Northeastern University

Assignment 3: Public Housing Inspections Star Schema

ALY6030: Data Warehousing and SQL

Lecturer: Dr. Adam Jones

Nam H Tran

May 11, 2025

1. How many facts are there in this dataset?

- COST_OF_INSPECTION_IN_DOLLARS (Additive): Dollar cost incurred for an inspection
- INSPECTION_SCORE (Non-additive or Semi-additive): Quality score given during inspection (0–100)

2. How many dimensions are there in this dataset?

There are 7 clear dimensions in this dataset.

- INSPECTION_ID: Primary key, uniquely identifies each record (inspection event) in the dataset
- PUBLIC_HOUSING_AGENCY_NAME: Which housing agency performed the inspection
- INSPECTED_DEVELOPMENT_ADDRESS: Street address of the development
- INSPECTED_DEVELOPMENT_CITY: City where the development is located
- INSPECTED_DEVELOPMENT_STATE: State where the development is located
- INSPECTION_DATE: When the inspection occurred

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 meet the analytical needs of both inspection-level detail and monthly summarized costs, we should use two types of fact tables in our star schema.

- Transaction fact table:
  - Purpose: Stores detailed, event-level data—one row per inspection.
  - Grain: Each row represents a single inspection event.
  - Enables analysis at the most granular level (score per inspection, cost per development, …)
  - Columns:
    - INSPECTION_ID: primary key
    - DATE_KEY: foreign key to Time Dimension
    - PHA_KEY: foreign key to Agency Dimension
    - DEVELOPMENT_KEY: foreign key to Development Dimension
    - LOCATION_KEY: foreign key to Location Dimension
    - INSPECTION_SCORE
    - INSPECTION_COST
- Periodic snapshot fact table:
  - Purpose: Stores monthly summaries of inspection costs.
  - Grain: One row per PHA per month (or per development per month, depending on need).
  - Enables high-level trend and budget analysis for leadership.
  - Columns:

- MONTH_KEY (foreign key to Time Dimension, e.g., YYYY-MM)
- PHA_KEY
- TOTAL_INSPECTION_COST
- AVERAGE_INSPECTION_SCORE
- INSPECTION_COUNT

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 this context, I believe SCD Type 2 is the most appropriate for PHAs.

- Captures Historical Changes:
  - PHA names and addresses change over time due to mergers, rebranding, or relocation.
  - SCD Type 2 maintains a full history by creating a new row for each change, preserving the original record.
- Captures Historical Changes
  - PHA names and addresses change over time due to mergers, rebranding, or relocation.
  - SCD Type 2 maintains a full history by creating a new row for each change, preserving the original record.
- Prevents Data Loss or Overwriting
  - Type 1 (overwriting) would erase history, which is unacceptable when tracking trends across organizational changes.
  - Type 3 only keeps limited history (e.g., previous value), which isn't sufficient if multiple changes happen.