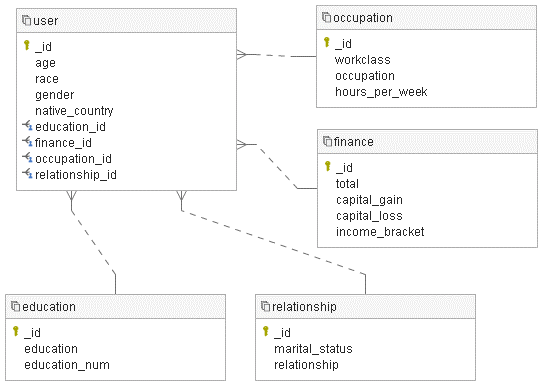
# **1.** Phân tích tập dữ liệu.

Từ bảng dữ liệu trong tập tin nguồn **US Adult Income.csv**, em sẽ nó thành các Collections sau:

* Collection Education chứa các thông tin liên quan đến học vấn, gồm các trường: **education** (trình độ học vấn)**, education\_num** (số thứ tự).
* Collection Occupation chứa các thông tin liên quan về nghề nghiệp, gồm các trường: **occupation** (nghề nghiệp), **workclass** (hình thức làm việc), **hours\_per\_week** (số giờ làm việc mỗi tuần).
* Collection Relationship chứa các thông tin về tình trạng hôn nhân, gồm các trường: **marital\_status** (Tình trạng hôn nhân), **relationship** (Mối quan hệ với chủ hộ).
* Collection Finance chứa các thông tin liên quan đến vấn đề tài chính, gồm các trường: **total** (tiền trong tài khoản), **income\_bracket (thuc nhập)**, **capital\_gain** (biến động tăng), **capital\_loss** (biến động giảm).
* Collection User chứa các trường: **age**, **race**, **gender**, **native\_country**, cùng các trường tham chiếu đến 4 collections kia.

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



Lược đồ trên được em dùng Dataedo để vẽ.

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

Đầu tiên, em import file US Adult Income.csv vào 1 collection trung gian có tên staging.

Sau đó, em chọn ra các cột thích hợp từ staging để import vào các collections: occupation, relationship, education và finance.

Em sử dụng $lookup để kết hợp staging và 4 collections kia để import vào user collection.

Chi tiết trong thư mục **Code\_import**.

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

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

db.user.aggregate

(

[

{$match: {gender: ' Female'}},

{$lookup:

{

from: 'occupation',

localField: 'occupation\_id',

foreignField: '\_id',

as: 'occupation\_info'

}

},

{$unwind: '$occupation\_info'},

{$match: {'occupation\_info.hours\_per\_week': {$gt: 30}}},

{$count: 'Women work over 30h per week'}

]

)

>8048

* Có bao nhiêu người ở Mỹ có mức thu nhập > 50K

db.user.aggregate

(

[

{$match: {native\_country: ' United-States'}},

{$lookup:

{

from: 'finance',

localField: 'finance\_id',

foreignField: '\_id',

as: 'finance\_info'

}

},

{$unwind: '$finance\_info'},

{$match: {'finance\_info.income\_bracket': ' >50K'}},

{$count: 'US citizen has income bracket over 50K'}

]

)

>7171

* Tính tổng số dư tài khoản của những người đang ở Mỹ.

db.user.aggregate

(

[

{$match: {native\_country: ' United-States'}},

{$lookup:

{

from: 'finance',

localField: 'finance\_id',

foreignField: '\_id',

as: 'finance\_info'

}

},

{$unwind: '$finance\_info'},

{$group:

{

\_id: '$native\_country',

'Total credist of US citizen': {$sum: '$finance\_info.total'}

}

},

{$project: {\_id: 0, 'Total credist of US citizen':1}}

]

)

> 5456803347

* Tính số giờ làm việc trung bình một tuần của những người có mức thu nhập <= 50K

db.user.aggregate

(

[

{$lookup:

{

from: 'finance',

localField: 'finance\_id',

foreignField: '\_id',

as: 'finance\_info'

}

},

{$unwind: '$finance\_info'},

{$match: {'finance\_info.income\_bracket': ' <=50K'}},

{$lookup:

{

from: 'occupation',

localField: 'occupation\_id',

foreignField: '\_id',

as: 'occupation\_info'

}

},

{$unwind: '$occupation\_info'},

{$group:

{

\_id: null,

'Average working hours per week': {$avg: '$occupation\_info.hours\_per\_week'}

}

},

{$project: {\_id: 0, 'Average working hours per week': {$round: ['$Average working hours per week',1]}}}

]

)

>38.8

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

db.user.aggregate

(

[

{$lookup:

{

from: 'finance',

localField: 'finance\_id',

foreignField: '\_id',

as: 'finance\_info'

}

},

{$unwind: '$finance\_info'},

{$match: {'finance\_info.total': {$gt: 100000}}},

{$lookup:

{

from: 'occupation',

localField: 'occupation\_id',

foreignField: '\_id',

as: 'occupation\_info'

}

},

{$unwind: '$occupation\_info'},

{$match: {'occupation\_info.hours\_per\_week': {$lt: 55}}},

{$count: 'total'}

]

)

>24200

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

Hai collections finance và occupation được em thêm 2 index lần lượt là {total, income\_bracket} và {hours\_per\_week} vì chúng thường xuyên được tham chiếu tới. Nó cũng giúp việc thêm dữ liệu vào user diễn ra nhanh hơn (35s thay vì 27 phút).