Solution Design

1. Solution
    a. Data collection
    b. Uploading data into S3
    c. Data cleaning in PySpark to handle missing values, duplicate records, and inconsistencies.
    d. Raw data transformation into structured format.
    e. Create the connection using Access Key and Secret Key.
    f. Create advanced analytics for the use cases using PySpark
    g. Visualize the use cases with graphs, bars and charts using visualization tools like Databricks.
    h. Store the analyzed results in AWS Redshift.
2. Use cases
    a. Analysing disease prevalence and claims data.
    b. Segmenting subscribers based on age group.
    c. Segmenting subscribers based on sub-group.
    d. Segmenting subscribers based on demographics group.
    e. Segmenting subscribers based on specific disease.
    f. Analysing the number of rejected claims.
    g. Segmenting subscribers based on sex.
    h. Analyzing policy subscription patterns.
3. Database design
    a. Tables Metadata Info with PK/FK relationship

DISEASE (Disease_ID **PK**, SubGrpID **FK**, Disease_name)

GROUP (Grp_ID **PK**, Country, Premium_written, Zipcode, Grp_Name, Grp_Type, City, Year)

GRPSUBGRP (Grp_ID **FK**, SubGrpID **FK**)

HOSPITAL (Hospiital_id **PK**, Hospital_name, City, State, Country)

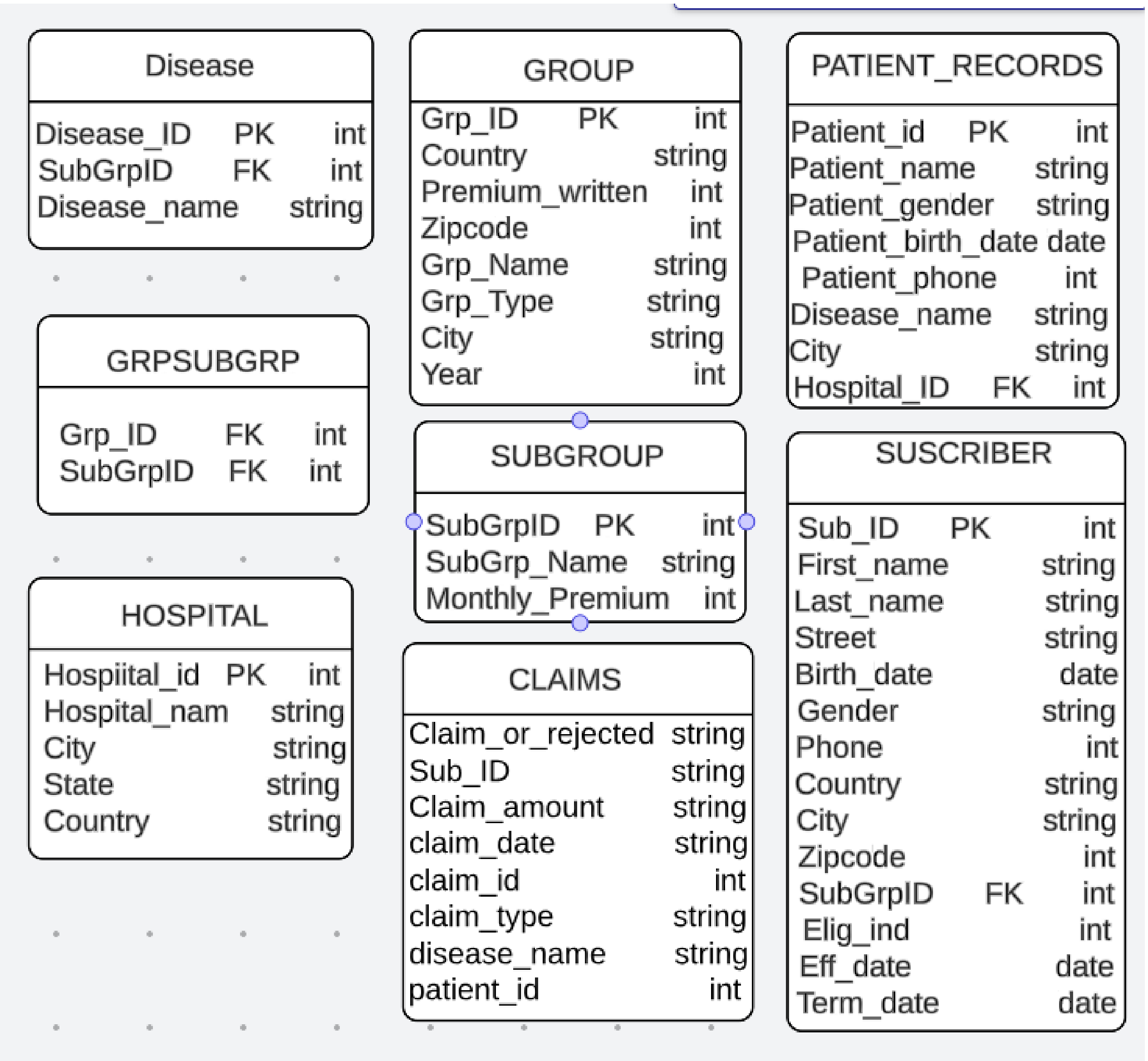
PATIENT_RECORDS (Patient_id **PK**, Patient_name, Patient_gender, Patient_birth_date, Patient_phone, Disease_name, City, Hospital_ID **FK**)

SUBGROUP (SubGrpID **PK**, SubGrp_Name, Monthly_Premium)

SUSCRIBER (Sub_ID **PK**, First_name, Last_name, Street, Birth_date, Gender, Phone, Country, City, Zipcode, SubGrpID **FK**, Elig_ind, Eff_date, Term_date)

CLAIMS (Claim_or_rejected, Sub_ID, Claim_amount, Claim_date, Claim_ID, Claim_type, disease_name, Patient_id)

b. ER Diagram



Disease		
Disease_ID	PK	int
SubGrpID	FK	int
Disease_name		string

GRPSUBGRP		
Grp_ID	FK	int
SubGrpID	FK	int

HOSPITAL		
Hospiital_id	PK	int
Hospital_nam		string
City		string
State		string
Country		string

GROUP		
Grp_ID	PK	int
Country		string
Premium_written		int
Zipcode		int
Grp_Name		string
Grp_Type		string
City		string
Year		int

SUBGROUP		
SubGrpID	PK	int
SubGrp_Name		string
Monthly_Premium		int

CLAIMS		
Claim_or_rejected		string
Sub_ID		string
Claim_amount		string
claim_date		string
claim_id		int
claim_type		string
disease_name		string
patient_id		int

PATIENT_RECORDS		
Patient_id	PK	int
Patient_name		string
Patient_gender		string
Patient_birth_date		date
Patient_phone		int
Disease_name		string
City		string
Hospital_ID	FK	int

SUSCRIBER		
Sub_ID	PK	int
First_name		string
Last_name		string
Street		string
Birth_date		date
Gender		string
Phone		int
Country		string
City		string
Zipcode		int
SubGrpID	FK	int
Elig_ind		int
Eff_date		date
Term_date		date

4. Technologies and Platforms to be used.

Listed are the technologies and platforms to be used in the project.

AWS: Amazon Web Services is a cloud computing platform.

S3: It is an object storage service provided by AWS.

Redshift: It is a data warehouse services managed by AWS.

Databricks: It is a unified data analytical platform for big data and AI.

PySpark: Apache spark is a python API used for processing large datasets.

Jira: It is a project management tool.

GitHub: It is a platform for hosting and collaborating on Git repositories.