MSiA-413 Introduction to Databases and Information Retrieval

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

Name 1: Alicia Burris

NetID 1: anb0847

Name 2: Naomi Kaduwela

NetID 2: nak133

Instructions

Due Date: Friday October 19, 11:59 pm

Question 1. Dataset Exploration (10 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. **(2.5 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.

Status: Yes - The file is too large to fully load in excel and exceeds the limit, thus only partial data from kaggle appears upon opening. This is why databases are better to use than excel and local storage.

Pokemon: No

b. For the dataset(s) that were successfully imported, please answer the following questions:

i. **(2.5 points)** What is the data set's name? Note: you have to be really precise with the name; after all, there may be multiple datasets at Kaggle.com with similar names.

Status: SF Bay Area Bike Share is the title on Kaggle > Anonymized bike trip data from August 2013 too August 2015

Pokemon: Pokemon with stats is the title on kaggle > 721 Pokemon with stats and types

ii. **(2.5 points)** How many rows does it have?

Status: 72.0M in kaggle, even though all do not download to be able to validate in excel

Pokemon: 800

iii. **(2.5 points)** How many columns does it have?

Status: 4

Pokemon: 13

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

Question 2. Data Types (12 points)

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

a. **(2 points)** 32-bit integer
b. **(2 points)** 64-bit integer
c. **(2 points)** fixed point (and specify the number of decimal places)
d. **(2 points)** floating point (either single or double precision)
e. **(2 points)** date and time in epoch seconds
f. **(2 points)** 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

a.) **32-bit integer**

i.) Yes

ii)

1.) **Data Set**: SF bay area bike share
2.) **Table**: status
3.) **Column**: station_id
4.) **Description**: This is the ID number associated with each bike station.
5.) **Example**: 2
6.) **Why data type**: It will always be a whole number and there is not a concern that the number of stations will exceed the range as there are only 84 thus far

b.) **64-bit integer**

i.) Yes

ii)

1.) **Data Set**: SF bay area bike share
2.) **Table**: sfpd_incidents
3.) **Column**: pdid
4.) **Description**: unique identifier for each incident during update and insert statements
5.) **Example**: 15037616304138
6.) **Why data type**: It will always be a whole number and the range of 32-bit int data type can hold integer values in the range of −2,147,483,648 to 2,147,483,647 so we know this could exceed the range, and thus 64 bit is ideal, as the range is 9,223,372,036,854,775,807,

c.) **Fixed point**

i.) Yes

ii)

1.) **Data Set**: SF bay area bike share

2.) **Table**: weather

3.) **Column**: mean_temperature

4.) **Description**: describes temperature for a specific day and zip code in the bay area

5.) **Example**: 70.0

6.) **Why data type**: Here we see the data values do not have many numbers after the decimal point and all numbers follow the same format, so we can save space by using fixed point as opposed to floating point here and only show temperature precision to that level granularity

d.) **Floating point -**

i.) Yes

ii)

1.) **Data Set**: SF bay area bike share

2.) **Table**: sfpd_incidents

3.) **Column**: latitude

4.) **Description**: describes latitudinal coordinate where the incident occured

5.) **Example**: 37.7775321935218

6.) **Why data type**: Latitude goes from 0-90, but there are many precision points after the decimal that are important to capture exactly as these are each specific different coordinate geography points. Thus, the number of digits before and after the decimal point vary. It is also good to use floating point here because we will have more precision as the values vary from 0-90 and the closer we are too 0, the more precision we have

e.) **Date and time in epoch seconds**

i.) Yes

ii)

1.) **Data Set**: SF Bay Area Bike Share

2.) **Table**: trip

3.) **Column**: start_date

4.) **Description**: Date/time associated with a given bike dock at a given station

5.) **Example**: `8/29/2013 14:13`

6.) **Why data type**: This date/time format conforms to the same level granularity (seconds) that we have in the data set given

f.) **Date and time in epoch microseconds**

i.) No

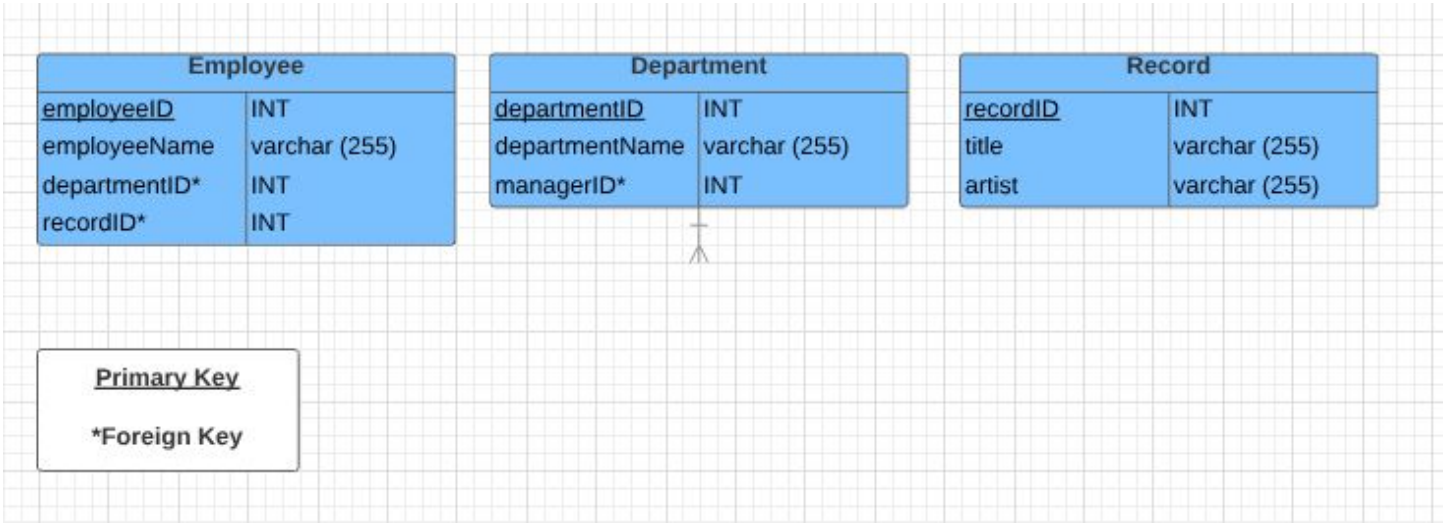
ii) We do not have the granularity (in microseconds) for this data, thus it would be a waste of space because we do not need that much precision.

Question 3. Database Normalization (10 points)

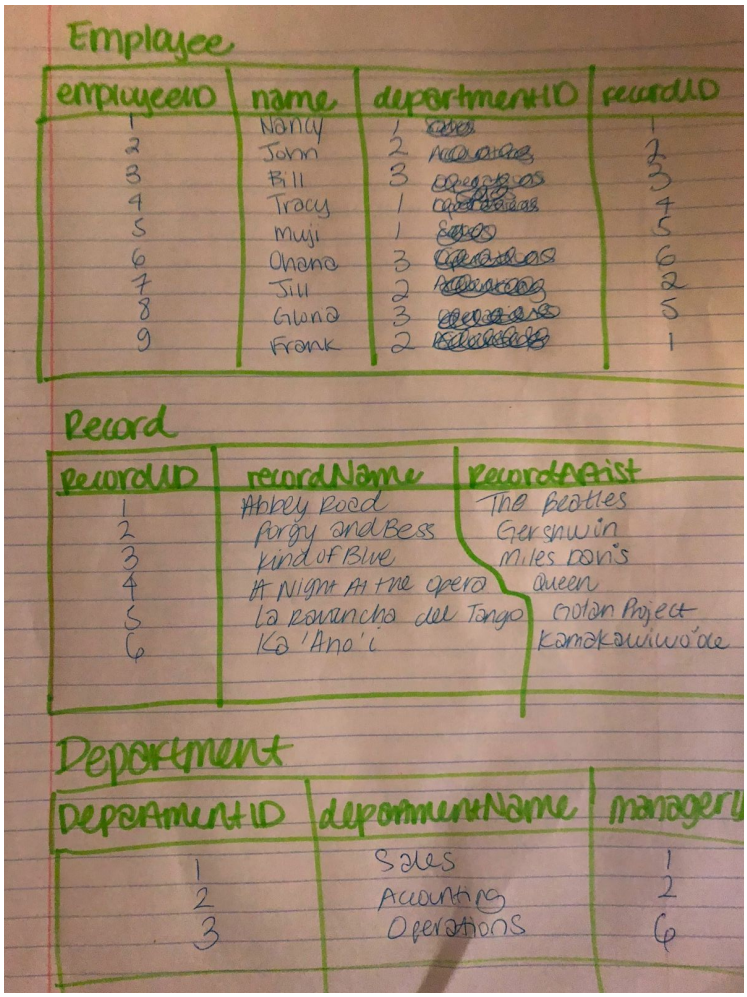
You work as a data analyst at a fishing/outdoors company. To identify new talents in database design, the company hosts an annual database schema competition. The winner takes home a commemorative statue known as the *Data Bass*. You won the competition last year, so a friend asked you to review his 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**.



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



Employee

employeeID	name	departmentID	recordID
1	Nancy	1	1
2	John	2	2
3	Bill	3	3
4	Tracy	1	4
5	Muji	1	5
6	Ohana	3	6
7	Jill	2	2
8	Gloria	3	5
9	Frank	2	1

Record

recordID	recordName	recordArtist
1	Abbey Road	The Beatles
2	Porgy and Bess	Gershwin
3	Kind of Blue	Miles Davis
4	A Night At the Opera	Queen
5	La Romancha del Tango	Gotan Project
6	Ka 'Ano'i	Kamakawiwo'ole

Department

departmentID	departmentName	managerID
1	Sales	1
2	Accounting	2
3	Operations	6

Question 4. ER Diagram (18 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(s) 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.

Assumptions:

1. We don't track authors that don't have books in this database as we are only interested in tracking those that do for this system
2. We assume that each new edition of a book would have a new ISBN number, and thus add a new row to the book table
3. We assume that publishers are only in this DB if they have published at least 1 book, else we do not consider them true publishers to track in the system
4. We assume each book can only have 1 publisher (per edition)
5. We assume that a shopping basket can exist without a customer (i.e. they are browsing without being logged into their customer account). We note this might lead to data quality constraints, so before a final purchase, we will ensure that the customer logs in so the ID can be captured into the data.
6. You can only create basket items once you have selected at least 1 ISBN