ALY 6030 Data warehousing and SQL

Assignment - Public Housing Inspections Star Schema

Author: Qing Shen

1. Facts in a dimensional model are typically numerical measures that can be analyzed. In this dataset, there are **two facts**:

   COST_OF_INSPECTION_IN_DOLLARS – This is an **additive fact**. It can be summed across different dimensions.

   INSPECTION_SCORE – This is a **semi-additive fact**, meaning it can be averaged or analyzed over some dimensions.

2. Based on the dataset, there are **six dimensions**. They are PUBLIC_HOUSING_AGENCY_NAME, INSPECTED_DEVELOPMENT_NAME, INSPECTED_DEVELOPMENT_ADDRESS, INSPECTED_DEVELOPMENT_CITY, INSPECTED_DEVELOPMENT_STATE, and INSPECTION_DATE.

3. Senior management needs to view inspection-level data and a periodic summary of inspection costs for each month. I plan to use two types of fact tables:

   - **Transactional Fact Table** – This table is used to record individual business transactions, where each row represents a single inspection record.

   - **Periodic Snapshot Fact Table** – This table stores aggregated business data over a specific period (e.g., daily, monthly). Each row represents the summary of inspection costs for a PHA within a given month.

4. Since PHA names and addresses change frequently due to mergers, we need to implement a Slowly Changing Dimension (SCD) strategy that preserves historical data while tracking changes effectively. The best choice is **SCD Type 2**, because this method tracks historical data by creating multiple records.

5. To create the file required by senior management from the existing data, first, create the Schema and Table in MySQL Workbench, then import the data into the table, as shown in the image.

```
-- Assignment — Public Housing Inspections Star Schema
-- Create Schema
CREATE DATABASE IF NOT EXISTS public_housing;
USE public_housing;

-- Create table
CREATE TABLE public_housing_inspection_data (
    INSPECTION_ID INT PRIMARY KEY,
    PUBLIC_HOUSING_AGENCY_NAME VARCHAR(255),
    COST_OF_INSPECTION_IN_DOLLARS INT,
    INSPECTED_DEVELOPMENT_NAME VARCHAR(255),
    INSPECTED_DEVELOPMENT_ADDRESS VARCHAR(255),
    INSPECTED_DEVELOPMENT_CITY VARCHAR(255),
    INSPECTED_DEVELOPMENT_STATE VARCHAR(10),
    INSPECTION_DATE VARCHAR(10),
    INSPECTION_SCORE INT
);
```

Senior management requires an analysis of PHAs that experienced an increase in inspection costs. The goal is to identify PHAs with rising costs and generate a structured dataset containing only relevant records. To achieve the requested analysis, the SQL query consists of the following steps:

1) Convert INSPECTION_DATE from text format to DATE type.

2) Use the LAG() function to retrieve the second most recent inspection date and cost for each PHA.

3) Compute the change in inspection costs (CHANGE_IN_COST) and the percentage increase (PERCENT_CHANGE_IN_COST).

4) Filter out PHAs that only had one inspection, ensuring that every PHA in the result has both a most recent and second most recent inspection record.

5) Filter the dataset to retain only PHAs where the most recent inspection cost is higher than the previous one.

6) Ensure that each PHA appears only once in the output by selecting only the most recent inspection per PHA.

7) Sort the final results by the highest percentage increase in inspection cost, providing management with insights into the PHAs experiencing the most significant cost growth.

The SQL queries:

```sql
WITH Inspection_MR AS (
    SELECT
        PUBLIC_HOUSING_AGENCY_NAME AS PHA_NAME,
        STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y') AS INSPECTION_DATE, -- Convert dates from TEXT to Date format
        COST_OF_INSPECTION_IN_DOLLARS AS MR_INSPECTION_COST, -- The most recent inspection cost
        -- Get the previous inspection date
        LAG(STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y'))
        OVER (
            PARTITION BY PUBLIC_HOUSING_AGENCY_NAME -- Group by PHA
            ORDER BY INSPECTION_DATE ASC -- Order by date in asc order so LAG() retrieves the earlier inspection
        ) AS SECOND_MR_INSPECTION_DATE,
        -- Get the previous inspection cost
        LAG(COST_OF_INSPECTION_IN_DOLLARS)
        OVER (
            PARTITION BY PUBLIC_HOUSING_AGENCY_NAME
            ORDER BY INSPECTION_DATE ASC
        ) AS SECOND_MR_INSPECTION_COST,
        -- Rank each PHA's inspections from most recent to oldest
        ROW_NUMBER()
        OVER (
            PARTITION BY PUBLIC_HOUSING_AGENCY_NAME
            -- Order by date in descending order so the most recent inspection is ranked 1
            ORDER BY STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y') DESC
        ) AS rn
    FROM public_housing_inspection_data
)

-- Select the required columns for final output
SELECT
    PHA_NAME,
    INSPECTION_DATE AS 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, -- Calculate cost difference
    ((MR_INSPECTION_COST - SECOND_MR_INSPECTION_COST) / SECOND_MR_INSPECTION_COST) * 100 AS PERCENT_CHANGE_IN_COST
FROM Inspection_MR
WHERE SECOND_MR_INSPECTION_COST IS NOT NULL  -- filter out PHAs that only performed one inspection
AND (MR_INSPECTION_COST - SECOND_MR_INSPECTION_COST) > 0 -- Only PHAs with increased costs are retained
AND rn = 1 -- Only keep the most recent inspection per PHA
ORDER BY PERCENT_CHANGE_IN_COST DESC; -- Sort by percentage increase in cost, highest first
```

A subset of the generated output is shown below.

PHA_NAME	MR_INSPECTION_DATE	MR_INSPECTION_COST	SECOND_MR_INSPECTION_DATE	SECOND_MR_INSPECTION_COST	CHANGE_IN_COST	PERCENT_CHANGE_IN_COST
Corpus Christi Housing Authority	2014-12-17	38669	2014-12-16	10185	28484	279.6662
Edison Housing Authority	2013-07-30	37351	2013-07-24	10103	27248	269.7021
Ithaca Housing Authority	2014-12-29	36958	2013-11-18	10335	26623	257.6004
COUNTY OF MONTEREY HSG AUTH	2014-09-05	37768	2014-08-04	12204	25564	209.4723
Housing Authority of Joliet	2013-08-22	38887	2013-08-19	12982	25905	199.5455
Housing Authority of Stanton	2014-06-17	34994	2014-05-07	11879	23115	194.5871
St. Clair County Housing Authori	2014-10-09	36561	2014-10-08	12475	24086	193.0741
HA City of Everett	2014-06-11	36090	2014-05-13	12424	23666	190.4862
Des Moines Municipal Housing Age	2014-07-17	39273	2014-07-11	14100	25173	178.5319
TAMPA HOUSING AUTHORITY	2014-07-02	33481	2014-07-01	12130	21351	176.0181
ASHTABULA METROPOLITAN HOU...	2014-06-03	37948	2014-04-24	13920	24028	172.6149
HA TROY	2014-08-05	39158	2014-05-06	14428	24730	171.4028
Irvington Housing Authority	2014-06-23	38981	2013-06-21	14488	24493	169.0572
Zanesville Metropolitan Housing	2014-09-23	28684	2014-09-22	10702	17982	168.0247
Brockton Housing Authority	2014-05-20	36225	2013-03-12	14155	22070	155.9166
Fayetteville Metropolitan Housin	2015-01-29	36300	2015-01-28	14237	22063	154.9694
Columbus Metropolitan Housing Au	2015-02-02	28892	2012-11-06	11436	17456	152.6408