Partition

**Partition là quá trình phân chia table thành các phân vùng nhỏ theo một quy tắc nào đó được gọi là partition function.**

1. Khi nào thì đánh partition trong 1 bảng

***Partition nên là lựa chọn cuối cùng khi muốn tối ưu hóa, tức là sau khi đã optimize câu query, sử dụng Index.***

Partition sẽ đem lại nhiều ý nghĩa nhất khi dữ liệu của bảng quá to hàng triệu bản ghi. Cụ thể hơn khi sử dụng Range Partition, khi dữ liệu quá lớn, muốn xóa các data cũ đi thì lựa chọn Range Partition theo time là vô cùng hợp lý.

Khi áp dụng Partition lên bất kỳ bảng nào thì nên nhớ đến một số hạn chế như: không thể sử dụng khóa ngoại, cẩn thận với khóa chính hay unique. Hãy đảm bảo điều kiện khi muốn sử dụng Partition.

**Nên cân nhắc đánh partition cho một bảng khi gặp các trường hợp sau:**

* **Bảng quá lớn**: Khi một bảng có hàng chục hoặc hàng trăm triệu bản ghi, việc truy vấn, xóa hoặc bảo trì sẽ trở nên rất chậm. Phân vùng giúp chia nhỏ dữ liệu, cho phép truy vấn chỉ cần quét một phần của bảng thay vì toàn bộ.
* **Thường xuyên thực hiện các thao tác theo dải**: Khi các truy vấn của bạn thường xuyên lọc dữ liệu theo một phạm vi ngày, tháng, năm hoặc một dải giá trị nào đó. Ví dụ: WHERE transaction\_date BETWEEN '2025-01-01' AND '2025-01-31'.
* **Thường xuyên xóa dữ liệu cũ**: Nếu ứng dụng của bạn cần xóa dữ liệu cũ (ví dụ: các bản ghi log sau 6 tháng), việc xóa một phân vùng sẽ nhanh hơn rất nhiều so với việc chạy một câu lệnh DELETE trên hàng triệu bản ghi.
* **Tăng hiệu suất truy vấn**: Nếu các truy vấn thường xuyên chỉ truy cập một phần nhỏ của bảng.

2. Đánh partition mang lại lợi ích gì cho truy vấn

Đánh partition (phân vùng) mang lại nhiều lợi ích cho truy vấn, chủ yếu tập trung vào việc **tối ưu hóa hiệu suất** và **quản lý dữ liệu hiệu quả hơn**. Lợi ích này đến từ việc chia một bảng lớn thành các phần nhỏ hơn, dễ dàng xử lý.

**Tối ưu hóa hiệu suất truy vấn**

* **Cải thiện tốc độ tìm kiếm**: Khi bạn truy vấn một bảng được phân vùng, cơ sở dữ liệu có thể loại bỏ các phân vùng không chứa dữ liệu bạn cần (gọi là **loại bỏ phân vùng**). Điều này có nghĩa là thay vì phải quét toàn bộ bảng khổng lồ, truy vấn chỉ cần quét một phần nhỏ hơn nhiều, giúp giảm đáng kể thời gian tìm kiếm.
* **Tăng hiệu quả của I/O (Input/Output)**: Bằng cách chỉ truy cập các phân vùng liên quan, cơ sở dữ liệu giảm thiểu lượng dữ liệu phải đọc từ đĩa. Điều này rất có lợi cho các truy vấn theo dải giá trị như ngày tháng.

**Ví dụ:** Một truy vấn tìm kiếm dữ liệu bán hàng của năm 2024 trên một bảng được phân vùng theo năm chỉ cần truy cập phân vùng chứa dữ liệu năm 2024 thay vì phải quét qua hàng tỷ bản ghi của nhiều năm khác.

**Quản lý dữ liệu hiệu quả hơn**

* **Dễ dàng bảo trì**: Các thao tác bảo trì như xóa dữ liệu cũ (ví dụ: các bản ghi lịch sử) trở nên nhanh hơn nhiều. Thay vì chạy một lệnh DELETE tốn kém trên hàng triệu bản ghi, bạn chỉ cần xóa toàn bộ một phân vùng, một thao tác hiệu quả hơn nhiều.
* **Giảm xung đột**: Trong môi trường có nhiều người dùng truy cập, việc phân vùng giúp giảm xung đột (contention) vì các truy vấn khác nhau có thể truy cập các phân vùng khác nhau mà không ảnh hưởng lẫn nhau.