**3. Triển khai bài toán K-means sử dụng tuần tự và Hadoop Mapreduce (song song).**

**a) Mô tả bài toán**

Tập dữ liệu chứa 2 tọa độ x và y sẽ được tạo ngẫu nhiên từ hàm makeblob của thư viện scikit-learn cùng với 3 cụm khởi tạo được chọn random từ tập dữ liệu. Dữ liệu sẽ được tạo theo quy mô từ 10000, 100000,1000000 dòng để kiểm thử về khả năng tăng tốc của MapReduce so với tuần tự.

**b) Triển khai tuần tự**

Dựa vào phần tóm tắt thuật toán ở phần 2, hoặc mã giả dưới đây để hình dung quá trình chạy của thuật toán K-means trong triển khai tuần tự.

Mã giả:

* Đầu vào:
  + data points: Danh sách điểm dữ liệu
  + K: K cụm ngẫu nhiên được khởi tạo.
* Đầu ra:
  + K: Danh sách (tọa độ, vị trí) của các cụm
* Thuật toán:
  + Khởi tạo giá trị ban đầu cho K cụm.
  + Lặp:
    - Với mỗi điểm dữ liệu, gán chúng vào cụm gần nhất (dùng Euclid).
    - Tính toán lại cụm mới bằng trung bình các điểm trong cụm.
  + Tới khi hội tụ.

Tiêu chí hội tụ:

Không thay đổi hoặc thay đổi rất nhỏ ở cụm mới và cụm cũ.

Đạt số lần lặp tối đa.

**c) Triển khai bằng Hadoop Mapreduce**

Để triển khai thuật toán K-means trên Hadoop Mapreduce, ta cần phải xác định rõ chức năng của thành phần Mapper và Reducer cho bài toán.

**Mapper:**

Mục tiêu: Gán các điểm dữ liệu vào cụm K gần nhất

Mã giả:

* Đọc các cụm khởi tạo từ file txt và lưu thành 1 mảng
* Với mỗi điểm dữ liệu:
  + Tính khoảng cách Euclid từ điểm đó tới mỗi cụm.
  + Gán điểm dữ liệu đó cho cụm gần nhất.
  + Xuất ra index của cụm và điểm dữ liệu đó (định dạng: <index\_K>\t<feature\_1>\t<feature\_2>\t...)

**Reducer:**

Mục tiêu: Tính toán lại tọa độ cụm mới dựa trên các điểm dữ liệu đã được gán cụm ở Mapper.

Mã giả:

* Với mỗi cụm:
  + Xử lý đầu vào của Mapper
  + Tính giá trị trung bình của tọa độ cho mỗi cụm dựa trên các điểm dữ liệu thuộc cùng 1 cụm.
  + Xuất ra các tọa độ cụm mới.

Và cuối cùng là kiểm tra xem tọa độ của cụm mới có khác so với cụm cũ không, nếu khác thì tiếp tục. Còn không sẽ dừng lại và kết thúc thuật toán.

Cụ thể có thể xem hình minh họa bên dưới để hình dung 1 cách tổng quát mô hình K-means chạy song song trên Hadoop sẽ như thế nào.

