Module 5: Public Housing Inspections Star Schema

Mansi Ved Patel (NUID 002459952)

College of Professional Studies, Northeastern University

ALY 6030: Course Name: Data Warehousing and SQL

Professor Adam Jones

Date: 30 Nov 2024

1. How many facts are there in this dataset?

The dataset contains the following facts:

- **Inspection Score**: The score assigned to a development during an inspection, expressed as a percentage or ratio from 0 to 100.
- **Inspection Cost**: The cost incurred for performing the inspection (in dollars).

So, **there are two facts**: **Inspection Score** and **Inspection Cost**.

2. What type of facts are they?

- **Inspection Score**: This is a **measure** that provides a quantitative evaluation of the development.
- **Inspection Cost**: This is also a **measure**, representing the monetary value spent on performing each inspection.

3. How many dimensions are there in this dataset?

There are **three dimensions** in the dataset:

- **PHA Name**: The name of the Public Housing Agency conducting the inspection.
- **Development Name & Address**: The specific development that is being inspected, which includes its address.
- **Inspection Date**: The date on which the inspection was conducted.

4. Which dimensions do you identify?

The identified dimensions are:

- **PHA Name**
- **Development Name & Address**
- **Inspection Date**

5. Which type(s) of fact tables would you use and why?

Based on the context provided, the two types of fact tables that might be used are:

- **Transactional Fact Table**: This table will record facts at the **inspection level**. It includes the individual inspection scores and costs for each development, tied to specific inspection dates.

The schema would include references to the dimensions (PHA, Development, Date) and the facts (Inspection Score and Inspection Cost).

- **Snapshot Fact Table**: For the periodic summary (monthly aggregation), a snapshot fact table can be used to store summarized data, such as the total inspection cost for each month. The fact here would be the monthly sum or average of inspection costs, and the dimension would be **Month**.

6. How should we handle the Slowly Changing Dimension (SCD) for PHA Names and Addresses?

Since PHA names and addresses change due to mergers and other factors, **SCD Type 2** should be used. In Type 2 SCD, historical data is preserved by creating a new record each time an update occurs, with an effective date range. This way, we can track the history of changes in PHA names and addresses while preserving historical data.

7. SQL Query for the Requested Analysis

Here's the SQL query:

```sql
-- Query to calculate changes in inspection costs
WITH Inspection_Costs AS (
    SELECT
        PUBLIC_HOUSING_AGENCY_NAME,
        INSPECTION_DATE,
        COST_OF_INSPECTION_IN_DOLLARS
    FROM public_housing_inspections.public_housing_inspection_data_corrected
),
Inspection_Pairs AS (
    SELECT
        a.PUBLIC_HOUSING_AGENCY_NAME,
        a.INSPECTION_DATE AS MR_INSPECTION_DATE,
        a.COST_OF_INSPECTION_IN_DOLLARS AS MR_INSPECTION_COST,
        b.INSPECTION_DATE AS SECOND_MR_INSPECTION_DATE,
        b.COST_OF_INSPECTION_IN_DOLLARS AS SECOND_MR_INSPECTION_COST
    FROM Inspection_Costs a
    JOIN Inspection_Costs b
        ON a.PUBLIC_HOUSING_AGENCY_NAME = b.PUBLIC_HOUSING_AGENCY_NAME
        AND a.INSPECTION_DATE > b.INSPECTION_DATE
),
```

```
Filtered_Inspection_Pairs AS (
    SELECT
        PUBLIC_HOUSING_AGENCY_NAME,
        MR_INSPECTION_DATE,
        MR_INSPECTION_COST,
        SECOND_MR_INSPECTION_DATE,
        SECOND_MR_INSPECTION_COST,
        (MR_INSPECTION_COST - SECOND_MR_INSPECTION_COST) AS CHANGE_IN_COST,
        ((MR_INSPECTION_COST - SECOND_MR_INSPECTION_COST) / SECOND_MR_INSPECTION_COST) * 100 AS PERCENT_CHANGE_IN_COST
    FROM Inspection_Pairs
    WHERE MR_INSPECTION_COST > SECOND_MR_INSPECTION_COST
)
SELECT DISTINCT
    PUBLIC_HOUSING_AGENCY_NAME,
    MR_INSPECTION_DATE,
    MR_INSPECTION_COST,
    SECOND_MR_INSPECTION_DATE,
    SECOND_MR_INSPECTION_COST,
    CHANGE_IN_COST,
    PERCENT_CHANGE_IN_COST
FROM Filtered_Inspection_Pairs;
```

Here is the output:

PUBLIC_HOUSING_AGENCY_NAME	MR_INSPECTION_DATE	MR_INSPECTION_COST	SECOND_MR_INSPECTION_DATE	SECOND_MR_INSPECTION_COST	CHANGE_IN_COST	PERCENT_CHANGE_IN_COST
Akron Metropolitan Housing Autho	2014-10-09 00:00:00	25593	2014-10-08 00:00:00	15626	9967	63.7847
Akron Metropolitan Housing Autho	2014-10-07 00:00:00	31822	2014-10-06 00:00:00	21776	10046	46.1334
Akron Metropolitan Housing Autho	2014-10-09 00:00:00	25593	2014-10-06 00:00:00	21776	3817	17.5285
Akron Metropolitan Housing Autho	2014-10-03 00:00:00	38714	2014-09-29 00:00:00	24450	14264	58.3395
Akron Metropolitan Housing Autho	2014-10-07 00:00:00	31822	2014-09-29 00:00:00	24450	7372	30.1513
Akron Metropolitan Housing Autho	2014-10-09 00:00:00	25593	2014-09-29 00:00:00	24450	1143	4.6748
Akron Metropolitan Housing Autho	2014-09-29 00:00:00	24450	2014-09-26 00:00:00	15079	9371	62.1460
Akron Metropolitan Housing Autho	2014-10-03 00:00:00	38714	2014-09-26 00:00:00	15079	23635	156.7412
Akron Metropolitan Housing Autho	2014-10-06 00:00:00	21776	2014-09-26 00:00:00	15079	6697	44.4128
Akron Metropolitan Housing Autho	2014-10-07 00:00:00	31822	2014-09-26 00:00:00	15079	16743	111.0352
Akron Metropolitan Housing Autho	2014-10-08 00:00:00	15626	2014-09-26 00:00:00	15079	547	3.6276