BlockX: A Secure, Decentralized Data Marketplace

George Chong gchong@ucsd.edu
Joey Kaminsky jkaminsky@ucsd.edu
Lincoln Ma lim007@ucsd.edu
Su Aye suaye@ucsd.edu

Sheffield Nolan
shieffieldnolan@franklintempleton.com

Abstract

The world of data analytics and blockchain technology is rapidly evolving, with the global blockchain market projected to reach $163.83 billion by 2029 and the data analytics market expected to hit $329.8 billion by 2030. With data being referred to as "the new oil," cyberattacks against local and cloud storage have reached new heights that surmounted to $530 million in reported losses in 2023 alone; this necessitates a new data and code storage method, one where data can be encrypted, securely stored, and easily accessible, all while being able to handle the increasing amount of data being stored.

A decentralized, secure data marketplace on the Ethereum Virtual Machine (EVM) meets these key requirements: it can securely–and optimally–store large amounts of data on the InterPlanetary File System (IPFS) using encryption techniques, on top of allowing seamless data-related transactions within minutes. Prior research attempted to instantiate this, but these versions neglected to account for potentially dangerous datasets, the exclusion of nonconsensual personal identifiable information (PII), and trustless integration of analytical capabilities that allow for secure, transparent data exchanges without compromising privacy.

To address this issue, we implemented BlockX: a decentralized data marketplace on the EVM with IPFS storage–optimized via encryption and verified with dangerous data and PII checks. This decentralized application is a safe and secure data ecosystem that aids the advancement of artificial intelligence and helps businesses and individuals with their data needs.

Code: https://github.com/gchongg/blockx

1 Introduction 3
2 Methods 4
3 Results 10
4 Discussion 11
5 Conclusion 12
References 14

1 Introduction

Blockchain technology, which gained widespread popularity at the introduction of Bitcoin in 2009, has transformed how transactions–of various forms such as cryptocurrency, assets, services, among other goods and services–are made by enabling secure, decentralized, and transparent systems without reliance on third-party entities. In many fields such as technology, healthcare, and finance, there is a need to upgrade the security pertaining to data storage and transactions; this is due to data breaches that have surmounted to $530 million in losses in 2023 alone–only including reported cases (Lynch, Cybercrimes). This does not even account for the damages caused by the leakage of personal identifiable information (PII). Traditional centralized systems are vulnerable to cyberattacks and unauthorized access that compromise sensitive data and lead to significant damages. The need for data is not going away either, as indicated by widespread digitization, expansion of the Internet of Things (IoT), and advances in artificial intelligence and other technologies. Therefore, as the volume of data held by individuals and companies grows and concerns around protecting those data increase, decentralized solutions offer a promising alternative, providing enhanced security, transparency, and user control. Specifically, there is a noticeable need for a quick and seamless data and data analytics decentralized marketplace that ensures protection from possible CSV-embedded attacks, all while accounting for potential security such as third-parties.

Deploying a decentralized marketplace that allows for the transparent transaction of data, where these items are searched for possible scams and PII, allows for a safe environment to trade data and data-related resources for both individuals and businesses. This can facilitate further innovation in an already rapidly changing field of artificial intelligence and can help both individuals and businesses with their current data-related problems. Furthermore, the use of IPFS and encryption methods adheres to the decentralized nature of blockchain and optimizes the storage space needed, leading to further data security and cheaper transactional costs.

In *Secured Document Storing Using Blockchain*, Jha et al. explored the idea of using blockchain and smart contracts to safely store and sell data. They sourced the need for this from the recent trend in overall data privacy concerns, where thousands of companies have had and reported data breaches that have resulted in millions of dollars of damage. Storing data in an Interplanetary File System (IPFS) and hosting the sale of data within the blockchain can provide the security measures needed to solve the issue of data breaches. IPFS uses distributed hash tables, quick switch, and self-certified file systems to safely store data and allow proper users to retrieve it (*Secured Document Storing Using Blockchain*). Pairing this with a user's private key on the blockchain, it is impossible to retrieve the data unless the user's private key is stolen. This method is much safer as it is easier to protect a private key compared to guarding the datasets in-house for companies. However, this project neglects to take into account potential security concerns associated with the users uploading the data; this includes scams (like empty datasets), unsolicited PII information included, and potential SQL injections. Therefore, this publication provides a solid background to the decentralized data marketplace we are instantiating, where we will expand

upon validating the data, in terms of ensuring there are no scams, nonconsensual PII data, nor dangerous contents.

In *Research on Data Transaction Security Based on Blockchain*, Jiang et al. also highlighted the need to transition to decentralized data marketplaces. Their focus on data privacy, both in terms of access to data and adhering to the decentralized nature of blockchain with IPFS storage, led them to use symmetric and homomorphic encryption on blockchain technology; these methods allow the marketplace to exclude third-parties via blockchain technology, only let proper owners access relevant data via homomorphic encryption, and optimize data storage size with symmetric encryption (*Research on Data Transaction Security Based on Blockchain*). While this marketplace improves on previous research and instantiations, a flawed assumption they use is that every data seller is trusted. Therefore, they do not explore the realm of protecting the buyer from potential attacks by the seller within the data; because the marketplace is publicly available–indicating that anyone can upload data and anyone can buy that data–it is imperative to adopt a system that protects buyers from SQL attacks, on top of protecting possible PII information contained.

Based on the history of blockchain and literature review, there is a real need for a safe and secure decentralized marketplace for data and data analysis. Thus, we created a data marketplace to facilitate the sharing of data. PII data and SQL injection detection and scrubbing is run on each dataset uploaded to the marketplace to ensure a safe transaction. On top of this, data is optimally compressed, encrypted, and then stored in IPFS to further protect the data. And when buyers locate the data they want to buy via the easy-to-use interface and filtering functionality, a quick transaction can be executed within minutes via a smart contract. This will be useful for both individuals and businesses; individuals can find data to use for personal reasons and sell data that they have spent time and effort collecting, whereas businesses can buy data based on business or research needs and sell valuable code and large datasets. This will allow businesses to build safe databases, enabling benefits such as improved AI integration and enhanced research capabilities.

2 Methods

Our data marketplace can be divided into four sections: interface and server to host the marketplace, publishing offerings, downloading data, and personal postings and purchases. We designed each of these sections with the core concepts of blockchain technology in mind so that our marketplace is a valid decentralized application; this includes utilizing smart contracts on a decentralized network, setting up trustless and transparent operations with open-source code, and prioritizing for safe and secure data transactions.

Interface

To construct our marketplace, we identified key characteristics of digital marketplaces: a home page dedicated to who they are, a location for sellers to upload their product, and

a digital catalog where buyers are able to buy what they want. Therefore, we constructed our interface is a similar format.

Home

We created a landing page that highlights our brand and includes a basic description of the marketplace–both a quick text description and a tutorial on how to use all of its different features. In addition, users can connect their wallet, which is the last key format to prepare users for using our marketplace.

Publish

The next page is dedicated to allowing users to upload their data. A simple-to-use interface was prioritized to not deter sellers that are unfamiliar with blockchain technology. This means our smart contract and data-checking analytics, while having their code included in the repository of the website, are hidden from the website interface. These functions, and their integration in our marketplace, align with blockchain principles of a decentralized, trustless, and open-source environment. Smart contracts operating on the Ethereum Virtual Machine (EVM) are the backbone for Ethereum blockchain technology. Furthermore, the Lambda functions operate in such a way that they uphold the trustless, decentralized, and secure nature of blockchain technologies–more details provided in the *Publishing Data* and *Downloading Details* sections. And inclusion of the code in the repository (as well as the code for the smart contract) creates transparency of how our marketplace operates. These functions facilitate a seamless and easy experience for the user, where they can upload a dataset in the matter of minutes.

Catalog

The third page is designed to display dataset offerings to potential buyers. With backend calls that read in datasets from the blockchain, users can explore potential datasets to purchase based on their needs; they have the ability to filter based on price and description. By reading straight from the blockchain, we uphold the 'no-third-parties' cornerstone of blockchain.

On this page, users can purchase data seamlessly. With a single click, they gain access to their chosen dataset within minutes, thanks to smart contract integration, which efficiently retrieves the data from IPFS. IPFS enhances blockchain's decentralized nature by providing secure, decentralized file storage, enabling rapid and reliable access to purchased datasets. To prevent unauthorized distribution, symmetric encryption safeguards the data, allowing only approved users–whose public address are registered–to decrypt and access it. Once a buyer purchases the dataset, their public address is added to the access list, ensuring that only they can retrieve and decrypt the data. The underlying principles behind these mechanisms will be explored in greater detail later.

My Datasets

The fourth page displays the user's previous purchases made on BlockX. This is done us-

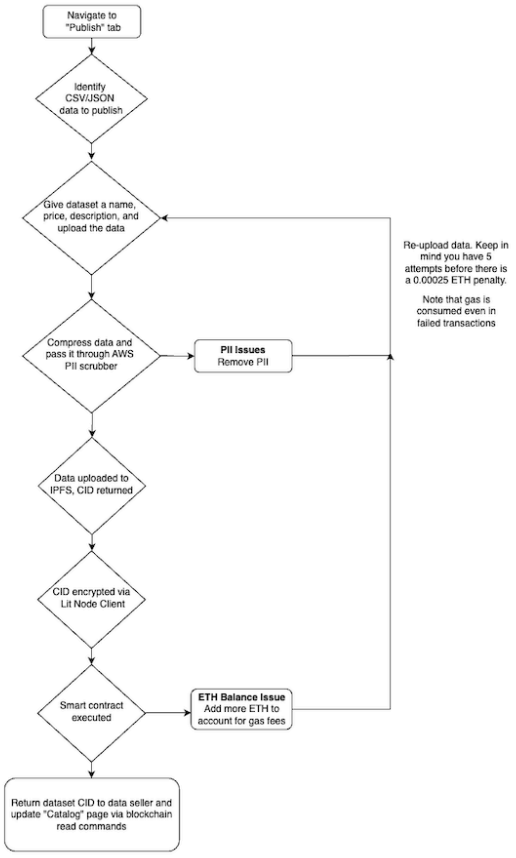
ing read calls for purchase transactions made by them, allowing them to re-download previously purchased data and code. This aligns with blockchain's core principles of transparency and immutability, ensuring that the transaction history remain verifiable and tamper-proof.

Listed Datasets

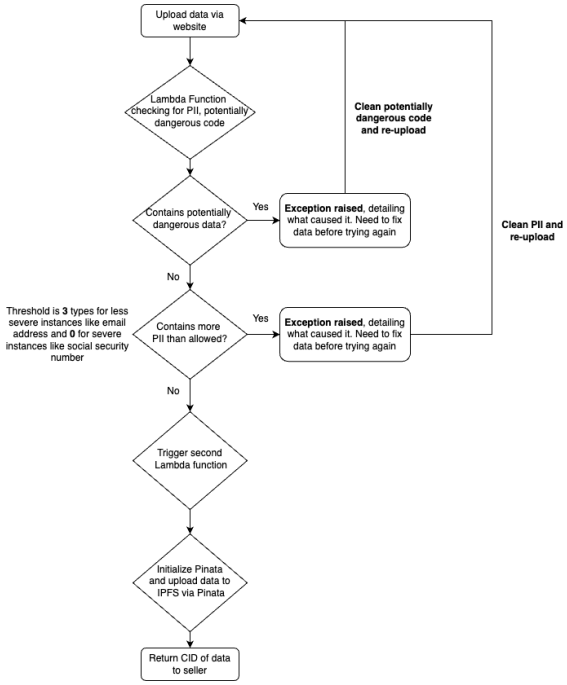
The fifth–and final–page displays which datasets the user have published to the market. Again, this is achieved with back-end calls that read in datasets from the blockchain. On this screen, sellers can also de-list data if they no longer want to sell it. While transactions (such as publishing data) cannot be removed from the blockchain, this action removes the dataset from the website's catalog. This allows users to remove old postings while still maintaining core blockchain principles.

Our marketplace is designed with a strong focus on user experience, ensuring a seamless and intuitive interface regardless of the user's familiarity with blockchain technology. The backend, featuring open-source smart contracts and Lambda functions, aligns with blockchain principles of decentralization, transparency, and trustlessness. As a result, this platform serves as a true embodiment of a decentralized application.

Publishing Data



"Publish Data" Smart Contract Diagram



Data Scan and Scrub, IPFS Upload Lambda Functions Diagram

As detailed in the smart contract diagram above, to upload data, users navigate to the "Publish" tab of our website. Once determining which CSV or JSON file to upload, users select that file and provide basic information, including dataset name, description, and price. Upon confirming this information, the user will *upload* the data. This executes the smart contract, where a function for checking and handling the data is executed prior. While the user will get a simple success message if the contract executed correctly along with their IPFS-created dataset content identifier (CID), a sophisticated process is carried out–as seen in the image regarding the Lambda functions above.

Once the user clicks *upload*, an AWS Lambda function is executed. Using AWS Lambda was chosen over on-chain analytics as we can still align with the principles of blockchain technologies and avoid the high costs associated with on-chain analytics. Using lambda functions for in-line data verification and scrubbing does not compromise decentralization because the processed data is directly pushed to IPFS, ensuring that it is stored and accessed only in a distributed environment. While Lambda functions execute within a centralized cloud service, they are only a transient step for data preprocessing that have no long-term reliance on centralized resources. Because the function is open-source and deterministic,

it aligns with these core principles of all blockchain applications. This approach balances practicality with decentralization by maintaining a lightweight, efficient processing pipeline that upholds the principles for a secure, trustless, and distributed data marketplace.

For the Lambda function, the dataset is operated on in-line, meaning that it is **never stored** in the cloud. This emphasizes the role that Lambda functions play in this scenario, where no data is permanently retained. This reinforces the trustless nature of blockchain decentralized applications. Within this function, the data is scrubbed for possible malicious and PII data. If malicious data–25+ different forms of SQL injections and suspicious words–is found, there is no room for scrubbing; thus, the function is halted with an exception and the dataset is not uploaded to the marketplace. There is a little room for error with PII. Using natural language processing, the PII checker checks for names, as well as other common forms of PII like phone numbers, email addresses, social security numbers, birth date, etc. If found, these cells are replaced with 'REDACTED'. If more than 3 different forms of PII are located, however, an exception is thrown to halt the function and prevent the data from being uploaded to the marketplace. This buffer is in place to account for honest PII-inclusion mistakes. However, more serious forms of PII like social security numbers are excluded from this buffer as they throw an exception on their own, preventing the upload to the catalog. Users are given 5 attempts per dataset before being assessed a 0.0025 ETH penalty for each following upload attempt; this gives users multiple chances to upload a clean dataset before being charged for their negligence. This is to account for the costs associated with running the Lambda function and acts as a deterrent for future reckless uploads.

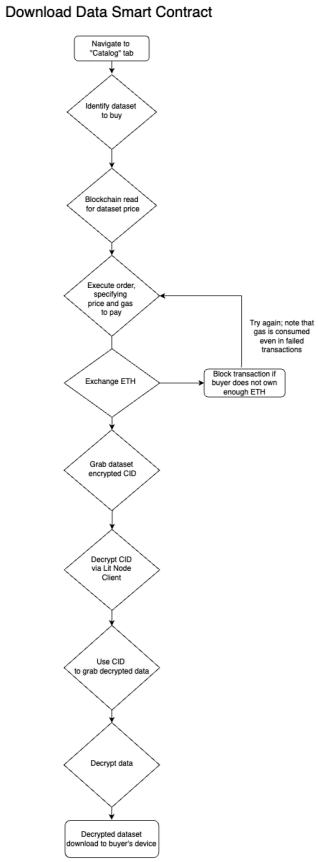
If the dataset is free of potentially dangerous data and PII, the data is partitioned and sent to IPFS nodes to be stored. Uploading to IPFS, specifically the decentralized storage of a file's data chunks, upholds the decentralized nature of our marketplace, where symmetric encryption is used to further protect data. Symmetric encryption uses a key–the seller's public address–to encrypt the data and dataset unique identifier in IPFS (CID); when someone buys the data, their public key is granted the ability to decrpyt the data. In the case where IPFS was compromised, the bad actor would also need the seller's or buyer's public address to decrypt the dataset's CID and then the data. While this information is public, the dataset's encrypted CID and seller's public address are not connected, indicating that it is unlikely that the data would be fully compromised. And only allowing buyers to access the data with their public key adds another level of security to prevent the data from being illegally distributed–as the buyer will be hesitant to allow other individuals to log into their account to access the dataset. This encryption protects the data if IPFS is ever compromised and from buyers with potentially malicious intent for purchasing the data.

Once the data passes the Lambda function, the dataset CID is returned for the smart contract to use. This triggers the smart contract, which uploads the dataset's descriptors and encrypted CID to the blockchain. Once it is processed, the user receives a "success" message declaring the completion of the contract execution.

There are limitations with using Lambda and IPFS. Lambda functions have an input limit of 6 MB and IPFS has high costs. With this in mind, users–as of now–are limited to uploading data under 6 MB. In addition, a marginal, transparent cost for repeated improper

use is charged to the seller to account for the costs associated with Lambda functions and IPFS. This is transferred to the account of the marketplace. Under the assumption that the average seller will take 1.5 times to upload the data (due to PII issues) and 1 in 10 users will have issues with 5 or more consecutive uploads, the improper use fee is 0.0025 ETH. This value was determined to be the minimum charge required to cover the costs associated with the Lambda functions and IPFS (Pinata). The data is not uploaded to the marketplace if the seller cannot afford this fee.

Downloading Data



"Downloading Data"
Smart Contract Diagram

As detailed in the *"Downloading Data" Smart Contract Diagram* above, to purchase and download data from the marketplace, buyers navigate to the "Catalog" tab of our website. Once determining which dataset to purchase, users select that file. Once verifying the price, the user will click *Buy Dataset* to purchase the data. This executes the smart contract, where a function for grabbing the encrypted data from IPFS is executed. The process of this function is detailed in the *IPFS Download Data Function* diagram.

If the seller does not possess enough Ether, the transaction is prevented from executing. If they do have enough, the smart contract executes and adds the dataset to 'My Datasets'; the user is also added to the Lit Node Client associated with the dataset. Once a user wants to download a dataset they bought, they navigate to 'My Datasets' and click 'Download dataset'. This executes a function that verifies that the user's address can access the dataset, decrypts the dataset CID then the dataset itself, and then downloads the dataset to the user's device. Through the addition of the Lit Node Client security feature, only the seller and purchasers of the dataset can access the dataset; this adds another layer of security to prevent purchasers from mass-sharing the data illegally.

Another aspect of the smart contract is its escrow functionality. Escrow is a smart contract that holds funds of a transaction until a certain condition is met. While the buyer is charged the cost of the data when buying it, only after a buyer confirms that they received the data they bought will the funds be exchanged to the data seller. This gives the buyer security from falsely marketed datasets. If a dataset is falsely advertised, the buyer can click a button next to the dataset in 'My Datasets' that says 'faulty data'. The contract will trigger a dispute resolution process where the buyer and seller must submit evidence to a decentralized arbitrator or a trusted third party. The arbitrator will assess the evidence and make a decision on whether the dataset is indeed faulty. If the dataset is found to be faulty, the funds will be refunded to the buyer. If the data is legit, the funds will automatically be released to the data seller. If a buyer does not specify whether they received a valid or faulty dataset within 7 days, the funds are automatically transacted to the data seller.

3 Results

We were able to create a functioning data marketplace with numerous nuances to provide a uniquely secure, easy-to-use marketplace. As highlighted, sellers can upload data using a simple interface, where the brunt of the work is handled by AWS Lambda functions and the smart contract; this includes scrubbing for PII, uploading to IPFS, encrypting the data and unique identifier, and adding the event to the blockchain. These functions and smart contract, on a small dataset, take less than a minute to execute. In addition, the PII-scrubber utilizes natural language.

There are controls to prevent abuse of the functions and storage; encryption of the data further protects the marketplace from harmful activities, on top of further securing the data on IPFS. Buyers can navigate the catalog of datatsets with the helps of the filters, and they can buy any dataset within the matter of minutes.

The design and implementation of our decentralized data marketplace included the development of the following key components:

Interface: The interface was designed with the goal of ensuring seamless and easy user interactions for sellers with varying experience with blockchain technology. The interface has minimal latency with the complicated smart contract logic hidden by a simple and easy-to-use form.

Publishing Data: The smart contract for publishing data was created to ensure secure, transparent data uploads. AWS Lambda function was used to execute quick and scalable PII scrubbing and data uploading to IPFS.

Catalog: Quick blockchain read calls were incorporated into the design to facilitate decentralized data retrieval. The smart contract for buying data was created to ensure secure, transparent, and trustless transactions.

My Datasets: Quick blockchain read calls were integrated into the design to facilitate decentralized data retrieval. In this instance, read calls are made to retrieve the individual's 'buy' smart contract executions. Users are able to quickly and seamlessly decrypt and download data previously purchased. This achieves our goal with minimal wait time.

Listed Datasets: Again, blockchain read calls were incorporated into the design to facilitate decentralized data retrieval. In this instance, calls are filtered to retrieve datasets from 'sell' smart contract executions from the individual.

4 Discussion

The Interface, Publishing Data, Downloading Data, and Retrieving Personal Transactions aspects work together to create a secure, decentralized, and trustless marketplace. The utilization of smart contracts to publish and download data adheres to the core principles of a decentralized application. Blockchain read calls, used to set up the catalog and personal listings, further emphasize the marketplace's reliance on decentralized calls. Finally, the Lambda function has no temporary storage and has open source code, providing transparency and upholding decentralization; these functions align with the trustless and secure–both in terms of encryption and removal of PII–nature of blockchain technologies through their deterministic, transparent methods and encryption offerings. These functionalities, combined, create a decentralized, safe data marketplace.

While Lambda functions execute within a centralized cloud service, they are only used as a small step for data preprocessing that have no long-term reliance on centralized resources. This approach balances practicality with decentralization by maintaining a lightweight, efficient processing pipeline that upholds the principles for a secure, trustless, and distributed data marketplace. The deterministic nature of the functions (verifiable by the open source code) and lack of file storage ensure the functions do not take away from the overall decentralized nature of the marketplace. Different decentralized analysis methods such as the Golem Network can be explored. This desire to explore other ways to conduct the analysis is backed by the limitation of the AWS Free Tier, where API Gateway calls can only wait 29 seconds for the Lambda functions to execute in their entirety; this greatly constricts the use of the Lambda functions to small datasets.

While the product is similar to other available decentralized marketplaces, ours offers a unique combination of security measures that ensures a safe marketplace. Through PII and malicious data (like SQL injection) detection and scrubbing, the platform ensures that

data does not contain sensitive information nor security threats. Therefore, users can be assured that the data they are buying from a seller is safe despite the anonymous nature of blockchain. In addition, the symmetric encryption and IPFS storage ensures the decentralized, safe storage of data for the marketplace. And utilizing Lit Node Client to control who has access to datasets–by determining who has the ability to decrypt the data and its identifier–protects the seller from having their data mass-downloaded; this also protects the marketplace from being used as cloud storage to spread files by another platform. The marketplace is also protected from abuse of Lambda functions by charging a fee for 5 or more failed executions. This ensures the resources are valued. These security measures combined protect the marketplace, buyer, and seller from potentially malicious activities, facilitating a safe medium for data exchanges.

A possible future improvement could be data-seller and overall market protection measures surrounding upload abuse. For instance, a purchaser of data can slightly modify the data and re-upload it as their own in an unauthorized manner. Even further, they could keep re-publishing data to flood the market with false data. This could be fatal to the market and lose the trust of the user-base. A future iteration of our decentralized marketplace can include both the reliability metrics to encourage sellers to share valid data.

Another shortcoming is our data size limitations. Due to Lambda functions and IPFS storage capacity, users are limited to a max file size of 6 MB. This is not enough to be a working data and code marketplace. A necessary improvement would be to increase our Pinata (IPFS) storage limit and work with AWS to increase the Lambda function input size limit.

Our decentralized marketplace has the potential to revolutionize how people buy, sell, and interact online. Our decentralized marketplace offers a safe and secure medium for data exchanges that relies only on minimal fees–in the form of gas for transactions and covering Lambda function and IPFS costs. While it offers many advantages, such as a simple and easy-to-use interface and not relying on third parties, it presents significant challenges in terms of scalability, governance, and potential misuse. Balancing decentralization with user experience, security, and compliance will be key to the success and widespread adoption of our marketplace.

5 Conclusion

Blockchain technology, since the introduction of Bitcoin in 2009, has revolutionized the way transactions are conducted by providing a secure, decentralized, and transparent alternative to centralized systems. As the need for data has increased dramatically, the demand for ways to mitigate risks like data breaches and unauthorized access has grown exponentially. The data marketplace discussed here addresses these challenges by strictly adhering the principles of blockchain technology: transparency, trustlessness, and decentralization.

Through the integration of smart contracts, IPFS, and AWS Lambda functions, the marketplace ensures secure transactions for both sellers and buyers. Sellers can confi-

dently share their datasets knowing that the encryption protocols prevent their data from being leaked. On the other end, buyers benefit from reliable, deterministic code scripts that protect them from accessing datasets with potentially harmful content or PII. The use of symmetric encryption and decentralized storage further bolsters data security and reduces transaction costs.

This platform bridges the gap between data and code accessibility and security, highlighting its value in the realm of decentralized marketplaces. Our approach balances advanced security and encryption, transparency, and decentralized storage, with a user-friendly experience. While there are shortcomings with seller and overall market protections and size- and cost-based limitations, these can be relatively quickly solved to improve the marketplace. The application exemplifies the potential of blockchain technology to create more secure, efficient systems, paving the way for a more secure digital economy.

References

[1] Jha, Sakshi, et al. “Secured Document Storing Using Blockchain.” *IJRTI*, International Journal for Research Trends and Innovation, 2022. Available at: https://www.ijrti.org/papers/IJRTI2205069.pdf.

[2] Jiang, Yongbo, et al. “Research on Data Transaction Security Based on Blockchain.” *MDPI*, Multidisciplinary Digital Publishing Institute, 8 Nov. 2022. Available at: https://www.mdpi.com/2078-2489/13/11/532.

[3] Lynch, Katherine. “Which Cybercrimes Generated the Largest Financial Losses in 2023?” *Verisk*, 7 June 2024. Available at: https://core.verisk.com/Insights/Emerging-Issues/Articles/2024/June/Week-2/2023-Cybercrime-Losses#.