# In-class exercises for Data Warehousing Practice

**Millennium College:**

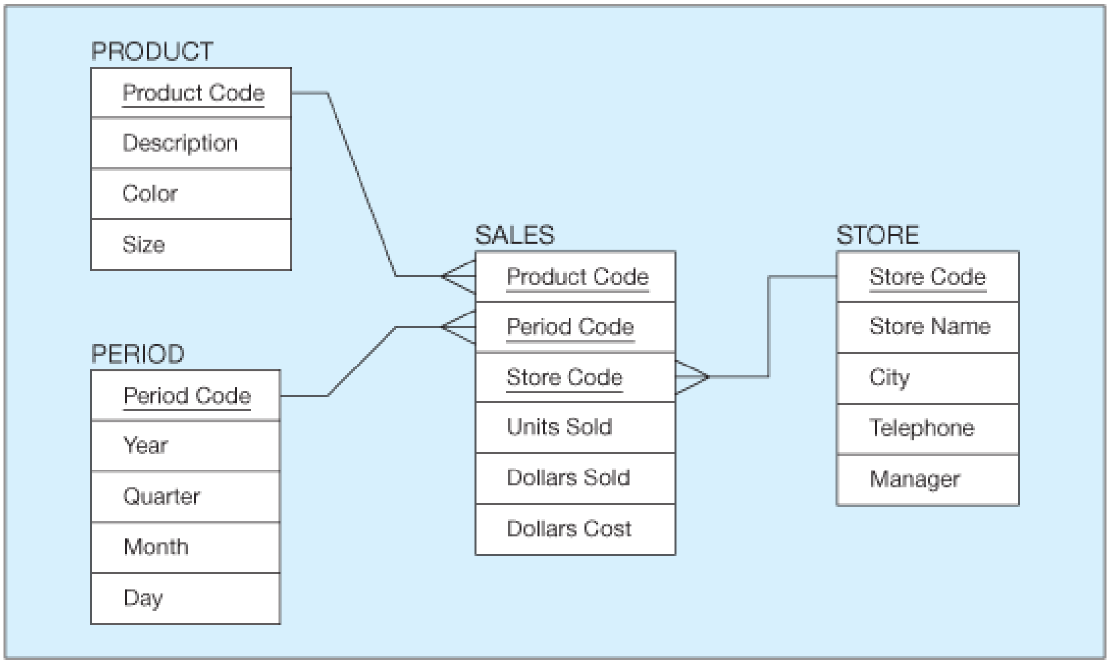
**Part 1:**

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

| | |
|---|---|
| CourseSection | Attributes: CourseID, SectionNumber, CourseName, Units, RoomID, and RoomCapacity. During a given semester, the college offers an average of 500 course sections. |
| Professor | Attributes: ProfID, ProfName, Title, DepartmentID, and DepartmentName. There are typically 200 professors at Millennium at any given time. |
| Student | Attributes: StudentID, StudentName, and Major. Each course section has an average of 40 students, and students typically take five courses per period. |
| Period | Attributes: SemesterID and Year. The database will contain data for 30 periods (a total of 10 years). |

The facts that are to be recorded in the fact table is Course Grade and CompletionDate.

a. Design a star schema for this problem. See the following figure for a format:



b. Estimate the number of rows in the fact table, using the assumptions stated previously.
c. Estimate the total size of the fact table (in bytes), assuming that each field has an average of 15 bytes.
d. If you didn't want to or didn't have to stick with a strict star schema for this data mart, how would you change the design? Why?
e. Various characteristics of sections, professors, and students change over time. How do you propose designing the star schema to allow for these changes? Why?

**Part 2:**

Having mastered the principles of normalization described in Chapter 4, you recognize immediately that the star schema you developed for Millennium College 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?

---

**Case-Study:**

Fitchwood Insurance Company, which is involved primarily in the sale 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 shared network drive consisting of approximately 600 different flat files. For the purposes of our case study, you can assume that 30 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 p.m. 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 30 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 9-24 was developed to describe the data to be used to populate the data mart.

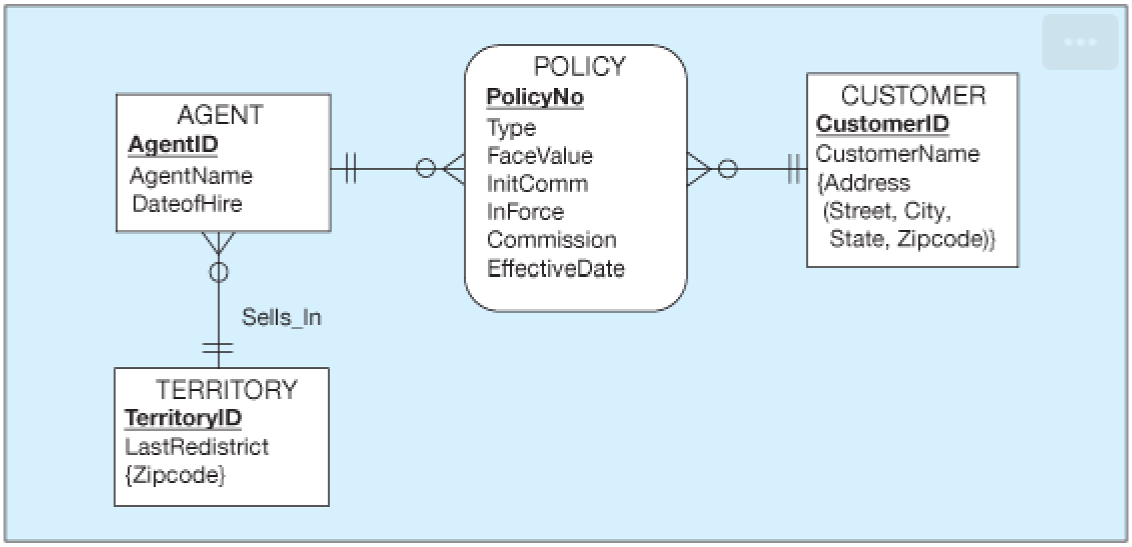
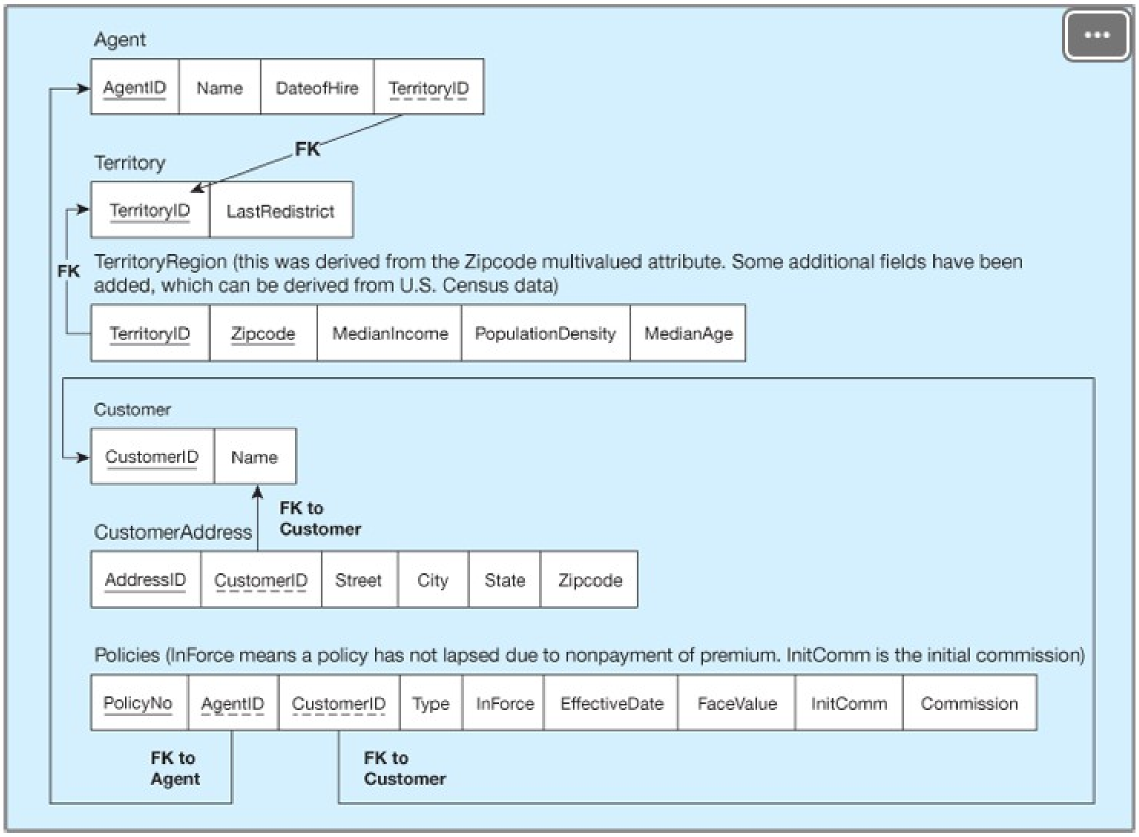


Figure 9-24 Fitchwood Insurance Company ERD

From this ERD, you get the set of relations shown in Figure 9-25. 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 listed here:

- 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.



Agent

| AgentID | Name | DateofHire | TerritoryID |
|---|---|---|---|

Territory

| TerritoryID | LastRedistrict |
|---|---|

TerritoryRegion (this was derived from the Zipcode multivalued attribute. Some additional fields have been added, which can be derived from U.S. Census data)

| TerritoryID | Zipcode | MedianIncome | PopulationDensity | MedianAge |
|---|---|---|---|---|

Customer

| CustomerID | Name |
|---|---|

CustomerAddress

| AddressID | CustomerID | Street | City | State | Zipcode |
|---|---|---|---|---|---|

Policies (InForce means a policy has not lapsed due to nonpayment of premium. InitComm is the initial commission)

| PolicyNo | AgentID | CustomerID | Type | InForce | EffectiveDate | FaceValue | InitComm | Commission |
|---|---|---|---|---|---|---|---|---|

## Relations for Fitchwood Insurance Company:

Commissions are paid to an agent on 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.

1. Create a star schema for this case study. How did you handle the time dimension?
2. Would you prefer to normalize (snowflake) the star schema of your answer to Problem and Exercise 9-38 [*Part 2 mentioned above]*? If so, how and why? Redesign the star schema to accommodate your accommodate your recommended changes.
3. Agents change territories over time. If necessary, redesign your answer to Problem and Exercise 9-47 to handle this changing dimensional data.
4. Customers may have relationships with one another (e.g., spouses or parents and children). Redesign your answer to Problem 9-48 to accommodate these relationships.
5. 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

in Figure 9-24? How often should the extraction process be performed? Should it be a static extract or an incremental extract?

6. What types of data pollution/cleansing problems might occur with the Fitchwood OLTP system data?
7. Research some tools that perform data scrubbing. What tool would you recommend for the Fitchwood Insurance Company?
8. What types of data transformations might be needed in order to build the Fitchwood data mart?
9. After some further analysis, you discover that the Commission field in the Policies table is updated yearly to reflect changes in the annual commission paid to agents on existing policies. Would knowing this information change the way in which you extract and load data into the data mart from the OLTP system?