Protecting User Data with Data Management Service and Isolation

Authors: Valerie Angulo, Adrienne Bouchie, Christopher Davidson, Jingni Liu

Introduction

Intro to User Data

Over the past decade, the most common type of data stolen has been User Data[1]. User data is any kind of data that is created or owned by an individual. This data could be a user's email address or login credentials, a 'like' or 'tweet' on social media, or a credit card number or bank account number. To understand the scope of this issue, we must first understand what type of data is being stored for each user when interacting with certain applications. Payment information and addresses are usually stored within e-commerce to encourage shopping, email addresses and cell phone numbers are typically stored for account recovery purposes, and sometimes social security numbers must be provided for services to verify identity, such as checking one's credit.

In addition to the obvious collection of user data, vast amounts of user data is stored in more concealed forms: When a user is searching for directions with Google Maps, Google tracks the user's location to determine traffic patterns in order to provide the fastest route to a given location. While you are interacting with friends on Facebook and Instagram, your 'likes', clicks, video views, follows, comments, and other user-triggered events are being tracked in order to better understand the user and provide tailored content based on the user's behavior. When listening to music on Spotify or watching a movie on Netflix, there is a constant feed suggesting new songs or movies that a user may like, based on their previous behavior. All of a user's activity within these applications is tracked in order to provide a better user experience.

Out of all the various kinds of user data however, Personally Identifiable Information (PII) is the type of data that businesses should handle with the most care. PII can be used to identify or single out an individual user, and if obtained by undesired parties, can lead to identity theft and credit fraud. This type of data is highly regarded to the individuals who own the data as well as companies utilizing this data to understand user behavior. Although there are regulations surrounding the use of PII, there are also a variety of vulnerabilities that enable attackers to abuse this PII.

Some Vulnerabilities of User Data

One vulnerability of user data is that user data collection is necessary. Businesses and organizations are incentivized to collect user data in order to provide an individualized user experience. This personalized service cannot be provided without knowing certains aspects about the user. Likewise, user information is needed in order to verify the user's identity via authentication. Additionally, user data is also necessary to improve the user's individual experience, like saving one's personal preferences or behavior in order to recommend additional services or provide useful information. Naturally, most companies typically have a

stockpile of user data, however the means of storing and protecting this abundance of user data are not fully understood nor well managed, often leading to data breaches.

Although it is impossible to avoid collecting user data, if this data is not securely protected millions of users can fall victim to data breaches or leaks. In the first three months of 2018 alone, there have been several significant data breaches, from MyFitnessPal exposing 150 million users emails and passwords[2], to Cambridge Analytica exposing 87 million Facebook users activity data[3]. Through this leak, Cambridge Analytica was able to understand the behaviors of over 80 million users by analyzing users' 'likes', comments, gender, and political views. This analysis allowed Cambridge Analytica to create specific content tailored to each user's interests, potentially having a serious impact on the 2017 presidential election through targeting misinformation to certain individuals[(3,4)]. It is clear that the amount of data companies have access to is growing exponentially in order to keep up with users' behavior. Despite this abundance of user data, many services and applications still fail to recognize or follow proper security protocols in storing user data, oftentimes due to the cost of implementing these standards, or unawareness of the standards altogether. Even if companies learn from the mistakes of other companies and implement new security standards in hopes of minimizing the risk of experiencing a data leak of their own, this approach is reactive, not proactive.

An additional vulnerability of user data is that many forms of user data are collected and stored together in a single location. For instance, with small social media websites, user data including PII and posts for an individual user are usually grouped together for quick access and easy implementation in a single location. When handling more precious user data such as credit card numbers, e-commerce sites often securely store transactional data through encryption, while still allowing fast access to all data for options like 1-click checkout. However, despite the potential value of this data, it is still often stored together in a single location in order to benefit from low latency. This exposes a vulnerability of user data due to the lack of isolation: Even though each individual piece of user information lacks significance on its own, such as a mailing address without a name, data breaches in the systems storing this user data can still be easily manipulated to identify a person.

Problem Statements

Problem 1: Failure to Follow Security Standards in Storing User Data

In many instances, data leaks could have been prevented if businesses merely followed security standards and kept an active awareness of changes in protocol. In the 2017 Equifax Breach for example, the flaw that allowed the attack to happen was disclosed in March of 2017, yet Equifax should have had enough time to patch this vulnerability before the attacks occurred in May[5]. The company should have followed basic security protocols, resulting in either having their security team quickly patch the problem efficiently, or make the decision to take the affected app offline immediately. During the Uber Data Breach of 2017, personal data, including names, email addresses, phone numbers, and driver's license numbers, was stored in an

unencrypted format, despite PII encryption being so well known, resulting in a large amount of user data being leaked through an easy to read plaintext format[6]. Incidentally, many businesses fail to follow proper security standards due to the cost of implementing these standards, mere laziness in properly implementing these standards, or just unawareness that these security standards even exist.

Problem 2: Misuse of Data

Another major problem, that relates to the issue of improperly following security standards, is the misuse of users personal data by unauthorized parties. Currently, there is no responsible data oversight methods in place; no one is policing the handling of user data to ensure that it is only used as intended by authorized parties. Failure to check authorization of data access increases the potential threat of an organization or individual being able to collect a large amount of varying data on one or multiple individuals, for their own personal gain or purposes. By gathering and analyzing fragmented data from multiple sources, sensitive information can be obtained from those whose data is collected.

Problem 3: Various Attacks on Data at Rest

Even if businesses properly implement security principles and handle user data with the utmost care, threat agents will always find a new attack paradigm: Encryptions can be decrypted, firewalls and anti-virus programs can be bypassed, and physical machines can be stolen from data centers. If data at rest is compromised or leakage is inevitable, how can we make the leaked data undesirable to threat agents?

Problem 4: Various Attacks on Data in Transit

Likewise, there are various attack mechanisms that target data in transit. Robust network security controls can be implemented, yet users can still fall for phishing attacks, thus breaking authentication. The successful implementation of these attacks depends on the user and their knowledge and practice of good security habits, and would be difficult if not impossible for a system to prevent.

Problem 5: Users are Kept in The Dark About Their Data

A major security problem prevalent with the collection of user data is that many users are neither aware nor in control of the usage of their personal data. Oftentimes users give away important data freely, skipping over an application's Terms of Agreements that allow companies to collect and use user data, or even allow the use of 'friends' data linked through other social media sites. Users should be aware about the kind and quantity of their data being collected, what and for how long their data is being used, where their data is being stored, and for how long that data is being stored. More often than not, there isn't much transparency between users and the data they have given out. Typically, users don't know who is still holding on to old data that was entered years ago, or even worse if that data is being sold to third parties. Although

some data is more sensitive than other types, such as SSN and credit card numbers, access to any type of user data can be very harmful to the parties responsible for housing this data. In the past decade, the amount of user data requested through websites and applications has grown astronomically. As more users turn to mobile devices to get directions, check in with friends, listen to music, or order new computer cases, personal data accumulates. So the question is, how can companies protect this ever growing user data?

Proposed Solution

To better protect user data, we propose a cloud storage based data management service that will help businesses in a cost effective manner.

Intro to our Data Management Service

We would like to provide a cloud data management service where users and businesses can manage all their data. Our data store system does not store any data on our side, rather it is the authentication and communication between businesses requesting to store their data and other cloud storage services. However, our system closely follows the best security practices, laws and regulations while making the storage decisions, and provides recommendations and guidelines to encourage clients to do so as well.

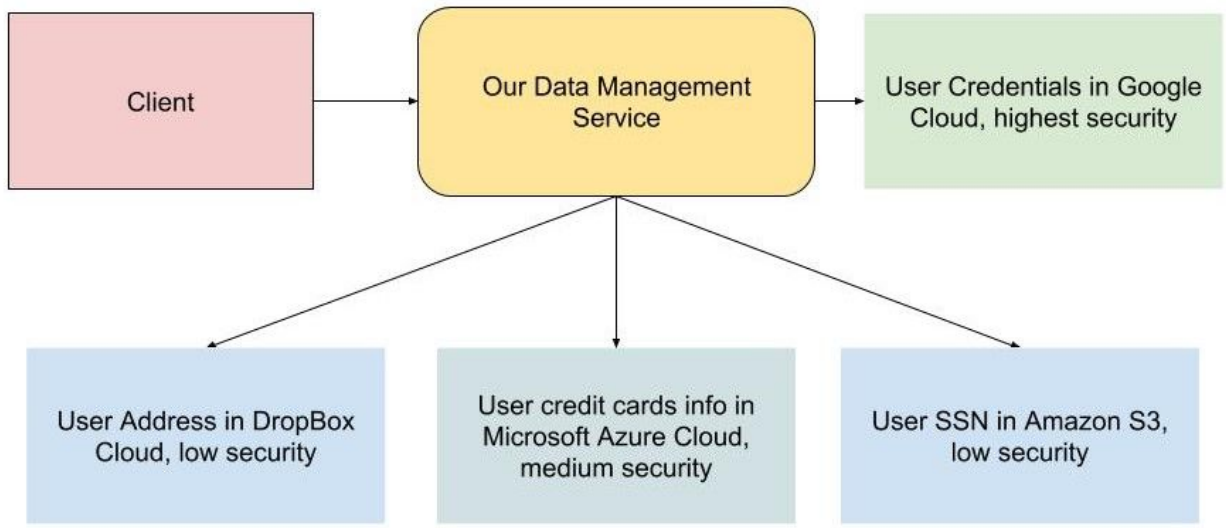
In this system, we can control where our data is stored, specifying for how long it is to be stored, as well as how to store it (e.g. encryption) by asking the client what data type they want to store. This management system will utilize various cloud storage services such as Amazon and Google clouds, and split personal data up into these different services based on an algorithm, making sure that clients aren't given the chance to store any single user's data in a single cloud. We are using this algorithm to ensure that even if one cloud is compromised, only one fragment of data per client would be stored on that system. Additionally, in our second implementation, no sole cloud will be designated as having more important data than another, ultimately making targeting one system over another useless.

Idea of Isolation

Our system provides an algorithm to split each user data and credentials to access this user into various systems based on the idea of isolation. Isolation takes advantage of the fact that individual data pieces are not meaningful if not connected to other pieces in a relational manner. For example, if one cloud storage service is attacked and only users' social security numbers are compromised, no issue arises because no other identifying information would be attached to it. It's a well-known fact the SSNs are 9 digit numbers that can be generated with ease, however they are not meaningful if they are not associated with a name or other personal information.

Simple Sample Implementation I

A very simple implementation of such a system is to store data types according to different cloud storage services, such as storing user credentials in Google Cloud, while also storing addresses in DropBox, credit cards numbers in Microsoft Azure Cloud, and SSN in Amazon S3. As demonstrated in the graph below, only the Data Management Service yellow box would be owned by us.



This simple solution, which utilizes clouds to store certain data based on the cloud services' level of security, is the easiest approach to implement, and would be the most basic option for those who utilize our service. It maintains security in keeping more personal data in higher security cloud systems. However, this method does increase risk of the cloud services with higher security being targeted more than the lower security cloud services, the assumption being that a cloud service with higher security would contain more valuable information. The benefits of this implementation would be less latency, where there will be target clouds for each data type, decreasing lag in randomizing data which could be an issue if there is a lot of data to manage. This method would also be the most cost effective, where fewer resources and less time would be needed to handle the data while still maintaining a solid level of security.

Clients would communicate with our data management service and request to store data, specifying which data types they are trying to store and what level of security they want the data to have. For example, users can decide whether the data being stored should be encrypted and what kind of encryption should be implemented. If clients are storing user credentials, their data will be put into Google Cloud with the level of security they requested. Our system can provide guidelines on how to set the security levels, and even force a client to implement minimum security. This would guarantee that our system is in compliance with security laws and regulations. For instance, if the data type is credit card information, the

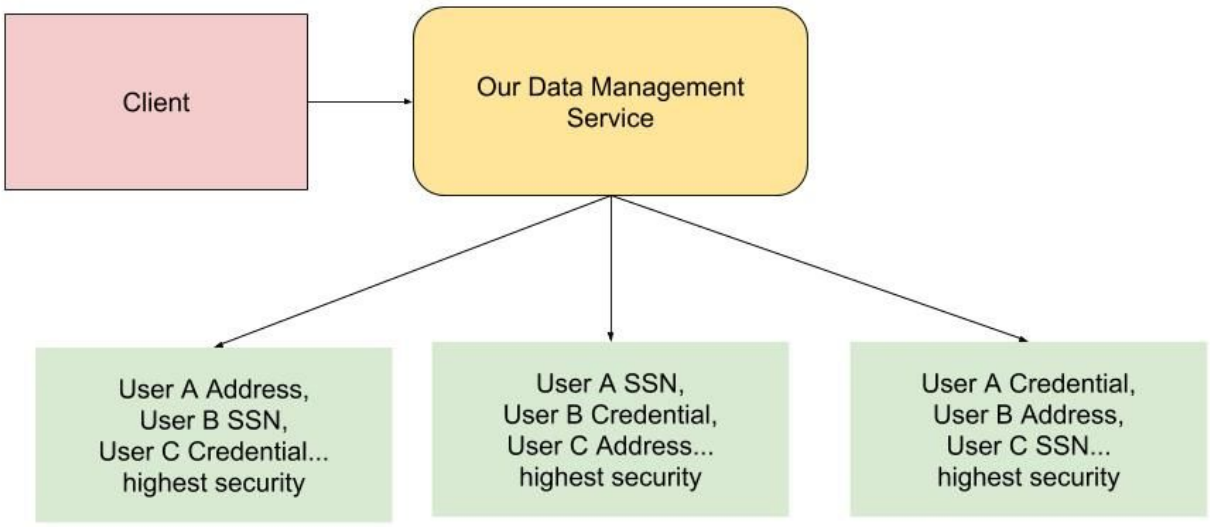
minimum security measures would include data encryption. If a user attempts to set a security level lower than the minimum security requirements of that data type, the system will warn the user or not allow them to continue without increasing the security level. We can also implement other features along with cloud storage management, such as implementing a “kill switch” by setting a expiration time in the storage.

Our system is cost effective and makes sure that clients follow security standards in the following ways:

- We update our system and storage backends following security standards, laws and regulations so that clients don’t have to. For example, the upcoming GDPR compliance taking effect on May 25th will require organizations to notify the local data protection authority and the owners of potentially compromised records of any security breaches within Europe[7]. Our service can either help generate a report, or send clients a notification about any data breach. We can also handle security flaws gracefully; if a flaw is found in one of the backend storage systems, we can quickly move the data to another storage system or patch the flaw for all of our communications.
- We allow different levels of security for different data types, which makes the storage process less expensive. We can also maximize some of the security levels for less through isolation, since such information breaches will do minimum to no harm to users if isolated pieces of data are compromised.
- We can partner with backend cloud storage services and negotiate deals due to the huge amount of data we are managing. Likewise, if we can produce an algorithm that minimizes costs to storage by selecting services wisely, such as storing low security data with the cheapest storage system that does not provide much security, it is more cost-effective than letting individual users or companies purchase storage on their own. This would be the profit for our business model.

Simple Sample Implementation II

There are also other implementations of our solution that would be more costly, but would provide more security by increasing randomness and isolation through use of the management systems algorithm. The algorithm will partition the clients data into various cloud services randomly, which would increase isolation overall so if one cloud was compromised, the data within it would be random fragments that couldn’t be compared with one another. If the same data types are stored in one cloud, patterns can more easily be found amongst the data since it is all unified in type, but our splitting algorithm would prevent this. The cloud services that would be used to store the clients’ data would also have high security levels, as we would not use the cheaper cloud services that provide lower security. This would prevent attackers from targeting cloud systems based on security level. Our solution with the algorithm could be a “premium” account to cover the costs of extra implementation measures and an increase in latency. This implementation, although more costly, provides higher security through increased isolation and randomization.



How it is a Solution to Problem 1: Failure to Follow Security Standards in Storing User Data

This proposal is a direct solution to businesses failing to follow security standards because our system provides and forces the implementation of such standards for our clients. By asking users to specify data type, we can better determine what kind of security level the data process should be in. It will be impossible to store unencrypted user PII into our backends, because when our system talks to the backend storages it will ask them to encrypt the information. Our clients will be in good coverage with successful implementation of our system and adherence to proper security standards.

How it is a Solution to Problem 2: Misuse of Data

While our method may not fully prevent the misuse of data, certain attributes of our method would deter it. A “kill switch” could be implemented by our system on top of the external storage backends, enabling a user to easily delete data that they do not want to be utilized, or if the app or service that has collected that data is no longer being used. This data would also be split up amongst various cloud services, ensuring that a threat agent would not be able to easily unify all of an individual’s forms of data. Furthermore, storing data throughout various cloud web services ensures that even if data was accessed, it would be anonymous random pieces of data that cannot be connected, rendering the accessed data meaningless. Finally, our system would also provide guidelines about proper data usage for clients.

How it is a Solution to Problem 3: Various Attacks on Data at Rest

Isolation is most helpful in preventing attacks on data at rest. While various attacks on data at rest can still occur, the consequence of any leaks in data loses significance. For

instance, in the simple sample implementation I diagram, if SSN in Amazon S3 is leaked, the leaked information only contains 9 digit numbers associated with some user ID as a key. No physical address, name, or email address is known and cannot be connected to the SSN, making these numbers meaningless. In order to connect the data together, at least one other cloud storage system would have to be comprised for the attacker to connect the stolen data to the individual user. However, that is a much more difficult task, as these different services would have different implementations and security focuses, making one attack less likely to work for every storage system. Even if physical machines are stolen from a service provider like Dropbox Cloud, the data at rest would be compromised, yet they bear no significance without being linked to other identifying data.

How it is Not a Solution to Problem 4: Various Attacks on Data in Transit

While our proposed solution does not provide protection against broken authentication caused by phishing, social engineering attacks, or other user targeted attacks, it does not exacerbate this problem either. Through phishing attacks or man-in-middle attacks, authentication can still be broken, thus making everything accessible. If that is the case, then no matter what system or mechanism we utilize, access would be granted. Therefore, handling broken authentication is not part of the focus of our system. That being said, our system would not worsen data leaks if authentication is broken, because there is still only one communication from the client to us, and our communication to backend cloud storage services will always be on the highest security level.

How it Can Be a Solution to Problem 5: Users are Kept in The Dark about Their Data

Our system will provide guidelines for clients to properly handle their user data. Naturally, a part of these guidelines will be helping users understand how their data can be accessed, but this is more like a suggestion rather than enforcement. Another possible means of protecting clients is if third parties try to access the data without first informing the client, then our management system would deny the request. This denial of access can be implemented if the client has a whitelist of third parties and their access levels, with our management system granting access based solely on this list.

Contributions to Field of Computer Security

Our solution teaches and enforces security standards on clients during the storage process. By asking the user what data type is being stored, we can provide guidelines on how to store the data and enforce minimum security on certain data types (e.g. PII). This method should help strengthen clients' understanding of practical security, as well as promote good security habits that they can apply anywhere. We placed emphasis on the importance of isolation, which solves many of our stated problems by removing the vulnerabilities of having grouped data. Isolated data has much less value than data that is unified, as isolated implementations are like stacked puzzle pieces where attackers would have to solve each step

to get useful information, rather than having to solve only one puzzle through a specific method. In this way, our system can also deter attackers, as the reward of obtaining personal information may not be worth the increased effort to collect isolated data and attempt to fit it together. Our system can raise the standard for secure data storage and if our solution is utilized by other organizations, attackers will have to increase their efforts past their current capabilities to obtain PII.

This solution would also help prevent unauthorized third parties from obtaining our clients' data without their permission. Only third parties that are whitelisted by a client can access certain data types, so a client will know exactly who is accessing what data. This will increase transparency in the usage of user data, which would be a great improvement. Our solution is straightforward and is in line with the principle of psychological acceptability as well. Data will be easy to manage and our algorithm will take care of isolating and randomizing a client's data among cloud storage systems in a secure manner, so that they do not have to worry about manually implementing security measures.

Conclusion

Collecting and storing user data has become a vital part of our technology driven world today. The amount of PII obtained continues to grow rapidly, as do the number of data breaches and leaks. Even with these data breaches, the current methods of storing user data still fail to uphold security standards and regulations, encouraging attackers to continue abusing the various vulnerabilities that exist. With many users kept in the dark about how their data is collected and stored, designing a method for securely storing this data is crucial. From this, we have designed a cloud based data management system to manage the secure storage of user data.

A fundamental part of data management is knowing what type of data is being stored and where this data lives. With our data management system, users will be able to choose what level of security they need based on the data they are storing. We provide the communication line between the cloud storage units and the owners of the data, ensuring that there is full transparency between users and companies who utilize this user data. By strictly following all security regulations, our data management system will reduce the risk of data breach by isolating user data into different cloud services. Although we cannot prevent attackers from exploiting our data storage backends, our isolation algorithm will ensure no PII is traceable to a single user if an attack were to happen. Implementing our data management system will greatly improve the ability to securely store user data, drastically reducing the impact of data breaches on society and elevate user's awareness of the security of their personal information.

References

1. Wallace, David. “The Biggest Data Breaches of the Past Decade [Infographic].” *Infographic Journal*, 20 Nov. 2017, infographicjournal.com/the-biggest-data-breaches-of-the-past-decade/.
2. Statt, Nick. “Under Armour Says 150 Million MyFitnessPal Accounts Compromised in Data Breach.” *The Verge*, The Verge, 29 Mar. 2018, www.theverge.com/2018/3/29/17177848/under-armour-myfitnesspal-data-breach-150-million-accounts-security.
3. Solon, Olivia. “'A Grand Illusion': Seven Days That Shattered Facebook's Facade.” *The Guardian*, Guardian News and Media, 24 Mar. 2018, www.theguardian.com/technology/2018/mar/24/cambridge-analytica-week-that-shattered-facebook-privacy.
4. Romano, Aja. “The Facebook Data Breach Wasn't a Hack. It Was a Wake-up Call.” *Vox*, Vox, 20 Mar. 2018, www.vox.com/2018/3/20/17138756/facebook-data-breach-cambridge-analytica-explained.
5. Larson, Selena. “Equifax Breach Impacted 2.5 Million More People than Originally Stated.”*CNNMoney*, Cable News Network, 2 Oct. 2017, money.cnn.com/2017/10/02/technology/business/equifax-million-more-impacted/index.html.
6. Wong, Julia Carrie. “Uber Concealed Massive Hack That Exposed Data of 57m Users and Drivers.” *The Guardian*, Guardian News and Media, 22 Nov. 2017, www.theguardian.com/technology/2017/nov/21/uber-data-hack-cyber-attack.
7. Nadeau, Michael. “What Is the GDPR, Its Requirements and Deadlines?” *CSO Online*, CSO, 16 Feb. 2018, www.csoonline.com/article/3202771/data-protection/general-data-protection-regulation-gdpr-requirements-deadlines-and-facts.html.
8. Pahl, Thomas B. “Stick with Security: Insights into FTC Investigations.” *Federal Trade Commission*, 26 Feb. 2018, www.ftc.gov/news-events/blogs/business-blog/2017/07/stick-security-insights-ftc-investigations.
9. “An Introduction to Data Protection.” *Https://Edri.org*, European Digital Rights, edri.org/files/paper06_datap.pdf.