УДК 004.62

РАЗРАБОТКА АРХИТЕКТУРЫ СИСТЕМЫ УПРАВЛЕНИЯ СЕМАНТИЧЕСКИМИ ДАННЫМИ, ОСНОВАННОЙ НА ТЕХНОЛОГИИ БЛОКЧЕЙН

Олимпиев Н.В., Жукова Н.А.

ФГАОУ ВО «Национальный исследовательский университет ИТМО», Санкт-Петербург, e-mail: 307702@niuitmo.ru

Статистические данные свидетельствуют о формировании тренда на рост производимых данных в мире, который способствует развитию систем управления данными. В статье рассмотрено применение технологии блокчейн для управления распределенными семантическими данными в системах управления данными Master Data Management. В работе сформулированы проблемы, связанные с управлением семантическими данными, распределенными по разным источникам, и описано, как технология блокчейн позволяет решить эти проблемы, обеспечивая согласованность и достоверность данных. Цель исследования заключается в разработке архитектуры системы управления семантическими данными с применением технологии блокчейн и с опорой на принципы управления основными данными (MDM). Предложен архитектурный подход к построению системы, основанный на использовании платформы Ethereum 2.0 и Proof of Stake (PoS), IPFS, смарт-контрактов и очередей сообщений, что позволяет осуществлять интеграцию с внешними системами, обеспечивать интероперабельность и децентрализацию при управлении основными данными. Для проверки согласованности онтологий используется моментальный снимок локальной базы данных Apache Jena, что повышает производительность и улучшает масштабируемость системы для работы с семантическими данными. Описанный подход решает проблему отсутствия исследований применимости блокчейна для концепции MDM с семантическими данными и имеет потенциал для использования в областях, где важна безопасность и целостность данных, таких как финансы, здравоохранение, государственное управление или логистика. Полученные результаты подтверждают потенциал блокчейна в управлении данными, но для оценки и реализации системы могут понадобиться дополнительные исследования.

Ключевые слова: семантические данные, блокчейн, Semantic Web, системы управления данными, Master Data Management, Ethereum, децентрализация, основные данные

DEVELOPMENT OF THE ARCHITECTURE OF A SEMANTIC DATA MANAGEMENT SYSTEM BASED ON BLOCKCHAIN TECHNOLOGY

Olimpiev N.V., Zhukova N.A.

ITMO University, Saint Petersburg, e-mail: 307702@niuitmo.ru

Statistical data indicate the formation of a trend for the growth of data produced in the world, which contributes to the development of data management systems. The paper considers the use of blockchain technology for managing distributed semantic data in Master Data Management data management systems. The paper formulates the problems associated with managing semantic data distributed across different sources, and describes how blockchain technology can solve these problems, ensuring data consistency and reliability. The purpose of the study is to develop the architecture of a semantic data management system using blockchain technology and based on the principles of master data management (MDM). An architectural approach to building a system based on the use of the Ethereum 2.0 platform and Proof of Stake (PoS), IPFS, smart contracts and a message queue is proposed, which allows integration with external systems, ensuring interoperability and decentralization when managing master data. To check the consistency of ontologies, a snapshot of the local Apache Jena database is used, which improves performance and improves the scalability of the system for working with semantic data. The described approach solves the problem of the lack of research on the applicability of blockchain for the concept of MDM with semantic data and has the potential to be used in areas where data security and integrity is important, such as finance, healthcare, public administration or logistics. The results obtained confirm the potential of the blockchain in data management, but more research may be needed to evaluate and implement the system.

Keywords: semantic data, blockchain, Semantic Web, data management systems, Master Data Management, Ethereum, decentralization, master data

По информации от компании Statista, специализирующейся на рыночных и потребительских данных, за 2021 г. объем данных по всему миру составил 79 зеттабайт [1]. К 2025 г. по тем же прогнозам общий объем данных достигнет 181 зеттабайт. Формируемый тренд побуждает исследователей развивать и адаптировать системы управления данными к растущим требованиям, а представителей крупного бизнеса – инвестировать в разработки и использовать системы для оптимизации своей работы. При этом с ростом объемов данных более актуальными становятся проблемы их безопасности и целостности, что актуально для предприятий, имеющих несколько филиалов и команд, управляющих локальными данными. Международная ассоциация управления данными (DAMA) определяет управление данными как способность планировать, контролировать и предоставлять информационные активы [2]. Для работы с распределенными данными среди корпораций востребован вид систем Master Data

Management (MDM), основанный на управлении «основными» данными в организации. MDM-системы призваны обеспечить единое и актуальное представление о сущностях и их связях в едином месте путем консолидации информации. По версии компании Gartner топ-5 лидеров рынка MDM в 2023 г. [3] выглядит следующим образом: PiLog MDRM, Intelligent Master Data Management Platform, Semarchy xDM, Stibo Systems MDM, TIBCO EBX. Системы MDM из рейтинга имеют высокие оценки, но сталкиваются с рядом сложностей, которые могут повлиять на их функциональность и эффективность. Существующие системы MDM зачастую основаны на табличной структуре данных, хранимых централизованно, что приводит к ограничениям безопасности и масштабируемости. Таким образом, существующие системы отвечают только частично требованиям по обеспечению скорости, качества данных и масштабируемости для работы в растущих организациях, поэтому разработка интегрированной системы MDM на базе блокчейна является актуальной задачей, требующей дополнительных исследований и разработок. Одним из способов решения описанных проблем является внедрение технологии блокчейн, позволяющей создавать распределенные системы управления данными с высоким уровнем безопасности и целостности. Как отмечают авторы статьи [4], технология блокчейн обеспечивает сдвиг парадигмы в оптимизации бизнес-процессов, обмене данными и совместимости в смежных отраслях, а также обеспечивает новый путь управления данными. Использование технологии возможно при построении систем управления данными для обеспечения безопасности данных и доступа к ним. В работе по теме внедрения блокчейна в системы управления данными [5] авторы выделяют возможные сценарии и преимущества использования технологии, подтверждающие интерес к теме. Рассматривая проблемы уязвимости, централизации и масштабируемости при управлении данными, авторы предлагают структуру управления совместной работой с данными на основе блокчейна с возможностью аутентификации пользователей, валидации данных, распределения нагрузки на узлы и использованием собственной цифровой валюты – datacoin. Открытым вопросом остается практическая применимость работы, поскольку в ней не сфокусировано внимание на конкретных механизмах, используемых блокчейн-платформах, виде системы управления и используемых в работе данных. В дополнение к вышеизложенному авторы работы [6] выделяют три уровня управления данными: архитектуру блокчейна, структуру данных блокчейна и механизм хранения данных блокчейна, а также приходят к выводу, что стандартный блокчейн, как правило, используется для цифровой валюты, гибридный блокчейн подходит для многоорганизационных сценариев, а блокчейны на основе DAG наиболее подходят для Интернета вещей. Несмотря на сделанные выводы, авторы также указывают на недостатки гибридного блокчейна и на то, что использование DAG все еще находится на ранних стадиях. Для обработки чрезмерной нагрузки данных предлагаются методы распределенного хранения и кодирования данных, а также создание дополнительной базы данных для запросов.

Еще одним инструментом в области управления и преобразования данных являются технологии Semantic Web: RDF, OWL, SPARQL, Linked Data. Технологии позволяют семантически обрабатывать данные и обеспечивают стандартизированный подход к их описанию и использованию, что способствует их более эффективной обработке и использованию. Однако управление семантическими данными является сложной задачей, связанной с необходимостью обеспечения их целостности, консистентности и актуальности, что возможно обеспечить использованием MDM-систем. Управление основными данными онтологий позволит использовать преимущества семантических технологий, а также позволит реализовать встраиваемую в рамки концепции Semantic Web MDM-систему. Поскольку семантические данные зачастую распределены, а их сбор требует работы с разнородными источниками, это приводит к проблемам с согласованностью и достоверностью данных. В то же время технология блокчейн позволяет обеспечить аккумуляцию разнородных данных, например авторы работы [7] для управления данными онтологий используют блокчейн, смарт-контракты Ethereum и сеть InterPlanetary File System (IPFS) для децентрализованного хранения данных. Такой стек позволяет избежать ограничений централизованного хранения данных путем распространения общедоступной онтологии среди пользователей. Несмотря на то, что статья демонстрирует эффективность использования блокчейн-технологии в комбинации с Semantic Web для управления данными онтологий, авторами отмечается, что время, необходимое для изменения смарт-контракта в сети Ethereum, может быть ограничивающим фактором, а реализованные проверки согласованности онтологии выполняются локально, и потенциальный злоумышленник

может обойти их. Однако предложенное решение не является полноценной системой управления данными, а скорее играет роль менеджера баз данных онтологий.

Анализ сведений, полученных из литературных источников, показывает, что наиболее известные системы MDM, а также используемые способы внедрения технологии блокчейн для управления данными имеют ряд ограничений. Во-первых, в открытых источниках нами не было найдено реализаций MDM, позволяющих взаимодействовать с семантическими данными, что ограничивает использование MDM-систем в рамках концепции сети Semantic Web. Во-вторых, исследователями по сей день не рассмотрены системы, основанные на концепции MDM и технологии блокчейн одновременно, несмотря на то, что симбиоз данных технологий позволяет решить проблемы с ограничениями существующих MDM-систем. Использование блокчейна и семантических технологий является возможным способом развития систем управления данными, эффективность которых по отдельности подтверждается результатами рассмотренных трудов. Следует заметить, что для объединения технологий в рамках единой MDM-системы необходим дополнительный анализ и аргументация подхода к их внедрению на основе имеющегося опыта. Цель проводимого исследования заключается в разработке архитектуры системы управления семантическими данными с применением технологии блокчейн и с опорой на принципы управления основными данными (MDM). Благодаря такому походу, онтологии могут быть использованы в MDM-системах для определения и классификации данных, а также для обеспечения точности и последовательности в данных, которые хранятся в системе. Они также могут использоваться для поддержки поиска и навигации по данным, которые хранятся в системе. При этом блокчейн позволяет обеспечить безопасность данных и масштабируемость системы, а принципы использования основных данных позволяют обеспечить согласованность и единство представления семантических данных в системе. В то время как варианты использования MDM и семантических технологий в организациях наиболее очевидны, одна из ключевых причин, по которой внедрение блокчейна все еще находится на начальных этапах, заключается в том, что его ценность для бизнеса еще не полностью признана.

Материалы и методы исследования

Для разработки архитектурного подхода использованы методы: декомпозиция требований, анализ и моделирование. Основные требования для разработки архитектуры включают в себя решение проблемы интероперабельности и обеспечение безопасного доступа к системе, а также гарантию качества получаемых данных. Их обеспечение возможно за счет принципа транзакционности при публикации данных, а также подтверждением транзакций с помощью использования блокчейна. На рис. 1 представлена архитектура системы управления данными, в основании которой лежат механизмы просмотра и управления результирующей онтологией, основанных на принципах Semantic Web и децентрализованном доступе к данным. При этом пользователи имеют возможность управлять онтологией в рамках в зависимости от предоставленных ролей, в том числе вносить собственные изменения в онтологию. Обеспечение и безопасность данных осуществляется с помощью реализованного контроля версий и ветвления по аналогии с системами контроля версий. Это позволяет вести журнал изменений, что в совокупности с децентрализацией позволяет обеспечить устойчивую безопасность, надежность и достоверность результирующих данных в онтологии. Основываясь на наработках из рассмотренных трудов, при реализации были учтены ошибки и рассмотрен успешный опыт авторов по теме работы. В основе инструмента управления онтологией используется блокчейн-платформа Ethereum, как и в подавляющем большинстве существующих решений для интеграции технологий Semantic Web с блокчейном. Такой выбор зачастую связан с тем, что решения полагаются на использование смарт-контрактов Ethereum.

Разрабатываемая система создана как обособленный центр управления и обработки данными, но при этом возможно внедрение инструмента в комплексную систему или сеть по контролю за данными в рамках концепции Semantic Web. Система управления семантическими данными состоит из следующих компонентов:

− точкой входа для пользователя при использовании инструмента является пользовательский интерфейс (User UI), с помощью которого осуществляется взаимодействие с приложением с помощью запросов;

− пользовательские запросы предполагают использование SPARQL, являющимся языком запросов, а также протоколом для передачи этих запросов и ответов на них: для взаимодействия с пользовательским интерфейсом бэкенд имеет конечную точку связи для SPARQL (SPARQL Endpoint);

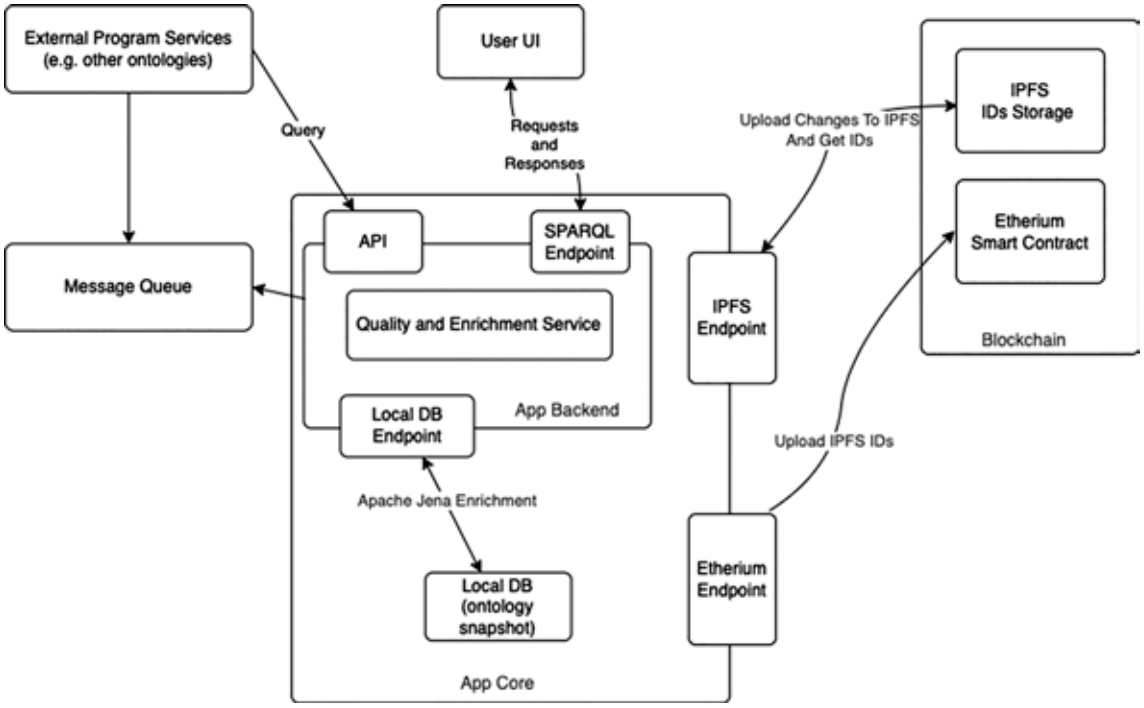

Рис. 1. Архитектура системы управления данными на основе Ethereum

− интеграционной точкой входа для взаимодействия с системой является интерфейс, взаимодействующий с внешними программными сервисами (API и External Program Services): вариантом стороннего сервиса может быть любой программный инструмент, предоставляющий посредством запросов данные, например, из сторонней онтологии;

− внешние сервисы могут взаимодействовать с очередями сообщений (Message Queue), которые служат способом взаимодействия с внешними системами, буферизуют входящие данные и обрабатывают их контролируемо, что обеспечивает стандартизированный способ интеграции данных других систем и выполнение проверки и преобразования данных;

− основой логики бэкенда приложения является сервис обеспечения качества и обогащения данных (Quality and Enrichment Service): сервис занимается выполнением операций CRUD, обеспечивает функциональность проверки консистентности и согласованности данных, а также проведения восстановления и обогащения данных; для реализации этих функций необходимы дополнительные проверки корректности изменений, такие как проверка синтаксиса запроса и согласованности онтологии после выполнения запроса, а также разработка и применение правил к отношениям в результирующих онтологиях;

− для возможности обогащения данных по аналогии с принципом MDM была воссоздана концепция правил качества, применяемых к данным: правилом качества является заранее заданный алгоритм действий на основе функции, который срабатывает при заданных конкретных данных условиях, в результате чего изменяет поступившие данные;

− приложение с помощью точки связи (Local DB Endpoint) взаимодействует с локальной базой данных, в которой хранится снимок текущей онтологии (Local DB) в тройной базе данных Apache Jena (TDB), которая поддерживает базы RDF, запросы SPARQL и повторное использование для проверки согласованности онтологии.

Использование семантических технологий (RDF и OWL) позволяет управлять данными в структурированном виде, обеспечивая их согласованность, точность и полноту. Семантические данные могут использоваться для автоматической идентификации и разрешения конфликтов данных, а также для обнаружения новых отношений между элементами данных, что улучшает качество и упрощает использование данных. Для работы с семантическими данными используется моментальный снимок локальной базы данных с тройной базой данных Apache Jena (TDB) и библиотекой Apache Jena для проверки согласованности онтологий. Использование моментального

снимка локальной базы данных повышает производительность системы и позволяет проверять данные без необходимости доступа к блокчейну Ethereum, что улучшает масштабируемость и снижает нагрузку на сеть Ethereum. При этом для реестра блокчейна используется оптимизированная стандартная архитектура, где блоки связываются хэшем родительского блока в хронологическом порядке. Блок транзакций состоит из заголовка и тела, где метаданные хранятся в заголовке, а транзакции в теле. Весь блокчейн-реестр представляет собой список цепочек блоков, где каждая цепочка состоит из последовательных блоков, связанных между собой хэш-связью. Используемая архитектура представлена на рис. 2.

На рис. 1 также выделено ядро приложения (App Core), которое состоит из бэкенда (App Backend) и локальной базы данных (Local DB), которые взаимодействуют между собой по описанным выше сценариям. Ядро приложения имеет две точки связи для взаимодействия с блокчейном: IPFS Endpoint и Ethereum Endpoint. Каждое произведенное изменение записывается в блокчейн, но, поскольку хранение больших документов в самом блокчейне неэффективно и дорого, данные об изменениях хранятся в сети IPFS, позволяющей разбивать большие файлы на более мелкие фрагменты, которые можно хранить и извлекать из разных узлов в сети для снижения нагрузки на блокчейн и повышения его масштабируемости за счет уменьшения объема хранимых данных. Изменения загружаются с помощью взаимодействия IPFS Endpoint приложения и хранилища данных (IPFS IDs Storage), в результате чего приложение сохраняет идентификатор содержимого файла, хранящегося в сети IPFS, в блокчейне Ethereum. Имея идентификаторы для данных, приложение связывается с сетью Ethereum для взаимодействия со смарт-контрактом (Ethereum Smart Contract) и загрузкой идентификаторов IPFS в блокчейн. Кроме того, IPFS интегрирован с Ethereum с помощью смарт-контрактов, что позволяет создавать децентрализованные приложения (dApps), которые могут получать доступ к файлам, хранящимся в IPFS, и управлять ими. Такой подход позволяет дополнительно использовать интеграционные механизмы, управлять метаданными файлов и контролировать доступ, в то время как сами файлы хранятся в IPFS. Использование смарт-контрактов также необходимо для управления жизненным циклом основных данных: от создания до проверки, обновления и удаления, а также для обеспечения соблюдения правил проверки данных, гарантируя точность и согласованность данных, хранящихся в блокчейне.

Для разрабатываемой системы использована платформа Ethereum 2.0. Ethereum в том виде, в каком он существовал в первой вариации, имеет ограничения, когда речь идет об обработке больших объемов данных и высокой пропускной способности транзакций, что подтверждают авторы работы [7]. Ethereum 1.0 использует механизм Proof of Work (PoW), который ограничивает количество обрабатываемых транзакций. Ethereum 2.0 использует механизм Proof of Stake (PoS), который обеспечивает более высокую пропускную способность для транзакций. В Ethereum 2.0 каждый участник, управляющий узлом, будет вознагражден за свой вклад в поддержание и улучшение системы данных.

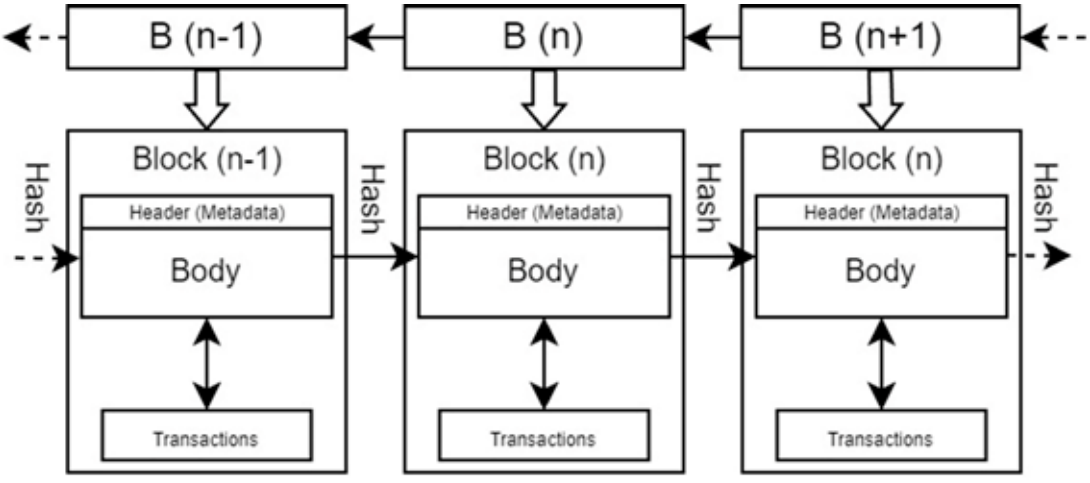

Рис. 2. Архитектура используемого стандартного блокчейна

Узлы, которые выполняют свои функции безупречно, будут награждены дополнительными эфирами, а узлы, которые не выполняют своих задач правильно, будут наказаны потерей своих эфиров. Более крупные участники будут иметь больше шансов быть выбранными для выполнения задач.

Результаты исследования и их обсуждение

Разработанная архитектура системы управления семантическими данными направлена на повышение масштабируемости и надежности MDM-систем. Для этого реализовано хранение файлов вне сети, что снижает нагрузку на Ethereum. Система предоставляет возможность обработки входящих данных и управления основными данными. Она также обеспечивает проверку данных с использованием семантических технологий и безопасность, благодаря использованию смарт-контрактов и децентрализованной, защищенной от несанкционированного доступа природе блокчейн-платформы. Таким образом, достигнута цель работы, в результате чего предложен архитектурный подход к решению известных проблем MDM-платформ: обеспечение качества данных, безопасности и масштабируемости системы; интеграция с внешними системами; обеспечение производительности. Для решения существующих проблем при проектировании системы были изучены существующие решения, и на основе этого были сформулированы ключевые особенности предложенной архитектуры:

− Консенсус и надежность: благодаря использованию технологии блокчейна, система позволяет обеспечивать надежность и согласованность семантических данных, что уменьшает возможность ошибок и сокращает время на проверку, оптимизируя бизнес-процессы.

− Управление доступом: организация управления доступом на уровне узлов блокчейна позволяет обеспечить защиту данных и предотвратить несанкционированный доступ.

− Надежная архитектура: система построена на платформе Ethereum 2.0, которая гарантирует интероперабельность, сохранность и защиту данных, а также позволяет децентрализованно хранить большой объем основных данных с использованием протокола IPFS. Архитектура системы разработана с целью обеспечения гибкости и интеграции ее с другими системами и сетями, что позволяет использовать ее в широком спектре областей знаний.

Заключение

Технология блокчейн является мощным инструментом для повышения безопасности и целостности данных в системах управления основными данными MDM. Однако внедрение этой технологии требует пересмотра архитектурного подхода к существующим системам. Разработанная в рамках работы архитектура системы призвана сохранить преимущества классических MDM-систем, но при этом предусматривает централизованное управление распределенными семантическими данными, обеспечивая масштабируемость и безопасность за счет внедрения блокчейна. Проектируемая система предназначена для создания узла по управлению и обработке данных с использованием семантических технологий, таких как RDF и SPARQL, а также блокчейн-технологии, платформы Ethereum и протокола IPFS. Целью практического применения блокчейн-MDM системы является улучшение процесса управления данными в производственных компаниях, которые заинтересованы в эффективном контроле своих поставщиков и контрактов. Используемый стек технологий прежде всего призван решить проблемы управления большими объемами данных, обеспечить гарантии безопасности и прозрачности данных, обеспечить интеграцию с внешними системами, обогатить и валидировать данные, а также предоставить возможность хранения файлов. Кроме того, благодаря совместному использованию технологии блокчейн и MDM-подхода, возможно устранение проблемы дублирования данных, что позволяет повысить оперативность и экономическую эффективность бизнеса. В целях решения проблем, связанных с возможным замедлением обработки больших объемов данных и масштабируемостью, в системе используются следующие технологии: алгоритм консенсуса Proof of Stake, технология шардинга и разделение хранимых данных с помощью протокола IPFS, что позволяет снизить нагрузку на систему. Возникающую в процессе внедрения блокчейна проблему стандартизации способен решить используемый в системе API и очереди сообщений.

Описанный архитектурный подход также призван решить проблему отсутствия исследований способов внедрения блокчейна для концепции MDM с использованием семантических данных. Разработка архитектуры, основанной на блокчейн-технологиях, имеет большой потенциал и может быть использована программными архитекторами и разработчиками в различных сферах де-

ятельности, где требуется высокий уровень безопасности и целостности данных, таких как финансы, здравоохранение, государственное управление и логистика. Для дальнейшей оценки практической эффективности технологий, описанных в статье, необходимо проведение дополнительных исследований, направленных на развитие и оптимизацию блокчейн-приложения для управления семантическими данными. Тем не менее результаты, полученные в ходе данного исследования, подтверждают значительный потенциал использования технологии блокчейн для управления данными.

Список литературы

1. Total data volume worldwide 2010–2025 / Statista. [Электронный ресурс]. URL: https://www.statista.com/statistics/871513/worldwide-data-created/ (дата обращения: 17.02.2023).

2. The DAMA Guide to the Data Management Body of Knowledge First Edition | Diego Fernandez Ayala – Academia.edu. [Электронный ресурс]. URL: https://www.academia.edu/19992490/The_DAMA_Guide_to_the_Data_Management_Body_of_Knowledge_First_Edition/ (дата обращения: 27.02.2023).

3. Master Data Management (MDM) Solutions Reviews 2023 / Gartner Peer Insights. [Электронный ресурс]. URL: https://www.gartner.com/reviews/market/master-data-management-solutions/ (дата обращения: 11.02.2023).

4. Zhang J., Wang F. Digital asset management system architecture based on blockchain for power grid big data // arXiv: Signal Processing. 2018. Vol. 16, Is. 8. P. 1–7.

5. Wen L., Zhang L. Application of Blockchain Technology in Data Management: Advantages and Solutions // Big Scientific Data Management. 2019. P. 239–254.

6. Wei Q., Li B., Chang W., Jia Z., Shen Z., Shao Z. A Survey of Blockchain Data Management Systems // ACM Transactions on Embedded Computing Systems. 2022. Vol. 21, Is. 25. P. 1–28.

7. Knez T., Gašperlin D., Bajec M., Žitnik S. Blockchain-Based Transaction Manager for Ontology Databases // Informatica. 2022. Vol. 33, Is. 2. P. 343–364.