Project Part 3: Data Warehouse Design 100 points

Overview

Over the last week or so we have been talking about data warehouses: their purpose, their characteristics, and their design. In this fourth assignment you will take the ClassDojo relational schema and use it to create the design for a data warehouse, or perhaps more precisely, a data mart that conforms to an overall data warehouse architecture.

Read through the assignment first, especially the kind of deliverables specified at the end, before you begin the assignment. Please note that not every step will produce deliverables but may be needed as a prerequisite to produce a deliverable in a next step.

The OLTP Schema

The starting point for this assignment is the relational schema for the ClassDojo scenario (from part 1, not the smaller version of part 2). Please note that we added a district table that defines to which district a school belongs as the address was decomposed.

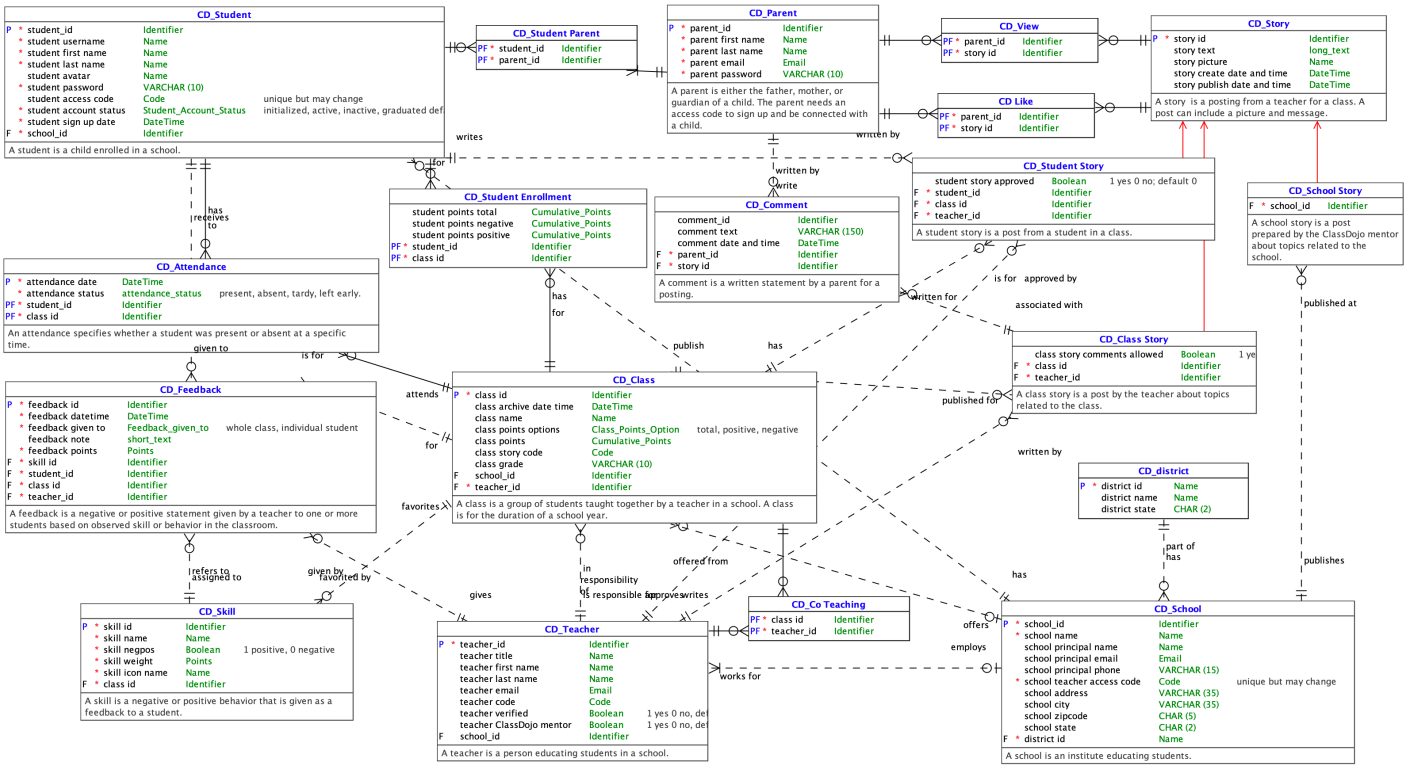

Figure 1 ClassDojo Schema

Project Part 3: Data Warehouse Design 100 points

The Data Warehouse Usage Requirements

The data warehouse team has already spent a good deal of time with end users and the following list shows the kinds of business questions that are most pressing.

1. List the number of feedbacks broken down by class and student for the school year 2020-2021 for one school. The output should list {classroom, student, number of feedback}.
2. List the number of positive and negative feedback broken down by teacher of one school in school year 2020-2021. The output should list {teacher, number of positive feedback, number of negative feedback}.
3. List the feedback points for one student for all the months in 2020. The output should list {month, class, feedback points}
4. List the number of feedbacks broken down by skill for one class for each academic quarter in academic year 2020-2021. The output should list {skill, quarter, number of feedbacks}
5. List the monthly attendance status for students in a class for school year 2020-2021. The output should list {student, month, status, number of attendance}
6. List the attendance status for students per school districts for schools in year 2020-2021. The output should list {district, status, number of attendance}

Design Tasks

Please note that the tasks below are designed to guide you through the star schema creation process. You will not hand in a deliverable for every task. Please see "Deliverables" after the description of the tasks.

Step #1 Identify Data Marts & Data Dimensions

Under normal circumstances, the first step would be to sketch the data warehouse bus architecture in the form of a table with columns representing dimensions and rows representing the key business processes that will each eventually be served by a data mart. In the case of the ClassDojo database, there are not many business processes that need to be supported – giving feedback, taking attendance, publishing stories – so the creation of such a table is not an exciting exercise:

Task #1: Given the questions above, identify the dimensions that your data warehouse is going to need.

Task #2: Define the dimensions. Since we are dealing with a single source of transaction data and a single data mart, there is no conforming of dimensions to be done. However, you should determine which attributes/columns belong in each dimension you identified in Task #1.

- you should create a new key, a surrogate key, for each dimension that is not any of the keys used in the existing "OLTP" system
- If classification hierarchies exist (e.g. day-week/month-year, or product-subcategory-category), de-normalize so that entire hierarchies exist within a single dimension. Hierarchies are often found where tables do not touch the fact directly and categorize another table, e.g., Order-Customer-Country-Subregion-Region, where country does not touch order directly.
- Once you have the fact tables, link them to all possible dimensions that match the grain.
- Certainly include attributes that are needed for the queries listed above. However, you may include additional descriptive attributes that are single-valued with respect to the dimension, if you think that they could be useful in future queries. Some attributes are not likely to be (e.g. user_password) and can be left out of the dimension. There's a bit of a gray area here.

Project Part 3: Data Warehouse Design 100 points

Step #3 Declare the Fact Table grain

You are now beginning to define your first star schema. Review the questions from the introduction. Decide what level of granularity is appropriate for each question.

Task #3: State the *level of granularity* of the fact table you are defining now. Define clearly *what exactly does a single row of the fact table* represent? For example, in the example given in class, for the fact table with the line-item level of granularity, "each row in the fact table represents the sale of a given product on a given order to a given customer on a given day."

Step #4 Choose the dimensions

Task #4: Indicate which dimensions will be included in the star schema whose fact table you identified in Step #3.

Step #5 Choose the facts

Task #5: Indicate the facts that will be included in the fact table identified in Task #3. Be sure that each fact matches the grain of the fact table. Note: There can exist some fact tables without facts. These are called "factless fact tables."

Step #6 Repeat

Repeat tasks #3-5 as many times as necessary to design all the star schemas needed to answer the questions in the introduction.

Queries

Write down the six SQL queries. These queries should provide answers to the questions defined in the Usage Requirements section.

Note: You do not need build tables on which to run your SQL. Just write out the SQL query. Of course, some students may choose to build tables to verify the query syntax.

Reflection

Write your reflection on the Discussion Board on Canvas. See instructions on Canvas.

Project Part 3: Data Warehouse Design 100 points

Deliverable

Please submit via Canvas the following documents:

- A pdf file:
  - A data model showing your star schemas drawn with Oracle Data Modeler. Your data model will include the fact and dimension tables and how they relate. Show clearly the name of each table and the columns it contains. You are **not** required to create or populate any tables.
- A pdf file:
  - For each fact table provide the following:
    - Granularity (e.g., Transaction, Transaction line, Snapshot)
    - Associated Dimensions (a list of all the dimensions connected to the fact table)
    - Grain (Definition of what exactly does a single row of the fact table represent)
  - The six SQL queries
- Your reflection in Canvas

Grading

Your grade will be computed in the following way:

Item	***Fraction of grade***
Star Schemas, Description of each fact table	60 points
Six Queries	30 points
Reflection	10 points