22/06/2022

DW&DM

MS. NUSRAT SHAHEEN

LAB PROJECT

Submitted By:

***Faiq Shahzad* (FA19-BCS-021)**

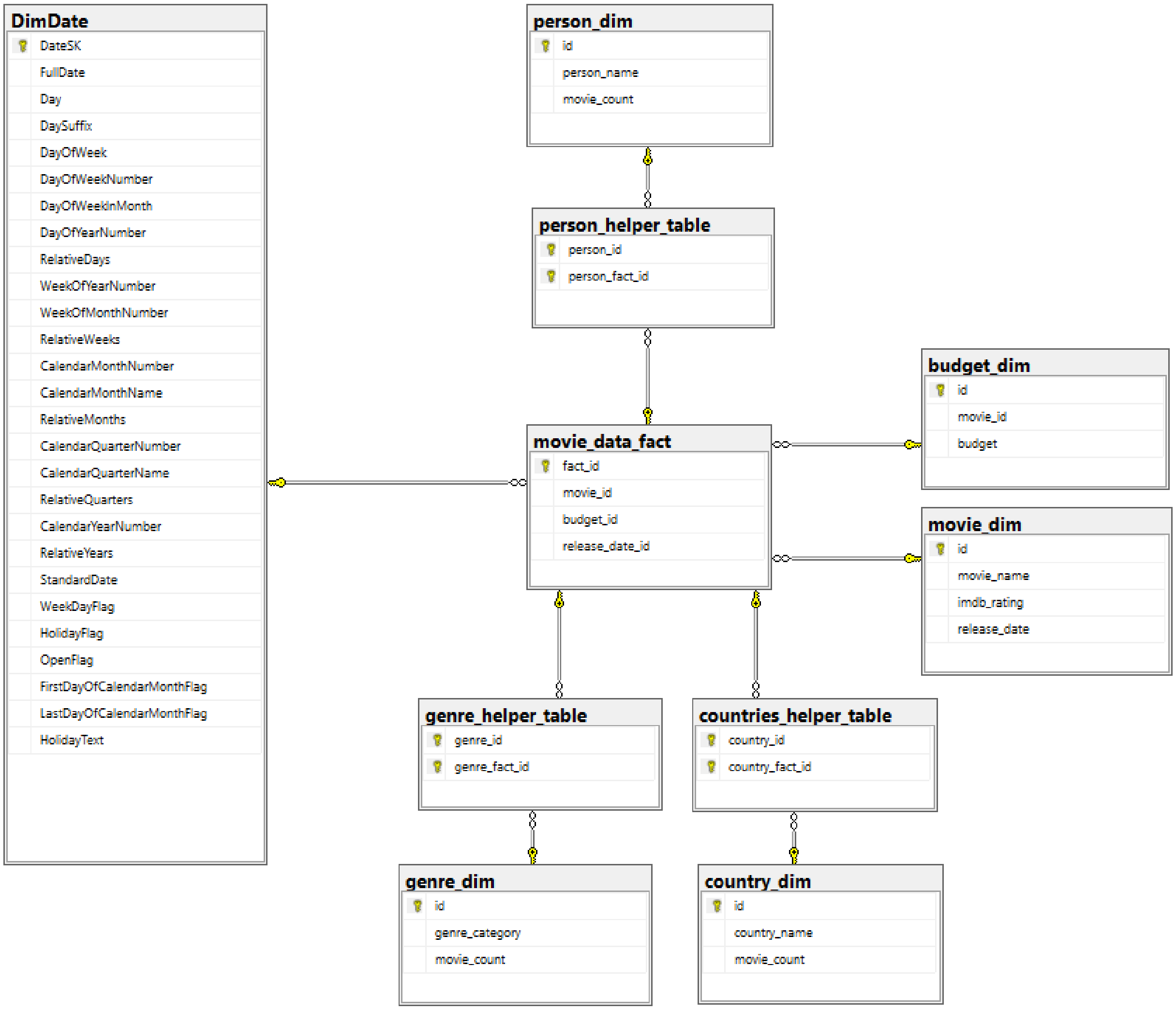
***Muhammad Ahmed* (FA19-BCS-041)**

***Muhammad Mubir Shami* (FA19-BCS-051)**

Business Questions:

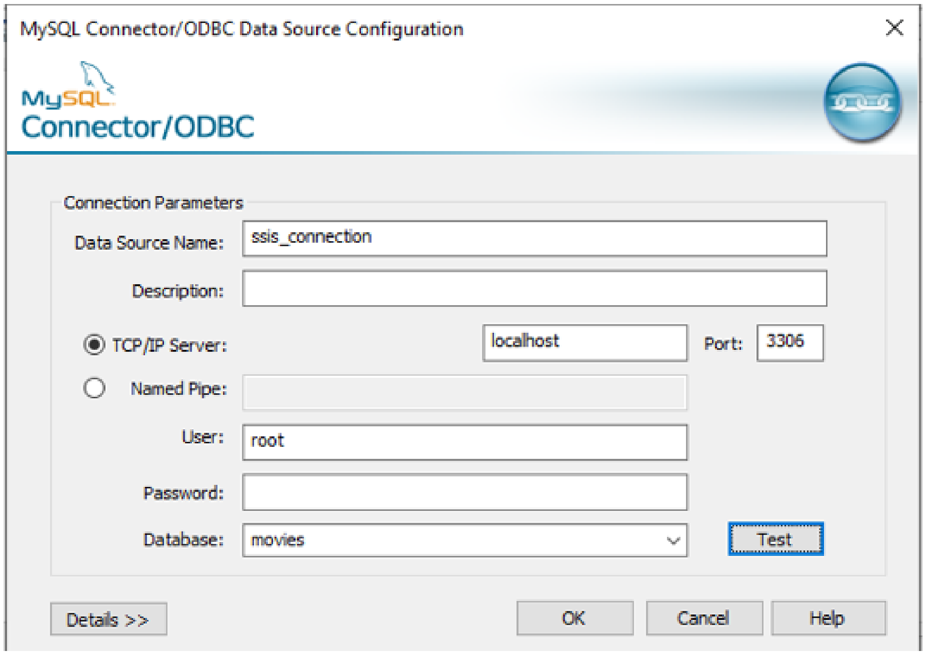
- What are the total number of movies in the specific genre?
- What are the total number of Movies produced in a specific country?
- What are the total number of Movies a person has participated in?
- What is the list of movies released in a specific year?
- What is the list of movies under a specific budget?

Dimension Model:



ETL:

Step 1: Creating Operational Database Connection:



Connection Parameters

Data Source Name: ssis_connection

Description:

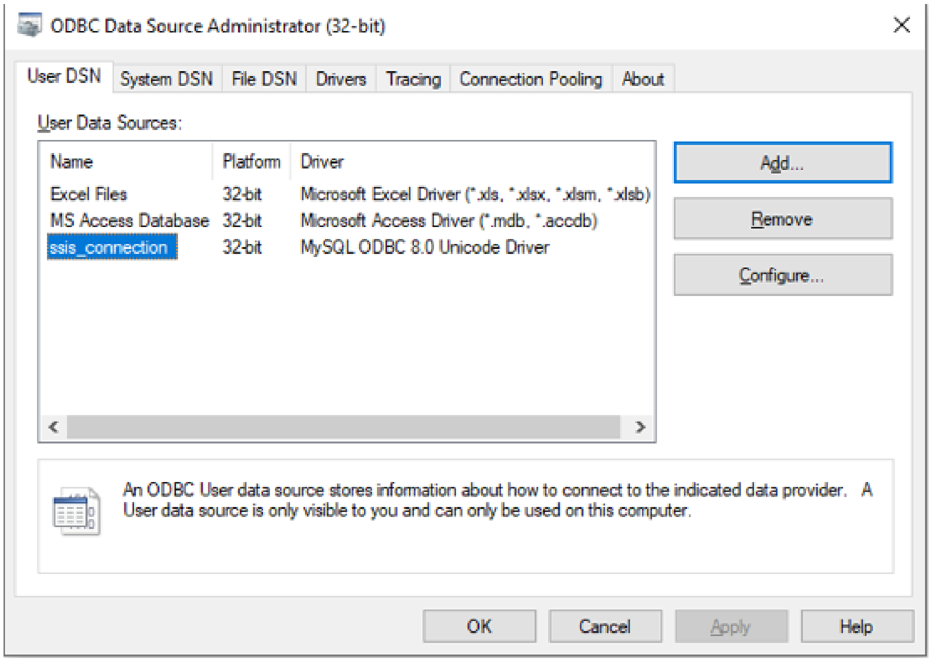
● TCP/IP Server: localhost Port: 3306

○ Named Pipe:

User: root

Password:

Database: movies



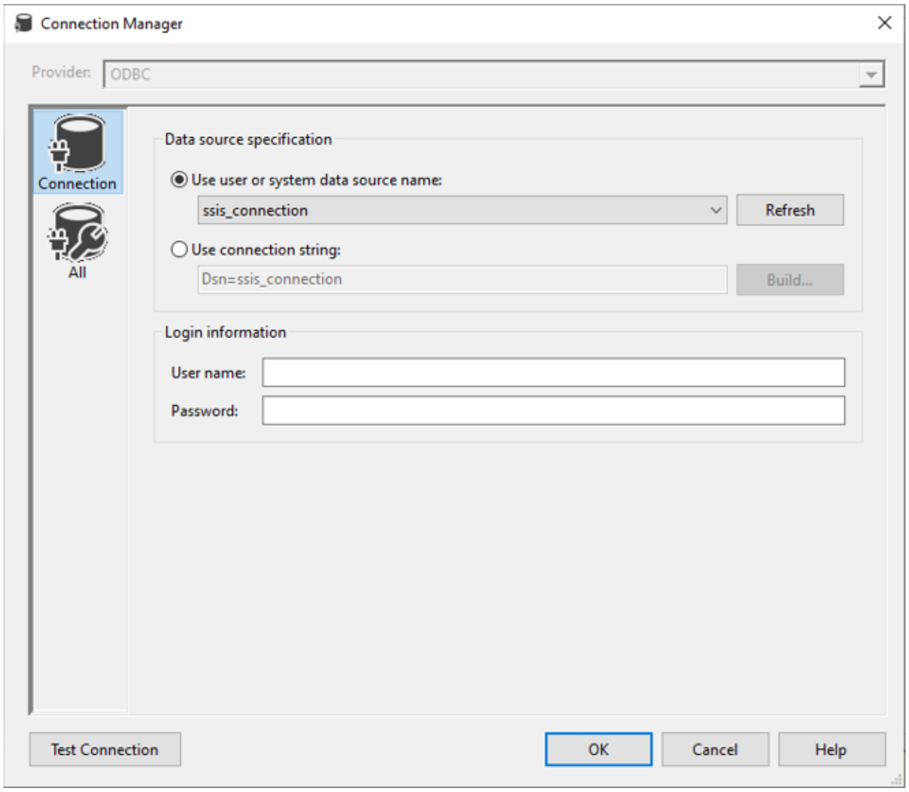
User Data Sources:

Name	Platform	Driver
Excel Files	32-bit	Microsoft Excel Driver (*.xls, *.xlsx, *.xlsm, *.xlsb)
MS Access Database	32-bit	Microsoft Access Driver (*.mdb, *.accdb)
ssis_connection	32-bit	MySQL ODBC 8.0 Unicode Driver

An ODBC User data source stores information about how to connect to the indicated data provider. A User data source is only visible to you and can only be used on this computer.

Step 2: Extraction:

1. Create Connections:



Provider: ODBC

Data source specification

● Use user or system data source name: ssis_connection

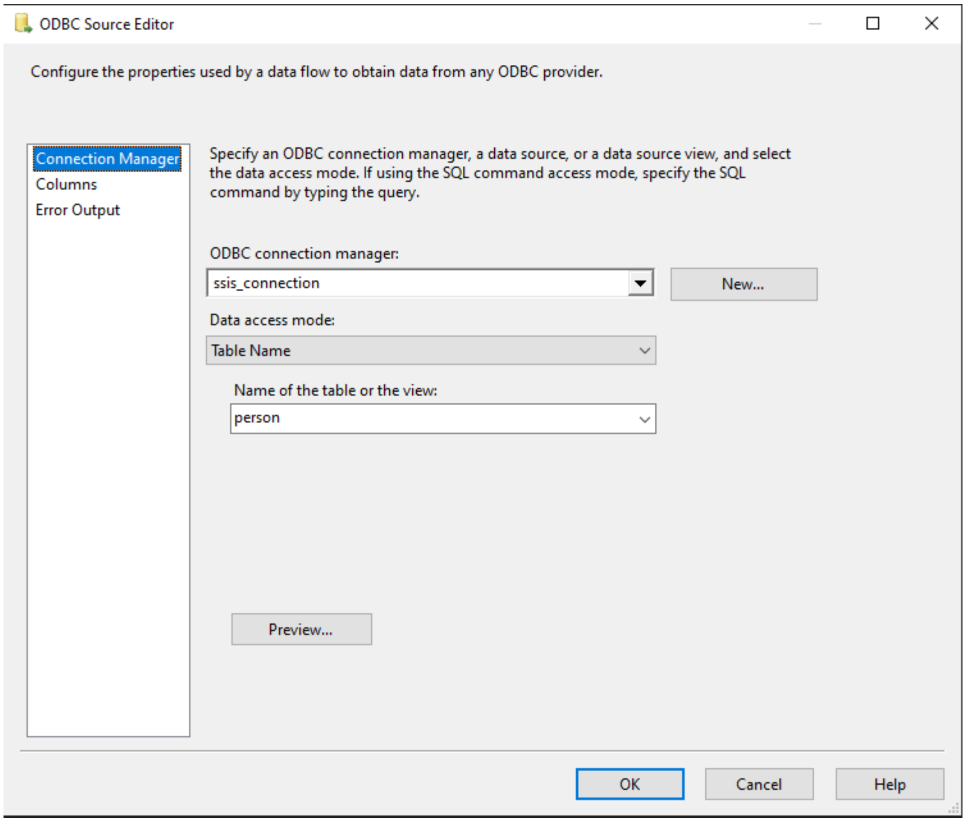
○ Use connection string: Dsn=ssis_connection

Login information

User name:

Password:

2. Extracting Table from Operation Database:



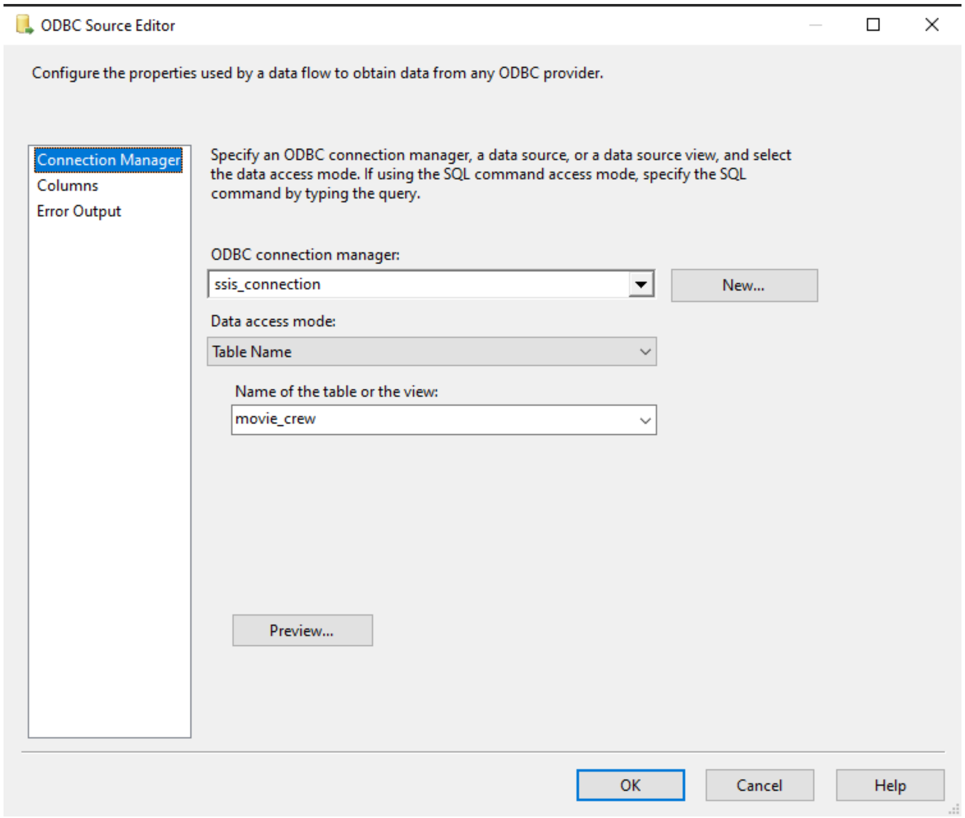
Configure the properties used by a data flow to obtain data from any ODBC provider.

Specify an ODBC connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command by typing the query.

ODBC connection manager: ssis_connection

Data access mode: Table Name

Name of the table or the view: person



ODBC Source Editor

Configure the properties used by a data flow to obtain data from any ODBC provider.

Specify an ODBC connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command by typing the query.

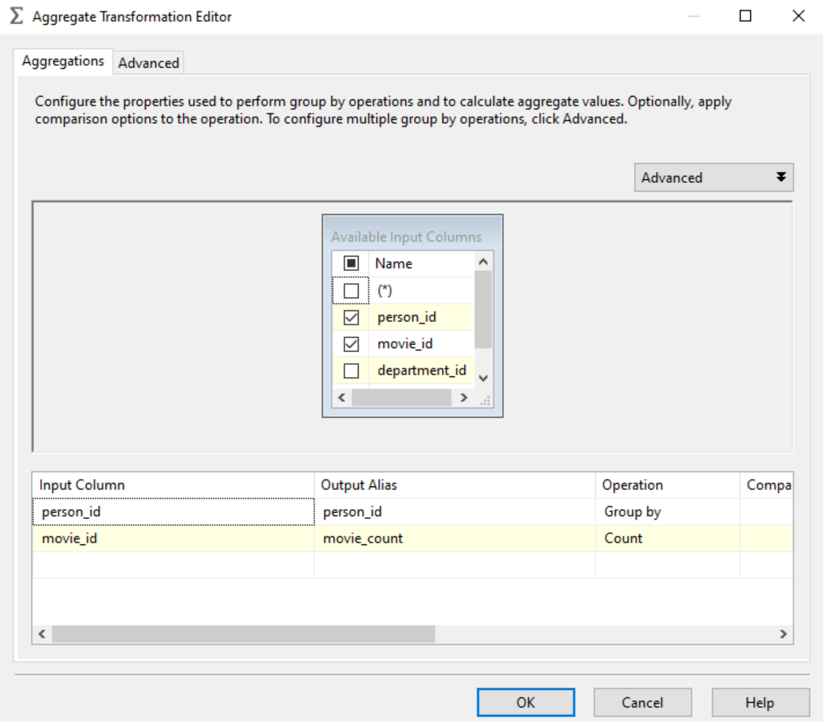
ODBC connection manager: ssis_connection

Data access mode: Table Name

Name of the table or the view: movie_crew

Step 3: Transformations:

1. Choose a Transformation to be performed on the Extracted Data such as:
   - **Aggregate (Group By)**



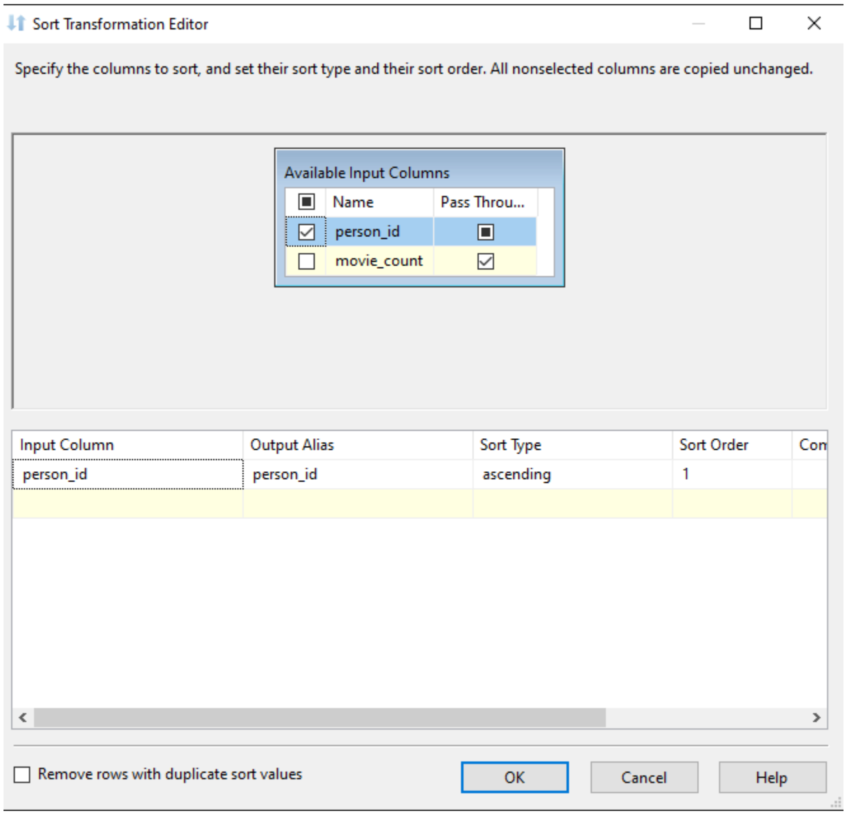
Aggregate Transformation Editor

Configure the properties used to perform group by operations and to calculate aggregate values. Optionally, apply comparison options to the operation. To configure multiple group by operations, click Advanced.

■	Name
☐	(*)
☑	person_id
☑	movie_id
☐	department_id

Input Column	Output Alias	Operation	Compa
person_id	person_id	Group by	
movie_id	movie_count	Count	

- **Sort**



Sort Transformation Editor

Specify the columns to sort, and set their sort type and their sort order. All nonselected columns are copied unchanged.

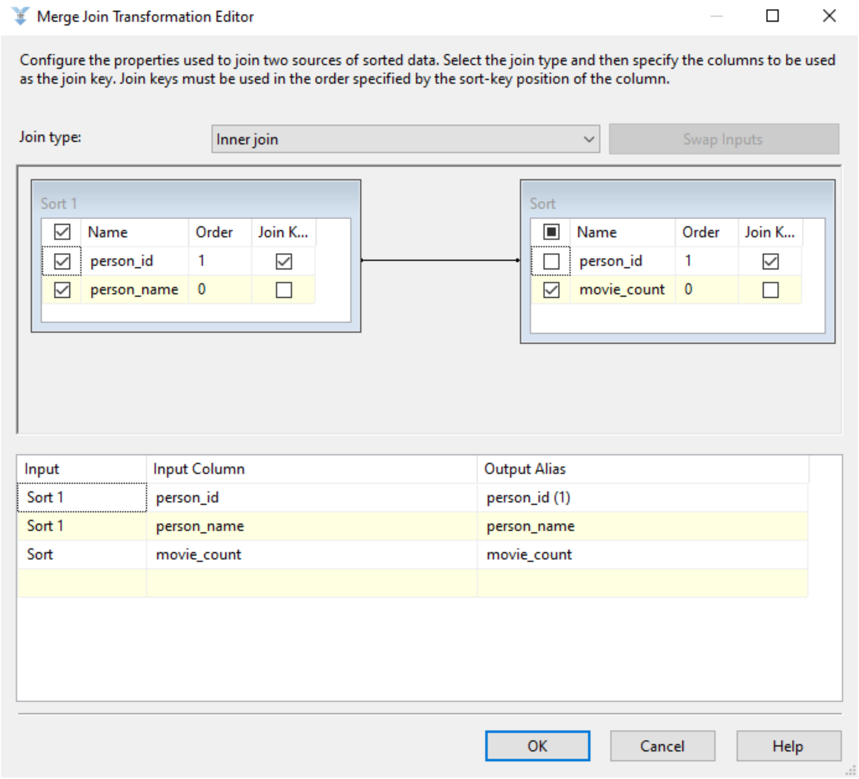
Available Input Columns

■	Name	Pass Throu...
☑	person_id	■
□	movie_count	☑

Input Column	Output Alias	Sort Type	Sort Order	Con
person_id	person_id	ascending	1	

□ Remove rows with duplicate sort values

- **Merge Join**



Merge Join Transformation Editor

Configure the properties used to join two sources of sorted data. Select the join type and then specify the columns to be used as the join key. Join keys must be used in the order specified by the sort-key position of the column.

Join type: Inner join

Sort 1

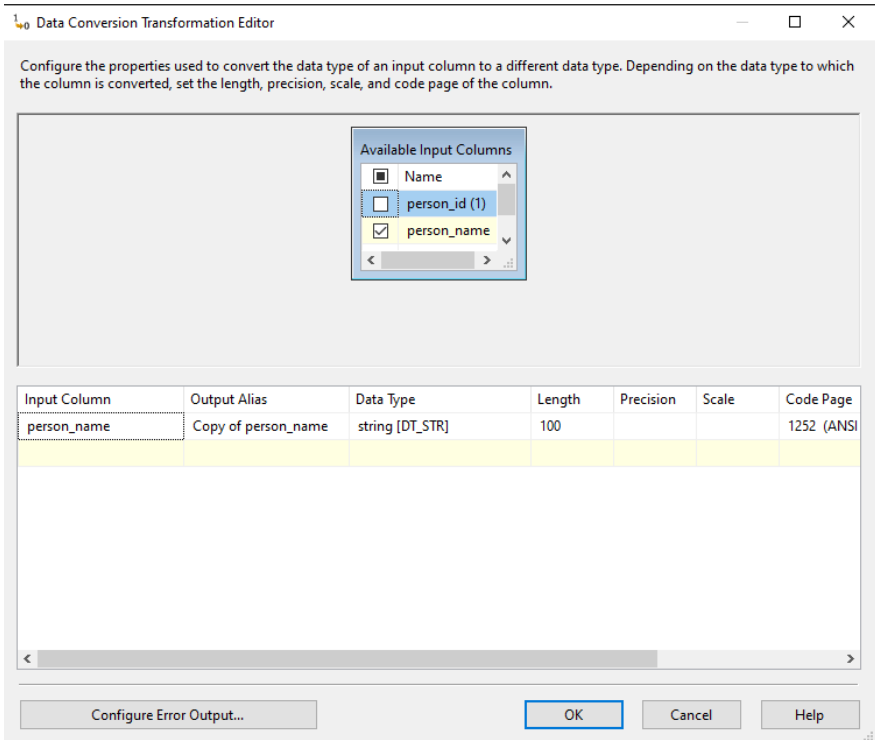
☑	Name	Order	Join K...
☑	person_id	1	☑
☑	person_name	0	□

Sort

■	Name	Order	Join K...
□	person_id	1	☑
☑	movie_count	0	□

Input	Input Column	Output Alias
Sort 1	person_id	person_id (1)
Sort 1	person_name	person_name
Sort	movie_count	movie_count

- **Data Conversion**



Data Conversion Transformation Editor

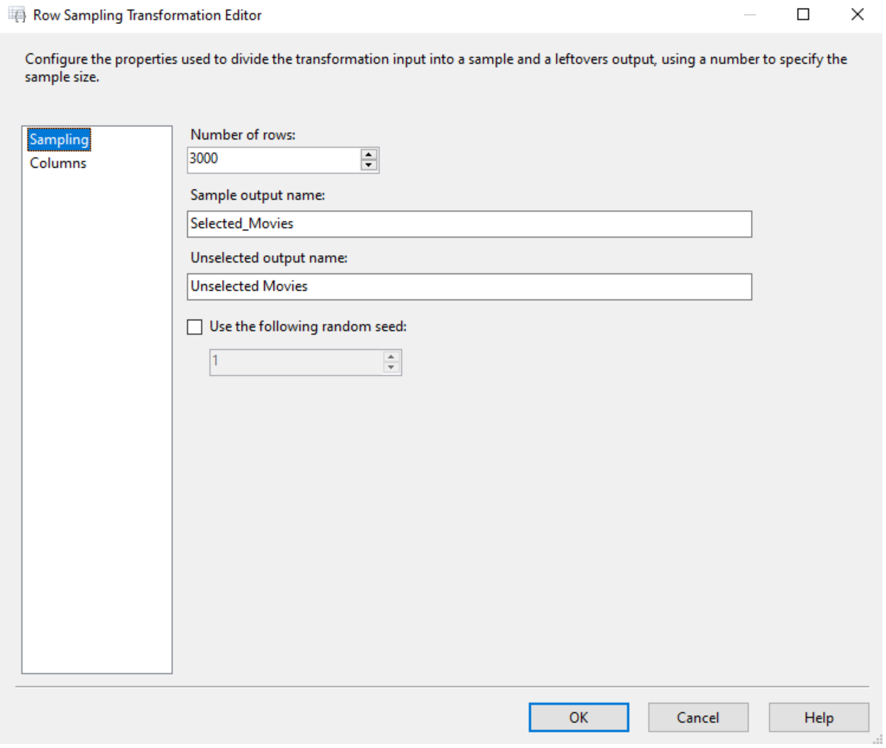
Configure the properties used to convert the data type of an input column to a different data type. Depending on the data type to which the column is converted, set the length, precision, scale, and code page of the column.

Available Input Columns

	Name
□	person_id (1)
☑	person_name

Input Column	Output Alias	Data Type	Length	Precision	Scale	Code Page
person_name	Copy of person_name	string [DT_STR]	100			1252 (ANSI

- **Row Sampling**



Row Sampling Transformation Editor

Configure the properties used to divide the transformation input into a sample and a leftovers output, using a number to specify the sample size.

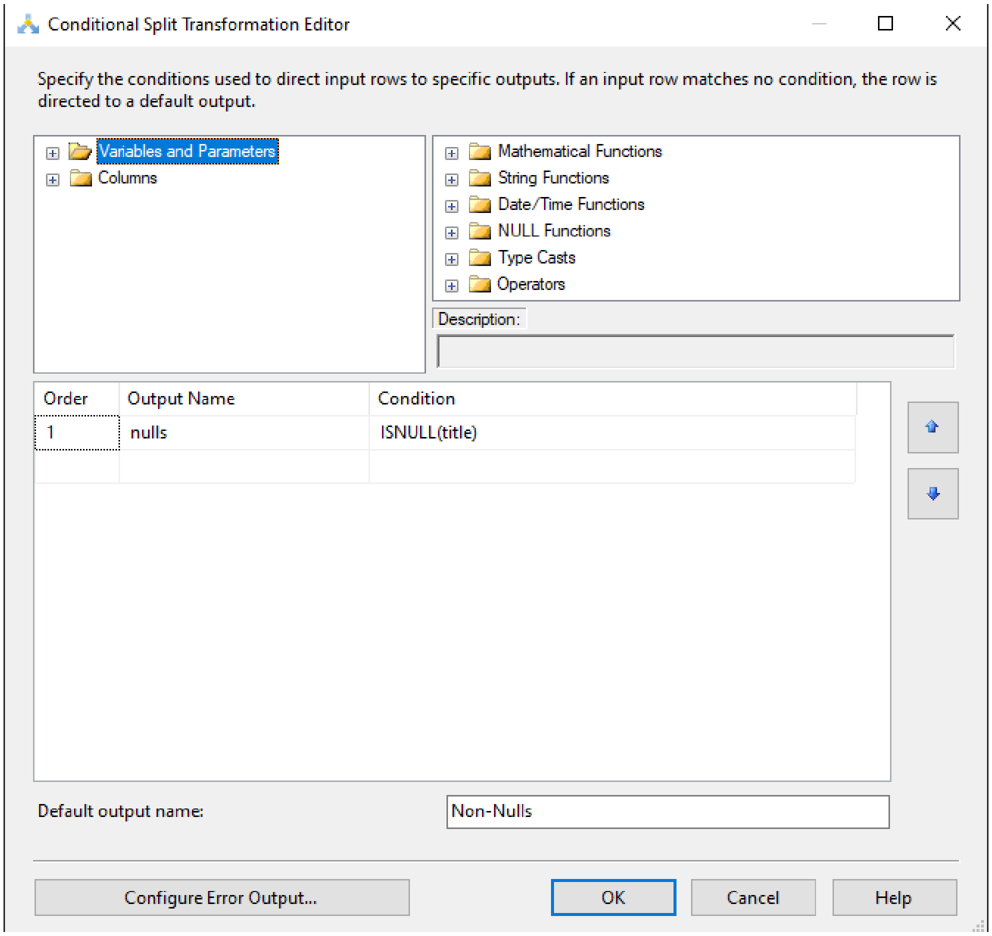
Number of rows: 3000

Sample output name: Selected_Movies

Unselected output name: Unselected Movies

□ Use the following random seed: 1

- **Conditional Split:**



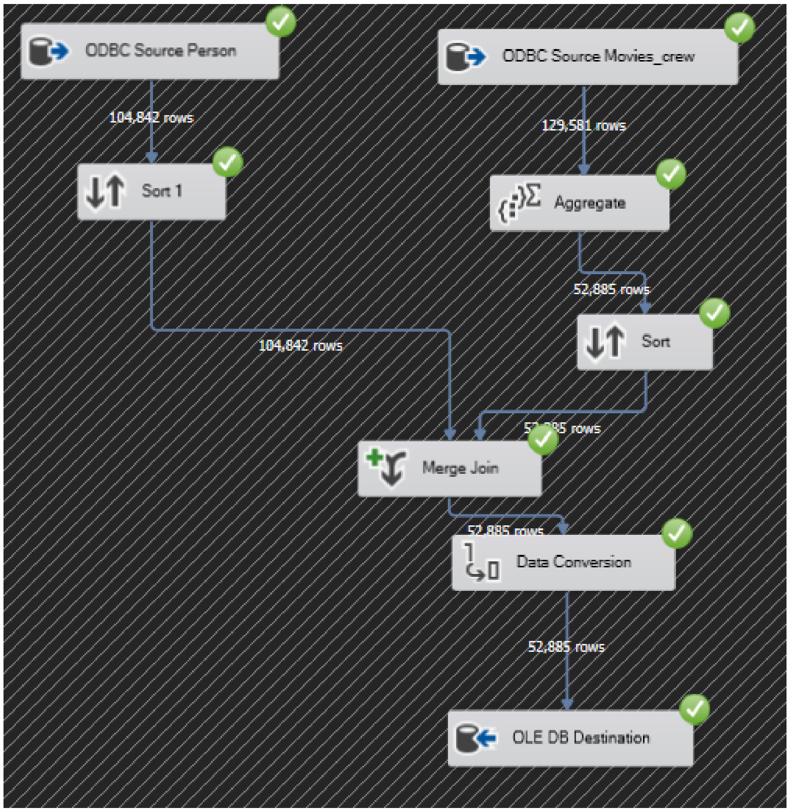
Specify the conditions used to direct input rows to specific outputs. If an input row matches no condition, the row is directed to a default output.

Order	Output Name	Condition
1	nulls	ISNULL(title)

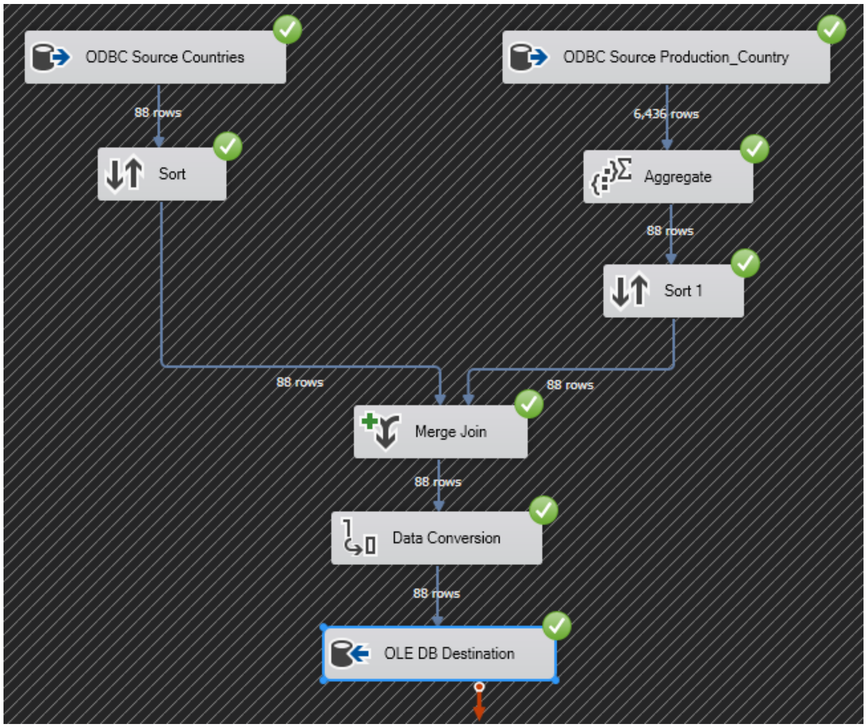
Default output name: Non-Nulls

2. Perform the Transformations:

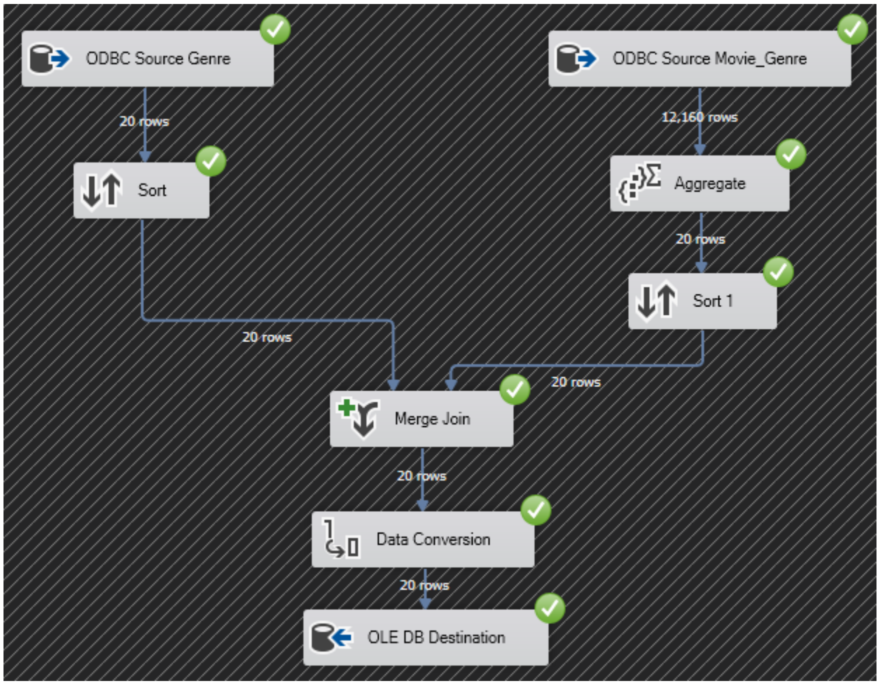
- **Person_Dim:**



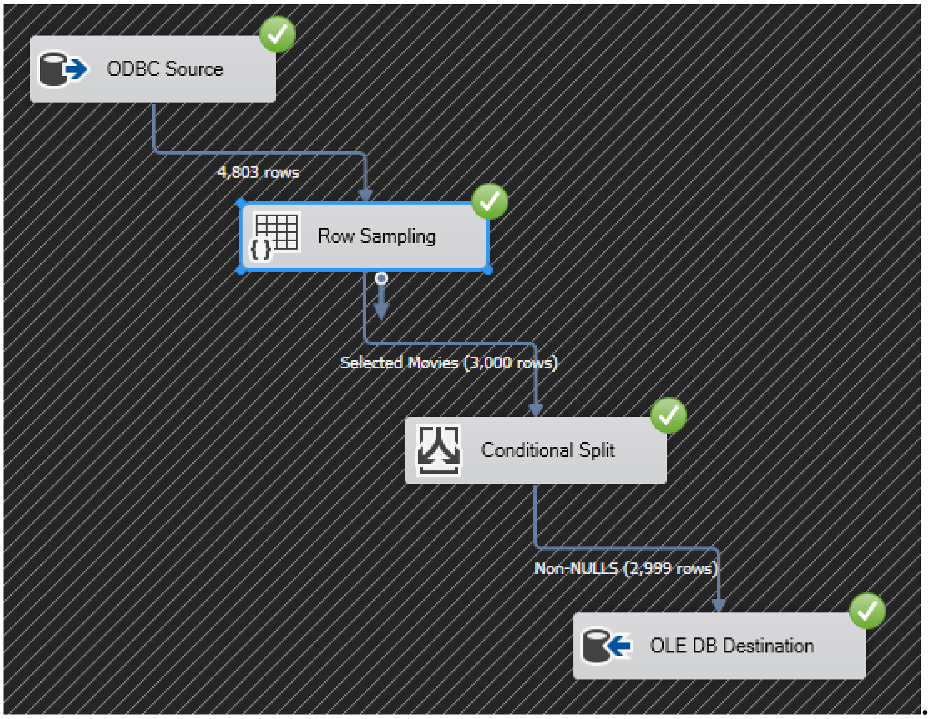
- **Countries_dim:**



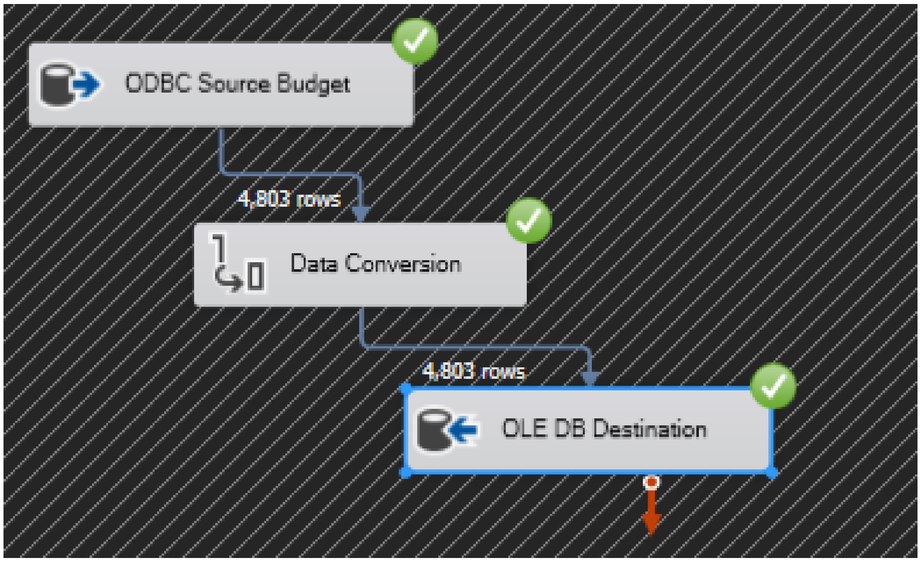
- **Genre_dim**



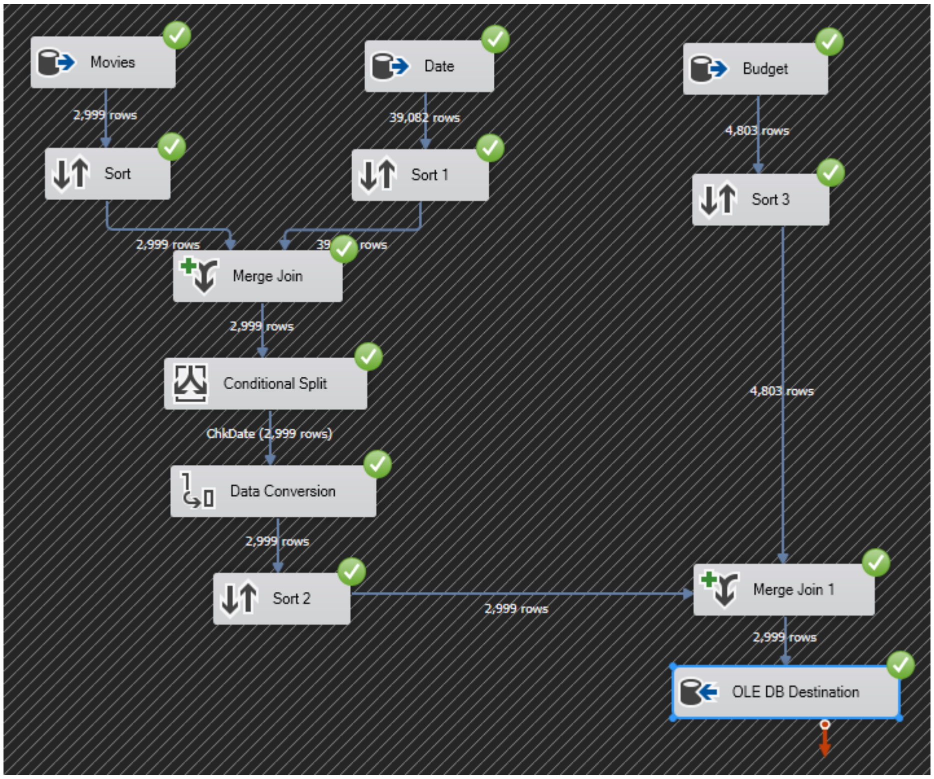
- **Movies_dim**



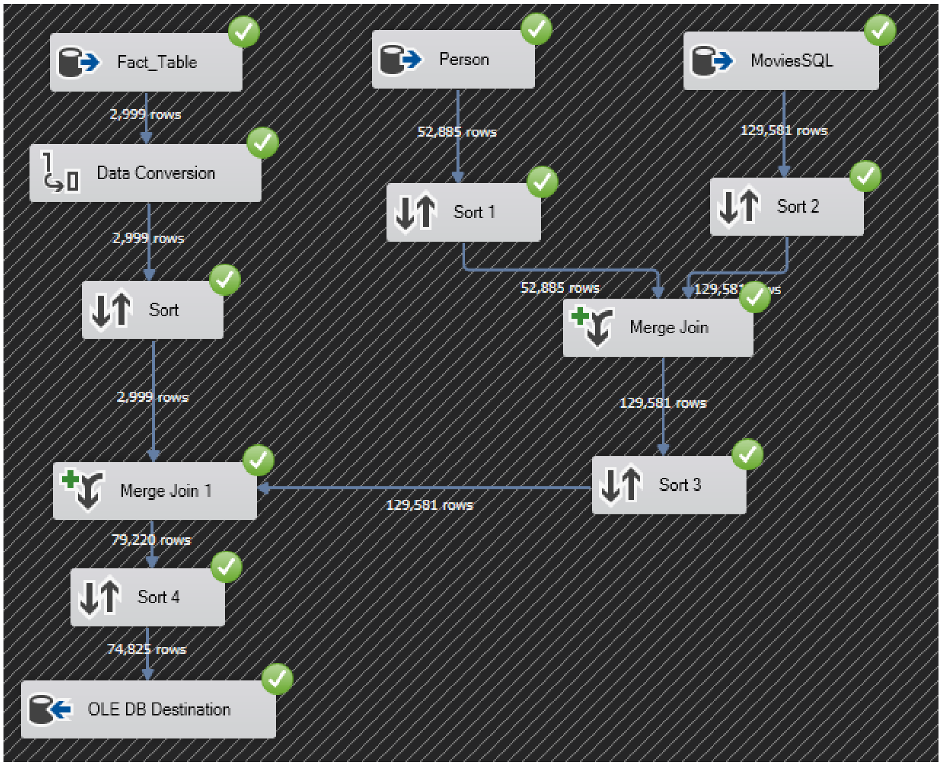
- **Budget_dim**



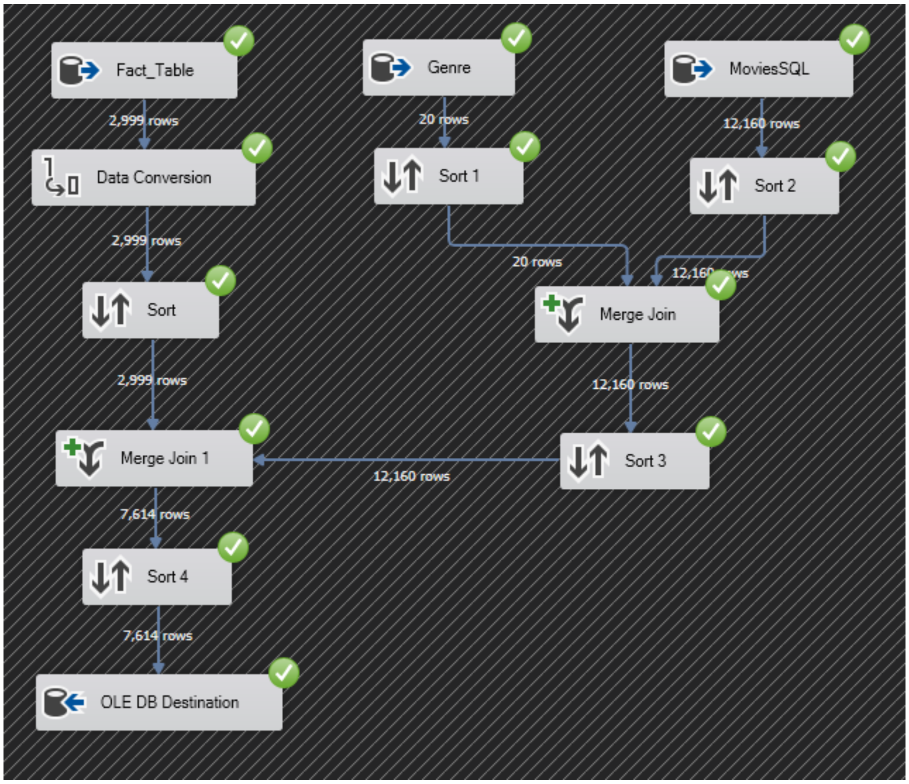
- **Movie_Data_Fact**



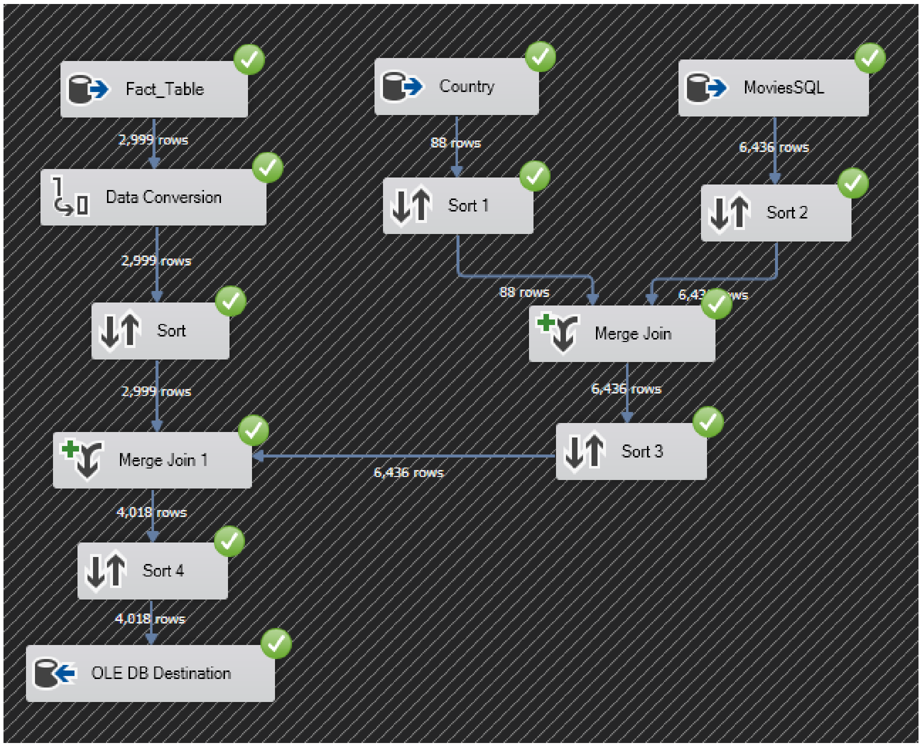
- **Person_Helper_Table**



- **Genre_Helper_Table**

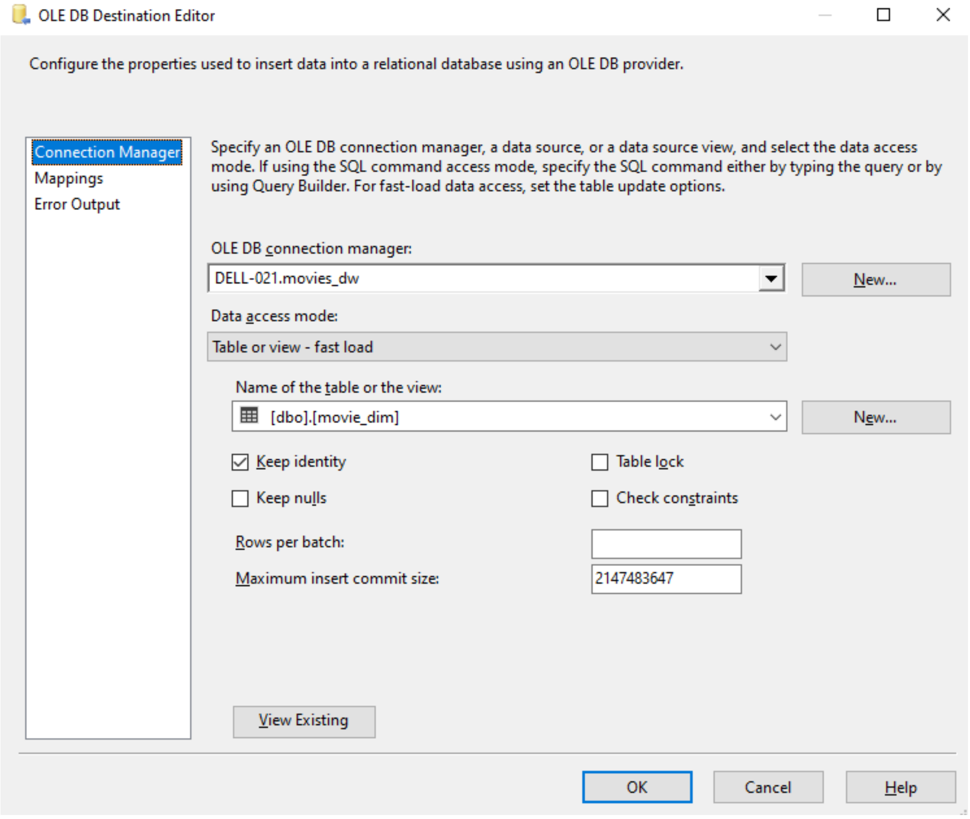


- **Country_Helper_Table**



Step 4: Loading:

1. Select the OLE DB from the destination panel
2. Connect to the desired table in the database



OLE DB Destination Editor

Configure the properties used to insert data into a relational database using an OLE DB provider.

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.

OLE DB connection manager: DELL-021.movies_dw

Data access mode: Table or view - fast load

Name of the table or the view: [dbo].[movie_dim]

☑ Keep identity

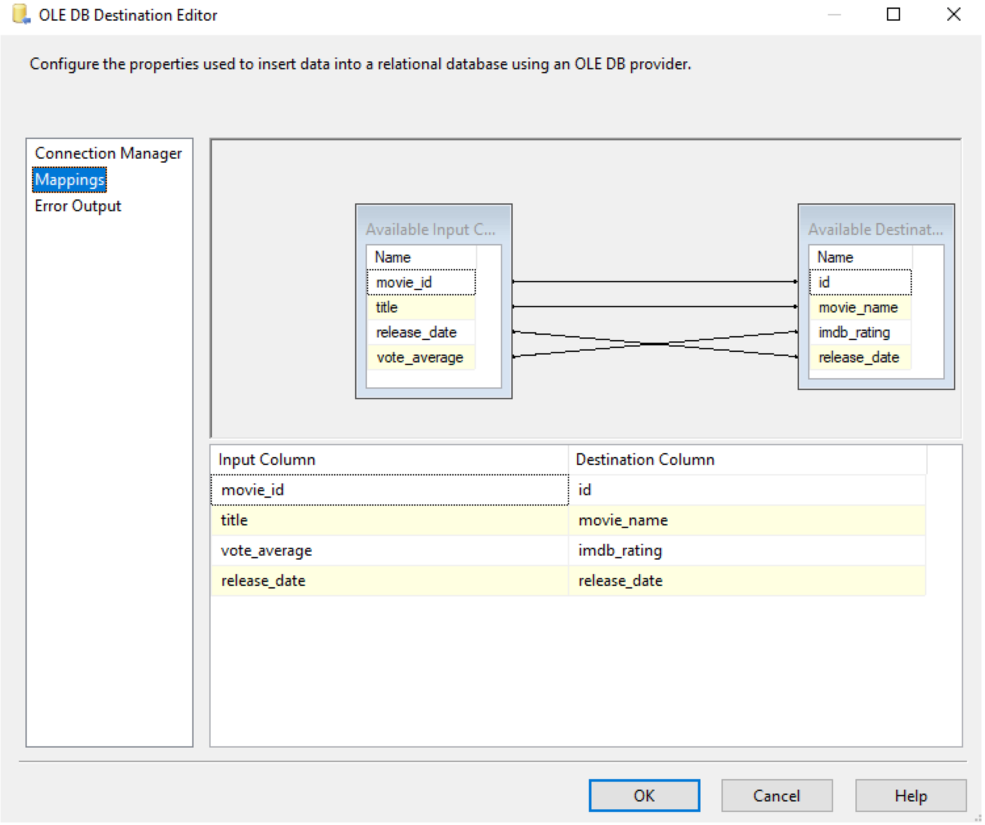
□ Table lock

□ Keep nulls

□ Check constraints

Rows per batch:

Maximum insert commit size: 2147483647



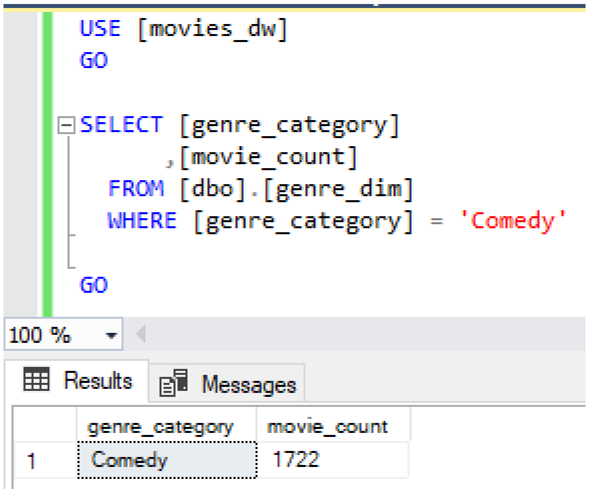
OLE DB Destination Editor

Configure the properties used to insert data into a relational database using an OLE DB provider.

Input Column	Destination Column
movie_id	id
title	movie_name
vote_average	imdb_rating
release_date	release_date

Business Queries:

- No of Movies in a Genre:



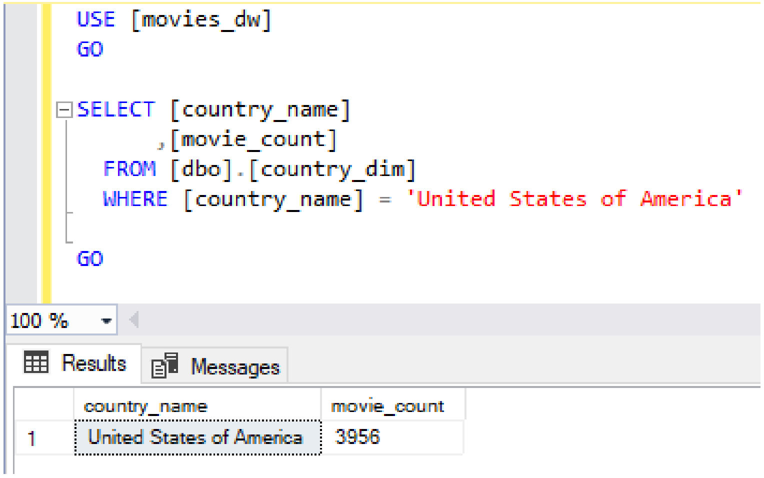
```
USE [movies_dw]
GO

SELECT [genre_category]
      ,[movie_count]
  FROM [dbo].[genre_dim]
  WHERE [genre_category] = 'Comedy'

GO
```

	genre_category	movie_count
1	Comedy	1722

- No of Movies Produced in a Country:



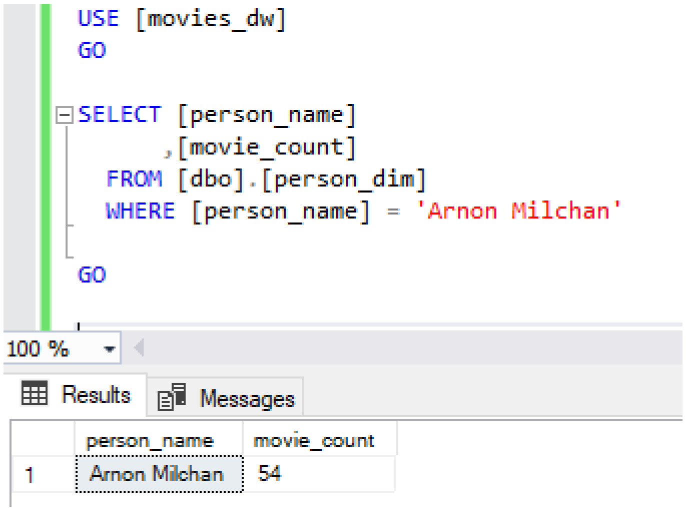
```
USE [movies_dw]
GO

SELECT [country_name]
      ,[movie_count]
  FROM [dbo].[country_dim]
  WHERE [country_name] = 'United States of America'

GO
```

	country_name	movie_count
1	United States of America	3956

- No of Movies respective of Cast:



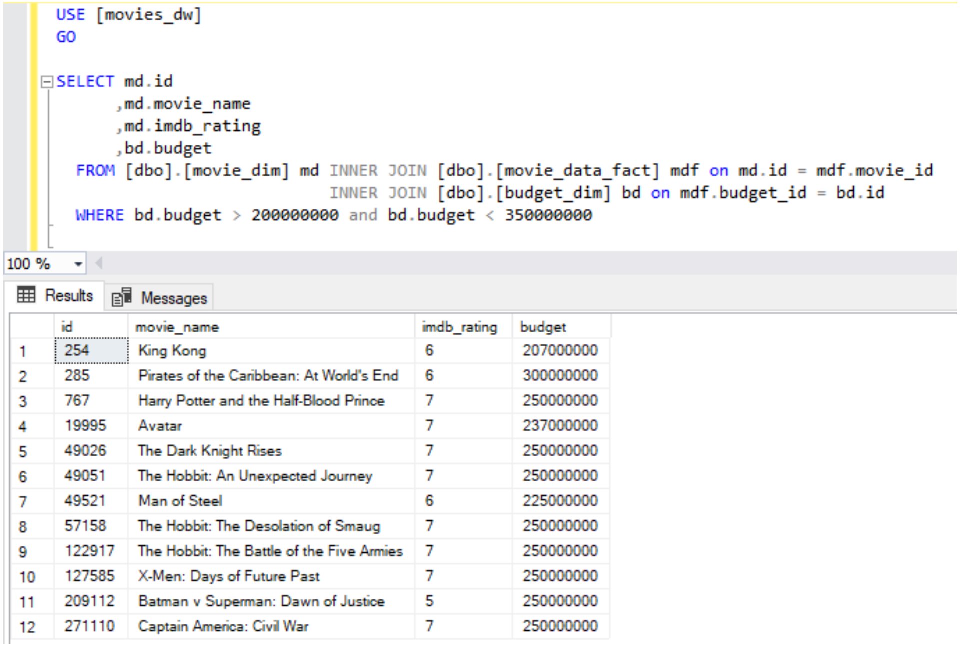
```
USE [movies_dw]
GO

SELECT [person_name]
      ,[movie_count]
  FROM [dbo].[person_dim]
  WHERE [person_name] = 'Arnon Milchan'

GO
```

	person_name	movie_count
1	Arnon Milchan	54

- **Movies in a Specific Range of Budget:**

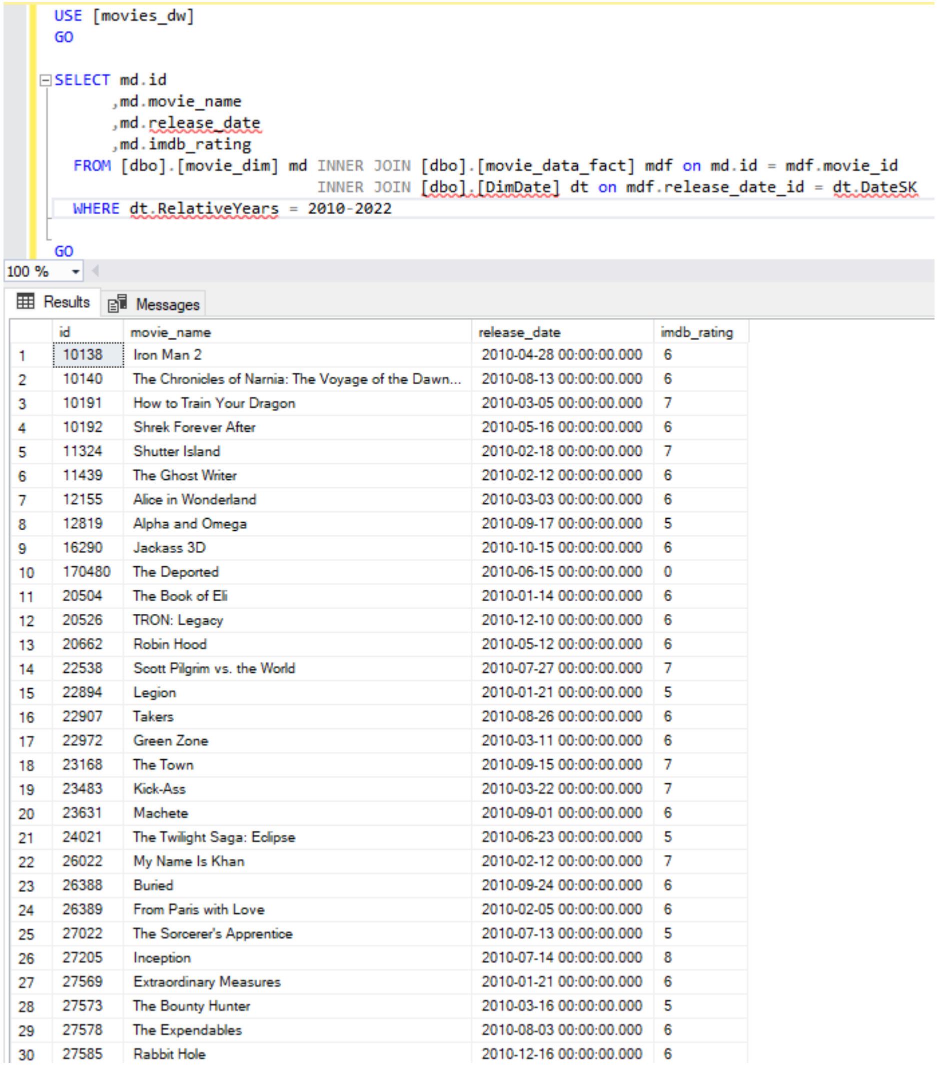


```
USE [movies_dw]
GO

SELECT md.id
      ,md.movie_name
      ,md.imdb_rating
      ,bd.budget
  FROM [dbo].[movie_dim] md INNER JOIN [dbo].[movie_data_fact] mdf on md.id = mdf.movie_id
                            INNER JOIN [dbo].[budget_dim] bd on mdf.budget_id = bd.id
  WHERE bd.budget > 200000000 and bd.budget < 350000000
```

	id	movie_name	imdb_rating	budget
1	254	King Kong	6	207000000
2	285	Pirates of the Caribbean: At World's End	6	300000000
3	767	Harry Potter and the Half-Blood Prince	7	250000000
4	19995	Avatar	7	237000000
5	49026	The Dark Knight Rises	7	250000000
6	49051	The Hobbit: An Unexpected Journey	7	250000000
7	49521	Man of Steel	6	225000000
8	57158	The Hobbit: The Desolation of Smaug	7	250000000
9	122917	The Hobbit: The Battle of the Five Armies	7	250000000
10	127585	X-Men: Days of Future Past	7	250000000
11	209112	Batman v Superman: Dawn of Justice	5	250000000
12	271110	Captain America: Civil War	7	250000000

- **Movies Released in a Specific Year:**



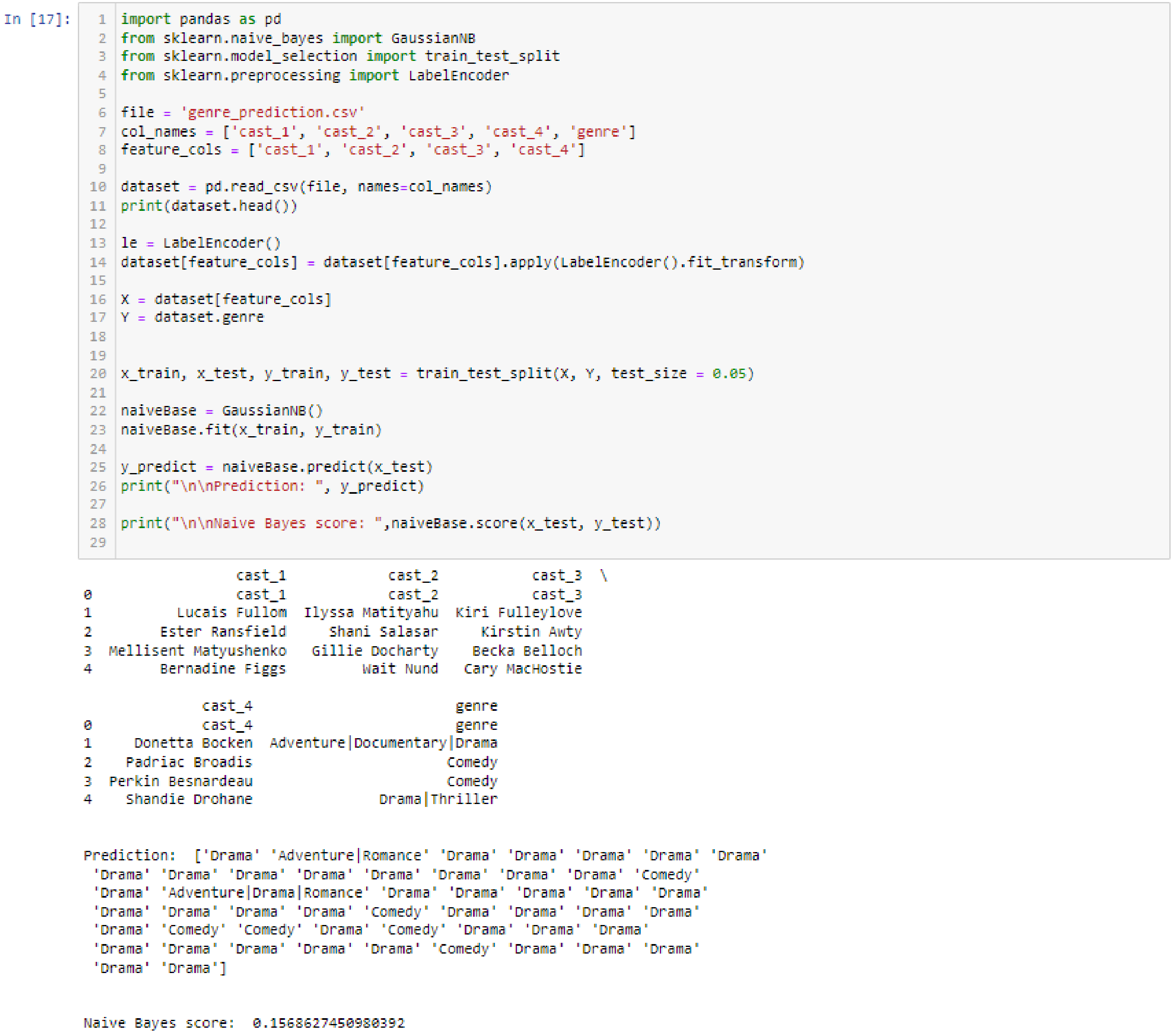
```
USE [movies_dw]
GO

SELECT md.id
      ,md.movie_name
      ,md.release_date
      ,md.imdb_rating
  FROM [dbo].[movie_dim] md INNER JOIN [dbo].[movie_data_fact] mdf on md.id = mdf.movie_id
                            INNER JOIN [dbo].[DimDate] dt on mdf.release_date_id = dt.DateSK
  WHERE dt.RelativeYears = 2010-2022

GO
```

	id	movie_name	release_date	imdb_rating
1	10138	Iron Man 2	2010-04-28 00:00:00.000	6
2	10140	The Chronicles of Narnia: The Voyage of the Dawn...	2010-08-13 00:00:00.000	6
3	10191	How to Train Your Dragon	2010-03-05 00:00:00.000	7
4	10192	Shrek Forever After	2010-05-16 00:00:00.000	6
5	11324	Shutter Island	2010-02-18 00:00:00.000	7
6	11439	The Ghost Writer	2010-02-12 00:00:00.000	6
7	12155	Alice in Wonderland	2010-03-03 00:00:00.000	6
8	12819	Alpha and Omega	2010-09-17 00:00:00.000	5
9	16290	Jackass 3D	2010-10-15 00:00:00.000	6
10	170480	The Deported	2010-06-15 00:00:00.000	0
11	20504	The Book of Eli	2010-01-14 00:00:00.000	6
12	20526	TRON: Legacy	2010-12-10 00:00:00.000	6
13	20662	Robin Hood	2010-05-12 00:00:00.000	6
14	22538	Scott Pilgrim vs. the World	2010-07-27 00:00:00.000	7
15	22894	Legion	2010-01-21 00:00:00.000	5
16	22907	Takers	2010-08-26 00:00:00.000	6
17	22972	Green Zone	2010-03-11 00:00:00.000	6
18	23168	The Town	2010-09-15 00:00:00.000	7
19	23483	Kick-Ass	2010-03-22 00:00:00.000	7
20	23631	Machete	2010-09-01 00:00:00.000	6
21	24021	The Twilight Saga: Eclipse	2010-06-23 00:00:00.000	5
22	26022	My Name Is Khan	2010-02-12 00:00:00.000	7
23	26388	Buried	2010-09-24 00:00:00.000	6
24	26389	From Paris with Love	2010-02-05 00:00:00.000	6
25	27022	The Sorcerer's Apprentice	2010-07-13 00:00:00.000	5
26	27205	Inception	2010-07-14 00:00:00.000	8
27	27569	Extraordinary Measures	2010-01-21 00:00:00.000	6
28	27573	The Bounty Hunter	2010-03-16 00:00:00.000	5
29	27578	The Expendables	2010-08-03 00:00:00.000	6
30	27585	Rabbit Hole	2010-12-16 00:00:00.000	6

NAÏVE BAYES:



In [17]:

```python
import pandas as pd
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder

file = 'genre_prediction.csv'
col_names = ['cast_1', 'cast_2', 'cast_3', 'cast_4', 'genre']
feature_cols = ['cast_1', 'cast_2', 'cast_3', 'cast_4']

dataset = pd.read_csv(file, names=col_names)
print(dataset.head())

le = LabelEncoder()
dataset[feature_cols] = dataset[feature_cols].apply(LabelEncoder().fit_transform)

X = dataset[feature_cols]
Y = dataset.genre


x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size = 0.05)

naiveBase = GaussianNB()
naiveBase.fit(x_train, y_train)

y_predict = naiveBase.predict(x_test)
print("\n\nPrediction: ", y_predict)

print("\n\nNaive Bayes score: ",naiveBase.score(x_test, y_test))

```

```
                  cast_1            cast_2           cast_3  \
0                 cast_1            cast_2           cast_3
1          Lucais Fullom  Ilyssa Matityahu  Kiri Fulleylove
2        Ester Ransfield     Shani Salasar     Kirstin Awty
3  Mellisent Matyushenko   Gillie Docharty    Becka Belloch
4        Bernadine Figgs         Wait Nund   Cary MacHostie

              cast_4                        genre
0             cast_4                        genre
1     Donetta Bocken  Adventure|Documentary|Drama
2    Padriac Broadis                       Comedy
3  Perkin Besnardeau                       Comedy
4    Shandie Drohane               Drama|Thriller


Prediction:  ['Drama' 'Adventure|Romance' 'Drama' 'Drama' 'Drama' 'Drama' 'Drama'
 'Drama' 'Drama' 'Drama' 'Drama' 'Drama' 'Drama' 'Drama' 'Drama' 'Comedy'
 'Drama' 'Adventure|Drama|Romance' 'Drama' 'Drama' 'Drama' 'Drama' 'Drama'
 'Drama' 'Drama' 'Drama' 'Drama' 'Comedy' 'Drama' 'Drama' 'Drama' 'Drama'
 'Drama' 'Comedy' 'Comedy' 'Drama' 'Comedy' 'Drama' 'Drama' 'Drama'
 'Drama' 'Drama' 'Drama' 'Drama' 'Drama' 'Comedy' 'Drama' 'Drama' 'Drama'
 'Drama' 'Drama']


Naive Bayes score:  0.1568627450980392
```