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

PHẦN 1: BÁO CÁO BÀI TẬP

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

1. Ý tưởng thực hiện

- Bài toán áp dụng kĩ thuật khái niệm cửa sổ trượt (sliding window) để tính tổng doanh thu trong 3 ngày liên tiếp (ngày hiện tại và 2 ngày trước đó) cho mỗi danh mục. Cửa sổ trượt từng ngày, từ ngày đầu tiên trong dữ liệu đến ngày cuối cùng.
- Mỗi khung cửa sổ sẽ bao gồm ngày hiện tại (record), ngày hôm qua và ngày hôm trước miễn là có số liệu. Do đó output sẽ có sự xuất hiện của hai timestamp không xuất hiện ở input.
- Ta thực hiện trượt cửa sổ để xác định ngày record, lấy thêm 2 ngày trước đó để filter và lọc ra các giá trị cần thiết, kết hợp với các filter status và category.

2. Mô tả chi tiết

- Mapper (SlidingWindowMapper):
  - Đọc dữ liệu với từng dòng từ file asr.csv.
  - Parse cột “Date” từ định dạng MM-dd-yy sang dd/MM/yyyy.
  - Lọc các đơn hàng có thuộc tính Status chứa “shipped”.
  - Tạo key dạng date, category và values là amount.
- Reducer (SlidingWindowReducer):
  - Giai đoạn Reduce:
    - Nhận key (date, category) và values (amount).
    - Tổng hợp amount theo category và date, lưu vào Tree<Map, Map<String,Double>>.
  - Giai đoạn Cleanup:
    - Sau khi xử lý hết dữ liệu, duyệt từ ngày đầu tiên đến ngày cuối cùng.
    - Với mỗi report_date, gọi emitWindowData để tính tổng và xuất kết quả.
  - Phương thức emitWindowData:
    - Lấy window 3 ngày (ngày hiện tại và 2 ngày trước).
    - Tính tổng revenue cho mỗi danh mục trong window.
    - Sắp xếp danh mục và xuất theo định dạng yêu cầu.
- Driver (SlidingWindowRevenue):
  - Cấu hình job MapReduce với các class Mapper, Reducer.
  - Nhận tham số đầu vào và đầu ra từ command line.

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

1. 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.

2. 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.

III. Exercise 2.3- Find pairs of overlapping shapes

1. 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.

2. 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.

IV. 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.	