Computational and Analytical Environment for Processing and Analysis of Geological Data

Vitaliy S. Eremenko
Vernadsky State Geological Museum RAS
Moscow, Russia
vitaer@gmail.com

Vera V. Naumova
Vernadsky State Geological Museum RAS
Moscow, Russia
v.naumova@sgm.ru

Abstract

This work describes the principles and technologies for creating a single access point to cloud computing to solve various scientific problems within the geological computing and analytical environment. The current approaches to the implementation of computing environments are briefly described. Based on the tasks facing researchers in the field of processing and analysis of geological information, requirements are formulated for external computing nodes and processing platforms. Based on the stated requirements, several methods and technologies are proposed. They allow to organize a single method of interaction with heterogeneous processing services. A structure for description of processing services for cataloging is proposed. The principles of operation of a monitoring system for processing services located in the environment services catalog are described.

1 Introduction

Conducting research in the field of geology involves working with heterogeneous geological information. Several basic types of data in geology can be distinguished: quantitative data (field measurement databases, etc.), spatial data (geological, topographic, geochemical, and other maps), satellite images, scientific publications, expert information, media information, etc. To process and analyze each type of data, you must use the appropriate software. In addition to financial investments, such software requires a certain amount of computing power and software environment. It also requires the user to have the skills in working with this software environment - installation and configuring of the components of the selected software and work with it.

With the development of information technology, many suppliers of software packages for data processing and analysis began to create analogues of their processing programs in the form of cloud services. Services are accessed using a web browser or special client-side software, and all calculations are performed on the side of the service provider, using its computing power and software environment.

2 Distributed computing environments

The use of external geographically distributed computing nodes for parallel data processing in one system is a kind of distributed computing environment. Such environments are developed by various business and scientific organizations to solve a certain class of problems. In 2006, Fedotov A.M. and co-authors presented a distributed information-analytical environment for research of ecological systems [Fed06]. The paper describes the main components of the environment, defines the categories of data, describes the model of the virtual environment, its architecture and the technologies used. In the work of Gordov E.P. et al. it is presented a project on creation of a thematic virtual research environment for analysis, assessment and forecasting the impact of global climate change [Gor16]. The main goal of the project being developed is to provide a free access to various sources of data and its processing services through web-browser. The article by Candela, L. et al. provides a general overview of existing virtual research environments. They identify general and distinctive features of various approaches for constructing such environments and discusses the problems that need to be addressed in this area [Can14]. Bychkov I.V. with co-authors developed and successfully released an environment of WPS-services for geodata processing [Byc14, Fjo16]. This environment supports calling of processing services built using the OGC Web

Processing Service (WPS) interface. The environment provides an ability to build processing chains (processing scripts) using the javascript language.
The aim of this work is to develop principles and technologies for single access point creation to cloud computing for solving various scientific problems within the computational and analytical environment in geology.

3 Basic requirements for computing nodes and processing platforms

Access to external computing nodes and processing platforms involves the use of various protocols and interaction interfaces. Unlike computing nodes, processing platforms offer the user not only individual processing services, but also additional tools for finding processing methods available on the platform, and the ability to view additional information about each processing algorithm available on the platform.

To integrate external computing nodes and processing platforms into a single computational and analytical environment, we have identified a number of requirements:

- External access via the Internet using a fixed IP address (or domain name) and port;
- Existence of the program interface of interaction (API) for external use. The API should be implemented using generally accepted interaction interfaces, for example, using the SOAP protocol or an interface based on the REST architectural style. It is also allowed to use other interfaces in the presence of software libraries of the most used programming languages, such as Java, Python, C;
- The ability to start processing or analysis procedures with specified parameters;
- The ability to receive the processing result in text or binary formats, including URL-link;
- The ability to work with remotely hosted data, or a mechanism for temporarily downloading data to a computing node for further work with them, or transferring them in text or binary formats as a processing parameter.

For processing platforms, it is also necessary to introduce a number of additional requirements:

- The ability to obtain a list of processing and analysis algorithms supported by the platform with a brief description of the basic principles of its operation;
- The ability to obtain more detailed information about a specific processing algorithm.

The use of geological information implies a number of requirements for the formats supported by the data environment.

List of data formats for each type of geological information:

- Quantitative data: CSV, XLS, XLSX
- Spatial data: HDF5, netCDF, GeoTIFF, Shapefile, KML, CDR
- Scientific publications: PDF
- Text data: TXT

4 Software platform

To interact with external nodes and processing platforms it is necessary to choose a software platform. This platform should have an ability to unify access to heterogeneous data processing and analysis services, considering the requirements.

One of the most promising and widely used approaches for heterogeneous data processing procedures execution is the use of a service of spatial data processing execution based on the international standard OGC Web Processing Service (WPS). This service standard provides the ability to run both individual processes and chains of these processes, executing them in serial or parallel mode, while transmitting the result of one or more processes as input parameters for another process [Xia09, Xia10].

The detailed analysis of software platforms that provide the ability to work with the WPS service was made by M. Ebrahim Poorazizi and Andrew J.S. Hunter [Poo14]. Software platforms such as 52 North, Deegree, GeoServer, PyWPS, and Zoo are considered in this paper. Based on the platform comparison table, the most suitable platform that meets the requirements we have developed is the GeoServer open source software platform. In addition to the ease of installation, configuration, and creation of new WPS processes, the GeoServer platform has extensive software support and well-described documentation. This software platform implements the OGC WPS interface standard version 1.0.0, and has an ability to add own WPS processes implemented in the Java programming language.

Using the general principles of building of distributed computational environments, we proposed the structure of a computational and analytical environment for geological research (Figure 1) [Ere18].

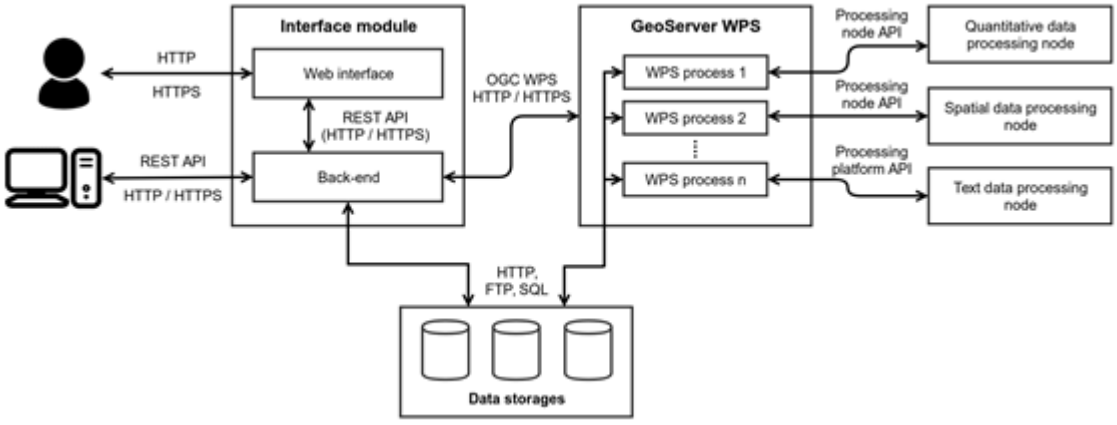

Figure 1: Architecture of a computational and analytical environment for processing geological information

5 Computing nodes and processing platforms

Based on the requirements we developed for integration into the computational and analytical environment, a number of computing nodes and processing platforms that provide open access to the processing and analysis of geological information were selected:

- Multidimensional data analysis. The computational unit includes a set of methods for multivariate analysis of quantitative data, such as factor analysis, cluster analysis, regression analysis, etc. [Pla18]. The programming language R was chosen as a component for implementing the module for statistical analysis of quantitative data. The interface for interacting with services is built using the Rserve module. The node was developed and maintained at SGM RAS.
- Satellite data processing. The computing node includes methods for processing of satellite data, such as calibration and spatial reference of satellite data [Dya16]. The node was developed and maintained at IACP FEB RAS.
- Processing of petrological and geochemical data. A processing platform has been developed at IPE RAS, which is an interactive base of methods for processing petrological and geochemical data [Iva16]. The platform provides services for constructing spidergrams, histograms, and classification diagrams; mineral identification service by their chemical composition; service for interpreting the composition of the mineral and decomposition into minerals, etc. The interface for interacting with services is based on the REST architecture.
- Structural analysis of publications. Interdisciplinary Centre for Mathematical and Computational Modelling (University of Warsaw, Poland) has developed a computational unit that performs procedure for extracting metadata from scientific publications [Tka15]. Metadata includes a list of authors, affiliation of an organization, abstracts, keywords, title, volume, year of issue, parsed bibliographic references, document section structure, section headings and paragraphs. The interface for interacting with services is based on the REST architecture.
- Natural language processing. At the University of Sheffield, as part of the GATE (General Architecture for Text Engineering) project, a computing node containing a number of text processing services for various languages was developed [Mya16]. For the processing of textual data in Russian, services are provided to determine the parts of speech of words, as well as the allocation of named entities, such as names and surnames, names of organizations, geographical names, dates, monetary units, etc. The interface for interacting with services is based on the REST architecture.
- Visualization of quantitative data. The Plotly project is an open data visualization platform. It provides the ability to visualize quantitative data in the form of tables and diagrams of various types (https://plot.ly). The interaction is carried out using an interface based on the REST architecture.

6 User interaction with the environment

To interact with the environment, a module has been developed. The module provides the user with a web interface for working with the environment through a web browser. The module has a program interface based on the REST architectural style for programmatic interaction in the system-system format.

When using the web-interface, the user selects the processing service he is interested in and fills in the parameters necessary to start this service. The user can add the source data for processing or analysis in several ways:

- upload from the user's computer - the user selects data on his PC, after which data is downloaded to the temporary storage of the environment and transferred to the appropriate processor.
- link to data from an external information system - the user indicates the URL address as source data. The environment supports working with links via the HTTP (s) and FTP protocols.
- data selection from the GeologyScience information system - the user selects data on the GeologyScience information system portal (http://geologyscience.ru), after which the identifiers of the selected data are transferred to the processing environment as source data for the specified service.

After processing is completed by an external computing node or processing platform, the environment publishes the result into a temporary storage. The result is shown to user in the form of a URL for download. The user can see the list of all orders with parameters and their results in the corresponding section of the web-interface of the environment.

7 Cataloging and monitoring services

To provide search and service information functions about services integrated with external computing nodes and processing platforms, a service presentation format was developed [Ere19]. The basis is the service description format in UDDI (Universal Description Discovery and Integration) registries. This format includes three main sections: description of the service, information about the supplier and technical information. The description section contains the service identifier, name, text description, keywords and scope. Supplier information includes the name of the supplier’s organization, contact person, contact address, contact phone number and website. The technical information section includes the IP address (or domain name) of the service, port number, interaction protocol, protocol version, description of the access interface, authorization information and the address of the access point.

Based on the proposed structure of the service description, a catalog of processing services has been developed. This catalog is a web service that allows you to search for services and get complete information about a service in XML and JSON formats.

To ensure a high level of reliability of the environment, a monitoring system has been developed for external geographically distributed services. The monitoring system is an independent software product that monitors the current state of computing nodes and processing platforms used in the environment, as well as the services provided by them, including checking the services for changes in work.

The monitoring system uses data from the service information section in the service catalog. Monitoring includes three main types of checks:

- Availability of a compute node or processing platform
- Availability of the processing service according to the specified interaction protocol
- Checking the service for changes using test requests

Thus, the monitoring system allows you to track changes from the provider of the computing node or processing platform to make the appropriate decision about the mode of access to services in the environment.

8 Conclusions

Using existing open computing nodes and data processing platforms, an approach is proposed to create a single access point to geological information processing services for scientific research. Based on the proposed approach, a computational and analytical environment (http://service.geologyscience.ru) is implemented. To monitor the availability of computing nodes and the services they provide, a monitoring system for geographically distributed processing services has been created (http://monitoring.geologyscience.ru).

9 Acknowledgements

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

[Fed06] Fedotov A.M., Barakhnin V.B., Guskov A.E., Molorodov Yu.I. Raspredelennaja informacionno-analiticheskaja sreda dlja issledovanij jekologicheskih sistem [Distributed information and analytical environment for environmental systems research]. Computational technologies, 2006. T.11. Specialist. no. p. 113-125. (In Russ.)

[Gor16] Gordov E. P., Krupchatnikov V. N., Okladnikov I. G. and Fazliev A. Z. Thematic virtual research environment for analysis, evaluation and prediction of global climate change impacts on the regional environment // Proc. SPIE 10035, 22nd International Symposium on Atmospheric and Ocean Optics: Atmospheric Physics, 100356J (29 November 2016); doi: 10.1117/12.2249118; https://doi.org/10.1117/12.2249118

[Can14] Candela, L., Castelli, D. and Pagano, P., 2013. Virtual Research Environments: An Overview and a Research Agenda. Data Science Journal, 12, p.GRDI75–GRDI81. DOI: http://doi.org/10.2481/dsj.GRDI-013

[Byc14] Bychkov I. V., Ruzhnikov G. M., Fjodorov R. K., Shumilov A. S. Komponenty sredy WPS-servisov obrabotki geodannyh [Components of WPS-services for geodata processing environment]. Vestnik NSU. Series: Information Technologies, 2014, vol. 12, no. 3, p. 16-24 (In Russ.)

[Fjo16] Fjodorov R. K., Bychkov I. V., Shumilov A. S., Ruzhnikov G. M. Sistema planirovanija i vypolnenija kompozicij veb-servisov v geterogennoj dinamicheskoj srede [System for planning and executing web service compositions in a heterogeneous dynamic environment]. Computational technologies, 2016. T. 21. Specialist. no. p. 18-35. (In Russ.)

[Xia09] Xiaoliang Meng, Fuling Bian, Yichun Xie Geospatial Services Chaining with Web Processing Service // Proceedings of the International Symposium on Intelligent Information Systems and Applications (IISA'09) Qingdao, P. R. China, Oct. 28-30, 2009, pp. 007-010

[Xia10] Xiaoliang Meng, Yichun Xie. Fuling Bian Distributed Geospatial Analysis through Web Processing Service: A Case Study of Earthquake Disaster Assessment // Journal Of Software. 2010. Vol. 5. No. 6

[Poo14] Poorazizi, M. Ebrahim and Hunter, Andrew J.S. Evaluation of Web Processing Service Frameworks // Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings: 2014, Vol. 14 , Article 5.

[Ere18] 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, Volume 18, Issue 6, doi:10.2205/2018ES000636.

[Pla18] Platonov K.A., Methods and Technologies for Integration and Processing of Territorially Distributed Quantitative Geological Information // Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018, p. 348-353.

[Dya16] Dyakov S.E. Cross-calibraton channels of ir-radiometers and sea surface termerature retrivial // Proceedings of the IV International Conference "Modern Information Technologies in Earth Sciences", Yuzhno-Sakhalinsk, August 7-11, 2016, p. 52.

[Iva16] Ivanov S.D. Interaktivnyj reestr geosensorov na osnove veb-prilozhenija [Interactive Web Application Based Geosensors Registry]. Computer research and modeling, 2016, vol. 8, no. 4, p. 621–632. (In Russ.)

[Tka15] Tkaczyk D., Szostek P., Fedoryszak M., Dendek P., Bolikowski L. CERMINE: automatic extraction of structured metadata from scientific literature. In International Journal on Document Analysis and Recognition, 2015, vol. 18, no. 4, p. 317-335, doi: 10.1007/s10032-015-0249-8.

[Mya16] Maynard D., Bontcheva K., Augenstein I. Synthesis Lectures on the Semantic Web: Theory and Technology, December 2016, Vol. 6, No. 2, p. 1-194.

[Ere19] Eremenko V. S., Naumova V. V. Sistema katalogizacii i monitoringa territorial'no raspredelennyh vychislitel'nyh uzlov v srede WPS servisov dlja reshenija geologicheskih zadach [The system of cataloging and monitoring geographically distributed computing nodes in the environment of WPS services for solving geological problems] // Vestnik NSU. Series: Information Technologies. 2019. T. 17, No 2. C. 39–48. DOI 10.25205/1818-7900-2019-17-2-39-48 (In Russ.)