****

**HỌC VIỆN NGÂN HÀNG**

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

****

**BÁO CÁO BÀI TẬP LỚN**

**MÔN KHO DỮ LIỆU VÀ KINH DOANH THÔNG MINH**

**ĐỀ TÀI: ỨNG DỤNG DATA WAREHOUSE VÀ BI TRONG VIỆC TỐI**

**ƯU HOÁ CHIẾN LƯỢC KINH DOANH CỦA NETFLIX**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn** | : T.S Đinh Trọng Hiếu |
| **Lớp** | : 241IS24A03 |
| **Nhóm** | : 12 |

**Hà Nội – 2024**

****

**HỌC VIỆN NGÂN HÀNG**

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

****

**BÁO CÁO BÀI TẬP LỚN**

**MÔN KHO DỮ LIỆU VÀ KINH DOANH THÔNG MINH**

**ĐỀ TÀI: ỨNG DỤNG DATA WAREHOUSE VÀ BI TRONG VIỆC TỐI**

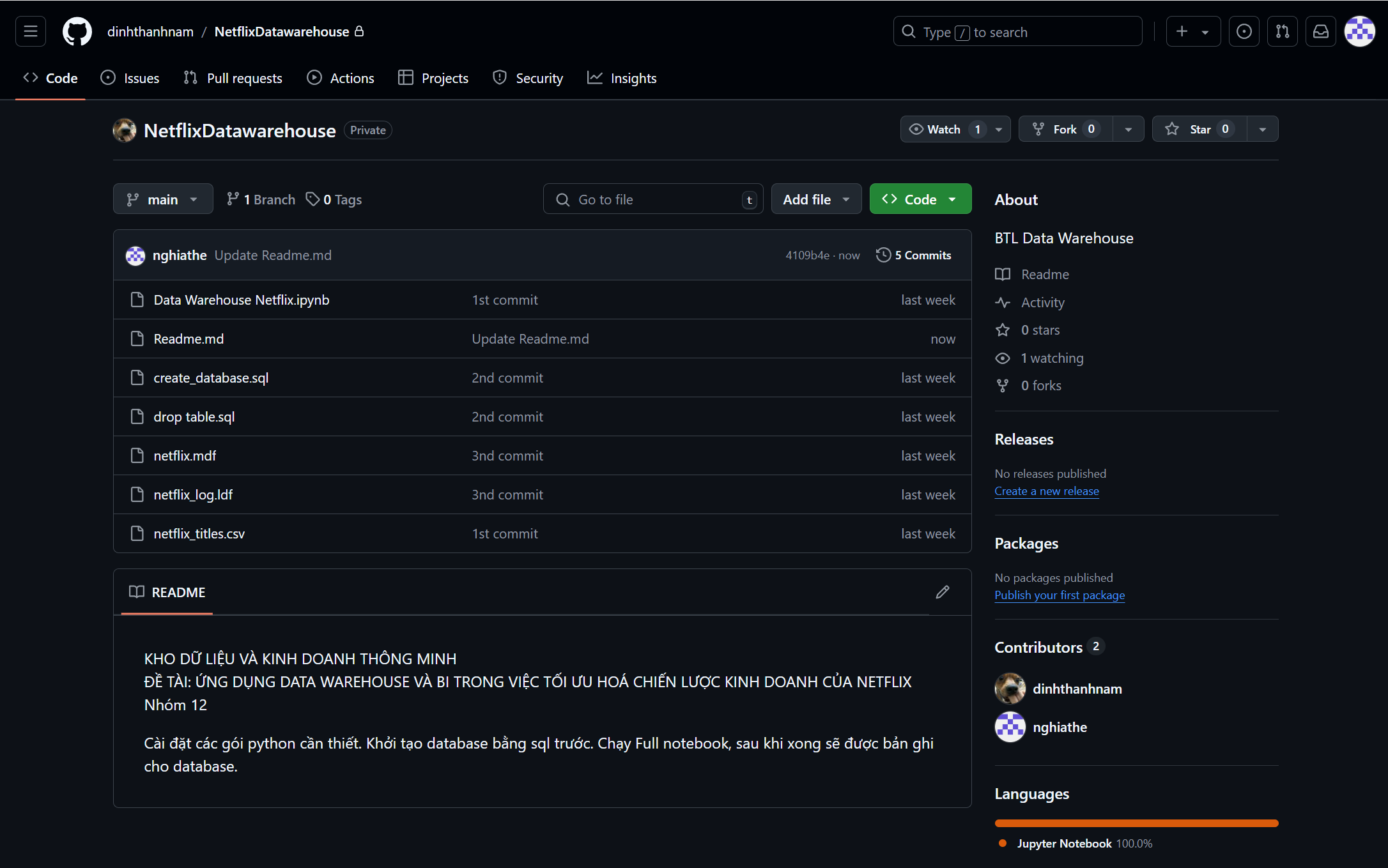
**ƯU HOÁ CHIẾN LƯỢC KINH DOANH CỦA NETFLIX**

|  |  |  |  |
| --- | --- | --- | --- |
| **Thành viên** | **Mã sinh viên** | **Tỉ lệ đóng góp** | **Xác nhận** |
| Đinh Thành Nam  (Nhóm trưởng) | 24A4042599 |  |  |
| Nguyễn Duy Hưng | 24A4042596 |  |  |
| Nguyễn Thế Nghĩa | 24A4042602 |  |  |
| Trần Quốc Việt |  |  |  |
| Trương Nhật Dũng |  |  |  |

**Hà Nội – 2024**

**HOẠT ĐỘNG NHÓM**

**1. Theo dõi tiến độ trên Github**

****

**2. Hoạt động nhóm**

Nhóm hoạt động thường xuyên qua kênh discord để trao đổi và phát triển bài tập:

# LỜI CẢM ƠN

Lời đầu tiên, nhóm 6 xin gửi lời cảm ơn đến Học viện Ngân hàng và khoa Công nghệ thông tin và kinh tế số đã tạo điều kiện cho chúng em và các bạn sinh viên có cơ hội được học tập trong một môi trường năng động, sáng tạo, tạo điều kiện nỗ lực và phát huy hết tiềm năng của mình. Nhóm xin gửi lời cảm ơn đặc biệt đến thầy Đinh Trọng Hiếu đã dành sự tâm huyết và nhiệt tình giảng dạy, giúp chúng em học được một môn học có ứng dụng cao trong ngành học và nghiệp vụ sau này, đáp ứng nhu cầu thực tiễn của các nhà tuyển dụng về kiến thức, kỹ năng cần có.

Tiếp nối các học phần trước đó, học phần này giúp sinh viên được thực hiện và quản lý một dự án CNTT của riêng mình, từ đó hiểu hơn về cách thức hoạt động và triển khai một dự án CNTT trong môi trường chuyên nghiệp. Song vì chưa có nhiều kinh nghiệm nên bài báo cáo sẽ không tránh khỏi những thiếu sót, kính mong cô nhận xét, góp ý để bài báo cáo của chúng em được hoàn thiện hơn, rút kinh nghiệm để làm tốt hơn trong tương lai.

Chúng em xin chân thành cảm ơn!

Nhóm sinh viên thực hiện

Nhóm 12

# LỜI CAM ĐOAN

Nhóm 12 xin cam đoan báo cáo này là sản phẩm nghiên cứu và thực hành của nhóm. Bài báo cáo đảm bảo tính liêm chính trong học tập, không đạo văn, gian lận, bịa đặt. Các thông tin tham khảo được trích dẫn nguồn đầy đủ và minh bạch.

Nhóm 12 xin chịu toàn bộ trách nhiệm nếu bài báo cáo vi phạm các điều trên.

Hà Nội, ngày 8 tháng 10 năm 2024

Đại diện nhóm 12

# MỤC LỤC

[LỜI CẢM ƠN 2](#_Toc179931526)

[LỜI CAM ĐOAN 3](#_Toc179931527)

[MỤC LỤC 4](#_Toc179931528)

[DANH MỤC BẢNG, HÌNH VẼ 5](#_Toc179931529)

[LỜI MỞ ĐẦU 6](#_Toc179931530)

[CHƯƠNG 1. TÌM HIỂU TỔNG QUAN VỀ HỆ THỐNG KINH DOANH THÔNG MINH (BUSINESS INTELLIGENCE) 7](#_Toc179931531)

[1. Khái niệm về Business Intelligence (BI) 7](#_Toc179931532)

[2. Các thành phần của Business Intelligence 7](#_Toc179931533)

[3. Cách thức hoạt động của Business Intelligence 7](#_Toc179931534)

[CHƯƠNG 2. GIỚI THIỆU VỀ NETFLIX 9](#_Toc179931535)

[1. Giới thiệu về Netflix 9](#_Toc179931536)

[2. Mô hình kinh doanh của Netflix 9](#_Toc179931537)

[3. Hoạt động nghiệp vụ của Netflix 10](#_Toc179931538)

[3.1. Sản xuất và cấp phép nội dung 10](#_Toc179931539)

[CHƯƠNG 3. XÁC ĐỊNH NGƯỜI SỬ DỤNG VÀ KHẢO SÁT VIỆC NHU THẬP, TỔ CHỨC, LƯU TRỮ DỮ LIỆU CỦA DOANH NGHIỆP 17](#_Toc179931540)

[1. Sơ đồ tổ chức của Netflix 17](#_Toc179931541)

[2. Người sử dụng BI (BI User) trong tổ chức Netflix 18](#_Toc179931542)

[3. Phân tích các quyết định kinh doanh của BI Users 20](#_Toc179931543)

[CHƯƠNG 4. XÂY DỰNG KHO DỮ LIỆU 22](#_Toc179931544)

[1. Data Warehouse là gì 22](#_Toc179931545)

[2. ETL trong xây dựng Data Warehouse 22](#_Toc179931546)

[3. Các công cụ hỗ trợ ETL 23](#_Toc179931547)

[4. Phương pháp mà nhóm áp dụng: Python thuần 23](#_Toc179931548)

[5. Phân tích dữ liệu 24](#_Toc179931549)

[6. Tạo các bảng tương ứng trong CSDL 25](#_Toc179931550)

[7. Thực hiện các thao tác ETL 27](#_Toc179931551)

[7.1. Extract & Transform 27](#_Toc179931552)

[7.2. Load 29](#_Toc179931553)

[CHƯƠNG 5. KẾT LUẬN 34](#_Toc179931554)

[TÀI LIỆU THAM KHẢO 35](#_Toc179931555)

# DANH MỤC BẢNG, HÌNH VẼ

[Bảng 1 Nghiệp vụ sản xuất và cấp phép nội dung 11](#_Toc179931296)

[Bảng 2 Nghiệp vụ phân phối và truyền phát nội dung 12](#_Toc179931297)

[Bảng 3 Nghiệp vụ phát triển và tối ưu hóa dịch vụ 14](#_Toc179931298)

[Bảng 4 Nghiệp vụ tiếp thị và phát triển người dùng 16](#_Toc179931299)

[Bảng 5 Người sử dụng BI (BI User) trong tổ chức Netflix 20](#_Toc179931300)

[Bảng 6 Phân tích các quyết định kinh doanh của BI Users 21](#_Toc179931301)

[Bảng 7 Trường dữ liệu 24](#_Toc179931302)

[Hình 1 Sơ đồ tổ chức 17](#_Toc179931342)

[Hình 2 Lược đồ cho kho dữ liệu 26](#_Toc179931343)

# LỜI MỞ ĐẦU

Trong thời đại cách mạng công nghệ 4.0, việc quản lý và khai thác dữ liệu trở thành một yếu tố quan trọng đối với sự phát triển bền vững của các doanh nghiệp. Netflix, một trong những công ty hàng đầu trong ngành công nghiệp giải trí trực tuyến, đã chứng tỏ sức mạnh của dữ liệu trong việc tối ưu hóa hoạt động kinh doanh và nâng cao trải nghiệm người dùng.

Dựa trên nhu cầu không ngừng tăng về việc phân tích dữ liệu, việc xây dựng một kho dữ liệu toàn diện, hiệu quả để quản lý, lưu trữ và truy xuất dữ liệu là vô cùng cần thiết. Kho dữ liệu này không chỉ giúp Netflix tối ưu hóa quá trình khuyến nghị phim cho người dùng, mà còn hỗ trợ việc phân tích hành vi người tiêu dùng, dự đoán xu hướng thị trường và ra quyết định chiến lược.

Trong bài tập lớn này, chúng em sẽ tập trung vào việc thiết kế và xây dựng một hệ thống kho dữ liệu cho Netflix. Hệ thống này sẽ được tối ưu hóa cho các mục tiêu như cải thiện hiệu quả truy xuất thông tin, hỗ trợ phân tích kinh doanh và cung cấp các công cụ báo cáo thông minh. Qua đó, bài tập không chỉ giúp người học củng cố kiến thức về kho dữ liệu, mà còn mang lại cái nhìn thực tế về ứng dụng công nghệ này trong ngành giải trí số.

# CHƯƠNG 1. TÌM HIỂU TỔNG QUAN VỀ HỆ THỐNG KINH DOANH THÔNG MINH (BUSINESS INTELLIGENCE)

## 1. Khái niệm về Business Intelligence (BI)

Business Intelligence (BI), hay **Kinh doanh thông minh**, là tập hợp các quy trình, công nghệ, công cụ và ứng dụng giúp doanh nghiệp thu thập, phân tích, xử lý và trực quan hóa dữ liệu từ các nguồn khác nhau. Mục tiêu của BI là cung cấp thông tin chi tiết có giá trị, hỗ trợ doanh nghiệp ra quyết định chiến lược và nâng cao hiệu quả hoạt động.

BI không chỉ đơn giản là thu thập dữ liệu mà còn là quá trình biến dữ liệu thô thành thông tin có ý nghĩa, giúp nhà quản lý hiểu rõ hơn về hoạt động kinh doanh, xu hướng thị trường và nhu cầu của khách hàng.

## 2. Các thành phần của Business Intelligence

BI bao gồm nhiều thành phần, mỗi thành phần đóng một vai trò quan trọng trong quy trình thu thập, xử lý và phân tích dữ liệu. Các thành phần chính của BI gồm:

* **Data Warehouse (Kho dữ liệu):** Kho dữ liệu là nơi lưu trữ các dữ liệu thô từ nhiều nguồn khác nhau trong một hệ thống tập trung, giúp cho việc truy xuất và phân tích dữ liệu trở nên dễ dàng hơn.
* **ETL (Extract, Transform, Load):** Là quy trình trích xuất, chuyển đổi và tải dữ liệu. ETL lấy dữ liệu từ các hệ thống nguồn (Extract), chuyển đổi thành dạng phù hợp với hệ thống đích (Transform), và sau đó tải dữ liệu vào kho dữ liệu (Load).
* **Data Mining (Khai phá dữ liệu):** Công cụ dùng để phát hiện các mẫu ẩn và mối quan hệ trong dữ liệu. Data mining hỗ trợ doanh nghiệp trong việc tìm kiếm các xu hướng tiềm ẩn, giúp dự đoán các kết quả kinh doanh trong tương lai.
* **OLAP (Online Analytical Processing):** OLAP là công cụ cho phép phân tích đa chiều dữ liệu theo các khía cạnh khác nhau, giúp các nhà quản lý có thể so sánh, đối chiếu dữ liệu từ nhiều góc độ.
* **Reporting & Dashboards:** Là công cụ tạo ra các báo cáo và bảng điều khiển trực quan, giúp nhà quản lý dễ dàng theo dõi hiệu suất hoạt động, các chỉ số chính và xu hướng kinh doanh.
* **Data Visualization (Trực quan hóa dữ liệu):** Sử dụng các biểu đồ, đồ thị, bản đồ và các công cụ hình ảnh khác để trình bày dữ liệu một cách dễ hiểu, giúp người sử dụng không chuyên sâu về dữ liệu cũng có thể nắm bắt được thông tin.

## 3. Cách thức hoạt động của Business Intelligence

Quy trình hoạt động của BI thường diễn ra qua các bước sau:

* **Bước 1: Thu thập dữ liệu từ nhiều nguồn khác nhau:** Dữ liệu có thể được lấy từ cơ sở dữ liệu nội bộ, ứng dụng CRM, hệ thống ERP, hoặc từ các nguồn dữ liệu bên ngoài như mạng xã hội, dữ liệu thị trường.
* **Bước 2: Trích xuất, chuyển đổi và tải dữ liệu (ETL):** Dữ liệu từ các nguồn sẽ được trích xuất, làm sạch và chuyển đổi thành định dạng chuẩn. Sau đó, dữ liệu sẽ được tải vào kho dữ liệu để lưu trữ và truy xuất khi cần thiết.
* **Bước 3: Phân tích và khai phá dữ liệu:** Sau khi dữ liệu được lưu trữ trong kho dữ liệu, các công cụ BI sẽ phân tích dữ liệu này để tìm ra các mẫu, xu hướng và mối quan hệ. Đây là giai đoạn quan trọng giúp biến dữ liệu thành thông tin giá trị.
* **Bước 4: Trực quan hóa và báo cáo:** Các kết quả phân tích sẽ được trình bày dưới dạng các báo cáo, bảng điều khiển hoặc biểu đồ, giúp người dùng có thể nhanh chóng hiểu được thông tin và đưa ra quyết định kinh doanh phù hợp.
* **Bước 5: Ra quyết định:** Dựa trên các báo cáo và phân tích, doanh nghiệp có thể đưa ra các quyết định chiến lược, tối ưu hóa hoạt động kinh doanh, cải thiện sản phẩm hoặc dịch vụ, và nâng cao sự hài lòng của khách hàng.

# CHƯƠNG 2. GIỚI THIỆU VỀ NETFLIX

## 1. Giới thiệu về Netflix

Netflix là một trong những nền tảng phát trực tuyến lớn nhất thế giới, cung cấp dịch vụ giải trí dưới dạng phim ảnh, chương trình truyền hình và nội dung gốc (original content). Được thành lập vào năm 1997 bởi Reed Hastings và Marc Randolph tại California, Mỹ, Netflix ban đầu chỉ là một dịch vụ cho thuê DVD qua thư. Tuy nhiên, từ năm 2007, Netflix đã chuyển hướng sang mô hình phát trực tuyến (streaming) và nhanh chóng trở thành một trong những tên tuổi hàng đầu trong ngành công nghiệp giải trí.

Tính đến hiện tại, Netflix có hàng triệu thuê bao trên toàn cầu, hoạt động tại hơn 190 quốc gia. Nền tảng này cung cấp nội dung đa dạng, từ phim điện ảnh, phim truyền hình, tài liệu cho đến các chương trình thực tế và sản xuất các nội dung gốc như **"Stranger Things," "The Crown"** và **"Money Heist."**

## 2. Mô hình kinh doanh của Netflix

Mô hình kinh doanh của Netflix chủ yếu dựa trên dịch vụ thuê bao (subscription-based model). Người dùng trả phí hàng tháng để truy cập vào thư viện nội dung số của Netflix, không có quảng cáo và có thể xem mọi lúc, mọi nơi, trên mọi thiết bị có kết nối internet. Có ba yếu tố chính trong mô hình kinh doanh của Netflix:

* **Dịch vụ thuê bao (Subscription):** Netflix cung cấp nhiều gói thuê bao khác nhau, phù hợp với nhu cầu và khả năng tài chính của người dùng. Các gói này có thể khác nhau về chất lượng video (SD, HD, 4K) và số lượng thiết bị có thể xem cùng lúc.
* **Nội dung gốc (Original Content):** Netflix đầu tư rất nhiều vào việc sản xuất nội dung gốc. Đây là một yếu tố then chốt giúp Netflix giữ chân khách hàng và cạnh tranh với các đối thủ như Disney+, Hulu, và Amazon Prime. Những bộ phim và chương trình truyền hình gốc của Netflix đã giúp công ty tạo ra sự khác biệt và thu hút đông đảo người xem.
* **Công nghệ và phân tích dữ liệu:** Netflix sử dụng các công cụ phân tích dữ liệu mạnh mẽ để hiểu rõ thói quen xem của người dùng, từ đó đề xuất các nội dung phù hợp cho từng cá nhân (recommendation engine). Điều này không chỉ tăng cường trải nghiệm người dùng mà còn giúp Netflix tối ưu hóa các khoản đầu tư vào sản xuất nội dung.

## 3. Hoạt động nghiệp vụ của Netflix

Các hoạt động nghiệp vụ của Netflix bao gồm một loạt các hoạt động quan trọng để duy trì và phát triển dịch vụ truyền phát nội dung trực tuyến của họ. Dưới đây là những hoạt động chính:

### 3.1. Sản xuất và cấp phép nội dung

* **Sản xuất nội dung gốc (Original Content):** Netflix sản xuất các bộ phim, chương trình truyền hình, và loạt phim gốc để thu hút người dùng và tạo lợi thế cạnh tranh.
* **Mua bản quyền nội dung:** Netflix mua quyền phát sóng từ các hãng phim, kênh truyền hình, và các đơn vị sản xuất khác để cung cấp thêm đa dạng nội dung cho người dùng.

|  |  |
| --- | --- |
| **Nhiệm vụ** | **Mô tả** |
| Tác vụ 1: Nghiên cứu thị trường | Thu thập và phân tích dữ liệu về xu hướng xem phim, chương trình giải trí của người dùng |
| Tác vụ 2: Lên kế hoạch sản xuất nội dung gốc | Xác định các dự án sản xuất phim và chương trình gốc, phát triển ý tưởng và kịch bản. |
| Tác vụ 3: Lựa chọn đối tác sản xuất | Lựa chọn các nhà sản xuất, đạo diễn, diễn viên, và các bên đối tác khác để hợp tác sản xuất. |
| Tác vụ 4: Sản xuất nội dung | Quay phim, biên tập, chỉnh sửa và hoàn thành các bộ phim, chương trình truyền hình. |
| Tác vụ 5: Quản lý bản quyền nội dung gốc | Đảm bảo quyền sở hữu trí tuệ đối với các sản phẩm gốc của Netflix được bảo vệ và quản lý hợp lý. |
| Tác vụ 6: Mua bản quyền nội dung từ bên thứ ba | Đàm phán với các hãng phim, đài truyền hình để mua quyền phát sóng nội dung của họ. |
| Tác vụ 7: Xác định nội dung phù hợp | Đánh giá chất lượng, xu hướng và tính phù hợp của nội dung bên thứ ba với nền tảng Netflix. |
| Tác vụ 8: Thương thảo hợp đồng | Đàm phán các điều khoản hợp đồng, phí cấp phép và điều kiện phát sóng với các đối tác nội dung. |
| Tác vụ 9: Địa phương hóa nội dung | Phiên dịch và địa phương hóa nội dung (phụ đề, lồng tiếng) để phù hợp với từng thị trường. |
| Tác vụ 10: Đảm bảo chất lượng nội dung | Kiểm tra và đánh giá chất lượng kỹ thuật của nội dung trước khi phát sóng (âm thanh, hình ảnh). |
| Tác vụ 11: Phân phối nội dung lên hệ thống | Tải và phân phối nội dung đã hoàn thiện lên hệ thống Netflix, chuẩn bị cho người dùng truy cập. |

Bảng 1 Nghiệp vụ sản xuất và cấp phép nội dung

**3.2. Phân phối và truyền phát nội dung**

* **Phân phối toàn cầu:** Netflix hoạt động trên nhiều quốc gia, cung cấp nội dung đa dạng, hỗ trợ nhiều ngôn ngữ, và cung cấp khả năng truy cập trên nhiều thiết bị như điện thoại, máy tính bảng, máy tính và TV.
* **Hạ tầng công nghệ:** Hệ thống máy chủ và mạng lưới phân phối nội dung (CDN) giúp truyền phát video mượt mà và chất lượng cao đến người dùng khắp thế giới.

|  |  |
| --- | --- |
| **Nhiệm vụ** | **Mô tả** |
| Tác vụ 1: Chuẩn bị nội dung | Netflix mua hoặc sản xuất nội dung (phim, chương trình, tài liệu). Các nội dung này phải được định dạng theo chuẩn để sẵn sàng phân phối, bao gồm xử lý chất lượng âm thanh và hình ảnh. |
| Tác vụ 2: Mã hóa nội dung | Nội dung được mã hóa thành các định dạng khác nhau (SD, HD, 4K) để phù hợp với nhiều thiết bị và kết nối internet khác nhau. Quá trình này còn bao gồm nén và tối ưu hóa kích thước tệp. |
| Tác vụ 3: Lưu trữ nội dung (Data Storage) | Sau khi mã hóa, nội dung được lưu trữ trong các kho dữ liệu phân tán của Netflix trên toàn thế giới. Hệ thống này sử dụng cơ sở hạ tầng điện toán đám mây và trung tâm dữ liệu. |
| Tác vụ 4: Triển khai qua CDN (Content Delivery Network) | Netflix sử dụng mạng phân phối nội dung (CDN) để đưa nội dung đến gần với người dùng hơn. Các bản sao nội dung được lưu trên các máy chủ tại nhiều địa điểm trên toàn cầu để giảm thời gian tải. |
| Tác vụ 5: Quản lý yêu cầu người dùng | Khi người dùng yêu cầu phát một nội dung, hệ thống Netflix sẽ kiểm tra yêu cầu và xác định máy chủ CDN gần nhất có sẵn bản sao của nội dung đó. |
| Tác vụ 6: Truyền phát nội dung | Sau khi xác định máy chủ CDN, nội dung sẽ được truyền phát từ máy chủ đó đến thiết bị của người dùng thông qua giao thức truyền phát (streaming). Dữ liệu được gửi đi theo từng phần nhỏ để giảm độ trễ và tránh mất mát dữ liệu. |
| Tác vụ 7 : Tối ưu hóa phát trực tuyến (Adaptive Streaming) | Trong quá trình truyền phát, Netflix điều chỉnh chất lượng video dựa trên tốc độ kết nối internet của người dùng. Điều này đảm bảo trải nghiệm liền mạch, giảm thiểu gián đoạn do băng thông yếu. |
| Tác vụ 8: Theo dõi và thu thập dữ liệu người dùng | Trong quá trình phát, Netflix thu thập thông tin về thói quen xem phim của người dùng (thời lượng, nội dung xem, tần suất). Dữ liệu này sẽ được sử dụng để cải thiện thuật toán gợi ý nội dung và trải nghiệm người dùng. |
| Tác vụ 9: Kết thúc phiên phát trực tuyến | Khi người dùng kết thúc xem, Netflix sẽ lưu lại lịch sử xem của người dùng và thông tin về điểm dừng của nội dung để tiếp tục cho lần sau. Phiên phát trực tuyến kết thúc, và hệ thống lưu trữ dữ liệu tạm thời của người dùng sẽ được giải phóng. |

Bảng 2 Nghiệp vụ phân phối và truyền phát nội dung

**3.3. Phát triển và tối ưu hóa dịch vụ**

* **Tối ưu hóa thuật toán đề xuất:** Netflix sử dụng công nghệ trí tuệ nhân tạo (AI) và học máy (machine learning) để cung cấp gợi ý cá nhân hóa dựa trên hành vi xem của người dùng.
* **Nghiên cứu và phát triển (R&D):** Đầu tư vào công nghệ để cải tiến chất lượng truyền phát, quản lý dữ liệu người dùng và đảm bảo trải nghiệm tốt hơn.

|  |  |
| --- | --- |
| Nhiệm vụ | Mô tả |
| Tác vụ 1: Thu thập dữ liệu người dùng | Netflix thu thập dữ liệu từ người dùng thông qua quá trình xem phim, lịch sử xem, hành vi trên nền tảng (thời gian dừng, chọn phim, đánh giá, yêu thích, thiết bị sử dụng, v.v.). |
| Tác vụ 2: Lưu trữ và quản lý dữ liệu | Dữ liệu thu thập được lưu trữ trong kho dữ liệu lớn của Netflix. Dữ liệu này bao gồm cả dữ liệu phi cấu trúc (hành vi người dùng) và dữ liệu cấu trúc (thông tin tài khoản, thiết bị). |
| Tác vụ 3: Phân tích dữ liệu người dùng | Netflix sử dụng các công cụ phân tích mạnh mẽ như Big Data, AI và Machine Learning để phân tích dữ liệu từ người dùng. Mục tiêu là tìm ra các xu hướng và mối quan hệ trong hành vi của người xem. |
| Tác vụ 4: Thuật toán đề xuất | Dựa trên kết quả phân tích, Netflix cải thiện các thuật toán gợi ý nội dung cá nhân hóa. Điều này giúp tăng khả năng giữ chân người dùng và cung cấp trải nghiệm phù hợp hơn. |
| Tác vụ 5: Tối ưu hóa giao diện người dùng (UI/UX) | Từ phản hồi và dữ liệu người dùng, Netflix liên tục điều chỉnh giao diện và trải nghiệm người dùng để nâng cao sự tiện lợi, dễ sử dụng và thân thiện hơn trên mọi thiết bị. |
| Tác vụ 6: Tối ưu hóa hiệu suất truyền phát | Đàm phán với các hãng phim, đài truyền hình để mua quyền phát sóng nội dung của họ. Netflix liên tục cải tiến hệ thống phát trực tuyến, sử dụng công nghệ như Adaptive Bitrate Streaming để đảm bảo chất lượng truyền phát ổn định trên mọi băng thông mạng và thiết bị. |
| Tác vụ 7: Kiểm tra và thử nghiệm A/B (A/B Testing) | Trước khi triển khai bất kỳ thay đổi nào, Netflix sử dụng thử nghiệm A/B để so sánh các phiên bản cải tiến với phiên bản hiện tại nhằm đánh giá hiệu quả của các cải tiến đó. |
| Tác vụ 8: Phản hồi từ người dùng | Netflix nhận phản hồi trực tiếp từ người dùng thông qua các đánh giá, khảo sát hoặc sự tương tác với nền tảng để hiểu thêm về trải nghiệm của họ và điều chỉnh dịch vụ cho phù hợp. |
| Tác vụ 9: Triển khai các cải tiến | Sau khi thử nghiệm thành công, Netflix triển khai các cải tiến mới trên toàn bộ nền tảng, bao gồm cập nhật thuật toán gợi ý, cải tiến giao diện, hoặc nâng cấp hạ tầng kỹ thuật. |
| Tác vụ 10: Giám sát và đánh giá hiệu suất | Sau khi triển khai, Netflix tiếp tục theo dõi hiệu suất của các thay đổi, đánh giá ảnh hưởng của chúng đến việc giữ chân người dùng, mức độ hài lòng và hiệu quả sử dụng dịch vụ. |

Bảng 3 Nghiệp vụ phát triển và tối ưu hóa dịch vụ

**3.4. Tiếp thị và phát triển người dùng**

* **Chiến lược quảng cáo và tiếp thị:** Netflix sử dụng các chiến dịch quảng cáo toàn cầu, nội dung truyền thông xã hội và hợp tác với các thương hiệu để thu hút người dùng mới.
* **Dịch vụ khách hàng và hỗ trợ:** Cung cấp hỗ trợ đa ngôn ngữ, giải quyết các vấn đề của người dùng liên quan đến tài khoản, thanh toán, và truyền phát.

|  |  |
| --- | --- |
| Tác vụ 1: Nghiên cứu thị trường và phân tích đối thủ | Netflix tiến hành nghiên cứu thị trường, phân tích nhu cầu và thói quen tiêu dùng ở các khu vực mục tiêu, đồng thời so sánh với các đối thủ như Disney+, Amazon Prime. |
| Tác vụ 2: Phân tích dữ liệu người dùng hiện tại | Dựa trên dữ liệu người dùng hiện tại, Netflix xác định các xu hướng tiêu dùng, sở thích nội dung và mức độ hài lòng, từ đó điều chỉnh chiến lược tiếp thị phù hợp. |
| Tác vụ 3: Xác định chiến lược tiếp thị | Netflix xây dựng chiến lược tiếp thị dựa trên các kết quả phân tích thị trường và dữ liệu nội bộ. Chiến lược này bao gồm quảng bá nội dung gốc, mở rộng thị trường, và tập trung vào các nhóm khách hàng mục tiêu. |
| Tác vụ 4: Phát triển nội dung quảng cáo | Nội dung quảng cáo được phát triển, bao gồm video, hình ảnh, và các thông điệp phù hợp với từng khu vực và nhóm đối tượng. Netflix tạo các chiến dịch quảng cáo dựa trên nội dung đang hot hoặc sắp ra mắt. |
| Tác vụ 5: Phân phối nội dung quảng cáo | Nội dung quảng cáo được triển khai trên các kênh truyền thông như Google, Facebook, YouTube, Twitter, TikTok, và các nền tảng truyền thông địa phương tùy thuộc vào khu vực mục tiêu. |
| Tác vụ 6: Quản lý chiến dịch quảng cáo | Netflix liên tục theo dõi và quản lý các chiến dịch quảng cáo, đánh giá hiệu quả của từng kênh truyền thông và điều chỉnh chiến dịch dựa trên các chỉ số như tỷ lệ tương tác, tỷ lệ chuyển đổi, và chi phí quảng cáo. |
| Tác vụ 7: Triển khai chiến lược khuyến mãi và ưu đãi | Netflix thực hiện các chương trình khuyến mãi như dùng thử miễn phí trong một khoảng thời gian hoặc giảm giá gói thuê bao nhằm thu hút người dùng mới. |
| Tác vụ 8: Theo dõi tương tác và phản hồi của người dùng tiềm năng | Netflix theo dõi phản hồi từ người dùng tiềm năng trên các kênh truyền thông và mạng xã hội, sử dụng các phản hồi này để tối ưu hóa nội dung quảng cáo và trải nghiệm của người dùng trước khi đăng ký. |
| Tác vụ 9: Chuyển đổi người dùng mới | Khi người dùng tiềm năng thể hiện sự quan tâm qua việc truy cập trang web hoặc tải ứng dụng, Netflix sẽ tối ưu hóa quy trình đăng ký, giúp người dùng mới dễ dàng trở thành thuê bao chính thức. |
| Tác vụ 10: Duy trì và tương tác với người dùng mới | Sau khi người dùng mới đăng ký, Netflix tiếp tục gửi thông báo, email cá nhân hóa và giới thiệu nội dung dựa trên sở thích để giữ chân người dùng và tăng tỷ lệ xem lại. |
| Tác vụ 11: Đánh giá và cải tiến chiến lược tiếp thị | Sau khi các chiến dịch kết thúc, Netflix đánh giá toàn diện hiệu quả chiến dịch, thu thập phản hồi và dữ liệu từ người dùng mới, sau đó sử dụng thông tin này để cải tiến chiến lược cho các chiến dịch tiếp theo. |

Bảng 4 Nghiệp vụ tiếp thị và phát triển người dùng

# CHƯƠNG 3. XÁC ĐỊNH NGƯỜI SỬ DỤNG VÀ KHẢO SÁT VIỆC NHU THẬP, TỔ CHỨC, LƯU TRỮ DỮ LIỆU CỦA DOANH NGHIỆP

## 1. Sơ đồ tổ chức của Netflix

Hình Sơ đồ tổ chức

Netflix duy trì một cơ cấu tổ chức nội bộ tương đối phẳng, tập trung vào sự sáng tạo và hiệu quả. Cơ cấu quyền lực phẳng này là nền tảng của văn hóa mở, nơi các bộ phận và nhân viên có thể chia sẻ thông tin và ý tưởng một cách tự do, từ đó thúc đẩy năng suất và đổi mới.

Ở cấp cao nhất, Giám đốc điều hành (CEO) chịu trách nhiệm lãnh đạo tổng thể công ty, làm việc cùng với Hội đồng quản trị, bao gồm các giám đốc độc lập và các kiểm toán viên nội bộ lẫn bên ngoài. Hội đồng quản trị đóng vai trò giám sát các chiến lược và hoạt động quan trọng của công ty. Các chức năng chính như Pháp lý (Legal), Tài chính (Finance), Tài năng (Talent), Nội dung (Content), Truyền thông (Communication) và Sản phẩm (Product) chịu trách nhiệm về các mảng hoạt động quan trọng của Netflix, phối hợp chặt chẽ với CEO để đạt được các mục tiêu chiến lược.

Ngoài ra, cơ cấu tổ chức của Netflix được chia thành các bộ phận hoạt động theo địa lý và chức năng:

* Geographical Divisions: Bao gồm Domestic Streaming và International Streaming, tập trung vào việc tối ưu hóa hoạt động tại các khu vực cụ thể.
* Product/Operations Divisions: Bao gồm Original Programming và Other Content, quản lý việc sản xuất và cung cấp nội dung trên nền tảng.

Các Giám đốc điều hành ở cấp quốc gia chịu trách nhiệm quản lý các hoạt động tại các lãnh thổ cụ thể, hợp tác chặt chẽ với trụ sở chính tại Los Gatos, California, để đảm bảo sự đồng nhất và hiệu quả trong toàn bộ hệ thống.

## 2. Người sử dụng BI (BI User) trong tổ chức Netflix

|  |  |  |  |
| --- | --- | --- | --- |
| **Chức vụ** | **Đặc điểm quyết định** | **Quản lý** | **Mức độ đưa ra quyết định** |
| Content Team | Phân tích dữ liệu nội dung hiện có (lượt xem, thể loại phổ biến, IMDB scores) | Quản lý nội dung; Tối ưu danh mục chương trình. | Chiến lược sản xuất và đầu tư nội dung. |
| Finance Team | Xác định mối quan hệ giữa IMDB scores và giá cổ phiếu; đánh giá ROI của nội dung | Quản lý ngân sách và lợi nhuận. | Chiến lược tài chính. |
| Product Team | Cải thiện giao diện và trải nghiệm người dùng dựa trên xu hướng thể loại và nội dung được yêu thích. | Quản lý trải nghiệm người dùng. | Chiến lược phát triển sản phẩm. |
| Domestic và International Streaming Divisions | Phân tích hành vi khách hàng theo từng khu vực để tối ưu danh mục phù hợp. | Quản lý theo khu vực địa lý. | Tối ưu hóa chiến lược địa phương. |
| Original Programming Team | Xác định sự thành công của nội dung gốc dựa trên IMDB scores và lượt xem; điều chỉnh kế hoạch sản xuất nội dung gốc. | Quản lý nội dung độc quyền. | Định hướng sản xuất chương trình. |
| Talent Management | Dựa vào thành công của các chương trình để tuyển chọn diễn viên, nhà sản xuất phù hợp. | Quản lý nhân sự trong sản xuất nội dung. | Tuyển dụng và tối ưu đội ngũ nhân tài. |
| Marketing/Communication Team | Sử dụng dữ liệu về thể loại và nội dung phổ biến để triển khai chiến dịch truyền thông hiệu quả. | Quản lý chiến lược marketing. | Chiến lược tiếp thị và quảng bá. |

Bảng Người sử dụng BI (BI User) trong tổ chức Netflix

Các BI Users chính trong dự án này là Content Team, Finance Team, và Original Programming Team, vì họ trực tiếp sử dụng dữ liệu để tối ưu nội dung và đánh giá tác động đến giá cổ phiếu.

Marketing Team và Product Team cũng sử dụng kết quả phân tích từ hệ thống BI để tối ưu hóa trải nghiệm người dùng.

## 3. Phân tích các quyết định kinh doanh của BI Users

|  |  |
| --- | --- |
| **Chức vụ** | **Phân tích** |
| Content Team | **Mục tiêu:** Tối ưu danh mục chương trình (thể loại, nội dung phổ biến) để thu hút và giữ chân người dùng.  **Dữ liệu BI** **cần quan tâm:** Lượt xem, thể loại phổ biến, IMDB scores, tỷ lệ hoàn thành chương trình, mức độ tương tác của khán giả. |
| Finance Team | **Mục tiêu**: Tăng hiệu quả chi phí đầu tư nội dung và tối ưu hóa lợi nhuận.  **Dữ liệu BI** **cần quan tâm:** ROI của các chương trình (chi phí sản xuất so với doanh thu tạo ra), mối quan hệ giữa IMDB scores và giá cổ phiếu, doanh thu theo nội dung/địa phương. |
| Product Team | **Mục tiêu:** Cải thiện trải nghiệm người dùng trên nền tảng (giao diện, cá nhân hóa nội dung).  **Dữ liệu BI** **cần quan tâm:** Thời gian xem trung bình, xu hướng nội dung được tìm kiếm, mức độ sử dụng các tính năng giao diện. |
| Original Programming Team | **Mục tiêu**: Định hướng sản xuất nội dung gốc có chất lượng cao và hấp dẫn khán giả.  **Dữ liệu BI cần quan tâm**: Hiệu suất của nội dung gốc (lượt xem, IMDB scores), so sánh với nội dung không gốc, phản hồi từ người dùng. |
| Domestic và International Streaming Divisions | **Mục tiêu**: Tối ưu hóa danh mục nội dung phù hợp theo từng khu vực để tăng mức độ hài lòng và lượng thuê bao ở mỗi địa phương.  **Dữ liệu BI cần quan tâm**: Xu hướng thể loại theo khu vực, hành vi người dùng (giờ xem, nội dung yêu thích), tỷ lệ rời bỏ của người dùng theo thị trường. |
| Marketing/Communication Team | **Mục tiêu**: Triển khai chiến dịch marketing hiệu quả để tăng nhận diện thương hiệu và số lượng thuê bao mới.  **Dữ liệu BI cần quan tâm**: Lượt xem theo chương trình được quảng bá, hiệu quả của các chiến dịch tiếp thị, dữ liệu nhân khẩu học của người dùng (độ tuổi, sở thích). |
| Talent Management | **Mục tiêu**: Tuyển dụng và giữ chân nhân tài phù hợp dựa trên thành công của các chương trình.  **Dữ liệu BI cần quan tâm**: Đánh giá hiệu suất của nhân sự liên quan đến nội dung (diễn viên, nhà sản xuất), phản hồi từ khán giả về chất lượng chương trình |

Bảng 6 Phân tích các quyết định kinh doanh của BI Users

# CHƯƠNG 4. XÂY DỰNG KHO DỮ LIỆU

## 1. Data Warehouse là gì

“Kho dữ liệu là một tập hợp dữ liệu hướng chủ đề, tích hợp không thay đổi, có tính thời gian và để hỗ trợ quá trình ra quyết định quản lý” – Theo W.H. Inmon

Data Warehouse (kho dữ liệu) là hệ quản trị cơ sở dữ liệu, chuyên dùng cho tạo báo cáo và phân tích dữ liệu. Hỗ trợ các truy vấn phức tạp, vừa là điểm tập trung dữ liệu từ nhiều nguồn khác nhau để có được thông tin phân tích đầy đủ nhất. Data Warehouse là tập hợp dữ liệu tương đối ổn định, không thay đổi, cập nhật theo thời gian, được tích hợp theo hướng chủ đề nhằm hỗ trợ quá trình tạo quyết định về mặt quản lý.

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 BI, SQL client hay bảng tính.

## 2. ETL trong xây dựng Data Warehouse

ETL (Extract, Transform, Load) là một quy trình cốt lõi trong việc xây dựng và duy trì một Data Warehouse, giúp tổ chức tích hợp và quản lý dữ liệu từ nhiều nguồn khác nhau. Quy trình này bao gồm ba bước chính:

* **Extract (Trích xuất)**: Bước đầu tiên trong quy trình ETL là trích xuất dữ liệu từ nhiều nguồn khác nhau. Những nguồn này có thể bao gồm các hệ thống giao dịch, cơ sở dữ liệu quan hệ, tệp tin Excel, API từ các ứng dụng khác nhau hoặc thậm chí là dữ liệu không cấu trúc. Việc trích xuất dữ liệu yêu cầu các công cụ và kỹ thuật để đảm bảo rằng dữ liệu được thu thập một cách chính xác và hiệu quả, đồng thời giảm thiểu ảnh hưởng đến hiệu suất của các nguồn dữ liệu.
* **Transform (Chuyển đổi)**: Sau khi dữ liệu được trích xuất, bước tiếp theo là chuyển đổi dữ liệu để phù hợp với cấu trúc và định dạng của Data Warehouse. Bước này có thể bao gồm nhiều quy trình như làm sạch dữ liệu (loại bỏ các lỗi và dữ liệu không chính xác), chuẩn hóa định dạng (đảm bảo tính đồng nhất của các trường dữ liệu), và tổng hợp dữ liệu (tính toán các chỉ số, như tổng, trung bình, v.v.). Quá trình chuyển đổi không chỉ giúp dữ liệu trở nên có tổ chức hơn mà còn tạo ra giá trị gia tăng cho dữ liệu, giúp người dùng có thể dễ dàng truy cập và phân tích.
* **Load (Tải)**: Cuối cùng, dữ liệu đã được chuyển đổi sẽ được tải vào kho dữ liệu. Quy trình tải có thể diễn ra theo nhiều cách, bao gồm tải toàn bộ (full load) hoặc tải gia tăng (incremental load), tùy thuộc vào nhu cầu và mục tiêu của tổ chức. Tải dữ liệu vào Data Warehouse cho phép tổ chức lưu trữ và truy xuất thông tin một cách hiệu quả, cung cấp một nền tảng cho các hoạt động phân tích và báo cáo.

Quy trình ETL không chỉ đơn thuần là một tập hợp các bước kỹ thuật, mà còn là một phần quan trọng trong chiến lược quản lý dữ liệu của bất kỳ tổ chức nào. Bằng cách đảm bảo rằng dữ liệu được trích xuất, chuyển đổi và tải một cách chính xác, tổ chức có thể nâng cao khả năng ra quyết định, tối ưu hóa quy trình kinh doanh, và tạo ra giá trị từ dữ liệu mà họ sở hữu.

## 3. Các công cụ hỗ trợ ETL

Có nhiều công cụ hỗ trợ cho quy trình ETL, từ các giải pháp thương mại đến mã nguồn mở, giúp thực hiện các bước trích xuất, chuyển đổi, và tải dữ liệu một cách tự động và hiệu quả hơn. Dưới đây là một số công cụ ETL phổ biến và đánh giá về ưu, nhược điểm của chúng:

**Apache Nifi:**

* Ưu điểm: Đây là một công cụ mã nguồn mở, có giao diện người dùng đồ họa dễ sử dụng, hỗ trợ kéo thả để xây dựng các quy trình ETL. Nifi cho phép quản lý và di chuyển dữ liệu giữa các hệ thống với khả năng theo dõi luồng dữ liệu theo thời gian thực.
* Nhược điểm: Mặc dù dễ sử dụng, Nifi có thể không phù hợp cho các tác vụ chuyển đổi dữ liệu phức tạp. Khả năng mở rộng của Nifi cũng có giới hạn trong các hệ thống lớn với lưu lượng dữ liệu cao.

**Microsoft SQL Server Integration Services (SSIS):**

* Ưu điểm: SSIS là một công cụ ETL mạnh mẽ và được tích hợp sẵn trong SQL Server. Nó cung cấp nhiều tính năng tiên tiến để làm sạch, chuyển đổi và tải dữ liệu vào Data Warehouse. Giao diện đồ họa của SSIS giúp người dùng dễ dàng thiết kế các quy trình ETL mà không cần viết nhiều mã.
* Nhược điểm: SSIS là một công cụ thương mại, nên chi phí có thể là rào cản đối với các tổ chức nhỏ. Ngoài ra, nó chỉ hỗ trợ tốt nhất trong hệ sinh thái Microsoft, khó tích hợp với các nền tảng khác.

## 4. Phương pháp mà nhóm áp dụng: Python thuần

Mặc dù có rất nhiều công cụ hỗ trợ ETL mạnh mẽ, nhóm chúng em đã quyết định sử dụng Python thuần để thực hiện các thao tác ETL. Lý do chính cho sự lựa chọn này là Python cung cấp sự linh hoạt và kiểm soát cao trong toàn bộ quy trình ETL.

* Tính linh hoạt: Python là một ngôn ngữ lập trình mạnh mẽ với nhiều thư viện hỗ trợ như `pandas` và `pyodbc`, giúp dễ dàng trích xuất dữ liệu từ nhiều nguồn khác nhau, thực hiện các phép chuyển đổi dữ liệu phức tạp, và tải chúng vào cơ sở dữ liệu theo cách tùy chỉnh. Điều này giúp nhóm có thể điều chỉnh từng bước của quy trình ETL để phù hợp với yêu cầu đặc thù của dự án.
* Kiểm soát chi tiết: Sử dụng Python thuần cho phép chúng em kiểm soát chặt chẽ mọi khía cạnh của quá trình xử lý dữ liệu, từ việc làm sạch đến chuyển đổi dữ liệu phức tạp. Điều này mang lại tính chính xác cao và khả năng tinh chỉnh khi cần thiết, đặc biệt khi làm việc với những tập dữ liệu đa dạng và không đồng nhất.
* Tính mở rộng: Python cũng cho phép dễ dàng mở rộng các quy trình ETL thông qua việc tích hợp với các thư viện mã nguồn mở khác và công cụ phân tích dữ liệu. Điều này giúp nhóm có thể dễ dàng bổ sung các tính năng mới mà không phải phụ thuộc vào các giới hạn của công cụ ETL thương mại.

## 5. Phân tích dữ liệu

**Dataset My CSV File**

Dataset "My CSV File" cung cấp thông tin chi tiết về danh mục phim và chương trình truyền hình trên Netflix, bao gồm các đặc điểm như đạo diễn, dàn diễn viên, thể loại nội dung, v.v. Dataset này thích hợp cho việc nghiên cứu khả năng thu hút khán giả và tính đa dạng trong nội dung.

|  |  |
| --- | --- |
| **Tên Trường** | **Ý Nghĩa** |
| **show\_id** | Mã duy nhất để xác định phim hoặc chương trình. |
| **type** | Loại nội dung: "Movie" (Phim) hoặc "TV Show" (Chương trình truyền hình). |
| **title** | Tên của phim hoặc chương trình truyền hình. |
| **director** | Đạo diễn phim hoặc chương trình truyền hình. |
| **cast** | Diễn viên chính tham gia phim hoặc chương trình. |
| **country** | Quốc gia sản xuất phim hoặc chương trình truyền hình. |
| **date\_added** | Ngày nội dung được thêm vào Netflix. |
| **release\_year** | Năm phát hành nội dung. |
| **rating** | Đánh giá phân loại độ tuổi cho nội dung. |
| **duration** | Thời lượng phim hoặc số mùa của chương trình truyền hình. |
| **listed\_in** | Thể loại nội dung (vd: Hành động, Hài kịch, Lãng mạn). |
| **description** | Tóm tắt nội dung chính của phim hoặc chương trình. |

**Dataset Netflix Originals**

Dataset "Netflix Originals" bao gồm thông tin chi tiết về các chương trình gốc do Netflix sản xuất, giúp cung cấp bức tranh toàn diện về danh mục nội dung độc quyền. Dataset này hữu ích cho việc nghiên cứu xu hướng sản xuất nội dung và mức độ ảnh hưởng của Netflix trên thị trường truyền hình.

|  |  |
| --- | --- |
| **Tên Trường** | **Ý Nghĩa** |
| **Title** | Tên của chương trình gốc của Netflix. |
| **Genre** | Thể loại của chương trình. |
| **Original Network** | Nền tảng/kênh gốc phát hành chương trình. |
| **Premiere** | Ngày ra mắt chương trình. |
| **Seasons** | Thông tin về số mùa và tập phim. |
| **Length** | Thời lượng trung bình của mỗi tập phim (phút). |
| **Netflix Exclusive Regions** | Khu vực Netflix phân phối độc quyền. |
| **Status** | Tình trạng hiện tại của chương trình (vd: Đã kết thúc, Đang chiếu). |

**Dataset NFLX**

Dataset "NFLX" cung cấp dữ liệu lịch sử về cổ phiếu Netflix, bao gồm mức giá giao dịch theo thời gian. Dataset này là nguồn quan trọng để phân tích xu hướng tài chính và dự báo mức giá trong tương lai. Các trường thông tin gồm các yếu tố sau:

|  |  |
| --- | --- |
| **Tên Trường** | **Ý Nghĩa** |
| **Date** | Ngày giao dịch của cổ phiếu. |
| **Open** | Mức giá mở cửa của cổ phiếu Netflix trong ngày. |
| **High** | Mức giá cao nhất của cổ phiếu Netflix trong ngày. |
| **Low** | Mức giá thấp nhất của cổ phiếu Netflix trong ngày. |
| **Close** | Mức giá đóng cửa của cổ phiếu Netflix trong ngày. |
| **Adj Close** | Mức giá đóng cửa đã được điều chỉnh (sau chia tách). |
| **Volume** | Số lượng cổ phiếu đã giao dịch trong ngày. |

## 6. Tạo các bảng tương ứng trong CSDL

import pymysql

db = pymysql.connect(

    host="localhost",

    user="root",

    password="",

    database="netflix\_datawarehouse"

)

cursor = db.cursor()

cursor.execute("DROP TABLE IF EXISTS show\_dim;")

createTable1 = """CREATE TABLE show\_dim(

                 show\_id INT not null,

                 type VARCHAR(255) null,

                 title VARCHAR(255) not null,

                 director VARCHAR(255) null,

                 cast MEDIUMTEXT null,

                 country VARCHAR(255) null,

                 duration VARCHAR(255) null,

                 description MEDIUMTEXT null,

                 PRIMARY KEY (show\_id, title));"""

cursor.execute(createTable1)

cursor.execute("DROP TABLE IF EXISTS original\_dim;")

createTable2 = """CREATE TABLE original\_dim(

                  original\_title VARCHAR(255) not null,

                  seasons VARCHAR(255) null,

                  duration VARCHAR (255) null,

                  regions VARCHAR (255) null,

                  status VARCHAR (255) null,

                  PRIMARY KEY (original\_title));"""

cursor.execute(createTable2)

cursor.execute("DROP TABLE IF EXISTS date\_dim;")

createTable3 = """CREATE TABLE date\_dim(

                  date\_id VARCHAR(50) not null,

                  date DATE null,

                  year INT null,

                  PRIMARY KEY (date\_id));"""

cursor.execute(createTable3)

cursor.execute("DROP TABLE IF EXISTS facts\_imdb\_rating;")

createTable5 = """

CREATE TABLE facts\_imdb\_rating (

    show\_id INT NOT NULL,

    title VARCHAR(255) NOT NULL,

    original\_title VARCHAR(255),

    date\_id VARCHAR(50),

    rating FLOAT,

    PRIMARY KEY (show\_id),

    FOREIGN KEY (show\_id) REFERENCES show\_dim(show\_id),

    FOREIGN KEY (original\_title) REFERENCES original\_dim(original\_title),

    FOREIGN KEY (date\_id) REFERENCES date\_dim(date\_id)

);

"""

cursor.execute(createTable5)

cursor.execute("DROP TABLE IF EXISTS genre\_dim;")

create\_genre\_dim = """

CREATE TABLE genre\_dim (

    genre\_id INT AUTO\_INCREMENT,

    genre\_name VARCHAR(255) NOT NULL,

    PRIMARY KEY (genre\_id)

);

"""

cursor.execute(create\_genre\_dim)

cursor.execute("DROP TABLE IF EXISTS show\_genre;")

create\_movie\_genre = """

CREATE TABLE show\_genre (

    show\_id INT NOT NULL,

    genre\_id INT NOT NULL,

    PRIMARY KEY (show\_id, genre\_id),

    FOREIGN KEY (show\_id) REFERENCES facts\_imdb\_rating(show\_id),

    FOREIGN KEY (genre\_id) REFERENCES genre\_dim(genre\_id)

);

"""

cursor.execute(create\_movie\_genre)

cursor.execute("DROP TABLE IF EXISTS facts\_stock\_prices;")

createTable6 = """CREATE TABLE facts\_stock\_prices(

                  date\_id VARCHAR(50) not null,

                  open FLOAT null,

                  high FLOAT null,

                  low FLOAT null,

                  close FLOAT null,

                  adj\_close FLOAT null,

                  volume FLOAT null,

                  PRIMARY KEY (date\_id),

                  FOREIGN KEY (date\_id) REFERENCES date\_dim(date\_id)

                  );"""

cursor.execute(createTable6)

**Lược đồ cho kho dữ liệu:**

## 7. Thực hiện các thao tác ETL

### 7.1. Extract & Transform

parse\_dates = ['date\_added']

netflix\_titles = pd.read\_csv('mycsvfile.csv', parse\_dates=parse\_dates)

netflix\_titles['date\_added'] = pd.to\_datetime(netflix\_titles['date\_added'], errors='coerce')

netflix\_titles['date\_added'] = netflix\_titles['date\_added'].dt.strftime("%Y-%m-%d")

fullList = netflix\_titles['title'].values.tolist()

print(netflix\_titles['date\_added'].head())

netflix\_titles['genres'] = netflix\_titles['listed\_in'].str.split(', ')

# Tạo danh sách các thể loại duy nhất

unique\_genres = set(genre for genres in netflix\_titles['genres'] for genre in genres)

genre\_dict = {genre: idx+1 for idx, genre in enumerate(unique\_genres)}

show\_genres = []

for \_, row in netflix\_titles.iterrows():

    movie\_id = row['show\_id']

    for genre in row['genres']:

        show\_genres.append((movie\_id, genre\_dict[genre]))

showDim = netflix\_titles.drop(['date\_added', 'release\_year', 'rating','listed\_in'], axis=1)

showDim = showDim.where((pd.notnull(showDim)), None)

showDimList = []

for row in showDim.values.tolist():

    showDimList.append(tuple(row))

netflix\_origionals = pd.read\_csv('netflix\_originals.csv')

netflix\_origionals = netflix\_origionals[['Title','Seasons','Length','Netflix Exclusive Regions','Status'

netflix\_origionals = netflix\_origionals.where((pd.notnull(netflix\_origionals)), None)

netflix\_origionals.drop\_duplicates(subset ="Title",

                     keep = 'first', inplace = True)

netflix\_origionalsList = [tuple(l) for l in netflix\_origionals.values.tolist()]

netflix\_stocks = pd.read\_csv('NFLX.csv', parse\_dates=['Date'])

dateDim = pd.DataFrame({'date': pd.date\_range(start='2002-05-23', end='2020-08-03')}) # create a dataframe that has dates ranging from 2002-05-23 to 2020-08-03

dateDim['date\_id'] = dateDim.index + 1 # create date\_id column and assign id numbers starting from 1

dateDim['date'] = dateDim['date'].dt.strftime("%Y-%m-%d") # format date column to YYYY-MM-DD

dateDim['year'] = pd.DatetimeIndex(dateDim['date']).year # using the year information from date colunn to create year column

dateDim = dateDim.reindex(columns=['date\_id','date','year']) # re-arrange the order of columns

#dateDim.date = pd.to\_datetime(dateDim.date)

dateDim.date\_id = dateDim.date\_id.astype(str) # convert data in date column to string

dateDimList = [tuple(l) for l in dateDim.values.tolist()] # format data into a list of tuples before inserting to database

# dateDimList: data for date\_dim table

dateDim.date = pd.to\_datetime(dateDim.date) # convert data in date column to datatype date

factsStockDF = pd.merge(netflix\_stocks, dateDim, left\_on='Date', right\_on='date', how='inner') # inner join dataframes netflix\_stocks and dataDim on date

factsStockDF = factsStockDF.drop(['Date', 'date', 'year'], axis=1) # delete columns that are not needed

factsStockDF = factsStockDF.reindex(columns=['date\_id','Open','High','Low','Close','Adj Close', 'Volume']) # re-arrange the order of columns

# factsStockDF['Date'] = factsStockDF['Date'].dt.strftime("%Y-%m-%d")

factsStockList = [tuple(l) for l in factsStockDF.values.tolist()] # format data into a list of tuples before inserting to database

# factsStockList: data for facts\_stock\_prices table

factsRating = pd.merge(netflix\_titles, netflix\_origionals, left\_on='title', right\_on='Title', how='left') # left join dataframes netflix\_titles and netflix\_originals on titles

factsRating.date\_added = pd.to\_datetime(factsRating.date\_added) # convert data in date\_added column to datatype date

factsRatingDF = pd.merge(factsRating, dateDim, left\_on='date\_added', right\_on='date', how='left') # left join dataframes factsRating and dateDim on date\_added/date

factsRatingDF = factsRatingDF[['show\_id','title','Title','date\_id','rating']] # select only columns needed

# factsRatingDF['date\_id'] = factsRatingDF['date\_id'].astype(str)

factsRatingDF = factsRatingDF.where((pd.notnull(factsRatingDF)), None) # convert empty values to "None" values

factsRatingList = [tuple(l) for l in factsRatingDF.values.tolist()] # format data into a list of tuples before inserting to database

#print(factsRatingList) # data for facts\_IMDB\_rating table

### 7.2. Load

insertData1 = "INSERT INTO show\_dim VALUES (%s,%s,%s,%s,%s,%s,%s,%s);" # insert data into table

cursor.executemany(insertData1, showDimList)

db.commit()

db.close() # close the connection to mysql

insertData2 = "INSERT INTO original\_dim VALUES (%s,%s,%s,%s,%s);" # insert data into table

cursor.executemany(insertData2, netflix\_origionalsList)

db.commit()

db.close() # close the connection to mysql

insertData3 = "INSERT INTO date\_dim VALUES (%s,%s,%s);" # insert data into table

cursor.executemany(insertData3, dateDimList)

db.commit()

db.close() # close the connection to mysql

insertData5 = "INSERT INTO facts\_imdb\_rating VALUES (%s,%s,%s,%s,%s);" # insert data into table

cursor.executemany(insertData5, factsRatingList)

db.commit()

db.close() # close the connection to mysql

insert\_genre\_dim = "INSERT INTO genre\_dim (genre\_name) VALUES (%s);"

cursor.executemany(insert\_genre\_dim, [(genre,) for genre in unique\_genres])

db.commit()

insert\_movie\_genre = "INSERT INTO show\_genre (show\_id, genre\_id) VALUES (%s, %s);"

cursor.executemany(insert\_movie\_genre, show\_genres)

db.commit()

# CHƯƠNG 5. PHÂN TÍCH MÔ HÌNH NHU QUAN HỆ GIỮA CÁC NHU CẦU PHÂN TÍCH CỦA BI USERS VÀ CÁC DATAVIEW

## ****1. BI Users và Nhu cầu phân tích****

BI users trong tổ chức (như Content Team, Finance Team, Marketing Team, v.v.) sử dụng dữ liệu để hỗ trợ các quyết định chiến lược. Dựa trên tài liệu, các nhu cầu phân tích bao gồm:

* **Content Team:** Tối ưu danh mục nội dung bằng cách phân tích lượt xem, thể loại phổ biến, và xếp hạng IMDb.
* **Finance Team:** Phân tích ROI nội dung và tác động của xếp hạng IMDb đến giá cổ phiếu.
* **Marketing Team:** Tối ưu chiến lược tiếp thị dựa trên dữ liệu nhân khẩu học và hiệu quả chiến dịch quảng bá.
* **Product Team:** Cải thiện trải nghiệm người dùng dựa trên thói quen và sở thích nội dung.

**2. Dataview và Hỗ trợ phân tích**

Dataview là các cấu trúc dữ liệu được thiết kế trong kho dữ liệu (data warehouse) để hỗ trợ các yêu cầu phân tích của BI users. Trong trường hợp này:

* **Fact Tables:** Lưu trữ dữ liệu chính liên quan đến các chỉ số phân tích, như:
  + **facts\_imdb\_rating:** Xếp hạng IMDb của nội dung.
  + **facts\_stock\_prices:** Giá cổ phiếu theo thời gian.
* **Dimension Tables:** Lưu trữ thông tin chi tiết để phân tích dữ liệu từ nhiều khía cạnh:
  + **show\_dim:** Chi tiết nội dung (tựa đề, đạo diễn, thể loại, quốc gia).
  + **date\_dim:** Thông tin thời gian (ngày, năm).
  + **genre\_dim:** Thể loại phim.

Các bảng này hỗ trợ việc phân tích đa chiều (OLAP), như:

* Phân tích xu hướng xếp hạng IMDb theo thời gian.
* So sánh ROI nội dung giữa các thể loại hoặc khu vực.
* Tối ưu hóa nội dung theo sở thích khách hàng ở các quốc gia khác nhau.

**3. Mối quan hệ giữa nhu cầu phân tích và dataview**

|  |  |  |
| --- | --- | --- |
| **Nhu cầu Phân Tích** | **Dataview Tương Ứng** | **Ví dụ** |
| **Phân tích xu hướng xếp hạng IMDb** | facts\_imdb\_rating, date\_dim | Xác định xu hướng xếp hạng IMDb của phim theo năm. |
| **Phân tích hiệu quả tài chính** | facts\_stock\_prices, show\_dim | Tìm mối tương quan giữa ROI và xếp hạng nội dung. |
| **Tối ưu danh mục nội dung** | show\_dim, genre\_dim, facts\_imdb\_rating | Xác định thể loại phim phổ biến nhất ở từng khu vực. |
| **Cải thiện trải nghiệm người dùng** | show\_dim, genre\_dim, facts\_imdb\_rating | Tối ưu hóa giao diện người dùng dựa trên thể loại yêu thích. |

**4. Áp dụng thực tế mối quan hệ giữa nhu cầu phân tích và Dataview**

Netflix tổ chức dữ liệu trong kho dữ liệu (Data Warehouse) theo mô hình **Kim Tự Tháp (Star Schema)**, bao gồm:

* **Bảng Fact:** Chứa các chỉ số phân tích chính như xếp hạng IMDb (facts\_imdb\_rating) và giá cổ phiếu (facts\_stock\_prices).
* **Bảng Dimension:** Cung cấp thông tin chi tiết (title, thể loại, thời gian) để phân tích đa chiều.

Netflix áp dụng quy trình **ETL (Extract, Transform, Load):**

1. **Trích xuất** dữ liệu từ các nguồn như IMDb và hệ thống nội bộ.
2. **Chuyển đổi:** Làm sạch, chuẩn hóa, và tính toán các chỉ số.
3. **Tải:** Nạp dữ liệu đã xử lý vào các bảng fact và dimension.

Dữ liệu từ dataview được sử dụng để:

* **Content Team:** Tối ưu danh mục nội dung và phân tích thể loại phổ biến.
* **Finance Team:** Đánh giá ROI và mối liên hệ giữa hiệu suất nội dung với giá cổ phiếu.
* **Marketing Team:** Xây dựng chiến dịch tiếp thị hiệu quả dựa trên sở thích người dùng.

# CHƯƠNG 5. KẾT LUẬN

Trong quá trình xây dựng Data Warehouse cho dữ liệu phim của Netflix, nhóm chúng em đã tiến hành các bước cần thiết để thu thập, làm sạch, biến đổi và tải dữ liệu (ETL) từ nguồn dữ liệu gốc vào hệ thống cơ sở dữ liệu phân tích. Cụ thể, nhóm đã hoàn thành các công việc sau:

**Thiết kế hệ thống Data Warehouse**:

Nhóm đã phân tích dữ liệu ban đầu và xây dựng mô hình cơ sở dữ liệu theo dạng Kim Tự Tháp (Star Schema), với bảng fact chính là **fact\_movie** chứa thông tin phim và các bảng dim xung quanh như **dim\_director**, **dim\_genre**, **dim\_cast**, **dim\_country**.

**Thực hiện quy trình ETL**:

* + **Extract (Trích xuất dữ liệu)**: Dữ liệu từ Netflix đã được tải vào hệ thống dưới dạng dataframe để chuẩn bị cho quá trình chuyển đổi.
  + **Transform (Biến đổi dữ liệu)**: Nhóm đã làm sạch dữ liệu, loại bỏ các giá trị trùng lặp và chuyển đổi các cột thành các định dạng phù hợp để nhập vào cơ sở dữ liệu. Chúng đã xử lý các cột như listed\_in, cast, country bằng cách chia tách và chuẩn hóa dữ liệu.
  + **Load (Nạp dữ liệu)**: Dữ liệu đã được nạp vào các bảng trong Data Warehouse, với việc quản lý quan hệ nhiều-nhiều giữa phim và các thông tin liên quan (đạo diễn, thể loại, diễn viên, quốc gia) thông qua các bảng liên kết.

**Xác định các báo cáo:**

Dữ liệu hiện tại cho phép chúng em thực hiện các báo cáo phân tích liên quan đến:

* **Thống kê phim theo thể loại**: Xác định thể loại nào phổ biến nhất trên Netflix.
* **Phân tích phim theo quốc gia sản xuất**: Đánh giá tần suất phim đến từ các quốc gia khác nhau.
* **Thống kê diễn viên và đạo diễn có nhiều phim nhất**: Phân tích xem những cá nhân nào đóng góp nhiều nhất vào nội dung của Netflix.
* **Báo cáo theo năm phát hành và độ dài phim**: Phân tích xu hướng sản xuất phim qua các năm và sự khác biệt về độ dài giữa các thể loại phim.

**Ý tưởng phát triển trong tương lai**:

Nhóm nhận thấy rằng những báo cáo có thể thực hiện được còn khá hạn chế, tuy nhiên, nhóm đã tìm được hướng đi phù hợp cho đề tài/bài toán này.

Để hoàn thiện hơn nữa ý tưởng của nhóm, ở bài tập cuối kì, nhóm dự định sẽ thu thập nhiều bộ dữ liệu hơn, mở rộng phạm vi báo cáo liên quan đến người dùng như tài khoản người dùng, trả phí subcription, v.v.

# TÀI LIỆU THAM KHẢO

Marketing, P. (không ngày tháng). Được truy lục từ https://pamarketing.vn/mo-hinh-kinh-doanh-cua-netflix-p1/