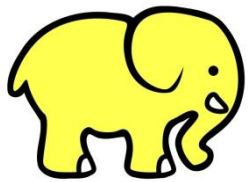
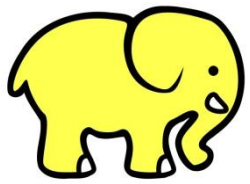
金象盃
全國大數據實務能力競賽



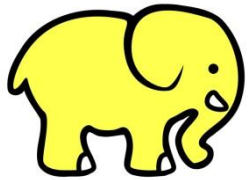
目錄

1. 競賽目的
2. 競賽資料集說明
3. 大數據平台的關鍵技術 (Hadoop)
4. 不會 Java 也能輕鬆處理大數據 (Pig)
5. 使用 SQL 語法分析大數據 (Hive)

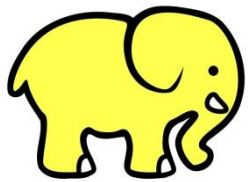


競賽目的

從 2014 年開始，企業逐漸由**資訊科技** (Information Technology) 邁向**資料科技** (Data Technology)，Hadoop 等技術發光發熱，大數據的價值蔚為風潮，資料分析師、資料工程師、資料科學家等職缺更是說明了資料科技的時代已經來臨。而2018年的現在，Hadoop 也已經是企業環境內不可或缺的關鍵要角之一，不但與傳統資訊技術 (BI、關聯式資料庫) 相輔相成，還能為物聯網、機器學習、人工智慧等技術提供莫大的助力。以 Hadoop 為基礎的大數據平台上，各種分析工具與日俱增，例如，Spark、Flink、HBase、Hive、Pig 等等，目前在國內業界又以「了解 SQL 就能上手的 Hive 」最為熱門。



競賽資料集說明



競賽資料來源 (一)

https://data.gov.tw/

競賽資料來源 (二)

https://segis.moi.gov.tw/

競賽資料集說明

競賽資料集清單	
資料中文名稱	原始資料來源
替代役役男訓練人數統計表	https://data.gov.tw/dataset/20445
各縣(市)警察(分)局暨所屬分駐(派出)所地址資料	https://data.gov.tw/dataset/5958
107年國民小學名錄	https://data.gov.tw/dataset/6087
107年幼兒園名錄	https://data.gov.tw/dataset/6086
大專校院各校科系別概況	https://data.gov.tw/dataset/9621
「ATM位置」查詢一覽表	https://data.gov.tw/dataset/24333
農產品交易行情	https://data.gov.tw/dataset/8066
A1類交通事故資料	https://data.gov.tw/dataset/12197
A2類交通事故資料	https://data.gov.tw/dataset/12197

競賽資料集說明

競賽資料集清單	
資料中文名稱	原始資料來源
內政部行事曆	https://data.gov.tw/dataset/26557
十六縣持卡人前十大國外消費金額及筆數(依簽帳筆數排名)	https://data.gov.tw/dataset/63029
十六縣居民跨縣市消費樣態	https://data.gov.tw/dataset/38315
連線紀錄	由成功大學與國網中心提供
攻擊者下載檔案紀錄	由成功大學與國網中心提供
107年6月行政區分齡兒童及少年性別人口統計_縣市	社會經濟資料服務平台，網址如下
https://segis.moi.gov.tw/STAT/Web/Platform/QueryInterface/STAT_QueryProductView.aspx?pid=BACF26E3C3F0E0CF73484A4AE2A91AF2&spid=7ED8D58E129BC680	

替代役役男訓練人數統計表

原始資料來源	https://data.gov.tw/dataset/20445			
資料中文名稱	替代役役男訓練人數統計表			
資料檔案名稱	20445.csv			
HDFS完整路徑	/dataset/20445/20445.csv			
資料總行數	200			
資料大小	3 KB			
資料說明	名稱為 20445.csv 的檔案，是93年到106年的替代役役男訓練人數統計表			
資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
year	年度	int	int	
phase	梯次	chararray	string	所有梯次均不重複；106_2，代表106梯次有第二梯次
trainees	替代役役男訓練人數	int	int	

各縣(市)警察(分)局暨所屬分駐(派出)所地址資料

原始資料來源	https://data.gov.tw/dataset/5958			
資料中文名稱	各縣(市)警察(分)局暨所屬分駐(派出)所地址資料			
資料檔案名稱	5958.csv			
HDFS完整路徑	/dataset/5958/5958.csv			
資料總行數	1695			
資料大小	96 KB			
資料說明	名稱為 5958.csv 的檔案，是各縣(市)警察(分)局暨所屬分駐(派出)所地址資料			
資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
name	名稱	chararray	string	
zip_code	郵遞區號	chararray	string	
address	地址	chararray	string	

107年國民小學名錄

原始資料來源	https://data.gov.tw/dataset/6087			
資料中文名稱	107年國民小學名錄			
資料檔案名稱	6087.csv			
HDFS完整路徑	/dataset/6087/6087.csv			
資料總行數	2633			
資料大小	301 KB			
資料說明	名稱為 6087.csv 的檔案，是107年的國民小學名錄			
資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
code	代碼	chararray	string	
school	學校名稱	chararray	string	
city	縣市名稱	chararray	string	
address	地址	chararray	string	
phone	電話	chararray	string	
url	網址	chararray	string	

107年幼兒園名錄

原始資料來源	https://data.gov.tw/dataset/6086
資料中文名稱	107年幼兒園名錄
資料檔案名稱	6086.csv
HDFS完整路徑	/dataset/6086/6086.csv
資料總行數	6720
資料大小	830 KB
資料說明	名稱為 6086.csv 的檔案，是106年的幼兒園名錄

資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
code	代碼	chararray	string	
school	學校名稱	chararray	string	
city	縣市名稱	chararray	string	
address	地址	chararray	string	
phone	電話	chararray	string	

大專校院各校科系別概況

原始資料來源	https://data.gov.tw/dataset/9621			
資料中文名稱	大專校院各校科系別概況			
資料檔案名稱	9621.csv			
HDFS完整路徑	/dataset/9621/9621.csv			
資料總行數	10081			
資料大小	676 KB			
資料說明	名稱為 9621.csv 的檔案，是大專校院各校科系別概況			
資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
city	縣市名稱	chararray	string	
school	學校名稱	chararray	string	
faculty	科系名稱	chararray	string	
courses	日間/進修別	chararray	string	
level	等級別	chararray	string	例如，學士、碩士、博士等
students	學生數	int	int	
teachers	教師數	int	int	

「ATM位置」查詢一覽表

原始資料來源	https://data.gov.tw/dataset/24333
資料中文名稱	「ATM位置」查詢一覽表
資料檔案名稱	24333.csv
HDFS完整路徑	/dataset/24333/24333.csv
資料總行數	22584
資料大小	2.1 MB
資料說明	名稱為 24333.csv 的檔案，是「ATM位置」查詢一覽表

資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
code	裝設金融機構代號	chararray	string	
name	裝設金融機構名稱	chararray	string	
location	裝設地點	chararray	string	例如，世貿中心展覽大樓
city	裝設縣市	chararray	string	例如，台北市
address	裝設地址	chararray	string	例如，台北市信義區信義路五段5號1樓

農產品交易行情

原始資料來源	https://data.gov.tw/dataset/8066			
資料中文名稱	農產品交易行情			
資料檔案名稱	8066.csv			
HDFS完整路徑	/dataset/8066/8066.csv			
資料總行數	1863			
資料大小	116 KB			
資料說明	名稱為 8066.csv 的檔案，是107年9月16日的農產品交易行情			
資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
date	交易日期	chararray	string	
crop_code	作物代號	chararray	string	
crop_name	作物名稱	chararray	string	
market_code	市場代號	chararray	string	
market_name	市場名稱	chararray	string	
upper_price	上價	float	float	
middle_price	中價	float	float	
lower_price	下價	float	float	
average_price	平均價	float	float	
trading_volume	交易量	int	int	

A1類交通事故資料

原始資料來源	https://data.gov.tw/dataset/12197			
資料中文名稱	A1類交通事故資料			
資料檔案名稱	12197_A1.csv			
HDFS完整路徑	/dataset/12197_A1/12197_A1.csv			
資料總行數	1285			
資料大小	89 KB			
資料說明	名稱為 12197_A1.csv 的檔案，是106年的A1類交通事故資料 A1類說明：造成人員當場或24小時內死亡之交通事故			
資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
time	發生時間	chararray	string	
injured	受傷人數	chararray	string	
dead	死亡人數	chararray	string	
vehicle	車種	chararray	string	

A2類交通事故資料

原始資料來源	https://data.gov.tw/dataset/12197
資料中文名稱	A2類交通事故資料
資料檔案名稱	12197_A2.csv
HDFS完整路徑	/dataset/12197_A2/12197_A2.csv
資料總行數	281634
資料大小	19 MB
資料說明	名稱為 12197_A2.csv 的檔案，是106年的A2類交通事故資料 A2類：造成人員受傷或超過24時死亡之交通事故

資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
time	發生時間	chararray	string	
injured	受傷人數	chararray	string	
dead	死亡人數	chararray	string	
vehicle	車種	chararray	string	

內政部行事曆

原始資料來源	https://data.gov.tw/dataset/26557			
資料中文名稱	內政部行事曆			
資料檔案名稱	26557.csv			
HDFS完整路徑	/dataset/26557/26557.csv			
資料總行數	730			
資料大小	17 KB			
資料說明	名稱為 26557.csv 的檔案，是內政部行事曆資料			
資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
day	日期	int	int	
day_of_week	星期	chararray	string	
is_holiday	是否放假	int	int	
holiday_desc	放假說明	chararray	string	
year	年分	int	int	

十六縣持卡人前十大國外消費金額及筆數(依簽帳筆數排名)

原始資料來源	https://data.gov.tw/dataset/63029
資料中文名稱	十六縣持卡人前十大國外消費金額及筆數(依簽帳筆數排名)
資料檔案名稱	63029.csv
HDFS完整路徑	/dataset/63029/63029.csv
資料總行數	
資料大小	
資料說明	名稱為 63029.csv 的檔案，是十六縣持卡人前十大國外消費金額及筆數(依簽帳筆數排名)

資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
year_month	年月	chararray	string	
country	國別	chararray	string	
tw_city	城市	chararray	string	
amt	金額	int	int	
count	筆數	int	int	

十六縣居民跨縣市消費樣態

原始資料來源	https://data.gov.tw/dataset/38315			
資料中文名稱	十六縣居民跨縣市消費樣態			
資料檔案名稱	38315.csv			
HDFS完整路徑	/dataset/38315/38315.csv			
資料總行數				
資料大小				
資料說明	名稱為 38315.csv 的檔案，是十六縣居民跨縣市消費樣態			
資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
year	年月	chararray	string	
area	地區	chararray	string	
type	類別	chararray	string	
cards	卡數	int	int	卡數係指持卡人帳單地址為該地區且於該期間內之有效總卡數
total_trans	總交易筆數	int	int	
total_amt	總交易金額[新台幣]	biginteger	bigint	總交易金額係合計前開有效卡數於該期間內全部交易金額
inter_city_trans	跨縣市交易筆數	int	int	「跨縣市」係指持卡人帳單地址郵遞區號與交易發生之特約商店郵遞區號位於不同縣市
inter_city_amt	跨縣市交易金額[新台幣]	biginteger	bigint	

連線紀錄

原始資料來源	由成功大學與國網中心提供，原始名稱為 connections			
資料中文名稱	連線紀錄			
資料檔案名稱	999001.tsv			
HDFS完整路徑	/dataset/999001/999001.tsv			
資料總行數	91961			
資料大小	6.82 MB			
資料說明	名稱為 999001.tsv 的檔案，是連線紀錄			
資料欄位說明 (每個欄位之間均已Tab鍵分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
connection	連線編號	int	int	
type	連線形式	chararray	string	
protocol	通訊協定	chararray	string	
service	服務程式	chararray	string	
timestamp	時間戳記	int	int	
local_host	本機位址	chararray	string	
local_port	本機通訊埠	int	int	
remote_host	遠端位址	chararray	string	
remote_port	遠端通訊埠	int	int	

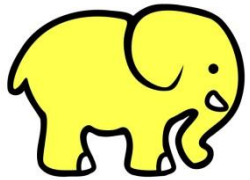
攻擊者下載檔案紀錄

原始資料來源	由成功大學與國網中心提供，原始名稱為 downloads			
資料中文名稱	攻擊者下載檔案紀錄			
資料檔案名稱	999002.tsv			
HDFS完整路徑	/dataset/999002/999002.tsv			
資料總行數	3530			
資料大小	272 KB			
資料說明	名稱為 999002.tsv 的檔案，是攻擊者下載檔案紀錄			
資料欄位說明 (每個欄位之間均已Tab鍵分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
download	下載記錄編號	int	int	
connection	連線編號	int	int	
download_url	下載網址	chararray	string	
download_md5_hash	下載檔案雜湊值	chararray	string	

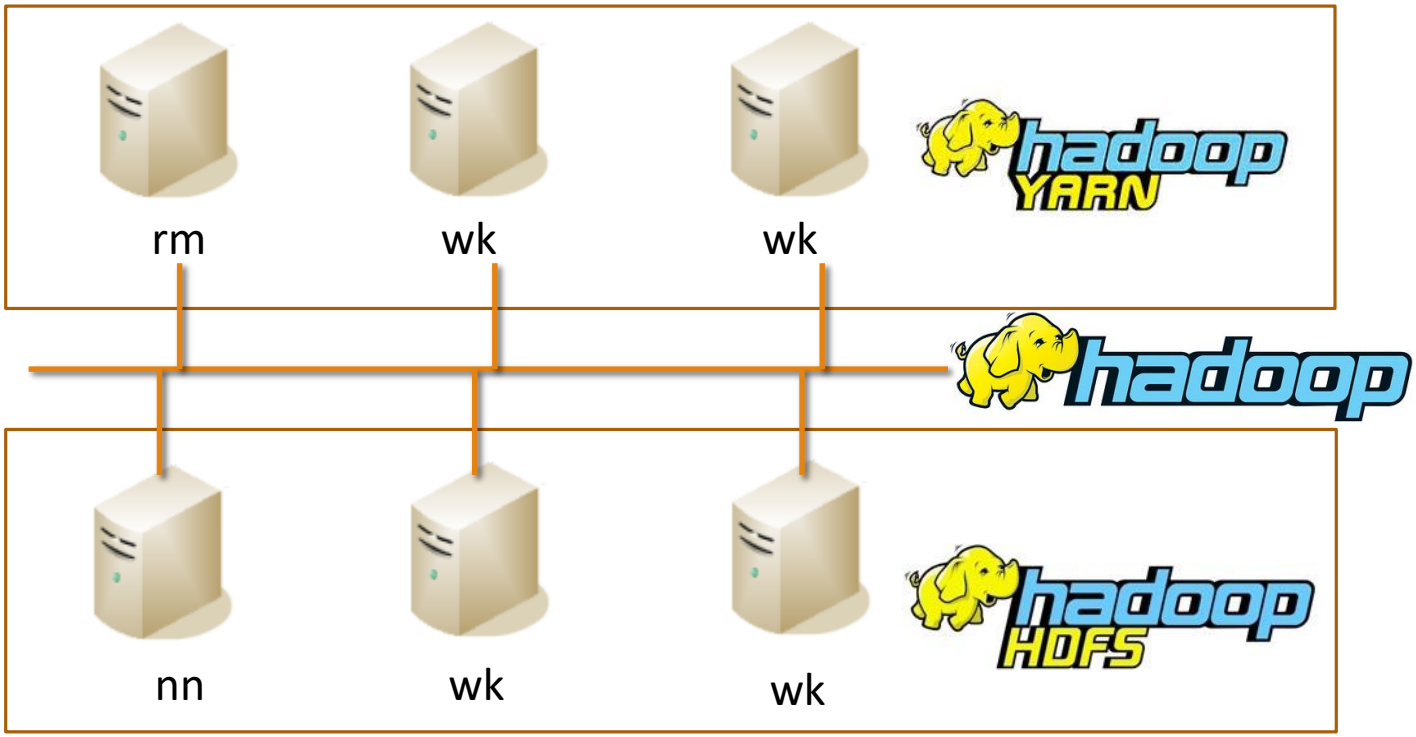
107年6月行政區分齡兒童及少年性別人口統計_縣市

原始資料來源	社會經濟資料服務平台			
資料中文名稱	107年6月行政區分齡兒童及少年性別人口統計_縣市			
資料檔案名稱	999003.csv			
HDFS完整路徑	/dataset/999003/999003.csv			
資料總行數	24			
資料大小	1 KB			
資料說明	名稱為 999003.csv 的檔案，是107年6月行政區分齡兒童及少年性別人口統計_縣市			
資料欄位說明 (每個欄位之間均已逗號分隔)				
欄位名稱 (英文)	欄位名稱 (中文)	資料型態 (Pig)	資料型態 (Hive)	備註
country_id	縣市代碼	int	int	
country	縣市名稱	chararray	string	
a0a5_cnt	0-5歲 兒童人口數	int	int	
a6a11_cnt	6-11歲 兒童人口數	int	int	
a12a17_cnt	12-17歲 少年人口數	int	int	

大數據平台的關鍵技術
Apache Hadoop

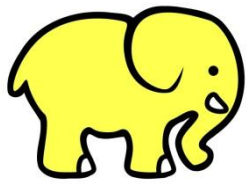


大數據平台的關鍵技術
Apache Hadoop

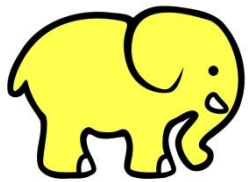


競賽答題系統

http://gec2018trn.azurewebsites.net/

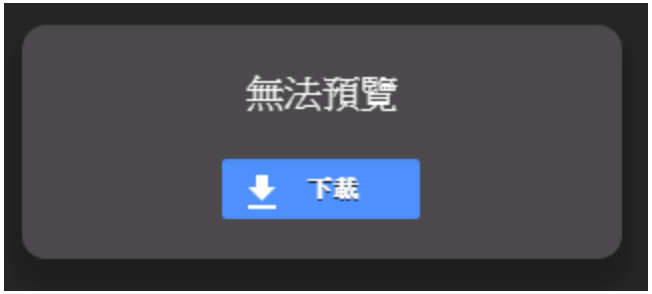


競賽環境操作
(VPN)

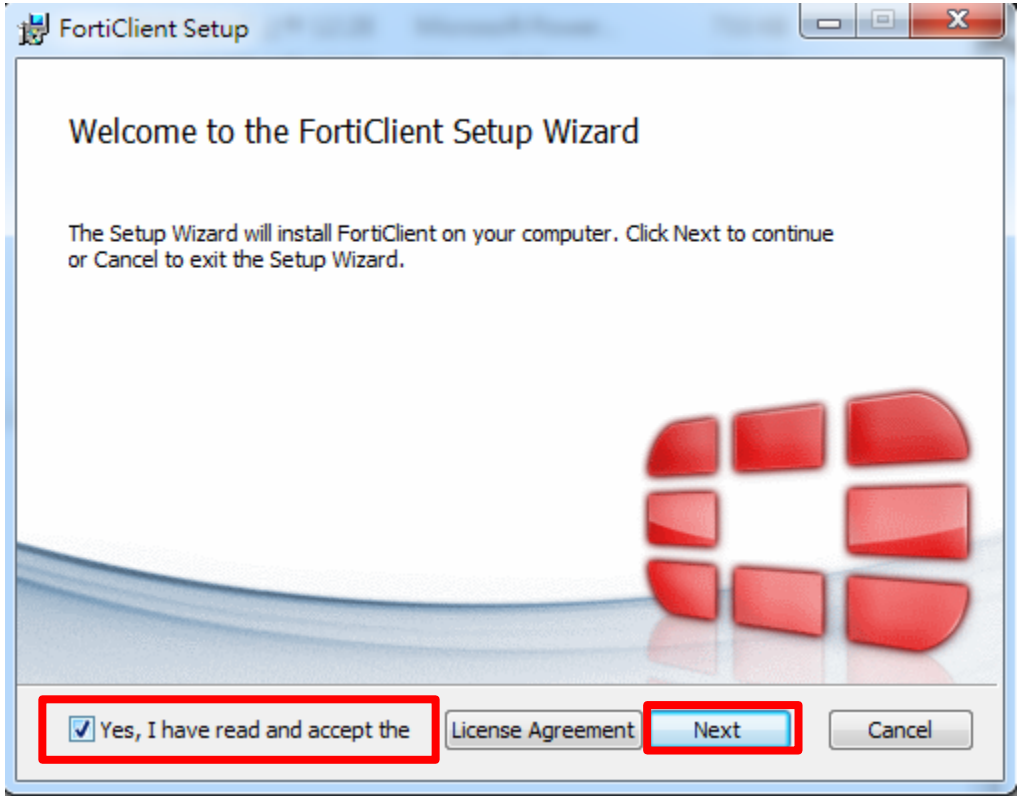


VPN 下載說明

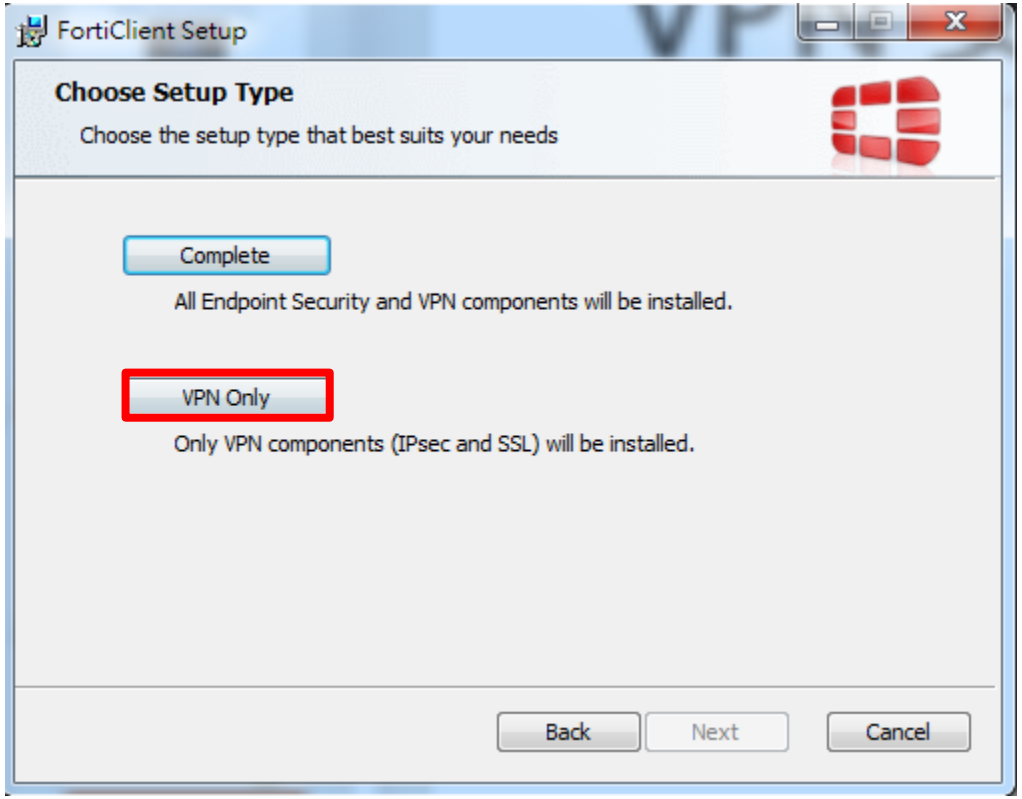
https://goo.gl/FSDv2n



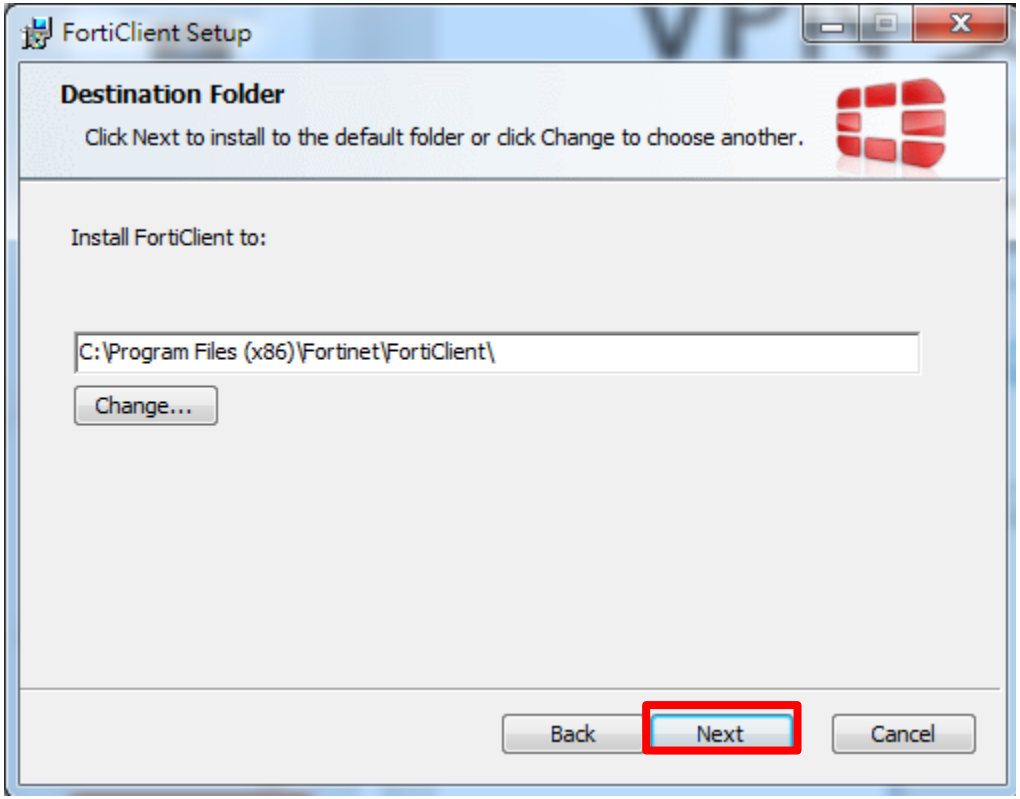
VPN 安裝 (一)


FortiClient Setup
Welcome to the FortiClient Setup Wizard
The Setup Wizard will install FortiClient on your computer. Click Next to continue or Cancel to exit the Setup Wizard.
Yes, I have read and accept the
License Agreement
Next
Cancel

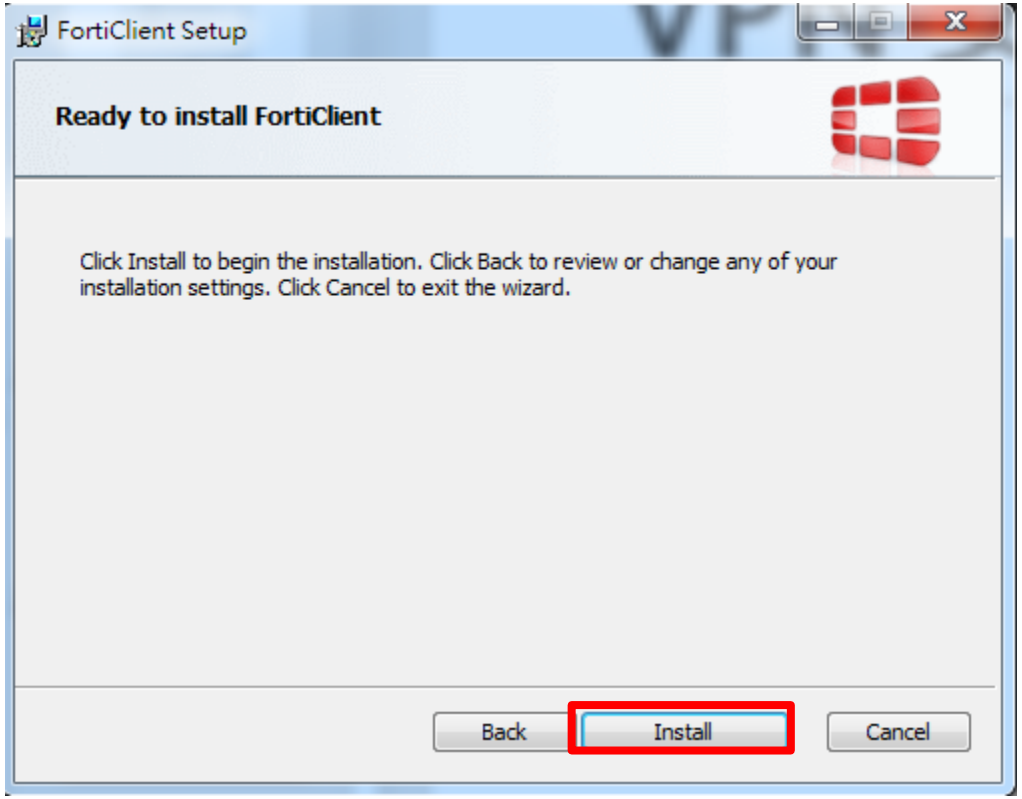
VPN 安裝 (二)


FortiClient Setup
Choose Setup Type
Choose the setup type that best suits your needs
Complete
All Endpoint Security and VPN components will be installed.
VPN Only
Only VPN components (IPsec and SSL) will be installed.
Back
Next
Cancel

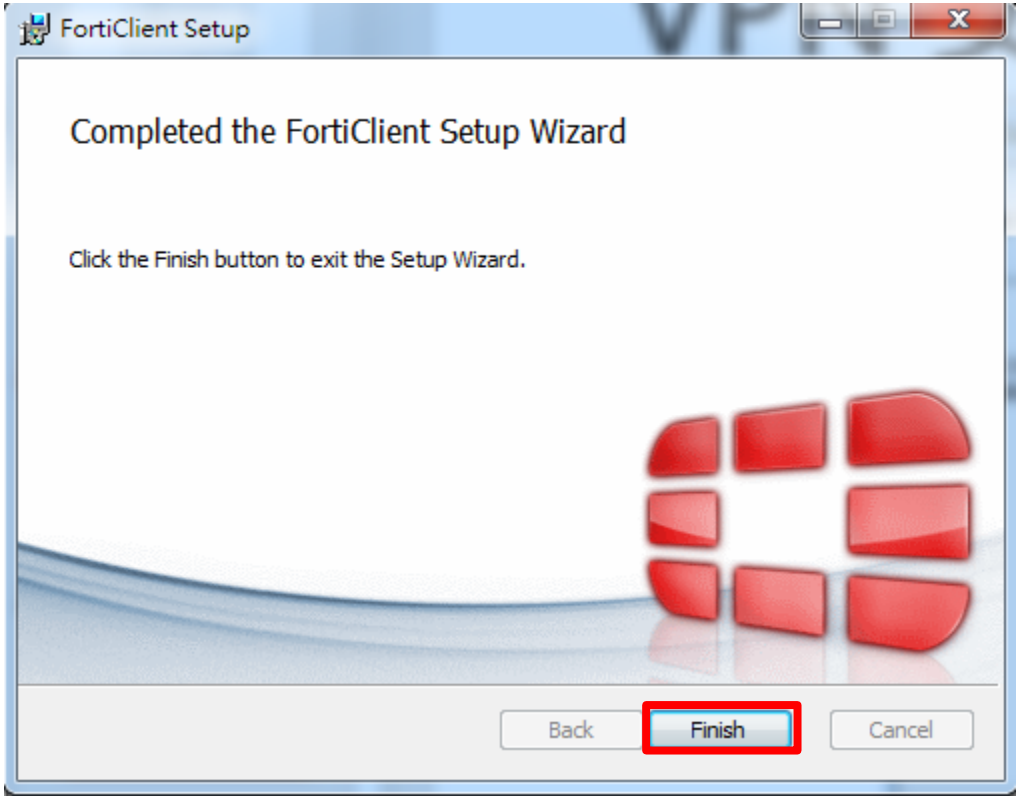
VPN 安裝 (三)


FortiClient Setup
Destination Folder
Click Next to install to the default folder or click Change to choose another.
Install FortiClient to:
C:\Program Files (x86)\Fortinet\FortiClient\
Change...
Back
Next
Cancel

VPN 安裝 (四)


FortiClient Setup
Ready to install FortiClient
Click Install to begin the installation. Click Back to review or change any of your installation settings. Click Cancel to exit the wizard.
Back
Install
Cancel

VPN 安裝 (五)


FortiClient Setup
Completed the FortiClient Setup Wizard
Click the Finish button to exit the Setup Wizard.
Back
Finish
Cancel

VPN 連線說明 (一)


FortiClient
File
Help
FortiClient Console
合規
不參加
遠程訪問
沒有VPN連接
配置 VPN
Security Features Not Installed
Get Protected

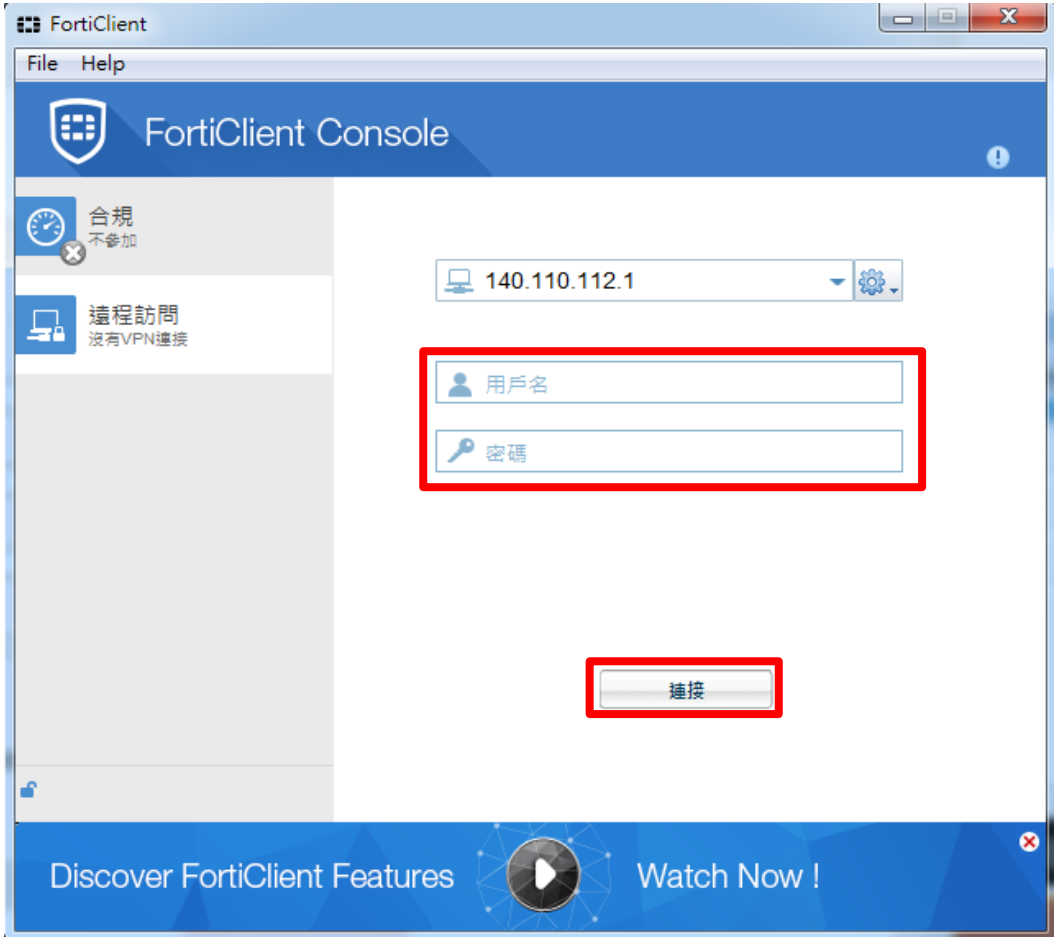
VPN 連線說明 (二)


FortiClient
File Help
新建 VPN
新 VPN 連接
SSL-VPN
IPsec VPN
連接名
140.110.112.1
描述
140.110.112.1
遠程網關
140.110.112.1
自定義端口
443
認証
登錄時提示
保存登錄名
客戶端証書
遇到無效的伺服器証書不提示
應用
關閉
Security Features Not Installed
Get Protected

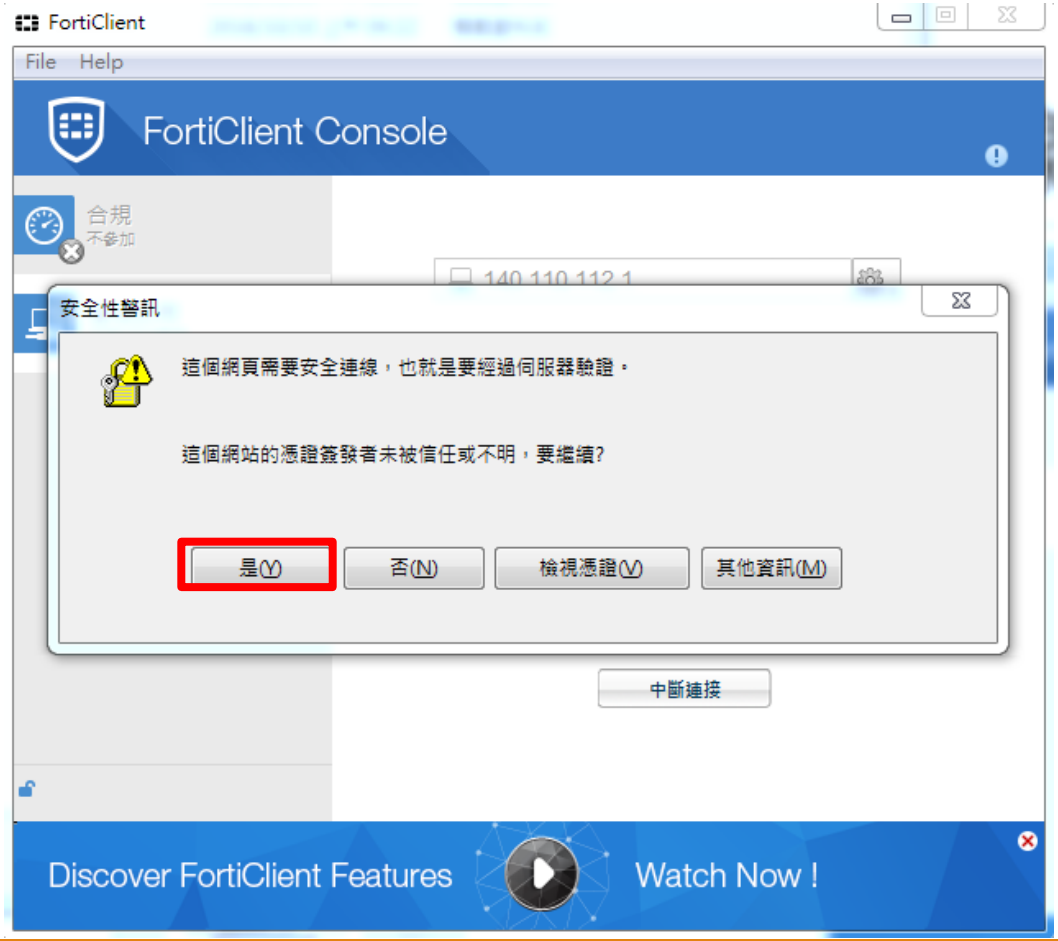
VPN 連線說明 (三)


FortiClient
File Help
140.110.112.1
SSL-VPN
編輯 VPN 連接
SSL-VPN
IPsec VPN
連接名
140.110.112.1
描述
140.110.112.1
遠程網關
140.110.112.1
自定義端口
443
認証
登錄時提示
保存登錄名
客戶端証書
遇到無效的伺服器証書不提示
應用
關閉
Security Features Not Installed
Get Protected

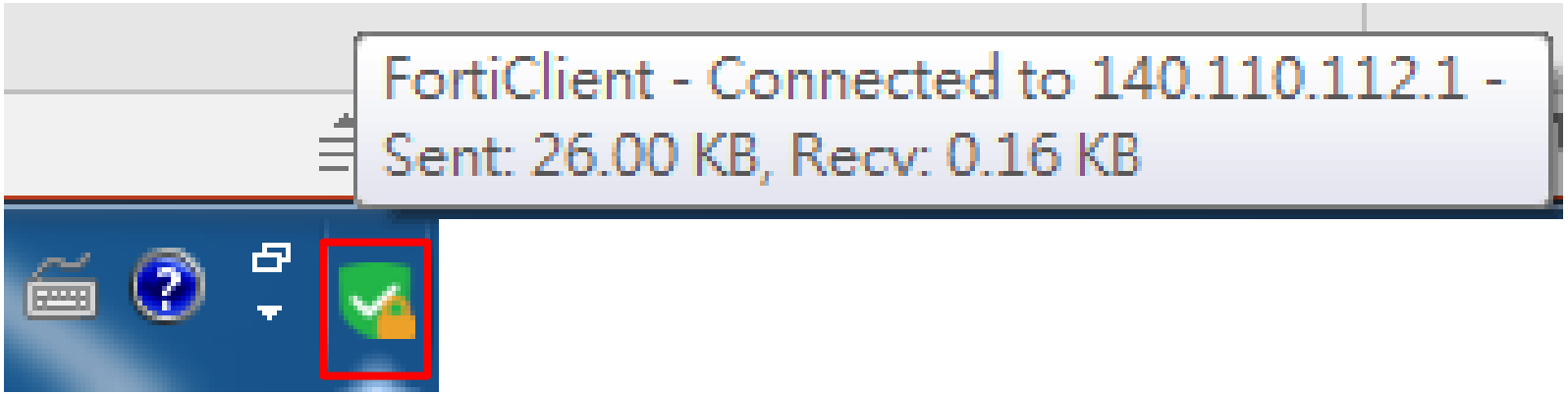
VPN 連線說明 (四)


FortiClient
File Help
FortiClient Console
合規
不參加
遠程訪問
沒有VPN連接
140.110.112.1
用戶名
密碼
連接
Discover FortiClient Features
Watch Now !

VPN 連線說明 (五)

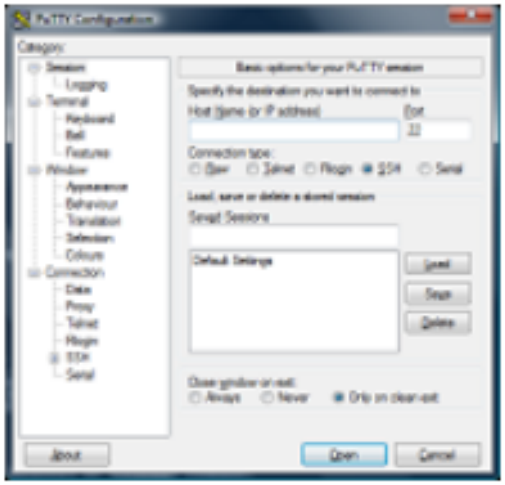

FortiClient
File Help
FortiClient Console
合規
不參加
安全性警訊
這個網頁需要安全連線，也就是要經過伺服器驗證。
這個網站的憑證簽發者未被信任或不明，要繼續?
是(Y)
否(N)
檢視憑證(V)
其他資訊(M)
中斷連接
Discover FortiClient Features
Watch Now !

VPN 連線說明 (六)


FortiClient - Connected to 140.110.112.1 -
Sent: 26.00 KB, Recv: 0.16 KB

Putty 下載說明

https://www.putty.org/

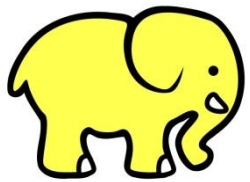


Download PuTTY

PuTTY is an SSH and telnet client, de
with source code and is developed ar

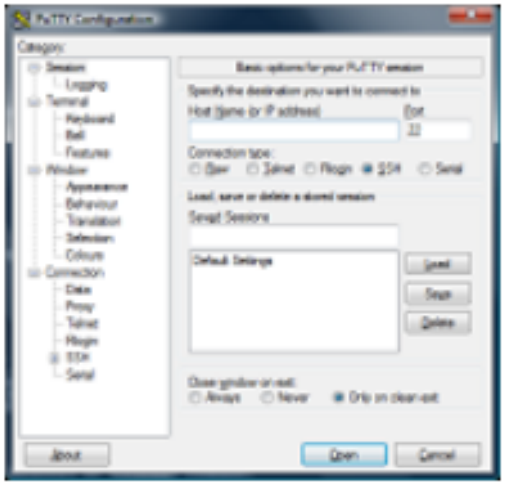
You can download PuTTY here.

競賽環境操作
(Putty)



Putty 下載說明

https://www.putty.org/

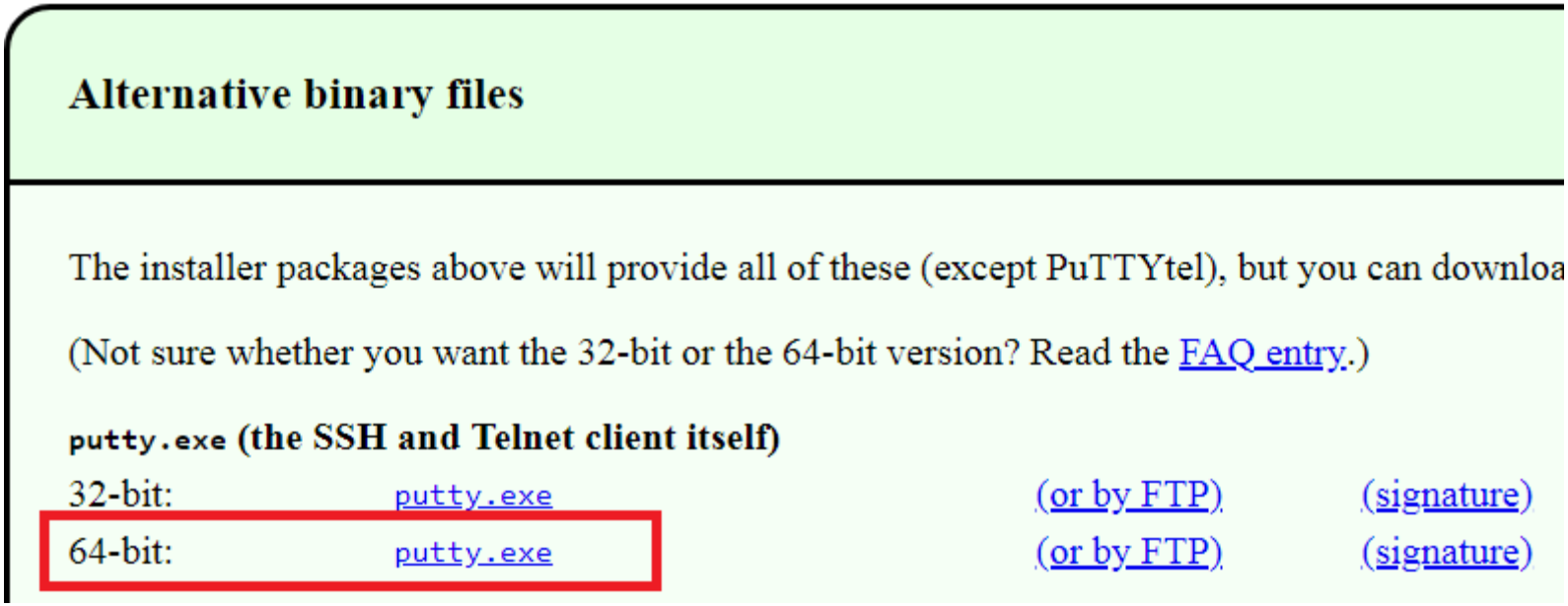


Download PuTTY

PuTTY is an SSH and telnet client, de
with source code and is developed ar

You can download PuTTY here.

Putty 下載說明



Alternative binary files

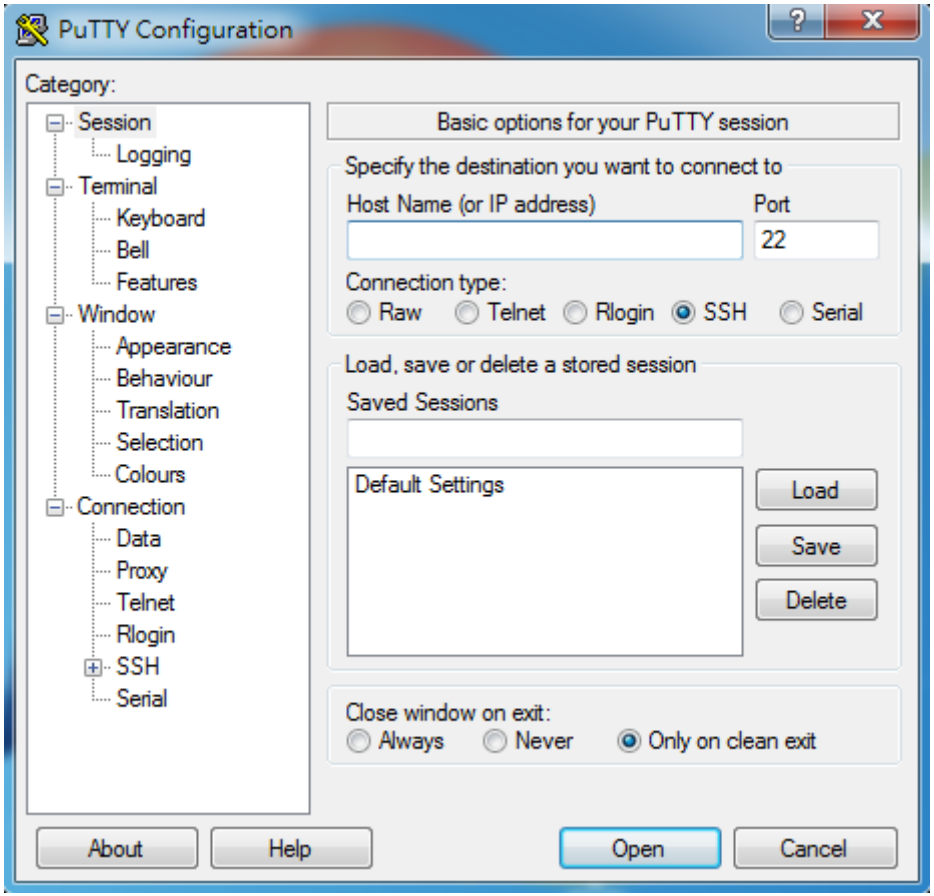
The installer packages above will provide all of these (except PuTTYtel), but you can downloa

(Not sure whether you want the 32-bit or the 64-bit version? Read the FAQ entry.)

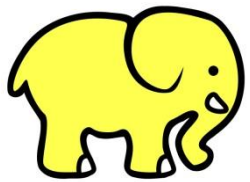
putty.exe (the SSH and Telnet client itself)

32-bit:	putty.exe	(or by FTP)	(signature)
64-bit:	putty.exe	(or by FTP)	(signature)

Putty 使用說明


PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Serial
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
22
Connection type:
Raw
Telnet
Rlogin
SSH
Serial
Load, save or delete a stored session
Saved Sessions
Default Settings
Load
Save
Delete
Close window on exit:
Always
Never
Only on clean exit
About
Help
Open
Cancel

不會 Java 也能輕鬆處理大數據 (Pig)



Pig

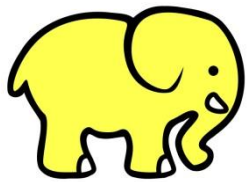
由於分散式運算的 Mapper/Reducer 是難以撰寫的因此也造成很多人的入門障礙

因此 Pig 提供了一個 Pig Latin 的語言並轉換成 Java 的 Map/Reduce 來執行大量的資料分析

競賽所用的資料集型態

int：整數
例如：12

chararray：字串或字串組
例如：美國



Pig 牛刀小試 (1-1)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t') as
(userid:chararray,gender:chararray,age:int,country:chararray,registered
:chararray);
grunt> DUMP data;
```

```
(user_000003,m,22,Canada,30-Oct-05)
(user_000008,m,23,Slovakia,28-Sep-06)
(user_000009,f,19,Canada,13-Jan-07)
(user_000010,m,19,Poland,4-May-06)
(user_000011,m,21,Finland,8-Sep-05)
(user_000012,f,28,Canada,30-Mar-05)
(user_000013,f,25,Romania,25-Sep-06)
(user_000017,m,22,Morocco,27-Aug-07)
(user_000019,f,29,Mexico,10-Nov-05)
(user_000020,f,27,United Kingdom,24-Jul-06)
```

Pig 牛刀小試 (1-2)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t') as
(userid:chararray,gender:chararray,age:int,country:chararray,registered
:chararray);
grunt> columnOne = FOREACH data GENERATE userid;
grunt> DUMP columnOne;

(user_000003)
(user_000008)
(user_000009)
(user_000010)
(user_000011)
(user_000012)
(user_000013)
(user_000017)
(user_000019)
(user_000020)
```

Pig 牛刀小試 (1-3)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t') as
(userid:chararray,gender:chararray,age:int,country:chararray,registered
:chararray);
grunt> columnOneTwo = FOREACH data GENERATE userid,gender;
grunt> DUMP columnOneTwo;

(user_000003,m)
(user_000008,m)
(user_000009,f)
(user_000010,m)
(user_000011,m)
(user_000012,f)
(user_000013,f)
(user_000017,m)
(user_000019,f)
(user_000020,f)
```

Pig 牛刀小試 (2-1)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t');
grunt> DUMP data;
```

```
(user_000003,m,22,Canada,30-Oct-05)
(user_000008,m,23,Slovakia,28-Sep-06)
(user_000009,f,19,Canada,13-Jan-07)
(user_000010,m,19,Poland,4-May-06)
(user_000011,m,21,Finland,8-Sep-05)
(user_000012,f,28,Canada,30-Mar-05)
(user_000013,f,25,Romania,25-Sep-06)
(user_000017,m,22,Morocco,27-Aug-07)
(user_000019,f,29,Mexico,10-Nov-05)
(user_000020,f,27,United Kingdom,24-Jul-06)
```

Pig 牛刀小試 (2-2)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t');
grunt> columnOne = FOREACH data GENERATE $0;
grunt> DUMP columnOne;

(user_000003)
(user_000008)
(user_000009)
(user_000010)
(user_000011)
(user_000012)
(user_000013)
(user_000017)
(user_000019)
(user_000020)
```

Pig 牛刀小試 (2-3)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t');
grunt> columnOneTwo = FOREACH data GENERATE $0,$1;
grunt> DUMP columnOneTwo;
```

```
(user_000003,m)
(user_000008,m)
(user_000009,f)
(user_000010,m)
(user_000011,m)
(user_000012,f)
(user_000013,f)
(user_000017,m)
(user_000019,f)
(user_000020,f)
```

Pig 牛刀小試 (3-1)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t');
grunt> LIMIT5 = LIMIT data 5;
grunt> DUMP LIMIT5;

(user_000003,m,22,Canada,30-Oct-05)
(user_000008,m,23,Slovakia,28-Sep-06)
(user_000009,f,19,Canada,13-Jan-07)
(user_000010,m,19,Poland,4-May-06)
(user_000011,m,21,Finland,8-Sep-05)
```

Pig 牛刀小試 (4-1)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t');
grunt> age = GROUP data BY $2;
grunt> DUMP age;

(3,{(user_000328,m,3,United States,Apr 24, 2006)})
(7,{(user_000071,m,7,Netherlands,Dec 13, 2005)})
(15,{(user_000107,m,15,Poland,Feb 1, 2006)})
(16,{(user_000215,m,16,Australia,Feb 22, 2006)})
(17,{(user_000346,f,17,Croatia,Apr 30,
2007),(user_000195,m,17,Poland,Jul 14,
2006),(user_000169,m,17,Canada,Jun 20,
2006),(user_000057,m,17,United Kingdom,Aug 27, 2006)})
```

Pig 牛刀小試 (4-2)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t');
grunt> age = GROUP data BY $2;
grunt> cAge = FOREACH age GENERATE $0,COUNT($1);
grunt> DUMP cAge;

(3,1)
(7,1)
(15,1)
(16,1)
(17,4)
(18,10)
(19,17)
(20,18)
```

Pig 牛刀小試 (4-3)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t');
grunt> age = GROUP data BY $2;
grunt> cAge = FOREACH age GENERATE $0,COUNT($1) as num;
grunt> cAge10 = FILTER cAge BY (num > 10);
grunt> DUMP cAge10

(18,13)
(19,19)
(20,19)
(21,35)
(22,34)
(23,20)
(24,20)
(25,17)
```

Pig 牛刀小試 (5-1)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t');
grunt> oage = ORDER data by $2 DESC;
grunt> DUMP oage;
```

(user_000258,f,75,Canada,21-Jun-06)
(user_000602,m,75,Taiwan,14-May-06)
(user_000071,m,7,Netherlands,13-Dec-05)
(user_000612,f,66,Taiwan,3-Mar-06)
(user_000630,f,65,China,27-Dec-04)
(user_000603,m,65,Taiwan,3-Oct-04)
(user_000607,f,63,Taiwan,3-Jun-06)
(user_000213,f,55,Germany,26-May-05)
(user_000638,f,52,China,25-Sep-05)

Pig 牛刀小試 (5-2)

```
grunt> data = LOAD '/dataset/999004/999004.tsv' USING PigStorage
('\t') as
(userid:chararray,gender:chararray,age:int,country:chararray,registered
:chararray);
grunt> oage = ORDER data by $2 DESC;
grunt> DUMP oage;

(user_000258,f,75,Canada,21-Jun-06)
(user_000602,m,75,Taiwan,14-May-06)
(user_000612,f,66,Taiwan,3-Mar-06)
(user_000630,f,65,China,27-Dec-04)
(user_000603,m,65,Taiwan,3-Oct-04)
(user_000607,f,63,Taiwan,3-Jun-06)
(user_000213,f,55,Germany,26-May-05)
(user_000638,f,52,China,25-Sep-05)
```

Pig 實戰演練 (一)

實戰一、男性與女性的會員數分別是多少？

使用資料集：/dataset/999004/999004.tsv

結果如下：
(f,137)
(m,182)

Pig 實戰演練 (二)

實戰二、會員數量排行，顯示前五名的國家？

使用資料集：/dataset/999004/999004.tsv

結果如下：
(Canada,60)
(Germany,35)
(China,34)
(Taiwan,20)
(Poland,19)

Pig 實戰演練 (三)

實戰三、年紀在20~30歲的人有多少位？

使用資料集：/dataset/999004/999004.tsv

結果如下：
(212)

Pig 實戰演練 (四)

實戰四、年紀最長者有幾人?

使用資料集：/dataset/999004/999004.tsv

結果如下：
(75,2)

Pig 實戰演練 (五)

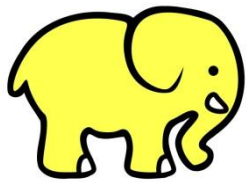
實戰五、熱門歌曲排行榜前十名?

使用資料集：/dataset/999005/999005.tsv

結果如下：

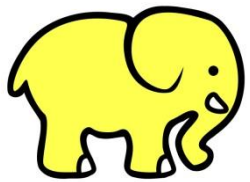
(Love Will Tear Us Apart,Boy Division,22)
(All I Need,Radiohead,19)
(Evil,Interpol,19)
(Hysteria,Muse,19)
(Such Great Heights,The Postal Service,19)
(No Surprises,Radiohead,18)
(Scar Tissue,Red Hot Chili Peppers,18)
(Smells Like Teen Spirit,Nirvana,18)
(All These Things That I'Ve Done,The Killers,17)
(Bodysnatchers,Radiohead,17)

使用 SQL 語法分析大數據(Hive)



使用 SQL 語法分析大數據 (Hive)

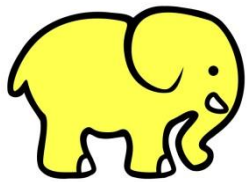
提供使用 HiveQL ，類似於 TRANSACT-SQL 的查詢語言，並轉換成 Java 的 Map/Reduce 來執行大量的資料分析



競賽所用的資料集型態

int：整數
例如：12

string：字串或字串組
例如：美國



Hive 牛刀小試 (1-1)

```
hive> CREATE EXTERNAL TABLE userprofile (
userid STRING,
gender STRING,
age INT,
country STRING,
registered STRING
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE LOCATION '/dataset/999004';
```

Hive 牛刀小試 (1-2)

```
hive> CREATE EXTERNAL TABLE userdemand (
userid STRING,
ctime STRING,
artname STRING,
traname STRING )
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE LOCATION '/dataset/999005';
```

Hive 牛刀小試 (1-3)

```
hive> show tables;

userdemand
userprofile
```

Hive 牛刀小試 (2-1)

```
hive> SELECT * FROM userprofile;

user_000003    m     22    Canada  30-Oct-05
user_000008    m     23    Slovakia      28-Sep-06
user_000009    f     19    Canada  13-Jan-07
user_000010    m     19    Poland  4-May-06
user_000011    m     21    Finland 8-Sep-05
user_000012    f     28    Canada  30-Mar-05
user_000013    f     25    Romania 25-Sep-06
user_000017    m     22    Morocco 27-Aug-07
user_000019    f     29    Mexico  10-Nov-05
user_000020    f     27    United Kingdom  24-Jul-06
```

Hive 牛刀小試 (2-2)

```
hive> SELECT userid,gender FROM userprofile;

user_000003     m
user_000008     m
user_000009     f
user_000010     m
user_000011     m
user_000012     f
user_000013     f
user_000017     m
user_000019     f
user_000020     f
```

Hive 牛刀小試 (3-1)

```
hive> SELECT * FROM userprofile LIMIT 5;

user_000003     m       22      Canada  30-Oct-05
user_000008     m       23      Slovakia        28-Sep-06
user_000009     f       19      Canada  13-Jan-07
user_000010     m       19      Poland  4-May-06
user_000011     m       21      Finland 8-Sep-05
```

Hive 牛刀小試 (4-1)

```
hive> SELECT age FROM userprofile GROUP BY age;

3
7
10
14
15
16
17
18
19
20
```

Hive 牛刀小試 (4-2)

```
hive> SELECT age,COUNT(age) FROM userprofile GROUP BY age;

3       1
7       1
10      2
14      1
15      2
16      1
17      4
18      13
19      19
20      19
```

Hive 牛刀小試 (4-3)

```
hive> SELECT * FROM (SELECT age,COUNT(age) as num FROM
userprofile GROUP BY age) agenum WHERE num > 10;
```

```
18    13
19    19
20    19
21    35
22    34
23    20
24    20
25    17
26    13
27    13
28    17
29    14
```

Hive 牛刀小試 (5-1)

```
hive> SELECT * FROM userprofile ORDER BY age DESC;

user_000258    f     75    Canada  21-Jun-06
user_000602    m      75     Taiwan  14-May-06
user_000612    f     66    Taiwan  3-Mar-06
user_000630    f     65    China   27-Dec-04
user_000603    m      65     Taiwan  3-Oct-04
user_000607    f     63    Taiwan  3-Jun-06
user_000213    f     55    Germany 26-May-05
user_000638    f     52    China   25-Sep-05
```

Hive 實戰演練 (一)

實戰一、男性與女性的會員數分別是多少？

使用資料集：/dataset/999004/999004.tsv

結果如下：
f　　137
m　　182

Hive 實戰演練 (二)

實戰二、會員數量排行，顯示前五名的國家？

使用資料集：/dataset/999004/999004.tsv

結果如下：
Canada 60
Germany 35
China 34
Taiwan 20
Poland 19

Hive 實戰演練 (三)

實戰三、年紀在20~30歲的人有多少位？

使用資料集：/dataset/999004/999004.tsv

結果如下：
212

Hive 實戰演練 (四)

實戰四、年紀最長者有幾人?

使用資料集：/dataset/999004/999004.tsv

結果如下：
75　2

Hive 實戰演練 (五)

實戰五、熱門歌曲排行榜前十名?

使用資料集：/dataset/999005/999005.tsv

結果如下：

Love Will Tear Us Apart Boy Division 22
All I Need Radiohead 19
Evil Interpol 19
Hysteria Muse 19
Such Great Heights The Postal Service 19
No Surprises Radiohead 18
Scar Tissue Red Hot Chili Peppers 18
Smells Like Teen Spirit Nirvana 18
All These Things That I'Ve Done The Killers 17
Bodysnatchers Radiohead 17