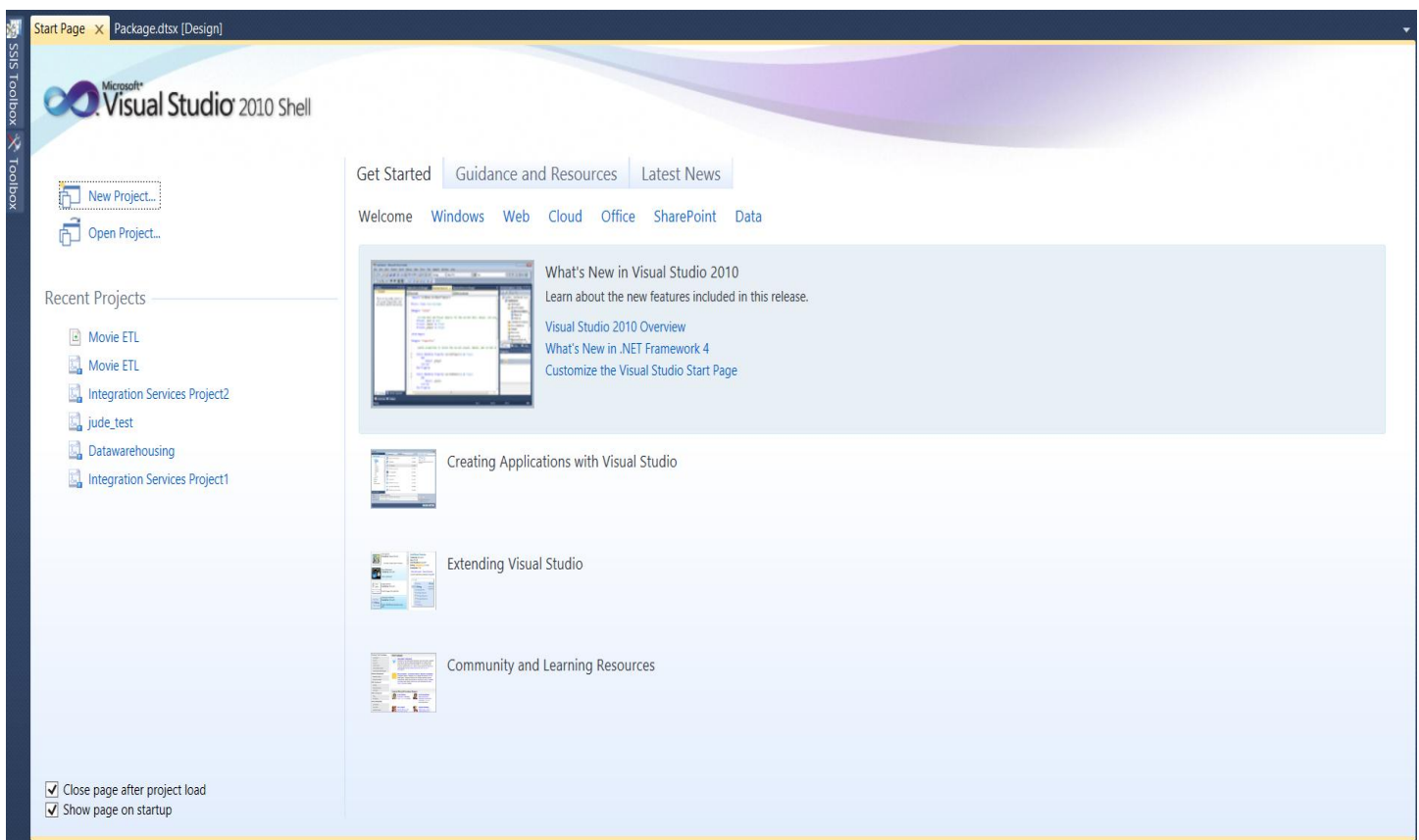


# “Movie Management System”

## ETL of the Application

---

*Mohamed Niyaz*



## 1. ETL Introduction

The term ETL stands for Extract, Transform and Load. ETL process involves extracting the data from the source systems. In many cases this is the most challenging aspect of ETL, since extracting data correctly sets the stage for how subsequent processes go further.

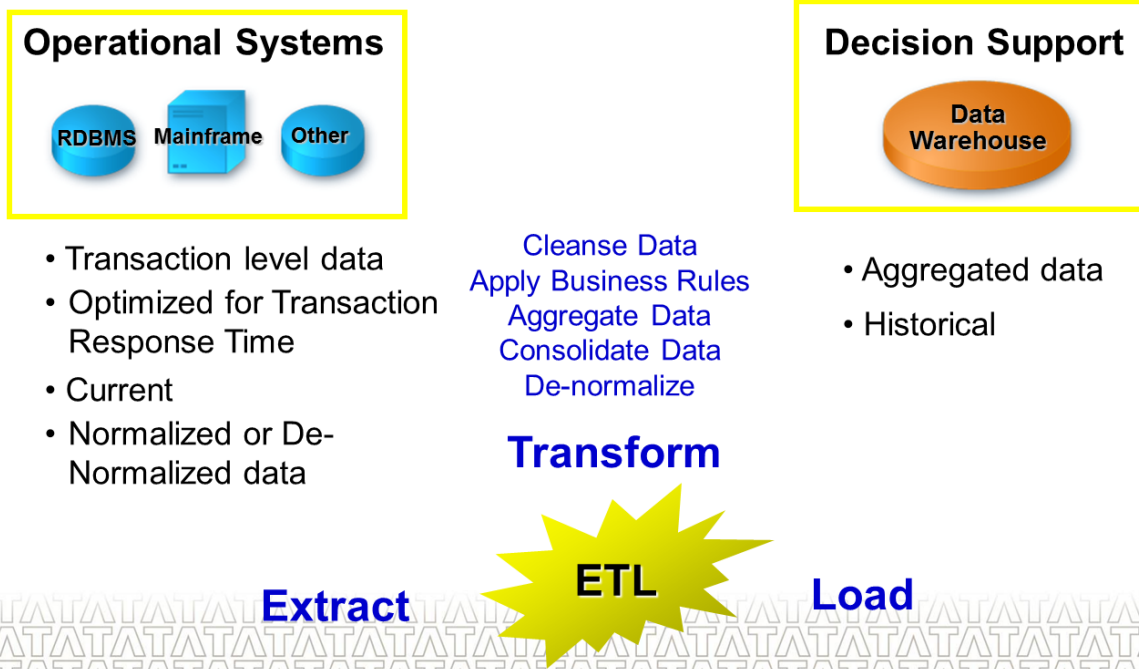
Most data warehousing projects consolidate data from different source systems. Each separate system may also use a different data organization/format. Common data source formats are relational databases and flat files, but may include non-relational database structures such as Information Management System (IMS) or other data structures such as Virtual Storage Access Method (VSAM) or Indexed Sequential Access Method (ISAM), or even fetching from outside sources such as through screen-scraping. In general, the goal of the extraction phase is to convert the data into a single format which is appropriate for transformation processing. An intrinsic part of the extraction involves the parsing of extracted data, resulting in a check if the data meets an expected pattern or structure. If not, the data may be rejected entirely or in part.

- **Source** : Amstat, Knomarix, autolab, amazon, myrrix.
- **File type** : Comma Separated files, Text files with Standard Delimiters
- **Extraction** : Files are extracted from the sources using a java batch files to a windows directory.
- **Transformation**: Files are cleansed and parsed through a set of validation processes. On completion of the task a cleansed file is create and moved to the inbound directory.
- **Loading** : The Files are picked from the inbound directory and loaded into the Microsoft SQL database through the built in Sql services.

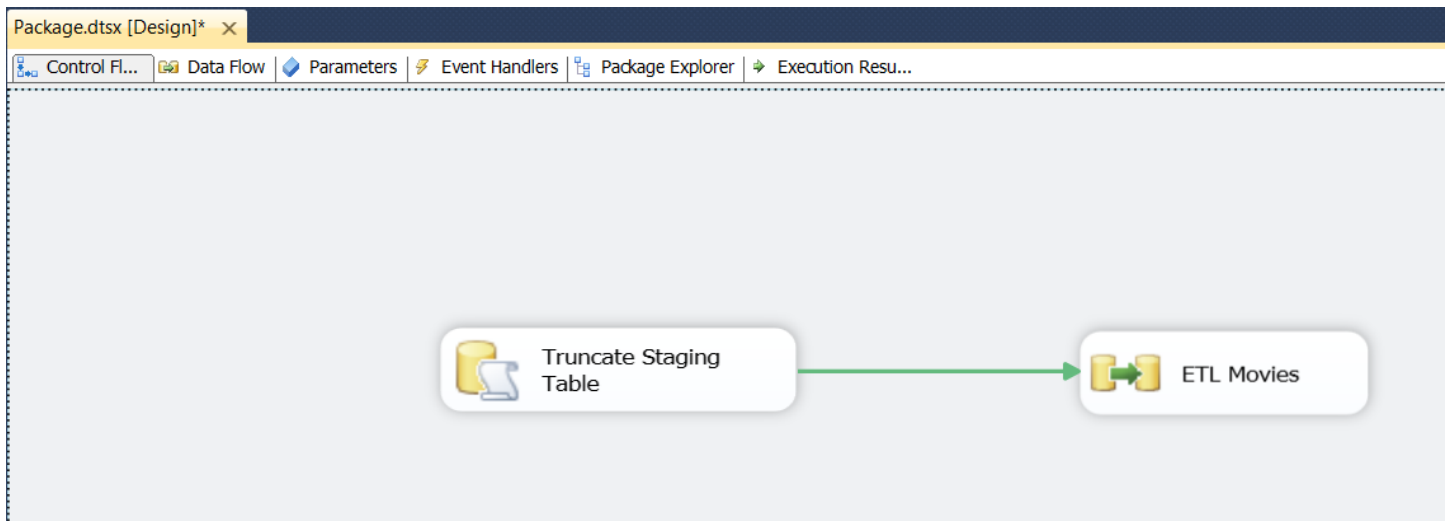
## 2. ETL Design

### Extract, Transform, and Load

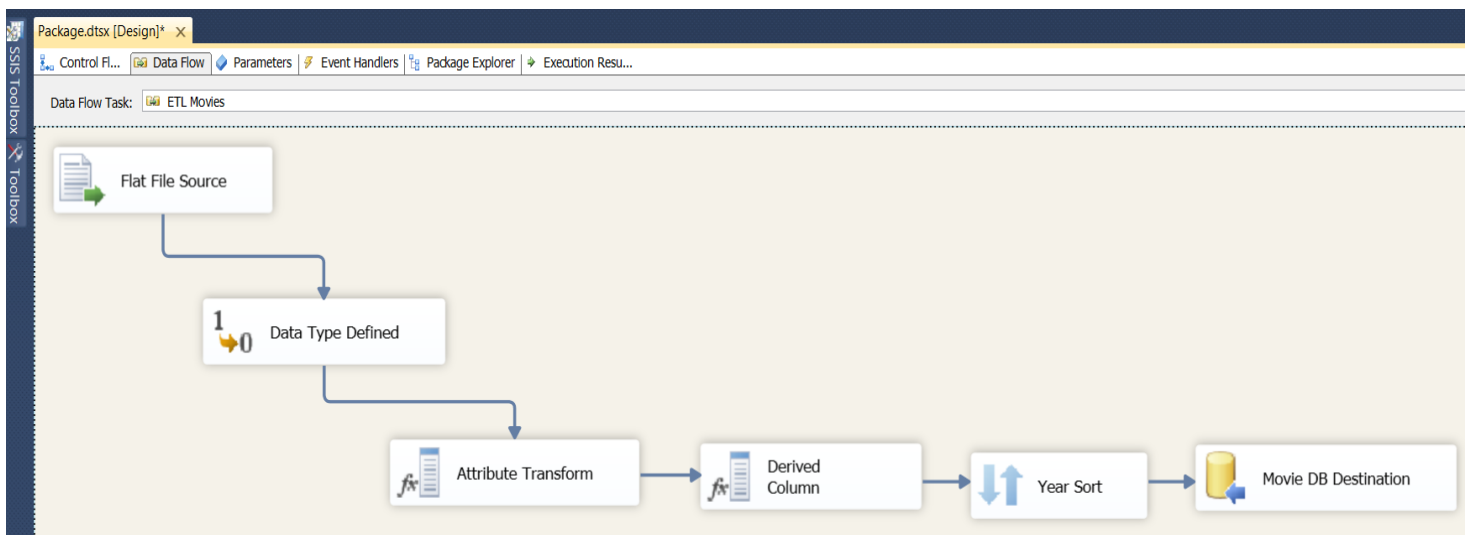
---



Our ETL Process is carried through **Microsoft SSIS** tool and below screen shot explains the high level overview of the **Control Task** ETL.



The below Screenshot depicts **data flow task** which is part of ETL process

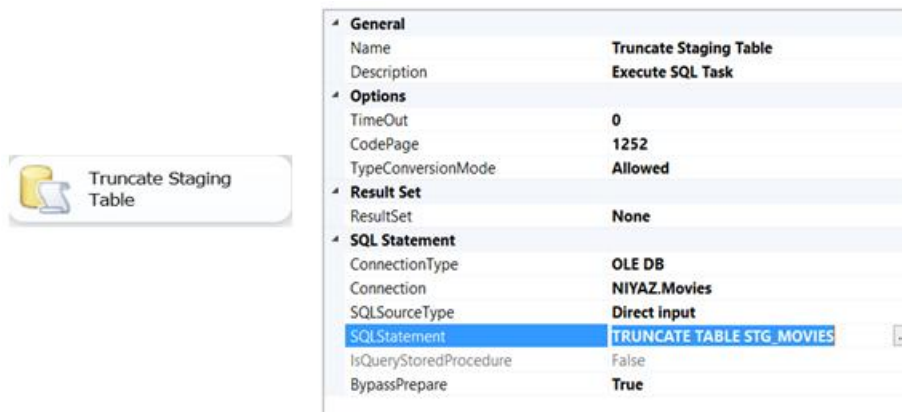


Source: MovieFile.csv from local pc.

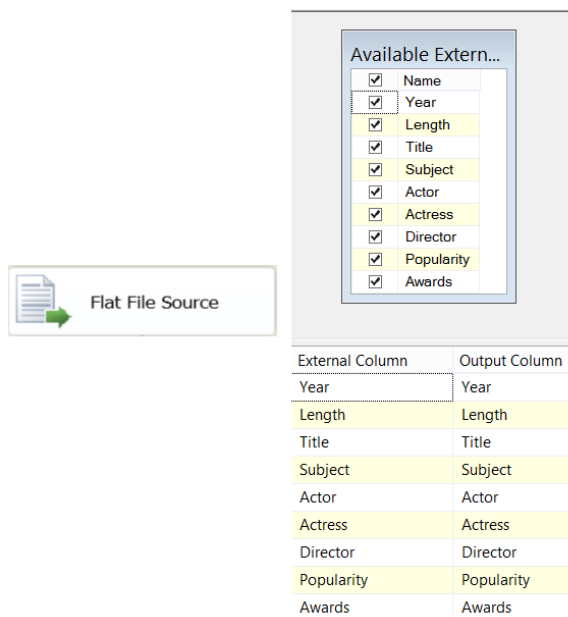
Destination: Movie database (Sql).

### 3. ETL Process

**Step1:** The initial ETL process commences by truncating the STG\_MOVIES table at Movie Database




**Step2:** A Data flow task is created and the Source connection is established.



The above screenshot explains the number of columns taken as an input and output.

**Step3:** All the columns imported from CSV file are string and the necessary data conversion is taken on depending on the columns. Here year, popularity, length are converted into INT.


**1**  Data Type Defined

Available Input ...

- ☒ Name
- ☒ Year
- ☒ Length
- ☒ Title
- ☒ Subject
- ☒ Actor
- ☒ Actress
- ☒ Director
- ☒ Popularity
- ☒ Awards

| Input Column | Output Alias  | Data Type            | Length | Precision | Scale | Code Page             |
|--------------|---------------|----------------------|--------|-----------|-------|-----------------------|
| Actor        | Actor_in      | string [DT_STR]      | 300    |           |       | 1252 (ANSI - Latin I) |
| Actress      | Actress_in    | string [DT_STR]      | 300    |           |       | 1252 (ANSI - Latin I) |
| Awards       | Awards_in     | string [DT_STR]      | 300    |           |       | 1252 (ANSI - Latin I) |
| Director     | Director_in   | string [DT_STR]      | 300    |           |       | 1252 (ANSI - Latin I) |
| Length       | Length_in     | Unicode string [D... | 300    |           |       |                       |
| Popularity   | Popularity_in | four-byte unsigne... |        |           |       |                       |
| Subject      | Subject_in    | string [DT_STR]      | 300    |           |       | 1252 (ANSI - Latin I) |
| Title        | Title_in      | string [DT_STR]      | 3000   |           |       | 1252 (ANSI - Latin I) |
| Year         | Year_in       | four-byte unsigne... |        |           |       |                       |

**Step4:** The data are validated and transform to business needs. A new column attribute date is added using GETDATE() function. All the columns are trimmed using LTRIM and RTRIM functions. The Null in each column are replaced using ISNULL, logical and conditional operators.

**fx**  Attribute Transform


Variables and Parameters  
Columns

Mathematical Functions  
String Functions  
Date/Time Functions  
NULL Functions  
Type Casts  
Operators

Description:

| Derived Column Name | Derived Column        | Expression  | Data Type                |
|---------------------|-----------------------|---|--------------------------|
| Date                | <add as new column>   | GETDATE()   | database timestamp [D... |
| Length_o            | <add as new column>   | ISNULL(Length_in)    (Length_in == "") ? "0" : Length_... | Unicode string [DT_WS... |
| Actor_in            | Replace 'Actor_in'    | LTRIM(RTRIM(UPPER(Actor_in)))                             | string [DT_STR]          |
| Actress_in          | Replace 'Actress_in'  | LTRIM(RTRIM(UPPER(Actress_in)))                           | string [DT_STR]          |
| Director_in         | Replace 'Director_in' | LTRIM(RTRIM(UPPER(Director_in)))                          | string [DT_STR]          |

**Step5:** The Null in each column are replaced using ISNULL, logical and conditional operators. A default value is provided in each query.

 Null Replacer

Variables and Parameters

Columns

Mathematical Functions

String Functions

Date/Time Functions

NULL Functions


Type Casts

Operators

Description:

| Derived Column Name | Derived Column      | Expression  | Data Type                | Length |
|---------------------|---------------------|---|--------------------------|--------|
| Actor_o             | <add as new column> | ISNULL(Actor_in)    (Actor_in == ""    Actor_in == "MISSING") ? "MANY" : Actor_in             | Unicode string [DT_WS... | 300    |
| Actress_o           | <add as new column> | ISNULL(Actress_in)    (Actress_in == ""    Actress_in == "MISSING") ? "MANY" : Actress_in     | Unicode string [DT_WS... | 300    |
| Director_o          | <add as new column> | ISNULL(Director_in)    (Director_in == ""    Director_in == "MISSING") ? "MANY" : Director_in | Unicode string [DT_WS... | 300    |

**Step6:** All Columns are sorted in ascending order using the Year column. The data's after this transformation are sorted.

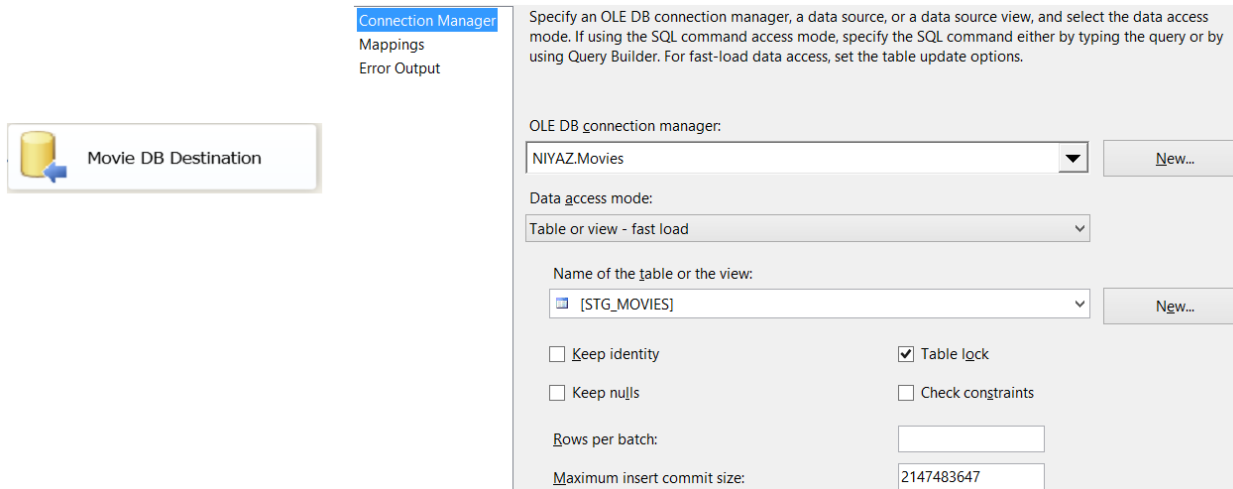
 Year Sort

Available Input Columns

| <input type="checkbox"/>            | Name         | Pass Through                        |
|-------------------------------------|--------------|-------------------------------------|
| <input checked="" type="checkbox"/> | Year         | <input type="checkbox"/>            |
| <input type="checkbox"/>            | Length       | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Title        | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Subject      | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Actor        | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Actress      | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Director     | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Popularity   | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Awards       | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Actor_in     | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Actress_in   | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Awards_in    | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Director_in  | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Length_in    | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Popularit... | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Subject_in   | <input checked="" type="checkbox"/> |
| <input type="checkbox"/>            | Title_in     | <input checked="" type="checkbox"/> |

| Input Column | Output Alias | Sort Type | Sort Order |
|--------------|--------------|-----------|------------|
| Year         | Year         | ascending | 1          |

**Step7:** The final step is to load the data's into the target destination which is Movies database. The connections are created using OLE DB drivers. The tool allows you to create a table directly or we can create a table in the DB directly.



Connection Manager

Mappings

Error Output

Movie DB Destination

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:  
NIYAZ.Movies

Data access mode:  
Table or view - fast load

Name of the table or the view:  
[STG\_MOVIES]

☐ Keep identity ☒ Table lock

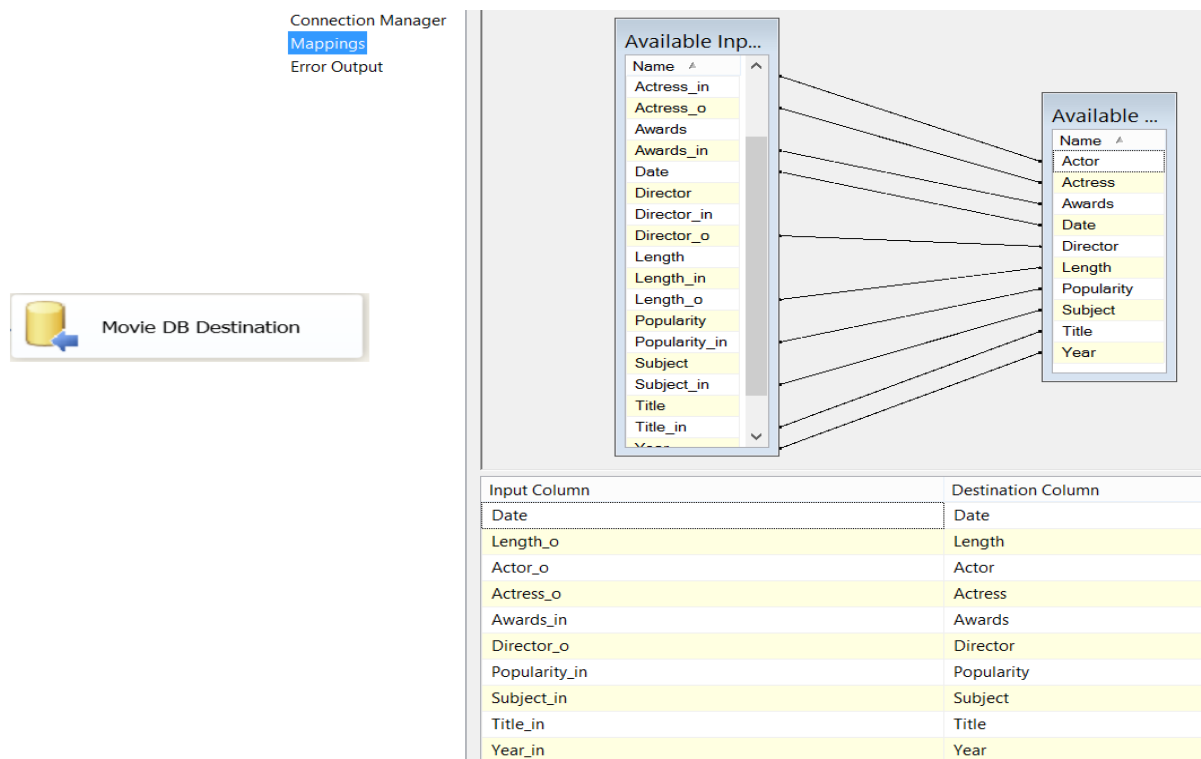
☐ Keep nulls ☐ Check constraints

Rows per batch:

Maximum insert commit size: 2147483647

The above screenshot explains the parameters used in target connection.

**Step8:** The input columns are chosen and mapped to the output columns



Connection Manager

Mappings

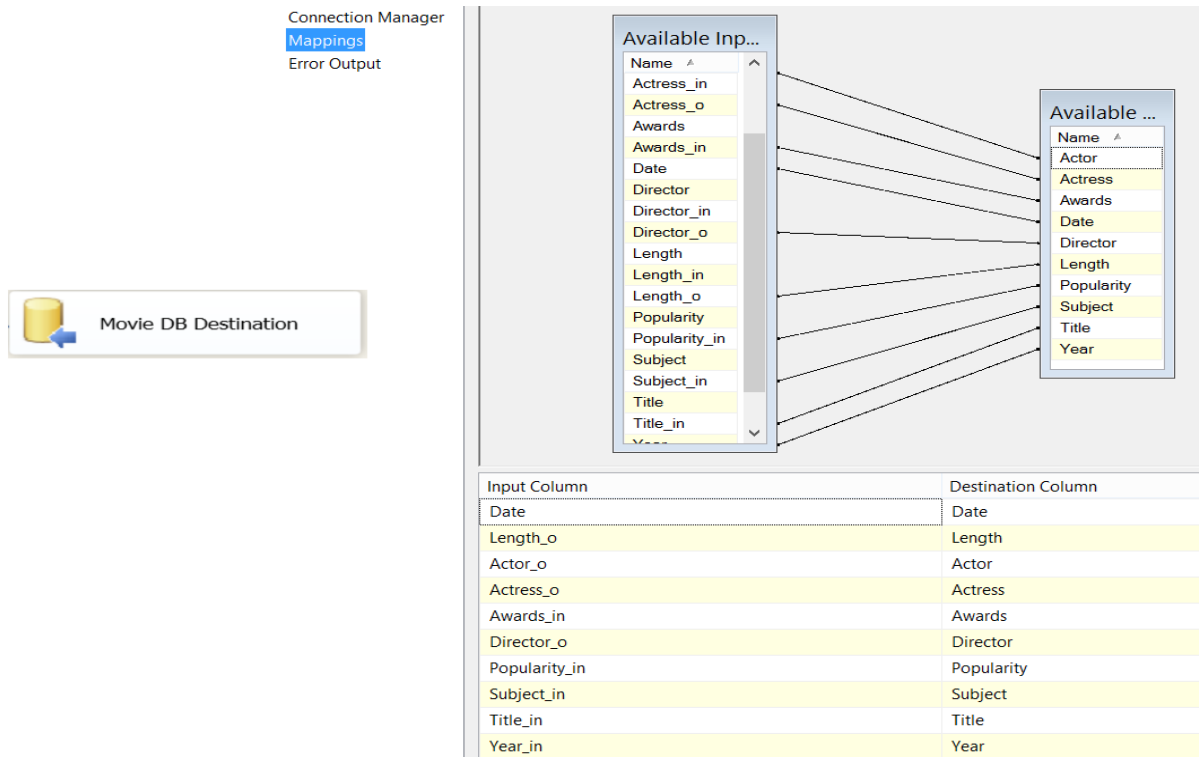
Error Output

Movie DB Destination

Available Input

Available Output

| Input Column  | Destination Column |
|---------------|--------------------|
| Date          | Date               |
| Length_o      | Length             |
| Actor_o       | Actor              |
| Actress_o     | Actress            |
| Awards_in     | Awards             |
| Director_o    | Director           |
| Popularity_in | Popularity         |
| Subject_in    | Subject            |
| Title_in      | Title              |
| Year_in       | Year               |

**Step9:** The input columns are chosen and mapped to the output columns**Step9:** The Source file in local disk before ETL

|    | A    | B      | C                                     | D       | E                  | F                   | G                         | H          | I      |
|----|------|--------|---------------------------------------|---------|--------------------|---------------------|---------------------------|------------|--------|
| 1  | Year | Length | Title                                 | Subject | Actor              | Actress             | Director                  | Popularity | Awards |
| 2  | 1990 | 125    | Wild at Heart                         | Drama   | Cage Nicolas       | Dern Laura          | Lynch David               | 6          | No     |
| 3  | 1961 | 120    | Goodbye Again                         | Drama   | Perkins Anthony    | Bergman Ingrid      | Litvak Anatole            | 6          | No     |
| 4  | 1990 | 135    | Hunt for Red October The              | Drama   | Connery Sean       |                     | McTiernan J               | 8          | No     |
| 5  | 1984 | 108    | Terminator The                        | Action  | Schwarzenegger A   | Hamilton Linda      | Cameron J                 | 17         | No     |
| 6  | 1991 | 136    | Terminator 2                          | Action  | Schwarzenegger A   | Hamilton Linda      | Cameron J                 | 8          | No     |
| 7  | 1993 | 65     | John Cleese on How to Irritate People | Comedy  | Cleese John        | Booth Connie        |                           | 62         | No     |
| 8  | 1987 | 103    | Au Revoir les Enfants                 | Drama   | Manesse Gaspard    | Racette Francine    | Malle Louis               | 35         | No     |
| 9  | 1983 | 128    | The Ballad of Narayama                | Drama   |                    | Missing             | Imamura Shohei            | 15         | No     |
| 10 | 1990 | 138    | Cyrano De Bergerac                    | Drama   | Depardieu Gerard   | Brochet Anne        | Rappeneau Jean-Paul       | 86         | No     |
| 11 | 1990 | 107    | Green Card                            | Comedy  | Depardieu Gerard   | MacDowell Andie     | Weir Peter                | 25         | No     |
| 12 | 1987 | 118    | Hope & Glory                          | War     | Hayman David       | Miles Sarah         | Boorman John              | 3          | No     |
| 13 | 1982 | 122    | Missing                               | Drama   | Lemmon Jack        | Spacek Sissy        | Costa-Gavras              | 30         | No     |
| 14 | 1986 | 125    | The Mission                           | Drama   | Niro Robert De     | Lunghi Cherie       | Joffe Roland              | 20         | No     |
| 15 | 1987 | 101    | My Life As a Dog                      | Comedy  | Glanzelius Anton   |                     | Hallstrom Lasse           | 21         | No     |
| 16 | 1984 | 150    | Paris Texas                           | Drama   | Stanton Harry Dean | Kinski Nastassia    | Wim Wenders               | 27         | No     |
| 17 | 1984 | 106    | Romancing the Stone                   | Action  | Douglas Michael    | Turner Kathleen     | Silvestri Robert Zemeckis | 83         | No     |
| 18 | 1982 | 120    | The State of Things                   | Drama   |                    | Isabelle Weingarten | Wenders Wim               | 40         | No     |
| 19 | 1986 | 98     | Summer                                | Comedy  | Gauthier Vincent   | Riviere Marie       | Rohmer Eric               | 11         | No     |
| 20 | 1955 | 108    | Smiles of a Summer Night              | Comedy  | Bjornstrand Gunnar | Jacobsson Ulla      | Bergman Ingmar            | 58         | No     |
| 21 | 1987 | 98     | Under the Sun of Satan                | Drama   | Depardieu Gerard   | Bonnaire Sandrine   | Pialat Maurice            | 45         | No     |
| 22 | 1985 | 105    | Vagabond                              | Drama   | Meril Macha        | Bonnaire Sandrine   | Varda Agnes               | 49         | No     |
| 23 | 1988 | 115    | Working Girl                          | Comedy  | Ford Harrison      | Griffith Melanie    | Nichols Mike              | 25         | No     |
| 24 | 1984 | 106    | A Year of the Quiet Sun               | Drama   | Wilson Scott       | Komorowska Maja     | Zanussi Krzystoff         | 78         | No     |



**Step10:** The Source file in database after ETL through Microsoft SSIS. The final modulated data's in the database.

The screenshot shows the Microsoft SQL Server Enterprise Manager interface. The left pane displays the 'Object Explorer' for a server named 'NIYAZ (SQL Server 11.0.2218 - NIYAZ\zackr)'. The right pane shows a query window titled 'SQLQuery1.sql - NIYAZ\zackriyaniyaz (55)' containing the following SQL script:

```

/***** Script for SelectTopNRows command from SSMS *****/
SELECT TOP 1000 [Actor]
      ,[Actress]
      ,[Awards]
      ,[Director]
      ,[Length]
      ,[Popularity]
      ,[Subject]
      ,[Title]
      ,[Year]
      ,[Date]
FROM [Movies].[dbo].[STG_MOVIES]
  
```

Below the query window, the 'Results' pane displays the output of the query as a table with 10 columns: Actor, Actress, Awards, Director, Length, Popularity, Subject, Title, Year, and Date. The results show the top 10 rows of data from the STG\_MOVIES table.

|   | Actor            | Actress          | Awards | Director         | Length | Popularity | Subject         | Title                        | Year | Date                    |
|---|------------------|------------------|--------|------------------|--------|------------|-----------------|------------------------------|------|-------------------------|
| 1 | SCHWARZENEGGER A | BERGMAN SANDAHL  | No     | MILIUS JOHN      | 128    | 45         | Action          | Conan the Barbarian          | 1982 | 2013-03-15 02:19:13.237 |
| 2 | HOPKINS ANTHONY  | MANY             | No     | MANY             | 208    | 84         | Drama           | Othello                      | 1982 | 2013-03-15 02:19:13.237 |
| 3 | DUVALL ROBERT    | HARPER TESS      | Yes    | BERESFORD BRUCE  | 93     | 61         | Drama           | Tender Mercies               | 1983 | 2013-03-15 02:19:13.247 |
| 4 | RUSSELL KURT     | STREEP MERYL     | No     | NICHOLS MIKE     | 131    | 52         | Drama           | Silkwood                     | 1983 | 2013-03-15 02:19:13.237 |
| 5 | NICHOLSON JACK   | MACLAINE SHIRLEY | Yes    | L JAMES          | 132    | 32         | Drama           | Terms of Endearment          | 1983 | 2013-03-15 02:19:13.247 |
| 6 | MURPHY EDDIE     | BERNHARD SANDRA  | No     | MANY             | 60     | 20         | Comedy          | Best of the Big Laff Off The | 1983 | 2013-03-15 02:19:13.237 |
| 7 | HAMILL MARK      | FISHER CARRIE    | No     | MARQUAND RICHARD | 132    | 4          | Science Fiction | Return of the Jedi           | 1983 | 2013-03-15 02:19:13.233 |
| 8 | JOLIVET PIERRE   | MANY             | No     | BESSON LUC       | 90     | 72         | Drama           | Le Dernier Combat            | 1983 | 2013-03-15 02:19:13.237 |
| 9 | CONNERY SEAN     | BASINGER KIM     | No     | KERSHNER IRVIN   | 134    | 8          | Action          | Never Say Never Again        | 1983 | 2013-03-15 02:19:13.233 |