1. Định nghĩa

* Partitioning (phân vùng) trong cơ sở dữ liệu là một kỹ thuật giúp chia một bảng lớn thành các phần nhỏ hơn gọi là partition (phân vùng), với mục tiêu cải thiện hiệu năng truy vấn, quản lý dữ liệu, và tối ưu hóa không gian lưu trữ.

1. Các loại parition phổ biến

* **Range Partitioning** : Phân chia dữ liệu dựa trên khoảng giá trị của một cột.
* **List Partitioning:** Phân chia dữ liệu dựa trên danh sách các giá trị cụ thể.

1. Các trường hợp sử dụng partition trong 1 bảng.

* **Bảng có dữ liệu lớn**: Khi một bảng có kích thước rất lớn (hàng triệu hoặc hàng tỷ bản ghi), việc phân vùng sẽ giúp giảm tải hệ thống trong việc xử lý truy vấn.
* **Quản lý dữ liệu theo thời gian**: Nếu bạn có bảng chứa dữ liệu theo ngày tháng, ví dụ như bảng logs hoặc bảng transactions, partition theo thời gian là một giải pháp tốt để dễ dàng quản lý và truy vấn dữ liệu theo từng giai đoạn.
* Nếu các truy vấn thường chỉ tìm kiếm dữ liệu trong một phạm vi nhỏ của bảng, việc đánh partition sẽ giúp truy vấn hiệu quả hơn.
* **Xóa dữ liệu cũ dễ dàng**: Phân vùng giúp bạn dễ dàng xóa dữ liệu cũ bằng cách xóa toàn bộ phân vùng mà không cần phải duyệt qua từng hàng để xóa.
* **Tối ưu hóa bảo trì**: Khi muốn thực hiện các tác vụ bảo trì chỉ trên một phần của dữ liệu thay vì toàn bộ bản.

1. Ví dụ :

CREATE TABLE transactions (

transaction\_id INT,

customer\_id INT,

transaction\_date DATE,

amount DECIMAL(10, 2)

)

PARTITION BY RANGE (YEAR(transaction\_date)) (

PARTITION p2020 VALUES LESS THAN (2021),

PARTITION p2021 VALUES LESS THAN (2022),

PARTITION p2022 VALUES LESS THAN (2023));

* Trong ví dụ trên ta chia nhỏ bảng traction thang thời gian của từng năm .
* Khi dữ liệu của bảng transactions tăng lên, việc truy vấn các giao dịch theo năm sẽ trở nên hiệu quả hơn, vì hệ thống chỉ cần truy vấn các partition tương ứng thay vì quét toàn bộ bảng.

1. Đánh partition mang lại lợi ích gì cho truy vấn?

* **Tăng hiệu suất truy vấn :** Khi thực hiện truy vấn, cơ sở dữ liệu chỉ cần quét qua những phân vùng có liên quan thay vì toàn bộ bảng dữ liệu. Điều này giúp giảm số lượng dữ liệu cần xử lý và tăng tốc độ truy vấn.
* **Quản lý dữ liệu dễ dàng hơn:**  Với bảng được phân vùng, việc xóa hoặc thêm dữ liệu theo từng phân vùng trở nên dễ dàng và nhanh chóng hơn. Bạn có thể ta toàn bộ một phân vùng mà không cần xóa từng hàng dữ liệu một cách thủ công.
* **Cân bằng tải** : Phân vùng dữ liệu cho phép phân phối dữ liệu qua nhiều ổ đĩa hoặc nút khác nhau (trong trường hợp hệ thống phân tán), giúp cân bằng tải và tránh tắc nghẽn khi nhiều truy vấn cùng thực hiện trên một tập dữ liệu lớn.
* **Tối ưu hóa bộ nhớ:** Khi chỉ truy vấn dữ liệu trong một số phân vùng nhất định, cơ sở dữ liệu có thể sử dụng bộ nhớ cache hiệu quả hơn, bởi vì nó chỉ cần nạp những phần dữ liệu có liên quan.
* **Tối ưu hóa bảo mật và phân quyền:** Phân vùng có thể giúp kiểm soát quyền truy cập dữ liệu tốt hơn. Bạn có thể áp dụng quyền truy cập khác nhau cho các phân vùng khác nhau, giúp hạn chế truy cập vào các phần dữ liệu không mong muốn.