Облачные вычисления для больших геопространственных данных

Cloud computing for big geospatial data

Буравцев / Buravtsev A.
Алексей Владимирович
(mister_j@mail.ru)
ФГБОУ ВО "МИРЭА – Российский технологический университет", заместитель директора.
г. Москва

Цветков / Tsvetkov V.
Виктор Яковлевич
(cvj2@mail.ru)
доктор технических наук, профессор,
почетный работник науки и техники РФ,
почетный работник высшей школы РФ.
АО «Научно-исследовательский и проектно-конструкторский институт информатизации, автоматизации и связи на железнодорожном транспорте»,
заместитель руководителя.
г. Москва

Ключевые слова: пространственные данные – spatial data; геопространственные данные – geospatial data; облачные вычисления – cloud computing; базы данных – databases; распределенные системы – distributed systems; геоданные – geodata.

Большие данные широко представлены при исследовании Земли из космоса при наземных исследованиях. Это дает основание говорить о больших пространственных данных, о больших геопространственных данных или о больших геоданных. В статье приведен анализ применения, облачных вычислений для решения задач обработки больших данных. Этот подход применим при решении других задач в области применения больших пространственных данных, таких как умные города.

Big data are widely represented in the study of the Earth from space during ground studies. This gives reason to talk about large spatial data, large geospatial data or large geodata. The article analyzes the use of cloud computing for solving problems of big data processing. It is shown that this approach supports the life cycle of big data processing, including management, access, analytical data extraction, modeling and forecasting. This approach is applicable in solving other problems in the application of large spatial data, such as smart cities.

Введение

Еще в 1967 году при получении фотоснимков с ИСЗ выяснилось, что для их обработки с использование вычислительных ресурсов того времени понадобится 5–10 лет [1]. Тогда эта проблема была специальной и о ней мало говорили. Но по существу эта проблема возникла впервые именно в области пространственных данных. Появление термина «большие данные» связывают с 2008 годом и работами Клиффорда Линча – редактора журнала Nature [2]. Но это означало лишь признание данной проблемы в широких слоях общества, в то время как в области вычислений и дистанционном зондировании эта проблема существовала задолго до этого. Появившаяся технология облачных вычислений [3] стала новым инструментом для обработки больших данных, включая большие геопространственные данные.

Наблюдение за Землей и имитационное моделирование создают ежедневно от терабайт до петабайт данных [4]. Нетрадиционные методы сбора пространственных данных, такие как социальные медиа, телефонные разговоры и беспилотные летательные аппараты, получают геопространственные данные с еще более высокими скоростями. В дополнение к большому объему (Volume) [5] геопространственные данные существуют в разнообразных (Variety) формах и форматах для различных приложений, их точность и неопределенность (Veracity) распространяются в широком диапазоне, при этом данные производятся с большой скоростью (Velocity) датчиками реального времени [6] (рис.1). Сами по себе большие данные бесполезны, если их нельзя обработать или привести в вид, удобный для анализа.

Рис. 1 показывает, что технология облачных вычислений создает необходимую поддержку обработки больших данных с помощью четырех V с целью получения «Ценности» для поддержки принятия решений, исследований и оптимизации операций в различных областях.

Благодаря уникальной информации и знаниям большие геопространственные данные могут быть использованы для повышения ценности научных исследований, разработки глобальных проектов и принятия бизнес-решений [7]. Они призваны обеспечить инновационное развитие, улучшение жизни и лучшее понимание геосистем. Это становится возможным, когда происходит трансформация из четырех V в пятую интеграционную V (ценность).

Однако такие преобразования создают серьезные проблемы для аналитики, разработки системной архитектуры и моделирования [8]. Например, первая проблема

заключается в том, как бороться с разнообразием и недостоверностью больших данных, чтобы создать объединенный набор данных, который может быть использован в единой системе поддержки принятия решений [9]. Другая проблема заключается в том, как обеспечить время обработки больших данных, сопоставимое со временем обработки обычных данных. Как создать масштабируемую и расширяемую вычислительную мощность, не зависящую от флуктуации передачи данных [10]. Поддержка аналитических функциональных возможностей или актуализация данных также создает серьезные проблемы для получения ценности от больших геоданных [11].

Технология облачных вычислений появилась в качестве новой парадигмы для предоставления вычислительной мощности в виде сервисной службы с пятью выгодными характеристиками (рис. 1, нижние два уровня):

а) быстрая и эластичная вычислительная мощность;

б) объединенная вычислительная мощность для эффективного и совместного использования ресурсов;

в) широкополосный сетевой доступ для быстрой связи;

г) доступ по требованию для вычислений в качестве сервисных служб;

д) pay-as-you-go (оплата по мере использования) за выделяемые мощности, используемые без существенных первоначальных затрат.

Эти характеристики облачных вычислений являются существенным преимуществом перед обычными вычислительными технологиями [3]. Сервисно-ориентированная архитектура применяется в облачных вычислениях и включает в себя специальный сервис «все как услугу», в частности инфраструктуру как услугу (IaaS), платформу как услугу (PaaS) и программное обеспечение как услугу (SaaS). Пересматривая возможности геоинформатики и концепции "цифровая Земля" (Digital Earth), облачные вычисления с использованием больших геоданных создают условия решения проблем с большими геопространственными данными в геодезических науках и в науках о Земле.

В настоящее время использование облачных вычислений для решения проблем больших данных находится в состоянии развития. Сложность задачи состоит в том, как пять существенных характеристик облачных вычислений могут помочь первым четырем *V* больших данных для получения ценности – пятого *V*.

Рассмотрим, как облачные вычисления поддерживают подобное преобразование на примере исследования климата. Этот опыт может быть легко использован в других областях, связанных с исследованиями окружающей среды и исследованиями технологий развития городов, такими как умные города [12]. Жизненный цикл больших геопространственных данных поддерживает следующие технологии: управление данными, интеллектуальный анализ данных, моделирование явлений, развитие сценариев. Проблемы решения больших данных описываются ключевыми характеристикам облачных вычислений.

Использование облачных вычислений для анализа климата

Изменения климата, такие как увеличение частоты проливных дождей и увеличение затопления рек, все чаще ставят под угрозу городскую инфраструктуру. Дополнительная деятельность человека (например, сжигание ископаемого топлива) в течение последних

Рис. 1. Технология облачных вычислений и больших данных

50 лет сильно повлияло на глобальный климат. Чтобы понять изменение климата и его воздействие на экологические и городские проблемы, были исследованы большие данные о климате. Эти данные были получены на основе ретроспективного анализа, текущих наблюдений и прогнозных данных. Например, типовой отчет Межправительственной группы экспертов по изменению климата (МГЭИК) содержит более 100 петабайт данных, а аналогичный отчет NASA на тему климатических данных к 2030 году имеет объем более 300 петабайт [13]. Естественно, что не только анализ, а попытка обозреть такой объем данных создает трудности для человеческого интеллекта. Эти данные различаются по форме, пространственно-временной разрешающей способности и цели исследования [14].

В то же время только "Большие данные" могут помочь улучшить понимание климатических явлений и помочь в определении того, как можно изменить последствия изменения климата для общества и экосистем, такие как обнаружение глобальных аномалий температуры и исследование пространственно-временного распределения экстремальных погодных явлений. Это особенно важно для сильно населенных регионов, например в мегаполисах [15, 16]. При использовании больших данных о климате возникает несколько проблем:

· объем и скорость больших климатических данных намного превышают емкость и вычислительную способность автономного компьютера;

· разнообразие климатических данных в формате и пространственно-временных разрешающих способностей затрудняет поиск удобного инструмента для анализа климатических данных;

· достоверность имитационного моделирования представляет собой проблему для ученых-климатологов из-за неопределенностей и качественно разных моделей.

Задачи, комбинирующие сложность объема, скорость, разнообразие и правдоподобие, могут быть решены с помощью облачных стратегий управления данными и сервисно-ориентированной архитектуры аналитических данных, чтобы помочь обрабатывать, анализировать и собирать климатические данные.

Расширенный пространственно-временной индекс для управления большими климатическими данными

Расширенным пространственно-временным индексом для управления большими климатическими данными называют специальную технологию, которая идеологически заимствована из теории баз данных и информационно поисковых систем.

Особенность климатических данных состоит в том, что сотни петабайт климатических данных могут управляться только в распределенной и масштабируемой вычислительной среде. Облачные вычисления являются именно такой средой. Облачные вычисления могут помочь в следующем: предоставление гибких виртуальных (VM) машин по требованию в соответствии с объемом климатических данных; автоматическое развертывание HDFS-распределенной файловой системы Hadoop на виртуальных машинах для создания распределенной файловой системы.

Большие данные в этом сервисе можно хранить в исходном формате, а не в текстовом, для уменьшения

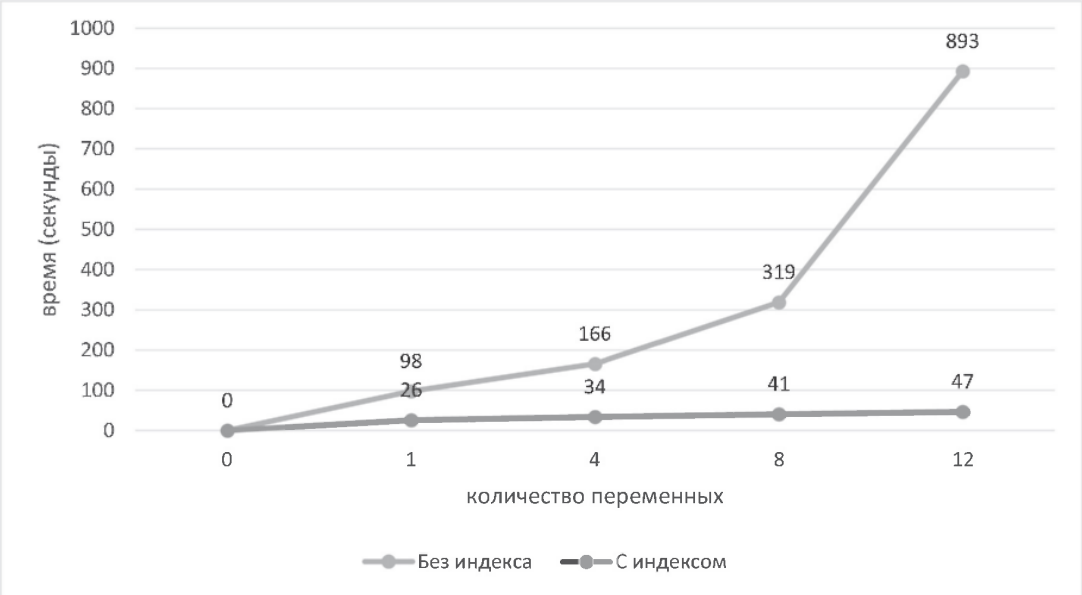

Рис. 2. Время выполнения ежедневного среднего глобального вычисления для разного количества переменных о климате

дискового пространства. Архитектура логических облачных данных построена для облегчения быстрой идентификации, доступа и анализа [17, 18]. Основная архитектура представляет собой пространственно-временной индекс для многомерных климатических данных, хранящихся на Hadoop Distributed Filesystem (HDFS). Индекс отображает содержимое данных на уровни байтов, файлов и узлов в HDFS. Девять компонентов используются для индекса и включают в себя: данные о пространстве, времени и форме, описывающие логическую информацию сетки данных, которая коррелирует с запросом данных, смещение в байтах, длину в байтах, код сжатия, список узлов и путь к файлу, идентифицирующий конкретное местоположение на HDFS. Этот индекс позволяет пользователям напрямую находить и получать данные с точным пространственно-временным и содержательным описанием.

Атрибуты пространства и времени в пространственно-временном индексе будут идентифицировать решетки, перекрывающиеся пространственно-временным ограничивающим прямоугольником. Атрибут списка узлов используется для доставки вычислительных программ к узлу, где хранятся сетки. Затем вычислительные программы могут считывать данные в виде потока данных с высокой локализацией данных в соответствии со смещением в байтах, длиной в байтах и атрибутами кода сжатия. Атрибуты формы и типа данных могут использоваться для изменения потока данных в многомерном массиве. Месячные данные Modern-Era Retrospective analysis for Research and Applications (MERRA) (около 90 ГБ) используются для оценки эффективности управления большими данными на основе облачных вычислений и пространственно-временной индексации. В этом эксперименте анализируется среднемесячное значение конкретных климатических переменных (путем изменения их количества) в заданном пространственно-временном диапазоне от 36 кластеров HDFS на основе виртуальных машин, связанных 1 гигабитным (Gbps) каналом. Каждый узел имеет следующую конфигурацию: восьми-ядерный процессор (2,60 ГГц), 16 ГБ оперативной памяти и операционная система CentOS 6.5. Результаты с использованием индекса и без него (рис. 2) [6] показывают, когда количество обрабатываемых переменных увеличивалось, то время выполнения без индекса тоже возрастало примерно в 9,1 раза, тогда как время выполнения с индексом увеличивалось только в 1,8 раза.

На основании временных ограничений, изменяемое количество виртуальных машин может быть предоставлено по требованию для завершения задач в течение определенного периода времени [17, 18, 19]. Таким образом, обслуживание по требованию и эластичность в сочетании с высоким уровнем управления эффективно учитывают большие климатические данные и аналитические требования.

Сервис «Все как услуга» для облегчения экспериментов по моделированию климата

Моделирование климата создает проблемы в связи с требованием большого количества вычислительных ресурсов для научных экспериментов. Эти ресурсы необходимы при имитационном анализе больших данных или при использовании большого количества имитационных моделей с различными входными данными модели. Технология облачных вычисление решает эту проблему следующим образом:

Рис. 3. Облачная сервисно-ориентированная система для изучения климатических моделей

· климатические модели могут быть опубликованы как услуга (MaaS – Model as a Service), и необходимое количество виртуальных машин может быть выделено по требованию конкретными конфигурациями моделей для каждого цикла моделирования групп данных;

· приложение развертывается как служба [20] с веб-порталом для поддержки работы модели и мониторинга;

· рабочий процесс с использованием различных аналитических средств и интуитивно понятным графическим интерфейсом используется как услуга (WaaS) [21].

Большая аналитика данных климата поддерживается технологией облачных вычислений на уровне вычислительной инфраструктуры. Архитектура облачной сервисно-ориентированной системы потока работ (процессов) для исследования климатического модуля включает в себя (рис. 3) [6]:

– службу моделирования данных, которая отвечает за компиляцию и запуск моделей на виртуальных машинах, которые предоставляются на основе моментального снимка системы, содержащей тематическую программную среду для запуска модели;

– службу мониторинга виртуальной машины, которая обеспечивает облачную платформу информацией о ее статусе для планирования ресурсов;

– службу анализа данных, которая подает выходные данные модели как входные данные для аналитики, одновременно анализируя данные параллельно во избежание проблем с производительностью.

Служба публикации данных позволяет пользователям получать доступ к результатам анализа в реальном времени через Интернет. Все эти службы управляются с помощью графического интерфейса пользователя, который позволяет ему одновременно перетаскивать и подключать сервисы для создания сложного рабочего процесса, чтобы система могла автоматически перейти к приложениям, указанным в рабочем процессе и работать в облаке с автоматически создаваемыми виртуальными машинами. В качестве примера можно привести исследования Ли [18], который построил ModelE как услугу, и эксперимент показал, что этот облачный метод уменьшил потребление времени в 10 раз по сравнению с традиционным методом.

Проблемы в исследованиях климата, решаемые технологией облачных вычислений, сводятся к следующим. Во-первых, большой объем климатических данных наблюдения и моделирования хранится в распределенной и масштабируемой среде, обеспечиваемой облачной платформой. Во-вторых, проблема разнообразия климатических данных рассматривается с использованием пространственно-временного индекса, чтобы объединить эти данные с аспектами пространства и времени. В-третьих, проблема разнообразия климатических моделей облегчается путем создания сервисно-ориентированной системы для упрощения настройки модели, ее запуска и анализа выходных данных. Эти методы могут быть распространены на другие геопространственные области, которые включают в себя данные высокой размерности и сложные модели, такие как дистанционное зондирование, обработка изображений и агент-моделирование экологических и городских событий.

Заключение

Большие геопространственные данные создают серьезные проблемы в обработке и обслуживании, но зато обеспечивают исчерпывающей информацией для поддержки принятия решений. Большие данные имеют свой жизненный цикл, в течение которого с ними производят следующие манипуляции: хранение, управление, анализ с их помощью, моделирование самих данных и моделирование процессов с помощью данных. Моделирование самих данных является особенностью больших геоданных и технической проблемой. Эта процедура имеет аналог в обновлении баз данных. Обновление моделей данных несколько иная процедура, которая присутствует в больших данных. Облачные вычисления помогают решать проблемы четырех *V* для получения ценности (Value) с пятью преимуществами технологии облачного вычисления: самообслуживание по требованию, широкополосный сетевой доступ, пул ресурсов, быстрая эластичность и измеряемое обслуживание. Задача хранения и управления большими геопространственными данными остается приоритетной. Существуют вопросы оптимизации различных СУБД в облачной среде для распределенного хранения и обработки. Организация больших геопространственных данных занимает промежуточное положение между организацией баз данных и коллекциями данных. Пространственно-временная обработка больших данных требует обработки в реальном времени, извлечения информации и автоматизации для извлечения информации и знаний. Необходимо разработать более масштабируемые методы пространственно-временного извлечения данных, чтобы использовать преимущества эластичных хранилищ и вычислительных ресурсов облачных платформ. Методы пространственно-временного анализа имеют большое значение, и они нуждаются в дальнейшем развитии и формализации для оптимизации облачных вычислений при обработке больших геопространственных данных. Использование технологий облачных вычислений для больших данных позволяет решать сложные задачи, такие как интеллектуальные города и интеллектуальные сообщества, аналитическое извлечения пространственных данных и извлечение знаний.

Литература

1. Космические исследования земных ресурсов. Методы и средства измерений и обработки информации. – М.: Наука, 1976. – 386 с.

2. Lynch, C. Bigdata: How do your data grow? / C. Lynch // Nature. – 2008. – Vol. 455, No. 7209. – P. 28–29.

3. Облачные платформы и сервисы / И.П. Дешко [и др.]. – М.: МАКС Пресс, 2016. – 80 с.

4.Geospatial cyberinfrastructure: Past, present and future / C. Yang [et al.] // Computers, Environment and Urban Systems. – 2010. – Vol. 34 (4). – P. 264–277.

5. Marr, B. Big Data: Using SMART Big Data. Analytics and metrics to make better decisions and improve performance / B. Marr. – Wiley, 2015. – 258 p.

6. Utilizing Cloud Computing to address big geospatial data challenges / C. Yang [et al.] // Computers, Environment and Urban Systems. – 2017. – Vol. 61. – P. 120–128.

7. Lee, J.G. Geospatial Big Data: Challenges and opportunities / J.G. Lee, M. Kang // Big Data Research. – 2015. – Vol. 2 (2). – P. 74–81.

8. Big Data and Cloud Computing: Innovation opportunities and challenges / C. Yang [et al.] // International Journal of Digital Earth. – 2017. – Vol. 10. – P. 13–53.

9. Kim, G. H. Big-data applications in the government sector / G.H. Kim, S. Trimi, J.H. Chung // Communications of the ACM. – 2014. – Vol. 57 (3). – P. 78–85.

10. Ammn, N. Big Data challenges / N. Ammn, M. Irfanuddin. // International Journal of Advanced Trends in Computer Science and Engineering. – 2013. – Vol. 2 (1). – P. 613–615.

11. Fan, J. Statistical analysis of Big Data on pharmacogenomics / J. Fan, H. Liu // Advanced Drug Delivery Reviews. – 2013. – Vol. 65 (7). – P. 987–1000.

12. Batty, M. Big Data, smart cities and city planning / M. Batty // Dialogues in Human Geography. – 2013. – Vol. 3 (3). – P. 274–279

13. Skytland, N. Big Data: What is NASA doing with Big Data today [Электронный ресурс] / N. Skytland. – Режим доступа: https://open.nasa.gov/blog/what-is-nasa-doing-with-big-data-today, свободный. – Загл. с экрана.

14. MERRA analytic services: Meeting the Big Data challenges of climate science through cloud-enabled climate analytics-as-a-service / J.L. Schnase [et al.] // Computers, Environment and Urban Systems. – 2017. – Vol. 61, P. B. – P. 198–211.

15. Das, M. Anomaly detection and spatio-temporal analysis of global climate system / M. Das, S. Parthasarathy // Proceedings of the third international workshop on knowledge discovery from sensor data ACM. – P. 142–150.

16. Debbage, N. The urban heat island effect and city contiguity / N. Debbage, J.M. Shepherd // Computers, Environment and Urban Systems. – 2015. – Vol. 54. – P. 181–194.

17. Automatic scaling Hadoop in the cloud forefficient process of big geospatial data / Z. Li [et al.] // ISPRS International Journal of Geo-Information. – 2016. – Vol. 5 (10). – P. 173.

18. A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce / Z. Li [et al.] // International Journal of Geographical Information Science. – 2017. – Vol. 31. – P. 17–35.

19. Contemporary computing technologies for processing big spatiotemporal data / C. Yang [et al.] // Space-time integration in geography and GIScience / ed. by Mei-Po Kwan [et al.]. – Netherlands.: Springer, 2015. – P. 327–351.

20. Lushbough, C. M. Life science data analysis workflow development using the bioextract server leveraging the iPlant collaborative cyberinfrastructure / C.M. Lushbough, E.Z. Gnimpieba, R. Dooley // Concurrency and Computation: Practice and Experience. – 2015. – Vol. 27 (2). – P. 408–419.

21. Kramer, M. A modular software architecture for processing of big geospatial data in the cloud / M. Kramer, I. Senner // Computers & Graphics. – 2015. – Vol. 49. – P. 69–81.