Contribution ID: **192** Type: **Sectional reports**

A new approach to the development of provenance metadata management systems for large scientific experiments

Tuesday, 11 September 2018 14:15 (15 minutes)

Provenance metadata (PMD) contain key information that is necessary to determine the origin, authorship and quality of corresponding data, their proper storage, correct using, and for interpretation and confirmation of relevant scientific results. The need for PMD is especially essential when big data are jointly processed by several research teams, which is a very common practice in many scientific areas of late. This requires a wide and intensive exchange of data and programs for their processing and analysis, covering long periods of time, during which both the data sources and the algorithms for their processing may be modified.

Although a number of projects have been implemented in recent years to create management systems for such metadata, but the vast majority of implemented solutions are centralized, which is poorly suited to current trends of working in distributed environments, open data access models, and the use of metadata by organizationally unrelated or loosely coupled communities of researchers.

We propose to solve this problem by using a new approach to creating a distributed registry of provenance metadata based on blockchain technology and smart contracts. In this work, the functional requirements for the PMD management system were formulated. Based on these requirements, we investigated the problem of the optimal choice of the type of blockchain for such a system, as well as the optimal choice of consensus algorithm for records ordering within the blockchain without participation of third-party trusted bodies. The architecture and algorithms of the system operation, as well as its interaction with the distributed storage resources management systems, are proposed. Specific use cases for the PMD management system are considered. A number of existing blockchain platforms are considered and the most preferable one is selected.

The results of this work are of particular importance in the big data era, when a full analysis of the results of experiments is often not possible for one team, so that many independent teams take part in their analysis.

The suggested approach is currently under implementation in SINP MSU in the framework of the project supported by the Russian Science Foundation (grant No 18-11-00075).

Primary author: Dr DEMICHEV, Andrey (SINP MSU)

Co-author: Dr KRYUKOV, Alexander (SINP MSU)

Presenter: Dr DEMICHEV, Andrey (SINP MSU)

Session Classification: 10. Databases, Distributed Storage systems, Datalakes

Track Classification: 10. Databases, Distributed Storage systems, Datalakes