Data Engineering Case Study: Sports Betting Data Warehouse

Background

Assume a fantasy world where Tipico is a sports betting platform that collects data on betting activities from users around the world. The platform allows both registered users and anonymous users to place bets on various sporting events. The company wants to build a data warehouse to enable efficient analytics on betting patterns, customer behavior, and sports performance.

As a data engineering candidate, your task is to design and implement a star schema data warehouse based on the raw data provided.

Raw Data Description

You have been provided with the following raw data files:

1. **raw_betting_activity.csv**: Contains all betting transactions, including both registered and anonymous users
2. **raw_matches.csv**: Contains information about sports matches
3. **raw_customers.csv**: Contains information about registered customers
4. **raw_sports.csv**: Contains reference data about different sports
5. **raw_leagues.csv**: Contains information about sports leagues

Your Task

1. **Data Exploration and Profiling**:

    - Analyze the raw data to understand its structure, quality, and relationships
    - Identify data quality issues and propose solutions

2. **Data Modeling**:

    - Design a star schema for the betting data warehouse
    - Define fact and dimension tables with appropriate granularity
    - Handle slowly changing dimensions where appropriate
    - Address the challenge of anonymous users in your design

3. **Data Transformation**:

    - Develop transformation logic to convert raw data into the star schema
    - Handle missing values, duplicates, and inconsistencies
    - Create surrogate keys for dimension tables
    - Implement type 1 or type 2 slowly changing dimensions as needed

4. **Implementation**:

    - Write SQL DDL statements to create the star schema tables

    - Write SQL or Python code to transform and load the data
    - Document your approach and any assumptions made

Expected Deliverables

1. **Star Schema Design**: Entity-relationship diagram showing fact and dimension tables
2. **Data Dictionary**: Description of all tables and columns in your star schema
3. **Transformation Logic**: SQL or Python code for ETL processes
4. **Data Quality Report**: Summary of data quality issues found and how they were addressed
5. **Sample Queries**: 3-5 analytical SQL queries that demonstrate the value of your star schema

Evaluation Criteria

Your solution will be evaluated based on:

1. **Correctness**: Does the star schema correctly model the domain?
2. **Completeness**: Are all requirements addressed?
3. **Data Quality**: How well are data quality issues handled?
4. **Performance**: Is the schema designed for efficient querying?
5. **Clarity**: Is the solution well-documented and easy to understand?

Final artifacts

1. Preferred languages to implement the code are Python and SQL.
2. You can provide your code as .py, txt or preferred option fully runnable Jupyter notebook.
3. Prepare a business friendly presentation to explain the implemented solution and the results of the following analytical queries:
    - Query 1: Daily Betting Volume by Sport
    - Query 2: Anonymous vs. Registered User Betting Patterns
    - Query 3: Match Popularity and Betting Performance
    - Query 4: Customer Segmentation by Betting Behavior
    - Query 5: Temporal Analysis of Betting Patterns

Hint

Use sqlite3 to create your database.

Good luck!