**Public Housing Inspections Star Schema**

Arun Radhakrishnan (002933307)

ALY6030: Data Warehousing and SQL, Northeastern University

Dr. Harpreet Sharma

21st March 2024

**Public Housing Inspections Star Schema**

With a focus on public housing inspections in the US, we undertake a thorough investigation of dimensional modelling and data analysis in this project. The foundation of our analysis is the dataset "public\_housing\_inspection," which includes vital elements like the names of Public Housing Agencies (PHAs), development specifics, and geographic data in addition to important information like inspection results and related expenses. Our overall goal is to create a dimensional model that is effective, takes mergers and name changes into account, and satisfies senior management's two distinct analytical requirements. This entails determining the important facts and dimensions using a suggested method that combines a Periodic Snapshot Fact Table with an Inspection-Level Fact Table. By using this tactical approach, we want to do more than just meet the needs for the current dataset.

1. **Number and Type of Facts:**

A item of data that we wish to examine is referred to as a "fact" in the dataset that has been provided. A fact table in the context of a dimensional model usually includes quantitative measures that we wish to compile and examine. Let's look at the information in the dataset:

**Number of Facts:** The number of rows in the Inspection Fact Table and the number of facts in this dataset match. The facts are made up of the individual inspection events represented by each row.

**Identified Facts:** Two important numerical measurements found in the dataset can be regarded as facts:

**COST\_OF\_INSPECTION\_IN\_DOLLARS:** This is the amount that is spent on each inspection. It is an additive fact since it is a quantitative measure that can be added up.

* **INSPECTION\_SCORE:**

This is an indication of each inspection's evaluation score. Although raw scores might not be additive in and of themselves, they can be combined using functions such as minimum, maximum, or average, which turns them into a semi-additive fact.

**Dimensions in the Dataset**

Dimensions are descriptive characteristics that provide the facts context in the dimensional model. The features that arrange and classify the data are either textual or categorical in nature. Let us ascertain and expound upon the dimensions present in the dataset:

* **PUBLIC\_HOUSING\_AGENCY\_NAME:**

Inspections are categorised in this dimension according to the names of public housing agencies. It reveals which organisation is in charge of each inspection.

* **INSPECTED\_DEVELOPMENT\_NAME:**

The names of the developments being examined are used to classify inspections in this dimension. It facilitates comprehension of the particular topics or tasks being looked at.

* **INSPECTED\_DEVELOPMENT\_LOCATION (combining city, state, and address):**

Although inspected\_development\_address may be considered a detailed attribute, a more comprehensive view is obtained by combining inspected\_development\_city, inspected\_development\_state, and inspected\_development\_address into a single dimension known as inspected\_development\_location.

* **INSPECTED\_DEVELOPMENT\_CITY:**

The dimension INSPECTED\_DEVELOPMENT\_CITY classifies inspections according to the cities in which the developments are situated. It provides information about how inspections are distributed among various cities.

* **INSPECTED\_DEVELOPMENT\_STATE:**

This dimension organises inspections according to the states in which the developments are situated, just like the city dimension does. It offers a more comprehensive geographic backdrop.

3) A dual-fact table technique is suggested to fulfil the senior management's desire for both periodic summaries of inspection expenses and comprehensive inspection-level data. With features like InspectionID, PHAID, DevelopmentID, InspectionDate, InspectionScore, and InspectionCost, the Inspection-Level Fact Table is made with granularity in mind. It maintains individual inspection data. This table enables flexible exploration of individual inspection events, auditing, and in-depth analysis. Parallel to this, the Summary Fact Table employs a periodic snapshot technique, using properties like SnapshotDate, PHAID, DevelopmentID, TotalInspections, TotalCost, and AverageScore to summarise inspection expenses on a monthly basis. Senior management reporting is streamlined by this table, which also offers high-level insights into monthly patterns and facilitates effective performance for periodic analyses. When these fact tables are integrated, a thorough solution that addresses both particular inspection details is guaranteed.

4) To handle this slowly changing dimension, a Type 2 Slowly Changing Dimension (SCD2) technique would probably be used in the context of changes in the names and addresses of public housing organisations, which often merge with other agencies.

**Justification:**

Changes to names and addresses can be tracked since Type 2 SCD keeps a historical record of changes over time.

With every modification, it adds new records, establishing a versioning system that documents the dimension's development.

**Preserving Historical Information:**

Preserving the historical context is essential for reporting and analysis when public housing agencies consolidate and change names and addresses.

In order to preserve a history of the agency names and addresses, Type 2 SCD makes sure that every change in name or address results in a new record.

**Flexibility for Analysis:**

The flexibility of Type 2 SCD allows for the analysis of data as it was at particular times. Understanding trends and patterns pertaining to public housing organisations can benefit from this.

**Auditing and Compliance:**

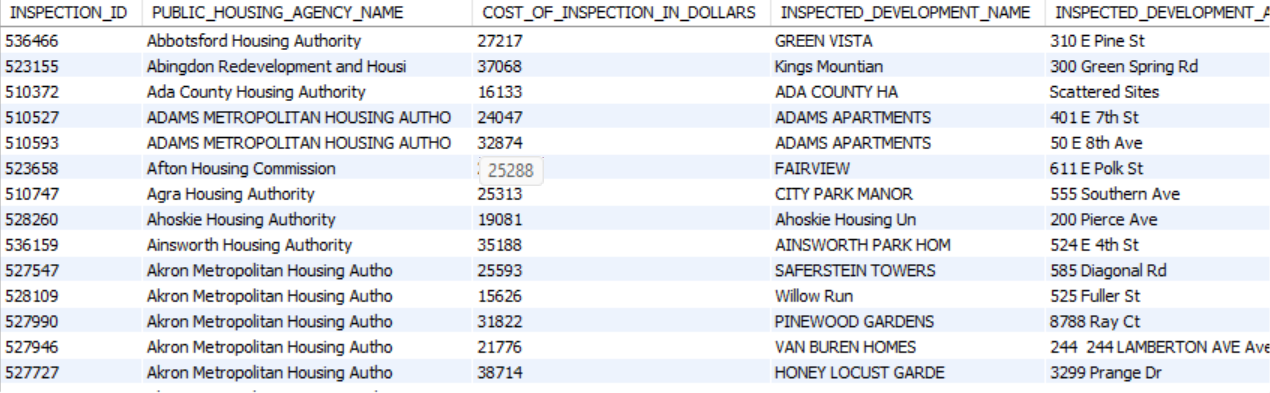
This method facilitates transparency in data evolution by maintaining an extensive record of changes, which satisfies auditing and compliance needs.

**Balancing Performance:**

Type 2 ensures speed in querying and reporting by striking a compromise between performance considerations and extensive historical recording, even though it adds more records.   
To sum up, by selecting a Type 2 Slowly Changing Dimension strategy, the data warehouse would be able to reliably record and handle alterations in public housing agency names and addresses over time, meeting the analytical needs of senior management with a historical view.

5)

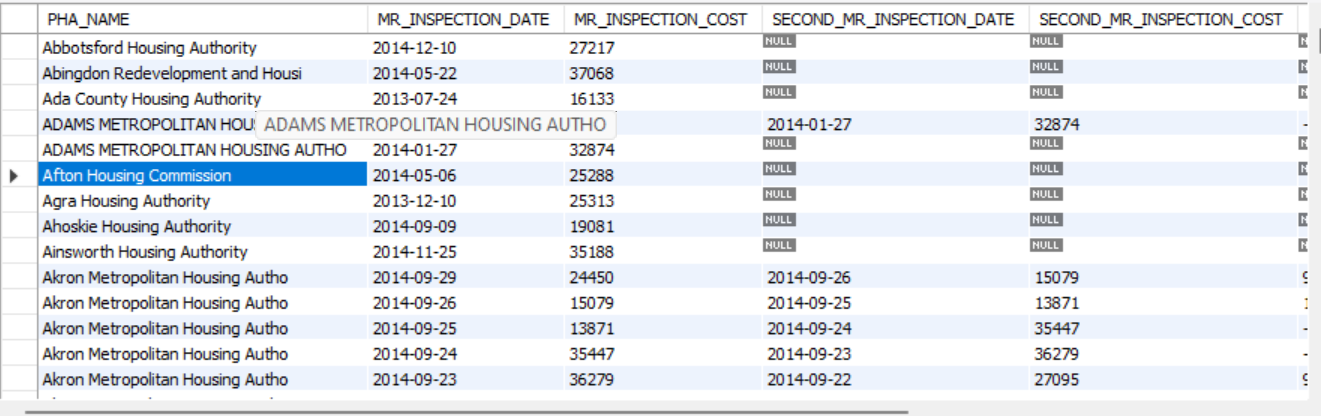
In Figure 1, The dataset has been successfully imported in the manner mentioned into our SQL workbench. The dataset can be seen in the screenshot below.



**Figure 1: Public housing Dataset**

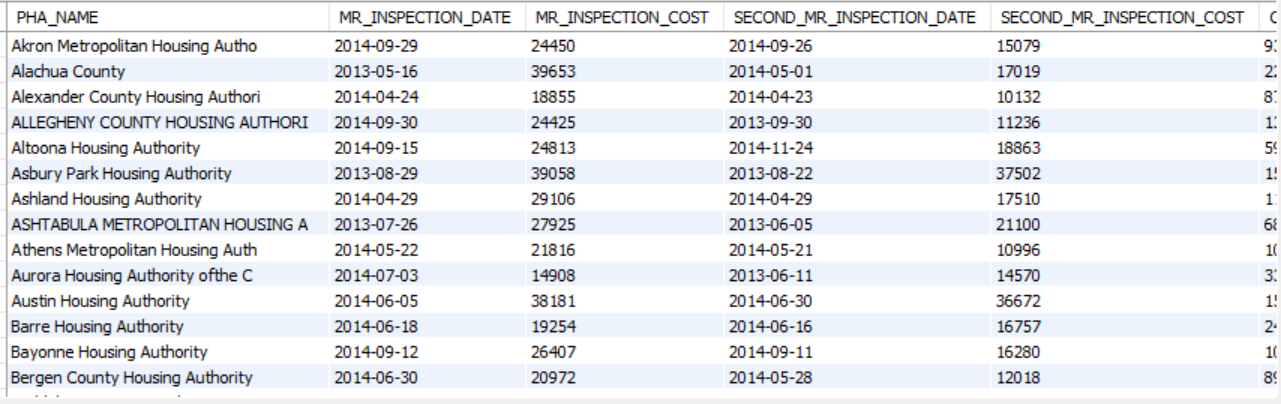
The dataset has been imported into the specified database in Figure 2, which corresponds to the dataset importation procedure. In compliance with the guidelines that were supplied, a question has been formulated using the lead function. The created query, as stated in the question, is intended to find public housing agencies that have seen an increase in the cost of doing inspections in their respective locations. By identifying and highlighting cases where inspection expenses have increased, this analytical method seeks to shed information on the financial dynamics of public housing organisations.

Top of Form



**Figure 2: Dataset with Null values**

In Figure 3, The careful procedure of removing noise and null values from the dataset shows how well-refined our output is. As instructed by the inquiry, we have limited our attention to public housing agencies (PHAs) that have experienced an increase in inspection expenses expressed in monetary terms. Moreover, in order to improve the accuracy of our analysis, every PHA is represented singularly and duplicate-free, adhering to the guidelines. Interestingly, PHAs that carried out just one examination were purposefully left out of the improved dataset. This customised strategy guarantees a more targeted and significant investigation of PHAs that saw a noticeable rise in the inspection cost, which helps to produce a more understandable and perceptive analytical result.



**Figure 3: Dataset with Noise Eliminated.**

**Reference:**

* Richardson, B. (2019, June 27). How to use Window functions in SQL Server. SQL Shack - Articles About Database Auditing, Server Performance, Data Recovery, and More. <https://www.sqlshack.com/use-window-functions-sql-server/>
* Joining a table value function to a MSSQL query. (n.d.). Stack Overflow. <https://stackoverflow.com/questions/699318/joining-a-table-value-function-to-a-mssql-query>
* Markingmyname. (2023, July 27). LEAD (Transact-SQL) - SQL Server. Microsoft Learn. https://learn.microsoft.com/en-us/sql/t-sql/functions/lead-transact-sql?view=sql-server-ver16