Implementing a Personal Container

Chris Elsmore

October 2011

Abstract

This document explains the motivation behind setting up a personal container style data store, and the advantages such a container can provide. Online services are becoming more popular and more numerous, however with this growth comes greater scattering of personal data. We discuss using a Personal Container allows this data to be stored where the user chooses, in the cloud or on a computer at home, preventing data loss in the closure of an online service. This also allows a user to more closely allow or deny access to their data, along with letting third parties run algorithms on their personal container and return results without exposes the raw information. Using an implementation of the Personal Container paradigm called Locker, we can setup infrastructure internally, and use it to provide detailed information to a user based on their personal and sensitive data without exposing the raw data and compromising privacy. We show an example of this infrastructure by deploying it within the University as a pilot study, to identify the benefits both to employees who can use the container to store information, and the university that can request access to this information for survey's, and future decision making.

1 Motivation

1.1 Current Practice & Limitations

Online services and social networking websites have surged in popularity and now enjoy vast user bases - Facebook currently has 800 Million active users[2], and Foursquare has acquired 100 Million Users in its 27 month lifetime[3]. Sites also store a significant amount of user generated content - members upload 8 years worth of video using YouTube, 250 million photos to Facebook and generate over 200 million tweets on Twitter each day[8],[6].

The users who generate this data must trust that these sites will continue to exist to continue accessing this content. In the case of photo sharing sites the user most likely has copies of the images elsewhere, but does not have their own copy of the metadata generated after upload for example conversations regarding photos. Sites like Twitter and Foursquare by their nature don't generate any data locally, and all tweet and checkin data is stored remotely by the service. If these sites were to disappear, each user would not have a copy of their information without manually accessing, often using custom code using an API, a high barrier of entry for the non-technical user.

Even if the users are technically savvy and willing to write code, or trust 3rd party applications that access data on their behalf, users with a wide reaching digital footprint will require many different apps accessing many different API endpoints to reach their data, all with their own idiosyncratic behaviour within standards such as OAuth[1], used for authentication. Access to this data is further burdened by the use of certain 3rd party applications which in the case of Facebook have been known to have security and privacy concerns [4] [7] [5] [1].

[1] More information available at `http://oauth.net/`

2 Personal Containers

2.1 Description

Personal Containers are designed to solve the problem of the fragmentation of user data throughout the social web, compute additional, higher resolution and more accurate personal data by combining multiple data sources, and facilitate tighter security around access to these data. They are basic data storage and access systems, and can range from simple databases of information on a single computer hosted by the user, to sophisticated systems that act as signposts, storing the location of all user data on multiple devices and services, and proxying requests to the relevant data sources.

2.2 Benefits

Personal Containers provide a central access point to and allow autonomous collection of a users data from multiple sources, with data such as photos, tweets, status updates, checkins etc. Since the container provides direct access to multiple sources of data, it is able to generate far higher resolution data than using data from just one service- for example using geo-tagged tweets, photos and checkins along with Google Latitude[2] data to provide highly accurate location information. It also allows faster and simpler access to multiple data, enable simple queries such as requesting all photos taken between two dates.

Personal Containers also ofter the possibility of increased security and data control. Personal data is by definition personal, and thus highly sensitive. The example above demonstrate the need to added security, when dealing with highly accurate and real-time location data for example. The personal container can use multiple factors to approve or deny requests based on traditional variables such as rate of access, and simple authentication, as well as time based restrictions such as validity windows where data access is only allowed for a certain time, and impose certain restrictions based on media type such as picture or location resolution.

The container can also allow results based on computation on personal container data to be accessed without allowing direct access to the data itself.

The data security befits are not only a matter of authentication, personal containers can act as sophisticated backup mechanisms of sensitive data, de-duplicating and federating throughout storage nodes such as personal computers at a users home or workplace, or instances running on cloud computing platforms such as Amazon's EC2[3].

[2] More information available at `http://google.com/latitude/`

[3] More information available at `http://aws.amazon.com/ec2/`

3 Implementation

3.1 Architecture

The Locker Project[4] aims to construct a personal container suitable for hosting on a users general use computer, as well as adapting to a cloud based platform or a dedicated machine hosted personally. It is written in Javascript on the Node.js[5] platform on the backend, and standard HTML and Javascript on the web-based user interface.

3.2 Infrastructure

Our deployment infrastructure involves multiple instances of the locker platform, installed on multiple virtual machines running a Linux OS with a secure proxy frontend. The locker codebase supports only single users with no authentication, so each instance is isolated within it's own virtual machine. A central Nginx[6] server handles SSL encryption, authentication and acts as a proxy to direct to the correct instance based on the hostname requested internally to the host machine.

The deployment is hosted by the university cloud computing service, as an example of an institution hosting secure personal data storage, in return for access to said data in an open, and transparent way, identifying to the user exactly which data will be used, and for what purpose.

[4] More information available at `http://lockerproject.org/`

[5] More information available at `http://nodejs.org/`

[6] More information available at `http://nginx.org/`

3.3 Advantages

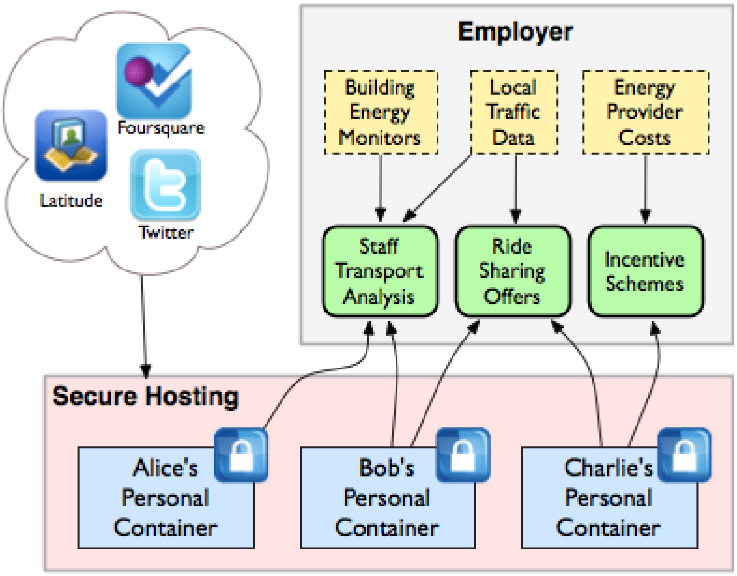

Figure 1: C-Aware Personal Container deployment layout.

As shown in figure 1, using this system an employer can offer a range of schemes to it's employees, as well as using this data for it's own projects. For example the employer can use this data to fill out it's own transport survey questionnaires, gathering data on how it's employees travel to and from work, and use this to help decide weather a new car park is needed, or better cycle routes on its property.

It also enables the university to deliver more personal benefits to its users, for example knowing how much and when a user uses their energy can be used to provide a price comparison feature that shows exactly how much a users energy bill would be with different suppliers.

References

[1] K. Bankston. Facebook's new privacy changes: The good, the bad, and the ugly. Available at https://www.eff.org/deeplinks/2009/12/facebooks-new-privacy-changes-good-bad-and-ugly, December 2009.

[2] Facebook. Facebook statistics page. Available at http://www.facebook.com/press/info.php?statistics, 2011.

[3] Foursquare. Foursquare reaches 10 million users. Available at https://foursquare.com/10million, June 2011.

[4] C. Morran. Report: All top 10 facebook apps leaking personal information. Available at http://consumerist.com/2010/10/report-all-top-10-facebook-apps-leaking-personal-information.html, October 2010.

[5] E. Steel and G. A. Fowler. Facebook in privacy breach. Available at http://online.wsj.com/article/SB10001424052702304772804575558484075236968.html, October 2010.

[6] Twitter. Twitter blog: Your world, more connected. Available at http://www.facebook.com/press/info.php?statistics, August 2011.

[7] Z. Whittaker. Facebook applications leak users' personal data to third parties. Available at \http://www.zdnet.com/blog/igeneration/facebook-applications-leak-users-personal-data-to-third-parties/9906, May 2011.

[8] YouTube. Youtube statistics page. Available at http://www.youtube.com/t/press_statistics, 2011.