Assignment 3: Public Housing Inspections Star Schema

Enzo Prado

College of Professional Studies, Northeastern University

ALY 6030: Data Warehousing & SQL

Professor Adam Jones

May 11th, 2025

Assignment 3: Public Housing Inspections Star Schema

Question 1: Facts

When reviewing the dataset, it's clear that we're working with a single primary fact table centered around inspection events. From this table, two measurable facts stand out: the cost of inspection and the inspection score. The inspection cost represents a classic additive fact, meaning it can be summed across different levels like time periods, developments, or public housing agencies without any issue. On the other hand, the inspection score behaves more like a semi-additive fact. While it's useful to analyze these scores over time or across developments, they aren't meant to be totaled averages or trends are more appropriate here. Understanding the nature of these facts helps ensure that we use the data in a way that preserves its meaning and supports effective reporting.

Question 2: Dimensions

When it comes to dimensions in this dataset, I identified four key ones that help provide context to our inspection facts. First, we have a Date dimension based on the inspection date, which allows us to view trends over time. Next is the Public Housing Agency dimension, which captures the agency responsible for each inspection and serves as a key organizational layer. Then there's the Development dimension, which includes the name, address, city, and state of the inspected property, giving us location-specific insights. Finally, I'd consider a Location dimension that breaks down the address components to allow for geographic analysis at different levels, such as city or state. Together, these dimensions form the backbone of a well-structured dimensional model that supports flexible and insightful analysis.

Question 3: Fact Table Recommendation

To best support both detailed and summary-level reporting, I would recommend designing two types of fact tables. The first would be a transactional fact table, which stores the individual inspection records. This allows us to see each event in detail, when it happened, what it cost, and how it scored. The second would be a periodic snapshot fact table, specifically designed to track the total cost of inspections over time, such as monthly summaries per agency. This second table is especially useful for high-level reporting and trend analysis, which senior management often relies on. Using both of these fact tables ensures we can answer both day-to-day operational questions and broader strategic ones.

Question 4: Dimension Recommendation

One important challenge in managing inspection data is that public housing agencies often go through changes, sometimes merging, sometimes being renamed. To handle this kind of evolving information, I recommend using Slowly Changing Dimension Type 2 (SCD Type 2). This approach doesn't just overwrite the old data; instead, it creates a new record for each change while keeping the historical version intact. That way, we preserve the timeline of each agency's identity and ensure that past inspections remain accurately tied to the agency's name at the time of the event. This is especially important for long-term reporting and analysis, where accuracy and historical context really matter.

Question 5: SQL and CSV

For For the final task, senior management asked to identify which Public Housing Agencies (PHAs) experienced an increase in their inspection costs over time. To address this, I

created a SQL query that makes use of window functions, specifically LEAD(), to retrieve the

second most recent inspection date and cost for each agency. This approach allows us to directly

compare the most recent inspection to the one that immediately preceded it. To ensure proper

functionality in MySQL Workbench, I also converted the inspection date from text to a

standardized SQL DATE format using STR_TO_DATE(). The query then calculates both the

absolute change and the percentage change in inspection cost, filtering out any agencies that had

only one inspection or did not see a cost increase. By assigning row numbers to each inspection

per agency and selecting only those with row_num = 1, the final result lists one row per PHA,

ensuring clarity and avoiding duplicates. The output includes exactly the fields management

requested dates and costs of the two most recent inspections, along with the cost difference and

percentage change. This is implemented in the EnzoPrado_Assignment3.sql file, and the

resulting subset of data has been saved as EnzoPrado_Assignment3.csv easier viewing by

stakeholders.

References

LearnSQL.com. (n.d.). *The LAG Function and the LEAD Function in SQL*. Retrieved May 11,

2025, from https://learnsql.com/blog/lead-and-lag-functions-in-sql/