MongoDB là một open-source document database ( cơ sở dữ liệu hướng tài liệu mã nguồn mở ) cung cấp high performance ( hiệu năng cao ), high availability ( tính sẵn sàng cao ), và automatic scaling ( mở rộng quy mô tự động ). ( <https://docs.mongodb.com/manual/introduction/> )

MongoDB là cơ sở dữ liệu cho các ứng dụng ngày nay, cho phép:

+ Tận dụng dữ liệu và công nghệ để tối đa hóa lợi thế cạnh tranh

+ GIảm thiểu rủi ro trong việc triển khai các nhiệm vụ quan trọng ( mission-critical deployments )

+ Tăng tốc time-to-value

+ Tổng chi phí của người sở hữu thấp hơn đáng kể

Với MongoDB, bạn có thể xây dựng các ứng dụng mà không thể với cơ sở dữ liệu quan hệ truyền thống vơi:

+ Fast, Iterative Development

+ Flexible Data Model

+ Multi-Datacenter Scalability

+ Integrated Feature Set

+ Lower TCO

+ Long-term Commitment

Document database ( Cơ sở dữ hiêu hướng tài liệu ): Một record ( bản ghi ) trong MongoDB là một document với cấu trúc dữ liệu bao gồm những cặp field và value. MongoDB documents thì tương tự như các đối tượng JSON ( JSON objects ). Những values của fields có thể bao gồm những documents khác, các arrays, hoặc arrays of documents.



Lợi thế của việc sử dụng documents:

+ Documents (i.e: Objects ) tương tác với những kiểu dữ liệu tự nhiên trong nhiều ngôn ngữ lập trình.

+ Embedded documents ( tài liệu nhúng ) và arrays ( mảng ) giảm sự cần thiết của các phép join tốn kém.

+ Dynamic schema ( Giản đồ động ) hỗ trợ fluent polymorphism ( đa hình linh hoạt ).

Cơ bản về sử dụng MongoDB:

Databases: trong Mongo, databases giữ các collections of documents ( tập hợp của các tài liệu ).

Chọn một database để sử dụng trong mongo shell: use <db>

Ví dụ: use myDB

Tạo một database: nếu database chưa tồn tại, MongoDB sẽ tạo database khi bạn chứa data lần đầu tiên cho database đó. Như vậy bạn có thể chuyển đến một non-existent database ( chưa tồn tại ) và hiện thực những thao tác sau trong mongo shell:

Ví dụ: use myNewDB

db.myNewCollection1. insert( { x: 1 } )

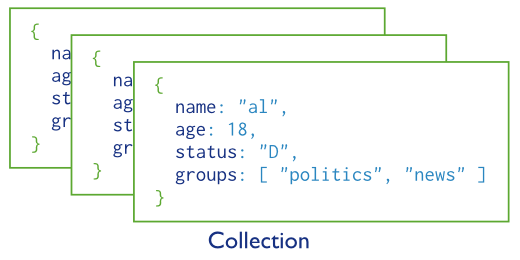
Thao tác insert() tạo cả database myNewDB và collection myNewCollection1 cho bạn nếu chúng chưa có.

Collections: MongoDB chứa documents trong collections. MongoDB cũng tạo collection khi bạn chứa dữ liệu đầu tiên vào collection đó.

Ví dụ: db.myNewCollection2. insert( { x: 1 } )

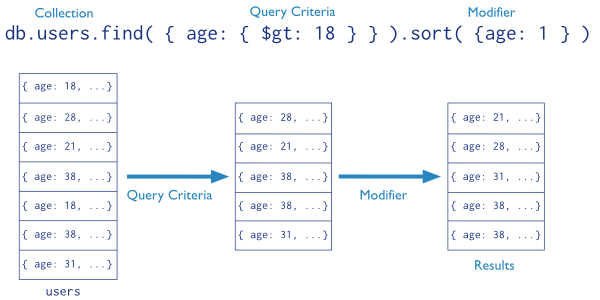
Db.myNewCollection3.createIndex( { y: 1} )

Cả 2 thao tác insert() và createIndex() đều tạo collection nếu chúng chưa tồn tại trước đó.



Basic MongoDB operations:

Query: Trong MongoDB một query chỉ đến một tập hợp các documents được đặc tả.

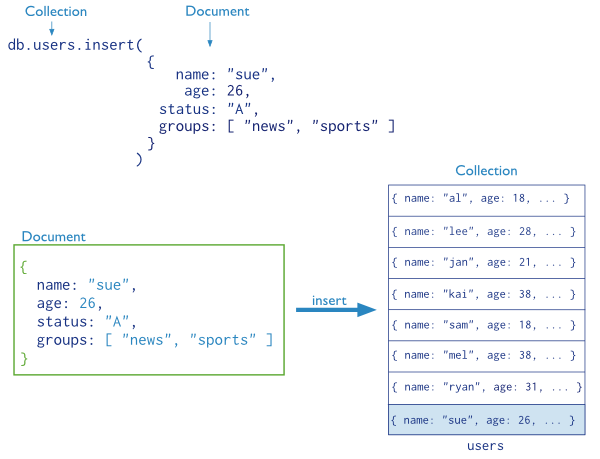


Data Modification: Data Modification là các thao tác create, update, delete dữ liệu. Trong MongoDB, những thao tác này thay đổi dữ liệu của một collection đơn. Với thao tác update và delete, bạn có thể đặc tả các tiêu chuẩn để chọn được những documents cần được update hoặc remove.

Insert: db.collection. insert( { <data> } )

Update: db.collection. update( { <criteria> } )

Delete: db.collection.remove( { <criteria> } )



Các thao tác trên mongo shell cơ bản khác:

db.collection.distinct(“field”) : tìm tất cả các value các nhau trong các documents thuộc một collection hiện tại của field.

db.dropDatabase(): xóa database đang truy cập hiện tại

Hadoop:

Dự án Apache Hadoop phát triển phần mềm mã nguồn mở ( open-source software ) với sự tin cậy, khả năng mở rộng, và tính toán phân tán.

Apache Hadoop software library là một framework cho phép xử lý phân tán các tập hợp dữ liệu lớn thông qua các clusters ( cụm ) các máy tính sử dụng các mô hình lập trình đơn giản. Nó được thiết kế để có thể mở rộng từ một server đơn đến hàn ngàn machines, với mỗi machine phục vụ lưu trữ và tính toán local. Thay vì dựa vào phần cứng để cung cấp high-availability, bản thân library được thiết kế để phát hiện và sử lý lỗi tại application layer ( tầng ứng dụng ), do việc cung cấp dịch vụ highly-available trên một cluster of computers ( cụm các máy tính ), không thể tránh khỏi một trong số đó có thể hoạt động không như ý. ( http://hadoop.apache.org/ )

Dự án bao gồm các module:

+ Hadoop Common: Các tiện ích ( utilities ) phổ biến hỗ trợ cho các module khác.

+ Hadoop Distributed File System ( HDFS ): một hệ thống dữ liệu phân tán cung cấp truy cập high-throughput ( thông lượng cao ) đến dữ liệu ứng dụng.

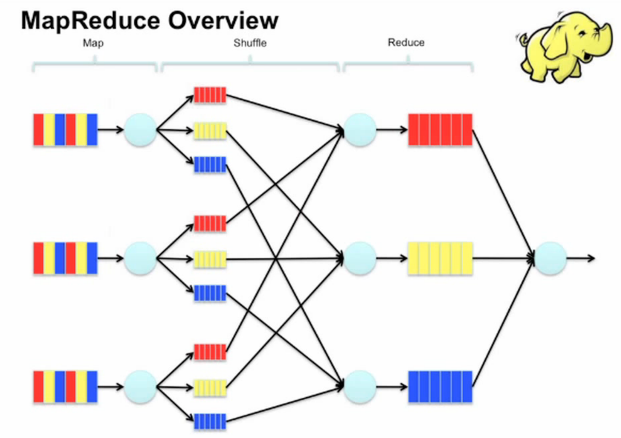
+ Hadoop YARN: một framework dùng trong định thời công việc ( job scheduling ) và quản lý tài nguyên ứng dụng ( cluster resource management )

+ Hadoop MapReduce: một hệ thống YARN-Based cho xử lý xong xong các tập hợp dữ liệu lớn

+ HBase, Hive, Pig, …

Hadoop MapReduce:

MapReduce hoạt động bằng cách phá vỡ quá trình xử lý thành 2 phases ( phần ): map phase và reduce phase. Mỗi phase có các cặp key-value được xem xét như input và output, có thể được chọn bởi người lập trình. Người lập trình cũng đặc tả hai functions: map function và reduce function.



HDFS:

HDFS là một hệ thống file được thiết kế để chứa những file rất lớn với streaming data access patterns ( mô hình truy cập dữ liệu trực tuyến ), chạy trên các cụm clusters of commodity hardware ( kiến trúc của HDFS ).

Một HDFS cluster có hai loại nodes chạy trong một mô hình master-worker ( master-worker pattern ). Một namenode ( the master ) và một số các datanodes ( workers ). Namenode quản lý filesystem namespace. Nó duy trì filesystem tree ( cây hệ thống file ) và metadata cho tất cả file và thư mục trong cây. Thông tin này được chứa liên tục trên ổ đĩa local trong hai files: namespace image và edit log.

