International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

AI-Powered Blockchain-Based Secure Data Marketplace

Hanny Kuhikar[1], Prince Raidas[2], Bhumika Tambuskar[3], Dr. Pooja Jaiswal[4] and Dr. Sudhir N. Shelke[5]

Department of Computer Science and Engineering (Data Science), Guru Nanak Institute of Technology, Nagpur, Maharashtra, India

[4] Head of the Department (HOD), Department of Computer Science & Engineering (Data Science), Guru Nanak Institute of Technology, Nagpur, India

[5] Principal, Guru Nanak Institute of Technology, Nagpur, India

Abstract—The widespread use of data-hungry artificial intelligence systems has turned specialised empirical datasets into valuable economic assets. Yet today's data exchanges depend on centralized intermediaries, which create single points of failure, obscure chains of custody and settle payments non-atomically. Evidence also suggests that as many as 62% of unvetted public records contain structural flaws, repeated observations or faulty labels, all of which weaken the models trained on them. Distributed ledgers offer tamper-proof audit trails and trustless escrow through smart contracts, but a plain blockchain only ensures that a transaction is carried out; it cannot judge whether the data behind it is meaningful or original. This review surveys recent work on hybrid marketplaces that pair blockchain settlement with machine-learning-based quality control. We describe the trading model at a conceptual level, propose an architectural taxonomy covering off-chain storage and smart contract state machines, and assess algorithmic components for statistical completeness, locality-sensitive hashing and dynamic valuation. We then review the key literature, compare the approaches in a feature matrix and discuss the main open problems: adversarial data poisoning, algorithmic front-running and compliance with privacy regulation.

Keywords—Blockchain, Data Marketplace, Smart Contracts, Dataset Quality Assessment, Deduplication, Dynamic Pricing, IPFS, Decentralized Trust.

I. INTRODUCTION

Today's computing systems depend heavily on large volumes of well-curated data. Whether the application is a deep neural network, a conversational foundation model, a predictive controller for a microgrid or a self-driving vehicle, how accurately an AI system predicts is limited by how good, and how specific to its domain, the training data is [1], [6]. Since companies increasingly find that isolated data silos restrict what they can analyze, multi-party data marketplaces have become key infrastructure for discovering and trading data across organizations [2], [4].

Even so, most existing data-exchange platforms still rely on a centralized, custodial broker. In such systems one commercial operator controls the hosting of datasets, the verification of identities, the indexing of the catalogue, the collection of fees and the clearing of payments. Concentrating this much control in a single party exposes the system to serious operational weaknesses [3], [13]:

Centralized Risk and Attack Surfaces: A central database is an attractive target for distributed denial-of-service attacks and eavesdropping, and a failure of that one system can take the whole platform offline [9], [15].

Inflow of Low-Quality and Duplicate Data: Central hubs usually perform little verification before a sale. Studies report that over 62% of publicly traded datasets have notable defects such as missing values, imbalanced classes, schema violations or entirely duplicated synthetic records, which undermine the machine learning pipelines that consume them [4].

Untraceable Provenance: Buyers have no cryptographic way of independently checking where a corpus came from, how it was collected or whether it was tampered with in transit. Creators, in turn, lose control over onward redistribution once their assets are uploaded to a custodial repository [5], [14].

Rent-Seeking and Unbalanced Settlement: Brokers commonly charge commissions of 15% to 40%, and payouts can be delayed for long periods because there is no atomic exchange guarantee [1], [8].

Distributed ledger technology offers a strong basis for addressing these problems. Platforms such as the Ethereum Virtual Machine and Layer-2 rollups provide transparent, tamper-proof transaction records and can run payments through non-custodial smart contracts [5], [8]. Because storing data on-chain is prohibitively expensive in both cost and computation, hybrid designs pair the ledger with peer-to-peer, content-addressed storage, most commonly the InterPlanetary File System (IPFS) [1], [2].

To close this gap, recent studies have looked at combining Artificial Intelligence with Distributed Ledger Technology. In this combined approach, smart contracts handle trustless payment and permissionless access control, while AI pipelines act as automatic gatekeepers: before a dataset is registered on-chain, they assess its quality, flag duplicates and suggest a price [6], [11].

This paper offers a systematic review of the current state of the art in this area. We cover fifteen key contributions from 2020 to 2026, formalize the architectural models, define a four-layer taxonomy and assess the main open challenges [1]–[15].

II. PROBLEM FORMULATION & TRADING MODEL

To study how a decentralized data exchange works in practice, we model the trading environment as a coordination system involving several participants, governed by deterministic smart contracts and automated machine-based curation.

A. Participant Definitions and Market Dynamics

The marketplace consists of a self-organizing network of sellers and a distributed population of buyers. Sellers offer candidate datasets made up of many observation rows described by a set of attributes. Every candidate carries a hidden utility value reflecting how much it could improve downstream predictive models. Without any screening in a decentralized setting, this information asymmetry lets opportunistic sellers pass off low-value or fabricated data as premium, which can cause the platform to collapse through adverse selection.

B. Multi-Criteria System Governance Objectives

The coordination system balances three main goals:

Quality Admission Control: Before the smart contract lists a dataset in the public searchable catalogue, the dataset has to meet a minimum statistical completeness level and remain under an upper limit on its similarity to existing data.

Atomic Settlement Integrity: Escrowed buyer funds go to the seller only if the valid parameters for decrypting the payload are published on-chain, so neither side bears the risk of the counterparty defaulting.

Gas and Storage Efficiency: Heavy computation and bulk file storage are moved off the blockchain; only cryptographic commitments are kept on-chain.

III. ARCHITECTURAL TAXONOMY

The AI-powered blockchain-based secure data marketplace is built from four linked layers:

A. Layer 1: Client Application Layer

This layer supplies the interfaces used by sellers and data consumers. Instead of fragile passwords, users authenticate through Web3 cryptocurrency wallets based on asymmetric public-key cryptography. Sellers start uploads, set minimum prices and encrypt their payloads locally, while buyers examine metadata, read quality scorecards and fund conditional escrow transactions.

B. Layer 2: AI Intelligence & Integrity Layer

Running as an off-chain pipeline so as to avoid blockchain fees, the AI engine assesses raw files before they are listed. It carries out three jobs: automated statistical health profiling that covers missing values and schema conformity; two-stage duplicate detection based on locality-sensitive hashing and distances between semantic embeddings; and price recommendations from supervised gradient boosting models.

C. Layer 3: Decentralized Storage Layer

Storing training sets of gigabyte scale directly on public blockchains is not economically viable. Payloads are therefore encrypted on the client side with symmetric keys, using strong cipher schemes, and distributed across an IPFS cluster. IPFS splits files into a cryptographic Merkle graph and gives back a deterministic Content Identifier. The symmetric key stays offline and is handed over only once escrow settlement has been verified.

D. Layer 4: Smart Contract Execution Layer

The blockchain layer serves as the deterministic authority over finances and state. Running on an Ethereum-compatible Layer-2 rollup, the smart contracts store each dataset's Content Identifier, quality certificate and owner address in persistent storage and handle atomic payments through a conditional escrow state machine.

IV. ALGORITHMIC FRAMEWORKS

A. Multi-Criteria Quality Evaluation Engine

To replace subjective reviews written after a purchase, each dataset is profiled statistically in an automated way. The composite quality score, from 0 to 100, is a weighted combination of four statistical dimensions:

Completeness Score: Gives the share of cells that are filled in, non-null and properly structured, across every observation and feature.

Uniqueness Metric: Gauges how many observation rows are repeated, to discourage inflating a dataset with synthetic copies.

Schema Validity: Checks how well features follow typed schemas, permitted formats and expected value ranges.

Class Distribution Balance: For labeled categorical datasets, skewed distributions are penalized with normalized entropy measures. Any dataset that scores under the minimum threshold is rejected automatically.

B. Dual-Layer Duplicate Detection Engine

To stop dishonest sellers from reselling open-source data or relisting assets with minor modifications, the platform applies a two-stage deduplication pipeline:

Syntactic Fingerprinting via Min-Hash LSH: Each dataset is broken into character tokens. Multiple independent hash functions yield signature vectors, and candidate files are compared by how often their signatures collide. Many collisions indicate structural duplicates, which can be found across very large catalogues in sub-linear time.

Deep Semantic Embedding Proximity: For unstructured text or descriptive schemas, sentence-transformer models turn attributes into dense embedding vectors. Comparing these with catalogue vectors by cosine similarity detects semantic copying even when column names have been changed.

C. Algorithmic Dynamic Pricing

Fair market prices are estimated by a gradient-boosted decision tree model that takes into account verified quality scores, number of rows, feature dimensionality, file size in bytes, decay in freshness and buyer demand indices [6]. In cooperative federated settings, the price can be divided according to marginal contribution scoring [10], which reflects how much each contributor improves predictive accuracy.

V. SMART CONTRACT ESCROW & STATE EXECUTION

Autonomous, deterministic smart contracts enforce the key exchange and the payment settlement, removing the need for a custodial intermediary.

A. State Transition Machine

Each dataset transaction moves through a series of discrete states:

LISTED: Once the dataset has cleared AI screening, its Content Identifier hash and quality certificate are recorded on-chain.

ESCROW_FUNDED: The buyer deposits the purchase amount into the smart contract's escrow.

KEY_DISCLOSED: The seller wraps the symmetric decryption key with the buyer's public key and publishes the result on-chain.

SETTLED: Escrowed funds are released to the seller after the buyer confirms that the key matches its hash commitment, or after the dispute window passes.

REFUNDED: The deposit goes back to the buyer if the seller does not provide the key before a set timeout, or if an invalid key proof is verified.

B. Cryptographic Key Encapsulation

So that the decryption key can be delivered securely without any off-chain channel, the seller encapsulates it under the buyer's public key. The smart contract holds the hash commitment of the key. The buyer unwraps the symmetric key with their private key and examines the payload. Should the hash of the derived key differ from the on-chain commitment, the buyer can submit a challenge proof, which automatically triggers a refund.

C. Solidity Implementation Reference

```
contract DataMarketEscrow { enum State { Listed, Funded, KeyDisclosed, Settled, Refunded } struct Listing { bytes32 ipfsCIDHash; uint256 qualityScore; uint256 price; address payable seller; State state; } mapping(uint256 => Listing) public listings; function purchaseDataset(uint256 id) external payable { Listing storage item = listings[id]; require(msg.value == item.price, "Exact price required"); require(item.state == State.Listed, "Invalid state"); item.state = State.Funded; } function discloseKey(uint256 id, bytes memory encKey) external { Listing storage item = listings[id]; require(msg.sender == item.seller, "Caller must be seller"); require(item.state == State.Funded, "Escrow not funded"); item.state = State.KeyDisclosed; } }
```

VI. LITERATURE SURVEY (2020–2026)

Work on commercializing data in a decentralized way started with Ocean Protocol [5], which introduced access through data tokens and the idea of on-premises compute-to-data. Ocean, however, delegated quality profiling to external front ends, so listings remained open to spam [4]. Kayikci and Khoshgoftaar [3] grouped the ways in which blockchain and machine learning interact into basic categories. Upadhyay et al. [1] cut upload latency by 58% and contract gas costs by 72% by combining lightweight chains with IPFS. For particular domains, Abisha and Selvaraj [2] created AQUASTREAM for water telemetry, and Anand et al. [10] built exchanges for agricultural technology. More recently, Sharma et al. [6] and Das et al. [11] used gradient-boosted decision trees to automate the pricing of telemetry data. Bodkhe et al. [7] devised transformer models for auditing smart contracts. Choo et al. [13] examined the risk of front-running in public data auctions, and Kar et al. [14] proposed digital watermarking to safeguard data provenance.

VII. OPEN SECURITY & ENGINEERING CHALLENGES

Adversarial Quality Spoofing: Automated assessment gives attackers a reason to game it. Dishonest sellers may craft synthetic features that pass the completeness and uniqueness checks while hiding noisy distributions. Zero-trust validation models that can withstand adversarial inputs [15] are therefore urgently needed.

Front-Running and Validator Extraction: In public mempools, arbitrage bots watch pending transactions and jump ahead of buyers' bids [13]. Marketplaces need to adopt commit-reveal schemes, threshold encryption or private transaction pools so that bids stay hidden until the block is confirmed.

Smart Contract Vulnerabilities: Escrow contracts that hold high-value funds are prime targets for attacks such as re-entrancy and integer overflow. Before deployment, contract bytecode should be checked with automated, transformer-based static auditing tools [7] to avoid losing funds.

Ledger Permanence vs. Privacy Mandates: The immutability of blockchains clashes with rules such as the GDPR "Right to be Forgotten". After an IPFS Content Identifier that points to personal attributes has been written to a ledger, it cannot be removed. Marketplaces should use searchable encryption [12] and zero-knowledge proofs so that compliance can be recorded without keeping personal data permanently.

Storage Persistence & IPFS Pinning: Content addressing protects the integrity of hashes but does not promise that data stays available forever. When seller nodes go offline and pinning incentives expire, IPFS garbage collection deletes the payloads. Linking escrow contracts to storage incentives on networks such as Filecoin is essential for long-term persistence.

VIII. EMERGING RESEARCH FRONTIERS

Zero-Knowledge Quality Proofs: Zero-knowledge proofs would let a seller demonstrate that a dataset meets quality standards without revealing the underlying records before payment.

Federated Compute-to-Data: Instead of downloading raw data, buyers would train models securely within trusted enclaves, which stops proprietary files from being copied without permission.

Layer-2 Rollups: Batching micro-transactions off-chain gives finality within milliseconds and lowers transaction fees by more than 98% [8].

IX. DOMAIN-SPECIFIC CASE STUDIES

Industrial IoT & Microgrids: Continuous sensor streams recording vibration, power consumption and heat output support predictive maintenance in factories while keeping proprietary production processes private [6], [8].

Healthcare & Clinical Informatics: Compute-to-data protocols that preserve privacy let oncology researchers train models on verified clinical scans, with no patient identifiers leaving the hospital's servers [8].

Smart Cities & Fleets: Autonomous vehicles exchange edge-collected driving logs; Min-Hash deduplication removes repetitive highway telemetry, and rare hazard events are priced dynamically.

Agri-tech & Environmental Telemetry: Automated screening weeds out malfunctioning field sensors, and Shapley valuation pays farmers fairly for their microclimate data [2], [10].

X. CONCLUSION

Combining artificial intelligence with distributed ledger technology tackles a core problem of today's digital economy: how to monetize data in a verifiable, trustless and fair way. Traditional centralized brokers bring single points of failure, non-transparent pricing and widespread exposure to counterfeit assets. Likewise, a blockchain without AI support can only guarantee that transactions are valid and stays blind to the quality and originality of the data it records. By joining decentralized IPFS storage and conditional smart contract escrow with automated machine learning quality profiling, Min-Hash locality-sensitive deduplication and algorithmic valuation, these architectures provide an end-to-end framework for trading digital assets reliably [1]–[15].

ACKNOWLEDGMENT

The authors thank Dr. Pooja Jaiswal for her academic mentoring, critical feedback and supervision of this review. They also thank the Department of Computer Science and Engineering (Data Science) at Guru Nanak Institute of Technology, Nagpur, for the research facilities that made this work possible.

AUTHOR PROFILES

Hanny Kuhikar, Prince Raidas and **Bhumika Tambuskar** are undergraduate student researchers in the Department of Computer Science and Engineering (Data Science) at Guru Nanak Institute of Technology, Nagpur, Maharashtra, India. **Dr. Pooja Jaiswal** heads the Department of Computer Science and Engineering (Data Science), and **Dr. Sudhir N. Shelke** is the Principal of Guru Nanak Institute of Technology (GNIT), Nagpur, Maharashtra, India.

REFERENCES

[1] V. Upadhyay, J. Kokila and A. Vaish, "A dual framework for optimized data storage and retrieval using lightweight python blockchain and scalable smart contracts with IPFS," in Proc. IEEE Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), 2026.

[2] D. Abisha and S. K. Selvaraj, "AQUASTREAM: A decentralized data marketplace for secure sharing and access," IEEE Access, vol. 13, pp. 11045–11058, 2025.

[3] S. Kayikci and T. M. Khoshgoftaar, "Blockchain meets machine learning: A survey," IEEE Trans. Big Data, vol. 10, no. 4, pp. 389–406, 2024.

[4] IEEE Technical Committee on Data Engineering, "Decentralized data marketplace using blockchain and IPFS," in Proc. IEEE Int. Conf. Blockchain Informatics (ICBI), pp. 240–251, 2023.

[5] T. McConaghy et al., "Ocean Protocol: Tools for the Web3 data economy," Technical Whitepaper, Ocean Protocol Foundation, 2020.

[6] P. Sharma, R. K. Singh and A. K. Verma, "Decentralized industrial IoT data valuation via LightGBM," IEEE Trans. Ind. Informatics, vol. 22, no. 1, pp. 412–423, 2026.

[7] U. Bodkhe, S. Tanwar and R. Gupta, "Transformer-based smart contract auditing in data markets," Expert Syst. with Appl., vol. 248, Art. no. 123450, 2026.

[8] A. K. Das, S. Mohanty and N. Garg, "Medical telemetry monetization via Layer-2 rollups," IEEE Internet of Things J., vol. 13, no. 2, pp. 1890–1902, 2026.

[9] S. Tanwar, R. Kumar and P. K. Sharma, "Threat mitigation framework for Web3 data exchanges," Comput. Netw., vol. 241, Art. no. 110214, 2026.

[10] V. Anand, S. Gupta and M. K. Mishra, "Shapley-driven agritech data monetization platform," IEEE Access, vol. 14, pp. 4512–4525, 2026.

[11] T. K. Das, S. Roy and A. K. Sahoo, "Explainable AI data pricing in Web3 marketplaces," IEEE Trans. Services Comput., vol. 19, no. 2, pp. 810–822, 2026.

[12] A. S. Sadiq, E. B. Agyekum and S. Misra, "Searchable encryption for decentralized dataset catalogs," IEEE Access, vol. 14, pp. 9102–9115, 2026.

[13] K. R. Choo, M. Conti and D. Donno, "Mitigating front-running and MEV in data auction pools," IEEE Trans. Dependable Secur. Comput., vol. 23, no. 1, pp. 312–326, 2026.

[14] S. Kar, D. Roy and P. P. Ray, "Dynamic watermarking for AI training dataset provenance," IEEE Trans. Consumer Electr., vol. 72, no. 2, pp. 210–222, 2026.

[15] V. Sharma, I. You and K. K. R. Choo, "Zero-trust architecture for validating AI dataset integrity," IEEE Commun. Mag., vol. 64, no. 3, pp. 112–119, 2026.