

Development of the computing node for processing satellite imagery and spatial data for earth sciences

Aleksey A. Zagumennov^{1,2}, Vera V. Naumova¹

¹*Vernadsky State Geological Museum of RAS, Moscow, Russia*

²*Institute of Automation and Control Processes of FEB RAS, Vladivostok, Russia*

Abstract

The work is devoted to the development of a computing node for processing satellite and spatial data for earth sciences by the example of its implementation as part of the Information and Analytical Environment to support scientific research in geology of the Vernadsky State Geological Museum (SGM RAS). The prerequisites for the creation of such a computing node and the requirements for it to solve geological problems are given. An overview of cloud platforms for access to satellite and spatial data and its processing has been presented. Based on the overview a conceptual diagram of a computing node has been proposed and the list of modern technologies required for building it has been determined. The developed node provides tools for searching data from external cloud providers, processing them with various built-in and custom algorithms, as well as tools for visualizing the results. It is an independent web service, although it is part of the Computational and Analytical Geological Environment of SGM RAS and is integrated with its services. This allows a wide range of users to access data and processing algorithms provided by computing node, including integrating it into other information systems as a third-party application for processing satellite and spatial data.

Keywords

Cloud services, geology, computing node, REST API, satellite imagery.

1. Introduction

The most of nowadays research in geosciences cannot be conducted without satellite or spatial data. Remote sensing with modern satellites provides information about Earth in many spectral ranges. Various physical parameters of the surface, ocean and atmosphere are calculated using this information. Satellite imagery is used in research in meteorology, oceanology, geology, and other earth sciences. Processed remote sensing data become crucial in industry, agriculture, territory management and other areas.


Problems and tasks that involve satellite imagery and spatial data processing require huge amount of storage space and computing power in terms of infrastructure, as well as a large number of special competencies in the field of geoinformatics in terms of qualification requirements. This fact along with the constant growth in the number of new satellites launched every year have led to the emergence of specialized cloud platforms and services to work with satellite imagery and spatial data. They reduce the cost of solving certain scientific and applied problems, providing access to data and giving tools and algorithms to process this data. Considering only

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

✉ truepikvic@gmail.com (A. A. Zagumennov)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

scientific problems one can notice that they define a number of basic requirements for such cloud platforms and services: the possibility of open access for scientific purposes, simplicity and flexibility of use for various scientific problems, the possibility of implementing custom processing algorithms, and reproducibility of results.

Three main problems arise working with satellite and spatial data:

- 1) spatiotemporal search for satellite and spatial data and access to them;
- 2) interactive data visualization and analysis;
- 3) data processing with standard or user-defined algorithms.

Modern cloud solutions in considered field partially or completely solve these problems. The problem of access to satellite imagery should be highlighted here. The development of satellite remote sensing and forming of various initiatives to provide access to satellite data made it possible to create distributed systems providing access to a wide variety of satellite and spatial data, and researchers and other users got access to data directly through the web interfaces, without the need to download them directly into local storage. Mentioned above problems of working with satellite and spatial data and impact of modern cloud platforms to their solution are discussed in the related work [1].

Information and Analytical Environment has been implemented to support scientific research in geology (<http://geologyscience.ru>) created few years ago operates in Vernadsky State Geological Museum of Russian Academy of Sciences (SGM RAS) providing a single entry point to various databases and cloud processing services for solving scientific problems in geology [2]. The information and analytical environment consists of services for accessing various types of data, including a service for accessing satellite data (<https://sputnik.geologyscience.ru>) and the computing and analytical environment for processing geological data (<https://service.geologyscience.ru>), the latter integrates various external computing nodes and cloud platforms for the processing of quantitative, spatial and text data [3].

In this paper, an approach to creating a computing node for processing satellite and spatial data in the ecosystem of the Information and Analytical Environment of the SGM RAS for solving geological problems is proposed to consider.

2. Overview of cloud platforms to work with satellite and spatial data

There are many cloud platforms and services to work with satellite and spatial data nowadays. They differ in functionality, type of satellite and spatial information provided, thematic focus, ease of use, cost. Many of the platforms provide an option to use them for research purposes, but often only a part of the functionality provided. Examples of several platforms, their architecture, and provided features are presented in the related work [4]. We will give an overview of the platforms from the point of the possibility of their use for scientific research involving satellite and spatial data. All platforms can be divided into two large classes: general purpose platforms and thematic ones.

General-purpose platforms allow to solve a wide range of tasks by means of access to different satellite and spatial data, variety of visualization tools, custom processing algorithms and the

mechanisms to share the results of work. The main general-purpose platforms are considered below.

Google Earth Engine (<https://earthengine.google.com>) is a cloud platform for analysis and visualization of geospatial data, free for the scientific purposes. The platform aggregates data from various satellite platforms and instruments. To work with this archive of satellite and spatial data, tools for visualization tools and development of processing algorithms are provided powered by the Google Cloud Platform.

Earth Observing System (<https://eos.com>) is a commercial cloud platform that provides access to a wide range of satellite data, including ultra-high resolution. The platform offers a wide range of visualization and analytics tools, as well as a set of common algorithms for processing satellite data for various tasks. To use it in scientific purposes it is required to sign special agreement.

Planet (<https://www.planet.com>) is another commercial platform from Planet, which is distinguished by the fact that the company has its own constellation of Earth observation satellites, which makes it possible to track in more detail all changes in various parameters of the regions of interest. In addition to valuable satellite data, the platform has advanced visualization and change monitoring tools, as well as integration tools with various satellite data processing applications. It is possible to use the platform and its data for research and educational purposes.

Descartes Labs (<https://www.descarteslabs.com>) is a commercial cloud platform that aggregates and prepares satellite data for further analysis from various data providers. It has modern tools for visualization and data analysis, as well as special workspaces where, using the computing power and the programming interface of the platform in the Python programming language, users can build their own satellite data processing workflows to solve their problems. It also provides an opportunity for free use in for scientific purposes.

ArcGIS Online (<https://www.esri.com/en-us/arcgis/products/arcgis-online/overview>) is a commercial cloud platform from a well-known developer of geographic information systems (GIS), which is primarily a cloud GIS. Provides a wide range of tools for working with geospatial data (raster and vector), including analytics and a Python API. It is possible to upload your data, share the results of your work, use the results of the work of other users of the platform.

Astraea (<https://astraea.earth>) is a commercial cloud platform which has a well-structured satellite and spatial data processing workflow as a distinctive feature.

1. Automatic continuous delivery of satellite images of the region of interest from the required satellites.
2. Custom processing algorithms creation using the JupyterLab environment in the Python programming language.
3. Algorithms scaling to cloud computing nodes using Big Data approaches.
4. Creation of automated analytical tasks using low/no-code approaches.

The platform provides an opportunity for cooperation in solving research and scientific problems.

Sentinel Hub (<https://www.sentinel-hub.com>) is a commercial platform with great opportunities for research and solving scientific problems. It provides access to data from satellites of the Sentinel and Landsat constellations, as well as satellites Meris and Proba-V, has flexible

visualization tools, and, like Google Earth Engine, allows users to implement their own scenarios for processing satellite data and provides a number of software interfaces for automated interaction with platform. It is also possible to use custom data.

Thematic platforms are aimed at solving a certain narrow range of tasks, as a rule, limited to a certain geographic region, providing the most relevant and thoroughly selected set of data, tools and algorithms for these tasks and region. Here are some examples of thematic platforms.

USGS Web Applications (<https://www.usgs.gov/products/data-and-tools/web-application>) is a collection of thematic web services from the US Geological Survey to solve a wide range of tasks in areas from geology to climate, mainly for the territory of USA. Web services use the USGS cloud data access platform and internal cloud infrastructure. All services and data access platform are free for scientific research.

NASA Earthdata Tools (<https://earthdata.nasa.gov/earth-observation-data/tools>) is a set of cloud-based tools from the US National Aerospace Agency, which provide the following functionality: data search and ordering, data preprocessing and filtering, geolocation and cartography, data visualization and analysis. Each of the services from this catalog is focused on solving a fixed range of tasks from certain domains. The services are also free and not limited just to the United States region.

Copernicus Marine Service (<https://marine.copernicus.eu>) is a cloud-based platform focused on research of the World Ocean, which is implemented as part of the European Union's Copernicus Program. The platform offers access to satellite and spatial data on the state of the ocean, as well as a range of tools for visualizing and analyzing various parameters of the ocean, thereby allowing to solve scientific and applied problems requiring this type of data. The platform is free and covers the entire World Ocean. There are similar thematic platforms within the Copernicus Program for climate, atmosphere and land studies.

Digital Earth Australia (<https://www.ga.gov.au/dea/products>) is a thematic cloud platform for monitoring various physical parameters in Australia. It combines satellite data and a set of thematic services that are focused on solving specific problems: from changing coastlines to determining freshwater reserves. Access to the platform is free.

Swiss Data Cube (<https://www.swissdatacube.org>) is a thematic cloud platform similar to the previous one, only focused on monitoring the territory of Switzerland.

Brazil Data Cube (<http://brazildatacube.org>) is a thematic cloud platform focused on monitoring the territory of Brazil.

The above review of cloud platforms for working with satellite and spatial data allows us to conclude that, to solve modern problems, such platforms, on the one hand, require the implementation of data access, processing and visualization capabilities, and on the other hand, simplicity and flexibility of use in modern thematic scientific tasks with the ability to define data sets and algorithms for their processing. At the same time, general-purpose platforms still have their own specifics and different goals, differences in the set of services provided, and the possibilities of using them for research purposes. This explains the large number of such platforms.

3. Computing node for processing satellite imagery and spatial data for earth sciences

Earth observation data are widely used in modern geology in a wide range of tasks: lineament analysis, minerals mapping, structural-tectonic analysis of deposits, monitoring of geodynamic processes, study of permafrost processes, study of the material composition of rocks, rational land-use, etc. Solving these problems requires a certain set of satellite and spatial data and algorithms for their processing, the composition of which is constantly changing with the development of methods for solving problems, also the problems themselves are changing. This defines certain requirements on tools for processing satellite and spatial data, which should combine the properties of both general-purpose and thematic platforms.

In terms of data access requirements, it is necessary to search and obtain data from various satellites and radiometers, as well as from spatial data catalogs. One solution is to create a data access gateway. Thus, the service for accessing satellite data of the Information and Analytical Environment of the SGM RAS (<http://sputnik.geologyscience.ru>) provides the search of satellite data from various providers: the US Geological Survey (<http://usgs.gov>), the Japan Aerospace Agency research (<https://www.eorc.jaxa.jp>), satellite center. Goddard (<https://oceancolor.gsfc.nasa.gov>), Scientific Center for Operational Monitoring of the Earth (<http://www.ntsomz.ru>), Center for Collective Use of Regional Satellite Environmental Monitoring of the Far Eastern Branch of the Russian Academy of Sciences (<http://satellite.dvo.ru>). The service provides access to the following types of satellite information: data from Landsat-7/8 satellites and Sentinel-2A/2B satellites; satellite topography data STRM and ALOS; data from meteorological satellites Aqua and Terra; data from satellites EO-1, OrbView-3, as well as from the Russian satellite Kanopus-V.

Another option for providing access to data is to connect to satellite and spatial data catalogs that implement the modern rapidly developing Spatiotemporal Asset Catalog (STAC) specification (<https://stacspec.org>). This specification regulates the rules for describing data and collections of spatiotemporal data to provide unified access to this data and navigation through it using a self-documented directory structure and data description. The main advantage of using this specification by data providers and their consumers is the uniformity of access to data without the need to change processing workflows and algorithms when adding new data types.

Modern work with satellite and spatial data takes place in web applications through interaction with a cartographic interface, into which the needed data is loaded, receiving from cloud platforms in real time. To provide such work with data, special storage formats and some preprocessing of real data are required to optimize delivery over the network, as well as render in the browser. For these purposes, the approaches offered by the Open Geospatial Consortium (OGC) (<https://ogcapi.ogc.org>) have been used for a long time: Web Map Service, Web Feature Service, Web Coverage Service, etc. But more recently, the Cloud-Optimized Geotiff (COG) standard (<https://www.cogeo.org>) has also begun to be used for these purposes. This standard adds to regular Geotiff files the ability to store overview images as well as smaller chunks of the original image for quick access maintaining backward compatibility. The only prerequisite for accessing such data over the network is support for HTTP GET Range requests by both the client and the server.

The overview of thematic cloud platforms for working with satellite and spatial data shows that an increasing number of such platforms are created using modern opensource software Open Data Cube (ODC) (<https://www.opendatacube.org>) [5]. ODC is an opensource suite of geospatial data management and analysis software built on top of other open technologies. It combines tools for cataloging data, direct access to data in the form of data cubes — multidimensional spatiotemporal arrays of measurements — and functions in the Python programming language to provide computations, including distributed ones. Thus, ODC can be the main system for general and thematic processing of satellite and spatial data of any size: from an individual researcher's workplace to a cloud platform [6].

The modern landscape of platforms to work with satellite and spatial data, as well as a set of technological solutions and standards in the field of geoinformatics, make it possible to implement a similar approach when developing a computing node for processing satellite and spatial data within the ecosystem of the Information and Analytical Environment of SGM RAS. The schematic diagram of the computing node is shown in Figure 1. Computing node consists of the following components:

- a cartographic web interface for user interaction with a computing node by searching for satellite and spatial data for the region and dates of interest, defining a processing algorithm, choosing a visualization and analysis methods for the result;
- a subsystem for processing and dispatching requests, which provides interaction between the web application and the rest of the computing node;
- a data access subsystem that provides interaction with STAC-catalogs of external cloud providers, a service for accessing satellite data of the SGM RAS, and contains service functions for working with the local ODC catalog, which is required for the ODC to work properly;
- data processing subsystem, which is based on the ODC platform, providing tools for satellite imagery processing algorithms implementation;
- a task queue, into which incoming requests for data processing are placed with an indication of the algorithm from the data processing subsystem, and which can track the status of tasks execution;
- task executors that process tasks from the queue in a distributed manner;
- local storage of data processing results in COG format and temporary files.

Calculations are performed in the Python environment using the ODC package and several auxiliary packages for working with geospatial data. The processing of incoming requests for data processing is performed by the FastAPI framework using the REST API implemented according to the OpenAPI standard (<http://spec.openapis.org/oas/v3.0.3>), using a queue of computational tasks based on the NoSQL Redis database.

Moreover, individual components of the system — the web interface, subsystems, the local ODC directory, a database with a task queue, task executors — are deployed using Docker containers. This architecture allows the processing of requests and heavy computation of large amounts of data to be separated providing fault tolerance and scalability of the considered computing node.

The current implementation of the proposed conceptual scheme is a prototype with a cartographic web interface that implements data search in the STAC catalog of the Landsat-8 satellite

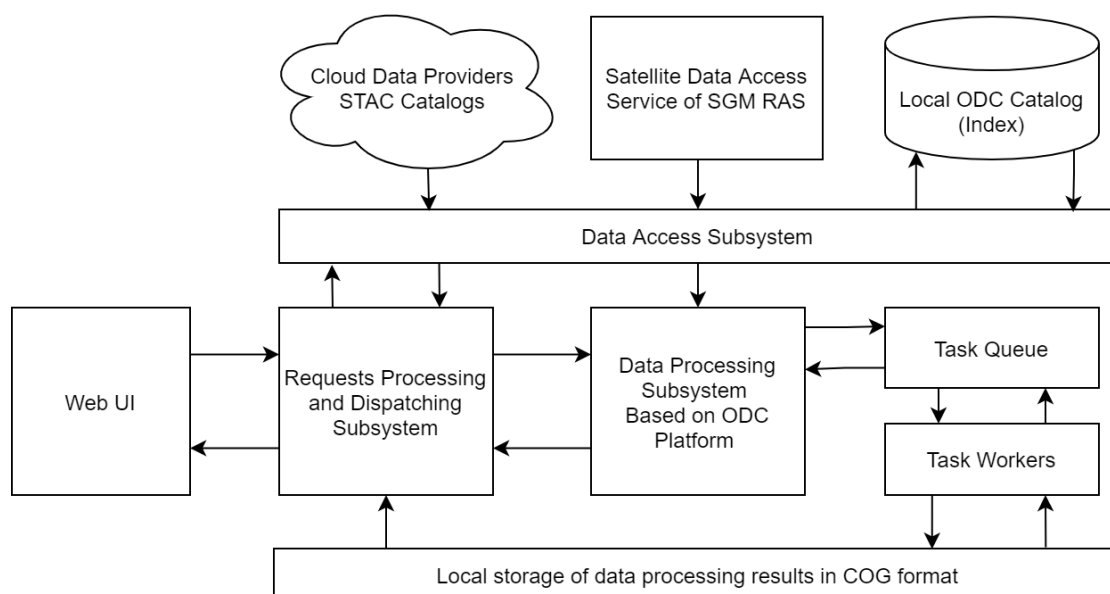


Figure 1: Schematic diagram of a computing node for processing satellite and spatial data.

(<https://landsat-stac.s3.amazonaws.com>), as well as algorithms for calculating various spectral indices. satellite channels.

4. Further work

Further development of the computing node is supposed to be carried out in three directions:

- increasing the functionality of the web interface by adding new visualization and data analysis tools;
- development of a data access subsystem through the ability to connect arbitrary STAC catalogs, as well as integration with the satellite data access service of SGM RAS;
- expansion of the list of algorithms provided by the data processing subsystem, as well as the implementation of the possibility of adding custom algorithms.

In addition, the approaches and modular architecture the computing node is based will make it possible to transform it into an independent cloud platform for solving a wide range of tasks for earth sciences. And the ease of use and flexibility of tuning for specific tasks will attract a wide range of scientists and researchers to its use.

Acknowledgments

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

References

- [1] Sudmanns M., Tiede D., Lang S., Bergstedt H., Trost G., Augustin H., Baraldi A., Blaschke T. Big Earth data: Disruptive changes in Earth observation data management and analysis? // *International Journal of Digital Earth*. 2020. Vol. 13. Is. 7. P. 832–850.
- [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. No. 6. ES6003. DOI:10.2205/2018ES000636.
- [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. Is. 1. P. 86–98. DOI:10.25743/ICT.2021.26.1.007.
- [4] Gomes V.C.F., Queiroz G.R., Ferreira K.R. An overview of platforms for big earth observation data management and analysis // *Remote Sensing*. 2020. Vol. 12. Is. 8. P. 1253.
- [5] Dhu T., Giuliani G., Juárez J., Kavvada A., Killough B., Merodio P., Minchin S., Ramage S. National open data cubes and their contribution to country-level development policies and practices // *Data*. 2019. Vol. 4. Is. 4. P. 144.
- [6] Giuliani G., Chatenoux B., Piller T., Moser F., Lacroix P. Data cube on demand (DCoD): Generating an earth observation data cube anywhere in the world // *International Journal of Applied Earth Observation and Geoinformation*. 2020. Vol. 87. DOI:10.1016/j.jag.2019.102035.