

Assignment 2: Xây dựng cơ sở dữ liệu MongoDB

1. Phân tích được tập dữ liệu.

Dựa vào tập dữ liệu được cung cấp, tôi sẽ thiết kế Database như sau:

Database: dep302_asm02

Collection:

user:

- Chứa các trường:
 - `_id`: ID của người dùng.
 - `age`: Tuổi
 - `gender`: Giới tính.
 - `race`: Chủng tộc.
 - `native_country`: Quốc tịch.
- Khóa ngoại:
 - `finance_id`: liên kết với `_id` của `finance`
 - `occupation_id`: liên kết với `_id` của `occupation`
 - `relationship_id`: liên kết với `_id` của `relationship`
 - `education_id`: liên kết với `_id` của `education`

education:

- Chứa các trường:
 - `_id`: ID của `education`
 - `education`: Trình độ học vấn.
 - `education_num`: Cấp độ học vấn.

occupation:

- Chứa các trường:
 - `_id`: ID của `occupation`
 - `occupation`: Nghề nghiệp.
 - `workclass`: Hình thức làm việc.
 - `hours_per_week`: Số giờ làm việc trong tuần.

relationship:

- Chứa các trường:
 - `_id`: ID của `relationship`
 - `relationship`: Mối quan hệ với chủ hộ.
 - `marital_status`: Tình trạng hôn nhân

finance:

- Chứa các trường:
 - `_id`: ID của `finance`
 - `total`: Tổng số tiền trong tài khoản của cá nhân hoặc của người giám hộ.
 - `capital_gain`: Biến động số dư tăng trong tài khoản.
 - `capital_loss`: Biến động số dư giảm trong tài khoản.
 - `income_bracket`: Mức thu nhập.

Liên kết giữa các collection:

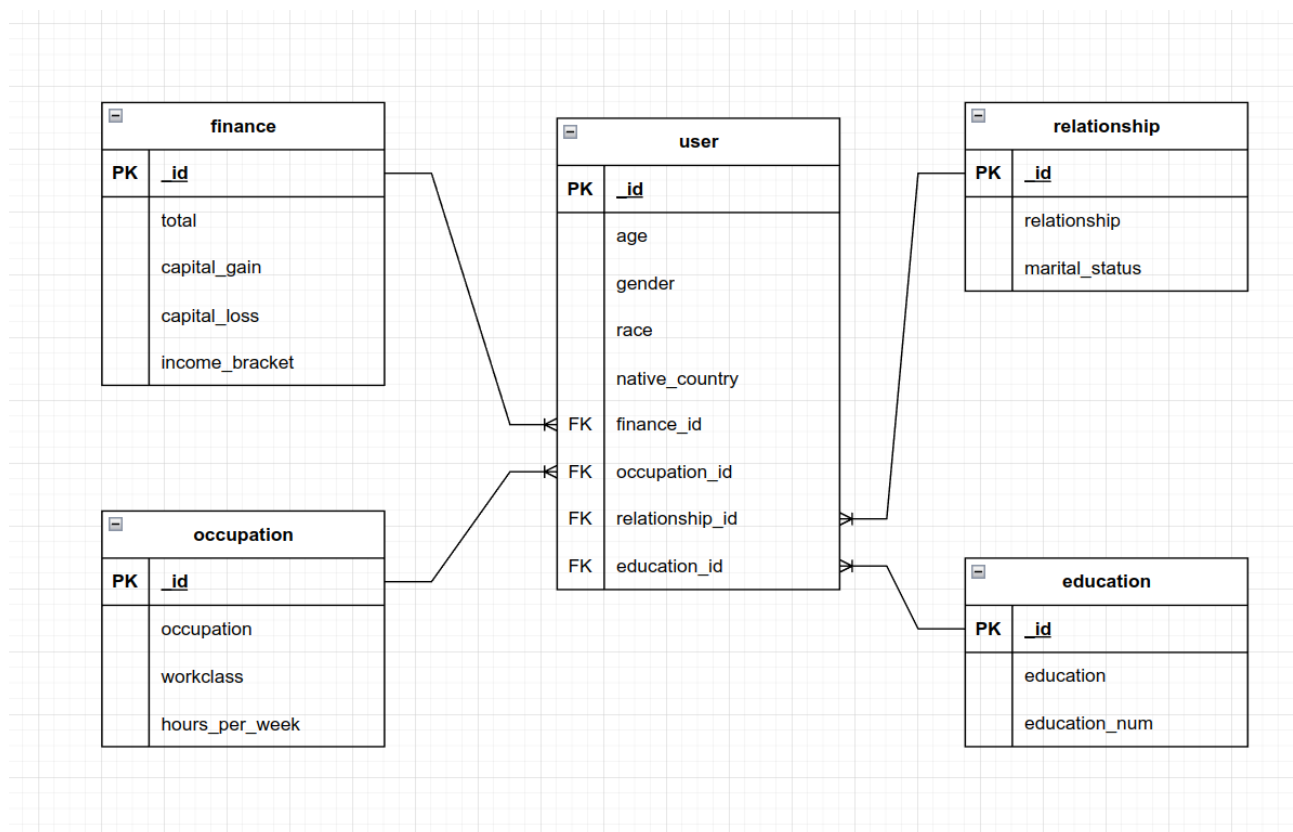
- Collection **user** sẽ liên kết với collection **education**, **occupation**, **relationship** và **finance** thông qua trường **_id** của từng collection.
- Collection **education** sẽ liên kết với collection **user** thông qua trường **education_id**.
- Collection **occupation** sẽ liên kết với collection **user** thông qua trường **occupation_id**.
- Collection **relationship** sẽ liên kết với collection **user** thông qua trường **relationship_id**.
- Collection **finance** sẽ liên kết với collection **user** thông qua trường **finance_id**.

Lý do thiết kế như vậy:

- Thiết kế này cho phép chúng ta lưu trữ tất cả các dữ liệu trong tập dữ liệu một cách đầy đủ và chính xác.
- Các collection được liên kết với nhau thông qua trường **_id** giúp chúng ta có thể dễ dàng truy vấn dữ liệu.

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

Dựa trên phân tích trên, ta thiết kế lược đồ cho database như sau:



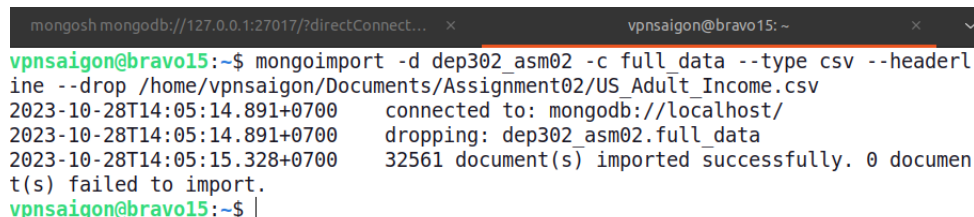
3. Viết được các câu lệnh để tạo Database theo như lược đồ đã thiết kế.

Các câu lệnh được trình bày chi tiết trong file **create_insert_db**. Dưới đây là một vài hình ảnh thực hiện.

```
test> use dep302_asm02
switched to db dep302_asm02
dep302_asm02> db.createCollection('full_data')
{ ok: 1 }
dep302_asm02> db.createCollection('user')
{ ok: 1 }
dep302_asm02> db.createCollection('education')
{ ok: 1 }
dep302_asm02> db.createCollection('occupation')
{ ok: 1 }
dep302_asm02> db.createCollection('relationship')
{ ok: 1 }
dep302_asm02> db.createCollection('finance')
{ ok: 1 }
dep302_asm02> show collections
education
finance
full_data
occupation
relationship
user
dep302_asm02> |
```

// Import file csv vào database MongoDB

```
mongoimport -d dep302_asm02 -c full_data --type csv --headerline --drop
/home/vpnsaigon/Documents/Assignment02/US_Adult_Income.csv
```



```
vpnsaigon@bravo15:~$ mongoimport -d dep302_asm02 -c full_data --type csv --headerline --drop /home/vpnsaigon/Documents/Assignment02/US_Adult_Income.csv
2023-10-28T14:05:14.891+0700 connected to: mongodb://localhost/
2023-10-28T14:05:14.891+0700 dropping: dep302_asm02.full_data
2023-10-28T14:05:15.328+0700 32561 document(s) imported successfully. 0 documents failed to import.
vpnsaigon@bravo15:~$ |
```

```
dep302_asm02> db.full_data.findOne()
{
  _id: ObjectId("653cb2aa5e7e2f725ef085c2"),
  age: 39,
  workclass: 'State-gov',
  total: 77516,
  education: 'Bachelors',
  education_num: 13,
  marital_status: 'Never-married',
  occupation: 'Adm-clerical',
  relationship: 'Not-in-family',
  race: 'White',
  gender: 'Male',
  capital_gain: 2174,
  capital_loss: 0,
  hours_per_week: 40,
  native_country: 'United-States',
  income_bracket: '<=50K'
}
dep302_asm02> |
```

4. Liệt kê được các Business Query (truy vấn nghiệp vụ)

Các câu lệnh được trình bày chi tiết trong file **bussiness_query**.

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

```
dep302_asm02> db.user.aggregate([
...   {
...     $lookup: {
...       from: "occupation",
...       localField: "occupation_id",
...       foreignField: "_id",
...       as: "occupation"
...     }
...   },
...   {
...     $match: {
...       "gender": "Female",
...       "occupation.hours_per_week": { $gt: 30 }
...     }
...   },
...   {
...     $count: "result"
...   }
... ])
[ { result: 8799 } ]
dep302_asm02> |
```

Đáp án: Có 8799 người.

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

```
dep302_asm02> db.user.aggregate([
...   {
...     $lookup: {
...       from: "finance",
...       localField: "finance_id",
...       foreignField: "_id",
...       as: "finance"
...     }
...   },
...   {
...     $match: {
...       native_country: "United-States",
...       "finance.income_bracket": ">50K"
...     }
...   },
...   {
...     $count: "result"
...   }
... ])
[ { result: 7150 } ]
dep302_asm02>
```

Đáp án: Có 7150 người.

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

```
dep302_asm02> db.user.aggregate([
...   {
...     $lookup: {
...       from: "finance",
...       localField: "finance_id",
...       foreignField: "_id",
...       as: "finance"
...     }
...   },
...   {
...     $unwind: '$finance'
...   },
...   {
...     $match: {
...       "native_country": "United-States"
...     }
...   },
...   {
...     $group: {
...       _id: null,
...       total_balance: { $sum: "$finance.total" }
...     }
...   }
... ])
[ { _id: null, total_balance: Long("5434891017") } ]
dep302_asm02> |
```

Đáp án: tổng số dư tài khoản là **5434891017**

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

```
dep302_asm02> db.user.aggregate([
...   {
...     $lookup: {
...       from: "occupation",
...       localField: "occupation_id",
...       foreignField: "_id",
...       as: "occupation"
...     }
...   },
...   {
...     $lookup: {
...       from: "finance",
...       localField: "finance_id",
...       foreignField: "_id",
...       as: "finance"
...     }
...   },
...   {
...     $unwind: '$occupation'
...   },
...   {
...     $match: {
...       'finance.income_bracket': "<=50K"
...     }
...   },
...   {
...     $group: {
...       _id: null,
...       hour_total: { $sum: '$occupation.hours_per_week' }
...     }
...   }
... ])
[ { _id: null, hour_total: 976086 } ]
dep302_asm02> |
```

Đáp án: là **976086** giờ

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

```
dep302_asm02> db.user.aggregate([
...   {
...     $lookup: {
...       from: "occupation",
...       localField: "occupation_id",
...       foreignField: "_id",
...       as: "occupation"
...     }
...   },
...   {
...     $lookup: {
...       from: "finance",
...       localField: "finance_id",
...       foreignField: "_id",
...       as: "finance"
...     }
...   },
...   {
...     $match: {
...       $and: [
...         {'finance.total': {$gt: 100000}},
...         {'occupation.hours_per_week': {$lt: 55}}
...       ]
...     }
...   },
...   {
...     $project: {
...       _id: 1,
...       age: 1,
...       gender: 1,
...       race: 1,
...       native_country: 1,
...       'finance.total': 1,
...       'occupation.hours_per_week': 1
...     }
...   }
... ])|
```

Đáp án:

```
{
  _id: ObjectId("653cc6232e3e9f9eb3d0f7c3"),
  age: 54,
  gender: 'Female',
  race: 'Black',
  native_country: 'United-States',
  occupation: [ { hours_per_week: 16 } ],
  finance: [ { total: 302146 } ]
},
{
  _id: ObjectId("653cc6232e3e9f9eb3d0f7c4"),
  age: 19,
  gender: 'Male',
  race: 'White',
  native_country: 'United-States',
  occupation: [ { hours_per_week: 40 } ],
  finance: [ { total: 168294 } ]
}
]
Type "it" for more
dep302_asm02> |
```

5. (Yêu cầu nâng cao) Xây dựng Index cho các Collection

Để tăng tốc cho việc truy vấn dữ liệu, bạn hãy thiết lập các chỉ mục cho từng Collection để truy vấn nhanh hơn. Đồng thời bạn cũng phải giải thích được tại sao lại xây dựng Index như vậy.

Dưới đây là các chỉ mục sẽ tạo cho các collection trong database:

Collection user:

- **_id**: Chỉ mục chính, được tạo tự động.
- **age**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên độ tuổi.
- **gender**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên giới tính.
- **race**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên chủng tộc.
- **native_country**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên quốc tịch.

Collection education:

- **_id**: Chỉ mục chính, được tạo tự động.
- **education**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên trình độ học vấn.
- **education_num**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên cấp độ học vấn.

Collection occupation:

- **_id**: Chỉ mục chính, được tạo tự động.
- **occupation**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên nghề nghiệp.
- **workclass**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên hình thức làm việc.
- **hours_per_week**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên số giờ làm việc trong tuần.

Collection finance:

- **_id**: Chỉ mục chính, được tạo tự động.
- **total**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên tổng số dư tài khoản.
- **capital_gain**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên biến động số dư tăng trong tài khoản.
- **capital_loss**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên biến động số dư giảm trong tài khoản.
- **income_bracket**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên mức thu nhập.

Collection relationship:

- **_id**: Chỉ mục chính, được tạo tự động.
- **relationship**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên mối quan hệ với chủ hộ.
- **marital_status**: Chỉ mục đơn, được sử dụng để truy vấn dữ liệu dựa trên Tình trạng hôn nhân

Có một số lý do chọn xây dựng các chỉ mục như vậy:

- **Để cải thiện hiệu suất của các truy vấn có điều kiện.** Các chỉ mục cho phép MongoDB tìm kiếm dữ liệu nhanh hơn nhiều so với khi không có chỉ mục.
- **Để cải thiện hiệu suất của các truy vấn tổng hợp.** Các chỉ mục có thể được sử dụng để cải thiện hiệu suất của các truy vấn tổng hợp, chẳng hạn như truy vấn tìm tổng hoặc trung bình của một trường.

```
// -----  
// 5. (Yêu cầu nâng cao) Xây dựng Index cho các Collection  
// Đề'tăng tốc cho việc truy vấn dữ liệu, bạn hãy thiết lập các  
chỉ mục cho từng Collection để'truy vấn nhanh hơn. Đồng thời bạn  
cũng phải giải thích được tại sao lại xây dựng Index như vậy.
```

```
db.user.createIndex({ age: 1 });  
db.user.createIndex({ gender: 1 });  
db.user.createIndex({ race: 1 });  
db.user.createIndex({ native_country: 1 });  
  
db.finance.createIndex({ total: 1 });  
db.finance.createIndex({ capital_gain: 1 });  
db.finance.createIndex({ capital_loss: 1 });  
db.finance.createIndex({ ncome_bracket: 1 });  
  
db.occupation.createIndex({ occupation: 1 });  
db.occupation.createIndex({ workclass: 1 });  
db.occupation.createIndex({ hours_per_week: 1 });  
  
db.education.createIndex({ education: 1 });  
db.education.createIndex({ education_num: 1 });  
  
db.relationship.createIndex({ relationship: 1 });  
db.relationship.createIndex({ marital_status: 1 });
```