УДК 004.550

СОЗДАНИЕ ЭКОСИСТЕМЫ ПО ОБРАБОТКЕ ДАННЫХ ДЛЯ НАУЧНЫХ ИССЛЕДОВАНИЙ В ОБЛАСТИ ГЕОЛОГИИ

В. С. Ерёменко[1][0000-0002-5250-5743], В. В. Наумова[2][0000-0002-3001-1638]

Государственный геологический музей им. В.И. Вернадского РАН, Москва

[1] vitaer@gmail.com, [2] naumova_new@mail.ru

Аннотация

Рассмотрены разнородные территориально-распределённые вычислительные системы по обработке геологических данных и подходы по организации взаимодействия с этими системами. Исследуемые системы классифицированы на несколько групп, исходя из принципов их функционирования и выбранных технологических решений. Для каждого типа систем приведено описание их основных свойств, включая возможные способы для взаимодействия.

На основе проведённого анализа предложен подход к организации единого рабочего пространства с доступом к разнородным территориально-распределённым вычислительным системам в рамках общей экосистемы. Описаны архитектура предлагаемой экосистемы и правила взаимодействия её участников. Продемонстрирован программный прототип, реализующий названные принципы, на примере нескольких разнородных систем по обработке геологической информации.

Ключевые слова*: вычислительно-аналитическая среда, облачные сервисы, веб-сервисы, программные платформы*

ВВЕДЕНИЕ

Использование внешних сервисов вместо пользовательских приложений позволяет проводить обработку данных на оборудовании, наиболее приспособленном для соответствующей задачи. Тем самым обработка данных происходит более эффективно, а пользователь получает возможность обрабатывать данные

наиболее актуальными версиями алгоритмов, используя для этого веб-интерфейс, без необходимости установки, настройки и поддержки программного обеспечения для обработки у себя на персональном компьютере.

В Государственном Геологическом Музее им. В.И. Вернадского РАН с 2018 года разрабатывается вычислительно-аналитическая Среда для поддержки научных исследований в геологии [1]. Она предоставляет пользователям единую точку доступа к территориально-распределённым вычислительным узлам для проведения научных исследований.

ПРОГРАММНЫЕ ИНСТРУМЕНТЫ ПО ОБРАБОТКЕ ДАННЫХ

Для организации взаимодействия с различными инструментами по анализу данных необходимо провести их классификацию по различным критериям. Можно выделить следующие основные критерии:

1. По типу взаимодействия:
    - Настольное (standalone) приложение;
    - Приложение «клиент-сервер»
        - Веб-приложение (клиент – веб-браузер);
        - Настольный клиент;
    - Программное API
        - Набор библиотек;
        - Веб-сервис (Web-API);
        - Сервис с собственным протоколом доступа;
2. По типу архитектуры вычислительного узла:
    - Виртуальный / традиционный сервер;
    - Облачная архитектура;
    - Суперкомпьютер;
    - Распределённая сеть вычислительных кластеров (GRID);
3. По типу инструмента обработки:

- Приложение;
- Программная платформа;
- Аппаратно-программная инфраструктура;

4. По разграничению доступа:
    - Публичный;
    - Частный;
    - Гибридный.

ВЕБ-СЕРВИСЫ

Подход к построению информационных систем с использованием веб-сервисов в качестве независимых компонентов со стандартизированными интерфейсами и протоколами взаимодействия называется сервис-ориентированной архитектурой (SOA – service-oriented architecture).

Наиболее часто встречаются программные комплексы на базе протокола SOAP (Simple Object Access Protocol) или с использованием интерфейсов на основе архитектуры REST (Representational State Transfer).

SOAP – это протокол обмена структурированными сообщениями в распределённой вычислительной среде. Согласно этому протоколу сообщения в формате XML помещаются в SOAP-конверт для дальнейшей передачи сообщения получателю. Чаще всего протокол SOAP реализуется на основе протокола HTTP (HyperText Transfer Protocol).

Более современный подход к построению веб-сервисов подразумевает использование REST-архитектуры. Как и большинство SOAP-сервисов, REST-сервисы реализуются на основе протокола HTTP. Обычно веб-сервисы на основе REST-архитектуры используют для взаимодействия сообщения в формате JSON (Javascript Object Notation).

Стандарт WSDL разработан для веб-сервисов на основе протокола SOAP (Simple Object Access Protocol), предназначенного для обмена XML (eXtensible Markup Language) сообщениями.

В настоящее время множество веб-сервисов по обработке и анализу научных данных использует другие технологии реализации интерфейсов, в частности, интерфейсы на основе REST-архитектуры.

ПЛАТФОРМЫ ОБРАБОТКИ И АНАЛИЗА

Развитие облачных сервисов в последнее десятилетие изменило подход к предоставлению различных информационных услуг. Согласно работе [2], облачные сервисы можно разделить на три основные группы:

- Программное обеспечение как сервис (Software as a Service, Saas);
- Платформа как сервис (Platform as a Service, PaaS);
- Инфраструктура как сервис (Infrastructure as a Service, IaaS).

Одним из крупнейших поставщиков сервисов по обработке пространственных данных является компания ESRI с линейкой продуктов ArcGIS (https://www.arcgis.com/). Облачная платформа ArcGIS Online позволяет пользователям получить доступ к различным функциям обработки и анализа пространственных данных непосредственно через веб-браузер, без необходимости установки и обслуживания локального программного обеспечения. Платформа позволяет загружать пользовательские данные непосредственно с компьютера либо из внешней базы пространственных данных для дальнейшей работы с этими данными в интерактивном режиме. Существует возможность реализации и публикации собственных методов обработки и анализа пространственных данных в общем каталоге сервисов обработки ArcGIS.

Одним из наиболее популярных программных продуктов обработки табличных данных в интерактивном режиме является Excel из пакета Microsoft Office. Этот программный продукт содержит ряд инструментов для редактирования данных, построения различных диаграмм, использования встроенных процедур анализа и создания собственных данных. Компания Microsoft разработала бесплатную облачную версию продукта Excel (https://www.office.com/). Зарегистрированный пользователь может использовать полноценную веб-версию для обработки данных, расположенных в облачном хранилище Microsoft One Drive.

Для анализа пространственных данных наиболее подходящим инструментом являются геоинформационные системы (ГИС), позволяющие в интерактивном режиме взаимодействовать с объектами на карте, применяя к отдельным

слоям и объектам доступные средства анализа. Наиболее подходящим для геологии облачным решением является ArcGIS Online, разрабатываемый компанией ESRI. Portal for ArcGIS – это инфраструктура ArcGIS Online, функционирующая в защищенной ИТ-среде или в частном облаке организации (под контролем сетевого экрана или в полностью изолированной локальной сети). Портал позволяет создавать карты, каталогизировать и анализировать пространственные данные с помощью удобного интуитивно понятного интерфейса. Создание и публикация интерактивных карт и приложений могут выполняться на любом устройстве при наличии браузера и доступа к интернету. Публикация данных в защищенном облаке Esri в виде кэшированных или динамических сервисов соответствует всем современным стандартам защиты информации с сохранением всех авторских прав.

Для анализа спутниковых данных одним из лидеров среди облачных сервисов является платформа Earth Engine компании Google (https://earthengine.google.com/). Эта платформа позволяет пользователю загружать собственные данные или использовать данные из каталога Earth Engine для дальнейшей обработки в интерактивном режиме. В каталоге содержатся продукты обработки данных радиометра Modis (спутники Aqua, Terra), спутников Sentinel-1A, Sentinel-1B, Sentinel-2A, Sentinel-2B, Landsat 8 и др. Earth Engine содержит ряд предустановленных алгоритмов анализа, а также инструменты для их создания, редактирования и запуска с использованием языков программирования Javascript и Python. Для работы пользователю необходимо наличие аккаунта Google. Для анализа и обработки можно использовать данные из облачного хранилища Google.

ВЫЧИСЛИТЕЛЬНО-АНАЛИТИЧЕСКИЕ СРЕДЫ

Развитие технологии облачных вычислений способствует появлению большого количества веб-сервисов по обработке разнородных данных. Среди них есть как коммерческие, так и открытые программные комплексы, разрабатываемые в научных организациях или различных интернет-сообществах.

Возможность использования таких программных комплексов в виде веб-сервисов позволяет создавать предметно-ориентированные виртуальные аналитические среды, объединяющие внутри себя набор веб-сервисов для решения

комплекса задач в рамках предметной области. Такие среды используют существующие внешние вычислительные узлы для обработки пользовательских данных, не требуя при этом больших вычислительных мощностей, в отличие от информационных систем, реализующих алгоритмы обработки данных внутри себя.

В настоящее время актуальной является разработка тематических вычислительно-аналитических сред, имеющих единые точки доступа к территориально распределенным вычислительно-аналитическим ресурсам, позволяющие в рамках единой системы решать различные задачи по обработке и анализу в заданной предметной области.

В [3] предложена концепция распределённой информационно-аналитической среды для исследований экологических систем: описана модель виртуальной среды, определены категории данных и объекты среды и приведен пример схемы среды с описанием используемых технологий.

В [4] представлен проект создания тематической виртуальной исследовательской среды для анализа, оценки и прогнозирования воздействия глобального изменения климата, который разрабатывается с целью обеспечения свободного доступа к различным ресурсам данных и службам обработки через веб-браузер.

В работе [5] дан общий обзор существующих виртуальных исследовательских сред, выделены общие и отличительные особенности различных подходов к построению таких сред и разобраны проблемы, которые необходимо решать в данной области.

В [6] разработана успешно функционирующая среда WPS-сервисов обработки геоданных. Она поддерживает вызов сервисов обработки, построенных с использованием интерфейса OGC WPS (Web Processing Service). Реализована возможность построения цепочек обработки с использованием языка javascript для формирования сценария обработки.

ИНТЕГРАЦИЯ С ИСПОЛЬЗОВАНИЕМ ЕДИНОЙ ПЛАТФОРМЫ

Для организации взаимодействия с внешними веб-сервисами, работающими по принципу запрос–ответ, была выбрана платформа на основе сервиса запуска процессов обработки пространственных данных на базе международного стандарта OGC Web Processing Service (WPS). Этот стандарт дает возможность за-

пуска как отдельных процессов обработки, так и цепочек этих процессов, выполняя их в последовательном или параллельном режимах, передавая при этом результат выполнения одного или нескольких процессов в качестве входных параметров для другого процесса.

ОРГАНИЗАЦИЯ ОБЩЕЙ ШИНЫ ДАННЫХ

Каждый из перечисленных облачных сервисов использует собственные уникальные протоколы взаимодействия, сильно затрудняя возможность интеграции сервиса в разрабатываемую вычислительную Среду на основе общего протокола доступа.

Некоторые облачные сервисы требуют наличия данных в их собственном хранилище. Например, при использовании Excel из MS Office Online данные должны находиться в пользовательском хранилище One Drive. Использование Earth Engine позволяет также использовать данные из облачного хранилища Google на аккаунте пользователя. Таким образом, для обеспечения «бесшовного» перехода между облачными сервисами необходимо разработать набор процедур для публикации выбранных пользователем данных в соответствующем пользовательском облачном хранилище.

Создание подобной процедуры стало возможным при использовании технологии веб-приложений, позволяющей запросить у пользователя разрешение на доступ к определённым возможностям пользовательского аккаунта различных поставщиков облачных сервисов.

Такая технология поддерживается Microsoft, Google, Yandex, ESRI и др. Перед использованием соответствующего сервиса пользователю предлагается загрузить данные для анализа в его персональное хранилище. Для этого пользователь проходит авторизацию на сайте поставщика сервиса, после чего приложение запрашивает у него разрешение на скачивание и публикацию данных в его хранилище. При выборе облачного сервиса другого поставщика пользователь имеет возможность перемещения данных для обработки в хранилище этого поставщика.

Таким образом, нами предложены подход и технологическое решение для организации единого пространства данных для различных поставщиков облачных сервисов [7] (рис. 1).

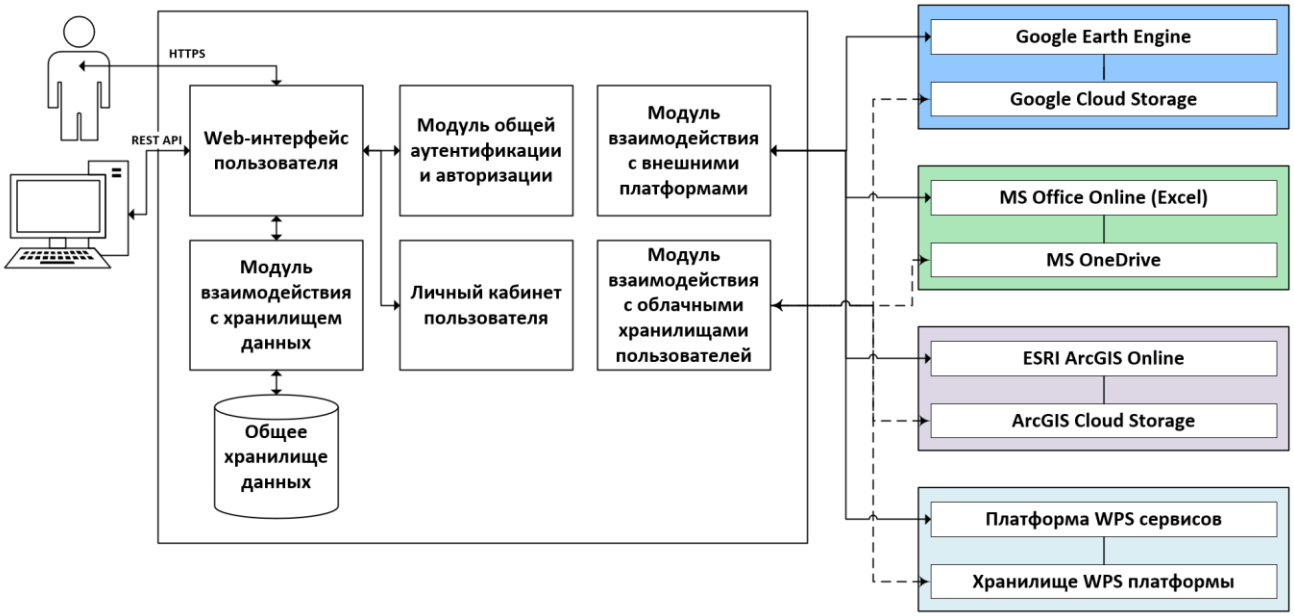


Рисунок 1. Функциональная схема вычислительно-аналитической среды

ДАЛЬНЕЙШЕЕ РАЗВИТИЕ СИСТЕМЫ

Для обеспечения более универсального подхода к обеспечению взаимодействия между различными инструментами анализа данных мы предлагаем использовать подход на основе создания общей шины платформ (рис. 2).

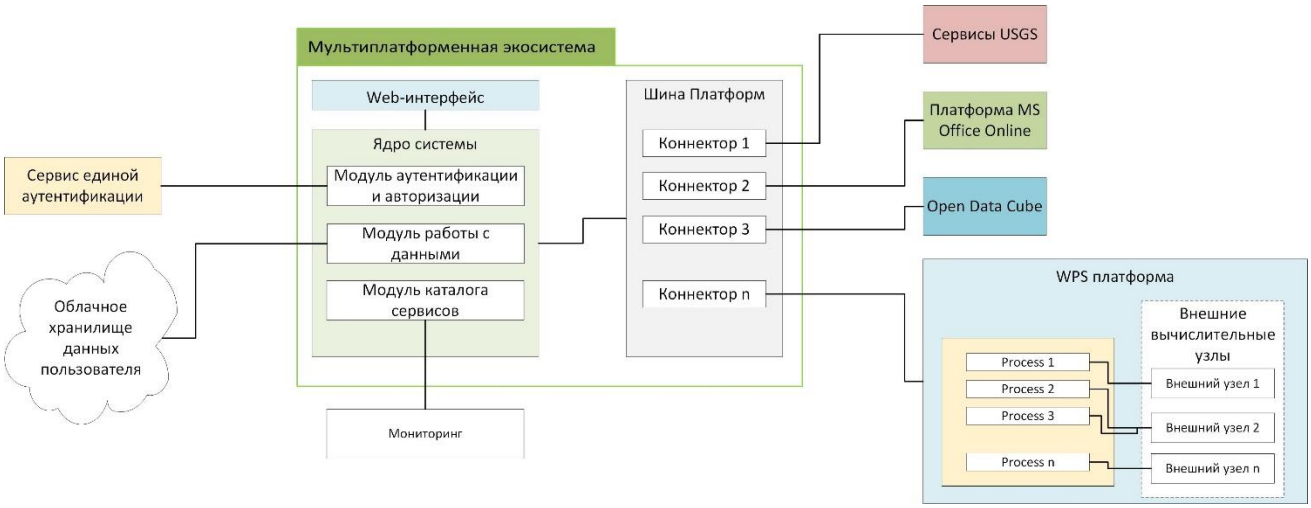


Рисунок 2. Схема предварительного проектирования системы

Такой подход подразумевает создание специализированных коннекторов для обеспечения взаимодействия с конкретным инструментом анализа данных. При этом все используемые данные, включая результаты обработки, передаются

между инструментами анализа через общую шину данных, как это реализуется в текущей версии системы.

Работы выполняются в рамках Государственного задания ГГМ РАН по Теме № 0140-2019-0005 «Разработка информационной среды интеграции данных естественнонаучных музеев и сервисов их обработки для наук о Земле», а также Государственной темы № 1021061009468-8-1.5.1 «Цифровая платформа интеграции и анализа геологических и музейных данных»

СПИСОК ЛИТЕРАТУРЫ

1. *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, Issue 6.

2. *Moser L., Thuraisingham B., Zhang J.* Services in the cloud // IEEE transactions on services computing. 2015. Vol. 8, No. 2.
https://doi.org/10.1007/978-3-319-41267-2_63

3. *Федотов А.М., Барахнин В.Б., Гуськов А.Е., Молородов Ю.И.* Распределенная информационно-аналитическая среда для исследований экологических систем // Вычислительные технологии. 2006. Т. 11. Спец. вып. С. 113–125.

4. *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); http://doi.org/10.1117/12.2249118

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

6. *Бычков И.В., Ружников Г.М., Фёдоров Р.К., Шумилов А.С.* Компоненты среды WPS-сервисов обработки геоданных // Вестник НГУ. Серия: Информационные технологии. 2014. Т. 12, Выпуск 3. С. 16–24.

7. *Eremenko V.S., Naumova V.V.* A multi-platform ecosystem for computing in Earth sciences // CEUR Workshop Proceedings. 2021. Vol. 3006. P. 67–73. http://doi.org/10.25743/SDM.2021.70.81.010

CREATING A DATA PROCESSING ECOSYSTEM FOR GEOLOGICAL RESEARCH

Vitaliy Eremenko[1][0000-0002-5250-5743], Vera Naumova[2][0000-0002-3001-1638]

Vernadsky State Geological Museum of the Russian Academy of Sciences, Moscow

[1] vitaer@gmail.com, [2] naumova_new@mail.ru

Abstract

This paper discusses heterogeneous geographically distributed computing systems for processing geological data and approaches to organizing interaction with these systems. The systems are classified by the authors into a number of groups based on the main functional capabilities and technological solutions. A description of the main properties for each type of systems is given, including possible ways for interaction.

An approach is proposed for organizing a single workspace with access to heterogeneous geographically distributed computing systems within the ecosystem developed by the authors. The architecture of the proposed solution and the rules of interaction for its participants are described. A software prototype is demonstrated that implements the described principles on the example of several heterogeneous systems for processing geological information.

Keywords*: computing and analytical environment, cloud services, web services, software platforms*

REFERENCES

1. 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, Issue 6.

2. Moser L., Thuraisingham B., Zhang J. Services in the cloud // IEEE transactions on services computing. 2015. Vol. 8, No. 2.

*https://doi.org/10.1007/978-3-319-41267-2_63*3. *Fedotov A.M., Barahnin V.B., Gus'kov A.E., Molorodov Yu.I.* Raspredelennaya informacionno-analiticheskaya sreda

dlya issledovanij ekologicheskih sistem // Vychislitel'nye tekhnologii. 2006. T. 11. Spec. vyp. S. 113–125.

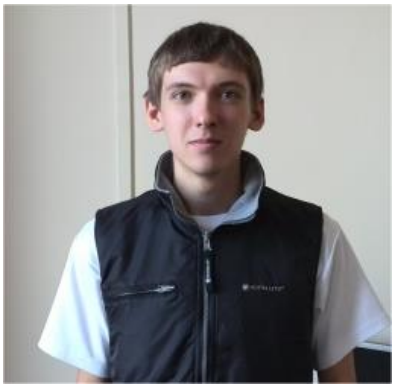
4. *Gordov E.P., Krupchatnikov V.N., Okladnikov I.G., and Fazliev A.Z.* Thematic virtual research environment for analysis, evaluation and prediction of global climate change impacts on the regional environment // Proc. SPIE 10035, 22nd International Symposium on Atmospheric and Ocean Optics: Atmospheric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

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

6. *Bychkov I.V., Ruzhnikov G.M., Fyodorov R.K., Shumilov A.S.* Komponenty sredy WPS-servisov obrabotki geodannyh // Vestnik NGU. Seriya: Informacionnye tekhnologii. 2014. T. 12, Vypusk 3. S. 16–24.

7. *Eremenko V.S., Naumova V.V.* A multi-platform ecosystem for computing in Earth sciences // CEUR Workshop Proceedings. 2021. Vol. 3006. P. 67–73. http://doi.org/10.25743/SDM.2021.70.81.010

СВЕДЕНИЯ ОБ АВТОРАХ



ЕРЁМЕНКО Виталий Сергеевич – младший научный сотрудник, Государственный геологический музей им. В.И. Вернадского РАН, Москва, Россия.

Vitaliy EREMENKO – Junior researcher of SGM scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

email: vitaer@gmail.com,

ORCID: 0000-0002-5250-5743

НАУМОВА Вера Викторовна – д. г.-м. н., г. н. с., зав. Научным отделом Государственного геологического музея им. В.И. Вернадского РАН, Москва.

Vera NAUMOVA – Prof., head of SGM scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

Email: naumova_new@mail.ru,

ORCID: 0000-0002-3001-1638

Материал поступил в редакцию 31 августа 2022 года