**4 SYSTEM ANALYSIS**

**4.1 PROBLEMS AND WEAKNESSES OF THE CURRENT SYSTEM**

Not a lot of work has been open-sourced for the general public to use and hence there is little to no awareness about the dangers of a data breach and how to protect your personal data online. Among the commonly prevalent data anonymization techniques such as roman cipher or substitution, masking and randomization, there a wide range of disadvantages:

* They can easily be decoded as they are relatively simple algorithms
* They can not be used for anonymization of sensitive data as they do not take quasi-identifiers into account
* The masked or randomised data loose their statistical significance as they can not be used as meaningful data

Although more sophisticated algorithms exist, they are mostly closed source and require significant time and or compute. However, several open source alternatives such as ARX often have limited capabilities and are often outdated

**4.2 REQUIREMENT OF NEW SYSTEM**

For the client, the patient's data holds the utmost priority as that data should not be seen by the malicious hackers and the technologies being used should be feasible and scalable without little to no overhead and should require little to no manual configuration. Clients also want complete control over their data and explicitly request that their data does not leave their system. Hence, it made the most sense to design the given solution

**4.3 SYSTEM FEASIBILITY**

**4.3.1 Economical Feasibility**

Economical feasibility of the system refers to the analysis of the finances of the current system. Whether or not after all costs considered the solution will be economically feasible or not. The beauty of the given solution is such that much of it’s feasibility lies with the user and the features that they want to use. E.g if the user is an avid snowflake user with large amounts of data in there, standard snowflake costs will apply regardless. However, the system is a lot more economically feasible E.g We convert all floating point values to Integers and hence the storage required for that column will be effectively reduced in half

**4.3.2 Feasibility with other systems**

The current solution is especially feasible with other systems as it supports a variety of sql databases and has snowflake integration support thus making it really easy for the users to import and export data. Moreover, it stores all the data at the grassroot level in the form of a pandas dataframe hence there is no need for conversion when using the data for analysis

**4.3.3 Research Feasibility**

The given system implements the data anonymization pipeline using the state-of-art methods and all of the implementation is open source meaning the users can add modify, improve and contribute to the development of the algorithms

**4.4 HOW DOES THE SYSTEM FITS IN THE CURRENT OBJECTIVE OF THE ORGANISATION**

The current system integrates really well with the overall objective of the organisation as this solution can be used by managers and the people at the upper management to preserve anonymity when asking for feedback and constructive criticism. Moreover, it can also be used to release the impressive feats achieved by the company without breaking the company-client confidentiality



Fig 4.4 A possible number of use cases of the current system with the objective of the organisation

**4.5 IMPLEMENTATION OF THE SOLUTION USING PRESENT SYSTEM AND TECHNOLOGIES**

The system is implemented using python, snowflake, powerbi, Minds Db and SQL

* Python: The reason for choosing python over any other languages such as Java or C++ was because the client has all of their systems running in python and hence it provides seamless integration. Morvoer, a lot of Machine Learning and analysis has been traditionally based on python hence designing the system in python would reduce the learning curve required for anonymizing the data
* Snowflake: Snowflake is a cloud as a service platform which provides data warehousing capabilities. A lot of organisations use snowflake these days and it supports a variety of data types and data integration platforms
* POWERBI: PowerBI is a tool by microsoft that is used to visualise the data automatically. With the help of power BI, we can generate beautiful charts and visualisations that would have been difficult otherwise
* Minds DB: Minds DB is a coming of age database solution that takes in data in the form of “AI Tables” and allows you to use pretty much any machine learning algorithm using a simple SQL query. Integration of this can be used potentially to automate the machine learning modelling process as well and allowing the developers to spend time on the things that really matter

**4.6 FEATURES AND FUNCTIONALITY OF THE SYSTEM**

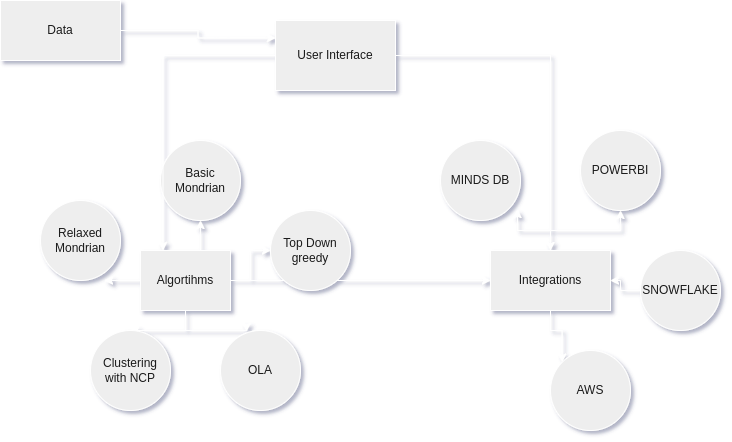


Fig 4.6 The flow and features of the system

Here are some of the salient features of the system:

* Is versatile, can be recalibrated to suit your business needs
* Supports a variety of analysis tools
* Stores and uses data in standard formats such as SQL tables and pandas dataframes
* Is open source
* Can be tweaked to meet a variety of data needs
* Can be extended to support a variety of services and softwares