Управление метаданными провенанса и правами доступа к данным в распределенных хранилищах на основе блокчейн-технологии и умных контрактов

В настоящее время осуществление крупных научных, инженерных и бизнес-проектов связано, как правило, с необходимостью хранения и обработки больших объемов данных. Это приводит к необходимости развивать новые, более экономичные и надежные, архитектуры и принципы работы информационных систем, в том числе систем хранения данных. Экстремальными вариантами архитектурных решений для таких систем являются полностью централизованные хранилища и хранилища на основе одноранговых P2P-сетей. Однако, во многих случаях такие решения оказываются неприемлемыми, например из-за их высокой стоимости или низкой надежности, а оптимальным является промежуточное между такими экстремальными вариантами решение. Для его осуществления организации — участники крупного проекта — объединяют свои локальные ресурсы хранения в единый распределенный пул и, при необходимости, дополнительно арендуют облачные ресурсы хранения, возможно у нескольких провайдеров. Особенно выгодным с экономической и технической точек зрения такое решение может оказаться в случае, когда появляется потребность в хранении больших объемов данных в течение ограниченного срока осуществления какой-либо проекта и в ситуации, когда проект объединяет многих организационно несвязанных между собой участников. В общем случае такой распределенный пул хранилищ образует динамически меняющуюся среду (по мере необходимости могут подключаться новые хранилища или отключаться ранее входившие в пул). Задача заключается в том, чтобы объединить все эти хранилища и данные в них в единую систему в динамически меняющейся среде, а также обеспечить реализацию взаимных политик доступа к данным участвующих сторон. Например, владелец файла с данными (пользователь, создавший эти данные или организация, которой они принадлежат) должен иметь возможность управлять правами доступа к нему для других пользователей. Это подразумевает наличие способов децентрализованного управления правами доступа к данным в такой динамически меняющейся среде, обеспечения консенсуса участвующих сторон относительно содержания и порядка операций с данными и обеспечения надежной неизменяемой записи истории совершенных операций, то есть метаданных провенанса (МДП), для разбора и разрешения возможных коллизий между участниками проекта, а также владельцами хранилищ. Коллизии могут быть связаны с вопросами приоритета при получении результатов обработки данных, использования результатов, нарушении прав доступа и т. п. Другими словами, необходимо обеспечить инструментарий для поддержки осуществления бизнес-процессов хранения и обмена данными в распределенной среде и при наличии административно несвязанных или слабо связанных организаций, участвующих в совместных проектах, или просто обменивающихся данными на определенных условиях.

Необходимо отметить, что хотя за последние годы был осуществлен целый ряд проектов по созданию систем для поддержки и управления метаданными, включая провенанс данных, но подавляющее большинство реализованных решений являются централизованными (см., например, обзор и ссылки в нем), что плохо соответствует случаю использования распределенной динамически меняющейся среды. С другой стороны, в последнее время в разных прикладных областях большую популярность приобрели — благодаря наличию ряда важных преимуществ — распределенные реестры на основе технологии блокчейна. В самое

последнее время на основе блокчейна появились разработки и для систем управления МДП (SmartProvenance, Provchain). Однако, они рассчитаны на работу с одним хранилищем, не решают проблему обеспечения бизнес-процесса обмена данными между административно различными организациями и управления доступом к данным.

В рамках данного проекта предложен новый подход к построению системы управления метаданными провенанса и правами доступа к данным, основанный на интеграции блокчейн-технологии, смарт-контрактов и управления данными на основе метаданных. Разработаны принципы и алгоритмы работы такой системы, названной **ProvHL** (Provenance HyperLedger), которая является отказоустойчивой, безопасной, надежной с точки зрения сохранности и защищенности записей метаданных провенанса от случайных или намеренных искажений. Исследованы вопросы оптимального выбора типа блокчейна для такой системы, а также выбора блокчейн-платформы. А именно, предложено использовать эксклюзивный (permissioned) тип блокчейна и блокчейн-платформу Hyperledger Fabric (HLF), на основе которой реализуется система ProvHL.

В настоящее время на базе НИИЯФ МГУ создан полигон, на котором развернут предварительный вариант прототипа ProvHL для реализации разработанных принципов и отработки алгоритмов работы системы. Создание системы ProvHL производственного уровня позволит существенно повысить качество и надежность результатов, получаемых на основе обработки и анализа данных в распределенной компьютерной среде.

Исследование выполняется за счет гранта Российского научного фонда (проект № 18-11-00075).

Managing provenance metadata and data access rights in distributed storage based on blockchain technology and smart contracts

Currently, the implementation of large scientific, engineering and business projects is associated, as a rule, with the need to store and process large amounts of data. This leads to development of new, more economical and reliable, architecture and operating principles of information systems, including storage systems. Extreme options for architectural solutions for the latter are fully centralized and fully decentralized (peer-to-peer, P2P) storages. However, often such solutions are unacceptable, in particular because of their high cost or low reliability. In many use cases, the intermediate solution between the fully centralized and fully decentralized ones may prove to be optimal [1]. To implement such a solution, organizations participating in a large project pool their local resources into a unified distributed storage and, if necessary, additionally rent cloud storage resources, possibly from several providers. Especially profitable from the economic and technical points of view, this solution can be in the case when there is a need to store large amounts of data during a limited period of the project implementation and in a situation where the project brings together many organizationally unrelated participants. In general, such a distributed storage pool forms a dynamically changing environment (the new storage can enter the pool and the other can leave it), and the local storages entering the pool can have different data management systems. The challenge is to combine all these storages and data in them into a single system in a dynamically changing environment, as well as ensure the implementation of reciprocal access policies to the data of the parties involved. For example, the owner of data (the user who created these data or the organization to which they belong) should be

able to manage access rights to them for other users. Another example is the ability of a cloud storage to grant access to data stored on it only to users from organizations that have paid for the provision of the storage services. This implies the availability of decentralized methods both for data access management in such a dynamically changing environment and for ensuring a reliable, immutable record of the history of committed transactions, that is, provenance metadata (PMD), for examination and resolving possible conflicts of project participants between themselves and with storage owners. Conflicts can be related to priority issues when obtaining the results of data processing, use of results, violation of access rights, etc.

In other words, it is necessary to provide tools to support the implementation of business processes of storage and exchange of scientific data in a distributed environment with administratively unrelated or loosely related organizations participating in joint projects or simply exchanging data on certain conditions. First of all, this requires a provenance metadata registry that is resistant to malicious changes as well as a method of ensuring consensus among participants in the business process about the content and order of transactions with data.

It should be noted that although a number of projects have been implemented in recent years to create systems for metadata storage and management, including the provenance of data, the vast majority of the implemented solutions are centralized, which is poorly suited to distributed dynamically changing environment, and the possibility of using metadata by organizationally unrelated research communities. On the other hand, in recent years, distributed registries based on blockchain technology have become very popular in various applied areas due to a number of important advantages. Most recently, on the basis of the blockchains, developments have also been appeared for the PMD management systems (SmartProvenance, Provchain). However, they are designed to work with one storage, do not solve the problem of providing business process for data exchange between administratively different organizations and data access management.

In the framework of this project we proposed a new approach to the construction of data management system and access rights in a distributed environment, based on the integration of blockchain technology, smart contracts and provenance metadata driven data management. Principles and operation algorithms have been developed for such a system, called **ProvHL** (Provenance HyperLedger), which is fault-tolerant, safe, reliable in terms of the safety and security of provenance metadata records from accidental or intentional distortion. The problems of the optimal choice of the type of blockchain for such a system, as well as the choice of blockchain platform have been investigated. Namely, it is proposed to use permissioned type of blockchain and the Hyperledger Fabric blockchain platform, on the basis of which the ProvHL system is implemented.

At present, a testbed has been created in SINP MSU, on which a preliminary version of the ProvHL prototype is deployed for implementation and elaboration of the principles and algorithms developed. In addition to simply maintaining the immutable ledger with provenance metadata the problem of distributed data access management including management of access rights is solved. The metadata is written to the blockchain beforehand, and data management systems refer to the blockchain and performs the transactions recorded there (metadata driven data management). Within the testbed, smart contracts is implemented to support basic operations with files (upload, download; copy to another storage) and with directories (create, delete, content listing). The deployment of a number of nontrivial endorsement policies of transactions with data by duly authorized parties within the distributed storage system was carried out. The connection to the network of remote storages (e.g., clouds) using http and ftp protocols is also simulated. Creating a production-level ProvHL system will improve the quality and reliability of the results, obtained on the basis of processing and analyzing data in distributed computer environments.

This work was funded by the Russian Science Foundation (grant No 18-11-00075).