Numbers Protocol

Building trust into data and control data privacy

Understand our tech and how to support our mission

Contents

1. Introduction

1.1 Overview

From medical research to urban planning, data has always played an important part in the development of human activities. Continued development of technology has eased the creation, edition, and transmission of data. Every second, data is being collected and processed at high volume and speed. Digital files are edited or copied infinite number of times and transferred with no restrictions. As users of technology, especially the internet, we consume data constantly. With the sheer amount of data that is generated and exists at any given time, identifying the source or confirming its authenticity is nearly impossible.

There is not a stable method of ensuring the integrity of data. Long established data information embedding XMP is the closest thing to a data integrity standard, however most digital assets that appear on the internet do not have this information intact. The lack of data control poses issues with data management and ultimately, the quality of the information. This has an adverse impact on all human activities, from misinformation provided by the media, to food supply chain, public health management, and personal affairs.

False data is easily proliferated and extremely difficult to contain after its been let loose. It can negatively affect our decision making and thought process. In more personal situations, incorrect data can bypass protective regulations to violate civil rights and privacy.

Data Authenticity is a major problem that needs to be solved. The solution to this problem can be boiled down to a few factors: awareness, increasing trust in data, empower individuals with their own (personal) data.

Awareness of misinformation is the first method to combat the misuse of data. As consumers of the data, we need to understand that the information in front of us may not be trustworthy. Performing due diligence before sharing on social media goes a long way to preserving the data ecosystem. Being aware of the data dilemma is limited because performing due diligence is time consuming and difficult to do without any verification framework.

A verification framework (covered in this paper) can be created to *increase the overall trust of data.* Technology has afforded us with many solutions in the past and will be used once again to establish the authenticity of data with a secure and verifiable digital fingerprint that ensures the informational content and context of data.

The last part of data authenticity issue is *empowering individuals with their own (personal) data.* Being responsible of you data increases the overall awareness of the importance of data. Having control of personal data will prevent misuse by untrustworthy organizations hoping to profit off your information.

1.2 Background

Numbers Protocol is a startup in Taipei, Taiwan established in 2018. According to University of Gothenburg, Taiwan is one of the top nations attacked by misinformation. This together with the news of Cambridge Analytica scandal spurred the team at Numbers launched out project to tackle the issue of misinformation and data integrity.

Numbers has key partnerships with Stanford University, National Taiwan University and HTC Exodus. Our work is also being mentored by Jonathan Dotan, fellow at Stanford Center for Blockchain Research and key contributor to the Starling Project hosted by Dotan and Shoah Foundation.

1.3 Numbers Mission

Bringing trust into data and empowering individuals with their own data is our mission. The past few years have brought to light the issue of data integrity and personal data because of notable events such as the 2016 US Presidential Election and Cambridge Analytica. At Numbers, we see the integrity of data as a paramount issue and hope our work to ensure trustworthiness in data through transparency and certainty will have a lasting positive effect on society.

1.3.1 Data Trustworthiness

Even though the discussion of data trustworthiness is often dependent on the use-case, the objective is always to determine whether the data is correct and useful. Trust is paramount whenever a third party relies on data provided by a third party. When data is incorrect, users will end up putting their trust in wrong.

There are many cases of data misuse in areas such as social media, journalism, public health, etc. The following are examples of data misuse in social media:

2020 US Elections

Modified image of presidential candidate Joe Biden to make him look older.

COVID-19 Pandemic

Diverging and fake information circulating about COVID-19 causes confusion regarding its origin, how it is spread, and damages to health. One of the early examples is the footage showing an alleged source of origin, which is the Wuhan wet market. However, reporters later pointed out that video shared by the media were taken in Indonesia.

In more serious cases, data can be fabricated in public health within the framework of a clinical trial. Some cases (1990 Dr. Poison Case & COVID-19 The Lancet's retracted paper on hydroxychloroquine) were brought to court, however other instances of data fabrication remain undetected.

Numbers attempts to remedy the information trust issue by establishing data context in order to clear society unnecessary doubt when it comes to data.

1.3.2 Personal Data

An extra layer to the data discussion is personal data. Many innovations and steps forward in medical and technological solutions have come off the backs of complete, informative data. However the revealing nature of data often comes at the cost of privacy as seen in Cambridge Analytica's successful engineering of Brexit and 2016 Presidential Election.

Despite having the rights to your data in many countries, individuals are unaware what personal information is collected and how it is processed. Systems are designed for authorities and businesses which often do not consider the individual making it difficult to exercise your rights.

Number Protocols trust that a human-centric system can empower individuals by providing transparency on who accessed your information, making it easier to track it, identify parties involved and exercise your rights.

2. Numbers Principles

2.1 Goals

Numbers solution hopes to create an open, transparent and traceable data ecosystem. This means our solution must (i) improve people live while preserving privacy, (ii) be transparent and inclusive, (iii) digitally robust, and (iv) leverage existing standards.

2.2 Human-Centric

Being human-centered means designing solutions that empower individuals. When it comes to (personal) data, it is key to give clarity and control back to the individual. Data systems solely built from an organizational perspective often forgets to consider its users. This makes it difficult for consumers of technology to exercise their data rights easily and effectively.

Data immutability is a key feature of Numbers solution designed to preserve the rights of the individual. By making captured data unchangeable, data integrity is preserved. At the same time, through the use of distributed ledger technology and decentralized storage, individuals have full control over whether to share their data, modify or forget it.

2.3 Fit Existing Standards

The solution will fit within existing and developing data standards such as long standing Extensible Metadata Platform (XMP) and newly established Content Authority Initiative (CAI). Additional existing technology frameworks include Secure Socket Layer (SSL).

2.4 Open Source

Open Source Software (**OSS**) refers to a software project that has source code that is publicly available to be viewed and modified by third parties. Open source core principles are freedom, transparency, no discrimination, and community and peer-reviewing. The idea of Open Source is not a new one as much of today's internet is powered by OSS.

There are numerous technical and business benefits for OSS, some of which include reliability, security, quality, performance, developer & test base, cost, and flexibility. With these factors in consideration, adopting the OSS model is an appropriate course of action.

2.4.1 Numbers Open Source Strategy

Numbers Protocol embraces the Open Source principles towards their projects and trust an Open Source approach improves our society in different ways, such as empowering individuals with freedoms, inclusivity, or fostering innovation.

To that end, Numbers has pursued and attained OpenChain 2.0 certification in June 2020 to indicate our open source compliance and resolve to adopt OSS friendly licenses such as MIT License and GPL-3.0 or later.

2.4.2 Contribution Mechanic

One of the main bottlenecks of OSS contribution are the legal terms that come with commercial products. Traditionally, Contributor License Agreements (CLA) is the industry standard for open source contributions, however its contractual nature discourages contribution from some developers who find entering legal terms and reviewing contracts restrictive.

To streamline contributions to Numbers projects, Developer Certificate of Origin (DCO) will be the adopted contribution mechanic. Contributors to OSS under DCO will certify that they have the right to submit the code they are contributing. This mechanic is easily added to the Git workflow and developers will be credited for their contributions.

2.5 Digital Security

Given the nature of data integrity work being done at Numbers, we are setting out to be ISO27001 certified, an international standard setting requirement for establishing, implementing, maintaining and continually improving information security management systems.

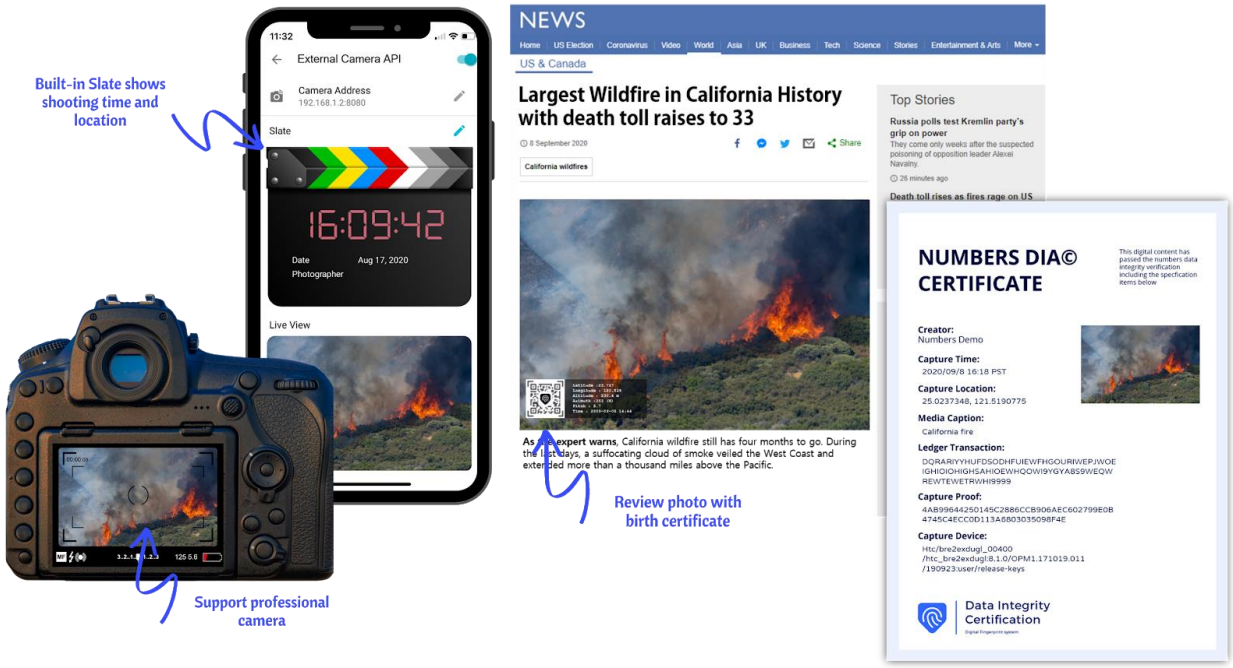
Appropriate digital security is mandatory to preserve data integrity. Being ISO27001 will confirm our compliance with high digital security standards.

3. User Cases

Data is so vast that the data integrity issue spans many areas. It would be difficult to cover all of them. Some interesting use cases of Number technology include Journalism and Digitial Signature.

3.1 Journalism

There is a lot of distrust in the news media industry due to increased awareness of fake news. Consumers of news have no means of verifying whether or not the images or reporting is reflective of what is actually happening. Numbers technology rectifies this issue through the use of digital proofs and the following workflow:

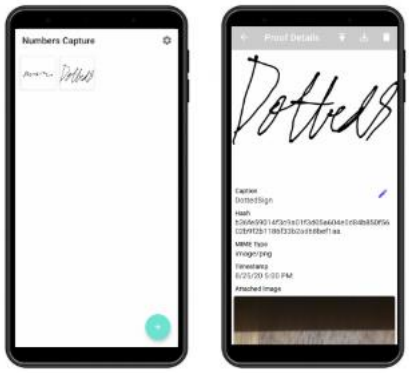


- Image / Video / Audio media (Digital Asset) is captured with a device (mobile, camera, etc).
- Digital proof is generated utilizing device sensors and device specs.
- Proof is signed with public and private keys, hashed, and uploaded to decentralized storage network.
- Upon the request of a trustworthy media source, access to the asset can be shared and downloaded from the decentralized network.

3.2 Digital Signature

Many contracts are signed over the web and currently there is no way to validate whether a digital signature is actually real. Utilizing Numbers Capture we can generate proofs with digital signatures making them verifiable with the following workflow:

Personal Seal Certificate IDA



Bring convenient and trustworthy signing experience

- Human-Centric
  GDPR compliant
- Users have full-control
- Integrity is preserved with immutable records
- Trace certificate use records

Generate Seal Certificate

1. Sign

2. Take a selfie

3. Create Seal Certificate

- Signature is created and selfie is captured by signer.
- Digital proof is generated utilized device sensors.
- Proofs are hashed and signed with public and private key.
- Upon publication, signature along with selfie and proofs are logged and certified.
- Signatures on contracts using Numbers solution can now be verified for certification.

4. Key Technologies

The data integrity solution will rely on a few key technologies. Cryptography is utilized to secure generated data. To ensure data has not been compromised, decentralized storage and distributed ledger technologies will be leveraged. In addition, we will conform to established data standards such as Extensible Metadata Platform (XMP) and Secure Socket Layer (SSL). On top of that our solution aims to be one of the first to follow the newly created 2019 Content Authority Initiative (CAI).

4.1 Decentralized Storage

Decentralized Storage is best described as a peer-to-peer (P2P) storage network. P2P networks change many aspects of internet storage. The first being instead of data being stored in one location, data is spread out over a vast network of computer systems connected with each other via the internet. This ensures accessibility of data. Data retrieval switches from location-based to content-based addressing. The reason why this is important is because data that is content-based has built in security because all content-based data is cryptographically hashed and accessed through this unique identifier instead of an IP address. The main decentralized storage player is Interplanetary File System (IPFS) as well as its incentive layer FileCoin.

4.2 Distributed Ledger Technology

DLT is a combination of pre-existing technologies, which can be summarized as a ledger system with triple-entry involving a protocol, cryptography techniques, peers, and (a certain level of) transparency. One of its benefits is the immutability of the transactional data.

4.3 Extensible Metadata Platform (XMP)

Long established metadata standard for describing the content and characters of a file upon its creation. Metadata data model is formatted as key-value pairs, structured values or list of values. Meaningful information such as title, description, searchable keywords, author, etc. are included in the embedded metadata in an easy to read format. XMP is also an ISO standard (16684-1).

4.4 Content Authority Initiative (CAI)

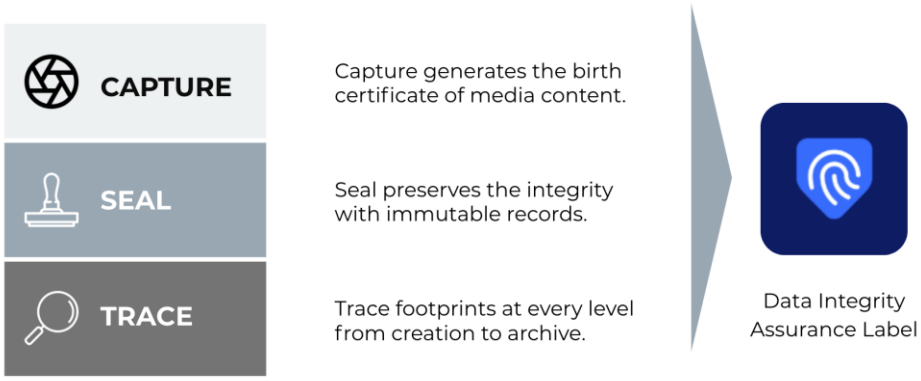
CAI is a developing industry standard for content attribution. This standard was born from the need to have digital assets to have metadata intact. Much of the assets on the internet today have incomplete or missing metadata leaving users to have to piece together the attribution information through ineffective and accurate means.

The proposed CAI workflow is the following: Creation of Digital Asset, Assertion created in JSON data structure following CAI specs, Assertions hashed, Assertions updated over the lifetime of asset.

4.5 Secure Socket Layer (SSL)

Secure Socket Layer (SSL) certificates is a protocol for web browsers and servers that allows for authentication, encryption and decryption of data. This is done with the use of public and private keys, which work together to establish a secure connection.

5. System Overview



Numbers solution is built based on the Starling Framework, the first open source framework for data integrity. Numbers hopes to address the issue of data integrity by leveraging traditional cryptography approaches along with blockchain technology in Capture, Seal, Trace approach in

order to provide data and content authenticity certifications and create traceable records at the beginning of the content lifecycle to help users claim ownership of their data.

5.1 Capture

The first stage in Numbers solution is known as Capture. At the end of the Capture stage, an Identifiable Digital Asset (IDA) will be created. Using a device (computer, mobile phone, etc), a digital asset (image, video, audio, etc.) is created at the request of the user. The device collects environmental data from sensors (accelerometer, GPS, etc.) and device data (operating system, device model, etc.) and compiles it as metadata.

5.2 Seal

IDA's are stored on a decentralized storage network like IPFS. IPFS is a decentralized storage solution that functions like a storage blockchain. The main reason for a decentralized storage solution is to ensure the availability of the IDA as traditional storage solutions are centralized and are susceptible to censorship. Peer-to-peer network and storing data with content-based addressing can ensure that the IDA remains available and secure.

5.3 Trace

Verify is the last stage of the process and is there to provide a means for certifying the integrity of the digital asset. When IDA's are published, they are logged in private ledgers which reviewers review to certify the authenticity of the IDA.

5.4 System Architecture

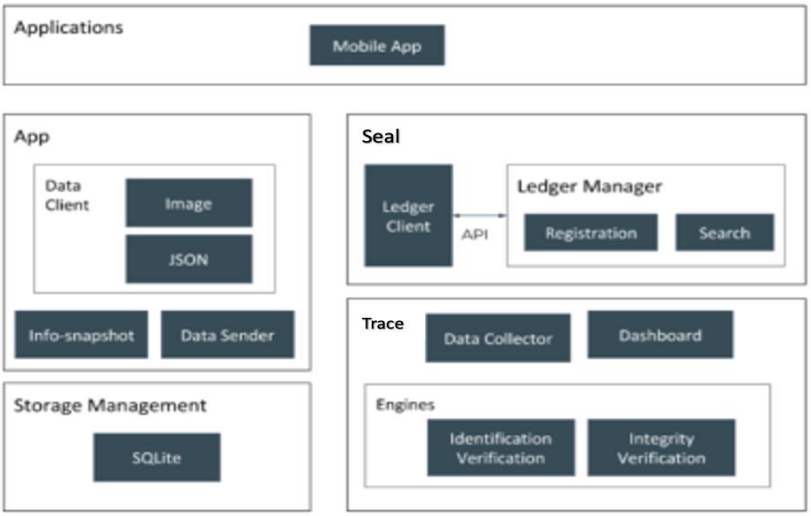


Figure 1: System Architecture

Figure 1 describes the overall Numbers solution system architecture. Data is captured by the app and metadata, following CAI specs, will be saved in unified JSON format. Users at that point will

have complete control over what to do with their data. They can store the data in local storage (ex: phone storage) or chose to make the data immutable by recording onto decentralized storage and ensuring its integrity with distributed ledger technology. Trace modules are called upon the users request.

5.5 Data Schematic Diagram

Part of Numbers Human-centric solution is to provide users with complete control over their own data. This means giving users the choice of what to do with their data and where to store it. In some scenarios, users may just want to keep their data in local storage. Figure 3 summaries the local storage data flow:

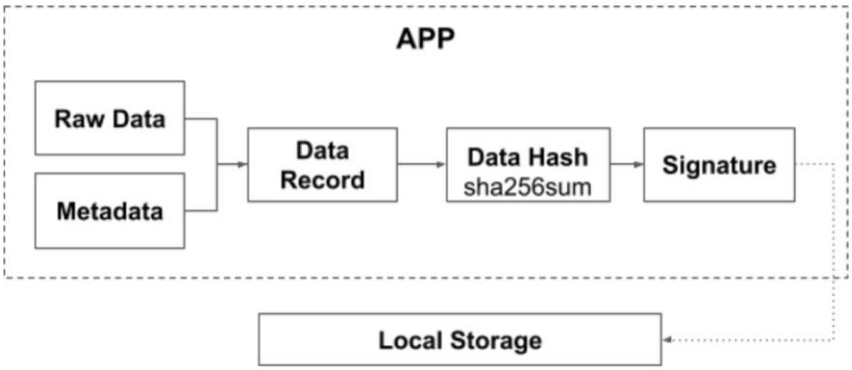


Figure 2: Local Storage Data Schematic

In situations where maintaining data authenticity is important, users have the option to have their data go through the decentralized ledger work flow summarized by Figure 3.

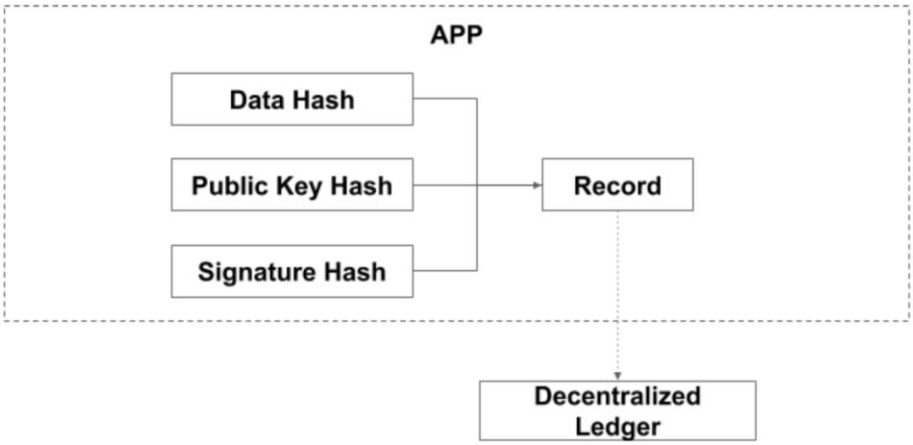


Figure 3: Decentralized Ledger Data Schematic

Data that go through Decentralized Ledger route are cryptographically hashed and signed with a public and private key. The package is wrapped up as a record and stored in decentralized storage

and logged on decentralized ledger. In other scenarios, users may just want to keep their data in local storage. Figure 3 summarizes the local storage data flow.

6. Technical Discussion

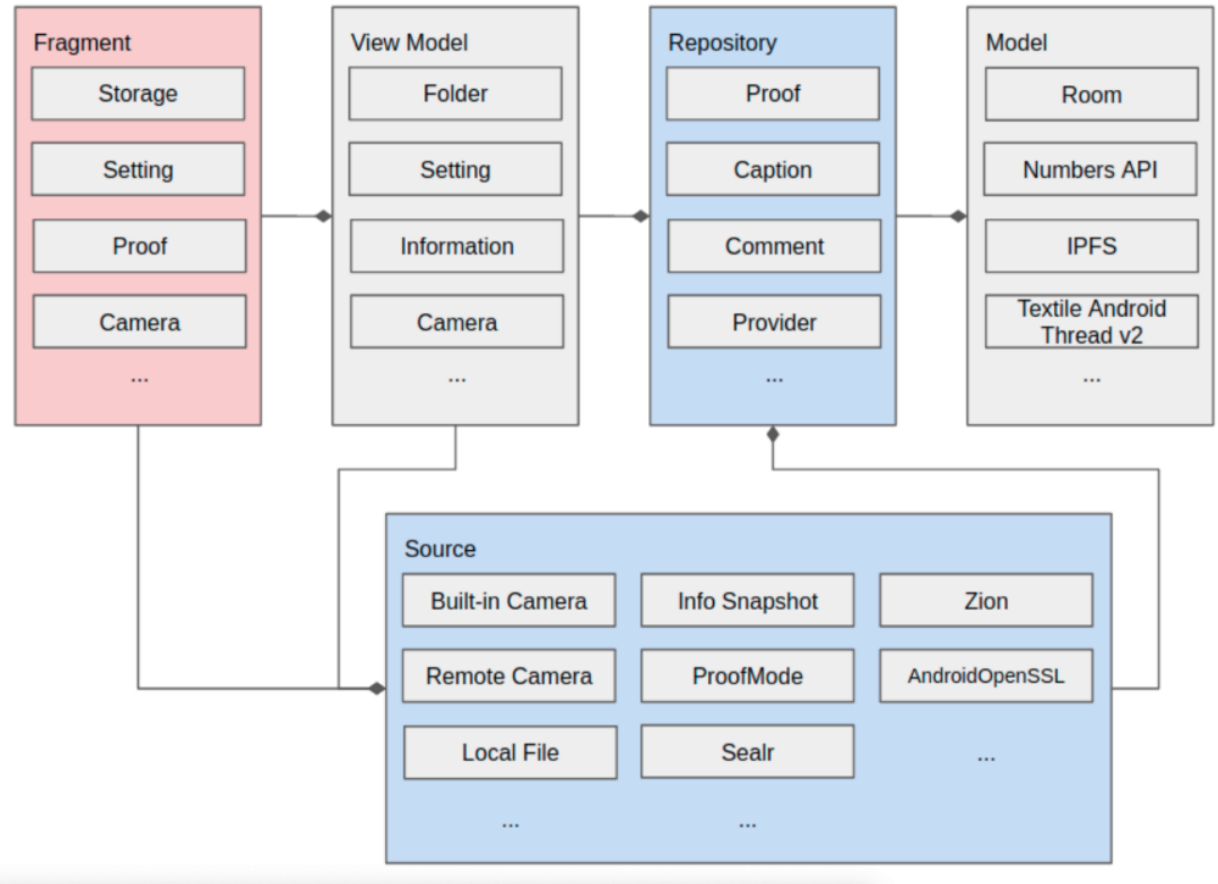
6.1 Technology Stack

Technology stack is subject to change and alteration. At the time of writing Numbers Capture Android application as the following technology stack:

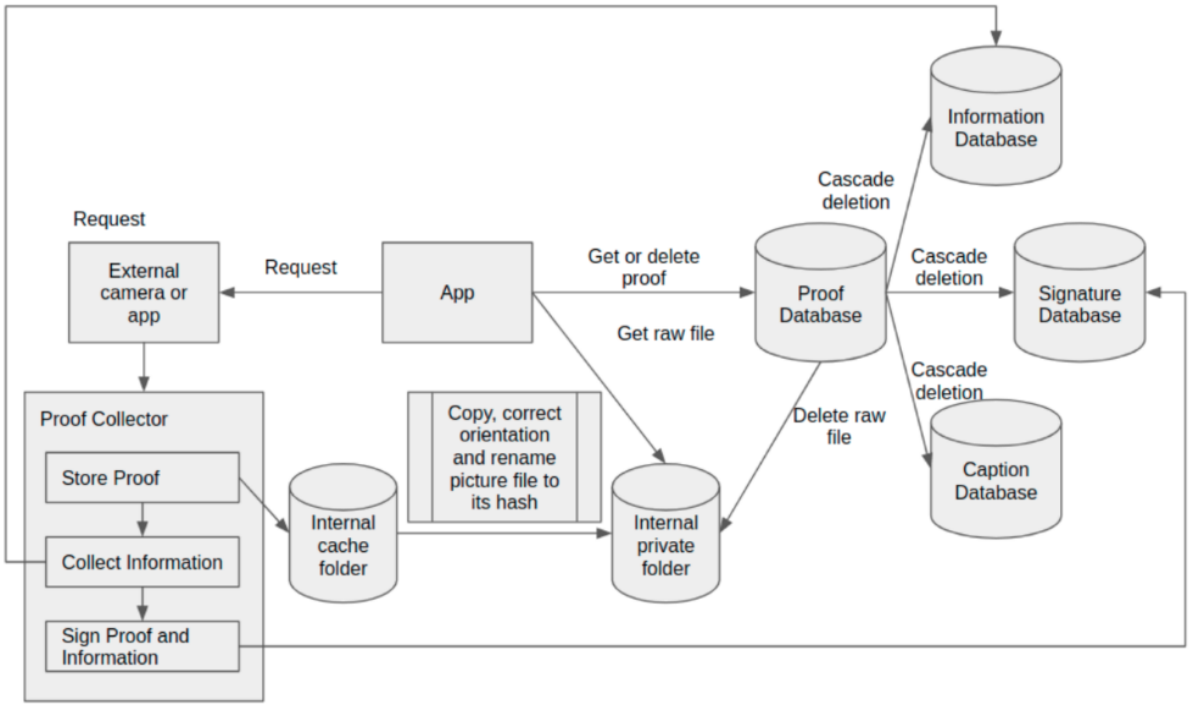
- Kotlin (Coroutine, Flow)
- AndroidX
- Model-View-ViewModel Architecture
- Android Architecture Component
- Material Component
- Koin
- Retrofit
- Coil

6.2 Architecture

Numbers Capture Android application has the following MVVM (Model-View-ViewModel) Architecture:


Fragment
Storage
Setting
Proof
Camera
...
View Model
Folder
Setting
Information
Camera
...
Repository
Proof
Caption
Comment
Provider
...
Model
Room
Numbers API
IPFS
Textile Android Thread v2
...
Source
Built-in Camera
Info Snapshot
Zion
Remote Camera
ProofMode
AndroidOpenSSL
Local File
Sealr
...
...
...

6.3 Data Flow



The figure above captures the following data flows for Numbers Capture:

- Create New Proof
- Get Proof
- Delete Proof

Proof Collector represents the data flow for generating metadata (proof, environmental data) and signing of IDA. The following summarizes the steps:

- Store proof raw file in internal directory
- Store proof hash in proof repository
- Collect Information
- Sign proof and its collected information

6.4 Serialization Schema

When a Digital Asset is captured, a proof is generated containing cryptographic hash code, file type and timestamp. Additional environmental data is collected and together with the proof make up the Digital Asset's metadata. Numbers Capture metadata schema is the following:

```
{
    proof: {
        hash: String,
        mimeType: String,
        timestamp: Long
    },
    information: [{
        proofHash: String,
        provider: String,
        name: String,
        value: String
    },
    ...
  ]
}
```

An example of this is:

```
{
   "proof":{
      "hash":"1837bc2c546d46c705204cf9f857b90b1dbffd2a7988451670119945ba39a10b",
      "mimeType":"image/jpeg",
      "timestamp":123456789
   },
   "information":[
      {
         "proofHash":"1837bc2c546d46c705204cf9f857b90b1dbffd2a7988451670119945ba39a10b",
         "provider":"ProofMode",
         "name":"Current Location",
         "value":"121.0, 23.0"
      },
      ...
   ]
}
```

In order to read and share, the metadata is signed with public and private keys. This is captured with the following schema:

```
[
    {
        proofHash: String,
        provider: String,
        signature: String,
        publicKey: String
    },
    ...
]
```

An example of this is:

```
[
  {
    "proofHash": "845ace0144620a18abf1d73c1dceaa51ea78cd5d791dbbbd2368d75260431bd9",
    "provider": "AndroidOpenSSL",
    "signature": "3046022100b96babf7fb1a374792ce47cebdf0b5a40166352a4e8aed2d8e84a04699898c72022100ec0646e318a0
    "publicKey": "3059301306072a8648ce3d020106082a8648ce3d030107034200043e4ba565aa9158b9aeafc1bb4a970bfc7fcdcc
  },
  {
    "proofHash": "845ace0144620a18abf1d73c1dceaa51ea78cd5d791dbbbd2368d75260431bd9",
    "provider": "Zion",
    "signature": "3045022100dbfe89fe13a4758f2124fc35d440d6ca8a6b3c3d72429a3a70b3a8146695c0db02204d8d387bba770d
    "publicKey": "Session:\n3059301306072a8648ce3d020106082a8648ce3d030107034200004b4b85c26384dda113f029cfb3b71
  },
  ...
]
```


7. Contributions / Next Steps

Numbers Protocol welcomes any individual interested in our project and passionate about bringing trust into data to contribute to our project. Contributions can range from code (new features, bug fixes, unit tests, etc.) to non-code contributions.

We follow GitFlow workflow and ask code contributors to write code in created feature-branches. Completed code will be submitted for review via Pull Request where it will undergo code review and discussion. Upon code review completion, contributions will be merged into the origin branch and credit will be given.

If you are not a developer, there are many non-code contributions that are welcome and encouraged. A successful project requires more than just code. Some non-code contributions include maintaining documentation as well as performing Quality Assurance (QA).

Numbers Protocol thanks you for your interest and looks forward to your contributions.