Northeastern University

Final Project Assignment — Pharmacy Claims

QINGRU YANG

ALY6030: Data Warehousing & SQL

Instructor: Zhi He

3/30/2022

Part 1 Normalization

Firstly, there are some attributes repeat in the dataset, therefore the dataset does not meet 1nf. (fill_date1, fill_date2, fill_date3, copay1, copay2, copay3, insurancepaid1, insurancepaid2, insurancepaid3). We need to rearrange the data to eliminate duplicate attributes to meet 1nf.

Secondly, there is Partial Dependency, therefore the dataset does not meet 2nf. To meet 2nf, I divided the table into multiple tables, which are 2-dimension tables (dim_member and dim_drug_ndc), and 1 fact table (fact_drug).

Finally, to meet 3nf, divide dim_drug_ndc table into 2 tables (dim_drug_form and dim_drug_brand_generic) to eliminate the Transitive Dependency.

Therefore, I divided into 1 fact table(fact_drug) and 4-dimension tables (dim_memebr, dim_drug, dim_drug_form, dim_drug_brand_generic).

member_id	member_first_name	member_last_name	member_birth_date	member_age	member_gender	drug_ndc	drug_name	drug_form_code	drug_form_desc	drug_brand_generic_code	drug_brand_generic_desc	fill_date	copay	insurancepaid
10001	David	Dennison	1946/6/14	72	M	43353084	Risperidone	TB	Tablet	1	Generic	2017/10/31	15	50
10002	John	Smith	1962/1/2	56	M	54569519	Amoxicillin	OS	Oral Solution	1	Generic	2018/6/14	50	130
10003	Jane	Doe	1982/5/4	36	F	54569382	Ambien	TB	Tablet	2	Brand	2017/12/30	35	250
10004	Elaine	Rogers	1983/10/12	34	F	18508530	Diprosone	TC	Topical Cream	1	Generic	2017/11/9	15	600
10001	David	Dennison	1946/6/14	72	M	54569382	Ambien	TB	Tablet	2	Brand	2018/1/15	20	650
10001	David	Dennison	1946/6/14	72	M	43353084	Risperidone	TB	Tablet	1	Generic	2018/2/22	15	48
10003	Jane	Doe	1982/5/4	36	F	54569382	Ambien	TB	Tablet	2	Brand	2018/5/16	35	322
10004	Elaine	Rogers	1983/10/12	34	F	18508530	Diprosone	TC	Topical Cream	1	Generic	2017/12/8	15	712
10001	David	Dennison	1946/6/14	72	M	54569382	Ambien	TB	Tablet	2	Brand	2018/2/14	20	648
10001	David	Dennison	1946/6/14	72	M	43353084	Risperidone	TB	Tablet	1	Generic	2018/5/8	15	55
10001	David	Dennison	1946/6/14	72	M	54569382	Ambien	TB	Tablet	2	Brand	2018/3/13	20	648

Questions:

Q1: For each fact variable in your fact table, what type of fact is it?

In the fact table, the fact variables are copay and insurancepaid. They are **additive** because they can be used with any aggregation function.

Q2: In your fact table, describe the grain in one sentence.

Each fact row is the patient's drug fill record, as well as their payment history.

Part 2 Primary and Foreign Key Setup in MySQL

Questions:

Q1: What are the primary keys you designated for each of your tables? For each PK, is it a natural key or a surrogate key?

Table	Primary Key	Type
Fact_drug	id	Surrogate key
Dim_member	Member_id	Natural key
Dim_drug	Drug_ndc	Natural key
Dim_drug_form	Drug_form_code	Natural key
Dim_drug_brand_generic	Drug_brand_generic_code	Natural key

Q2: What are the foreign keys you designated for each of your tables? For each FK, which table did you reference where that FK is listed as the PK?

The member_id is PK of the dim_member table, and drug_ndc is PK of dim_drug_ndc table. The drug_brand_generic_code is PK of dim_drug_brand_generic table, and drug_form_code is PK of dim_drug_form_code. In fact_drug table, the FK are member_id, drug_ndc, drug_brand_generic_code and drug_form_code.

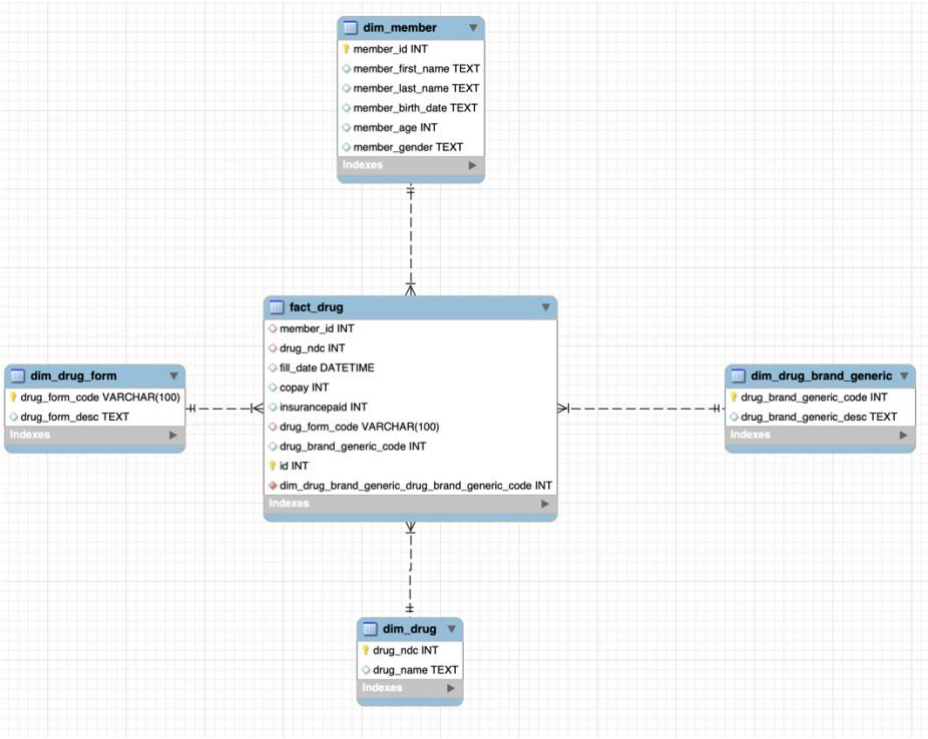
Q3: For each FK, what did you tell MySQL to in case of deletion or update (CASCADE, SET NULL, or RESTRICT)? Why did you select the option that you did for each FK?

I choose the SET NULL option for each FK because for this option when we delete the parent table row, it will set the column value to NULL in the child table. In this case it

is easier to make the change. However, when the parent changes, CASCADE will propagate the change. If we delete a row, the rows in the constrained table that reference that row will also be deleted. If there are child rows that reference the value of the parent row, RESTRICT will cause you to be unable to delete the given parent row.

Part 3 Entity Relationship Diagram

The ERD shows below:



Part 4 Analytics and Reporting

Q1: How many prescriptions were filled for the drug Ambien?

From below table, we can see that there are 5 prescriptions filled for Ambien.

drug_name	number_prescri...
Amoxicillin	1
Diprosone	2
Risperidone	3
Ambien	5

Q2: How many unique members are over 50 years of age? And how many prescriptions did they fill?

age_group	member_numbers	sum_copay	sum_insurancepaid	number_prescriptions
<50	2	100	1884	4
50+	2	155	2229	7

There are 2 unique members are over 50 years of old. And 7 prescriptions filled.

Q3: For member ID 10004, what was the drug name listed on their most recent fill date? How much did their insurance pay for that medication?

member_id	member_first_na...	member_last_na...	drug_name	fill_date	insurancepaid
10001	David	Dennison	Risperidone	2017-10-31	50
10001	David	Dennison	Ambien	2018-01-15	650
10001	David	Dennison	Ambien	2018-02-14	648
10001	David	Dennison	Risperidone	2018-02-22	48
10001	David	Dennison	Ambien	2018-03-13	648
10001	David	Dennison	Risperidone	2018-05-08	55
10002	John	Smith	Amoxicillin	2018-06-14	130
10003	Jane	Doe	Ambien	2017-12-30	250
10003	Jane	Doe	Ambien	2018-05-16	322
10004	Elaine	Rogers	Diprosone	2017-11-09	600
10004	Elaine	Rogers	Diprosone	2017-12-08	712

drug_name	fill_date	insurancepaid	member_id	member_first_na...	member_last_na...	fill_times
Risperidone	2018-05-08	55	10001	David	Dennison	1
Amoxicillin	2018-06-14	130	10002	John	Smith	1
Ambien	2018-05-16	322	10003	Jane	Doe	1
Diprosone	2017-12-08	712	10004	Elaine	Rogers	1

From above table, we can see that: For member_id 10004, the drug named **Diprosone** listed on their most recent fill date. member_id 10001, the insurance paid is 55. member_id 10002, the insurance paid is 130. member_id 10003, the insurance paid is 322.member_id 10004, the insurance paid is 712.