DNSC 6305- Data Management

Group Assignment 2 – Fall 2023

Participation Details:

Submission Date: November 2, 2023, 6:45 pm
Group Lead for the Assignment: Maneesh Tekwani

All Participants

Student Name	Question No	Group Participation in discussions(min)/week	Final submission date	Percent Participation in discussions (0-100)	Percent Participation in final proof reading and editing
Anukshan Ghosh	E, S, P1, P2	110	10/31/2023	100	100
Allison Ko	E, P2, P3, P4	120	11/1/2023	100	100
Sean Vaghedi	E, S, P1, P2	110	10/31/2023	100	100
Alex Le	E, S, P1, P2, P4	110	11/1/2023	100	100
Zaheer Soleh	E, P2, P3, P4	120	11/1/2023	100	100

Key:
E: ER Diagram
S: Schema
P: Problems in DB (P1, P2, P3, P4)

The group have used the following tools for discussion:

1. Blackboard Discussion → DMFA- Group 1 → Group Discussion Board
2. WhatsApp Group
3. Email
4. Virtual Calls (Google Meet)

Part I – Question 3: What is your general opinion on the data files provided? Would you consider the files to be tidy and organized? Provide any suggestions on improving the design of the data files.

Our group found the data dictionary to be a helpful, tidy, and organized resource to guide us in understanding the Entities and Attributes within them. Furthermore, the Data Dictionary also provided us information on the attributes, their data types (numeric, character, varchar), and whether any attributes were nullable or not.

However, when opening the specific csv files, we realize that there were some discrepancies between the Data Dictionary and csv files:

1. Attributes like Offense_Type_ID wasn't available.
2. An Entity needed for our analysis REF_Race wasn't available.
3. VICTIM_OFFENDER_ID (which was a primary key for Entity VICTIM_OFFENDER_REL) was null.
4. Some Entities didn't have primary keys:
   - VICTIM_OFFENDER_REL Entity
   - VICTIM_OFFENSE Entity
5. OFFENDER and VICTIM Entities had character values (instances of 'NS' and 'BB' in the tuple/rows) within the age_num attribute for both.

Therefore, we made the changes below to improve the design of the data files:

1. Within the Arrestee, Offense, and Offense_Type Entities, the attribute Offense_Type_ID wasn't available, so our group decided to use Offense_Code instead, since that is a unique attribute within the Offense_Type Table and could be used for the relationship between Arrestee and Offense_Type (Offense_Type_ID as a FK and a Primary Key for the Offense_Type Entity) and Offense and Offense Type.

2. Our group noticed in the Arrestee, Offender and Victim Entities, that each of these respective entities were referring to REF_RACE, which meant that the attribute 'race_id' was a Foreign Key for each of them. We looked for REF_RACE in the Data Dictionary; however, it wasn't there, so we decided to look REF Race.csv File and identified the different attributes, attribute types, data types, whether it was nullable, and wrote comments. Please refer below to the table, which we recommend adding to the data dictionary for the purpose of representing REF_RACE as a new Entity with its attributes. We represented all attributes within our ER Diagram since there are no Foreign Keys.

3. Our group noticed that within the VICTIM_OFFENDER_REL Entity, VICTIM_OFFENDER_ID, which was meant to hold an Internal Unique ID or Primary Key for the Entity was filled with null values. This meant that we couldn't consider it as a primary key for the Entity. Instead, we assumed, that the Primary Key for the VICTIM_OFFENDER_REL Entity would be a composite of its Foreign Keys. The New Primary Key was VICTIM_ID, OFFENDER_ID, RELATIONSHIP_ID. Doing so, helped later when we were creating the Entity tables in PostgreSQL.

4. Our group noticed that within the VICTIM_OFFENSE Entity, there were two foreign keys and no primary key, so we made a primary key with the foreign key's composite. The

New Primary Key was VICTIM_ID, OFFENSE_ID. Doing so, helped later when we were creating the Entity tables in PostgreSQL.

a. Further Explanation on this Entity: VICTIM_OFFENSE Entity seemed like an Entity in our Data Dictionary; however, when examining its relationship with VICTIM and OFFENSE Entities we found it to be in a Many-to-Many relationship, so we changed VICTIM_OFFENSE to be relation in our ER instead of an Entity.

5. The OFFENDER and VICTIM Entities within the csv files had instances with ‘NS’ and ‘BB’ values for the age_num attribute. This wasn’t a problem until we tried bulk loading the data in these two entities. For both entities, the attribute age_num would only take the data type numeric and since ‘NS’ and ‘BB’ are characters, Postgre didn’t let us proceed until all tuples in the age_num attribute was numeric with a data size limit of 3 integers. We fixed this by using the !awk command to drop the instances of character values and replace them with Null values instead. We decided to do this to align with the requirements of the Data Dictionary and use age_num for numeric calculations if we need to do so in the future.

Addition to the Data Dictionary for Entity REF_RACE:

TABLE NAME	COLUMN NAME	DATA TYPE	DATA SIZE LIMIT	NULLABLE	COMMENTS
REF_RACE	RACE_ID	NUMBER	2	N	Internal Unique ID for this Race ID
	RACE_CODE	CHARACTER	2	N	NIBRS Race_Code
	RACE_DESC	VARCHAR2	100	N	NIBRS Race_DESC
	SORT_ORDER	NUMBER	2	N	NIBRS Sort_Order for Race ID
	START YEAR	NUMBER	4	Y	Year when that specific RACE_ID classification was introduced for grouping by RACE
	END YEAR	NUMBER	4	Y	Year when that specific RACE_ID classification was discontinued for grouping by RACE
	NOTES	VARCHAR2	100	Y	Notes to provide more clarity on specific attributes within the RACE Entity