Data Warehousing & Business Intelligence (IT)

3rd Year, 1st Semester

Assignment 1

Submitted to

Sri Lanka Institute of Information Technology

IT20126124

Kumarasinghe S

Weekday Batch

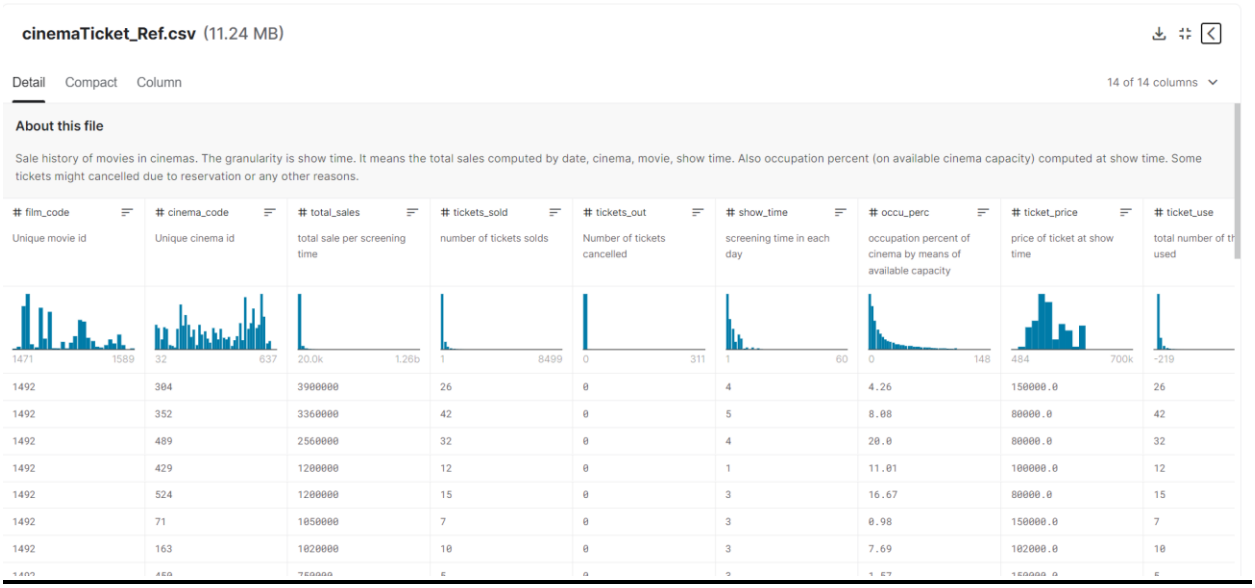
Contents

Step 01: Data set selection

I have selected Kaggle cinema tickets as data set. It consists of one CSV fie with sufficient data with 14 columns. Furthermore, I have partitioned the main large CSV file into small sub-CSV files. The sub-CSV files consist of new IDs. And, I have manually modified some data records according to the requirements.

The data set was initiated with sufficient data, according to the assignment criteria. It has more the 10,000 unique values and it is enriched with transactional data and data hierarchies.

Data Set https://www.kaggle.com/datasets/arashnic/cinema-ticket



cinemaTicket_Ref.csv (11.24 MB)

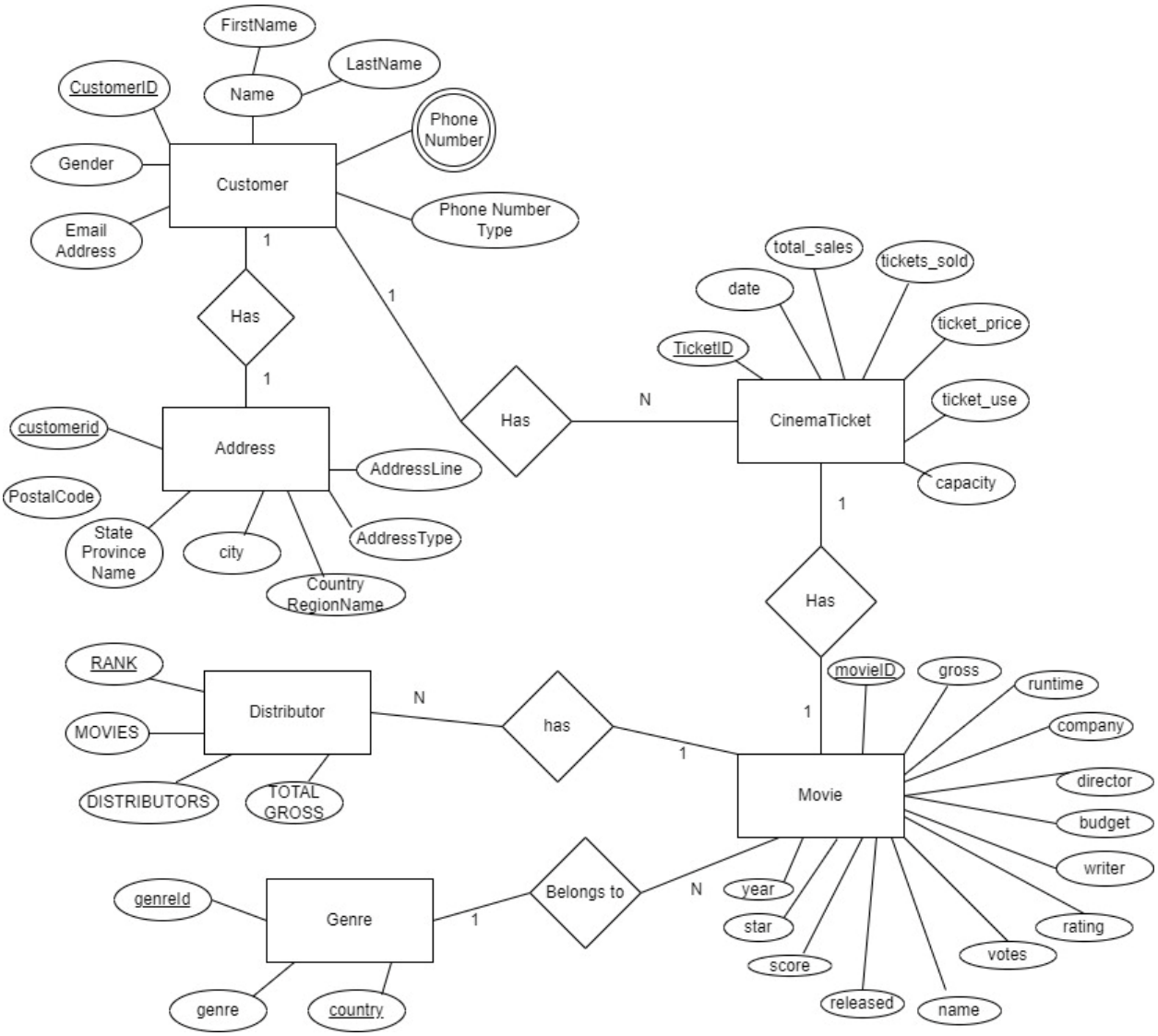
Detail | Compact | Column — 14 of 14 columns

About this file

Sale history of movies in cinemas. The granularity is show time. It means the total sales computed by date, cinema, movie, show time. Also occupation percent (on available cinema capacity) computed at show time. Some tickets might cancelled due to reservation or any other reasons.

# film_code	# cinema_code	# total_sales	# tickets_sold	# tickets_out	# show_time	# occu_perc	# ticket_price	# ticket_use
Unique movie id	Unique cinema id	total sale per screening time	number of tickets solds	Number of tickets cancelled	screening time in each day	occupation percent of cinema by means of available capacity	price of ticket at show time	total number of th used
1471 – 1589	32 – 637	20.0k – 1.26b	1 – 8499	0 – 311	1 – 60	0 – 148	484 – 700k	-219
1492	304	3900000	26	0	4	4.26	150000.0	26
1492	352	3360000	42	0	5	8.08	80000.0	42
1492	489	2560000	32	0	4	20.0	80000.0	32
1492	429	1200000	12	0	1	11.01	100000.0	12
1492	524	1200000	15	0	3	16.67	80000.0	15
1492	71	1050000	7	0	3	0.98	150000.0	7
1492	163	1020000	10	0	3	7.69	102000.0	10

ER Diagram -


FirstName
LastName
Name
CustomerID
Phone Number
Gender
Customer
Email Address
Phone Number Type
1
Has
1
1
Has
N
CinemaTicket
total_sales
tickets_sold
date
ticket_price
TicketID
ticket_use
capacity
customerid
Address
AddressLine
PostalCode
State Province Name
city
AddressType
Country RegionName
1
Has
RANK
Distributor
N
has
1
movieID
gross
runtime
company
MOVIES
DISTRIBUTORS
TOTAL GROSS
Movie
director
budget
writer
rating
votes
genreId
Genre
1
Belongs to
N
year
star
score
released
name
genre
country

Step 02: Preparation of Data Sources

First, main data set was separated into sub-CSV files (CinemaCustomer, cinemaTicke, movies, Genre and TopDistributors) and categorized related data into same csv files. Then csv files were imported into the tables which were in the newly created database called IT20126124_SourceDB (Except CustomerAddress)

And Customer Address details saved into text file format. This file contains all the customer Address Information.

	A	B	C	D	E	F	G	H	I
1	CustomerID	FirstName	LastName	Gender	PhoneNumber	PhoneNumberType	EmailAddress		
2	13643	Ronald	Sai	M	1 (11) 500 555-0174	Cell	ronald7@adventure-works.com		
3	29128	Alyssa	Reed	F	966-555-0118	Home	alyssa28@adventure-works.com		
4	26207	Kelvin	Zeng	M	1 (11) 500 555-0135	Cell	kelvin41@adventure-works.com		
5	12612	Shannon	Torres	M	1 (11) 500 555-0180	Cell	shannon33@adventure-works.com		
6	14776	Ronald	Arthur	M	1 (11) 500 555-0189	Home	ronald8@adventure-works.com		
7	18609	Alyssa	Cook	F	931-555-0112	Cell	alyssa29@adventure-works.com		
8	11664	Alyssa	Morgan	F	144-555-0113	Cell	alyssa30@adventure-works.com		
9	25828	Ronald	Madan	M	1 (11) 500 555-0180	Cell	ronald9@adventure-works.com		
10	16467	Ronald	Srini	M	1 (11) 500 555-0159	Cell	ronald10@adventure-works.com		
11	15677	Alyssa	Bradley	F	1 (11) 500 555-0127	Cell	alyssa31@adventure-works.com		
12	25072	Alyssa	Murphy	F	698-555-0138	Cell	alyssa32@adventure-works.com		
13	24383	Ronald	Prasad	M	1 (11) 500 555-0162	Home	ronald11@adventure-works.com		
14	26595	Alyssa	Bailey	F	715-555-0165	Home	alyssa33@adventure-works.com		
15	19901	Ronald	Sara	M	1 (11) 500 555-0154	Cell	ronald12@adventure-works.com		
16	26256	Ronald	Rana	M	722-555-0126	Cell	ronald13@adventure-works.com		
17	26195	Alyssa	Rivera	F	188-555-0142	Home	alyssa34@adventure-works.com		
18	17077	Ronald	Raman	M	1 (11) 500 555-0186	Home	ronald14@adventure-works.com		
19	12171	Alyssa	Cooper	F	491-555-0143	Cell	alyssa35@adventure-works.com		
20	24810	Alyssa	Richardson	F	387-555-0174	Home	alyssa36@adventure-works.com		
21	27733	Ronald	Subram	M	1 (11) 500 555-0148	Home	ronald15@adventure-works.com		
22	11083	Alyssa	Cox	F	561-555-0140	Cell	alyssa37@adventure-works.com		
23	11513	Alyssa	Howard	F	805-555-0188	Home	alyssa38@adventure-works.com		
24	17649	Ronald	Mehta	M	1 (11) 500 555-0138	Cell	ronald16@adventure-works.com		
25	19268	Alyssa	Ward	F	1 (11) 500 555-0196	Cell	alyssa39@adventure-works.com		
26	25720	Ronald	Garcia	M	1 (11) 500 555-0195	Home	ronald17@adventure-works.com		

CinemaCustomer.csv

	A	B	C	D	E	F	G	H	I	J
1	TicketID	total_sales	tickets_sold	ticket_price	ticket_use	capacity	date	CustomerID	movieID	RANKID
2	1001	3900000	26	150000	26	610.3286	5/5/2018	13643	501	1
3	1002	3360000	42	80000	42	519.802	5/5/2018	29128	502	2
4	1003	2560000	32	80000	32	160	5/5/2018	26207	503	3
5	1004	1200000	12	100000	12	108.9918	5/5/2018	12612	504	4
6	1005	1200000	15	80000	15	89.982	5/5/2018	14776	505	5
7	1006	1050000	7	150000	7	714.2857	5/5/2018	18609	506	6
8	1007	1020000	10	102000	10	130.039	5/5/2018	11664	507	7
9	1008	750000	5	150000	5	318.4713	5/5/2018	25828	508	8
10	1009	750000	11	68181.81818	11	1157.895	5/5/2018	16467	509	9
11	1010	600000	4	150000	4	258.0645	5/5/2018	15677	510	10
12	1011	480000	6	80000	6	1363.636	5/5/2018	25072	511	11
13	1012	480000	4	120000	4	135.1351	5/5/2018	24383	512	12
14	1013	400000	5	80000	5	943.3962	5/5/2018	26595	513	13
15	1014	300000	2	150000	2	800	5/5/2018	19901	514	14
16	1015	240000	2	120000	2	98.03922	5/5/2018	26256	515	15
17	1016	16500000	112	147321.4286	112	611.0202	5/4/2018	26195	516	16
18	1017	13950000	93	150000	93	879.8486	5/4/2018	17077	517	17
19	1018	10200000	68	150000	68	796.2529	5/4/2018	12171	518	18
20	1019	6600000	44	150000	44	716.6124	5/4/2018	24810	519	19
21	1020	3360000	31	108387.0968	31	125	5/4/2018	27733	520	20
22	1021	3000000	20	150000	20	258.0645	5/4/2018	11083	521	21
23	1022	2400000	16	150000	16	138.05	5/4/2018	11513	522	22
24	1023	1800000	12	150000	12	90.02251	5/4/2018	17649	523	23
25	1024	1680000	14	120000	13	165.0943	5/4/2018	19268	524	24
26	1025	1400000	17	82352.94118	17	580.2048	5/4/2018	25720	525	25

cinemaTicke.csv

	A	B	C	D	E	F	G	H	I	J	K	L
1	movieID	name	rating	year	released	score	votes	director	writer	star	company	genreId
2	501	The Shining	R	1980	June 13, 1980 (United S	8.4	927000	Stanley Kubrick	Stephen King	Jack Nicholson	Warner Bros.	201
3	502	The Blue Lagoon	R	1980	July 2, 1980 (United Sta	5.8	65000	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	Columbia Pictures	203
4	503	Star Wars: Episode V - The Em	PG	1980	June 20, 1980 (United S	8.7	1200000	Irvin Kershner	Leigh Brackett	Mark Hamill	Lucasfilm	205
5	504	Airplane!	PG	1980	July 2, 1980 (United Sta	7.7	221000	Jim Abrahams	Jim Abrahams	Robert Hays	Paramount Pictures	207
6	505	Caddyshack	R	1980	July 25, 1980 (United St	7.3	108000	Harold Ramis	Brian Doyle-Murray	Chevy Chase	Orion Pictures	209
7	506	Friday the 13th	R	1980	May 9, 1980 (United Sta	6.4	123000	Sean S. Cunningha	Victor Miller	Betsy Palmer	Paramount Pictures	211
8	507	The Blues Brothers	R	1980	June 20, 1980 (United S	7.9	188000	John Landis	Dan Aykroyd	John Belushi	Universal Pictures	213
9	508	Raging Bull	R	1980	December 19, 1980 (Un	8.2	330000	Martin Scorsese	Jake LaMotta	Robert De Niro	Chartoff-Winkler Pro	215
10	509	Superman II	PG	1980	June 19, 1981 (United S	6.8	101000	Richard Lester	Jerry Siegel	Gene Hackman	Dovemead Films	217
11	510	The Long Riders	R	1980	May 16, 1980 (United S	7	10000	Walter Hill	Bill Bryden	David Carradine	United Artists	219
12	511	Any Which Way You Can	PG	1980	December 17, 1980 (Un	6.1	18000	Buddy Van Horn	Stanford Sherman	Clint Eastwood	The Malpaso Compa	221
13	512	The Gods Must Be Crazy	PG	1980	October 26, 1984 (Unite	7.3	54000	Jamie Uys	Jamie Uys	N!xau	C.A.T. Films	223
14	513	Popeye	PG	1980	December 12, 1980 (Un	5.3	30000	Robert Altman	Jules Feiffer	Robin Williams	Paramount Pictures	225
15	514	Ordinary People	R	1980	September 19, 1980 (Ur	7.7	49000	Robert Redford	Judith Guest	Donald Sutherlanc	Paramount Pictures	227
16	515	Dressed to Kill	R	1980	July 25, 1980 (United St	7.1	37000	Brian De Palma	Brian De Palma	Michael Caine	Filmways Pictures	229
17	516	Somewhere in Time	PG	1980	October 3, 1980 (Unitec	7.2	27000	Jeannot Szwarc	Richard Matheson	Christopher Reeve	Rastar Pictures	231
18	517	Fame	R	1980	May 16, 1980 (United S	6.6	21000	Alan Parker	Christopher Gore	Eddie Barth	Metro-Goldwyn-May	233
19	518	9 to 5	PG	1980	December 19, 1980 (Un	6.9	29000	Colin Higgins	Patricia Resnick	Jane Fonda	IPC Films	235
20	519	The Fog	R	1980	February 8, 1980 (Unite	6.8	66000	John Carpenter	John Carpenter	Adrienne Barbeau	AVCO Embassy Pictu	237
21	520	Stir Crazy	R	1980	December 12, 1980 (Un	6.8	26000	Sidney Poitier	Bruce Jay Friedman	Gene Wilder	Columbia Pictures	239
22	521	Cruising	R	1980	February 15, 1980 (Unit	6.5	20000	William Friedkin	William Friedkin	Al Pacino	Lorimar Film Entertai	241
23	522	Heaven's Gate	R	1980	April 24, 1981 (United S	6.8	14000	Michael Cimino	Michael Cimino	Kris Kristofferson	Partisan Productions	243
24	523	The Final Countdown	PG	1980	August 1, 1980 (United S	6.7	22000	Don Taylor	Thomas Hunter	Kirk Douglas	Bryna Productions	245
25	524	Xanadu	PG	1980	August 8, 1980 (United S	5.3	12000	Robert Greenwald	Richard Christian Danus	Olivia Newton-Joh	Universal Pictures	247
26	525	Urban Cowboy	PG	1980	June 6, 1980 (United Sta	6.4	14000	James Bridges	Aaron Latham	John Travolta	Paramount Pictures	249
27	526	Altered States	R	1980	December 25, 1980 (Un	6.9	33000	Ken Russell	Paddy Chayefsky	William Hurt	Warner Bros.	251
28	527	Little Darlings	R	1980	March 21, 1980 (United	6.5	5100	Ron Maxwell	Kimi Peck	Tatum O'Neal	Stephen Friedman/Ki	253

movies.csv

	A	B	C	D
1	genreId	genre	country	
2	201	Drama	United Kingdom	
3	203	Adventure	United States	
4	205	Action	United States	
5	207	Comedy	United States	
6	209	Comedy	United States	
7	211	Horror	United States	
8	213	Action	United States	
9	215	Biography	United States	
10	217	Action	United States	
11	219	Biography	United States	
12	221	Action	United States	
13	223	Adventure	South Africa	
14	225	Adventure	United States	
15	227	Drama	United States	
16	229	Crime	United States	
17	231	Drama	United States	
18	233	Drama	United States	
19	235	Comedy	United States	
20	237	Horror	United States	
21	239	Comedy	United States	
22	241	Crime	West Germany	
23	243	Adventure	United States	
24	245	Action	United States	
25	247	Fantasy	United States	

Genre.csv

	A
1	RANKDISTRIBUTORSMOVIESTOTAL GROSS
2	1Walt Disney588"$40
3	2Warner Bros.824"$36
4	3Sony Pictures747"$29
5	4Universal535"$28
6	520th Century Fox525"$25
7	6Paramount Pictures493"$24
8	7Lionsgate426"$9
9	8New Line209"$6
10	9Dreamworks SKG77"$4
11	10Miramax385"$3

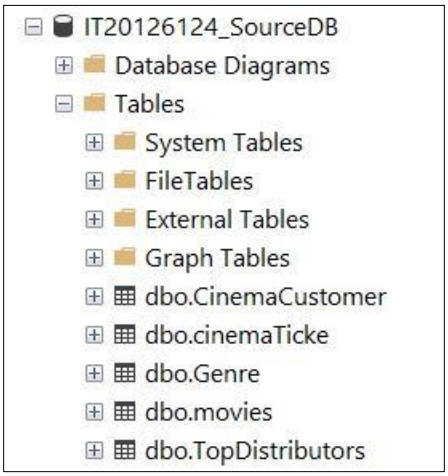
TopDistributors.csv

CustomerAddress - Notepad

File Edit Format View Help

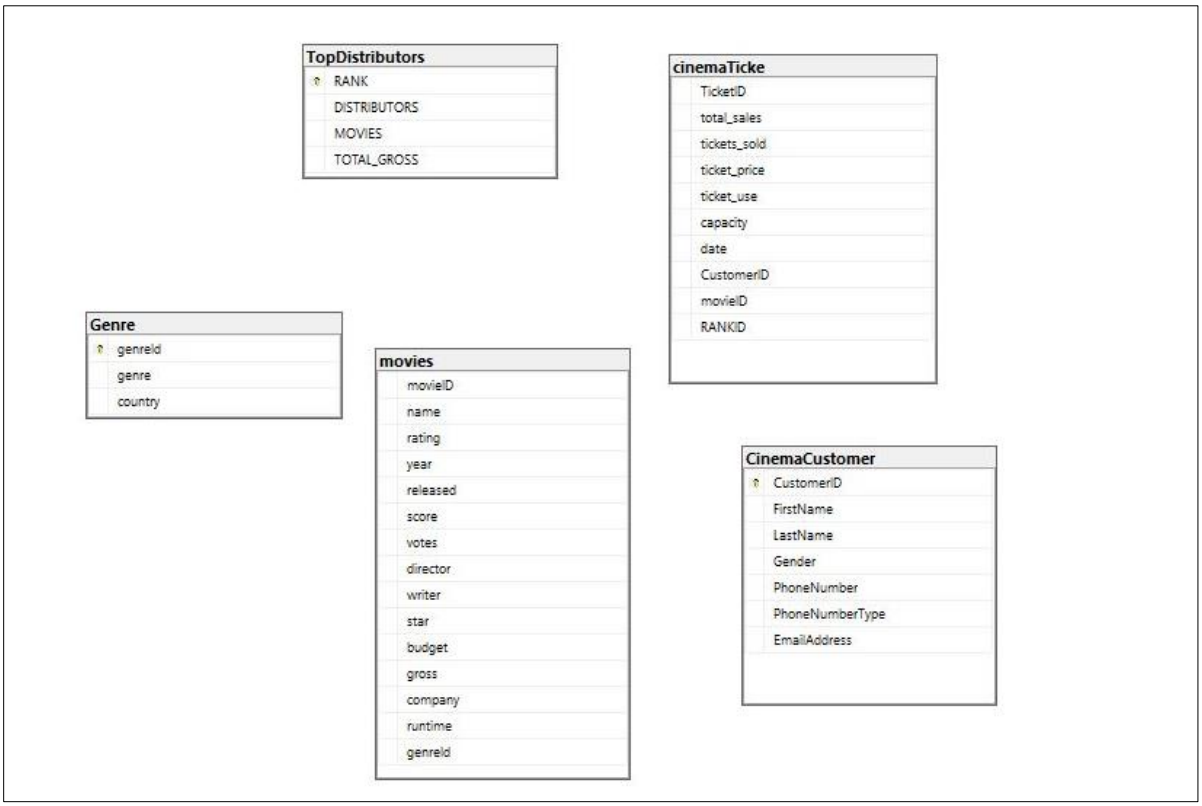
customerid	AddressType	AddressLine1	AddressLine2	City	StateProvinceName	PostalCode	CountryRegionName
23899	Home	00, rue Saint-Lazare	NULL	Dunkerque	Nord	59140	France
22458	Home	02, place de Fontenoy	NULL	Verrieres Le Buisson	Essonne	91370	France
21058	Home	035, boulevard du Montparnasse	NULL	Verrieres Le Buisson	Essonne	91370	France
26871	Home	081, boulevard du Montparnasse	NULL	Saint-Denis	Seine Saint Denis	93400	France
24676	Home	081, boulevard du Montparnasse	NULL	Seattle	Washington	98104	United States
17740	Home	084, boulevard du Montparnasse	NULL	Les Ulis	Essonne	91940	France
15017	Home	1 Smiling Tree Court	Space 55	Los Angeles	California	90012	United States
20553	Home	1, allée des Princes	NULL	Courbevoie	Hauts de Seine	92400	France
12649	Home	1, avenue des Champs-Elysées	NULL	Paris	Seine (Paris)	75017	France
12632	Home	1, boulevard Beau Marchais	NULL	Sèvres	Hauts de Seine	92310	France
12639	Home	1, cours Mirabeau	NULL	Roncq	Nord	59223	France
12718	Home	1, place Beaubernard	NULL	Paris	Seine (Paris)	75003	France
20910	Home	1, place Beaubernard	NULL	Paris	Seine (Paris)	75009	France
14766	Home	1, place Beaubernard	NULL	Saint-Denis	Seine Saint Denis	93400	France
27335	Home	1, place Beaubernard	NULL	Tremblay-en-France	Seine Saint Denis	93290	France
29450	Home	1, place de Brazaville	NULL	Colomiers	Garonne (Haute)	31770	France
13773	Home	1, place de Brazaville	NULL	Dunkerque	Nord	59140	France
17084	Home	1, place de Brazaville	NULL	Lille	Nord	59000	France
12510	Home	1, rue de Courtaboeuf	NULL	Lieusaint	Seine et Marne	77127	France
14959	Home	1, rue de Courtaboeuf	NULL	Lille	Nord	59000	France
24095	Home	1, rue de Courtaboeuf	NULL	Versailles	Yveline	78000	France
16439	Home	1, rue de Fontfroide	NULL	Paris	Seine (Paris)	75003	France
14689	Home	1, rue de l´Avenir	NULL	Chatou	Yveline	78400	France
27772	Home	1, rue de l´Avenir	NULL	Morangis	Essonne	91420	France
20827	Home	1, rue de la Cavalerie	NULL	Paris La Defense	Hauts de Seine	92081	France
14135	Home	1, rue de la Cavalerie	NULL	Roncq	Nord	59223	France
13531	Home	1, rue de la Cavalerie	NULL	Sèvres	Hauts de Seine	92310	France
14144	Home	1, rue de la Centenaire	NULL	Cergy	Val d'Oise	95000	France
22800	Home	1, rue de la Centenaire	NULL	Colombes	Hauts de Seine	92700	France
19320	Home	1, rue de la Centenaire	NULL	Les Ulis	Essonne	91940	France
14923	Home	1, rue de la Centenaire	NULL	Lille	Nord	59000	France
19076	Home	1, rue de la Centenaire	NULL	Roubaix	Nord	59100	France
14845	Home	1, rue de la Centenaire	NULL	Saint-Denis	Seine Saint Denis	93400	France
21080	Home	1, rue de la Centenaire	NULL	Versailles	Yveline	78000	France
19295	Home	1, rue de Maubeuge	NULL	Metz	Moselle	57000	France
15651	Home	1, rue de Maubeuge	NULL	Morangis	Essonne	91420	France
23973	Home	1, rue de Maubeuge	NULL	Paris	Seine (Paris)	75007	France
21132	Home	1, rue de Maubeuge	NULL	Paris	Seine (Paris)	75009	France
16596	Home	1, rue de Maubeuge	NULL	Saint Germain en Laye	Yveline	78100	France

CustomerAddress.txt



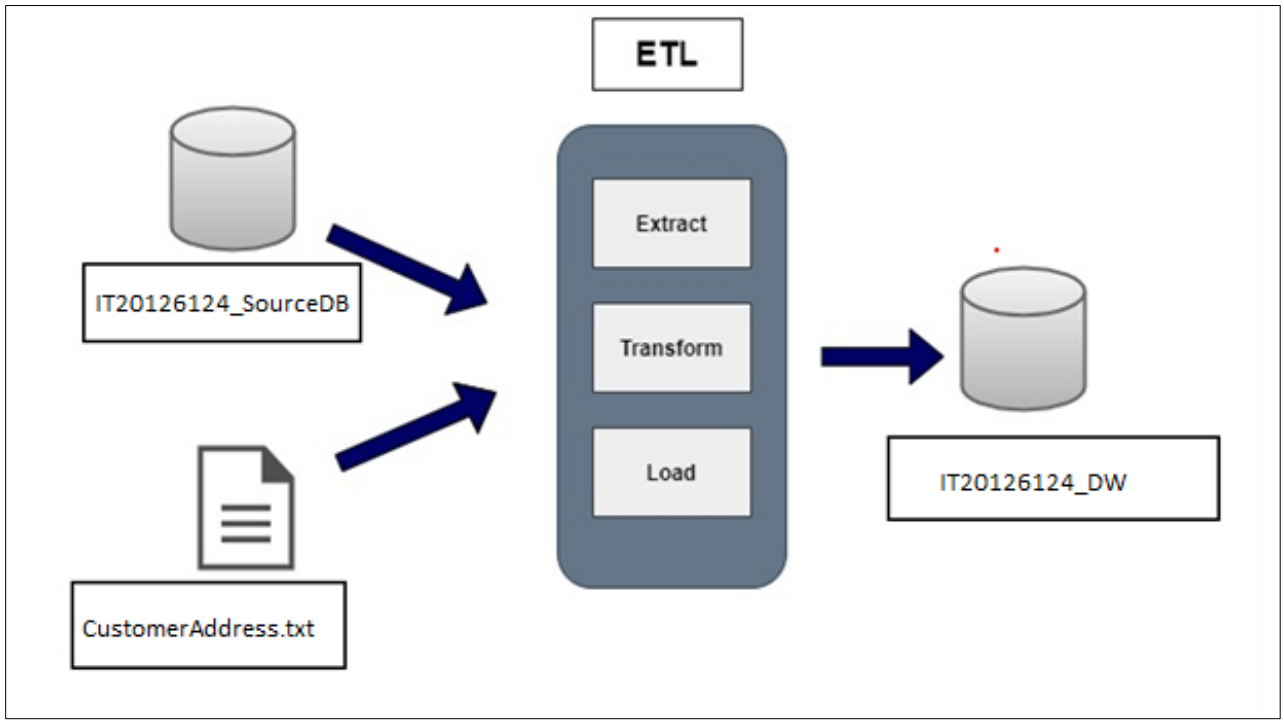
IT20126124_SourceDB

Database Diagram for IT20126124_SourceDB:



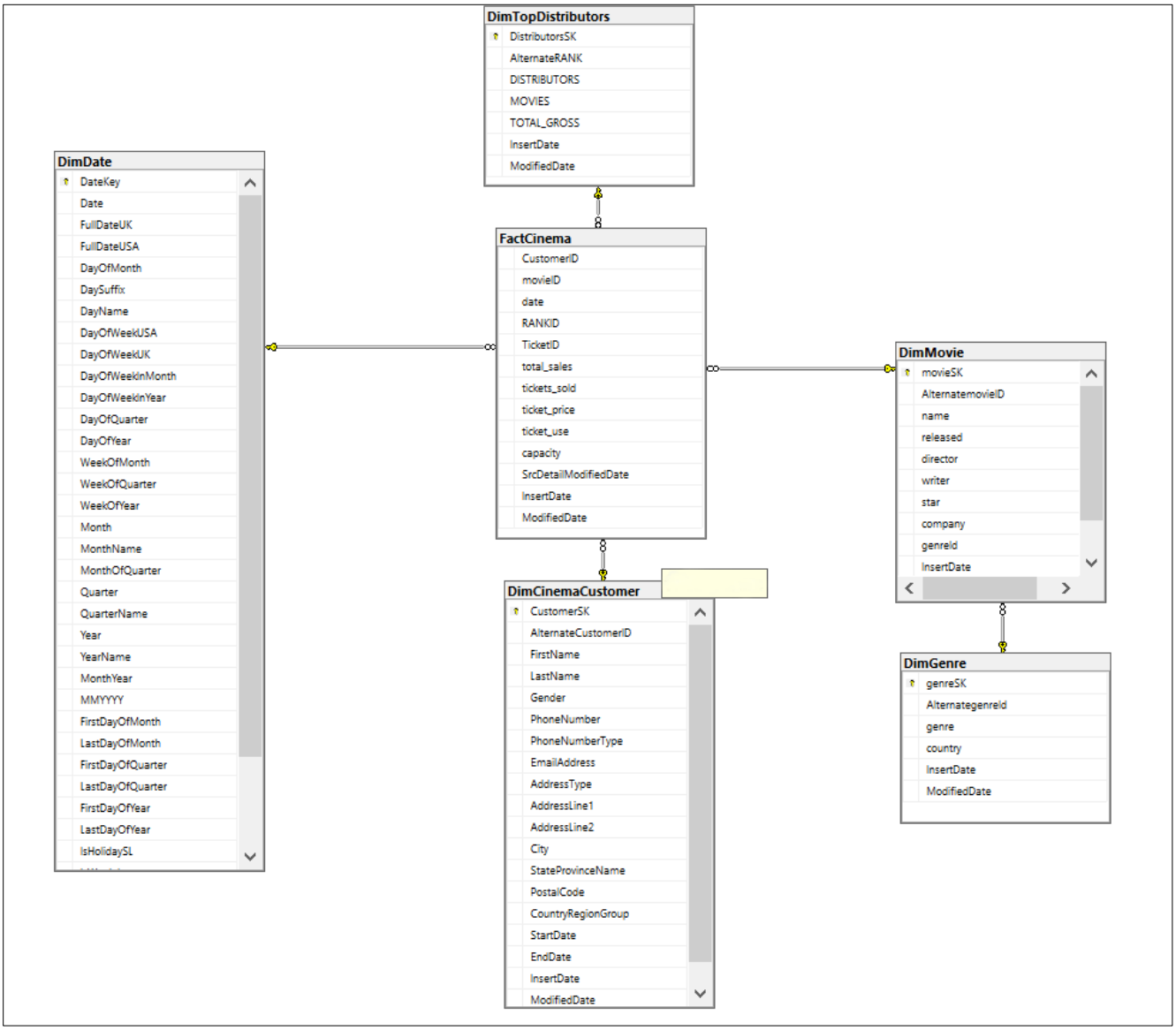
IT20126124_SourceDB

Step 03 – Solution Architecture



Using different processes, architectures, and technologies we can manage data from various sources and convert them to business insights to make decisions, analysis data and report building. This will bring new dimension to the data as well.

Step 04: Data Warehouse design & development



Assumptions

- Here I have used snowflake schema for data warehouse design and add 4 dimensions apart from the date dimension.
- I have taken Dim CinemaCustomer as slowly changing dimension, customer address and phone number can change time to time, and we need to keep track of their historical address details.

Before creating the cinema fact table & other dimensions, started by creating the Date dimension using ‘DateMaster.sql’ file code dimension using ‘DateMaster.sql’ file code

DESKTOP-I2IVLM...mCinemaCustomer

Column Name	Data Type	Allow Nulls
CustomerSK	int	☐
AlternateCustomerID	int	☑
FirstName	nvarchar(50)	☑
LastName	nvarchar(50)	☑
Gender	nvarchar(1)	☑
PhoneNumber	nvarchar(25)	☑
PhoneNumberType	nvarchar(50)	☑
EmailAddress	nvarchar(50)	☑
AddressType	nvarchar(50)	☑
AddressLine1	nvarchar(60)	☑
AddressLine2	nvarchar(60)	☑
City	nvarchar(30)	☑
StateProvinceName	nvarchar(50)	☑
PostalCode	nvarchar(15)	☑
CountryRegionGroup	nvarchar(50)	☑
StartDate	datetime	☑
EndDate	datetime	☑
InsertDate	datetime	☑
ModifiedDate	datetime	☑
		☐

DESKTOP-I2IVLMO...- dbo.DimMovie | DESKTOP-I2IVLM...mCinemaCustomer

Column Name	Data Type	Allow Nulls
movieSK	int	☐
AlternatemovieID	int	☑
name	nvarchar(100)	☑
released	nvarchar(100)	☑
director	nvarchar(100)	☑
writer	nvarchar(100)	☑
star	nvarchar(100)	☑
company	nvarchar(100)	☑
genreId	int	☑
InsertDate	datetime	☑
ModifiedDate	datetime	☑
		☐

DESKTOP-I2IVLMO...- dbo.DimGenre | DESKTOP-I2IVLMO...- dbo.DimMovie | DESK

Column Name	Data Type	Allow Nulls
genreSK	int	☐
AlternategenreId	int	☑
genre	nvarchar(50)	☑
country	nvarchar(50)	☑
InsertDate	datetime	☑
ModifiedDate	datetime	☑
		☐

DESKTOP-I2IVLMO....mTopDistributors | DESKTOP-I2IVLMO...- dbo.DimGenre | DES

Column Name	Data Type	Allow Nulls
DistributorsSK	int	☐
AlternateRANK	int	☑
DISTRIBUTORS	nvarchar(50)	☑
MOVIES	int	☑
TOTAL_GROSS	money	☑
InsertDate	datetime	☑
ModifiedDate	datetime	☑
		☐

DESKTOP-I2IVLMO....- dbo.FactCinema | DESKTOP-I2IVLMO....mTopDistributors | DE

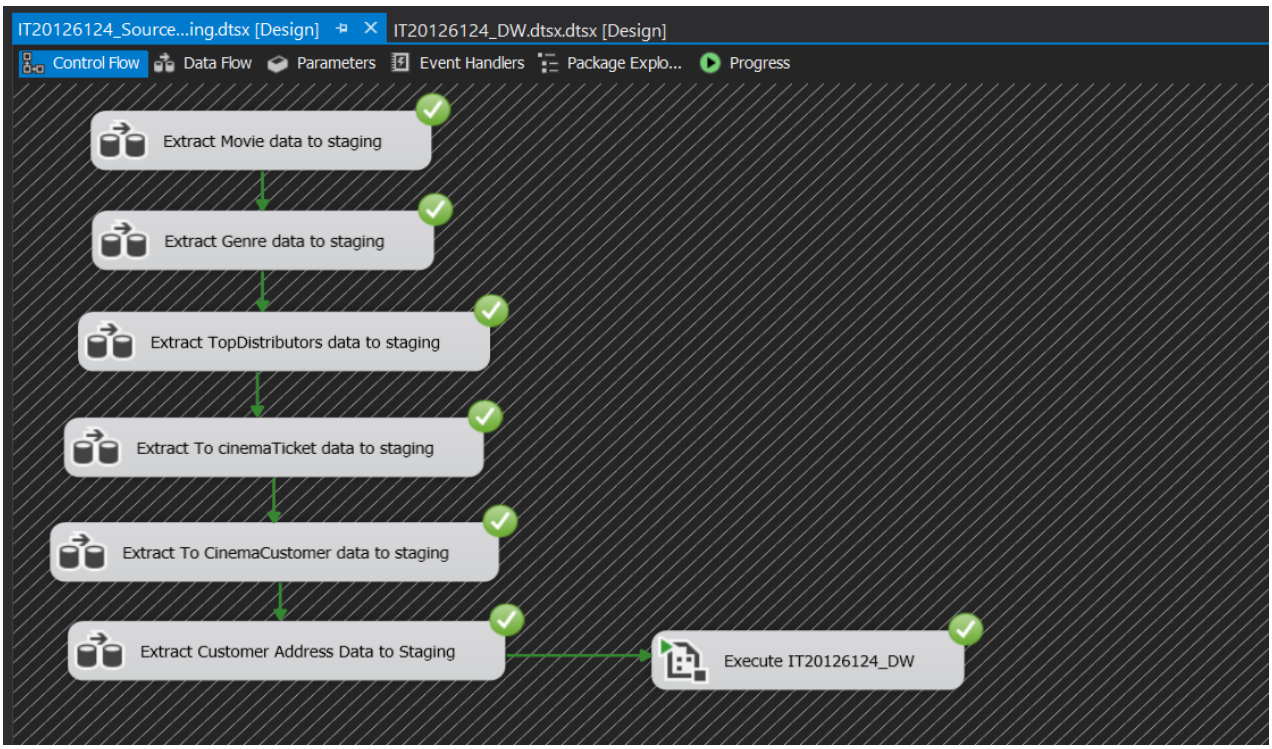
Column Name	Data Type	Allow Nulls
CustomerID	int	☑
movieID	int	☑
date	int	☑
RANKID	int	☑
TicketID	int	☑
total_sales	int	☑
tickets_sold	int	☑
ticket_price	int	☑
ticket_use	int	☑
capacity	float	☑
SrcDetailModifiedDate	datetime	☑
InsertDate	datetime	☑
ModifiedDate	datetime	☑
		☐

DESKTOP-I2IVLMO...W - dbo.DimDate | DESKTOP-I2IVLMO....- dbo.FactCinema | DES

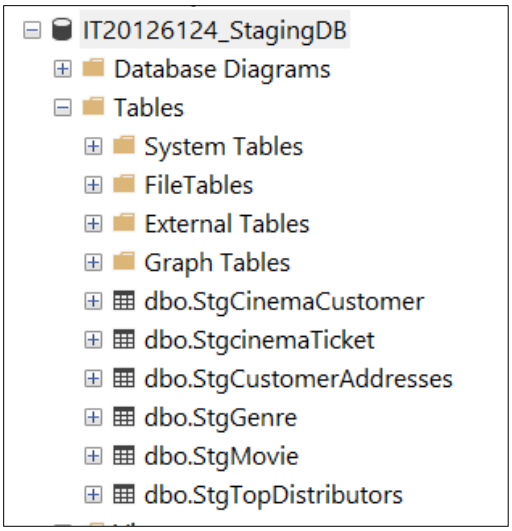
Column Name	Data Type	Allow Nulls
DateKey	int	☐
Date	datetime	☑
FullDateUK	char(10)	☑
FullDateUSA	char(10)	☑
DayOfMonth	varchar(2)	☑
DaySuffix	varchar(4)	☑
DayName	varchar(9)	☑
DayOfWeekUSA	char(1)	☑
DayOfWeekUK	char(1)	☑
DayOfWeekInMonth	varchar(2)	☑
DayOfWeekInYear	varchar(2)	☑
DayOfQuarter	varchar(3)	☑
DayOfYear	varchar(3)	☑
WeekOfMonth	varchar(1)	☑
WeekOfQuarter	varchar(2)	☑
WeekOfYear	varchar(2)	☑
Month	varchar(2)	☑
MonthName	varchar(9)	☑
MonthOfQuarter	varchar(2)	☑
Quarter	char(1)	☑
QuarterName	varchar(9)	☑
Year	char(4)	☑
YearName	char(7)	☑
MonthYear	char(10)	☑
MMYYYY	char(6)	☑

MMYYYY	char(6)	☑
FirstDayOfMonth	date	☑
LastDayOfMonth	date	☑
FirstDayOfQuarter	date	☑
LastDayOfQuarter	date	☑
FirstDayOfYear	date	☑
LastDayOfYear	date	☑
IsHolidaySL	bit	☑
IsWeekday	bit	☑
HolidaySL	varchar(50)	☑
isCurrentDay	int	☑
isDataAvailable	int	☑
isLatestDataAvailable	int	☑
		☐

Step 05: ETL Development

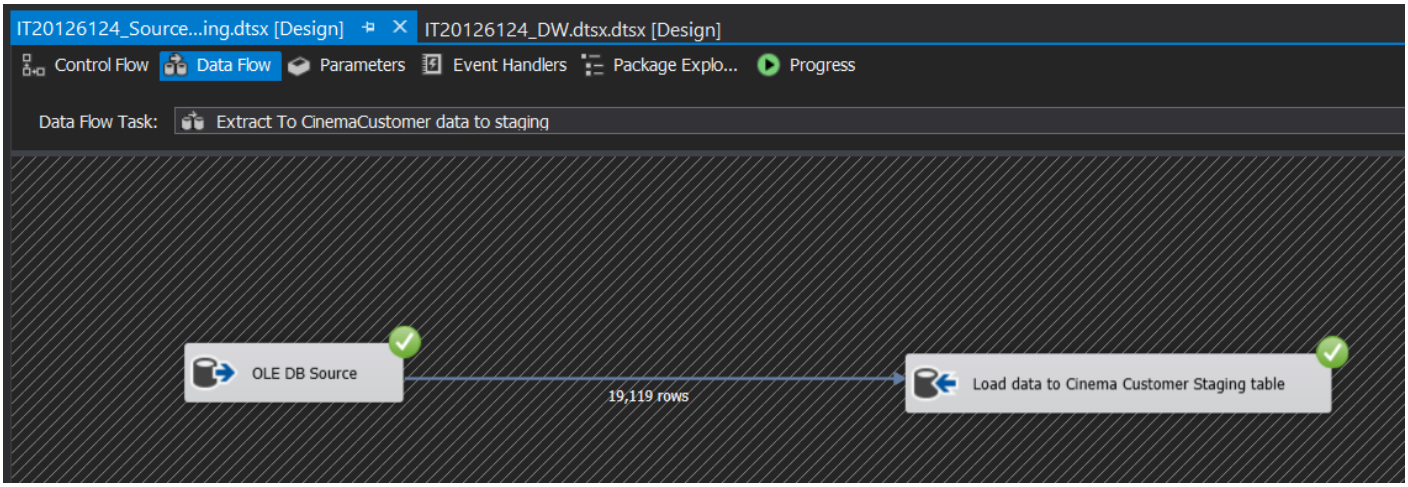


First extracted all the data from the tables which are in the IT20126124_SourceDB and CustomerAddress.txt to the separate staging DB called IT20126124_StagingDB as shown in the below using SQL Server Integration Service Software.

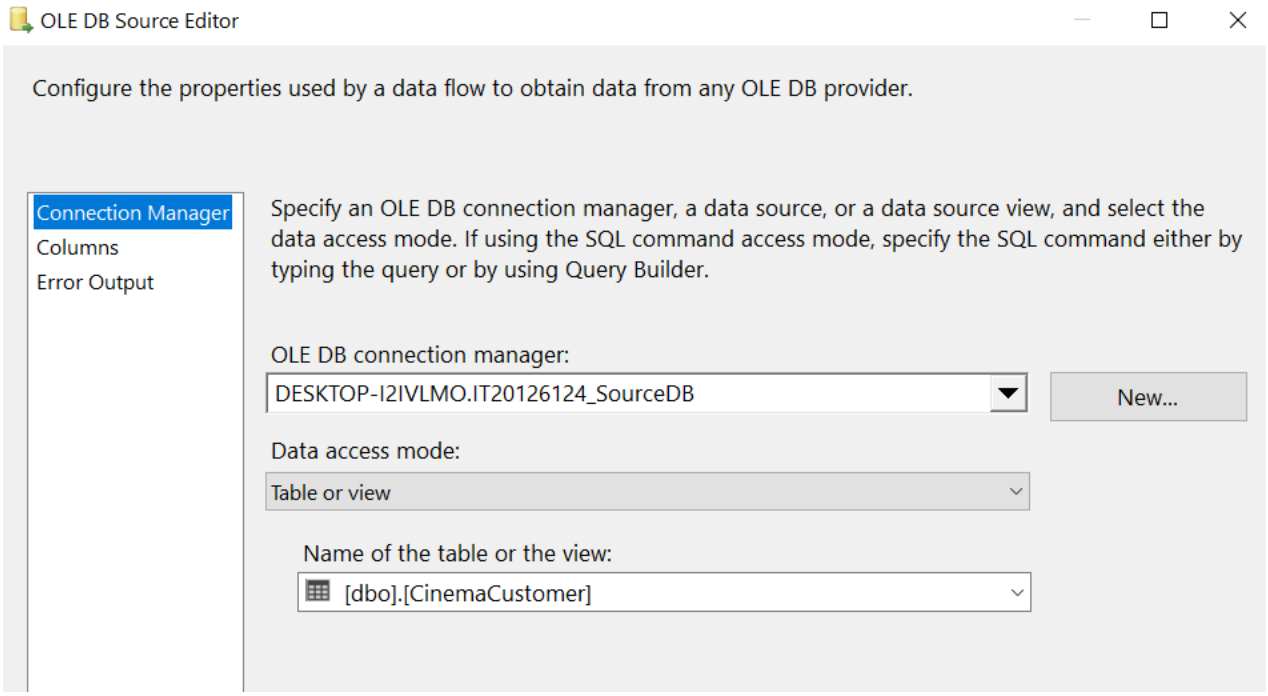



```
IT20126124_StagingDB
```

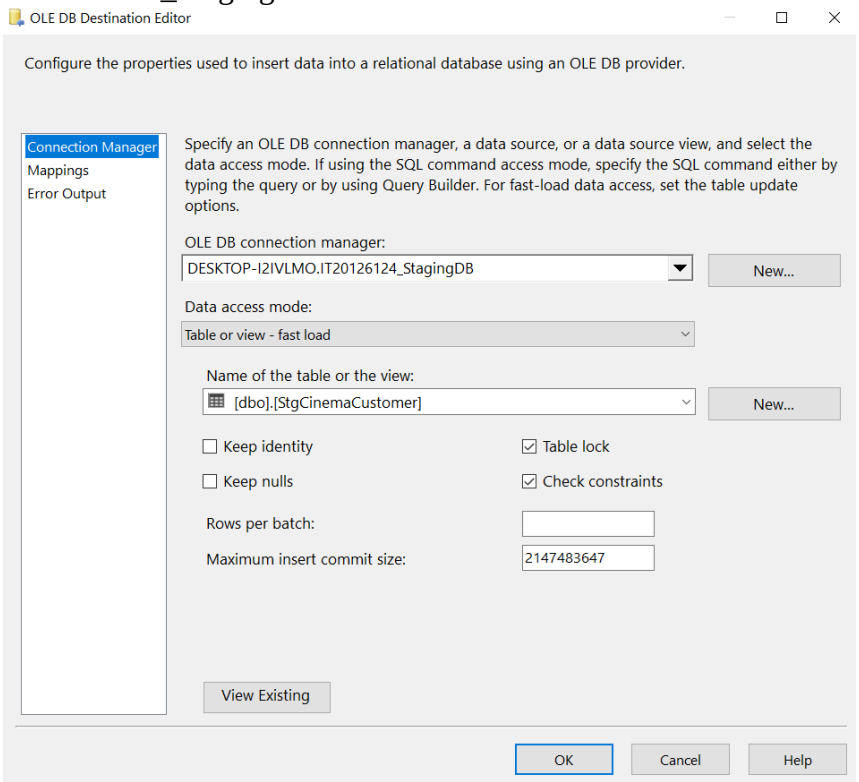
1. Extract cinema customer details



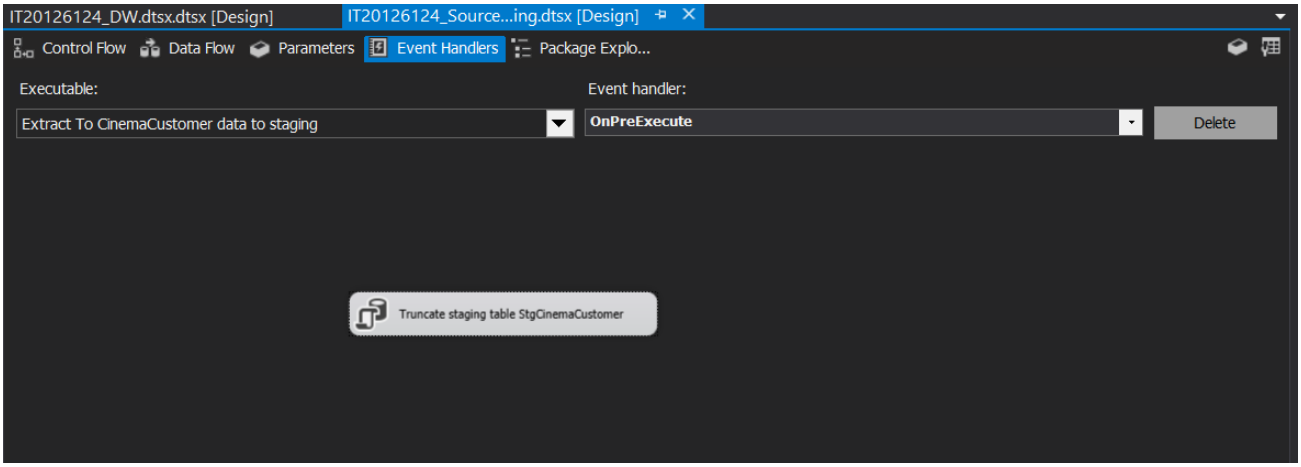
Used OLEDB data source as dbo.CinemaCustomer table in IT20126124_SourceDB



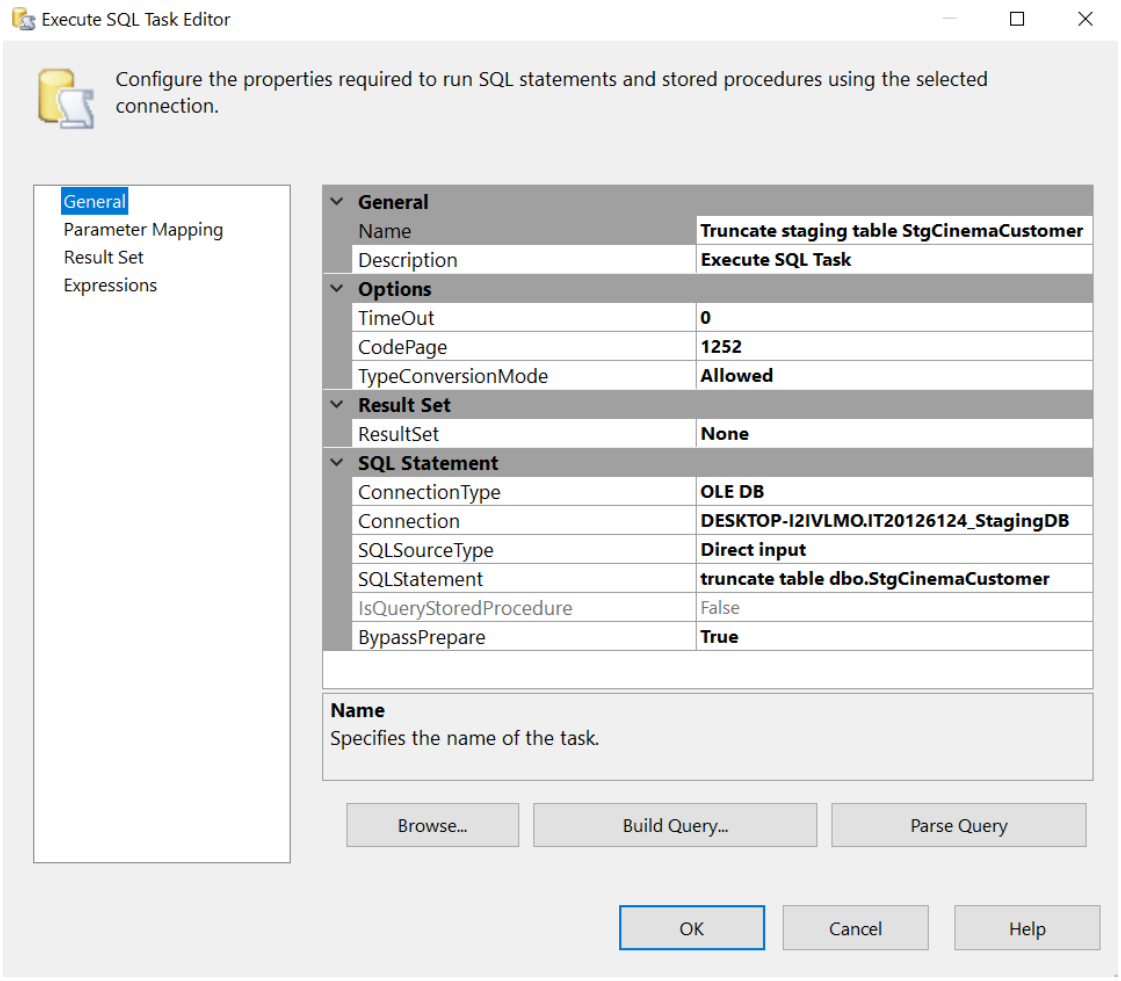
OLE DB destination for create new table as dbo.StgCinemaCustomer in the IT20126124_StagingDB



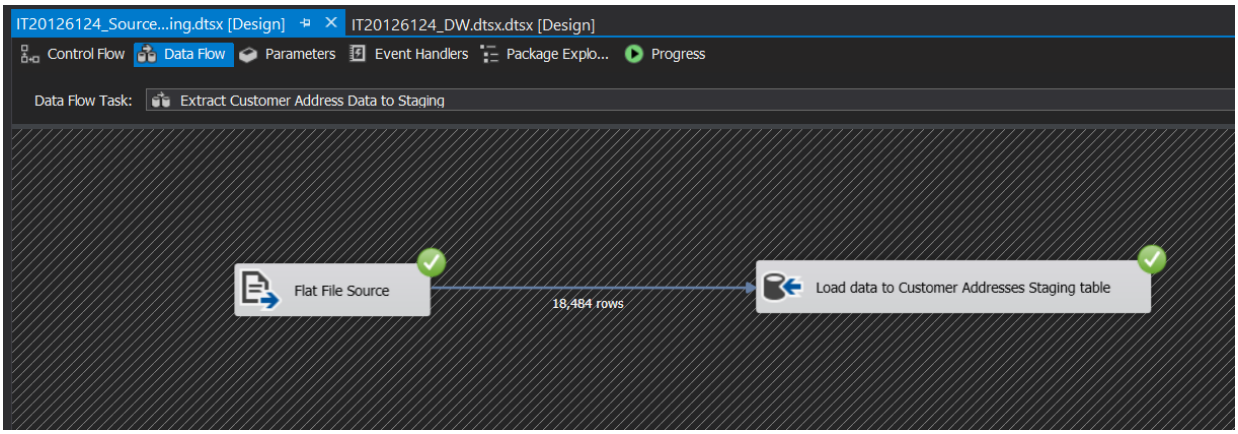
Event Handlers



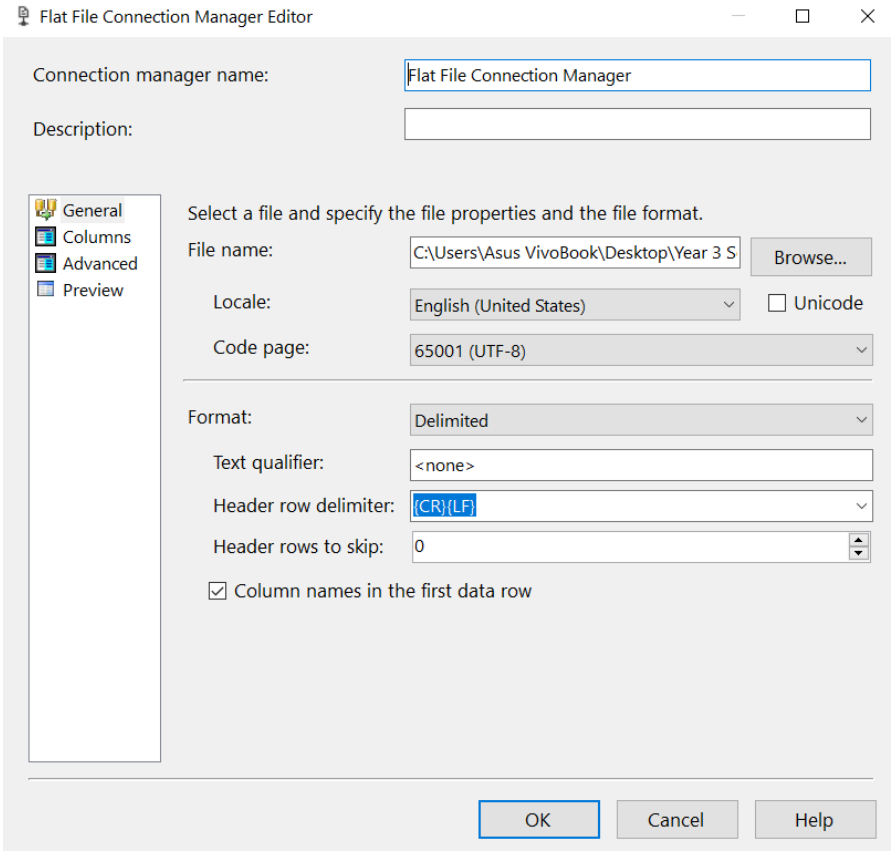
Used Execute SQL Task SSIS tool Truncate table for SQL command as truncate table dbo.StgCinemaCustomer in the IT20126124_StagingDB.



Customer Address Details Extraction (Data Flow)



Used Flat file Source SSIS tool, to extract CustomerAddress.txt data

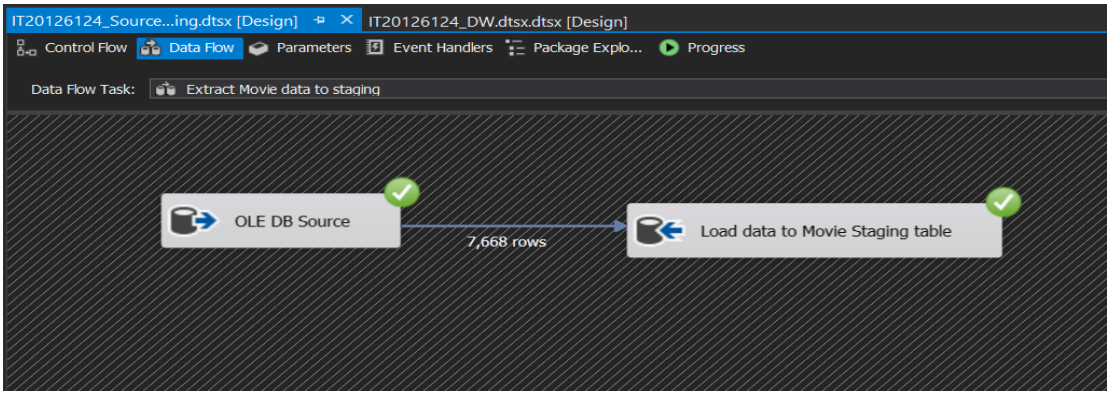


Used OLE DB Destination SSIS tool to create new table as StgCustomerAddresses to load text file's data into IT20126124_StagingDB

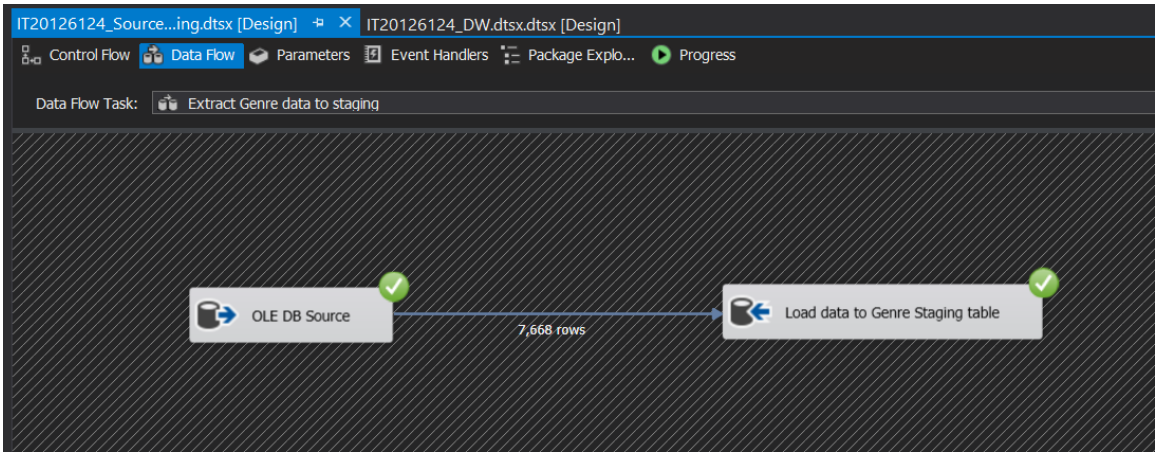
Used Execute SQL Task SSIS tools Truncate table for SQL command as truncate table

Note: Followed exact process to extract other source tables data in to staging

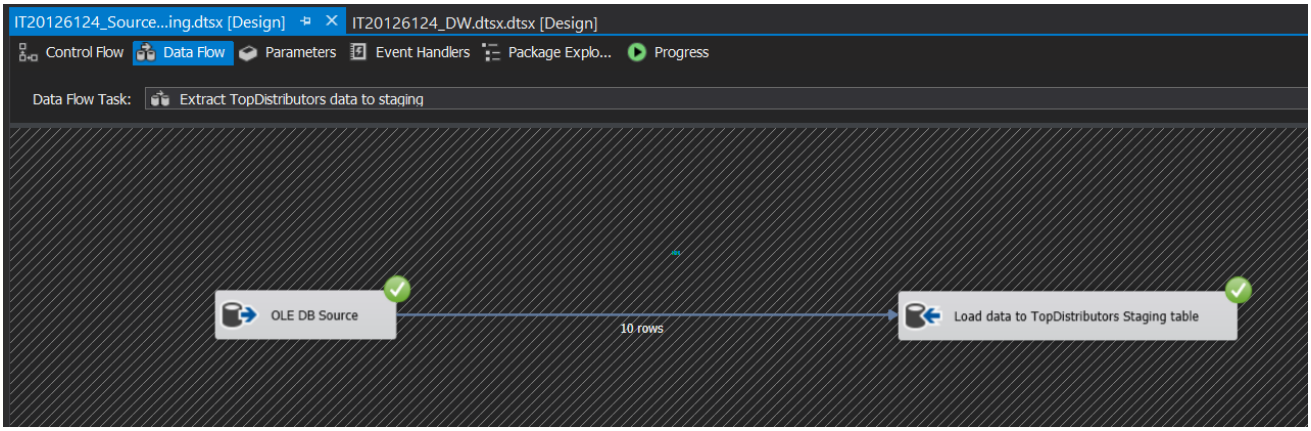
2. Extract Movie data to staging



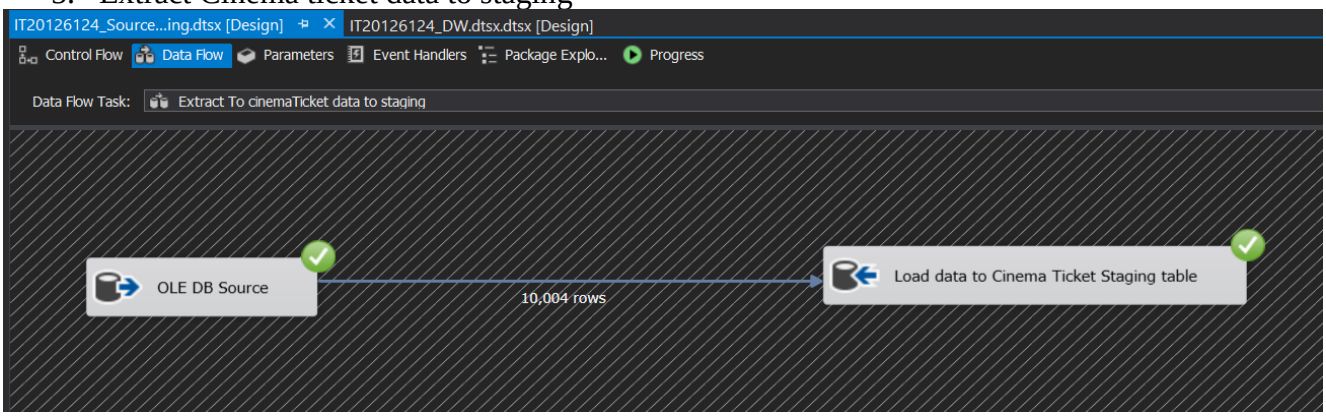
3. Extract Genre data to staging



4. Extract Top distributors data to staging



5. Extract Cinema ticket data to staging

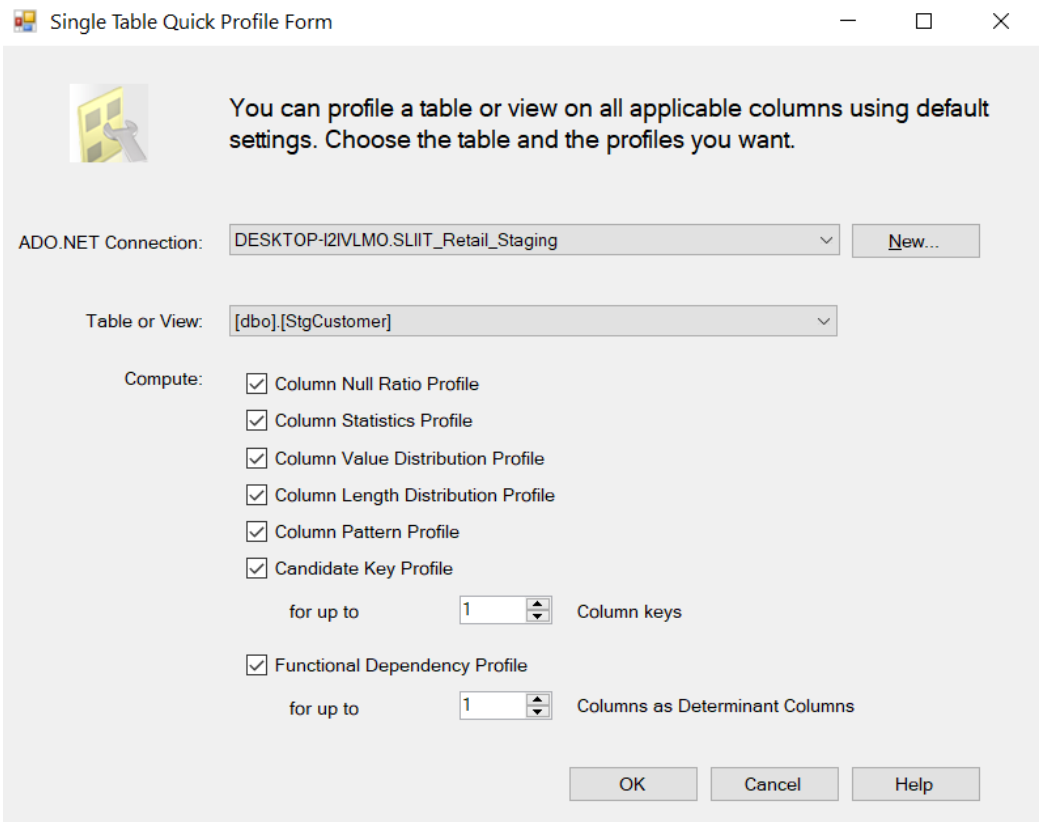


Data Profiling

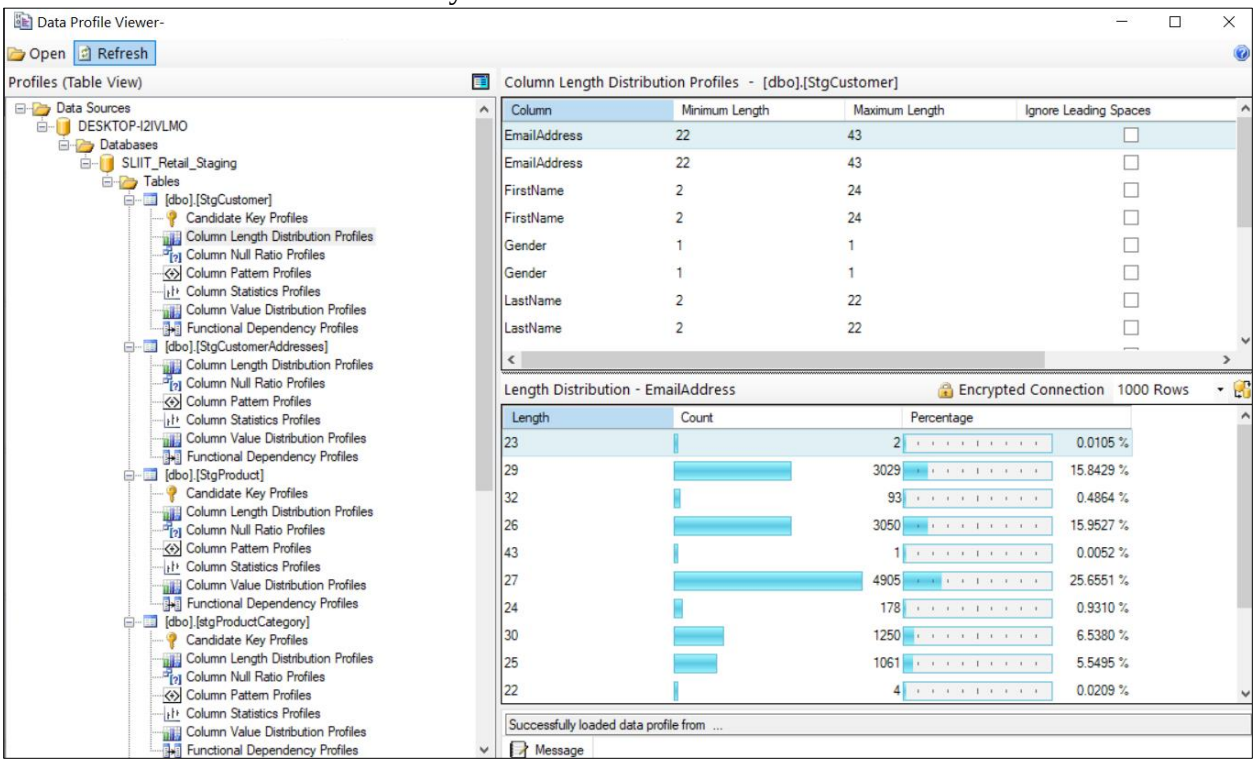
Staging table data was used to analyzed and determined what types of transformation was needed to perform on the data.

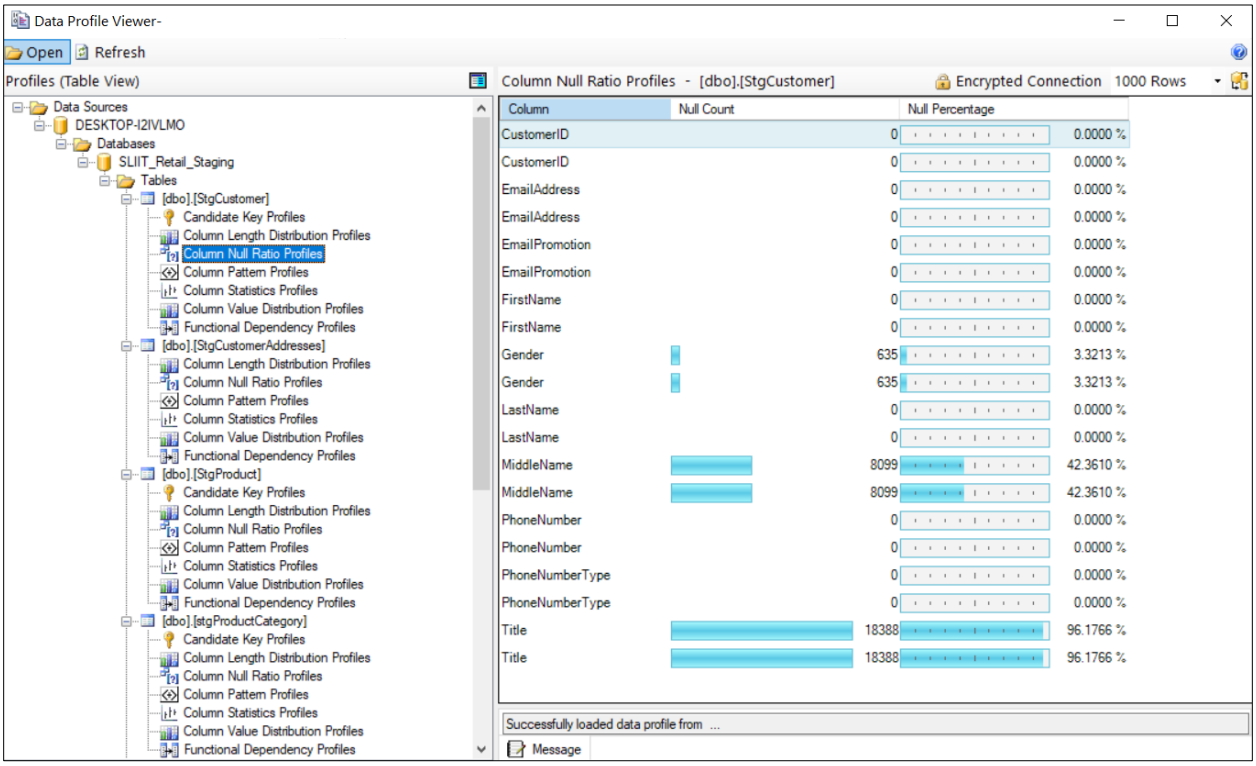
- Right click on SSIS Packages and New SSIS Package selected.
- In the Control Flow of Data_Profiling.dtsx, drag and drop the Data Profiling Task and double click to open the configuration.
- Clicked on Quick Profile button to open Single Table Quick Profile Form.
- Clicked on New button and create the connection to IT20126124_StagingDB
- From the Table or View dropdown, StgCustomer table selected.

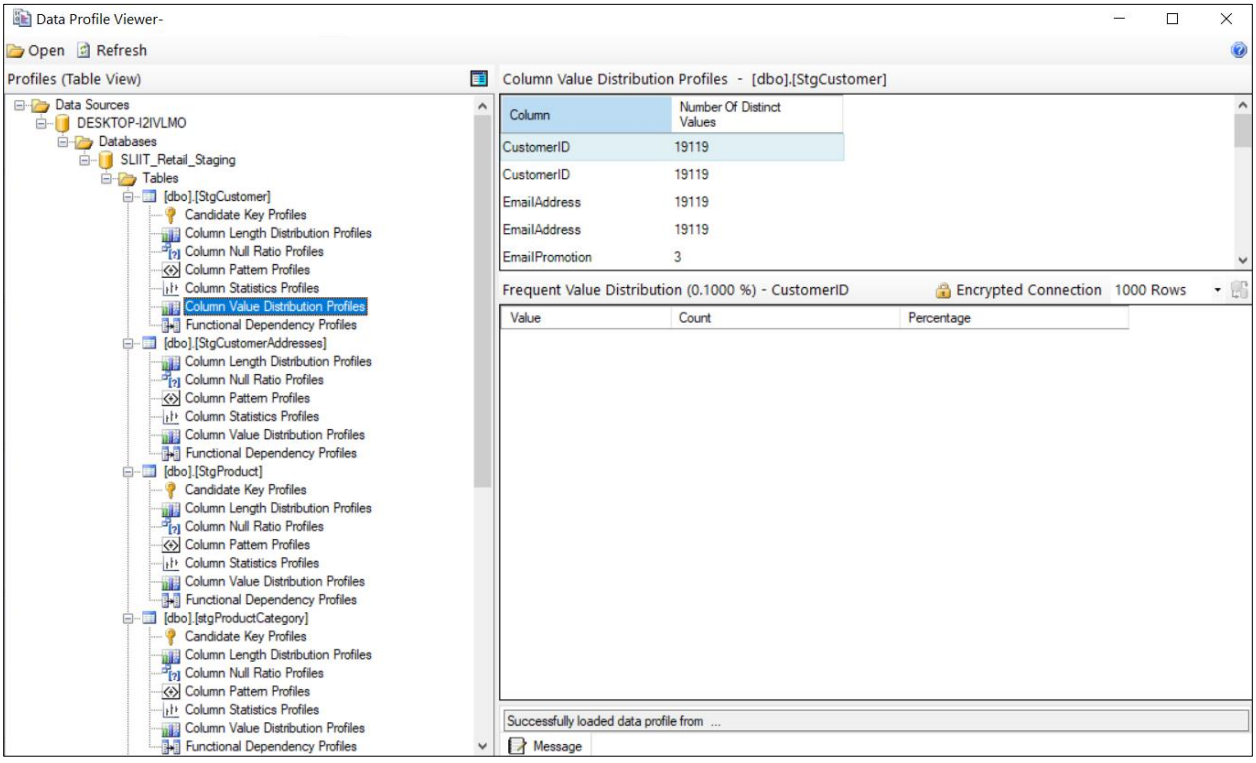
Selected all check boxes and click on OK button to complete the configuration



- Save the package and Run the Data profiling Pack to profile the stgCustomer
- Once the green tick appeared, double click the Data Profiling Task and Click on the Open Profile Viewer to view the analyzed data




Data Profile Viewer-
Open
Refresh
Profiles (Table View)
Data Sources
DESKTOP-I2IVLMO
Databases
SLIIT_Retail_Staging
Tables
[dbo].[StgCustomer]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[StgCustomerAddresses]
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[StgProduct]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[stgProductCategory]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
Column Null Ratio Profiles - [dbo].[StgCustomer]
Encrypted Connection
1000 Rows
Column
Null Count
Null Percentage
CustomerID 0 0.0000 %
CustomerID 0 0.0000 %
EmailAddress 0 0.0000 %
EmailAddress 0 0.0000 %
EmailPromotion 0 0.0000 %
EmailPromotion 0 0.0000 %
FirstName 0 0.0000 %
FirstName 0 0.0000 %
Gender 635 3.3213 %
Gender 635 3.3213 %
LastName 0 0.0000 %
LastName 0 0.0000 %
MiddleName 8099 42.3610 %
MiddleName 8099 42.3610 %
PhoneNumber 0 0.0000 %
PhoneNumber 0 0.0000 %
PhoneNumberType 0 0.0000 %
PhoneNumberType 0 0.0000 %
Title 18388 96.1766 %
Title 18388 96.1766 %
Successfully loaded data profile from ...
Message


Data Profile Viewer-
Open
Refresh
Profiles (Table View)
Data Sources
DESKTOP-I2IVLMO
Databases
SLIIT_Retail_Staging
Tables
[dbo].[StgCustomer]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[StgCustomerAddresses]
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[StgProduct]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
[dbo].[stgProductCategory]
Candidate Key Profiles
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Functional Dependency Profiles
Column Value Distribution Profiles - [dbo].[StgCustomer]
Column
Number Of Distinct Values
CustomerID 19119
CustomerID 19119
EmailAddress 19119
EmailAddress 19119
EmailPromotion 3
Frequent Value Distribution (0.1000 %) - CustomerID
Encrypted Connection
1000 Rows
Value
Count
Percentage
Successfully loaded data profile from ...
Message

Data Transformation

Customer Data Transformation was created by using below mentioned steps

• Created new package called IT20126124_DW.dtsx.dtsx.

• Then Dragged and dropped a Data Flow Task, renamed it as Transform and Load Customer Data and go the Data Flow tab.

• Dragged and dropped OLE DB Source, renamed as Extract from Extract from CinemCustomer Staging and configure it to access the StgCinemaCustomer table.

• And I used another OLE DB Source, renamed as Extract from Extract from Customer Addresses Staging and configure it to access the StgCustomerAddresses table. And selected all the columns.

• Then I Dragged and dropped two Sort items and connect each OLE DB Source to them.

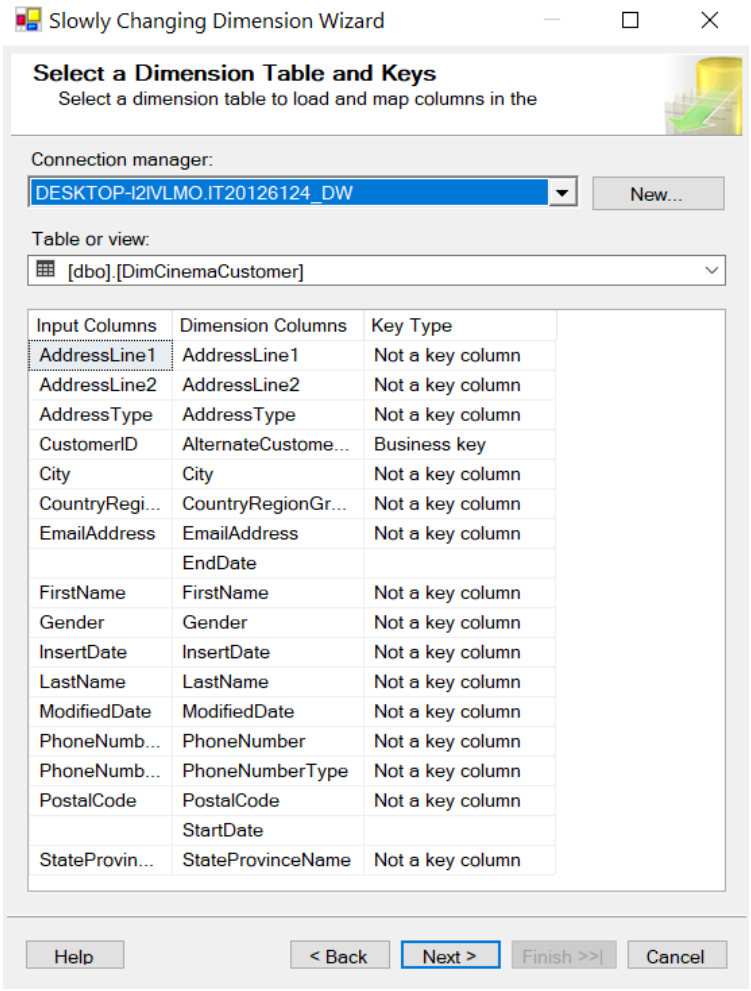
• After that I Double click Sort that is connected to Extract from CinemCustomer Staging and select CustomerID as the Sort option by ticking on the checkbox in from of CustomerID Then I did the same for the other Sort item connected to StgCustomerAddresses.

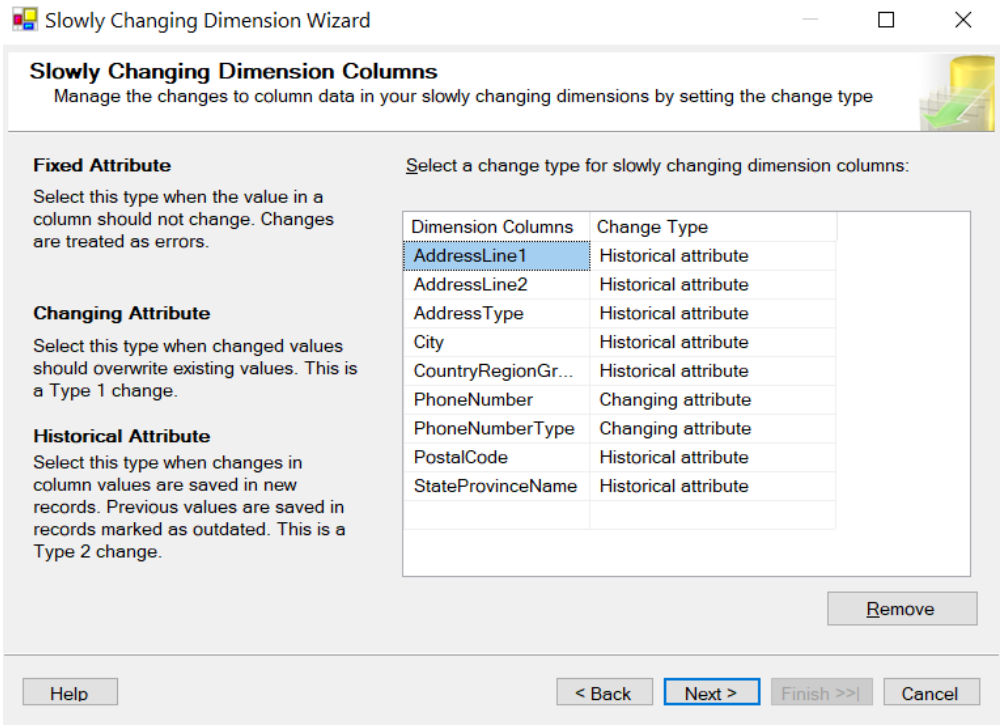
• Dragged and dropped Merge Join and link above two sort items to the Merge join.

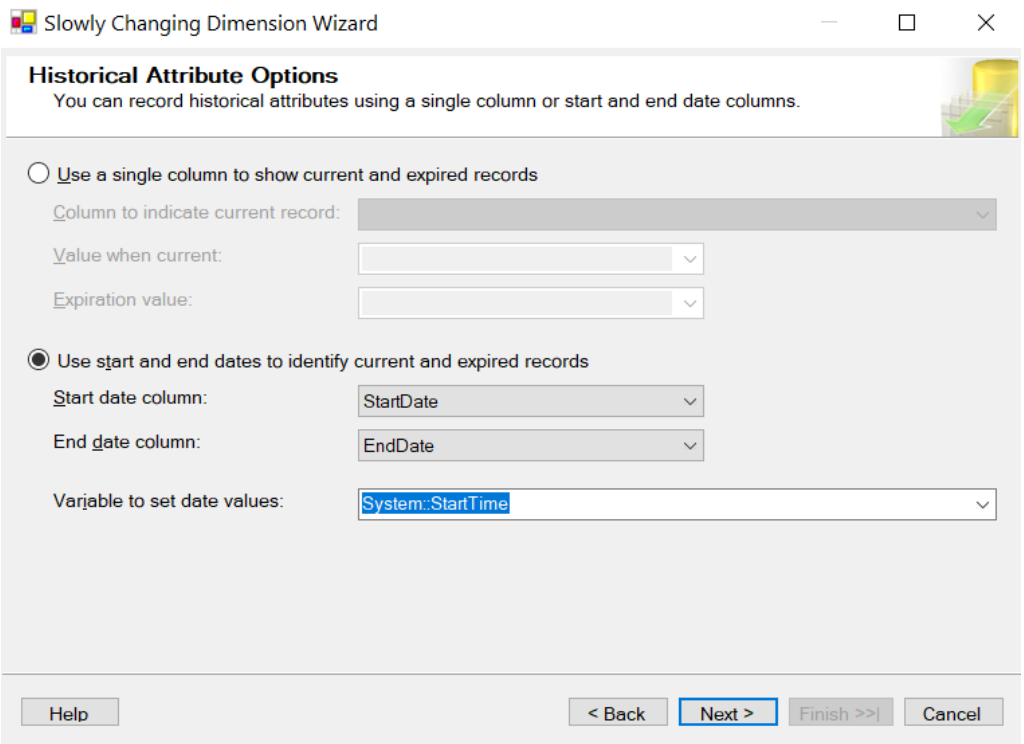
• In the Input Output Selection popup, I have selected Merge Join Left Input.

• After that I dragged and dropped Slowly Changing Dimension item and connect the last Merge Join to that.

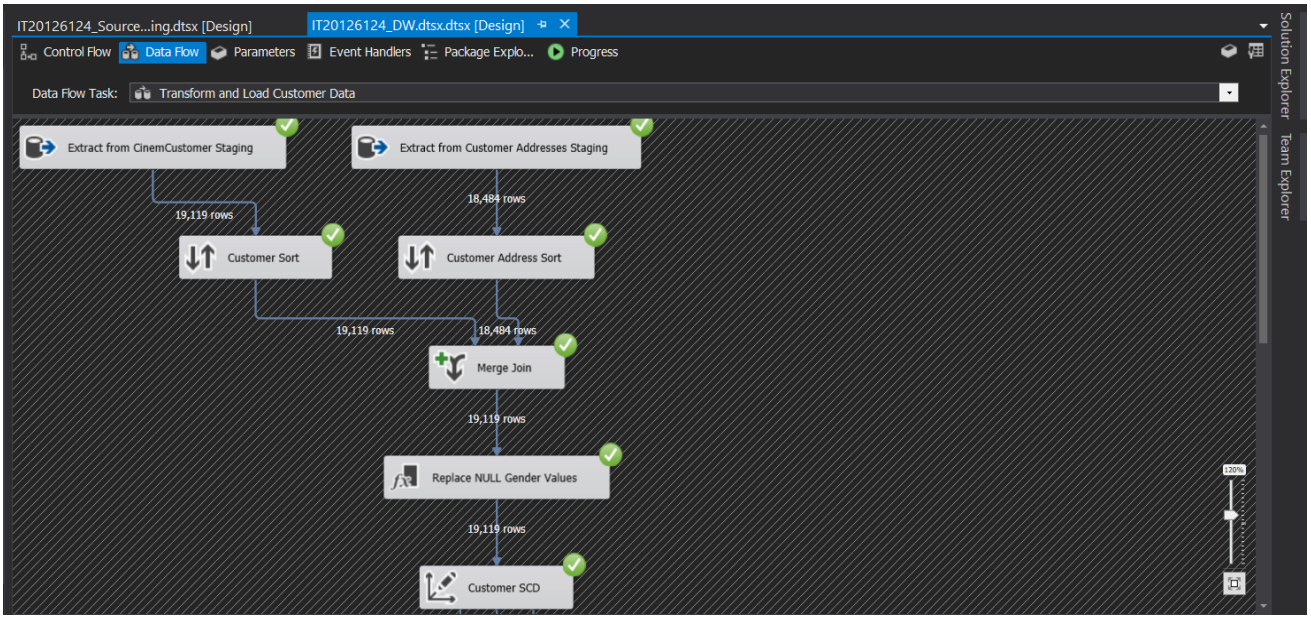
• In the SCD Configuration Wizard I set the configurations as below

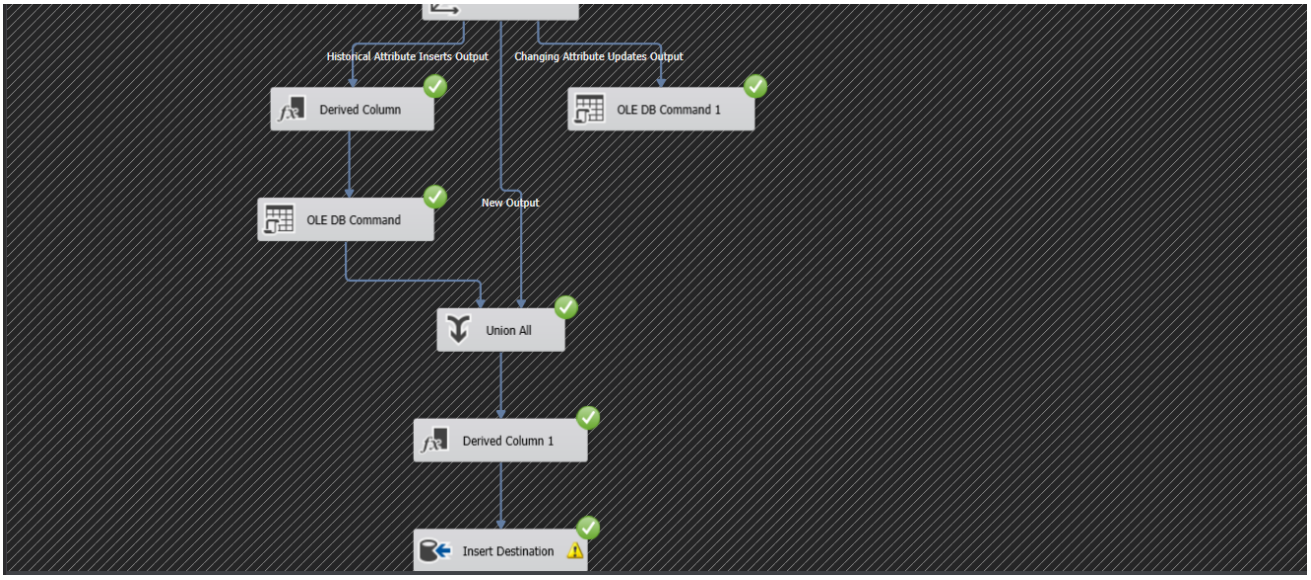

Slowly Changing Dimension Wizard
Select a Dimension Table and Keys
Select a dimension table to load and map columns in the
Connection manager:
DESKTOP-I2IVLMO.IT20126124_DW
New...
Table or view:
[dbo].[DimCinemaCustomer]
Input Columns | Dimension Columns | Key Type
AddressLine1 | AddressLine1 | Not a key column
AddressLine2 | AddressLine2 | Not a key column
AddressType | AddressType | Not a key column
CustomerID | AlternateCustome... | Business key
City | City | Not a key column
CountryRegi... | CountryRegionGr... | Not a key column
EmailAddress | EmailAddress | Not a key column
| EndDate |
FirstName | FirstName | Not a key column
Gender | Gender | Not a key column
InsertDate | InsertDate | Not a key column
LastName | LastName | Not a key column
ModifiedDate | ModifiedDate | Not a key column
PhoneNumb... | PhoneNumber | Not a key column
PhoneNumb... | PhoneNumberType | Not a key column
PostalCode | PostalCode | Not a key column
| StartDate |
StateProvin... | StateProvinceName | Not a key column
Help
< Back
Next >
Finish >>|
Cancel


Slowly Changing Dimension Wizard
Slowly Changing Dimension Columns
Manage the changes to column data in your slowly changing dimensions by setting the change type
Fixed Attribute
Select this type when the value in a column should not change. Changes are treated as errors.
Changing Attribute
Select this type when changed values should overwrite existing values. This is a Type 1 change.
Historical Attribute
Select this type when changes in column values are saved in new records. Previous values are saved in records marked as outdated. This is a Type 2 change.
Select a change type for slowly changing dimension columns:
Dimension Columns | Change Type
AddressLine1 | Historical attribute
AddressLine2 | Historical attribute
AddressType | Historical attribute
City | Historical attribute
CountryRegionGr... | Historical attribute
PhoneNumber | Changing attribute
PhoneNumberType | Changing attribute
PostalCode | Historical attribute
StateProvinceName | Historical attribute
Remove
Help
< Back
Next >
Finish >>|
Cancel



Once All Configurations done properly, it will automatically create the Slowly Changing Dimension as shown below.

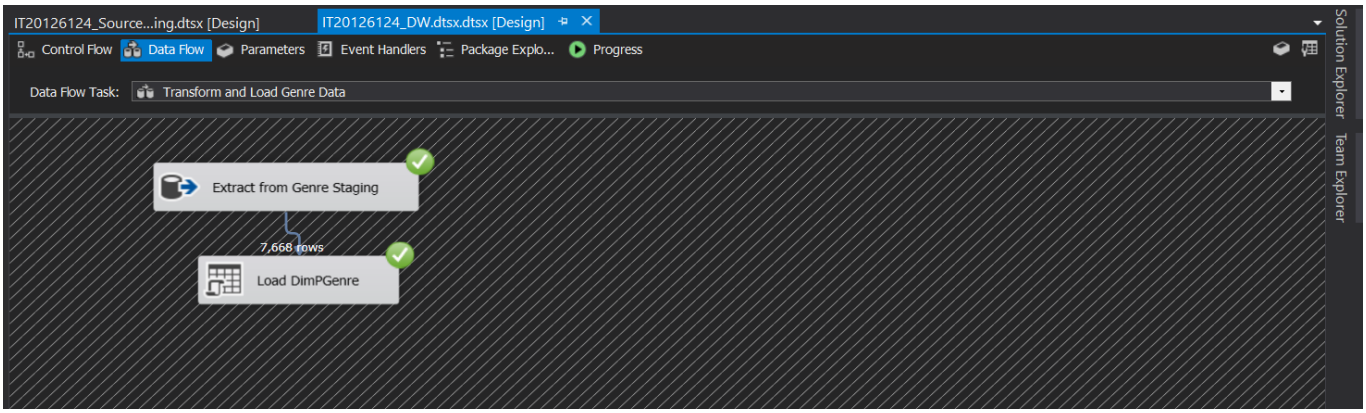




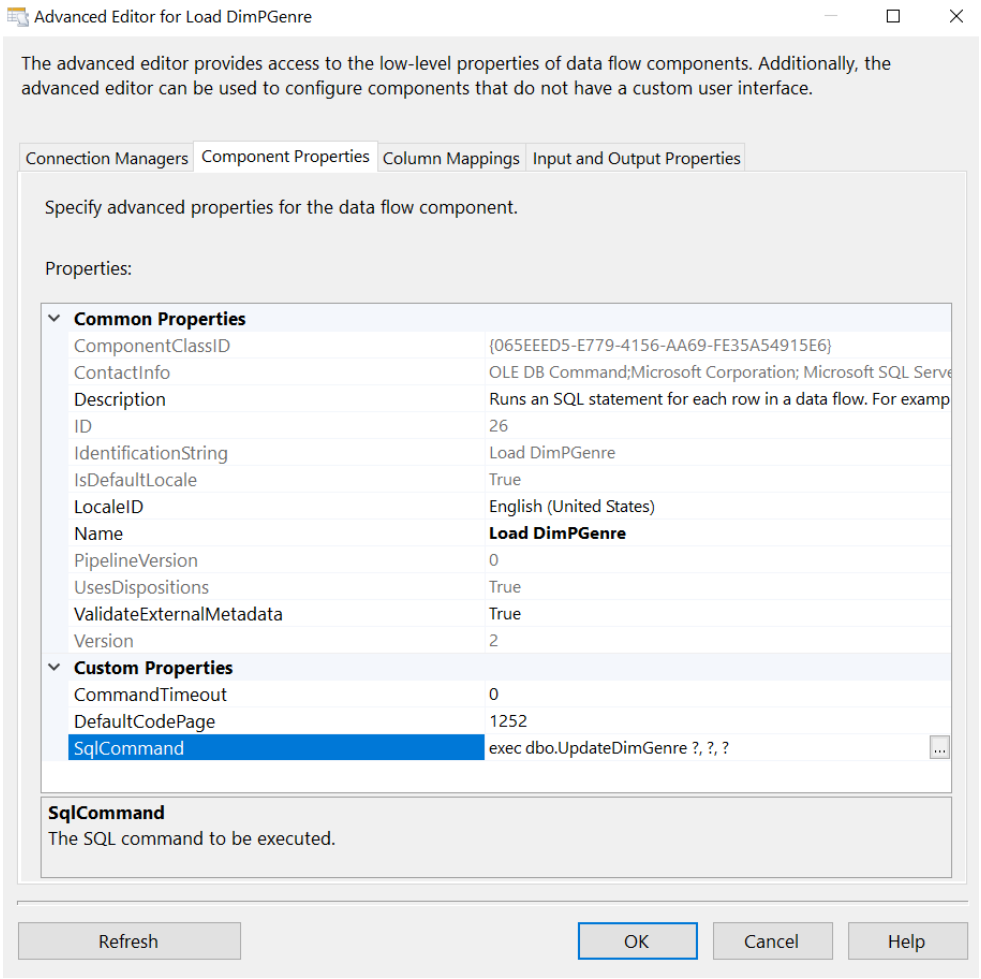
Transform and Load Genre Data

First created a Procedure called UpdateDimGenre and executed in the IT20126124_DW database

```
CREATE PROCEDURE dbo.UpdateDimGenre
@genreId int,
@genre nvarchar(50),
@country nvarchar(50)
AS BEGIN
if not exists (select genreSK
from [dbo].[DimGenre]
where AlternategenreId = @genreId) BEGIN
insert into dbo.DimGenre
(AlternategenreId, genre, country, InsertDate, ModifiedDate)
values
(@genreId, @genre, @country, GETDATE(), GETDATE()) END;
if exists (select genreSK
from dbo.DimGenre
where AlternategenreId = @genreId) BEGIN
update dbo.DimGenre
set genre = @genre ,
ModifiedDate = GETDATE()
where AlternategenreId = @genreId END;
END;
```



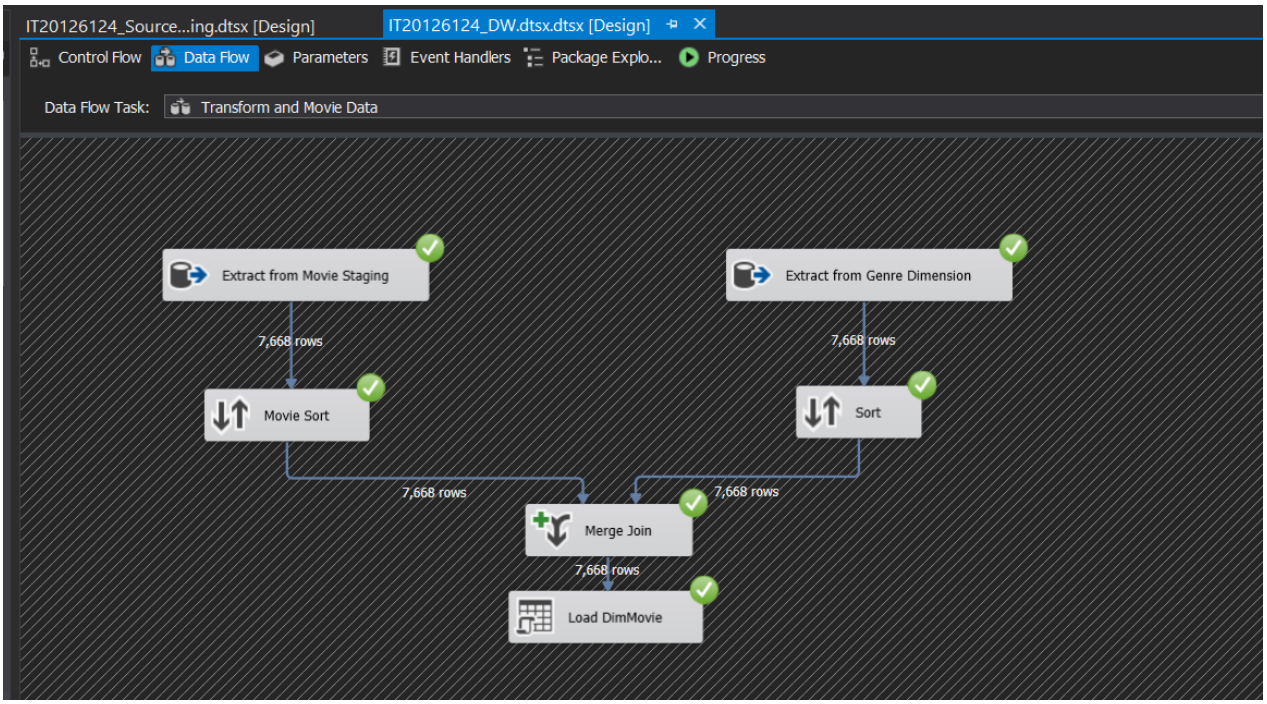
OLE DB Command SSIS tool used to execute, UpdateDimGenre procedure, it used to insert data into staging `Genre` to Dim`Genre` without data duplication.



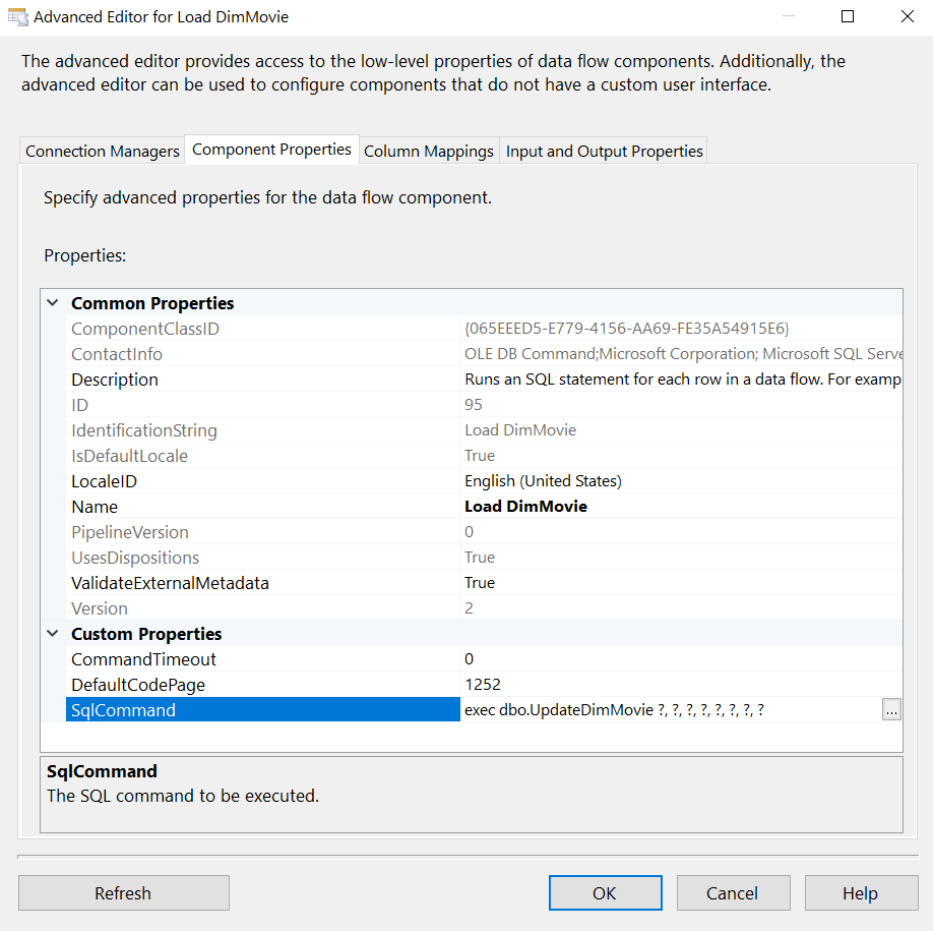
Transform and Load Movie Data

First created a Procedure called `UpdateDimMovie` and executed in the IT20126124_DW database

```
CREATE PROCEDURE dbo.UpdateDimMovie
@movieID int,
@name nvarchar(100),

@released nvarchar(100),
@director nvarchar(100),
@writer nvarchar(100),
@star nvarchar(100),
@company nvarchar(100),
@genreId int

AS BEGIN
if not exists (select movieSK
from [dbo].[DimMovie]
where AlternatemovieID = @movieID) BEGIN
insert into dbo.DimMovie
(AlternatemovieID, name, released, director, writer, star, company, genreId,
InsertDate, ModifiedDate)
values
(@movieID, @name, @released, @director, @writer, @star, @company, @genreId,
GETDATE(), GETDATE()) END;


if exists (select movieSK
from dbo.DimMovie
where AlternatemovieID = @movieID) BEGIN
update dbo.DimMovie
set name = @name ,
ModifiedDate = GETDATE()
where AlternatemovieID = @movieID END;
END;
```

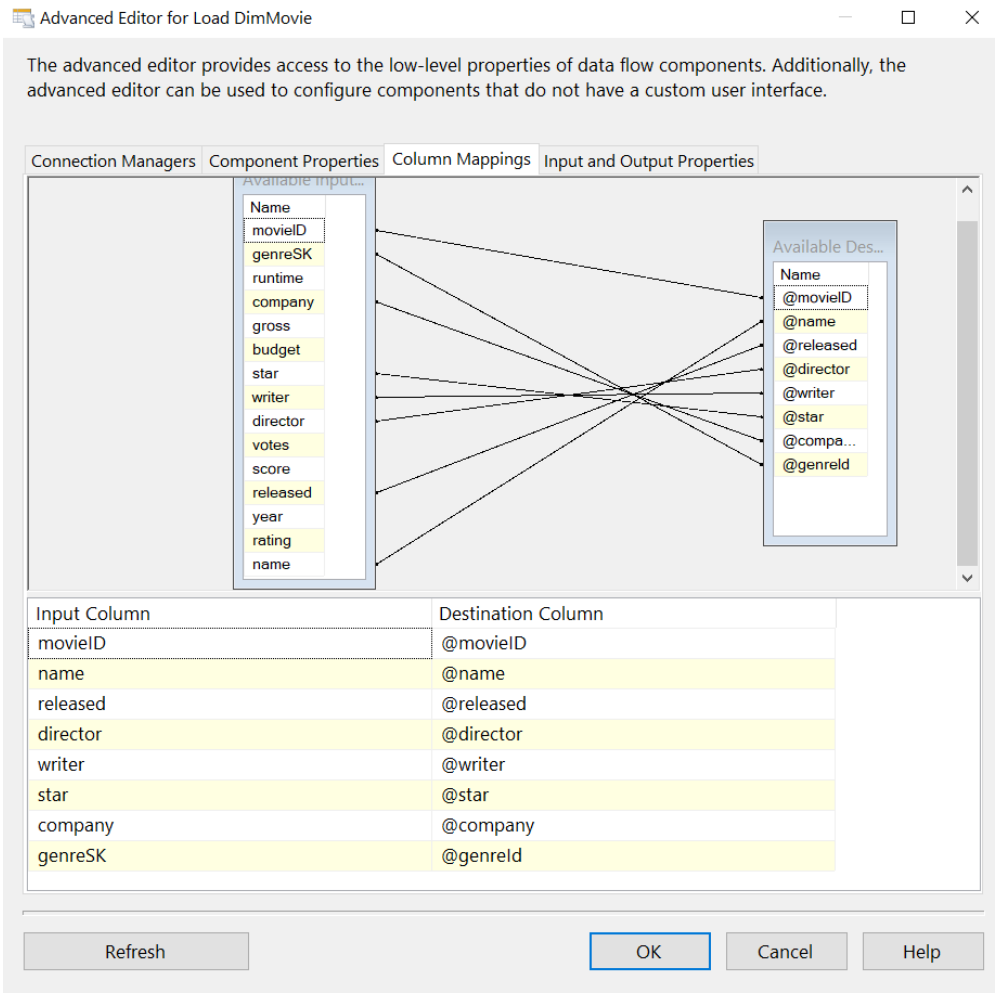


I Drag and drop OLE DB Command task, rename It as Load DimMovie link it with merge join.
After double clicking on Load DimMovie we can view Advanced Editor for Load DimMovie window.
Then used bellow code to execute above stored Procedures.



Above Stored Procedure ensure no duplicates are entered into the data warehouse table 'DimMovie'. If there is an existing record, it will be updated with the latest record coming from staging table 'StgMovie' else, if it is a new record, just insert it.

In Column Mappings tab map, the columns to the variables accordingly. here I had map genreSK as the input for @genreID.

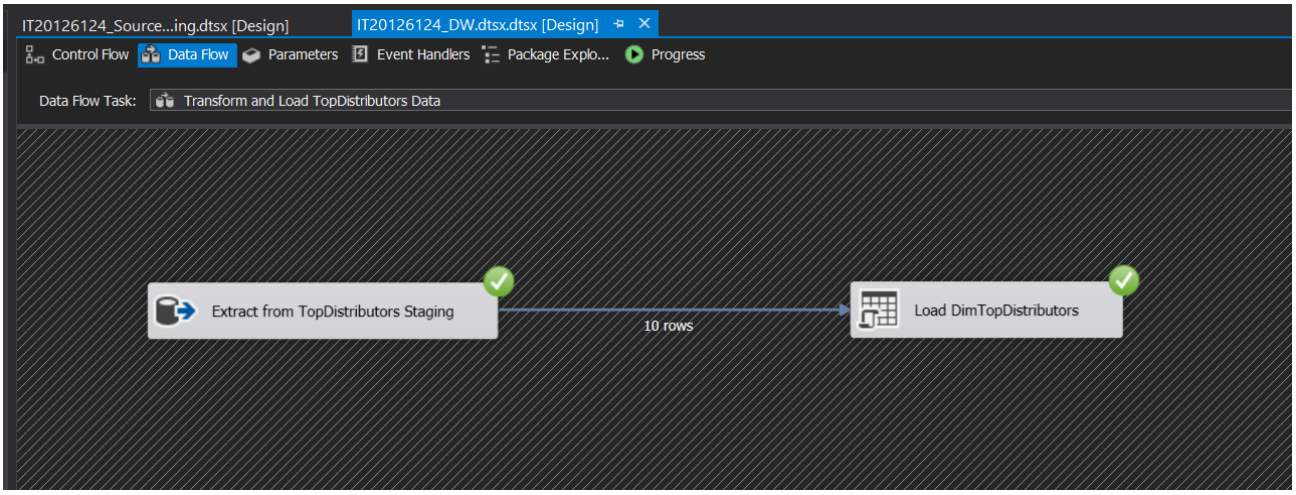


Transform and Load TopDistributors Data

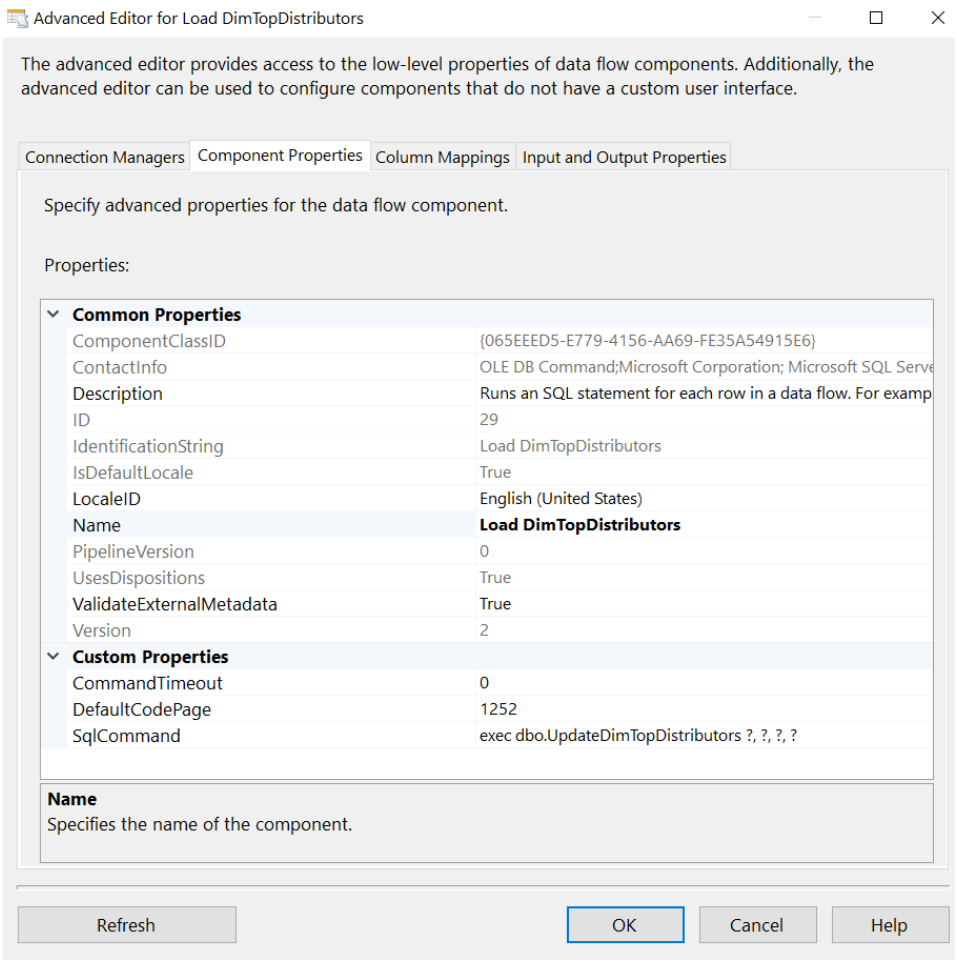
First created a Procedure called UpdateDimTopDistributors and executed in the IT20126124_DW database

```
CREATE PROCEDURE dbo.UpdateDimTopDistributors
@RANK int,
@DISTRIBUTORS nvarchar(50),
@MOVIES int,
@TOTAL_GROSS money
AS BEGIN
if not exists (select DistributorsSK
from [dbo].[DimTopDistributors]
where AlternateRANK = @RANK) BEGIN
insert into dbo.DimTopDistributors
(AlternateRANK, DISTRIBUTORS, MOVIES, TOTAL_GROSS, InsertDate, ModifiedDate)
values
(@RANK, @DISTRIBUTORS, @MOVIES, @TOTAL_GROSS, GETDATE(), GETDATE())
```

```
END;
if exists (select DistributorsSK
from dbo.DimTopDistributors
where AlternateRANK = @RANK) BEGIN
update dbo.DimTopDistributors
set DISTRIBUTORS = @DISTRIBUTORS ,
ModifiedDate = GETDATE()
where AlternateRANK = @RANK END;
END;
```



OLE DB Command SSIS tool used to execute, UpdateDimTopDistributors procedure, it used to insert data into staging TopDistributors to DimTopDistributors without data duplication.



Load to FactSales Fact

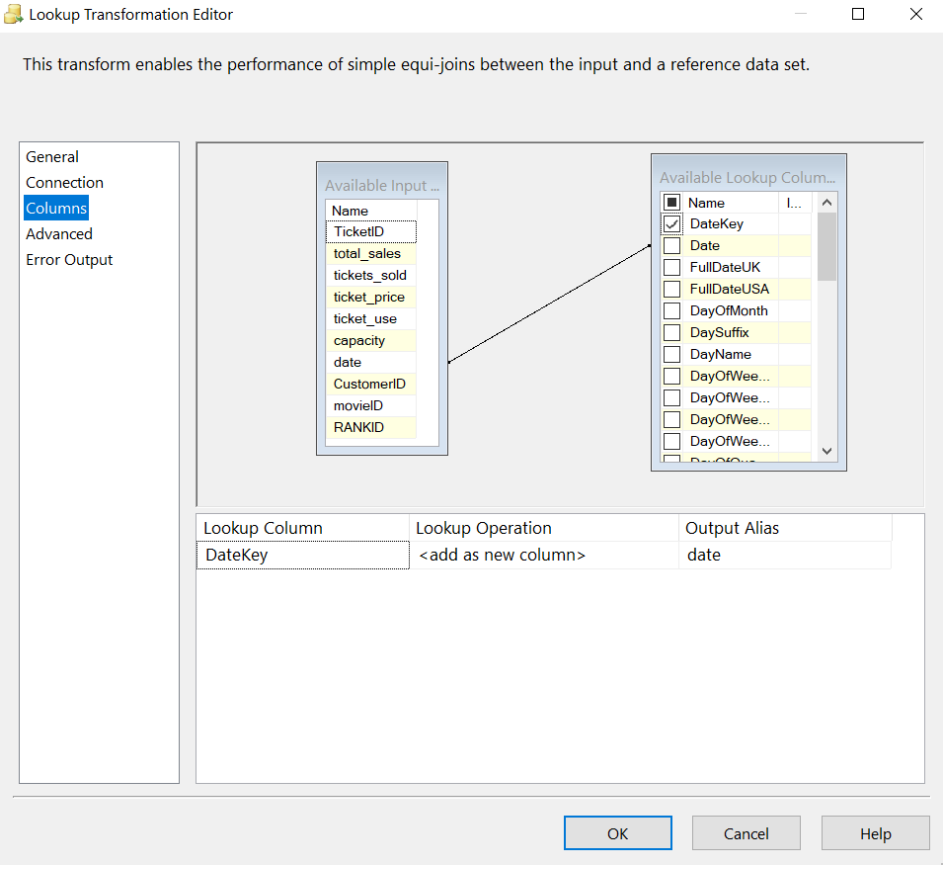
In the IT20126124_DW.dtsx, add another Data Flow Task and join the previous data flow task with the new data flow task.

Renamed the new Data Flow Task as Transform and Load FactCinema

Then I dragged and dropped the OLE DB source and configure it to fetch data from dbo. StgcinemaTicket table

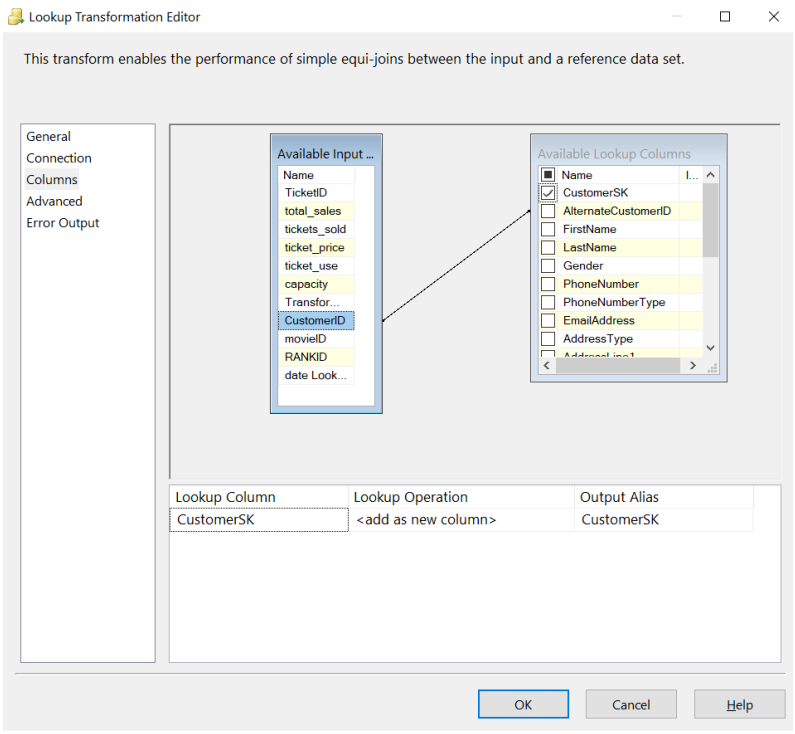
Effective to Date Lookup –

Here I map Effective to Date Input with DateKey lookup to obtain date key.

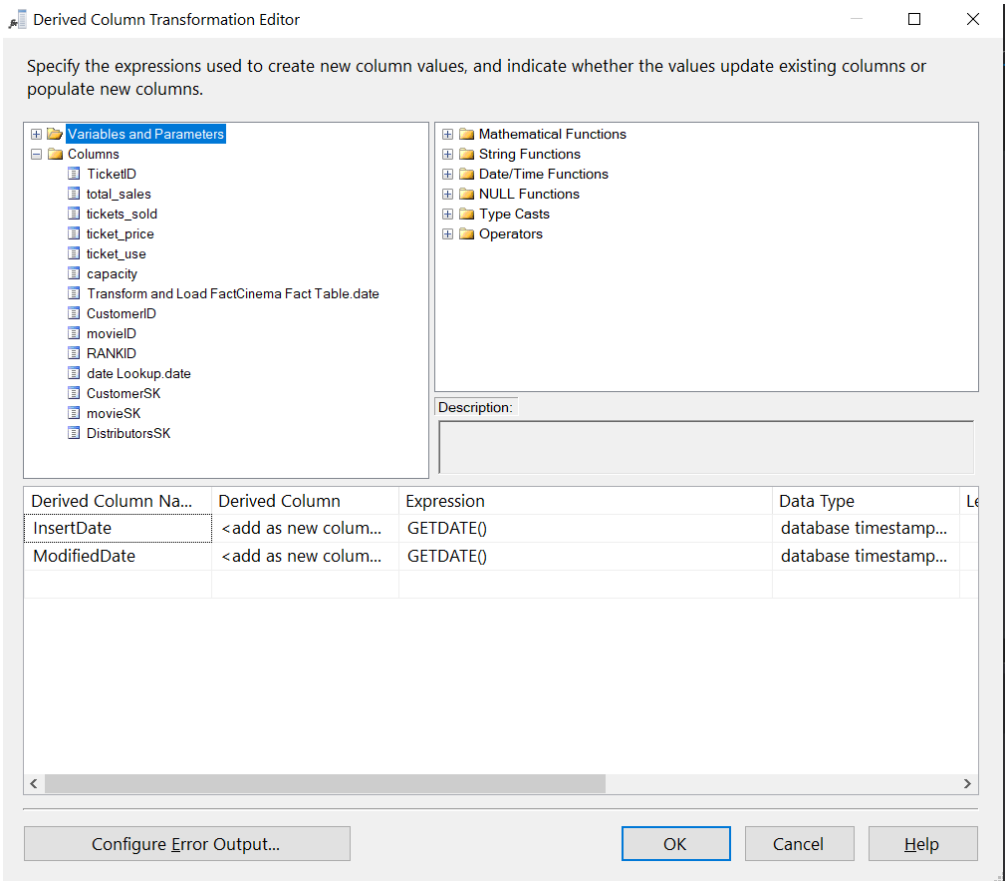


Customer Lookup

Here I map CustomerID with AlternateCustomerID to obtain CustomerSK

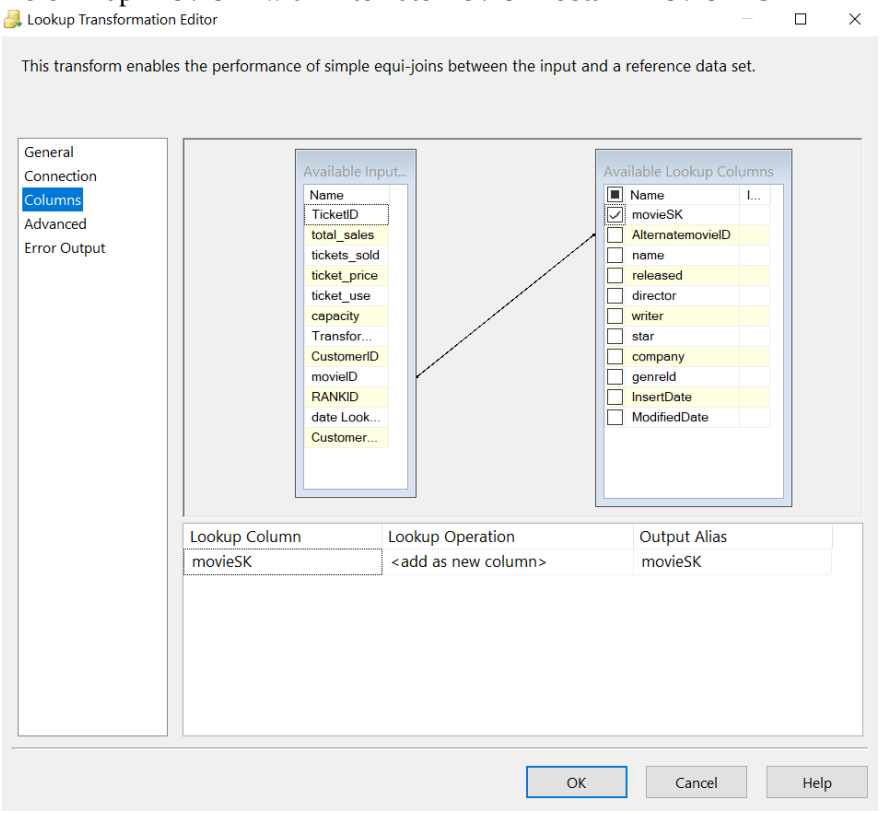


Derived Insert and Modified Date Fields –



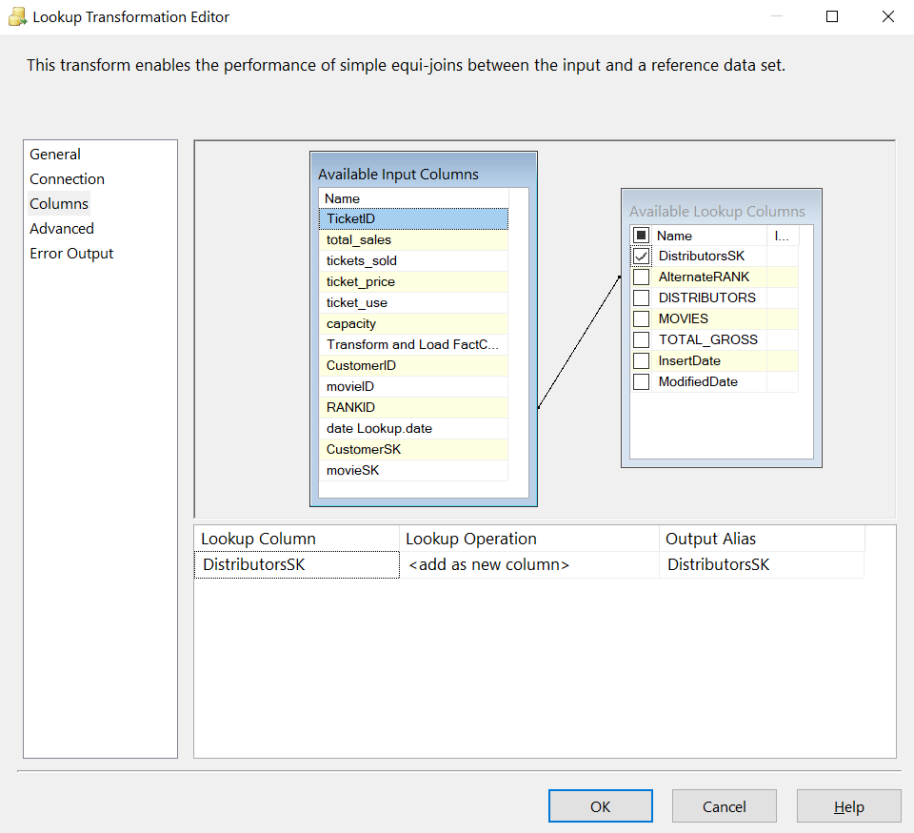
Movie Lookup

Here I map MovieID with AlternateMovieID obtain MovieID SK



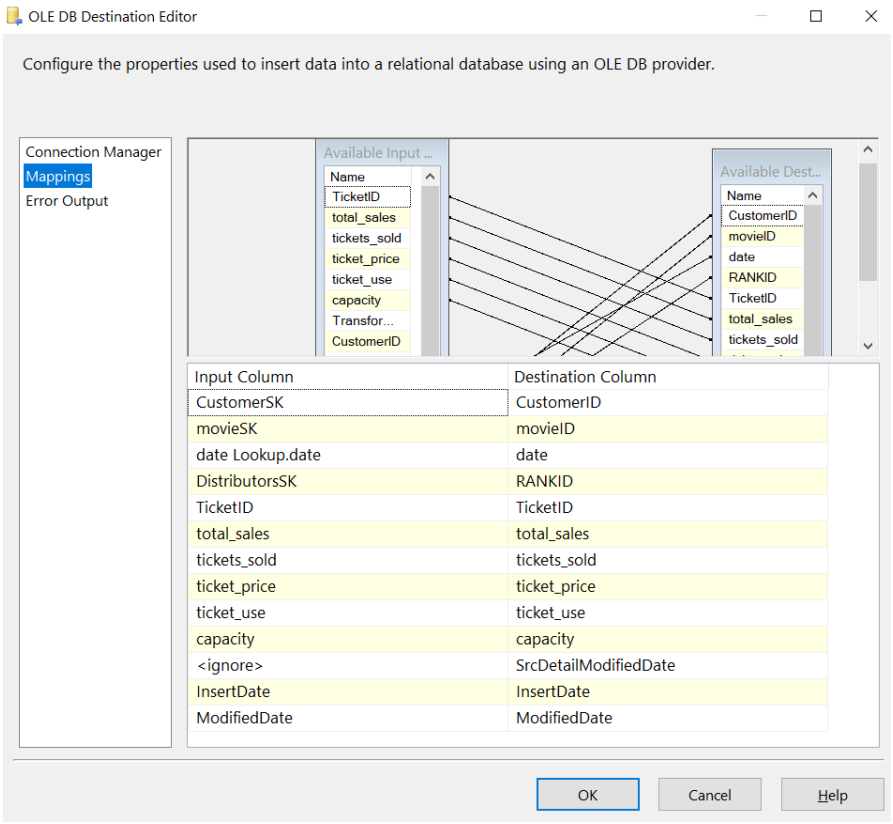
DimTopDistributors Lookup

Here I map TicketID with Alternate TicketID obtain DistributorSK

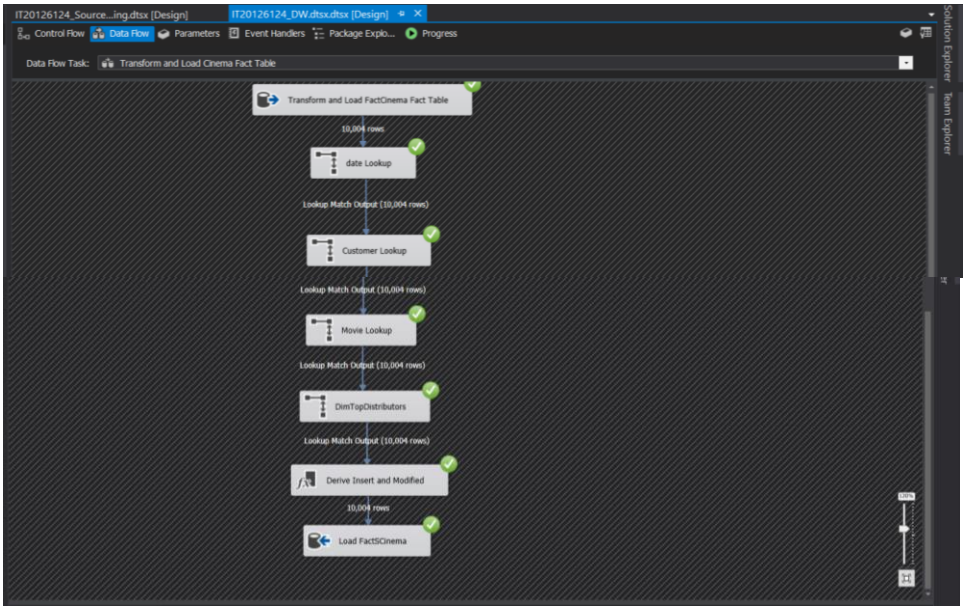


Take a another OLE DB Destination component and rename it as Load FactSCinema
And connected it with above ‘Derive Insert and modified Date Fields’

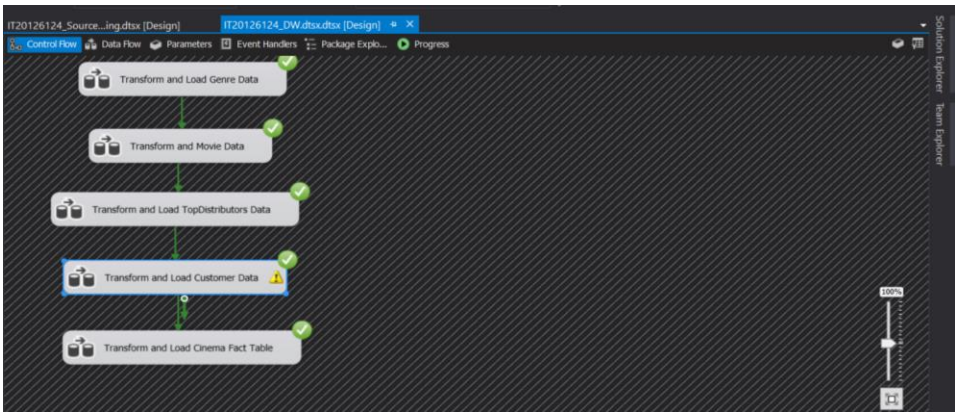
I mapped the input columns with Destination Columns as below.



Outcome of the Transform and Load Cinema Fact Table:



Final IT20126124_DW.dtsx package Control Flow:



At the end of the staging I have connected the data warehousing package to the end of the data staging package using an execute package task editor.
So, when executing the staging it will execute the data warehousing package as well

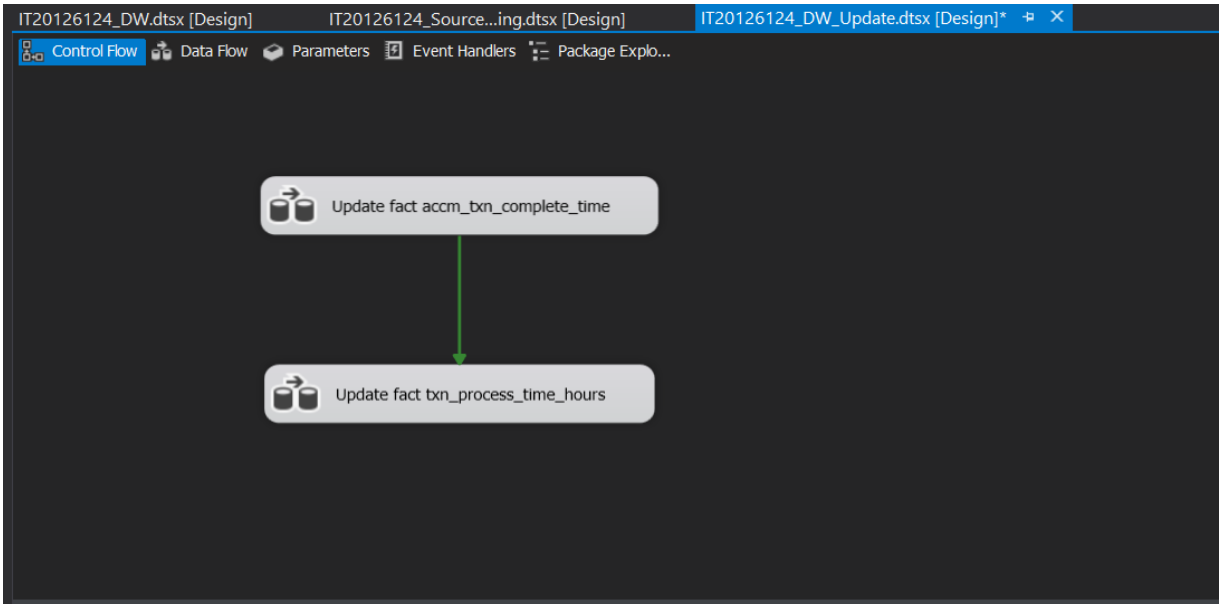
Step 06: ETL development – Accumulating fact tables

Extended the original FactCinema table by adding accm_txn_create_time, accm_txn_complete_time, txn_process_time_hours columns. Kept accm_txn_create_time to be equal to the current system date when you load data to your fact table and Prepared a separate data set for accm_txn_complete_time in csv file format.

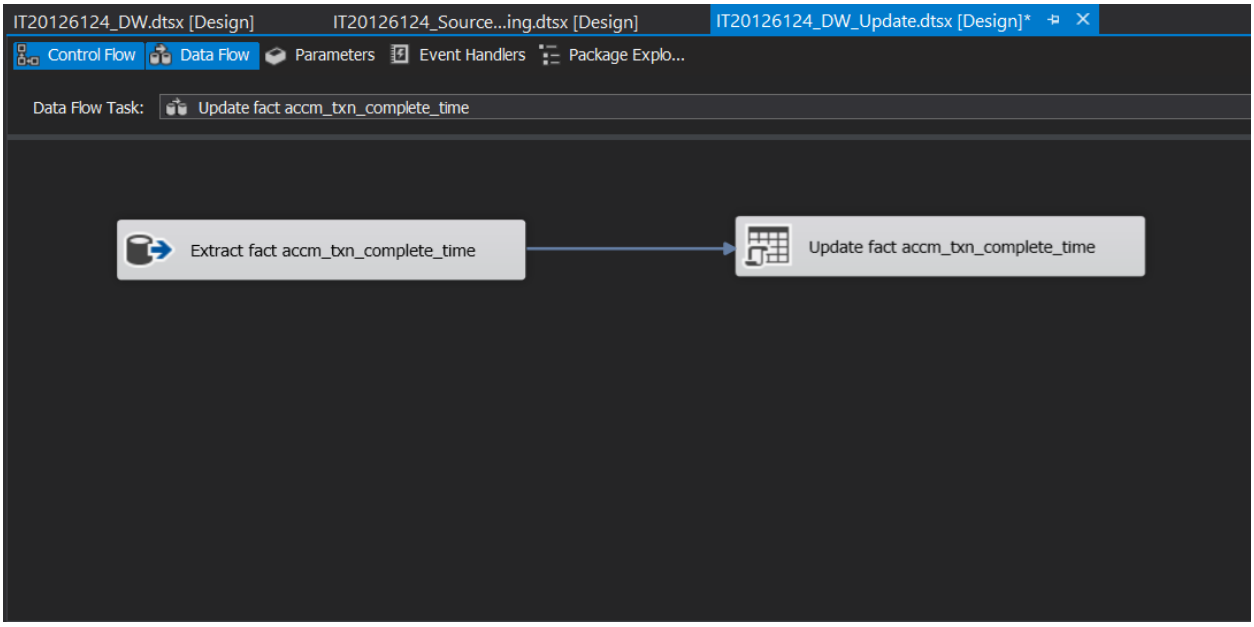
DESKTOP-I2IVLMO....- dbo.FactCinema | SQLQuery28.sql -...sus VivoBook (66)) | SC

Column Name	Data Type	Allow Nulls
CustomerID	int	☑
movieID	int	☑
date	int	☑
RANKID	int	☑
TicketID	int	☑
total_sales	int	☑
tickets_sold	int	☑
ticket_price	int	☑
ticket_use	int	☑
capacity	float	☑
SrcDetailModifiedDate	datetime	☑
InsertDate	datetime	☑
ModifiedDate	datetime	☑
accm_txn_create_time	datetime	☑
accm_txn_complete_time	datetime	☑
txn_process_time_hours	int	☑
		☐

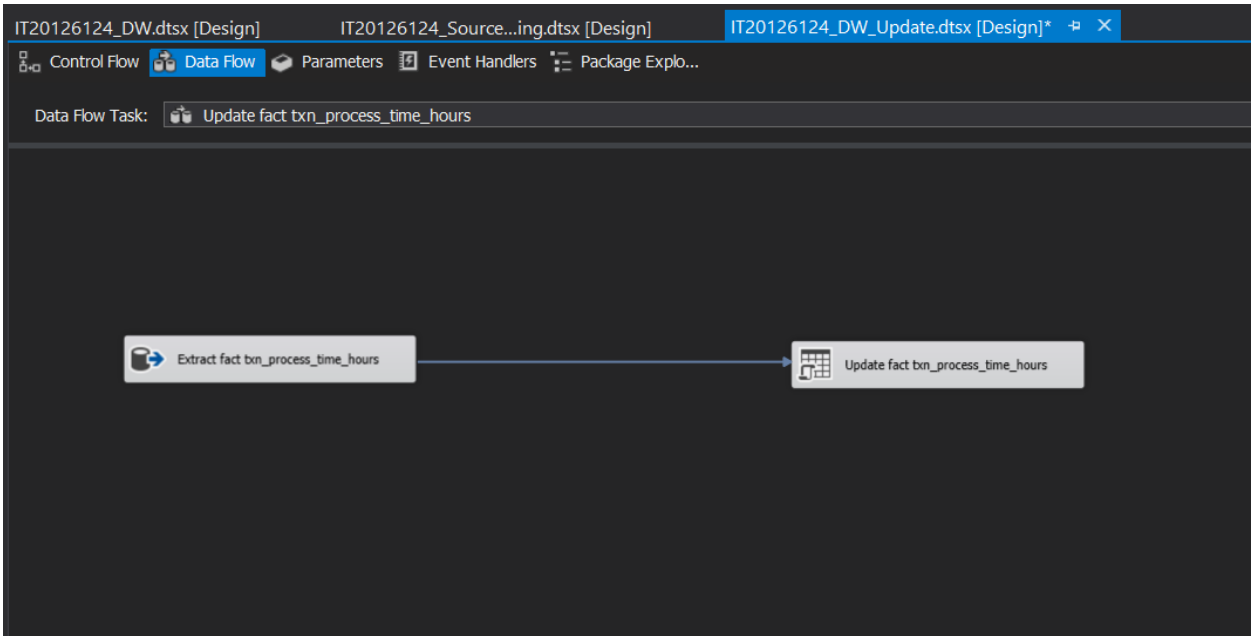
Created a specific SSIS package called IT20126124_DW_Update and created a data flow to update FactCinema Coverage table by adding a new column called accm_txn_complete_time and txn_process_time_hours.



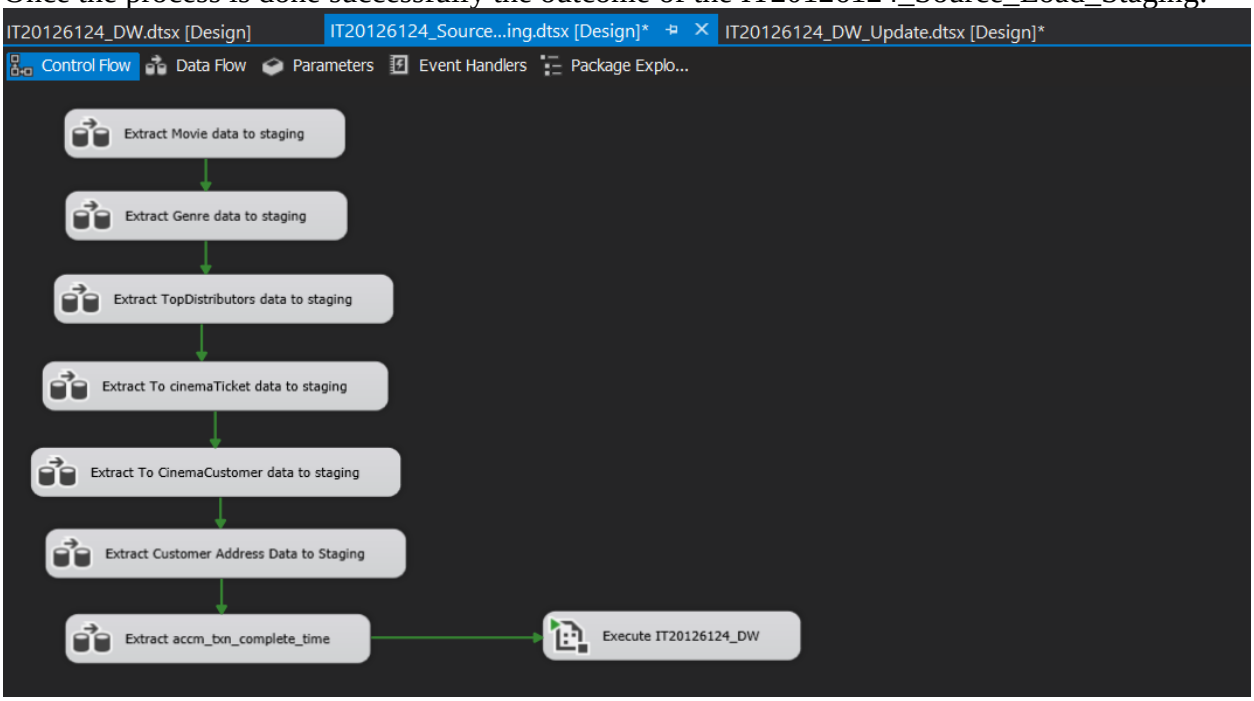
Then Transform and Load data to accm_txn_complete_time to update its relevant column in the FactCinema



Then we extract data from the FactCinema to update the txn_process_time_hours column in the fact table.



Once the process is done successfully the outcome of the IT20126124_Source_Load_Staging:



Step 6 Process:

As mentioned in the assignment guidelines first created the csv file that contains natural key of fact table, transaction complete time (accm_txn_complete_time) and extract data from the created csv file into staging (IT20126124_Source_Load_Staging) as a table (accm_txn_complete_time).

After that create accm_txn_create_time, accm_txn_complete_time, txn_process_time_hours

three columns in the FactCinema Coverage Fact table.

As the next step create a derived column in a “Transform and Load FactCinema Table to DW” data flow and map it with accm_txn_complete_time as mentioned below.

Next create new SSIS package to load the accm_txn_complete_time and txn_process_time_hours. After that I created a data flow task to map the accm_txn_complete_time in fact table.

Then I created a data flow task to map the txn_process_time_hours in fact table.

END