Luis Medina

012020203

Data Management – D597

Task 1: Relational Database Design and Implementation

**Part 1: Design Document**

**Part A:**

1. Describe a business problem that can be solved with a database solution and is in alignment with the chosen scenario.

HealthFit Innovations has been experiencing some challenges with its increasing volume, variety and velocity of health-related data collected. The way the platform collects data is through various sources that include wearables, electronic health records (EHR’s), medical imaging systems, and patient-reported outcomes. With the current way it is, it lacks the ability to scale and flexibility, which can lead to problems. For example, problems with integration and data processing failures.

1. Purpose a data structure to solve the identified business problem.

To solve the identified business problem, a relation database structure is proposed and modeled through an Entity Relationship Diagram (ERD). By creating an ERD, it will be able to support scalable, secure, and flexible integration of the data collected from the health devices and patient-reported inputs. The database that will be created will include these tables to support the data, fitness records, patients, medical conditions, medications, allergies, appointments, and trackers. Each table will have a primary and most of the tables will have a foreign key to connect them to other tables. By doing this, it will reduce data redundancy, enforce data integrity and support fast querying regardless of the size of the data.

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

Using a database solution will provide HealthFit Innovations with scalable, structured, and integrated solutions for managing lots of data, that will be collected from the wearables. It will also, improve performance and provide support with any analytical work for the future.

1. Explain how the business data will be used within the database solution.

All the data collected from the wearables that the customers are wearing, will have a fast and reliable way to store all the data that is gathered. It will also make easy to access the data using SQL commands to retrieve information that an analyst will need.

**Part B: Data Model**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Part C: Objects and Storage**

The database that was created in the previous section will consist of seven main tables, each defined with a primary key and a foreign key to enforce data integrity and normalization. This database scheme is meant to store structured data that will be collected from health devices and any other form of data collection from HealthFit.

1. Database Tables
   1. Fitness Records
      1. Id (Primary key INT), brand name, device type, model name, color, display, strap material (varchar)
   2. Patients
      1. Patient id (Primary Key), name (varchar), date of birth(Date), gender(Varchar)
   3. Medical Conditions
      1. Id(Primary key, INT), patient id (Foreign key to patients), medical conditions (Text)
   4. Medications
      1. Id (Primary Key, INT), patient Id (Foreign key), medications (Text)
   5. Allergies
      1. Id (Primary Key, INT), patient Id (Foreign key), allergies (Text)
   6. Appointments
      1. Id (Primary Key, INT), patient Id (Foreign key), last appointment date (Date)
   7. Trackers
      1. Id (Primary Key, INT), patient Id (Foreign Key), model name (VARCHAR).
2. File attributes and storage considerations
   1. Data types:
      1. Numeric values (e.g. prices, rating) are stored as Decimal for precision, Text values such as names and model names use VARCHAR, while the longer text use TEXT, then Dates are stored using the DATE type for proper range filters.
   2. Primary key and Indexing
      1. All the tables have a foreign key (id) are going to be indexed automatically, foreign keys are going to be used for enforcing referential integrity and to help speed up the joins on the different tables.
   3. Scalability
      1. Tables appointments and patients could be partitioned by date for improved performance

**Part D: Scalability**

Like I mentioned before that as HealthFit’s user base continues to grow and the value collected from the devices increases, it’s extremely important that the database supports both vertical and horizontal scalability. As, the data continues to flow in from the devices the ERDs offer a great opportunity through the tables created can make the data more manageable, then with primary keys and foreign keys, all the tables will be connected making easy to gather or combine data. Another example is with the tables Health\_metric and activity\_log can be partitioned independently, as they will grow quickly.

**Part E: Privacy and Security**

Privacy and security are going to be one of the most important parts that need to be addressed, as it stores sensitive personal health information. To ensure the confidentiality, integrity, and availability of the data, here are some of the security measures that we will go through to protect the customers data.

First, it will be accessing control, meaning that role-based access control will be used, so that people that should not see the data cannot see it. For example, patients can only access their own records, healthcare providers can only access records of patients under their case and not just be able to look at any record. Administrators will have more access for system maintenance and data audits. There will also be auditing and logging of the data and what is done in the database. An example would be that every interaction with the sensitive tables will be logged to gather information about who accessed the data, how it was modified, and when it happened. Lastly, there will be backups set up for all the data in case of an emergency and some of the data is lost. The backups will be automatically done daily and stored securely. Backup will be tested to ensure that the data can be restored quickly.

**Part 2: Implementation**

**F1: Database Instance**

A screenshot of a computer

AI-generated content may be incorrect.

Now I will create the SQL script to import the data records from the first scenario CSV files into the database.

A screenshot of a computer

AI-generated content may be incorrect.

**F2: Insert Records**

Here we are going to insert the fitness data into the fitness table using the ‘copy’ method. Here is a screenshot of the script to copy the data and then ‘SELECT’ statement to make sure the data was copied into the table correctly.

A black screen with white text

AI-generated content may be incorrect.

Now going to use insert statements to fill in the other tables and here is a screenshot showing some of the insert statements.

A screenshot of a computer program

AI-generated content may be incorrect.

**F3: Queries**

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Here are the 3 scripts that I created for part F3 to help answer the business problem that was stated in the beginning of this document. Gathering information about battery life is extremely helpful to keeping devices running and having no down time. Then gathering information about popular trackers will give us a better idea of what tracker people prefer to get a better idea of what is wanted in the market.

**F4: Optimization**

For this part I am going to apply the optimization techniques I learned from the book provided in the class notes for this class. Then, after I will rerun the scripts from the previous section to see how much the optimization helped with the query speed.

**A screen shot of a computer program

AI-generated content may be incorrect.**

A screenshot of a computer

AI-generated content may be incorrect.

Here is me running the first scrip again and comparing it to the first screenshot the speed slightly increased compared to the first one, now to look at the other two.

**A screenshot of a computer

AI-generated content may be incorrect.**

Now here is the screenshot for the second query and comparing it to the first one there is more of a difference in the speeds. Finally, I will run the third script and compare how they differ.

A screenshot of a computer

AI-generated content may be incorrect.

Looking at the last picture we can see that there is a slight difference in run times between the two. Although, it’s not a big difference we still noticed that there is a difference in run time after applying the optimization techniques.

**H: Sources**

For the sources used it was mostly the course material that was provided and the webinars that I attend throughout the course.

Data Analytics: Perform Fast and Efficient Data Analysis with the Power of SQL was the book used for a lot of the techniques that were needed throughout the project.