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	BIGINT	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 (BIGINT)
    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)

Anomalies

No date columns in source data

- Fact tables (order_fact) are normally partitioned by a natural date field (e.g., order_date).
- The current source data does not include any date column (customer_created_date,order_date or shipment_date), so the fact table is created as non-partitioned.
- This limits incremental refresh and time-series reporting.