BÀI TẬP TRÊN LỚP
MÔN HỌC: HỆ PHÂN TÁN

CHƯƠNG 5: NHÂN BẢN VÀ NHẤT QUÁN DỮ LIỆU

HỌ TÊN SV: TRẦN TRUNG PHONG MSSV: 20210676

MÃ LỚP: 149501 MÃ HỌC PHẦN: IT4611

Câu hỏi 1: Tại sao phải thực hiện nhân bản dữ liệu?

Trả lời:

Phải thực hiện nhân bản dữ liệu để:

- Tăng tín tin cậy, tính sẵn sàng cho hệ thống: Trong quá trình đọc hoặc ghi dữ liệu, nếu dữ liệu đó bị lỗi hay vì một nguyên nhân nào đó mà không thể dùng được, ta có thể dùng ngay bản sao dữ liệu đó để hệ thống không phải dừng lại và tránh được tình trạng sử dụng các dữ liệu không chính xác.
- Tăng hiệu năng của hệ thống: cần tăng quy mô của hệ thống cả về số lượng và phạm vi địa lý.

Câu hỏi 2: 1. Xét một kho dữ liệu phân tán với 5 tiến trình độc lập P1, P2, P3, P4, và P5. Mỗi tiến trình chỉ tác động lên được bản sao cục bộ riêng của mình. Các bản sao cục bộ kết nối thành kho dữ liệu phân tán. Xét các tiến trình chỉ tương tác (ghi, đọc) lên thành phần dữ liệu x ở bản sao cục bộ riêng của mình. Hoạt động của mô hình ở các thời điểm t tương ứng, các thao tác được thực hiện như sau:

t1: P1 ghi giá trị a

t2: P3 đọc giá trị a

t3: P2 ghi giá trị b và P3 ghi giá trị c

t4: P5 đọc được giá trị b

t5: P4 và P5 đều đọc được giá trị a

t6: P4 đọc được giá trị b

t7: P4 và P5 đọc được giá trị c (biết rằng ti < ti+1 với i=(0..6)

Câu hỏi:

a) Mô hình trên có thoả mãn nhất quán nhân quả không? Giải thích.

b) Mô hình trên có thoả mãn nhất quán tuần tự không? Giải thích.

Trả lời:

Bảng minh họa:

	t1	t2	t3	t4	t5	t6	t7
P1	W(x)a						
P2			W(x)b				
P3		R(x)a	W(x)c				
P4					R(x)a	R(x)b	R(x)c
P5				R(x)b	R(x)a		R(x)c

a) P1 ghi a và P3 ghi c có quan hệ nhân quả, P4 và P5 đều đọc được a và c theo đúng thứ tự nên bảng minh họa trên nhất quán nhân quả.

b) P4 đọc được a trước b, tuy nhiên P5 lại ngược lại là đọc được b trước a nên bảng minh họa trên không nhất quán tuần tự.

Câu hỏi 3: Conit là gì? Nếu đặt kích thước Conit quá lớn thì sẽ gây ra vấn đề gì? Tương tự với kích thước Conit quá nhỏ?

Trả lời:

- Conit (Consistency Unit) là đơn vị nhất quán dùng để xác định tính nhất quán.
- Nếu kích thước conit quá lớn thì sẽ gây ra vấn đề các bản sao sớm rơi vào tình trạng không nhất quán, Middleware liên tục phải hoạt động để tránh các bản sao liên tục rơi vào trạng thái không nhất quán.
- Nếu kích thước conit quá nhỏ thì sẽ gây ra vấn đề số lượng conit nhiều nên sẽ rất phức tạp trong việc quản lý.

Câu hỏi 4: Tại sao *nhất quán nhân quả* có tính nhất quán yếu hơn nhất quán tuần tự? Cho ví dụ để làm rõ điều này.

Trả lời:

- Thống nhất nhân quả có tính thống nhất yếu hơn thống nhất tuần tự bởi vì:
    - Thống nhất nhân quả: Chỉ đảm bảo thứ tự cho các sự kiện có quan hệ nhân quả.
    - Thống nhất tuần tự: Kết quả luôn như nhau nếu thứ tự các thao tác cục bộ của một tiến trình không thay đổi trong thứ tự thực hiện trên kho dữ liệu.
- Ví dụ: Nếu tiến trình P1 thực hiện W(x)a, P2 thực hiện W(x)b. Thì việc tiến trình P3 và P4 thực hiện R(x)a trước hay R(x) trước đều được đối với nhân quả còn với thống nhất tuần tự thì không, chỉ có thể R(x)b trước rời tới R(x)a theo thời gian.

Câu hỏi 5: Vấn đề của mô hình Eventual Consistency là gì? Từ đó đưa ra định nghĩa mô hình nhất quán hướng client.

Trả lời:

Vấn đề của mô hình Eventual Consistency là khi client thực hiện cập nhật tại một bản sao và trong khoảng thời gian ngắn lại chuyển sang bản sao khác. Trường hợp client là thiết bị di động, việc thực hiện yêu cầu gặp khó khăn hơn nên cần đảm bảo các bản sao luôn nhất quán khi mà client thay đổi về vị trí vật lý.

➔ Mô hình nhất quán hướng client đảm bảo nhất quán cho các truy cập của 1 client đơn vào kho dữ liệu nhưng không đảm bảo nhất quán cho các client khác nhau.

Câu hỏi 6: Một ngân hàng quyết định sử dụng dịch vụ CDN (Content Delivery Network) của một công ty mới khởi nghiệp cung cấp.

a) Với bước đặt máy chủ, công ty chọn thuật toán chọn đặt các máy chủ bản sao (replica) dựa trên khoảng cách với các chi nhánh ngân hàng. Hãy đề xuất thuật toán chọn đặt k replica với N vị trí có thể đặt máy chủ. Biết rằng đây là thuật toán dựa trên khoảng cách và công ty biết trước các vị trí các chi nhánh ngân hàng.

b) Với thuật toán để quản lý nội dung dữ liệu ở các replica, công ty quyết định chọn thuật toán dựa trên bản sao kích hoạt bởi server (server-initiated replicas). Hãy mô tả cơ chế đó với việc xem xét một đơn vị dữ liệu X là thông tin tài khoản một người dùng cùng với 2 ngưỡng là del(X) và rep(X).

c) Liên quan đến giao thức đảm bảo nhất quán, công ty quyết định chọn giao thức ghi trên các bản sao (replicated write), tuy nhiên công ty băn khoăn giữa giao thức nhân bản tích cực và giao thức nhân bản dựa trên túc số. Bạn hãy giúp công ty lựa chọn giao thức phù hợp bằng việc so sánh 2 giao thức trên với việc chỉ ra ưu nhược điểm của chúng.

Trả lời:

a) Đề xuất thuật toán: Dựa vào khoảng cách tới các client giúp giảm tối đa khoảng cách trung bình từ các bản sao tới các bản sao khác và client. Sau khi đã lựa chọn (k-1) bản sao sẽ xác định bản sao thứ k để khoảng cách trung bình từ bản sao đến các client là nhỏ nhất, khoảng cách có thể được đo bằng thời gian hoặc các tiêu chí khác.

b) Mô tả cơ chế: Mỗi server theo dõi số lượng và nguồn gốc các truy cập vào từng phần tử dữ liệu. Với ngưỡng dell(X): khi số lượng truy cập vào phần tử dữ liệu X xuống dưới giá trị ngưỡng dell(X) thì phần tử đó sẽ bị xóa khỏi server, nếu làm như vậy số lượng các bản sao giảm xuống dẫn đến tăng tải cho các server khác, từ đó cần có cơ chế đảm bảo tồn tại 1 bản sao của các phần tử dữ liệu.

c)

	Nhân bản tích cực	Nhân bản dựa trên túc số
Ý tưởng	Có một tiến trình quảng bá thao tác cập nhật đến tất cả các bản sao khác mỗi khi có thay đổi, khi cần đọc dữ liệu chỉ cần đọc trên bản sao cục bộ	Chỉ cập nhật một số bản sao thay vì tất cả
Ưu điểm	Nhất quán mạnh	Giảm lãng phí
Nhược điểm	Gây lãng phí khi cập nhật tất cả bản sao kể cả những bản sao không sử dụng đến dữ liệu được cập nhật	Cơ chế Write-Quorum và Read-Quorum ảnh hưởng đến hiệu năng của hệ thống

Câu hỏi 7: Liên quan đến các mô hình nhất quán hướng dữ liệu và các mô hình nhất quán hướng người dùng:

a. Giải thích vắn tắt ý tưởng của 2 loại mô hình nhất quán hướng dữ liệu trên.

b. Một công ty startup mới mở chuyên triển khai thương mại hóa dịch vụ CDN (Content Delivery Network) cho 2 loại hình dịch vụ là thư điện tử và WWW. Để đảm bảo nhất quán dữ liệu cho 2 loại dịch vụ đó thì tầng middleware sẽ áp dụng mô hình nhất quán dữ liệu nào (ở câu a) cho mỗi loại dịch vụ trên? Giải thích.

c. Công ty đó triển khai 3000 server bản sao vật lý và chọn hình thức nhân bản dữ liệu dựa trên túc số (quorum) với Nw = 1600 và Nr = 1100. Vậy hệ thống đó sẽ tránh được xung đột đọc-ghi và xung đột ghi-ghi hay không? Giải thích.

Trả lời:

a) - Mô hình nhất quán hướng dữ liệu*:* kết quả của các hoạt động đọc và ghi được thực hiện trên dữ liệu có thể được sao chép sang các máy chủ bản sao khác nhau để đảm bảo dữ liệu được nhất quán giữa các bản sao.
   - Mô hình nhất quán hướng người dùng*:* mô hình này không xử lý các cập nhật ngay nhưng sẽ đảm bảo nhất quán cho 1 client đơn khi truy cập các bản sao khác nhau từ các vị trí khác nhau.

b) - Dịch vụ WWW: chủ yếu các tiến trình thực hiện đọc, rất ít tiến trình thực hiện cập nhật vì thế có thể dùng mô hình nhất quán hướng dữ liệu.
   - Dịch vụ thư điện tử: người dùng thực hiện rất nhiều thao tác ghi, vì thế nên dùng mô hình nhất quán hướng người dùng.

c) $Nr + Nw = 2700 < 3000$

   → Trong 1100 bản sao được đọc có thể không có bản sao nào được cập nhật nên có thể xảy ra xung đột đọc-ghi.

   $Nw = 1600 > 3000/2$

→ Tránh được xung đột ghi-ghi vì khi cập nhật 1600 bản sao chắc chắn sẽ phát hiện được thay đổi từ ít nhất 1600*2-3000 = 200 bản sao khác.