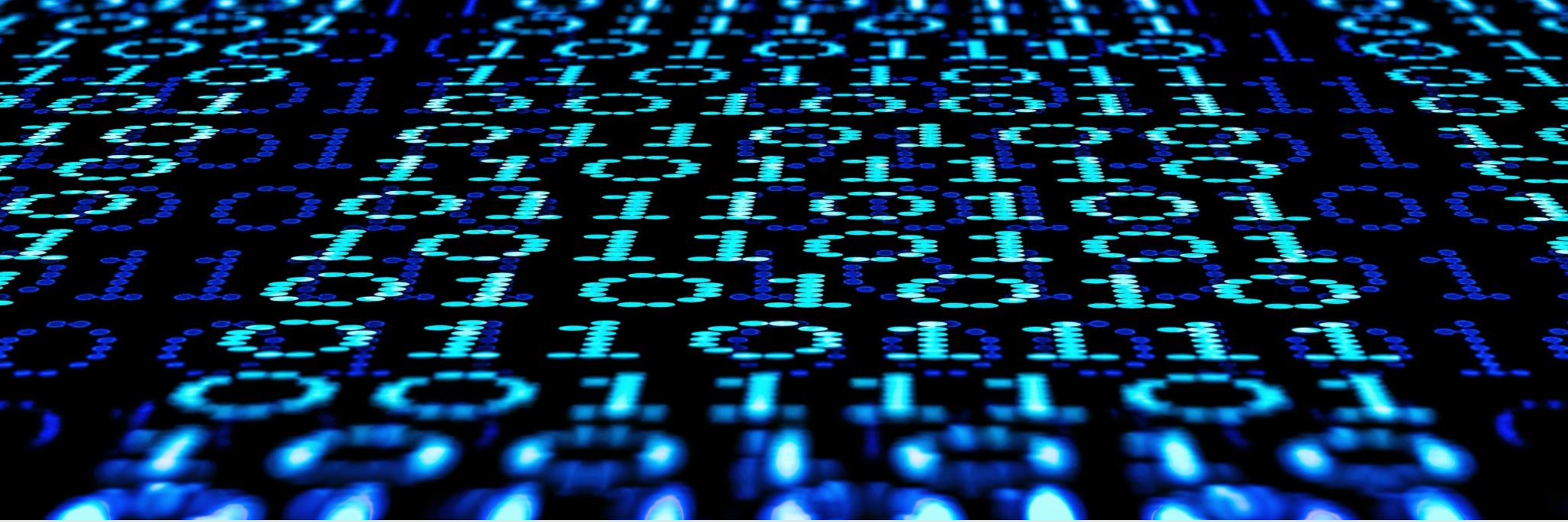


Privacy shield –The Data Anonymizer

Data Anonymizer

- Overview : This document contains 5 sections, section-1 is about introduction to the project, section-2 and 3 contains the design and development of project, section-4 contains the user manual and section-5 contains future work

1. Introduction

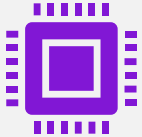
1.1. Context

The digital world of now is composed of personalized data services and there is a huge demand for data, especially for personal data

Large volumes of data can now be employed quite efficiently to gain new insights on a certain phenomenon, and data mining algorithms can give a lot of information about persons, events or entities

The purposes for which one requires meaningful personal data are very diversified, for example, research or public health policy purposes, to develop new services and products, to enhance the efficiency and effectiveness of new drugs, or even to simply condition people's behaviour, when they make a purchase on an online store

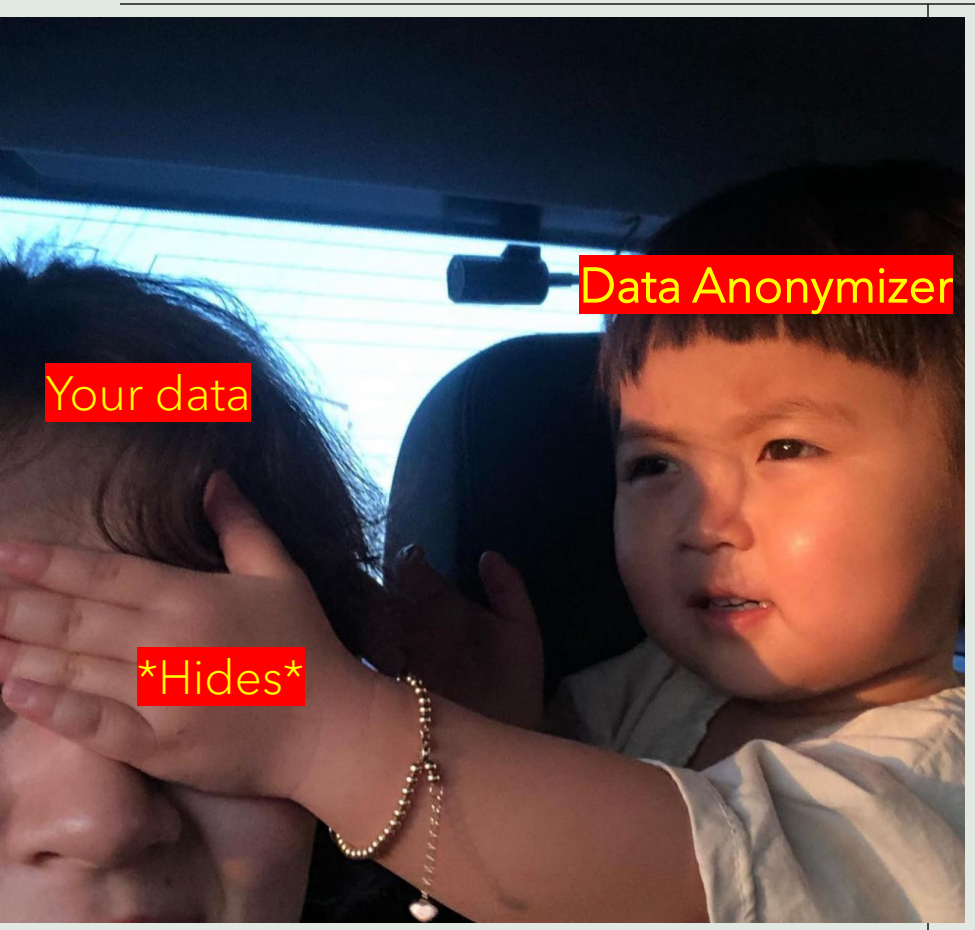
1.2. Motivation



In recent years, escalation of technology led to an increase in the capability to record and store personal data about consumers and individuals

To mitigate these issues, some de-identification methodologies have recently been proposed that, in some well controlled circumstances, allow for the re-use of personal data in privacypreserving ways

We will present a web tool that provides a way to easily anonymize data



Anonymization

Anonymity, means simply that a person is not identifiable

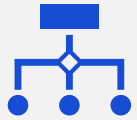
The anonymization process is intended to irreversibly remove the association between an individual and some information that can identify this person

Personal data is anonymized to protect the privacy of subjects when storing or disclosing data

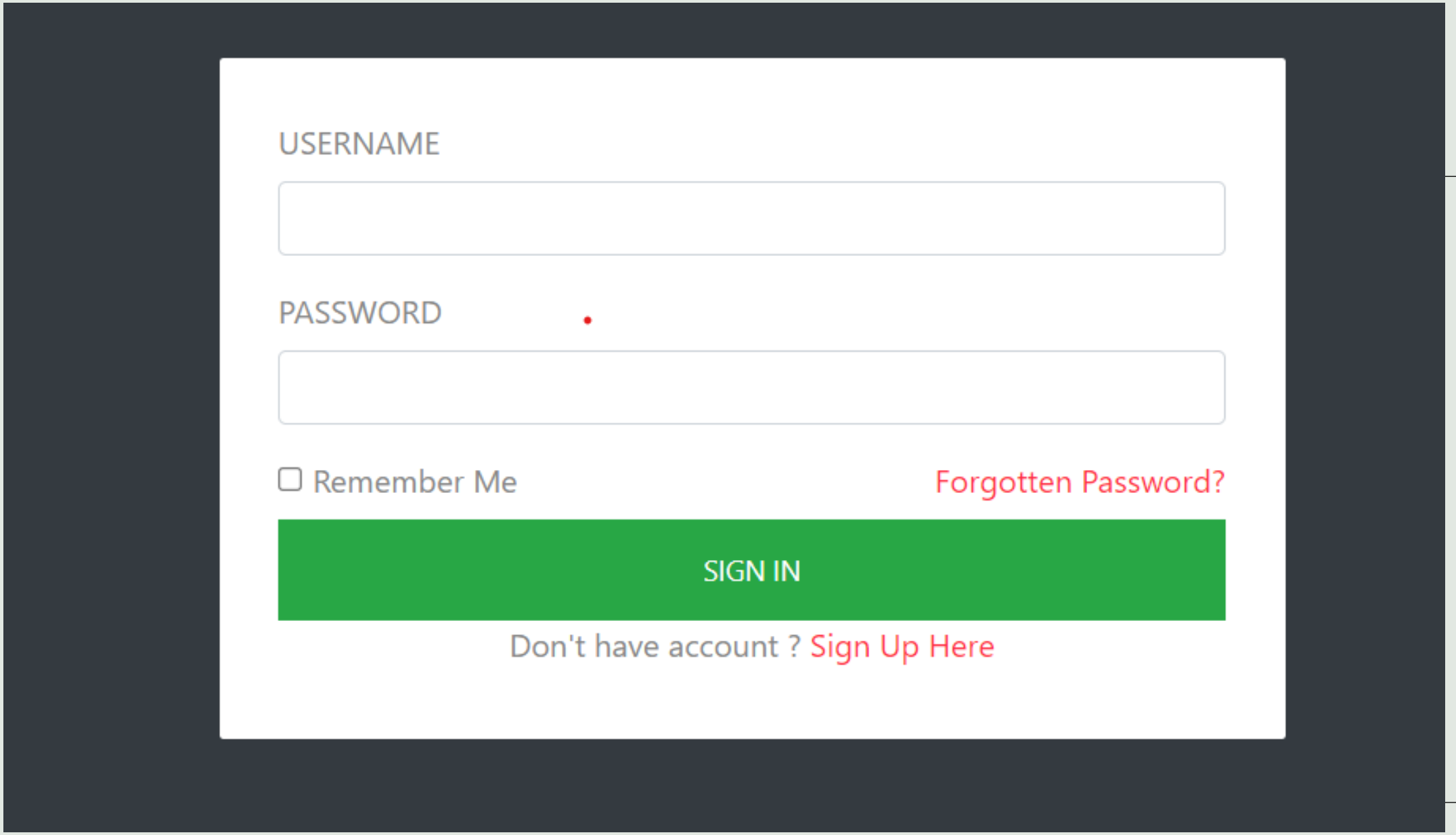
1.3. Scope of Project

The main objective of this project is to create web application which make it easy for a user to anonymize the data

It means the user can upload the dataset and can choose configuration or create a new configuration with in the web interface for de-identification, after that the de-identified or anonymised data can be downloaded



The important task of this project was to create an interface for the user such that a new configuration can be developed on the dataset

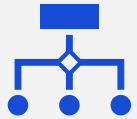

USERNAME
PASSWORD
Remember Me
Forgotten Password?
SIGN IN
Don't have account ? Sign Up Here

Design and Development

2. Architecture

In order to give the web application some structure, a layered architecture was created, grouping the different components in sections

This layered separation is often used on web platforms, on this particular architecture is separated in three layers



Each of these layers is responsible for a particular and unique processing, they communicate with each other through different frame works and languages, so, aggregating all those tiers in a single framework simplifies hugely this interconnection process

Presentation Layer

This layer contains the user oriented functionalities, they are responsible for managing user interaction with the system, and generally consists of components that provide a common bridge to the business layer core

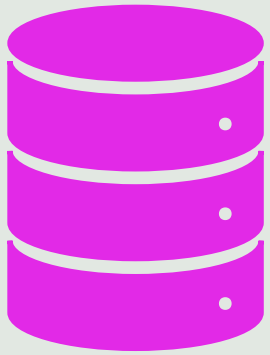
Presentation tier components allow users to interact with the application

Business Tier

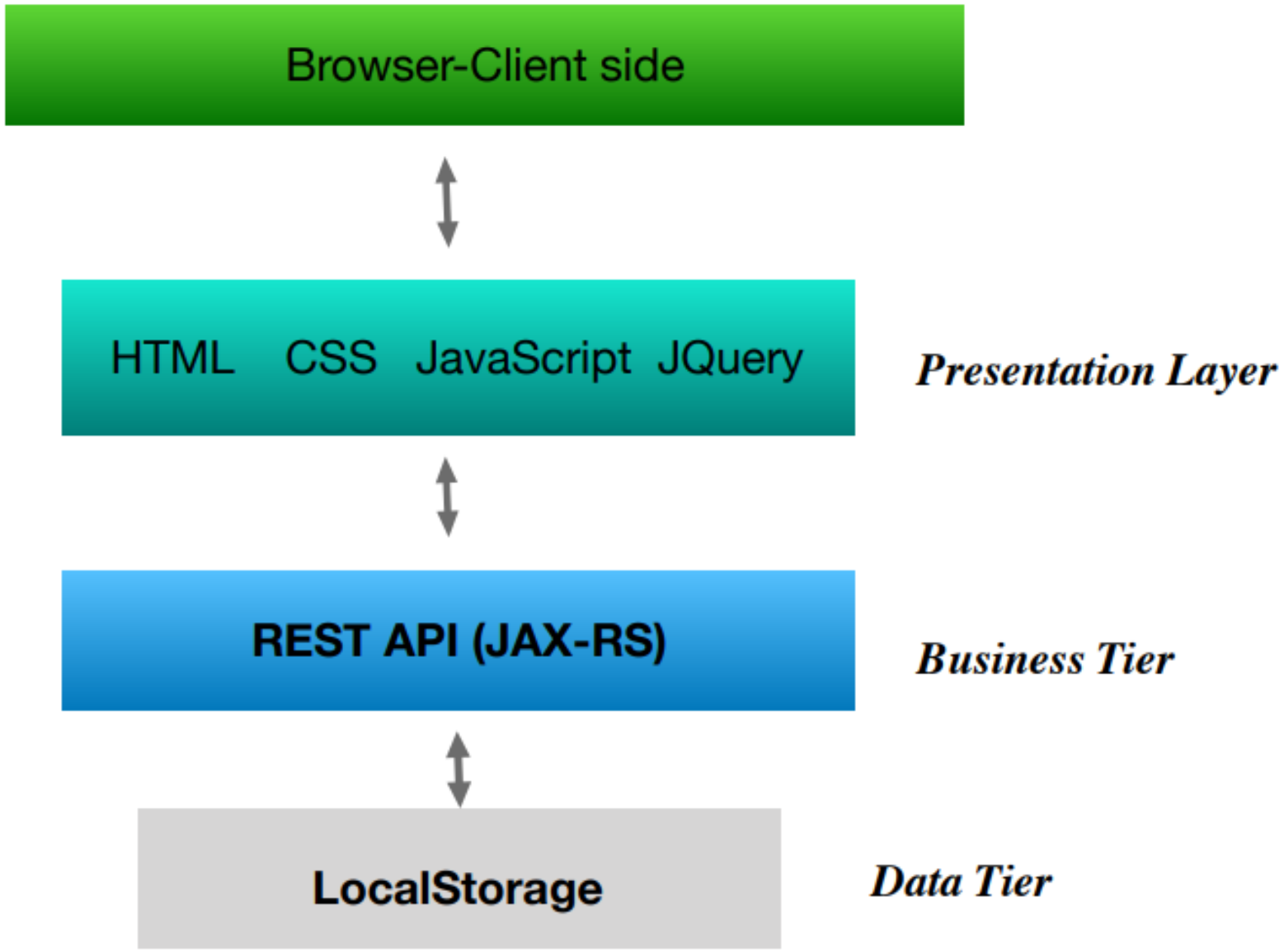
This layer implements the core functionality of the system and encapsulates the relevant business logic

It generally consists of elements that execute rules and separates them from the user interface and data access

Data Tier



The data tier provides access to the databases and storage devices used by a three-tier applications. It is responsible for retrieving data and transforming into a suitable format for the rest of the application. Essentially, this layer stores and retrieves information to the business tier for processing and eventually to the presentation tier. On adapting the ARX API there was no need to use a database, because saving the files locally was sufficient enough to perform the anonymization.


Browser-Client side
HTML CSS JavaScript JQuery
Presentation Layer
REST API (JAX-RS)
Business Tier
LocalStorage
Data Tier

Technologies

Client-side

- In the application front-end, we used five technologies: HyperText Markup Language (HTML), Cascading Style Sheets (CSS), JavaScript, Jquery, and Asynchronous JavaScript and XML (AJAX). The web page presented on the user's browser was built on HTML, created from scratch, exhibiting a simple but intuitive interface that guides the user through a configuration process that sometimes could be a complex task. In order to make the HTML more attractive, we applied CSS, giving a better visual presentation to the interface.
- Another technology used in order to send and receive data from the server, is called JavaScript Object Notation (JSON), and is useful for parsing JavaScript objects into text. This is needed when the client wants to exchange data with the server because this data has to be in text format. JSON converts JavaScript objects into text, and vice-versa, the data is saved as pairs of keys/values and they are kept as an array of strings. This technology was used in all the AJAX requests created, parsing the data sent to the server, or making the reverse process for the received array of data.

Server-side

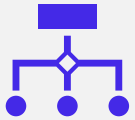
JERSEY :Jersey RESTful Web Services framework is open source, production quality framework for developing RESTful Web Services in Java that provides support for JAX-RS APIs

REST: A web application without resources don't serve for much, they are usually something that can be stored on a computer, e.g., an electronic document, a dataset or the result of executing an algorithm

Tomcat: To deploy the services and resources, apache tomcat is being used as server

ARX API

The ARX open source tool allows the user to change structured sensitive personal data, using Statical Disclosure Control , into data that can be shared securely



ARX API has been chosen for anonymization process because of its simplicity and organised code, with a good documentation, plus a deductive online javadoc with all packages and classes included in the API

The ARXConfiguration defines a group of settings that are sent to the ARXAnonymizer, this configuration allows multiple parameters, offering the user an opportunity to create suitable de-identified datasets

3.4. Project Folder Structure and Dependencies

The data that is uploaded to the server or data anonymizer will reside in a directory named Anonymization_resources of root folder

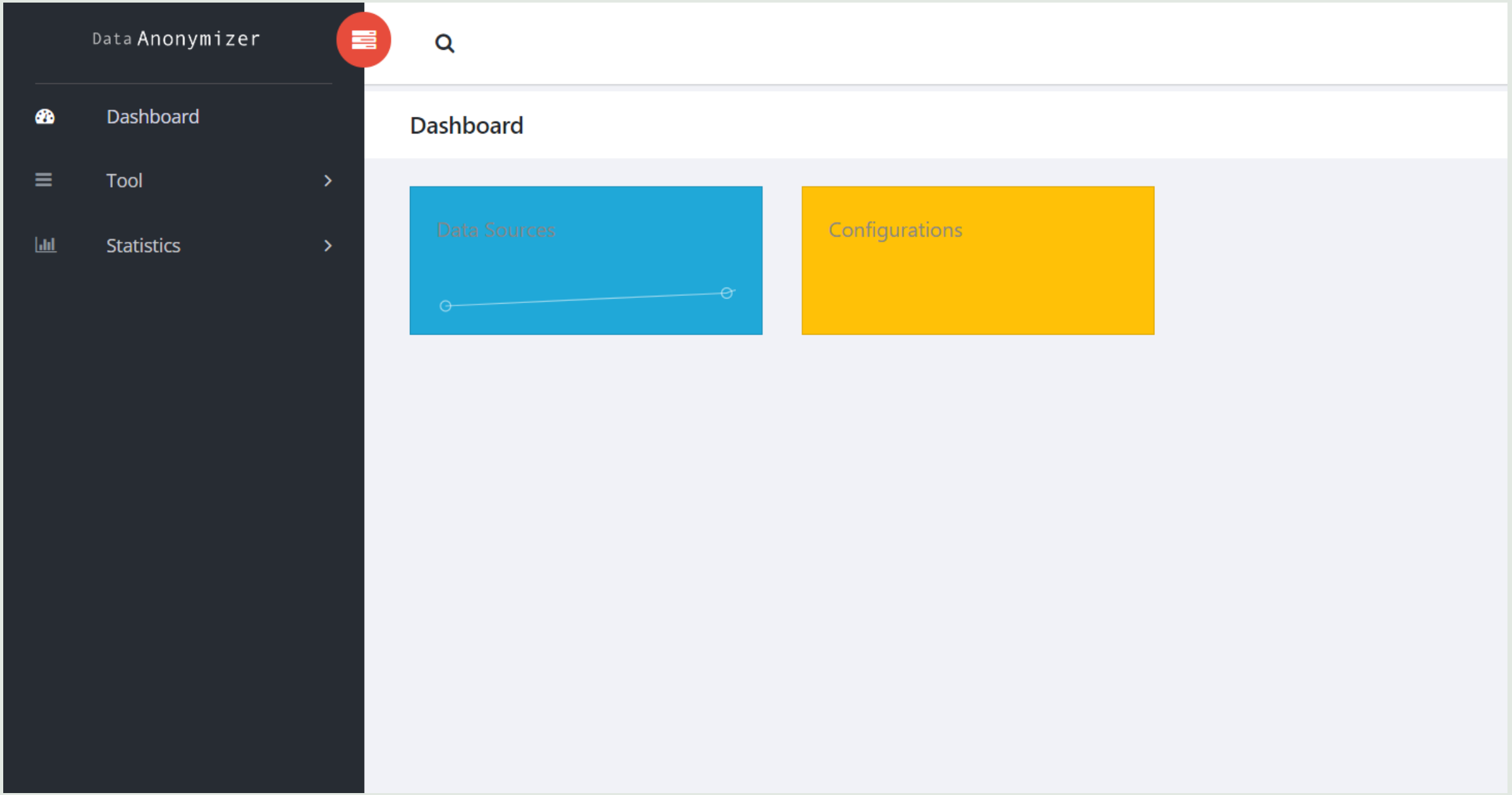
Inside the directory it again contains folders named data sources, configurations, hierarchies for their respective files

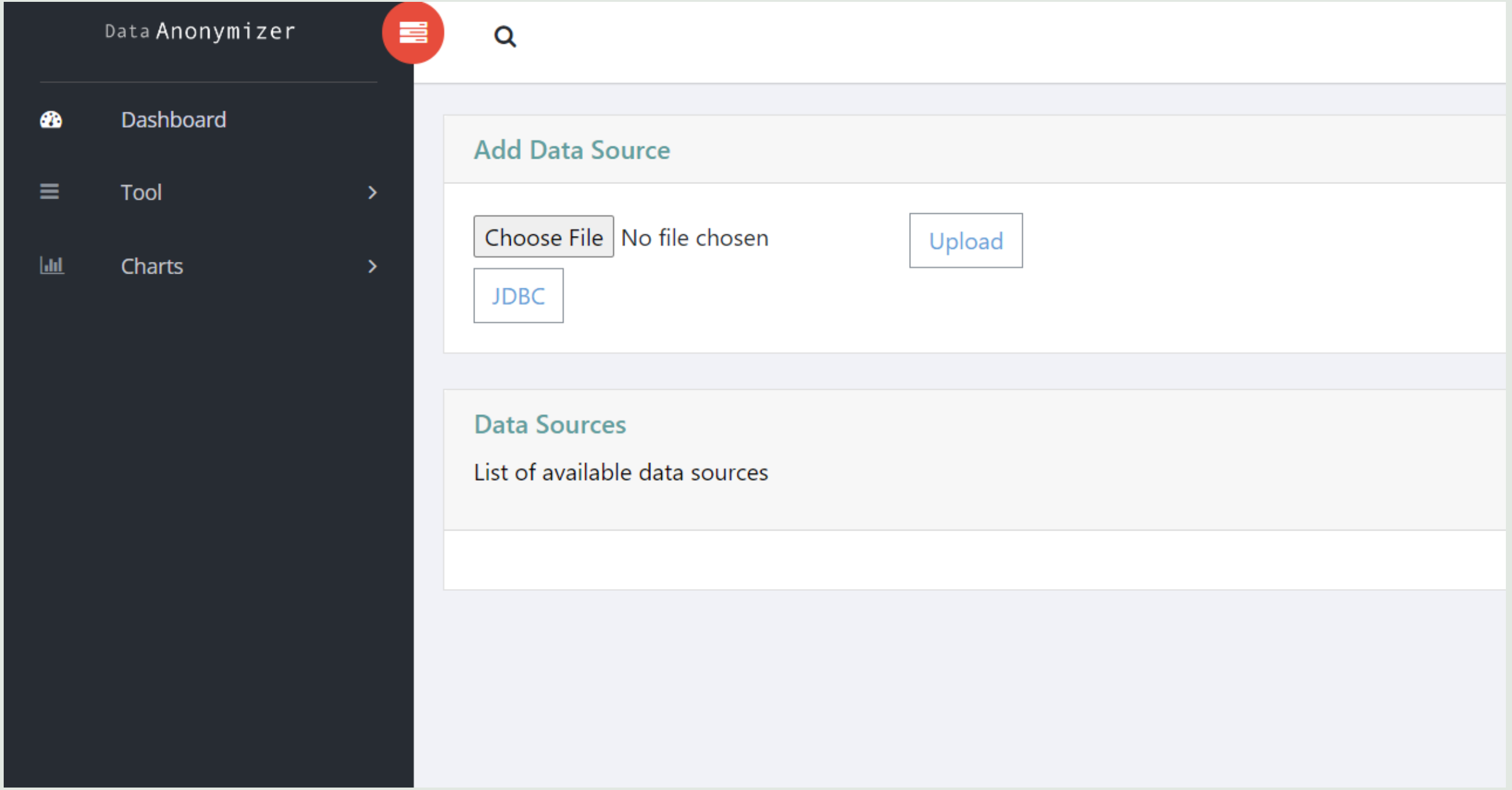
Also the output folder contains the recent anonymized file

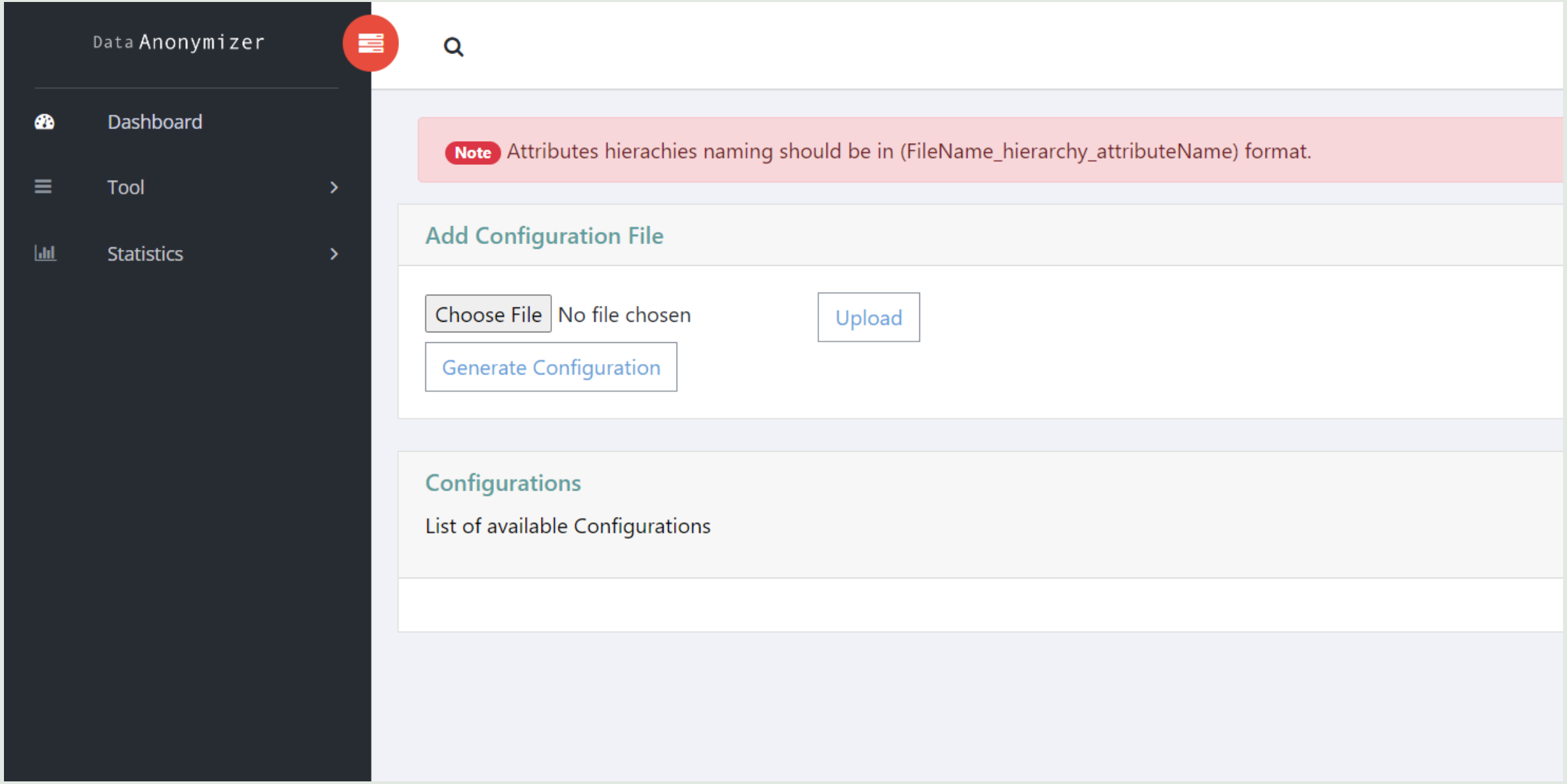
4. Interface/User Manual

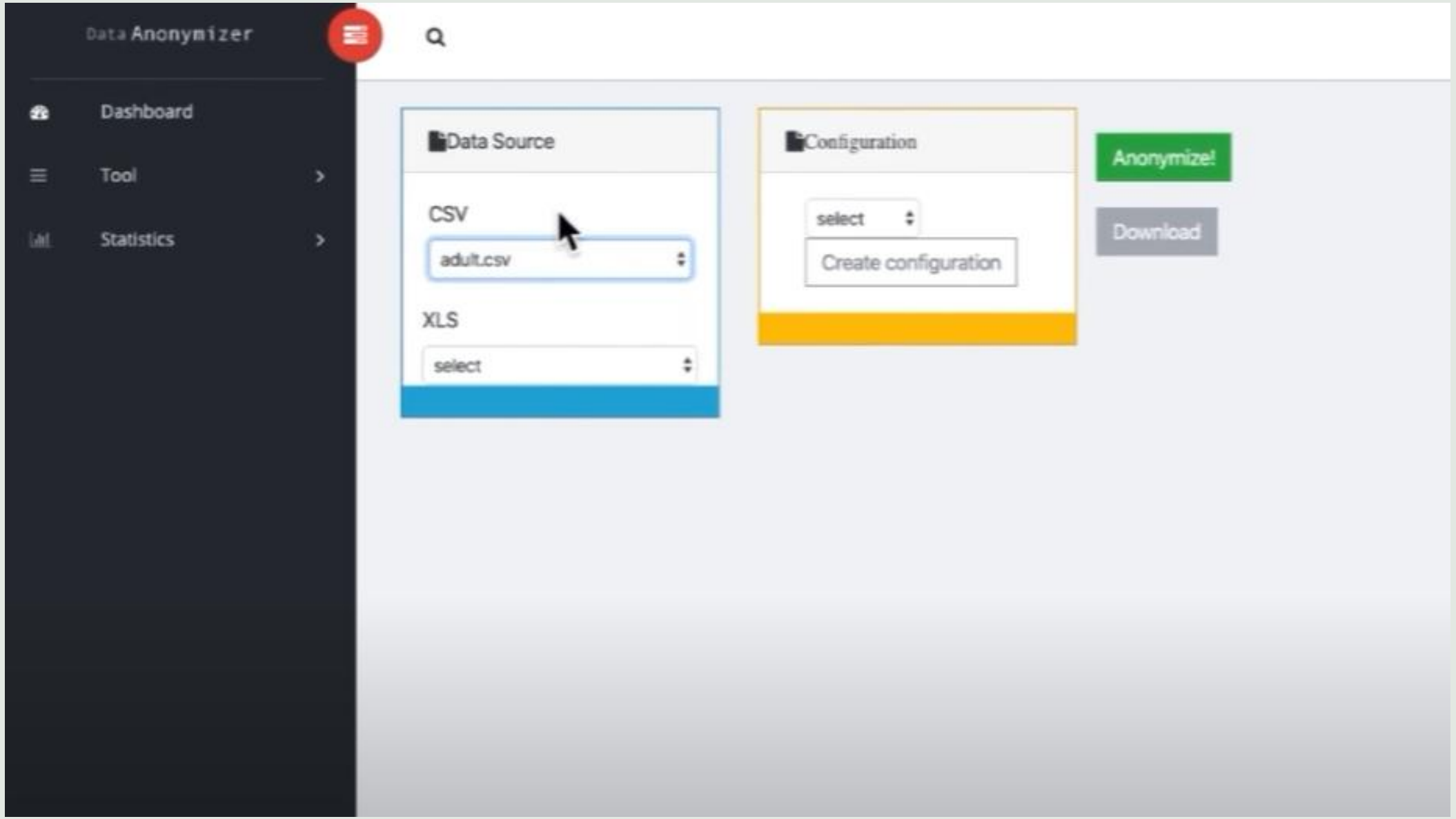
Data Sources: By clicking the data sources card the user will be redirected to the data sources page

Configurations: By clicking the configurations card from the index page, the user will be redirected towards configurations page


Data Anonymizer
Dashboard
Tool
Statistics
Dashboard
Data Sources
Configurations


Data Anonymizer
Dashboard
Tool
Charts
Add Data Source
Choose File
No file chosen
Upload
JDBC
Data Sources
List of available data sources


Data Anonymizer
Dashboard
Tool
Statistics
Note Attributes hierachies naming should be in (FileName_hierarchy_attributeName) format.
Add Configuration File
Choose File No file chosen
Upload
Generate Configuration
Configurations
List of available Configurations


Data Anonymizer
Dashboard
Tool
Statistics
Data Source
CSV
adult.csv
XLS
select
Configuration
select
Create configuration
Anonymize!
Download

4.3. Privacy Model

Typically most of the anonymization algorithms are based on Generalization and Suppression of QID attributes

They are

k-Anonymity

This algorithm requires that each QID tuple appear in at least k-1 records, this will ensure at minimal, that released data processed with k-anonymity will be difficult to re-identify

The QIDs contains information that is more likely to find over the dataset, so this type of attribute is more vulnerable to re-identification

k-anonymity uses generalization and suppression methods

l-diversity

This algorithm requires a high entropy on the distribution of SAs for each QID

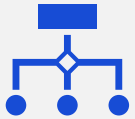
The main idea behind l-diversity, is the well balancing dispersion of SAs between all the groups included on the datasets

Vulnerability- Similarity attacks, that happens because it considers the diversity of SAs in the group, but it is not concerned with the semantic proximity of the values

t-closeness

Due to the fact that previous privacy models had some vulnerabilities, a new one emerged, the t-closeness algorithm

It requires that the distribution of a SA in any equivalence class must be similar to the attributes distribution in the overall dataset, this way, the chances of learning individual's information are lower

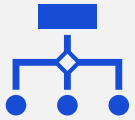


In order to introduce and manage gaps between values of SAs, tcloseness uses the Earth Mover Distance metric , receiving a precise distance between the two distributions

Hierarchies

Defining the intervals is important in the configuration process, all the solutions below use this mechanism, but sometimes is not clear what their function is, neither the right way to configure them

Hierarchies are normally used for categorical attributes, to increase the utility of anonymized datasets, "categorisation" is often combined with tuple suppression, i.e., data records inconsistent with privacy criteria are automatically removed from the dataset



Data and generalization hierarchies can be imported from many different types, providing compatibility with a wide range of data processing tools

Hierarchies

- In this tool the hierarchies can be supplied by the user in the form of ‘CSV’ files, which each file contains the column headers specifying the level fo the hierarchy and rows contain intervals or generalised attributes

	A	B	C	D	E	F	G	H	I	J	K
1	sex;age;race;marital-status;education;native-country;workclass;occupation;salary-class										
2	Male;39;White;Never-married;Bachelors;United-States;State-gov;Adm-clerical;<=50K										
3	Male;50;White;Married-civ-spouse;Bachelors;United-States;Self-emp-not-inc;Exec-managerial;<=50K										
4	Male;38;White;Divorced;HS-grad;United-States;Private;Handlers-cleaners;<=50K										
5	Male;53;Black;Married-civ-spouse;11th;United-States;Private;Handlers-cleaners;<=50K										
6	Female;28;Black;Married-civ-spouse;Bachelors;Cuba;Private;Prof-specialty;<=50K										
7	Female;37;White;Married-civ-spouse;Masters;United-States;Private;Exec-managerial;<=50K										
8	Female;49;Black;Married-spouse-absent;9th;Jamaica;Private;Other-service;<=50K										
9	Male;52;White;Married-civ-spouse;HS-grad;United-States;Self-emp-not-inc;Exec-managerial;>50K										
10	Female;31;White;Never-married;Masters;United-States;Private;Prof-specialty;>50K										
11	Male;42;White;Married-civ-spouse;Bachelors;United-States;Private;Exec-managerial;>50K										
12	Male;37;Black;Married-civ-spouse;Some-college;United-States;Private;Exec-managerial;>50K										
13	Male;30;Asian-Pac-Islander;Married-civ-spouse;Bachelors;India;State-gov;Prof-specialty;>50K										
14	Female;23;White;Never-married;Bachelors;United-States;Private;Adm-clerical;<=50K										
15	Male;32;Black;Never-married;Assoc-acdm;United-States;Private;Sales;<=50K										
16	Male;34;Amer-Indian-Eskimo;Married-civ-spouse;7th-8th;Mexico;Private;Transport-moving;<=50K										
17	Male;25;White;Never-married;HS-grad;United-States;Self-emp-not-inc;Farming-fishing;<=50K										
18	Male;32;White;Never-married;HS-grad;United-States;Private;Machine-op-inspct;<=50K										
19	Male;38;White;Married-civ-spouse;11th;United-States;Private;Sales;<=50K										
20	Female;43;White;Divorced;Masters;United-States;Self-emp-not-inc;Exec-managerial;>50K										
21	Male;40;White;Married-civ-spouse;Doctorate;United-States;Private;Prof-specialty;>50K										
22	Female;54;Black;Separated;HS-grad;United-States;Private;Other-service;<=50K										
23	Male;35;Black;Married-civ-spouse;9th;United-States;Federal-gov;Farming-fishing;<=50K										
24	Male;43;White;Married-civ-spouse;11th;United-States;Private;Transport-moving;<=50K										
25	Female;59;White;Divorced;HS-grad;United-States;Private;Tech-support;<=50K										
26	Male;56;White;Married-civ-spouse;Bachelors;United-States;Local-gov;Tech-support;>50K										
27	Male;19;White;Never-married;HS-grad;United-States;Private;Craft-repair;<=50K										
28	Male;39;White;Divorced;HS-grad;United-States;Private;Exec-managerial;<=50K										
29	Male;49;White;Married-civ-spouse;HS-grad;United-States;Private;Craft-repair;<=50K										

adult

Adult hierarchy age

	A	B	C
1	1;0-4;0-9;0-19;*		
2	2;0-4;0-9;0-19;*		
3	3;0-4;0-9;0-19;*		
4	4;0-4;0-9;0-19;*		
5	5;0-4;0-9;0-19;*		
6	6;5-9;0-9;0-19;*		
7	7;5-9;0-9;0-19;*		
8	8;5-9;0-9;0-19;*		
9	9;5-9;0-9;0-19;*		
10	10;5-9;0-9;0-19;*		
11	11;10-14;10-19;0-19;*		
12	12;10-14;10-19;0-19;*		
13	13;10-14;10-19;0-19;*		
14	14;10-14;10-19;0-19;*		
15	15;10-14;10-19;0-19;*		
16	16;15-19;10-19;0-19;*		
17	17;15-19;10-19;0-19;*		
18	18;15-19;10-19;0-19;*		
19	19;15-19;10-19;0-19;*		
20	20;15-19;10-19;0-19;*		
21	21;20-24;20-29;20-39;*		
22	22;20-24;20-29;20-39;*		
23	23;20-24;20-29;20-39;*		
24	24;20-24;20-29;20-39;*		
25	25;20-24;20-29;20-39;*		
26	26;25-29;20-29;20-39;*		
27	27;25-29;20-29;20-39;*		
28	28;25-29;20-29;20-39;*		
29	29;25-29;20-29;20-39;*		

adult_hierarchy_age

Adult hierarchy native country

1	United-States;North America;*			
2	Cambodia;Asia;*			
3	England;Europe;*			
4	Puerto-Rico;North America;*			
5	Canada;North America;*			
6	Germany;Europe;*			
7	Outlying-US(Guam-USVI-etc);North America;*			
8	India;Asia;*			
9	Japan;Asia;*			
10	Greece;Europe;*			
11	South;Africa;*			
12	China;Asia;*			
13	Cuba;North America;*			
14	Iran;Asia;*			
15	Honduras;North America;*			
16	Philippines;Asia;*			
17	Italy;Europe;*			
18	Poland;Europe;*			
19	Jamaica;North America;*			
20	Vietnam;Asia;*			
21	Mexico;North America;*			
22	Portugal;Europe;*			
23	Ireland;Europe;*			
24	France;Europe;*			
25	Dominican-Republic;North America;*			
26	Laos;Asia;*			
27	Ecuador;South America;*			
28	Taiwan;Asia;*			
29	Haiti;North America;*			

adult_hierarchy_native-country

Attack scenarios

Identity Re-identification:

- Attack Scenario: Malicious actors attempt to re-identify individuals in the anonymized dataset to discover sensitive information.
- Prevention: Techniques like k-anonymity, l-diversity, and t-closeness introduce generalization and suppression methods, making it challenging to identify specific individuals.

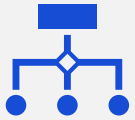
Homogeneity Attacks:

- Attack Scenario: Attackers exploit patterns in the data to reveal that a particular attribute is the same for a group of individuals.
- Prevention: k-Anonymity helps protect against homogeneity attacks by ensuring that each quasi-identifier tuple appears in at least k-1 records, reducing the risk of revealing identical sensitive information.

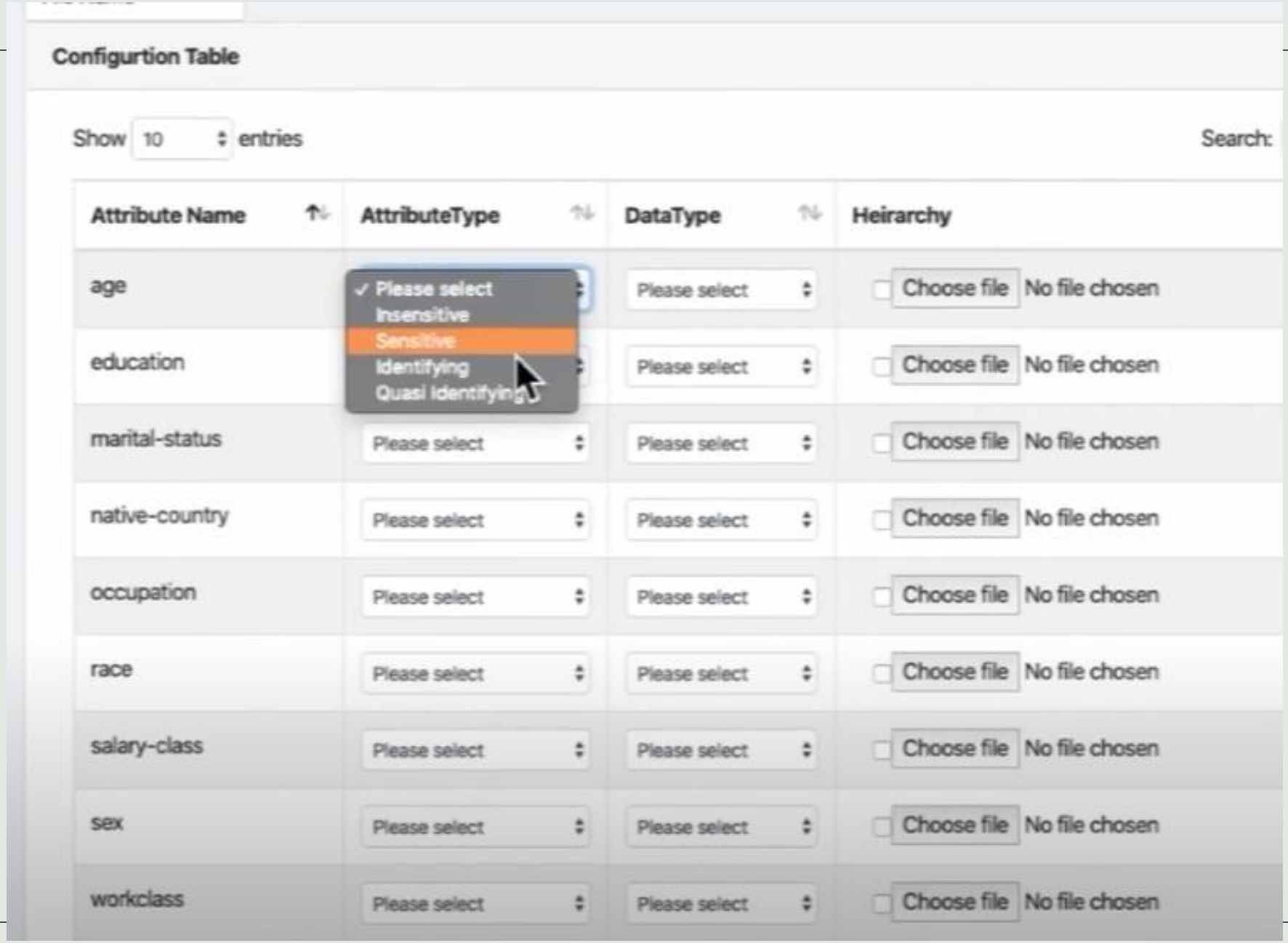
4.4. Anonymize

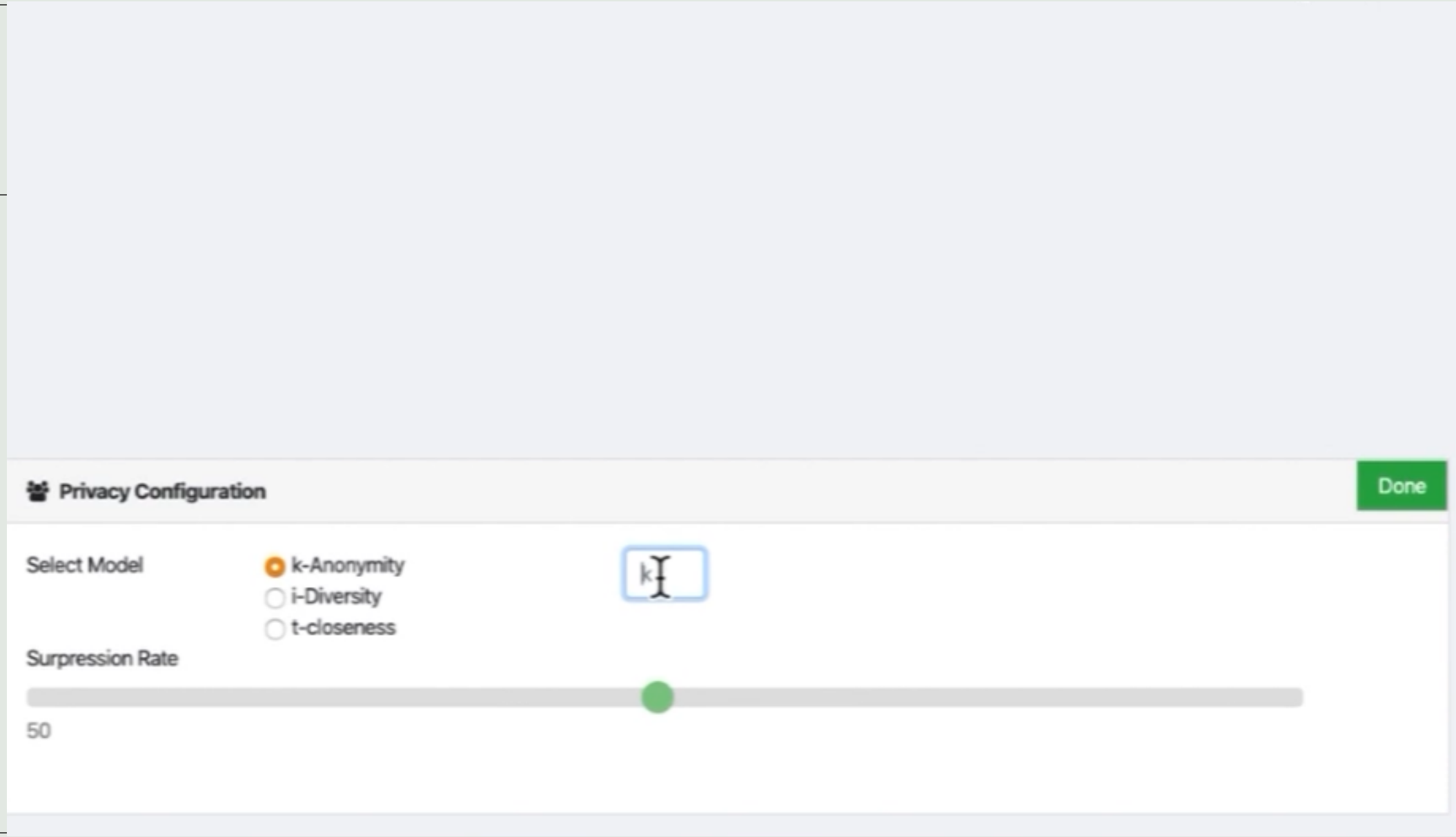
When the user has both Data source and configuration selected or in place, then by clicking the anonymize button the anonymization process can be initiated

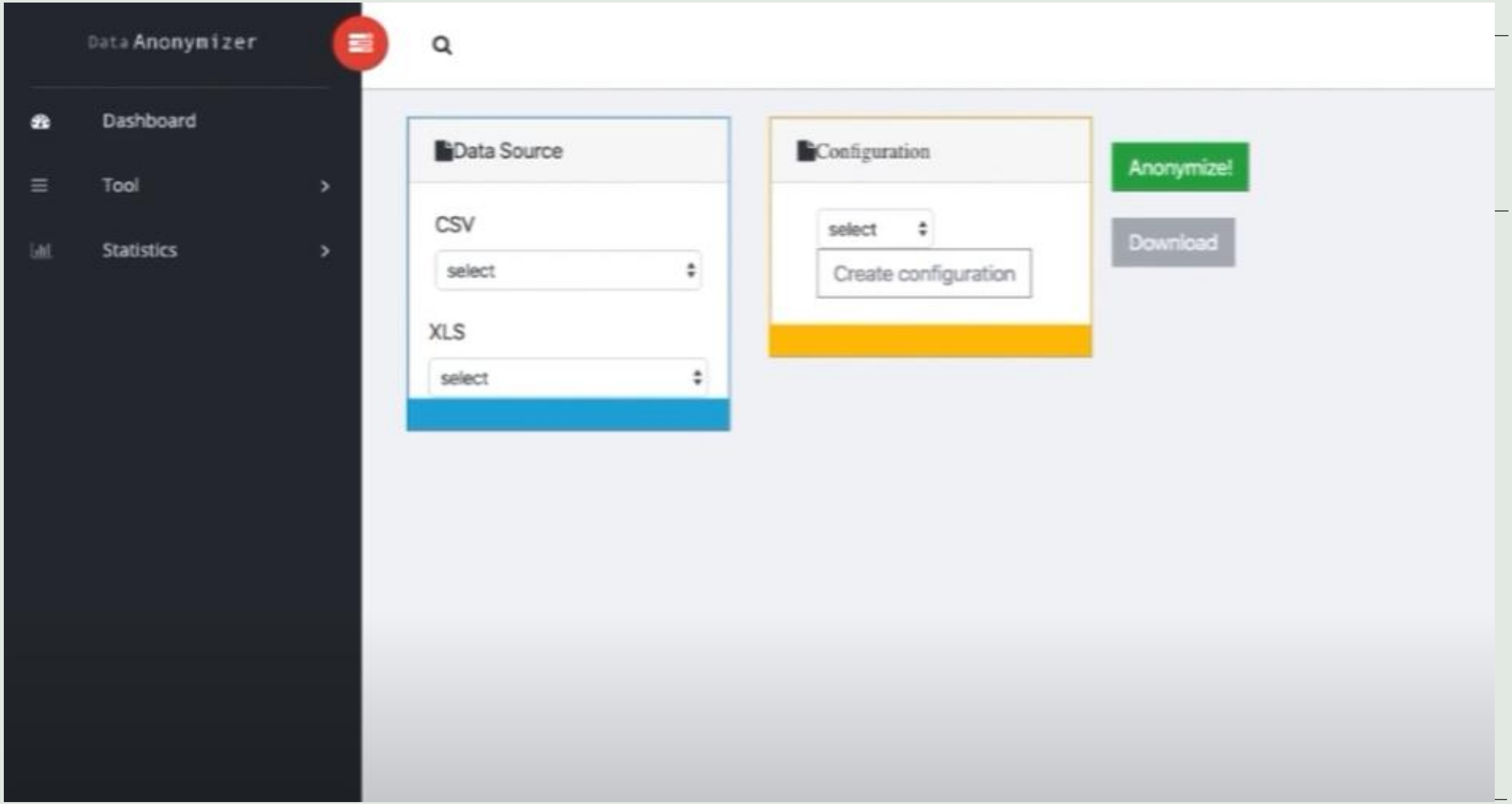
After the anonymization the user can download the anonymized data by clicking the ‘Download’ button which is below on the ‘anonymize’ button



After the anonymization process the Data Source file will be automatically deleted from the server in order for the system to maintain privacy


Configurtion Table
Show 10 entries
Search:
Attribute Name
AttributeType
DataType
Heirarchy
age
education
marital-status
native-country
occupation
race
salary-class
sex
workclass
Please select
Insensitive
Sensitive
Identifying
Quasi Identifying
Choose file
No file chosen


Privacy Configuration
Done
Select Model
k-Anonymity
i-Diversity
t-closeness
k
Surpression Rate
50


Data Anonymizer
Dashboard
Tool
Statistics
Data Source
CSV
select
XLS
select
Configuration
select
Create configuration
Anonymize!
Download

	A	B	C	D	E	F	G	H	I
1	age	sex	native-cou	marital-sta	race	education	salary-clas	workclass	occupation
2	20-39	Male	North Ame	Never-mar	White	Higher edu	<=50K	State-gov	Other
3	40-59	Male	North Ame	Married-ci	White	Higher edu	<=50K	Self-emp-r	Nontechnical
4	20-39	Male	North Ame	Divorced	White	Secondary	<=50K	Private	Nontechnical
5	40-59	Male	North Ame	Married-ci	Black	Secondary	<=50K	Private	Nontechnical
6	*	*	*	*	*	*	*	*	*
7	20-39	Female	North Ame	Married-ci	White	Higher edu	<=50K	Private	Nontechnical
8	40-59	Female	North Ame	Married-sp	Black	Secondary	<=50K	Private	Other
9	40-59	Male	North Ame	Married-ci	White	Secondary	>50K	Self-emp-r	Nontechnical
10	20-39	Female	North Ame	Never-mar	White	Higher edu	>50K	Private	Technical
11	40-59	Male	North Ame	Married-ci	White	Higher edu	>50K	Private	Nontechnical
12	20-39	Male	North Ame	Married-ci	Black	Higher edu	>50K	Private	Nontechnical
13	20-39	Male	Asia	Married-ci	Asian-Pac-	Higher edu	>50K	State-gov	Technical
14	20-39	Female	North Ame	Never-mar	White	Higher edu	<=50K	Private	Other
15	20-39	Male	North Ame	Never-mar	Black	Higher edu	<=50K	Private	Nontechnical
16	20-39	Male	North Ame	Married-ci	Amer-India	Secondary	<=50K	Private	Other
17	20-39	Male	North Ame	Never-mar	White	Secondary	<=50K	Self-emp-r	Other
18	20-39	Male	North Ame	Never-mar	White	Secondary	<=50K	Private	Technical
19	20-39	Male	North Ame	Married-ci	White	Secondary	<=50K	Private	Nontechnical
20	*	*	*	*	*	*	*	*	*
21	20-39	Male	North Ame	Married-ci	White	Higher edu	>50K	Private	Technical
22	40-59	Female	North Ame	Separated	Black	Secondary	<=50K	Private	Other
23	20-39	Male	North Ame	Married-ci	Black	Secondary	<=50K	Federal-go	Other
24	40-59	Male	North Ame	Married-ci	White	Secondary	<=50K	Private	Other
25	40-59	Female	North Ame	Divorced	White	Secondary	<=50K	Private	Technical
26	40-59	Male	North Ame	Married-ci	White	Higher edu	>50K	Local-gov	Technical
27	0-19	Male	North Ame	Never-mar	White	Secondary	<=50K	Private	Technical
28	20-39	Male	North Ame	Divorced	White	Secondary	<=50K	Private	Nontechnical
29	40-59	Male	North Ame	Married-ci	White	Secondary	<=50K	Private	Technical

adult_output

5. Future work

Improving Interface design- The design of this version is composed of single wizard which contains all the features

Database storage- In this version the files uploaded by the user are being stored in a local storage due to ARX API in the server which is a kind of limitation

Input and output views- When the file has been uploaded by the user, there is no option in the present system to view the input

Hierarchy Creation- Similar to creating the configuration there can be an interface for each attribute to specify the intervals or levels or generalisations

ARX - Open Source Data Anonymization Software

A Web Anonymizer Platform for Datasets with Personal Information