09/08/2017

Julia Kvetnaya

Foundations of Data Curation

UIUC

# Assignment 1: Relational schema design exercise

Overview

The setting is an auto dealer. In this company, there are 3 departments, including:

Inventory

Sales

Customer relations

Right now, each department manages their information differently.

These departments would like to integrate their data into a shared database, in order to be able to answer questions like, “What engine is in Customer Smith’s car?”. It’s a challenge to answer right now because the Sales department has information about which car Customer Smith bought; the Inventory department tracks which engines are in which cars; and the Customer Relations has totally separate, slightly redundant information about Customer Smith.

You have been asked to design and populate a database that is effective and efficient for all departments. You are given data from each department:

File A (Inventory): Document relating models to styles, power trains, quantities, and individual cars, etc.

File B (Sales): Document relating people to cars, prices, trade-ins, etc.

File C (Customer relations): Document relating people to personal information, information about services, warranties, etc.

## Solution

Description of each file and information shared between files.

File A (Inventory): Document relating models to styles, power trains, quantities, and individual cars, etc.

This file is in txt format. It uses Tab as the delimiter between individual fields. Words within fields are separated by a single space. There are some inconsistencies that are described below.

Fields:

1) Index - order index of the cars in the inventory. Numerical.

2) VIN - alpha-numeric string containing numbers and upper case letters.

3) Year - 4 digit year.

4) Make - one word string, starting with capital letter. There are 3 examples in the file, “Ford”, “Tesla”, and “Toyota”.

5) Model - could be a couple of words, each word starting with capital letter, usually separated by a single space. However, the first make in the file contains a Tab between the first and the second word, and space between the second and third words. There are four examples in the file, “Expedition King Ranch”, “Explorer Eddie Bauer”, “Model S” and “Prius”.

6) Drive - following letter combinations are used: “4WD”, “AWD”, “FWD”.

7) Color - a capitalized string. There are 6 examples in the file, “White (Pearl)”, “White”, “Gray”, “Blue”, “Silver”, “Black”.

8) Doors: number of doors as a digit followed by a single space and the word “door”. The word “door” is sometimes capitalized and sometimes not. It can also be followed by a space separated capitalized type of body. There are two examples in the file, “Sedan” and “Hatchback”

9) Engine: string field, where each word starts with capital letter. There are 2 examples in the file, “Internal Combustion”, “Electric”.

10) Price: string value in quotes, starts with a dollar sign and ends with a single space after the last digit.

File B (Sales): Document relating people to cars, prices, trade-ins, etc.

This file is in cvs format. In this file the first line contains fields names, each record starts on the new line, and all fields are separated by commas.

Fields:

1) ID - numerical, incremental number.

2) LastName - capitalized string.

3) FirstName - capitalized string.

4) MI - single capital letter. Optional.

5) Address - number address and street name.

6) City - capitalized string.

7) State - standard 2 capital letter string code. Missing for some records.

8) Country - string. In this example all customers are located in the USA. Missing for some records.

9) SaleDate - date in date format.

10) Model - a combination of Make and Model. Each word is capitalized, multiple words are separated by a single space.

11) Year - four digit year.

12) Color - capitalized string. There are 6 examples in the file, “White (Pearl)”, “White”, “Gray”, “Blue”, “Silver”, “Black”.

13) Engine - string value. Each word is capitalized, multiple words are separated by a space. Two values are present: “ Combustion Engine” and “Electric”.

14) VIN - alpha-numeric string.

15) MSRP - string starting with a dollar sign containing numbers, commas (separating thousands) and decimal point.

16) Discount - string value. Empty if no discount was applied. The following values are present: “EndofYear”, “First Time Driver”, “Repeat Customer”, “Senior Citizen”.

17) TradeIn - string value. For the customers that had trade ins contains word “Yes”. It is empty for all other records.

18) TradeInValue - price format. Exists for the customers that had trade ins. Matches for the customers for whom “TradeIn” field is Yes. Empty for other records.

19) PurchasePrice - price format. Empty for Draco Malfoy record which seems to be an error.

20) RepeatCustomer - string. It is “Yes” for repeat customers and is empty for all other records.

File C (Customer relations): Document relating people to personal information, information about services, warranties, etc.

This file is in doc format. In this file individual fields are separated by Tabs, words within each field are separated by a single space, and each new record starts on a new line.

Fields:

1) Last name - capitalized string, followed by Tab.

2) First name - capitalized string, followed by Tab.

3) Middle initial - one capital letter, missing in some cases and seems to be optional, followed by newline.

4) Street number - numerical, followed by a single space.

5) Street name - each word capitalized, multiple words separated by a single space, followed by newline.

6) Occupation - capitalized string. Multiple words separated by a single space, followed by newline.

7) Note - capitalized string, usually refers to financing options, missing for roughly half of the entries. Followed by 2 newlines to indicate end of a single customer record.

The information that is shared between the files:

Files A and B share the following information:

1)Make/Model

2) MSRP

3) Year

4) Color

5) VIN

6) Engine

Files B and C share the following information:

1)Last Name

2) First Name

3) MI

4) Street Address

5) City

6) State

7) Coutnry

Files A and C don’t share any information.

**Account of my process for creating the database scheme and tables**.

1) How did you decide to represent the data in the way that you did?

In order to take out redundancies, to normalize data and to make it independent I decided to organize data around 4 entities, namely, Customer, Car, Sales and TradeIns.

CustomerTable includes information from File C (Customer): last name, first name, middle initial, address, state, zip code, occupation and a note. I also added RepeatCustomer field that used to belong to File B (Sales) because it seems to be an attribute of the client. I made this field to have predefined values of Y and N.

CarTable is based on the information from File A (Inventory). I added one new field Body since it seems to be a separate characteristic. Also, the fields Make, Model, Drive, Color and Engine need to be standardized. Each field should come with a predefined set of values to maintain consistency. Also, I changed the meaning of the Inventory. CarTable contains information about all cars that were ever sold or traded in to this dealership. At this point this table doesn’t contain information about current status of the car. Just looking at this table we can’t say whether the car is currently in the dealership or it has been sold. This information can be derived by combining information from SalesTable and TradinTable.

TradeinTable contains trade in information from File B (Sales). The reason I decided to put it into a separate table is because this dealership seem to be getting inventory in 2 ways, from manufacturers and by trade ins. Trade in is sort of an opposite of a sale, and it seems to be its own entity. In order to populate this table with example records, I created fictitious cars that were traded in by customers during sales.

SalesTable received its information from File B (Sales). However, since Customer, Car and Tradein data have already been stored in there designated tables, SalesTable only uses foreign keys from those tables, and uses Date and Price info that are pertinent to a particular sale. Also, I left the DiscountType field in this table. However, the data for this field need to be standardized and limited to a particular set. The current data should be limited to “EndOfYear”, “FirstTimeDriver”, “RepeatCustomer”, and “SeniorCitizen”.

2) Did you leave out any information? If so, why?

I took out the following information:

1) Took out TradeIn field from SalesTable. This used to be a Yes/Empty field. Now the records that had TradeIns have a foreign key from TradeinTable, and the ones that had no trade ins have this field empty.

2) Took out RepeatCustomer field from SalesTable. RepeatCustomer became an attribute in CustomerTable since it seems to be pertinent to Customer.

3) Sales file used to contain record pertaining Draco Malfoy. However, this record did not have a purchase price. In my database schema such inconsistency is impossible since each sale must have purchase price. Thus for the sake of the example I could either make up a price or just take out the record. I chose to take out the record.

4) I took out the word “doors” from the field describing the number of doors in the vehicle. It seems that it is an unnecessary information since already contained in the name of this field. I only the left the integer number.

3) Why did you choose certain things as attributes? As keys?

Each of the 4 tables that I created contains a primary key which is an incremental index of records in this table. I had a dilemma about CarTable. VIN is a unique number and could be chosen as a primary key. However, I decided to use the index rather than VIN for the sake of consistency and, perhaps, unforeseen future changes.

CustomerTable and CarTable are independent from all other tables and don’t use foreign keys. In addition to a primary key, TradeinTable uses CustomerID and CarID as foreign keys. It records which customer traded in which car. Thus, for trade in to occur, CustomerTable and CarTable need to be populated with pertinent data.

SalesTable uses CustomerID, CarID and TradeinID as foreign keys. It records which customer purchased what car and which car was traded in during this deal. Thus, in order for a sale to take place, CustomerTable, CarTable and TradeinTable need to be populated with pertinent data.

4) What were the hardest decisions you had to make in this design process?

There were 3 decisions that I struggled with.

1) I chose to create TradeinTable. The reason it was a difficult decision is because I made an assumption that this dealership sells cars that were traded in. From the data that we received there was no evidence of this. In the original data all cars that were sold by the dealership were new and supplied by manufacturers.

2) The second decision that I struggled with was that currently there is no plan of when the data are removed from any tables. Every car that ever passes through this dealership will remain the database forever and ever. What if the customer purchases the car, then trades in back in, then buys it again in a couple of years? I suppose, that this information will be maintained in a collection of Sales and Tradein records, but the solution seems a little clunky.

3) The third decision that I struggled with is that I didn’t leave any easy way to know if the car is currently in the dealership. I didn’t do it to preserve maximum normalization and data independence. This information can be obtained by combining data from SalesTable and TradeinTable. However, again, it is clunky.

5) How does your schema design support data independence?

In this schema I believe that I achieved maximum normalization of the data. I completely removed all redundancies. Through the use of primary and foreign keys and multiple constraints I believe that I achieve data independence.

6) How may your schema design support the overarching goals of data curation (revisit objectives and activities of Week 1)?

Collection:

In my schema, I manage proper collection of data by employing the following techniques:

1) Defining proper types of input values.

2) Standardizing input values (State, Country, Repeat, Make, Model, Drive, Color, Body, Engine, Discount Type).

3) Putting constraints on which values could be empty and which cannot.

4) Putting constraints on functional dependencies.

Organization:

In my schema, I manage organization in the following way:

1) I defined a proper schema for the data.

2) By organizing data around Customer, Car, Trade-ins and Sales I use abstraction indirection to manage data.

3) Through defining constraints on inputs and functional dependencies I identified and used relevant standards for both syntax and semantics.

4) I specified and documented proper document schema attributes.

Storage:

In this exercise I don’t employ Storage activity.

Preservation:

In my schema I employ Preservation activity in the following way:

I defined the meaning and described the type of each field in the database.

Discoverability:

In this exercise I don’t employ Discoverability.

Access:

In this exercise I don’t employ Access.

Workflow:

In this exercise I don’t employ Workflow.

Identification:

In this exercise I support Identification by carefully describing all fields, their meaning, possible values and use.

Integration:

In this exercise I integrated data from 3 different departments using 3 different data models. First I carefully documented each original data format. Then I found similarities. Afterwards, I made an appropriate data model to accommodate data from all sources, and also address the new requirements. Lastly, I documented the new schema, its relation to the old data models, and all changes that I made along the way.

Reformatting:

Since I was combining data from 3 data sources, normalizing and standardizing the values, I reformatted some some of the values and documented all changes.

Reproducibility:

In this exercise I employed Reproducibility by documenting processing and analysis.

Sharing:

In this exercise I didn’t employ Sharing activity.

Communication:

In this exercise I didn’t employ Communication activity.

Provenance:

In this exercise I employed Provenance but carefully documenting original formats and data models, the new data formats and all changes that happened while transforming data from old data models to new data model.

Modification:

In this exercise I didn’t employ Modification activity.

Compliance:

In this exercise I didn’t employ Compliance activity.

Security:

In this exercise I didn’t employ Security activity.

7) Which curation activities could enhance or sustain the database for future discovery and use for new purposes? What additional activities would you recommend?

1) Preservation: Maintain and document preservation strategy. Document semantics and syntax documentation. Generate and preserve all metadata needed to ensure that the data is usable and understandable, and can be authenticated and audited for provenance.

2) Discoverability: Develop metadata to support search for and finding relevant data in relevant formats.

3) Access: Maintain systems, tools, and metadata that support efficient and reliable retrieval and distribution of data.

4) Workflow: Well-maintained scripts should be developed and used to document as well as execute data transformations.

5) Communication: Support representation, publishing, and visualizations that provide insight.

6) Modification: Support management of corrections and updates.