# horizontal lineDatabase

Project 2

I confirm that this is my own work and that use of material from other sources, including the Internet, has been properly and fully acknowledged and referenced.

|  |  |
| --- | --- |
| Name: | Pang, Jinhao |
| Date: | 2022.11.23 |
| NYU ID: | N19475049 |
| Course Section Number: | csci-ga.2433-001 |



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

**Professor’s Comments:**

|  |
| --- |
|  |

**Data Lake**

The data collection can start from the dataset which has integrated information of a person in each record. For example, one record could have the diseases the person gets, gender, if have different factors to cause chronic diseases, age, country, etc. The most detailed dataset can be used as a base set to forecast chronic diseases, but it is still insufficient to cover all of the factors. Except to expend the attributes stored in the dataset, we can collect other factors that roughly estimate the potential effects on the diseases. For example, it may be hard to know if a person burns the midnight oil a lot or not and how it counts for the prediction in this detailed dataset, but we can find a dataset where we can estimate its effects on chronic diseases. Then we can take all the factors into account by a specific algorithm to integrate them. As for data late, we can extract the attributes from the unstructured data to construct it into structured data. Take images as an example, if they are used to check the situation of a specific position, we can transform it into the number of tumors or their diameter. For those images and voices that keep the original form would be better, just to store them in the database as unstructured data.

**Logical Schema optimization**

The Entity-relationship conceptual model developed in project 1 contains the attributes and keys, which is actually a logical schema. Hence, the previous work should not have these features and no more modifications will be made to the structure of the model. Here shows a simple example of schema transformation in Figure 1.

Diagram

Description automatically generated

Figure Logical schema of customer

Compared to the conceptual model, the logical schema has attributes and keys. In the conceptual model, Customer has alias, address, images, and relation, they can all be designed in customer entity, while in the logical schema, they have to become another entity because of the 2NF and 3NF.

**Reference Architecture**

Information Architecture can be suitable for the insurance company. First, it addresses the information and data aspects of the Business Unit Reference Architecture. The aim is to use hybrid data as part of their business, so this architecture is suitable to handle information in business and includes various operations in both structured and unstructured formats. The requirement is to predict chronic diseases based on various factors. According to the summary of information architecture, the update frequency of transactional data stores, and operational data stores is in real-time. Reporting and Analysis layer contains the tools that enable query and analysis of the data in the various data stores. Hence, it will be suitable for the company to update the datasets and modify the model to predict the result in real-time.

**Cloud platform**

The database used is MySQL deployed on Microsoft Azure.

Graphical user interface, application

Description automatically generated

Figure MySQL instance

This is the sample dataset of a specific kidney chronic disease stored in csv, which is a structured data that can be stored directly in the database.

Application, table, Excel

Description automatically generated

Figure Sample data

Some unstructured data:

Map

Description automatically generated

Figure Distribution