A vision for integrated Data Authentication and Data Integrity[1]

aka Data Provenance and Traceability

Gary Mawdsley CEO Lockular Limited

May 2019[2]

[1] A precursor to the detailed 2021-2022 paper on the Provenance Tracking Filesystem

[2] Updated October 2023 to refresh the language and to add Why does Data Authentication and Data Integrity matter?

This white paper delineates an approach to tackling the challenge of Data Authentication, with a specific focus on critical infrastructure. It's important to note that this approach is applicable across all sectors and extends into the realm of personal computing.

.

Standard Definitions

Data Authentication is the widely accepted concept for confirming, without a doubt, the origin of data. It entails validating the source of data to guarantee that it remains unaltered and originates from a trustworthy entity.

The term used to establish the soundness of data from its source and to ensure its completeness without any omissions is referred to as "data integrity". Data integrity practices concentrate on upholding the precision and uniformity of data throughout its entire lifecycle, thwarting unauthorised modifications or deletions.

Leveraging Blockchain

Techniques like checksums, hashing, and data validation are commonly employed for the purpose of confirming data integrity. Achieving data integrity often involves the use of methods such as digital signatures, certificates, and cryptographic approaches. However, these methods, when used in isolation, do not provide full protection against tampering.

To establish a tamper-resistant and auditable record of data, one effective approach is to record data hashes on a blockchain. Blockchain technology is renowned for its immutability due to its decentralised and distributed ledger, making it extremely challenging to alter records once they are added. The viability of Bitcoin, for instance, heavily relies on this characteristic.

A note on blockchain performance: We employ the Polkadot blockchain platform, which has been purpose-built to tackle various performance and scalability issues encountered by conventional blockchain networks.

The blockchain serves as the underlying framework for an economic model where the cost of Data Authentication is inherently embedded in the governance structure of the blockchain, thus eliminating the need for licensing fees. Moreover, the amount of data written to the blockchain is minimal, as only summary hashes of

the audit are recorded. This results in a cost-effective and scalable solution that also addresses privacy concerns.

Furthermore, the utilisation of Para-Chains and Oracles enhances scalability. Para-Chains are parallel blockchains that can seamlessly interoperate through an exchange hub known as a Relay-Chain. Oracles, on the other hand, serve as off-chain data stores and can be based on any persistent database technology. What's crucial is that the data integrity of information within the Oracles is safeguarded through correlation with blockchain hashes derived from the data.

It achieves enhanced performance and scalability through its distinct architectural approach and a range of innovative features, including multi-chain functionality, a shared security model, seamless interoperability, upgradability, the utilisation of Nominated Proof of Stake, horizontal scalability, efficient cross-chain communication, and the effective integration of correlated Oracles.

The problem

Numerous forms of auditing and tracking are in existence, with widespread usage within the legal profession and the IT industry, particularly in the realm of source control management. Nevertheless, it's worth noting that, at present, there is no foolproof method to prevent a situation where a malicious actor, in a position of authority, can manipulate historical data while concealing their actions.

Moreover, a universally accepted approach is lacking, and there is a growing recognition of the necessity to safeguard and account for a broad spectrum of data, ranging from critical infrastructure information to personal data. Notably, Apple has taken strides towards enhancing personal privacy by integrating alerts and providing daily reports on vulnerabilities and exposure. However, this approach remains proprietary and somewhat specialised.

Our perspective is that the consensus architecture of blockchain technology offers a trustless framework for governing data usage and ensuring data integrity.

Lockular solution architecture

The current solution is built upon two fundamental concepts: the capability to manipulate (break down and reconstruct) data for long-term storage and a set of guidelines (a protocol) for the disassembly and reassembly of data.

These guidelines consist of multiple stages and incorporate additional security measures by utilising ephemeral (temporary) keys throughout the data transformation and exchange process.

Both of these concepts serve as the foundational technologies within Lockular.

The disassembly and reassembly technology functions as a pipeline, offering features such as pseudonymisation, standard encryption, and cryptographic slicing similar to Shamir secret sharing.

[The configurability of the pipeline allows substitution of opera-

tors. For example, we are presently considering quantum safety by switching out AES-256 as the default encryption engine and replacing with a lattice based engine. The lattice approach is combinatorial in nature and we believe in the very general sense this has similar difficulties to NP-complete problems.

Material point here is NP-complete is believed to be quantum safe.

The protocol technology manages the sequencing and protection of data between its creation, usage and storage.

While each of these technologies can address various use cases on its own or in combination, our initial implementation integrates them into a filesystem, providing a wide range of applicability. In the virtual landscape of cloud computing, particularly within Kubernetes, this approach allows for many potential adaptations to specific workloads through custom Kubernetes storage classes and, consequently, custom filesystems. We believe that integrating this protocol into the core of cloud storage will enable numerous applications and use cases to seamlessly adopt robust data integrity capabilities.

These are the touch points inside clouds that offer the prospect of a ubiquitious provenance tracking filesystem

As data is written to, and read from this filesystem, access audit records are generated, hashed, and then these hashes are recorded along with manifest data onto a blockchain. These hashes and manifests comprehensively represent the audit data. The audit is fundamentally stored in conventional databases with corresponding hashes on the blockchain. The hashes do not reveal the semantic content of the data and therefore allows for the use of a public blockchain to achieve the highest level of consensus and immutability.

To retrospectively examine the audit data for reporting purposes, information is cross-referenced between the blockchain data and the off-chain audit records contained within the Oracles, resulting in an assessment.

The outcome is an immutable log that attests to the data's integrity, disclosing precise actions and the involvement of individuals or workload processes. Our present solution is employed satisfying the data integrity requirements of critical infrastructure.

To play devils advocate: Why does Data Authentication and Data Integrity matter?

There are a number of crucial reasons:

- Trustworthiness: Data integrity ensures that data is accurate and reliable. When data integrity is compromised, it can erode trust in the data and the systems that produce and manage it. Trustworthy data is essential for decision-making, analysis, and maintaining credibility.
- Compliance and Regulation: Many industries and organisations

are subject to regulations and compliance requirements that mandate data integrity. Failing to maintain data integrity can result in legal and financial consequences.

- Data Quality: Data with high integrity is of better quality. Accurate and complete data is essential for generating meaningful insights, conducting analytics, and making informed decisions.
- Preventing Errors: Data integrity measures help detect and prevent errors, whether they are accidental or intentional. This reduces the risk of making decisions based on incorrect information.
- Security: Data integrity is closely linked to data security. Ensuring that data remains unchanged and uncorrupted is a fundamental aspect of protecting sensitive information from unauthorised access and tampering.
- Reputation and Customer Trust: Businesses and organisations that maintain strong data integrity earn the trust of their customers and stakeholders. Data breaches or data quality issues can damage an organisation's reputation.
- Operational Efficiency: Data integrity contributes to operational efficiency by reducing the need for data correction and cleanup. Clean and reliable data streamlines processes and workflows.
- Historical Accuracy: Maintaining data integrity ensures that historical data remains accurate and can be relied upon for historical analysis, audits, and compliance purposes.

In summary, data integrity is vital for ensuring that data can be trusted, used effectively, and is compliant with regulations. It underpins many aspects of data management, data analytics, and decision-making in organisations and industries across the board.

When the worst happens

No system is impenetrable and so data will get hacked and stolen. There is significant merit in having information about what data was seen or stolen in the event of a data breach or hack. This information is valuable for several reasons:

- Mitigation: Understanding what specific data was accessed or stolen allows an organisation to take targeted actions to mitigate the impact. For example, they can focus on securing the compromised data, notifying affected parties, and implementing additional security measures for the exposed data.

- Notification and Compliance: Many data protection regulations, such as GDPR and CCPA, require organisations to notify affected individuals and authorities in the event of a data breach. Knowing what data was exposed is essential for fulfilling these notification requirements accurately and promptly.
- Risk Assessment: It helps in assessing the potential risks and consequences associated with the breach. Different types of data have varying levels of sensitivity, and knowing what was exposed allows organisations to prioritise their response efforts.
- Forensic Analysis: Detailed information about what data was accessed can be valuable for forensic analysis. It can help in identifying the methods used by the attackers, their motives, and the extent of the breach.
- Legal and Regulatory Compliance: Having a clear record of what data was compromised can assist in legal proceedings, investigations, and regulatory compliance. It demonstrates transparency and a commitment to addressing the breach responsibly.
- Security Improvements: The knowledge of what data was targeted can inform future security measures and strategies. Organisations can use this information to enhance their cybersecurity defences and protect against similar attacks in the future.

In summary, understanding the specifics of a data breach, including what data was accessed or stolen, is crucial for both immediate incident response and long-term cybersecurity planning. It enables organisations to take appropriate actions to minimise the impact, comply with legal requirements, and strengthen their overall security posture.

Having a robust approach to data integrity, monitoring, and auditing in place could have potentially helped in the case of Edward Snowden's leak of classified information from the U.S. National Security Agency (NSA). Snowden's actions resulted in the exposure of a vast amount of sensitive government information. Here's how improved data integrity practices might have made a difference:

- Identification of Data Access: With proper data monitoring and auditing in place, the NSA could have detected unauthorised access to classified information in near real-time. This could have raised red flags and triggered an investigation into who accessed the data and why.
- Data Classification: Data classification systems help organisations categorise data based on its sensitivity. If the exposed data had

been appropriately classified, it would have been easier to determine the severity of the breach and which specific documents or information were compromised.

- Access Controls: Strong access controls, including role-based access and the principle of least privilege, can limit the number of individuals who have access to highly sensitive data. This reduces the risk of a single insider gaining access to and leaking extensive amounts of data.
- Immutable Audit Trails: Detailed audit trails can provide a record of who accessed what data and when. This information can be invaluable in investigations and forensics to understand the extent of the breach and track the actions of insiders.
- Data Loss Prevention (DLP): DLP solutions can help prevent unauthorised data transfer or leakage by monitoring and blocking sensitive data transfers.

While strong data integrity practices can enhance security and assist in detecting and responding to insider threats, it's important to note that no system is entirely foolproof. Insiders with privileged access and knowledge can sometimes find ways to circumvent security measures.

Nonetheless, a well-implemented data integrity and security program can significantly reduce the risk of such incidents and facilitate a more effective response when they do occur.

References