Bài thực hành số 2. Sử dụng RDD

Mục tiêu: làm quen với RDD và các thao tác trên RDD.

Nhắc lại: tạo SparkContex:

```
from pyspark import SparkContext
sc = SparkContext()
```

Bài 1. Tạo RDD và phân vùng dữ liệu

Thực hành việc tạo RDD và phân vùng cho các tập dữ liệu sau:

a) Một danh sách các số

Sử dụng chế độ tự động phân vùng:

```
rdd1 = sc.parallelize([1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
n = rdd1.getNumPartitions()
print(n)
result = rdd1.glom().collect()
print(result)
```

Kết quả:

```
4
[[1, 2], [3, 4], [5, 6], [7, 8, 9, 10]]
```

Cố định phân vùng trong chương trình:

```
rdd1 = sc.parallelize([1, 2, 3, 4, 5, 6, 7, 8, 9, 10], 3)
n = rdd1.getNumPartitions()
print(n)
result = rdd1.glom().collect()
print(result)
```

Kết quả:

```
3
[[1, 2, 3], [4, 5, 6], [7, 8, 9, 10]]
```

b) Một xâu ký tự

c) Một danh sách các xâu ký tự

```

```

d) Một dict

```
rdd2 = sc.parallelize({
    'hello': 'xin chào',
    'goodbye': 'tạm biệt',
    'thank you': 'cảm ơn',
    'please': 'làm ơn',
    'yes': 'vâng',
    'no': 'không'
}, 3)
result2 = rdd2.glom().collect()
print(result2)
```

Kết quả:

```
[['hello', 'goodbye'], ['thank you', 'please'], ['yes', 'no']]
```

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

```
data = {
    'hello': 'xin chào',
    'goodbye': 'tạm biệt',
    'thank you': 'cảm ơn',
    'please': 'làm ơn',
    'yes': 'vâng',
    'no': 'không'
}
rdd3 = sc.parallelize(data.items(), 3)
result3 = rdd3.glom().collect()
print(result3)
```

Kết quả:

```
[[('hello', 'xin chào'), ('goodbye', 'tạm biệt')],
 [('thank you', 'cảm ơn'), ('please', 'làm ơn')],
 [('yes', 'vâng'), ('no', 'không')]]
```

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

f) Một text file

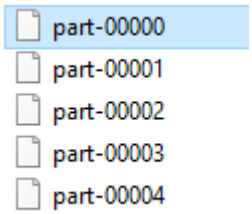
g) Một thư mục chứa các text file

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.

```
rdd4 = sc.parallelize(range(1000), 5)
rdd5 = rdd4.filter(lambda n: n%3 ==0 and n%9 != 0)
rdd5.saveAsTextFile('F:/Spark/Bai2')
result4 = rdd5.glom().collect()
print(result4)
```

Kết quả sẽ ghi ra 5 file:



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'.

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.
