УДК 528.91
DOI: 10.33764/2618-981X-2020-1-2-82-87

ИСПОЛЬЗОВАНИЕ РАСПРЕДЕЛЕННЫХ СУБД ДЛЯ ОБРАБОТКИ ПРОСТРАНСТВЕННЫХ ДАННЫХ

Алексей Александрович Колесников
Сибирский государственный университет геосистем и технологий, 630108, Россия, г. Новосибирск, ул. Плахотного, 10, кандидат технических наук, доцент, тел. (913)725-09-28, e-mail: alexeykw@mail.ru

Елена Владимировна Комиссарова
Сибирский государственный университет геосистем и технологий, 630108, Россия, г. Новосибирск, ул. Плахотного, 10, кандидат технических наук, доцент, тел. (913)710-85-60, e-mail: komissarova_e@mail.ru

Иван Васильевич Жданов
Сибирский государственный университет геосистем и технологий, 630108, Россия, г. Новосибирск, ул. Плахотного, 10, магистрант, тел. (383)361-06-35, e-mail: zhdanov190496@gmail.com

В настоящий момент объемы данных увеличиваются в геометрической прогрессии. Геопространственные данные являются одним из основных элементов концепции больших данных. Существует очень большое количество инструментов для анализа больших данных, но далеко не все они учитывают особенности и обладают возможностями обрабатывать геопространственные данные. В статье рассматриваются три платформы с открытым исходным кодом, такие как Hadoop Spatial, GeoSpark, GeoFlink для работы с геопространственными данными очень больших объемов. Рассмотрены их архитектура, достоинства и недостатки, зависимость от времени выполнения и объема использованных данных. Также выполнена оценка обработки с точки зрения, как потоковых, так и пакетных данных. Эксперименты выполнялись на наборах растровых и векторных данных, представляющих собой спутниковые снимки в видимом диапазоне, индексы NDVI и NDWI, климатические показатели (снежный покров, интенсивность осадков, температура поверхности), данные из Open Street Map на территории Новосибирской и Иркутской областей.

Ключевые слова: распределенные СУБД, распределенная обработка, ДЗЗ, растровые данные, климатические данные.

USE OF DISTRIBUTED DBMS FOR SPATIAL DATA PROCESSING

Aleksey A. Kolesnikov
Siberian State University of Geosystems and Technologies, 10, Plakhotnogo St., Novosibirsk, 630108, Russia, Ph. D., Associate Professor, phone: (913)725-09-28, e-mail: alexeykw@mail.ru

Elena V. Komissarova
Siberian State University of Geosystems and Technologies, 10, Plakhotnogo St., Novosibirsk, 630108, Russia, Ph. D., Associate Professor, phone: (913)710-85-60, e-mail: komissarova_e@mail.ru

Ivan V. Zhdanov
Siberian State University of Geosystems and Technologies, 10, Plakhotnogo St., Novosibirsk, 630108, Russia, Graduate Student, phone: (913)725-09-28, e-mail: zhdanov190496@gmail.com

Currently, data volumes are growing exponentially. Geospatial data is one of the main elements of the concept of Big data. There is a very large number of tools for analyzing Big data, but not all of them take into account the features and have the ability to process geospatial data. The article discusses three popular open analytical tools Hadoop Spatial, GeoSpark, GeoFlink for working with geospatial data of very large volumes. Their architectures, advantages and disadvantages, depending on the execution time and the amount of data used are considered. Processing evaluations were also performed in terms of both streaming and packet data. The experiments were carried out on raster and vector data sets, which are satellite imagery in the visible range, NDVI and NDWI indices, climate indicators (snow cover, precipitation intensity, surface temperature), data from the Open Street Map in the Novosibirsk and Irkutsk Regions.

Key words: distributed DBMS, distributed processing, remote sensing, raster data, climate data.

Введение

Из-за огромного объема данных становится все сложнее их своевременно собирать, обрабатывать, анализировать и принимать решения. Одной из основных составляющих этого объема являются данные с пространственной привязкой. Например, архивы данных дистанционного зондирования EOSDIS увеличиваются со скоростью 4 ТБ в день. Архив спутниковых данных НАСА превышает 37 ПБ и постоянно увеличивается. Источниками этих данных являются аэрокосмическое дистанционное зондирование, цифровые камеры беспилотных летательных аппаратов, сенсоры интернета вещей, тахеометрическая съемка, лазерные сканеры и т. д. В таких случаях очень важно использовать специализированные инструменты, которые позволят своевременно и корректно обрабатывать данные даже очень большого объема [1–3]. Эти инструменты основываются на технологиях распределенного хранения и вычислений, map reduce и т.д. Платформы с открытым исходным кодом, такие как Apache Hadoop и Apache Spark, используют эти технологии и позволяют выполнять вычисления и анализ на наборах данных очень большого объема [4, 5]. Но, в варианте по умолчанию, эти платформы не ориентированы на обработку и не учитывают все особенности пространственных данных, в них также отсутствуют простые инструменты для интеграции в популярные настольные ГИС [6]. Целью исследования был базовый анализ существующих сегодня инструментов с открытым исходным кодом для распределенной обработки пространственных данных, выделение их особенностей, достоинств, недостатков, оценка эффективности применения на примере обработки набора данных дистанционного зондирования и векторной картографической основы на территории Новосибирской и Иркутской областей.

Методы и материалы

Традиционные подходы используют мощность вычислительных станций для обработки данных, но при этом они могут масштабироваться только вертикально (что всегда затратно и возможности сильно ограничены аппаратной платформой) и поэтому в какой-то момент физически не могут справиться

с непрерывным ростом объема обрабатываемых данных [7–9]. Эту проблему чаще всего решают с помощью технологий распределенных вычислений, поскольку они практически неограниченно масштабируемы горизонтально [9, 10]. В этом подходе огромный массив данных разбивается на более мелкие части, что позволяет получить распределенное хранилище. По тому же принципу вычисления разделяются на отдельную обработку каждой из частей данных, и каждый узел выполняет свою подзадачу, в итоге объединение результатов всех узлов дает конечный результат, который возвращается приложению, таким образом достигается параллельная и распределенная обработка [11–13]. Для оценки эффективности такой технологии были взяты наиболее популярные программные пакеты для распределенной обработки пространственных данных с открытым исходным кодом: Hadoop Spatial, GeoSpark и GeoFlink [13,14].

Поскольку распределенная обработка изначально подразумевает большой объем исходных данных, то в качестве базовой файловой системы для этих пакетов была использована распределенная файловая система Hadoop (HDFS). Ее управляющий узел разбивает весь входящий поток данных на фрагменты, которые максимально равномерно распределяются по всему имеющемуся пулу хранилищ. Обработка данных выполняется отдельно для каждого фрагмента на том сервере, где он хранится модулем mapper (фаза map), затем полученные результаты объединяются модулем reducer (фаза reduce) и только окончательный результат отправляется на управляющий узел [15–18]. Схематично эти процессы представлены на рисунке 1. Такой подход реализуют все выбранные программные пакеты, поскольку объем передаваемой для обработки программы практически всегда значительно меньше, чем объем обрабатываемых данных [19–21].

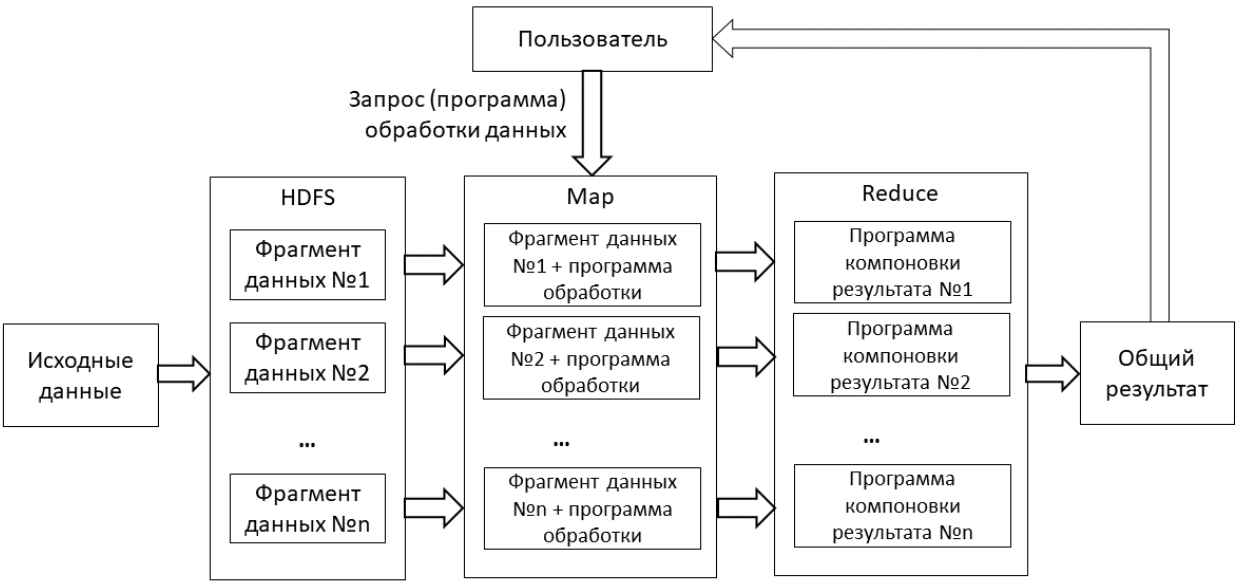

Рис. 1. Схема используемой технологии распределенной обработки данных

Для обработки были использованы наборы растровых данных со спутников Sentinel-2 и MODIS объемами 4 и 36 гигабайт на территории Новосибирской и Иркутской областей. Были использованы данные видимого диапазона, вычисленные индексы NDVI и NDWI сервиса Sentinel Hub, климатические показатели (снежный покров, интенсивность осадков, температура поверхности) на основе данных MODIS. На каждом фрагменте Sentinel-2 векторизуются данные относящиеся к гидрографии. Из данных MODIS для каждого фрагмента извлекается суммарное значение показателя.

Результаты исследования

В результате исследования были выявлены следующие особенности программного обеспечения. Spatial Hadoop, поскольку является только расширением базовой функциональности Hadoop, упрощает переход от стандартных типов данных HDFS к обработке пространственных данных, но содержит, в большинстве случаев, только базовые операции над векторными данными. GeoSpark базируется на инструментарии Apache Spark, что обеспечивает гораздо большую скорость выполнения операций по сравнению со стандартными методами Hadoop (и, соответственно, Spatial Hadoop), поддерживает работу со сторонними библиотеками, в том числе для пространственного анализа и машинного обучения (базовый же набор функций практически идентичен Spatial Hadoop). Основное отличие GeoFlink от GeoSpark и Spatial Hadoop это возможность обработки потоковых пространственных данных, позволяя обрабатывать, например, данные метеостанций в реальном времени и генерировать ежедневные отчеты.

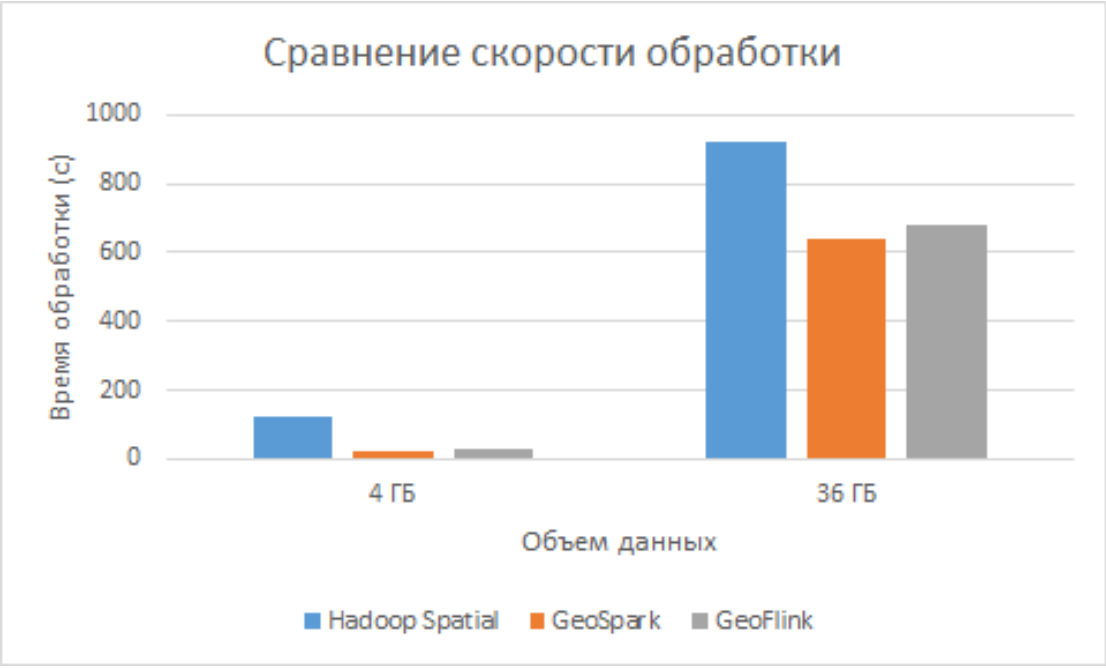

Рис. 2. Сравнение скорости обработки данных

Обсуждение

Различия в результатах обусловлены тем, что GeoFlink и GeoSpark максимально задействуют оперативную память для расчетов и когда ее объем заканчивается (для экспериментов использовались серверы с объемом памяти 2 Гб), то задействуется дисковый накопитель и происходит значительное уменьшение скорости обработки. Эту проблему можно решить подбором оптимальных настроек размера обрабатываемого фрагмента. Что и планируется провести в дальнейших исследованиях. Кроме этого, следующим шагом будет рассмотрение возможностей визуализации и управления задачами из настольных ГИС. В рамках исследования для визуализации использовался программный пакет HadoopViz.

Заключение

В результате анализа различных программных пакетов для распределенной обработки пространственных данных была доказана эффективность такого подхода по сравнению с локальной обработкой. Основными проблемами его внедрения являются сложность в настройке и формировании задач для вычислений и недостаточность инструментов интеграции с наиболее популярными геоинформационными системами. Поскольку большинство стандартных инструментов для обработки и визуализации данных проанализированных программных пакетов используют web-интерфейс, то наиболее перспективным направлением является их интеграция с Web-ГИС.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Yu J., Wu J., Sarwat M. Geospark: a cluster computing framework for processing large-scale spatial data // В сборнике: 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2015. С. 70.

2. Eldawy A., Mokbel M. F. Spatial Hadoop: a map reduce framework for spatial data // В сборнике: IEEE 31st International Conference on Data Engineering. 2015. С. 1352–1363.

3. Dubey H., Samaddar. A. B., Gupta R. D., Barik R. K., Ray P.K. FogGIS: Fog computing for geospatial big data analytics // В сборнике: 3rd IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics. 2016. С. 613-618.

4. Wang A. F., Vo H., Lee R., Liu Q., Zhang X., Saltz J. Hadoop GIS: a high performance spatial data warehousing system over map reduce // В сборнике: VLDB Endow. 2013. Т. 6. № 11. С. 1009-1020. DOI: 10.14778/2536222.2536227

5. Yu J., Zhang Z., Sarwat M. Spatial data management in Apache Spark: the GeoSpark perspective and beyond // GeoInformatica. 2019. Т. 23, № 1. С. 37–78. DOI: 10.1007/s10707-018-0330-9

6. Рыкин И.С., Паниди Е.А. Определение границ и продолжительности вегетационных сезонов по данным рядов спутниковых съёмок с высоким временным разрешением - применение облачных вычислений. В сборнике: Геоматика: образование, теория и практика материалы международной научно-практической конференции, посвященной 50-летию кафедры геодезии и космоаэрокартографии и 85-летию факультета географии и геоинформатики БГУ. отв. ред. А. П. Романкевич. 2019. С. 118-122.

7. Lee J., Kang M. Geospatial big data: challenges and opportunities // Big data research. 2015. Т. 2. № 2. С. 74–81.

8. Ma Y., Wu H., Wang L., Huang B., Ranjan R., Zomaya A., Jie W. Remote sensing big data computing: challenges and opportunities // Future Generation Computer Systems. 2015. Т. 51, С. 47–60.

9. Jin X., Wah B.W., Cheng X., Wang Y. Significance and challenges of big data research // Big data research. 2015. Т. 2. №. 2. С. 59–64.

10. Hughes J. N., Annex A., Eichelberger C. N., Fox A., Hulbert A., Ronquest M. GeoMesa: a distributed architecture for spatio-temporal fusion // Geospatial Informatics, Fusion, and Motion Video Analytics V. 2015. Т. 9473. С. 128–140.

11. You S., Zhang J., Gruenwald L. Large-scale spatial join query processing in cloud // В сборнике: 31st IEEE International Conference on Data Engineering Workshops. 2015. С. 34–41.

12. Baig F., Vo H., Kurc T. M., Saltz J. H., Wang F. SparkGIS: Resource aware efficient in-memory spatial query processing // В сборнике: 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2017, Redondo Beach, CA, USA. 2017. С. 28:1–28:10. DOI: 10.1145/3139958.3140019

13. Zhang F., Zheng Y., Xu D., Du Z., Wang Y., Liu R., Ye X. Realtime spatial queries for moving objects using storm topology // ISPRS International Journal of Geo-Information. 2016. Т. 5, № 10. С. 178.

14. Hadjieleftheriou M., Manolopoulos Y., Theodoridis Y., Tsotras V. J. R-Trees: A Dynamic Index Structure for Spatial Searching // Cham: Springer International Publishing. 2017. С. 1805–1817.

15. Sharma T., Shokeen V., Mathur S. Distributed processing of satellite images on hadoop to generate normalized difference vegetation index images // International Conference on Computing, Communication, Control and Automation (ICCUBEA). 2017. С. 1-5.

16. Потапов В.П., Попов С.Е., Костылев М.А. Информационно-вычислительная система массивно-параллельной обработки радарных данных в среде Apache Spark. // Вычислительные технологии. 2018. Т. 23. № 4. С. 110-123.

17. Xu C. Big data analytic frameworks for GIS (Amazon EC2, Hadoop, Spark) // Comprehensive Geographic Inform. Syst. 2017. Т. 1. С. 148-152.

18. Безпалов В.В., Лочан С.А., Федюнин Д.В., Иванов А.В., Автономова С.А. Большие данные и возможности их использования при разработке коммуникативной стратегии предприятий регионального промышленного комплекса // Вестник Алтайской академии экономики и права. 2020. № 1-2. С. 28-34.

19. Ахметьянова А.И., Исмагилова А.С. Преимущества и использование науки о данных и больших данных // В сборнике: Наука о данных Материалы международной научно-практической конференции. 2020. С. 35-37.

20. Решетников В.И., Голубчиков Е.А., Пятлин А.В., Кузин А.К., Киев В.А., Шабров Н.Н., Журавлёв А.С., Гусева Е.К. Применение средств анализа больших данных в проблеме визуализации результатов решения задач газодинамики большой размерности // Научная визуализация. 2020. Т. 12. № 1. С. 83-89.

21. Григорьев Ю.А., Пролетарская В.А. Модель процессов выполнения запросов к хранилищу данных на платформе параллельных вычислений Spark // Информатика и системы управления. 2019. № 1 (59). С. 3-17.