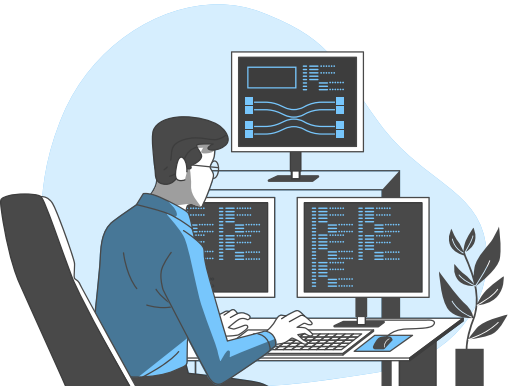
第二組-Spark

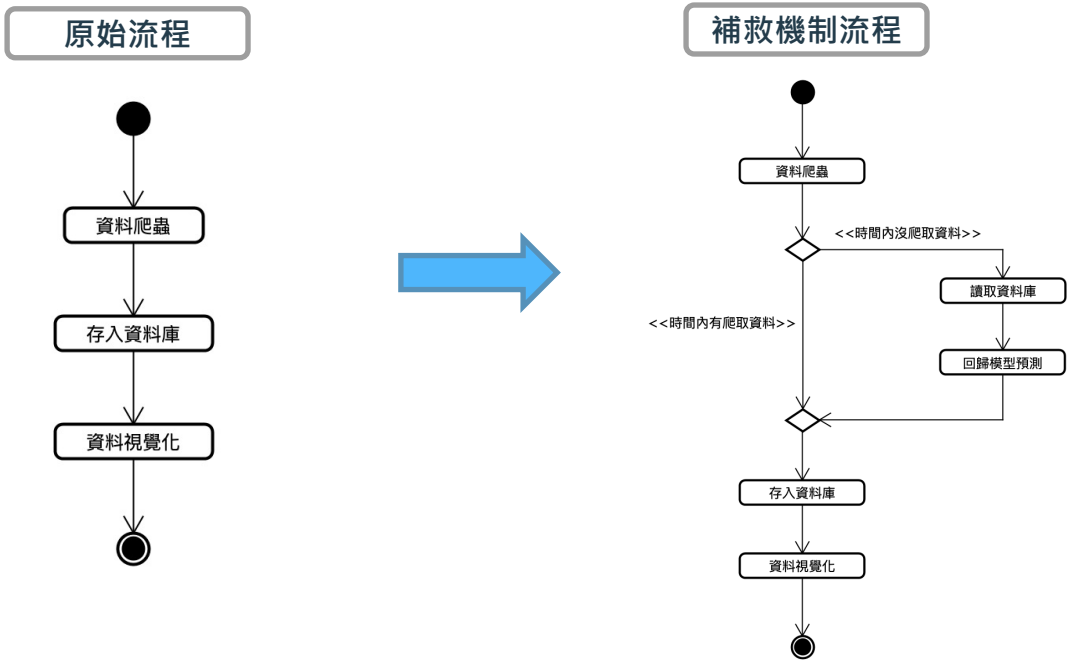
Data Remediation by Spark for Web Crawler



資碩計一	110753157	吳仁凱
資碩計一	110753165	張修誠
資碩計一	110753202	何彥南
資碩工一	110753117	莊崴宇
資碩工一	110753135	姚惠馨
資管四甲	107306009	吳泓澈

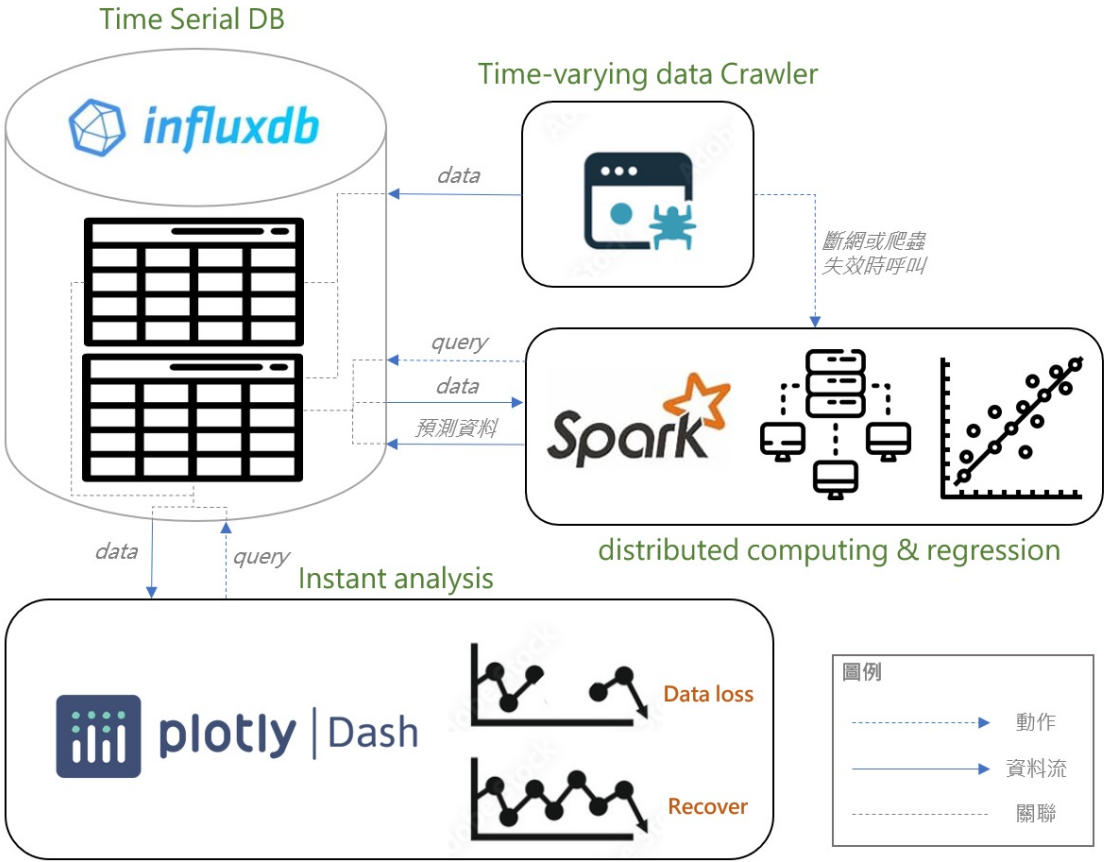
專案介紹

- 功能性：網路爬蟲之缺失值補救機制與視覺化
- 非功能性：自動補足爬蟲缺失資料



目錄

API展示圖


Time Serial DB
influxdb
Time-varying data Crawler
data
斷網或爬蟲
失效時呼叫
query
data
預測資料
Spark
data
query
distributed computing & regression
Instant analysis
plotly | Dash
Data loss
Recover
圖例
動作
資料流
關聯

01

Web Crawler

Web Crawler

- FinMind API

  → 金融開源資料庫

- 時間序列資料

  → 每二十秒call一次API

- 每日股票資訊

  → 欄位資訊：
  amount、average_price、buy_price、buy_volume、change_price、Change_rate、**close**、high、low、open、sell_price、sell_volume、total_amount、total_volume、volume、volume_ratio、yesterday_volume、**date**、stock_id、TickType

  → 以資料時間和收盤價進行視覺化

參考資料：FinMind API

02

Data Prediction

MLlib

- 當爬蟲出現問題時
  → Call MLlib並傳送未成功爬取資料之時間

- LinearRegression作為預測模型
  → 以未成功爬取資料時間點之前的資料進行預測

- 預測資料設定
  → his_num：預測資料筆數，default = 10
  → his_win：線性迴歸變數之各組預測資料筆數，default = 3

EX. 我們有一個 1~12 的時間序列輸入

	His1	His2	Target
train	1	2	3
	2	3	4
	3	4	5
	4	5	6
	5	6	7
	6	7	8
	7	8	9
test	8	9	10
	9	10	11
	10	11	12
predict	11	12	*

預測程式碼：Spark_Prediction

03

Database

InfluxDB

- 時間序列資料庫
  → 適合儲存即時、不重複資料
- 爬蟲資料寫入兩個Table
  → web_crawler_data存放爬蟲資料
  → prediction_data存放爬蟲資料和預測資料
- 支援多種程式語言
  → 透過API即可直接取得資料庫資料

Read.py 說明

```
1  #引入influxdb套件
2  from influxdb import InfluxDBClient
3
4  #與DB建立連線
5  #InfluxDBClient(資料庫IP,資料庫PORT,帳號,密碼,DB名稱)
6  client = InfluxDBClient('                ',8086,'','','stock_data')
7
8  #撈取爬蟲TABLE的資料
9  web_crawler_data = client.query('select * from web_crawler_data')
10 #撈取預測TABLE的資料
11 prediction_data = client.query('select * from prediction_data')
12
13 #對從爬蟲TABLE爬取的資料進行整理 並印出
14 print(list(web_crawler_data.get_points()))
15 print("------------")
16 #對從預測TABLE爬取的資料進行整理 並印出
17 print(list(prediction_data.get_points()))
18 print("------------")
19 #對從預測TABLE爬取的資料進行整理(抓出最後一個欄位的value) 並印出
20 print(list(prediction_data.get_points())[-1]['value'])
21
```

資料庫IP

此為DB名稱

此為Table名稱

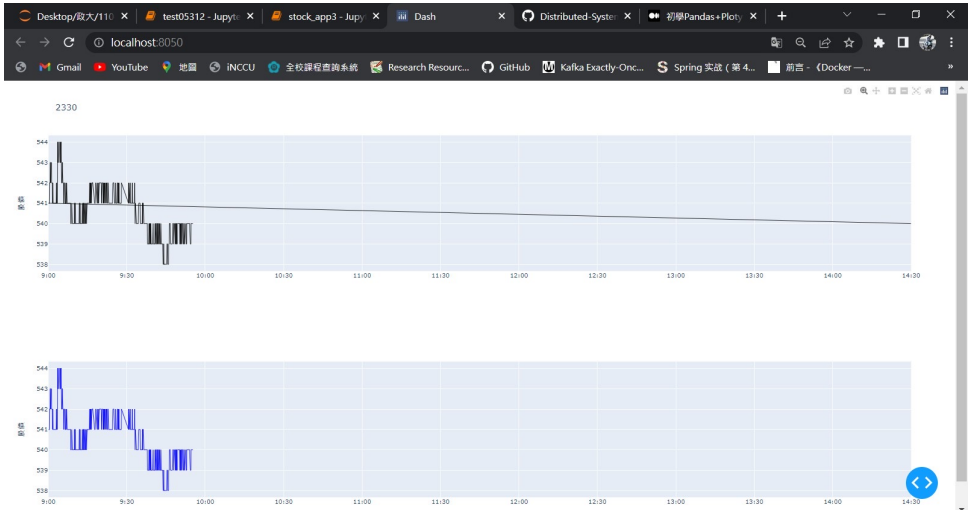
influxdb

04

Data Visualization

Visualization

- Plotly API
  → 將DataFrame繪製輸出視覺化圖表
- Dash API
  → 將視覺化圖表部署至網站上
- 每五秒從Database更新一次資料



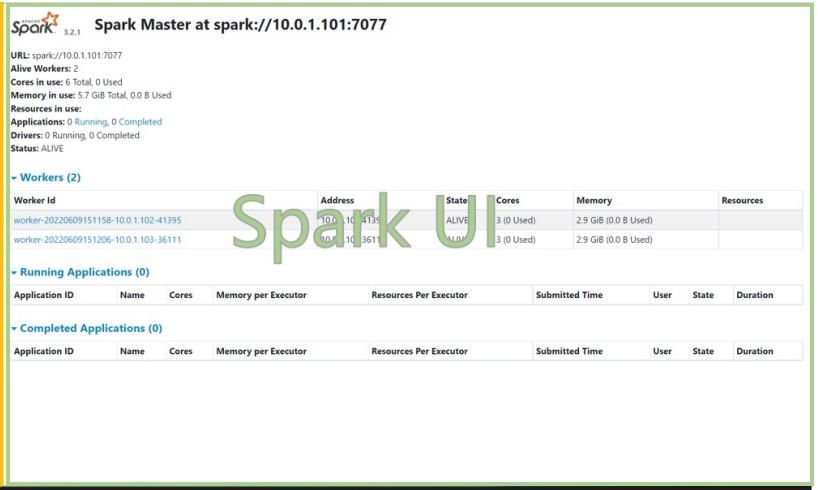
參考資料：[Visualization with Plotly and Dash]

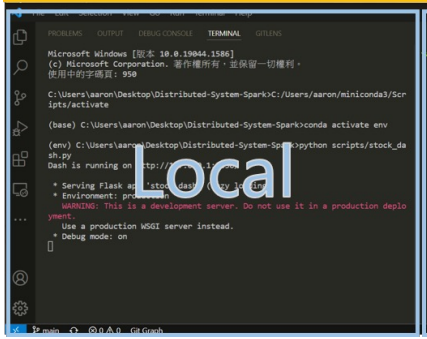
05

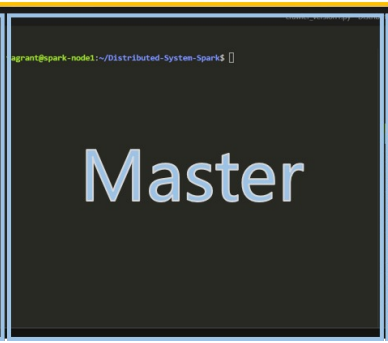
Demo

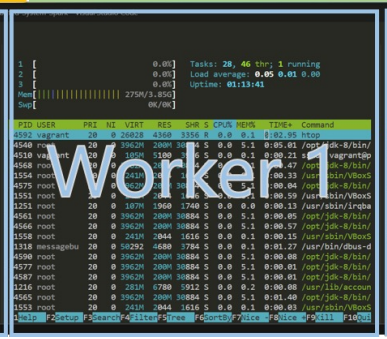
Demo Video

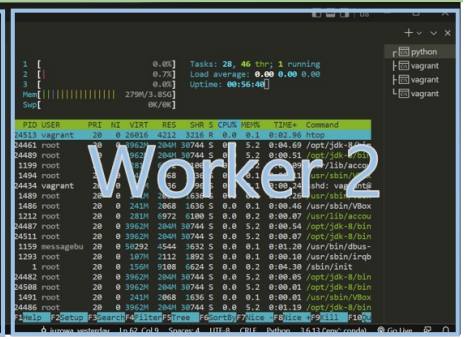

Dashboard


Spark Master at spark://10.0.1.101:7077
URL: spark://10.0.1.101:7077
Alive Workers: 2
Cores in use: 6 Total, 0 Used
Memory in use: 5.7 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE
Workers (2)
Running Applications (0)
Completed Applications (0)
Spark UI


Local


Master


Worker 1


Worker 2

Thanks!

Demo影片網址：https://www.youtube.com/watch?v=76fHSBQzw_Y

Github網址：https://github.com/C-WeiYu/Distributed-System-Spark