -stat /user/hadoop/output

INFO: Informed scheduler that task WordCount_user_hadoop_inp_user_hadoop_out_home_hadoop_iDr_10ee23b60a has status DO

NE

INFO: Done scheduling tasks

INFO: Running Worker with 1 processes

DEBUG: Asking scheduler for work...

DEBUG: Done

DEBUG: There are no more tasks to run at this time

INFO: Worker Worker(salt=695430859, workers=1, host=hadoop-2274802010480, username=hadoop, pid=20892) was stopped. Shutting do wn Keep-Alive thread

INFO:

===== Luigi Execution Summary =====
```

Luigi_mapreduce.py:

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/Luigi$ python2 luigi_mapreduce.py --i
nput-file hdfs:///user/hadoop/input.txt --output-file hdfs:///user/hadoop/output/ --local-scheduler
No handlers could be found for logger "luigi-interface"
DEBUG: Checking if WordCount(input_file=hdfs:///user/hadoop/input.txt, output_file=hdfs:///user/hadoop/output/) is complete
DEBUG: Running file existence check: hadoop fs -stat hdfs:///user/hadoop/output/
INFO: Informed scheduler that task WordCount_hdfs___user_had_hdfs___user_had_c64bc50fc7 has status DONE
INFO: Done scheduling tasks
INFO: Running Worker with 1 processes
DEBUG: Asking scheduler for work...
DEBUG: Done
DEBUG: There are no more tasks to run at this time
INFO: Worker Worker(salt=587456706, workers=1, host=hadoop-2274802010480, username=hadoop, pid=36491) was stopped. Shutting do
wn Keep-Alive thread
INFO:
==== Luigi Execution Summary =====
Scheduled 1 tasks of which:
  1 complete ones were encountered:
     - 1 WordCount(input_file=hdfs:///user/hadoop/input.txt, output_file=hdfs:///user/hadoop/output/)
Did not run any tasks
This progress looks :) because there were no failed tasks or missing dependencies
==== Luigi Execution Summary =====
```

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/Luigi$ hdfs dfs -cat /user/hadoop/output/part-*

be

the

jack

over

juick

jumped

nimble

candlestick
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/Luigi$

hdfs dfs -cat /user/hadoop/output/part-New / value / value
```

LAB3: MapReduce

HadoopStreaming

Mapper.py:

Reducer.py:

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/MapReduce/HadoopStreaming$ hdfs dfs - cat /user/hadoop/input.txt | python2 mapper.py | sort > mapper_output.txt cat mapper_output.txt | python2 reducer.py be 2 candlestick 1 jack 3 jumped 1 nimble 1 over 1 quick 1 the 1 hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/MapReduce/HadoopStreaming$
```

Mrjob

Word count.py

```
TERMINAL PORTS (2)
                                                                                                                                            PROBLEMS 14 OUTPUT DEBUG CONSOLE
"INSPECTIONS, A49102, TRANS-Highways"
"INSPECTOR"
                        199
"INSPECTOR, A06011, Housing"
"INSPECTOR, A06013, Housing"
"INSPECTOR, A06014, Housing"
"INSPECTOR, A06038, Housing"
"INSPECTOR, A06043, Housing"
                                                 12
"INSPECTOR, A06083, Housing
                                                 11
"INSPECTOR, A14003, FIN-Collections"
"INSPECTOR, A50804, DPW-Water"
"INSTALLATI,A49506,TRANS-Highways"
"INSTALLER" 133
"INSTRUCT,A49102,TRANS-Highways"
"INSTRUCTOR,A04009,R&P-Recreation"
"INSTRUCTOR, A04010, R&P-Recreation"
"INSTRUCTOR, A06709, Housing" 2

"INSTRUCTOR, A065200, HLTH-Health" 2

"INSTRUCTOR, A67002, R&P-Administration"

"INSTRUCTOR, A67003, R&P-Administration"

"INSTRUCTOR, A67005, R&P-Administration"
"INSTRUCTOR, A6701, R&P-Administration"
"INSTRUCTOR, A6701, R&P-Administration"
"INSTRUCTOR, A67012, R&P-Administration"
"INSTRUCTOR, A68005, R&P-Parks" 17
"INSTRUCTOR, A68011, R&P-Parks" 1
"INSTRUCTOR, B68040, R&P-Parks"
"INSTRUCTOR, B68051, R&P-Parks"
"INSTRUCTOR, P04001, R&P-Recreation"
Removing temp directory /tmp/word_count.hadoop.20250329.163409.891759...
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/MapReduce/mrjob$ python2 word_count.p
y salaries.csv
```

Top_salary.py

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/MapReduce/mrjob$ python2 top_salary.p
y salaries.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/top_salary.hadoop.20250329.163513.936739
Running step 1 of 1...
Counters: 1
                       missing gross=3223
job output is in /tmp/top_salary.hadoop.20250329.163513.936739/output
Streaming final output from /tmp/top_salary.hadoop.20250329.163513.936739/output...
"salary" [163200.0, "\"Chow,Rudolph S\",DIRECTOR PUBLIC WORKS,A41101,DPW-Administration ,02/01/2011,$163200.00,$145513.
"salary
79"]
"salary"
                       [163200.0, "\"Nilson,George A\",CITY SOLICITOR,A30001,Law Department ,01/16/2007,$163200.00,$164332.32"]
[163365.0, "\"Rawlings-Blake,Stephanie C\",MAYOR,A01001,Mayors Office ,12/07/1995,$163365.00,$161219.24"]
[165000.0, "\"Ford,Niles R\",EXECUTIVE LEVEL III,A64006,Fire Department ,01/15/2014,$165000.00,$69807.64"]
[172000.0, "\"Parthemos,Kaliope\",EXECUTIVE LEVEL III,A01020,Mayor's Office ,12/26/2006,$172000.00,$154654.39"
"salary'
"salary"
"salary"
                       [187200.0, "\"Swift,Michael\",CONTRACT SERV SPEC II,A02003,City Council ,05/19/2008,$187200.00,$3510.00"]
[190000.0, "\"Black,Harry E\",EXECUTIVE LEVEL III,A23001,FIN-Admin & Budgets ,01/30/2012,$190000.00,$188328.50
"salary"
"salary
"salary"
                       [193800.0, "\"Batts,Anthony W\",EXECUTIVE LEVEL III,A99390,Police Department ,09/25/2012,$193800.00,$193653.69
"salary"
                       [200000.0, "\"Charles,Ronnie E\",EXECUTIVE LEVEL III,A83001,HR-Human Resources ,07/05/2012,$200000.00,$185741.
```

Avg.py:

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/MapReduce/mrjob$ python2 avg.py salar
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/avg.hadoop.20250329.163640.264128
Running step 1 of 2...
Counters: 65
        depts
                  COMP-Audits=40
                  COMP-Communication Ser=30
                  COMP-Comptroller's O=12
                  COMP-Real Estate=8
                  Circuit Court=140
                  City Council=93
                  Civil Rights & Wage Enfor=12
                  Council Services=6
                  DPW-Administration=116
                  DPW-Solid Waste=255
                  DPW-Solid Waste (wkly)=475
                  DPW-Water & Waste Water=1491
                  ERS/EOS Admin=33
                  Elections=8
                  Enoch Pratt Free Library=463
                  FIN-Acct & Payroll=52
                  FIN-Admin & Budgets=31
                  FIN-Collections=129
                  FIN-Purchasing=52
```

Lab4: Spark

Word_count.py

hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/Spark\$

Text_search.py:

('jumped', 1) ('over', 1) ('the', 1)

```
hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/Spark$ spark-submit text_search.py
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/jars/spark-unsafe_2.12-3.1.2.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations WARNING: All illegal access operations will be denied in a future release
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
25/03/30 01:36:25 INFO SparkContext: Running Spark version 3.1.2
25/03/30 01:36:25 INFO ResourceUtils:
25/03/30 01:36:25 INFO ResourceUtils: No custom resources configured for spark.driver.
25/03/30 01:36:25 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cores, amount: 1.0)
25/03/30 01:36:35 INFO ResourceProfile: initialized the cores amount: 1.0)
, vendor: ), task resources: Map(cpus -> name. cpus, amount. 25/03/30 01:36:25 INFO ResourceProfile: Limiting resource is cpu
25/03/30 01:36:25 INFO ResourceProfileManager: Added ResourceProfile id: 0
Toy Story (1995)
25/03/30 01:36:29 INFO SparkUI: Stopped Spark web UI at http://hadoop-2274802010480:4040
 25/03/30 01:36:29 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/03/30 01:36:29 INFO MemoryStore: MemoryStore cleared 25/03/30 01:36:29 INFO BlockManager: BlockManager stopped
 25/03/30 01:36:29 INFO BlockManagerMaster: BlockManagerMaster stopped
 25/03/30 01:36:29 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
 25/03/30 01:36:29 INFO SparkContext: Successfully stopped SparkContext
 25/03/30 01:36:30 INFO ShutdownHookManager: Shutdown hook called
 25/03/30 01:36:30 INFO ShutdownHookManager: Deleting directory /tmp/spark-d4a7bd54-d60c-492a-a303-da64fcc9240a/pyspark-4c0bb80
 9-462c-4e34-aa47-a5948db51944
 25/03/30 01:36:30 INFO ShutdownHookManager: Deleting directory /tmp/spark-d4a7bd54-d60c-492a-a303-da64fcc9240a
25/03/30 01:36:30 INFO ShutdownHookManager: Deleting directory /tmp/spark-f2d13367-7f18-4396-a11e-b9c022cb9435 hadoop@hadoop-2274802010480:~/iDragonCloud/71ITDS40303/Thuc_hanh/HadoopWithPython/python/Spark$
△ 11 (♠) 4
                                                     ♦ Zachary Radtka (9 years ago) Ln 28, Col 50 Spaces: 3 UTF-8 LF
                                                                                                                     1 Matching titles found:
 Dead Man Walking (1995)
 25/03/30 01:38:01 INFO SparkUI: Stopped Spark web UI at http://hadoop-2274802010480:4040
 25/03/30 01:38:01 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
 25/03/30 01:38:01 INFO MemoryStore: MemoryStore cleared
 25/03/30 01:38:01 INFO BlockManager: BlockManager stopped
 25/03/30 01:38:01 INFO BlockManagerMaster: BlockManagerMaster stopped
 25/03/30 01:38:01 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoor
```