Logo

Description automatically generated

**ALY 6030: Data Warehouse and SQL**

**Assignment 3: Public Housing Inspection**

**NUID: 002938209**

**Email: jain.abhin@northeastern.edu**

**Submitted by: Abhinav Jain**

**Date: 12/11/2022**

**Submitted to: Prof. Yoni Dvorkis**

**Introduction**

1. Answer the questions below.

* How many facts are there in this dataset?

Answer: In this given dataset there are two facts.

* Which facts do you identify?

Answer: In this given dataset two facts are identified as follows:

1. “COST\_OF\_INSPECTION\_IN\_DOLLARS”

2. “INSPECTION\_SCORE”

* For the facts that you identify, what type of facts are they?

Answer: In this given dataset two type facts are identified as follows:

1. “COST\_OF\_INSPECTION\_IN\_DOLLARS”: **Additive**

2. “INSPECTION\_SCORE” : **Non-Additive**

2. Answer the questions below.

* How many dimensions are there in this dataset?

Answer: In this given dataset there are seven (7) dimensions.

* Which dimensions do you identify?

Answer: In this given dataset the dimensions which are identified in the dataset are as follows:

INSPECTION\_ID,

PUBLIC\_HOUSING\_AGENCY\_NAME,

INSPECTED\_DEVELOPMENT\_NAME,

INSPECTED\_DEVELOPMENT\_ADDRESS,

INSPECTED\_DEVELOPMENT\_CITY,

INSPECTED\_DEVELOPMENT\_STATE,

INSPECTION\_DATE

3. Answer the question below.

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?

Answer: We can utilize a transaction fact table to store the smallest units of data, which will make performing aggregation easier while considering the needs of senior management. We will be able to discover the results per inspection using this technique. I assume the Periodic Snapshot fact table will be the most appropriate to collect for each month in the instance of the periodic summary of the inspection cost. The current condition, or the time, is stored in this periodic table of instantaneous data for the designated regular intervals. As a result, the monthly interval is set for the table at the lowest level. Finally, from the periodic snapshot, top management will also be able to see the inspection cost for the month.

4. Answer the question below.

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 context, how would handle this slowly changing dimension? Select from types 0,1,2, or 3 from the Kimball reading. Justify your answer.

Answer: It is important for us to have the ability to add new records that specifically match the type 2 of Slowly Changing Dimensions in the provided challenge since we are required to add new data that results from mergers with other agencies. We are aware that this process works in the direction of adding new rows to the dimension table using a new surrogate key. We can keep the historical records of the various agencies as the current record for the same department and update the most latest names and addresses in the warehouse.

5. Address the scenario below.

Finally, Senior Management is interested in a subset of this data, for only those PHAs that saw an *increase* in the $$ cost of performing an inspection in their jurisdiction. Since none of them are SQL programmers, they’ve asked your help in performing this analysis by providing a file as your final deliverable with the following columns:

Note that MR stands for “most recent”:

* **PHA\_NAME,**
* **MR\_INSPECTION\_DATE,**
* **MR\_INSPECTION\_COST,**
* **SECOND\_MR\_INSPECTION\_DATE,**
* **SECOND\_MR\_INSPECTION\_COST,**
* **CHANGE\_IN\_COST**
* **PERCENT\_CHANGE\_IN\_COST**

Management has asked that you perform this function using lead or lag functions in SQL.

However, they’re concerned that the files when imported into MySQL Workbench may not properly refer to dates using the correct format. If that is the case, they’ve asked you to investigate how best to convert dates from TEXT to Date format so that the lead/lag functions work as expected.

They’ve also asked that you filter your dataset to only those PHAs that saw an increase in $$ cost, and that you only list the PHA once with no duplicates to avoid noisy data.

Naturally, this would also require you to filter out PHAs that only performed one inspection, so they’ve asked you to remove those as well.

Answer:

Table

Description automatically generated