BỘ GIÁO DỤC VÀ ĐÀO TẠO



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

NHẬP MÔN DỮ LIỆU LỚN

**BÁO CÁO BÀI TẬP MÔN HỌC**

**NỘI DUNG: LAB 02 - ADVANCED HADOOP & SPARK STRUCTURED APIs**

**Giảng viên hướng dẫn** : Huỳnh Lâm Hải Đăng

**Lớp** : CQ2022/21

**Sinh viên thực hiện** : Trương Tiến Anh- 22120017

Nguyễn Minh Cường- 22120043

Đinh Viết Lợi- 22120188

Nguyễn Trần Lợi-22120190

*Hồ Chí Minh, ngày 8 tháng 4 năm 2025*

# MỤC LỤC

[MỤC LỤC 1](#_Toc195009574)

[PHẦN 1: BÁO CÁO BÀI TẬP 2](#_Toc195009575)

[I. Exercise 2.1- Calculate revenue in the last 3 days for each country 2](#_Toc195009576)

[1. Ý tưởng thực hiện 2](#_Toc195009577)

[2. Mô tả chi tiết 2](#_Toc195009578)

[II. Exercise 2.2- Calculate the number of products sold in the last 7 days of each SKU, report every Monday 2](#_Toc195009579)

[1. Ý tưởng thực hiện 2](#_Toc195009580)

[2. Mô tả chi tiết 2](#_Toc195009581)

[III. Exercise 2.3- Find pairs of overlapping shapes 2](#_Toc195009582)

[1. Ý tưởng thực hiện 2](#_Toc195009583)

[2. Mô tả chi tiết 2](#_Toc195009584)

[IV. References 2](#_Toc195009585)

[PHẦN 2: BÁO CÁO NHÓM 3](#_Toc195009586)

# PHẦN 1: BÁO CÁO BÀI TẬP

## Exercise 2.1- Calculate revenue in the last 3 days for each country

### **Ý tưởng thực hiện**

### **Mô tả chi tiết**

### **Hướng dẫn thực thi chương trình**

## Exercise 2.2- Calculate the number of products sold in the last 7 days of each SKU, report every Monday

### **Ý tưởng thực hiện**

* + Report date thực hiện các báo cáo “in the last 7 days”, tức là ngày thứ hai mỗi tuần sẽ thực hiện bản báo cáo của 7 ngày tuần trước- từ sáng thứ 2 tuần trước tới tối chủ nhật tuần trước.
  + Với mỗi ngày trong input, ta tính ngày thực hiện báo cáo của ngày này.
  + Ta lọc ra các record có “Status” là shipped.
  + Nhóm các record có chung ngày báo cáo và SKU để thu được kết quả.

### **Mô tả chi tiết**

* + Tạo Park Session và import các thư viện cần thiết.
  + Đọc asr.csv và kiểm tra tính chất của dữ liệu về kiểu dữ liệu, các record.
  + Sắp xếp lại các record theo thứ tự của Date, SKU và lọc các record có Status là Shipped.
  + Tính ngày thực hiện report đối với mỗi record từ cột “Date”, các ngày thuộc cùng một tuần sẽ có cùng “report\_date”
  + GroupBy 2 cột “report\_date” và “SKU”, tính tổng cột Qty bằng hàm sum và lưu tên cột giá trị mới này là total\_quantity.

## Exercise 2.3- Find pairs of overlapping shapes

### **Ý tưởng thực hiện**

* + Xem các điểm tạo thành tứ giác bao lồi.
  + Chứng minh các tứ giác này là hình chữ nhật.
  + Sau đó lọc lấy 4 giá trị x\_min, x\_max, y\_min và y\_max.
  + Thực hiện kiểm tra overlap giữa hai hình bằng cách kiểm ra x\_min\_a > x\_max\_b hoặc y\_min\_a > y\_min\_b hoặc ngược lại.

### **Mô tả chi tiết**

* + Import các thư viện cần thiết và tạo Spark Session.
  + Đọc file, kiểm tra kiểu dữ liệu và chuyển shape\_id thành dạng số.
  + Viết hàm kiểm tra là hình chữ nhật không bằng cách tính các tích vô hướng của các vector cạnh.
  + Lọc lấy 4 giá trị x\_min, x\_max, y\_min và y\_max của từng hình.
  + Hợp hai bảng dữ liệu lại với nhau sao cho shape\_id\_2 > shape\_id\_1 và đổi tên cột thành "shape\_1", "x\_min\_1", "x\_max\_1", "y\_min\_1", "y\_max\_1", "shape\_2", "x\_min\_2", "x\_max\_2", "y\_min\_2", "y\_max\_2".
  + Viết hàm kiểm tra overlap và thực hiện kiểm tra sau đó lọc lấy các dòng có giá trị "true" ở cột is\_overlap. Cách thức kiểm tra overlap dựa vào việc kiểm tra vị trí tương đối của min max hoành độ và tung độ từng hình.
  + Chọn cột shape\_1 và shape\_2 sau đó xuất file.

## References

1. Apache Spark, ‘Quickstart: DataFrame’ [Trực tuyến]. Đường dẫn: <https://spark.apache.org/docs/latest/api/python/getting_started/quickstart_df.html>

# PHẦN 2: BÁO CÁO NHÓM

* + Phương thức trao đổi và làm việc:
    - Nhóm yêu cầu tất cả thành viên đều tham gia giải tất cả bài tập nhằm mục đích bao quát trường hợp, nắm bắt lỗi và học tập hiệu quả.
    - Nhóm tổ chức họp nhóm trao đổi sau 4 ngày làm việc cá nhân kết hợp trao đổi liên tục hằng ngày.
    - Phiên họp nhằm kiểm tra kết quả, ý tưởng giải quyết cho từng vấn đề.
    - Giao công việc cho từng thành viên trên workspace của nhóm sau khi thống nhất kết quả.
  + Bảng phân công công việc:

|  |  |  |
| --- | --- | --- |
| Thành viên | Công việc | Lưu ý |
| Trương Tiến Anh | Thực hiện exercise 1 và tóm tắt ý tưởng, quy trình. |  |
| Nguyễn Minh Cường | Thực hiện exercise 2 và tóm tắt ý tưởng, quy trình. |  |
| Đinh Viết Lợi | Thực hiện exercise 1 bằng spark. Điều hành công việc nhóm, viết báo cáo. | Kết quả bài 1 do có sự khác biệt về kết quả giữa các thành viên nên cần thực hiện bằng nhiều phương pháp và thống nhất diễn giải. |
| Nguyễn Trần Lợi | Thực hiện exercise 3 và tóm tắt ý tưởng, quy trình. |  |