Caitlin Atkins
D597 – Data Management
August 16, 2024

Task 1: Relational Database Design and Implementation

Part 1: Design Document

A. Select one of the provided scenarios to complete the following:

Scenario 1

1. **Describe a business problem that can be solved with a database solution and aligns with the chosen scenario.**

HealthFit Innovations is a new company with hopes to scale out and introduce new users to the product. Because the company is so new and needs more specificity in the business data, they might need to understand their current demographics fully.

By referring to the business data, the company will be able to determine if there are any relevant relationships. There will be no opportunity to determine if there are commonalities in the preferred choice of fitness watch to current medical conditions, medications, frequency of doctor visits, allergies, age, and gender.

When this problem is presented, HealthFit Innovations will need help marketing to specific demographics and preparing the databases for scalability.

2. **Propose a data structure to solve the identified business problem.**

The Entity Relationship Diagram (ERD) is a great data structure that can be used to help solve existing business problems.

The ERD will give HealthFit Innovations the ability to identify current relationships that may exist between the two sets of business data. The data structure can also offer flexibility in developing new databases as the company introduces new attributes and users to the program (Coronel, C., & Morris, S., 2023).

With the ERD's abilities, the company will be able to learn about the marketable demographics and prepare as the company expands simultaneously.

3. **Justify why a database solution will solve the identified business problem.**

The business's lack of demographical marketing knowledge can be solved using ERDs, PostgreSQL, and pgAdmin.

By using these programs, HealthFit Innovations can correctly identify target demographics, preferences in fitness trackers, and customer satisfaction.

PostgreSQL and pgAdmin are open source, allowing anyone to access the document quickly, especially since these programs are designed for single and multiple users. The data in these databases are distributed and centrally located (Coronel, C., & Morris, S., 2023). These are all benefits of using PostgreSQL for the business.

Beyond this project's scope, these programs could be used continuously throughout the marketing study to measure user engagement, marketing effectiveness, and other relevant queries.

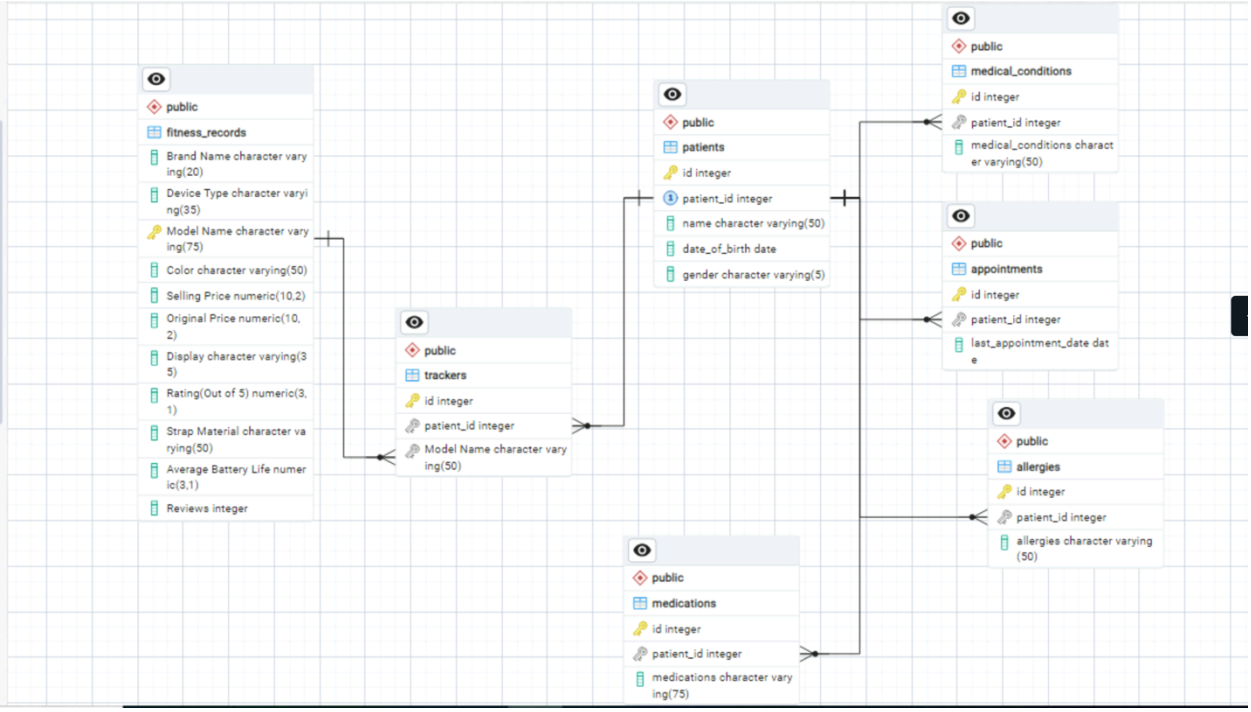
4. Explain how the database solution will use the business data.

Business data will initially be used during the development of tables in pgAdmin. The data will be loaded into the tables and then used to develop an ERD.

Once completed, the relationships between the business data tables will be identified and reviewed. Based on the existing relationships, queries can be formed and used to help solve the business problem.

Queries that may relate to the patient's identifying qualities, the type of fitness trackers chosen, the fitness tracker rating based on the brand, and existing medical conditions, medications, and allergies.

B. Create a logical data model for storing data in the database solution.



C. Describe the database objects and storage, identifying the file attributes within the database solution.

There are seven tables in total. These are fitness records, trackers, patients, medications, medical conditions, appointments, and allergies.

Fitness_records have 11 columns titled Brand Name, Device Type, Model Name, Color, Selling Price, Original Price, Display, Rating (Out of 5), Strap Material, Average Battery Life, and Reviews. Model Name, Brand Name, Device Type, Color, Display, and Strap Material are all VARCHAR datatypes. Selling Price, Original Price, Rating (Out of 5), and Average Battery Life are DECIMAL datatypes. Finally, Reviews is an INT datatype. The primary key in this table is the Model Name. Model Name connects to the foreign key in "trackers."

Trackers have an ID as an INT, patient_id as an INT, and Model Name as VARCHAR. The primary key here is id, which is a surrogate key. Surrogate keys are unique identifiers that give all tables a primary key. Beyond providing a unique identification in the table, surrogate keys play no other roles. The tracker's foreign key is patient_id and connects to the Patient table.

Patients hold ID (INT), patient_id (INT), name (VARCHAR), date_of_birth (DATE), and gender (VARCHAR). ID is the primary key, and patient_id is the unique key. Patient_id is the foreign key in medications, medical_conditions, appointments, and allergies. Additionally, the primary keys in these tables are IDs.

Medications contain ID (INT), patient_id (INT), and medications (VARCHAR). Medical_conditions contain ID (INT), patient_id (INT), and medical_conditions (VARCHAR). Appointments contain ID (INT), patient_id (INT), and last_appointment_date (DATE). Finally, allergies contain ID (INT), patient_id (INT), and allergies (VARCHAR).

D. Discuss how the proposed database design addresses scalability concerns, including strategies that align with the chosen scenario.

ERDs are beneficial for HealthFit Innovations because they fully address scalability concerns. As previously mentioned, ERDs offer a chance to expand as the data entered expands. This is done through table divisions and partitions. As data is entered and new relationships, foreign, and primary keys are created, new tables can be designed to make the information more manageable.

Other potential strategies that can be used with ERDs include normalization and denormalization. These techniques reduce data redundancies that might make the data analysis difficult (Coronel, C., & Morris, S., 2023).

Beyond that, horizontal and vertical scaling can be used to ensure that the actual programs can handle the expansion of data. The database can be shared across multiple services, and if needed, the hardware used may need to be updated to handle large amounts of data.

E. Outline the privacy and security measures that should be implemented in the proposed database design.

Because the information handled for HealthFit Innovations contains users' personal health-related information, it is essential to understand the privacy and security measures implemented in the database.

First, limited access must be implemented. Only those who need to handle the data should have complete access. Beyond that, only the minimum amount of data should be provided so that the employees at the company can complete their jobs. In addition, the data should be handled within a specific timeframe (Segalla, M., & Rouziès, D., 2023).

Data anonymity techniques should be used when shared with the company or third parties. By doing this, the information cannot be used to identify the individual users (Segalla, M., & Rouziès, D., 2023). If this is not possible, pseudonyms can be used, making connecting the information with the users harder.

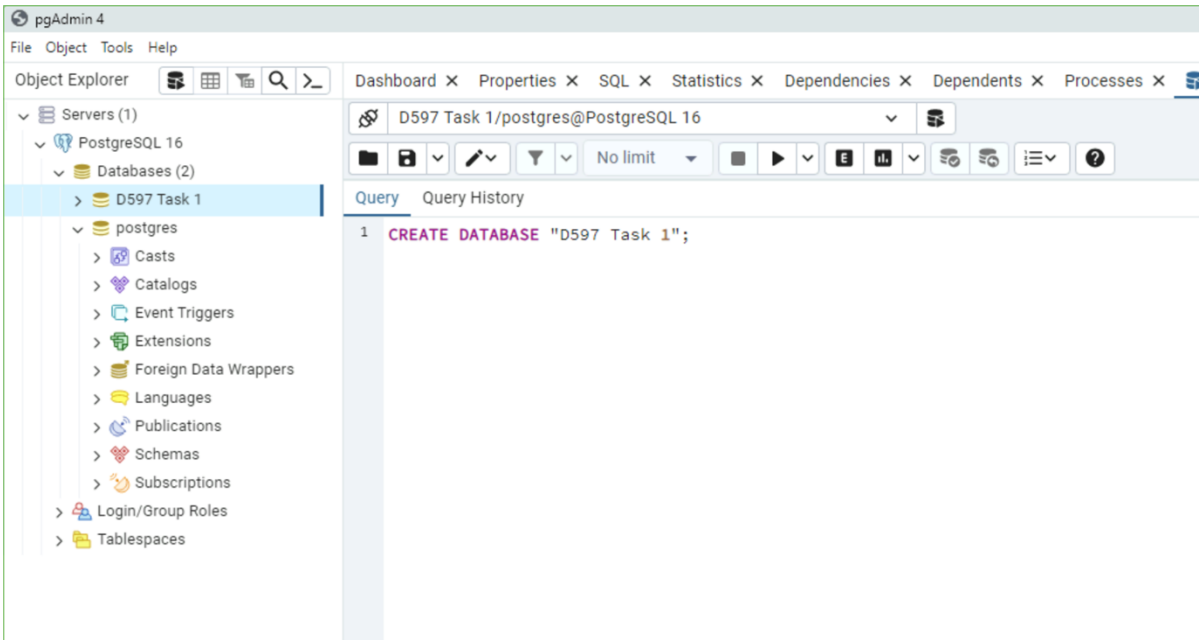
Part 2: Implementation

Note: The data files for each scenario are located in a folder titled "D597 Datasets" on the desktop of the WGU Virtual Lab environment. Pull the files from "Task 1" related to your chosen scenario.

Note: Submit screenshots showing successfully run queries from the WGU Virtual Lab for each prompt with your design document.

F. Implement the proposed database design in the WGU Virtual Lab environment by completing the following:

1. **Write a script to create a database instance named "D597 Task 1" using the appropriate query language based on the logical data model in part B. Provide a screenshot showing the script and the database instance in the platform.**



2. **Write a script to import the data records from the chosen scenario CSV files into the database instance. Provide a screenshot showing the script and the data correctly inserted or mapped into the database.**

```sql

CREATE TABLE fitness_records (
    "Brand Name" varchar(20),
    "Device Type" varchar(35),
    "Model Name" varchar(75),
    "Color" varchar(50),
    "Selling Price" varchar(10),
    "Original Price" varchar(10),
    "Display" varchar(35),
    "Rating(Out of 5)" DECIMAL(3,1),
    "Strap Material" varchar(50),
    "Average Battery Life(in days)" int,
    "Reviews" int
);


SELECT * FROM fitness_records

```

Data Output | Messages | Notifications

	Brand Name character varying (20)	Device Type character varying (35)	Model Name character varying (75)	Color character varying (50)
1	Xiaomi	FitnessBand	Smart Band 5	Black
2	Xiaomi	FitnessBand	Smart Band 4	Black
3	Xiaomi	FitnessBand	HMSH01GE	Black
4	Xiaomi	FitnessBand	Smart Band 5	Black
5	Xiaomi	FitnessBand	Band 3	Black
6	Xiaomi	FitnessBand	Band - HRX Edition	Black
7	Xiaomi	FitnessBand	Band 2	Black
8	Xiaomi	Smartwatch	Revolve	Black
9	Xiaomi	Smartwatch	RevolveActive	Black
10	Xiaomi	FitnessBand	Smart Band 3i	Black
11	OnePlus	FitnessBand	Steven Harrington Edition Band	Blue

Total rows: 565 of 565 | Query complete 00:00:00.581 | Ln 1, Col 15

```sql


CREATE TABLE medical_records(
    patient_id int,
    name varchar(50),
    date_of_birth text,
    gender varchar(5),
    medical_conditions varchar(50),
    medications varchar(75),
    allergies varchar(50),
    last_appointment_date text,
    tracker varchar(50)
);


SELECT * FROM medical_records

```

Data Output | Messages | Notifications

	patient_id integer	name character varying (50)	date_of_birth date	gender character varying (5)	medical_conditions character varying (50)	medications character varying (75)	allergies character varying
1	1	Scott Webb	1967-04-28	F	None	No	None
2	2	Rachel Frederick	1977-04-04	M	None	No	None
3	3	Eric Kline	1926-05-18	F	Watch	Yes	None
4	4	James Rodriguez	1954-07-20	M	None	No	None
5	5	David Scott	2015-12-07	M	Mild	Yes	None
6	6	Dawn Roach	1967-06-02	M	Mild	Yes	None
7	7	Mary Harris	2013-08-27	M	Mild	Yes	None
8	8	Sandy Brown	1927-06-23	F	Watch	Yes	None
9	9	Kenneth Johnson	2011-04-15	F	Mild	Yes	None
10	10	Nicole Henderson	1978-10-14	F	Mild	Yes	None
11	11	Brooke Nolan	2020-05-18	F	Mild	Yes	None
12	12	Kristen Weaver	1931-10-10	F	Watch	Yes	None
13	13	Brian Brown	2019-05-15	F	Mild	Yes	None
14	14	Reginald Perez	1998-05-26	M	Mild	Yes	None
15	15	Laura Lopez	1934-10-06	F	Watch	Yes	None

Total rows: 1000 of 100000 | Query complete 00:00:00.518 | Ln 1, Col 1

3. Write a script for three **queries to retrieve specific information from the database to help solve the identified business problem. Provide a screenshot showing the script for *each* query successfully executed.**

```
139 --TOP SELLING PRODUCTS
140 SELECT "Brand Name", "Model Name",
141 COUNT(*) AS total_sales
142 FROM fitness_records
143 GROUP BY "Brand Name", "Model Name"
144 ORDER BY total_sales DESC;
145
```

Data Output | Messages | Notifications

	Brand Name character varying (20)	Model Name character varying (75)	total_sales bigint
1	FitBit	Charge 2	8
2	FOSSIL	Gen 5 Julianna HR	6
3	FOSSIL	Jacqueline Hybrid	5
4	huami	Amazfit Bip	5
5	GARMIN	Forerunner	4
6	FitBit	Versa	4
7	FitBit	Surge	4
8	SAMSUNG	Galaxy Classic 4	4
9	FOSSIL	Sport	4
10	huami	Amazfit GTS	4
11	FOSSIL	Marshall	4
12	huami	Amazfit Verge	4

```sql
--CUSTOMER TRACKING PREFERENCE
SELECT fitness_records."Brand Name",
COUNT(*) AS count
FROM trackers
JOIN patients ON trackers.patient_id = patients.patient_id
JOIN fitness_records ON trackers."Model Name" = fitness_records."Model Name"
JOIN medical_conditions ON patients.patient_id = medical_conditions.patient_id
WHERE medical_conditions IS NOT NULL
GROUP BY fitness_records."Brand Name"
ORDER BY count DESC;
```

Data Output | Messages | Notifications

	Brand Name character varying (20)	count bigint
1	huami	99377
2	Honor	66248
3	realme	9287
4	boAt	6970
5	Huawei	6957
6	Infinix	6956
7	Xiaomi	4655
8	Oppo	2336

```sql
--PRICE VS. RATING
SELECT "Selling Price", "Rating(Out of 5)"
FROM fitness_records
ORDER BY "Selling Price";
```

Data Output | Messages | Notifications

	Selling Price numeric (10,2)	Rating(Out of 5) numeric (3,1)
1	1195.00	4.1
2	1270.00	4.2
3	1290.00	4.0
4	1299.00	4.2
5	1395.00	4.1
6	1450.00	4.0
7	1499.00	3.8
8	1499.00	4.0
9	1595.00	4.0
10	1645.00	4.0
11	1649.00	3.2
12	1699.00	3.9
13	1699.00	4.1
14	1699.00	3.8

5. Apply optimization techniques to improve the run time of your queries from part F3, providing output results via a screenshot.

```sql
CREATE INDEX idx_patient_id ON trackers(patient_id);
CREATE INDEX idx_selling_price ON fitness_records("Selling Price");
CREATE INDEX idx_model_name ON trackers("Model Name");
```

```sql
--TOP SELLING PRODUCTS
SELECT "Brand Name", "Model Name",
COUNT(*) AS total_sales
FROM fitness_records
GROUP BY "Brand Name", "Model Name"
ORDER BY total_sales DESC;
```

Data Output | Messages | Notifications

```
Successfully run. Total query runtime: 104 msec.
392 rows affected.
```

```sql
--CUSTOMER TRACKING PREFERENCE
SELECT fitness_records."Brand Name",
COUNT(*) AS count
FROM trackers
JOIN patients ON trackers.patient_id = patients.patient_id
JOIN fitness_records ON trackers."Model Name" = fitness_records."Model Name"
JOIN medical_conditions ON patients.patient_id = medical_conditions.patient_id
WHERE medical_conditions IS NOT NULL
GROUP BY fitness_records."Brand Name"
ORDER BY count DESC;
```

Data Output | Messages | Notifications

```
Successfully run. Total query runtime: 483 msec.
8 rows affected.
```

```sql
--PRICE VS. RATING
SELECT "Selling Price", "Rating(Out of 5)"
FROM fitness_records
ORDER BY "Selling Price";
```

Data Output | Messages | Notifications

```
Successfully run. Total query runtime: 121 msec.
565 rows affected.
```

H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

Coronel, C., & Morris, S. (2023). Database Systems. Cengage Learning, Inc. *Database Systems* (14th ed., pp.1–32). https://ebooks.cenreader.com/#!/reader/63a8f021-e709-4941-9e1a-116bb401a808/page/7d2b325ac7a78cb777a9f50093980c83

Coronel, C., & Morris, S. (2023). Entity Relationship (ER) Modeling. Cengage Learning, Inc. *Database Systems* (14th ed., pp. 108-161). https://ebooks.cenreader.com/#!/reader/63a8f021-e709-4941-9e1a-116bb401a808/page/7d2b325ac7a78cb777a9f50093980c83

Coronel, C., & Morris, S. (2023). The Relational Database Model. Cengage Learning, Inc. *Database Systems* (14th ed., pp. 65–107). https://ebooks.cenreader.com/#!/reader/63a8f021-e709-4941-9e1a-116bb401a808/page/7d2b325ac7a78cb777a9f50093980c83

Segalla, M., & Rouziès, D. (2023). The Ethics of Managing People's Data. Harvard Business Review, 101(4), 86–94.