Your boss in IT wants you to answer the following questions based on your inspection of the dataset.

1

a. How many facts are there in this dataset?

There are two main facts in this dataset:

1. Cost of Inspection in Dollars
2. Inspection Score

b. Which facts do you identify?

The facts identified from the dataset are:

1. Cost of Inspection in Dollars: This is the cost incurred for performing each inspection.
2. Inspection Score: This represents the outcome of the inspection, scored as a numerical value.

c. What type of facts are they?

1. Cost of Inspection in Dollars: This is a monetary fact because it represents the actual cost associated with performing the inspection.
2. Inspection Score: This is a qualitative fact that is represented in numeric form. It reflects the result or quality of the inspection.

These facts represent the core measurable quantities of the dataset that provide quantitative data for analysis.

2

a. How many dimensions are there in this dataset?

There are six dimensions in this dataset.

b. Which dimensions do you identify?

The dimensions identified from the dataset are:

1. Public Housing Agency Name: Identifies the public housing agency conducting the inspection.
2. Inspected Development Name: Refers to the name of the development or property being inspected.

3. Inspected Development Address: The physical address of the property that is being inspected.
4. Inspected Development City: The city where the inspected development is located.
5. Inspected Development State: The state where the inspected development is located.
6. Inspection Date: The date the inspection took place.

These dimensions provide descriptive attributes that contextualize the facts (inspection costs and scores) and allow for filtering and aggregation in analysis.

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?

The data should be stored in two types of fact tables:

1. Transactional Fact Table (Inspection-Level Detail)

This table would capture the detailed data for each individual inspection. It would store records for each inspection performed, including the associated cost and inspection score. This table would be particularly useful for analyzing individual inspection events and their characteristics.

Why use a transactional fact table?

- It allows for detailed analysis at the inspection level.
- It can track specific attributes of each inspection (e.g., inspection cost, inspection score).
- Provides flexibility for detailed reporting or drill-down analysis at the individual inspection level.

Fields in this table:

- FACT_ID (Primary Key)
- PUBLIC_HOUSING_AGENCY_NAME
- INSPECTED_DEVELOPMENT_NAME
- INSPECTED_DEVELOPMENT_ADDRESS
- INSPECTED_DEVELOPMENT_CITY
- INSPECTED_DEVELOPMENT_STATE
- INSPECTION_DATE
- COST_OF_INSPECTION_IN_DOLLARS
- INSPECTION_SCORE

2. Periodic Snapshot Fact Table (Monthly Summary)

This table would provide a summary of inspection costs on a monthly basis. It would aggregate the inspection costs for each month, which would be useful for senior management to track and compare costs over time. The periodic snapshot would give a clear view of the inspection costs by month, making it easier for senior management to analyze trends.

Why use a periodic snapshot fact table?

- It provides a higher-level view of trends and totals over time, making it easier to track and summarize monthly costs.
- It simplifies reporting for senior management, as they can view aggregated data for each month, without needing to analyze every single inspection.
- It reduces the need for real-time aggregation of data and can speed up queries related to overall trends.

Fields in this table:

- FACT_ID (Primary Key)
- PUBLIC_HOUSING_AGENCY_NAME
- MONTH
- TOTAL_INSPECTION_COST
- TOTAL_INSPECTION_SCORE

Conclusion:

By storing the data in both transactional fact tables and periodic snapshot fact tables, we can satisfy both the need for detailed inspection-level analysis and for summarized monthly trends. The transactional fact table provides granularity, while the periodic snapshot table allows for aggregated analysis, which will be useful for senior management's periodic reviews of inspection costs.

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

To handle the slowly changing dimension (SCD) for Public Housing Agency Names and Addresses, where changes occur due to mergers or name/address updates, SCD Type 2 would be the most appropriate choice.

Justification for Using SCD Type 2:

SCD Type 2 is the best option because:

- **Historical Tracking:** Type 2 allows to preserve historical data by creating a new record whenever a change occurs (e.g., a merger or name/address change). This is important because it ensures that the historical context of each public housing agency is maintained in the data, even after changes in the agency name or address.
- **Auditing Changes:** Since public housing agencies are frequently merging or changing names, using SCD Type 2 ensures that can track when these changes happened and maintain a complete history of the different versions of agency names and addresses over time. This is especially important for accurate reporting and analysis, where users need to see data as it existed at different points in time.
- **Ensuring Data Integrity:** With SCD Type 2, each record of the public housing agency will have a start and end date (or a version flag), indicating when a particular name/address was valid. This prevents the loss of information, as older names or addresses are retained alongside the current ones.

How SCD Type 2 Works:

- When a public housing agency changes its name or address, a new record is created in the dimension table with the updated values.
- The old record is marked as inactive (e.g., using an end date or a version flag) while the new record becomes active (using a start date or current flag).
- By using this approach, I can track each agency's name/address over time, even if they merge or change.

Example of SCD Type 2:

For instance, if a public housing agency named "Abbotsford Housing Authority" merges with another agency and changes its name to "Abbotsford Metropolitan Housing Authority," a new record would be created with the updated name, while the old record would be marked with an end date. This allows me to track the history of the agency's name, ensuring that past inspections tied to the old name can still be correctly attributed to the right agency.

In conclusion, SCD Type 2 will best address the needs for tracking changes in public housing agency names and addresses while maintaining a full historical record, ensuring that senior management has the most accurate and complete data for analysis.

Reference

LAG / lead functions. Sisense Community. (2024, February 23). https://community.sisense.com/t5/knowledge-base/lag-lead-functions/ta-p/508#:~:text=LAG%20%26%20LEAD%20Functions%20are%20functions,further%20down%20the%20result%20set.

Shana, ShanaShana 7011 gold badge11 silver badge88 bronze badges, Ifeanyi ChukwuIfeanyi Chukwu 3, & Ray KrungkaewRay Krungkaew 6. (1960, June 1). *What is the best way prevent Duplicate Records in a SQL server database*. Stack Overflow. https://stackoverflow.com/questions/30050482/what-is-the-best-way-prevent-duplicate-records-in-a-sql-server-database

PySquirrelPySquirrel 3911 silver badge66 bronze badges, & Justin CaveJustin Cave 231k2525 gold badges377377 silver badges392392 bronze badges. (1967, May 1). *Lag function over dates*. Stack Overflow. https://stackoverflow.com/questions/71651951/lag-function-over-dates