Aerospike

Kiến trúc tổng quan:

* Kiến trúc Client-server
* Các layer bên client: bao gồm các thư viện mã nguồn mở, APIs, track nodes (các nút để theo dõi dữ liệu nằm ở đâu trong cụm).
* Các layer bên server:
  + Real-time Transaction Engine: Đảm bảo tính nhất quán về khóa chính (Primary Key Consistency); tối ưu hóa CPU, xử lý đa nhân và đa luồng; tối ưu hóa cấp phát bộ nhớ; bộ lập lịch và ưu tiên cho các giao tác.
  + Data distribution: phân vùng bộ nhớ; mapping dữ liệu. Gồm có các module:
    - Cluster Management: Dùng thuật toán Paxos-based gossip-voting process để theo dõi dữ liệu, xác định những node nào là 1 phần của cluster; đồng bộ hóa “heartbeat” để quản lý việc giao tiếp giữa các cluster.
    - Data Rebalancer: Quản lý việc “di cư” dữ liệu và tái cân bằng dữ liệu.
    - Transaciton Processing Module: Quản lý việc đồng bộ hóa các bản sao dữ liệu.
    - Resolution of Duplicate Data: Xử lý mâu thuẫn giữa các bản sao.
    - Ngoài ra, có thể cài đặt thêm Cross-Datacenter Replication để quản lý các bản sao dữ liệu không đồng bộ từ nhiều hệ thống khác nhau.
  + Data Storage: Quản lý index, dữ liệu, …

Mô hình dữ liệu (Data model):

* Không cần định nghĩa trước lược đồ cơ sở dữ liệu.
* Dữ liệu được nhóm với nhau qua namespaces, hay có thể hiểu là database trong RDBMS system.
* Trong namespace, dữ liệu được chia thành các sets và records.
* Mỗi record gồm: Key và bins. Key là khóa duy nhất, đi kèm với khóa này là bins lưu các giá trị. Trong bins có thể bao gồm nhiều trường dữ liệu với các kiểu khác nhau: integers, doubles, strings, byte arrays, BLOBs, maps, lists, GeoJSON, HyperLogLog.
* Các record trong cùng set cũng có thể có cấu trúc bins khác nhau.
* Bên cạnh việc truy vấn bằng cách tra khóa chính, Aerospike còn hỗ trợ parallel scans, secondary indexing, geospatial indexing and filtering, user-defined functions.

Quản lý lưu trữ (Storage Management):

Cấu hình Homogeneous:

* All DRAM: Tất cả chỉ mục (indexes) và dữ liệu được lưu trong DRAM (dynamic RAM). Dữ liệu được trong disk chỉ để dự phòng.
* All PMEM: Tất cả chỉ mục và dữ liệu được lưu trong PMEM (persistent memory).
* All FLASH: Tất cả chỉ mục và dữ liệu được lưu trong Flash.

Cấu hình Hybrid memory:

* Lưu dữ liệu trong Flash, chỉ mục trong DRAM hoặc PMEM.
* Giảm thiểu tối đa độ trễ khi đọc do disk I/O.

Phân bổ dữ liệu (Data distribution):

* Mỗi namespaces được chia thành 4096 phân vùng (partitions), được phân tán đều khắp các cluster node.
* Phân bổ record vào các phân vùng: Sử dụng thuật toán hash RIPEMD, Aerospike hash khóa chính của một record thành một giá trị (hash value) có 160 bit. Sau đó, Aerospike sử dụng 12 bit của giá trị này để xác định phân vùng mà record này thuộc về.
* Phân bổ phân vùng vào các nodes:
  + Sử dụng một thuật toán hash ngẫu nhiên để đảm bảo các phân vùng được phân đều vào các nodes. Ngoài ra, tất cả node là cặp, khi một node gặp sự cố sẽ không ảnh hưởng tới toàn bộ database.
  + Khi một node mới được thêm hoặc xóa bỏ khỏi cluster, một cluster mới sẽ được tạo, và các phân vùng sẽ được chia đều vào các node.
* Tất cả dữ liệu được phân bổ đều và ngẫu nhiên vào các node, nên sẽ không tạo thành các hotspots, bottlenecks (một node phải chịu truy vấn nhiều hơn các node khác), giúp cải thiện hiệu suất và giảm thiểu rủi ro.

Sao chép và đồng bộ hóa dữ liệu (Data Replication and Synchronization)

Để đảm bảo tính khả dụng và độ tin cậy cao, Aerospike sẽ tạo các bản sao của các phân vùng trên một hoặc nhiều node. Mỗi node có thể chứa data chính (data master) và bản sao của các data master của các node khác.

Người dùng có thể cấu hình chỉ số mở rộng (replication factor). Chỉ số này càng cao, càng nhiều bản sao sẽ được tạo ra, và độ tin cậy sẽ càng cao, tuy nhiên hiệu suất sẽ thấp hơn do mỗi truy vấn ghi phải được thực hiện trên tất cả bản sao. Ngoài ra, chỉ số này không được vượt quá số node có trong cluster.

Ví dụ:

* Khi chỉ số mở rộng = 1 (không tạo bản sao): Giả sử trong cluster có 4 node, mỗi node sẽ chứa 1024 phân vùng ngẫu nhiên trong tổng cộng 4096 phân vùng. Tất cả node đều sẽ là data master.
* Khi chỉ số mở rộng =2 (một bản chính, một bản sao): Lúc này, mỗi node sẽ chứa 1024 phân vùng là data master, và 1024 phân vùng là một phần bản sao (replica) của các node còn lại.

Đặc điểm

At scale: Phù hợp với tất cả quy mô, từ terabytes tới cả petabytes dữ liệu.

All the time: Tính khả dụng và tính nhất quán của dữ liệu cao.

Multi-site clustering (chia thành nhiều cụm địa điểm): Một node cũng như là một cụm, có thể được phân tán ở nhiều địa điểm. Một địa điểm có thể là 1 rack trong 1 datacenter, hoặc là cả datacenter, một availability zone trong cloud region, hay cả cloud region.

* Đảm bảo 100% tính khả dụng (availability), cho dù có 1 địa điểm xảy ra sự cố.
* Khi kết hợp với tính năng Strong Consistency, cơ sơ dữ liệu đảm bảo không mất mát dữ liệu với tất cả câu lệnh write.
* Có một khoảng thời gian bị trễ nhỏ khi truy vấn khoảng từ 2 đến 100 milliseconds, tùy thuộc vào khoảng cách giữa các địa điểm.

Strong Consistency: Là một cơ sở dữ liệu phân tán, Aerospike đảm bảo tất cả truy vấn đọc và ghi đều được thực thi theo lịch biểu khả tuần tự. Ngoài ra, Aerospike có hỗ trợ cài đặt giảm tính nhất quán của dữ liệu để đổi lại hiệu suất và tính khả dụng cao hơn nếu người dùng muốn.