Sean McLean

ALY 6030

Module 5 Assignment

Public Housing Inspections Star Schema

1. How many facts are there in this dataset? Which facts do you identify? For the facts that you identify, what type of facts are they?

In the provided dataset, there are two facts that were identified:

1. **INSPECTION_SCORE**: This represents the inspection score for each development, shown as a percentage with a range from 0 to 100. It is a quantitative measurement of the development's condition that is used to assess the performance of the development. This type of fact is semi-additive, and while it can be averaged across different inspections, summing the scores would probably not make sense as it represents a ratio or percentage.
2. **COST_OF_INSPECTION_IN_DOLLARS**: This shows the cost incurred to perform each inspection and is expressed in dollars. It is a quantitative measurement of the resources used to conduct the inspection, and this type of fact is additive. It can be summed across different periods or categories to get total costs, making it suitable for aggregation.

Overall, the facts in a dataset are usually numerical measurements or events that can be analyzed in a data warehouse. They often reside in fact tables and involve quantifiable data like sales amounts or transaction counts. The facts are typically classified into three types:

- **Additive facts**: These can be summed across different dimensions like with total sales or total units sold.
- **Semi-additive facts**: These can be summed across some dimensions but not others like with inventory levels.
- **Non-additive facts**: These cannot be summed across any dimension like with ratios or percentages.

2. How many dimensions are there in this dataset? Which dimensions do you identify?

In this dataset there are six dimensions identified, and they all have descriptive attributes that provide context for the facts, helping to categorize or segment the data. The dimensions in this dataset include:

1. **PUBLIC_HOUSING_AGENCY_NAME**: This dimension describes the organization conducting the inspection.
2. **INSPECTED_DEVELOPMENT_NAME**: This dimension identifies the specific development being inspected.
3. **INSPECTED_DEVELOPMENT_ADDRESS**: This provides the address of the inspected development.
4. **INSPECTED_DEVELOPMENT_CITY**: This provides the city where the inspected development is located.

5. **INSPECTED_DEVELOPMENT_STATE**: This provides the state where the inspected development is located.
6. **INSPECTION_DATE**: This dimension provides the date of the inspection.

In dimensional modeling, the dimensions are stored in dimension tables and help provide valuable context to the facts. They could include attributes such as time, product, region, or customer which are related to the facts. They can also be categorized based on how they handle changes over time as seen in slowly changing dimensions. These can be classified as:

- **Type 0**: Dimensions that don't change over time.
- **Type 1**: Dimensions that overwrite the old value with a new value.
- **Type 2**: Dimensions that preserve all historical changes by adding new records with a new surrogate key.
- **Type 3**: Dimensions that store only the current and previous values in separate columns.

When assigning primary keys to dimension tables, surrogate keys are often preferred because they avoid the complexities of natural keys across different systems. The surrogate keys help simplify the management of dimension tables, especially when dealing with fact and dimension tables.

3. 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?

When provided the scenario where senior management is interested in tracking both inspection-level data and periodic summaries of inspection costs by month, the appropriate fact tables to use would be:

1. **Transactional Fact Table**: This table can store detailed inspection data at the individual inspection level. Each record would represent a specific inspection, capturing detailed attributes such as inspection cost, date, and related dimensions like inspector, property, etc.
2. **Snapshot Fact Table**: This table would store periodic summaries of inspection costs, typically at the monthly level that would facilitate aggregation and historical analysis. Each row would represent a month for a specific grouping like with public housing agency for example, and the metrics would entail total inspection costs and other aggregated data. These two types of fact tables would meet the need for both detailed and summary reporting.

4. 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.

The fact that senior management is concerned with frequent changes in the names and addresses of public housing agencies due to the mergers, the most effective approach to handle this would be type 2 from the Kimball reading. This method is ideal for tracking historical changes in attributes over time. In this situation when a public housing agency undergoes a merger or changes its name or address, a new record would be added to the dimension table with a new surrogate key. This approach preserves the historical data like previous names and addresses while also capturing the current details of the agency. Using type 2 ensures that historical information remains available for analysis, providing accurate reporting over time. And by creating a new record with a new surrogate key, referential integrity is maintained while retaining the original data.