Name	CWID

Final Exam

May 4th, 2020
8:00-11:00

CS520 - Data Integration, Warehousing, and Provenance

Please leave this empty!

1.1 □ 1.2 □ Sum □

Instructions

- **When writing a query, write the query in a way that it would work over all possible database instances and not just for the given example instance!**

Consider the following datawarehouse schema (star schema) and partial example instance. There is a single fact table (`sales`) about sales of items. Each row in this fact table stores the quantity of a certain product (e.g., 3 Samson Galaxy phones) sold at a particular location and time to a particular customer. There are four dimension tables corresponding to the following dimensions:

- **Time** with three levels (year, month, day)
- **Location** with four levels (state, city, zip, street)
- **Customer** with one level (name). Customers are associated with a location through a foreign key (attribute CLID) to dimension **Location**.
- **Product** with three levels (category, brand, pname, price) where pname is the finest granularity and brand and category are not comparable (some brands can have products from multiple categories and categories obviously can contain have products from different brands). The same holds for price and brand and price and category.

sales

TID	LID	CID	PID	numItems
1	4	1	1	15
2	1	5	2	10
100	1	76	4	22
...	...	...	...	...

timeDim

TID	year	month	day
1	2010	1	1
2	2010	1	2
...	...	...	...
...	2018	5	1

locationDim

LID	state	city	zip	street
1	Illinois	Chicago	60616	10 W 31st
2	Illinois	Chicago	60615	900 Cottage Grove
3	Lousiana	New Orleans	42345	12 Mark street
...	...	...	...	...

customerDim

CID	cname	CLID
1	Noekig	1
2	Prokig	3
...	...	...

productDim

PID	category	brand	pname	price
1	computers	Apple	MacBook	1300
2	computers	Dell	Inspire	1000
3	smartphones	Samsung	Galaxy 1	600
...	...	...	...	...

Hints:

- Attributes with black background form the primary key of a relation (e.g., `PID` for relation `productDim`)
- Attributes `LID`, `TID`, `PID`, and `CID` in the fact table are foreign keys to the dimension tables
- Attribute `CLID` of relation `customerDim` is a foreign key to relation `locationDim`

Part 1.1 Data Warehousing (Total: 60 Points)

Recall that you should write all queries according to the schema and not according to the example instance.

Question 1.1.1 (8 Points)

Write an SQL query that returns for each city, the 3 zip codes with the highest total revenue (calculated as the sum over the number of items sold multiplied by their price).

Question 1.1.2 (8 Points)

Write an SQL query that returns for each product category and year, the accumulative sum of the revenue over the months, i.e., the result for Feb 2010 would be the sum of revenue for 2010 Jan and Feb, for Mar 2010 return the sum of 2010 Jan, Feb, Mar, and so on.

Question 1.1.3 (8 Points)

Write an SQL query that returns the total revenue for each level of the time dimension.

Question 1.1.4 (9 Points)

Write an SQL query that returns the three customers per state (location of the customer) that generated the largest revenue in their state. For example, for the result for Illinois only include customers from Illinois and only consider the revenue that they produced (by buying items) in Illinois.

Question 1.1.5 (9 Points)

Return the three years with the largest growth factor. The growth factor for a year should be calculated as the sum of the revenue for that year divided by the average of the sums of revenue for the three preceding years.

Question 1.1.6 (9 Points)

Write an SQL query that returns for brands Apple, Dell, and IBM the number of years where items of this brand sold during the year produced the greatest revenue among all brands.

Question 1.1.7 (9 Points)

Write an SQL query that returns for each customer the percentage of items they brought per brand. For example, if a customer buys 10 Apple products and 50 products in total then you should return 20% as the fraction for Apple for this customer.

Part 1.2 Provenance (Total: 40 Points)

For each of the queries shown in the following compute the provenance of all of their result tuples produced over the database shown below. Calculate provenance for these provenance models:

- Why-Provenance
- Minimal Why-Provenance
- Provenance Polynomials

Before showing the provenance, first write down the results of the query and label the result tuples t_1, t_2, $\ldots$, t_n.

Consider the following database schema and instance:

hotel

hname	location	owner	roomscnt	
Astor	Chicago	Bob	20	h_1
Blackstone	Chicago	Bob	120	h_2
Seaside	Miami	Alice	12	h_3

booking

hotel	tourist	startdate	enddate	rooms	rate	
Astor	Peter	2004-01-01	2004-01-03	1	230	b_1
Astor	Tilda	2004-01-05	2004-01-06	1	200	b_2
Seaside	Peter	2004-02-20	2004-02-30	2	75	b_3
Blackstone	Alice	2004-03-01	2004-03-05	1	140	b_4

person

name	age	location	
Peter	43	Chicago	p_1
Bob	24	Madison	p_2
Alice	25	Chicago	p_3
Tilda	32	Miami	p_4

location

loc	state	country	
Chicago	43	USA	l_1
Madison	24	USA	l_2
Miami	25	USA	l_3

Question 1.2.1 (7 Points)

$$\pi_{location}(\sigma_{roomcnt>15}(hotel))$$

Question 1.2.2 (9 Points)

$$\begin{aligned}
q_1 &\overset{def}{=} booking \bowtie_{hotel=hname} hotel \bowtie_{loc=location} location\\
q_2 &\overset{def}{=} \rho_{plocation \leftarrow location}(person) \bowtie_{plocation=ploc} \rho_{ploc \leftarrow loc, pstate \leftarrow state, pcountry \leftarrow country}(location)\\
q &\overset{def}{=} \pi_{hname,plocation,pcountry,location,country}(q_1 \bowtie_{tourist=name} q_2)
\end{aligned}$$

Question 1.2.3 (8 Points)

$$q \overset{def}{=} \pi_{location}(\sigma_{country=USA}(person \bowtie_{location=loc} location))$$

Question 1.2.4 (8 Points)

$$\begin{aligned} q_1 &\overset{def}{=} \pi_{hotel,date}(\rho_{date \leftarrow startdate}(booking)) \\ q_2 &\overset{def}{=} \pi_{hotel,date}(\rho_{date \leftarrow enddate}(booking)) \\ q &\overset{def}{=} q_1 \cup q_2 \end{aligned}$$

Question 1.2.5 (8 Points)

$$q \overset{def}{=} \pi_{hname,name}(person \bowtie hotel)$$