Databox: Privacy aware personal data processing

We are all subject to large quantities of data being gathered about us. This data is often processed to extract inferences and value by the companies that hold our data. Databox a novel platform designed to provide domestic users with a privacy-preserving option for data storage while still allowing processing by third parties.

Databox aims to put users back in control of their data by holding it locally on hardware they control rather than in siloed cloud-based systems. Databox is motivated by recent moves to strengthen data privacy regulations. In light of this, storing, managing and maintaining personal data now carries a significant burden and risk for data processors. Consumers are also becoming more aware of the consequences of sharing and uncontrolled collection of their personal data.

Databox provides a platform where data subjects and the data processors can extract value from data in an accountable and transparent manner. Data subjects are able to collate, curate, mediate access to their personal data. Additionally, data processes are able to request access to locally process data on a user's Databox to extract the information they require without having to hold the data on their servers.

Conceptually Databox separates data collection (performed by Drivers) from data processing (performed by Apps) and provides the socio-technical infrastructure to enable data sharing with the minimum leakage of personal data. The point at witch Apps and Drivers are installed onto a Databox provides an important interaction opportunity where the purposes for processing can be explained, data exported surfaced and informed consent sort. Databox requires that all Drivers and Apps provide a Manifest describing what data they produce, require access to and plan to export. These Manifests encode the permissions required for an App to function and can be used to express to the user how their data will be used before any data is processed or exported.

Databox uses Docker containers to sandbox components. Docker containers provide an easy and flexible way to deploy code in any language, along with any required libraries, onto a Databox. Communication between containers takes place over the Docker network, which is tightly controlled. The Databox Core-Network only connects components that require access to each other and limits access to the local area network and the wider Internet.

Within Databox, all data storage and retrieval are performed through a trusted system data store. The store provides Git-based storage and an audit trail for all actions performed on the data they contain. Stores are used as intermediaries between Drivers and Apps. This creates a point of audit where all transactions can be monitored. Fine-grained access to data is controlled through the used of macaroons, minted by the Arbiter (a trusted system component that holds the granted permissions for all components) and verified by stores.

Databox currently a work-in-progress with a research implementation available on GitHub. We are in the processes of testing the ideas encapsulated in the platform with end users and industry. We look forward to sharing a more detailed system overview and receiving feedback on our design choices.