Chạy CD-HIT với ngưỡng giống nhau 90% (0.9) trên tập all_proteins.fasta

```
cd-hit -i all_proteins.fasta -o clustered_proteins.fasta -c 0.9 -n 5
```

Kết quả sau khi thực hiện lệnh này là một file đầu ra *clustered_proteins.fasta* chứa tập hợp rút gọn các chuỗi protein sao cho không còn các chuỗi gần giống nhau vượt quá ngưỡng 90% về mặt trình tự.

Ngoài ra, CD-HIT cũng sinh ra một file *clustered_proteins.fasta.clstr* chứa thông tin về cách các chuỗi được phân cụm, có thể dùng để truy ngược chuỗi gốc nào thuộc về cụm nào trong quá trình phân tích kết quả.

Sử dụng: Toàn bộ các bước triển khai hệ thống, bao gồm: tạo database, upload lên HDFS, thực thi job Hadoop, và lấy kết quả đầu ra... vẫn được giữ nguyên như khi chạy với dữ liệu thô. Điểm duy nhất thay đổi là tập dữ liệu đầu vào: thay vì sử dụng file all_proteins.fasta chứa chuỗi gốc, sử dụng file clustered_proteins.fasta đã được thực hiện cluster.