Module -1 -2 Solutions

1.Suppose that a data warehouse consists of the four dimensions, *date, spectator, location*, and *game*, and the two measures, *count* and *charge*, where *charge* is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

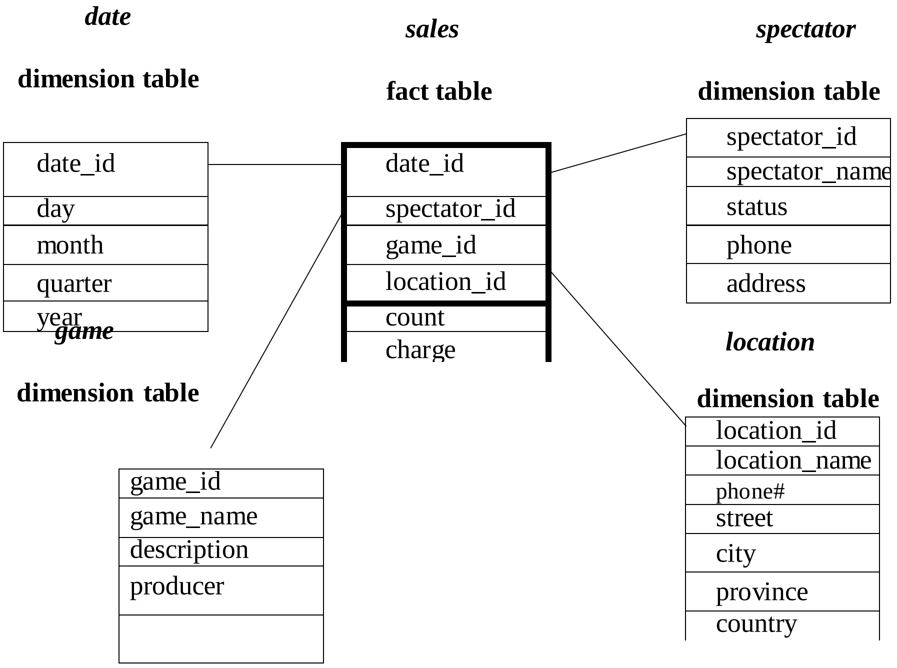
(a) Draw a *star schema* diagram for the data warehouse.

(b) Starting with the base cuboid [*date, spectator, location, game*], what specific *OLAP operations* should one perform in order to list the total charge paid by student spectators at GM_Place in 2010?

Answer:

(a) Draw a *star schema* diagram for the data warehouse.

A star schema is shown in Figure



(b) Starting with the base cuboid [*date, spectator, location, game*], what specific *OLAP operations* should one perform in order to list the total charge paid by student spectators at GM_Place in 2010?

The specific OLAP operations to be performed are:

- Roll-up on *date* from *date_id* to *year*.
- Roll-up on *game* from *game_id* to all.
- Roll-up on *location* from *location_id* to location_name.
- Roll-up on *spectator* from *spectator_id* to *status*.
- Dice with *status="students"*, *location_name="GM_Place"*, and *year = 2010*.

2. Suppose that a data warehouse for *Big-University* consists of the following four dimensions: *student, course, semester*, and *instructor*, and two measures *count* and *avg_grade*. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg_grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg_grade* stores the average grade for the given combination.

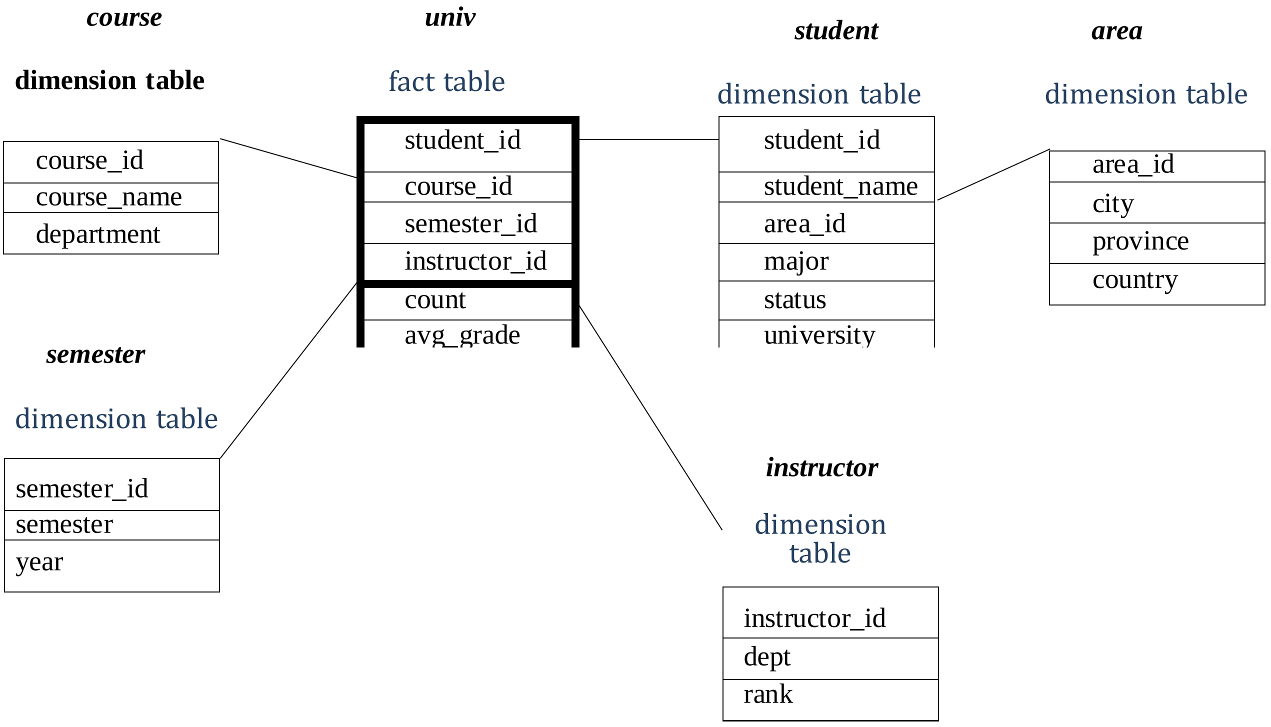
(a) Draw a *snowflake schema* diagram for the data warehouse.

(b) Starting with the base cuboid [*student, course, semester, instructor*], what specific *OLAP operations* (e.g., roll-up from *semester* to *year*) should one perform in order to list the average grade of *CS* courses for each *Big-University* student.

Answer:

(a) Draw a *snowflake schema* diagram for the data warehouse.

A snowflake schema is shown in Figure 4.2.



(a) Starting with the *base cuboid* [*student, course, semester, instructor*], what specific *OLAP operations* (e.g., roll-up from *semester* to *year*) should one perform in order to list the average grade of *CS* courses for each *Big-University* student.

The specific OLAP operations to be performed are:

- Roll-up on course from *course_id* to *department*.
- Roll-up on semester from *semester_id* to *all*.
- Slice for *course="CS"* .

(b) If each dimension has five levels (including all), such as *student* < *major* < *status* < *university* <

■

all, how many cuboids will this cube contain (including the base and apex cuboids)?

This cube will contain $5^4 = 625$ cuboids.

3. For the attribute *age data*: 13, 15, 16, 16, 19, 20,20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) Use *smoothing by bin means* to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

Answer:

The following steps are required to smooth the above data using smoothing by bin means with a bin depth of 3.

- **Step 1:** Sort the data. (This step is not required here as the data are already sorted.)
- **Step 2:** Partition the data into equidepth bins of depth 3.

Bin 1: 13, 15, 16 Bin 2: 16, 19, 20 Bin 3: 20, 21, 22
Bin 4: 22, 25, 25 Bin 5: 25, 25, 30 Bin 6: 33, 33, 35
Bin 7: 35, 35, 35 Bin 8: 36, 40, 45 Bin 9: 46, 52, 70

- **Step 3:** Calculate the arithmetic mean of each bin.
- **Step 4:** Replace each of the values in each bin by the arithmetic mean calculated for the bin.

Bin 1: 142/3, 142/3, 142/3 Bin 2: 181/3, 181/3, 181/3 Bin 3: 21, 21, 21
Bin 4: 24, 24, 24 Bin 5: 262/3, 262/3, 262/3 Bin 6: 332/3, 332/3, 332/3
Bin 7: 35, 35, 35 Bin 8: 401/3, 401/3, 401/3 Bin 9: 56, 56, 56

This method smooths a sorted data value by consulting to its "neighborhood". It performs *local* smoothing.

(b) How might you determine *outliers* in the data?

Outliers in the data may be detected by clustering, where similar values are organized into groups, or 'clusters'. Values that fall outside of the set of clusters may be considered outliers. Alternatively, a combination of computer and human inspection can be used where a predetermined data distribution is implemented to allow the computer to identify possible outliers. These possible outliers can then be verified by human inspection with much less effort than would be required to verify the entire initial data set.

4. Use the methods below to *normalize* the following group of data:

200, 300, 400, 600, 1000

(a) min-max normalization by setting *min* = 0 and *max* = 1

(b) z-score normalization

(c) z-score normalization using the mean absolute deviation instead of standard deviation

(d) normalization by decimal scaling

Answer:

(a) *min-max normalization* by setting *min* = 0 and *max* = 1 get the new value by computing

$$v'_i = \frac{v_i - 200}{1000 - 200}(1 - 0) + 0.$$

The normalized data are: 0, 0.125, 0.25, 0.5, 1

(b) *z-score normalization* using the *mean absolute deviation* instead of standard deviation replaces
σ_A with s_A, where

$$s_A = \frac{1}{5}(|200 - 500| + |300 - 500| + ... + |1000 - 500|) = 240$$

The normalized data are:

$$-1.25, -0.833, -0.417, 0.417, 2.08$$

(c) The smallest integer j such that $Max(|\frac{v_i}{10^j}|) < 1$ is 3. After *normalization by decimal scaling*, the data becomes

$$0.2, 0.3, 0.4, 0.6, 1.0$$

■

5. Using the data for *age* 13, 15, 16, 16, 19, 20,20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70., answer the following:

(a) Use min-max normalization to transform the value 35 for *age* onto the range [0.0, 1.0].

(b) Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age*
is 12.94 years.

(c) Use normalization by decimal scaling to transform the value 35 for *age*.

Answer:

(a) Use min-max normalization to transform the value 35 for *age* onto the range [0.0, 1.0].
Using the corresponding equation with min_A = 13, max_A = 70, *new* min_A = 0, *new* max_A = 1.0,then v = 35 is transformed to v' = 0.39.

(b) Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age*
is 12.94 years.
Using the corresponding equation where A = 809/27 = 29.96 and σ_A = 12.94, then v = 35 istransformed to v' = 0.39.

(c) Use normalization by decimal scaling to transform the value 35 for *age*.
Using the corresponding equation where j = 2, v = 35 is transformed to v' = 0.35.

■

6. Using the data for *age* and *body fat* given answer the following:

(a) Normalize the two attributes based on *z-score normalization*.
(b) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two at-tributes positively or negatively correlated? Compute their covariance

Answer:

(a) Normalize the two variables based on *z-score normalization*.

age	23	23	27	27	39	41	47	49	50
z-age	-1.83	-1.83	-1.51	-1.51	-0.58	-0.42	0.04	0.20	0.28

%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
z-%fat	-2.14	-0.25	-2.33	-1.22	0.29	-0.32	-0.15	-0.18	0.27
age	52	54	54	56	57	58	58	60	61
z-age	0.43	0.59	0.59	0.74	0.82	0.90	0.90	1.06	1.13
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7
z-%fat	0.65	1.53	0.0	0.51	0.16	0.59	0.46	1.38	0.77

(b) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two vari-ables positively or negatively correlated?

The *correlation coefficient* is 0.82. The variables are positively correlated.

■

7. Suppose a group of 12 *sales price* records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Partition them into three bins by each of the following methods.

(a) equal-frequency (equidepth) partitioning

(b) equal-width partitioning

(c) clustering

Answer:

(a) equal-frequency (equidepth) partitioning Partition the data into equidepth bins of depth 4:

Bin 1: 1: 5, 10, 11, 13 Bin 2: 15, 35, 50, 55 Bin 3: 72, 92, 204, 215

(b) equal-width partitioning

Partitioning the data into 3 equi-width bins will require the width to be (215 5)/3 = 70. Weget:

Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72 Bin 2: 92 Bin 3: 204, 215

(c) clustering

Using *K*-means clustering to partition the data into three bins we get:

Bin 1: 5, 10, 11, 13, 15, 35 Bin 2: 50, 55, 72, 92 Bin 3: 204, 215

8. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the *Euclidean distance* between the two objects.

(b) Compute the *Manhattan distance* between the two objects.

(c) Compute the *Minkowski distance* between the two objects, using *h* = 3.

Answer:

(a) Compute the *Euclidean distance* between the two objects.
The Euclidean distance is computed using Equation (2.6).
Therefore, we have $\sqrt{(22-20)^2+(1-0)^2+(42-36)^2+(10-8)^2} = \sqrt{45} = 6.7082$.

(b) Compute the *Manhattan distance* between the two objects.
The Manhattan distance is computed using Equation (2.7). Therefore, we have $|22-20|+|1-0|+|42-36|+|10-8| = 11$.

(c) Compute the *Minkowski distance* between the two objects, using $h = 3$.
The Minkowski disance is

$$d(i,j) = \sqrt[h]{|x_{i1}-x_{j1}|^h + |x_{i2}-x_{j2}|^h + \cdot + |x_{ip}-x_{jp}|^h} \tag{2.10}$$

where h is a real number such that $h \geq 1$.
Therefore, with $h = 3$, we have $\sqrt[3]{|22-20|^3+|1-0|^3+|42-36|^3+|10-8|^3} = \sqrt[3]{233} = 6.1534$.