Privacy-Aware Infrastructure for Managing Personal Data

Personal Data Arbitering within the Databox Framework

Yousef Amar, Hamed Haddadi
Queen Mary University of London
{y.amar, hamed.haddadi}@qmul.ac.uk

Richard Mortier
University of Cambridge
richard.mortier@cl.cam.ac.uk

ABSTRACT

In recent times, we have seen a proliferation of personal data. This can be attributed not just to a larger proportion of our lives moving online, but also through the rise of ubiquitous sensing through mobile and IoT devices. Alongside this surge, concerns over privacy, trust, and security are expressed more and more as different parties attempt to take advantage of this rich assortment of data.

The Databox seeks to enable all the advantages of personal data analytics while at the same time enforcing **accountability** and **control** in order to protect a user's privacy. In this work, we propose and delineate a personal networked device that allows users to **collate**, **curate**, and **mediate** their personal data.

CCS Concepts

•**Security and privacy** → **Privacy protections; Information flow control;**

Keywords

Personal Data, Privacy, Networks

1. INTRODUCTION

Over the past decade alone, the world has seen an explosion in the quantity of personal data people produce daily [4]. As storage space for our online and social media data cheapens, and sensors become more and more ubiquitous through wearables, mobile, and IoT devices, this quantity continues to increase exponentially.

SIGCOMM '16, August 22–26, 2016, Florianopolis, Brazil
 ISBN 978-1-4503-4193-6/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2934872.2959054

Meanwhile, privacy concerns are becoming an ever more common theme in the modern zeitgeist. The market forces that drive the current rate of technological development seem to have long outpaced the rate at which new legislation comes into effect that protect rights such as an individual's privacy. As people release more data into the wild, occurrences of breaches of privacy are similarly rising at alarming rates [3].

There is a distinct need to rebalance the distribution of power between data subjects and data controllers, while at the same time not impinging on the main advantages that full control of data provides to controllers. There have been various fragmentary attempts at addressing this need, but these are generally fraught with the shortcomings of centralisation and/or implicitly overexpose personal data to third parties [1].

In this work, we propose and delineate a solution that not only covers these issues, but additionally provides the means to more effortlessly draw on a multitude of vastly differing sources of data out of the box to augment or complement analytics: the Databox [1]. It solves this need for privacy by enforcing **accountability** and **control** by design, and provides a means for users to **collate**, **curate**, and **mediate** their personal data.

2. APPROACH

The Databox is a personal networked device with a form factor comparable to a home router [1]. It is configured to be able to access a user's personal data from a variety of sources. These include but are not limited to online data (social media, online banking, email), mobile data (motion sensors, GPS), and IoT devices (temperature sensors, light sensors, wearables).

This data is what is usually collected by third parties for analytics. In most cases however, a third party only needs a limited set of results derived from a dataset, and not the dataset itself. The Databox serves as a platform upon which the processing of personal data can be done locally, in the context of Databox *apps*, and only emit the necessary results to a third party, thus protecting the privacy of the user by design.

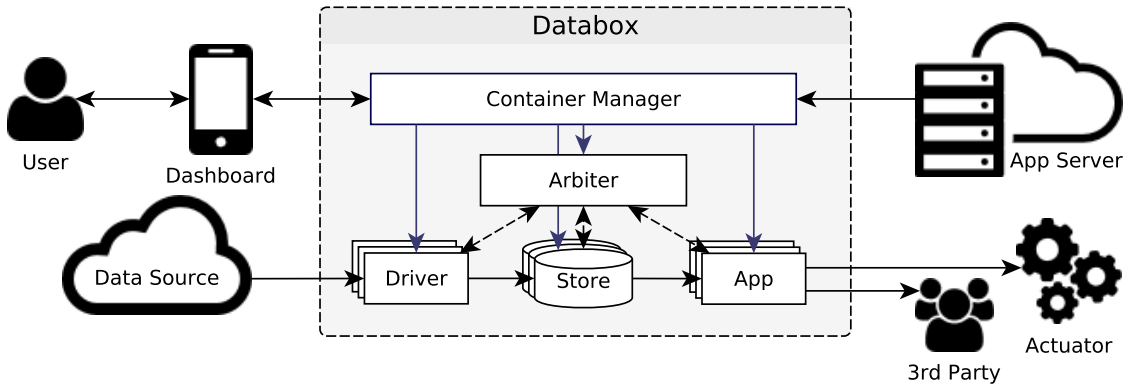

Figure 1: A High-level Overview of Databox Components

2.1 Databox Architecture

The Databox architecture, depicted at a high level in figure 1, is spilt into a number of discrete components. External data sources are interfaced with through data *drivers*, which can make data available to *apps* for processing, by writing them to system-managed data *stores*. It is important to note that although data may be stored internally, a Databox should *not* be considered a data silo, but rather a platform through which data can be accessed from arbitrary sources and processed locally.

Apps, loaded from a remote app store, each declare a *manifest* – a set of app metadata that include a definition of possible permutations of privacy permissions. This information is presented to the user in the form of a service level agreement (SLA) alongside the intended uses of these data sources, and risks and inferences. Through the SLA, a user is able to "negotiate" and control access to match their subjective privacy needs. Access control extends to more specific granularity restrictions such as how long a data source is accessible, at what frequency it can be accessed, and how much data can be read, among other source-specific dimensions of granularity.

This agreement is enforced by the *arbiter* at the heart of the OS. Its primary purpose is to mint refinable bearer tokens verifiable by *stores*. Beyond that, the *arbiter* is responsible for logging all flow of data to enforce accountability. Based on the information provided by the *arbiter*, a user has a full overall view of what happens to their data and why, and can react accordingly by controlling access – especially in cases where a third party is involved (such as market research firms, universities, government, hospitals, etc.) – through a *dashboard* on any client device.

Databox architecture is privacy-aware by design. All apps are isolated in sandboxed, virtualised containers and communication with the outside world is heavily restricted. This, coupled with the open-source nature of the project and heavy scrutineering of third party apps, ensures that a user's privacy is well protected.

3. CONCLUSION

The speed at which the way we create, manage, and consume data is changing, mean that this area of research is only just beginning to gain momentum that warrants a great deal of discourse. One can imagine large scale statistics and research being done through Databoxes; "self-filling surveys". To data subjects, the advantages of control over privacy and personal analytics is obvious. To data controllers, economic implications are pervasive; heightened privacy means subjects are less averse to their data being processed.

Preliminary prototypal evaluation demonstrates viable scalability, and paves way for an SDK and widespread development, that will ultimately return power over personal data to the individual and spawn a novel paradigm of personal data analytics.

4. ACKNOWLEDGMENTS

This work was funded in part by the EPSRC Databox project, EP/N028260/1, in collaboration with The University of Nottingham and others [2].

References

[1] A. Chaudhry, J. Crowcroft, H. Howard, A. Madhavapeddy, R. Mortier, H. Haddadi, and D. McAuley. Personal data: thinking inside the box. In *Proceedings of the fifth decennial aarhus conference on critical alternatives*. Aarhus University Press, 2015, pages 29–32.

[2] Databox project – EPSRC project on privacy-aware personal data platform. URL: http://www.databoxproject.uk/ (visited on 04/07/2016).

[3] Identity Theft Resource Center. Itrc breach statistics 2005 - 2015. 2016. URL: http://www.idtheftcenter.org/images/breach/2005to2015multiyear.pdf (visited on 03/09/2016).

[4] J. James. Data never sleeps 2.0. 2014. URL: https://www.domo.com/blog/2014/04/data-never-sleeps-2-0/ (visited on 03/09/2016).