Star Schema (Printable Textbook Edition)

This guide teaches Star Schema from fundamentals through an end-to-end OLTP → ETL → OLAP build, including a 100K+ dataset generator and SCD Type 2 history tracking.

1. Dimensional Modeling Basics

Dimensional modeling separates analytic data into facts (events) and dimensions (context). The most common dimensional model is the Star Schema: one fact table connected to several dimension tables.

- **Fact table**: metrics at a declared grain (e.g., one row per order line).
- **Dimensions**: descriptive attributes used to filter/group results (product, customer, date, store).
- **Grain**: the meaning of one fact row; it must be defined before ETL begins.

2. OLTP vs OLAP (Why We Need Different Schemas)

- OLTP systems are optimized for many small writes and strict consistency (normalized tables).
- OLAP systems are optimized for scans, joins, and aggregations over large historical datasets.
- Star schemas reduce join complexity and speed up BI-style queries.

3. Example: E-Commerce Star Schema

A practical grain is **one row per order item**. The fact table records measures such as quantity and sales amount; dimensions describe the customer, product, store, and date.

```
FACT_SALES(date_key, customer_key, product_key, store_key, order_id, order_item_id, quantity, unit_pri

DIM_DATE(date_key, full_date, year, quarter, month, weekday_name, ...)
DIM_CUSTOMER(customer_key, customer_id, country, effective_date, expiry_date, is_current, ...)
DIM_PRODUCT(product_key, product_id, category, brand, ...)
DIM_STORE(store_key, store_id, region, ...)
```

4. ETL Ordering (Best Practice)

- Load or generate DIM_DATE.
- Load Type 1 dimensions (product, store).
- Load Type 2 dimensions (customer history).
- Load the fact table with surrogate key lookups.
- Validate referential integrity and row counts.

5. Slowly Changing Dimensions (SCD Type 2)

SCD Type 2 preserves history by expiring the current record and inserting a new version. This enables reporting that reflects attributes as they were at the time of the fact event.

```
-- Expire current row
UPDATE dim_customer
SET expiry_date = CURDATE() - INTERVAL 1 DAY,
    is_current  = FALSE
WHERE customer_id = 101 AND is_current = TRUE;

-- Insert new version
INSERT INTO dim_customer(customer_id, first_name, last_name, email, country, effective_date, expiry_da
VALUES (101, 'John', 'Smith', 'john@example.com', 'Canada', CURDATE(), '9999-12-31', TRUE);
```

6. Star vs Snowflake

- **Star**: denormalized dimensions, fewer joins, simpler and faster for BI.
- **Snowflake**: normalized dimensions, more joins, less redundancy, sometimes easier governance.
- Rule of thumb: prefer Star for dashboards and self-serve analytics.

7. OLAP Query Examples

```
-- Revenue by year
SELECT d.year, SUM(f.sales_amount) AS revenue
FROM fact_sales f
JOIN dim_date d ON d.date_key = f.date_key
GROUP BY d.year
ORDER BY d.year;

-- Top 10 products by revenue
SELECT p.product_name, SUM(f.sales_amount) AS revenue
FROM fact_sales f
JOIN dim_product p ON p.product_key = f.product_key
GROUP BY p.product_name
ORDER BY revenue DESC
LIMIT 10;
```

Appendix: What’s Included

- 01_oltp_100k_data_generator_mysql.sql (OLTP schema + data)
- 02_dw_star_schema_mysql.sql (DW star schema)
- 03_etl_oltp_to_dw_mysql.sql (ETL initial load)
- 04_olap_queries_star_schema.sql (query pack)
- 05b_lab_workbook_only_exercises_solutions.md (labs)
- 06_star_schema_teaching_slides.pptx (slides)
- 07_star_schema_textbook.pdf (this document)