Exercises Data Warehousing
Dimensional Modelling

Solutions

To draw the schema, the tool "Indyco Builder" can be used. The link to the tool and personal licenses for all students are available from the course website. The tool is also installed in the computer lab.

1. **Record Label.** A record label wants to keep track of all contracts they have with bands, records they are producing and the sales of these records. Currently they are only keeping data of their ongoing contracts; no historical information is kept. Therefore it is decided to construct a data warehouse for collecting and storing historical information. With the data warehouse the company wants to analyze its sales. Based on conversations with the managers of the company, you were able to compile the following description of the available data.

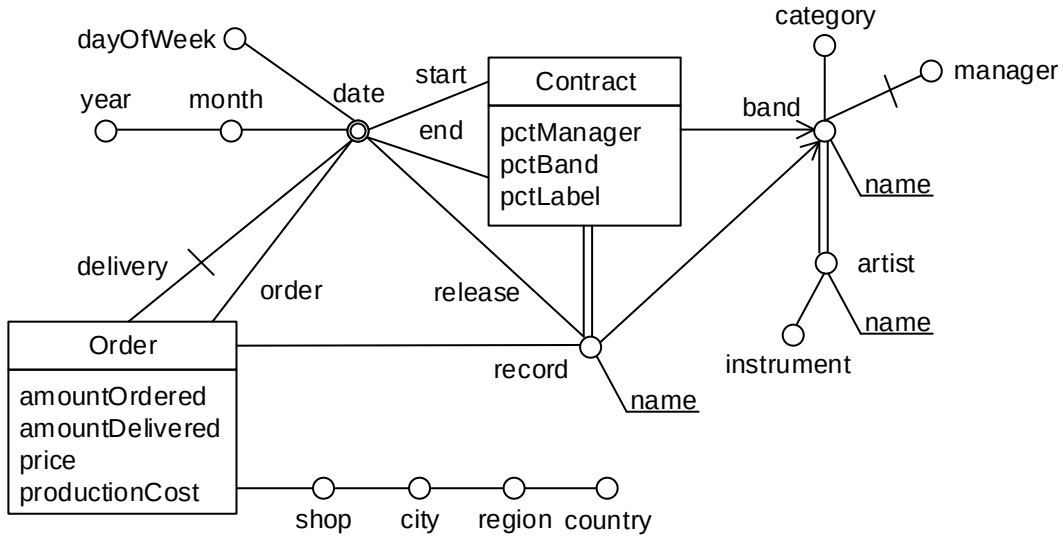
   The record label has several bands under contract. Each band has a name, a main category of music it plays, and one or more artists. A band may have a manager, but does not have to. Bands without a manager must have one of the artists as their main representative. An artist has one main instrument, which can also be his/her voice. Over time, bands can split and attract or replace group members. Bands make records which are physically produced and distributed by the record label. Records, including those that are not yet finished, have a title and are identified by a special international code. For every record the price to produce it and its release date are stored. The record label has contracts with the bands. For every contract start and end data are registered as well as for which records it holds, what percentage of sales goes to the manager of the group (if present), how much to the group, and how much to the record label. A contract can apply to multiple records, but every record falls under exactly one contract. Shops can order records at the record label. The date, amount, and agreed price of these orders are recorded, as well as the number of records that were sent to the shop in the end (sometimes demand outweighs supply) and the date on which the order was fulfilled. For every shop, the record label has information about the spatial location, including: city, country, and region of the shop.

   Some examples of the types of analyses the record label wants to perform based on the data warehouse are the following:

   - Which record/group/artist has given the record label the highest/lowest gain/return on investment per week/month/year.
   - What seller/city/country has the highest number of demands that could not be met?
   - Which record/group/artist has the highest number of demands?

   Make a dimensional fact model for a data warehouse according to the above description.

Solution



Common mistakes: dimensional attributes that are incorrectly placed into a hierarchy (for instance, roll-up from band to record); many-to-many relations modeled as one-to-many; models with only one fact usually did not work: contracts are not a property of sales, not vice versa. These are two different subjects in the data warehouse, hence requiring two facts. A third fact, for record, or the production of a record could be considered.

2. **Loan Data Warehouse.** A bank wants to build a data warehouse for storing and analyzing data about all loans issued by them.

   Every loan has one or more borrowers, a starting date, a type (e.g., fixed rate or one of different types of variable rate), the branch of the bank where the loan was issued, the interest rate at the start of the loan, and the amount. For every loan the purpose of the loan is recorded; e.g., to buy a car, a house, a personal loan, ... When a borrower applies for the loan, different discounts on the interest rate may be awarded; e.g., fidelity discount, discount because the borrower also bought some additional insurances, VIP discount, etc. For one loan, multiple discounts may apply. The amount of discount is independent of the branch. Every discount that has been awarded needs to be stored. When the loan ends, this is stored as well, together with an indication if the loan was fully repaid or the borrower defaulted.

   For the borrowers, their date of birth, family status, monthly income, number of children and address is stored.

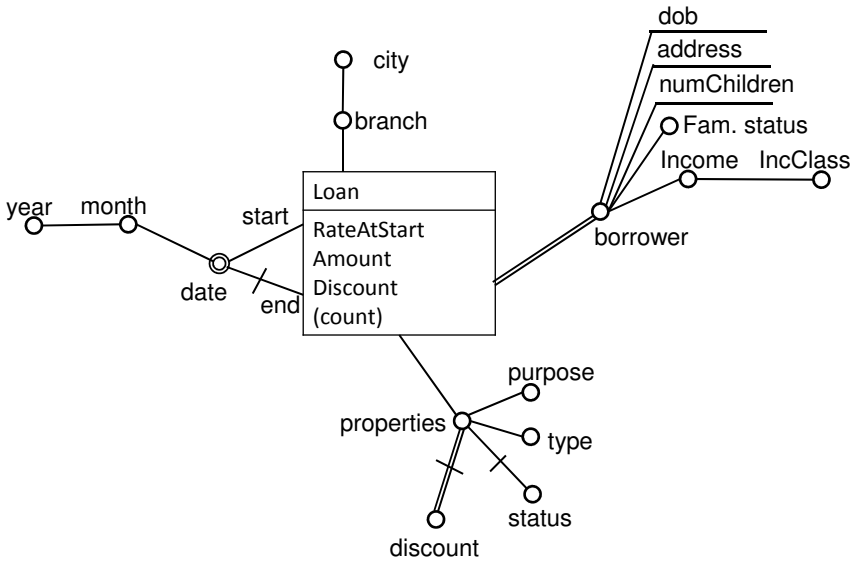
   The following questions are prototypical for the type of query analysts want to answer based on the data warehouse:

   - Give the average interest rate before discount at the start of the loan, per loan type and branch.
   - For all branches, give the minimum, maximum and average interest rate per loan type and purpose.
   - Give the number of loans per branch and per amount category. The amount category depends on predefined thresholds; amounts are divided into the following classes: `very high`, `high`, `medium`, `low`, and `very low`.
   - Give the percentage of defaulted loans per year and per city of the branch where the loan was issued.

Make a dimensional fact model for a data warehouse according to the above description.

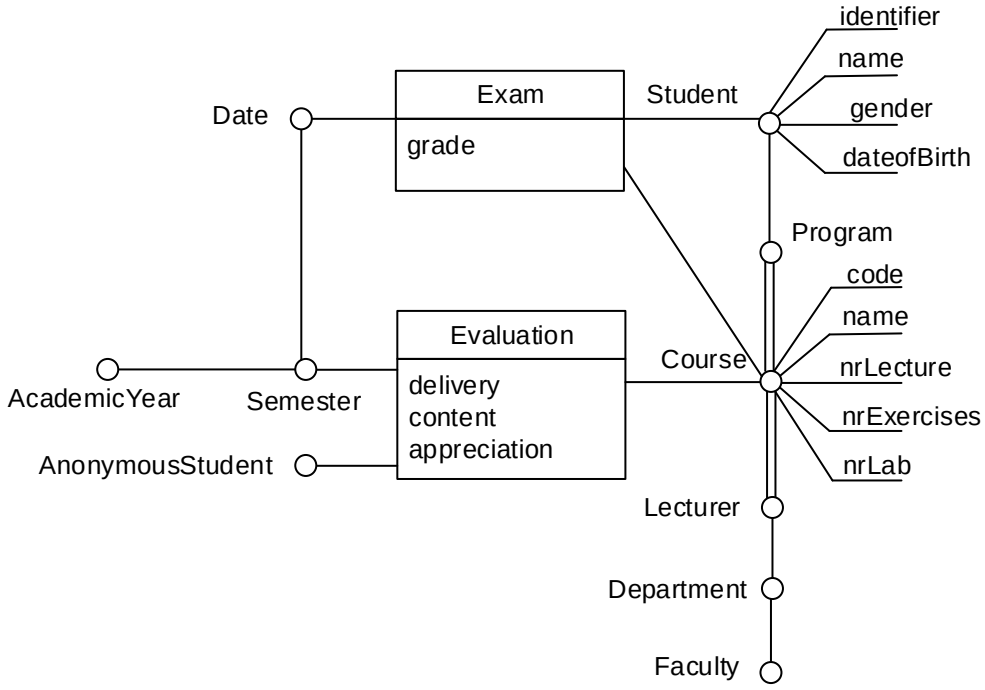
Solution *There is not a single correct answer; different choices may lead to different models. The level of detail given in this explanation was not necessarily expected in your answer.*

A first important observation is what will be the subjects around which we will be building our data warehouse. In this case the most natural choice is "loan." Another option could have been "loan event"; e.g., change in interest rate. This other option, however, has a couple of disadvantages: given the prototypical queries, loan is a more natural choice. Also, it is hard to define meaningful ways to aggregate measures such as amount over these events.

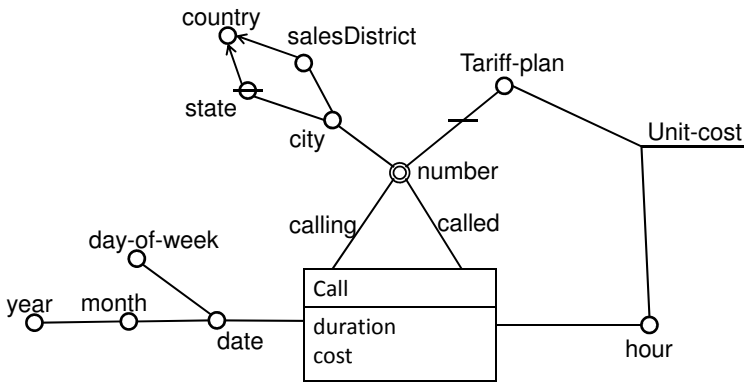


3. **Course Evaluation** A university wants to analyze the results of its students and the quality of its courses over time. For this purpose they want to create a data warehouse that will store the results of all students. A student is registered for a program which consists of several courses taught by one or more lecturers. Every course has a course code, a name, the number of lecturing hours and the number of hours for exercises and lab sessions. A lecturer belongs to a department and a department to a faculty. A student can take one or more exams for a course. In case of multiple exam trials, all results should be stored. For a student, his or her name, student identifier, gender, and date of birth are stored. Furthermore, every course offering is evaluated (anonymously) by the students. This evaluation results in three scores: one for course delivery, one for course content and one for overall appreciation. Based on the data in the data warehouse it should be possible to evaluate the exam results by student, by program, by academic year, by course, by department, etc. Furthermore courses will be evaluated on the basis of their evaluations by year, by lecturer, by program, etc.

Solution:



4. Consider the following DFM schema.



How big would the roll-up lattice be for this schema? Draw the roll-up lattice for the dimensions calling and date only.

Solution:
Size of the roll-up lattice:

- calling and called: both 13 grouping sets

  (ALL, country, state, sales-dist, state&sales-dist, city, number, tariff-plan, country&tariff-plan, state&tariff-plan, sales-district&tariff-plan, state&sales-dist&tariff-plan, city&tariff-plan)

- hour: 2 (ALL and hour)
- date: 7

  (All, Year, Day-of-week, Year&Day-of-week, Month, Month&Day-of-week, Date)

Total: $13 \times 13 \times 2 \times 7 = 2366$ ways to roll-up the data