Assignment-3(Public Housing Inspections Star Schema)

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

Facts are quantitative, measurable data points that provide insight into key performance metrics. In this dataset, the primary facts are:

1) **Cost of Inspection in Dollars** – Represents the monetary cost associated with performing an inspection.

    - **Type**: Additive Fact (can be summed across different inspections and time periods).
    - **Example Use Case**: Total inspection cost incurred by an agency over a given year.

2) **Inspection Score** – Represents the score assigned to a housing development based on inspection results.

    - **Type**: Semi-Additive Fact (can be averaged but not summed across inspections).
    - **Example Use Case**: Average inspection score for all inspections conducted in a particular city.

These facts will be stored in fact tables to facilitate reporting and trend analysis.

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

Dimensions provide descriptive context for facts, allowing users to analyze data across different perspectives. The key dimensions in this dataset include:

- **Inspection ID** – Unique identifier for each inspection.
- **Public Housing Agency Name** – The name of the agency responsible for a given housing development.
- **Inspected Development Name** – Name of the housing development being inspected.
- **Inspected Development Address** – Physical address of the development, which includes:
    - **City** – The city where the development is located.
    - **State** – The state where the development is located.
- **Inspection Date** – The date when the inspection was conducted, which allows for time-based analysis.

These dimensions will be used to filter, group, and aggregate fact data for reporting and analytics.

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?

Senior management is interested in viewing the identified facts (**COST_OF_INSPECTION_IN_DOLLARS** and **INSPECTION_SCORE**) at both the inspection level and as a periodic summary of inspection costs for each month. To meet this requirement, we will use two types of fact tables: a Transaction Fact Table and a Periodic Snapshot Fact Table.

1) **Transaction fact table: inspection_fact**

- This table records each individual inspection event as it occurs.
- It includes inspection_id (unique identifier), inspection_date, public_housing_agency_id, inspected_development_id, and measurable facts such as cost_of_inspection_in_dollars and inspection_score.
- This table supports detailed analysis, such as identifying costly inspections, tracking low scores, and filtering inspections by agency or location.

2) **Periodic snapshot fact table: monthly_inspection_cost_summary_fact**

- This table aggregates inspection data at a monthly level to support trend analysis.
- It includes year, month, public_housing_agency_id, total_cost_of_inspections, and average_inspection_score per agency.
- This table enables efficient reporting of monthly inspection costs and agency performance trends without needing to query individual inspection records.

By implementing both **inspection_fact** as a Transaction Fact Table and **monthly_inspection_cost_summary_fact** as a **Periodic Snapshot Fact Table**, we ensure flexibility: analysts can drill down into specific inspections when needed, while executives can quickly assess financial and performance trends without processing massive transactional data.

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.

The dataset contains public housing inspection records with fields such as **PUBLIC_HOUSING_AGENCY_NAME, INSPECTED_DEVELOPMENT_ADDRESS, INSPECTION_DATE**, and other related details. Since senior management is concerned about changes in public housing agency names due to frequent mergers, we should handle this as a Slowly Changing Dimension (SCD) Type 2.

Justification for SCD Type 2:

1. **Preserves Historical Data:** Instead of overwriting agency names and addresses, we create a new record with a different surrogate key for each change. This ensures that previous records remain available for historical analysis.
2. **Effectively Manages Mergers:** When agencies merge or are renamed, the system creates a new version of the record while keeping the old data intact, ensuring that past inspections remain linked to their original agencies.
3. **Supports Long-Term Reporting and Trend Analysis:** By maintaining historical records, management can analyze past agency performance and track changes over time.

5. 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"):

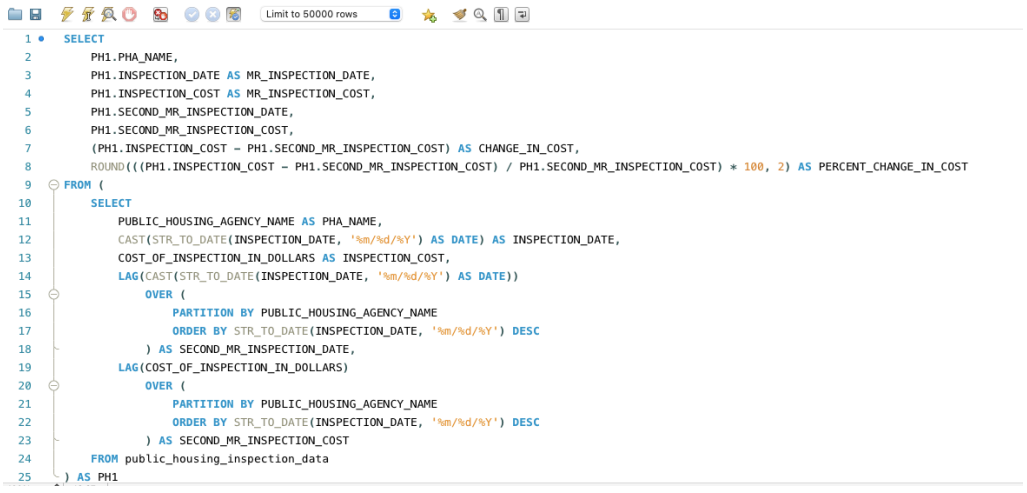
- PHA_NAME,
- MR_INSPECTION_DATE,
- MR_INSPECTION_COST,
- SECOND_MR_INSPECTION_DATE,
- SECOND_MR_INSPECTION_COST,
- CHANGE_IN_COST
- PERCENT_CHANGE_IN_COST

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.



```sql
) AS PH1
JOIN (
    SELECT
        PUBLIC_HOUSING_AGENCY_NAME AS PHA_NAME,
        MAX(STR_TO_DATE(INSPECTION_DATE, '%m/%d/%Y')) AS LatestInspectionDate
    FROM public_housing_inspection_data
    GROUP BY PUBLIC_HOUSING_AGENCY_NAME
) AS PH2
ON PH1.PHA_NAME = PH2.PHA_NAME AND PH1.INSPECTION_DATE = PH2.LatestInspectionDate
WHERE PH1.SECOND_MR_INSPECTION_COST IS NOT NULL
AND PH1.INSPECTION_COST > PH1.SECOND_MR_INSPECTION_COST;
```

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	PERCENT_CHANGE_
ALTOONA HOUSING AUTHORITY	2014-11-24	25750	2014-11-24	18863	6887	36.51
Benton Harbor Housing Commission	2014-10-27	36524	2014-10-27	18026	18498	102.62
Cameron County Housing Authority	2013-10-31	21230	2013-10-31	18388	2842	15.46
City of Clay Center	2014-04-29	37716	2014-04-29	33064	4652	14.07
CLEARWATER HOUSING AUTHORITY	2014-07-10	33812	2014-07-10	15199	18613	122.46
Dover Housing Authority	2015-01-21	11644	2015-01-21	10979	665	6.06
Glens Falls Housing Authority	2014-05-12	31495	2014-05-12	20458	11037	53.95
HA NORTHPORT	2014-10-01	29759	2014-10-01	14302	15457	108.08
HA NORTHPORT	2014-10-01	37900	2014-10-01	15644	22256	142.27
Highlands Housing Authority	2013-08-28	34991	2013-08-28	25037	9954	39.76
Housing Authority of City of Day	2014-09-11	33620	2014-09-11	29677	3943	13.29
Housing Authority of Hamilton, A	2014-04-07	28557	2014-04-07	17152	11405	66.49
HOUSING AUTHORITY OF LAKE CHARLE	2013-09-27	37191	2013-09-27	29419	7772	26.42
Housing Authority of St. James P	2013-09-19	25901	2013-09-19	25649	252	0.98
Housing Authority of the County	2015-01-29	38839	2015-01-29	11355	27484	242.04
HRA of WINONA, MINNESOTA	2014-05-20	23401	2014-05-20	10966	12435	113.40
KINGS COUNTY HOUSING AUTH	2013-07-23	19919	2013-07-23	11578	8341	72.04
Kingsville Housing Authority	2014-04-22	34329	2014-04-22	11274	23055	204.50
Marquette Housing Commission	2014-11-19	29475	2014-11-19	11063	18412	166.43
New York City Housing Authority	2015-01-26	29453	2015-01-26	19269	10184	52.85
Roanoke Redevelopment & Housing	2015-01-29	19596	2015-01-29	10466	9130	87.23
San Diego Housing Commission	2014-08-13	30651	2014-08-13	27052	3599	13.30
Sault Ste Marie Housing Commissi	2014-07-01	35408	2014-07-01	18010	17398	96.60
Stevens Point Housing Authority	2014-04-24	30108	2014-04-24	20528	9580	46.67
VIRGIN ISLANDS HOUSING AUTHORITY	2014-06-10	27935	2014-06-10	21586	6349	29.41
Winchester Housing Authority	2014-10-06	36821	2014-10-06	24034	12787	53.20