Sri Lanka Institute of Information Technology

Data Warehousing & Business Intelligence

Assignment 01
IT Number – IT20083182
Submitted by – Senadeera N. A. J. N.
Batch – Year 03 Semester 01 (Y3S1.5.1 (DS))

Contents

01.Data set selection

1.1 Data set name: **Hotel Reviews**

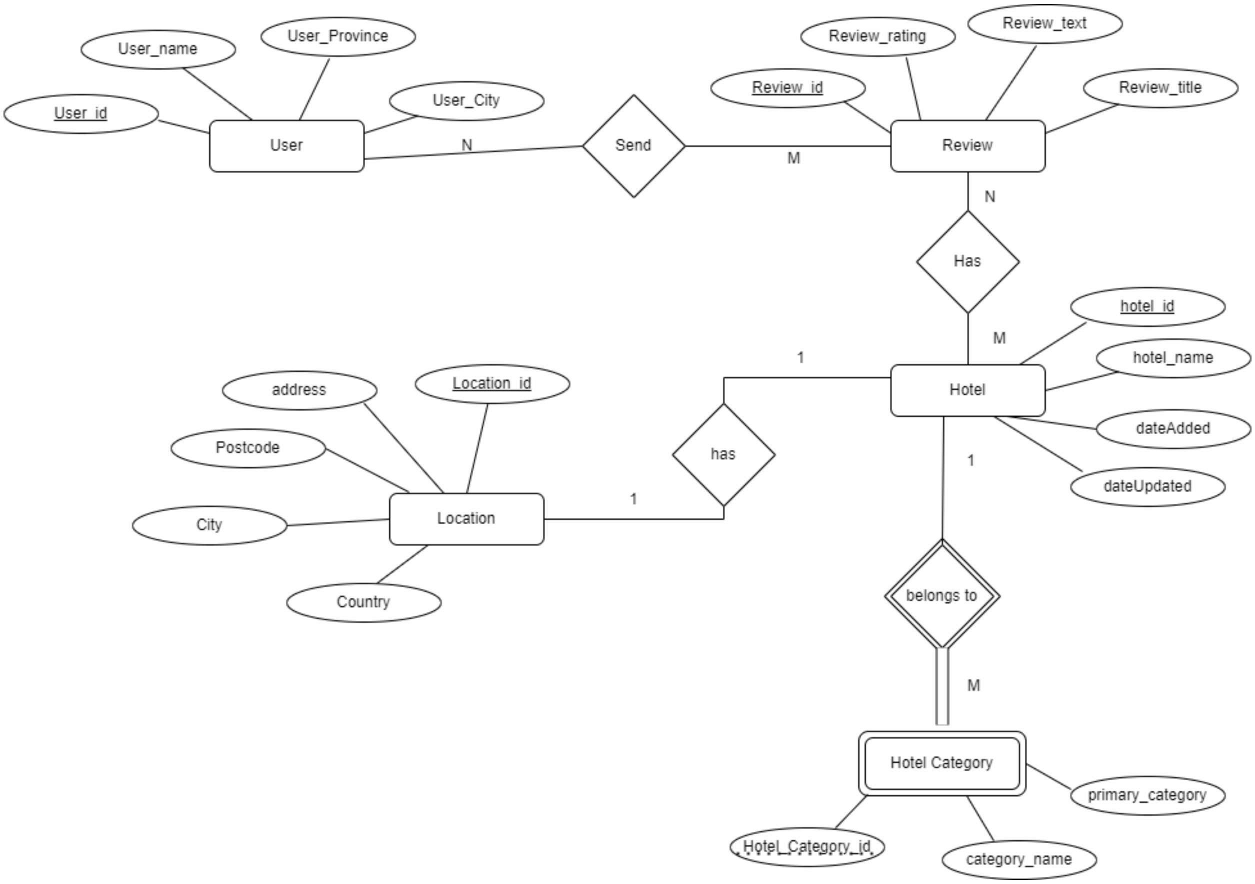
Provided by: kaggle.com
Source link: https://www.kaggle.com/datasets/datafiniti/hotel-reviews?select=Datafiniti_Hotel_Reviews_Jun19.csv

1.2 About Dataset:

This is a list of 1,000 hotels and their reviews provided by Datafiniti's Business Database. The dataset includes hotel location, name, rating, review data, title, username, and more.
You can use this data to compare hotel reviews on a state-by-state basis; experiment with sentiment scoring and other natural language processing techniques. The review data lets you correlate keywords in the review text with ratings. E.g.:

- What are the bottom and top states for hotel reviews by average rating?
- What is the correlation between a state's population and their number of hotel reviews?
- What is the correlation between a state's tourism budget and their number of hotel reviews?

1.3 ER Diagram



02.Preparation of data

All the data sources are provided in csv format by the web site. In preparation of data sources, some changes have done for the source format (some columns were added, separated into another table) of the given files as converting into text files and importing csv files into a source database.

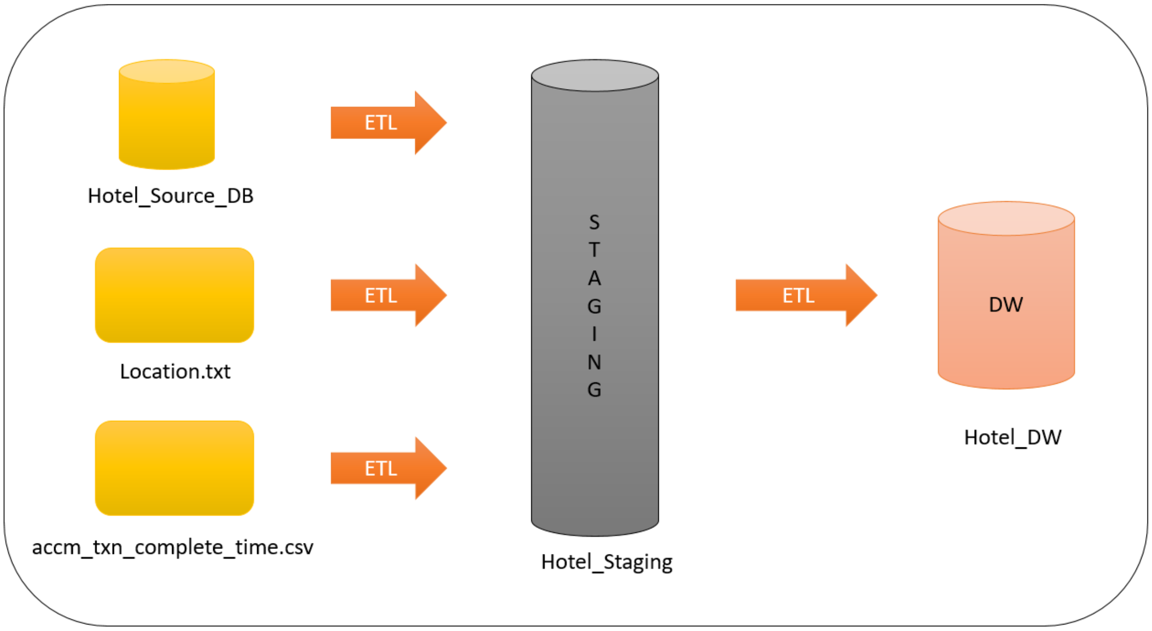
Final State of Preparation of the source data formats before Transforming data =>

1. CSV files (.csv)
    - Hotel CSV File.
    - Hotel Category CSV File.
    - User CSV File.
    - Review CSV File.
    - accm_txn_complete_time CSV

These csv files are imported into SSMS, database created as **Hotel_Source_DB** database.

2. Text file (.txt)
    - Location text file

03.Solution Architecture



Hotel_Staging.

- accm_txn_complete_time
- stgHotel
- stgHotalCategory
- stgReview
- StgUser
- stgLocation

Hotel_DW

- DimDate
- DimHotelCategory
- DimLocation
- DimReview
- DimUser
- FactHotel

Architecture Components.

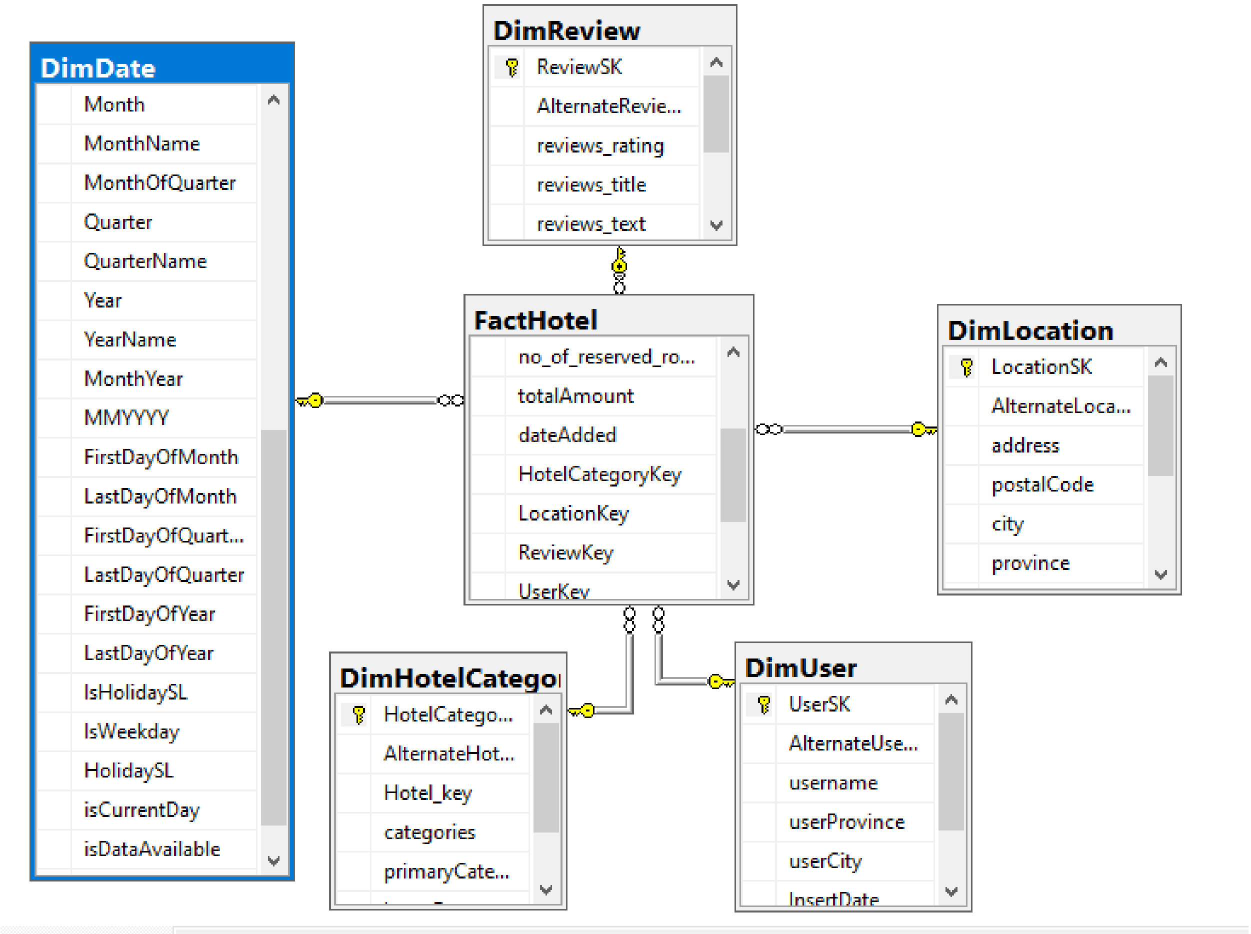
- Data Sources.
  Operational System (**Accumulating**).
  External Sources.

- Extract, Transform and Load.
  Extract – reading data from source systems.
  Transform – Combine data from multiple sources, De-duplicating.

- Data Warehouse
  EDW and Data Mart.
  Dimensional Modeling- Facts and Dimensions.
  Many schemas – In here I use star schema.

04.Data Warehouse Design & Development

Relational Diagram – Star Schema

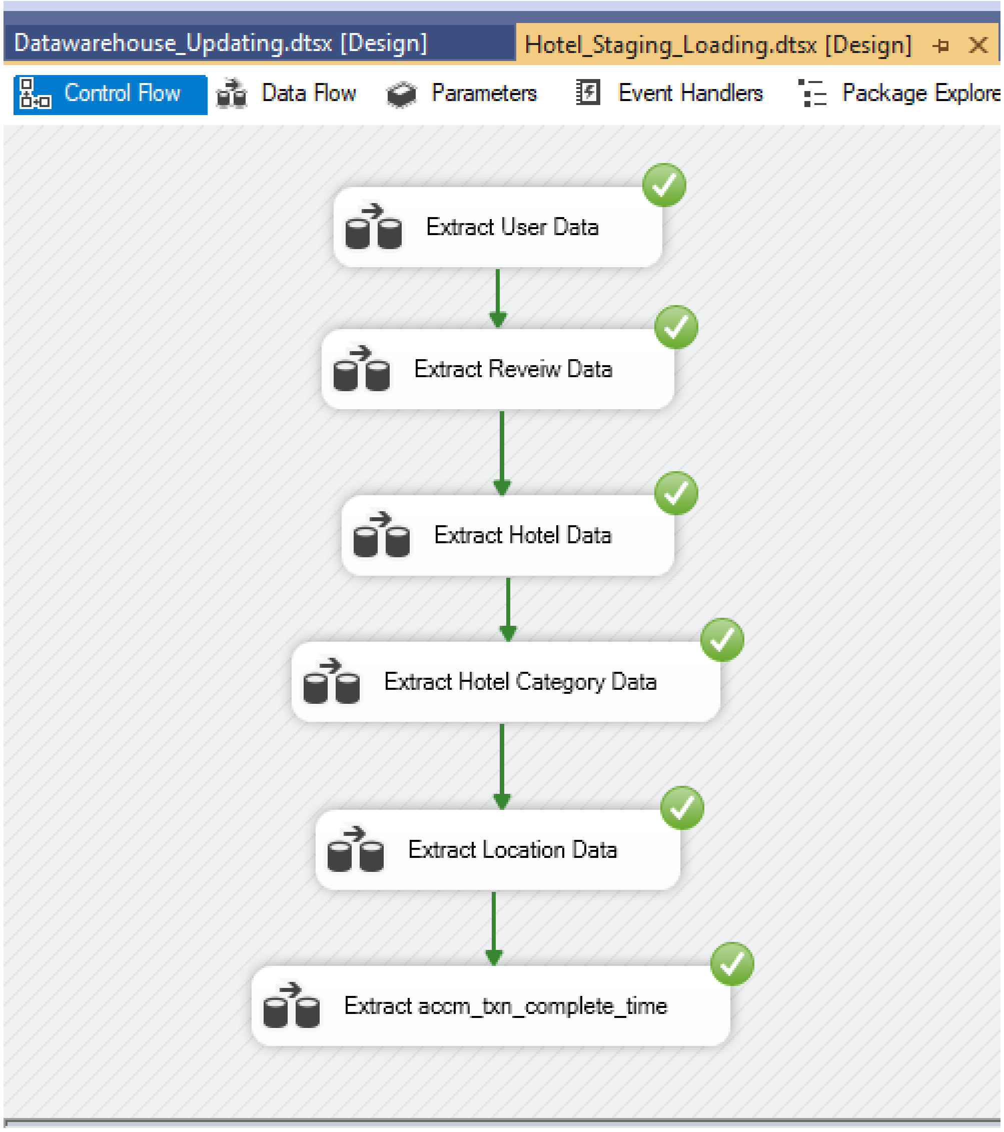


DimLocation is **slowly changing dimensions**. Address and city may be changed in future. Therefore, I get it as slowly changing attribute.

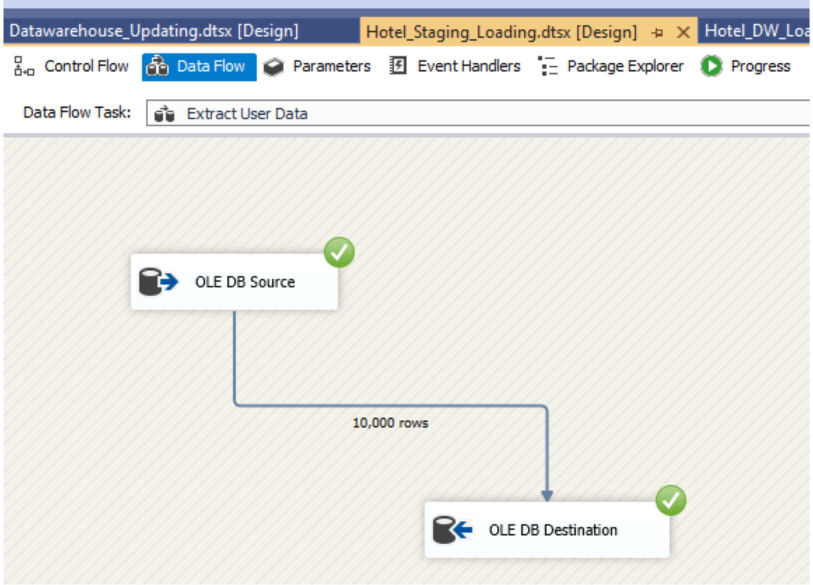
Address -> PostalCode -> City -> Province ->Country
This is the Hierarchies (Location table.)

05.ETL Development

5.1 ETL – Source to Staging

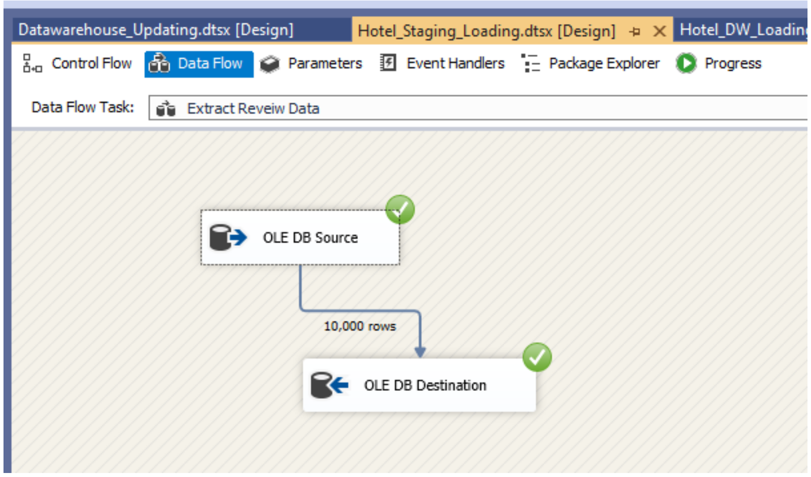


5.1.1 Load data User to staging



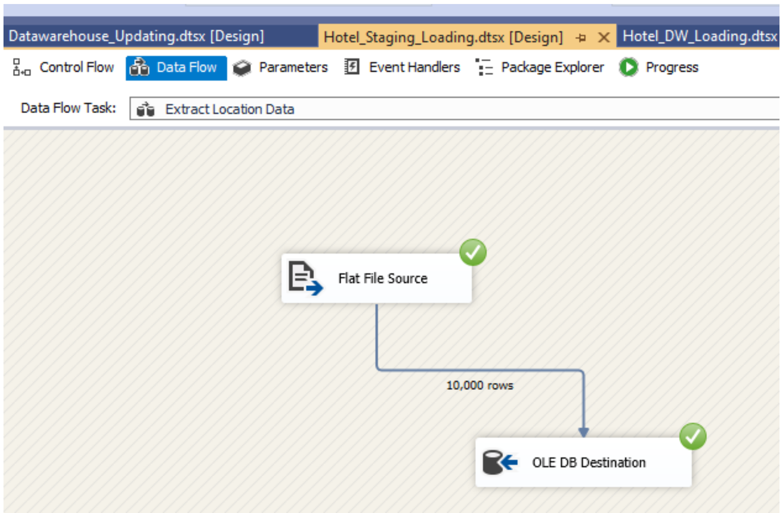
Data Flow Task: Extract User Data

5.1.2 Load data Review to staging

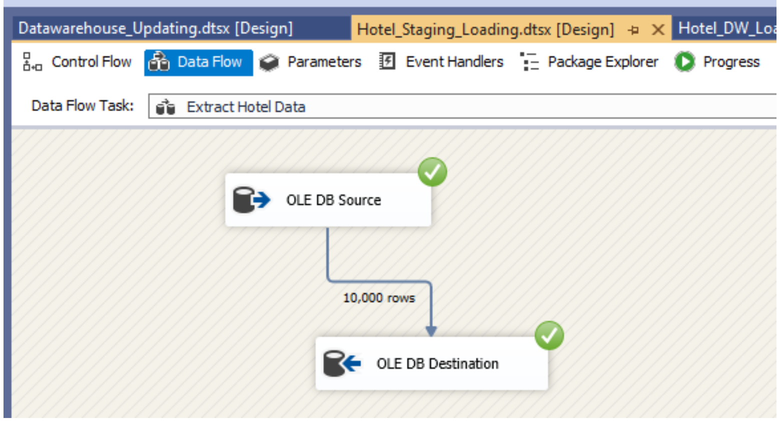


Data Flow Task: Extract Reveiw Data

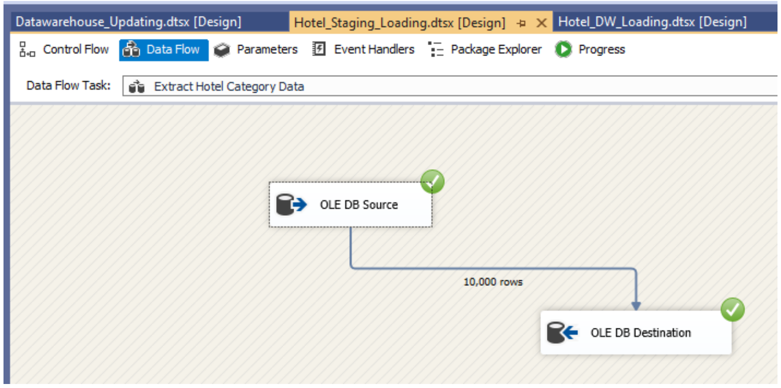
5.1.3 Load data Location to staging (.txt file)



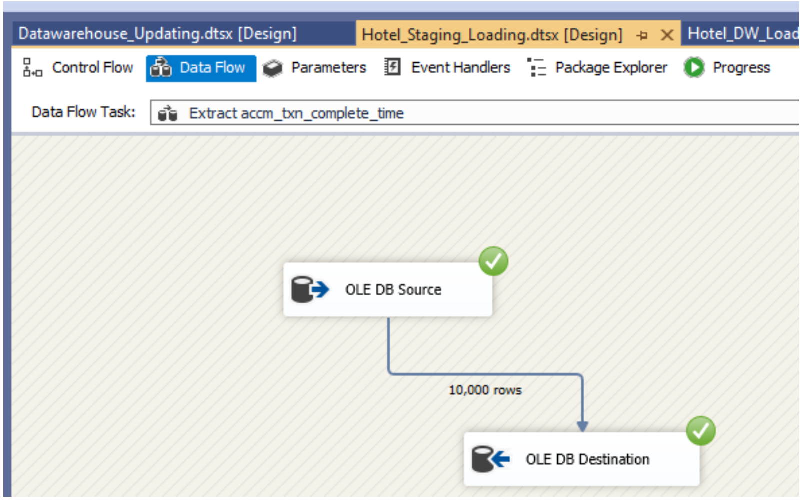
5.1.4 Load data Hotel to staging



5.1.5 Load data Hotel Category to staging


Datawarehouse_Updating.dtsx [Design]
Hotel_Staging_Loading.dtsx [Design]
Hotel_DW_Loading.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Data Flow Task:
Extract Hotel Category Data
OLE DB Source
10,000 rows
OLE DB Destination

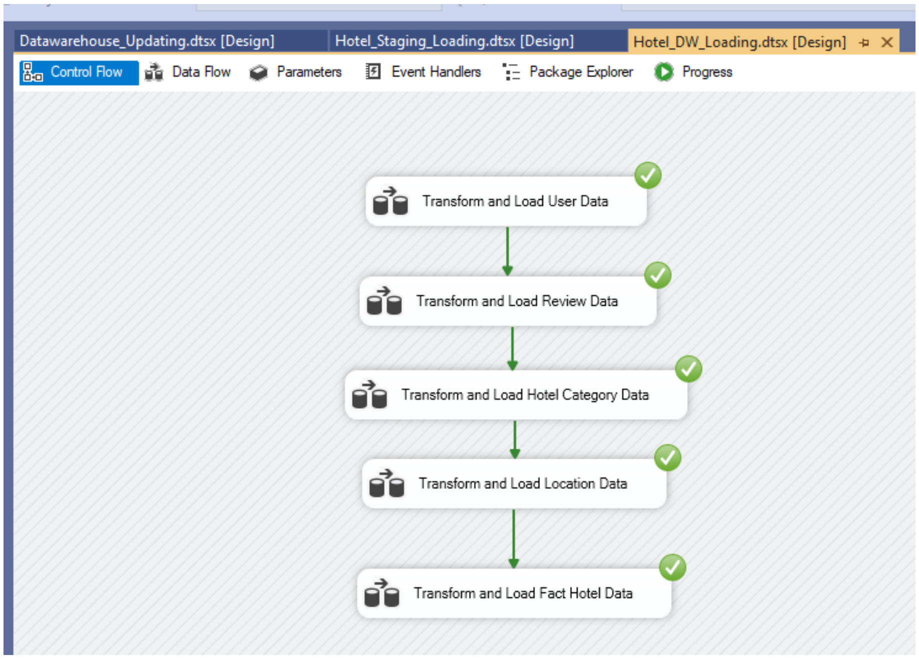
5.1.6 Load data accm_txn_complete_time to staging



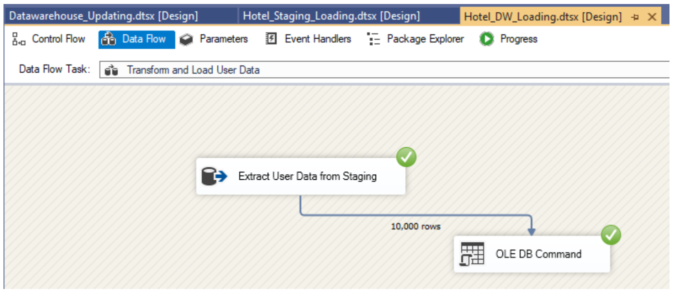
Data Flow Task: Extract accm_txn_complete_time

06.Staging to DW

6.1 ETL System to Datawarehouse


Datawarehouse_Updating.dtsx [Design]
Hotel_Staging_Loading.dtsx [Design]
Hotel_DW_Loading.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Transform and Load User Data
Transform and Load Review Data
Transform and Load Hotel Category Data
Transform and Load Location Data
Transform and Load Fact Hotel Data

6.1.1 Transfer and Load DimUser Data from staging



Datawarehouse_Updating.dtsx [Design] | Hotel_Staging_Loading.dtsx [Design] | Hotel_DW_Loading.dtsx [Design]

Control Flow | Data Flow | Parameters | Event Handlers | Package Explorer | Progress

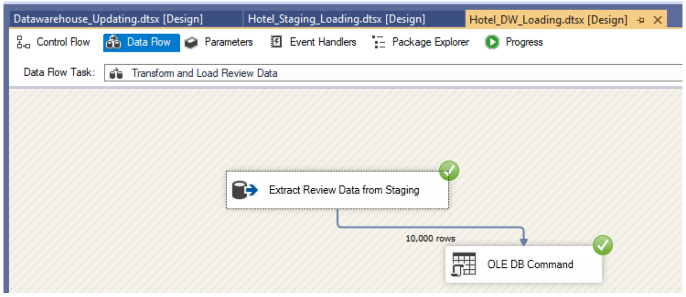
Data Flow Task: Transform and Load User Data

Extract User Data from Staging

10,000 rows

OLE DB Command

6.1.2 Transfer and Load DimReview Data from staging



Datawarehouse_Updating.dtsx [Design] | Hotel_Staging_Loading.dtsx [Design] | Hotel_DW_Loading.dtsx [Design]

Control Flow | Data Flow | Parameters | Event Handlers | Package Explorer | Progress

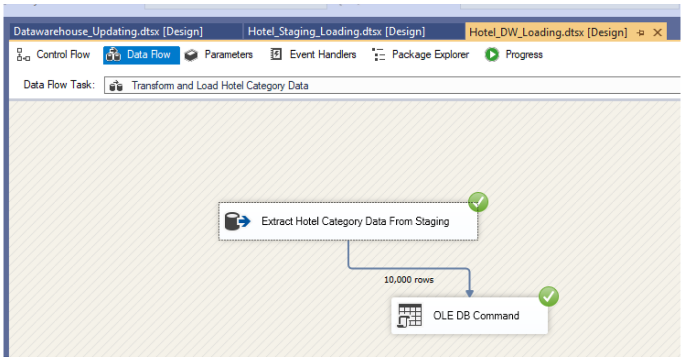
Data Flow Task: Transform and Load Review Data

Extract Review Data from Staging

10,000 rows

OLE DB Command

6.1.3 Transfer and Load DimHotelCategory Data from staging



Datawarehouse_Updating.dtsx [Design] | Hotel_Staging_Loading.dtsx [Design] | Hotel_DW_Loading.dtsx [Design]

Control Flow | Data Flow | Parameters | Event Handlers | Package Explorer | Progress

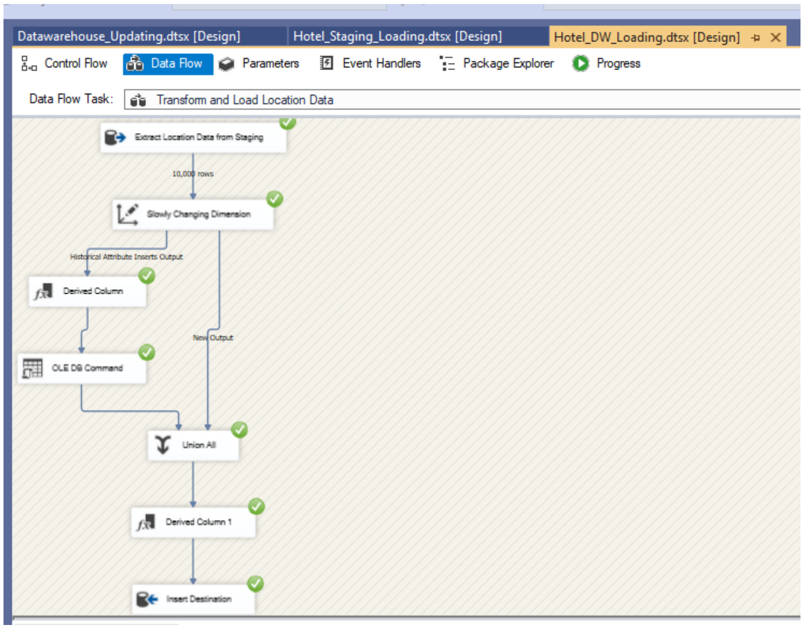
Data Flow Task: Transform and Load Hotel Category Data

Extract Hotel Category Data From Staging

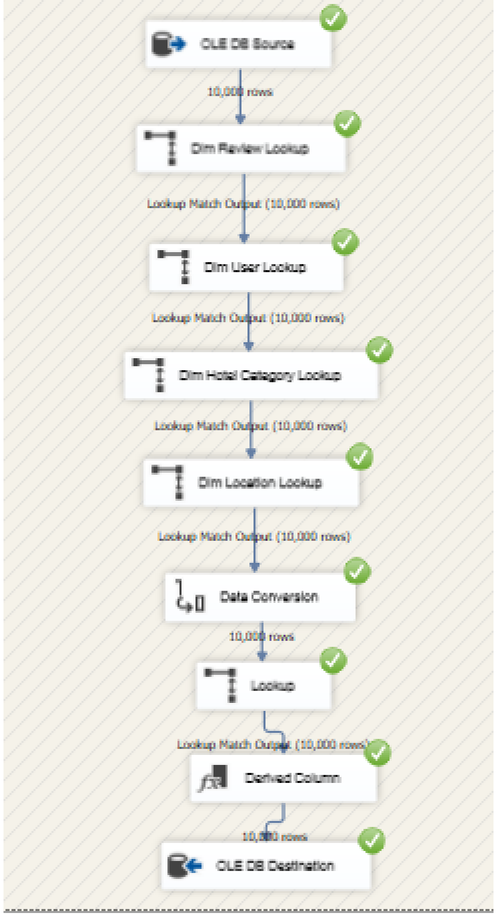
10,000 rows

OLE DB Command

6.1.4 Transfer and Load DimLocation Data from staging (Slowly changing dimension)


Datawarehouse_Updating.dtsx [Design]
Hotel_Staging_Loading.dtsx [Design]
Hotel_DW_Loading.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Data Flow Task:
Transform and Load Location Data
Extract Location Data from Staging
10,000 rows
Slowly Changing Dimension
Historical Attribute Inserts Output
Derived Column
New Output
OLE DB Command
Union All
Derived Column 1
Insert Destination

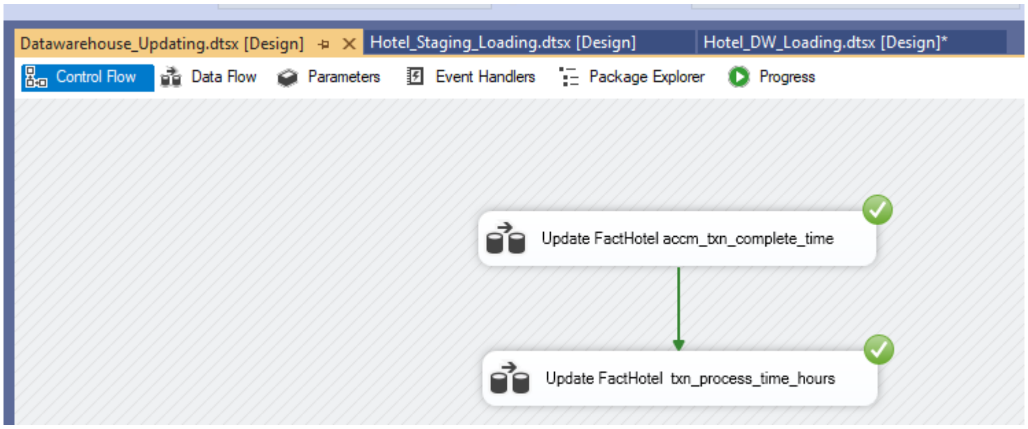
6.1.5 Load FactHotel Data from staging


OLE DB Source
10,000 rows
Dim Review Lookup
Lookup Match Output (10,000 rows)
Dim User Lookup
Lookup Match Output (10,000 rows)
Dim Hotel Category Lookup
Lookup Match Output (10,000 rows)
Dim Location Lookup
Lookup Match Output (10,000 rows)
Data Conversion
10,000 rows
Lookup
Lookup Match Output (10,000 rows)
Derived Column
10,000 rows
OLE DB Destination

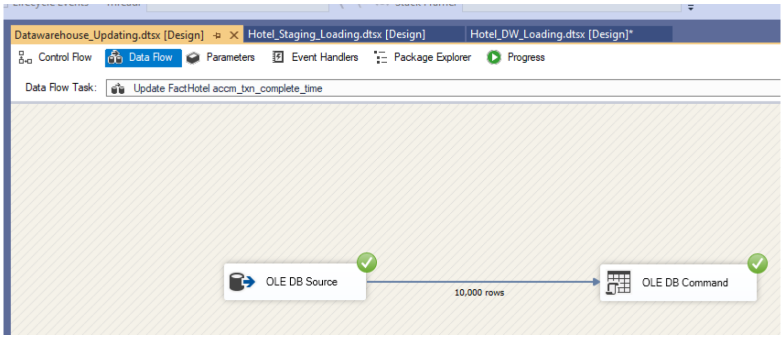
07.Datawarehouse Updating

In order to creating Accumulated fact table I created a new SSIS package and updated accm_txn_complete_time and txn_process_time_hours.

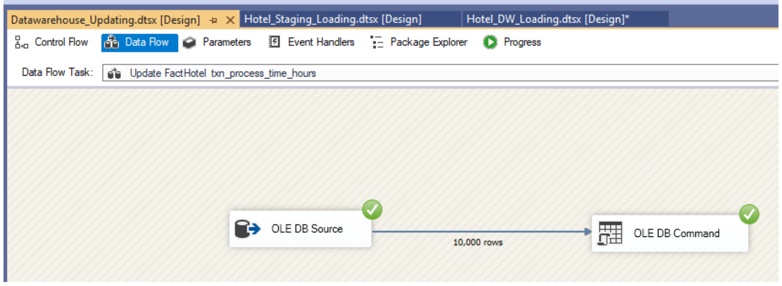
7.1 Datawarehouse updating


Datawarehouse_Updating.dtsx [Design]
Hotel_Staging_Loading.dtsx [Design]
Hotel_DW_Loading.dtsx [Design]*
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Update FactHotel accm_txn_complete_time
Update FactHotel txn_process_time_hours

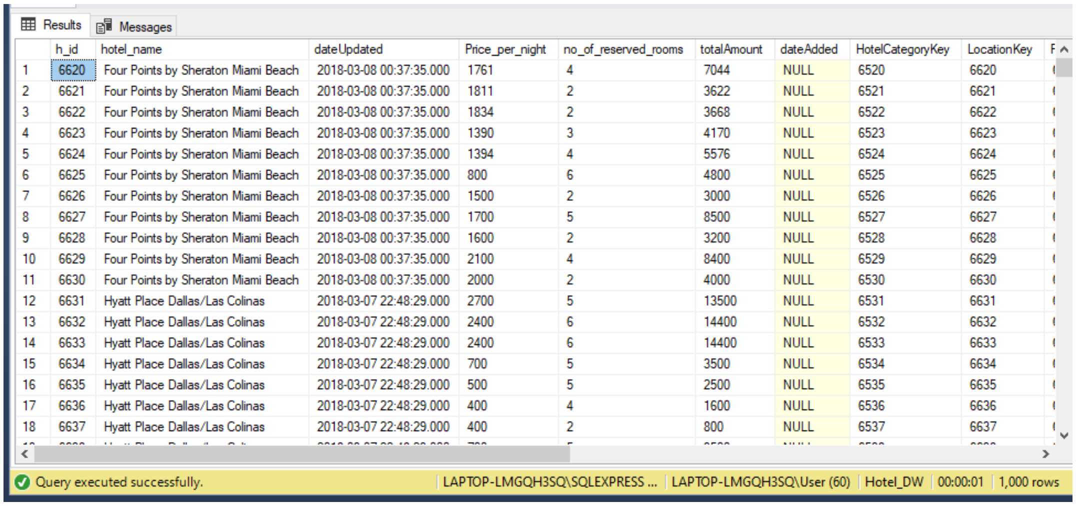
7.1.1 Update FactHotel accm_txn_complete_time


Datawarehouse_Updating.dtsx [Design]
Hotel_Staging_Loading.dtsx [Design]
Hotel_DW_Loading.dtsx [Design]*
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Data Flow Task:
Update FactHotel accm_txn_complete_time
OLE DB Source
10,000 rows
OLE DB Command

7.1.2 Update FactHotel txn_process_time_hours

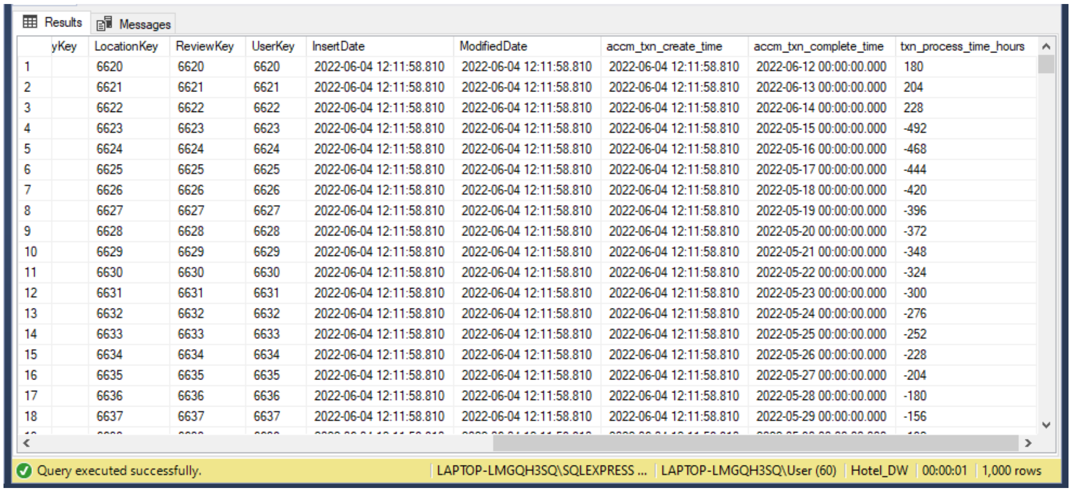

Datawarehouse_Updating.dtsx [Design]
Hotel_Staging_Loading.dtsx [Design]
Hotel_DW_Loading.dtsx [Design]*
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
Data Flow Task:
Update FactHotel txn_process_time_hours
OLE DB Source
10,000 rows
OLE DB Command

7.2 Accumulated Fact Table (FactHotel)



	h_id	hotel_name	dateUpdated	Price_per_night	no_of_reserved_rooms	totalAmount	dateAdded	HotelCategoryKey	LocationKey
1	6620	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	1761	4	7044	NULL	6520	6620
2	6621	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	1811	2	3622	NULL	6521	6621
3	6622	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	1834	2	3668	NULL	6522	6622
4	6623	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	1390	3	4170	NULL	6523	6623
5	6624	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	1394	4	5576	NULL	6524	6624
6	6625	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	800	6	4800	NULL	6525	6625
7	6626	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	1500	2	3000	NULL	6526	6626
8	6627	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	1700	5	8500	NULL	6527	6627
9	6628	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	1600	2	3200	NULL	6528	6628
10	6629	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	2100	4	8400	NULL	6529	6629
11	6630	Four Points by Sheraton Miami Beach	2018-03-08 00:37:35.000	2000	2	4000	NULL	6530	6630
12	6631	Hyatt Place Dallas/Las Colinas	2018-03-07 22:48:29.000	2700	5	13500	NULL	6531	6631
13	6632	Hyatt Place Dallas/Las Colinas	2018-03-07 22:48:29.000	2400	6	14400	NULL	6532	6632
14	6633	Hyatt Place Dallas/Las Colinas	2018-03-07 22:48:29.000	2400	6	14400	NULL	6533	6633
15	6634	Hyatt Place Dallas/Las Colinas	2018-03-07 22:48:29.000	700	5	3500	NULL	6534	6634
16	6635	Hyatt Place Dallas/Las Colinas	2018-03-07 22:48:29.000	500	5	2500	NULL	6535	6635
17	6636	Hyatt Place Dallas/Las Colinas	2018-03-07 22:48:29.000	400	4	1600	NULL	6536	6636
18	6637	Hyatt Place Dallas/Las Colinas	2018-03-07 22:48:29.000	400	2	800	NULL	6537	6637

Query executed successfully. | LAPTOP-LMGQH3SQ\SQLEXPRESS ... | LAPTOP-LMGQH3SQ\User (60) | Hotel_DW | 00:00:01 | 1,000 rows



	yKey	LocationKey	ReviewKey	UserKey	InsertDate	ModifiedDate	accm_txn_create_time	accm_txn_complete_time	txn_process_time_hours
1		6620	6620	6620	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-12 00:00:00.000	180
2		6621	6621	6621	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-13 00:00:00.000	204
3		6622	6622	6622	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-14 00:00:00.000	228
4		6623	6623	6623	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-15 00:00:00.000	-492
5		6624	6624	6624	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-16 00:00:00.000	-468
6		6625	6625	6625	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-17 00:00:00.000	-444
7		6626	6626	6626	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-18 00:00:00.000	-420
8		6627	6627	6627	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-19 00:00:00.000	-396
9		6628	6628	6628	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-20 00:00:00.000	-372
10		6629	6629	6629	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-21 00:00:00.000	-348
11		6630	6630	6630	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-22 00:00:00.000	-324
12		6631	6631	6631	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-23 00:00:00.000	-300
13		6632	6632	6632	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-24 00:00:00.000	-276
14		6633	6633	6633	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-25 00:00:00.000	-252
15		6634	6634	6634	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-26 00:00:00.000	-228
16		6635	6635	6635	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-27 00:00:00.000	-204
17		6636	6636	6636	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-28 00:00:00.000	-180
18		6637	6637	6637	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-06-04 12:11:58.810	2022-05-29 00:00:00.000	-156

```
Query executed successfully.    LAPTOP-LMGQH3SQ\SQLEXPRESS ... | LAPTOP-LMGQH3SQ\User (60) | Hotel_DW | 00:00:01 | 1,000 rows
```