Diabetes
Database Schema Design Report

Prepared by

[Team 10 – Mellitus]

Date: 22/11/2024

Table of Contents

1. 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.

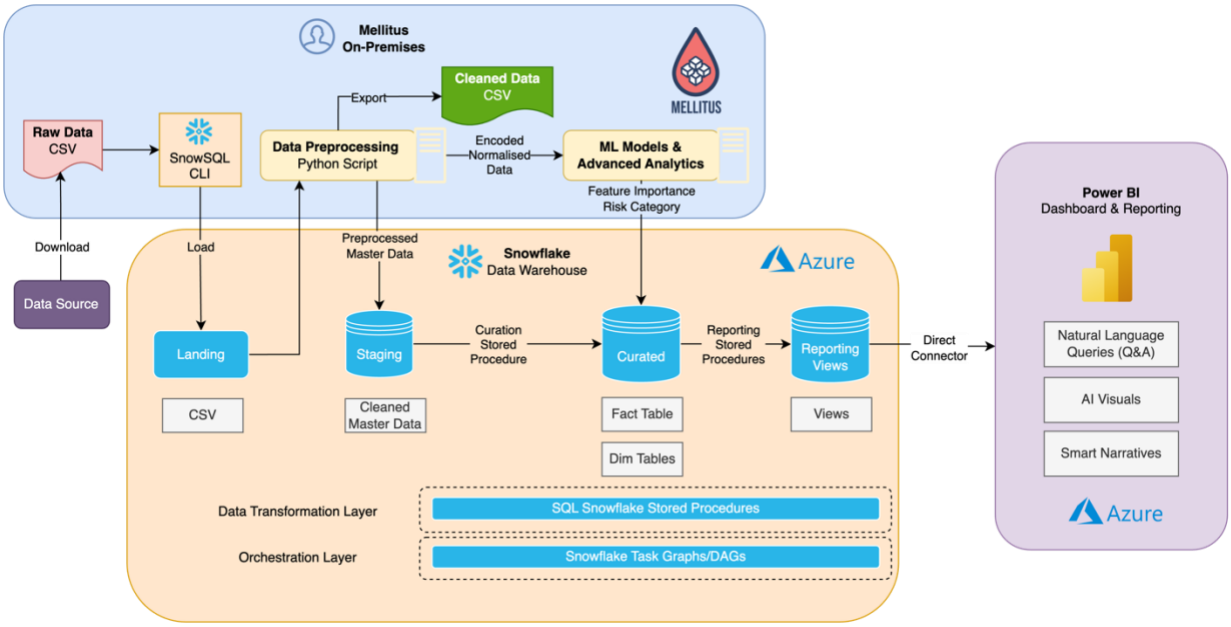
1.3 Overview of the Schema

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.

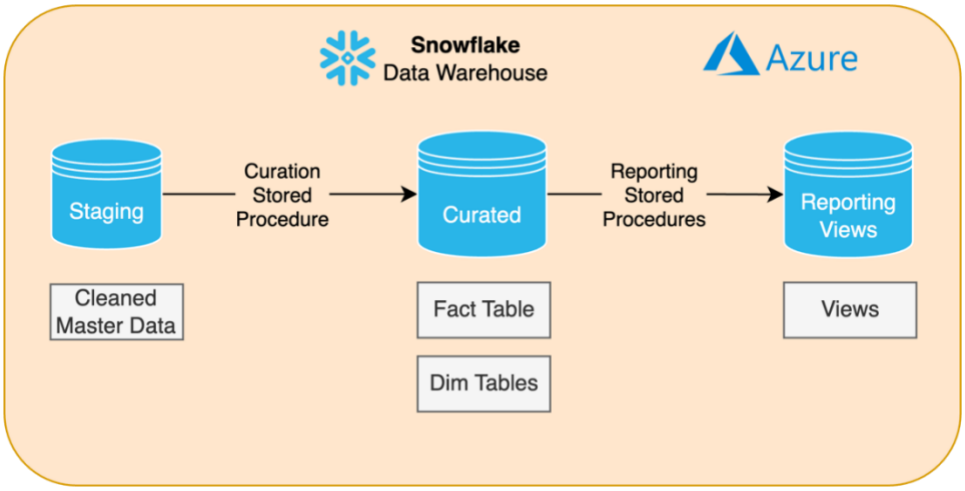
2. Relationships and Diagram

2.1 E2E High Level Architecture

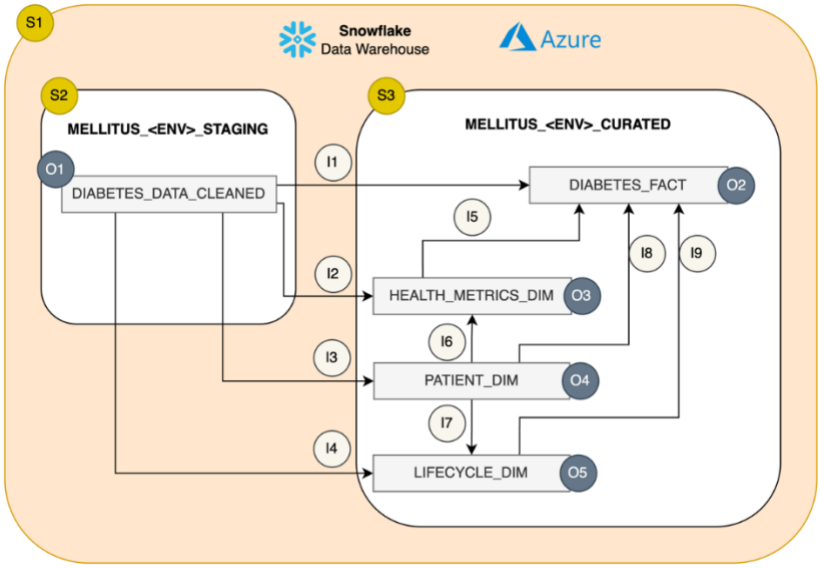


2.2 Data Layer Design

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.



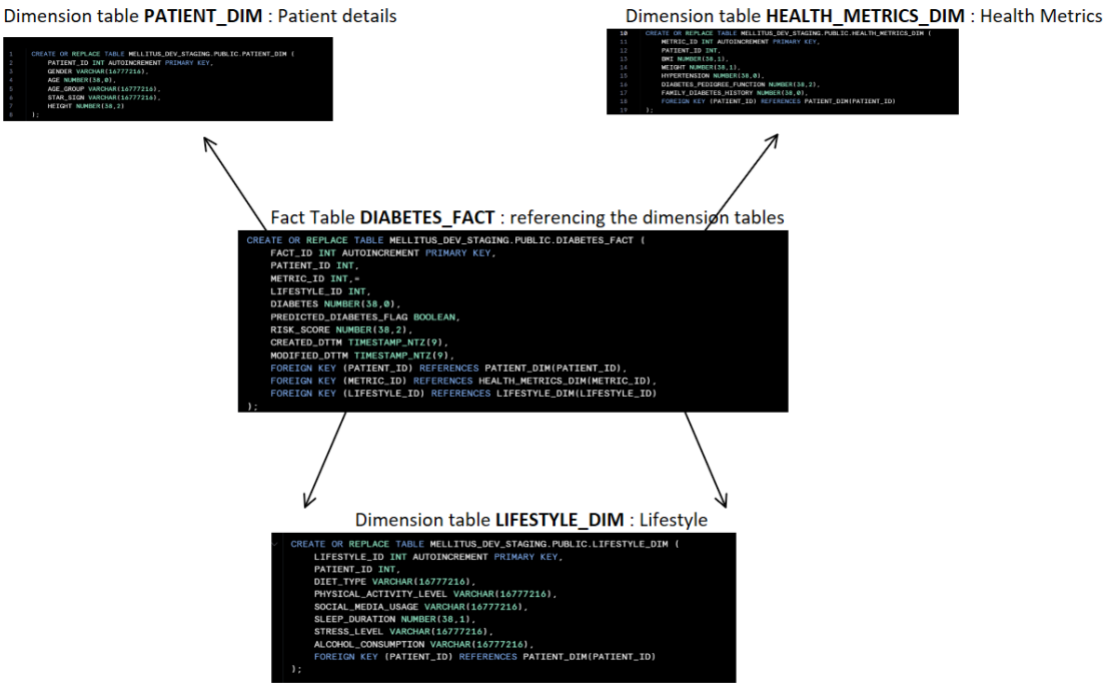
2.3 Curated Data Flow Diagram



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

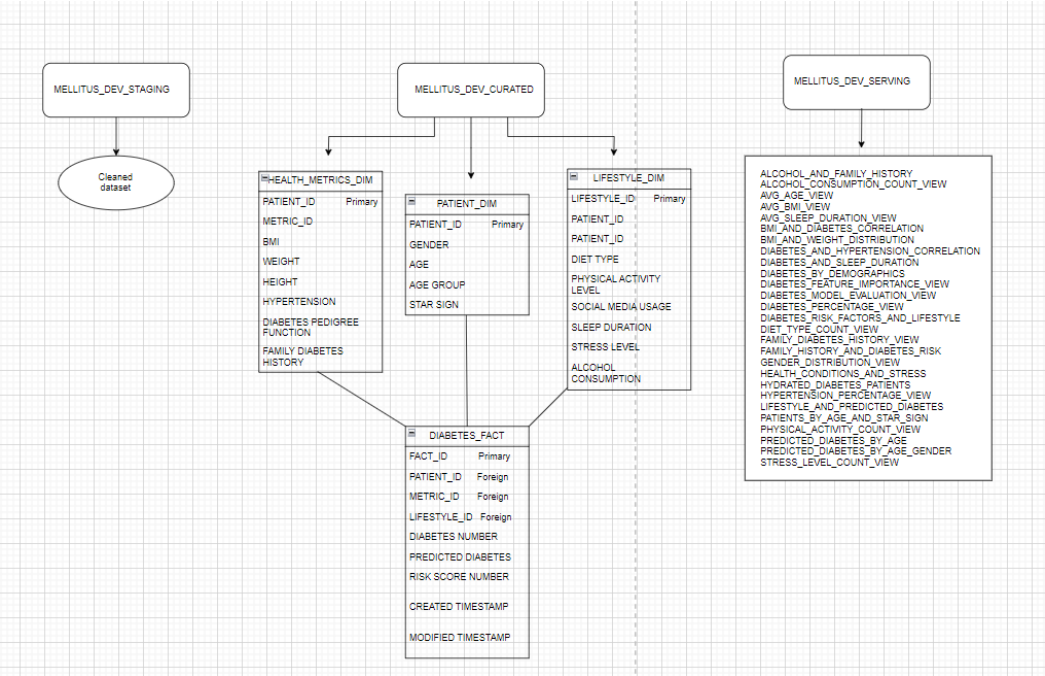
2.4 Star Schema (basic outline)

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

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

4. **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.