**Project Part 2**

Team member: Ziyu Lin, Lu Huang, Qin Du

**1. Put together a small data lake to compile various relevant datasets and identify how insights**

**that may be extracted from the data collected will feed into the EDA that was designed in the**

**first part of the project.**

**Overview:**

We established a data lake to compile various datasets that will assist the insurance company in forecasting chronic disease trends and optimizing insurance products. The data lake uses Azure Data Lake Storage Gen2 as the primary storage platform, with Azure Data Factory for ETL (Extract, Transform, Load) processes to ensure data quality and consistency.

**Dataset Description:**

The dataset includes 5,000 synthetic patient records designed to simulate real-world chronic disease patterns and their correlation with demographic, socioeconomic, and lifestyle factors. Each record includes:

* SSN (unique identifier)
* Disease Types: Chronic respiratory disease, Type 2 diabetes, heart disease, arthritis, obesity, etc.
* Demographics: Age group, gender, race, education level, income level.
* Lifestyle Factors: Smoking, alcohol consumption, exercise frequency, diet quality.
* Health Indicators: BMI category, sleep hours, stress levels.

The data incorporates realistic correlations between different variables, such as the higher prevalence of certain diseases in older age groups, the relationship between lifestyle choices (smoking, exercise, diet) and health outcomes, and the connection between education level and income. The disease types range from common chronic conditions like hypertension and diabetes to more specific conditions like fibromyalgia and sleep apnea. The detailed columns are as follows:

|  |  |  |
| --- | --- | --- |
| Parameter Name | Values | Description |
| SSN | XXX-XX-XXXX | Social Security Number, Unique Identifier |
| Disease\_Type | Chronic Respiratory Disease, Heart Disease, Type 2 Diabetes, Breast Cancer, Lung Cancer, Colorectal Cancer, Hypertension, Obesity, Arthritis, Depression, Chronic Kidney Disease, Liver Disease, Multiple Sclerosis, Fibromyalgia, Sleep Apnea | Type of Chronic Disease |
| Age\_Group | 18-30 (15%), 31-45 (25%), 46-60 (35%), 61+ (25%) | Age Distribution |
| Gender | Male, Female, Other | Gender Identity |
| Race | Caucasian, Hispanic, Native American, Black, Asian, Other | Racial Background |
| Education\_Level | Less than high school, High school graduate, Some college, Bachelor's degree, Graduate degree | Educational Attainment |
| Income\_Level | Less than $30,000, $30,000-$50,000, $50,000-$75,000, $75,000-$100,000, $100,000+ | Annual Income |
| Exercise\_Frequency | Never, 1-2 times/week, 3-4 times/week, 5+ times/week | Physical Activity Level |
| Diet\_Quality | Poor, Fair, Good, Excellent | Dietary Habits |
| Sleep\_Hours | Less than 6, 6-7, 7-8, 8+ | Sleep Duration |
| Smoking | Never, Former, Current | Smoking Status |
| Alcohol\_Consumption | Never, Occasional, Regular, Heavy | Alcohol Usage |
| BMI\_Category | Underweight, Normal, Overweight, Obese | Body Mass Index Category |
| Stress\_Level | Low, Medium, High | Stress Assessment |
| Family\_History | Yes, No | Family Medical History |
| Occupation | Office/Desk Job, Manual Labor, Healthcare, Education, Service Industry, Retired, Unemployed | Employment Type |
| Insurance\_Type | Private, Medicare, Medicaid, Uninsured | Healthcare Coverage |
| State | Illinois, California, Florida, North Carolina, Michigan, New York, Georgia, Pennsylvania, Texas, Ohio, Virginia, Washington, Arizona, Massachusetts, Colorado | US State of Residence |

**How Insights Feed into EDA:**

The data collected feeds into the Exploratory Data Analysis (EDA) conducted earlier, focusing on identifying patterns related to chronic disease prevalence among different demographic groups. Key insights include the impact of lifestyle factors on disease outcomes, which can guide insurance policy adjustments and customer segmentation.

**2. Create/generate and optimize a logical database schema for the structured data used by the**

**insurance company at hand based on the conceptual model created earlier. The logical**

**database schema should inter-relate structured and unstructured data and support the hybrid**

**data model/data lake required for the business to capture insights and drive decisions. Please**

**note that you should not focus on generating actual insights in this part of the project. Please**

**follow the steps in section 3 below (i.e., EDA Logical Schema Optimization) for details on**

**how to create/generate and optimize your logical schema.**

The ER model designed in part 1 is translated into a logical schema shown below. It supports the insurance company's hybrid data model, integrating structured and unstructured data for efficient decision-making.

**Schema Overview:**

Besides from the entities and relationships built from part 1, it also includes an entity named ‘HEALTH\_RECORD’ to capture customers’ health related information, ensuring streamlined operations and risk analysis.

**Schema Optimization:**

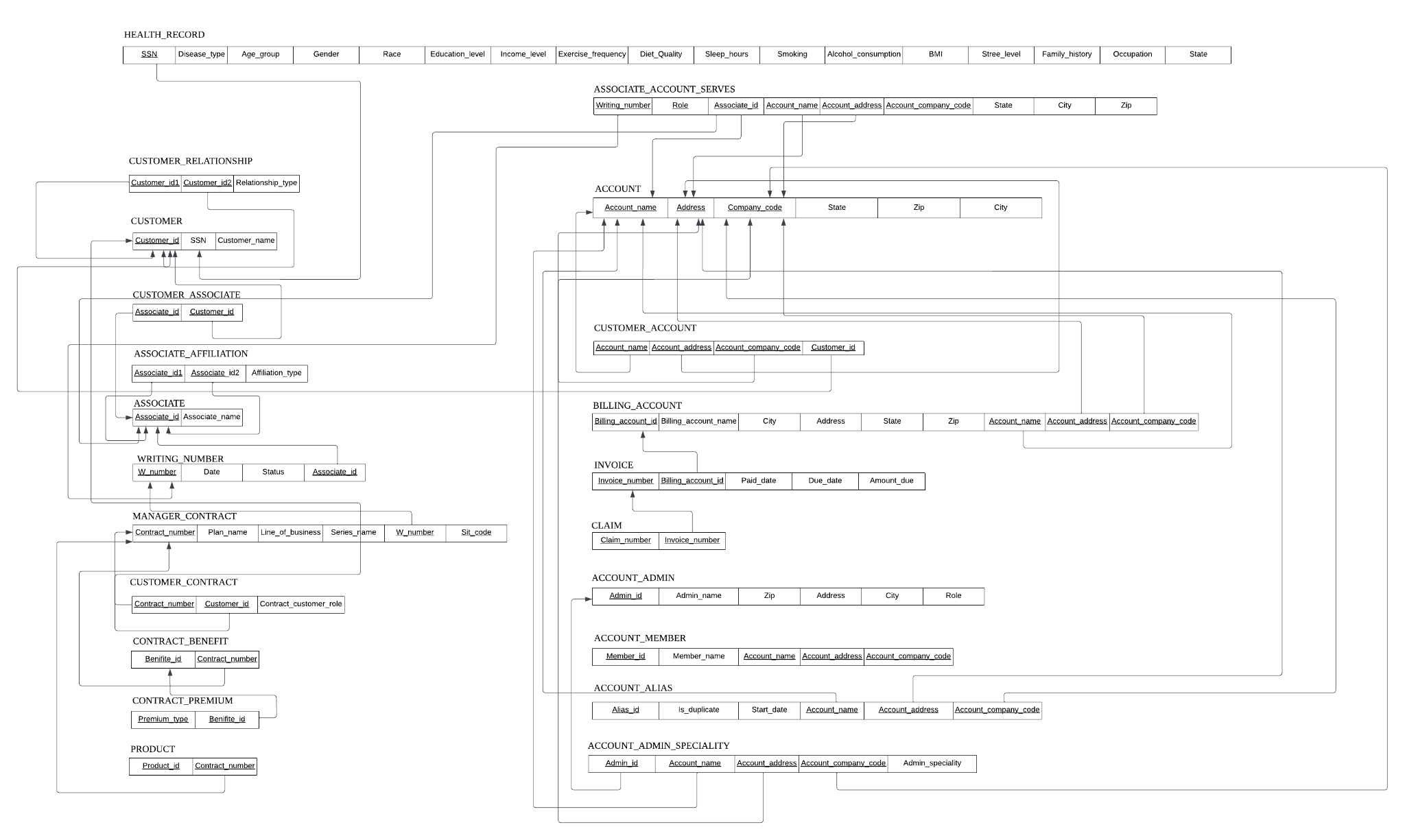
The logical database schema designed for the insurance company has been carefully optimized to ensure efficient data management, scalability, and support for both structured and unstructured data. The schema addresses key business needs, such as managing customer information, health records, accounts, invoices, and contracts, while enabling advanced analytics capabilities.

1. Normalization (3NF):

The schema has been normalized to minimize redundancy and maintain data integrity by separating entities like Customer and Customer\_Account.

1. Indexing for Query Speed:

Indexes have been added to high-use fields (Customer\_id, Invoice\_number) and composite indexes to optimize complex filters.

****

**3. Elaborate on the reference architecture that is most suitable for the insurance company to use**

**in order to leverage hybrid data as part of their business.**

We designed a scalable architecture using Azure Cloud to support data ingestion, processing, and analytics.

**Key Components:**

1. **Data Sources:**

All data is structured (e.g., demographics, health records, claims) and stored in Azure SQL Database.

1. **Data Processing:**

Azure Data Factory automates data ingestion from CSV files and databases, performing ETL to ensure data quality.

1. **Data Storage & Analytics:**

Azure SQL Database handles structured data storage.

Azure Synapse Analytics supports large-scale data queries and reporting.

Azure Analysis Services enables multidimensional analysis for quick insights.

1. **Data Governance & Security:**

Azure Purview for metadata management.

Azure Active Directory for access control.

Azure Key Vault secures sensitive data.

Azure Monitor ensures compliance and security monitoring.

### **Data Flow:**

1. Ingest data using Data Factory into Azure SQL Database.
2. Clean and transform data using ETL pipelines.
3. Analyze data in Synapse Analytics for insights.
4. Secure data using Key Vault and manage access with Active Directory.

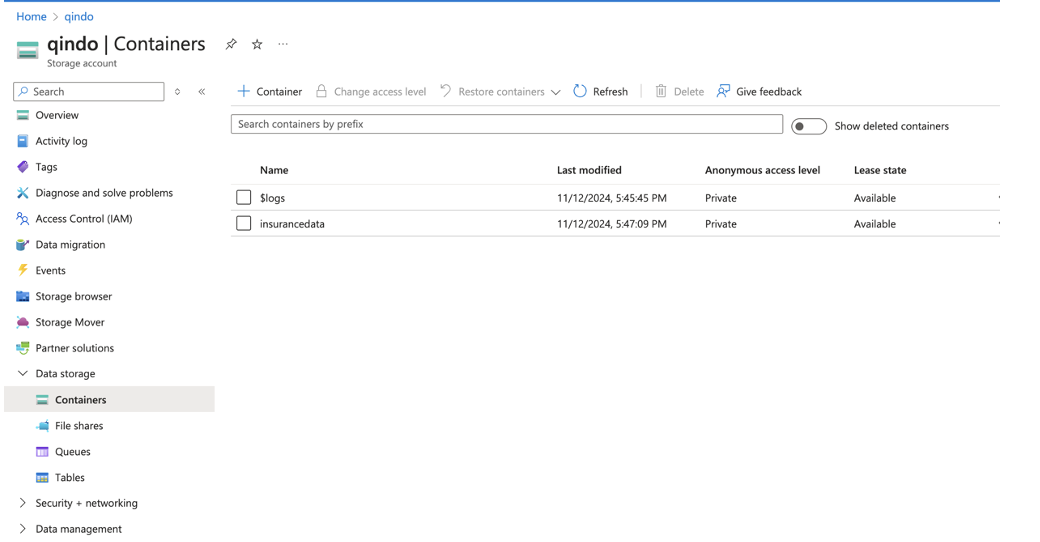
### **Benefits**

1. Efficient Processing: Fast data ingestion and analytics with Azure tools.
2. Scalable: Easily handles growing data volumes.
3. Secure: Robust encryption and access controls ensure data protection.

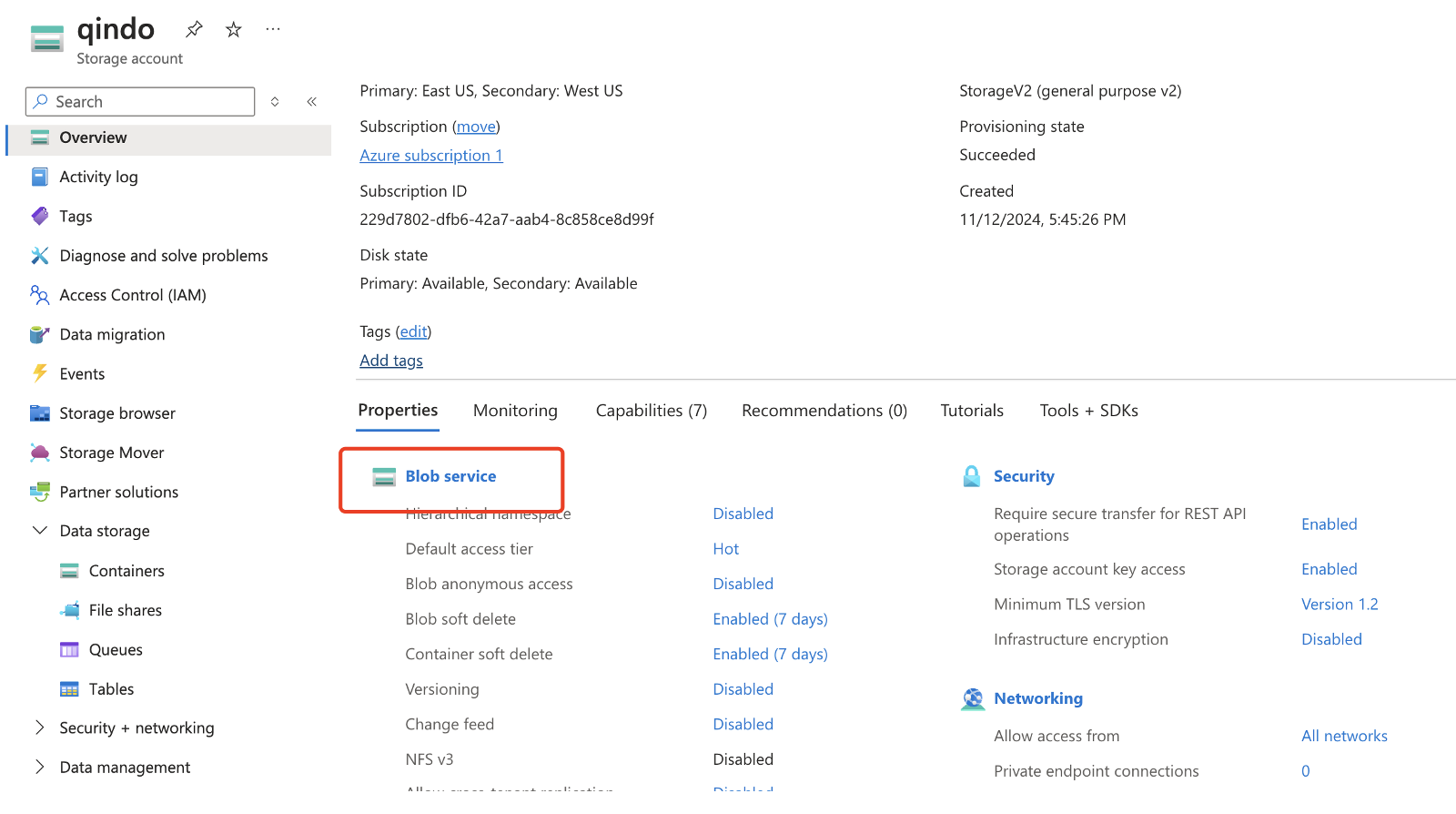
**4. Leverage a cloud platform (e.g., Microsoft Azure Cloud) to store and manage (all or part of)**

**your hybrid data.**

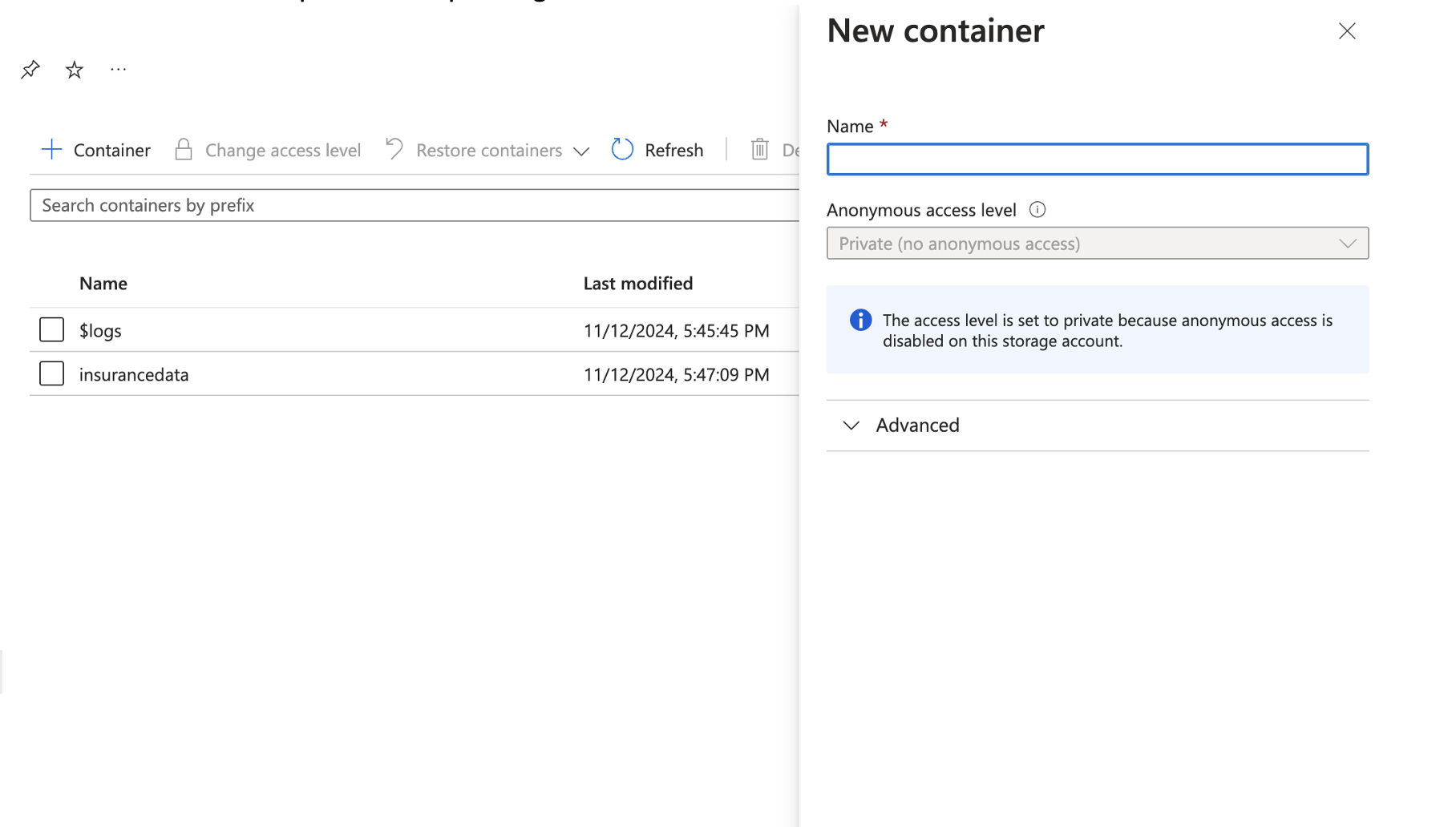
1. Create a storage account in Data Lake Gen 2 and initialize a data lake resource group

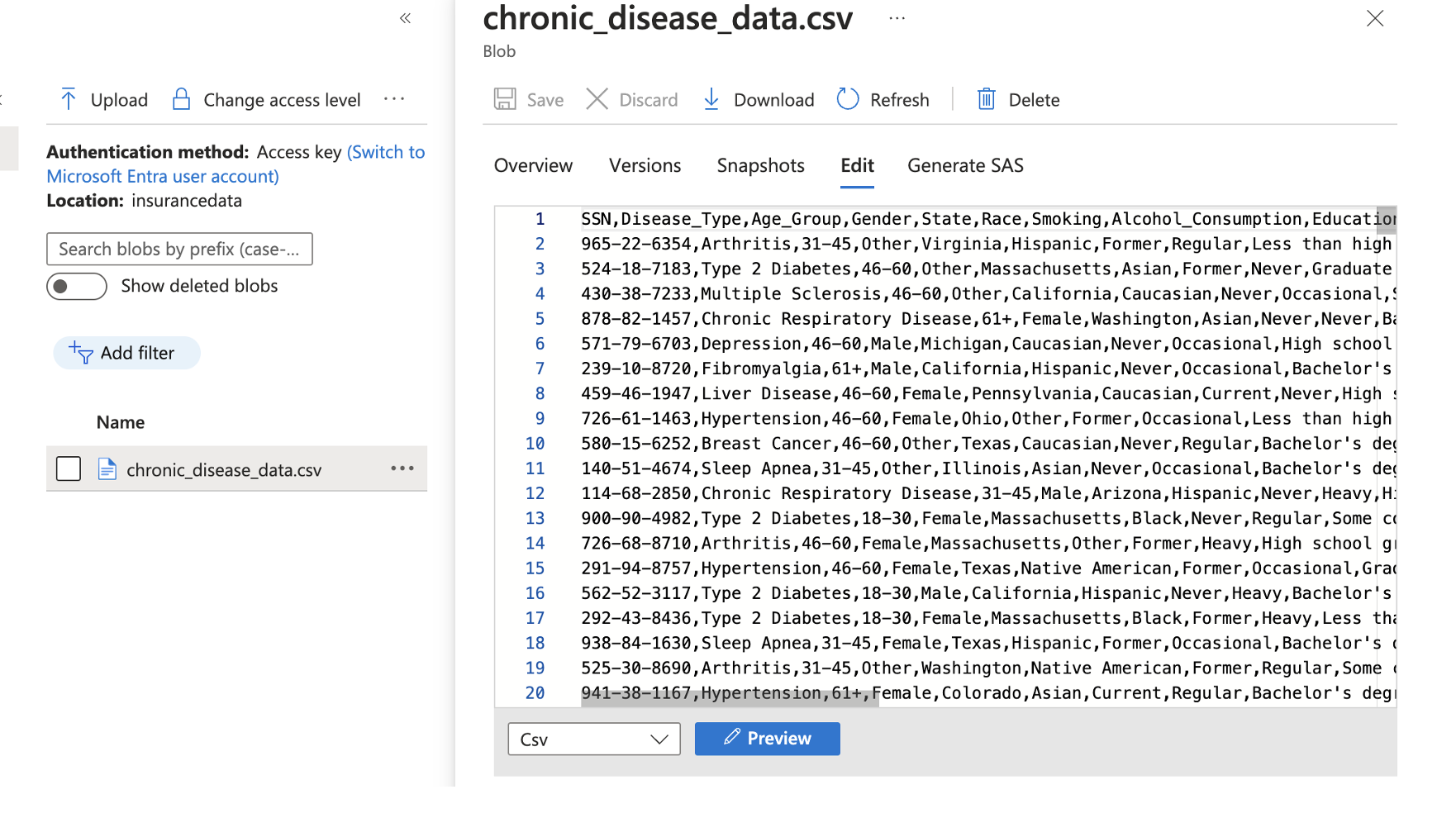
****

2. Create a new container and store and manager insurance data in the Blob service

****

3. Create a data table and upload corresponding data

****

****