


XÂY DỰNG CƠ SỞ DỮ LIỆU LƯU TRỮ DỮ LIỆU VỀ THU NHẬP CỦA NGƯỜI LỚN TẠI MỸ

ỨNG DỤNG HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU PHI QUAN HỆ MONGODB



 Phạm Thái Quang Nguyên
FX14245

THIẾT KẾ LƯỢC ĐỒ ERD VÀ CẤU TRÚC COLLECTIONS



Dựa vào file dữ liệu có sẵn, dữ liệu sẽ được chia thành 5 thực thể.

Person: chứa những thông tin cơ bản như tuổi, chủng tộc, quốc tịch.

Education: chứa các thông tin liên quan đến học vấn.

Occupation: chứa các thông tin liên quan đến nghề nghiệp.

Relationship: chứa các thông tin về tình trạng hôn nhân.

Finance: chứa các thông tin về tài chính.

Collection và các mối quan hệ

Trong 5 thực thể kể trên, giữa Finance và Person tương đương mối quan hệ 1 – 1 vì nhiều người khó có thể có cùng đồng thời cả số dư tài khoản, biến động số dư tăng và biến động số dư giảm bởi sự chi tiêu của cũng như thu nhập của mỗi người là khác nhau. Do đó, việc lưu trữ thông tin về tài chính (Finance) vào trong các bản ghi (Document) của thực thể con người (Person) sẽ tối ưu truy vấn .

→ Ta có collection Person lưu những thông tin cơ bản và 1 nested Document chứa thông tin Finance.

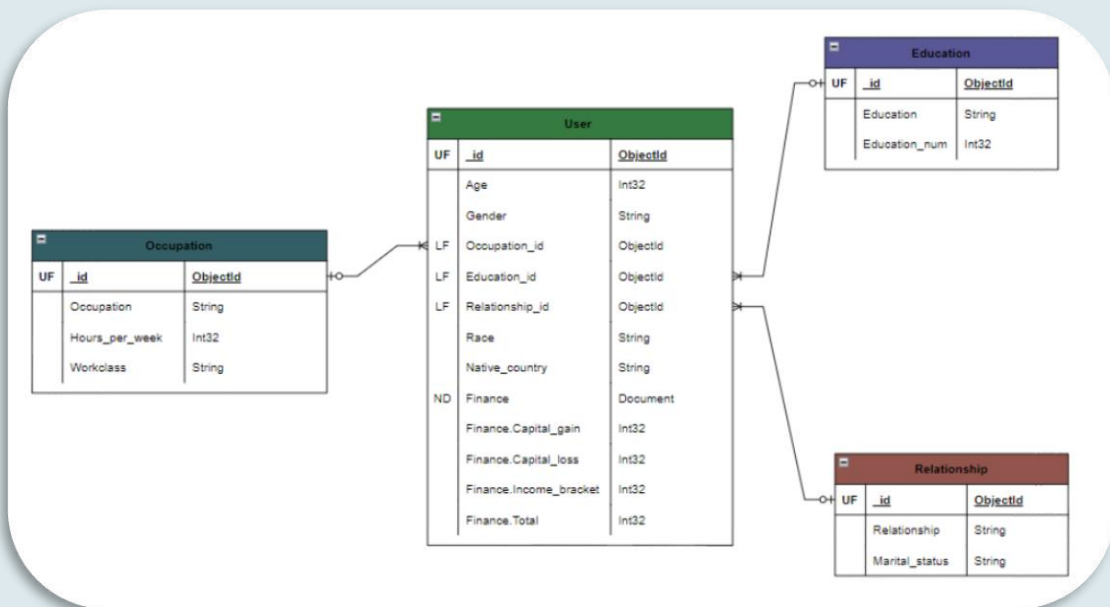
Một người thì chỉ có một bộ thông tin học vấn nhưng một bộ thông tin học vấn thì có thể giống nhau bởi nhiều người. Hay nói cách khác là nhiều người có thể có trình độ học vấn giống y hệt nhau và một người chỉ có 1 bộ dữ liệu tương ứng để thể hiện cho trình độ học vấn của họ, đây thể hiện cho mối quan hệ 1 – n. Bởi vì dữ liệu về học vấn không nhiều nên lưu ở collection riêng biệt sẽ tránh việc trùng lặp dữ liệu và tiết kiệm vùng nhớ.

→ Collection Person và collection Education quan hệ một - nhiều thông qua 1 field trung gian.

→ Tương tự ta cũng có đối với Occupation và Relationship, đều là quan hệ một - nhiều với collection Person với phía nhiều ở Person.

Lược đồ ERD

Những trường (Field) được chọn làm trường trung gian đều là trường _id của collection ở phía một của quan hệ 1 – n , thiết kế schema như vậy sẽ tận dụng được tính độc nhất (unique) của _id và tiết kiệm bộ nhớ so với việc tạo trường mới.



XÂY DỰNG PIPELINE ĐỂ TÁCH DỮ LIỆU VÀO TỪNG COLLECTIONS

Nhập dữ liệu

Để chuyển dữ liệu vào database, ta có thể sử dụng mongo Compass (GUI) hoặc dùng mongoshell (CLI) Di chuyển đến thư mục chứa file dữ liệu, ở đây là data.csv ta sử dụng lệnh mongoimport. Có thể thêm tham số --username và --password với điều kiện bật chế độ authorization để đảm bảo tính bảo mật.



```
mongoimport --type csv -d <database> -c <collection> --headerline --drop <file>
mongoimport --type csv -d US_adult_income -c raw --headerline --drop data.csv
```

Tách dữ liệu (chi tiết được trình bày trong file detach.js trong thư mục code) 🔍

- Bước đầu tiên là duyệt qua toàn bộ document trong collection raw và lấy những trường thông tin cần thiết cho từng collection Education, Occupation và Relationship.
- Bước tiếp theo là nhóm các documents trong từng collection Education, Occupation và Relationship để xóa những bản documents có dữ liệu bị trùng lặp. (sử dụng toán tử \$group)
- Bước cuối cùng là duyệt lại tất cả các document trong collection raw một lần nữa đồng thời tạo những trường trung gian (Occupation_Id, Education_Id và Relationship_Id) tương ứng với dữ liệu có trong những bản ghi của các collection đó.
- Sau cùng có thể xóa collection raw nếu muốn.

MỘT SỐ TRUY VẤN NGHIỆP VỤ VÀ ĐÁNH CHỈ MỤC NÂNG CAO

Một số truy vấn nghiệp vụ (chi tiết được trình bày trong file query.js trong thư mục code)

- Có bao nhiêu người là Nữ và làm việc nhiều hơn 30 tiếng / tuần ?
- Có bao nhiêu người ở Mỹ có mức thu nhập > 50K
- Tính tổng số dư tài khoản của những người đang ở Mỹ.
- Tính tổng số giờ làm việc một tuần của những người có mức thu nhập <= 50K
- Tìm những người có tổng số tiền trong tài khoản > 100000 và có số giờ làm việc hàng tuần < 55.

Đánh chỉ mục nâng cao (chi tiết được trình bày trong file index.js trong thư mục code)

- Sử dụng Compound Index và Unique Index để đánh chỉ mục cho tất cả các trường trong collection Education, Occupation và Relationship. Như vậy sẽ tránh việc dữ liệu thêm vào bị trùng lặp và bởi vì mỗi collection chỉ có từ 2 đến 3 và dữ liệu có thể lưu của những collection này là ít nên sẽ không tốn quá nhiều bộ nhớ cho chỉ mục.
- Sử dụng Partial Index cho collection Person với điều kiện native_country : “United-States” sẽ giúp tiết kiệm thời gian truy vấn dữ liệu về những người mang quốc tịch Mỹ, vốn chiếm phần nhiều trong dữ liệu, ngoài ra đặt chỉ mục trên trường chủng tộc (race) sẽ dễ dàng phân loại dữ liệu đối với nước Mỹ vốn đã có rất nhiều dân tộc nhập cư tới.