****

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NÔNG LÂM**

**KHOA: CÔNG NGHỆ THÔNG TIN**



------

BÁO CÁO MÔN HỌC  
Môn học: Data Warehouse

Sinh viên thực hiện: Nguyễn Cường Huy

Mã số sinh viên: 21130377

Lớp: DH21DTB

Link github: https://github.com/jussahug/DH\_Nhom11.git

**MỤC LỤC**

**[MÔ TẢ QUY TRÌNH THIẾT KẾ HỆ THỐNG DATA WAREHOUSE 3](#_Toc10408)**

**[I. Giới Thiệu Sơ Lược: 3](#_Toc23525)**

[1. Giới thiệu: 3](#_Toc28402)

[2. Chi tiết quy trình thiết kế hệ thống Data Warehouse: 3](#_Toc17640)

**[II. Mô Tả Dữ Liệu: 5](#_Toc5475)**

[1. Bảng phimChieuRap: 5](#_Toc28112)

[2. Bảng Log: 6](#_Toc1056)

[3. Bảng Config: 6](#_Toc10968)

**[III. Quy trình thực hiện: 6](#_Toc24675)**

[1. Transform: 6](#_Toc20125)

[2. Warehouse: 7](#_Toc29587)

**MÔ TẢ QUY TRÌNH THIẾT KẾ HỆ THỐNG DATA WAREHOUSE**

# Giới Thiệu Sơ Lược:

## Giới thiệu:

Quy trình thiết kế hệ thống Data Warehouse (DWH) là một bước quan trọng trong việc xây dựng nền tảng quản lý dữ liệu tập trung, phục vụ nhu cầu phân tích và ra quyết định trong tổ chức. Data Warehouse không chỉ là nơi lưu trữ dữ liệu mà còn đóng vai trò như một công cụ hỗ trợ doanh nghiệp khai thác thông tin một cách hiệu quả và chính xác.

Quy trình này bao gồm các giai đoạn chính như xác định yêu cầu nghiệp vụ, thiết kế mô hình dữ liệu, trích xuất - chuyển đổi - nạp dữ liệu (ETL), và tối ưu hóa hiệu năng truy vấn. Mỗi bước trong quy trình đều được thực hiện một cách khoa học và chặt chẽ nhằm đảm bảo hệ thống được xây dựng với độ tin cậy cao, linh hoạt và dễ dàng mở rộng khi có nhu cầu.

Thiết kế DWH cần phải đáp ứng được các tiêu chí về tính nhất quán, khả năng chịu tải lớn, cũng như khả năng tích hợp với các nguồn dữ liệu đa dạng. Đồng thời, nó cũng phải đảm bảo khả năng cung cấp thông tin nhanh chóng và chính xác cho các báo cáo và phân tích.

Trong phần tiếp theo, quy trình này sẽ được mô tả chi tiết để làm rõ các bước cần thiết, công cụ và phương pháp áp dụng, cũng như cách thức tổ chức dữ liệu trong hệ thống DWH nhằm đạt được mục tiêu của tổ chức.

## Chi tiết quy trình thiết kế hệ thống Data Warehouse:

A/ Xác định yêu cầu nghiệp vụ:  
 Giai đoạn đầu tiên trong quy trình thiết kế hệ thống DWH là thu thập và phân tích yêu cầu từ phía các bên liên quan. Điều này bao gồm việc hiểu rõ mục tiêu kinh doanh, các loại báo cáo cần thiết, và các chỉ số hiệu suất chính (KPIs) mà tổ chức muốn theo dõi.

- Hoạt động chính:

+ Phỏng vấn các bên liên quan (stakeholders).

+ Xác định các nguồn dữ liệu hiện có.

+ Ưu tiên hóa các yêu cầu phân tích dữ liệu.

- Kết quả đầu ra: Tài liệu yêu cầu nghiệp vụ (Business Requirement Document - BRD) làm cơ sở cho việc thiết kế hệ thống.

B/ Thiết kế mô hình dữ liệu:  
 Sau khi xác định yêu cầu, bước tiếp theo là thiết kế mô hình dữ liệu phù hợp. Mô hình dữ liệu trong DWH thường được chia thành hai loại chính:

- Mô hình quan hệ (ER): Tập trung vào tổ chức dữ liệu theo dạng bảng với các mối quan hệ phức tạp.

- Mô hình ngôi sao (Star Schema) và tuyết (Snowflake Schema): Tập trung vào tối ưu hóa cho truy vấn và phân tích dữ liệu.

- Hoạt động chính:

+ Xác định các bảng sự kiện (Fact tables) và bảng chiều (Dimension tables).

+ Định nghĩa các khóa chính, khóa ngoại, và mối quan hệ giữa các bảng.

+ Tối ưu hóa mô hình để hỗ trợ các truy vấn phức tạp.

- Kết quả đầu ra: Mô hình dữ liệu vật lý và logic.

C/Xây dựng quy trình ETL (Trích xuất - Chuyển đổi - Nạp dữ liệu)  
 Đây là bước chuyển đổi dữ liệu từ các hệ thống nguồn vào Data Warehouse. Quy trình ETL cần được thiết kế để đảm bảo dữ liệu được làm sạch, chuẩn hóa và sẵn sàng cho phân tích.

- Hoạt động chính:

+Trích xuất (Extract): Kết nối và lấy dữ liệu từ các hệ thống nguồn (CRM, ERP, cơ sở dữ liệu SQL, API, v.v.).

+ Chuyển đổi (Transform): Làm sạch, chuẩn hóa, và biến đổi dữ liệu để phù hợp với mô hình DWH.

+Nạp dữ liệu (Load): Đưa dữ liệu đã xử lý vào các bảng Fact và Dimension.

- Kết quả đầu ra: Quy trình ETL tự động và tài liệu chi tiết về dữ liệu nguồn và đích.

D/ Tích hợp và tối ưu hóa hiệu năng hệ thống:  
 Sau khi hoàn thành các bước thiết kế, dữ liệu trong hệ thống cần được tối ưu hóa để hỗ trợ truy vấn nhanh chóng và chính xác. Điều này bao gồm:

- Tối ưu hóa cấu trúc bảng và chỉ số.

- Sử dụng các công cụ lập chỉ mục, phân vùng dữ liệu (partitioning), và sao chép dữ liệu (replication).

- Kiểm tra hiệu năng và tinh chỉnh các truy vấn SQL.

E/ Triển khai và kiểm thử hệ thống:  
 Trước khi đưa vào sử dụng, hệ thống cần được triển khai trong môi trường thử nghiệm và kiểm tra toàn diện.

- Hoạt động chính:

+ Kiểm thử dữ liệu (Data validation).

+ Kiểm thử hiệu năng (Performance testing).

+ Kiểm thử các báo cáo và phân tích.

- Kết quả đầu ra: Hệ thống DWH hoạt động ổn định và đáp ứng yêu cầu ban đầu.

F/ Quản lý và bảo trì hệ thống:  
 Sau khi triển khai, hệ thống DWH cần được duy trì và quản lý định kỳ để đảm bảo hoạt động ổn định.

- Hoạt động chính:

+ Cập nhật quy trình ETL khi có nguồn dữ liệu mới.

+ Giám sát hiệu năng và bảo trì hệ thống.

+ Đảm bảo an toàn và bảo mật dữ liệu.

# Mô Tả Dữ Liệu:

## Bảng phimChieuRap:

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Kiểu dữ liệu | Mô tả | Ví dụ |
| id | INT | Mã phim chiếu rạp, là khóa chính, tự động tăng. | 1 |
| name\_vn | NVARCHAR(255) | Tên phim bằng tiếng Việ | Avengers: Endgame |
| director | NVARCHAR(255) | Tên đạo diễn của phim | Anthony Russo |
| actor | NVARCHAR(255) | Các diễn viên tham gia trong phim | Robert Downey Jr., Chris Hemsworth |
| limitage\_vn | INT | Giới hạn độ tuổi của người xem phim | 13 |
| country\_name\_vn | INT | Mã quốc gia của phim (tham chiếu đến bảng quốc gia) | 1 |
| brief\_vn | NVARCHAR(MAX) | Mô tả ngắn gọn về phim bằng tiếng Việt | Cuộc chiến cuối cùng chống lại kẻ thù lớn nhất của loài người |
| image | NVARCHAR(255) | URL ảnh đại diện của phim | http://images/avengers\_endgame.jpg |
| release\_date | DATE | Ngày phát hành phim | 2023-04-25 |
| end\_date | DATE | Ngày kết thúc chiếu phim | 2023-07-30 |
| duration | INT | Thời gian chiếu phim (dạng số nguyên, tính bằng phút) | 180 |
| created\_at | DATETIME | Ngày tạo bản ghi | 2023-04-01 |
| updated\_at | DATETIME | Ngày cập nhật bản ghi | 2023-04-01 |

## Bảng Log:

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Kiểu dữ liệu | Mô tả | Ví dụ |
| id | INT | Mã log, là khóa chính, tự động tăng | 1 |
| id\_config | INT | Mã cấu hình tương ứng (tham chiếu đến bảng config) | 1 |
| fileName | NVARCHAR(255) | Tên file liên quan đến log. | movie\_data\_import.csv |
| count | INT | Số lượng file được xử lý hoặc ghi nhận. | 1000 |
| status | NVARCHAR(50) | Trạng thái của file (ví dụ: thành công, lỗi, v.v.) | Success |
| file\_size | INT | Kích thước của file (đơn vị byte) | 204800 |
| date\_create | DATE | Ngày tạo log | 2023-04-01 |
| date\_update | DATE | Ngày cập nhật log | 2023-04-01 |

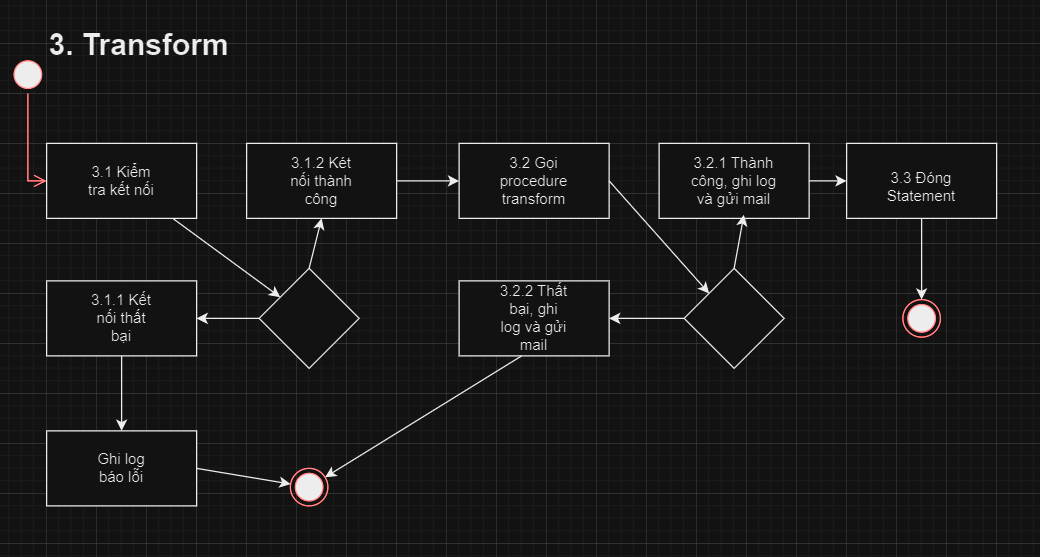
## Bảng Config:

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Kiểu dữ liệu | Mô tả | Ví dụ |
| id | INT | Mã cấu hình, là khóa chính, tự động tăng | 1 |
| name | NVARCHAR(255) | Tên cấu hình | Movie Data Import |
| source | NVARCHAR(255) | Nguồn dữ liệu hoặc hệ thống gốc | API |
| source\_fileLocation | NVARCHAR(255) | Vị trí của file nguồn | https://data/source/movies/ |
| des\_fileLocation | NVARCHAR(255) | Vị trí của file đích | https://data/destination/movies/ |
| created\_date | DATETIME | Ngày tạo cấu hình | 2023-04-01 |

# Quy trình thực hiện:

## Transform:

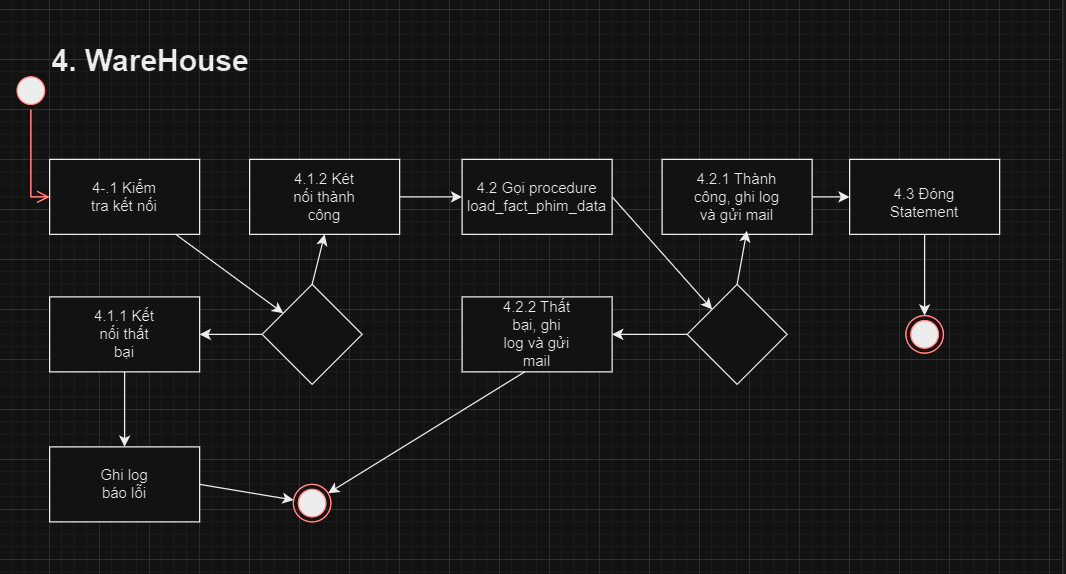
- Workflow:



- Mô tả: Quá trình transform xử lý và chuyển đổi dữ liệu từ nguồn ban đầu để phù hợp với Data Warehouse. Trong mã nguồn, hàm transformData() gọi procedure transform\_phim\_data() để xử lý dữ liệu phim. Sau khi thành công, hệ thống ghi log và gửi email thông báo "Data Transformation Success". Nếu có lỗi, email "Data Transformation Failure" sẽ được gửi.

## Warehouse:

- Workflow:



- Mô tả: Quá trình load dữ liệu tải dữ liệu đã được transform vào Data Warehouse. Hàm loadDataToWarehouse() gọi procedure load\_fact\_phim\_data() để thực hiện việc này. Sau khi hoàn tất, email "Data Load Success" được gửi. Nếu lỗi xảy ra, email "Data Load Failure" sẽ được gửi. Cả hai bước đều bao gồm việc ghi log và gửi email thông báo về tình trạng thành công hoặc lỗi.