Multistage Cryptonet

Secure anonymous data analytics on the cloud

Gowtham, Naveen, Abhishek, Tripti

Problem Statement

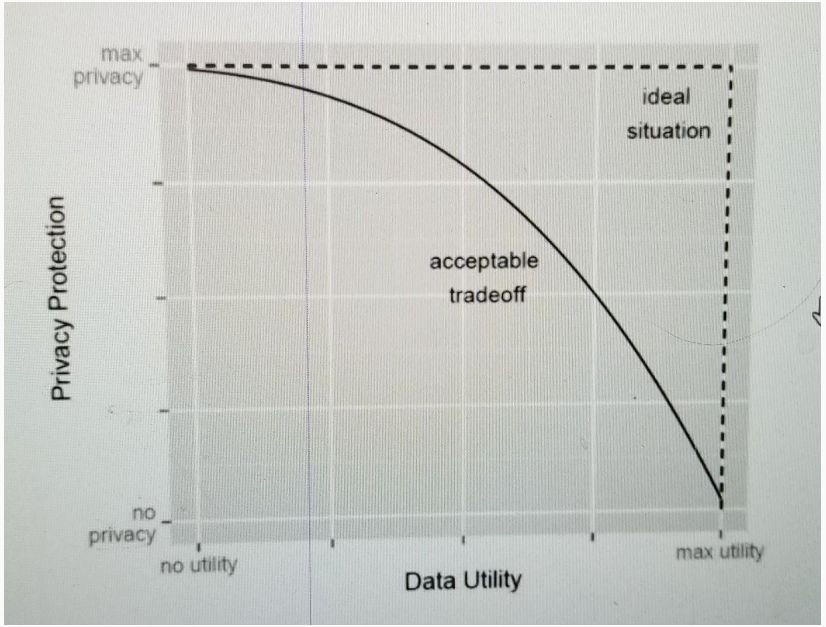
How do we enable secure and anonymous sharing of sensitive data?

-

Methodologies Explored

Homomorphic Encryption

Multistage Cryptonet



How we Solve this problem?

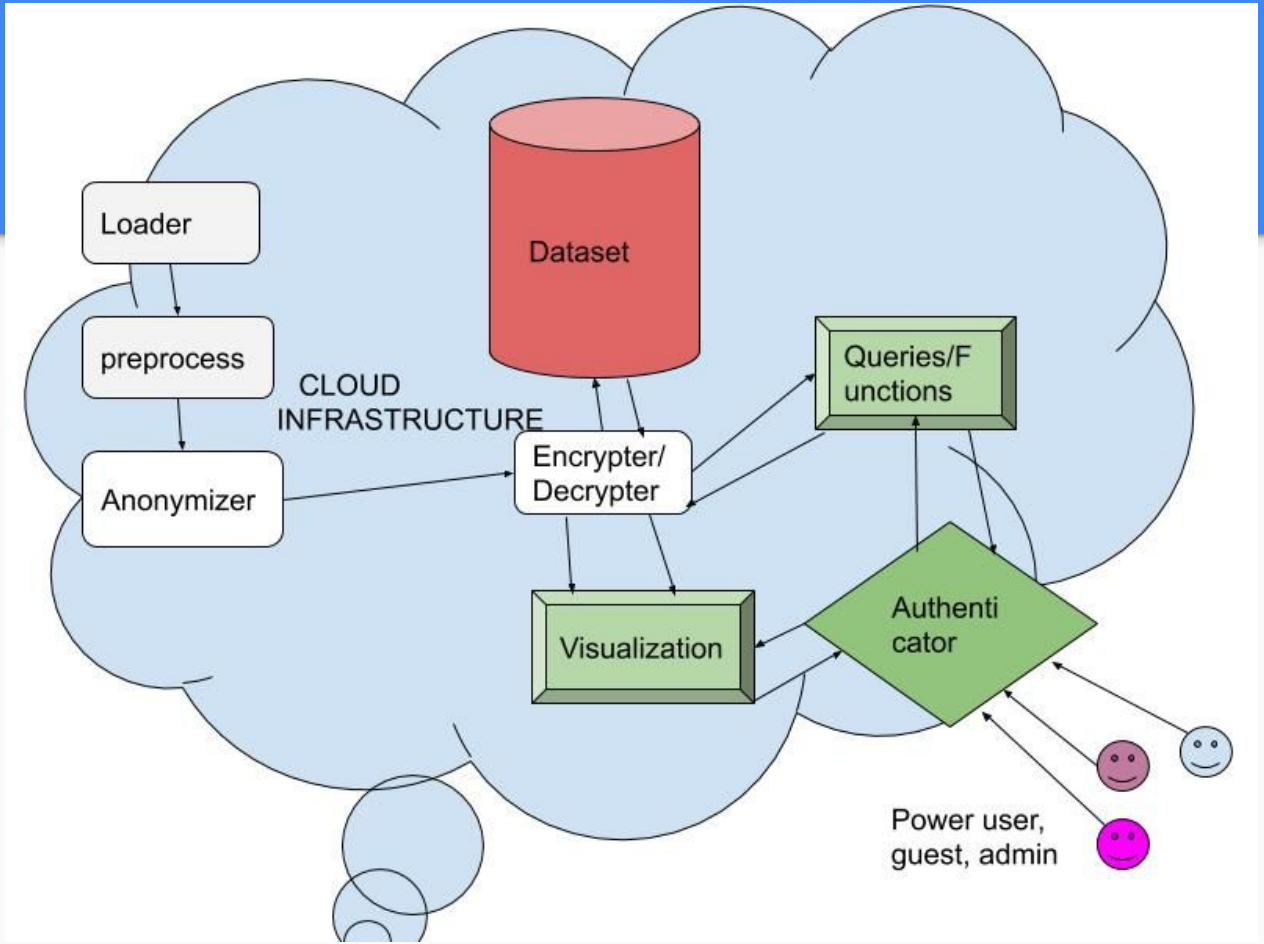
We don’t share the data, Researchers use **cryptonet solution** to perform operations. Instead of sending the data we ask users to work on the platform.

Contributors directly submit their data through load module.

Researchers can access data indirectly by using visualization and Functions module.

Architecture

Multistage Cryptonet



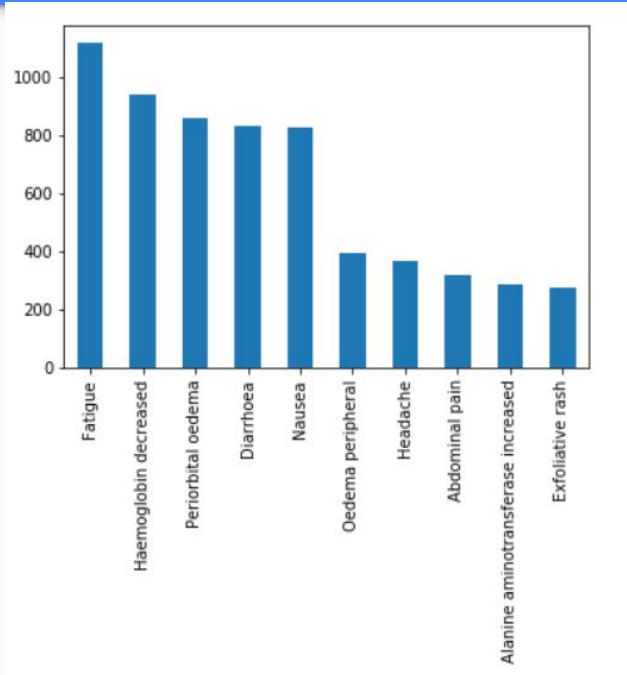
Modules Load, Preprocess, Anonymizer

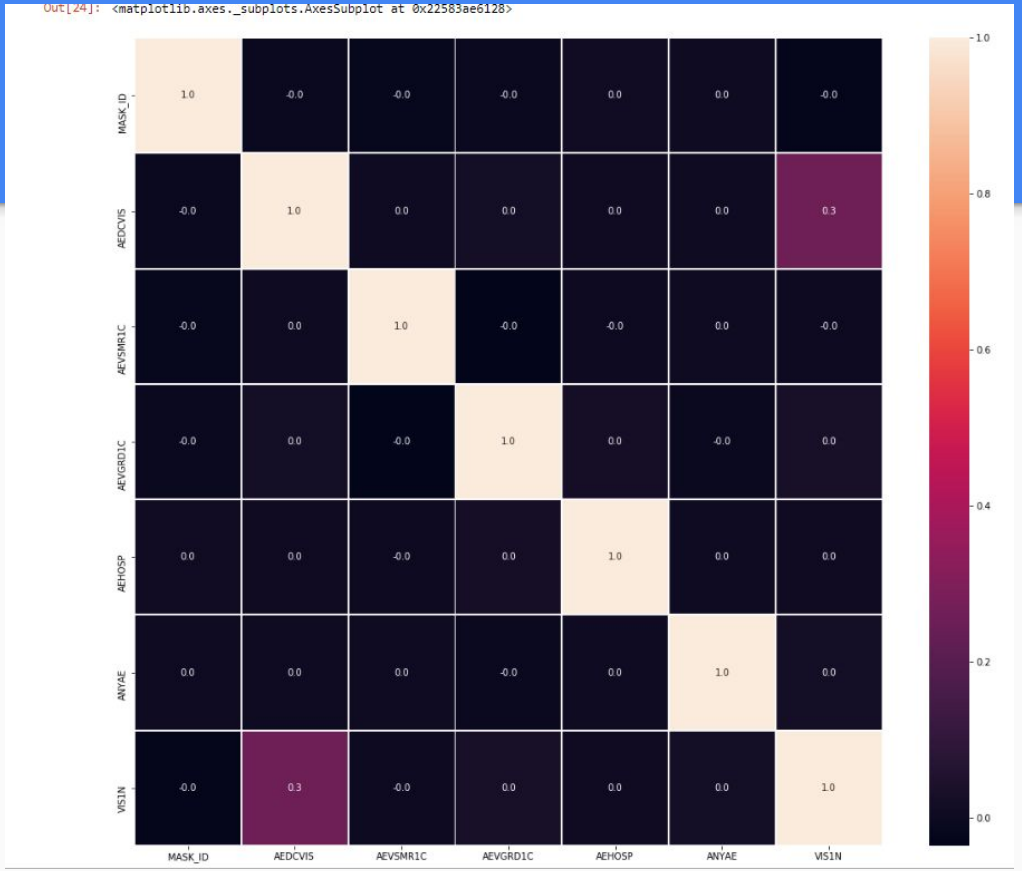
Load module is used to ingest via csv, json, parquet etc data formats. T(C) O(n)

Preprocessing module is used to impute missing values, generate synthetic samples(SMOTE), detect and remove outliers etc T(C) O(n3)

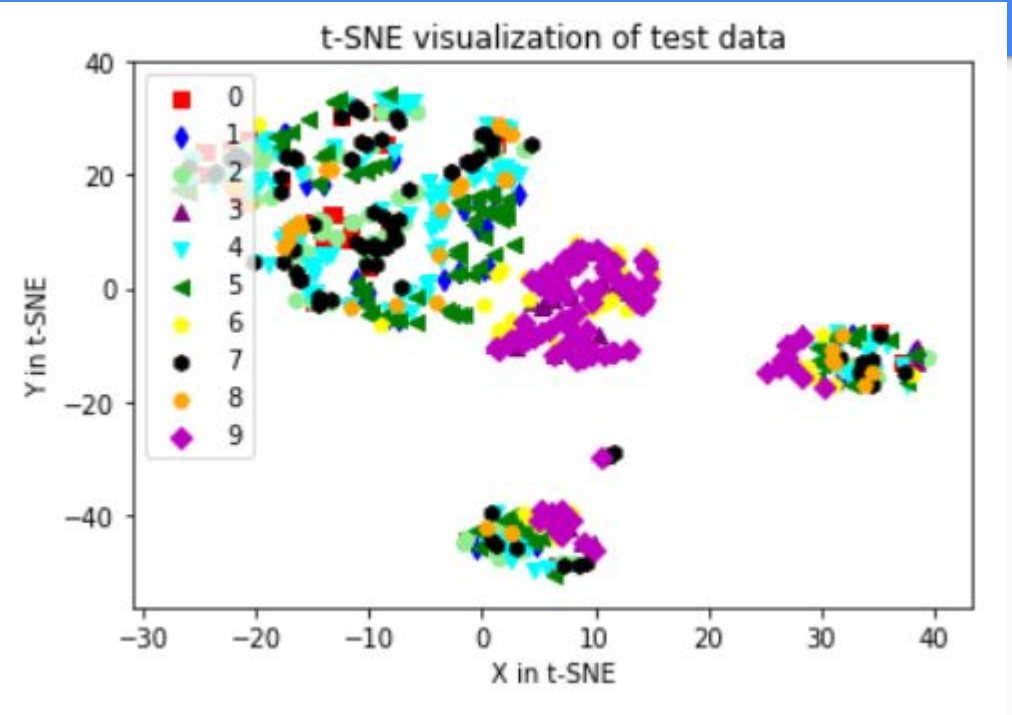
In the Anonymizer module we use Faker to generate phone numbers, names, SSN etc. The semantic relationship is always maintained. T(C) O(n2)

Visualization





Visualization



Encryption/Decryption

Bcrypt is used for encryption and decryption.Hashing with salt padding is used for added security.

Use homomorphic encryption techniques to encrypt the underlying dataset. Data is encrypted in the source (cloud) and all operations run on the data without decrypting.Mitigates data misuse and privacy concerns at source (cloud). Involves trading performance for security / risk free privacy

In cloud operations platform automatically translates data operations to homomorphic operations and decrypts the result.

Example 5 + 6 = 11, 5 -> e, 6 -> f, e + f = o , o ->11 (response)

Authenticator

Researchers are granted access through multi-factor authentication to the data visualization and data operations platform.

Secure Operations : Researchers send requests to perform operations on data and get responses from the cloud without ever seeing the underlying data.

Policy : Data at source (cloud) is safe, enforced through data privacy policies

Time Constrained token access key

Authenticator

Our proposed cloud based solution has two levels of risk mitigation :

Level 2 - Hides underlying dataset from researchers but displays what the legends are.

Level 1- Hides underlying data set AND Legend from users. Suitable for high risk applications like rare disease research etc. Involves tradeoff between performance and risk free privacy.

References

1] Smote: Synthetic Minority Over-sampling Technique

N. Chawla-K. Bowyer-L. Hall-W. Kegelmeyer -

https://jair.org/index.php/jair/article/view/10302

2] Microsoft Seal: Fast and Easy-to-use Homomorphic Encryption Library

https://www.microsoft.com/en-us/research/project/microsoft-seal/

3] Bcrypt https://www.npmjs.com/package/bcrypt

Thank You !

Questions ?