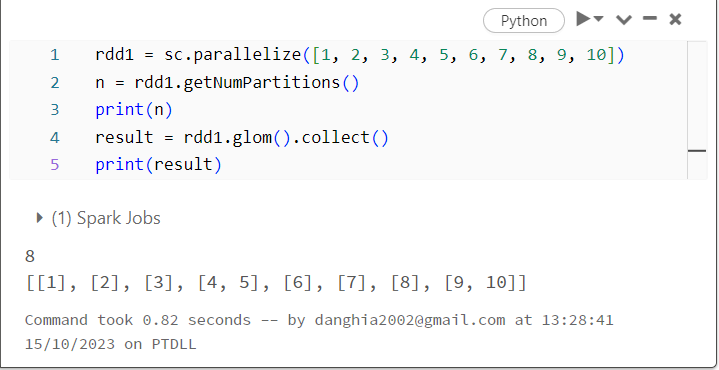
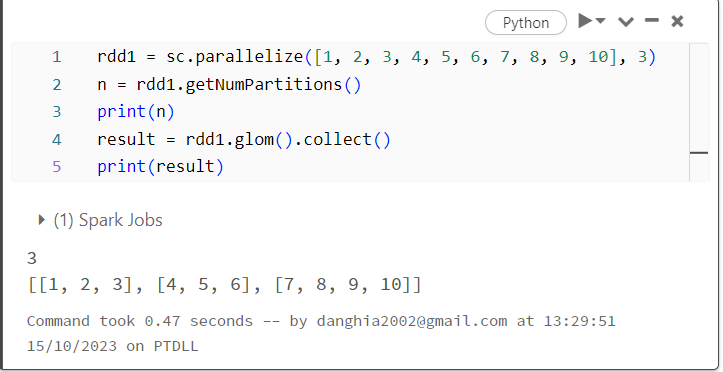
Bài thực hành số 2. Sử dụng RDD

Bài 1. Tạo RDD và phân vùng dữ liệu

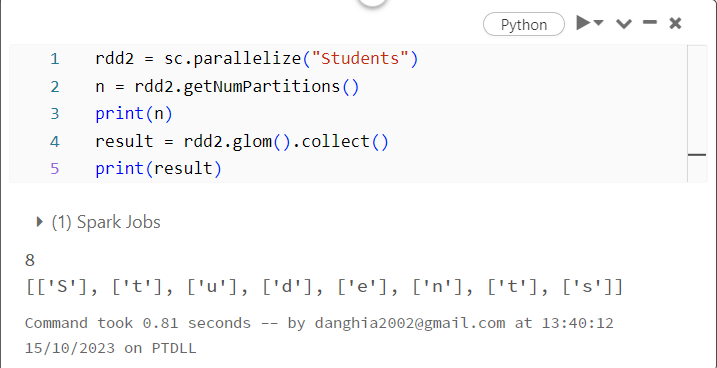
1. Một danh sách các số



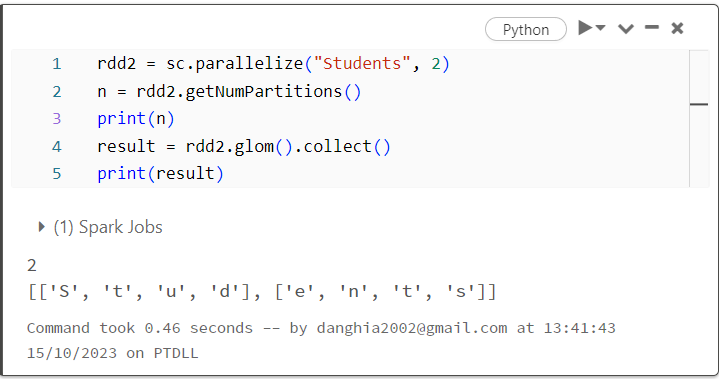
Cố định phân vùng trong chương trình:



1. Một xâu ký tự



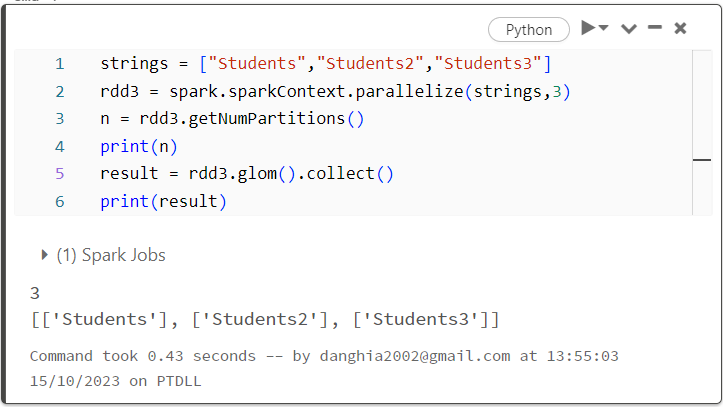
Cố định phân vùng trong chương trình:



1. Một danh Sách các Xâu ký tự



Cố định phân vùng trong chương trình:



1. Một dict



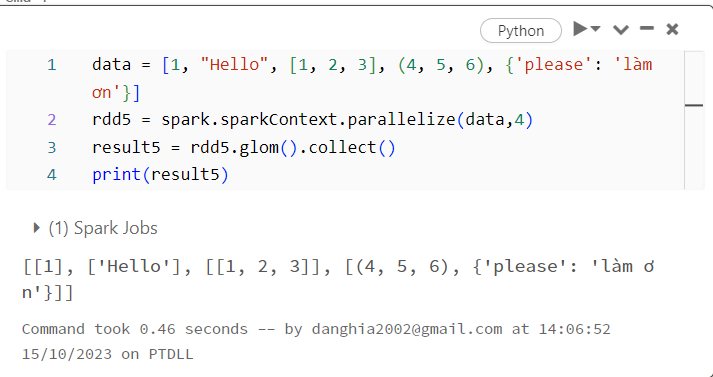
Cố định phân vùng trong chương trình:

Pyspark không hỗ trợ đầy đủ cho kiểu dict, để sử dụng được dict phải chuyển dict sang

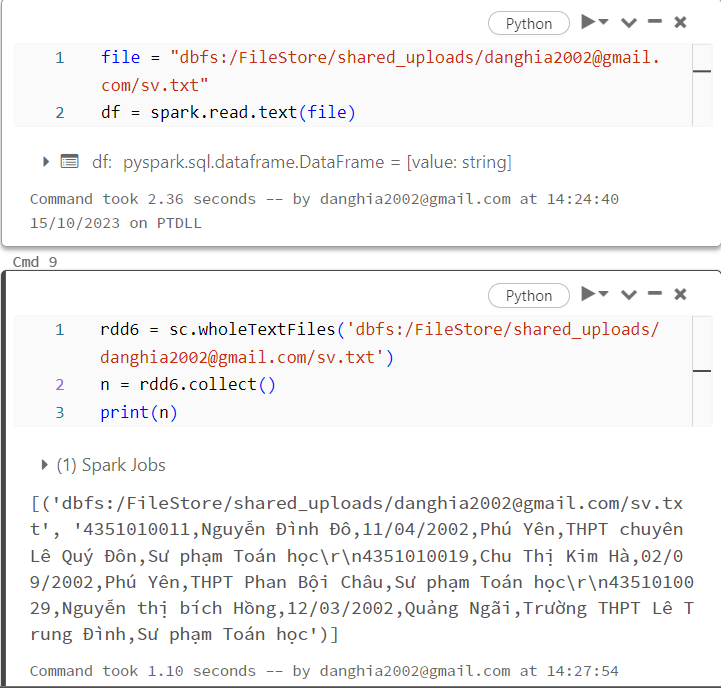
list các tuple.



1. Một tập hợp các đối tượng hỗn hợp gồm số, xâu, list, tuple, dict



1. Một text file



Bài 4. Đếm sinh viên

Cho tệp sinhvien.csv có cấu trúc: mã sinh viên, họ tên, ngày sinh, quê quán, trường THPT, tên ngành học.

Viết chương trình trên Spark đếm số sinh viên có quê quán ở Bình Định.

# Đọc tệp CSV với schema tùy chỉnh

from pyspark.sql.types import StructType, StructField, StringType

# Định nghĩa schema cho DataFrame

custom\_schema = StructType([

StructField("maSV", StringType(), True),

StructField("hoTen", StringType(), True),

StructField("ngaySinh", StringType(), True),

StructField("queQuan", StringType(), True),

StructField("truongTHPT", StringType(), True),

StructField("nganhHoc", StringType(), True)

])

# Đọc tệp CSV với schema tùy chỉnh

df = spark.read.csv("dbfs:/FileStore/shared\_uploads/thuydangthanh2704@gmail.com/SinhVien.csv", schema=custom\_schema, header=False)

# Đếm số sinh viên có quê quán ở Bình Định

count\_binh\_dinh = df.filter(df["queQuan"] == "Bình Định").count()

# In kết quả

print("Số sinh viên có quê quán ở Bình Định:", count\_binh\_dinh)

A screenshot of a computer program

Description automatically generated