УДК 004.62

РЕАЛИЗАЦИЯ ПРИНЦИПОВ ТЕХНОЛОГИИ БЛОКЧЕЙН ДЛЯ ОРГАНИЗАЦИИ БЕЗОПАСНОГО ОБМЕНА ДАННЫМИ В СЕТИ.

Петрин Д.А.[1], Белов Ю.С.[1]

[1] Калужский филиал ФГБОУ ВО «Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)», Калуга, e-mail: fn1-kf@mail.ru;

В данной статье был произведен обзор новой полностью распределенной платформы для обмена данными TSAR (Trustless data ShARing platform) как одной из возможностей применения технологии блокчейн (blockchain). В частности, была представлена архитектура платформы TSAR, механизмы публикации, поиска и обмена данными, а также механизм обеспечения подлинности данных; рассмотрены основные модули, их предназначение и особенности реализации. Также рассматривались две модели использования данных – с ограниченным и неограниченным доступом к данным. Для каждой из моделей приведена соответствующая схема рабочего процесса обмена данными. Были обозначены концепции децентрализованного хранения и защиты данных в платформе TSAR. Даётся обоснование применения метаданных как средства описания исходных данных, загружаемых на данную платформу. Описывается процесс проверки записи данных, поступившей в ответ на запрос пользователя к соответствующему URL адресу. Рассматривается схема построения набора ключевых слов, который используется для поиска данных, а также процесс сопоставления его с тегами метаданных. Делается предположение о дальнейших перспективах внедрения и использования платформы TSAR, а также ее значимости для конечных пользователей.

Ключевые слова: блокчейн, TSAR, метаданные, интерфейс TSAR, защита и неизменяемость данных.

IMPLEMENTATION OF THE BLOCKCHAIN TECHONOLOGY PRINCIPLES FOR THE ORGANIZATION OF SECURE DATA EXCHANGE IN THE NETWORK.

Petrin D.A.[1], Belov Y.S.[1]

[1]Moscow State Technical University n.a. Bauman (National Research University), Kaluga Branch, Kaluga, e-mail: fn1kf@mail.ru

This article was reviewed the new, fully distributed data exchange platform TSAR (Trustless data ShARing platform) as one of the opportunities for applying blockchain technology. In particular, the architecture of the TSAR platform was presented, the mechanisms for publishing, data retrieval and exchange, as well as the mechanism for ensuring the authenticity of data; the main modules were considered, their purpose and specifics of the implementation. Also there were considered two models use of data - with limited and unlimited access to the data. For each model is given the corresponding scheme of the workflow of data exchange. The concepts of decentralized storage and data protection in the TSAR platform were outlined. The justification for using of metadata as a means of describing the source data, loaded on this platform, is given. The process of verifying data recording is described, that were received in response to the user's request for the corresponding URL. The scheme for constructing a set of keywords, that is used to find data, as well as the process of matching it with metadata tags, is considering. It is assumed that there are further prospects for the introduction and use of the TSAR platform, as well as its significance for end users.

Keywords: blockchain, TSAR, metadata, TSAR interface, data protection and invariability.

Введение. В настоящее время сеть Интернет имеет очень широкое распространение, занимая крайне важное место в нашей деятельности. Это один из главных инструментов для сотрудничества и взаимодействия миллионов людей. Люди постоянно публикуют в сети какие-либо данные, предоставляя к ним доступ другим. Объем данных в Интернете растет с каждым днём. Для их обработки и анализа существует большое количество различных сервисов в частности облачных [1,2]. Но вместе с тем растет и количество угроз, которые могут нанести значительный ущерб владельцам данных и тем, кто их использует. Кроме

того, не всегда можно гарантировать достоверность данных и их источника. К тому же при размещении данных в сети необходимо воспользоваться какой-либо платформой или сервисом для обмена данными. С связи с этим так же встает вопрос о “доверии” этому сервису.

Именно поэтому безопасность и достоверность данных являются ключевыми вопросами, с которыми приходится сталкиваться в процессе информационного обмена в сети. Существуют различные технологии, обеспечивающие надежность данных. Одной из наиболее перспективных является технология блокчейн.

Технология блокчейн. Блокчейн представляет собой последовательную цепочку связанных блоков, содержащих информацию. Копии цепочек реплицируются на несколько независимых компьютеров [4]. Блок цепочки состоит из заголовка и списка транзакций, в виде которых и представлена информация. Заголовок содержит хеш, хеш предыдущего блока[9], хеш транзакции и дополнительную служебную информацию. Каждый блок содержит информацию о предыдущем блоке [10], поэтому возможно осуществить проверку блоков. Все блоки выстроены в одну цепочку. Таким образом, цепочка блоков осуществляет хранение информации.

Блокчейн также называют технологией распределенного реестра учета. Данная технология включает в себя следующие компоненты:

1. одноранговые сети;
2. распределенное хранение данных;
3. криптографическую защиту; [3]

Реализация принципов блокчейн обеспечивает:

- высокую надежность системы (принципы децентрализации и распределенности);
- возможность доступа к данным системы только для зарегистрированных пользователей (принципы безопасности и защищенности);
- доступность информации о проведении транзакций всем пользователям того или иного сервиса. При этом детали транзакций могут быть закрыты от публичного доступа специальным шифрованием.
- возможность осуществления взаимодействия без участия посредников, подлинность транзакций в системе проверяют непосредственно ее участники;
- невозможность изменения записанной информации. Новые блоки в цепочке создаются постоянно, каждый вновь созданный блок содержит группу накопившихся за последнее время и упорядоченных записей (транзакций). После проверки и согласования с участниками сети блок присоединяется к концу цепочки. Последующее его видоизменение невозможно.[3]

Именно на этих принципах и строится разработанная в Китае новая полностью распределённая платформа для обмена данными TSAR (Trustless data ShARing platform) [7], которая и будет рассмотрена в данной статье.

Архитектура TSAR. Платформа TSAR имеет архитектуру, которая отражена на рис. 1.

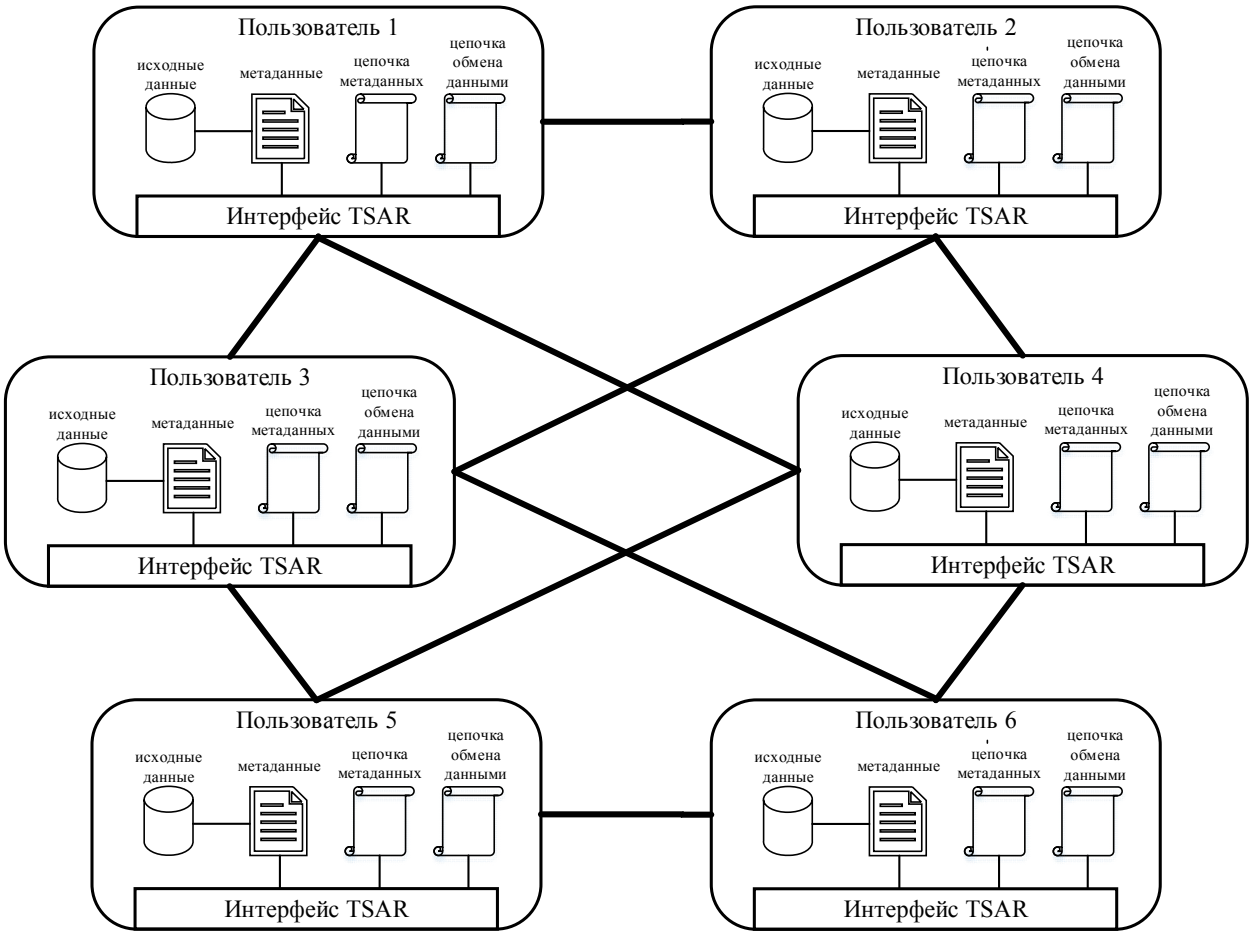

Рис. 1 Архитектура платформы TSAR

Каждый пользователь TSAR использует пять локальных компонентов: исходные данные, метаданные, цепочка метаданных, цепочка обмена данными и интерфейс TSAR. Последний предназначен для выполнения трёх сетевых функций: поиск, публикацию и обмен данными. Если пользователь владеет некоторыми данными, которые будут использоваться совместно, ему необходимо уведомить других пользователей в сети TSAR, что имеется часть недавно опубликованных данных.

Процесс публикации данных связан с компонентами исходных данных, метаданными и цепочкой метаданных. Исходные данные, хранящиеся локально, преобразуются в метаданные, а метаданные добавляются в цепочку метаданных, которая доступна всем пользователям сети TSAR. Цепочка метаданных представляет собой блокчейн [8]. Она хранит метаданные в виде транзакций.

Функция поиска данных требуется в том случае, если пользователь хочет найти данные по некоторым ключевым словам.

Если пользователь хочет получить определенные данные, возникает необходимость в функции обмена данными. Рассмотрим подробнее каждую из этих трёх функций.

Публикация данных. Процесс публикации данных в TSAR можно разделить на 3 шага:

1. упаковка исходных данных в запись с соответствующей сигнатурой
2. трансляция и проверка записи данных
3. синхронизация цепочки метаданных [8]

Входными данными процедуры публикации являются данные пользователя. Они могут иметь большой размер, измеряемый гигабайтами или даже терабайтами.

Если исходные данные публикуются напрямую, почти невозможно гарантировать их авторство. Кроме того, это создает огромную нагрузку на сеть. С этой целью в TSAR определён специальный тип данных - метаданные - для описания и публикации исходных данных [8]. Метаданные содержат схему данных, набор ключевых слов, небольшое количество выборочных данных, время получения и размер данных. Формат метаданных определяется так, чтобы полностью описать исходные данные и обеспечить высокопроизводительный поиск. Размер метаданных составляет около нескольких сотен килобайт. По сравнению с исходными данными огромного размера метаданные значительно уменьшают нагрузку на сеть. После преобразования пользовательских данных в метаданные они публикуются на HTTP-сервере. Таким образом, каждый в сети может просматривать метаданные через соответствующий URL. Кроме того, опубликованные метаданные не могут быть изменены другими.

Однако, существует еще две проблемы. Первая заключается в том, как осведомить других пользователей в сети о недавно опубликованных данных. Вторая проблема - как гарантировать тот факт, что метаданные не модифицируются на сервере. Для решения этих двух проблем TSAR использует механизм цепочки метаданных для децентрализованной регистрации публикаций данных.

После того, как пользователь генерирует URL-адрес для отображения метаданных, с помощью записи, которая содержит идентификатор пользователя, контрольную сумму исходных и метаданных, а также URL-адрес, создаются сами метаданные. Запись данных зашифровывается с использованием личного ключа пользователя и передается по всей сети с использованием интерфейса TSAR [8].

Когда пользователь получает запись данных, она проверяется следующим образом:

1. идентификация пользователя, использующего сигнатуру в записи;
2. получение ключа от пользователя, опубликовавшего данные;

3. использование открытого ключа для дешифрования записи данных;
4. проверка соответствия формата данных;
5. проверка соответствия подписи издателю;
6. проверка доступности URL-адреса, содержащегося в записи данных;
7. проверка определимости метаданных по URL-адресу;
8. сравнение контрольной суммы метаданных с контрольной суммой, содержащейся в записи данных [8];

Условия проверяются один за другим. Если очередное условие не удовлетворено, запись данных будет прервана. Если запись данных проверена пользователем, она будет помещена в локальный пул метаданных. Однако попадание в пул не означает, что запись данных будет опубликована. С определённой фиксированной частотой метаданные из пула будут упакованы в цепочку метаданных. Если запись данных упакована в цепочку, она публикуется. Каждый узел в сети будет синхронизировать цепочку метаданных.

Поиск данных. Следующая функция для рассмотрения – поиск данных. Для поиска платформе TSAR не нужен центральный сервер. Как упоминалось ранее, все метаданные будут публиковаться в цепочке метаданных, и на текущий запрос будет отвечать собственно клиент системы в соответствии с цепочкой метаданных. Процедура поиска данных включает следующие шаги:

1. Синхронизация цепочек метаданных клиента
2. Расширение слов и поиск сходства
3. Извлечение данных и отображение результатов. [8]

Для пользователей, которые вошли в систему, будет выполнена синхронизация цепочки метаданных. Однако новый пользователь или же пользователь, который хочет найти нужные ему данные без публикации данных, не синхронизирует цепочку метаданных на своем клиенте. Процедура синхронизации цепочки метаданных пользователя такая же, как и при публикации данных. При этом синхронизация общей цепочки может потребовать существенных временных затрат.

В каждом блоке цепочки метаданных есть некоторые ключевые слова (теги) для описания семантики данных. Процесс извлечения данных заключается в обратном отслеживании метаданных и сопоставлении запроса этим ключевым словам.

Чаще всего запрос пользователя короткий и содержит очень мало информации. Поэтому только лишь слова запроса не могут дать нужные результаты. Существует общий метод решения этой проблемы - расширить слова запроса с помощью дополнительных знаний. Такие знания могут быть предоставлены с помощью специальных сервисов. В TSAR для этой цели используется сервис WordNet [8]. WordNet — это электронный словарь для

английского языка, разработанный в Принстонском университете [7]. В нем содержатся синонимы, антонимы, определения слов и т. д. В системе TSAR для каждого слова запроса извлекаются синонимы этого слова с помощью WordNet, и затем все вместе используются в качестве слов запроса. Чтобы избежать различных форм слова, например, «дом» и «дома», TSAR использует платформу NLTK (Natural Language Toolkit) [6], чтобы получить основу каждого ключевого слова [8]. В итоге получается набор ключевых слов, который TSAR использует для поиска связанных с ними данных.

Метаданные в блоках цепочки содержат очень мало тегов. Это сделано для того, чтобы в одном блоке могло поместиться больше метаданных. При прямом сравнении тегов с ключевыми словами запроса сложно найти семантическую взаимосвязь. Поэтому в TSAR теги метаданных расширяются с использованием вышеописанного для слов запроса метода. После расширения получаются финальные теги. Затем используется коэффициент Жаккара [5] для вычисления схожести между ключевыми словами запроса и тегами метаданных [8].

Извлечение данных. Извлечение данных в TSAR аналогично извлечению данных в поисковой системе. В результате извлечения возвращается список данных, которые семантически подобны запросу пользователя. Для ускорения процесса поиска и удовлетворения принципу локальности, для каждого пользователя создается кэш записей последних результатов поиска [8]. При этом данным назначаются ранги. Получив предварительные семантически схожие данные, ранги результатов перезаписываются. В конечном итоге данные с меньшими номерами рангов становятся более востребованными. Здесь же действует следующий принцип: чем раньше опубликованы данные, тем менее важными они будут [8].

Цель распределенной платформы обмена данными заключается в обеспечении достоверности данных и надежности обмена. В TSAR для реализации этих требований предусмотрен модуль обмена данными [8]. Для обеспечения управляемости совместного использования данных с помощью этого модуля можно устанавливать различные разрешения для пользователей, запрашивающих данные, через их идентификаторы. Использование данных делится на две модели в соответствии с этим идентификатором.

Модели использования данных. Первая модель – неограниченное использование данных. Пользователи с неограниченным доступом к данным осуществляют поиск через цепочку метаданных, извлекают данные и отправляют владельцу запрос на передачу. После того как запрос был одобрен владельцем, предполагаемые данные могут быть отправлены запрашивающему. Чтобы гарантировать достоверность полученных данных, в интерфейсе TSAR задействована функция проверки [8].

Вторая модель – ограниченное использование данных. Пользователи с ограниченным доступом к данным не могут напрямую обращаться к исходным данным владельца, но могут получить желаемый результат обработки, отправив соответствующий запрос владельцу данных.

Защита данных в TSAR осуществляется за счет использования цепочки обмена данными [8]. В рамках этого механизма система хранит запись о совместном использовании данных в цепочке. Эти записи можно использовать для отслеживания операций совместного использования данных.

Архитектура цепочки обмена данными следует концепциям блокчейн. Запись о совместном использовании данных содержит следующую информацию:

1. владелец данных, запрашивающий данные и их сигнатуры
2. указатель метаданных, код подтверждения и URL-адрес данных
3. время осуществления совместного использования и разрешения
4. дополнительные сведения [8]

Для пользователей с неограниченным доступом к данным рабочий процесс показан на рис. 2. Пользователь B публикует в сети контракт на запрос и передачу данных A1. После получения запроса пользователь A проверяет и отправляет зашифрованные данные A1 пользователю B. При этом A помещает ключ шифрования данных A1 вместе с подписанным контрактом в блок цепочки обмена данными. После того, как этот блок будет аутентифицирован, B сможет получить ключ шифрования и расшифровать данные A1.

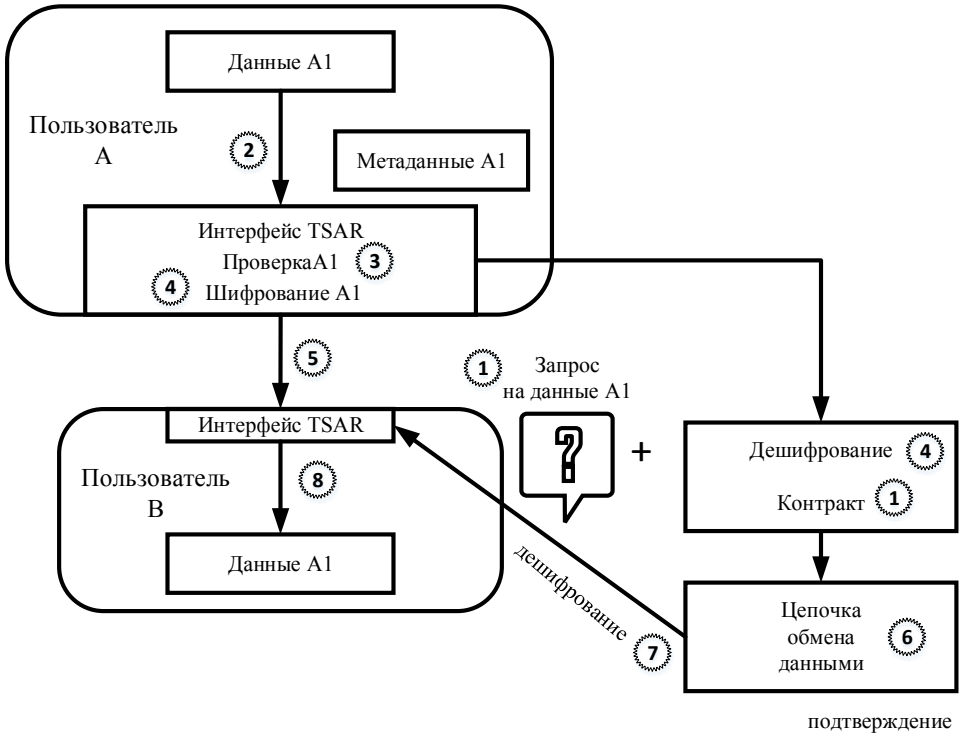

Рис.2 Рабочий процесс для модели неограниченного использования данных

Для пользователей с ограниченным доступом к данным схема рабочего процесса представлена на рис.3. Пользователь B публикует в сети запрос и контракт на получение данных A1, а затем отправляет код и результаты тестовой выборки данных. После этого пользователь A обрабатывает данные A1 с этим кодом и передает результат обработки на B в зашифрованном виде через интерфейс TSAR. В то же время пользователь A упаковывает ключ шифрования вместе с подписанным контрактом в блок цепочки обмена данными. Если блок с этой записью будет аутентифицирован, пользователь B с помощью ключа шифрования сможет расшифровать данные A1.

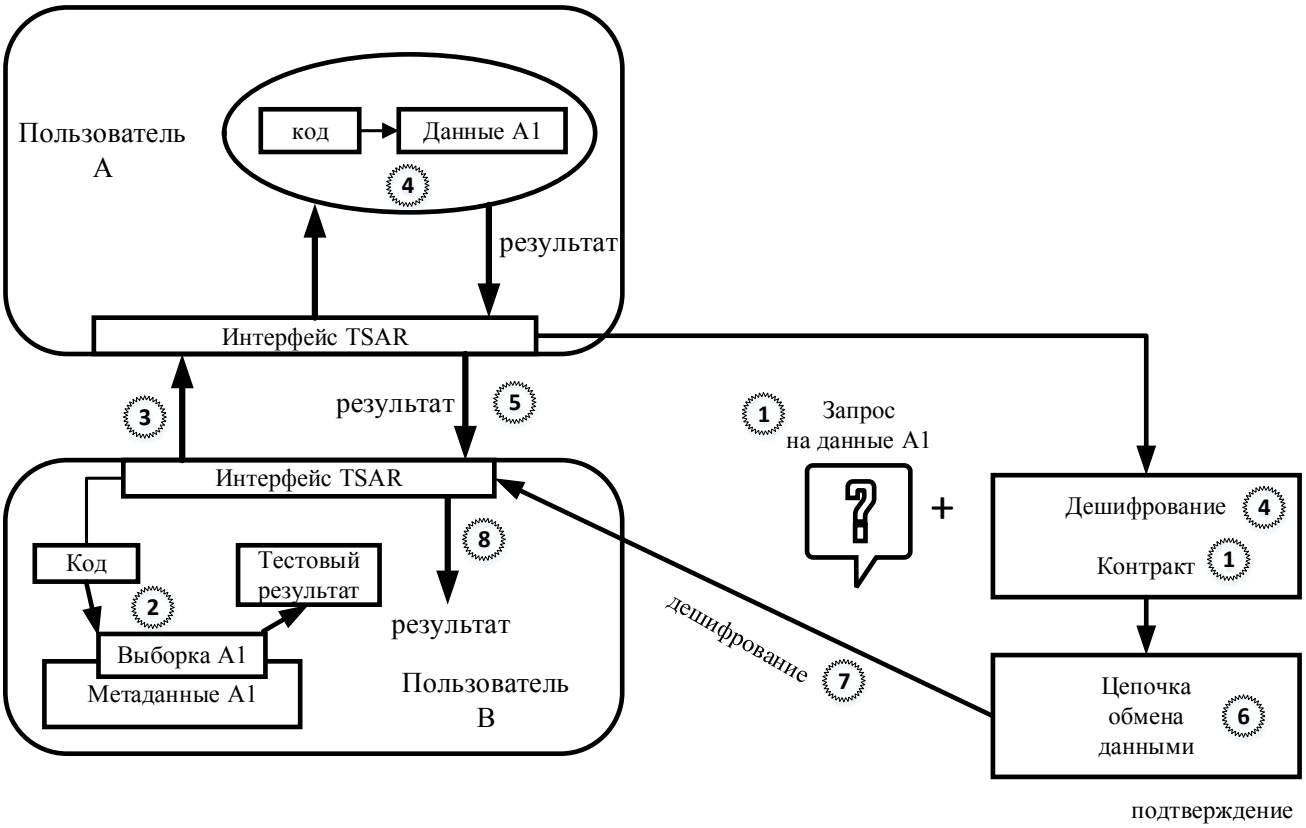

Рис.3 Рабочий процесс для модели ограниченного использования данных

Заключение. В заключение стоит отметить, что платформа TSAR на данный момент находится на этапе тестирования и оценки производительности. Но уже сейчас можно сказать, что данный проект имеет хорошие предпосылки для будущего внедрения и использования. Ведь в традиционных платформах пользователи должны загружать свои данные для обмена на централизованный сервер (или сервера), который будет предоставлять к ним доступ. Таким образом, пользователи становятся сильно зависимыми от надежности владельца платформы. Ведь никто полностью не может гарантировать того, что загружаемые данные не модифицируются и имеют должный уровень защищенности. Именно поэтому при проектировании и разработке платформы TSAR ключевым решением стал отказ от

концепции централизованного сервера для хранения данных. Вместо этого в TSAR используется цепочка метаданных и цепочка обмена данными, которые основаны на технологии блокчейн. Поэтому при должном уровне поддержки со стороны компании Huawei Technologes. Co. Ltd данный проект в ближайшее время может стать одной из самых перспективных разработок, которая позволит пользователям в будущем повысить безопасность своих данных в сети.

Список литературы

[1] Аксютина Е.М., Белов Ю.С. ИСПОЛЬЗОВАНИЕ ОБЛАЧНЫХ ТЕХНОЛОГИЙ ДЛЯ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ Электронный журнал: наука, техника и образование . 2016. № 2 (6). С. 67-73.

[2] Аксютина Е.М., Белов Ю.С. ОБЗОР АРХИТЕКТУР И МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ Электронный журнал: наука, техника и образование. 2016. № 1 (5). С. 134-141.

[3] Малявкина Л.И., Савина А.Г., Смагина И.В. КОНЦЕПТУАЛЬНЫЕ ОСНОВЫ ТЕХНОЛОГИИ BLOCKCHAIN И ПРОБЛЕМЫ ЕЕ ВНЕДРЕНИЯ В ЦИФРОВУЮ ЭКОНОМИКУ РОССИИ 2017,URL https://e.lanbook.com/reader/journalArticle/422204/#1

[4] Википедия, статья «Блокчейн» [Электронный ресурс]. – Режим доступа https://ru.wikipedia.org/wiki/Блокчейн#Реализация_в_системе_Биткойн (дата обращения 25.09.2018)

[5] Википедия, статья «Коэффициент Жаккара» [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Коэффициент_Жаккара (дата обращения 25.09.2018)

[6] Платформа Natural Language Toolkit - [Электронный ресурс]. – Режим доступа https://www.nltk.org (дата обращения 25.09.2018)

[7] Электронный словарь Wordnet - [Электронный ресурс]. – Режим доступа:https://wordnet.princeton.edu (дата обращения 25.09.2018)

[8] Hanqing Wu, Jiannong Cao, Shan Jiang, Ruosong Yang, Yanni Yang, Jianfei He. TSAR: a fully-distributed Trustless data ShARing platform. Conference: The Third IEEE Workshop on Smart Service Systems (SmartSys 2018), At Taormina, Sicily, Italy Available from: https://www.researchgate.net/publication/324820346_TSAR_a_fully-distributed_Trustless_data_ShARing_platform (дата обращения 27.09.2018).

[9] Suyash Gupta and Mohammad Sadoghi Department of Computer Science, University of California, Davis, Davis, CA, USA, 2018, Blockchain Transaction Processing URL https://www.researchgate.net/publication/325116198_Blockchain_Transaction_Processing

[10] Zibin Zheng, Shaoan Xie, Hongning Dai, Xiangping Chen, and Huaimin Wang An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends 2017 IEEE 6th International Congress on Big Data URL https://www.researchgate.net/publication/318131748_An_Overview_of_Blockchain_Technology_Architecture_Consensus_and_Future_Trends