**Healthcare Insurance Data Pipeline for Competitive Analysis and Revenue Enhancement**

**Solution Overview**: The primary goal of this solution is to create a collection of data pipelines using the Big Data Ecosystem to increase the revenue of a Health Care Insurance firm by monitoring client behavior. This study will aid in tracking consumers' situations, allowing the corporation to deliver customized offers and earn royalties based on previous policy subscriptions**.**

**Use Cases:** The solution will address the following use cases:

* The health issue with the most claims.
* Members of any grouping who are younger than thirty
* Maximum Subgroups in Groups
* Hospitals with the Highest Patient Volume
* Subgroups Most Frequently Subscribed
* The total number of denied claims
* Cities with the Most Charges
* Subscriptions to Government vs. Private Policies
* Average Monthly Subscription Premium Paid by Members
* The most profitable policy group
* Admissions of Cancer Patients Under the Age of 18
* Insurance Without Cash Patients with Over Rs. 50,000 in Claims
* Female Patients Having Knee Surgery Over 40

**Database Design:** Tables and Relationships

* Patients Table
* Primary Key: patient\_id
* Columns: patient\_id, patient\_name,patient\_gender, patient\_birth\_date, patient\_phone
* Subscribers Table
* Primary Key: subscriber\_id
* Foreign Key: patient\_id(relates to patient table)
* Columns: subscriber\_id,first\_name,last\_name,city

**Technologies and Platforms:** The solution will handle and analyze healthcare insurance data efficiently by combining sophisticated technologies and platforms. AWS S3 will act as the data storage layer, allowing for the input and storage of massive amounts of raw data. AWS Redshift will be used for scalable data warehousing, allowing for rapid querying and reporting on processed data. Databricks, a unified analytics platform, will be utilized to run PySpark jobs for data transformation, purification, and analysis, creating an environment that is ideal for big data processing. Jira will make project management easier by assisting in the organization of sprints, tasks, and user stories for effective execution, while AWS EMR Studio will offer extra support for carrying out distributed data processing operations in the cloud. Lastly, PyCharm and GitHub will be utilized for codebase development collaboration and version control, guaranteeing seamless code integration and administration across the project lifecycle. Scalable processing, smooth integration, and productive teamwork will all be made possible by this technology stack.