

巨量資料探勘與應用

HW3

一、組員名單與分工

資工碩一 112598017 楊淨雯

資工碩一 112598028 余珮綺

二、環境設定


Multi-node on spark

master 和 worker 皆為 windows, 一樣都要先照之前的作業把 python、spark、hadoop 安裝好

master: 140.124.181.42

worker 1: 192.168.56.1, 4 cores, 8GB Used

worker 2: 140.124.181.73, 8 cores, 16GB Used

 3.5.0 Spark Master at spark://140.124.181.42:7077

URL: spark://140.124.181.42:7077
Alive Workers: 2
Cores in use: 24 Total, 0 Used
Memory in use: 45.7 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 2 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20231207172704-192.168.56.1-58068	192.168.56.1:58068	ALIVE	16 (0 Used)	14.8 GiB (0.0 B Used)	
worker-20231207172839-140.124.181.73-60930	140.124.181.73:60930	ALIVE	8 (0 Used)	30.9 GiB (0.0 B Used)	

Running Applications (0)

Completed Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20231207174847-0001	hw3	12	8.0 GiB		2023/12/07 17:48:47	ACLab	FINISHED	2 s
app-20231207174412-0000	hw3	12	8.0 GiB		2023/12/07 17:44:12	ACLab	FINISHED	3 s

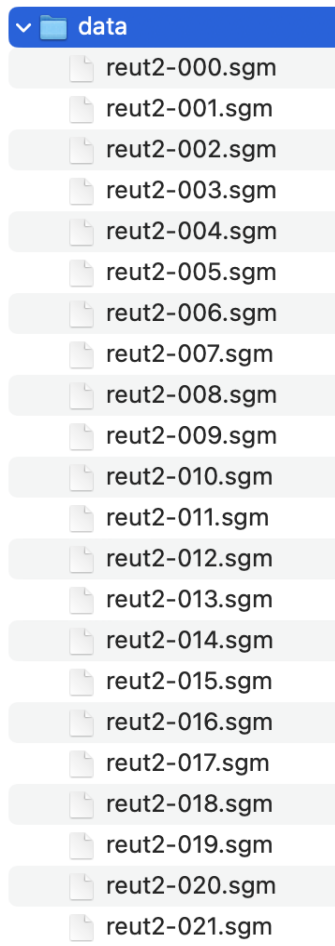
三、實驗執行說明

Step1. 連線

- Master:
以 Admin 開啟 cmd, 到 spark \ bin 資料夾, 輸入指令:
spark-class org.apache.spark.deploy.master.Master
直到出現 Alive 即可
- worker:
以 Admin 開啟 cmd, 到 spark \ bin 資料夾, 輸入指令:
spark-class org.apache.spark.deploy.worker.Worker spark://<master_ip>:7077

Step2. 執行

- 確保指定路徑中的資料夾裡有以下這些檔案



- 打開 cmd，進入 spark \ bin 資料夾
- 輸入 `spark-submit --master spark://http://<master_ip>:7077 --executor-cores 4 --executor-memory 8g <code_file_name>`

附註：

在我們的電腦上

<master_ip> 為 140.124.181.42

<code_file_name> 為 hw3.py

畫底線部分為分配worker資源

要下載作業有用到的額外套件：Pandas

四、實驗結果

1.

成功將 Char_Mat 表格輸出成 CSV 檔案，檔案路徑為：out_csv

[illegible]

2.

id	hashValues
5653	[[1.08574447E7], [5811999.0], [66322.0], [1.10471527E7], [1975240.0], [17465684.0], [37008575.0], [5826871.0], [1.7321749E7], [8246349.0]]
5654	[[31318344E7], [1.021093.0], [1.007995E7], [1.4050304E7], [2.27124269E8], [6.2137093E7], [1.95750418E8], [2.8056988E7], [7.4856499E7], [1.4067882E7]]
5663	[[47000147.0], [4.184361E7], [1821893.0], [2.211099E7], [217149.0], [6.2137093E7], [5.5343504E7], [2.8098554E7], [1.4097967E7], [7614353.0]]
5613	[[1.1560588E7], [4.184361E7], [1.1890949E7], [1851159.0], [2351749.0], [2.224054E7], [3276900.0], [2.8096698E7], [1.4085557E7], [7614353.0]]
5659	[[1.1560588E7], [2.3218665E7], [9.3832974E7], [2.1281858E7], [2.27214269E8], [6.2137093E7], [5.5343504E7], [1.8064438E7], [1.4097967E7], [1.2448607E7]]
5625	[[1.337706E7], [3.2218665E7], [1821093.0], [5.9207815E7], [2351749.0], [2.5901728E7], [2.9939877E7], [1.06825335E8], [1937525.0], [7614353.0]]
5654	[[6.5616666E7], [3.338171E7], [1821093.0], [4.8050304E7], [6.4911025E7], [6.2137093E7], [2.9939877E7], [2.1171599E7], [5.0600338E7], [3.8602292E7]]
5591	[[8706948.0], [4.8233827E7], [2.01874046E8], [2.221099E7], [4.5312963E7], [6.2137093E7], [1.11820583E8], [2.00028322E8], [7.4856499E7], [3.8602292E7]]
5652	[[47000147.0], [6492643.0], [1821093.0], [2.211099E7], [1.2560691E7], [2133347.0], [1.962809.0], [9611057.0], [4.2105483E7], [1.3728436E7]]
5650	[[47000147.0], [1.1560588E7], [4.547375E7], [1.8070250E7], [1.312009E7], [1.740310E7], [1.917000E7], [6.53327E7], [1.917000E7], [3.3728753E7]]
5650	[[47000147.0], [1.1561999.0], [5561642.0], [2087194.0], [1.6125585E7], [831001.0], [1.0749727E7], [1966810.0], [2617160.0], [1.9897272E7]]
5769	[[6033349.0], [1.7642949E7], [8.9724253E7], [4.1532126E7], [1.4512848E7], [6.195327E7], [5.057066E7], [4.2037797E7], [7.4856499E7], [7614353.0]]
5709	[[1.4655511E8], [9.5311173E7], [6.8658334E7], [2.8166688E8], [4.30677149E8], [2.8997244E8], [1.99750418E8], [2.70839285E8], [7.4856499E7], [165688.0]]
5734	[[6093567.0], [3.2218665E7], [1314157.0], [1.7053105E7], [2.2511749E8], [6.2137093E7], [5.534351E7], [1.68400238E8], [3779421.0], [8.8303108E7]]
5683	[[1.337706E7], [9.6560133E7], [1821093.0], [2.14712362E8], [8.0534143E7], [1.36067811E8], [1.02812973E8], [3.60101957E8], [7.4856499E7], [7.021437E8]]
8315	[[1.1560588E7], [1.0553342E7], [2883131.0], [1851159.0], [3.419379E7], [2133347.0], [7391115.0], [7493395.0], [4478681.0], [5087799.0]]
5704	[[1.1560588E7], [2.5630285E7], [1821093.0], [1.0022702E7], [2351749.0], [2.2242045E7], [5.5343504E7], [947744.0], [1.4097967E7], [1.4067882E7]]
5704	[[1.1560588E7], [1.342142E7], [4.8050304E7], [2.27214269E8], [6.2137093.0], [1.975503.0], [2.9939877E7], [1.926871E7], [7.4856499E7], [4.9550506E7]]
5747	[[8.2284915E7], [5.5899417E7], [1821093.0], [1.011802E7], [2.211099E7], [2.2242045E7], [3.7928084E7], [1.02016588E8], [2.7270179E7], [8.8303108E7]]
5738	[[47000147.0], [1695040.0], [1.0182055E7], [2.211099E7], [6407660.0], [1.4922679E7], [1.3788098E7], [1.11859987E8], [3.6199727E7], [2140246.0]]

only showing top 20 rows

3.

+-----+-----+	
	hash count
+-----+-----+	
[3.2867133E7]	2
[1.5545498E7]	2
[7731514.0]	2
[7.4627246E7]	3
[1.6571645E7]	4
[1.3728635E7]	2
[6212580.0]	5
[1.9857206E7]	12
[1.714317E7]	17
[7.1483169E7]	8
[3.7458825E7]	5
[6404769.0]	10
[2661413.0]	4
[9077028.0]	2
[9135768.0]	2
[1.2684948E7]	3
[5.9734509E7]	2
[1689089.0]	2
[1.0006493E8]	22
[3.504417E7]	7
+-----+-----+	
only showing top 20 rows	