Sri Lanka Institute of Information Technology

IT3021 - Data Warehousing and Business Intelligence
Assignment 1: Final Report
Name: Widyasekara S.P
Student ID: IT22132482

Dataset Selection

The project used a dataset based on an Online Retail Store scenario. The dataset includes customer information, product details, order records, and order completion timestamps. The dataset is semi-realistic and derived from transactional data resembling a retail business. The goal was to convert OLTP-style raw data into a dimensional data warehouse for reporting and business intelligence.
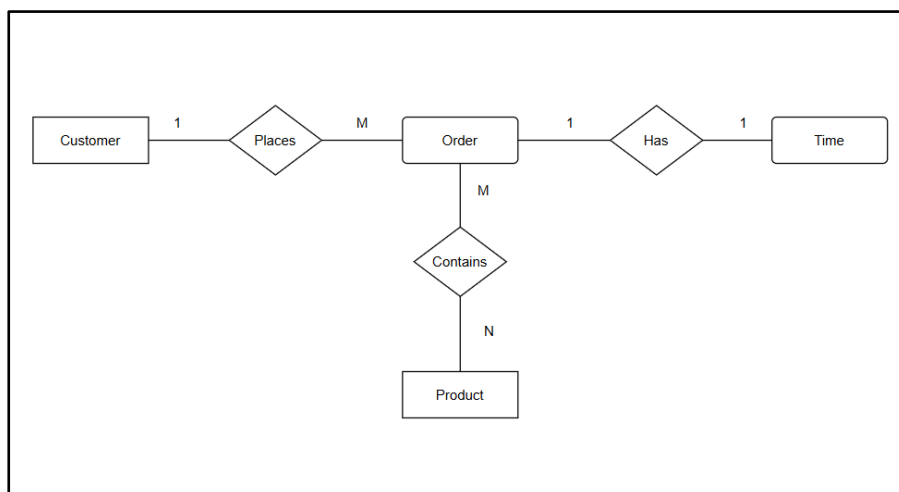
Tables Used:

- **Customer**: CustomerID, Country

- **Product**: ProductID, Description, UnitPrice

- **Orders**: OrderID, CustomerID, InvoiceDate (loaded to SQL Server)

- **OrderDetails**: OrderID, ProductID, Quantity, UnitPrice

Fact Table:

- **FactSales**: Captures metrics for each sale along with timestamps

ER Diagram

Preparation of Data Sources

To simulate a real-world scenario, multiple data source types were used and connected to SSIS:
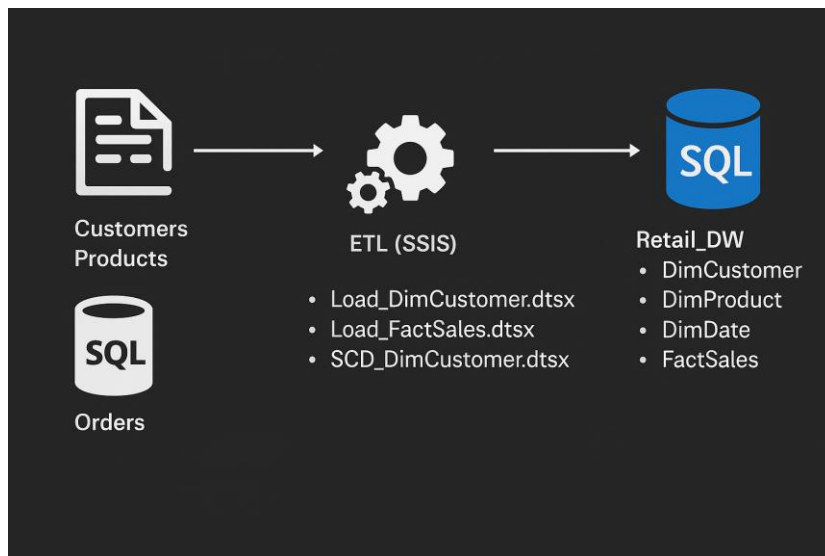
- **Customer**
    - Format: CSV
    - Connected via Flat File Connection Manager

- **Product**
    - Format: CSV
    - Connected via Flat File Connection Manager

- **Orders**
    - Format: SQL Table (imported via Import Wizard)
    - Connected via OLE DB Source in SSIS

- **OrderDetails**
    - Format: CSV
    - Connected via Flat File Source

- **CompleteTimes**
    - Format: CSV (used to update transaction complete times)
    - Connected via Flat File Connection Manager

Appropriate connection managers were configured for each format, including:

- First row as column names

- Column delimiter and text qualifier settings

- Proper data type configuration

Solution Architecture Design

This project used ETL flow in SSIS to load and transform data from multiple sources into a star-schema dimensional model hosted in a SQL Server database.

Data Sources:

- Customers.csv

- Products.csv

- OrderDetails.csv

- Orders (SQL Table)

- CompleteTimes.csv

ETL Process:

- Built using SQL Server Integration Services

- Extracted, cleaned, transformed, and loaded data

- Applied lookups, derived columns, sort and merge joins

- Loaded into data warehouse

Data Warehouse:

- Database: Retail_DW

- Dimension Tables: DimCustomer, DimProduct, DimDate

- Fact Table: FactSales

Data Warehouse Design and Development
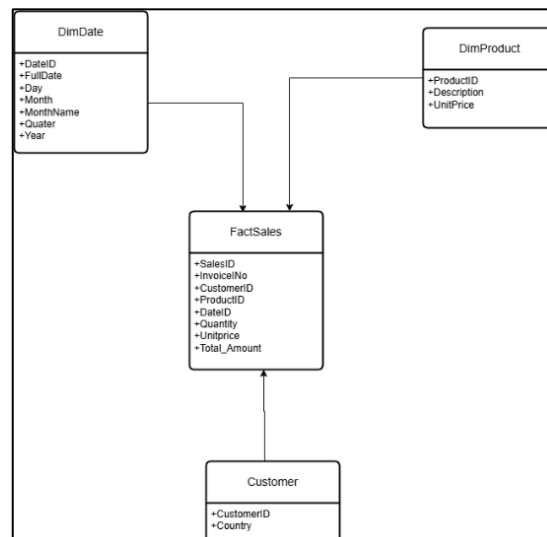
The Retail data warehouse follows a **Star Schema** consisting of:

- **DimCustomer**: CustomerID, Country, SCD tracking fields

- **DimProduct**: ProductID, Description, UnitPrice

- **DimDate**: DateID, FullDate, Year, Month, etc.

- **FactSales**: SalesID, CustomerID (FK), ProductID (FK), DateID (FK), Quantity, UnitPrice, TotalAmount, accm_txn_create_time, accm_txn_complete_time, txn_process_time_hours

These tables are connected via surrogate keys to facilitate fast querying and historical analysis.

ETL Development

Developed using SSIS with the following packages:

- **Load_DimCustomer.dtsx**

- **Load_DimProduct.dtsx**

- **Load_FactSales.dtsx**

- **SCD_DimCustomer.dtsx**
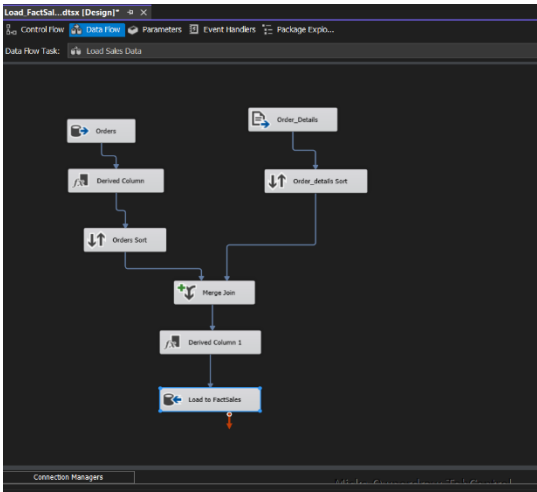
- **Update_FactSales_CompleteTime.dtsx**

Key Transformations:

- **Flat File Source**: CSV imports for Customer, Product, OrderDetails, Complete Times

- **OLE DB Source**: SQL Server table Orders

- **Derived Column**: Calculate TotalAmount, DateID, and timestamps

- **Sort + Merge Join**: Join OrderDetails with Orders

- **Lookup**: Map CustomerID, ProductID, DateID

- **Data Conversion**: Fix type mismatches

- **OLE DB Command**: Update FactSales completion times

Load_DimCustomer.dtsx

Load_DimProduct.dtsx

Load_FactSales.dtsx

SCD_DimCustomer.dtsx

Slowly Changing Dimension (SCD)

Implemented **SCD Type 2** on DimCustomer for the Country column. Tracked changes using:

- StartDate, EndDate, IsCurrent

Configured using SSIS SCD Wizard with business key = CustomerID.

Accumulating Fact Table

FactSales includes:

- accm_txn_create_time: Captured via Derived Column (GETDATE())

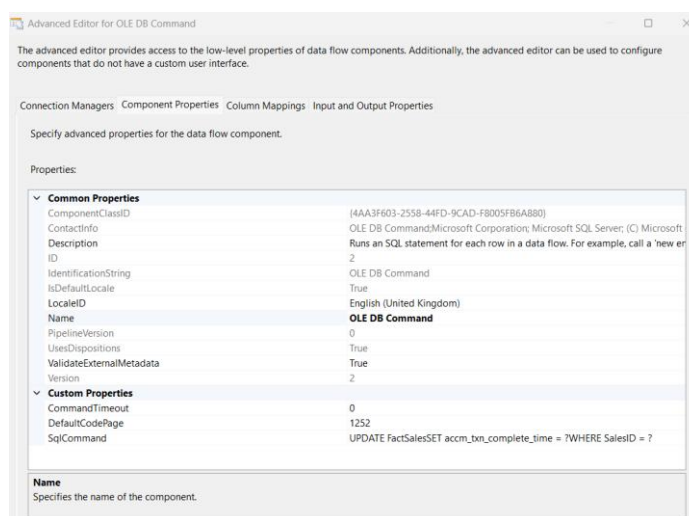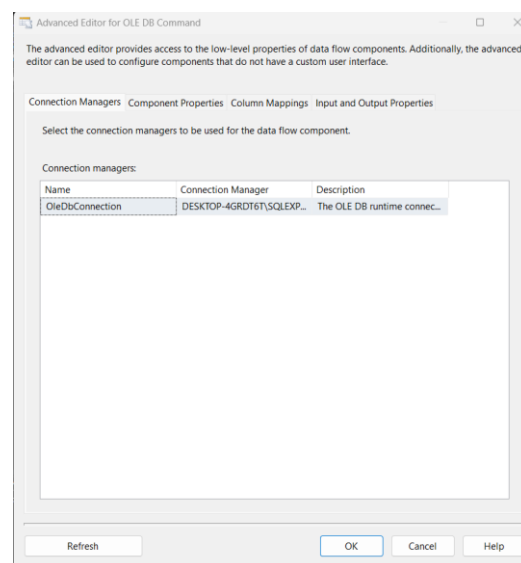- accm_txn_complete_time: Updated from CompleteTimes.csv via OLE DB Command

- txn_process_time_hours: Calculated as difference between complete and create time using DATEDIFF in Derived Column
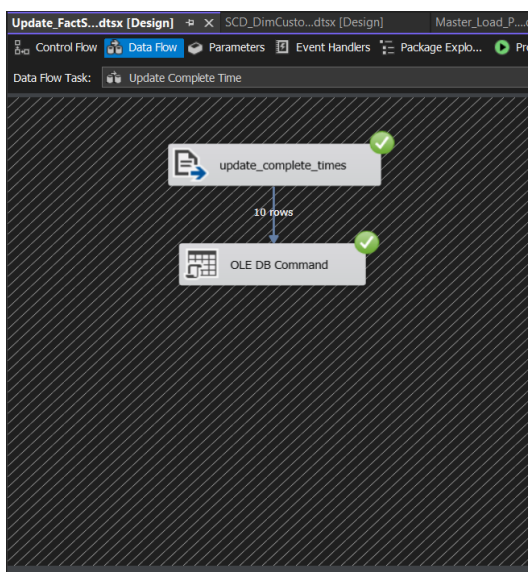
Data Validation

Validation was performed via SQL queries:

- Checked for NULLs and duplicates

- Confirmed correct mappings of foreign keys

- Verified calculated fields such as TotalAmount and DateID

Validation Queries:

SQLQuery2.sql - DE...4GRDT6T\Savi (59))* + × SQLQuery1.sql - DE...4GRDT6T\S

`SELECT TOP 10 * FROM DimCustomer;`

100 %

Results | Messages

	CustomerID	Country	StartDate	EndDate	IsCurrent	CustomerSK
1	12346	United Kingdom	NULL	NULL	NULL	1
2	12347	Iceland	NULL	NULL	NULL	2
3	12348	Finland	NULL	NULL	NULL	3
4	12349	Italy	NULL	NULL	NULL	4
5	12350	Norway	NULL	NULL	NULL	5
6	12352	Norway	NULL	NULL	NULL	6
7	12353	Bahrain	NULL	NULL	NULL	7
8	12354	Spain	NULL	NULL	NULL	8
9	12355	Bahrain	NULL	NULL	NULL	9
10	12356	Portugal	NULL	NULL	NULL	10

DimProduct.sql - D...4GRDT6T\Savi (55))* + × SQLQuery2.sql - D

`select top 10 * from DimProduct`

100 %

Results | Messages

	ProductID	Description	UnitPrice
1	10002	INFLATABLE POLITICAL GLOBE	0.85
2	10080	GROOVY CACTUS INFLATABLE	0.85
3	10120	DOGGY RUBBER	0.21
4	10123C	HEARTS WRAPPING TAPE	0.65
5	10124A	SPOTS ON RED BOOKCOVER TAPE	0.42
6	10124G	ARMY CAMO BOOKCOVER TAPE	0.42
7	10125	MINI FUNKY DESIGN TAPES	0.85
8	10133	COLOURING PENCILS BROWN TUBE	0.85
9	10135	COLOURING PENCILS BROWN TUBE	1.25
10	11001	ASSTD DESIGN RACING CAR PEN	1.69

```
SQLQuery1.sql - DE...4GRDT6T\Savi (58))*  ⌐ ×
    SELECT TOP 10 * FROM DimDate;
```

100 % ▾ ◂

Results Messages

	DateID	FullDate	Day	Month	MonthName	Quarter	Year
1	20101201	2010-12-01	1	12	December	4	2010
2	20101202	2010-12-02	2	12	December	4	2010
3	20101203	2010-12-03	3	12	December	4	2010
4	20101205	2010-12-05	5	12	December	4	2010
5	20101206	2010-12-06	6	12	December	4	2010
6	20101207	2010-12-07	7	12	December	4	2010
7	20101208	2010-12-08	8	12	December	4	2010
8	20101209	2010-12-09	9	12	December	4	2010
9	20101210	2010-12-10	10	12	December	4	2010
10	20101212	2010-12-12	12	12	December	4	2010


```
SQLQuery1.sql - DE...4GRDT6T\Savi (58))*  ⌐ ×
    SELECT TOP 10 * FROM FactSales;
```

100 % ▾ ◂

Results Messages

	SalesID	InvoiceNo	CustomerID	ProductID	DateID	Quantity	UnitPrice	TotalAmount	accm_txn_create_time	accm_txn_complete_time	txn_process_time_hours
1	1	536365	17850	84029G	20101201	6	3.39	20.34	2025-04-18 01:11:09.897	2025-04-20 01:11:00.000	48
2	2	536366	17850	22632	20101201	6	1.85	11.10	2025-04-18 01:11:09.897	2025-04-20 01:11:00.000	48
3	3	536367	13047	84879	20101201	32	1.69	54.08	2025-04-18 01:11:09.897	2025-04-22 01:11:00.000	96
4	4	536368	13047	22913	20101201	3	4.95	14.85	2025-04-18 01:11:09.897	2025-04-20 01:11:00.000	48
5	5	536369	13047	21756	20101201	3	5.95	17.85	2025-04-18 01:11:09.897	2025-04-23 01:11:00.000	120
6	6	536370	12583	21035	20101201	18	2.95	53.10	2025-04-18 01:11:09.897	2025-04-20 01:11:00.000	48
7	7	536371	13748	22086	20101201	80	2.55	204.00	2025-04-18 01:11:09.897	2025-04-20 01:11:00.000	48
8	8	536372	17850	22632	20101201	6	1.85	11.10	2025-04-18 01:11:09.897	2025-04-22 01:11:00.000	96
9	9	536373	17850	21068	20101201	6	1.06	6.36	2025-04-18 01:11:09.897	2025-04-21 01:11:00.000	72
10	10	536374	15100	21258	20101201	32	10.95	350.40	2025-04-18 01:11:09.897	2025-04-19 01:11:00.000	24

Conclusion

This project successfully demonstrated the implementation of a retail sales data warehouse using:

- Star Schema modeling

- ETL development with multiple data sources

- Slowly Changing Dimensions

- Accumulating fact table