ASSIGMENT 2

**1.  Phân tích được tập dữ liệu**

-Từ các trường của tập dữ liệu US Adult Income chứa các thông tin về mức thu thập và các thông tin cá nhân của một số người ở nước Mỹ vào năm 1994, Em có thể chia thành các collection sau:

Collection: users

* age: độ tuổi (integer)
* gender: giới tính (string)
* race: chủng tộc (string)
* native\_country: quốc tịch (string)
* education\_id: tham chiếu đến collection "educations"
* occupatios\_id: tham chiếu đến collection "occupations"
* finance\_id : tham chiếu đến collection "finances"
* relationship\_id: tham chiếu đến collection "relationships"

Collection: educations

* education: học vấn (string)
* education\_num: cấp độ học vấn (integer)

Collection: occupations

* occupation: nghề nghiệp (string)
* workclass: Hình thức làm việc.(string)
* hours\_per\_week: số giờ làm việc trong tuần (integer)

Collection: relationships

* marital\_status: tình trạng hôn nhân (string)
* relationship: mối quan hệ với chủ hộ (string)

Collection: finances

* total: tổng số tiền trong tài khoản của cá nhân hoặc của người giám hộ (integer)
* capital\_gain: biến động số dư tăng trong tài khoản (integer)
* capital\_loss: biến động số dư giảm trong tài khoản (integer)
* income\_bracket: mức thu nhập (string)

**2. Thiết kế được lược đồ của Database dựa trên các phân tích.**

1. Mối quan hệ giữa collection "users" và collection "educations":

-Một người dùng có thể có một hoặc nhiều thông tin về học vấn (1-n).

1. Mối quan hệ giữa collection "users" và collection "occupations":

-Một người dùng có thể có một hoặc nhiều thông tin về nghề nghiệp (1-n).

1. Mối quan hệ giữa collection "users" và collection "relationships":

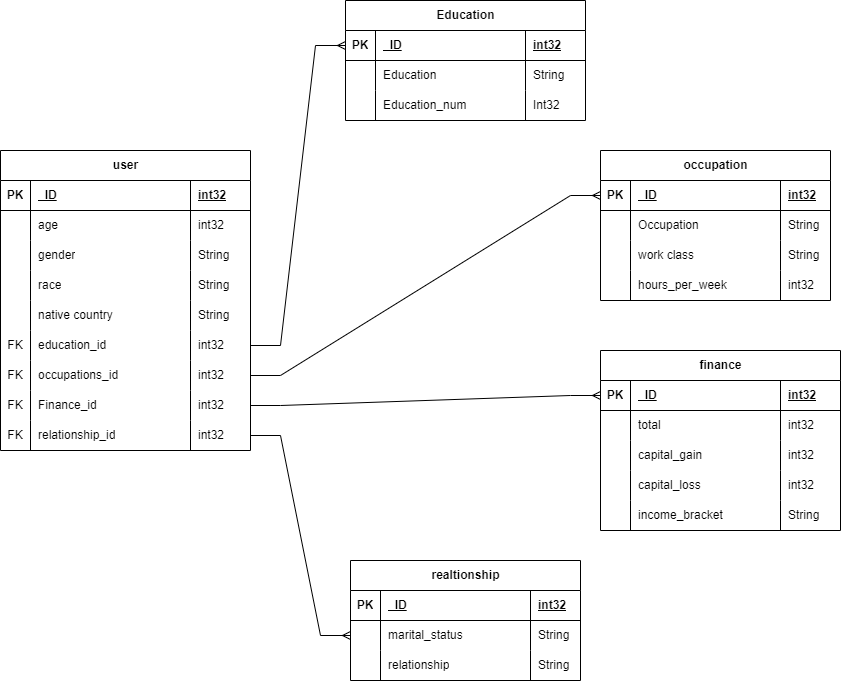
-Một người dùng có thể có một hoặc nhiều thông tin về mối quan hệ (1-n).

1. Mối quan hệ giữa collection "users" và collection "finances":

-Một người dùng có thể có một hoặc nhiều thông tin về tài chính (1-n).

Tóm lại, các mối quan hệ trong hệ thống của bạn có thể được mô tả như sau:

Mối quan hệ 1-n (one-to-many): users-educations, users-occupations, users-relationships, users-finances.



Note : em có để một hình ERD.png cho thầy cô dễ theo dõi chấm điểm , công cụ em sử dụng để vẽ là draw.io

**3. Viết được các câu lệnh để tạo Database theo như lược đồ đã thiết kế.**

**Sau đây là phần chú thích vị trí để bài làm của các phần để dễ theo dõi và chấm điểm:**

* **Ngôn ngữ sử Dụng: Python**
* get\_database.py : sử dụng để chia data vào các collection
* insert\_data.py : sử dụng để thêm 1 hoặc n đối tượng vào data, tự động chia vào các collection sau thi thêm các giá trị
* db.dep302\_asm2: chứa các kết quả collection sau khi đã chia dữ liệu

**4. Liệt kê được các Business Query (truy vấn nghiệp vụ) và viết các câu lệnh để thực hiện các truy vấn đó.**

Câu truy vấn 1: Có bao nhiêu người là Nữ và làm việc nhiều hơn 30 tiếng / tuần ?

db.users.aggregate([

{

$lookup: {

from: "finances",

localField: "finance\_id",

foreignField: "\_id",

as: "finance"

}

},

{

$match: {

gender: " Female",

"finance.hours\_per\_week": { $gt: 30 }

}

},

{

$count: "count"

}

]);

=> count:8048

Câu truy vấn 2: Có bao nhiêu người ở Mỹ có mức thu nhập > 50K.

db.users.aggregate([

{

$lookup: {

from: "finances",

localField: "finance\_id",

foreignField: "\_id",

as: "finance"

}

},

{

$match: {

native\_country: " United-States",

"finance.income\_bracket": " >50K"

}

},

{

$count: "count"

}

]);

=> count: 7171

Câu truy vấn 3: Tính tổng số dư tài khoản của những người đang ở Mỹ.

db.users.aggregate([

{

$lookup: {

from: "finances",

localField: "finance\_id",

foreignField: "\_id",

as: "finance"

}

},

{

$match: {

native\_country: " United-States"

}

},

{

$unwind: "$finance"

},

{

$group: {

\_id: null,

total\_balance: { $sum: "$finance.total" }

}

},

{

$project: {

\_id: 0,

total\_balance: 1

}

}

])

=>total\_balance: Long("5456803347")

Câu truy vấn 4: Tính tổng số giờ làm việc một tuần của những người có mức thu nhập <= 50K.

db.users.aggregate([

{

$lookup: {

from: "finances",

localField: "finance\_id",

foreignField: "\_id",

as: "finance"

}

},

{

$unwind: "$finance"

},

{

$match: {

"finance.income\_bracket": " <=50K"

}

},

{

$lookup: {

from: "occupations",

localField: "occupation\_id",

foreignField: "\_id",

as: "occupation"

}

},

{

$unwind: "$occupation"

},

{

$group: {

\_id: null,

total\_hours\_per\_week: { $sum: "$occupation.hours\_per\_week" }

}

},

{

$project: {

\_id: 0,

total\_hours\_per\_week: 1

}

}

])

=>total\_hours\_per\_week: 960130

Câu truy vấn 5: Tìm những người có tổng số tiền trong tài khoản > 100000 và có số giờ làm việc hàng tuần < 55.

db.users.aggregate([

{

$lookup: {

from: "finances",

localField: "finance\_id",

foreignField: "\_id",

as: "finance"

}

},

{

$unwind: "$finance"

},

{

$match: {

"finance.total": { $gt: 100000 },

}

},

{

$lookup: {

from: "occupations",

localField: "occupation\_id",

foreignField: "\_id",

as: "occupation"

}

},

{

$unwind: "$occupation"

},

{

$match: {

"occupation.hours\_per\_week": { $lt: 55 }

}

}

])