

Assignment 1: Hadoop MapReduce exercises

611121203 張茗溱

- 打開終端機輸入指令 start-all.sh 打開 hadoop

```
hadoop@ubuntu22: ~  
hadoop@ubuntu22:~$ start-all.sh  
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.  
WARNING: This is not a recommended production deployment configuration.  
WARNING: Use CTRL-C to abort.  
Starting namenodes on [localhost]  
Starting datanodes  
Starting secondary namenodes [ubuntu22]  
Starting resourcemanager  
Starting nodemanagers
```

1. Sorting algorithm

- Input data：將隨機 50 組數字以逗點隔開寫入 data.txt

```
1 225, 413, 88, 322, 17, 95, 271, 157, 249, 33, 444, 118, 382, 74, 197, 310, 490, 43, 280, 131, 419, 60, 366, 181, 277, 29, 462, 106, 234, 69, 395, 126, 490,  
241, 153, 8, 398, 200, 463, 355, 274, 122, 50, 442, 310, 183, 32, 437, 107, 4905
```

- 打開位置在 sorting 目錄上的終端機並在 hdfs 上建立新的目錄 HW1

```
hadoop@ubuntu22: ~/HW/sorting  
hadoop@ubuntu22:~/HW/sorting$ hdfs dfs -mkdir /HW1
```

- 在 HW1 目錄上建立一個新目錄 sorting

```
hadoop@ubuntu22:~/HW/sorting$ hdfs dfs -mkdir /HW1/sorting
```

- 將本地上的 data.txt 複製到 hdfs 裡的/HW1/sorting 上

```
hadoop@ubuntu22:~/HW/sorting$ hdfs dfs -copyFromLocal data.txt /HW1/sorting
```

- 打開瀏覽器並前往 localhost:9870 裡查看 sorting 資料夾是否被建立

Browse Directory

/HW1 Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	Oct 23 16:50	0	0 B	searching
drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 20:02	0	0 B	sorting
drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 19:53	0	0 B	tfidf

Showing 1 to 3 of 3 entries Previous 1 Next

- 點選 sorting 資料夾查看是否 input 資料已成功放上

Browse Directory

/HW1/sorting Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	236 B	Oct 20 20:02	1	128 MB	data.txt
drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 20:03	0	0 B	result

Showing 1 to 2 of 2 entries Previous 1 Next

- 撰寫 run.sh

```
run.sh
~/HW/sorting
1 hadoop jar '/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4-jar' \
2 -mapper 'python mapper.py' \
3 -file /home/hadoop/HW/sorting/mapper.py \
4 -reducer 'python reducer.py' \
5 -file /home/hadoop/HW/sorting/reducer.py \
6 -input hdfs://HW1/sorting/data.txt \
7 -output /HW1/sorting/result
```

- 執行 run.sh

hadoop@ubuntu22:~/HW/sorting\$./run.sh

- 到 localhost:9870 網站上/HW1/sorting/result 資料夾下查看結果
 - _SUCCESS 為成功執行訊息
 - part-0000 為程式運行結果

/HW1/sorting/result Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Oct 20 20:03	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	237 B	Oct 20 20:03	1	128 MB	part-0000

Showing 1 to 2 of 2 entries Previous 1 Next

- Output data：input 的 50 個數字由小到大排序

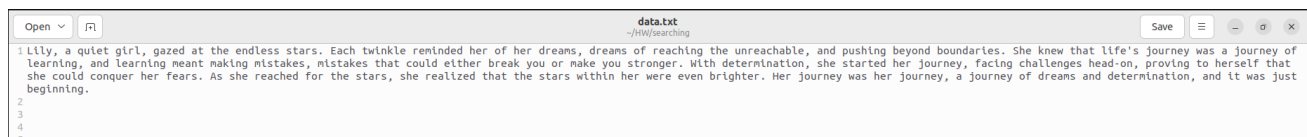
```

1 8
2 17
3 29
4 32
5 33
6 43
7 50
8 60
9 69
10 74
11 88
12 95
13 106
14 107
15 118
16 122
17 126
18 131
19 153
20 157
21 181
22 183
23 197
24 200
25 225
26 234
27 241
28 249
29 271
30 274
31 277
32 280
33 310
34 310
35 322
36 355
37 366
38 382
39 395
40 398
41 413
42 419
43 437
44 442
45 444
46 462
47 463

```

2. Searching algorithm

- Input data：將一段文章寫入 data.txt



- 在 HW1 目錄上建立一個新目錄 searching

```
hadoop@ubuntu22:~/HW/searching$ hdfs dfs -mkdir /HW1/searching
```

- 將本地上的 data.txt 複製到 hdfs 裡的/HW1/searching 上

```
hadoop@ubuntu22:~/HW/searching$ hdfs dfs -copyFromLocal data.txt /HW1/searching
```

- 打開瀏覽器並前往 localhost:9870 裡查看 searching 資料夾是否被建立

localhost:9870/explorer.html#/HW1

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/HW1 Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	Oct 23 16:50	0	0 B	searching
drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 20:02	0	0 B	sorting
drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 19:53	0	0 B	tfidf

Showing 1 to 3 of 3 entries Previous 1 Next

- 點選 searching 資料夾查看是否 input 資料已成功放上

/HW1/searching Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	622 B	Oct 20 19:59	1	128 MB	data.txt
drwxr-xr-x	hadoop	supergroup	0 B	Oct 23 16:50	0	0 B	result

Showing 1 to 2 of 2 entries Previous 1 Next

- 撰寫 run.sh

```
1 hadoop jar '/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar' \
2 -mapper 'python mapper.py' \
3 -file /home/hadoop/HW/searching/mapper.py \
4 -reducer 'python reducer.py' \
5 -file /home/hadoop/HW/searching/reducer.py \
6 -input hdfs://HW1/searching/data.txt \
7 -output /HW1/searching/result
```

- 執行 run.sh

```
hadoop@ubuntu22:~/HW/searching$ ./run.sh
```

- 到 localhost:9870 網站上/HW1/searching/result 資料夾下查看結果
 - _SUCCESS 為成功執行訊息
 - part-0000 為程式運行結果

Browse Directory

/HW1/searching/result Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	Oct 23 16:50	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	273 B	Oct 23 16:50	1	128 MB	part-0000

Showing 1 to 2 of 2 entries Previous 1 Next

- Output data：秀出查詢字在哪幾句句子裡

```

2
3 This article has your searching word "she":
4 [1] With determination, she started her journey, facing challenges head-on, proving to herself that she could conquer her fears.
5 [2] As she reached for the stars, she realized that the stars within her were even brighter.

```

3. TF-IDF computation algorithm

- Input data：將十段文章前分別標入 Document i 以表示為第 i 個文章，並寫入 data.txt 裡

```

data.txt
~/HW/tfidf/Job1
1 Document1: The quick brown fox jumps over the lazy dog in a leap of sheer determination and agility. The sun shines brightly in the sky. It's a beautiful day.
2 Document2: A brown fox is a quick fox. A quick fox is sly. The lazy dog sleeps in the sun, enjoying the warmth.
3 Document3: The sun is shining brightly, making the day beautiful. A sunny day lifts everyone's spirits. The dog rests under the tree, looking content.
4 Document4: The cat chases the squirrel up a tall oak tree. It's a sight to behold. The oak tree stands tall, providing shelter for various animals.
5 Document5: Birds chirp in the morning, creating a soothing melody. The river flows calmly, reflecting the blue sky. Nature is a wonder to behold.
6 Document6: A quiet library is a peaceful place to read and study. Books offer knowledge and adventures. The library's silence is its charm.
7 Document7: The scientist conducts experiments in the laboratory, seeking answers to complex questions. Science is a pursuit of understanding the world.
8 Document8: Music fills the concert hall with a symphony of sounds. The orchestra plays harmoniously, captivating the audience.
9 Document9: The chef prepares a gourmet meal, using fresh ingredients. Culinary art combines flavors, creating a masterpiece.
10 Document10: Artists express themselves through paintings, sculptures, and other mediums. Art is a reflection of the soul's creativity and emotions.
11
12

```

- 在 HW1 目錄上建立一個新目錄 tfidf

```

hadoop@ubuntu22: ~/HW/tfidf
hadoop@ubuntu22:~/HW/tfidf$ hdfs dfs -mkdir /HW1/tfidf

```

- 將本地上的 data.txt 複製到 hdfs 裡的/HW1/tfidf 上

```

hadoop@ubuntu22: ~/HW/tfidf/Job1
hadoop@ubuntu22:~/HW/tfidf/Job1$ hdfs dfs -copyFromLocal data.txt /HW1/tfidf

```

- 打開瀏覽器並前往 localhost:9870 裡查看 tfidf 資料夾是否被建立





localhost:9870/explorer.html#/HW1

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/HW1

Go!















Show

25

 entries

Search:

<input type="checkbox"/>	 Permission	 Owner	 Group	 Size	 Last Modified	 Replication	 Block Size	 Name	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Oct 23 16:50	0	0 B	searching	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 20:02	0	0 B	sorting	
<input checked="" type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 19:53	0	0 B	tfidf	

Showing 1 to 3 of 3 entries

Previous

1

Next

- 點選 tfidf 資料夾查看是否 input 資料已成功放上

Browse Directory

/HW1/tfidf

Go!

Show

25

entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	1.38 KB	Oct 20 19:21	1	128 MB	data.txt	<input type="checkbox"/>
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 19:22	0	0 B	result1	<input type="checkbox"/>
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 19:22	0	0 B	result2	<input type="checkbox"/>
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 19:32	0	0 B	result3	<input type="checkbox"/>
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Oct 20 19:53	0	0 B	result4	<input type="checkbox"/>

Showing 1 to 5 of 5 entries

Previous

1

Next

- 撰寫 run1.sh、run2.sh、run3.sh、run4.sh(因為 tfidf 拆成 4 個 job 做)
- run1.sh 的 input 為 data.txt，run2.sh、run3.sh、run4.sh 的 input 則為前一個 job 的 output

```

Open  [icon] run1.sh
~/HW/tfidf/Job1
1 hadoop jar '/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar' \
2 -mapper 'python mapper1.py' \
3 -file /home/hadoop/HW/tfidf/Job1/mapper1.py \
4 -reducer 'python reducer1.py' \
5 -file /home/hadoop/HW/tfidf/Job1/reducer1.py \
6 -input hdfs:/HW1/tfidf/data.txt \
7 -output /HW1/tfidf/result1

Open  [icon] *run2.sh
~/HW/tfidf/Job2
1 hadoop jar '/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar' \
2 -mapper 'python mapper2.py' \
3 -file /home/hadoop/HW/tfidf/Job2/mapper2.py \
4 -reducer 'python reducer2.py' \
5 -file /home/hadoop/HW/tfidf/Job2/reducer2.py \
6 -input hdfs:/HW1/tfidf/result1/part-00000 \
7 -output /HW1/tfidf/result2

Open  [icon] run3.sh
~/HW/tfidf/Job3
1 hadoop jar '/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar' \
2 -mapper 'python mapper3.py' \
3 -file /home/hadoop/HW/tfidf/Job3/mapper3.py \
4 -reducer 'python reducer3.py' \
5 -file /home/hadoop/HW/tfidf/Job3/reducer3.py \
6 -input hdfs:/HW1/tfidf/result2/part-00000 \
7 -output /HW1/tfidf/result3

Open  [icon] run4.sh
~/HW/tfidf/Job4
1 hadoop jar '/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar' \
2 -mapper 'python mapper4.py' \
3 -file /home/hadoop/HW/tfidf/Job4/mapper4.py \
4 -reducer 'python reducer4.py' \
5 -file /home/hadoop/HW/tfidf/Job4/reducer4.py \
6 -input hdfs:/HW1/tfidf/result3/part-00000 \
7 -output /HW1/tfidf/result4

```

- 執行 run1.sh、run2.sh、run3.sh、run4.sh

```
hadoop@ubuntu22:~/HW/tfidf/Job1$ ./run1.sh
```

```
hadoop@ubuntu22:~/HW/tfidf/Job2$ ./run2.sh
```

```
hadoop@ubuntu22:~/HW/tfidf/Job3$ ./run3.sh
```

```
hadoop@ubuntu22:~/HW/tfidf/Job4$ ./run4.sh
```

- 到 localhost:9870 網站上/HW1/tfidf/result1 (result2/result3/result4)資料夾下查看結果
 - `_SUCCESS` 為成功執行訊息
 - `part-0000` 為程式運行結果

Show entries

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	0 B	Oct 20 19:22	1	128 MB	_SUCCESS	
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	3.37 KB	Oct 20 19:22	1	128 MB	part-00000	

Showing 1 to 2 of 2 entries

- Output1 :
 - ✓ The 1st column : Word
 - ✓ The 2nd column : Document name
 - ✓ The 3rd column : Number of this word in this document

1	A	Document6	1	
2	A	Document2	2	
3	A	Document3	1	
4	Art	Document10	1	
5	Artists	Document10	1	
6	Birds	Document5	1	
7	Books	Document6	1	
8	Culinary	Document9	1	1
9	It's	Document1	1	
10	It's	Document4	1	
11	Music	Document8	1	
12	Nature	Document5	1	
13	Science	Document7	1	
14	The	Document9	1	
15	The	Document6	1	
16	The	Document7	1	
17	The	Document8	1	
18	The	Document4	2	
19	The	Document1	2	
20	The	Document5	1	
21	The	Document2	1	
22	The	Document3	2	
23	a	Document5	2	
24	a	Document4	2	
25	a	Document1	2	
26	a	Document2	1	
27	a	Document9	2	
28	a	Document7	1	
29	a	Document6	1	
30	a	Document8	1	
31	a	Document10	1	
32	adventures	Document6	1	1
33	agility	Document1	1	
34	and	Document1	1	
35	and	Document10	2	
36	and	Document6	2	
37	animals	Document4	1	
38	answers	Document7	1	
39	art	Document9	1	
40	audience	Document8	1	1
41	beautiful	Document3	1	1
42	beautiful	Document1	1	1
43	behold	Document5	1	
44	behold	Document4	1	
45	blue	Document5	1	
46	brightly	Document3	1	1
47	brightly	Document1	1	1
48	brown	Document2	1	
49	brown	Document1	1	

- Output2 :
 - ✓ The 1st column : Word
 - ✓ The 2nd column : Document name
 - ✓ The 3rd column : Number of this word in this document
 - ✓ The 4th column : The total number of this document

1	leap,Document1	1,28	
2	the,Document1	2,28	
3	quick,Document1	1,28	
4	shines,Document1		1,28
5	jumps,Document1	1,28	
6	sheer,Document1	1,28	
7	lazy,Document1	1,28	
8	sun,Document1	1,28	
9	sky,Document1	1,28	
10	over,Document1	1,28	
11	of,Document1	1,28	
12	agility,Document1		1,28
13	in,Document1	2,28	
14	It's,Document1	1,28	
15	fox,Document1	1,28	
16	The,Document1	2,28	
17	dog,Document1	1,28	
18	and,Document1	1,28	
19	determination,Document1	1,28	
20	day,Document1	1,28	
21	a,Document1	2,28	
22	brown,Document1	1,28	
23	brightly,Document1		1,28
24	beautiful,Document1		1,28
25	and,Document10	2,19	
26	express,Document10		1,19
27	a,Document10	1,19	
28	Art,Document10	1,19	
29	creativity,Document10		1,19
30	emotions,Document10		1,19
31	is,Document10	1,19	
32	Artists,Document10		1,19
33	soul's,Document10		1,19
34	other,Document10		1,19
35	reflection,Document10		1,19
36	the,Document10	1,19	
37	paintings,Document10		1,19
38	of,Document10	1,19	
39	mediums,Document10		1,19
40	through,Document10		1,19
41	sculptures,Document10		1,19
42	themselves,Document10		1,19
43	sleeps,Document2		1,22
44	warmth,Document2		1,22
45	lazy,Document2	1,22	
46	the,Document2	2,22	
47	sun,Document2	1,22	
48	quick,Document2	2,22	
49	slv,Document2	1,22	

- Output3 :

- ✓ The 1st column : Word
- ✓ The 2nd column : Document name
- ✓ The 3rd column : Number of this word in this document
- ✓ The 4th column : The total number of this document
- ✓ The 5th column : Number of this word in all documents
- ✓ The 6th column : Total number of documents

1	A,Document6	1,22,3	10	
2	A,Document2	2,22,3	10	
3	A,Document3	1,23,3	10	
4	Art,Document10	1,19,1	10	
5	Artists,Document10	1,19,1	10	
6	Birds,Document5	1,23,1	10	
7	Books,Document6	1,22,1	10	
8	Culinary,Document9	1,16,1	10	
9	It's,Document1	1,28,2	10	
10	It's,Document4	1,25,2	10	
11	Music,Document8	1,17,1	10	
12	Nature,Document5	1,23,1	10	
13	Science,Document7	1,20,1	10	
14	The,Document5	1,23,9	10	
15	The,Document8	1,17,9	10	
16	The,Document9	1,16,9	10	
17	The,Document7	1,20,9	10	
18	The,Document6	1,22,9	10	
19	The,Document1	2,28,9	10	
20	The,Document2	1,22,9	10	
21	The,Document4	2,25,9	10	
22	The,Document3	2,23,9	10	
23	a,Document2	1,22,9	10	
24	a,Document10	1,19,9	10	
25	a,Document1	2,28,9	10	
26	a,Document7	1,20,9	10	
27	a,Document4	2,25,9	10	
28	a,Document5	2,23,9	10	
29	a,Document9	2,16,9	10	
30	a,Document6	1,22,9	10	
31	a,Document8	1,17,9	10	
32	adventures,Document6	1,22,1	10	
33	agility,Document1	1,28,1	10	
34	and,Document1	1,28,3	10	
35	and,Document10	2,19,3	10	
36	and,Document6	2,22,3	10	
37	animals,Document4	1,25,1	10	
38	answers,Document7	1,20,1	10	
39	art,Document9	1,16,1	10	
40	audience,Document8	1,17,1	10	
41	beautiful,Document1	1,28,2	10	
42	beautiful,Document3	1,23,2	10	
43	behold,Document4	1,25,2	10	
44	behold,Document5	1,23,2	10	
45	blue,Document5	1,23,1	10	
46	brightly,Document3	1,23,2	10	
47	brightly,Document1	1,28,2	10	
48	brown,Document1	1,28,2	10	
49	brown,Document2	1,22,2	10	

- Output4 :

✓ The 1st column : Word

✓ The 2nd column : Document name

✓ The 3rd column : TF-IDF

1	A	Document6	0.05472603656026982
2	A	Document2	0.10945207312053964
3	A	Document3	0.05234664366634505
4	Art	Document10	0.12118868910494977
5	Artists	Document10	0.12118868910494977
6	Birds	Document5	0.1001123953475672
7	Books	Document6	0.10466295877245664
8	Culinary	Document9	0.14391156831212787
9	It's	Document1	0.05747992544407501
10	It's	Document4	0.064377516497364
11	Music	Document8	0.13544618194082622
12	Nature	Document5	0.1001123953475672
13	Science	Document7	0.1151292546497023
14	The	Document5	0.0045808919851228844
15	The	Document8	0.006197677391636844
16	The	Document9	0.006585032228614147
17	The	Document7	0.005268025782891318
18	The	Document6	0.004789114348083016
19	The	Document1	0.007525751118416167
20	The	Document2	0.004789114348083016
21	The	Document4	0.008428841252626109
22	The	Document3	0.009161783970245769
23	a	Document10	0.005545290297780334
24	a	Document2	0.004789114348083016
25	a	Document1	0.007525751118416167
26	a	Document7	0.005268025782891318
27	a	Document4	0.008428841252626109
28	a	Document5	0.009161783970245769
29	a	Document9	0.013170064457228294
30	a	Document6	0.004789114348083016
31	a	Document8	0.006197677391636844
32	adventures	Document6	0.10466295877245664
33	agility	Document1	0.08223518189264449
34	and	Document1	0.042999028725926286
35	and	Document10	0.12673397940273012
36	and	Document6	0.10945207312053964
37	animals	Document4	0.09210340371976183
38	answers	Document7	0.1151292546497023
39	art	Document9	0.14391156831212787
40	audience	Document8	0.13544618194082622
41	beautiful	Document1	0.05747992544407501
42	beautiful	Document3	0.06997556141017827
43	behold	Document4	0.064377516497364
44	behold	Document5	0.06997556141017827
45	blue	Document5	0.1001123953475672
46	brightly	Document3	0.06997556141017827
47	brightly	Document1	0.05747992544407501
48	brown	Document1	0.05747992544407501
49	brown	Document2	0.07315626874700457