University of South-Eastern Norway

Data Dictionary

BID3000: Business Intelligence and Data Warehouse

Professor Peyman Teymoori

Autumn 2025

Candidate no.:7003 | Lidor Shachar
Candidate no.:7014 | Andreas Moen
Candidate no.:7007 | Lars Andreas Strand

Brazilian E-Commerce

Public Dataset by Olist

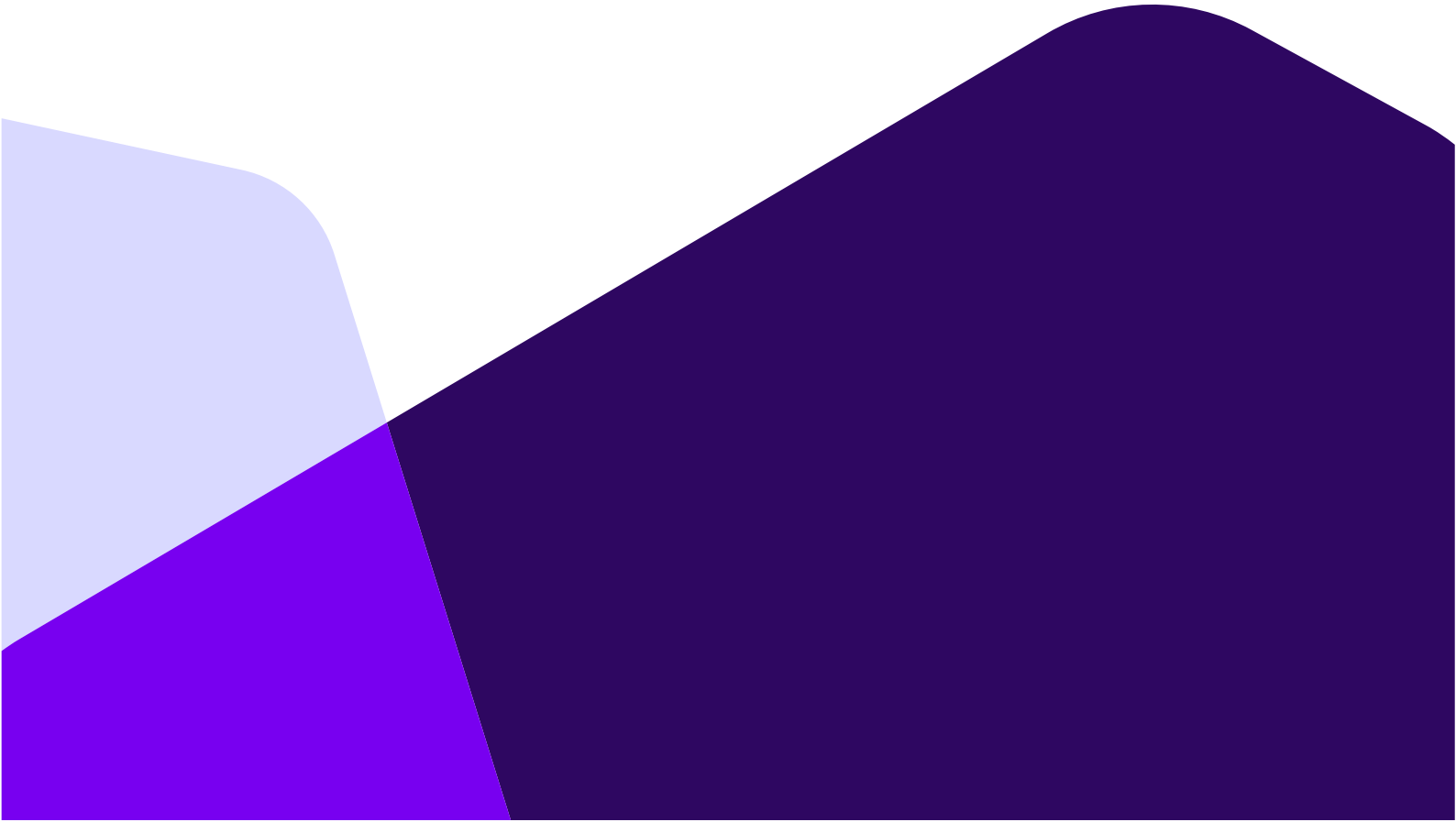

Table of Contents

Overview

In this data dictionary document, we display the data warehouse structure we created for Olist – Brazilian E-commerce, that includes:

- Tables
- Columns
- Relationships
- Business rules

Schema design

We implemented a star schema architecture with three fact tables and seven dimension tables, while attempting to optimize it for Online Analytical Processing (OLAP). That would allow us to gain business insights and access information faster (faster queries).

Grain definition

- **FACT_SALES:** One row per order item (order_id + order_item_id)
- **FACT_DELIVERY_PERFORMANCE:** One row per order (order_id)
- **FACT_CUSTOMER_REVIEWS:** One row per review (review_id)

Dimension Tables

DIM_DATE – Date Dimension

Column Name	Data Type	Nullable	PK/FK	Description	Source	Sample Values
date_key	SERIAL	No	PK	Surrogate key	Generated	1, 2, 3...
full_date	DATE	No	UK	Calendar date	Calculated	2017-01-01, 2018-12-31
day_of_week	INTEGER	Yes	-	Day number (1=Sunday)	Calculated	1, 2, 3, 4, 5, 6, 7
day_name	VARCHAR(10)	No	-	Day name	Calculated	Sunday, Monday, Tuesday
day_of_month	INTEGER	Yes	-	Day of month (1-31)	Calculated	1, 15, 31
month_number	INTEGER	Yes	-	Month number (1-12)	Calculated	1, 6, 12
month_name	VARCHAR(10)	No	-	Month name	Calculated	January, June, December
quarter_number	INTEGER	Yes	-	Quarter (1-4)	Calculated	1, 2, 3, 4
year_number	INTEGER	Yes	-	Year	Calculated	2017, 2018
is_weekend	BOOLEAN	No	-	Weekend flag	Calculated	TRUE, FALSE

Business purpose: The date dimension centralizes date and time for all analysis.

DIM_GEOGRAPHY - Geographic Dimension

Column Name	Data Type	Nullable	PK/FK	Description	Source	Sample Values
geography_key	SERIAL	No	PK	Surrogate key	Generated	1, 2, 3...
zip_code	VARCHAR(10)	No	UK	Brazilian ZIP code	geolocation_dataset.csv	01037-010, 01310-100
city	VARCHAR(100)	No	-	City name	geolocation_dataset.csv	sao paulo, rio de janeiro
state	VARCHAR(2)	No	-	State abbreviation	geolocation_dataset.csv	SP, RJ, MG
region	VARCHAR(20)	No	-	Geographic region	Derived from state	Southeast, South, Northeast
latitude	DECIMAL(10,7)	Yes	-	Latitude coordinate	geolocation_dataset.csv	23.550519,22.906846 7
longitude	DECIMAL(10,7)	Yes	-	Longitude coordinate	geolocation_dataset.csv	46.633309, 43.1728965

Business purpose: Geographic analysis and regional reporting.
SCD Type: Type 1 (overwrite).

DIM_CUSTOMER - Customer Dimension (SCD Type 2)

Column Name	Data Type	Nullable	PK/FK	Description	Source	Sample Values
customer_key	SERIAL	No	PK	Surrogate key	Generated	1, 2, 3...
customer_id	VARCHAR(50)	No	-	Business key	customers_dataset.csv	06b8999e2fba1a1fbc88172c00ba8bc7
customer_unique_id	VARCHAR(50)	No	-	Unique customer ID	customers_dataset.csv	861eff4711a542e4b93843c6dd7febb0
customer_zip_code	VARCHAR(10)	No	-	Customer ZIP code	customers_dataset.csv	14409-113, 09210-580
customer_city	VARCHAR(100)	No	-	Customer city	customers_dataset.csv	São Paulo, Franca
customer_state	VARCHAR(2)	No	-	Customer state	customers_dataset.csv	SP, RJ, MG
customer_region	VARCHAR(20)	No	-	Customer region	Derived	Southeast, South
effective_date	DATE	No	-	SCD start date	ETL Process	2024-01-01
expiry_date	DATE	Yes	-	SCD end date	ETL Process	9999-12-31
is_current	BOOLEAN	Yes	-	Current record flag	ETL Process	TRUE, FALSE

Business purpose: Customer dimension with historical tracking.
SCD Type: Type 2 (historical tracking).

DIM_PRODUCT - Product Dimension

Column Name	Data Type	Nullable	PK/FK	Description	Source	Sample Values
product_key	SERIAL	No	PK	Surrogate key	Generated	1, 2, 3...
product_id	VARCHAR (50)	No	UK	Product identifier	products_dataset.csv	1e9e8ef04dbcff4541ed26657ea7a6cd
product_category_portuguese	VARCHAR (100)	Yes	-	Category in Portuguese	products_dataset.csv	cama_mesa_banho, esporte_lazer
product_category_english	VARCHAR (100)	Yes	-	Category in English	category_translation.csv	bed_bath_table, sports_leisure
product_name_length	INTEGER	Yes	-	Product name length	products_dataset.csv	58, 42, 76
product_photos_qty	INTEGER	Yes	-	Number of photos	products_dataset.csv	1, 2, 3, 4
product_weight_grams	INTEGER	Yes	-	Weight in grams	products_dataset.csv	740, 1000, 225
product_length_cm	INTEGER	Yes	-	Length in cm	products_dataset.csv	16, 25, 30
product_height_cm	INTEGER	Yes	-	Height in cm	products_dataset.csv	10, 15, 20
product_width_cm	INTEGER	Yes	-	Width in cm	products_dataset.csv	14, 18, 25
product_volume_cm3	INTEGER	Yes	-	Calculated volume	ETL Calculated	2240, 6750, 15000
product_size_category	VARCHAR (20)	Yes	-	Size classification	ETL Calculated	Small, Medium, Large, XLarge
product_category_level_1	VARCHAR (50)	Yes	-	Top-level category	ETL Calculated	Home & Garden, Sports
has_complete_attributes	BOOLEAN	Yes	-	Data completeness flag	ETL Calculated	TRUE, FALSE

Business purpose: Product catalog with derived analytics attributes.

SCD Type: Type 1 (overwrite).

DIM_SELLER - Seller Dimension

Column Name	Data Type	Nullable	PK/FK	Description	Source	Sample Values
seller_key	SERIAL	No	PK	Surrogate key	Generated	1, 2, 3...
seller_id	VARCHAR(50)	No	UK	Seller identifier	sellers_dataset.csv	3442f8959a84dea7ee197c632cb2df15
seller_zip_code	VARCHAR(10)	No	-	Seller ZIP code	sellers_dataset.csv	13023-101, 04038-001
seller_city	VARCHAR(100)	No	-	Seller city	sellers_dataset.csv	Campinas, São Paulo
seller_state	VARCHAR(2)	No	-	Seller state	sellers_dataset.csv	SP, RJ, MG
seller_region	VARCHAR(20)	No	-	Seller region	Derived	Southeast, South

Business purpose: Seller information for marketplace analysis.
SCD Type: Type 1 (overwrite).

DIM_PAYMENT_TYPE - Payment Method Dimension

Column Name	Data Type	Nullable	PK/FK	Description	Source	Sample Values
payment_type_key	SERIAL	No	PK	Surrogate key	Generated	1, 2, 3, 4
payment_type	VARCHAR(20)	No	UK	Payment method	payments_dataset.csv	credit_card, boleto, voucher, debit_card
payment_method_category	VARCHAR(20)	No	-	Payment category	ETL Derived	Card, Cash, Digital, Other

Business purpose: Payment method classification for financial analysis.
SCD Type: Type 1 (overwrite)

DIM_ORDER_STATUS - Order Status Dimension

Column Name	Data Type	Nullable	PK/FK	Description	Source	Sample Values
order_status_key	SERIAL	No	PK	Surrogate key	Generated	1, 2, 3...
order_status	VARCHAR(20)	No	UK	Order status	orders_dataset.csv	delivered, shipped, processing, canceled
status_category	VARCHAR(20)	No	-	Status grouping	ETL Derived	Completed, In_Progress, Cancelled
is_final_status	BOOLEAN	Yes	-	Final status flag	ETL Derived	TRUE, FALSE

Business purpose: Order lifecycle tracking and analysis.

SCD Type: Type 1 (overwrite).

Fact Tables

FACT_SALES - Sales Transactions

Column Name	Data Type	Nullable	PK/FK	Description	Source	Sample Values
sales_key	SERIAL	No	PK	Surrogate key	Generated	1, 2, 3...
order_id	VARCHAR (50)	No	BK	Order identifier	order_items_dataset.csv	e481f51cbdc54678b7cc49136f2d6af7
order_item_id	INTEGER	No	BK	Item sequence in order	order_items_dataset.csv	1, 2, 3
customer_key	INTEGER	No	FK	→ dim_customer	Lookup	1001, 1002, 1003
product_key	INTEGER	No	FK	→ dim_product	Lookup	501, 502, 503
seller_key	INTEGER	No	FK	→ dim_seller	Lookup	201, 202, 203
purchase_date_key	INTEGER	No	FK	→ dim_date	Lookup	715, 716, 717
payment_type_key	INTEGER	No	FK	→ dim_payment_type	Lookup	1, 2, 3, 4
item_price	DECIMAL(10,2)	No	Measure	Item price in BRL	order_items_dataset.csv	58.90, 239.90, 199.99
freight_value	DECIMAL(10,2)	No	Measure	Shipping cost in BRL	order_items_dataset.csv	13.29, 0.00, 25.49
total_item_value	DECIMAL(10,2)	No	Measure	Price + freight	ETL Calculated	72.19, 239.90, 225.48
payment_value	DECIMAL(10,2)	No	Measure	Payment amount	payments_dataset.csv	72.19, 239.90, 225.48
quantity	INTEGER	Yes	Measure	Item quantity	order_items_dataset.csv	1, 2, 1

Business purpose: Detailed sales transactions for revenue and profitability analysis.
Grain: One row per order item (order_id + order_item_id).

FACT_DELIVERY_PERFORMANCE - Order Delivery Performance

Column Name	Data Type	Nullable	PK/FK	Description	Source	Sample Values
delivery_key	SERIAL	No	PK	Surrogate key	Generated	1, 2, 3...
order_id	VARCHAR (50)	No	BK/UK	Order identifier	orders_dataset.csv	e481f51cbdc54678b7cc49136f2d6af7
customer_key	INTEGER	No	FK	→ dim_customer	Lookup	1001, 1002, 1003
seller_key	INTEGER	No	FK	→ dim_seller	Lookup	201, 202, 203
order_date_key	INTEGER	No	FK	→ dim_date (order date)	Lookup	715, 716, 717
approved_date_key	INTEGER	Yes	FK	→ dim_date (approval date)	Lookup	716, 717, 718
delivered_date_key	INTEGER	Yes	FK	→ dim_date (delivery date)	Lookup	725, 726, 727
est_delivery_date_key	INTEGER	Yes	FK	→ dim_date (estimated delivery)	Lookup	724, 725, 726
order_status_key	INTEGER	No	FK	→ dim_order_status	Lookup	1, 2, 3, 4
order_value	DECIMAL(10,2)	No	Measure	Total order value in BRL	Aggregated from order_items	158.90, 439.90, 299.99
freight_value	DECIMAL(10,2)	No	Measure	Total shipping cost in BRL	Aggregated from order_items	23.29, 15.10, 35.49
item_count	INTEGER	No	Measure	Number of items in order	Count from order_items	1, 2, 3, 5
estimated_delivery_days	INTEGER	Yes	Measure	Est. delivery time (days)	ETL Calculated	7, 14, 21, 30

actual_delivery_days	INTEGER	Yes	Measure	Actual delivery time (days)	ETL Calculated	5, 12, 25, 35
delivery_delay_days	INTEGER	Yes	Measure	Delay vs estimate (days)	ETL Calculated	-2, -2, 4, 5
delivery_performance_score	DECIMAL(3,2)	Yes	Measure	Performance score (0.00-1.00)	ETL Calculated	1.00, 0.85, 0.60, 0.40
is_on_time	BOOLEAN	Yes	Flag	On-time delivery flag	ETL Calculated	TRUE, FALSE
is_delivered	BOOLEAN	Yes	Flag	Successfully delivered flag	ETL Calculated	TRUE, FALSE

Business purpose: Order fulfillment and delivery performance analysis.
Grain: One row per order (order_id).

FACT_CUSTOMER_REVIEWS - Customer Review Data

Column Name	Data Type	Nullable	PK/FK	Description	Source	Sample Values
review_key	SERIAL	No	PK	Surrogate key	Generated	1, 2, 3...
review_id	VARCHAR(50)	No	BK/UK	Review identifier	order_reviews_dataset.csv	7bc2406110b926393aa56f80a40eba40
order_id	VARCHAR(50)	No	BK	Order identifier	order_reviews_dataset.csv	73fc7af87114b39712e6da79b0a377eb
customer_key	INTEGER	No	FK	→ dim_customer	Lookup	1001, 1002, 1003
product_key	INTEGER	No	FK	→ dim_product	Lookup	501, 502, 503
seller_key	INTEGER	No	FK	→ dim_seller	Lookup	201, 202, 203
review_date_key	INTEGER	No	FK	→ dim_date (review creation)	Lookup	715, 716, 717
review_score	INTEGER	No	Measure	Review rating (1-5)	order_reviews_dataset.csv	1, 2, 3, 4, 5
review_score_category	VARCHAR(10)	Yes	Flag	Score classification	ETL Calculated	Poor, Fair, Good, Excellent
is_positive_review	BOOLEAN	Yes	Flag	Positive review flag (≥4)	ETL Calculated	TRUE, FALSE
is_negative_review	BOOLEAN	Yes	Flag	Negative review flag (≤2)	ETL Calculated	TRUE, FALSE
has_comment	BOOLEAN	Yes	Flag	Comment presence flag	ETL Calculated	TRUE, FALSE

Business purpose: Customer satisfaction and review sentiment analysis.

Grain: One row per review (review_id).

Source-to-Target Mapping

Dimension Loading

Source File	Source Column	ETL Transformation	Target Table	Target Column
Date Generation				
ETL Generated	Date range 2016-2020	Date calculations	dim_date	All columns
Geography				
geolocation_dataset.csv	geolocation_zip_code_prefix	Pad zeros + format	dim_geography	zip_code
geolocation_dataset.csv	geolocation_city	Direct mapping	dim_geography	city
geolocation_dataset.csv	geolocation_state	Direct mapping	dim_geography	state
ETL Logic	State-to-region mapping	Case statement	dim_geography	region
Customer				
customers_dataset.csv	customer_id	Direct mapping	dim_customer	customer_id
customers_dataset.csv	customer_unique_id	Direct mapping	dim_customer	customer_unique_id
customers_dataset.csv	customer_zip_code_prefix	Pad zeros + format	dim_customer	customer_zip_code
customers_dataset.csv	customer_city	Direct mapping	dim_customer	customer_city
customers_dataset.csv	customer_state	Direct mapping	dim_customer	customer_state
ETL Logic	State-to-region mapping	Case statement	dim_customer	customer_region
Product				
products_dataset.csv	product_id	Direct mapping	dim_product	product_id
products_dataset.csv	product_category_name	Direct mapping	dim_product	product_category_portuguese

category_translation.csv	product_category_name_english	Lookup translation	dim_product	product_category_english
products_dataset.csv	product_weight_g	Direct mapping	dim_product	product_weight_grams
ETL Calculation	length * width * height	Multiplication	dim_product	product_volume_cm3
ETL Logic	Volume-based classification	Case statement	dim_product	product_size_category

Fact Loading

Source File	Source Column	ETL Transformation	Target Table	Target Column
Sales Fact				
order_items_dataset.csv	order_id	Direct mapping	fact_sales	order_id
order_items_dataset.csv	order_item_id	Direct mapping	fact_sales	order_item_id
order_items_dataset.csv	price	Direct mapping	fact_sales	item_price
order_items_dataset.csv	freight_value	Direct mapping	fact_sales	freight_value
ETL Calculation	price + freight_value	Addition	fact_sales	total_item_value
orders_dataset.csv	order_purchase_timestamp	Date lookup	fact_sales	purchase_date_key
payments_dataset.csv	payment_type	Payment type lookup	fact_sales	payment_type_key
Delivery Fact				
orders_dataset.csv	order_id	Direct mapping	fact_delivery_performance	order_id
orders_dataset.csv	order_purchase_timestamp	Date lookup	fact_delivery_performance	order_date_key
orders_dataset.csv	order_approved_at	Date lookup	fact_delivery_performance	approved_date_key
orders_dataset.csv	order_delivered_customer_date	Date lookup	fact_delivery_performance	delivered_date_key
orders_dataset.csv	order_estimated_delivery_date	Date lookup	fact_delivery_performance	est_delivery_date_key
ETL Aggregation	SUM(price + freight) by order	Group by order_id	fact_delivery_performance	order_value
ETL Calculation	delivered_date - order_date	Date arithmetic	fact_delivery_performance	actual_delivery_days
Reviews Fact				
order_reviews_dataset.csv	review_id	Direct mapping	fact_customer_reviews	review_id

order_reviews_dataset.csv	order_id	Direct mapping	fact_customer_reviews	order_id
order_reviews_dataset.csv	review_score	Direct mapping	fact_customer_reviews	review_score
order_reviews_dataset.csv	review_creation_date	Date lookup	fact_customer_reviews	review_date_key
ETL Logic	Score categorization	Case statement	fact_customer_reviews	review_score_category
ETL Logic	review_score >= 4	Boolean logic	fact_customer_reviews	is_positive_review

Indexes and Performance Optimization

Dimension Indexes

Index Name	Table	Columns	Purpose
idx_dim_date_full_date	dim_date	full_date	Date lookup performance
idx_dim_date_year_month	dim_date	year_number, month_number	Time-based grouping
idx_dim_date_quarter	dim_date	quarter_number, year_number	Quarterly analysis
idx_dim_geography_state	dim_geography	state	Geographic filtering
idx_dim_geography_region	dim_geography	region	Regional analysis
idx_dim_customer_unique_id	dim_customer	customer_unique_id	Customer lookup
idx_dim_customer_current	dim_customer	is_current (WHERE TRUE)	SCD Type 2 current records
idx_dim_product_category_eng	dim_product	product_category_english	Category analysis
idx_dim_seller_state	dim_seller	seller_state	Seller geographic analysis

Fact Table Indexes

Index Name	Table	Columns	Purpose
idx_fact_sales_date_customer	fact_sales	purchase_date_key, customer_key	Time-series customer analysis
idx_fact_sales_product	fact_sales	product_key	Product performance queries
idx_fact_sales_seller	fact_sales	seller_key	Seller performance queries
idx_fact_sales_order_id	fact_sales	order_id	Order-based joins
idx_fact_delivery_performance_date_status	fact_delivery_performance	order_date_key, order_status_key	Delivery funnel analysis
idx_fact_delivery_on_time	fact_delivery_performance	is_on_time	Performance filtering
idx_fact_reviews_customer	fact_customer_reviews	customer_key	Customer satisfaction analysis
idx_fact_reviews_score	fact_customer_reviews	review_score	Review filtering

Constraints and Business Rules

Data Quality Constraints

Constraint Name	Table	Rule	Business Purpose
chk_sales_total_calc	fact_sales	total_item_value = item_price + freight_value	Ensure calculation accuracy
chk_delivery_dates	fact_delivery_performance	delivered_date >= order_date AND est_delivery_date >= order_date	Prevent illogical date sequences
chk_customer_dates	dim_customer	effective_date <= expiry_date	SCD Type 2 date integrity
uq_sales_grain	fact_sales	UNIQUE(order_id, order_item_id)	Enforce fact table grain
uq_geography_zip	dim_geography	UNIQUE(zip_code)	One record per ZIP code

Referential Integrity

Parent Table	Child Table	Foreign Key	Constraint
dim_customer	fact_sales	customer_key	NOT NULL
dim_product	fact_sales	product_key	NOT NULL
dim_seller	fact_sales	seller_key	NOT NULL
dim_date	fact_sales	purchase_date_key	NOT NULL
dim_payment_type	fact_sales	payment_type_key	NOT NULL
dim_customer	fact_delivery_performance	customer_key	NOT NULL
dim_date	fact_delivery_performance	order_date_key	NOT NULL
dim_order_status	fact_delivery_performance	order_status_key	NOT NULL
dim_customer	fact_customer_reviews	customer_key	NOT NULL
dim_product	fact_customer_reviews	product_key	NOT NULL
dim_date	fact_customer_reviews	review_date_key	NOT NULL

Business Rules and Calculations

Derived Measures

Calculation	Formula	Business Rule
Total Item Value	item_price + freight_value	Complete cost per item
Product Volume	length_cm * width_cm * height_cm	Physical size calculation
Delivery Delay	actual_delivery_days - estimated_delivery_days	Performance measurement
Review Category	CASE WHEN score <= 2 THEN 'Poor' WHEN score = 3 THEN 'Fair' WHEN score = 4 THEN 'Good' ELSE 'Excellent' END	Sentiment classification

Business Flags

Flag	Logic	Purpose
is_positive_review	review_score >= 4	Positive sentiment identification
is_negative_review	review_score <= 2	Negative sentiment identification
is_on_time	actual_delivery_days <= estimated_delivery_days	Delivery performance
is_weekend	day_of_week IN (1, 7)	Weekend sales analysis
has_complete_attributes	All key product attributes NOT NULL	Data quality flag

Summary

Source Systems: 9 CSV files from Olist Brazilian e-commerce dataset
ETL Tool: Pentaho Data Integration (PDI)
Target System: PostgreSQL Data Warehouse
Data Quality: Comprehensive validation and cleansing rules implemented
SCD Implementation: Type 1 (most dimensions) and Type 2 (customer dimension)