CSC12107 – HTTT PHỤC VỤ TRÍ TUỆ KINH DOANH

ĐỒ ÁN MÔN HỌC

DOANMH: XÂY DỰNG VÀ KHAI THÁC KDL

I. Thông tin chung

Mã số bài tập:	DOANMH
Thời lượng dự kiến:	10 tuần
Deadline nộp bài:	31/12/2021
Hình thức:	Bài tập nhóm
Hình thức nộp bài:	Nộp qua Moodle môn học
GV phụ trách:	Hồ Thị Hoàng Vy, Tiết Gia Hồng
Thông tin liên lạc với GV:	hthvy@fit.hcmus.edu.vn, tghong@fit.hcmus.edu.vn

II. Chuẩn đầu ra cần đạt

Bài tập này nhằm mục tiêu đạt được các chuẩn đầu ra sau:

- G3.3 Thiết kế lược đồ chuẩn hoá, đa chiều (sao, bông tuyết) dựa vào dữ liệu hệ thống tác vụ và yêu cầu phân tích từ tình huống cho trước
- G5.1 Triển khai quy trình ETL để rút trích dữ liệu từ nhiều nguồn, biến đổi, làm sạch dữ liệu, nạp dữ liệu vào KDL sử dụng công cụ SSIS
- G5.2 Xây dựng KDL đa chiều sử dụng công cụ SSAS và giải thích được lựa chọn phép toán OLAP phù hợp đối với 1 số yêu cầu phân tích.
- G5.3 Sử dụng một số công cụ biểu diễn dữ liệu (SSRS, powerBI, excel…) để biểu diễn kết quả phân tích, khai thác được (report, dashboard…)
- G5.4 Sử dụng SSAS và áp dụng các kỹ thuật mining tích hợp để thực hiện khai thác dữ liệu từ KDL xây dựng được.

III. Các yêu cầu & quy định chi tiết cho bài nộp

Xây dựng và phân tích dữ liệu về Covid-19 trong các năm 2020 - 2022.

- **Mô tả dữ liệu:** Mô tả ý nghĩa các thuộc tính của các nguồn dữ liệu sau:

Tên thuộc tính	Mô tả
Date	Ngày báo cáo
PHU ID	Định danh của đơn vị chăm sóc y tế cộng đồng
At least one dose_cumulative	Số người tiêm được ít nhất 1 mũi
Second_dose_cumulative	Số người tiêm được 1 mũi

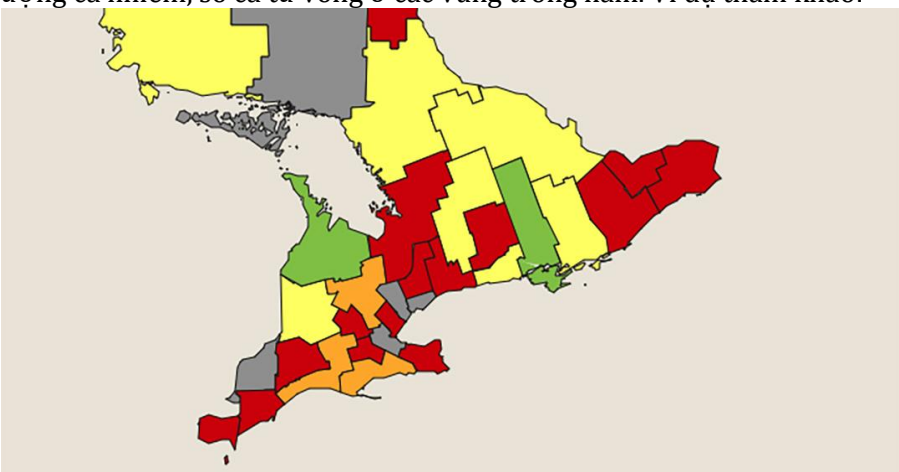
fully_vaccinated_cumulative	Số người tiêm đủ vaccin. Tiêm đầy đủ nghĩa là: - Tiêm 1 mũi Janssen (Johnson & Johnson) - Tiêm 2 mũi trong danh mục vaccin được Bộ y tế Canada phê duyệt - Tiêm 1 mũi trong danh mục được Bộ ý tế phê duyệt + 1 mũi trong danh mục không được phê duyệt - Tiêm 3 mũi vaccin thuộc loại bất kỳ
third_dose_cumulative	Số người tiêm được 3 mũi (tập con của số người tiêm đủ)
Reporting_PHU	Các PHU được báo cáo
Reporting_PHU_Address	Địa chỉ PHU được báo cáo
Reporting_PHU_City	Thành phố của các PHU được báo cáo.
Reporting_PHU_Postal_Code	Mã bưu điện của PHU được báo cáo
Reporting_PHU_Latitude	Vĩ tuyến PHU
Reporting_PHU_Longitude	Kinh tuyến PHU
outbreak_group	Cơ sở bùng phát dịch: - 1 Congregate Care - Chăm sóc cộng đồng - 2 Congregate Living - Lưu trú cộng đồng - 3 Education - Giáo dục - 4 Workplace - Nơi làm việc - 5 Recreational - Cơ sở giải trí - 6 Other/Unknown - Không xác định
number_ongoing_outbreaks	Số đợt bùng phát đang diễn ra
row_id	Mã dòng
age_group	Nhóm tuổi, được phân loại gồm: - 5 to 11 years old - 12 to 17 year olds - 18 to 29 years old - 30 to 39 years old - 40 to 49 years old - 50 to 59 years old - 60 to 69 years old - 70 to 79 years old - 80 years and older - Adults_18plus - Ontario_12plus - Ontario_5plus - Undisclosed_or_missing
gender	Giới tính bệnh nhân
exposure	Phơi nhiễm - Outbreak - Bùng phát - Close Contact - Liên hệ chặt chẽ - Not Reported - Không được báo cáo

	- Travel-Related - Du lịch
case_status	Trạng thái ca nhiễm - Recovered - Phục hồi - Deceased - Tử vong - Active - Điều trị tích cực
outcome	Kết quả: - Resolved - Điều trị - Fatal - Tử vong
specimenDate	Ngày lấy mẫu
TestReported Date	Ngày trả kết quả
CaseAcquisition info	Thông tin ca nhiễm: - CC: dương tính xác định được nguồn lây (closed contact) - No known Epi-link: dương tính không rõ nguồn lây - OB: bùng phát (Outbreak) - Travel
AccurateEpisode Dt	Ngày khởi phát
OutbreakRelated	Có liên quan đến đợt bùng phát

Tên tập tin	**Mô tả**
Compiled_COVID-19_Case_Deatails	Dữ liệu ca nhiễm của tất cả các tỉnh bang ở canada
Cases Report	Dữ liệu ca nhiễm của tỉnh bang Ontario
Vaccines_by_age_phu	Dữ liệu tiêm vắc-xin tại các đơn vị chăm sóc sức khỏe của Ontario
Public health unit	Dữ liệu các đơn vị chăm sóc sức khỏe của Ontario
ongoing_outbreadks_phu	Dữ liệu về việc bùng phát dịch tại các đơn vị chăm sóc sức khỏe của Ontario
Public Health Units GROUP	Nhóm các PHU

- **Thiết kế kho dữ liệu (KDL), tổng hợp, nạp dữ liệu các nguồn vào KDL và thiết kế, xây dựng Cube**:
  Gợi ý:
  - Mapping các nguồn dữ liệu trên và đề xuất giải pháp xây dựng **Geography** dimension với phân cấp: City > PHU_Group > PHU
  - Chuyển đổi dữ liệu ngày tháng sao cho có thể tạo được Date dimension với phân cấp chiều: Year > Quarter > Month > Day
  - Xác định và thiết kế các phân cấp chiều khác để đáp ứng yêu cầu OLAP và report

- **<u>OLAP và Report</u>**:
  1. Thống kê **Số ca nhiễm, số ca tử vong, số ca phục hồi** của dịch Covid-19 theo từng **PHU** trong từng năm.
  2. Thống kê **Mức Độ Nghiêm Trọng** (*tiêu chí nghiêm trọng sinh viên tự định nghĩa*) của dịch Covid-19 theo **PHU** và theo các **Quý trong từng năm.**
  3. Thống kê tổng số người tử vong theo **Giới Tính và Nhóm Tuổi** theo các năm.
  4. Thống kê số ca nhiễm, tử vong theo **Mức Độ Nghiêm Trọng** theo **Ngày Trong Tháng** của các năm.
  5. Thống kê số ca nhiễm, tử vong theo **Mức Độ Nghiêm Trọng, khu vực** (PHU_Group, City)**,** và **số người đã được tiêm vaccin** trong các năm.
  6. Thống kê số ca nhiễm theo **Mức Độ Nghiêm Trọng, nhóm bùng phát** của từng khu vực trong các năm
  7. Sinh viên tự thiết kế những bảng thống kê khác để có thêm nhiều chiều đánh giá số ca nhiễm và tử vong ở Ontario.
  8. Xây dựng đồ thị/ biểu đồ cho các bảng thống kê ở trên.
  9. [Data Visualization] Dùng regional map để biểu diễn trực quan (bằng màu sắc) số lượng ca nhiễm, số ca tử vong ở các vùng trong năm. Ví dụ tham khảo:



- **<u>Data Mining</u>:** Gợi ý:
  - Sử dụng thuật toán mining để xác định các luật (pattern), ví dụ ở vùng nào, vào thời điểm nào, nhóm tuổi nào, nhóm người nào,... thường dễ nhiễm, tử vong.
  - Sinh viên tự đề xuất các yêu cầu phân tích khác, lựa chọn mô hình phù hợp.

IV. Cách đánh giá

- Vấn đáp giữa kỳ: ETL process (data flow, data cleaning, ETL data from source to DW)
- Vấn đáp cuối kỳ: Project hoàn chỉnh (khai thác KDL với report, olap, tạo job tự động định kỳ thực hiện ETL)

V. Tài liệu tham khảo

VI. Các quy định khác

- Sinh viên nộp project lên drive và nộp link lên moodle nếu project vượt quá dung lượng cho phép
- Project bao gồm:

- .Doc: file báo cáo phân tích và thiết kế KDL cùng các phân tích khai thác dữ liệu, kết quả thực hiện, thông tin phân công và kết quả đạt được mỗi thành viên
- Source: script tạo csdl, project ETL, mining…
- Minh chứng làm việc của từng thành viên trong nhóm