

FACULTY OF ENGINEERING AND BASIC SCIENCES
ACADEMIC PROGRAM: DATA ENGINEERING AND ARTIFICIAL INTELLIGENCE

COURSE: ETL (G01)
LAB-3. Dimensional Data Modeling

1. Context

A retail company wants to strengthen its decision-making process through a **Business Intelligence (BI)** system. The sales management team has requested the implementation of a **Data Warehouse** based on **dimensional modeling**, enabling the calculation of **Key Performance Indicators (KPIs)** and the creation of visual analytics to support strategic decisions.

The analytics team must design and implement a **complete ETL pipeline**, starting from a transactional data source (OLTP), transforming the data, and loading it into a dimensional Data Warehouse (OLAP).

2. Learning Objectives

By completing this lab, students will be able to:

- Design a star schema dimensional model.
- Implement a complete ETL pipeline using Python and SQL.
- Load transformed data into a Data Warehouse.
- Compute business KPIs from fact and dimension tables.
- Build visualizations based on dimensional data.
- Understand the difference between staging data and analytical data.

3. Business Requirements

The management team wants to answer the following strategic questions:

1. What is the sales volume and revenue per product category?
2. Which sales channels (physical store vs online) generate the highest revenue?
3. How do sales evolve over time (monthly trends)?
4. Which brands are the most profitable?

4. Source Data (OLTP)

The data originates from a transactional system (OLTP) and is structured in normalized tables.

Products Table

Field	Type	Description
product_id (PK)	Integer	Unique identifier of the product

Field	Type	Description
name	Text	Product name
category	Text	Main product category
brand	Text	Product brand
unit_price	Decimal	Product list price
unit_cost	Decimal	Product acquisition cost

Customers Table

Field	Type	Description
customer_id (PK)	Integer	Unique identifier of the customer
name	Text	Full name
city	Text	City
country	Text	Country
age	Integer	Customer age

Channels Table

Field	Type	Description
channel_id (PK)	Integer	Unique identifier of the sales channel
channel	Text	Sales channel (Physical Store, Online)

Sales Table

Field	Type	Description
sale_id (PK)	Integer	Unique identifier of the sale
sale_date	Date	Sale date
product_id (FK)	Integer	Reference to the sold product
customer_id (FK)	Integer	Reference to the customer
channel_id (FK)	Integer	Reference to the channel
quantity	Integer	Number of units sold

Field	Type	Description
unit_price_sale	Decimal	Unit price applied in the sale (after discount)

5. Dataset Conditions

Each group must **generate a synthetic dataset** that meets the following conditions:

- At least **200 sales records**.
- **4 consecutive months** of sales data.
- At least **20 customers** from **3 different countries**.
- **3 sales channels** (2 physical stores + 1 online).
- At least **4 brands** and **4 product categories**.

6. ETL Pipeline Requirements

Students must implement a complete ETL pipeline with the following layers:

a. Extract (E)

- Read raw transactional data from CSV files.
- Validate schema and basic data types.
- Store extracted data in a staging layer.

b. Transform (T)

- Clean and standardize data (dates, text fields, numeric types).
- Create derived attributes (month, year, total_amount, profit).
- Generate surrogate keys for dimensions.
- Prepare data for dimensional modeling.

c. Load (L)

- Load dimension tables first.
- Load the fact table using foreign keys.
- Ensure referential integrity in the Data Warehouse.

7. Activities

Activity 1: Define KPIs and Business Understanding

- Define **KPIs** (include the ones required by management and at least two additional).
- For each KPI:
 - Provide the calculation formula.
 - Specify the required fact and dimension tables.
 - Propose an appropriate visualization type.
 - Justify its business value.

Activity 2: Dimensional Model Design

- a. Design a star schema including:**
 - At least one fact table.
 - Relevant dimension tables (Product, Customer, Channel, Date).
- b. Define:**
 - Grain of the fact table.
 - Surrogate keys.
 - Measures and dimensions.
- c. Provide a diagram and written explanation.**

Suggested Grain for This Lab:

One row in the fact table represents one product sold to one customer, through one channel, on one specific date.

Activity 3: ETL Implementation

Implement the ETL pipeline using Python and SQL:

Extract

- Read raw CSV files.

Transform

- Create a **Date Dimension** (year, month, quarter).
- Standardize categorical values.
- Calculate:
 - Total sales amount.
 - Profit (optional but recommended).

Load

- Create the Data Warehouse schema.
- Load dimension tables.
- Load the fact table.
- Explain the **loading order** and key management strategy.

Activity 4: KPI Deployment & Visualization

- a. Query the Data Warehouse using SQL.
- b. Load query results into Pandas.
- c. Compute the KPIs.
- d. Build visualizations using:
 - Matplotlib / Seaborn / Power BI / Tableau (any one).
- e. Interpret the results from a business perspective.

8. Project Structure

```
etl_lab_3/
|
|--- data/
|   |--- raw/
```

```

  └── warehouse/
    ├── etl/
    │   ├── extract.py
    │   ├── transform.py
    │   └── load.py
    ├── sql/
    │   ├── create_tables.sql
    │   └── queries.sql
    └── visualization/
        └── kpi_dashboard.py

```

README.md

9. Deliverables

a. Technical Report

- Dimensional model (diagram + explanation).
- KPI definitions and formulas.
- ETL design explanation.
- SQL queries.
- Reflection on how AI tools were used.

b. Python Project

- Well-structured ETL pipeline.
- Synthetic data generated.
- Data Warehouse database file.

3. Presentation

- Business problem overview.
- Dimensional model explanation.
- ETL pipeline demonstration.
- KPI results and insights.

10. Group Work

- Groups of 2 to 4 students.
- One submission per group
- Suggested roles:
 - **Product Owner.**
 - **Data Engineer**
 - **BI Analyst.**

11. Business Scenarios

- a. **Appliance store** (household electronics). (Groups: 1, 6)
- b. **Technology store** (gadgets and IT products). (Groups: 2, 7)
- c. **Online bookstore** (physical and e-books). (Groups: 3, 8)
- d. **Grocery chain** (food and consumer goods). (Groups: 4, 9)
- e. **Clothing store** (clothing retail company). (Groups: 5, 10)

Número de ID	Grupo
1144108706	8
1114873918	1
1110293996	2
1126003708	6
1109185806	7
1112040828	3
1112045720	3
1126844478	5
1112040616	8
1114542796	1
1112044531	5
1110289536	6
1110041641	9
1059236663	4
1110294473	2
1110287118	1
1107843663	9
1111481197	4
1112391393	1
1001237121	7
1107843747	10
1087804322	3
1110366547	5
1105369599	7
1110042234	3
1110043593	4
1109664941	10
1110040708	10
1114881135	8
1111665738	4