

雲端計算與網路 Homework

Hadoop

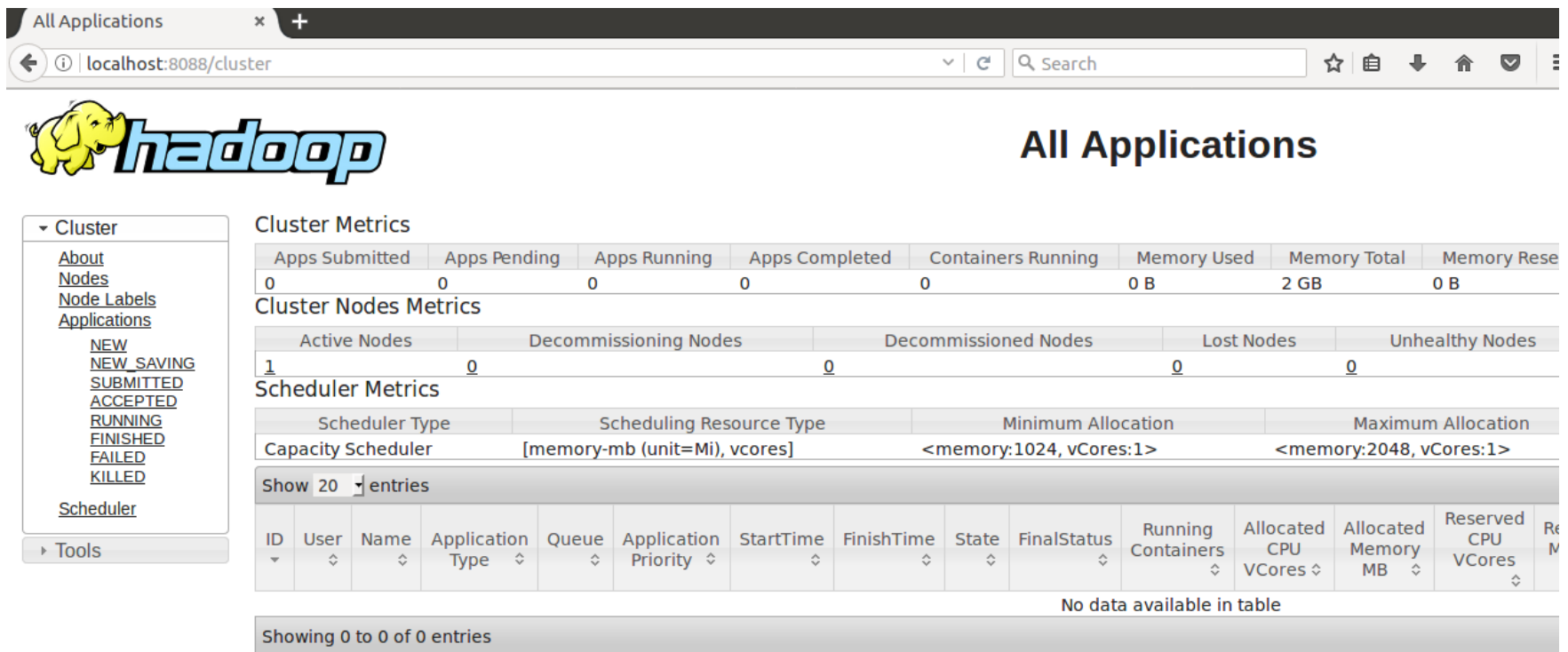
- Install Hadoop (30%)

\$ jps

```
cjp@cjp-VirtualBox:~$ jps
4630  DataNode
5127  ResourceManager
5292  NodeManager
4461  NameNode
4893  SecondaryNameNode
5839  Jps
cjp@cjp-VirtualBox:~$
```

Hadoop

- Using your browser to connect to
 - `http://localhost:8088` (screen shots)



The screenshot displays the Hadoop web interface at `localhost:8088`. The page title is "All Applications". On the left, there is a navigation menu with links for "Cluster", "About", "Nodes", "Node Labels", "Applications", and "Scheduler". The "Cluster" section is expanded, showing a list of application states: NEW, NEW_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, and KILLED. The "Scheduler" section is also visible.

The main content area displays several metrics:

- Cluster Metrics:** A table showing the number of applications in various states and memory usage.
- Cluster Nodes Metrics:** A table showing the number of active, decommissioning, decommissioned, lost, and unhealthy nodes.
- Scheduler Metrics:** A table showing the scheduler type, scheduling resource type, and minimum/maximum allocations.

Below the metrics, there is a table of application entries. The table has columns for ID, User, Name, Application Type, Queue, Application Priority, StartTime, FinishTime, State, FinalStatus, Running Containers, Allocated CPU VCores, Allocated Memory MB, Reserved CPU VCores, and Reserved Memory MB. The table is currently empty, showing "Showing 0 to 0 of 0 entries".

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Rese
0	0	0	0	0	0 B	2 GB	0 B

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

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=M), vcores]	<memory:1024, vCores:1>	<memory:2048, vCores:1>

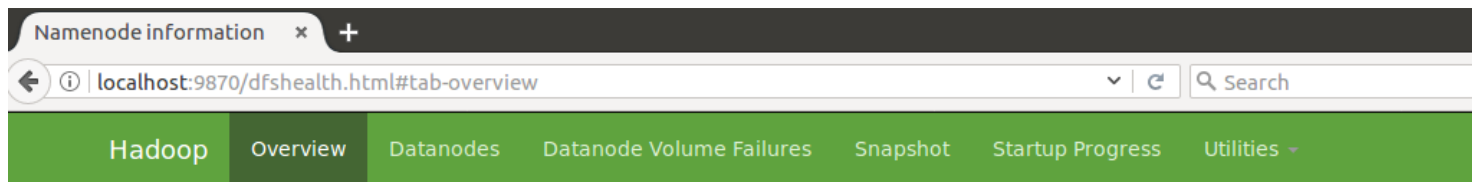
ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB
----	------	------	------------------	-------	----------------------	-----------	------------	-------	-------------	--------------------	----------------------	---------------------	---------------------	--------------------

No data available in table

Showing 0 to 0 of 0 entries

Hadoop

- Using your browser to connect to
 - <http://localhost:9870/> (screen shots)



Overview 'localhost:9000' (active)

Started:	Tue Mar 20 03:17:50 -0700 2018
Version:	3.0.0, rc25427ceca461ee979d30edd7a4b0f50718e6533
Compiled:	Fri Dec 08 11:16:00 -0800 2017 by andrew from branch-3.0.0
Cluster ID:	CID-17f54c6c-1210-48b4-974c-030aa0372c54
Block Pool ID:	BP-335749269-127.0.1.1-1521540992806

Summary

Security is off.

Hadoop

- Run WordCount.java (10%)

- Input :

- file1.txt
 - file2.txt
 - file3.txt

```
cjp@cjp-VirtualBox:~$ hdfs dfs -cat /data/file1.txt
2020-09-29 16:47:01,422 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
aaa bbb ccc
ddd
eee

cjp@cjp-VirtualBox:~$ hdfs dfs -cat /data/file2.txt
2020-09-29 16:53:06,379 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
aa zzz
yyy
xxx
bbb
ii oo

cjp@cjp-VirtualBox:~$ hdfs dfs -cat /data/file3.txt
2020-09-29 16:53:16,963 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
a ttt
ccc
bbb
kkk
jjj
```

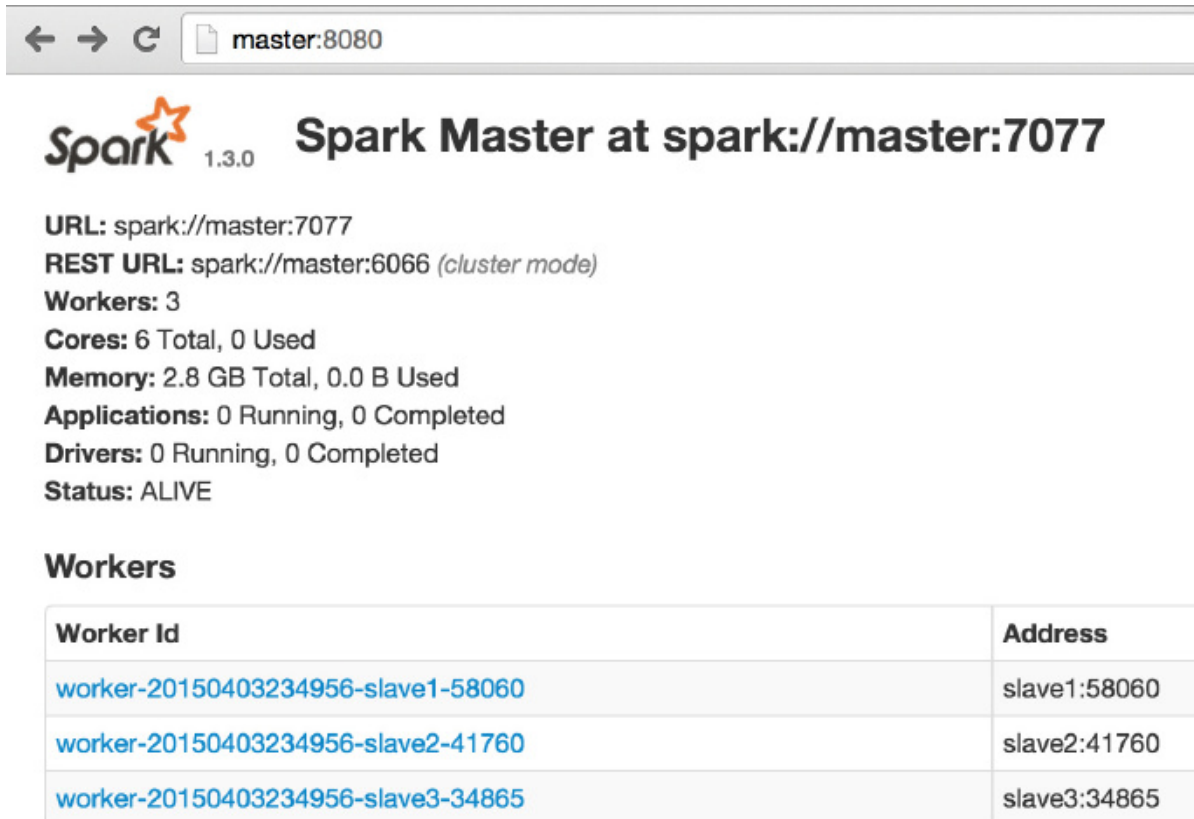
Hadoop

- Output : part-r-00000

```
cjp@cjp-VirtualBox: ~  
File Edit View Search Terminal Help  
cjp@cjp-VirtualBox:~$ hdfs dfs -ls /out/p0  
Found 2 items  
-rw-r--r--    1 cjp supergroup          0 2020-09-29 22:17 /out/p0/_SUCCESS  
-rw-r--r--    1 cjp supergroup       85 2020-09-29 22:17 /out/p0/part-r-00000  
cjp@cjp-VirtualBox:~$ hdfs dfs -cat /out/p0/part-r-00000  
2020-09-29 22:18:38,492 INFO sasl.SaslDataTransferClient: SASL encryption trust  
check: localhostTrusted = false, remoteHostTrusted = false  
a          1  
aa         1  
aaa        1  
bbb        3  
ccc        2  
ddd        1  
eee        1  
ii         1  
jjj        1  
kkk        1  
oo         1  
ttt        1  
xxx        1  
yyy        1  
zzz        1  
cjp@cjp-VirtualBox:~$
```

Spark

- Install Spark (30%)
- Using your browser to connect to
 - <http://master:8080> (screen shots)



Spark 1.3.0 Spark Master at spark://master:7077

URL: spark://master:7077
REST URL: spark://master:6066 (*cluster mode*)
Workers: 3
Cores: 6 Total, 0 Used
Memory: 2.8 GB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Worker Id	Address
worker-20150403234956-slave1-58060	slave1:58060
worker-20150403234956-slave2-41760	slave2:41760
worker-20150403234956-slave3-34865	slave3:34865

Spark

- Run wordcount.py (10%)
 - Input : file1.txt, file2.txt, file3.txt
 - Output : part-000000

```
cjp@cjp-VirtualBox:~$ hdfs dfs -ls /out/p1
Found 2 items
-rw-r--r--    3 cjp supergroup          0 2020-09-29 17:04 /out/p1/_SUCCESS
-rw-r--r--    3 cjp supergroup       160 2020-09-29 17:04 /out/p1/part-000000
cjp@cjp-VirtualBox:~$ hdfs dfs -cat /out/p1/part-000000
2020-09-29 17:05:31,031 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHost
Trusted = false, remoteHostTrusted = false
('aaa', 1)
('xxx', 1)
('kkk', 1)
('jjj', 1)
('bbb', 3)
('eee', 1)
('zzz', 1)
('yyy', 1)
('ii', 1)
('ttt', 1)
('ccc', 2)
('ddd', 1)
('aa', 1)
('oo', 1)
('a', 1)
```


Spark

- Sort words in ascending order (5%)
 - Input : file1.txt, file2.txt, file3.txt
 - Output : part-00000

```
cjp@cjp-VirtualBox:~$ hdfs dfs -cat /out/p3/part-00000
2020-09-29 17:20:36,600 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localhostTrusted = false, remoteHostTrusted = false
('a', 1)
('aa', 1)
('aaa', 1)
('bbb', 3)
('ccc', 2)
('ddd', 1)
('eee', 1)
('ii', 1)
('jjj', 1)
('kkk', 1)
('oo', 1)
('ttt', 1)
('xxx', 1)
('yyy', 1)
('zzz', 1)
```

Spark

- Sort word count in descending order (5%)
 - Input : file1.txt, file2.txt, file3.txt
 - Output : part-00000

```
cjp@cjp-VirtualBox:~$ hdfs dfs -cat /out/p2/part-00000
2020-09-29 17:17:14,884 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localhostTrusted = false, remoteHostTrusted = false
('bbb', 3)
('ccc', 2)
('aaa', 1)
('xxx', 1)
('kkk', 1)
('jjj', 1)
('eee', 1)
('zzz', 1)
('yyy', 1)
('ii', 1)
('ttt', 1)
('ddd', 1)
('aa', 1)
('oo', 1)
('a', 1)
```

Spark

- Inverted index (10%)
 - Input : file1.txt, file2.txt, file3.txt
 - Output : part-00000

```
cjp@cjp-VirtualBox:~$ hdfs dfs -ls /out/p4
Found 2 items
-rw-r--r--   3 cjp supergroup      0 2020-09-29 17:23 /out/p4/_SUCCESS
-rw-r--r--   3 cjp supergroup    898 2020-09-29 17:23 /out/p4/part-00000
cjp@cjp-VirtualBox:~$ hdfs dfs -cat /out/p4/part-00000
2020-09-29 17:24:26,279 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
(u'aa', [u'hdfs://localhost:9000/data/file2.txt'])
(u'oo', [u'hdfs://localhost:9000/data/file2.txt'])
(u'aaa', [u'hdfs://localhost:9000/data/file1.txt'])
(u'zzz', [u'hdfs://localhost:9000/data/file2.txt'])
(u'xxx', [u'hdfs://localhost:9000/data/file2.txt'])
(u'yyy', [u'hdfs://localhost:9000/data/file2.txt'])
(u'bbb', [u'hdfs://localhost:9000/data/file1.txt', u'hdfs://localhost:9000/data/file2.txt', u'hdfs://localhost:9000/data/file3.txt'])
(u'ttt', [u'hdfs://localhost:9000/data/file3.txt'])
(u'ii', [u'hdfs://localhost:9000/data/file2.txt'])
(u'kkk', [u'hdfs://localhost:9000/data/file3.txt'])
(u'jjj', [u'hdfs://localhost:9000/data/file3.txt'])
(u'a', [u'hdfs://localhost:9000/data/file3.txt'])
(u'eee', [u'hdfs://localhost:9000/data/file1.txt'])
(u'ccc', [u'hdfs://localhost:9000/data/file1.txt', u'hdfs://localhost:9000/data/file3.txt'])
(u'ddd', [u'hdfs://localhost:9000/data/file1.txt'])
```

Homework

- Submit **code**, **output file** and **word** to nchuwccclab@gmail.com
 - screen shots of results
 - explain in detail what parts were done and the results

學號姓名.zip (ex: 9876543210陳OO.zip)

Word.doc

Output

p1, p2, p3, p4, p5

Source code

p1.java, p2.py, p3.py, p4.py, p5.py

- **Deadline 2020/11/04 23:59**