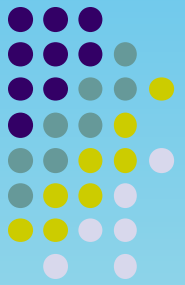
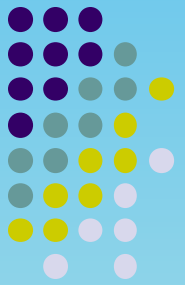


網路資料搜集器設計與應用實務

Pei Chen, BoShion Hwang, Sun Wu

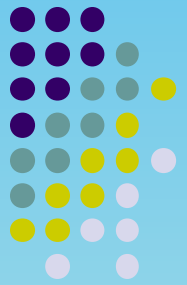
20170420



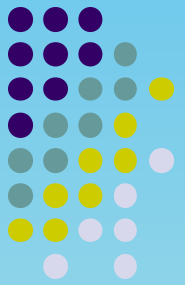
BigData | 巨量資料 | 大數據很熱

資料那裏來？

資料那裏來？



政府單位？
大企業？

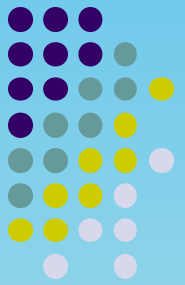


私有資料很難取得

那就抓公開的資料吧！

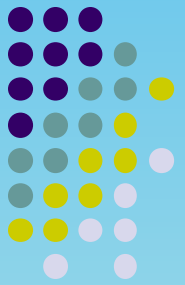
資料來源主要分

社群大站例如 Facebook
一般 WWW 網站



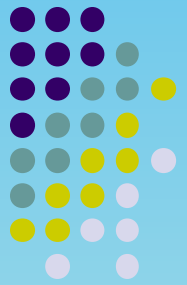
兩種 BOT

透過 API 例如 FB Graph API
透過 HTTP GET 抓網頁、透過網頁解析產生連結



兩種 BOT

透過 API 例如 FB Graph API
透過 HTTP GET 抓網頁、透過網頁解析產生連結



FB BOT

Graph API

粉絲專頁

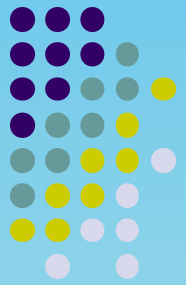
貼文

回應

FB BOT

蒐集粉專節點
蒐集粉專資訊
蒐集粉專貼文
蒐集貼文的回應

Ookon Cloud Server 1.0



FB BOT 結構

分散式架構

Crawler Master

Seed Master

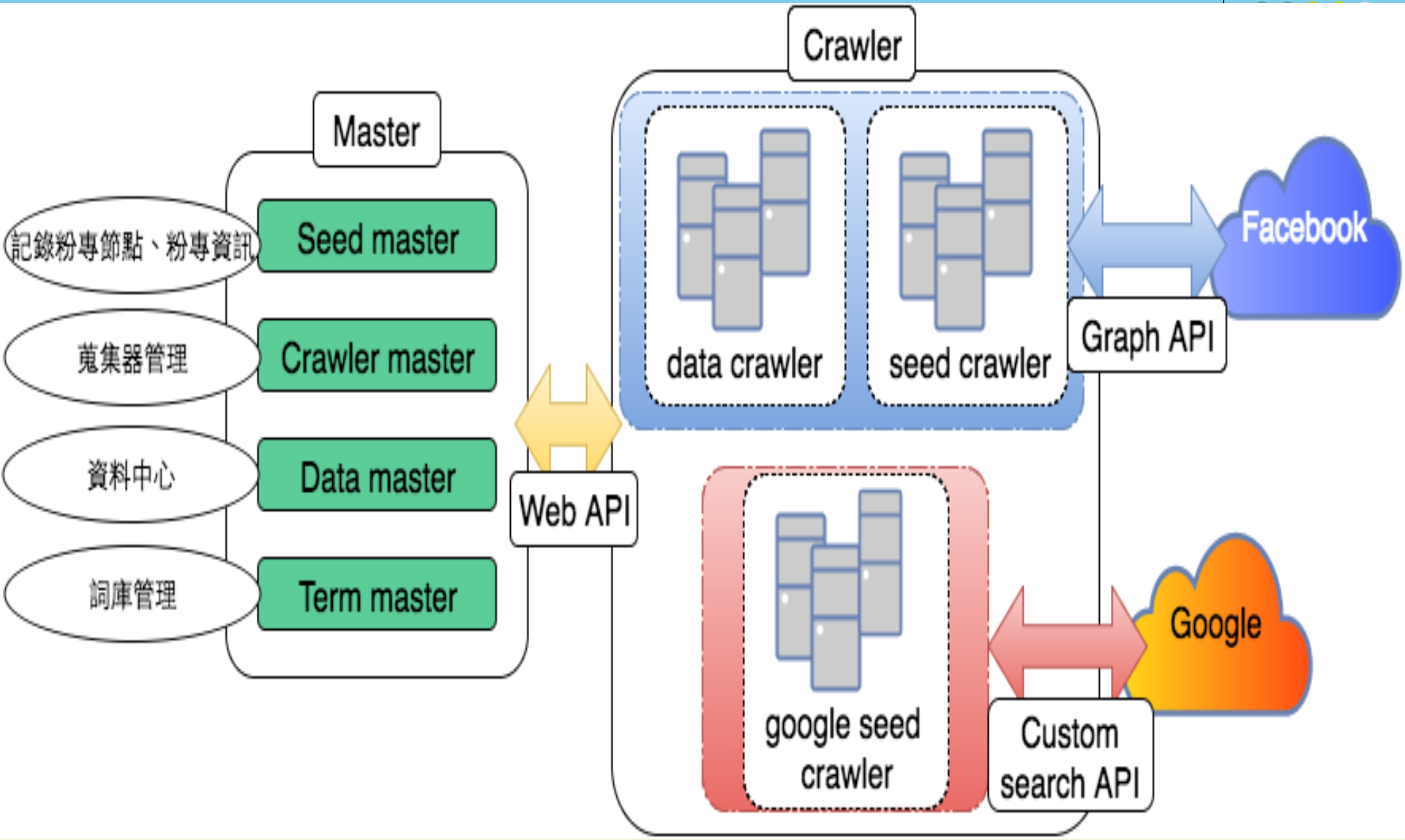
Data Master

Term Master

Seed Crawler

Data Crawler

```
{
  "data": [
    {
      "id": "109249609124014_1195124387203192",
      "link": "https://goo.gl/AfImwK",
      "message": "【世界公民島專欄／謝易軒】
曾是最高機密的地點竟然是某部電影的拍攝地！
#AlphaGo #英國 #台灣 #模仿遊戲",
      "description": "今年，AlphaGo因為打敗世界圍棋棋王而聲名大噪，但你可能不知道，在Google的背後，這個人工智慧軟體是...",
      "name": "英國經驗：如何製造出擊敗人類的人工智慧",
      "created_time": "2016-07-26T15:40:00+0000",
      "from": {
        "name": "Yahoo!奇摩新聞",
        "id": "109249609124014"
      }
    },
    {
      "id": "109249609124014_1195332033849094",
      "link": "https://goo.gl/Y8zewF",
      "message": "構想引發反彈，此案將暫緩上路
#北宜 #濱海公路 #國道5號 #大貨車",
      "description": "（中央社記者沈如峰宜蘭縣26日電）為紓解國道5號暑假塞車問題，宜蘭縣政府與公路總局日前考慮試辦北宜、濱海公路假日管制大貨車北上通行，但這項構想引發反彈 ...",
      "name": "濱海北宜假日禁行大貨車 決定喊卡",
      "created_time": "2016-07-26T15:34:58+0000",
```


Crawler
Master
記錄粉專節點、粉專資訊
Seed master
蒐集器管理
Crawler master
資料中心
Data master
詞庫管理
Term master
Web API
data crawler
seed crawler
Graph API
Facebook
google seed crawler
Custom search API
Google

FB 抓取實務：

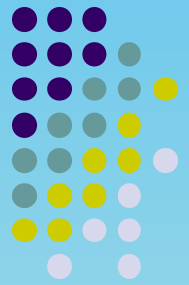
台灣：約 100 萬粉專，接近 4 億貼文
全球：目前蒐集四千多萬粉專節點
已經蒐集一百多億貼文
…..

FB 抓取實務：

透過 Hinet 家用頻寬 如 100M/40M
透過 Google Cloud Platform
共約 20 台機器。
每天可抓取十億 + 個貼文

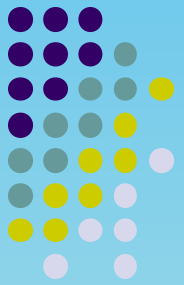
FB 抓取實務：

Data Center 設於一般主機
透過 Google Cloud Platform
抓取後，壓縮後分時段傳回
四百億個貼文，壓縮完後大約 10T
(未壓縮前約 50T)



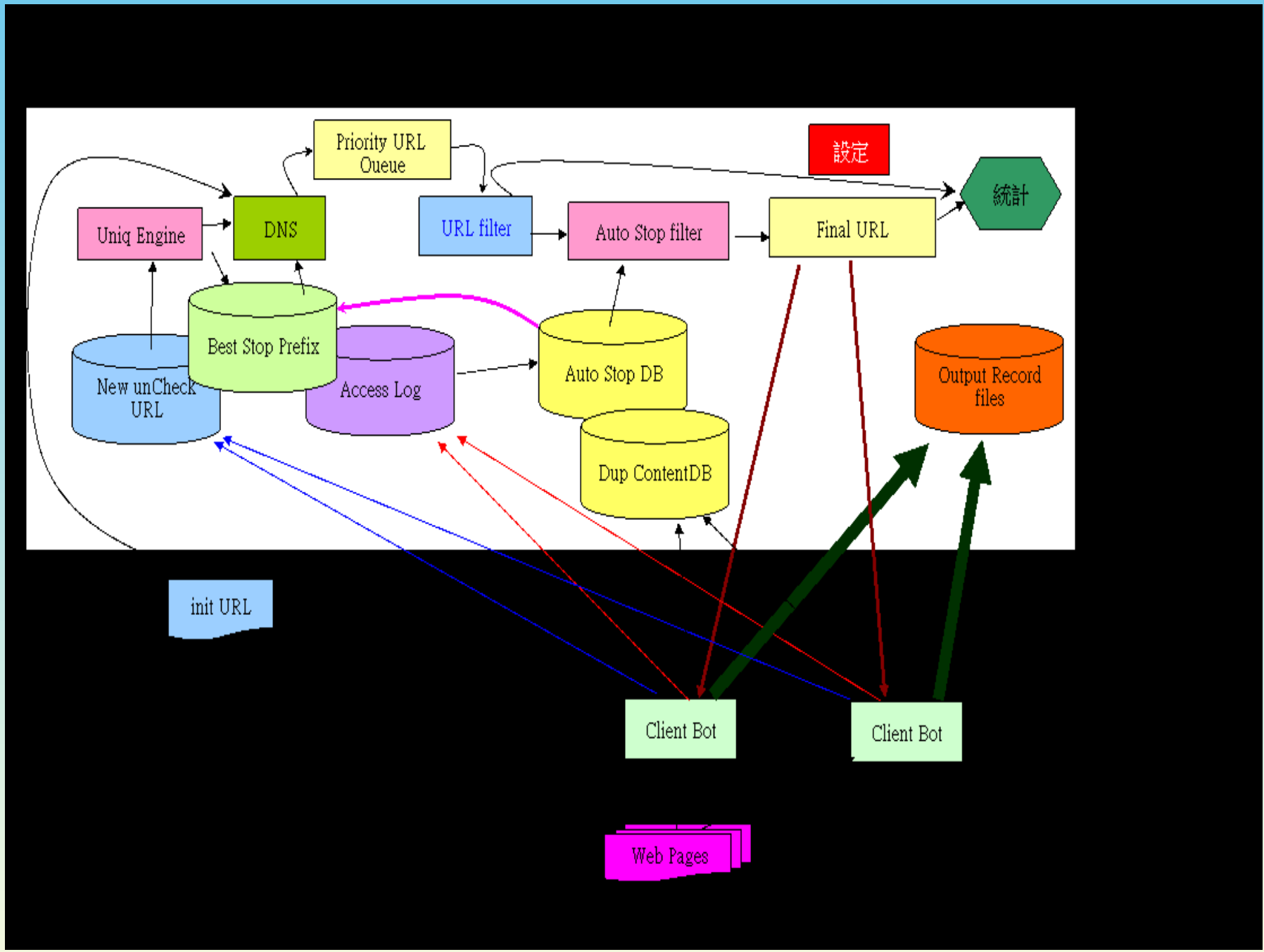
WEB BOT

Crawler Master
Data Master
Crawler Client

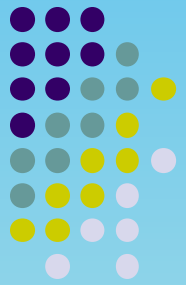


issues

如何抓齊全，
如何高效率抓取，
那些連結要抓，那些不要抓
如何避免重複抓取？
如何過濾 spams ？
如何抓特定主題？
如何更新抓取？


Priority URL Oueue
設定
統計
Uniq Engine
DNS
URL filter
Auto Stop filter
Final URL
Best Stop Prefix
New unCheck URL
Access Log
Auto Stop DB
Output Record files
Dup ContentDB
init URL
Client Bot
Client Bot
Web Pages

抓取實務
一台機器一天可抓取約 1000 萬
抓取台灣大 約 5 億網頁
約兩千萬筆新聞類網頁



抓取實務

每天抓取
新聞類資料（透過 RSS)
社群新訊 例如 PTT, FB, Dcard

應用

熱門新聞、焦點訊息
搜尋引擎
資訊擷取
輿情分析
資料採礦
趨勢分析
資安資料分析
….

應用

熱門新聞、焦點訊息
搜尋引擎
資訊擷取
輿情分析
資料採礦
趨勢分析
資安資料分析