Sample Project: Designing and Implementing a Star Schema DataMart in Snowflake using Talend

Project Objective:

Build a **Star Schema** data mart model by creating fact tables and multiple-dimensional tables using **Talend in Snowflake**. This project will demonstrate how to extract, transform, and load (ETL) data from various sources into a Snowflake target database designed for analytical reporting.

Step-by-Step Breakdown:

1. Data Source Identification and Analysis

- Identify data sources: transactional data (sales, orders) for the **Fact table** and reference data (products, customers, time, and regions) for the **Dimension tables**.
- Define the data types, granularity, and relationships between the fact and dimension tables.

2. Star Schema Design

- **Fact Table**: Sales data including measures such as sales_amount, units_sold, and discounts.
- **Dimension Tables**:
    - **Product Dimension**: product_id, product_name, category, price
    - **Customer Dimension**: customer_id, customer_name, gender, region
    - **Time Dimension**: date_id, day, month, quarter, year
    - **Region Dimension**: region_id, region_name, country

3. ETL Process Development in Talend

- **3.1. Extraction**:
    - Use Talend connectors (tJDBCInput, tFileInputDelimited, tRESTClient) to pull data from source systems (databases, files, APIs).
    - Extract transactional sales data for the fact table and master data for dimensions.
- **3.2. Transformation**:
    - Use tMap to join and transform the raw data.

    - Apply necessary transformations: date parsing, lookups (e.g., to enrich transactional data with product, customer, and region details).
    - Ensure foreign key relationships between fact and dimension tables.
- **3.3. Loading into Snowflake**:
    - Load the dimension tables first using tSnowflakeOutput, ensuring all necessary columns (primary keys) are populated.
    - After the dimension tables are loaded, populate the fact table using tSnowflakeOutput, ensuring correct foreign key relationships are established.

4. Data Load Validation and Integrity Checks

- Validate the consistency of the foreign key relationships between the fact and dimension tables using tSchemaComplianceCheck or custom queries.
- Implement data load validation by checking row counts, data types, and constraints using Talend components (tAssert, tLogRow).

5. Data Quality Checks

- Implement data quality checks for missing or inconsistent data in the dimension tables (e.g., products without valid categories).
- Use Talend Data Quality components (tDqRule, tMatchGroup) to enforce business rules and ensure data accuracy.

6. Scheduling and Automation

- Schedule the Talend jobs for periodic data extraction and loading using **Talend Management Console (TMC)**.
- Automate error handling and logging using tDie and tLogCatcher for continuous monitoring.

Project Outcome:

- A fully implemented **Star Schema** data mart in **Snowflake** with a **Fact table** (Sales) and **Dimension tables** (Product, Customer, Time, and Region).
- The data will be transformed and validated through Talend and loaded into Snowflake for analytical reporting and BI tool integration.