Name Chopra Aman, Ranjan Nidhi Date: 23rd November 2022

NYU NetID: ac8511, nr2387

Section: CSCI-GA.2433-001

**Project Part 2**

We Aman Chopra and Nidhi Ranjan, hereby certify by submitting this project Part 1 that all the efforts put into this part are our own. We have referred the project support materials.

**Total in points** (100 points total): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Professor’s Comments:**

**A picture containing arrow

Description automatically generated**

**Affirmation of my Independent Effort:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Description:**

We will be tackling four sub-parts in this part. In the first part, we will be talking about the datasets we have chosen that could be used by the insurance company. In the second part, we will make the logical design of the ER diagram we made in part 1 of this project. We would also talk about the optimizations and normalizations we did. In the third part, we will briefly discuss the reference architecture that is most suitable for the insurance company. In part 4, we would show how we leveraged the cloud platform to store and manage the data.  
  
**Part 1:  
  
Datasets used:**a) Diabetes Dataset  
<https://www.kaggle.com/datasets/mathchi/diabetes-data-set?resource=download>  
  
This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

* Pregnancies: Number of times pregnant
* Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
* BloodPressure: Diastolic blood pressure (mm Hg)
* SkinThickness: Triceps skin fold thickness (mm)
* Insulin: 2-Hour serum insulin (mu U/ml)
* BMI: Body mass index (weight in kg/(height in m)^2)
* DiabetesPedigreeFunction: Diabetes pedigree function
* Age: Age (years)
* Outcome: Class variable (0 or 1)

b) Cardiovascular Study Dataset

<https://www.kaggle.com/datasets/christofel04/cardiovascular-study-dataset-predict-heart-disea>

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression.

**Demographic**:  
• **Sex**: male or female("M" or "F")  
• **Age**: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)  
Behavioral  
• **is\_smoking**: whether or not the patient is a current smoker ("YES" or "NO")  
• **Cigs Per Day**: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)  
Medical( history)  
• **BP Meds**: whether or not the patient was on blood pressure medication (Nominal)  
• **Prevalent Stroke**: whether or not the patient had previously had a stroke (Nominal)  
• **Prevalent Hyp**: whether or not the patient was hypertensive (Nominal)  
• **Diabetes**: whether or not the patient had diabetes (Nominal)  
Medical(current)  
• **Tot Chol**: total cholesterol level (Continuous)  
• **Sys BP**: systolic blood pressure (Continuous)  
• **Dia BP**: diastolic blood pressure (Continuous)  
• **BMI**: Body Mass Index (Continuous)  
• **Heart Rate**: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)  
• **Glucose**: glucose level (Continuous)  
Predict variable (desired target)  
• **10 year risk of coronary heart disease CHD**(binary: “1”, means “Yes”, “0” means “No”)

c) Obesity Levels and Life Style  
<https://www.kaggle.com/code/mpwolke/obesity-levels-life-style/data>

# Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico

@attribute Gender {Female,Male}

@attribute Age numeric

@attribute Height numeric

@attribute Weight numeric

@attribute family\_history\_with\_overweight {yes,no}

@attribute FAVC {yes,no}

@attribute FCVC numeric

@attribute NCP numeric

@attribute CAEC {no,Sometimes,Frequently,Always}

@attribute SMOKE {yes,no}

@attribute CH2O numeric

@attribute SCC {yes,no}

@attribute FAF numeric

@attribute TUE numeric

@attribute CALC {no,Sometimes,Frequently,Always}

@attribute MTRANS {Automobile,Motorbike,Bike,Public\_Transportation,Walking}

@attribute NObeyesdad {Insufficient\_Weight,Normal\_Weight,Overweight\_Level\_I,Overweight\_Level\_II,Obesity\_Type\_I,Obesity\_Type\_II,Obesity\_Type\_III}

**Problem Statement:**

We have saved the datasets in AWS S3 buckets and compiled the data to get relevant features. We have taken the following features:

1. BP
2. Skin Thickness
3. Age
4. BMI
5. Gender
6. Cigs Per Day
7. Glucose
8. Family\_history\_with\_overweight

These features are passed to the customer table which can be seen in the Logical diagram below. This along with the number of claims and accidents done by the customer can be used by the insurance company to help forecast chronic disease. Additionally, the suggestion of chronic disease might indicate the chances of claims and the number of accidents to be higher.

**Part 2:**

We will be preparing the optimized Logical Database Schema:

**Normalization:**

**It** is a database design technique that reduces data redundancy and eliminates undesirable characteristics like Insertion, Update, and Deletion Anomalies. Normalization rules divide larger tables into smaller tables and link them using relationships. The purpose of Normalization in SQL is to eliminate redundant (repetitive) data and ensure data is stored logically.

In our Schema, each table cell contains a single value and is unique except the ones with the phone number as phone numbers are multivalued attributes. **To make the schema in 1NF, we have created a separate phone numbers table for Customers and Employees as shown below.**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

The table contains the customer or Employee Id along with the Phone numbers as shown.

After this change, **all the entities are in 1NF**

In our Schema, there are no composite keys and the single Column Primary Key does not functionally dependent on any subset of the candidate key relation, **Hence all the entities are in 2NF also.**

Some of the entities are not in 3NF as they have transitive functional dependencies. For instance, the Salutation can be derived from Gender. To make the relation in 3NF, we have created a new Salutation table with Salutation ID and Salutation which can be used by EMPLOYEE and CUSTOMER tables as shown below. We have also removed the derived attribute AGE as it can be calculated from DOB.

Graphical user interface, text, application

Description automatically generated

**Optimizations**:

1. The first optimization we did was to remove multivalued attributes and created a separate table as part of 1NF.
2. Secondly, we tried to tackle the weak entity by passing the primary key of the Strong entity as a Foreign Key to the weak entity. We can see below that We have passed the Cust\_Id to History and Vehicle table as the Foreign Key.

Graphical user interface, application

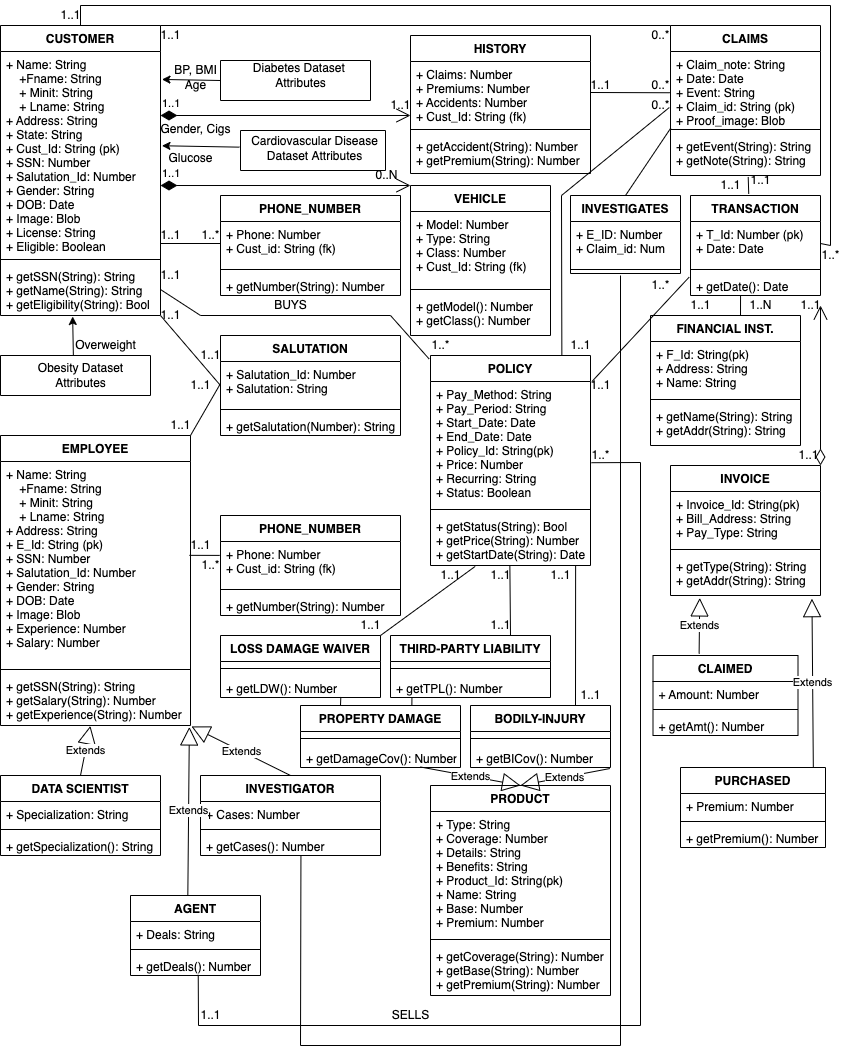
Description automatically generated

1. To tackle the M:N relation, we have created a separate table INVESTIGATES which is used CLAIMS and INVESTIGATOR.

Graphical user interface, text, application

Description automatically generated

1. In this optimization, we have taken the three datasets in the Data Lake and stored them in the AWS S3. We have taken attributes from these datasets and have fed them to the Customer table as shown below. This can be used to get insights and forecast chronic disease probability based on the Insurance dataset.

**Logical Database Schema:**

**Part 3:**

<https://www.ibm.com/cloud/architecture/architectures/insurance/reference-architecture/>

Using container architecture for business automation.

Data and AI capabilities to analyze the data.  
  
Non-Functional Requirements:

1. **Security:** Every connection in and out of the enterprise must be secured. All data in transit must be sent over secure protocols and any data that is stored in the cloud or data center must be encrypted. Customers, claims adjusters, and agents must be authenticated and checked for authorization before they're allowed access to business functions. All controls for audits and compliance must be provided.
2. **Response Time:** Speed is synonymous with a better customer experience. Insurance architectures require minimal latency because decisions must be made quickly whether the communication is between customers, adjusters, agents, or ecosystem partners.
3. **Connectivity:** Insurance cloud platforms must provide secured connectivity with many entities:

* The agents' infrastructure
* Ecosystem partners such as utilities, rental car companies, hotels, the DMV, credit agencies, medical providers, payers, auto repair facilities, home repair contractors, cars, and homes
* Enterprise and ISV applications that run on premises

Availability, maintainability, and scalability must also be addressed in an insurance architecture.

**Part 4:**

To store our data, we have made use of Amazon Web Service. We have saved the datasets in the S3 buckets.

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated