Problems and Exercises

1. Examine the three tables with student data shown in Figure 11-1. Design a single table format that will hold all of the data (nonredundantly) that are contained in these three tables. Choose column names that you believe are most appropriate for these data.
2. The following table shows some simple student data as of the date 06/20/2004:

Key	*Name*	*Major*
001	Amy	Music
002	Tom	Business
003	Sue	Art
004	Joe	Math
005	Ann	Engineering

The following transactions occur on 06/21/2004:

- Student 004 changes major from 'Math' to 'Business.'
- Student 005 is deleted from the file.
- New student 006 is added to the file: Name is 'Jim,' Major is 'Phys Ed.'

The following transactions occur on 06/22/2004:

- Student 003 changes major from 'Art' to 'History.'
- Student 006 changes major from 'Phys Ed' to 'Basket Weaving.'

Your assignment is in two parts:

a. Construct tables for 06/21/2004 and 06/22/2004 reflecting these transactions, assume that the data are transient (refer to Figure 11-8).

b. Construct tables for 06/21/2004 and 06/22/2004 reflecting these transactions, assume that the data are periodic (refer to Figure 11-9).

3. Millennium College wants you to help them design a star schema to record grades for courses completed by students. There are four dimension tables, with attributes as follows:

- **Course_Section**. Attributes: Course_ID, Section_Number, Course_Name, Units, Room_ID, Room_Capacity. During a given semester the college offers an average of 500 course sections.
- **Professor**. Attributes: Prof_ID, Prof_Name, Title, Department_ID, Department_Name.
- **Student**. Attributes: Student_ID, Student_Name, Major. Each course section has an average of forty students.
- **Period**. Attributes: Semester_ID, Year. The database will contain data for thirty periods (a total of 10 years).

The only fact that is to be recorded in the fact table is Course_Grade.

a. Design a star schema for this problem. See Figure 11-14 for the format you should follow.

b. Estimate the number of rows in the fact table, using the assumptions stated above.

c. Estimate the total size of the fact table (in bytes), assuming that each field has an average of five bytes.

d. Various characteristics of sections, professors, and students change over time. How do you propose designing the star schema to allow for these changes? Why?

4. Having mastered the principles of normalization described in Chapter 5, you recognize immediately that the star schema you developed for Millennium College (Problem and Exercise 3) is not in third normal form. Using these principles, convert the star schema to a snowflake schema. What impact (if any) does this have on the size of the fact table for this problem?
5. You are to construct a star schema for Simplified Automobile Insurance Company (see Kimball, 1996b, for a more realistic example). The relevant dimensions, dimension attributes, and dimension sizes are as follows:

- **Insured Party**. Attributes: Insured_Party_ID, Name. There is an average of two insured parties for each policy and covered item.
- **Coverage Item**. Attributes: Coverage_Key, Description. There is an average of ten covered items per policy.
- **Agent**. Attributes: Agent_ID, Agent_Name. There is one agent for each policy and covered item.
- **Policy**. Attributes: Policy_ID, Type. The company has approximately one million policies at the present time.
- **Period**. Attributes: Date_Key, Fiscal_Period.

Facts to be recorded for each combination of these dimensions are Policy_Premium, Deductible, and Number_of_Transactions.

a. Design a star schema for this problem. See Figure 11-14 for the format you should follow.

b. Estimate the number of rows in the fact table, using the assumptions stated above.

c. Estimate the total size of the fact table (in bytes), assuming an average of five bytes per field.

6. Simplified Automobile Insurance Company would like to add a Claims dimension to its star schema (see Problem and Exercise 5). Attributes of Claim are Claim_ID, Claim_Description, and Claim_Type. Attributes of the fact table are now Policy_Premium, Deductible, and Monthly_Claim__Total.

a. Extend the star schema from Problem and Exercise 5 to include these new data.

b. Calculate the estimated number of rows in the fact table, assuming that the company experiences an average of 2,000 claims per month.

7. Millennium College (see Problem and Exercise 3) now wants to include new data about course sections: the department offering the course, the academic unit to which the department reports, and the budget unit to which the department is assigned. Change your answer to Problem and Exercise 3 to accommodate these new data requirements. Explain why you implemented the changes in the star schema the way you did.
8. As mentioned in the chapter, Kimball (1997), Inmon (1997 and 2000), and Armstrong (2000) have debated the merits of independent and dependent data marts and normalized

versus denormalized data marts. Obtain copies of these articles from your library or from online sources and summarize the arguments made by each side of this debate. See also **www.intelligententerprise.com/030917/615warehouse1_1.shtml** for a recent article clarifying the Kimball position.

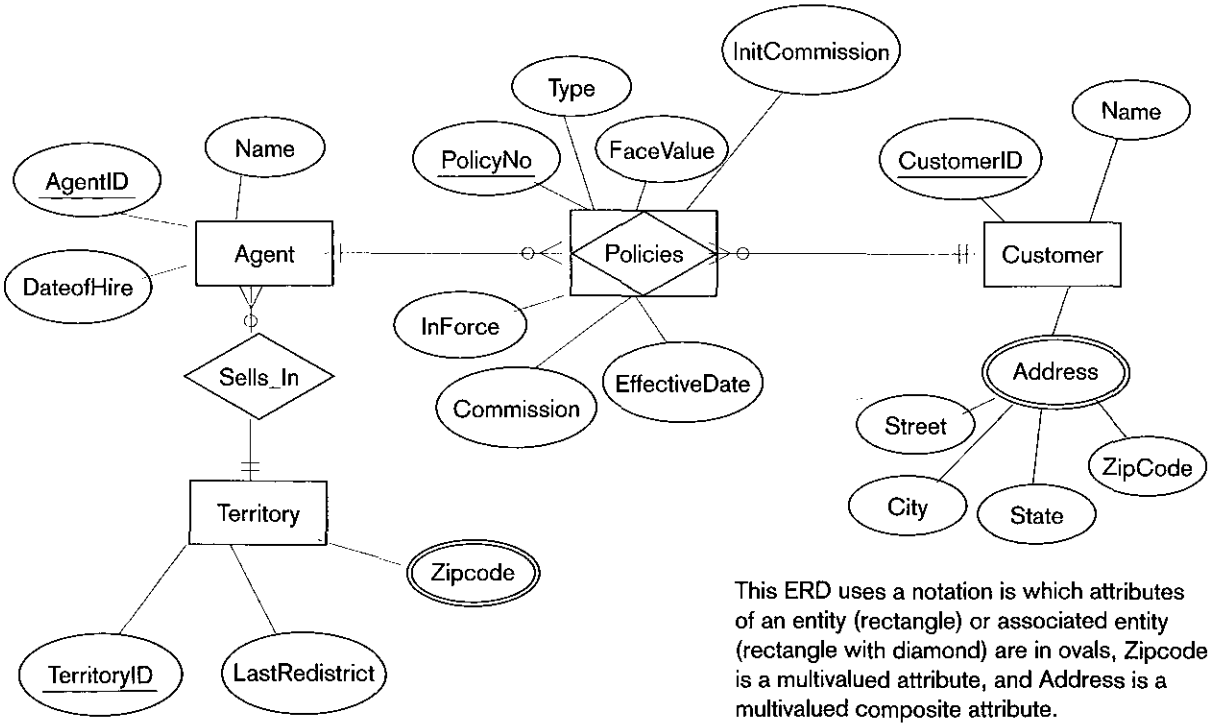
9. A food manufacturing company needs a data mart to summarize facts about orders to move goods. Some orders transfer goods internally, some are sales to customers, some are purchases from vendors, and some are returns of goods from customers. The company needs to treat customers, vendors, plants, and storage locations as distinct dimensions that can be involved at both ends of a movement event. For each type of destination or origin, the company wants to know the type of location (i.e., customer, vendor, etc.), name, city, and state. Facts about each movement include dollar volume moved, cost of movement, and revenue collected from the move (if any, and this can be negative for a return). Design a star-type schema to represent this data mart. Hint: After you design a typical star schema, think about how you might simplify the design through the use of generalization.
10. Visit **www.ralphkimball.com** and locate Kimball University Design Tip #37. Study this design tip and draw the dimensional model for the recommended design for a "pipeline" application for university admissions.
11. Visit **www.teradatastudentnetwork.com** and download the dimensional modeling tool located under the software section. Use this tool to draw your answers to Problems and Exercises 3, 5, 6, and 9. Write a report that comments on the usefulness of this modeling tool. What other features would you like the tool to have?
12. Visit the Problems and Exercises material on **www.prenhall.com/hoffer** for Chapter 11 and answer the data warehouse and processing questions located there. These questions deal with a data warehouse for Pine Valley Furniture; there are questions for you to design a data mart for Pine Valley and to write some queries against an instance of this data mart.

Problems 13–20 are based upon the Fitchwood Insurance Company case study, which is described below.

Fitchwood Insurance Company, which is primarily involved in the sales of annuity products, would like to design a data mart for its sales and marketing organization. Presently, the OLTP system is a legacy system residing on a Novell network consisting of approximately 600 different flat files. For the purposes of our case study, we can assume that thirty different flat files are going to be used for the data mart. Some of these flat files are transaction files that change constantly. The OLTP system is shut down overnight on Friday evening beginning at 6 PM for backup. During that time, the flat files are copied to another server, an extraction process is run, and the extracts are sent via FTP to a UNIX server. A process is run on the UNIX server to load the extracts into Oracle and rebuild the star schema. For the initial loading of the data mart, all information from the thirty files was extracted and loaded. On a weekly basis, only additions and updates will be included in the extracts.

Although the data contained in the OLTP system are broad, the sales and marketing organization would like to focus on the sales data only. After substantial analysis, the ERD shown in Figure 11-25 was developed to describe the data to be used to populate the data mart.

Figure 11-25
Fitchwood Insurance Company ERD



This ERD uses a notation is which attributes of an entity (rectangle) or associated entity (rectangle with diamond) are in ovals, Zipcode is a multivalued attribute, and Address is a multivalued composite attribute.

From this ERD, we get the set of relations shown in Figure 11-26. Sales and marketing is interested in viewing all sales data by territory, effective date, type of policy, and face value. In addition, the data mart should be able to provide reporting by individual agent on sales as well as commissions earned. Occasionally, the sales territories are revised (i.e., zip codes are added or deleted). The Last Redistrict attribute of the Territory table is used to store the date of the last revision. Some sample queries and reports are shown below:

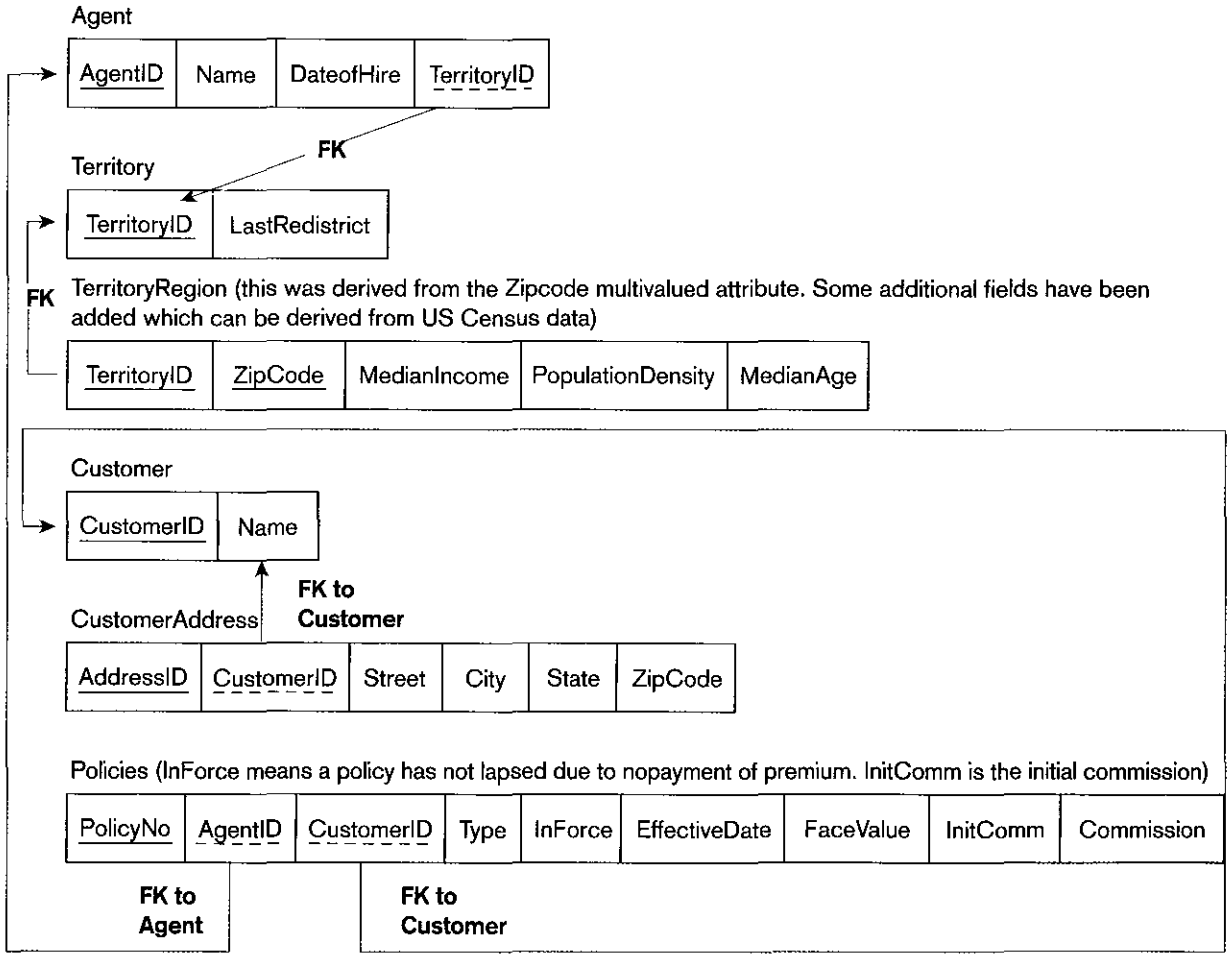
- Total sales per month by territory by type of policy
- Total sales per quarter by territory by type of policy
- Total sales per month by agent by type of policy
- Total sales per month by agent by zip code
- Total face value of policies by month of effective date
- Total face value of policies by month of effective date by agent
- Total face value of policies by quarter of effective date
- Total number of policies in force by agent
- Total number of policies not in force by agent
- Total face value of all policies sold by an individual agent
- Total initial commission paid on all policies to an agent
- Total initial commission paid on policies sold in a given month by agent
- Total commissions earned by month by agent
- Top selling agent by territory by month

Commissions are paid to an agent upon the initial sale of a policy. The InitComm field of the policy table contains the percentage of the face value paid as an initial commission. The Commission field contains a percentage that is paid each month as long as a policy remains active or in force. Each month, commissions are calculated by computing the sum of the commission on each individual policy that is in force for an agent.

13. The OLTP system data for the Fitchwood Insurance Company is in a series of flat files. What process do you envision would be needed in order to extract the data and create the ERD shown above? How often should the extraction process be performed? Should it be a static extract or an incremental extract?

14. What types of data pollution problems might occur with the Fitchwood OLTP system data?

Figure 11-26
Relations for Fitchwood Insurance Company



15. Research some tools that perform data scrubbing. What tool would you recommend for the Fitchwood Insurance Company?
16. What types of data transformations might be needed in order to build the Fitchwood data mart?
17. After some further analysis, you discover that the commission field in the Policies table is updated on a yearly basis to reflect changes in the yearly commission paid to agents on existing policies. Would this change the way in which you extract and load data into the data mart from the OLTP system?
18. Create a star schema for this case study. How did you handle the time dimension?
19. Management would like to use the data mart for drill-down online reporting. For example, a sales manager might want to view a report of total sales for an agent by month and then drill-down into the individual types of policies to see how sales were broken down by type of policy. What type of tool would you recommend for this? What additional tables, other than those required by the tool for administration, might need to be added to the data mart?
20. Do you see any opportunities for data mining using the Fitchwood data mart? Research data mining tools and recommend one or two for use with the data mart.

Field Exercises

1. Visit an organization that has developed a data warehouse and interview the data administrator or other key participant. Discuss the following issues:
   a. How satisfied are users with the data warehouse? In what ways has it improved their decision making?
   b. Does the warehouse employ a two-tier or three-tier architecture?
   c. Does the architecture employ one or more data marts? If so, are they dependent or independent?
   d. What end-user tools are employed? Is data mining used?
   e. What were the main obstacles or difficulties overcome in developing the data warehouse environment?
2. Visit the following Web sites. Browse these sites for additional information on data warehouse topics, including case examples of warehouse implementations, descriptions of the latest warehouse-related products, and announcements of conferences and other events.
   a. The Data Warehousing Institute: **www.tdwi.org**
   b. Knowledge Discovery Mine: **www.kdnuggets.com**
   c. Data Mining Institute: **www.datamining.org**
   d. Data Warehousing Knowledge Center: **www.dataware housing.org**
   e. An electronic data warehousing journal: **www.tdan.com**

References

Agosta, L. 2003. "Data Warehouse Refresh Rates." *DM Review* 13,6 (June): 49.

Armstrong, R. 1997. "A Rebuttal to the Dimensional Modeling Manifesto." A white paper produced by NCR Corporation.

Armstrong, R. 2000. "Avoiding Data Mart Traps." *Teradata Review* (Summer): 32–37.

Chisholm, M. 2000. "A New Understanding of Reference Data." *DM Review* 10,10 (October): 60, 84–85.

Devlin, B. 1997. *Data Warehouse: From Architecture to Implementation.* Reading, MA: Addison-Wesley Longman.

Devlin, B., and P. Murphy. 1988. "An Architecture for a Business Information System." *IBM Systems Journal* 27,1 (March): 60–80.

Dyché, J. 2000. *e-Data: Turning Data into Information with Data Warehousing.* Reading, MA: Addison-Wesley.

Eckerson, W., and C. White. 2003. *Evaluating ETL and Data Integration Platforms.* The Data Warehouse Institute. Available at **www.tdwi.org** under Research Reports.

English, L. P. 1999. *Improving Data Warehouse and Business Information Quality.* New York: Wiley.

Hackathorn, R. 1993. *Enterprise Database Connectivity.* New York: Wiley.

Hackathorn, R. 2002. "Current Practices in Active Data Warehousing." Available at **www.teradata.com** under White Papers.

Hays, C. 2004. "What They Know About You." *New York Times.* November 14: section 3, page 1.

Imhoff, C. 1998. "The Operational Data Store: Hammering Away." *DM Review* 8,7 (July): Available at **www.dmreview.com/master.cfm?NavID=55&EdID=470**.

Imhoff, C. 1999. "The Corporate Information Factory." *DM Review* 9,12 (December): Available at **www.dmreview.com/master.cfm?NavID=55&EdID=1667**.

Inmon, B. 1997. "Iterative Development in the Data Warehouse." *DM Review* 7,11 (November): 16, 17.

Inmon, W. 1998. "The Operational Data Store: Designing the Operational Data Store." *DM Review* 8,7 (July): Available at **www.dmreview.com/master.cfm?NavID=55&EdID=469**.

Inmon, W. 1999. "What Happens When You Have Built the Data Mart First?" *TDAN.* Available at **www.tdan.com/i012fe02.htm**.

Inmon, W. 2000. "The Problem with Dimensional Modeling." *DM Review* 10,5 (May): 68–70.

Inmon, W. H., and R. D. Hackathorn. 1994. *Using the Data Warehouse.* New York: Wiley.