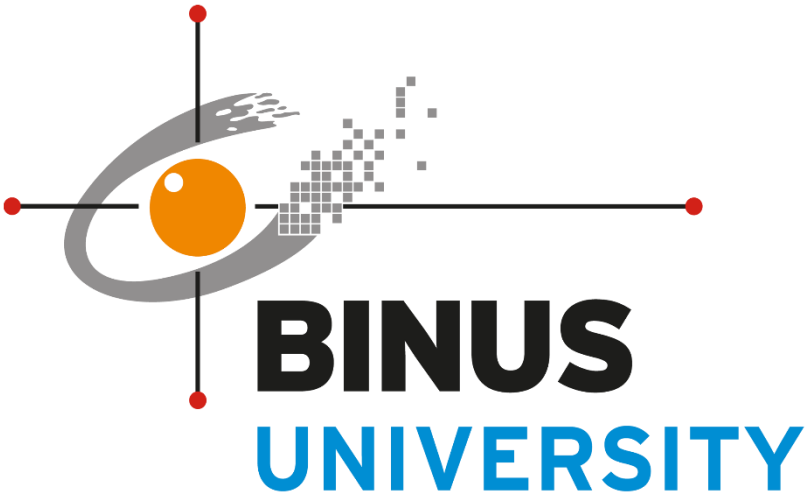
COMP6799001 – Database Technology

AoL: Final Report



By:

Benedictus Yogatama Favian Satyajati – 2602106364 – LT01

Dean Hans Felandio Setiadi Saputra – 2602094206 – LT01

Vincensius Theodorico Kenzie – 2602103974 – LT01

Justin Tjandra – 2602095902 – LT01

Bina Nusantara University

2023/2024

Chapter 1 – Case Description

The case that we used was taken from the website data.world, a website that provides a unified view of data, resources, and knowledges alike for database uses. More specifically, we have taken a data with the title of **Health Facility General Information** which can be accessed using the following link: https://data.world/healthdatany/vn5v-hh5r/workspace/intro. The dataset basically shows the location of health care facilities, and programs from the HFIS (Health Facilities Information System). The facilities mentioned contains hospitals, nursing homes, diagnostic treatment centers, midwifey birth centers, certified home health care agencies, licensed home care services agencies, long term home health care programs, hospices, and licensed adult care facilities which was last updated on November 16, 2023, and was provided by the New York State Department of Health themselves. This dataset consists of 6,141 rows of data and 36 distinct columns that define the data. The following table contains the **name** of each column and their respective **description** along with their **data types**.

Column Name	Description	Data Type
Facility_ID	Site specific facility identification number.	Number
Facility_Name	Current facility name.	Plain Text
Short_Description	Abbreviated Name for Type of Facility	Plain Text
Description	Type of Facility	Plain Text
Facility_Open_Date	Date facility opened	Date & Time
Facility_Address_1	Address Line 1 of Facility	Plain Text
Facility_Address_2	Address Line 2 of Facility (optional)	Plain Text
Facility_City	City of Facility	Plain Text
Facility_State	State of Facility (will always be New York State)	Plain Text
Facility_Zip_Code	Zip code of facility	Plain Text
Facility_Phone_Number	Phone Number of facility	Plain Text
Facility_Fax_Number	Fax Number of facility	Plain Text
Facility_Website	Website of facility	Plain Text
Facility_County_Code	Gazetteer code for county facility is located in	Number
Facility_County	County where facility is located	Plain Text
Regional_Office_ID	Identification Number of NYSDOH Regional Office	Number
Regional_Office	Name of NYSDOH Regional Office for Facility	Plain Text
Main_Site_Name	Name of facility's main site. (if the facility is an extension clinic)	Plain Text

Main_Site_Facility_ID	Facility identification number of main site	Number
Operating_Certificate_Number	Operating certificate number for facility	Plain Text
Operator_Name	Name of current operator of facility	Plain Text
Operator_Address_1	Address Line 1 of current operator	Plain Text
Operator_Address_2	Address Line 2 of current operator	Plain Text
Operator_City	City of current operator	Plain Text
Operator_State	State of current operator	Plain Text
Operator_Zip_Code	Zip code of current operator	Plain Text
Cooperator_Name	Name of current cooperator (optional)	Plain Text
Cooperator_Address_1	Address line 1 of current cooperator	Plain Text
Cooperator_Address_2	Address line 2 of current cooperator	Plain Text
Cooperator_City	City of current cooperator	Plain Text
Cooperator_State	State of current cooperator	Plain Text
Cooperator_Zip_Code	Zip code of current cooperator	Plain Text
Ownership_Type	Type of legal ownership	Plain Text
Facility_Lattitude	System for representing healthcare facilities on map.	Number
Facility_Longitude	System for representing healthcare facilities on map.	Number
Facility_Location		Location

Table 1 - Column Descriptions

In database, in one of the structures of it (Based on Elmasri & Navethe, Mini-World → Requirements & Analysis → Conceptual Design → Logical Design → Physical Design), there exist a way of illustrating a database in a Conceptual Design called an ERD. ERD stands for Entity Relational Diagram, and it has 2 most common ways of drawing; one is called Chen's Notation and the other is called Crow's Foot. We will be using Chen's Notation in this report which consists of an Entity, their Attributes, and a Relationship that connects one or more Entities. Entity is basically a noun whereas attributes are the properties of an entity. On the other hand, a relationship defines the verb of the entities. There are many types of entities, attributes, and relationships in an ERD, here are some of them and their respective brief description:

Entity

Entities are objects in a database and is symbolized with a rectangle in an ERD.

- Entity

A normal entity does not have any special properties and only symbolizes the object contained within it. Symbolized with a rectangle with only one outline.

- Weak Entity

A weak entity is an entity that cannot be identified uniqely by its own attributes and instead depends on another entity's attribute(s). Symbolized with a rectangle with two outlines. The relationship that defines the relationship of this entity and another entity is also drawn with two outlines.

Attribute

Attributes are properties of an entity and is symbolized with an oval in an ERD.

- Attribute

Shows a normal property of an entity without any special properties. Drawn with a simple oval.

- Key Attribute

Shows the key to the entity. Drawn with an oval with an underline under its text.

- Multivalued Attribute

Shows that the specific attribute has many values. For example, for a entity called *Film*, it might have a multivalued attribute called *Genre*. Drawn with an oval with two outlines.

- Derived Attribute

Shows that the attribute obtains its' value from another attribute. For example, for an entity called *Person*, there could be an attribute called *DateOfBirth* and a derived attribute called *Age* which calculates its value on *DateOfBirth*. Drawn with an oval with a dashed outline.

Relationship

Relationship is a way of showing the connection between one or more entities and is symbolized with a diamond in an ERD and a line that connects it with the entity.

Still in the relationship topic, there is also a property called cardinality and participation. Cardinality shows the degree of the relationship while Participation shows how much (in parts) of the entity takes part in the relationship.

Cardinality

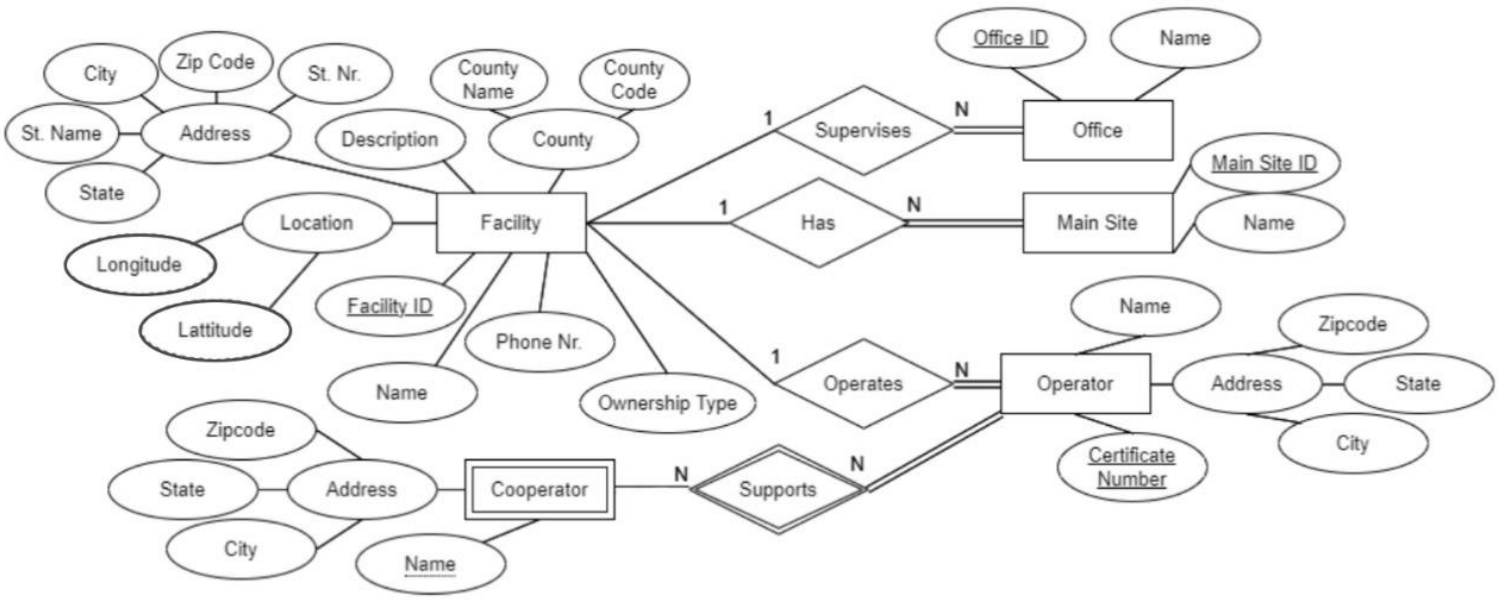
- 1:1 (One-to-One) defines that the relationship is done by 0 or 1 part of the entity in both ends.

- 1:M / M:1 (One-to-Many or Many-to-One) defines that the relationship is done by 0 or 1 part of one of the entities and 0 or *n* (many) parts of the other end of the entity.
- M/N (Many-to-Many) defines that the relationship is done by 0 or *n* (many) parts of both ends of the relationship.

Participation

- Total Participation defines that all the parts of the entity are involved in the relationship. (If cardinality is 1, then the possibility is always 1 and never 0; if cardinality is M, then minimum is 2).
- Partial Participation defines that not all parts of the entity might be involved in the relationship. (If cardinality is 1, then the possibility is 1 or 0; if cardinality is M, then the possibility is 0 to *n* parts).

By analyzing the dataset, here's the ERD of Chen's Notation that we created of the dataset:



Chapter 2 – Indentifying Problem

Database

In its simplest understanding, a database is a system of **integrated collection of data**. The size of any database has the possiblity of growing to a very large size. Databases are also modelled after real world enterprises and are created based on their usage. Usually, a database will consist of several entities and a relationship that defines the connection between two or more entities. Someone or a corporation might consider the use of database because of several reasons. Those reasons being that the world is now currently ongoing a shift from the so-called computational era to become more of an information era. Other than that, databases are also used because as time goes by, datasets will also experience a significant increase in terms of their diversity and volume as also mentioned earlier. A database will also decrease the chance of a collision happening because it takes concurrency control into account. All-in-all, the use of database will appeal to the end-users and the Database Administrator themselves because it makes the act of designing logical/physical schema much easier, the security handling and authorization is much more secure, and if a crash should happen, the crash recovery system is much more reliable.

Database for the Dataset

In this case specifically, database is pretty much needed mainly because of the sheer immense size of the dataset of more than 6,000 number of rows; the use of database will give the dataset a sense of scalability with its evergrowing size. Moreover, A database will guarantee a clearer and more reliable data organization system, ease of access, and querying. This dataset also needs to use a database to ensure data integrity which enforces unique identification codes that prevent duplicate entries and maintain data consistency. Because this dataset is of public information, it is also valid that this dataset is a data that represents and shares a state's condition to the public that represents the state and the people of the state so the security and access control of the data needs to also be controlled, maintained, and managed by a database administrator of some kind. Futhermore, suppose there happens an error or data loss to the dataset, the data contained in a database can be recovered and backed up easily. A database is also needed for the dataset because of the concept of **ACID** in database. ACID stands for **Atomicity**, **Consistency**, **Isolation**, and **Durability** and is the backbone of a database. ACID ensures that in the case of errors, anomalies, and other similar things, the database stands valid, and no significant data loss or collision might happen. Atomicity means that every statement/query in a database such as *read, write, update, delete, drop*, etc is executed as a

single unit that cannot be divided further. This means that only two possibilites might occur: either the statement is executed as a whole and fully, or it is not executed at all. This property is meant to reduce the risk of data loss and corruption. Consistency (as its' name suggests) is a property where every statement/transaction that might alter the table in some way is executed in a predictable and predetermined way. Isolation means that when there are many users are accessing the database at the same time, whether it'd be reading or writing, the property of isolation will *isolate* every transaction of those users to prevent one statement affecting the other statement and to prevent collision. This process is supported by what is called the **DBMS Locking Mechanism**. This process will make it seem like the process is happening all at the same time, whereas in reality, the processes are *isolated* from one another. Lastly, durability means that every transaction that occurs within the database is recorded thoroughly and saved in the database's logs. Every query, failures, errors, etc are also recorded. This process ensures the stability and integrity of the data while also keeping a backup of the data.

Chapter 3 – Data Description and Anomalies

The dataset that we chose is still presented in one of the most primitive and inefficient data storing methods, which is univerasally called the UNF (Unnormalized Form). In its simplest form, a dataset is prone to have many anomalies and errors that are caused by many factors; but as each normalization step is performed, the data will become much more organized, secure, and less dependent.

Samples and Description

By using the following query, we can extract 20 random rows from the dataset.

```
SELECT *
FROM health_facility_general_information_1
ORDER BY RAND()
LIMIT 20;
```

Here is the dataset produced from said query:

facility_id	facility_name	short_description	description	...	ownership_type	facility_latitude	facility_longitude
1	Albany Medical Center Hospital	HOSP	Hospital		Not for Profit Corporation	42.65337	-73.773834
2	Albany Medical Center - South Clinical Campus	HOSP	Hospital		Not for Profit Corporation	42.645485	-73.77829
12	Oswego Hospital - Alvin L Krakau Comm Mtl Health Center Div	HOSP	Hospital		Not for Profit Corporation	43.450497	-76.490562
183	Department of Behavioral and Community Health	DTC	Diagnostic and Treatment Center		County	41.703957	-73.922379
2965	Nascentia Health at Home	LTHHCP	Long Term Home Health Care Program		Not for Profit Corporation	43.053761	-76.172157
2966	Visiting Nurse Service of New York Home Care	LTHHCP	Long Term Home Health Care Program		Not for Profit Corporation	40.747776	-73.988197
2971	Upper Hudson Planned Parenthood, Inc.	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.258457	-73.7658
2974	Leroy Village Green Residential Health Care Facility, Inc	NH	Residential Health Care Facility - SNF		Business Corporation	42.971077	-77.994492
2975	Center for Comprehensive Health Practice Inc	DTC	Diagnostic and Treatment Center		Not for Profit Corporation	40.785938	-73.945763
2994	Marathon Rural Health Center Clinic	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.443099	-76.029541
2995	Cincinnatus Rural Health Center Clinic	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.542934	-75.895599
3002	Island Rehabilitative Services Corp	DTC	Diagnostic and Treatment Center		Business Corporation	40.5826	-74.084587
3012	Buffalo Center for Rehabilitation and Nursing	NH	Residential Health Care Facility - SNF		LLC	42.911777	-78.870277
3014	Family Planning of South Central New York, Inc. at Norwich	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.534122	-75.524529
184	Planned Parenthood of the Mid-Hudson Valley, Inc.	DTC	Diagnostic and Treatment Center		Not for Profit Corporation	41.701019	-73.929726
3019	Little Sisters of the Assumption Family Health	CHHA	Certified Home Health Agency		Not for Profit Corporation	40.795704	-73.936668
3024	Samaritan Health Service Ellenville	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	41.744904	-74.483742
3032	Royal Care Certified Home Health Care, LLC	CHHA	Certified Home Health Agency		LLC	40.770457	-73.836388
3041	Good Samaritan Nursing Home	NH	Residential Health Care Facility - SNF		Not for Profit Corporation	40.725567	-73.079498
3044	Brooklyn Plaza Medical Center	DTC	Diagnostic and Treatment Center		Not for Profit Corporation	40.687092	-73.97654

*Because the dataset is too large, you can find the dataset from the query and the steps of Normalization in the next chapter in the following link: AoL_Kelompok1.xlsx (must use Bina Nusantara account)

This dataset, which is still in its' most primitive form, is prone to have many anomalies when a query is implemented into its' database. Some of these anomalies may have come from technical anomalies such as **insertion anomalies, update anomalies, deletion anomalies, to relational inconsistencies and redundancies.**

1. **Insertion Anomaly**
Insertion anomaly is an anomaly in Database that happens when an insertion of data into a column of the database is also coupled with the insertion of other unrelated data in another column.

In this case of database for example, an insertion anomaly may occur when we are trying to insert a new facility to the database, but we may also be obligated to insert into an unrelated column of say, regional office which may or may not be related to the records of the facility.

2. **Update Anomaly**
An update anomaly is an anomaly in Database where problems might happen in the Database after we insert some values into a certain column in Database. This anomaly can lead to a more serious problem of data inconsistencies, or worse, discrepancies of data.

In this case specifically, an Update Anomaly may occur when we want to update the address of an Operator. Because an operator may have the same name as another operator, changing the address of one row of operators might not lead to the changing of the address of the other rows with the same name.

3. **Deletion Anomaly**
A deletion anomaly may occur in a primitive database when the deletion of certain data will result in an intentional loss of another value/column that might be necessary. This can lead to the loss of data in our Database which is not favorable for the entire team using the Database, especially for the client (if any).

In our case, if we remove data from a certain column, say from the facility_id column, there is no guarantee that the values of other related data are also deleted. This can produce a weird and unfavorable result of having an awkwardly empty rows in certain rows of our database.

4. **Redundancy and Relational Inconsistency**
A data redundancy, as its' name suggests, is an anomaly that might happen where a value in a column is shown/inserted more than one time that might lead to inconsistencies among the data set. On the other hand, relational inconsistencies happen where the relationship of certain columns are not defined in the first place, leading to a database that is not relational at all that could lead us further into other anomalies previously mentioned.

All the previously mentioned anomalies might happen to our dataset when we are implementing certain queries. To combat/prevent this, we can reform the dataset with certain techniques. One of these techniques is the **normalization** of our database (which will be presented in the next chapter). Normalization of a database is a method to reform database to

reduce redundancy and to ensure data integrity where the data will be well-maintained. Another techinque is to implement the use of Integrity Constraints and to determine the keys of our database (which is also done in the conversion from UNF to 1NF). All-in-all, the use of **Database Normalization** is the most effective and versatile way of ensuring that the possibility of previously mentioned anomalies is kept to a minimum.

Chapter 4 – Database Normalization

For reference, here is the link to the table complete with it's Normalization steps:
AOL – Kelompok 1

There are 3 main steps in common way of doing database normalization. Each of those steps will produce a database that is more normalized from the previous step. These databases are called normal forms, these forms range from 1NF, 2NF, to 3NF. Obviously, there are other normal forms like BCNF (Boyce-Codd Normal Forms) etc; but generally, a database in 3NF is enough. A database in a larger normal form is guaranteed to have the properties of the previous forms. For example, a database that is already in 2NF has the properties of 1NF database and a 3NF database also has the properties of both 2NF and 1NF databases.

There are certain rules that must be fulfilled in each step of normalization to transform said database into their respective normal forms. Here is a brief description of each step and their rules.

UNF → 1NF
To transform a default (UNF) table/database to it's 1NF, a database must:

1. Have an atomic value for in each field. If a field contains an array, collection, or groups of values, they must be divided into different rows or tables depending on the case.
2. Remove every field that is obtained by calculations. For example, if a field contains a value which was obtained by summing up a column (e.g., Price, thus Total Price), that sum field must be removed.
3. Have a key decided (can be primary key or composite key).

1NF → 2NF
To transform a 1NF table to it's 2NF, a database must:

1. Be in 1NF first.
2. Remove any partial dependencies between each column which may lead to some anomalies.

Partial dependency is defined as: (A, B) → C and A → C. In other words, partial dependency is where a whole key defines another column, but part of the key still defines that same column. This must be removed in 2NF such that only (A, B) → C; there shouldn't be a case where A → C or B → C.

2NF → 3NF
To transform a 2NF table to it's 3NF, a database must:

1. Be in 2NF first.
2. Remove any transitive dependencies between each column which may lead to some anomalies.

Transitive dependency is defined as: A → B and B → C. In other words, transitive dependency is where column A defines column B, but B also defines column C. This must be removed in 3NF such that if A → B, B → Ø.

1NF Form

Here is the result of converting from UNF → 1NF for the database.

facility_id	facility_name	short_description	description	...	ownership_type	facility_latitude	facility_longitude
1	Albany Medical Center Hospital	HOSP	Hospital		Not for Profit Corporation	42.65337	-73.773834
2	Albany Medical Center - South Clinical Campus	HOSP	Hospital		Not for Profit Corporation	42.645485	-73.77829
12	Oswego Hospital - Alvin L Krakau Comm Mtl Health Center Div	HOSP	Hospital		Not for Profit Corporation	43.450497	-76.490562
183	Department of Behavioral and Community Health	DTC	Diagnostic and Treatment Center		County	41.703957	-73.922379
2965	Nascentia Health at Home	LTHHCP	Long Term Home Health Care Program		Not for Profit Corporation	43.053761	-76.172157
2966	Visiting Nurse Service of New York Home Care	LTHHCP	Long Term Home Health Care Program		Not for Profit Corporation	40.747776	-73.988197
2971	Upper Hudson Planned Parenthood, Inc.	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.258457	-73.7658
2974	Leroy Village Green Residential Health Care Facility, Inc	NH	Residential Health Care Facility - SNF		Business Corporation	42.971077	-77.994492
2975	Center for Comprehensive Health Practice Inc	DTC	Diagnostic and Treatment Center		Not for Profit Corporation	40.785938	-73.945763
2994	Marathon Rural Health Center Clinic	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.443099	-76.029541
2995	Cincinnatus Rural Health Center Clinic	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.542934	-75.895599
3002	Island Rehabilitative Services Corp	DTC	Diagnostic and Treatment Center		Business Corporation	40.5826	-74.084587
3012	Buffalo Center for Rehabilitation and Nursing	NH	Residential Health Care Facility - SNF		LLC	42.911777	-78.870277
3014	Family Planning of South Central New York, Inc. at Norwich	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.534122	-75.524529
184	Planned Parenthood of the Mid-Hudson Valley, Inc.	DTC	Diagnostic and Treatment Center		Not for Profit Corporation	41.701019	-73.929726
3019	Little Sisters of the Assumption Family Health	CHHA	Certified Home Health Agency		Not for Profit Corporation	40.795704	-73.936668
3024	Samaritan Health Service Ellenville	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	41.744904	-74.483742
3032	Royal Care Certified Home Health Care, LLC	CHHA	Certified Home Health Agency		LLC	40.770457	-73.836388
3041	Good Samaritan Nursing Home	NH	Residential Health Care Facility - SNF		Not for Profit Corporation	40.725567	-73.079498
3044	Brooklyn Plaza Medical Center	DTC	Diagnostic and Treatment Center		Not for Profit Corporation	40.687092	-73.97654

Logical Schema of the 1NF table:

Facility(<u>facility_id, facility_name</u>, short_description, description, facility_open_date, facility_address_1, facility_city, facility_state, facility_zip_code, facility_phone_number, facility_fax_number, facility_county_code, facility_county, regional_office_id, regional_office, main_site_name, main_site_facility_id, operating_certificate_number, operator_name, operator_address_1, operator_city, operator_state, operator_zip_code, cooperator_name, cooperator_address, cooperator_city, cooperator_state, cooperator_zip_code, ownership_type, facility_latitude, facility_longitude)

From here onwards, the logical schema is shown such that <u>Underlined Words</u> refers to the Primary Key of the table; Words with asterisks (*) refers to the Foreign Key of the table.

In the 1NF table above, *facility_location* and facility_*location_2* table is removed because it is obtained from concatenating *facility_latitude* and *facility_longitude;* hence it consists

of/obtained from another column nor it's atomic. In the 1NF tables, the column *facility_id* and *facility_name* serves as primary key for the whole table.

2NF Form

Here is the result of converting from 1NF → 2NF for the database.

Facility

facility_id	facility_name	short_description	description	...	ownership_type	facility_latitude	facility_longitude
1	Albany Medical Center Hospital	HOSP	Hospital		Not for Profit Corporation	42.65337	-73.773834
2	Albany Medical Center - South Clinical Campus	HOSP	Hospital		Not for Profit Corporation	42.645485	-73.77829
12	Oswego Hospital - Alvin L Krakau Comm Mtl Health Center Div	HOSP	Hospital		Not for Profit Corporation	43.450497	-76.490562
183	Department of Behavioral and Community Health	DTC	Diagnostic and Treatment Center		County	41.703957	-73.922379
2965	Nascentia Health at Home	LTHHCP	Long Term Home Health Care Program		Not for Profit Corporation	43.053761	-76.172157
2966	Visiting Nurse Service of New York Home Care	LTHHCP	Long Term Home Health Care Program		Not for Profit Corporation	40.747776	-73.988197
2971	Upper Hudson Planned Parenthood, Inc.	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.258457	-73.7658
2974	Leroy Village Green Residential Health Care Facility, Inc	NH	Residential Health Care Facility - SNF		Business Corporation	42.971077	-77.994492
2975	Center for Comprehensive Health Practice Inc	DTC	Diagnostic and Treatment Center		Not for Profit Corporation	40.785938	-73.945763
2994	Marathon Rural Health Center Clinic	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.443099	-76.029541
2995	Cincinnatus Rural Health Center Clinic	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.542934	-75.895599
3002	Island Rehabilitative Services Corp	DTC	Diagnostic and Treatment Center		Business Corporation	40.5826	-74.084587
3012	Buffalo Center for Rehabilitation and Nursing	NH	Residential Health Care Facility - SNF		LLC	42.911777	-78.870277
3014	Family Planning of South Central New York, Inc. at Norwich	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	42.534122	-75.524529
184	Planned Parenthood of the Mid-Hudson Valley, Inc.	DTC	Diagnostic and Treatment Center		Not for Profit Corporation	41.701019	-73.929726
3019	Little Sisters of the Assumption Family Health	CHHA	Certified Home Health Agency		Not for Profit Corporation	40.795704	-73.936668
3024	Samaritan Health Service Ellenville	DTC-EC	Diagnostic and Treatment Center Extension Clinic		Not for Profit Corporation	41.744904	-74.483742
3032	Royal Care Certified Home Health Care, LLC	CHHA	Certified Home Health Agency		LLC	40.770457	-73.836388
3041	Good Samaritan Nursing Home	NH	Residential Health Care Facility - SNF		Not for Profit Corporation	40.725567	-73.079498
3044	Brooklyn Plaza Medical Center	DTC	Diagnostic and Treatment Center		Not for Profit Corporation	40.687092	-73.97654

Logical Schema of 2NF table:

Facility(facility_id, facility_name, short_description, description, facility_open_date, facility_address_1, facility_city, facility_state, facility_zip_code, facility_phone_number, facility_fax_number, facility_county_code, facility_county, regional_office_id, regional_office, main_site_name, main_site_facility_id, operating_certificate_number, operator_name, operator_address_1, operator_city, operator_state, operator_zip_code, cooperator_name, cooperator_address, cooperator_city, cooperator_state, cooperator_zip_code, ownership_type, facility_latitude, facility_longitude)

The 2NF of the table is unexpectedly the same as the previous normal form of the table (1NF) which by default is already 2NF because it doesn't have any partial dependancies to the primary key.

3NF Form

Here is the result of converting from 2NF → 3NF for the database.

Facility

facility_id	facility_name	short_description	…	cooperator_name	ownership_type	facility_latitude	facility_longitude
1	Albany Medical Center Hospital	HOSP		Albany Medical Center	Not for Profit Corporation	42,65337	-73,773834
2	Albany Medical Center - South Clinical Campus	HOSP		Albany Medical Center	Not for Profit Corporation	42,645485	-73,77829
12	Oswego Hospital - Alvin L Krakau Comm Mtl Health Center Div	HOSP			Not for Profit Corporation	43,450497	-76,490562
183	Department of Behavioral and Community Health	DTC			County	41,703957	-73,922379
2965	Nascentia Health at Home	LTHHCP			Not for Profit Corporation	43,053761	-76,172157
2966	Visiting Nurse Service of New York Home Care	LTHHCP			Not for Profit Corporation	40,747776	-73,988197
2971	Upper Hudson Planned Parenthood, Inc.	DTC-EC			Not for Profit Corporation	42,258457	-73,7658
2974	Leroy Village Green Residential Health Care Facility, Inc	NH			Business Corporation	42,971077	-77,994492
2975	Center for Comprehensive Health Practice Inc	DTC			Not for Profit Corporation	40,785938	-73,945763
2994	Marathon Rural Health Center Clinic	DTC-EC			Not for Profit Corporation	42,443099	-76,029541
2995	Cincinnatus Rural Health Center Clinic	DTC-EC	…		Not for Profit Corporation	42,542934	-75,895599
3002	Island Rehabilitative Services Corp	DTC			Business Corporation	40,5826	-74,084587
3012	Buffalo Center for Rehabilitation and Nursing	NH			LLC	42,911777	-78,870277
3014	Family Planning of South Central New York, Inc. at Norwich	DTC-EC			Not for Profit Corporation	42,534122	-75,524529
184	Planned Parenthood of the Mid-Hudson Valley, Inc.	DTC			Not for Profit Corporation	41,701019	-73,929726
3019	Little Sisters of the Assumption Family Health	CHHA			Not for Profit Corporation	40,795704	-73,936668
3024	Samaritan Health Service Ellenville	DTC-EC			Not for Profit Corporation	41,744904	-74,483742
3032	Royal Care Certified Home Health Care, LLC	CHHA			LLC	40,770457	-73,836388
3041	Good Samaritan Nursing Home	NH		Catholic Health System of Long Island, Inc.	Not for Profit Corporation	40,725567	-73,079498
3044	Brooklyn Plaza Medical Center	DTC			Not for Profit Corporation	40,687092	-73,97654

Descriptions

short_description	description
CHHA	Certified Home Health Agency
DTC	Diagnostic and Treatment Center
DTC-EC	Diagnostic and Treatment Center Extension Clinic
HOSP	Hospital
LTHHCP	Long Term Home Health Care Program
NH	Residential Health Care Facility - SNF

Facility County

facility_county_code	facility_county
1	Albany
8	Chenango
10	Columbia
11	Cortland
13	Dutchess
14	Erie
18	Genesee
33	Onondaga
37	Oswego
51	Suffolk
55	Ulster
7093	New York
7095	Kings
7096	Queens
7097	Richmond

Regional Office

regional_office_id	regional_office
1	Western Regional Office - Buffalo
3	Central New York Regional Office
4	Capital District Regional Office
5	Metropolitan Area Regional Office - New York City
6	Metropolitan Area Regional Office - New Rochelle
7	Metropolitan Area Regional Office - Long Island

Main Site

main_site_facility_id	main_site_name
1	Albany Medical Center Hospital
727	Oswego Hospital
10	Upper Hudson Planned Parenthood Inc
4963	Family Health Network of Central New York Inc
742	Family Planning of South Central New York, Inc.
6067	Damian Family Care Center

Operator

operating_certificate_number	operator_name	operator_address_1	operator_city	operator_state	operator_zip_code
0101000H	Albany Medical Center Hospital	New Scotland Avenue	Albany	New York	12208
0101204R	Upper Hudson Planned Parenthood Inc	855 Central Avenue	Albany	New York	12210
1101201R	Family Health Network of Central New York Inc	11 Avena Avenue	Cortland	New York	13045
1302201R	Department of Behavioral and Community Health	29 North Hamilton Street	Poughkeepsie	New York	12601

1302207R	Planned Parenthood of the Mid-Hudson Valley, Inc.	85 Market Street	Poughkeepsie	New York	12601
1401341N	Delaware Operations Associates LLC	1014 Delaware Avenue	Buffalo	New York	14209
1823300N	Leroy Village Green Residential Health Care Facility, Inc	10 Munson Street	Leroy	New York	14482
3301902L	Visiting Nurse Association of Central New York Inc	1050 West Genesee Street	Syracuse	New York	13204
3702000H	Oswego Hospital Inc	110 West Sixth Street	Oswego	New York	13126
3801202R	Family Planning of South Central New York, Inc.	37 Dietz Street	Oneonta	New York	13820
5154310N	Good Samaritan Hospital Medical Center	1000 Montauk Highway	West Islip	New York	11795
7001250R	Brooklyn Plaza Medical Center	50 Greene Avenue	Brooklyn	New York	11238
7002134R	Center for Comprehensive Health Practice Inc	163 East 97th Street	New York	New York	10029
7002645	Little Sisters of the Assumption Family Health	426 E 119th St	New York	New York	10035
7002911L	Visiting Nurse Service of New York Home Care II	220 E 42nd Street	New York	New York	10017
7003246R	Project Samaritan Health Services Inc	137-50 Jamaica Avenue	Jamaica	New York	11435
7003618	Royal Care Certified Home Health Care, LLC	6323 14th Avenue	Brooklyn	New York	11219
7004204R	Island Rehabilitative Services Corp	470 Seaview Avenue	Staten Island	New York	10305

Cooperator

cooperator_name	cooperator_address	cooperator_city	cooperator_state	cooperator_zip_code
Albany Medical Center	New Scotland Avenue	Albany	New York	12208
Catholic Health System of Long Island, Inc.	1 Huntington Quadrangle	Melville	New York	11747

Logical Schema of the 3NF table:

Facility(facility_id, facility_name, *short_description, description, facility_open_date, facility_address_1, facility_city, facility_state, facility_zip_code, facility_phone_number, facility_fax_number, *facility_county_code, facility_county, *regional_office_id, regional_office, main_site_name, *main_site_facility_id, *operating_certificate_number, operator_name, operator_address_1, operator_city, operator_state, operator_zip_code, *cooperator_name, cooperator_address, cooperator_city, cooperator_state, cooperator_zip_code, ownership_type, facility_latitude, facility_longitude)

Descriptions(short_description, description)

Facility_county(facility_county_code, facility_county)

Regional Office(regional_office_id, regional_office)

Main Site(main_site_facility_id, main_site_name)

Operator(operating_certificate_number, operator_name, operator_address_1, operator_city, operator_state, operator_zip_code)

Cooperator(cooperator_name, cooperator_address, cooperator_city, cooperator_state, cooperator_zip_code)

As we can see, there is a significant difference between the steps before and the 3NF in this case. In the update of 2NF to 3NF, we remove all the transitive dependencies of the table, resulting in the separation of tables that defines the dependancy between other tables. The

database is split into Facility, Descriptions, Facility County, Regional Office, Main Site, Operator, and Cooperator. Because in the main table each category (Regional Office, Facility County, etc) is dependent on the Primary Key of the table (*facility_id* & *facility_name*) but they also have their own properties (For example the Main Site category has *main_site_facility_id* and *main_site_name*), it can be said that they have a transitive dependency going on between them. This cannot persist in 3NF, hence the splitting of the tables.

Chapter 5 – Conclusion

In this report, which the dataset we've borrowed from data.world and is titled **Health Facility General Information** shows the location of health care facilities, and programs from the HFIS (Health Facilities Information System). This dataset consists of 6,141 rows of data and 36 distinct columns. With the immense size of the dataset, the use of the database is necessary. It ensures the integrity of the data through unique identification codes, preventing duplicates and maintaining consistency. Security and access control is also vital for public datasets representing a state's condition and needs constant and vigilant management by the database administrator. Additionally, the database facilitates seamless data recovery and backup, addressing potential errors or losses. Overall, embracing a database solution enhances the reliability, accessibility, and security of the dataset, aligning with the demands of our evolving information era. Before implementing the dataset into the database system, the dataset must be normalized. The dataset that we chose is still in unnormalized form because it hasn't fulfilled the rules of a 1NF database, hence many errors and anomalies might occur. In this database scenario, inserting a new facility may bring about an insertion anomaly, requiring us to add data to an unrelated column, like regional office, potentially creating confusion in record relationships. Furthermore, updating an operator's address could lead to an update anomaly, especially when operators share the same name – altering one row may not impact others with identical names. Deletion anomalies add another layer of complexity; removing data from the facility_id column doesn't ensure related values are deleted, resulting in awkwardly empty rows. Data redundancy, where a value appears more than once in a column, introduces inconsistencies, while relational inconsistencies arise when certain column relationships are undefined, disrupting the database's relational structure. These challenges highlight the need for effective solutions such as normalization to address anomalies, reduce redundancy, and maintain a well-organized and coherent dataset. So, to prevent those anomalies from happening in the database system we need to normalize the dataset, preferably to one of the highest forms of normalization, 3NF (amongst other also high forms such as BCNF).

In the First Normal Form, we found that one of the rows (*facility_location*) was obtained from the concatination of *facility_longitude* and *facility_lattitude,* thus *facility_location* was removed. Other than that, we did not find any redundant, repeating, or duplicating row and column, hence no other changes were made in the data. Also still in the first normal form, we opted that the key to the table is a composite between (*facility_id, facility_name*) For the Second Normal Form, since the data does not contain any partial dependencies or other composite keys, no changes were made from the First Normalization to the Second. For the Third and last Normal Form, we removed any remaining transitive dependencies on the data so that all fields are dependent only on the primary key. To achieve the Third Normal Form, we separated the tables for facilities, descriptions, regional office, facility county, main site, operator, and cooperator. To sum it up, by carefully organizing the dataset to follow certain rules, we've fixed issues like mistakes and unnecessary repeats. This makes the health facility information more solid and reliable when put into a database. It helps prevent problems when adding, changing, or deleting data. Now, the dataset is well-organized, making it easier to access and more secure, meeting the needs of today's information world while also fulfilling the ACID (Atomicity, Consistency, Isolation, Durability) concept of Database itself.