УДК 004.75 DOI: 10.21782/GIPR0206-1619-2016-6(156-159)

Ю. В. АВРАМЕНКО[1], А. С. ШУМИЛОВ[2]

[1] Иркутский научный центр СО РАН,
664033, Иркутск, ул. Лермонтова, 134, Россия, avramenko@icc.ru
[2] Институт динамики систем и теории управления им. В. М. Матросова СО РАН,
664033, Иркутск, ул. Лермонтова, 134, Россия, alexshumilov@yahoo.com

СПЕЦИФИКАЦИИ ОБРАБОТЧИКОВ РАСТРОВЫХ ИЗОБРАЖЕНИЙ В РАМКАХ МОДЕЛИ MAPREDUCE

По мере развития информационных технологий возрастают объемы обрабатываемой информации, что требует как увеличения аппаратных мощностей, так и нахождения новых подходов к более эффективной обработке данных. В статье предлагается метод обработки растровых изображений на основе использования модели распределенных вычислений MapReduce. Модель MapReduce предполагает разбиение исходного массива данных на части с помощью операции Map, отправку частей на обработку и сбор результатов с помощью операции Reduce.

Сервис-ориентированная среда распределенных сервисов Института динамики систем и теории управления (ИДСТУ) СО РАН также сталкивается с проблемами обработки больших массивов данных, в частности растровых. Для повышения скорости обработки растровых пространственных данных в пределах сервис-ориентированной среды было реализовано распределение растровых изображений между узлами вычислительной сети с помощью обработчиков для операций Map и Reduce. Распределение фрагментов растровых изображений осуществляется с помощью спецификаций — правил распределения и сбора обработанных данных.

Механизм создания и использования спецификаций интегрирован в информационную систему Геопортала ИДСТУ СО РАН. Геопортал дает возможность централизованно выполнять распределенные сервисы. Использование спецификаций при выполнении сервисов позволяет эффективно применять доступные вычислительные мощности.

Предлагаемый метод позволяет использовать инструменты пространственного анализа растровых изображений в распределенной вычислительной среде без их модификации. Выполнение распределенных сервисов, работающих с большими объемами растровых данных, в рамках модели MapReduce уменьшает время выполнения сервисов при максимальном использовании имеющихся аппаратных мощностей.

Ключевые слова: *MapReduce, WPS, SVM, GEOTIFF, растровые данные, обработка изображений.*

YU. V. AVRAMENKO[1] AND A. S. SHUMILOV[2]

[1] Irkutsk Scientific Center SB RAS,
664033, Irkutsk, Lermontova str., 134, Russia, avramenko@icc.ru
[2] V. M. Matrosov Institute for System Dynamics and Control Theory SB RAS,
664033, Irkutsk, Lermontova str., 134, Russia, alexshumilov@yahoo.com

SPECIFICATIONS OF RASTER IMAGES PROCESSORS WITHIN THE MAPREDUCE MODEL

As the information technologies are actively developing, the volume of data that needs to be processed constantly increases, which requires keeping hard- and software technologically advanced and finding new approaches to data processing. Based on the distributed computations model MapReduce, the original method of raster images processing is proposed in this paper. The MapReduce model proposes to split initial dataset into pieces using Map operation, process these data pieces and gather all results using Reduce operation.

Service-oriented distributed environment of ISDCT SB RAS also faces problems of large data volumes processing, particularly processing the raster images. In order to increase the processing speed of spatial data within the service-oriented infrastructure, the distribution of raster images among computational nodes was organized. Mapping of the raster images is implemented using the specifications — sets of rules of how the data should be split and gathered.

The mechanism of definition and application of specifications is implemented as a part of ISDCT SB RAS Geoportal. The Geoportal allows executing distributed services in a centralized way. The use of specification during the service execution allows to effectively utilize the available computational resources.

The proposed approach allows using the instruments for spatial analysis of raster images within the distributed environment without theirs modification. Execution of distributed services that work with large volumes of spatial data within the MapReduce model allows decreasing the overall services execution time and using available computing resources at higher rates.

Keywords: *MapReduce, WPS, SVM, GEOTIFF, spatial data, image processing.*

ВВЕДЕНИЕ

В области геоинформационных технологий происходит рост объемов доступной для обработки информации. Часто возникает ситуация, когда увеличение объемов обрабатываемых данных затрудняет их обработку в силу значительных временных затрат или более высоких требований к аппаратной части. Для сокращения времени выполнения вычислений или более рационального использования аппаратных мощностей разрабатываются новые и развиваются существующие подходы выполнения распределенных вычислений.

Один из наиболее популярных подходов обработки данных основан на использовании модели распределенных вычислений MapReduce [1]. Свободно доступная реализация MapReduce — Apache Hadoop [2] позволяет осуществлять контроль и управление вычислительными узлами, а также предоставляет такие средства, как распределенную файловую систему (единое файловое пространство для выполняемых сервисов) и собственную распределенную СУБД. Расширение Spatial Hadoop [3] ориентировано на работу с большими массивами данных. Особый интерес представляет гибридный подход Hadoop DB [4], состоящий из распределенной СУБД Postgre SQL и Apache Hadoop. Суть Hadoop DB состоит в связывании нескольких одноузловых систем баз данных с использованием Apache Hadoop в качестве координатора задач и сетевого коммуникационного слоя MapReduce.

В работе [5] авторы предлагают на основе Apache Hadoop и MapReduce систему обработки изображений с автоматическим распараллеливанием данных между вычислительными узлами. Система состоит из двух частей Apache Hadoop и реализованных обработчиков изображений с аппаратно программным интерфейсом (API), встроенных в пакет Image Processing Library. Система ориентирована на обработку коллекции независимых изображений. В статье [6] авторы использовали Hadoop-GIS для обработки пространственных данных. Входные данные содержат множество полигональных, точечных или других объектов. Данные разделяются на блоки и распределяются между вычислительными узлами. При таком подходе возникает ситуация, когда один объект содержится в двух и более блоках одновременно, и необходимо определить, к какому блоку его отнести. Для решения этой задачи авторы предлагают два способа. Первый — исключить эти объекты из обработки и потерять малую часть данных. Второй — произвести дополнительные вычисления по обработке конфликтных ситуаций. В этом случае достигается точный результат, а общее время работы увеличивается незначительно. Основное отличие настоящей работы заключается в том, что в силу специфики распределенной системы нет возможности настроить общую распределенную систему хранения и передачи данных. Предлагаемый подход по обработке конфликтных ситуаций в [2] является перспективным, он будет более подробно рассмотрен, улучшен и адаптирован под цели настоящей работы.

Существует широкий выбор программных систем геообработки, имеющихся на рынке. Однако, несмотря на обилие систем, реализующих программную модель MapReduce, остается открытым вопрос применения модели MapReduce для программных систем, ее не поддерживающих, для обработки пространственных данных. Все чаще эти программные системы реализуют в виде Web-сервисов, но не решается вопрос разделения и сборки пространственных данных, управления распределенным вычислением для применения этих систем без программирования. Для решения данной проблемы предлагается метод обработки растровых изображений в рамках модели распределенных вычислений MapReduce, который позволяет использовать инструменты пространственной обработки в распределенной вычислительной среде без их модификации. В ИДСТУ СО РАН был предложен оригинальный способ контроля выполнения сервисов [7] в виде функций на языке JavaScript, которые можно использовать в JavaScript-сценариях наравне со стандартными конструкциями языка. Этот способ [7] позволяет выполнять сервисы в автоматическом режиме, а также поддерживает длительно выполняющиеся сервисы и передачу данных.

ПОСТАНОВКА ЗАДАЧИ

В ИДСТУ СО РАН, в рамках Интеграционной программы ИНЦ СО РАН, ведется разработка и развитие Геопортала, одной из функций которого является предоставление инструментов пространственной обработки данных в виде WPS-сервисов (Web Processing Service). WPS — это стандарт интерфейса Web-сервисов, реализующих пространственную обработку растровых и векторных данных, а также доступ к пакетам геомоделирования, инструментам статистики и обработки через Интернет.

Приведем краткое описание некоторых из них. Сервис обработки данных радарной топографической съемки (Shuttle Radar Topography Mission, SRTM) применяется во многих задачах, в частности,

для вычисления уклона (slope) и экспозиции (aspect). Уклон представляет скорость изменения высоты для каждой ячейки цифровой модели рельефа Digital Elevation Model (DEM). Экспозиция устанавливает направление уклона максимальной скорости изменения значений от конкретной ячейки до соседних. Сервис вычислений вегетационного индекса (Normalized Difference Vegetation Index, NDVI) применяется для определения параметров растительности в данном пикселе снимка.

Рассмотрим особенности обработки пространственных данных указанными инструментами. Некоторым инструментам пространственного анализа для корректной работы достаточно обработать каждый пиксель входного растра независимо от других пикселей. В этом случае достаточно будет разделить входные данные на N равных частей и произвести обработку, затем собрать полученные данные. Другие инструменты работают с каждым пикселем входного растра и его окрестностью, поэтому для корректной обработки входные данные приходится разбивать с некоторым перекрытием и определять, как поступать с повторяющимися результатами на шаге Reduce, т. е. усреднить результат, выбрать экстремальное значение, объединить, вычесть и т. д.

Операции над пространственными данными, которые используются в Map и Reduce, повторяются для различных инструментов геообработки ввиду общности обрабатываемых данных. В настоящей работе предлагается метод, включающий обработчики для операций Map и Reduce и спецификации, на основе которых будет происходить процесс распределения и сбора данных среди вычислительных узлов.

ОБРАБОТКА РАСТРОВЫХ ИЗОБРАЖЕНИЙ В РАМКАХ МОДЕЛИ РАСПРЕДЕЛЕННЫХ ВЫЧИСЛЕНИЙ MAPREDUCE

Метод обработки растровых изображений основывается на реализованном в ИДСТУ СО РАН способе контроля процесса выполнения сервисов [7]. Программно Map и Reduce обработчики представляют собой библиотеки, встраиваемые в модуль выполнения сценариев WPS-сервисов. При выполнении сценария модуль определяет, какому вызову сервиса сопоставляется спецификация. При наличии спецификации происходит анализ вычислительных узлов, поддерживающих выполнение данного сервиса. В зависимости от настроек распределения входных данных, определенных в спецификации, производится разделение входных данных с последующим вызовом копий сервисов на удаленных серверах. Модуль выполнения сценариев последовательно опрашивает выполняемые копии сервисов, и как только последняя копия сервиса завершает свою работу, все результаты работы копий скачиваются модулем, происходит процесс сборки результата в соответствии с правилами, определенными в спецификации. Файлы, получающиеся в результате процесса сборки, передаются дальнейшим участникам сценария. На рисунке дана схема работы метода обработки растровых изображений.

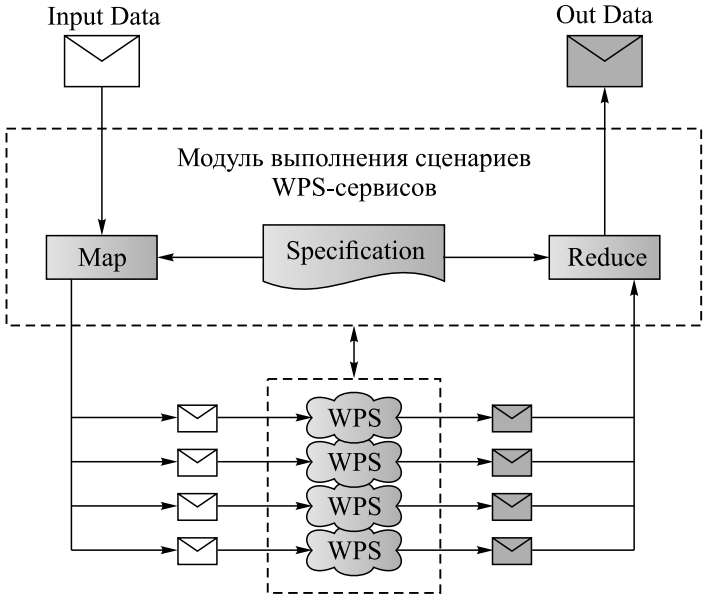

Схема метода обработки растровых изображений в рамках модели MapReduce.

Приведем описание некоторых элементов этого метода. Обработчик операции Map включает реализованные функции чтения спецификаций, на их основе формируются параметры для распределения растровых данных между вычислительными узлами. Для разделения данных формируются параметры запуска утилиты GDAL TRANSLATE, предназначенной для конвертации растров. Обработчик операции Reduce включает реализованные функции чтения спецификаций, реализованные обработчики сбора данных. Обработчики данных выполняют стандартные функции обработки конфликтных ситуаций, возникающих в процессе сбора данных. Конфликтные ситуации возникают, например, при сборе частей мозаики растра в одно целое. К таким ситуациям можно отнести поступление повторяющихся или неоднозначных данных. В этом случае обработчик применяет к ним операцию, указанную в спецификации.

Спецификации написаны в формате Java Script Object Notation (JSON). Настройки спецификаций позволяют указывать минимальные и максимальные размеры ячейки для обработки, предоставляя операции Map самостоятельно определять размер ячеек для оптимальной загрузки вычислительных узлов вызываемыми сервисами (в таком случае обработчику также сообщается число вычислительных узлов). Расчет размера ячейки производится на основе стратегии равномерной загрузки вычислительных узлов, т. е. обработчик подбирает размер ячейки в рамках заданных ограничений таким образом, чтобы занять все доступные узлы. Спецификации для операции Map содержат следующую информацию: ширина и высота ячейки данных, ширина полосы перекрывающихся пикселей для соседних ячеек. Спецификации для операции Reduce содержат название метода, применяемого на шаге сбора полученных результатов для обработки перекрывающихся пикселей.

ЗАКЛЮЧЕНИЕ

В настоящей статье предложен метод автоматизации параллельного выполнения инструментов пространственного анализа растровых изображений в рамках модели распределенных вычислений MapReduce. Отличительной чертой метода является возможность использования инструментов обработки пространственных данных в распределенной вычислительной среде без их модификации. Разработаны и реализованы обработчики для операций Map и Reduce, которые позволяют управлять процессом распределения и сбора обрабатываемых данных независимо от инструментов пространственной обработки. На основе спецификаций можно указать способ распределения и сбора обрабатываемых данных. Предлагаемый метод не требователен к вычислительным узлам.

Работа выполнена в рамках Интеграционной программы ИНЦ СО РАН «Фундаментальные исследования и прорывные технологии как основа опережающего развития Байкальского региона и его межрегиональных связей».

СПИСОК ЛИТЕРАТУРЫ

1. **Dean J., Ghemawat S.** MapReduce: Simplified data processing on large clusters // Sixth Symposium on Operating System Design and Implementation. — San Francisco, USA, 2004. — P. 120–127.
2. **Hadoop** // Apache Hadoop Tutorials [Электронный ресурс]. — http://hadoop.apache.org (дата обращения 18.04.2012).
3. **Spatial** hadoop workshop // Spatial tools [Электронный ресурс]. — http://spatialhadoop.cs.umn.edu/ (дата обращения 18.04.2012).
4. **Abouzeid A., Bajda-Pawlikowski K., Abadi D.** Hadoop DB: An architectural hybrid of MapReduce and DBMS technologies for workloads // Proceed. of the 35th VLDB Conference. — Lyon, France, 2009. — P. 84–90.
5. **Созыкин А. В., Гольдштейн М. Л.** Система обработки изображений с автоматическим распараллеливанием на основе MapReduce // Вестн. Южно-Уральского гос. ун-та. — 2012. — № 27 (286). — С. 109–118.
6. **Aji A., Wang F., Vo H.** Hadoop: GIS: A high performance spatial data warehousing system over MapReduce // The 39th Intern. Conference on Very Large Data Bases. — Trento, Italy, 2013. — Vol. 6, N 11. — P. 1009–1020.
7. **Фёдоров Р. К., Шумилов А. С.** WPS-сервисы пространственного анализа состояния окружающей среды и природных ресурсов // Инфраструктура науч. информ. ресурсов и систем. — 2014. — Т. 2. — С. 66–74.

Поступила в редакцию 21 октября 2016 г.