KULLIYYAH OF INFORMATION & COMMUNICATION TECHNOLOGY

INFO 4311 DATA WAREHOUSING

END-OF-SEMESTER ASSESSMENT 1

SEMESTER 1, 2024/2025 SESSION

SECTION 2

NORTHWIND

PREPARED BY:

NAME	MATRIC NO.
Muhammad Hafiz Faruqi bin Md Saifuddin	2217241

LECTURER: ASSOC. PROF. DR. LILI MARZIANA BT. ABDULLAH

Part 1: Dataset Selection

Selected Dataset: Northwind

Dataset Description:

The Northwind database is a sample database that was created by Microsoft and used as the basis for their tutorials in a variety of database products for decades. The Northwind dataset simulates the operations of a global company that supplies products across various regions. The dataset is commonly used for database and analytics training and is structured to include key business areas such as customers, orders, products, and employees (*yugabyteDB,* 2024).

Dataset Source:

Originally crafted by Microsoft, the Northwind database has served as a sample database foundation for their instructional materials across a range of database products for numerous years (*Kaggle Northwind Database*, 2024). The Northwind database is publicly available and widely used in training environments for relational databases and analytics.

Key Tables and Attributes:

The Northwind dataset includes sample data for the following.

1. **Categories:** Product categories, such as beverages, dairy products, and grains.
2. **Customers:** Contains customer details like CustomerID, CompanyName, ContactName, Address, City, Country, etc.
3. **Employees:** Employee information, such as names, titles, and territories.
4. **Order_Details:** Contains line-item details for each order, such as OrderID, ProductID, Quantity, UnitPrice, Discount, etc.
5. **Suppliers:** Details of suppliers, such as company names and contact information.
6. **Products:** Contains product information such as ProductID, ProductName, SupplierID, CategoryID, UnitPrice, UnitsInStock, etc.
7. **Orders:** Contains details about each order, such as OrderID, CustomerID, EmployeeID, OrderDate, ShipVia, Freight, etc.

Analytical Opportunities

The Northwind dataset provides a rich source of data for various analytical purposes, such as:

- Understanding sales performance by product, region, and time.
- Identifying high-value customers and their purchase behaviours.
- Analysing supplier performance and dependency on key suppliers.
- Monitoring employee contributions to regional or overall sales.

Analytical Goals:

The focus of the data mart will be on **"Sales Analysis."** Specifically, the data mart will aim to answer the following analytical questions:

1. What are the top-selling products?
2. Which products generate the most revenue?
3. What are the sales trends across different periods (monthly, quarterly, yearly)?
4. Who are the highest-value customers based on purchase history?

Part 2: Dimensional Modelling

Bus Matrix

Data Warehouse Bus Matrix

Business Process	Fact Table	Granularity	Facts	Products	Customers	Orders	Date
Sales Analysis	FactOrderSales	One row represents an individual product sold in an order	Revenue, Quantity, Discount	X	X	X	X

Dimensional Modelling Workbook

FactOrderSales: contain transactional sales data

Table Name : FactOrderSales
Table Type : Fact
Table Description Fact table for one row of each product sold in an order

				Target						Source			
Column Name	Description	Unknown Member	Example Values	Data Type	Size	Key	FK To	Null	Default Valu	Source System	Source Table	Source Field Name	Source Data Type
SalesID	Surrogate primary key	-1	1,2,3	int		PK		N		DW			
DateKey	FK to DimDate	-1	20060106	int		FK	DimDate	N		DW	DimDate	DateKey	int
ProductKey	FK to DimProducts	-1	1,2,3	int		FK	DimProducts	N		Northwind_DB	Orders	ProductID	int
CustomerKey	FK to DimCustomers	-1	1,2,3	int		FK	DimCustomers	N		Northwind_DB	Orders	CustomerID	int
OrderKey	FK to DimOrders	-1	1,2,3	int		FK	DimOrders	N		Northwind_DB	Orders	OrderID	int
Revenue	Total revenue for the product in the order	-1	1250.50, 2240.44	money				N	0	Northwind_DB	Order Details	UnitPrice * Quantity * (1 - Discount)	money
Quantity	Quantity of the product sold	-1	50 , 200	int				N	0	Northwind_DB	Order Details	Quantity	int
Discount	Discount applied to the products	-1	0.15,0.20	float				N	0	Northwind_DB	Order Details	Discount	float

DimProducts : Contains attributes related to the products sold.

Table Name : DimProducts
Table Type : Dimension
Table Description : Details of the products

				Target						Source			
Column Name	Description	Unknown Member	Example Values	Data Type	Size	Key	FK To	Null	Default Value	Source System	Source Table	Source Field Name	Source Data Type
ProductKey	Surrogate primary key(unique identifier for each product)	-1	1, 2, 3, ...	int		PK		N		Derived			
ProductID	Business key from source system (aka natural key)	-1	101,102,103, ...	int				N		Northwind_DB	Products	ProductID	int
ProductName	Name of the product	Unknown	Apple, Ikura	nvarchar	255			N		Northwind_DB	Products	ProductName	varchar
QuantityPerUnit	Description of product's packaging	Unknown	10 boxes, 12 ml bottl	nvarchar	50			N		Northwind_DB	Products	QuantityPerUnit	varchar
UnitPrice	Price per unit of the product	-1	19.99, 249.99, 1.99	money				N	0	Northwind_DB	Products	UnitPrice	money
UnitsInStock	Quantity of the product in stock	-1	10,20,500	int	50			N	0	Northwind_DB	Products	UnitsInStock	int
UnitsOnOrder	Quantity of the product currently on order	-1	3,10,44	int	50			N	0	Northwind_DB	Products	UnitsOnOrder	int
Discontinued	Indicates if product is discontinued	0	1(TRUE), 0 (FALSE)	boolean				N	0	Northwind_DB	Products	Discontinued	boolean

DimCustomers: Contains customer-related details

Table Name :	DimCustomers
Table Type :	Dimension
Table Description	Details of Customers

				Target						Source			
Column Name	**Description**	**Unknown Member**	**Example Values**	**Data Type**	**Size**	**Key**	**FK To**	**Null**	**Default Value**	**Source System**	**Source Table**	**Source Field Name**	**Source Data Type**
CustomerKey	Surrogate primary key	-1	1, 2, 3, ...	int		PK		N		Derived			
CustomerID	Business key from source system (aka natural key)	-1	ALFKI,ANTON	int				N		Northwind_DB	Customers	Customer_ID	int
CompanyName	Name of the company	Unknown	Alfreds, Antonio	nvarchar	255			N		Northwind_DB	Customers	CompanyName	varchar
ContactName	Name of the primary contact person	Unknown	Maria, Anna	nvarchar	255			N		Northwind_DB	Customers	ContactName	varchar
ContactTitle	Title of the contact person	Unknown	Owner, Sales Mana	nvarchar	255			N		Northwind_DB	Customers	ContactTitle	varchar
Address	Address of the customer	Unknown	Forsterstr. 57	nvarchar	255			N		Northwind_DB	Customers	Address	varchar
City	City of customer located	Unknown	Berlin, Bern	nvarchar	100			N		Northwind_DB	Customers	City	varchar
Region	Region of the customer	Unknown	SP, DF, CA	nvarchar	50			N		Northwind_DB	Customers	Region	varchar
Country	Country of the customer	Unknown	Germany, Thailar	nvarchar	100			N		Northwind_DB	Customers	Country	varchar

DimOrders: Contains order-related information, linking to order-specific details (e.g., shipping method)

Table Name :	DimOrders
Table Type :	Dimension
Table Description	Details of orders

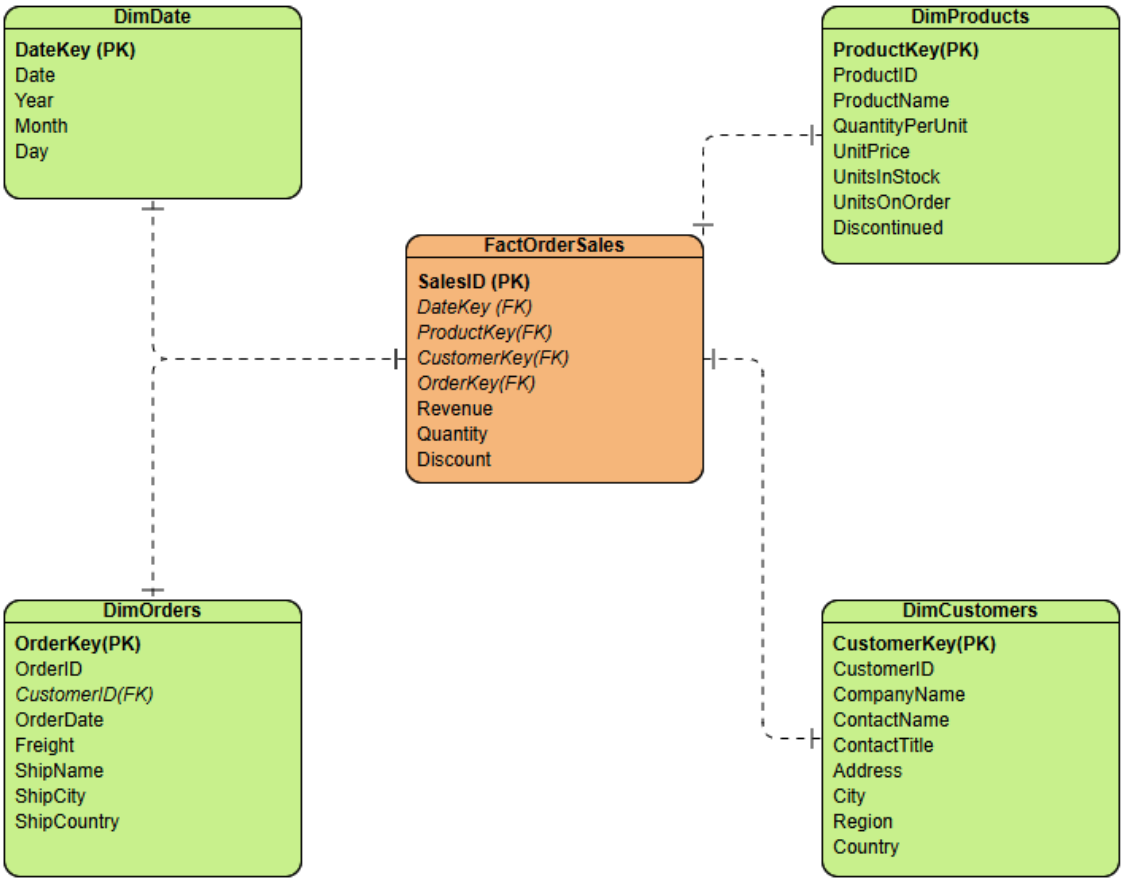
				Target						Source			
Column Name	**Description**	**Unknown Member**	**Example Value**	**Data Type**	**Size**	**Key**	**FK To**	**Null**	**Default Value**	**Source System**	**Source Table**	**Source Field Name**	**Source Data Type**
OrderKey	Surrogate primary key	-1	1, 2, 3, ...	int		PK		N		Derived			
OrderID	Business key from source system (aka natural key)	-1	10248; 10345	int				N		Northwind_DB	Orders	OrderID	int
CustomerID	References to the customer place the order	-1	ALFKI, ANTON	nvarchar	50	FK	DimCustomers	N		Northwind_DB	Orders	CustomerID	varchar
OrderDate	Date when order was placed	1/1/1900	3/5/1996	date				N		Northwind_DB	Orders	OrderDate	date
Freight	Freight cost associated with the order	0	32.33, 111.34	decimal	10,2			N		Northwind_DB	Orders	Freight	decimal
ShipName	Name of the recipient	Unknown	Alfred, Alfredo	nvarchar	255			N		Northwind_DB	Orders	ShipName	varchar
ShipCity	City of shipping address	Unknown	Berlin, Bern	nvarchar	100			N		Northwind_DB	Orders	ShipCity	varchar
ShipCountry	Country of shipping address	Unknown	Germany, Japar	nvarchar	100			N		Northwind_DB	Orders	ShipCountry	varchar

DimDate: Time-based dimension for analysis by date (e.g., year, month, day)

Table Name :	DimDate
Table Type :	Dimension
Table Description	Standard Date Dimension

				Target						Source			
Column Name	**Description**	**Unknown Member**	**Example Value**	**Data Type**	**Size**	**Key**	**FK To**	**Null**	**Default Value**	**Source System**	**Source Table**	**Source Field Name**	**Source Data Type**
DateKey	Surrogate primary key YYYYMMDD	-1	20060106	int		PK		N		Derived			
Date	Business key from source system (aka natural key)	-1	1/6/2006	date				N		Derived			date
Year	Year of the date	0	2006	int				N		Derived			int
Month	Month of the date	0	1	int				N		Derived			int
Day	Day part of the date	0	6	int				N		Derived			int

Star Schema Diagram (Visual Paradigm)



This design ensures that the Sales Analysis data mart can provide meaningful insights into product sales, customer behaviour, and time-based trends. The Star Schema provides a clear and efficient structure for reporting and analysis, while the chosen dimensions cover key aspects of the business process, enabling detailed and actionable insights for decision-making.

Justification for Sales Analysis Data Mart Design

The Sales Analysis data mart is designed to provide insights into the sales performance of products, with a focus on customer behaviour, order details, and time. The design follows a Star Schema to ensure simplicity, speed, and flexibility in querying. This structure uses a central Fact Table connected to four surrounding Dimension Tables to support detailed analysis of sales data.

The FactOrderSales table serves as the central fact table in the schema, containing key sales metrics such as Revenue, Quantity, and Discount. These facts are derived from individual sales transactions, where each row represents a product sold in a specific order. The Revenue is calculated as **UnitPrice * Quantity * (1 - Discount)**, enabling direct analysis of total sales for each product. The FactOrderSales table links to the dimension tables through foreign keys, allowing for aggregation and filtering based on product, customer, time, and order.

The granularity of the FactOrderSales table is set at the level of individual products sold in an order. This provides detailed data that can be aggregated at higher levels, such as total sales per product, customer, or time.

Dimension Tables

- **DimProducts**: Contains product details, such as names, prices, stock unit and product continuation. This dimension enables analysis of sales by product, helping to identify top sellers and performance across categories.
- **DimCustomers**: stores customer information, enabling insights into customer behaviour and purchase history. It supports the analysis of sales patterns by customer demographics (location, region, etc.).
- **DimOrders**: Holds order-related details, such as order dates and shipping methods. It allows for analysis of sales over time and helps track the efficiency of order processing.
- **DimDate**: Focuses on time, breaking down sales data by specific dates, months, and years. This dimension is essential for time-based analysis, such as comparing sales across different time periods.

Part 3: ETL Pipeline Design

ETL Rule in Dimensional Modelling Workbook

FactOrderSales

Table Name :	FactOrderSales				
Table Type :	Fact				
Table Description	Fact table for one row of each product sold in an order				
Column Name	**Description**	**Unknown Member**	**Example Values**	**SCD Type**	**ETL Rule**
SalesID	Surrogate primary key	-1	1,2,3		A surrogate primary key, auto-generated during the load process.
DateKey	Fk to DimDate	-1	20060106		Map the OrderDate from the source table Orders to the appropriate DimDate.
ProductKey	FK to DimProducts	-1	1,2,3		From Products table, map ProductID to DimP
CustomerKey	FK to DimCustomers	-1	1,2,3		rom Customers table, map CustomerID to Di
OrderKey	FK to DimOrders	-1	1,2,3		Calculation: hours worked * hourly rate
Revenue	Total revenue for the product in the order	-1	1250.50, 2240.44		Revenue = UnitPrice * Quantity * (1 - Discount)
Quantity	Quantity of the product sold	-1	50 , 200		Directly copied from OrderDetails.
Discount	Discount applied to the products	-1	0.15,0.20		Directly copied from OrderDetails.

DimProducts

Table Name :	DimProducts				
Table Type :	Dimension				
Table Description	Details of the products				
Column Name	**Description**	**Unknown Member**	**Example Values**	**SCD Type**	**ETL Rule**
ProductKey	Surrogate primary key(unique identifier for each product)	-1	1, 2, 3, ...		Auto generated
ProductID	Business key from source system (aka natural key)	-1	101,102,103, ...		Extract ProductID as-is
ProductName	Name of the product	Unknown	Apple, Ikura		Extract ProductName as-is
QuantityPerUnit	Description of product's packaging	Unknown	10 boxes, 12 ml bottles		Extract QuantityPerUnit as-is
UnitPrice	Price per unit of the product	-1	19.99, 249.99, 1.99		Extract UnitPrice as-is
UnitsInStock	Quantity of the product in stock	-1	10,20,500		Extract UnitsInStock as-is
UnitsOnOrder	Quantity of the product currently on order	-1	3,10,44		Extract UnitsOnOrder as-is
Discontinued	Indicates if product is discontinued	0	1(TRUE), 0 (FALSE)		Extract Discontinued as-is

DimCustomers

Table Name :	DimCustomers				
Table Type :	Dimension				
Table Description :	Details of Customers				
Column Name	**Description**	**Unknown Member**	**Example Values**	**SCD Type**	**ETL Rule**
CustomerKey	Surrogate primary key	-1	1, 2, 3, ...		Auto generated
CustomerID	Business key from source system (aka natural key)	-1	ALFKI,ANTON		Extract CustomerID as-is
CompanyName	Name of the company	Unknown	Alfreds, Antonio		Extract CompanyName as-is
ContactName	Name of the primary contact person	Unknown	Maria, Anna		Extract ContactName as-is
ContactTitle	Title of the contact person	Unknown	Owner, Sales Manager		Extract ContactTitle as-is
Address	Address of the customer	Unknown	Forsterstr. 57		Extract Address as-is
City	City of customer located	Unknown	Berlin, Bern		Extract City as-is
Region	Region of the customer	Unknown	SP, DF, CA		Extract Region as-is
Country	Country of the customer	Unknown	Germany, Thailand		Extract Country as-is

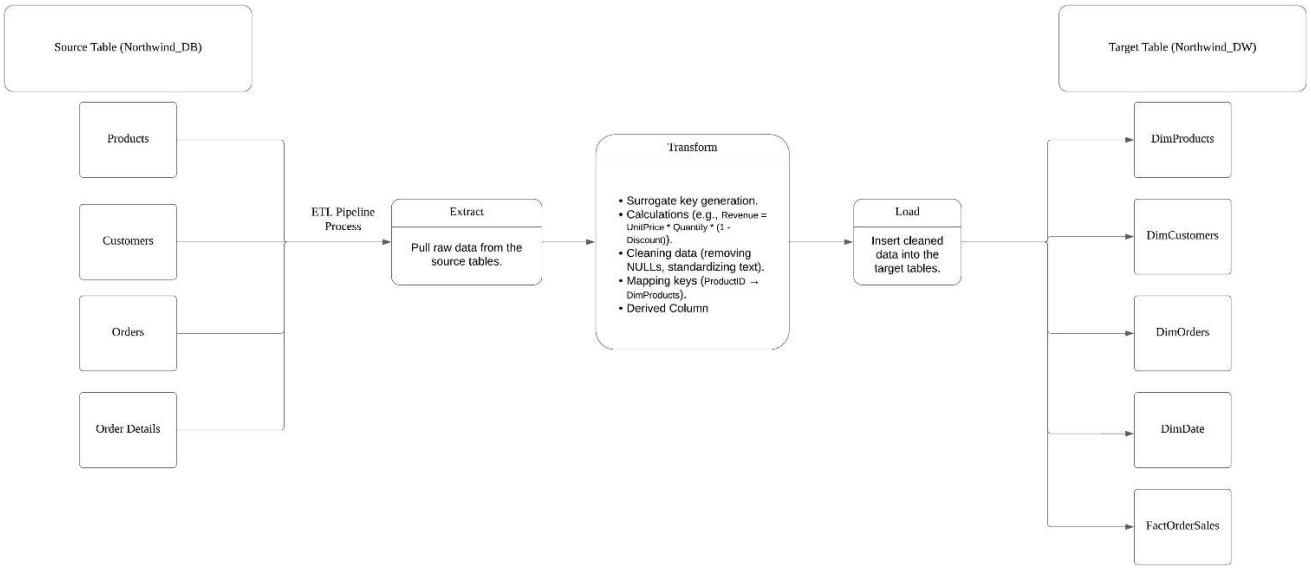
DimOrders

Table Name :	DimOrders				
Table Type :	Dimension				
Table Description :	Details of orders				
Column Name	**Description**	**Unknown Member**	**Example Values**	**SCD Type**	**ETL Rule**
OrderKey	Surrogate primary key	-1	1, 2, 3, ...		Auto generated
OrderID	Business key from source system (aka natural key)	-1	10248; 10345		Extract OrderID as-is
CustomerID	References to the customer place the order	-1	ALFKI, ANTON		Extract CustomerID as-is
OrderDate	Date when order was placed	1/1/1900	3/5/1996		Extract directly from Orders table.
Freight	Freight cost associated with the order	0	32.33, 111.34		Extract Freight as-is
ShipName	Name of the recipient	Unknown	Alfred, Alfredo		Extract ShipName as-is
ShipCity	City of shipping address	Unknown	Berlin, Bern		Extract ShipCity as-is
ShipCountry	Country of shipping address	Unknown	Germany, Japan		Extract ShipCountry as-is

DimDate

Table Name :	DimDate				
Table Type :	Dimension				
Table Description	Standard Date Dimension				
Column Name	**Description**	**Unknown Member**	**Example Value**	**SCD Type**	**ETL Rule**
DateKey	Surrogate primary key YYYYMMDD	-1	20060106		Convert Date in datetime to int of format YYYYMMDD
Date	Business key from source system (aka natural key)	-1	1/6/2006		Use Date as-is
Year	Year of the date	0	2006		Use function Year to extract Year
Month	Month of the date	0	1		Use function Month to extract Month
Day	Day part of the date	0	6		Use function Day to extract Day

ETL Plan Diagram Sources -ETL – Target (LucidChart)



Explanation of the ETL Design Choices

1.Extraction

Source System: Northwind_DB(providing data from tables like Orders, OrderDetails, Products, and Customers).. By focusing only on the necessary attributes, we minimize unnecessary data movement, improving ETL performance and reducing load times.

Key Design Choices:

- Only essential tables and attributes were extracted to ensure that the focus is on the data needed for the sales analysis.
- Orders and OrderDetails tables provide the transactional data for the FactOrderSales table, capturing key sales measures.
- Products and Customers tables are used to support the creation of dimension tables for analysing sales across different product categories and customer profiles.
- Filter the extracted data to avoid loading unnecessary records (e.g., exclude discontinued products).
- OLE DB Source was selected to efficiently extract data from the relational database.

2.Transformation

In the transformation phase, raw data is cleaned, standardized, and transformed to fit the dimensional model. This step is critical to ensure the data is ready for analysis. Key transformations included:

- **Data Conversion Transformation:** Ensure proper data types (e.g., convert dates to consistent formats, and handle data types like int, money, varchar).
- **Derived Column Transformation:** Calculated the Revenue measure in the fact table (UnitPrice * Quantity * (1 - Discount)) and for generating surrogate keys (e.g., ProductKey, CustomerKey) for the dimension tables.

- **Lookup Transformation:** Match keys from source data to the corresponding dimension (e.g., look up ProductID from the source dataset to find ProductKey in the DimProducts). This transformation helps link the fact table with dimension tables (using surrogate keys).
- **Conditional Split Transformation:** Excluded irrelevant data, such as discontinued products, from dimensions. Standardize attributes like address formats or customer names (data cleaning).

3.Loading

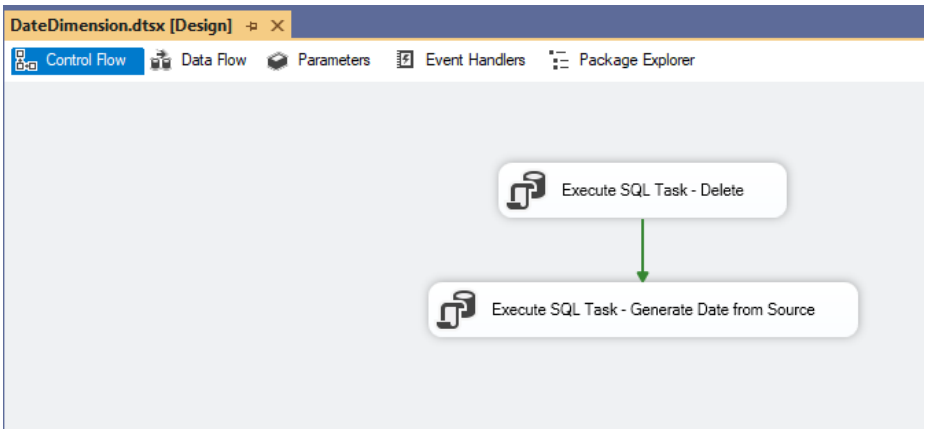
The transformed data was loaded into the dimensional model, adhering to the star schema design. Surrogate keys were generated for dimension tables, and foreign keys in the fact table were mapped to these surrogate keys. Fact and dimension tables were loaded incrementally to handle future updates efficiently.

Key Design Choices:

- A star schema was chosen because of its simplicity, efficiency, and ease of use for analytical queries. The fact table contains the measurable data (e.g., Revenue, Quantity, and Discount), while the dimension tables store descriptive attributes.
- Data was loaded incrementally to efficiently handle future updates and additions to the dataset. This ensures that new data is appended without requiring a full reload.
- The Destination Transformation in SSIS is used to load data into the fact and dimension tables, with foreign keys in the fact table pointing to the surrogate keys in the dimension tables.
- Start by loading the DimProducts, DimCustomers, DimOrders, and DimDate tables with their corresponding surrogate keys.
- Load FactOrderSales last, ensuring all foreign keys (ProductKey, CustomerKey, OrderKey, DateKey) are correctly mapped from the dimension tables.

Data Flow Diagram - Strategy for ETL Pipeline Package (SSIS)

DateDimension.dtsx



1. Execute SQL Task

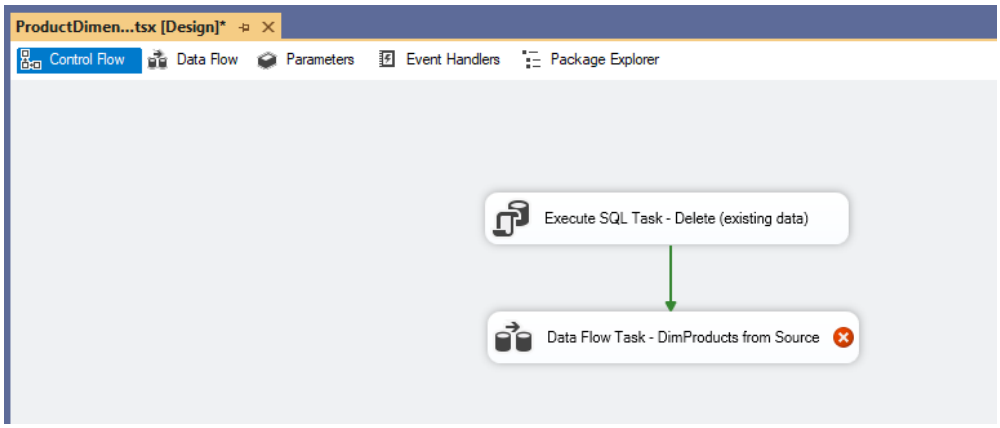
- **Purpose**: Remove existing records in DimDate to ensure clean loading of new dates.

2. Execute SQL Task – Generate Dates

- **Purpose**: Populate the DimDate table with a range of dates with SQL Script.

The chosen strategy leverages an Execute SQL Task to generate and populate the DimDate table.

ProductDimension.dtsx

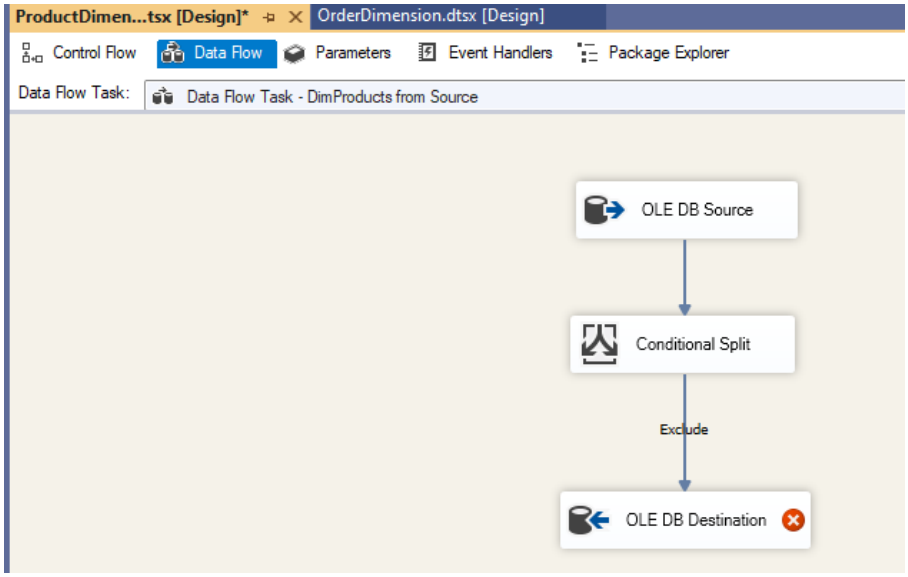


1. Execute SQL Task

- **Purpose**: Remove existing records in DimProducts to ensure a clean slate before loading updated data.

2. Data Flow Task

- This task extracts data from the Products table in Northwind_DB, transforms it, and loads it into DimProducts in Northwind_DW.



OLE DB Source

- **Purpose**: Extract product data from the Products table in the Northwind_DB database.

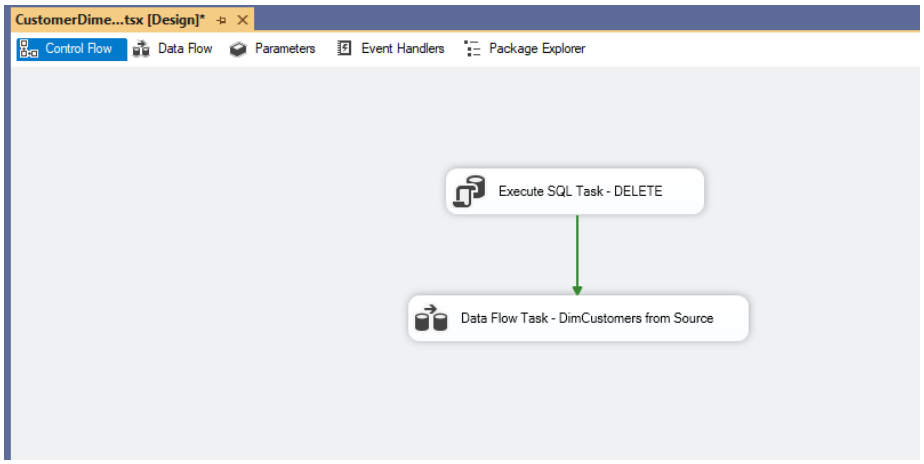
Conditional Split Transformation

- **Purpose**: Exclude records for discontinued products.
- **Condition:** Discontinued == FALSE

OLE DB Destination

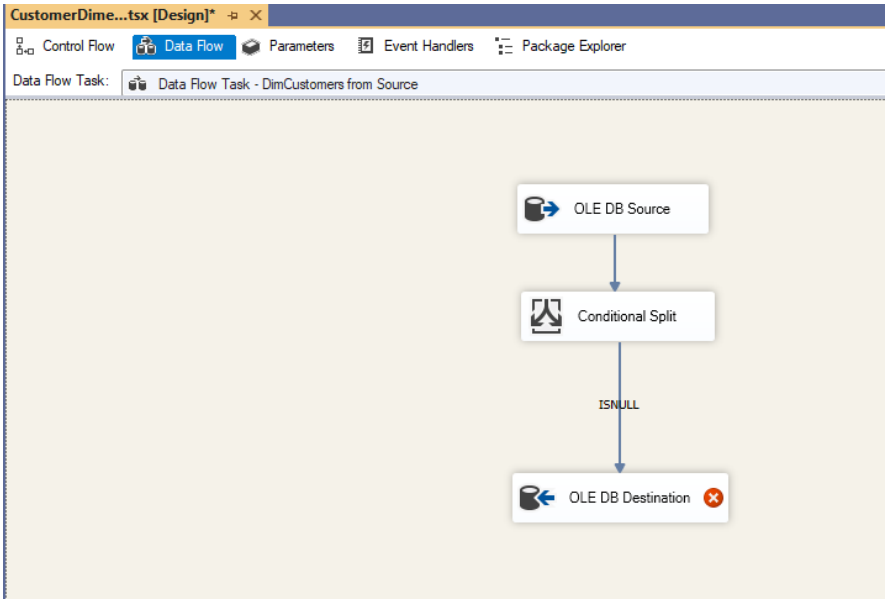
- **Purpose**: Load the transformed product data into the DimProducts table in Northwind_DW.

CustomerDimension.dtsx



1. Execute SQL Task

- **Purpose**: Clear the existing data from the DimCustomers table in the Northwind_DW database to prepare for a new data load.



2. Data Flow Task

This task manages the data extraction, transformation, and loading for the DimCustomers table.

1. OLE DB Source

   Purpose: Extract customer data from the Customers table in the Northwind_DB database.

2. Conditional Split Transformation

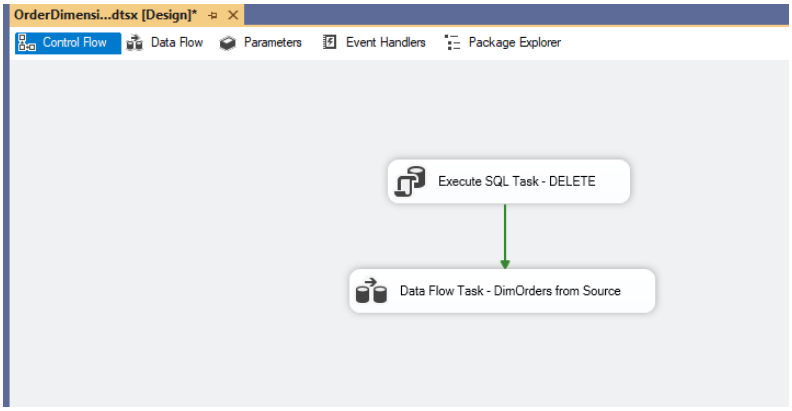
   Purpose: Exclude irrelevant data, such as customers with incomplete address information (e.g., missing Country or City).

   Condition : ISNULL(Country) || ISNULL(City) ? "Invalid" : "Valid"

3. OLE DB Destination

   Purpose: Load the transformed and filtered data into the DimCustomers table in the Northwind_DW database.

OrderDimension.dtsx

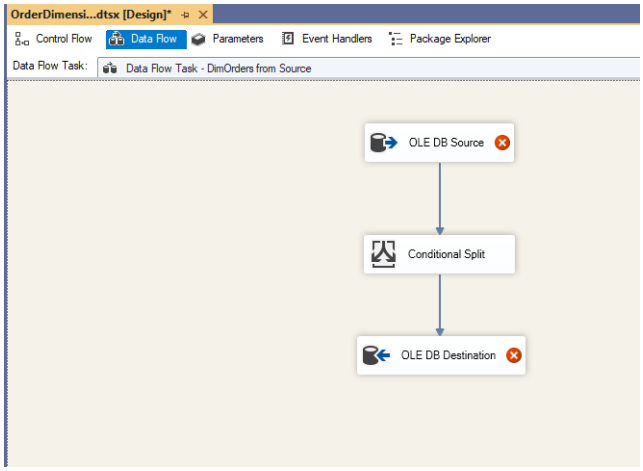


1. Execute SQL Task

- **Purpose**: Clear the existing data from the DimOrders table in the Northwind_DW database to ensure no duplicate or outdated data exists.

2. Data Flow Task

This task extracts, transforms, and loads data for the DimOrders table.



OLE DB Source

- **Purpose**: Extract order data from the Orders table in the Northwind_DB database.

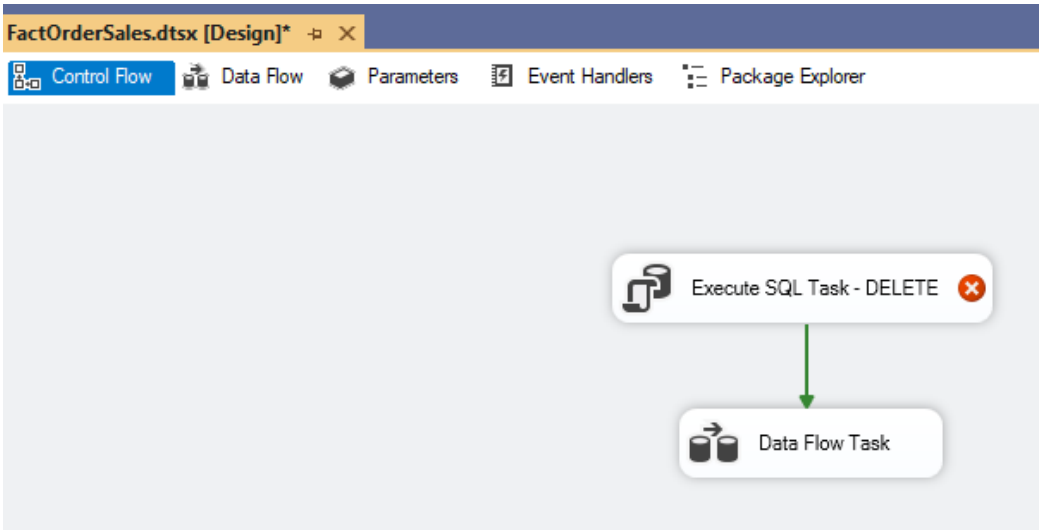
Conditional Split Transformation

- **Purpose**: Exclude orders with invalid data, such as missing CustomerID or OrderDate.
- **Condition:** ISNULL(CustomerID) || ISNULL(OrderDate) ? "Invalid" : "Valid"

OLE DB Destination

- **Purpose**: Load the processed data into the DimOrders table in the Northwind_DW database.

FactOrderSales.dtsx

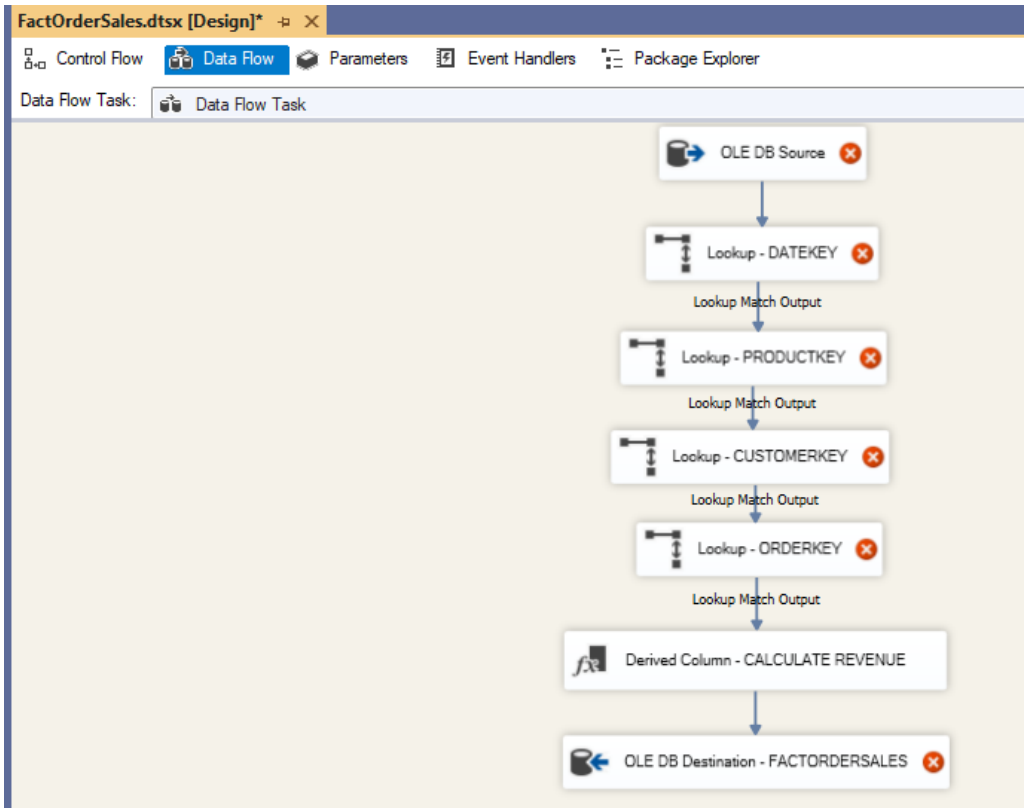


1. Execute SQL Task

- **Purpose**: Clear the existing records in the FactOrderSales table to avoid data duplication.

2. Data Flow Task – Load Data into FactOrderSales

- **Purpose**: Extract, transform, and load data from the source (Orders, OrderDetails, Products, Customers) into the fact table in the data warehouse.



1. OLE DB Source

- **Purpose**: Extract data from Orders, OrderDetails, Products, and Customers in Northwind_DB.

2. Lookup Transformations

- **Purpose**: Match surrogate keys from the dimension tables (DimOrders, DimProducts, DimCustomers, DimDate).

Lookup OrderKey:

- Input: OrderID from the source.
- Output: OrderKey from DimOrders.

Lookup ProductKey:

- Input: ProductID from the source.
- Output: ProductKey from DimProducts.

Lookup CustomerKey:

- Input: CustomerID from the source.
- Output: CustomerKey from DimCustomers.

Lookup DateKey:

- Input: OrderDate from the source.
- Output: DateKey from DimDate.

3. Derived Column Transformation

- **Purpose**: Create calculated fields, e.g., Revenue.
- **Expression**: [UnitPrice] * [Quantity] * (1 - [Discount])

4. OLE DB Destination

- **Purpose**: Load the transformed data into the FactOrderSales table in Northwind_DW.

References

CSV Northwind database. (2024, January 11). Kaggle. https://www.kaggle.com/datasets/cleveranjosqlik/csv-northwind-database

Northwind sample database. (2024, December 18). YugabyteDB Docs. https://docs.yugabyte.com/preview/sample-data/northwind/

Appendices

Bus Matrix and Dimensional Modelling Workbook

Dimensional Modelling Workbook with ETL Rule

Star Schema Diagram

Data Flow Diagram