ALY 6030 Data Warehousing and SQL

Assignment 3 – Public Housing Inspections Star Schema

Shraddha Gupte

CPS - MPSA, Northeastern University

11th May, 2025

ALY6030 Data Warehousing and SQL

A. Introduction:

Today's discussion centers on analyzing a public housing inspection dataset to support senior management’s decision-making through data modeling and reporting. We'll explore how to identify and structure facts and dimensions from the dataset, apply dimensional modeling techniques such as slowly changing dimensions, and create SQL-based solutions to generate meaningful insights. Key highlights include building fact and dimension tables, summarizing inspection costs over time, and using analytical functions like ROW_NUMBER() and STR_TO_DATE() to track changes in inspection costs. The goal is to deliver a clear, actionable subset of data focused on PHAs with increased inspection costs tailored specifically for non-technical stakeholders.

1. How many facts are there in this dataset?

This dataset revolves around public housing inspections and contains one main fact: the Inspection event. The key measurable, quantitative data associated with this fact includes the INSPECTION_SCORE, which represents the performance or quality of the inspected public housing development, the COST_OF_INSPECTION_IN_DOLLARS, a financial metric indicating the cost to perform the inspection, and the INSPECTION_DATE, which serves as the temporal element of the inspection event. Together, these elements define the core fact table, Public_data, which captures the details of each inspection, including its performance, cost, and timing.

- **Which facts do you identify?**
- **For the facts that you identify, what type of facts are they?**

Based on the attributes in the dataset, two key facts can be identified. The first is the Inspection Score, which serves as a performance rating of the inspected public housing development. This is a numeric value within the 0–100 range and is classified as an Additive Fact because it can be aggregated across different dimensions such as developments, agencies, or time. For example, the average inspection score could be calculated per agency or per year.

The second key fact is the Cost of Inspection (in Dollars), which represents the dollar amount spent to conduct each inspection. This is also a numeric value (e.g., 250, 1200) and is an Additive Fact, as it can be summed across various dimensions such as time, agencies, or regions. An example of aggregation would be the total cost of inspections calculated by state or month. Both of these facts are additive, meaning they can be aggregated and analyzed in various ways to uncover insights.

2. **How many dimensions are there in this dataset?**
   - **Which dimensions do you identify?**

In this dataset, there are four key dimensions that provide context to the facts of inspection scores and inspection costs.

The first dimension is the **Date Dimension**, which represents the temporal aspect of the inspection, including attributes such as the inspection date, year, month, and day. This dimension enables analysis of inspection data over time, allowing for trends to be identified, such as changes in inspection scores or costs over different periods.

The second dimension is the **Public Housing Agency (PHA) Dimension**, which represents the agency that conducted the inspection. This dimension allows for comparisons of inspection scores and costs across different public housing agencies.

The third dimension is the **Development Dimension,** which captures the details of the housing development being inspected, including the development name, address, city, and state. This dimension provides insights into the specific developments being inspected and allows for comparisons across different locations.

Finally, the **Location Dimension** is another key dimension, encompassing the city and state of the inspected development. It enables geographic analysis of the inspections, helping to compare inspection results across different cities or states.

Together, these four dimensions—Date, Public Housing Agency, Development, and Location—provide the necessary context for detailed analysis and aggregation of the facts in the dataset.

3. **Senior management is interested in viewing the facts identified above, at both the inspection level, as well as a periodic summary of inspection costs for each month. Based on this context, if you were to store these data in a set of fact tables, which type (or types) of fact tables would you use and why?**

To address the needs of senior management for viewing the facts at both the inspection level and as a periodic summary, we would create two types of fact tables: a detailed (transactional) fact table for the inspection-level data and a summary fact table for the periodic aggregation of inspection costs (monthly summary).

a. Transactional Fact Table (Inspection Level) – refer sql code

The **Transactional Fact Table** would store each individual inspection event. This table would capture detailed metrics such as the inspection score and inspection cost, along with the foreign keys linking to the dimensions (Date, Public Housing Agency, Development, Location).
This table stores the inspection-level details such as inspection scores and costs. It also includes foreign keys (agency_id, development_id, and location_id) that link to the respective dimensions. It is a transactional fact table because it records each individual inspection event.
Transactional Fact Table (Inspection Level) is designed to capture granular data for each individual inspection, allowing for detailed analysis and drill-down into inspection details by development, agency, or location.

b. Summary Fact Table (Monthly Summary)

The **Summary Fact Table** would aggregate the inspection costs at the monthly level, providing a summary of the total inspection costs for each month. This type of fact table is typically used for high-level reporting, allowing senior management to track trends over time.
Here, the summary_id attribute aggregates the inspection costs and counts per month. This summary fact table would include metrics such as the total cost of inspections for each month (total_cost_of_inspection_in_dollars) and the number of inspections conducted in that month (total_inspection_count). The summary_id is a foreign key reference to the Date Dimension, where a record for each month exists.
Summary Fact Table is designed to aggregate data at a higher level, providing a summarized view of inspection costs and counts, which is useful for trend analysis and reporting purposes, especially for senior management who may be interested in viewing periodic performance.
By implementing both transactional and summary fact tables, we offer both detailed insights at the individual inspection level and high-level summaries for trends and comparisons over time.

4. **Senior management is also concerned with changes in the names and addresses of the public housing agency names since they tend to get merged with other agencies on a frequent basis. Based on this, how should we handle this slowly changing dimension?**

For handling changes in Public Housing Agency names and addresses, the best approach is to use a Slowly Changing Dimension Type 2 (SCD Type 2).
SCD Type 2 tracks the historical evolution of dimension data by creating a new row for each change, preserving full history. Since senior management is concerned with tracking mergers, name changes, and address updates over time, we want to retain previous values rather than overwrite them.
This way, when an agency name or address changes, we:

- Keep the old record for history,
- Insert a new row with the updated data,
- Associate each inspection event with the correct version of the agency as it existed at the time.

5. **Finally, senior management is interested in a subset of this data, for only those PHAs that saw an increase in the cost of performing an inspection in their jurisdiction… Naturally, this would also require you to filter out PHAs that only performed one inspection, so they've asked you to remove those as well.**

   The goal is to identify only those PHAs where the inspection cost has increased between the two most recent inspections. To ensure accuracy, we also need to exclude PHAs that have had only one inspection, as no comparison would be possible for them. Moreover, it's crucial to ensure that the inspection dates are in the correct format, as any discrepancies in date formatting can lead to errors when using window functions. Once the comparisons are made, the final output should include the following columns: PHA name, the most recent inspection date and cost, the second-most recent inspection date and cost, the change in cost, and the percentage change in cost. These steps are crucial to provide management with clear and actionable insights on cost trends for each PHA.

B. **Conclusion:**

   In summary, our exploration of the public housing inspection data focused on transforming raw records into a well-structured dimensional model to enable meaningful analysis. We identified key facts such as inspection scores and costs, and defined relevant dimensions including development, location, and time. We addressed slowly changing dimensions using Type 2 to preserve historical accuracy, and we applied SQL analytic functions to isolate public housing agencies with rising inspection costs. The resulting dataset offers management a clear, digestible view of cost trends, enabling data-driven decisions for resource allocation and oversight. This exercise demonstrates how dimensional modeling and SQL techniques can bridge the gap between raw data and strategic insight.