Stephen Morvant

EPBI 414 HW 4

**Part 1**

1. Large table with several columns
   1. Index for finding people from state, city, order by gender
      1. I would advise to query state, city, gender in that order. This first grabs the state(s) needed, of which cities will be a subset. Then grab city(s) of since gender is a subset of the people in the city(s). This means that you progressively grab smaller chunks until you have what you need.
   2. Index of race, gender query of white female is faster than just male. Why? Change how?
      1. It is due to the fact that the database is indexed by race then gender so looking for a race then gender first shrinks the total number of people significantly before looking for gender, so there are less data points that need to be checked by the database. When querying by just gender, every single person, regardless of race, needs to be checked and either grabbed or rejected, which is a lot of work for the database.
      2. I would change the indexing to put gender first so that instead of needing to check all of the data points for gender, it already has the gender indexed and can just grab all of the males very quickly, which eliminates needing to even look at the females, much less checking and rejecting them.
   3. Why can PK of just SSN be useful?
      1. Since SSN are always unique to a person, they are great for making primary keys. An example of a place where this is useful is for credit card companies. This can ensure that they don’t give credit to someone who has poor credit or is committing identity theft.
2. ACID in real world (using <http://beginnersbook.com/2015/04/acid-properties-in-dbms/> )
   1. ACID important and why
      1. In banking ACID is very important, especially when transferring between accounts. You need atomicity so that both a withdrawal and a deposit occur for each transaction so you aren’t magically gaining or missing money anywhere. The database needs consistency so it can make each checking account transaction to use a unique check number. Banks need isolation so tellers can be give a balance even when a withdrawal is happening at the same time. Durability is needed so that even if the bank loses power, etc. it still knows how much each account should have after all previous transactions.
   2. ACID less important and why
      1. In a research study where you gather race information or other factors, sometimes the data collector needs to input something that is incongruent with the initial parameters such subjects refusing to give race or not having a gender that is simply male/female.
3. Transaction isolation importance (slide 31)
   1. Since multi-site clinical trials can have inputs from many different people at once and people may be grabbing data to make some statistics to look at while other people are inputting data. If the transactions are not isolated, then the person who is grabbing statistics may end up starting to calculate something before something is inputted while the last value(s) used in the calculation is updated in the middle which causes the statistics to likely be completely incorrect.
4. Why does val\_sum violate 3NF
   1. Since val\_sum depends on two non- PK factors, it violates 3NF. Facts in a table cannot be references to (or depend upon) non-PK factors.

**Part 2**

fact\_procedure\_events

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| procedure\_event\_id  (PK) | visit | clinic | doctor | patient | amount\_billed | amount\_received | total\_costs |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

The procedure\_event\_id, amount\_billed, amount\_received, and total\_costs are all facts that come from original procedure\_events table. The dimensions are all below in tables that would be drawn into a star schema if I had the requisite abilities. (I do not though.) Each dimension in the fact table is named as follows: visit🡪 fact\_procedure\_events.visit; clinic 🡪 fact\_procedure\_events.clinic; …

dim\_procedure 🡪 original procedure\_events table

|  |  |  |  |
| --- | --- | --- | --- |
| procedure\_code (PK) | procedure\_name | procedure\_specialty | outpatient\_procedure |
|  |  |  |  |
|  |  |  |  |

dim\_visit🡪fact\_procedure\_events.visit

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| visit\_id (PK) | doctor\_id | patient\_id | clinic\_id | visit\_date |
|  |  |  |  |  |
|  |  |  |  |  |

dim\_clinic🡪 fact\_procedure\_events.clinic

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| clinic\_id (PK) | clinic\_name | clinic\_street | clinic\_city | clinic\_state | clinic\_zip | clinic\_group |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

dim\_doctor🡪 fact\_procedure\_events.doctor

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| doctor\_id (PK) | first\_name | last\_name | primary\_clinic\_id | year\_joined\_system | primary\_specialty |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

dim\_patient🡪 fact\_procedure\_events.patient

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| patient\_id (PK) | pcp\_id | first\_name | last\_name | date\_of\_birth | home\_street | home\_city | home\_state | home\_zip |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

1. Does the fact table meet 2NF?
   1. Yes because the primary key (procedure\_event\_id) is only one column. You cannot violate 2NF without multiple columns for the primary key. To violate, a non- PK field must depend on only part of the PK.
2. Does the fact table meet 3NF?
   1. If total\_cost is calculated using amount\_billed and amount\_recieved, then it fails 3NF since both of the latter are not part of the primary key for that table. Since doctors probably depend on clinic (they only work at certain one(s), then it fails 3NF. If there is dependency of the patients upon which doctors they see or which clinics they go to, then it fails 3NF. If all of those assumptions are wrong, then the fact table is 3NF compliant, but I doubt it
3. Which dimensions meet 2NF and why? Which meet 3NF and why?
   1. All meet 2NF because the primary key for each is only one column.
   2. Clinics and patients fail because zip depends on state, visits pass because nothing is dependent on anything else in the table except the primary key, and the same can be said for doctors
4. If 6th added, why use date\_id? If you did use this, why considered degenerate?
   1. Date dimension tables are great for dealing with missing rows as well as making it easier for the computer to handle since dates are often hard for computers to deal with.
   2. It could be considered degenerate if its only attributes were already used or incorporated in the visits dimension, thus preventing it from having unique information.