CYBERSEC 2023 臺灣資安大會 | 5/9 - 5/11 臺北南港展覽二館

BRING SECURITY TO

資料匿名化, 安心用資料

古永忠 博士

ycku@pgsql.tw

 # Da*a Se*ity**
 # Inf**tion Go*er*an*e**
 # *erso*al Idenfiable Infor***ion**

PostgreSQ

古永忠 (linkedin.com/in/ycku/)

國立臺灣大學
資訊工程研究所博士

資訊工業策進會工程師

DATABASE SYSTEM

研究者

實踐者

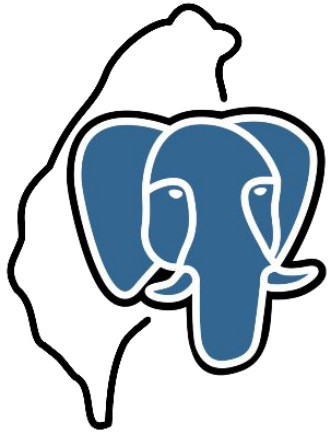
倡議者

行政院主計處研究員

國泰人壽系統資訊部副理

PostgreSQL 台灣使用者社群召集人

PostgreSQL Taiwan



正體中文使用手冊
docs.postgresql.tw
協作

PostgreSQL.TW社團
fb.com/groups/pgsql.tw
討論

開源人年會
coscup.org/2023
倡議

個人資料(Personally Identifiable Information)

- 個人資料保護法 - 第 2 條 - 第一項
  - https://law.moj.gov.tw/LawClass/LawSingle.aspx?pcode=i0050021&flno=2
  - 個人資料:指自然人之姓名、出生年月日、國民身分證統一編號、護照號碼、特徵、指紋、婚姻、家庭、教育、職業、病歷、醫療、基因、性生活、健康檢 查、犯罪前科、聯絡方式、財務情況、社會活動及其他得以直接或間接方式識別該個人之資料。
- 個人資料保護法 - 第 41 條 - 第一項
  - https://law.moj.gov.tw/LawClass/LawSingle.aspx?pcode=I0050021&flno=41
  - 意圖為自己或第三人不法之利益或損害他人之利益，而違反第六條第一項、第十五條、第十六條、第十九條、第二十條第一項規定，或中央目的事業主管機關依第二十一條限制國際傳輸之命令或處分，足生損害於他人者，處五年以下有期徒刑，得併科新臺幣一百萬元以下罰金。

不要用最安全

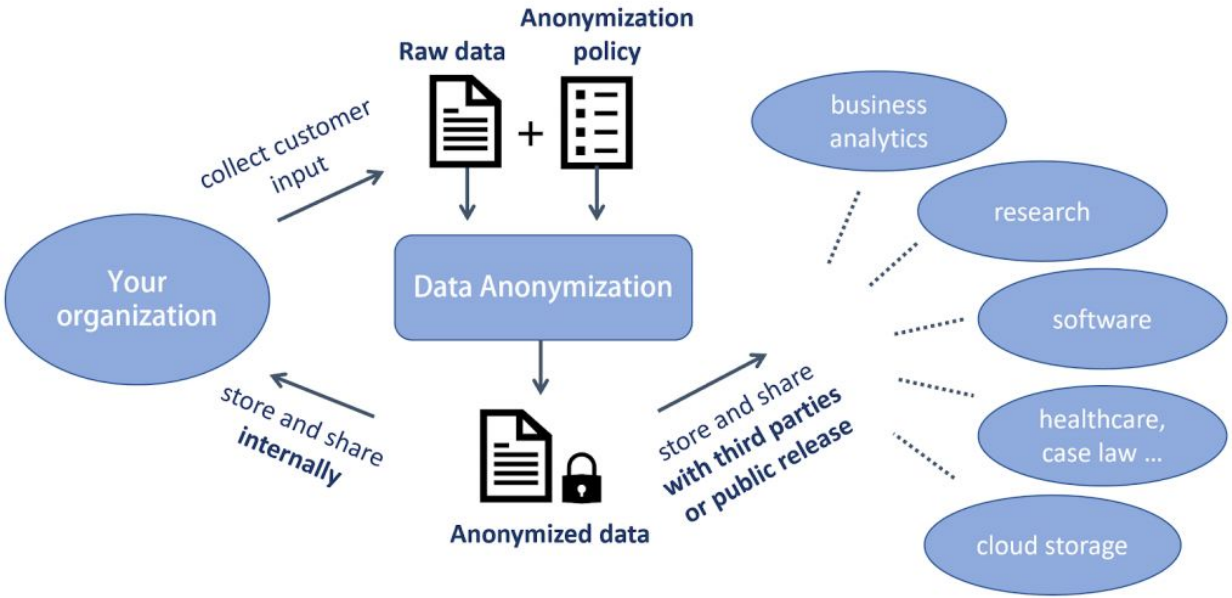
PostgreSQ

案例

- iRent洩個資 雙北各裁罰9萬｜強化資安| 產經| 聯合新聞網
- 國內個資外洩詐騙案頻傳, 2021年報案件數劇增, 誠品書店、東森購物與王品集團最嚴重
- 前科累累！因超過5 億名Facebook 用戶資料外洩, Meta 再次遭重罰2.76 億美元- INSIDE
- Three years of GDPR: the biggest fines so far - BBC News
- 涉違法蒐集用戶個資 南韓重罰Google、Meta｜ 公視新聞網PNN

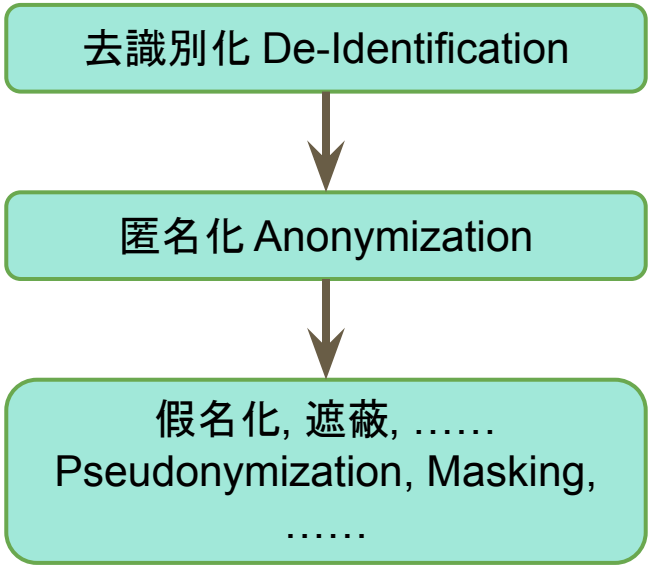
- 個人資料值多少？
  - 參議員Warner得出2018年美國大型科技公司、資料處理公司等, 其一條個人資料大約價值240美元, 或是一個月 20美元。
  - 預因此估至2022年, 一條個資的價 值大約可以攀升至600美元。

資料匿名化(Data Anonymization)



European Commission: Data Anonymization

個資操作分類


去識別化 De-Identification
匿名化 Anonymization
假名化, 遮蔽, ……
Pseudonymization, Masking,
……

資料匿名化的意義

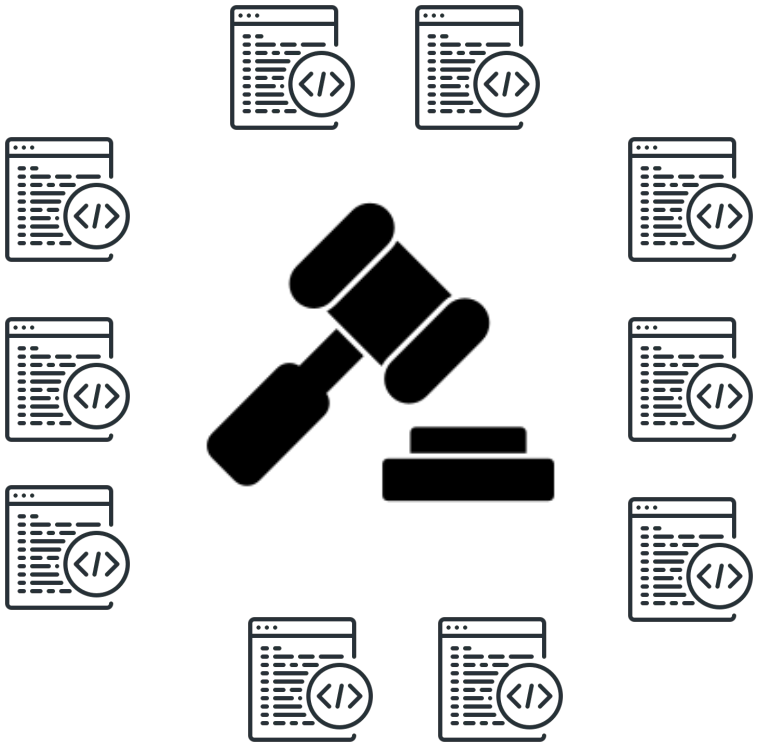
- **罰金避免**
  - 降低營運成本。
  - 避免負面商譽。
- **減少接觸原始資料的風險**
  - 可積極地增加原始資料區域的安全性。
  - 原始資料庫的效能增加。
- **積極地使用資料**
  - 增加業務擴展的可能性。
  - 增加公司及客戶利益。
- **保護同仁, 避免無心之過**
  - 減少人為錯誤的維護成本。
  - 減少繁瑣的程序, 增進工作效率。

資料匿名化(Anomymization) v.s. 資料遮蔽(Masking)

- **資料匿名化 (理想目的)**
  - 資料匿名化是指將個人資料轉換為無法識別特定自然人身份的形式。
  - 不再被視為個人資料, 因為無法透過這些數據來識別特定自然人。
  - 不幸外洩時風險最低。
  - 資料可利用性不高。
- **資料遮蔽 (現實方法)**
  - 將原始個人資料的敏感部分部分進行虛擬化、刪除或替換。
  - 仍然視為個人資料, 這些數據仍然可能被進一步分析以識別特定自然人。
  - 不幸外洩時風險較低。
  - 保持資料可利用性的平衡。

法風稽為核心

- 區分何者為個資。
- 建立良好制度。
- IT 同仁只是輔助。
- 資料匿名化只是一種邏輯。
- 互相跨域學習是合理的實務做法。
- 資料分析需考慮標的為遮蔽資料。
- 法風稽決策明確，並為主要決策者。



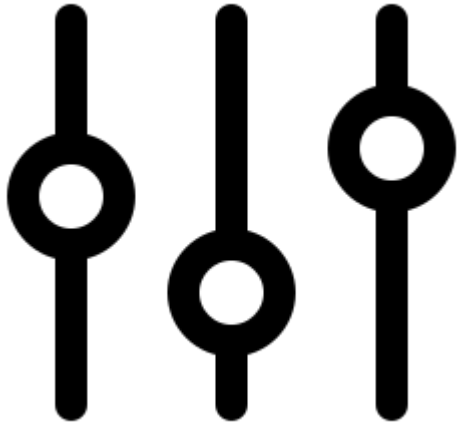
區分資料環境

- **原始資料環境**
  - 資料交易應用
- 限制
  - 操作權限限制
  - 操作環境限制
  - 操作來源限制
  - 操作目的限制
  - 操作形式限制
  - 操作範圍限制
  - 操作數量限制
  - 操作軌跡限制

- **資料遮蔽環境**
  - 資料分析
  - 資料測試
  - 程式開發
  - .....
- 限制
  - 仍然屬於不可外洩資料
  - 去識別化 只是用小毛巾遮臉 | 政府作為
- 因為不需要接觸原始資料環境，自然減少外洩風險。

資料遮蔽環境

- 遮蔽後資料難以運算還原
- 資料庫結構相容
  - 資料欄位大小
  - Primary Key / Foreign Key
  - JOIN
  - 內容規則維持
  - 資料操作不能因為遮蔽而改變
- 可分析性維持
  - 次序
  - 範圍
  - 分類
- 遮蔽運算效率高
  - 可處理大量資料



保持資料可利用性的平衡

ISO 27002 - Guidance – Data Masking Principles

- https://www.isms.online/iso-27002/control-8-11-data-masking/
- 實行資料遮蔽時，僅向任何使用它的人透露最少量的資料。
- 「混淆（隱藏）」某些資料，並且僅允許人員存取與他們相關的部分。
- 遵循特定的法律和監管準則來實作資料遮蔽行為。
- 在實施假名化的情況下，用於「解除遮蔽」的資料演算法應確保其安全。

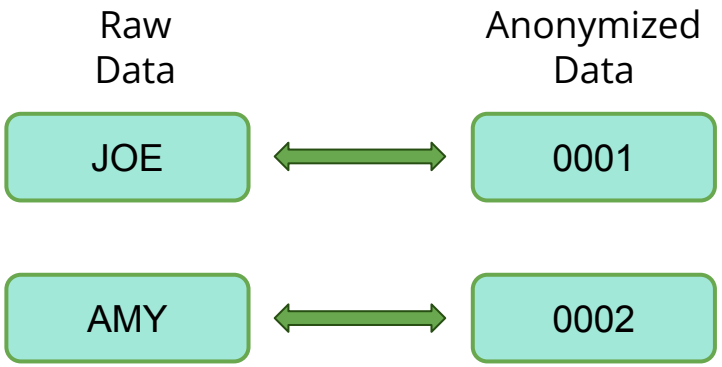
Data Anonymization Techniques

- Pseudonymization
- Generalization
- Data swapping
- Data perturbation
- Synthetic data
- Data masking

- 保持資料的可利用性
- 根據資料利用的情況選擇匿名方法
  - 資料分析
  - 資料測試
  - 程式開發
  - .....
- 不一定僅能輸出一套匿名資料
- 資料越好用，通常越不安全

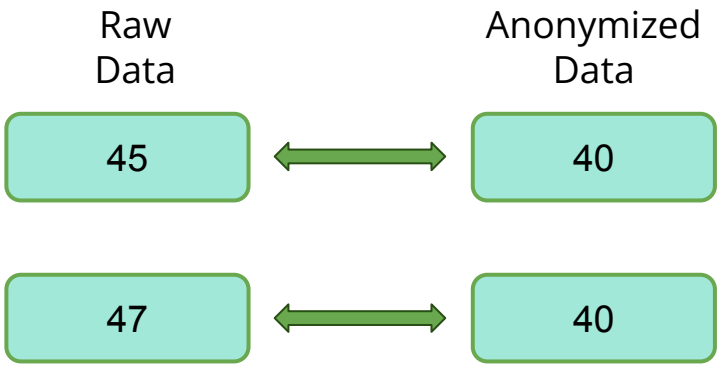
Pseudonymization

- 替換為虛擬的對應內容
  - 例:會員ID
- 優點
  - 無法單獨以轉換後 內容推得原始內容
- 缺點
  - 需記錄轉換 內容
  - 規則太複雜的運算效率可能不符合經濟效益
  - 資料可反 查, 要好好保護



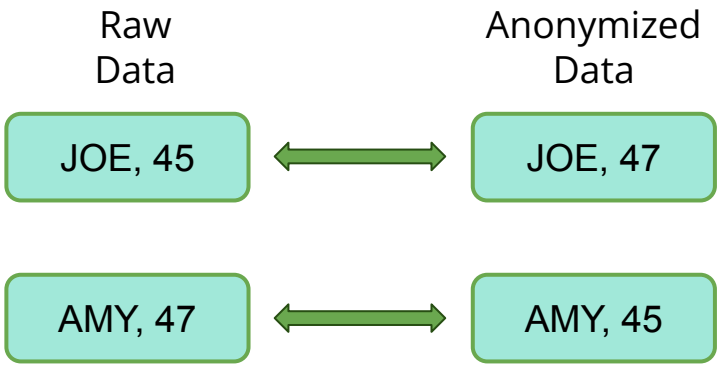
Generalization

- 替換為精確度較低的內容
- 優點
  - 運算快速且無法反解
- 缺點
  - 資料分析的精確度也會降低
  - 可能會測不到有問題的特殊 值



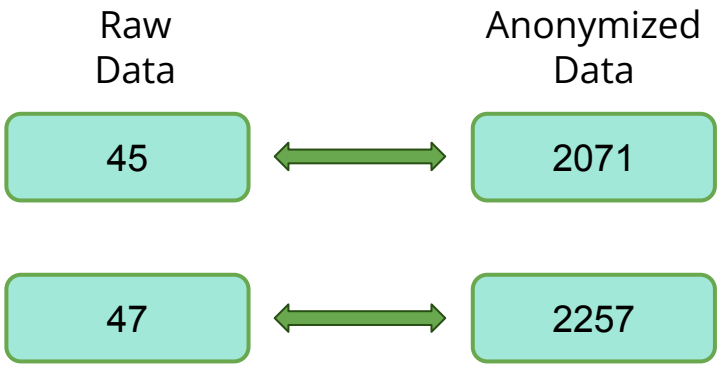
Data swapping

- 與另一筆資料交換內容
- 優點
  - 不影響統計性分析
- 缺點
  - 資料太少時可能會被重組還原
  - 無法進行相關性分析



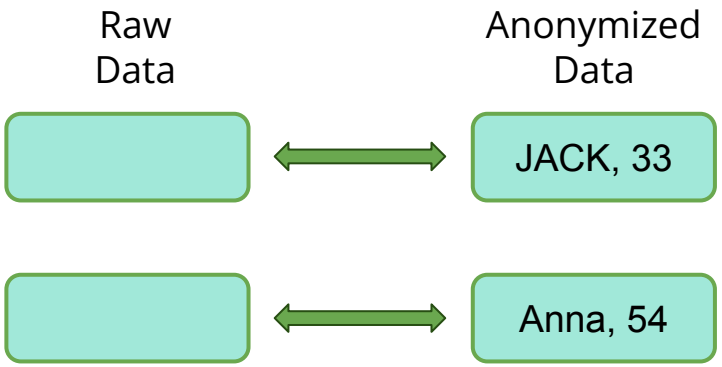
Data perturbation

- 加入雜訊或轉換資料空間
- 優點
  - 混淆性大幅增加
  - 統計運算需進行額外校正
- 缺點
  - 資料長度增加
  - 運算複雜度高



Synthetic data

- 以真實資料的模型產出完全的假資料
- 優點
  - 運算速度快
  - 不會被破解
  - 適用於功能性測試
- 缺點
  - 主觀性可能不信任資料可用性
  - 須額外設計適用的統計模型以接近真實資料
  - 無法測試真實的特殊 值



Data masking

- 資料半真半假，可調控安全性，兼顧執行效率和資料可利用性。
- 自建遮蔽函數，自行套用欄位（自由度最高）
  - 進行 In-Database Processing，以兼顧效率與安全性
- PostgreSQL Anonymizer
  - 處理個人資訊的 PostgreSQL extension
  - Static Masking
    - 對所有人產生同樣的遮蔽結果。
  - Dynamic Masking
    - 依使用者權限產生不同的遮蔽結果。
  - 規則整合於 DDL 宣告之中，容易與開發連動。
- PostgreSQL Faker
  - 製造完全的假資料

- 古○忠
- 新竹縣湖口鄉XXXXX
- C1ooooooX
  - 程式能判斷是否為合法的遮蔽資料
  - 程式能相容執行遮蔽資料與原始資料

PostgreSQL Anonymizer

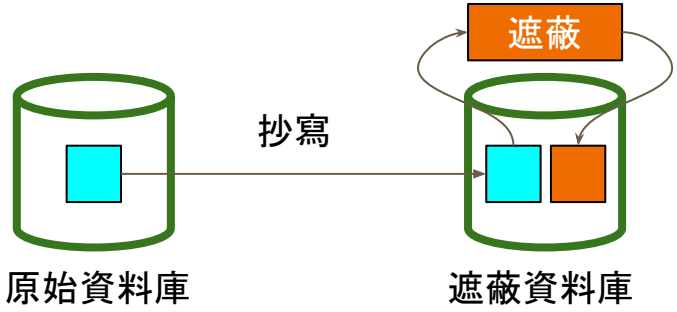
```
SELECT * FROM customer;

 id  |   full_name      |   birth    |    employer    | zipcode | fk_shop
-----+------------------+------------+----------------+---------+---------
 911 | Chuck Norris     | 1940-03-10 | Texas Rangers  | 75001   | 12
 112 | David Hasselhoff | 1952-07-17 | Baywatch       | 90001   | 423
```

Masked

```
SELECT * FROM customer;

 id  |  full_name  |   birth    |      employer       | zipcode | fk_shop
-----+-------------+------------+---------------------+---------+---------
 911 | jesse Kosel | 1940-03-10 | Marigold Properties | 62172   |      12
 312 | leolin Bose | 1952-07-17 | Inventure           | 20026   |     423
```

```
SECURITY LABEL FOR anon ON COLUMN
customer.full_name IS 'MASKED WITH
FUNCTION anon.fake_first_name() ||
'' '' || anon.fake_last_name()';

SECURITY LABEL FOR anon ON COLUMN
customer.employer IS 'MASKED WITH
FUNCTION anon.fake_company()';

SECURITY LABEL FOR anon ON COLUMN
customer.zipcode IS 'MASKED WITH
FUNCTION anon.random_zip()';
```

遮蔽流程



- 原始資料庫負擔低。
- 抄寫出來的資料可多次遮蔽。
- 遮蔽流程長。
- 遮蔽過程仍有外洩風險。

- 原始資料庫負擔高。
- 節省空間使用。
- 確保資料離開原始資料庫時，已具備一定的資料安全性。

PostgreSQ

原始資料庫/遮蔽資料庫為同一個資料庫？

- 不建議。
- 完全不需要抄寫，內容最即時。
- 權限控制必須要十分小心。

PostgreSQ

Database function and permission

- 儘可能使用資料庫的原生語言撰寫，執行效率高。
  - 使用 pl/pgsql
- SECURITY DEFINER
  - Function 定義是公開的。
  - Table 權限是可限制的。
  - 執行者僅能得到執行結果，無法取得 TABLE 內的參數。

```
REVOKE select on TABLE keytable FROM public;

GRANT execute on FUNCTION maskid to public;
```

```
CREATE FUNCTION maskid(id TEXT)
RETURNS TEXT AS $$
DECLARE
        mask_key TEXT;
        masked    TEXT;
BEGIN
        SELECT  value FROM keytable WHERE
key='mymaskkey' INTO mask_key;
        masked='XXXXXXXX';
        RETURN masked;
END;
$$  LANGUAGE plpgsql
     SECURITY DEFINER;
```

小結

- 消極:避免罰金。
- 積極:安全地使用資料→積極地使用資料→增加公司獲利。
- 進行 In-Database Processing, 以兼顧效率與安全性
- 依權限及使用情境套用遮蔽方法。
- 依系統資源決定遮蔽流程。
- 儘可能使用資料庫處理, 減少資料落地。
- 系統化資安政策, 完備資料治理, IT 只是輔助。

Thanks

THE DATABASE WHERE I SELECT