# The task at Hand:

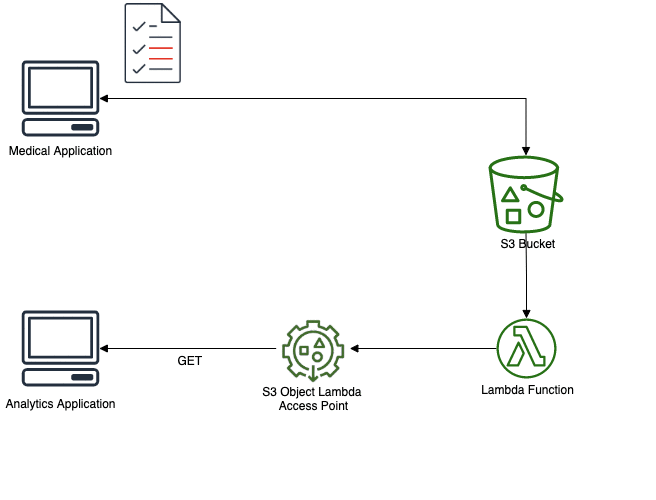
Assuming that I am working for a hospital, I have a lot of patient data with me that needs to be analysed. This data can be used to train complex machine learning models, for evaluation purposes and also understand more about the patients. However, before handing off the data to the analysts or developers, the middle man needs to come in and make sure that no sensitive information from the data is being leaked since the data might have the patient's name, address etc. The goal of the project is to anonymize the data such that no sensitive information of the patient is leaked and or can be traced back to the user AND making sure that the data is still statistically significant to draw conclusions from it. Once the solution has been implemented, the solution needs to be deployed on aws along with the database and end to end pipeline solution needs to be presented that is both effective and cost efficient.

Note: We already know the schema of the data and the format of the data is not going to change, we can expect new data but even that will follow the schema

## Proposed solution 1:

### Data Anonymization using s3 object Lambda

Technical diagram



Explanation

One of the first things that came to my mind when I heard of databases was s3. S3 is an AWS service which is an object based storage system which has a 5TB limitation for storage, hence it is virtually

Why it wont work/ feedback from the mentor

Proposed solution 2:

Diagram

Explanation

Why it wont work

Proposed solution 3

Diagram

Explanation

Why it will/wont work

Components of the diagram

Tech stack