TRƯỜNG ĐẠI HỌC THĂNG LONG

**KHOA CÔNG NGHỆ THÔNG TIN**

**A red and blue logo

Description automatically generated**

**BÀI TẬP LỚN**

**TÌM HIỂU VỀ DATAWAREHOUSE CHO CÔNG TY LĨNH VỰC ĐIỆN ẢNH**

**GIÁO VIÊN HƯỚNG DẪN: SINH VIÊN THỰC HIỆN:**

**CTI045 - VŨ MINH HOÀNG A41529 – ĐỖ THỊ LAN**

**A43193 – LÊ THỊ MỸ HẢO**

**A41641 – VŨ KIỀU TRANG**

**HÀ NỘI – 2024**

**MỤC LỤC**

[PHẦN 1. Mô Tả bài toán 7](#_Toc157897058)

[PHẦN 2. Cơ sở lý thuyết 8](#_Toc157897059)

[2.1. DataWarehouse 8](#_Toc157897060)

[2.2. Star Schema 8](#_Toc157897061)

[2.3. ETL 8](#_Toc157897062)

[PHẦN 3. Giải quyết bài toán 9](#_Toc157897063)

[3.1. Datawarehouse Process 9](#_Toc157897064)

[3.2. Business Process 9](#_Toc157897065)

[3.2.1. Khái niệm 9](#_Toc157897066)

[3.2.2. Xây dựng Business Process 9](#_Toc157897067)

[3.2.2.1 Quy trình tài chính 9](#_Toc157897068)

[3.2.2.2 Quy trình kinh doanh 10](#_Toc157897069)

[3.2.2.3 Quy trình nhân sự 11](#_Toc157897070)

[3.2.2.4 Quy trình marketing 12](#_Toc157897071)

[3.2.2.5 Quy trình sản xuất 12](#_Toc157897072)

[3.3. Thiết kế Star Schema 13](#_Toc157897073)

[3.3.1. Quy trình tài chính 14](#_Toc157897074)

[3.3.1.1 Quản lý hóa đơn 14](#_Toc157897075)

[3.3.1.2 Quản lý ngân sách 14](#_Toc157897076)

[3.3.1.3 Quản lý nhà đầu tư 15](#_Toc157897077)

[3.3.1.4 Quản lý thu chi và ngân quỹ 15](#_Toc157897078)

[3.3.1.5 Báo cáo tài chính 16](#_Toc157897079)

[3.3.2. Quy trình kinh doanh 17](#_Toc157897080)

[3.3.2.1 Phân tích và nghiên cứu thị trường 17](#_Toc157897081)

[3.3.2.2 Phát triển ý tưởng và kịch bản 17](#_Toc157897082)

[3.3.2.3 Lập kế hoạch sản xuất 18](#_Toc157897083)

[3.3.2.4 Chiến lược Phát hành và quảng bá 19](#_Toc157897084)

[**3.3.2.5** **Báo cáo kinh doanh** 19](#_Toc157897085)

[3.3.3. Quy trình nhân sự 20](#_Toc157897086)

[3.3.3.1 Tuyển dụng và tuyển chọn 20](#_Toc157897087)

[3.3.3.2 Đào tạo và phát triển nhân sự 21](#_Toc157897088)

[3.3.3.3 Quản lý hiệu suất 21](#_Toc157897089)

[3.3.3.4 Chăm sóc nhân sự và phúc lợi: 22](#_Toc157897090)

[3.3.3.5 Báo cáo nhân sự 23](#_Toc157897091)

[3.3.4. Quy trình marketing 23](#_Toc157897092)

[3.3.4.1 Nghiên cứu và phân tích thị trường 23](#_Toc157897093)

[3.3.4.2 Quảng cáo và quảng bá 24](#_Toc157897094)

[3.3.4.3 Xác định chiến lược tiếp thị 25](#_Toc157897095)

[3.3.4.4 Quản lý truyền thông xã hội 25](#_Toc157897096)

[3.3.4.5 Đánh giá và đo lường hiệu suất 26](#_Toc157897097)

[3.3.5. Quy trình sản xuất 27](#_Toc157897098)

[3.3.5.1 Lên ý tưởng 27](#_Toc157897099)

[3.3.5.2 Tiền kì 27](#_Toc157897100)

[3.3.5.3 Sản xuất 28](#_Toc157897101)

[3.3.5.4 Hậu kì 29](#_Toc157897102)

[3.3.5.5 Phát hành 30](#_Toc157897103)

[3.3.5.6 Đánh giá 31](#_Toc157897104)

[3.4. Xây dựng SQL cho các business process 32](#_Toc157897105)

[3.4.1. Quy trình sản xuất 32](#_Toc157897106)

[3.4.2. Quy trình tài chính 32](#_Toc157897107)

[3.4.3. Quy trình kinh doanh 33](#_Toc157897108)

[3.4.4. Quy trình nhân sự 33](#_Toc157897109)

[3.4.5. Quy trình Marketing 34](#_Toc157897110)

[3.5. Xây dựng ETL cho bài toán 34](#_Toc157897111)

[3.5.1. Làm sạch dữ liệu 34](#_Toc157897112)

[3.5.2. Xây dựng Dim và Fact 35](#_Toc157897113)

[3.5.3. Chi tiết từng bảng 35](#_Toc157897114)

[3.5.3.1 Dim\_Product 35](#_Toc157897115)

[3.5.3.2 Dim\_Audience 36](#_Toc157897116)

[3.5.3.3 Dim\_Profit 36](#_Toc157897117)

[3.5.3.4 Fact\_PhatHanh 36](#_Toc157897118)

[PHẦN 4. Demo Xây dựng etl 37](#_Toc157897119)

[4.1. Mục đích 37](#_Toc157897120)

[4.2. Extract 37](#_Toc157897121)

[4.3. Transform 40](#_Toc157897122)

[4.4. Load 44](#_Toc157897123)

[PHẦN 5. Phân chia công việc 1](#_Toc157897124)

**DANH MỤC MINH HỌA**

[Ảnh 3.1. Quy trình tài chính 9](#_Toc157894260)

[Ảnh 3.2. Quy trình kinh doanh 10](#_Toc157894261)

[Ảnh 3.3. Quy trình nhân sự 11](#_Toc157894262)

[Ảnh 3.4. Quy trình Marketing 12](#_Toc157894263)

[Ảnh 3.5. Quy trình sản xuất 12](#_Toc157894264)

[Ảnh 3.6. Quản lý hóa đơn 14](#_Toc157894265)

[Ảnh 3.7. Quản lý ngân sách 14](#_Toc157894266)

[Ảnh 3.8. Quản lý nhà đầu tư 15](#_Toc157894267)

[Ảnh 3.9. Quản lý thu chi và ngân quỹ 15](#_Toc157894268)

[Ảnh 3.10. Báo cáo tài chính 16](#_Toc157894269)

[Ảnh 3.11. Phân tích và nghiên cứu thị trường 17](#_Toc157894270)

[Ảnh 3.12. Phát triển ý tưởng và kịch bản 17](#_Toc157894271)

[Ảnh 3.13. Lập kế hoạch sản xuất 18](#_Toc157894272)

[Ảnh 3.14. Chiến lược phát hành và quảng bá 19](#_Toc157894273)

[Ảnh 3.15. Báo cáo kinh doanh 19](#_Toc157894274)

[Ảnh 3.16. Tuyển dụng và tuyển chọn 20](#_Toc157894275)

[Ảnh 3.17. Đào tạo và phát triển nhân sự 21](#_Toc157894276)

[Ảnh 3.18. Quản lý hiệu suất 21](#_Toc157894277)

[Ảnh 3.19. Chăm sóc nhân sự và phúc lợi 22](#_Toc157894278)

[Ảnh 3.20. Báo cáo nhân sự 23](#_Toc157894279)

[Ảnh 3.21. Nghiên cứu và phân tích thị trường 23](#_Toc157894280)

[Ảnh 3.22. Quảng cáo và quảng bá 24](#_Toc157894281)

[Ảnh 3.23. Xác định chiến lược tiếp thị 25](#_Toc157894282)

[Ảnh 3.24. Quản lý truyền thông xã hội 25](#_Toc157894283)

[Ảnh 3.25. Đánh giá và đo lường hiệu suất 26](#_Toc157894284)

[Ảnh 3.26. Lên ý tưởng 27](#_Toc157894285)

[Ảnh 3.27. Tiền kì 27](#_Toc157894286)

[Ảnh 3.28. Sản xuất 28](#_Toc157894287)

[Ảnh 3.29. Hậu kì 29](#_Toc157894288)

[Ảnh 3.30. Phát hành 30](#_Toc157894289)

[Ảnh 3.31. Đánh giá 31](#_Toc157894290)

[Ảnh 3.32. SQL cho quy trình sản xuất 32](#_Toc157894291)

[Ảnh 3.33. SQL cho quy trình tài chính 32](#_Toc157894292)

[Ảnh 3.34. SQL cho quy trình kinh doanh 33](#_Toc157894293)

[Ảnh 3.35. SQL cho quy trình nhân sự 33](#_Toc157894294)

[Ảnh 3.36. SQL cho quy trình Marketing 34](#_Toc157894295)

[Ảnh 3.37. ETL trong giai đoạn làm sạch dữ liệu 34](#_Toc157894296)

[Ảnh 3.38. Thiết kế ETL cho dim và fact 35](#_Toc157894297)

[Ảnh 3.39. Thiết kế chi tiết ETL cho bảng Dim\_Product 35](#_Toc157894298)

[Ảnh 3.40. Thiết kế bảng Dim\_Audience 36](#_Toc157894299)

[Ảnh 3.41. Thiết kế bảng Dim\_Profit 36](#_Toc157894300)

[Ảnh 3.42. Thiết kế bảng Fact\_PhatHanh 36](#_Toc157894301)

[Ảnh 4.1. Quy trình ETL 37](#_Toc157894302)

[Ảnh 4.2. Thư viện sử dụng để lập trình 38](#_Toc157894303)

[Ảnh 4.3. Hàm extract\_transform\_load để đọc các file 38](#_Toc157894304)

[Ảnh 4.4. Dataframes wiki\_movies\_df 39](#_Toc157894305)

[Ảnh 4.5. Dataframes kaggle\_metadata 39](#_Toc157894306)

[Ảnh 4.6. Dataframes ratings 40](#_Toc157894307)

[Ảnh 4.7. Sử dụng hàm clean\_movie và change\_column\_name 41](#_Toc157894308)

[Ảnh 4.8. Kết nối với database 44](#_Toc157894309)

[Ảnh 4.9. Kết quả thực hiện quy trình ETL 45](#_Toc157894310)

# Mô Tả bài toán

Lĩnh vực làm phim là một ngành nghệ thuật và công nghiệp tạo ra các tác phẩm nghệ thuật chuyển động được. Nó không chỉ là một hình thức giải trí mà còn là một phương tiện mạnh mẽ để truyền đạt thông điệp, tạo ra ảnh hưởng và kể chuyện. Mỗi dự án phim bắt đầu với một ý tưởng sáng tạo và sau đó là một quy trình phức tạp bắt đầu để biến ý tưởng đó thành một sản phẩm hoàn chỉnh.

Theo số liệu 2011, chỉ riêng 3 nước Ấn Độ, Nigera, Mỹ đã sản xuất lên đến 3.070 bộ phim. Vậy làm thế nào để lưu trữ các bộ phim này? Làm thế nào để biết được đâu là dữ liệu quan trọng và đâu là dữ liệu nháp cần bỏ? Muốn quản lí vấn đề này một cách hiệu quả ta cần xây dựng một kho dữ liệu (DataWarehouse) để lọc các dữ liệu thô từ tệp dữ liệu thành các dữ liệu có ích và lưu vào trong kho. Thiết kế kho dữ liệu (DataWarehouse) với mục đích giảm thiểu thời gian tìm kiếm dữ liệu và lưu dữ liệu một cách hiệu quả.

Xây dựng DataWarehouse lưu trữ hiệu quả cần:

* Phân tích nhu cầu doanh nghiệp
* Thiết kế quy trình nghiệp vụ cho doanh nghiệp
* Thiết kế Star Schema cho từng nghiệp vụ
* Thu thập dữ liệu
* Xây dựng ETL
* Phát triển và triển khai

# Cơ sở lý thuyết

## DataWarehouse

Data warehouse (DW) là một hệ thống lưu trữ dữ liệu từ nhiều nguồn, nhiều môi trường khác nhau như: phần mềm bán hàng, kế toán, nhân sự hay hệ thống lõi ngân hàng, … giúp tăng cường hiệu suất của các truy vấn cho báo cáo và phân tích.

Data Warehouse hoạt động như một kho lưu trữ trung tâm. Dữ liệu đi vào kho dữ liệu từ hệ thống giao dịch và các cơ sở dữ liệu liên quan khác. Sau đó, dữ liệu được xử lý, chuyển đổi để người dùng có thể truy cập những dữ liệu này thông qua công cụ Business Intelligence, SQL client hay bảng tính.

## Star Schema

Star Schema là một lược đồ quan hệ trong đó một lược đồ quan hệ có thiết kế đại diện cho một mô hình dữ liệu đa chiều. Star Schema là lược đồ kho dữ liệu rõ ràng. Nó được gọi là giản đồ sao vì biểu đồ mối quan hệ thực thể của lược đồ này mô phỏng một ngôi sao, với các điểm, phân kỳ từ một bảng trung tâm. Trung tâm của lược đồ bao gồm một bảng dữ kiện lớn và các điểm của ngôi sao là bảng kích thước.

Đặc điểm Star Schema: Star Schema rất thích hợp cho việc thiết kế cơ sở dữ liệu kho dữ liệu vì các tính năng sau:

* Tạo ra một cơ sở dữ liệu chuẩn hóa DE có thể nhanh chóng cung cấp các phản hồi truy vấn.
* Cung cấp một thiết kế linh hoạt có thể dễ dàng thay đổi hoặc thêm vào trong suốt chu kỳ phát triển và khi cơ sở dữ liệu phát triển.
* Cung cấp sự song song trong thiết kế về cách người dùng cuối thường nghĩ và sử dụng dữ liệu.
* Làm giảm sự phức tạp của siêu dữ liệu cho cả nhà phát triển và người dùng cuối.

## ETL

Trích xuất, chuyển đổi và tải (ETL) là quá trình kết hợp dữ liệu từ nhiều nguồn thành một kho lưu trữ lớn trung tâm được gọi là kho dữ liệu. ETL sử dụng hệ thống các quy tắc kinh doanh để làm sạch và sắp xếp dữ liệu thô, sau đó chuẩn bị dữ liệu cho quá trình lưu trữ, phân tích và máy học (ML).

# Giải quyết bài toán

## Datawarehouse Process

Data Warehouse Process là quá trình tổ chức, tích hợp và lưu trữ dữ liệu từ nhiều nguồn khác nhau vào một kho dữ liệu tập trung, nhằm hỗ trợ việc phân tích và ra quyết định trong doanh nghiệp. Quá trình này bao gồm các bước như trích xuất dữ liệu từ nguồn, biến đổi để đảm bảo đồng nhất và chất lượng, sau đó nạp dữ liệu vào kho dữ liệu để phục vụ nhu cầu phân tích và báo cáo.

## Business Process

### Khái niệm

Business Process là một chuỗi các hoạt động được liên kết với nhau nhằm hoàn thành một mục tiêu nào đó của tổ chức. Những hoạt động này được chỉ định cho các bên liên quan. Họ sẽ để thực hiện một công việc cụ thể với mục đích cung cấp sản phẩm hoặc dịch vụ cho khách hàng.

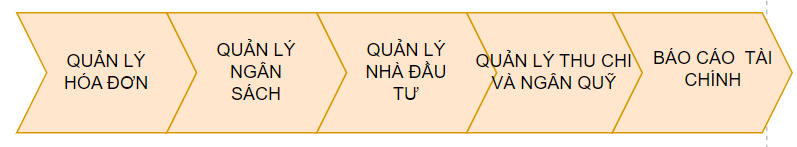
Ưu điểm của Business Process:

* Cải thiện năng suất làm việc
* Tiết kiệm chi phí hoạt động
* Đề ra những bước tiến đột phá
* Giảm thiểu rủi ro

### Xây dựng Business Process

Mỗi dự án phim bắt đầu với một ý tưởng sáng tạo và sau đó một quy trình phức tạp bắt đầu để biến ý tưởng đó thành một sản phẩm hoàn chỉnh. Các bước chính trong quá trình sản xuất phim bao gồm.

#### Quy trình tài chính



Ảnh 3.1. Quy trình tài chính

* Quản lý hóa đơn:
* Quản lý hóa đơn cho phép theo dõi chặt chẽ chi phí và nhận biết bất kỳ chi phí không mong muốn nào. Điều này giúp quản lý và nhóm sản xuất có khả năng kiểm soát ngân sách và đưa ra quyết định dựa trên thông tin chi tiết.
* Quản lý ngân sách:
* Đề suất ngân sách để sản xuất phim
* Ngân sách được duy trì và kiểm soát một cách chặt chẽ, giúp dự án phim tiến triển một cách có tổ chức và thành công.
* Quản lý nhà đầu tư:
* Đảm bảo mối quan hệ tích cực với nhà đầu tư, giúp họ hiểu rõ về tiến trình tài chính của dự án và đồng thuận với quản lý chi phí, tăng cường sự minh bạch và tin tưởng, đồng thời bảo vệ lợi ích của cả hai bên.
* Quản lý thu chi ngân sách
* Theo dõi các khoản thu nhập và chi phí giúp doanh nghiệp hiểu rõ hơn về nơi tiêu tiền và làm thế nào để tối ưu hóa nguồn lực tài chính.
* Báo cáo tài chính
* Công cụ quan trọng để chia sẻ thông tin với người đọc bên ngoại như cổ đông, đối tác kinh doanh, hay các bên liên quan khác.

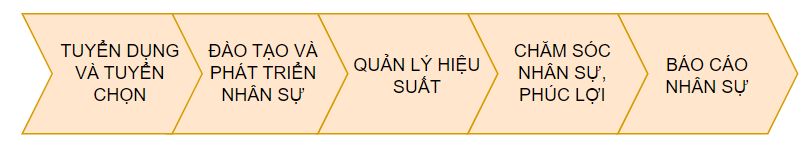
#### Quy trình kinh doanh



Ảnh 3.2. Quy trình kinh doanh

* Phân tích và nghiên cứu thị trường:
* Xác định thị trường mục tiêu và tìm hiểu về xu hướng, yêu cầu của khán giả. Nghiên cứu ý tưởng và xác định tiềm năng thị trường cho dự án phim.
* Phát triển ý tưởng và kịch bản:
* Bắt đầu với việc phát triển ý tưởng thành kịch bản. Quá trình này liên quan đến việc viết, chỉnh sửa và phát triển câu chuyện, nhân vật và các yếu tố tác phẩm.
* Lập kế hoạch sản xuất:
* Xác định các yếu tố sản xuất như địa điểm quay, lịch trình sản xuất, ngân sách, đội ngũ sản xuất, và các nguồn lực khác.
* Chiến lược và phát hành quảng bá:
* Phát triển chiến lược tiếp thị và quảng bá để thu hút sự chú ý từ khán giả. Bao gồm cả việc tạo trailer, quảng cáo, và sự kiện quảng bá.
* Báo cáo kinh doanh:
* Tạo báo cáo đánh giá kinh doanh

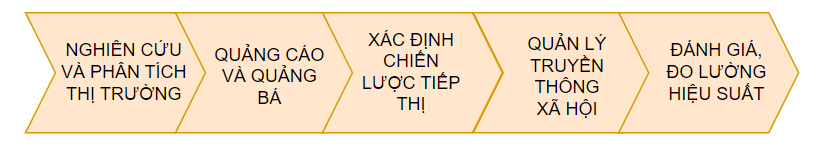
#### Quy trình nhân sự



Ảnh 3.3. Quy trình nhân sự

* Tuyển dụng và tuyển chọn:
* Tìm kiếm, thu hút và chọn lựa những cá nhân có kỹ năng và kinh nghiệm phù hợp để tham gia vào dự án phim. Bao gồm việc xây dựng hồ sơ, phỏng vấn và chọn lựa ứng viên.
* Đào tạo và phát triển nhân sự:
* Cung cấp đào tạo và phát triển kỹ năng cho nhân viên để họ có thể hoàn thành nhiệm vụ của mình hiệu quả.
* Quản lý hiệu suất:
* Theo dõi và đánh giá hiệu suất của nhân viên theo thời gian. Điều này bao gồm việc thiết lập mục tiêu, đánh giá định kỳ và cung cấp phản hồi để cải thiện hiệu suất làm việc.
* Chăm sóc nhân sự, phúc lợi:
* Chăm sóc nhân sự và tạo các cơ hội phúc lợi cho nhân sự
* Báo cáo nhân sự:
* Báo cáo chi tiết năng suất làm việc của từng bộ phận, phòng ban

#### Quy trình marketing



Ảnh 3.4. Quy trình Marketing

* Nghiên cứu và phân tích thị trường:
* Hiểu rõ đối tượng khán giả mục tiêu, sở thích, xu hướng, và yêu cầu giúp định hình chiến lược marketing phù hợp và tối ưu hóa hiệu suất chiến dịch.
* Quảng cáo và quảng bá:
* Tổ chức sự kiện đặc biệt và chiến dịch quảng cáo nhằm tăng cường sự nhận thức về dự án. Điều này có thể bao gồm tham gia các liên hoan phim, buổi họp báo, hay chiến dịch quảng cáo đa kênh để thu hút sự chú ý.
* Xác định chiến lược tiếp thị:
* Dựa trên thông tin từ nghiên cứu thị trường, xây dựng chiến lược tiếp thị để xác định cách tiếp cận và tương tác với đối tượng khán giả
* Điều này có thể bao gồm quảng cáo truyền hình, truyền thông xã hội, sự kiện, và chiến dịch tiếp thị nội dung.
* Quản lý truyền thông xã hội:
* Sử dụng mạng xã hội và các nền tảng trực tuyến khác để xây dựng cộng đồng fan hâm mộ
* Đánh giá, đo lường hiệu suất:
* Sử dụng các công cụ phân tích để theo dõi hiệu suất chiến dịch tiếp thị. Đánh giá số liệu về lượt xem, tương tác, và phản hồi để điều chỉnh chiến lược nếu cần thiết và đảm bảo hiệu quả cao nhất.

#### Quy trình sản xuất

A close up of a sign

Description automatically generated

Ảnh 3.5. Quy trình sản xuất

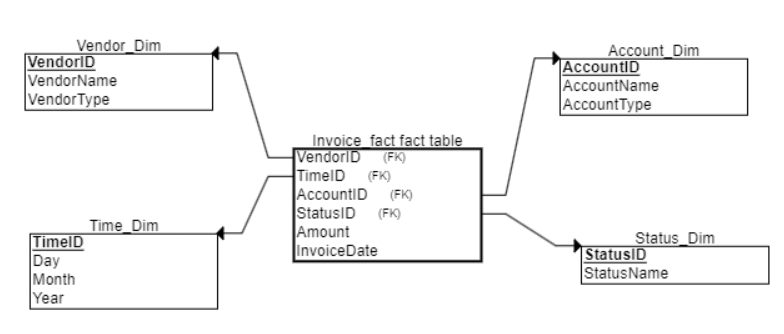
* Lên ý tưởng
* Tạo ý tưởng đưa ra một khái niệm hay ý tưởng cơ bản cho dự án phim.
* Lên kịch bản sơ lược cho bộ phim
* Ý tưởng có thể xuất phát từ các tác giả, đạo diễn, hoặc nhóm sáng tạo.
* Tiền kì
* Đề xuất ngân sách và kế hoạch sản xuất được xây dựng.
* Cần quảng bá và thu hút đối tác tài trợ hoặc nhà đầu tư.
* Sản xuất
* Quay phim, thu âm và thực hiện các công việc sản xuất.
* Bao gồm việc làm việc với đội ngũ diễn viên, đạo diễn, nhóm sản xuất và những người nghệ sĩ khác.
* Hậu kì
* Chỉnh sửa, thiết kế âm thanh, hiệu ứng hình ảnh và các công việc hậu kì khác.
* Chuẩn bị cho quá trình phát hành.
* Phát hành
* Xác định các kênh phát sóng hoặc rạp chiếu phim để phát hành.
* Tiếp thị và quảng cáo để thu hút khán giả.
* Đánh giá
* Người xem và giới phê bình đánh giá và đưa ra ý kiến về tác phẩm.
* Điều này có thể ảnh hưởng đến thành công của phim và cũng có thể có tác động lâu dài đến sự nghiệp của những người tham gia.

## Thiết kế Star Schema

Thiết kế Star Schema cho từng quy trình trong Business Process:

### Quy trình tài chính

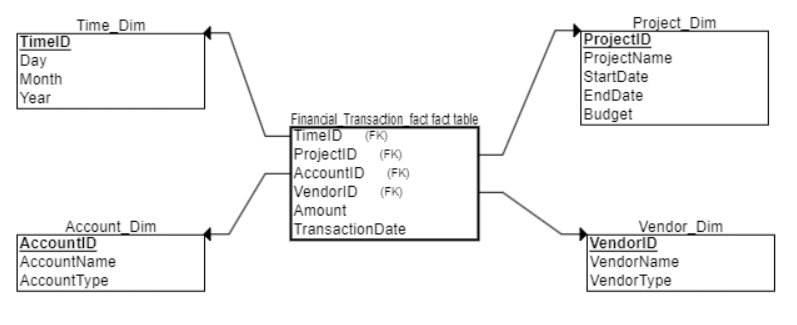
#### Quản lý hóa đơn



Ảnh 3.6. Quản lý hóa đơn

* Invoice\_Fact: @VendorID, @AccountID, @TimeID, Status, Amount)
* Vendor\_Dim: VendorID, VendorName, VendorType
* Account\_Dim: AccountID**,** AccountName, AccountType
* Time\_Dim: TimeID, Year, Month, Day
* Payment\_Status\_Dim: StatusID, StatusName

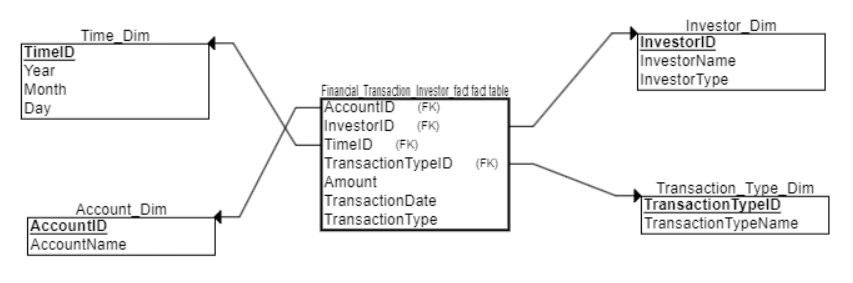
#### Quản lý ngân sách



Ảnh 3.7. Quản lý ngân sách

* Financial\_Transaction\_Fact: @ProjectID, @VendorID, @AccountID, @TimeID, Amount, TransactionDate
* Project\_Dim: ProjectID, ProjectName, StartDate, EndDate, Budget
* Vendor\_Dim: VendorID, VendorName, VendorType
* Account\_Dim: AccountID, AccountName, AccountType
* Time\_Dim: TimeID, Year, Month, Day

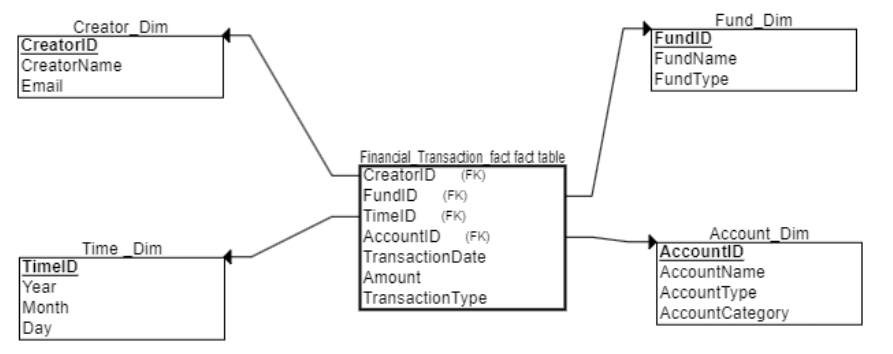
#### Quản lý nhà đầu tư



Ảnh 3.8. Quản lý nhà đầu tư

* Financial\_Transaction\_Investor\_Fact: @AccountID, @TimeID, TransactionDate, Amount, @TransactionTypeID, TransactionType
* Investor\_Dim: InvestorID, InvestorName, InvestorType
* Transaction\_Type\_Dim: TransactionTypeID, TransactionTypeName
* Time\_Dim:TimeID, Year, Month, Day
* Account\_Dim: AccountID, AccountName

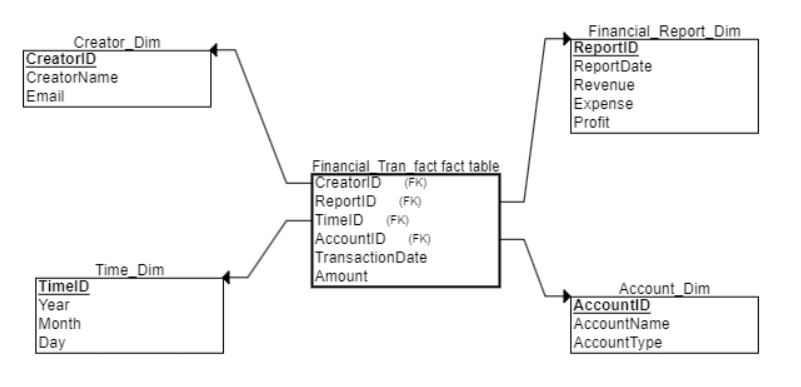
#### Quản lý thu chi và ngân quỹ



Ảnh 3.9. Quản lý thu chi và ngân quỹ

* Financial\_Transaction\_Fact: @CreatorID, @FundID, @TimeID, @AccountID, DateAmount, TransactionDate, Amount, TransactionType
* Account\_Dim: AccountID, AccountName, AccountType, AccountCategory
* Time Dimension: TimeID, Year, Month, Day
* Fund\_Dim: FundID, FundName, FundType
* Creator\_Dim: CreatorID, CreatorName, Email

#### Báo cáo tài chính

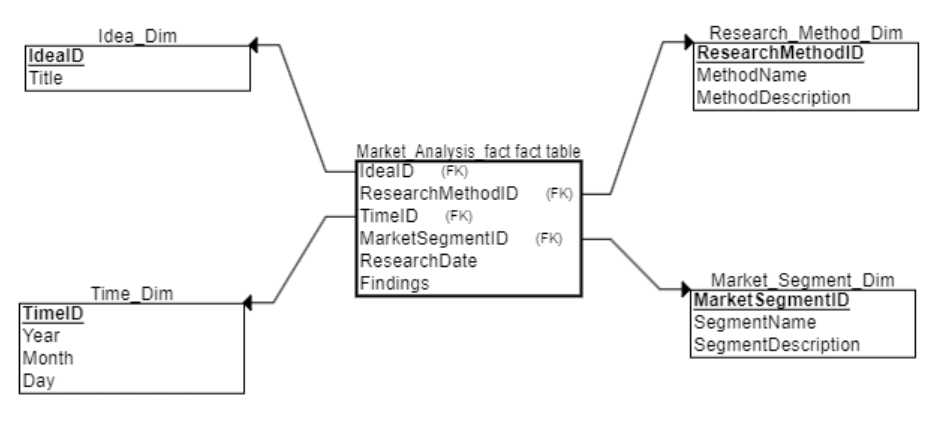


Ảnh 3.10. Báo cáo tài chính

* Financial\_Transaction\_Fact: @CreatorID,@TimeID, TransactionDate, Amount, @AccountID, @ReportID
* Creator\_Dim: CreatorID, CreatorName, Email
* Account\_Dim: AccountID, AccountName, AccountType
* Time\_Dim: TimeID, Year, Month, Day
* Financial\_Report\_Dim: ReportID, ReportDate, Revenue, Expense, Profit

### Quy trình kinh doanh

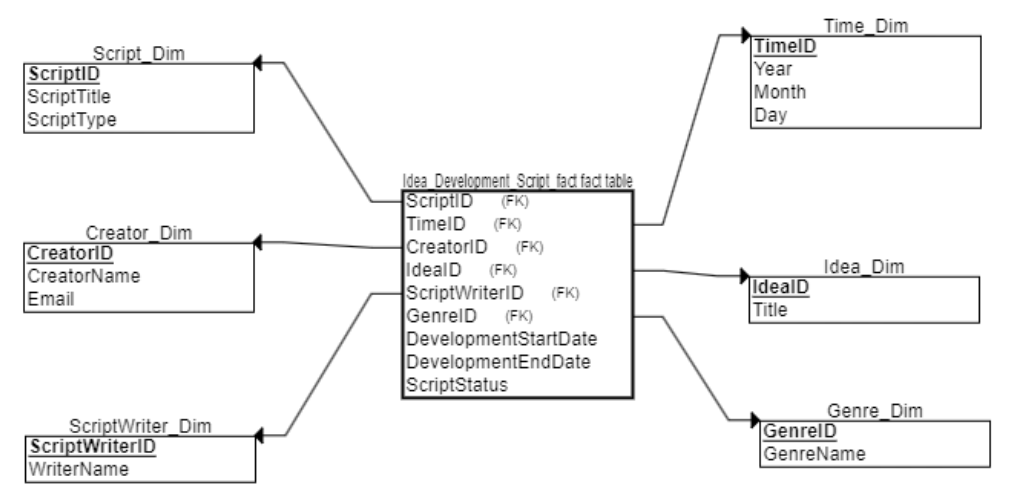
#### Phân tích và nghiên cứu thị trường



Ảnh 3.11. Phân tích và nghiên cứu thị trường

* Market\_Analysis\_Fact: @IdeaID, @MarketSegmentID, ResearchDate, @ResearchMethodID, Findings, @TimeID
* Idea\_Dim: IdeaID, Title
* Market\_Segment\_Dim: MarketSegmentID, SegmentName, SegmentDescription,
* Research\_Method\_Dim: ResearchMethodID, MethodName, MethodDescription
* Time\_Dim: TimeID, Year, Month, Day

#### Phát triển ý tưởng và kịch bản



Ảnh 3.12. Phát triển ý tưởng và kịch bản

* Idea\_Development\_Script\_Fact: @IdeaID, @ScriptID, DevelopmentStartDate, DevelopmentEndDate, ScriptStatus, @CreatorID,@ScriptWriterID, @GenreID, @TimeID
* Idea\_Dim: IdeaID, Title
* Creator\_Dim: CreatorID, CreatorName, Email
* Script\_Dim: ScriptID, ScriptTitle, ScriptType
* ScriptWriter\_Dim: ScriptWriterID, WriterName
* Genre\_Dim: GenreID, GenreName
* Time\_Dim: TimeID, Year, Month, Day

#### Lập kế hoạch sản xuất

A diagram with text and lines

Description automatically generated with medium confidence

Ảnh 3.13. Lập kế hoạch sản xuất

* Production\_Plan\_Fact: @PlanID, @IdeaID, @CrewID, @CreatorID, @LocationID, ProductionStartDate, ProductionEndDate, Budget, ActualCost
* Idea\_Dim: IdeaID, Title
* Creator\_Dim: CreatorID, CreatorName, Email
* Location\_Dim: LocationID, LocationName
* Time\_Dim: TimeID, Year, Month, Day
* Crew\_Dim: CrewID, CrewName, Role

#### Chiến lược Phát hành và quảng bá

A diagram of a data flow

Description automatically generated

Ảnh 3.14. Chiến lược phát hành và quảng bá

* Distribution\_Promotion\_Fact: @MovieID, @PromotionID, @TimeID ReleaseDate, BoxOfficeRevenue, @DistributionChannelID
* Movie\_Dim: MovieID, Title, ReleaseDate, Genre, Director
* DistributionChannel\_Dim: DistributionChannelID, ChannelName
* Promotion\_Dim:PromotionID, PromotionType, PromotionBudget, PromotionStartDate, PromotionEndDate
* Time\_Dim: TimeID, Year, Month, Day

#### **Báo cáo kinh doanh**

A diagram of a business report

Description automatically generated

Ảnh 3.15. Báo cáo kinh doanh

* Business\_Report\_Fact: @IdeaID, @CreatorID, @TimeID, TransactionDate, Revenue, Expense, Profit
* Idea\_Dim: IdeaID, Title
* Creator\_Dim: CreatorID, CreatorName, Email,
* Time\_Dim: TimeID, Year, Month, Day

### Quy trình nhân sự

#### Tuyển dụng và tuyển chọn

A diagram of a job

Description automatically generated

Ảnh 3.16. Tuyển dụng và tuyển chọn

* Recruitment\_Fact: @TimeID, @CandidateID, @PositionID, @RecruiterID, ApplicationDate, InterviewDate, OfferDate, HireDate
* Candidate\_Dim: CandidateID, CandidateName, Email, Phone, EducationLevel, Experience
* Position\_Dim: PositionID, PositionTitle, Department, JobDescription
* Recruiter\_Dim: RecruiterID, RecruiterName, RecruiterDepartment
* Time\_Dim: TimeID, Year, Month, Day

#### Đào tạo và phát triển nhân sự

A diagram of a training program

Description automatically generated

Ảnh 3.17. Đào tạo và phát triển nhân sự

* Training\_Development\_Fact: @EmployeeID, @DepartmentID, @CategoryID @CourseID, @TrainerID, @PositionID, TrainingDate, Duration
* Employee\_Dim: EmployeeID, EmployeeName
* Course\_Dim: CourseID, CourseName
* Department\_Dim: DepartmentID, DepartmentName
* Position\_Dim: PositionID, PositionName
* Course\_Category\_Dim: CategoryID, CategoryName
* Trainer\_Dim: TrainerID, TrainerName

#### Quản lý hiệu suất

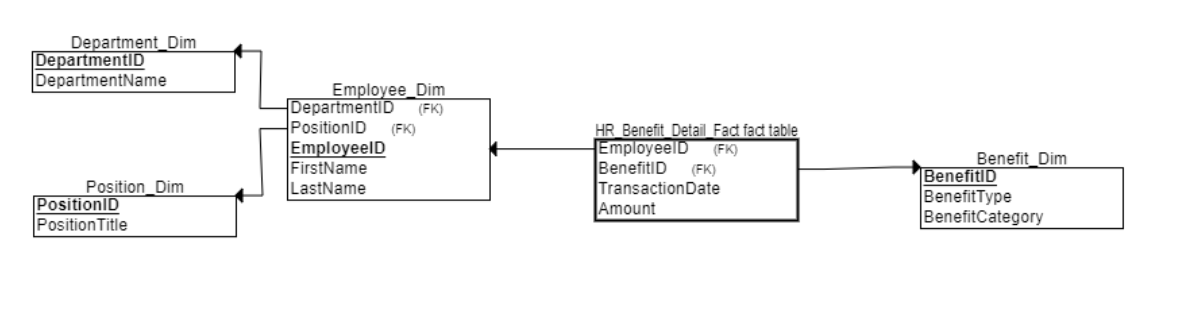
A diagram of a company

Description automatically generated

Ảnh 3.18. Quản lý hiệu suất

* Performance\_Evaluation\_Fact:@EmployeeID, EvaluationDate, Rating, @ManagerID, @TimeID, @PositionID, @DepartmentID
* Employee\_Dim: EmployeeID, EmployeeName
* Manager\_Dim: ManagerID, ManagerName
* Department\_Dim: DepartmentID, DepartmentName
* Position\_Dim : PositionID, PositionName
* Time\_Dim: TimeID, Year, Month, Day

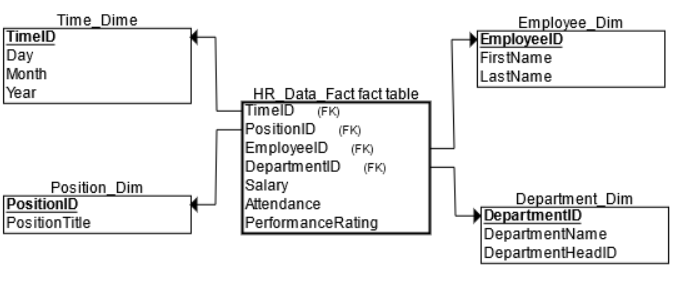
#### Chăm sóc nhân sự và phúc lợi:



Ảnh 3.19. Chăm sóc nhân sự và phúc lợi

* HR\_Benefit\_Detail\_Fact: @EmployeeID, @BenefitID, TransactionDate, Amount
* Employee\_Dim: EmployeeID, FirstName, LastName, @DepartmentID, @PositionID
* Department\_Dim: DepartmentID, DepartmentName
* Position\_Dim: PositionID, PositionTitle,
* Benefit\_Dim: BenefitID, BenefitType, BenefitCategori
* Time\_Dim: TimeID, Year, Month, Day

#### Báo cáo nhân sự

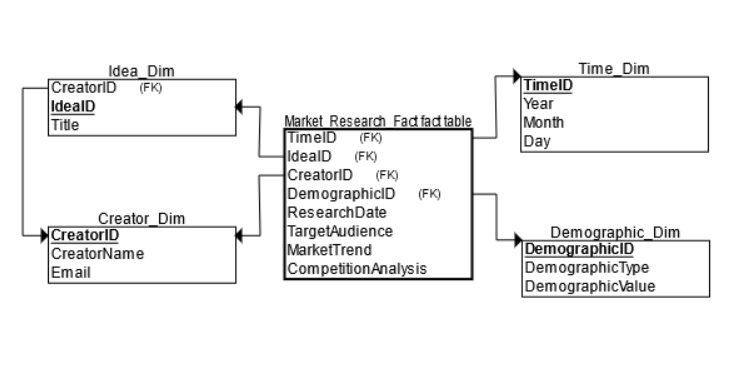


Ảnh 3.20. Báo cáo nhân sự

* HR\_Data\_Fact: @EmployeeID, @DepartmentID, @PositionID, @TimeID, Salary, Attendance, PerformanceRating
* Employee\_Dim: EmployeeID, FirstName, LastName
* Department\_Dim: DepartmentID, DepartmentName
* Position\_Dim: PositionID, PositionTitle,
* Time\_Dim: TimeID, Year, Month, Day

### Quy trình marketing

#### Nghiên cứu và phân tích thị trường



Ảnh 3.21. Nghiên cứu và phân tích thị trường

* Market\_Research\_Fact: @IdeaID, @TimeID, @CreatorID, @DemogrophicID, ResearchDate, TargetAudience, MarketTrend, CompetitionAnalysis
* Idea\_Dim: IdeaID, Title, @CreatorID
* Creator\_Dim: CreatorID, CreatorName, Email
* Time\_Dim: TimeID, Year, Month, Day
* Demographic\_Dim: DemographicID, DemographicType, DemographicValue

#### Quảng cáo và quảng bá

A diagram of a company

Description automatically generated

Ảnh 3.22. Quảng cáo và quảng bá

* Advertising\_Campaign\_Fact: @IdeaID, @CreatorID, @TimeID, @MediaID, @LocationID, Cost, RevenueGenerated
* Idea\_Dim: IdeaID, Title, @CreatorID
* Creator\_Dim: CreatorID, CreatorName, Email
* Time\_Dim: TimeID, Year, Month, Day
* Location\_Dim: LocationID, LocationName,
* Media\_Dim: MediaID, MediaName

#### Xác định chiến lược tiếp thị

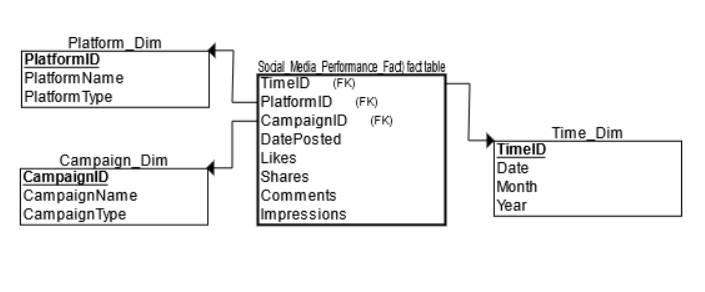
A diagram of a business process

Description automatically generated

Ảnh 3.23. Xác định chiến lược tiếp thị

* Marketing\_Strategy\_Fact: @IdeaID, @CreatorID, @TimeID, @ChannelID, @AudienceID, StrategyName, StartDate, EndDate, Budget
* Idea\_Dim: IdeaID, Title, @CreatorID
* Creator\_Dim: CreatorID, CreatorName, Email
* Time\_Dim: TimeID, Year, Month, Day
* Channel\_Dim: ChannelID, ChannelName,
* Audience\_Dim: AudienceID, AudienceName

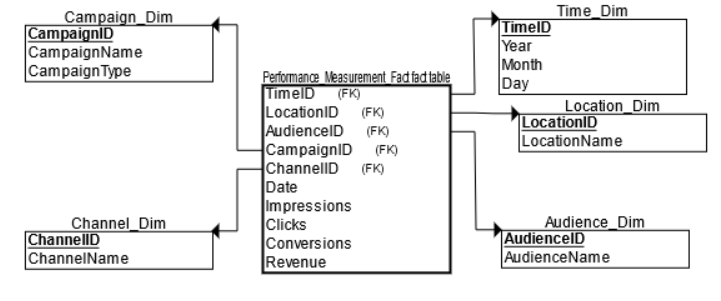
#### Quản lý truyền thông xã hội



Ảnh 3.24. Quản lý truyền thông xã hội

* Social\_Media\_Performance\_Fact: @TimeID, @CampaignID, @PlatformID, DatePosted, Likes, Shares, Comments, Impressions
* Campaign\_Dim: CampaignID, CampaignName, CampaignType
* Platform\_Dim: PlatformID, PlatformName, PlatformType
* Time\_Dim: TimeID, Year, Month, Day

#### Đánh giá và đo lường hiệu suất



Ảnh 3.25. Đánh giá và đo lường hiệu suất

* Performance\_Measurement\_Fact: @TimeID, @LocationID, @AudienceID, @ChannelID @CampaignID, Date, Impressions, Clicks, Conversions, Revenue
* Campaign\_Dim: CampaignID, CampaignName, CampaignType
* Channel\_Dim: ChannelID, ChannelName,
* Time\_Dim: TimeID, Year, Month, Day
* Location\_Dim : LocationID, LocationName
* Audience\_Dim : AudienceID, AudienceName

### Quy trình sản xuất

#### Lên ý tưởng

A diagram of a diagram

Description automatically generated

Ảnh 3.26. Lên ý tưởng

* Idea\_fact fact: @LocationID**, @**StatusID**, @**CreatorID**, @**TimeID**, @**GenreID, Title, Description, Date\_Created)
* Location\_Dim: LocationID, LocationName
* Status\_Dim: StatusID, StatusName
* Genre\_Dim: GenreID, GenreName
* Creator\_Dim: CreatorID, CreatorName, Email
* Time\_Dim: TimeID, Year, Month, Day

#### Tiền kì

A diagram of a budget

Description automatically generated

Ảnh 3.27. Tiền kì

* Budget\_Proposal\_fact fact: @IdeaID**, @**TimeID**, @**LocationID**, @**CategoryID**, @**StatusID, BudgetAmount, Date\_Submitted
* Idea\_Dim: IdeaID, Title
* Location\_Dim: LocationID, LocationName)
* Status\_Dim: StatusID, StatusName
* Time\_Dim: TimeID, Year, Day, Month
* Budget\_Category\_Dim: CategoryID, CategoryName

#### Sản xuất

A diagram of a company

Description automatically generated

Ảnh 3.28. Sản xuất

* Production\_Order\_Fact: @LocationID**, @**StatusID**, @**CrewID**, @**IdeaID**, @**EquipmentID**, @**TimeID, BudgetSpent, StartTime, EndTime)
* Location\_Dim: LocationID, LocationName
* Idea\_Dim: IdeaID, Title
* Time\_Dim: TimeID, Year, Month, Day
* Status\_Dim: StatusID, StatusName
* Crew\_Dim: CrewID, CrewName, Role
* Equiment\_Dim: EquimentID, EquimentName, Type

#### Hậu kì

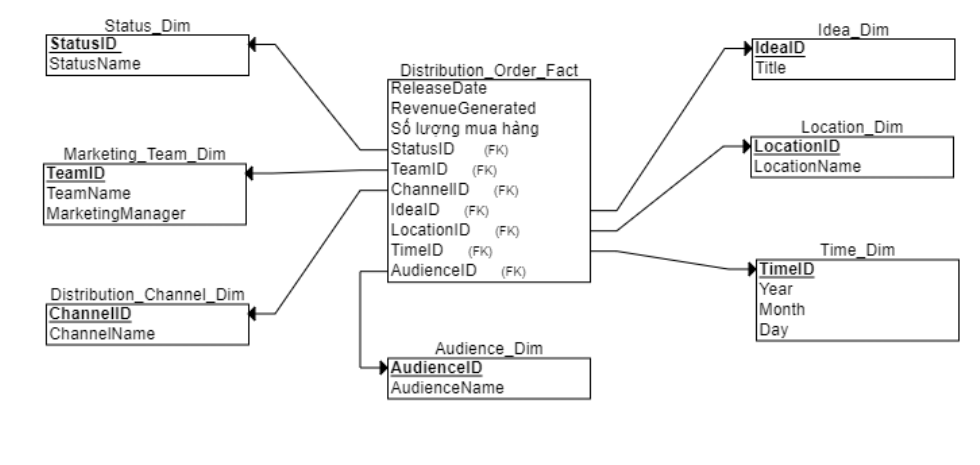
A diagram of a computer

Description automatically generated

Ảnh 3.29. Hậu kì

* Post\_Production\_Order\_Fact: @LocationID**, @**IdeaID**, @**TimeID, @StatusID, @TeamID, @SoundID, @VFXID, StarTime, EndTime, BudgetSpent)
* Location\_Dim: LocationID, LocationName
* Idea\_Dim: IdeaID, Title
* Time\_Dim: TimeID, Year, Month, Day
* Status\_Dim: StatusID, StatusName
* Editing\_Team\_Dim: TeamID, TeamName, EditorName
* Sound\_Design\_Dim: SoundID, SoundName, EngineerName
* VFX\_Dim: VFXID, VFXName, ArtisName

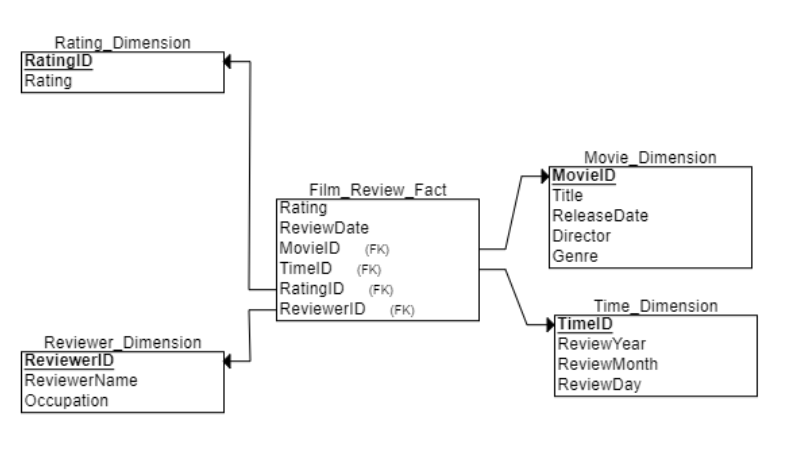
#### Phát hành



Ảnh 3.30. Phát hành

* Distribution\_Order\_Fact: @OrderID, @IdeaID, @LocationID, RevenueGenerated, @StatusID
* Idea\_Dim: IdeaID, Title
* Location\_Dim : LocationID, LocationName
* Time\_Dim: TimeID, Year, Month, Day
* Status\_Dim: StatusID, StatusName
* Marketing\_Team\_Dim: TeamID, TeamName, MarketingManager
* Distribution\_Channel\_Dim: ChannelID, ChannelName
* Audience\_Dim: AudienceID, AudienceName

#### Đánh giá

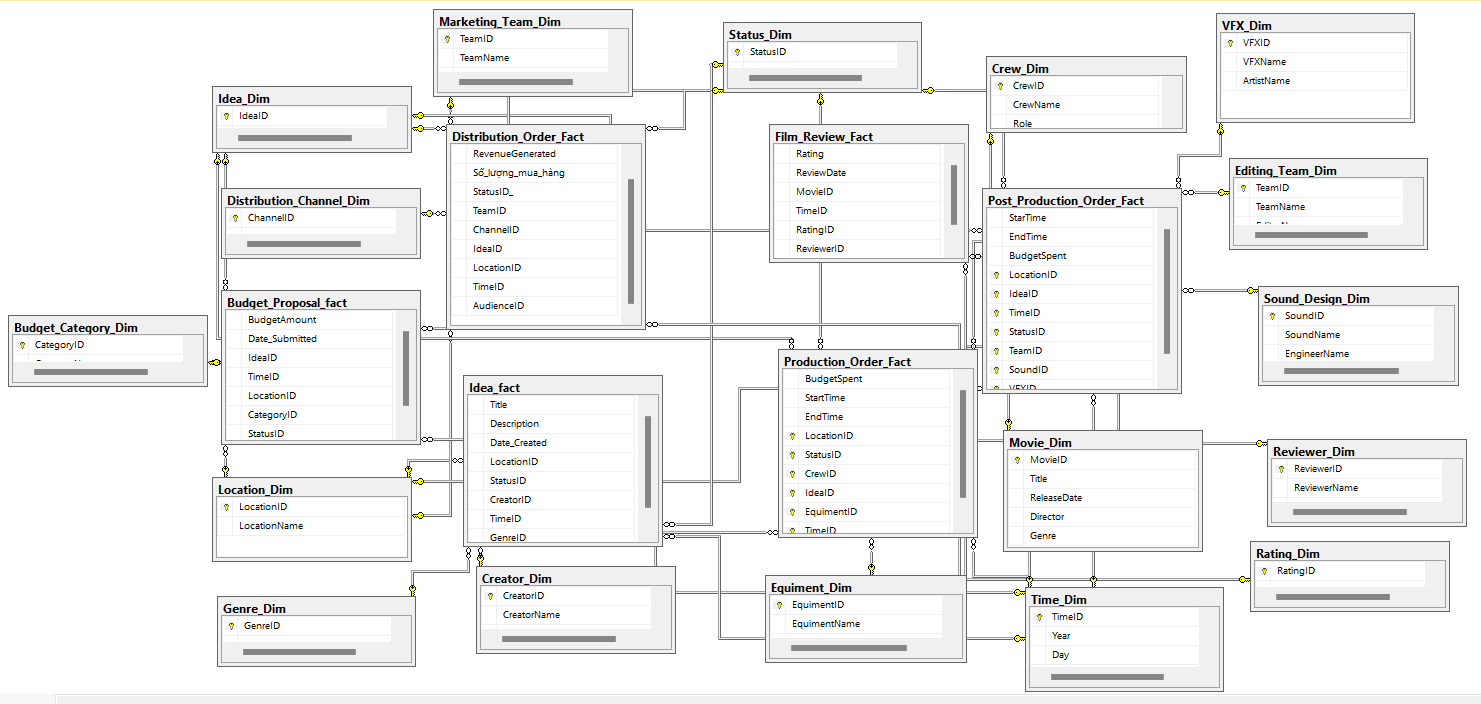


Ảnh 3.31. Đánh giá

* Film\_Review\_Fact: @ReviewID, @MovieID, @ReviewerID, Rating, ReviewDate
* Movie\_Dimension: MovieID, Title, ReleaseDate, Director, Genre
* Reviewer\_Dimension: ReviewerID, ReviewerName, Occupation
* Rating\_Dimension: RatingID, Rating
* Time\_Dimension: TimeID, ReviewYear, ReviewMonth, ReviewDay

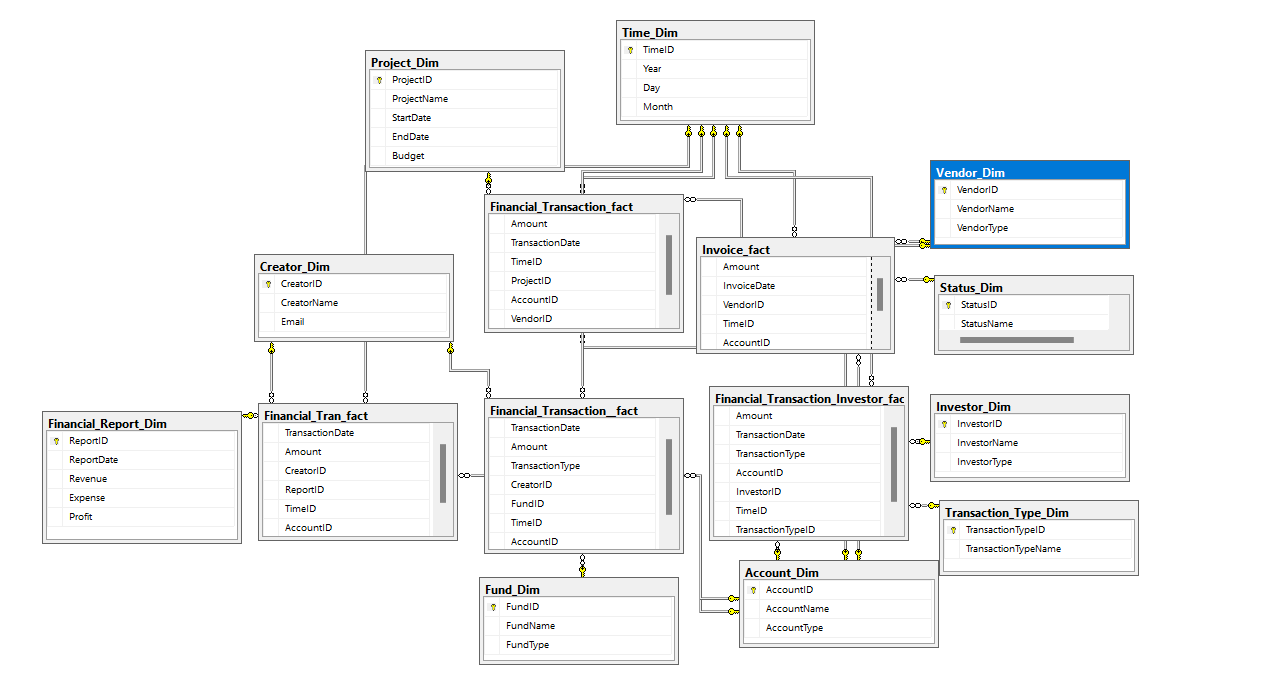
## Xây dựng SQL cho các business process

### Quy trình sản xuất



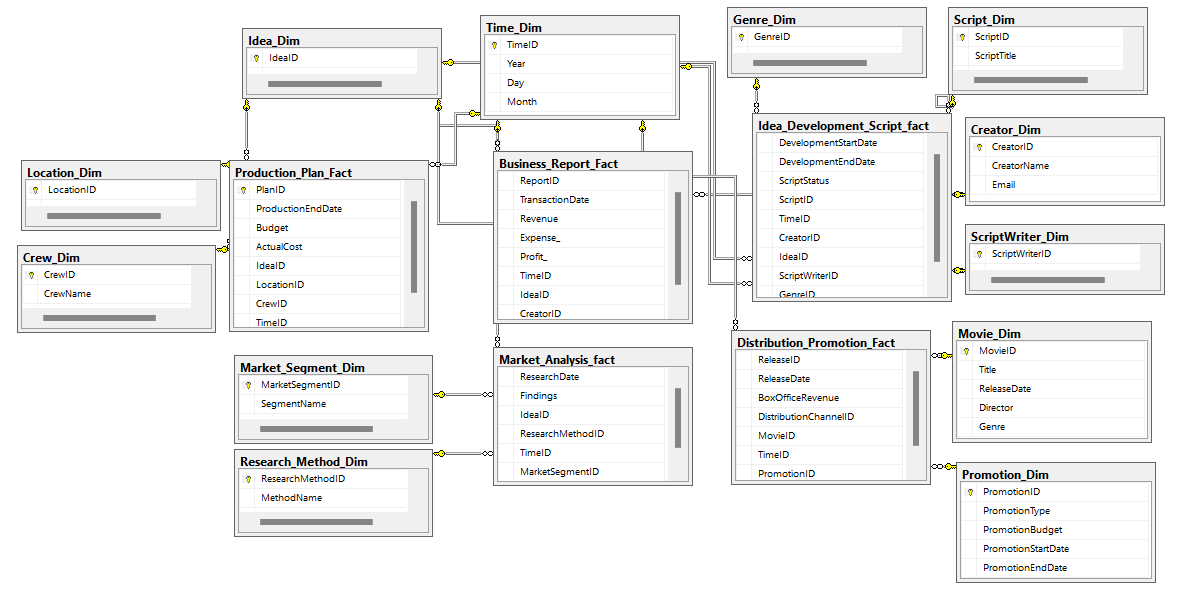
Ảnh 3.32. SQL cho quy trình sản xuất

### Quy trình tài chính



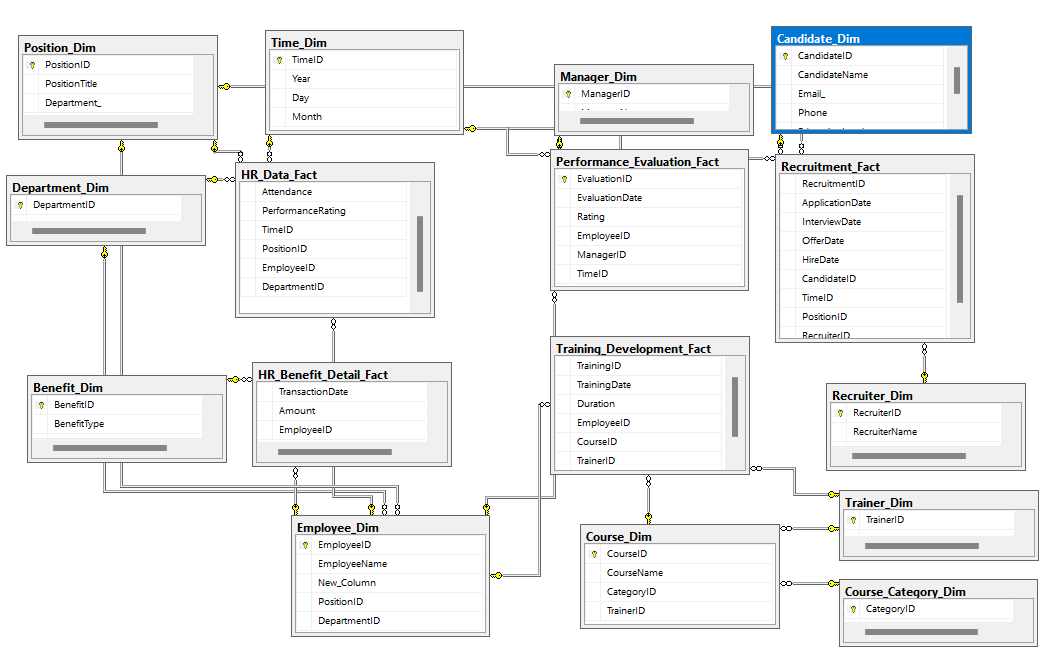
Ảnh 3.33. SQL cho quy trình tài chính

### Quy trình kinh doanh



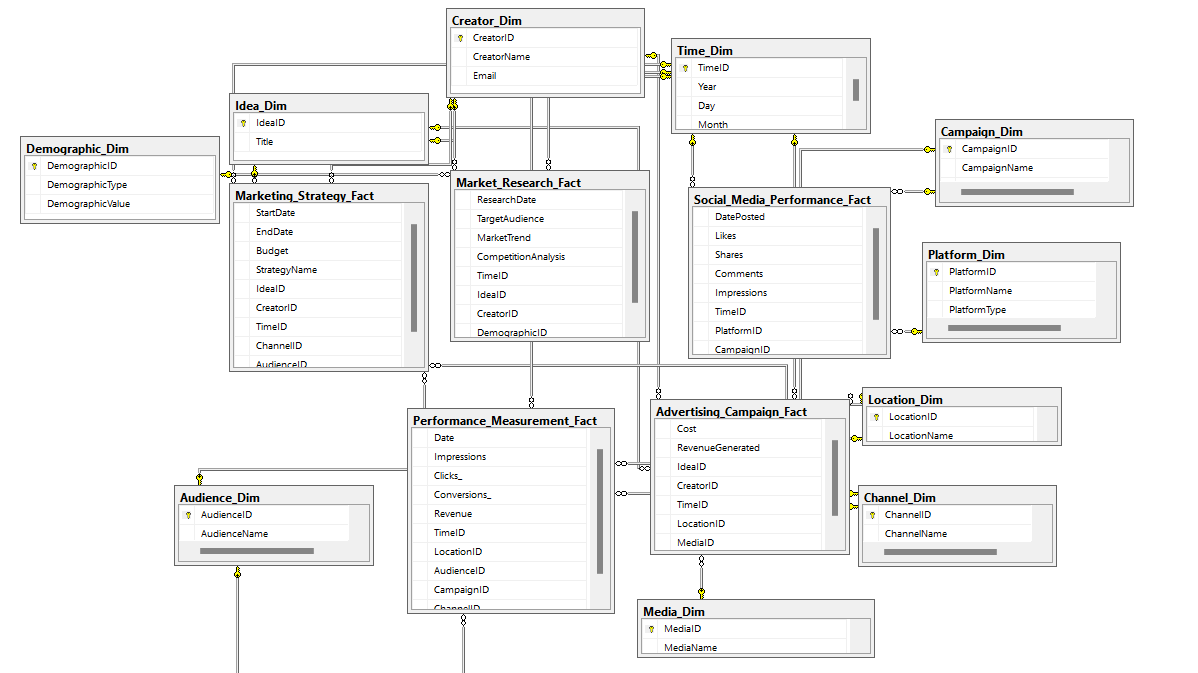
Ảnh 3.34. SQL cho quy trình kinh doanh

### Quy trình nhân sự



Ảnh 3.35. SQL cho quy trình nhân sự

### Quy trình Marketing



Ảnh 3.36. SQL cho quy trình Marketing

## Xây dựng ETL cho bài toán

Công ty chúng em muốn hướng đến là công ty về điện ảnh, bao gồm rất nhiều các quy trình như đã nêu trên. Do đó để thuận tiện cho việc demo chúng em chỉ thiết kế xây dựng ETL cho phát hành trong quy trình sản xuất.

### Làm sạch dữ liệu

A screen shot of a computer

Description automatically generated

Ảnh 3.37. ETL trong giai đoạn làm sạch dữ liệu

### Xây dựng Dim và Fact

A screenshot of a computer

Description automatically generated

Ảnh 3.38. Thiết kế ETL cho dim và fact

### Chi tiết từng bảng

#### Dim\_Product

A computer screen shot of a computer

Description automatically generated

Ảnh 3.39. Thiết kế chi tiết ETL cho bảng Dim\_Product

#### Dim\_Audience

A diagram of a flowchart

Description automatically generated

Ảnh 3.40. Thiết kế bảng Dim\_Audience

#### Dim\_Profit

A diagram of a flowchart

Description automatically generated

Ảnh 3.41. Thiết kế bảng Dim\_Profit

#### Fact\_PhatHanh

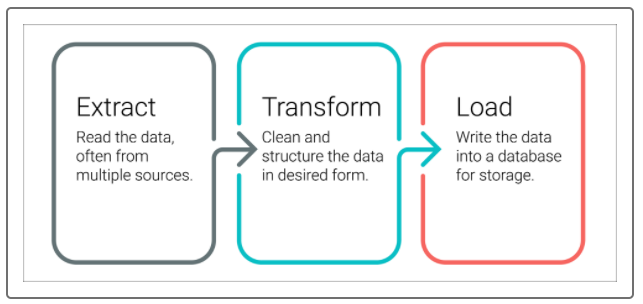
A diagram of a product

Description automatically generated

Ảnh 3.42. Thiết kế bảng Fact\_PhatHanh

# Demo Xây dựng etl

Thực hành xây dựng quy trình ETL cho phần phát hành trong quy trình sản xuất



Ảnh 4.1. Quy trình ETL

## Mục đích

Để cung cấp tập dữ liệu có tổ chức và sạch sẽ, dự án này tập trung vào quy trình trích xuất, chuyển đổi và tải bao gồm:

* Trích xuất dữ liệu từ hai nguồn khác nhau.
* Web scrape của Wikipedia cho tất cả các bộ phim phát hành từ năm 1990
* Dữ liệu từ trang web Kaggle để biết dữ liệu xếp hạng.
* Chuyển đổi dữ liệu bằng mô-đun Jupyter Notebook, Python, Pandas và Python RegEx.
* Tải dữ liệu bằng PostgreSQL và pgAdmin để lưu trữ tập dữ liệu được làm sạch cuối cùng.

Về dữ liệu đã sử dụng 3 tập là:

* Dữ liệu Movies\_metadata.csv có 45.572 dòng và 24 cột
* Dữ liệu Ratings.csv có 100.837 dòng và 4 cột
* Dữ liệu Wikipedia-movies.json có 7.312 dòng và 17 cột

## Extract

* Thiết lập các thư viện cần dùng

A screenshot of a computer

Description automatically generated

Ảnh 4.2. Thư viện sử dụng để lập trình

* Tạo hàm đối số để đọc dữ liệu từ 3 file đã thu thập được

A screenshot of a computer program

Description automatically generated

Ảnh 4.3. Hàm extract\_transform\_load để đọc các file

* Đặt Dataframes và hiển thị các bảng

A screenshot of a computer

Description automatically generated

Ảnh 4.4. Dataframes wiki\_movies\_df

A screenshot of a computer

Description automatically generated

Ảnh 4.5. Dataframes kaggle\_metadata

A screenshot of a computer

Description automatically generated

Ảnh 4.6. Dataframes ratings

## Transform

Sử dụng các hàm như:

* Clean\_movie để kết hợp các dữ liệu rải rác của các ngôn ngữ thay thế thành một cột alt\_titles.
* Chang\_column\_name để tổ chức các tên cột thành mẫu nhất quán
* Các phương thức apply () và map () kết hợp với các hàm lambda để ép kiểu dữ liệu.

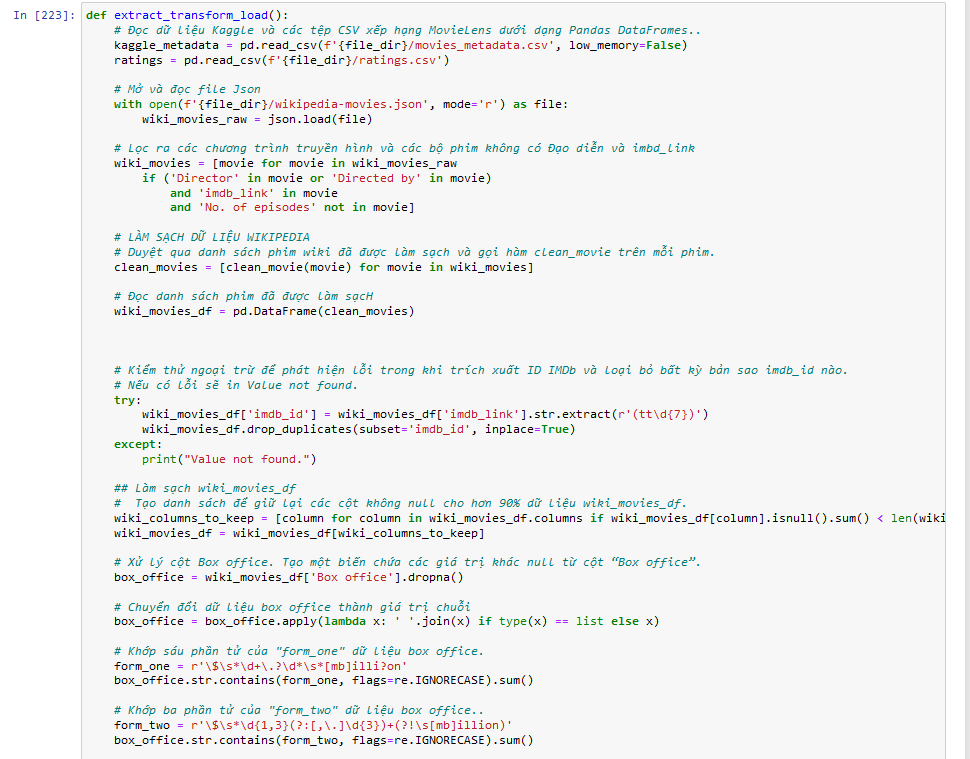
A screenshot of a computer program

Description automatically generated

Ảnh 4.7. Sử dụng hàm clean\_movie và change\_column\_name

Sử dụng hàm extract\_transform\_load để làm sạch dữ liệu Kaggle và wikipedia bao gồm:

* Thay đổi kiểu dữ liệu, sử dụng các toán tử so sánh các phương thức pd.to\_numeric, astype () và python cho các kiểu Boolean.
* Điền các giá trị còn thiếu và lọc các cột không mong muốn.
* Hợp nhất các khung dữ liệu bằng phương pháp pd\_merge.



A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

## Load

* Kết nối với cơ sở dữ liệu bằng thư viện giả kim thuật và phương thức to\_sql.
* Quá trình ETL hoàn chỉnh có thể được thực hiện với một hàm duy nhất extract\_transform\_load gọi.

A screenshot of a computer code

Description automatically generated

Ảnh 4.8. Kết nối với database

A screenshot of a computer code

Description automatically generated

* Kết quả dữ liệu khi được truyền vào PostgreSQL
* Dữ liệu được truyền vào PostgreSQL và có thể sử dụng được
* Dữ liệu được truyền vào database Movie\_data bao gồm 2 bảng là Movies và Ratings

A screenshot of a computer

Description automatically generated

Ảnh 4.9. Kết quả thực hiện quy trình ETL

# Phân chia công việc

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Mã sinh viên** | **Họ và tên** | **Chức vụ** | **Nhiệm vụ** | **Hoàn thành** |
| 1 | A41529 | Đỗ Thị Lan | Trường nhóm | Xây dựng business process, thiết kế star schemae cho marketing và sản xuất. Demo Xây dựng ETL | 100% |
| 2 | A41641 | Vũ Kiều Trang | Thành viên | Thiết kế star schema cho nhân sự. Thiết kế ETL cho bài toán. Tổng hợp CSDL database | 100% |
| 3 | A43193 | Lê Thị Mỹ Hảo | Thành viên | Thiết kế star schema cho tài chính và kinh doanh. Làm word và powerpoint | 100% |