

系統設計的...

RC

Scenario

如果想要建立一個服務

給予客戶ID服務回傳相關資訊，如右邊所列，然後需於秒級內回覆。

那請問，我們要如何開始呢？

很不幸，基本資料，資產狀況，近來事件 都不是位於同一個資料庫，且目前資料定義可能不太相同

基本資料 姓名，性別，年齡，...

資產狀況 存款，股票投資金額，基金投資金額，...

近來事件 信用卡消費，提款時間，...



第一步會是



第二步會是



第三步會是



我的第一步

跟使用者再次確認

需求



我的第二步

特性決定使用元件

- 秒級內回覆
- 水平擴充容易
- 各查詢服務依賴性低
- 動態決定每個欲查詢ID查詢路徑



我的第三步

系統架構圖

系統監控機制

資料流程圖



那我們今天就來實做以下情境

- | | | | |
|-----|--------|-------|------------------------|
| 1 | 給予一個ID | 2.3.a | 查詢04/17 02:00-04:00的事件 |
| 2.1 | 查詢基本資料 | 2.3.b | 查詢04/17 04:00-06:00的事件 |
| 2.2 | 查詢資產資料 | 3.1 | 彙整資料並儲存至 DB |

我使用的工具



<http://kafka-python.readthedocs.io/en/master/>

```
>>> from kafka import KafkaProducer
>>> producer = KafkaProducer(bootstrap_servers='localhost:1234')
>>> for _ in range(100):
...     producer.send('foobar', b'some_message_bytes')
```

生產者

```
>>> from kafka import KafkaConsumer
>>> consumer = KafkaConsumer('my_favorite_topic')
>>> for msg in consumer:
...     print (msg)
```

消費者

notebook time

pymongo
kaka-python

Data

人



身分證字號

```
{ "姓名": RC,  
  "性別": 男性,  
  "年齡": 36,  
  "存款": 26,000,  
  "基金": 13,111,  
  "股票": 341,  
  ...  
  ...  
}
```

資料



資料序號

人



身分證字號

資料

```
{ "姓名": RC,  
  "性別": 男性,  
  "年齡": 36,  
  "存款": 26,000,  
  "基金": 13,111,  
  "股票": 341,  
  ...  
  ...  
}
```

1 照key值順序排列

2 串接 value 值

3 對其取 MD5 運算



資料序號

人



DNA序列

```
{ "姓名": RC,  
  "性别": 男性,  
  "年龄": 36,  
  "存款": 26,000,  
  "基金": 13,111,  
  "股票": 341,  
  ...  
  ...  
}
```

?

資料



資料DNA

Categorical Variables

as the original

Numeric Variables

Normalization -> Ranking

性別	有無信用卡	AUM	最近提款次數											
M	1	080	0	N	N	Y	0	10	099	81	4	31		
M	1	015	1	N	N	Y	0	10	099	81	4	31		
M	1	015	0	N	N	Y	0	10	099	80	4	31		
M	1	015	2	N	N	Y	0	10	099	78	2	31		
M	1	015	0	Y	Y	N	0	10	099	78	2	31		

Rolling Hash Algorithm

vs.

BLAST

[https://zh.wikipedia.org/wiki/BLAST_\(生物資訊學\)](https://zh.wikipedia.org/wiki/BLAST_(生物資訊學))

Rolling Hash Algorithm

local sensitive hash algorithm

MD5 Hash

abcdef

e80b5017098950fc58aad83c8c14978e

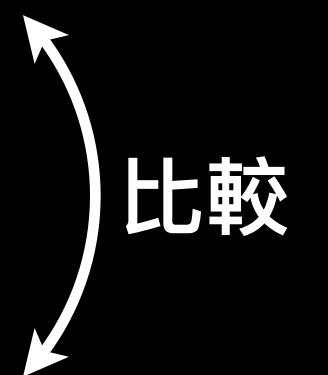
bcdefg

02b04a9c28417ea6b6657b638497e63e

Fuzzy Hash

394,398,402

398,402,406



BLAST

Sequence - Sensitive

1

Query sequence: PQGEFG

Word 1: PQG

Word 2: QGE

Word 3: GEF

Word 4: EFG

2

Query sequence: R P P Q G L F

Database sequence: D P P E G V V

Exact match is scanned.

Score: -2 7 7 2 6 1 -1

HSP

Optimal accumulated score = $7+7+2+6+1 = 23$

3

Query sequence

Newly joined region,
then it is extended to be
an HSP region.

Distance $< A$

Database sequence

可能應用

- 快速找尋相似的人
 - 省卻複雜的 SQL 撰寫？
 - Ex.,
 - 我知道「Roger」是潛力VIP, 那麼我想要拿他去黏「我們資料庫」跟他相像的人
 - 過去用將屬性向量化，做距離計算
 - 以後給予 DNA 直接搜尋
- 建立模型效益的 **baseline**
 - 用目標 DNA (Rule Based) 找到的那群人，理論上應該有比亂猜好
 - 那麼模型應該要比這方法好

界

界忙

界 忙 瞞