

Н. Карпетыанц¹✉

Методы обработки и хранения данных транзакций сети Bitcoin

¹Национальный исследовательский ядерный университет «МИФИ»,
г. Москва, Российская Федерация
e-mail: nkarapetyants@mephi.ru

Аннотация. Сеть Bitcoin и лежащая в ее основе технология распределенного реестра представляют собой значимый элемент глобальной финансово-технической инфраструктуры, мониторинг которой представляет практический интерес для специалистов в области информационной безопасности в связи с устойчивым ростом объема незаконных криптовалютных операций. Существующие реализации систем работы с транзакционными данными зачастую опираются на разрозненные инструменты без системного разграничения этапов конвейера обработки и моделей организации хранения, что снижает отказоустойчивость сбора гетерогенной информации и эффективность последующих аналитических запросов. Целью настоящей работы является систематизация и обоснование выбора методов агрегации, потоковой обработки и многоуровневого хранения транзакционных данных сети Bitcoin для повышения эффективности их аналитического исследования в задачах информационной безопасности. В работе предложена архитектура отказоустойчивого конвейера обработки данных, формализующая этапы от первичной агрегации информации из узла Bitcoin Core и внешних OSINT-источников до финальной стандартизации, а также обоснован мультимодельный подход к хранению, сочетающий реляционную, колоночную, графовую, резидентную и объектную модели организации данных. Результаты показывают, что выбор инструментов определяется типом данных и характером аналитических запросов. Использование Apache Kafka, Flink, NiFi и Airflow позволяют реализовать буферизацию, потоковую нормализацию, маршрутизацию и оркестрацию конвейера, а комбинация СУБД PostgreSQL, ClickHouse, Neo4j, Redis и хранилище MinIO в совокупности покрывают весь спектр транзакционных и аналитических запросов. Практическая значимость работы состоит в обосновании архитектуры конвейера на базе комбинации инструментов вендора Apache и мультимодельной стратегии хранения, минимизирующей задержки потоковой нормализации и обеспечивающей горизонтальную масштабируемость. Направления дальнейших исследований включают разработку алгоритмов машинного обучения для автоматизированной кластеризации адресов и выявления сложных паттернов транзакционной активности.

Ключевые слова: Bitcoin, Blockchain, KYT, Базы данных, Обработка данных

N. Karapetyants¹✉

Methods for processing and storing bitcoin network transaction data

¹National Research Nuclear University MEPhI, Moscow, Russian Federation
e-mail: nkarapetyants@mephi.ru

Abstract. Abstract. The Bitcoin network and its underlying distributed ledger technology constitute a significant component of the global financial and technical infrastructure, the monitoring of which is of practical interest to information security specialists in light of the steady growth in the volume of illicit cryptocurrency operations. Existing implementations of transaction data processing systems often rely on fragmented tools without a systematic separation of pipeline processing stages and storage organization models, which reduces the fault tolerance of heterogeneous data collection and the

efficiency of subsequent analytical queries. The aim of this study is to systematize and substantiate the selection of methods for the aggregation, stream processing, and multi-tier storage of Bitcoin network transaction data in order to improve the effectiveness of their analytical investigation in information security tasks. The paper proposes the architecture of a fault-tolerant data processing pipeline that formalizes the stages from the primary aggregation of information from a Bitcoin Core node and external OSINT sources to final standardization, and substantiates a multi-model storage approach combining relational, columnar, graph, in-memory, and object data organization models. The results show that the choice of tools is determined by the data type and the nature of analytical queries. The use of Apache Kafka, Flink, NiFi, and Airflow enables the implementation of buffering, stream normalization, routing, and pipeline orchestration, while the combination of the PostgreSQL, ClickHouse, Neo4j, and Redis DBMS together with the MinIO storage collectively covers the full range of transactional and analytical queries. The practical significance of the work lies in the substantiation of a pipeline architecture based on a combination of Apache vendor tools and a multi-model storage strategy that minimizes stream normalization latencies and ensures horizontal scalability. Future research directions include the development of machine learning algorithms for the automated clustering of addresses and the detection of complex patterns of transactional activity.

Keywords: Bitcoin, Blockchain, KYT, ETL, Databases, Data processing

Введение

В современных условиях сеть Bitcoin и лежащая в ее основе технология распределенного реестра трансформировались в значимый элемент глобальной финансово-технической инфраструктуры. Популярность цифровых активов обусловлена возможностью проведения круглосуточных трансграничных платежей с относительно низкими издержками, где ключевую роль играют стейблкоины как инструменты с высокой ликвидностью. Широкая доступность и неизменяемость данных в публичных блокчейнах обеспечивают высокий уровень прозрачности, позволяя любому его участнику проводить транзакционный анализ.

Развитие экосистем криптовалют сопровождаются появлением новых угроз информационной безопасности. Согласно отчетным данным за 2025 год, совокупный объем незаконных криптовалютных операций достиг 158 миллиардов долларов США, что в абсолютном выражении является историческим максимумом. Тем не менее, относительная доля криминальной активности в общем обороте рынка демонстрирует тенденцию к снижению и составляет 1,2% (против 1,3% годом ранее). Особую проблему в сфере кибербезопасности представляют хакерские атаки и программные эксплойты: в 2025 году в результате 150 инцидентов было похищено активов на сумму 2,87 миллиарда долларов [1, 2]. Подобная статистика подчеркивает необходимость совершенствования средств защиты, использующих методы сбора, обработки, хранения и анализа транзакционных данных.

Нормативно-правовая база в области регулирования криптовалют развивается в направлении ужесточения надзора за финансовыми потоками и внедрения международных стандартов противодействия отмыванию доходов. Современные регуляторные подходы ориентированы на использование аналитических инструментов блокчейна для идентификации подозрительных паттернов и пресечения незаконной деятельности на ранних стадиях. Адаптация законодательства

к архитектурным особенностям блокчейна направлена на минимизацию рисков киберпреступности путем установления требований к операторам по обеспечению сквозной прозрачности всех операций с цифровыми активами.

Исследования в области данной проблематики осуществляется ведущими учеными такими как Басыня Е.А., Стрельцов А.С., Французова Г.А., Мелихова Ю. А., Анаденков Д. В., Милосердова Л. П., Гапанович А. М., Талай Е. В., Виларков Н. В., Бочаров М. И. [3 – 8].

В работе [9] был разработан виртуальный испытательный стенд для автоматизации процессов сбора, обработки, хранения и анализа транзакционных данных сети Bitcoin. В текущем исследовании рассматриваются подходы расширяющие возможности обработки и хранения данных транзакций сети Bitcoin.

Методы и материалы

Процесс агрегации первичных данных из сети Bitcoin (через ZMQ и RPC-интерфейсы узла Bitcoin Core) [9] и внешних интернет-источников (форумов, OSINT-отчетов, баз вредоносного ПО) требует создания отказоустойчивого слоя промежуточного хранения. В качестве основного инструмента агрегации и первичной буферизации обосновано использование Apache Kafka [10]. Данная платформа с открытым исходным кодом позволяет объединять разнородные потоки данных в единый поток событий, обеспечивая высокую пропускную способность и гарантированную доставку сообщений. Это необходимо для предотвращения потери данных в условиях пиковых нагрузок в блокчейн-сети и при нестабильной скорости получения ответов от веб-ресурсов.

Поточная обработка данных в режиме реального времени делегируется Apache Flink [11]. Этот инструмент позволяет проводить мгновенную нормализацию и очистку поступающих транзакций и текстовых отчетов. В отличие от аналогов, Flink обеспечивает развитую поддержку состояний, что необходимо для выявления паттернов подозрительной активности в режиме мягкого реального времени – например, для сопоставления адреса из текущего блока с упоминанием в базе вредоносного ПО или на форумах. Это позволяет структурировать данные еще до их попадания в постоянное хранилище, сокращая задержки при последующем анализе.

Для специфической задачи структуризации и предварительной очистки неструктурированной информации из открытых источников (тексты форумов, отчеты, логи вредоносного ПО) целесообразно использование Apache NiFi [12]. Данный инструмент предоставляет расширенные возможности по управлению потоками данных и автоматизации маршрутизации между различными парсерами и нормализаторами. Преимущество NiFi заключается в его гибкости при работе с изменчивыми форматами данных из сети Интернет и возможности визуального контроля происхождения данных.

Координация и автоматизация всех этапов обработки – от первичной агрегации до финальной трансформации может осуществляться оркестратором Apache Airflow [13]. Он позволяет формализовать зависимости между процессами в виде направленных ациклических графов. Применение Airflow обосновано необходимостью строгого контроля последовательности операций очистки и структуризации, а также

обеспечения отказоустойчивости всей системы за счет механизмов автоматического перезапуска задач и мониторинга состояния конвейеров данных.

Завершающим этапом обработки является приведение данных к стандартизированным структурам (например, в форматах Apache Avro или Apache Parquet) [14] перед их последующей передачей в системы хранения. Этот процесс включает окончательную трансформацию атрибутов и проверку схем данных на соответствие заданным правилам. Использование бинарных форматов с поддержкой схем обосновано необходимостью минимизации объема подготовленных данных и обеспечения их высокой совместимости с любыми аналитическими инструментами, которые будут задействованы на следующих этапах.

Исследование методов хранения данных для последующего анализа сети Bitcoin обусловлено необходимостью обработки разнородной информации: структурированных данных из Bitcoin Core (блоки, транзакции) и неструктурированных данных из внешних источников (форумы, отчеты о мошенничестве, сигнатуры вредоносного ПО). Использование различных моделей организации данных позволяет оптимизировать выполнение как транзакционных, так и аналитических запросов в рамках единой экосистемы.

Реализация строковой организации данных (реляционной модели) наиболее эффективно осуществляется посредством СУБД PostgreSQL [15]. Данный инструмент с открытым исходным кодом оптимален для хранения метаданных, информации о пользователях форумов и структурированных отчетов о мошенничестве. PostgreSQL поддерживает строгую согласованность (ACID), систему индексации и возможность построения сложных связей между сущностями внешних источников и конкретными идентификаторами в блокчейне. PostgreSQL выступает в роли основного реестра для нормализованных данных, требующих высокой точности и частого обновления отдельных записей.

Анализ массивов исторических данных требует колоночной организации, которую предоставляет СУБД ClickHouse [16]. Блокчейн-данные, такие как значения входов/выходов и временные метки транзакций, характеризуются высокой избыточностью, что позволяет ClickHouse достигать значительной степени сжатия и обеспечивать высокую скорость выполнения агрегирующих запросов. Использование этого инструмента позволяет производить быстрое сканирование миллиардов записей для выявления аномалий в объемах передаваемых средств без критической нагрузки на основные транзакционные базы.

Исследование топологии сети и сложных взаимосвязей между адресами опирается на графовую организацию данных, реализуемую в Neo4j [17]. В контексте Bitcoin транзакции и адреса естественным образом формируют направленный граф. Применение Neo4j обосновано эффективностью алгоритмов обхода графа для выявления цепочек перемещения активов, кластеризации адресов, принадлежащих одному субъекту, и отслеживания пути криптовалют от источника (например, адреса, связанного с вредоносным ПО) до конечного получателя. Графовая модель позволяет выполнять запросы на поиск кратчайших путей, которые крайне ресурсозатратны в традиционных реляционных СУБД.

Высокая скорость доступа к критически важным и часто запрашиваемым объектам может быть обеспечена за счет резидентного хранения (in-memory) на

базе Redis [18]. Данный инструмент лучше подходит для кэширования актуального набора неизрасходованных выходов (UTXO), оперативного мониторинга сети в реальном времени и хранения промежуточных результатов обработки данных из блокчейн-обозревателей. Redis обеспечивает минимальные задержки доступа к данным, что является критическим фактором для систем оперативного обнаружения подозрительной активности, где время реакции на инцидент строго ограничено.

Систематизация больших объемов неструктурированной информации, таких как бинарные файлы блоков, дампы веб-страниц или файлы вредоносного ПО, возлагается на объектно-ориентированное хранилище MinIO [19]. Это высокопроизводительное S3-совместимое решение с открытым кодом позволяет масштабировать хранилище горизонтально и эффективно управлять объектами любого размера без жесткой схемы данных.

Результаты и обсуждения

В результате проведенных исследований методов обработки и хранения данных транзакций сети Bitcoin была сформирована сводная таблица 1, наглядно демонстрирующая распределение технических средств в соответствии с их функциональным назначением, типом решаемых задач и характеристиками обрабатываемой информации. Данная систематизация позволяет сопоставить возможности современных Big Data инструментов с архитектурными требованиями системы, обеспечивая комплексный подход к анализу как строго структурированных блоков транзакций, так и нелинейных данных из внешних источников.

Внедрение конвейера на базе Kafka и Flink позволяет обеспечить минимальные задержки при потоковой нормализации данных, а связка NiFi и Airflow дает возможность автоматизировать отказоустойчивый сбор неструктурированной информации из внешних источников. Применение бинарных форматов Avro и Parquet способствует оптимизации последующего анализа за счет механизмов строгой типизации и высокого сжатия данных, что минимизирует нагрузку на вычислительные ресурсы.

Таблица 1. Средства обработки и хранения данных транзакций сети Bitcoin

Инструмент	Решаемая задача	Тип обрабатываемых данных
Apache Kafka	Промежуточное хранение, буферизация и агрегация потоков	Разнородные события из узла Bitcoin Core и внешних API
Apache Flink	Поточная обработка, нормализация и фильтрация	Транзакции в реальном времени, текстовые оповещения
Apache NiFi	Управление маршрутизацией и первичная очистка данных	Неструктурированные данные (форумы, OSINT-отчеты)
Apache Airflow	Оркестрация рабочих процессов и контроль последовательности задач	Метаданные конвейеров, зависимости между этапами

Окончание табл. 1

Инструмент	Решаемая задача	Тип обрабатываемых данных
PostgreSQL	Хранение структурированных реестров и метаданных	Информация о пользователях, отчеты о мошенничестве
ClickHouse	Аналитическая обработка больших исторических массивов	Временные метки, значения входов/выходов транзакций
Neo4j	Исследование топологии сети и графовых взаимосвязей	Адреса, связи между транзакциями, кластеры субъектов
Redis	Резидентное хранение для обеспечения минимальных задержек	Актуальный набор UTXO, оперативные результаты мониторинга
MinIO	Объектное хранение неструктурированных массивов	Бинарные дампы блоков, файлы вредоносного ПО, PDF-отчеты
Avro / Parquet	Финальная структуризация и оптимизация хранения	Подготовленные аналитические датасеты со строгой схемой

В части хранения применяется мультимодельный подход, где выбор модели данных продиктован спецификой аналитических запросов. PostgreSQL обеспечивает необходимую согласованность метаданных, ClickHouse позволяет достичь кратного ускорения агрегирующих запросов, а Neo4j способствует эффективному выявлению графовых связей и кластеризации адресов. Использование Redis и MinIO позволяет сбалансировать скорость доступа к оперативным данным и общую горизонтальную масштабируемость системы.

Заключение

В ходе проведенного исследования были рассмотрены методы агрегации, обработки и многоуровневого хранения данных сети Bitcoin и сопутствующих внешних источников. Практическим результатом работы является обоснование архитектуры высокопроизводительного конвейера на базе Apache Kafka, Flink и Airflow, а также применение мультимодального подхода хранения данных, объединяющей реляционный, колоночный и графовый методы хранения данных. Сформированный перечень инструментов и методов позволяет обеспечить отказоустойчивую нормализацию гетерогенных данных, преобразуя сырые транзакции и неструктурированные данные в стандартизированные форматы Avro и Parquet для последующего анализа.

В рамках дальнейших исследований планируется реализация алгоритмов машинного обучения на базе подготовленных массивов данных для автоматизированной кластеризации адресов и выявления сложных паттернов транзакционной активности, связанных с угрозами информационной безопасности.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. The Chainalysis 2026 Crypto Crime Report [Электронный ресурс] // Chainalysis. URL: <https://www.chainalysis.com/reports/crypto-crime-2026/> (дата обращения: 28.04.2026).

2. 2026 Crypto Crime Report [Электронный ресурс] // TRM. URL: <https://www.trmlabs.com/reports-and-whitepapers/2026-crypto-crime-report> (дата обращения: 28.04.2026).
3. Basynya E. A., Karapetyants N., Karapetyants M. Bitcoin Transaction Analysis System //Programming and Computer Software. – 2024. – Т. 50. – №. Suppl 2. – С. S104-S112.
4. Стрельцов А. С., Французова Г. А., Басыня Е. А. Разработка системы сбора, обработки, анализа, идентификации корреляции событий информационной инфраструктуры предприятия //Системы анализа и обработки данных. – 2023. – №. 1 (89). – С. 101-113.
5. Мелихова Ю. А., Анаденков Д. В. Блокчейн bitcoin: принципы работы и особенности //Проблемы развития современного общества. – 2023. – С. 287-289.
6. Милосердова Л. П. Системные аспекты обработки транзакций в блокчейн-узлах: конкурентность и управление памятью //Актуальные вопросы современной науки 3. – 2026. – С. 30.
7. Гапанович А. М., Талай Е. В. О криптокошельках, транзакциях и аресте криптовалюты //ББК 67.408 я43 О75. – 2024. – С. 324.
8. Вилаков Н. В., Бочаров М. И. Способы защиты криптовалютных блокчейн систем и систематизация угроз //Системная инженерия и информационные технологии. – 2025. – Т. 7. – №. 2 (21). – С. 143-154.
9. Mun H., Lee Y. Bitsql: A sql-based bitcoin analysis system //2022 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). – IEEE, 2022. – С. 1-8.
10. Raptis T. P., Passarella A. A survey on networked data streaming with apache kafka //IEEE access. – 2023. – Т. 11. – С. 85333-85350.
11. Mezati M., Aouria I. Flink-ML: machine learning in Apache Flink Flink-ML: aprendizado de máquina no Apache Flink Flink-ML: aprendizagem automático en Apache Flink //Brazilian Journal of Technology. – 2024. – Т. 7. – №. 4. – С. 01-18.
12. Carthen C. et al. Orchestrating apache nifi/minifi within a spatial data pipeline //2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA). – IEEE, 2023. – С. 366-371.
13. Haines S. Workflow orchestration with apache airflow //Modern Data Engineering with Apache Spark: A Hands-On Guide for Building Mission-Critical Streaming Applications. – Berkeley, CA : Apress, 2022. – С. 255-295.
14. Nelluri S. R., Saldanha F. A. A. Mastering Big Data Formats: ORC, Parquet, Avro, Iceberg, and the Strategy of Selection //International Journal of Computer Trends and Technology. – 2025. – Т. 73. – №. 1. – С. 44-50.
15. Salunke S. V., Ouda A. A performance benchmark for the PostgreSQL and MySQL databases //Future Internet. – 2024. – Т. 16. – №. 10. – С. 382.
16. Schulze R. et al. Clickhouse-lightning fast analytics for everyone //Proceedings of the VLDB Endowment. – 2024. – Т. 17. – №. 12. – С. 3731-3744.
17. Naha R. T., Zhang K. Cryptocurrencies forensics with real-time intelligence and graph database: A comprehensive review //2024 6th International Conference on Blockchain Computing and Applications (BCCA). – IEEE, 2024. – С. 1-12.
18. Abu Kausar M., Nasar M., Soosaimanickam A. A study of performance and comparison of nosql databases: MongoDB, cassandra, and redis using ycsb //Indian Journal of Science and Technology. – 2022. – Т. 15. – №. 31. – С. 1532-1540.
19. Malhotra S. et al. Evaluating fault tolerance and scalability in distributed file systems: A case study of gfs, hdfs, and minio //arXiv preprint arXiv:2502.01981. – 2025.

© Н. Карпетьянц, 2026