SQL 的五十道練習

> *簡介*

數據交點 | 郭耀仁 yaojenkuo@datainpoint.com

什麼是 SQL？

SQL（發音為 ess-que-ell 或 sequel）全名為 Structured Query Language，是一個能夠針對資料庫進行「資料操作」的語言。

SQL 按照使用目的可以再細分為三類

1. **資料操作語言**（Data Manipulation Language, DML）
2. 資料定義語言（Data Definition Language, DDL）
3. 資料控制語言（Data Control Language, DCL）

解釋定義中出現的名詞

- 什麼是**資料操作**？
- 什麼是資料庫？
- 什麼是資料庫管理系統？

「資料操作」涵蓋了 CRUD 這四個動詞：

- 創造 **C**reate
- 查詢 **R**ead
- 更新 **U**pdate
- 刪除 **D**elete

舉例來說，在使用任何的網頁或手機應用程式時，我們的滑鼠點擊與手勢觸控都會被轉換成 CRUD：

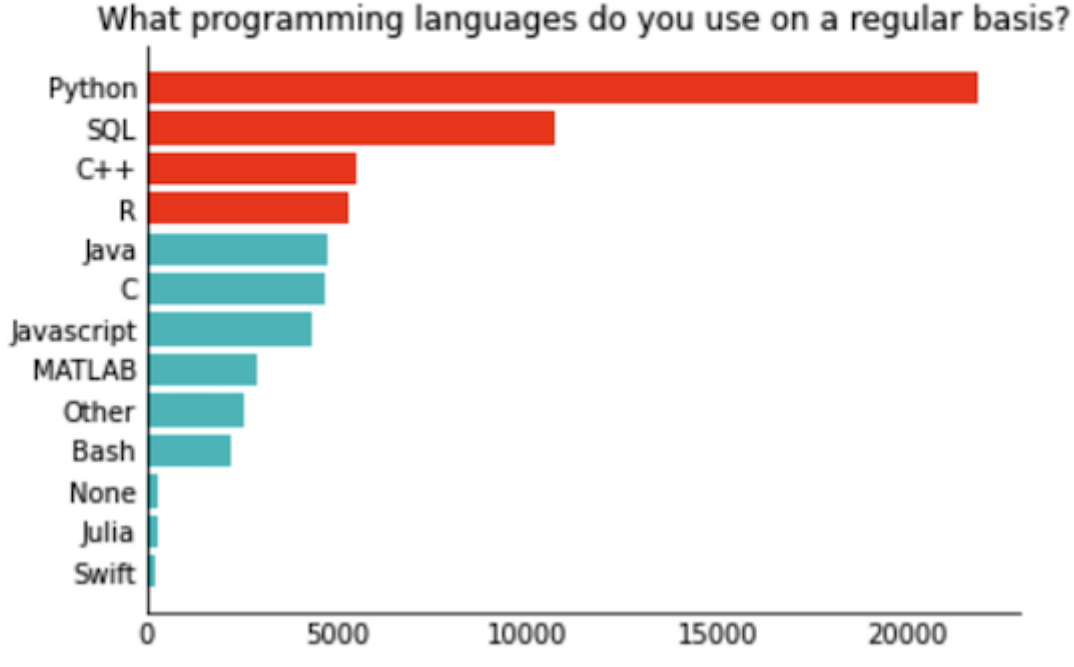
- 創造 **C**reate：發佈新的動態。
- 查詢 **R**ead：瀏覽追蹤對象的動態。
- 更新 **U**pdate：編輯先前發佈動態的內容。
- 刪除 **D**elete：撤掉先前所發佈的動態。

為什麼要學習 SQL？

SQL 是一個歷久彌新的語言，早於 1970 年代問世，50 年後今日仍然是資料科學與軟體開發從業者最重要的技能之一。

2021 Kaggle ML&DS Survey 中 SQL 在資料科學家日常使用語言中排名第二。

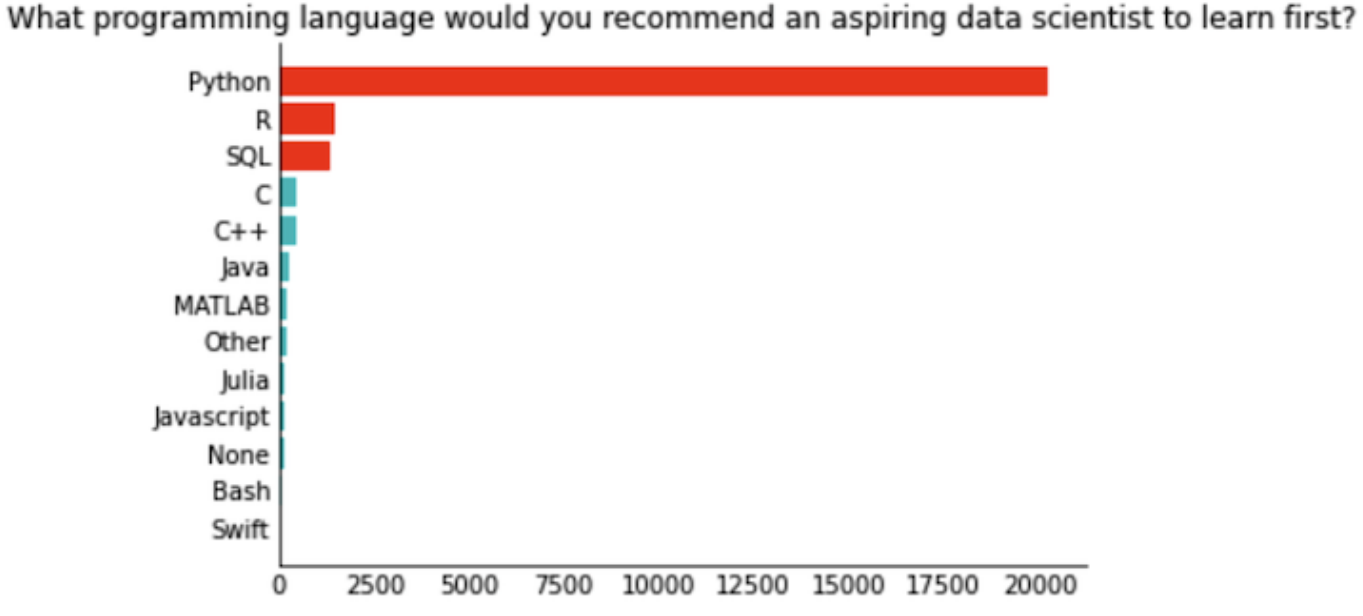
In [13]:
```
kaggle_survey.plot_survey_summary("Q7", n=4)
```



來源：https://www.kaggle.com/yaojenkuo/analyzing-kaggle-survey-in-a-more-structured-way

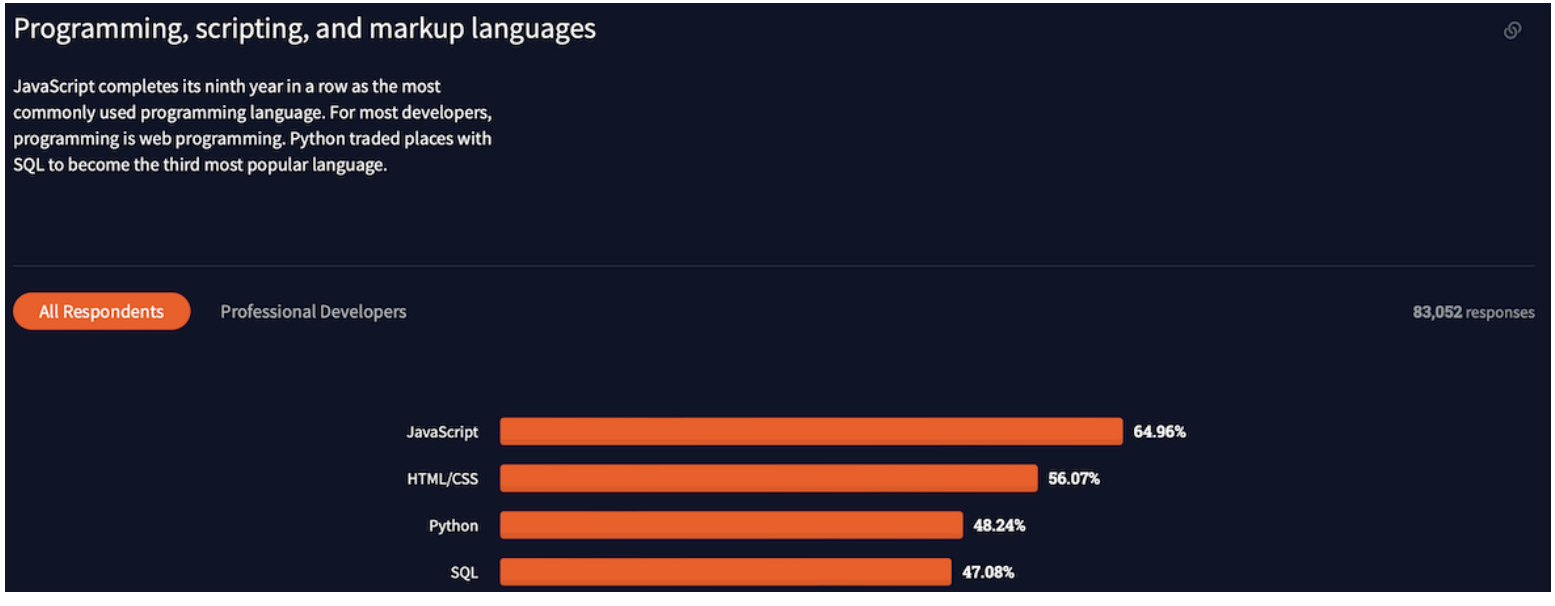
2021 Kaggle ML&DS Survey 中 SQL 在資料科學家推薦學習語言中排名第三。

```
In [14]: kaggle_survey.plot_survey_summary("Q8")
```



來源：https://www.kaggle.com/yaojenkuo/analyzing-kaggle-survey-in-a-more-structured-way

Stack Overflow 2021 Developer Survey 中 SQL 在軟體工程師受歡迎技術中排名第四。



來源：https://insights.stackoverflow.com/survey/2021

解釋定義中出現的名詞（續）

- 什麼是資料操作？
- 什麼是**資料庫**？
- 什麼是資料庫管理系統？

什麼是資料庫？

資料庫（Database）是儲存在電腦中的資料集合，我們可以透過撰寫 SQL 有效率地對資料庫中的數據進行「資料操作」。

什麼樣的資料集合能夠格被稱為資料庫呢？

具有兩個特徵的資料集合被稱為資料庫：

1. 觀測值必須具有屬性。
2. 資料集合必須具備有元資料（Metadata）。

不具有屬性的資料觀測值

In [3]:

```
show_without_attributes()
```

Out[3]:

```
array([[1, 'The Shawshank Redemption', 1994, 9.3, 'Frank Dara
bont', 142],
       [2, 'The Godfather', 1972, 9.2, 'Francis Ford Coppol
a', 175],
       [3, 'The Godfather: Part II', 1974, 9.0, 'Francis Ford
Coppola',
        202],
       [4, 'The Dark Knight', 2008, 9.0, 'Christopher Nolan',
152],
       [5, '12 Angry Men', 1957, 9.0, 'Sidney Lumet', 96]], d
type=object)
```

具有屬性的資料觀測值

In [5]:

```
show_with_attributes()
```

Out[5]:

	id	title	release_year	rating	director	runtime
0	1	The Shawshank Redemption	1994	9.3	Frank Darabont	142
1	2	The Godfather	1972	9.2	Francis Ford Coppola	175
2	3	The Godfather: Part II	1974	9.0	Francis Ford Coppola	202
3	4	The Dark Knight	2008	9.0	Christopher Nolan	152
4	5	12 Angry Men	1957	9.0	Sidney Lumet	96

元資料（Metadata）常見的解釋為「data about data」、「描述資料的資料」。

In [7]:

```
show_metadata()
```

Out[7]:

	cid	name	type	notnull	dflt_value	pk
0	0	id	INTEGER	0	None	1
1	1	title	TEXT	0	None	0
2	2	release_year	INTEGER	0	None	0
3	3	rating	REAL	0	None	0
4	4	director	TEXT	0	None	0
5	5	runtime	INTEGER	0	None	0

為什麼資料庫是重要的？

- 對於資料科學而言，資料庫是常見的資料來源。
- 對於軟體開發而言，資料庫是應用程式與系統不可或缺的元素。
- 對於生活應用而言，資料庫無所不在，小至手機的通話紀錄與通訊錄、大至銀行的存款資訊與交易資訊，背後都有資料庫在運作。

解釋定義中出現的名詞（續）

- 什麼是資料操作？
- 什麼是資料庫？
- 什麼是**資料庫管理系統**？

什麼是資料庫管理系統？

DBMS 全名為 DataBase Management System，透過資料庫管理系統，SQL 將能「自動化」且「規模化」地對資料庫進行 CRUD 的「資料操作」。

自動化的體現

不需要透過人工、客服就能夠發佈新的動態、瀏覽追蹤對象的動態、編輯先前發佈動態的內容與撤掉先前所發佈的動態。

規模化的體現

能夠讓成千上萬個使用者同時發佈新的動態、瀏覽追蹤對象的動態、編輯先前發佈動態的內容與撤掉先前所發佈的動態。

資料庫管理系統可以分為兩大類：

1. 關聯式資料庫管理系統（RDBMS, Relational Database Management System）
2. 非關聯式資料庫管理系統（NoSQL DBMS, Not only SQL Database Management System）

RDBMS vs. NoSQL DBMS

- 結構化 vs. 非結構化。
- 表格 vs. 文件、鍵值。
- 效率 vs. 彈性。

常見的關聯式資料庫管理系統

- 甲骨文（Oracle）的 Oracle Database
- 微軟（Microsoft）的 SQL Server
- 國際商業機器（IBM）的 DB2
- **開放原始碼的 SQLite**
- 開放原始碼的 MySQL
- 開放原始碼的 PostgreSQL

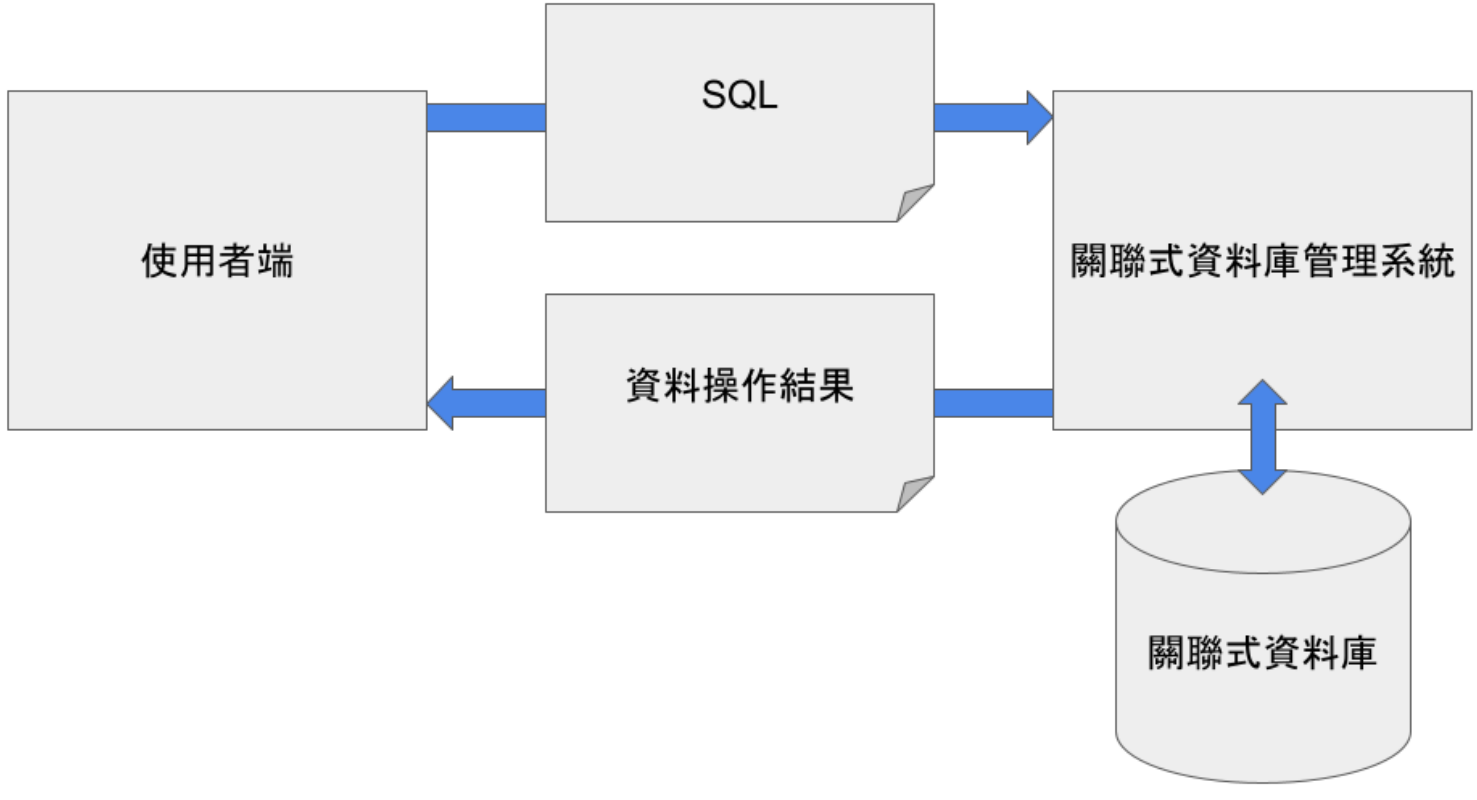
常見的非關聯式資料庫管理系統

- 甲骨文（Oracle）的 Coherence
- 微軟（Microsoft）的 Azure Cosmos DB
- 國際商業機器（IBM）的 Domino
- 開放原始碼的 MongoDB
- 開放原始碼的 Elasticsearch
- 開放原始碼的 Cassandra

課程使用開放原始碼的 SQLite

以常見關聯式資料庫管理系統都支援的標準 SQL 基本語法為主。

SQL、關聯式資料庫與關聯式資料庫管理系統的示意圖



重點統整

- SQL 全名為 Structured Query Language，是能夠針對資料庫進行「資料操作」的語言。
- 資料庫是儲存在電腦中的資料集合，具有兩個特徵：
    - 資料觀測值具有屬性。
    - 儲存有元資料（Metadata）。
- 透過資料庫管理系統，SQL 能**自動化**且**規模化**地對資料庫進行「資料操作」。