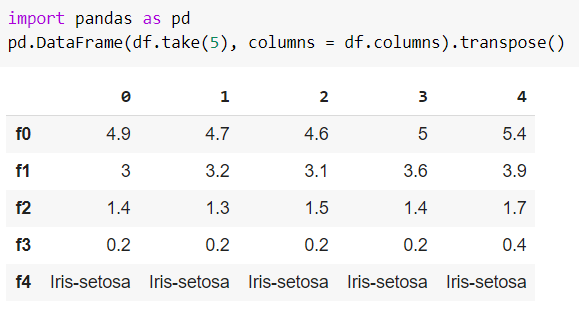
**DATAFRAMES**

DataFrames là một tập hợp phân tán của các hàng dưới các cột được đặt tên. Về mặt khái niệm, nó tương đương với một bảng trong cơ sở dữ liệu quan hệ, một trang tính Excel có tiêu đề cột hoặc một khung dữ liệu trong R / Python, nhưng với các tối ưu hóa phong phú hơn. [1]

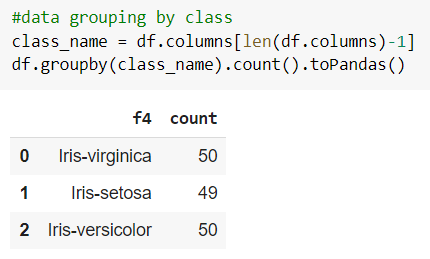


Vì việc sử dụng RDD còn nhiều hạn chế nên DataFrames ra đời với nhiều ưu điểm như:

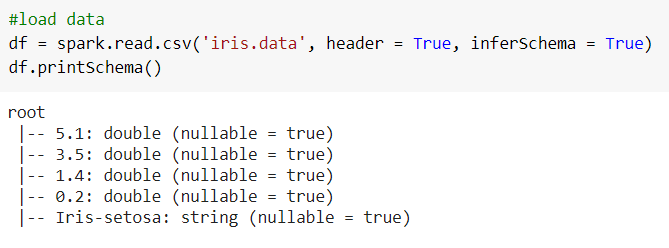
* Tương tự như SQL gồm đầy đủ chức năng như select, where ... đặc biệt là join với các DataFrames khác.
* Sử dụng các method như filter, select để trích xuất dữ liệu theo cột, hàng.
* Thêm cột dễ dàng với UDF (User Defined Function).
* Được hỗ trợ Pivot (Spark 1.6 trở lên) hữu ích cho việc lập bảng biểu, báo cáo. [2]

Tóm lại, ta thấy những lợi ích này chính là tính đơn giản, dễ sử dụng, truy vấn tốt mà DataFrames mang lại.

DataFrames được thiết kế để xử lý một tập hợp lớn dữ liệu có cấu trúc cũng như bán cấu trúc (nhiều định dạng dữ liệu có thể đồng nhất hoặc không đồng nhất). Nó còn hỗ trợ các phương thức phức tạp để lấy và xử lý dữ liệu bao gồm các hoạt động như chọn các hàng, cột và ô theo tên hoặc theo số, lọc ra các hàng, nhóm các dữ liệu… [3]



Dataframes có thể được tạo từ nhiều cách như RDD, hay trực tiếp từ file csv, json.



**TÀI LIỆU THAM KHẢO**

1. <https://ichi.pro/vi/vi-du-ve-viec-su-dung-apache-spark-voi-pyspark-bang-python-267611095265298>
2. https://blog.vietnamlab.vn/xu-ly-du-lieu-voi-spark-dataframe/
3. <https://helpex.vn/article/huong-dan-pyspark-dataframe-gioi-thieu-ve-dataframes-5c6b21e6ae03f628d053c29e>