EE4791 Database Systems -Tutorial 11 and Sample Answer

1. In a college, the following are four groups of data used to analyze grades achieved by students:

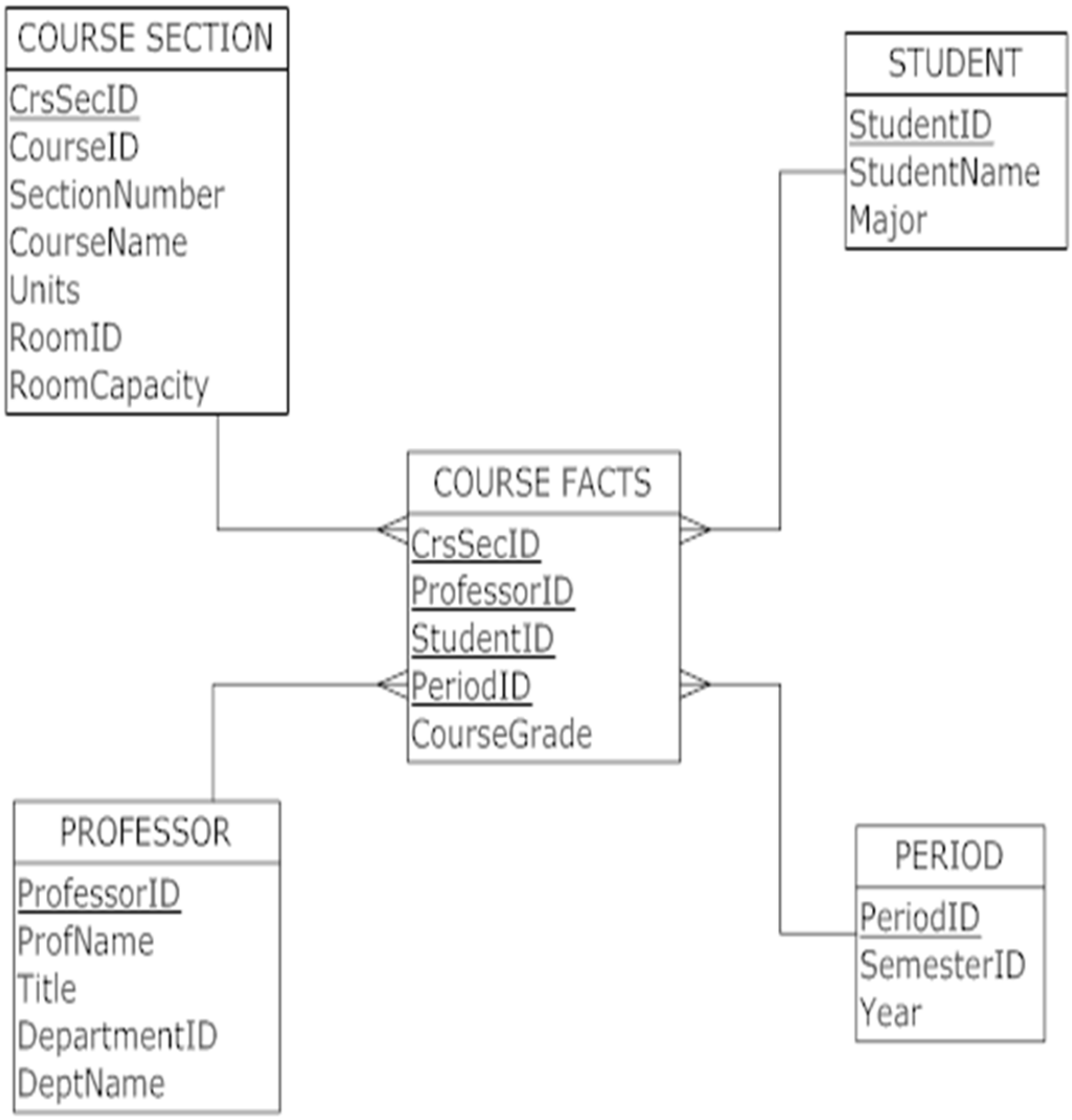
   - CourseSection Attributes: CourseID, SectionNumber, CourseName, Units, RoomID, and RoomCapacity. During a given semester, the college offers an average of 500 course sections.
   - Professor Attributes: ProfID, ProfName, Title, DepartmentID, and DepartmentName. There are typically 200 professors at the college at any given time. On the average, 2 professors teach one course section.
   - 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 only fact that is to be recorded in the fact table is CourseGrade. Do the following:

   a. Design a basic star schema for the loading the above-mentioned data in a data mart.
   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 5 bytes.

Answer

(a) Star Schema



b. 500 course sections x 2 professors per section x 40 students per section x 30 periods (i.e., 3 semesters per year) = 1, 200,000 rows

c. 1, 200,000 rows * 5 fields per row * 5 bytes per field = 30,000,000 bytes

2. In additional to the information given in Question 1, we further assume that:

- RoomID uniquely identifies RoomCapacity.
- CourseID uniquely identifies CourseNumber, CourseName and Units.
- DepartmentID uniques identifies DepartmentName

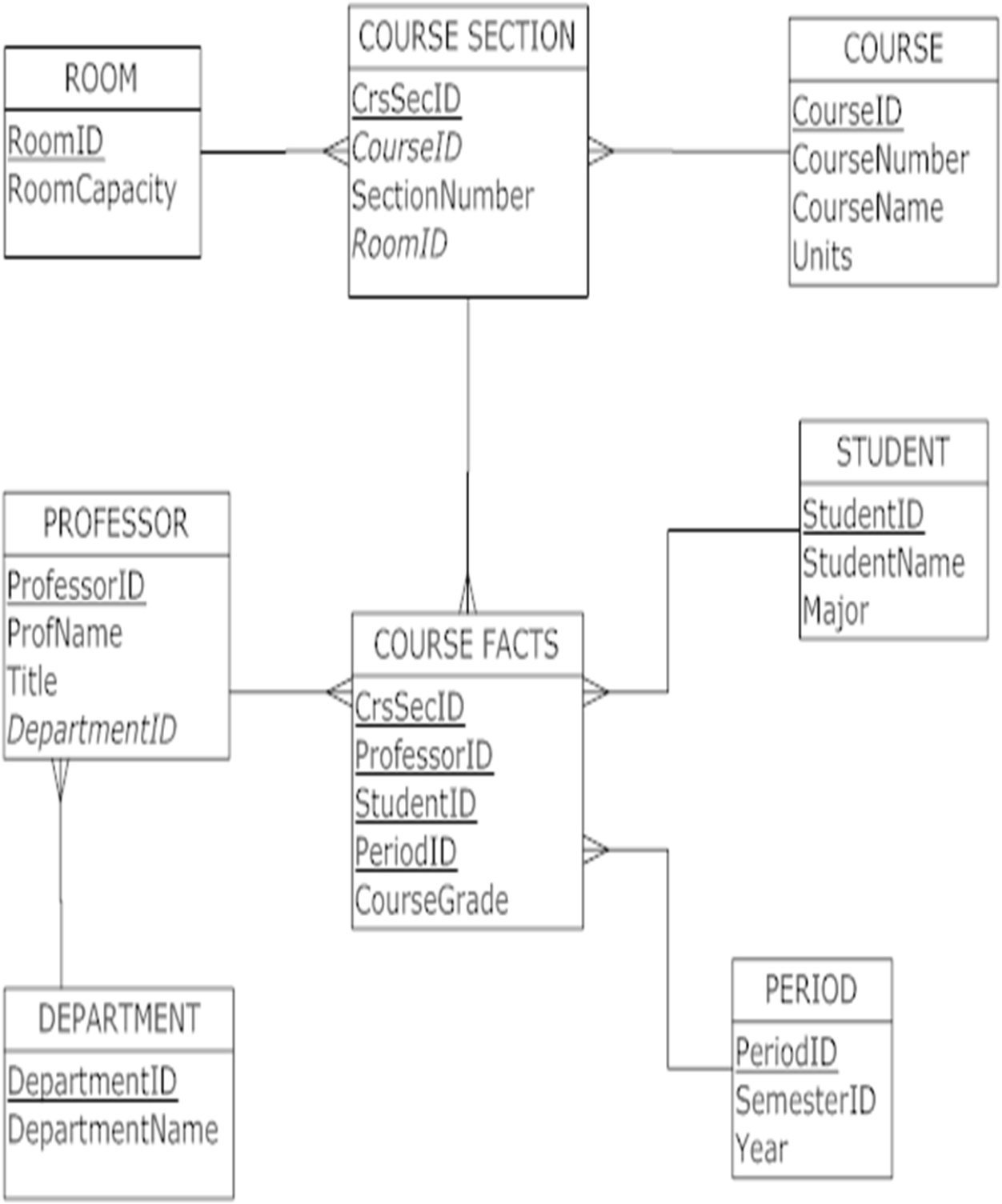
Discuss the advantages and disadvantages in using normalized dimension tables in star schemas for data marts. Based on these further assumptions, normalize the dimensional tables in the star schema designed in Question 1.

Answer

Advantages: Using normalized dimension tables will save storage due as no data will be duplicated.

Disadvantages: Using normalized dimension tables may affect the performance of some queries due to the joining of information from more tables together.

Normalized Star Schema

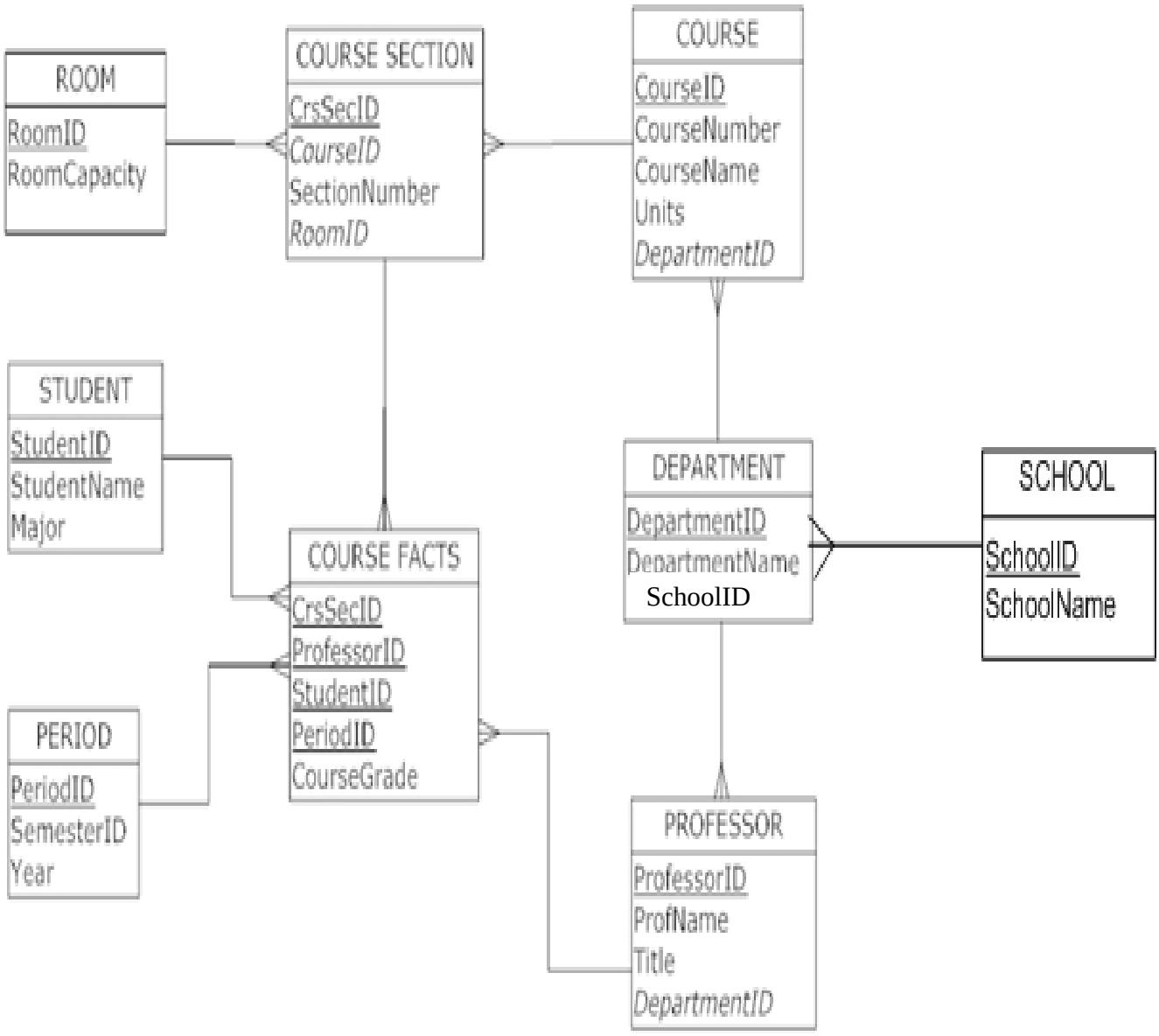


3. Assuming that the college stated in Question 1 now wants to include the following new data about course sections:

    - The department offering the course
    - The school to which the department reports

    Change the star schema designed in Question 2 to cater for the new data.

Answer



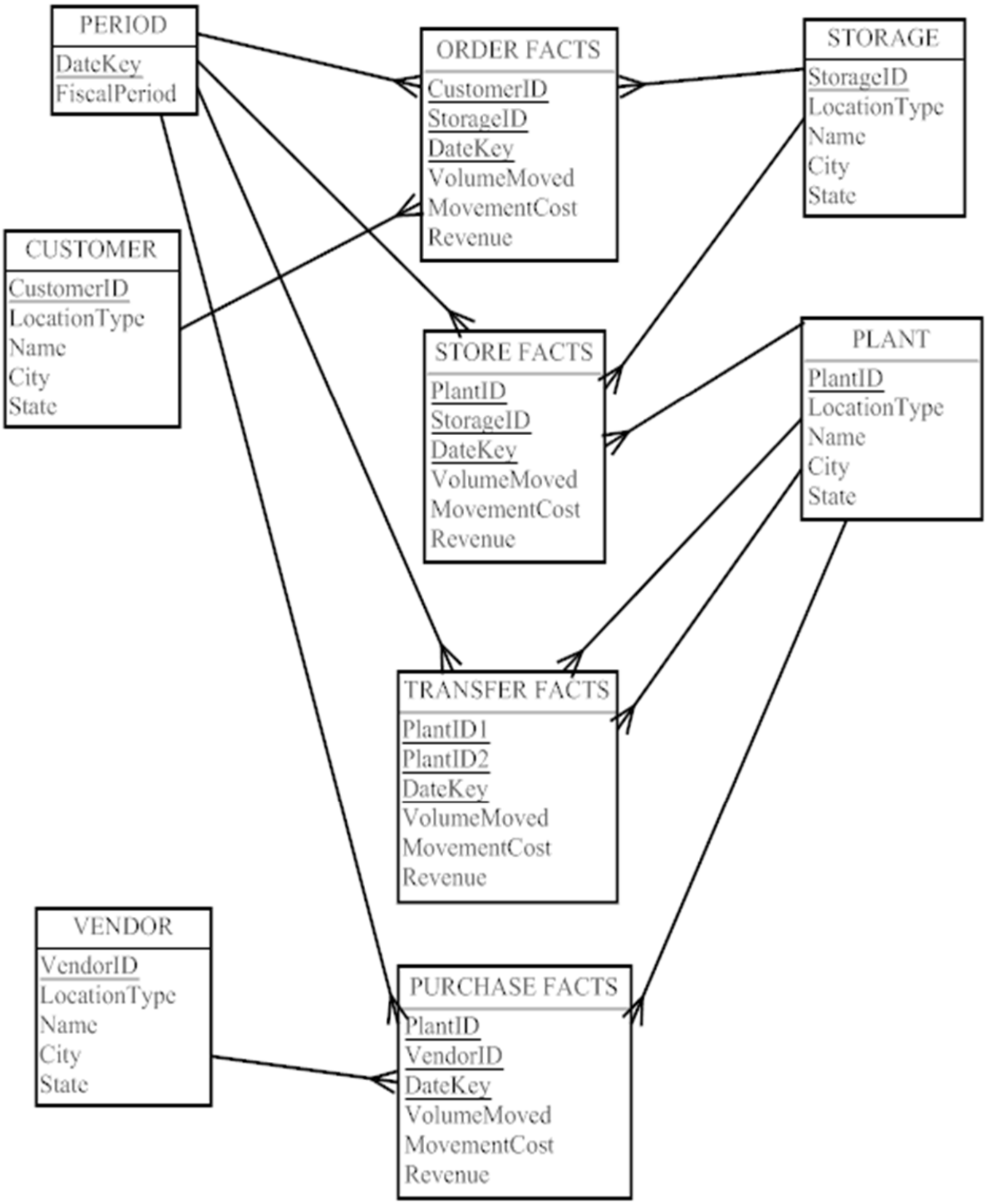
4. A food manufacturing company needs a data mart to summarize facts about the following type of orders to move goods:

    a) Transfer goods internally, between plants and from plants to storage
    b) Sales to customers from storage locations
    c) Purchases from vendors to plants
    d) Returns of goods from customers to storage locations

    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 and volume moved, cost of movement, and revenue collected from the move (if any, and this can be negative for a return). Design a star schema to represent this data mart directly. Simplify the resulting star schema through generalization.

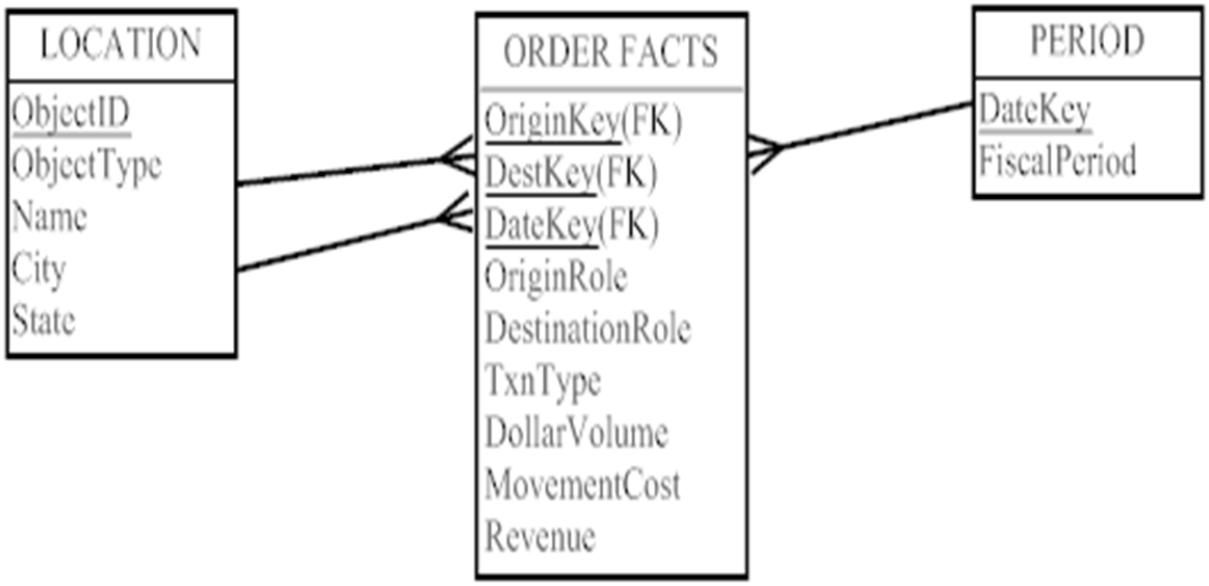
Answer

Star Schema designed directly



Star Schema through generalization

We could simplify this star schema substantially by re-designing the fact table to act more generically. Essentially, the re-designed fact table contains an origin key and a destination key. For example, if a customer was to purchase some items, the origin key would be a storage ID and the destination key would be the CustomerID. OriginRole and DestinationRole will be the name of the role for each fact related to the Location ObjectID (e.g., customer, vendor, plant). TxnType will label the type of transaction that occurred (e.g., sale, return, etc.). Also, note the use of an ObjectID as the surrogate key for the Location dimension.



5. An international pharmaceutical company operates a network of 300 chain drug stores all over the world. The company is setting up a drug data warehouse to store information for a period of 10 years drug sales analysis. The total sales of drugs (Total_Items_Sold and Total_Sales_Value) per day for each drug and for each store should be kept in the data warehouse. There is an average of 50 different drugs sold by each store per day. Data for the data warehouse are extracted from the company database. There are three relevant tables in the database:

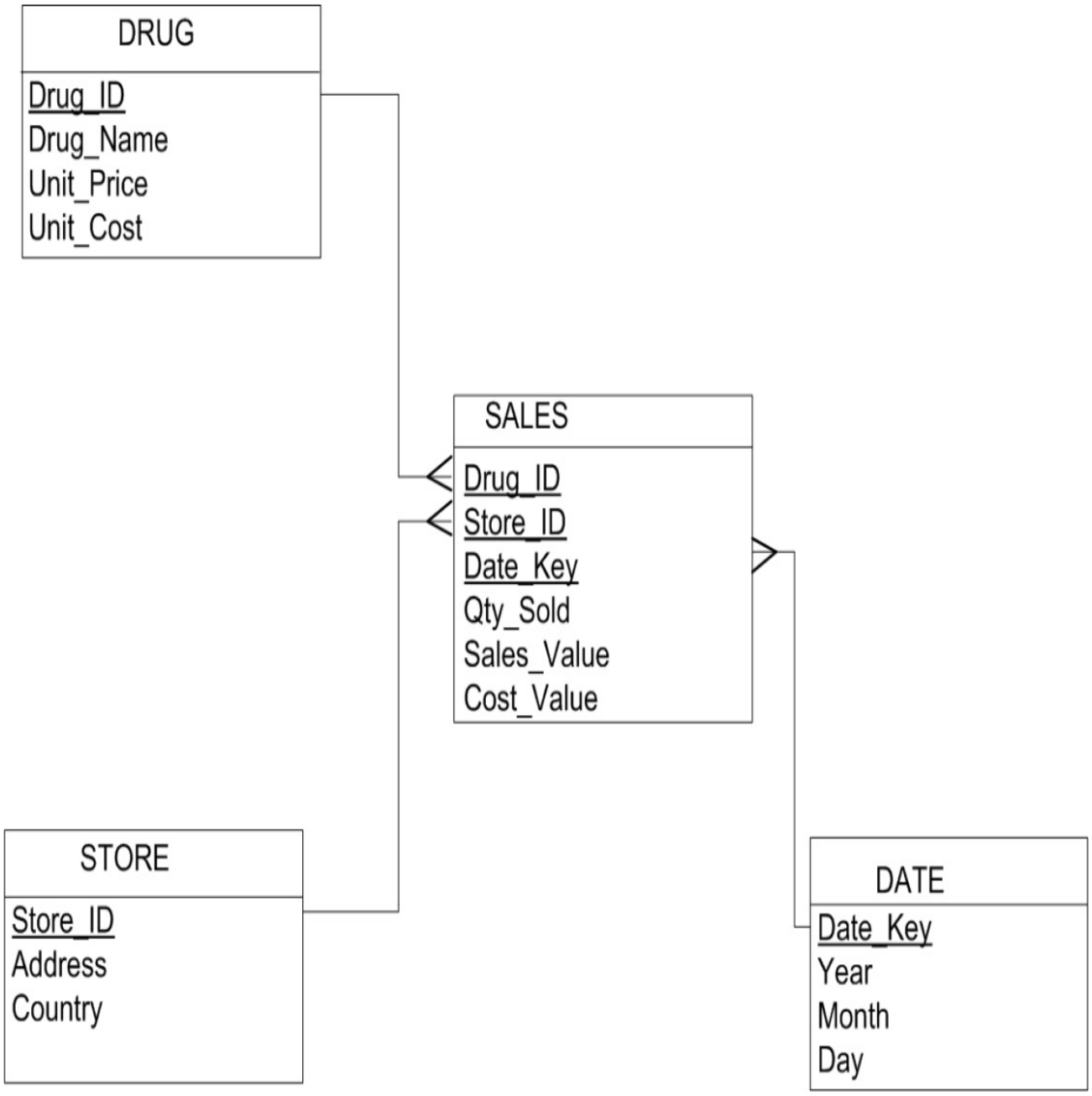
   DRUG(Drug_ID, Drug_Name, Package_Dosage, Unit_Price, Unit_Cost)
   SALES(Drug_ID, Store_ID, Sales_Date, Qty_Sold, Prescription_Details)
   STORE(Store_ID, Address, Comuntry)

   Do the following:

   a) Design and draw a schema to represent the data warehouse accurately for the company.
   b) Estimate the number of rows in the fact table in part 5(a).

Answer

(a) Star Schema for the data warehouse



(b) Star Schema for the data warehouse

The estimated no of rows = 300*50*365*10
= 54,750,000 rows