Data modeling and pipelining

This section tests your data engineering capabilities. You can attempt this section or the next (system design) section of the quiz.

Assume you are developing an application to bring out the latest in COVID related clinical trials. While there are any number of web sites providing dashboards on COVID related trials, none of them provide an easy way for users to get answers to questions of common interest such as:

1. What are the leading candidates for vaccination/treatment as evidenced by:
1. results obtained so far,
2. number and nature of subjects enrolled in the trials,
3. source of funding for the trials,
4. experience of the biotech/pharma company that is the innovator,
5. mechanism of action for the proposed drug/vaccine
2. What are the limitations of leading vaccine/therapy candidates such as:
1. Demography they've been tested on: E.g., age group, gender, race/nationality, social strata
2. Ease of distribution and administration: E.g., number of visits to the clinic needed per patient, storage conditions needed to keep the drug potent and stable, devices needed for administering the medicine etc.
3. Manufacturing capacity available to ramp-up production
3. What are the co-morbidities that complicate COVID treatment and what treatment options are being tested in each case?

In the questions below, you'll be asked to design a data pipeline and a data model needed to create an application that will answer the above questions.

Through this submission, you are expected to demonstrate your understanding of relational data modeling principles and your ability to grasp specifics of the domain you are modeling.

To understand the basics of data modeling, we suggest using a formal text book (such as https://opentextbc.ca/dbdesign01/) or course (such as https://www.coursera.org/learn/database-management/).

To understand the domain to model: Read the descriptions at clinicaltrials.gov for a few trials such as: https://clinicaltrials.gov/ct2/show/NCT04794946. Use the glossary at https://clinicaltrials.gov/ct2/about-studies/glossary for any clinical trials related terms you do not understand.

Question 1:

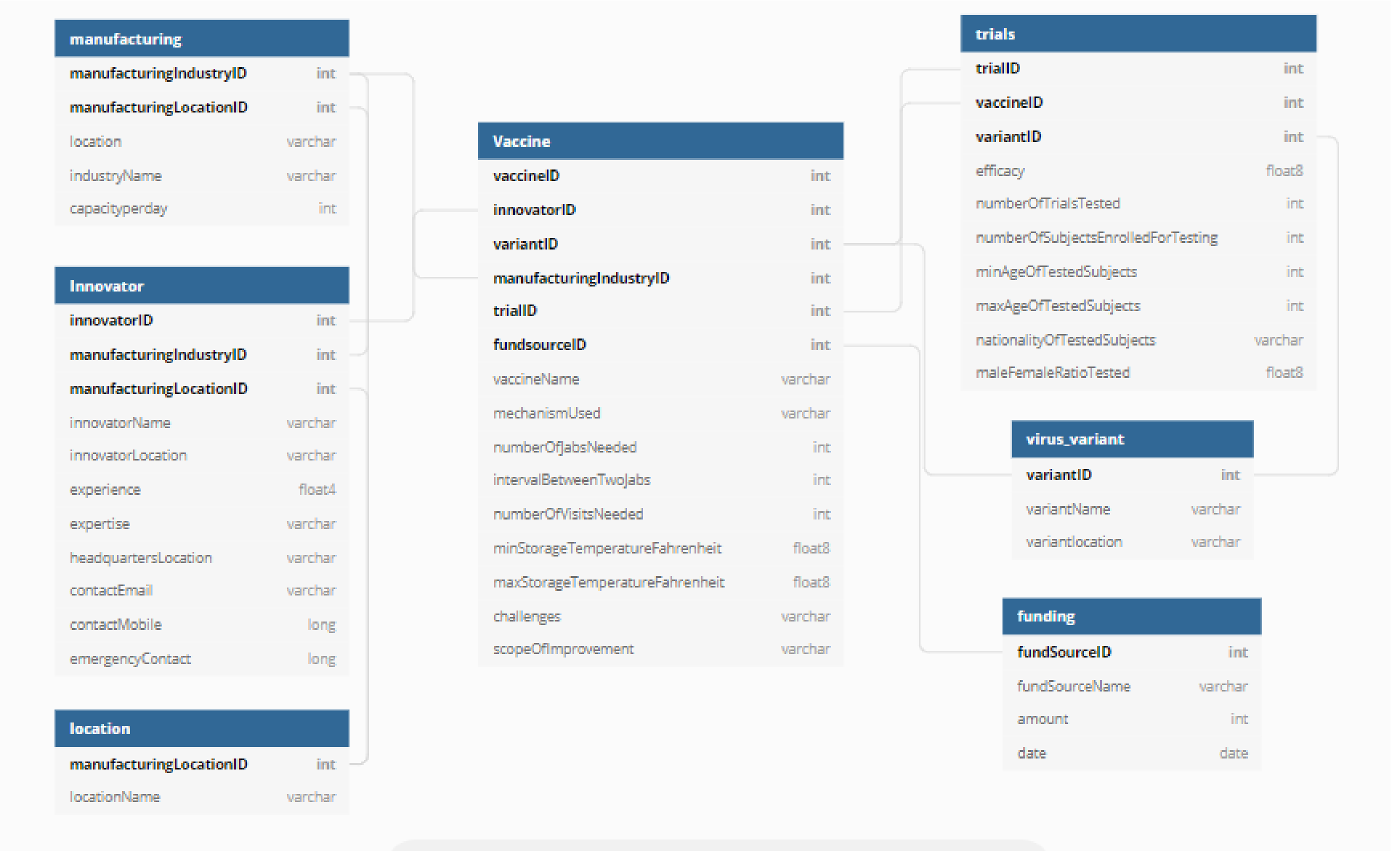
Using a diagramming/modeling tool such as the one at https://dbdiagram.io, draw an entity-relationship (ER) diagram providing a data model for the proposed application. Provide here a link to your work.

-- -----------------------------------

ER Model/Diagram Link:

https://dbdiagram.io/d/60b6523db29a09603d178023

Entity-Relationship model:



manufacturing	
manufacturingIndustryID	int
manufacturingLocationID	int
location	varchar
industryName	varchar
capacityperday	int

Innovator	
innovatorID	int
manufacturingIndustryID	int
manufacturingLocationID	int
innovatorName	varchar
innovatorLocation	varchar
experience	float4
expertise	varchar
headquartersLocation	varchar
contactEmail	varchar
contactMobile	long
emergencyContact	long

location	
manufacturingLocationID	int
locationName	varchar

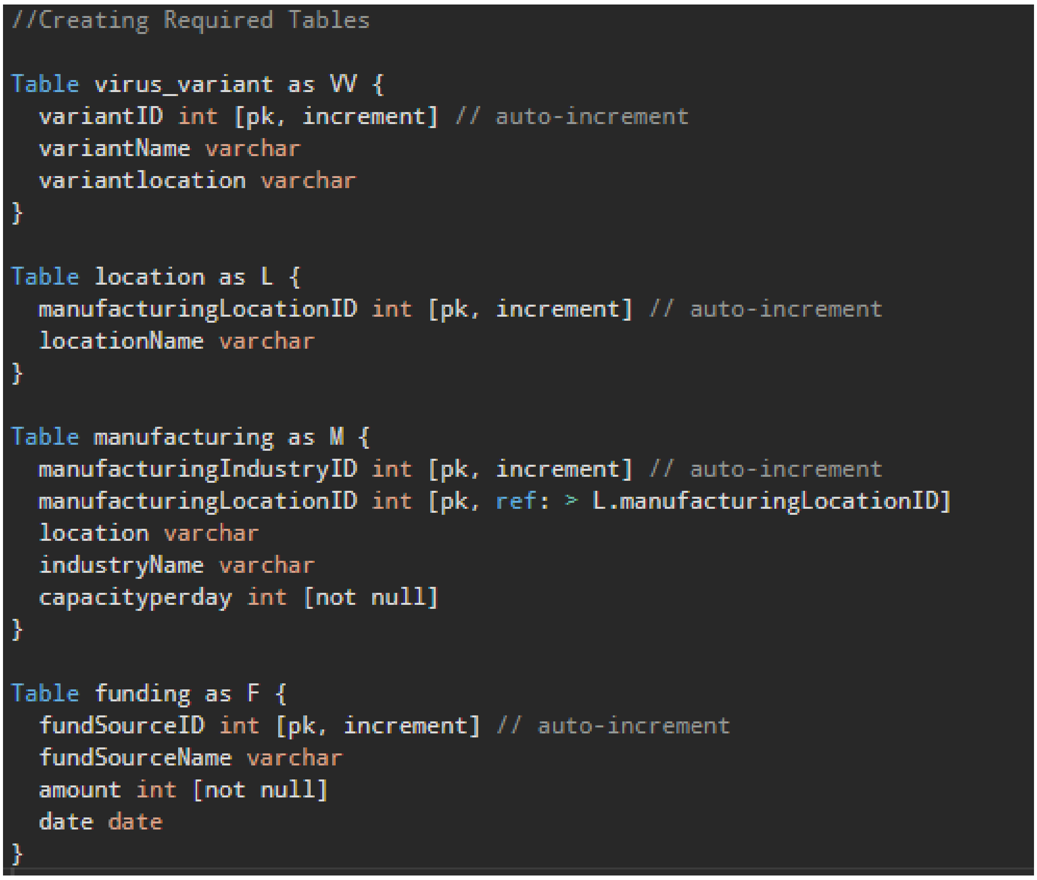
Vaccine	
vaccineID	int
innovatorID	int
variantID	int
manufacturingIndustryID	int
trialID	int
fundsourceID	int
vaccineName	varchar
mechanismUsed	varchar
numberOfJabsNeeded	int
intervalBetweenTwoJabs	int
numberOfVisitsNeeded	int
minStorageTemperatureFahrenheit	float8
maxStorageTemperatureFahrenheit	float8
challenges	varchar
scopeOfImprovement	varchar

trials	
trialID	int
vaccineID	int
variantID	int
efficacy	float8
numberOfTrialsTested	int
numberOfSubjectsEnrolledForTesting	int
minAgeOfTestedSubjects	int
maxAgeOfTestedSubjects	int
nationalityOfTestedSubjects	varchar
maleFemaleRatioTested	float8

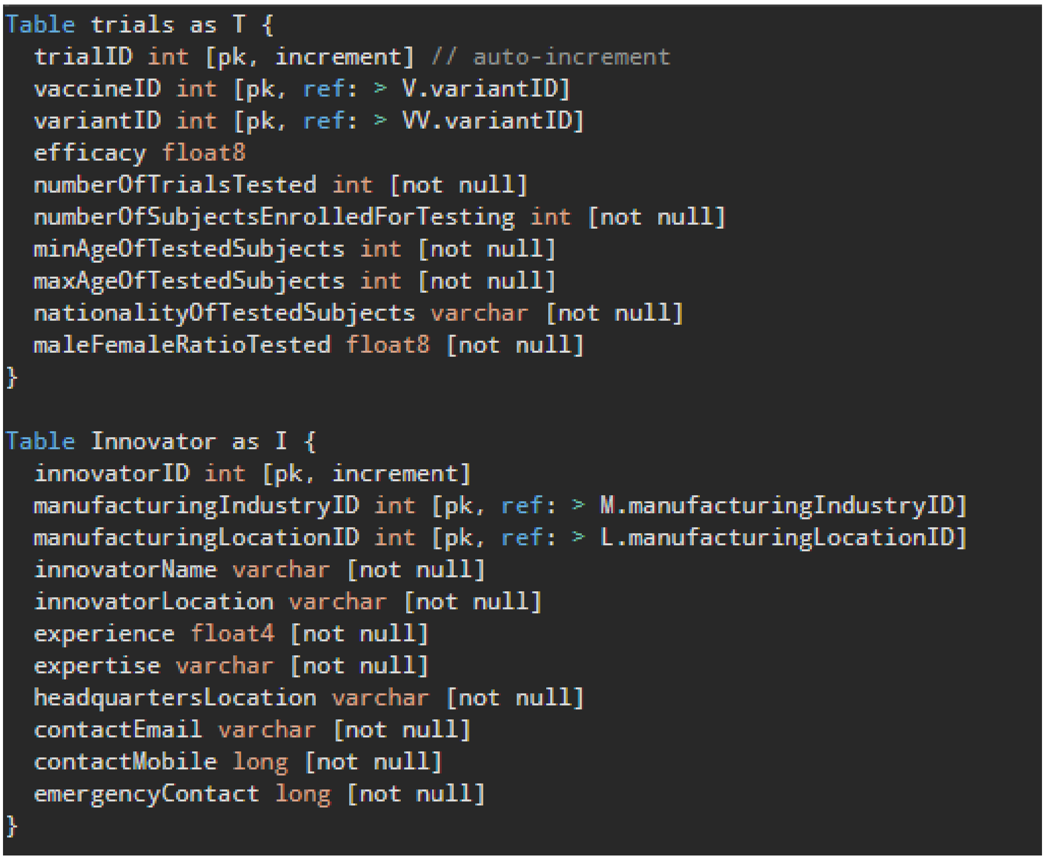
virus_variant	
variantID	int
variantName	varchar
variantlocation	varchar

funding	
fundSourceID	int
fundSourceName	varchar
amount	int
date	date

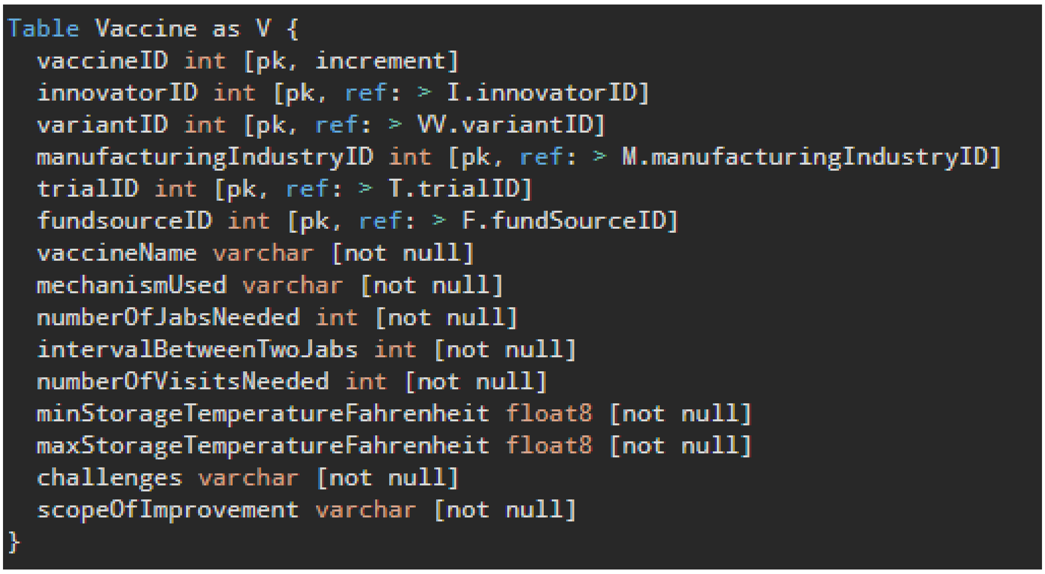
Script for ER model:



```
//Creating Required Tables

Table virus_variant as VV {
  variantID int [pk, increment] // auto-increment
  variantName varchar
  variantlocation varchar
}

Table location as L {
  manufacturingLocationID int [pk, increment] // auto-increment
  locationName varchar
}

Table manufacturing as M {
  manufacturingIndustryID int [pk, increment] // auto-increment
  manufacturingLocationID int [pk, ref: > L.manufacturingLocationID]
  location varchar
  industryName varchar
  capacityperday int [not null]
}

Table funding as F {
  fundSourceID int [pk, increment] // auto-increment
  fundSourceName varchar
  amount int [not null]
  date date
}
```



```
Table trials as T {
  trialID int [pk, increment] // auto-increment
  vaccineID int [pk, ref: > V.variantID]
  variantID int [pk, ref: > VV.variantID]
  efficacy float8
  numberOfTrialsTested int [not null]
  numberOfSubjectsEnrolledForTesting int [not null]
  minAgeOfTestedSubjects int [not null]
  maxAgeOfTestedSubjects int [not null]
  nationalityOfTestedSubjects varchar [not null]
  maleFemaleRatioTested float8 [not null]
}

Table Innovator as I {
  innovatorID int [pk, increment]
  manufacturingIndustryID int [pk, ref: > M.manufacturingIndustryID]
  manufacturingLocationID int [pk, ref: > L.manufacturingLocationID]
  innovatorName varchar [not null]
  innovatorLocation varchar [not null]
  experience float4 [not null]
  expertise varchar [not null]
  headquartersLocation varchar [not null]
  contactEmail varchar [not null]
  contactMobile long [not null]
  emergencyContact long [not null]
}
```



```
Table Vaccine as V {
  vaccineID int [pk, increment]
  innovatorID int [pk, ref: > I.innovatorID]
  variantID int [pk, ref: > VV.variantID]
  manufacturingIndustryID int [pk, ref: > M.manufacturingIndustryID]
  trialID int [pk, ref: > T.trialID]
  fundsourceID int [pk, ref: > F.fundSourceID]
  vaccineName varchar [not null]
  mechanismUsed varchar [not null]
  numberOfJabsNeeded int [not null]
  intervalBetweenTwoJabs int [not null]
  numberOfVisitsNeeded int [not null]
  minStorageTemperatureFahrenheit float8 [not null]
  maxStorageTemperatureFahrenheit float8 [not null]
  challenges varchar [not null]
  scopeOfImprovement varchar [not null]
}
```

Brief Explanation:

1. I have created 7 tables for this activity, namely: **vaccine**, **trials**, **innovator**, **manufacturing**, **virus_variant**, **location** & **funding**.
2. I have made all the fields as **[not null]** as all the data here is extremely important and having sparse table is not at all recommended in the medical domain.
3. I have tried to follow all the fundamental rules of ER modelling and database creation.
4. I have tried to maintain normalization up to **3NF** as higher normalization levels may lead to loss of data.
5. This is just a first-hand implementation of the system and I can surely come up with a more detailed ER model if given more time.