

Họ và tên: Dương Nguyễn Trường Sang

Lớp: CNTT4

MSSV: 1250080158

Hệ quản trị lựa chọn: Apache Cassandra

1. Giới thiệu tổng quan và Lịch sử

Qua tìm hiểu từ nguồn NoSQL Databases – From Basics to Enterprise Use, em chọn Apache Cassandra để trình bày vì đây là hệ quản trị tiêu biểu cho dòng Wide-Column Store.

- Lịch sử ra đời: Cassandra được khởi xướng bởi các kỹ sư tại Facebook vào năm 2008 nhằm giải quyết bài toán lưu trữ và tìm kiếm tin nhắn cực kỳ đồ sộ tại thời điểm đó.
- Đơn vị quản lý: Sau khi được Facebook mở nguồn, dự án này đã trở thành một dự án trọng điểm (Top-level project) của Apache Software Foundation.

2. Mô hình lưu trữ và Ngôn ngữ thao tác

- Mô hình Wide-Column: Khác với các bảng dữ liệu SQL truyền thống có cấu trúc cứng nhắc, Cassandra cho phép dữ liệu có cấu trúc lỏng lẻo (linh hoạt). Điều này có nghĩa là trong cùng một bảng, mỗi hàng có thể có các cột khác nhau tùy theo nhu cầu thực tế, giúp việc phát triển ứng dụng rất nhanh chóng mà không cần tốn công chỉnh sửa lược đồ (schema) quá nhiều.
- Ngôn ngữ CQL: Hệ thống sử dụng ngôn ngữ CQL (Cassandra Query Language). Theo em thấy, cú pháp của nó cực kỳ giống SQL (ví dụ: `SELECT * FROM users`), điều này giúp những người đã biết SQL như sinh viên chúng em tiếp cận rất nhanh.

3. Mô phỏng cơ chế lưu trữ và truy vấn phân tán (Phần quan trọng)

Để hiểu rõ cách Cassandra vận hành, em xin mô phỏng qua hệ thống "Quản lý đơn hàng thương mại điện tử".

A. Cơ chế phân tán dữ liệu (Consistent Hashing)

Cassandra sử dụng một vòng tròn mã băm (Hash Ring) để quản lý các Node (máy chủ).

1. Gán mã băm (Hash): Giả sử mỗi đơn hàng có một mã Order_ID. Khi lưu vào, hệ thống sẽ băm mã này ra một con số (ví dụ: băm "ORD-001" ra số 450).
2. Định vị Node: Trên vòng tròn, Node 1 quản lý dải số 0-300, Node 2 quản lý 301-600. Vì mã băm là 450, đơn hàng này sẽ tự động được lưu vào Node 2.

B. Cơ chế truy vấn (Query Path)

Khi em thực hiện câu lệnh `SELECT * FROM Orders WHERE Order_ID = 'ORD-001'`:

- Em có thể gửi yêu cầu đến bất kỳ máy chủ nào trong cụm (ví dụ gửi đến Node 3). Node 3 lúc này đóng vai trò là "người điều phối".
- Node 3 tính toán mã băm của khóa tìm kiếm, biết được dữ liệu nằm ở Node 2. Nó sẽ liên hệ Node 2 để lấy dữ liệu về và trả cho em.
- Đặc điểm nổi bật: Nhờ cơ chế này, hệ thống không bao giờ bị "chết" hoàn toàn nếu một máy chủ hỏng, vì các máy khác vẫn có thể điều phối dữ liệu cho nhau (không có Master duy nhất).