# Lab 03: MapReduce hóa với

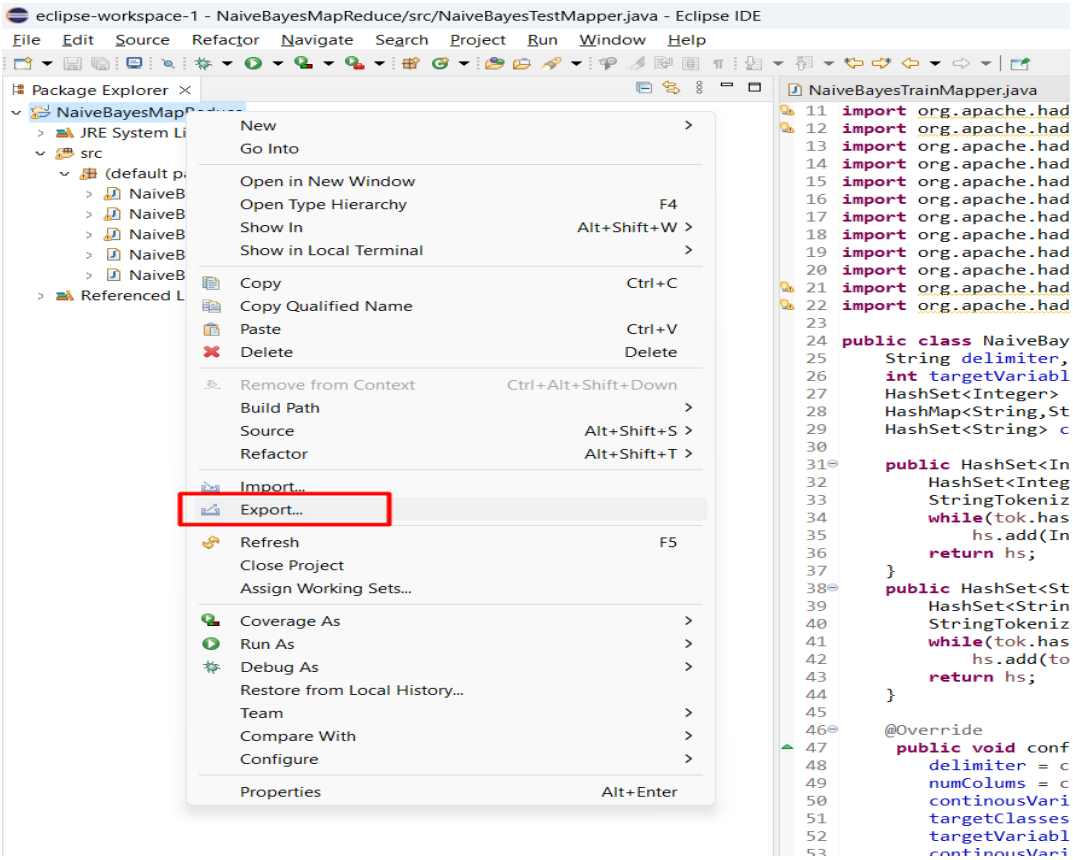
**Mục tiêu:**

- Tìm Hiểu Hadoop MapReduce (Tài liệu trên lớp Tuần 2)
- Mô hình lập trình MapReduce (Tài liệu trên lớp Tuần 2)
- Tìm hiểu MapReduce hóa với các thuật toán xử lý dữ liệu Naïve Bayes, K-Means
- Triển khai xây dựng cài đặt Hadoop MapReduce hóa với các thuật toán xử lý dữ liệu.
- Kiểm tra kết quả thực hiện trên màn hình giao diện ứng dụng, xem tiến trình của tác vụ trong bảng điều khiển và một URL để xem thông tin chi tiết hơn về tác vụ.
- Submit kết quả thực hiện Project lên hệ thống (LMS Canvas, Github, jupyter notebook…)

**Hướng dẫn:**

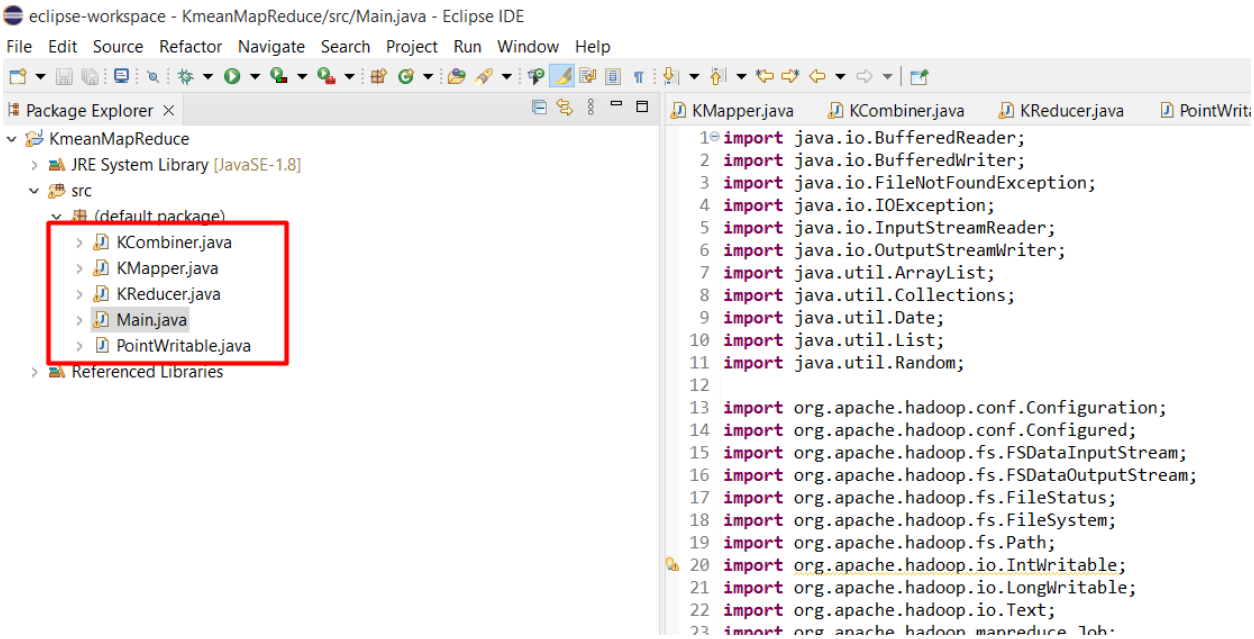
## 1. MapRedure hóa với thuật toán Naïve Bayes

- MapReduce - Phân loại nhị phân Naive Bayes
- Lập trình MapReduce cho bài toán phân loại hoa Iris sử dụng thuật toán Naive Bayes




Dữ liệu: https://www.kaggle.com/datasets/uciml/iris

## 2. MapRedure hóa với thuật toán K-Means




## 3. Bài tập về MapReduce hóa trên Hadoop

Classifier sử dụng MapReduce trên Hadoop Triển khai Naive Bayes được thực hiện bằng MapReduce và triển khai Local Machine, tập dữ liệu là DBPedia. Các bước thực hiện:

1. Naive_Bayes_classifier.ipynb Tập lệnh python trên Jupyter Notebook triển khai bộ phân loại Naive Bayes và tính toán thời gian tính toán và độ chính xác trên tập dữ liệu đào tạo, phát triển và thử nghiệm trên Local Machine.
2. Naive_Bayes_classifier.py Tập lệnh python triển khai bộ phân loại Naive Bayes và tính toán thời gian tính toán và độ chính xác trên tập dữ liệu đào tạo, phát triển và thử nghiệm trên Local Machine.
3. Python_mapreduce.py Tập lệnh python triển khai thuật toán Naive Bayes trên tập dữ liệu đào tạo, thử nghiệm và phát triển và ghi lại độ chính xác tương ứng cùng với thời gian đào tạo thuật toán. Từ điển được chuẩn bị trên nền tảng hadoop mapreduce.
4. mapper.py Tập lệnh python mapper được sử dụng để lập bản đồ bằng cách sử dụng hadoop stream để tạo đầu ra luồng (nhãn, từ, 1).
5. reducer.py Tập lệnh python reducer được sử dụng để lập bản đồ bằng hadoop stream để tạo luồng đầu ra (nhãn, từ, số lượng).
6. dictionary.pickle Tệp Pickle để lưu trữ từ điển để mô hình không phải được đào tạo mỗi lần.
7. log.txt Tệp nhật ký chứa các bản ghi nhật ký hadoop.

Link: https://github.com/dolaram/Naive-Bayes-using-MapReduce