

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN



ĐỒ ÁN CUỐI KỲ KHO DỮ LIỆU VÀ OLAP
ĐỀ TÀI

PHÂN TÍCH DỮ LIỆU PHIM ĐIỆN ẢNH
TỪ NGUỒN IMDB VÀ TMDB

GV: Nguyễn Thị Kim Phụng

Mã lớp: IS217.P12

Nguyễn Hồng Phát 22521072

Hồ Chí Minh, tháng 12 2024

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

....., ngày tháng năm 2024

Người nhận xét
(Ký tên và ghi rõ họ tên)

LỜI CẢM ƠN

Kính gửi giảng viên Nguyễn Thị Kim Phụng,

Em xin gửi lời cảm ơn sâu sắc đến quý thầy cô đã hỗ trợ em trong quá trình thực hiện đồ án. Với tinh thần tận tâm và trách nhiệm cao, quý thầy cô đã giúp đỡ em vượt qua những khó khăn, thách thức và hoàn thành tốt công việc nghiên cứu đồ án.

Đầu tiên, em xin gửi lời cảm ơn đến quý thầy cô đã truyền đạt kiến thức chuyên môn một cách rõ ràng, cặn kẽ và tận tình. Những giải đáp thắc mắc, những lời khuyên hữu ích và những phản hồi chân thành của quý thầy cô đã giúp đỡ em nắm bắt được kiến thức một cách nhanh chóng và hiệu quả hơn.

Ngoài ra, em cũng muốn bày tỏ lòng biết ơn đến quý thầy cô về sự hỗ trợ quan trọng trong việc chỉnh sửa, đánh giá và phê duyệt bản đồ án của em. Những lời nhận xét, sửa chữa và đề xuất cải thiện của quý thầy cô đã giúp em hiểu được những sai sót và thực hiện đồ án của mình tốt hơn.

Em tin rằng, những kiến thức và kinh nghiệm mà quý thầy cô đã truyền đạt sẽ giúp em phát triển nghề nghiệp và trở thành những người có ích cho xã hội. Một lần nữa, em xin chân thành cảm ơn quý thầy cô vì tất cả những điều tốt đẹp mà quý thầy cô đã mang đến cho em trong suốt thời gian qua.

Kính chúc quý thầy cô sức khỏe, thành công và hạnh phúc!

Trân trọng,

Nguyễn Hồng Phát

MỤC LỤC

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI.....	6
1.1 Lý do chọn đề tài	6
1.2 Giới thiệu về Dataset	7
1.2.1 Thông tin về Dataset	7
1.2.2 Tác giả	7
1.2.3 Số dòng, số cột và thời gian thu thập	7
1.2.4 Nguồn Dataset	7
1.2.5 Tiền xử lý và làm sạch dữ liệu.....	7
1.3 Mô tả chi tiết các cột thuộc tính của kho dữ liệu.....	10
1.4 Thiết kế kho dữ liệu	12
1.4.1 Thiết kế lược đồ hình sao.....	12
1.4.2 Mô tả chi tiết về bảng Fact.....	12
1.4.3 Mô tả chi tiết về các bảng Dimension.....	14
1.5 Các câu truy vấn.....	15
CHƯƠNG 2. XÂY DỰNG KHO DỮ LIỆU (SSIS)	17
2.1 Chuẩn bị các công cụ.....	17
2.2 Chuẩn bị cơ sở dữ liệu	19
2.3 Tạo mới project SSIS	21
2.4 Tạo bảng Dim và bảng Fact.....	23
2.4.1 Bảng Dim Date	29
2.4.2 Bảng Dim Language	37
2.4.3 Bảng Dim Genres List	40
2.4.4 Bảng Dim Movie	42
2.4.5 Bảng Dim Country	45
2.4.6 Bảng Dim Director	47
2.4.7 Bảng Dim Company	49
2.4.8 Tạo bảng Fact Movie	51
2.4.9 Tạo các khôi lệnh SQL	81
2.4.10 Thực thi project và kết quả SSIS	85
2.4.11 Kiểm tra dữ liệu các bảng	86
CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU SSAS	90
3.1 Chuẩn bị các công cụ.....	90
3.1.1 Cài đặt Microsoft Analysis Services Projects	90
3.1.2 Cài đặt Analysis Services	92
3.2 Tạo mới Project SSAS.....	100
3.3 Thiết lập nguồn dữ liệu (Data Sources)	101
3.4 Thiết lập khung nhìn dữ liệu nguồn (Data Source Views).....	105
3.5 Thiết lập các khôi (Cube).....	109
3.5.1 Tạo Cube và Dimension.....	109
3.5.2 Thêm thuộc tính vào Dimension	113
3.5.3 Xác định các độ đo.....	118

3.5.4	Phân cấp các bảng chiều	119
3.6 Thực hiện các câu truy vấn		124
3.6.1	Câu truy vấn 1: Top 5 năm có tổng doanh thu cao nhất.....	124
3.6.2	Câu truy vấn 2: Top 10 bộ phim có kinh phí sản xuất cao nhất, xếp giảm dần	129
3.6.3	Câu truy vấn 3: Top 10 phim có nhiều lượt đánh giá nhất theo thứ tự giảm dần của độ phổ biến.....	135
3.6.4	Câu truy vấn 4: Truy vấn các tháng của mỗi năm 2020 có doanh thu hơn 10 triệu, sắp xếp theo thứ tự tăng dần trong tháng	141
3.6.5	Câu truy vấn 5: Liệt kê top 3 quốc gia sản xuất có doanh thu cao nhất trong từng năm.....	145
3.6.6	Câu truy vấn 6: Truy vấn các quốc gia có doanh thu nằm trong top 10% của năm 2012 và đồng thời cũng nằm trong top 10% của năm 2013	150
3.6.7	Câu truy vấn 7: Top 7 các quốc gia có tổng doanh thu cao nhất với điểm đánh giá trung bình trên 6.5 trong năm 2014	154
3.6.8	Câu truy vấn 8: Mỗi tháng trong năm 2018, liệt kê phim nổi tiếng nhất trong từng tháng	159
3.6.9	Câu truy vấn 9: Độ nổi tiếng của Top 10 công ty sản xuất có số lượng phim ít nhất	163
3.6.10	Câu truy vấn 10: Thông kê số lượng phim, số lượt bình chọn và điểm đánh giá trung bình được sản xuất bởi các nước công ty Jules Jordan Video, Naughty America, Marvel Studios, Pixar và Digital Playground	167
3.6.11	Câu truy vấn 11: Thông kê số lượng phim, tổng số lượt đánh giá, và độ phổ biến theo thể loại phim và các ngôn ngữ khác nhau (Việt, Hàn, Trung, Nhật)	171
3.6.12	Câu truy vấn 12: Tổng kinh phí đầu tư hàng tháng và cả năm của các phim được sản xuất tại Mỹ với thời gian công chiếu là từ năm 2017 đến năm 2019	176
3.6.13	Câu truy vấn 13: Thông kê số lượng phim, tổng số lượt đánh giá, điểm đánh giá trung bình và độ phổ biến của các bộ phim có ngôn ngữ gốc là tiếng Anh và Pháp và được phát hành từ năm 2016 đến 2023 và có công ty sản xuất tại Mỹ	181
3.6.14	Câu truy vấn 14: Tính tổng số lượng phim của các công ty sản xuất theo các tháng từ năm 2014 đến năm 2017	185
3.6.15	Câu truy vấn 15: Tính tổng số lượng phim theo thể loại và đạo diễn	189
CHƯƠNG 4. QUÁ TRÌNH KHAI THÁC DỮ LIỆU (DATA MINING).....		194
4.1 Mô tả bộ dữ liệu.....		194
4.2 Mô tả bài toán.....		195
4.2.1	Mô tả bài toán	195
4.2.2	Mục tiêu	195
4.3 Khám phá cơ bản bộ dữ liệu		196
4.4 Xử lý đặc trưng cho mô hình máy học		198
4.4.1	Cột revenue.....	198
4.4.2	Cột vote_average	199
4.4.3	Cột runtime	200
4.4.4	Cột title	201
4.4.5	Cột budget.....	202
4.4.6	Cột popularity	203
4.4.7	Cột vote_count.....	204
4.4.8	Cột overview_sentiment	204
4.4.9	Cột status	206
4.4.10	Cột adult	207
4.4.11	Cột production_countries.....	207
4.4.12	Cột spoken_languages	209
4.4.13	Cột original_language.....	210
4.4.14	Cột release_date.....	211
4.4.15	Cột genres_list	214
4.4.16	Xử lý các đặc trưng số (Numerical Feature Transformation)	215
4.4.17	Lựa chọn các đặc trưng (Feature Selection).....	216
4.5 Tiền xử lý dữ liệu.....		217

4.6 Mô hình máy học phân loại	218
4.6.1 Catboost Classifier	218
4.6.2 Decision Tree Classifier.....	219
4.6.3 LightGBM Classifier	222
4.6.4 Độ đo đánh giá.....	224
4.6.5 Kết luận.....	227
Tài liệu tham khảo	238

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

Chương này sẽ là phần giới thiệu tổng quát về đề tài “PHÂN TÍCH DỮ LIỆU PHIM ĐIỆN ẢNH TỪ NGUỒN IMDB VÀ TMDB”. Nội dung của chương sẽ bao gồm: lý do chọn đề tài, giới thiệu về dataset và thiết kế kho dữ liệu cùng với các câu truy vấn.

1.1 Lý do chọn đề tài

Phân tích dữ liệu phim điện ảnh từ nguồn IMDb và TMDB là một đề tài có tính ứng dụng cao trong việc nghiên cứu và hiểu rõ hơn về xu hướng, sở thích của khán giả, cũng như các yếu tố ảnh hưởng đến thành công của một bộ phim. Đây là một lĩnh vực được quan tâm rất nhiều trong ngành công nghiệp giải trí, bởi điện ảnh không chỉ là một loại hình nghệ thuật mà còn là một ngành kinh tế lớn với sức ảnh hưởng sâu rộng trên toàn cầu.

Việc phân tích dữ liệu từ IMDb và TMDB sẽ giúp có cái nhìn toàn diện về các khía cạnh như doanh thu, đánh giá của khán giả, xu hướng thể loại, và các yếu tố làm nên thành công của phim. Thông qua việc xử lý và trực quan hóa dữ liệu, có thể đưa ra các nhận định hữu ích cho nhà sản xuất, đạo diễn và các đơn vị phát hành phim, từ đó tối ưu hóa chiến lược sản xuất và quảng bá để thu hút khán giả hiệu quả hơn.

Ngoài ra, nghiên cứu này còn giúp hiểu sâu hơn về sự khác biệt giữa các thị trường, cách mà các yếu tố văn hóa và xã hội ảnh hưởng đến thị hiếu khán giả, cũng như cách sử dụng dữ liệu để dự đoán xu hướng trong tương lai. Điều này không chỉ có giá trị thực tiễn trong ngành điện ảnh mà còn đóng góp vào sự phát triển của các mô hình phân tích dữ liệu lớn trong lĩnh vực giải trí.

Vì vậy, em chọn đề tài phân tích dữ liệu phim điện ảnh từ nguồn IMDb và TMDB vì đây là một đề tài thú vị, mang tính ứng dụng cao, và có ý nghĩa thiết thực trong việc phát triển ngành công nghiệp điện ảnh.

1.2 Giới thiệu về Dataset

1.2.1 Thông tin về Dataset

Tên dataset: **IMDB & TMDB Movie Metadata Big Dataset (over 1M)**

The screenshot shows the Kaggle interface. On the left, there's a sidebar with navigation links like Home, Competitions, Datasets, Models, Code, Discussions, Learn, More, Your Work, and a list of datasets including 'IMDB & TMDB Movie Metadata Big Dataset (over 1M)' and 'VN_MOTO_DATASET'. The main content area displays the dataset details for 'IMDB & TMDB Movie Metadata Big Dataset (over 1M)'. It includes a title, subtitle ('A Comprehensive Dataset Featuring Detailed Metadata of Movies (IMDB, TMDB).'), a Data Card tab, and other sections like 'About Dataset' which provide information about the title, subtitle, detailed description, overview, and various metrics such as Usability (7.65), License (MIT), and Expected update frequency (Not specified). There are also sections for Tags and a preview image featuring the IMDb logo.

Hình 1.1 Dataset gốc

- Dataset là tập dữ liệu toàn diện kết hợp thông tin phong phú về phim từ cả IMDB và TMDB, mang đến một nguồn tài nguyên đa dạng dành cho những người yêu thích điện ảnh, các nhà khoa học dữ liệu, và các nhà nghiên cứu.
- Ngày cập nhật: 24/11/2024.

1.2.2 Tác giả

- Tên tác giả: Shubham Chandra

1.2.3 Số dòng, số cột và thời gian thu thập

- Bộ dữ liệu gồm: 42 cột thuộc tính và 1.072.255 dòng dữ liệu.
- Dữ liệu sau khi lọc còn: cột thuộc tính và dòng dữ liệu.
- Dataset được thu thập liên tục từ nguồn IMDb và TMDB.
- Dữ liệu phim được thu thập từ IMDB và TMDB, bao gồm các thông tin chi tiết về phim như thể loại, doanh thu, đánh giá từ khán giả và các đặc điểm khác liên quan đến ngành công nghiệp điện ảnh. Thông tin chi tiết về quá trình thu thập và xử lý dữ liệu này có thể được tìm thấy trên trang tài nguyên của Internet Movie Database (IMDB) và The Movie Database (TMDB).

1.2.4 Nguồn Dataset

- Nguồn: <https://www.kaggle.com/datasets/shubhamchandra235/imdb-and-tmdb-movie-metadata-big-dataset-1m>

1.2.5 Tiền xử lý và làm sạch dữ liệu

SVTH: Nguyễn Hồng Phát

Kho dữ liệu và OLAP - IS217.P12

Bước 1: Tải các thư viện cần thiết:

```
import numpy as np
import polars as pl
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime
```

Bước 2: Tải và đọc bộ dữ liệu:

```
from google.colab import drive
drive.mount('/gdrive')

path = '/gdrive/MyDrive/IS217.P12_22521072/Data/IMDB TMDB Movie Metadata Big Dataset (1M).csv'
df = pl.read_csv(path)
df = df.to_pandas()
```

Bước 3: Chuyển đổi dữ liệu sang dạng phù hợp:

```
df['release_date'] = pd.to_datetime(df['release_date'], errors='coerce')
# Biến đổi cột release_date thành kiểu datetime
# Thay giá trị không hợp lệ hoặc không thể chuyển đổi bằng NaT (Not a Time)
```

Bước 4: Kiểm tra tỷ lệ dữ liệu bị thiếu trong các trường dữ liệu:

```
summary = pd.DataFrame({
    "Data Type": df.dtypes, # Kiểu dữ liệu của mỗi cột
    "Missing (%)": df.isnull().mean() * 100, # Tỷ lệ phần trăm giá trị thiếu
    "Unique Values": df.nunique(), # Số lượng giá trị duy nhất trong mỗi cột
})

# Vẽ biểu đồ barh cho các giá trị thiếu (% thiếu)
plt.figure(figsize=(12, 8)) # Giảm chiều rộng của biểu đồ để chữ không bị tràn ra
missing_values = df.isnull().mean() * 100 # Tính tỷ lệ phần trăm giá trị thiếu
missing_values = missing_values[missing_values > 0] # Chỉ lấy các cột có giá trị thiếu

# Vẽ biểu đồ cột ngang
bars = plt.bart(missing_values.index, missing_values.values, color='skyblue')
plt.title('Missing Values in Dataset Features (%)', fontsize=15)
plt.xlabel('Percentage of Missing Values (%)', fontsize=13)
plt.grid(axis='x', linestyle='--', alpha=0.7)

# Hiển thị tỷ lệ phần trăm ngay giữa mỗi cột
for i, bar in enumerate(bars):
    # Tính vị trí trung tâm của cột
    width = bar.get_width()
    plt.text(width / 2, bar.get_y() + bar.get_height() / 2, f'{width:.2f}%', color='black', ha='center', va='center', fontsize=13)
```

Kho dữ liệu và OLAP - IS217.P12

```
# Tăng cỡ của các nhãn trục x và y để dễ đọc
```

```
plt.xticks(fontsize=13)
```

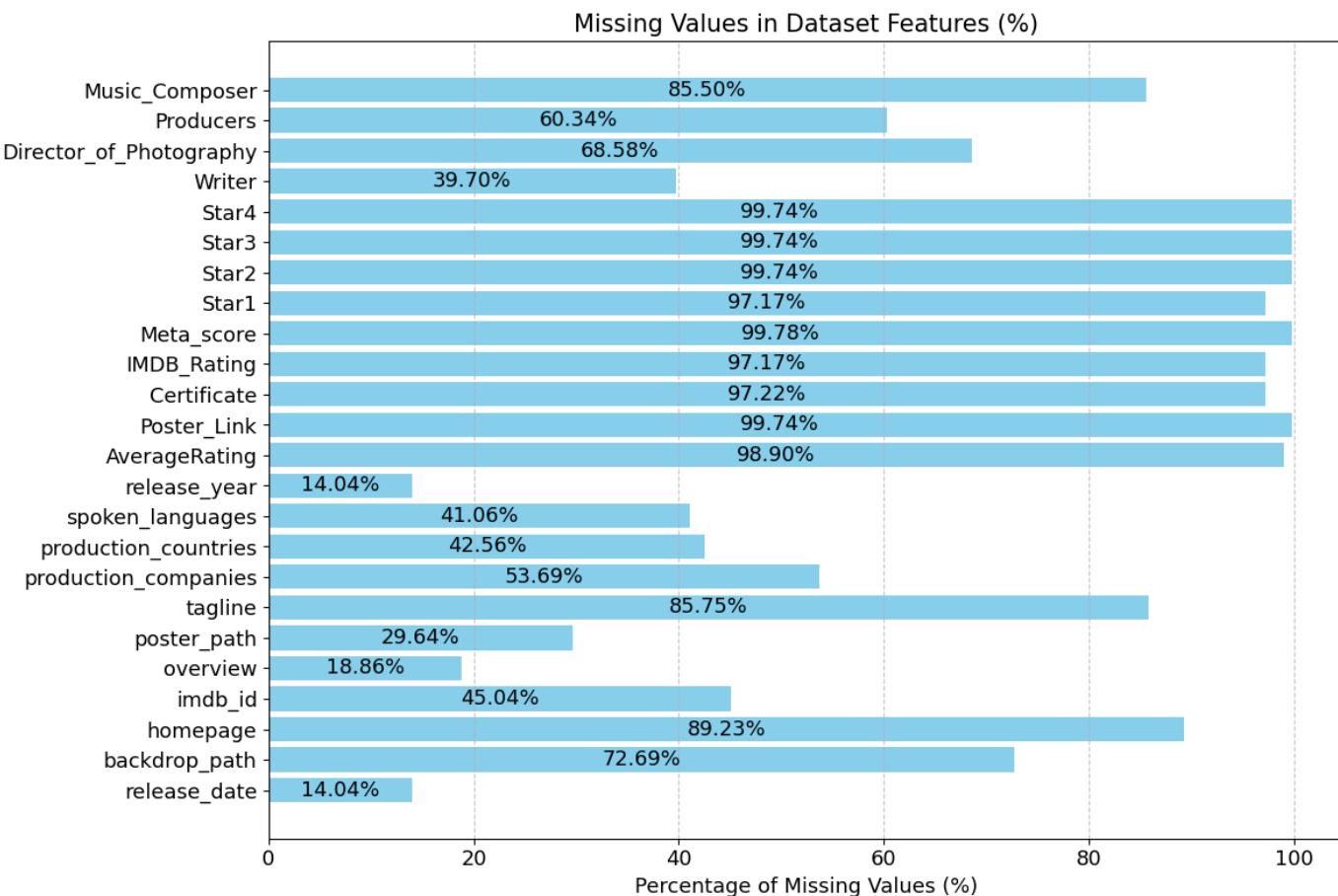
```
plt.yticks(fontsize=13)
```

```
# Tự động điều chỉnh bố cục để các nhãn không bị cắt
```

```
plt.tight_layout()
```

```
# Hiển thị biểu đồ
```

```
plt.show()
```



Hình 1.2 Tỷ lệ dữ liệu bị thiếu trong các trường dữ liệu

Bước 5: Loại bỏ các cột có tỉ lệ dữ liệu bị thiếu trên 50% nhưng em sẽ giữ lại cột production_companies để dùng tạo bảng trong DataWarehouse:

```
missing_values = df.isnull().mean() * 100
```

```
# Xác định các cột cần loại bỏ do thiếu trên 50%, trừ 'production_companies'
```

```
columns_to_drop_due_to_missing = missing_values[(missing_values > 50) & (missing_values.index != 'production_companies')].index
```

```
df = df.drop(columns=columns_to_drop_due_to_missing, errors='ignore')
```

Bước 6: Loại bỏ các dòng có id và title bị trùng:

SVTH: Nguyễn Hồng Phát

Kho dữ liệu và OLAP - IS217.P12

```
df = df.drop_duplicates(subset=['id'])
df = df.drop_duplicates(subset=['title'], keep='first')
```

Bước 7: Loại bỏ các dòng có kích thước quá lớn (hơn 255 ký tự):

```
max_length = 255
condition = True

for col in df.select_dtypes(include='object').columns:
    condition &= df[col].apply(lambda x: len(str(x)) <= max_length)

df = df[condition]
```

Bước 8: Xóa các dòng không cần thiết và loại bỏ các dòng Null, trùng:

```
columns_to_drop = ['poster_path', 'overview',
'original_title','all_combined_keywords','keywords','overview_sentiment','Cast_list','Writer','adult','release_year']
df = df.drop(columns=columns_to_drop, errors='ignore')
df = df.dropna()
df = df.drop_duplicates()
```

Bước 9: Lọc dữ liệu chứa thông tin các phim từ năm 2000 trở đi sẽ được dùng để tạo datawarehouse:

```
df = df[df['release_date'].dt.year >= 2000]
```

Bước 10: Lưu lại bộ dữ liệu:

```
df.to_csv('Cleaned_Dataset.csv', index=False)
```

Bộ dữ liệu sau khi làm sạch: Gồm 17 cột và 17167 dòng.

1.3 Mô tả chi tiết các cột thuộc tính của kho dữ liệu

STT	Tên cột	Kiểu dữ liệu	Ý nghĩa
1	id	int	Mã định danh duy nhất cho mỗi phim trong TMDB
2	title	varchar	Tên chính thức của phim
3	vote_average	float	Đánh giá trung bình của phim theo thang điểm từ 0 đến 10
4	vote_count	int	Số phiếu đánh giá bộ phim

Kho dữ liệu và OLAP - IS217.P12

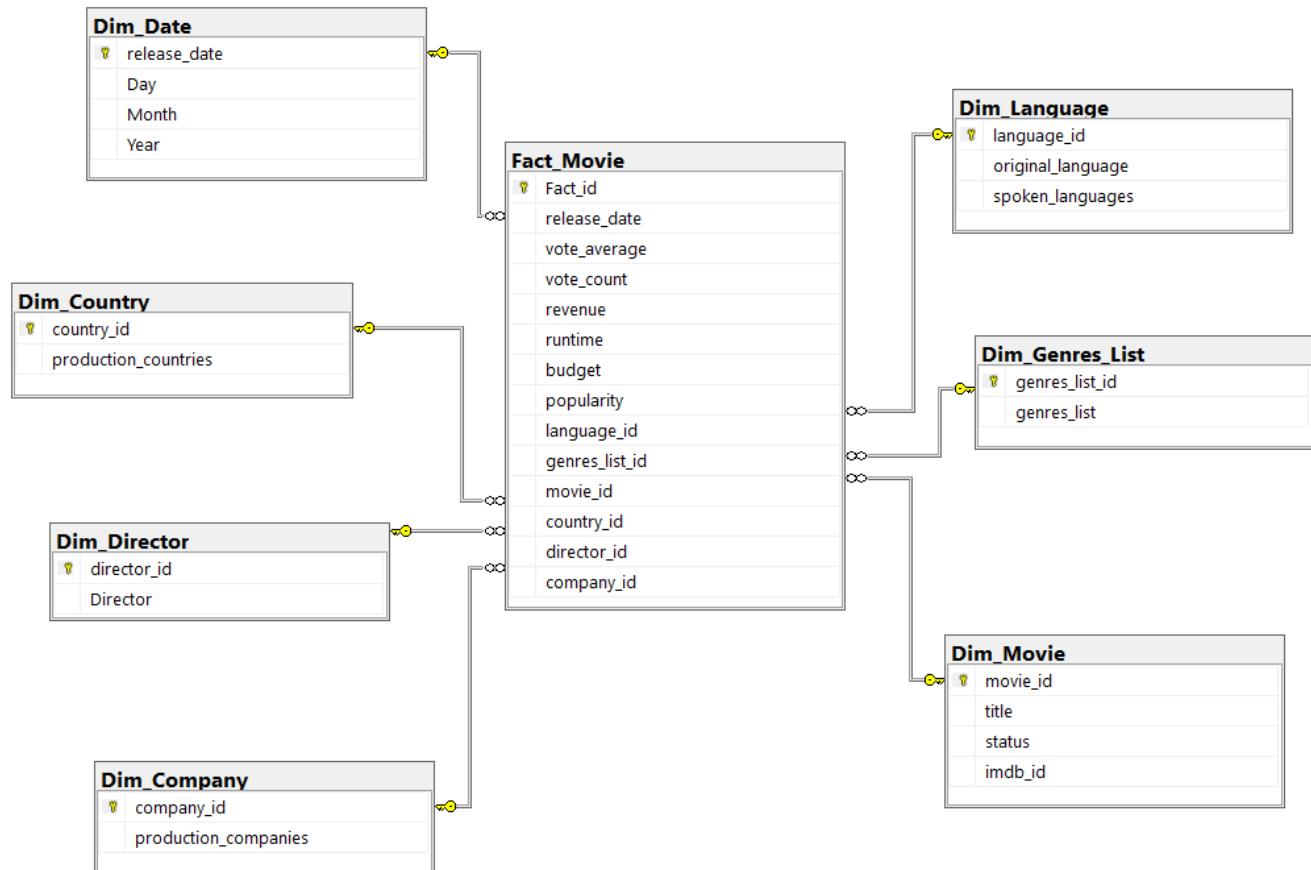
5	status	varchar	Trạng thái hiện tại của phim
6	release_date	datetime	Ngày phim chính thức ra mắt
7	revenue	int	Doanh thu của bộ phim
8	runtime	int	Thời lượng của phim
9	budget	int	Ngân sách sản xuất của bộ phim
10	imdb_id	varchar	ID của phim trên nền tảng IMDb
11	original_language	varchar	Ngôn ngữ ban đầu của bộ phim
12	popularity	float	Điểm phổ biến của phim
13	production_companies	varchar	Công ty tham gia sản xuất
14	production_countries	varchar	Quốc gia tham gia sản xuất
15	spoken_languages	varchar	Các ngôn ngữ được sử dụng trong bộ phim
16	Director	varchar	Đạo diễn của bộ phim
17	genres_list	varchar	Thể loại của bộ phim

Bảng 1 Thuộc tính dataset sau khi làm sạch

1.4 Thiết kế kho dữ liệu

Bộ dữ liệu sau khi làm sạch thì còn 17167 dòng dữ liệu để thực hiện thiết kế kho dữ liệu.

1.4.1 Thiết kế lược đồ hình sao



Hình 1.3 Lược đồ hình sao

1.4.2 Mô tả chi tiết về bảng Fact

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	id	Int	PK	Mã định danh duy nhất của bảng
2	release_date	Datetime	FK	Ngày phim chính thức ra mắt

Kho dữ liệu và OLAP - IS217.P12

3	language_id	Int	FK	Mã định danh duy nhất cho ngôn ngữ của mỗi phim
4	genres_list_id	Int	FK	Mã định danh duy nhất cho danh sách các thể loại của mỗi phim
5	country_id	int	FK	Mã định danh duy nhất cho danh sách các quốc gia tham gia sản xuất của mỗi phim
6	company_id	Int	FK	Mã định danh duy nhất cho danh sách các công ty tham gia sản xuất của mỗi phim
7	director_id	Int	FK	Mã định danh duy nhất cho danh sách các đạo diễn của mỗi phim
8	movie_id	Int	FK	Mã định danh duy nhất cho mỗi phim
9	vote_average	float		Đánh giá trung bình của phim theo thang điểm từ 0 đến 10
10	vote_count	int		Số phiếu đánh giá bộ phim
11	revenue	int		Doanh thu của bộ phim
12	runtime	int		Thời lượng của phim
13	budget	budget		Ngân sách sản xuất của bộ phim
14	popularity	float		Điểm phổ biến của phim

Bảng 2 Chi tiết bảng Fact

1.4.3 Mô tả chi tiết về các bảng Dimension

1.4.3.1 Dim Date

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	release_date	Datetime	PK	Thời gian phim được công chiếu Định dạng: MM/DD/YYYY hh:mm
2	day	Int		Lấy ngày trong thời gian phim công chiếu
3	month	Int		Lấy tháng trong thời gian phim công chiếu
4	year	Int		Lấy năm trong thời gian phim công chiếu

Bảng 3 Chi tiết bảng Dim Date

1.4.3.2 Dim Language

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	language_id	Int	FK	Mã định danh duy nhất cho ngôn ngữ của mỗi phim
2	original_language	Varchar		Ngôn ngữ ban đầu của bộ phim
3	spoken_languages	Varchar		Ngôn ngữ được dùng trong phim

Bảng 4 Chi tiết bảng Dim Language

1.4.3.3 Dim Company

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	company_id	Int	PK	Số nhận dạng cho các công ty sản xuất
2	production_companies	Varchar		Tên các công ty sản xuất

Bảng 5 Chi tiết bảng Dim Company

1.4.3.4 Dim Country

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	country_id	Int	PK	Số nhận dạng cho các quốc gia sản xuất
2	production_countries	Varchar		Quốc gia tham gia sản xuất

Bảng 6 Chi tiết bảng Dim Country

1.4.3.5 Dim Genres List

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	genres_list_id	Int	PK	Số nhận dạng cho thể loại của bộ phim
2	genres_list	Varchar		Thể loại của bộ phim

Bảng 7 Chi tiết bảng Dim Genres List

1.4.3.6 Dim Director

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	director_id	Int	PK	Số nhận dạng cho các đạo diễn của bộ phim
2	Director	Varchar		Đạo diễn của bộ phim

Bảng 8 Chi tiết bảng Dim Director

1.4.3.7 Dim Movie

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa
1	movie_id	Int	PK	Số nhận dạng cho bộ phim
2	title	Varchar		Tên chính thức của phim
3	status	Varchar		Trạng thái hiện tại của phim
4	imdb_id	Varchar		ID của phim trên nền tảng IMDb

Bảng 9 Chi tiết bảng Dim Movie

1.5 Các câu truy vấn

Câu 1: Top 5 năm có tổng doanh thu cao nhất.

Câu 2: Top 10 bộ phim có kinh phí sản xuất cao nhất, xếp giảm dần.

Câu 3: Top 10 phim có nhiều lượt đánh giá nhất theo thứ tự giảm dần của độ phổ biến.

Câu 4: Truy vấn các tháng của mỗi năm 2020 có doanh thu hơn 10 triệu, sắp xếp theo thứ tự tăng dần trong tháng.

Câu 5: Liệt kê top 3 quốc gia sản xuất phim có doanh thu cao nhất trong từng năm.

Câu 6: Truy vấn các quốc gia có doanh thu nằm trong top 10% của năm 2012 và đồng thời cũng nằm trong top 10% của năm 2013.

Câu 7: Top 7 các quốc gia có tổng doanh thu trên 10 triệu và điểm đánh giá trung bình trên 6.5 trong năm 2014.

Câu 8: Mỗi tháng trong năm 2018, liệt kê phim nổi tiếng nhất trong từng tháng.

Câu 9: Độ nổi tiếng của Top 10 công ty sản xuất có số lượng phim ít nhất.

Câu 10: Thông kê số lượng phim, số lượt bình chọn, điểm đánh giá trung bình và mức độ phổ biến được sản xuất bởi các nước công ty Jules Jordan Video, Naughty America, Marvel Studios, Pixar và Digital Playground.

Câu 11: Thông kê số lượng phim, tổng số lượt đánh giá, và độ phổ biến theo thể loại phim và các ngôn ngữ khác nhau (Việt, Hàn, Trung, Nhật).

Câu 12: Tổng kinh phí đầu tư hàng tháng và cả năm của các phim được sản xuất tại Mỹ với thời gian công chiếu là từ năm 2017 đến năm 2019.

Câu 13: Thông kê số lượng phim, tổng số lượt đánh giá, điểm đánh giá trung bình và độ phổ biến của các bộ phim có ngôn ngữ gốc là tiếng Anh và Pháp và được phát hành từ năm 2016 đến 2023 và có công ty sản xuất tại Mỹ.

Câu 14: Top 10 công ty sản xuất có doanh thu cao nhất trong năm 2020 kèm theo thể loại.

Câu 15: Tính tổng số lượng phim theo thể loại và đạo diễn.

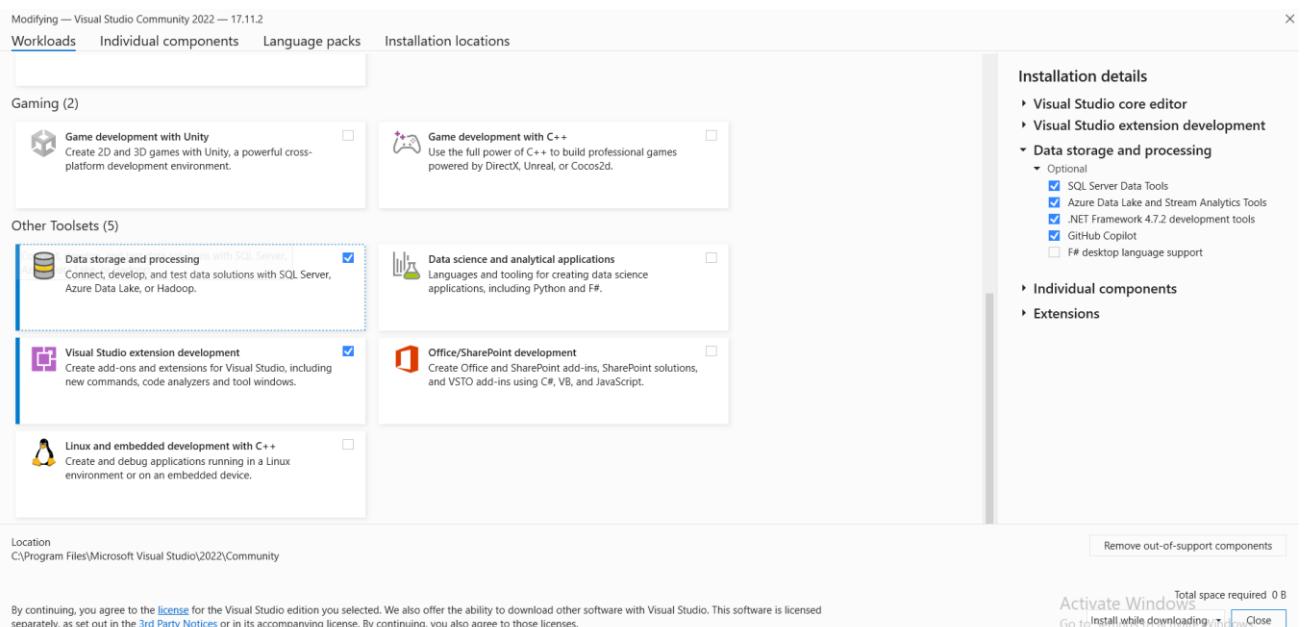
CHƯƠNG 2. XÂY DỰNG KHO DỮ LIỆU (SSIS)

2.1 Chuẩn bị các công cụ

Để thực hiện quá trình SSIS ta cần chuẩn bị và cài đặt các công cụ sau:

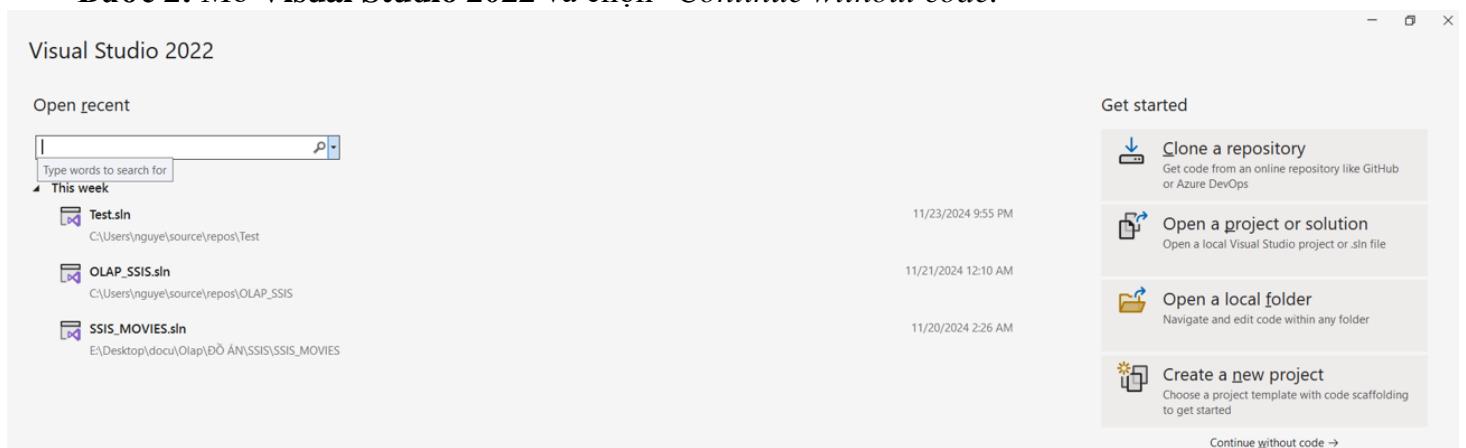
- Visual Studio Community 2022
- SQL Server Integration Services Project

Bước 1: Tải [Visual Studio Community 2022](#) về máy. Trong lúc cài đặt, chọn mục “Data storage and processing” để cài đặt SQL Server Data Tools. Sau đó chọn Install để tiến hành cài đặt.



Hình 2.1 Cài đặt Toolsets

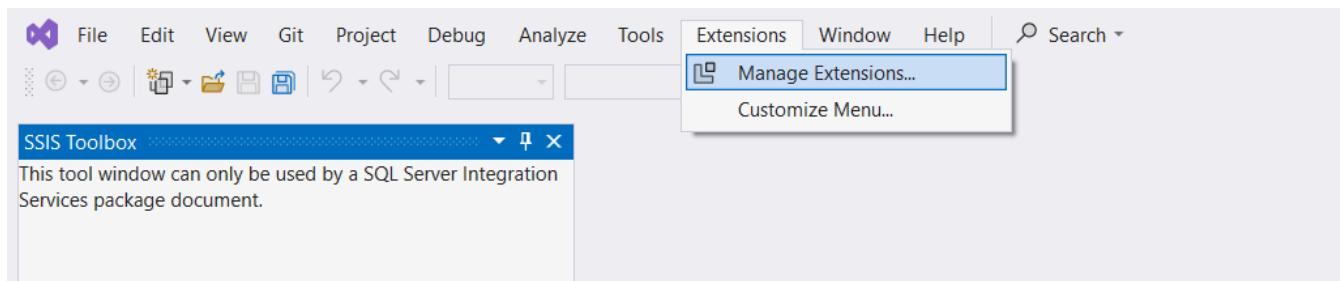
Bước 2: Mở [Visual Studio 2022](#) và chọn “Continue without code.”



Hình 2.2 Khởi động Visual Studio 2022

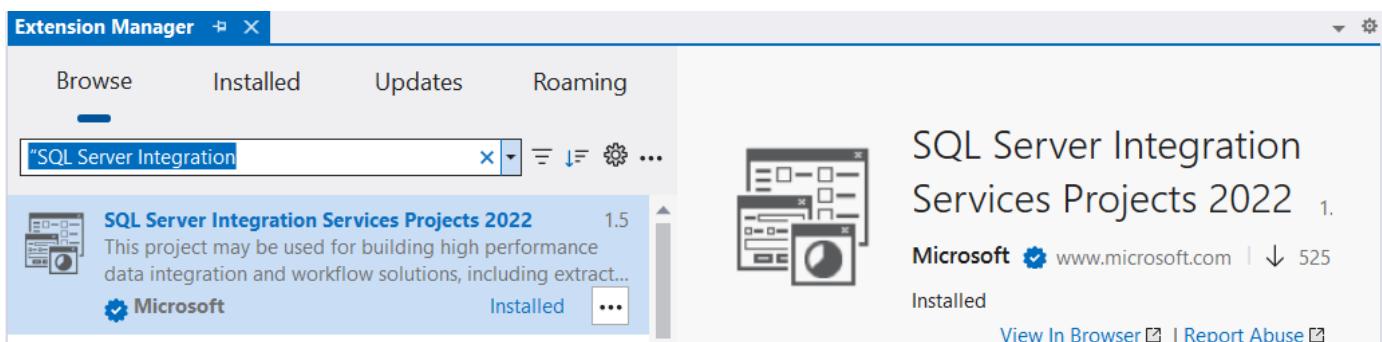
Kho dữ liệu và OLAP - IS217.P12

Bước 3: Trong giao diện chính , chọn "Extensions" ở thanh công cụ phía trên, sau đó chọn "Manage Extensions"



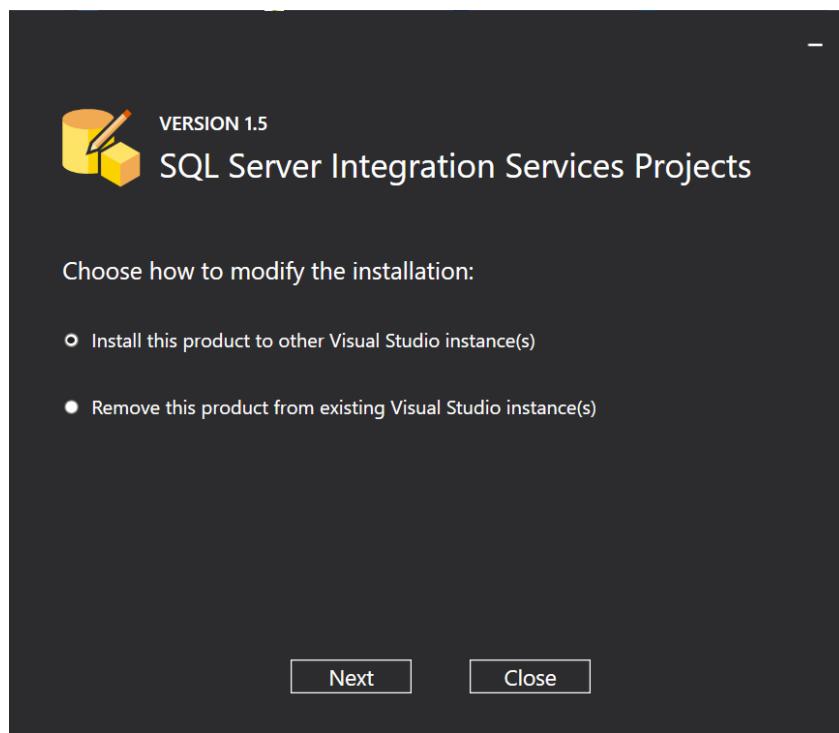
Hình 2.3 Cài đặt Extention

Bước 4: Tìm và tải về công cụ **SQL Server Integration Services Projects**.



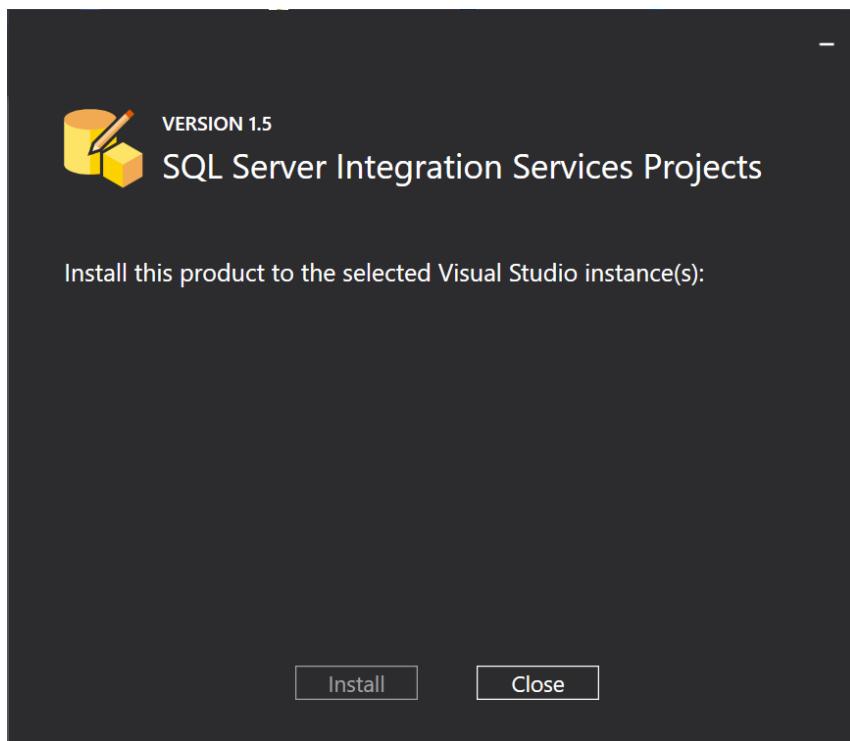
Hình 2.4 Cài đặt SQL Server Integration Services Projects 2022

Bước 5: Mở file vừa tải xuống, ta được giao diện như hình, chọn “Next” để tiếp tục.



Hình 2.5 Mở SQL Server Integration Services Projects 2022

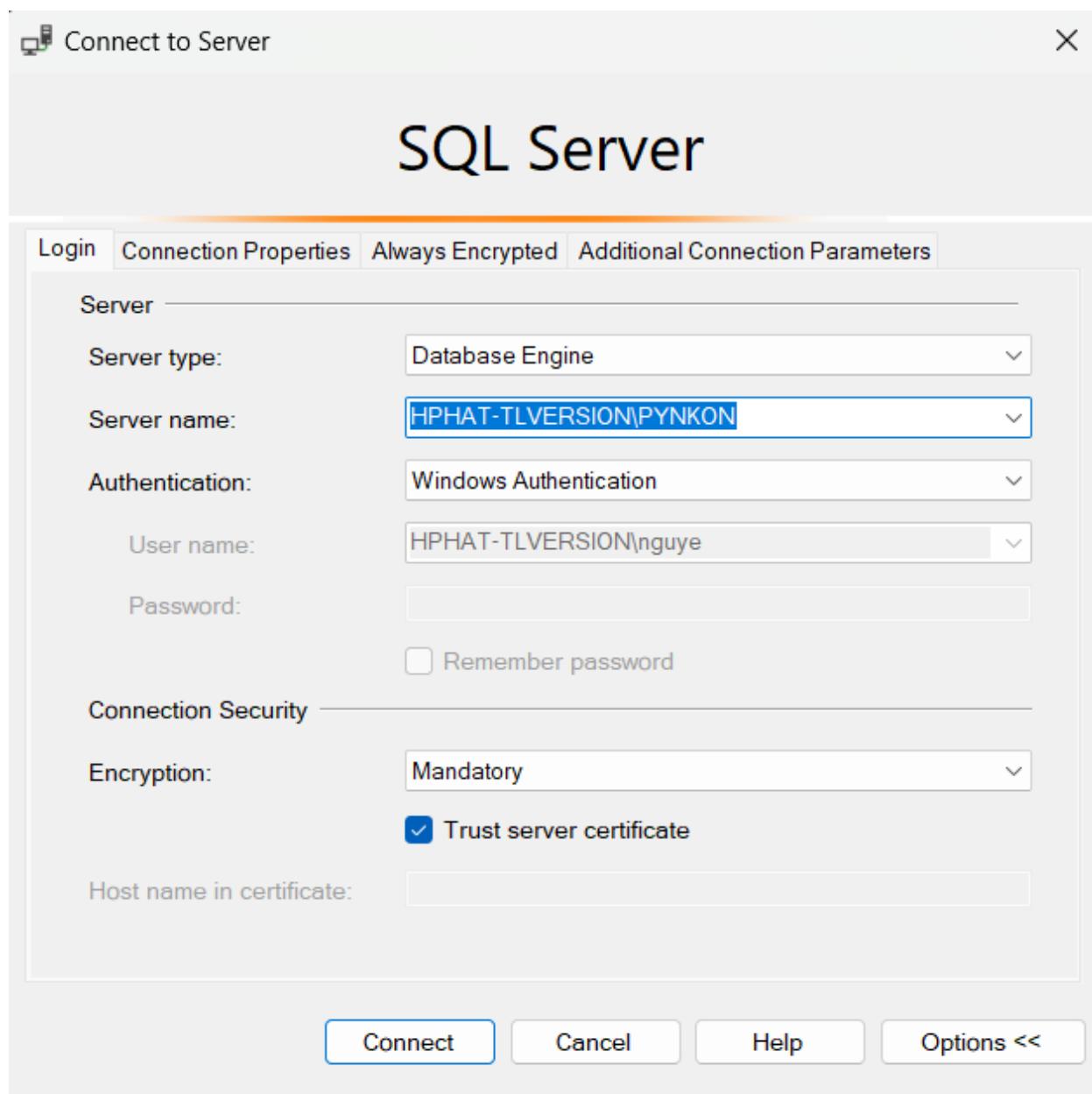
Bước 6: Tick chọn vào ô Visual Studio 2022 và chọn “Install” để tiến hành cài đặt



Hình 2.6 Install SQL Server Integration Services Projects 2022

2.2 Chuẩn bị cơ sở dữ liệu

Bước 1: Mở SQL Server 2022 và kết nối với server bằng tài khoản user của windows (Windows Authentiacation)



Hình 2.7 Kết nối với SQL server

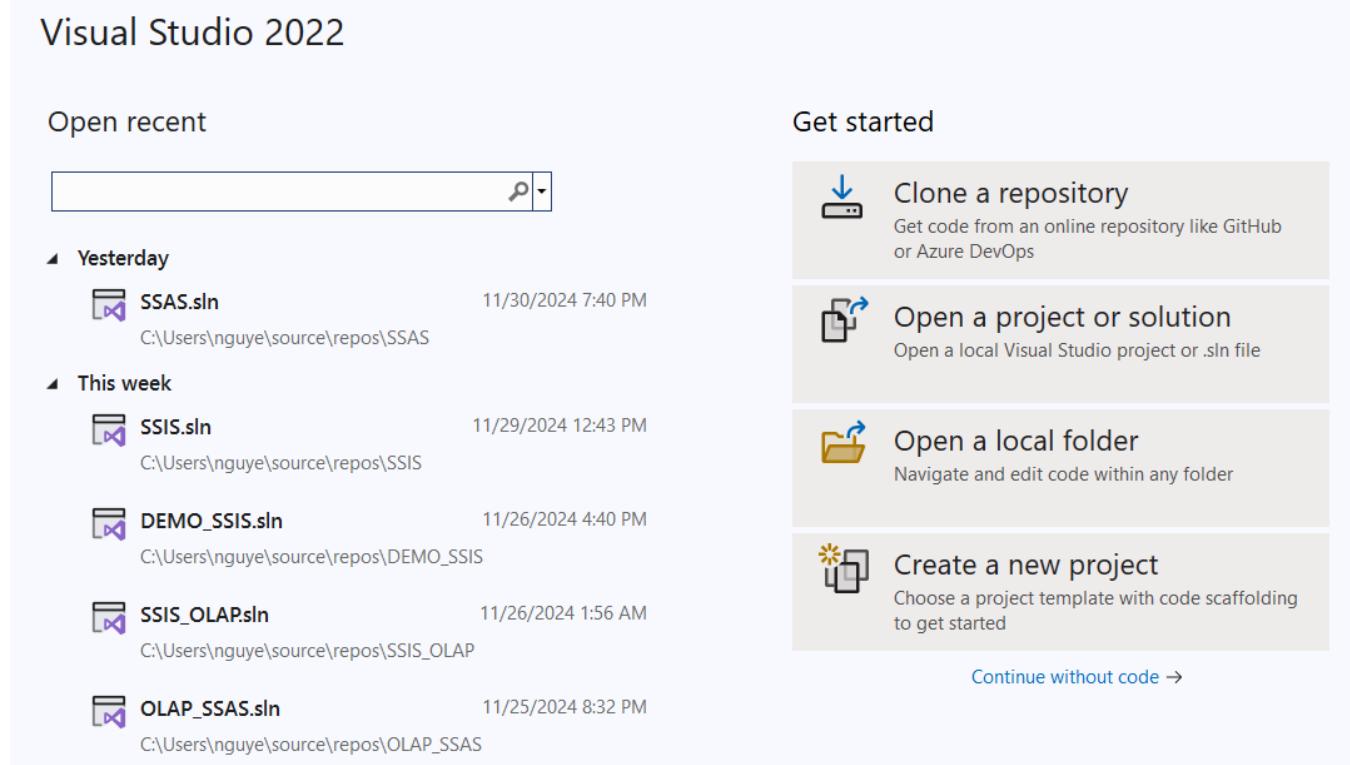
Bước 2: Khởi tạo một cơ sở dữ liệu có tên là **Movie_wh**, đây là nơi lưu các bảng Dim và bảng Fact cùng dữ liệu của các bảng đó.

```
SQLQuery1.sql - H...ERSION\nguye (51)* CREATE DATABASE Movie_wh use Movie_wh
```

Hình 2.8 Khởi tạo cơ sở dữ liệu

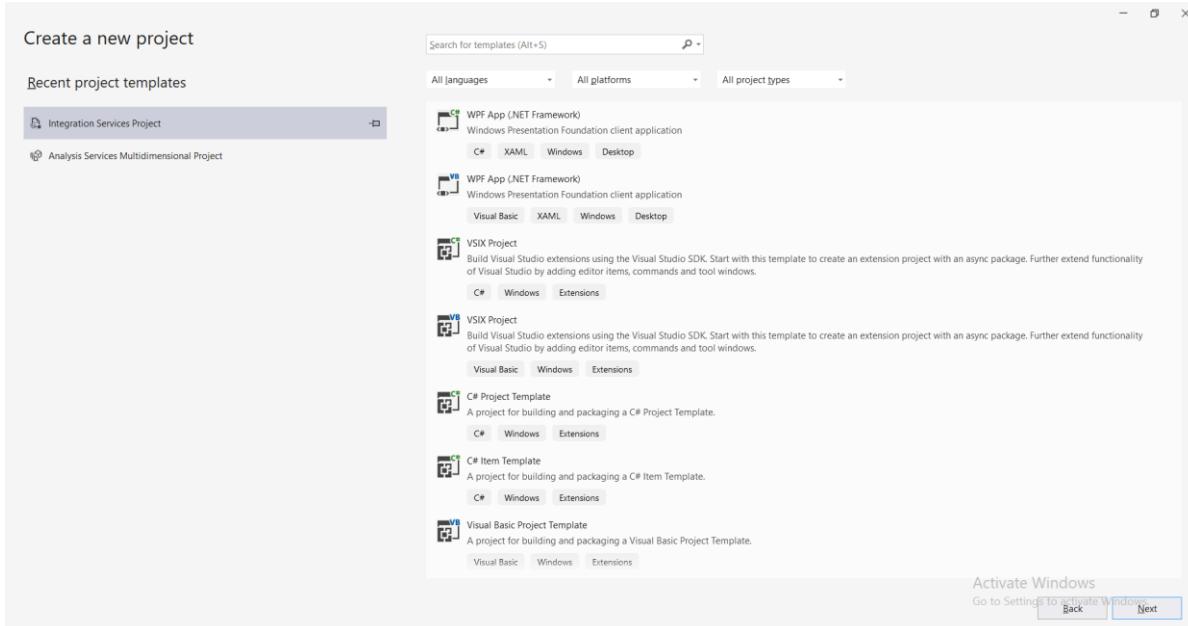
2.3 Tạo mới project SSIS

Bước 1: Mở Visual Studio 2022 và chọn “Create a new project”.



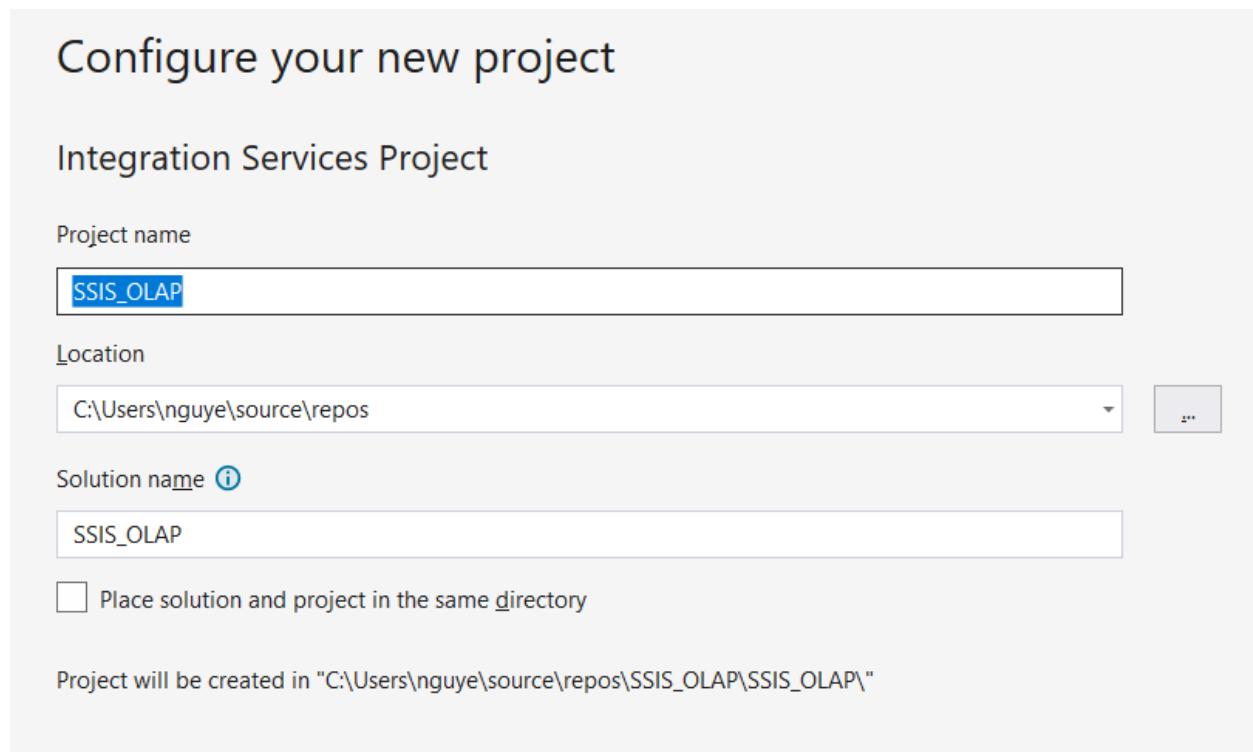
Hình 2.9 Khởi tạo Project mới

Bước 2: Chọn Integration Services Project và chọn Next



Hình 2.10 Khởi tạo Integration Services Projec

Bước 3: Đặt tên và thiết lập đường dẫn cho Project. Sau đó chọn **Create**.

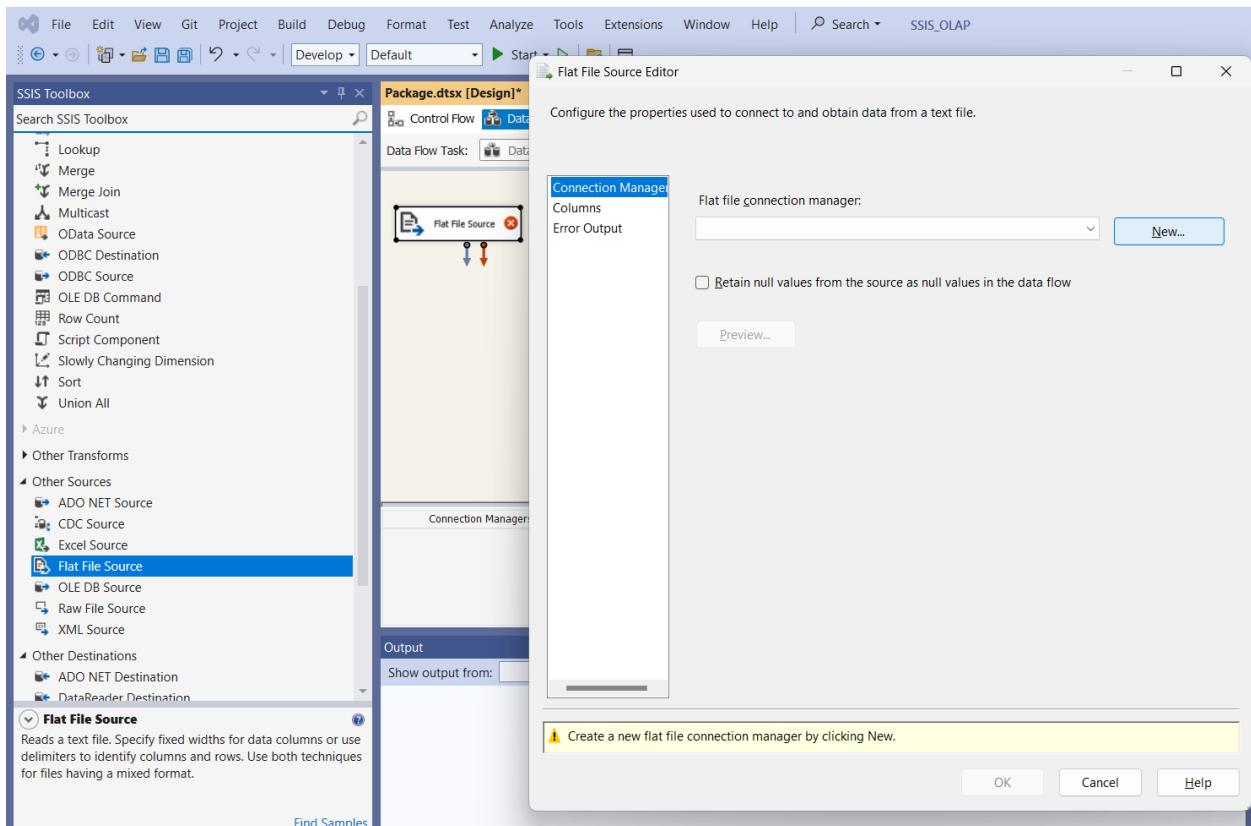


Hình 2.11 Đặt tên Project

2.4 Tạo bảng Dim và bảng Fact

Trước khi tiến hành chia Dimension và bảng Fact, ta cần load dữ liệu gốc từ file.csv vào Data Flow:

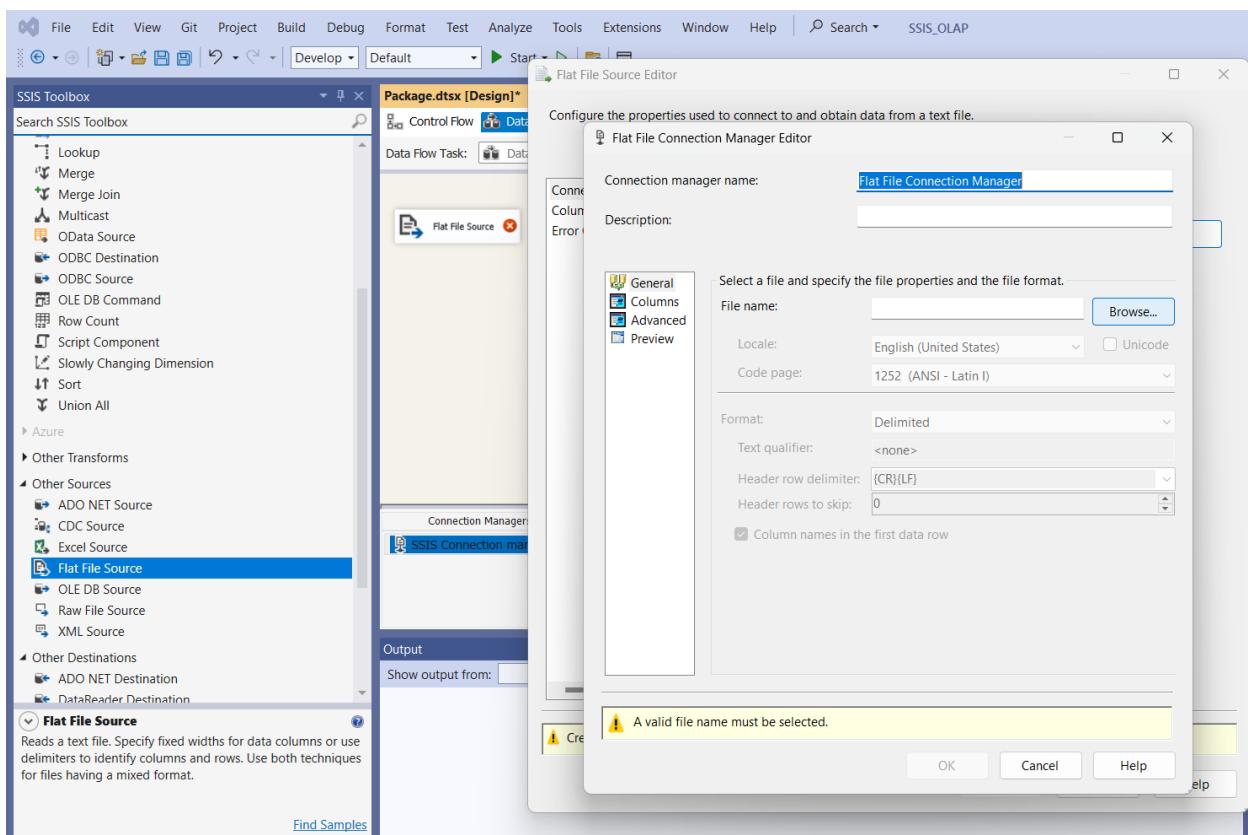
Bước 1: Trong Data Flow, tạo một đối tượng Flat File Source để lấy dữ liệu gốc từ file.csv. Chọn New để tạo một *Flat File Connection Manager*



Hình 2.12 Tạo Flat File Source

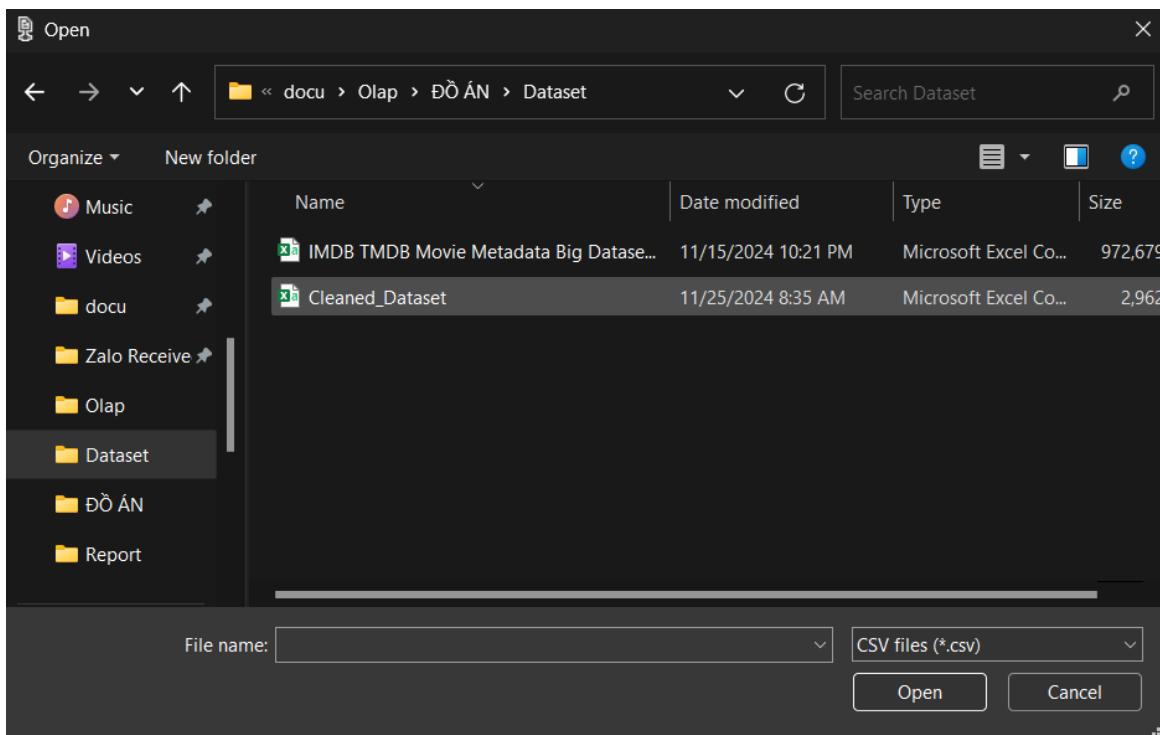
Bước 2: Chọn nút **Browse...** để tải lên file dữ liệu gốc lưu trong máy.

Kho dữ liệu và OLAP - IS217.P12



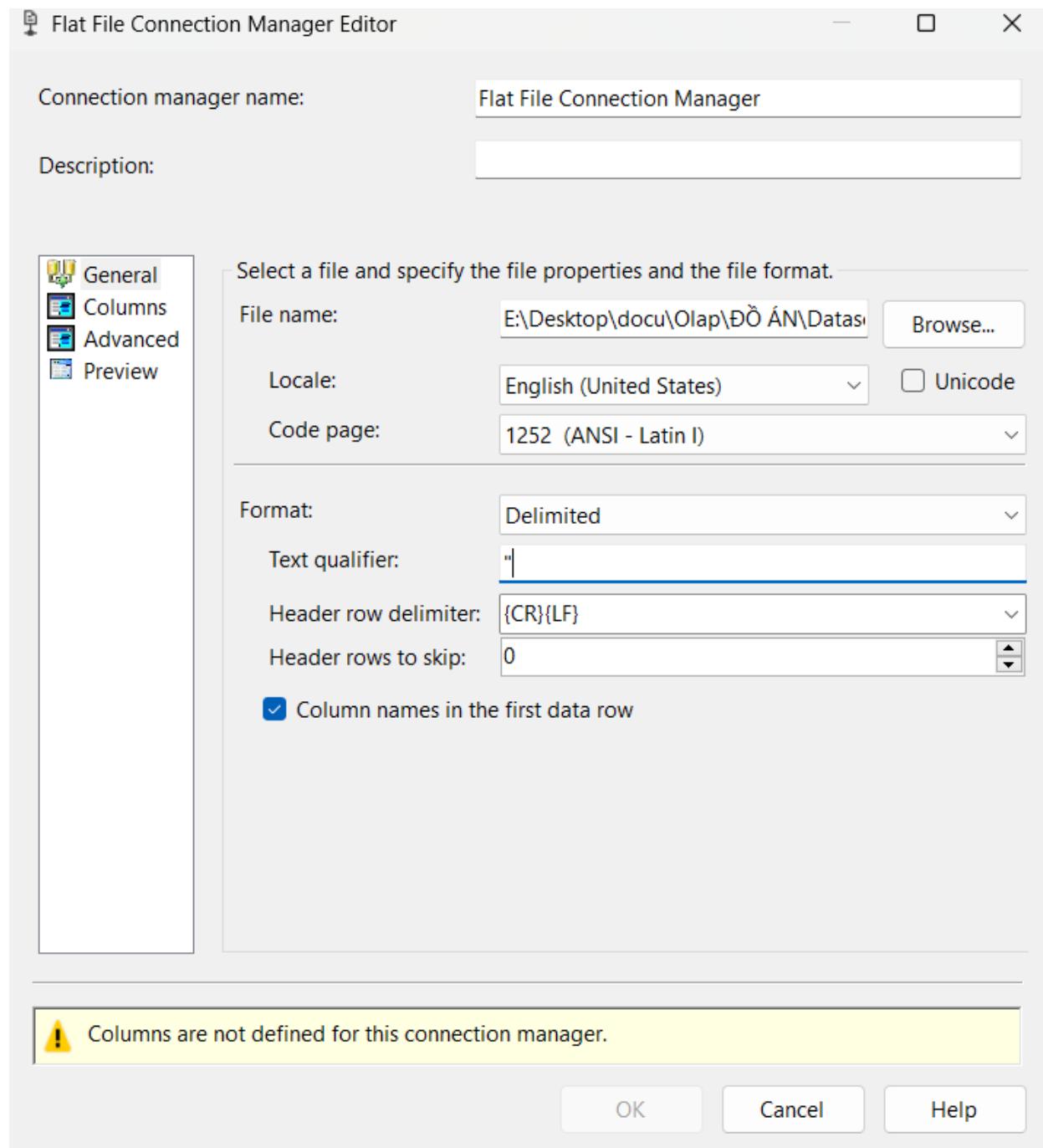
Hình 2.13 Mở Browse để chọn Dataset

Bước 3: Chọn file dữ liệu .csv và chọn Open.



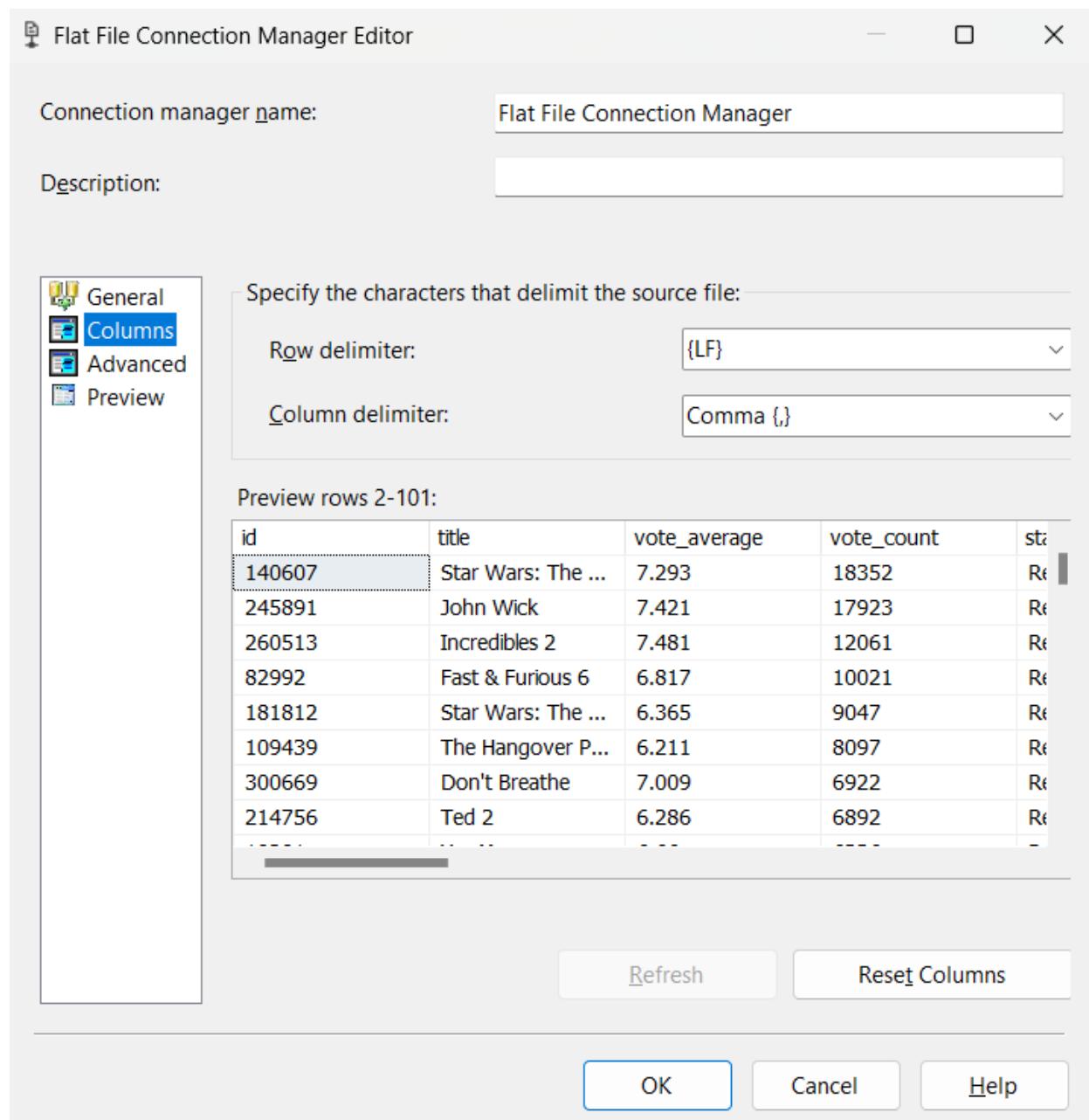
Hình 2.14 Chọn file dữ liệu .csv

Bước 4: Thêm dấu “ vào ô *Text qualifier* để tránh hiện tượng các dữ liệu ở cột này bị lêch sang cột khác.



Hình 2.15 Kiểm tra Text qualifier

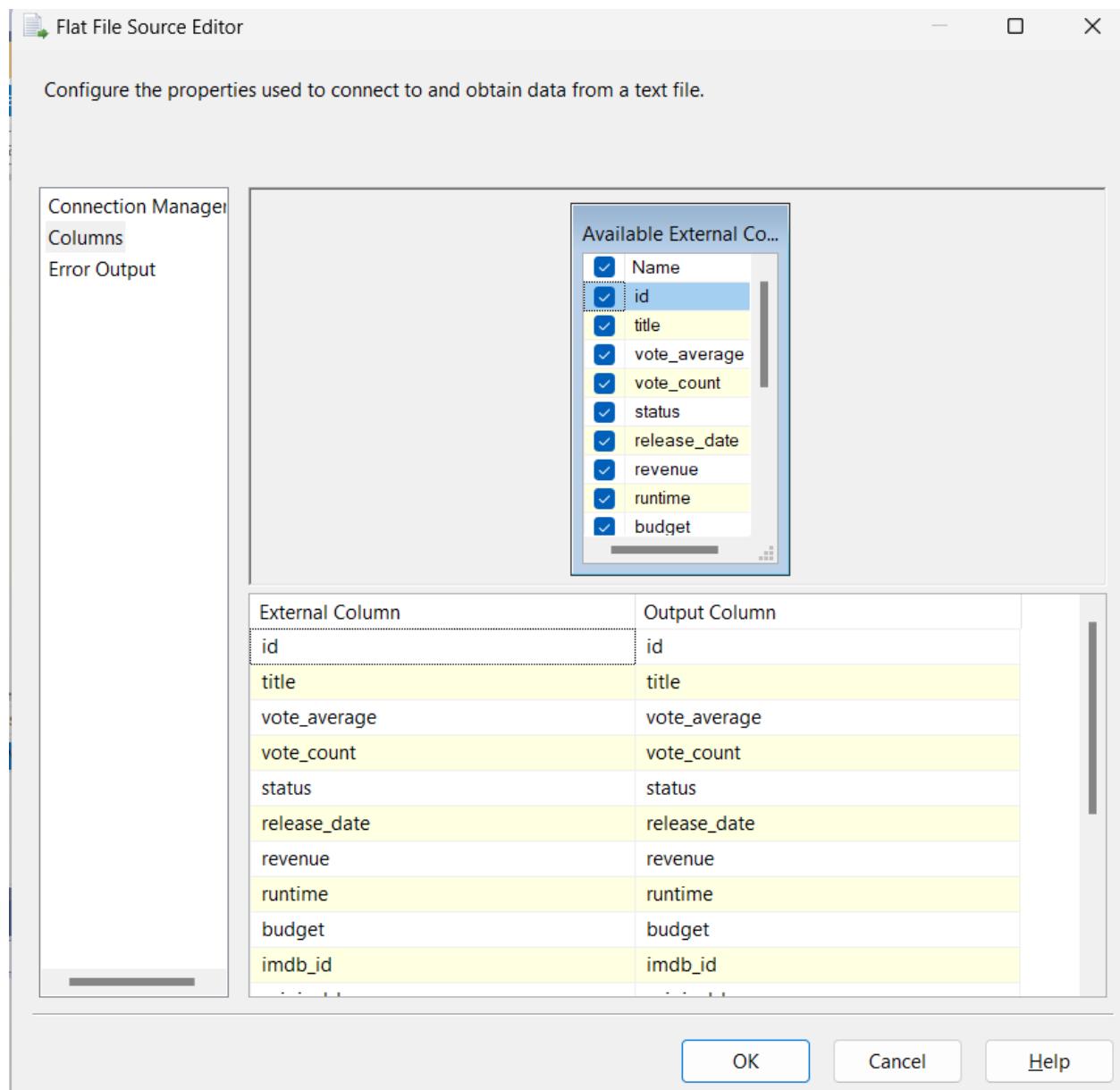
Bước 5: Kiểm tra các cột dữ liệu ở *Columns*.



Hình 2.16 Kiểm tra các cột dữ liệu

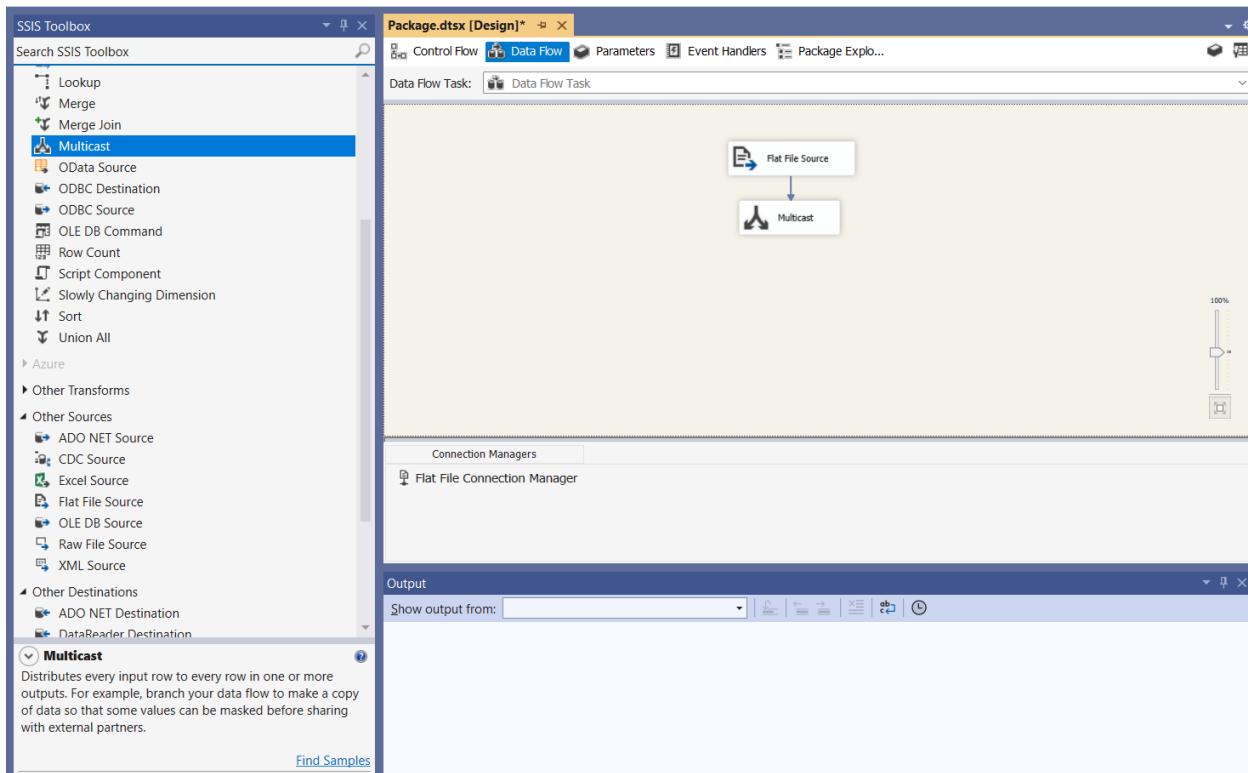
Bước 5: Click chọn **OK** và kiểm tra lần nữa các cột dữ liệu ở dạng danh sách. Nhấn **OK** lần nữa để tiến hành hoàn tất quá trình load dữ liệu và Flat File Source.

Kho dữ liệu và OLAP - IS217.P12



Hình 2.17 Nhấn OK để hoàn tất quá trình tải dữ liệu gốc lên

Bước 6: Tạo Multicast để phân tán dữ liệu từ Flat File Source đến các Dimension. Tiếp theo kết nối Flat File Source và Multicast.

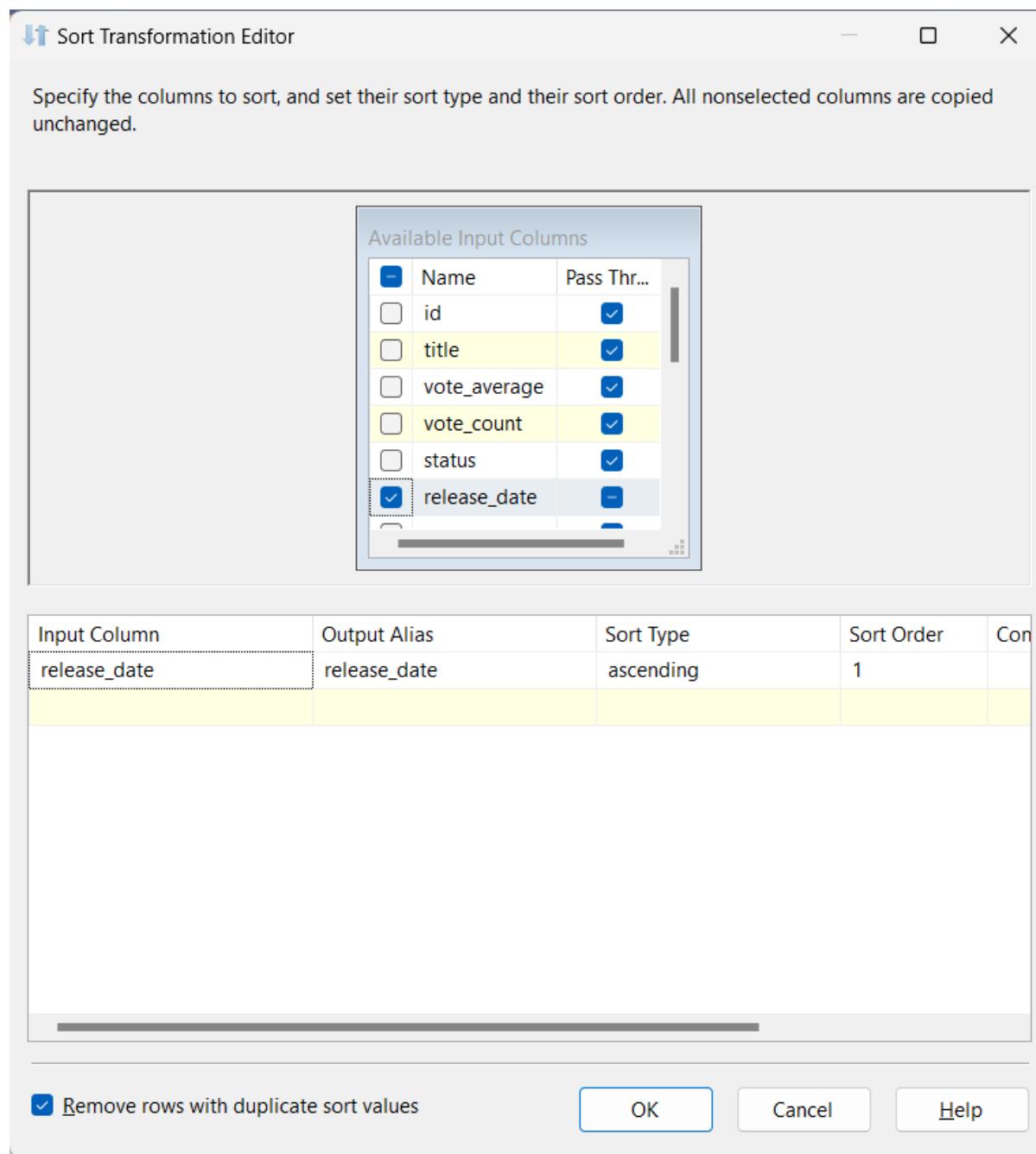


Hình 2.18 Tạo Multicast để phân tán dữ liệu từ Flat File Source đến các Dimension

2.4.1 Bảng Dim Date

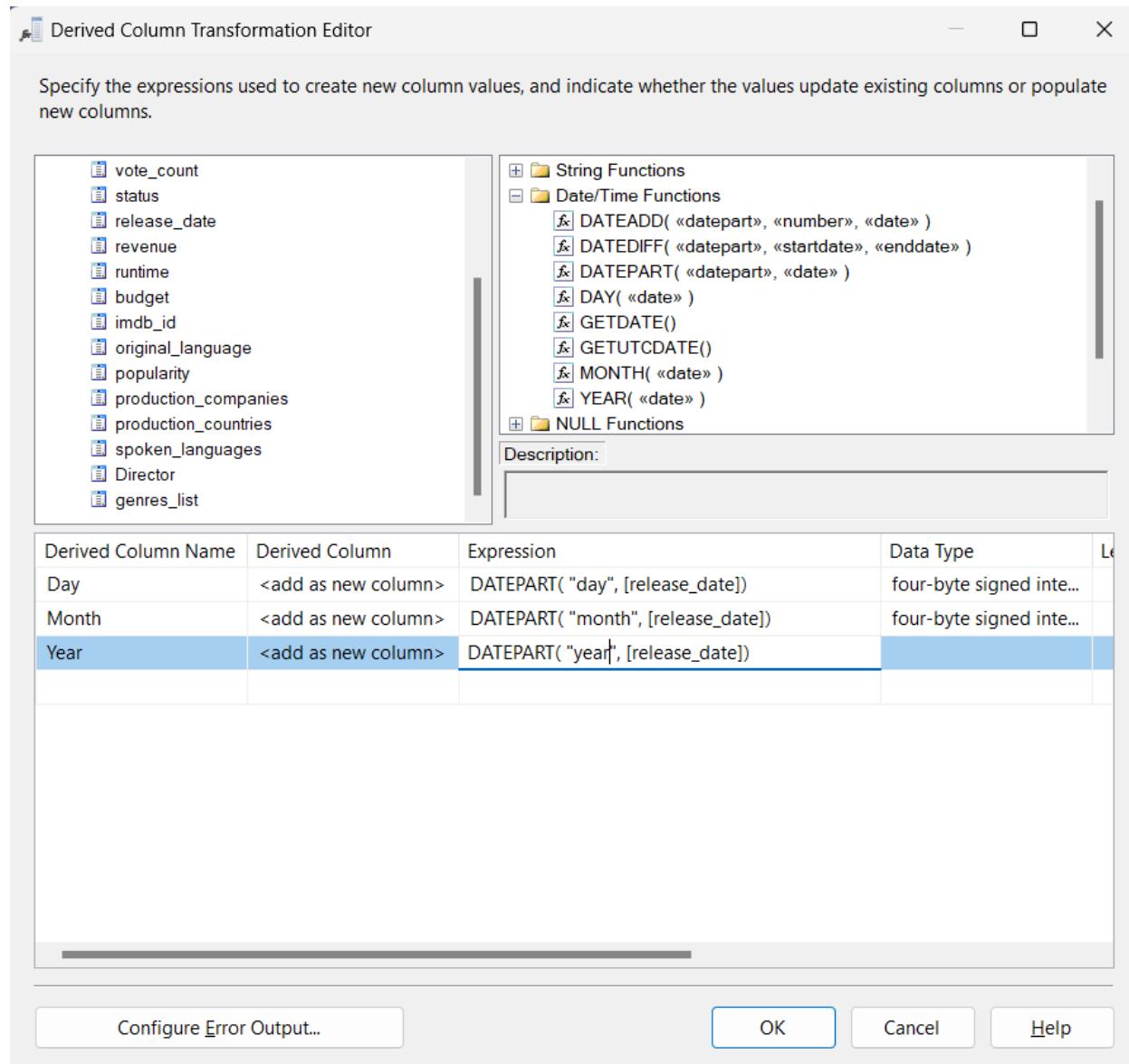
Bước 1: Tạo mới một Sort có tên là Sort_Dim_Date để lấy ra các cột dữ liệu cần thiết cho Dim_Date. Nhấn chuột phải và nhấp **Edit** để chọn **release_date** làm cột dữ liệu cho **Sort_Dim_Date**

Tick chọn **Remove rows with duplicate sort values** xoá các dòng dữ liệu trùng nhau, sau đó chọn **OK**.



Hình 2.19 Chọn release_date Làm cột dữ liệu cho Sort_Dim_Date

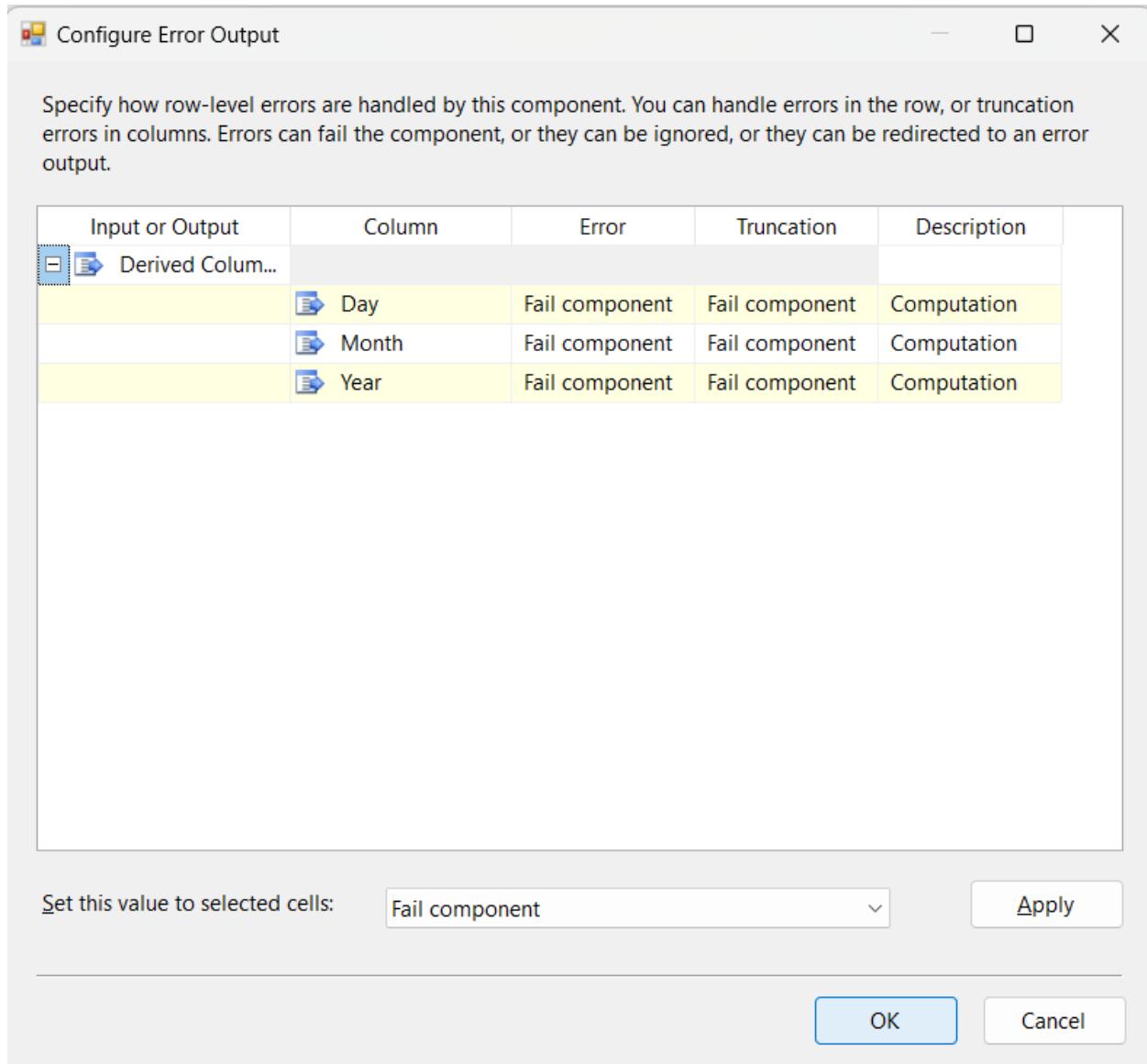
Bước 2: Thêm thành phần **Derived Column** và chọn **Edit** để chia cột dữ liệu. Ta thấy từ cột **release_date** được chia thành 3 cột tương ứng với lược đồ dữ liệu bảng Dim_Date.



Hình 2.20 Chọn vào edit Derived Column để chia cột release_date

Bước 3: Mở *Configure Error Output* để kiểm tra việc chia cột. Nhấn **OK** để hoàn tất quá trình chia cột dữ liệu release_date.

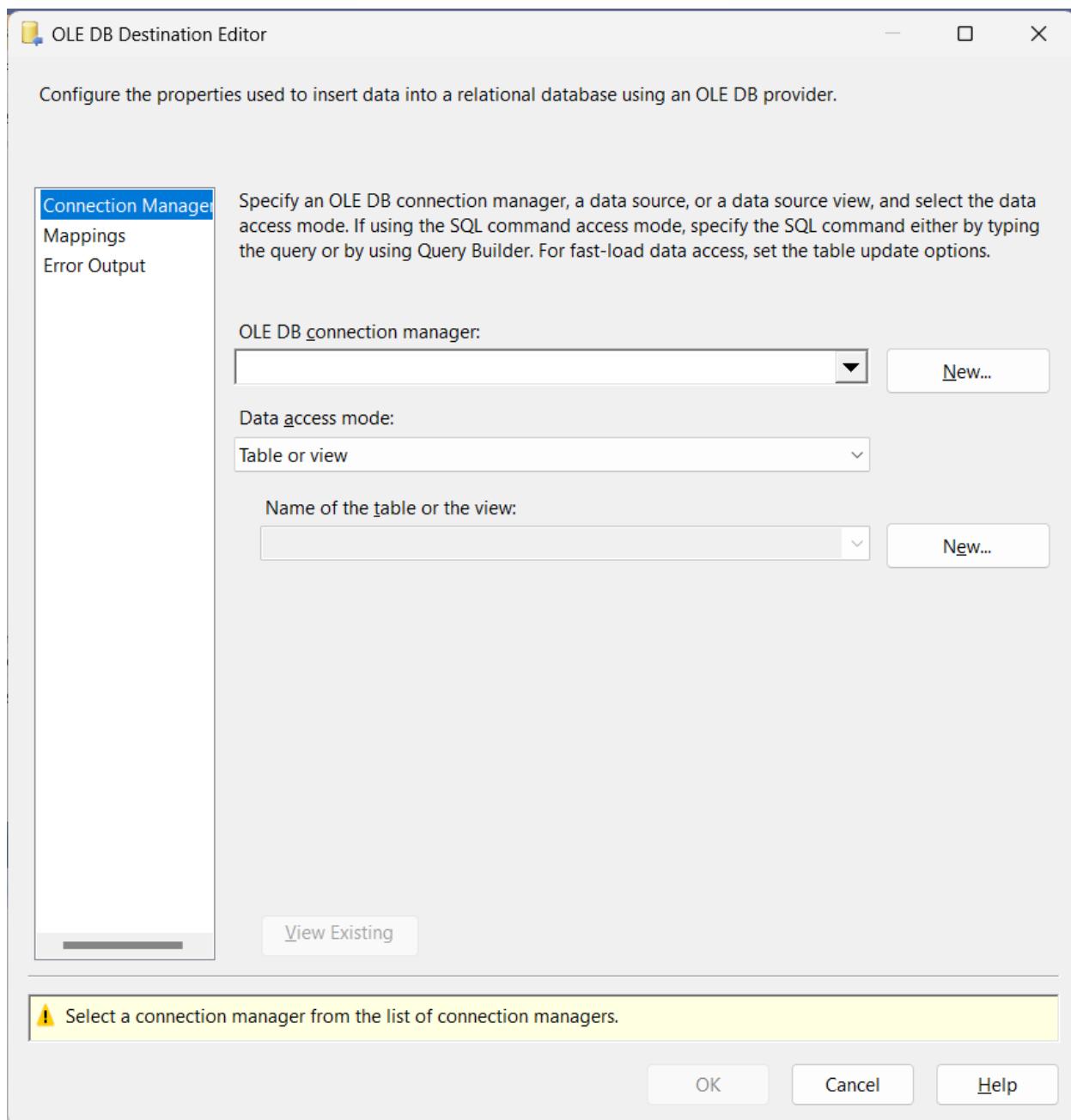
Kho dữ liệu và OLAP - IS217.P12



Hình 2.21 Kiểm tra Configure Error Output

Bước 4: Tạo Dim_Date từ một **OLE DB Destination**. Double click vào OLE DB Destination này để tạo một connection mới đến MS SQL Server.

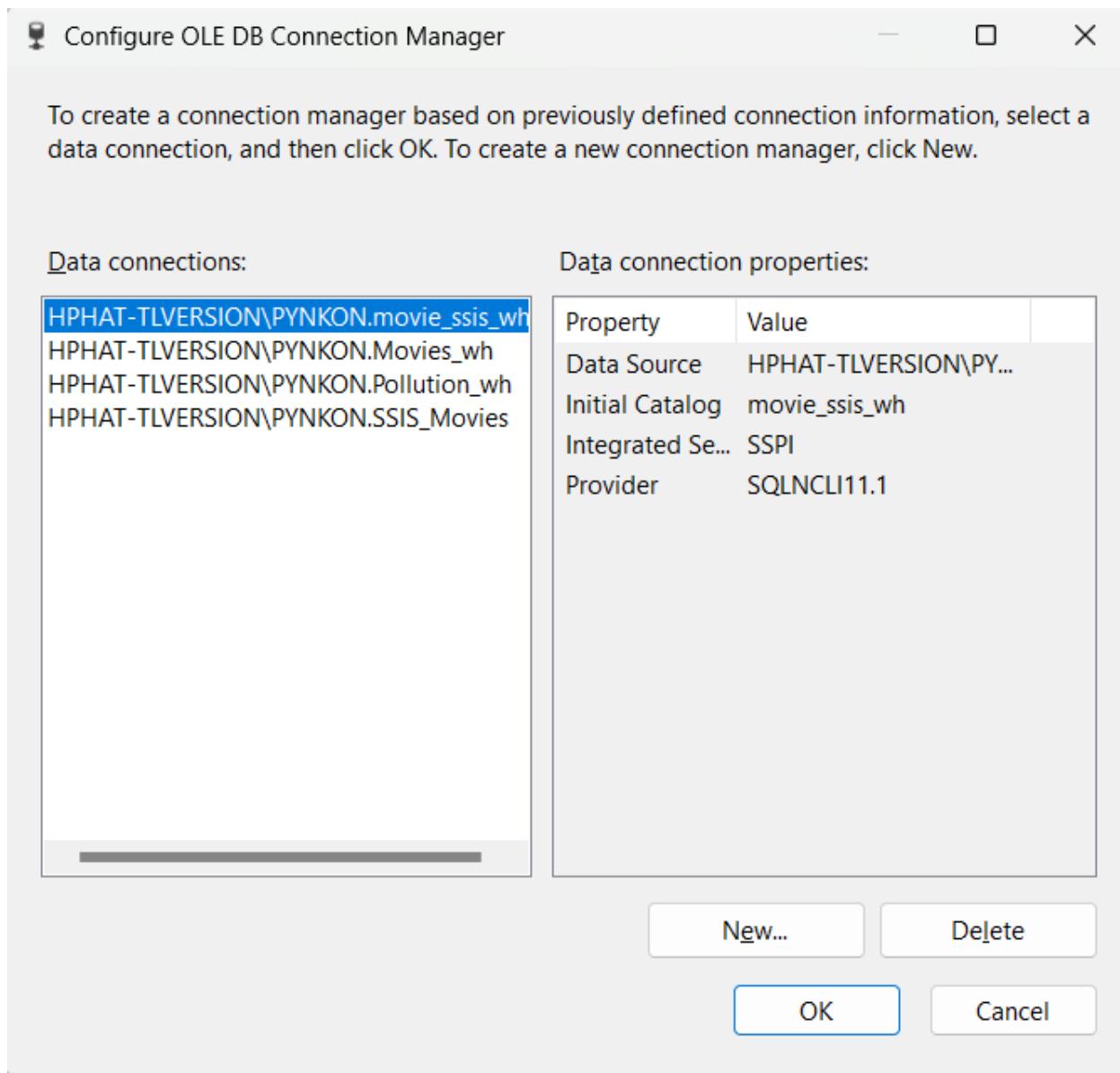
Kho dữ liệu và OLAP - IS217.P12



Hình 2.22.1 Tạo mới OLE DB Connection Manager

Kho dữ liệu và OLAP - IS217.P12

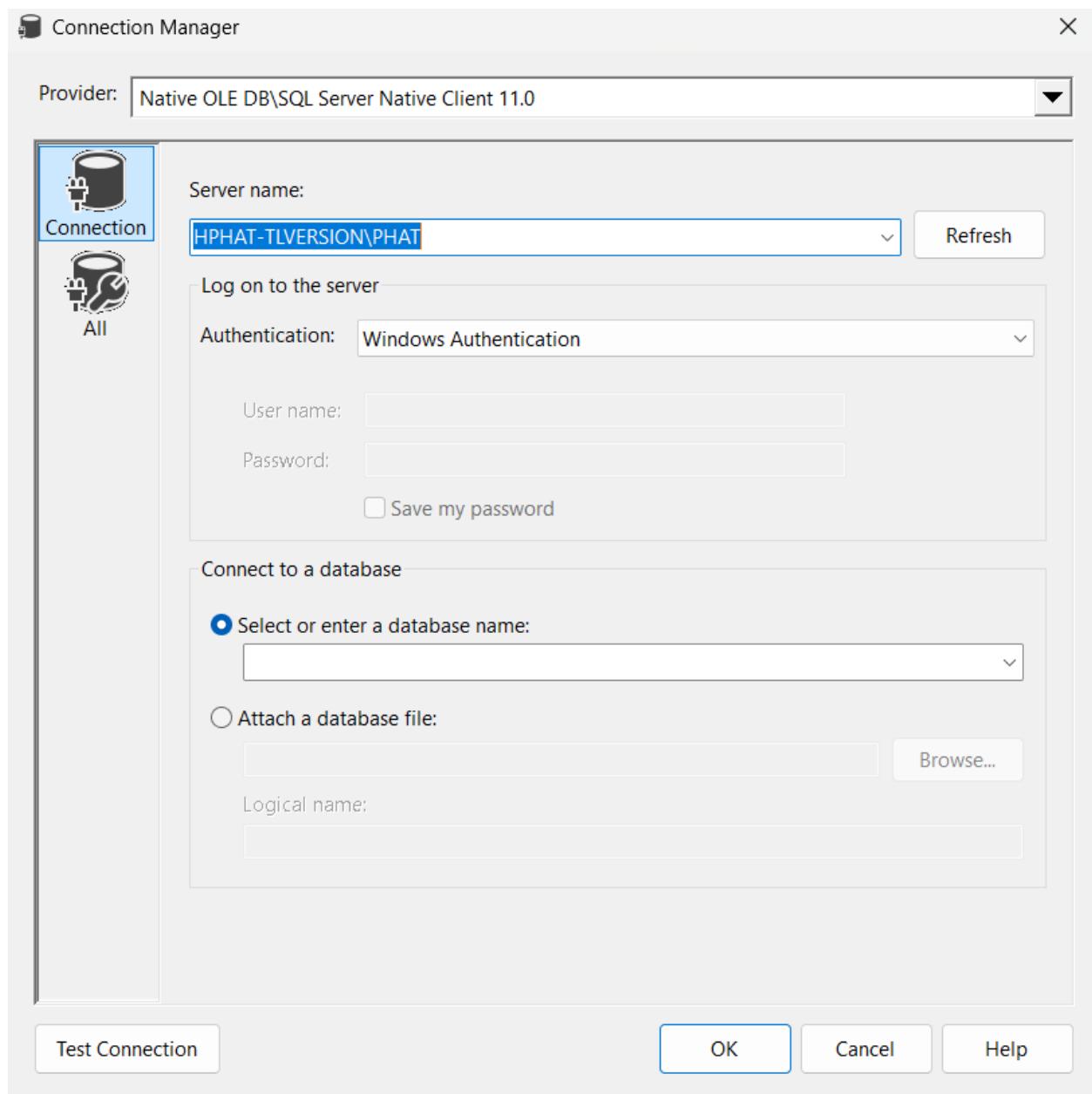
Tiếp tục chọn **New...** để tạo một connection mới:



Hình 2.22.2 Tạo mới OLE DB Connection Manager

Bước 5: Chọn tên **server name** trùng với server name MS SQL Server để ta có thể kết nối đến datawarehouse **Movie_wh**. Kết nối đến server bằng tài khoản window mặc định (**Windows Authentication**)

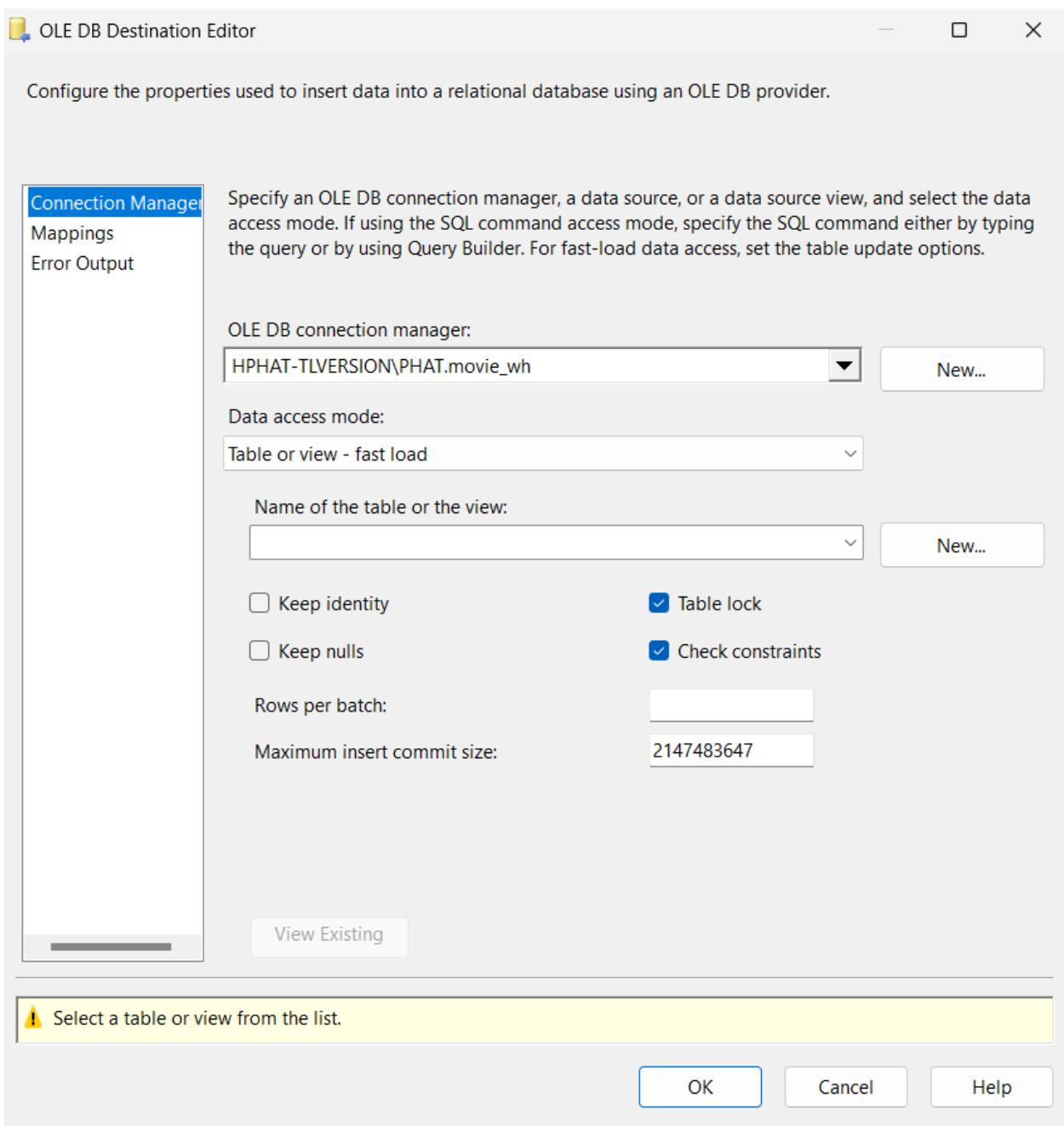
Kho dữ liệu và OLAP - IS217.P12



Hình 2.23 Kết nối đến MS SQL Server

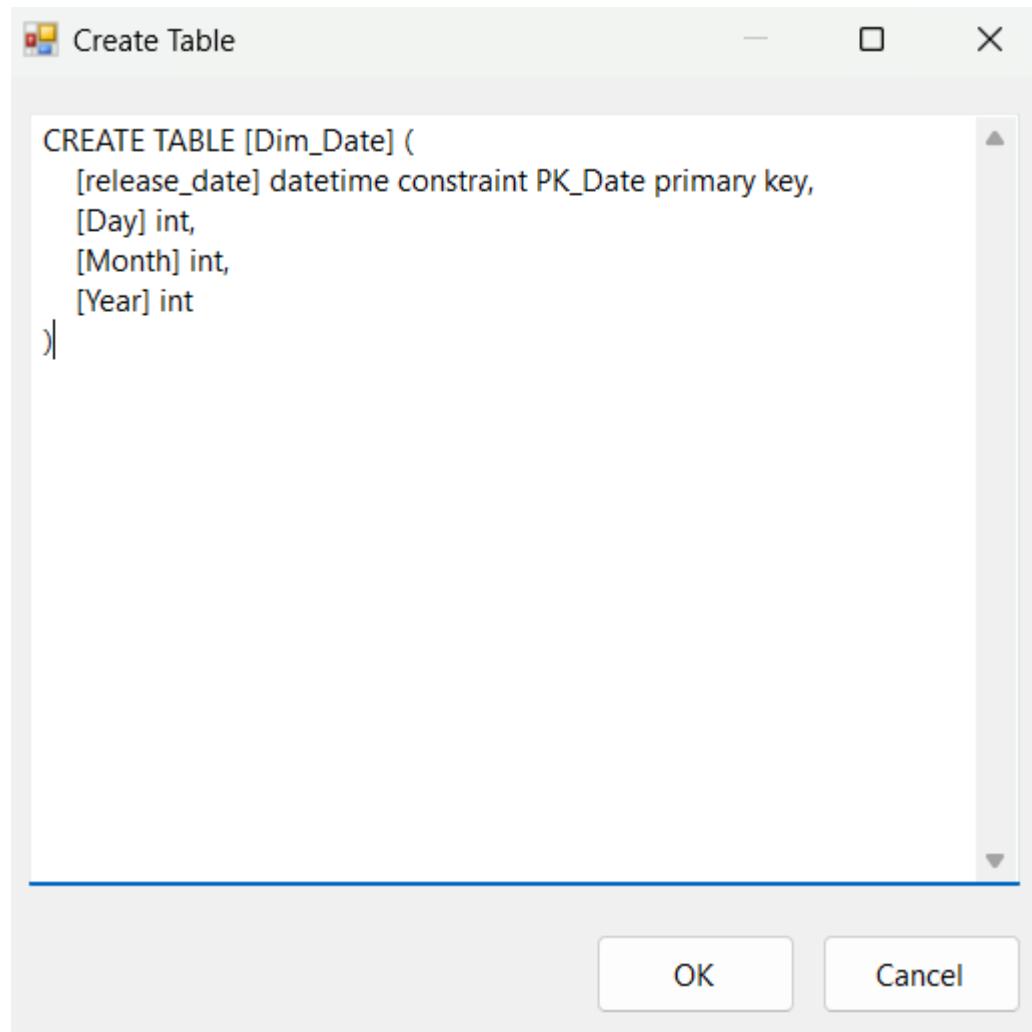
Bước 6: Chọn connection vừa tạo đến MS SQL Server và nhấn **OK**.

Kho dữ liệu và OLAP - IS217.P12



Hình 2.24 Kết nối với connection vừa tạo đến MS SQL Server

Bước 7: Chọn New... để tạo bảng Dim Date



Hình 2.25 Tạo bảng Dim Date

Nội dung câu lệnh SQL tạo bảng Dim_Date như sau:

```
CREATE TABLE [Dim_Date] (
    [release_date] datetime constraint PK_Date primary key,
    [Day] int,
    [Month] int,
    [Year] int
)
```

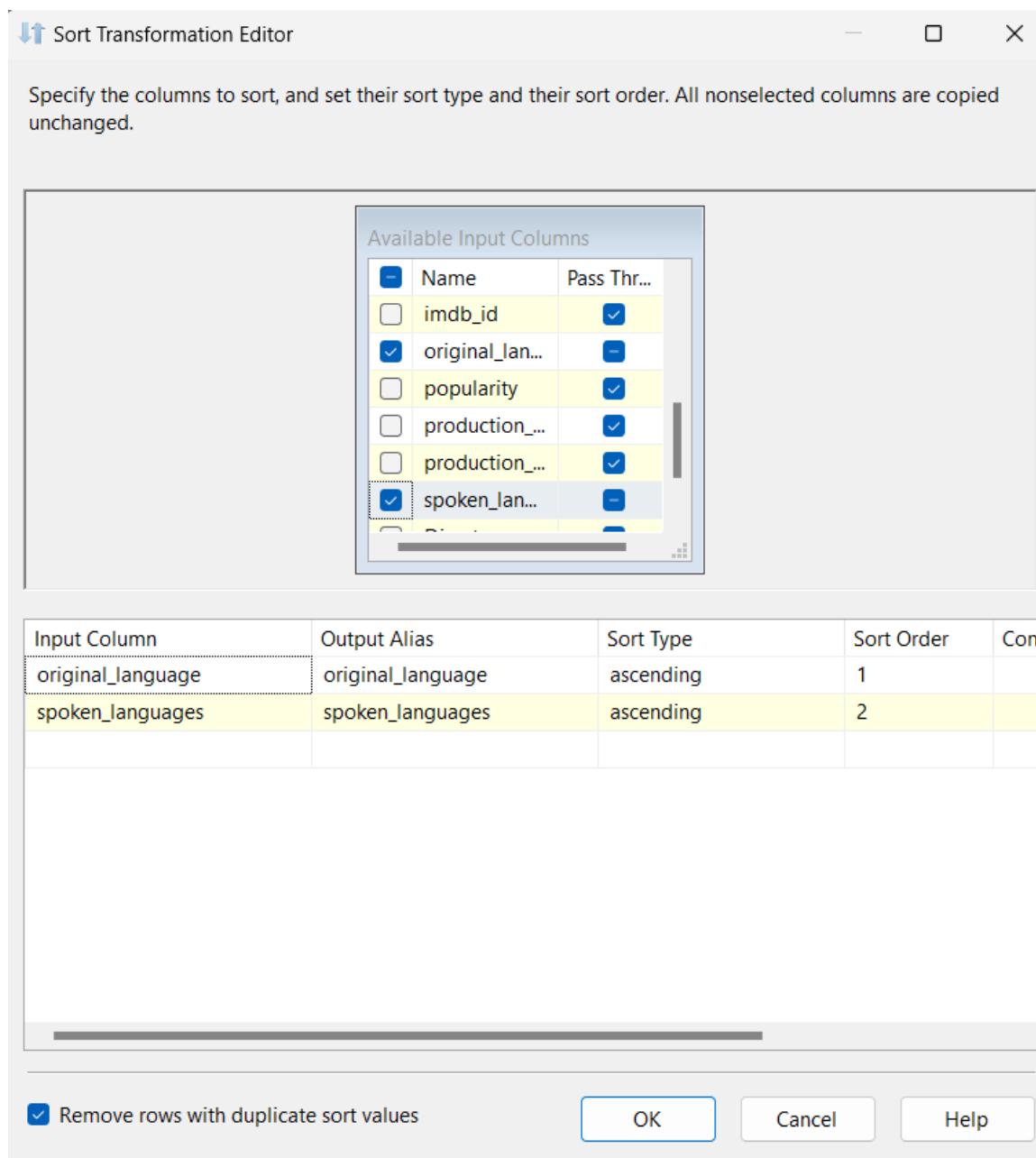
Bước 8: Tiếp đến ta cần chọn mục **Mappings** để xem xét việc ánh xạ các cột dữ liệu
Chọn **OK** để hoàn tất thiết lập.

2.4.2 Bảng Dim Language

Bước 1: Chọn một Sort và đổi tên để tạo ra Sort_Dim_Language

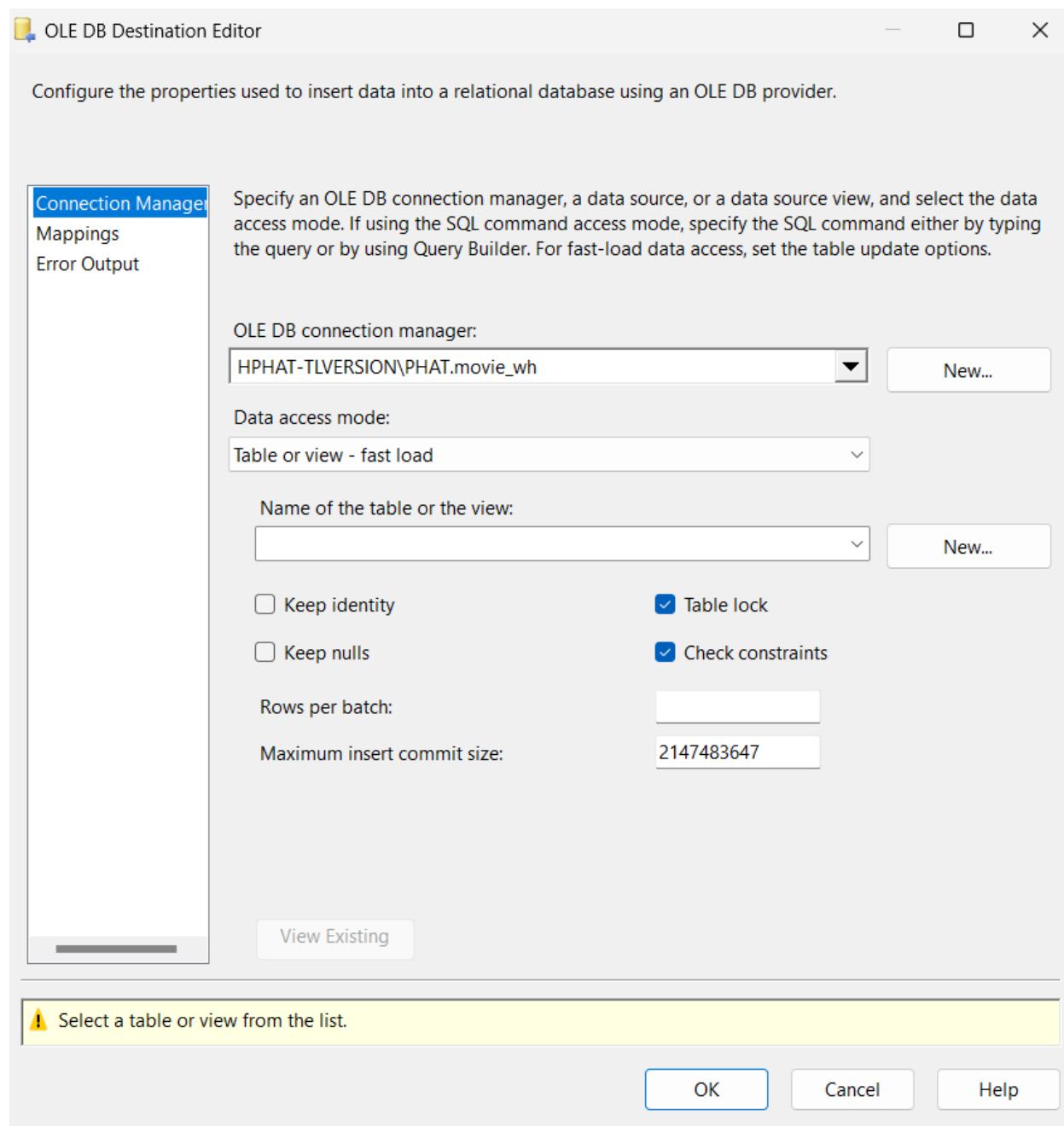
Bước 2: Click chuột và Sort_Dim_Language, chọn Edit: chọn cột spoken_language và original_language để đổ dữ liệu vào Sort_Dim_Language

Tick chọn Remove rows with duplicate sort values để xoá đi các dòng dữ liệu trùng nhau và sau đó chọn OK.



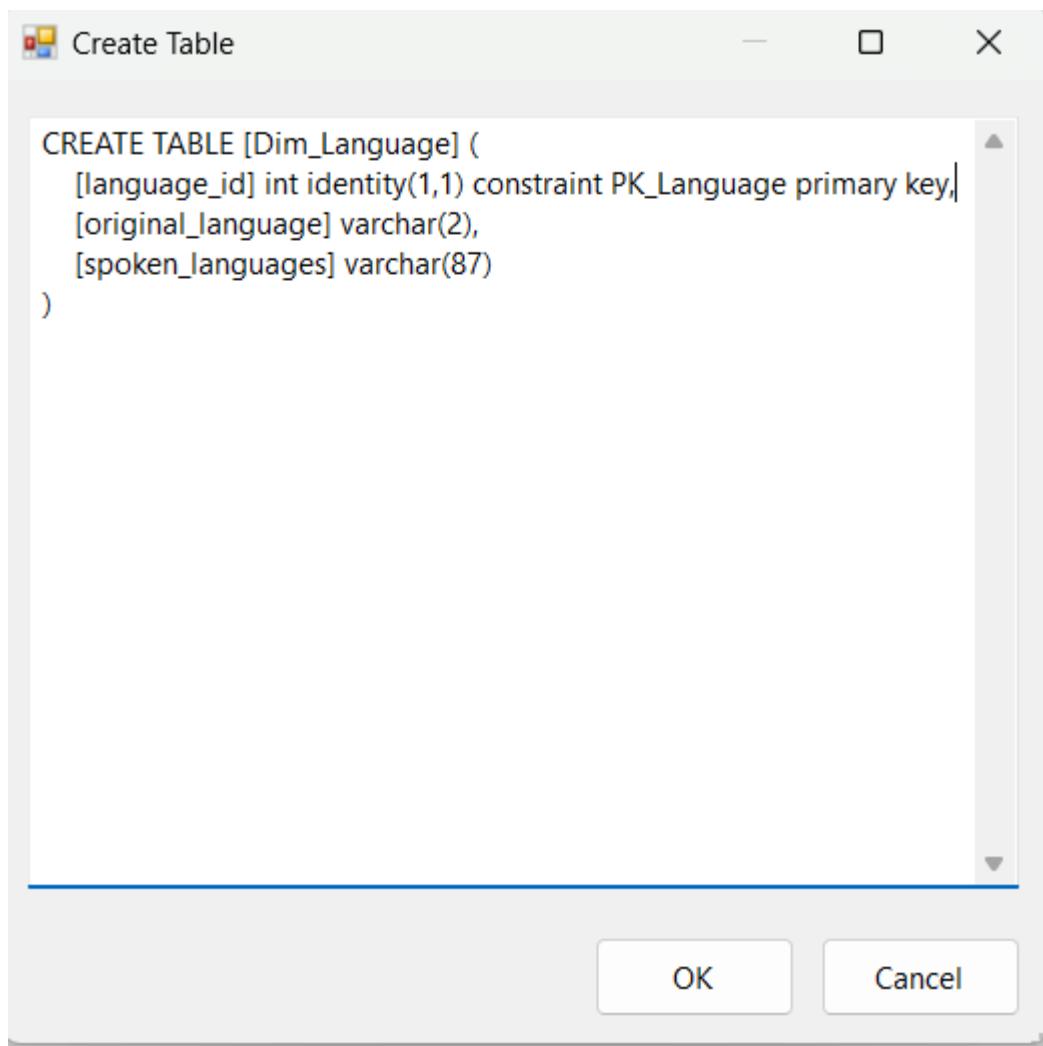
Hình 2.26 chọn cột spoken_language và original_language để đổ dữ liệu vào Sort_Dim_Language

Bước 3: Tạo mới một **OLE DB Destination** để đỡ dữ liệu gốc sau khi đã được xử lý vào trong kho dữ liệu **Movie_wh**



Hình 2.27 Kết nối đến database Movie_wh

Bước 4: Connection đến kho dữ liệu đã được tạo khi tạo **Dim_Time**, vì vậy ta chỉ cần chọn **New...** để tạo bảng **Dim_Language**



Hình 2.28 Tạo bảng Dim Language

Nội dung câu lệnh SQL tạo bảng Dim_Language như sau:

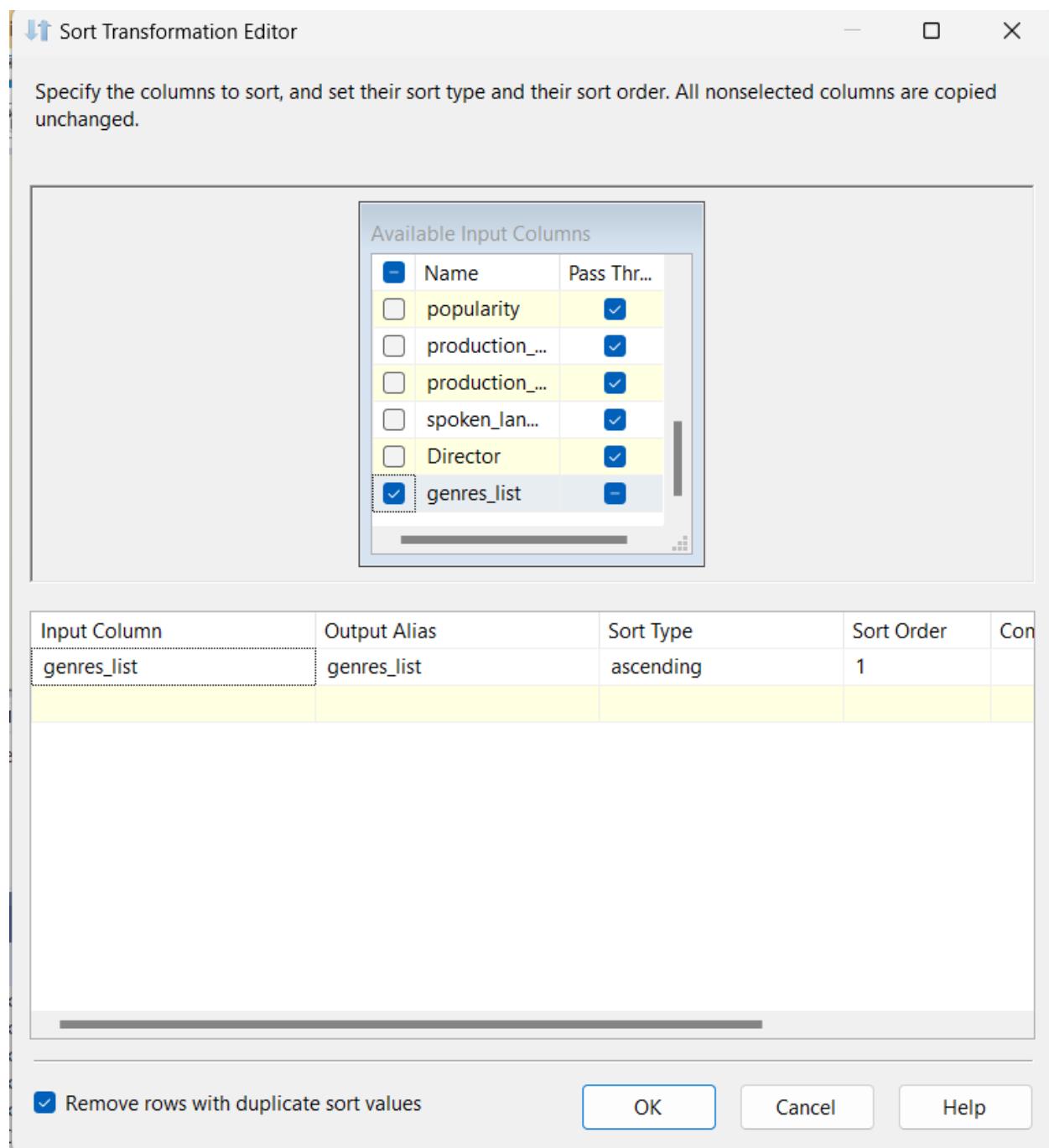
```
CREATE TABLE [Dim_Language] (
    [language_id] int identity(1,1) constraint PK_Language primary key,
    [original_language] varchar(2),
    [spoken_languages] varchar(87)
)
```

Bước 5: Tiếp tục ta cần chọn mục **Mappings** để xem xét việc ánh xạ các cột dữ liệu
Chọn **OK** để hoàn tất thiết lập.

2.4.3 Bảng Dim Genres List

Bước 1: Chọn một Sort để tạo ra Sort_Dim_Genres_List cho Dim_Genres_List.

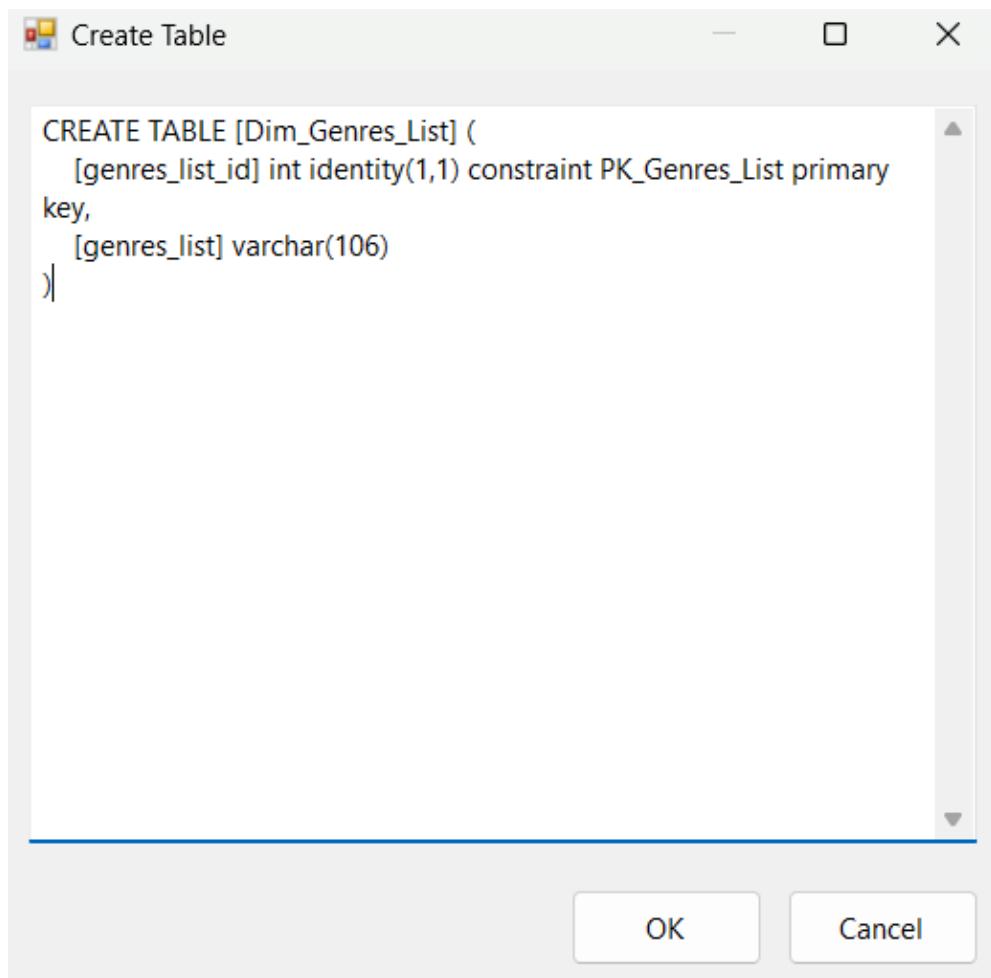
Bước 2: Click chuột phải vào Sort_Dim_Genres_List, chọn **Edit**: chọn cột Genres List làm cột đỡ dữ liệu vào Sort_Dim_Genres_List.



Hình 2.29 Chọn cột Genres List làm cột đỡ dữ liệu vào Sort_Dim_Genres_List

Bước 3: Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong bảng **Dim_Genres_List**.

Bước 4: Connection đến kho dữ liệu đã được tạo khi tạo **Dim_Date**, vì vậy ta chỉ cần chọn New... để tạo bảng **Dim_Genres_List**



Hình 2.30 Tạo bảng Dim Genres List

Nội dung câu lệnh SQL tạo bảng Dim_Genres_List như sau:

```
CREATE TABLE [Dim_Genres_List] (
    [genres_list_id] int identity(1,1) constraint PK_Genres_List
    primary key,
    [genres_list] varchar(106) )
```

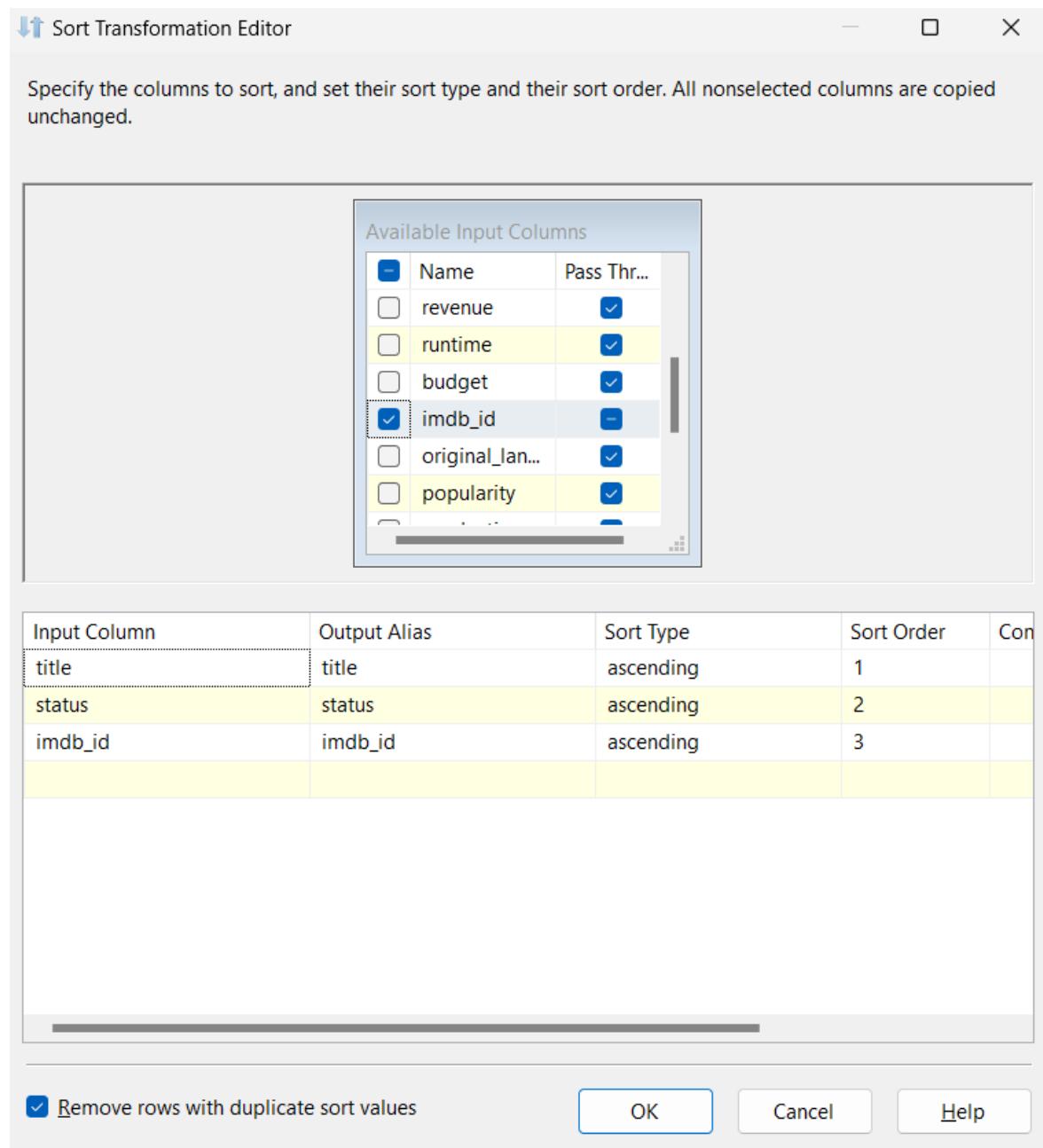
Bước 5: Tiếp đến ta cần chọn mục **Mappings** để xem xét việc ánh xạ các cột dữ liệu Chọn **OK** để hoàn tất thiết lập.

2.4.4 Bảng Dim Movie

Bước 1: Chọn một Sort để tạo ra Sort_Dim_Movie

Bước 2: Click chuột phải vào **Sort_Dim_Movie**, chọn **Edit**: lần lượt chọn các cột **title, status, imdb_id, overview** làm các cột để đổ dữ liệu vào **Sort_Dim_Movie**.

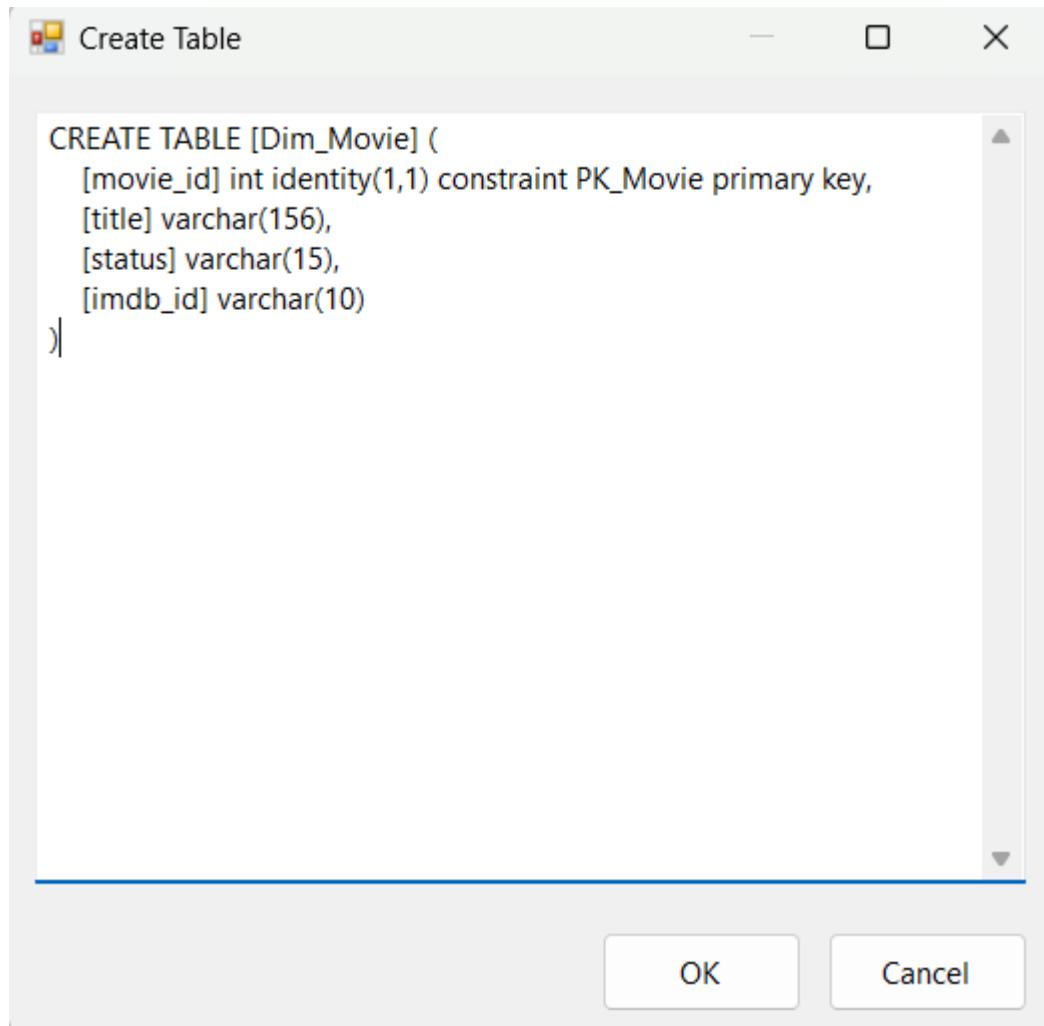
Tick chọn **Remove row with duplicate sort values** xoá đi các dòng dữ liệu trùng nhau và sau đó chọn **OK**.



Hình 2.31 Chọn các cột title, status và imdb_id, làm các cột để đổ dữ liệu vào Sort_Dim_Movie

Bước 3: Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong bảng **Dim_Movie**

Bước 4: . Connection đến kho dữ liệu đã được tạo khi tạo **Dim_Date**, vì vậy ta chỉ cần Chọn **New...** để tạo bảng **Dim_Movie**

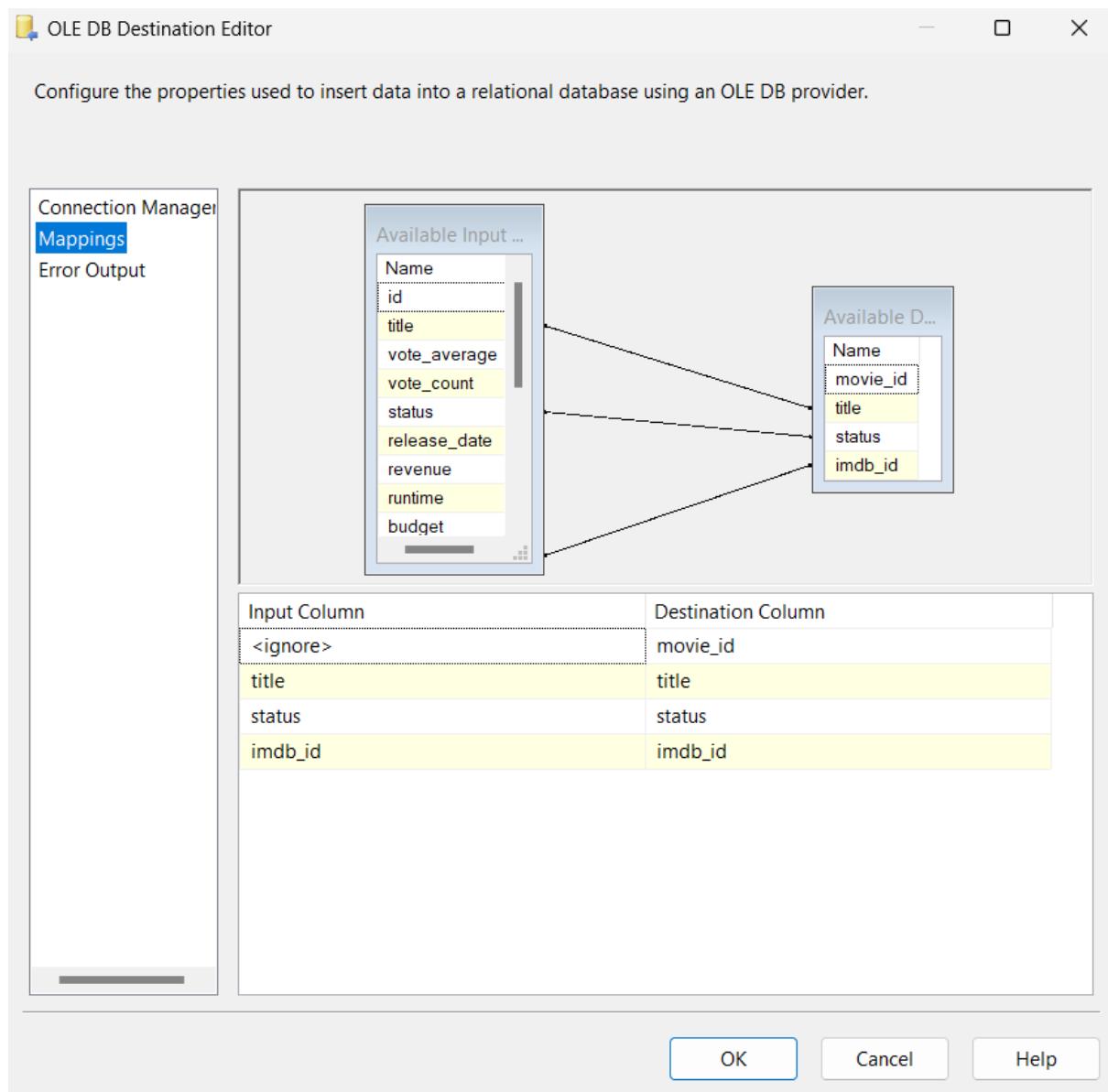


Hình 2.32 Tạo bảng Dim Movie

Nội dung câu lệnh SQL tạo bảng Dim_Movie như sau:

```
CREATE TABLE [Dim_Movie] (
    [movie_id] int identity(1,1) constraint PK_Movie primary key,
    [title] varchar(156),
    [status] varchar(15),
    [imdb_id] varchar(10), )
```

Bước 5: Tiếp đến ta cần chọn mục **Mappings** để xem xét việc ánh xạ các cột dữ liệu



Hình 2.33 Kiểm tra Mappings

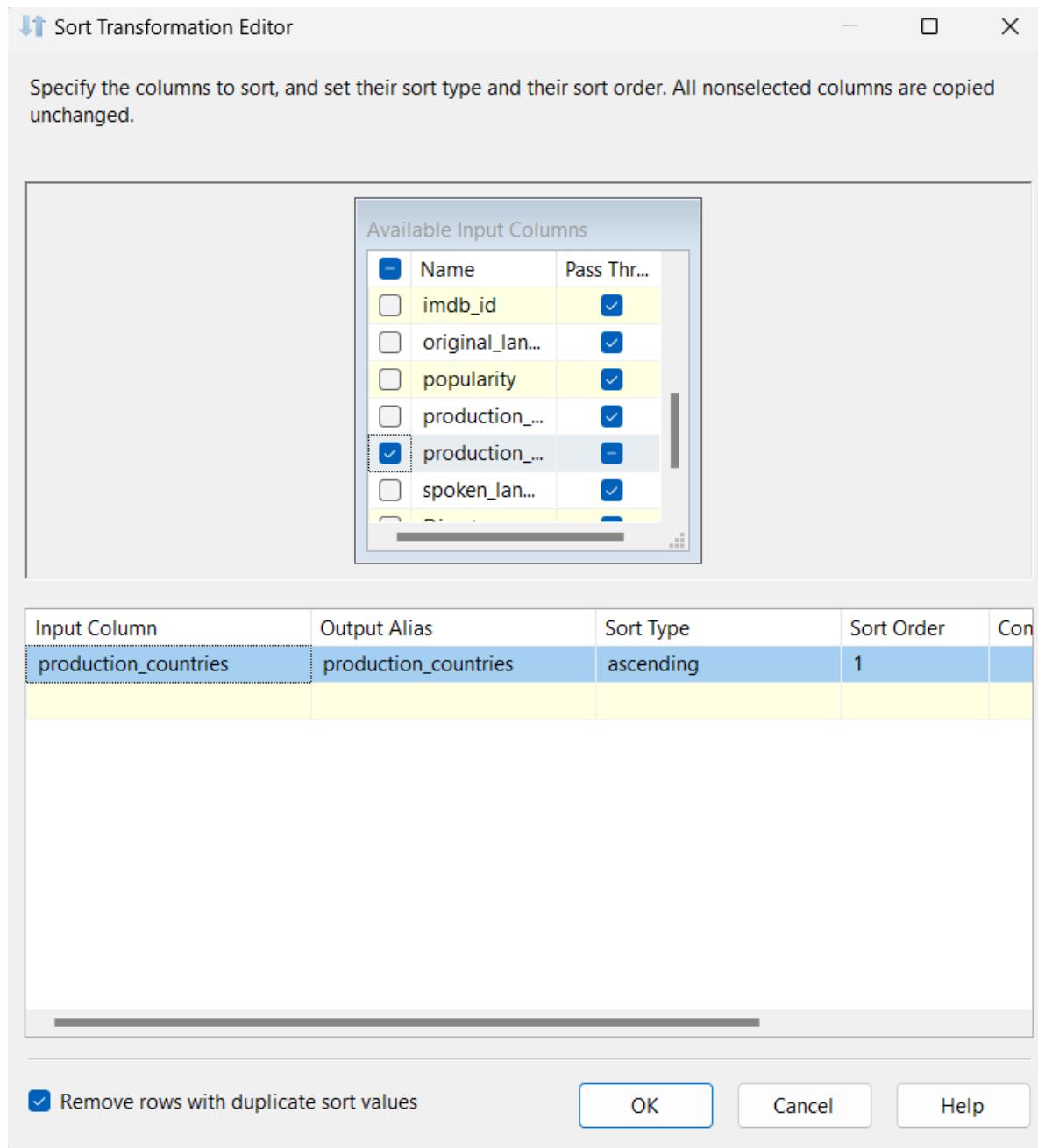
Chọn **OK** để hoàn tất thiết lập.

2.4.5 Bảng Dim Country

Bước 1: Chọn một Sort để tạo ra Sort_Dim_Country.

Bước 2: Click chuột phải và Sort_Dim_Country, chọn Edit: là cột production_countries làm cột dữ liệu để đổ dữ liệu vào Sort_Dim_Country.

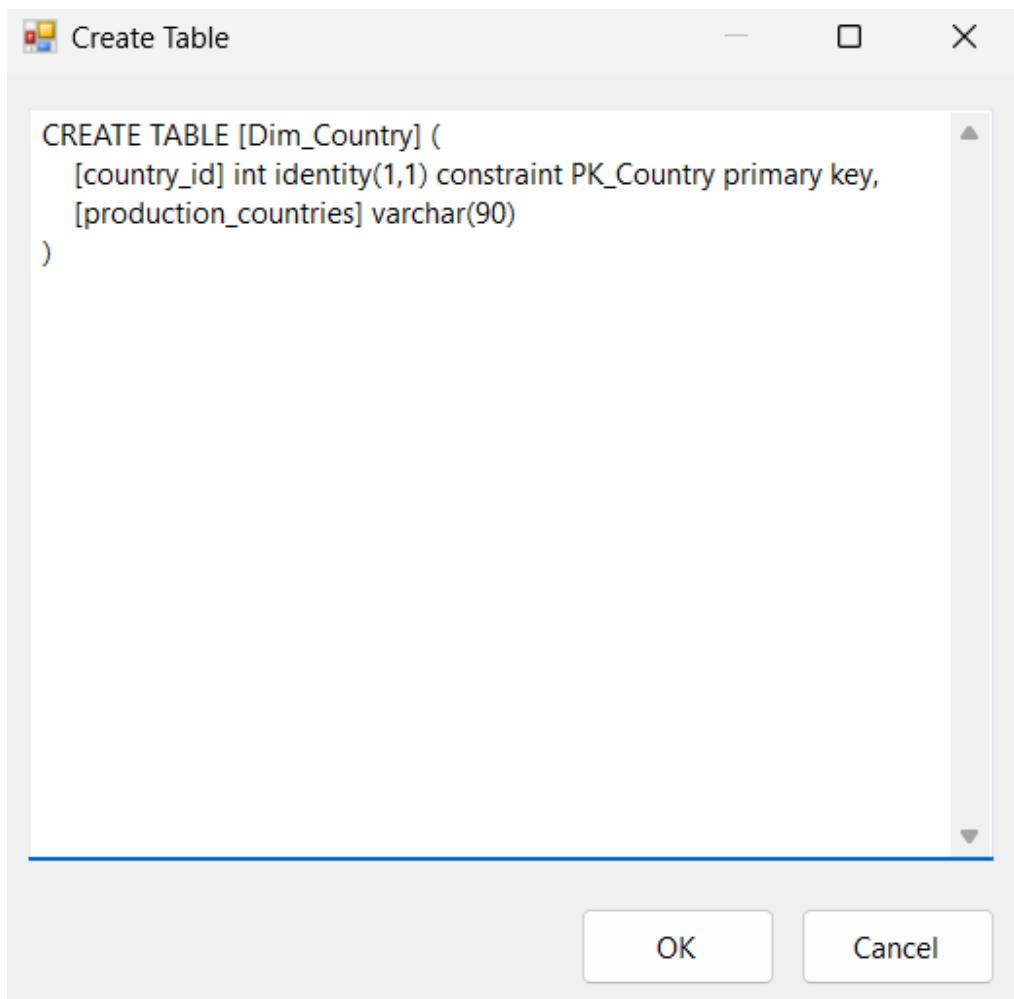
Tick chọn Remove rows with duplicate sort values xoá đi các dòng dữ liệu trùng nhau và sau đó chọn OK.



Hình 2.34 Chọn cột production_countries làm cột dữ liệu để đổ dữ liệu vào Sort_Dim_Country

Bước 3: Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong bảng Dim_Country.

Bước 4: Chọn **New...** để tạo bảng Dim_Country



Hình 2.35 Tạo bảng Dim Country

Nội dung câu lệnh SQL tạo bảng Dim_Country như sau:

```
CREATE TABLE [Dim_Country] (
    [country_id] int identity(1,1) constraint PK_Country primary key,
    [production_countries] varchar(90) )
```

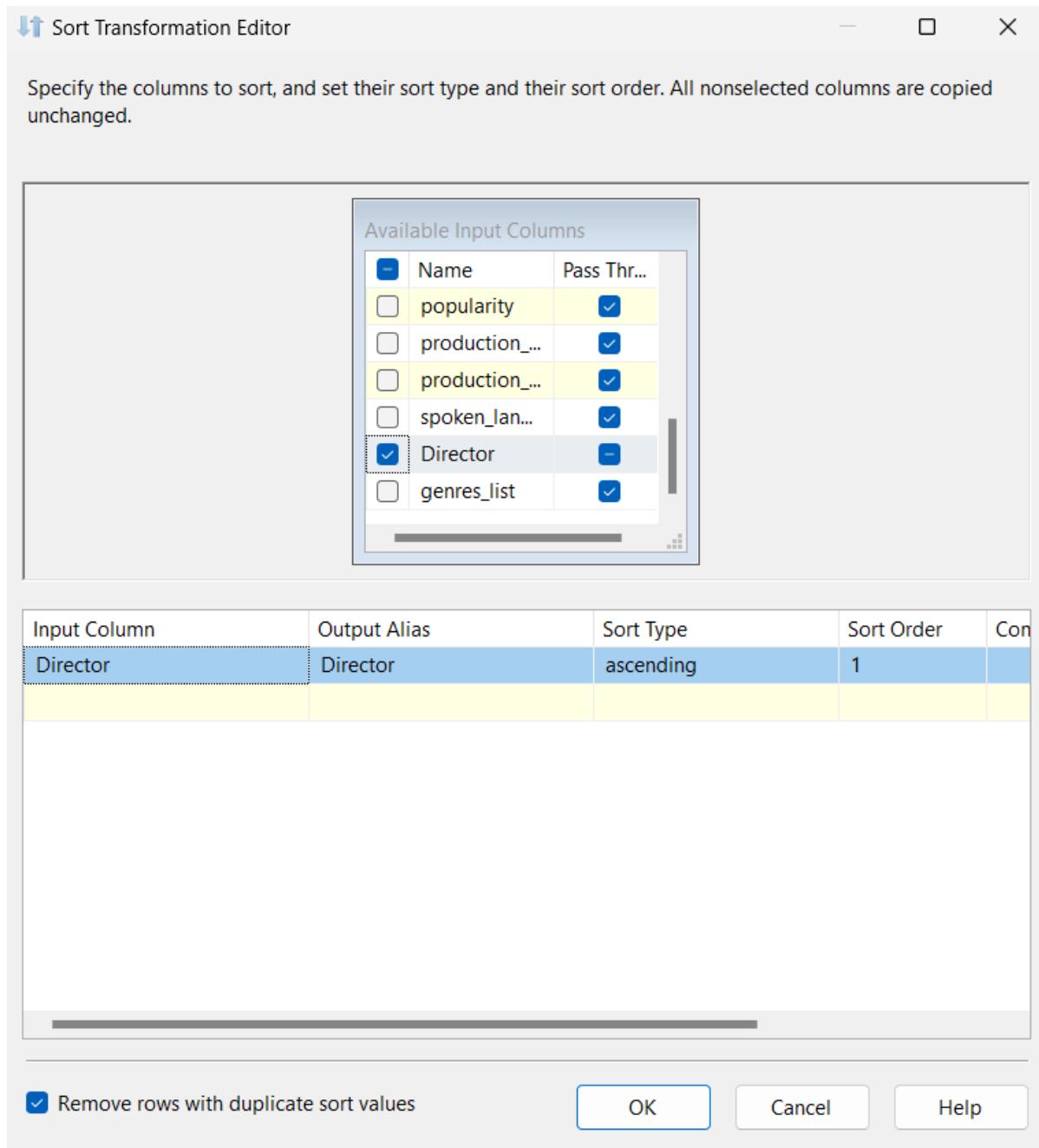
Bước 5: Tiếp đến ta cần chọn mục **Mappings** để xem xét việc ánh xạ các cột dữ liệu
Chọn **OK** để hoàn tất thiết lập.

2.4.6 Bảng Dim Director

Bước 1: Chọn một Sort để tạo ra **Sort_Dim_Director**.

Bước 2: Click chuột phải và Sort_Dim_Director, chọn **Edit**: là cột **director** làm cột dữ liệu để đổ dữ liệu vào **Sort_Dim_Director**.

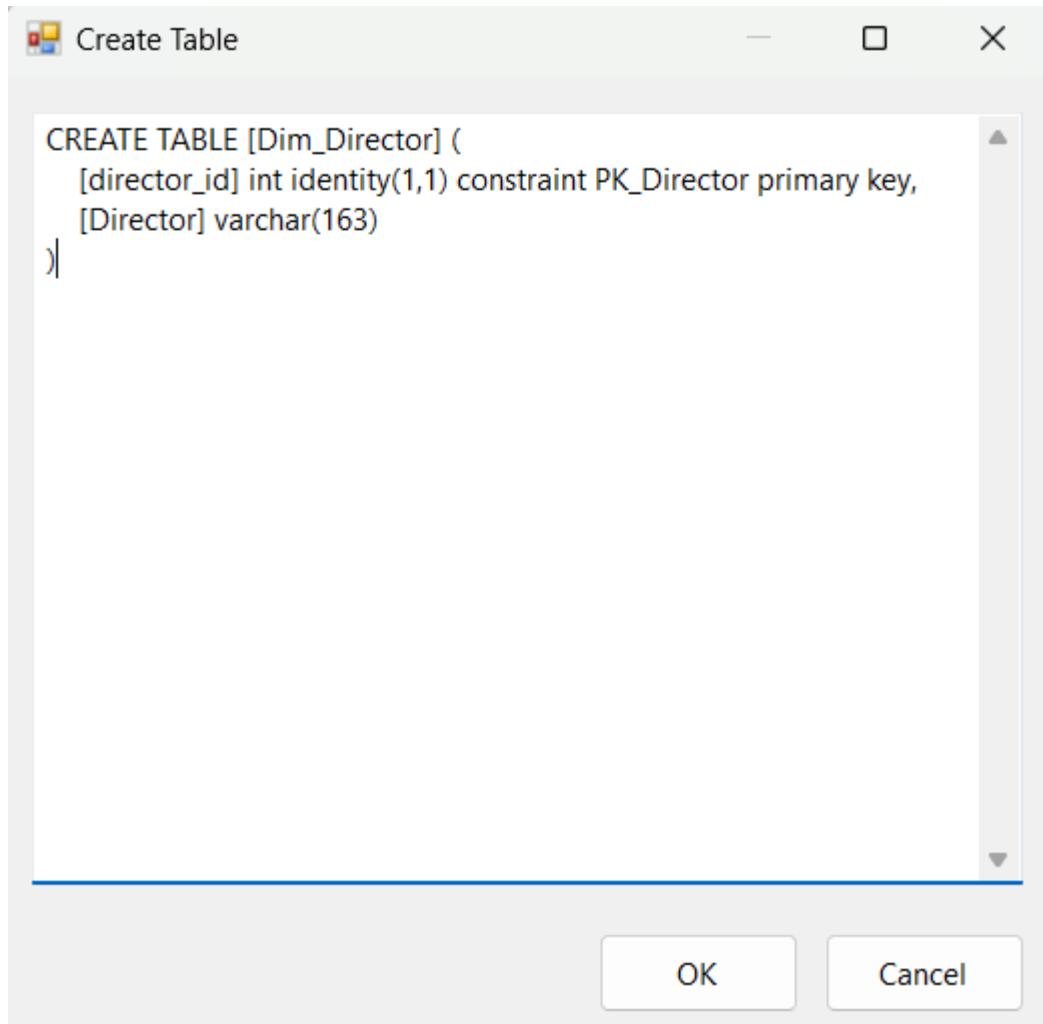
Tick chọn **Remove rows with duplicate sort values** xoá đi các dòng dữ liệu trùng nhau và sau đó chọn **OK**.



Hình 2.36 Chọn cột Director làm cột dữ liệu để đổ dữ liệu vào Sort_Dim_Director

Bước 3: Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong bảng Dim_Director.

Bước 4: Chọn **New...** để tạo bảng Dim_Director



Hình 2.37 Tạo bảng Dim Director

Nội dung câu lệnh SQL tạo bảng Dim_Director như sau:

```
CREATE TABLE [Dim_Director] (
    [director_id] int identity(1,1) constraint PK_Director primary key,
    [Director] varchar(163) )
```

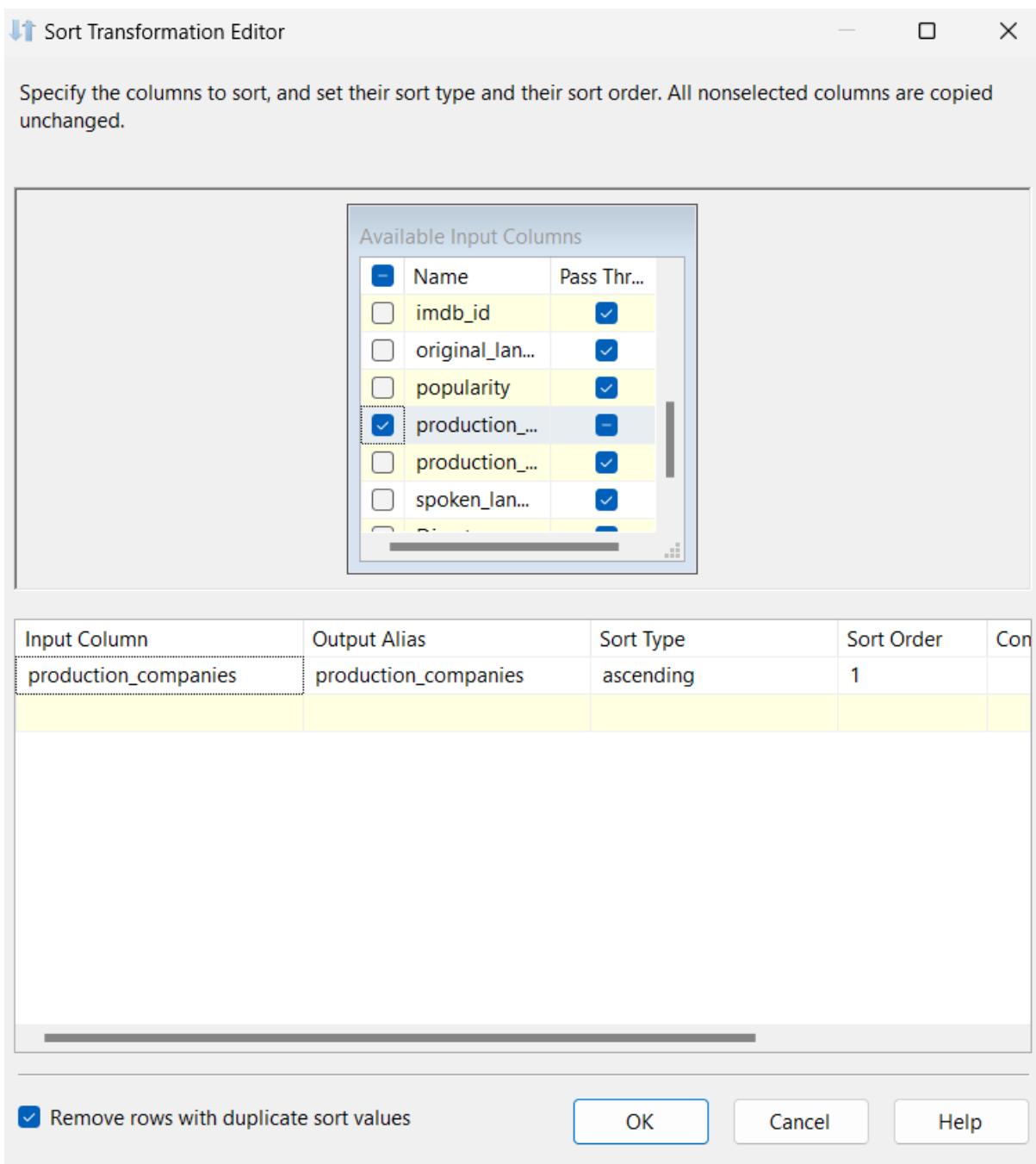
Bước 5: Tiếp đến ta cần chọn mục **Mappings** để xem xét việc ánh xạ các cột dữ liệu
Chọn **OK** để hoàn tất thiết lập.

2.4.7 Bảng Dim Company

Bước 1: Chọn một Sort để tạo ra **Sort_Dim_Company**.

Bước 2: Click chuột phải và **Sort_Dim_Actor**, chọn **Edit** cột **production_companies** làm cột dữ liệu để đổ dữ liệu vào **Sort_Dim_Company**.

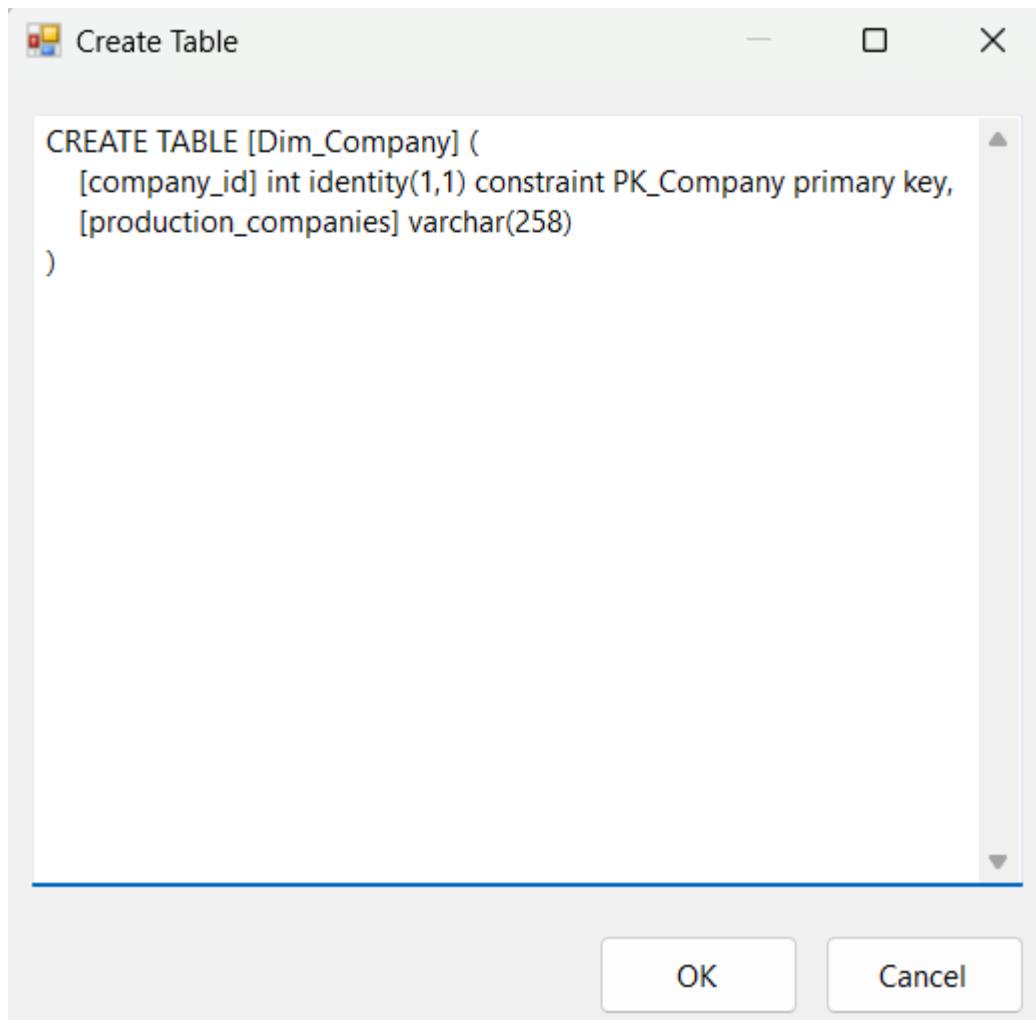
Tick chọn **Remove rows with duplicate sort values** xoá đi các dòng dữ liệu trùng nhau và sau đó chọn **OK**.



Hình 2.38 Chọn cột *production_companies* làm cột dữ liệu để đổ dữ liệu vào *Sort_Dim_Company*

Bước 3: Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong bảng Dim_Company

Bước 4: Chọn **New...** để tạo bảng Dim_Company



Hình 2.39 Tạo bảng Dim Company

Nội dung câu lệnh SQL tạo bảng Dim_Company như sau:

```
CREATE TABLE [Dim_Company] (
    [production_companies_id] int identity(1,1) constraint PK_Company primary key,
    [production_companies] varchar(258)
```

Bước 5: Tiếp đến ta cần chọn mục **Mappings** để xem xét việc ánh xạ các cột dữ liệu

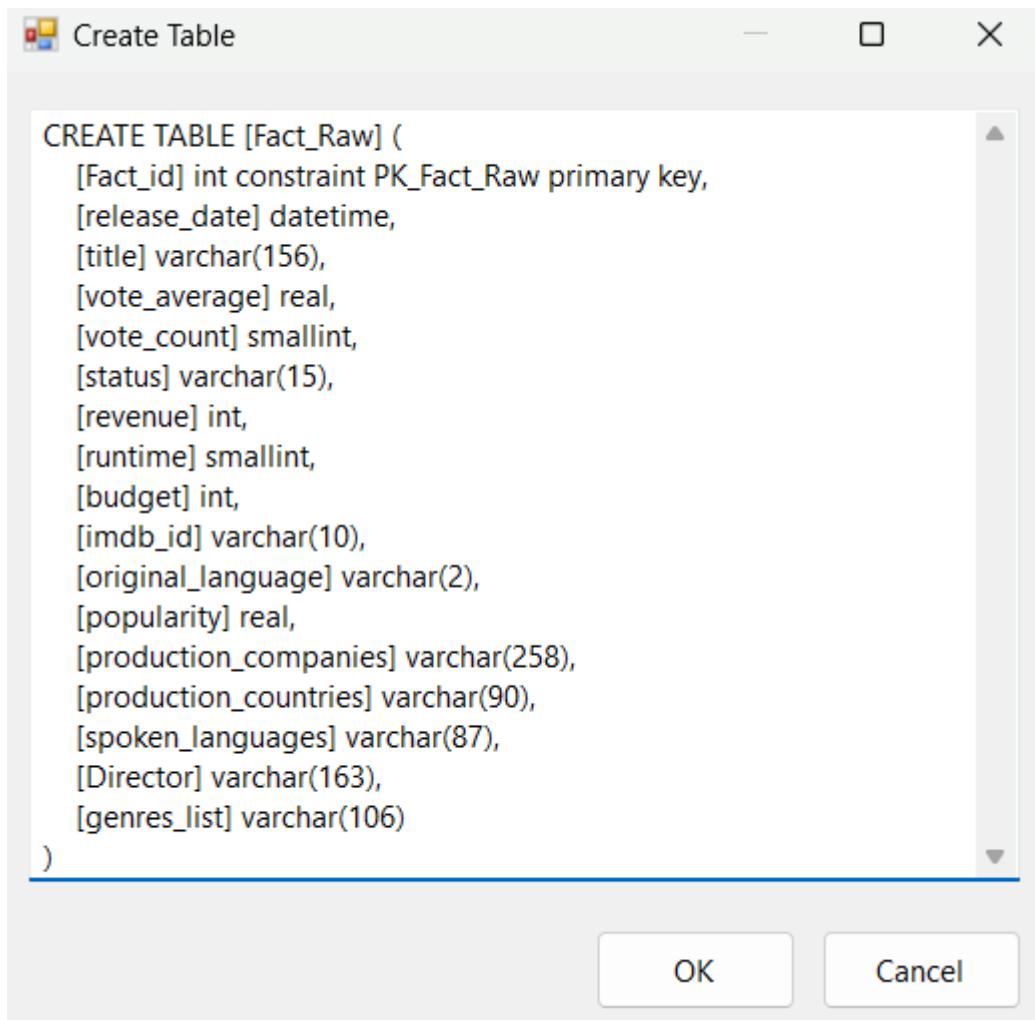
Chọn **OK** để hoàn tất thiết lập.

2.4.8 Tạo bảng Fact Movie

2.4.8.1 Tạo bảng Fact Raw

Bước 1: Tiến hành tạo bảng **Fact** và đặt tên là **Fact_Raw** từ một **OLE DB Destination**

Bước 2: Click chuột phải và chọn **Edit** để tạo bảng **Fact_Raw** có các cột là tất cả các cột từ dữ liệu gốc và chứa tất cả các dòng dữ liệu .



Hình 2.40 Tạo bảng Fact Raw

Bước 3: Tiếp đến ta cần chọn mục **Mappings** để xem xét việc ánh xạ các cột dữ liệu. Tathấy ID của bảng **Fact_Raw** cũng chính là ID của tập dữ liệu nên ta chọn ánh xạ cột id trong Input Column vào cột id của bảng **Fact_Raw**

Cuối cùng nhấn nút **OK** để hoàn tất quá trình tạo bảng.

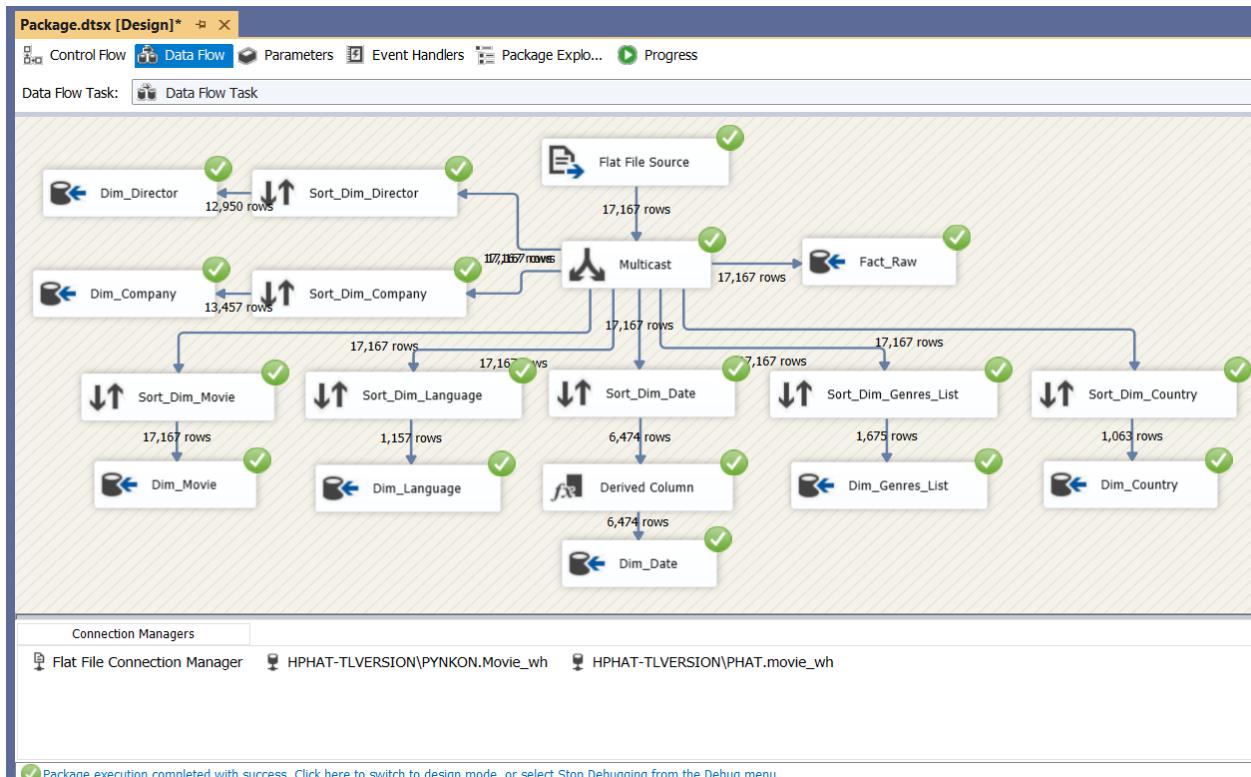
Tiếp theo đây ta sẽ thực hiện quá trình lần lượt loại bỏ các cột dữ liệu trùng của bảng Fact với các Dimension, thực hiện thêm khóa ngoại vào bảng Fact nhằm thu gọn bảng Fact, tối ưu hóa quá trình phân tích dữ liệu.

Nội dung câu lệnh SQL tạo bảng Fact_Raw như sau:

```
CREATE TABLE [Fact_Raw] (
    [Fact_id] int constraint PK_Fact primary key,
    [release_date] datetime,
    [title] varchar(156),
    [vote_average] real,
    [vote_count] smallint,
    [status] varchar(15),
    [revenue] int,
    [runtime] smallint,
    [budget] int,
    [imdb_id] varchar(10),
    [original_language] varchar(2),
    [popularity] real,
    [production_countries] varchar(258),
    [production_countries] varchar(90),
    [spoken_languages] varchar(87),
    [Director] varchar(163),
    [genres_list] varchar(106), )
```

Kho dữ liệu và OLAP - IS217.P12

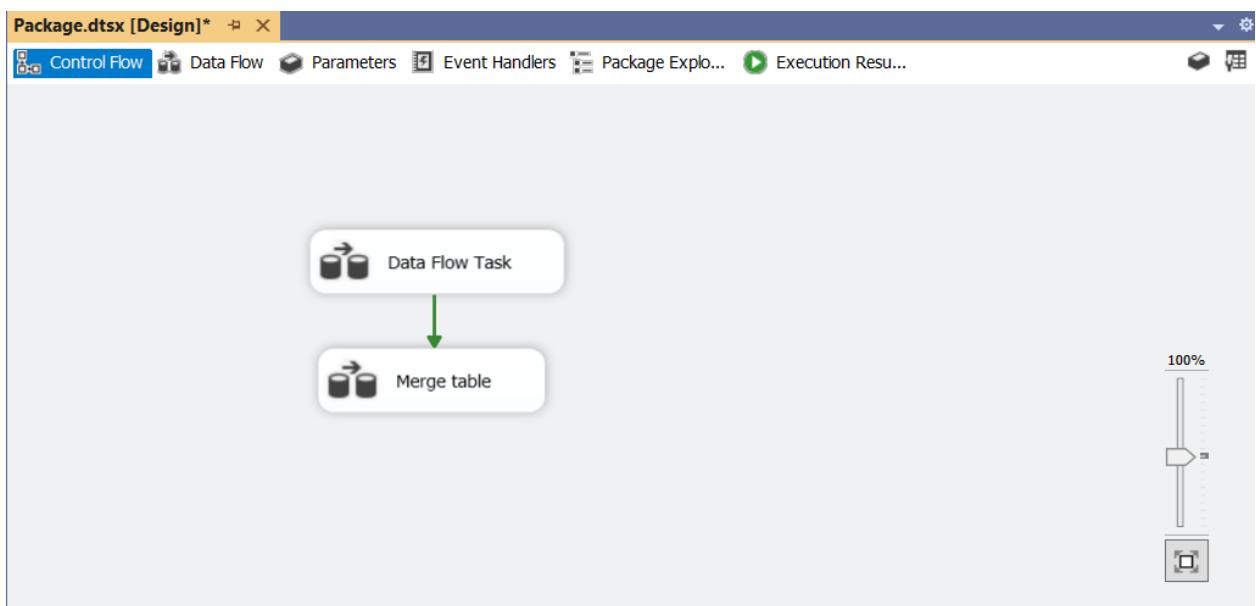
Kết quả sau khi thực thi *Data Flow Task*:



Hình 2.41 Kết quả thực thi Data Flow Task

2.4.8.2 Merge Fact Raw và các bảng Dim khác vào Fact Movie

Bước 1: Ở tab Control Flow, tạo Data Flow Task và đổi tên Data Flow Task vừa tạo là “Merge table”.



Hình 2.42 Tạo Data Flow Task “Merge table”

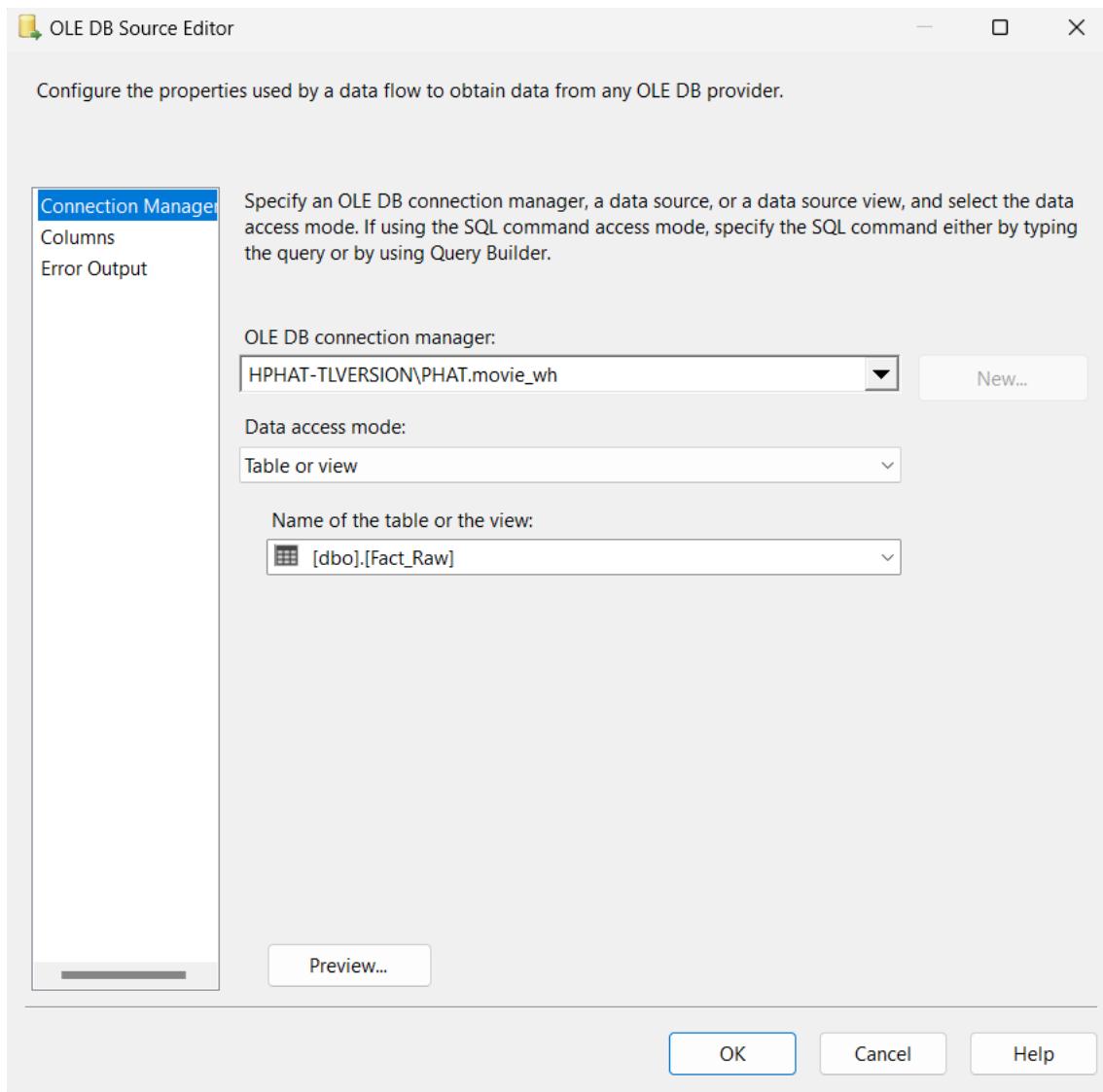
SVTH: Nguyễn Hồng Phát

Bước 2: Click chuột phải vào Data Flow Task nói trên và chọn **Edit**, trong tab **DataFlow** ta tạo 2 **OLE DB Source** và đổi tên **Fact_Raw** và **Dim_Language**



Hình 2.43 Tạo OLE DB Source cho Fact Raw và Dim Language

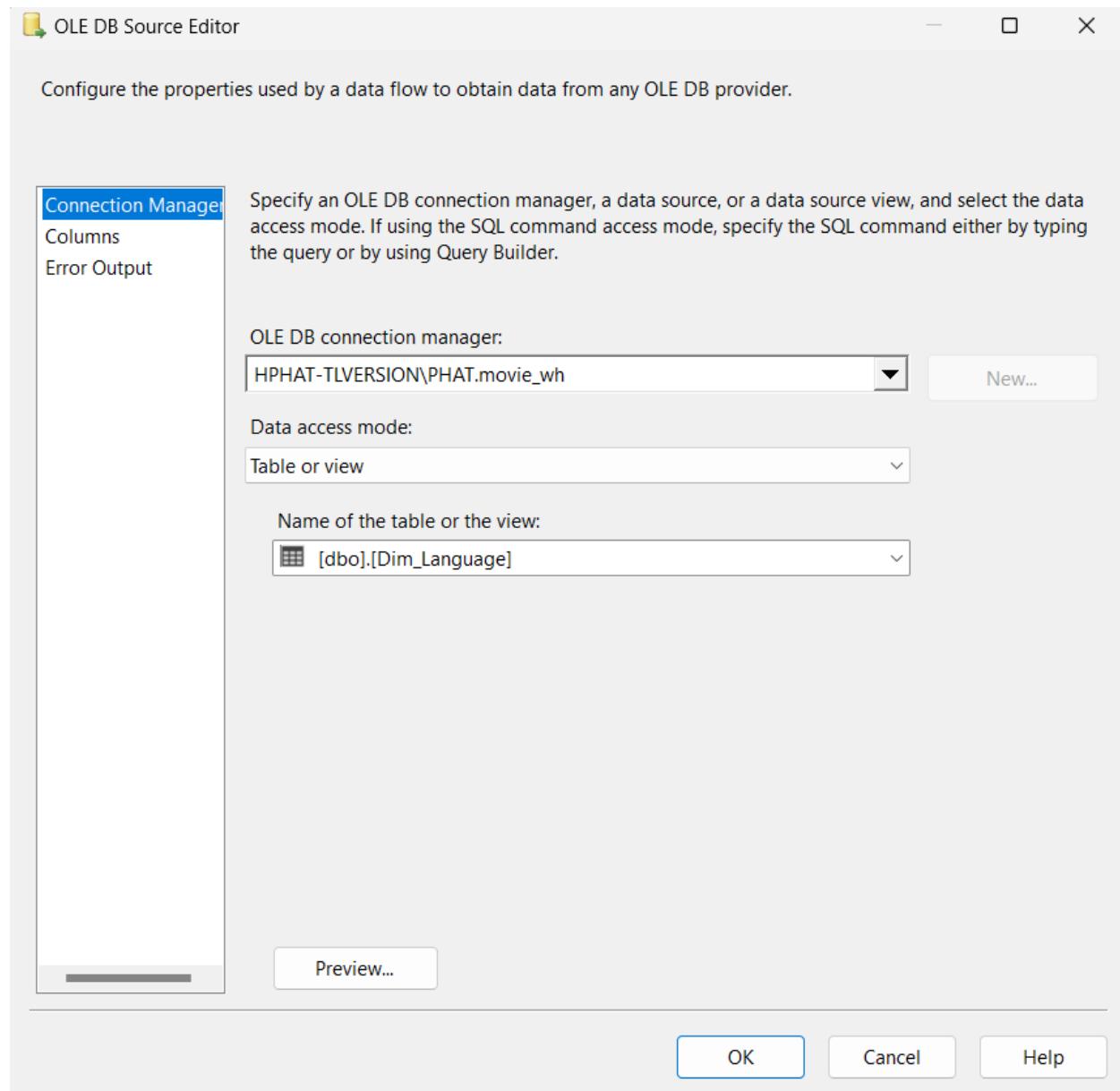
Bước 3: Click chuột phải chọn **Edit**, sau đó chọn bảng **Fact_Raw** đã tạo trước đó làm data source cho bảng **Fact_Raw** mới này.



Hình 2.44 Chọn Fact_Raw làm data source cho Fact_Raw vừa tạo

Bước 4: Chọn mục **Columns** để xem xét các cột được ánh xạ. Nhấn **OK**.

Bước 5: Tương tự thực hiện chọn ánh xạ cột cho **Dim_Language**



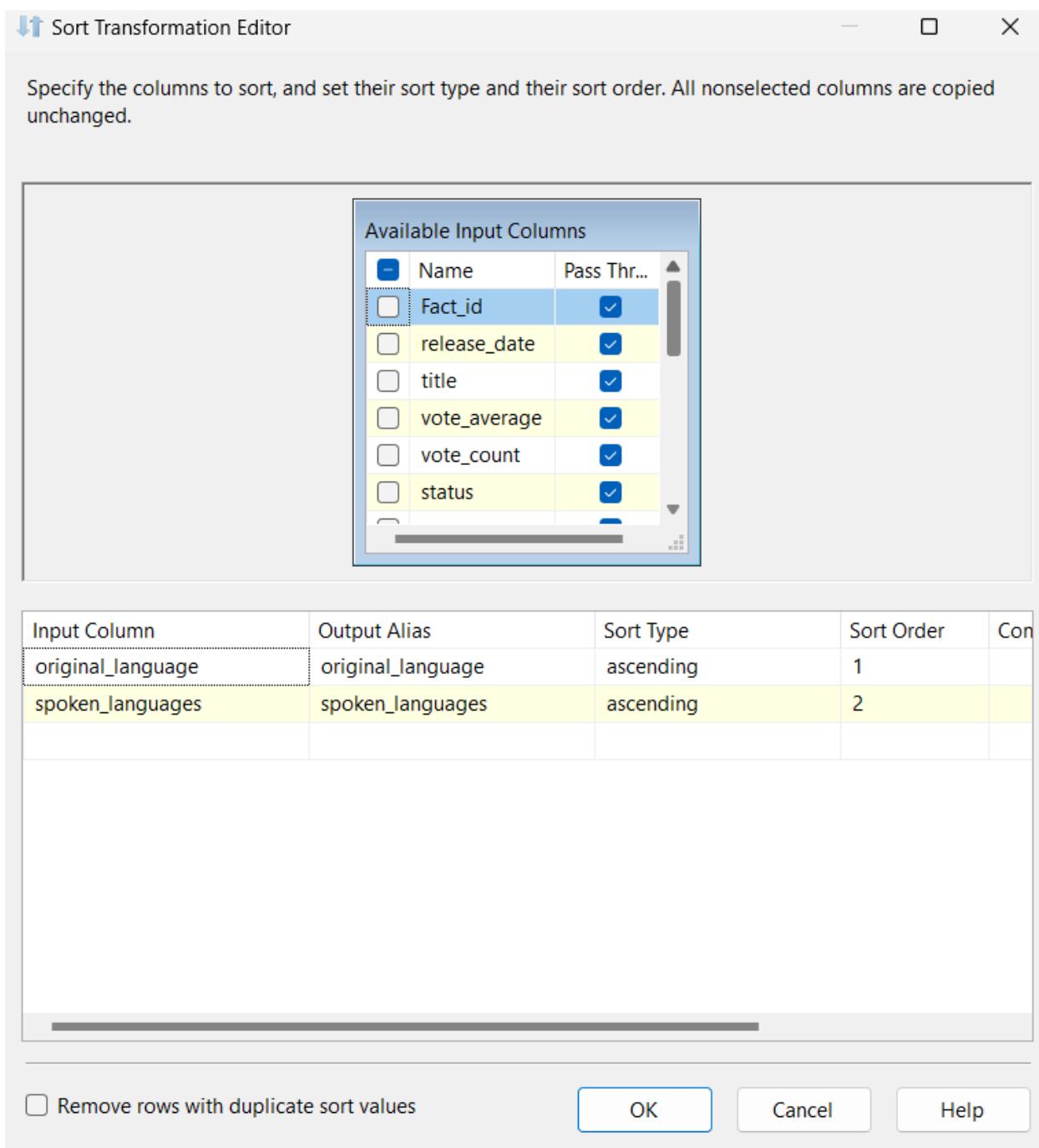
Hình 2.45 Chọn Dim_Language làm data source cho OLE DB Source “Dim_Language”

Bước 6: Tạo 2 Sort là **Sort** và **Sort1** tương ứng với mỗi Source.

Bước 7: Ở **Sort**, click chuột phải chọn **Edit** và chọn các cột **original_language**, **spoken_languages** để chuẩn bị cho quá trình merge.

* **Lưu ý:** Sort ở Fact Raw không nhấn **Remove rows with duplicate sort values**.

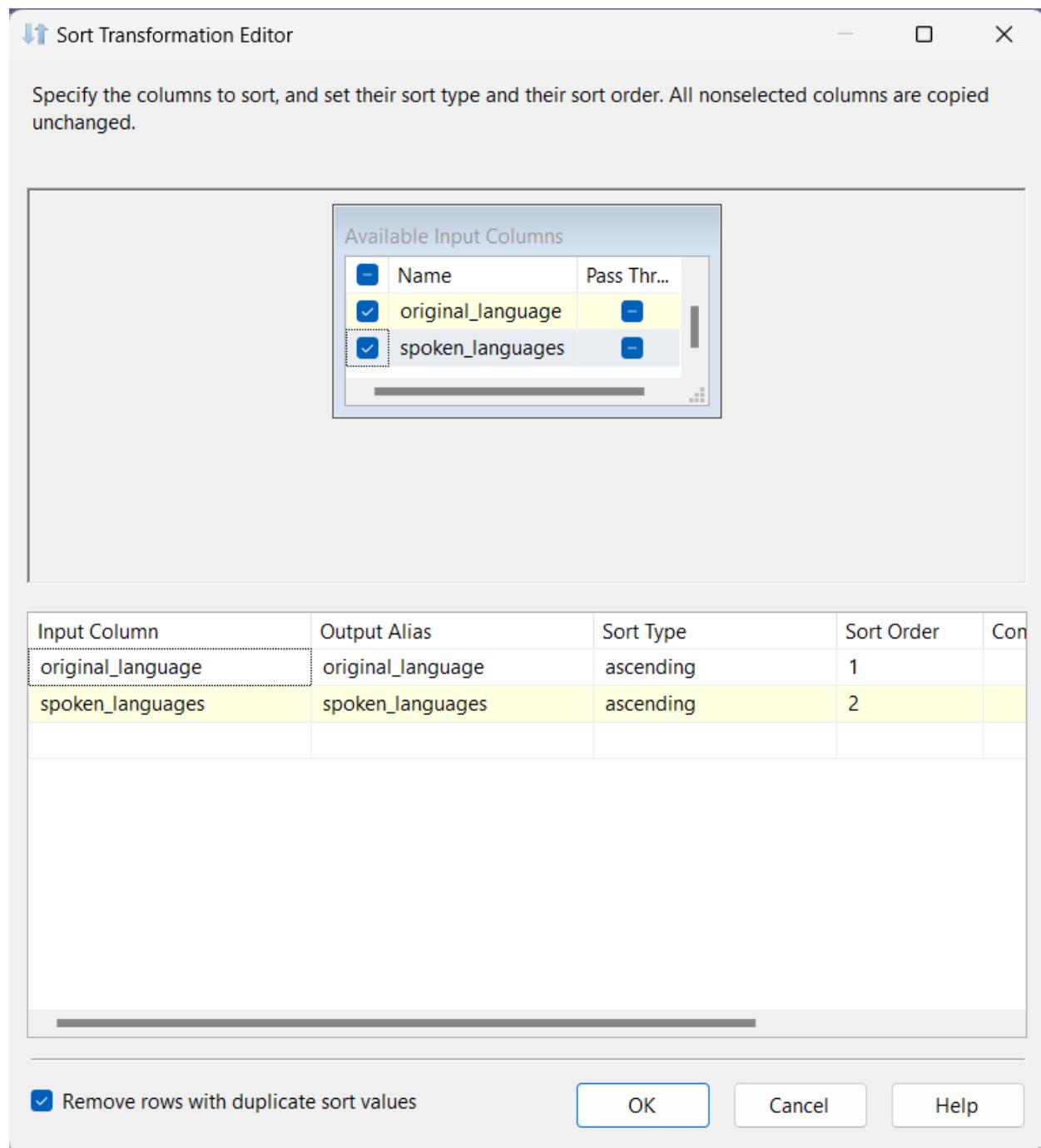
Kho dữ liệu và OLAP - IS217.P12



Hình 2.46 Chọn cột original_language, spoken_languages để đổ dữ liệu vào cho Sort

Bước 8: Tạo một Merge Join và nối với Sort, tiếp theo ta chọn Merge Join Left Input để giữ lại toàn bộ các dòng trong bảng Fact_Raw.

Bước 9: Tương tự ta chọn các cột **original_language, spoken_languages** cho Sort1



Hình 2.47 Chọn cột original_language, spoken_languages để đổ dữ liệu vào cho Sort1

Bước 10: Nối Sort1 với Merge Join

Chuột phải vào **Merge Join** và nhấn **Edit**, ở đây ta tick chọn tất cả các cột của **Sort** nhưng không lấy 2 thuộc tính **original_language** và **spoken_languages**

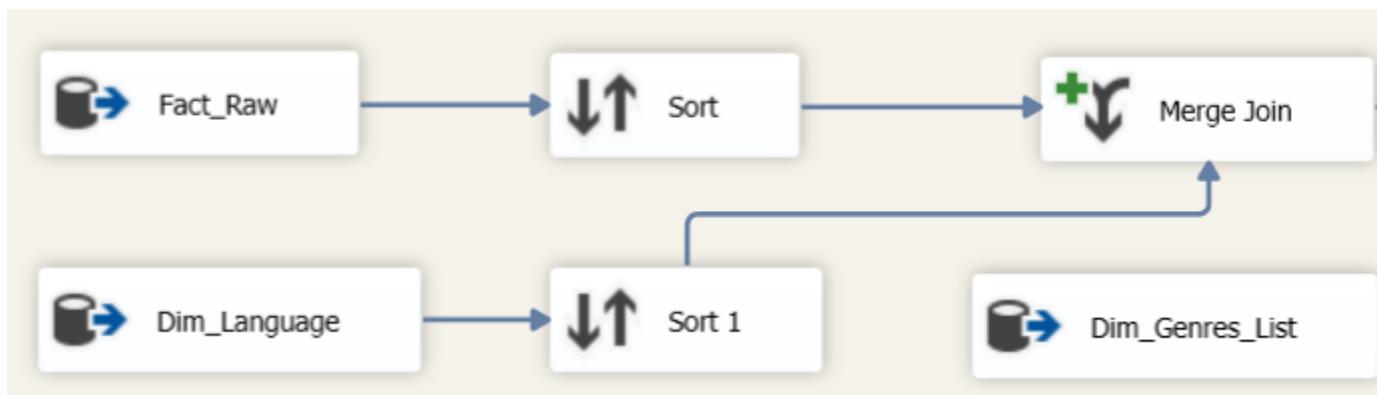
Tiếp theo ta chọn **language_id** ở **Sort1** để merge vào **Fact_Raw**

Kết quả sau khi merge là bảng **Fact_Raw** không còn 2 thuộc tính **original_language** và **spoken_languages** và có thêm 1 thuộc tính mới là **language_id**

Input	Input Column	Output Alias
Sort	budget	budget
Sort	imdb_id	imdb_id
Sort	popularity	popularity
Sort	production_companies	production_companies
Sort	production_countries	production_countries
Sort	Director	Director
Sort	genres_list	genres_list
Sort 1	language_id	language_id

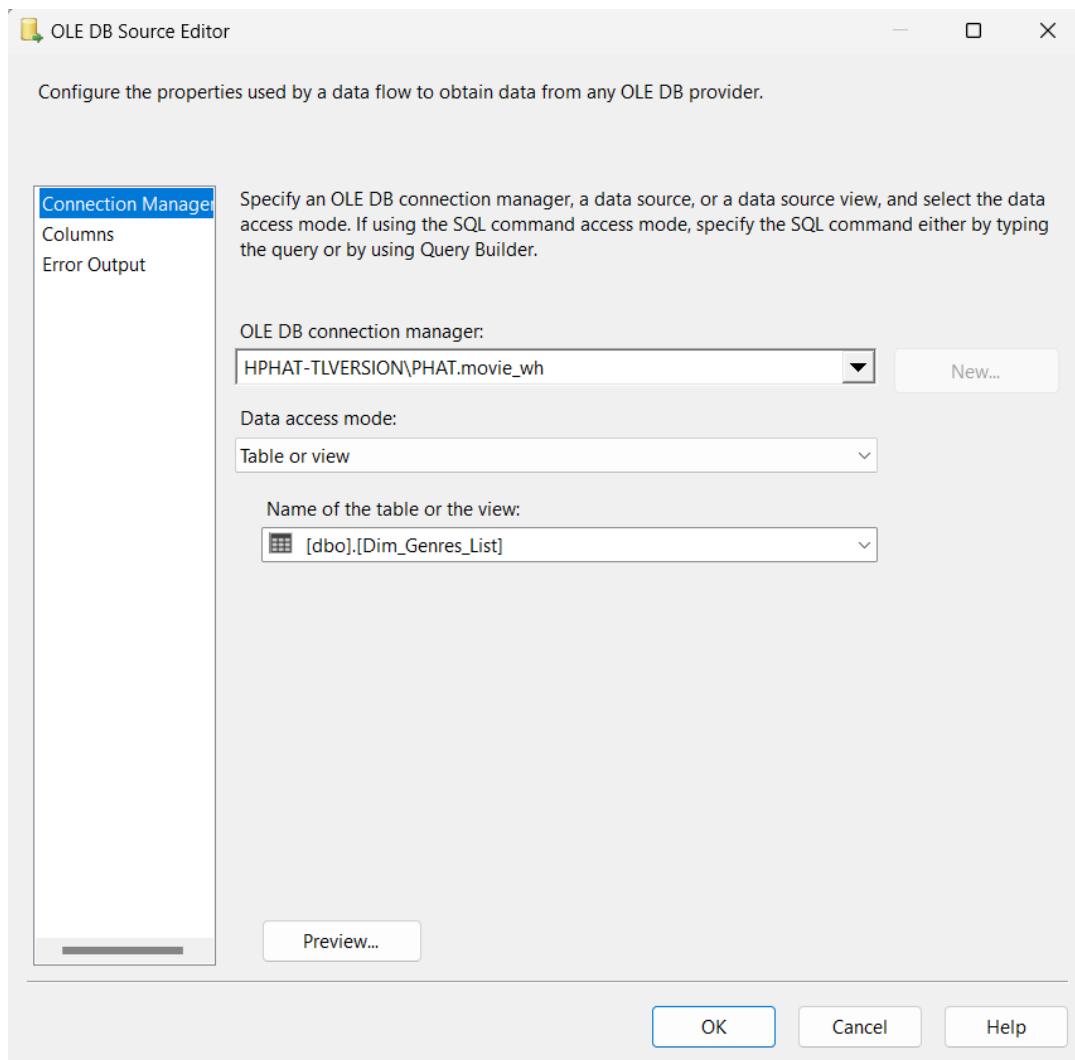
Hình 2.48 Merge Join Fact Raw với Dim Languae

Bước 11: Tạo OLE DB Source và đổi tên Dim_Genres_List.



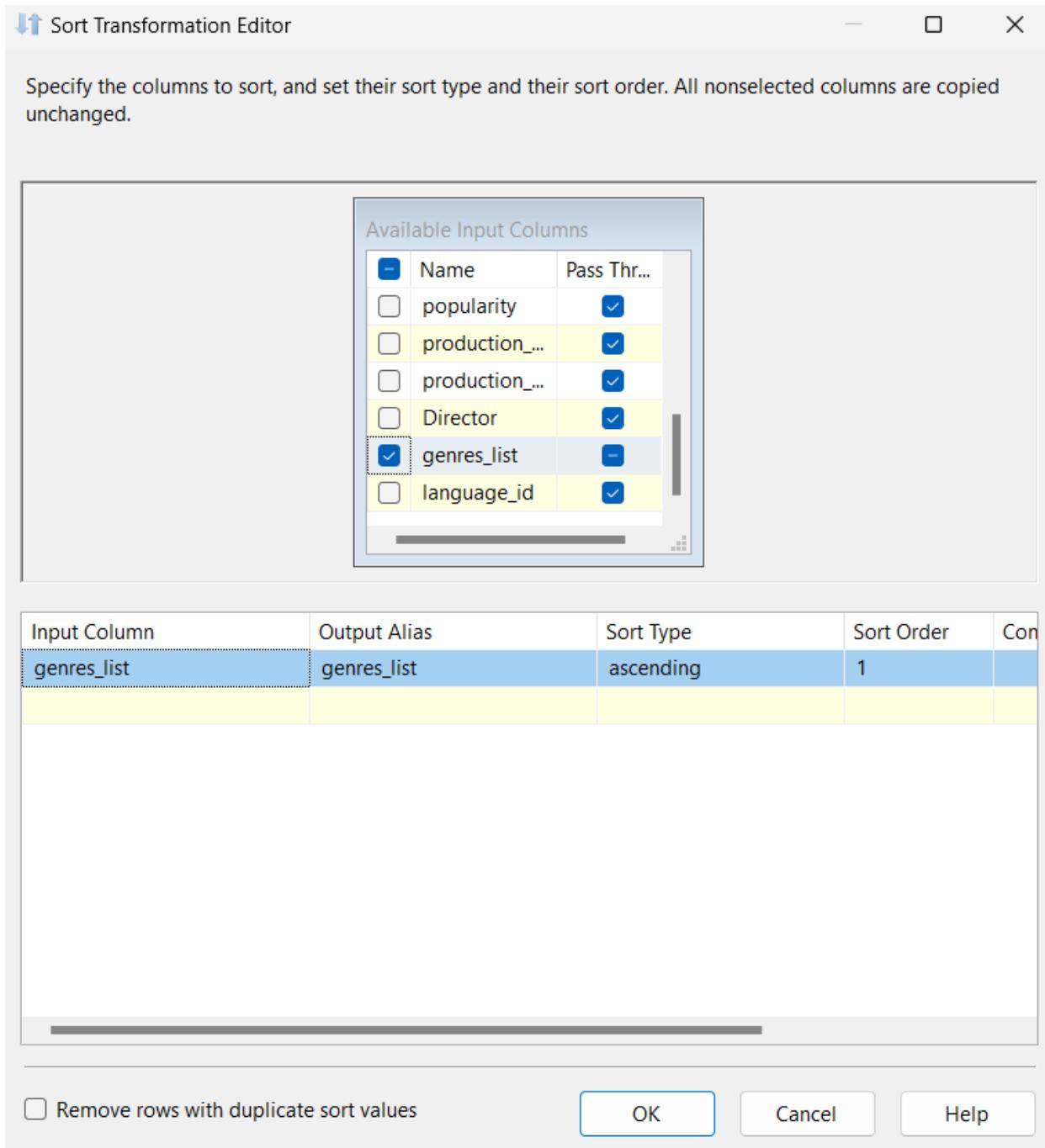
Hình 2.49 Tạo OLE DB Source cho Dim_Genres_List

Bước 12: Mở Dim_Genres_List, Mở Edit, chọn Dim_Genres_List để đổ dữ liệu vào OLE DB Source Dim_Genres_List vừa tạo.



Hình 2.50 Chọn Dim_Genres_List để đổ dữ liệu vào OLE DB Source Dim_Genres_List

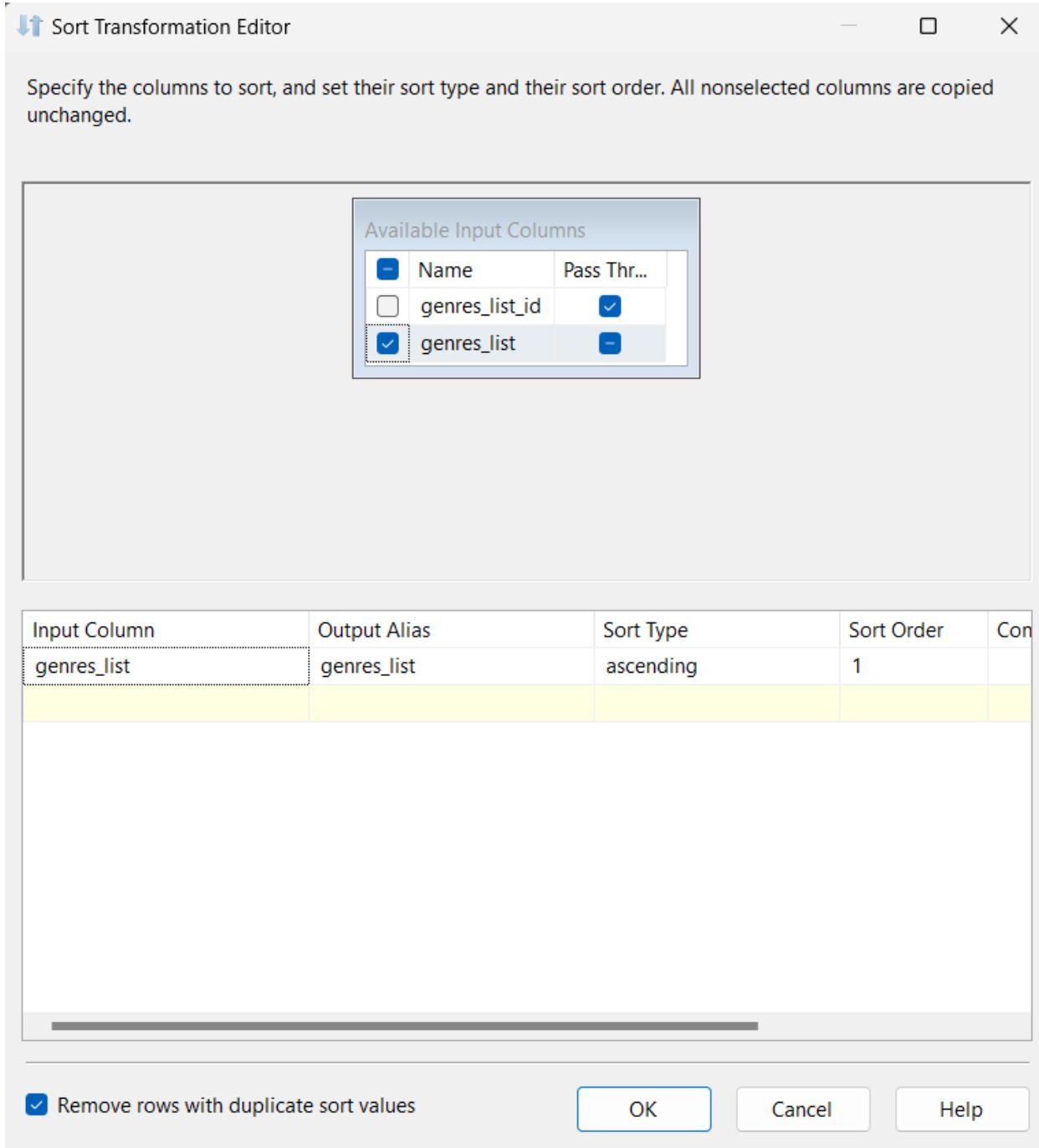
Bước 13: Tạo 2 Sort tương ứng với Dim_Genres_List và Merge Join. Ở Sort2, click chuột phải chọn Edit và chọn các cột original_language, spoken_languages để chuẩn bị cho quá trình merge.



Hình 2.51 chọn các cột original_language, spoken_languages để đồ dữ liệu vào Sort2

Bước 14: Tạo **Merge Join 1** và nối với **Sort2**, tiếp theo chọn **Merge Join Left Input** để giữ lại toàn bộ các dòng trong bảng **Merge Join**.

Bước 15: Tương tự ta chọn các cột **genres_list** cho **Sort3**



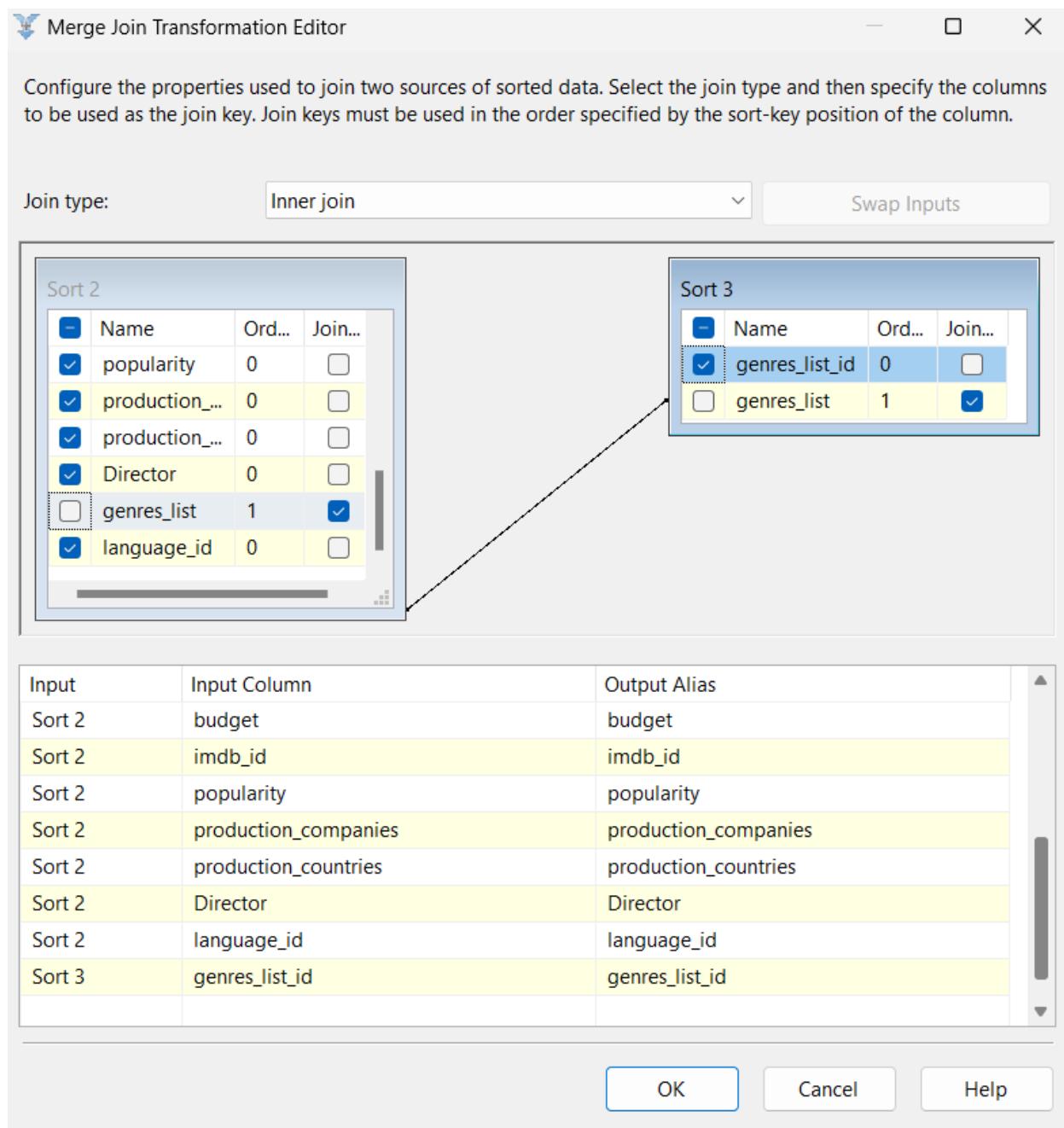
Hình 2.52 chọn các cột *original_language*, *spoken_languages* để đổ dữ liệu vào *Sort3*

Bước 16: Nối Sort3 với Merge Join 1

Chuột phải vào **Merge Join 1** và nhấn **Edit**, ở đây ta tick chọn tất cả các cột của **Sort2** nhưng không lấy thuộc tính **genres_list**.

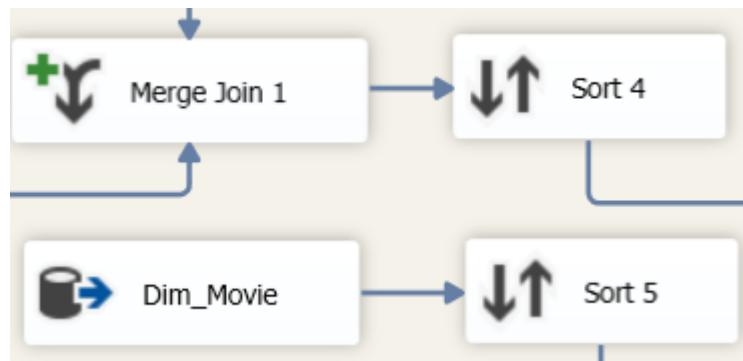
Tiếp theo ta chọn **genres_list_id** ở **Sort3** để merge vào **Fact Raw**.

Kết quả sau khi merge là bảng **Fact_Raw** không còn thuộc tính **genres_list** và có thêm 1 thuộc tính mới là **genres_list_id**.



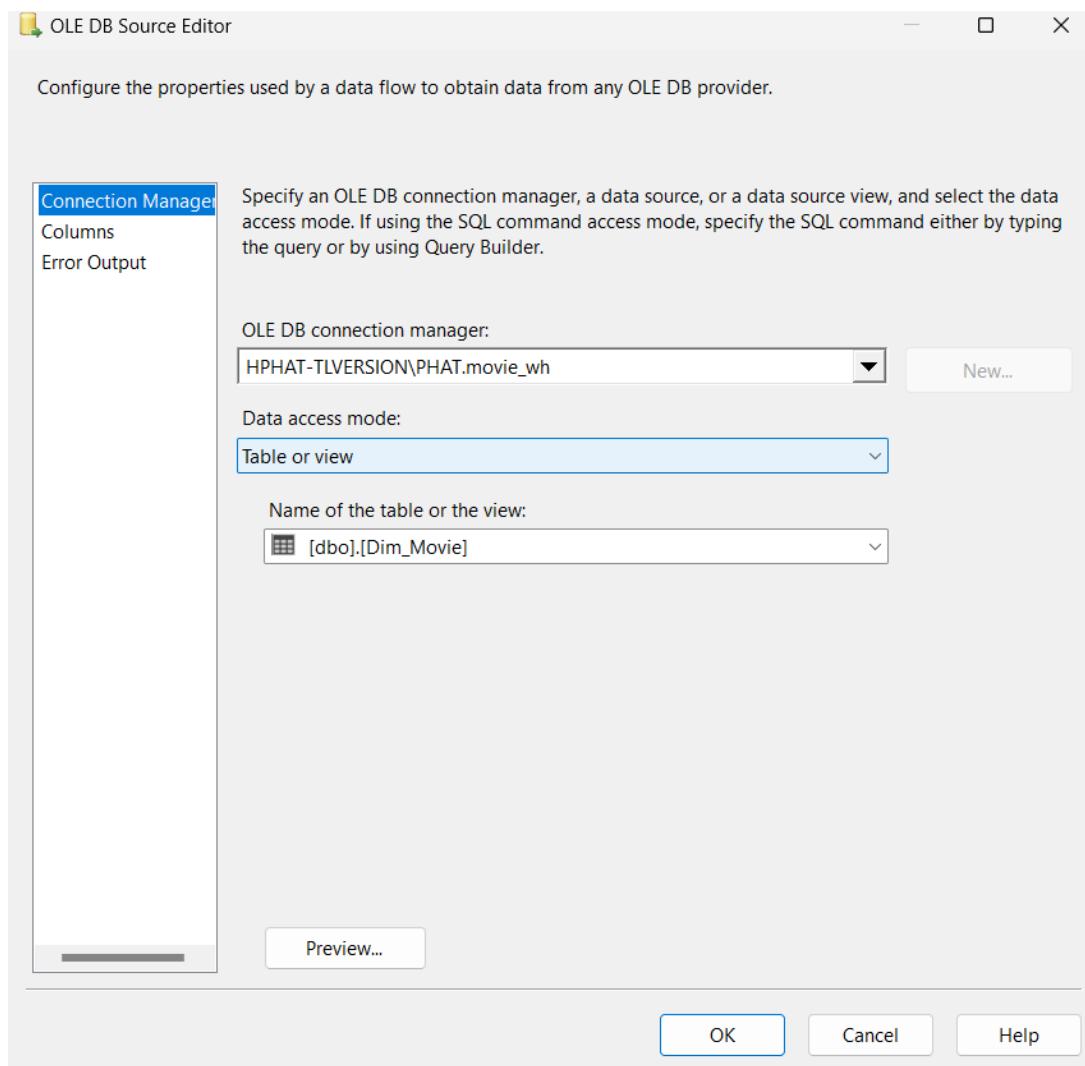
Hình 2.53 Merge Join Fact Raw với Dim Genres List

Bước 17: Tạo **OLE DB Source** và đổi tên **Dim_Movie** Tạo 2 **Sort** tương ứng với **Dim_Movie** và **Merge Join 1**.



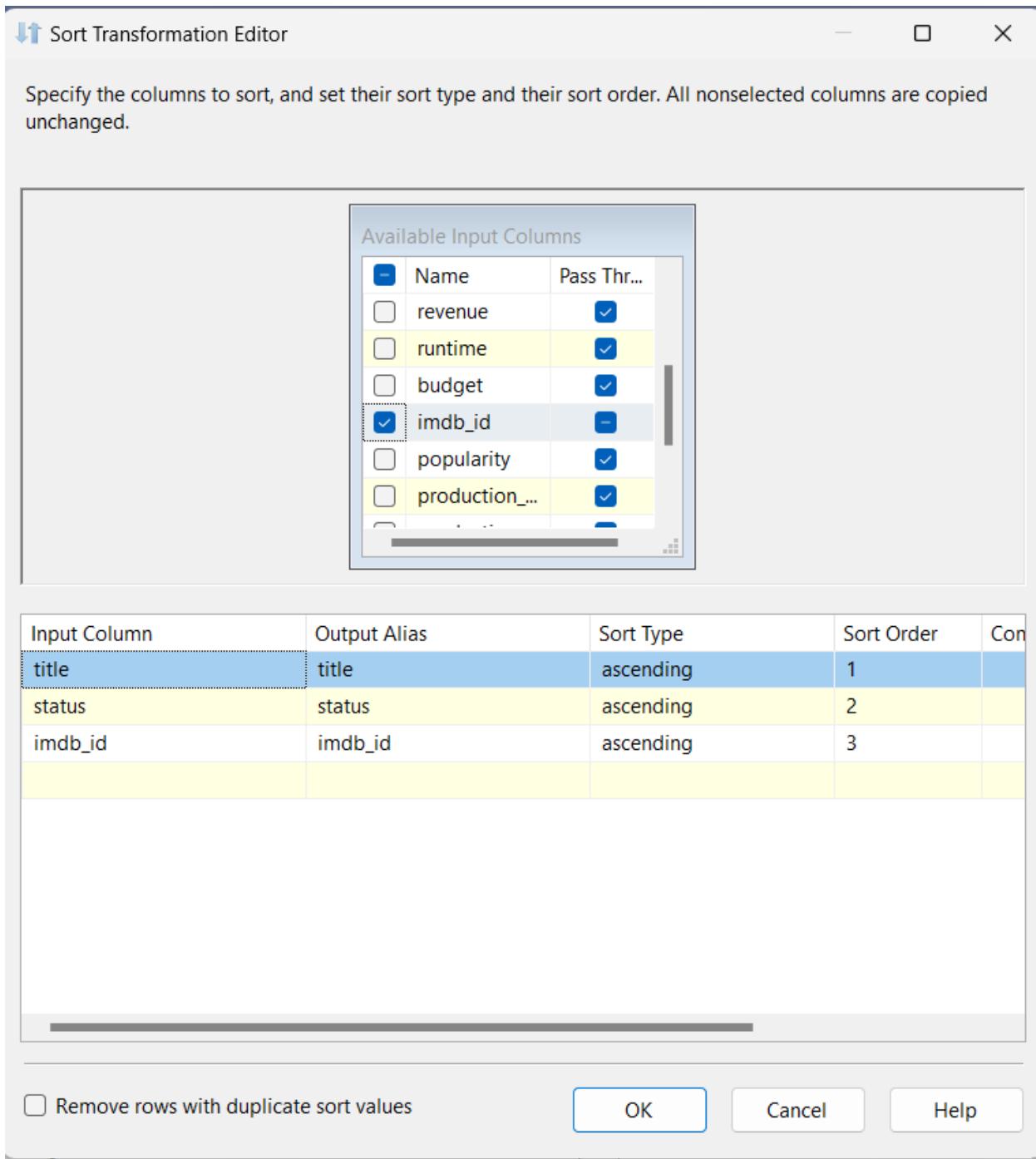
Hình 2.54 Tạo OLE DB Source cho Dim_Movie

Bước 18: Mở **Dim_Movie** Mở **Edit**, chọn **Dim_Movie** để đổ dữ liệu vào **OLE DB Source Dim_Movie** vừa tạo.



Hình 2.55 Chọn Dim_Movie để đổ dữ liệu vào OLE DB Source Dim_Movie

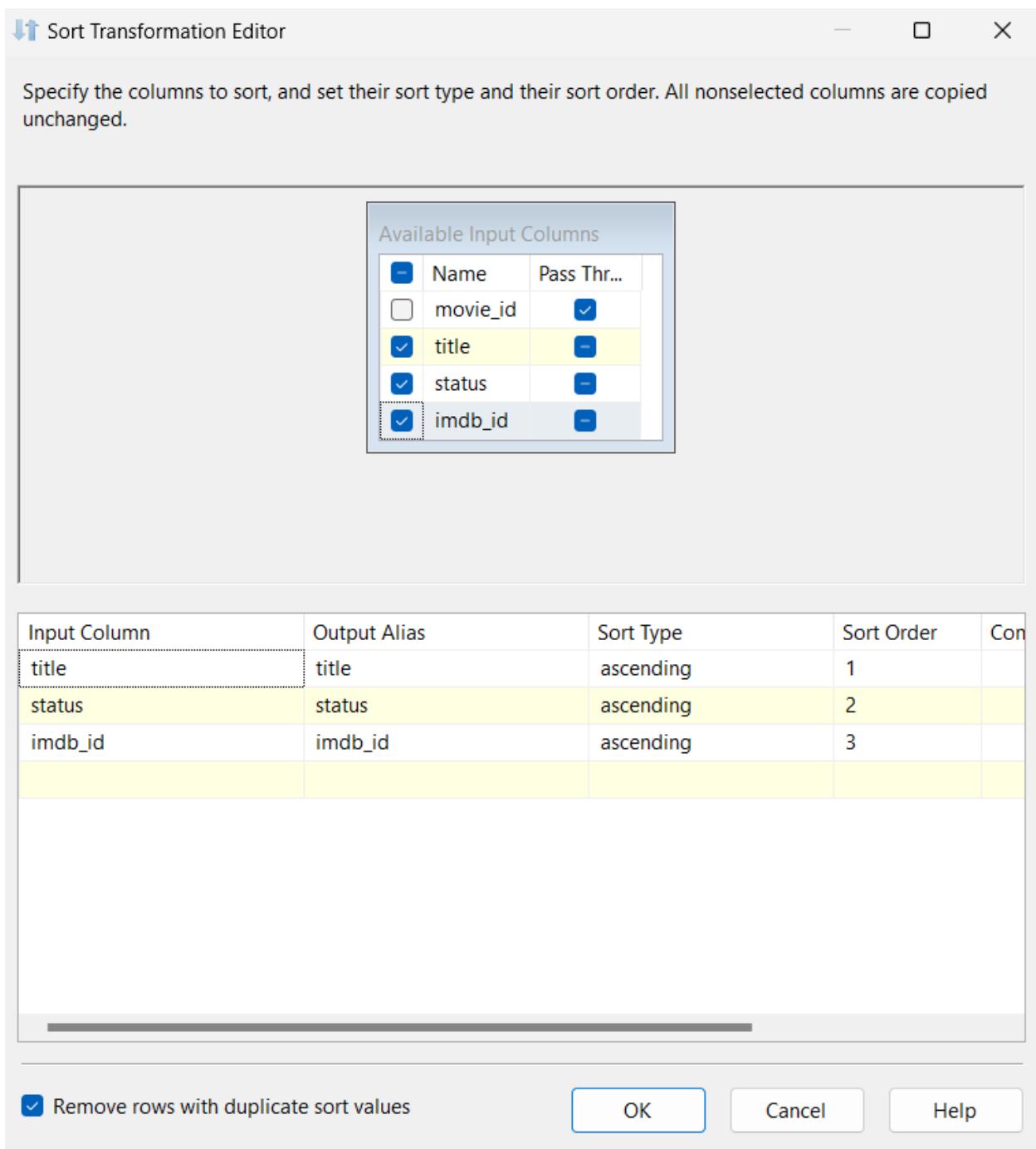
Bước 19: Ở Sort4, click chuột phải chọn **Edit** và chọn các cột **title, status, imdb_id** để chuẩn bị cho quá trình merge.



Hình 2.56 chọn các cột title, status, imdb_id để đổ dữ liệu vào Sort4

Bước 20: Tạo Merge Join 2 và nối với Sort4, tiếp theo chọn Merge Join Left Input để giữ lại toàn bộ các dòng trong bảng Merge Join 2.

Bước 21: Tương tự ta chọn các cột **title, status, imdb_id** cho Sort5



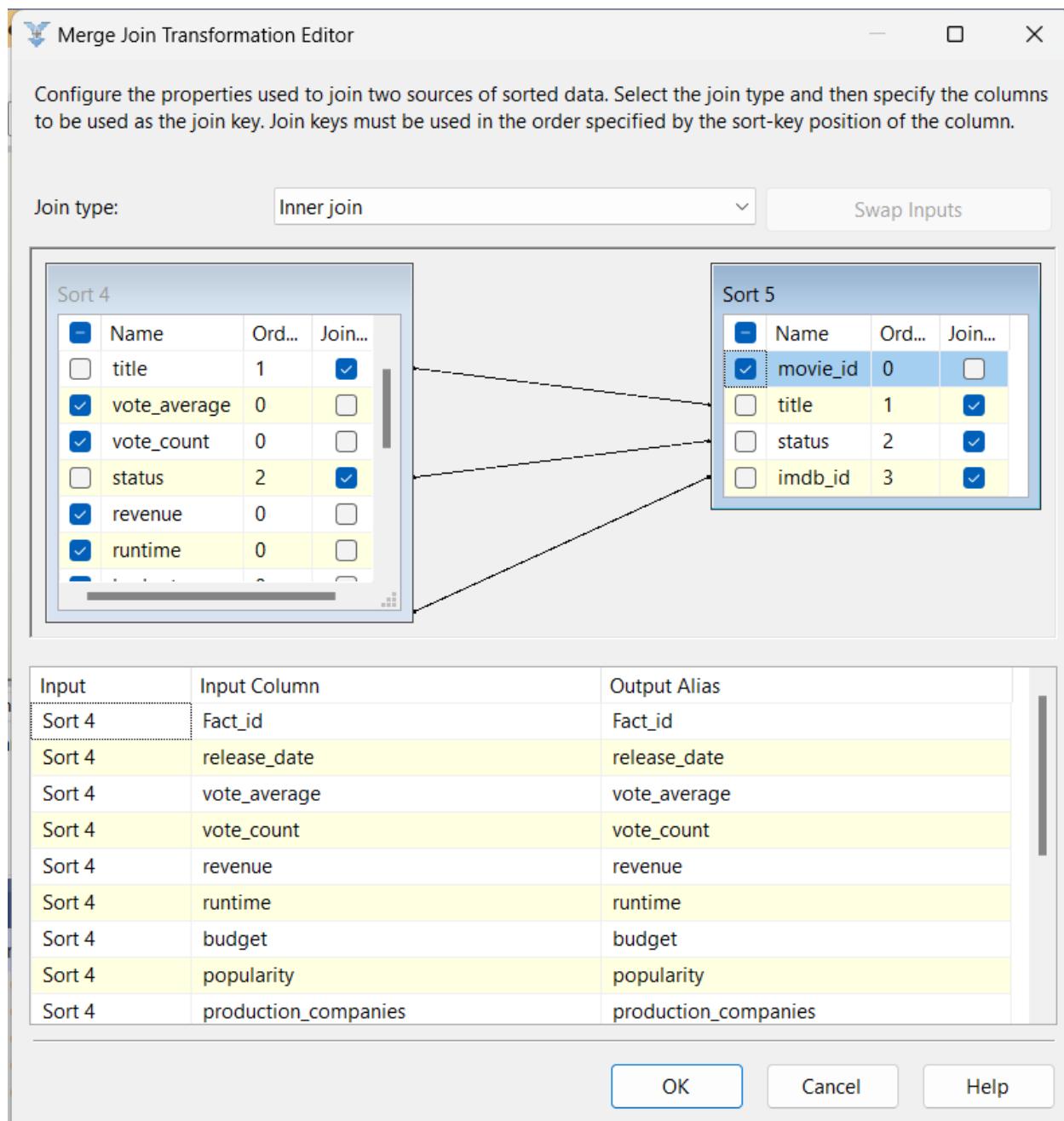
Hình 2.57 chọn các cột title, status, imdb_id để đổ dữ liệu vào Sort5

Bước 22: Nối Sort5 với Merge Join 2

Chuột phải vào **Merge Join 2** và nhấn **Edit**, ở đây ta tick chọn tất cả các cột của **Sort4** nhưng không lấy các thuộc tính **title, status, imdb_id**.

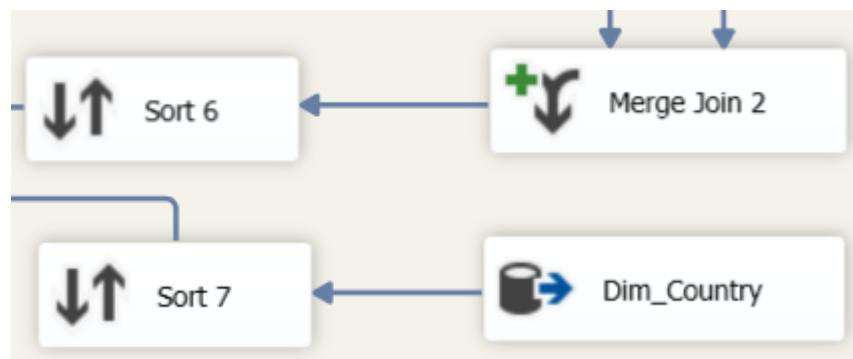
Tiếp theo ta chọn **movie_id** ở **Sort5** để merge vào **Fact Raw**.

Kết quả sau khi merge là bảng Fact_Raw không còn các thuộc tính **title, status, imdb_id** và có thêm 1 thuộc tính mới là **movie_id**.



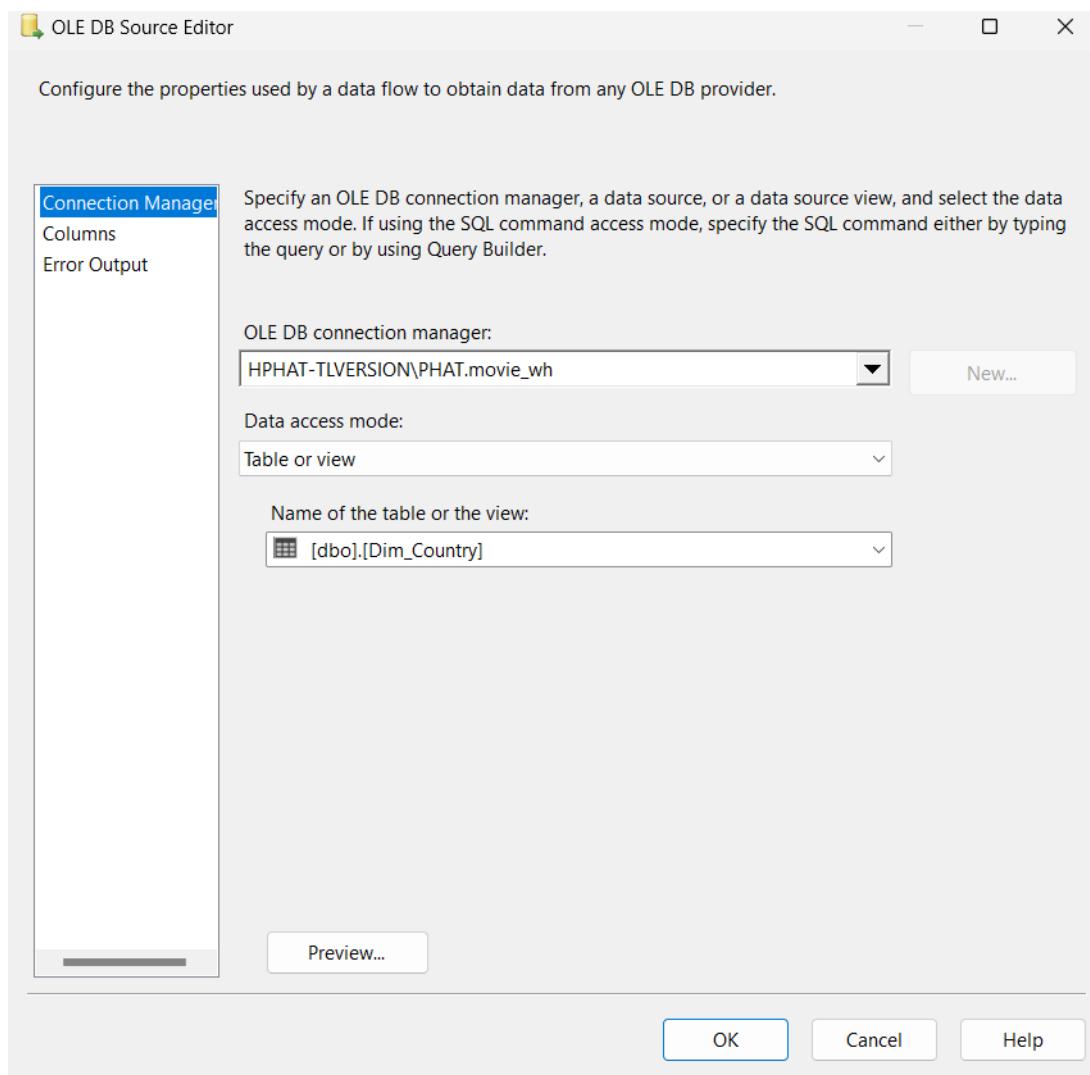
Hình 2.58 Merge Join Fact Raw với Dim Movie

Bước 23: Tạo **OLE DB Source** và đổi tên **Dim_Country** Tạo 2 **Sort** tương ứng với **Dim_Country** và **Merge Join 2**.



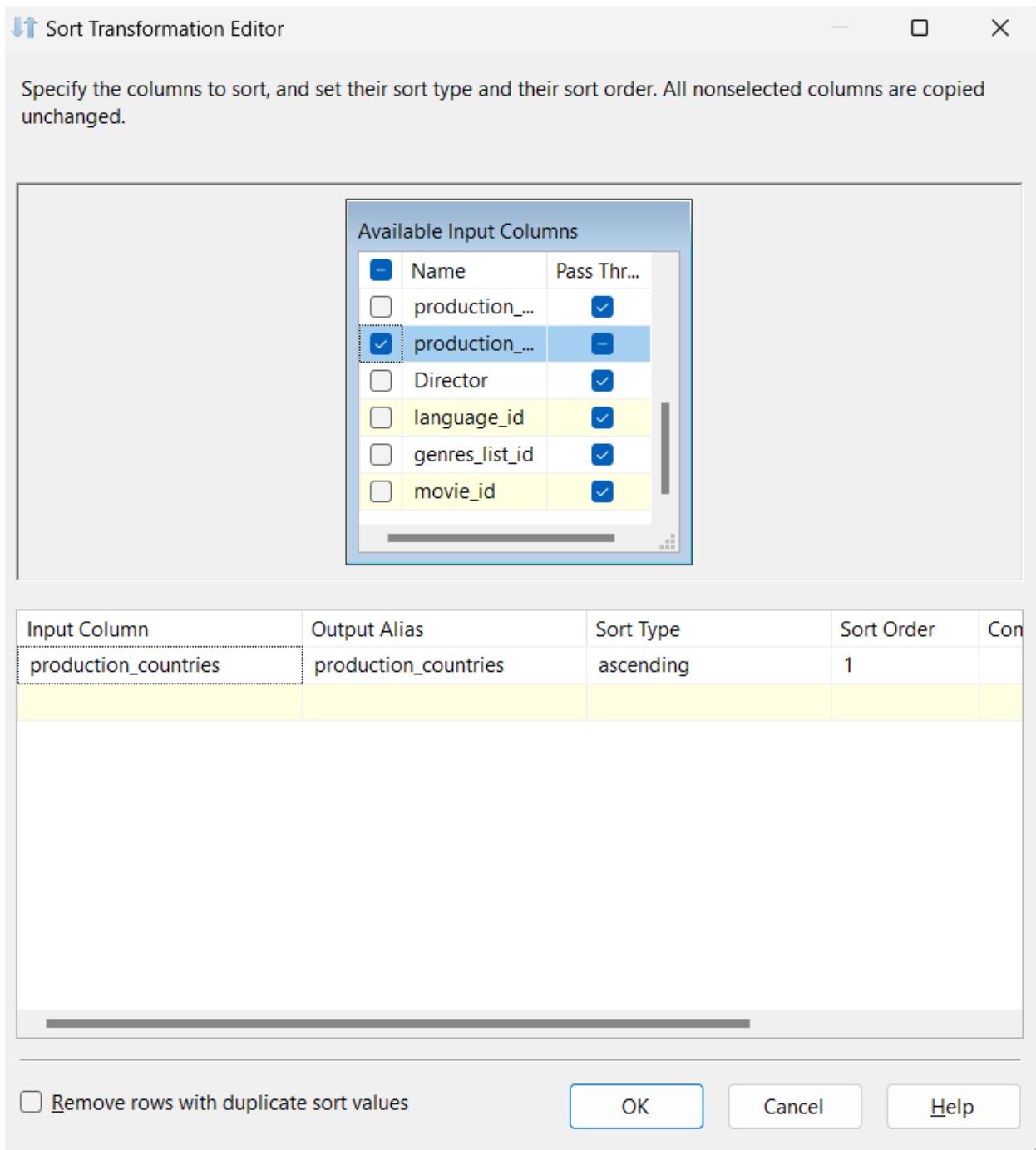
Hình 2.59 Tạo OLE DB Source cho Dim_Country

Bước 24: Mở **Dim_Country** Mở **Edit**, chọn **Dim_Country** để đổ dữ liệu vào **OLE DB Source Dim_Country** vừa tạo.



Hình 2.60 Chọn Dim_Country để đổ dữ liệu vào OLE DB Source Dim_Country

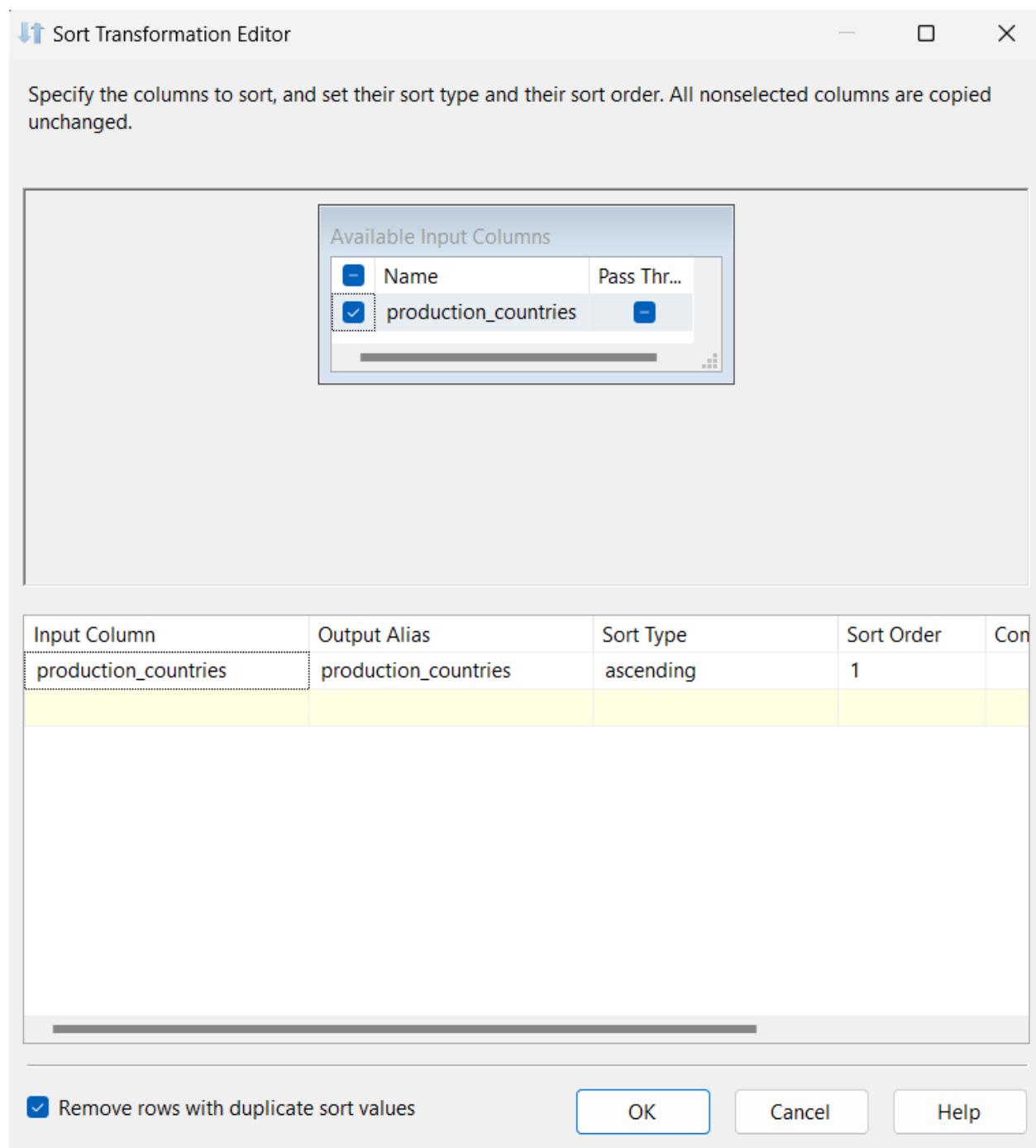
Bước 25: Ở Sort6, click chuột phải chọn **Edit** và chọn cột **production_countries** để chuẩn bị cho quá trình merge.



Hình 2.61 chọn cột production_countries để đổ dữ liệu vào Sort6

Bước 26: Tạo **Merge Join 3** và nối với **Sort6**, tiếp theo chọn **Merge Join Left Input** để giữ lại toàn bộ các dòng trong bảng **Merge Join 3**.

Bước 27: Tương tự ta chọn cột **production_countries** cho Sort7



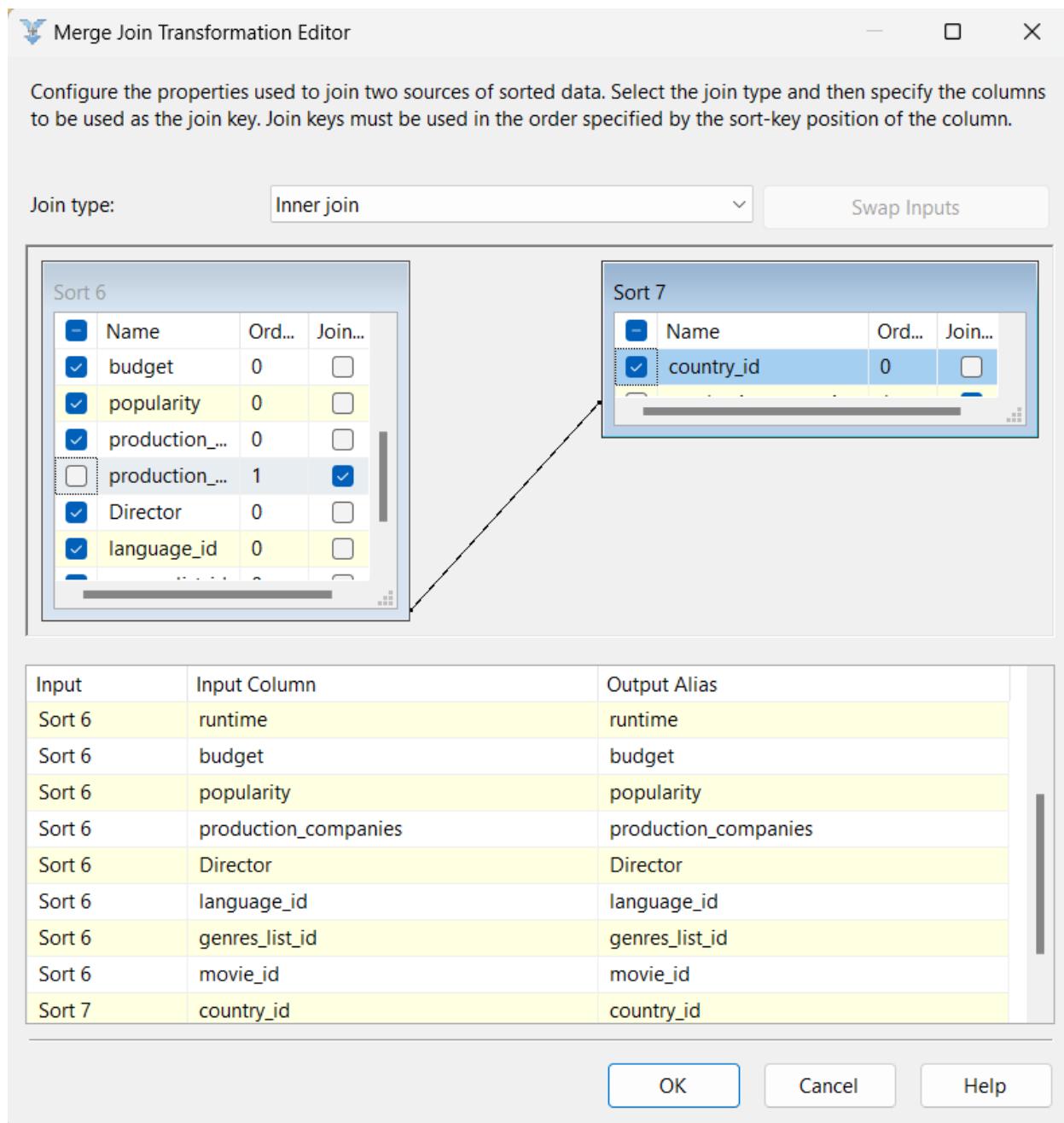
Hình 2.62 chọn cột production_countries để đổ dữ liệu vào Sort7

Bước 28: Nối Sort7 với Merge Join 3

Chuột phải vào **Merge Join 3** và nhấn **Edit**, ở đây ta tick chọn tất cả các cột của **Sort6** nhưng không lấy thuộc tính **production_countries**.

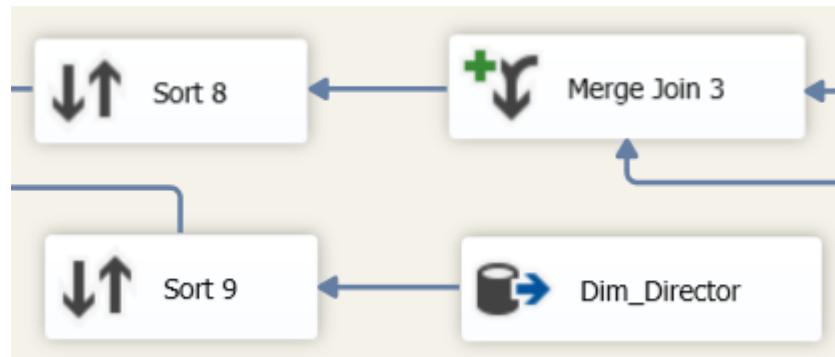
Tiếp theo ta chọn **country_id** ở **Sort7** để merge vào **Fact Raw**.

Kết quả sau khi merge là bảng Fact_Raw không còn thuộc tính **production_countries** và có thêm 1 thuộc tính mới là **country_id**.



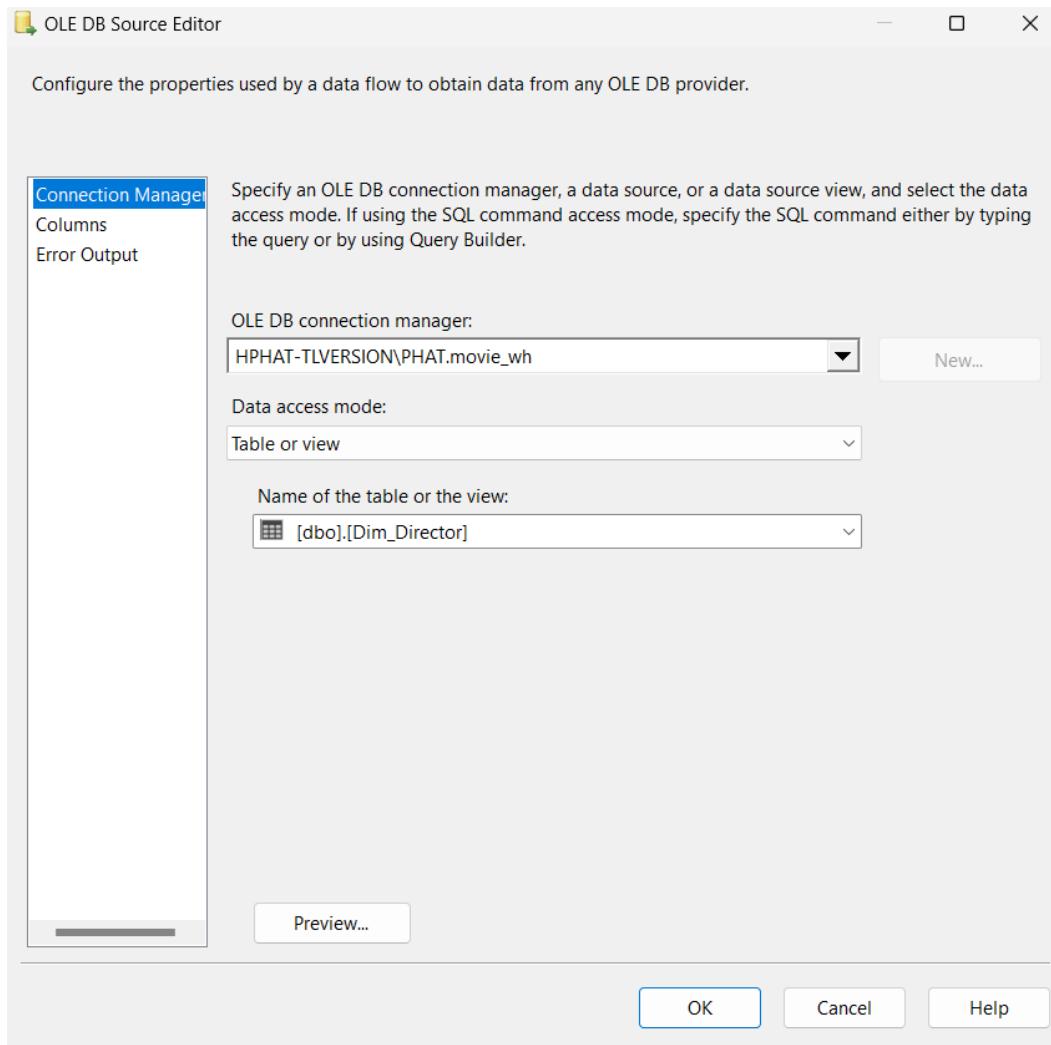
Hình 2.63 Merge Join Fact Raw với Dim Country

Bước 29: Tạo **OLE DB Source** và đổi tên **Dim_Director**. Tạo 2 **Sort** tương ứng với **Dim_Director** và **Merge Join 3**.



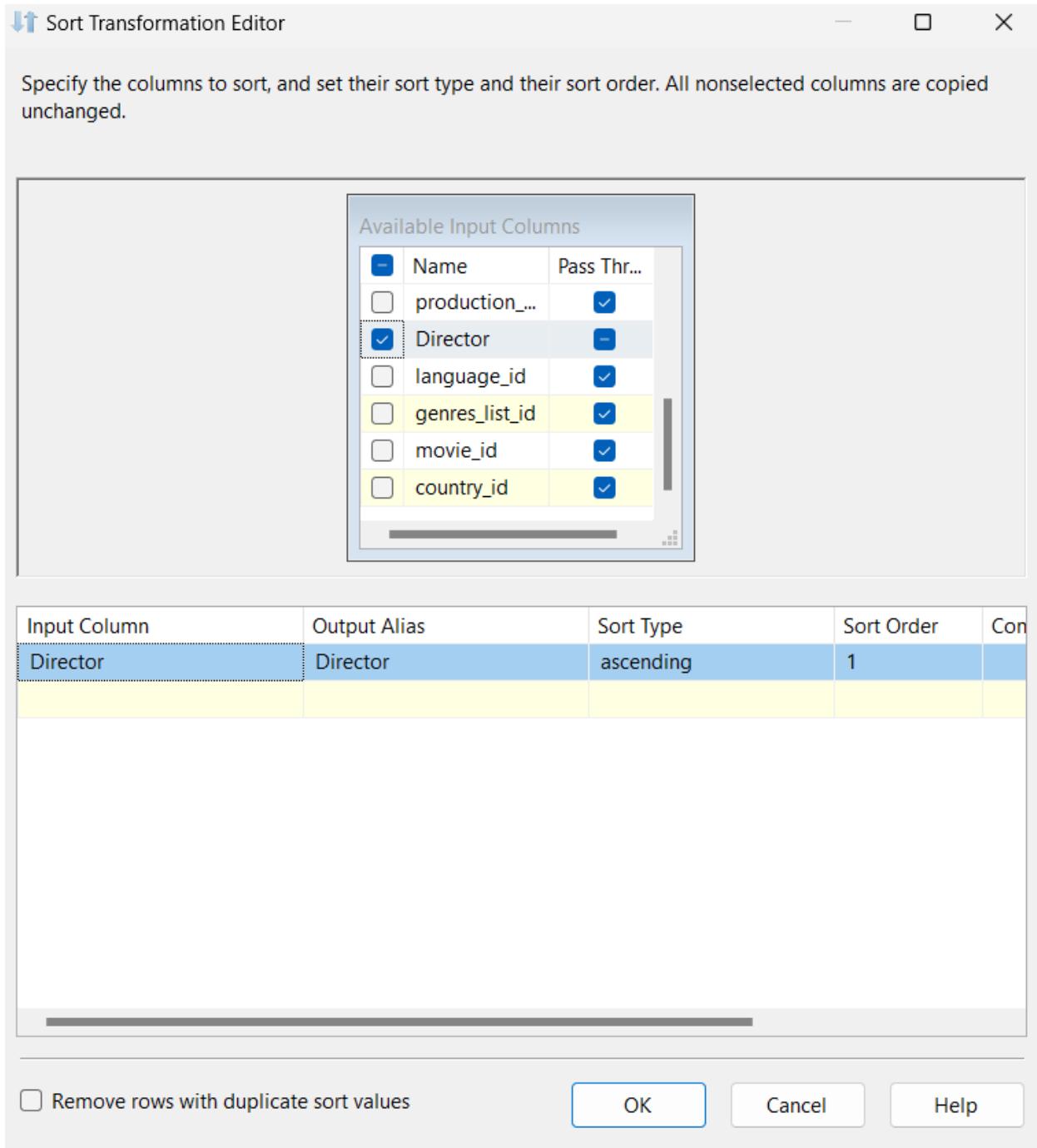
Hình 2.64 Tạo OLE DB Source cho Dim_Director

Bước 30: Mở **Dim_Director** Mở **Edit**, chọn **Dim_Director** để đổ dữ liệu vào **OLE DB Source Dim_Director** vừa tạo.



Hình 2.65 Chọn Dim_Director để đổ dữ liệu vào OLE DB Source Dim_Director

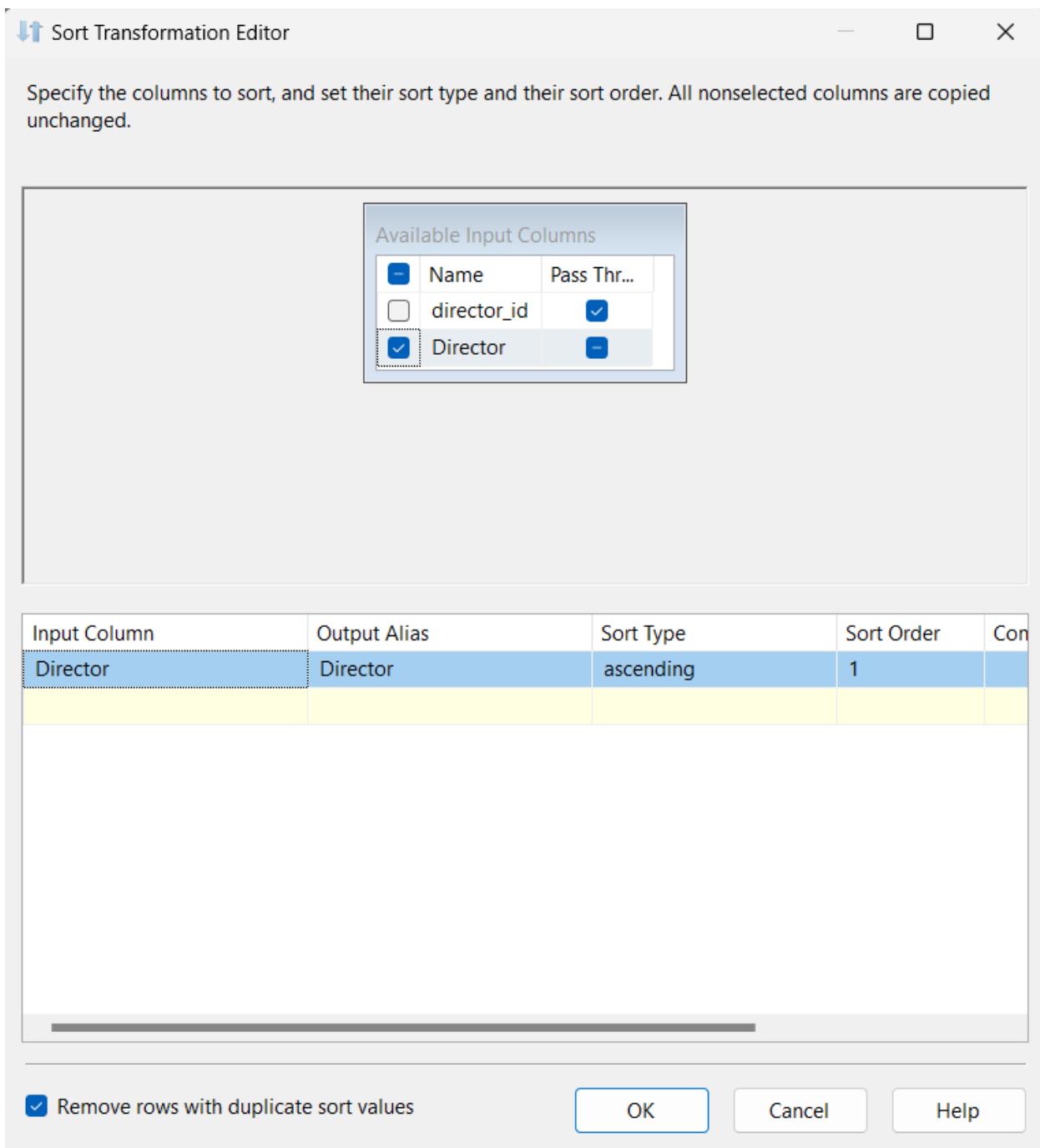
Bước 31: Ở Sort8, click chuột phải chọn **Edit** và chọn cột **Director** để chuẩn bị cho quá trình merge.



Hình 2.66 chọn cột Director để đổ dữ liệu vào Sort8

Bước 32: Tạo **Merge Join 4** và nối với **Sort8**, tiếp theo chọn **Merge Join Left Input** để giữ lại toàn bộ các dòng trong bảng **Merge Join 4**.

Bước 33: Tương tự ta chọn cột **Director** cho **Sort9**



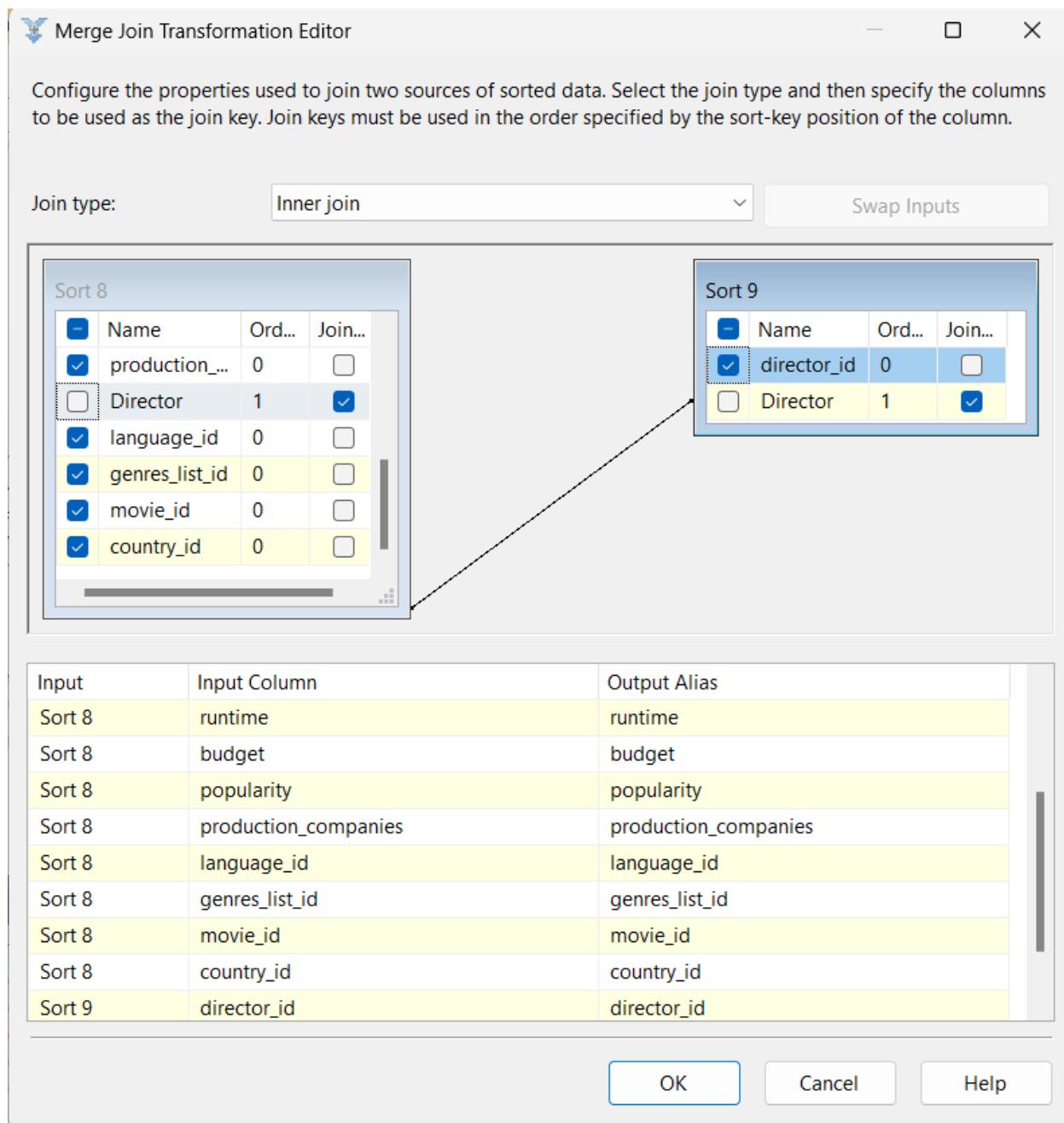
Hình 2.67 chọn cột Director để đổ dữ liệu vào Sort9

Bước 34: Nối Sort9 với Merge Join 4

Chuột phải vào **Merge Join 4** và nhấn **Edit**, ở đây ta tick chọn tất cả các cột của **Sort8** nhưng không lấy thuộc tính **Director**.

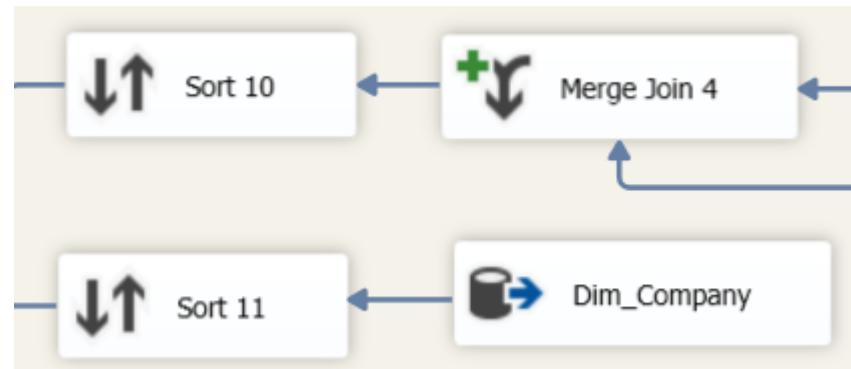
Tiếp theo ta chọn **director_id** ở **Sort9** để merge vào **Fact Raw**.

Kết quả sau khi merge là bảng **Fact_Raw** không còn thuộc tính **Director** và có thêm 1 thuộc tính mới là **director_id**.



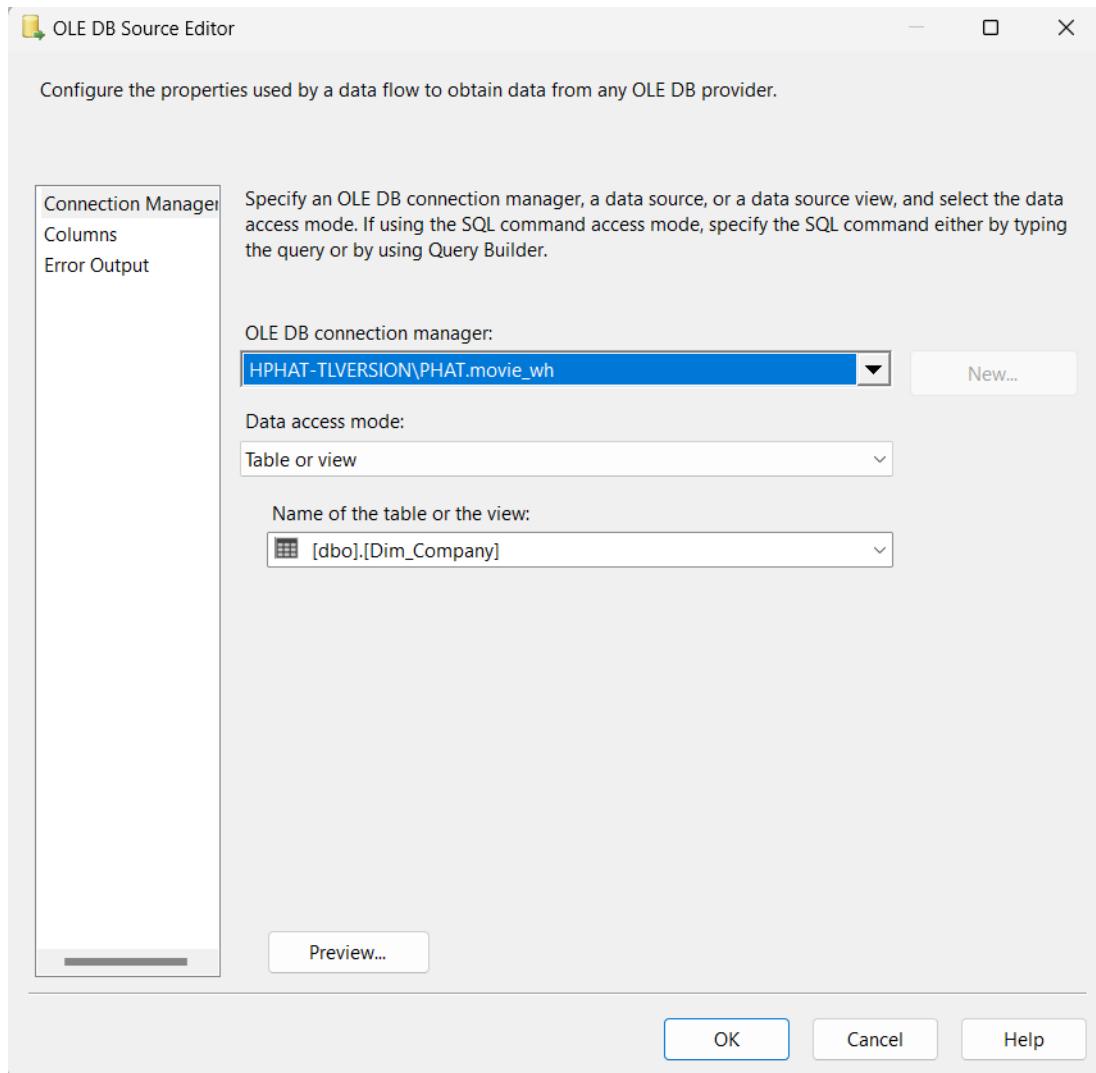
Hình 2.68 Merge Join Fact Raw với Dim Director

Bước 35: Tạo **OLE DB Source** và đổi tên **Dim_Company**. Tạo 2 **Sort** tương ứng với **Dim_Company** và **Merge Join 4**.



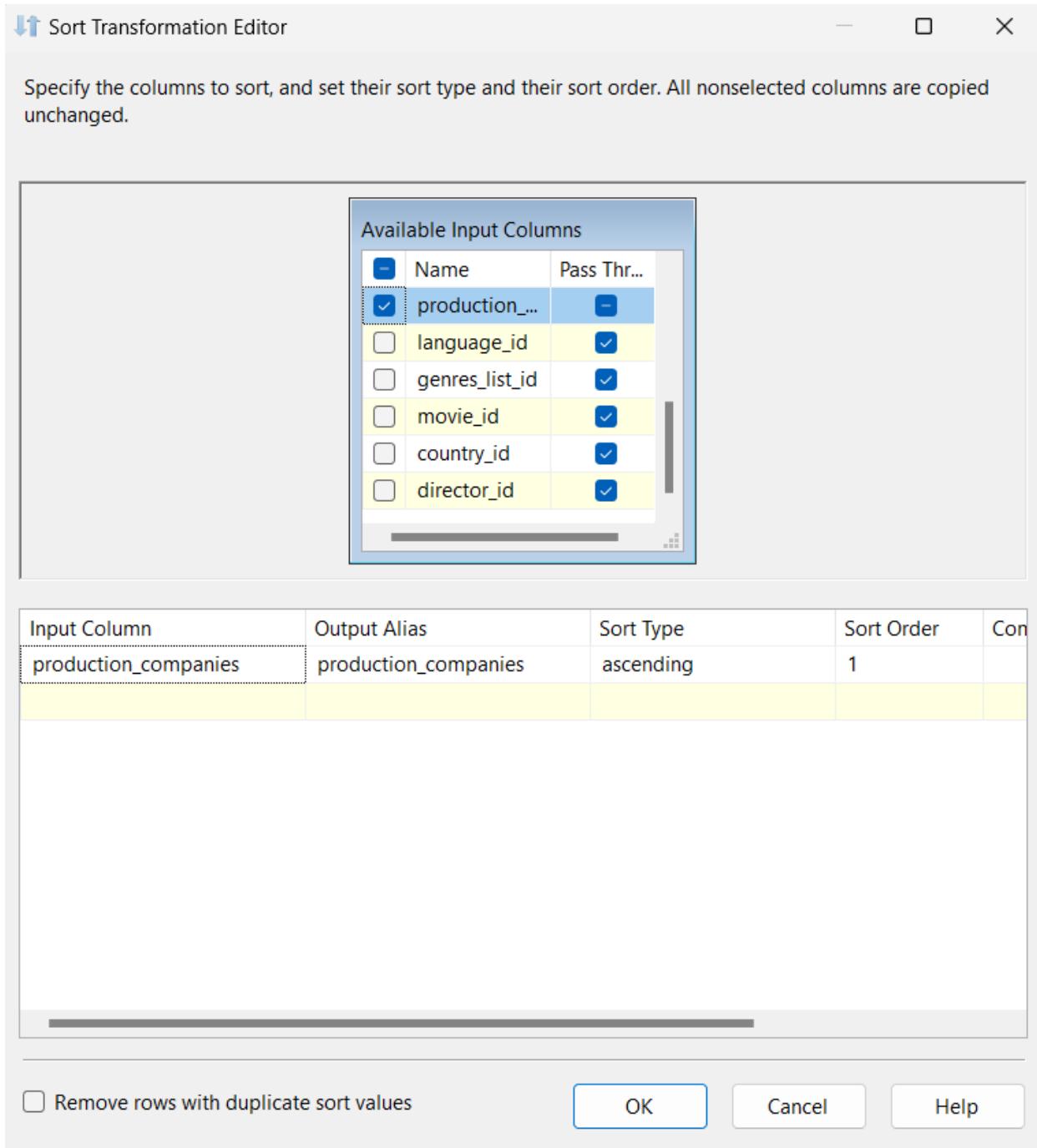
Hình 2.69 Tạo OLE DB Source cho Dim_Company

Bước 36: Mở **Dim_Company** Mở **Edit**, chọn **Dim_Company** để đổ dữ liệu vào **OLE DB Source Dim_Company** vừa tạo.



Hình 2.70 Chọn Dim_Company để đổ dữ liệu vào OLE DB Source Dim_Company

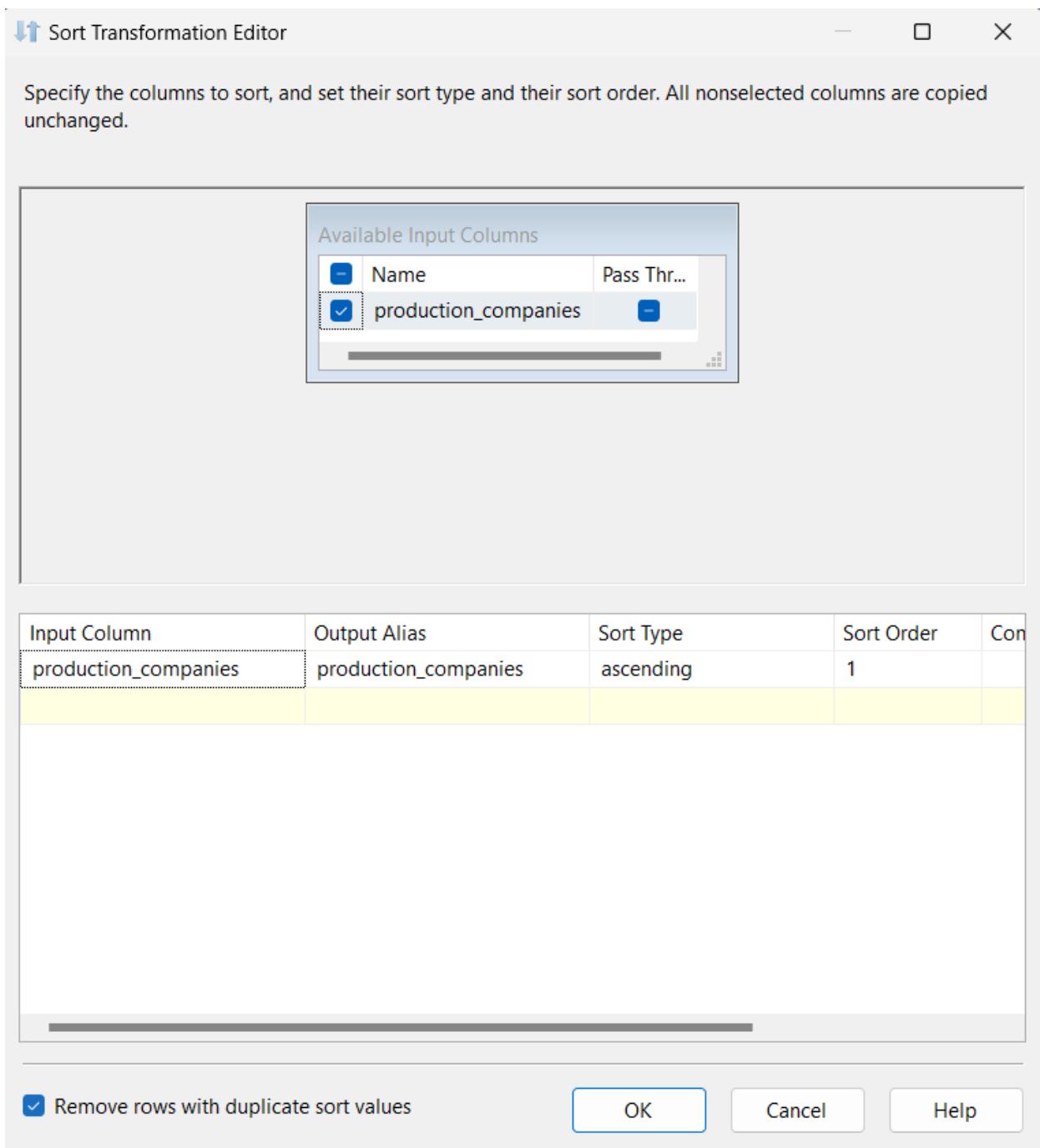
Bước 37: Ở Sort10, click chuột phải chọn **Edit** và chọn cột **production_companies** để chuẩn bị cho quá trình merge.



Hình 2.71 chọn cột production_companies để đổ dữ liệu vào Sort10

Bước 38: Tạo Merge Join 5 và nối với Sort10, tiếp theo chọn **Merge Join Left Input** để giữ lại toàn bộ các dòng trong bảng **Merge Join 5**.

Bước 39: Tương tự ta chọn cột **production_companies** cho Sort11



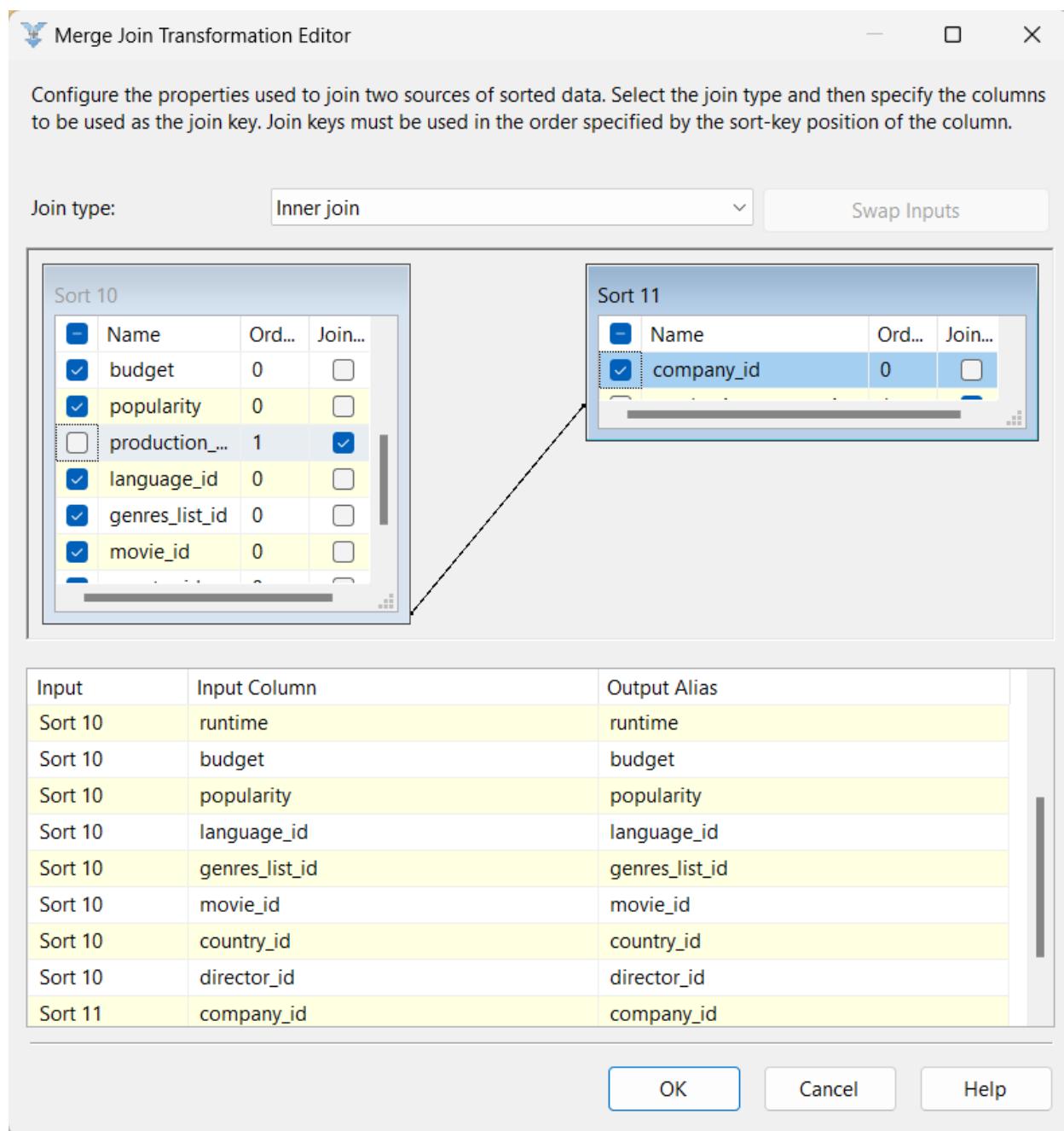
Hình 2.72 chọn cột production_companies để đổ dữ liệu vào Sort11

Bước 40: Nối Sort11 với Merge Join 5

Chuột phải vào **Merge Join 5** và nhấn **Edit**, ở đây ta tick chọn tất cả các cột của **Sort10** nhưng không lấy thuộc tính **production_companies**.

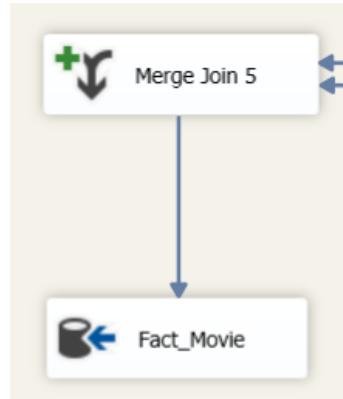
Tiếp theo ta chọn **company_id** ở **Sort11** để merge vào **Fact Raw**.

Kết quả sau khi merge là bảng Fact_Raw không còn thuộc tính **production_companies** và có thêm 1 thuộc tính mới là **company_id**.



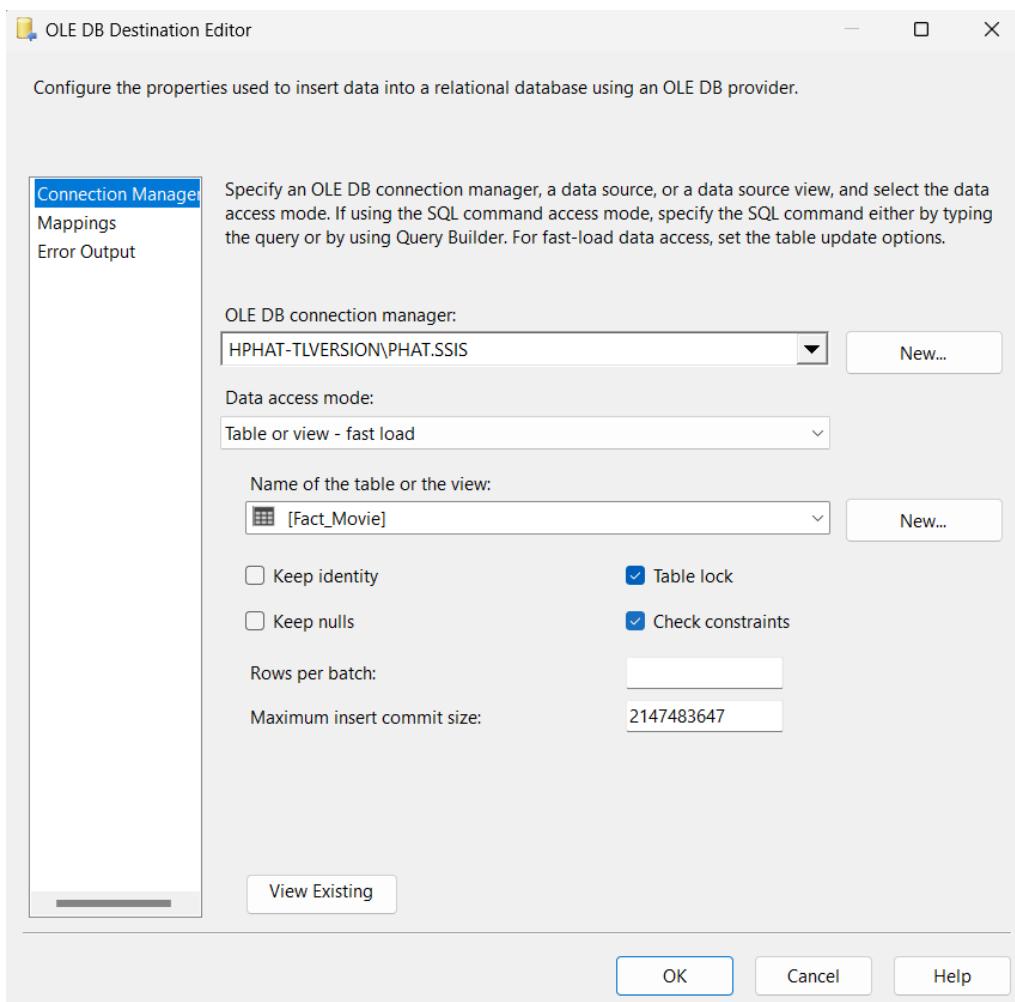
Hình 2.73 Merge Join Fact Raw với Dim Company

Bước 41: Tạo **OLE DB Destination** và đặt tên là **Fact_Movie**. Ta nối **Merge Join 5** vào **Fact_Movie**.



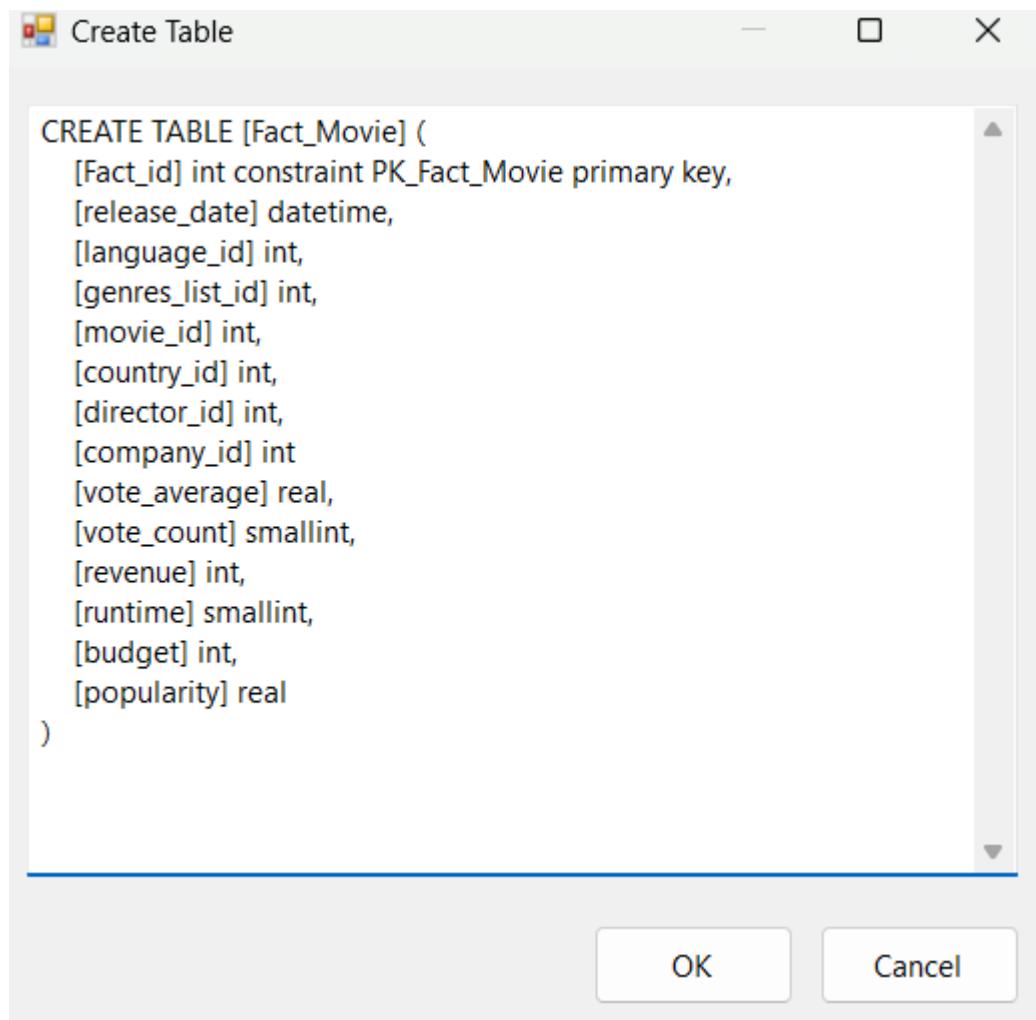
Hình 2.74 Tạo OLE DB Destination “Fact_Movie”

Bước 42: Chọn **Edit Fact_Movie**. Kiểm tra xem là kết nối đúng MS SQL Server không.



Hình 2.75 Kiểm tra kết nối server của Fact_Movie

Bước 43: Nhấn **New** để tạo bảng Fact Movie.

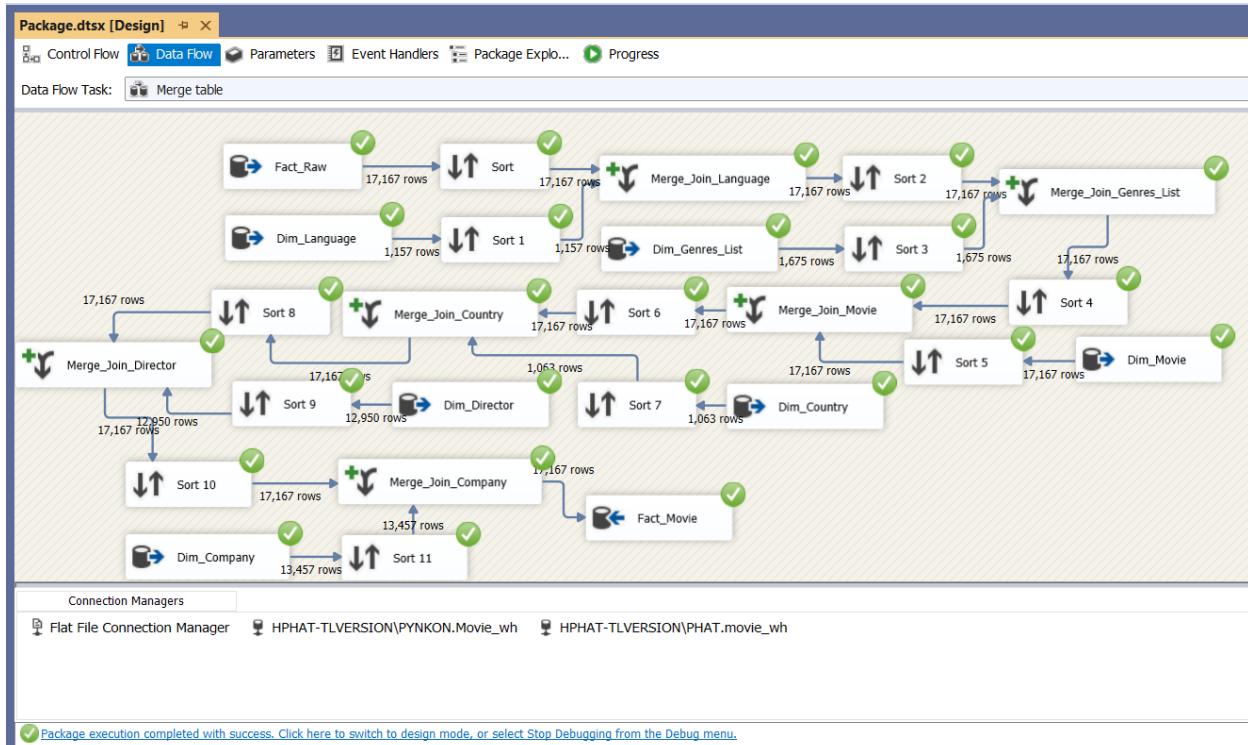


Hình 2.76 Tạo bảng Fact Movie

Bước 44: Tiếp đến ta cần chọn mục **Mappings** để xem xét việc ánh xạ các cột dữ liệu.
Cuối cùng nhấn nút **OK** để hoàn tất quá trình tạo bảng.

Kết quả sau khi thực thi khối **Merge table**

Kho dữ liệu và OLAP - IS217.P12



Hình 2.77 Kết quả thực thi khối Merge table

2.4.9 Tạo các khối lệnh SQL

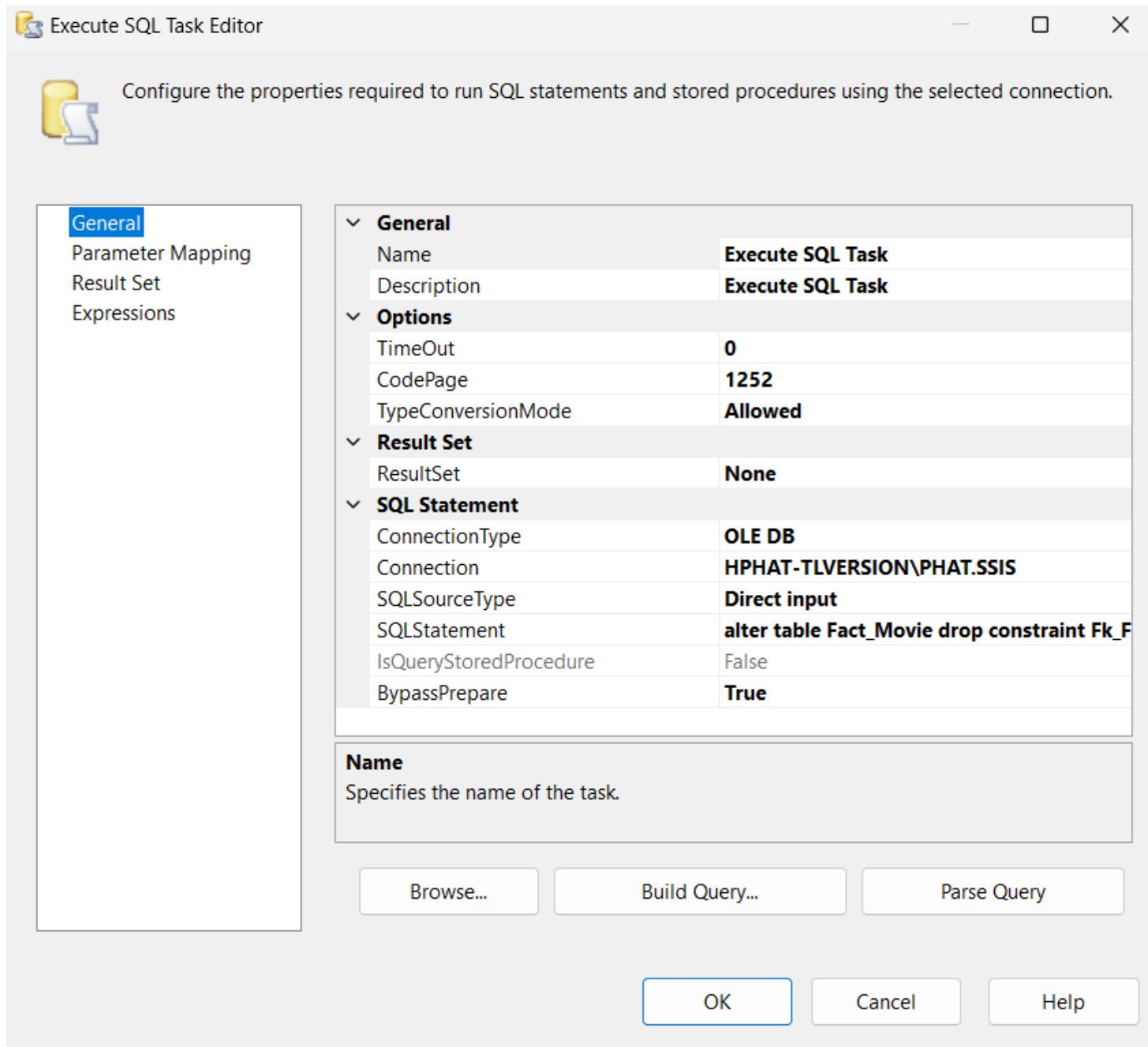
2.4.9.1 Tạo lệnh SQL để reset môi khi chạy SSIS

Bước 1: Tạo khối Execute SQL Task và kết nối với khối Data Flow Task.



Hình 2.78 Tạo khối Execute SQL Task

Bước 2. Nhấn chuột phải vào khối **Execute SQL Task** này và chọn **Edit**. Ở ô **Connection**, chọn connection đã thiết lập đến data warehouse trong SQL Server.

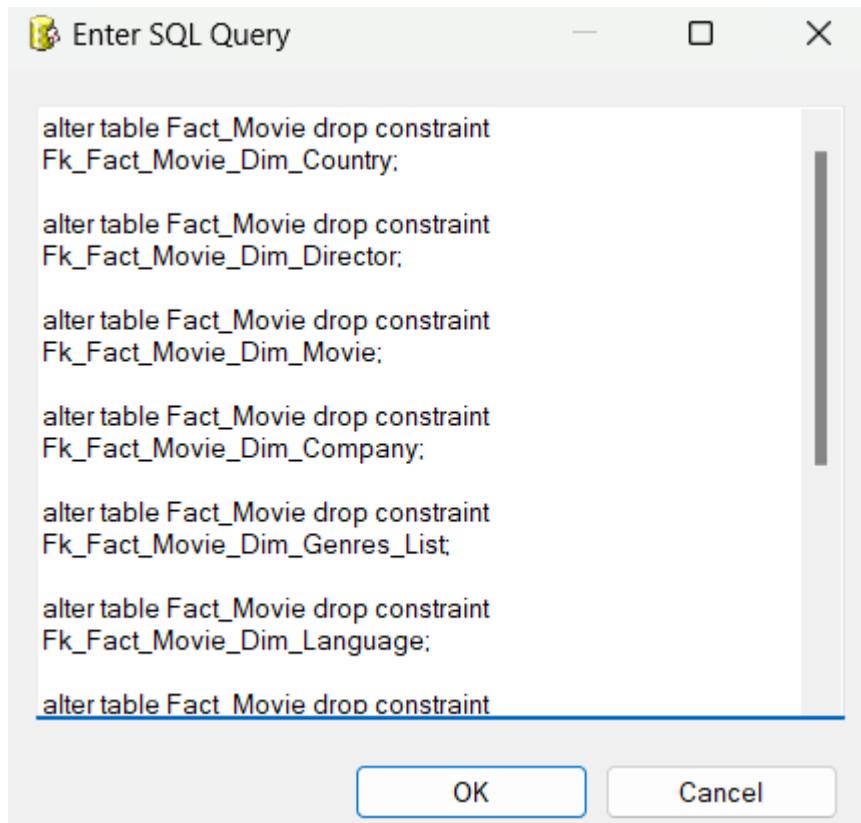


Hình 2.79 Kết nối đến data warehouse trong SQL Server

Bước 3: Ở ô **SQLStatement**, thêm các câu truy vấn SQL thực hiện reset mỗi khi chạy chương trình SSIS.

Nhấn **OK** để hoàn tất quá trình.

SVTH: Nguyễn Hồng Phát



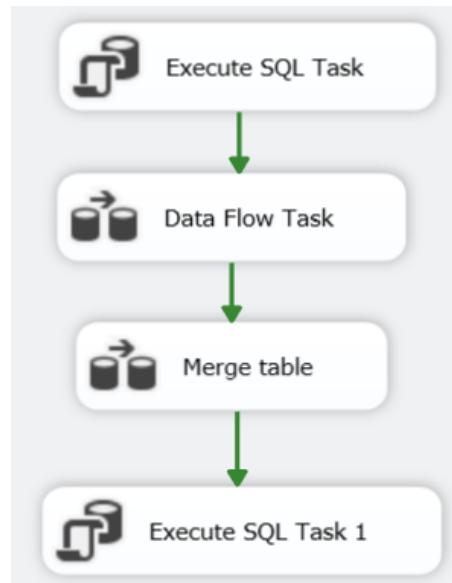
```
alter table Fact_Movie drop constraint Fk_Fact_Movie_Dim_Country;
alter table Fact_Movie drop constraint Fk_Fact_Movie_Dim_Director;
alter table Fact_Movie drop constraint Fk_Fact_Movie_Dim_Movie;
alter table Fact_Movie drop constraint Fk_Fact_Movie_Dim_Company;
alter table Fact_Movie drop constraint Fk_Fact_Movie_Dim_Genres_List;
alter table Fact_Movie drop constraint Fk_Fact_Movie_Dim_Language;
alter table Fact_Movie drop constraint
```

OK Cancel

Hình 2.80 Lệnh SQL reset

2.4.9.2 Tạo khóa ngoại từ các Dimension đến Fact Movie

Bước 1: Tạo khối Execute SQL Task 1 và kết nối Merge table với Execute SQL Task 1 .



Hình 2.81 Tạo khối Execute SQL Task 1

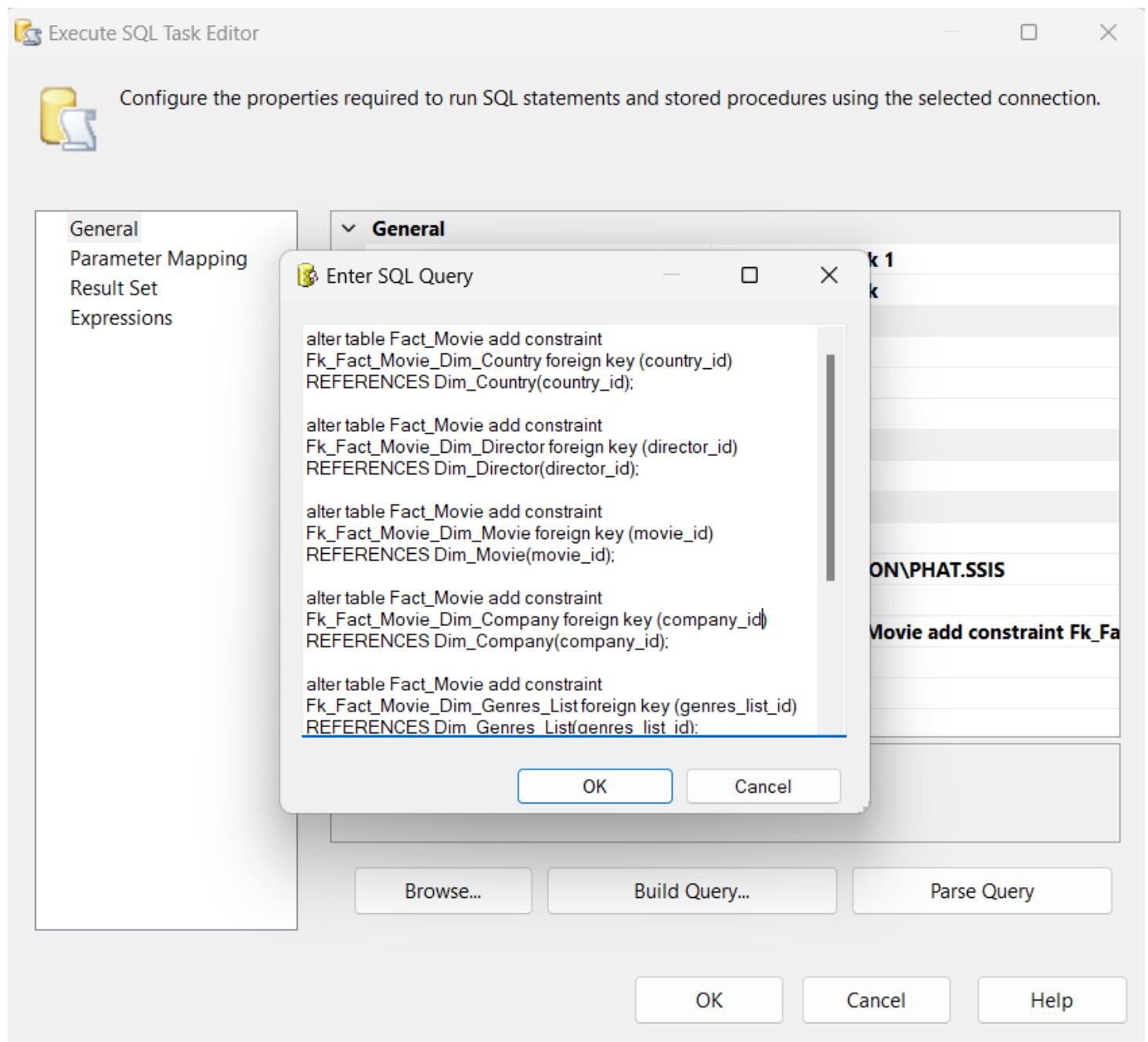
Bước 2. Nhấn chuột phải vào khối Execute SQL Task 1 này và chọn Edit. Ở ô Connection,

Kho dữ liệu và OLAP - IS217.P12

chọn connection đã thiết lập đến data warehouse trong SQL Server.

Bước 3: Ở ô SQLStatement, thêm các câu truy vấn SQL thực hiện tạo khóa ngoại từ các Dimension đến Fact Movie

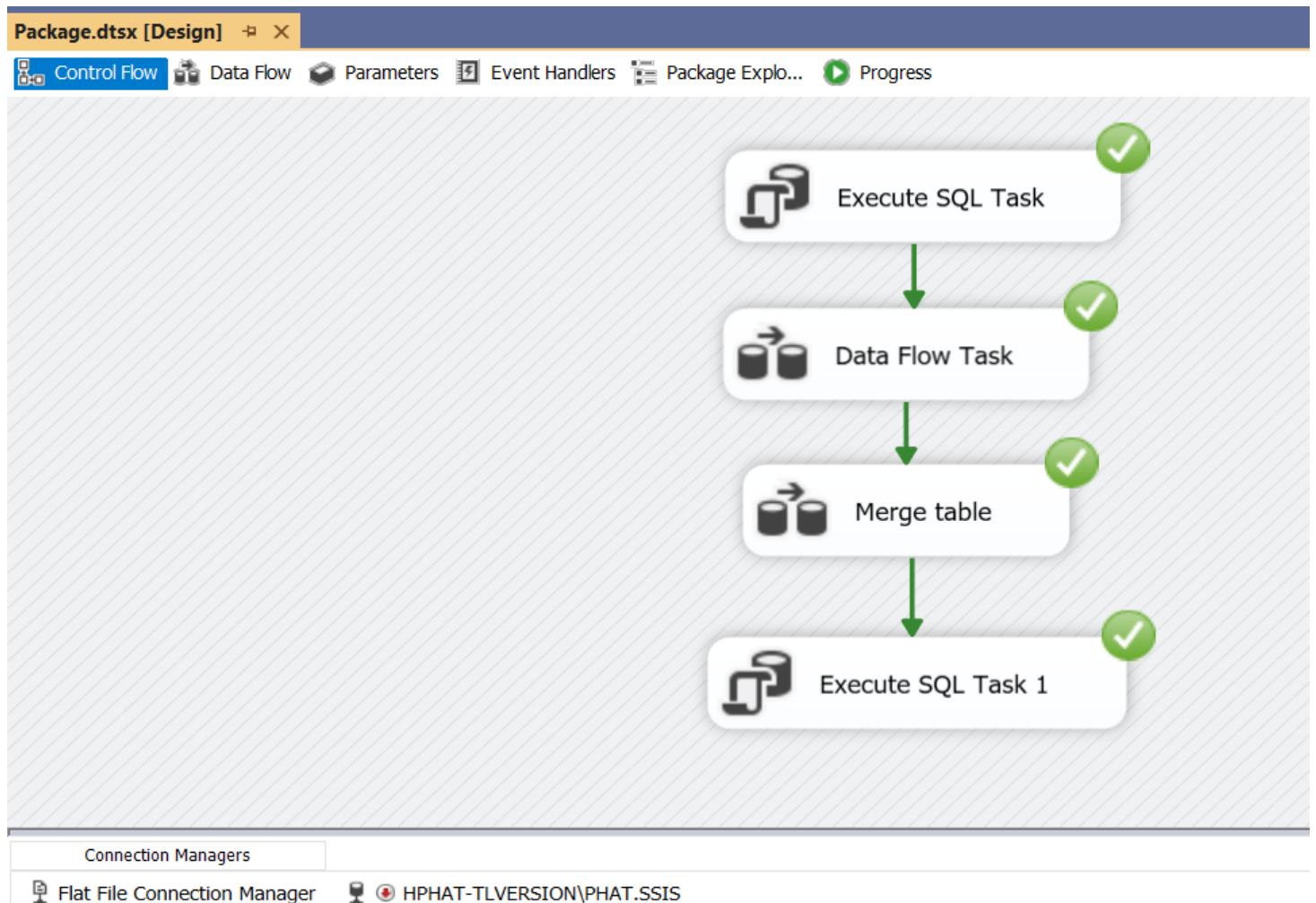
Nhấn **OK** để hoàn tất quá trình.



Hình 2.82 Lệnh SQL tạo khóa ngoại

2.4.10 Thực thi project và kết quả SSIS

Nhấn nút Start trên thanh menu để tiến hành thực thi project :



Hình 2.83 Kết quả thực thi project

2.4.11 Kiểm tra dữ liệu các bảng

2.4.11.1 Kiểm tra dữ liệu bảng Movie

	movie_id	title	status	imdb_id
1	1	#1 Cheerleader Camp	Released	tt1637976
2	2	#eagoraoque	Released	tt13368114
3	3	#FollowFriday	Released	tt5233106
4	4	#Jowable	Released	tt10850892
5	5	#MyEscape	Released	tt6390434
6	6	#Stuck	Released	tt2075318
7	7	\$5 a Day	Released	tt1024733
8	8	& Jara Hatke	Released	tt7288662
9	9	(A) Typical Couple	Released	tt2645592
10	10	(If I Can Sing a Song About) Ligatures	Released	tt1925394
11	11	(NULL)	Released	tt3572736
12	12	(Romance) in the Digital Age	Released	tt5525418
13	13	... e se domani	Released	tt0438466
14	14	...So Goes the Nation	Released	tt0432337
15	15	/andragogy./	Released	tt23643918
16	16	@home	Released	tt2772832

Hình 2.84 Dữ liệu bảng Dim Movie

2.4.11.2 Kiểm tra dữ liệu bảng Country

	country_id	production_countries
1	1	Afghanistan
2	2	Afghanistan, Iraq, Italy, Syrian Arab Republic, Unit...
3	3	Afghanistan, United Kingdom
4	4	Albania
5	5	Albania, Australia
6	6	Albania, Cyprus, Greece
7	7	Albania, Denmark, Italy, United Kingdom, United ...
8	8	Albania, France, Austria
9	9	Albania, Germany
10	10	Albania, Greece, Serbia

Hình 2.85 Dữ liệu bảng Dim Country

2.4.11.3 Kiểm tra dữ liệu bảng Company

	company_id	production_companies
1	1	#1NFLUENCE Production
2	2	#LetsDoeit
3	3	(i£¼)i–iŠ¤é¹..i–i–T'
4	4	[f.u.c.]Film, Penny Lane Film, Beleza Film, HIMHI...
5	5	+1 Filmes
6	6	01 Start Adam Ustynowicz
7	7	01010101 films, Coletivo BinÅ¡rio
8	8	011 Productions
9	9	0708 Films
10	10	1 Production Film
11	11	1/27 Pictures
12	12	1:1.3 Entertainments
13	13	1+1 Production
14	14	10 West Studios, A Really Good Home Pictures
15	15	10:15! Productions, Tripode Productions
16	16	100% Halal

Hình 2.86 Dữ liệu bảng Dim Company

2.4.11.4 Kiểm tra dữ liệu bảng Director

	director_id	Director
1	1	(LA)HORDE, Marine Brutti, Arthur Harel, Jonathan ...
2	2	A Couple' A Cowboys
3	3	A K Lohithadas
4	4	A K Sajan
5	5	Å arÅ«nas Bartas
6	6	Å emsudin RadonÅž
7	7	Å pela ÅŒadeÅ%
8	8	Å tefan VorÅ¾Åjek
9	9	A. Bhimsingh
10	10	A. L. Vijay
11	11	A. Venkatesh
12	12	A.D. Calvo
13	13	A.J. Hall
14	14	A.J. Serrano
15	15	A.S. Ravi Kumar Chowdary
16	16	A.V. Bramble
17	17	A.V. Rockwell

Hình 2.87 Dữ liệu bảng Dim Director

2.4.11.5 Kiểm tra dữ liệu bảng Date

	release_date	Day	Month	Year
1	2000-01-01 00:00:00.000	1	1	2000
2	2000-01-02 00:00:00.000	2	1	2000
3	2000-01-08 00:00:00.000	8	1	2000
4	2000-01-09 00:00:00.000	9	1	2000
5	2000-01-14 00:00:00.000	14	1	2000
6	2000-01-15 00:00:00.000	15	1	2000
7	2000-01-18 00:00:00.000	18	1	2000
8	2000-01-20 00:00:00.000	20	1	2000
9	2000-01-25 00:00:00.000	25	1	2000
10	2000-01-28 00:00:00.000	28	1	2000
11	2000-01-29 00:00:00.000	29	1	2000
12	2000-02-01 00:00:00.000	1	2	2000
13	2000-02-03 00:00:00.000	3	2	2000
14	2000-02-05 00:00:00.000	5	2	2000
15	2000-02-07 00:00:00.000	7	2	2000
16	2000-02-11 00:00:00.000	11	2	2000

Hình 2.88 Dữ liệu bảng Dim Date

2.4.11.6 Kiểm tra dữ liệu bảng Genres List

	genres_list_id	genres_list
1	1	['Action', 'Action', 'Crime']
2	2	['Action', 'Adventure', 'Animation', 'Drama']
3	3	['Action', 'Adventure', 'Animation', 'Family', 'Co...']
4	4	['Action', 'Adventure', 'Animation', 'Family', 'Fan...']
5	5	['Action', 'Adventure', 'Animation', 'Family']
6	6	['Action', 'Adventure', 'Animation', 'Fantasy', 'Co...']
7	7	['Action', 'Adventure', 'Animation', 'Fantasy']
8	8	['Action', 'Adventure', 'Animation']
9	9	['Action', 'Adventure', 'Comedy', 'Crime', 'Anima...']
10	10	['Action', 'Adventure', 'Comedy', 'Drama', 'Anim...']
11	11	['Action', 'Adventure', 'Comedy', 'Drama', 'Fami...']
12	12	['Action', 'Adventure', 'Comedy', 'Fantasy', 'Ro...']
13	13	['Action', 'Adventure', 'Comedy', 'Horror', 'Scienc...']
14	14	['Action', 'Adventure', 'Comedy', 'Science Fictio...']
15	15	['Action', 'Adventure', 'Comedy']
16	16	['Action', 'Adventure', 'Crime', 'Drama', 'Mystery...']

Hình 2.89 Dữ liệu bảng Dim Genres List

2.4.11.7 Kiểm tra dữ liệu bảng Language

	language_id	original_language	spoken_languages
1	1	ab	English
2	2	af	Afrikaans
3	3	ak	Akan, English
4	4	am	Amharic
5	5	ar	Arabic
6	6	ar	Arabic, English
7	7	ar	Arabic, French
8	8	ar	Arabic, Greek
9	9	ar	Arabic, Hebrew
10	10	ar	Arabic, Hebrew, English
11	11	ar	Arabic, Spanish
12	12	ar	English, Arabic
13	13	ar	English, Turkish, Arabic

Hình 2.90 Dữ liệu bảng Dim Language

2.4.11.8 Kiểm tra dữ liệu bảng Fact Movie

	Fact_id	release_date	vote_average	vote_count	revenue	runtime	budget	popularity	language_id	genres_list_id	movie_id	country_id	director_id	company_id
1	83	2003-10-26 00:00:00.000	5.508	981	54667954	79	130000	12.827	228	934	10497	1028	7597	9367
2	189	2014-08-20 00:00:00.000	6.4	3539	39407616	102	65000000	52.627	228	636	12518	1028	8537	10028
3	286	2006-10-03 00:00:00.000	6.826	46	0	110	0	2.468	115	760	5527	540	431	7065
4	1596	2006-07-04 00:00:00.000	5.831	239	0	88	0	10.545	228	1225	11111	228	4481	9066
5	1647	2003-01-31 00:00:00.000	6.348	1580	101191884	115	46000000	13.178	326	177	14843	1028	10344	1715
6	1956	2002-09-20 00:00:00.000	6.004	275	0	103	3500000	8.12	228	784	5528	69	10568	3926
7	2029	2001-11-21 00:00:00.000	5.952	427	24272365	108	0	9.689	599	627	13334	439	1836	11573
8	2084	2001-09-06 00:00:00.000	5.873	366	0	93	13000000	13.52	335	499	2241	1024	12593	5304
9	2096	2000-08-08 00:00:00.000	4.6	10	0	94	0	2.436	228	159	15992	1024	232	9248
10	2100	2001-10-19 00:00:00.000	7.14	1123	27642707	131	72000000	19.533	228	110	14350	1028	1626	3598
11	2144	2001-04-19 00:00:00.000	5.815	311	13500000	93	18000000	7.112	228	501	10438	1028	3327	12569
12	2332	2000-03-25 00:00:00.000	6.231	1583	64400000	82	10500000	17.224	632	74	13376	439	5854	984
13	2859	2007-03-09 00:00:00.000	6.5	21	0	88	0	0.84	115	627	5224	96	10465	2940
14	2882	2007-03-21 00:00:00.000	6.599	372	0	97	0	11.798	599	1368	6471	439	10442	11732
15	3152	2000-02-25 00:00:00.000	5.3	20	0	92	0	7.56	228	576	2150	1028	8324	9971
16	3178	2000-01-29 00:00:00.000	5	21	0	93	0	5.751	228	962	1920	1028	7740	7507
17	3631	2005-05-11 00:00:00.000	5.611	27	875898	101	12000000	5.027	297	962	13346	678	10128	4237
18	3982	2007-09-04 00:00:00.000	6.1	27	0	98	0	2.94	747	668	15116	698	4194	5847
19	4204	2008-09-06 00:00:00.000	5.568	74	0	98	0	4.452	228	806	7	1028	1146	2235
20	4213	2007-05-20 00:00:00.000	6.3	10	0	96	0	2.363	115	962	3413	540	808	5464
21	4252	2003-06-12 00:00:00.000	6.216	88	18000000	175	4600000	8.727	683	527	2927	649	10469	3599
22	4284	2005-12-04 00:00:00.000	3.8	8	0	88	0	1.177	228	1232	12042	1028	6493	11821
23	4896	2004-01-28 00:00:00.000	6.583	30	0	85	40000	1.53	100	535	11381	540	11307	10463
24	5206	2006-09-21 00:00:00.000	6.612	246	0	95	3000000	6.679	84	627	13703	381	8053	13391
25	5246	2003-11-28 00:00:00.000	0	0	0	90	0	1.091	228	1638	2399	1028	383	13091
26	5523	2006-11-26 00:00:00.000	8.6	6	0	92	0	1.339	228	941	6161	265	1971	4762

Hình 2.91 Dữ liệu bảng Fact Movie

CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU SSAS

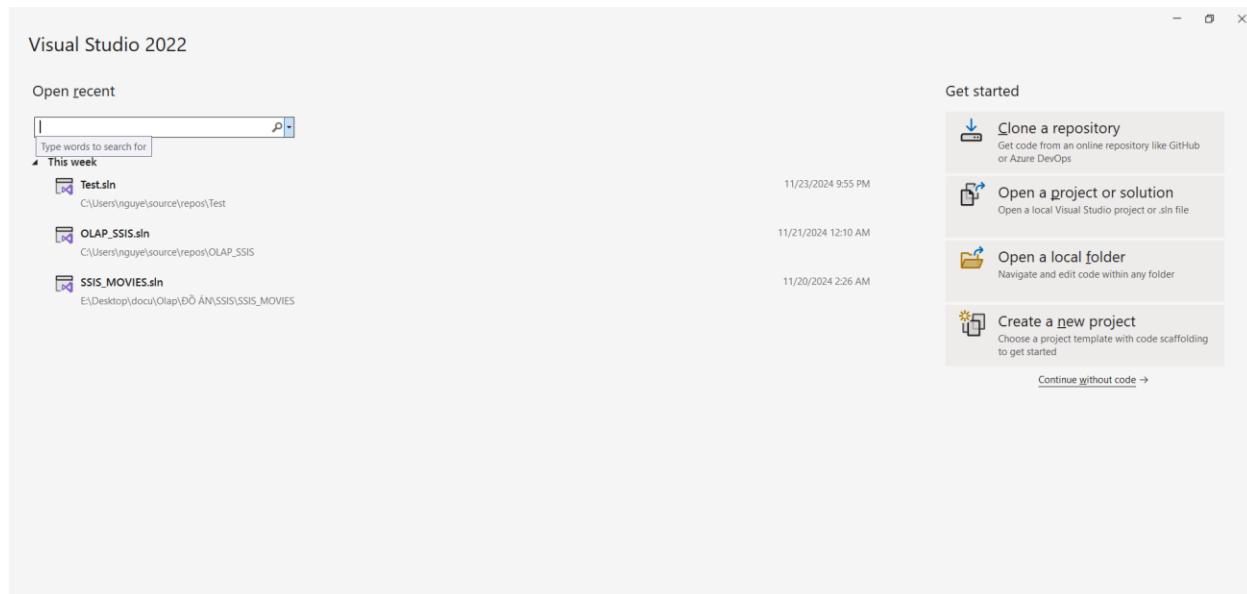
3.1 Chuẩn bị các công cụ

Để thực hiện được quá trình SSAS ta cần chuẩn bị và cài đặt các công cụ sau:

- + Microsoft SQL Server có cài đặt Analysis Services.
- + Microsoft Analysis Services Projects

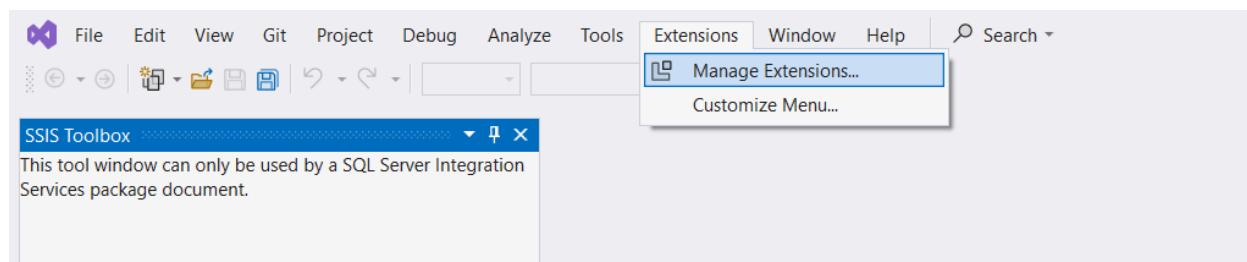
3.1.1 Cài đặt Microsoft Analysis Services Projects

Bước 1: Mở Visual Studio 2022 và chọn "Continue without code"



Hình 3.1 Mở Visual Studio 2022

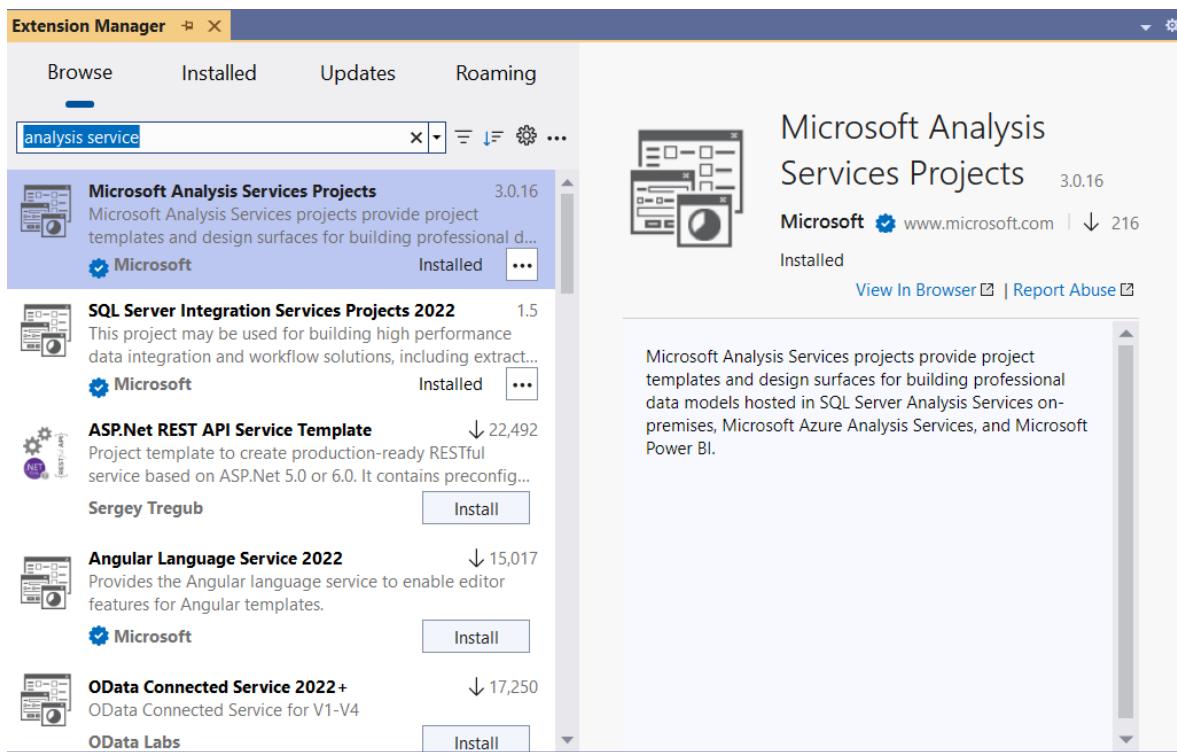
Bước 2: Trong giao diện chính , click chọn "Extensions" > "Manage Extensions"



Hình 3.2 Mở Manage Extensions

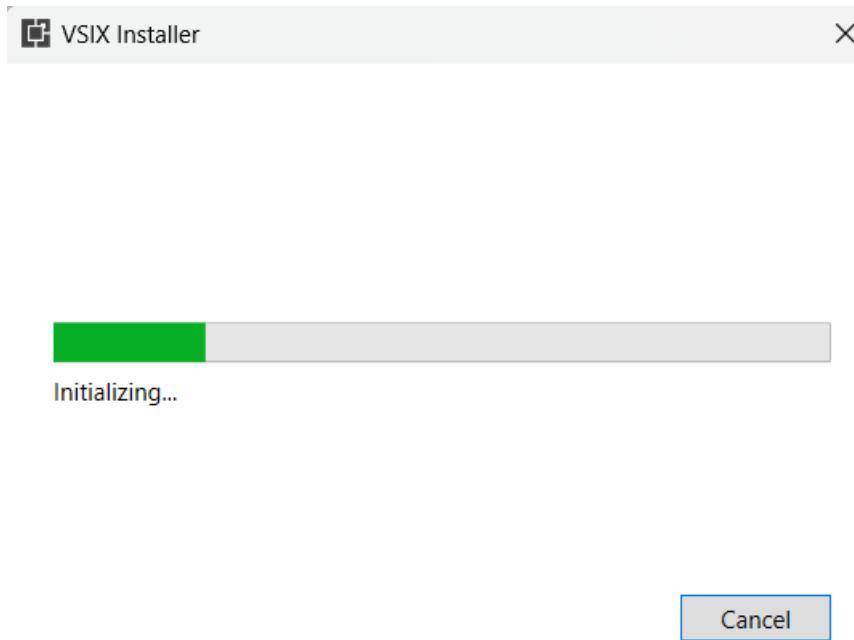
Bước 3: Tìm và tải về công cụ Microsoft Analysis Services Projects 2022.

Kho dữ liệu và OLAP - IS217.P12



Hình 3.3 Tải Microsoft Analysis Services Projects 2022

Bước 4: Sau khi cài đặt xong, ta cần đóng cửa sổ Visual Studio để công cụ được cập nhật vào Visual Studio. Sau khi đóng tất cả cửa sổ Visual Studio, hộp thoại VSIX Installer được khởi tạo.



Hình 3.4 VSIX Installer được khởi tạo

Bước 5: Sau khi đã khởi tạo xong, ta chọn Modify để đồng ý với các điều khoản khi cài đặt công SVTH: Nguyễn Hồng Phát

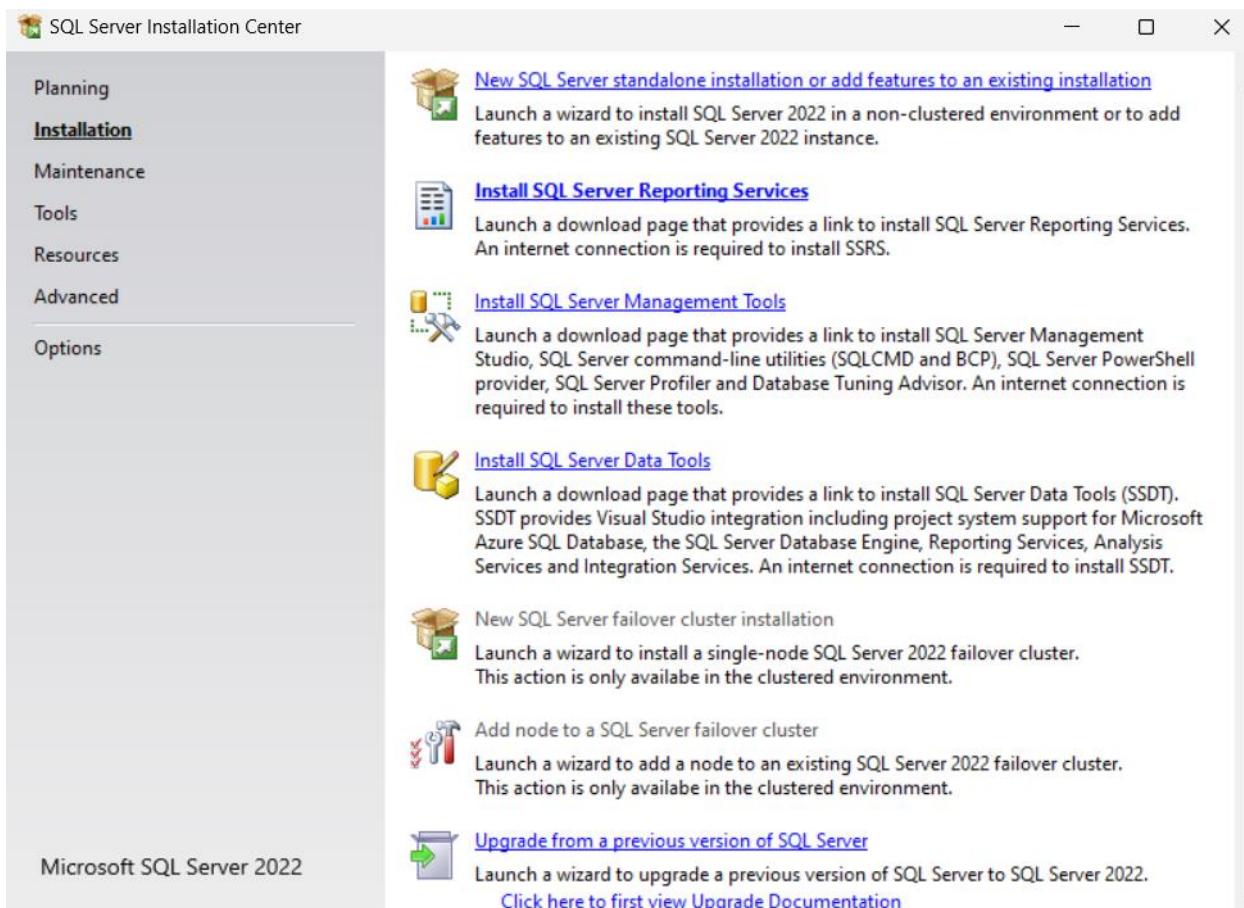
cụ Microsoft Analysis Services Projects.

Bước 6: Sau khi cài đặt hoàn tất, ta chọn Close để đóng cửa sổ VSIX Installer.

3.1.2 Cài đặt Analysis Services

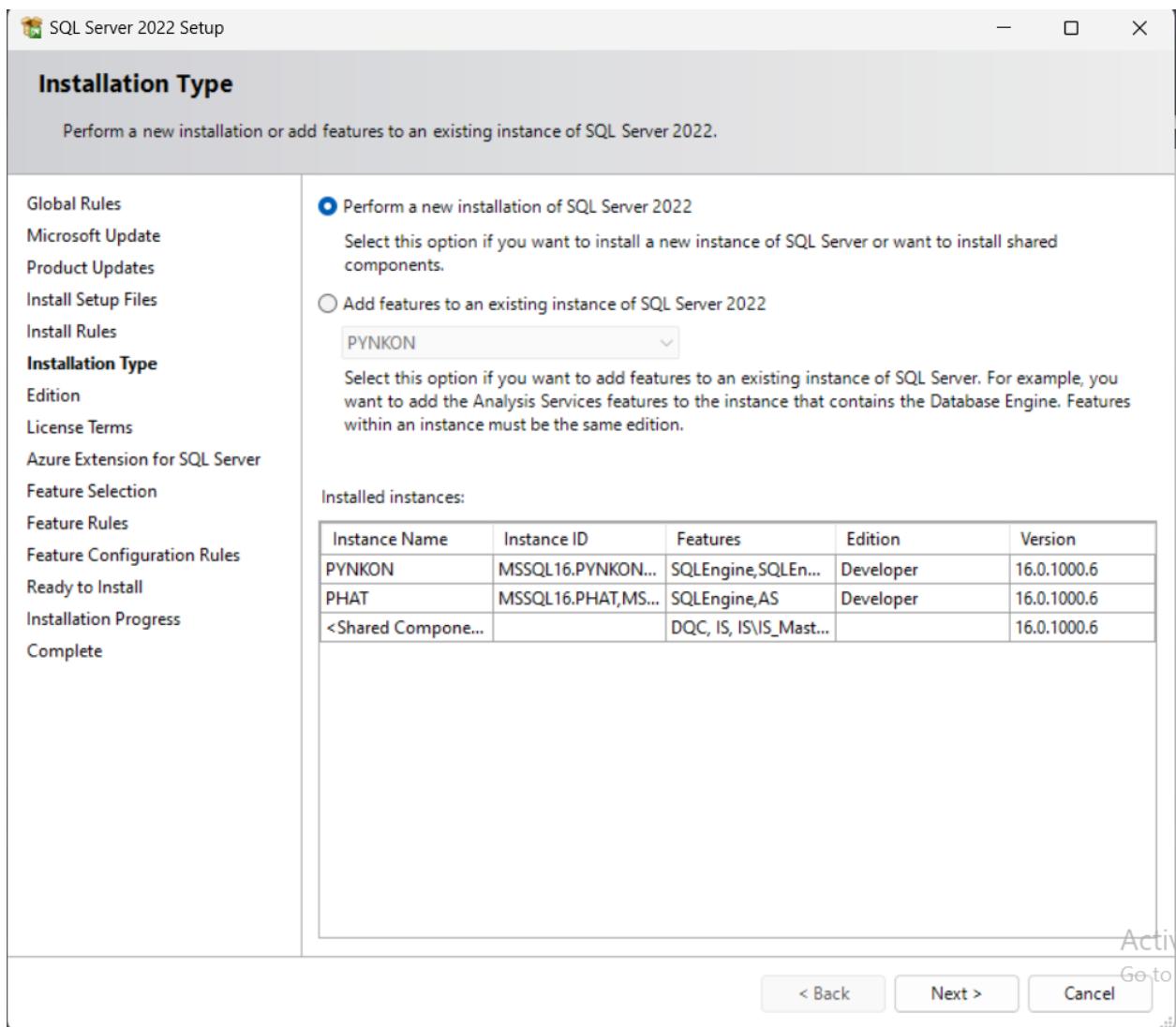
Bước 1: Mở SQL Server Installation Center.

Bước 2: Chọn mục **Installation**, tại mục Installation chọn **New SQL Server standalone or add features to an existing installation**.



Hình 3.5 Mở SQL Server standalone or add features to an existing installation

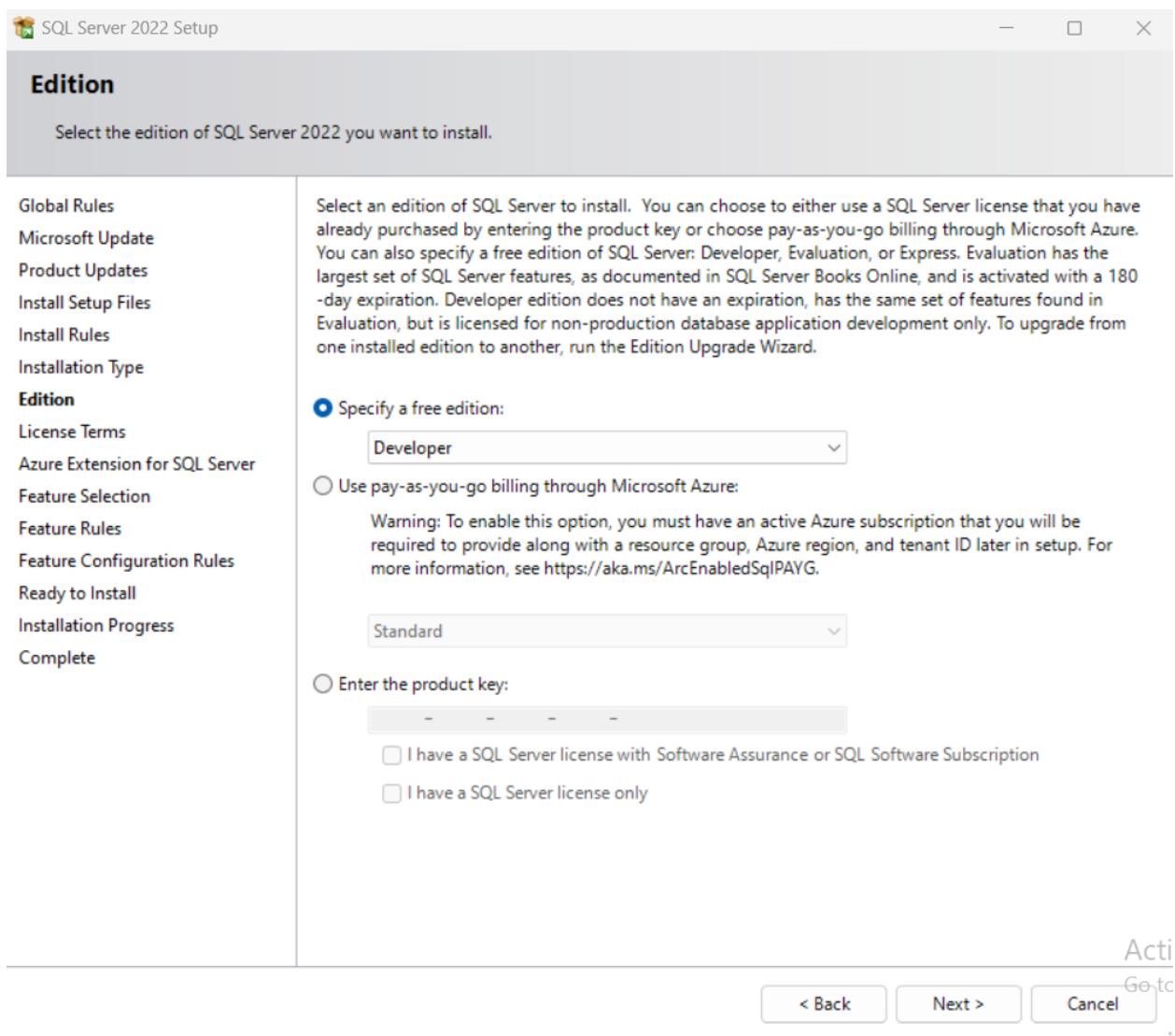
Bước 3: Hộp thoại mới xuất hiện, ta chọn **Next** đến tab **Installation Type**. Nếu mới cài MS SQL Server hoặc muốn tại một Server mới thì ta chọn **Perform a new installation of SQL Server 2022**. Còn không thì ta chọn **Add features to an existing instance of SQL Server 2022**. Sau đó tiếp tục chọn **Next**.



Hình 3.6 Tab Installation Type

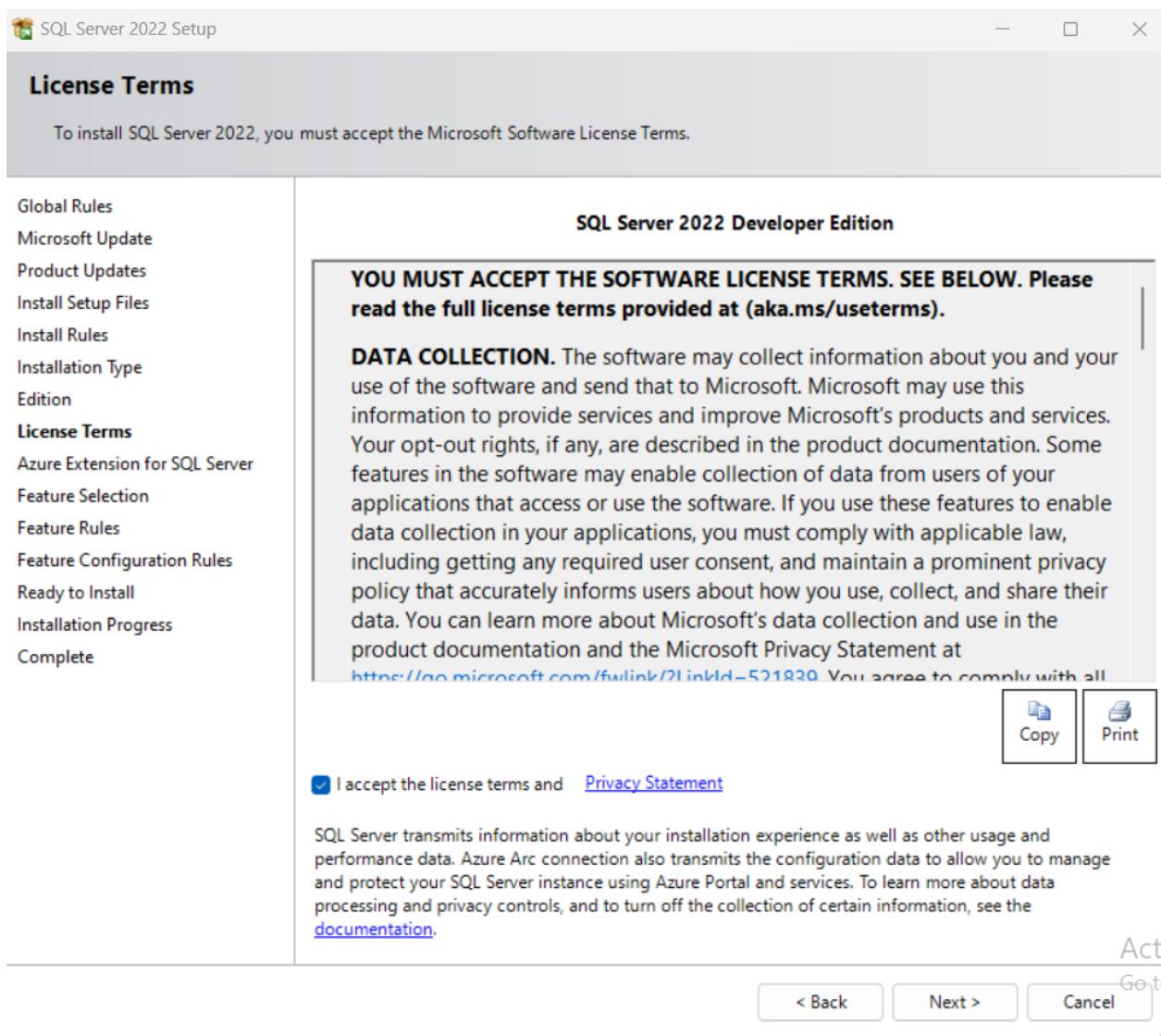
Bước 4: Tới tab **Edition**. Ở mục **Specify a free edition** ta chọn **Developer**. Sau đó tiếp tục chọn **Next**.

Kho dữ liệu và OLAP - IS217.P12



Hình 3.7 Tab Edition

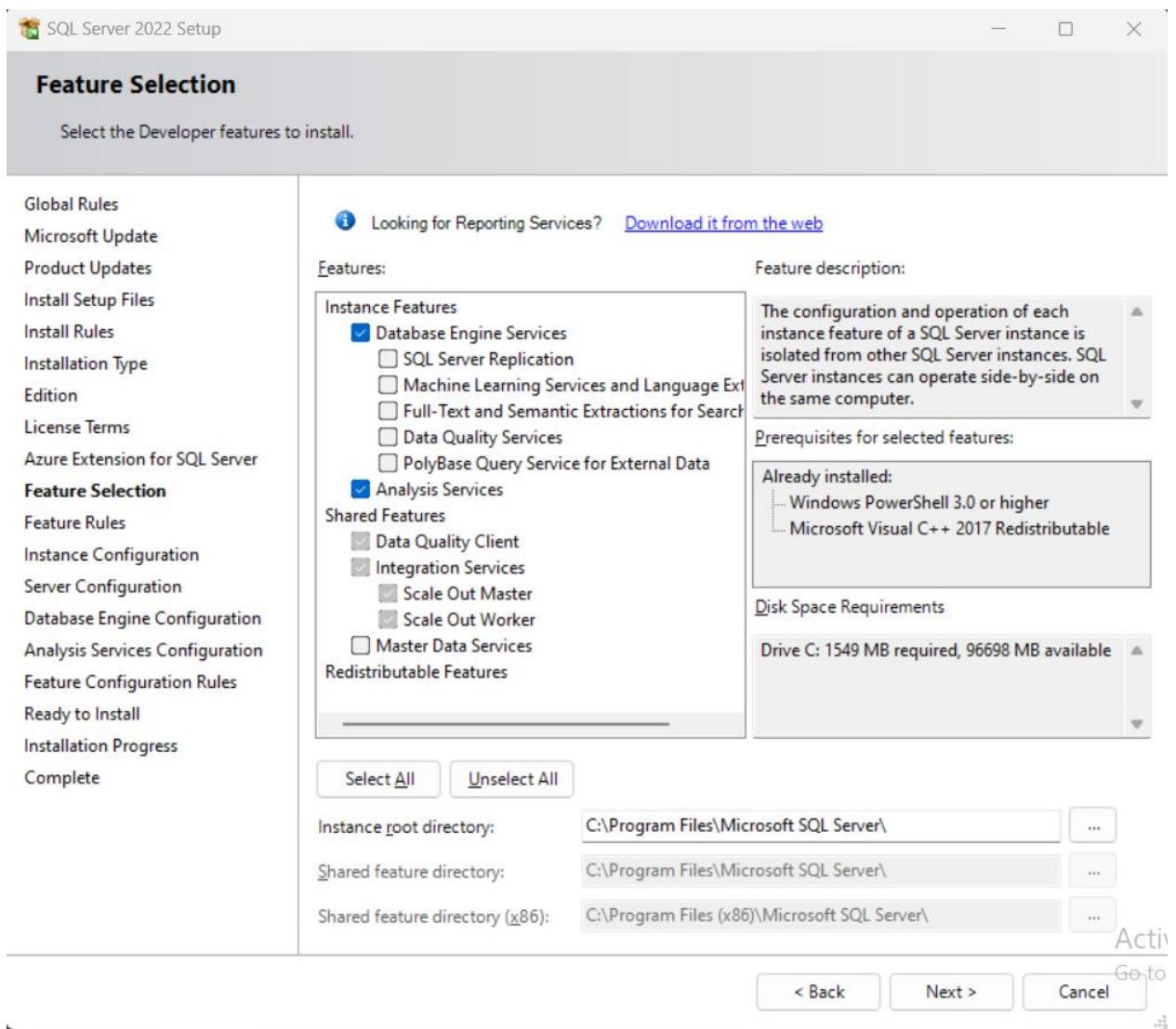
Bước 5: Ở tab License Terms, ta tích vào I accept the license terms and Privacy Statement. Sau đó tiếp tục chọn Next.



Hình 3.8 Tab License Terms

Bước 6: Ở tab Feature Selection ta chọn Database Engine Services và Analysis Services. Sau đó tiếp tục chọn Next.

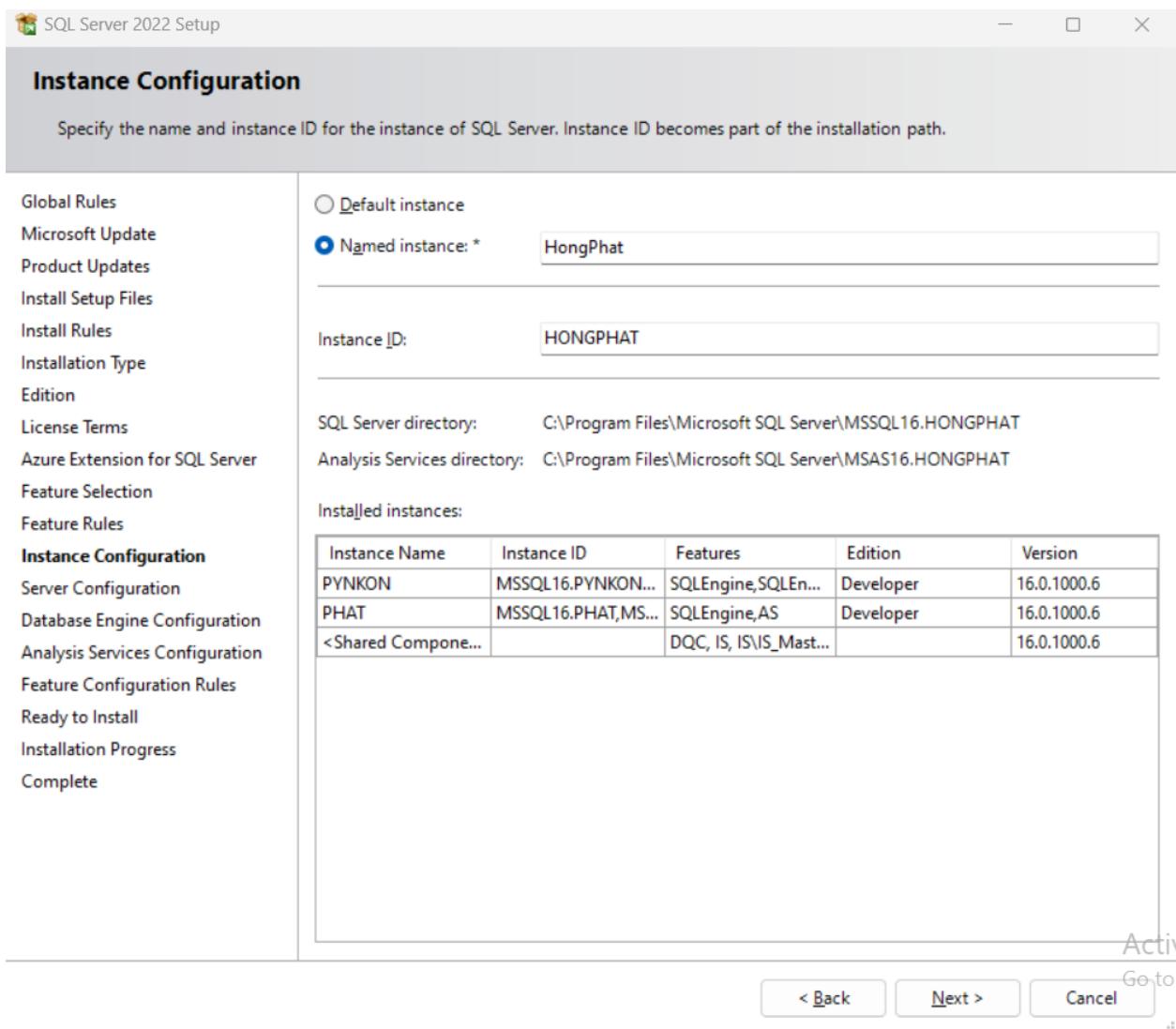
Kho dữ liệu và OLAP - IS217.P12



Hình 3.9 Tab Feature Selection

Bước 9: Chọn **Named instance**, sau đó ta đặt tên cho instance vừa tạo. Sau đó tiếp tục chọn **Next**.

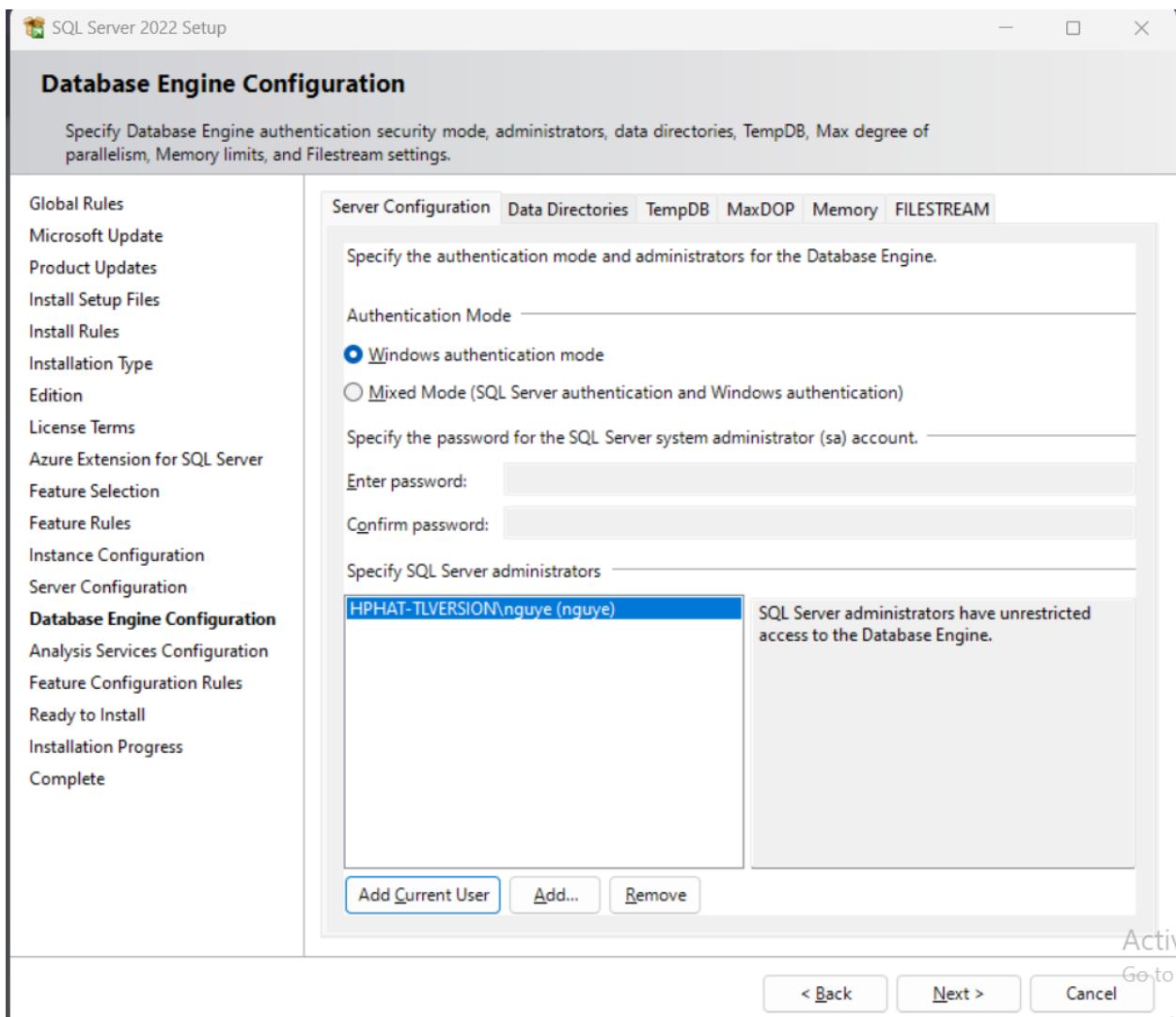
Kho dữ liệu và OLAP - IS217.P12



Hình 3.10 Tab Instance Configuration

Bước 10: Chọn Windows authentication mode. Tiếp theo, chọn Add Current User. Sau khi user xuất hiện, ta chọn vào tên user và chọn Next.

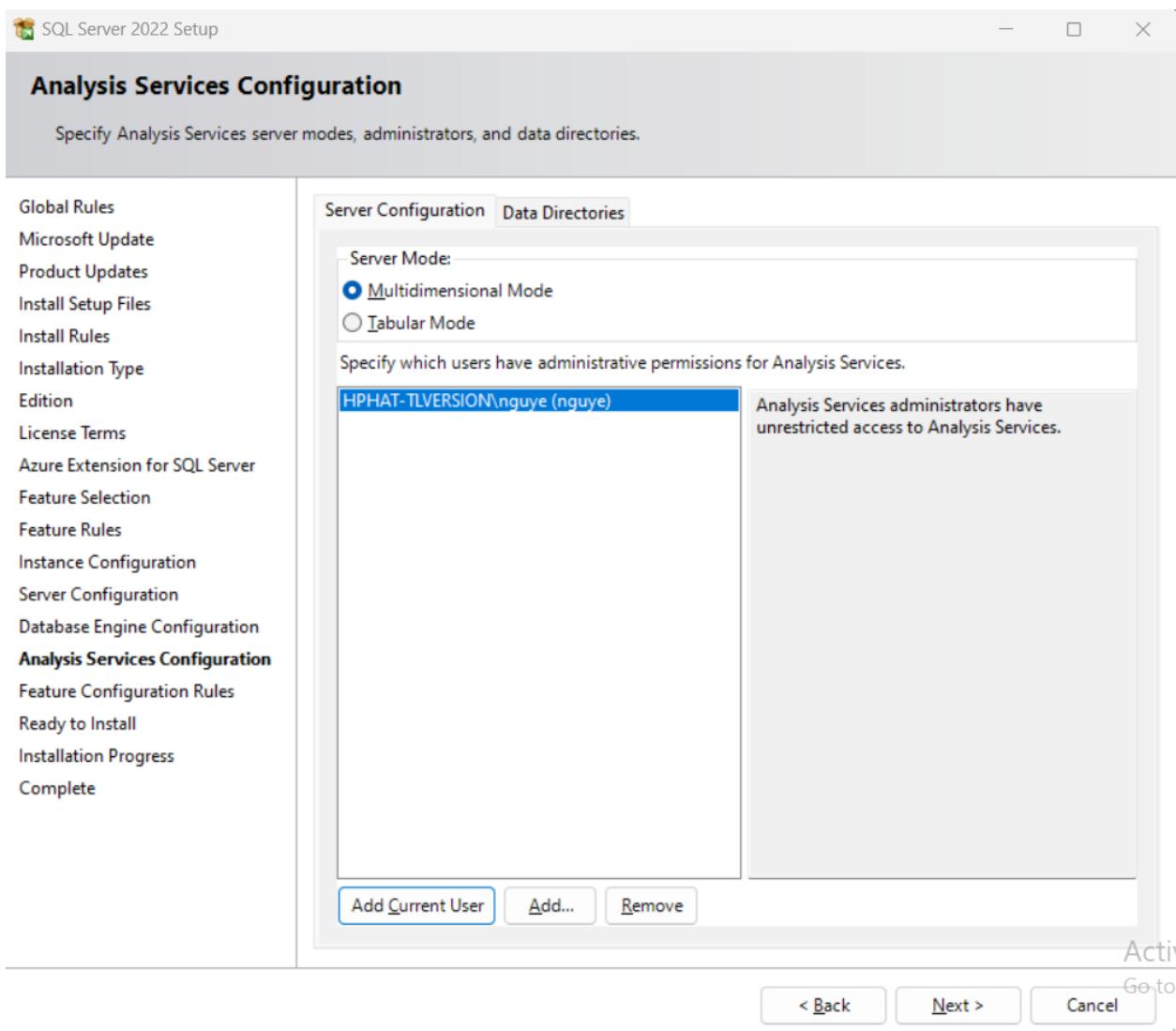
Kho dữ liệu và OLAP - IS217.P12



Hình 3.11 Tab Database Engine Configuration

Bước 11: Chọn **Multidimensional Mode**. Tiếp theo, chọn **Add Current User**. Sau khi user xuất hiện, ta chọn vào tên user và chọn **Next**.

Kho dữ liệu và OLAP - IS217.P12



Hình 3.12 Tab Analysis Services Configuration

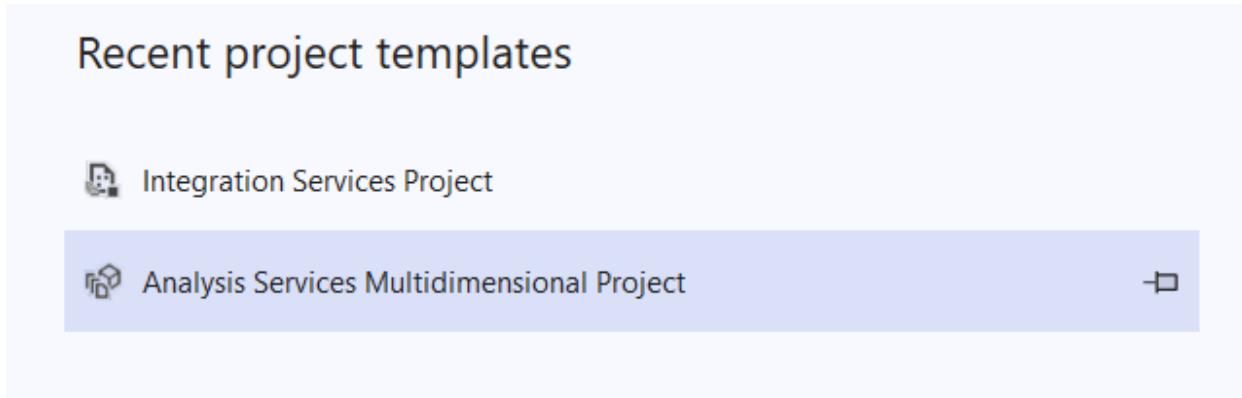
Bước 12: Kiểm tra lại các cài đặt và chọn **Install** để tải.

Bước 13: Chọn **Close** để hoàn tất cài đặt.

3.2 Tạo mới Project SSAS

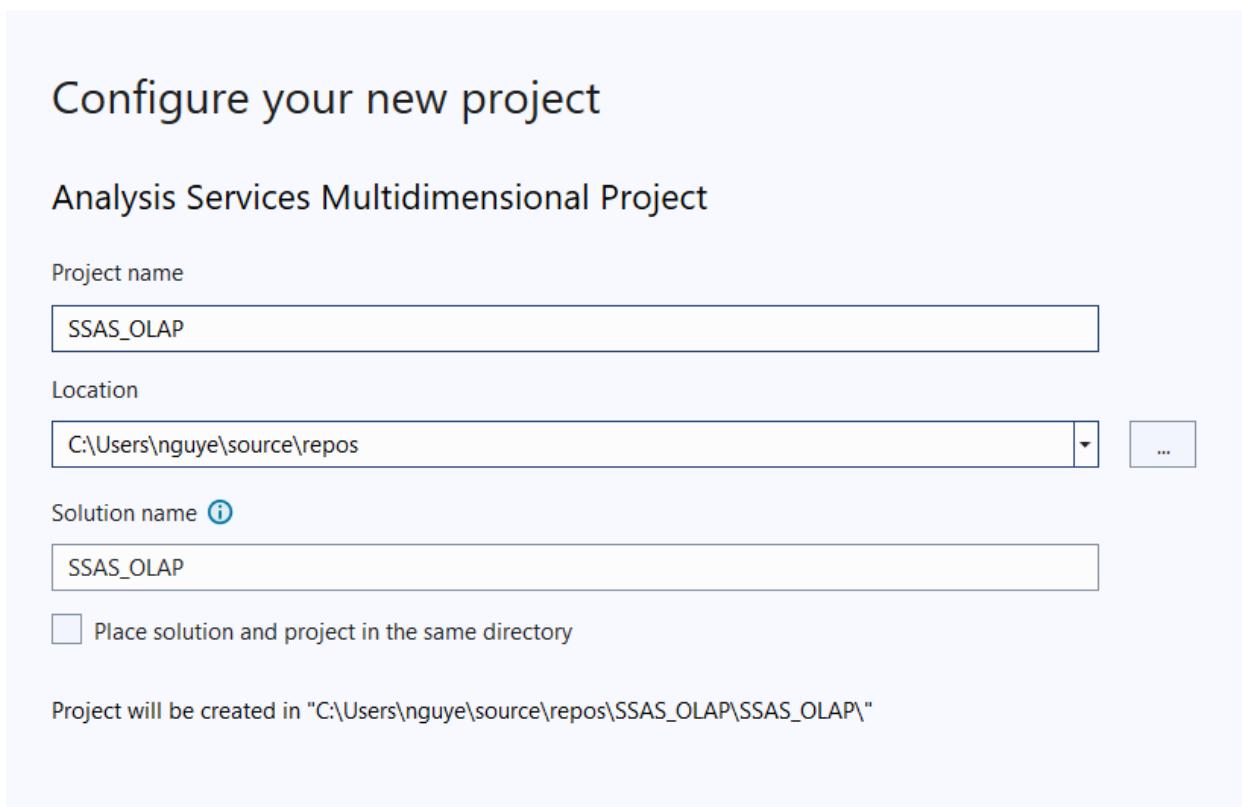
Bước 1: Mở Visual Studio và chọn “Create a new project”.

Bước 2: Chọn **Analysis Services Multidimensional Project** và chọn **Next**.



Hình 3.13 Chọn Analysis Services Multidimensional Project

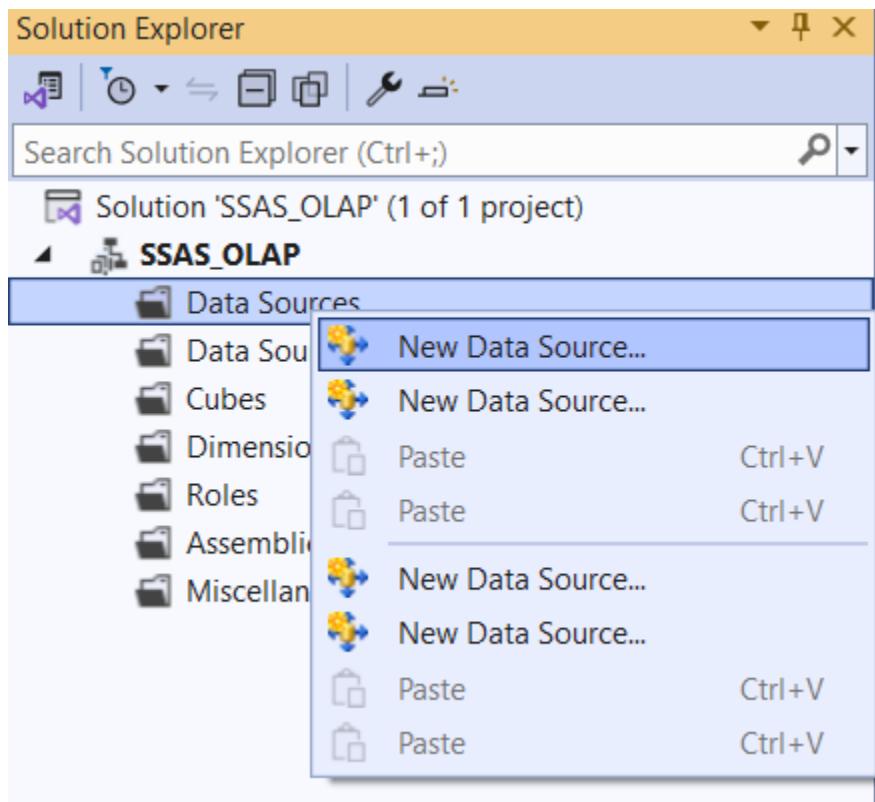
Bước 3: Đặt tên và thiết lập đường dẫn cho Project. Sau đó chọn **Create**.



Hình 3.14 Tạo Project SSAS

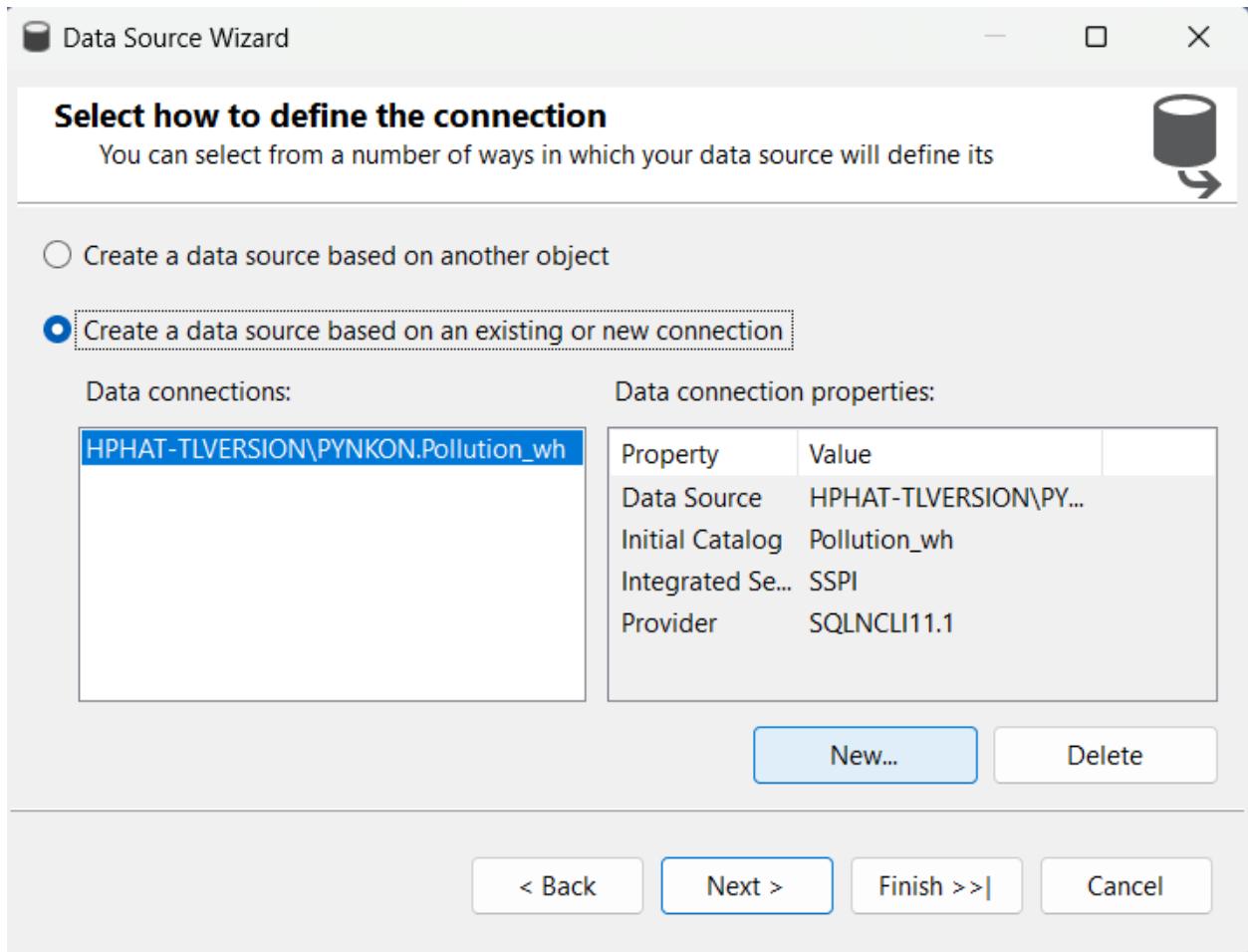
3.3 Thiết lập nguồn dữ liệu (Data Sources)

Bước 1: Bên góc phải màn hình, phần **Solution Explorer** nhấn chuột phải vào **DataSources** và chọn **New Data Source**.



Hình 3.15 New Data Source

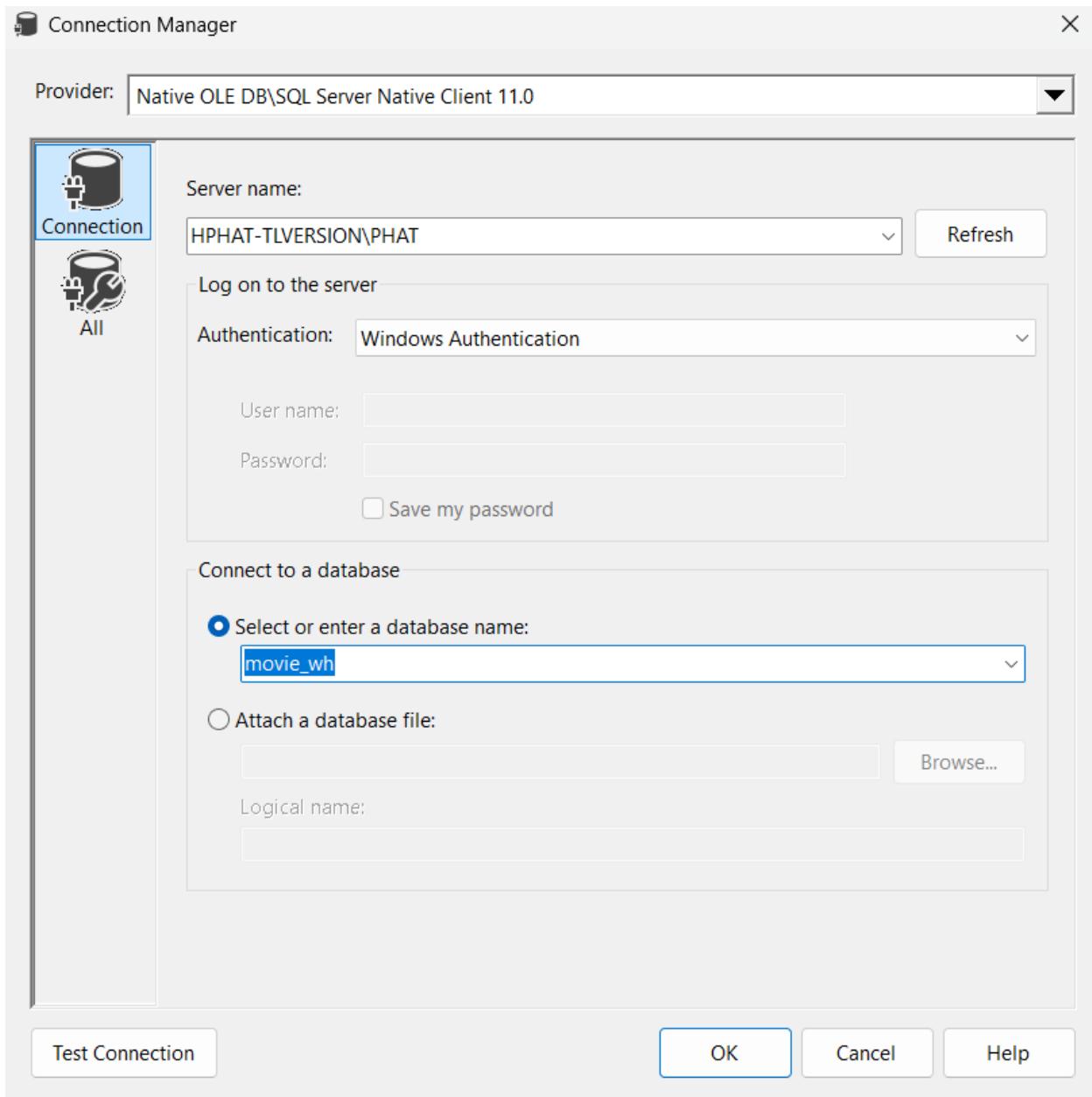
Bước 2: Tích vào **Create a data source based on an existing or new connection** và chọn **New**.



Hình 3.16 Thiết lập kết nối đến cơ sở dữ liệu

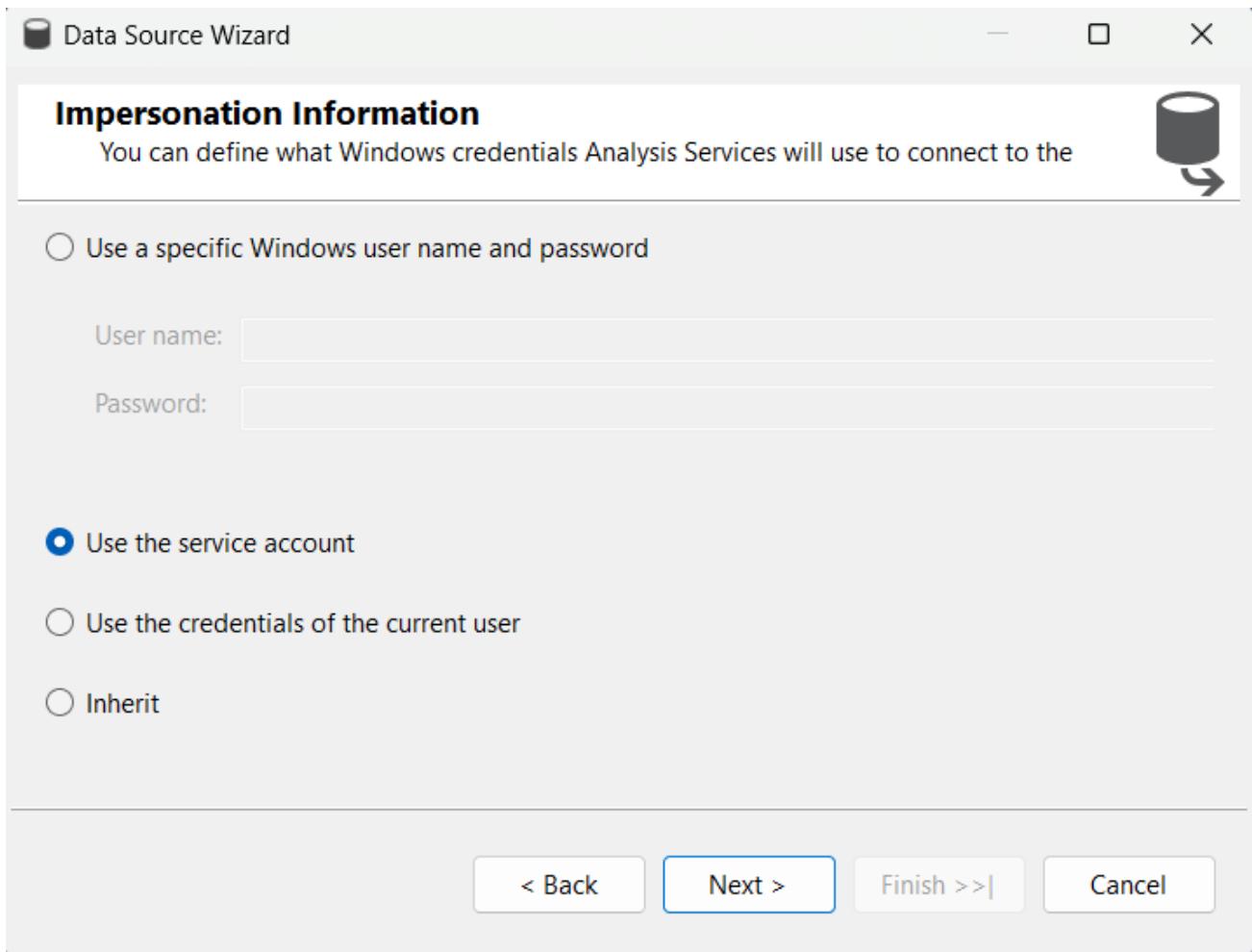
Bước 3: Chọn **Server name** và chọn **Movie_wh**. Nhấn **OK** để hoàn tất quá trình kết nối

Kho dữ liệu và OLAP - IS217.P12



Hình 3.17 Kết nối đến Movie_wh

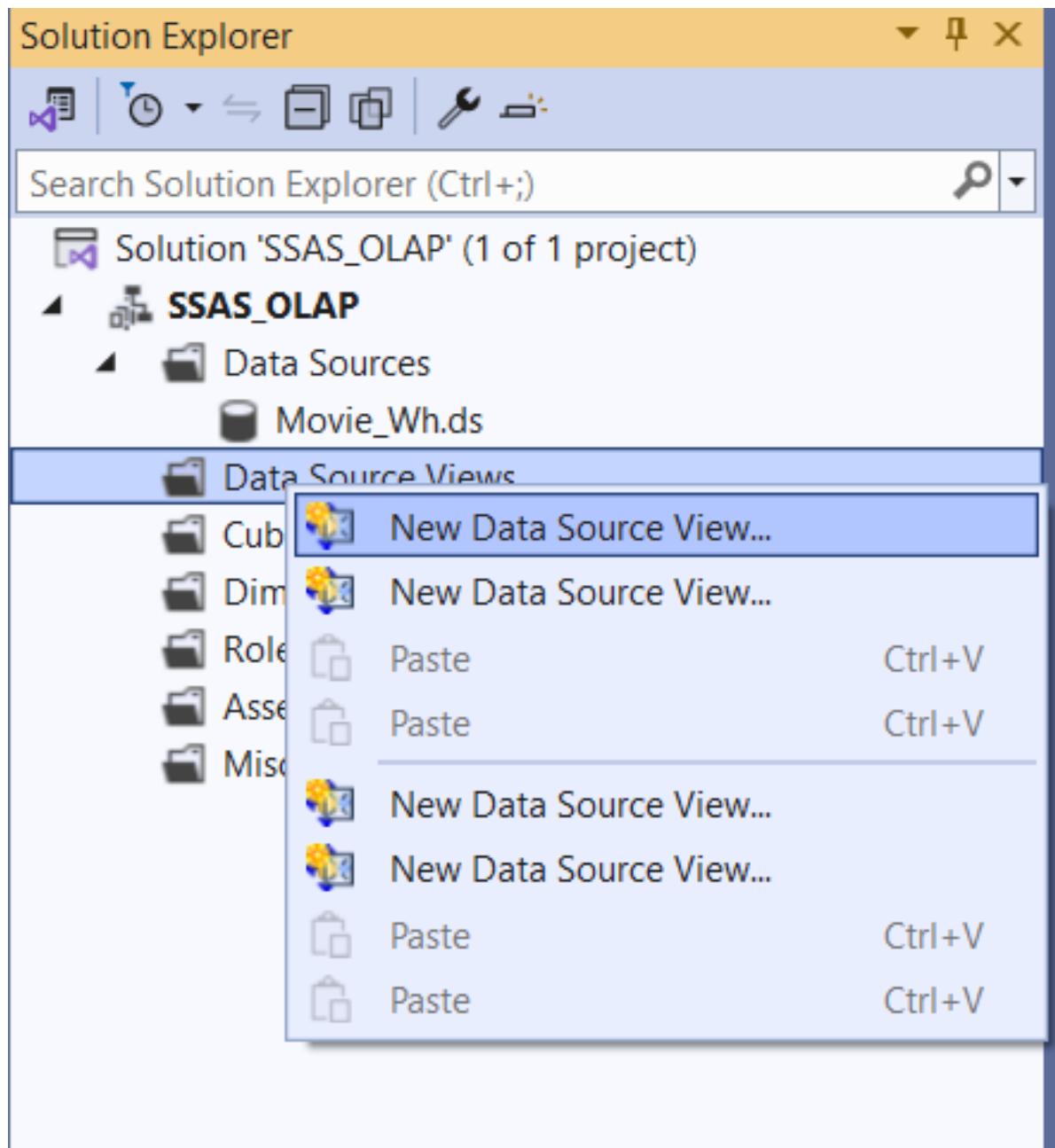
Bước 4: Chọn **Use the server account**. Nhấn **Next** và chọn **Finish**



Hình 3.18 Chọn Use the server account

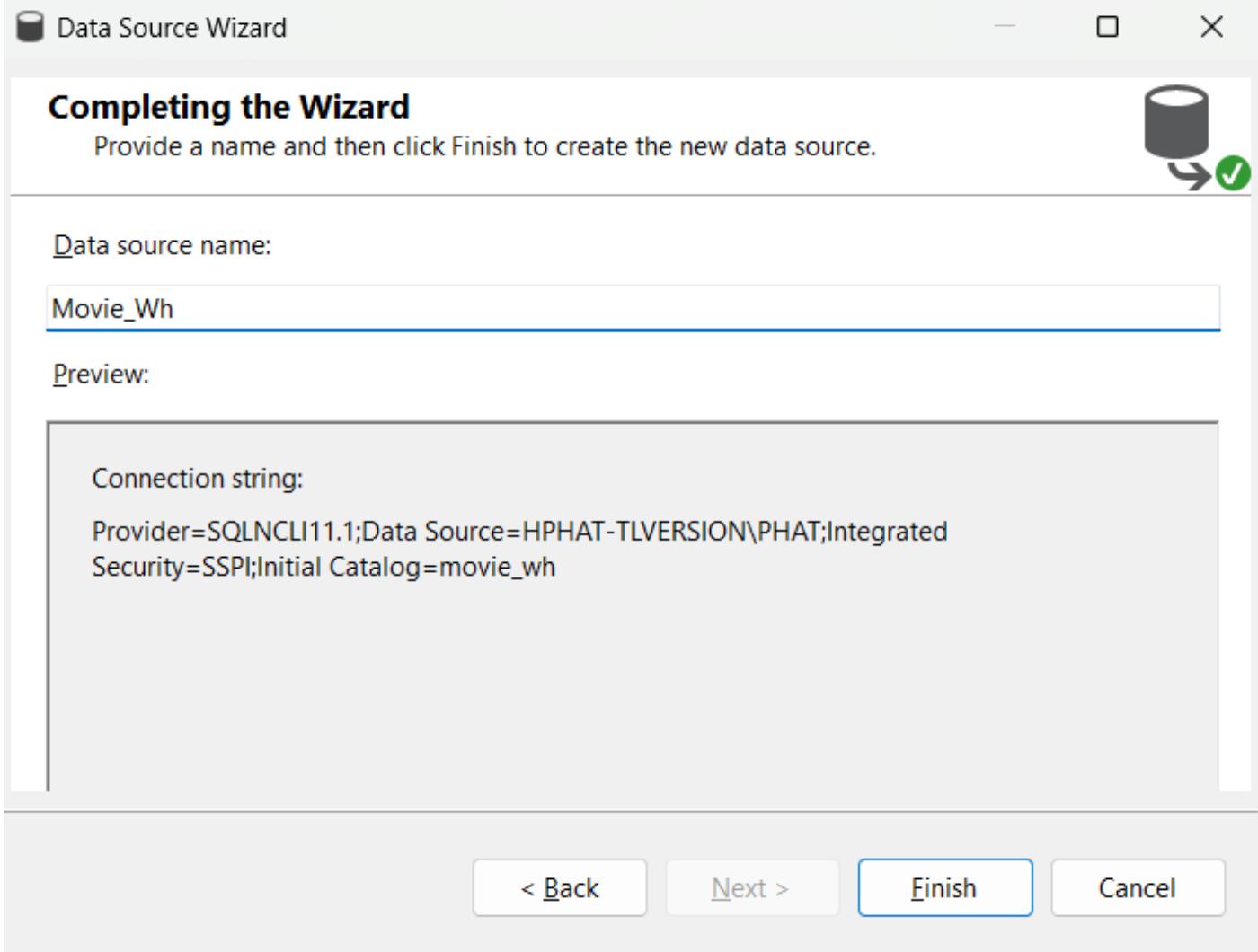
3.4 Thiết lập khung nhìn dữ liệu nguồn (Data Source Views)

Bước 1: Bên góc phải màn hình, phần **Solution Explorer** nhấn chuột phải vào **DataSources View** và chọn **New Data Source View**.



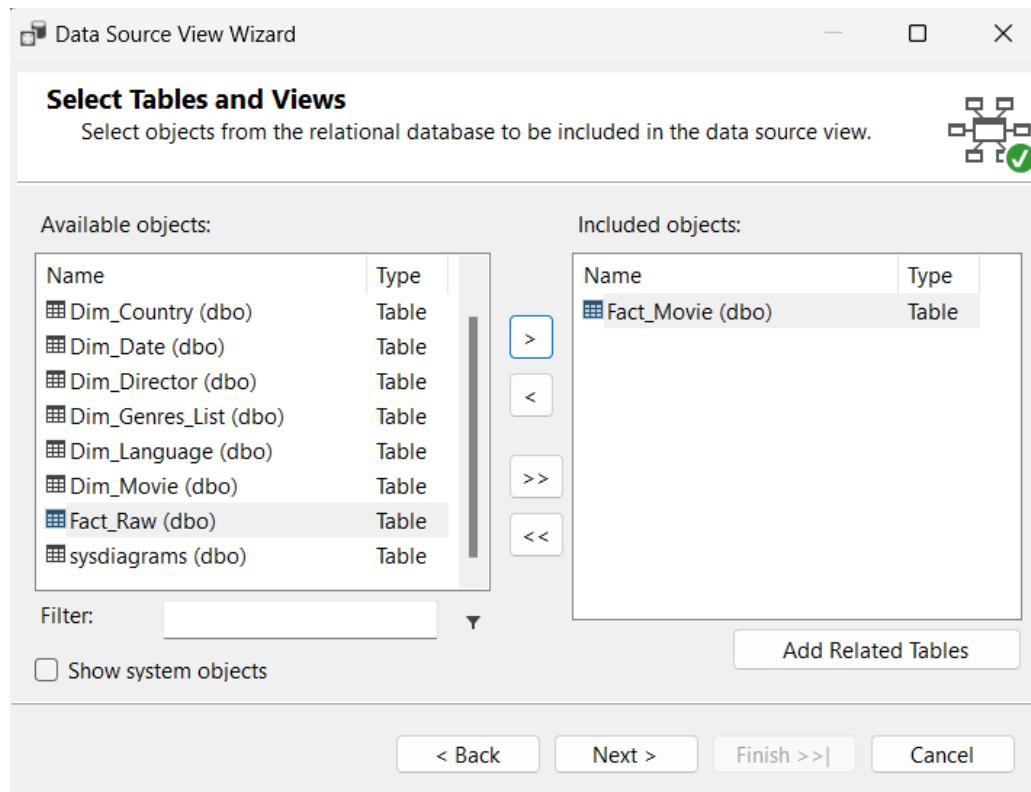
Hình 3.19 Chọn New Data Source View

Bước 2: Chọn kho dữ liệu **Movie_wh** và nhấn **Next**



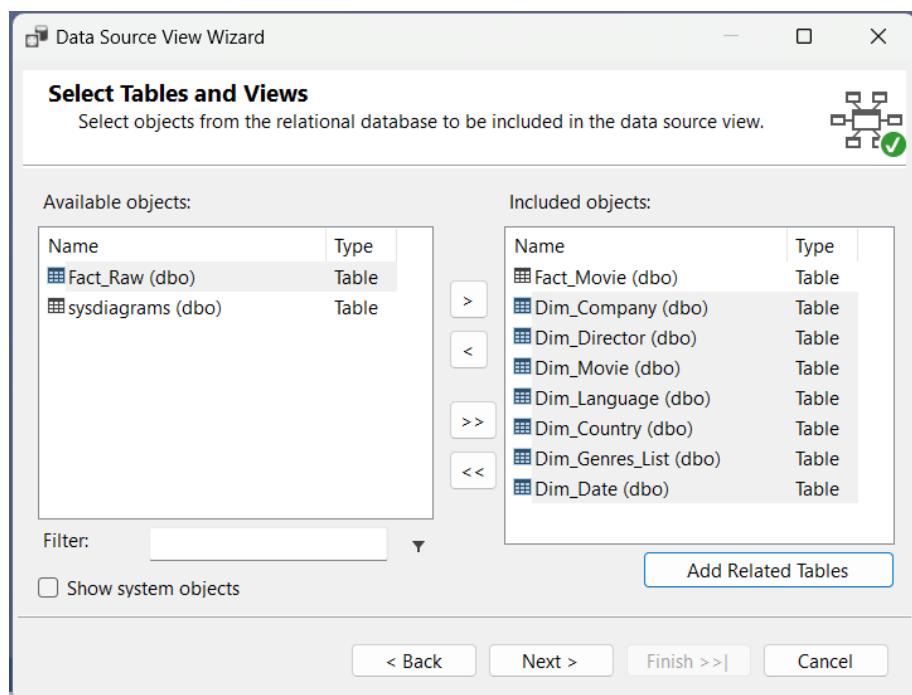
Hình 3.20 Chọn Movie_wh

Bước 3: Kéo bảng Fact Movie qua ô Included objects và nhấn Add Related Tables.



Hình 3.21 Kéo Fact Movie qua Included objects

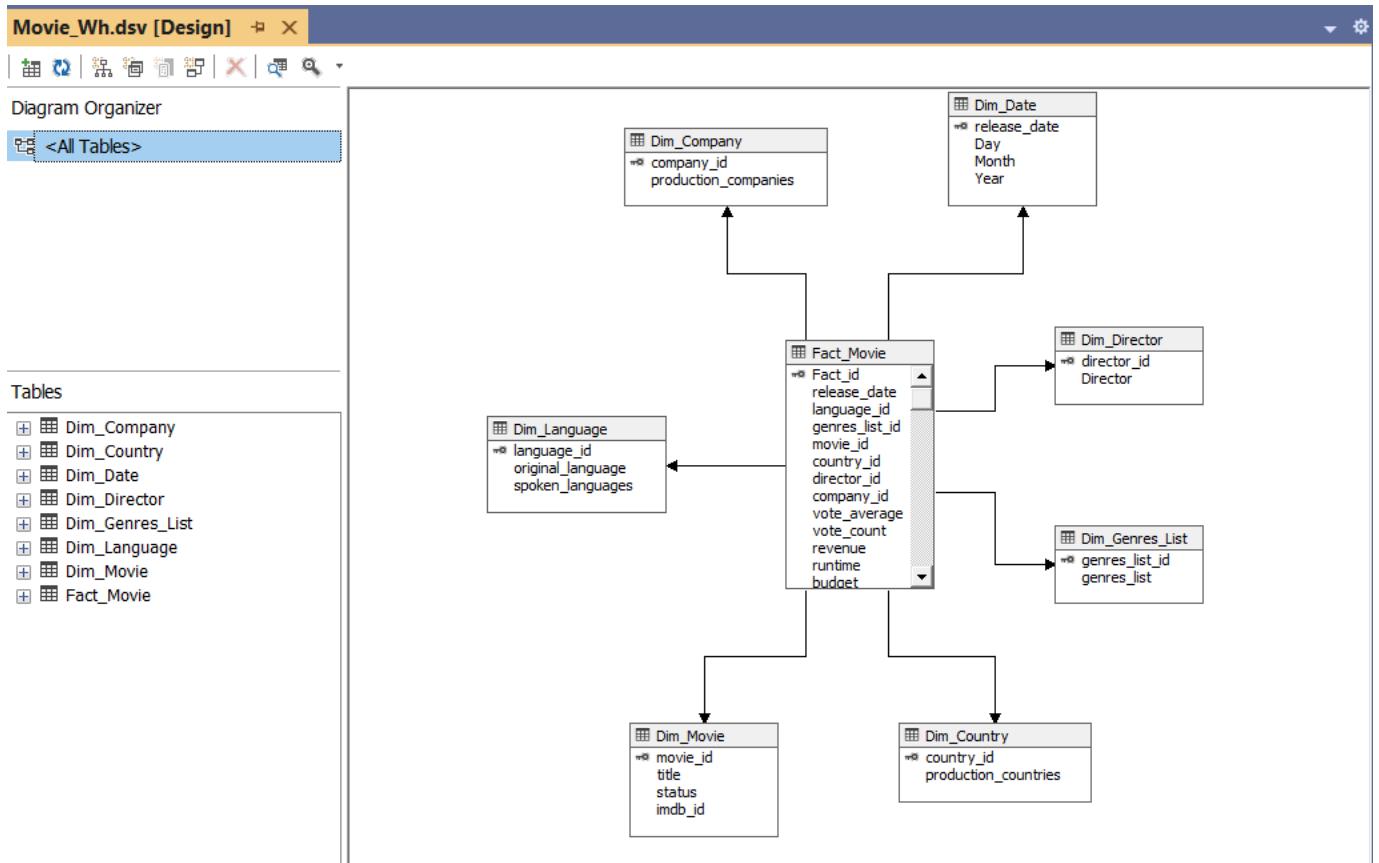
Bước 4: Các bảng Dim sẽ tự động được di chuyển sang và nhấn Next.



Hình 3.22 Các bảng tự động di chuyển sang

Kho dữ liệu và OLAP - IS217.P12

Bước 5: Nhấn **Finish** để kết thúc quá trình tạo Data Source Views và kết quả sẽ được hiển thị như hình.

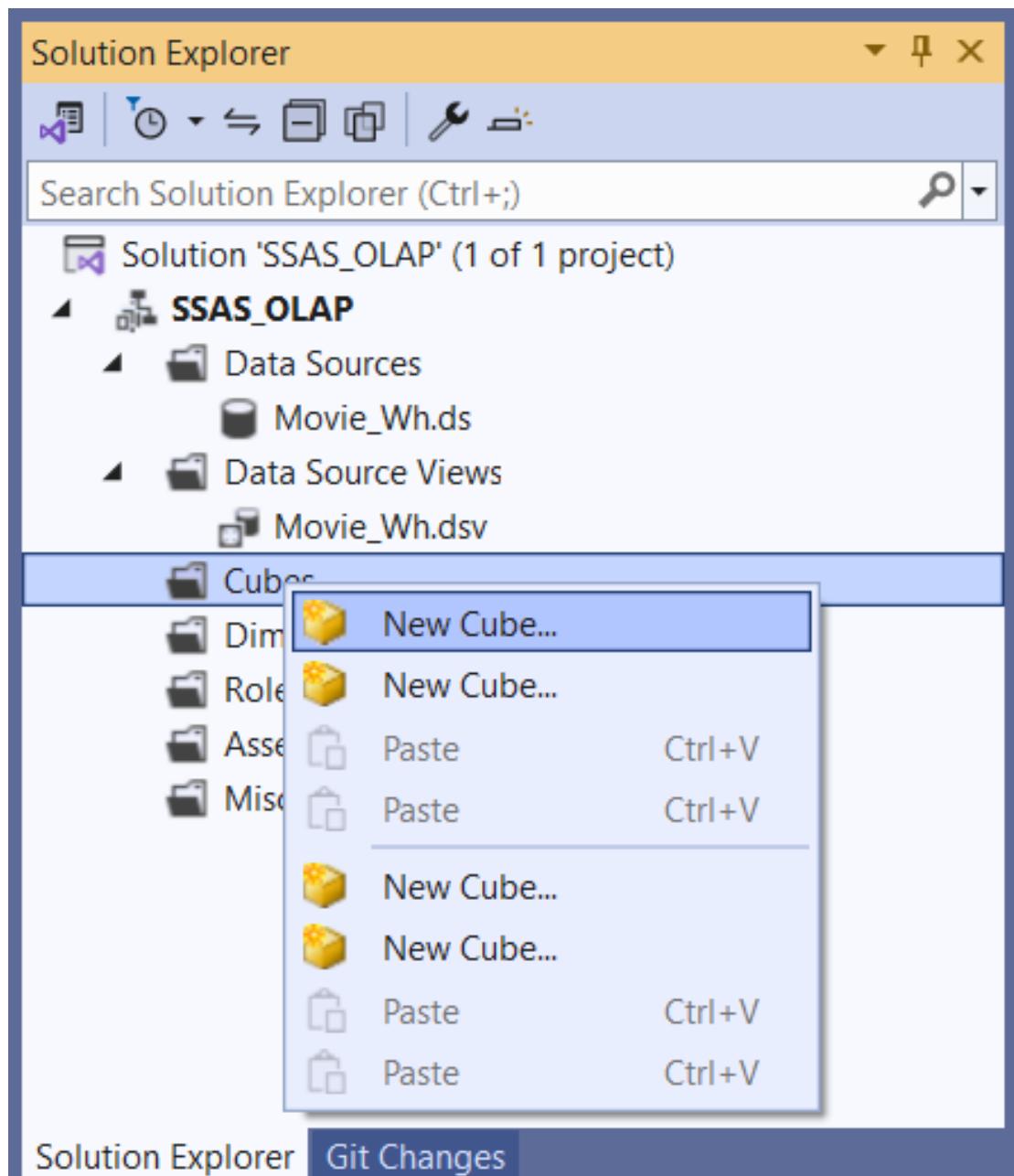


Hình 3.23 Kết quả sau khi tạo Data Source Views

3.5 Thiết lập các khối (Cube)

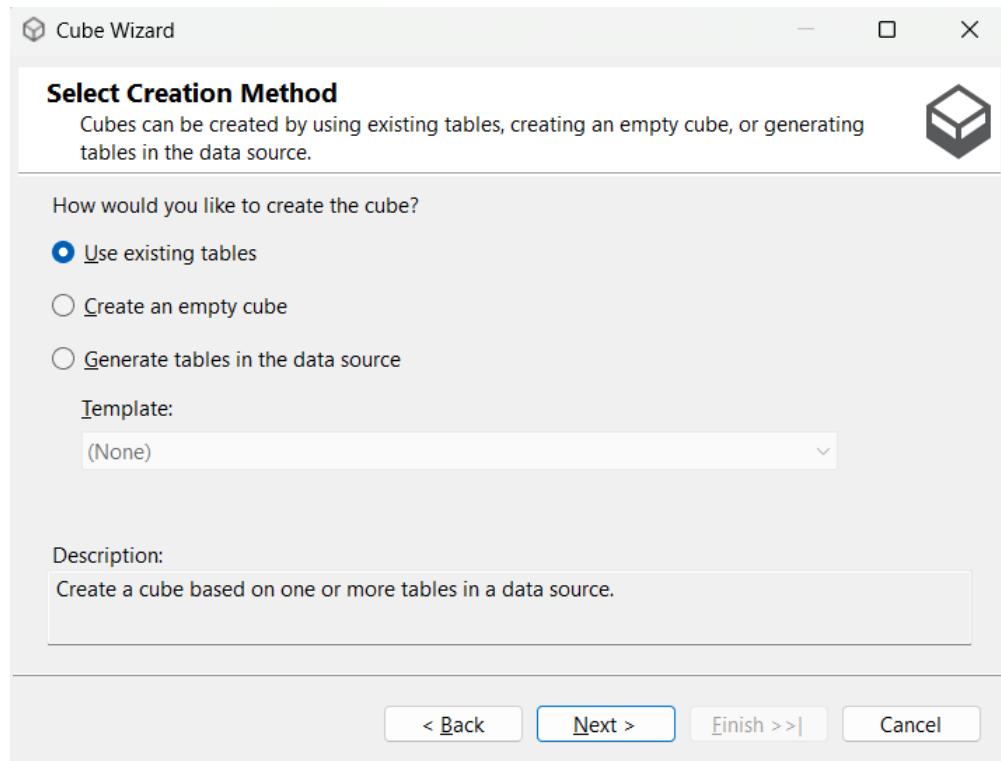
3.5.1 Tạo Cube và Dimension

Bước 1: Bên góc phải màn hình, phần **Solution Explorer** nhấn chuột phải vào **Cubes** và chọn **New Cube**.



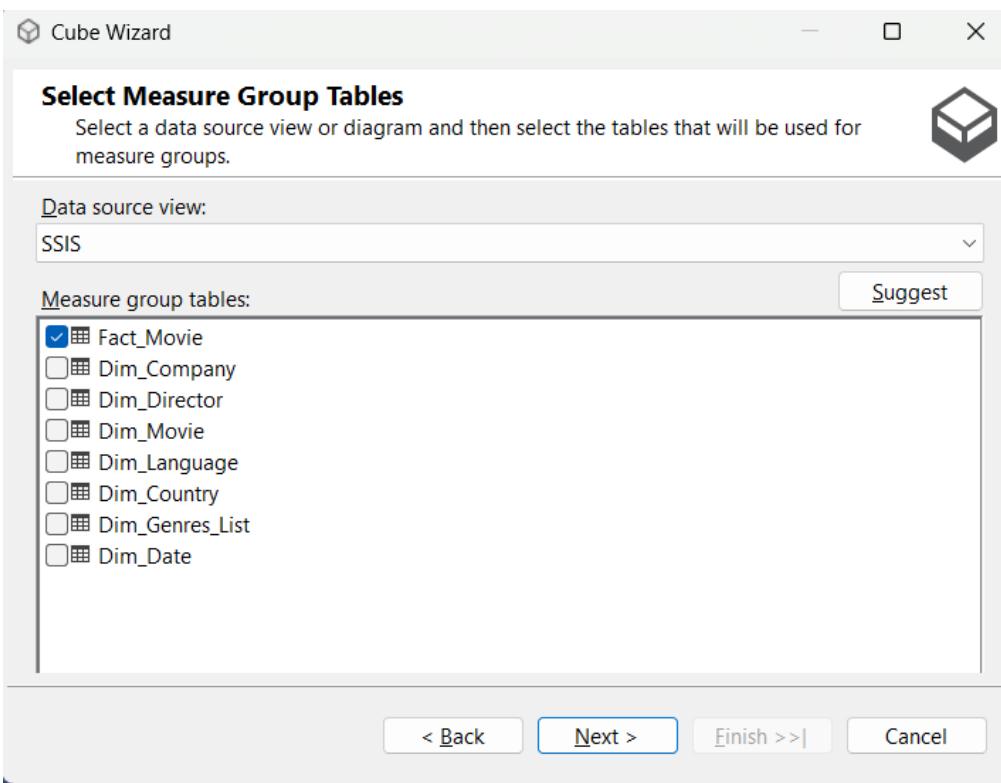
Hình 3.24 Chọn New Cube

Bước 2: Chọn **use an existing table** và nhấn **Next**.



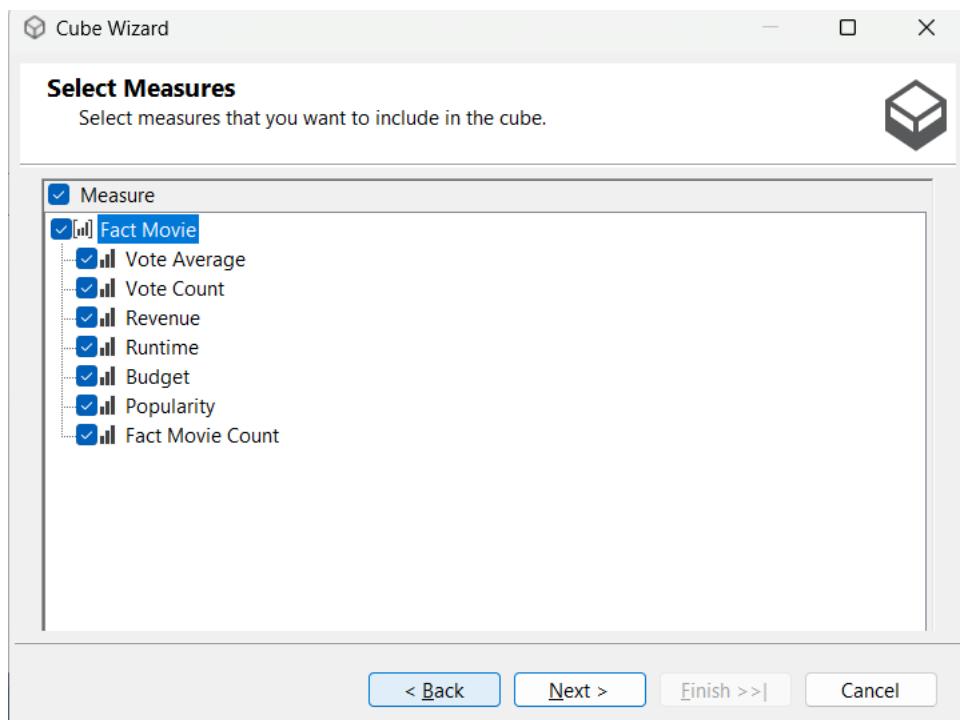
Hình 3.25 Chọn use an existing table

Bước 3: Tiếp tục chọn bảng **Fact_Movie** và nhấn **Next**.



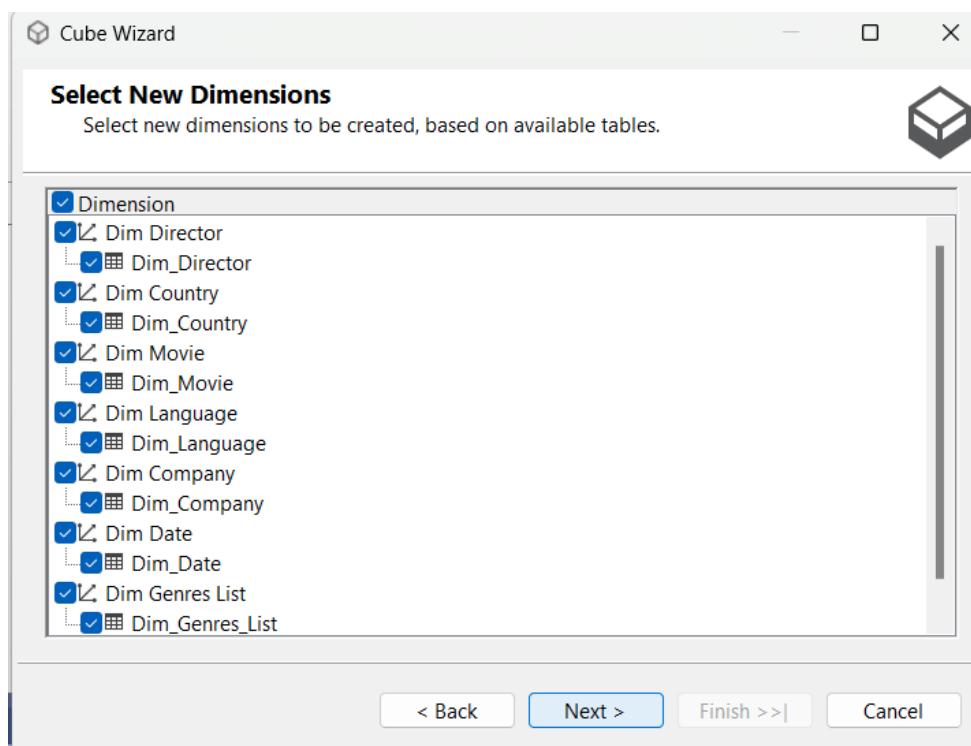
Hình 3.26 Chon Fact_Movie

Bước 4: Các độ đo được tạo tự động, để mặc định và nhấn Next.



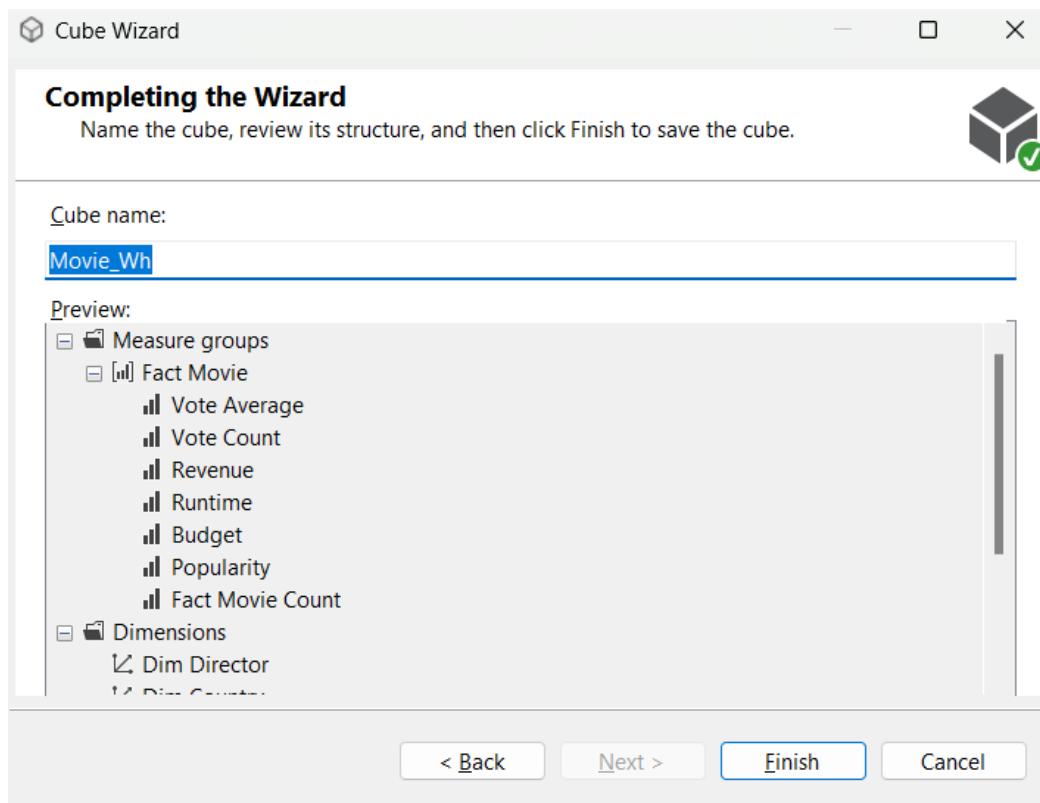
Hình 3.27 Chọn Measures

Bước 5: Để mặc định và nhấn Next.

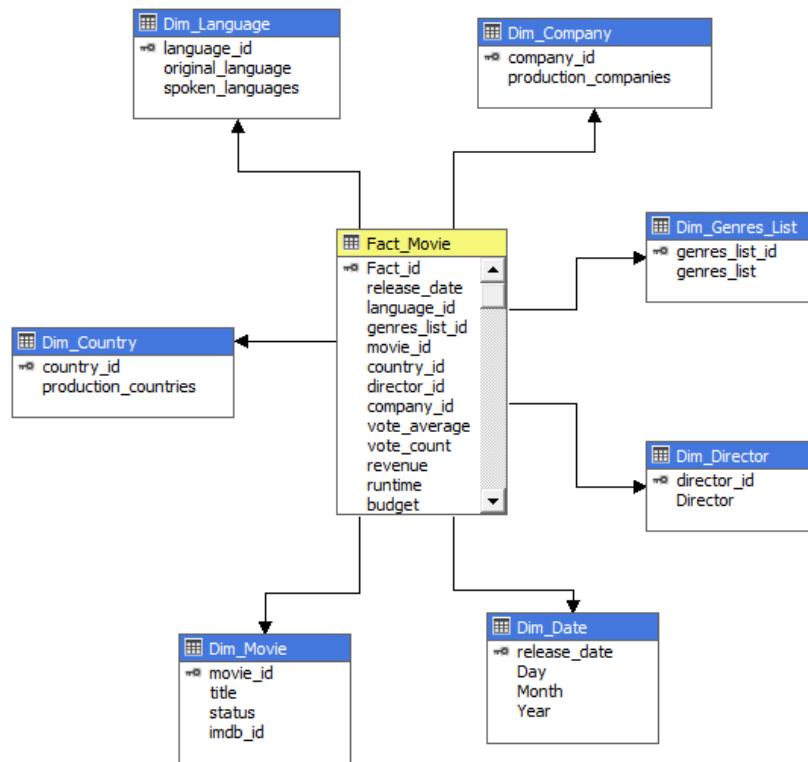


Hình 3.28 Tab Select New Dimensions

Bước 6: Nhấn **Finish**. Kết quả sẽ được như hình.



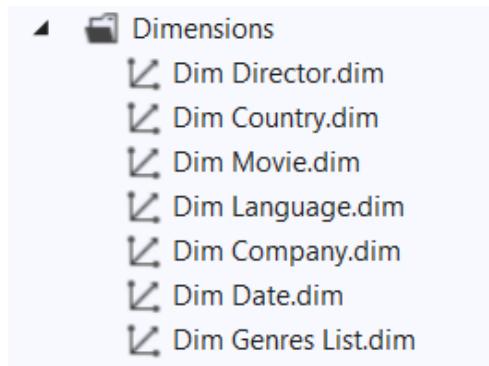
Hình 3.29 Nhập Cube name



Hình 3.30 Kết quả sau khi thiếp lập Cube

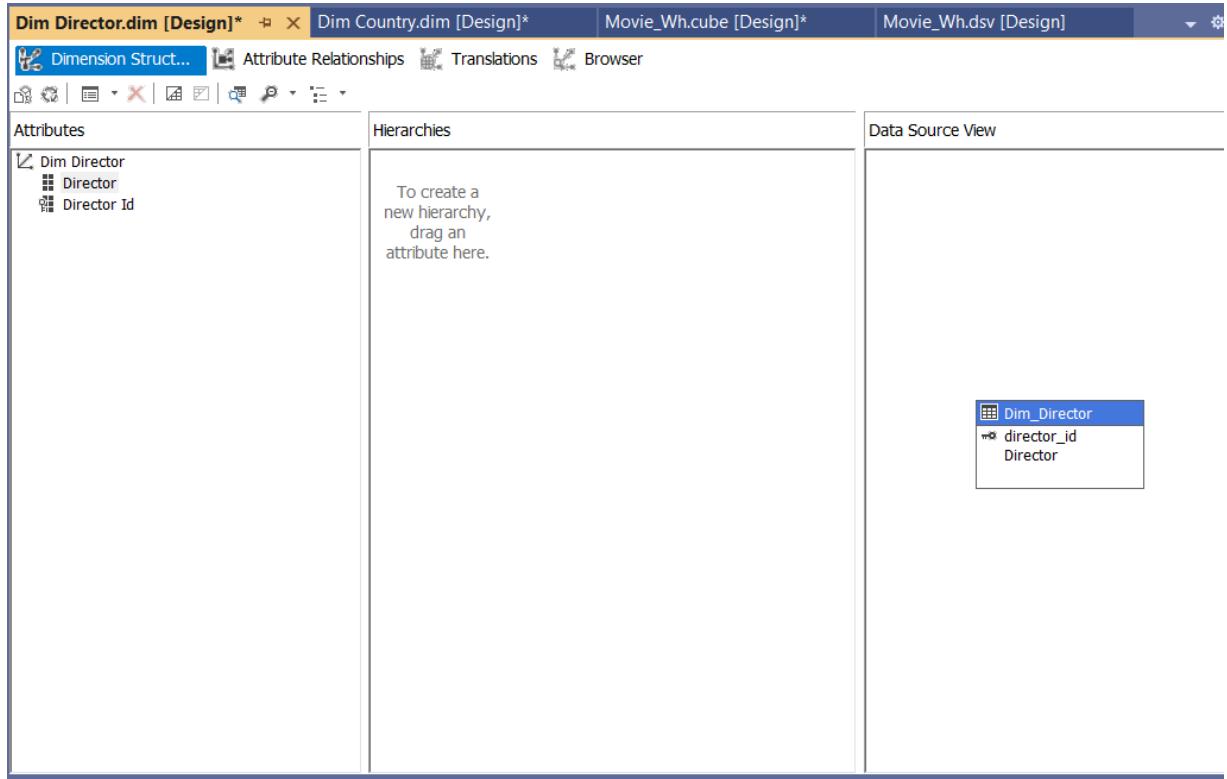
3.5.2 Thêm thuộc tính vào Dimension

Bước 1: Ở Solution Explorer, nhấp đúp chuột trái vào từng bảng Dimension để tiến hành thêm các thuộc tính.



Hình 3.31 Dimensions

Bước 2: Ở bảng Dim Director, kéo thả các thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



Hình 3.32 Thuộc tính bảng Dim Director

Bước 3: Ở bảng Dim Country, kéo thả các thuộc tính cần thiết ở cột Data Source View sang cột Attributes.

Kho dữ liệu và OLAP - IS217.P12

The screenshot shows the SSAS Dimension Designer interface for the 'Dim Country.dim [Design]' dimension. The top navigation bar includes tabs for 'Dim Country.dim [Design]', 'Movie_Wh.cube [Design]', and 'Movie_Wh.dsv [Design]'. The 'Dimension Struct...' tab is selected. The interface is divided into three main panes: 'Attributes' (left), 'Hierarchies' (center), and 'Data Source View' (right). The 'Attributes' pane lists the dimension's attributes: 'Dim Country' (selected), 'Country Id', and 'Production Countries'. The 'Hierarchies' pane contains a placeholder message: 'To create a new hierarchy, drag an attribute here.' The 'Data Source View' pane displays the data source view for the dimension, showing a table named 'Dim_Country' with two columns: 'country_id' and 'production_countries'.

Hình 3.33 Thuộc tính bảng Dim Country

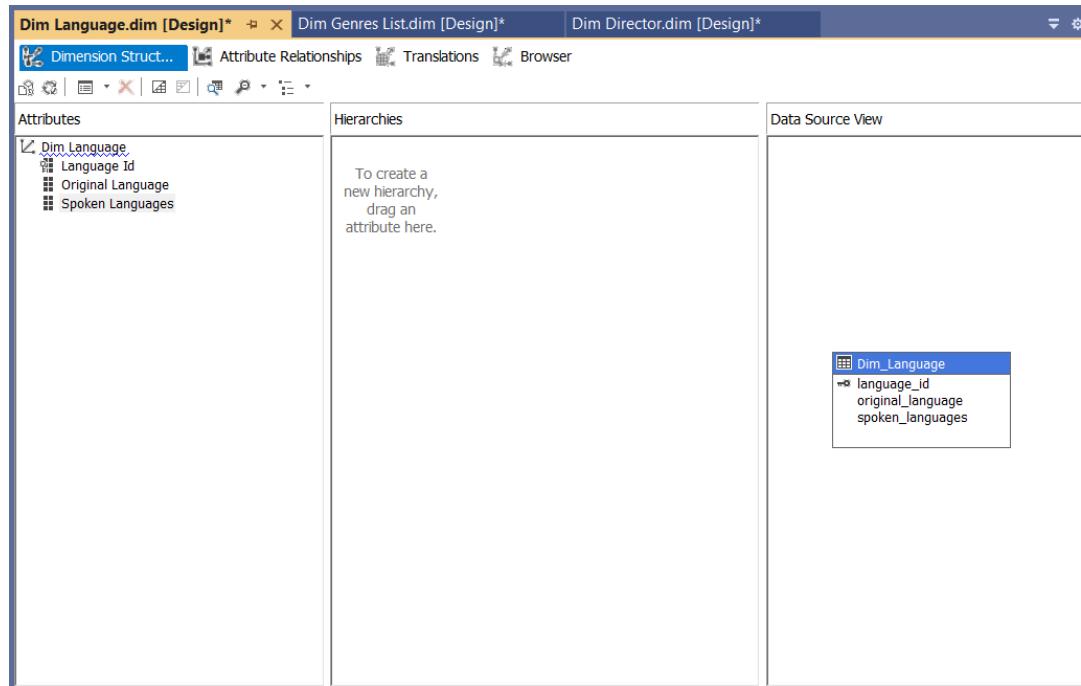
Bước 4: Ở bảng Dim Genres List, kéo thả các thuộc tính cần thiết ở cột Data Source View sang cột Attributes.

The screenshot shows the SSAS Dimension Designer interface for the 'Dim Genres List.dim [Design]' dimension. The top navigation bar includes tabs for 'Dim Genres List.dim [Design]', 'Dim Director.dim [Design]', and 'Dim Country.dim [Design]'. The 'Dimension Struct...' tab is selected. The interface is divided into three main panes: 'Attributes' (left), 'Hierarchies' (center), and 'Data Source View' (right). The 'Attributes' pane lists the dimension's attributes: 'Dim Genres List' (selected), 'Genres List', and 'Genres List Id'. The 'Hierarchies' pane contains a placeholder message: 'To create a new hierarchy, drag an attribute here.' The 'Data Source View' pane displays the data source view for the dimension, showing a table named 'Dim_Genres_List' with two columns: 'genres_list_id' and 'genres_list'.

Hình 3.34 Thuộc tính bảng Dim Genres List

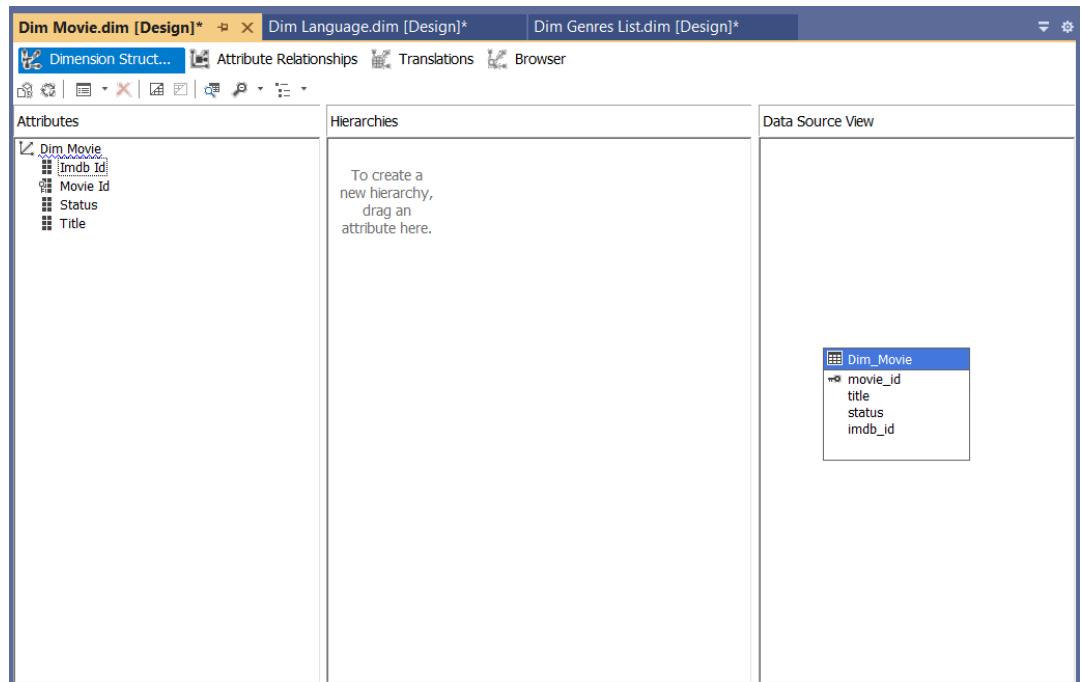
Kho dữ liệu và OLAP - IS217.P12

Bước 5: Ở bảng Dim Language, kéo thả các thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



Hình 3.35 Thuộc tính bảng Dim Language

Bước 6: Ở bảng Dim Movie, kéo thả các thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



Hình 3.36 Thuộc tính bảng Dim Movie

Kho dữ liệu và OLAP - IS217.P12

Bước 7: Ở bảng Dim Date, kéo thả các thuộc tính cần thiết ở cột Data Source View sang cột Attributes.

The screenshot shows the SSAS Dimension Designer interface for the Dim Date dimension. The top navigation bar includes tabs for Dim Date.dim [Design], Dim Movie.dim [Design], Dim Language.dim [Design], and Dim Genres List.dim [Design]. The left pane, titled 'Attributes', lists the attributes of the Dim Date dimension: Day, Month, Release Date, and Year. The middle pane, titled 'Hierarchies', contains a placeholder message: 'To create a new hierarchy, drag an attribute here.' The right pane, titled 'Data Source View', displays a table named 'Dim_Date' with four columns: release_date, Day, Month, and Year.

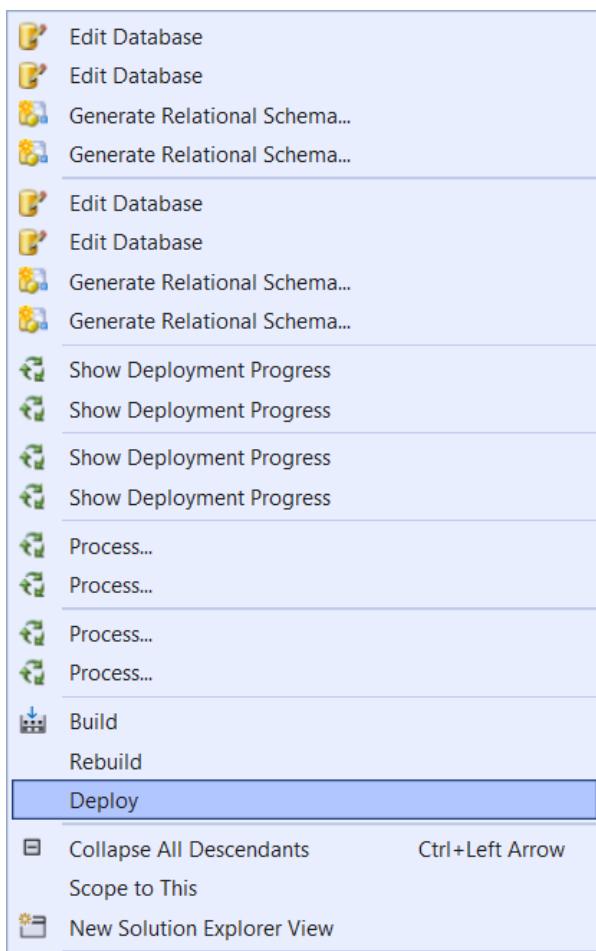
Hình 3.37 Thuộc tính bảng Dim Date

Bước 8: Ở bảng Dim Company, kéo thả các thuộc tính cần thiết ở cột Data Source View sang cột Attributes.

The screenshot shows the SSAS Dimension Designer interface for the Dim Company dimension. The top navigation bar includes tabs for Dim Company.dim [Design], Dim Date.dim [Design], Dim Movie.dim [Design], and Dim Language.dim [Design]. The left pane, titled 'Attributes', lists the attributes of the Dim Company dimension: Company Id and Production Companies. The middle pane, titled 'Hierarchies', contains a placeholder message: 'To create a new hierarchy, drag an attribute here.' The right pane, titled 'Data Source View', displays a table named 'Dim_Company' with two columns: company_id and production_companies.

Hình 3.38 Thuộc tính bảng Dim Company

Bước 9: Nhấn Deploy để chạy dự án SSAS.



Hình 3.39 Nhấn Deploy để chạy dự án SSAS

The screenshot displays the 'Deployment Progress' window for the 'OLAP_SSAS' database. The window title is 'Deployment Progress - OLAP_SSAS'. It shows deployment details for the server 'HPHAT-TLVERSION\PHAT' and the database 'OLAP_SSAS'. The main pane lists the deployment steps, starting with 'Command' and then 'Processing Database 'OLAP_SSAS'' completed. Below this, it details the processing of various dimensions: 'Dim Company', 'Dim Country', 'Dim Date', 'Dim Director', 'Dim Genres List', 'Dim Language', and 'Dim Movie'. Each dimension entry includes a start time (11/25/2024 8:38:06 PM), end time (11/25/2024 8:38:07 PM), and duration (0:00:00).

Hình 3.40 Kết quả sau khi nhấn Deploy

3.5.3 Xác định các độ đo

Bước 1: Mở khối Cube

Bước 2: Chọn Show Measures Grid để hiển thị chi tiết các độ đo.

Measures				
	Name	Measure Group	Data Type	Aggregation
📊	Vote Average	~~~~~	Double	AverageOfC...
📊	Vote Count	~~~~~	Integer	Sum
📊	Revenue	~~~~~	Integer	Sum
📊	Runtime	~~~~~	Integer	Sum
📊	Budget	~~~~~	Integer	Sum
📊	Popularity	~~~~~	Double	AverageOfC...
📊	Fact Movie Count	Fact Movie	Integer	Count

Hình 3.41 Bảng các độ đo ban đầu

Bước 3: Thiếp lập một số thay đổi như hình.

Measures				
	Name	Measure Group	Data Type	Aggregation
📊	Vote Average	Fact Movie	Single	AverageOfC...
📊	Vote Count		Integer	Sum
📊	Revenue		Integer	Sum
📊	Runtime		Integer	Sum
📊	Budget		Integer	Sum
📊	Popularity		Single	AverageOfC...
📊	Number of Movies		Integer	Count

Hình 3.42 Bảng các độ đo sau khi thay đổi

Bước 4: Deploy lại project

Deployment Progress - OLAP_SSAS

Server : HPHAT-TLVERSION\PHAT
Database : OLAP_SSAS

Command

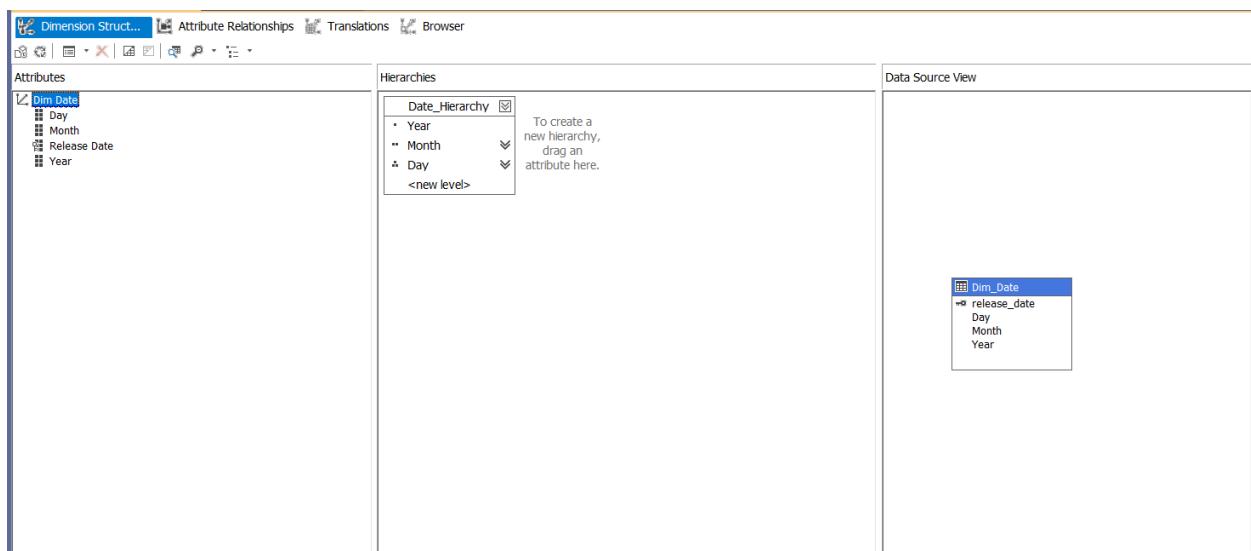
- Command
 - Processing Database 'OLAP_SSAS' completed.
Start time: 11/25/2024 9:53:14 PM; End time: 11/25/2024 9:53:14 PM; Duration: 0:00:00
 - Processing Cube 'Movie Wh' completed.
Start time: 11/25/2024 9:53:14 PM; End time: 11/25/2024 9:53:14 PM; Duration: 0:00:00
 - Processing Measure Group 'Fact Movie' completed.

Hình 3.43 Kết quả sau khi Deploy Project

3.5.4 Phân cấp các bảng chiều

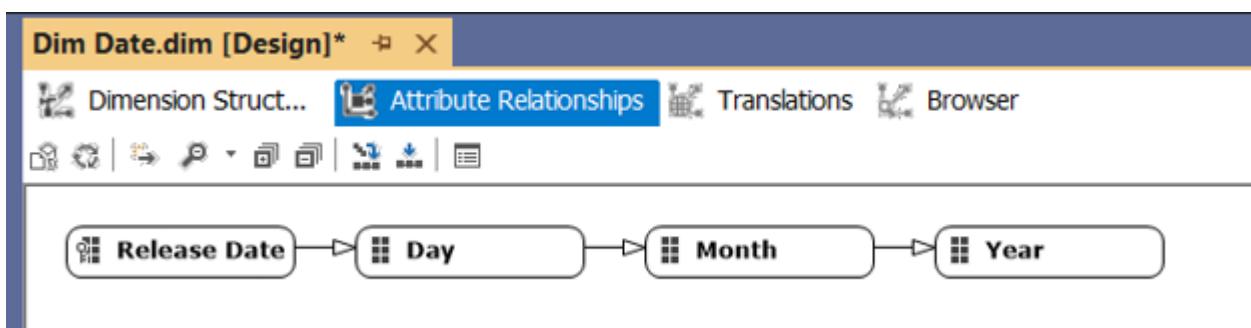
Phân cấp bảng Dim Date

Bước 1: Kéo thả các thuộc tính year, month, day sang cột Hierarchies. Với thứ tự phân cấp từ cao tới thấp nhất và đổi tên Hierarchy thành Date_Hierarchy.



Hình 3.44 Phân cấp Date_Hierarchy

Bước 2: Chuyển sang tab Attribute Relationships để tiến hành định nghĩa Attribute Relationships. Tiến hành kéo thả phân cấp từ nhỏ đến lớn theo thứ tự từ trái sang phải.

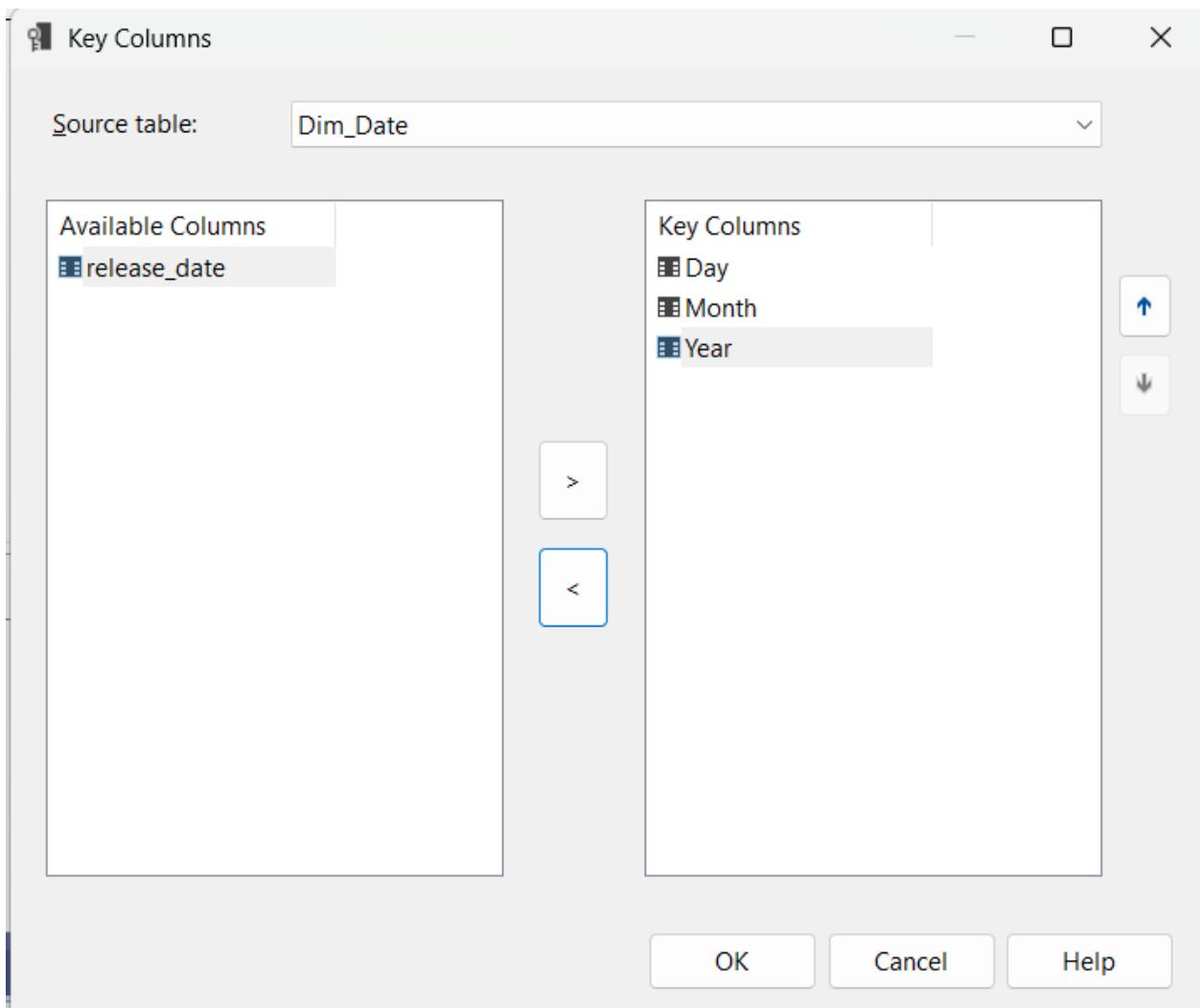


Hình 3.45 Phân cấp Day -> Month -> Year

Bước 3: Chính khóa cột (Key Columns) và tên cột (Name Column) của thuộc tính **Day**. Vì thuộc tính **Day** là thuộc tính cấp nhỏ nhất nên sẽ lấy khóa cột gồm chính nó và những thuộc tính cấp cao hơn.

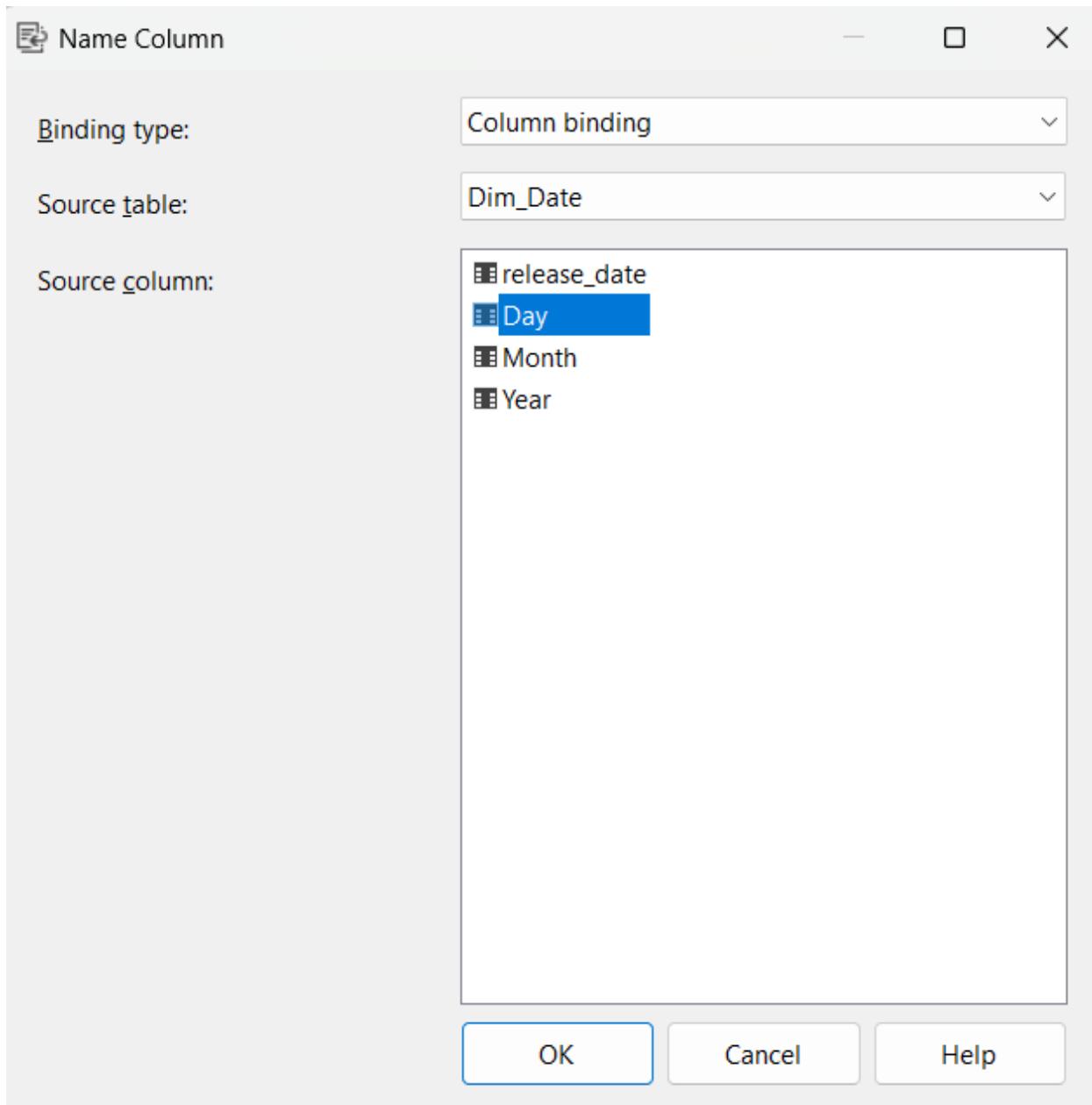
Tại cửa sổ Properties của thuộc tính Day, chọn **Key Columns**.

Thêm các thuộc tính cấp cao hơn vào Key Columns, sau đó OK để hoàn tất.



Hình 3.46 Chỉnh Key Columns cho Day

Tại cửa sổ Properties của thuộc tính **Day**, ta chọn **Name Column** và chọn tên thuộc tính là Day



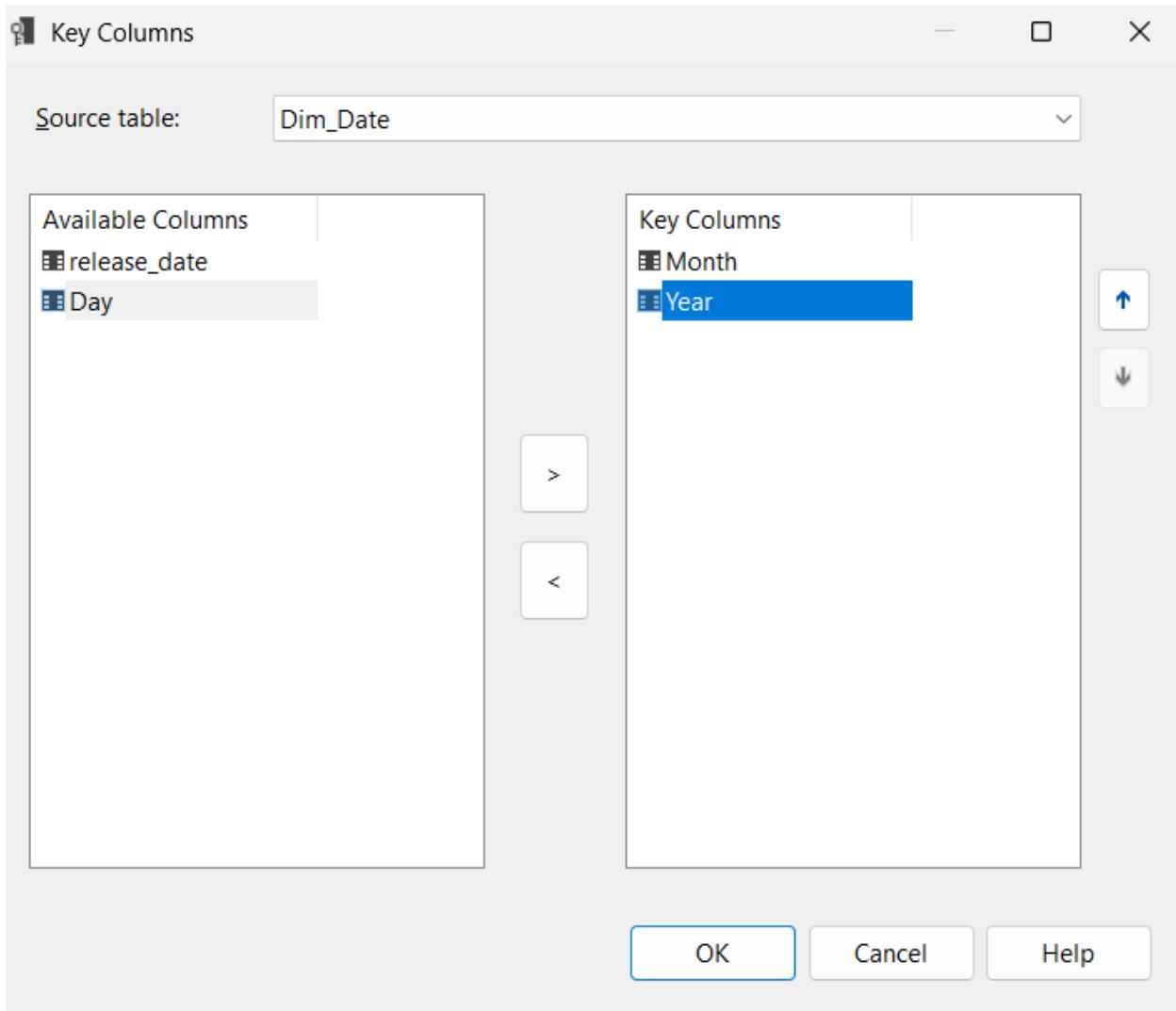
Hình 3.47 Chính Name Column cho Day

Bước 4: Chính khóa cột (**KeyColumns**) và tên cột (**Name Column**) của thuộc tính **Month**. Vì thuộc tính **Month** là thuộc tính cấp nhỏ hơn Year nên sẽ lấy khóa cột gồm chính nó và những thuộc tính cấp cao hơn.

Tại cửa sổ **Properties** của thuộc tính **Month**, chọn **Key Columns**.

Thêm các những thuộc tính cấp cao hơn vào KeyColumns, sau đó chọn OK để hoàn tất

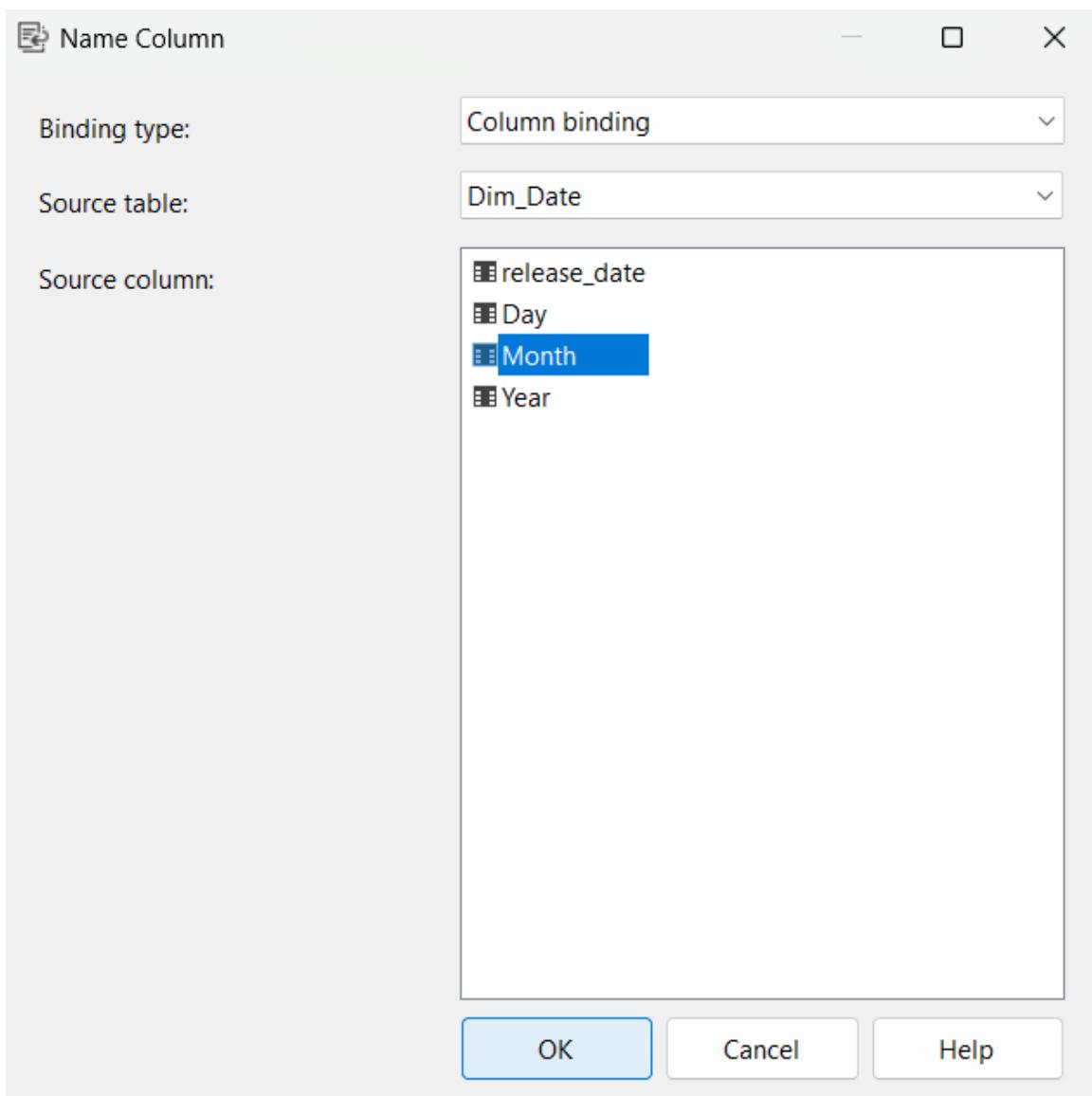
Kho dữ liệu và OLAP - IS217.P12



Hình 3.48 Chỉnh Key Columns cho Month

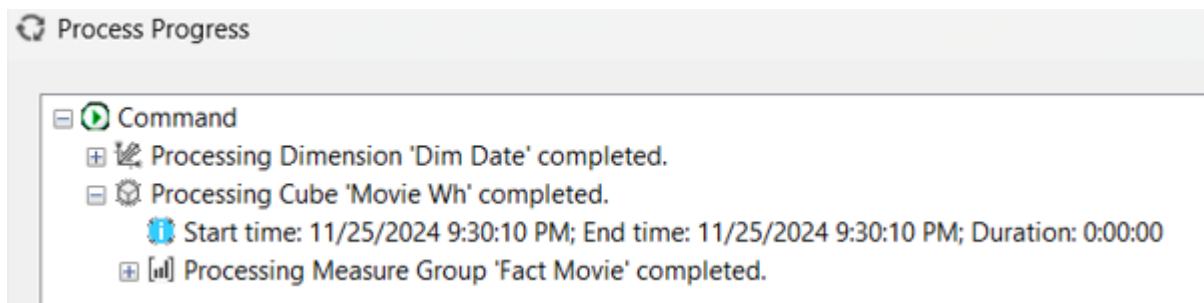
Tại cửa sổ Properties của thuộc tính Month, ta chọn Name Column và chọn tên thuộc tính là Month

Kho dữ liệu và OLAP - IS217.P12



Hình 3.49 Chỉnh Name Column cho Month

Bước 5: Nhấn Deploy để chạy dự án. Khi deploy thành công, hệ thống sẽ hiển thị như hình sau và bắt đầu thực hiện các câu truy vấn.



Hình 3.50 Deploy dự án

3.6 Thực hiện các câu truy vấn

3.6.1 Câu truy vấn 1: Top 5 năm có tổng doanh thu cao nhất

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Tạo NameSet [Cau_1] biểu thức như sau:

The screenshot shows the 'Script Organizer' window. On the left, a tree view lists various scripts, with '[Cau_1]' selected and highlighted in blue. On the right, a detailed view of '[Cau_1]' is shown. The 'Name:' field contains '[Cau_1]'. The 'Expression' section contains the MDX code: `SUBSET(ORDER([Dim Date].[Year].Members,[Measures].[Revenue], BDESC),0,5)`. A green checkmark indicates 'No issues found'. The 'Additional Properties' section shows 'Type: Dynamic' and an empty 'Display folder:' field.

Hình 3.51 Nameset [Cau_1]

Bước 2: Kéo thả thuộc tính [Year] trong bảng [Dim Date] và độ đo [Revenue] vào cửa sổ thực thi.

Sau đó kéo NameSet [Cau_1] vào Dimension

The screenshot shows the 'Analysis Services Browser' interface. On the left, there are navigation panes for 'SSIS', 'Metadata', and 'Measure Group'. The main area displays a table with columns: Dimension, Hierarchy, Operator, Filter Expression, and Parameters. One row is selected with 'Dim Date' in the Dimension column, 'Year' in the Hierarchy column, 'In' in the Operator column, and 'Cau_1' in the Filter Expression column. Below the table, there are tabs for 'Year' and 'Revenue', and a link 'Click to execute the query.'

Hình 3.52 Thiết lập truy vấn câu 1 ở Browser

Bước 3: “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Kho dữ liệu và OLAP - IS217.P12

Dimension	Hierarchy	Operator	Filter Expression
Dim Date	Year	In	Cau_1
Year	Revenue		
2014	1508677136		
2016	1038373075		
2017	1300636712		
2018	1807816778		
2019	1852990871		

Hình 3.53 Kết quả truy vấn câu 1 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

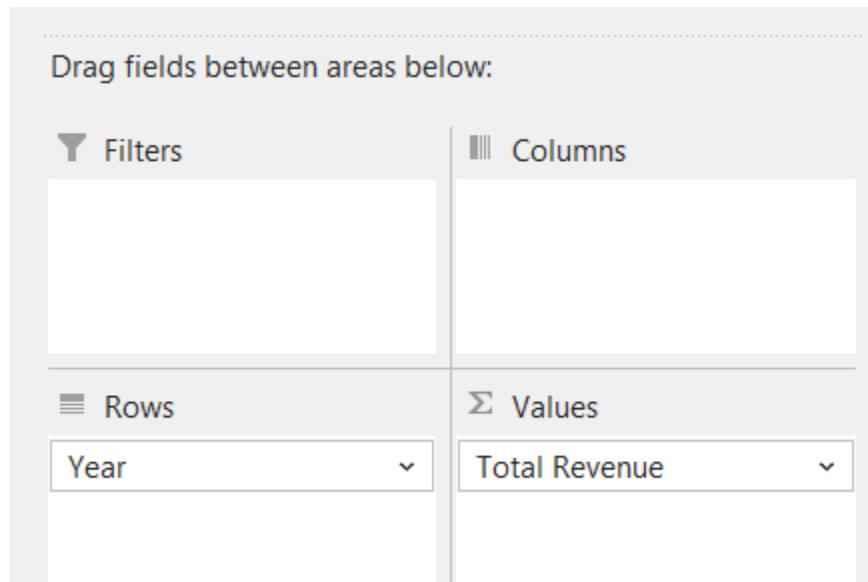
```
SELECT
    {[Measures].[Revenue]} ON COLUMNS,
    SUBSET(
        ORDER(
            [Dim Date].[Year].Members,
            [Measures].[Revenue], BDESC), 0, 5
    ) ON ROWS
FROM [SSIS];
```

Messages		Results
		Revenue
2019	1852990871	
2018	1807816778	
2014	1508677136	
2017	1300636712	
2016	1038373075	

Hình 3.54 Kết quả truy vấn câu 1 ở MSSQ

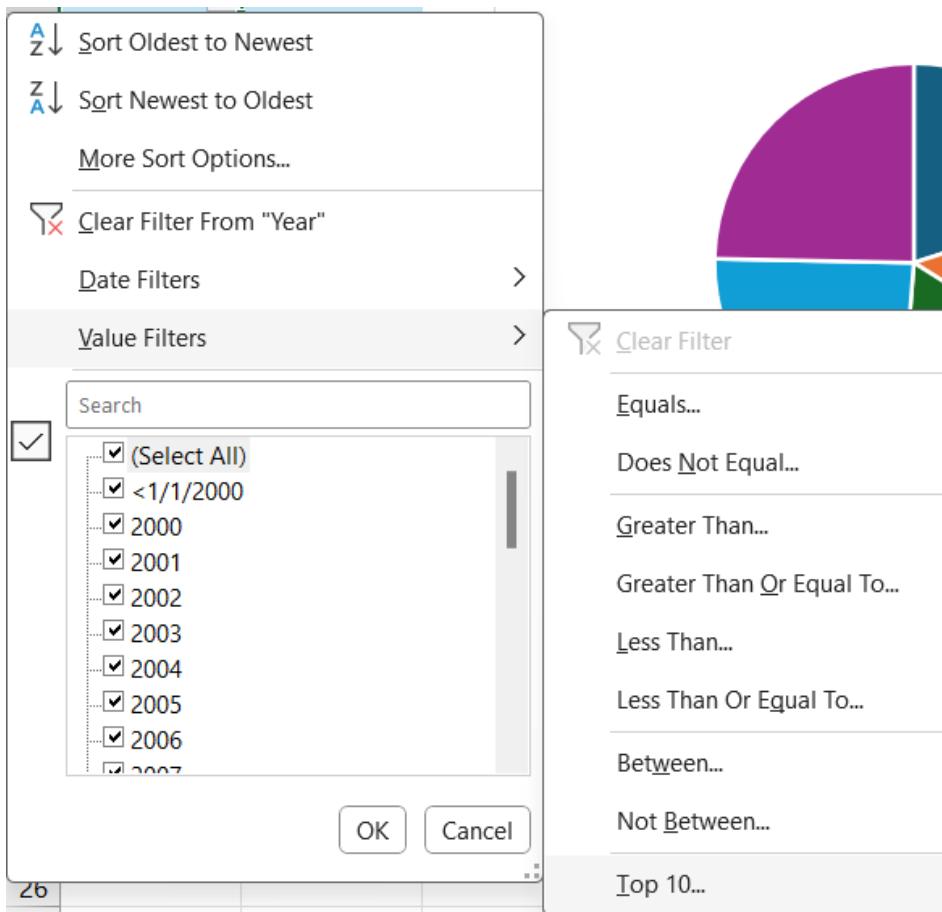
Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo [Revenue] trong bảng [Fact Movie] xuống ô **Values** và kéo [Year] trong bảng [Dim Date] qua ô **Rows**.



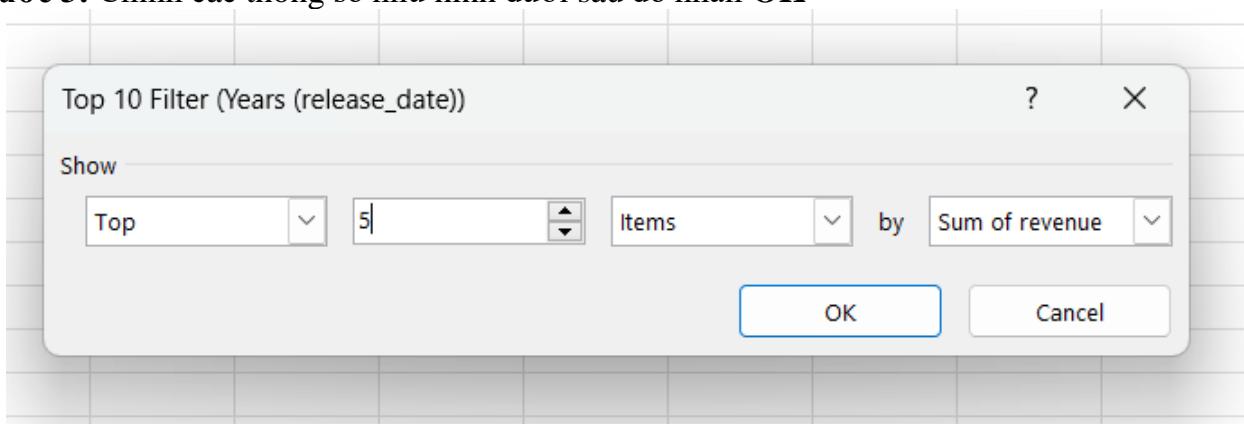
Hình 3.55 Thiết lập PivotTable Fields của câu 1

Bước 2: Click biểu tượng cạnh **Row Labels** chọn **Value Filters**, sau đó chọn **Top 10**



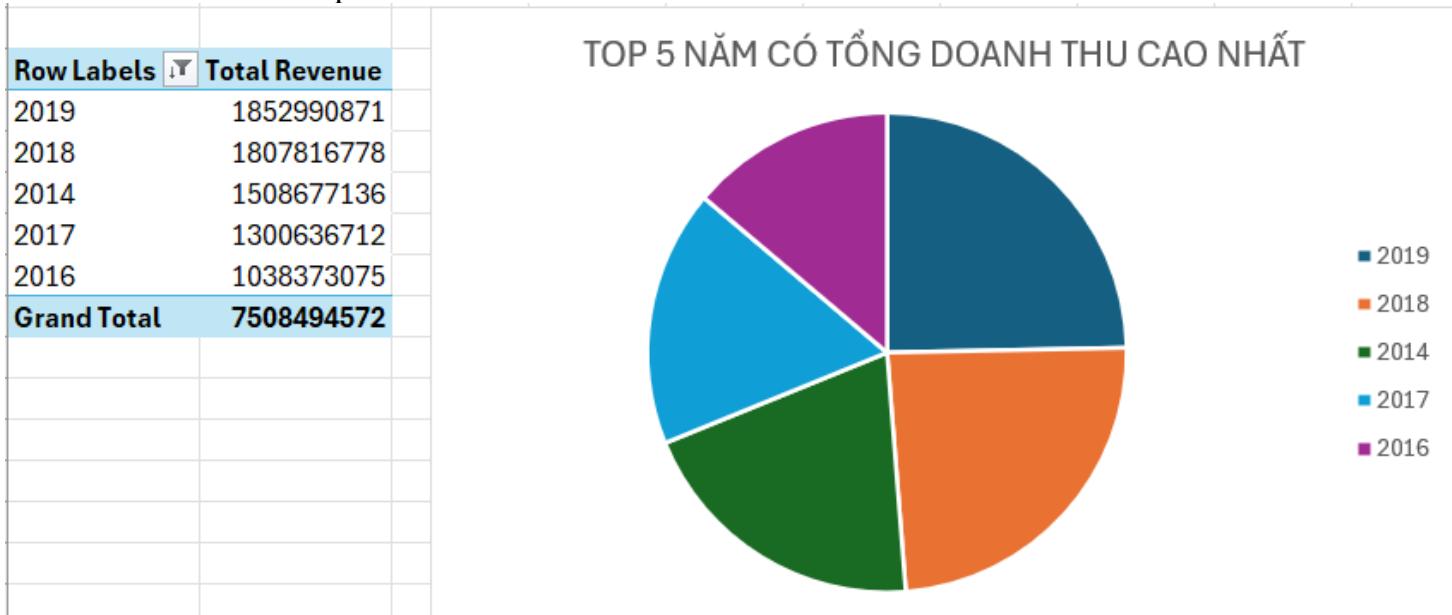
Hình 3.56 Điều chỉnh Value Filters

Bước 3: Chính các thông số như hình dưới sau đó nhấn **OK**



Hình 3.57 Điều chỉnh Top 5 Filter

Bước 4: Xem kết quả

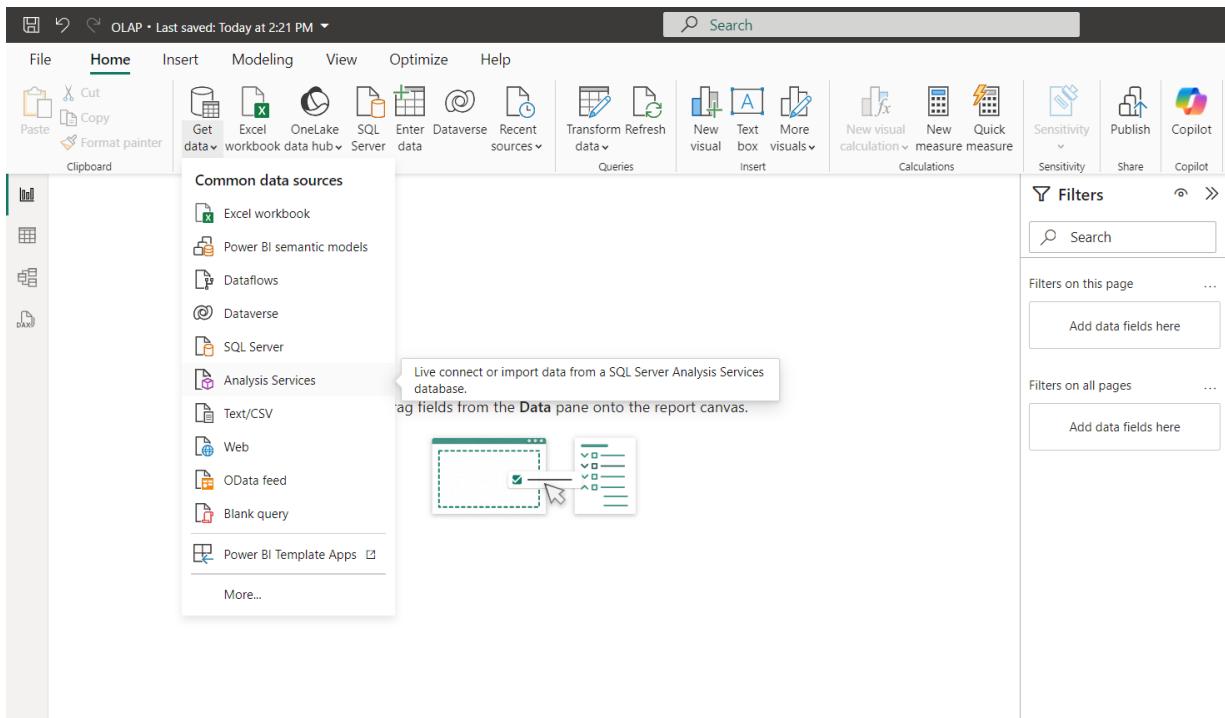


Hình 3.58 Kết quả truy vấn câu 1 ở Excel

Thực hiện trong Power BI

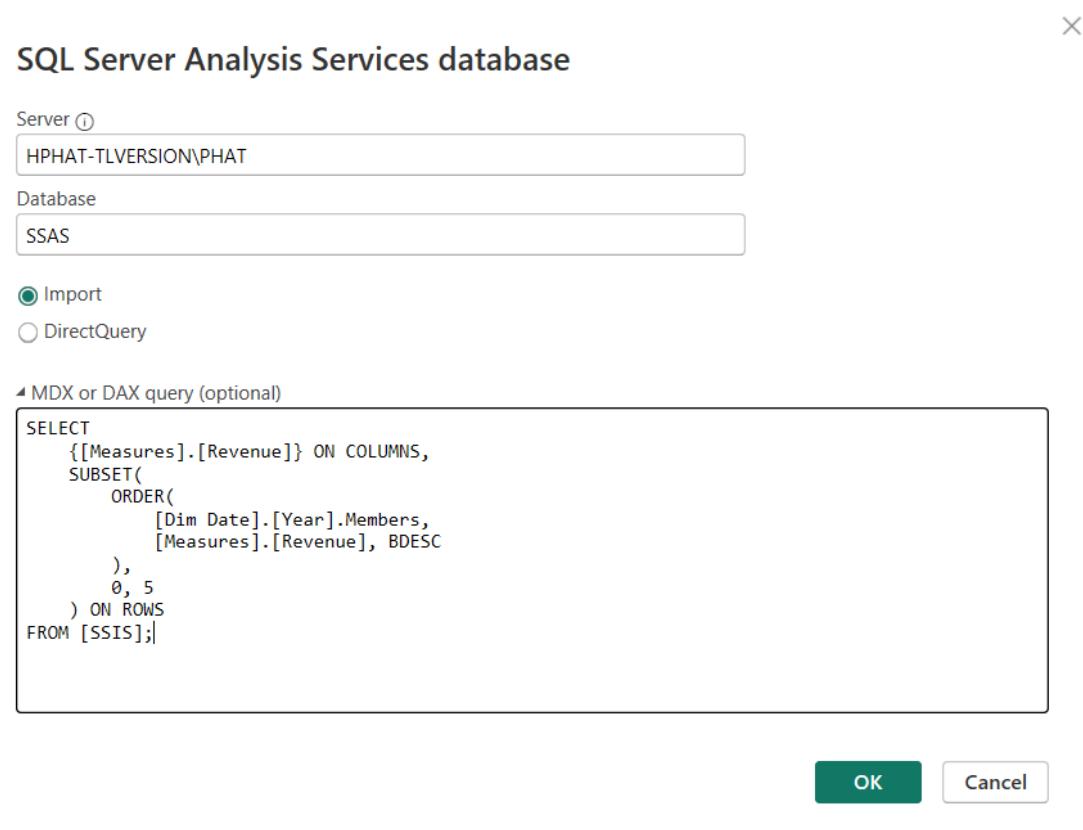
Bước 1: Trong Tab **Home**, nhấn vào **Get data** và chọn **Analysis Services**.

Kho dữ liệu và OLAP - IS217.P12



Hình 3.59 Kết nối Analysis Services để lấy data

Bước 2: Nhập Server và Database đã tạo ở SSAS. Chọn Import và nhập truy vấn MDX.



Hình 3.60 Thiết lập truy vấn MDX của câu 1 ở Power BI

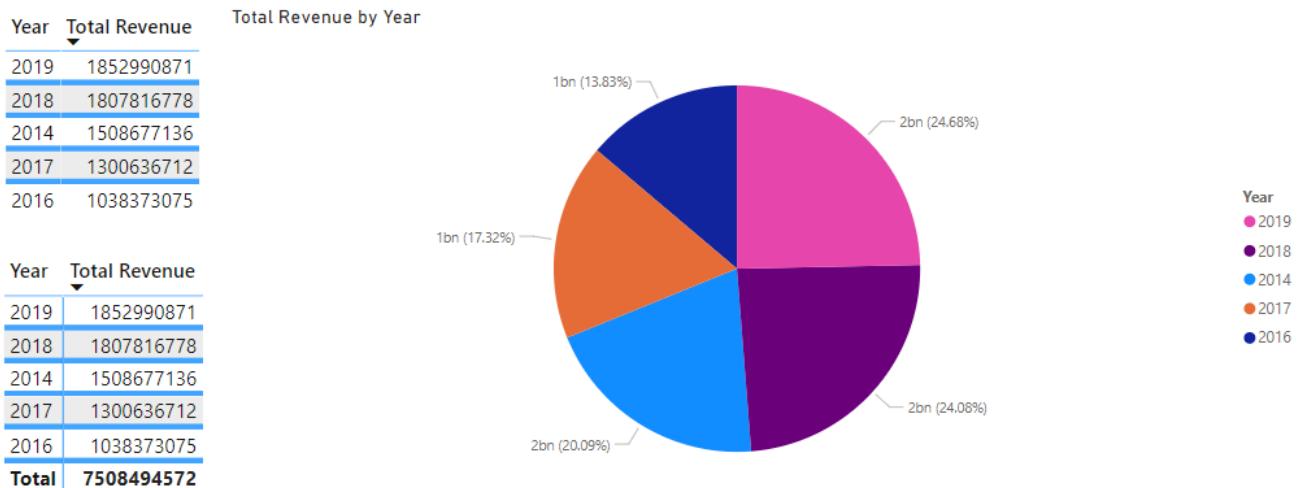
Bước 3: Ở Tab **Table view** điều chỉnh tên cột và kiểu dữ liệu phù hợp.

The screenshot shows the Microsoft Power BI ribbon with the 'Table tools' tab selected. Under the 'Column tools' section, the 'Name' is set to 'Total Revenue' and 'Data type' is set to 'Decimal number'. The 'Format' dropdown shows '\$' and '0' decimal places. In the 'Properties' group, 'Summarization' is set to 'Don't summarize' and 'Data category' is set to 'Uncategorized'. Below the ribbon, a table structure pane displays a table with columns 'Year' and 'Total Revenue'. The table contains five rows of data: 2019 (1852990871), 2018 (1807816778), 2014 (1508677136), 2017 (1300636712), and 2016 (1038373075). To the right of the table is a 'Data' pane showing a query named 'Query1' with a single row 'Total Revenue' under the 'Year' column.

Hình 3.61 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 1

Bước 4: Ở Tab **Report view** tạo **Visualizations** kiểu Table, Matrix và Chart.

TOP 5 NĂM CÓ TỔNG DOANH THU LỚN NHẤT



Hình 3.62 Kết quả của truy vấn câu 1 ở Power BI

3.6.2 Câu truy vấn 2: Top 10 bộ phim có kinh phí sản xuất cao nhất, xếp giảm dần

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Tạo NameSet [Cau_2] biểu thức như sau:

Kho dữ liệu và OLAP - IS217.P12

Name: [Cau_2]

Expression:

```
TOPCOUNT([Dim Movie].[Title].[Title].Members, 10, [Measures].[Budget])
```

No issues found

Additional Properties:

Type: Dynamic

Display folder:

Hình 3.63 Nameset [Cau_2]

Bước 2: Kéo thả thuộc tính [Title] trong bảng [Dim Movie] và độ đo [Budget] vào cửa sổ thực thi.

Sau đó kéo NameSet [Cau_2] vào Dimension

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Movie	Title	In	Cau_2	<input type="checkbox"/> <input type="checkbox"/>

Hình 3.64 Thiết lập truy vấn câu 2 ở Browser

Bước 3: “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Kho dữ liệu và OLAP - IS217.P12

The screenshot shows the Microsoft Analysis Services cube editor in Visual Studio 2022. On the left, there's a navigation pane with 'SSIS' selected, showing 'Measures', 'Fact Movie' (with 'Budget' under it), 'Calculated Members', and 'KPIs'. In the center, there's a large table grid titled 'Title' and 'Budget'. The table lists various movie titles with their corresponding budgets. A filter expression 'Cau_2' is applied to the 'Dim Movie' dimension. The table data is as follows:

Title	Budget
Double Wedding	104002432
Fast & Furious 6	160000000
Gods of Egypt	140000000
Incredibles 2	200000000
Lemony Snicket's A Series of Unfortunate Events	140000000
Mars Needs Moms	150000000
Noah	125000000
Star Wars: The Force Awakens	245000000
Star Wars: The Rise of Skywalker	250000000
Turbo	135000000

Hình 3.65 Kết quả truy vấn câu 2 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```
SELECT
{[Measures].[Budget]} ON COLUMNS,
TOPCOUNT(
    [Dim Movie].[Title].[Title].Members,
    10,
    [Measures].[Budget]
) ON ROWS
FROM [SSIS];
```

The screenshot shows the 'Results' tab in SQL Server Management Studio (MSSQ). It displays the same list of movies and their budgets as the previous screenshot, ordered by budget from highest to lowest. The table has two columns: 'Title' and 'Budget'.

Title	Budget
Star Wars: The Rise of Skywalker	250000000
Star Wars: The Force Awakens	245000000
Incredibles 2	200000000
Fast & Furious 6	160000000
Mars Needs Moms	150000000
Gods of Egypt	140000000
Lemony Snicket's A Series of Unfortunate Events	140000000
Turbo	135000000
Noah	125000000
Double Wedding	104002432

Hình 3.66 Kết quả truy vấn câu 2 ở MSSQ

Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo [Budget] trong bảng [Fact Movie] xuống ô

SVTH: Nguyễn Hồng Phát

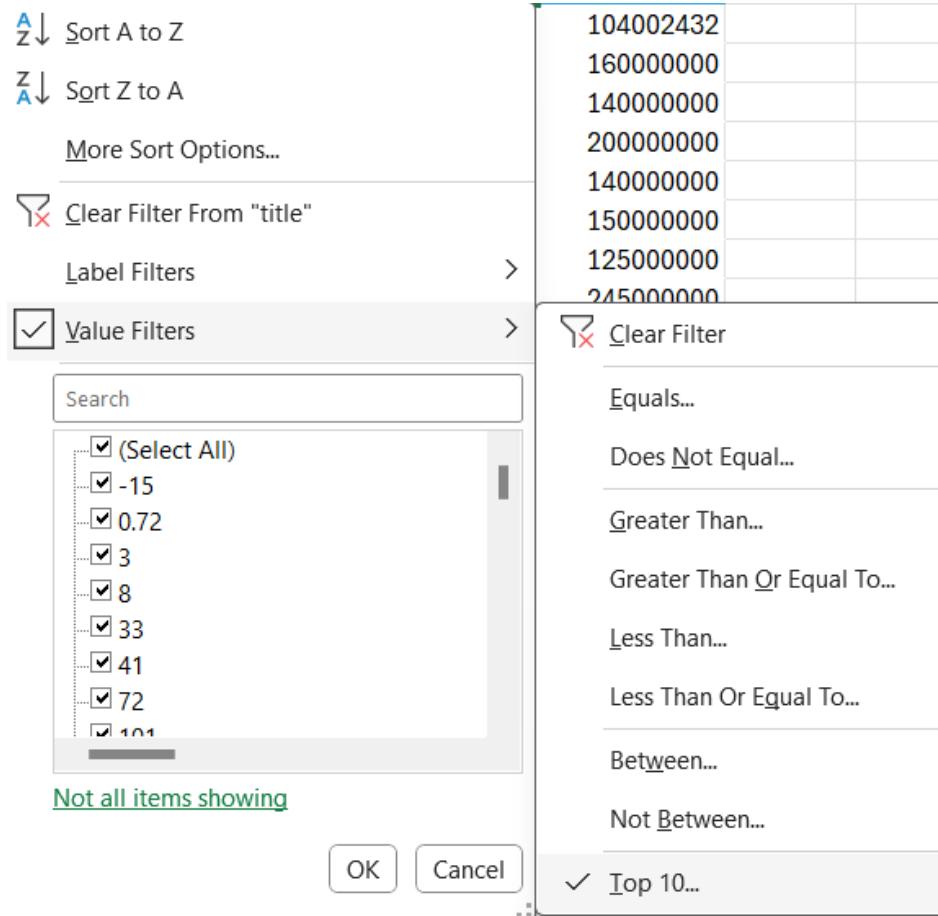
Values và kéo [title] trong bảng [Dim Movie] qua ô **Rows**.

Drag fields between areas below:

Filters	Columns
Rows	Σ Values
'title'	'Sum of budget'

Hình 3.67 Thiết lập PivotTable Fields của câu 2

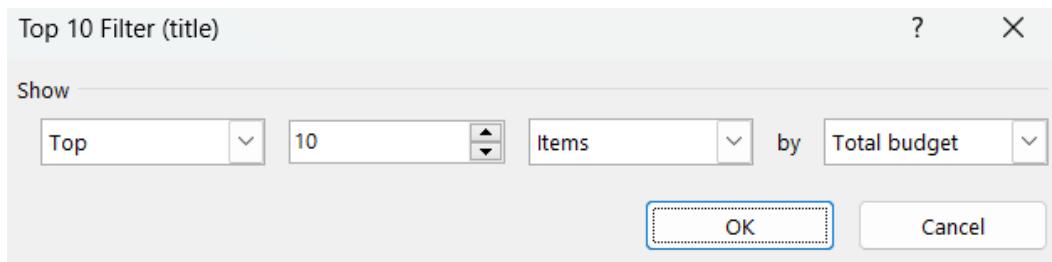
Bước 2: Click biểu tượng cạnh **Row Labels** chọn **Value Filters**, sau đó chọn **Top 10**



Hình 3.68 Điều chỉnh Value Filters

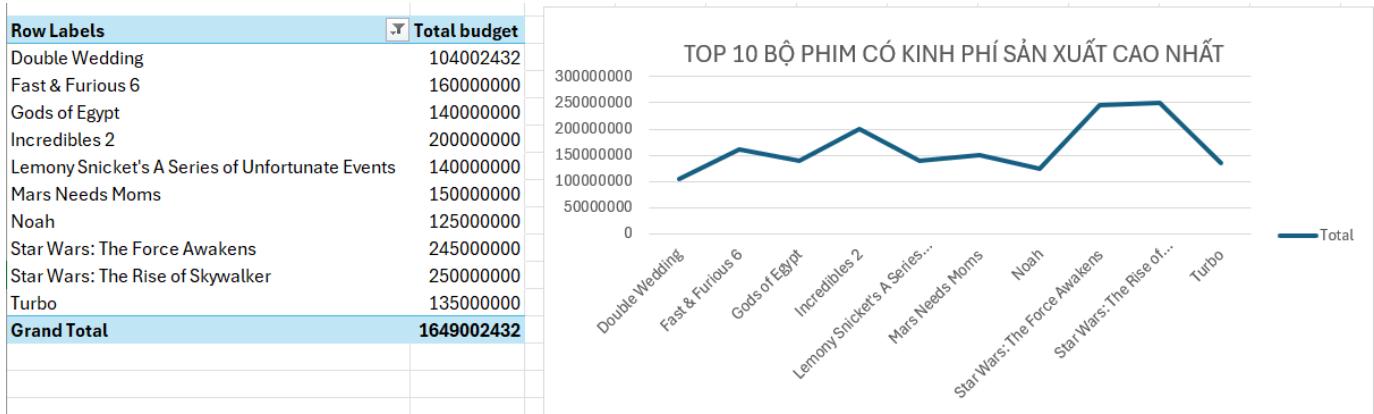
Kho dữ liệu và OLAP - IS217.P12

Bước 3: Chính các thông số như hình dưới sau đó nhấn **OK**



Hình 3.69 Điều chỉnh Top 10 Filter

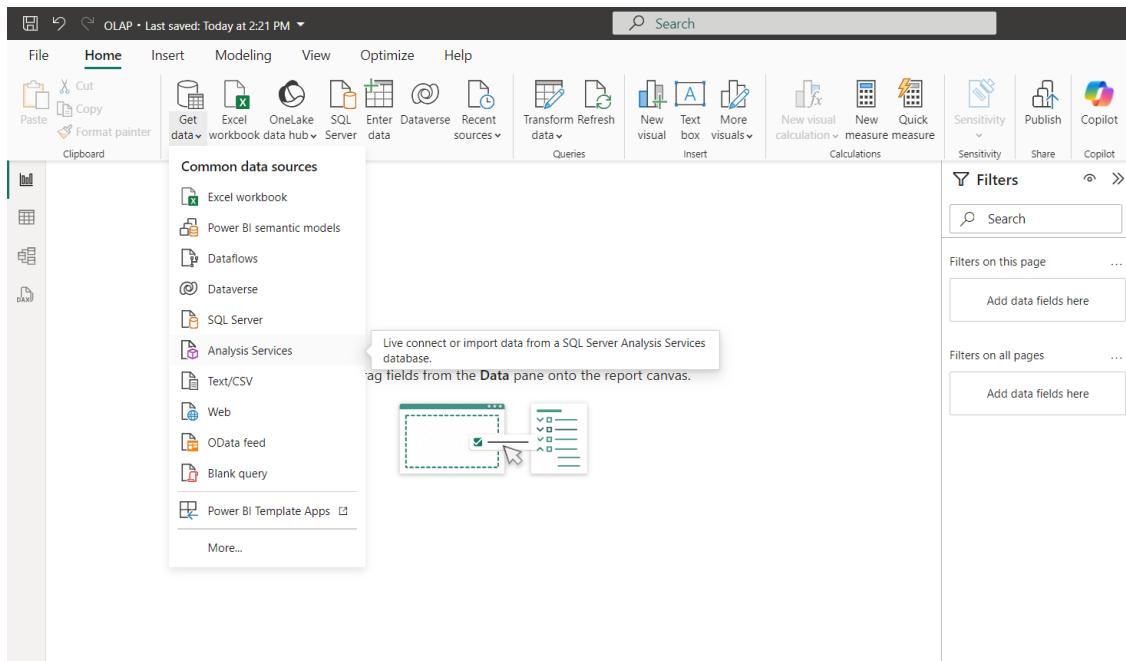
Bước 4: Xem kết quả



Hình 3.70 Kết quả truy vấn câu 2 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấn vào **Get data** và chọn **Analysis Services**.

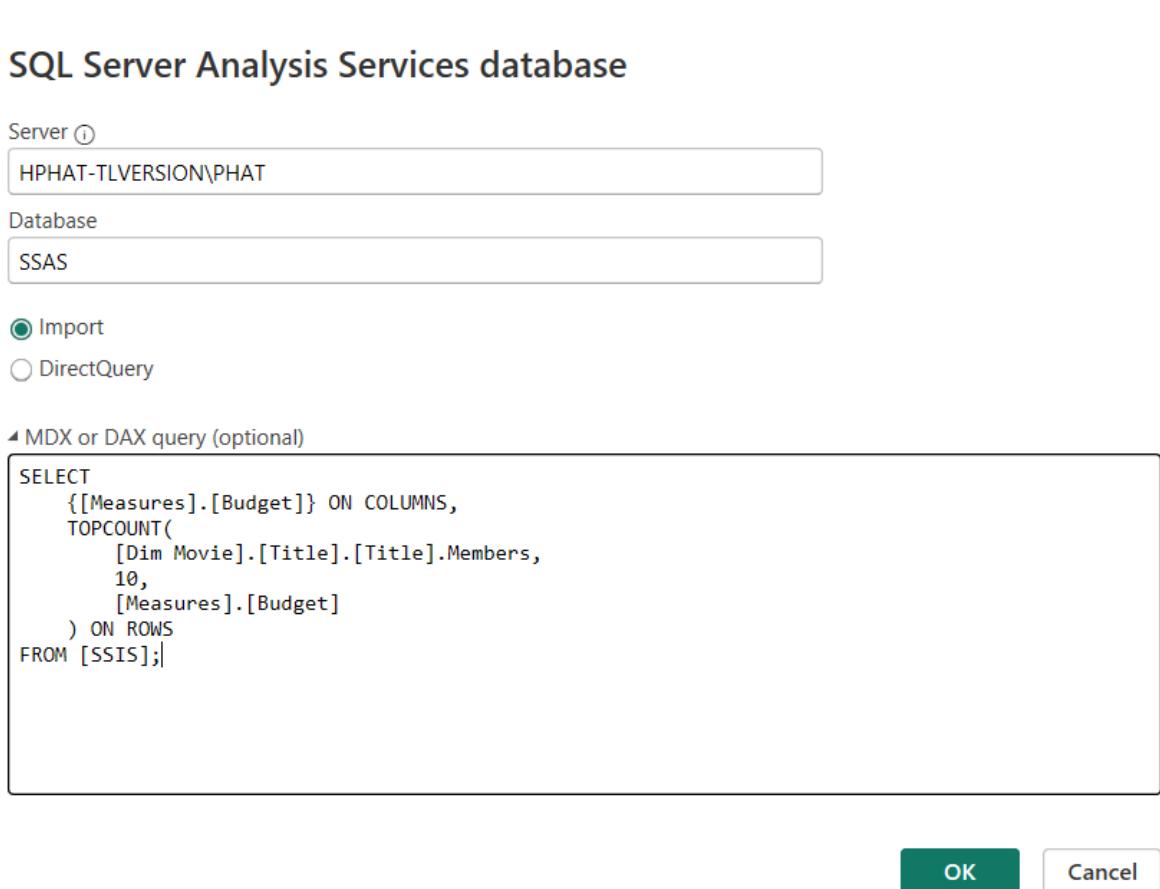


Hình 3.71 Kết nối Analysis Services để lấy data

SVTH: Nguyễn Hồng Phát

Kho dữ liệu và OLAP - IS217.P12

Bước 2: Nhập Server và Database đã tạo ở SSAS. Chọn Import và nhập truy vấn MDX.



Hình 3.72 Thiết lập truy vấn MDX của câu 2 ở Power BI

Bước 3: Ở Tab Table view điều chỉnh tên cột và kiểu dữ liệu phù hợp.

File Home Help Table tools Column tools Share

Name Total Budget Data type Decimal number \$. , , 0

Structure Format Properties

Summarization Sum Data category Uncategorized

Sort by column Sort Data groups Groups Manage Relationships New column Calculations

Title	Total Budget
Star Wars: The Rise of Skywalker	250000000
Star Wars: The Force Awakens	245000000
Incredibles 2	200000000
Fast & Furious 6	160000000
Mars Needs Moms	150000000
Gods of Egypt	140000000
Lemony Snicket's A Series of Unfortunate Events	140000000
Turbo	135000000
Noah	125000000
Double Wedding	104002432

Data

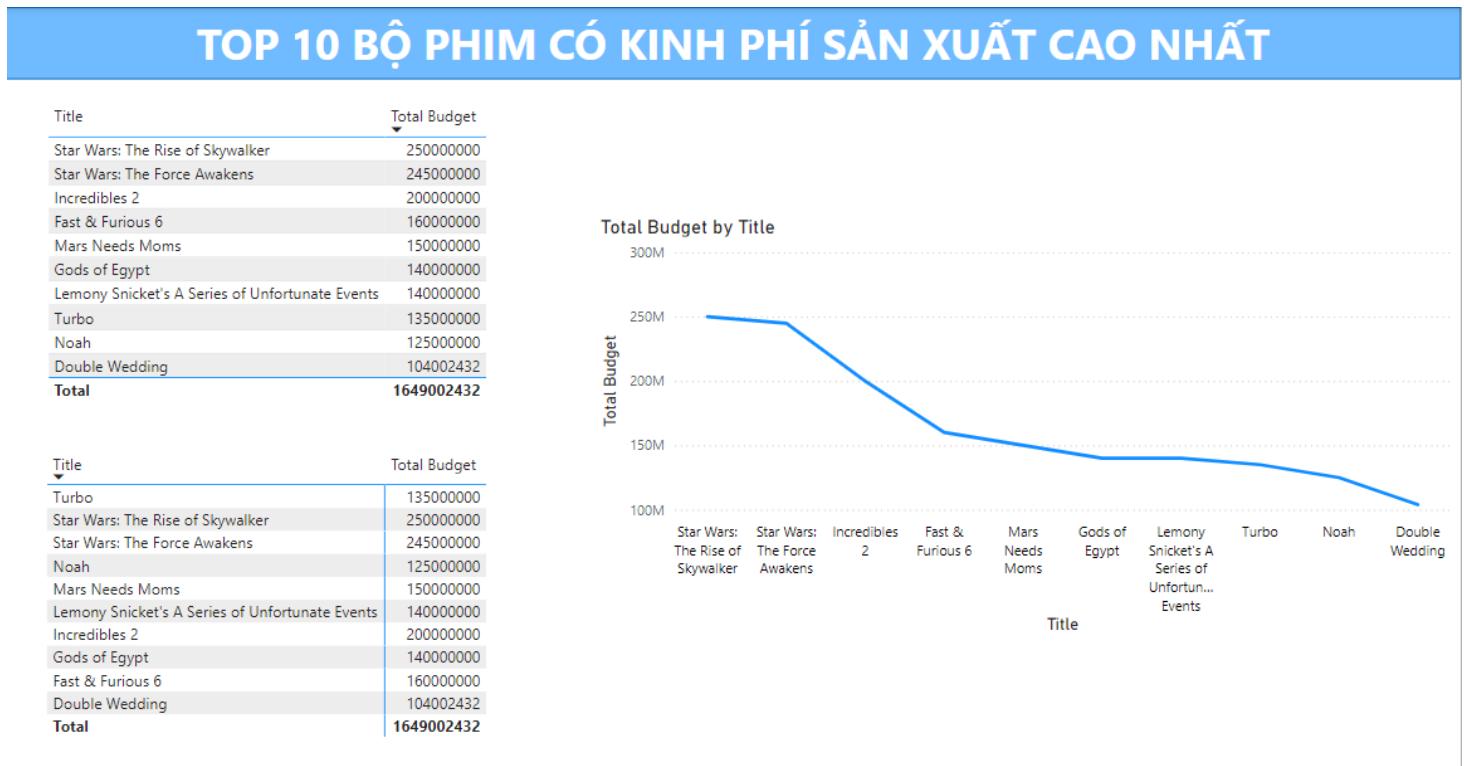
Search

Query1
Total Revenue
Year

Query2
Title
Σ Total Budget

Hình 3.73 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 2

Bước 4: Ở Tab Report view tạo Visualizations kiểu Table, Matrix và Chart.



Hình 3.74 Kết quả của truy vấn câu 2 ở Power BI

3.6.3 Câu truy vấn 3: Top 10 phim có nhiều lượt đánh giá nhất theo thứ tự giảm dần của độ phổ biến

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Tạo NameSet [Cau_3] biểu thức như sau:

Name:
Cau_3

Expression

```
SUBSET([ORDER([Dim Movie].[Title].[Title].Members,[Measures].[Vote Count], BDESC), 0, 10)
```

Ln: 1 Ch: 8 SPC CRLF

Additional Properties

Type: Dynamic

Display folder:

Hình 3.75 Nameset [Cau_3]

Bước 2: Kéo thả thuộc tính [Title] trong bảng [Dim Movie] và độ đo [Popularity] và [Vote Count] vào cửa sổ thực thi. Sau đó kéo NameSet [Cau_3] vào Dimension

Bước 3: “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Dimension	Hierarchy	Operator	Filter Expression
Dim Movie	>Title	In	Cau_3
<Select dimension>			

Title	Popularity	Vote Count
Don't Breathe	24.9080009460449	6922
Fast & Furious 6	6.26300001144409	10021
Incredibles 2	49.2150001525879	12061
John Wick	55.7389984130859	17923
Noah	28.8819999694824	5846
Star Wars: The Force Awakens	66.7720031738281	18352
Star Wars: The Rise of Skywalker	89.6989974975586	9047
Ted 2	58.25	6892
The Hangover Part III	48.6290016174316	8097
Yes Man	26.2600002288818	6336

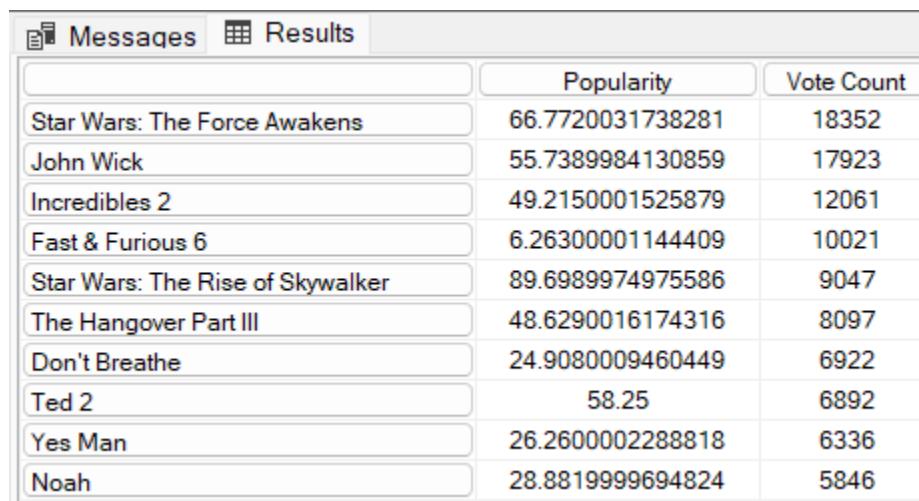
Hình 3.76 Kết quả truy vấn câu 3 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```

SELECT
    {[Measures].[Popularity],[Measures].[Vote Count]} ON COLUMNS,
    SUBSET(
        ORDER(
            [Dim Movie].[Title].[Title].Members,
            [Measures].[Vote Count], BDESC
        ), 0, 10) ON ROWS
FROM [SSIS];
  
```

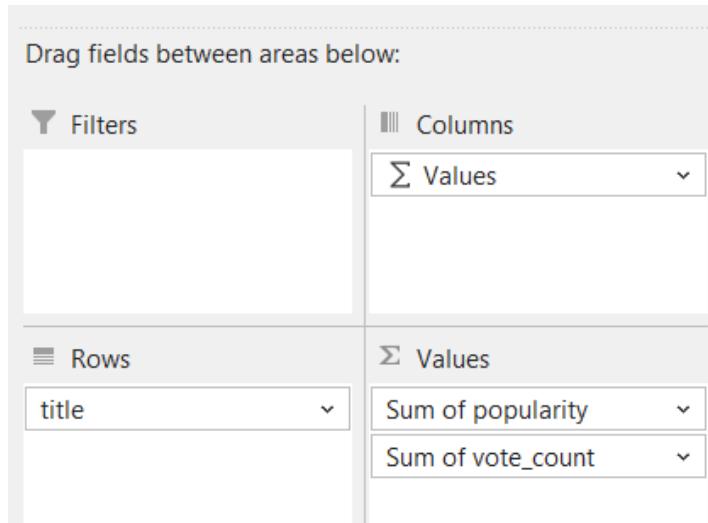


	Popularity	Vote Count
Star Wars: The Force Awakens	66.7720031738281	18352
John Wick	55.7389984130859	17923
Incredibles 2	49.2150001525879	12061
Fast & Furious 6	6.26300001144409	10021
Star Wars: The Rise of Skywalker	89.6989974975586	9047
The Hangover Part III	48.6290016174316	8097
Don't Breathe	24.9080009460449	6922
Ted 2	58.25	6892
Yes Man	26.2600002288818	6336
Noah	28.8819999694824	5846

Hình 3.77 Kết quả truy vấn câu 3 ở MSSQ

Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo [popularity] và [vote_count] trong bảng [Fact Movie] xuống ô **Values** và kéo [title] trong bảng [Dim Movie] qua ô **Rows**.



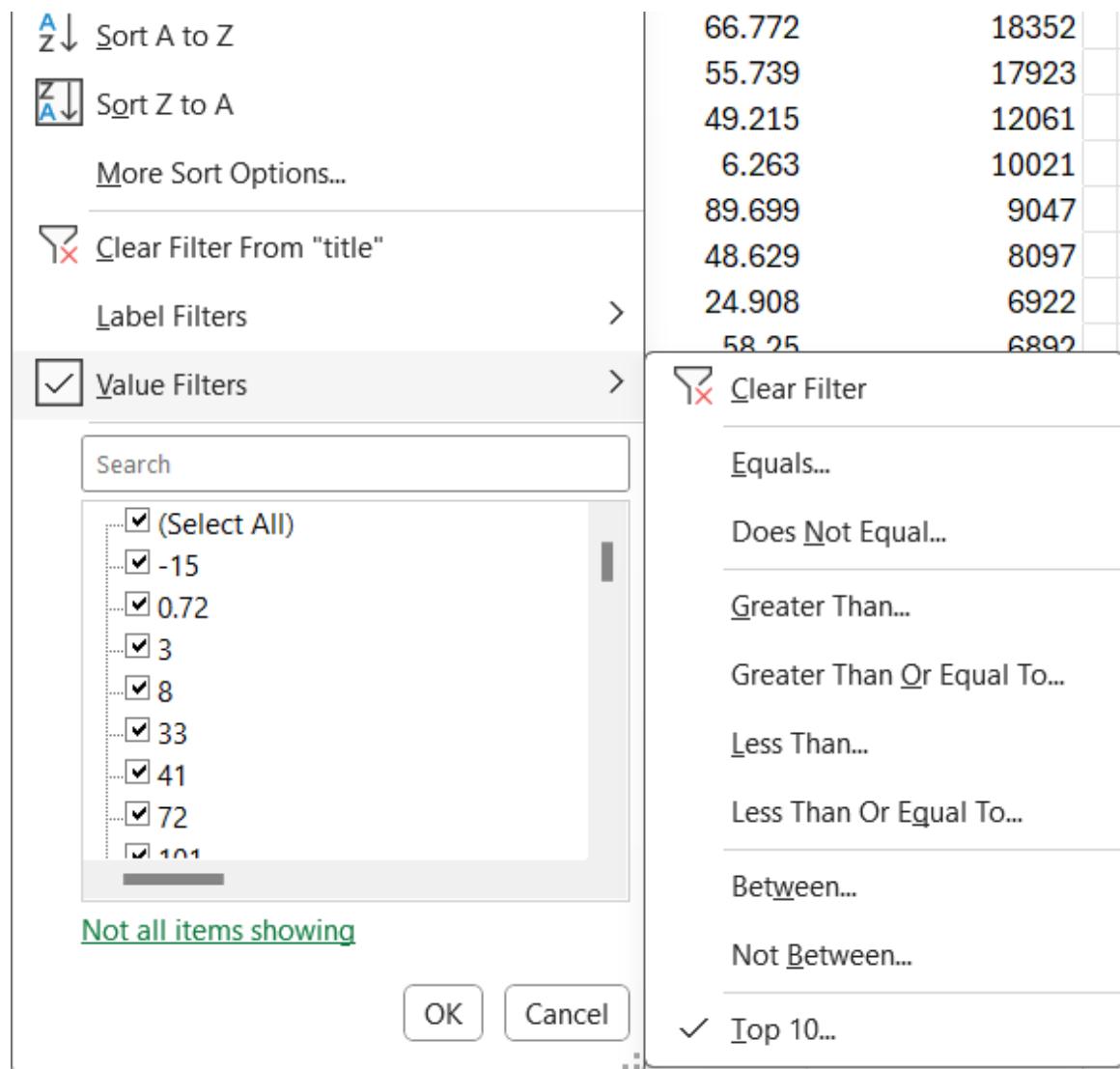
The screenshot shows the 'PivotTable Fields' dialog box with the following settings:

- Drag fields between areas below:** This section is empty.
- Filters:** No fields are listed.
- Columns:** A single field Σ Values is selected.
- Rows:** The field title is selected.
- Σ Values:** Two fields are listed: Sum of popularity and Sum of vote_count.

Hình 3.78 Thiết lập PivotTable Fields của câu 3

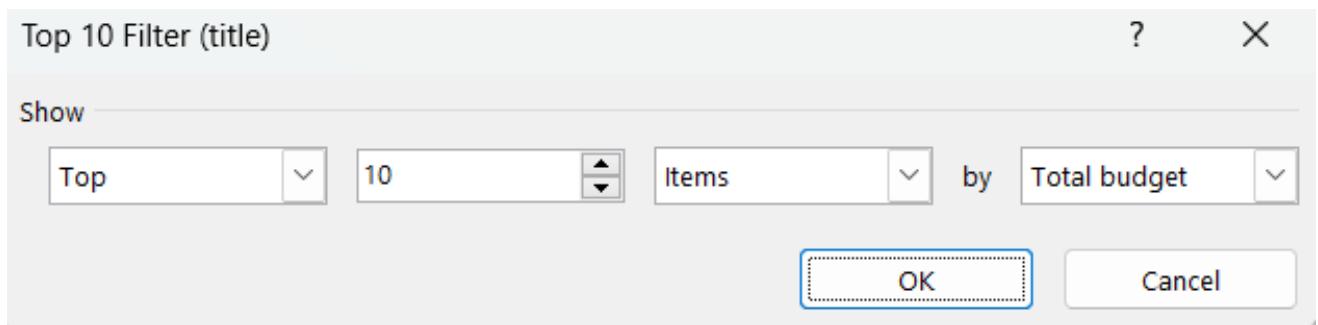
Bước 2: Click biểu tượng cạnh **Row Labels** chọn **Value Filters**, sau đó chọn **Top 10**

Kho dữ liệu và OLAP - IS217.P12



Hình 3.79 Điều chỉnh Value Filters

Bước 3: Chính các thông số như hình dưới sau đó nhấn **OK**



Hình 3.80 Điều chỉnh Top 10 Filter

Bước 4: Xem kết quả

Kho dữ liệu và OLAP - IS217.P12



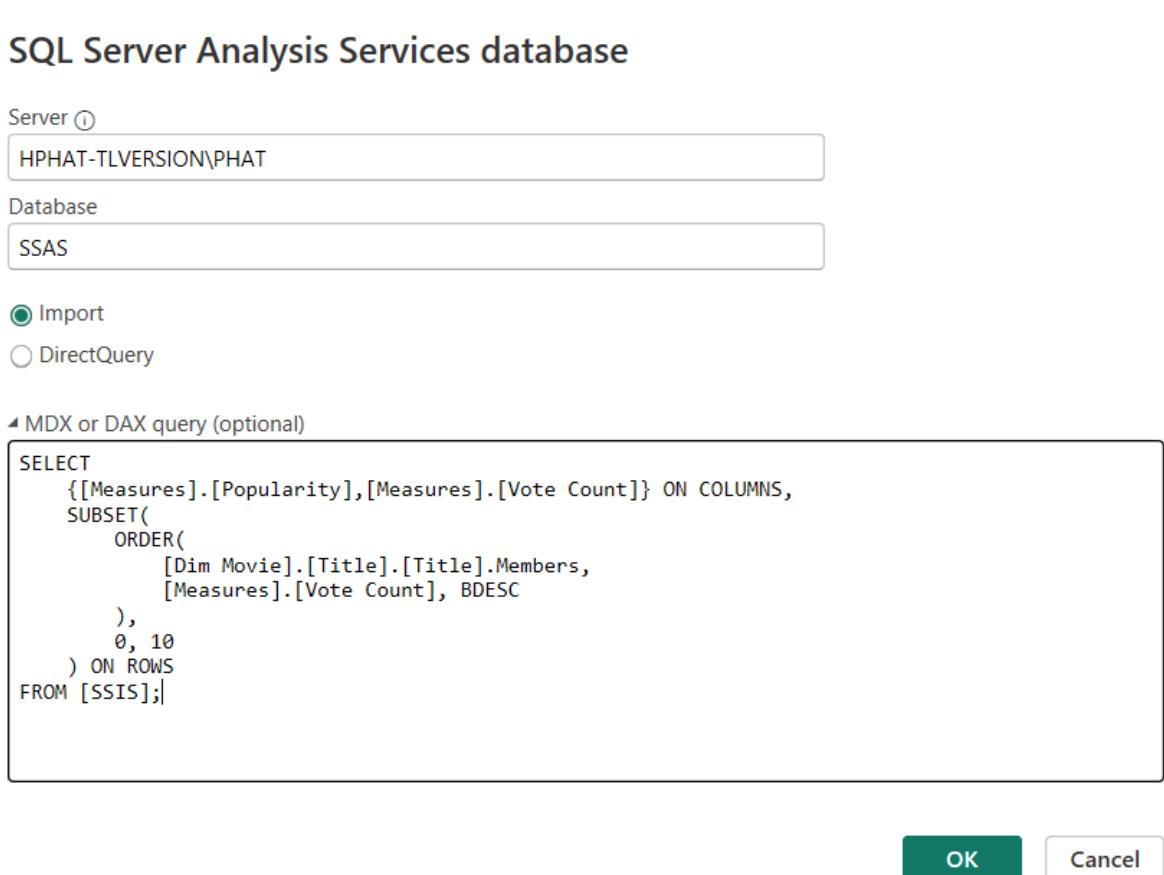
Hình 3.81 Kết quả truy vấn câu 3 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấn vào **Get data** và chọn **Analysis Services**.

Bước 2: Nhập **Server** và **Database** đã tạo ở SSAS. Chọn **Import** và nhập truy vấn MDX.

Kho dữ liệu và OLAP - IS217.P12



Hình 3.82 Thiết lập truy vấn MDX của câu 3 ở Power BI

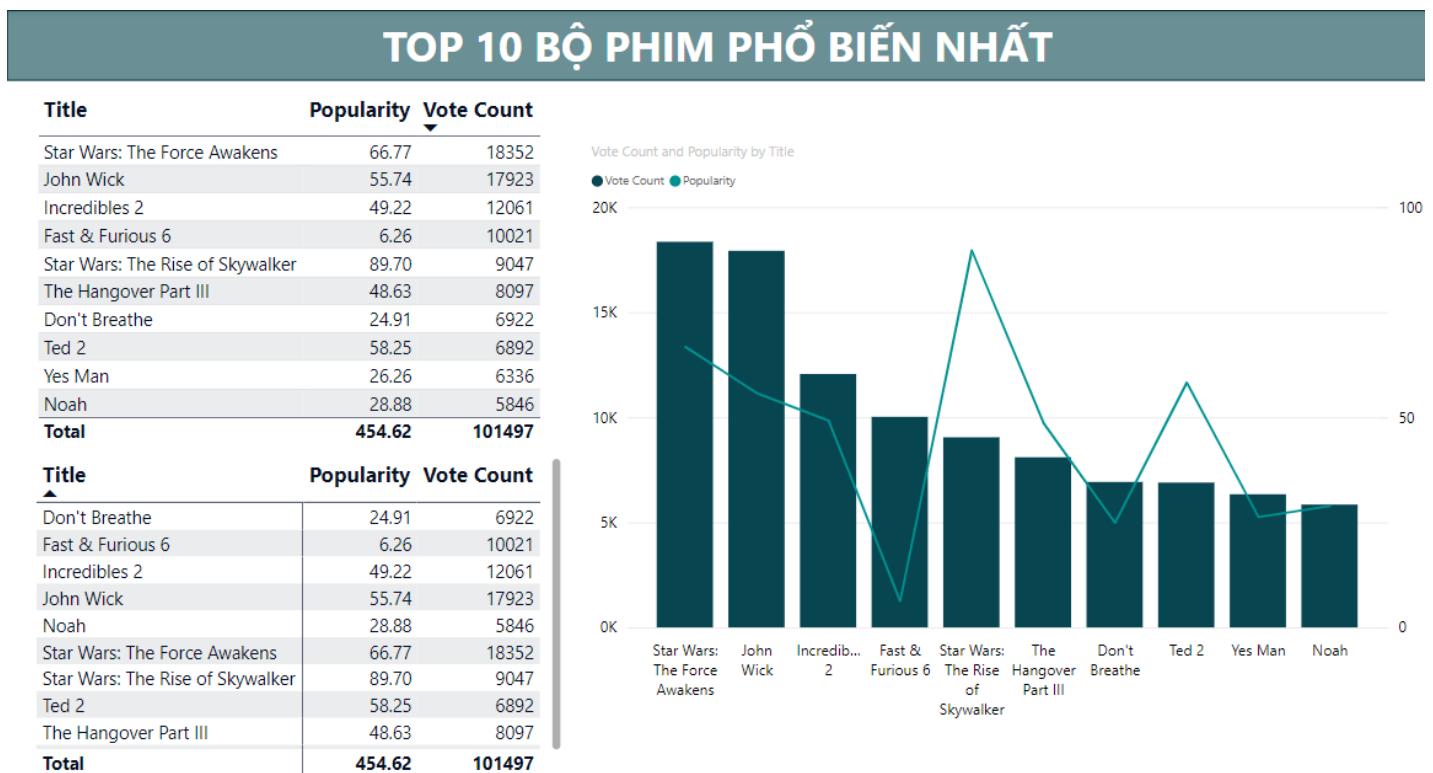
Bước 3: Ở Tab Table view điều chỉnh tên cột và kiểu dữ liệu phù hợp.

The screenshot shows the 'Table tools' tab selected in the ribbon. The 'Column tools' section is active, showing settings for 'Name' (Popularity), 'Format' (Decimal number), 'Summarization' (Sum), and other properties like Data type (Decimal number). The main table view displays movie titles and their popularity and vote counts. To the right, the 'Data' pane shows the columns: Popularity, Title, and Vote Count.

Title	Popularity	Vote Count
Star Wars: The Force Awakens	66.77	18352
John Wick	55.74	17923
Incredibles 2	49.22	12061
Fast & Furious 6	6.26	10021
Star Wars: The Rise of Skywalker	89.70	9047
The Hangover Part III	48.63	8097
Don't Breathe	24.91	6922
Ted 2	58.25	6892
Yes Man	26.26	6336
Noah	28.88	5846

Hình 3.83 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 3

Bước 4: Ở Tab Report view tạo Visualizations kiểu Table, Matrix và Chart.



Hình 3.84 Kết quả của truy vấn câu 3 ở Power BI

3.6.4 Câu truy vấn 4: Truy vấn các tháng của mỗi năm 2020 có doanh thu hơn 10 triệu, sắp xếp theo thứ tự tăng dần trong tháng

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Ở Browser, phần Dimension, chọn điều kiện truy vấn: ở [Dim Date] là năm 2022 và ở [Dim Movie] là [title] và chọn custom như ảnh bên dưới.

Dimension	Hierarchy	Operator	Filter Expression
Dim Date	Year	Equal	{ 2020 }
Dim Movie	Title	Custom	ORDER(FILTER(DrillDownMember([Dim Date].[Date_Hierarchy].[Month], [Dim Date].[Date_Hierarchy].[Year]) * [Dim Movie].[Title].[Title].Members, [Measures].[Revenue] > 10000000), [Measures].[Revenue], ASC)

Expression Builder

```
Expression:
ORDER(
    FILTER(
        DrillDownMember(
            [Dim Date].[Date_Hierarchy].[Month],
            [Dim Date].[Date_Hierarchy].[Year]
        ) * [Dim Movie].[Title].[Title].Members,
        [Measures].[Revenue] > 10000000
    ),
    [Measures].[Revenue],
    ASC
)
```

Hình 3.85 Thiết lập điều kiện cho truy vấn câu 4

Bước 2: Kéo thả thuộc tính [Title] trong bảng [Dim Movie], [Month] trong bảng [Dim Date] và độ

Kho dữ liệu và OLAP - IS217.P12

đo [Revenue] vào cửa sổ thực thi.

Bước 3: “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Dimension	Hierarchy	Operator	Filter Expression
Dim Date	Year	Equal	{ 2020 }
Dim Movie	Title	Custom	ORDER(FILTER(DrillDownMember([Dim Date].[Da...)
<Select dimension>			
Month	Title	Revenue	
1	Street Dancer 3D	12589965	
1	The Turning	19428166	
2	Malang	11136444	
3	I Still Believe	16069730	
4	The New Mutants	49169594	
9	After We Collided	42000000	
9	Honest Thief	31220247	

Hình 3.86 Kết quả truy vấn câu 4 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

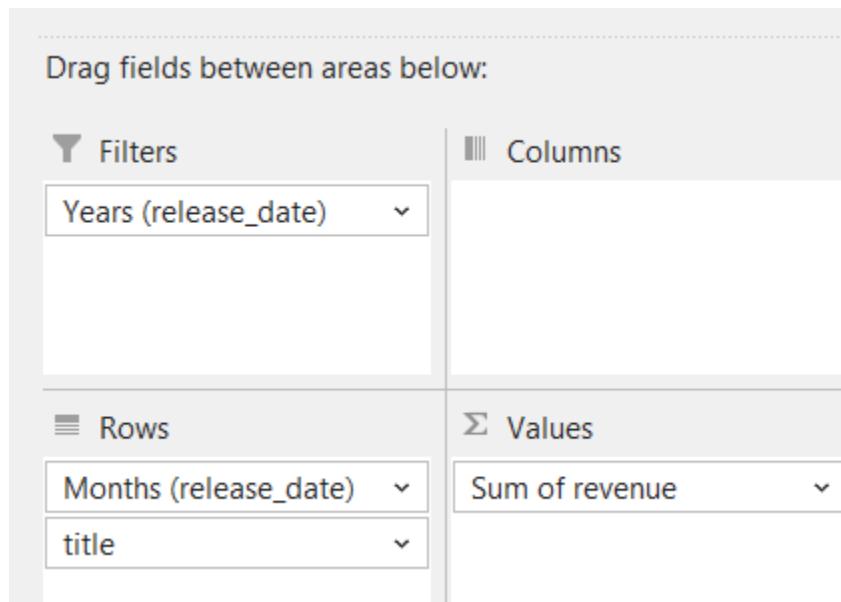
```
SELECT
    [Measures].[Revenue] ON COLUMNS,
    ORDER(
        FILTER(
            DrillDownMember(
                [Dim Date].[Date_Hierarchy].[Month],
                [Dim Date].[Date_Hierarchy].[Year]
            ) * [Dim Movie].[Title].[Title].Members,
            [Measures].[Revenue] > 10000000
        ),
        [Measures].[Revenue],
        ASC
    ) ON ROWS
FROM [SSIS]
WHERE [Dim Date].[Year].&[2020];
```

		Revenue
1	Street Dancer 3D	12589965
1	The Turning	19428166
2	Malang	11136444
3	I Still Believe	16069730
4	The New Mutants	49169594
9	Honest Thief	31220247
9	After We Collided	42000000

Hình 3.87 Kết quả truy vấn câu 4 ở MSSQ

Thực hiện trong Excel

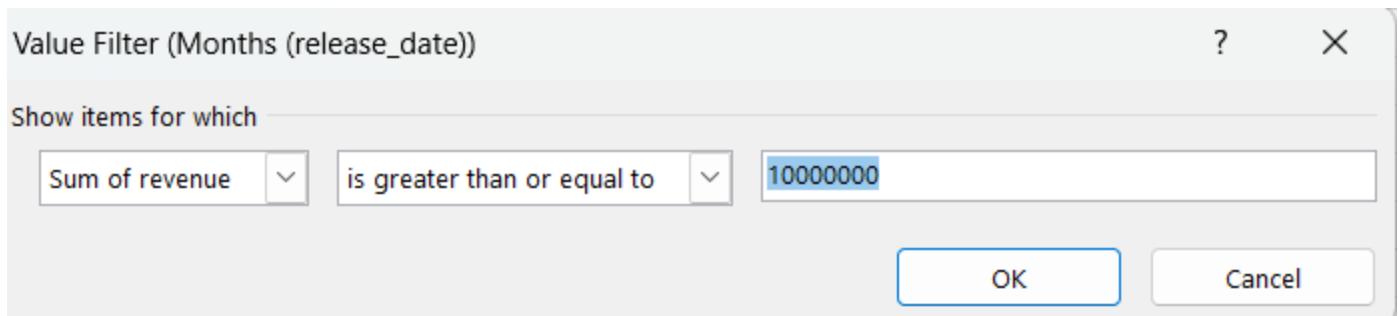
Bước 1: Trong PivotTable Fields, kéo [revenue] trong bảng [Fact Movie] xuống ô **Values** và kéo [title] và [Months (release_date)] qua ô **Rows**. Kéo [Years (release_date)] vào **Filters**.



Hình 3.88 Thiết lập PivotTable Fields của câu 4

Bước 2: Click biểu tượng cạnh **Row Labels** chọn **Value Filters**, sau đó chọn **Greater Than Or Equal To**

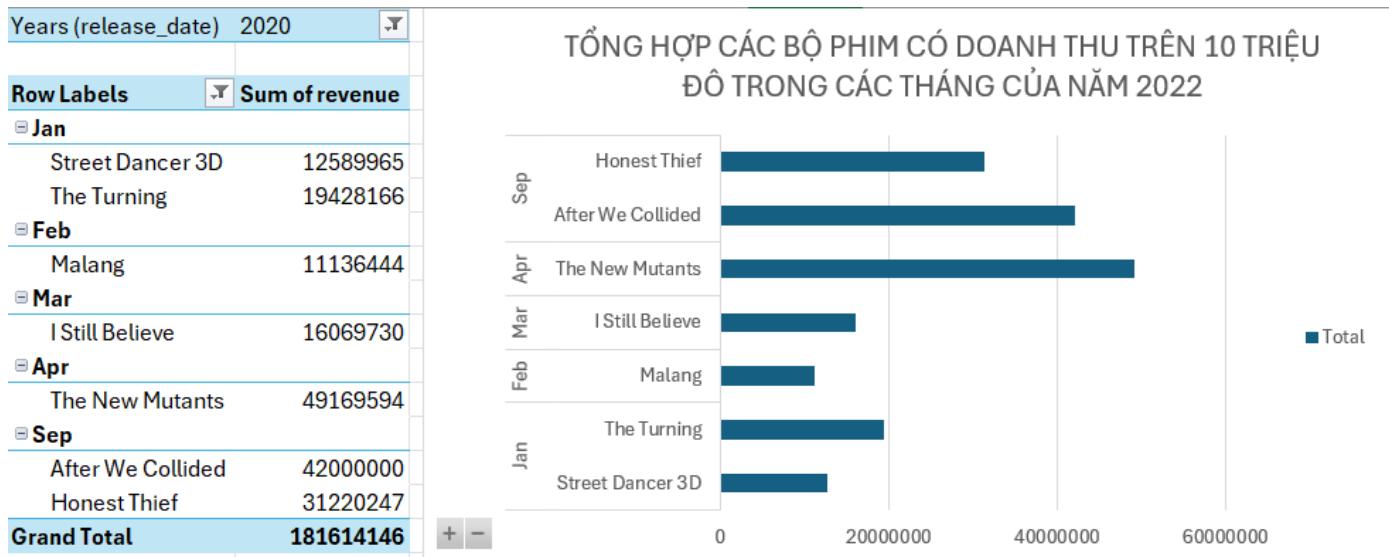
Bước 3: Chỉnh các thông số như hình dưới sau đó nhấn **OK**



Hình 3.89 Điều chỉnh Greater Than Or Equal To 10 triệu

Bước 4: Xem kết quả

Kho dữ liệu và OLAP - IS217.P12

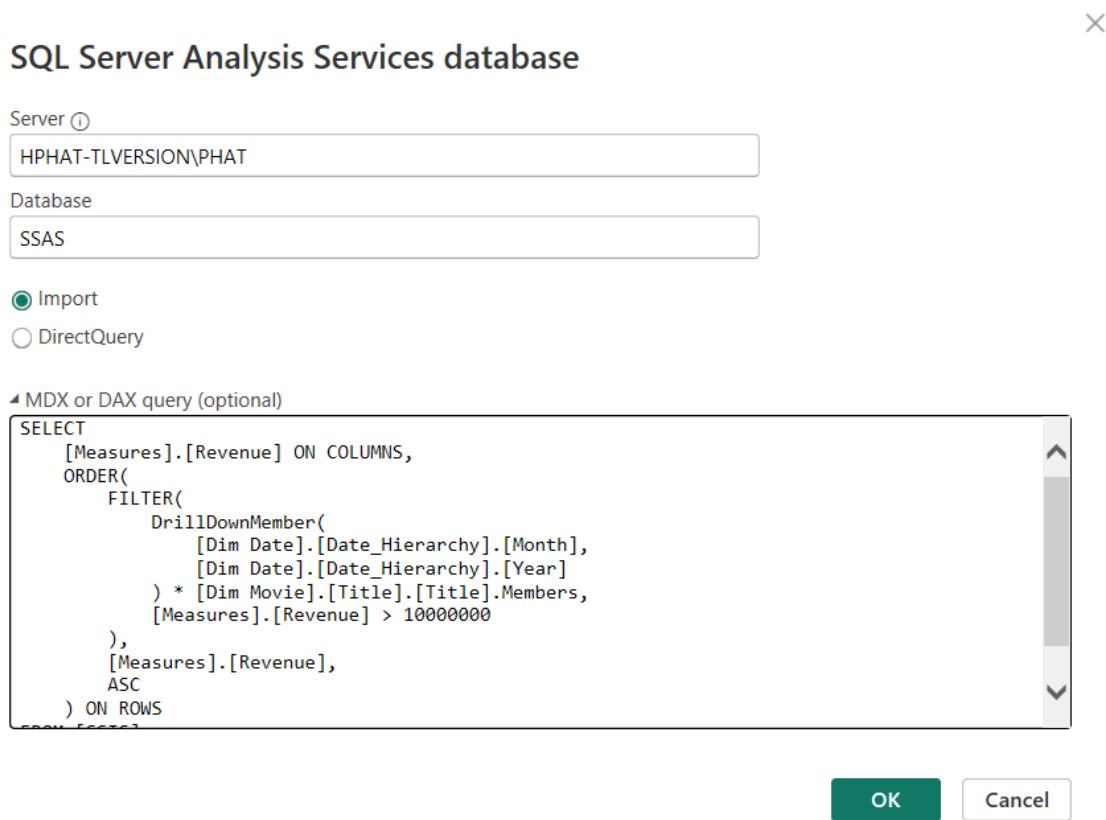


Hình 3.90 Kết quả truy vấn câu 4 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấn vào Get data và chọn Analysis Services.

Bước 2: Nhập Server và Database đã tạo ở SSAS. Chọn Import và nhập truy vấn MDX.



Hình 3.91 Thiết lập truy vấn MDX của câu 4 ở Power BI

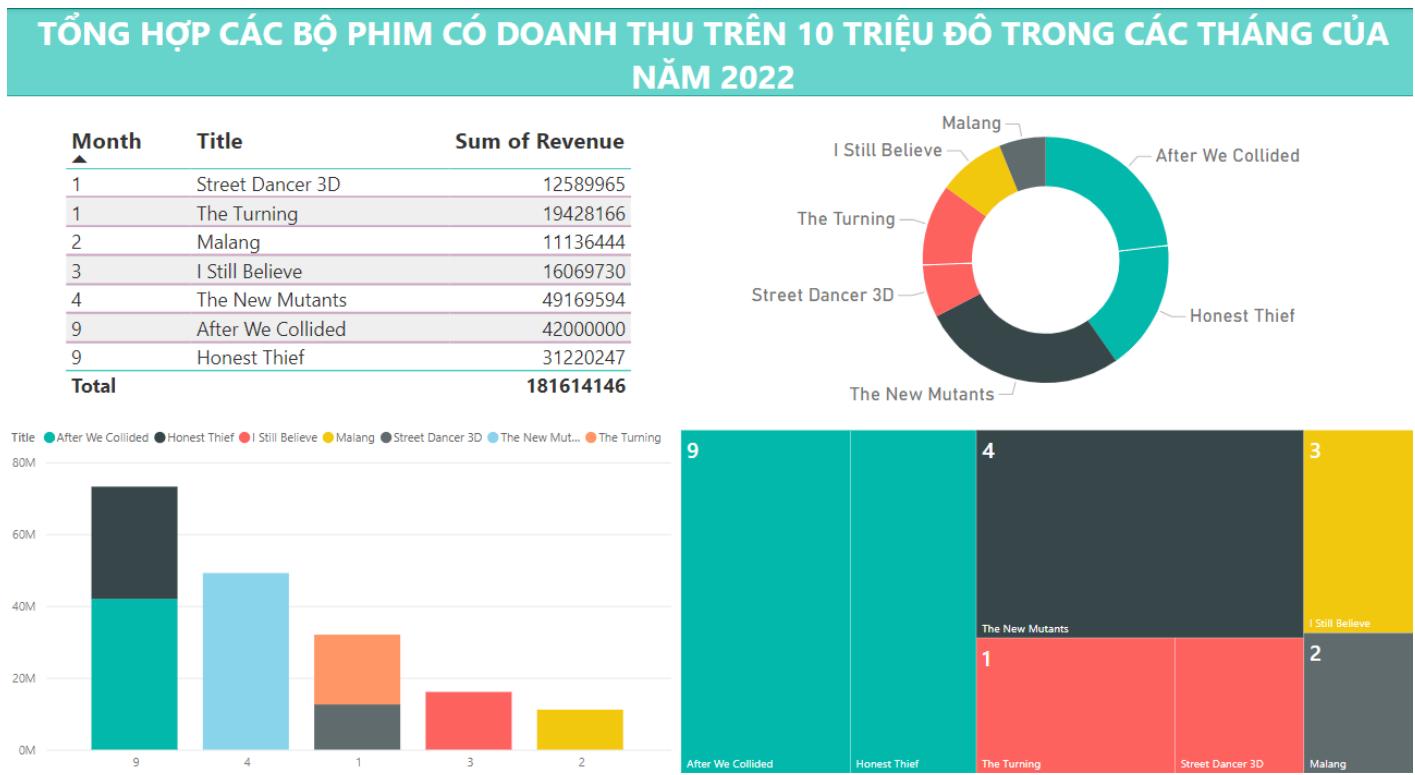
SVTH: Nguyễn Hồng Phát

Kho dữ liệu và OLAP - IS217.P12

Bước 3: Ở Tab **Table view** điều chỉnh tên cột và kiểu dữ liệu phù hợp.

Hình 3.92 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 4

Bước 4: Ở Tab **Report view** tạo **Visualizations** cho câu truy vấn.



Hình 3.93 Kết quả của truy vấn câu 4 ở Power BI

3.6.5 Câu truy vấn 5: Liệt kê top 3 quốc gia sản xuất có doanh thu cao nhất trong từng năm

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Ở Browser, phần **Dimension**, chọn điều kiện truy vấn: ở [Dim Country] chọn [Production Countries] và chọn custom như ảnh bên dưới.

Kho dữ liệu và OLAP - IS217.P12



Hình 3.94 Thiết lập điều kiện cho truy vấn câu 5

Bước 2: Kéo thả thuộc tính [Production Countries] trong bảng [Dim Country], [Year] trong bảng [Dim Date] và độ đo [Revenue] vào cửa sổ thực thi.

Bước 3: “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Dimension	Hierarchy	Operator	Filter Expression
Dim Country	Production Countries	Custom	GENERATE([Dim Date].[Year].Members, HEAD(...))
<Select dimension>			
Year			
2000	Australia, United States of America	Revenue	33463969
2000	France		64400000
2000	United States of America		83803717
2001	Germany, United States of America		40222729
2001	Switzerland, United States of America		29419291
2001	United States of America		111347552
2002	Germany, United States of America		65736816
2002	United Kingdom, United States of A...		7775138
2002	United States of America		31590839
2003	Germany, United States of America		19322135
2003	India		18083051
2003	United States of America		163496033
2004	Canada, United States of America		30120671
2004	Germany, Ireland, United Kingdom, ...		30031874

Hình 3.95 Kết quả truy vấn câu 5 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```

SELECT
    [Measures].[Revenue] ON COLUMNS,
    NON EMPTY
    GENERATE(
        [Dim Date].[Year].[Year].Members,
        HEAD(
            ORDER(
                [Dim Date].[Year].currentmember*[Dim
Country].[Production Countries].Children,
                [Measures].[Revenue], BDESC
            ),3 )
    ) ON ROWS
FROM [SSIS];
  
```

Messages		Results
		Revenue
1	Street Dancer 3D	12589965
1	The Turning	19428166
2	Malang	11136444
3	I Still Believe	16069730
4	The New Mutants	49169594
9	Honest Thief	31220247
9	After We Collided	42000000

Hình 3.96 Kết quả truy vấn câu 5 ở MSSQ

Thực hiện trong Excel

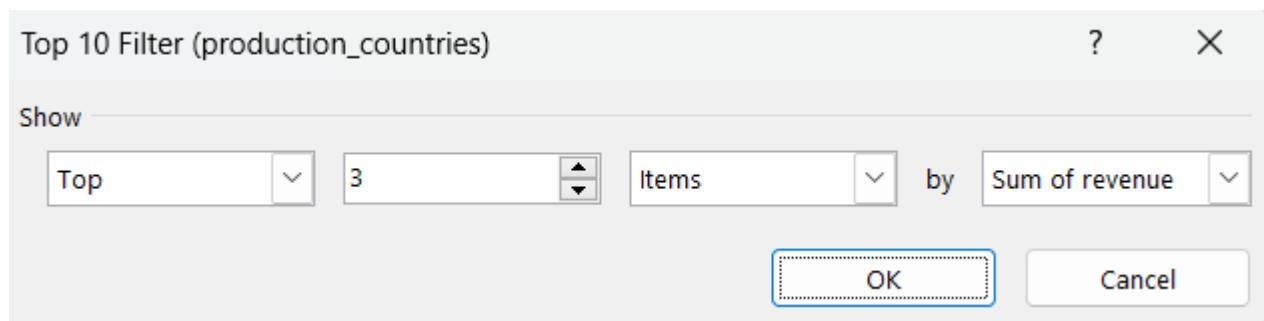
Bước 1: Trong PivotTable Fields, kéo [revenue] trong bảng [Fact Movie] xuống ô **Values** và kéo [production_companies] và [Years (release_date)] qua ô **Rows**.

The screenshot shows the 'PivotTable Fields' pane in Excel. At the top, it says 'Drag fields between areas below:' with four main sections: 'Filters' (empty), 'Columns' (empty), 'Rows' (containing 'Years (release_date)' and 'production_countries'), and 'Values' (containing 'Sum of revenue').

Hình 3.97 Thiết lập PivotTable Fields của câu 5

Bước 2: Click biểu tượng cạnh **Row Labels** chọn **Value Filters**, sau đó chọn **Top 10**

Bước 3: Chính các thông số như hình dưới sau đó nhấn **OK**



Hình 3.98 Điều chỉnh Top 3

Bước 4: Xem kết quả

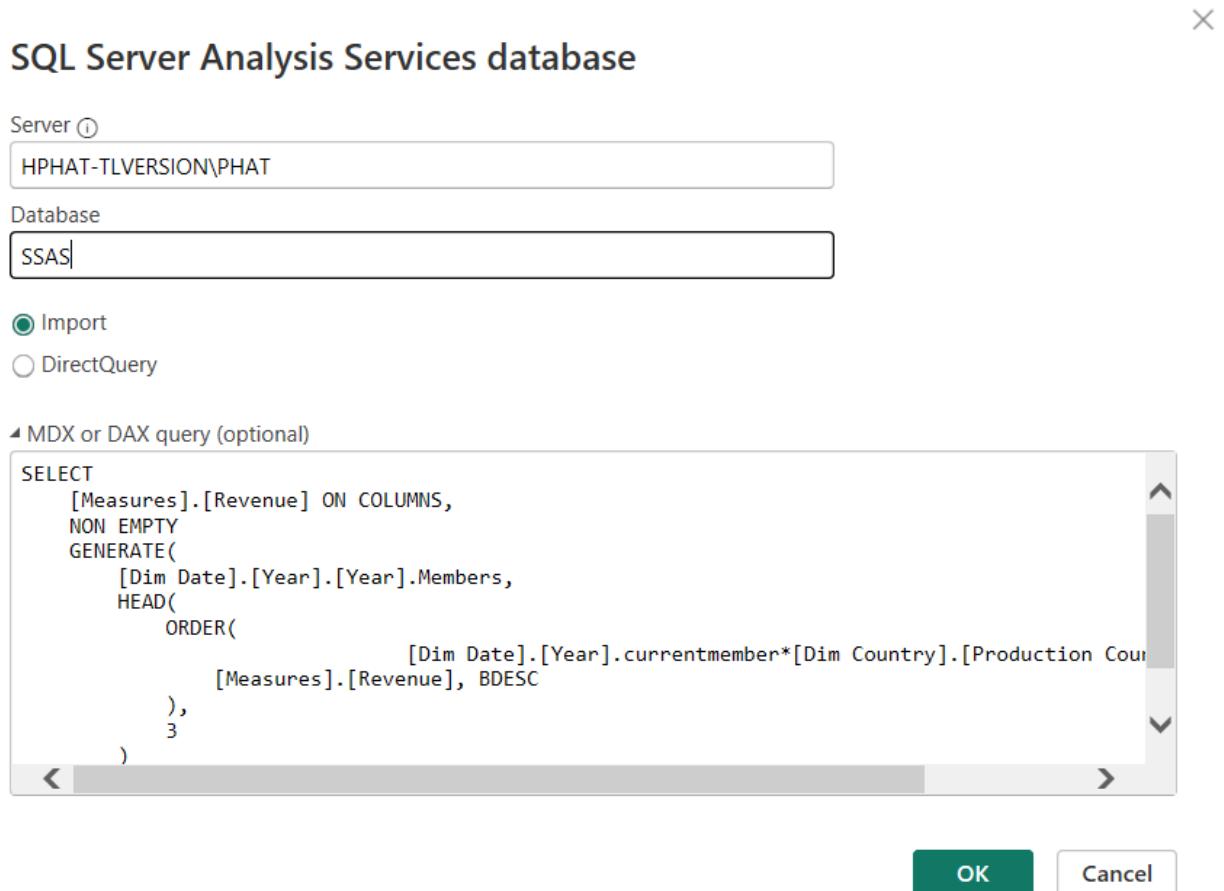
Row Labels	Sum of revenue
2000	
Australia, United States of America	33463969
France	64400000
United States of America	83803717
2001	
Germany, United States of America	40222729
Switzerland, United States of America	29419291
United States of America	111347552
2002	
Germany, United States of America	65736816
United Kingdom, United States of America	7775138
United States of America	31590839
2003	
Germany, United States of America	19322135
India	18083051
United States of America	163496033
2004	
Canada, United States of America	30120671
Germany, Ireland, United Kingdom, United States of America	30031874
Germany, United States of America	223073645
2005	

Hình 3.99 Kết quả truy vấn câu 5 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấn vào Get data và chọn Analysis Services.

Bước 2: Nhập Server và Database đã tạo ở SSAS. Chọn Import và nhập truy vấn MDX.



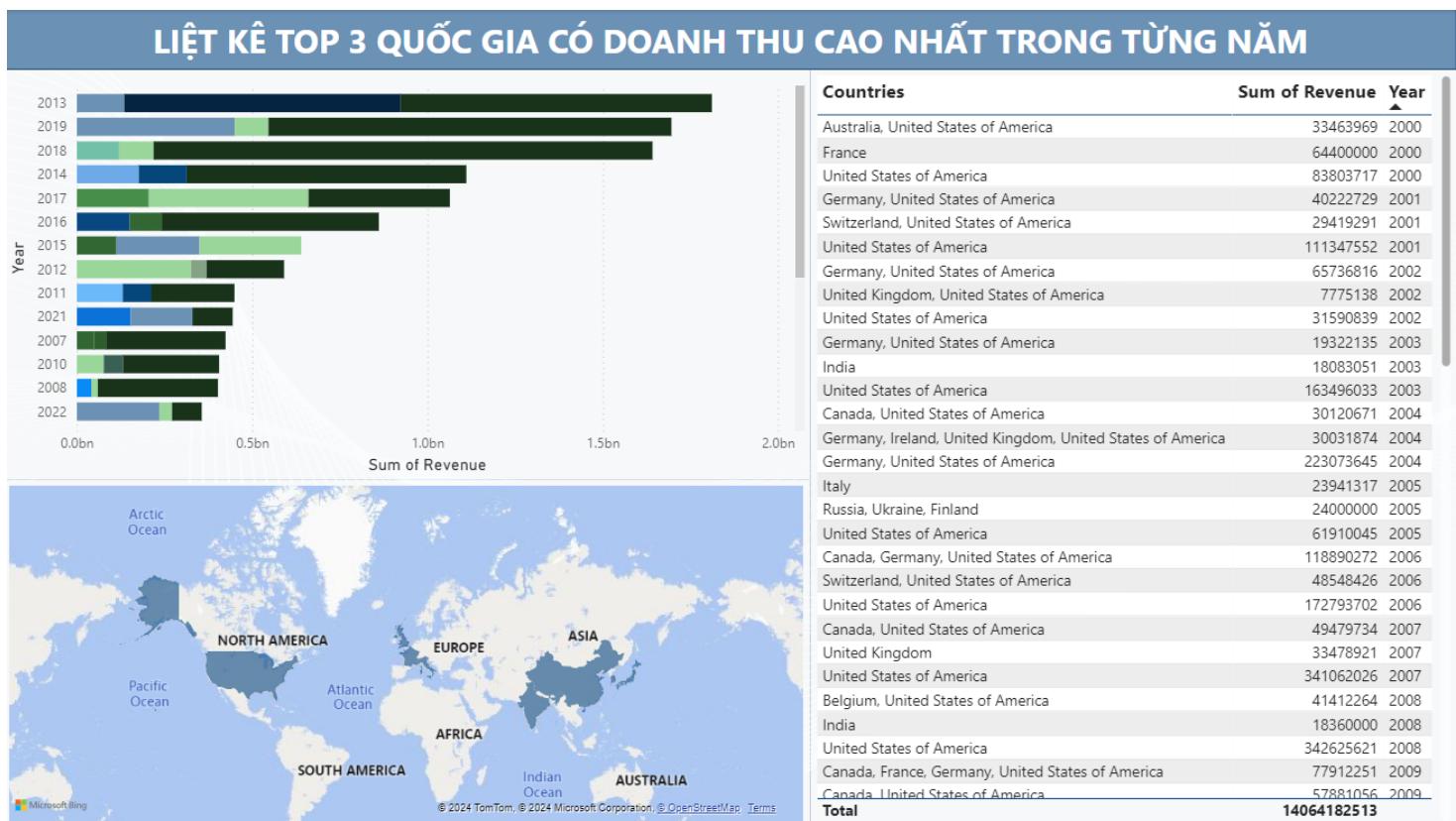
Hình 3.100 Thiết lập truy vấn MDX của câu 5 ở Power BI

Bước 3: Ở Tab Table view điều chỉnh tên cột và kiểu dữ liệu phù hợp.

Year	Countries	Revenue
2000	United States of America	83803717
2000	France	64400000
2000	Australia, United States of America	33463969
2001	United States of America	111347552
2001	Germany, United States of America	40222729
2001	Switzerland, United States of America	29419291
2002	Germany, United States of America	65736816
2002	United States of America	31590839
2002	United Kingdom, United States of America	7775138
2003	United States of America	163496033
2003	Germany, United States of America	19322135
2003	India	18083051
2004	Germany, United States of America	223073645
2004	Canada, United States of America	30120671

Hình 3.101 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 5

Bước 4: Ở Tab Report view tạo Visualizations cho câu truy vấn.



Hình 3.102 Kết quả của truy vấn câu 5 ở Power BI

3.6.6 Câu truy vấn 6: Truy vấn các quốc gia có doanh thu nằm trong top 10% của năm 2012 và đồng thời cũng nằm trong top 10% của năm 2013

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Tạo NameSet [Cau_6] biểu thức như sau:

Name: [Cau_6]

Expression:

```

INTERSECT(
    TOPPERCENT(
        [Dim Country].[Production Countries].Children,
        10,
        ([Measures].[Revenue], [Dim Date].[Year].&[2014])
    ),
    TOPPERCENT(
        [Dim Country].[Production Countries].Children,
        10,
        ([Measures].[Revenue], [Dim Date].[Year].&[2014])
    )
)

```

No issues found

Ln: 12 Ch: 6 SPC CRLF

Hình 3.103 Nameset [Cau_6]

Bước 2: Kéo thả thuộc tính [Production Countries] trong bảng [Dim Country], [Year] trong bảng

SVTH: Nguyễn Hồng Phát

Kho dữ liệu và OLAP - IS217.P12

[Dim Date] và độ đo [Revenue] vào cửa sổ thực thi. Sau đó kéo NameSet [Cau_6] vào Dimension.

Bước 3: “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Dimension	Hierarchy	Operator	Filter Expression
Dim Country	Production Countries	In	Cau_6
Dim Date	Year	Equal	{ 2013, 2014 }
<Select dimension>			
Year	Production Countries	Revenue	
2013	United States of America	887336558	
2014	United States of America	797838898	

Hình 3.104 Kết quả truy vấn câu 6 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```
SELECT
{([Measures].[Revenue], [Dim Date].[Year].&[2013]),
 ([Measures].[Revenue], [Dim Date].[Year].&[2014])} ON COLUMNS,
INTERSECT(
    TOPPERCENT(
        [Dim Country].[Production Countries].Children,
        10,
        ([Measures].[Revenue], [Dim Date].[Year].&[2013])
    ),
    TOPPERCENT(
        [Dim Country].[Production Countries].Children,
        10,
        ([Measures].[Revenue], [Dim Date].[Year].&[2014])
    )
) ON ROWS
FROM [SSIS];
```

Messages	Results
	Revenue
	2013
United States of America	887336558
	Revenue
	2014
United States of America	797838898

Hình 3.105 Kết quả truy vấn câu 6 ở MSSQ

Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo [revenue] trong bảng [Fact Movie] xuống ô **Values** và kéo [production_countries] và [Years (release_date)] qua ô **Rows**.

Drag fields between areas below:

Filters

Columns

Rows

Years (release_date)
production_countries

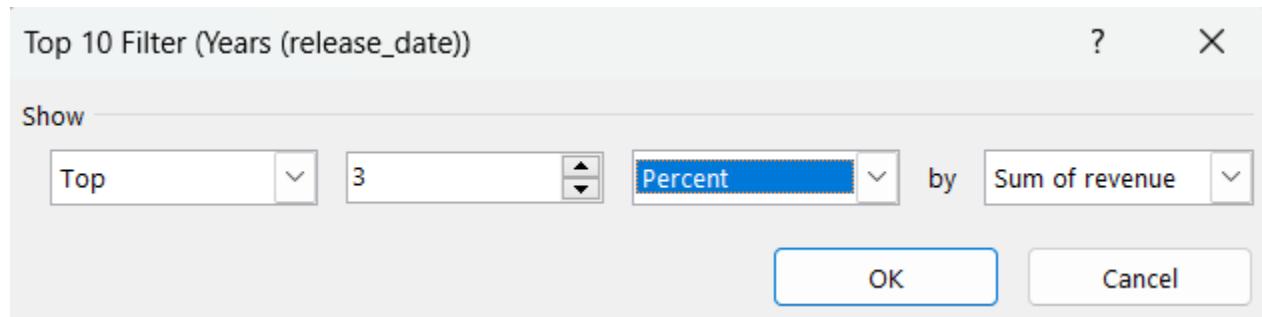
Values

Sum of revenue

Hình 3.106 Thiết lập PivotTable Fields của câu 6

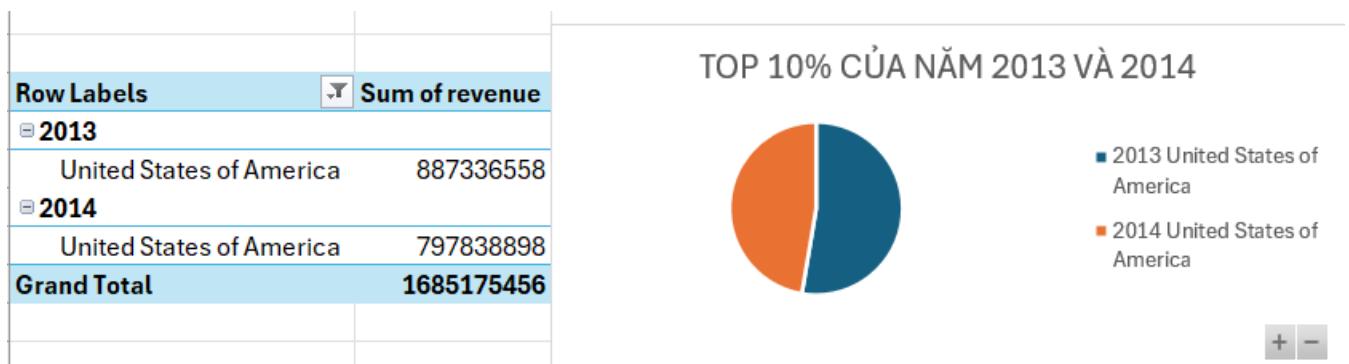
Bước 2: Trong **Row Labels** chọn năm 2013 và 2014. Click biểu tượng cạnh **Row Labels** chọn **Value Filters**, sau đó chọn **Top 10**.

Bước 3: Chính các thông số như hình dưới sau đó nhấn **OK**



Hình 3.107 Điều chỉnh Top 3 by Percent

Bước 4: Xem kết quả



Hình 3.108 Kết quả truy vấn câu 6 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab **Home**, nhấn vào **Get data** và chọn **Analysis Services**.

Bước 2: Nhập **Server** và **Database** đã tạo ở SSAS. Chọn **Import** và nhập truy vấn MDX.

SQL Server Analysis Services database

```

SELECT
    {[Measures].[Revenue], [Dim Date].[Year].&[2013]},
    {[Measures].[Revenue], [Dim Date].[Year].&[2014]} ON COLUMNS,
    INTERSECT(
        TOPPERCENT(
            [Dim Country].[Production Countries].Children,
            10,
            {[Measures].[Revenue], [Dim Date].[Year].&[2013]})
        ),
        TOPPERCENT(
            [Dim Country].[Production Countries].Children,
            10,
            {[Measures].[Revenue], [Dim Date].[Year].&[2014]}) )
    
```

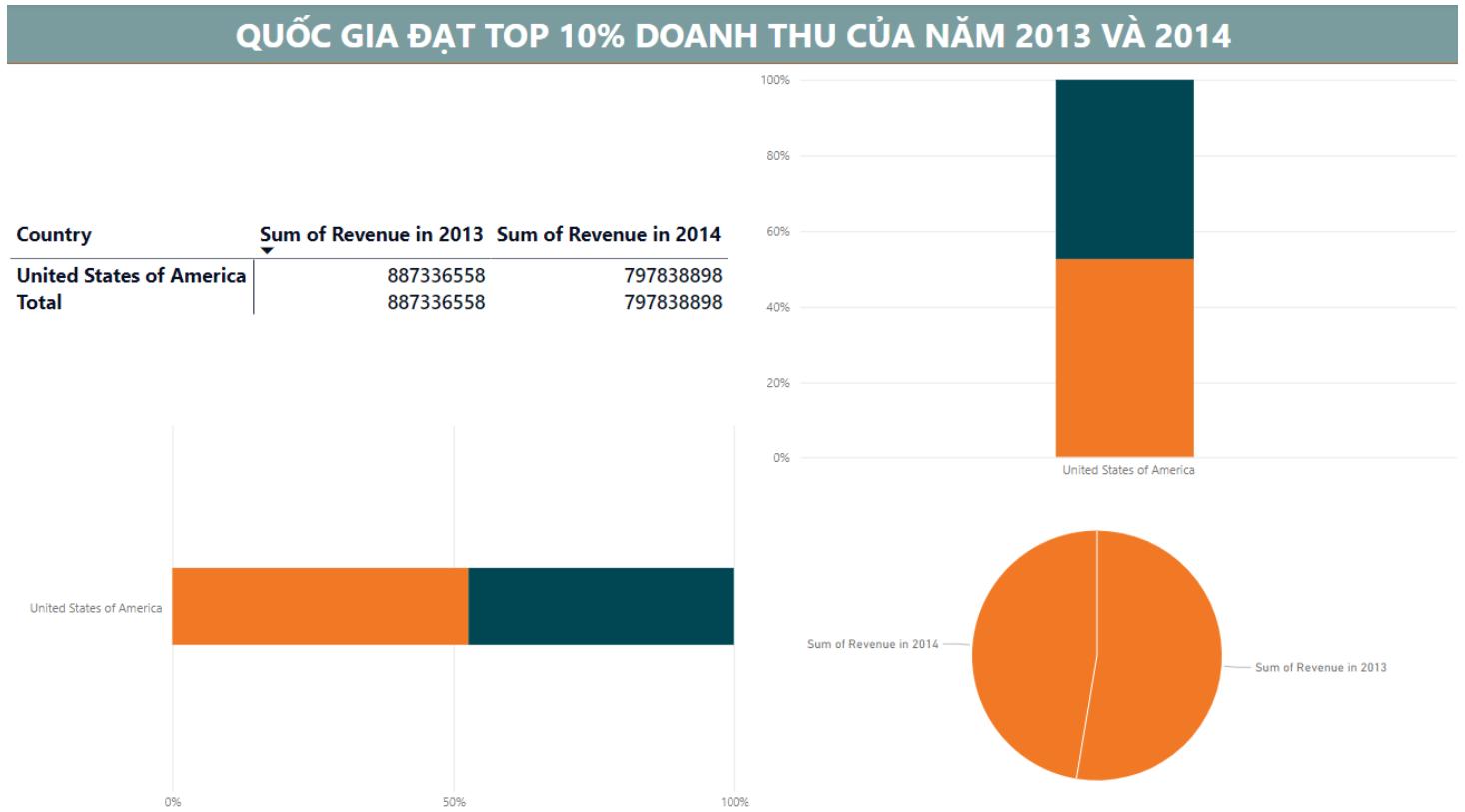
Hình 3.109 Thiết lập truy vấn MDX của câu 6 ở Power BI

Bước 3: Ở Tab **Table view** điều chỉnh tên cột và kiểu dữ liệu phù hợp.

Country	Revenue in 2013	Revenue in 2014		
United States of America	887336558	797838898		

Hình 3.110 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 6

Bước 4: Ở Tab **Report view** tạo **Visualizations** cho câu truy vấn.



Hình 3.111 Kết quả của truy vấn câu 6 ở Power BI

3.6.7 Câu truy vấn 7: Top 7 các quốc gia có tổng doanh thu cao nhất với điểm đánh giá trung bình trên 6.5 trong năm 2014

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Tạo NameSet [Cau_7] biểu thức như sau:

Name: [Cau_7]

Expression:

```
topcount(
    filter(
        [Dim Country].[Production Countries].[Production Countries].MEMBERS,
        [Measures].[Revenue] > 10000000 AND [Measures].[AverageRating] > 6.5
    ),
    7,
    [Measures].[Revenue]
)
```

No issues found

Ln: 8 Ch: 4 Col: 10 TABS CRLF

Hình 3.112 Nameset [Cau_7]

Tạo measures mới: Nhấn New Calculated Member và nhập biểu thức như sau:

Kho dữ liệu và OLAP - IS217.P12

Name: AverageRating

Parent Properties

Parent hierarchy: Measures Change

Parent member:

Expression

```
SUM([Dim Movie].[Title].[Title].Members,[Measures].[Vote Average] * [Measures].[Vote Count]) / SUM([Dim Movie].[Title].[Title].Members),  
No issues found
```

Ln: 2 Ch: 1 SPC CRLF

Hình 3.113 Tạo measures AverageRating

Bước 2: Kéo thả thuộc tính [Production Countries] trong bảng [Dim Country] và độ đo [Revenue], [AverageRating] vào cửa sổ thực thi. Sau đó kéo NameSet [Cau_7] vào **Dimension**. Ở **Dimension** chọn [Dim Date] và chọn [Year] là 2014.

Bước 3: “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Dimension	Hierarchy	Operator	Filter Expression
Dim Date	Year	Equal	{ 2014 }
Dim Country	Production Countries	In	Cau_7
<Select dimension>			
Production Countries		Revenue	AverageRating
Belgium, France		42830578	6.786038572474...
France		176404493	6.532222783803...
India		84194445	6.564308171523...
India, United States of America, United Arab Emirates		89514453	7.306000232696...
Japan		91865582	7.131280410349...
South Korea		136644258	7.029248088645...
United States of America		797838898	6.537162234767...

Hình 3.114 Kết quả truy vấn câu 7 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

Kho dữ liệu và OLAP - IS217.P12

```
WITH MEMBER [Measures].[Average Rating]
AS SUM(
    [Dim Movie].[Title].[Title].Members,
    [Measures].[Vote Average] * [Measures].[Vote Count]) / SUM(
    [Dim Movie].[Title].[Title].Members,
    [Measures].[Vote Count])
select {[Measures].[Revenue],[Measures].[Average Rating]} on columns,
non empty
topcount(
filter(
[Dim Country].[Production Countries].[Production
Countries].MEMBERS,
[Measures].[Revenue] > 10000000 AND
[Measures].[AverageRating] > 6.5),7, [Measures].[Revenue]
) ON ROWS
FROM [SSIS]
```

	Revenue	AverageRating
United States of America	797838898	6.53716223476795
France	176404493	6.53222278380359
South Korea	136644258	7.02924808864504
Japan	91865582	7.13128041034907
India, United States of America, United Arab Emirates	89514453	7.30600023269653
India	84194445	6.56430817152324
Belgium, France	42830578	6.78603857247404

Hình 3.115 Kết quả truy vấn câu 7 ở MSSQ

Thực hiện trong Excel

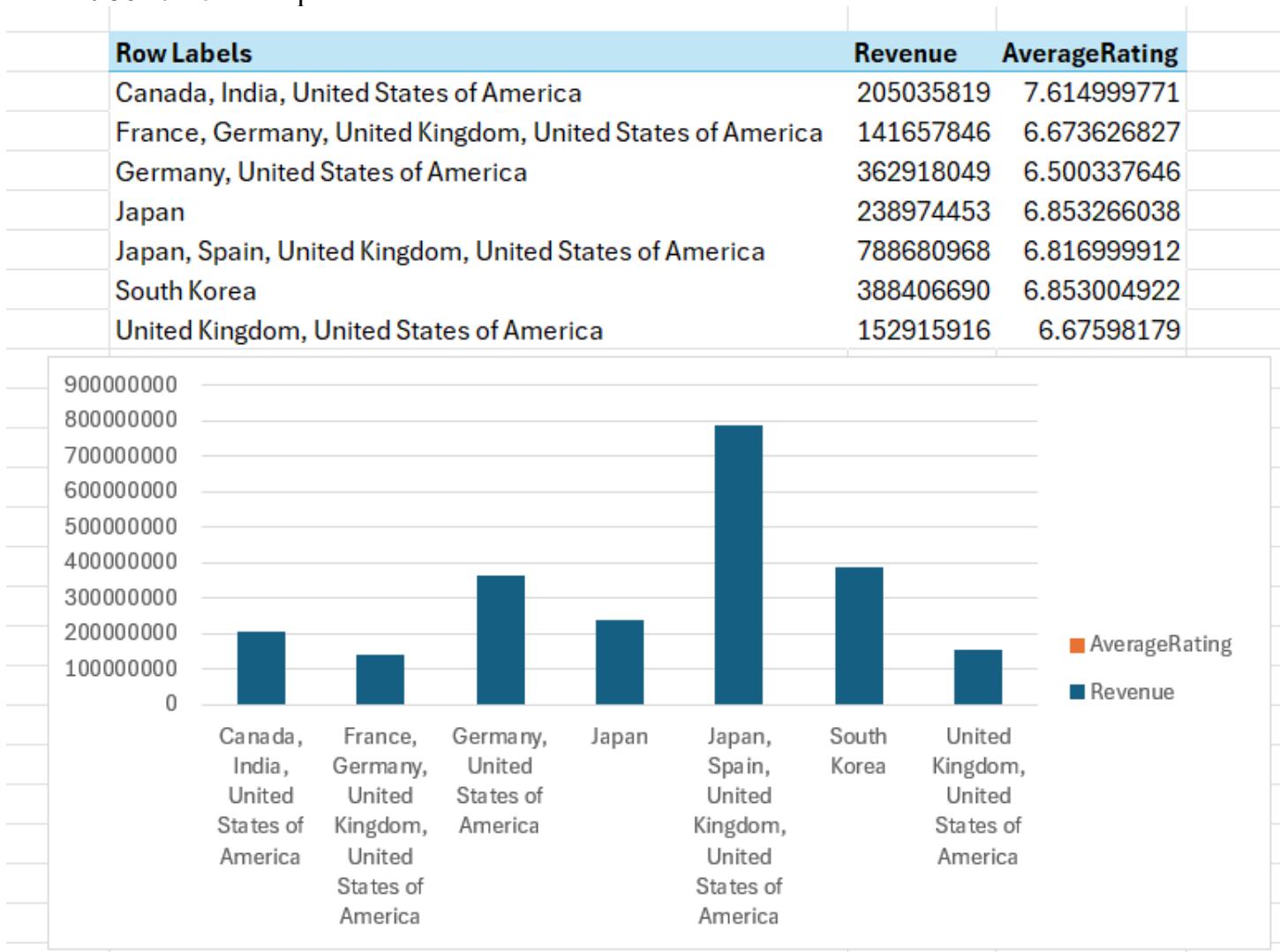
Bước 1: Trong PivotTable Fields, kéo [Revenue] và [AverageRating]trong bảng [Fact Movie] xuống ô **Values** và kéo [Cau_7] trong [Production Countries] vào ô **Rows**.

The screenshot shows the PivotTable Fields pane with the following configuration:

- Fields List:** Shows the hierarchy of fields:
 - Production Countries
 - Sets
 - Cau_6
 - Cau_7** (selected)
 - Dim Date
 - Date_Hierarchy
 - More Fields
- Drag fields between areas below:** A horizontal line with arrows indicating where fields can be moved.
- Filters:** An empty area.
- Columns:** An empty area.
- Rows:** Cau_7 is selected.
- Values:** Contains two fields:
 - \sum Values (Revenue)
 - AverageRating
 - Revenue

Hình 3.116 Thiết lập PivotTable Fields của câu 7

Bước 2: Xem kết quả



Hình 3.117 Kết quả truy vấn câu 7 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấn vào **Get data** và chọn **Analysis Services**.

Bước 2: Nhập **Server** và **Database** đã tạo ở SSAS. Chọn **Import** và nhập truy vấn MDX.

SQL Server Analysis Services database

Server ⓘ
HPHAT-TLVERSION\PHAT

Database
SSAS

Import
 DirectQuery

▲ MDX or DAX query (optional)

```
select {[Measures].[Revenue],[Measures].[AverageRating]} on columns,
       non empty
       topcount(
           filter(
               [Dim Country].[Production Countries].[Production Countries].M
               [Measures].[Revenue] > 10000000 AND [Measures].[AverageRating]
           ),
           7,
           [Measures].[Revenue]
       ) ON ROWS
FROM [SSIS]
WHERE ([Dim Date].[Year].&[2014]);
```

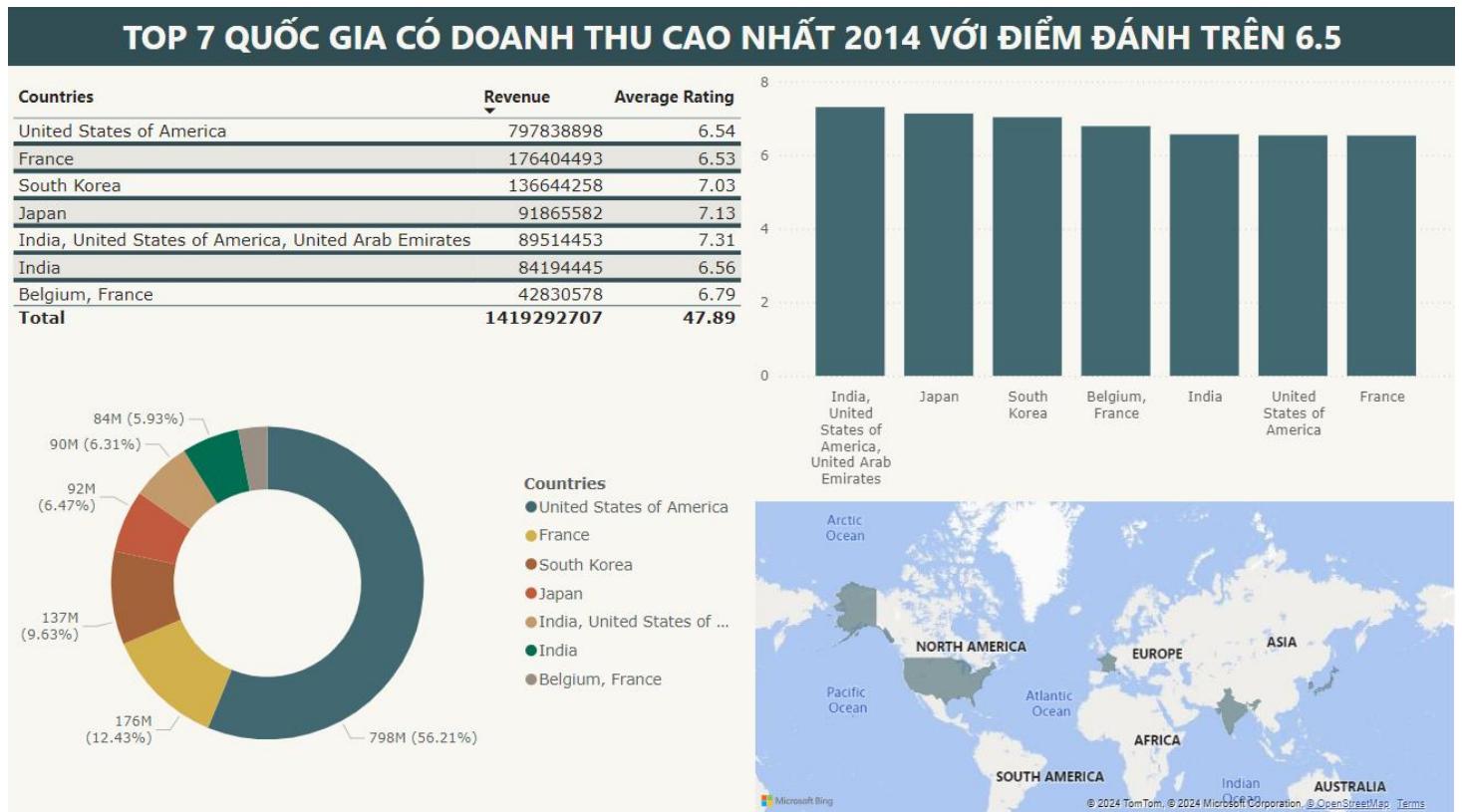
Hình 3.118 Thiết lập truy vấn MDX của câu 7 ở Power BI

Bước 3: Ở Tab Table view điều chỉnh tên cột và kiểu dữ liệu phù hợp.

Countries	Revenue	Average Rating
United States of America	797838898	6.54
France	176404493	6.53
South Korea	136644258	7.03
Japan	91865582	7.13
India, United States of America, United Arab Emirates	89514453	7.31
India	84194445	6.56
Belgium, France	42830578	6.79

Hình 3.119 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 7

Bước 4: Ở Tab Report view tạo Visualizations cho câu truy vấn.



Hình 3.120 Kết quả của truy vấn câu 7 ở Power BI

3.6.8 Câu truy vấn 8: Mỗi tháng trong năm 2018, liệt kê phim nổi tiếng nhất trong từng tháng

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Ở Browser, phần Dimension, chọn điều kiện truy vấn: ở [Dim Date] là năm 2018 và ở [Dim Movie] là [title] và chọn custom như ảnh bên dưới.

Dimension	Hierarchy	Operator	Filter Expression
Dim Date	Year	Equal	{ 2018 }
Dim Movie	Title	Custom	GENERATE([Dim Date].[Date_Hierarchy].[Month].Members,

Expression:

```
GENERATE( [Dim Date].[Date_Hierarchy].[Month].Members, TOPCOUNT( [Dim Date].[Month].currentmember*[Dim Movie].[Title].[Title].Members, 1, ([Measures].[Popularity], [Dim Date].[Month].CurrentMember) ) )
```

Hình 3.121 Thiết lập điều kiện cho truy vấn câu 8

Bước 2: Kéo thả thuộc tính [Title] trong bảng [Dim Movie], [Month] trong bảng [Dim Date] và độ đo [Popularity] vào cửa sổ thực thi.

Bước 3: “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Kho dữ liệu và OLAP - IS217.P12

Dimension	Hierarchy	Operator	Filter Expression
Dim Date	# Year	Equal	{ 2018 }
Dim Movie	# Title	Custom	GENERATE([Dim Date].[Date_Hierarchy].[Month].Members, ...)
<Select dimension>			

Month	Title	Popularity
1	Terrifier	43.382999420166
10	Blacked Raw V11	11.0550003051758
11	The Marine 6: Close Quarters	14.6319999694824
12	Backtrace	18.9209995269775
2	Accident Man	14.0979995727539
3	Charming	12.6999998092651
4	Overboard	20.121000289917
5	The Con Is On	13.4270000457764
6	Incredibles 2	49.2150001525879
7	Uncharted: Live Action Fan Film	6.88700008392334
8	Mia and the White Lion	18.4069995880127
9	The Bad Seed	15.7489995956421

Hình 3.122 Kết quả truy vấn câu 8 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```

SELECT {[Measures].[Popularity]} ON COLUMNS,
NON EMPTY GENERATE(
    [Dim Date].[Date_Hierarchy].[Month].Members,
    TOPCOUNT(
        [Dim Date].[Month].currentmember*[Dim Movie].[Title].[Title].Members,
        1,
        ([Measures].[Popularity], [Dim Date].[Month].CurrentMember)
    ) ON ROWS
FROM [SSIS]
WHERE ([Dim Date].[Year].&[2018]);

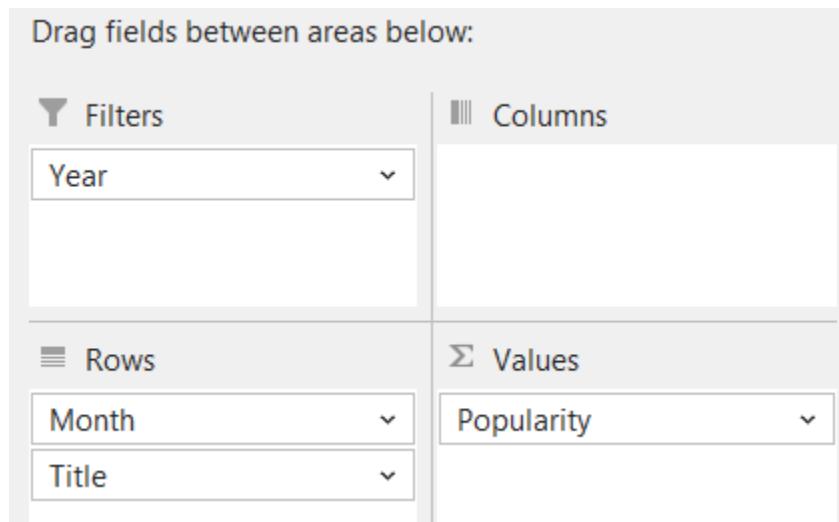
```

Messages		Results
		Popularity
1	Terrifier	43.382999420166
10	Blacked Raw V11	11.0550003051758
11	The Marine 6: Close Quarters	14.6319999694824
12	Backtrace	18.9209995269775
2	Accident Man	14.0979995727539
3	Charming	12.6999998092651
4	Overboard	20.121000289917
5	The Con Is On	13.4270000457764
6	Incredibles 2	49.2150001525879
7	Uncharted: Live Action Fan Film	6.88700008392334
8	Mia and the White Lion	18.4069995880127
9	The Bad Seed	15.7489995956421

Hình 3.123 Kết quả truy vấn câu 8 ở MSSQ

Thực hiện trong Excel

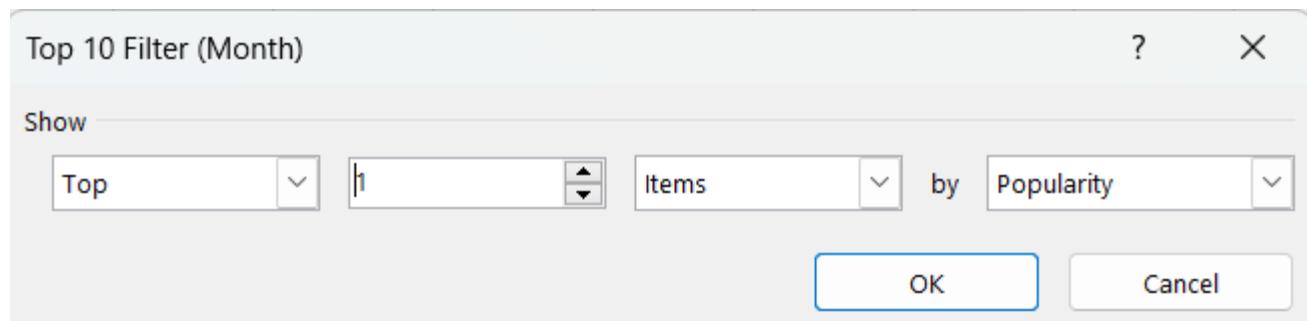
Bước 1: Trong PivotTable Fields, kéo [popularity] trong bảng [Fact Movie] xuống ô **Values** và kéo [Title] và [Month] qua ô **Rows**, kéo [Year] vào **Filters**.



Hình 3.124 Thiết lập PivotTable Fields của câu 8

Bước 2: Trong **Filters** chọn năm 2018. Click biểu tượng cạnh **Row Labels** chọn **Value Filters**, sau đó chọn **Top 1**.

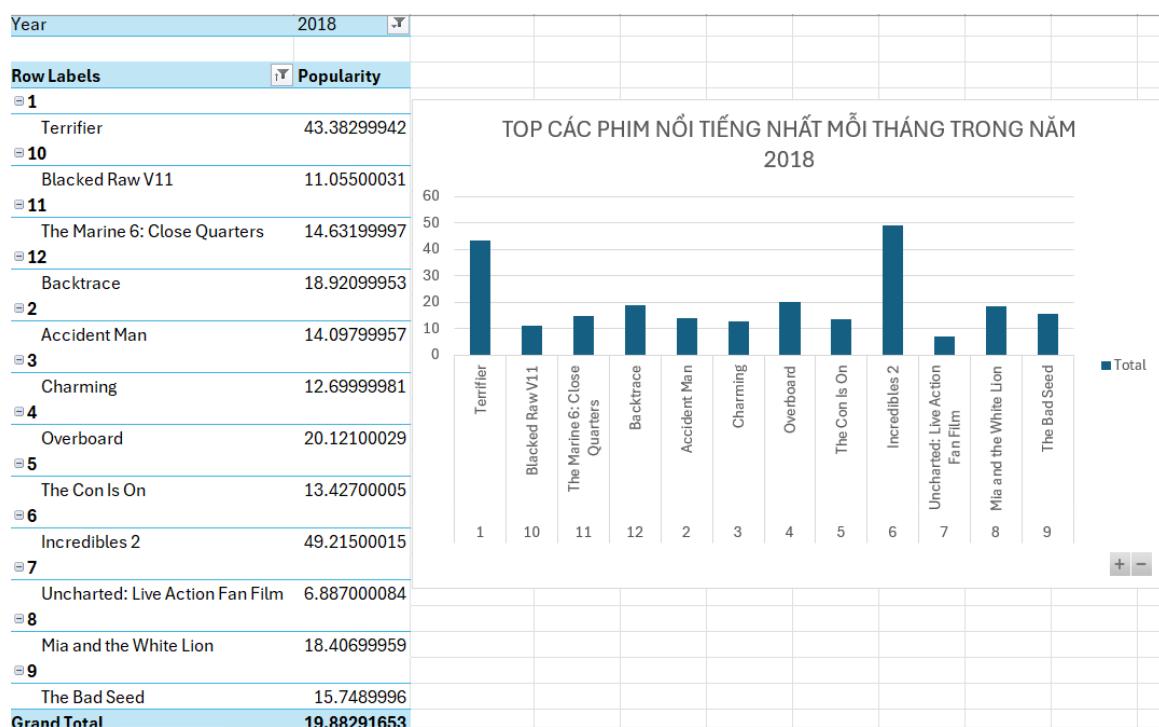
Bước 3: Điều chỉnh các thông số như hình dưới sau đó nhấn **OK**



Hình 3.125 Điều chỉnh Top 1 per Month

Bước 4: Xem kết quả

Kho dữ liệu và OLAP - IS217.P12



Hình 3.126 Kết quả truy vấn câu 8 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấn vào **Get data** và chọn **Analysis Services**.

Bước 2: Nhập Server và Database đã tạo ở SSAS. Chọn **Import** và nhập truy vấn MDX.

SQL Server Analysis Services database

Server

Database

Import

DirectQuery

MDX or DAX query (optional)

```

SELECT
    {[Measures].[Popularity]} ON COLUMNS,
    NON EMPTY GENERATE(
        [Dim Date].[Date_Hierarchy].[Month].Members,
        TOPCOUNT(
            [Dim Date].[Month].currentmember*[Dim Movie].[Title].[Title].Members,
            1,
            ([Measures].[Popularity], [Dim Date].[Month].CurrentMember)
        )
    ) ON ROWS
FROM [SSIS]
WHERE ([Dim Date].[Year].&[2018]);

```

Hình 3.127 Thiết lập truy vấn MDX của câu 8 ở Power BI

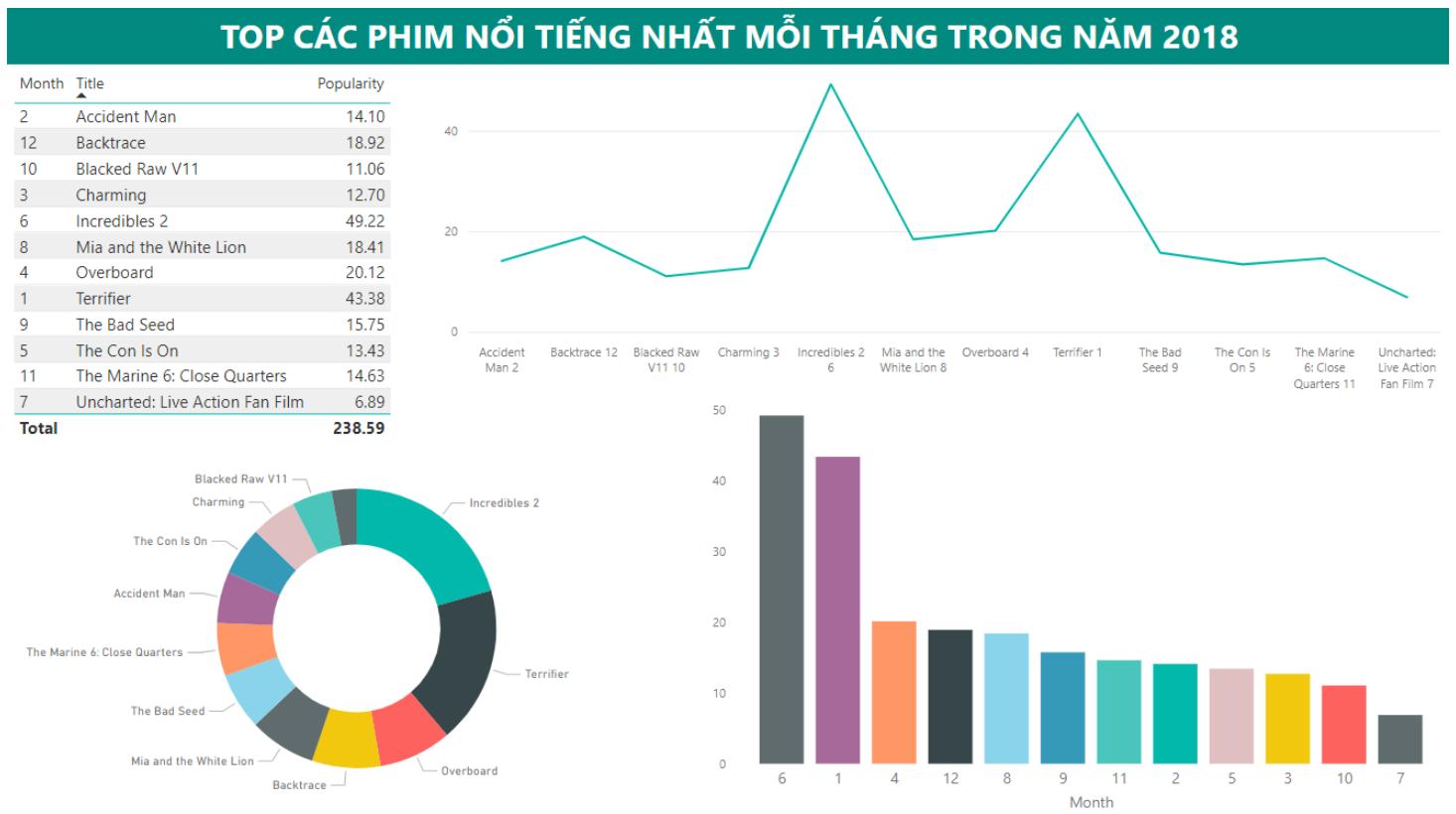
Kho dữ liệu và OLAP - IS217.P12

Bước 3: Ở Tab Table view điều chỉnh tên cột và kiểu dữ liệu phù hợp.

The screenshot shows the Power BI interface in 'Table tools' mode. On the left, a table of movie data is displayed with columns: Month, Title, and Popularity. The 'Popularity' column is selected. The top ribbon shows 'Table tools' is active. The 'Column tools' tab is open, showing settings for 'Name' (Popularity), 'Data type' (Decimal number), 'Format' (Decimal number with 2 decimal places), and 'Summarization' (Sum). To the right, there are buttons for 'Properties', 'Sort by column', 'Data groups', 'Manage relationships', and 'New column'. A 'Data' pane on the right lists various queries and their properties, such as 'Month' and 'Title'.

Hình 3.128 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 8

Bước 4: Ở Tab Report view tạo Visualizations cho câu truy vấn.



Hình 3.129 Kết quả của truy vấn câu 8 ở Power BI

3.6.9 Câu truy vấn 9: Độ nổi tiếng của Top 10 công ty sản xuất có số lượng phim ít nhất Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Tạo Nameset [Cau_9] như hình.

SVTH: Nguyễn Hồng Phát

Kho dữ liệu và OLAP - IS217.P12

Name: Cau_9

Expression

```
BOTTOMCOUNT(
    [Dim Company].[Production Companies].[Production Companies].Members,
    10,
    [Measures].[Number of Movies]
)
No issues found
```

Ln: 5 Ch: 7 SPC CRLF

Additional Properties

Type: Dynamic

Display folder:

Hình 3.130 Thiết lập điều kiện cho truy vấn câu 9

Bước 2: Kéo thả thuộc tính [Production Companies] trong bảng [Dim Company], độ đo [Number of Movies] và độ đo [Popularity] vào cửa sổ thực thi.

Bước 3: “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Dimension	Hierarchy	Operator	Filter Expression
Dim Company	Production Companies	In	Cau_9
<Select dimension>			
<hr/>			
Production Companies		Number of Movies	Popularity
Zuccherarte [IT]		1	0.600000023841858
Zuckerfilm		1	0.959999978542328
Zum Goldenen Lamm Filmproduktion		1	0.600000023841858
Zum Goldenen Lamm Filmproduktion, ARTE		1	0.600000023841858
Zvonko Produktion, Svenska Filminstitutet		1	0.600000023841858
Zwart Arbeid		1	1.31400001049042
Zycropolis Productions, Beall Productions		1	1.17499995231628
Z'Yeux Noirs Movies, K'Ien Productions		1	0.600000023841858
ZyPIX Productions		1	0.600000023841858

Hình 3.131 Kết quả truy vấn câu 9 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```
SELECT
{[Measures].[Number of Movies],[Measures].[Popularity]} ON COLUMNS,
NON EMPTY
BOTTOMCOUNT(
    [Dim Company].[Production Companies].[Production Companies].Members,
    10,
    [Measures].[Number of Movies]
) ON ROWS
FROM [SSIS];
```

	Number of Movies	Popularity
ZyPiX Productions	1	0.600000023841858
Z'Yeux Noirs Movies, K'len Productions	1	0.600000023841858
Zycopolis Productions, Beall Productions	1	1.17499995231628
Zwart Arbeid	1	1.31400001049042
Zvonko Produktion, Svenska Filminstitutet	1	0.600000023841858
Zum Goldenen Lamm Filmproduktion, ARTE	1	0.600000023841858
Zum Goldenen Lamm Filmproduktion	1	0.600000023841858
Zuckerfilm	1	0.959999978542328
Zuccherrarte [IT]	1	0.600000023841858

Hình 3.132 Kết quả truy vấn câu 9 ở MSSQ

Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo [popularity] trong bảng [Fact Movie] xuống ô **Values** và kéo [Title] và [Month] qua ô **Rows**, kéo [Year] vào **Filters**.

The screenshot shows the Microsoft Power BI PivotTable Fields pane. In the top-left area, under 'Sets', 'Cau_9' is selected. In the bottom-right area, under 'Values', 'Number of Movies' and 'Popularity' are listed. The 'Columns' section shows 'Σ Values'. The 'Rows' section shows 'Cau_9'.

Hình 3.125 Thiết lập PivotTable Fields của câu 9

Bước 4: Xem kết quả

Row Labels	Number of Movies	Popularity
Zuccherarte [IT]	1	0.600000024
Zuckerfilm	1	0.959999979
Zum Goldenen Lamm Filmproduktion	1	0.600000024
Zum Goldenen Lamm Filmproduktion, ARTE	1	0.600000024
Zvonko Produktion, Svenska Filminstitutet	1	0.600000024
Zwart Arbeid	1	1.31400001
Zycopolis Productions, Beall Productions	1	1.174999952
Z'Yeux Noirs Movies, K'len Productions	1	0.600000024
ZyPiX Productions	1	0.600000024

Hình 3.127 Kết quả truy vấn câu 9 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấn vào **Get data** và chọn **Analysis Services**.

Bước 2: Nhập **Server** và **Database** đã tạo ở SSAS. Chọn **Import** và nhập truy vấn MDX.

SQL Server Analysis Services database

The screenshot shows the 'Get data' configuration window for an Analysis Services database. It includes fields for 'Server' (set to 'PHAT-TLVERSION\PHAT'), 'Database' (set to 'SSAS'), and a radio button for 'Import' (which is selected). Below these, there is a section for 'MDX or DAX query (optional)' containing the following MDX query:

```

SELECT
    {[Measures].[Number of Movies],[Measures].[Popularity]} ON COLUMNS,
    NON EMPTY
    BOTTOMCOUNT(
        [Dim Company].[Production Companies].[Production Companies].Members,
        10,
        [Measures].[Number of Movies]
    ) ON ROWS
FROM [SSIS];

```

Hình 3.128 Thiết lập truy vấn MDX của câu 9 ở Power BI

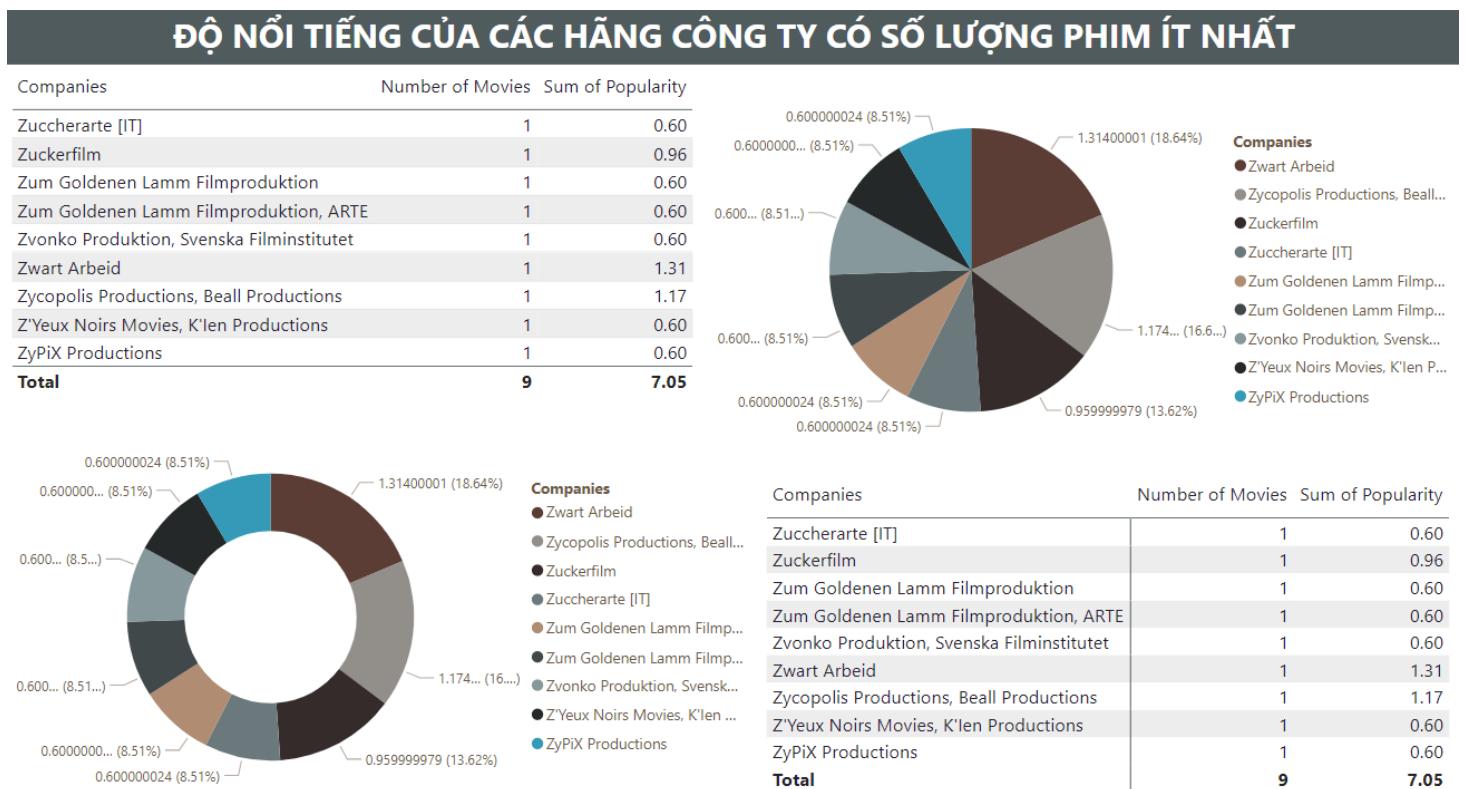
Bước 3: Ở Tab **Table view** điều chỉnh tên cột và kiểu dữ liệu phù hợp.

Kho dữ liệu và OLAP - IS217.P12

The screenshot shows the Power BI interface with the 'Column tools' tab selected. In the 'Structure' section, there is a table titled 'Companies' with columns 'Number of Movies' and 'Popularity'. The 'Number of Movies' column has its name changed to 'Number of Movies' and its data type set to 'Whole number'. The 'Popularity' column has its name changed to 'Popularity' and its data category set to 'Uncategorized'. The 'Formatting' section shows the current format as '\$ 0,00'. The 'Properties' section includes options for summarization ('Sum'), sorting ('Sort by column'), grouping ('Data groups'), managing relationships ('Manage relationships'), and creating new columns ('New column'). The 'Calculations' section is also visible. On the right side, the 'Data' pane shows a tree view of queries: 'Query1' (Total Revenue, Year), 'Query2', 'Query3', 'Query4', 'Query5', 'Query6', 'Query7', 'Query8', 'Query9' (Companies), and 'Number of Movies' and 'Popularity' under 'Companies'.

Hình 3.129 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 9

Bước 4: Ở Tab Report view tạo Visualizations cho câu truy vấn.



Hình 3.130 Kết quả của truy vấn câu 9 ở Power BI

3.6.10 Câu truy vấn 10: Thông kê số lượng phim, số lượt bình chọn và điểm đánh giá trung bình được sản xuất bởi các nước công ty Jules Jordan Video, Naughty America, Marvel Studios, Pixar và Digital Playground

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Tạo Nameset [Cau_10] như hình.

SVTH: Nguyễn Hồng Phát

Kho dữ liệu và OLAP - IS217.P12

Name:

[Cau_10]

❖ Expression

```
TOPCOUNT({  
    [Dim Company].[Production Companies].[Jules Jordan Video],  
    [Dim Company].[Production Companies].[Naughty America],  
    [Dim Company].[Production Companies].[Marvel Studios],  
    [Dim Company].[Production Companies].[Pixar]}
```

✓ No issues found

Ln: 6 Ch: 42 SPC CRLF

Hình 3.131 Thiết lập điều kiện cho truy vấn câu 10

Bước 2: Kéo thả các mục ở **Dimension** và các thông tin như hình và nhấn “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Dimension	Hierarchy	Operator	Filter Expression
Dim Company	Production Companies	In	Cau_10
<Select dimension>			
Production Companies	Number of Movies	Vote Count	AverageRating
Jules Jordan Video	36	31	7.80322582490983
Marvel Studios	5	1249	7.02546203833184
Naughty America	159	54	6.7851852134422
Pixar	10	1237	6.69142595258757
Popularity			

Hình 3.132 Kết quả truy vấn câu 10 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```
SELECT  
{ [Measures].[Number of Movies],  
[Measures].[Vote Count],  
[Measures].[AverageRating],  
[Measures].[Popularity]} ON COLUMNS,  
TOPCOUNT({  
    [Dim Company].[Production Companies].[Jules Jordan Video],  
    [Dim Company].[Production Companies].[Naughty America],  
    [Dim Company].[Production Companies].[Marvel Studios],  
    [Dim Company].[Production Companies].[Pixar],  
    [Dim Company].[Production Companies].[Digital Playground]  
}, 5, [Measures].[Number of Movies]) ON ROWS  
FROM [SSIS];
```

	Number of Movies	Vote Count	AverageRating	Popularity
Naughty America	159	54	6.7851852134422	0.788954563342132
Jules Jordan Video	36	31	7.80322582490983	0.682027791937192
Digital Playground	33	65	7.10001528813289	0.981757584846381
Pixar	10	1237	6.69142595258757	9.54644441604614
Marvel Studios	5	1249	7.02546203833184	27.7239997982979

Hình 3.133 Kết quả truy vấn câu 10 ở MSSQ

Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo [popularity] trong bảng [Fact Movie] xuống ô **Values** và kéo [Title] và [Month] qua ô **Rows**, kéo [Year] vào **Filters**.

The screenshot shows the Microsoft Power BI Fields pane. At the top, there is a search bar and a 'Sets' section containing a selected set named 'Cau_10'. Below this, the 'Filters' section contains a dropdown for 'Cau_10'. The 'Columns' section has a 'Σ Values' button. The 'Rows' section also contains a dropdown for 'Cau_10'. The 'Values' section lists three measures: 'Number of Movies', 'Vote Count', and 'AverageRating'. The overall interface is used to build a PivotTable for the specified question.

Hình 3.134 Thiết lập PivotTable Fields của câu 10

Bước 4: Xem kết quả

Row Labels	Number of Movies	Vote Count	AverageRating	Popularity
Jules Jordan Video	36	31	7.803225825	0.682027792
Marvel Studios	5	1249	7.025462038	27.7239998
Naughty America	159	54	6.785185213	0.788954563
Pixar	10	1237	6.691425953	9.546444416

Hình 3.135 Kết quả truy vấn câu 10 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab **Home**, nhấn vào **Get data** và chọn **Analysis Services**.

Bước 2: Nhập **Server** và **Database** đã tạo ở SSAS. Chọn **Import** và nhập truy vấn MDX.

SQL Server Analysis Services database

The screenshot shows the 'Get data' dialog for Power BI. It has the following fields filled in:

- Server:** HPHAT-TLVERSION\PHAT
- Database:** SSAS
- Import/DirectQuery:** Import (selected)

Below these fields, there is a section titled "MDX or DAX query (optional)" containing the following MDX query:

```

SELECT
    { [Measures].[Number of Movies],
      [Measures].[Vote Count],
      [Measures].[Average Rating],
      [Measures].[Popularity] } ON COLUMNS,
    TOPCOUNT(
        [Dim Company].[Production Companies].[Jules Jordan Video],
        [Dim Company].[Production Companies].[Naughty America],
        [Dim Company].[Production Companies].[Marvel Studios],
        [Dim Company].[Production Companies].[Pixar],
        [Dim Company].[Production Companies].[Digital Playground]
    ), 5, [Measures].[Number of Movies]) ON ROWS
FROM [SSIS];
  
```

Hình 3.136 Thiết lập truy vấn MDX của câu 10 ở Power BI

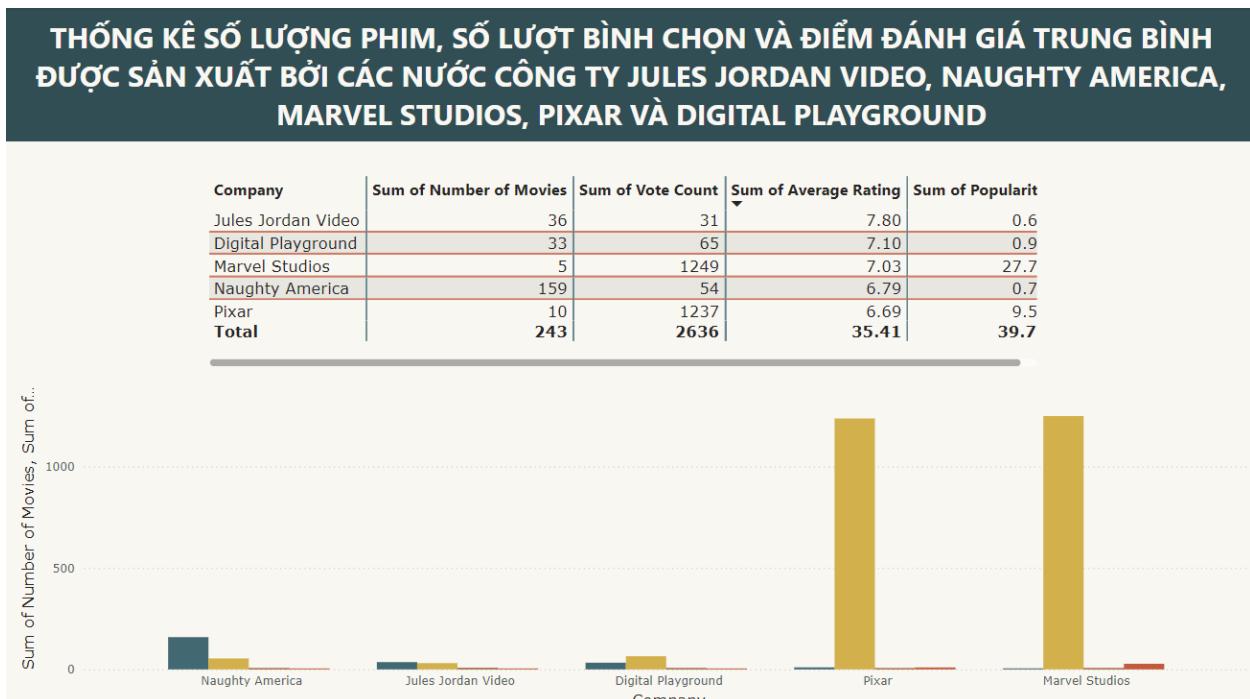
Bước 3: Ở Tab **Table view** điều chỉnh tên cột và kiểu dữ liệu phù hợp.

Kho dữ liệu và OLAP - IS217.P12

The screenshot shows the Microsoft Power BI 'Column tools' ribbon tab selected. A table is displayed with columns: Company, Number of Movies, Vote Count, Average Rating, and Popularity. The 'Popularity' column is currently selected for editing, with its properties shown in the ribbon: Name (Popularity), Data type (Decimal number), Format (\$, #,##,###,##0.00), Summarization (Sum), Data category (Uncategorized), Sort by column (Sort), Data groups (Groups), Manage relationships (Relationships), New column (New column), and Calculations (Calculations).

Hình 3.137 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 10

Bước 4: Ở Tab Report view tạo Visualizations cho câu truy vấn.



Hình 3.138 Kết quả của truy vấn câu 10 ở Power BI

3.6.11 Câu truy vấn 11: Thông kê số lượng phim, tổng số lượt đánh giá, và độ phổ biến theo thể loại phim và các ngôn ngữ khác nhau (Việt, Hàn, Trung, Nhật)

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Kéo thả các mục ở Dimension và các thông tin như hình và nhấn “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Kho dữ liệu và OLAP - IS217.P12

Dimension	Hierarchy	Operator	Filter Expression	
Dim Language	Original Language	Equal	{ ja, cn, vi, ko }	
<Select dimension>				
Genres List	Original Language	Number of Movies	Vote Count	Popularity
['Action', 'Adventure', 'Comedy', 'Crime', 'Animation']	ja	1	7	4.41400003433228
['Action', 'Adventure', 'Comedy', 'Drama', 'Animation']	ja	1	126	21.9979991912842
['Action', 'Adventure', 'Drama']	ja	1	1	2.09200000762939
['Action', 'Adventure', 'Fantasy']	cn	1	72	11.6280002593994
['Action', 'Adventure', 'Fantasy']	ja	2	171	25.8910006284714
['Action', 'Adventure', 'Science Fiction', 'Fantasy']	ja	2	12	4.36999988555908
['Action', 'Adventure', 'Science Fiction']	ja	1	5	1.49199998378754
['Action', 'Animation', 'Comedy', 'TV Movie', 'Crime']	ja	1	27	7.18800020217896
['Action', 'Animation', 'Comedy']	ja	2	34	3.755499958992
['Action', 'Animation', 'Drama', 'History']	ja	1	8	2.48300004005432
['Action', 'Animation', 'Science Fiction', 'Fantasy']	ja	4	58	6.24425005912781
['Action', 'Animation', 'Science Fiction', 'War']	ja	1	14	4.50099992752075
['Action', 'Animation', 'Science Fiction']	ja	1	21	5.43499994277954
['Action', 'Animation']	ja	2	8	3.19199991226196

Hình 3.139 Kết quả truy vấn câu 11 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```

SELECT
{[Measures].[Number of Movies],
 [Measures].[Vote Count],
 [Measures].[Popularity]} ON COLUMNS,
NON EMPTY
CROSSJOIN(
    [Dim Genres List].[Genres List].[Genres List].MEMBERS,
    {[Dim Language].[Original Language].&[vi],
     [Dim Language].[Original Language].&[ko],
     [Dim Language].[Original Language].&[cn],
     [Dim Language].[Original Language].&[ja]}) ON ROWS
FROM [SSIS];

```

			Number of Movies	Vote Count	Popularity
['Action', 'Adventure', 'Comedy', 'Crime', 'Animation']	ja	1	7	4.41400003433228	
['Action', 'Adventure', 'Comedy', 'Drama', 'Animation']	ja	1	126	21.9979991912842	
['Action', 'Adventure', 'Drama']	ja	1	1	2.09200000762939	
['Action', 'Adventure', 'Fantasy']	cn	1	72	11.6280002593994	
['Action', 'Adventure', 'Fantasy']	ja	2	171	25.8910006284714	
['Action', 'Adventure', 'Science Fiction', 'Fantasy']	ja	2	12	4.36999988555908	
['Action', 'Adventure', 'Science Fiction']	ja	1	5	1.49199998378754	
['Action', 'Animation', 'Comedy', 'TV Movie', 'Crime']	ja	1	27	7.18800020217896	
['Action', 'Animation', 'Comedy']	ja	2	34	3.755499958992	
['Action', 'Animation', 'Drama', 'History']	ja	1	8	2.48300004005432	

Hình 3.140 Kết quả truy vấn câu 11 ở MSSQ

Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo các độ đo và cột như hình.

The screenshot shows the 'PivotTable Fields' ribbon in Excel. At the top, there is a tree view of fields from two dimensions: 'Dim Director' and 'Dim Genres List'. Under 'Dim Director', 'Director' and 'Director Id' are listed. Under 'Dim Genres List', 'Genres List' (with a checked checkbox) and 'Genres List Id' are listed. Below the tree view, a message says 'Drag fields between areas below:'. The ribbon is divided into four main sections: 'Filters' (empty), 'Columns' (labeled 'Σ Values'), 'Rows' (containing 'Original Language' and 'Genres List' dropdowns), and 'Values' (containing 'Vote Count', 'Popularity', and 'Number of Movies' dropdowns). The 'Values' section has a small downward arrow icon indicating it can be expanded.

Hình 3.141 Thiết lập PivotTable Fields của câu 11

Bước 4: Xem kết quả

Kho dữ liệu và OLAP - IS217.P12

Row Labels		Vote Count	Popularity	Number of Movies
cn				
['Action', 'Adventure', 'Fantasy']	72	11.6280003		1
['Action', 'Comedy']	35	5.55999994		1
['Action', 'Crime', 'Drama', 'Mystery']	2	4.77199984		1
['Action', 'Crime', 'Drama', 'Thriller']	0	1.26199996		1
['Action', 'Crime']	102	4.65966654		3
['Action', 'Drama', 'Crime', 'Adventure']	156	8.95899963		1
['Action', 'Drama', 'Crime']	241	13.6190004		1
['Action', 'Drama', 'Thriller', 'Crime']	133	11.2440004		1
['Action', 'History', 'Drama']	111	14.7659998		1
['Action']	20	3.64433324		3
['Comedy', 'Drama', 'Family']	6	5.28599977		1
['Comedy', 'Drama']	23	5.57200003		1
['Comedy', 'Horror']	0	1.39999998		1
['Comedy', 'Romance']	20	4.06650007		2
['Comedy']	4	1.7385		2
['Crime', 'Thriller']	10	1.62899995		1
['Drama', 'Action', 'Thriller']	34	10.0039997		1
['Drama', 'Comedy']	18	2.22850001		2
['Drama', 'Crime', 'Thriller', 'Action']	73	5.81899977		1
['Drama', 'Horror', 'Mystery']	8	1.24899995		1
['Drama', 'Music', 'Romance']	45	7.00500011		1
['Drama', 'Romance']	45	4.5630001		3
['Drama']	216	3.64599999		8
['History']	12	4.35300016		1
['Horror']	28	2.13899997		4
['Music']	0	1.07299995		1
['Romance', 'Action']	12	5.49300003		1
['Romance', 'Comedy']	1	1.016		2
['Romance', 'Drama']	12	3.79300004		2
['Romance']	2	1.40650004		2
['Science Fiction', 'Action']	16	2.40199995		1
['Thriller', 'Comedy', 'Horror']	68	7.875		1
['Unknown']	2	0.68200001		2
ja				
['Action', 'Adventure', 'Comedy', 'Crime', 'Animation']	7	4.41400003		1
['Action', 'Adventure', 'Comedy', 'Drama', 'Animation']	126	21.9979992		1
['Action', 'Adventure', 'Drama']	1	2.09200001		1
['Action', 'Adventure', 'Fantasy']	171	25.8910006		2
['Action', 'Adventure', 'Science Fiction', 'Fantasy']	12	4.36999989		2
['Action', 'Adventure', 'Science Fiction']	5	1.49199998		1

Hình 3.143 Kết quả truy vấn câu 11 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấp vào Get data và chọn Analysis Services.

Bước 2: Nhập Server và Database đã tạo ở SSAS. Chọn Import và nhập truy vấn MDX.

SQL Server Analysis Services database

```

SELECT
    {[Measures].[Number of Movies],
     [Measures].[Vote Count],
     [Measures].[Popularity]} ON COLUMNS,
    NON EMPTY
    CROSSJOIN(
        [Dim Genres List].[Genres List].[Genres List].MEMBERS,
        {[Dim Language].[Original Language].&[vi],
         [Dim Language].[Original Language].&[ko],
         [Dim Language].[Original Language].&[cn],
         [Dim Language].[Original Language].&[ja]}
    ) ON ROWS
FROM [SSIS];

```

Hình 3.144 Thiết lập truy vấn MDX của câu 11 ở Power BI

Bước 3: Vào Add Columns chọn Conditional Format để thêm 1 cột thể hiện ngôn ngữ gốc thay vì kí tự như cột ban đầu.

Hình 3.145 Tạo một cột mới để thể hiện rõ ngôn ngữ gốc

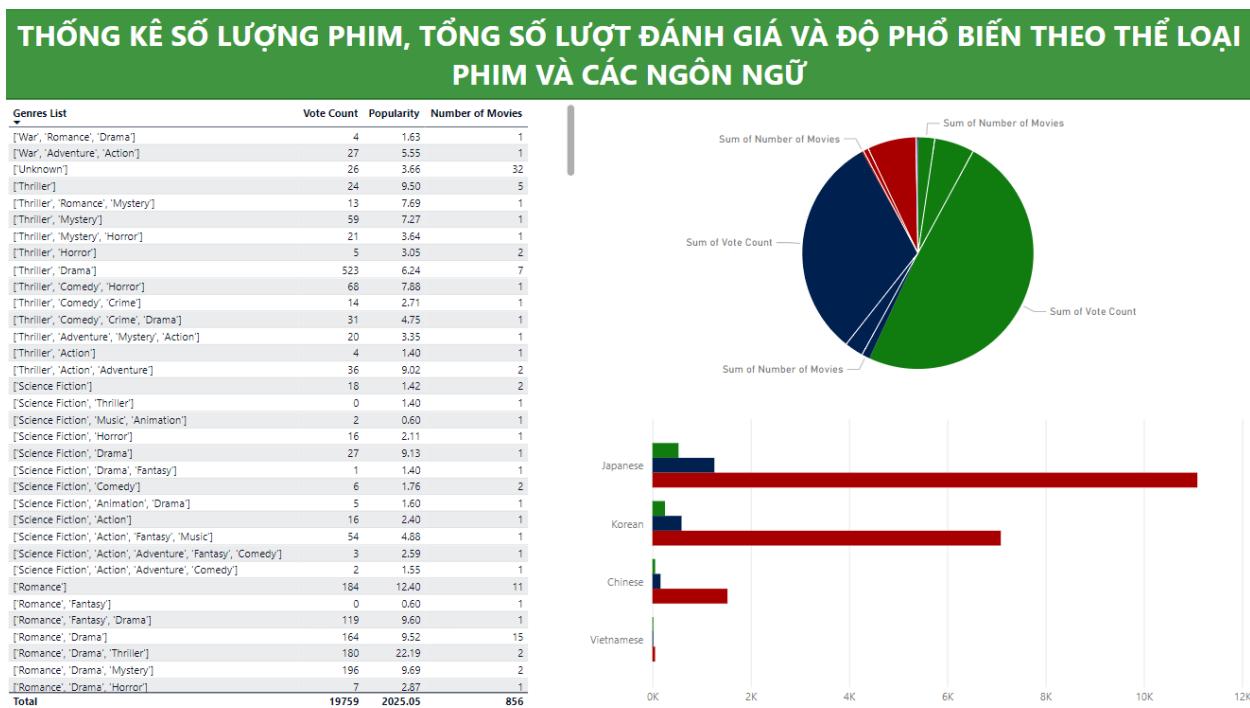
Bước 4: Ở Tab Table view điều chỉnh tên cột và kiểu dữ liệu phù hợp.

Kho dữ liệu và OLAP - IS217.P12

The screenshot shows the Power BI ribbon with the 'Column tools' tab selected. Under 'Structure', there's a 'Genres List' dropdown and a table view showing movie genres like 'Adventure', 'Comedy', 'Crime', etc., with their corresponding Japanese codes (ja). Under 'Formatting', there are options for 'Name' (Number of Movies), 'Format' (\$ %), 'Summarization' (Sum), 'Data category' (Uncategorized), 'Sort by column' (Sort), 'Data groups' (Groups), 'Manage relationships' (Relationships), and 'New column' (Calculations). The main area displays a table of movie data with columns: Number of Movies, Vote Count, Popularity, and Custom. The 'Custom' column contains Japanese values such as 4.414000034, 21.99799919, etc. A sidebar on the right lists various queries and dimensions.

Hình 3.146 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 11

Bước 5: Ở Tab Report view tạo Visualizations cho câu truy vấn.



Hình 3.147 Kết quả của truy vấn câu 11 ở Power BI

3.6.12 Câu truy vấn 12: Tổng kinh phí đầu tư hàng tháng và cả năm của các phim được sản xuất tại Mỹ với thời gian công chiếu là từ năm 2017 đến năm 2019

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Tạo Nameset [Cau_12] như ảnh.

SVTH: Nguyễn Hồng Phát

Kho dữ liệu và OLAP - IS217.P12

Name:

[Cau_12]

Expression

```
DrillDownLevel(  
    [Dim Date].[Date_Hierarchy].[Year].&[2017] : [Dim Date].[Date_Hierarchy].[Year].&[2019])
```

✓ No issues found

Ln: 2 Ch: 95 SPC CRLF

Hình 3.148 Tạo Nameset [Cau_12]

Bước 2: Kéo thả các mục ở **Dimension** và các thông tin như hình và nhấn “**Click to execute the query**” để thực thi câu lệnh và được kết quả như sau.

Dimension	Hierarchy	Operator	Filter Expression
Dim Date	▪ Date_Hierarchy	In	Cau_12
Dim Country	▪ Production Countries	Equal	{ United States of America }
<Select dimension>			

Year	Month	Budget
2017	1	10200000
2017	10	15189600
2017	11	13002075
2017	12	936030
2017	2	22431080
2017	3	2040200
2017	4	25050000
2017	5	3000000
2017	6	11017095
2017	7	60000000
2017	8	3010030
2017	9	95107000
2018	1	36180350
2018	10	3766500

Hình 3.149 Kết quả truy vấn câu 12 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```
SELECT  
    [Measures].[Budget] ON COLUMNS,  
    DrillDownLevel(  
        [Dim Date].[Date_Hierarchy].[Year].&[2017] : [Dim  
Date].[Date_Hierarchy].[Year].&[2019]  
    ) ON ROWS  
FROM [SSIS]  
WHERE ([Dim Country].[Production Countries].&[United States of America]);
```

Budget	
2017	260983110
1	10200000
10	15189600
11	13002075
12	936030
2	22431080
3	2040200
4	25050000
5	3000000
6	11017095
7	60000000
8	3010030
9	95107000
2018	322643029

Hình 3.150 Kết quả truy vấn câu 12 ở MSSQ

Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo các độ đo và cột như hình.

The screenshot shows the PivotTable Fields pane in Excel. On the left, under 'Dim Country', 'Production Countries' is selected. In the main area, 'Production Countries' is in the 'Filters' section, 'Year' and 'Month' are in the 'Rows' section, and 'Budget' is in the 'Values' section. The 'Columns' section is empty.

Hình 3.151 Thiết lập PivotTable Fields của câu 12

Bước 4: Xem kết quả

2017	
1	10200000
10	15189600
11	13002075
12	936030
2	22431080
3	2040200
4	25050000
5	3000000
6	11017095
7	60000000
8	3010030
9	95107000

2018	
1	36180350
10	3766500
11	5210000
12	13679879
2	12000
3	20500000
4	14820300
5	28002000
6	200060000
7	12200
8	395000
9	4800

2019	

Hình 3.152 Kết quả truy vấn câu 12 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấn vào **Get data** và chọn **Analysis Services**.

Bước 2: Nhập **Server** và **Database** đã tạo ở SSAS. Chọn **Import** và nhập truy vấn MDX.

SQL Server Analysis Services database

Server ⓘ
HPHAT-TLVERSION\PHAT

Database
SSAS

Import
 DirectQuery

▲ MDX or DAX query (optional)

```
SELECT
    [Measures].[Budget] ON COLUMNS,
    DrillDownLevel(
        [Dim Date].[Date_Hierarchy].[Year].&[2017] : [Dim Date].[Date_Hierarchy].[Year].&[2019]
    ) ON ROWS
FROM [SSIS]
WHERE ([Dim Country].[Production Countries].&[United States of America]);
```

Hình 3.153 Thiết lập truy vấn MDX của câu 12 ở Power BI

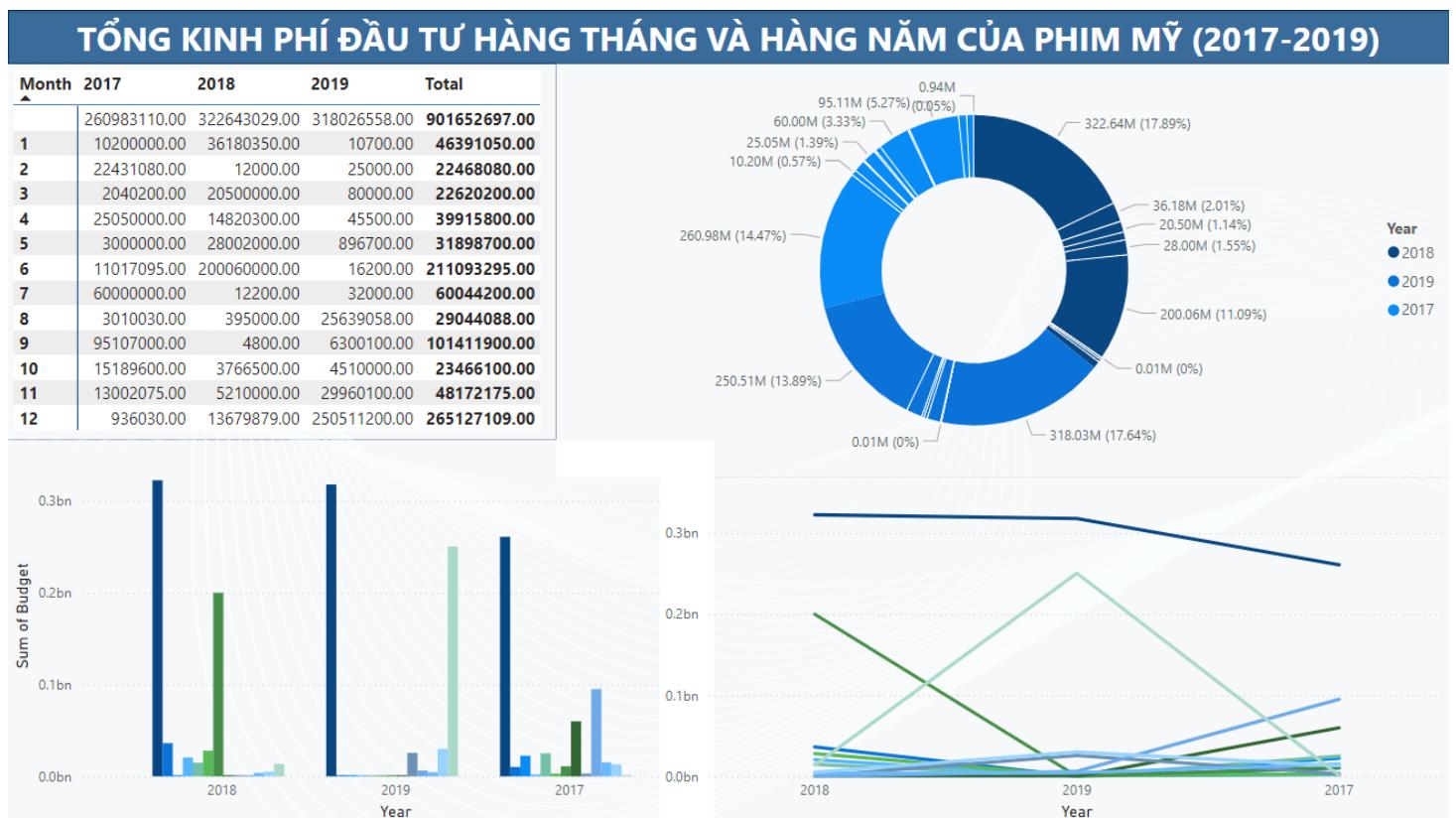
Bước 4: Ở Tab Table view điều chỉnh tên cột và kiểu dữ liệu phù hợp.

Year	Month	Budget
2017		260983110.00
2017	1	10200000.00
2017	10	15189600.00
2017	11	13002075.00
2017	12	936030.00
2017	2	22431080.00
2017	3	2040200.00
2017	4	25050000.00
2017	5	3000000.00
2017	6	11017095.00
2017	7	6000000.00
2017	8	3010030.00
2017	9	95107000.00
2018		322643029.00
2018	1	36180350.00
2018	10	3766500.00
2018	11	5210000.00
2018	12	13679879.00
2018	2	12000.00
2018	3	20500000.00
2018	4	14820300.00
2018	5	28002000.00
2018	6	200060000.00
2018	7	12200.00
2018	8	395000.00
2018	9	18000.00

Hình 3.154 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 12

Bước 5: Ở Tab Report view tạo Visualizations cho câu truy vấn.

SVTH: Nguyễn Hồng Phát



Hình 3.155 Kết quả của truy vấn câu 12 ở Power BI

3.6.13 Câu truy vấn 13: Thống kê số lượng phim, tổng số lượt đánh giá, điểm đánh giá trung bình và độ phổ biến của các bộ phim có ngôn ngữ gốc là tiếng Anh và Pháp và được phát hành từ năm 2016 đến 2023 và có công ty sản xuất tại Mỹ

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Kéo thả các mục ở Dimension và các thông tin như hình và nhấn “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Dimension	Hierarchy	Operator	Filter Expression		
Dim Date	# Year	Equal	{ 2016, 2017, 2018, 2020, 2019, 2021, 2022, 2023 }		
Dim Country	# Production Countries	Equal	{ United States of America }		
Dim Language	# Original Language	Equal	{ fr, en }		
Year	Original Language	Number of Movies	Vote Co...	AverageRating	Popularity
2016	en	251	33222	6.54127985062128	4.40519528417192
2017	en	330	30658	6.17200406747552	5.50914494628492
2018	en	307	30552	6.76764867906016	4.29123835798372
2019	en	312	23295	6.50529566755825	5.8676349088717
2020	en	269	23660	6.67529853872596	7.56989361231144
2021	en	318	15101	6.83173620406385	5.45578350080657
2022	en	276	8949	6.49648078117494	6.72560216278158
2023	fr	1	70	5.1710000038147	7.05600023269653
	en	215	2876	6.52029339909056	11.4831350139669

Hình 3.156 Kết quả truy vấn câu 13 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```
SELECT
    {[Measures].[Number of Movies],[Measures].[Vote Count],
    [Measures].[AverageRating],[Measures].[Popularity]} ON COLUMNS,
    Non empty
    ([Dim Date].[Year].&[2016] : [Dim Date].[Year].&[2023],
    {[Dim Language].[Original Language].&[fr],
    [Dim Language].[Original Language].&[en]}) ON ROWS
FROM [SSIS]
WHERE ([Dim Country].[Production Countries].&[United States of America]);
```

		Number of Movies	Vote Count	AverageRating	Popularity
2016	en	251	33222	6.54127985062128	4.40519528417192
2017	en	330	30658	6.17200406747552	5.50914494628492
2018	en	307	30552	6.76764867906016	4.29123835798372
2019	en	312	23295	6.50529566755825	5.8676349088717
2020	en	269	23660	6.67529853872596	7.56989361231144
2021	en	318	15101	6.83173620406385	5.45578350080657
2022	fr	1	70	5.1710000038147	7.05600023269653
2022	en	276	8949	6.49648078117494	6.72560216278158
2023	en	215	2876	6.52029339909056	11.4831350139669

Hình 3.157 Kết quả truy vấn câu 13 ở MSSQ

Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo các độ đo và cột như hình.

Kho dữ liệu và OLAP - IS217.P12

The screenshot shows the 'PivotTable Fields' pane in Excel. At the top, there is a list of fields with checkboxes: Budget (unchecked), Number of Movies (checked), Popularity (checked), Revenue (unchecked), Runtime (unchecked), Vote Average (unchecked), and Vote Count (checked). Below this, a message says 'Drag fields between areas below:'. There are four main sections: 'Filters' (Production Countries dropdown), 'Columns' (Σ Values dropdown), 'Rows' (Year and Original Language dropdowns), and 'Values' (Number of Movies, Vote Count, and AverageRating dropdowns).

Hình 3.158 Thiết lập PivotTable Fields của câu 13

Bước 4: Xem kết quả

Production Countries	United States of America					
Row Labels	Number of Movies	Vote Count	AverageRating	Popularity		
2016						
en	251	33222	6.541279851	4.405195284		
2017						
en	330	30658	6.172004067	5.509144946		
2018						
en	307	30552	6.767648679	4.291238358		
2019						
en	312	23295	6.505295668	5.867634909		
2020						
en	269	23660	6.675298539	7.569893612		
2021						
en	318	15101	6.831736204	5.455783501		
2022						
en	276	8949	6.496480781	6.725602163		
fr	1	70	5.171000004	7.056000233		
2023						
en	215	2876	6.520293399	11.48313501		
Grand Total	2279	168383	6.553578253	6.269234784		

Hình 3.159 Kết quả truy vấn câu 13 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấn vào Get data và chọn Analysis Services.

Bước 2: Nhập Server và Database đã tạo ở SSAS. Chọn Import và nhập truy vấn MDX.

SQL Server Analysis Services database

The screenshot shows the 'Get data' dialog in Power BI. It has fields for 'Server' (set to 'HPHAT-TLVERSION\PHAT') and 'Database' (set to 'SSAS'). Below these, there are two radio buttons: 'Import' (selected) and 'DirectQuery'. A large text area labeled 'MDX or DAX query (optional)' contains the following MDX query:

```

SELECT
    {[Measures].[Number of Movies], [Measures].[Vote Count], [Measures].[AverageRating], [Measures].[Non empty]}
    ON COLUMNS
    {[Dim Date].[Year].&[2016] : [Dim Date].[Year].&[2023],
     {[Dim Language].[Original Language].&[fr],
      [Dim Language].[Original Language].&[en]}} ON ROWS
FROM [SSIS]
WHERE ([Dim Country].[Production Countries].&[United States of America]);
  
```

Hình 3.160 Thiết lập truy vấn MDX của câu 13 ở Power BI

Bước 3: Vào Add Columns chọn Conditional Format để thêm 1 cột thể hiện ngôn ngữ gốc thay vì kí tự như cột ban đầu.

The screenshot shows the 'Add Column' dialog in Power BI. On the left, a list of queries shows 'Query13' selected. The main area displays the 'Add Conditional Column' dialog. It has a table with two rows under 'If' and one row under 'Else If'. The first 'If' row has 'Column Name' set to '[Dim Language].[...]', 'Operator' set to 'equals', 'Value' set to 'ABC 123' with a dropdown arrow pointing to 'fr', and 'Output' set to 'ABC 123' with a dropdown arrow pointing to 'France'. The 'Else If' row has 'Column Name' set to '[Dim Language].[...]', 'Operator' set to 'equals', 'Value' set to 'ABC 123' with a dropdown arrow pointing to 'en', and 'Output' set to 'ABC 123' with a dropdown arrow pointing to 'English'. Below these rows is a button 'Add Clause'. To the right of the dialog, the 'Properties' pane shows 'Name: Query13' and the 'Applied Steps' pane shows 'Query13 > Renamed Columns'. Buttons 'OK' and 'Cancel' are at the bottom right of the dialog.

Hình 3.161 Tạo một cột mới để thể hiện rõ ngôn ngữ gốc

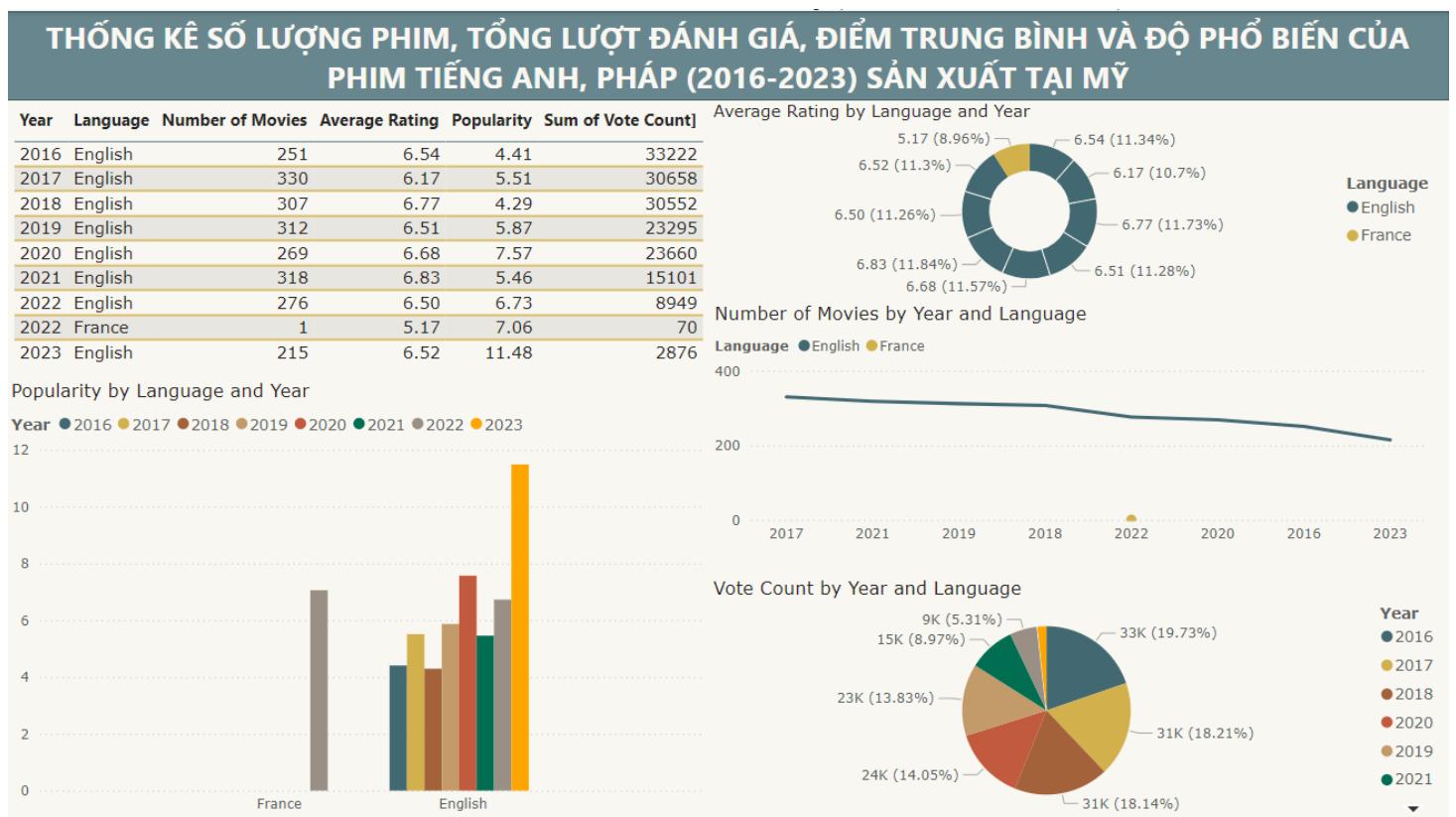
Kho dữ liệu và OLAP - IS217.P12

Bước 4: Ở Tab Table view điều chỉnh tên cột và kiểu dữ liệu phù hợp.

The screenshot shows the Power BI interface with the 'Table tools' tab selected. A table is displayed with the following columns: [Dim Language].[Original Language].[Original Language].[MEMBER_CAPTION], Number of Movies, Vote Count, Average Rating, Popularity, and Language. The 'Popularity' column is currently selected, indicated by a green border around its header. The 'Language' column is also visible. The 'Data' pane on the right shows various measures and dimensions, with 'Popularity' being one of the selected items.

Hình 3.162 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 13

Bước 5: Ở Tab Report view tạo Visualizations cho câu truy vấn.



Hình 3.163 Kết quả của truy vấn câu 13 ở Power BI

3.6.14 Câu truy vấn 14: Tính tổng số lượng phim của các công ty sản xuất theo các tháng từ năm 2014 đến năm 2017

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Kho dữ liệu và OLAP - IS217.P12

Bước 1: Kéo thả các mục ở **Dimension** và các thông tin như hình và nhấn “**Click to execute the query**” để thực thi câu lệnh và được kết quả như sau.

Dimension	Hierarchy	Operator	Filter Expression
Dim Date	Year	Equal	{ 2014, 2015, 2016, 2017 }
<Select dimension>			

Production Companies	Month	Number of Movies
#1NFLUENCE Production	1	1
011 Productions	12	1
011 Productions	8	3
0708 Films	7	1
1/27 Pictures	11	1
100% Halal	9	1
100% Halal, VICE Media	2	1
100Film	12	1
1201	10	1
13 Entertainment, Visinema Pictures	7	1
14 Reels Entertainment	2	1
14 Reels Entertainment, Eros International	9	1
1515 Productions	12	1

Hình 3.164 Kết quả truy vấn câu 14 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```

SELECT
{[Measures].[Number of Movies]} ON COLUMNS,
NONEMPTY(
CROSSJOIN(
[Dim Company].[Production Companies].[Production Companies].Members,
[Dim Date].[Month].[Month].Members
),
{[Measures].[Revenue]}
) ON ROWS
FROM [SSIS]
WHERE ([Dim Date].[Year].&[2014] : [Dim Date].[Year].&[2017]);

```

		Number of Movies
#1NFLUENCE Production	1	1
011 Productions	12	1
011 Productions	8	3
0708 Films	7	1
1/27 Pictures	11	1
100% Halal	9	1
100% Halal, VICE Media	2	1
100Film	12	1
1201	10	1
13 Entertainment, Visinema Pictures	7	1

Hình 3.165 Kết quả truy vấn câu 14 ở MSSQ

Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo các độ đo và cột như hình.

The screenshot shows the PivotTable Fields pane. At the top, under 'Dim Company', 'Production Companies' is selected. Below it, under 'Dim Country', 'Production Countries' is listed. In the center, there's a message: 'Drag fields between areas below:'. On the left, under 'Filters', 'Year' is selected. On the right, under 'Columns', nothing is selected. At the bottom, under 'Rows', 'Production Companies' and 'Month' are listed. Under 'Values', 'Number of Movies' is selected.

Hình 3.166 Thiết lập PivotTable Fields của câu 14

Bước 4: Xem kết quả

Row Labels	Number of Movies
#1NFLUENCE Production	1
011 Productions	12
0708 Films	7
1/27 Pictures	11
100% Halal	9
100% Halal, VICE Media	2
100Film	12
1201	10
13 Entertainment, Visinema Pictures	7
14 Reels Entertainment	2
14 Reels Entertainment, Eros International	1

Hình 3.167 Kết quả truy vấn câu 14 ở Excel

Thực hiện trong Power BI

Bước 1: Trong Tab Home, nhấn vào Get data và chọn Analysis Services.

Bước 2: Nhập Server và Database đã tạo ở SSAS. Chọn Import và nhập truy vấn MDX.

SQL Server Analysis Services database

The screenshot shows the 'Get data' configuration dialog for connecting to an Analysis Services database. The 'Server' field contains 'HPHAT-TLVERSION\PHAT'. The 'Database' field contains 'SSAS'. The 'Import' radio button is selected, while 'DirectQuery' is unselected. Below the fields, there is a section titled 'MDX or DAX query (optional)' containing the following MDX query:

```

SELECT
    {[Measures].[Number of Movies]} ON COLUMNS,
    NONEMPTY(
        CROSSJOIN(
            [Dim Company].[Production Companies].[Production Companies].Members,
            [Dim Date].[Month].[Month].Members
        ),
        {[Measures].[Revenue]}
    ) ON ROWS
FROM [SSIS]
WHERE ([Dim Date].[Year].&[2014] : [Dim Date].[Year].&[2017]);

```

Hình 3.168 Thiết lập truy vấn MDX của câu 14 ở Power BI

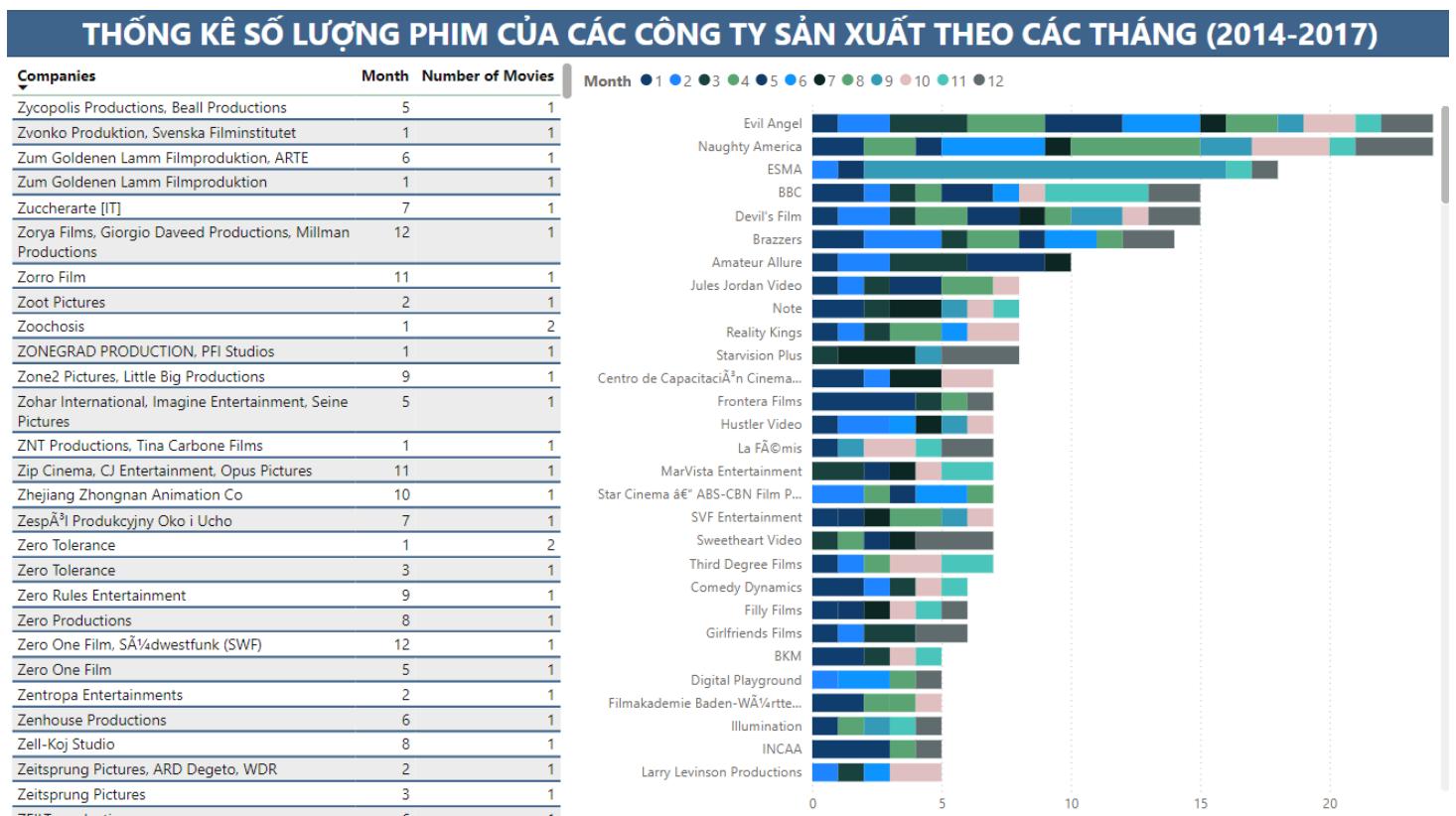
Bước 3: Ở Tab Table view điều chỉnh tên cột và kiểu dữ liệu phù hợp.

The screenshot shows the 'Table tools' ribbon with the 'Column tools' tab selected. The 'Name' dropdown is set to 'Popularity' and the 'Data type' dropdown is set to 'Decimal number'. Other tabs include 'Format' (with options like '\$', '%', '#,##0.00'), 'Summarization' (with 'Sum' selected), 'Data category' (set to 'Uncategorized'), 'Sort by column' (with a downward arrow icon), 'Data groups' (with a 'Groups' icon), 'Manage relationships' (with a 'Relationships' icon), and 'New column' (with a 'Calculations' icon). The main area displays a table with columns: [Dim Language].[Original Language].[Original Language].[MEMBER_CAPTION], Number of Movies, Vote Count, Average Rating, Popularity, and Language. The 'Popularity' column is currently highlighted. The bottom right corner shows a 'Data' pane with a search bar and a list of queries: Query1, Query10, Query11, Query12, and Query13, along with some calculated columns and measures.

Hình 3.169 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 14

Bước 5: Ở Tab Report view tạo Visualizations cho câu truy vấn.

SVTH: Nguyễn Hồng Phát



Hình 3.170 Kết quả của truy vấn câu 14 ở Power BI

3.6.15 Câu truy vấn 15: Tính tổng số lượng phim theo thể loại và đạo diễn

Thực hiện trong Visual Studio 2022 - Truy vấn Manual

Bước 1: Kéo thả các mục ở **Dimension** và các thông tin như hình và nhấn “Click to execute the query” để thực thi câu lệnh và được kết quả như sau.

Genres List	Director	Number of Movies
['Action', 'Action', 'Crime']	Kate Perry	1
['Action', 'Adventure', 'Animation', 'Drama']	Randal Kleiser	1
['Action', 'Adventure', 'Animation', 'Family', 'Comedy', 'TV Movie']	Michael Shea	1
['Action', 'Adventure', 'Animation', 'Family', 'Fantasy']	Sirus Alvand	1
['Action', 'Adventure', 'Animation', 'Family']	Clint Eastwood	1
['Action', 'Adventure', 'Animation', 'Fantasy', 'Comedy', 'Science Fiction']	Joyce Bernal	1
['Action', 'Adventure', 'Animation', 'Fantasy']	Allen Walls	1
['Action', 'Adventure', 'Animation', 'Fantasy']	Noah Nicholson	1
['Action', 'Adventure', 'Animation']	Gerry Chiniquy	1
['Action', 'Adventure', 'Comedy', 'Crime', 'Animation']	Ahammed Khabeer	1
['Action', 'Adventure', 'Comedy', 'Drama', 'Animation']	Allan Harmon	1
['Action', 'Adventure', 'Comedy', 'Drama', 'Family']	Bizhan M. Tong	1
['Action', 'Adventure', 'Comedy', 'Fantasy', 'Romance', 'Crime']	Clive Rees	1

Hình 3.171 Kết quả truy vấn câu 15 ở Visual Studio 2022

Thực hiện trong MSSQ - Truy vấn MDX

Câu truy vấn MDX:

```
SELECT {[Measures].[Number of Movies]} ON COLUMNS,
NON EMPTY
CROSSJOIN(
    [Dim Genres List].[Genres List].[Genres List].Members,
    [Dim Director].[Director].[Director].Members
) ON ROWS
FROM [SSIS];
```

Messages Results		Number of Movies
['Action', 'Action', 'Crime']	Kate Perry	1
['Action', 'Adventure', 'Animation', 'Drama']	Randal Kleiser	1
['Action', 'Adventure', 'Animation', 'Family', 'Comedy', 'TV Movie']	Michael Shea	1
['Action', 'Adventure', 'Animation', 'Family', 'Fantasy']	Sirus Alvand	1
['Action', 'Adventure', 'Animation', 'Family']	Clint Eastwood	1
['Action', 'Adventure', 'Animation', 'Fantasy', 'Comedy', 'Science Ficti...']	Joyce Buel	1
['Action', 'Adventure', 'Animation', 'Fantasy']	Allen Walls	1
['Action', 'Adventure', 'Animation', 'Fantasy']	Noah Nicholson	1
['Action', 'Adventure', 'Animation']	Gerry Chiniquy	1
['Action', 'Adventure', 'Comedy', 'Crime', 'Animation']	Ahammed Khabeer	1
['Action', 'Adventure', 'Comedy', 'Drama', 'Animation']	Allan Harmon	1
['Action', 'Adventure', 'Comedy', 'Drama', 'Family']	Bizhan M. Tong	1
['Action', 'Adventure', 'Comedy', 'Fantasy', 'Romance', 'Crime']	Clive Rees	1
['Action', 'Adventure', 'Comedy', 'Horror', 'Science Fiction']	Steven Zaillian	1
['Action', 'Adventure', 'Comedy', 'Science Fiction', 'Thriller']	Benjamin Vu	1

Hình 3.172 Kết quả truy vấn câu 15 ở MSSQ

Thực hiện trong Excel

Bước 1: Trong PivotTable Fields, kéo các độ đo và cột như hình.

The screenshot shows the 'PivotTable Fields' pane in Excel. At the top, it says 'Drag fields between areas below:'. Below this are four sections: 'Filters', 'Columns', 'Rows', and 'Values'. The 'Rows' section contains two items: 'Genres List' and 'Director'. The 'Values' section contains one item: 'Number of Movies'.

Hình 3.173 Thiết lập PivotTable Fields của câu 15

Bước 4: Xem kết quả

Row Labels	Number of Movies
■ ['Action', 'Action', 'Crime'] Kate Perry	1
■ ['Action', 'Adventure', 'Animation', 'Drama'] Randal Kleiser	1
■ ['Action', 'Adventure', 'Animation', 'Family', 'Comedy', 'TV Movie'] Michael Shea	1
■ ['Action', 'Adventure', 'Animation', 'Family', 'Fantasy'] Sirus Alvand	1
■ ['Action', 'Adventure', 'Animation', 'Family'] Clint Eastwood	1
■ ['Action', 'Adventure', 'Animation', 'Fantasy', 'Comedy', 'Science Fiction'] Joyce Buel	1
■ ['Action', 'Adventure', 'Animation', 'Fantasy'] Allen Walls	1
■ ['Action', 'Adventure', 'Animation'] Noah Nicholson	1
■ ['Action', 'Adventure', 'Animation'] Gerry Chiniquy	1
■ ['Action', 'Adventure', 'Comedy', 'Crime', 'Animation'] Ahammed Khabeer	1
■ ['Action', 'Adventure', 'Comedy', 'Drama', 'Animation'] Allan Harmon	1
■ ['Action', 'Adventure', 'Comedy', 'Drama', 'Family'] Bizhan M. Tong	1
■ ['Action', 'Adventure', 'Comedy', 'Fantasy', 'Romance', 'Crime'] Clive Rees	1

Hình 3.174 Kết quả truy vấn câu 15 ở Excel

Thực hiện trong Power BI**Bước 1:** Trong Tab Home, nhấp vào **Get data** và chọn **Analysis Services**.**Bước 2:** Nhập **Server** và **Database** đã tạo ở SSAS. Chọn **Import** và nhập truy vấn MDX.

SQL Server Analysis Services database

Server ⓘ
HPHAT-TLVERSION\PHAT

Database
SSAS

Import
 DirectQuery

▲ MDX or DAX query (optional)

```
SELECT
    {[Measures].[Number of Movies]} ON COLUMNS,
    NON EMPTY
    CROSSJOIN(
        [Dim Genres List].[Genres List].[Genres List].Members,
        [Dim Director].[Director].[Director].Members
    ) ON ROWS
FROM [SSIS];
```

Hình 3.175 Thiết lập truy vấn MDX của câu 15 ở Power BI

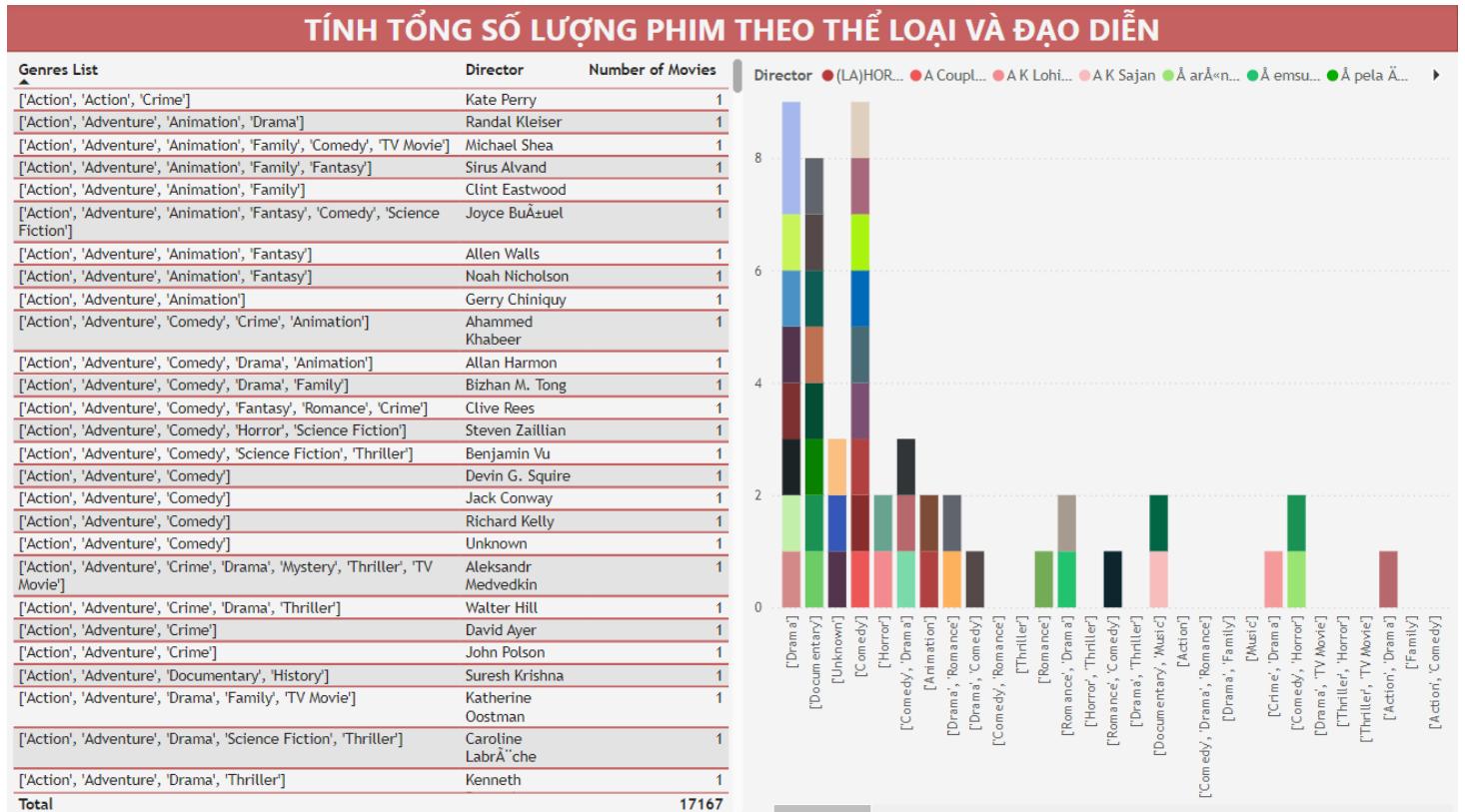
Bước 3: Ở Tab Table view điều chỉnh tên cột và kiểu dữ liệu phù hợp.

Genres List	Director	Number of Movies
[Drama]	A. Bhimsingh	1
[Drama]	A.J. Hall	1
[Drama]	Àetin A'nanÃš	1
[Drama]	Ã‰douard Bergeon	1
[Drama]	Ãœ Duy PhÃ¢c	1
[Drama]	Aarav Ramnani	1
[Drama]	Aaron Boltz	1
[Drama]	Aaron Nee, Adam Nee	1
[Drama]	Aaron Rottinghaus	1
[Drama]	Aaron Seltzer	1
[Drama]	Abba Makama	1
[Drama]	Abdul sameeh	1
...

Hình 3.176 Điều chỉnh tên cột và kiểu dữ liệu của truy vấn câu 15

Bước 5: Ở Tab Report view tạo Visualizations cho câu truy vấn.

Kho dữ liệu và OLAP - IS217.P12



Hình 3.177 Kết quả của truy vấn câu 15 ở Power BI

CHƯƠNG 4. QUÁ TRÌNH KHAI THÁC DỮ LIỆU (DATA MINING)

4.1 Mô tả bộ dữ liệu

STT	Tên cột	Kiểu dữ liệu	Ý nghĩa
1	id	int	Mã định danh duy nhất cho mỗi phim trong TMDB
2	title	varchar	Tên chính thức của phim
3	vote_average	float	Đánh giá trung bình của phim theo thang điểm từ 0 đến 10
4	vote_count	int	Số phiếu đánh giá bộ phim
5	status	varchar	Trạng thái hiện tại của phim
6	release_date	datetime	Ngày phim chính thức ra mắt
7	revenue	int	Doanh thu của bộ phim
8	runtime	int	Thời lượng của phim
9	budget	int	Ngân sách sản xuất của bộ phim
10	imdb_id	varchar	ID của phim trên nền tảng IMDb
11	original_language	varchar	Ngôn ngữ ban đầu của bộ phim

12	popularity	float	Điểm phổ biến của phim
13	production_companies	varchar	Công ty tham gia sản xuất
14	production_countries	varchar	Quốc gia tham gia sản xuất
15	spoken_languages	varchar	Các ngôn ngữ được sử dụng trong bộ phim
16	Director	varchar	Đạo diễn của bộ phim
17	genres_list	varchar	Thể loại của bộ phim
18	adult	bool	Thể hiện bộ phim chỉ phù hợp cho người xem thành niên
19	overview_sentiment	float	Cảm xúc tổng quan trong phần mô tả phim

4.2 Mô tả bài toán

4.2.1 Mô tả bài toán

Với bộ dữ liệu phim điện ảnh gồm nhiều các thuộc tính từ nguồn IMDb và TMDb, ở phần khai phá dữ liệu (**Data Mining**), em sẽ tập trung vào việc khai thác dữ liệu để dự đoán mức độ hài lòng của khán giả đối với các bộ phim.

Bài toán tập trung vào việc phân tích và dự đoán mức độ hài lòng dựa trên các đặc trưng liên quan đến phim. Mức độ hài lòng của khán giả được thể hiện thông qua **đánh giá trung bình** (*vote_average*) và các đặc trưng khác như **số phiếu đánh giá** (*vote_count*), **doanh thu** (*revenue*), **ngân sách sản xuất** (*budget*), **thời lượng phim** (*runtime*), và các yếu tố khác (thể loại, cảm xúc mô tả, ngôn ngữ,...).

4.2.2 Mục tiêu

Mục tiêu là phân loại các bộ phim vào **3 nhóm** mức độ hài lòng dựa trên đánh giá trung bình của khán giả (*vote_average*). Các nhóm này có thể được phân loại theo thang điểm đánh giá như sau:

- **Nhóm 1 (Flop):** Các bộ phim có điểm đánh giá trung bình dưới 5.
- **Nhóm 2 (Average):** Các bộ phim có điểm đánh giá trung bình từ 5 trở lên.
- **Nhóm 3 (Hit):** Các bộ phim có điểm đánh giá trung bình từ 8 trở lên.

Mô hình sẽ sử dụng các đặc trưng mà tiếp theo đây em sẽ phân tích và làm rõ.

4.3 Khám phá cơ bản bộ dữ liệu

Bước 1: Thiết lập các thư viện cần thiết.

```
!pip install catboost
!pip install shap
import numpy as np
import polars as pl
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime
from sklearn.metrics import silhouette_score
from sklearn.metrics import davies_bouldin_score
from scipy.stats import pearsonr
from sklearn.impute import KNNImputer
from sklearn.compose import make_column_transformer
from sklearn.pipeline import Pipeline, make_pipeline
import sklearn.linear_model as skl_lm
import sklearn.preprocessing as skl_pre
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.cluster import KMeans
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.model_selection import GridSearchCV, KFold
from sklearn.preprocessing import StandardScaler, MinMaxScaler, PowerTransformer,
RobustScaler, Normalizer
from sklearn.metrics import accuracy_score, f1_score, precision_score,
recall_score, confusion_matrix, ConfusionMatrixDisplay
from sklearn.ensemble import RandomForestClassifier
from lightgbm import LGBMClassifier
from catboost import CatBoostClassifier
import shap
```

Bước 2: Thông tin cơ bản của bộ dữ liệu

Kho dữ liệu và OLAP - IS217.P12

```
[ ] df_rate.info()
→ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 1072255 entries, 0 to 1072254
Data columns (total 42 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   id               1072255 non-null  int64  
 1   title             1072255 non-null  object  
 2   vote_average      1072255 non-null  float64 
 3   vote_count        1072255 non-null  int64  
 4   status             1072255 non-null  object  
 5   release_date      921699 non-null  datetime64[ns]
 6   revenue            1072255 non-null  float64 
 7   runtime            1072255 non-null  int64  
 8   adult              1072255 non-null  bool    
 9   backdrop_path     292826 non-null  object  
10   budget             1072255 non-null  float64 
11   homepage           115519 non-null  object  
12   imdb_id            589364 non-null  object  
13   original_language 1072255 non-null  object  
14   original_title     1072255 non-null  object  
15   overview            870074 non-null  object  
16   popularity          1072255 non-null  float64 
17   poster_path         754417 non-null  object  
18   tagline             152783 non-null  object  
19   production_companies 496572 non-null  object  
20   production_countries 615879 non-null  object  
21   spoken_languages    631998 non-null  object  
22   keywords             1072255 non-null  object  
23   release_year        921699 non-null  float64 
24   Director            1072255 non-null  object  
25   AverageRating       11751 non-null   float64 
26   Poster_Link          2770 non-null   object  
27   Certificate          29835 non-null  object  
28   IMDB_Rating          30312 non-null  float64 
29   Meta_score           2331 non-null   float64 
30   Star1                30313 non-null  object  
31   Star2                2770 non-null   object  
32   Star3                2770 non-null   object  
33   Star4                2770 non-null   object  
34   Writer                646580 non-null  object  
35   Director_of_Photography 336853 non-null  object  
36   Producers             425308 non-null  object  
37   Music_Composer        155476 non-null  object  
38   genres_list            1072255 non-null  object  
39   Cast_list              1072255 non-null  object  
40   overview_sentiment    1072255 non-null  float64 
41   all_combined_keywords 1072255 non-null  object  
dtypes: bool(1), datetime64[ns](1), float64(9), int64(3), object(28)
```

Hình 4.1 Thông tin cơ bản của bộ dữ liệu

Bước 3: Loại bỏ các cột bị thiếu trên 50% và chuyển đổi các cột sang kiểu dữ liệu phù hợp.

```
# Loại bỏ các cột thiếu hơn 50% dữ liệu
df_rate.dropna(thresh=int(0.5*len(df_rate)), axis=1, inplace=True)

# Chuyển đổi dữ liệu sang dạng phù hợp
df_rate['revenue'] = df_rate['revenue'].astype(float)
df_rate['budget'] = df_rate['budget'].astype(float)

df_rate['release_date'] = pd.to_datetime(df_rate['release_date'], errors='coerce')

# Loại bỏ các dòng có NaN hoặc giá trị vô hạn trong 'release_year'
df_rate = df_rate[~df_rate['release_year'].isna() &
~df_rate['release_year'].isin([np.inf, -np.inf])]
df_rate['release_year'] = df_rate['release_year'].astype(int)
```

Bước 4: Loại bỏ các cột không cần thiết.

```
columns2drop = ['imdb_id', 'Writer', 'Director', 'Cast_list', 'original_title'
```

```

        , 'all_combined_keywords' , 'overview', 'poster_path', 'keywords', 'id']

df_rate = df_rate.drop(columns=columns2drop, errors='ignore')

```

Bước 5: Tính toán các thống kê cơ bản cho các trường có thể tính toán được (dữ liệu dạng số).

```

[ ] columns_to_analyze = ['vote_average', 'vote_count', 'revenue', 'runtime', 'budget', 'popularity']

statistics = df[columns_to_analyze].describe()

print(statistics)

    vote_average   vote_count      revenue      runtime      budget \
count  17167.000000  17167.000000  1.716700e+04  17167.000000  1.716700e+04
mean     4.239156     43.654978   1.176313e+06    78.061106   5.707823e+05
std      3.086769    332.524327   2.340053e+07    48.757295   5.483747e+06
min      0.000000     0.000000   0.000000e+00     0.000000   0.000000e+00
25%     0.000000     0.000000   0.000000e+00     34.000000   0.000000e+00
50%     5.300000     2.000000   0.000000e+00     88.000000   0.000000e+00
75%     6.600000    11.000000   0.000000e+00    105.000000   0.000000e+00
max     10.000000   18352.000000  2.068224e+09    993.000000   2.500000e+08

           popularity
count  17167.000000
mean     2.834187
std      9.881448
min      0.000000
25%     0.600000
50%     1.137000
75%     2.566000
max     810.288000

```

Hình 4.2 Tính toán thống kê cơ bản

4.4 Xử lý đặc trưng cho mô hình máy học

- Xử lý đặc trưng cho mô hình máy học (**Feature Engineering for Machine Learning**) là một bước quan trọng trong quy trình xây dựng mô hình, đóng vai trò quyết định đến hiệu quả và độ chính xác của dự đoán.
- Quá trình này bao gồm việc chọn lọc, chuyển đổi, và tạo ra các đặc trưng phù hợp từ dữ liệu để tối ưu hóa khả năng học của thuật toán.
- Mục tiêu chính là làm nổi bật những thông tin quan trọng nhất từ dữ liệu, đồng thời loại bỏ các yếu tố gây nhiễu hoặc không liên quan.
- Những kỹ thuật phổ biến trong xử lý đặc trưng bao gồm mã hóa dữ liệu phân loại, chuẩn hóa dữ liệu số, giảm chiều dữ liệu, và tạo đặc trưng mới.
- Ở bước này, em sẽ phân tích từng thuộc tính từ đó chọn ra các cột cần thiết để đưa vào mô hình máy học.

4.4.1 Cột revenue

```
zero_revenue = df_rate[df_rate['revenue'] == 0].shape[0]/df_rate.shape[0] * 100
```

➔ Phim có doanh thu là 0: **97.99%**

Trong phần xử lý đặc trưng của **Revenue** (doanh thu), nhận thấy rằng một phần lớn các bộ phim trong bộ dữ liệu có doanh thu bằng 0. Cụ thể, **97.99%** số bộ phim có **doanh thu là 0**, điều này có thể do một số bộ phim chưa phát hành, hoặc dữ liệu không đầy đủ.

Các bộ phim chưa phát hành (với trạng thái không phải 'Released') có thể có doanh thu bằng 0, vì chúng chưa tạo ra doanh thu.

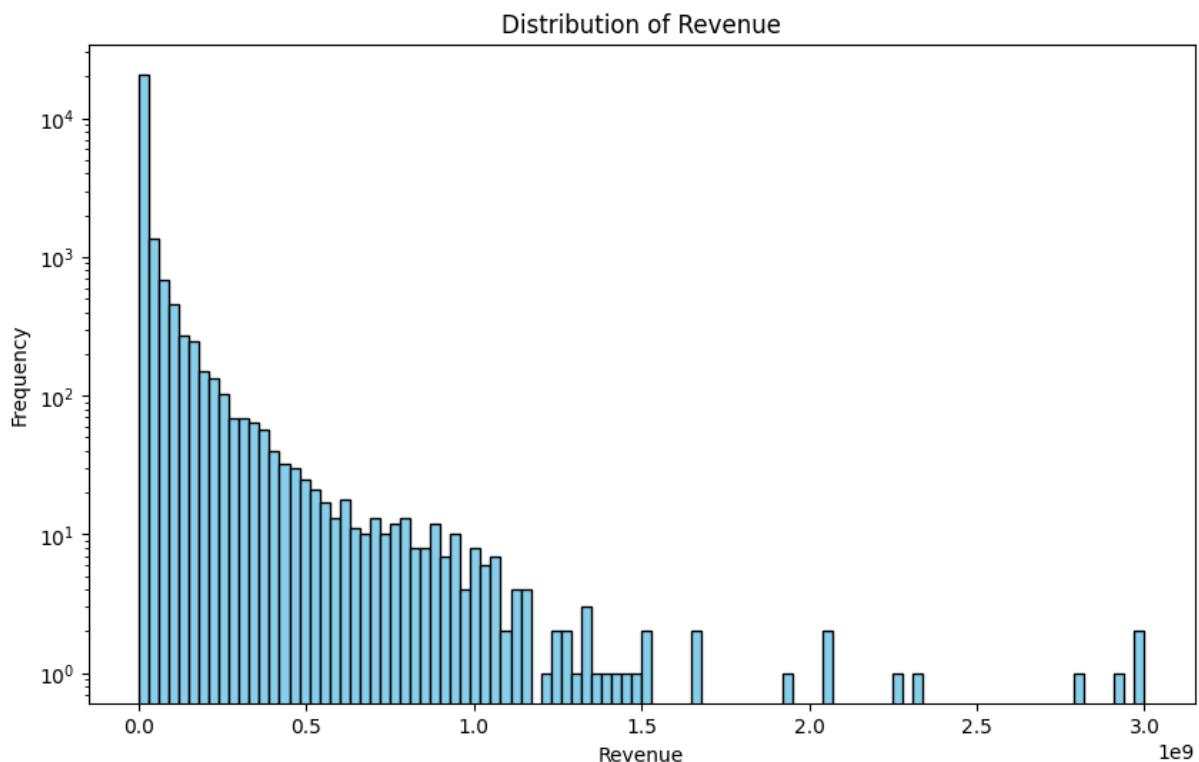
Tuy nhiên, đối với các bộ phim đã phát hành (trạng thái là 'Released'), doanh thu không thể bằng 0. Điều này có thể phản ánh một vấn đề trong dữ liệu hoặc một bộ phim không thành công, không có doanh thu nào.

Do đó, bộ dữ liệu sẽ được lọc theo quy tắc sau:

- Các bộ phim chưa phát hành sẽ được giữ lại bất kể doanh thu có bằng 0 hay không.
- Các bộ phim đã phát hành phải có doanh thu khác 0.

```
df_rate = df_rate[((df_rate['revenue'] >= 0) & (df_rate['status'] != 'Released')) | ((df_rate['revenue'] != 0) & (df_rate['status'] == 'Released'))]
```

➔ Số dòng dữ liệu sau khi xử lý dữ liệu ngoại lai: **24861**



Hình 4.3 Phân phối của cột revenue

4.4.2 Cột vote_average

Trong quá trình chuẩn bị dữ liệu, cột **vote_average** (điểm đánh giá trung bình) được chọn làm biến mục tiêu để phân loại mức độ hài lòng của khán giả đối với các bộ phim.

```
(df_rate[ (df_rate['vote_average'] == 0) ].shape[0])
```

➔ Phim có rating là 0: **8411**

Trong bộ dữ liệu, có một số lượng đáng kể các bộ phim có điểm đánh giá trung bình (vote_average) bằng 0

Điểm đánh giá bằng 0 thường có thể xảy ra vì:

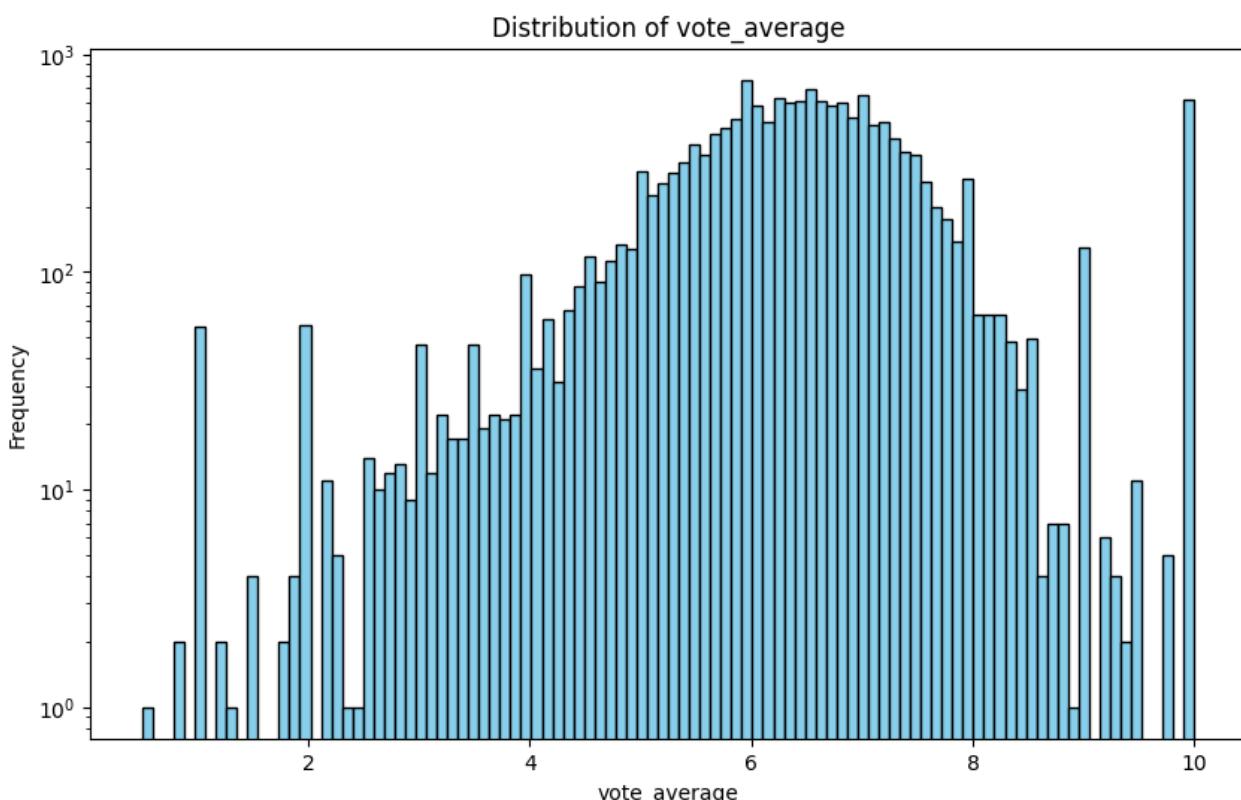
- Phim chưa nhận được đánh giá từ khán giả.
- Lỗi hoặc thiếu dữ liệu.

Vì vote_average là cột mục tiêu để phân loại, các giá trị bằng 0 không có ý nghĩa cho việc học máy. Những phim này không cung cấp thông tin hữu ích cho mô hình dự đoán mức độ hài lòng của khán giả.

Do đó, loại bỏ các bộ phim có điểm đánh giá bằng 0 ra khỏi bộ dữ liệu.

```
df_rate = df_rate[df_rate['vote_average'] != 0]
```

➔ Số dòng dữ liệu sau khi loại bỏ các phim có đánh giá là 0: **16450**.



Hình 4.4 Phân phối của cột vote_average

4.4.3 Cột runtime

200

Cột **runtime** (thời lượng phim) phản ánh thời gian chạy của bộ phim, được xem là một đặc trưng quan trọng vì thời lượng có thể ảnh hưởng đến mức độ hài lòng của khán giả.

```
print(f"Phim có thời lượng là 0: {df_rate[df_rate['runtime'] == 0].shape[0]}")
print(f"Phim có thời lượng là 5h: {df_rate[df_rate['runtime'] > 300].shape[0]}")
```

➔ Phim có thời lượng là 0: **461**

➔ Phim có thời lượng là 5 tiếng: **16**

Có **461 bộ phim** có thời lượng bằng **0 phút**, điều này không hợp lý. Nguyên nhân có thể do dữ liệu thiếu hoặc lỗi nhập liệu. Ngoài ra, có **16 bộ phim** có thời lượng lớn hơn **300 phút** (tương đương 5 tiếng), một giá trị rất hiếm thấy và không thực tế đối với hầu hết các bộ phim.

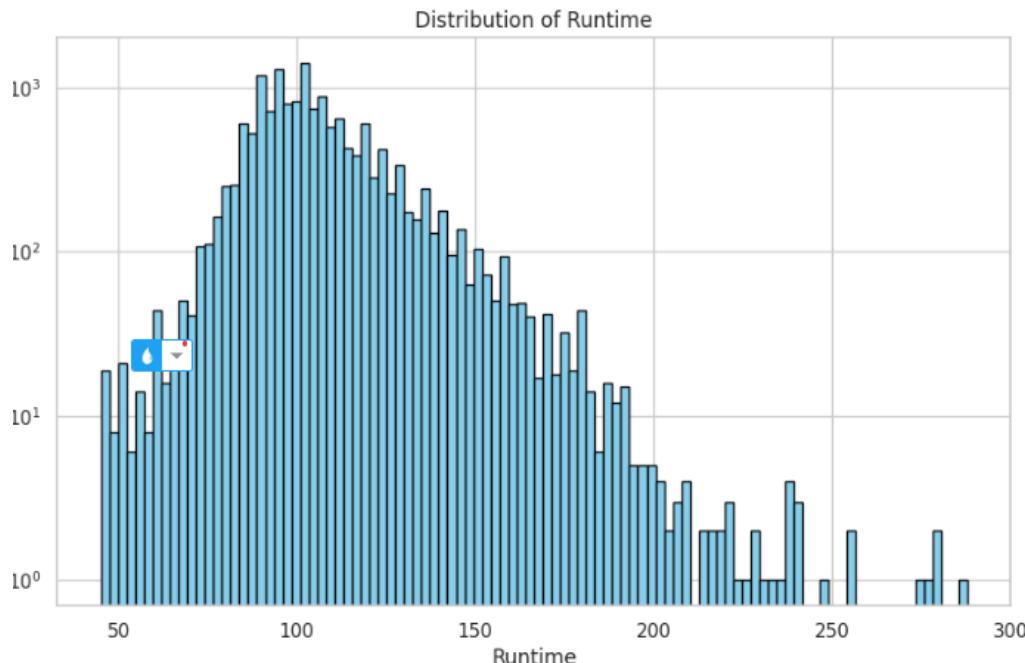
Đối với các phim có thời lượng bằng 0, giá trị này được thay thế bằng **giá trị trung vị (median)** của các phim có thời lượng hợp lệ (lớn hơn 0). Trung vị là một lựa chọn tốt vì nó ít bị ảnh hưởng bởi các giá trị ngoại lai so với trung bình.

```
median_runtime = df_rate['runtime'][df_rate['runtime'] > 0].median()
df_rate['runtime'] = df_rate['runtime'].replace(0, median_runtime)
```

lọc các phim có thời lượng nằm trong khoảng hợp lý từ **45 phút** đến dưới **300 phút**, loại bỏ các phim quá ngắn (không đủ thông tin để đánh giá) hoặc quá dài (không phổ biến).

```
df_rate = df_rate[(df_rate['runtime'] >= 45) & (df_rate['runtime'] < 300)]
print(f"Phim có thời lượng từ 45p đến 300p: {df_rate.shape[0]}")
```

➔ Phim có thời lượng từ 45p đến 300p: **15928**



Hình 4.5 Phân phối của cột runtime

4.4.4 Cột title

Phân tích nội dung tiêu đề trực tiếp đòi hỏi xử lý ngôn ngữ tự nhiên phức tạp, tốn kém và thiếu thông tin phân loại rõ ràng. Do đó, đặc trưng **title_length** (độ dài tiêu đề) được tạo ra vì tính đơn giản, dễ tính toán và tiềm năng phản ánh xu hướng đặt tên, chiến lược tiếp thị hoặc đối tượng khán giả của phim, giúp mô hình dự đoán hiệu quả mà không làm tăng độ phức tạp.

```
df_rate['title_length'] = df_rate['title'].apply(len)
print(df_rate['title_length'].describe())
```

count	15928.000000
mean	15.888749
std	9.569325
min	1.000000
25%	10.000000
50%	14.000000
75%	20.000000
max	213.000000
Name:	title_length, dtype: float64

Hình 4.6 Mô tả thống kê của cột title_length

Tiêu đề quá ngắn hoặc quá dài có thể ảnh hưởng đến mức độ phổ biến của phim (thể hiện qua cột popularity hoặc vote_average).

Độ dài tiêu đề có thể đóng vai trò như một đặc trưng phân biệt giữa các loại phim hoặc nhóm khán giả mục tiêu.

4.4.5 Cột budget

Cột budget đại diện cho kinh phí sản xuất của bộ phim, là một đặc trưng quan trọng vì nó thường phản ánh mức độ đầu tư và ảnh hưởng đến chất lượng cũng như mức độ hài lòng của khán giả.

```
zero_budget = df_rate[df_rate['budget'] == 0].shape[0] / df_rate.shape[0] * 100
print(f"Phim có kinh phí là 0: {zero_budget:.2f}%")
```

➔ Phim có kinh phí là 0: **36.05%**

```
budget_10k = df_rate[df_rate['budget'] < 10000].shape[0]
print(f"Phim có kinh phí dưới 10k: {budget_10k}")
```

➔ Phim có kinh phí dưới 10k: **6309**

Có **36.05%** phim có kinh phí bằng 0. Điều này không hợp lý và có thể do dữ liệu bị thiếu hoặc nhập sai. Có **6,309** phim có kinh phí dưới 10,000 (rất thấp), điều này không phản ánh thực tế đối với hầu hết các bộ phim chiếu rạp.

Các bộ phim có kinh phí bằng 0 được thay thế bằng **giá trị trung vị** của các bộ phim có kinh phí lớn hơn 0. Trung vị giúp giảm ảnh hưởng của các giá trị ngoại lai.

```
median_budget = df_rate.loc[df_rate['budget'] > 0, 'budget'].median()
df_rate['re_budget'] = df_rate['budget'].replace(0, median_budget)
```

```
print(f"Số dòng dữ liệu sau khi thay thế các phim có kinh phí bằng 0 bằng trung  
vị: {df_rate.shape[0]}")
```

Các giá trị kinh phí nhỏ hơn **10,000** được thay thế bằng giá trị NaN để biểu thị dữ liệu không đáng tin cậy.

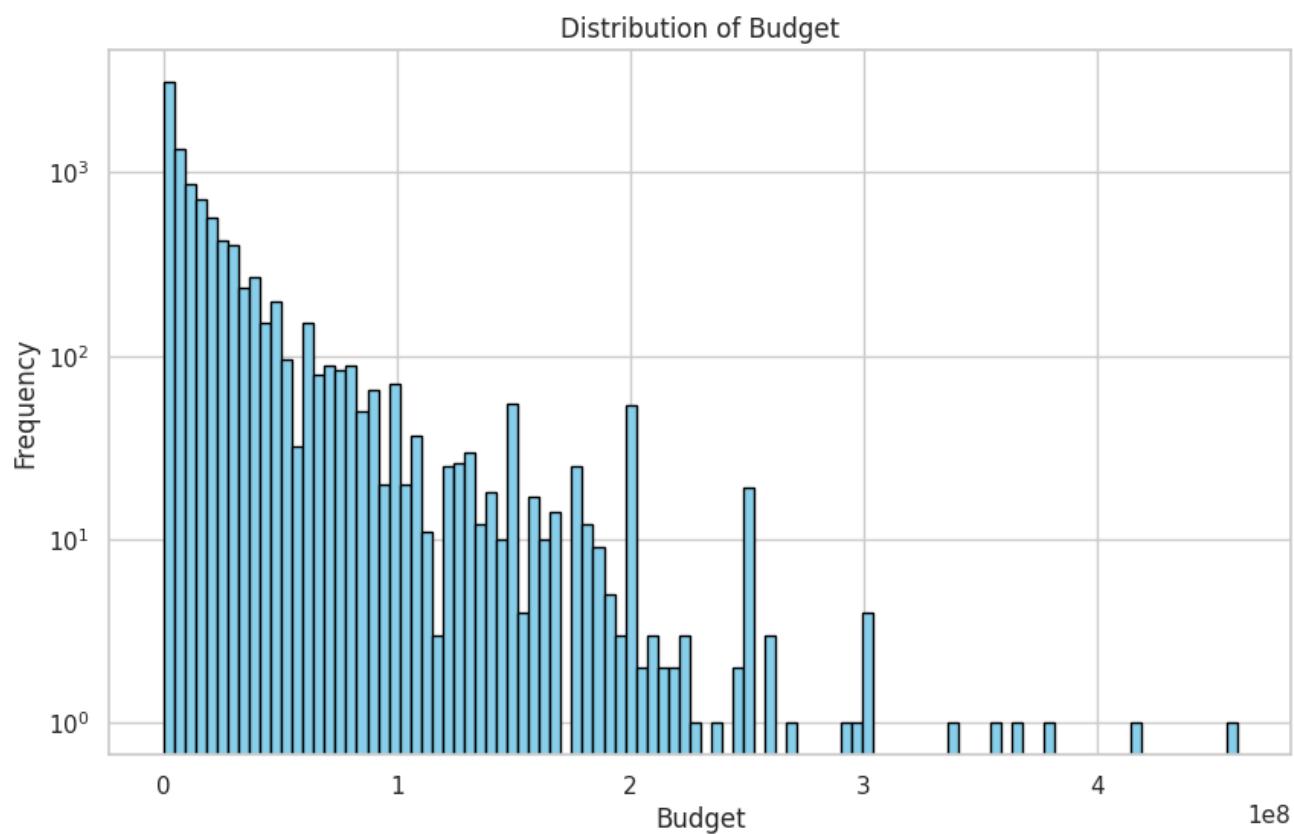
```
df_rate.loc[df_rate['budget'] <= 10000, 'budget'] = np.nan
```

KNN Imputer được sử dụng để dự đoán giá trị kinh phí dựa trên các cột liên quan như **title_length**, **vote_average**, **vote_count**, và **runtime**. Phương pháp này tận dụng mối quan hệ giữa các đặc trưng để điền giá trị phù hợp.

```
imputer = KNNImputer(n_neighbors=5)
df_rate['re_budget'] = imputer.fit_transform(df_rate[['title_length',
'vote_average', 'vote_count', 'runtime']])
```

Các giá trị nhỏ hơn **10,000** được thay thế lại bằng **giá trị trung vị** để đảm bảo dữ liệu nhất quán.

```
median = df_rate.loc[df_rate['budget'] >= 10000, 'budget'].median()
df_rate["re_budget"] = df_rate["budget"].mask(df_rate["budget"] < 10000, median)
```



Hình 4.7 Phân phối của cột budget

4.4.6 Cột popularity

Cột **popularity** đại diện cho mức độ phổ biến của bộ phim, thường được đo bằng lượt xem, đánh giá hoặc sự quan tâm của khán giả.

Ở cột này tạm thời chưa xử lý mà sẽ thực hiện biến đổi một lượt với các đặc trưng số khác ở phần sau.

4.4.7 Cột vote_count

Cột vote_count thể hiện số lượt đánh giá của khán giả cho mỗi bộ phim.

```
print(f"Phim có vote và rating đều là 0: {df_rate[((df_rate['vote_count'] == 0) & (df_rate['vote_average'] == 0))].shape[0]}")  
print(f"Phim có vote là 0: {df_rate[(df_rate['vote_count'] == 0)].shape[0]}")
```

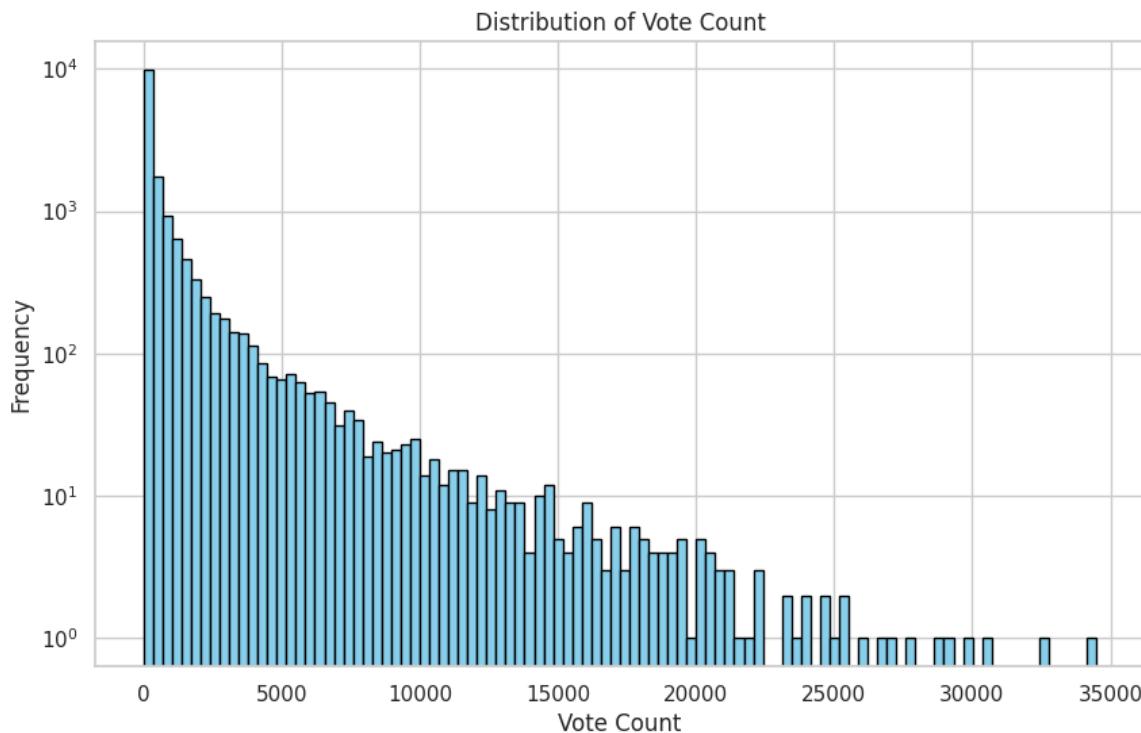
→ Phim có vote và rating đều là 0: 0

→ Phim có vote là 0: 3

Có một số phim có **vote_count** bằng 0, nghĩa là không có lượt đánh giá nào từ khán giả. Trong trường hợp này, cột **vote_average** (đánh giá trung bình) cũng sẽ không có giá trị hợp lệ, gây ảnh hưởng đến quá trình phân tích.

Các phim có **vote_count** bằng 0 sẽ được loại bỏ, vì không có thông tin nào về lượt đánh giá. Việc này giúp giảm thiểu dữ liệu thiếu và đảm bảo tính chính xác của mô hình dự đoán.

```
df_rate = df_rate[df_rate['vote_count'] > 0]  
print(f"Số lượng phim có số vote lớn hơn không 0: {df_rate.shape[0]}")
```



Hình 4.8 Phân phối của cột vote_count

4.4.8 Cột overview_sentiment

Cột **overview_sentiment** biểu thị cảm xúc của các bộ phim dựa trên phần mô tả (overview). Để tìm hiểu thêm về các phân nhóm cảm xúc trong dữ liệu, sử dụng thuật toán **KMeans** để phân nhóm

các phim vào các cụm khác nhau, dựa trên giá trị **overview_sentiment**.

Để tìm số lượng cụm tối ưu, sử dụng **phương pháp elbow**. Phương pháp này giúp xác định số cụm sao cho điểm **inertia** (độ phân tán trong cụm) giảm mạnh. Sử dụng thuật toán **KMeans** với số lượng cụm là 3, gán nhãn cụm cho mỗi bộ phim vào cột **sentiment_cluster**.

```
X = df_rate[['overview_sentiment']]
inertia = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X)
    inertia.append(kmeans.inertia_)

kmeans = KMeans(n_clusters=3, random_state=42)
df_rate['sentiment_cluster'] = kmeans.fit_predict(X)
```

Sử dụng các độ đo sau để đánh giá hiệu suất phân nhóm:

1. Silhouette Score

```
silhouette_avg = silhouette_score(X, kmeans.labels_)
print(f"Silhouette Score: {silhouette_avg}")
```

➔ Silhouette Score: 0.5559049388829899

Điểm gần 0.5 cho thấy phân nhóm hợp lý nhưng có thể cải thiện. Chỉ ra rằng các cụm có sự phân biệt tốt nhưng không hoàn hảo.

2. Inertia

```
print(f"Inertia: {kmeans.inertia_}")
```

➔ Inertia: 193.51950334353415

Giá trị này không quá cao, cho thấy các cụm có độ phân tán vừa phải. Có thể giảm Inertia bằng cách điều chỉnh số lượng cụm.

3. Davies-Bouldin Index

```
db_index = davies_bouldin_score(X, kmeans.labels_)
print(f"Davies-Bouldin Index: {db_index}")
```

➔ Davies-Bouldin Index: 0.5783751633350815

Giá trị thấp cho thấy các cụm phân biệt rõ ràng và không bị chồng lấn nhiều.

Phân tích giá trị **overview_sentiment** trung bình của mỗi cụm để hiểu rõ hơn về các nhóm cảm xúc khác nhau.

```
cluster_means = df_rate.groupby('sentiment_cluster')['overview_sentiment'].mean()
print(f"cluster_means: {cluster_means}")
```

➔ 0 0.323889

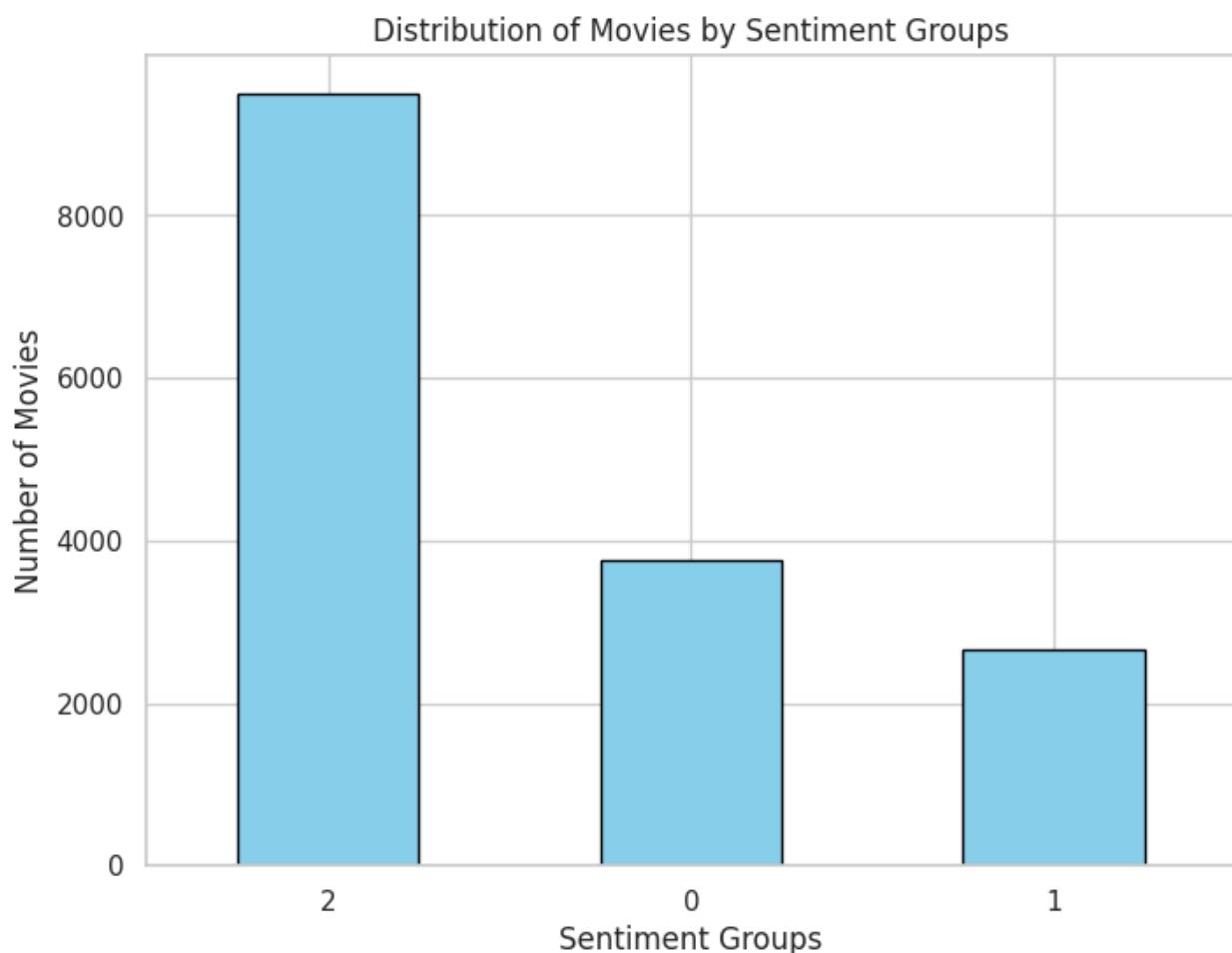
➔ 1 -0.271971

➔ 2 0.028858

Qua phân tích, ta có được kết quả sau:

- Cụm 0: Các phim có cảm xúc trung bình dương (0.32).
- Cụm 1: Các phim có cảm xúc trung bình âm (-0.27).
- Cụm 2: Các phim có cảm xúc trung bình gần như trung lập (0.03).

Thông qua việc sử dụng **KMeans** và phương pháp **elbow**, chúng ta đã phân nhóm các bộ phim thành 3 cụm cảm xúc, với các chỉ số đánh giá như **Silhouette Score** và **Davies-Bouldin Index** cho thấy phân nhóm này hợp lý. Cụ thể, mỗi cụm phản ánh một mức độ cảm xúc khác nhau của các bộ phim, giúp chúng ta hiểu rõ hơn về xu hướng cảm xúc của các bộ phim trong tập dữ liệu.



Hình 4.9 Phân phối của các cụm của sentiment_cluster

4.4.9 Cột status

```
df_rate['status'].value_counts(normalize=True) * 100
```

➔ Proportion of status: 100% Released

Do cột **status** chỉ chứa một giá trị duy nhất là '**Released**' và không cung cấp thông tin phân loại

hữu ích nên sẽ loại bỏ cột này khỏi dữ liệu.

4.4.10 Cột adult

Cột **adult** trong bộ dữ liệu thể hiện thông tin về độ tuổi phù hợp của bộ phim, với giá trị **True** cho các bộ phim dành cho người lớn và **False** cho các bộ phim không dành cho người lớn.

Cột **adult** chứa các giá trị **True** hoặc **False**. Để dễ dàng sử dụng trong mô hình học máy, giá trị này cần được chuyển đổi thành **1** (cho phim dành cho người lớn) và **0** (cho phim không dành cho người lớn).

```
df_rate['adult'] = df_rate['adult'].map(lambda x: 1 if (x == True) else 0)
print(df_rate['adult'].value_counts())
→ 0    15894
→ 1     31
```

4.4.11 Cột production_countries

Cột **production_countries** có thể chứa giá trị thiếu (NaN) và danh sách các quốc gia tham gia sản xuất phim. Các giá trị này cần được xử lý để có thể sử dụng trong mô hình học máy.

Đầu tiên, kiểm tra số lượng giá trị thiếu trong cột **production_countries**. Sau đó, thay thế các giá trị thiếu bằng chuỗi 'Unknown' để giữ cho bộ dữ liệu không bị thiếu giá trị

```
df_rate['production_countries'] =
df_rate['production_countries'].fillna('Unknown')
```

Tiếp theo, chuyển mỗi giá trị trong cột **production_countries** thành một danh sách các quốc gia bằng cách tách chuỗi bằng dấu phẩy.

```
df_rate['re_production_countries'] = df_rate['production_countries'].apply(lambda x: [genre.strip() for genre in x.split(',')])
```

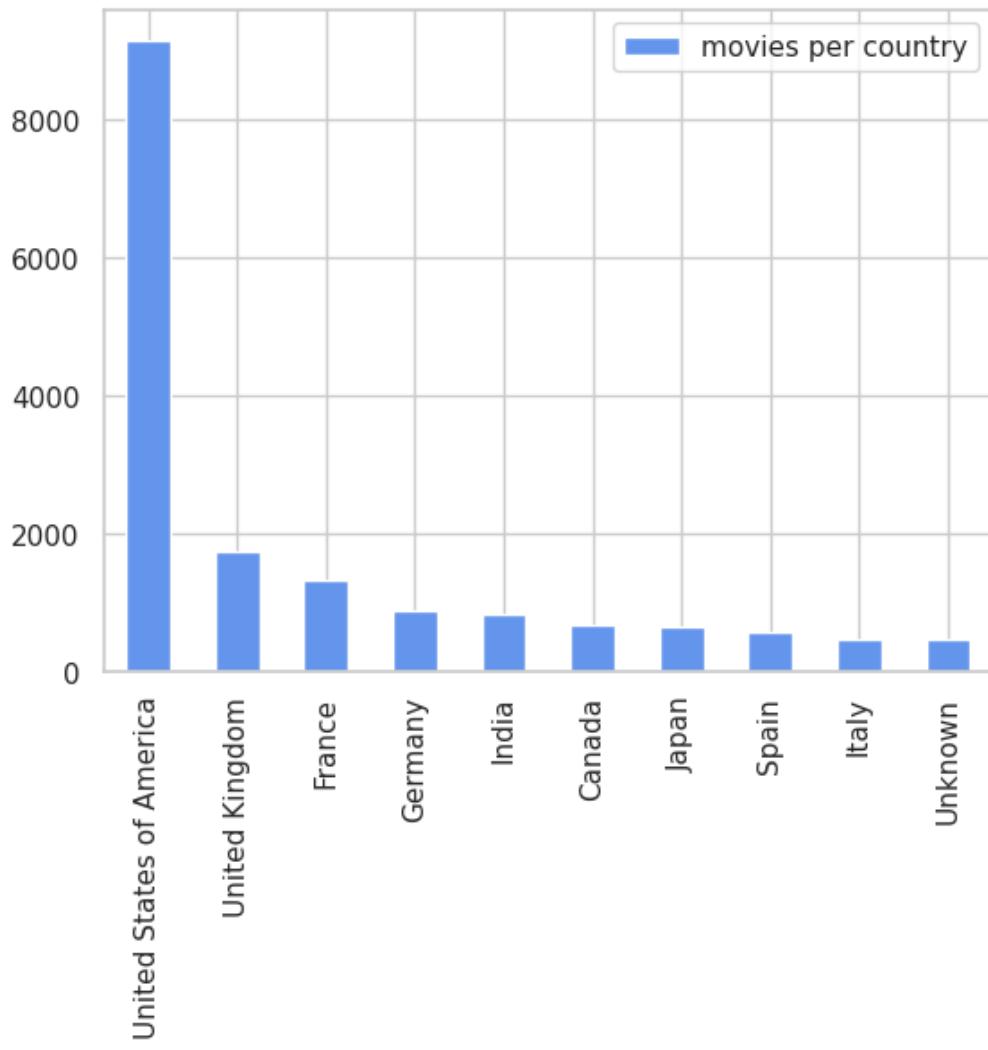
Sử dụng một từ điển **countriesDict** để đếm số lượng phim được sản xuất tại từng quốc gia. Mỗi quốc gia sẽ là một khóa trong từ điển, và giá trị của khóa sẽ là số lượng phim sản xuất tại quốc gia đó. Sau đó, chuyển từ điển **countriesDict** thành một DataFrame để dễ dàng phân tích và trực quan hóa.

```
countriesDict = {}
for element in df_rate["re_production_countries"].values:
    for country in element:
        if country not in countriesDict:
            countriesDict[country] = 1
        else:
            countriesDict[country] += 1

countries_train = pd.DataFrame.from_dict(countriesDict, orient='index',
columns=["movies per country"])
```

Kho dữ liệu và OLAP - IS217.P12

```
countries_train.sort_values(by="movies per country",  
ascending=False).head(10).plot.bar(color='cornflowerblue')
```



Hình 4.10 Biểu đồ thê hiện số lượng phim theo quốc gia

Tiếp theo, tạo cột **num_production_countries** để đếm số lượng quốc gia sản xuất của từng bộ phim. Nếu cột **production_countries** chứa giá trị **None**, thì sẽ gán số quốc gia sản xuất là **0**.

```
df_rate['num_production_countries'] =  
df_rate['production_countries'].apply(lambda x: len(x) if x[0] != 'None' else  
0)
```

Tạo một cột **is_produced_in_US** để kiểm tra xem bộ phim có được sản xuất ở Mỹ hay không, dựa trên việc xem liệu '**United States of America**' có xuất hiện trong danh sách các quốc gia hay không.

```
df_rate['is_produced_in_US'] = df_rate['production_countries'].apply(lambda x:  
'United States of America' in x)  
df_rate['is_produced_in_US'] = df_rate['is_produced_in_US'].astype(int)
```

4.4.12 Cột spoken_languages

Cột **spoken_languages** chứa danh sách các ngôn ngữ được sử dụng trong phim.

Giá trị thiếu trong cột **spoken_languages** được thay thế bằng một chuỗi rỗng (""), để đảm bảo dữ liệu không gây lỗi khi thực hiện các bước xử lý tiếp theo.

```
df_rate['spoken_languages'] = df_rate['spoken_languages'].fillna('')
```

Cột **spoken_languages** chứa các ngôn ngữ dưới dạng danh sách chuỗi phân tách bởi dấu phẩy.

Sử dụng phương pháp `.apply()`, chúng ta chuyển đổi chuỗi thành danh sách các ngôn ngữ và loại bỏ khoảng trắng thừa.

```
df_rate['re_spoken_languages'] = df_rate['spoken_languages'].apply(lambda x: [genre.strip() for genre in x.split(',')])
```

Một từ điển (`languagesDict`) được tạo ra để đếm số lượng phim sử dụng mỗi ngôn ngữ.

```
languagesDict = {}
for element in df_rate["re_spoken_languages"].values:
    for name in element:
        if name not in languagesDict:
            languagesDict[name] = 1
        else:
            languagesDict[name] += 1

sns.set(rc={'figure.figsize':(12,6)})

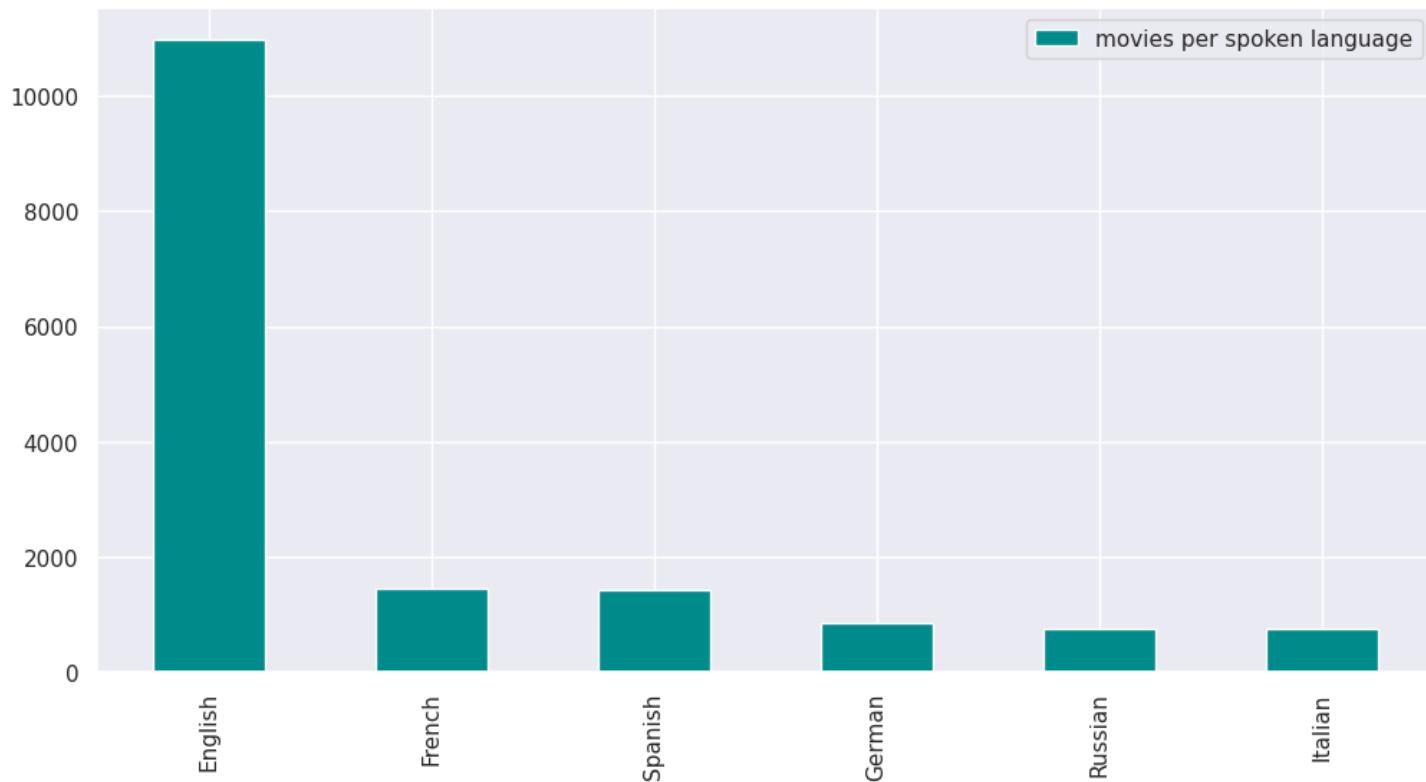
languages_train = pd.DataFrame.from_dict(languagesDict, orient='index',
columns=["movies per spoken language"])
languages_train.sort_values(by="movies per spoken language",
ascending=False).head(6).plot.bar(color='darkcyan')
languages_train.columns = ["number_of_languages"]
```

Số lượng ngôn ngữ sử dụng trong mỗi phim - Một cột mới, **num_spoken_languages**, được thêm vào để đếm số lượng ngôn ngữ trong mỗi phim. - Nếu danh sách ngôn ngữ rỗng, giá trị sẽ được đặt là 0.

```
df_rate['num_spoken_languages'] = df_rate.re_spoken_languages.apply(lambda x: len(x) if x[0] != '' else 0)
```

Một cột nhị phân mới, **is_spoken_in_English**, được tạo ra để đánh dấu phim có sử dụng tiếng Anh. Nếu English xuất hiện trong danh sách ngôn ngữ, giá trị sẽ là 1, ngược lại là 0.

```
df_rate['is_spoken_in_English'] = df_rate['spoken_languages'].apply(lambda x: 1 if 'English' in x else 0)
```



Hình 4.11 Biểu đồ thể hiện số lượng phim theo ngôn ngữ được dùng trong phim

4.4.13 Cột original_language

Cột **original_language** biểu thị ngôn ngữ gốc của từng bộ phim.

Sử dụng phương pháp `.value_counts()` để đếm số lượng phim theo từng ngôn ngữ gốc và liệt kê 10 ngôn ngữ phổ biến nhất.

```
print("Counts of each original language:")
print(df_rate['original_language'].value_counts()[:10])
```

```
→ en 10497
→ fr 633
→ es 620
→ ja 464
→ ru 416
→ hi 394
→ zh 375
→ it 274
→ de 261
→ ko 243
```

Kết quả cho thấy ngôn ngữ **English (en)** chiếm ưu thế với **10,497** phim, tiếp theo là **French (fr)**, **Spanish (es)**, và các ngôn ngữ khác.

Một ngưỡng (**language_threshold**) được đặt để xác định các ngôn ngữ phổ biến, với giá trị mặc định là 500. Các ngôn ngữ xuất hiện ít hơn ngưỡng này sẽ không được gán mã hóa One-Hot.

```
language_threshold = 500

popular_languages = df_rate['original_language'].value_counts()
popular_languages = popular_languages[popular_languages >=
language_threshold].index.tolist()
```

Áp dụng **One-Hot Encoding** cho cột **original_language**, trong đó mỗi ngôn ngữ phổ biến được biểu thị bằng một cột nhị phân riêng. Ví dụ: Nếu bộ phim có ngôn ngữ gốc là English, giá trị của cột lang_en sẽ là 1, và các cột ngôn ngữ khác sẽ là 0.

```
one_hot_encoded = pd.get_dummies(df_rate['original_language'], prefix='lang')

one_hot_encoded = one_hot_encoded[['lang_' + lang for lang in popular_languages
if 'lang_' + lang in one_hot_encoded.columns]].astype(int)

df_rate = pd.concat([df_rate, one_hot_encoded], axis=1)
```

4.4.14 Cột release_date

Thuộc tính **release_date** (ngày phát hành) là một trong những yếu tố quan trọng ảnh hưởng đến doanh thu và lượt đánh giá phim. Việc trích xuất thông tin từ thuộc tính này sẽ cung cấp nhiều khía cạnh hữu ích như:

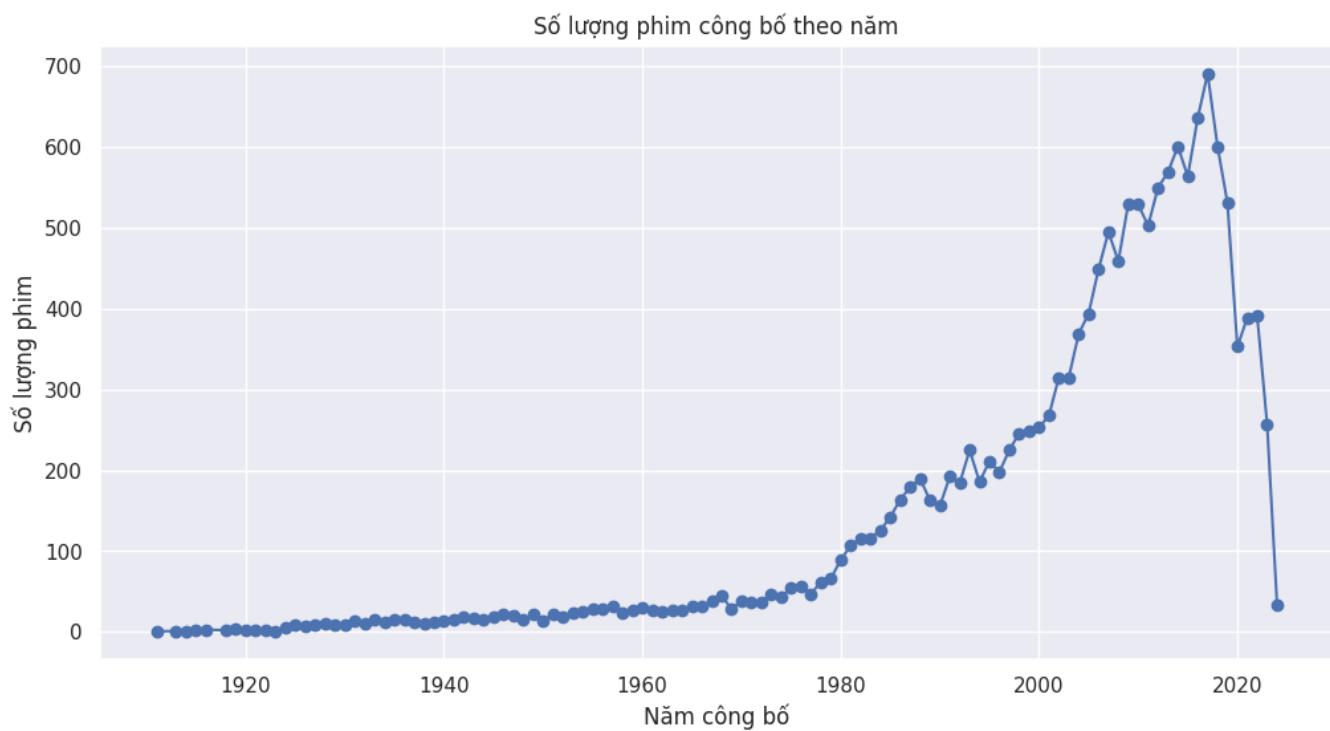
- Năm phát hành.
- Ngày trong tuần, tháng trong năm.
- Mùa phát hành.
- Thập kỷ phát hành.
- Mối quan hệ giữa ngày phát hành và các yếu tố khác như doanh thu hoặc lượt đánh giá.

Sử dụng **.value_counts()** để tính số lượng phim phát hành theo từng năm. Dùng **dt.weekday** để trích xuất ngày trong tuần (0: Monday, ..., 6: Sunday) và **dt.month** trích xuất tháng phát hành.

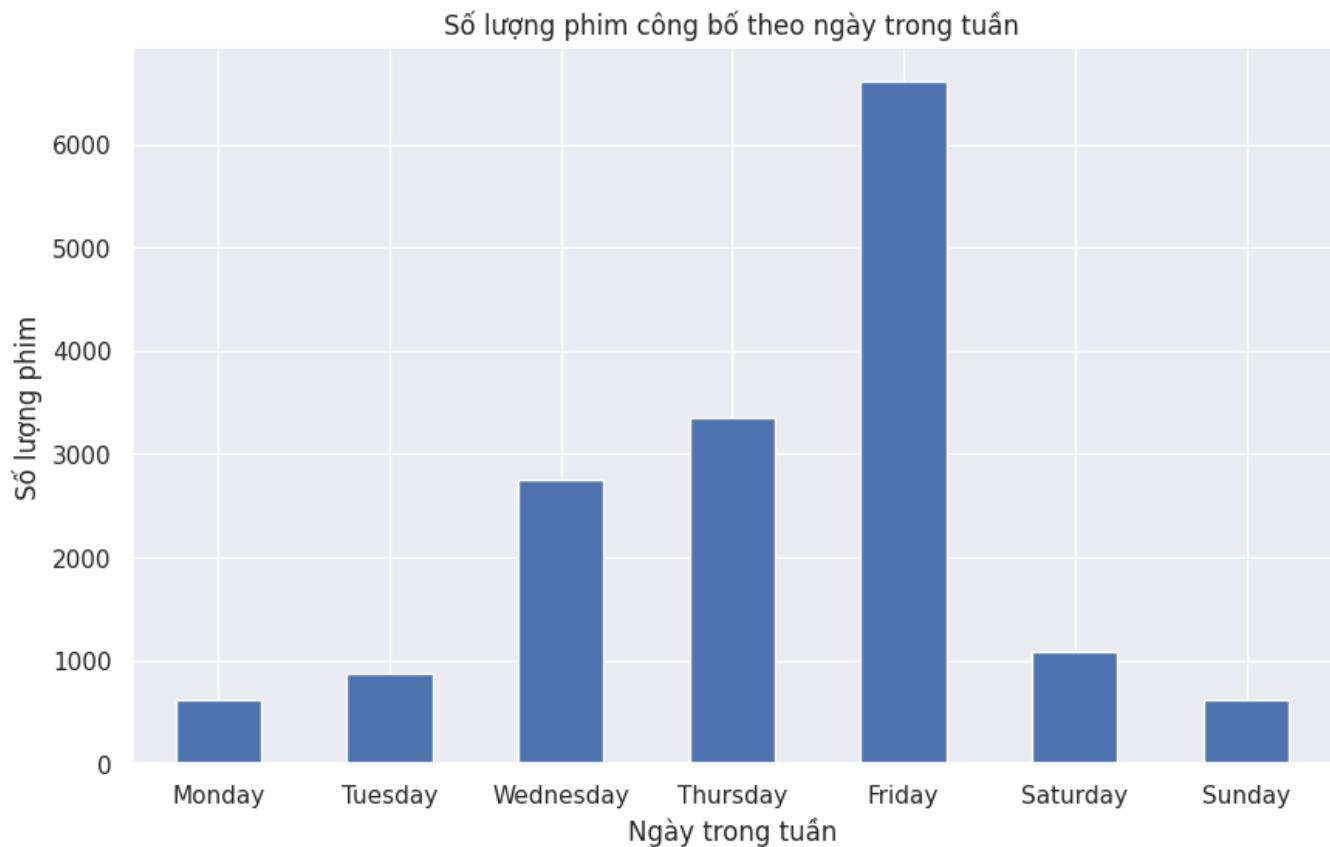
```
release_year_counts = df_rate['release_year'].value_counts().sort_index()

df_rate['release_weekday'] = df_rate['release_date'].dt.weekday

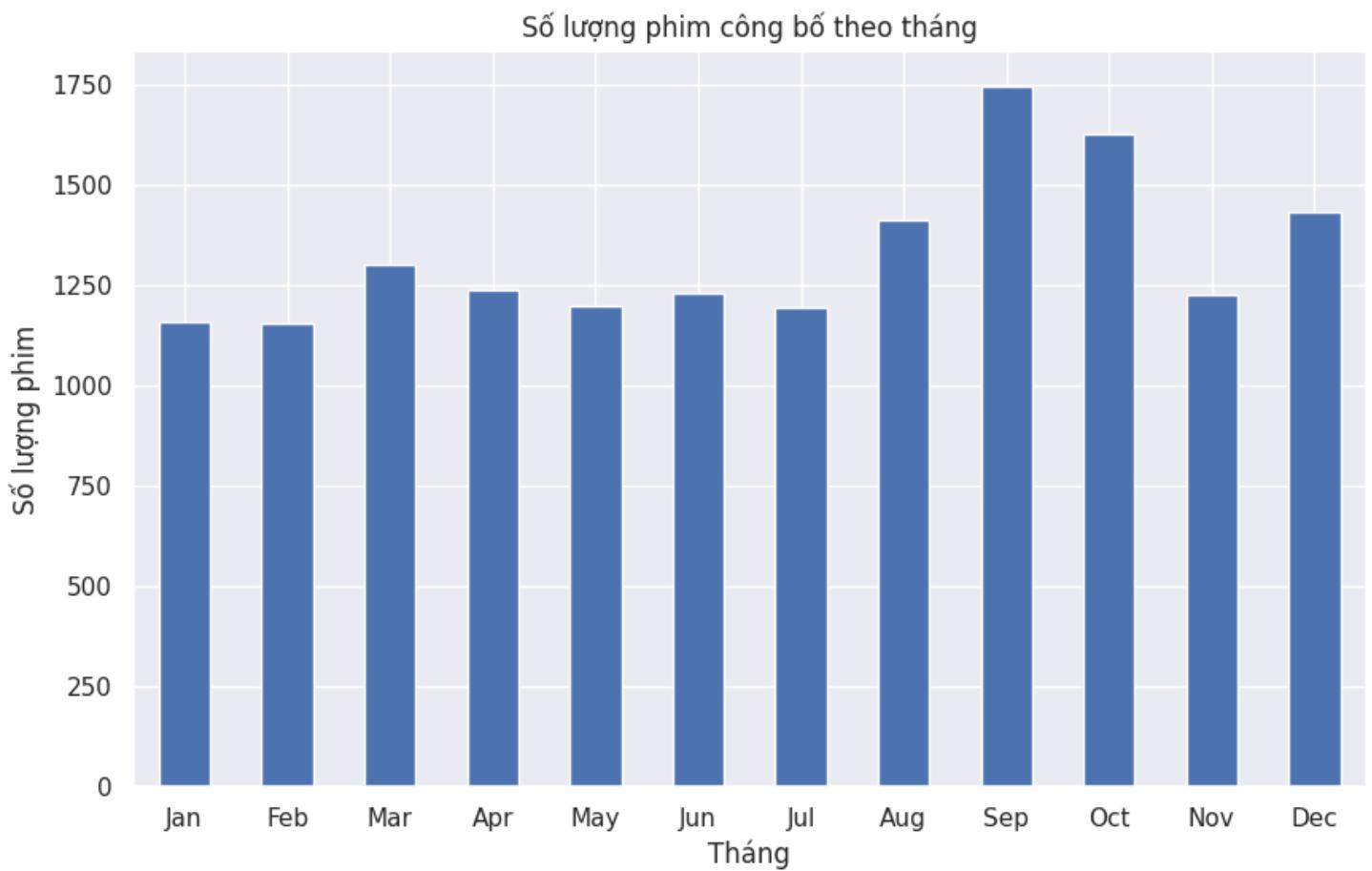
df_rate['release_month'] = df_rate['release_date'].dt.month
```



Hình 4.12 Biểu đồ thể hiện số lượng phim phát hành mỗi năm



Hình 4.13 Biểu đồ thể hiện số lượng phim phát hành theo ngày trong tuần



Hình 4.14 Biểu đồ thể hiện số lượng phim phát hành theo tháng

Dữ liệu cho thấy xu hướng tăng mạnh về số lượng phim phát hành trong các năm gần đây.

Một số giai đoạn có sự giảm sút, có thể do yếu tố lịch sử hoặc công nghệ.

Biểu đồ cột cho thấy xu hướng phát hành phim theo **từng ngày**: **Thứ Sáu** và **Thứ Bảy** là hai ngày phổ biến nhất để phát hành phim. Điều này phản ánh thói quen tiêu dùng, khi khán giả có xu hướng xem phim vào cuối tuần.

Biểu đồ cột hiển thị phân bố số lượng phim theo **từng tháng**: **Tháng 7** và **Tháng 12** là thời điểm cao điểm phát hành, có thể do trùng với các kỳ nghỉ lớn (mùa hè và lễ cuối năm).

Tạo các đặc trưng về **Mùa**, **Thập kỷ** và **Ngày trong tuần**:

```
def convert_decade(x):
    try:
        return str(int(str(x)[:4])) // 10) + '0s'
    except:
        return None
df_rate['release_decade'] = df_rate['release_year'].apply(convert_decade)
```

```

def get_season(month):
    if month in [3, 4, 5]:
        return 'Spring'
    elif month in [6, 7, 8]:
        return 'Summer'
    elif month in [9, 10, 11]:
        return 'Fall'
    else:
        return 'Winter'

df_rate['season'] = df_rate['release_date'].dt.month.apply(get_season)

df_rate['release_weekday'] = df_rate['release_weekday'].map({
0: 'Monday', 1: 'Tuesday', 2: 'Wednesday',
3: 'Thursday', 4: 'Friday', 5: 'Saturday', 6: 'Sunday'})

df_rate = pd.get_dummies(df_rate, columns=['release_weekday'],
prefix='release', drop_first=False)

```

4.4.15 Cột genres_list

Cột genres_list chứa thông tin về các thể loại phim dưới dạng danh sách. Việc xử lý và tạo đặc trưng từ cột này nhằm chuyển đổi dữ liệu dạng văn bản thành dạng số, giúp mô hình máy học dễ dàng khai thác thông tin và cải thiện hiệu suất.

Tạo đặc trưng mới - **num_genres_list**, mục tiêu là đếm số lượng thể loại được liệt kê trong cột genres_list. Nếu danh sách không rỗng, số lượng thể loại sẽ được tính bằng hàm len(x). Còn danh sách rỗng, giá trị sẽ được gán là 0.

```

df_rate['num_genres_list'] = df_rate.genres_list.apply(lambda x: len(x) if
x[0] != '' else 0)

```

Chuyển đổi dữ liệu văn bản trong cột genres_list thành các đặc trưng số, sử dụng kỹ thuật Bag of Words với CountVectorizer:

- **Làm sạch dữ liệu:**
 - Dữ liệu trong **genres_list** được chuyển đổi thành danh sách các từ.
 - Các ký tự đặc biệt (như /, +, &, -, !) được thay thế bằng dấu gạch dưới (_), giúp chuẩn hóa dữ liệu.
- **Ghép chuỗi:** Sau khi chuẩn hóa, các từ trong danh sách được ghép lại thành một chuỗi duy nhất cho mỗi dòng.
- **Vector hóa:**
 - Sử dụng **CountVectorizer**, các thể loại phổ biến nhất (xuất hiện ít nhất min_df lần,

trong trường hợp này là 200 lần) được chọn làm đặc trưng.

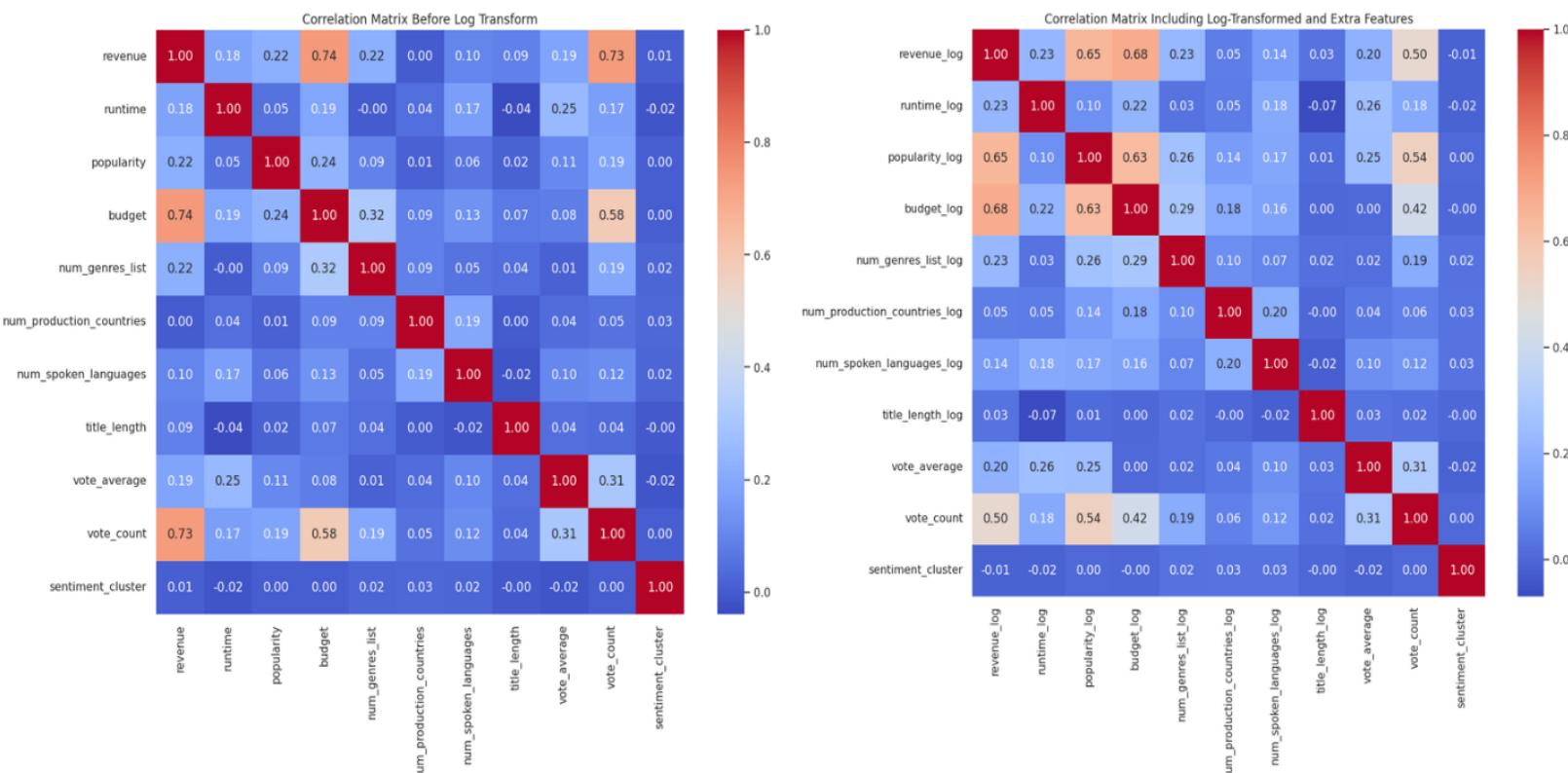
- Kết quả vector hóa là một DataFrame với các cột đặc trưng mới, mỗi cột đại diện cho một thể loại phim.
- **Gộp dữ liệu:** Các cột vector hóa được thêm vào DataFrame gốc, sau đó cột **genres_list** gốc bị loại bỏ.

4.4.16 Xử lý các đặc trưng số (Numerical Feature Transformation)

Đối với các đặc trưng số, áp dụng **log transformation** có thể giúp cải thiện hiệu suất của mô hình học máy, đặc biệt là khi dữ liệu có sự phân phối không đồng đều (**skewed distribution**). Các giá trị có phạm vi lớn, chẳng hạn như doanh thu (revenue), ngân sách (budget), hay độ phổ biến (popularity), có thể gây khó khăn cho mô hình.

Log transformation giúp làm giảm độ lệch (**skewness**) của các dữ liệu này và giúp mô hình dễ dàng học được các mối quan hệ hơn.

Các đặc trưng cần biến đổi: Dữ liệu số như **revenue**, **runtime**, **popularity**, **budget**, **num_genres_list**, **num_production_countries**, **num_spoken_languages**, và **title_length** sẽ được áp dụng biến đổi log để chuẩn hóa các phân phối.



Hình 4.15 So sánh giữa các đặc trưng số trước và sau khi log transformation

4.4.17 Lựa chọn các đặc trưng (Feature Selection)

4.4.17.1 Phân loại đặc trưng

Các cột không có tác dụng trực tiếp trong việc xây dựng các mô hình hoặc không liên quan đến phân tích dữ liệu đã được loại bỏ để làm sạch dữ liệu. Cụ thể, chúng ta loại bỏ các cột như: 'title', 'original_language', 'production_countries', 'spoken_languages', 'release_year', 'overview_sentiment', 're_production_countries', và 're_spoken_languages'.

Cột **vote_average** trong bộ dữ liệu thể hiện điểm đánh giá trung bình của phim, nhưng giá trị này có thể rất khó sử dụng trực tiếp trong các mô hình học máy, nhất là khi mô hình yêu cầu các đặc trưng phân loại (**categorical features**). Để giải quyết vấn đề này, chúng ta có thể chuyển đổi điểm đánh giá thành các nhóm phân loại để mô hình có thể dễ dàng học được.

Phim sẽ được phân loại thành ba nhóm dựa trên giá trị của vote_average:

- **Hit:** Các phim có đánh giá từ 8 trở lên.
- **Average:** Các phim có đánh giá từ 5 đến dưới 8.
- **Flop:** Các phim có đánh giá dưới 5.

Phân loại này giúp mô hình nhận diện các nhóm phim khác nhau dễ dàng hơn, và có thể tạo ra một ảnh hưởng tích cực trong việc dự đoán các yếu tố liên quan đến thành công của phim.

Sau khi thực hiện xử lý đặc trưng thì chúng ta có được tổng cộng **9** đặc trưng số và **49** đặc trưng phân loại.

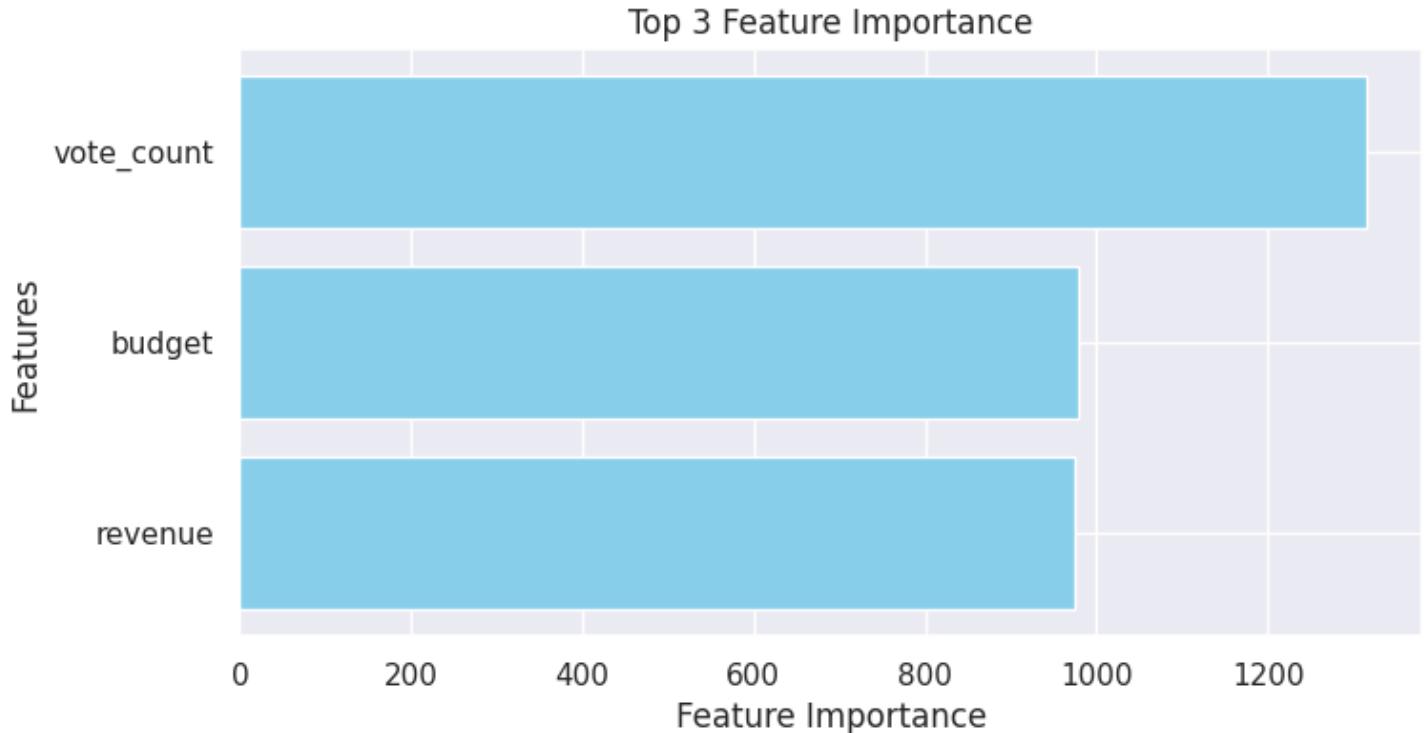
Các đặc trưng phân loại (**categorical features**): ['adult', 'sentiment_cluster', 'is_produced_in_US', 'is_spoken_in_English', 'lang_en', 'lang_fr', 'lang_es', 'release_month', 'release_Friday', 'release_Monday', 'release_Saturday', 'release_Sunday', 'release_Thursday', 'release_Tuesday', 'release_Wednesday', 'release_1910s', 'release_1920s', 'release_1930s', 'release_1940s', 'release_1950s', 'release_1960s', 'release_1970s', 'release_1980s', 'release_1990s', 'release_2000s', 'release_2010s', 'release_2020s', 'season_Fall', 'season_Spring', 'season_Summer', 'season_Winter', 'genres_list_Action', 'genres_list_Adventure', 'genres_list_Animation', 'genres_list_Comedy', 'genres_list_Crime', 'genres_list_Documentary', 'genres_list_Drama', 'genres_list_Family', 'genres_list_Fantasy', 'genres_list_History', 'genres_list_Horror', 'genres_list_Music', 'genres_list_Mystery', 'genres_list_Romance', 'genres_list_Science_Fiction', 'genres_list_Thriller', 'genres_list_War', 'genres_list_Western'].

Các đặc trưng số (**numerical features**): ['vote_count', 'revenue_log', 'runtime_log', 'popularity_log', 'budget_log', 'num_genres_list_log', 'num_production_countries_log', 'num_spoken_languages_log', 'title_length_log'].

Cột mục tiêu là [vote_average_category].

4.4.17.2 Lựa chọn đặc trưng

Từ các đặc trưng đã tìm và phân loại được ở quá trình xử lý đặc trưng. Nhóm tiến hành sử dụng mô hình **Tree-based Models** là **LightGBM Classifier** để xác định 3 đặc trưng quan trọng nhất để sử dụng cho việc xây dựng mô hình phân loại.



Hình 4.16 Top 3 features

Kết quả cho thấy 3 đặc trưng quan trọng nhất là “vote_count”, “budget” và “revenue”. Nhóm tiến hành xây dựng và đánh giá mô hình phân loại mức độ hài lòng của khán giả với phim điện ảnh dựa trên các đặc trưng:

- **vote_count**: Số lượt đánh giá.
- **budget**: Ngân sách làm phim.
- **revenue**: Doanh thu phim.

4.5 Tiết xử lý dữ liệu

Bước 1: Thiết lập:

- Đầu vào (X): Gồm 3 đặc trưng: vote_count, budget, revenue.
- Đầu ra (y): Phân loại mức độ hài lòng của khán giả (vote_average_category).

Bước 2: Chia tập dữ liệu:

```
x_train, x_temp, y_train, y_temp = train_test_split(X, y, test_size=0.2,
random_state=42)
x_test, x_valid, y_test, y_valid = train_test_split(x_temp, y_temp,
test_size=0.5, random_state=42)
```

Tập huấn luyện (X_{train} , y_{train}): 80% dữ liệu, dùng để huấn luyện mô hình.

Tập kiểm định (X_{valid} , y_{valid}): 10% dữ liệu, dùng để tối ưu hóa tham số và tránh overfitting.

Tập kiểm tra (X_{test} , y_{test}): 10% dữ liệu, dùng để đánh giá hiệu suất cuối cùng của mô hình.

Bước 3: Chuẩn hóa dữ liệu:

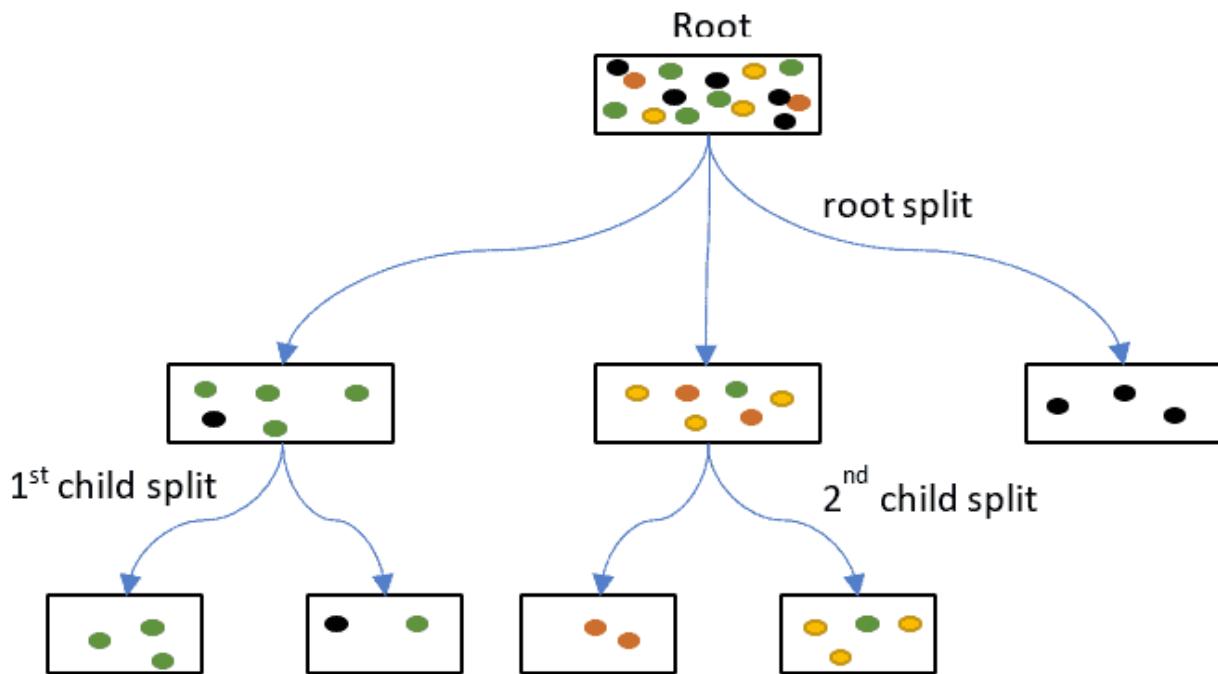
```
scaler = RobustScaler()
X_train = scaler.fit_transform(X_train)
X_valid = scaler.transform(X_valid)
X_test = scaler.transform(X_test)
```

RobustScaler được lựa chọn vì khả năng xử lý tốt dữ liệu có **outliers** (giá trị ngoại lai), tránh ảnh hưởng lớn đến kết quả chuẩn hóa.

4.6 Mô hình máy học phân loại

4.6.1 Catboost Classifier

4.6.1.1 Tổng quan về Catboost Classifier



CatBoost là một thuật toán học máy mạnh mẽ, thuộc họ Boosting, được sử dụng cho cả bài toán hồi quy và phân loại. CatBoost xây dựng các mô hình dựa trên việc kết hợp nhiều cây quyết định để tối ưu hóa hiệu quả dự đoán.

Ưu điểm:

- Giảm hiện tượng overfitting.
- Hiệu suất cao.
- Làm việc với dữ liệu không cân bằng.

Nhược điểm:

- Cần nhiều tài nguyên.
- Độ phức tạp thuật toán.
- Cần chọn siêu tham số cẩn thận.

4.6.1.2 Xây dựng mô hình Catboost Classifier

Bước 1: Mô hình Catboost cơ bản.

```
cb_model = CatBoostClassifier(verbose=0)
cb_model.fit(X_train, y_train)
```

Bước 2: Tìm tham số tối ưu với Grid Search.

```
param_grid_cb = {
    'iterations': [100, 200, 300],
    'learning_rate': [0.01, 0.05, 0.1],
    'depth': [4, 6, 8] }
```

Các tham số được xem xét bao gồm:

- **iterations:** Số vòng lặp huấn luyện, với các giá trị [100, 200, 300].
- **learning_rate:** Tốc độ học, với các giá trị [0.01, 0.05, 0.1].
- **depth:** Độ sâu của cây, với các giá trị [4, 6, 8].

```
grid_cb = GridSearchCV(estimator=CatBoostClassifier(silent=True,
random_state=42), param_grid=param_grid_cb, cv=5, scoring='accuracy')
grid_cb.fit(X_train, y_train)
```

Sử dụng lớp GridSearchCV để tìm tổ hợp tham số tốt nhất từ param_grid_cb.

Bước 3: Đánh giá mô hình Catboost với Cross-Validation.

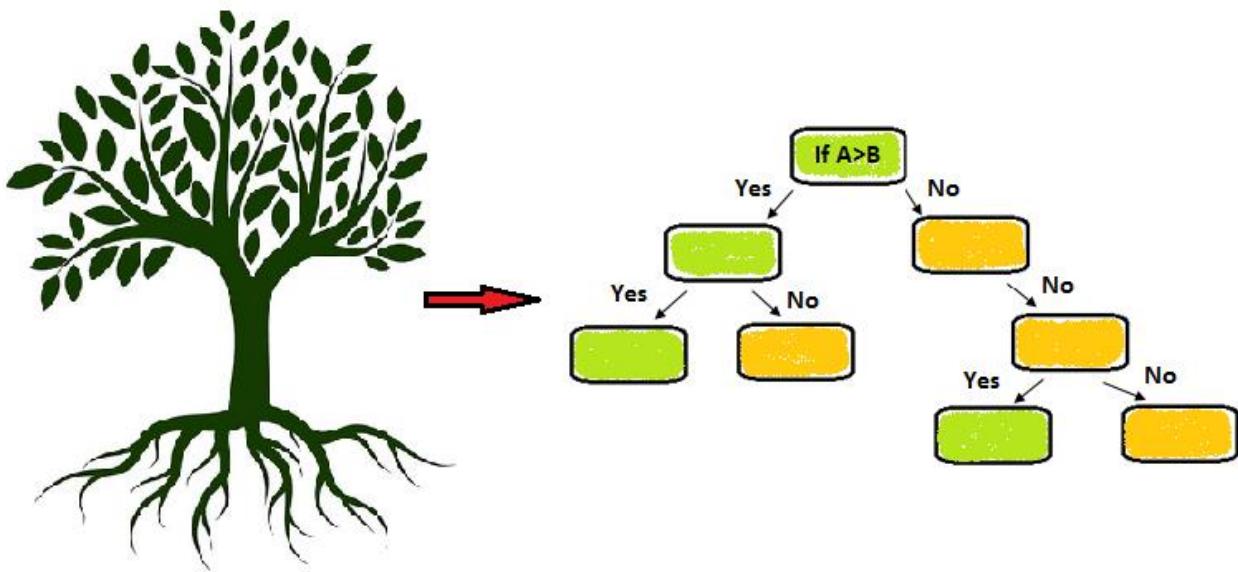
```
pipeline_tree = Pipeline([
    ('estimator', grid_cb.best_estimator_)
])

cv = KFold(n_splits=10, shuffle=True, random_state=42)
scores_tree = cross_val_score(pipeline_tree, X, y, cv=cv, scoring='accuracy')
```

Sử dụng Cross-Validation giúp đánh giá mô hình một cách tổng quan, tránh hiện tượng overfitting hoặc underfitting.

4.6.2 Decision Tree Classifier

4.6.2.1 Tổng quan về Decision Tree Classifier



Cây quyết định là một mô hình dự đoán được biểu diễn dưới dạng cây, được dùng trong bài toán hồi quy và phân loại.

Mỗi nút trong cây đại diện cho một đặc trưng (feature), mỗi nhánh đại diện cho một giá trị của đặc trưng đó, và mỗi lá (leaf) đại diện cho một nhãn (label) hoặc giá trị đầu ra.

Ưu điểm:

- Dễ hiểu, trực quan.
- Không cần xử lý trước nhiều dữ liệu.
- Cây quyết định có thể xử lý các loại dữ liệu khác nhau, bao gồm cả dữ liệu số và dữ liệu dạng chữ.

Nhược điểm:

- Dễ bị overfitting.
- Độ nhạy cao với thay đổi nhỏ trong dữ liệu.
- Cây quyết định có thể gặp khó khăn trong việc nắm bắt và biểu diễn các mối quan hệ phức tạp và phi tuyến tính giữa các đặc điểm.

4.6.2.2 Xây dựng mô hình Decision Tree Classifier

Bước 1: Mô hình Decision Tree cơ bản.

```
dt_model = DecisionTreeClassifier()  
dt_model.fit(X_train, y_train)
```

Bước 2: Tìm tham số tối ưu với Grid Search.

```
param_grid_tree = {  
    'criterion': ['gini', 'entropy'],  
    'max_depth': [5, 10, 20, 30],  
    'min_samples_split': [1, 2, 5, 10],  
    'min_samples_leaf': [1, 2, 4, 10]}
```

Các tham số được xem xét bao gồm:

- **criterion:** Hàm đánh giá chất lượng của việc phân chia dữ liệu, với 2 lựa chọn learning_rate:
 - 'gini': Chỉ số Gini.
 - 'entropy': Entropy thông tin.
- **max_depth:** Độ sâu tối đa của cây, với các giá trị [5, 10, 20, 30].
- **min_samples_split:** Số lượng mẫu tối thiểu cần thiết để chia một nút, với các giá trị [1, 2, 5, 10].
- **min_samples_leaf:** Số lượng mẫu tối thiểu tại một lá, với các giá trị [1, 2, 4, 10].

```
grid_tree = GridSearchCV(DecisionTreeClassifier(random_state=42),  
param_grid_tree, cv=5, scoring='accuracy')  
grid_tree.fit(X_train, y_train)
```

Sử dụng lớp GridSearchCV để tìm tổ hợp tham số tốt nhất từ param_grid_tree.

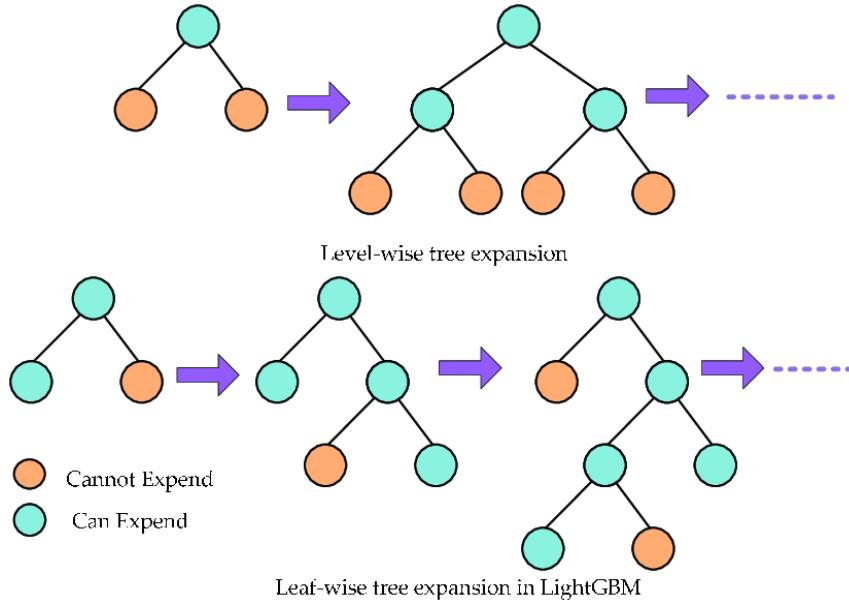
Bước 3: Đánh giá mô hình Decision Tree với Cross-Validation.

```
pipeline_tree = Pipeline([  
    ('estimator', grid_tree.best_estimator_)])  
  
cv = KFold(n_splits=10, shuffle=True, random_state=42)  
scores_tree = cross_val_score(pipeline_tree, X, y, cv=cv, scoring='accuracy')
```

Sử dụng Cross-Validation giúp đánh giá mô hình một cách tổng quan, tránh hiện tượng overfitting hoặc underfitting.

4.6.3 LightGBM Classifier

4.6.3.1 Tổng quan về LightGBM Classifier



LightGBM (Light Gradient Boosting Machine) là một thuật toán học máy dựa trên phương pháp boosting, được thiết kế để tối ưu hóa tốc độ và hiệu quả bộ nhớ.

Sử dụng Gradient Boosting: Cải tiến so với các phiên bản Gradient Boosting khác với nhiều tối ưu hóa về tốc độ và bộ nhớ.

Ưu điểm:

- Tốc độ và hiệu quả bộ nhớ cao.
- Khả năng khai quát tốt.
- Xử lý dữ liệu lớn.
- Hỗ trợ nhiều hàm mục tiêu.

Nhược điểm:

- Cần nhiều tài nguyên.
- Độ phức tạp thuật toán.
- Nhạy cảm với dữ liệu nhiễu.

4.6.3.2 Xây dựng mô hình LightGBM Classifier

Bước 1: Mô hình LightGBM cơ bản.

```
lgb_model = LGBMClassifier(verbose=-1)
lgb_model.fit(X_train, y_train)
```

Bước 2: Tìm tham số tối ưu với Grid Search.

```
param_grid_lgbm = {  
    'n_estimators': [100, 200, 500],  
    'learning_rate': [0.01, 0.1, 0.2],  
    'max_depth': [10, 20, 30],  
    'num_leaves': [10, 20, 50, 100],  
    'min_split_gain': [0.05, 0.1, 1],  
    'colsample_bytree': [0.5, 0.8, 1.0],  
    'subsample': [0.5, 0.8, 1.0],  
    'min_child_samples': [80, 100, 150, 200] }
```

Các tham số được xem xét bao gồm:

- **n_estimators**: Số lượng cây trong mô hình (100, 200, 500).
- **learning_rate**: Tốc độ học, ảnh hưởng đến bước cập nhật của thuật toán (0.01, 0.1, 0.2).
- **max_depth**: Độ sâu tối đa của cây, giới hạn mức phân chia (10, 20, 30).
- **num_leaves**: Số lượng lá trong mỗi cây (10, 20, 50, 100).
- **min_split_gain**: Tăng cường tối thiểu cần thiết để phân chia một nút (0.05, 0.1, 1).
- **colsample_bytree**: Tỷ lệ cột (features) được chọn ngẫu nhiên cho mỗi cây (0.5, 0.8, 1.0).
- **subsample**: Tỷ lệ mẫu (samples) được chọn ngẫu nhiên cho mỗi cây (0.5, 0.8, 1.0).
- **min_child_samples**: Số lượng mẫu tối thiểu để phân chia một nút (80, 100, 150, 200).

```
grid_lgbm = GridSearchCV(  
    estimator=LGBMClassifier(random_state=42, verbose=-1),  
    param_grid=param_grid_lgbm,  
    cv=5,  
    scoring='accuracy',  
    n_jobs=-1  
)  
grid_lgbm.fit(X_train, y_train)
```

Sử dụng lớp GridSearchCV để tìm tổ hợp tham số tốt nhất từ param_grid_lgbm.

Bước 3: Đánh giá mô hình LightGBM với Cross-Validation.

```
pipeline_lgbm = Pipeline([  
    ('estimator', grid_lgbm.best_estimator_)  
)  
  
cv = KFold(n_splits=10, shuffle=True, random_state=42)  
scores_lgbm = cross_val_score(pipeline_lgbm, X, y, cv=cv, scoring='accuracy')
```

Sử dụng Cross-Validation giúp đánh giá mô hình một cách tổng quan, tránh hiện tượng overfitting hoặc underfitting.

4.6.4 Độ đo đánh giá

4.6.4.1 Accuracy

Accuracy là một phép đo quan trọng và phổ biến trong bài toán phân loại, giúp đánh giá hiệu suất của một mô hình dự đoán bằng cách tính toán tỷ lệ giữa số lượng dự đoán chính xác và tổng số lượng mẫu dữ liệu.

Công thức: Accuracy = (Số lượng dự đoán đúng) / (Tổng số dự đoán)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Ưu điểm:

- Dễ hiểu và dễ diễn giải: Accuracy là một metric đơn giản và trực quan, dễ dàng để hiểu và giải thích cho cả những người không chuyên về kỹ thuật.
- Phổ biến và được sử dụng rộng rãi: Do tính đơn giản và dễ hiểu, accuracy được sử dụng rộng rãi trong nhiều bài toán phân loại.

Nhược điểm:

- Không hiệu quả với tập dữ liệu không cân bằng: Nếu tập dữ liệu của bạn có sự mất cân bằng giữa các lớp (ví dụ: 90% dữ liệu thuộc lớp A và 10% dữ liệu thuộc lớp B), một mô hình luôn dự đoán lớp A sẽ có accuracy cao (90%) mặc dù nó không học được gì cả. Trong trường hợp này, accuracy không phản ánh đúng hiệu suất thực sự của mô hình.
- Không phân biệt giữa các loại lỗi: Accuracy chỉ quan tâm đến việc dự đoán đúng hay sai, mà không phân biệt giữa các loại lỗi. Ví dụ, trong bài toán phân loại ung thư, việc chẩn đoán nhầm một người khỏe mạnh là bị ung thư (false positive) có thể ít nghiêm trọng hơn việc chẩn đoán nhầm một người bị ung thư là khỏe mạnh (false negative). Accuracy không thể hiện được sự khác biệt này.

4.6.4.2 Precision

Precision là một metric được sử dụng để đánh giá hiệu suất của một mô hình phân loại, đặc biệt hữu ích khi chi phí của false positive (dương tính giả) cao. Nó tập trung vào việc đo lường, trong số những mẫu mà mô hình dự đoán là tích cực, có bao nhiêu mẫu thực sự là tích cực.

Công thức tính: Precision = (Số lượng true positive) / (Số lượng true positive + Số lượng false positive)

$$Precision = \frac{TP}{TP + FP}$$

Ưu điểm:

- Hữu ích khi chi phí của false positive cao: Trong ví dụ phát hiện spam, một false positive (phân loại nhầm email bình thường là spam) có thể khiến người dùng bỏ lỡ thông tin quan trọng. Precision giúp đánh giá khả năng mô hình tránh được những lỗi này.
- Tập trung vào độ chính xác của dự đoán tích cực: Precision chỉ quan tâm đến những mẫu được dự đoán là tích cực, giúp đánh giá độ tin cậy của những dự đoán này.

Nhược điểm:

- Không xem xét false negative: Precision không tính đến số lượng false negative (âm tính giả), tức là những mẫu thực sự là tích cực nhưng bị mô hình phân loại nhầm là âm tính. Trong một số trường hợp, false negative cũng có thể rất nghiêm trọng.
- Có thể bị ảnh hưởng bởi ngưỡng phân loại: Giá trị Precision có thể thay đổi tùy thuộc vào ngưỡng phân loại được sử dụng.

4.6.4.3 Recall

Recall, còn được gọi là Sensitivity hoặc True Positive Rate, là một metric được sử dụng để đánh giá hiệu suất của một mô hình phân loại. Nó đặc biệt hữu ích khi chi phí của false negative (âm tính giả) cao. Recall tập trung vào việc đo lường, trong số tất cả các mẫu thực sự là tích cực, có bao nhiêu mẫu được mô hình dự đoán đúng là tích cực.

Công thức tính: Recall = (Số lượng true positive) / (Số lượng true positive + Số lượng false negative)

$$Recall = \frac{TP}{TP + FN}$$

Ưu điểm:

- Hữu ích khi chi phí của false negative cao: Trong ví dụ chẩn đoán bệnh, một false negative (chẩn đoán nhầm người bệnh là khỏe mạnh) có thể dẫn đến việc trì hoãn điều trị và hậu quả nghiêm trọng. Recall giúp đánh giá khả năng mô hình tránh được những lỗi này.
- Tập trung vào việc phát hiện tất cả các mẫu tích cực: Recall đo lường khả năng mô hình tìm ra tất cả các trường hợp thực sự là tích cực, bất kể mô hình có dự đoán nhầm một số mẫu âm tính thành tích cực hay không.

Nhược điểm:

- Không xem xét false positive: Recall không tính đến số lượng false positive, tức là những mẫu thực sự là âm tính nhưng bị mô hình phân loại nhầm là tích cực.
- Có thể bị ảnh hưởng bởi ngưỡng phân loại: Giá trị Recall có thể thay đổi tùy thuộc vào ngưỡng phân loại được sử dụng.

4.6.4.4 F1-Score

F1-score là một metric được sử dụng để đánh giá hiệu suất của một mô hình phân loại. Nó là trung bình điều hòa giữa Precision (Độ chính xác) và Recall (Độ phủ), cung cấp một thước đo cân bằng hơn khi cả false positive và false negative đều quan trọng. F1-score đặc biệt hữu ích khi bạn cần tìm một sự cân bằng giữa Precision và Recall, và không muốn thiên vị về một metric nào.

Công thức tính: $F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Ưu điểm:

- Cân bằng giữa Precision và Recall: F1-score cung cấp một thước đo cân bằng, xem xét cả false positive và false negative.
- Hữu ích khi tập dữ liệu không cân bằng: Khi tỉ lệ giữa các lớp không đồng đều, F1-score cung cấp một đánh giá hiệu suất hơn so với Accuracy.
- Dễ dàng so sánh giữa các mô hình: F1-score là một giá trị đơn, dễ dàng so sánh hiệu suất giữa các mô hình khác nhau.

Nhược điểm:

- Khó diễn giải hơn Accuracy: F1-score phức tạp hơn Accuracy và có thể khó giải thích cho những người không chuyên về kỹ thuật.
- Không thể hiện được mức độ mất cân bằng giữa Precision và Recall: Hai mô hình có thể có cùng F1-score nhưng có sự phân bố Precision và Recall rất khác nhau.

4.6.4.5 K Fold cross-validation

K-Fold Cross-Validation là một kỹ thuật đánh giá mô hình được sử dụng rộng rãi trong học máy để ước lượng chính xác hơn hiệu suất của mô hình trên dữ liệu chưa được nhìn thấy. Nó giúp giảm thiểu sự phụ thuộc vào việc chia dữ liệu train/test cụ thể và cung cấp một đánh giá tổng quát hơn về khả năng tổng quát hóa của mô hình.

Cách thức hoạt động:

- Chia dữ liệu: Tập dữ liệu được chia thành K folds (phần) có kích thước bằng nhau.

- Huấn luyện và đánh giá: Mô hình được huấn luyện K lần. Trong mỗi lần, một fold được sử dụng làm tập kiểm tra (validation set), và K-1 folds còn lại được sử dụng làm tập huấn luyện (training set). Hiệu suất của mô hình được đánh giá trên fold kiểm tra.
- Trung bình kết quả: Sau K lần huấn luyện và đánh giá, kết quả (ví dụ: accuracy, F1-score) từ mỗi fold được trung bình lại để có được một ước lượng tổng quát về hiệu suất của mô hình.

Ưu điểm:

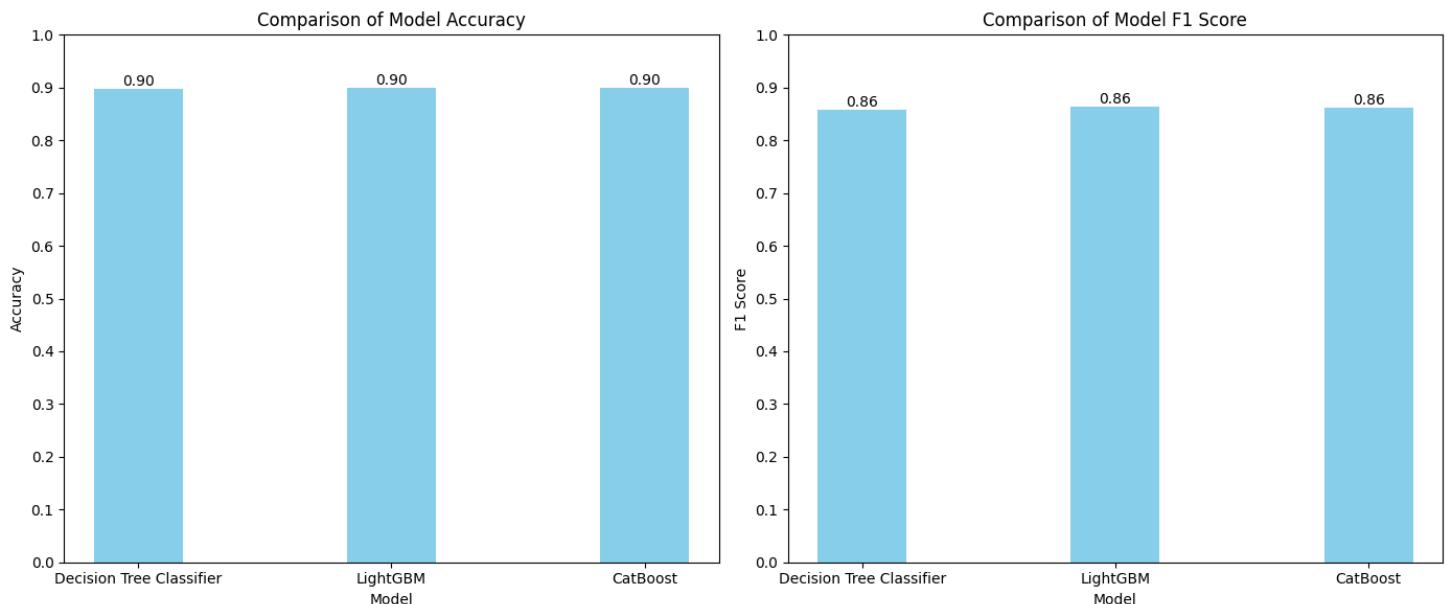
- Đánh giá chính xác hơn: Sử dụng toàn bộ dữ liệu cho cả huấn luyện và kiểm tra, giúp đánh giá hiệu suất mô hình một cách toàn diện hơn.
- Giảm thiểu bias: Giảm sự phụ thuộc vào cách chia dữ liệu train/test ngẫu nhiên.
- Tận dụng tối đa dữ liệu: Mọi điểm dữ liệu đều được sử dụng cho cả huấn luyện và kiểm tra.

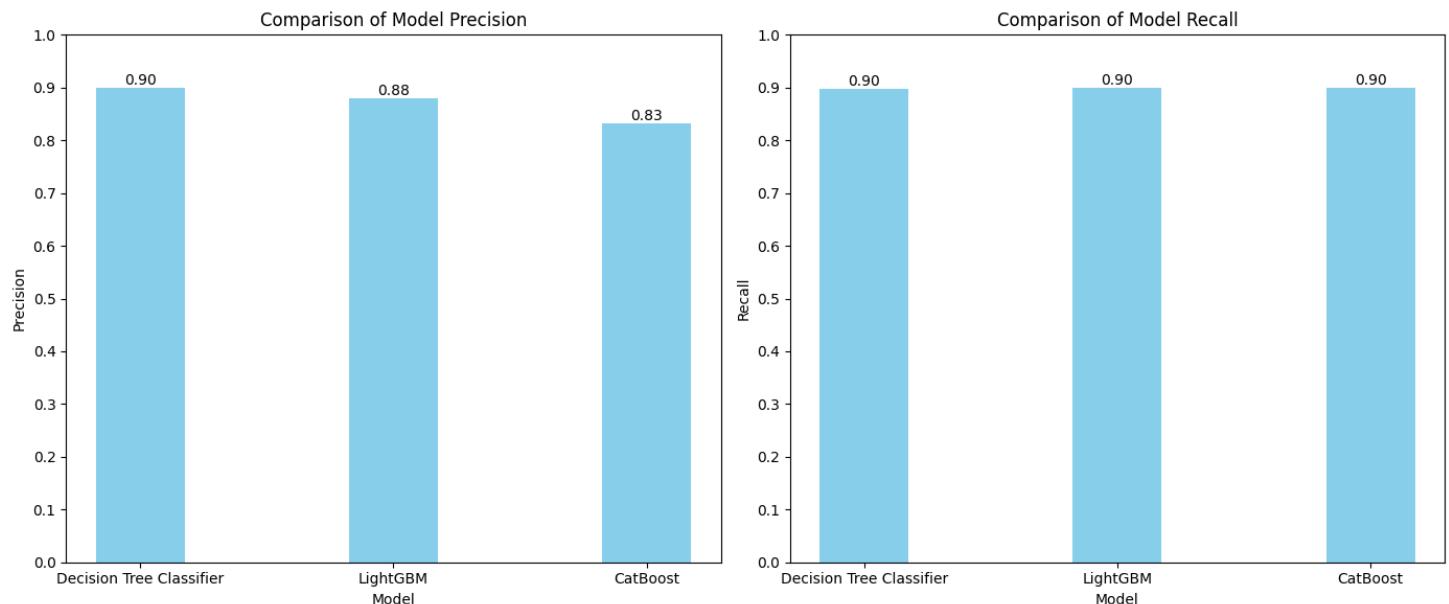
Nhược điểm:

- Tốn thời gian tính toán: Huấn luyện mô hình K lần có thể tốn nhiều thời gian, đặc biệt với tập dữ liệu lớn và mô hình phức tạp.
- Độ phức tạp: Việc triển khai K-Fold Cross-Validation có thể phức tạp hơn so với việc chỉ sử dụng một lần chia train/test.

4.6.5 Kết luận

4.6.5.1 Đánh giá kết quả mô hình





Hình 4.17 So sánh accuracy, f1_score, precision và recall của 3 mô hình

Accuracy:

- CatBoost và LightGBM có độ chính xác cao nhất, đạt 89.95%, cao hơn một chút so với Decision Tree (89.74%).
- Chênh lệch về Accuracy giữa các mô hình rất nhỏ, nhưng CatBoost và LightGBM có ưu thế hơn Decision Tree.

Precision:

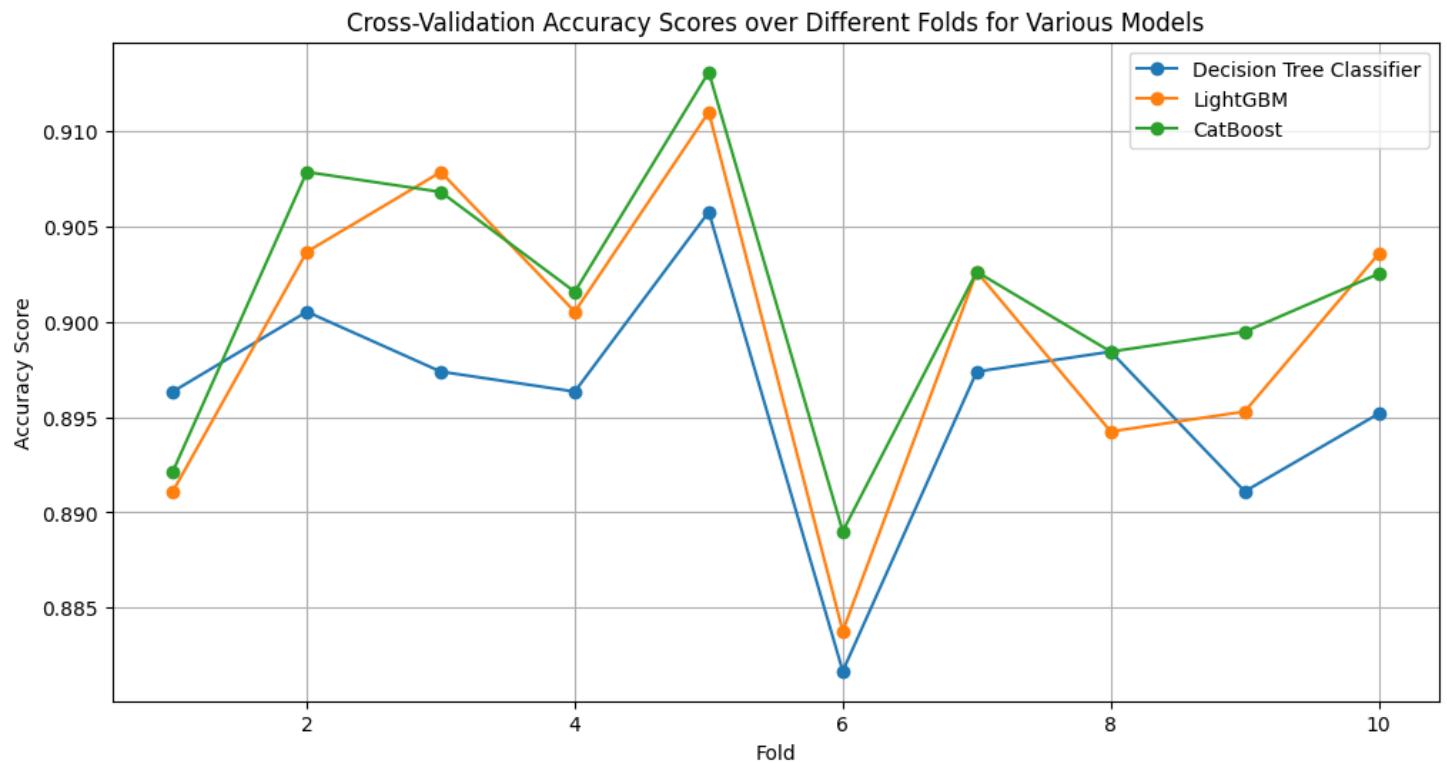
- Decision Tree có Precision cao nhất (90.04%), cho thấy mô hình ít gặp lỗi dương tính giả hơn so với hai mô hình còn lại.
- LightGBM đạt Precision khá cao (88.04%), trong khi CatBoost thấp nhất (83.23%).

Recall:

- CatBoost và LightGBM đạt Recall cao nhất (89.95%), cho thấy khả năng phát hiện các mẫu dương tính rất tốt.
- Decision Tree có Recall thấp hơn một chút (89.74%).

F1 Score:

- LightGBM có F1 Score cao nhất (86.40%), nhờ cân bằng tốt giữa Precision và Recall.
- CatBoost đứng thứ hai với F1 Score 86.17%, và Decision Tree xếp cuối với 85.89%.

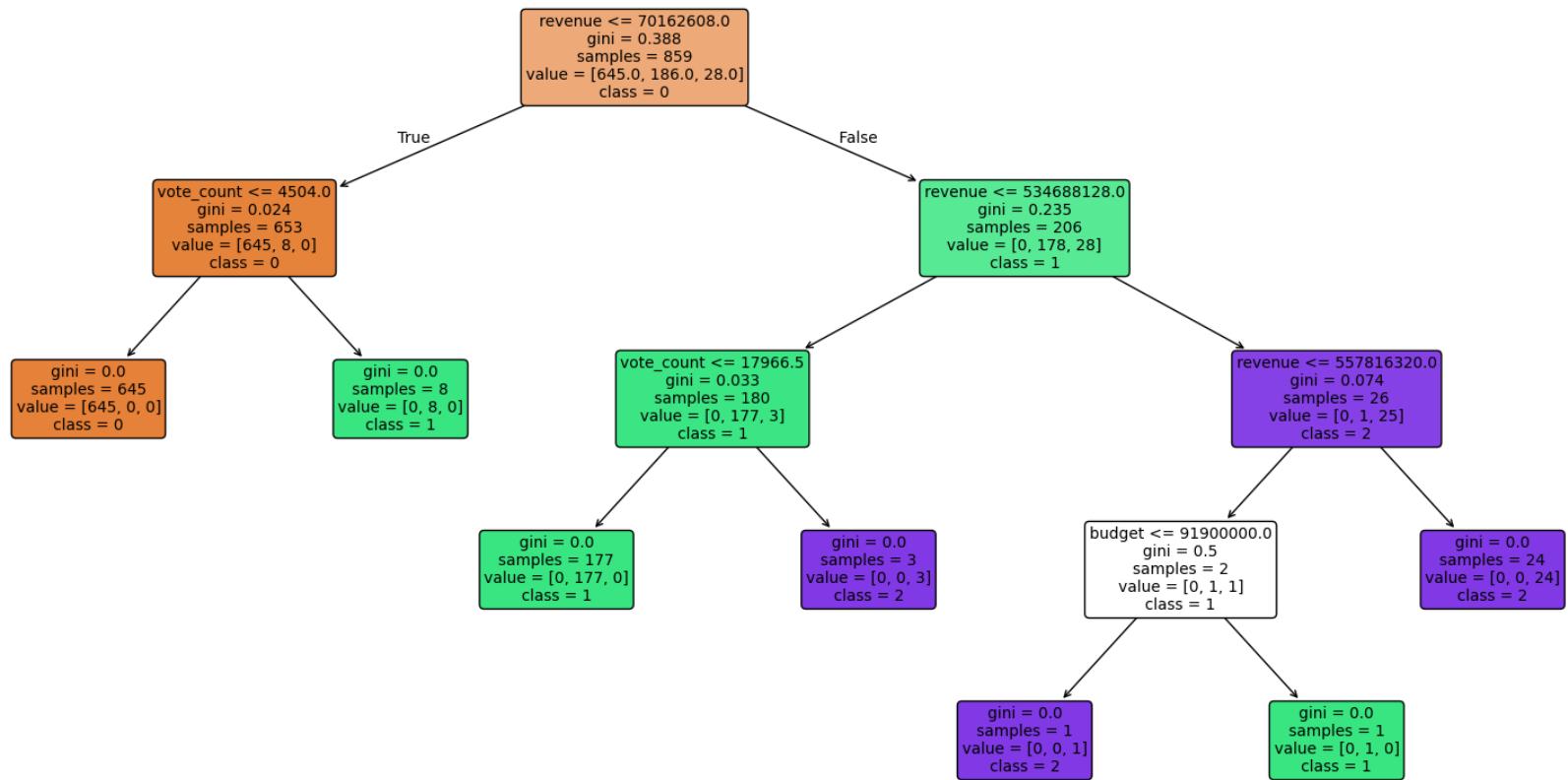


Hình 4.18 So sánh cross validation của 3 mô hình

Nhận xét:

- **CatBoost:** Có độ chính xác trung bình cao nhất trong ba mô hình.
- **LightGBM:** Có hiệu suất khá tốt, xếp sau CatBoost.
- **Decision Tree Classifier:** Mặc dù có độ ổn định thấp nhưng vẫn đạt được độ chính xác ổn định.

4.6.5.2 Nhận xét về đặc trưng nổi bật của từng nhãn



Hình 4.19 Decision Tree Rules

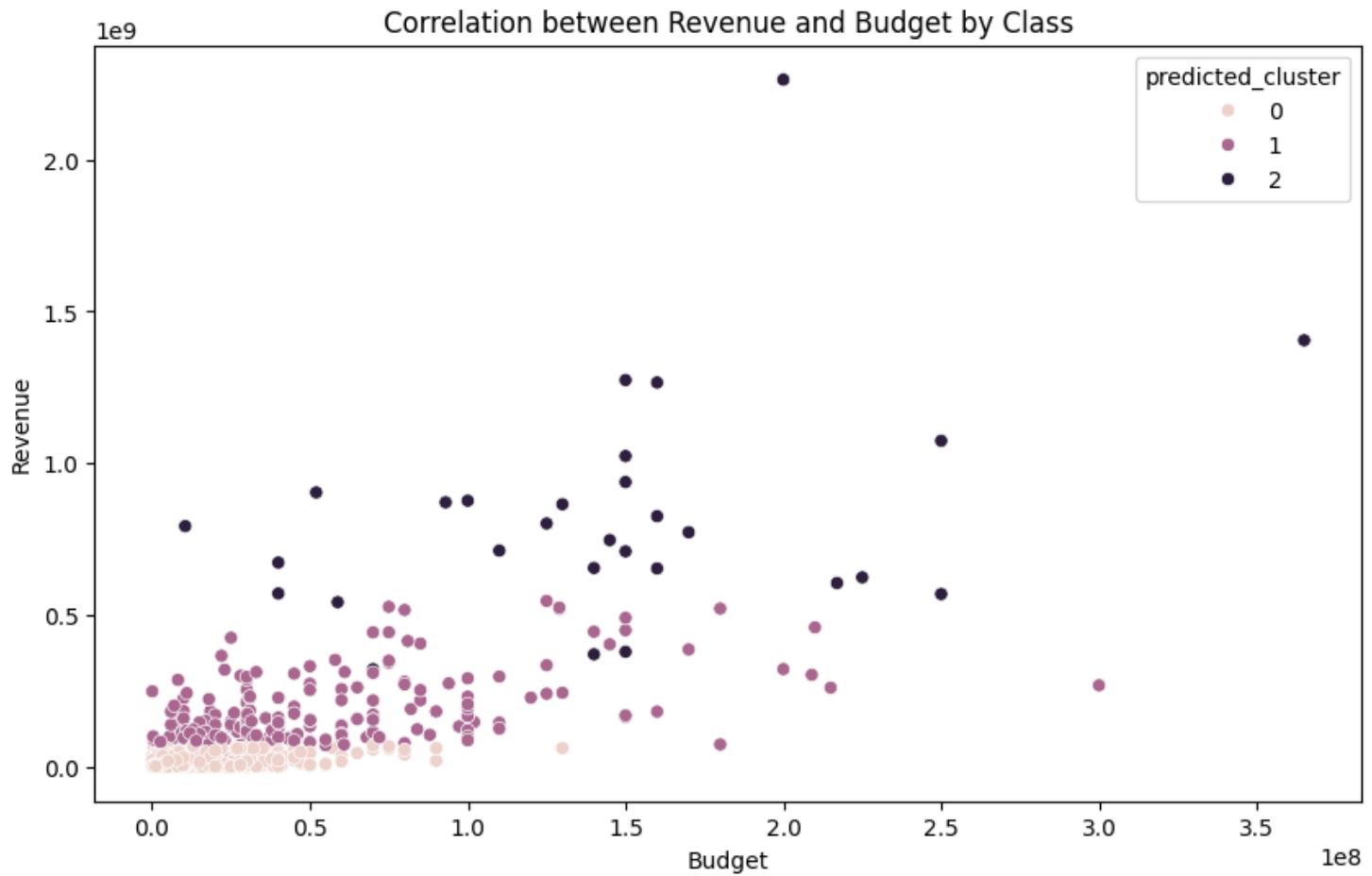
Biến mục tiêu của bộ dữ liệu được phân thành 3 nhóm:

- **Hit:** Các phim có đánh giá từ 8 trở lên – class 2.
- **Average:** Các phim có đánh giá từ 5 đến dưới 8 – class 1.
- **Flop:** Các phim có đánh giá dưới 5 – class 0.

Từ biểu đồ cây trên, nhóm rút ra được tập luật sau:

1. Nếu $\text{revenue} \leq 70162608.0$:
 - o Nếu $\text{vote_count} \leq 4504.0$:
 - Kết quả là class 0.
 - o Nếu $\text{vote_count} > 4504.0$:
 - Kết quả là class 1.
2. Nếu $\text{revenue} > 70162608.0$:
 - o Nếu $\text{revenue} \leq 534688128.0$:
 - Nếu $\text{vote_count} \leq 17966.5$:

- Kết quả là class 1.
- Nếu $\text{vote_count} > 17966.5$:
 - Kết quả là class 2.
- Nếu $\text{revenue} > 534688128.0$:
 - Nếu $\text{revenue} \leq 557816320$:
 - Nếu $\text{budget} \leq 91900000$:
 - Kết quả là class 2.
 - Nếu $\text{budget} > 91900000$:
 - Kết quả là class 1.
 - Nếu $\text{revenue} > 557816320$:
 - Kết quả là class 2.

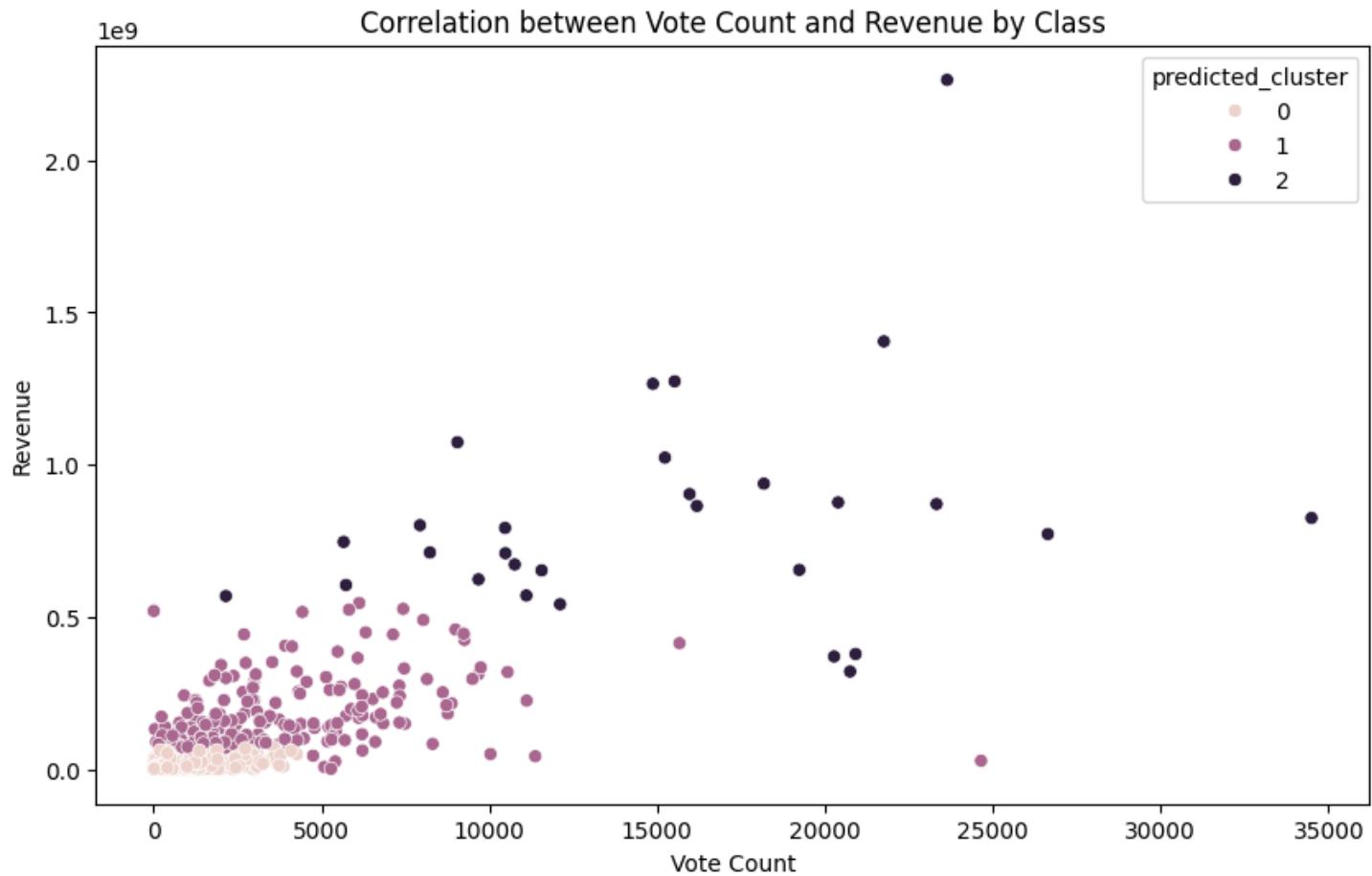


Hình 4.20 Sự tương quan giữa doanh thu và ngân sách giữa các lớp

Nhận xét:

- Nhóm 0 – Flop:
 - + Các điểm dữ liệu chủ yếu tập trung rất gần gốc tọa độ, thể hiện rằng cả ngân sách (Budget) và doanh thu (Revenue) đều thấp.
 - + Nhiều điểm dữ liệu tập trung gần gũi, cho thấy những dự án trong nhóm này thường không có đầu tư lớn và kết quả cũng hạn chế.
 - + Các điểm dữ liệu ở revenue tập trung ở mức rất thấp mặc dù các điểm dữ liệu có sự phân bố rải rác nhiều hơn ở budget. Điều đó cho thấy dù có các phim đầu tư nhưng doanh thu vẫn rất thấp
 - + Đặc biệt, khoảng cách giữa ngân sách và doanh thu rất lớn, thể hiện rằng nhóm này gồm các phim bị lỗ nặng.
 - ⇒ Ngân sách thấp hoặc cao nhưng doanh thu đều rất thấp, chứng tỏ hiệu suất kém, tập trung gần gốc tọa độ.
- Nhóm 1 – Average:
 - + Ngân sách từ trung bình đến cao, nhưng doanh thu lại không tương ứng cao.
 - + Nhóm này có sự phân bố rộng hơn so với nhóm 0.
 - + Có một lượng điểm dữ liệu rải rác, cho thấy sự linh hoạt nhưng không ổn định trong việc đạt doanh thu dù có ngân sách kha khá.
 - + Các phim trong nhóm này thường có hiệu suất tốt hơn, với sự tương quan rõ ràng hơn giữa ngân sách và doanh thu. Tuy nhiên, vẫn có những điểm dữ liệu cho thấy sự không hiệu quả, khi một số phim có ngân sách lớn nhưng doanh thu không tương ứng.
 - ⇒ Phân bố rộng hơn, có mối tương quan giữa ngân sách và doanh thu nhưng không ổn định ở một số trường hợp.
- Nhóm 2 – Hit:
 - + Nhóm này thể hiện rõ sự thành công vượt trội với các điểm dữ liệu có doanh thu rất cao.
 - + Ngân sách của các phim trong nhóm này cũng lớn hơn nhiều.
 - + Sự tương quan mạnh mẽ giữa ngân sách và doanh thu trong nhóm này là điểm đặc biệt, cho thấy rằng đầu tư lớn thường mang lại doanh thu lớn (khác với nhóm 0,1 với các phim mặc dù đầu tư tương đối nhưng doanh thu thấp hoặc rất thấp).
 - + Có các điểm dữ liệu doanh thu cực cao nhưng ngân sách lại trung bình cho thấy lợi nhuận khổng lồ của các phim thuộc nhóm 2.

⇒ Doanh thu rất cao, ngân sách lớn, mối quan hệ chặt chẽ giữa đầu tư và lợi nhuận, tập trung ở vùng trên.



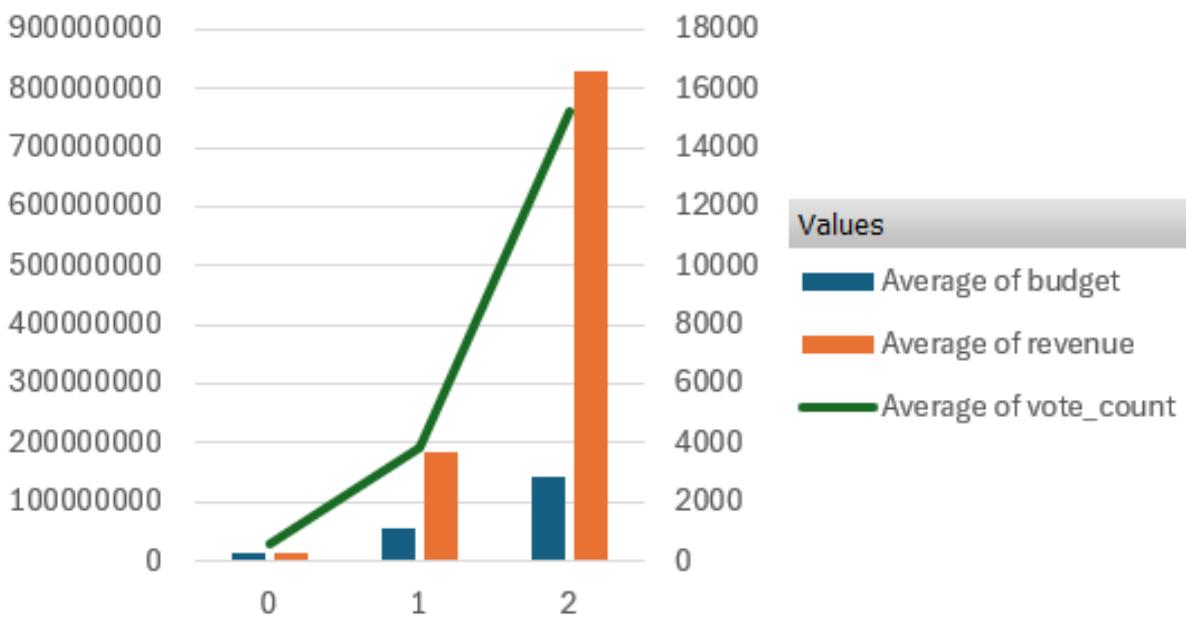
Hình 4.21 Sự tương quan giữa doanh thu và lượt đánh giá giữa các lớp

Nhận xét:

- Nhóm 0 – Flop:
 - + Nhóm này có số lượng điểm dữ liệu nhiều nhất, phân bố chủ yếu ở những giá trị thấp của cả số lượng phiếu và doanh thu.
 - + Hầu hết các bộ dữ liệu trong nhóm này có doanh thu dưới 0.5 tỷ với số lượng phiếu thường dưới 10,000, cho thấy đây là nhóm phim có doanh thu và sự quan tâm thấp nhất.
 - ⇒ Nhóm 0 đại diện cho các bộ phim ít được khán giả chú ý, thể hiện doanh thu kém và sự tương tác hạn chế từ người xem.
- Nhóm 1 – Average:
 - + Dữ liệu trong nhóm này phân bố rộng hơn so với nhóm 0, với một số điểm có doanh

thu từ 0.5 tỷ đến 1 tỷ.

- + Số lượng phiếu của các bộ phim này thường nằm trong khoảng từ 5,000 đến 20,000, cho thấy sự quan tâm tương đối từ khán giả.
- ⇒ Nhóm 1 phản ánh những bộ phim có thành công vừa phải, nhận được sự chú ý cao hơn nhóm 0 nhưng vẫn chưa đạt tới tầm cao của nhóm 2.
- Nhóm 2 – Hit:
 - + Nhóm này có số lượng điểm dữ liệu tập trung ở các vùng cao của cả "Vote Count" và "Revenue".
 - + Doanh thu của các bộ phim trong nhóm này thường vượt ngưỡng 1 tỷ, với số lượng phiếu có thể lên tới 35,000, cho thấy sức thu hút mạnh mẽ và thành công vượt trội.
 - ⇒ Doanh thu của các bộ phim trong nhóm này thường vượt ngưỡng 1 tỷ, với số lượng phiếu có thể lên tới 35,000, cho thấy sức thu hút mạnh mẽ và thành công vượt trội.



Hình 4.22 Biểu đồ thể hiện sự tương quan giữa 3 đặc trưng của mỗi lớp

Nhận xét chung:

- Biểu đồ cho thấy rõ sự tăng trưởng đồng bộ giữa ngân sách, doanh thu và số lượng phiếu từ nhóm 0 đến nhóm 2. Nhóm 0 có đặc trưng yếu kém về cả ba yếu tố, nhóm 1 có sự cải thiện vừa phải, trong khi nhóm 2 nổi bật với thành công vượt trội trong tất cả các chỉ số. Điều này chỉ ra rằng đầu tư lớn hơn vào sản xuất và tiếp thị thường dẫn đến sự tăng trong doanh thu và sự quan tâm từ khán giả.

Row Labels	Average of vote_count	Average of budget	Average of revenue
0	591.8879668	12013590.34	14934888.11
Grand Total	591.8879668	12013590.34	14934888.11

Row Labels	Max of vote_count	Max of budget	Max of revenue
0	4261	130000000	69633110
Grand Total	4261	130000000	69633110

Row Labels	Min of vote_count	Min of budget	Min of revenue
0	1	11475	3
Grand Total	1	11475	3

Hình 4.23 Giá trị average, max và min của các đặc trưng tại class 0 – Flop

Nhận xét:

- Vote Count (Số lượng đánh giá): Trung bình 591.89, dao động từ 1 đến 4,261. Điều này cho thấy rằng các phim thuộc nhóm "Flop" nhận được sự quan tâm rất ít từ khán giả.
- Budget (Ngân sách): Trung bình 12,013,590.34, với ngân sách dao động từ rất thấp (11,475) đến cao nhất 130,000,000. Điều này thể hiện rằng hầu hết các phim "Flop" đều thuộc nhóm phim có ngân sách thấp, nhưng vẫn có một vài phim có ngân sách lớn.
- Revenue (Doanh thu): Trung bình chỉ 14,934,888.11, với doanh thu thấp nhất là 3 và cao nhất là 69,633,110. Điều này chỉ ra rằng nhóm phim "Flop" thường không tạo ra doanh thu lớn.
- Sự tương quan giữa các đặc trưng:
 - + Vote Count và Revenue: Có mối quan hệ tích cực nhẹ, khi số lượng đánh giá càng nhiều thì doanh thu cũng có xu hướng tăng. Tương quan dương nhẹ, nhưng mức độ ảnh hưởng của số lượng đánh giá đến doanh thu không mạnh.
 - + Budget và Revenue: Tương quan thấp, vì ngay cả với ngân sách cao nhất (130 triệu USD), doanh thu cũng chỉ đạt tối đa gần 70 triệu USD. Trong khi đó, Budget rất thấp (chỉ 11,475) gần như không tạo ra doanh thu (3 USD). Tương quan rất yếu, cho thấy ngân sách không phải yếu tố quyết định chính trong nhóm này.
 - + Phim nhóm "Flop" có sự chênh lệch lớn giữa ngân sách và doanh thu. Ngay cả Vote Count tối đa (4,261), doanh thu cũng không vượt qua mức trung bình của các nhóm cao hơn.
- ⇒ Doanh thu phản ánh rõ ràng nhất sự thất bại của phim trong nhóm "Flop", bất chấp ngân sách đầu tư và sự quan tâm của khán giả.

Row Labels	Average of vote_count	Average of budget	Average of revenue
1	3818.527094	55267859.11	182712706.4
Grand Total	3818.527094	55267859.11	182712706.4
Row Labels	Max of vote_count	Max of budget	Max of revenue
1	24649	300000000	546388105
Grand Total	24649	300000000	546388105
Row Labels	Min of vote_count	Min of budget	Min of revenue
1	1	60000	1940906
Grand Total	1	60000	1940906

Hình 4.24 Giá trị average, max và min của các đặc trưng tại class 1 – Average

Nhận xét:

- Vote Count: Trung bình 3818.53, dao động từ 1 đến 24,649. Đây là sự gia tăng rõ rệt so với nhóm "Flop", chứng minh rằng nhóm này nhận được sự quan tâm tốt hơn từ khán giả.
- Budget: Trung bình 55,267,859.11, dao động từ 60,000 đến 300,000,000. Ngân sách nhóm này trải dài từ thấp đến rất cao, cho thấy các phim thuộc nhóm Average có được sự đầu tư nhất định.
- Revenue: Trung bình 182,712,706.4, dao động từ 1,940,906 đến 546,388,105. Doanh thu nhóm này có sự cải thiện lớn so với nhóm "Flop" và đạt mức khá cao ở một số phim.
- Sự tương quan giữa các đặc trưng:
 - + Vote Count và Revenue: Tương quan rõ rệt hơn so với nhóm "Flop", khi số lượng đánh giá tăng cao thì doanh thu cũng tăng mạnh. Điều này có thể giải thích bởi sự lan tỏa tốt hơn và chất lượng phim tương đối ổn.
 - + Budget và Revenue: Tương quan dương đáng kể, khi ngân sách tăng thì doanh thu cũng có xu hướng tăng. Tuy nhiên, doanh thu không nổi bật hẳn so với ngân sách.
 - + Vote Count trung bình tăng mạnh lên 3818.53, và sự gia tăng này trực tiếp đẩy doanh thu trung bình lên 182.7 triệu USD.
 - + Budget trung bình tăng lên 55.2 triệu USD, và phim có ngân sách cao nhất (300 triệu USD) thường tạo doanh thu cao. Tuy nhiên, vẫn có ngoại lệ với ngân sách thấp nhất (60,000 USD) nhưng doanh thu thấp nhất lại tận (1.9 triệu USD).
- ⇒ Số lượng đánh giá là yếu tố tác động mạnh nhất trong nhóm "Average", giúp doanh thu tăng trưởng đáng kể khi khán giả quan tâm nhiều hơn.

Row Labels	Average of vote_count	Average of budget	Average of revenue
2	15248.93103	143493103.4	830655761.2
Grand Total	15248.93103	143493103.4	830655761.2
Row Labels	Max of vote_count	Max of budget	Max of revenue
2	34495	365000000	2264162353
Grand Total	34495	365000000	2264162353
Row Labels	Min of vote_count	Min of budget	Min of revenue
2	2148	10500000	321457747
Grand Total	2148	10500000	321457747

Hình 4.25 Giá trị average, max và min của các đặc trưng tại class 2 - Hit

Nhận xét:

- Vote Count: Trung bình 15,248.93, dao động từ 2,148 đến 34,495. Nhóm này thu hút sự quan tâm lớn từ khán giả với số lượng đánh giá cao nhất trong cả 3 lớp.
- Budget: Trung bình 143,493,103.4, dao động từ 10,500,000 đến 365,000,000. Các phim thuộc nhóm "Hit" thường có ngân sách lớn, thể hiện sự đầu tư mạnh mẽ từ nhà sản xuất.
- Revenue: Trung bình 830,655,761.2, dao động từ 321,457,747 đến 2,264,162,353. Doanh thu cực kỳ ấn tượng, đặc biệt có những phim vượt mức 2 tỷ USD.
- Sự tương quan giữa các đặc trưng:
 - + Vote Count và Revenue: Tương quan chặt chẽ, khi số lượng đánh giá tăng thì doanh thu cũng tăng đáng kể, cho thấy sự thành công lan tỏa mạnh mẽ của nhóm phim này.
 - + Budget và Revenue: Tương quan dương rất mạnh, chứng minh rằng ngân sách lớn được sử dụng hiệu quả trong việc sản xuất và quảng bá, dẫn đến thành công vượt bậc.
- ⇒ Ngân sách là yếu tố tác động mạnh nhất trong nhóm "Hit", giúp tối ưu hóa cả doanh thu và số lượng đánh giá.

Tài liệu tham khảo

- [1] Ayesha Siddique1, Muhammad Kamran Abid. Muhammad Fuzail, Naeem Aslam, “Movies Rating Prediction Using Supervised Machine Learning Techniques”, January 2024.
Available:
https://www.researchgate.net/publication/378242486_Movies_Rating_Prediction_using_Supervised_Machine_Learning_Techniques
- [2] Brain John, “When to Choose CatBoost Over XGBoost or LightGBM” October, 2024.
[Trực tuyến]. Available: <https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm>
- [3] CatBoost (n.d.). CatBoostClassifier. CatBoost.
Available: https://catboost.ai/docs/en/concepts/python-reference_catboostclassifier
- [4] M Saraee, S White, J Eccleston, “A Data Mining Approach To Analysis And Prediction Of Movie Ratings”, 2004.
Available: <https://www.witpress.com/elibrary/wit-transactions-on-information-and-communication-technologies/33/14248>
- [5] Hồ Thị Ngọc Huyền và Phạm Huỳnh Mỹ Hạnh, *Tài liệu Seminar Câu truy vấn ngôn ngữ MDX*.