Name ____ Zhang Xinge,QianZhang__ Date:

__2023.11.1___

NYU ID: ____N10837293, N19756113_____

Net ID: XZ4513, QZ2570

Course Section: _____ CSCI-GA.2433-001_____

Project #2

Total in points (100 points total): _____

Professor's Comments:

Affirmation of my Independent Effort:

_____ZHANG XINGE_Zhang Qian_____

(Sign here)

README:

This assignment was completed by a group consisting of two students Xinge Zhang and Qian Zhang. Each member is considered to contribute equal effort to this solution.

2 Unstructured data collections

1

In this section, we plan to use the datasets Medical Insurance Premium Prediction as our data lake.

2

Please go to section **3 EDA Logical Schema Optimization.**

3

Currently, we believe that using a traditional SQL relational database is the most reasonable choice. For the purpose of predicting diseases, structured data such as the smoking status shown in the above image is more important compared to unstructured data like user photos. Additionally, SQL databases are more compatible with potential future machine learning algorithms, enabling us to achieve the prediction goal. Overall, using an SQL relational database allows us to leverage the advantages of structured data while the disadvantage of inconveniently storing unstructured data in this project is not that significant.

4

In this section, we deploy a SQL database on Microsoft Azure Clould. This is a traditional relational SQL database based on the logical schema we created in section 3 suitable to store structured data. However, it's now private and could only be connected from our group member's IP address. But we also deploy a NoSQL database and data lake on mongoDB, since it's more suitable to handle unstructured data such as image.
Everyone could access this mongoDB using VS Code extensions "mongoDB for VSCode" the connection string is "mongodb+srv://qz2570:maimaiQn@dbsproject1.ynuaav7.mongodb.net/".

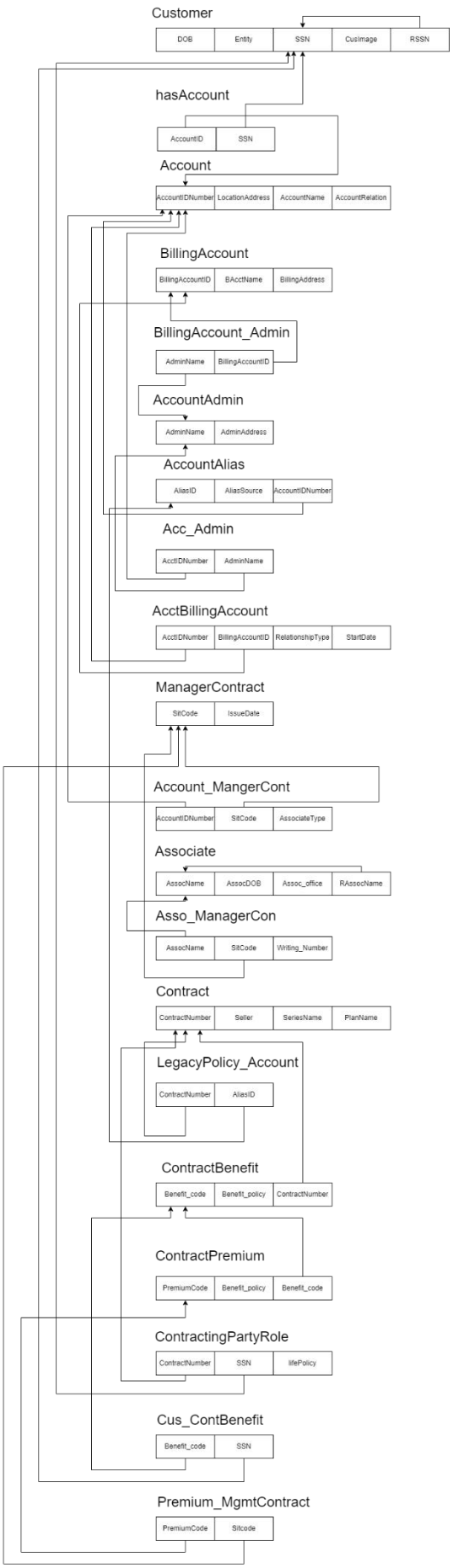
To be honest, we don’t know which database is more suitable for this project since we don’t know the project overall goals. So we will test the database and choose a proper one in the remaining part of this project.

3 EDA Logical Schema Optimization

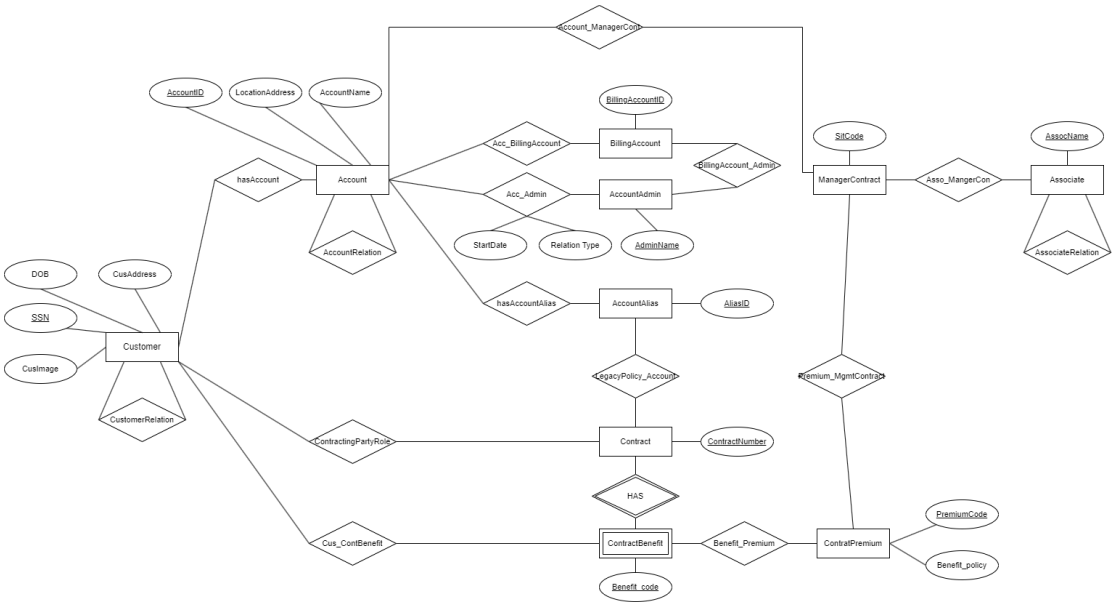
1

Create and/or generate a logical schema that corresponds to the entity-relationship conceptual model developed earlier in the first part of the project. Please note that the tool you used to create the conceptual model may provide support to facilitate the generation of a logical schema for a database system of your choice. Please make sure that you select the database system target that corresponds to the database product you plan to use to manage and store data as part of your project solution.

In this part we create a relational database schema illustrated below. Due to the large size of the schema diagram and the limitations of the PDF page size, the image displayed here is not clear. For a high-resolution image, please refer to the files we have provided in PNG, SVG, or HTML formats. We have also included the original draw.io project file.


Customer
DOB
Entity
SSN
CusImage
RSSN
hasAccount
AccountID
SSN
Account
AccountIDNumber
LocationAddress
AccountName
AccountRelation
BillingAccount
BillingAccountID
BAcctName
BillingAddress
BillingAccount_Admin
AdminName
BillingAccountID
AccountAdmin
AdminName
AdminAddress
AccountAlias
AliasID
AliasSource
AccountIDNumber
Acc_Admin
AcctIDNumber
AdminName
AcctBillingAccount
AcctIDNumber
BillingAccountID
RelationshipType
StartDate
ManagerContract
SitCode
IssueDate
Account_MangerCont
AccountIDNumber
SitCode
AssociateType
Associate
AssocName
AssocDOB
Assoc_office
RAssocName
Asso_ManagerCon
AssocName
SitCode
Writing_Number
Contract
ContractNumber
Seller
SeriesName
PlanName
LegacyPolicy_Account
ContractNumber
AliasID
ContractBenefit
Benefit_code
Benefit_policy
ContractNumber
ContractPremium
PremiumCode
Benefit_policy
Benefit_code
ContractingPartyRole
ContractNumber
SSN
lifePolicy
Cus_ContBenefit
Benefit_code
SSN
Premium_MgmtContract
PremiumCode
Sitcode

This relational database is based on a simplified version ER diagram of project part 1, due to the extreme complexity the original ER diagram was, we must abstract that into a new ER diagram like this:

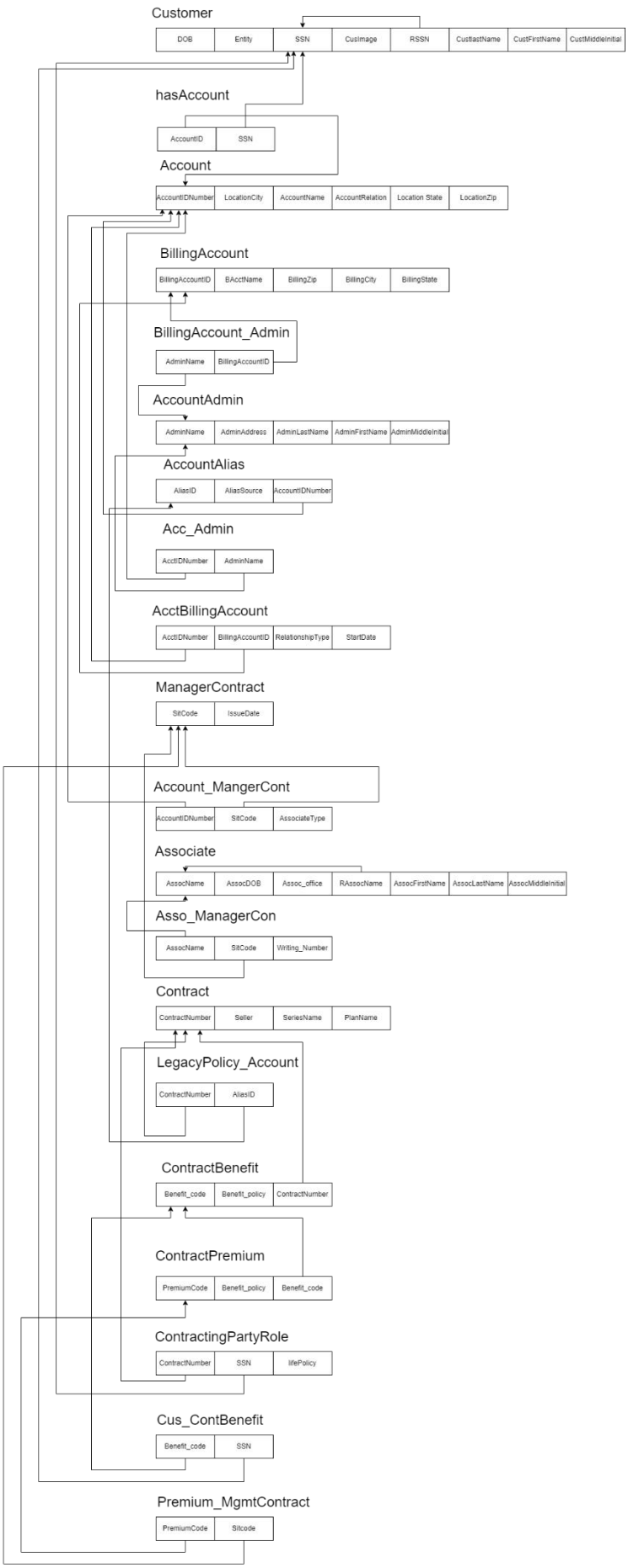


After careful consideration, we believe that this new image is simpler, more elegant, readable, and contains all the important information from the original image. So we will use this ER diagram for the remaining parts of this project.

2

Optimize your logical schema and provide details as to the reasoning behind each one of the optimization steps you are taking. Please note that optimization in this context includes normalization as well as extensions required to inter-related structured and unstructured data.

In this part we provide a optimized version of the above relational schema and we name this one “relational schema 3NF”. First, we decompose multivalued attribute such as customer name into three attributes cusFirstName, cusLastName, and cusMiddleInitial to satisfy 1NF. After this, we carefully check this relational schema and find in every relation, there are no partial dependency nor transition dependency. It may due to in most of these relations, candidate keys are likely to contain only one attribute, which means partial dependency issue has a little chance to occur in this condition. So after normalization, this logical schema look like this:


Customer
DOB
Entity
SSN
CusImage
RSSN
CustlastName
CustFirstName
CustMiddleInitial
hasAccount
AccountID
SSN
Account
AccountIDNumber
LocationCity
AccountName
AccountRelation
Location State
LocationZip
BillingAccount
BillingAccountID
BAcctName
BillingZip
BillingCity
BillingState
BillingAccount_Admin
AdminName
BillingAccountID
AccountAdmin
AdminName
AdminAddress
AdminLastName
AdminFirstName
AdminMiddleInitial
AccountAlias
AliasID
AliasSource
AccountIDNumber
Acc_Admin
AcctIDNumber
AdminName
AcctBillingAccount
AcctIDNumber
BillingAccountID
RelationshipType
StartDate
ManagerContract
SitCode
IssueDate
Account_MangerCont
AccountIDNumber
SitCode
AssociateType
Associate
AssocName
AssocDOB
Assoc_office
RAssocName
AssocFirstName
AssocLastName
AssocMiddleInitial
Asso_ManagerCon
AssocName
SitCode
Writing_Number
Contract
ContractNumber
Seller
SeriesName
PlanName
LegacyPolicy_Account
ContractNumber
AliasID
ContractBenefit
Benefit_code
Benefit_policy
ContractNumber
ContractPremium
PremiumCode
Benefit_policy
Benefit_code
ContractingPartyRole
ContractNumber
SSN
lifePolicy
Cus_ContBenefit
Benefit_code
SSN
Premium_MgmtContract
PremiumCode
Sitcode

To maximize extensibility for this schema, we also design some attributes such as cusImage for Customer to store potential unstructured data. With this attribute, we could easily extend this relation database to NoSQL database we deployed on mongoDB with image URL. Which make inter-related structured and unstructured data possible to handle.