Assignment 3

Public Housing Inspections Star Schema

INSPECTION_ID	INSPECTED_DEVELOPMENT_NAME	INSPECTED_DEVELOPMENT_STATE
PUBLIC_HOUSING_AGENCY_NAME	INSPECTED_DEVELOPMENT_ADDRESS	INSPECTION_DATE
COST_OF_INSPECTION_IN_DOLLARS	INSPECTED_DEVELOPMENT_CITY	INSPECTION_SCORE

Before any queries operation, the raw data set has been preprocessing and cleaned.

Q1: How many facts are there in this dataset?

1. **Which facts do you identify?**

Two mains facts include:

Inspection Cost: INSPECTION_SCORE

Inspection Score: COST_OF_INSPECTION_IN_DOLLARS

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

Facts identified from the dataset

Column Name	Fact Type	Description
INSPECTION_SCORE	Additive Fact	A numeric score can be aggregated
COST_OF_INSPECTION_IN_DOLLARS	Additive Fact	A monetary value can be summed across

Q2: How many dimensions are there in this dataset?

Which dimensions do you identify?

6 dimensions

Column Name	Variable Type	Description
PUBLIC_HOUSING_AGENCY_NAME	Dimension	organization
INSPECTED_DEVELOPMENT_NAME	Dimension	location/entity
INSPECTED_DEVELOPMENT_ADDRESS	Dimension	location
INSPECTED_DEVELOPMENT_CITY	Dimension	location
INSPECTED_DEVELOPMENT_STATE	Dimension	location
INSPECTION_DATE	Dimension	time

Q3: Senior management is interested in viewing the facts identified above, at both the inspection level and 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?

Type of Table	Purpose	Detail level
Transactional Fact Table	Store each inspection event	Per inspection
Periodic Snapshot Table	Store a monthly summary of inspection costs	Per month per agency

1. Transactional Fact Table

Purpose: Stores individual inspection records, the most granular level.

Reason: The TransactionFactTable stores data about individual inspections, capturing events or transactions that occur at a granular level. Each row represents a single inspection with attributes such as the cost, inspection score, and development details. This is a typical transactional fact table because it records detailed, event-level data, usually with a focus on operational metrics.

Contents:

INSPECTION_ID (Primary Key)

PUBLIC_HOUSING_AGENCY_NAME

COST_OF_INSPECTION_IN_DOLLARS

INSPECTED_DEVELOPMENT_NAME

INSPECTED_DEVELOPMENT_ADDRESS

INSPECTED_DEVELOPMENT_CITY

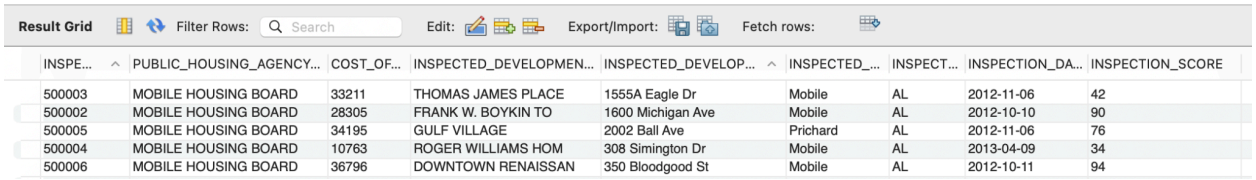
INSPECTED_DEVELOPMENT_STATE

INSPECTION_DATE

INSPECTION_SCORE

```
LOAD DATA LOCAL INFILE
'/Users/sherry_yu/ALY6030/MOD5/public_housing_inspection_data_2.csv'
 INTO TABLE `TransactionFactTable`
```

```
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES
   (INSPECTION_ID,
   PUBLIC_HOUSING_AGENCY_NAME,
   COST_OF_INSPECTION_IN_DOLLARS,
   INSPECTED_DEVELOPMENT_NAME ,
   INSPECTED_DEVELOPMENT_ADDRESS,
   INSPECTED_DEVELOPMENT_CITY,
   INSPECTED_DEVELOPMENT_STATE,
   @INSPECTION_DATE, INSPECTION_SCORE)
   SET INSPECTION_DATE = STR_TO_DATE(@INSPECTION_DATE, '%m/%d/%Y');
```



Result Grid | Filter Rows: Search | Edit: | Export/Import: | Fetch rows:

INSPE...	PUBLIC_HOUSING_AGENCY...	COST_OF...	INSPECTED_DEVELOPMEN...	INSPECTED_DEVELOP...	INSPECTED_...	INSPECT...	INSPECTION_DA...	INSPECTION_SCORE
500003	MOBILE HOUSING BOARD	33211	THOMAS JAMES PLACE	1555A Eagle Dr	Mobile	AL	2012-11-06	42
500002	MOBILE HOUSING BOARD	28305	FRANK W. BOYKIN TO	1600 Michigan Ave	Mobile	AL	2012-10-10	90
500005	MOBILE HOUSING BOARD	34195	GULF VILLAGE	2002 Ball Ave	Prichard	AL	2012-11-06	76
500004	MOBILE HOUSING BOARD	10763	ROGER WILLIAMS HOM	308 Simington Dr	Mobile	AL	2013-04-09	34
500006	MOBILE HOUSING BOARD	36796	DOWNTOWN RENAISSAN	350 Bloodgood St	Mobile	AL	2012-10-11	94

2. Periodic Snapshot Fact Table

Purpose: To provide a summarized view of inspection costs by month.

Reason: The PeriodicSnapshotFactTable stores aggregated data for a specific time period. Here, the data is summarized at a higher level (e.g., the total cost of inspections for a public housing agency as of a snapshot date). This type of fact table is used to track metrics over time at regular intervals (daily, weekly, monthly, etc.) and allows us to analyze trends over time without recalculating aggregates from transactional data.

Contents:

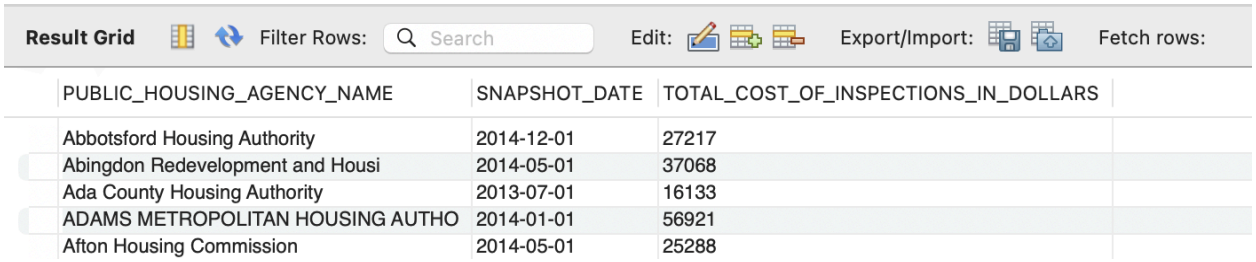
PUBLIC_HOUSING_AGENCY_NAME(Primary Key)

SNAPSHOT_DATE

TOTAL_COST_OF_INSPECTIONS_IN_DOLLARS

```
-- Insert info into PeriodicSnapshotFactTable
INSERT IGNORE INTO `PeriodicSnapshotFactTable`
(PUBLIC_HOUSING_AGENCY_NAME, SNAPSHOT_DATE,
TOTAL_COST_OF_INSPECTIONS_IN_DOLLARS)
SELECT
    PUBLIC_HOUSING_AGENCY_NAME,
    DATE_FORMAT(INSPECTION_DATE, '%Y-%m-01') AS SNAPSHOT_DATE,
    SUM(COST_OF_INSPECTION_IN_DOLLARS) AS TOTAL_COST_OF_INSPECTIONS_IN_DOLLARS
FROM
    `TransactionFactTable`
GROUP BY
    PUBLIC_HOUSING_AGENCY_NAME, SNAPSHOT_DATE
```

```
ORDER BY
    PUBLIC_HOUSING_AGENCY_NAME, SNAPSHOT_DATE;
```



Result Grid Filter Rows: Search Edit: Export/Import: Fetch rows:

PUBLIC_HOUSING_AGENCY_NAME	SNAPSHOT_DATE	TOTAL_COST_OF_INSPECTIONS_IN_DOLLARS
Abbotsford Housing Authority	2014-12-01	27217
Abingdon Redevelopment and Housi	2014-05-01	37068
Ada County Housing Authority	2013-07-01	16133
ADAMS METROPOLITAN HOUSING AUTHO	2014-01-01	56921
Afton Housing Commission	2014-05-01	25288

Q4: 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.

1. **The Four Types of Slowly Changing Dimensions**

Type	Description
Type 0	No changes allowed, retains original data
Type 1	Overwrites old data with new data
Type 2	Creates a new row with a new surrogate key for each change
Type 3	Adds new columns to track limited history

Type 2 SCD is recommended.

2. **Justification**

Why Type 2?

Type 2 tracks historical changes by creating a new row in the dimension table every time a change occurs (e.g., agency name or address changes). It allows us to preserve history, so we can analyze facts in the context of what the agency was called or where it was located at the time of the event. This is especially useful for time-trend reporting, audit trails, and understanding how mergers or name changes affected operational data over time.

How It work?

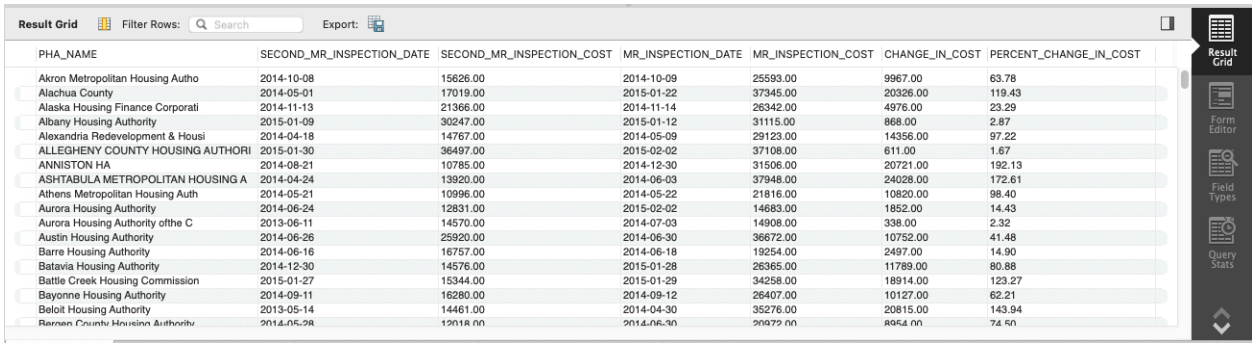
Each time a housing agency's name or address changes, a new version of the agency is added to the dimension table with: A new surrogate key, updated name and address, and validity period.

So, in summary, we choose to use Type 2 SCD for the public housing agency dimension to: Preserve historical accuracy, enable time-trend and before/after merger analysis, meet senior management's need for historical insight into agency name/address changes.

Q5: New Scenario

```
-- Rank Inspections and Calculate Previous Inspection Data
WITH InspectionRanked AS (
    SELECT PUBLIC_HOUSING_AGENCY_NAME AS PHA_NAME,
           INSPECTION_DATE, COST_OF_INSPECTION_IN_DOLLARS AS
INSPECTION_COST,
           LAG(INSPECTION_DATE) OVER (
               PARTITION BY PUBLIC_HOUSING_AGENCY_NAME
               ORDER BY INSPECTION_DATE)
               AS SECOND_MR_INSPECTION_DATE,
           LAG(COST_OF_INSPECTION_IN_DOLLARS) OVER (
               PARTITION BY PUBLIC_HOUSING_AGENCY_NAME
               ORDER BY INSPECTION_DATE)
               AS SECOND_MR_INSPECTION_COST,
           ROW_NUMBER() OVER (
           PARTITION BY PUBLIC_HOUSING_AGENCY_NAME
           ORDER BY INSPECTION_DATE DESC)
           AS RN
    FROM public_housing_inspection
)
-- Retrieve data from the CTE InspectionRanked with the specified
conditions
SELECT
    PHA_NAME,
    SECOND_MR_INSPECTION_DATE,
    SECOND_MR_INSPECTION_COST,
    INSPECTION_DATE AS MR_INSPECTION_DATE,
    INSPECTION_COST AS MR_INSPECTION_COST,
    (INSPECTION_COST - SECOND_MR_INSPECTION_COST) AS CHANGE_IN_COST,
```

```
    ROUND((INSPECTION_COST - SECOND_MR_INSPECTION_COST) /
SECOND_MR_INSPECTION_COST * 100, 2) AS PERCENT_CHANGE_IN_COST
FROM InspectionRanked
WHERE SECOND_MR_INSPECTION_COST IS NOT NULL
AND INSPECTION_COST > SECOND_MR_INSPECTION_COST
-- Only the most recent inspection record
AND RN = 1;
```



Result Grid | Filter Rows: Search | Export:

PHA_NAME	SECOND_MR_INSPECTION_DATE	SECOND_MR_INSPECTION_COST	MR_INSPECTION_DATE	MR_INSPECTION_COST	CHANGE_IN_COST	PERCENT_CHANGE_IN_COST
Akron Metropolitan Housing Autho	2014-10-08	15626.00	2014-10-09	25593.00	9967.00	63.78
Alachua County	2014-05-01	17019.00	2015-01-22	37345.00	20326.00	119.43
Alaska Housing Finance Corporati	2014-11-13	21366.00	2014-11-14	26342.00	4976.00	23.29
Albany Housing Authority	2015-01-09	30247.00	2015-01-12	31115.00	868.00	2.87
Alexandria Redevelopment & Housi	2014-04-18	14767.00	2014-05-09	29123.00	14356.00	97.22
ALLEGHENY COUNTY HOUSING AUTHORI	2015-01-30	36497.00	2015-02-02	37108.00	611.00	1.67
ANNISTON HA	2014-08-21	10785.00	2014-12-30	31506.00	20721.00	192.13
ASHTABULA METROPOLITAN HOUSING A	2014-04-24	13920.00	2014-06-03	37948.00	24028.00	172.61
Athens Metropolitan Housing Auth	2014-05-21	10996.00	2014-05-22	21816.00	10820.00	98.40
Aurora Housing Authority	2014-06-24	12831.00	2015-02-02	14683.00	1852.00	14.43
Aurora Housing Authority ofthe C	2013-06-11	14570.00	2014-07-03	14908.00	338.00	2.32
Austin Housing Authority	2014-06-26	25920.00	2014-06-30	36672.00	10752.00	41.48
Barre Housing Authority	2014-06-16	16757.00	2014-06-18	19254.00	2497.00	14.90
Batavia Housing Authority	2014-12-30	14576.00	2015-01-28	26365.00	11789.00	80.88
Battle Creek Housing Commission	2015-01-27	15344.00	2015-01-29	34258.00	18914.00	123.27
Bayonne Housing Authority	2014-09-11	16280.00	2014-09-12	26407.00	10127.00	62.21
Beloit Housing Authority	2013-05-14	14461.00	2014-04-30	35276.00	20815.00	143.94
Bergen County Housing Authority	2014-05-28	12018.00	2014-06-30	20972.00	8954.00	74.50

Result Grid | Form Editor | Field Types | Query Stats

Reference

Faisal, S., & Sarwar, M. (2014). Handling slowly changing dimensions in data warehouses. *Journal of Systems and Software*, 94, 151-160.

Kimball, R., & Ross, M. (2013). The data warehouse toolkit: The definitive guide to dimensional modeling. *John Wiley & Sons*.

Santos, R. J., & Bernardino, J. (2008, September). Real-time data warehouse loading methodology. *In Proceedings of the 2008 international symposium on Database engineering & applications* (pp. 49-58).