**TỔNG HỢP KIẾN THỨC**

CÔNG CỤ TRONG KHOA HỌC DỮ LIỆU

# 1. Ngôn ngữ chính trong Khoa học Dữ liệu

Trong lĩnh vực Khoa học Dữ liệu, có ba ngôn ngữ phổ biến được sử dụng là Python, R và SQL.

* Python là ngôn ngữ lập trình linh hoạt và dễ học, được sử dụng rộng rãi trong nhiều lĩnh vực, từ phát triển web đến khoa học dữ liệu.
* R là một ngôn ngữ và môi trường phát triển chủ yếu dùng trong phân tích thống kê và lập trình dữ liệu, có nhiều gói mở rộng mạnh mẽ và được ưa chuộng trong lĩnh vực khoa học dữ liệu và thống kê.
* SQL (Structured Query Language) là ngôn ngữ lập trình được sử dụng để tương tác với các hệ quản trị cơ sở dữ liệu quan hệ, cho phép truy vấn, cập nhật và quản lý dữ liệu.

***Tại sao Python lại được sử dụng nhiều nhất?***

* Dễ học và sử dụng: Python có cú pháp đơn giản và dễ hiểu, giúp người mới bắt đầu nhanh chóng tiếp cận và làm việc hiệu quả.
* Python có một cộng đồng lớn và năng động, cung cấp nhiều thư viện mạnh mẽ như:
  + Pandas: Xử lý và phân tích dữ liệu.
  + Numpy: Xử lý các mảng và hàm tính toán Đại số tuyến tính.
  + Matplotlib: Trực quan hóa dữ liệu, tạo biểu đồ, đồ thị.
  + Scikit-learn: Thư viện học máy (classification, regression, clustering, ...).
  + PyTorch: Xây dựng mô hình cho Deep Learning.
  + …
* Đa năng: Python không chỉ được sử dụng trong Khoa học Dữ liệu mà còn trong nhiều lĩnh vực khác như phát triển web, trí tuệ nhân tạo, và automation, tạo ra sự linh hoạt và tiện ích cho người sử dụng.

# 2. SQL và Loại Dữ Liệu

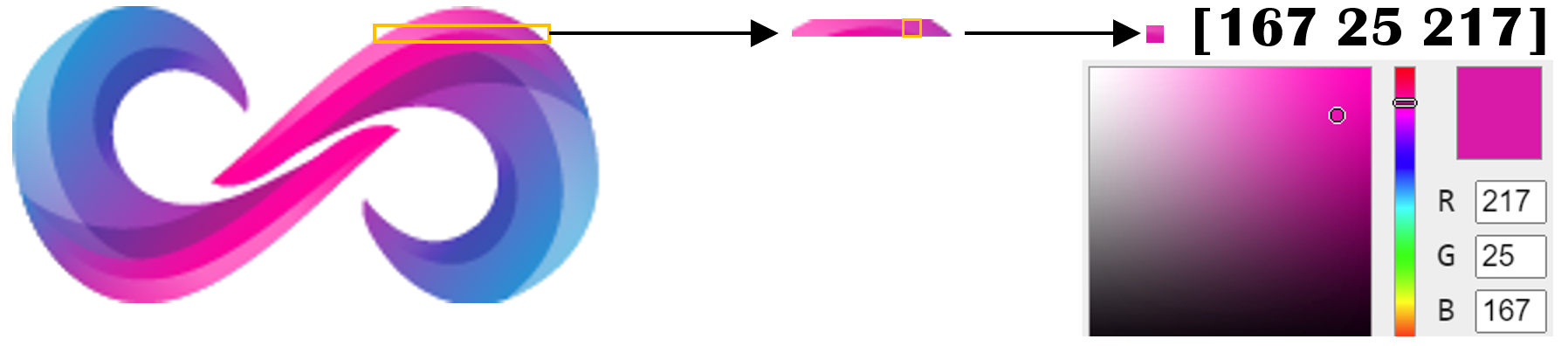
Trong lĩnh vực Khoa học Dữ liệu, SQL (Structured Query Language) đóng một vai trò quan trọng trong việc quản lý và tương tác với cơ sở dữ liệu quan hệ. Đây là một ngôn ngữ lập trình mạnh mẽ được sử dụng để truy vấn, cập nhật và quản lý dữ liệu trong các hệ thống cơ sở dữ liệu quan hệ như MySQL, PostgreSQL, SQLite, và Oracle.

Trong SQL, dữ liệu có thể được phân loại thành ba loại chính dựa trên cấu trúc và tổ chức của nó:

1. Dữ liệu có cấu trúc: Loại dữ liệu được tổ chức rõ ràng như chia thành các hàng và cột. Ví dụ các bảng Excel, các bảng trong cơ sở dữ liệu SQL, …

| **Bảng trong cơ sở dữ liệu SQL** | **Bảng Excel** |
| --- | --- |

2. Dữ liệu phi cấu trúc: Đây là loại dữ liệu không thể tổ chức, sắp xếp theo một cấu trúc cụ thể. Thay vào đó, nó thường bao gồm các đối tượng không cố định như văn bản, video, hình ảnh, hoặc dữ liệu âm thanh.



**Mã màu RGB là dạng dữ liệu không được tổ chức theo cách cố định**

3. Dữ liệu bán cấu trúc: Loại dữ liệu này không tuân theo một cấu trúc cụ thể nhưng vẫn chứa một số cấu trúc mô tả, cho phép tổ chức và lưu trữ một cách linh hoạt. Các định dạng dữ liệu như JSON, XML, hoặc tệp nén là các ví dụ điển hình của dữ liệu bán cấu trúc.

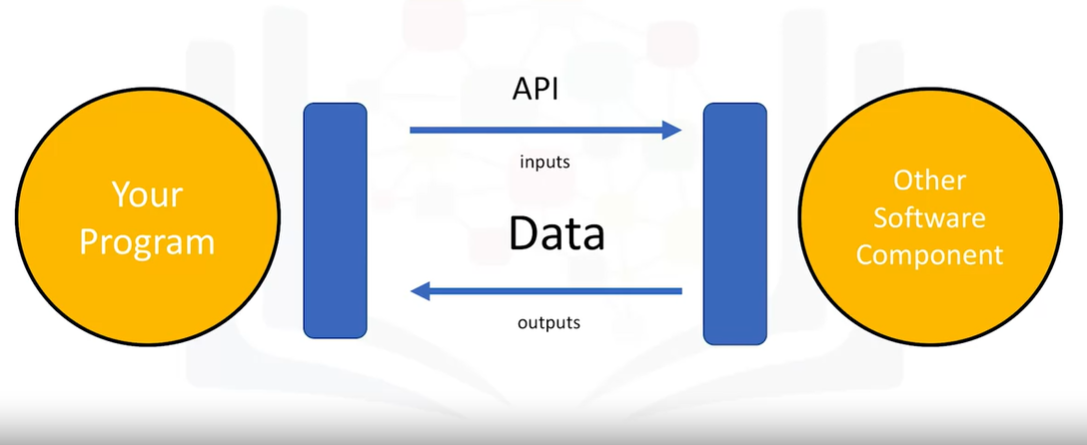


**Tệp JSON**

# 

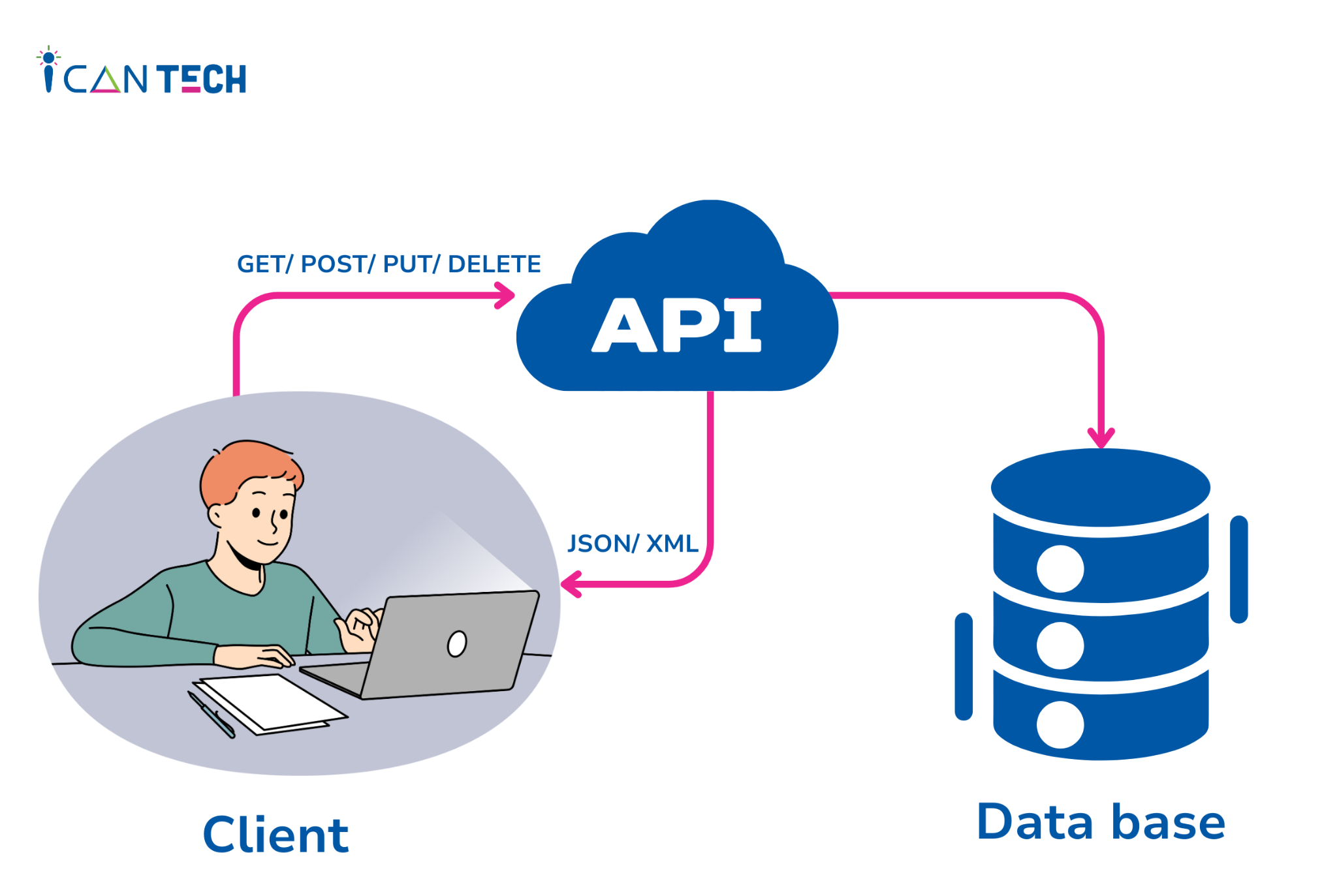
# 3. API (Application Programming Interface)

API cho phép các phần mềm giao tiếp với nhau, cung cấp các hàm để phần mềm khác có thể truy cập và sử dụng chức năng của hàm đó.



***REST API (REpresentational State Transfer API):***

* Cho phép các phần mềm giao tiếp với nhau thông qua Internet.
* Cơ chế hoạt động: Yêu cầu của khách hàng được gửi đến REST API, REST API sẽ xử lý yêu cầu đó và trả về kết quả cho khách hàng.



Ví dụ:

- cURL (Client URL) là một command line tool được sử dụng để kiểm thử các REST API một cách nhanh chóng.