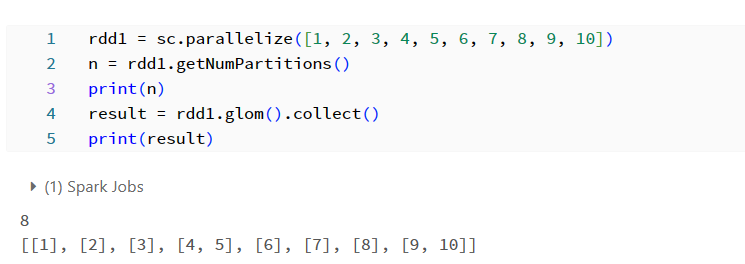
**Bài thực hành số**

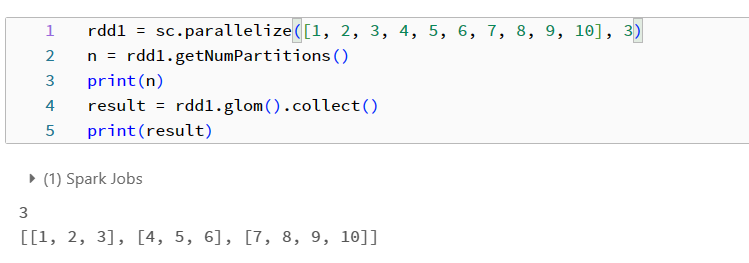
**02**

**Bài 1. Tạo RDD và phân vùng dữ liệu**

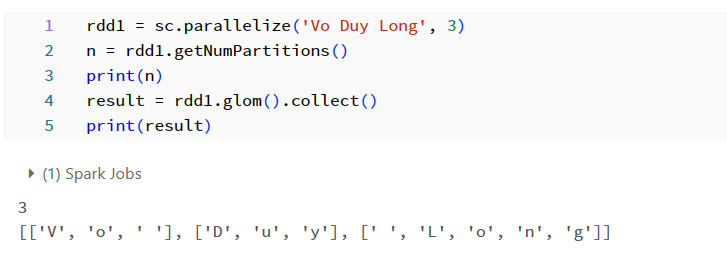
a. Một danh sách các số



- Cố định phân vùng trong chương trình:



b. Một xâu kí tự:



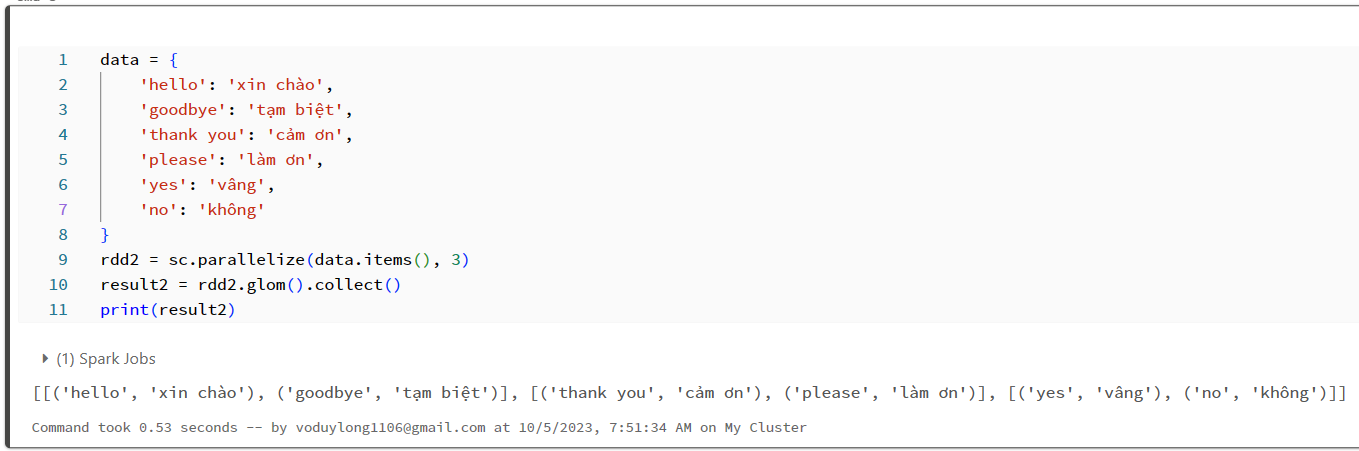
c. Một danh sách các xâu:



d. Một dict:



- Chuyển dict sang một list các tuple



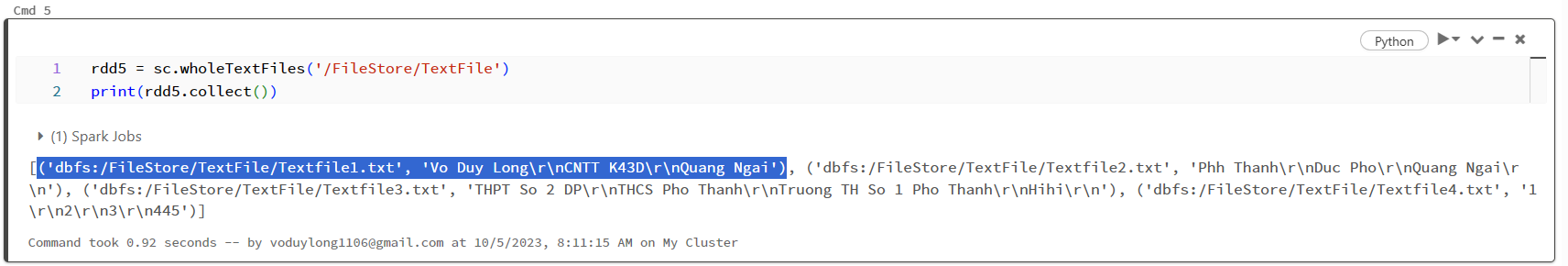
e. Một tập hợp các đối tượng hỗn hợp gồm số, xâu, list, tuple, dict



f. Một text file



g. Một thư mục chứa các text file



**Bài 2. Lọc các RDD**

Tạo một list các số nguyên từ 0 đến 9999 chia thành 5 phân vùng. Sau đó lọc lấy

các số chia hết cho 3 mà không chia hết cho 9 trên từng phân vùng. Xem kết quả các số còn lại ở từng phân vùng. Ghi các số ở từng phân vùng thư mục Bai2.

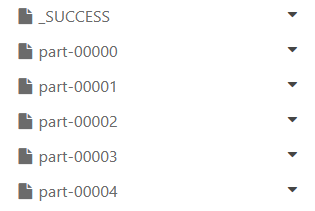
rdd6 = sc.parallelize(range(10000), 5)

rdd7 = rdd6.filter(lambda n: n%3 == 0 and n%9 != 0)

rdd7.saveAsTextFile('/FileStore/SaveFile')

result7 = rdd7.glom().collect()

print(result7)



**Bài 3. Lọc các dòng của tệp văn bản.**

Đọc tệp readme.md trong thư mục spark rồi ghi ra tệp text những dòng có chứa

chữ ‘Spark’.

rdd\_bai3 = sc.textFile('/FileStore/TextFile/README.md')

filter\_spark = rdd\_bai3.filter(lambda line: "Spark" in line)

filter\_spark.saveAsTextFile("/FileStore/SaveFileSparkText")



**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.

rdd\_bai4 = sc.textFile('/FileStore/SinhVien/SinhVien.csv')

count = rdd\_bai4.filter(lambda line: "Bình Định" in line.split(',')[3]).count()

print("Số lượng:", count)



Link bài làm: <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/345358637122346/1394488951951253/5017236833160125/latest.html>