Data Analytics Platform using Docker (Lab手冊)

Ching-Hao, Eric, Mao Ph. D. & Ro, Lucas, Ko
chmao@iii.org.tw

Outline

- 課程介紹 15 mins
- 資安威脅趨勢 1 hours 15 mins
  - 企業內、雲端及IoT
- Lab (4 hours)
  - 使用Docker建構資料分析環境
    - Lab 1: 快速建立屬於自己的Docker資料分析環境
  - 了解Elasticsearch
    - Lab 2: 使用ElasticSearch熟悉資安日誌
  - 了解LogStash連結資料
    - Lab 3: 導入遠端日誌
  - 動手簡單的資料分析IRIS分類問題
    - Lab 4: Iris 分類演練
  - 動手玩玩Apache Access Log
    - Lab 5: Apache Access log 分群
  - 動手分析Google Drive Access Log
    - Lab 6: Google Drive Access Log分群與視覺化分析

Tutorial List

- 資安威脅簡介 (投影片)
  https://www.dropbox.com/sh/bzesgstziawoay4/AABB6iIY6v3OtNBAHje9zap9a?dl=0
- IoT資安威脅與分析 (投影片)
  https://www.dropbox.com/sh/bzesgstziawoay4/AABB6iIY6v3OtNBAHje9zap9a?dl=0
- 惡意程式行為分析 (投影片)
  https://www.dropbox.com/sh/bzesgstziawoay4/AABB6iIY6v3OtNBAHje9zap9a?dl=0
- Lab手冊 (投影片)

Introduction

- Virtual Machine提供一個pre-install的整合環境，提供系統隔絕，適合進行”惡意程式分析”的”動靜態行為分析”
  - Kali-Linux即為資安滲透測試常用工具 (https://www.kali.org/)
  - REMNUX則常用於Malware分析 (https://remnux.org/)
- Data Analytics是一個ecosystem，包含：ETL(Extraction Transform Loading)、Repository(SQL, NoSQL)、Computation(Hadoop, HDFS)等
- 要用哪一種語言整合？
- 本課程採用Python- Python被高度使用於Security與Data Analytics領域

Introduction

- 惡意行為分析可從:
  - Intrusion Detection System Log
  - Firewall Log
  - Host Activity Event
  - Process
  - Proxy Log
  - Web Access Log
  - Netflow
  - DNS log
  - Cloud log
- 資料分析在資安威脅行為分析中，如何扮演角色？
  - 收整資料、統計分析、行為塑模、威脅預測

這些Log可以做什麼？- IDS Logs

- Intrusion Detection System (IDS) 中文為入侵偵測系統，可配置於主機(Host IDS)或是網路閘道口(Network IDS)
- 對於已知攻擊的特徵，如網路行為(使用的port, protocol, 字串特徵, 特定傳送內容)，然而若不會調校，**虛警率非常高**
- 通常會有以下資訊：事件名稱,來源IP, 來源Port, 目標IP,目標Port,時間戳記
- 過去十多年，非常多研究聚焦於分析IDS的Log (not a novel direction)
  - 降低需警率 (False Alarm Reduction)
  - 關聯多步驟攻擊 (Multi-steps Attacks)
  - 找出潛在攻擊行為 (困難)
- 適合的資安防護體系：SOC (security operation center), CERT, ISAC...

這些Log可以做什麼？- Firewall Logs

- Firewall 防火牆，進行連線行為阻擋功能，針對已知連線黑名單進行阻擋，是資安防護非常倚重的角色
- 防火牆是非常好的執行者，然而缺乏分析能力，對於未知的網路特徵，如：IP，Domain Name等，缺乏分析能力
- 防火牆會產生以下資訊：來源IP、來源Port、目標IP、目標Port、動作(Permit/Deny)等
- Firewall相關資料分析也十多年，著重於大規模連線行為的關聯
  - 誘捕系統攻擊結構特徵分析
  - 骨幹網路大規模攻擊結構
  - 殭屍網路結構分析
  - 網路特徵分析

今天課程會有LAB

這些Log可以做什麼？Web Server 存取日誌

- Web Server存取日誌，隨著網頁應用服務發展，藉由Web Server存取日誌，可進行HTTP應用層的惡意行為偵測
- Web Server存取日誌通常有: 來源IP、存取路徑、存取狀態、存取時間，另外可額外紀錄存取大小、HTTP標題資訊等
- 通常log過多Web Server資訊，會影響效能，Web Server存取日誌僅能聚焦於應用層惡意行為(Script)
- Web Server日誌分析約從2005年開始
  - 透過多模型字串分析 URL的PATH資訊，找出潛在具有 XSS或是SQL Injection的攻擊行為
  - 藉由圖論方法建出存取關聯圖
  - 透過分類器判斷 request的風險程度

這些Log可以做什麼？Proxy日誌

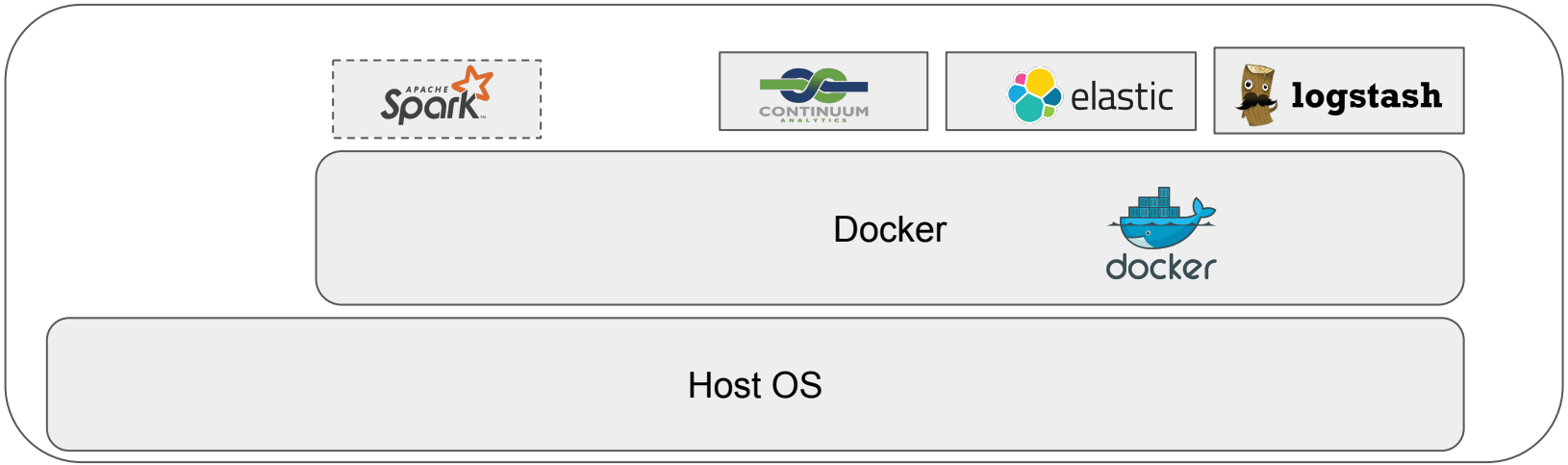
- Proxy日誌通常指的是網頁代理伺服器，佈建於企業網路出口，可記錄企業內電腦上網行為
- 上網行為對於追蹤惡意程式的重要性：
  - 惡意程式會透過URL連線方式，與外部C&C Server溝通
  - 可發現釣魚網站會誘導使用者連線到釣魚網頁
  - 可找出惡意掃描以及攻擊的行為
- Proxy包含：source_ip, 連線目標的URL, 時間戳記 (可能有時候會帶AP), 瀏覽器資訊等
- Proxy日誌分析從2010年後逐漸受到重視，除了看是否有上可疑網頁，異常連網行為也非常重要
  - 分析惡意程式在內部的行為
  - 關聯其他日誌如：AD、IDS、FW，以帳號風險評估的角度分析

這些Log可以做什麼？Active Directory日誌

- Active Directory是由作業系統在使用端與伺服端所產生的活動日誌，可用來分析是否有潛藏的惡意程式行為，包含：異常權限取得行為、異常活動行為
- Active Directory日誌廣泛用於電腦事件處理時，追查惡意行為，日誌數量非常大，值得近年來巨量資料分析技術成熟，被用於資安監控
- Active Directory包含：時間戳記、AD帳號、來源IP、事件名稱、子事件名稱等欄位
- 這兩三年研究略有提到，但資安大廠相繼投入Active Directory流量與日誌的研究

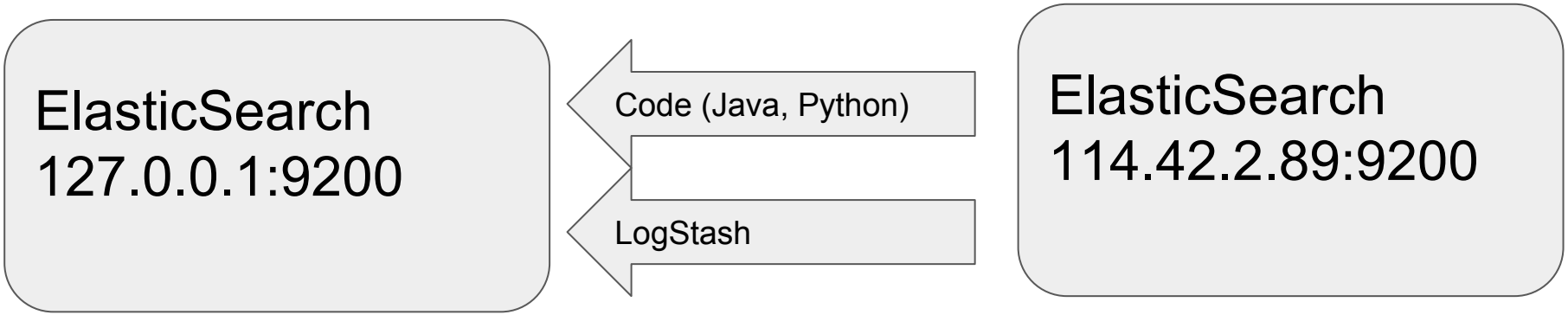
Analytics Docker in this Course

Remote Elasticsearch
(114.32.24.166:9200)



參考:
Docker ELK: https://github.com/deviantony/docker-elk
Docker Spark/Anaconda: https://github.com/LogBaseInc/spark-anaconda

Analytics Docker in this Course



Cloud Audit Log Dataset

- ElasticSearch Source
  - http://114.32.24.166:9200/_plugin/head/
  - IP: 114.32.24.166 Port: 9200
- Google Audit Log with two index

Install LogStash ElasticSearch on your Host

- LogStash: https://www.elastic.co/products/logstash
  - https://download.elastic.co/logstash/logstash/logstash-2.3.4.zip
- ElasticSearch:
  - https://download.elastic.co/elasticsearch/release/org/elasticsearch/distribution/zip/elasticsearch/2.3.5/elasticsearch-2.3.5.zip

Lab 1: 快速建立屬於自己的Docker資料分析環境

- 安裝Docker (either Docker 原生 or Docker Toolbox)
- 下載yml檔 (Docker的Compose file reference)
- 確認是否可以登入
  - 開發環境：iPython Notebook
  - 資料庫：Elasticsearch
  - 資料拋轉：Logstash
- 透過terminal確認images與containers

安裝Docker

- 安裝Docker
  - Windows: https://docs.docker.com/engine/installation/windows/
    - 下載路徑: https://download.docker.com/win/stable/InstallDocker.msi
  - Mac: https://docs.docker.com/engine/installation/mac/
    - 下載路徑: https://download.docker.com/mac/stable/Docker.dmg
- Docker Toolbox (如果原生Docker無法安裝，請安裝這個)
  - URL: https://www.docker.com/products/docker-toolbox
  - 下載路徑 :https://github.com/docker/toolbox/releases/download/v1.12.0/DockerToolbox-1.12.0.exe

下載YML檔 (方法1)

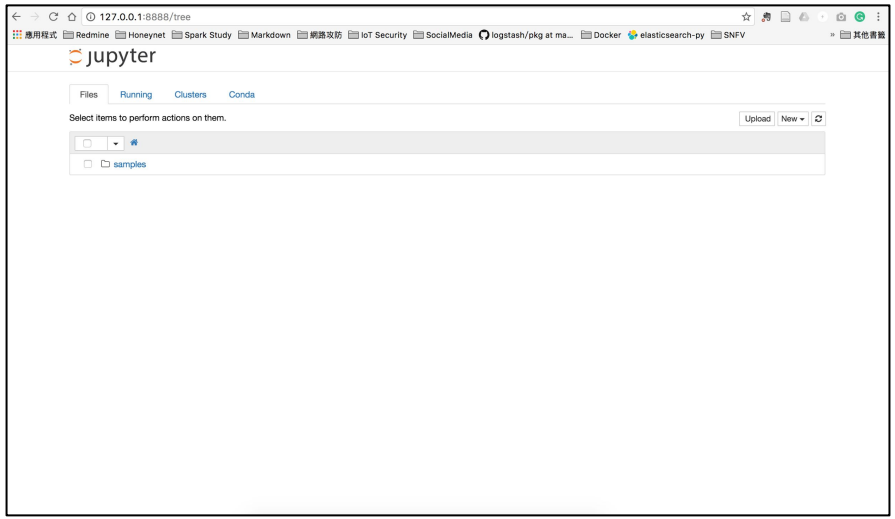
- 下載YML網址 :https://dl.dropboxusercontent.com/u/23229197/NTU_course2016/yml/jupyter%2Belasticsearch%2Blogstash.zip
- 解壓縮取得: docker-compose.yml
- 執行:docker-compose up
- Done

下載Docker VirtualBox for Docker

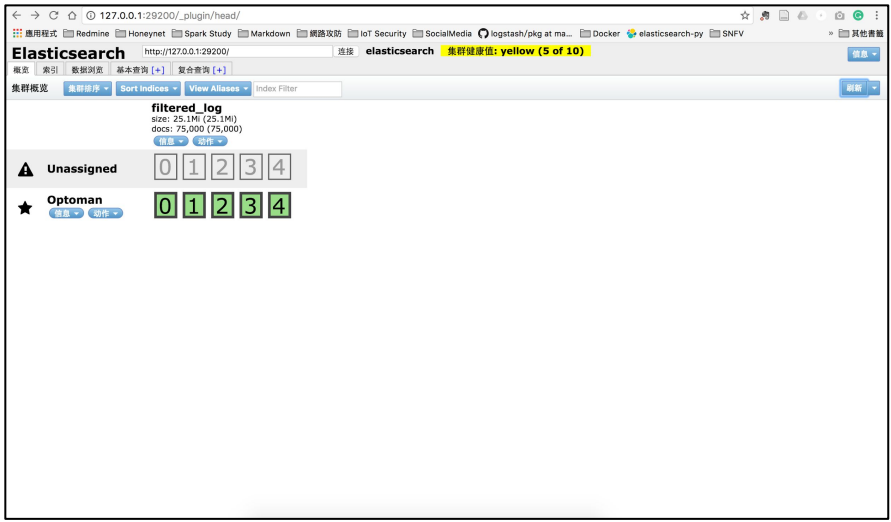
- 下載 VirtualBox VM for Docker：https://goo.gl/1z2uLX
- VM帳號 / 密碼：docker / docker
- docker compose 設定檔案： docker-compose.yml
- 啟用集群container指令：
- cd /home/docker/jupyter+elasticsearch+logstash/
- vim logstash.config
- docker-compose up
- 啟用後連線至：http://192.168.xxx.xxx:8888

確認能否正常登入

Anaconda iPython Notebook:
http://127.0.0.1:8888



Elasticsearch: http://127.0.0.1:29200
Head: http://127.0.0.1:29200/_plugin/head



透過Terminal確認Docker images與containers

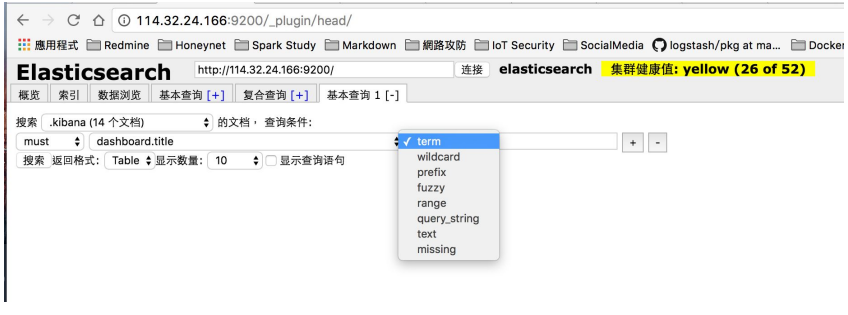
- Anaconda, ElasticSearch
  - 請下載YML檔:
    - https://drive.google.com/drive/folders/0B6CSO_BFk1BRUFVDbkdKa21zVU0?usp=sharing
    - docker-compose.yml
  - 下載後:
    - docker-compose up
  - 原生的Docker
    - ### To connect the website of Jupyter
    - http://localhost:8888/tree
    - ### To connect the head plugin of elasticsearch
    - http://localhost:29200/_plugin/head/
  - Docker Toolkit
    - localhost -> 192.168.99.100 (預設, 例外: 以docker拿到的IP為準)

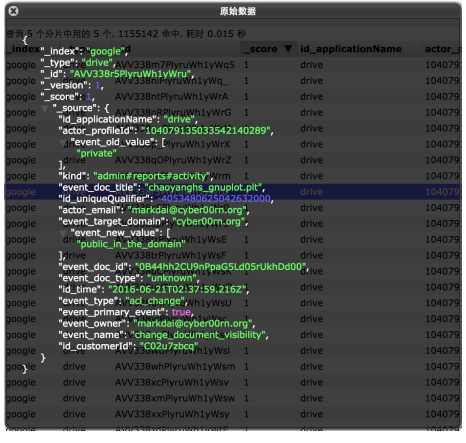
透過terminal確認images與containers

- 想查詢有哪些images
  - docker images
- 想查詢有哪些正在運作的containers
  - docker ps
- 想查詢所有的containers (停掉的containers也會列出)
  - docker ps -a
- 想刪掉Container
  - docker rm [CONTAONER_ID]
- 想刪掉image
  - docker rmi [IMAGE_ID]
- 刪掉image前，要先確認Container已經先刪掉
- 一次快刪
  - docker rm $(docker ps -a -q)
  - docker rmi $(docker images -q)

Lab 2: 使用Elasticsearch熟悉日誌

- 請參考此份講義：ElasticSearch安裝與操作 (https://bigdataanalytics2014.files.wordpress.com/2015/03/elasticsearch.pptx)
- 進入head管理介面 (http://114.32.24.166:9200/_plugin/head/)
  - 了解shard與replica set意義
  - 進入數據瀏覽，選擇 index -> google，選擇type -> drive (有_____筆資料)
  - 點選紀錄，了解資料結構 (如右圖)
  - 進入基本查詢，演練不同類型 查詢方式 (如下圖)





Lab 3: 導入遠端日誌

- 任務:將位於IP: 114.32.24.166 Port: 9200, index- filtered_log 導入位於container裡的Elasticsearch, 他的位置是IP: 127.0.0.1 Port: 29200
- 請參考LogStash講義- https://www.dropbox.com/s/a7fh7p9iaqeib6k/logstash.zip?dl=0
- 了解LogStash Conf檔撰寫
  - Elasticsearch -> Elasticsearch, 參考範例 :https://dl.dropboxusercontent.com/u/23229197/NTU_course2016/logstash/es-to-es.conf
  - Apache Log to Redis or Elasticsearch: http://www.wklken.me/posts/2015/05/08/elk-data-collect.html#logstash-shipper

Lab 3: 導入遠端日誌 (cont.)

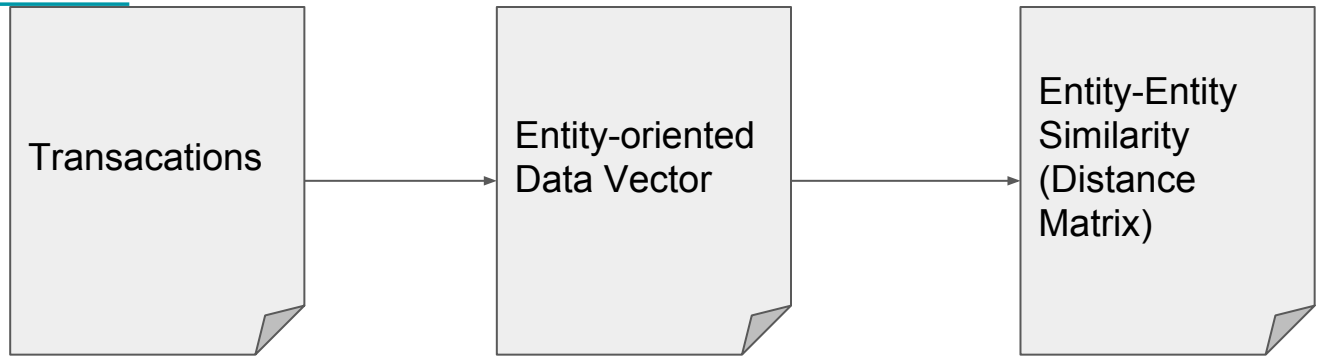
```
input {
  elasticsearch{
      hosts=>"114.32.24.166"
      index=>"filtered_log"
  }
}
output {
  elasticsearch{
      hosts=>"127.0.0.1:29200"
      index=>"filtered_log"
  }
  stdout {
    codec => rubydebug
  }
}
```

Lab 4: Iris分類演練

- 請到iPython Notebook, http://127.0.0.1:8888/tree/samples/example-data-science-notebook
  - 範例程式：Example Machine Learning Notebook.ipynb (點兩下)
  - 請看一下iris-data.csv的格式
  - 直接執行，了解分析的 code
- 請登入iPython notebook中，下載另外一隻iris-data.csv的資料分析程式
  - 先確認container 的 id, 例如: ${CONTAINER_ID}
  - docker exec -i -t ${CONTAINER_ID} /bash/bin
  - wget https://dl.dropboxusercontent.com/u/23229197/NTU_course2016/pyes_google/bokeh_iris_plot_practice.ipynb
  - 將下載後的bokeh_iris_plot_practice.ipynb放到/sample底下
  - 透過iPython Notebook開啟

Lab 5: Apache Log 分群

- 一般而言，資安日誌都是交易紀錄，缺乏”角度”
  - 何謂角度？
    - 想從來源主機的角度？目標主機的角度？事件觸發的角度？
    - 要不要考量時間？
    - 要不要考量分析個體與個體之間的群體關係？ (Social Network Analysis)
- 下載範例 :https://www.dropbox.com/sh/92uw0e7u11ouu9b/AAD-o8hRvoqXlUZxYlk3_Uu9a?dl=0



Lab 6: Google Drive Access Log分群與視覺化分析

- 範例程式:
  https://www.dropbox.com/sh/wwp49xxcgvifzk6/AAAgNvd-fnqkIcHya53nAAfDa?dl=0
- 任務
  - 將位於IP: 114.32.24.166 Port: 9200, index- Google 導入 container裡的ES
  - 演練Python如何取得es資料, 如何查詢資料
  - 演練如何將資料轉成矩陣
  - 將資料透過MDS以及K-means分群
  - 視覺化分析結果, 透過Bokeh