

Technical Story for Data Engineer

Project Overview

Our task is to implement a Snowflake schema that will be the foundation of our sales data analysis. This schema will enable us to run efficient queries, maintain data integrity, and reduce redundancy.

Data Model

Fact Table: Order Fact Table

This is our central table, capturing all transactional data related to sales. Key columns include Order_ID, Customer_ID, Product_ID, Amount, and Shipping_ID.

Dimension Tables

1. **Customer Table:** Contains all customer-related details. Each customer is linked to a country in the Country Table.
2. **Country Table:** Normalizes country information to avoid repetition.
3. **Product Table:** Stores product information.
4. **Shipping Table:** Tracks the shipping status of each order.

Relationships

- **Customer Table ↔ Order Fact Table:** One-to-Many
- **Country Table ↔ Customer Table:** One-to-Many
- **Product Table ↔ Order Fact Table:** One-to-Many
- **Shipping Table ↔ Order Fact Table:** One-to-Many

ETL Process

1. **Extract:** Pull data from the source systems (CRM, ERP, Logistics).
2. **Transform:** Normalize the data and clean up inconsistencies.
3. **Load:** Load the data into the Snowflake schema, starting with dimension tables.

Performance Tips

- Consider indexing Customer_ID, Product_ID, and Shipping_ID in the fact table.
- Partition the fact table by date or Order_ID for better performance.