

**Module 6 Assignment - Final Project Assignment — Pharmacy Claims**

**ALY6030 – Data Warehousing & SQL**

**CRN: 20361**

**Manmitha Pantangi**

**Email:** [**pantangi.m@northeastern.edu**](mailto:pantangi.m@northeastern.edu)

**Part 1 – Normalization**

**Introduction**

Efficiently handling pharmacy claims data is crucial for insurance firms that rely on Pharmacy Benefit Managers (PBMs). To enhance data analysis and reporting workflows, it's vital to establish a normalized database schema that conforms to Third Normal Form (3NF) principles. This project aims to restructure raw pharmacy claims data into a series of relational tables, facilitating easier querying and more meaningful analysis.

**Data Overview**

The dataset contains detailed sample pharmacy claims data for made-up members, including a wide variety of fields such as member information (ID, first name, last name, birth date, age, gender), drug information (NDC, name, form code, form description, strength, quantity, brand or generic), and claim details (fill dates, copays, and insurance paid amounts).

**Steps to Normalize the Data:**

To convert the sample data into a set of relational tables that meet the 3rd Normal Form (3NF) standards and create a star schema, we'll follow a structured approach. This involves identifying the fact and dimension tables, ensuring we remove any redundancy and that each table represents a single concept.

* **Identify Dimension Tables:** These tables contain attributes that describe the business entities. Based on the provided data, potential dimension tables include:
  + **Member Dimension (dim\_member):** Contains unique member information such as member\_id, member\_first\_name, member\_last\_name, member\_birth\_date, member\_age, and member\_gender.
  + **Drug Dimension (dim\_drug):** Includes details about the drugs like drug\_ndc, drug\_name, drug\_form\_code, drug\_form\_desc, drug\_strength, and drug\_brand\_generic\_desc.
* **Identify the Fact Table:** This table focuses on the transactions or events, in this case, pharmacy claims. The fact table will include:
  + **Pharmacy Claims Fact (fact\_claims):** Contains keys that link to the dimension tables and measurable, numeric data related to the pharmacy claims such as fill\_date1, fill\_date2, fill\_date3, copay1, copay2, copay3, insurancepaid1, insurancepaid2, and insurancepaid3.
* **Normalization:** Ensure the data meets 3NF standards by:
  + **Eliminating Redundant Data:** For members appearing more than once due to multiple drug fills, ensure their personal information is stored once in the Member Dimension.
  + **Removing Partial Dependency:** All non-primary key columns in each table should only depend on the primary key.
  + **Removing Transitive Dependency:** Non-primary key columns should not depend on other non-primary key columns.
* **Schema Rationalization:** The selected schema segregates data into dimension (Member, Drug) and fact (Prescription) tables. This arrangement clearly distinguishes between entities and their characteristics, guaranteeing scalability, reducing data duplication, and enhancing the efficiency of query execution.

**1. Fact Variable Types:**

* **Additive Facts:** These are numeric facts that can be summed up through all dimensions in the fact table. Examples include copay1, copay2, copay3, insurancepaid1, insurancepaid2, and insurancepaid3.
* **Semi-Additive Facts:** These facts can be summed up for some dimensions but not others. In this case, since we aren't creating a separate date dimension, all facts remain additive.
* **Non-Additive Facts:** There are no clear non-additive facts in this dataset, as most numeric values can be aggregated in some form.

**2. Fact Table Grain:** Each row in the Pharmacy Claims Fact table represents a single pharmacy claim event for a drug filled by a member on specific fill dates, including the copay amounts and insurance payments for up to three fills.

**Part 2 — Primary and Foreign Key Setup in MySQL**

**1. Primary Keys Designation:**

**Member Table:** member\_id as a surrogate key. Chosen because while it could naturally occur, assigning it as a surrogate key ensures each member has a unique identifier, avoiding potential issues with duplicate names or birthdates.

**Drug Table:** drug\_ndc as a natural key. The National Drug Code (NDC) naturally serves as a unique identifier for each drug.

**Claims Table:** A surrogate key, perhaps claim\_id, since claims themselves may not have a natural unique identifier across all dimensions.

**2. Foreign Keys Designation:**

**Prescription Table:**

* FKs: member\_id and drug\_ndc.
* member\_id references the PK in the Member table.
* drug\_ndc references the PK in the Drug table.

**3. Actions for FKs on Deletion or Update:**

* **member\_id:** **SET NULL** on deletion to maintain prescription records even if the member data is removed, ensuring data integrity without losing the prescription history. Preserves the integrity of prescription records by not deleting them even if the member data is removed, which is important for historical data analysis.
* **drug\_ndc: RESTRICT** on deletion to prevent deletion of a drug that has associated prescriptions, ensuring referential integrity. Prevents accidental deletion of drugs that are linked to prescriptions, ensuring that all prescription data remains consistent and valid.
* **Update Actions: CASCADE** for both member\_id and drug\_ndc to automatically update related prescription records if a member's ID or a drug's NDC is changed. Ensures that any changes to member IDs or drug NDCs are reflected in the prescription records, maintaining data consistency without manual updates.

**Part 3 — Entity Relationship Diagram (ERD)**

A screenshot of a computer

Description automatically generated

**Part 4 — Analytics and Reporting**

1. **Number of Prescriptions by Drug Name**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**How many prescriptions were filled for the drug Ambien? –** 5

1. **Total Prescriptions and Unique Members by Age Group**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**How many unique members are over 65 years of age? –** 3

**How many prescriptions did they fill? –** 5

1. **Most Recent Prescription Fill Date and Insurance Paid**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**For member ID 10003, what was the drug name listed on their recent fill date? -** Ambien

**How much did their insurance pay for that medication? –** 322

**References**

GfG. (2023e, December 4). *Third normal form (3NF)*. GeeksforGeeks. <https://www.geeksforgeeks.org/third-normal-form-3nf/>

WilliamDAssafMSFT. (2023, September 26). *Primary and foreign key constraints - SQL server*. Microsoft Learn. <https://learn.microsoft.com/en-us/sql/relational-databases/tables/primary-and-foreign-key-constraints?view=sql-server-ver16>

*What is Star Schema?* (n.d.-b). Databricks. <https://www.databricks.com/glossary/star-schema#:~:text=A%20star%20schema%20is%20used,like%20transaction%20amounts%20and%20quantities>).

*MySQL CASE Function*. (n.d.). <https://www.w3schools.com/sql/func_mysql_case.asp>

Simplilearn. (2023b, June 6). *Fact Table vs. Dimension Table - Differences Between The Two*. Simplilearn.com. <https://www.simplilearn.com/fact-table-vs-dimension-table-article#:~:text=Fact%20tables%20and%20dimension%20tables%20play%20different%20but%20important%20roles,data%20analysis%20and%20decision%2Dmaking>.