Heinz College of Information Systems and Public Policy
Carnegie Mellon University

Capstone Project: Digital Identity

Final Report

Industrial Client : Tata Consultancy Services Ltd

Faculty Advisor: Prof. Michael McCarthy

Team:

Muhammad Waqas Aziz
Poojitha Prasad
Nandini Nerurkar
Vinit Shah
Shunqi Wang

For years, marketers and technology companies have crept further into the homes and habits of people, arguing all the while that there is a fair and voluntary exchange of data taking place. The argument is, the more they know about us, the more they can show us the products that we really want. Producers are trading very specific information about their lives in order to receive this kind of personalized advertising and marketing. This arises a set of crucial questions – Are people aware of how and where their personal data is being used? Is there a way in which people can directly track and control the sharing of their data? Thus, data sharing and privacy is a huge concern nowadays. We see numerous cases where private data is shared without the users' consent and the digital ecosystem is too complex for users to comprehend and manage their own digital footprints.

Following terms will be constantly used throughout the document:

Producer: Someone who produces the data and owns it.

Consumer: Someone/entity who asks for the data from producer.

Third-party: An entity that asks consumer for producer's data.

Not long ago, the world witnessed a string of revelations about how Facebook mishandled data from its users, even sharing their private messages with large corporations like Spotify and Netflix [1]. Similarly, a study done by MD Researcher John Torous [2] reveals that out of the top 26 mental health applications, 92% share data with third party applications and half of them do not disclose this information to the user.

Even if the users are made aware of where their data is shared, the digital ecosystem is too complex for them to keep track of what information is being shared and how they can change their data sharing permissions. With the current system, a user can log-in to numerous platforms using his/her Gmail, Facebook, GitHub or phone number. Thus, there is a dire need to standardize the process where the user can utilize a single identifier for logging-in across various platforms or websites.

The solution that we propose for all these issues is to introduce and implement the concept of Digital Identity. A Digital Identity consists of the user's personal information, health records, shopping transactions, bank details, etc. A Digital Identifier (DID) is a unique 16-character alphanumeric string associated with the Digital Identity. This number points to an address in the blockchain containing the public key of the producer. The DID is used to retrieve the elements of the Digital Identity, via protocols that we have proposed in this document.

This Digital Identifier would be managed through a mobile application or application(app) which stores the private key of the producer. This app would act as an interface for the producer to authorize actions, view past data access activities, and control permissions to access data. The mobile application would be installed on the producer's device which is considered as the "Trusted Device". The access to this database would be restricted to the outside world through protocols discussed in the document.

The system is designed in such a way that will give the user/producer complete autonomy and control over his data. If a company wants to access your personal records, they will have to send

you a request through your DID. Once you receive a request, it's entirely up to you to decide how much information you want to share with the company. You can also authorize the company with a 'blank cheque' authorization i.e. giving the company the authority to share data with any 3rd party companies without asking you for your permission. In addition to this, at any point, you feel the need to stop sharing your data with the company, you can simply do that using your app. Thus, our solution solves the issue of tracking, authorization and complication of data sharing.

The process can be divided into 4 distinct processes:

- Registration
- Authentication
- Permission
- Data Retrieval

Registration stage refers to the situation where the producers create and store the Digital Identifier (DID) securely. We use biometrics to ensure that a unique DID is generated per producer. We also ask the producer to set a password for the DID. This password is required in the unlikely event that the producer's Trusted Device is compromised, and the producer loses access to the DID, private key, and the app. For the application at producer's end, we have used Android Studio to develop a Java-based application. The producer's database has been implemented using Redis to store the authentication token and MongoDB to store the records like logs and permission sets.

Authentication stage helps the consumer to verify the authenticity of the producer (and vice versa) and establishes secure protocol for communication between them. By the end of this stage, the producer will have a record of an Authentication Token (auth-token) along with the consumer's DID in his/her personal data store. The consumer will have the same auth-token associated with producer's DID. This auth-token expires after a brief amount of time and authentication process is required to be restarted. The technologies used for this stage are HTML5, CSS3 and JS to create a web-based UI from which the consumer will send the authentication request to producer. We have also used blockchain from which the consumer will get the public key of the producer (used for message encryption). Finally, to setup a database at the consumer's end, we have used Redis to store the authentication token. The consumer and the producer communicate using REST APIs developed on a Java-based framework, Spring Boot.

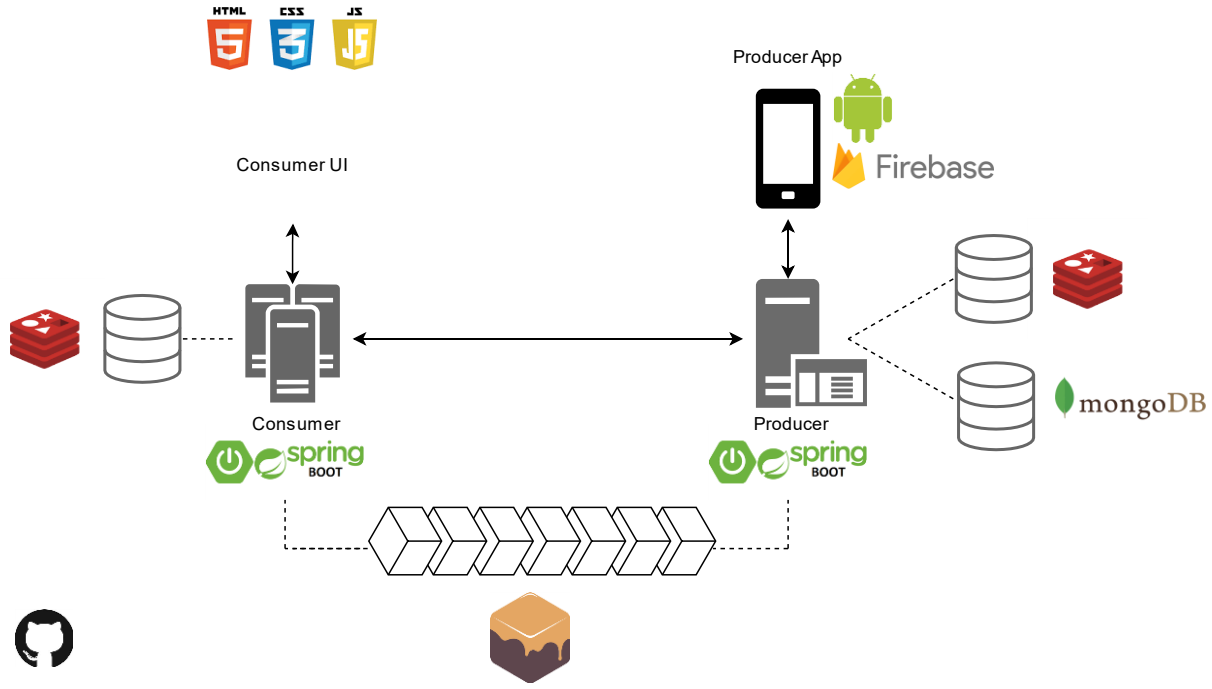
Permission stage refers to the situation where a permission set is established between producer and consumer. The permission set is a JSON document that contains information about the producer DID, consumer DID, the data categories approved by the producer and the access control, read, write or shareable, granted to the consumer. The permission-set will be stored on a MongoDB database. By using the permission set, consumers and producers can easily check what kind of data they are sharing/using. The third-party sharing check is implemented by including a 'shareable' field in the JSON permission-set. If that is set true by producer, then the consumer can share it with third party consumers without seeking consent from the producer.

Data Retrieval stage refers to the situation where data is fetched from the database based on the agreed permission-set. The consumer accesses the API route using the auth-token to get producer's

data. The producer verifies the data requested against the permission-set stored in MongoDB and sends the appropriate data to the consumer.

With the proposed architecture, we are able to solve two key issues with the current digital ecosystem - lack of autonomy and complexity. The single DID makes it easier for the user to make decisions regarding his personal data. Instead of using multiple identifiers, we strive to develop an architecture that encapsulates the entire entity of a person in a single 16-character string. Moreover, our solution provides user complete autonomy and control over his data. Through Digital Identifier (DID), a unique 16-character alphanumeric string, a user can selectively give permissions to any entity for data sharing and can revoke it at any given time.

Implementation and technologies involved:



The diagram represents the architecture of the prototype. It shows the different components, their interactions, and the technologies used for the development.

The Blockchain portion has been developed using Ganache. Ganache is a personal blockchain for Ethereum development. It has a smart contract which is used to store a map between the DID and the public key of different producers and consumers.

The producer has an Android application on their Trusted device. This uses Firebase for push notifications and it communicates with a server. The personal database of the producer is on MongoDB which has different collections like the logs, permission set, and the personal data or the Digital Identity. This data is stored in an encrypted format using the public key of the producer

so it can only be decrypted by the producer using the Trusted device where the private key is stored. In future, this can be changed to use symmetric key as the data grows.

The consumer has a simple UI to interact with the producer. The consumer has a Redis store which stores the authentication token of different producers after trust is established with them

References :

(1) New York Times, 2018, Facebook Faces Broadened Federal Investigations (https://www.nytimes.com/2018/07/02/technology/facebook-federal-investigations.html?module=inline)
(2) Rachel Becker ‘Do I trust the person who made the app, and do I understand where this data is going?’ (https://www.theverge.com/2019/4/20/18508382/apps-mental-health-smoking-cessation-data-sharing-privacy-facebook-google-advertising)