students
@deti
2022

# Crowdsourcing Smart Home Data

Martinho Tavares, Diogo Monteiro, Camila Fonseca, Rodrigo Lima
Orientador: Prof. Diogo Gomes

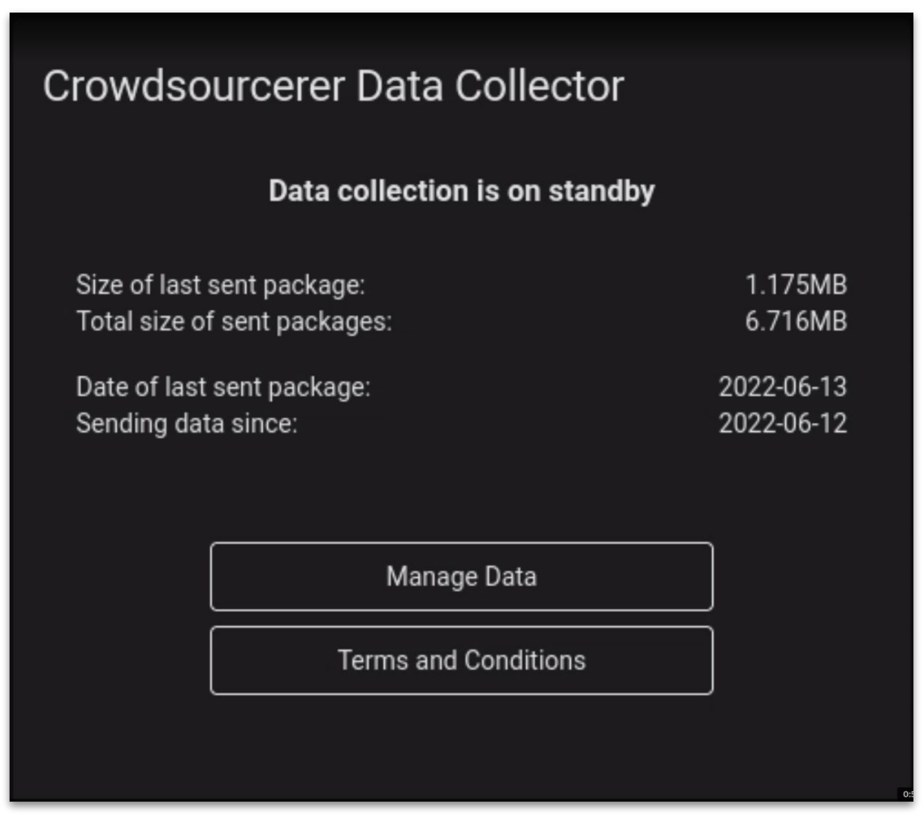
Projeto em Engenharia Informática, 3º ano, LEI.

## Abstract

The market for smart homes is large with huge growth expected in the near future. Smart homes are homes with automation systems monitoring and controlling home elements such as lighting, temperature, and appliances.

In spite of a large number of existing smart homes and their predicted growth, few public statistics and data sets exist on their use on a day-to-day basis. Instead, most studies and analytics focus on smart home adoption and market value. It may be difficult and biased to conduct research on the specifics of smart home usage, depending on the methods used to collect the data.

This realization led to the creation of the "Crowdsourcing Smart Home Data" project, which provides a starting point for future home automation research by assembling data sets compiled from the data of real smart home users.



Crowdsourcerer Data Collector

Data collection is on standby

Size of last sent package: 1.175MB
Total size of sent packages: 6.716MB

Date of last sent package: 2022-06-13
Sending data since: 2022-06-12

Fig 1 - Custom Lovelace card

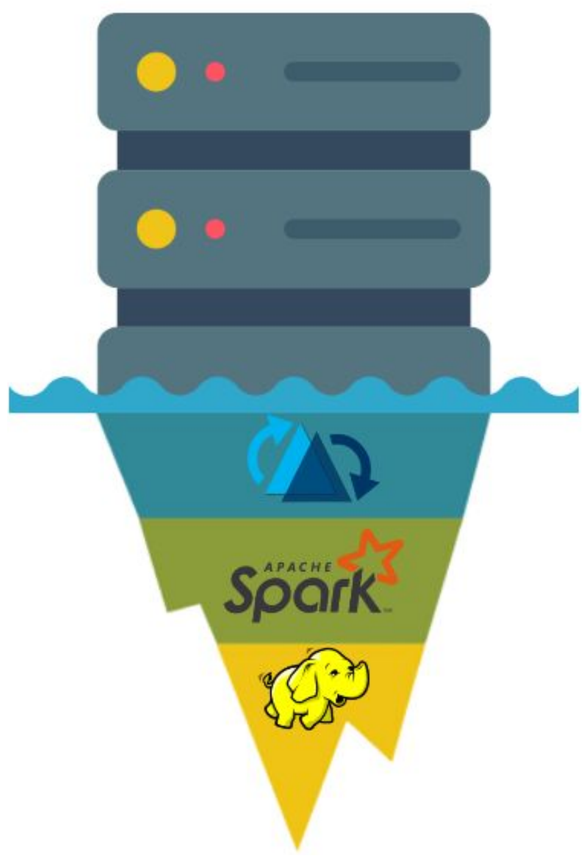


Fig 2- Data Lake Tech Stack

## Objectives

This project’s main goal is to obtain a dataset for further research on the topic at hand, and allow for it to be queried, analyzed, and exported. In order to accomplish this, we used:

- A data lake as storage with schema evolution;
- A Home Assistant integration to collect this data;
- A Dashboard to display metrics related to the platforms’ status;
- A means to export the data in CKAN-compliant formats, so that it can be readily used in future work;
- And an API to allow querying the data stored in the data lake.

Furthermore, the data we collect should be put through an anonymization process in order to remove Personally Identifying Information, to ensure the volunteers’ privacy.

## Methodology

Using HDFS, YARN, Spark, and Hudi, the Data lake is the storage component of the project.

A Flask-based Export API using PySpark enables CKAN-compliant data exportation.

Grafana allows administrators to query and visualize data related to the system itself.

The Query API, using Pushgateway, allows the user to query data via Prometheus.

Data is cleaned using the Scrubadub library, and sent to Data Lake encrypted with SSL.

Home Assistant records a large size of data in its MySQL database, which we retrieve.

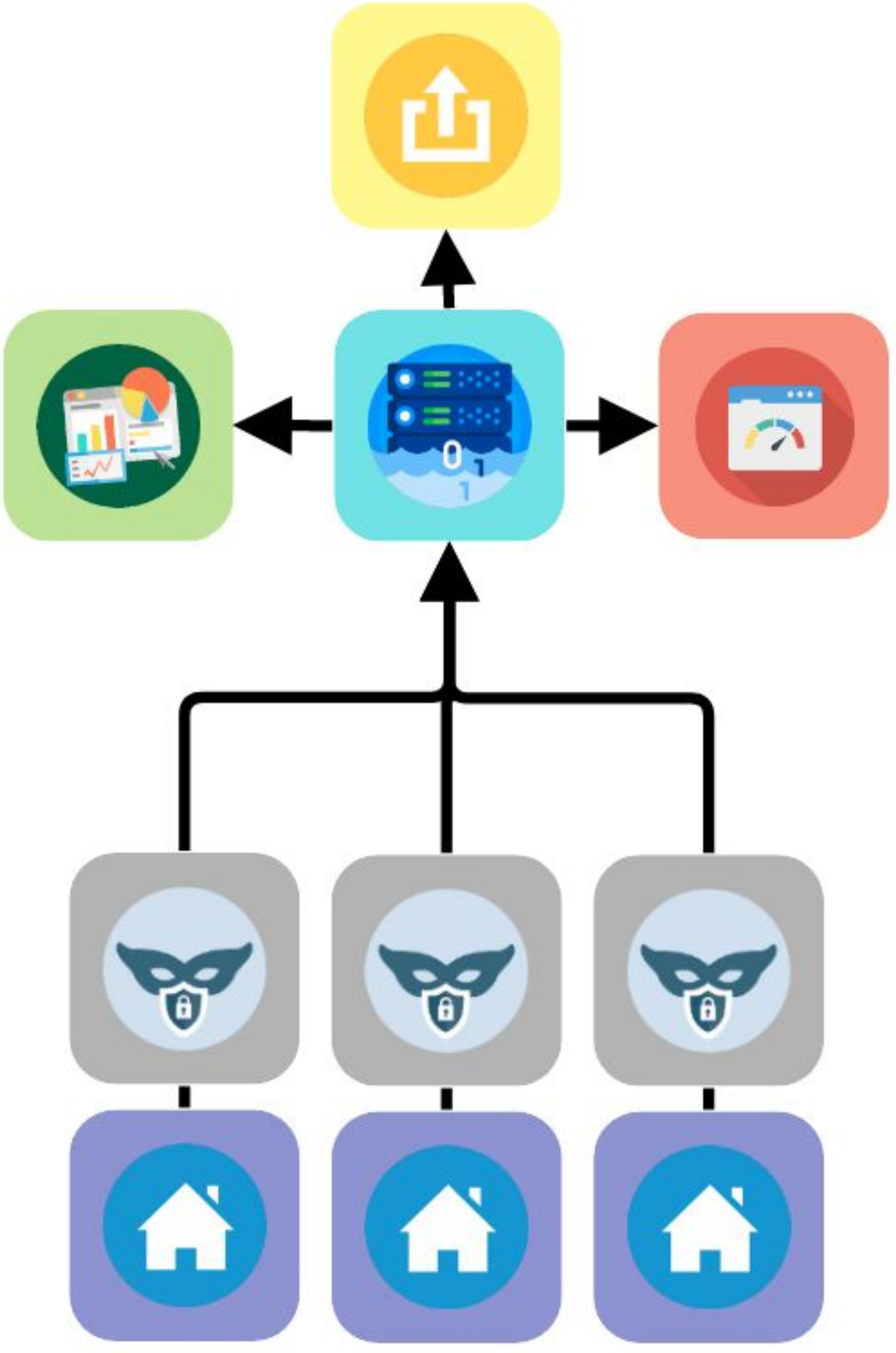


Fig 3- High Level Architecture

## Conclusions

Our initial effort was to create an integration for the open-source Home Assistant project. We would gather data from this, and later analyze and store it.
After that, we developed a platform to store the provided data, built using Hadoop, alongside other technologies.
We later developed querying methods with support for visualization and exportation, while bridging the integration and storage.
Finally, we hardened security, improved stability, and created a web documentation support system. Despite the many hardships faced, we ended up with a minimally functional system, available for further extension to improve the anonymity of its data.

CrowdSorcerer
Organisation@Github

instituto de
telecomunicações

universidade de aveiro
theoria poiesis praxis

deti departamento de electrónica,
telecomunicações e informática