Crowdsourcing Smart Home Data
Abstract

Team Leader - Martinho Tavares[1], Frontend Dev - Diogo Monteiro[1], Backend Dev - Camila Fonseca[1], DevOps Master - Rodrigo Lima[1], and Supervisor - Diogo Gomes[1]

[1]DETI, University of Aveiro

***Abstract*—The main objective of this project is to, through a crowdsourcing mechanism integrated in smart homes, gather anonymous information from them. The base of this project is the Open Source Software Home Assistant and its Portuguese community.**

The project involves the development of a custom component for Home Assistant to extract data and collect informed consent from the user, a Data Lake to house the smart home data, and a Dashboard for analyzing the platform health and overall data status.

With this, the project aims to generate value by providing the collected data as datasets for further analysis, such as energy consumption statistics. Since we are dealing with information that can be considered sensitive, providing informed decision making (choosing what to share), right to be forgotten, anonymization techniques and data security are important concerns that were taken into account while designing this system.

Various challenges were faced during development, which led to certain decisions being taken in order to manage risk. Despite this, we ended up with a functional system, although with performance issues that need to be taken into consideration, and simple anonymization approaches that require further work for usage in a production environment.

***Keywords*—** Crowdsource, Smart home, Data lake, Anonymization, Home Assistant

I. Introduction

Smart homes are homes with automation systems that monitor and control home elements such as lighting, temperature and appliances. Despite the large growth that smart homes have had in recent years[1], few datasets exist of their daily usage that incorporate a large set of covered users and/or features. With the intention of providing datasets that are not limited in this regard, we attempted to develop a crowdsourcing solution to aggregate data from various smart homes who volunteer to offer their data.

For this, we built a system for extracting data from a smart home application into a data lake, by building an extension to a smart home system that volunteers may install to collect data, a data lake structure to store it, and a set of APIs for data insertion, querying and extraction into datasets.

Our solution is open-source, taking advantage of other open-source tools and technologies, in order to be transparent in our collection and treatment of data.

Since the data may allow for identification of the users, anonymization and privacy procedures have to be applied in order to secure their data. Because of this, compliance with the General Data Protection Regulation (GDPR) is a requirement that had to be carefully followed.

II. Procedure

We decided to develop extensions to the Home Assistant smart home application, which included:

- a Data collector component that obtains the data from the different sensors and integrations installed in Home Assistant, allowing configuration of which sensors to include for data upload
- a Dashboard card that allows users to provide informed consent to data collection, visualize their information about their sent data and opt-out from collection

For the data lake we opted for storing the data in a Hudi table hosted on an Hadoop Distributed File System (HDFS). Interaction with the HDFS is done using Apache Spark with Hudi.

The status of the data lake platform, as well as information of the data within it, is presented in a Grafana Dashboard through different metrics that were created in a Prometheus time series database.

The 3 APIs developed are as follows:

- Ingest API: ingestion of data into the data lake. A Python server extended from a server stub generated from an OpenAPI documentation file, using PySpark for communication with the Data Lake and Redis for caching
- Query API: provider of metrics to the Grafana Dashboard, which is the role taken by Prometheus
- Export API: exportation of the data lake's data into CKAN compliant datasets, allowing filtering by time interval, sensor type and unit of measurement. It's built with a similar structure as the Ingest API, a Python server with PySpark for querying the Hudi table containing the smart home data

III. Discussion

Implementing crowdsourcing of potentially sensitive data by using open-source projects for which we lacked familiarity posed a set of challenges in terms of project management to reduce risk and time spent on fixing issues, which is evidenced by the changes to the project's roadmap.

In order to solve some of the problems we faced, we had to resort to other tools and follow compromises, as well as slightly changing some requirements, such as the Dashboard card location and specifications of the Query API.

Adoption of methodologies such as pair programming could have attenuated the issues we met, and prevented the isolation of work currently verified, where the different modules were largely implemented by a single person who has all the knowledge of that part of the project.

IV. Conclusion

The complexity of anonymizing publicly acessible data, coupled with the novelty the technologies had to the team brought difficulty into implementing all the requirements that were set in the beginning.

Some requirements could not have been fully met, especially the one concerning anonymity of the users' data. Since this is a complex problem that requires careful analysis of the data stored and the data that is publicly exported in the datasets, we had trouble implementing a solution minimally compliant with this requirement.

Promoting more involvement of each member into the various project modules seemed counter-intuitive for timely deliveries, but could have been the best approach for an agile workflow resistant to blocking issues.

Despite this, we still ended up with a functional system for crowdsourcing data from Home Assistant instances to a data lake, with a dashboard presenting the platform's metrics and the ability to export the data to a CKAN compliant dataset. This system is up for further development, especially in terms of performance improvements and deeper anonymity application.

References

[1] M. Armstrong, "The market for smart home devices is expected to boom over the next 5 years," Jun. 2022, [Online; accessed 5. Jun. 2022]. [Online]. Available: https://www.weforum.org/agenda/2022/04/homes-smart-tech-market