Diabetes

Database Schema Design Report

Prepared by

[Team 10 – Mellitus]



Table of Contents

[1. Introduction 3](#_Toc183186437)

[1.1 Background 3](#_Toc183186438)

[1.2 Objectives of the Schema Design 3](#_Toc183186439)

[1.3 Overview of the Schema 3](#_Toc183186440)

[2. Relationships and Diagram 4](#_Toc183186441)

[2.1 E2E High Level Architecture 4](#_Toc183186442)

[2.2 Data Layer Design 4](#_Toc183186443)

[2.3 Curated Data Flow Diagram 5](#_Toc183186444)

[2.4 Star Schema (basic outline) 6](#_Toc183186445)

[2.5 Diagram Representation (Logical Relationships) 7](#_Toc183186446)

[3.0 Tables and Relationships 7](#_Toc183186447)

[3.1. Patient\_Dim Table 7](#_Toc183186448)

[3.2. Health Metrics Dimension Table (Health\_Metrics\_Dim) 7](#_Toc183186449)

[3.3. Lifestyle Dimension Table (Lifestyle\_Dim) 8](#_Toc183186450)

[3.4. Fact Table (DIABETES\_FACT) 8](#_Toc183186451)

[4.0. Justification of Design 9](#_Toc183186452)

[5.0 Conclusion 9](#_Toc183186453)

# Introduction

### 1.1 Background

This report provides a comprehensive explanation of the database schema used for analyzing diabetes-related data. The design follows a **star schema for** optimal performance and simplicity, combining a central fact table and supporting dimension tables. This design facilitates queries, analytical processing, and understanding relationships between patient demographics, health metrics, lifestyle factors, and diabetes indicators. The schema is centered on the DIABETES\_FACT table, supported by dimension tables (Patient\_Dim, Health\_Metrics\_Dim, Lifestyle\_Dim). This design facilitates detailed analysis while ensuring data normalization and clarity.

### 1.2 Objectives of the Schema Design

1. **Support analytical queries** for diabetes prediction and insights.
2. **Ensure data integrity and scalability** through normalization.
3. **Minimize redundancy** while maintaining a user-friendly structure.
4. Enable **efficient joins** between tables using surrogate keys.

### 1.3 Overview of the Schema

The database schema consists of:

* **One Fact Table**: **DIABETES\_FACT** — the central table that captures key relationships and diabetes status.
* **Three Dimension Tables**: **Patient\_Dim, Health\_Metrics\_Dim, Lifestyle\_Dim** — providing descriptive details linked to the fact table.

# Relationships and Diagram

### 2.1 E2E High Level Architecture

A diagram of a data processing process

Description automatically generated

### 2.2 Data Layer Design

The data pipeline design below uses a Staging Layer in Snowflake to store cleaned master data, which is transformed into fact and dimension tables in the Curated Layer via curation stored procedures. The curated data is further processed into Reporting Views for analytical use, leveraging stored procedures to generate Azure-hosted reporting outputs.

A diagram of data storage

Description automatically generated

### 2.3 Curated Data Flow Diagram

A diagram of data warehouse

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Type | Name | Description |
| S1 | System | Snowflake Data Warehouse | Stores all raw/cleaned/curated data for Mellitus (Diabetes Data) |
| S2 | System | Staging Layer | Stores cleaned master dataset |
| S3 | System | Curated Layer | Stores curated tables |
| O1 | Object | DIABETES\_DATA\_CLEANED | Cleaned master dataset |
| O2 | Object | DIABETES\_FACT | Diabetes fact table |
| O3 | Object | HEALTH\_METRICS\_DIM | Health metrics dimension table |
| O4 | Object | PATIENT\_DIM | Patient dimension table |
| O5 | Object | LIFECYCLE\_DIM | Lifecycle dimension table |
| I1 | Interface | Staging to Diabetes Fact | Data load from staging table to fact table |
| I2 | Interface | Staging to Health Metrics Dim | Data load from staging table to dim table |
| I3 | Interface | Staging to Patient Dim | Data load from staging table to dim table |
| I4 | Interface | Staging to Lifecycle Dim | Data load from staging table to dim table |
| I5 | Interface | Health Metrics Dim to Diabetes Fact | Dim table to dim table data transfer |
| I6 | Interface | Patient Dim to Health Metrics Dim | Dim table to dim table data transfer |
| I7 | Interface | Patient Dim to Lifecycle Dim | Dim table to dim table data transfer |
| I8 | Interface | Patient Dim to Diabetes Fact | Dim table to dim table data transfer |
| I9 | Interface | Lifecycle Dim to Diabetes Fact | Dim table to dim table data transfer |

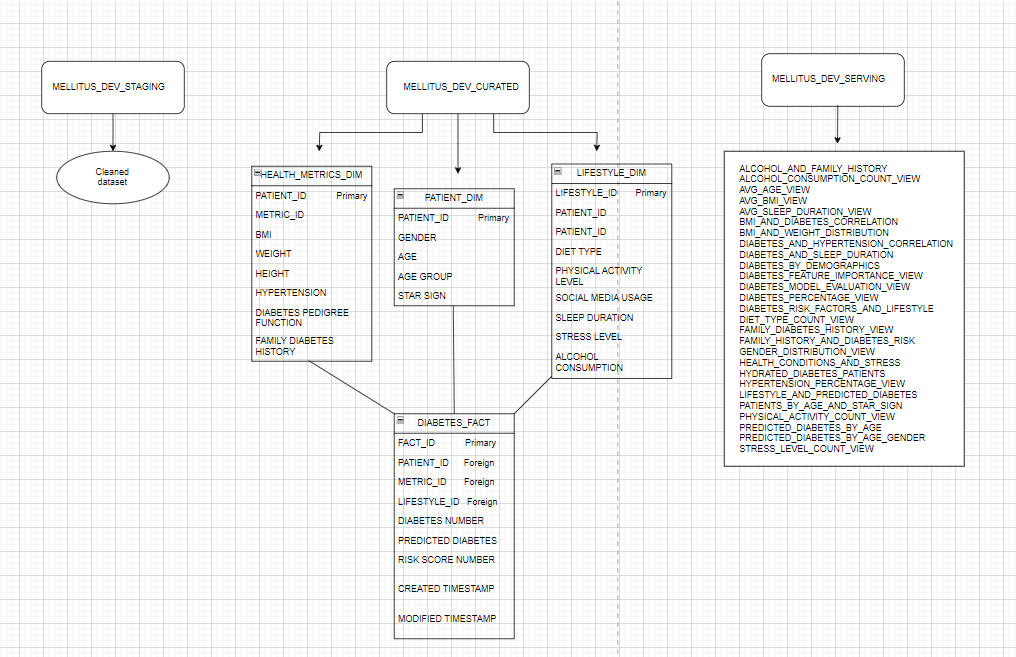
### 2.4 Star Schema (basic outline)

The schema follows the **star schema** model, with DIABETES\_FACT as the central table connected to dimensions. This design supports efficient queries, enabling analytics like calculating average BMI, tracking diabetes prevalence, and exploring lifestyle correlations.



### 2.5 Diagram Representation (Logical Relationships)

**Purpose**: Outlines the relationships between the databases and the tables/views held by them.



# 3.0 Tables and Relationships

### 3.1. Patient\_Dim Table

**Purpose**: Captures patient-specific demographic and background information.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| PATIENT\_ID | NUMBER (38,0) | Primary key; unique identifier for patients. |
| GENDER | VARCHAR (10) | Patient's gender. |
| AGE | NUMBER (38,0) | Patient's age. |
| AGE\_GROUP | VARCHAR (20) | Categorized age group (e.g., 18-25). |
| STAR\_SIGN | VARCHAR (20) | Patient's star sign. |

#### Relationships:

* Linked to DIABETES\_FACT through PATIENT\_ID.

### 3.2. Health Metrics Dimension Table (Health\_Metrics\_Dim)

**Purpose**: Stores clinical measurements and health-related data.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| METRIC\_ID | NUMBER (38,0) | Primary key; unique identifier for metrics. |
| PATIENT\_ID | NUMBER (38,0) | Foreign key; links to Patient\_Dim. |
| BMI | NUMBER (38,1) | Body Mass Index of the patient. |
| WEIGHT | NUMBER (38,1) | Weight of the patient (kg). |
| HEIGHT | NUMBER (38,2) | Height of the patient (m). |
| HYPERTENSION | NUMBER (38,0) | 1 if patient has hypertension, 0 otherwise. |
| DIABETES\_PEDIGREE\_FUNCTION | NUMBER (38,2) | Genetic predisposition score for diabetes. |
| FAMILY\_DIABETES\_HISTORY | NUMBER (38,0) | 1 if diabetes history exists, 0 otherwise. |
| PREGNANCIES | NUMBER (38,0) | Number of pregnancies (if applicable). |

#### Relationships:

* Linked to DIABETES\_FACT through METRIC\_ID.

### 3.3. Lifestyle Dimension Table (Lifestyle\_Dim)

**Purpose:** Captures lifestyle choices and habits that may impact health.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| LIFESTYLE\_ID | NUMBER (38,0) | Primary key; unique identifier for lifestyle data. |
| PATIENT\_ID | NUMBER (38,0) | Foreign key; links to Patient\_Dim. |
| DIET\_TYPE | VARCHAR (255) | Type of diet followed by the patient. |
| PHYSICAL\_ACTIVITY\_LEVEL | VARCHAR (255) | Level of physical activity. |
| SOCIAL\_MEDIA\_USAGE | VARCHAR (255) | Social media usage habits. |
| SLEEP\_DURATION | NUMBER (10,1) | Average hours of sleep per day. |
| STRESS\_LEVEL | VARCHAR (255) | Stress level (e.g., low, medium, high). |
| ALCOHOL\_CONSUMPTION | VARCHAR (255) | Frequency of alcohol consumption. |

#### Relationships:

* Linked to DIABETES\_FACT through LIFESTYLE\_ID.

### 3.4. Fact Table (DIABETES\_FACT)

**Purpose**: Combines references to all dimensions and stores key measures.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| FACT\_ID | NUMBER (38,0) | Primary key; unique identifier for records. |
| PATIENT\_ID | NUMBER (38,0) | Foreign key; links to Patient\_Dim. |
| METRIC\_ID | NUMBER (38,0) | Foreign key; links to Health\_Metrics\_Dim. |
| LIFESTYLE\_ID | NUMBER (38,0) | Foreign key; links to Lifestyle\_Dim. |
| DIABETES | NUMBER (38,0) | 1 if patient has diabetes, 0 otherwise. |
| PREDICTED\_DIABETES\_FLAG | BOOLEAN | Indicates predicted diabetes status. |
| CREATED\_DTTM | TIMESTAMP\_NTZ (9) | Record creation timestamp. |
| MODIFIED\_DTTM | TIMESTAMP\_NTZ (9) | Last modification timestamp. |

#### Relationships:

* Integrates all dimensions for analytical queries.

# 4.0. Justification of Design

1. **Star Schema for Analytics**:

The star schema minimizes joins and improves query performance, which is critical for analytical workloads.

1. **Dimensional Structure**:
   1. Dimensions hold descriptive attributes, ensuring data normalization and avoiding redundancy.
   2. This structure supports slicing and dicing for metrics like average BMI or stress level distribution.
2. **Scalability**:

The schema allows easy addition of new dimensions or fact metrics, enabling future expansion.

1. **Efficiency**:

By centralizing metrics in DIABETES\_FACT, analysts can run complex queries without traversing multiple tables.

# 5.0 Conclusion

This schema design provides a robust framework for analyzing diabetes data, aligning with the goal of analysing the impact of the disease on the human body. By organising data into well-defined dimensions and a central fact table, the schema enables efficient and insightful analysis, empowering data-driven decisions for healthcare advancements.

Using a star schema for this dataset offers several advantages, including simplified querying and improved performance because the centralised fact table connected to dimension tables minimises the number of joins, making queries faster and more efficient. Its intuitive, denormalised structure enhances understandability, allowing analysts to easily comprehend and navigate the data for quicker insights. Additionally, the clear separation of facts and dimensions supports efficient aggregation and streamlined reporting processes. Furthermore, the schema provides scalability and flexibility by easily accommodating additional dimensions or measures, enabling the dataset to grow and adapt to evolving analytical needs. These benefits make the star schema ideal for organizing and analysing the dataset’s diverse attributes effectively.