

大數據分析工具Hive

2018金象盃

Hive 命令類型

Hive 所使用的指令稱為 HiveQL,執行可以簡單分成二個步驟

建立表格 HiveQL

HiveQL

讀取	From
篩選條件	Where, Having
資料處理	GROUP, inner JOIN, outer JOIN, UNION
彙總運算	AVG, COUNT, MAX, MIN,
數學運算	ABS, ROUND,
字串處理	SUBSTRING
顯示資料	Select
排序	Order by ASC DESC



Apache Hive 交談模式

dsa50@cla50:~\$ hive ←── 開啟Hive交談模式

hive> quit; ◆ 退出Hive交談模式

dsa50@cla50:~\$



競賽所用的資料集型態

int:整數

例如:12

string:字串

例如:美國



Hive 建立表格

```
Create External Table movies(
id int,
name String,
year int,
rating float,
duration int
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/dataset/movies_data';
```



Hive 檢查表格是否建立成功

hive> show tables;

movies





請建立userprofile.tsv Hive表格

資料樣式

user_000003	m	22	United States	Oct 30, 2005
user_000008	m	23	SlovakiaSep 28,	2006

HQL語法分析

HQL寫法

1.SELECT ←



程式裡面一定要出現

3.WHERE

2.FROM

- 4. GROUP BY
- 5. HAVING
- 6.ORDER BY

HQL程式處理方式

- 1.FROM
- 2. WHERE
- 3. GROUP BY
- 4. HAVING
- 5.SELECT
- 6. ORDER BY



載入資料集並顯示前5筆原始資料

\$ hive

hive> select * from movies limit 5;

- 1 The Nightmare Before Christmas 1993 3.9 4568
- 2 The Mummy 1932 3.5 4388
- 3 Orphans of the Storm 1921 3.2 9062
- 4 The Object of Beauty 1991 2.8 6150
- 5 Night Tide 1963 2.8 5126



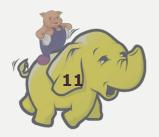


如何顯示全部的資料?

1920年共有幾部電影?

hive>select * from movies where year=1920

```
62 Dr. Jekyll and Mr. Hyde 1920 3.2 4679
79 The Mark of Zorro 1920 3.1 6433
189 Way Down East 1920 3.0 8985
2744 One Arabian Night 1920 2.5 6203
2745 Anna Boleyn 1920 2.4 7118
42676 Headin' Home 1920 2.7 4390
```





2009年共有幾部電影?

「只」列出電影名稱

hive>select name from movies;

:::

Lady Gaga & The Muppets' Holiday Spectacular

Sunset Strip

Silver Bells

Winter Wonderland

Top Gear: Series 19: Africa Special

Fireplace For Your Home: Crackling Fireplace with Music

Kate Plus Ei8ht

Kate Plus Ei8ht: Season 1



將電影由年份遠到近排序

hive>select * from movies order by year ASC;

```
:::
```

```
(45575, Caillou: Season 5: Big Time Caillou & Other Stories: Blast Off to Space, 2013, 1551) (44561, In APPropriate Comedy, 2013, 2.6, 5020) (49561, The Square (Trailer), 2014, 3.6, 154)
```



列出片長大於2小時的電影資料

hive>select * from movies where duration > 7200;

:::

(45108, Settai, 2013, 2.8, **7390**)

(46832, WWE for All Mankind: The Life & Career of Mick Foley,

2013,4.2,**8066**)

(46949, Example Short XII 23976 2 Hour, 2010, 2.6, 7370)

(48714, WWE: The Top 25 Rivalries in Wrestling History, 2013, 3.7, **10228**)

(49055, Theeya Velai Seyyanum Kumaru, 2013, 3.3, **8239**)





列出片長大於2.5小時且 評價界於3.5~4.5之間 的電影資料

列出電影資料集的評價、電影名稱欄位

hive>select rating, name from movies;

```
(2.8, Winter Wonderland)
(,Top Gear: Series 19: Africa Special)
(,Fireplace For Your Home: Crackling Fireplace with Music)
(2.7, Kate Plus Ei8ht)
(2.7, Kate Plus Ei8ht: Season 1)
```



列出有評價的資料

hive>select rating, name from movies where rating is not null;

(3.0, Sunset Strip)

(3.5, Silver Bells)

(2.8, Winter Wonderland)

(2.7,Kate Plus Ei8ht)

(2.7, Kate Plus Ei8ht: Season 1)





列出電影資料集的片長、 評價欄位並過濾掉所有空值 ex. (片長,評價)

列出總電影比數小於50部的是哪幾年

hive>select * from (select year,count(*) as num from movies
group by year) movies_1 where num < 50</pre>

```
1946 6
1947 9
1948 13
1949 9
1950 10
1951 33
```



列出總電影比數大於50部小於100部的有哪幾年

hive>select * from (select year,count(*) as num from movies
group by year) movies_1 where num > 50 and num < 100;</pre>

:::

```
1956 60
1957 98
1958 73
1959 87
1963 88
```



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:userprofile.tsv

結果如下:

f 110

m 155



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:userprofile.tsv

結果如下:

United States 60

United Kingdom 35

Poland 19

Turkey 14

Canada 12



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:userprofile.tsv

結果如下:

189



Hive 實戰演練 (四)

實戰四、年紀最長者有幾人?

使用資料集:userprofile.tsv

結果如下:

75 1



Hive 實戰演練 (五)

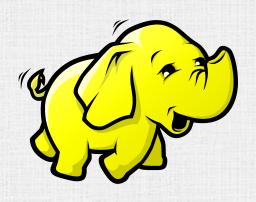
實戰五、熱門歌曲排行榜前十名?

使用資料集:userdemand.tsv

結果如下:

Karma Police Radiohead 95
Paranoid Android Radiohead 92
Creep Radiohead 91
Somebody Told Me The Killers 91
Wonderwall Oasis 89
No Surprises Radiohead 86
Take Me Out Franz Ferdinand 86
Love Will Tear Us Apart Boy Division 85
Such Great Heights The Postal Service 85
Smells Like Teen Spirit Nirvana 84





競賽資料集說明

2018金象盃說明會

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			
L. CTATAALLON				

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/20445/2	0445.csv			
資料總行數	200				
資料大小	3 KB				
資料說明	名稱為 20445.cs	v 的檔案,是	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	欄位名稱	資料型態	資料型態	備註	
(英文)	(中文)	(Pig)	(Hive)		
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	int	int		
	人數				

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/	<u>/dataset/5958</u>		
資料中文名稱	各縣(市)警察(分)局]暨所屬分駐()	底出)所地址資	料
資料檔案名稱	5958.csv			
HDFS完整路徑	/dataset/5958/5958	8.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	欄位名稱	資料型態	資料型態	備註
(英文)	(中文)	(Pig)	(Hive)	
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/6087/60	87.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1	07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	欄位名稱	資料型態	資料型態	備註	
(英文)	(中文)	(Pig)	(Hive)		
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.tv	w/dataset/60	86		
資料中文名稱	107年幼兒園名錄	ξ			
資料檔案名稱	6086.csv				
HDFS完整路徑	/dataset/6086/60	86.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1	06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	欄位名稱	資料型態	資料型態	備註	
(英文)	(中文)	(Pig)	(Hive)		
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/9621/962	1.csv		
資料總行數	10081			
資料大小	676 KB			
資料說明	名稱為 9621.csv 的	内檔案,是大專	厚校院各校科	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	欄位名稱	資料型態	資料型態	備註
(英文)	(中文)	(Pig)	(Hive)	
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	
				44 (14 4 3

「ATM位置」查詢一覽表

原始資料來源	https://data.gov.tw/c	https://data.gov.tw/dataset/24333				
資料中文名稱	「ATM位置」查詢-					
資料檔案名稱	24333.csv					
HDFS完整路徑	/dataset/24333/243	33.csv				
資料總行數	22584					
資料大小	2.1 MB					
資料說明	名稱為 24333.csv 的	内檔案,是「A	ATM位置」查詢	旬一覽表		
	資料欄位說明	(每個欄位之間	間均已逗號分鬥	鬲)		
欄位名稱	欄位名稱	資料型態	資料型態	備註		
(英文)	(中文)	(Pig)	(Hive)			
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		

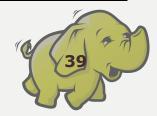


農產品交易行情

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.cs	SV						
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				
	資料欄位說日	明 (每個欄位之間	間均已逗號分隔)					
欄位名稱	欄位名稱	資料型態	資料型態	備註				
(英文)	(中文)	(Pig)	(Hive)					
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							
average_price	平均價	float	float	de con				
trading_volume	交易量	int	int	(4/38)				

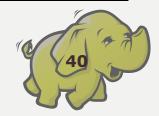
A1類交通事故資料

原始資料來源	https://data.gov.tw/dataset/12197				
資料中文名稱	A1類交通事故資制	科			
資料檔案名稱	12197_A1.csv				
HDFS完整路徑	/dataset/12197_A	\1/12197_A	I.csv		
資料總行數	1285				
資料大小	89 KB				
資料說明	名稱為 12197_A	1.csv 的檔案	,是106年的	JA1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	欄位名稱	資料型態	資料型態	備註	
(英文)	(中文)	(Pig)	(Hive)		
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		



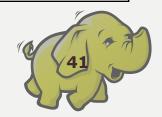
A2類交通事故資料

		/ / / / / / / / / / / / / / / / / / / /	407		
原始資料來源	https://data.gov.tw/dataset/12197				
資料中文名稱	A2類交通事故資制				
資料檔案名稱	12197_A2.csv				
HDFS完整路徑	/dataset/12197_ <i>F</i>	\2/12197_A2	2.csv		
資料總行數	281634				
資料大小	19 MB				
資料說明	名稱為 12197_A2	2.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	欄位名稱	資料型態	資料型態	備註	
(英文)	(中文)	(Pig)	(Hive)		
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.tv	w/datacat/26	557			
		Widalaseli 20	<u> </u>			
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/26557/2	6557.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/	 		
	資料欄位說明 (每個欄位之間均已逗號分隔)					
欄位名稱	欄位名稱	資料型態	資料型態	備註		
(英文)	(中文)	(Pig)	(Hive)			
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	一大國外消費	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/63029/6	3029.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	欄位名稱	資料型態	資料型態	備註	
(英文)	(中文)	(Pig)	(Hive)		
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,是十六縣周	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	欄位名稱	資料型態	資料型態	備註			
(英文)	(中文)	(Pig)	(Hive)				
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/999001/99	9001.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,是連絡	 泉紀錄		
	資料欄位說明	(每個欄位之間]均已Tab鍵分[隔)	
欄位名稱	欄位名稱	資料型態	資料型態	備註	
(英文)	(中文)	(Pig)	(Hive)		
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		

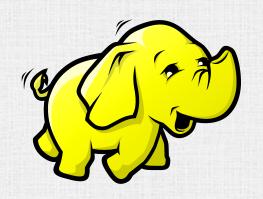
攻擊者下載檔案紀錄

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/999002/9	99002.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	常案紀錄		
	資料欄位說明 (每	個欄位之間均	已 Tab鍵 分隔	Ā)		
欄位名稱	欄位名稱	資料型態	資料型態	備註		
(英文)	(中文)	(Pig)	(Hive)			
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

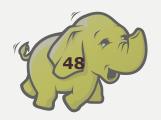
原始資料來源	社會經濟資料服務平台					
資料中文名稱	107年6月行政區分齡	兒童及少年性別	川人口統計_縣市	<u> </u>		
資料檔案名稱	999003.csv					
HDFS完整路徑	/dataset/999003/999	003.csv				
資料總行數	24					
資料大小	1 KB					
資料說明	名稱為 999003.csv 的	 勺檔案,是107章	 ∓6月行政區分齒	令兒童及少年性別人口統計_縣		
	市					
	資料欄位說明	月 (每個欄位之間	引均已逗號分隔) 11均已逗號分隔)			
欄位名稱	欄位名稱	資料型態	資料型態	備註		
(英文)	(中文)	(Pig)	(Hive)			
country_id	縣市代碼	int	int			
country	縣市名稱	chararray	string			
a0a5_cnt	0-5歳	int	int			
	兒童人口數					
a6a11_cnt	6-11歳	int	int			
	兒童人口數					
a12a17_cnt	12-17歲	int	int			
	少年人口數					



競賽環境操作(VPN)

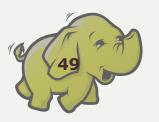
2018金象盃說明會

http://gec2018trn.azurewebsites.net/



https://goo.gl/FSDv2n



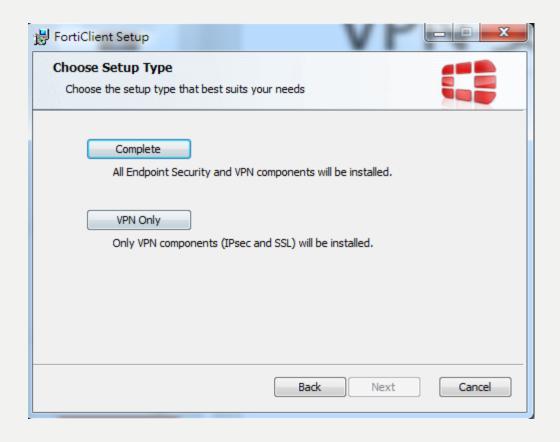


VPN 安裝 (一)



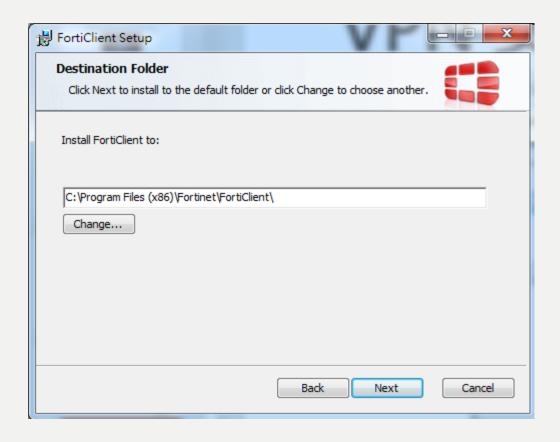


VPN 安裝 (二)



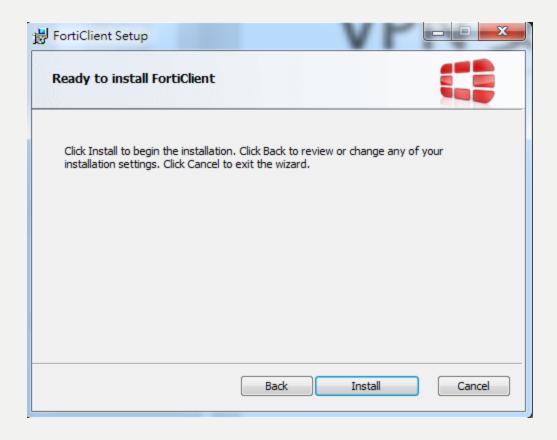


VPN 安裝 (三)



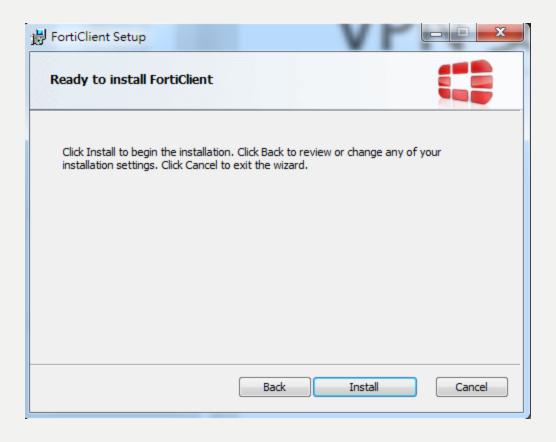


VPN 安裝 (四)



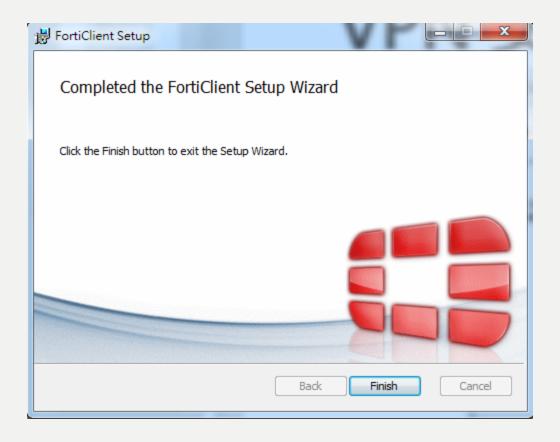


VPN 安裝 (四)



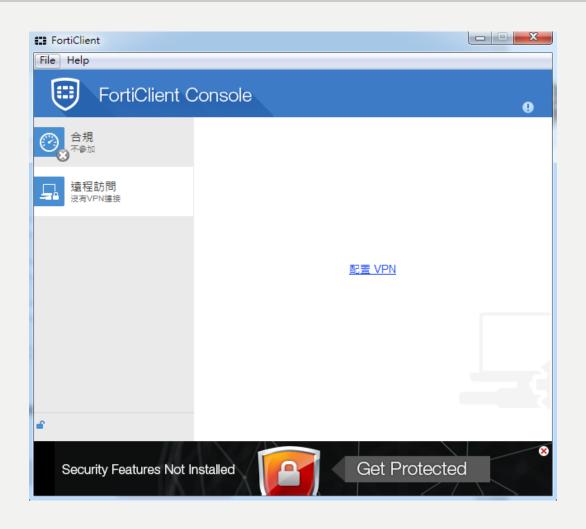


VPN 安裝 (五)





VPN 連線說明 (一)





VPN 連線說明 (二)



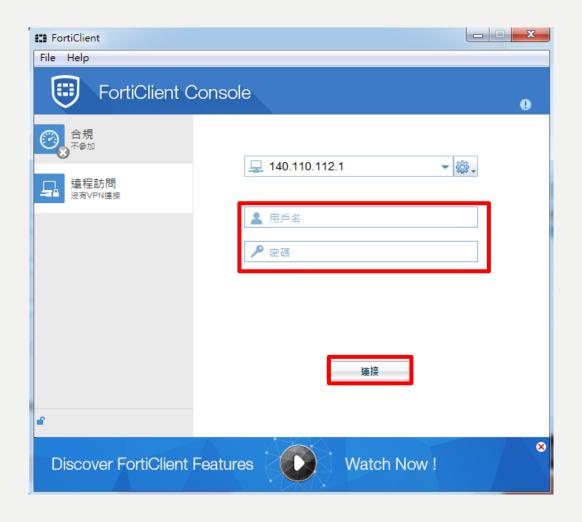


VPN 連線說明 (三)



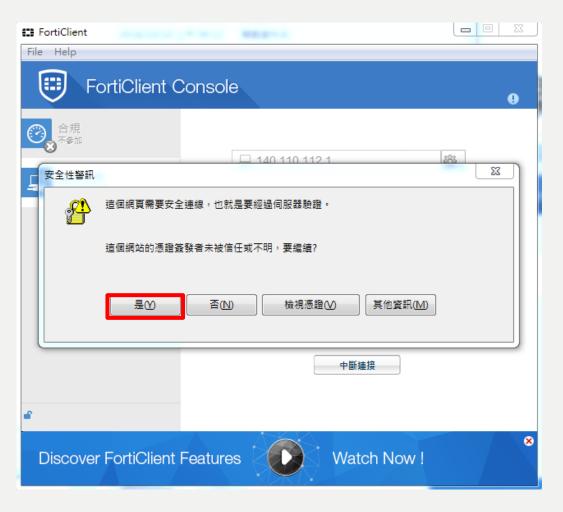


VPN 連線說明 (四)





VPN 連線說明 (五)





VPN 連線說明 (六)



