NARLabs
國家實驗研究院

邁向PetaByte級惡意程式知識庫

國家高速網路與計算中心

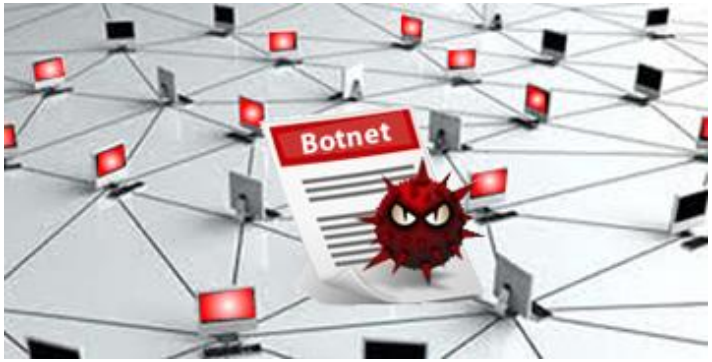
安興彥 副研究員

邁向PetaByte級惡意程式知識庫

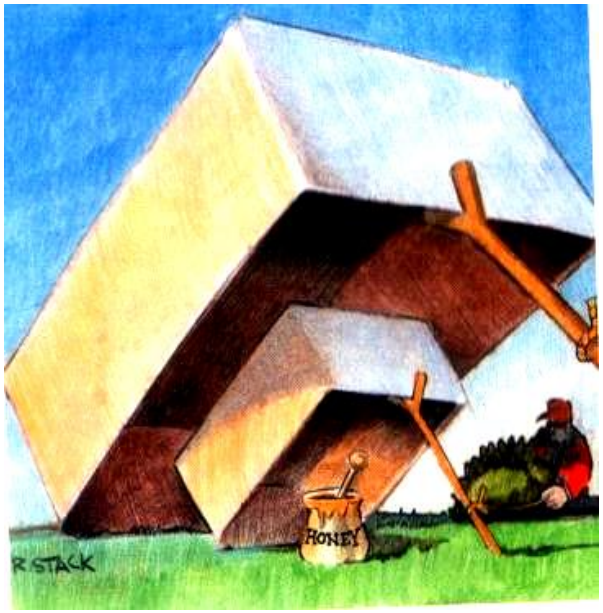
- 資安威脅大資料
  - 誘捕系統日誌
  - 惡意程式樣本
  - 動態分析報告
- Big data 來了怎麼辦？
- 大資料儲存問題
- NoSQL, MongoDB, GridFS
- HDFS？

資安威脅大資料

殭屍網路 vs. 誘捕網路



VS.



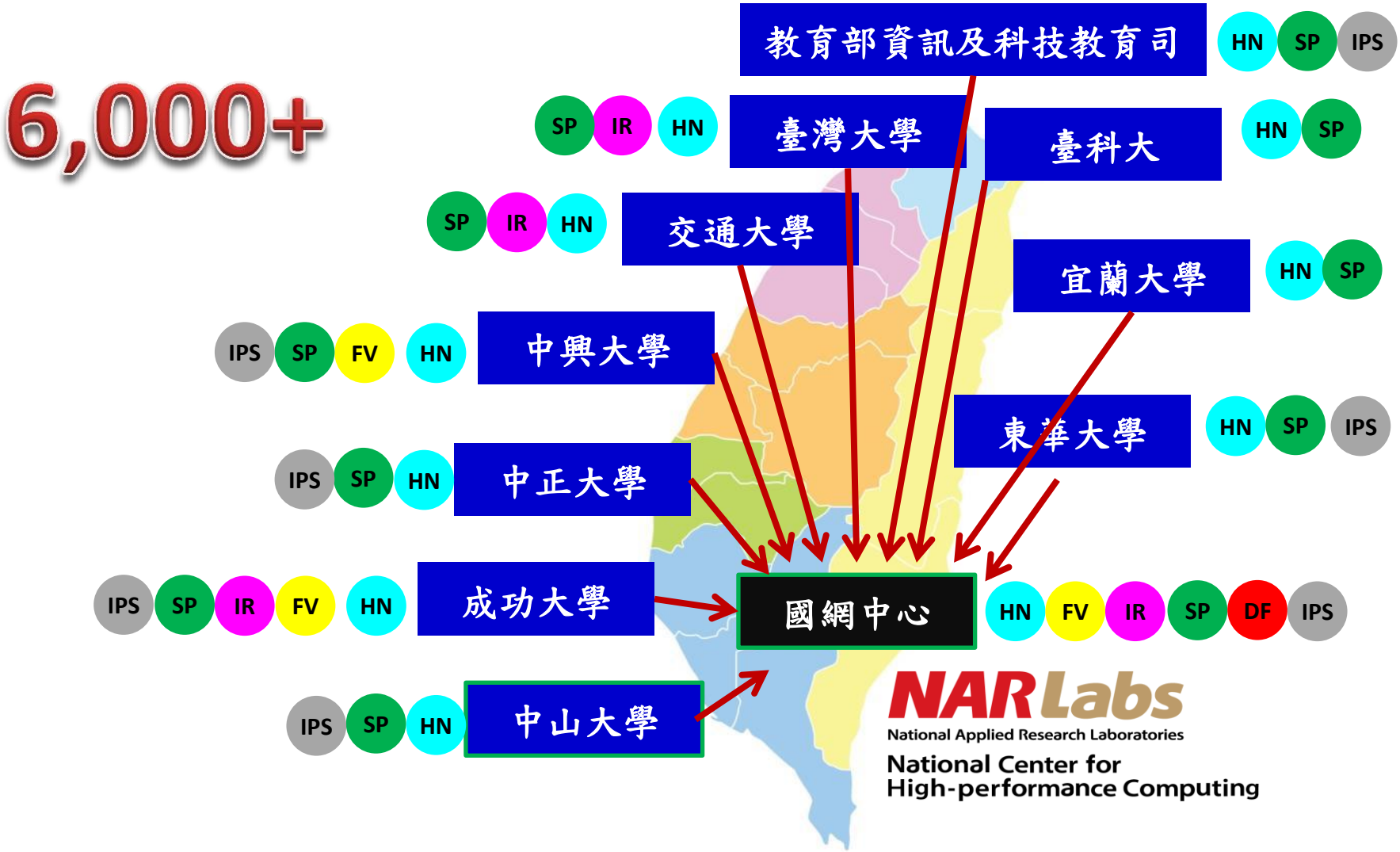
- 無營運價值
- 低度/無安全防護
- 引誘駭客入侵
- 蒐集駭客的活動與行為

圖片來源: http://cybershafarat.com/2015/05/07/education-on-honeypots-sharif-university-courses-on-honeypot-detection/

Honeynet on TANet

NARLabs

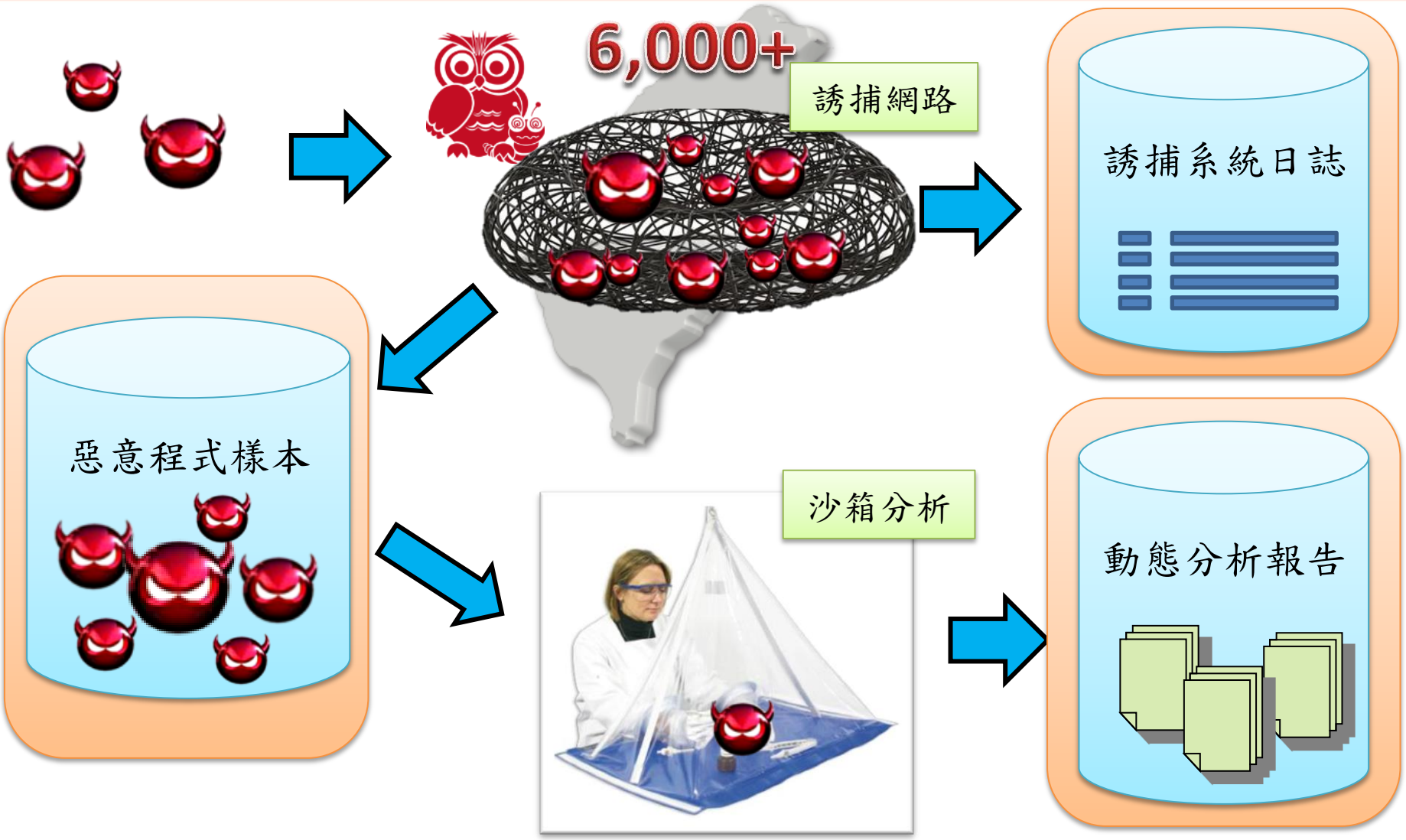
承諾·熱情·創新



資安威脅大資料

NARLabs

承諾·熱情·創新



圖片來源: http://www.erlab.com/fr/20-23-boite-a-gants-pour-l-investigation-biologique-captair-pyramid.html

誘捕系統日誌

NARLabs

承諾·熱情·創新

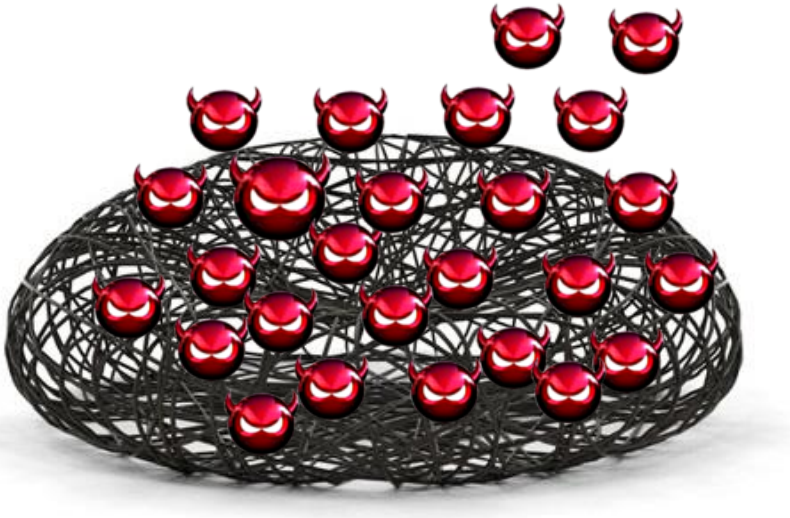
- 總事件量: 33 億筆
- 成長量: 210 萬筆/天
- 磁碟空間: 1.6 TB

通報國內外
即時處理

* 截自 2015/10 止

惡意程式樣本

- 惡意程式: 1250萬隻
- 成長量: 1.2萬隻/天
- 磁碟空間: 6.3 TB



* 截自 2015/10 止

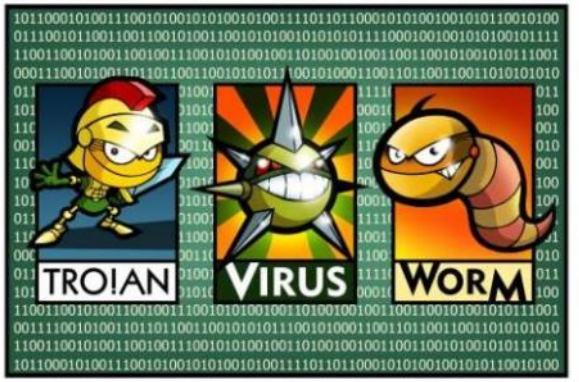
動態分析報告

NARLabs

承諾·熱情·創新

- 沙箱測試分析
  - 隔離的環境
  - 惡意程式分類
  - 網路ip、攻擊行為
  - 更動的檔案

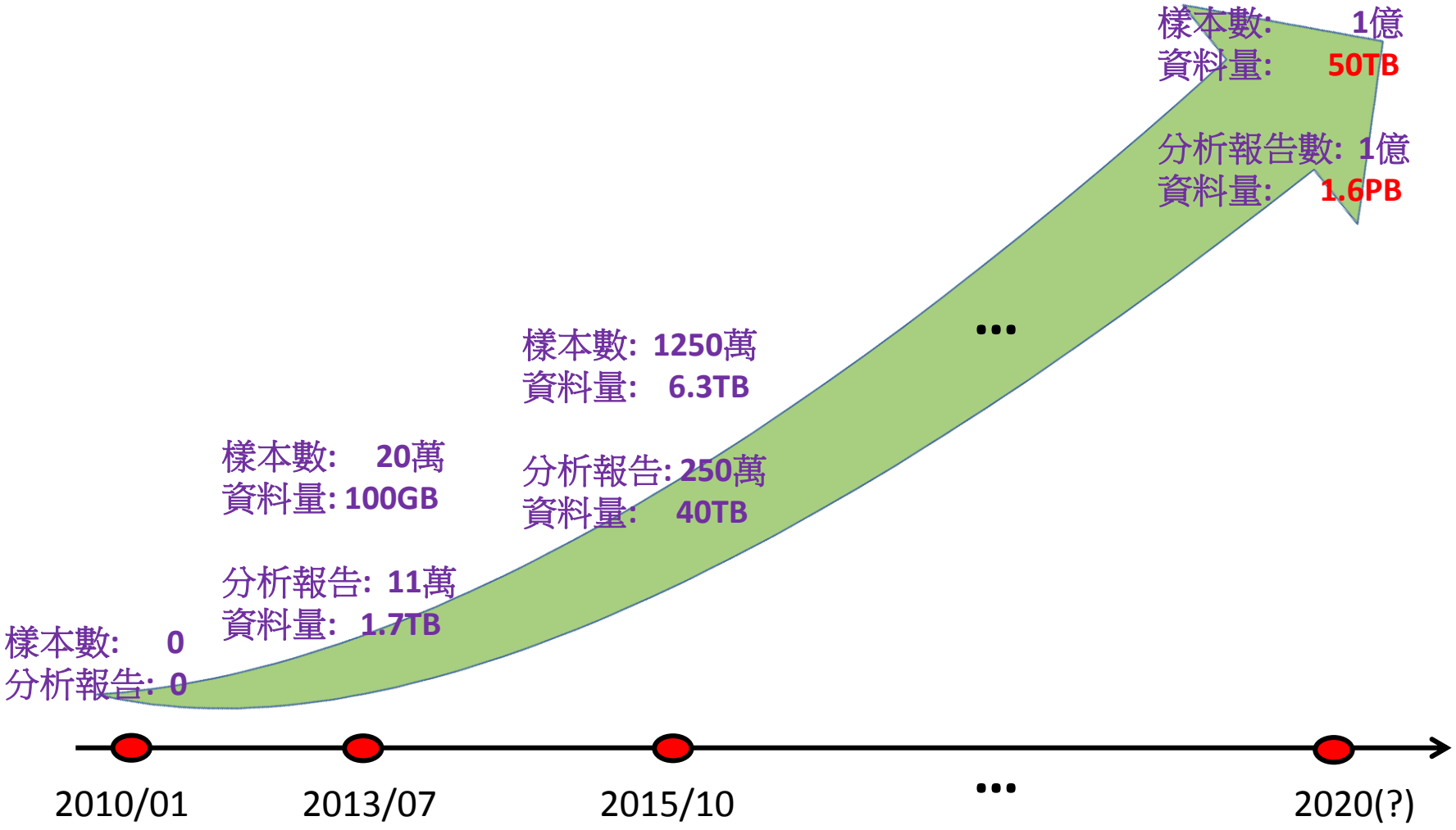
- 分析報告: 250萬份
- 成長量: 1萬份/天
- 磁碟空間: 40 TB



* 截自 2015/10 止

圖片來源: http://www.erlab.com/fr/20-23-boite-a-gants-pour-l-investigation-biologique-captair-pyramid.html

惡意程式樣本快速累積


樣本數: 0
分析報告: 0
樣本數: 20萬
資料量: 100GB
分析報告: 11萬
資料量: 1.7TB
樣本數: 1250萬
資料量: 6.3TB
分析報告: 250萬
資料量: 40TB
...
樣本數: 1億
資料量: 50TB
分析報告數: 1億
資料量: 1.6PB
2010/01
2013/07
2015/10
...
2020(?)

Big Data 來了怎麼辦？

颱風來了怎麼辦?

- 颱風就是要泛舟呀，不然要幹嘛?

圖片來源: http://www.ettoday.net/news/20150709/532400.htm

Big Data來了怎麼辦?

- Big Data 就是要Splunk 呀，不然要幹嘛?
  - 第一家Nasdaq 上市的 big data 公司
  - 可處理PetaByte (PB) 等級資料
  - ~~超好用!! 超好用!! 超好用!!~~ 很貴! 很貴! 很貴!

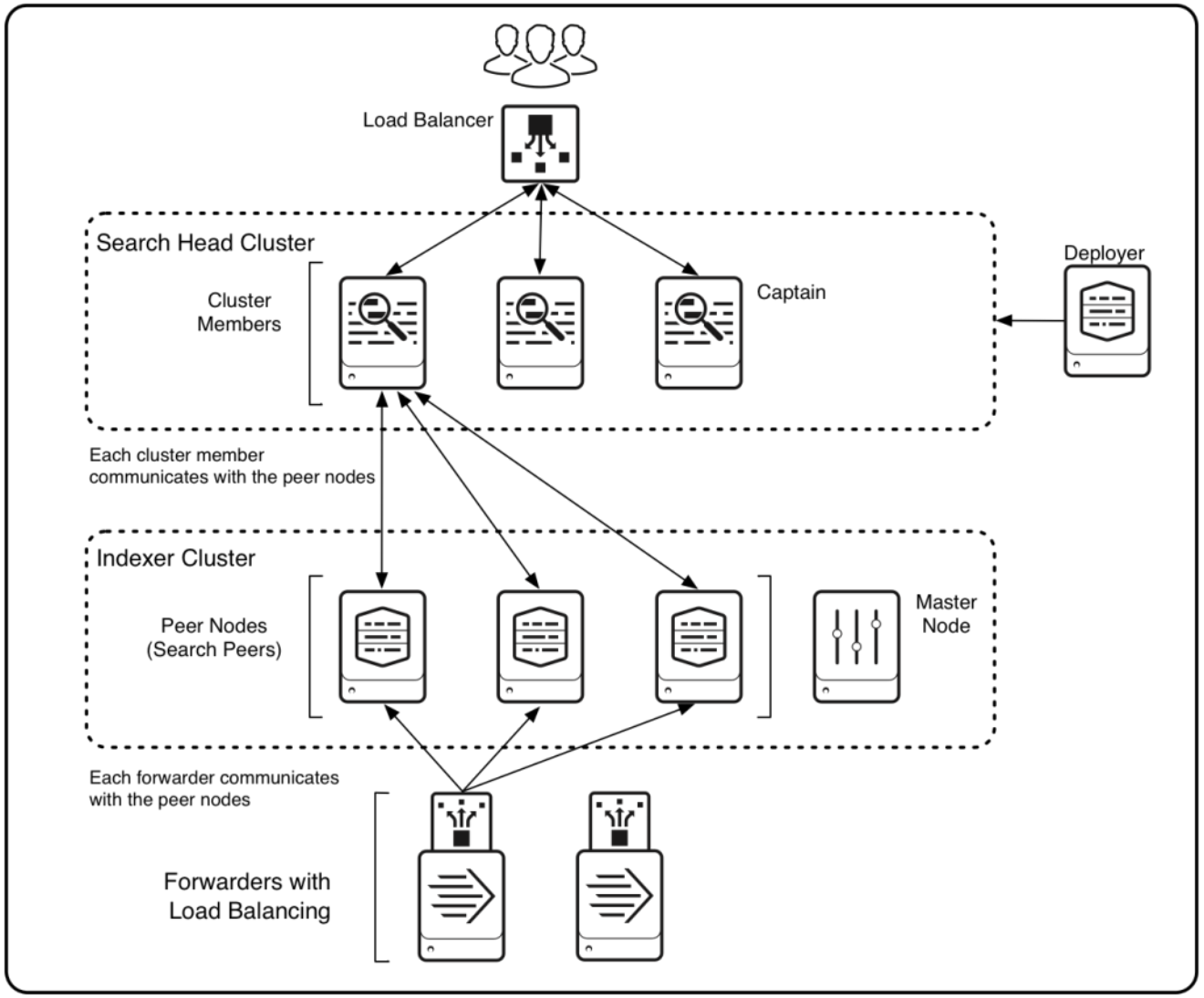
圖片來源: http://www.splunk.com/

Splunk架構

Search Head

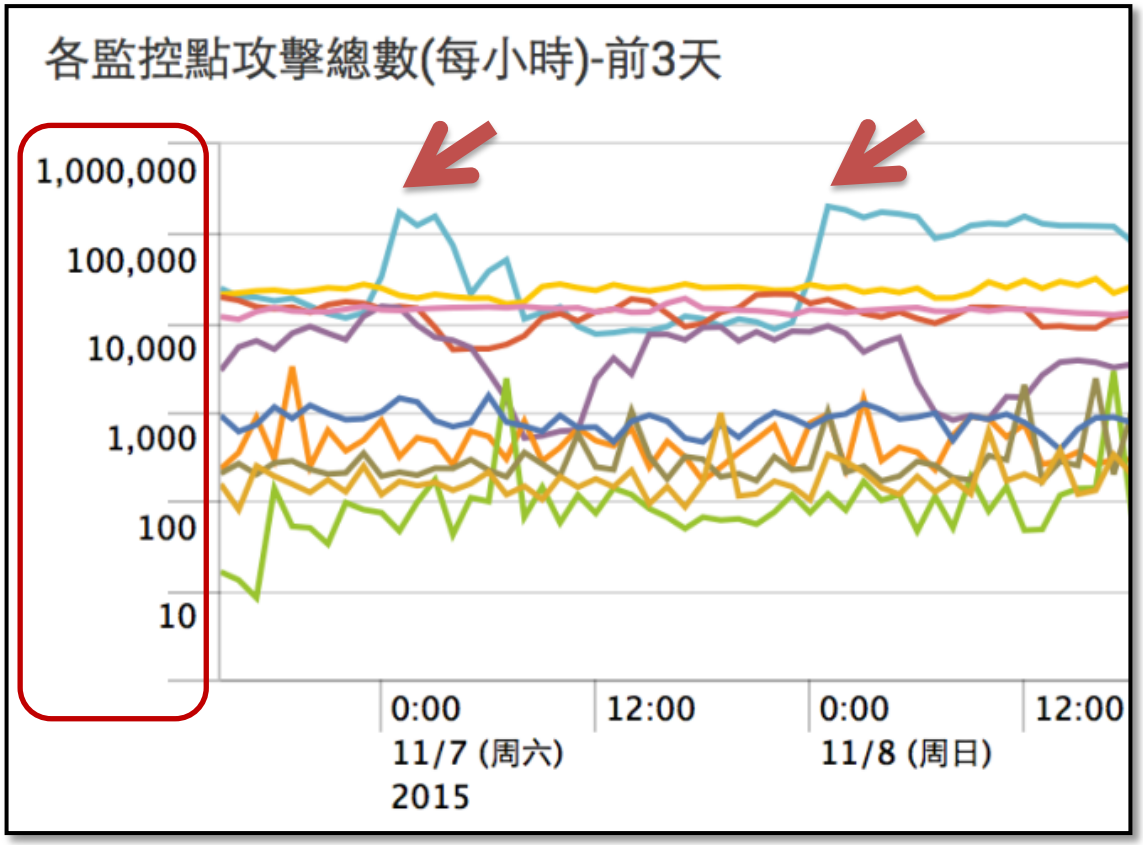
Indexer

Forwarder

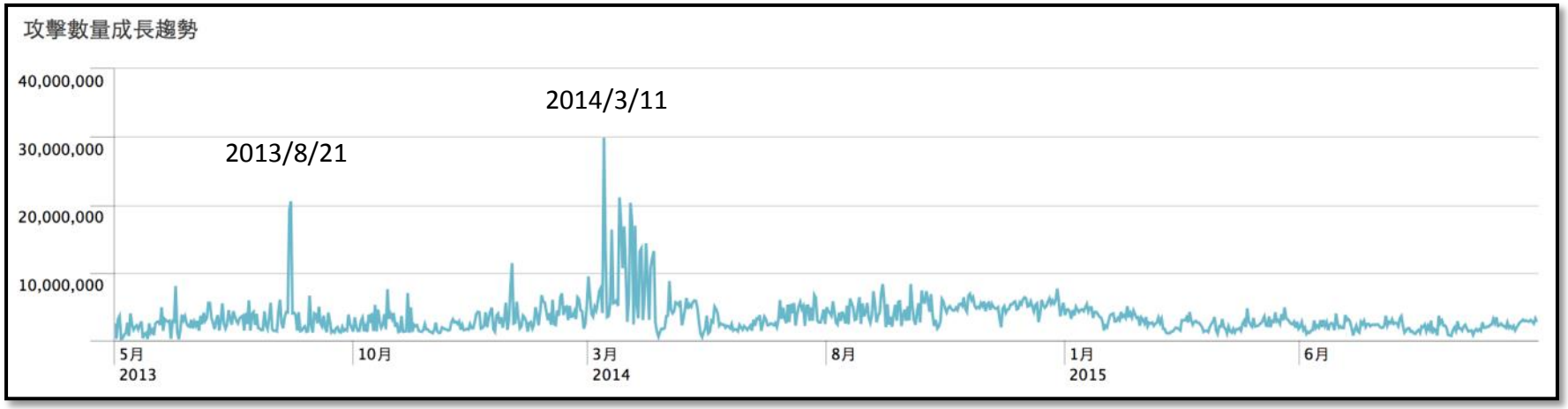


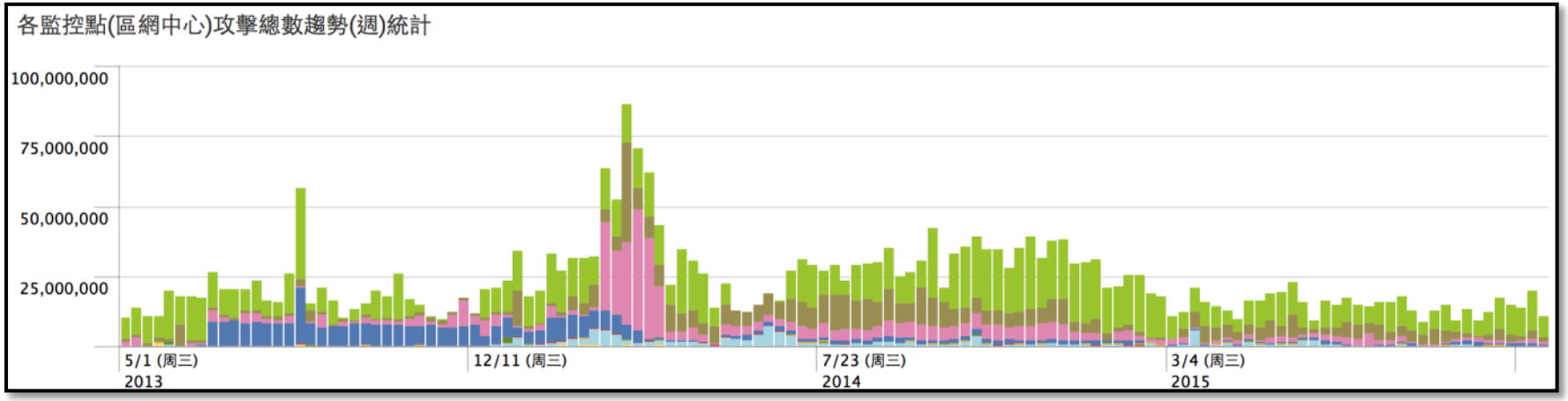
圖片來源: http://docs.splunk.com/Documentation/Splunk/6.2.3/DistSearch/SHCandindexercluster

即時監控

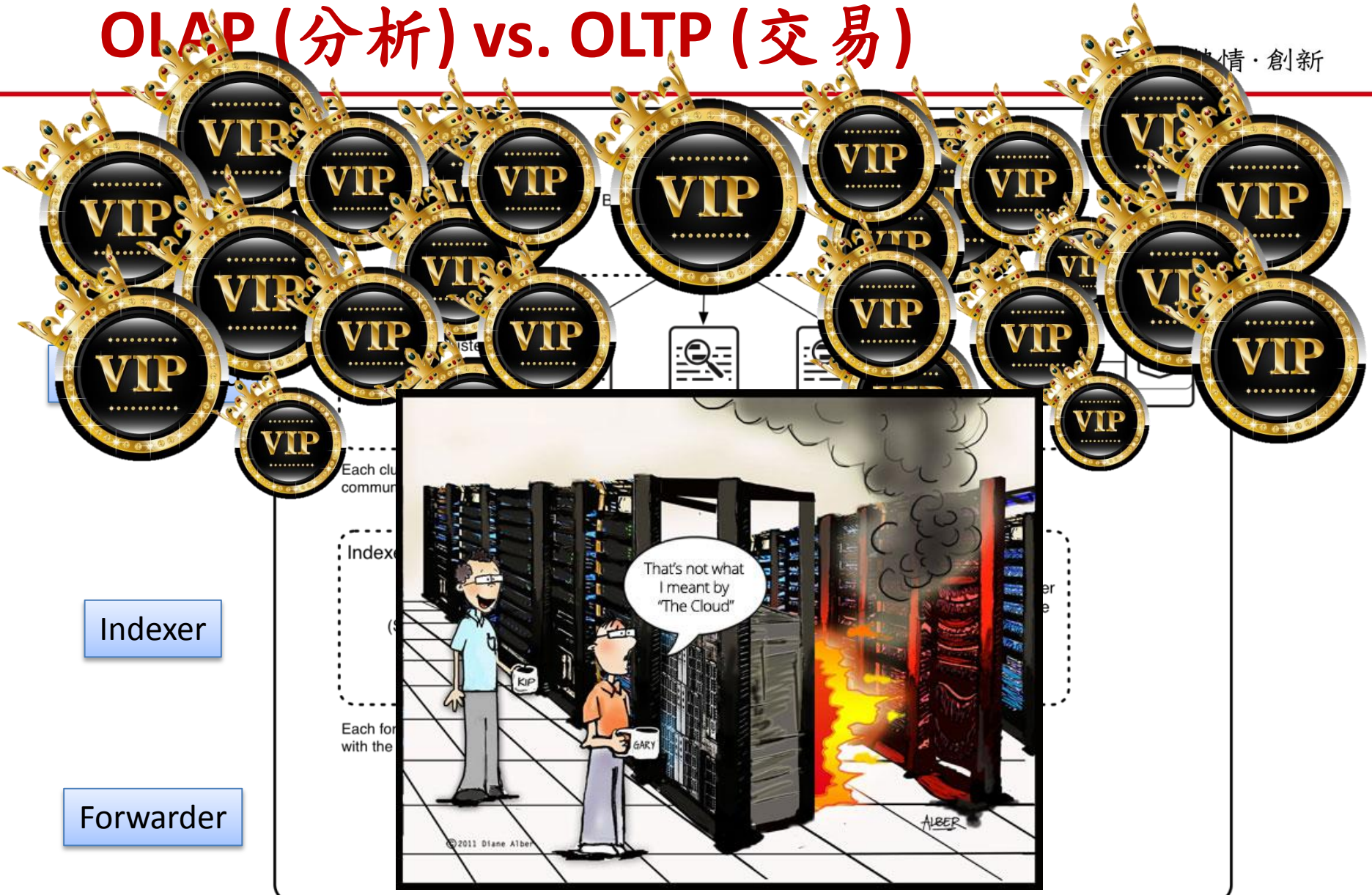

各監控點攻擊總數(每小時)-前3天
1,000,000
100,000
10,000
1,000
100
10
0:00
11/7 (周六)
2015
12:00
0:00
11/8 (周日)
12:00

攻擊成長趨勢分析


攻擊數量成長趨勢
40,000,000
30,000,000
20,000,000
10,000,000
2014/3/11
2013/8/21
5月
2013
10月
3月
2014
8月
1月
2015
6月


各監控點(區網中心)攻擊總數趨勢(週)統計
100,000,000
75,000,000
50,000,000
25,000,000
5/1 (周三)
2013
12/11 (周三)
7/23 (周三)
2014
3/4 (周三)
2015

OLAP (分析) vs. OLTP (交易)



Indexer

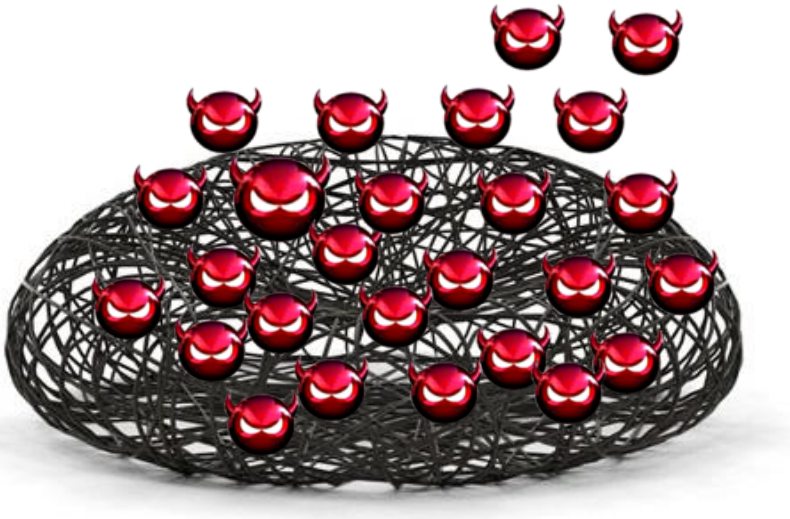
Forwarder

圖片來源: http://docs.splunk.com/Documentation/Splunk/6.2.3/DistSearch/SHCandindexercluster

大資料儲存問題

惡意程式樣本

- 惡意程式: 1250萬隻
- 成長量: 1.2萬隻/天
- 磁碟空間: 6.3 TB



NFS大資料儲存問題

- List all:
  - 11時25分 (檔案數: 700萬)

```
ls /index/md5
```

- Find prefix:
  - 5分2秒 (符合數: 154)

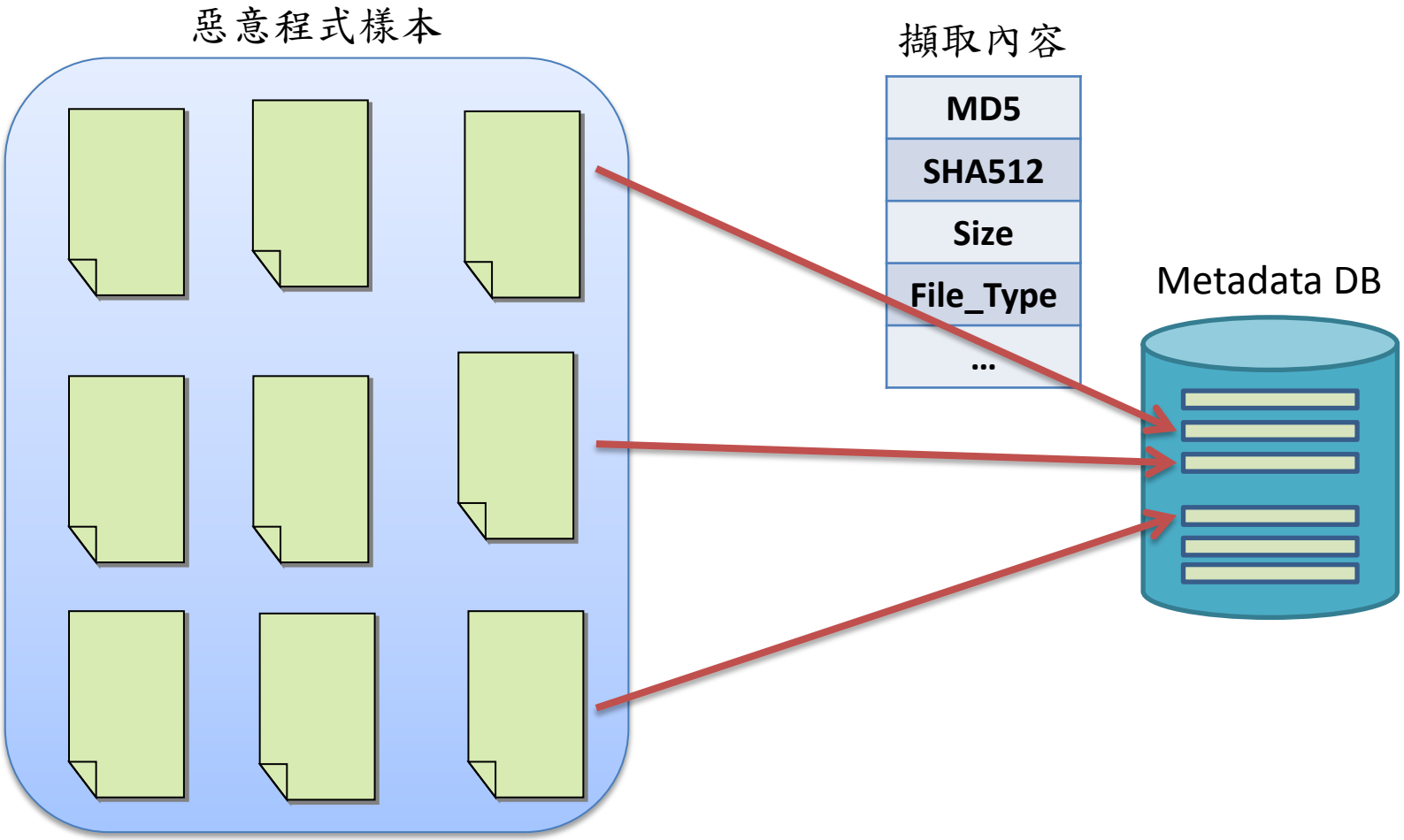
```
ls /index/md5/aaaa*
```

- Find prefix:
  - (符合數: 65萬)
  - bash: /bin/ls: Argument list too long

```
ls /index/md5/a*
```

```
ls /index/md5/a4d33ecc2484d951ee7a0db7996b3cf0
 /index/md5/a4d3401de957a230dd71b552add96e90
 …
 /index/md5/affffac91fd14926193fe06639cf9370
```

建立DB(惡意程式知識庫)



惡意程式知識庫效能 (MySQL)

- Find all:
  - 3分55秒 (檔案數: 700萬)

```
SELECT count(*) FROM `malware`
```

- Find prefix: aaaa*
  - 0.05秒 (符合數: 154)

```
SELECT count(*) FROM `malware`
WHERE md5 LIKE 'aaaa%';
```

- Find prefix: a*
  - 3.9秒 (符合數: 65萬)

```
SELECT count(*) FROM `malware`
WHERE md5 LIKE 'a%';
```

- 已建立 Index

NoSQL, MongoDB, GridFS

MongoDB 資料庫簡介

- NoSQL (Not Only SQL)
  - 不用事先建立 DB schema
- 儲存格式：BSON (Binary JSON)
- 高效能 OLTP (On-Line Transaction Processing)
- 強大的水平擴充能力
  - Sharding (分割/平行運算/加速)
  - Replication (備援)
  - Petabyte (PB)-scale
- 2015資料庫排名: 第4名
- 缺點: 缺乏 SQL join、複雜交易

Rank Nov 2015	Rank Oct 2015	Rank Nov 2014	DBMS
1.	1.	1.	Oracle
2.	2.	2.	MySQL
3.	3.	3.	Microsoft SQL Server
4.	4.	↑ 5.	MongoDB
5.	5.	↓ 4.	PostgreSQL
15.	15.	15.	HBase

圖片來源: http://db-engines.com/en/ranking

Document-based vs. Tables

NARLabs

承諾·熱情·創新

- 正規化 vs. 反正規化

```
{
  "info": {
    "category": "file",
    "package": "",
    "started": "2015-01-14 01:28:45",
    "custom": "",
    "machine": {
      "shutdown_on": "2015-01-14 01:31:16",
      "started_on": "2015-01-14 01:28:45",
      "manager": "VirtualBox",
      "label": "cuckoo2",
      "id": 544,
      "name": "cuckoo2"
    },
    "ended": "2015-01-14 01:31:17",
    "version": "X",
    "duration": 152,
    "id": 543
  },
  "virustotal": {
    "scans": {
      "Bkav": {
        "detected": false,
        "version": "1.3.0.4959",
        "result": null,
        "update": "20140602"
      },
      "MicroWorld-eScan": {
        "detected": false,
        "version": "12.0.250.0",
```

MD5	Host_name	Dst_ip
7208754c4c53cac2b7a308b151e83240	c1.applicationgrabb.com	54.xxx.xxx.99
7208754c4c53cac2b7a308b151e83240	r1.dirgreatbestepicl.info	54.xxx.xxx.99

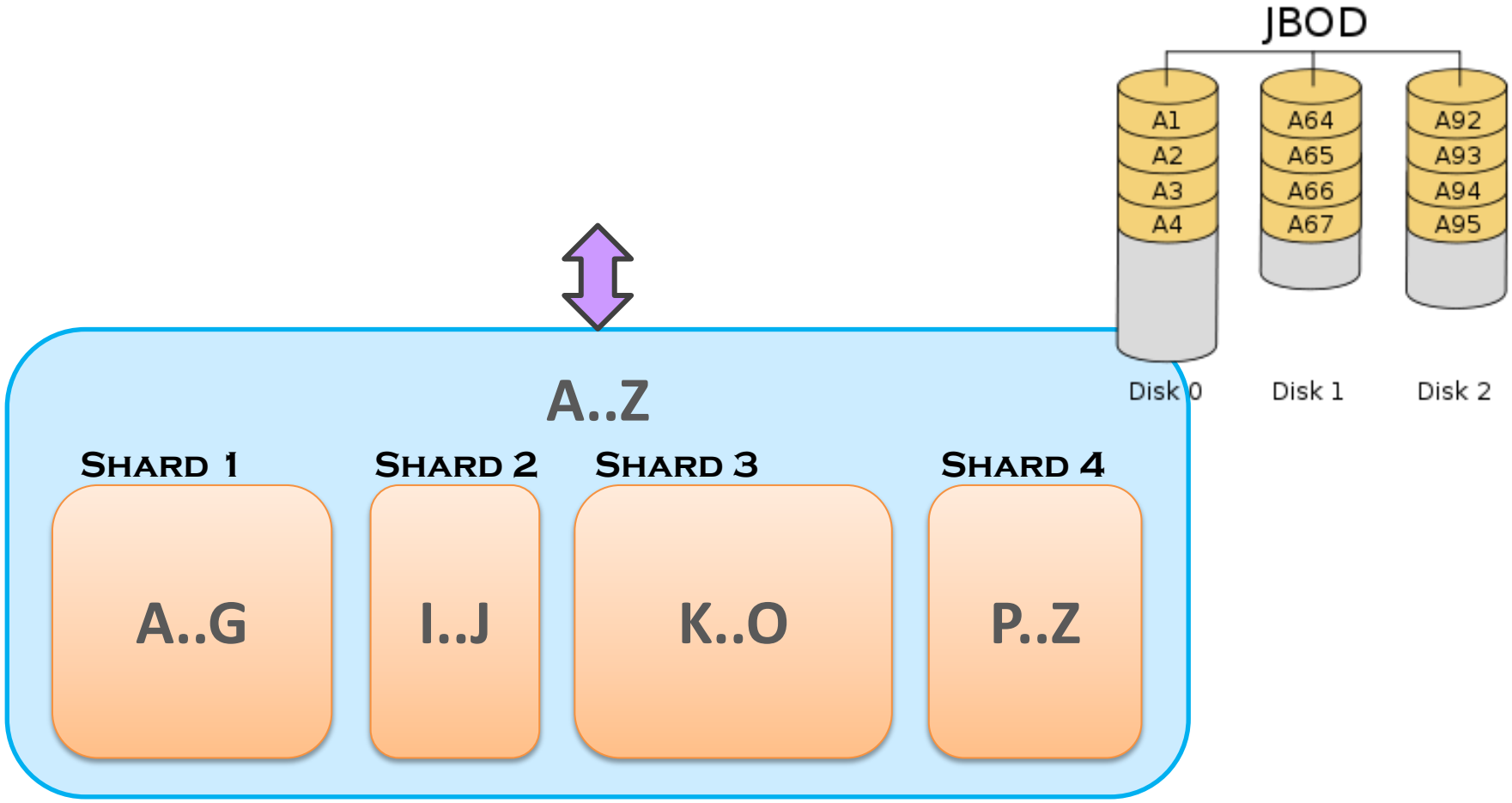
MD5	Scanner	Result
7208754c4c53cac2b7a308b151e83240	nProtect	Trojan.HTML.Iframe.T
7208754c4c53cac2b7a308b151e83240	McAfee	JS/Iframe.gen.af
7208754c4c53cac2b7a308b151e83240	Avast	HTML:Iframe-PE [Trj]

MD5	Src_ip	Src_port	Dst_ip	Dst_port
7208754c4c53cac2b7a308b151e83240	192.168.56.102	54820	70.xxx.xxx.223	14955
7208754c4c53cac2b7a308b151e83240	192.168.56.102	54820	46.xxx.xxx.250	24044
7208754c4c53cac2b7a308b151e83240	192.168.56.102	54820	213.xxx.xxx.10	55209

MongoDB Sharding

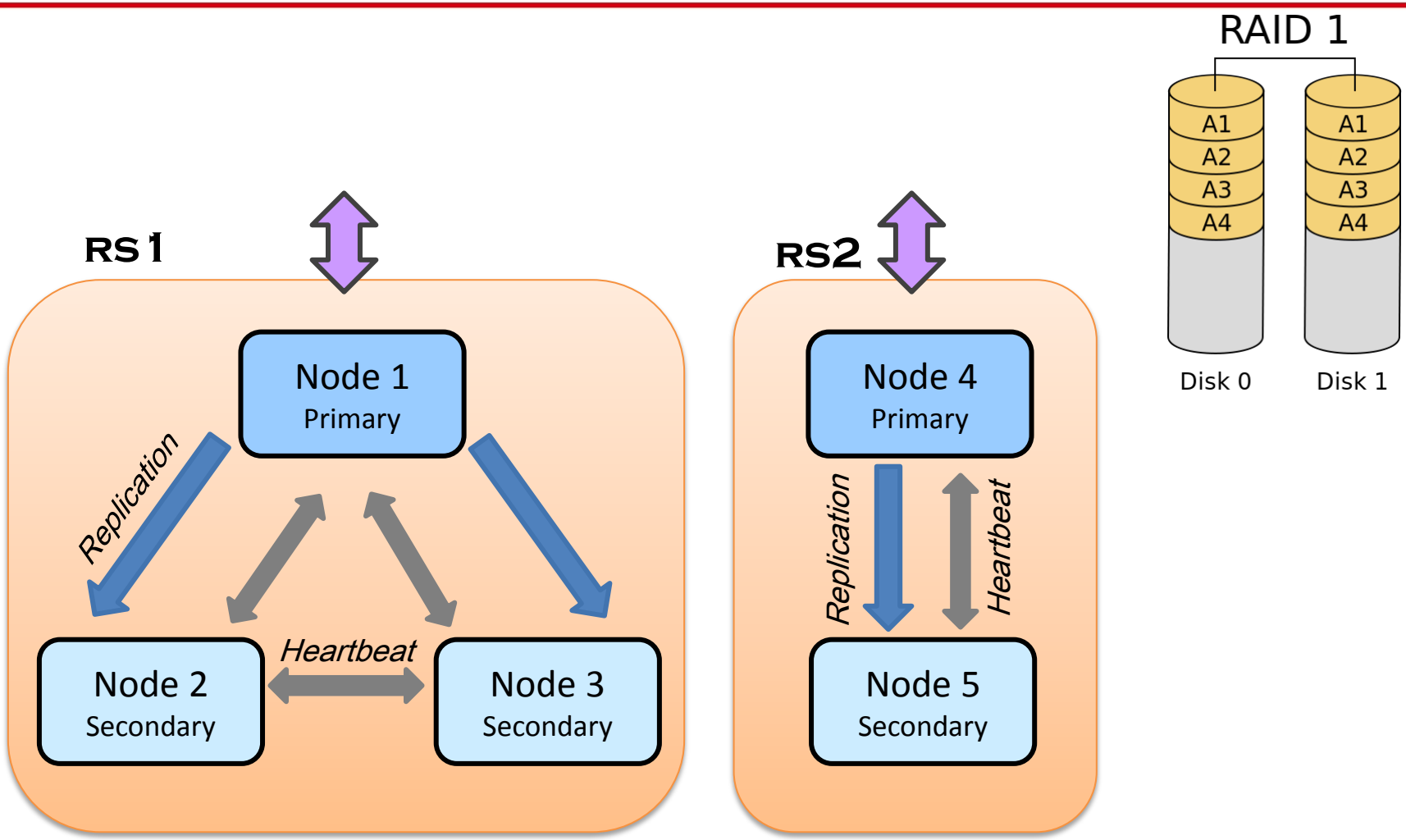
NARLabs

承諾·熱情·創新



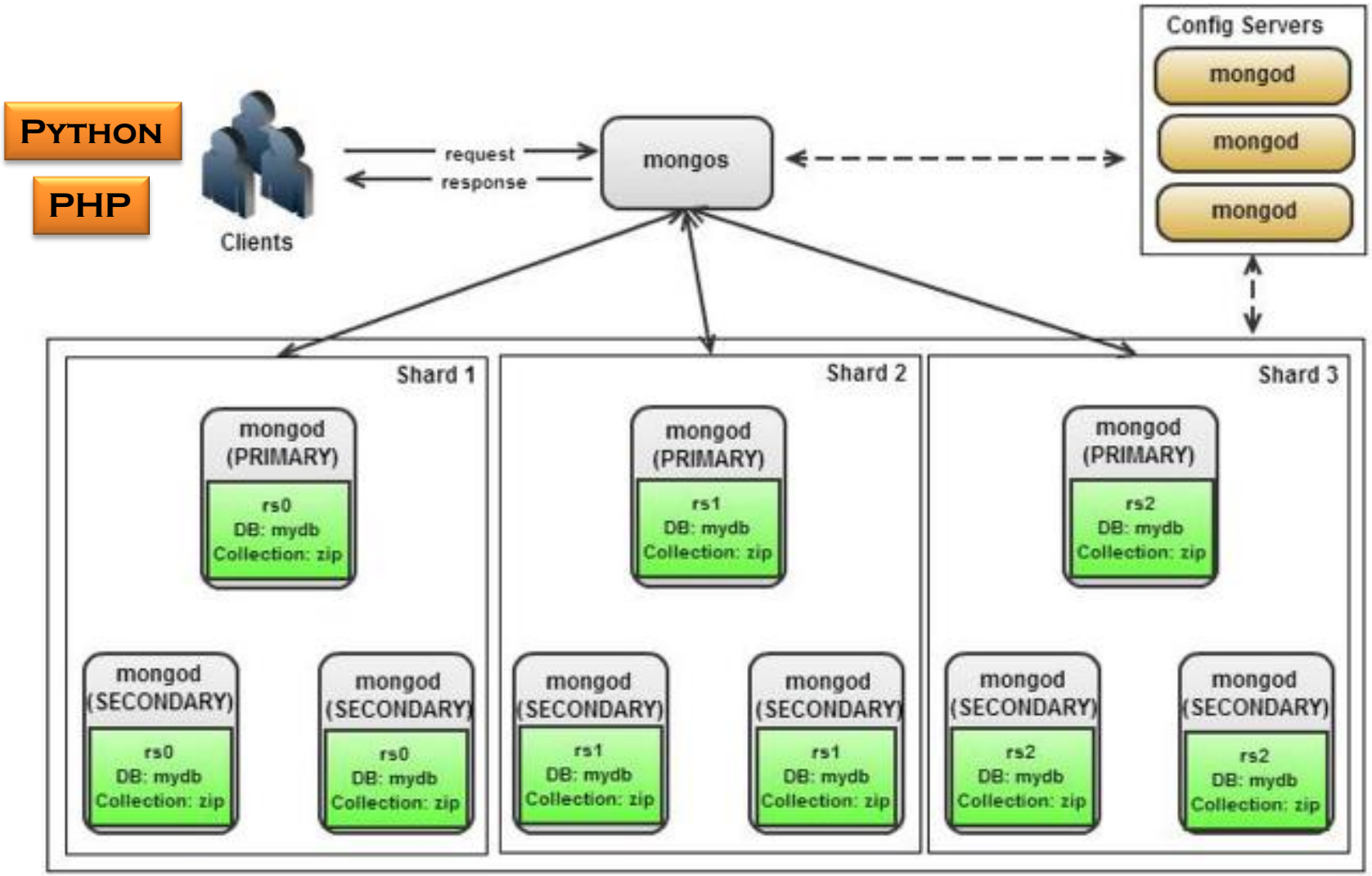
http://uk.wikipedia.org

MongoDB Replica Set



http://uk.wikipedia.org

MongoDB Sharded Cluster



http://severalnines.com/blog/turning-mongodb-replica-set-sharded-cluster

惡意程式知識庫效能 (MongoDB)

NARLabs
承諾 · 熱情 · 創新

- Find all:
  - 0.001秒 (檔案數: 700萬)

```
db.malware.count()
```

- Find prefix: aaaa*
  - 0.05秒 (符合數: 154)

```
db.malware.count(
  {md5 : {$gt: 'aaaa', $lt: 'aaab'} } )
```

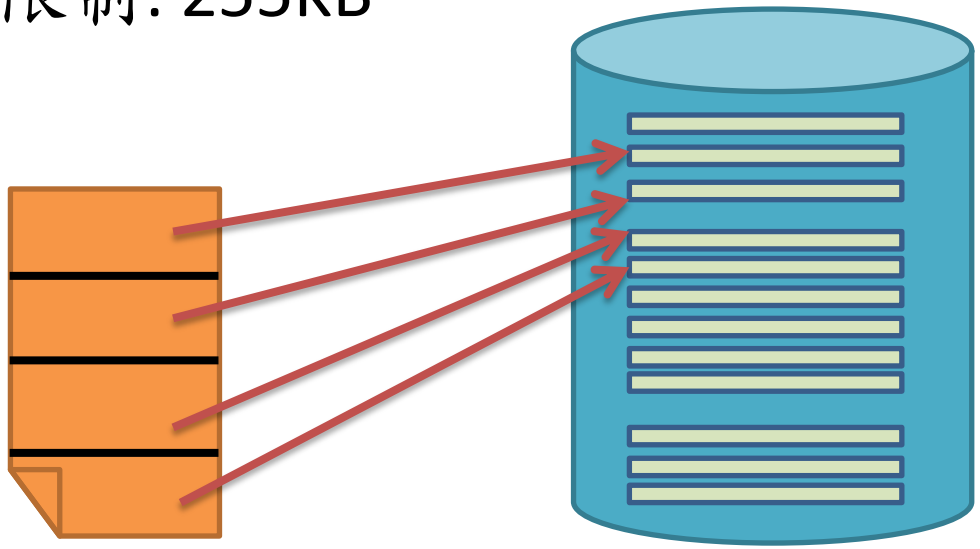
- Find prefix: a*
  - 0.31秒 (符合數: 65萬)

```
db.malware.count(
  {md5 : {$gt: 'a', $lt: 'b'} } )
```

- 已建立 Index

GridFS: MongoDB 附贈功能

- MongoDB 文件大小限制: 16 MB
- GridFS
  - 將檔案切成多個碎片 (Chunk)
  - 每個碎片當作一個文件
  - 預設 Chunk 大小限制: 255KB



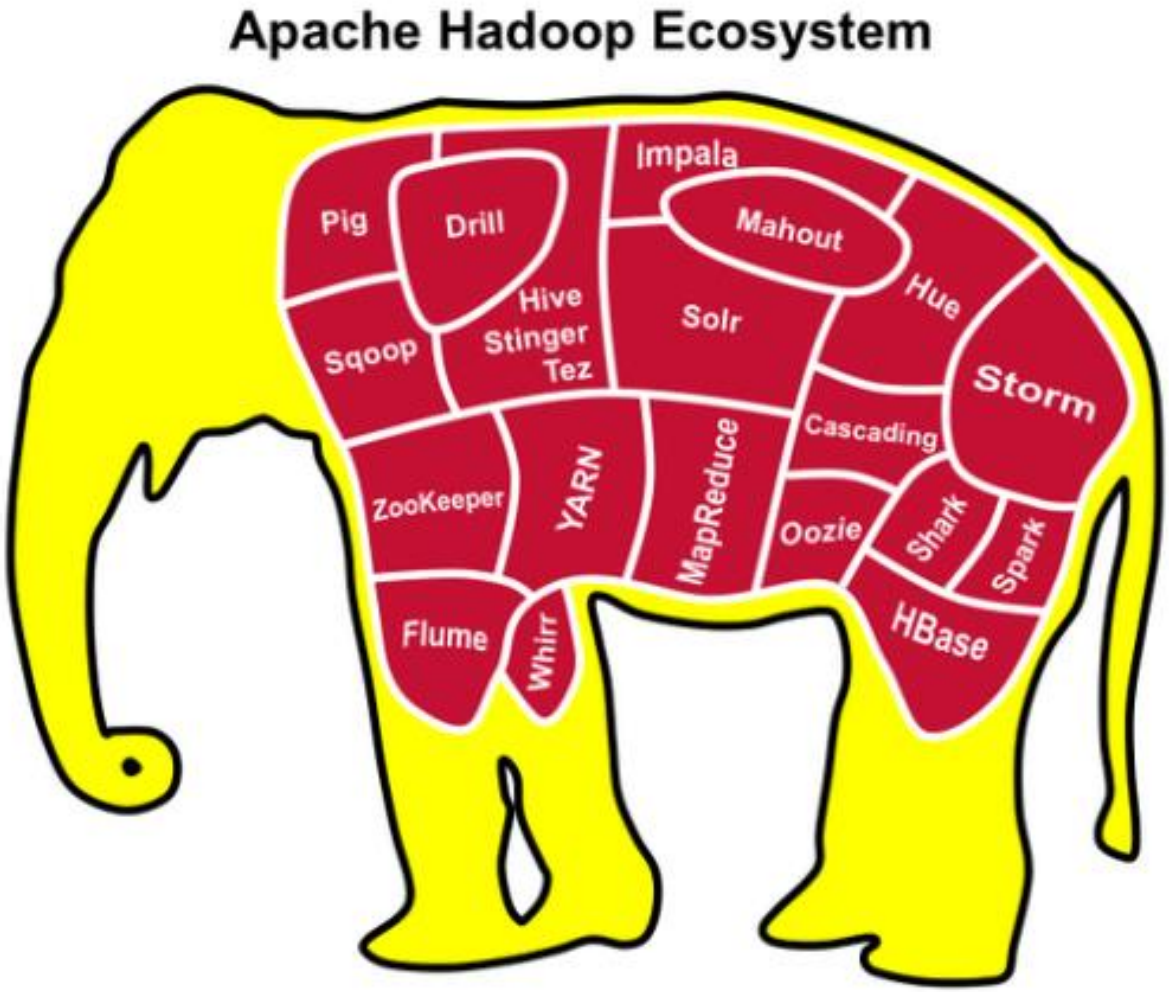
Hadoop? HDFS?

Hadoop Ecosystem

February 19, 2014

NARLabs

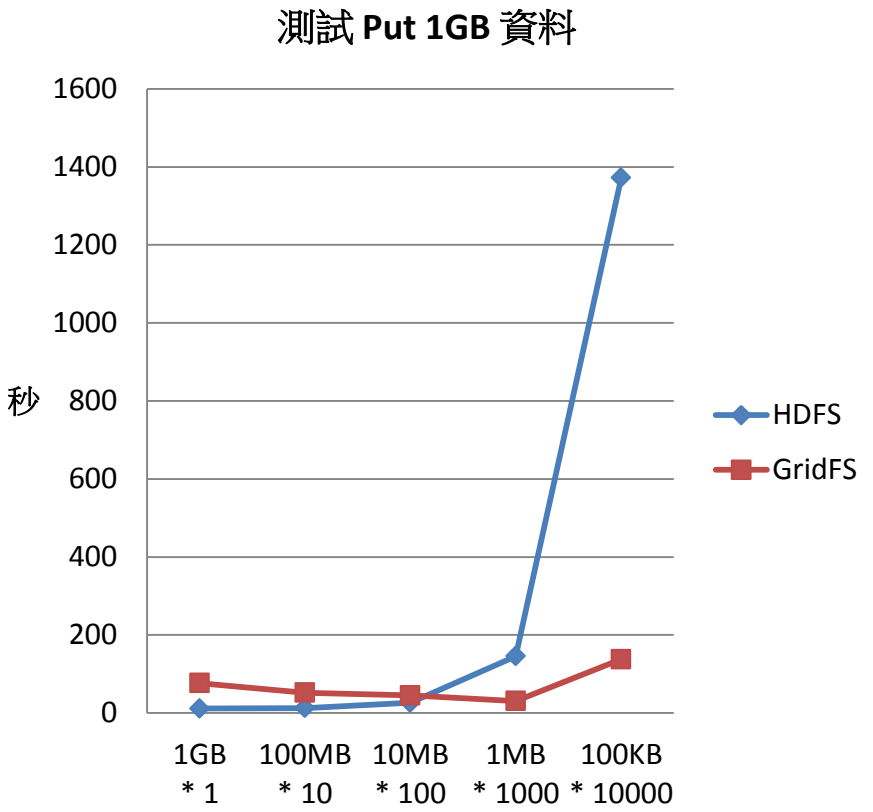
承諾・熱情・創新



https://aryannava.com/2014/02/19/apache-hadoop-ecosystem/hadoopecosystem/

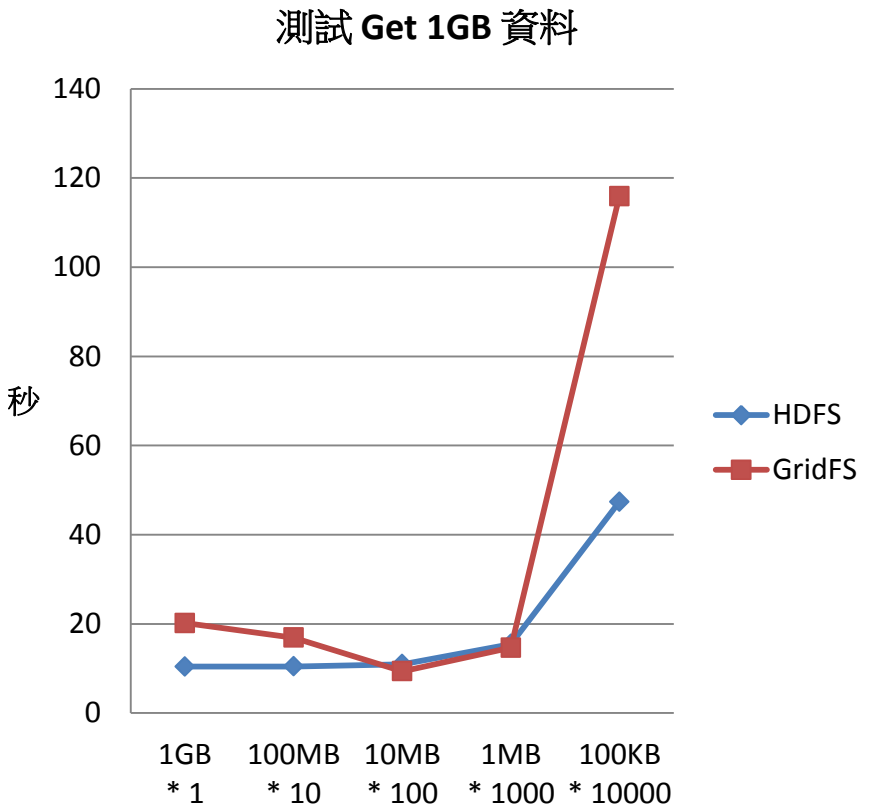
HDFS/GridFS 大量寫入機制測試

測試 Put 1GB 資料		
檔案分割大小	HDFS	GridFS
1GB * 1	11.472	76.805
100MB * 10	12.744	51.872
10MB * 100	26.261	45.041
1MB * 1000	145.489	30.688
100KB * 10000	1372.342	137.665

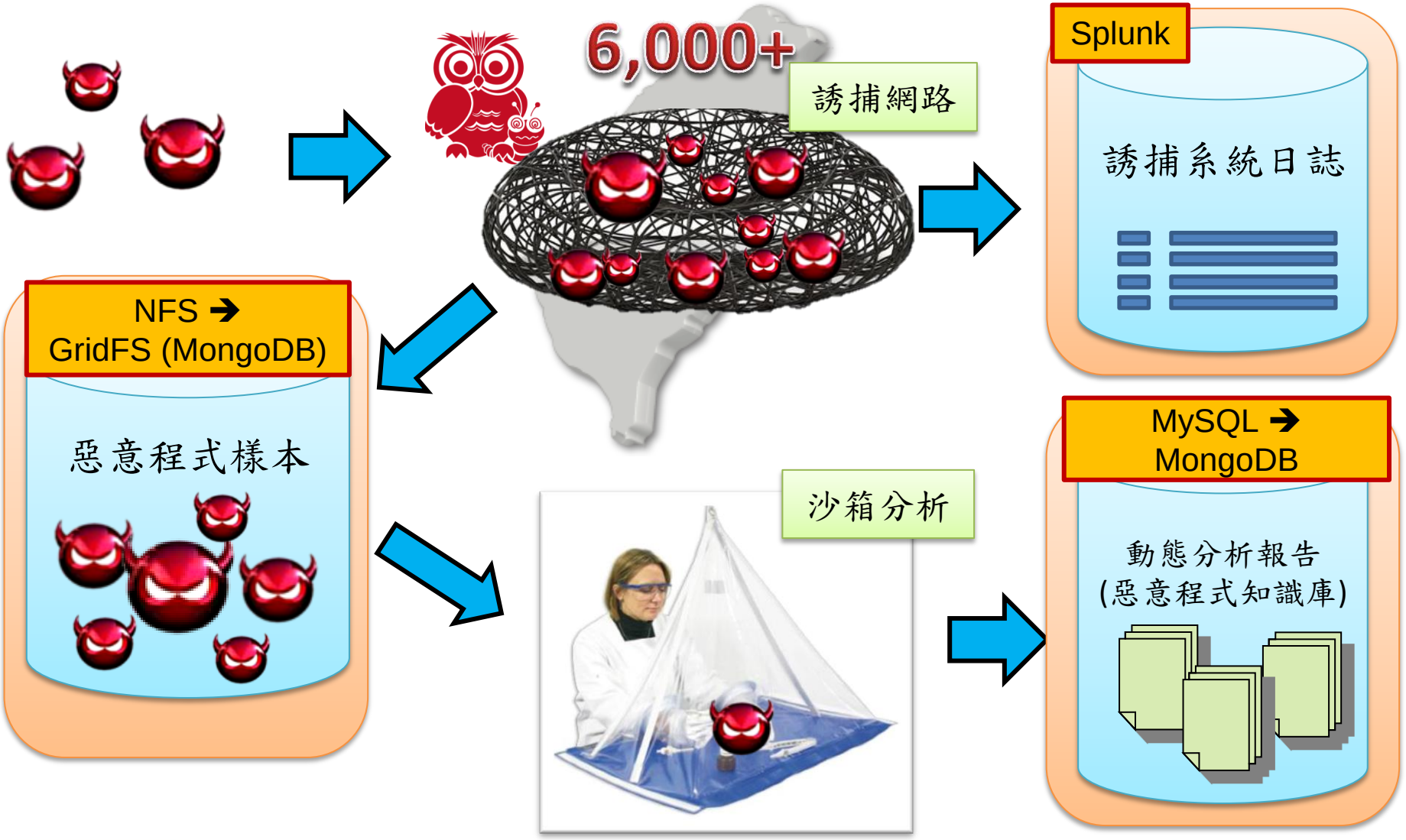


HDFS/GridFS 大量讀取機制測試

測試 Get 1GB 資料		
檔案分割大小	HDFS	GridFS
1GB * 1	10.414	20.155
100MB * 10	10.42	16.916
10MB * 100	10.931	9.367
1MB * 1000	15.475	14.631
100KB * 10000	47.366	115.934



資安威脅大資料架構


6,000+
誘捕網路
Splunk
誘捕系統日誌
NFS ➔
GridFS (MongoDB)
惡意程式樣本
沙箱分析
MySQL ➔
MongoDB
動態分析報告
(惡意程式知識庫)

簡報結束
謝謝您