Sri Lanka Institute of Information Technology

Retail Transactional Dataset
Insights into Consumer Behaviour and Operations

Submitted on: 2025 05 01

Program: **BSc (Hons) in Information Technology**
Specialization: **Data Science**
Module: **IT3021 – Data Warehousing and Business Intelligence**
Assignment **1 – Year 3 Semester 2, 2025**

Prepared by: G**amage D.M.G.P.K**
Student ID: **IT22188472**

Table of Contents

Data Set Selection

Dataset Overview

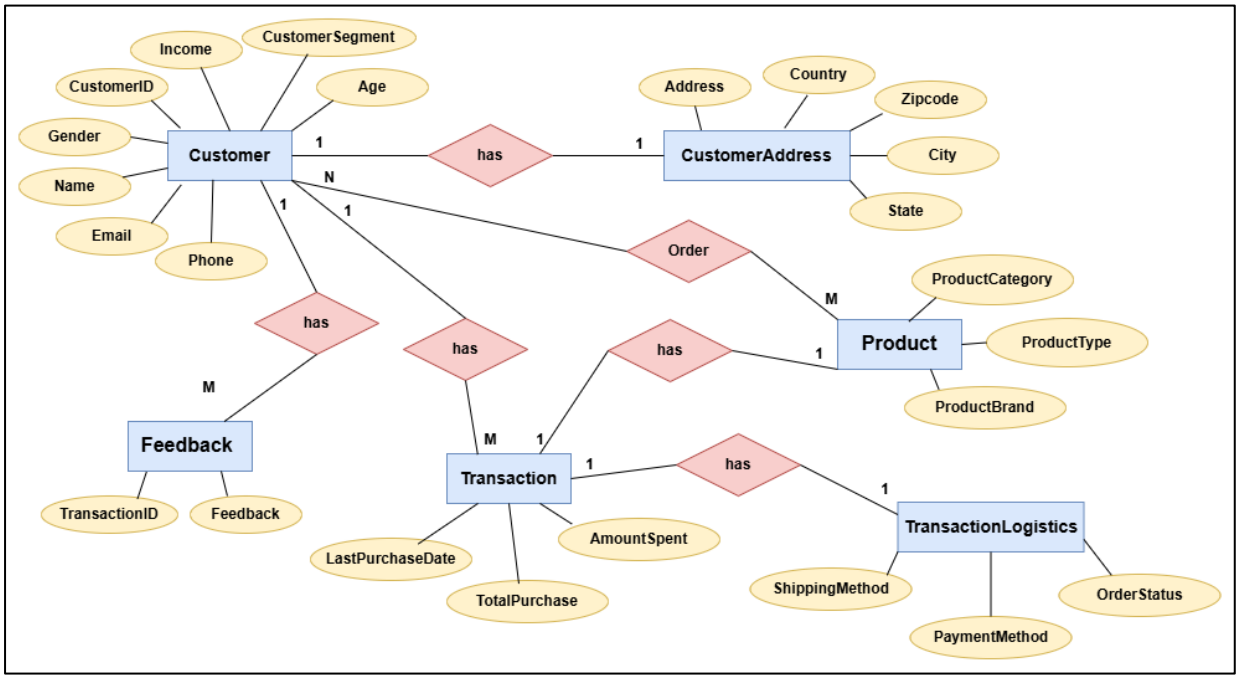
The dataset includes a detailed collection of retail transaction data, which supports complete DW/BI functionality within the retail system. The dataset extends over one year while holding thousands of records among multiple entities and reaching complete volume and temporal requirements through synthesised data.

The data organization supports analysis of four key domains:

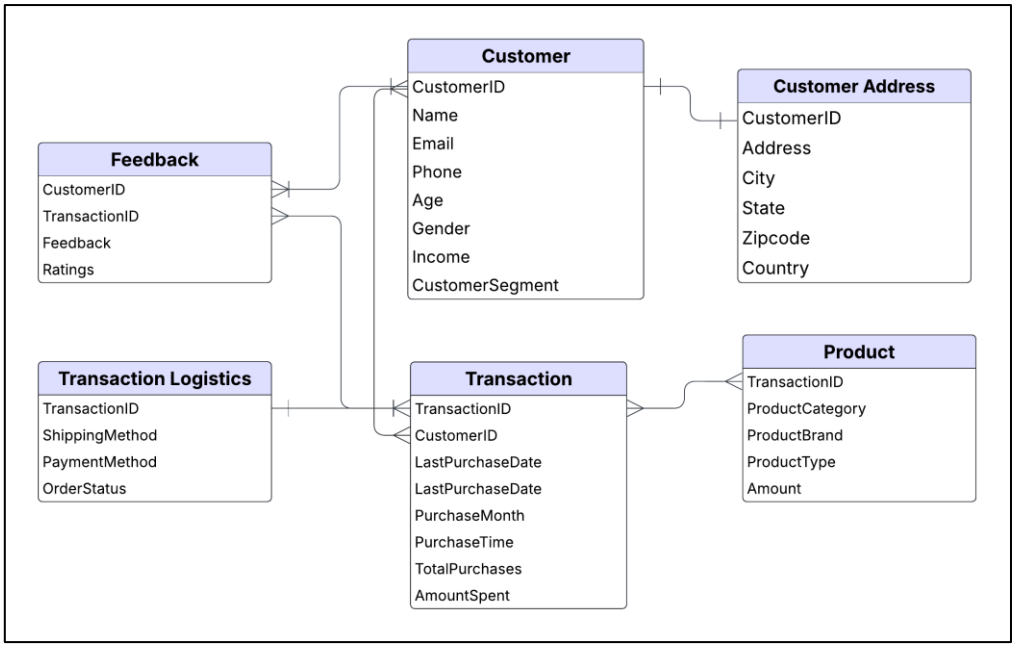
- Customer demographics and segmentation
- Purchasing behaviors over time
- Product preferences and trends
- Transaction processing and logistics
- Customer satisfaction through feedback

The dataset is highly suitable for building OLAP cubes while enabling ETL transformations and dimensional model creation for reporting and decisionmaking purposes.

ER Diagram



ER Diagram



Diagram

Customer

The entity contains complete profiles about customers including their information.

- ❖ CustomerID (PK): Unique identifier for each customer.

Transaction

Captures individual transactions or cumulative behavior per customer.

- ❖ TransactionID (PK): Unique transaction identifier.
- ❖ CustomerID (FK) that refers to the involved customer.

Product

The table stores details of items which are present in transactions.

- ProductID (PK): Unique identifier for each product.

Feedback

Captures customer experience data.

- FeedbackID (PK): Unique identifier for each feedback record.
- TransactionID number provides a reference to its linked transaction record.

TransactionLogistics

The table stores information about delivery solutions and payment processing for each transaction.

- TransactionID (PK/FK): Shared key with Transaction table.

CustomerAddress

The entity contains complete customers Addresses.

- ❖ CustomerID (FK) that refers to the involved customer.

❖ Why This Dataset is Suitable for DW/BI

1. **OLTP Design**: The dataset is normalised and suitable for transformation into a dimensional model.
2. **Data Variety**: Includes demographic, transactional, product, and operational data.
3. **Multiple Sources**: Can be split into multiple files/tables for simulation

   CSV Files - Customer.csv / Feedback.csv
   Test Files - Product.txt / CustomerAddress.txt / TransactionLogistics.txt
   SQL Files – Transaction.sql

4. **BI Readiness**: Includes fields suitable for building hierarchies, dimensions, and accumulating fact tables.
5. **Realism & Scalability**: Synthetic data added to ensure year-round transactions with meaningful relationships.

Preparation of data sources

Data Sources Used:

1. CSV Files:
   - Customers.csv
   - feedback.csv

2. SQL Server Database:
   - Transactions.sql
3. Text Files:
   - Product.txt
   - CustomerAddress.txt
   - TransactionLogistics.txt

Customer-related details are separated into a CSV file, while transactional and logistics details are stored in a SQL database. This allows the simulation of real-world data integration scenarios using multiple formats.

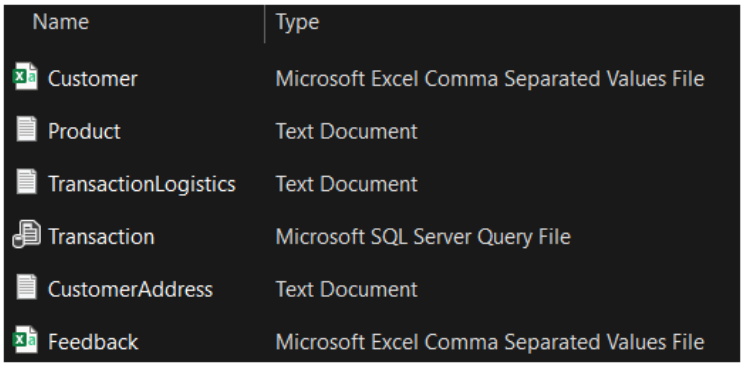


Table Structure

1. Customer Table

	COLUMN_NAME	DATA_TYPE	Length	IS_NULLABLE
1	CustomerID	varchar	50	YES
2	Name	varchar	50	YES
3	Email	varchar	50	YES
4	Phone	varchar	50	YES
5	Age	varchar	50	YES
6	Gender	varchar	50	YES
7	Income	varchar	50	YES
8	CustomerSegment	varchar	50	YES

2. Customer Address Table

	COLUMN_NAME	DATA_TYPE	Length	IS_NULLABLE
1	Customer_ID	varchar	50	YES
2	Address	varchar	50	YES
3	City	varchar	50	YES
4	State	varchar	50	YES
5	Zipcode	varchar	50	YES
6	Country	varchar	50	YES

3. Feedback Table

	COLUMN_NAME	DATA_TYPE	Length	IS_NULLABLE
1	CustomerID	varchar	50	YES
2	TransactionID	varchar	50	YES
3	Feedback	varchar	50	YES

4. Product Table

	COLUMN_NAME	DATA_TYPE	Length	IS_NULLABLE
1	TransactionID	varchar	50	YES
2	ProductCategory	varchar	50	YES
3	ProductBrand	varchar	50	YES
4	ProductType	varchar	50	YES

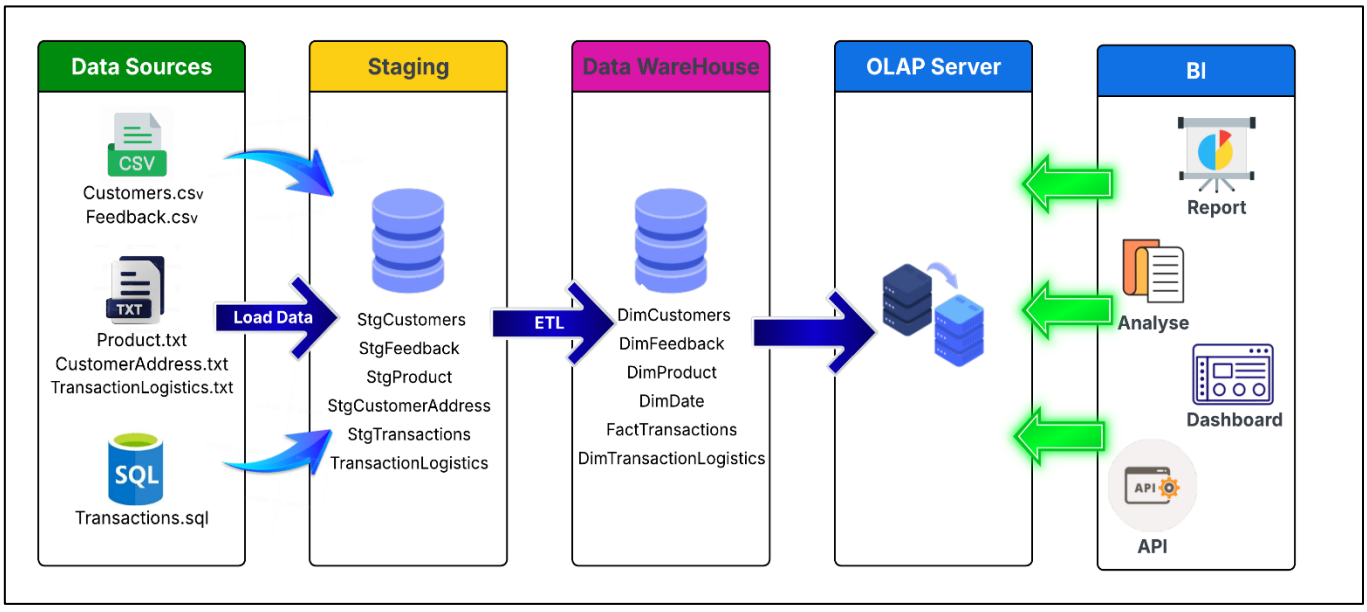
5. Transaction Table

	COLUMN_NAME	DATA_TYPE	Length	IS_NULLABLE
1	TransactionID	varchar	50	YES
2	CustomerID	varchar	50	YES
3	LastPurchaseDate	date	NULL	YES
4	TotalPurchases	int	NULL	YES
5	AmountSpent	decimal	NULL	YES

6. Transaction Logistics

	COLUMN_NAME	DATA_TYPE	Length	IS_NULLABLE
1	TransactionID	varchar	50	YES
2	ShippingMethod	varchar	50	YES
3	PaymentMethod	varchar	50	YES
4	OrderStatus	varchar	50	YES

Solution Architecture



Solution Architecture

Component	Description
Data Sources	- **CSV Files**: Contain Customer, Product, and Feedback data. - **SQL Server DB**: Stores transactional and logistics data. Simulates a mixed-source operational environment.
ETL Layer (SSIS)	- Extracts data from heterogeneous sources. - Transforms: cleans data, derives new fields, handles Slowly Changing Dimensions (SCD). - Loads data into dimension and fact tables.
Data Warehouse	- Follows a **star schema**. - Includes fact and dimension tables. - Supports aggregation, drill-down, and OLAP queries.
OLAP Layer	- Multidimensional cubes (SSAS) support slicing, dicing, and pivot-based analysis. - Enhances performance for complex analytical queries.
BI Reporting Layer	- **Power BI / SSRS** used to visualize KPIS, trends, and operational insights. - Provides dashboards for decision makers.

Data warehouse design & development

1. Design

A **Star Schema** was selected due to its simplicity and performance benefits for querying and reporting. It consists of one central **Fact Table** surrounded by **Dimension Tables**.

The database design used the star schema model with these arrangements:
Order data undergoes storage in the FactOrder table to store metrics.

Dimension Table :

- DimCustomer
- DimProduct
- DimDate
- DimFeedback
- DimTransactionLogistics

Fact Table:

- FactTransaction

Slowly Changing Dimension

DimCustomer -

- Hierarchical Dimensions - Address / City / Country / State / Zipcode
- Changing Attributes - CustomerSegment / Phone

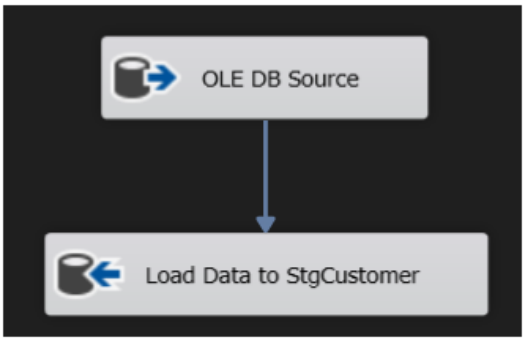
The model enables fast aggregation operations alongside quick data selection through business-specific attributes. The integration between Primary and foreign keys works to ensure data refers to the correct entities. The database schema execution utilized SQL Server scripts.

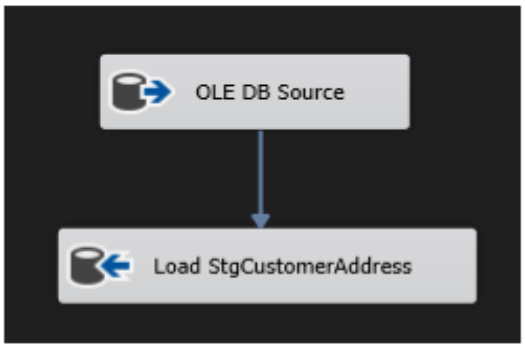
2. ETL Development

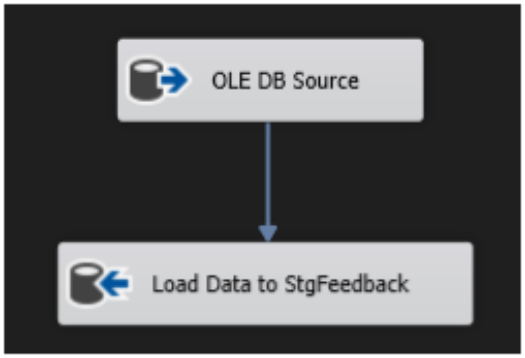
ETL was developed using SQL Server Integration Services (SSIS). Key tasks included:

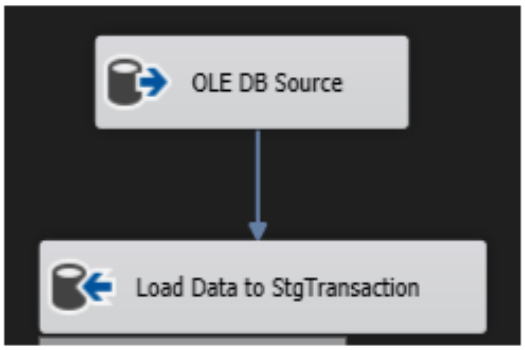
- **Flat File Source**: Loaded CSV data.
- **OLE DB Source**: Extracted data from SQL tables.
- **Derived Column**: Used to handle NULL values and standardize missing entries.
- **Lookup & Conditional Split**: Validated data consistency.
- **OLE DB Destination**: Loaded data into dimensional tables.
  Detailed logging, error handling, and transformation logic were implemented for clean data flow.

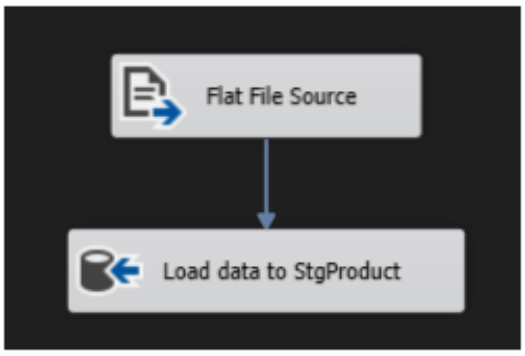
1. Data Extraction & Load into Staging tables

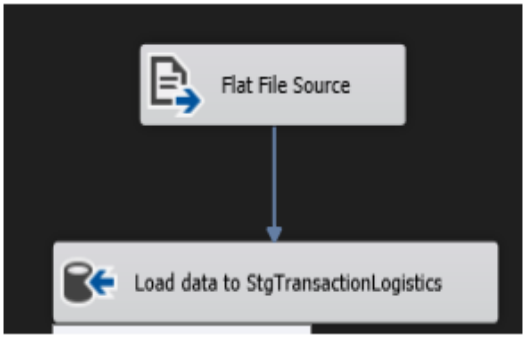




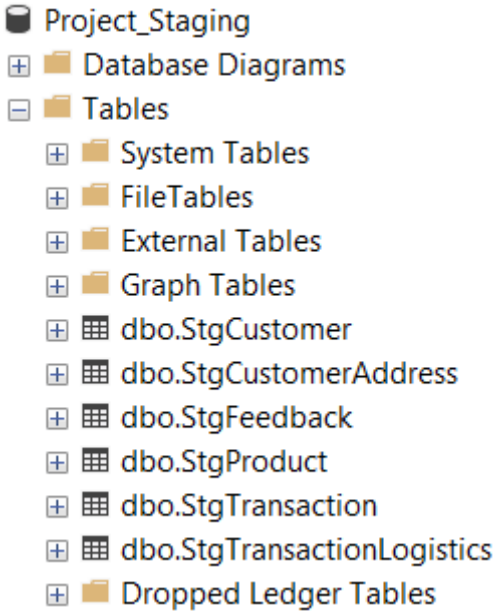






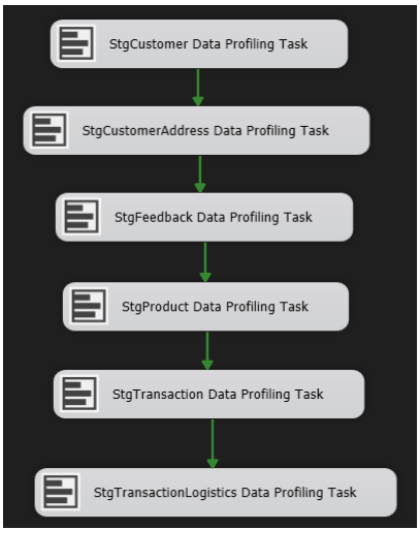


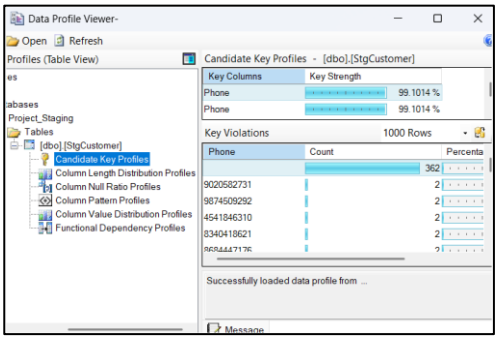
Implement staging tables to temporarily hold raw extracted data.

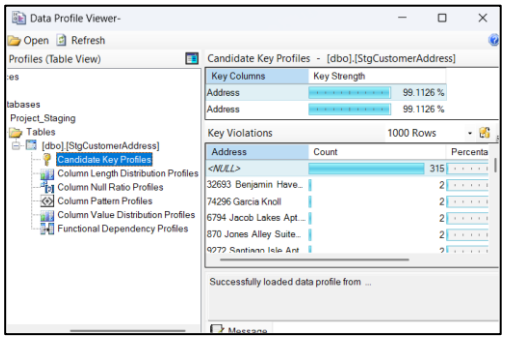


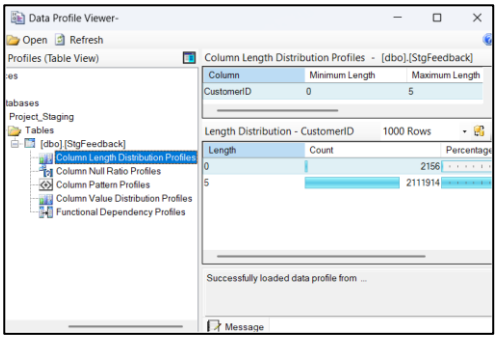
2. Data Profiling

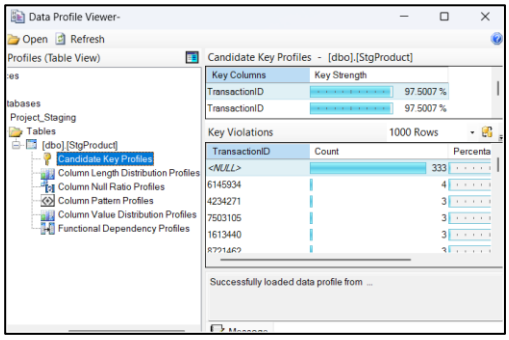
Data Profiling provides the means of analyzing large amount of data using different kind of processes. In this step, null values, repeated values and quality of the data is checked.

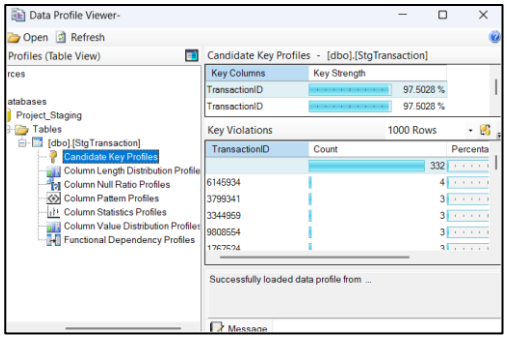


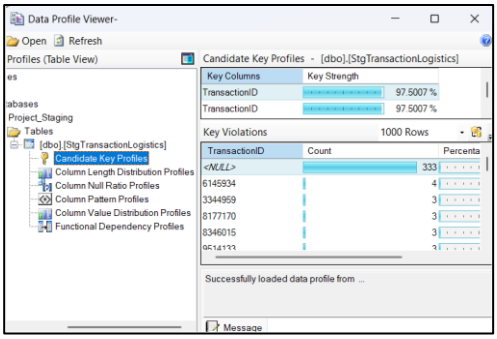




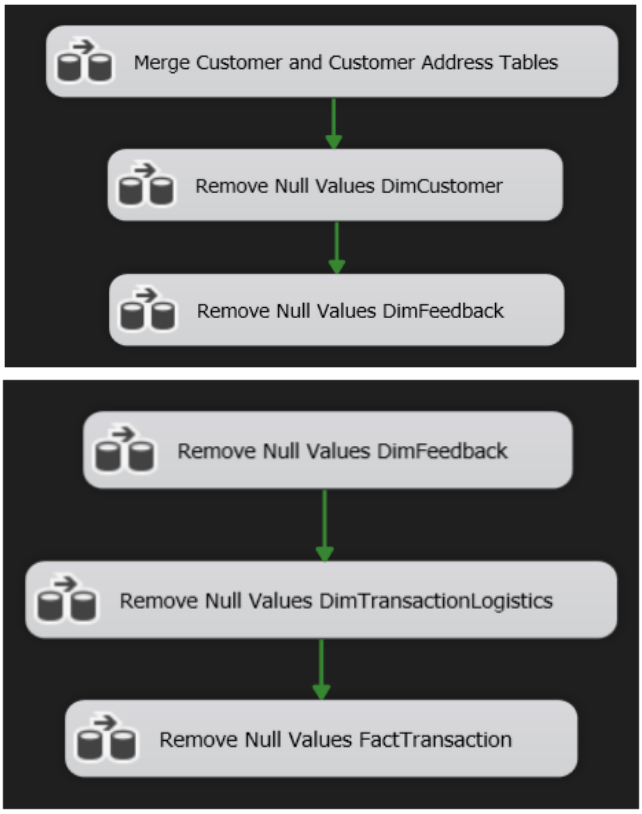




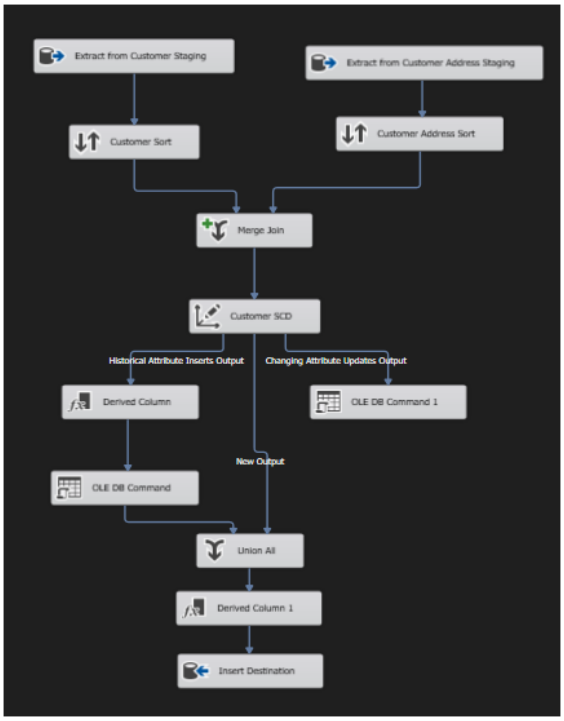




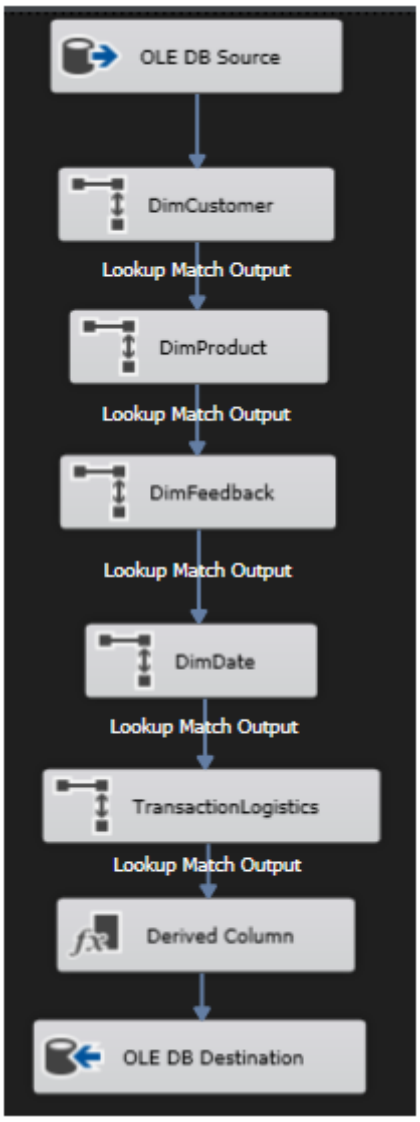
3. **Clean and transform data using Derived Columns.**
4. **Handle missing/null values by replacing them with a default.**
5. **Load transformed data into dimensional and fact tables in the data warehouse.**



6. **Apply Slowly Changing Dimension (SCD) logic where necessary.**



7. **Extend the fact table with accumulating snapshot columns for transaction tracking.**



❖ **Database Tables**

DimCustomer

	CustomerSK	CustomerID	Name	Email	Phone	Age	Gender	Income	CustomerSegment	Address
1	1		Jessica Shah	Joseph8@gmail.com	8151638900	37	Female	Low	Premium	Unknown
2	2	10000	Robert Cook	Caroline60@gmail.com	8428883216	64	Female	Low	Regular	136 Perkins Street
3	3	10001	Rebecca Lee	Michael61@gmail.com	5831371332	46	Male	Low	Regular	50931 Wilson Lodge
4	4	10002	Regina Dickson	Rita67@gmail.com	8342574825	46	Female	High	Premium	56489 Clark Forks
5	5	10003	Natalie Gonzalez	Ray27@gmail.com	7112898015	26	Male	High	Regular	79113 Jarvis Ridge
6	6	10004	William Orr	Michelle65@gmail.com	5661361904	23	Male	High	Regular	03001 Nelson Common
7	7	10005	Theresa Sheppard	Carla31@gmail.com	6238294350	20	Male	Medium	Regular	23756 Green Junction Apt. 328
8	8	10006	John Nelson	Elizabeth4@gmail.com	8216354167	46	Male	Low	Premium	59095 Long Radial
9	9	10007	Scott Carson	Kristine52@gmail.com	1288003188	65	Female	Low	Regular	243 Rebecca Loop
10	10	10008	Amanda Williams	Joshua35@gmail.com	7460772065	46	Male	Low	Regular	2352 Michael Locks
11	11	10009	Kimberly Robinson	Stephanie50@gmail.com	5181015344	32	Female	Medium	Regular	976 Lisa Shoal Apt. 741
12	12	10010	Juan Navarro	Steven56@gmail.com	1815801639	26	Male	High	New	236 Burton Plaza
13	13	10011	Amanda Collins	Daniel76@gmail.com	8497778385	63	Male	High	Regular	2552 Davis Circles Suite 958
14	14	10012	Alyssa Mcconnell	David58@gmail.com	8129098871	23	Female	High	Regular	1359 Douglas Wells
15	15	10014	Eric Johnson	Allison73@gmail.com	7258302571	21	Female	Medium	New	8333 White Union Apt. 108
16	16	10015	Holly Santos	Joseph51@gmail.com	1143516552	60	Male	Low	New	626 Hooper Inlet Suite 331
17	17	10016	Jesus Crane	Robert2@gmail.com	5078894303	34	Female	High	Premium	8809 Amber Course

City	State	Zipcode	Country	StartDate	EndDate	InsertDate	ModifiedDate
Unknown	Unknown	00000	Unknown	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Phoenix	North Carolina	28468	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Melbourne	New South Wales	21143	Australia	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Jacksonville	Oregon	97146	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
San Francisco	Maine	29266	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Boston	Georgia	51868	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Chicago	Connecticut	10132	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Regina	Ontario	62986	Canada	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Columbus	New Hampshire	03608	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Las Vegas	West Virginia	25747	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Launceston	New South Wales	87013	Australia	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Cleveland	Arkansas	71642	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
St. John's	Ontario	63567	Canada	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Boston	Georgia	84142	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Leicester	England	13761	UK	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Mesa	New Jersey	07918	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430
Chicago	Connecticut	36036	USA	2025-05-01 11:39:40.000	9999-12-31 00:00:00.000	2025-05-01 11:40:01.430	2025-05-01 11:40:01.430

DimDate

	DateKey	FullDate	Day	Month	Year	DayOfWeek	DayName	MonthName	Quarter	IsWeekend	WeekOfYear
1	20100101	2010-01-01	1	1	2010	6	Friday	January	1	0	1
2	20100102	2010-01-02	2	1	2010	7	Saturday	January	1	1	1
3	20100103	2010-01-03	3	1	2010	1	Sunday	January	1	1	2
4	20100104	2010-01-04	4	1	2010	2	Monday	January	1	0	2
5	20100105	2010-01-05	5	1	2010	3	Tuesday	January	1	0	2
6	20100106	2010-01-06	6	1	2010	4	Wednesday	January	1	0	2
7	20100107	2010-01-07	7	1	2010	5	Thursday	January	1	0	2
8	20100108	2010-01-08	8	1	2010	6	Friday	January	1	0	2
9	20100109	2010-01-09	9	1	2010	7	Saturday	January	1	1	2
10	20100110	2010-01-10	10	1	2010	1	Sunday	January	1	1	3
11	20100111	2010-01-11	11	1	2010	2	Monday	January	1	0	3
12	20100112	2010-01-12	12	1	2010	3	Tuesday	January	1	0	3
13	20100113	2010-01-13	13	1	2010	4	Wednesday	January	1	0	3
14	20100114	2010-01-14	14	1	2010	5	Thursday	January	1	0	3

DimFeedback

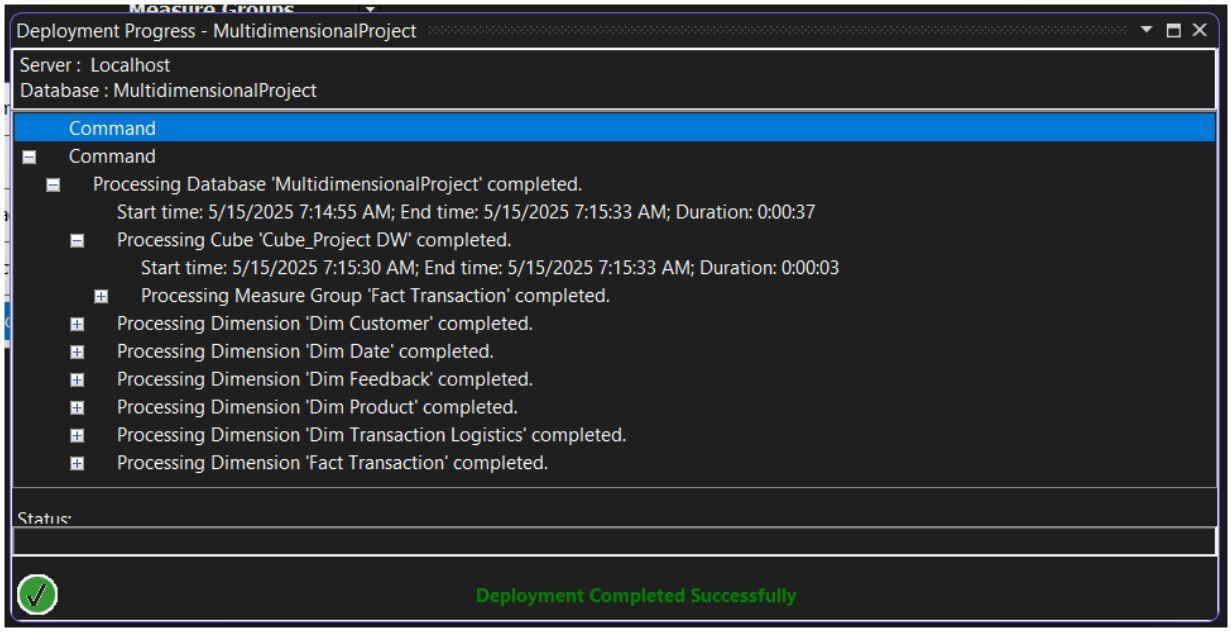
	FeedbackSK	TransactionID	CustomerID	Feedback	InsertDate	ModifiedDate
1	1	1443012	32766	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
2	2	3817065	82907	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
3	3	1683941	21724	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
4	4	3991832	92131	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
5	5	6705147	38447	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
6	6	3076394	98504	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
7	7	3475753	51468	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
8	8	9802103	90536	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
9	9	9684608	76962	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
10	10	7113548	35580	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
11	11	2954060	80397	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
12	12	4170673	75385	Good	2025-05-01 19:33:24.770	2025-05-01 19:33:24.770
13	13	4817675	90300	Good	2025-05-01 19:33:24.717	2025-05-01 19:33:24.717
14	14	9079779	30128	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687
15	15	9969397	12106	Good	2025-05-01 19:33:24.687	2025-05-01 19:33:24.687

8. Accumulating Fact Table with Transaction Duration

The final output is the FactTransaction table, which contains accumulated values and consolidated data from multiple source tables, providing a comprehensive view of transactions, customer details, and transaction metrics.

	TransactionSK	CustomerSK	ProductSK	DateSK	LogisticsSK	FeedbackSK	TotalPurchases	AmountSpent	TransactionID	InsertDate	ModifiedDate
1	1	56975	113519	20100928	113429	293496	7	1829.66	7523717	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
2	2	53522	113520	20100921	113430	293497	10	4015.95	9040009	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
3	3	31548	113521	20101122	113431	293498	5	2065.46	5001332	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
4	4	38628	113522	20100129	113432	293499	8	2141.47	3938104	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
5	5	29699	113523	20101203	113433	293500	4	1997.88	2279825	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
6	6	38281	113524	20100731	113434	293501	3	1368.20	8337964	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
7	7	41641	113525	20100917	113435	293502	6	87.75	7169735	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
8	8	77668	113526	20100113	113436	293503	1	449.66	8551076	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
9	9	55225	113527	20100312	113437	293504	2	868.98	3151632	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
10	10	39732	113528	20100912	113438	293505	1	151.79	2618710	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
11	11	5909	113529	20100711	113439	293506	5	655.63	2988065	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
12	12	10972	113530	20100905	113440	293507	6	277.92	1838553	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
13	13	4326	113531	20101106	113441	293508	8	80.88	7906896	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
14	14	31066	113532	20100412	113442	293509	1	250.65	5511620	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
15	15	84371	96365	20100322	113443	276285	10	3335.54	1771009	2025-05-15 01:32:38.207	2025-05-15 01:32:38.207
16	16	25892	113534	20100511	113444	293511	7	2526.88	3265804	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
17	17	20633	113535	20100502	113445	293512	5	1784.14	2313030	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920
18	18	76273	113536	20101203	113446	293513	4	767.68	7725514	2025-05-14 22:44:04.920	2025-05-14 22:44:04.920

Cube Deployment



***