



Module 3:

Assignment 3: Public Housing Inspections Star Schema

ALY6030-Data Warehousing & SQL

Submitted by: Trupal Chaudhari

Submitted to: Adam Jones

Nov 22, 2024

Introduction

The assignment will delve into public housing inspections in the United States. I am tasked with delivering a dimensional model for the inspection data, and providing key insights to Senior Management. The data is a flat file of inspections that includes but is not limited to the PHA name, the name and address of the development being inspected, the inspection date. The dataset includes a ratio or percentage score for the inspection between 0 and 100, and the cost associated with each inspection measured in dollars (\$).

Question 1

There are two in the dataset, one is "COST_OF_INSPECTION_IN_DOLLARS", another is "INSPECTION_SCORE".

The attribute "COST_OF_INSPECTION_IN_DOLLARS" describes the dollar amount associated with the cost of inspection for each individual INSPECTION_ID. "INSPECTION_SCORE" describes the score given to each INSPECTION_ID at the time of inspection processes.

The "COST_OF_INSPECTION_IN_DOLLARS" is additive fact because it can be aggregated or summed across any one of the dimensions linked to the fact tableNortheastern.

That means we can roll up or summarize the cost of the inspections by different dimensions, including time by month or by year, by location by region or city, or any other relevant dimension, without any loss of data integrity.

Well, the "INSPECTION_SCORE" is a non-additive fact because its capability to be aggregated or summed across certain dimensions but not all.

Question 2

There are seven dimensions. Since primary key also consider as dimensions, so the dimensions are: Inspection_id, Public_Housing_Agency_Name, Inspected_Development_Nname, Inspected_Development_Address, Inspected_Development_City, Inspected_Development__S, and Inspection_Date.

Question 3

To effectively store the data in a set of fact tables for the given context, we would likely use two types of fact tables:

The Inspection Fact Table comprises comprehensive details regarding each inspection, including the inspection ID, date, score, and inspection cost. Each row within this table represents a singular inspection event. This type of fact table enables senior management to access detailed insights into each inspection at an individual level.

Periodic Summary Fact Table: It serves to consolidate inspection costs on a monthly basis, featuring aggregated metrics like total inspection expenses and potentially the average inspection cost per month. Each row in this table represents the inspection costs for a particular month.

Periodic Summary fact table allows senior management to analyze inspection costs over time at a higher level of granularity. By using both types of fact tables, senior management can view inspection data at both the detailed inspection level and the summarized monthly level, providing them with comprehensive insights into inspection activities and costs.

Question 4

Slowly Changing Dimension-SCD would be the best approach. Type 2 SCD enables the capture of historical data where a new row is added in the dimension table in case of any change, and the previous rows are kept as such. This captures all types of changes and retains the history for analysis. In this way, Type 2 SCD captures the history of the change in the public housing agencies over time. Analysis of trends in the agency names and addresses is atomized. Historical data is always intact and accessible for reporting or decision-making whenever needed through the use of surrogate keys. In all, Type 2 SCD meets the challenge of solving the dynamism problem in data from public housing agencies and at the same time meeting the requirements for historical analysis and preservation of data.

Type 3 SCD may not be the most suitable type in this case, as it will store only the immediately preceding value along with the current one. If there are more than one changes in the agency name or address, only the most recent and one previously kept value would be stored, older changes being overwritten. This could result in the loss of historical data and reduce the ability to analyse the trend over time comprehensively.

Question 5

Firstly, we use this query to shows out the information about PUBLIC_HOUSING_AGENCY_NAME, INSPECTION_DATE, COST_OF_INSPECTION_IN_DOLLARS from table ‘public_housing_inspection_data’.

```
-  
4  #  
5 • SELECT PUBLIC_HOUSING_AGENCY_NAME, INSPECTION_DATE, COST_OF_INSPECTION_IN_DOLLARS  
6   FROM public_housing_inspection_data;  
7
```

Result Grid Filter Rows: Search Export: Fetch rows:

PUBLIC_HOUSING_AGENCY_NAME	INSPECTION_DATE	COST_OF_INSPECTION_IN_DOLLARS
▶ Abbotsford Housing Authority	12/10/2014	27217
▶ Abingdon Redevelopment and Housi	5/22/2014	37068
▶ Ada County Housing Authority	7/24/2013	16133
▶ ADAMS METROPOLITAN HOUSING AUTHO	1/28/2014	24047
▶ ADAMS METROPOLITAN HOUSING AUTHO	1/27/2014	32874
▶ Afton Housing Commission	5/6/2014	25288
▶ Agra Housing Authority	12/10/2013	25313
▶ Ahoskie Housing Authority	9/9/2014	19081
▶ Ainsworth Housing Authority	11/25/2014	35188
▶ Akron Metropolitan Housing Autho	10/9/2014	25593
▶ Akron Metropolitan Housing Autho	10/8/2014	15626
▶ Akron Metropolitan Housing Autho	10/7/2014	31822
▶ Akron Metropolitan Housing Autho	10/6/2014	21776
▶ Akron Metropolitan Housing Autho	10/3/2014	38714
▶ Akron Metropolitan Housing Autho	9/29/2014	24450
▶ Akron Metropolitan Housing Autho	9/26/2014	15079
▶ Akron Metropolitan Housing Autho	9/25/2014	13871
▶ Akron Metropolitan Housing Autho	9/24/2014	35447
▶ Akron Metropolitan Housing Autho	9/23/2014	36279
▶ Akron Metropolitan Housing Autho	9/22/2014	27095
▶ Akron Metropolitan Housing Autho	9/22/2014	15830
▶ Akron Metropolitan Housing Autho	8/29/2014	11882
▶ Akron Metropolitan Housing Autho	8/27/2014	22289
▶ Akron Metropolitan Housing Autho	8/26/2014	20772
▶ Akron Metropolitan Housing Autho	8/26/2014	17208
▶ Akron Metropolitan Housing Autho	8/25/2014	22730
▶ Akron Metropolitan Housing Autho	8/8/2014	11472

Then, I rename the PUBLIC_HOUSING_AGENCY_NAME column to PHA_NAME and converts the INSPECTION_DATE column into a date format using the ‘STR_TO_DATE’ function. The COST_OF_INSPECTION_IN_DOLLARS column is selected as-is and renamed to INSPECTION_COST. The result are shows below:

```
-  
9 • SELECT  
0   PUBLIC_HOUSING_AGENCY_NAME AS PHA_NAME,  
1   STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y') AS "DATE",  
2   COST_OF_INSPECTION_IN_DOLLARS AS INSPECTION_COST  
3   FROM public_housing_inspection_data;
```

PHA_NAME	DATE	INSPECTION_COST
▶ Abbotsford Housing Authority	2014-12-10	27217
Abingdon Redevelopment and Housi	2014-05-22	37068
Ada County Housing Authority	2013-07-24	16133
ADAMS METROPOLITAN HOUSING AUTHO	2014-01-28	24047
ADAMS METROPOLITAN HOUSING AUTHO	2014-01-27	32874
Afton Housing Commission	2014-05-06	25288
Agra Housing Authority	2013-12-10	25313
Ahoskie Housing Authority	2014-09-09	19081
Ainsworth Housing Authority	2014-11-25	35188
Akron Metropolitan Housing Autho	2014-10-09	25593
Akron Metropolitan Housing Autho	2014-10-08	15626
Akron Metropolitan Housing Autho	2014-10-07	31822
Akron Metropolitan Housing Autho	2014-10-06	21776
Akron Metropolitan Housing Autho	2014-10-03	38714
Akron Metropolitan Housing Autho	2014-09-29	24450
Akron Metropolitan Housing Autho	2014-09-26	15079
Akron Metropolitan Housing Autho	2014-09-25	13871
Akron Metropolitan Housing Autho	2014-09-24	35447
Akron Metropolitan Housing Autho	2014-09-23	36279

I compute various metrics concerning inspections, including the date and cost of each inspection, as well as the date and cost of the preceding inspection. Additionally, I calculate the change in cost between consecutive inspections, along with the percentage change in cost.

The final query selects all columns from the CTE where there's an increase in inspection cost ($\text{CHANGE_IN_COST} > 0$) and the cost of the second most recent inspection is the maximum for each agency. The LAG function is used to access data from the previous row within the same partition (grouped by agency name) to facilitate these calculations. The queries essentially analyze inspection data for public housing agencies, examining changes in inspection costs over time and identifying instances where costs have increased. The result shows 241 columns.

```

• SELECT
    PUBLIC_HOUSING_AGENCY_NAME AS PHA_NAME,
    STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y') AS "DATE",
    COST_OF_INSPECTION_IN_DOLLARS AS INSPECTION_COST
  FROM public_housing_inspection_data;
• WITH CTE AS (
    SELECT
        PUBLIC_HOUSING_AGENCY_NAME AS PHA_NAME,
        STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y') AS MR_INSPECTION_DATE,
        COST_OF_INSPECTION_IN_DOLLARS AS MR_INSPECTION_COST,
        LAG(STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y')) OVER(PARTITION BY PUBLIC_HOUSING_AGENCY_NAME ORDER BY STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y')) AS SECOND_MR_INSPECTION_DATE,
        LAG(COST_OF_INSPECTION_IN_DOLLARS) OVER(PARTITION BY PUBLIC_HOUSING_AGENCY_NAME ORDER BY STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y')) AS SECOND_MR_INSPECTION_COST
      FROM public_housing_inspection_data
),
CostIncrease AS (
    SELECT
        PHA_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 CTE
    WHERE rn = 1 AND MR_INSPECTION_COST > SECOND_MR_INSPECTION_COST
)
SELECT
    PHA_NAME,
    MR_INSPECTION_DATE,
    MR_INSPECTION_COST,
    SECOND_MR_INSPECTION_DATE,
    SECOND_MR_INSPECTION_COST,
    CHANGE_IN_COST,
    PERCENT_CHANGE_IN_COST

```

100% | 34:10 |

Result Grid Filter Rows: Search Export:

PHA_NAME	MR_INSPECTION_DATE	MR_INSPECTION_COST	SECOND_MR_INSPECTION_DATE	SECOND_MR_INSPECTION_COST	CHANGE_IN_COST
Akron Metropolitan Housing Autho	2014-10-09	25593	2014-10-08	15626	9967
Alachua County	2015-01-22	37345	2014-05-01	17019	20326
Alaska Housing Finance Corporati	2014-11-14	26342	2014-11-13	21366	4976
Albany Housing Authority	2015-01-12	31115	2015-01-09	30247	868
Alexandria Redevelopment & Housi	2014-05-09	29123	2014-04-18	14767	14356
ALLEGIENY COUNTY HOUSING AUTHORI	2015-02-02	37108	2015-02-02	36454	654
ANNISTON HA	2014-12-30	31506	2014-08-21	10785	20721
ASHTABULA METROPOLITAN HOUSING A	2014-06-03	37948	2014-04-24	13920	24028
Athens Metropolitan Housing Auth	2014-05-22	21816	2014-05-21	10996	10820
Aurora Housing Authority	2015-02-02	14683	2014-06-24	12831	1852
Aurora Housing Authority ofte C	2014-07-03	14908	2013-06-11	14570	338
Austin Housing Authority	2014-06-30	36672	2014-06-26	25920	10752
Barre Housing Authority	2014-06-18	19254	2014-06-16	16757	2497
Batavia Housing Authority	2015-01-28	26365	2014-12-30	14576	11789
Battle Creek Housing Commission	2015-01-29	34258	2015-01-27	15344	18914
Bayonne Housing Authority	2014-09-12	26407	2014-09-11	16280	10127
Beloit Housing Authority	2014-04-30	35276	2013-05-14	14461	20815
Bergen County Housing Authority	2014-06-30	20972	2014-05-28	12018	8954
Binghamton Housing Authority	2014-10-06	29731	2013-08-26	11087	18644
Bloomfield Housing Authority	2015-01-27	39447	2014-04-21	30705	8742
BLUE EARTH COUNTY EDA	2015-01-15	37189	2015-01-14	18784	18405
BOAZ HOUSING AUTHORITY	2014-04-07	22334	2014-04-03	12740	9594
Boulder Housing Partners	2014-05-08	19869	2014-05-07	19550	319
Bradford County Housing Authorit	2014-10-29	19262	2014-10-28	10512	8750
Bristol Housing Authority	2014-10-07	39542	2014-09-26	27324	12218
Brockton Housing Authority	2014-05-20	36225	2014-05-08	35206	1019
Bureau County Housing Authority	2014-08-21	37026	2014-06-09	16452	20574
Burleigh County Housing Authorit	2014-05-01	11930	2014-04-29	10301	1629

Conclusion

In this way, we have derived from our analysis of the 'public_housing_inspection_data' dataset certain valuable insights into the dynamics of the inspections carried out by public housing agencies. We segregated fact from dimension columns and made use of advanced SQL techniques like window functions to offer comprehensive insights into inspection costs and agency changes. The analysis here not only justifies the structure of the dataset but also delivers certain key insights for senior management to make informed decisions about public housing initiatives.

Reference

Decisionworks. (n.d.) . Dimensional Modeling: The Kimball Method.

<https://decisionworks.com/dimensional-data-warehouse-business-intelligence-courses/education-training-classes/dimensional-modeling-kimball/>

Gegner, S. (n.d.). SCD Type 2. <https://www.iri.com/blog/vldb-operations/scd-type-2/>

Gegner, S. (n.d.). Introduction to Slowly Changing Dimensions (SCD).

<https://www.iri.com/blog/vldb-operations/introduction-to-slowly-changing-dimensions-scd/>