Health Care Insurance Data Analysis (Solution Design Document)

Solution

Data ingestion and preparation:

- Utilize databricks on the Community Edition to ingest competitor company sample data via the catalog.
- Implement data validation and initial cleaning using PySpark DataFrames to handle null values and duplicates.

Data transformation and analysis:

- Apply PySpark transformations to clean and preprocess datasets (Patients_records.csv, subscriber.csv, claims.json, grpsubgrp.csv) to ensure data quality and consistency.
- Use PySpark and Python for DataFrame operations to perform necessary analytics (e.g., identifying diseases with the maximum claims).

Data storage and result compilation:

- Develop PySpark jobs to save cleaned and transformed data into Databricks File System (DBFS) within Databricks.
- Generate separate PySpark scripts for each analytical query or requirement to create structured outputs.

Use Cases

- Healthcare insurance policy analysis based on customer behavior.
- Personalized offer customization based on customer demographics.
- Revenue enhancement through targeted marketing strategies.
- Risk assessment and profitability analysis of insurance policy groups.
- Regulatory compliance monitoring and reporting.

Database Design

Tables and Their Relationships: Cleaned Datasets

Patients_records.csv

Columns:

sub_id (PK), first_name, last_name, Street, Birth_date, Gender, Phone, Country, City, Zip Code, Subgrp_id (FK references grpsubgrp.csv (SubGrp_ID)), Elig_ind, eff_date, term_date

Primary Key: sub_id

Foreign Key: Subgrp_id references grpsubgrp.csv (SubGrp_ID)

claims.json

Columns:

claim_id (PK), patient_id (FK references Patients_records.csv), disease_name, SUB_ID (FK reference subscriber.csv), Claim_Or_Rejected, claim_type, claim_amount, claim_date

Primary Key: claim_id

Foreign Keys:

patient_id references Patients_records.csv

SUB_ID references subscriber.csv(sub_id)

subscriber.csv

Columns:

sub_id (PK), first_name, last_name, Street, Birth_date, Gender, Phone, Country, City, Zip Code, Subgrp_id (FK references grpsubgrp.csv (SubGrp_ID)), Elig_ind, eff_date, term_date

Primary Key: sub_id

Foreign Key: Subgrp_id references grpsubgrp.csv (SubGrp_ID)

grpsubgrp.csv

Columns:

SubGrp_ID (PK), Grp_Id

Primary Key: SubGrp_ID

Foreign Key: None

Technologies and Platforms

- Databricks (Community Edition for development and testing)
- PySpark (for data processing and analytics)
- Python (for scripting and data manipulation)
- GitHub (for version control and collaboration)