雲端運算相關應用
(Based on Hadoop)

陳威宇
格網技術組
waue@nchc.org.tw

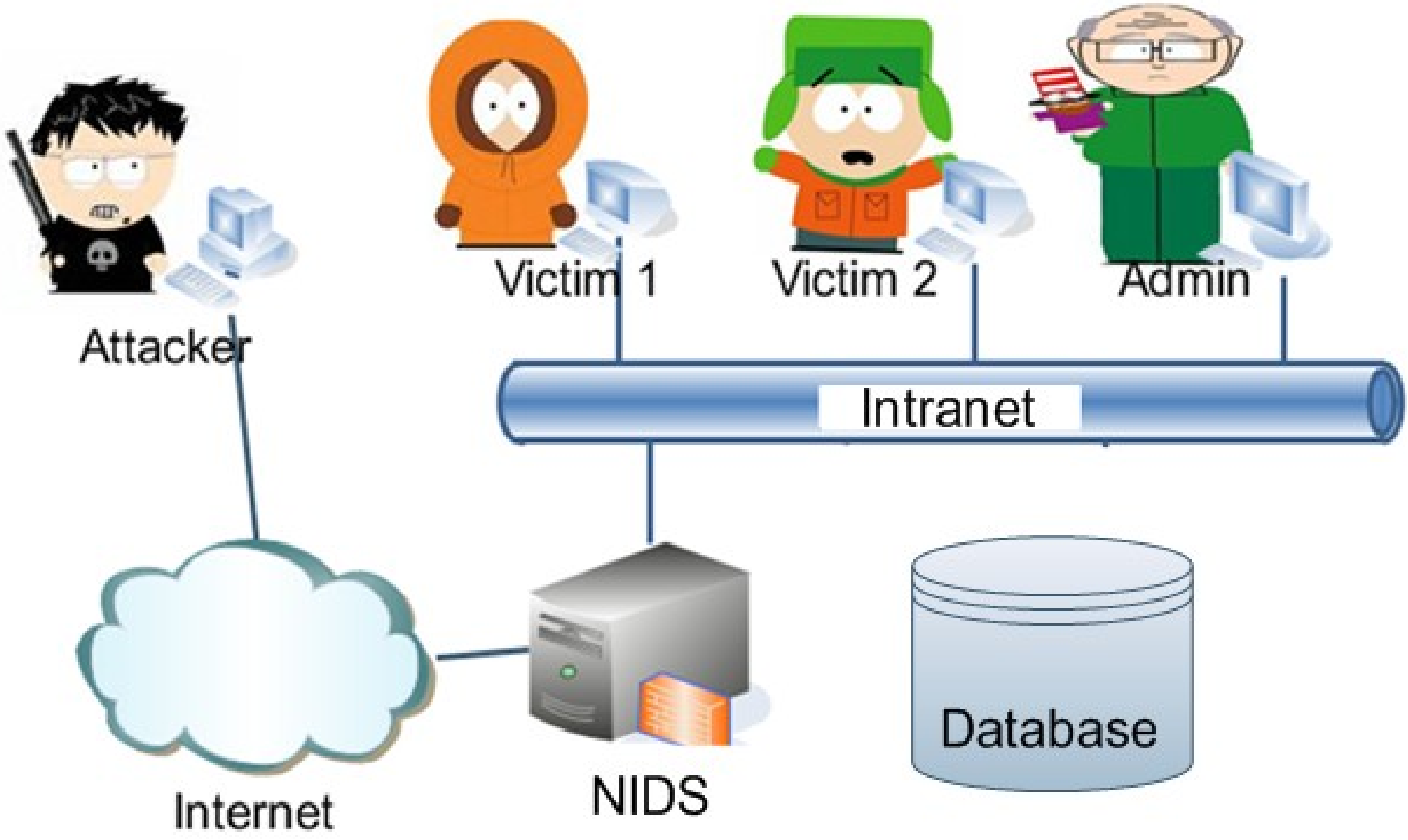
Two Topics

- **ICAS : IDS-Log Analysis System Based on Hadoop and HBase**
- **NutchEz : An Easy Way to Crawl Web Pages by Nutch**

ICAS

IDS-Log Analysis System Based on Hadoop and HBase

網路型入侵偵測系統


Attacker
Victim 1
Victim 2
Admin
Intranet
Internet
NIDS
Database

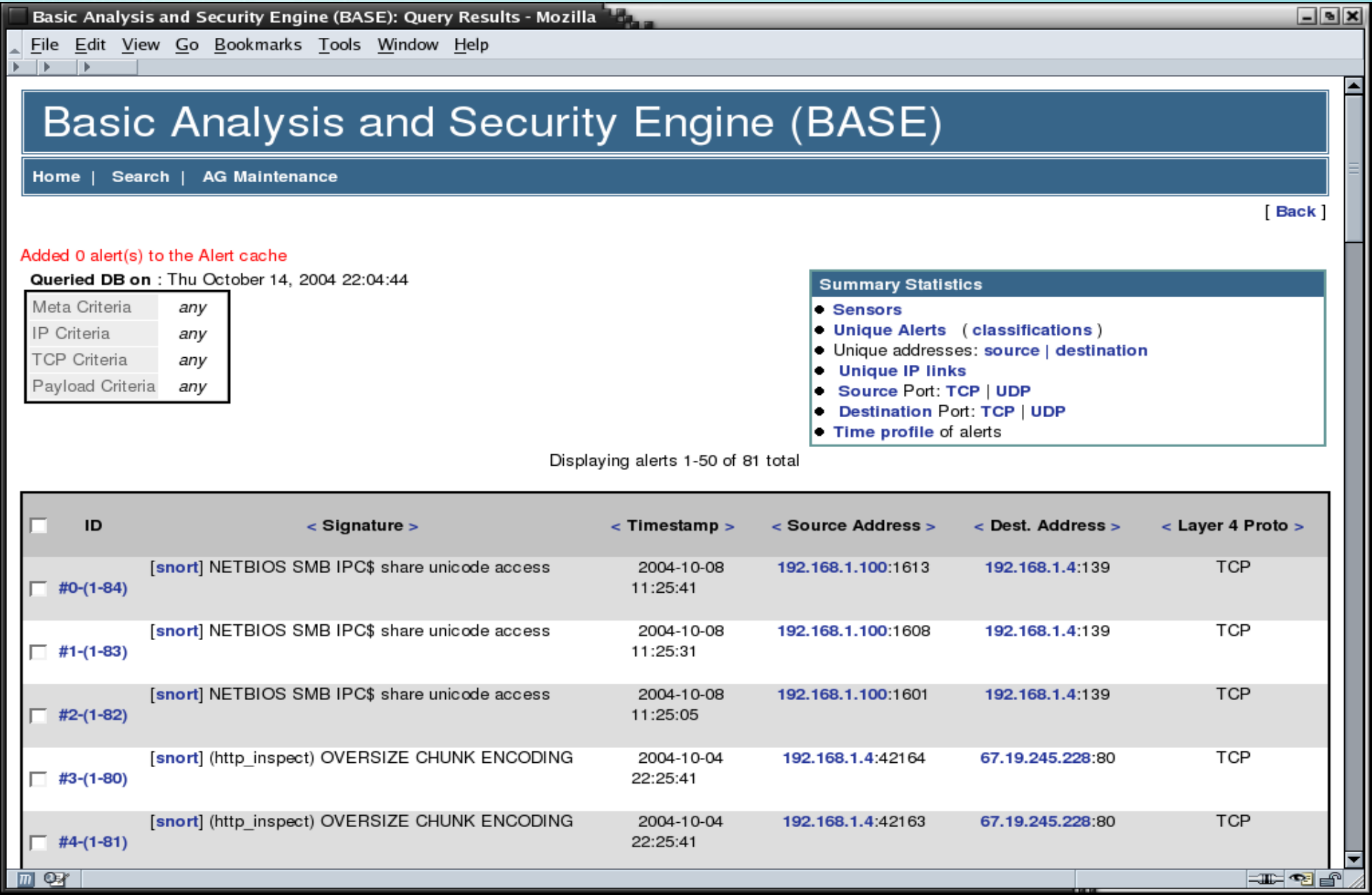
警訊格式

[**] [1:538:15] NETBIOS SMB IPC$ unicode share access [**]
[Classification: Generic Protocol Command Decode] [Priority: 3]
09/04-17:53:56.363811 168.150.177.165:1051 -> 168.150.177.166:139
TCP TTL:128 TOS:0x0 ID:4000 IpLen:20 DgmLen:138 DF
AP Seq: 0x2E589B8 Ack: 0x642D47F9 Win: 0x4241 TcpLen: 20

[**] [1:1917:6] SCAN UPnP service discover attempt [**]
[Classification: Detection of a Network Scan] [Priority: 3]
09/04-17:53:56.385573 168.150.177.164:1032 -> 239.255.255.250:1900
UDP TTL:1 TOS:0x0 ID:80 IpLen:20 DgmLen:161
Len: 133

[**] [1:1917:6] SCAN UPnP service discover attempt [**]
[Classification: Detection of a Network Scan] [Priority: 3]
09/04-17:53:56.386910 168.150.177.164:1032 -> 239.255.255.250:1900
UDP TTL:1 TOS:0x0 ID:82 IpLen:20 DgmLen:161
Len: 133
……

Network IDS Interface



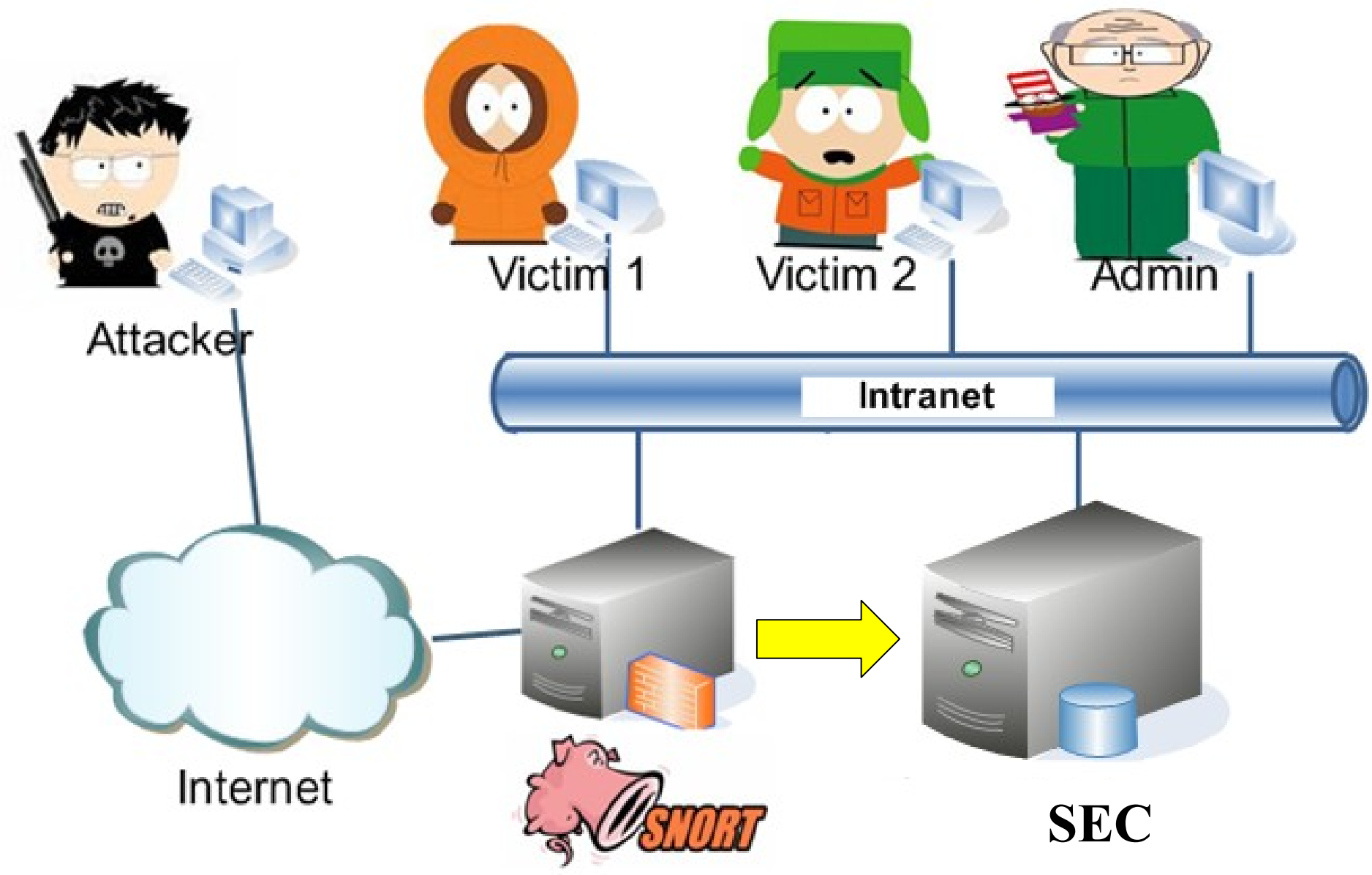
These Events are MIS's Nightmare !!!!

1. 重複的資訊太多
2. 難以瞭解全部的事件
3. 易忽略重要的訊息

The Security Events Center

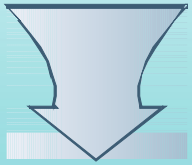
- 資訊安全事件中心
  - 收集、整合、關聯惡意入侵警訊，於一個提供資安事故訊息呈現的平台
- 主要功能
  - 收集資訊
  - 分析並整合事件

SEC Overview



Alert Merge Example

Destination IP	Attack Signature	Source IP	Destination Port	Source Port	Packet Protocol	Timestamp
Host_1	Trojan	Sip1	80	4077	tcp	T1
Host_1	Trojan	Sip2	80	4077	tcp	T2
Host_1	Trojan	Sip1	443	5002	tcp	T3
Host_2	Trojan	Sip1	443	5002	tcp	T4
Host_3	D.D.O.S	Sip3	53	6007	udp	T5
Host_3	D.D.O.S	Sip4	53	6008	tcp	T5
Host_3	D.D.O.S	Sip5	53	6007	udp	T5
Host_3	D.D.O.S	Sip6	53	6008	tcp	T5



Key		Values				
Host_1	Trojan	Sip1,Sip2	80,443	4077,5002	tcp	T1,T2,T3
Host_2	Trojan	Sip1	443	5002	tcp	T4
Host_3	D.D.O.S.	Sip3,Sip4,Sip5 ,Sip6	53	6007,6008	tcp, udp	T5

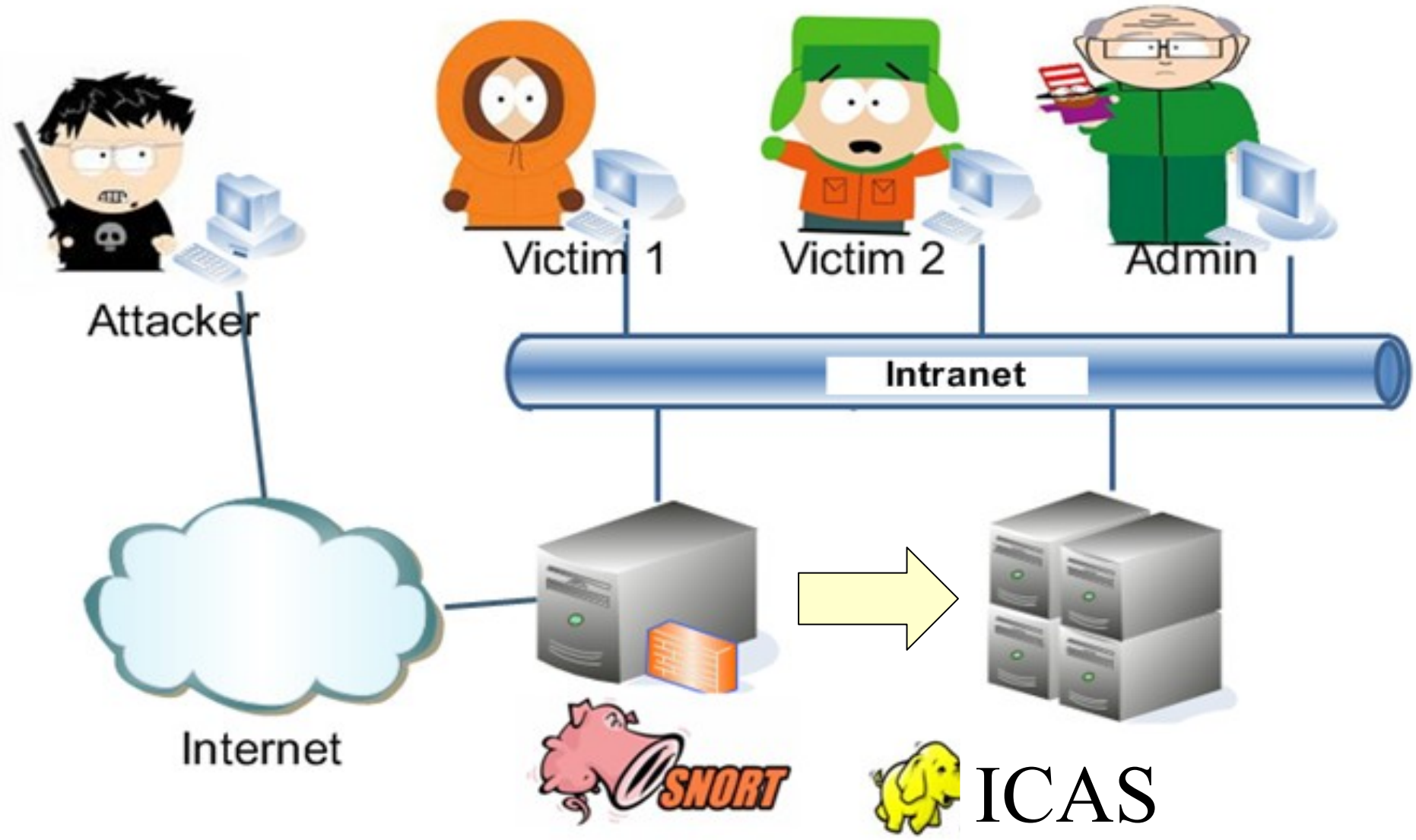
What's problem about the SEC ?

1. 大量的資料將導致效能變差
2. 資料庫毀損
3. 執行分析時，系統資源忙碌

ICAS

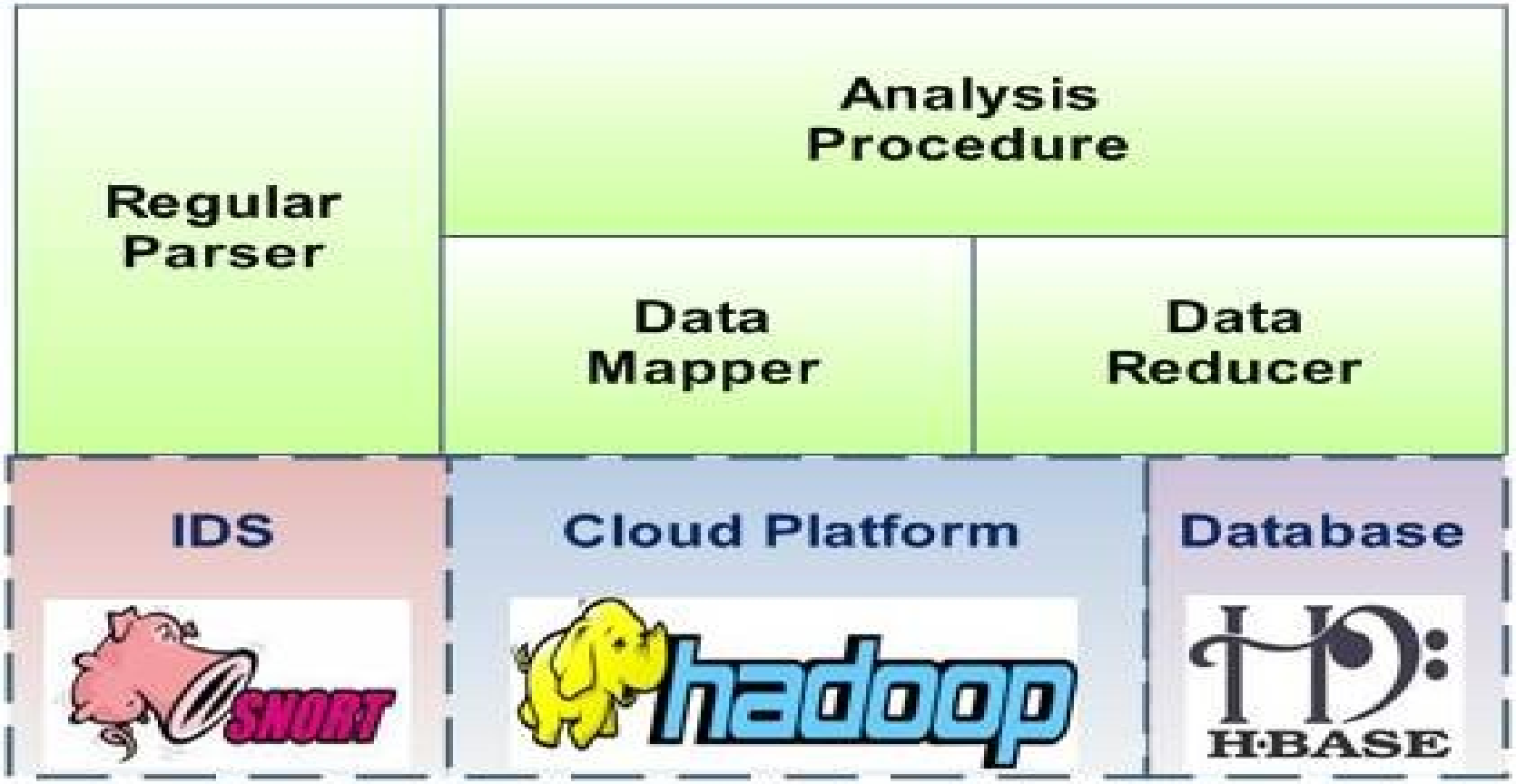
- ICAS, *IDS Cloud Analysis System*
- 透過雲端運算
  - Higher capability
  - Fault tolerance
- 主要分析功能
  - Reducing redundancy
  - Merge relation

ICAS Overview

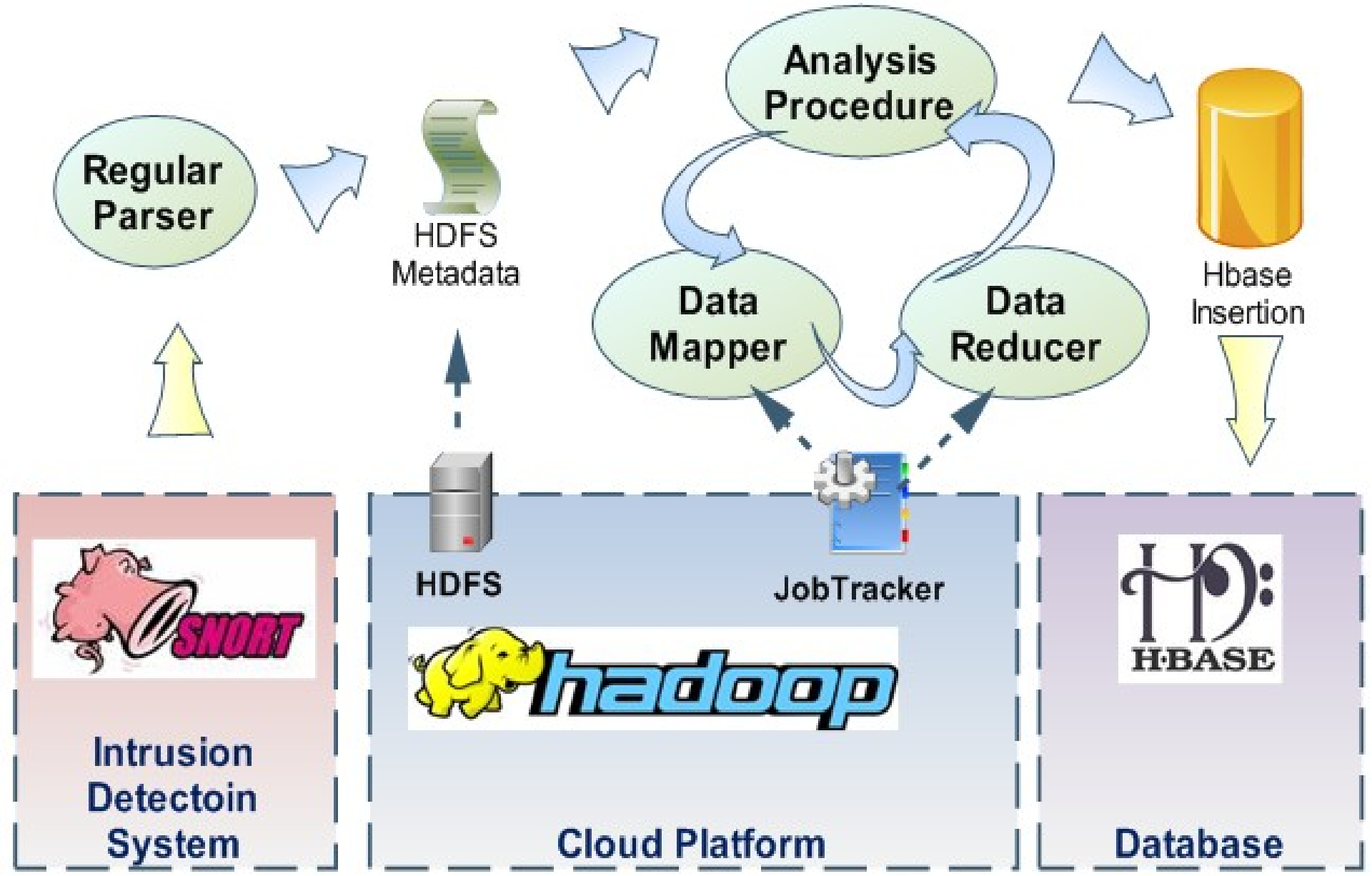


System Architecture

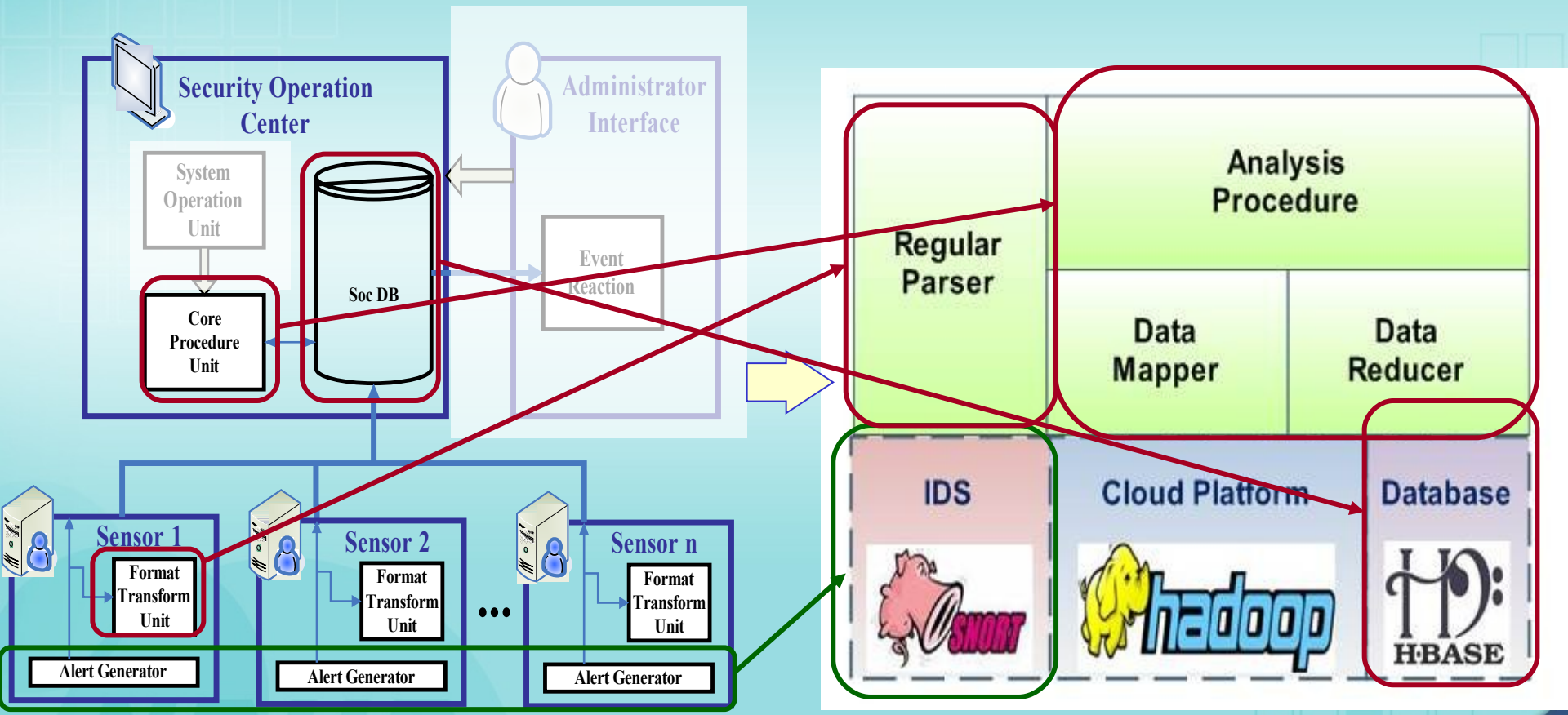
ICAS Component Overview



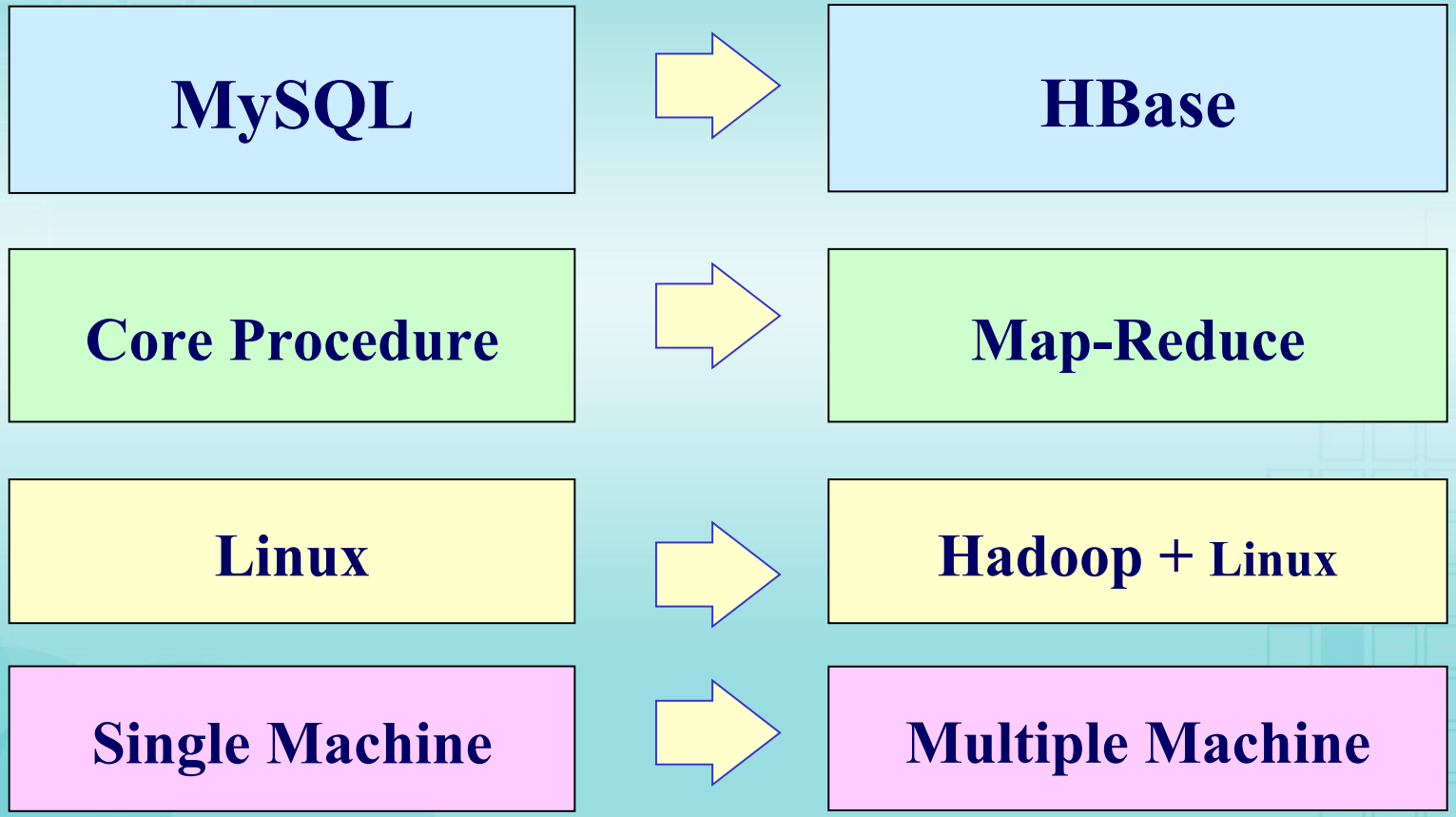
Program Procedure



Change SEC to ICAS (Architecture)

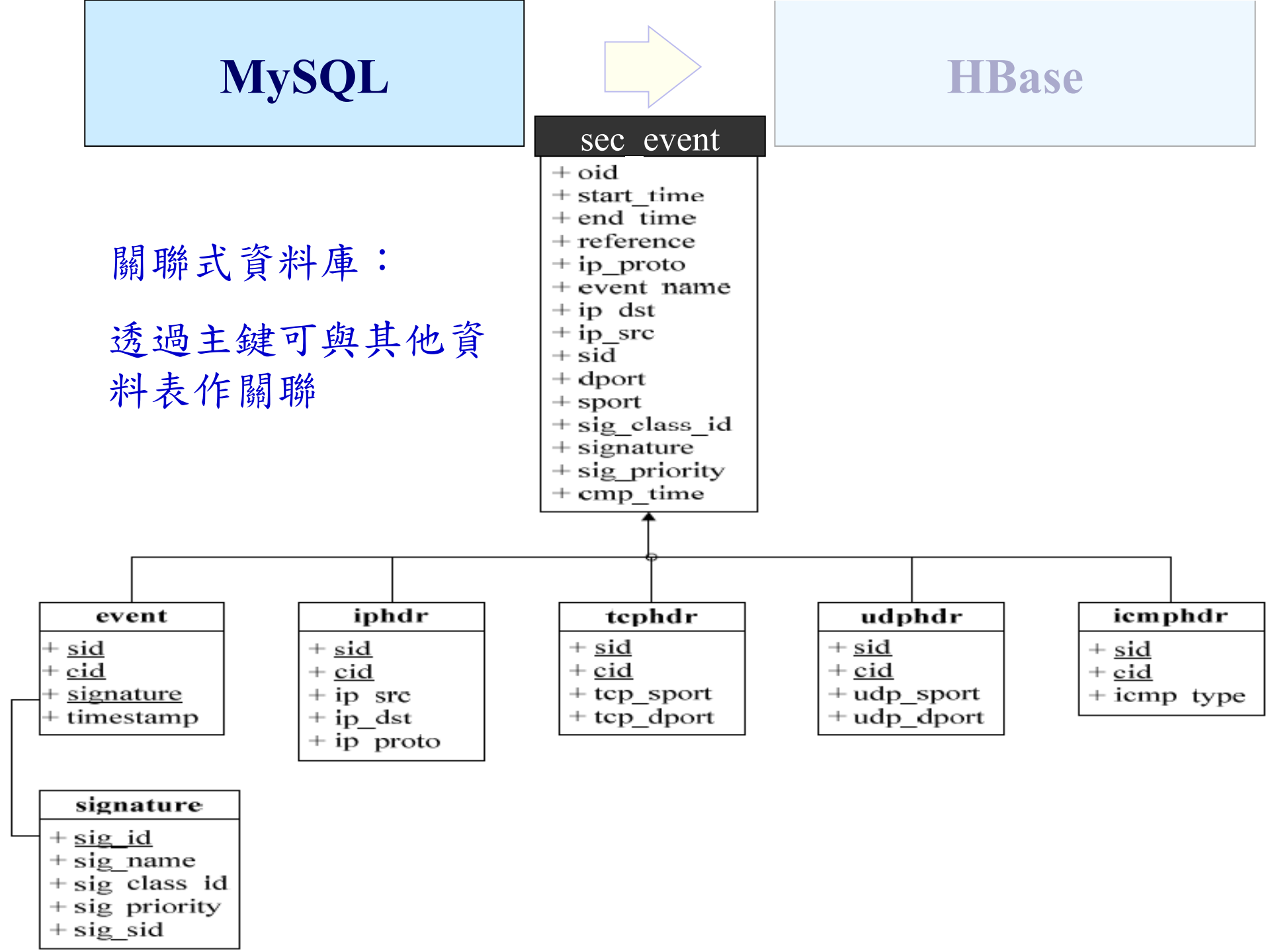


Change SEC to ICAS (components)



- **Format Transfer Unit**
  - Setup Snort logging to MySQL
  - Setup MySQL client logging to remote MySQL server
- **Core Procedure Unit**
  - Fuse redundant data
  - Merge data as event
- **Program language**
  - Shell & PHP

- **Regular Parser**
  - Parsing original snort log and transfer to HDFS (hadoop file system)
- **Analysis Procedure**
  - Dispatch job if pool is not empty and insert the result into database
- **Data Mapper**
  - <key, value> mapping
- **Data Reducer**
  - <“key1”, value1…valueN>
  - <“key2”, value1…valueN>


MySQL
HBase
sec_event
+ oid
+ start_time
+ end_time
+ reference
+ ip_proto
+ event_name
+ ip_dst
+ ip_src
+ sid
+ dport
+ sport
+ sig_class_id
+ signature
+ sig_priority
+ cmp_time
關聯式資料庫：
透過主鍵可與其他資料表作關聯
event
+ sid
+ cid
+ signature
+ timestamp
iphdr
+ sid
+ cid
+ ip_src
+ ip_dst
+ ip_proto
tcphdr
+ sid
+ cid
+ tcp_sport
+ tcp_dport
udphdr
+ sid
+ cid
+ udp_sport
+ udp_dport
icmphdr
+ sid
+ cid
+ icmp_type
signature
+ sig_id
+ sig_name
+ sig_class_id
+ sig_priority
+ sig_sid

MySQL → HBase

Row Key	Time Stamp	Column "signature:"	Column "Infor:"		Column " SourceIP :"
(Destination IP) dIP1	t1	sig1	Infor :Port	p1,p2..	sIP1
	t2	sig2	Infor : ... (the other info)	o1,o2…	sIP2
dIP2	…	…	Infor :…	..	…
…	…	…	…	..	…

雲端資料庫：

格式為三個維度(Row Key, TimeStamp, Column)

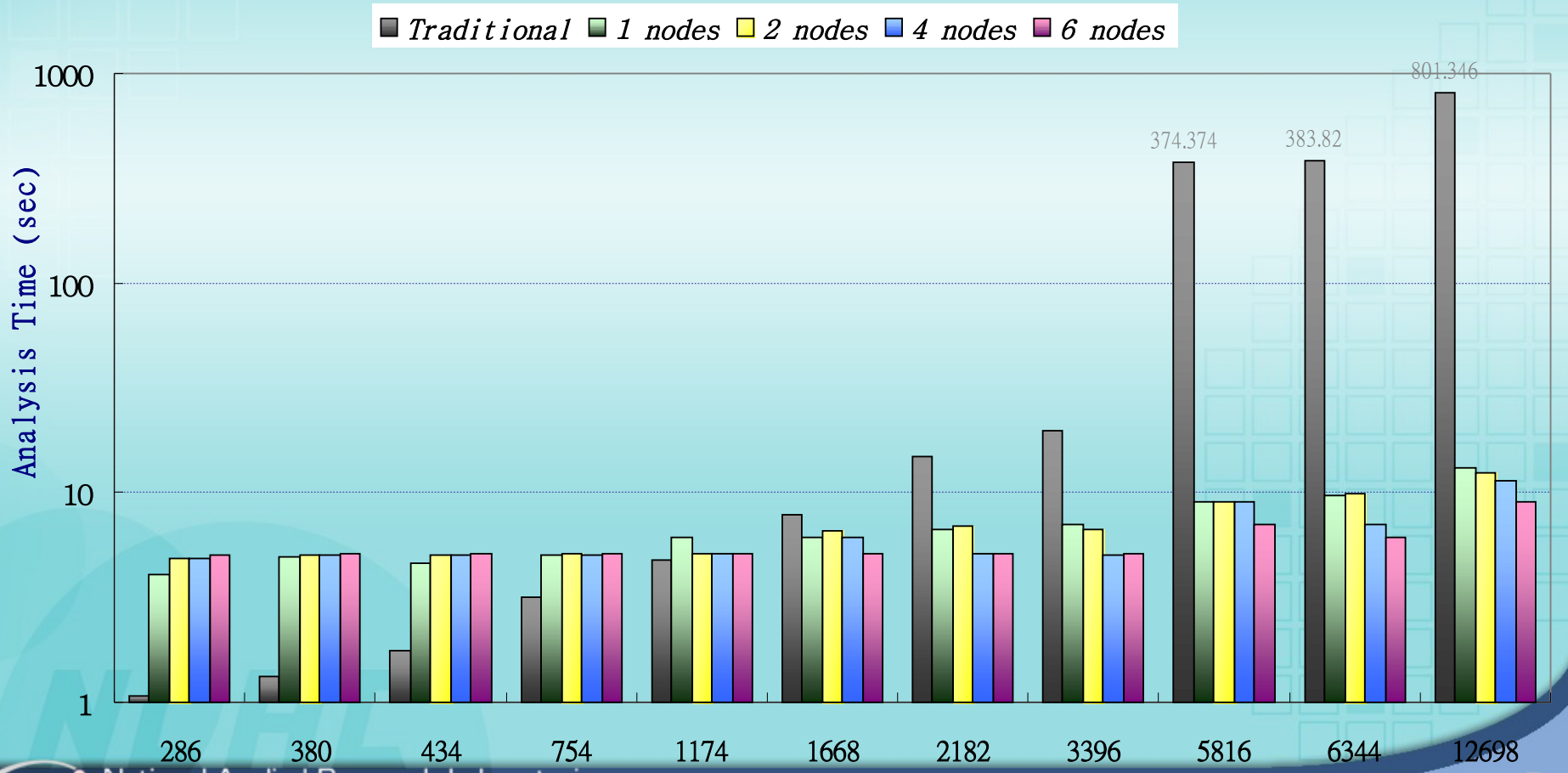
搭配雲端運算架構

實驗環境

- **Machine: X6**
  - CPU : Intel quad-core, Memory : 2g,
- **OS :** Linux : Ubuntu 8.04 server
- **Software : version**
  - Hadoop : 0.16.4
  - Hbase : 0.1.3
  - Java : 6
- **Alerts Data Sets**
  - MIT Lincoln Laboratory, Lincoln Lab Data Sets
  - Computer Security group at UCDavis, tcpdump file

Experimental Result

The Calculation Time of Each Number of Data Sets



Experimental Result

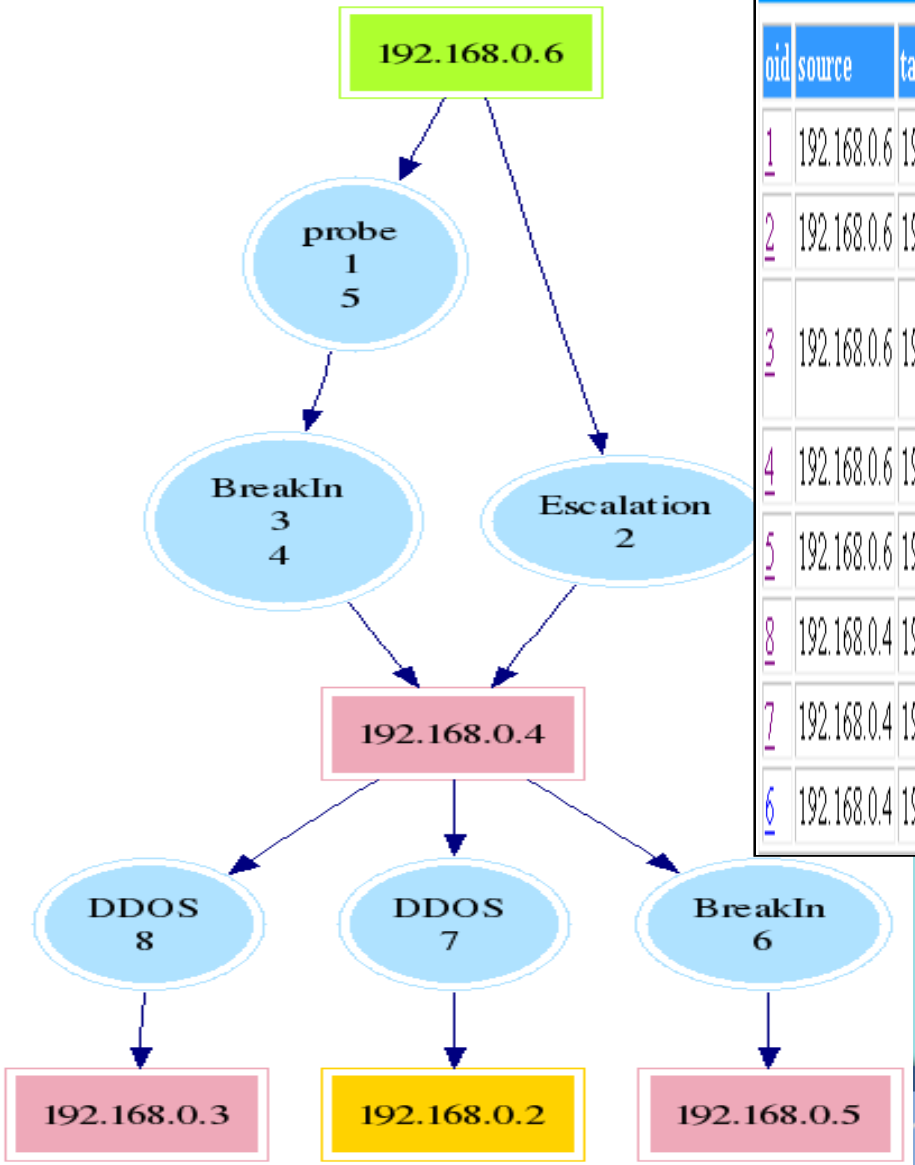
Throughput Data Overall

Origianl Alerts	Analysis Time (sec)					Results	Reduction Rate
	Traditional	1 nodes	2 nodes	4 nodes	6 nodes		
286	1.068	4.087	4.869	4.864	5.077	30	89.51%
380	1.333	4.94	5.069	5.067	5.097	11	97.11%
434	1.76	4.61	5.066	5.068	5.09	9	97.93%
754	3.145	5.066	5.079	5.038	5.096	16	97.88%
1174	4.73	6.066	5.093	5.089	5.097	33	97.19%
1668	7.909	6.07	6.56	6.071	5.082	16	99.04%
2182	14.949	6.671	6.95	5.166	5.088	16	99.27%
3396	19.901	7.053	6.654	5.076	5.091	68	98.00%
5816	374.374	9.081	9.076	9.07	7.076	66	98.87%
6344	383.82	9.68	9.872	7.069	6.069	72	98.87%
12698	801.346	13.096	12.367	11.367	9.083	36	99.72%

ICAS：結論

- 由實驗結果可看出，雲端運算處理資料格式相似且資料量大的情況下，能展現其效益，並提供高容錯率、低獨占系統資源、多工作同時執行等能力
- ICAS 的特性適用於 Map/Reduce 演算法，故即使都是一個運算節點的環境下， ICAS 也在大資料量的分析有較好得效率
- Hadoop 不適用要求即時性高、或是 latency 低的系統，且每個版本的 API 差異大
- 關聯式資料庫對小量資料的讀寫的效率較好，並且支援的語言也較多（如下頁）

ICAS：結論 (2)



associate ticket !! total number = 8

oid	source	target	class	signature name
1	192.168.0.6	192.168.0.4	probe	NETBIOS SMB-DS IPC$ unicode share access
2	192.168.0.6	192.168.0.4	Escalation	SHELLCODE x86 0x90 unicode NOOP
3	192.168.0.6	192.168.0.4	BreakIn	NETBIOS SMB-DS lsass DsRolerUpgradeDownlevelServer WriteAndX unicode little endian overflow attempt
4	192.168.0.6	192.168.0.4	BreakIn	NETBIOS SMB-DS lsass DsRolerUpgradeDownlevelServer unicode little endian overflow attempt
5	192.168.0.6	192.168.0.4	probe	MISC MS Terminal server request
8	192.168.0.4	192.168.0.3	DOS	DDOS Trin00 Master to Daemon default password attempt
7	192.168.0.4	192.168.0.2	DOS	DDOS Trin00 Master to Daemon default password attempt
6	192.168.0.4	192.168.0.5	BreakIn	NETBIOS DCERPC ISystemActivator path overflow attempt little endian unicode

NutchEz :

An Easy Way to Crawl Web Pages by Nutch

公司內部文件問題

- 有些內部資料雖放在網路上，但不適合對外公開，僅在內部網路中的員工可以讀取
  - 搜尋引擎 ..X => 靠印象找資料 ..O
    新人…囧 rz
- 方法：
  - 建立資料庫文件查詢系統： MIS=> 資料庫
  - 用分類法建立樹狀資料結構：容易誤會
- 以上缺點：無法全文查詢

解決辦法

- 建立屬於公司內部的搜尋引擎
  - 解析網頁內容
    - 支援各種網頁格式　html, php, jsp…
  - 統一的搜尋窗口
    - 不同網站於不同主機，同一窗口
    - 不用選擇資料類別
  - 成本小
  - 無痛
  - 保密

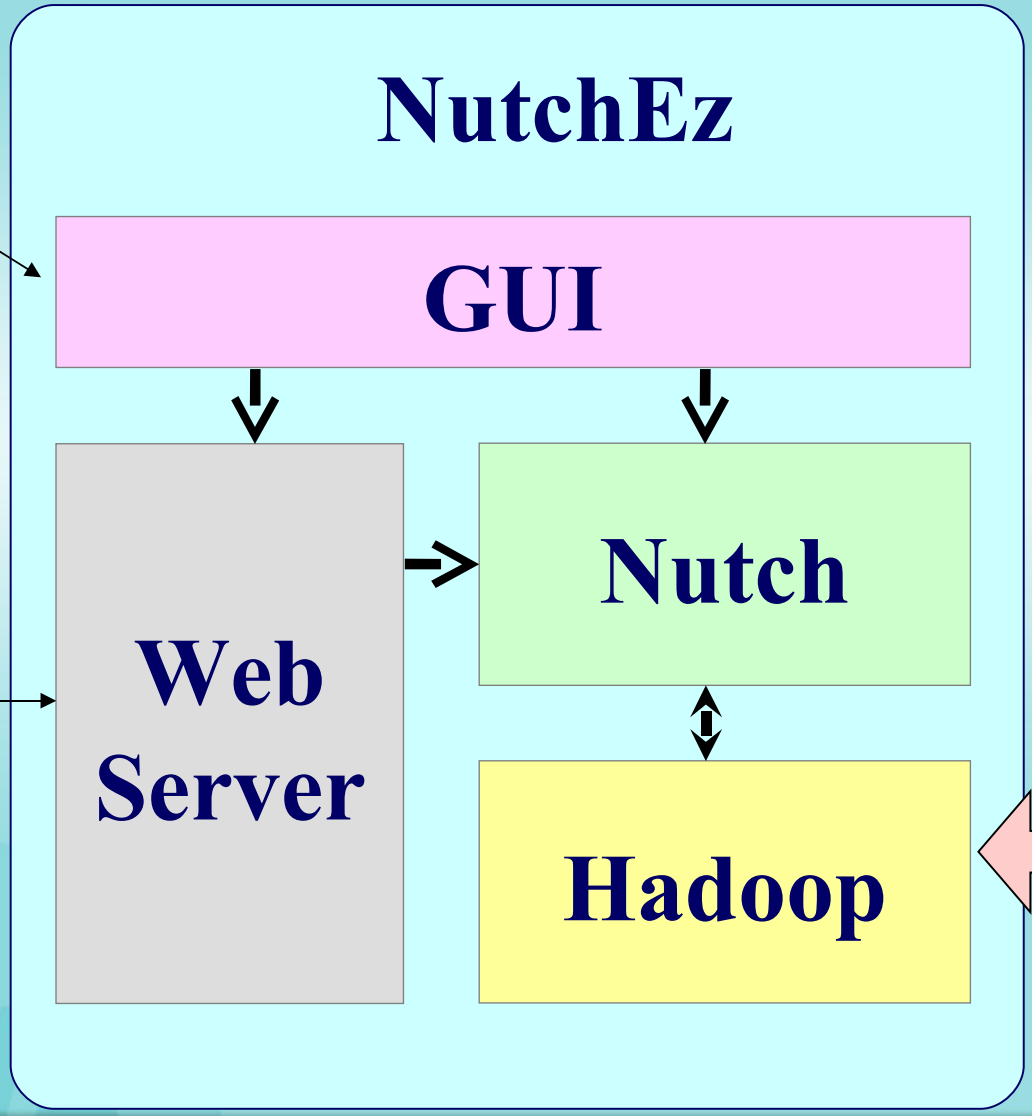
NutchEz

- 全自動的搜尋解決方案
- 安裝簡單、操作方便
- 效率高、支援格式多、功能強大
- 開放原始碼
- http://trac.nchc.org.tw/cloud/wiki/NutchEz

NutchEz 系統架構

管理者

使用者



What's Nutch

- 以 Java 來實做的 open source 搜索引擎
- 與 Hadoop 為同一創始者
- 以 Hadoop 為運算平台
- 目標：
  - 一個月抓取幾十億網頁
  - 為這些網頁維護索引
  - 對索引文件進行每秒上千次的搜索
  - 提供精準的搜索結果
  - 以最小的成本運作

NutchEz ： Nutch 的整合套件

- 簡易
  - 安裝與操作都很簡便
- 透明
  - Opensource ，資訊不隱藏
- 廣泛
  - 可分析不同檔案格式
- 隱私
  - 可應用於搜尋內部資料
- 客製化
  - 可設計成專用的 data mining 工具

可分析的格式與網路協定

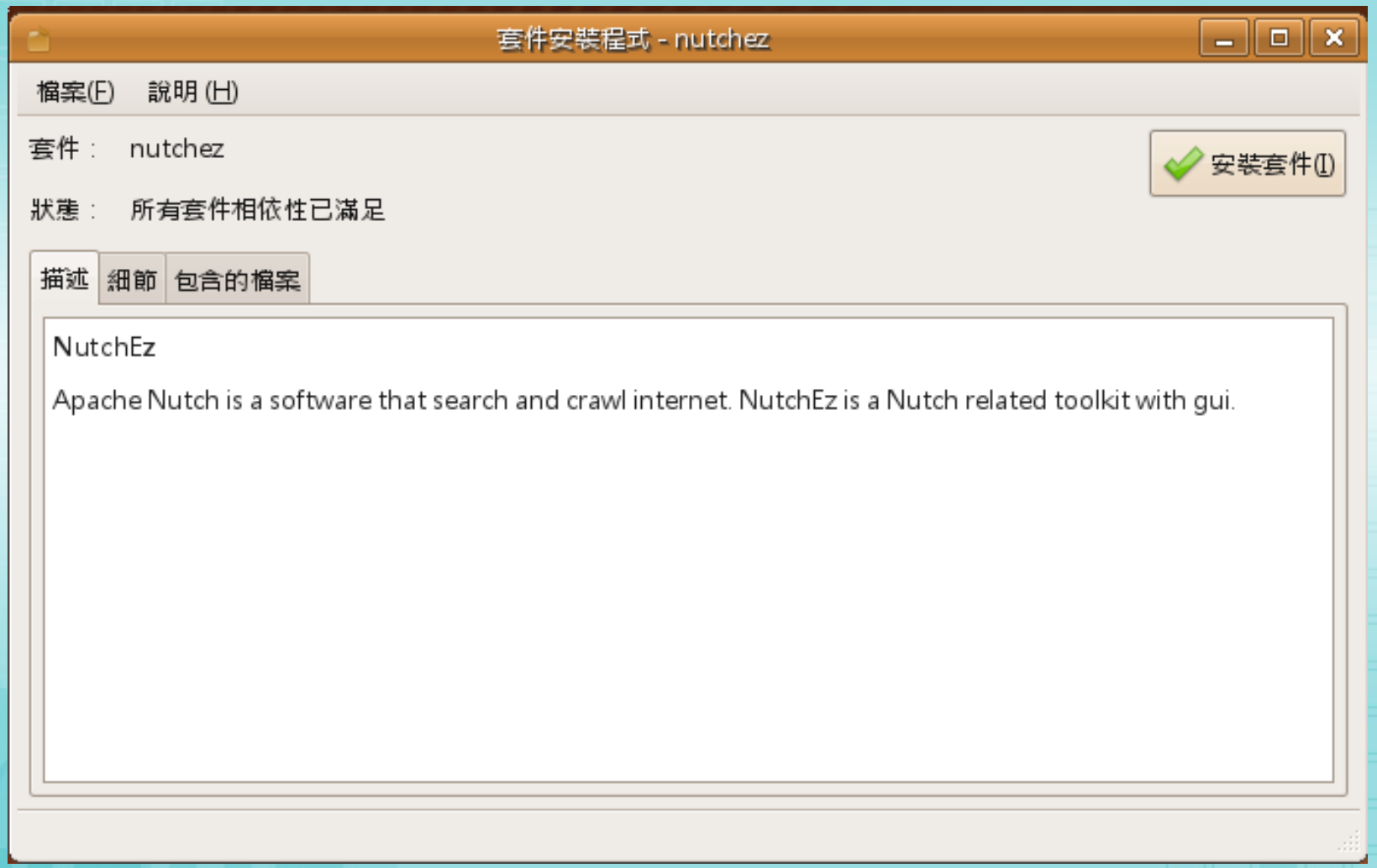
分析檔案格式

text	msword
ext	msexcel
html	msppt
js	pdf
mp3	rss
zip	openoffice
rtf	swf

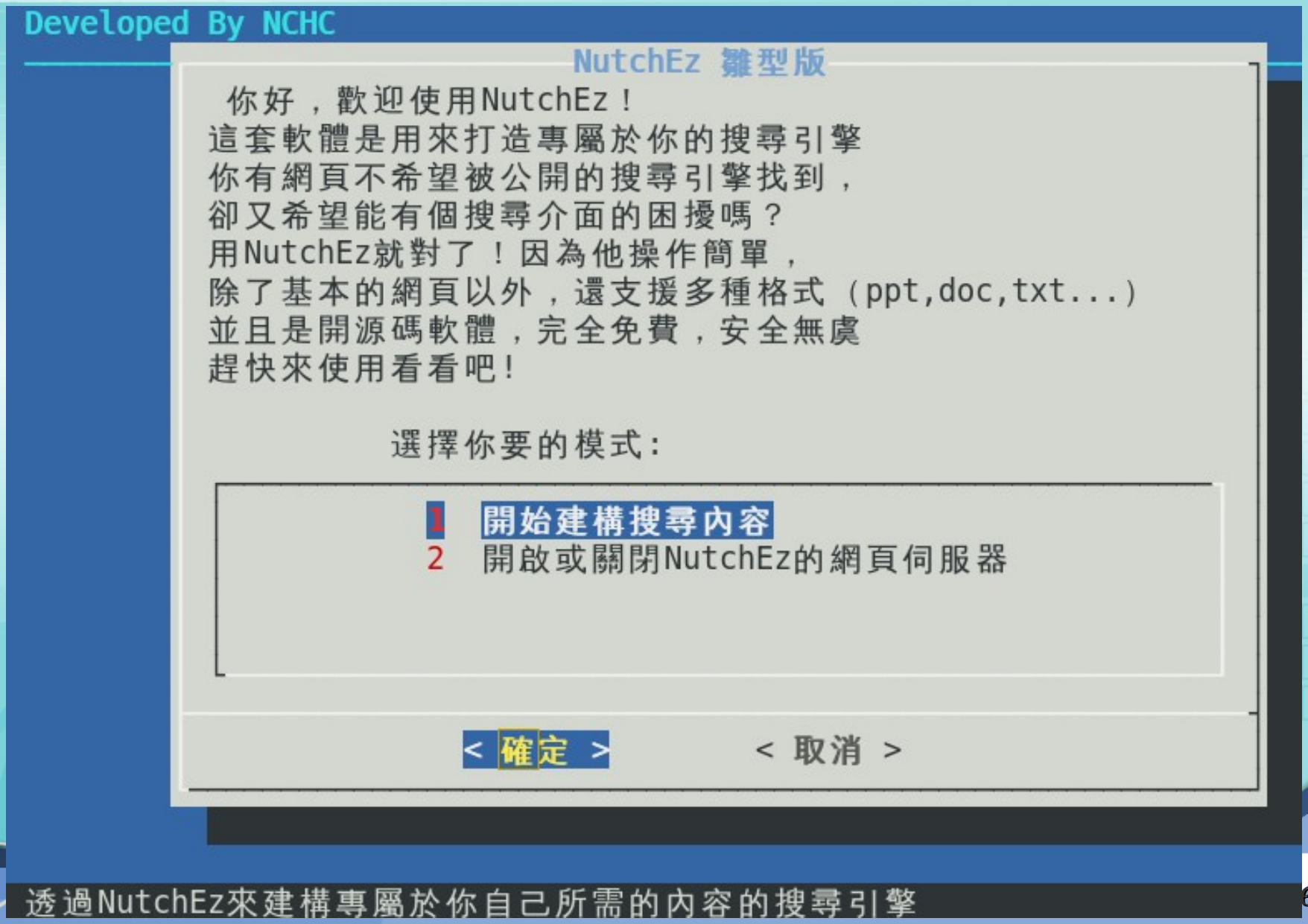
網路協定

- file
- ftp
- http
- https

安裝

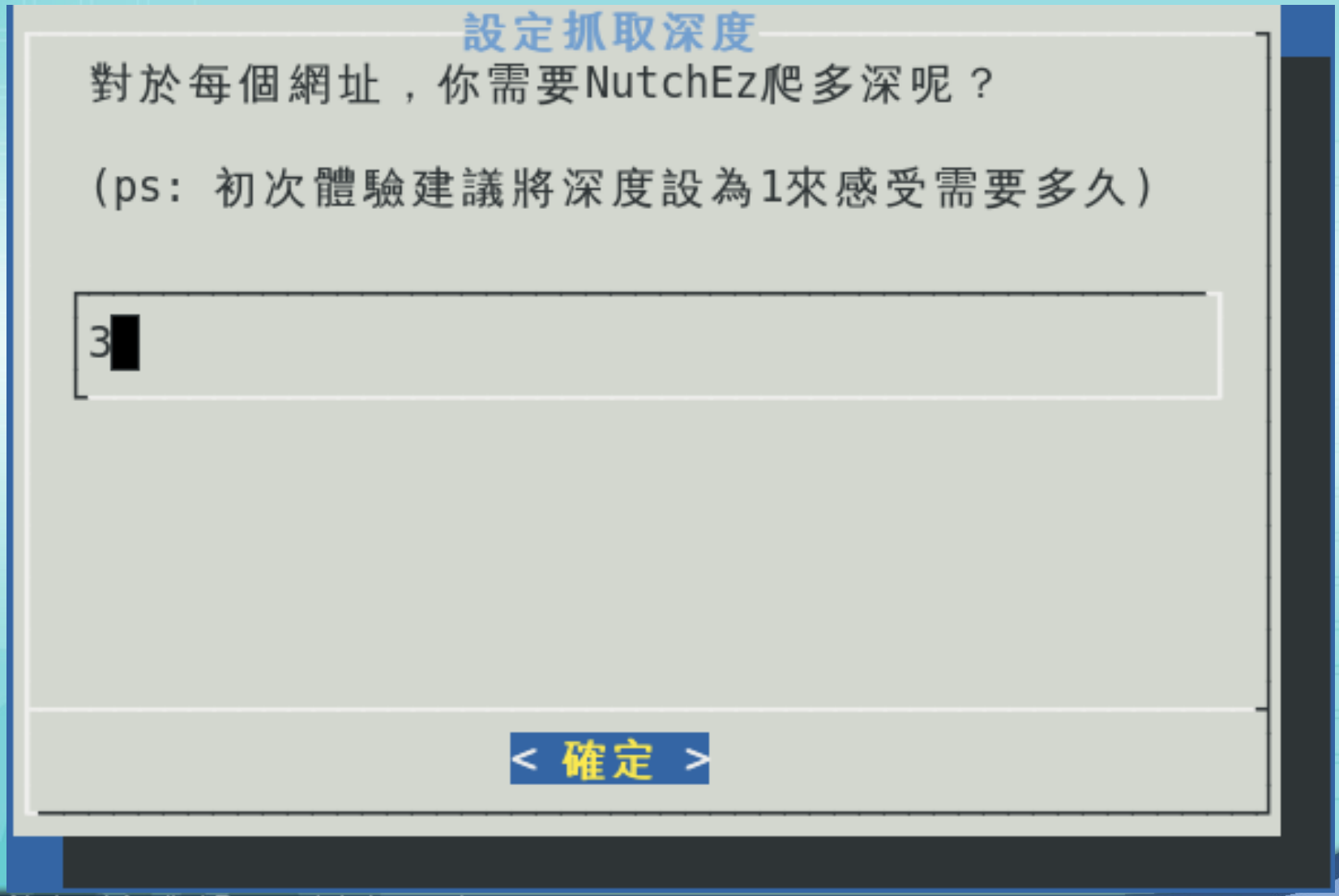

套件安裝程式 - nutchez
檔案(F)
說明(H)
套件： nutchez
安裝套件(I)
狀態： 所有套件相依性已滿足
描述
細節
包含的檔案
NutchEz
Apache Nutch is a software that search and crawl internet. NutchEz is a Nutch related toolkit with gui.

使用 - 建構搜尋內容 (1/5)


Developed By NCHC
NutchEz 雛型版
你好，歡迎使用NutchEz！
這套軟體是用來打造專屬於你的搜尋引擎
你有網頁不希望被公開的搜尋引擎找到，
卻又希望能有個搜尋介面的困擾嗎？
用NutchEz就對了！因為他操作簡單，
除了基本的網頁以外，還支援多種格式（ppt,doc,txt...）
並且是開源碼軟體，完全免費，安全無虞
趕快來使用看看吧!
選擇你要的模式:
1 開始建構搜尋內容
2 開啟或關閉NutchEz的網頁伺服器
< 確定 >
< 取消 >

使用 - 建構搜尋內容 (2/5)

使用 - 建構搜尋內容 (3/5)


設定抓取深度
對於每個網址，你需要NutchEz爬多深呢？
(ps: 初次體驗建議將深度設為1來感受需要多久)
3
< 確定 >

使用 - 建構搜尋內容 (4/5)


設定網頁伺服器
你希望NutchEz將網頁伺服器開在哪個port
(ps: 請選擇一個沒用到的port以免造成衝突
也請盡量不要設成80以免造成你誤以為是apache的混淆)
8080
< 確定 >

使用 - 建構搜尋內容 (5/5)

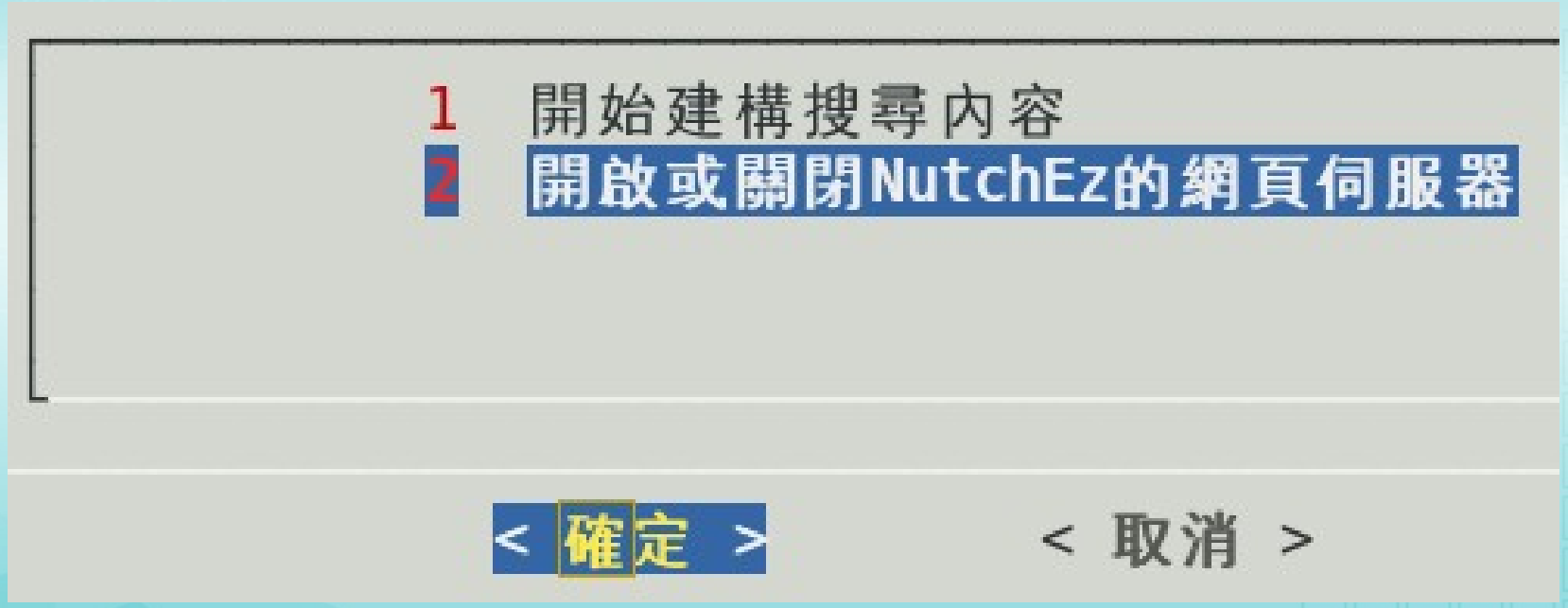
```
 1. 你所選擇要爬取的網址為 :
http://www.hadoop.tw/

 2. 對於這個爬網機器人，你取名為 :
nutchez

 3. 爬網的深度，你設定為 :
3

 4. NutchEz將會把你的搜尋結果呈現在這個Port :
8080

 5. 是否要清除上一次的收尋結果繼續搜尋 :
YES
```

< ok > <reset> <exit >

Running …

```
09/06/09 18:13:49 INFO crawl.Crawl: crawl started in: /home/waue/.nutchez/search
09/06/09 18:13:49 INFO crawl.Crawl: rootUrlDir = /home/waue/.nutchez/urls
09/06/09 18:13:49 INFO crawl.Crawl: threads = 10
09/06/09 18:13:49 INFO crawl.Crawl: depth = 2
09/06/09 18:13:49 INFO crawl.Injector: Injector: starting
09/06/09 18:13:49 INFO crawl.Injector: Injector: crawlDb: /home/waue/.nutchez/search/crawld
09/06/09 18:13:49 INFO crawl.Injector: Injector: urlDir: /home/waue/.nutchez/urls
09/06/09 18:13:49 INFO crawl.Injector: Injector: Converting injected urls to crawl db entri
09/06/09 18:13:49 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker
09/06/09 18:13:49 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments
mplement Tool for the same.
09/06/09 18:13:50 INFO mapred.FileInputFormat: Total input paths to process : 1
09/06/09 18:13:50 INFO mapred.JobClient: Running job: job_local_0001
09/06/09 18:13:50 INFO mapred.FileInputFormat: Total input paths to process : 1
09/06/09 18:13:50 INFO mapred.MapTask: numReduceTasks: 1
09/06/09 18:13:50 INFO mapred.MapTask: io.sort.mb = 100
09/06/09 18:13:50 INFO mapred.MapTask: data buffer = 79691776/99614720
09/06/09 18:13:50 INFO mapred.MapTask: record buffer = 262144/327680
09/06/09 18:13:50 INFO plugin.PluginRepository: Plugins: looking in: /opt/nutch/plugins
09/06/09 18:13:50 WARN plugin.PluginRepository: java.io.FileNotFoundException: /opt/nutch/p
(No such file or directory)
09/06/09 18:13:50 INFO plugin.PluginRepository: Plugin Auto-activation mode: [true]
09/06/09 18:13:50 INFO plugin.PluginRepository: Registered Plugins:
09/06/09 18:13:50 INFO plugin.PluginRepository:         Pdf Parse Plug-in (parse-pdf)
09/06/09 18:13:50 INFO plugin.PluginRepository:         Jakarta POI - Java API To Access Mi
ib-jakarta-poi)
09/06/09 18:13:50 INFO plugin.PluginRepository:         HTTP Framework (lib-http)
09/06/09 18:13:50 INFO plugin.PluginRepository:         More Indexing Filter (index-more)
09/06/09 18:13:50 INFO plugin.PluginRepository:         Regex URL Filter (urlfilter-regex)
09/06/09 18:13:50 INFO plugin.PluginRepository:         More Query Filter (query-more)
```

控制 – 網頁伺服器



實例：

- 機器： CPU Quad 4 2.4G / 4G mem
- 運作時系統平均使用率：
  - CPU 19% 、 MEM 20%
- 搜尋內容：
  - 699 doc, 322 pdf, 9 ppt, 13 odt.
- 費時： 11 min
- Demo: http://secuse.nchc.org.tw:8080

結論： NutchEz

- Nutchez 是一套 Opensource 的搜尋引擎套件，核心為強大的 Nutch ，建構於 Hadoop 之上
- 適用於建立內部資料的索引、分析各種檔案格式，且不會存放原始檔案
- 目前不支援搜索需登入帳號密碼的網站，也無提供搜尋後統計資料

Thank You !

&

Question ?