Project Proposal.md

A Blockchain based file verification system.

The Problem

Document verification has always been a major problem and road block for many key organisations and institutes which includes but not limited to banks, healthcare, education, law making and enforcement. The authenticity of a digital file is always in question as they can be easily replicated and modified.

Currently, the industry uses file hashes (like MD5, SHA1,2,3) to determine if a file has been tampered with or not. But there is nothing guaranteeing that the original hash of a file published by the author of a file has not been tampered with itself i.e. the source itself is corrupted. Such attacks are easy to carry out as the current infrastructure supporting the delivery of software services over the internet is highly centralised with major corporations storing their data in large data centres across the the globe, which makes them a single point of failure and an obvious target of an attack both physically and over the cyberspace.

The Solution

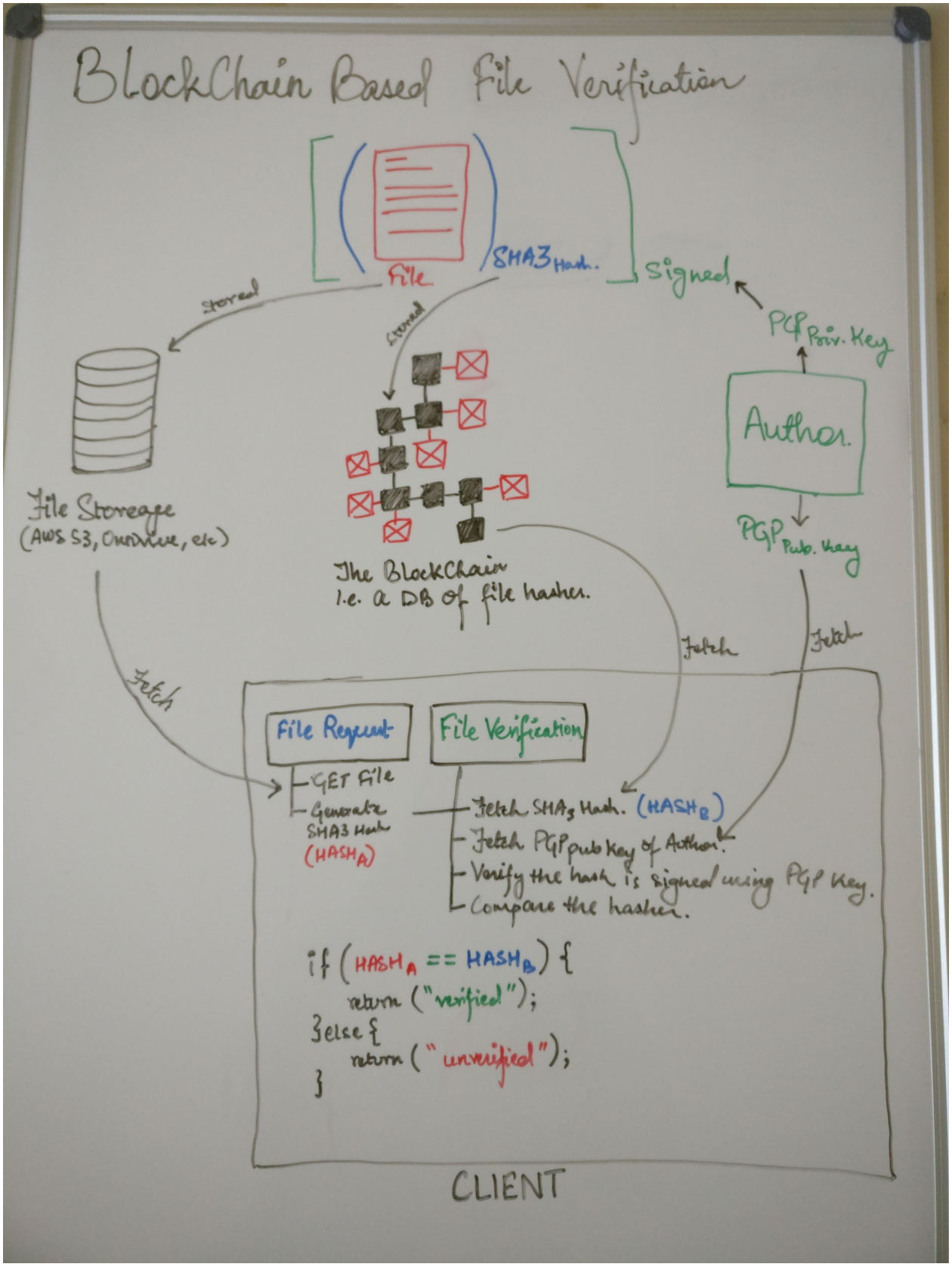
I propose that file hash signed by the author's RSA key be stored in a decentralised database running on a system of consensus.

Enter the Blockchain

Crypto-currencies are all the buzz these days, everyone from tech gurus to main stream media is talking about it. But as a computer scientist the most interesting part about crypto-currencies is the tech-stack on which it is build i.e. the blockchain. There are many definitions of what a blockchain is, but to put it very simply, it is decentralised database system which distributes the whole database to every single node participating in the network and maintains its integrity via a system of consensus between the nodes. ***Anders Brownworth*** *has a great example with working demo here: https://anders.com/blockchain/*

I intend to use the blockchain as a decentralised database of the above mentioned file hashes which can act as the front-end for multiple client facing solutions.

Rough Sketch of the system



The File

As shown in the figure above, the system generates the SHA3 hash of the file, requests the author of the file to sign it using their private RSA key. This signed SHA3 hash is then stored on the blockchain.

The Author

An author in the context of this system is anyone who is owner/creator of a file. When an author register's for an account, the system generates a 4096-bit RSA key pair. These keys are then used to sign and verify the file hashes in the future.

File Storage

The files can be stored in any user preferred locations, i.e. local file system or cloud based storage solutions.

The Blockchain

This maintains a decentralised and tamper resistant database of all the file hashes for every file per user. This not only helps us to determine the authenticity of the file but also the ownership, date, time, and location of creation, which has the potential of providing irrefutable evidence for many digital file related disputes, including but not limited to accreditation, identity, patent, etc.

The Client

This is where we actually give access to the database to the client and make it accessible to everyone with ease. Lets take the example of University of Birmingham. The university exists as an author on the system, the user wants to find their degree certificate from the university. Their reference is their student ID, once they enter the ID:

- the system fetches the file (based on the provided storage location)
- generates its SHA3 hash
- then it requests the hash of the same file from the blockchain database
- it then requests the public RSA key of the author
- uses the public RSA key of the author to verify that the hash received from the blockchain is signed by the author
    - if yes, then compare the two hashes to verify the file
        - if they are equal, then the file is verified
        - else, the file is unverified and potentially tampered with
    - else, the file hash on the blockchain is not right and signed by the author, in this case a notification is sent to the author, to take further action.

Why not store the files directly on the Blockchain?

Initially I thought of storing the files directly on the blockchain, though it sounds doable for simple text documents, it becomes increasingly infeasible for audio, video and other complex files. Since every participating node in the blockchain receives the complete copy of the database it becomes highly expensive for small to medium size nodes on the network to store this ever increasing database of actual files and also the security of these files becomes a major risk, which makes file encryption and key storage another issue. Though these issues can be somewhat addressed when using a private blockchain over a public one, they still remain a major concern when considering the fact that blockchains currently doesn't scale. ***Preethi Kasireddy*** *has covered this in great detail in her article: Blockchains don't scale. Not today, at least. But there's hope.*

Benefits of the system

Since, we are not storing the files in any particular location, it makes the distribution and storage of the said file much simpler for the author. This file can essentially be stored anywhere, from local to cloud based storage solutions and distributed using any medium, it can always be verified using the verification system.

The system outlined so far is in its initial stage and I'm still refining it and nailing down some of the specifics.