MSiA-413 Introduction to Databases and Information Retrieval

Homework 2: Data modeling: Data Sets, Normalization, and ER Diagrams

Ayush Agarwal

Name 1:

scg1143

NetID 1:

Lowan (Sydney) Li

Name 2:

chr0390

NetID 2:

Instructions

You should submit this homework assignment via Canvas. Acceptable formats are word files, text files, and pdf files. Paper submissions are not allowed and they will receive an automatic zero.

As explained during lecture and in the syllabus, assignments are done in groups. The groups have been created and assigned. Each group needs to submit only one assignment (i.e., there is no need for both partners to submit individually the same homework assignment).

Each group can submit solutions multiple times (for example, you may discover an error in your earlier submission and choose to submit a new solution set). We will grade only the last submission and ignore earlier ones.

Make sure you submit your solutions before the deadline. The policies governing academic integrity, tardiness and penalties are detailed in the syllabus.

Question 1. Dataset Exploration (6 points)

Download the data sets using the links below and import them to either Excel, Numbers, or another spreadsheet processing program of your choice. Below you can find a tutorial on how to download a data set and import it to Excel or Numbers. Then, proceed to answer the following questions:

- **a. (3 points)** Did you encounter any problems in importing any of these datasets into a spreadsheet? If yes, describe which dataset(s) you encountered the problem with, and explain the reasons you believe it failed to be imported.
- **b.** For the dataset(s) that were successfully imported, please answer the following questions:
  - **i. (1 point)** What is the data set's name? You have to be really precise with the name; after all, there may be multiple datasets at Kaggle.com with similar names. Find the name that uniquely identifies it.
  - **ii. (1 points)** How many rows does it have?
  - **iii. (1 points)** How many columns does it have?

Link to Project in Kaggle	**CSV file to download**
https://www.kaggle.com/benhamner/sf-bay-area-bike-share	status.csv
https://www.kaggle.com/abcsds/pokemon	pokemon.csv

How to download a data set from Kaggle

The links above point to two data predictive modelling and analytics projects on Kaggle. You can read the project and data description on the corresponding project overview page. If you click the project data page, you will see a list of comma-separated values (CSV) files and other file types on the left hand side. When you click on any file, you can preview the first 100 columns, and read the column metadata or column metrics. There is a download button to download that CSV file to your computer. Once downloading is done, you can follow the guidelines below to import the dataset into the spreadsheet program of your choice.

How to import a CSV file into Excel

Go to the "File" main menu and select the submenu "import" to import a csv file in Excel. Choose the appropriate file type to import the file (CSV), and then click the import button. Next, select which file to import, and click "Get Data". Most of the CSV files are delimited by commas to separate each column, and you can import it starting at row *n* if that file is too large. After that, there are a few options to opt for a proper column data format, and a spreadsheet which you want to put the data. It is recommended import the new data into a new sheet. When you are done with all file configurations, click "Finish". Excel will try to import the CSV file, or throw an exception if any error happens.

How to open a CSV file in Number

For Numbers, you can choose which CSV file to open, and the program will do it for you. That’ s it.

Solution :

I encountered issues while loading the status.csv file into spreadsheet software in MacOS. One possible reason is the size of the status.csv file which is ~1.9GB and this will be too computational for the software too load. Also, the file contains ~71million rows which is way larger than the capacity of the spreadsheet which is ~1 million. However, I still imported the data set, but one of the columns (‘time’) does not show any information, only showing ####.

For the Pokemon.csv file, I successfully imported the data.
The name of the data set is Pokemon. There are 13 columns and 800 rows without including the header row.

Question 2. Data Type Exploration (6 points)

Assume the datasets provided below. Consider the following data formats:

- **a. (1 point)** 32-bit integer
- **b. (1 point)** 64-bit integer
- **c. (1 point)** fixed point (and specify the number of decimal places)
- **d. (1 point)** floating point (either single or double precision)
- **e. (1 point)** date and time in epoch seconds
- **f. (1 point)** date and time in epoch microseconds

For each one of the data formats above, please answer the following:

- i. Is there a column in one of these datasets that would be **best** stored in that format? (yes/no)
- ii. If yes, please provide
  1. the data set
  2. the table name
  3. the column name
  4. a one-sentence description of the column
  5. an example of the data in the column
  6. the reason why your chosen data type is appropriate
- iii. If no, explain why not (1-2 sentences)

Data sets:

1. https://www.kaggle.com/benhamner/sf-bay-area-bike-share
2. https://www.kaggle.com/datasf/san-francisco

Solution

a) 32-bit integer

- a. Yes
  - ○ San Francisco Open Data
  - ○ bikeshare_stations
  - ○ dockcount
  - ○ The Number of total docks at each station
  - ○ 11
  - ○ All the numbers of total docks here are integers, which match with the data type. All of them fall within the range of values that a 32-bit signed integer can represent. In a 32-bit integer system, the range typically extends from -2,147,483,648 to 2,147,483,647.
- b. N/A

b) 64-bit integer

- **a.** No
- **b.** N/A
- **c.** Since all integers in the datasets are below the 32-bit limit, opting for a 64-bit representation would result in unnecessary resource consumption.

c) fixed point

- **a.** Yes

- **i.** SF Bay Area Bike Share
- **ii.** weather.csv
- **iii.** max_temperature
- **iv.** The maximum temperature of each day
- **v.** 73.0
- **vi.** Max Temperature should be a good fit for a fixed-point representation because it represents a real-world quantity that often doesn't require high precision beyond a certain number of decimal places. Normally, the precision needed is limited to one decimal place, 73.0, for example.

b. N/A

d) floating point

a. Yes

- **i.** San Francisco Open Data
- **ii.** bikeshare_stations
- **iii.** Longitude
- **iv.** The longitude of each bikeshare stations
- **v.** -121.894902
- **vi.** The longitude data type is typically well-suited for single-precision floating-point representation. Single-precision can represent values with a precision of roughly 7 decimal digits, which is usually sufficient for longitude coordinates, as they typically have up to 6 decimal places of precision.

b. N/A

e) date and time in epoch seconds

a. Yes

- **i.** San Francisco Open Data
- **ii.** bikeshare_trips
- **iii.** start_date
- **iv.** The start date of trip with date and time, in PST
- **v.** 06/11/2016 08:19:00
- **vi.** The data type matches, as this column is a datetime variable. We only need to record down to the seconds in the format of MM/DD/YYYY HH:MM:SS. Nothing further we need to store.

b. N/A

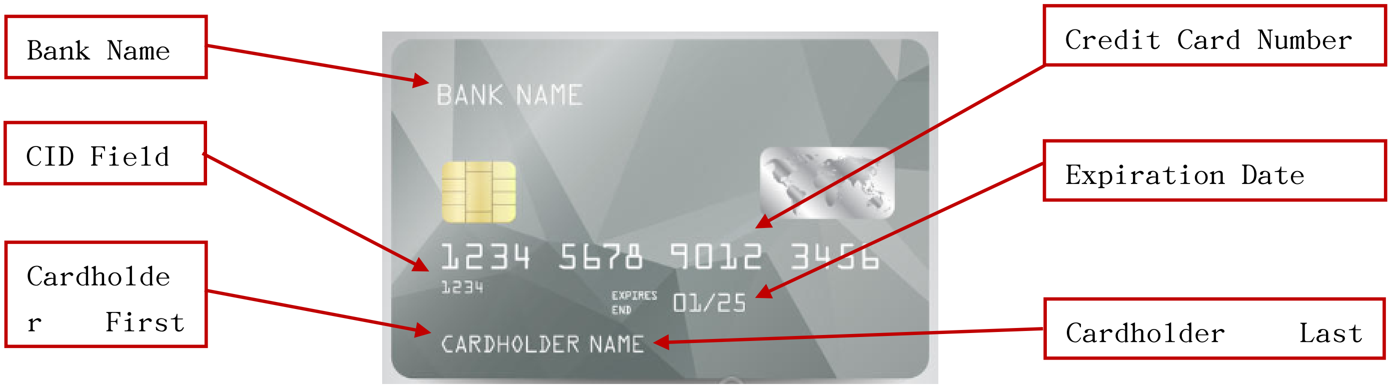
f) date and time in epoch microseconds

a. No

b. N/A

c. All the columns listed can be represented in epoch seconds, and none of them need a level of accuracy finer than one second. Microseconds are not needed in this situation

Question 3. Data Types (8 points)



You are building a database for a credit card company. You need to select the best data types for the various parts of a credit card shown above. The database software you use supports the following types:

- *32-bit signed integer*: can store all integers between -2,147,483,648 and 2,147,483,647.
- *32-bit floating point*: can store numbers in the scientific notation with about 7 decimal digits of precision and exponent between 10^{-38} and 10^{+38}. It can precisely store all integers ≤ 16,777,215.
- *Epoch seconds*: a 32-bit unsigned integer that represents data and time by the number of seconds since midnight on Jan 1, 1970 in London.
- *UTF-8*: text, of any length.

You must use one of these four data types to store each of the values below. Each value should be stored in a single data element. Which of these types is the best to store:

a. **(1 point)** The bank name?

Answer:

UTF-8 (text, of any length)

Explanation: Bank names can vary in length, and using UTF-8 allows for storing text of any length.

b. **(1 point)** The CID field (4-digit decimal number)?

Answer:

32-bit signed integer

Explanation: The CID field is a 4-digit decimal number, and a 32-bit signed integer can accurately store this range of values

c. **(1 point)** The cardholder last name?

Answer:

UTF-8 (text, of any length)

Explanation: Last names can vary in length, and using UTF-8 allows for storing text of any length.

d. **(1 point)** The expiration date?

Answer:

Epoch seconds (32-bit unsigned integer)

Explanation: Epoch seconds is suitable for representing date and time, and it can be used to store the expiration date as seconds since a specific date.

e. **(2 points)** The credit card number (16-digit decimal number)?

Answer:

UTF-8

Explanation: It will be UTF-8, since 16 digit decimal number will not be accomated in 32 bit signed integer.

f. **(2 points)** The balance of the credit card?

NOTE: the balance is in USD, it is guaranteed to stay between -1,000,000.00 and +1,000,000.00 (negative values indicate credit), and must be accurate down to one cent (i.e., 1/100th of a dollar).

Answer:

32-bit floating point

Question 4. Database Normalization (10 points)

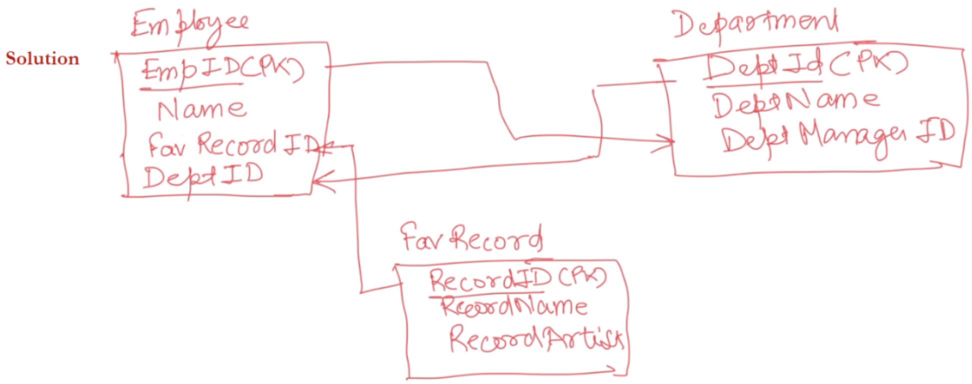
You work as a data analyst at a fishing/outdoors company. To identify new talents in database design, the company hosts an annual database schema competition. The winner takes home a commemorative statue known as the *Data Bass*. You won the competition last year, so a friend asked you to review his submission. Unfortunately, your friend did not take MSiA-413 and put all his data in a single table, shown below:

Employees_Database					
Empl. ID	*Name*	*Favorite Record*	*Department*	*Dept. Manager ID*	*Fav. Record’s Artist*
1	Nancy	Abbey Road	Sales	1	The Beatles

2	John	Porgy and Bess	Accounting	2	Gershwin
3	Bill	Kind of Blue	Operations	6	Miles Davis
4	Tracy	A Night At The Opera	Sales	1	Queen
5	Muji	La Revancha del Tango	Sales	1	Gotan Project
6	Ohana	Ka 'Ano'i	Operations	6	Kamakawiwo'ole
7	Jill	Porgy and Bess	Accounting	2	Gershwin
8	Gloria	La Revancha del Tango	Operations	6	Gotan Project
9	Frank	Abbey Road	Accounting	2	The Beatles

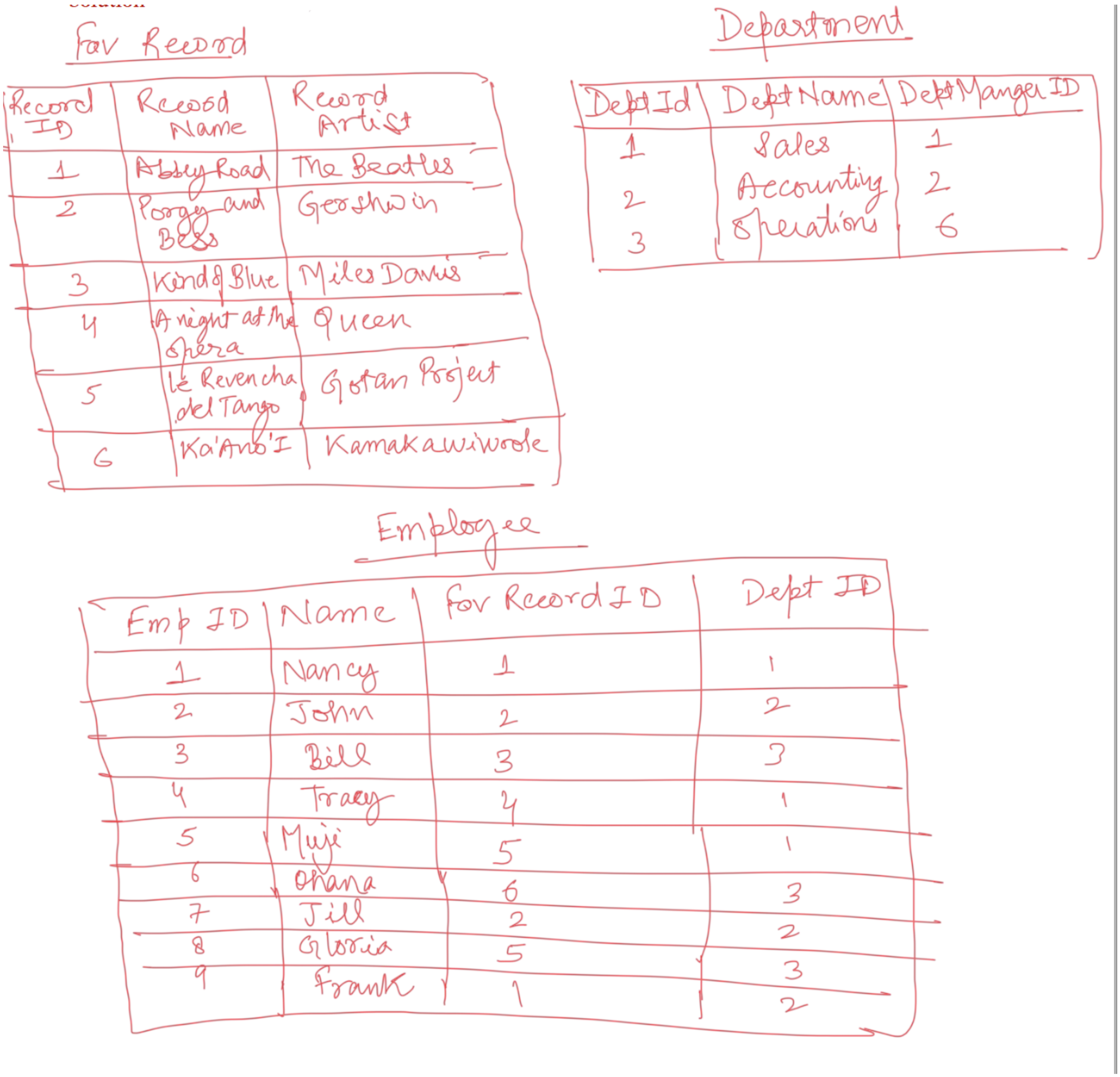
a. **(6 points)** Help him by normalizing the database to remove redundancy. Show the normalized database **schema**.

Solution:



b. (4 points) Show the current **instance** of the database in the normalized schema.

Solution



Fav Record

Record ID	Record Name	Record Artist
1	Abbey Road	The Beatles
2	Porgy and Bess	Gershwin
3	Kind of Blue	Miles Davis
4	A night at the Opera	Queen
5	Lé Revencha del Tango	Gotan Project
6	Ka'Ano'I	Kamakawiwoole

Department

Dept Id	Dept Name	Dept Manger ID
1	Sales	1
2	Accounting	2
3	Operations	6

Employee

Emp ID	Name	Fav Record ID	Dept ID
1	Nancy	1	1
2	John	2	2
3	Bill	3	3
4	Tracy	4	1
5	Muji	5	1
6	Ohana	6	3
7	Jill	2	2
8	Gloria	5	3
9	Frank	1	2

Question 5. ER Diagram (20 points)

The main entities that participate in an online bookstore enterprise are as follows:

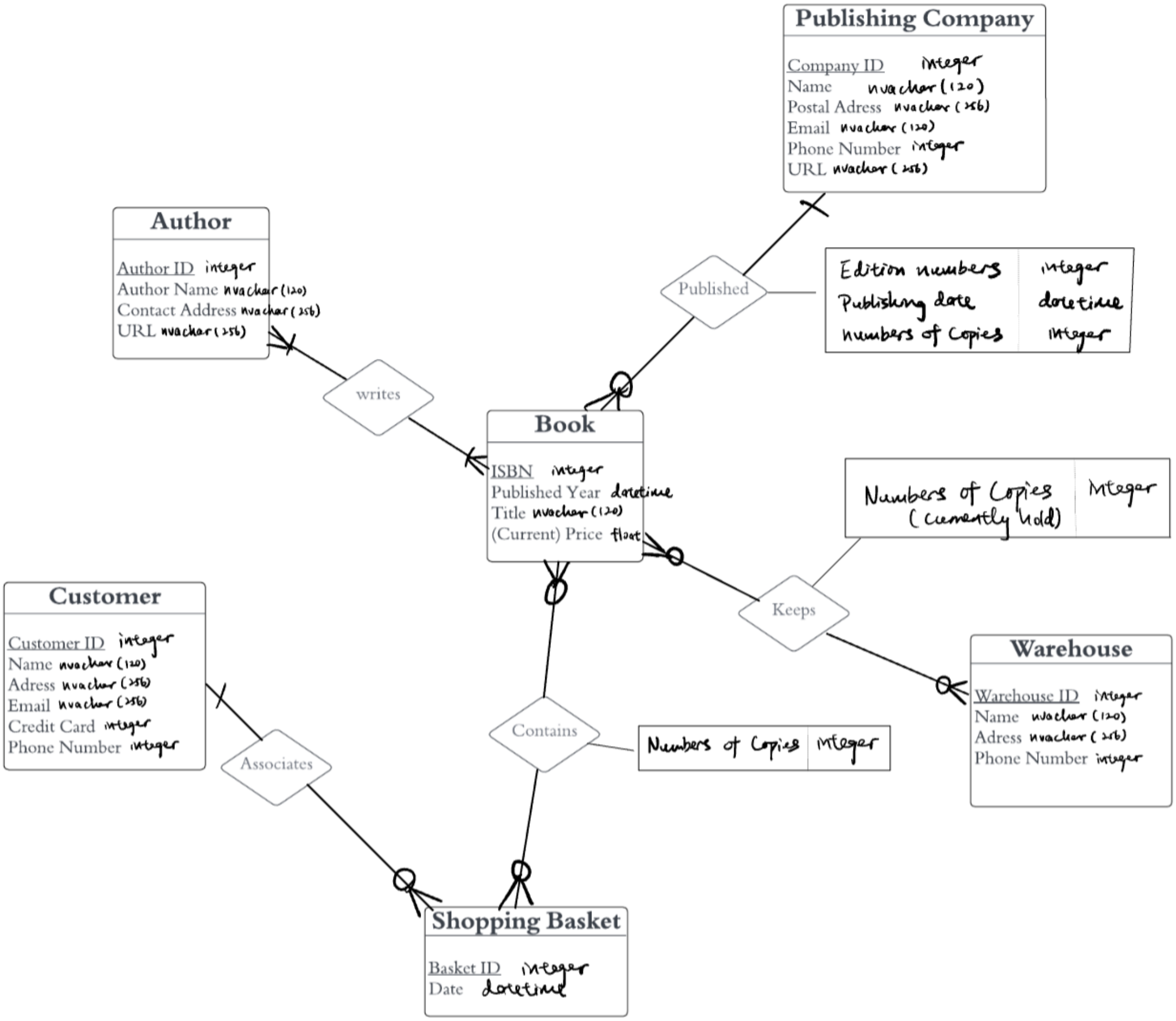
- A book has the information about the year that it was published, its title, the (current) price and its ISBN number. **Assume that ISBN is the unique number assigned to each edition/version of the book.**
- An author has information which includes his/her name and contact-address, along with a URL.
- Each publishing house/company has a name, postal address, phone number, email and URL.
- Each customer has a customer ID, name, address, email, credit card, and phone number, and **each customer must provide only one set of information.**
- A particular "shopping session" is typically recorded as a shopping basket, which is assigned a unique basket ID and has the information about the date of the given shopping session.
- Since this is an online bookstore, there must be physical locations where (copies of) the books are stored. A given warehouse has its address, name, and phone number available.

The associations among the various entities listed above are as follows:

- Each book is written by some author(s).
- Each book is published by a particular publishing house and information is kept about the publishing date, the edition number, and the number of copies.
- Each shopping basket is associated with a particular customer.
- Each shopping basket may contain several books and even several copies of a particular book.
- Each warehouse keeps/stocks different books, and for each book it also records the number of copies that it currently has.

Please draw an ER Diagram that models the online bookstore according to the information and rules provided above. If you make any assumptions along the way, please write them down. Note: there are a number of online tools to draw ER diagrams (e.g., https://www.lucidchart.com).

Solution


Publishing Company
Company ID integer
Name nvachar(120)
Postal Adress nvachar(256)
Email nvachar(120)
Phone Number integer
URL nvachar(256)
Published
Edition numbers integer
Publishing date datetime
numbers of Copies integer
Author
Author ID integer
Author Name nvachar(120)
Contact Address nvachar(256)
URL nvachar(256)
writes
Book
ISBN integer
Published Year datetime
Title nvachar(120)
(Current) Price float
Numbers of Copies (currently hold) Integer
Keeps
Warehouse
Warehouse ID integer
Name nvachar(120)
Adress nvachar(256)
Phone Number integer
Customer
Customer ID integer
Name nvachar(120)
Adress nvachar(256)
Email nvachar(256)
Credit Card integer
Phone Number integer
Associates
Contains
Numbers of Copies integer
Shopping Basket
Basket ID integer
Date datetime

Assumptions/Explanations:

- For books, I decided to use the ISBN as the primary key since its uniqueness for each book.
- Next, I created Company ID for each publishing company to be its primary key, since the company name can be repeated. Same reason for the Warehouse and Customer, I also created the Warehouse ID and the Customer ID as the primary keys.
- For Shopping Basket, I used the Basket ID as the primary key since each basket has its only number.
- Warehouse keeps Books. The association (keeps/stocks) has additional information about the numbers of copies a warehouse currently holds.
- Publishing Company publishes books. The association (published) has additional information about the edition numbers, publishing date, and numbers of copies.
- Shopping Basket contains books. The association (contains) has additional information about the numbers of copies.
- A warehouse can keep zero, one, or many books. A book can be kept in zero, one, or many warehouses.
- A shopping basket can contain zero, one, or many books. A book can be in zero, one, or many shopping baskets.
- Each customer can have many shopping baskets but one shopping basket is only associated with one customer.
- A book can be written by one or more authors. An author can write one or more books.
- Publishing companies can publish many books, but a book can be only published by one publishing company because of the legal and contractual aspects of the publishing industry.