



# 結合Spark與Hadoop 於雲端平台進行網路 異常流量偵測

桃園區網會議  
104/11/5

國立中央大學 電算中心

許時準  
周小慧

[center20@cc.ncu.edu.tw](mailto:center20@cc.ncu.edu.tw)

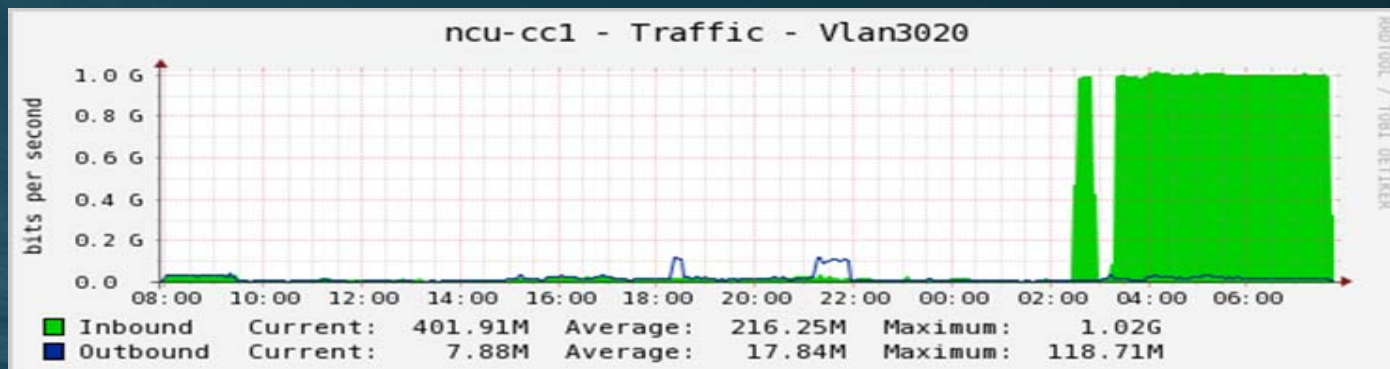
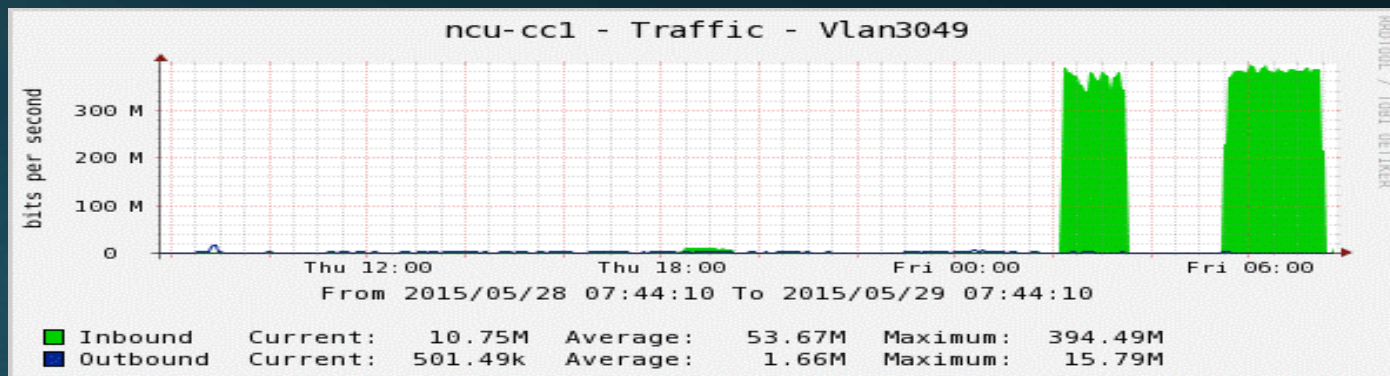


# 報告大綱

- 網路異常流量分析
- Apache Spark
- 桃園區網偵測系統架構
- Spark 模組
- Hadoop 模組
- 測試結果
- 結論



# 網路異常流量分析





## 網路異常流量分析

1. 利用Server、網路設備弱點入侵的駭客攻擊。
2. 點選惡意連結中毒或被開後門。
3. 下載遭植入木馬及後門的軟體。
4. 執行有問題的郵件附檔而致使自己電腦感染病毒、蠕蟲。

## 網路異常流量分析

- 網管可以利用CISCO 路由器的 Netflow 資料找出蛛絲馬跡

srcIP	dstIP	prot	srcPort	dstPort	octets	packets
220.123.30.252	140.115.222.75	17	24933	8322	321	1
180.153.97.14	140.115.1.31	17	53	59354	203	1
59.115.225.220	140.115.153.235	6	2227	12149	92	2
74.125.203.102	140.115.189.67	6	80	2453	20410	27
168.95.1.1	140.115.126.250	17	53	13080	179	1
222.59.54.193	140.115.41.95	6	3549	59574	314	5
74.125.23.138	140.115.236.202	6	443	2663	4759	11
74.125.203.138	140.115.65.206	6	443	33026	1335	4
123.223.24.109	140.115.56.160	6	27546	8265	52	1



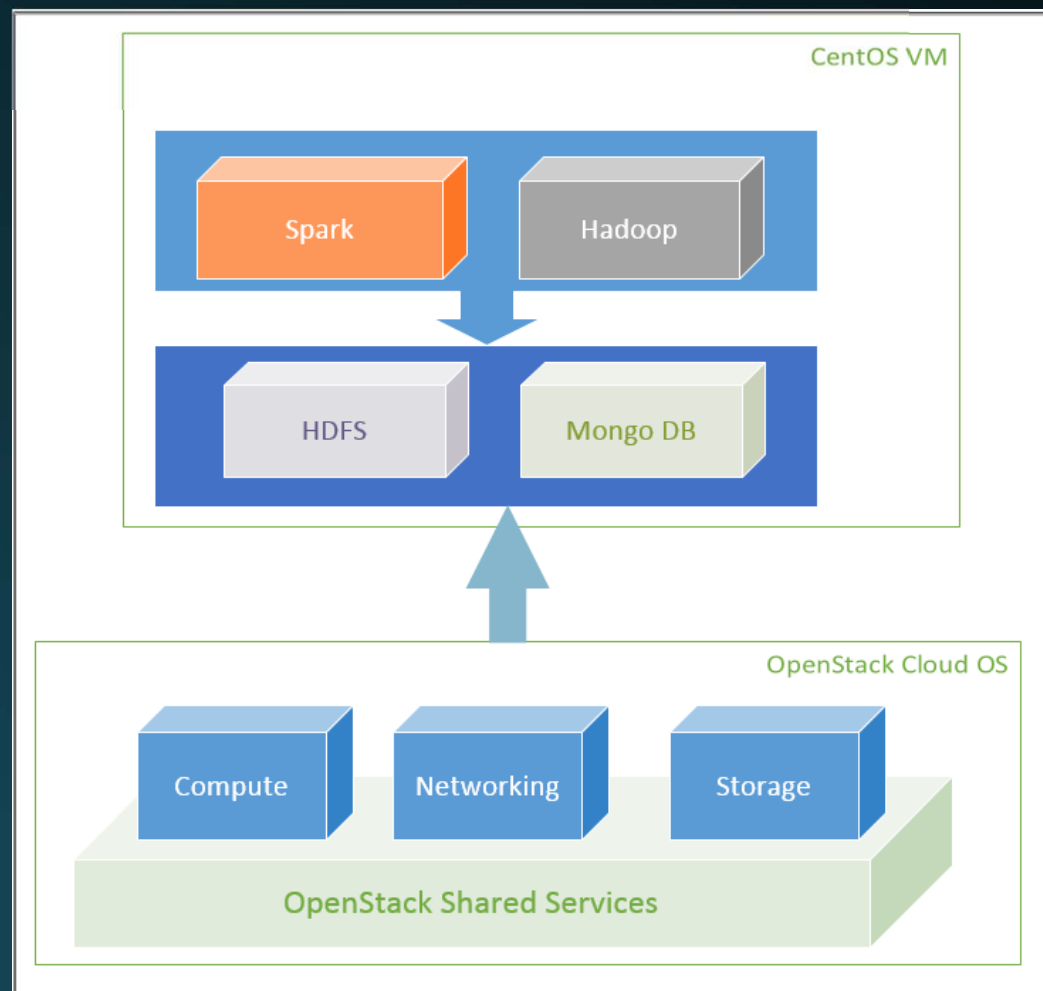


# Apache Spark

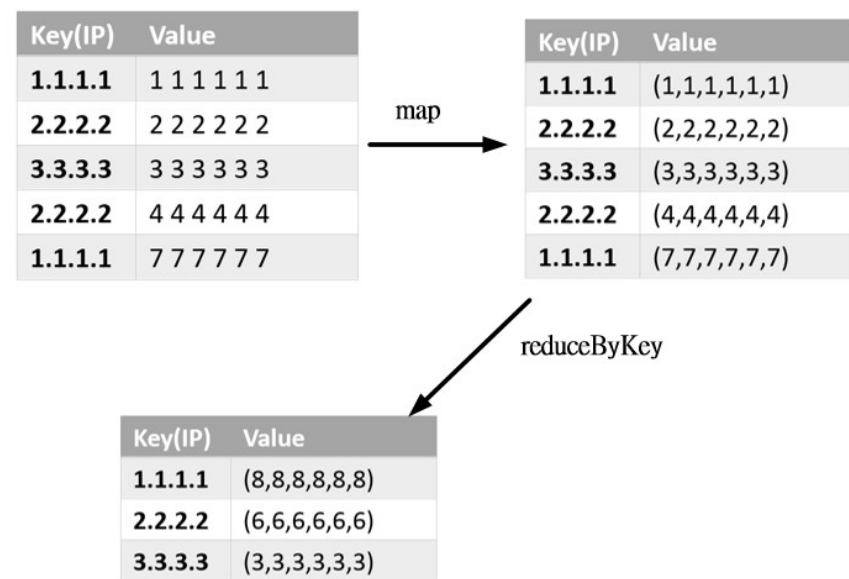
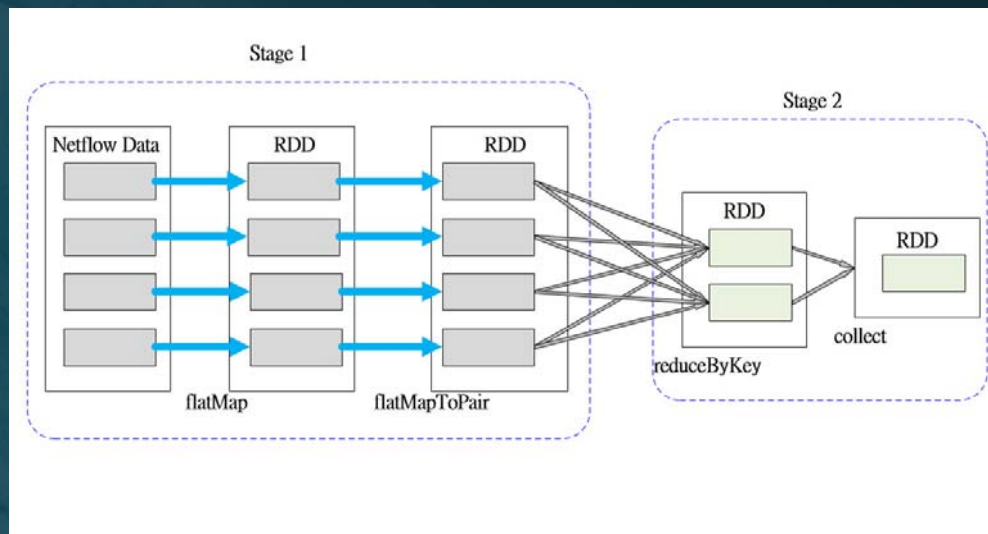
- Apache Spark是一個可快速運算處理巨量資料的運算框架，Spark使用了記憶體內運算技術，能在資料尚未寫入硬碟時在記憶體內分析運算。
- 在 2014 Gray Sort Benchmark 競賽中，Spark 在23分鐘內完成100TB的資料排序，而之前的紀錄是由Hadoop 所創下的72分鐘。Spark 使用的機器只有Hadoop 的1/10，速度卻達到Hadoop 的三倍以上。

# 桃園區網偵測系統架構

- 每10分鐘的即時網路流量偵測處理由Spark 模組進行，下載核心路由器之Netflow 資料，針對異常的攻擊行為特性，篩選出異常的來源主機，並將資料存入 Mongo DB
- Hadoop 模組負責以每小時的頻率累計彙整異常的網路流量，並篩選TopN異常流量主機資料



# Spark 模組





## Spark模組處理後之資料型態

srcIP@prot	sum_in:sum_out:cnt_in:cnt_out:pkt_in:pkt_out
74.125.203.136@17	0:2836185:0:112:0:2958
74.125.203.136@6	0:1867105:0:413:0:3793
74.125.203.138@17	0:14560853:0:630:0:17070
74.125.203.138@6	0:22674792:0:2009:0:31098
74.125.203.139@17	0:17421876:0:625:0:18890
74.125.203.139@6	0:14348897:0:1945:0:24920
74.125.203.141@17	0:15936:0:6:0:21

- srcIP@prot(來源IP及協定)
- sum\_in(輸入位元數), sum\_out(輸出位元數)
- cnt\_in(輸入連接量), cnt\_out(輸出連接量)
- pkt\_in(輸入封包量), pkt\_out(輸出封包量)

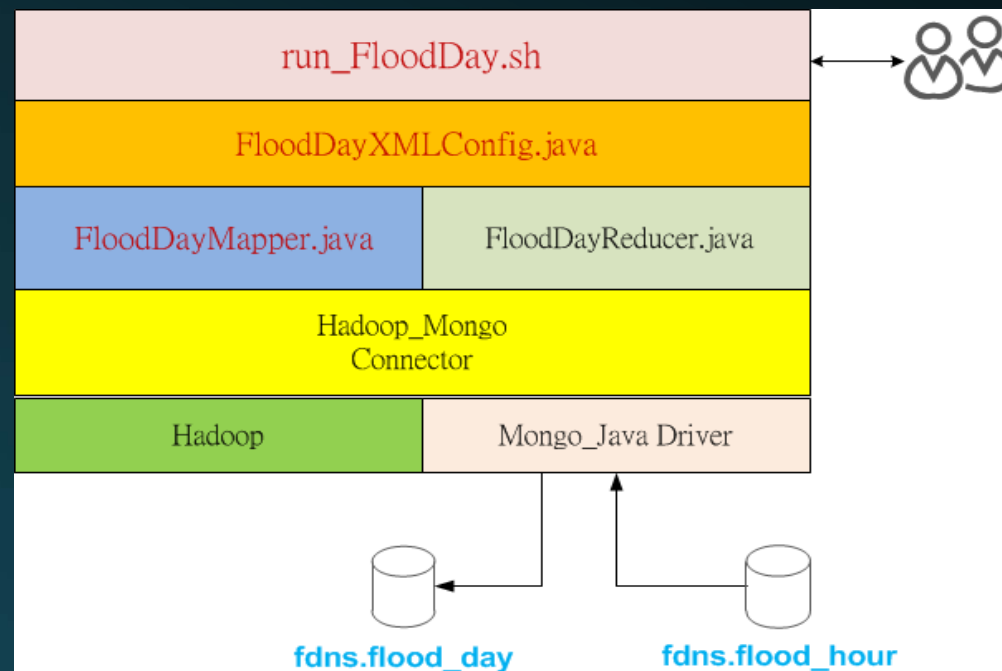


# UDP Flooding 偵測條件

1. appServ flooding (per-10-minute)
  - ✓ Output Packet Size = 1500 B/pkt
  - ✓ Output flow count > 10000
2. Bandwidth Flooding
  - ✓ Output Udp Traffic > 5 GB
3. Resource Flooding
  - ✓ Output Packet Size < 80 B/pkt
  - ✓ Output flow count > 100000
4. Connection Flooding
  - ✓ Output flow count > 100000
5. Packet Flooding
  - ✓ Output packet count > 1000000

# Hadoop 模組

- Apache Hadoop是近年來處理巨量資料最常使用的技術。
- Hadoop 架構包括Hadoop Kernel、MapReduce、Hadoop Distributed File System (HDFS)。






# Hadoop Map Reduce


This is probably the most common error that people make with any sort of data access technology. "Just give me the whole thing, I'll make sense of it on the client side." In the RDBMS world, this would be expressed as "SELECT \* FROM Orders". The problem with such queries is they can potentially...

**Map**



This	1
is	1
probably	1
the	1
most	1
common	1
error	1

**Reduce**



this	2
is	2
probably	1
the	4
most	1
common	1
error	1

# 偵測系統 <http://192.192.227.83/Fdns/>



## 中央大學 結合Spark與Hadoop的網路流量偵測

[\[連線學校 MRTG流量\]](#) [\[IPv6 MRTG流量\]](#) [\[Links 連線狀態偵測\]](#) [\[網管工具箱\]](#)

### TopN 流量

TopN 流量排行

TopN 流量 (小時)

### UDP Flooding 流量監看

UDP Flooding 流量

### UDP 詳細流量

UDP 流量排行

Udp 流量 (小時)

Udp 流量 (10分鐘)

### Pscan 異常

Pscan 異常流量排行

Pscan 異常流量 (小時)

Pscan 異常流量 (10分鐘)

### TopP 封包量

TopP 封包量排行

TopP 封包量 (小時)

TopP 封包量 (10分鐘)

### TopC 連接量

TopC 連接數量排行

TopC 連接量 (小時)

TopC 連接量 (10分鐘)

### TCP 異常流量偵測

密碼猜測 流量

Pandora 流量 (111.111.111.111)

### TopC 連接量排行 (10分鐘)

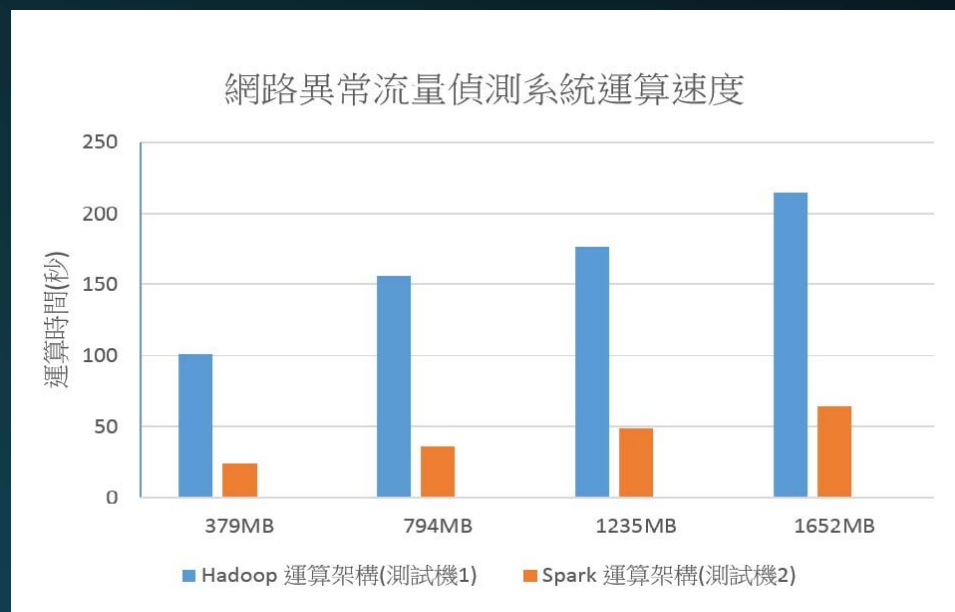
Keyword:

IP 位址	輸入流量(MB)	輸出流量	輸入連接數	輸出連接數	輸入封包長度	輸出封包長度	輸入封包量	輸出封包量	紀錄時間
106.10.150.171@	0	7	0	22501	318	72	0	109075	10-19 09:40
123.240.151.46@	5	5	18878	18378	65	77	77508	74381	10-19 09:40
124.237.78.119@	0	2	16832	47153	50	46	18153	47428	10-19 09:40
140.115.152.64@	14	22	25848	14889	117	394	127226	57405	10-19 09:40
140.115.153.36@	2	3	5096	15706	257	134	9616	23403	10-19 09:40
140.115.155.36@	11	20	15276	9458	59	570	194463	36748	10-19 09:40
140.115.155.54@	6	0	134920	0	46	165	145117	0	10-19 09:40
140.115.171.231@	1	2	4108	15026	272	129	5691	21705	10-19 09:40
140.115.171.232@	2	2	4597	10055	293	140	10132	17916	10-19 09:40
140.115.171.232@	203	133	26430	12933	701	497	290301	268522	10-19 09:40
140.115.171.233@	53	60	27614	10765	376	315	142372	192454	10-19 09:40
140.115.197.163@	1	0	10507	0	53	1454	22145	0	10-19 09:40
140.115.197.179@	1	0	10501	0	53	0	22211	0	10-19 09:40
140.115.197.35@	1	0	10479	0	53	0	21948	0	10-19 09:40
140.115.197.51@	1	0	10145	0	53	0	21308	0	10-19 09:40
140.115.205.70@	2	8	5057	12971	143	291	15764	27815	10-19 09:40
140.115.21.193@	1	3	3703	15091	270	126	5774	23787	10-19 09:40
140.115.21.193@	14	7	15941	18417	237	114	63038	64659	10-19 09:40
140.115.82.114@	7	11	14811	17979	141	305	50177	39252	10-19 09:40


國立中央大學 電算中心

# Hadoop架構及Spark架構運算速度比較

No	Netflow 資料量	Hadoop (測試機1)	Spark (測試機2)
1	379MB	101 sec	24 sec
2	794MB	156 sec	36 sec
3	1235MB	176 sec	49 sec
4	1652MB	215 sec	64 sec







## 結論

- 使用Spark協同Hadoop架構偵測網路異常流量的平台。Spark模組負責即時的運算，可以充分利用 Spark in-memory computing特性，在大量的Netflow Data中快速篩選出異常網路行為的主機。Hadoop模組則處理大量批次作業，以每小時進行大量資料分析作業。
- 實驗結果顯示以Spark 結合 Hadoop架構其處理效能比起原Hadoop架構快了4倍以上，有著相當大的改進。

*The End*

