**Supply Chain Data Warehouse (ETL) Project Documentation**

**Project Overview:**

This document outlines the systematic approach to building a robust data warehouse for supply chain analytics. The project leverages a Medallion Architecture (Bronze, Silver, Gold layers) to ensure data quality, consistency, and optimal structure for business intelligence. The primary goal is to transform raw supply chain data into a well-organized star schema, facilitating efficient reporting and analytical insights.

**Step 1: Data Sourcing and Initial Exploration**

* **Data Source:** Kaggle Supply Chain Dataset
* **Initial Data Profile:**
  + Total Columns: 54
  + Total Rows: 180,519

**Step 2: Medallion Architecture Implementation**

**A. Bronze Layer (Staging Area)**

The Bronze Layer acts as the raw data ingestion zone, capturing data in its original form.

* **Data Ingestion:**
  + Data loaded directly from source without transformation.
* **Adjustments:**
  + Removed Product Image column due to loading errors.
  + Added LoadTime\_to\_StagingArea column to record ingestion timestamp.
* **Resulting Data Profile:**
  + Columns: 53
  + Rows: 180,519
* **Show Column Names & Data Types:**

| **Column Name** | **Data Type** |
| --- | --- |
| Type | VARCHAR(255) |
| Days for shipping (real) | VARCHAR(255) |
| Days for shipment (scheduled) | VARCHAR(255) |
| Benefit per order | VARCHAR(255) |
| Sales per customer | VARCHAR(255) |
| Delivery Status | VARCHAR(255) |
| Late\_delivery\_risk | VARCHAR(255) |
| Category Id | VARCHAR(255) |
| Category Name | VARCHAR(255) |
| Customer City | VARCHAR(255) |
| Customer Country | VARCHAR(255) |
| Customer Email | VARCHAR(255) |
| Customer Fname | VARCHAR(255) |
| Customer Id | VARCHAR(255) |
| Customer Lname | VARCHAR(255) |
| Customer Password | VARCHAR(255) |
| Customer Segment | VARCHAR(255) |
| Customer State | VARCHAR(255) |
| Customer Street | VARCHAR(255) |
| Customer Zipcode | VARCHAR(255) |
| Department Id | VARCHAR(255) |
| Department Name | VARCHAR(255) |
| Latitude | VARCHAR(255) |
| Longitude | VARCHAR(255) |
| Market | VARCHAR(255) |
| Order City | VARCHAR(255) |
| Order Country | VARCHAR(255) |
| Order Customer Id | VARCHAR(255) |
| order date (DateOrders) | DATETIME |
| Order Id | VARCHAR(255) |
| Order Item Cardprod Id | VARCHAR(255) |
| Order Item Discount | VARCHAR(255) |
| Order Item Discount Rate | VARCHAR(255) |
| Order Item Id | VARCHAR(255) |
| Order Item Product Price | VARCHAR(255) |
| Order Item Profit Ratio | VARCHAR(255) |
| Order Item Quantity | VARCHAR(255) |
| Sales | VARCHAR(255) |
| Order Item Total | VARCHAR(255) |
| Order Profit Per Order | VARCHAR(255) |
| Order Region | VARCHAR(255) |
| Order State | VARCHAR(255) |
| Order Status | VARCHAR(255) |
| Order Zipcode | VARCHAR(255) |
| Product Card Id | VARCHAR(255) |
| Product Category Id | VARCHAR(255) |
| Product Description | VARCHAR(255) |
| Product Name | VARCHAR(255) |
| Product Price | VARCHAR(255) |
| Product Status | VARCHAR(255) |
| shipping date (DateOrders) | DATETIME |
| Shipping Mode | VARCHAR(255) |
| LoadTime\_to\_StagingArea | VARCHAR(255) |

**B. Silver Layer (Transformation Layer)**

The Silver Layer focuses on data cleansing, enrichment, and conformance to prepare the data for analytics.

* **Data Cleaning:**
  + Removed columns irrelevant or sensitive to analysis:
    - Customer Email, Customer Password, Product Description, Product Status
  + Resolved column duplication by selecting preferred naming conventions:
    - OrderItemCardProdID (preferred over ProductCardID)
    - OrderCustomerID (preferred over CustomerID)
    - CategoryID (preferred over ProductCategoryID)
  + Columns removed in total: 7
* **Data Enrichment:**
  + Added calculated and derived columns:
    - Customer Full Name (concatenation of first and last name)
    - LocationID (unique location identifier)
    - LoadTimeToTransformArea (timestamp for Silver Layer load)
* **Transformed Data Profile:**
  + Columns: 49 (46 original + 3 new)
  + Rows: 180,519
* **Transformed Column Names & Data Types:**

| **Column Name** | **Data Type** |
| --- | --- |
| PaymentType | nvarchar |
| ShippingDaysReal | int |
| ShipmentDaysScheduled | int |
| BenefitPerOrder | numeric |
| SalesPerCustomer | numeric |
| DeliveryStatus | nvarchar |
| LateDeliveryRisk | int |
| CategoryName | nvarchar |
| CustomerCity | nvarchar |
| CustomerCountry | nvarchar |
| CustomerFname | nvarchar |
| CustomerId | int |
| CustomerLname | nvarchar |
| CustomerSegment | nvarchar |
| CustomerState | nvarchar |
| CustomerStreet | nvarchar |
| CustomerZipcode | int |
| DepartmentId | int |
| DepartmentName | nvarchar |
| Latitude | numeric |
| Longitude | numeric |
| Market | nvarchar |
| OrderCity | nvarchar |
| OrderCountry | nvarchar |
| OrderDateDateOrders | datetime |
| OrderId | int |
| OrderItemDiscount | numeric |
| OrderItemDiscountRate | numeric |
| OrderItemId | int |
| OrderItemProductPrice | numeric |
| OrderItemProfitRatio | numeric |
| OrderItemQuantity | int |
| Sales | numeric |
| OrderItemTotal | numeric |
| OrderProfitPerOrder | numeric |
| OrderRegion | nvarchar |
| OrderState | nvarchar |
| OrderStatus | nvarchar |
| OrderZipcode | int |
| ProductCardId | int |
| ProductCategoryId | int |
| ProductName | nvarchar |
| ProductPrice | numeric |
| ShippingDate | datetime |
| ShippingMode | nvarchar |
| LoadTimeToStagingArea | datetime |
| CustomerFullName | nvarchar |
| LocationID | bigint |
| LoadTimeToTransformData | nvarchar |

**Step 3: Data Modeling (Gold Layer)**

The Gold Layer contains curated and conformed data modeled into a **star schema** for optimal BI performance.

**A. Dimension Tables**

| **Ref** | **Dimension Table** | **Primary Key** | **Key Attributes** |
| --- | --- | --- | --- |
| 1 | DimCustomer | CustomerKey | Cus\_FName, Cus\_LName, Cus\_City, Cus\_Country, Cus\_Segment, etc. |
| 2 | DimStore | StoreKey | DepartmentID, Dep\_Name |
| 3 | DimOrderStatus | OrderStatusKey | OrderStatus |
| 4 | DimCategory | CategoryKey | CategoryName |
| 5 | DimProduct | ProductKey | ProductName, ProductStatus, ProductPrice, ProductCategoryID |
| 6 | DimActualShipping | ActShippingKey | ShippingDaysReal |
| 7 | DimScheduledShipping | SchShippingKey | ShipmentDays |
| 8 | DimRisk | LateDeliveryRiskKey | LateDeliveryRisk |
| 9 | DimDiscountRate | DiscountRateKey | DiscountRate |
| 10 | DimDeliveryLocation | DeliveryLocKey | OrderCity, OrderState, OrderRegion, OrderCountry, Market |
| 11 | DimDeliveryStatus | DeliveryStatusKey | DeliveryStatus |
| 12 | DimProfitRatio | ProfitRatioKey | OrderProfitRatio |
| 13 | DimShippingMode | ShippingModeKey | ShippingMode |
| 14 | DimPayment | PaymentKey | PaymentType |
| 15 | DimStoreLocation | StoreLocKey | Latitude, Longitude, (Concat LocationID) |
| 16 | DimDate | DateKey | Year, Month, Day, Quarter, DayOfWeek, WeekOfYear |

**B. Fact Table: FactSupplyChain**

* + Columns: 40
  + Rows: 180,519

**1. Column Categorization**

We’ll classify your columns into:

* **Quantitative columns** – for numerical analysis.
* **Categorical/Dimensional columns** – to group, segment, or categorize data.
* **Date/Time columns** – for time-based analysis.
* **Keys/IDs** – mainly for joins.

**2. Quantitative Columns (Measures)**

These are your numerical columns used for calculations like totals, averages, ratios, etc.

| **Column Name** | **Description** |
| --- | --- |
| BenefitPerOrder | Financial benefit per order |
| SalesPerCustomer | Aggregated sales per customer |
| OrderItemDiscount | Discount amount per item |
| OrderItemDiscountRate | Discount rate on item |
| OrderItemProductPrice | Item price |
| OrderItemQuantity | Quantity ordered |
| Sales | Total sales amount |
| OrderItemTotal | Total order item value |
| OrderProfitPerOrder | Profit per order |
| ProductPrice | Product price |
| ShippingDaysReal | Actual shipping days |
| ShipmentDaysScheduled | Scheduled shipping days |
| LateDeliveryRisk | Delivery risk indicator |
| DiscountRate | Applied discount rate |
| OrderItemProfitRatio | Profit ratio per item |

**3. Categorical / Dimensional Columns**

Used for slicing and dicing the data (group by).

| **Column Name** | **Description** |
| --- | --- |
| CustomerKey | FK to DimCustomer |
| DepartmentId | FK to DimStore (store department) |
| StoreKey | FK to DimStore |
| OrderStatusKey | FK to DimOrderStatus |
| ProductCategoryId | FK to DimCategory |
| CategoryKey | FK to DimCategory |
| ProductKey | FK to DimProduct |
| LateDeliveryRiskKey | FK to DimRisk |
| DiscountRateKey | FK to DimDiscountRate |
| DeliveryKey | FK to DimDeliveryLocation |
| DeliveryStatusKey | FK to DimDeliveryStatus |
| ProfitRatioKey | FK to DimProfitRatio |
| ShippingModeKey | FK to DimShippingMode |
| PaymentTypeKey | FK to DimPayment |
| StoreLocKey | FK to DimStoreLocation |
| OrderStatusKey | Order status |

**4. Date/Time Columns**

Useful for tracking when orders were placed, shipped, or loaded into the system.

| **Column Name** | **Description** |
| --- | --- |
| OrderDateDateOrders | FK to DimDate (order date) |
| ShippingDate | Actual date of shipping |
| LoadTimeToTransformData | Data transformation timestamp |
| LoadTimeToStagingArea | Data ingestion timestamp |

**5. Identifiers / Keys**

These are primary or surrogate keys for entities.

| **Column Name** | **Description** |
| --- | --- |
| OrderId | Order identifier |
| OrderItemId | Order item identifier |
| ProductCardId | Product card identifier |
| CustomerId | Customer identifier |

**6. Suggested Measures to Create**

To support **category-wise** insights and analysis, you can derive the following measures:

| **Measure Name** | **Calculation** |
| --- | --- |
| **Total Sales** | SUM(Sales) |
| **Total Orders** | COUNT(DISTINCT OrderId) |
| **Total Items Sold** | SUM(OrderItemQuantity) |
| **Average Discount Rate** | AVG(OrderItemDiscountRate) |
| **Total Discount Given** | SUM(OrderItemDiscount) |
| **Profit Margin (%)** | (SUM(OrderProfitPerOrder) / SUM(Sales)) \* 100 |
| **Average Order Value** | SUM(Sales) / COUNT(DISTINCT OrderId) |
| **Total Benefit** | SUM(BenefitPerOrder) |
| **Late Delivery Count** | COUNT(\*) WHERE LateDeliveryRisk = 1 |
| **Average Shipping Delay (days)** | AVG(ShippingDaysReal - ShipmentDaysScheduled) |
| **Average Product Price** | AVG(OrderItemProductPrice) |

**7. Show the Fact Table**:

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| BenefitPerOrder | numeric | Financial benefit per order |
| SalesPerCustomer | numeric | Sales aggregated per customer |
| OrderItemDiscount | numeric | Discount amount on order items |
| OrderItemDiscountRate | numeric | Discount rate |
| OrderItemProductPrice | numeric | Price per order item |
| OrderItemQuantity | int | Quantity ordered |
| Sales | numeric | Total sales amount |
| OrderItemTotal | numeric | Total order item value |
| OrderProfitPerOrder | numeric | Profit per order |
| ProductPrice | numeric | Product price |
| ShippingDaysReal | int | Actual shipping days |
| ShipmentDaysScheduled | int | Scheduled shipping days |
| LateDeliveryRisk | int | Indicator of late delivery risk |
| DiscountRate | numeric | Discount rate applied |
| OrderItemProfitRatio | numeric | Profit ratio on order items |
| CustomerKey | int | FK to DimCustomer |
| DepartmentId | int | FK to DimStore |
| StoreKey | int | FK to DimStore |
| OrderStatusKey | int | FK to DimOrderStatus |
| ProductCategoryId | int | FK to DimCategory |
| CategoryKey | int | FK to DimCategory |
| ProductKey | int | FK to DimProduct |
| ActShippingKey | int | FK to DimActualShipping |
| SchShippingKey | int | FK to DimScheduledShipping |
| LateDeliveryRiskKey | int | FK to DimRisk |
| DiscountRateKey | int | FK to DimDiscountRate |
| DeliveryKey | int | FK to DimDeliveryLocation |
| DeliveryStatusKey | int | FK to DimDeliveryStatus |
| ProfitRatioKey | int | FK to DimProfitRatio |
| ShippingModeKey | int | FK to DimShippingMode |
| PaymentTypeKey | int | FK to DimPayment |
| StoreLocKey | int | FK to DimStoreLocation |
| OrderDateDateOrders | bigint | Foreign key to DimDate |
| ShippingDate | datetime | Date of shipping |
| LoadTimeToTransformData | datetime | Timestamp of data transformation |
| LoadTimeToStagingArea | datetime | Timestamp of data ingestion |
| OrderId | int | Order identifier |
| OrderItemId | int | Order item identifier |
| ProductCardId | int | Product card identifier |
| CustomerId | int | Customer identifier |

**Conclusion:**

The Supply Chain Data Warehouse project successfully transforms raw operational data into a high-performing analytical foundation using a layered Medallion Architecture (Bronze, Silver, and Gold). Each phase of the architecture plays a distinct and critical role:

* **The Bronze Layer** ensures accurate data capture and traceability by preserving the original state of the data with minimal interference.
* **The Silver Layer** enhances the data by cleansing, enriching, and standardizing it, ensuring high data quality and consistency.
* **The Gold Layer** delivers a business-friendly star schema that supports efficient querying, robust reporting, and actionable insights for supply chain performance.

By applying structured transformations and creating dimensionally modeled data, the warehouse empowers business users and analysts with a clear, navigable view of key supply chain metrics—such as sales performance, shipping efficiency, customer segmentation, product profitability, and delivery risk.

Furthermore, with the creation of derived measures like **Profit Margin**, **Average Order Value**, and **Late Delivery Count**, the warehouse enables advanced analytics and supports data-driven decision-making.

This warehouse can be seamlessly integrated with BI tools like Power BI, Tableau, or Looker to provide interactive dashboards and KPI tracking. Additionally, the modular architecture and clearly defined data lineage facilitate ongoing scalability, governance, and maintenance, making it a future-ready solution for evolving supply chain analytics needs.