

## CHAPTER 20



# Data Analysis and Mining

### Practice Exercises

- 20.1 Describe benefits and drawbacks of a source-driven architecture for gathering of data at a data warehouse, as compared to a destination-driven architecture.

**Answer:** In a destination-driven architecture for gathering data, data transfers from the data sources to the data-warehouse are based on demand from the warehouse, whereas in a source-driven architecture, the transfers are initiated by each source.

The benefits of a source-driven architecture are

- Data can be propagated to the destination as soon as it becomes available. For a destination-driven architecture to collect data as soon as it is available, the warehouse would have to probe the sources frequently, leading to a high overhead.
- The source does not have to keep historical information. As soon as data is updated, the source can send an update message to the destination and forget the history of the updates. In contrast, in a destination-driven architecture, each source has to maintain a history of data which have not yet been collected by the data warehouse. Thus storage requirements at the source are lower for a source-driven architecture.

On the other hand, a destination-driven architecture has the following advantages.

- In a source-driven architecture, the source has to be active and must handle error conditions such as not being able to contact the warehouse for some time. It is easier to implement passive sources, and a single active warehouse. In a destination-driven architecture, each source is required to provide only a basic functionality of executing queries.
- The warehouse has more control on when to carry out data gathering activities, and when to process user queries; it is not a good

idea to perform both simultaneously, since they may conflict on locks.

- 20.2 Why is column-oriented storage potentially advantageous in a database system that supports a data warehouse?

**Answer:** No Answer

- 20.3 Suppose that there are two classification rules, one that says that people with salaries between \$10,000 and \$20,000 have a credit rating of *good*, and another that says that people with salaries between \$20,000 and \$30,000 have a credit rating of *good*. Under what conditions can the rules be replaced, without any loss of information, by a single rule that says people with salaries between \$10,000 and \$30,000 have a credit rating of *good*?

**Answer:** Consider the following pair of rules and their confidence levels :

No.	Rule	Conf.
1.	$\forall \text{ persons } P, 10000 < P.\text{salary} \leq 20000 \Rightarrow P.\text{credit} = \text{good}$	60%
2.	$\forall \text{ persons } P, 20000 < P.\text{salary} \leq 30000 \Rightarrow P.\text{credit} = \text{good}$	90%

The new rule has to be assigned a confidence-level which is between the confidence-levels for rules 1 and 2. Replacing the original rules by the new rule will result in a loss of confidence-level information for classifying persons, since we cannot distinguish the confidence levels of people earning between 10000 and 20000 from those of people earning between 20000 and 30000. Therefore we can combine the two rules without loss of information only if their confidences are the same.

- 20.4 Consider the schema depicted in Figure 20.2. Give an SQL query to summarize sales numbers and price by store and date, along with the hierarchies on store and date.

**Answer:** query:

```
select store-id, city, state, country,
       date, month, quarter, year,
       sum(number), sum(price)
from sales, store, date
where sales.store-id = store.store-id and
      sales.date = date.date
group by rollup(country, state, city, store-id),
         rollup(year, quarter, month, date)
```

- 20.5 Consider a classification problem where the classifier predicts whether a person has a particular disease. Suppose that 95% of the people tested do not suffer from the disease. (That is, *pos* corresponds to 5% and *neg* to 95% of the test cases.) Consider the following classifiers:

- Classifier  $C_1$  which always predicts negative (a rather useless classifier of course).
- Classifier  $C_2$  which predicts positive in 80% of the cases where the person actually has the disease, but also predicts positive in 5% of the cases where the person does not have the disease.
- Classifier  $C_3$  which predicts positive in 95% of the cases where the person actually has the disease, but also predicts positive in 20% of the cases where the person does not have the disease.

Given the above classifiers, answer the following questions.

- a. For each of the above classifiers, compute the accuracy, precision, recall and specificity.
- b. If you intend to use the results of classification to perform further screening for the disease, how would you choose between the classifiers.
- c. On the other hand, if you intend to use the result of classification to start medication, where the medication could have harmful effects if given to someone who does not have the disease, how would you choose between the classifiers?

**Answer:** No Answer

