

Sri Lanka Institute of Information Technology



Data Warehousing & Business Intelligence

Assignment 01

Submitted by:

Gunathilaka U.K.R.S.W(it20188818)

Contents

1.Data Set Selection.....	3
Background	3
Content	3
ER Diagram.....	4
2. Preparation of Data Sources.	5
3. Solution Architecture.....	6
4. Data warehouse design & development.....	8
5. ETL Development.....	9
5.1 ETL –Source To Staging	9
5.1.1 Load data User to staging	9
5.1.2 Load data Review to staging	10
5.1.3 Load data Hotel to staging	10
5.1.4 Load data Hotel Category to staging.....	11
5.1.5 Load data location to staging(.txt file)	11
5.1.6 Load data accm_txn_complete_time to staging.....	12
6. Staging To DW.	13
6.1 ETL System to Datawarehouse.....	13
6.1.1. Transfer and Load DimUser Data from staging.....	13
6.1.2. Transfer and Load DimReview Data from staging	14
6.1.3. Transfer and Load DimHotelCategory Data from staging	14
6.1.4. Transfer and Load DimLocation Data from staging (Slowly changing diamention).....	15
6.1.5. Load FactHotel Data from staging	16
7. Datawarehouse Updating	17
7.1. Datawarehouse updating.....	17
7.1.1 Update factHotel accm_txn_complete_time	18
7.1.2 Update factHotel txn_process_time_hours.....	18
7.2 Accumulated Fact Table (FactHotel)	19

1.Data Set Selection.

Background

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.:

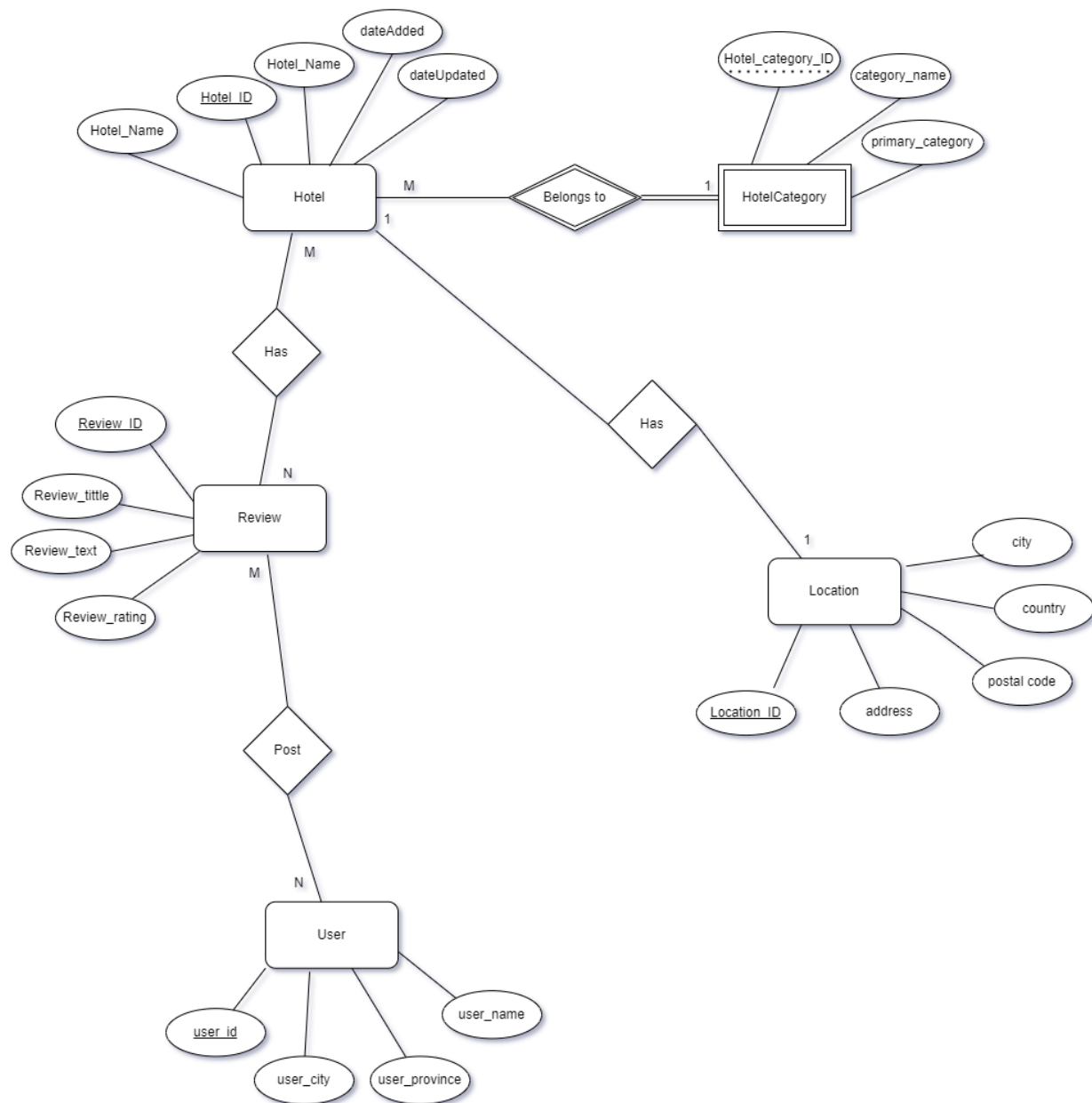
- What are the bottom and top states for hotel reviews by average rating?
- What is the correlation between a state's population and their number of hotel reviews?
- What is the correlation between a state's tourism budget and their number of hotel reviews?

Content

The data set was downloaded from -:

[https://www.kaggle.com/datasets/datafiniti/hotel-reviews?select=Datafiniti Hotel Reviews.csv](https://www.kaggle.com/datasets/datafiniti/hotel-reviews?select=Datafiniti+Hotel+Reviews.csv)

ER Diagram



2. Preparation of Data Sources.

In order to data extraction need to prepare the data sources. From my main data source, I have extracted to type of data sources.

1. Text file (.txt)

- Location text file

2. CSV files (.csv)

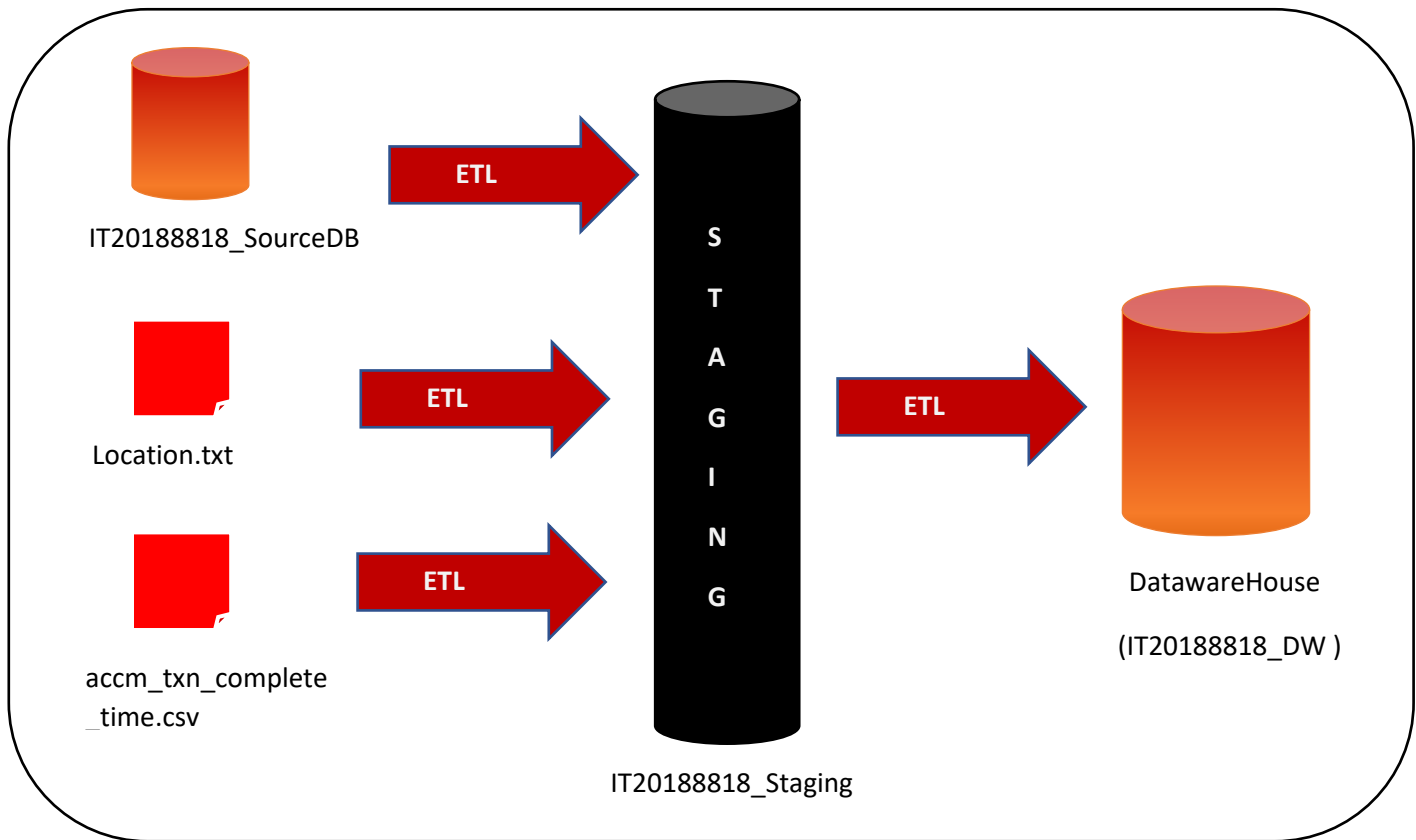
- Hotel CSV File.
- Hotel Category CSV File.
- User CSV File.
- Review CSV File.
- accm_txn_complete_time CSV

From these dataset, I import above csv files to SQL SERVER MANAGEMENT STUDIO and create the database called **IT20188818_sourceDB** Database.

IT20188818_sourceDB have following tables -:

- accm_txn_complete_time table.
- Hoteltbl Table.
- HotelCategorytbl Table.
- Usertbl Table.
- Reviewtbl Table.

Text file –This text file include all the hotel address details including address, Postal code ,city, province and the country.



3. Solution Architecture.

3.1 IT20188818 StagingDB.

- accm_txn_complete_time
- stgHotel
- stgHotelCategory
- stgReview
- StgUser
- stgLocation

3.2 IT20188818 DW

- DimDate.
- DimHotelCategory.
- DimLocation.
- DimReview.

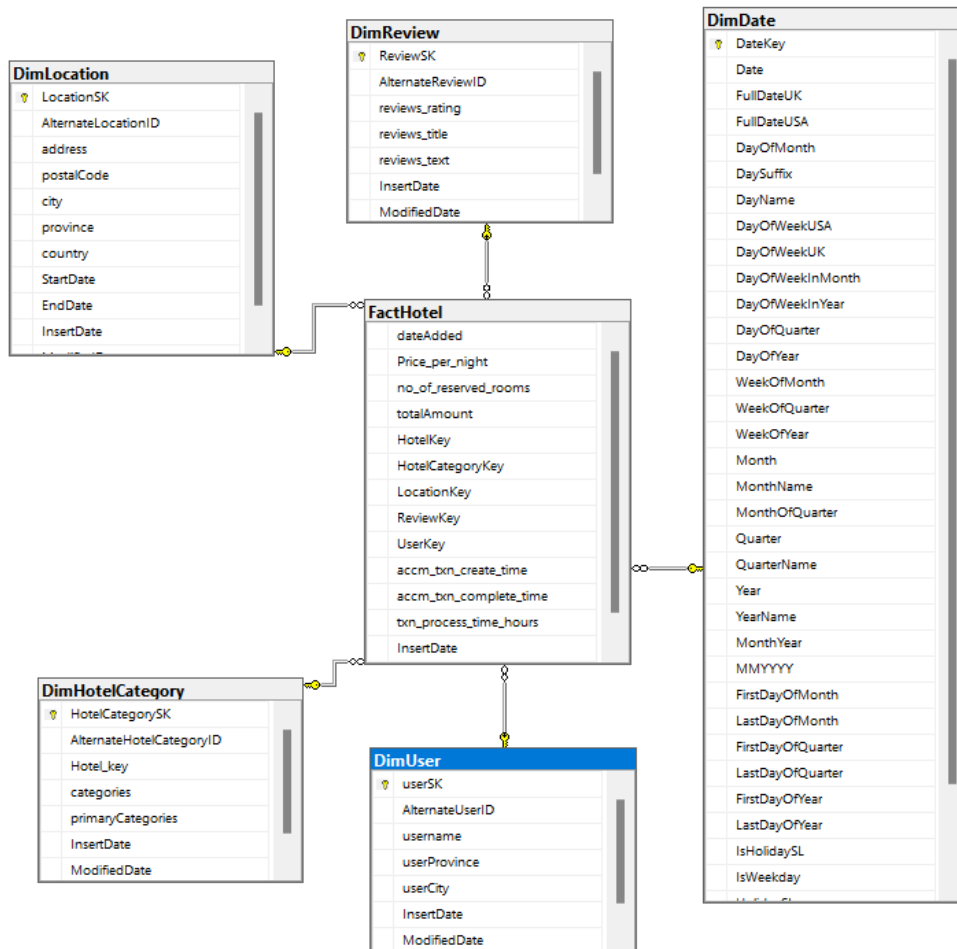
- DimUser.
- FactHotel.

3.3 Architure 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.

4. Data warehouse design & development.

4.1 Relational Diagram – Star Schema.

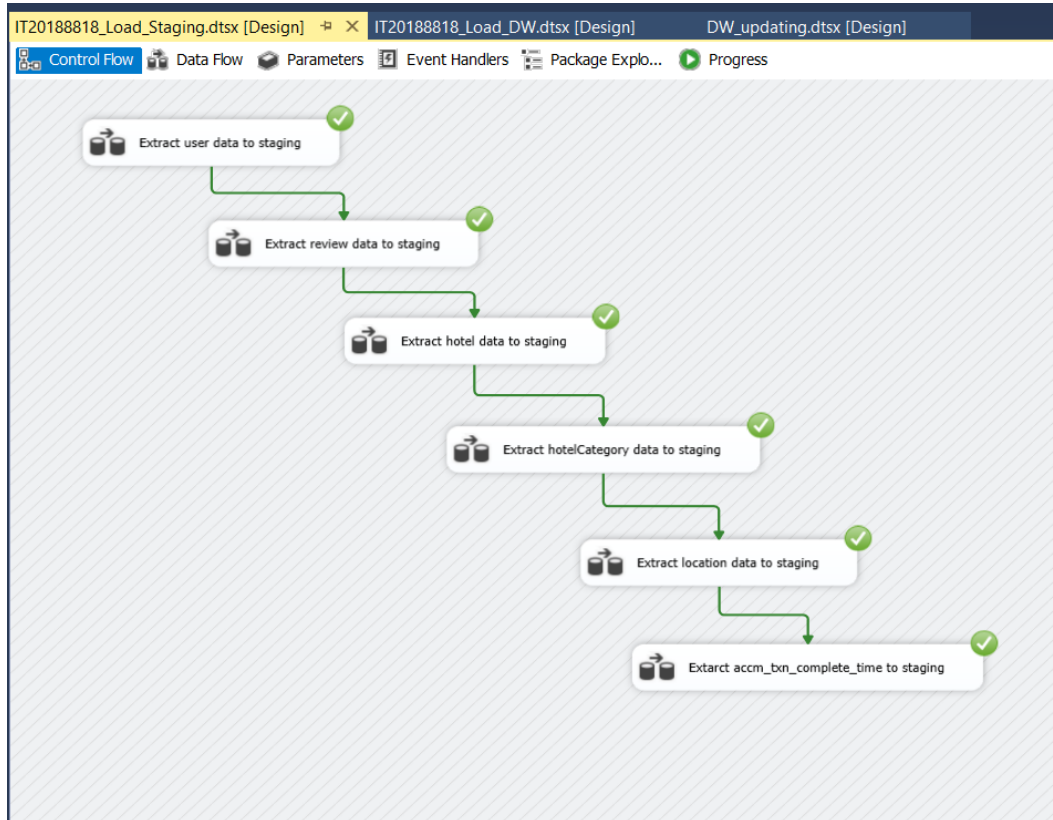


DimLocation is slowly changing dimension. Address and city may be changed in future. Therefore, I get it as slowly changing attribute.

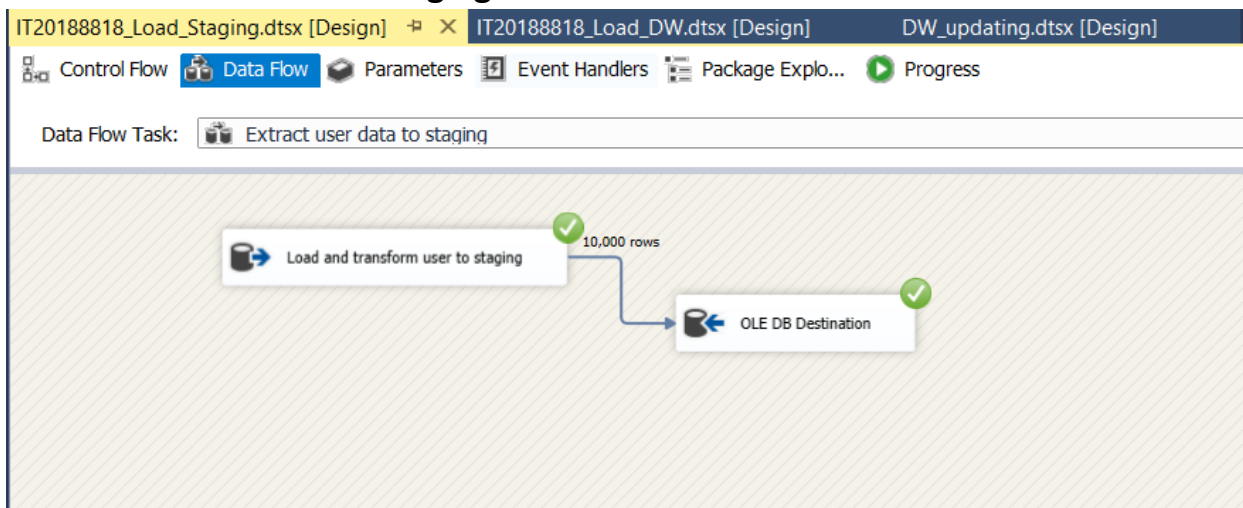
Address -> PostalCode -> City -> Province -> Country This is the Hierachies (Location table.)

5. ETL Development.

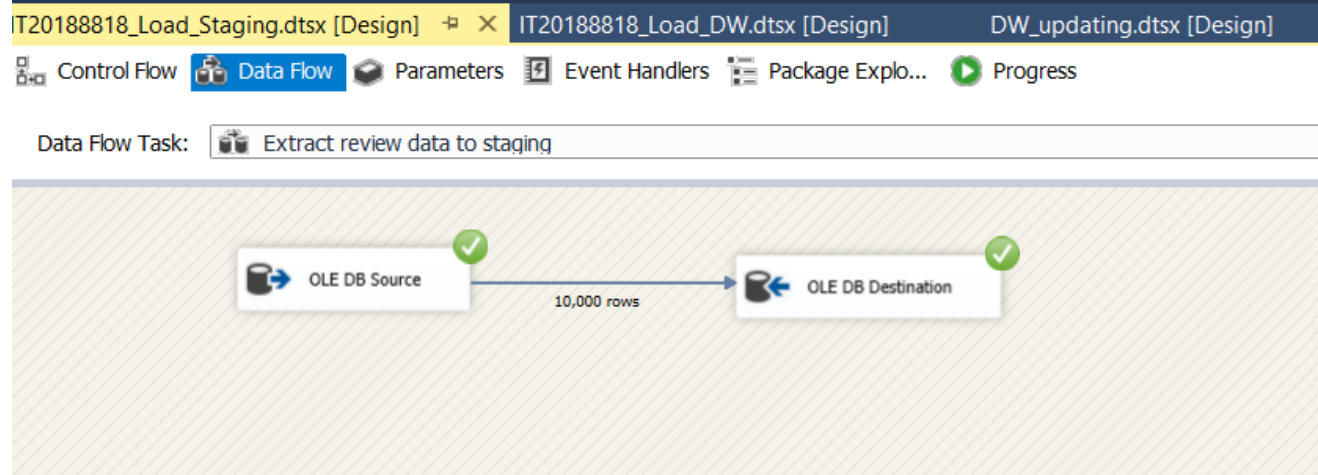
5.1 ETL –Source To Staging



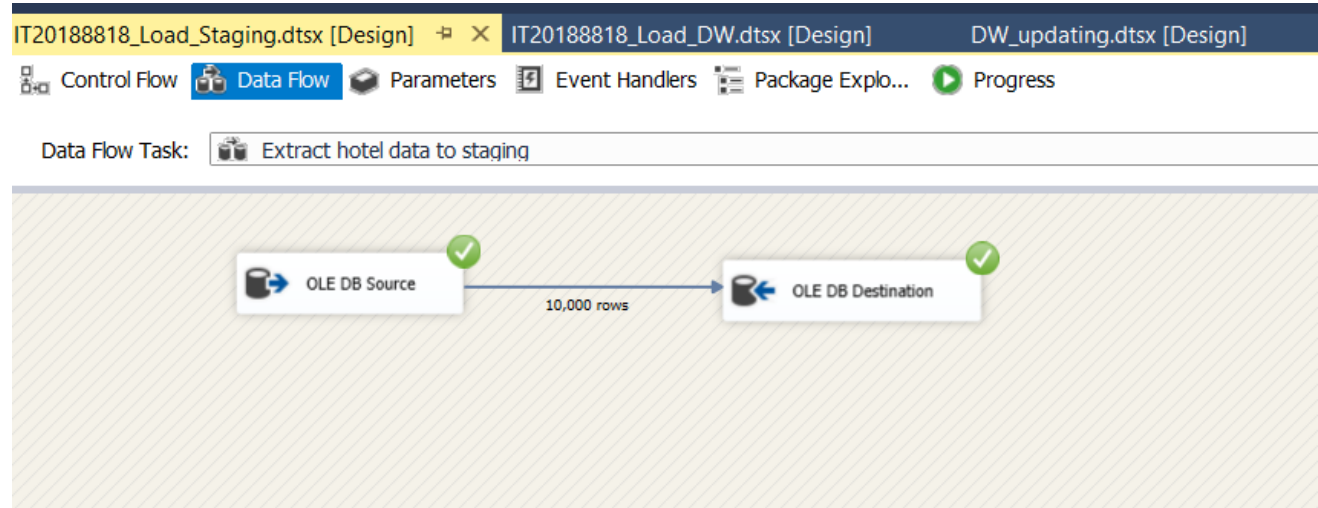
5.1.1 Load data User to staging



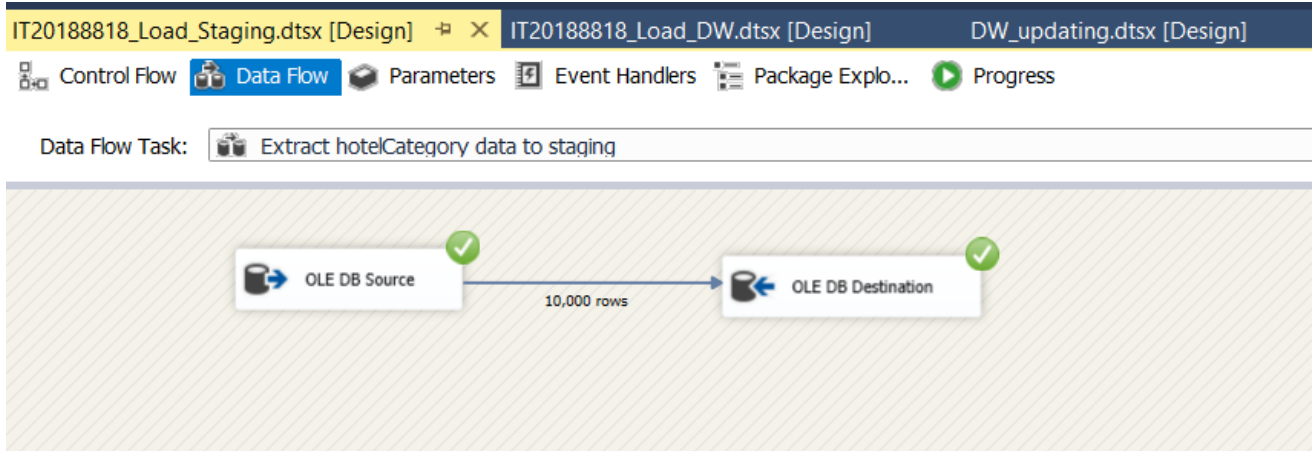
5.1.2 Load data Review to staging



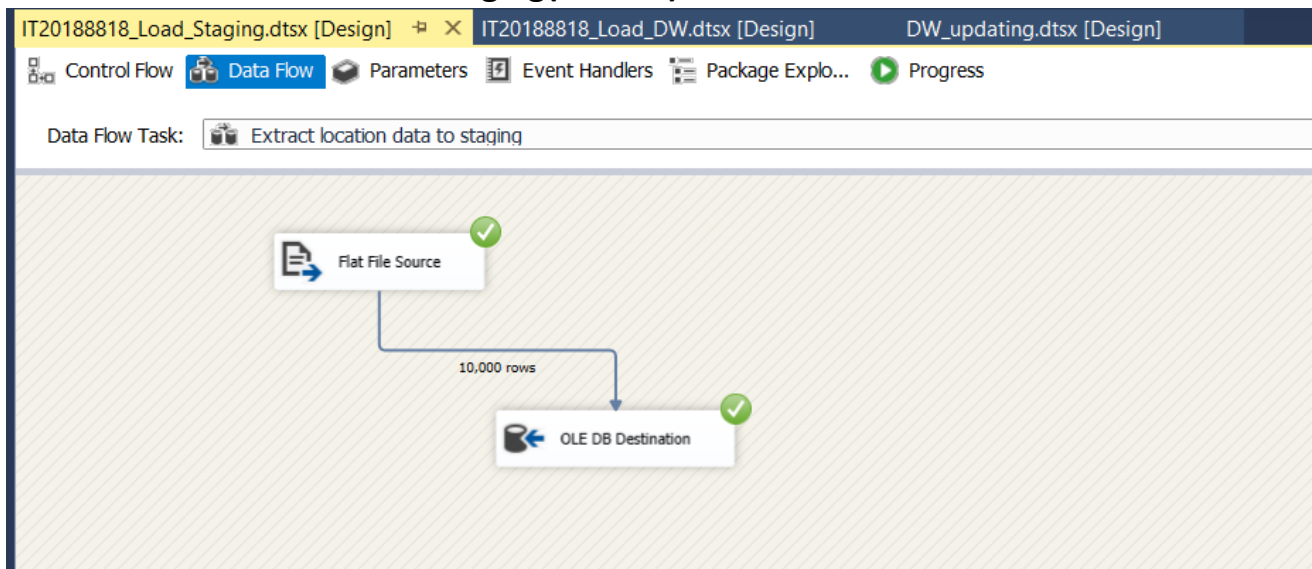
5.1.3 Load data Hotel to staging



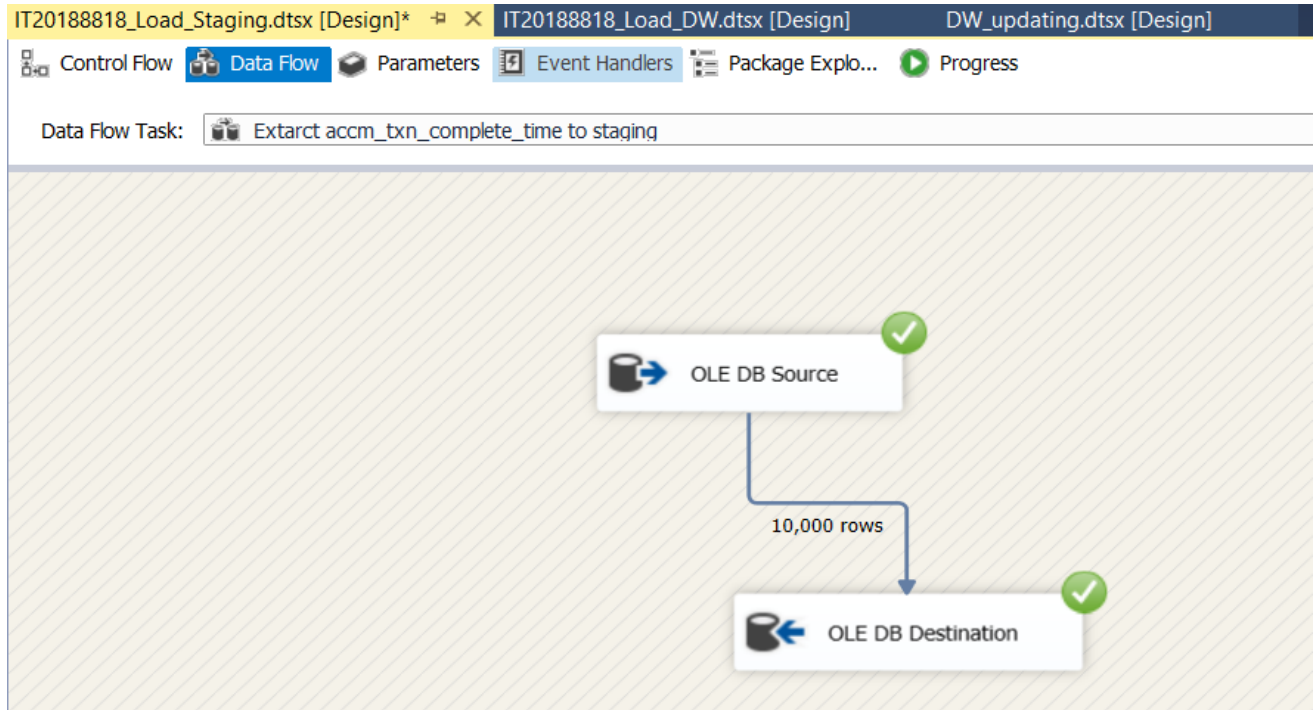
5.1.4 Load data Hotel Category to staging



5.1.5 Load data location to staging(.txt file)

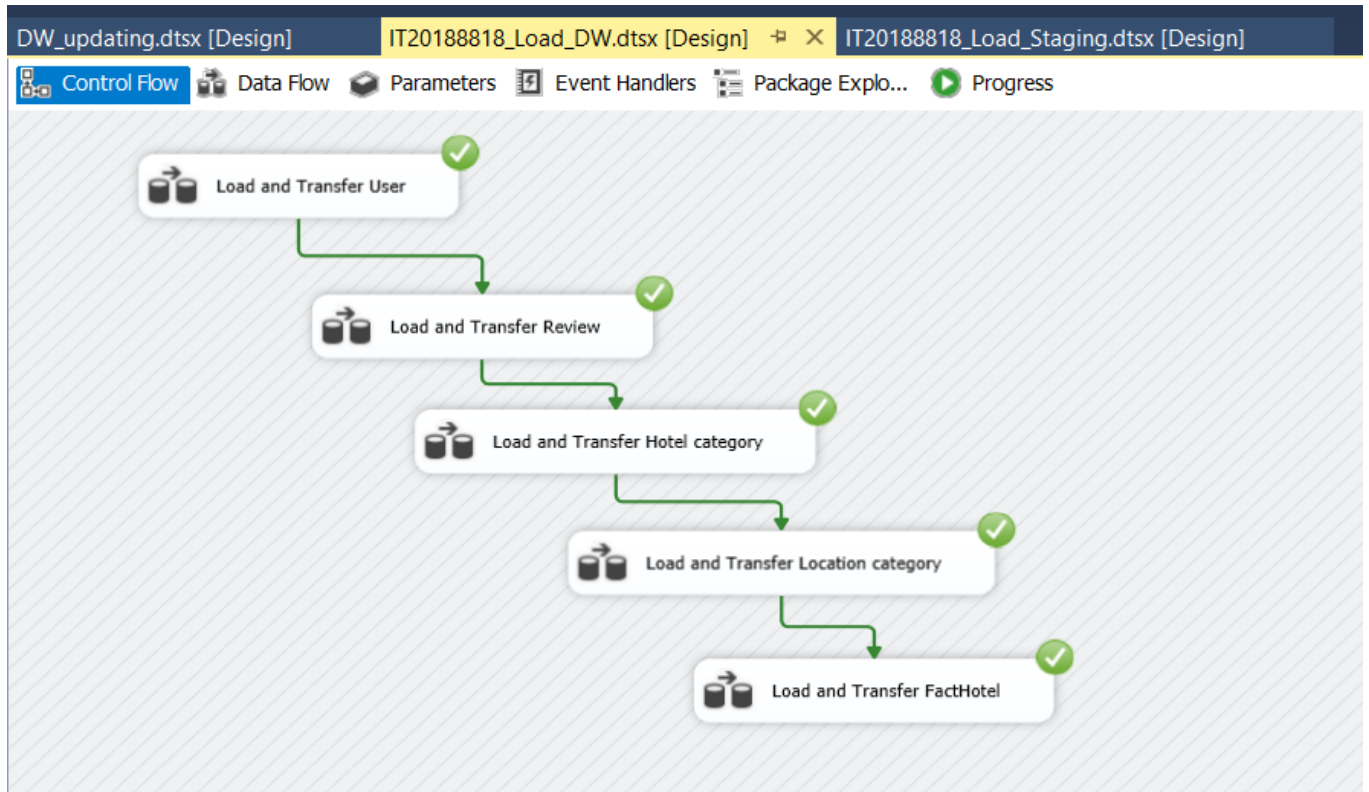


5.1.6 Load data accm_txn_complete_time to staging

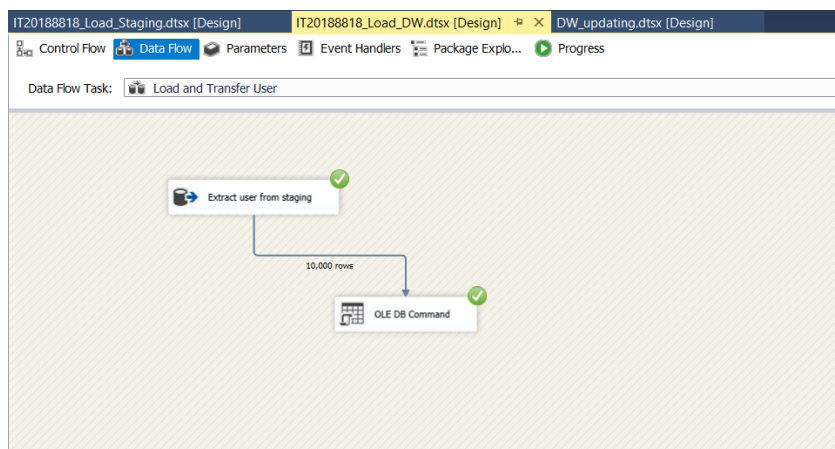


6. Staging To DW.

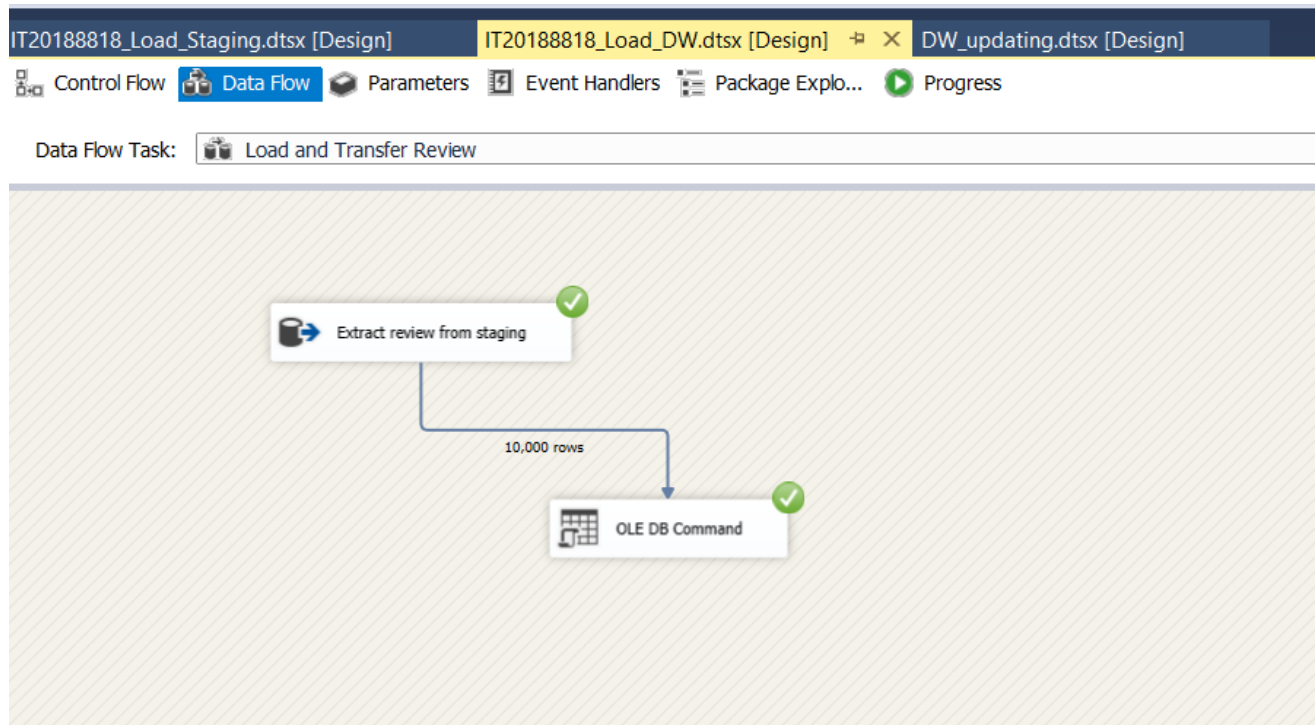
6.1 ETL System to Datawarehouse



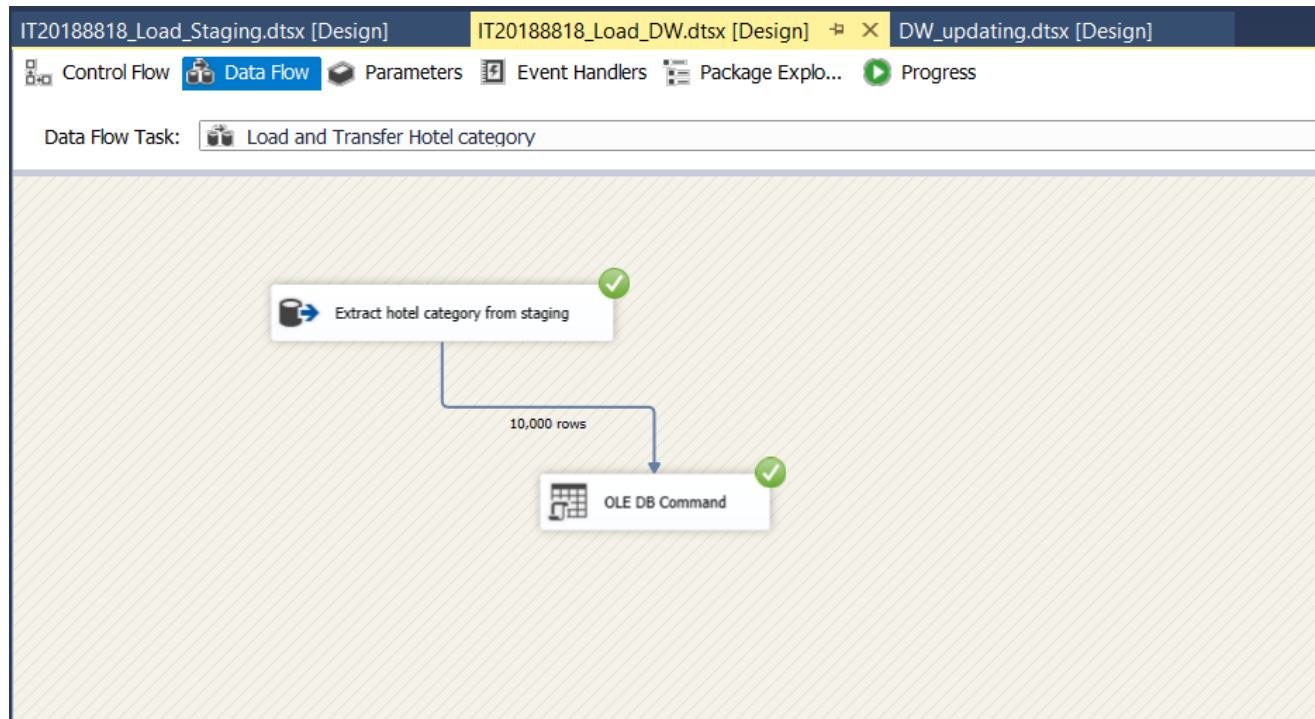
6.1.1. Transfer and Load DimUser Data from staging



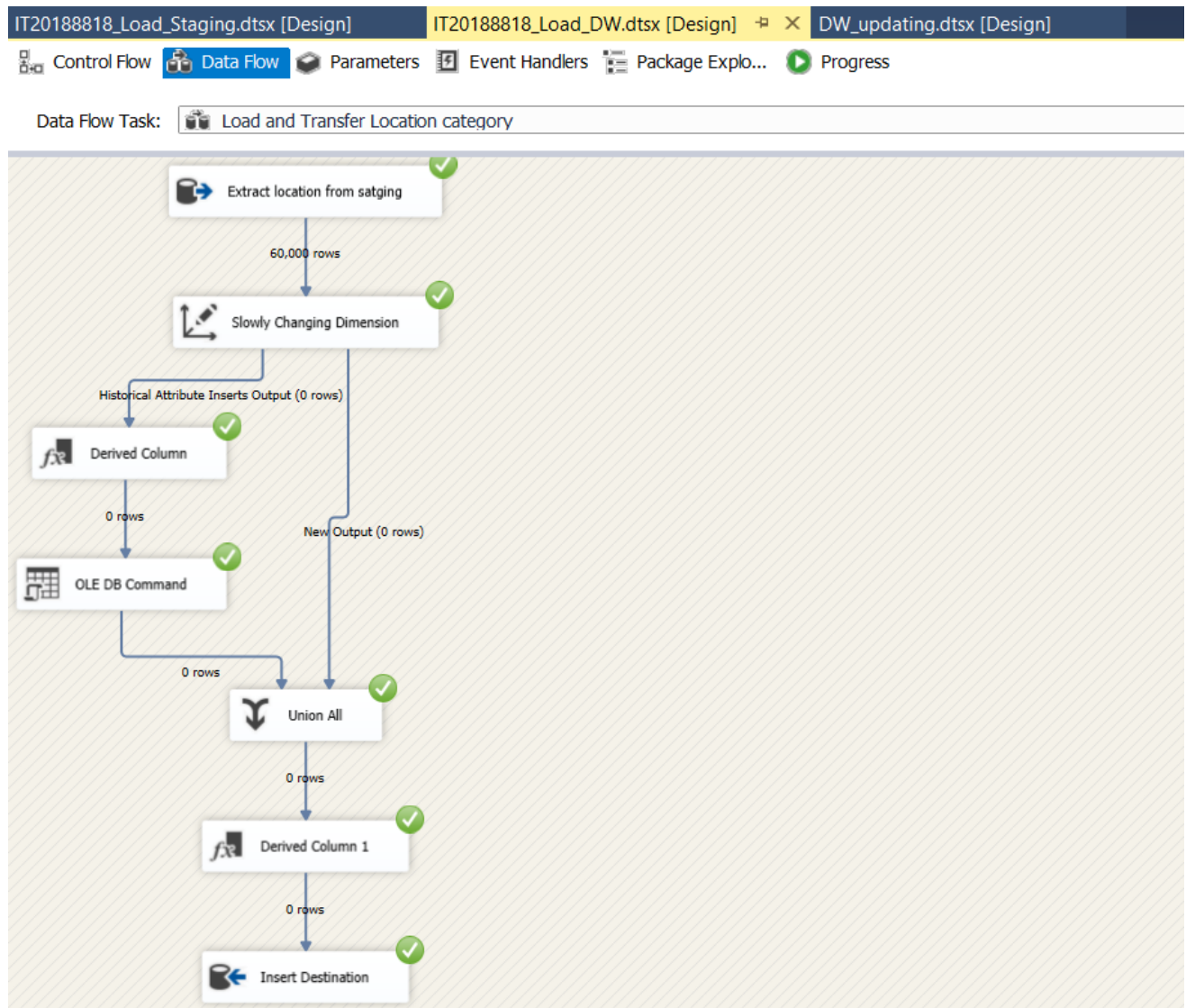
6.1.2. Transfer and Load DimReview Data from staging



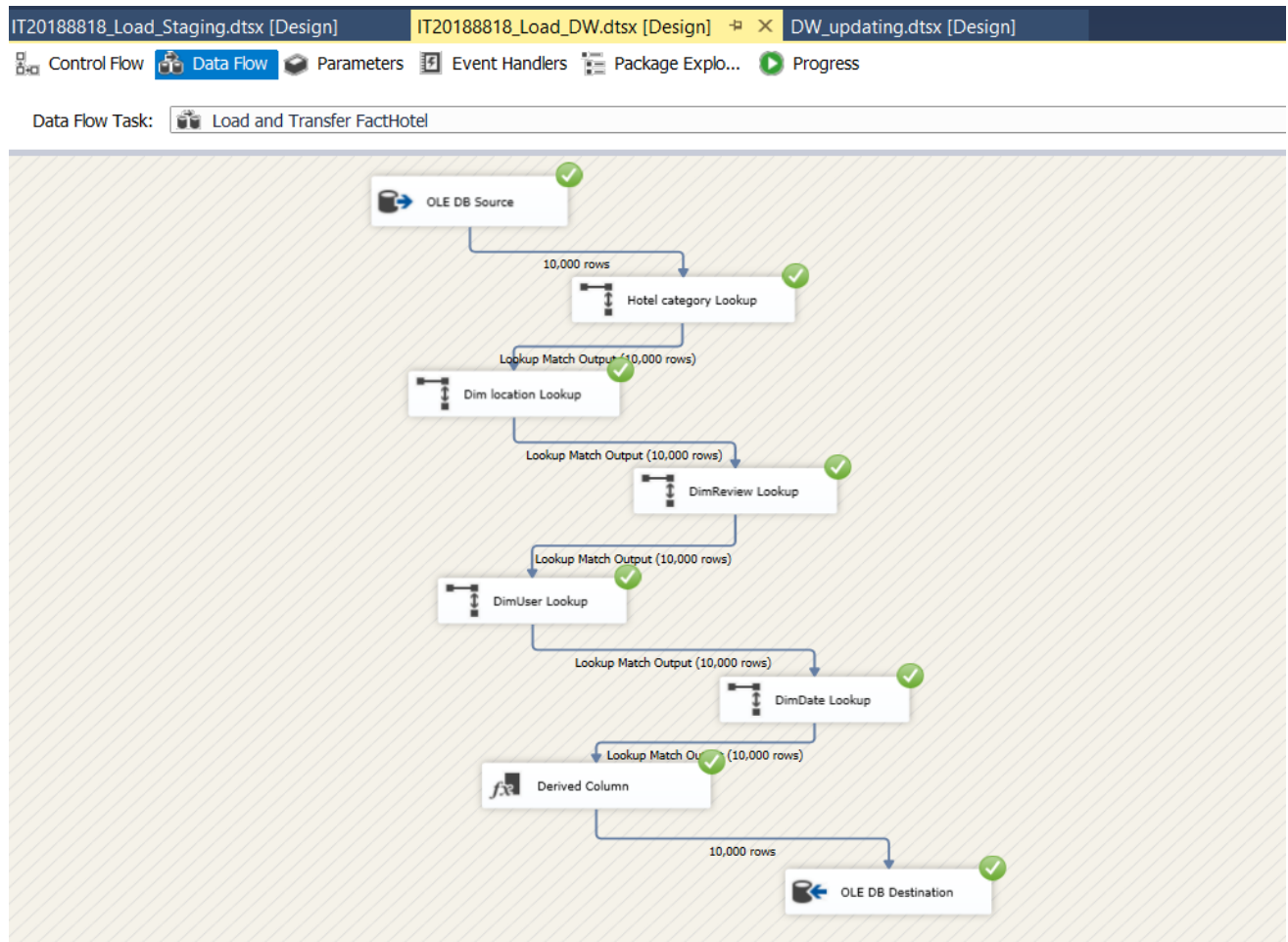
6.1.3. Transfer and Load DimHotelCategory Data from staging



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



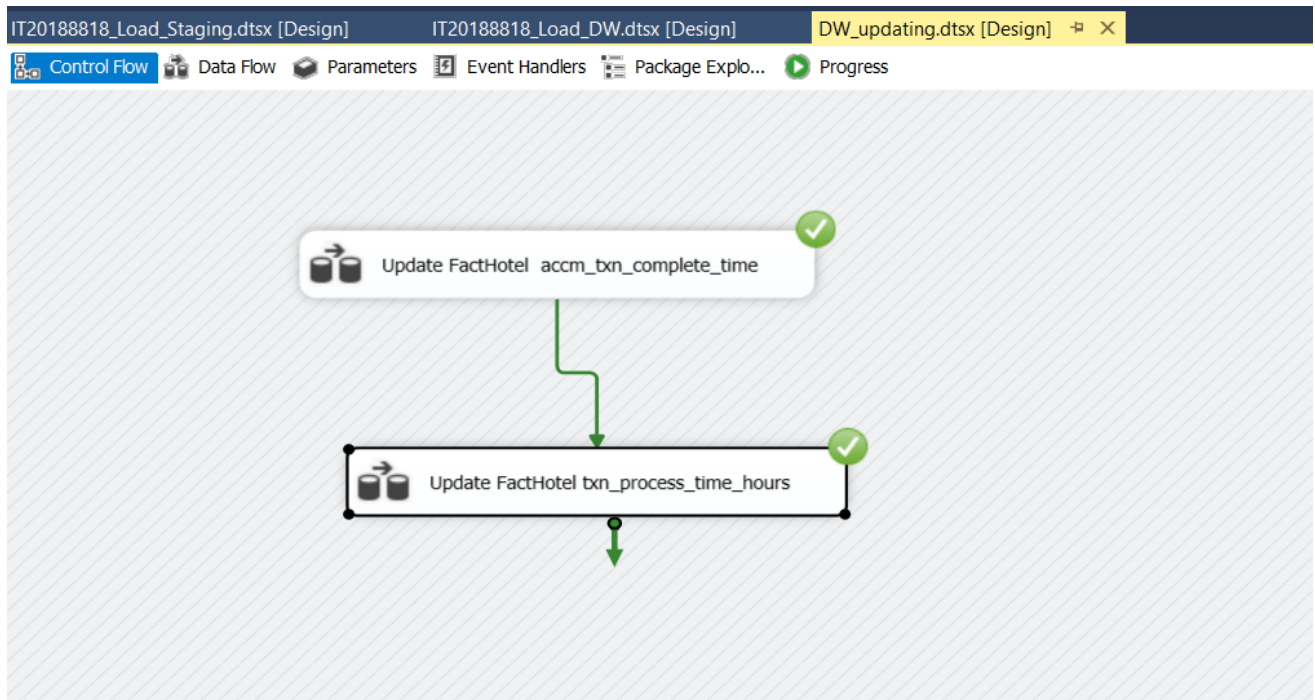
6.1.5. Load FactHotel Data from staging



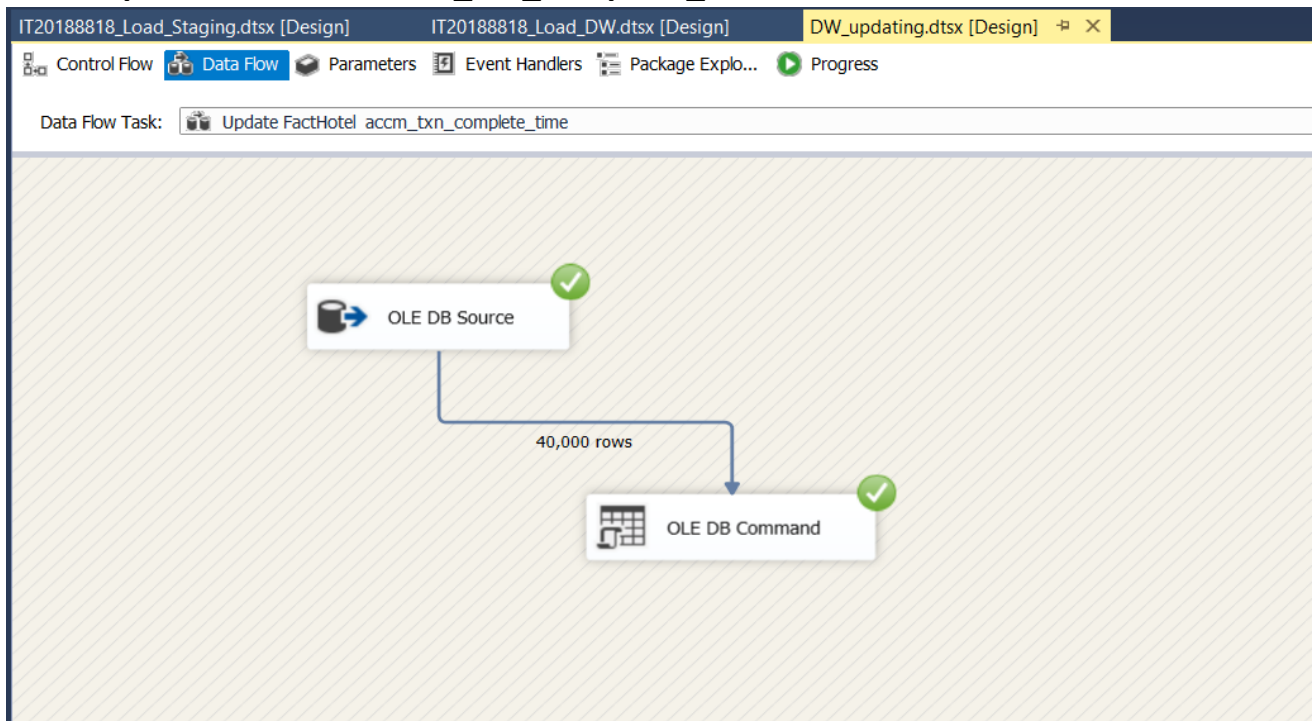
7. Datawarehouse Updating

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

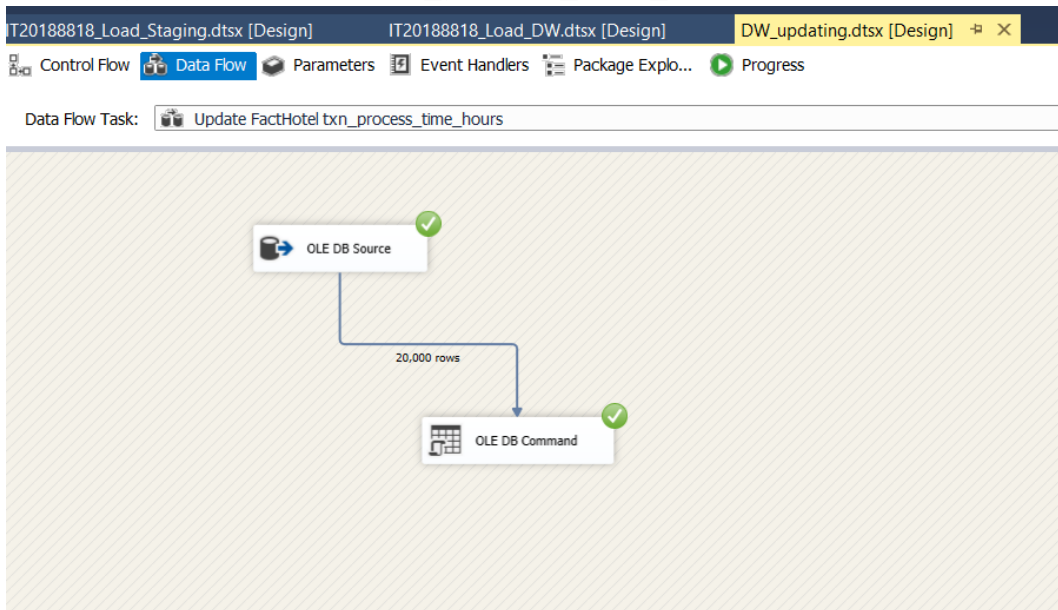
7.1. Datawarehouse updating



7.1.1 Update factHotel accm_txn_complete_time



7.1.2 Update factHotel txn_process_time_hours



7.2 Accumulated Fact Table (FactHotel)

90 %

Results Messages

	h_id	hotel_name	dateUpdated	Price_per_night	no_of_reserved_rooms	totalAmount	dateAdded	HotelCategoryKey	LocationKey	ReviewKey	UserKey	InsertDate	Modified
1	1	Rancho Valencia Resort Spa	2018-09-11 02:36:27.000	50	3	150	NULL	1	1	1	1	2022-05-14 17:30:38.157	2022-05
2	2	Rancho Valencia Resort Spa	2018-09-11 02:36:27.000	50	2	100	NULL	2	2	2	2	2022-05-14 17:30:38.157	2022-05
3	3	Rancho Valencia Resort Spa	2018-09-11 02:36:27.000	50	2	100	NULL	3	3	3	3	2022-05-14 17:30:38.157	2022-05
4	4	Aloft Arundel Mills	2018-09-11 02:36:16.000	50	2	100	NULL	4	4	4	4	2022-05-14 17:30:38.157	2022-05
5	5	Aloft Arundel Mills	2018-09-11 02:36:16.000	50	2	100	NULL	5	5	5	5	2022-05-14 17:30:38.157	2022-05
6	6	Aloft Arundel Mills	2018-09-11 02:36:16.000	50	2	100	NULL	6	6	6	6	2022-05-14 17:30:38.157	2022-05
7	7	Aloft Arundel Mills	2018-09-11 02:36:16.000	50	1	50	NULL	7	7	7	7	2022-05-14 17:30:38.157	2022-05
8	8	Aloft Arundel Mills	2018-09-11 02:36:16.000	50	4	200	NULL	8	8	8	8	2022-05-14 17:30:38.157	2022-05
9	9	Aloft Arundel Mills	2018-09-11 02:36:16.000	50	2	100	NULL	9	9	9	9	2022-05-14 17:30:38.157	2022-05
10	10	Hampton Inn Suites PortlandVancouver	2018-09-11 02:36:09.000	50	2	100	NULL	10	10	10	10	2022-05-14 17:30:38.157	2022-05
11	11	Hampton Inn Suites PortlandVancouver	2018-09-11 02:36:09.000	50	2	100	NULL	11	11	11	11	2022-05-14 17:30:38.157	2022-05
12	12	Hampton Inn Suites PortlandVancouver	2018-09-11 02:36:09.000	50	4	200	NULL	12	12	12	12	2022-05-14 17:30:38.157	2022-05
13	13	Hampton Inn Suites PortlandVancouver	2018-09-11 02:36:09.000	50	2	100	NULL	13	13	13	13	2022-05-14 17:30:38.157	2022-05
14	14	Hampton Inn Suites PortlandVancouver	2018-09-11 02:36:09.000	50	3	150	NULL	14	14	14	14	2022-05-14 17:30:38.157	2022-05
15	15	Hampton Inn Suites PortlandVancouver	2018-09-11 02:36:09.000	50	4	200	NULL	15	15	15	15	2022-05-14 17:30:38.157	2022-05
16	16	Hotel Phillips	2018-09-11 02:35:27.000	50	3	150	NULL	16	16	16	16	2022-05-14 17:30:38.157	2022-05

Query executed successfully.

LAPTOP-5CADPCEG (15.0 RTM) | LAPTOP-5CADPCEG\sudara... | IT20188818_DW | 00:00:03 | 1,000 rows

ModifiedDate	accm_txn_create_time	accm_txn_complete_time	txn_process_time_hours
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-14 00:00:00.000	-17
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-15 00:00:00.000	7
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-16 00:00:00.000	31
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-17 00:00:00.000	55
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-18 00:00:00.000	79
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-19 00:00:00.000	103
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-20 00:00:00.000	127
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-21 00:00:00.000	151
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-22 00:00:00.000	175
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-23 00:00:00.000	199
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-24 00:00:00.000	223
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-25 00:00:00.000	247
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-26 00:00:00.000	271
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-27 00:00:00.000	295
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-28 00:00:00.000	319
2022-05-14 17:30:38.157	2022-05-14 17:30:38.157	2022-05-29 00:00:00.000	343

CADPCEG (15.0 RTM) | LAPTOP-5CADPCEG\sudara... | IT20188818_DW | 00:00:03 | 1,000 rows

