Sri Lanka Institute of Information Technology

Data Warehousing & Business Intelligence

Assignment 01

IT22918192

Rathnayaka S J

Year 03 semester 2- WE.DS.01.02

01. Data set selection

Dataset- Hotel reviews
Data source- Hotel Reviews

The Datafiniti Hotel Reviews dataset on Kaggle comprises information on 1,000 hotels, including their locations, names, ratings, review data, titles, usernames, and more. This dataset is suitable for various data analysis and machine learning tasks, such as sentiment analysis and natural language processing. It can be utilized to compare hotel reviews across different states and to experiment with sentiment scoring techniques. The dataset is provided in CSV format, facilitating easy integration into data warehousing and business intelligence projects. Its comprehensive structure makes it a valuable resource for exploring customer feedback and enhancing hotel services
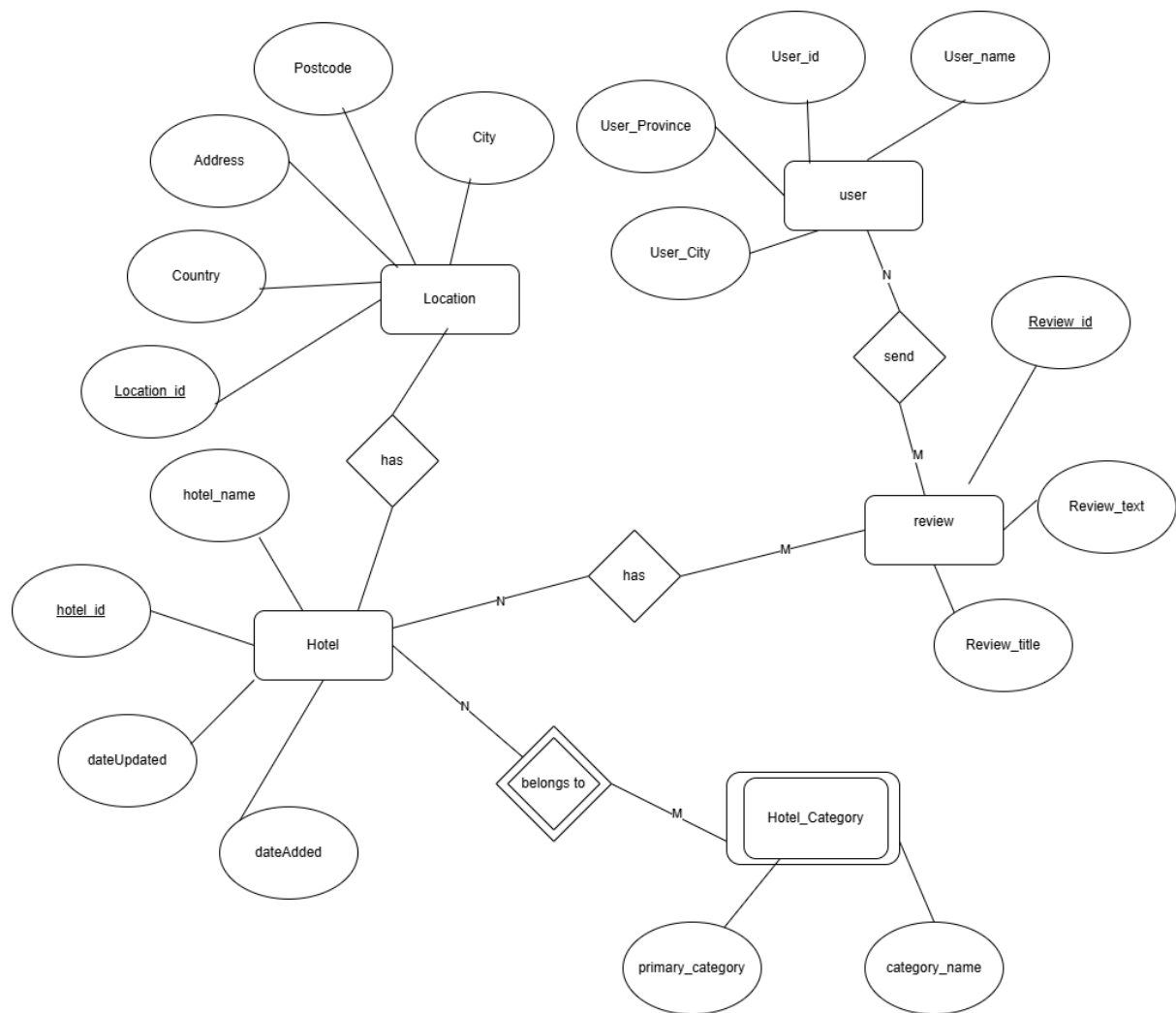
02.Preparation of data

All the data sources were initially provided in CSV format by the website. During the data preparation phase, modifications were made to the original files, such as adding new columns and splitting some information into separate tables. The files were then either converted into text files or imported directly into a source database.

Final Format of Source Data Before Transformation:

1. **CSV Files (.csv)**
   - Hotel data CSV file
   - Hotel Category CSV file
   - User data CSV file
   - Review data CSV file
   - Accommodation transaction complete time CSV file

These CSV files were imported into SQL Server Management Studio (SSMS) and organized within a newly created database named **Hotel_Source_DB**.

2. **Text File (.txt)**
   • Location data stored in a text file

03. Solution Architecture

Hotel_Staging.

- complete_time
- stgHotel
- stgHotalCategory
- stgReview
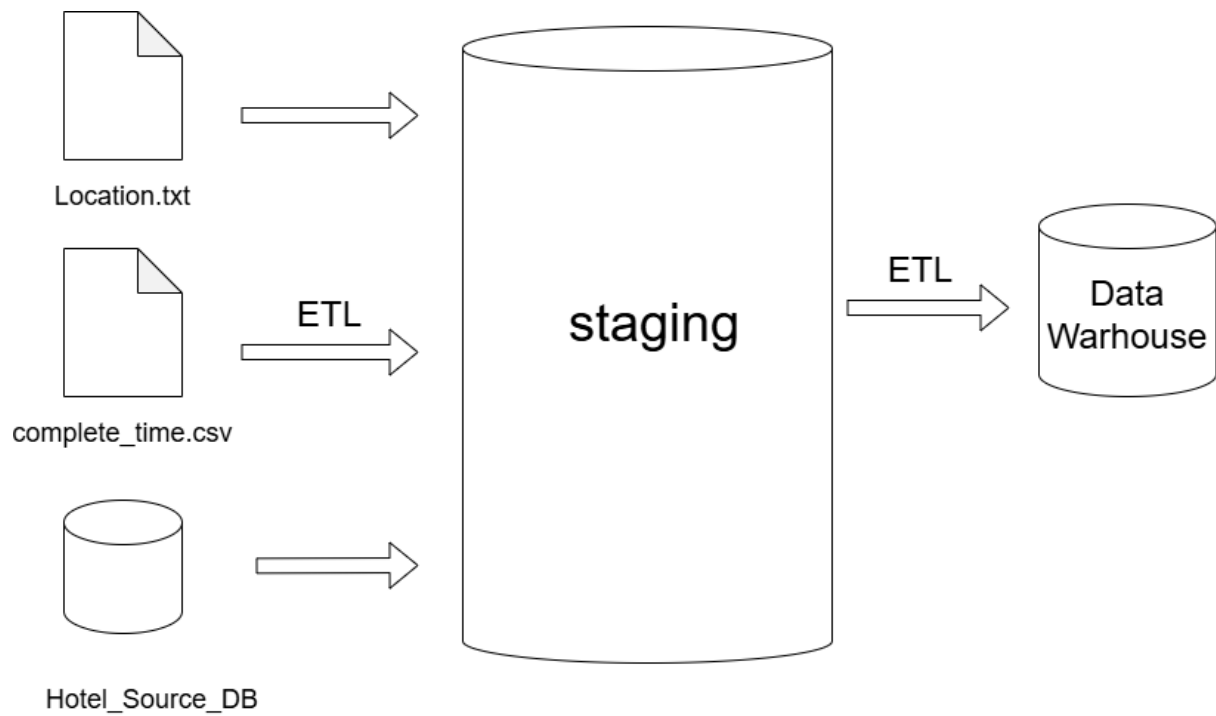- StgUser
- stgLocation

Hotel_DW

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

Architecture Components.

- **Data Sources.**
  - Transactional processing systems (operational databases)
  - Third-party data providers and external datasets

- **Extract, Transform and Load.**
  - Extraction phase- Retrieving information from source database
  - Conversion phase- Merging datasets, eliminating duplicates
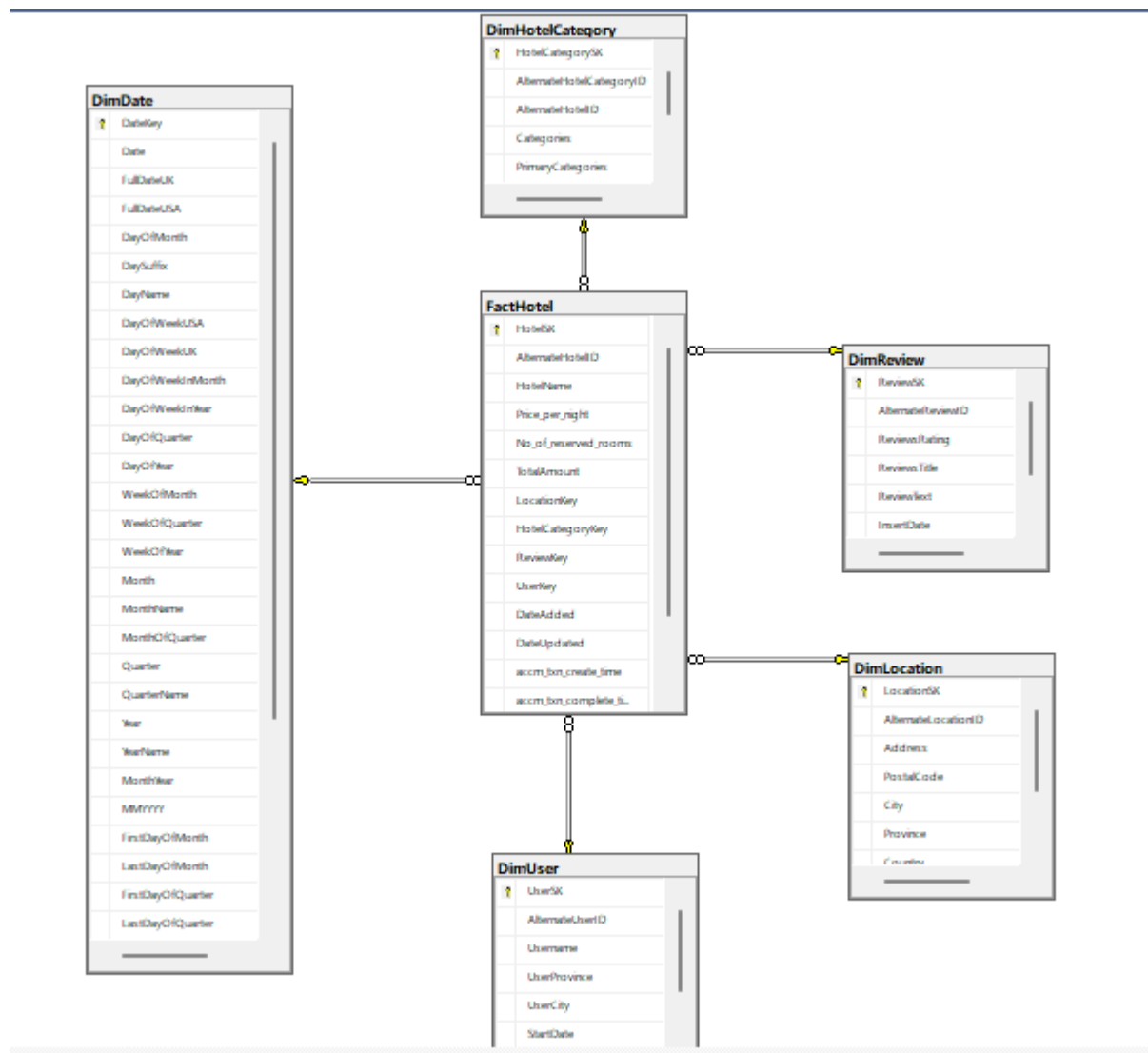 - Quality control- Standardizing and cleansing information

- Data Warehouse
  - EDW and Data Mart.
  - Dimensional Modeling- Facts and Dimensions.
  - Schema approach- Utilizing star-based structure for relationships

04. Data Warehouse Design & Development

Relational Diagram – Star Schema

DimHotelCategory
- HotelCategorySK
- AlternateHotelCategoryID
- AlternateHotelID
- Categories
- PrimaryCategories

DimDate
- DateKey
- Date
- FullDateUK
- FullDateUSA
- DayOfMonth
- DaySuffix
- DayName
- DayOfWeekUSA
- DayOfWeekUK
- DayOfWeekInMonth
- DayOfWeekInYear
- DayOfQuarter
- DayOfYear
- WeekOfMonth
- WeekOfQuarter
- WeekOfYear
- Month
- MonthName
- MonthOfQuarter
- Quarter
- QuarterName
- Year
- YearName
- MonthYear
- MMYYYY
- FirstDayOfMonth
- LastDayOfMonth
- FirstDayOfQuarter
- LastDayOfQuarter

FactHotel
- HotelSK
- AlternateHotelID
- HotelName
- Price_per_night
- No_of_reserved_rooms
- TotalAmount
- LocationKey
- HotelCategoryKey
- ReviewKey
- UserKey
- DateAdded
- DateUpdated
- accm_txn_create_time
- accm_txn_complete_ti...

DimReview
- ReviewSK
- AlternateReviewID
- ReviewRating
- ReviewTitle
- ReviewText
- InsertDate

DimLocation
- LocationSK
- AlternateLocationID
- Address
- PostalCode
- City
- Province
- Country

DimUser
- UserSK
- AlternateUserID
- Username
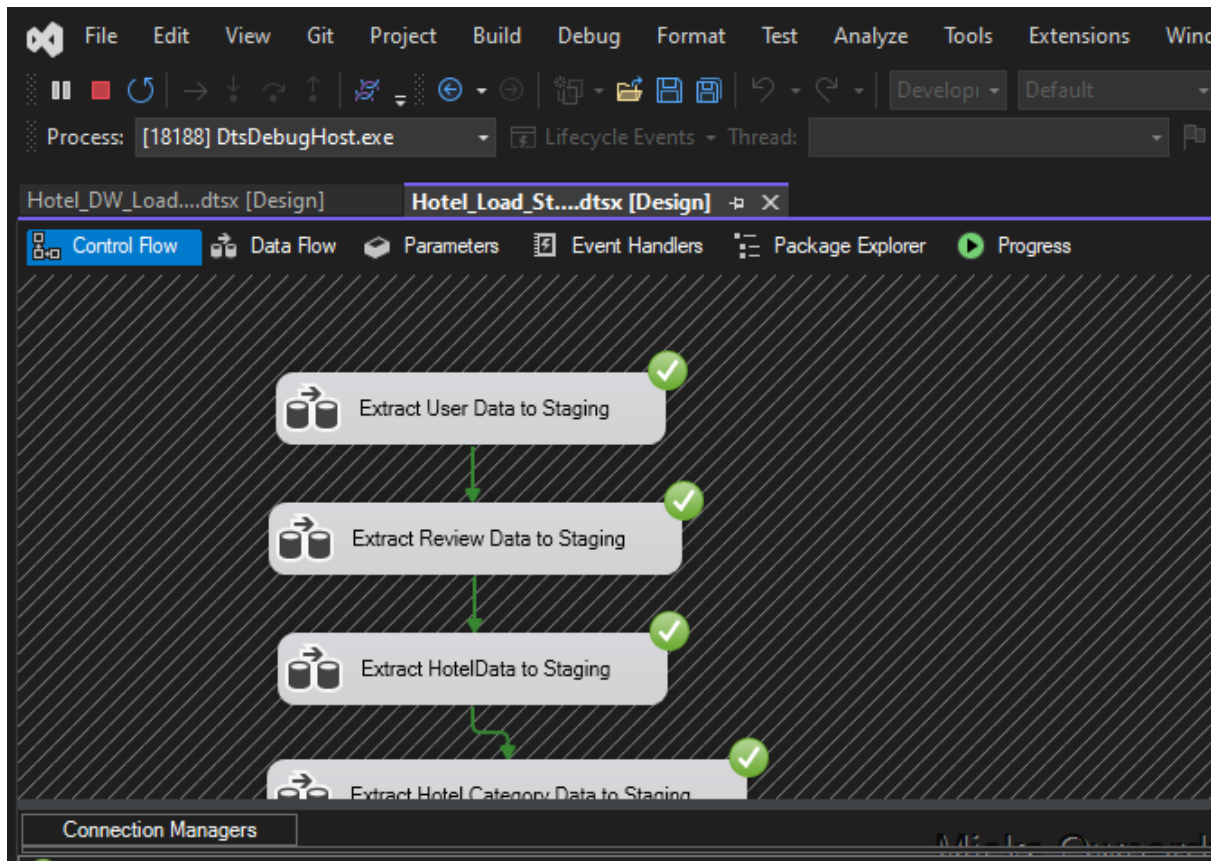- UserProvince
- UserCity
- StartDate

- This schema consists with 5 dimensional tables and 1 fact table.
- I considered DimLocation as a slowly changing dimensions because addresses can be changed in future.

05.ETL Development
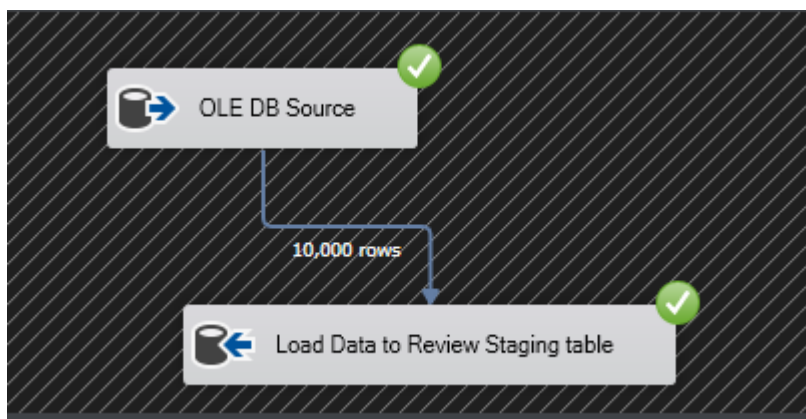
5.1 Staging tables
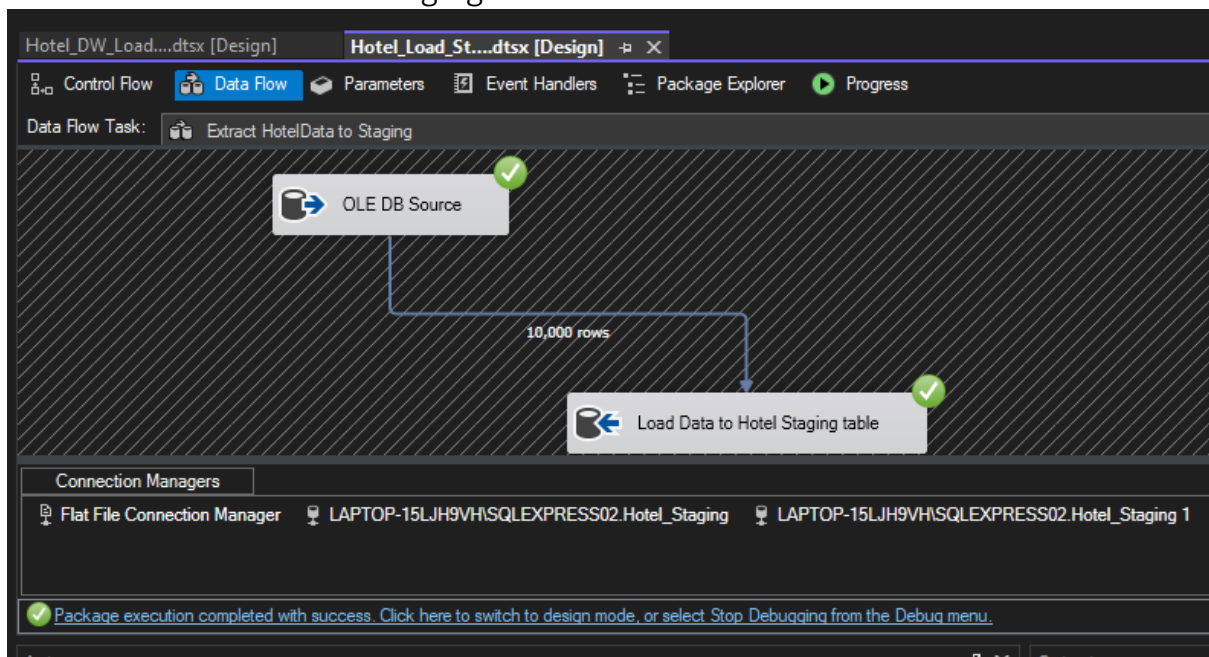
- Load the tables from data sources to staging

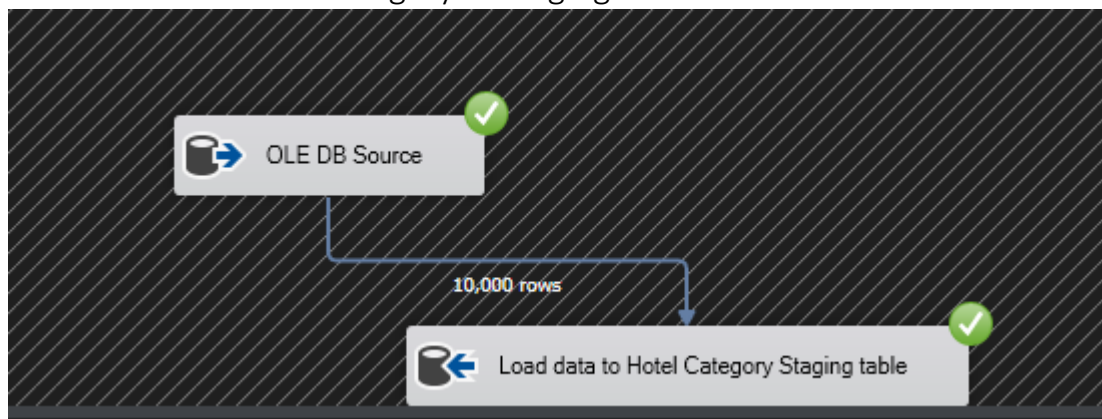### 5.1.1 Load data User to staging
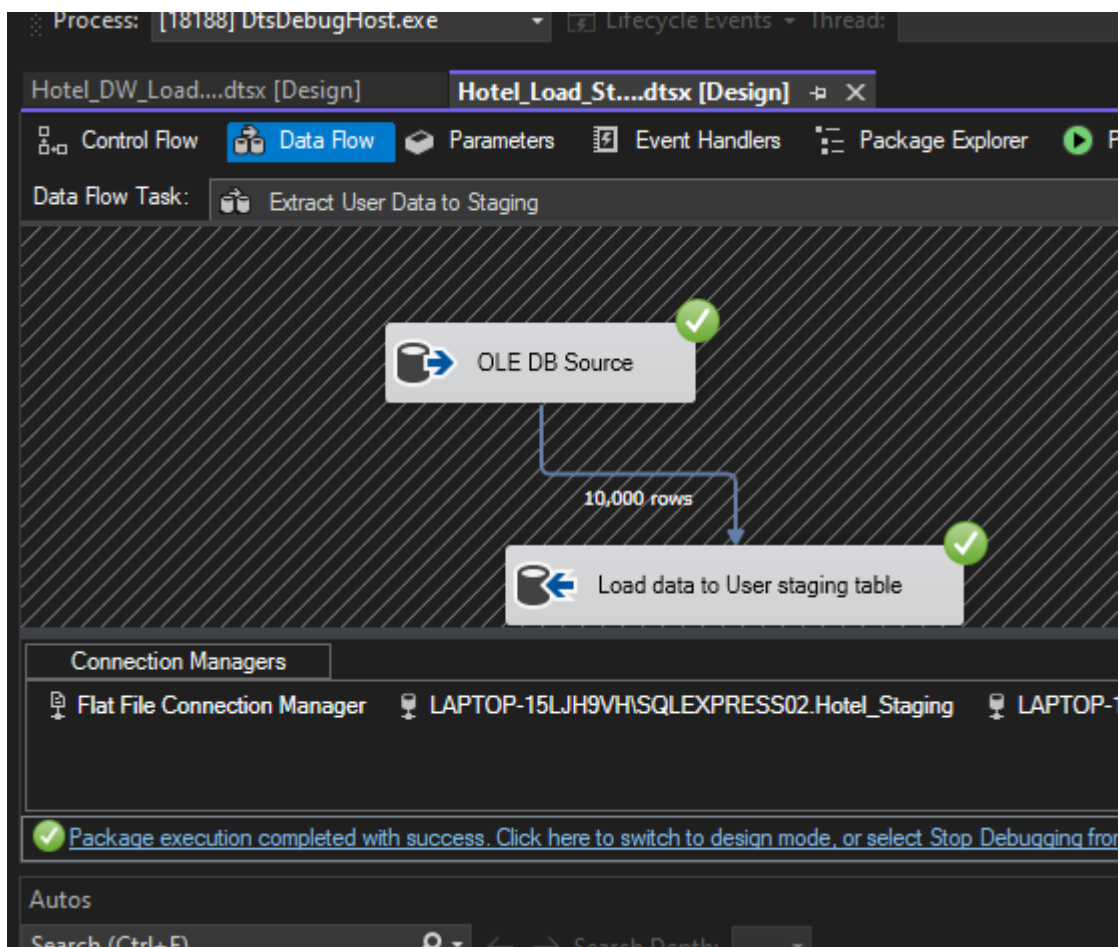
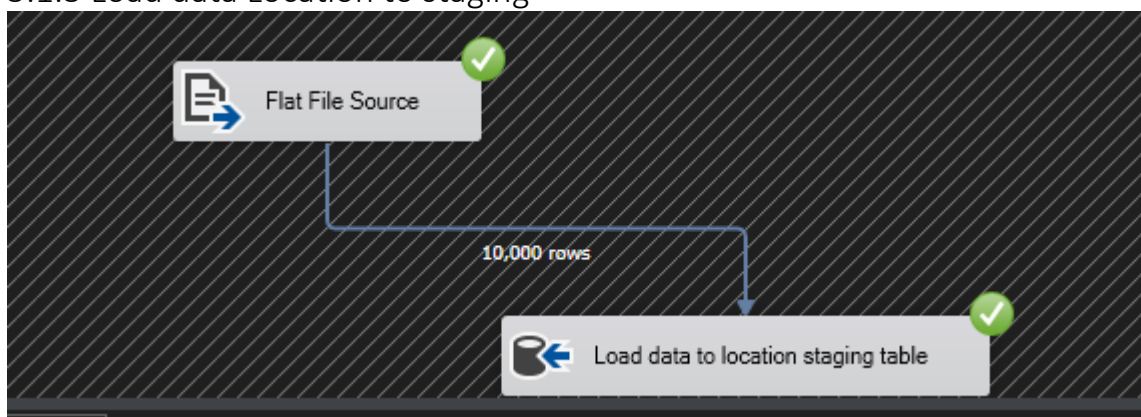### 5.1.2 Load data Review to staging

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

5.1.6 Load data complete_time to staging
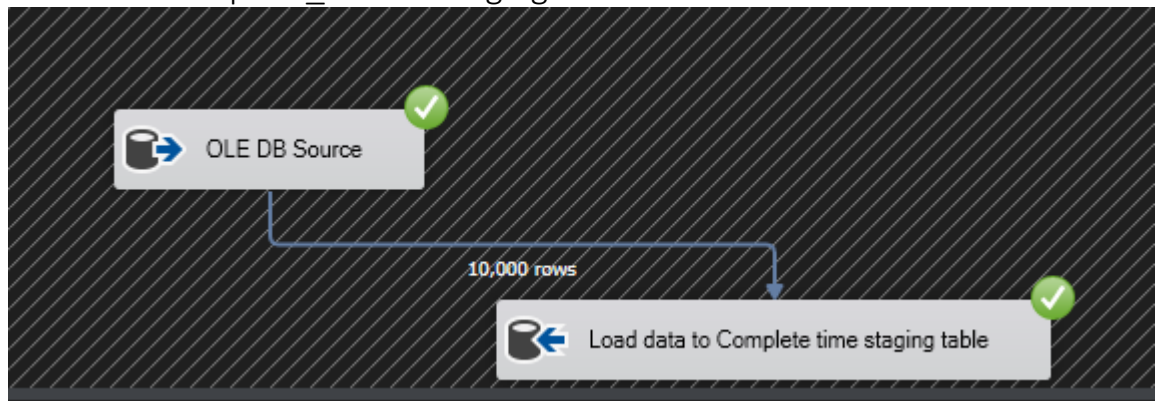
06.Staging to DW

- Loading the staging tables to datawarehouse.

6.1 ETL System to Datawarehouse
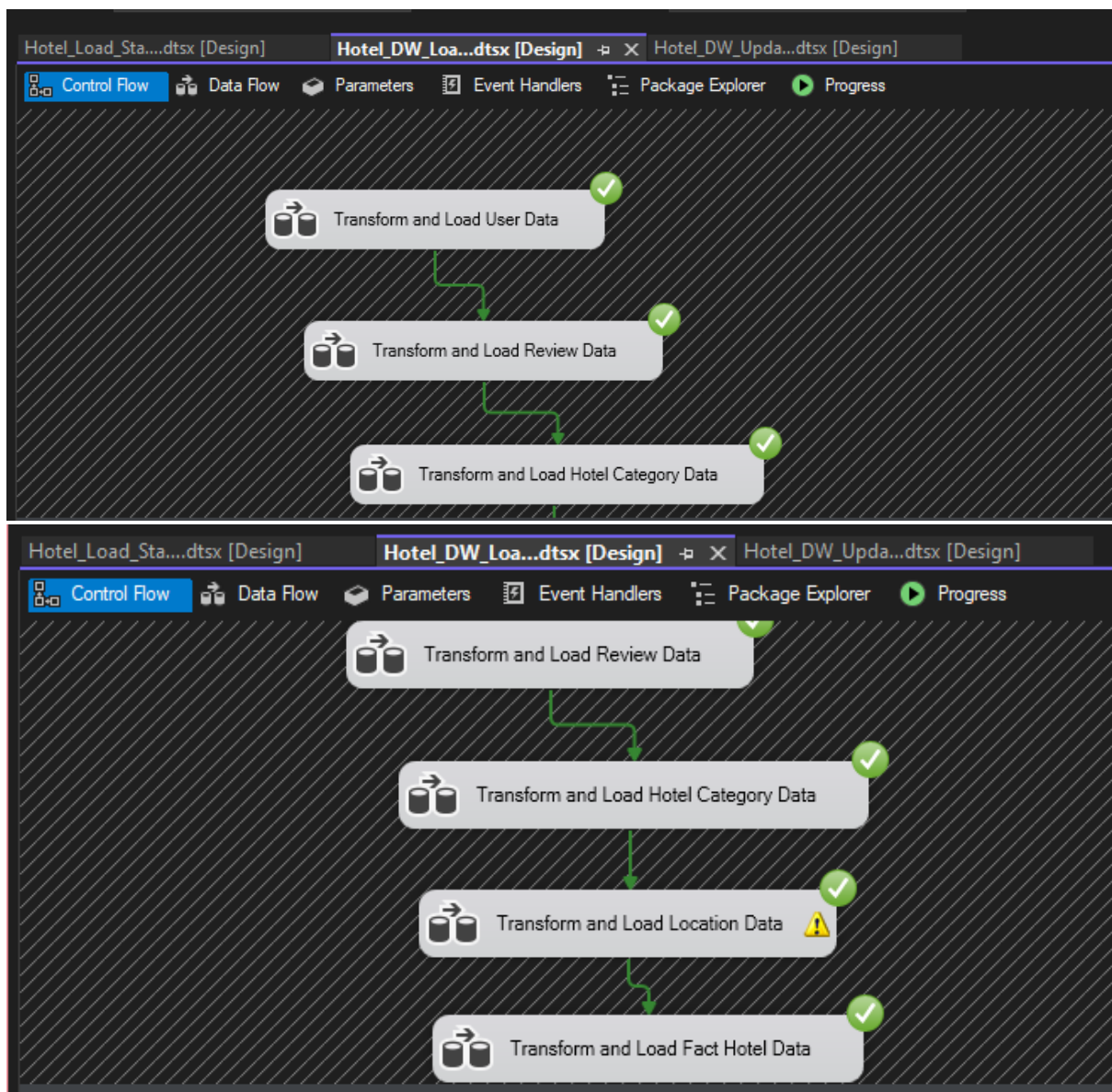
6.1.1 Transfer and Load DimUser Data from staging
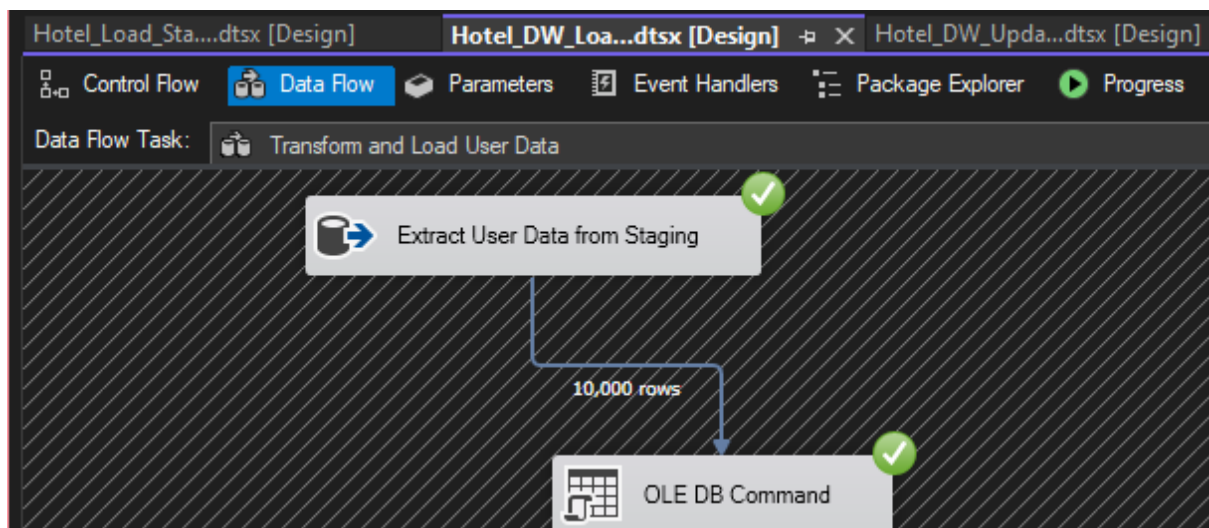
6.1.2 Transfer and Load DimReview Data from staging
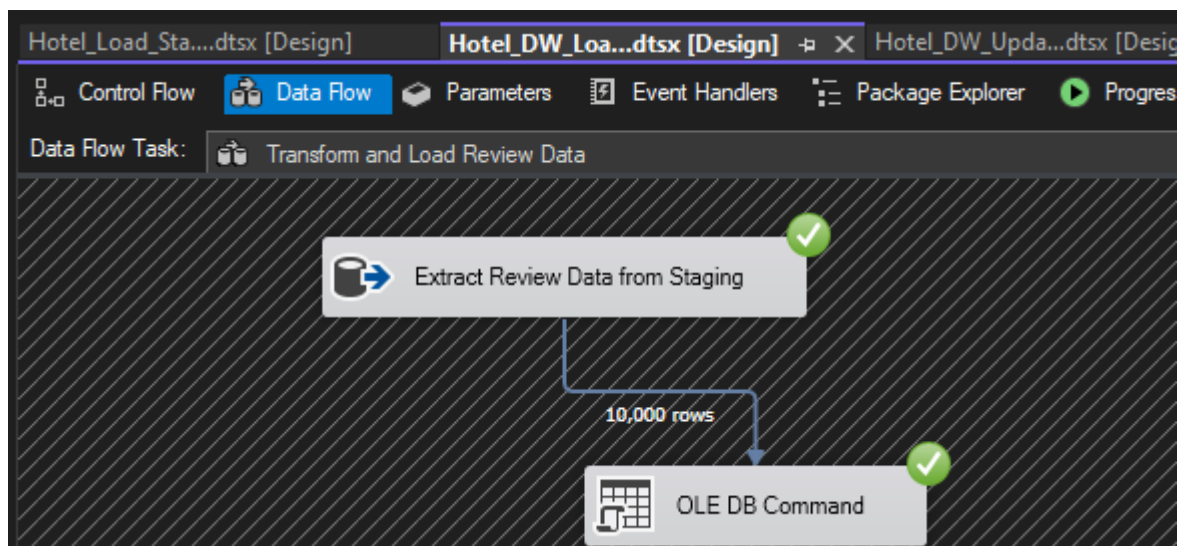
6.1.3 Transfer and Load DimHotelCategory Data from staging

6.1.4 Transfer and Load DimLocation Data from staging

6.1.5 Load FactHotel Data from staging

07. Datawarehouse Updating

7.1.1 Update FactHotel accm_txn_complete_time

7.1.2 Update FactHotel txn_process_time_hours

08. Accumulated Fact Table

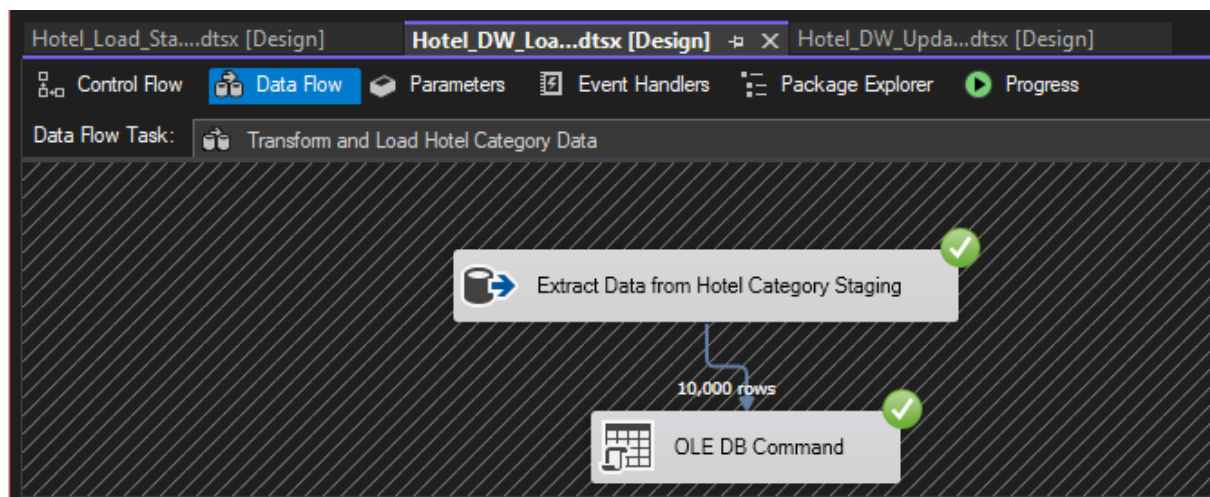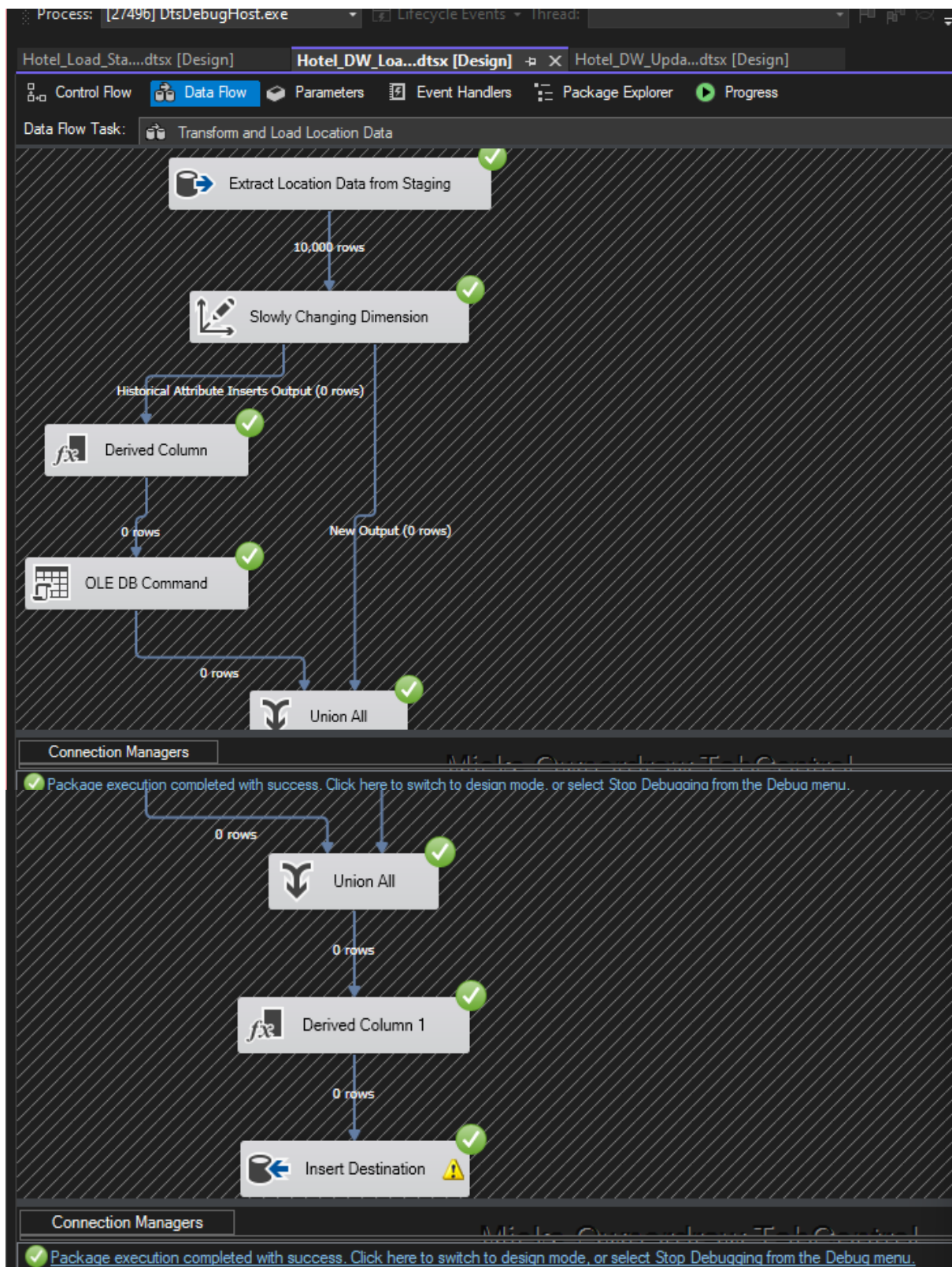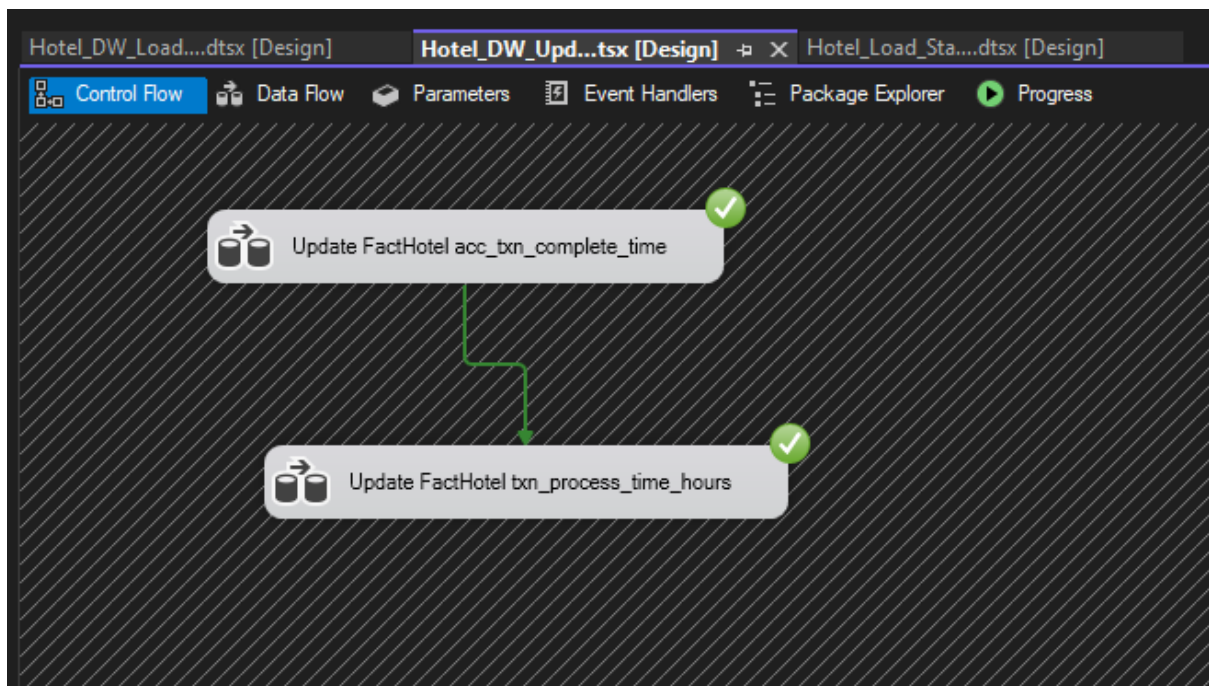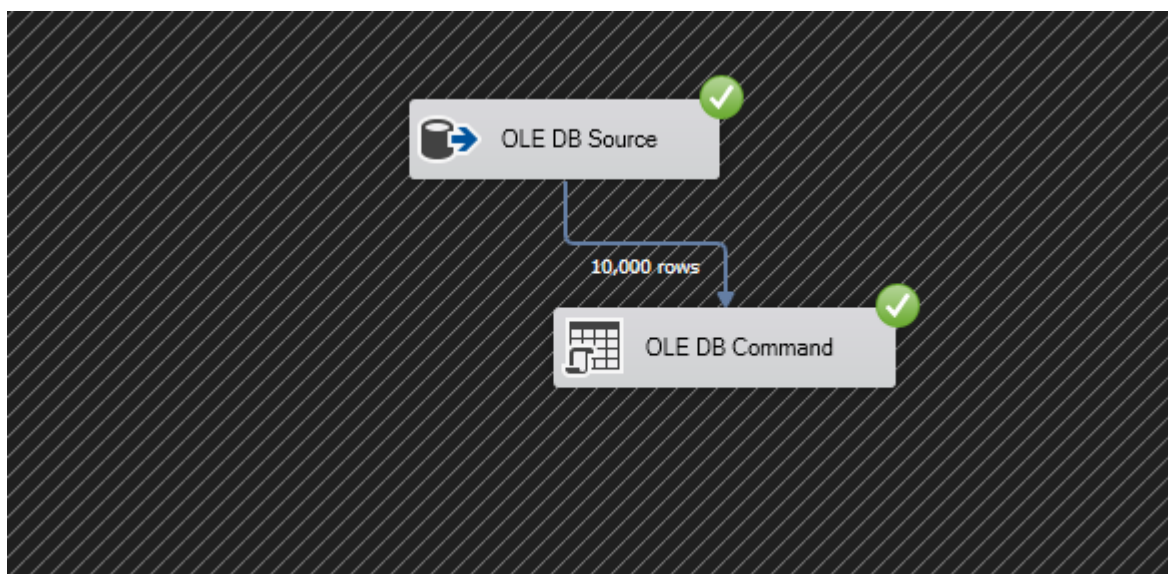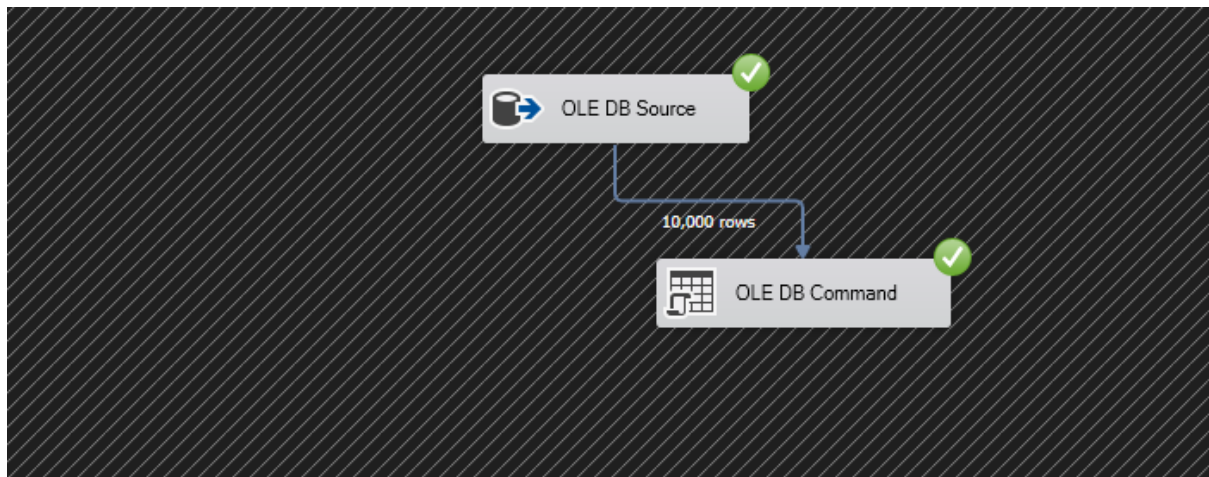|	HotelSK	AlternateHotelID	HotelName	Price_per_night	No_of_reserved_rooms	TotalAmount	LocationKey	HotelCategoryKey	ReviewKey	UserKey	DateAdded	DateUpdated
183	183	183	Fremont Hotel & Casino	25.00	2	50.00	183	84	6840	183	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
184	184	184	Main Street Station Casino Brewery H...	25.00	3	75.00	184	85	6841	184	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000
185	185	185	Fremont Hotel & Casino	25.00	2	50.00	185	86	6842	185	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
186	186	186	Main Street Station Casino Brewery H...	25.00	2	50.00	186	87	6843	186	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000
187	187	187	Fremont Hotel & Casino	25.00	3	75.00	187	88	6844	187	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
188	188	188	Fremont Hotel & Casino	24.75	7	173.25	188	89	6845	188	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
189	189	189	Fremont Hotel & Casino	23.00	3	69.00	189	90	6846	189	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
190	190	190	Fremont Hotel & Casino	23.00	2	46.00	190	91	6847	190	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
191	191	191	Fremont Hotel & Casino	23.00	2	46.00	191	92	6848	191	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
192	192	192	Fremont Hotel & Casino	23.79	2	47.59	192	93	6849	192	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
193	193	193	Fremont Hotel & Casino	24.00	2	48.00	193	94	6850	193	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
194	194	194	Fremont Hotel & Casino	25.00	2	50.00	194	95	6851	194	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
195	195	195	Fremont Hotel & Casino	25.75	2	51.50	195	96	6852	195	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
196	196	196	Fremont Hotel & Casino	26.75	3	80.25	196	97	6853	196	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
197	197	197	Main Street Station Casino Brewery H...	27.00	4	108.00	197	98	6854	197	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000
198	198	198	Fremont Hotel & Casino	26.00	4	104.00	198	99	6855	198	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
199	199	199	Main Street Station Casino Brewery H...	26.00	4	104.00	199	100	6856	199	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000
200	200	200	Main Street Station Casino Brewery H...	25.25	4	101.00	200	101	6857	200	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000
201	201	201	Main Street Station Casino Brewery H...	25.00	5	125.00	201	102	6858	201	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000
202	202	202	Fremont Hotel & Casino	25.75	3	77.25	202	103	6859	202	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
203	203	203	Fremont Hotel & Casino	25.00	4	100.00	203	104	6860	203	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
204	204	204	Fremont Hotel & Casino	24.50	3	73.50	204	105	6861	204	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000
205	205	205	Fremont Hotel & Casino	24.25	3	72.75	205	106	6862	205	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000

Query executed successfully.   LAPTOP-15LJH9VH\SQLEXPRESS0... | LAPTOP-15LJH9VH\Sathya... | Hotel_DW | 00:00:00 | 1,000 rows

	r_night	No_of_reserved_rooms	TotalAmount	LocationKey	HotelCategoryKey	ReviewKey	UserKey	DateAdded	DateUpdated	accm_txn_create_time	accm_txn_complete_time	txn_process_time_hours
183	183	2	50.00	183	84	6840	183	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-06-10 00:00:00.000	964
184	184	3	75.00	184	85	6841	184	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-06-11 00:00:00.000	988
185	185	2	50.00	185	86	6842	185	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-06-12 00:00:00.000	1012
186	186	2	50.00	186	87	6843	186	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-06-13 00:00:00.000	1036
187	187	3	75.00	187	88	6844	187	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-06-14 00:00:00.000	1060
188	188	7	173.25	188	89	6845	188	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-14 00:00:00.000	316
189	189	3	69.00	189	90	6846	189	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-15 00:00:00.000	340
190	190	2	46.00	190	91	6847	190	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-16 00:00:00.000	364
191	191	2	46.00	191	92	6848	191	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-17 00:00:00.000	388
192	192	2	47.59	192	93	6849	192	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-18 00:00:00.000	412
193	193	2	48.00	193	94	6850	193	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-19 00:00:00.000	436
194	194	2	50.00	194	95	6851	194	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-20 00:00:00.000	460
195	195	2	51.50	195	96	6852	195	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-21 00:00:00.000	484
196	196	3	80.25	196	97	6853	196	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-22 00:00:00.000	508
197	197	4	108.00	197	98	6854	197	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-23 00:00:00.000	532
198	198	4	104.00	198	99	6855	198	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-24 00:00:00.000	556
199	199	4	104.00	199	100	6856	199	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-25 00:00:00.000	580
200	200	4	101.00	200	101	6857	200	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-26 00:00:00.000	604
201	201	5	125.00	201	102	6858	201	2018-04-09 05:56:34.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-27 00:00:00.000	628
202	202	3	77.25	202	103	6859	202	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-28 00:00:00.000	652
203	203	4	100.00	203	104	6860	203	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-29 00:00:00.000	676
204	204	3	73.50	204	105	6861	204	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-30 00:00:00.000	700
205	205	3	72.75	205	106	6862	205	2018-04-08 12:17:12.000	2018-09-04 21:27:52.000	2025-04-30 20:02:32.353	2025-05-31 00:00:00.000	724

Query executed successfully.   LAPTOP-15LJH9VH\SQLEXPRESS0... | LAPTOP-15LJH9VH\Sathya... | Hotel_DW | 00:00:00 | 1,000 rows