# Crowdsourcing Smart Home Data

## Technical presentation

Group 5

# The Team

| | |
|---|---|
| **Team Leader** | - Martinho Tavares |
| **Frontend Dev** | - Diogo Monteiro |
| **Backend Dev** | - Camila Fonseca |
| **DevOps Master** | - Rodrigo Lima |

---

| | |
|---|---|
| **Coordinator** | - Diogo Gomes |

# What?

Smart homes have been on the rise, and will continue to grow in the foreseeable future.

# Why?

- ❏ Lack of real data about smart home usage (information about the house itself, rather than statistics)
- ❏ No central and automated way of gathering data from many sources
- ❏ Few datasets provide the information we're looking for

# How?

- ❏ Collect smart home usage data from volunteers
- ❏ Aggregate and store collected information
- ❏ Respect user privacy and anonymity
- ❏ Export data in CKAN compliant formats
- ❏ Visualize data in a web-based dashboard

# State of the Art

## *UK-DALE Dataset*

Open-access dataset from the UK recording Domestic Appliance-Level Electricity.

Issues:

- ❏ Limited region scope, only evaluates UK;
- ❏ Limited entities scope, only five houses were evaluated;
- ❏ Lack of data visualization;
- ❏ Lack of real-time data updates;

Source: https://ukerc.rl.ac.uk/DC/cgi-bin/edc_search.pl?GoButton=Detail&WantComp=41

## *CASAS Datasets*

A collection of datasets from the CASAS project of Washington State University

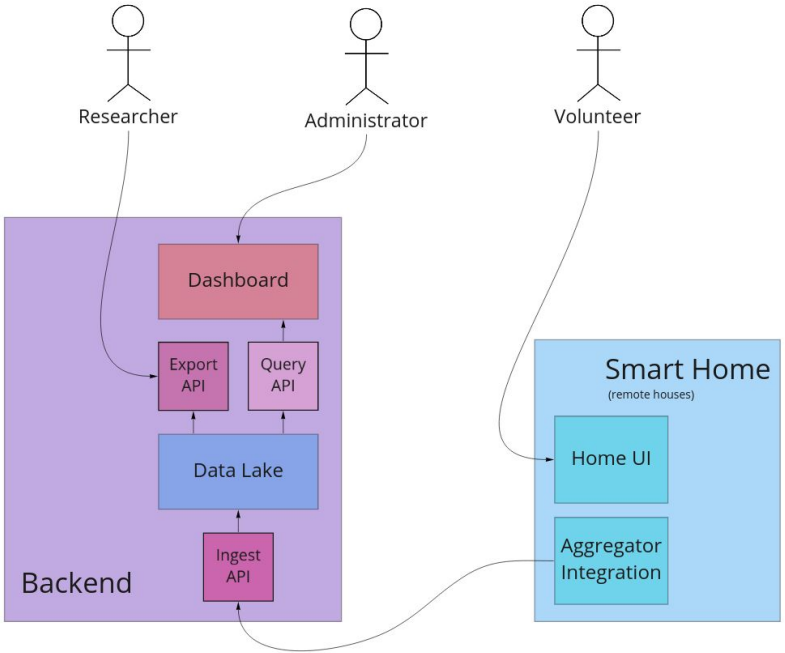
Issues:

- ❏ Most datasets pertain to a single house;
- ❏ Many datasets are over 10 years old now;
- ❏ Lack of data visualization;
- ❏ Lack of real-time data updates;

Source: http://casas.wsu.edu/datasets/

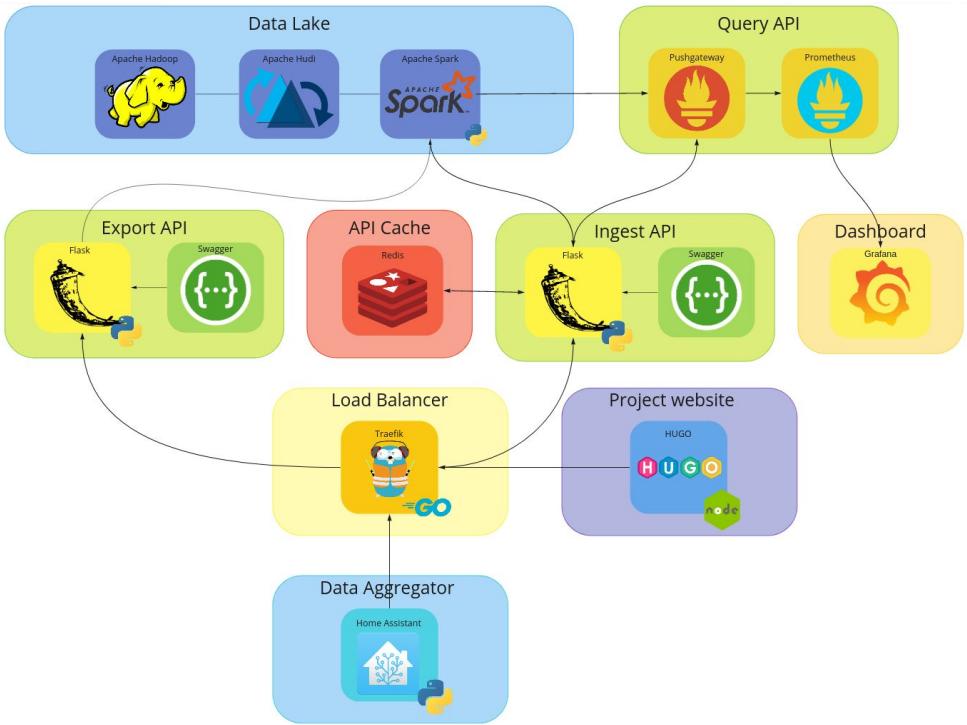
# Architecture

*Logical*



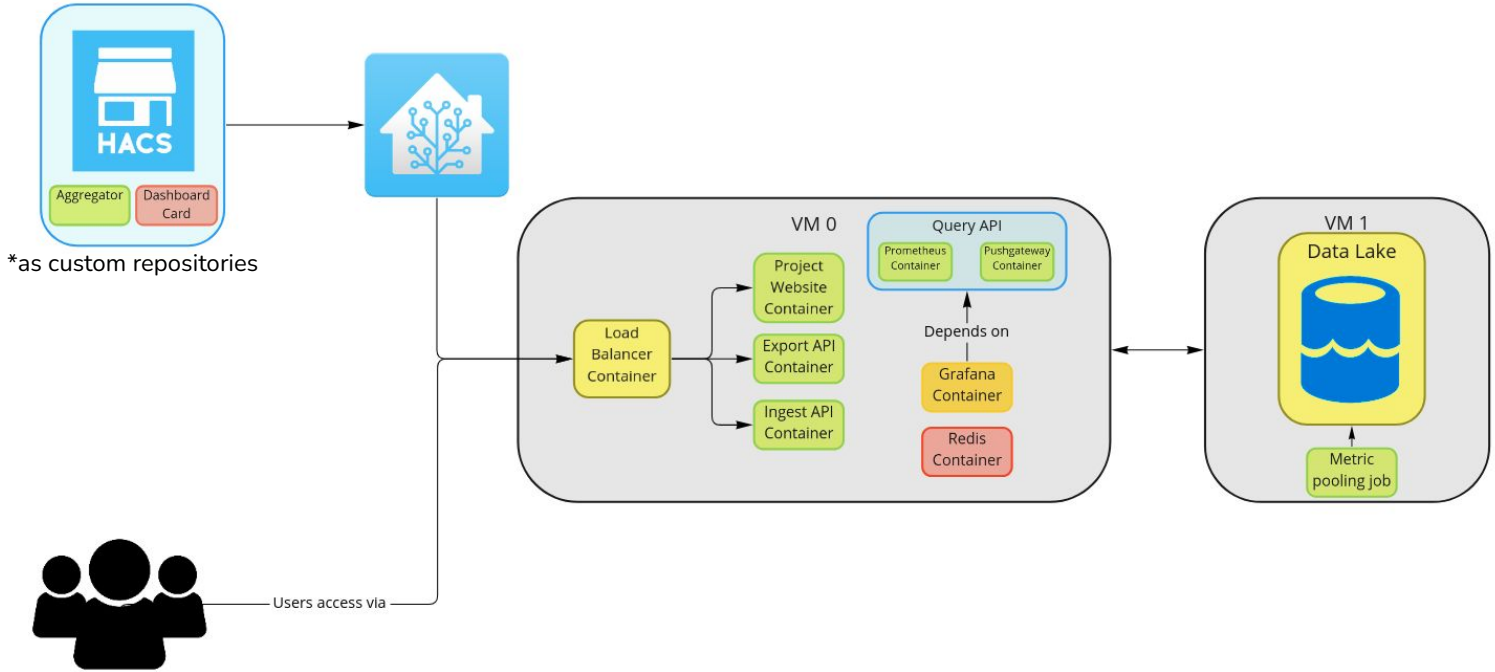
# Architecture

*Implementation*



# Architecture

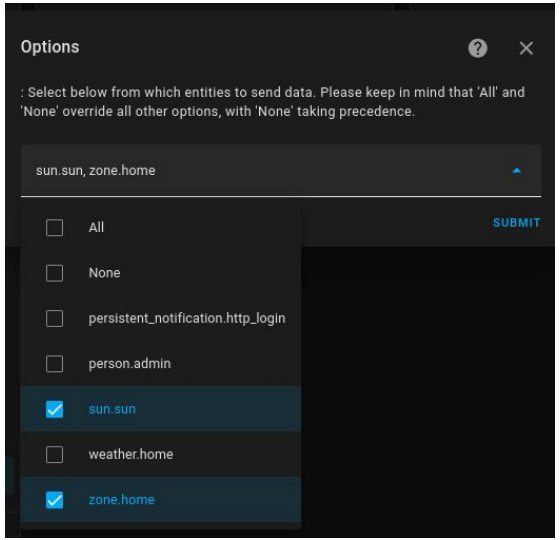
*Deployment*



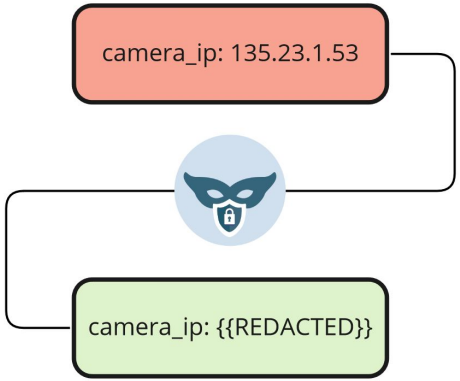
# Home Assistant

*Aggregator*

The user can add sensors to a whitelist, to control what data is sent.




The data is scrubbed of Personally Identifying Information using the python library Scrubadub...




...and some custom methods.

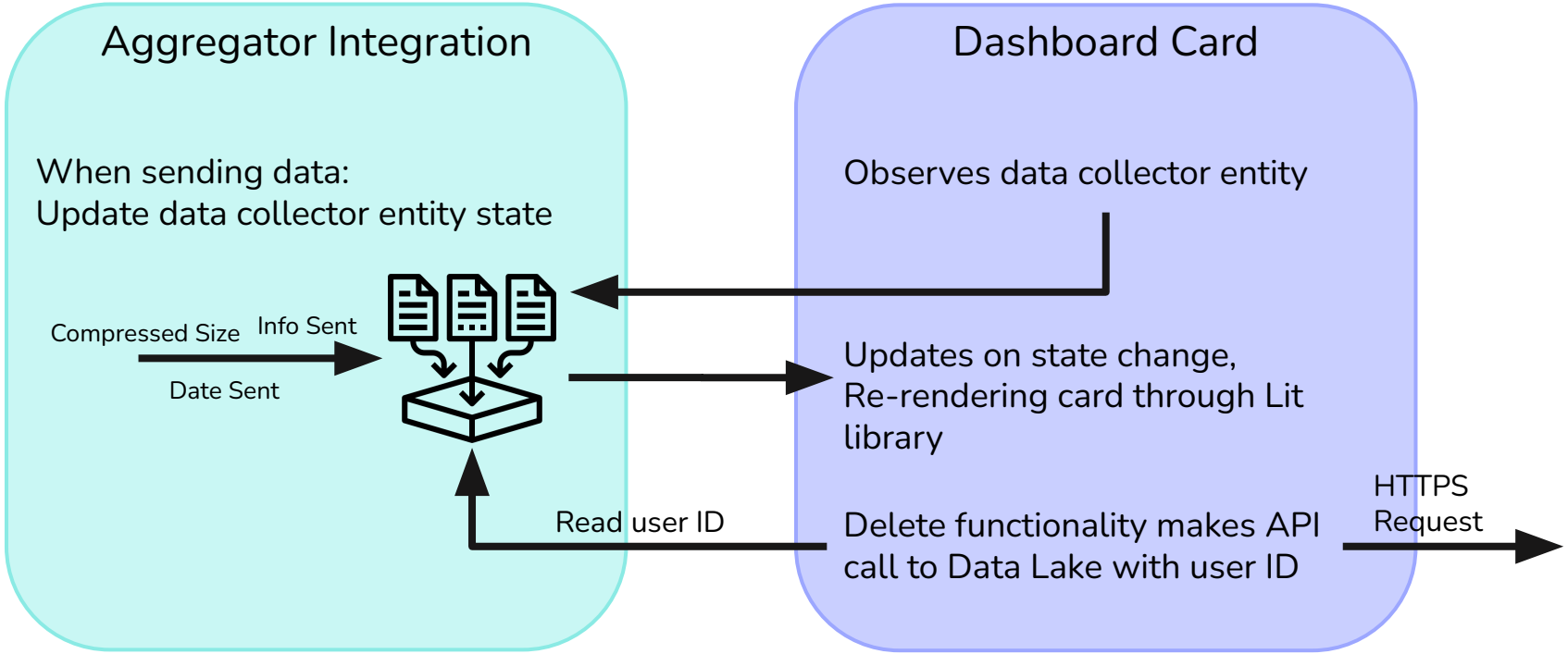
Data filtered:

- IP Addresses
- Postal Codes
- Email Addresses
- Phone Numbers
- URLs
- Geo-Coordinates
- User IDs (For HA)

The data is sent periodically at a set time, that is randomized for each user to avoid overloading the Data Lake

# Home Assistant

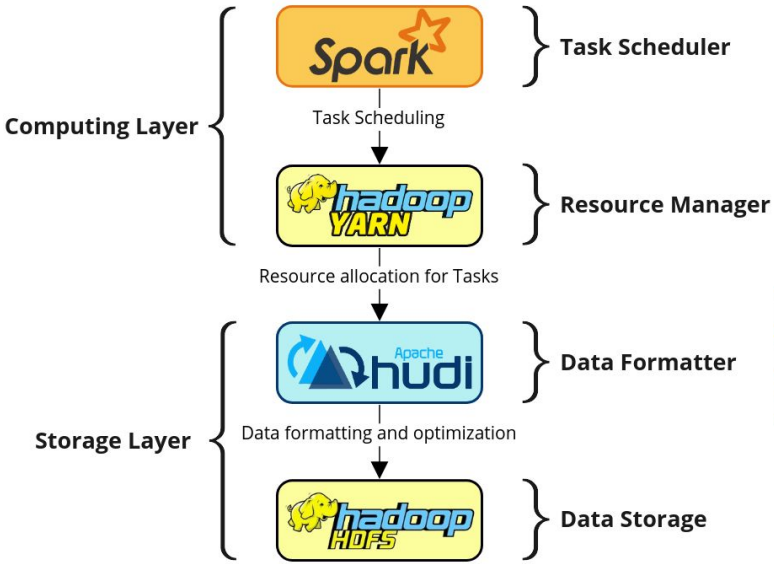
*Dashboard card*



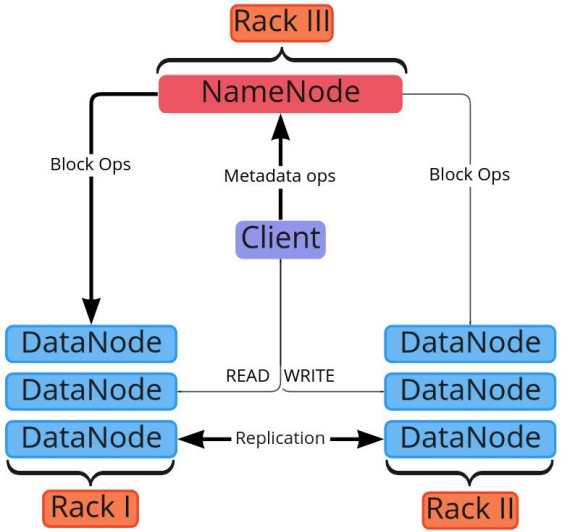
# Data lake

*Storage*

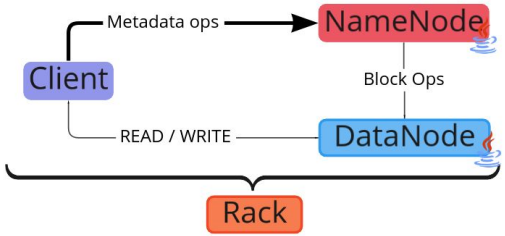
## Data Lake Architecture




## Normal Installation

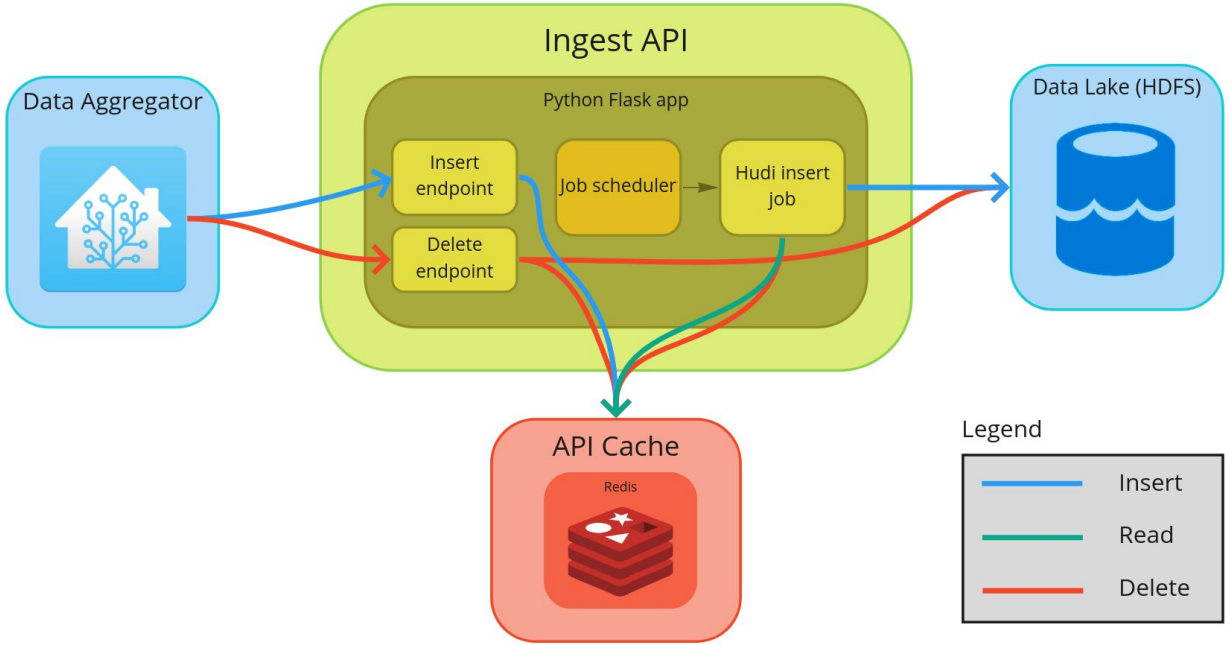



## Single Node Installation



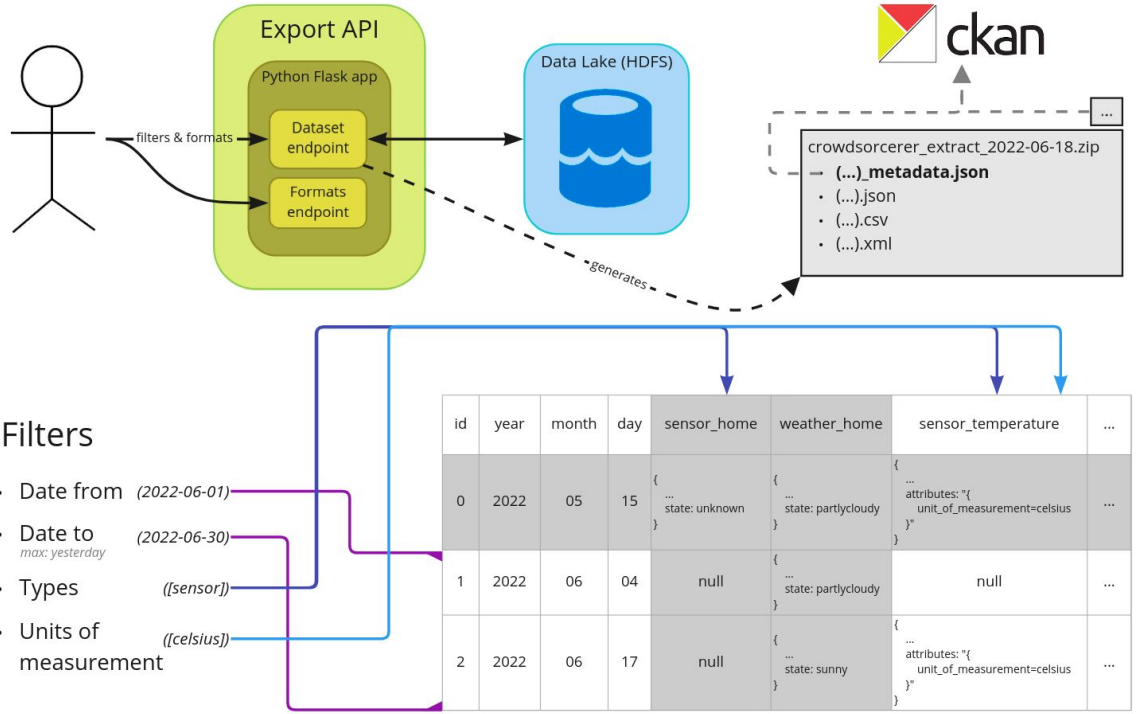
# Data lake

*Ingestion*



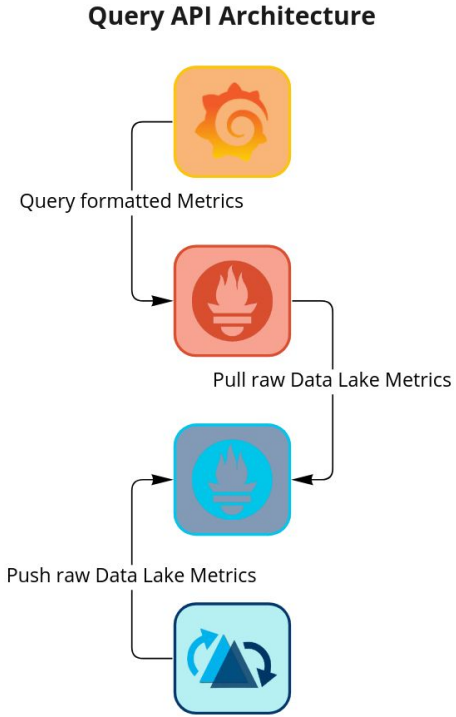
# Data lake

*Exportation*



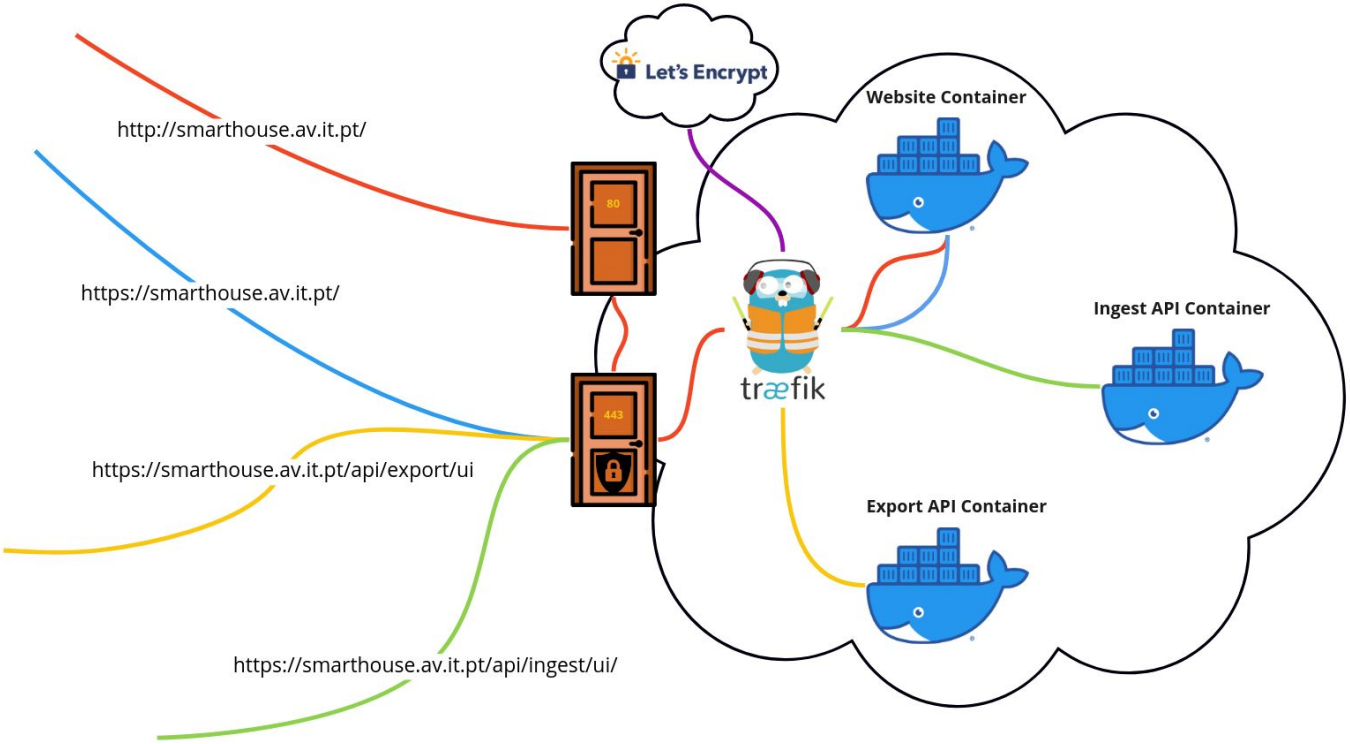
# Data lake

*Metrics dashboard*




# Data lake

*Reverse proxy*



# Discussion

*Home Assistant*

### Aggregator - Problems

- Sparse documentation for Integration development led to delays and a severe underestimation of the work needed;
- “Invisible” bugs and errors, of critical severity;
- Necessity to fork the anonymization library due to version conflicts;
- Testing in varied hardware revealed severe performance issues that led to filters being scraped;

### Lovelace / Dashboard Card

- Trouble starting development due to misguided efforts;
- More limited functionality than expected led to some features being scrapped;
- Separate installation from Aggregator is somewhat cumbersome;

# Discussion

*Data lake*

Core problems:

- Lack of experience with the technology stack;
- Lack of similar documented approaches;
- Misleading documentation;

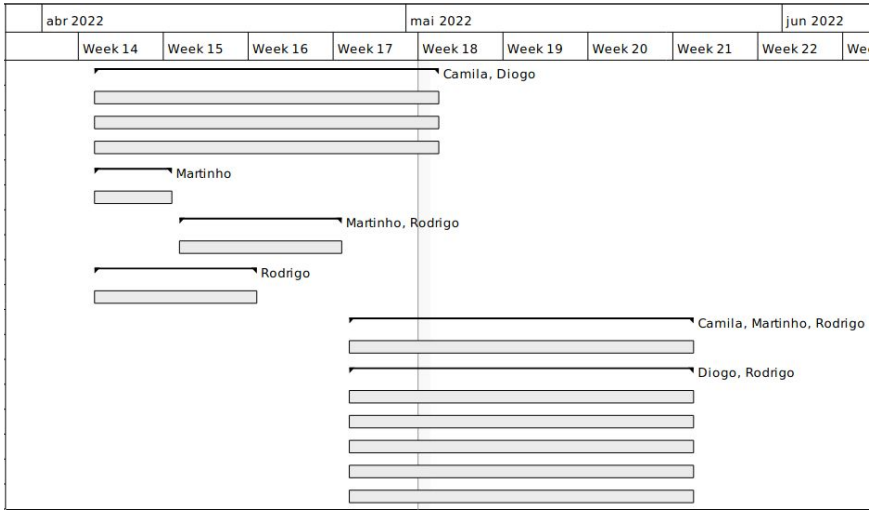
Proposed solutions:

- Use of MongoDB instead of the current Data Lake;
- Improved inception phase

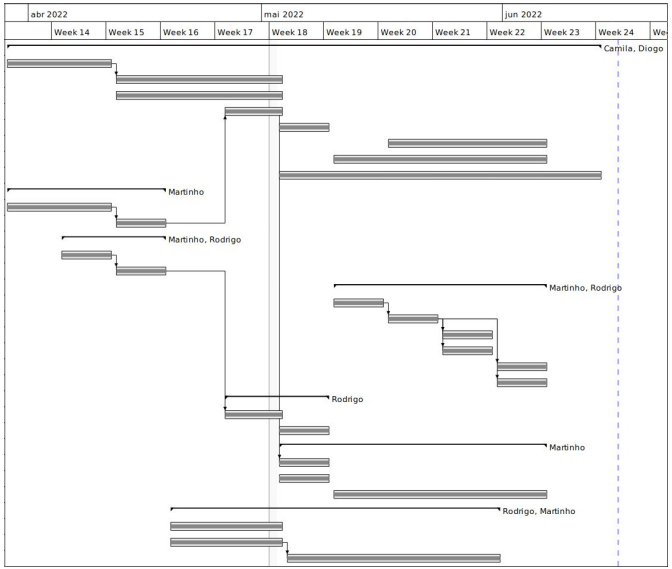
# Discussion

*Project management*

## Roadmap at inception




## Roadmap at transition



# Conclusion

Despite the troubles faced, the system is minimally functional. However, some things have to be worked on:

- Increase scalability
- Better performance both on HA and backend
- Deeper attention and care for protecting user privacy

This project warned us to apply better risk management, as most of the development time was spent on learning and fixing problems

# Thank you