Phase 3 Documentation

Date	26-10-2023
Team ID	4162
Project Name	Data Warehousing with IBM Cloud Db2 Warehouse

Table of Contents

1.Introduction

In this document we building a data warehouse using IBM Cloud and Python, particularly with the Pandas library, and employing ETL (Extract, Transform, Load) processes is a powerful way to centralize and optimize data management within an organization. In this introductory overview, we'll explore the key components and concepts of this data warehousing approach.

2.Problem Statement

Objective: To bring together data from various sources, perform advanced data integration and transformation, and provide data architects with the tools to explore, analyse, and deliver actionable data for informed decision-making.

Data: We have multiple datasets containing various features for diabetes prediction (such as glucose, insulin, BMI, pregnancies etc). These data will be used to integrate and transform into a robust data warehouse.

3.Data Warehouse Structure:

3.1. Define the schema

Schema Definition:

- **Pregnancies (INT):** This column stores the number of times a person has been pregnant. It is of type INT, representing whole numbers.
- **Glucose (INT):** This column stores the plasma glucose concentration after a 2-hour oral glucose tolerance test. It is of type INT.
- **BloodPressure (INT):** This column stores the diastolic blood pressure (mm Hg). It is of type INT.
- **SkinThickness (INT):** This column stores the thickness of the skinfold of the triceps (mm). It is of type INT.
- **Insulin (INT):** This column stores the 2-hour serum insulin (mu U/ml). It is of type INT.
- **BMI (FLOAT):** This column stores the Body Mass Index (weight in kg/(height in m)^2). It is of type FLOAT, allowing decimal values.
- **DiabetesPedigreeFunction (FLOAT):** This column stores a diabetes pedigree function which represents the likelihood of diabetes based on family history. It is of type FLOAT.
- **Age (INT):** This column stores the age of the person (years). It is of type INT.
- **Outcome (BINARY):** This column stores binary values representing the outcome of whether a person has diabetes or not. The exact representation of these binary values (e.g., 0 and 1) would need to be defined based on the specific context of the dataset. It is of type BINARY.

3.2.Structure of the data warehouse tables:

CREATE TABLE diabetes_data: This line initiates the creation of a new table named diabetes_data in the database.

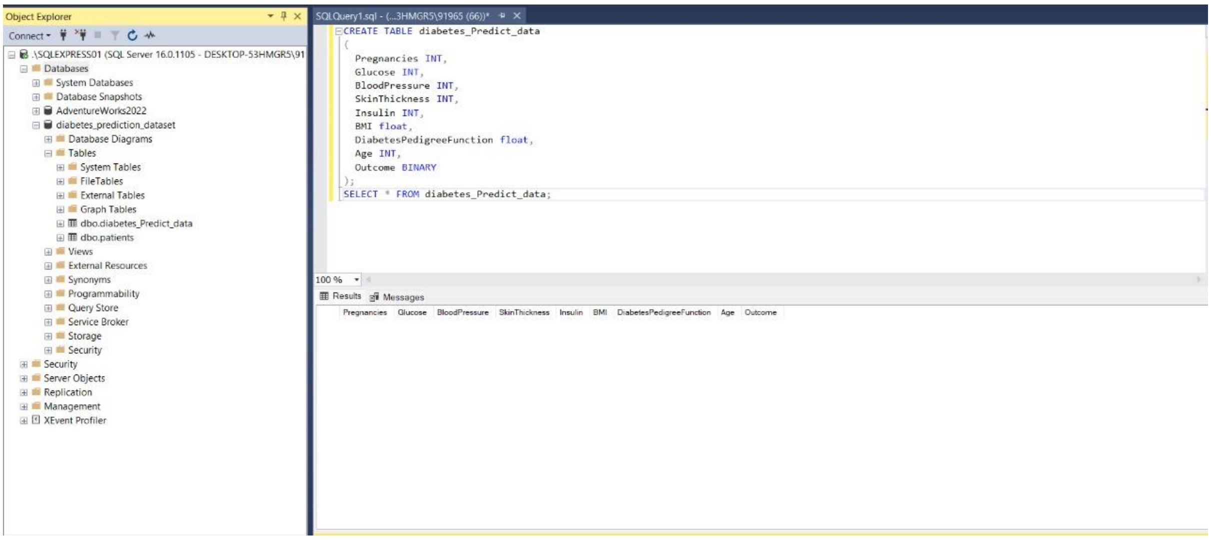
SQL QUERY:

```
 CREATE TABLE diabetes_data
(
Pregnancies INT,
Glucose INT,
BloodPressure INT,
SkinThickness INT,
    Insulin INT,
    BMI float,
    DiabetesPedigreeFunction float,
    Age INT,
    Outcome BINARY
    );
SELECT * FROM diabetes_data;
```

SELECT: This statement is used to select data from the table. which means it selects all columns from the specified table.

FROM diabetes_data: This part of the statement specifies the source table from which data is being selected, which is diabetes_data.

The purpose of the script is to create a table that can store data related to diabetes, and the **SELECT * FROM diabetes_data** statement is used to retrieve all the records (rows) from this table.



```
CREATE TABLE diabetes_Predict_data
(
  Pregnancies INT,
  Glucose INT,
  BloodPressure INT,
  SkinThickness INT,
  Insulin INT,
  BMI float,
  DiabetesPedigreeFunction float,
  Age INT,
  Outcome BINARY
);
SELECT * FROM diabetes_Predict_data;
```

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome

4. Data Integration:

- The objective of this project is to classify whether someone has diabetes or not.
- Dataset consists of several Medical Variables (Independent) and one Outcome Variable (Dependent)
- The independent variables in this data set are: -'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin','BMI', 'DiabetesPedigreeFunction', 'Age'
- The outcome variable value is either 1 or 0 indicating whether a person has diabetes (1) or not (0).

Dataset: https://www.kaggle.com/code/mvanshika/diabetes-prediction

5.Conclusion:

The SQL script and the schema for the diabetes_data table are designed to facilitate the storage and organization of health-related data, particularly related to diabetes. The table structure includes various columns, each with specific data types and meanings, to capture relevant information for analysis and research. The schema includes important health indicators, such as glucose levels, blood pressure, BMI, and age, which are crucial in understanding and assessing diabetes risk.