

# Data Model Documentation

## Dimension Tables

### 1. customer\_dim

**Purpose:** Customer master data dimension  
**Location:** s3://analytics-development/test\_data/customer\_dim/  
**Format:** Parquet with Snappy compression  
**Grain:** One row per unique customer

Column Name	Data Type	Description	Source	Business Rules
customer_id	BIGINT	Unique customer identifier	stg_customers_raw.Customer_ID	Primary key, not null
first_name	VARCHAR	Customer first name	stg_customers_raw.First	Trimmed, not null
last_name	VARCHAR	Customer last name	stg_customers_raw.Last	Trimmed, not null
age	INTEGER	Customer age	stg_customers_raw.Age	Converted from string, nullable
country	VARCHAR	Customer country	stg_customers_raw.Country	Trimmed, not null

#### Data Quality Rules

- customer\_id must be unique and not null
- Names must be trimmed and not empty
- Age must be valid integer
- Country must be valid and not empty

### 2. order\_dim

**Purpose:** Order details dimension  
**Location:** s3://analytics-development/test\_data/order\_dim/  
**Format:** Parquet with Snappy compression  
**Grain:** One row per order item

Column Name	Data Type	Description	Source	Business Rules
order_id	BIGINT	Unique order identifier	stg_orders_raw.order_id	Primary key, not null
item	VARCHAR	Product/item name	stg_orders_raw.item	Trimmed, not null
amount	DECIMAL(18,2)	Item amount	stg_orders_raw.amount	Must be positive, not null
customer_id	BIGINT	Customer reference	stg_orders_raw.customer_id	Foreign key to customer_dim

#### Data Quality Rules

- order\_id must be unique and not null
- item must be trimmed and not empty
- amount must be positive decimal value
- customer\_id must reference valid customer

### 3. shipping\_dim

**Purpose:** Shipping and logistics dimension

**Location:** s3://analytics-development/test\_data/shipping\_dim/

**Format:** Parquet with Snappy compression

**Grain:** One row per shipping record

Column Name	Data Type	Description	Source	Business Rules
shipping_id	VARCHAR	Unique shipping identifier	stg_shipping_raw.Shipping_ID	Primary key, not null
shipping_status	VARCHAR	Delivery status	stg_shipping_raw.Status	Trimmed, not null
customer_id	BIGINT	Customer reference	stg_shipping_raw.Customer_ID	Foreign key to customer_dim

#### Data Quality Rules

- shipping\_id must be unique and not null
- shipping\_status must be valid status value
- customer\_id must reference valid customer

# Fact Tables

## 1. order\_fact

**Purpose:** Order summary facts and metrics

**Location:** s3://analytics-development/test\_data/order\_fact/

**Format:** Parquet with Snappy compression

**Grain:** One row per order

Column Name	Data Type	Description	Source	Business Rules
order_id	BIGINT	Order identifier	order_dim.order_id	Primary key, not null
customer_id	BIGINT	Customer reference	order_dim.customer_id	Foreign key to customer_dim
items_count	INTEGER	Number of items in order	Calculated	Must be positive
order_total	DECIMAL(18,2)	Total order amount	Calculated	Sum of item amounts
is_customer_missing	BOOLEAN	Data quality flag	Calculated	True if customer_id is null

### Measures

- items\_count: Count of items per order
- order\_total: Sum of all item amounts per order

### Data Quality Rules

- order\_id must be unique and not null
- items\_count must be positive integer
- order\_total must be positive decimal
- Quality flag indicates data integrity issues

## Relationships

### Primary Relationships

1. customer\_dim.customer\_id ← order\_dim.customer\_id (1:Many)
2. customer\_dim.customer\_id ← shipping\_dim.customer\_id (1:Many)

3. **order\_dim.order\_id** → **order\_fact.order\_id** (1:1)
4. **customer\_dim.customer\_id** ← **order\_fact.customer\_id** (1:Many)

## Referential Integrity

- All foreign key relationships are maintained
  - Data quality flags identify orphaned records
  - Consistent data types across related tables
- 

## Data Lineage

### Source to Target Mapping

stg\_customers\_raw → customer\_dim

Customer\_ID → customer\_id (BIGINT)  
First → first\_name (VARCHAR, trimmed)  
Last → last\_name (VARCHAR, trimmed)  
Age → age (INTEGER, converted)  
Country → country (VARCHAR, trimmed)

stg\_orders\_raw → order\_dim → order\_fact

order\_id → order\_id (BIGINT)  
item → item (VARCHAR, trimmed)  
amount → amount (DECIMAL, converted)  
customer\_id → customer\_id (BIGINT)  
Aggregated to order\_fact metrics

stg\_shipping\_raw → shipping\_dim

Shipping\_ID → shipping\_id (VARCHAR)  
Status → shipping\_status (VARCHAR, trimmed)  
Customer\_ID → customer\_id (BIGINT)

## Business Rules

### Data Quality Standards

1. **Completeness:** All required fields must be populated
2. **Accuracy:** Data must be valid and consistent
3. **Consistency:** Data types and formats must be standardized
4. **Timeliness:** Data should be refreshed regularly
5. **Integrity:** Referential relationships must be maintained

## **Naming Conventions**

- Table names: {entity}\_dim for dimensions, {entity}\_fact for facts
- Column names: snake\_case format
- Data types: Consistent across related tables
- Compression: Snappy for optimal performance

## **Usage Guidelines**

### **Analytics Queries**

- Use dimension tables for filtering and grouping
- Use fact tables for measures and aggregations
- Join dimensions to facts for comprehensive analysis
- Leverage data quality flags for data validation

### **Assumed Reporting Requirements**

- Revenue and orders (total, by customer, by item)
- AOV (average order value), items per order
- Customer 360 (lifetime value)
- Shipping performance (delivered / failed / in-transit counts, delivery time)
- Reconciliations: source totals → warehouse totals
- Ad-hoc drilldowns (by customer, item, country)