BÁO CÁO TÌM HIỂU HỆ THỐNG PHÂN TÁN

Môn học: Ứng dụng phân tán
Họ và tên: Dương Công Minh

MSSV: 22010009

Chủ đề: Tìm hiểu cơ chế hoạt động và triển khai cụm Database PostgreSQL và MongoDB

1. Giới thiệu chung về hệ thống phân tán

Hệ thống phân tán (Distributed System) là tập hợp nhiều máy tính độc lập, kết nối với nhau thông qua mạng và phối hợp để thực hiện một nhiệm vụ chung. Đối với cơ sở dữ liệu, hệ thống phân tán giúp:

- Tăng khả năng mở rộng (Scalability)
- Tăng độ sẵn sàng (High Availability)
- Giảm rủi ro khi xảy ra lỗi phần cứng

Trong báo cáo này, hai hệ quản trị cơ sở dữ liệu phổ biến là **PostgreSQL** và **MongoDB** sẽ được phân tích dưới góc độ hoạt động và cơ chế phân tán, đồng thời triển khai thử nghiệm cụm database bằng Docker.

---

2. PostgreSQL trong hệ thống phân tán

2.1 Tổng quan PostgreSQL

PostgreSQL là hệ quản trị cơ sở dữ liệu quan hệ (RDBMS), sử dụng mô hình Client–Server và tuân thủ đầy đủ các tính chất **ACID**. PostgreSQL thường được sử dụng cho các hệ thống nghiệp vụ yêu cầu tính nhất quán dữ liệu cao.

Mặc định, PostgreSQL hoạt động như một cơ sở dữ liệu tập trung. Để sử dụng trong hệ thống phân tán, PostgreSQL cần mở rộng thông qua các cơ chế sao chép và phân mảnh dữ liệu.

2.2 Cơ chế hoạt động phân tán của PostgreSQL

2.2.1 Streaming Replication

Streaming Replication là cơ chế sao chép dữ liệu ở mức WAL (Write-Ahead Logging):

- Một node **Primary (Master)** chịu trách nhiệm ghi dữ liệu
- Một hoặc nhiều node **Standby (Replica)** nhận WAL và đồng bộ dữ liệu

- Replica chỉ hỗ trợ đọc (read-only)

Mô hình này giúp tăng độ sẵn sàng và khả năng đọc song song.

2.2.2 Logical Replication

Logical Replication cho phép sao chép dữ liệu ở mức bảng (table-level), hỗ trợ:

- Sao chép chọn lọc dữ liệu
- Phù hợp với kiến trúc microservices

2.2.3 Sharding trong PostgreSQL

PostgreSQL không hỗ trợ sharding một cách native. Để phân mảnh dữ liệu, cần sử dụng:

- Citus Extension
- Hoặc xử lý sharding ở tầng ứng dụng

2.3 Đánh giá PostgreSQL theo CAP Theorem

- **Consistency (Nhất quán):** Cao
- **Availability (Sẵn sàng):** Cao (khi có replica)
- **Partition Tolerance:** Trung bình

PostgreSQL thiên về mô hình **CP** trong CAP Theorem.

2.4 Triển khai PostgreSQL Cluster (Thử nghiệm)

Trong bài thực hành, PostgreSQL được triển khai dưới dạng **Master–Slave (Streaming Replication)** bằng Docker.

Cấu trúc thư mục:

Postgres_Cluster/

└── docker-compose.yml

Cụm gồm:

- 1 Primary node
- 1 Replica node

Quá trình thử nghiệm:

1. Khởi động cluster bằng Docker Compose
2. Ghi dữ liệu tại Primary

3. Kiểm tra dữ liệu được đồng bộ sang Replica

Kết quả cho thấy dữ liệu được sao chép chính xác và hệ thống vẫn duy trì khả năng đọc khi Primary gặp sự cố.

---

3. MongoDB trong hệ thống phân tán

3.1 Tổng quan MongoDB

MongoDB là hệ quản trị cơ sở dữ liệu NoSQL, lưu trữ dữ liệu dưới dạng **document (BSON/JSON)**. MongoDB được thiết kế hướng tới hệ thống phân tán ngay từ đầu, phù hợp cho các ứng dụng lớn, dữ liệu linh hoạt.

3.2 Cơ chế hoạt động phân tán của MongoDB

3.2.1 Replica Set

Replica Set là cơ chế đảm bảo High Availability:

- 1 **Primary** node
- N **Secondary** node
- Tự động bầu chọn Primary mới khi node chính gặp sự cố (Failover)

Replica Set giúp MongoDB hoạt động ổn định trong môi trường phân tán.

3.2.2 Sharding

Sharding là điểm mạnh lớn nhất của MongoDB:

- Dữ liệu được phân mảnh theo **Shard Key**
- Hỗ trợ scale ngang rất tốt

Các thành phần chính:

- **Shard**: nơi lưu trữ dữ liệu
- **Config Server**: lưu metadata
- **mongos**: router điều phối truy vấn

3.3 Đánh giá MongoDB theo CAP Theorem

- **Consistency:** Eventual Consistency
- **Availability:** Rất cao
- **Partition Tolerance:** Cao

MongoDB thiên về mô hình **AP** trong CAP Theorem.

3.4 Triển khai MongoDB Cluster (Thử nghiệm)

Trong bài thực hành, MongoDB được triển khai dưới dạng **Replica Set (3 nodes)** bằng Docker.

Cấu trúc thư mục:

MongoDB_Cluster/

└── docker-compose.yml

Quá trình thử nghiệm:

1. Khởi động 3 node MongoDB
2. Khởi tạo Replica Set
3. Thực hiện ghi dữ liệu
4. Tắt Primary và quan sát quá trình failover

Kết quả cho thấy MongoDB tự động bầu chọn Primary mới và hệ thống vẫn hoạt động bình thường.

---

4. So sánh PostgreSQL và MongoDB trong hệ phân tán

Tiêu chí	PostgreSQL	MongoDB
Kiểu dữ liệu	Quan hệ	Document
Hỗ trợ phân tán	Qua extension	Native
Sharding	Phức tạp	Đơn giản
Consistency	Strong	Eventual
Khả năng scale ghi	Hạn chế	Tốt

---

5. Kết luận

PostgreSQL và MongoDB đều có thể được sử dụng trong hệ thống phân tán nhưng phù hợp với các bài toán khác nhau:

- PostgreSQL phù hợp cho các hệ thống nghiệp vụ yêu cầu tính nhất quán cao
- MongoDB phù hợp cho hệ thống lớn, dữ liệu linh hoạt và cần mở rộng nhanh

Việc triển khai thử nghiệm cho thấy MongoDB có lợi thế rõ rệt về khả năng phân tán native, trong khi PostgreSQL mạnh về độ tin cậy và tính toàn vẹn dữ liệu.

Tài liệu tham khảo

- PostgreSQL Documentation
- MongoDB Documentation
- CAP Theorem