Concept Paper: Network Node Model versus Data Lake Model

Alex Renz
Eric V. Trappen

Introduction

In the realm of energy efficiency and sustainability, securing data from devices, systems, and operational analytics is crucial for ensuring trust, auditability, and compliance. Furthermore, the use of AI to normalize, analyze (i.e. anomaly detection) or to make predictions warrants data trust models to establish an audit trail of how data gets transformed.

As we seek to align our technology strategy with customer project needs, where we might initially engage around data science, we explore alternative approaches to trusted data management.

This paper discusses two distinct approaches to data confidence: a **Node Network Model** and a **Data Lake Model**, evaluating their respective advantages, challenges, and implications for managing energy data in a decentralized, trustworthy manner.

Approach 1: Node Network Model

Description:

Under this model, each participant (like electric utilities, manufacturers) operates a node within a decentralized network. This node not only stores all internal data from various sources (equipment, systems of record) but also performs data normalization, aggregation, and analysis.

- **Data Storage and Analysis:** All data handling occurs within the node, ensuring that data does not leave the secure environment of the participant until necessary.
- **Data Verification:** Hashes of the data are distributed across other nodes in the network. This hash distribution creates a verifiable chain of custody, where data provenance can be traced back to its source with corresponding trust levels.
- **Controlled Data Sharing:** Select data can be shared with 3rd parties or by providing a user on the system with associated access rights. This would include auditors who can be granted access for the purpose of auditing and certification. Data can also be shared with 3rd parties in a controlled manner, such as by providing an access key to a data stream or role- and user-based access control. Access could also be controlled in a decentralized fashion using verifiable credentials linked to blockchain-based identities for people, organizations and devices.
- **Encryption:** Where actual data is shared on the blockchain, data must be encrypted. Where data is shared outside of the blockchain

Pros:

- **High Security:** Data remains within controlled environments, reducing risks of external breaches.
- **Decentralized Control:** Each participant manages their data integrity, fostering trust and reducing reliance on third-party data lakes.
- **Immediate Auditability:** Auditors can verify data directly from the node, with full access to data provenance and trust levels.

Cons:
- **Scalability Issues:** As the network grows, managing and synchronizing numerous nodes could become complex.
- **Resource Intensive:** Each node must have robust computational resources for data processing and security.
- **Interoperability:** Different participants might use different systems, complicating data exchange and standardization.

Approach 2: Data Lake Model

Description:
In this model, data from various sources is centralized into a data lake, external to the blockchain node network.

- **Data Storage:** Raw data is stored in the data lake, which operates independently of blockchain software.
- **Data Security and Verification:** Data integrity is ensured by hashing all data into the blockchain network. Metadata, including links to data sources, device IDs, and digital twins, is stored in the blockchain, providing an immutable record of each data point's origin and trust level.
- **Access Control:** Access to the data lake is managed through sophisticated access controls, allowing only authorized audits or verifications.

Pros:
- **Scalability:** A centralized data lake can handle large volumes of data more efficiently than a decentralized node network.
- **Flexibility:** It's easier to integrate new data sources or upgrade data handling technologies without affecting the blockchain network.
- **Cost Efficiency:** Centralized storage might reduce redundancy in data storage across multiple nodes.

Cons:
- **Security Concerns:** Centralizing data could make it a single point of failure or attack if not secured adequately.
- **Trust Issues:** The separation of data storage from the blockchain might raise questions about data integrity unless robust linking and hashing mechanisms are in place.

- **Audit Complexity:** While verification is possible, auditors might find it challenging to access or correlate data without direct node interaction.

Comparative Discussion:

- **Security vs. Accessibility:** The node network offers superior security at the cost of complexity in management, while the data lake model prioritizes accessibility and scalability but may compromise on security perception.
- **Data Provenance:** Both models can secure data provenance, but in the node network, this information is inherently more integrated with the data handling process.
- **Compliance and Audit:** Node networks might provide a more straightforward audit trail due to the integrated nature of data handling and storage. The data lake model requires careful management of metadata to ensure auditability.
- **Operational Efficiency:** Data lakes could be more operationally efficient for large-scale data analytics, but node networks ensure that data manipulation and analysis are under the direct control of the data owner.

Other Considerations:

Another consideration beyond the data confidence as such is the implementation of a repeatable methodology that represents a logical and scalable approach to realizing energy savings and other sustainability measures. GoldStandard and Verra require project developers to establish and follow a methodology as part of their certification schemes. Our goal should be to create scalable methodologies that combine sustainability measures with data trust and data science to achieve high degrees of automation and scale to increase impact and monetization.

Conclusion:

The choice between these models depends largely on the balance needed between security, control, scalability, and auditability. The Node Network Model might appeal to entities prioritizing data control and security, while the Data Lake Model could be favored where scalability and centralized management of data are paramount. A hybrid approach, where sensitive data remains in nodes while less critical data resides in a lake, might also be considered to leverage the strengths of both systems. This decision should be guided by the specific requirements of data security, regulatory compliance, and operational efficiency in the context of energy efficiency and sustainability initiatives.