**LAB 4**

***Ứng dụng Spark để triển khai các mô hình học máy (machine learning) phân tán***

## **Spark Machine Learning Library**

## Spark cung cấp thư viện MLLib (Machine Learning Library) cho phép triển khai các dự án machine learning phân tán. Các tính năng của MLLib bao gồm:

## Một số bài toán ML tiêu biểu: phân lớp (classification), hồi qui (regression), phân cụm (clustering), collaborative filtering;

## Xử lý đặc trưng (featurization): trích xuất (feature extraction), biến đổi (transformation), giảm chiều (dimensionality reduction), chọn lọc (feature selection);

## Pipelines: các công cụ xây dựng, tinh chỉnh và đánh giá mô hình ML;

## Các tiện ích: xử lý dữ liệu, lưu/nạp mô hình, v.v..

## Bên cạnh thư viện MLLib (RDD-based API), hiện nay Spark cung cấp thư viện ML mới (DataFrame-based API) với nhiều tính năng dễ sử dụng hơn MLLib.

## **Thuật toán K-means với PySpark**

* Khai báo các thư viện

***import pandas as pd***

***import numpy as np***

***from pyspark.ml.clustering import KMeans***

***from pyspark.sql import SparkSession***

***from pyspark.ml.feature import VectorAssembler***

* Khởi tạo Spark Session

# Create new Spark session

Spark phiên bản 2.0 trở đi cung cấp lớp SparkSession để khởi tạo các chức năng của Spark. Sau khi tạo Spark Session, người dùng có thể lập trình với RDD, DataFrame và Dataset.

**spark = SparkSession.builder.appName("Distributed KMeans Example").getOrCreate()**

* Tạo dữ liệu dạng DataFrame
* Với Spark Context, người dùng phải tạo mới RDD sau đó lập trình xử lý trên RDD. Spark Session cung cấp các API xử lý với các loại dữ liệu khác, gồm cả DataFrame.
* Để nạp dữ liệu iris vào Spark cluster, có thể thực hiện như sau: trước hết đọc dữ liệu đưa vào DataFrame dùng thư viện pandas, sau sử dụng phương phương thức spark.createDataFrame() để nạp dữ liệu vào Spark DataFrame.

**iris\_data =** **spark.createDataFrame(pd.read\_csv('/home/user/Downloads/iris.csv', header=None, names=['sepal-length', 'sepal-width', 'petal-length', 'petal-width','label']))**

**print("First 10 rows:")**

**iris\_data.show(10)ức spark.createDataFrame() để nạp dữ liệu vào Spark DataFrame.**

* Tạo vector đặc trưng từ DataFrame

**assembler = VectorAssembler(inputCols = ["sepal-length", "sepal-width", "petal-length", "petal-width"], outputCol="features")**

**irisFeatures = assembler.transform(iris\_data)**

**irisFeatures.show(5)**

Lúc này dữ liệu đã sẵn sàng để đưa vào huấn luyện mô hình ML (ở đây là K-means). Gỉa sử chọn số cụm là 3, tiến hành huấn luyện mô hình K-means với PySpark bằng các lệnh sau:

# Huấn luyện mô hình K-means với K=3 trên toàn bộ tập dữ liệu

**kmeans = KMeans().setK(3).setSeed(0)**

**model = kmeans.fit(irisFeatures)**

Hiển thị kết quả

**# In ra tâm điểm của các cụm**

**centers = model.clusterCenters()**

**np.set\_printoptions(precision=2)**

**print("Centers learned by Spark ML: ")**

**for center in centers:**

**print(center)**

**spark.stop()**

## **Bài tập: *So sánh với kết quả trên với chạy K-means bằng thư viện Scikit-learn***