Sinh viên thực hiện: 19127304 – Trần Khải Trúc

19127471 – Lương Kiếm Minh

19127649 – Tô Thanh Tuấn

GV phụ trách: Cô Hồ Thị Hoàng Vy

**Đồ án/bài tập môn học – HTTT PHỤC VỤ TRÍ TUỆ KINH DOANH**

**HỌC KỲ I – NĂM HỌC 2021-2022**

**HỆ THỐNG THÔNG TIN**

**PHỤC VỤ TRÍ TUỆ KINH DOANH**

**BỘ MÔN HỆ THỐNG THÔNG TIN – KHOA CÔNG NGHỆ THÔNG TIN**

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TPHCM**

**Đồ án thực hành**

**BẢNG THÔNG TIN CHI TIẾT NHÓM**

|  |  |  |  |
| --- | --- | --- | --- |
| **Mã nhóm:** | 19HTTT2#05 | | |
| **Số lượng:** | **3 thành viên** | | |
| **MSSV** | **Họ tên** | **Email** | **Điện thoại** |
| 19127304 | Trần Khải Trúc | 19127304@student.hcmus.edu.vn |  |
| 19127471 | Lương Kiếm Minh | 19127474@student.hcmus.edu.vn |  |
| 19127649 | Tô Thanh Tuấn | 19127649@student.hcmus.edu.vn |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Bảng phân công & đánh giá hoàn thành công việc** | | | | |
| **Công việc thực hiện** | **Thành viên thực hiện** | **Phần trăm công việc** | **Đánh giá của nhóm** |
| **Yêu cầu 1:**  - Mô tả dữ liệu  - Biến đổi dữ liệu  - Làm sạch dữ liệu  - Chuẩn bị nguồn dữ liệu  **Yêu cầu 2:**  - Thiết kế database Stage  - Thiết kế database NDS  **Yêu cầu 3:**  - Nạp dữ liệu Source to Stage  - Nạp dữ liệu Stage to NDS | 19127304 – Trần Khải Trúc | 50% | 10/10 |
| **Yêu cầu 1:**  - Mô tả dữ liệu  - Chuẩn bị nguồn dữ liệu  **Yêu cầu 2:**  - Thiết kế database Stage  - Thiết kế database NDS  **Yêu cầu 3:**  - Nạp dữ liệu Stage to NDS (demo) | 19127471 – Lương Kiếm Minh | 30% | 8/10 |
| **Yêu cầu 1:**  - Mô tả dữ liệu  **Yêu cầu 2:**  - Thiết kế database Stage  - Thiết kế database NDS | 19127649 – Tô Thanh Tuấn | 20% | 6/10 |

**Mục lục**

[**A.** **Yêu cầu của Đồ án/Bài tập** 3](#_Toc121312601)

[**B.** **Kết quả** 3](#_Toc121312602)

[**Yêu cầu 1: Biến đổi, làm sạch dữ liệu** 3](#_Toc121312603)

[**1.** **Mô tả dữ liệu** 3](#_Toc121312604)

[**2.** **Biến đổi dữ liệu** 5](#_Toc121312605)

[**3.** **Làm sạch dữ liệu** 6](#_Toc121312606)

[**4.** **Chuẩn bị nguồn dữ liệu** 7](#_Toc121312607)

[**Yêu cầu 2: Thiết kế kho dữ liệu** 8](#_Toc121312608)

[**1.** **Database Stage** 8](#_Toc121312609)

[**2.** **Database NDS** 9](#_Toc121312610)

[**Yêu cầu 3: Nạp dữ liệu các nguồn vào KDL** 10](#_Toc121312611)

[**1.** **Source to Stage** 10](#_Toc121312612)

[**2.** **Stage to NDS** 16](#_Toc121312613)

# **Yêu cầu của Đồ án/Bài tập**

|  |  |
| --- | --- |
| **Loại bài tập** | **Lý thuyết 🗹 Thực hành 🗹 Đồ án Bài tập** |
| **Ngày bắt đầu** | **31/10/2022** |
| **Ngày kết thúc** | **03/12/2022** |

# **Kết quả**

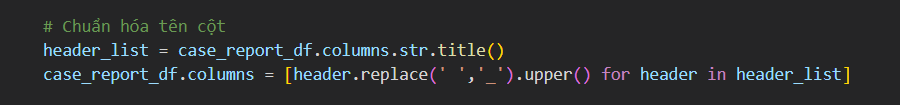
# **Yêu cầu 1: Biến đổi, làm sạch dữ liệu**

## **Mô tả dữ liệu**

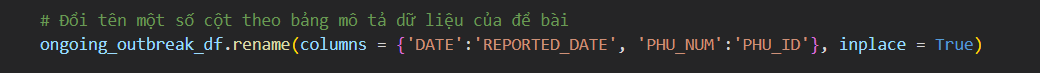
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CASE REPORT** | **Dữ liệu ca nhiễm của tỉnh bang** | | | |
| **Tên thuộc tính** | **Ý nghĩa thuộc tính** | | | |
| Outcome | Kết quả ca nhiễm | | | |
| Age Group | Nhóm tuổi | | | |
| Gender | Giới tính bệnh nhân | | | |
| Reporting PHU | Các cơ sở y tế được báo cáo | | | |
| Specimen Date | Ngày lấy mẫu | | | |
| Reported Date | Ngày báo cáo | | | |
| Reporting PHU City | Thành phố của các cơ sở y tế được báo cáo | | | |
| Test Reported Date | Ngày trả kết quả | | | |
| Case Acquisition info | Thông tin ca nhiễm | | | |
| Accurate Episode DT | Ngày khởi phát | | | |
| Reporting PHU Address | Địa chỉ cơ sở y tế được báo cáo | | | |
| Reporting PHU Website | Địa chỉ trang web của cơ sở y tế được báo cáo | | | |
| Outbreak Related | Có liên quan đến đợt bùng phát | | | |
| Reporting PHU Latitude | Vĩ tuyến của cơ sở y tế | | | |
| Reporting PHU Longitude | Kinh tuyến của cơ sở y tế | | | |
| Reporting PHU Postal Code | Mã bưu điện của cơ sở y tế được báo cáo | | | |
| **CASE DETAIL** | **Dữ liệu ca nhiễm của tất cả các tỉnh bang** | | | |
| **Tên thuộc tính** | **Ý nghĩa thuộc tính** | | | |
| Object ID | Mã chi tiết ca nhiễm | | | |
| Row ID | Mã dòng | | | |
| Reported Date | Ngày báo cáo | | | |
| Reporting PHU | Các cơ sở y tế được báo cáo | | | |
| Age Group | Nhóm tuổi | | | |
| Gender | Giới tính bệnh nhân | | | |
| Exposure | Tình trạng phơi nhiễm | | | |
| Case Status | Trạng thái ca nhiễm | | | |
| Province | Khu vực cơ sở y tế | | | |
| **ONGOING OUTBREAK** | | **Dữ liệu về việc bùng phát dịch tại các đơn vị chăm sóc**  **sức khỏe** | | |
| **Tên thuộc tính** | | **Ý nghĩa thuộc tính** | | |
| Reported Date | | Ngày báo cáo | | |
| PHU ID | | Định danh của đơn vị chăm sóc y tế cộng đồng | | |
| Outbreak Group | | Cơ sở bùng phát dịch | | |
| Number Ongoing Outbreaks | | Số đợt bùng phát đang diễn ra | | |
| **PUBLIC HEALTH UNIT** | | **Dữ liệu các đơn vị chăm sóc sức khỏe** | | |
| **Tên thuộc tính** | | **Ý nghĩa thuộc tính** | | |
| PHU ID | | Định danh của đơn vị chăm sóc y tế cộng đồng | | |
| Reporting PHU | Các cơ sở y tế được báo cáo | | | |
| Reporting PHU Address | Địa chỉ cơ sở y tế được báo cáo | | | |
| Reporting PHU City | Thành phố của các cơ sở y tế được báo cáo | | | |
| Reporting PHU Postal Code | Mã bưu điện của cơ sở y tế được báo cáo | | | |
| Reporting PHU Website | Địa chỉ trang web của cơ sở y tế được báo cáo | | | |
| Reporting PHU Latitude | Vĩ tuyến của cơ sở y tế | | | |
| Reporting PHU Longitude | Kinh tuyến của cơ sở y tế | | | |
| **PHU GROUP** | | **Nhóm các PHU** | | |
| **Tên thuộc tính** | | **Ý nghĩa thuộc tính** | | |
| Reporting PHU Group | Nhóm các cơ sở y tế | | | |
| Reporting PHU City | Thành phố của các cơ sở y tế được báo cáo | | | |
| Reporting PHU | Các cơ sở y tế được báo cáo | | | |
| **VACCINES** | | | | **Dữ liệu tiêm vaccine tại các đơn vị chăm sóc sức khỏe** |
| **Tên thuộc tính** | | | | **Ý nghĩa thuộc tính** |
| Reported Date | | | | Ngày báo cáo |
| PHU ID | | | | Định danh của đơn vị chăm sóc y tế cộng đồng |
| Age Group | | | Nhóm tuổi | |
| At least one dose cumulative | | | Số người tiêm được ít nhất 1 mũi | |
| Second dose cumulative | | | Số người tiêm được 2 mũi | |
| Fully vaccinated cumulative | | | Số người tiêm đủ vaccine | |
| Third dose cumulative | | | Số người tiêm được 3 mũi | |

## **Biến đổi dữ liệu**

* Sử dụng thư viện pandas của python để biến đổi dữ liệu cho khớp các bảng với nhau và làm sạch các dòng dữ liệu bị trùng lắp.
* ***Chuẩn hóa tên cột***
* Ví dụ ở bảng Case Report, thay những khoảng trống bằng dấu gạch dưới để tránh bị lỗi khi gọi tên. Thực hiện tương tư với các bảng còn lại.



* Ví dụ ở bảng Ongoing Outbreak, đổi tên một số cột theo chuẩn của đề bài để khớp với cột ở các bảng khác. Thực hiện tương tư với các bảng còn lại.



* ***Chuẩn hóa dữ liệu***
* Ví dụ ở bảng Case Report, đổi giá trị của một số dòng dữ liệu theo chuẩn của đề bài để khớp với các dữ liệu có **cùng ý nghĩa nhưng khác giá trị** ở những bảng khác . Việc chuẩn hóa dữ liệu này **không làm thay đổi ý nghĩa** của dữ liệu. Thực hiện tương tư với các bảng còn lại.

Graphical user interface, text

Description automatically generated

* Đối với kiểu dữ liệu ngày tháng sẽ có nhiều định dạng khác nhau có thể gây khó khăn khi thực hiện tìm kiếm, truy suất hoặc cũng có thể gây lỗi khi chạy chương trình nạp dữ liệu từ nguồn vào KDL. Nên thực hiện biến đổi định dạng của các cột có kiểu dữ liệu ngày tháng để dữ liệu ở các nguồn được đồng bộ.

Text

Description automatically generated

## **Làm sạch dữ liệu**

* Sử dụng thư viện pandas của python để biến đổi dữ liệu cho khớp các bảng với nhau và làm sạch các dòng dữ liệu bị trùng lắp.
* Với các bảng có dữ liệu bị trùng lắp sẽ gây lỗi khi nạp dữ liệu. Vì vậy cần thực hiện kiểm tra trùng lắp và xử lí.

Text

Description automatically generated with medium confidence

* Xử lý các dòng bị trùng lắp dữ liệu bằng cách xóa.

Text

Description automatically generated

## **Chuẩn bị nguồn dữ liệu**

* Đưa các nguồn dữ liệu đã được xử lí và làm sạch về database để chuẩn bị cho việc nạp nguồn dữ liệu vào KDL.

# **Yêu cầu 2: Thiết kế kho dữ liệu**

## **Database Stage**

Graphical user interface, application, table

Description automatically generated

* Database ở giai đoạn Stage là ánh xạ 1-1 với các nguồn nên cấu trúc của bảng ở Stage tương tự như các bảng ở Source System.
* Ở Stage không có bất kì ràng buộc hay chỉ mục nào trên các thuộc tính của bảng.

## **Database NDS**

Graphical user interface, application, table

Description automatically generated

* Dữ liệu khi vào NDS phải là dữ liệu sạch đã được chuẩn hóa, dùng để phân tích dữ liệu. Thông thường lược đồ ở giai đoạn NDS phải đạt từ dạng chuẩn 3 trở lên.
  + Ở bảng **CASE\_REPORT**: loại bỏ các thuộc tính ***Reporting\_PHU***, ***Reporting\_PHU\_City***, ***Reporting\_PHU\_Address***, ***Reporting\_PHU\_Website***, ***Reporting\_PHU\_Latitude***, ***Reporting\_PHU\_Longitude***, **Reporting\_PHU\_Postal\_Code** có thể suy dẫn từ thuộc tính **PHU\_ID** được thêm vào và tham chiếu đến **PHU\_ID** của bảng **PHU**.
  + Ở bảng **CASE\_DETAIL**: loại bỏ các thuộc tính ***Reporting\_PHU***, ***Province*** có thể suy dẫn từ thuộc tính **PHU\_ID** được thêm vào và tham chiếu đến **PHU\_ID** của bảng **PHU**.
  + Ở bảng **PHU\_GROUP**: loại bỏ các thuộc tính ***Reporting\_PHU***, ***Reporting\_PHU\_City*** có thể suy dẫn từ thuộc tính **PHU\_ID** được thêm vào và tham chiếu đến **PHU\_ID** của bảng **PHU**.
  + Ở bảng **PHU**: thêm thuộc tính ***Reporting\_PHU\_Province*** dựa vào thuộc tính **Province** của bảng **CASE\_DETAIL**.
* Các dòng dữ liệu ở giai đoạn NDS đều được đánh dấu thời gian.
  + Thêm các thuộc tính lưu giá trị ngày tạo **Created\_Date** và ngày cập nhật **Updated\_Date** ở tất cả các bảng.
* Vì khóa ở nguồn có thể thay đổi và gây nên việc mất dữ liệu nên cần sử dụng khóa đại diện.
  + Thêm **ID** cho từng dòng dữ liệu như các khóa thay thế ở tất cả các bảng.

## **Database DDS**

* Phân tích yêu cầu:
  + Sự kiện:
    - Phân tích số ca nhiễm Covid-19
    - Phân tích sự kiện tiêm vaccine của các cơ sở y tế
  + Bối cảnh sự kiện:
    - Ai: Người nhiễm bệnh
    - Cái gì: Bệnh bùng phát
    - Khi nào: Ngày báo cáo
    - Ở đâu: Các đơn vị chăm sóc sức khỏe
  + Độ đo:
    - Mức độ nghiêm trọng
    - Số người tiêm được ít nhất 1 mũi/ 2 mũi/ 3 mũi/ đầy đủ.
* Mô hình hóa:
  + Fact table:
    - Các giá trị có sẵn từ nguồn:
      * CASE\_DETAIL: Case\_Status
      * VACCINES: One\_Dose, Second\_Dose, Fully\_Vaccinated, Third\_Dose
    - Cấp chi tiết dữ liệu:
      * Một dòng trong CASE\_DETAIL\_fact tương ứng với mỗi ca nhiễm được ghi nhận tại cơ sở y tế.
      * Một dòng trong VACCINE\_fact tương ứng với mỗi nhóm người bệnh cùng độ tuổi đã tiêm vaccine của 1 cơ sở y tế.
  + Dimension tables:
    - Date dimension: Year > Quarter > Month > Day
    - Geography dimension:
      * Cần lưu trữ giá trị cũ, tương lai có thể thay đổi phân bố nhưng hiếm xảy ra → SCD loại 3 – lưu trữ giá trị cũ theo cột, do không thay đổi nhiều thuộc tính.
      * Cần lưu nhiều thông tin PHU trong PHU\_Group → Chuẩn hóa thành các bảng con (snowflake schema).
    - Case dimension:
      * Cần lưu lại giá trị cũ → SCD loại 2 – lưu trữ giá trị cũ theo dòng.
    - Outbreak dimension:
      * Không yêu cầu lưu giá trị cũ → SCD loại 1 – ghi chồng giá trị cũ.
  + Phân cấp dữ liệu:

Diagram

Description automatically generated

* Lược đồ DDS: có 2 bảng fact và tái sử dụng các chiều → Mô hình chòm sao (Galaxy schema)

Diagram

Description automatically generated

# **Yêu cầu 3: Nạp dữ liệu các nguồn vào KDL**

## **Source to Stage**

* ***Case Report***
* Dữ liệu của source Case Report và stage CASE REPORT trước khi thực hiện ETL.

Graphical user interface, application, table

Description automatically generated

* Xóa dữ liệu của bảng Case Report để nạp vào những dòng dữ liệu mới cho mỗi lần ETL.
* Load dữ liệu từ Source vào bảng Case Report trong Stage.

Graphical user interface, application

Description automatically generated A screenshot of a phone

Description automatically generated with low confidence

* Thực hiện ETL để đưa dữ liệu vào Stage thành công.
* Kết quả:

A screenshot of a computer

Description automatically generated

* ***Case Detail***
* Dữ liệu của source Case Detail và stage CASE DETAIL trước khi thực hiện ETL.

***Graphical user interface, table

Description automatically generated***

* Xóa dữ liệu của bảng Case Detail để nạp vào những dòng dữ liệu mới cho mỗi lần ETL.
* Load dữ liệu từ Source vào bảng Case Detail trong Stage.

***Graphical user interface, text, application, chat or text message

Description automatically generated*** Graphical user interface, text, application, chat or text message

Description automatically generated

* Thực hiện ETL để đưa dữ liệu vào Stage thành công.
* Kết quả:

***Graphical user interface, application, table

Description automatically generated***

* ***Ongoing Outbreak***
* Dữ liệu của source Ongoing Outbreak và stage ONGOING OUTBREAK trước khi ETL.

***Graphical user interface, application

Description automatically generated***

* Xóa dữ liệu bảng Ongoing Outbreak để nạp vào những dòng dữ liệu mới cho mỗi lần ETL.
* Load dữ liệu từ Source vào bảng Ongoing Outbreak trong Stage.

***Graphical user interface, application

Description automatically generated*** ***Graphical user interface, application

Description automatically generated***

* Thực hiện ETL để đưa dữ liệu vào Stage thành công.
* Kết quả:

***Graphical user interface, application

Description automatically generated***

* ***PHU***
* Dữ liệu của source PHU và stage PHU trước khi thực hiện ETL.

***Graphical user interface, application

Description automatically generated***

* Xóa dữ liệu của bảng PHU để nạp vào những dòng dữ liệu mới cho mỗi lần ETL.
* Load dữ liệu từ Source vào bảng PHU trong Stage.

***Graphical user interface, application

Description automatically generated Graphical user interface, text, application, chat or text message

Description automatically generated***

* Thực hiện ETL để đưa dữ liệu vào Stage thành công.
* Kết quả:

***Graphical user interface, application

Description automatically generated***

* ***PHU Group***
* Dữ liệu của source PHU Group và stage PHU GROUP trước khi thực hiện ETL.

***Graphical user interface, text, application

Description automatically generated***

* Xóa dữ liệu của bảng PHU Group để nạp vào những dòng dữ liệu mới cho mỗi lần ETL.
* Load dữ liệu từ Source vào bảng PHU Group trong Stage.

***Graphical user interface, text, application, chat or text message

Description automatically generated*** ***Graphical user interface, text, application

Description automatically generated***

* Thực hiện ETL để đưa dữ liệu vào Stage thành công.
* Kết quả:

***Graphical user interface, text, application

Description automatically generated***

* ***Vaccines***
* Dữ liệu của source Vaccines và stage VACCINES trước khi thực hiện ETL.

***Table

Description automatically generated***

* Xóa dữ liệu của bảng Vaccines để nạp vào những dòng dữ liệu mới cho mỗi lần ETL.
* Load dữ liệu từ Source vào bảng Vaccines trong Stage.

***Graphical user interface, application

Description automatically generated*** ***Graphical user interface, text, application, chat or text message

Description automatically generated***

* Thực hiện ETL để đưa dữ liệu vào Stage thành công.
* Kết quả:

***Graphical user interface, application, table

Description automatically generated***

## **Stage to NDS**

* ***Case Report***
* Dữ liệu của stage CASE REPORT và NDS CASE REPORT trước khi thực hiện ETL.

***Graphical user interface, application

Description automatically generated***

* Load dữ liệu bảng Case Report ở Stage.
* Thêm khóa ngoại tham chiếu đến PHU\_ID của bảng PHU.
* Lookup CaseReport\_ID:
  + Nếu có tồn tại, thực hiện Update dữ liệu.
  + Nếu không tồn tại, thực hiện Insert dữ liệu vào NDS.

***A screenshot of a video game

Description automatically generated***

* Thực hiện ETL để đưa dữ liệu vào NDS thành công.
* Kết quả:

***Table

Description automatically generated***

* ***Case Detail***
* Dữ liệu của stage CASE DETAIL và NDS CASE DETAIL trước khi thực hiện ETL.

***Graphical user interface, table

Description automatically generated***

* Load dữ liệu bảng Case Detail ở Stage.
* Thêm khóa ngoại tham chiếu đến PHU\_ID của bảng PHU.
* Lookup CaseDetail\_ID:
  + Nếu có tồn tại, thực hiện Update dữ liệu.
  + Nếu không tồn tại, thực hiện Insert dữ liệu vào NDS.

***A screenshot of a video game

Description automatically generated***

* Thực hiện ETL để đưa dữ liệu vào NDS thành công.
* Kết quả:

***Graphical user interface, application, table

Description automatically generated***

* ***Ongoing Outbreak***
* Dữ liệu của stage ONGOING OUTBREAK và NDS ONGOING OUTBREAK trước khi ETL.

***Graphical user interface, text, application, email

Description automatically generated***

* Load dữ liệu bảng Ongoing Outbreak ở Stage.
* Lookup Outbreak\_ID:
  + Nếu có tồn tại, thực hiện Update dữ liệu.
  + Nếu không tồn tại, thực hiện Insert dữ liệu vào NDS.

***Graphical user interface, application

Description automatically generated***

* Thực hiện ETL để đưa dữ liệu vào NDS thành công.
* Kết quả:

***Graphical user interface, application, table

Description automatically generated***

* ***PHU***
* Dữ liệu của stage PHU và NDS PHU trước khi thực hiện ETL.

***Graphical user interface, application

Description automatically generated***

* Load dữ liệu bảng PHU ở Stage.
* Lookup Reporting\_PHU\_Province đến Province của bảng Case\_Detail.
* Lookup PHU\_ID:
  + Nếu có tồn tại, thực hiện Update dữ liệu.
  + Nếu không tồn tại, thực hiện Insert dữ liệu vào NDS.

***A screenshot of a video game

Description automatically generated***

* Thực hiện ETL để đưa dữ liệu vào NDS thành công.
* Kết quả:

***Graphical user interface, application

Description automatically generated***

* ***PHU Group***
* Dữ liệu của stage PHU GROUP và NDS PHU GROUP trước khi thực hiện ETL.

***Graphical user interface, application

Description automatically generated***

* Load dữ liệu bảng PHU Group ở Stage.
* Thêm khóa ngoại tham chiếu đến PHU\_ID của bảng PHU.
* Lookup PHUGroup\_ID:
  + Nếu có tồn tại, thực hiện Update dữ liệu.
  + Nếu không tồn tại, thực hiện Insert dữ liệu vào NDS.

***Graphical user interface, application

Description automatically generated***

* Thực hiện ETL để đưa dữ liệu vào NDS thành công.
* Kết quả:

***Graphical user interface, text, application, email

Description automatically generated***

* ***Vaccines***
* Dữ liệu của stage VACCINES và NDS VACCINES trước khi thực hiện ETL.

Graphical user interface, application

Description automatically generated

* Load dữ liệu bảng Vaccines ở Stage.
* Lookup Vaccines\_ID:
  + Nếu có tồn tại, thực hiện Update dữ liệu.
  + Nếu không tồn tại, thực hiện Insert dữ liệu vào NDS.

Graphical user interface, application

Description automatically generated

* Thực hiện ETL để đưa dữ liệu vào NDS thành công.
* Kết quả:

Graphical user interface, application, table

Description automatically generated