# Assessment 2 – Xây dựng cơ sở dữ liệu MongoDB

## Mô tả dự án

Dự án xây dựng cơ sở dữ liệu MongoDB từ một tập dữ liệu csv có sẵn.

Email tác giả: [baodkfx23194@funix.edu.vn](mailto:baodkfx23194@funix.edu.vn)

Mục lục

[Assessment 2 – Xây dựng cơ sở dữ liệu MongoDB 1](#_Toc150572212)

[Mô tả dự án 1](#_Toc150572213)

[Tổng quan dự án 1](#_Toc150572214)

[Chi tiết dự án 2](#_Toc150572215)

[1. Phân tích tập dữ liệu 2](#_Toc150572216)

[2. Thiết kế lược đồ quan hệ thực thể cho cơ sở dữ liệu 2](#_Toc150572217)

[3. Commands for creating mongoDB database: 3](#_Toc150572218)

[4. Các truy vấp nghiệp vụ: 3](#_Toc150572219)

[5. (Nâng cao) Xây dựng index cho các collection 4](#_Toc150572220)

## Tổng quan dự án

* Phân tích được tập dữ liệu.
* Thiết kế được lược đồ của Database dựa trên các phân tích.
* Hiểu được sự khác nhau khi thiết kế Database giữa SQL Server và MongoDB.
* Viết được các câu lệnh để tạo Database theo như lược đồ đã thiết kế.
* Liệt kê được các Business Query (truy vấn nghiệp vụ) và viết các câu lệnh để thực hiện các truy vấn đó.
* Diễn đạt cách đánh index giúp cải thiện tốc độ truy vấn nhưng đảm bảo không chiếm quá nhiều bộ nhớ.

## Chi tiết dự án

### Phân tích tập dữ liệu

Collection **user**: chứa các thông tin cơ bản như *age* (tuổi), *gender* (giới tính), *race* (chủng tộc), *native\_country* (quốc tịch).

Collection **education:** chứa các thông tin liên quan đến *education* (học vấn) được định danh bởi *education\_num*.

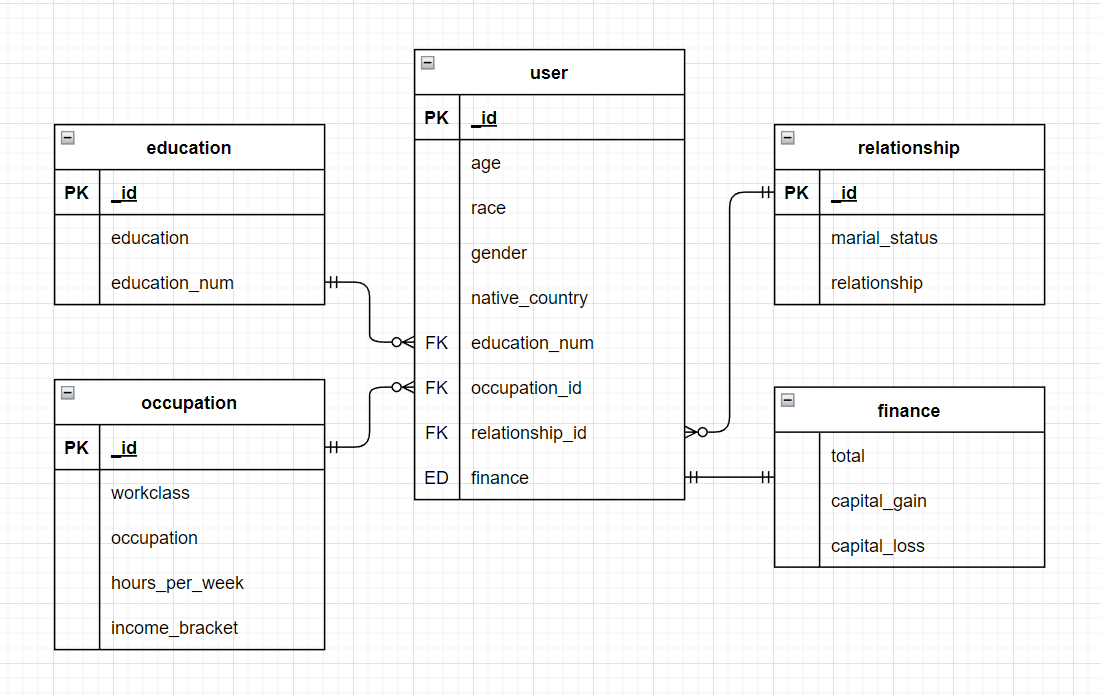
Collection **occupation:** chứa các thông tin liên quan về *occupation* (nghề nghiệp), *workclass* (loại công việc), *hours\_per\_week* (số tiếng làm việc trong tuần), *income\_bracket* (mức lương <=50k hoặc >50k)

Collection **relationship:** chứa các thông tin về *marial\_status* (tình trạng hôn nhân) và *relationship* ( mối quan hệ với chủ hộ).

**finance:** chứa các thông tin liên quan đến vấn đề tài chính như *capital\_gain* (số tiền được nhận), *capital\_loss* (số tiền chi tiêu), *total* (số tiền trong tài khoản của cá nhân hoặc người giám hộ), đây là một ***Embedded documents*** nên sẽ trở thành 1 trường của **user** document

### Thiết kế lược đồ quan hệ thực thể cho cơ sở dữ liệu

File diagram có tên là: “erd.dio”



ERD được thiết kế theo mô hình ngôi sao Star Schema

Vì dữ liệu không chứa định danh người dùng, nên hầu như các mối quan hệ giữa các thực thể này là 1-to-many dựa theo từng dòng trong tệp dữ liệu

Các mối quan hệ giữa các thực thể trong cơ sở dữ liệu này:

* User – Education : 1-to-many, 1 người có 1 học vấn, 1 học vấn có thể có nhiều người
* User – Occupation: 1-to-many, 1 người có 1 nghề nghiệp, 1 nghề nghiệp có thể có nhiều người
* User – Relationship: 1-to-many, 1 người có 1 mối quan hệ, 1 mối quan hệ có thể có nhiều người
* User – Finance: 1-to-1, 1 người có 1 dữ liệu tài chính, 1 dữ liệu tài chính sẽ chỉ được bảo mật cho 1 người

### Commands for creating mongoDB database:

* Câu lệnh để nhập file csv vào cơ sở dữ liệu mongoDB

|  |
| --- |
| cmd |
| cd Downloads/Assessment2  mongoimport /db:asm2 /collection:full /uri:mongodb://localhost:27017 /file:USADUL~1.CSV /type:csv /headerline |

* Câu lệnh để nhập script ‘initial.mongodb.js’ vào cơ sở dữ liệu mongoDB

|  |
| --- |
| mongosh |
| load(“initial.mongodb.js”);  loadEducation();  loadOccupation();  loadRelationship();  loadUser(); |

* Khi script chạy xong màn hình sẽ hiện thế này, ví dụ cho function **loadUser()**

A screen shot of a computer

Description automatically generated

### Các truy vấp nghiệp vụ:

* Danh sách truy vấn nghiệp vụ:
  + Top 5 quốc gia có tổng số dư tài khoản lớn nhất
  + Tổng số người trong từng nhóm tuổi và nghề nghiệp
  + Số lượng người làm việc nhiều hơn 40 giờ/tuần theo mức thu nhập
  + Bình quân số tiền trong tài khoản và bình quân số giờ làm việc theo nghề nghiệp và mức giáo dục
  + Khoảng tuổi và và giờ làm việc mỗi tuần theo các khoảng phân loại theo tổng số dư trong tài khoản đối với các nghề nghiệp cá nhân
* File script các câu lệnh thực hiện các truy vấn trên: ‘query.mongodb.js’

### (Nâng cao) Xây dựng index cho các collection

* Collection `**education**`, để tăng tốc tìm kiếm cho mối quan hệ với collection **user** ta sẽ đánh unique index trên field `***education\_num***`.
* Collection `**occupation**`:
  + Ta tạo index cho các fields `***workclass***`, `***hours\_per\_weeks***`, `***income\_bracket***`, `***occupation***` vì chúng sẽ được truy vấn đồng thời với nhau khi load initial.
  + Tạo thêm 1 index ở trường `***hours\_per\_week***` vì đây là trường được truy vấn khá nhiều trong các nghiệp vụ.
  + Một partial index được tạo cho trường `***workclass***` với điều kiện workclass=*Private* để phục vụ cho truy vấn 5.
* Collection `**relationship**`, 1 combine index cho 2 trường `***marital\_status***` và `***relationship***`
* Collection `**user**`:
  + 1 combine index cho các trường `***age***`, `***gender***`, `***race***`, `***native\_country***`, `***finance.total***`, `***finance.capital\_gain***` và `***finance.capital\_loss***`.
  + Các index hỗ trợ cho việc $**lookup**: `***education\_num***`, `***occupation\_id***`, `***relationship\_id***`.
  + Index phục vụ cho hầu hết các truy vấn: `***finance.total***`.
  + Partial Index được cho trường được dự đoán sẽ sử dụng nhiều `***native\_country***` khi người ở “*United-States*”
* Chạy command sau trên mongosh để chạy file tạo index ‘index.mongodb.js’:

|  |
| --- |
| **mongosh** |
| load(“index.mongodb.js”); |

* Sau khi chạy xong màn hình sẽ hiện tương tự thế này

