Sri Lanka Institute of Information Technology

Data Warehousing & Business Intelligence

Assignment 01

by:

Samarakoon S.M.D.H.

Contents

1.Data Set Selection.

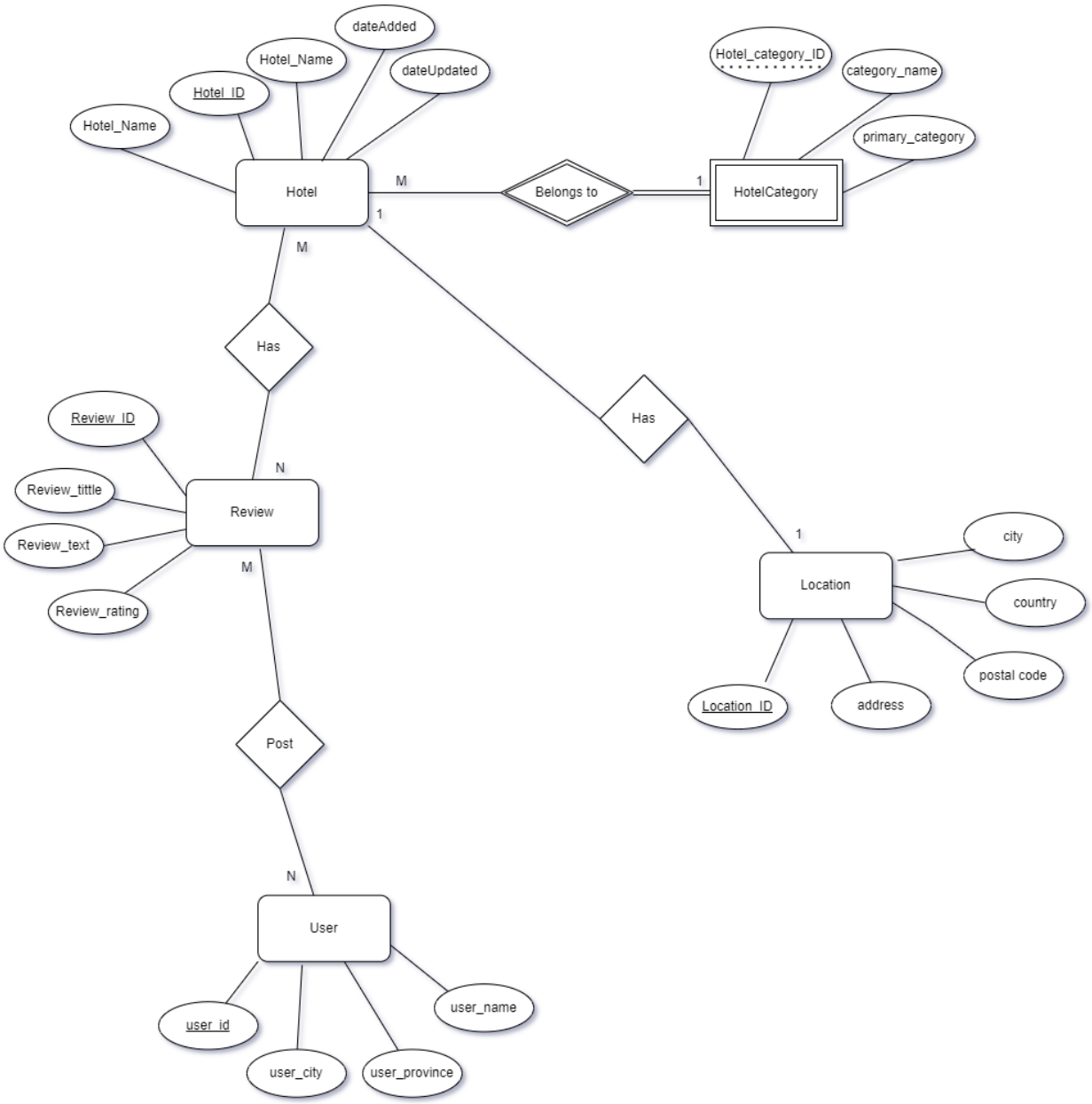
Background

This is a data set which gathered details about co-operate travels focusing flights and hotels selected. Link to the mentioned data set is given below-:

https://www.kaggle.com/datasets/datafiniti/hotel-reviews?select=Datafiniti_Hotel_Reviews.csv

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.

ER Diagram


dateAdded
Hotel_Name
dateUpdated
Hotel_ID
Hotel_Name
Hotel
M
Belongs to
1
HotelCategory
Hotel_category_ID
category_name
primary_category
1
M
Has
Has
Review_ID
N
Review_tittle
Review
Review_text
M
Review_rating
1
city
Location
country
postal code
Location_ID
address
Post
N
User
user_id
user_name
user_city
user_province

2. Preparation of Data Sources.

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)

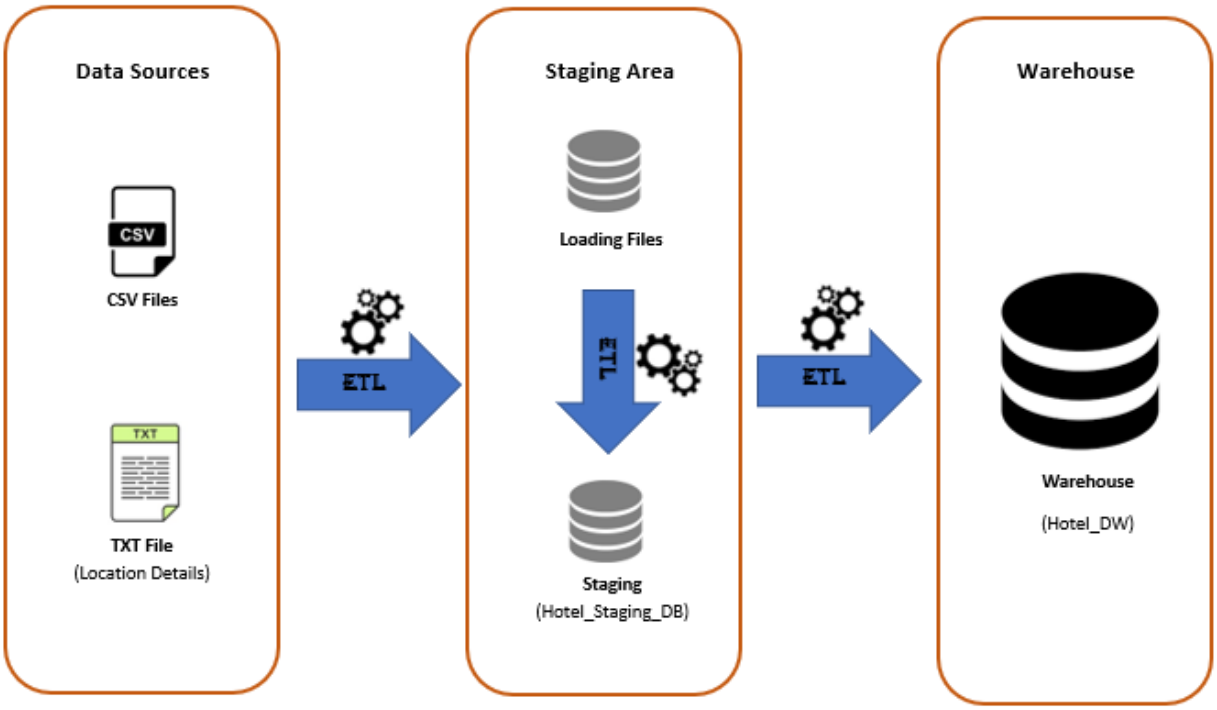
- Hotel CSV File.
- Hotel Category CSV File.
- User CSV File.
- Review CSV File.
- accm_txn_complete_time CSV

Hotel_SourceDB have following tables -:

- accm_txn_complete_time table.
- Hoteltable Table.
- HotelCategory Table.
- Usertable Table.
- Reviewtable 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 Hotel_StagingDB.

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

3.Hotel_DW

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

3.3 Architure Components.

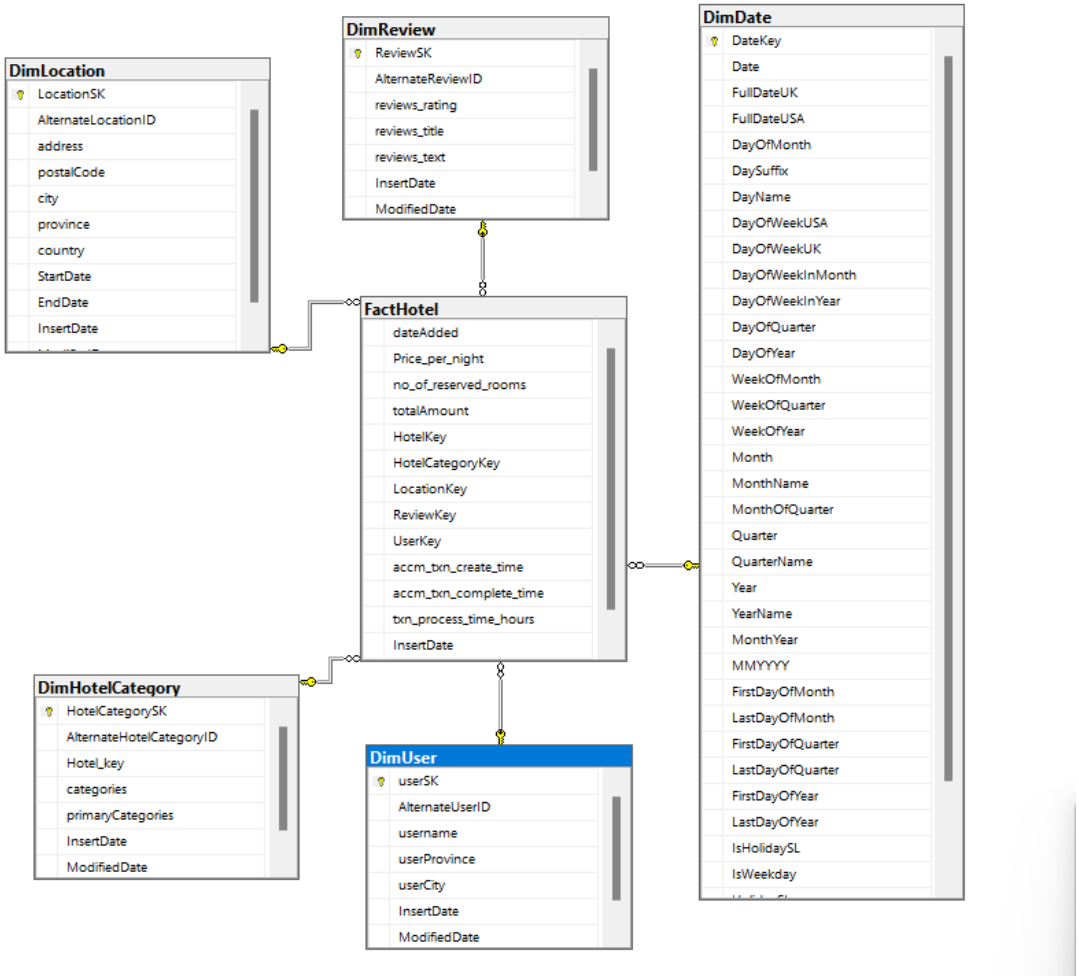
- Data Sources.
  Operational System(**Accumulating**).
  External Sources.

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

- Data Warehouse
  EDW and Data Mart.
  Dimensional Modeling- Facts and Dimensions.
  schema – Star schema.

4. Data warehouse design & development.

4.1 Relational Diagram – Star Schema.

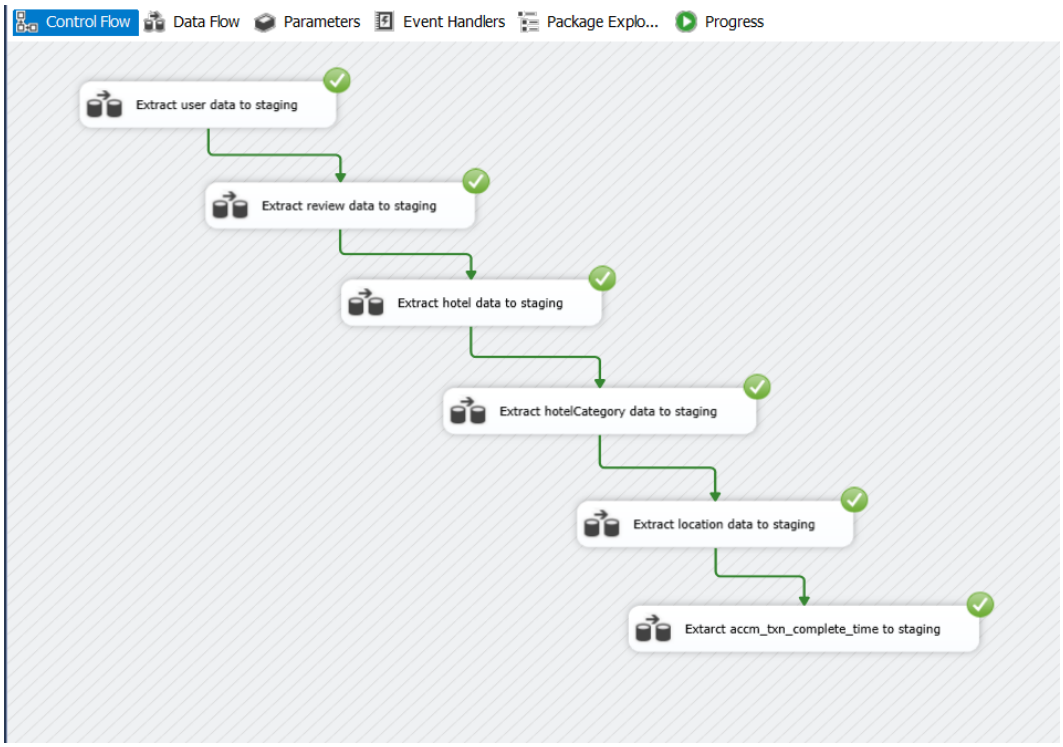


DimLocation is **slowly changing dimention**. Address and city may be changed in future. Therefore, I get it as slowly changing attribute.

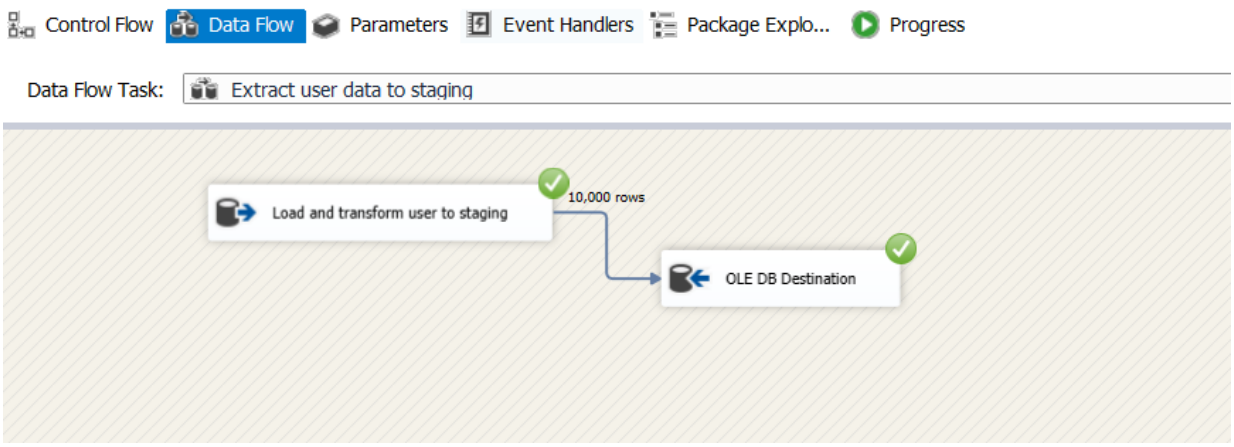
Address -> PostalCode -> City -> Province ->Country This is the Hierarchies

5. ETL Development.

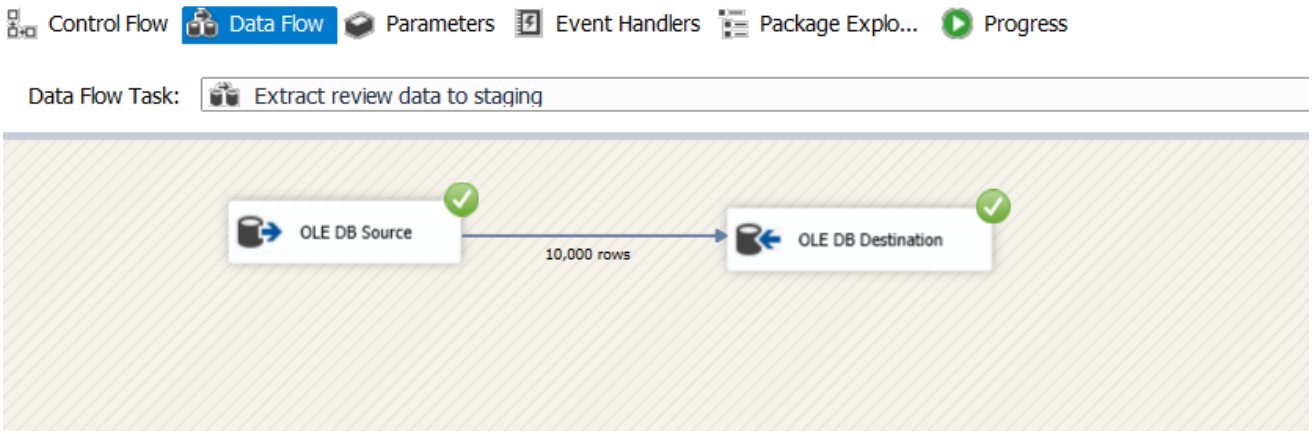
5.1 ETL –Source To Staging



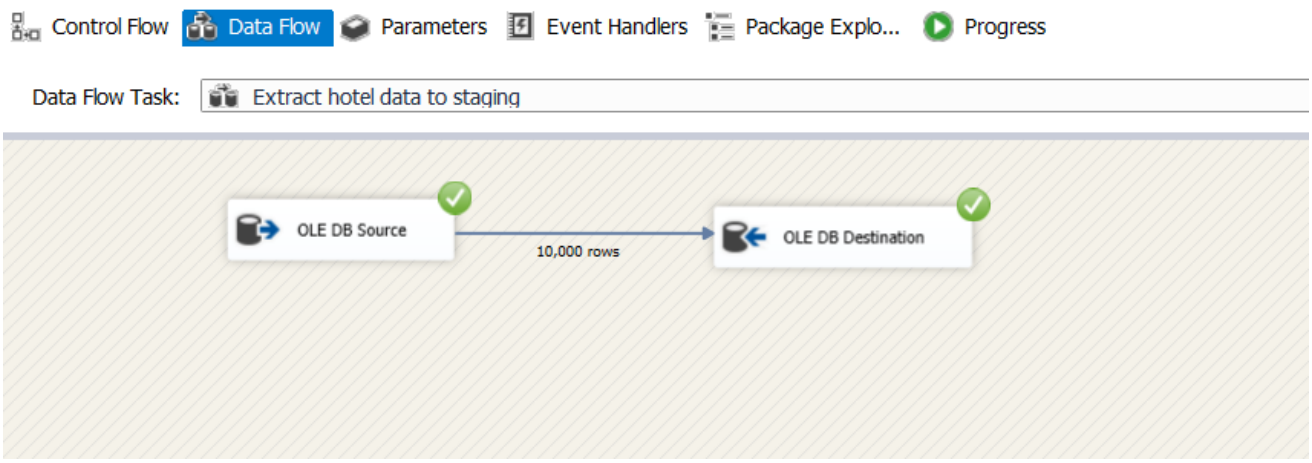
5.1.1 Load data User to staging



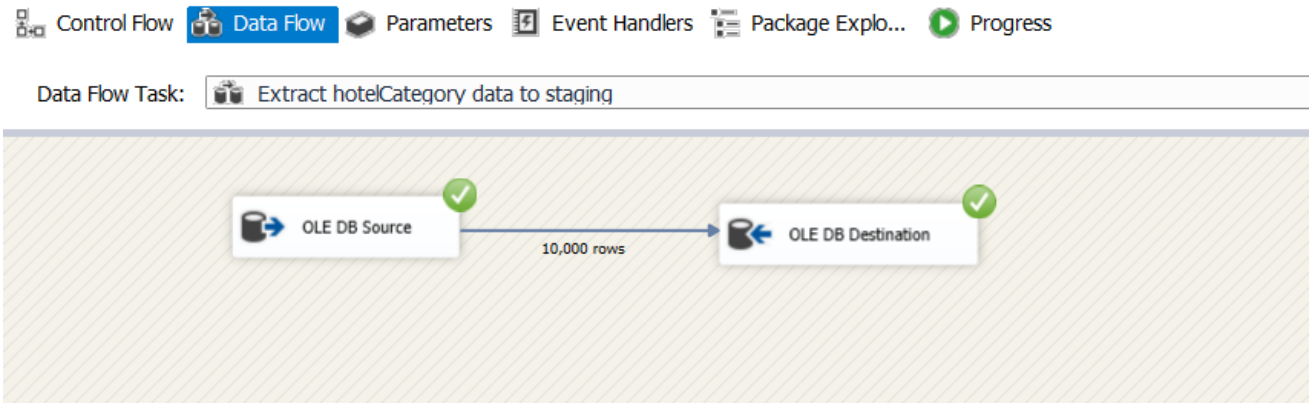
5.1.2 Load data Review to staging



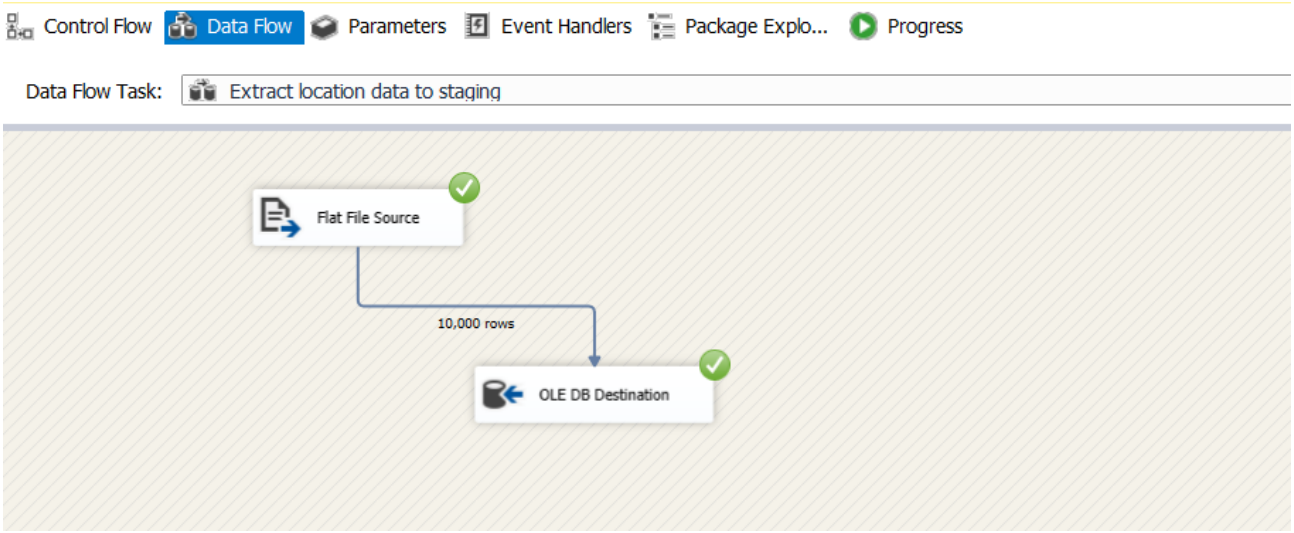
5.1.3 Load data Hotel to staging



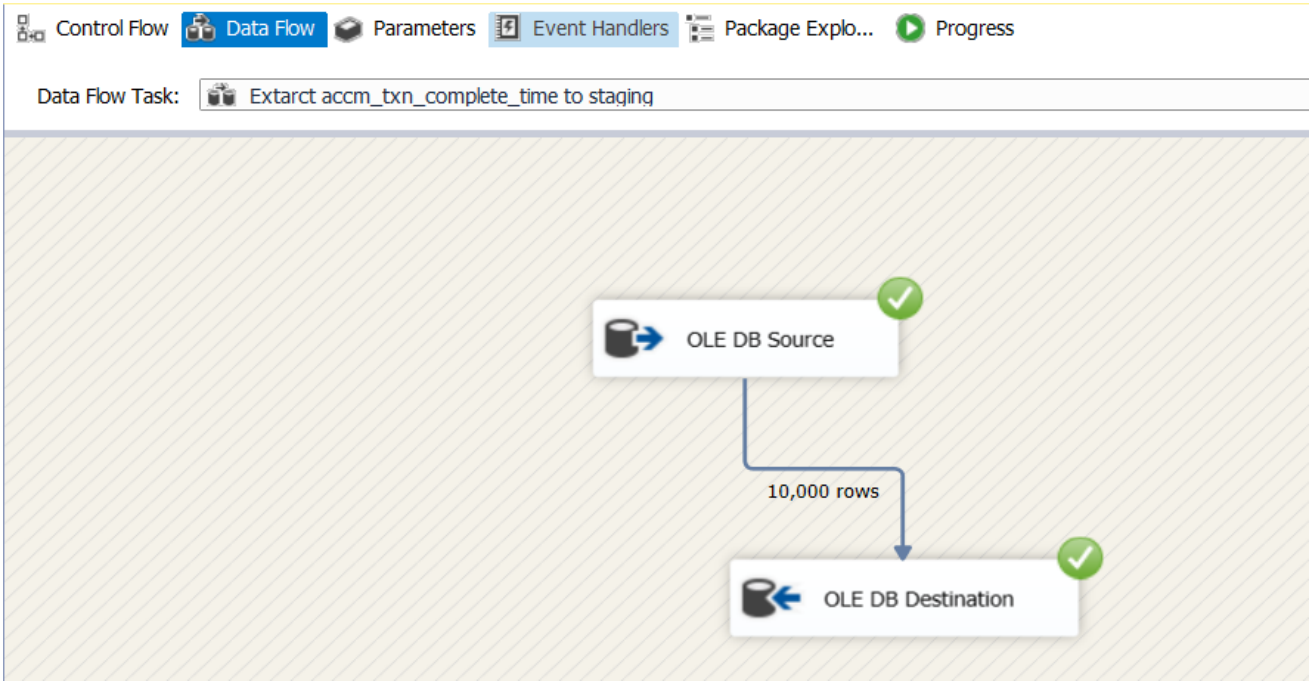
5.1.4 Load data Hotel Category to staging



5.1.5 Load data location to staging(.txt file)

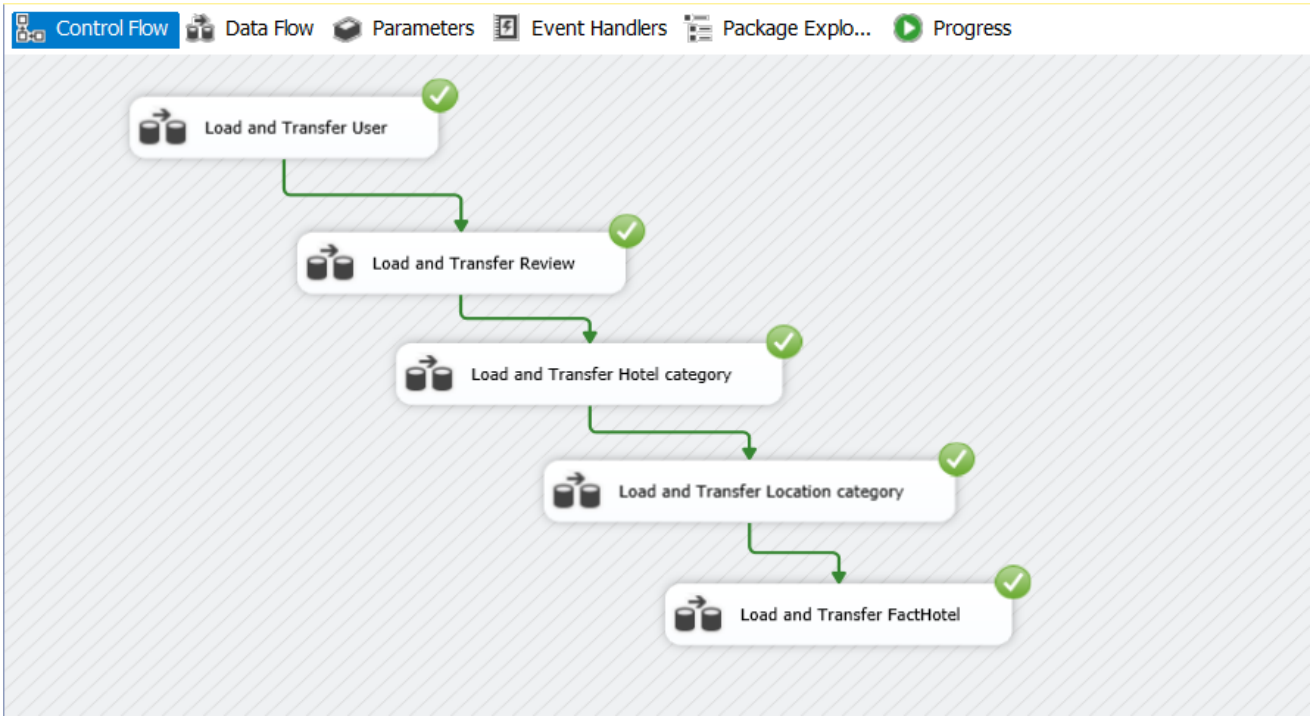


5.1.6 Load data accm_txn_complete_time to staging

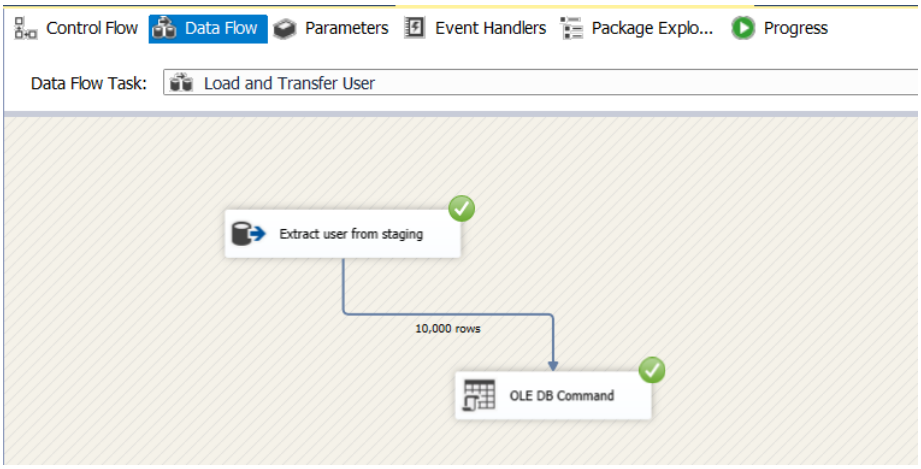


6. Staging To DW.

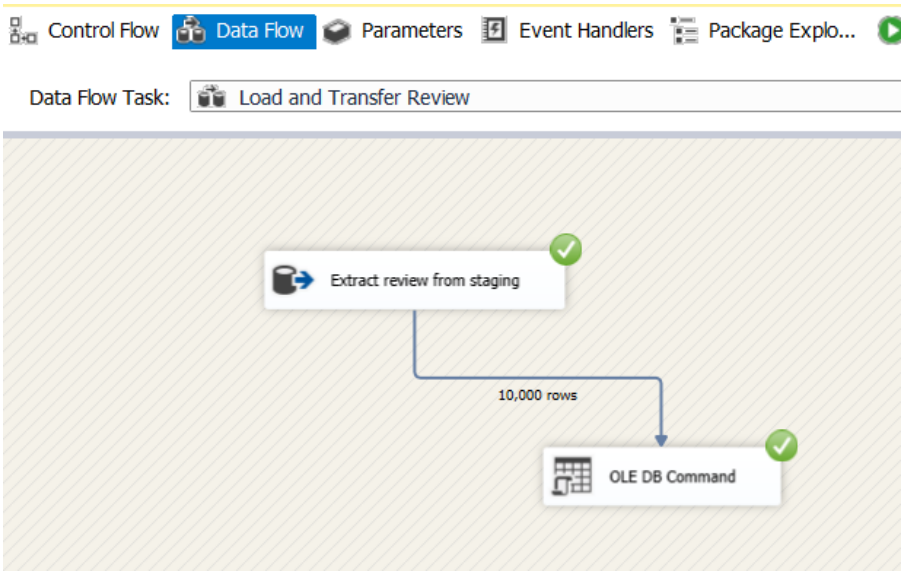
6.1 ETL System to Datawarehouse



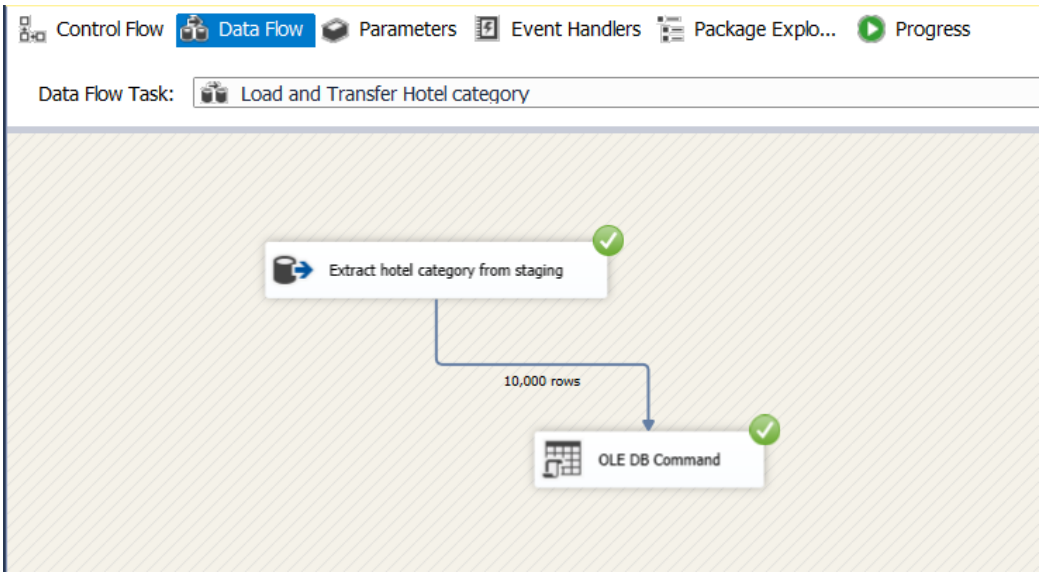
6.1.1. Transfer and Load DimUser Data from staging



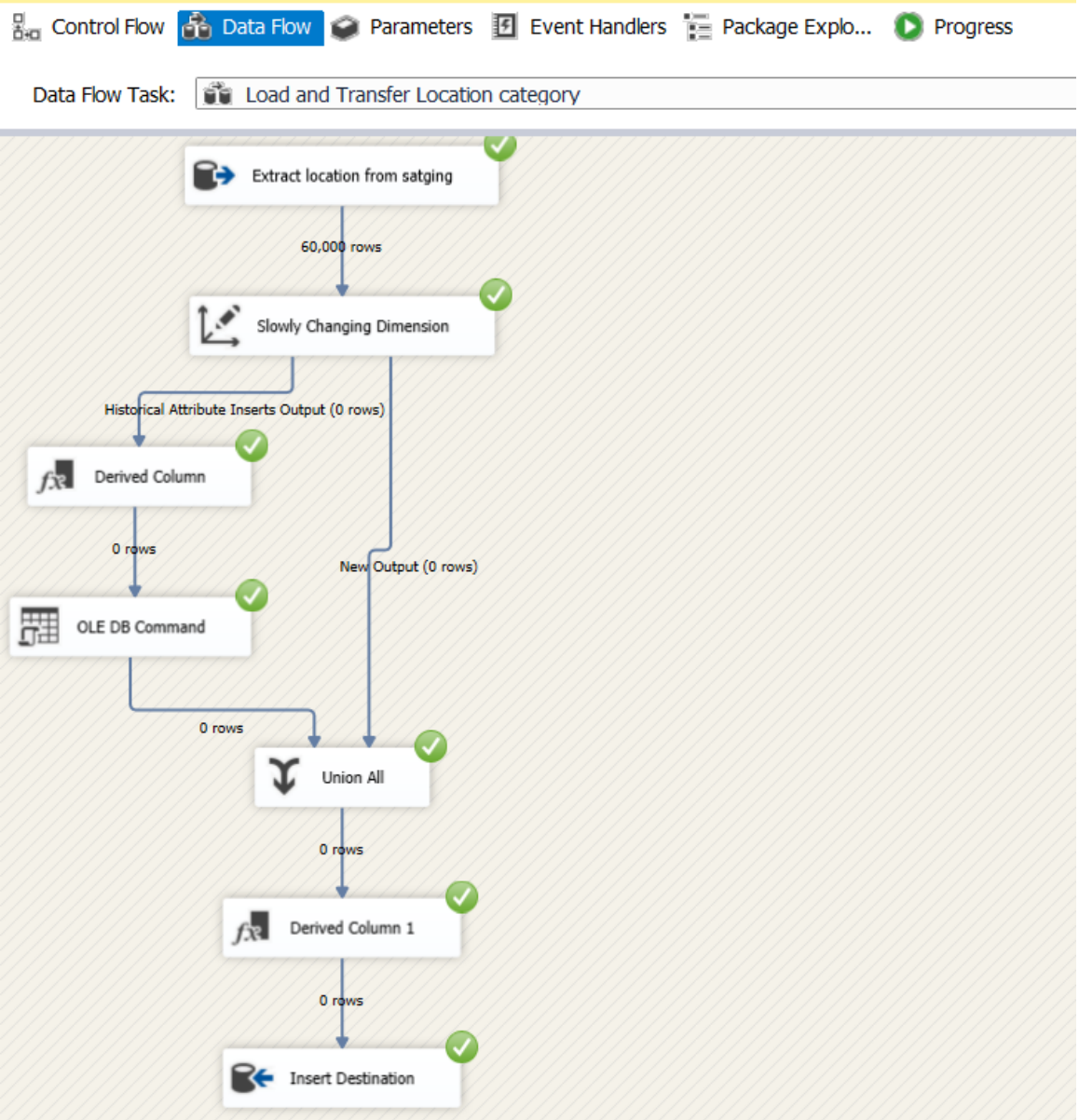
6.1.2. Transfer and Load DimReview Data from staging



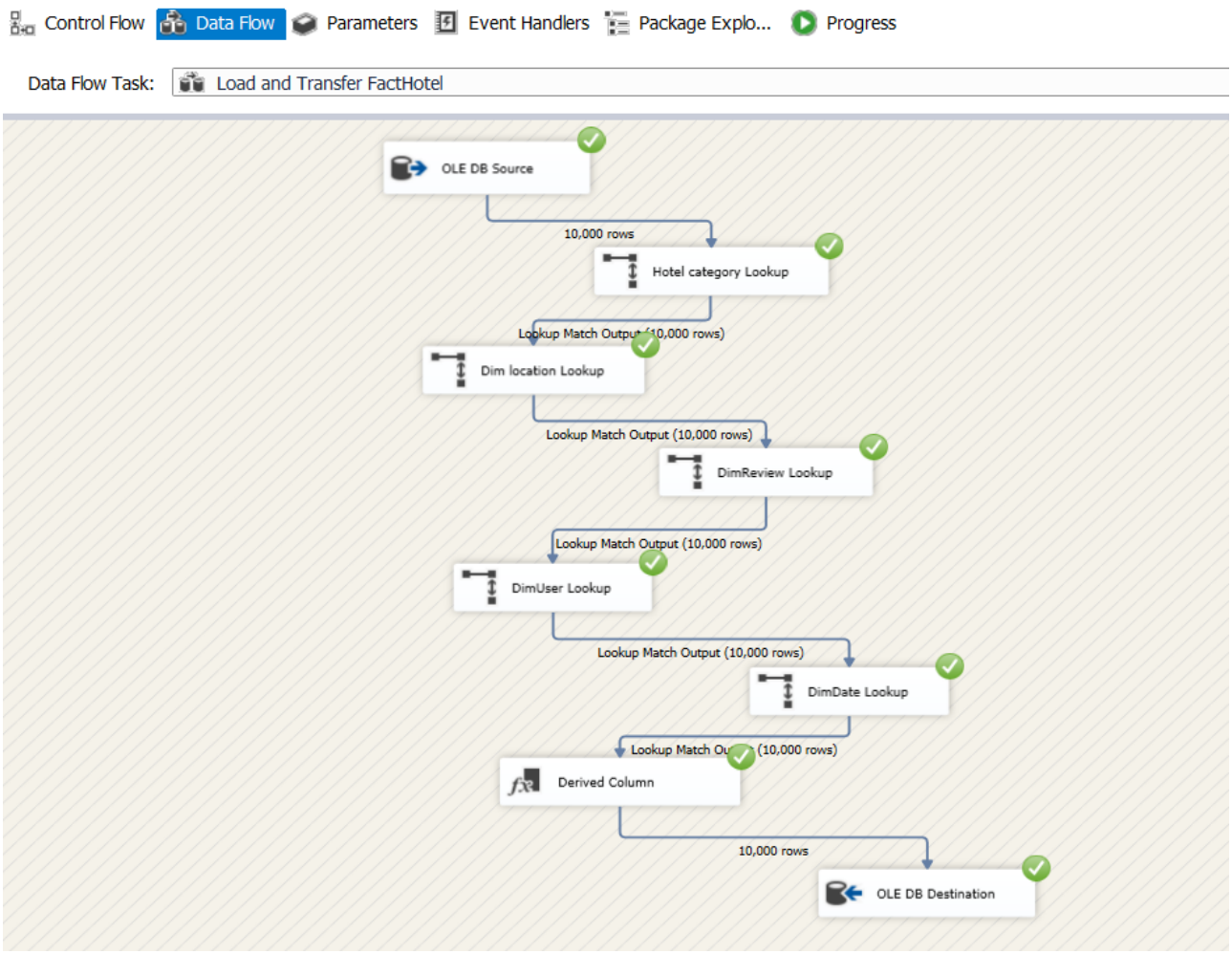
6.1.3. Transfer and Load DimHotelCategory Data from staging



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



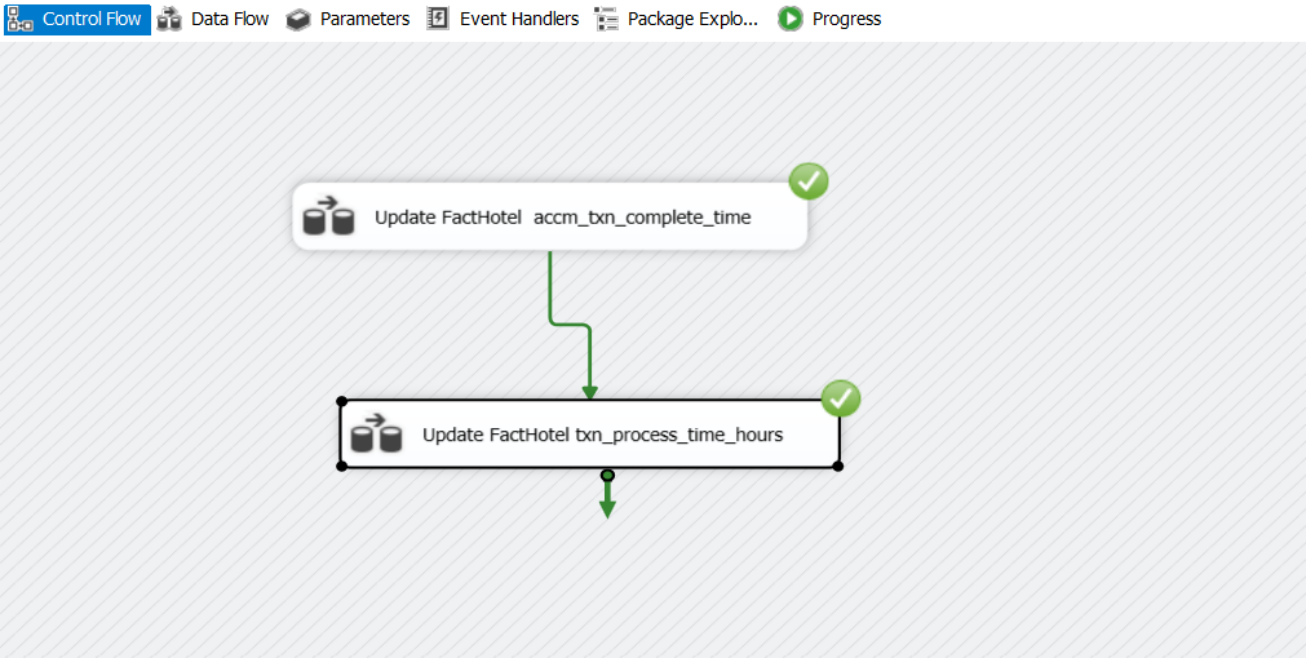
6.1.5. Load FactHotel Data from staging



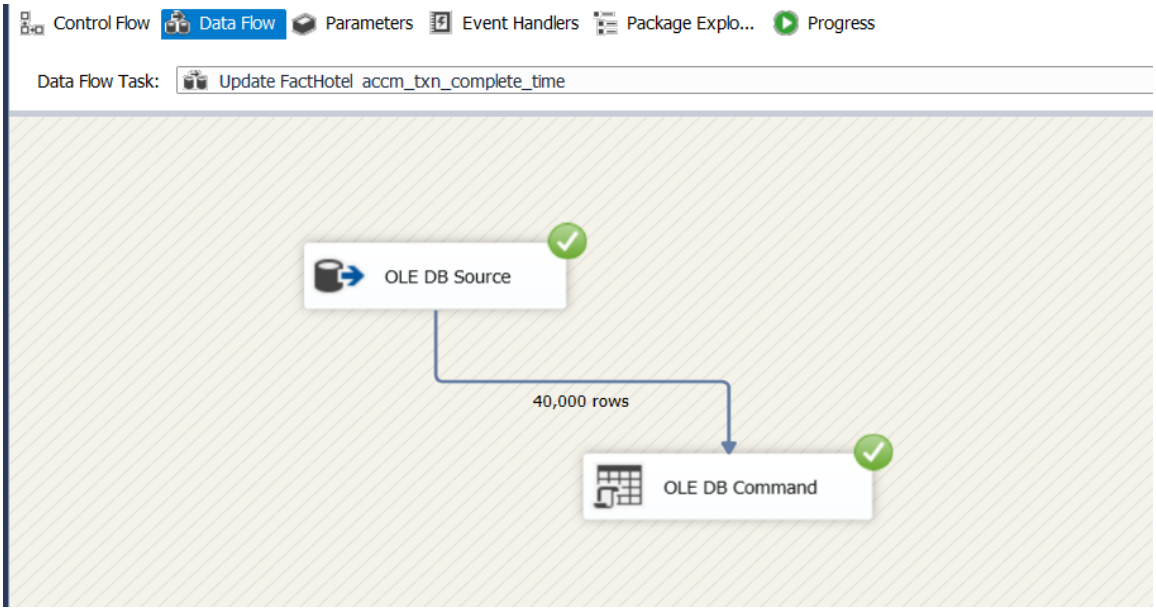
7. Datawarehouse Updating

To creating accumulated fact table I created a new SSIS package and updated accm_txn_complete_time and txn_process_time_hours.

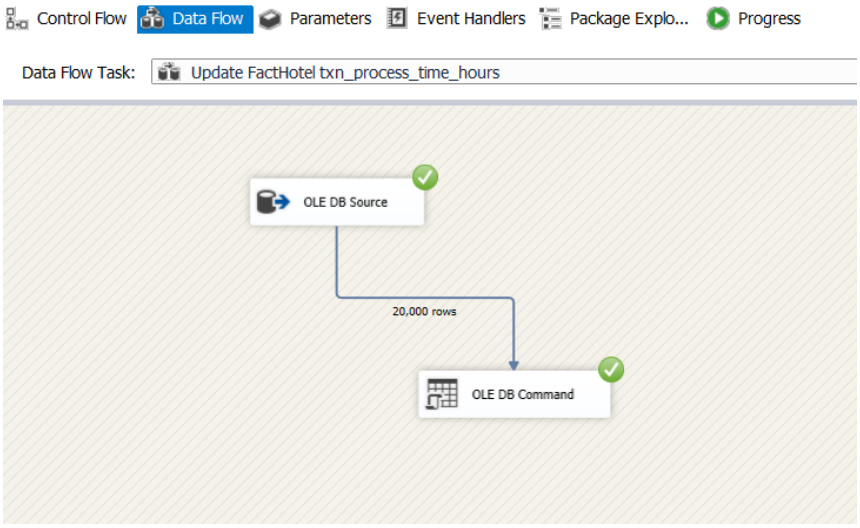
7.1. Datawarehouse updating



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

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

[End]