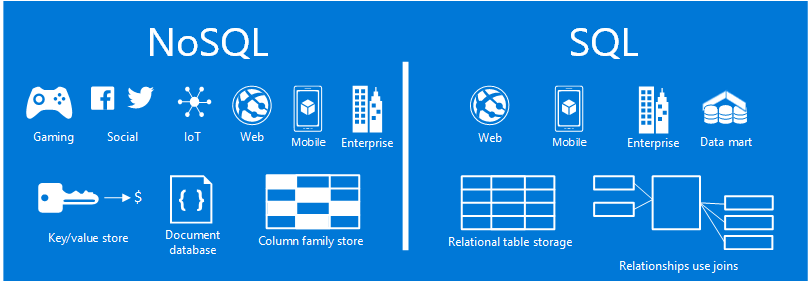
# Giới thiệu về NoSQL database

Cơ Sở dữ liệu quan hệ được thiết kế cho những mô hình cần đảm bảo tính chặt chẽ và dữ liệu không quá lớn, trong khi các dịch vụ mạng xã hội lại có một lượng lớn dữ liệu và được cập nhật liên tục do số lượng người dùng quá nhiều. Do đó cơ sở dữ liệu NOSQL sinh ra để giải quyết các vấn đề mà RDBMS (A relational database management system) đã bộc lộ những yếu kém như: tốc độ thực thi, khả năng lưu trữ, các nghiệp vụ phức tạp (như phân trang, đánh chỉ mục …) Nhờ vậy giải pháp sử dụng cơ sở dữ liệu NOSQL sẽ mang lại một chi phí thấp hơn nếu so sánh với RDBMS truyền thống.



## Định Nghĩa

NoSQL là một xu hướng cơ sở dữ liệu mà không dùng dữ liệu quan hệ để quản lý dữ liệu trong lĩnh vực phần mềm. NOSQL có nghĩa là Non-Relational – không ràng buộc. Tuy nhên, thuật ngữ đó ít phổ biến hơn và ngày nay người ta thường dịch thành Not Only SQL – không chỉ SQL.

NOSQL được xem như thế hệ database kế tiếp của RDBMS, là một thế hệ cơ sở dữ liệu Non-relational (không ràng buộc), distributed (phân tán), open source (mã nguồn mở), horizontal scalable (khả năng mở rộng theo chiều ngang), có độ chịu tải, lỗi cao.

## Một Số Thuật Ngữ Trong NoSQL

* Tính ràng buộc (Relational): Thuật ngữ để mô tả tính ràng buộc giữa các bảng trong cơ sở dữ liệu quan hệ (MySQL, SqlSever, PostgreSQL…).
* Không ràng buộc (Non-relational): Dữ liệu sẽ không có các ràng buộc giữa các bảng nữa, mà dữ liệu sẽ ở dạng Json hoặc Bson.
* Khả năng mở rộng (High Scalability): Khi hệ thống lớn lên, ta có thể bổ sung thêm các Nodes mới, các Sever mới để chia tải hoặc chia dữ liệu. Để hệ thống không bị quá tải.
* Khả năng mở rộng theo chiều dọc (Vertical Scalable / Scale Up): Là việc nâng cấp phần cứng hệ thống bằng việc nâng cấp RAM, hay bộ nhớ.
* Khả năng mở rộng theo chiều ngang (Horizontal Scalable / Scale-Out): Là việc bổ sung phần cứng tránh sự hoạt động quá tải của hệ thống.
* Phân tán dữ liệu (Distributed Data): Là việc mô tả dữ liệu được phân tán ở các địa điểm khác nhau.
* Triển khai linh hoạt (Deployment Flexibilitty): Dễ dàng mở rộng thêm các nodes (Severs) mà không ảnh hưởng đến hoạt động của hệ thống.
* Tính sẵn sàng (High Availability): Hệ thống sẽ không bị ảnh hưởng khi một node bị trục trặc.
* Nhất quán cuối (Eventual Consistency): Khi ta đưa một dữ liệu mới vào một node của hệ thống, dữ lệu sẽ được lan truyền sang các node khác của hệ thống và cuối cùng tất cả node sẽ được đồng bộ.
* Lưu trữ tốt (Durability).

**Khi làm việc với Nosql ta sẽ gặp một số khái niệm sau:**

* FIELDS: Tương đương với khái niệm Columns trong SQL
* Documents: Thay thế khái niệm Rows trong SQL. Đây cũng chính là khái niệm làm nên sự khác biệt giữa NOSQL và SQL, 1 document chưa số cột (fields) không cố định trong khi 1 row thì số cột (columns) là định sẵn trước.
* Collection: Tương đương với khái niệm table trong SQL. Một Collection là tập hợp các documents.
* Key-Value: Cặp khóa - giá trị được dùng để lưu trữ dữ liệu trong NOSQL
* Cursor: Tạm dịch là con trỏ, sử dụng cursor để lấy dữ liệu từ database.

Trong các hệ cơ sở dữ liệu quan hệ, các cột được định nghĩa theo bảng, còn với hệ cơ sở dữ liệu không ràng buộc, các cột được định nghĩa ở mỗi document. Bởi thế, các document quản lý gần như tất cả, các collection không cần quản lý chặt chẽ những gì đang xảy ra trong nó nữa.

|  |  |
| --- | --- |
| RDBMS | NOSQL |
| Columns  Row  Table  Schema | Fields  Documents  Collection  Free Schema |

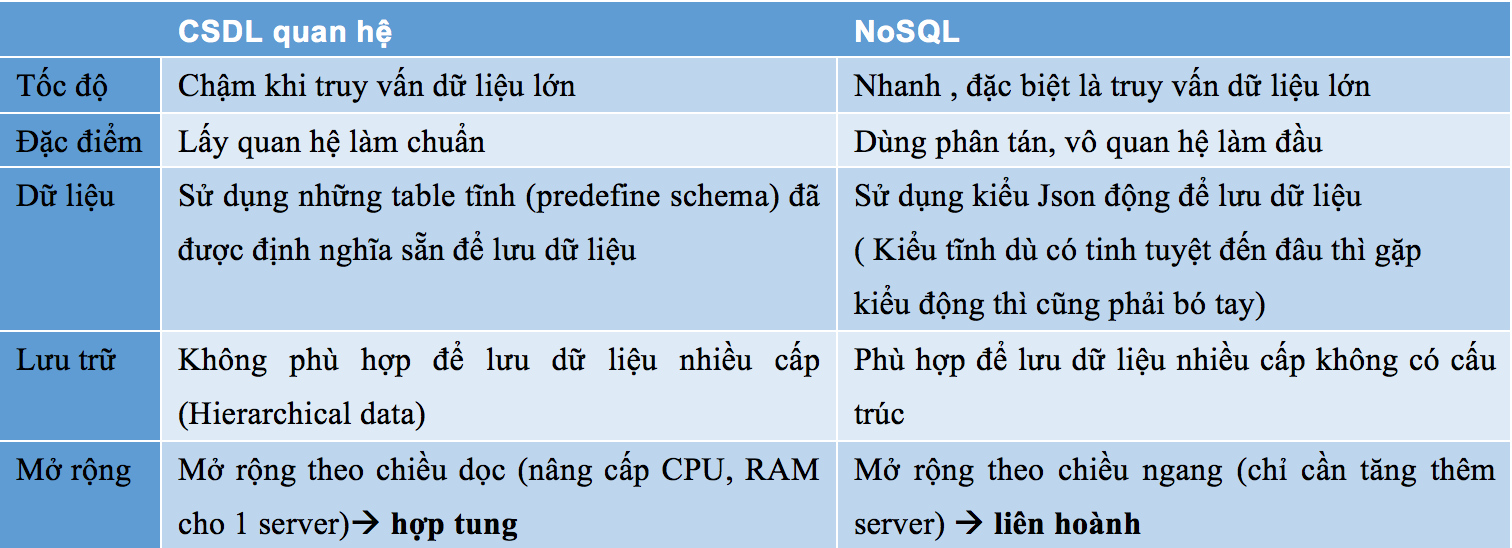
## Đặc Điểm

- NoSQL lưu trữ dữ liệu của mình theo dạng cặp giá trị “key – value”. Sử dụng số lượng lớn các node để lưu trữ thông tin.

- Chấp nhận dữ liệu bị trùng lặp do một số node sẽ lưu cùng thông tin giống nhau.

- Phi quan hệ – không có ràng buộc nào cho việc nhất quán dữ liệu.

- Có hiệu suất cao (High performance) và tính sẵn sàng cao (High availability).



Apache Cassandra là một trong các hệ thống cơ sở dữ liệu mã nguồn mở phổ biến nhất. Nó được thiết kế với mục đích xử lý lượng lớn dữ liệu được lưu trữ trong nhiều máy chủ phân bố khắp nơi song song đó cung cấp khả năng mở rộng cao và sẵn có mà không gặp một lỗi nào. Các hệ thống Cassandra có thể mở rộng nhiều trung tâm dữ liệu, cho phép độ trễ thấp cho tất cả các máy kết nối.