**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: mongo script**
* Load\_collection.js : sử dụng để chia data vào các collection.
* db.dep302\_asm2: chứa các kết quả collection sau khi đã chia dữ liệu.
* full.json : tập dữ liệu lớn để từ đó load vào các collection users, educations , occupations, finances, relationships.

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

**Định nghĩa:**

Xây dựng các chỉ mục (indexes) trong MongoDB có thể giúp tăng tốc đáng kể cho việc truy vấn dữ liệu và cải thiện hiệu suất của hệ thống. Các chỉ mục được sử dụng để tìm kiếm và truy xuất dữ liệu một cách hiệu quả, giúp giảm thiểu số lượng bản ghi phải duyệt qua để tìm kiếm.

Dưới đây là phần tạo index cho các attribute cho từng collection

1. Collection `educations`:

db.educations.createIndex({ education: 1 }); # theo trình độ học vấn

2. Collection `occupations`:

db.occupations.createIndex({ occupation: 1 }); # theo nghề nghiệp

3. Collection `relationships`:

db.relationships.createIndex({ marital\_status: 1 }); #theo mối quan hệ

4. Collection `finances`:

db.finances.createIndex({ income\_bracket: 1 });# theo thu nhập

5. Collection `users`:

db.users.createIndex({ age: 1 }); theo độ tuổi

Mục đích : lấy dữ liệu lọc dữ liệu nhanh hơn hiệu quả hơn bằng cách tạo các mục , các attribute có tần xuất xuất hiện nhiều , có thể kếp hợp nhiều truy vấn ví dụ:

* truy vấn người đã kết hôn trên 20 tuổi có thu nhập <=50
* truy vấn những người là Adm clerical có độ tuổi trên 20

ngoài ra còn có thể chọn theo độ phân tán tùy vào bộ dữ liệu để tạo index cho atteibute từ đó đễ khoanh vùng, phân tích , giúp giảm thiểu số lượng bản ghi phải duyệt qua để tìm kiếm.

Cách thức chọn :

* theo tần xuất xuất hiện trong các truy vấn
* theo độ phân tán dữ liệu

Dưới dây là các biểu đồ cột từ đó đánh giá độ phân tán từ đó chọn ra các Attribute để tạo index em có để các biểu dồ ở các hình ghi chú ở foder bổ xung để dễ dàng theo dõi.