CPSC 304 Tutorial 10

Tristan Rice, q7w9a, 25886145

Question 1

(a) Define a star schema for this data warehouse [5 Marks]

Sales Fact table

- type (FK)
- cid (FK)
- CalendarKey (FK)
- UnitsSold: integer
- GrossRevenue: integer

Calendar Dimension

- *CalendarKey*: integer
- day: integer
- week: integer
- month: integer
- quarter: integer
- year: integer

Customer Dimension

- *cid*: integer
- city: string
- zip: string
- rating: integer
- salary: enum(very low, low, medium, high, very high)

ProductCategory Dimension

- *type*: string
- category: string

(b) Define a snowflake schema for this data warehouse [5 Marks]

According to Wikipedia, a star schema is just a special case of snowflake schema. Thus, the answer to 1.a applies here as well.

Sales Fact table

- type (FK)
- cid (FK)
- CalendarKey (FK)
- UnitsSold: integer
- GrossRevenue: integer

Calendar Dimension

- *CalendarKey*: integer
- day: integer
- weekID (FK)

Week Dimension

- *weekID*: integer
- week: integer
- monthID (FK)

Month Dimension

- *monthID*: integer
- month: integer
- quarterID (FK)

Quarter Dimension

- *quarterID*: integer
- quarter: integer
- yearID (FK)

Year Dimension

- *yearID*: integer
- year: integer

Customer Dimension

- *cid*: integer
- cityID (FK)
- zipID (FK)
- ratingID (FK)
- salaryID (FK)

City Dimension

- *cityID*: integer
- city: string

ZIP Dimension

- *zipID*: integer
- zip: string

Rating Dimension

- *ratingID*: integer
- rating: integer

Salary Dimension

- *salaryID*: integer
- salary: enum(very low, low, medium, high, very high)

ProductCategory Dimension

- *type*: string
- categoryID (FK)

Category Dimension

- *categoryID*: integer
- category: string

(c) Assume the chosen schema is star schema. How many different cuboids (cubes and sub-cubes) can be there in this warehouse? [2 Marks]

Number of cuboids is given by:

$$T = \Pi_{i=1}^{n}(L_i + 1)$$

We have a rather high level of dimensionality since there can be an arbitrary number of ratings, city, zip code, product type, product category, and years.

The fixed dimensions are:

- salary with five levels
- day, 7
- week, 52
- month, 12
- quarter, 4

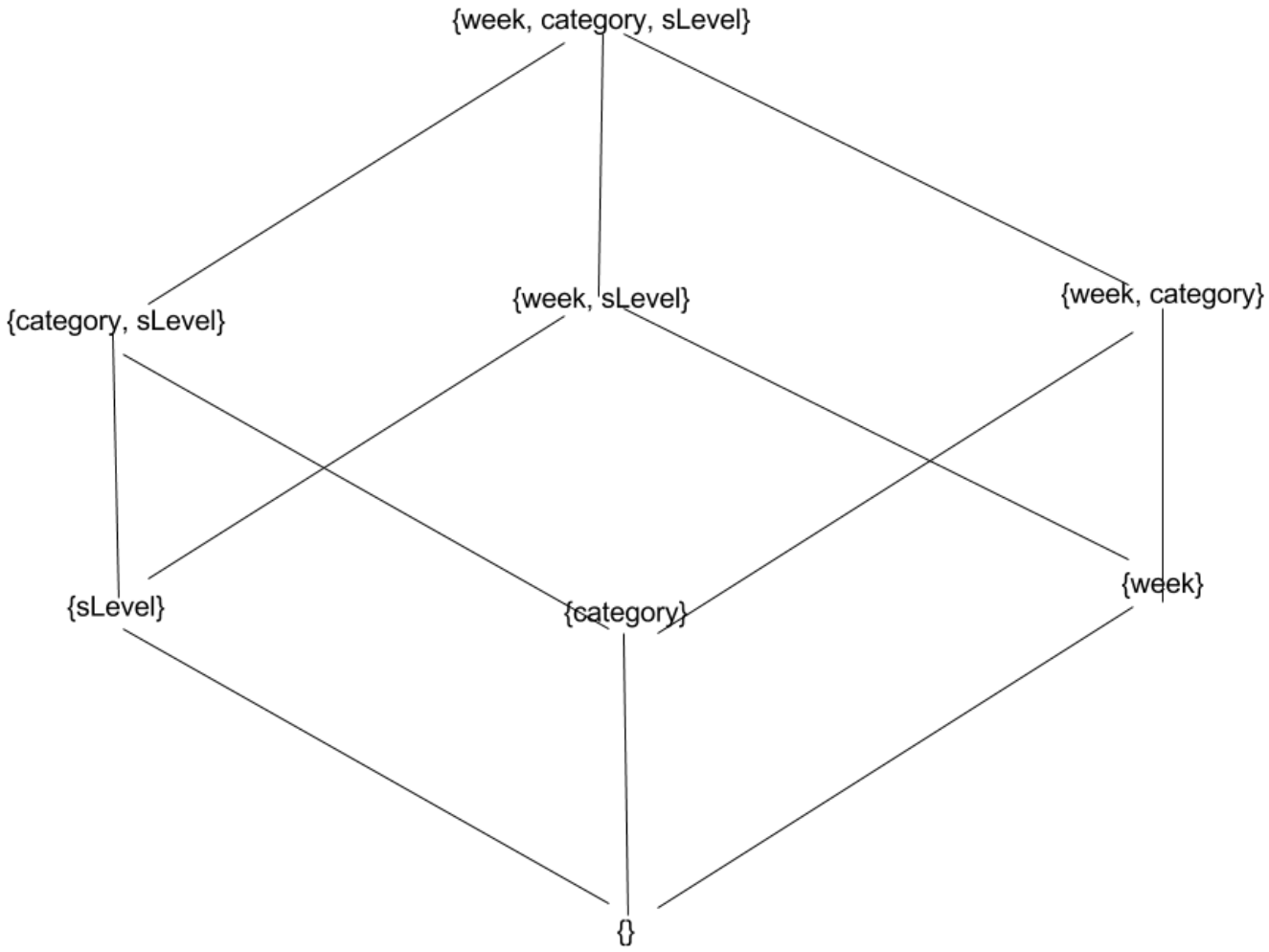
Thus the number of cuboids is:

$$(1+5)(1+\#ratings)(1+\#cities)(1+\#zipcodes)(1+\#producttype)(1+\#productcategory)(1+7)(1+52)(1+12)(1+4)(1+\#years)$$

(d) Using SQL's CUBE operator define a pivoting operation that aggregates the sales of the year 2016 on weeks, product category and salary level. [5 Marks]

```
SELECT weeks, category, salary, sum(GrossRevenue)
FROM s Sales, c Calendar, p ProductCategory, cust Customer
GROUP BY weeks, category, salary
WHERE s.type = p.type AND s.CalendarKey = c.CalendarKey AND s.cid = cust.cid
WITH CUBE
```

(e) Draw the lattice with the aggregations performed by the operator CUBE(week, category, sLevel) [5 Marks]



Question 2

Using HRU.

View	Iteration 1	Iteration 2
{c,p}	6M * 4 = 24M	Materialized
{p,s}	2M * 4 = 8M	2M * 4 = 8M
{c,s}	4M * 4 = 16M	4M * 4 = 16M
{c}	6.5M * 2 = 11M	0.5M * 2 = 1M
{s}	8M * 2 = 16M	8M * 2 = 16M
{p}	6.8M * 2 = 13.6M	0.8M * 2 = 1.6M
{}	10M - 1	4M - 1

Best view to materialize is {c,p}. The second best view to materialize is tied between {c,s} and {s}.

Question 3

a) Grouping the customers of a company according to their profitability.

This is a datamining task since it's extracting valuable information from existing data.

b) Predicting the outcomes of tossing a (fair) pair of dice.

This is not a data mining task since it's simple probability and doesn't require any stored data.

c) Predicting the future stock price of a company using historical data.

This is a data mining task since it requires a vast amount of historical data and algorithms to produce any meaningful output.

Question 4

Transaction ID	Items
T1	Milk, Bread, Mayo
T2	Milk, Bread
T3	Milk, Egg, Cheese
T4	Cheese, Egg
T5	Cheese, Mayo
T6	Milk, Egg, Cheese

1) Trace the Apriori algorithm on the below transactions with minsup = 33.3% and minconf = 60%.

Minsup = 33.3% = 2/6

Kth pass	Candidate k itemset	supported value	Frequent k itemset1
1	{Milk}	4 / 6	Yes
1	{Bread}	2 / 6	Yes
1	{Mayo}	2 / 6	Yes
1	{Cheese}	4 / 6	Yes
1	{Egg}	3 / 6	Yes
2	{Milk, Bread}	2 / 6	Yes
2	{Milk, Mayo}	1 / 6	No
2	{Milk, Egg}	2 / 6	Yes
2	{Milk, Cheese}	2 / 6	Yes
2	{Bread, Mayo}	1 / 6	No
2	{Mayo, Cheese}	1 / 6	No
2	{Cheese, Egg}	3 / 6	Yes
3	{Milk, Cheese, Egg}	2 / 6	Yes

2) Final list of frequent item set

- {Milk, Bread}
- {Milk, Egg}
- {Milk, Cheese}
- {Cheese, Egg}
- {Milk, Cheese, Egg}

3) List of association rules with support and confidence value

Item sets	Generated Rules	Support values	Conf value
{Milk, Bread}	{Milk} $\rightarrow$ {Bread}	2 / 6	2/4 = 50%
{Milk, Bread}	{Bread} $\rightarrow$ {Milk}	2 / 6	2/2 = 100%
{Milk, Egg}	{Milk} $\rightarrow$ {Egg}	2 / 6	2/4 = 50%
{Milk, Egg}	{Egg} $\rightarrow$ {Milk}	2 / 6	2/3 = 66.6%
{Milk, Cheese}	{Milk} $\rightarrow$ {Cheese}	2 / 6	2/4 = 50%

Item sets	Generated Rules	Support values	Conf value
{Milk, Cheese}	{Cheese} → {Milk}	2 / 6	2/4 = 50%
{Cheese, Egg}	{Cheese} → {Egg}	3 / 6	3/4 = 75%
{Cheese, Egg}	{Egg} → {Cheese}	3 / 6	3/3 = 100%
{Milk, Cheese, Egg}	{Milk, Cheese} → {Egg}	2 / 6	2/2 = 100%
{Milk, Cheese, Egg}	{Milk, Egg} → {Cheese}	2 / 6	2/2 = 100%
{Milk, Cheese, Egg}	{Cheese, Egg} → {Milk}	2 / 6	2/3 = 66.6%

4) List of valid association rules based on minconf =60% (sorted by conf value)

Association Rules	Conf
{Bread} → {Milk}	2/2 = 100%
{Egg} → {Cheese}	3/3 = 100%
{Milk, Cheese} → {Egg}	2/2 = 100%
{Milk, Egg} → {Cheese}	2/2 = 100%
{Cheese} → {Egg}	3/4 = 75%
{Cheese, Egg} → {Milk}	2/3 = 66.6%
{Egg} → {Milk}	2/3 = 66.6%