Transparency of **data processing** within **data trustee** of sleep studies

presented by **Buwei Liao** (buweiliao@gmail.com)

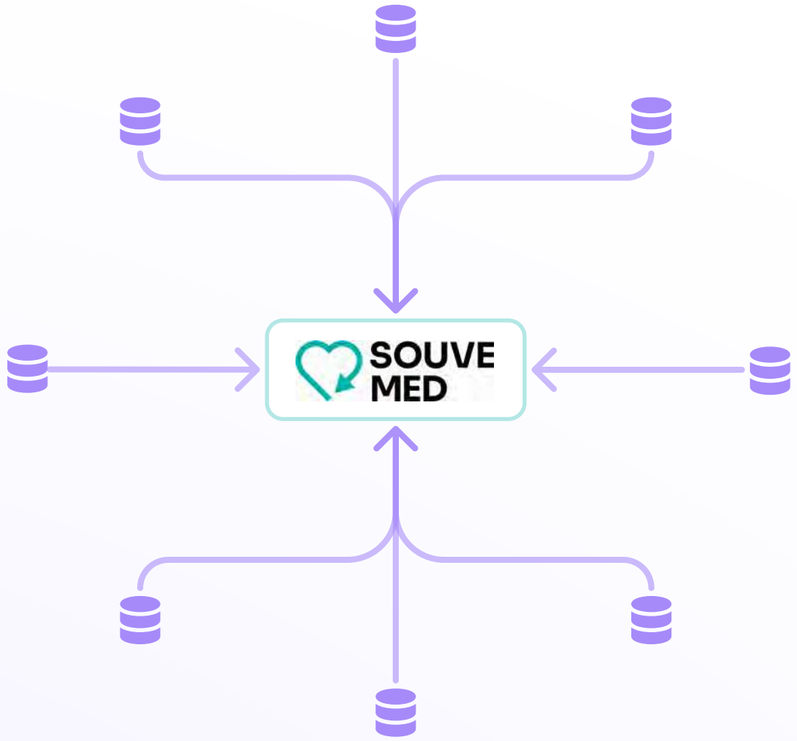
30/06/2023

What is **data trustee**

1. Data trustee connects with different data silos, they are distributed among different clinics or labs originally.

2. Researchers can directly use the data prepared by the platform.

3. Great impact on digital innovation: big data analysis, machine learning, artificial intelligence.



SouveMed

SOUVE MED

a **data trustee platform** for sleep studies

Urgency of **sleep hygiene**

- Sleep deprivation
- Sleep disorders
- Insomnia
- Sleep apnea
- Narcolepsy

37% of people between 20 and 39 years old reported short sleep duration

National Center For Biotechnology Information

20% of all car crash accidents and injuries are associated with sleepiness

Centers for Disease Control and Prevention

Sleep deprivation increases risk of obesity

Harvard School of Public Health

35% of Europeans struggles to get a full night's rest

STADA Health Report 2022

4% reported falling asleep or nodding off while driving in the last 30 days

CDC

90 million people in the U.S. have reported snoring problems

Yale Medicine

1 billion adults around the world experience obstructive sleep apnea.

National Library of Medicine

13% of men reported havin obstructive sleep apnea.

National Library of Medicine

50% of insomnia cases result from anxiety, depression, or psychological stress.

90% of obstructive sleep apnea cases go undiagnosed.

Yale Medicine

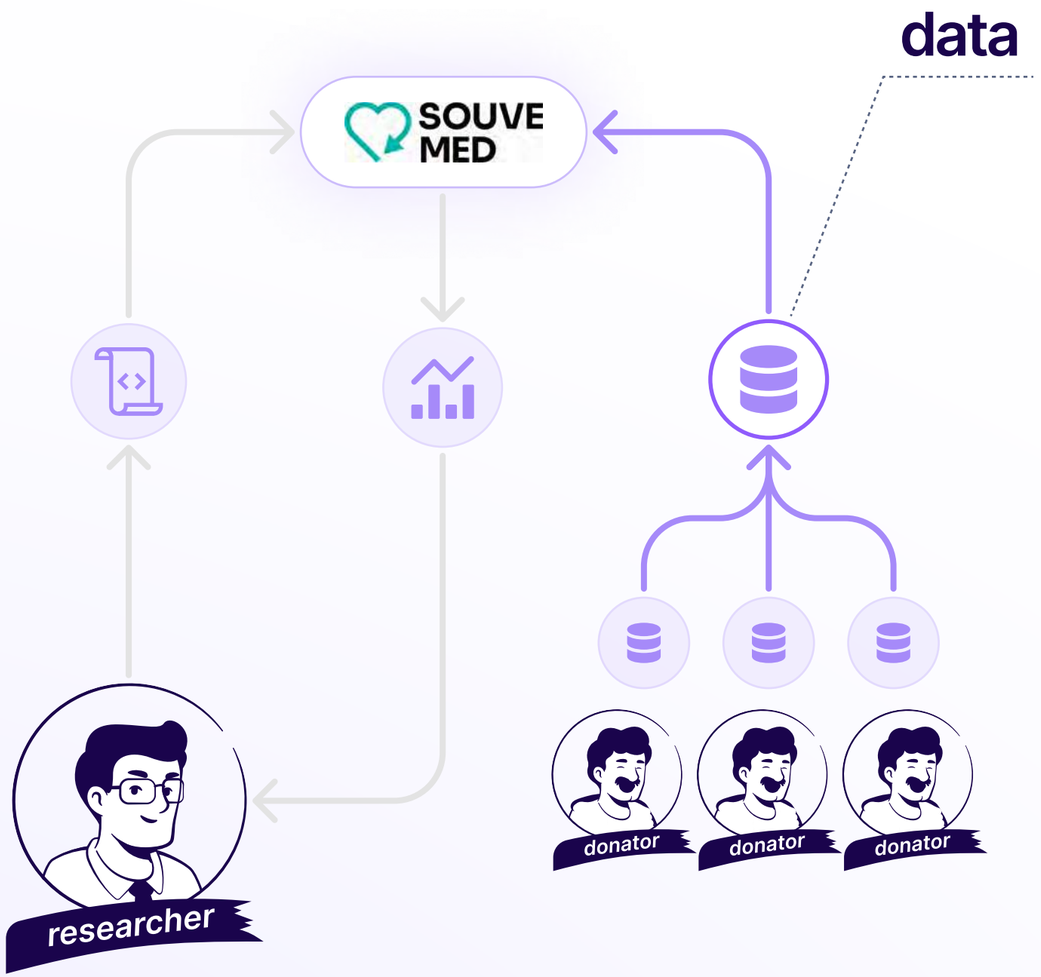
60% of Narcolepsy patients were misdiagnosed as obstructive sleep apnea or depression.

How **SouveMed** works

1. SouveMed build the data pipeline, and gathers required data from different clinics.

2. Researchers submit data analysis algorithm to the blackbox.

3. After the blackbox have finished the analysis, researchers will get the result.

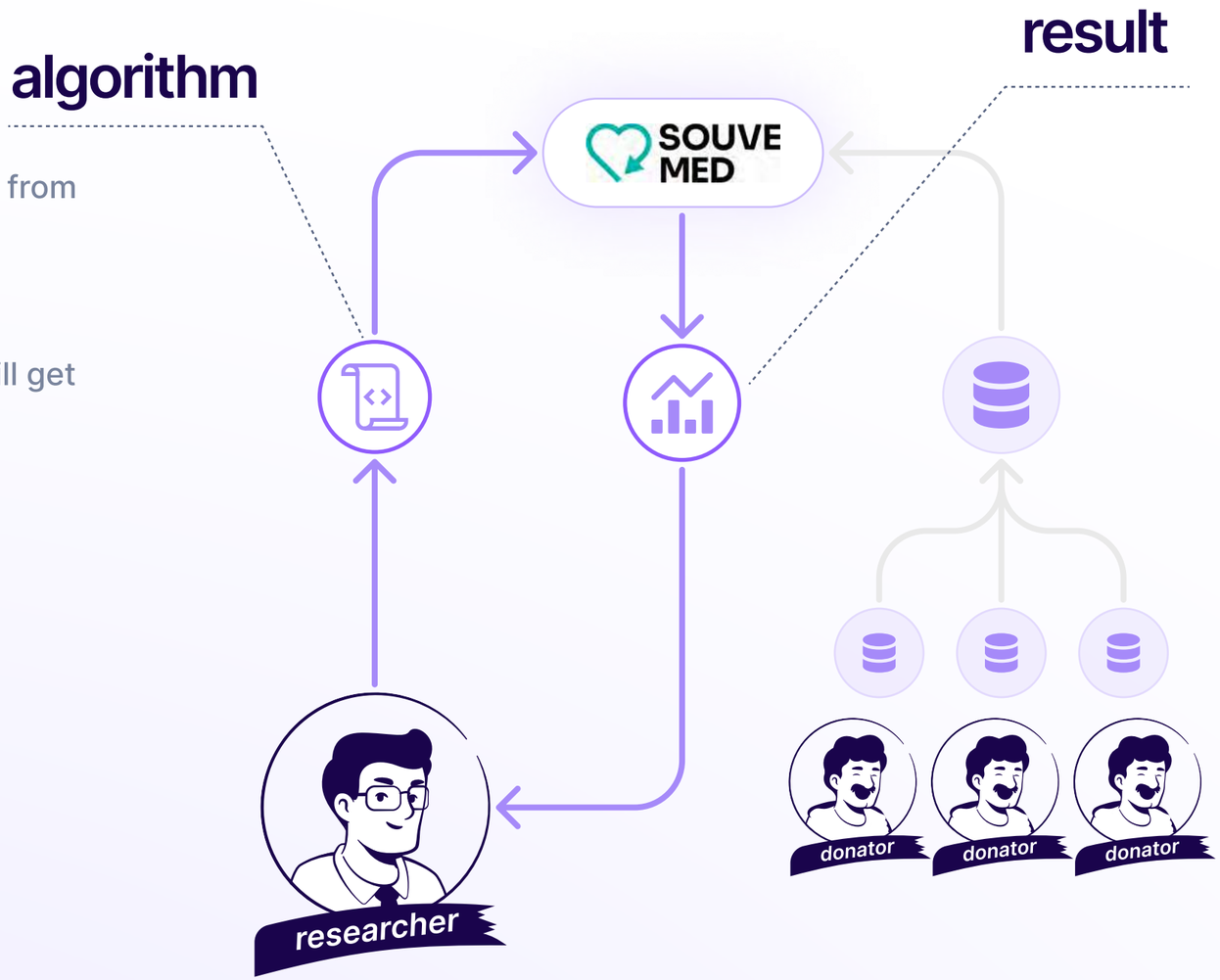


How **SouveMed** works

1. SouveMed build the data pipeline, and gathers required data from different clinics.

2. Researchers submit data analysis algorithm to the blackbox.

3. After the blackbox have finished the analysis, researchers will get the result.



Why **transparency** matters

1. GDPR, **General Data Protection Regulation**
- GDPR requires businesses to protect the personal data and privacy of EU citizens.

2. DGA, **Data Governance Act**
- DGA encorages data sharing through novel intermediaries.
- DGA will be officially applicable from September 2023.

3. BDSG, **Federal Data Protection Act**
- Governs the exposure of personal data in the national level.

GDPR ➜ https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
DGA ➜ https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868
BDSG ➜ https://www.gesetze-im-internet.de/bdsg_2018/

BDSG

DGA

GDPR

Research questions

1

Authentic audit logs

Cyber attacks, internal attackers deter the authenticity of audit logs, making it unable to truly reflect data processing events on the platform.

So we need a technologically robust solution to ensure authenticity of audit logs.

2

User requirements

Different stakeholders are connected to the data trustee platform. Each of them have their specific requirements for “transparency”.

3

Effectiveness of information disclosure

The more friendly the message is, the more efficient it is communicated to the user.

Log information is often cyptic for average users. It’s crucial to convey these information in a user-friendly way.

My overall schedule

Literature review	Concept design	Implementation	Evaluation	Writing
1/2 ~ 15/3	16/3 ~ 15/5	16/5 ~ 15/6	16/6 ~ 30/6	1/7 ~ 30/7

Literature review

Search string

(log OR audit OR provenance OR transparen OR accountab OR repudia)
AND
(secur OR immutab OR tamper OR distribut OR forensic OR blockchain OR cryptograph)

Literature filtering

First round

title, keywords and abstract

```
Wiley Online Library: 78 → 4
IEEE Xplore:          60 → 28
ACM DL:               48 → 19
ResearchGate:        152 → 28
ScienceDirect:        69 → 15
Scopus:              319 → 44
Springer Link:       123 → 10
```

114 items in total after deduplication

Second round

quality and subtopic of fulltext

```
1. Systematic review      → 8
2. Hardware-based         → 14
3. Cryptography-based     → 16
4. Blockchain-based       → 29
5. 3rd party service      → 1
Some connected papers     → 11
```

Taxonomy development

Covered phases

Generation ➜ Transmission ➜ Storage

Attacks

- truncation attack
- delayed detection attack
- reorder attack
- insertion attack
- modification attack

Security measures

- forward secure
- data encryption
- secure log retrieval
- public verifiability

paper	scheme name	technical means		
			forward secure	data encryption
(Schneier & Kelsey, 1998)(Schneier & Kelsey, 1999)	/	cryptograpphy	×	✓
(Bellare & Yee, 1997)	/	cryptograpphy (one-way hash chain, evolving symmetric keys, MACs)	✓	✓
(Ma & Tsudik, 2009)	FssAgg	cryptography (forward-secure signatures and aggregate signatures)	✓	×
(Holt, 2006)	Logcrypt	cryptography (identity-based encryption)	✓	✓
(Yavuz et al., 2012)	LogFAS		✓	×
(Yavuz & Ning, 2009)	BAF	cryptography (blind aggregate forward)	✓	✓
(Yavuz et al., 2012)	Fi-BAF	cryptography (blind aggregate forward)	✓	✓
(Kampanakis & Yavuz, 2015)	BAFi	cryptography (blind aggregate forward)	✓	✓
(Hartung et al., 2017)	/	cryptography (fault-tolerant forward-secure sequential aggregate signature)	✓	✓
(Wang & Zheng, 2003)	/	hardware (WORM device)		
(Chong et al., 2003)	/	hardware (iButton) + cryptography	×	✓
(Sinha et al., 2014)	/	hardware (TPM) + cryptography	✓	✓
(Karande et al., 2017)	SGX-Log	hardware (Intel SGX) + cryptography	✓	✓
(Shepherd et al., 2017)	EmLog	hardware (TEE) + cryptography	✓	✓
(Accorsi, 2011)	BBox	hardware (TPM) + cryptography (symmetric keys)		
(Zawoad et al., 2013)	SecLaaS	third party service (cloud function) + cryptography		✓
(Snodgrass et al., 2004)	/	third party service (notary service) + cryptography		
(Cucurull & Puiggalí, 2016)	/	cryptography (MAC hash chain) + blockchain (store hash value of checkpoint)	✓	✓

State of the art

1st generation

(1998 ~ 2009)

Focus on security of logs stored on local logging server.

Based on cryptographic method to construct a robust hash chain, but unable to avoid **single point of failure**.

(Schneier & Kelsey, 1998) (Schneier & Kelsey, 1999) (Bellare & Yee, 1997) (Ma & Tsudik, 2009)

2nd generation

(2009 ~ 2015)

Combines cryptography and **secure handware** such as Intel SGX, TPM, or TEE together to make the entire logging system more robust.

(Sinha et al., 2014) (Karande et al., 2017) (Chong et al., 2003) (Karande et al., 2017)

3rd generation

(2015 ~ 2022)

The recent mainstream solutions have achieved **immutability** with the help of Blockchain and smart contract. But most of them focus only on technology and failed to mention how to effectively concey log information.

Concept design

Requirement engineering

Concept design

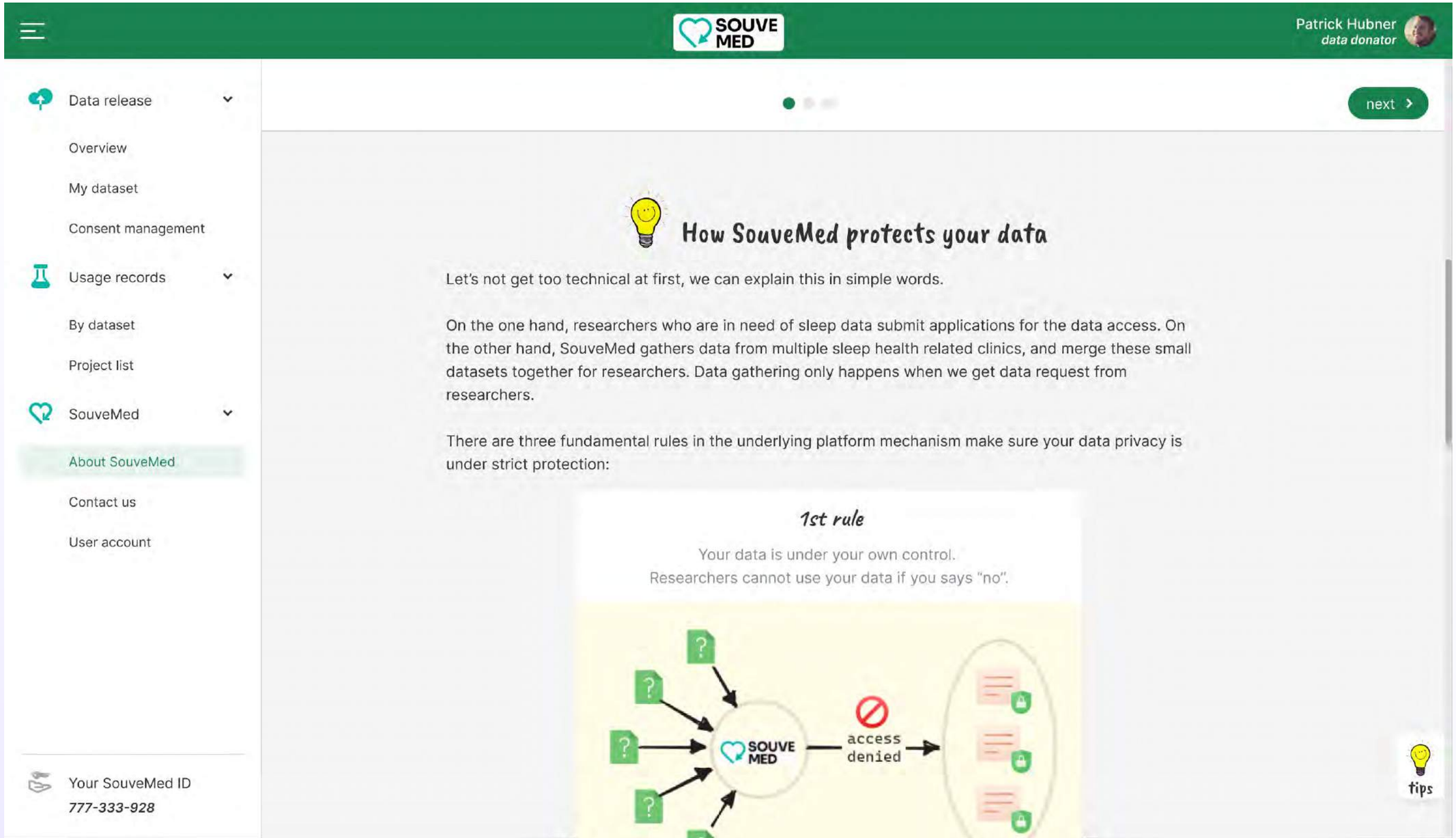
Requirement engineering

- What's included in my dataset?
- How will the data be processed by the platform?
- Who is using the data?
- How to adjust preferences?

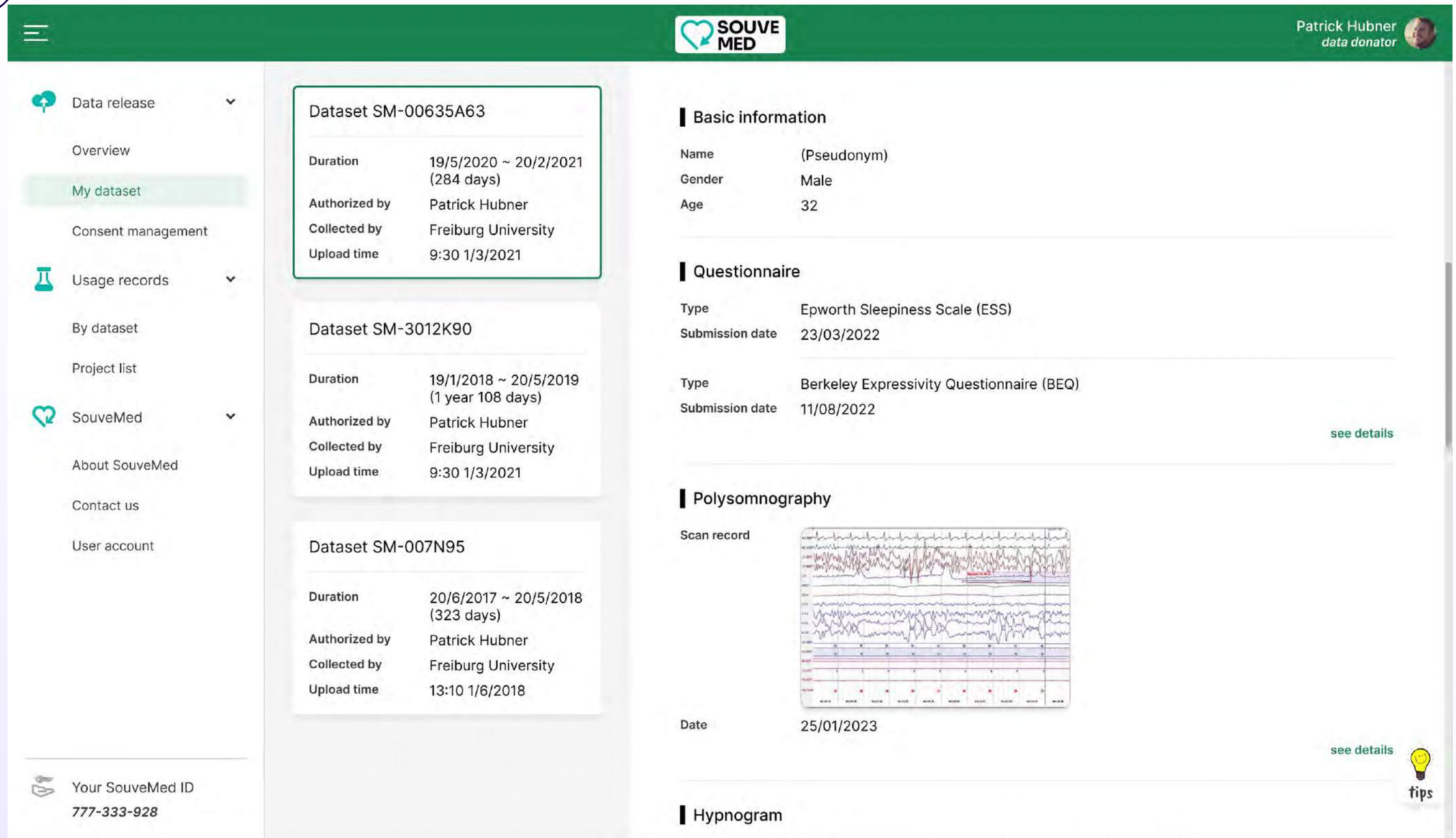
- The third party incldues legal enforcement agencies, technology enthusiasts and peer researchers from the community.
- Whether the platform has been compliant to laws?
- Reproduce research result.

- Lookup usage records.
- How to check the validity of proofs while facing disputes from the 3rd party or data donator?

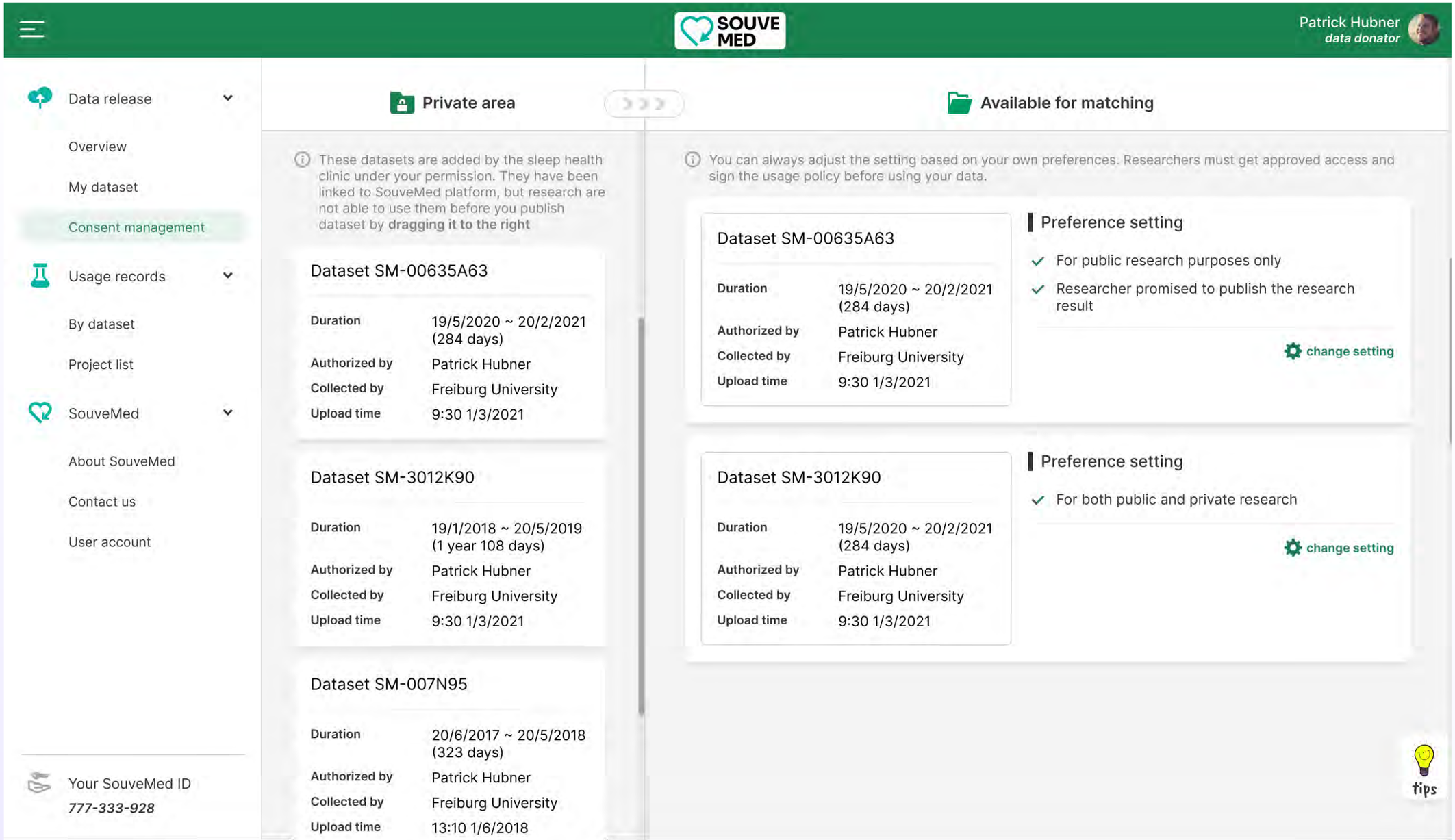
UI prototype



UI prototype



UI prototype



UI prototype

SOUVE MED

Patrick Hubner
data donator

- Data release
  - Overview
  - My dataset
  - Consent management
- Usage records
  - By dataset
  - Project list
- SouveMed
  - About SouveMed
  - Contact us
  - User account

Your SouveMed ID
777-333-928

Search
input keywords

Dataset
Dataset SM-3012K90

Research note
- [ ] Available
- [ ] Not available

Data category
- [ ] Questionnaire
- [ ] Polysomnography
- [ ] Hypnogram

Project type
- [x] Public research
- [ ] Private research

Date of usage
from date ~ to date

8 search results Dataset SM-3012K90 × Public research ×

Project acronym	Project title	Organization	Status	Data category	Research note
Tracker3	Relation of sleep-disordered breathing to cardiovascular disease risk factors	Sleep track GmbH	In progress	questionnaire, hypnogram	Not available
MSSE	Ambulatory sleep scoring using accelerometers-distinguishing between nonwear and sleep/ wake states	Med Sci Sports	Ended	questionnaire, polysomnogra phy	view online
EBASO	Measure the correlation of age, blood pressure, heart rate and sleep	University of Freiburg	Ended	questionnaire, polysomnogra phy	view online
FSSG	Further Validation of Actigraphy for Sleep Studies	University Hospital Rechts der Isar	Ended	questionnaire	view online
H.E.A.R.T	The corticothalamic system in sleep	FZI Forschungszent rum Informatik	In progress	questionnaire, polysomnogra phy, hypnogram	Not available
DSMG	Cognitive Performance, Sleepiness, and Mood in Partially Sleep Deprived Adolescents: The Need for Sleep Study	University Hospital Heidelberg	Ended	questionnaire, polysomnogra phy	view online
DeepSleep	Use of home sleep studies for diagnosis of the sleep apnoea/ hypopnoea syndrome.	Universitätsklini kum Carl Gustav Carus Dresden	Ended	questionnaire	view online
SleepTight	To investigate the effects of sleep restriction on cognitive performance, subjective sleepiness, and mood in	Nordwest Hospital, Frankfurt	In progress	questionnaire, polysomnogra phy, hypnogram	view online

tips

UI prototype

SOUVE MED

Home | Project list | How to audit | Contact

back to detail page

Verify blockchain hash

Blockchain record

Platform	Ethereum blockchain
Transaction hash	0xDB65702A9b26f8a643a31a4c84b9392589e03D7c
Block id	17266044

secured by ETH

Hash input

Organization	*Sleep track GmbH*
Project URL	https://souvemed.de/project/72hhadf-j83hasdfj-09ahdfj
Experiment URL	https://souvemed.de/experiment/a3biusdf-sdhfuba-q2bnd
Timestamp	10:13:56 07/05/2023 UTC+8
Algorithm file path	https://docker.io/project/ajsidf7723/experiment-636
Algorithm file hash	0xD7S7eUds9f89ahndfui883jad7fyhij92nmxc

copy all

Online validation

Input

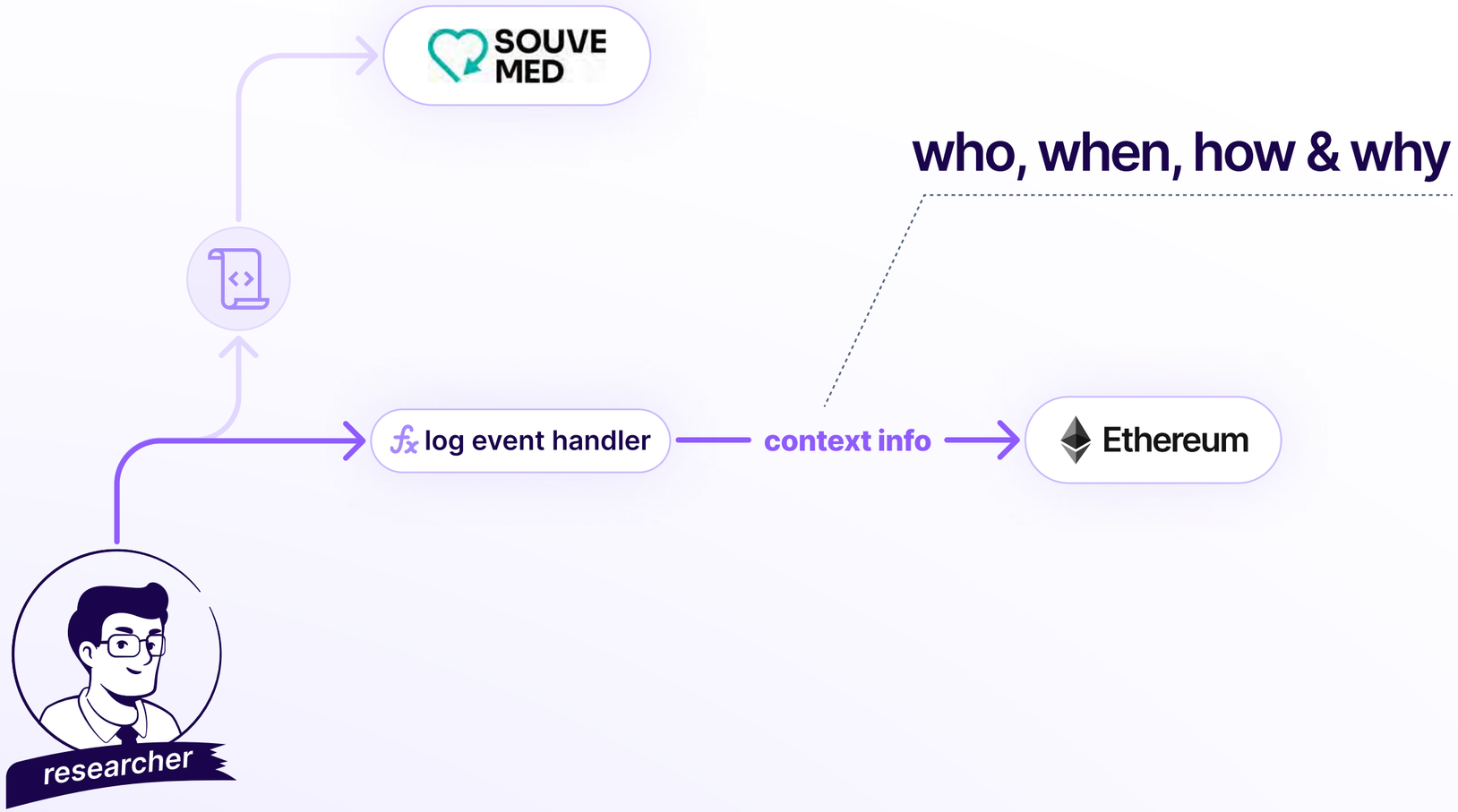

```
{
  "organization": "Sleep track GmbH",
  "projectURL": "https://souvemed.de/project/72hhadf-j83hasdfj-09ahdfj",
  "experimentURL": "https://souvemed.de/experiment/a3biusdf-sdhfuba-
q2bnd",
  "timestamp": "10:13:56 07/05/2023 UTC+8",
  "algorithm": {
    "filePath": "https://docker.io/project/ajsidf7723/experiment-636",
    "fileHash": "0xD7S7eUds9f89ahndfui883jad7fyhij92nmxc"
  }
}
```


json

Hash function Keccak256

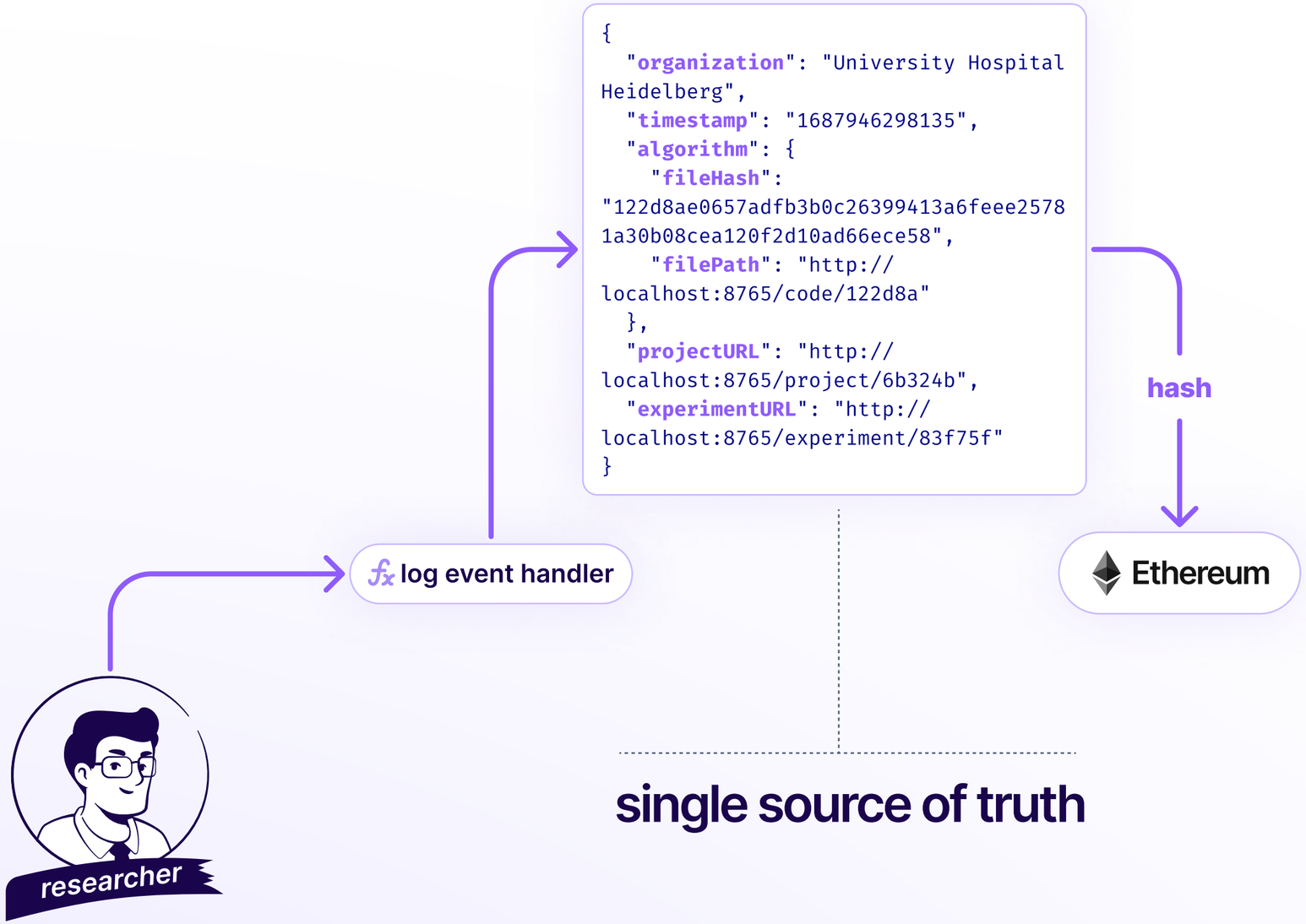
Implementation

1. Each time the researcher runs an experiment, the log event handler will be triggered.
2. The log event handler gather context information about this experiment.
3. The context information includes who, when, how & why the data is being used.
4. The context information will be used to generate a hash.
5. The hash will be send to the smart contract and be notarized.



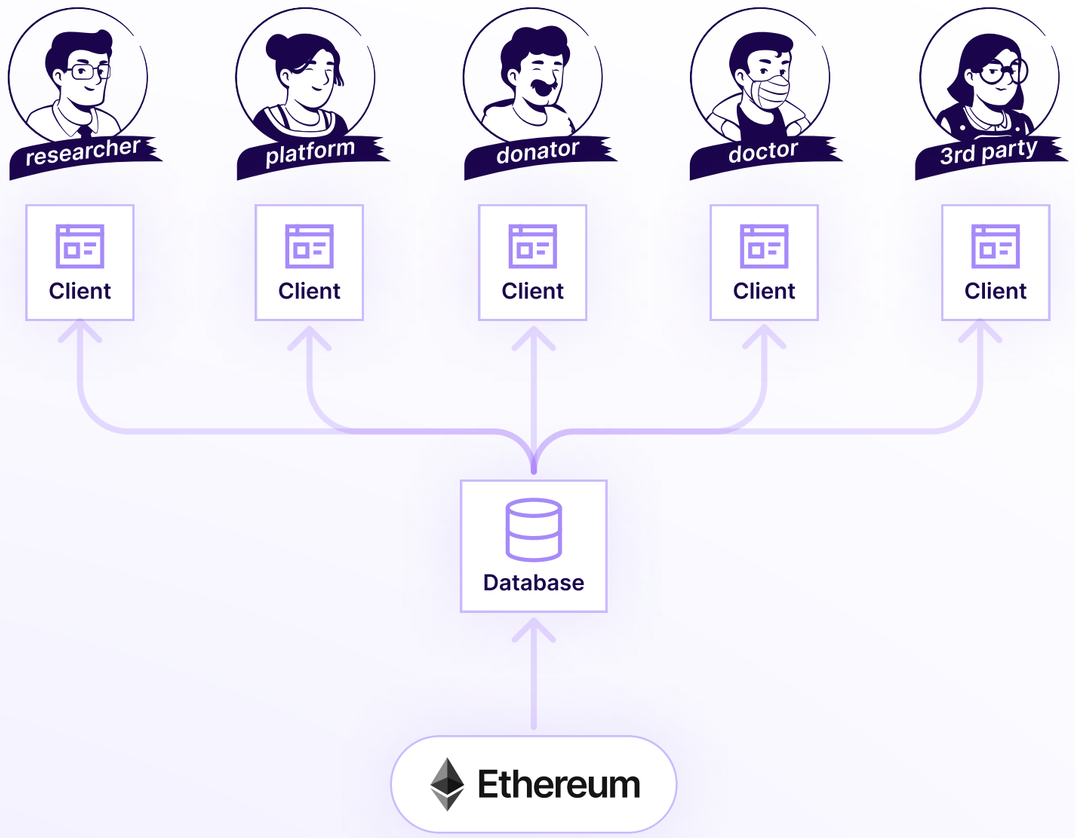
Implementation

1. Each time the researcher runs an experiment, the log event handler will be triggered.
2. The log event handler gather context information about this experiment.
3. The context information includes who, when, how & why the data is being used.
4. The context information will be used to generate a hash.
5. The hash will be send to the smart contract and be notarized.

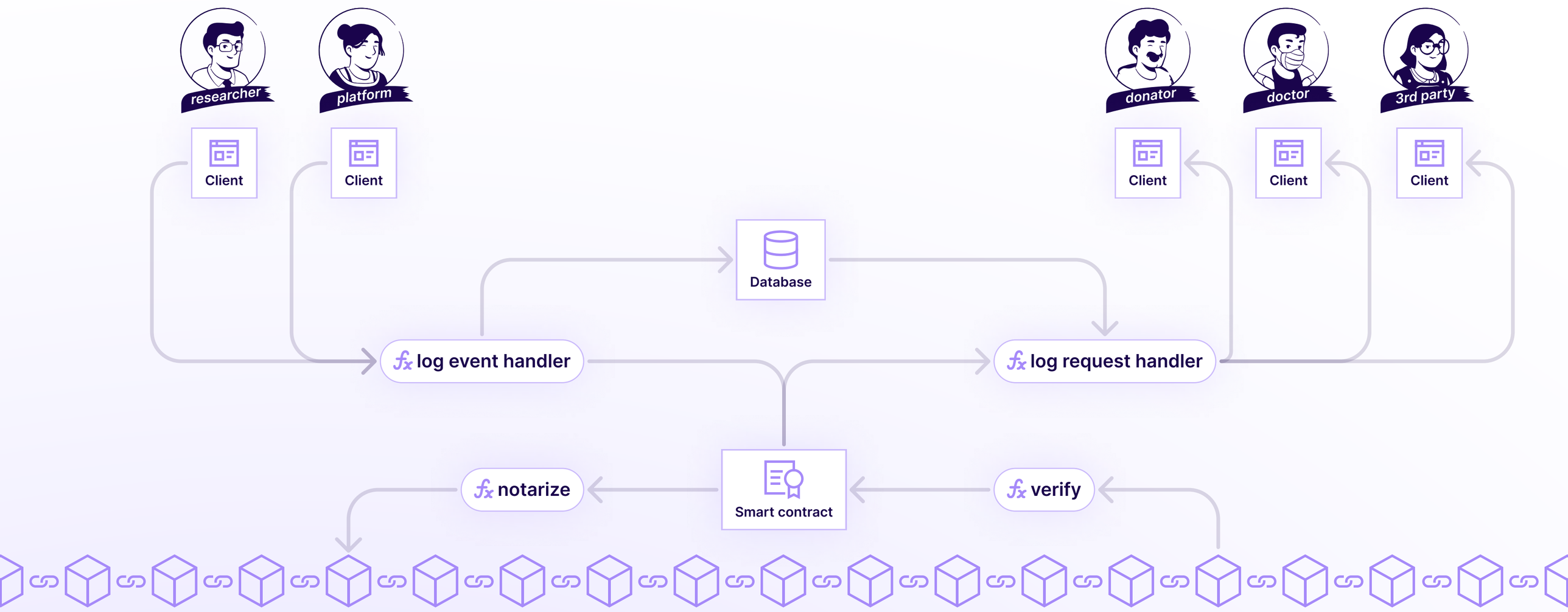


Implementation

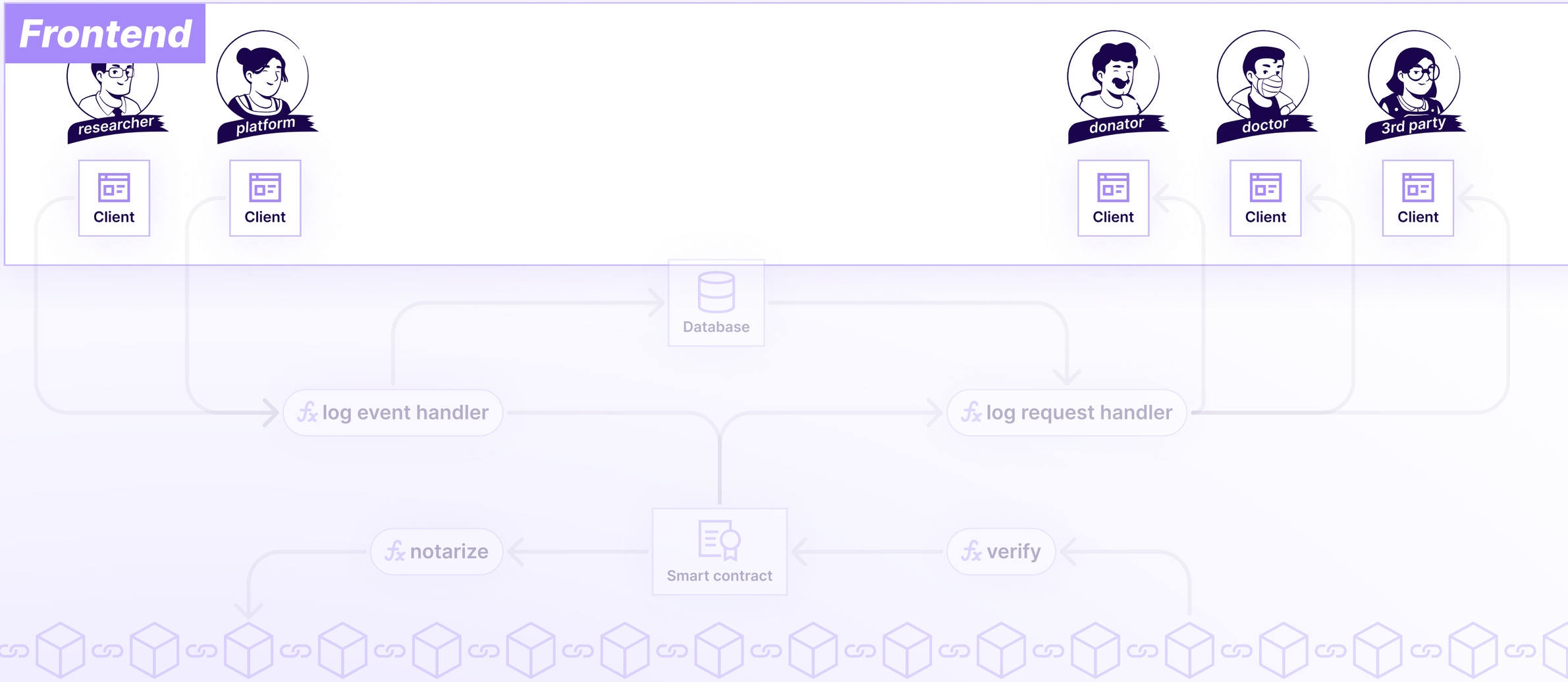
- Dedicated client apps for different stakeholders.
- Able to check the validity of log events.
- Professional user, for example technology enhthusiasts from the community should be able to verify the logs by themselves.



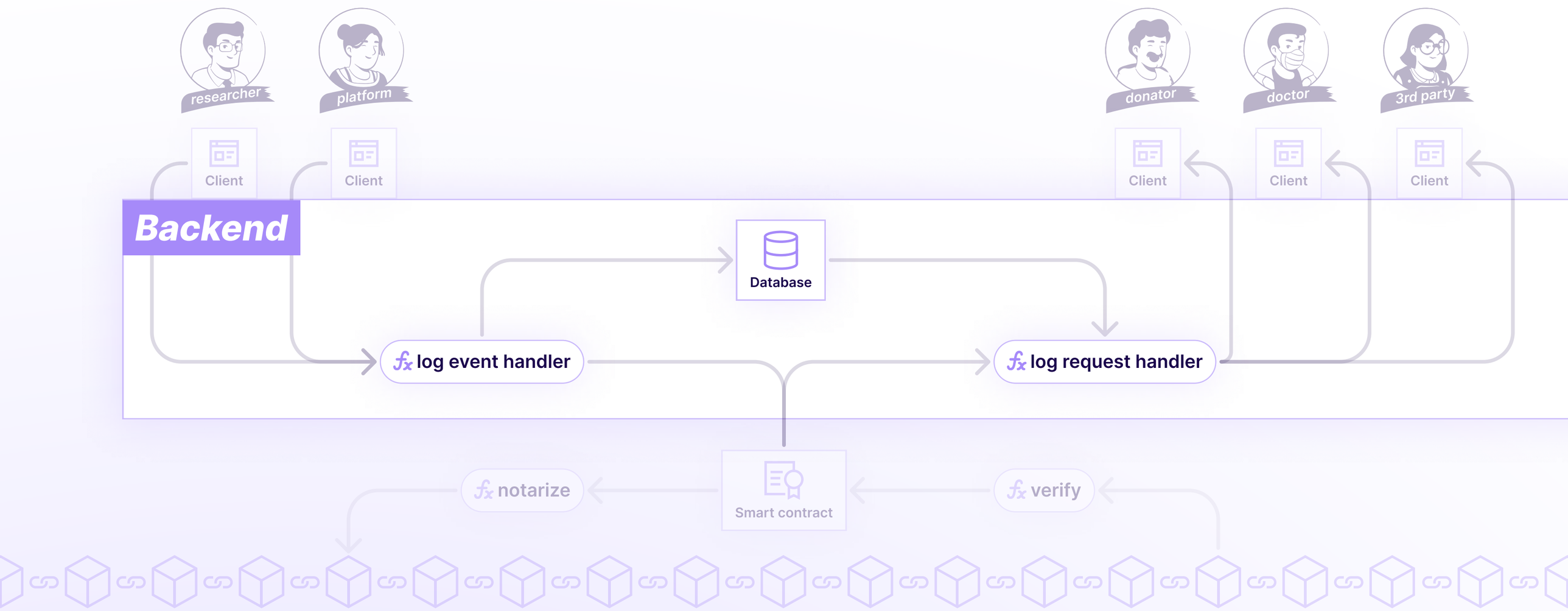
3-layer technology architecture



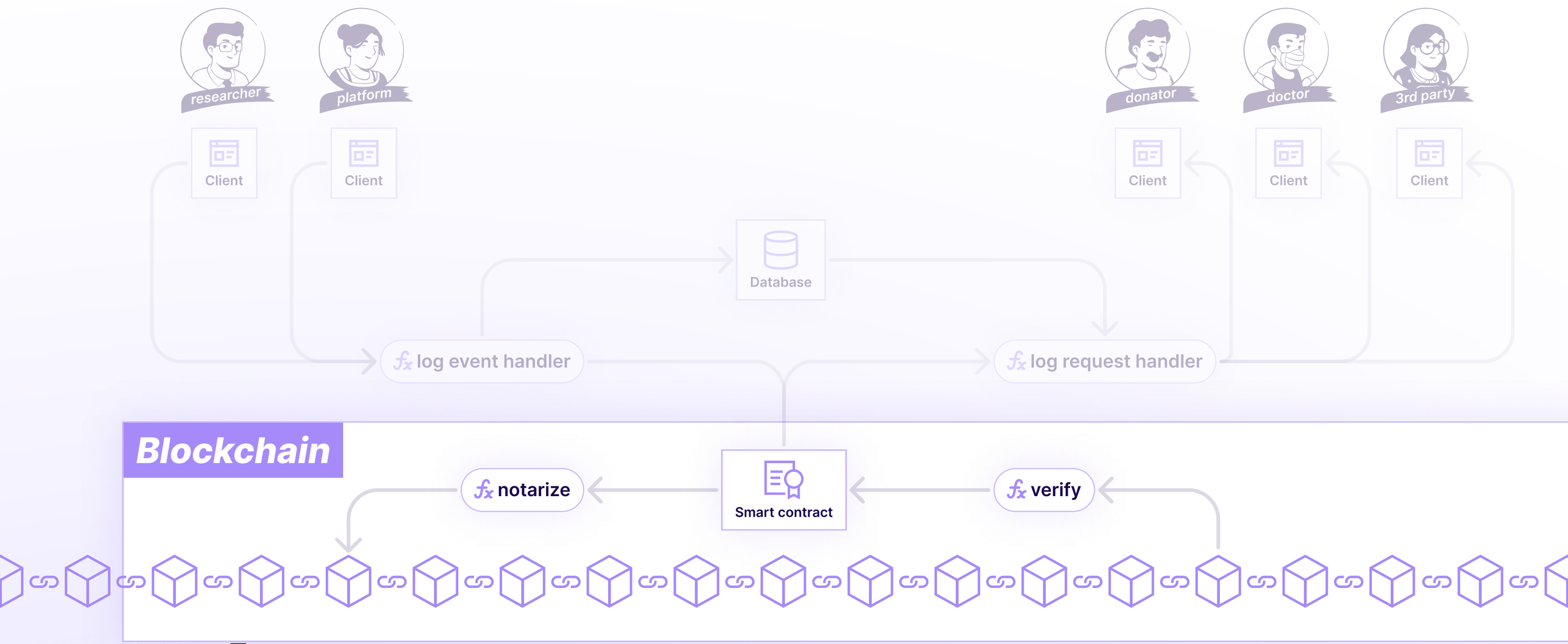
3-layer technology architecture



3-layer technology architecture



3-layer technology architecture



Implementation

Technology stack

- React.js to build frontend UI, tremor and tailwindcss to add styles.
- Next.js to handle requests from frontend.
  Prisma to connect with the database.
  Ethers to interact with smart contract.
- PostgreSQL as the database.
- Ethereum is one of the most secure and stable public blockchain.
  Solidity to write smart contracts.

Frontend	React.js, tremor, tailwindcss
Backend	NEXT.JS, Prisma, ethers
Storage	postgreSQL
Immutability	Solidity, Ethereum

Demo of the mockup system

https://extropy.dev/loglock

Immutable logging (PoC)

This prototype is a proof of concept that demonstrates how Ethereum blockchain could be used to immutablize log information.

Statistic | Project list | Mock data | Verification tool

Filter options

Project type: Select...

Organization: Select...

Keywords: Type...

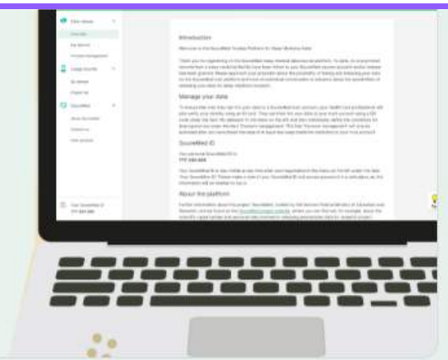
Acronym	Project title	Project type	Project URL	Experiment URL
Med-5	Further Validation of Actigraphy for Sleep Studies	Public	Visit project	Visit experiment
HealthMed	Cognitive Performance, Sleepiness, and Mood in Partially Sleep Deprived Adolescents: The Need for Sleep Study	Private	Visit project	Visit experiment
FastNight	To investigate the effects of sleep restriction on cognitive performance, subjective sleepiness, and mood in adolescents.	Private	Visit project	Visit experiment
SleepSci	Sleep and circadian rhythms in health and disease	Public	Visit project	Visit experiment

Evaluating the prototype & mockup system

https://forms.gle/JEzsB8SJC3Ab4X2RA

Test the prototype of SouveMed platform

SOUVE MED



About this survey

Hello,

My name is Buwei Liao, I study information system engineering and management at Karlsruhe Institute of Technology. I'm currently writing a master thesis about "Transparency of data processing within data trustee platform of sleep research" at FZI (Research Center for Information Technology).

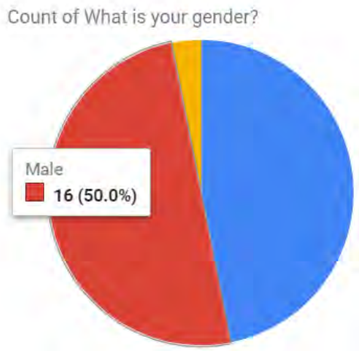
For this purpose I created a web application (prototype), and hope to gather real feedbacks with the help from you.

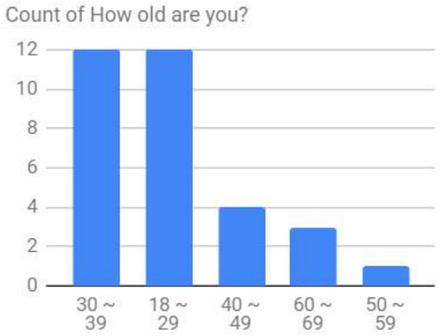
Note: The prototype is designed for web, so it makes sense if you test them on a relatively larger

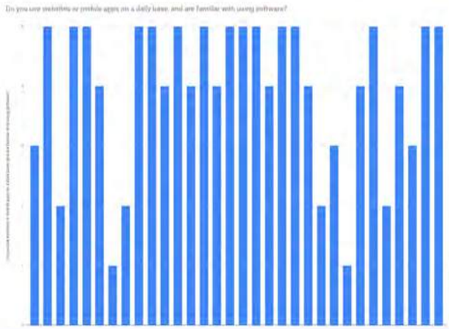
1. general information
2. tasks
3. System Usability Scale (SUS)

+

TeamViewer







Task 1

About SouveMed & My dataset

81.2%

Task 2

Consent management

84.3%

Task 3 & Task 4

Usage records: two different modes

82.2%

32 feedbacks
5 live session

How transparent do you think SouveMed platform is?

9.37

System Usability Scale

77.1

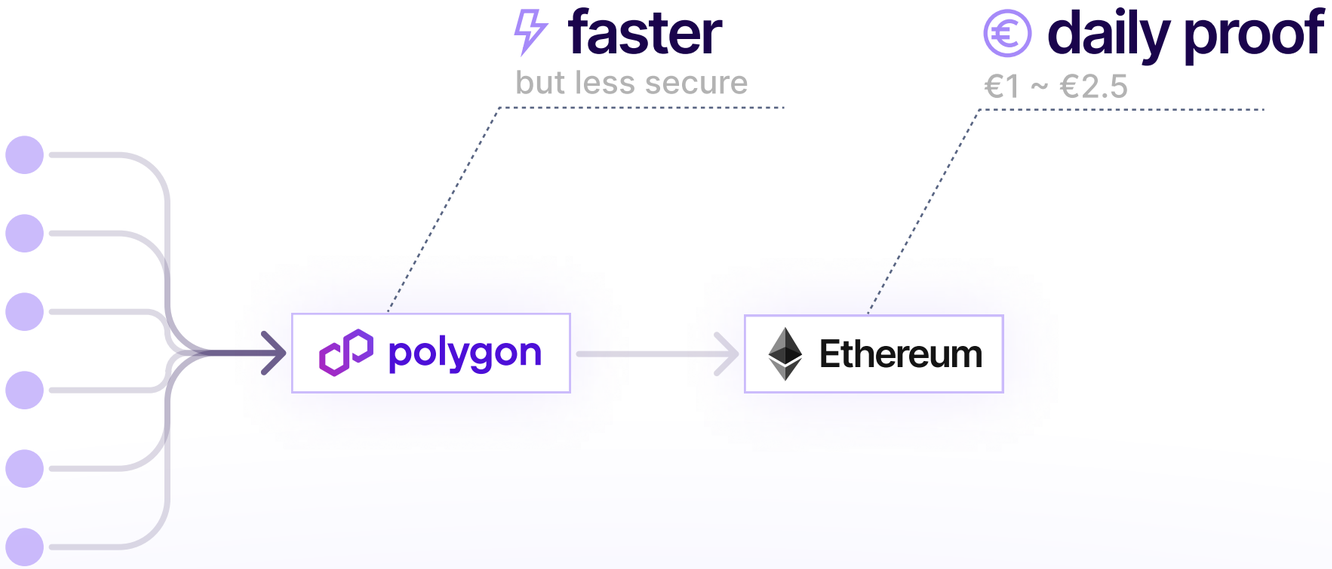
Future improvements

Low throughput

Ethereum is the second most secure public blockchain available, but it also take a little bit longer and high gas fee to process each log event.

Apply a Layer 2 blockchain and make it collaborate with Ethereum.

1



weakness of the system

Future improvements

Partial proof

There are 4 primary process might cause security problems: gathering data consent, signing usage policy, building data pipeline, and using data. Currently we are focusing on the first part.

Build proofs for entire life cycle.

2

- Gathering data consent from donators
- Signing usage policy by researchers
- Building data pipeline by the platform
- Consuming data by researchers

Ethereum

weakness of the system

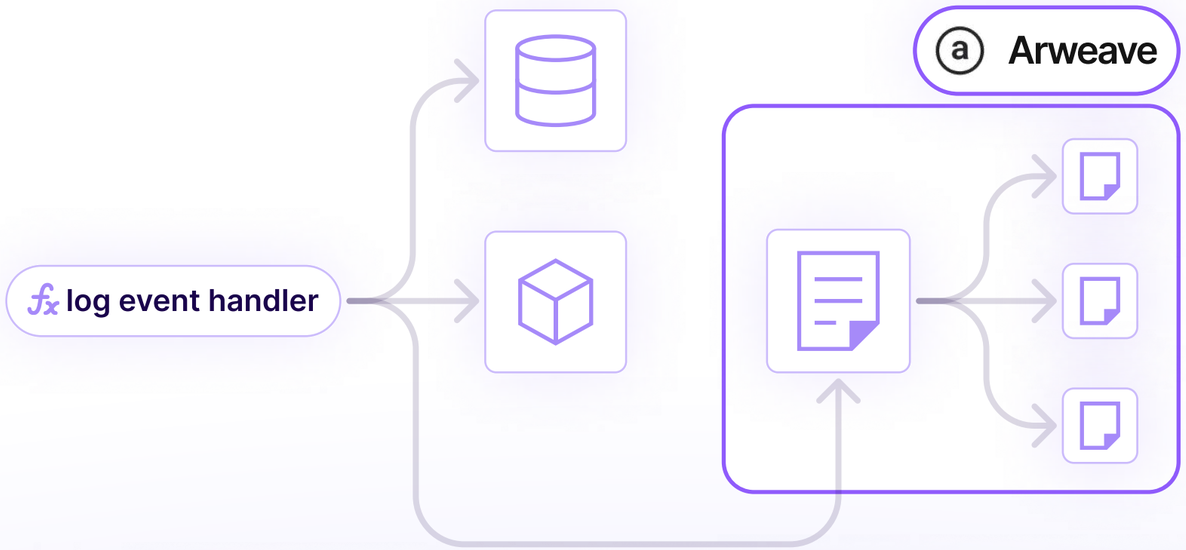
Future improvements

Insider attack

Our solution is based on the assumption that we can trust the database of the platform. We assume it's secure and won't be modified out of unjustified reasons.

Use a 3rd party decentralized permanent storage service like Arweave

3



weakness of the system

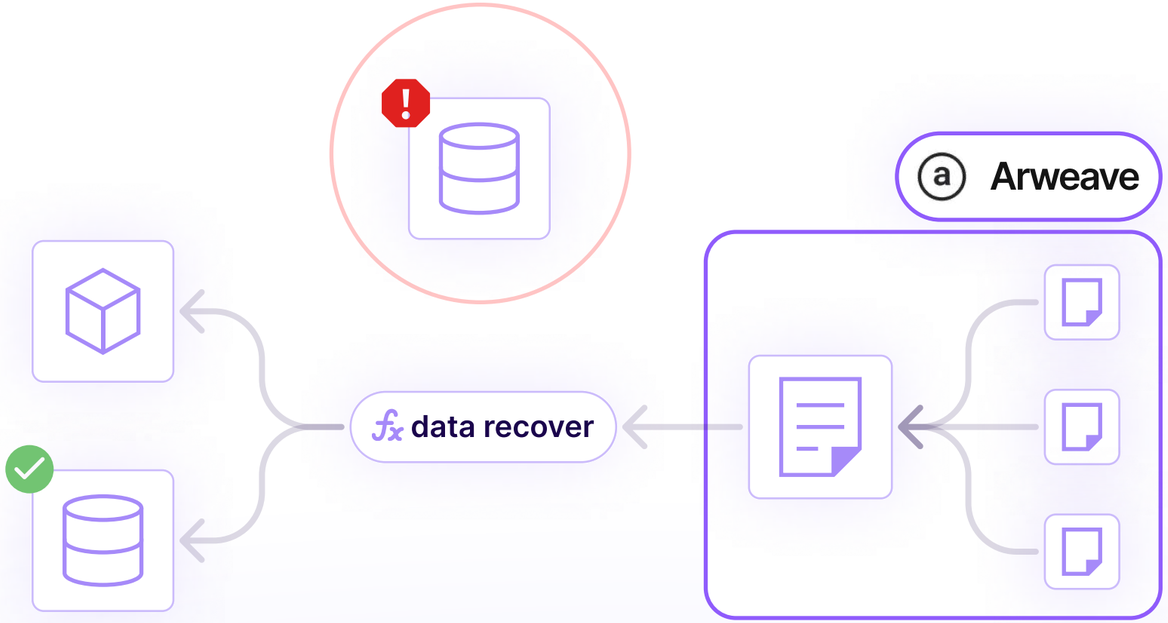
Future improvements

Insider attack

Our solution is based on the assumption that we can trust the database of the platform. We assume it's secure and won't be modified out of unjustified reasons.

Use a 3rd party decentralized permanent storage service like Arweave

3



Thanks for listening!

presented by **Buwei Liao** (buweiliao@gmail.com)

30/06/2023