**Data Warehousing and SQL : Assignment 3**

Siddhant Sapte

ALY 6030

CRN: 21008

**Question 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?

* There are two facts in the dataset, COST\_OF\_INSPECTION\_IN\_DOLLARS and INSPECTION\_SCORE.

COST\_OF\_INSPECTION\_IN\_DOLLARS is an additive fact as we can calculate, sum, min, max, average and perform other aggregation functions.

INSPECTION\_SCORE is semi-additive as we only need to know the min, max and average of the scores across states and other dimensions.

**Question 2**

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

* We have considered two dimension tables, first table contains the inspection details, such as,

INSPECTION\_ID,

PUBLIC\_HOUSING\_AGENCY\_NAME,

INSPECTION\_DATE

The other table will contain the details where the inspection is performed,

INSPECTED\_DEVELOPMENT\_NAME,

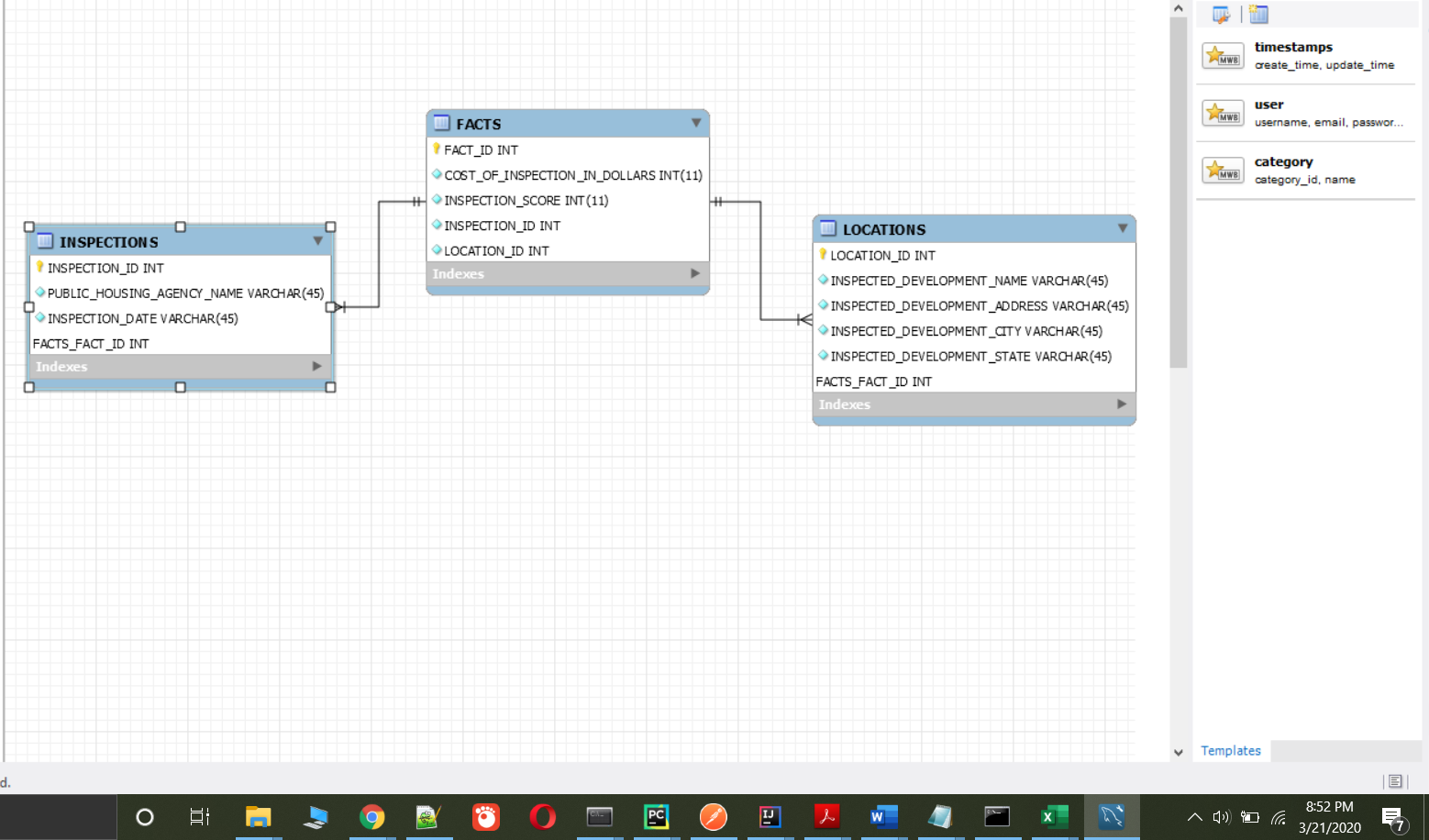
INSPECTED\_DEVELOPMENT\_ADDRESS,

INSPECTED\_DEVELOPMENT\_CITY,

INSPECTED\_DEVELOPMENT\_STATE

**Question 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?



In the fact table I would use COST\_OF\_INSPECTION\_IN\_DOLLARS and INSPECTION\_DATE from the INSPECTIONS table to get the monthly inspection cost. This is additive type of fact table.

**Question 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.

* We will handle the slowly changing dimension as per the type 2 from the Kimball reading. As per the above ER diagram, in the fact table, INSPECTION\_ID and LOCATION\_ID will act as a foreign key to the two dimension table. When we will add new data in the dimensions table, we need to update the fact table as well.

**Question 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.

Answer:

select \* from public\_housing\_inspections;

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

from (

select PUBLIC\_HOUSING\_AGENCY\_NAME as PHA\_NAME,

DATE\_OF\_INSPEC as MR\_INSPECTION\_DATE,

COST\_OF\_INSPECTION\_IN\_DOLLARS as MR\_INSPECTION\_COST,

lead(DATE\_OF\_INSPEC,1) over (partition by public\_housing\_agency\_name order by DATE\_OF\_INSPEC desc) as SECOND\_MR\_INSPECTION\_DATE,

lead(COST\_OF\_INSPECTION\_IN\_DOLLARS,1) over (partition by public\_housing\_agency\_name order by DATE\_OF\_INSPEC desc) as SECOND\_MR\_INSPECTION\_COST,

COST\_OF\_INSPECTION\_IN\_DOLLARS - lead(COST\_OF\_INSPECTION\_IN\_DOLLARS,1) over (partition by public\_housing\_agency\_name order by DATE\_OF\_INSPEC desc) as CHANGE\_IN\_COST,

(((COST\_OF\_INSPECTION\_IN\_DOLLARS) - (lead(COST\_OF\_INSPECTION\_IN\_DOLLARS,1) over (partition by PUBLIC\_HOUSING\_AGENCY\_NAME order by DATE\_OF\_INSPEC desc)))

/(lead(COST\_OF\_INSPECTION\_IN\_DOLLARS,1) over (partition by PUBLIC\_HOUSING\_AGENCY\_NAME order by DATE\_OF\_INSPEC desc)))

\* 100 as PERCENT\_CHANGE\_IN\_COST,

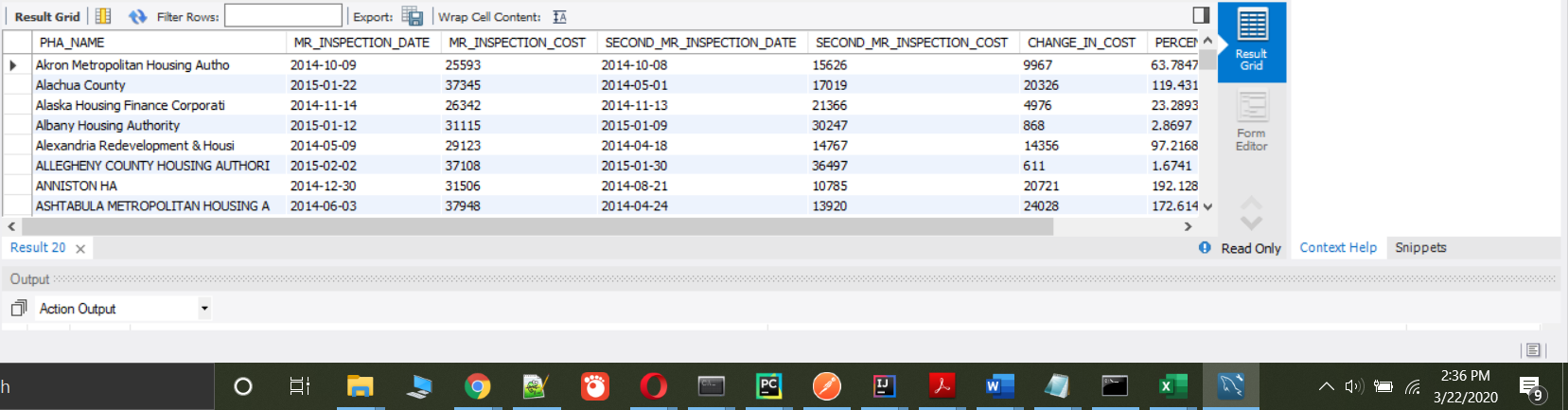
row\_number() over(partition by PUBLIC\_HOUSING\_AGENCY\_NAME) as sit

from

(select \*, STR\_TO\_DATE(INSPECTION\_DATE, '%m/%d/%Y') AS DATE\_OF\_INSPEC

from public\_housing\_inspections)as pha)as pha\_final

where PERCENT\_CHANGE\_IN\_COST > 0 and (sit = 1);



**Reference:**

1 Understanding Star Schemas. (n.d.). Retrieved from <http://gkmc.utah.edu/ebis_class/2003s/Oracle/DMB26/A73318/schemas.htm>