

# Business Document: Data Warehouse Design Proposal

## 1. Data Model Concept

The proposed data model is based on a dimensional star schema, looking for being useful for fast analytical queries and business insights.

Divided in dimension tables, that have the data from the entities and the fact tables, that will be the ones that link the content of the dimension tables.

### Fact Tables

- Fact\_Orders (Orders/Sales):
  - Order ID
  - Customer ID
  - Product ID
  - Quantity Ordered
  - Gross Price
  - Net Price (after discount)
  - Total = Quantity × Net Price
  
- Fact\_Delivery:
  - Delivery ID
  - Scheduled Delivery - Date
  - Actual Delivery - Date
  - Delay in Minutes
  - On-Time Delivery. (Yes/No)

### Dimension Tables

- Dim\_Customer:
  - Customer ID
  - Name
  - Type

- Dim\_Product:
  - Material ID
  - Product Name
  - Base Price
- Dim\_Delivery:
  - Delivery ID
  - Date
  - Time
  - Day of Week
- Dim\_Order:
  - Order ID
  - Order Date
  - Status

With this model, the business can efficiently analyze:

- Most sold products
- Average product prices
- Revenue per customer
- On-time performance of logistics

## 2. Architecture Concept

The data warehouse architecture follows a layered approach, ensuring scalability, maintainability, and clear data order.

### Source Systems

- Internal Excel files (orders, deliveries)
- External FTP data (truck delivery confirmations)

### Architecture Layers

1. Source Layer: Raw data ingestion from Excel and FTP
2. Transformation Layer:
  - Joins across orders, deliveries, confirmations.

- Data cleaning and format normalization
- Calculation of key metrics (totals, delays)

3. Data Warehouse Layer:

- Fact and dimension tables built using star schema
- Optimized for analytical reporting

4. Presentation Layer (BI Tools)

- Fact and dimension tables built using star schema
- Optimized for analytical reporting

3. Key Challenges to Resolve

Challenge	Description
Data integration from mixed sources	Combining structured Excel files and FTP feeds with differing formats
Date and time format inconsistencies	Mixed formats across systems (e.g., dd/mm/yyyy vs. mm/dd/yyyy)
Linking deliveries to orders	One delivery may include multiple order lines
Transporter identification	Not explicitly available in the source data
Partial deliveries	Orders may be fulfilled across several deliveries
Discount and pricing logic	Ensuring accurate net price calculation based on quantity and discount
Handling missing data	Some orders may not yet have delivery records

4. Open Questions

Open Question	Reason
What is the expected data refresh frequency?	Determines whether ETL should be daily, hourly, or event-based
Which BI tool will be used?	Impacts dashboard design and performance, or company structure
How to treat partially delivered orders?	Define business logic for order completeness

Amparo Iglesias

April 2025

<b>Will analysis by region or geography be required?</b>	May require geographic dimensions or external mapping
<b>What should be done with undelivered orders?</b>	Should they be excluded or marked as "open"?
<b>How could routes or deliveries be optimized?</b>	Find new routes, ways to ship or reduce unsuccessful shipments.