A multi-platform ecosystem for computing in Earth sciences

Vitaliy S. Eremenko[1], Vera V. Naumova[1]

[1]*Vernadsky State Geological Museum RAS, Moscow, Russia*

Abstract

Analysis of diverse data in geosciences requires access to various processing tools, including specialized software packages, proprietary algorithms, GIS systems, web services, etc. Such tools require from the user a certain level of skills to work with them, form a software environment, the availability of the necessary computing power, and sometimes significant time and financial costs.

With the development of information technology, more and more software products, including professional software packages for data processing, are provided to users in the format of various cloud services and platforms. In such systems, computation takes place on the side of the service provider and is accessed through a web browser. The emergence of such open-access services and platforms makes it possible to organize a single workspace for a researcher with the ability to analyze his own data using various processing methods, including tools traditionally used in earth sciences.

The report is devoted to the development of an approach for the integration of various tools for processing heterogeneous data with open access within a single multi-platform ecosystem. The software system developed based on the proposed approach is demonstrated. The report describes software modules that implement the functions of access to processing and analysis tools, as well as service modules for system administration, component monitoring and event logging. Services and processing platforms integrated into the ecosystem are considered, as well as scenarios for solving some geological problems.

Keywords

Web services, cloud services, processing of geological data.

1. Introduction

Since 2017, at the Vernadsky State Geological Museum RAS, an information and analytical environment is being developed to support scientific research in geology (http://geologyscience.ru) [1]. The environment provides a single point of access to various types of geological information on the territory of Russia, as well as a set of tools for processing and analyzing this data.

Processing and analyzing data in geology requires the application of a large number of different algorithms, processing procedures and corresponding software solutions. With the development of information technology, approaches to the organization of such processing have changed. Solving the problem of analyzing large amounts of data or tasks requiring the use of resource-intensive methods required the acceleration of software computations. To solve such problems, both parallel computing using supercomputers and distributed computing using

SDM-2021: All-Russian conference, August 24–27, 2021, Novosibirsk, Russia

vitaer@gmail.com (V. S. Eremenko)

a large number of computing devices combined into a single computing system, such as GRID systems, were used. However, access to such systems is limited.

The advent of cloud services has allowed users to gain easy access to the computing resources they are interested in. The use of external services instead of custom applications allows data processing on the equipment that is most suitable for the corresponding tasks. Thus, data processing is more efficient, and the user gets the opportunity to process data with the most current versions of algorithms, using a web interface for this, without the need to install, configure and maintain software for processing on his personal computer. Due to the continuous growth in the amount of geological data and tools for their analysis, it becomes necessary to organize a single workspace for a geologist-researcher with the ability to use open access tools for analyzing geological information available in the world.

The purpose of the computing block of the information and analytical environment is to provide researchers with access to geographically distributed services for processing and analyzing geological information (https://service.geologyscience.ru) [2, 3].

Basic functional requirements:

- ability to use external open information processing services in unlimited quantities;
- working with data provided by users and open information systems;
- cataloging external processing services.

2. Implementation

To organize a single mechanism for interacting with web services, the OGC Web Processing Service (WPS, http://www.opengeospatial.org/standards/wps) interface is used, developed by the international organization for standardization of the Open Geospatial Consortium. This interface is implemented on the basis of the HTTP protocol (HTTPS), is widely used in various scientific data processing systems [4]; a variety of software products support the operation with geoinformational services.

Basic requirements for this type of service:

- Permanent IP address (or domain name) and a port with external access via the Internet.
- Client program interface (API) and libraries implementing it in the Java language, or via the HTTP/HTTPS protocol (REST or SOAP);
- Ability to start the processing procedure with the specified parameters;
- Ability to obtain the result of processing in text or binary formats, including in the form of URL links;
- Ability to work with user data in one of the following ways:
    - reading remotely placed data by URL address;
    - temporary loading data to the computing node;
    - transferring data in binary format as a processing parameter.

Open source software package GeoServer (http://geoserver.org) was selected to create and host your own WPS processes. Thus, for each external service, a separate WPS process is created with the corresponding startup parameters. Using WPS as an access interface to remote services allows you to execute several computational processes sequentially. Thus, outcomes of one or several processes can be used as input parameters for another process, thereby providing the possibility to combine the results of processing different types of data.

Using a common access protocol to interact with cloud services that provide users with interactive access through a web interface is difficult due to the heterogeneity of the access protocols used in each platform. However, some common properties of platforms, such as the presence of centralized data storages and single authentication systems, make it possible to organize interaction by transferring data for analysis to the data storages corresponding to the platforms on behalf of a specific user. The implementation of such a mechanism has become possible when using web-application technology, which allows you to request permission from the user to access certain capabilities of the user account of various cloud service providers. This technology is supported by Microsoft, Google, Yandex, ESRI, etc. Before using the corresponding service, the user needs to upload data for analysis to his personal storage. The user is authorized on the website of the service provider, after which the application asks the user for permission to download and publish data to its storage. The user has the ability to move data for processing from one storage to another when choosing a cloud service from different providers.

Thus, we have proposed an approach and a technological solution for organizing a single data space for various cloud service providers.

A monitoring module has been developed to track the state of geographically distributed components of the ecosystem [5].

The following general types of tests are presented.

A. Checking the availability of a remote site.
B. Checking service performance at a remote site using the required communication protocol.
C. Checking for changes in the operation of the service based on test requests.

The module of the catalog of external services for processing and analyzing geological information is used as a data source for monitoring.

3. Computing capabilities

The developed software ecosystem provides access to the following services and processing platforms.

- Computing node "Multidimensional methods of data analysis", developed at the State Geological Museum of the Russian Academy of Sciences, which allows you to process tabular data by various methods of data analysis with setting their parameters and visualizing the results. The computational node includes such groups of methods as statistical analysis, regression analysis, factor analysis, clustering, machine learning, visualization methods, and others. Calculations are performed in the Python environment using well-known data processing packages: Scikit-learn, Pandas, Matlplotlib and others. The processing of incoming requests for processing is carried out by the Flask framework through the REST API using the task queue, implemented on the basis of the NoSQL Redis database. This architecture allows the processing of requests and heavy computations of large amounts of data to be separated, which provides fault tolerance and scalability of the node. In the future, it is planned to expand the number of data processing methods with the involvement of specialized processing packages for solving geological tasks.

— Petrological and geochemical data processing. An interactive database of methods for processing petrological and geochemical data has been developed at the Schmidt Institute of Physics of the Earth of RAS [6]. This system provides services for constructing spidergrams, histograms and classification diagrams; service for identification of minerals by their chemical composition; service for the interpretation of the composition of the mineral and decomposition into minals, etc. The interface for interaction with services is based on REST architecture.

— Structural analysis of publications. The Interdisciplinary Center for Mathematical and Computational Modeling (University of Warsaw, Poland) has developed a service for extracting metadata from scientific publications [7]. Metadata includes authors, affiliation, abstract, keywords, journal title, volume, year of issue, parsed bibliographic references, document section structure, section headings and paragraphs. The interface for interaction with services is built on the basis of REST architecture.

— Natural language processing. At the University of Sheffield, the GATE (General Architecture for Text Engineering) project has developed a number of services for processing text data for various languages [8]. For processing textual data in Russian, services are provided to determine the parts of speech of words, as well as to highlight named entities, such as names and surnames, names of organizations, geographical names, dates, monetary units, etc. The interface for interaction with services is based on REST architecture.

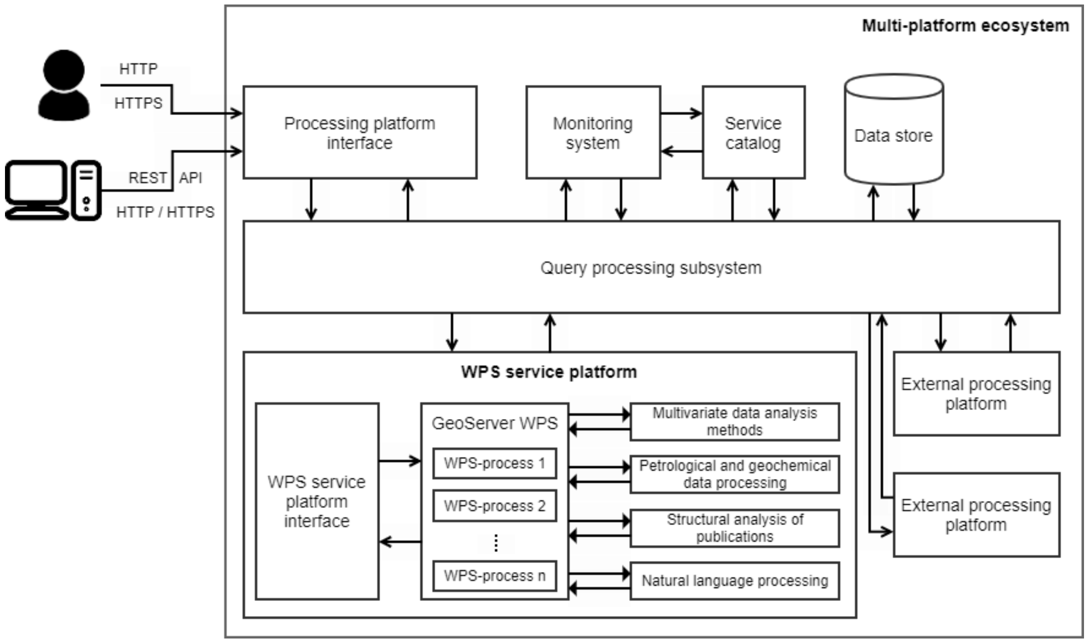

Multi-platform ecosystem
HTTP
HTTPS
REST API
HTTP / HTTPS
Processing platform interface
Monitoring system
Service catalog
Data store
Query processing subsystem
WPS service platform
WPS service platform interface
GeoServer WPS
WPS-process 1
WPS-process 2
WPS-process n
Multivariate data analysis methods
Petrological and geochemical data processing
Structural analysis of publications
Natural language processing
External processing platform
External processing platform

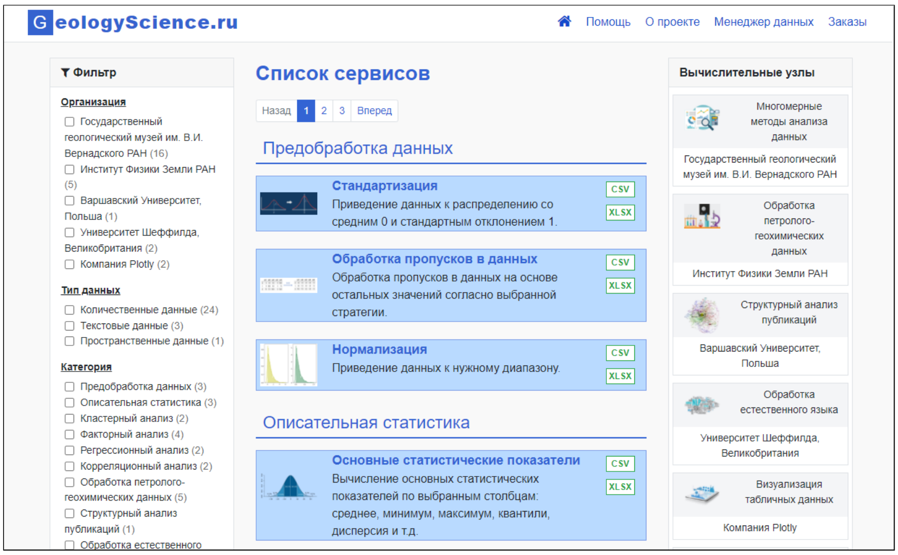
Figure 1: General functional diagram of a multi-platform ecosystem.

- Microsoft Office Online is a cloud-based platform that allows users to work with web versions of software products such as Word, Excel, PowerPoint, OneNote. Microsoft Excel is one of the key tools for working with tabular data in geology (https://www.office.com/launch/excel). It contains tools for viewing and editing tabular data, as well as a set of analytical functions and tools for building various types of charts.
- ArcGIS Online (https://www.arcgis.com) is a cloud-based mapping and analysis solution. The ArcGIS platform allows users to work with 2D and 3D data to explore and visualize it. One of the key features is the ability for multiple users to collaborate on the same data. The platform provides tools for creating web maps, 3D scenes and notebooks. Using ArcGIS Notebook allows you to access Python resources to perform analysis, automate workflows, and visualize data.
- Google Earth Engine is a satellite data analysis platform (https://earthengine.google.com). This platform allows the user to upload their own data or use data from the Earth Engine catalog for further processing in an interactive mode. The catalog contains data processing products for the Modis radiometer (Aqua, Terra satellites), Sentinel-1A, Sentinel-1B, Sentinel-2A, Sentinel-2B, Landsat 8, etc. creation, editing and launching using Javascript and Python programming languages. To work, the user needs a Google account. For analysis and processing, you can use data from the Google cloud storage.

The general functional diagram of the multi-platform ecosystem is shown in Figure 1.

4. Testing

Testing of the system functions of the processing platform took place with the automatic testing tools provided by the framework used in the development.



Фильтр

Организация

□ Государственный геологический музей им. В.И. Вернадского РАН (16)
□ Институт Физики Земли РАН (5)
□ Варшавский Университет, Польша (1)
□ Университет Шеффилда, Великобритания (2)
□ Компания Plotly (2)

Тип данных

□ Количественные данные (24)
□ Текстовые данные (3)
□ Пространственные данные (1)

Категория

□ Предобработка данных (3)
□ Описательная статистика (3)
□ Кластерный анализ (2)
□ Факторный анализ (4)
□ Регрессионный анализ (2)
□ Корреляционный анализ (2)
□ Обработка петролого-геохимических данных (5)
□ Структурный анализ публикаций (1)
□ Обработка естественного

Список сервисов

Предобработка данных

Стандартизация
Приведение данных к распределению со средним 0 и стандартным отклонением 1.

Обработка пропусков в данных
Обработка пропусков в данных на основе остальных значений согласно выбранной стратегии.

Нормализация
Приведение данных к нужному диапазону.

Описательная статистика

Основные статистические показатели
Вычисление основных статистических показателей по выбранным столбцам: среднее, минимум, максимум, квантили, дисперсия и т.д.

Вычислительные узлы

Многомерные методы анализа данных
Государственный геологический музей им. В.И. Вернадского РАН

Обработка петролого-геохимических данных
Институт Физики Земли РАН

Структурный анализ публикаций
Варшавский Университет, Польша

Обработка естественного языка
Университет Шеффилда, Великобритания

Визуализация табличных данных
Компания Plotly

Figure 2: Service catalog interface for geological data analysis.

List of tested system functions:

- interaction with the platform data store;
- interaction with the service catalog;
- interaction with WPS services;
- interaction with third-party data stores.

The testing of the services presented on the platform was carried out in manual mode using the initial data, input parameters and expected results based on data taken from the sources corresponding to the service topic. So, for example, to test the operation of the computing node services using multidimensional methods of data analysis, the materials of the book by J. Davis "Statistics and analysis of geological data" were used.

5. Conclusion

This ecosystem is being developed to provide geological research with modern methods and tools for analyzing geological information, which implies a further increase in the number of processing tools and their varieties provided by the platform.

The principles of the ecosystem being developed can be used in the future to create various digital computing platforms to support and accompany scientific research.

Acknowledgments

The study is supported by the Government contract No. 0140-2019-0005 with SGM RAS "Development of an information environment for integrating data from natural science museums and their processing services for Earth sciences".

References

[1] Naumova V.V., Platonov K.A., Eremenko V.S., Patuk M.I., Dyakov S.E. Information and analytical environment for supporting scientific research in geology: current state and development prospects // Proceedings of the XVII International Conference "Distributed Information and Computing Resources (DICR-2019)". 2019. P. 139–147. (In Russ.)

[2] Eremenko V.S., Naumova V.V., Platonov K.A., Dyakov S.E., Eremenko A.S. The main components of a distributed computational and analytical environment for the scientific study of geological systems // Russian Journal of Earth Sciences. 2018. Vol. 18. Is. 6.

[3] Eremenko V.S., Naumova V.V., Zagumennov A.A., Bulov S.V. Cloud technologies for development of geographically distributed computational and analytical geological environment // Computational Technologies. 2021. Vol. 26. No. 1. P. 86–98.

[4] Bychkov I.V., Ruzhnikov G.M., Fjodorov R.K., Shumilov A.S. Components of WPS-services for geodata processing environment // Vestnik NSU. Series: Information Technologies. 2014. Vol. 12. No. 3. P. 16–24. (In Russ.)

[5] Eremenko V.S., Naumova V.V. A system for cataloging and monitoring geographically distributed computing nodes in an environment of WPS services for solving geological problems // Vestnik NSU. Series: Information Technologies. 2019. Vol. 17. No. 2. P. 39–48. (In Russ.).

[6] Ivanov S.D. Interactive web application based geosensors registry // Computer Research and Modeling. 2016. Vol. 8. No. 4. P. 621–632. (In Russ.)

[7] Tkaczyk D., Szostek P., Fedoryszak M., Dendek P., Bolikowski L. CERMINE: Automatic extraction of structured metadata from scientific literature // International Journal on Document Analysis and Recognition. 2015. Vol. 18. No. 4. P. 317–335.

[8] Maynard D., Bontcheva K., Augenstein I. Synthesis lectures on the semantic web: Theory and technology // December 2016. Vol. 6. No. 2. P. 1–194.