Data Integration & ETL "DataStage" for TradeXPress

Table of Contents:

Overview

Project Purpose

The main purpose of this project is to implement data engineering methodologies and tools to integrate data from various sources for TradeXPress, a global trade company. The project aims to create an efficient schema for querying the integrated data and to support the Business Intelligence (BI) team in gaining actionable insights from the data.

Scope

The scope of the project includes:

- Designing and implementing an Entity-Relationship (ER) Diagram for the database.
- Using Microsoft SQL Server for database implementation.
- Transforming data using Microsoft Excel to meet the database's schema requirements.
- Inserting transformed data into the database using the import-export wizard in Microsoft SQL Server.
- Creating a comprehensive ETL (Extract, Transform, Load) process.
- Establishing dimension tables to provide context and detail for analysis.
- Creating a fact table to capture key business metrics.
- Implementing a star schema for efficient querying and analysis.
- Utilizing Tableau Public for analysis and reporting.

Objectives

The key objectives of the project are:

- To integrate data from multiple sources to provide a unified view of the business's performance.

- To create a schema that allows for efficient querying and analysis of the integrated data.
- To support the BI team in gaining insights and making informed decisions based on the data.
- To enhance data-driven decision-making processes within TradeXPress.

Stakeholders

The stakeholders involved in the project include:

- The BI team, who will use the integrated data for analysis and reporting.
- The IT team, who will be responsible for implementing the data engineering methodologies and tools.
- TradeXPress management, who will benefit from the insights generated by the project.

Business Process, Units, and KPIs:

Business Process:

The business process for TradeXPress involves managing orders, customers, employees, products, categories, shippers, and shipping details. The process includes receiving orders from customers, processing them, managing inventory, and shipping products to customers. It also involves managing customer relationships, employee information, and shipping logistics.

Business Units:

1. Order Management Unit: Responsible for receiving and processing orders, managing inventory, and ensuring timely delivery.
2. Customer Management Unit: Handles customer inquiries, manages customer accounts, and maintains customer satisfaction.
3. Employee Management Unit: Manages employee information, including job roles, locations, and reporting relationships.
4. Product Management Unit: Manages product information, including pricing, availability, and categorization.
5. Shipping Management Unit: Manages shipping logistics, including selecting shippers, tracking shipments, and ensuring on-time delivery.

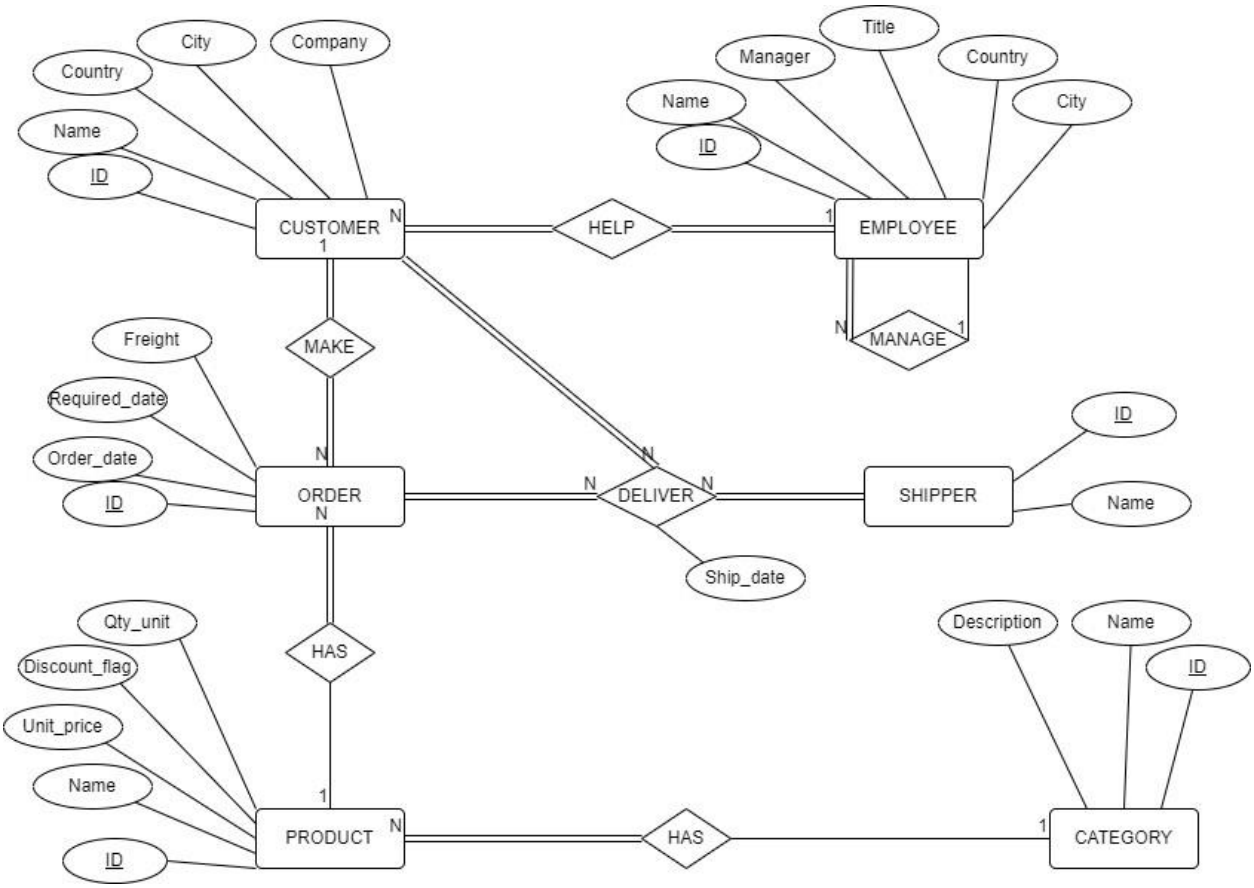
Key Performance Indicators (KPIs):

1. Order Fulfillment Rate: Percentage of orders fulfilled on time.
2. Customer Satisfaction Score: Measure of customer satisfaction based on feedback and surveys.
3. Inventory Turnover Ratio: Number of times inventory is sold and replaced in a given period.
4. Employee Efficiency Ratio: Measure of employee productivity and efficiency.
5. Shipping Accuracy Rate: Percentage of orders shipped accurately and on time.
6. Revenue Growth Rate: Rate of increase in revenue over a specific period.
7. Customer Retention Rate: Percentage of customers retained over a specific period.
8. Average Order Value: Average value of orders placed by customers.

These KPIs help evaluate the performance of the business units and overall business process, identify areas for improvement, and make informed decisions to drive business growth and success.

Database Design:

ERD:



Mapping:

The database design for the TradeXPress project consists of several tables that store information related to orders, customers, employees, products, categories, shippers, and shipping details. Below is a description of each table and its columns:

Orders

- **orderID**: Primary key, unique identifier for each order.
- **orderDate**: Date when the order was placed.
- **requiredDate**: Date when the order is required to be fulfilled.
- **freight**: Shipping cost for the order.

Customers

- **customerID**: Primary key, unique identifier for each customer.
- **contactName**: Name of the customer contact person.
- **contactTitle**: Title of the customer contact person.
- **companyName**: Name of the customer company.
- **city**: City where the customer is located.
- **country**: Country where the customer is located.

CustomerOrders

- **orderID**: Foreign key referencing the Orders table.
- **customerID**: Foreign key referencing the Customers table.

CustomerEmployee

- **customerID**: Foreign key referencing the Customers table.
- **employeeID**: Foreign key referencing the Employees table.

Employees

- **employeeID**: Primary key, unique identifier for each employee.
- **employeeName**: Name of the employee.
- **title**: Job title of the employee.
- **city**: City where the employee is located.
- **country**: Country where the employee is located.
- **reportsTo**: ID of the employee's supervisor.
- **managerID**: ID of the employee's manager.

Order_Details

- **orderID**: Foreign key referencing the Orders table.
- **productID**: Foreign key referencing the Products table.
- **quantity**: Quantity of the product ordered.
- **discount**: Discount applied to the product.

Products

- **productID**: Primary key, unique identifier for each product.
- **productName**: Name of the product.
- **quantityPerUnit**: Quantity per unit of the product.
- **unitPrice**: Price per unit of the product.
- **discontinued**: Indicator if the product has been discontinued.
- **categoryID**: Foreign key referencing the Categories table.

Categories

- **categoryID**: Primary key, unique identifier for each category.
- **categoryName**: Name of the category.
- **description**: Description of the category.

Shippers

- **shipperID**: Primary key, unique identifier for each shipper.
- **companyName**: Name of the shipping company.

Shipping

- **shipperID**: Foreign key referencing the Shippers table.
- **orderID**: Foreign key referencing the Orders table.
- **customerID**: Foreign key referencing the Customers table.
- **shippedDate**: Date when the order was shipped.

Database schema:

This database design allows for efficient storage and retrieval of data related to orders, customers, employees, products, categories, shippers, and shipping details, enabling the BI team to perform in-depth analysis and reporting.

Data Warehouse Modeling (Star Schema) with Periodic Fact Table

In the star schema for the TradeXPress project, the fact table will be the **Orders** table, which contains the primary data about orders. The dimension tables will provide additional context to the orders data. Below is the star schema model for the TradeXPress project:

In addition to the existing star schema model for the TradeXPress project, a periodic fact table is introduced to provide summarized data over specific time periods. This table can help the business in analyzing trends and patterns over time. Below is the updated star schema model with the periodic fact table:

Fact Tables

Orders (Transactional Fact Table)

- **orderID** (Primary Key): Unique identifier for each order.
- **orderDate**: Date when the order was placed.
- **requiredDate**: Date when the order is required to be fulfilled.
- **freight**: Shipping cost for the order.
- **customerID** (Foreign Key): Reference to the Customers dimension table.
- **employeeID** (Foreign Key): Reference to the Employees dimension table.
- **shipperID** (Foreign Key): Reference to the Shippers dimension table.

Orders_Summary (Periodic Fact Table)

- **summaryID** (Primary Key): Unique identifier for each summary record.
- **Order_date**: Day for which the data is summarized.
- **Customer_id**: Customer for which data is summarized by this day.
- **totalOrders**: Total number of orders in the month.
- **totalSales**: Total sales amount in the month.

Dimension Tables

Customers

- **customerID** (Primary Key): Unique identifier for each customer.
- **contactName**: Name of the customer contact person.
- **contactTitle**: Title of the customer contact person.
- **companyName**: Name of the customer company.
- **city**: City where the customer is located.
- **country**: Country where the customer is located.

Employees

- **employeeID** (Primary Key): Unique identifier for each employee.
- **employeeName**: Name of the employee.
- **title**: Job title of the employee.
- **reportsTo**: ID of the employee's supervisor.
- **managerID**: ID of the employee's manager.

Territory

- **territoryID** (Primary Key): Unique identifier for each territory.
- **city**: City in the territory.
- **country**: Country in the territory.

Shippers

- **shipperID** (Primary Key): Unique identifier for each shipper.
- **companyName**: Name of the shipping company.

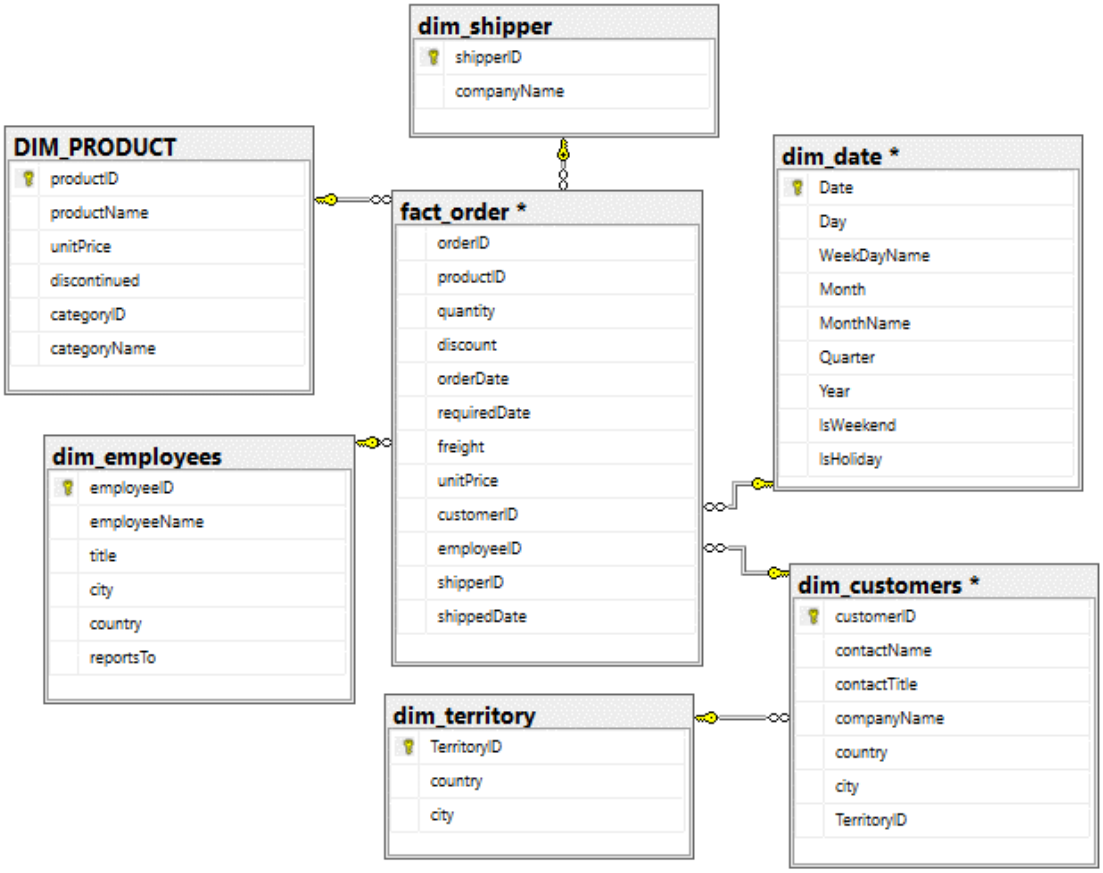
Products

- **productID** (Primary Key): Unique identifier for each product.
- **productName**: Name of the product.
- **quantityPerUnit**: Quantity per unit of the product.
- **unitPrice**: Price per unit of the product.
- **discontinued**: Indicator if the product has been discontinued.
- **categoryID**: Reference to the Categories dimension table.

Categories

- **categoryID** (Primary Key): Unique identifier for each category.
- **categoryName**: Name of the category.
- **description**: Description of the category.

Star Schema:



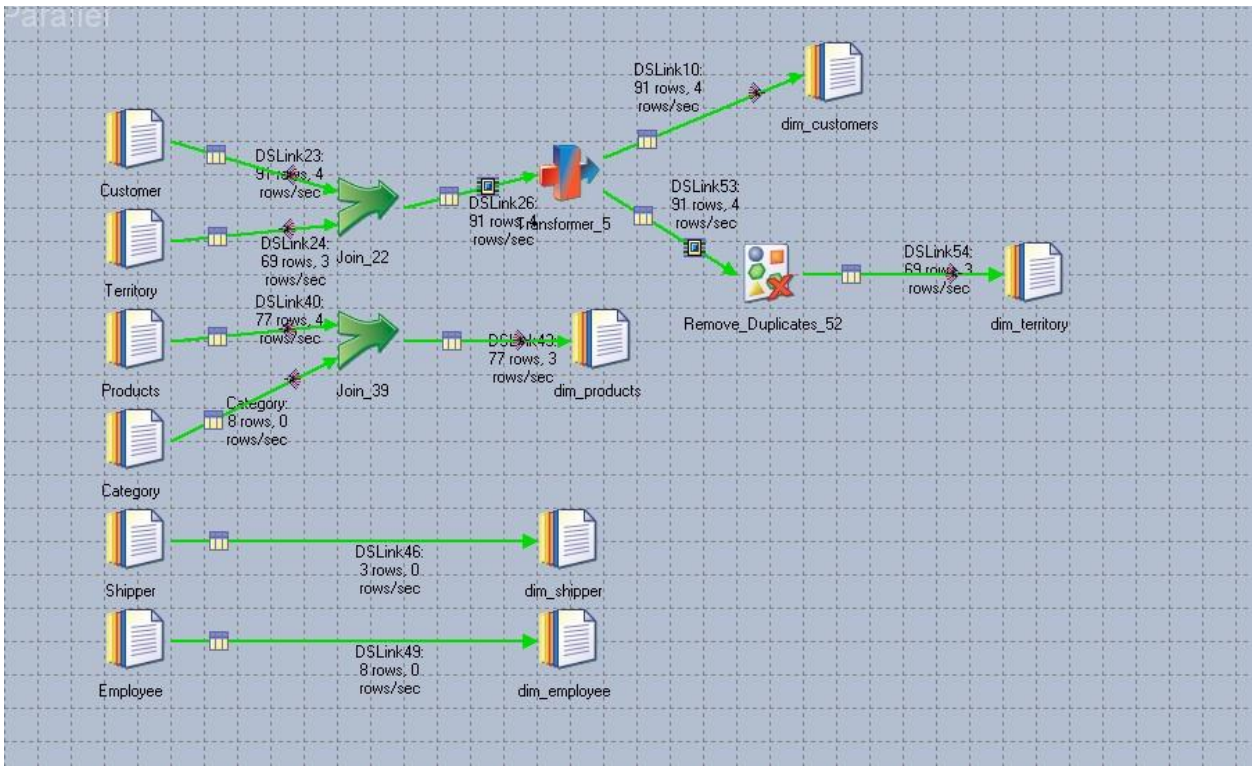
The periodic fact table Orders_Summary provides summarized data over specific time periods (e.g., month, year) and can help the business in analyzing trends, identifying seasonal patterns, and making informed decisions based on the summarized data.

ETL Process Details:

Parallel DataStage Jobs

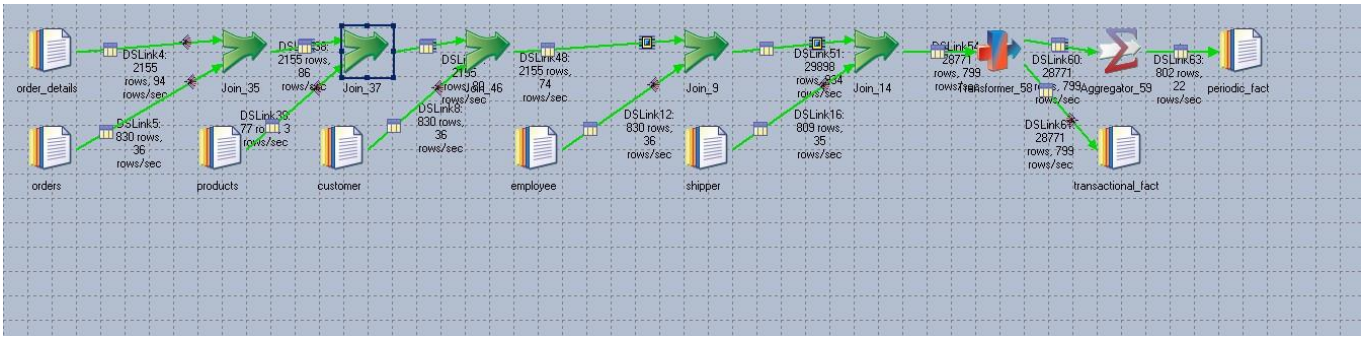
The ETL process for the TradeXPress project involved the use of two parallel DataStage jobs:

1. **Dimension Driving Job**:



This job was responsible for driving the dimensions. It extracted data from the source files, performed necessary transformations (such as cleansing, deduplication, and formatting), and loaded the data into the dimension tables (e.g., Customers, Employees, Shippers).

2. **Transactional and Periodic Fact Derivation Job**:



This job derived the transactional fact table and the periodic fact table. It extracted data from the source files, performed joins, filters, and aggregations to derive the required metrics for the fact tables, and loaded the data into these tables (e.g., Orders, Orders_Summary).

Data Cleansing

During the data cleansing phase, duplicate records were removed from the dataset to ensure data integrity and consistency.

Data Mapping

Data mapping involved using joins, filters, constraints, and aggregation to transform and map the data from the source to the target system. This process ensured that the data was formatted correctly and met the business requirements.

Data Enrichment

Data enrichment was achieved through aggregation to drive periodic facts in the periodic fact table. For example, aggregating sales data to calculate total sales for each month or year.

Data Quality

Data quality was maintained through various measures such as validation rules, data profiling, and data cleansing. These measures ensured that the data was accurate, complete, and consistent.

Conclusion:

Key Findings and Insights

The ETL process using IBM DataStage for the TradeXPress project has successfully integrated data from flat files, transformed it using various techniques such as joins, filters, and aggregation, and build a star schema as the target. This process has provided valuable insights into the business's performance, including sales trends, product performance, key customer identification, shipping costs and delays, employees' performance, and market analysis.