De-identifying PII Data

De-identification of data in bioinformatics removes or obscures personally identifiable information (PII) from datasets in research. This protects the privacy of individuals whose data is being used and complies with legal and ethical requirements.

Why De-identification Matters :

Protects Individual Privacy

Ensures that individuals cannot be identified from the data being used in research, protecting their privacy and preventing potential harm or discrimination.

Legal And Ethical Compliance

Many countries have laws and regulations requiring de-identifying personal data, such as the General Data Protection Regulation (GDPR). Failure to comply with these laws can result in legal and financial penalties.

Facilitates Data Sharing

De-identified data can be shared more easily and widely, enabling collaboration between researchers and institutions to accelerate scientific progress.

Improves Data Quality

Removing PII from datasets can improve data quality, reducing the risk of errors or biases arising from personal information.

Enables Secondary Use Of Data

De-identified data can be used for secondary purposes beyond the original study, such as meta-analyses, systematic reviews, and machine learning applications.

The Backstory

A global life sciences company wanted to implement a comprehensive de-identification strategy to protect patient privacy and enable secure data exposure. At the same time, they had to develop an end-to-end solution that could automate the entire ETL process from data ingestion to transformation and loading. The company required a robust de-identification of PII data involving multimodal pipelines that could handle structured clinical data in BigQuery and raw molecular data.

The Challenge

De-identification is not foolproof. It can result in a loss of valuable information that could be used in future studies. Plus, there is no consensus on de-identification standards or best practices so different methods can create inconsistencies in the data. Moreover, de-identification can be computationally complex when dealing with large datasets.

In this case, the life sciences company faced the challenge of combining and creating a longitudinal patient dataset for research use cases. In addition, the large data sets came in different formats, such as Bam, Vcfs, FastQ, and CSVs.

The Solution

The life sciences company needed to de-identify and ingest multimodal data in a highly regulated environment. Our team implemented a modern data architecture known as data lake house on Google Cloud to enable that. The architecture consolidated data in a centralized data lake using Google Cloud Storage (GCS). After that, our team used a BigQuery data warehouse for data processing and analysis

The Results

The pipeline's overall performance and reliability increased, making it a more effective de-identification and ingestion tool. We shortened the ingestion pipeline, which resulted in a 30% faster run time. The refactored pipelines now ingest over 100,000 files, and the refactored code is more reliable.