ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

УДК 004.75
doi: 10.18101/2304-5728-2017-2-12-19

ОБРАБОТКА ВЕКТОРНЫХ ДАННЫХ С ПОМОЩЬЮ СПЕЦИФИКАЦИЙ В СООТВЕТСТВИИ С МОДЕЛЬЮ MAPREDUCE[1]

© Федоров Роман Константинович
кандидат технических наук, ведущий научный сотрудник,
Институт динамики систем и теории управления
им. В. М. Матросова СО РАН
Россия, 664033, Иркутск, ул. Лермонтова, 134
E-mail: fedorov@icc.ru

© Авраменко Юрий Владимирович
программист, Институт динамики систем и теории управления
им. В. М. Матросова СО РАН
Россия, 664033, Иркутск, ул. Лермонтова, 134
E-mail: avramenko@icc.ru

© Шумилов Александр Сергеевич
аспирант, Институт динамики систем и теории управления
им. В. М. Матросова СО РАН
Россия, 664033, Иркутск, ул. Лермонтова, 134
E-mail: shumilov@icc.ru

Одним из подходов к ускорению обработки данных является применение модели распределенных вычислений MapReduce. В данной модели исходные данные распределяются между вычислительными узлами, обрабатываются, а затем собираются в результирующий массив данных. Существующие программные средства, реализующие MapReduce, не учитывают специфику обработки пространственных данных. В целях уменьшения времени выполнения сервисов и равномерного использования доступных вычислительных ресурсов локальной облачной инфраструктуры ИДСТУ СО РАН в рамках модели MapReduce, предложены реализации операций, управляемых спецификациями, для параллельного выполнения сервисов. В данной работе рассматривается применение спецификаций для сбора результатов обработки пространственных данных в векторном формате на примере веб-сервиса идентификации объектов.
Ключевые слова: MapReduce; WPS; SHAPE; spatial data; image processing.

[1] Работа выполнена при финансовой поддержке РФФИ, номер гранта 16-07-00110-мол_а; 16-57-44034_монг_а; 16-07-00411_а.

Введение

В современном мире информационных технологий постоянно увеличиваются объемы создаваемой, хранимой и обрабатываемой информации. Для ее обработки и публикации все чаще используется сервис-ориентированный подход, заключающийся в предоставлении вычислительных алгоритмов и источников данных в виде веб-сервисов, доступных через сеть Интернет. Веб-сервисы могут быть развернуты как в пределах локальной облачной инфраструктуры, так и на удаленных серверах, которые могут находиться в любой точке сети Интернет. Часто сервис может находиться на нескольких вычислительных узлах. Соответственно для ускорения обработки данных некоторыми сервисами можно применить модель MapReduce, где исходные данные распределяются между вычислительными узлами, обрабатываются, а затем собираются в результирующий массив данных. В рамках модели MapReduce необходимо реализовывать операции map и reduce, которые изменяются в зависимости от решаемой задачи и данных. Для пространственных данных можно выделить типовые реализации операций map и reduce [1], которые могут применяться для многих сервисов, работающих с такими данными. В статье приводится реализация операций map и reduce в рамках Геопортал ИДСТУ СО РАН [2] для пространственных данных на примере сервиса идентификации объектов на спутниковых снимках. Геопортал ИДСТУ СО РАН представляет собой веб-приложение, ориентированное на обработку и хранение геоинформационных данных. Для выполнения операций map и reduce над данными реализован набор обработчиков, работа которых определяется с помощью спецификаций, заданных в виде JSON файлов. Спецификации содержат команды для обработчиков операций map и reduce. Любой сервис, зарегистрированный на Геопортале, может иметь определенную для него спецификацию, задающую способ разбивки входных (операция map) и склейки результирующих данных (операция reduce). Ранее спецификации были определены только для растровых данных.

Цель работы

В ИДСТУ СО РАН разработан алгоритм идентификации объектов на спутниковых снимках, реализованный в виде веб-сервиса с интерфейсом стандарта WPS (Web Processing Service) и зарегистрированный на Геопортале. Веб-сервис идентификации принимает на вход растровые файлы спутниковых снимков высокого разрешения и возвращает векторный файл с найденными полигональными объектами, для каждого из которых определено значение функции энергии. Время работы веб-сервиса на изображении размером 1 000x1 000 пикселей может составлять от нескольких до десятков минут. В целях уменьшения времени работы сервиса и равномерного использования доступных вычислительных ресурсов локальной облачной инфраструктуры ИДСТУ СО РАН выполнение разработанного веб-сервиса было распараллелено в рамках

модели MapReduce, то есть входные данные, в частности, спутниковый снимок, делится на некоторое количество частей, которые впоследствии передаются в копии веб-сервиса, запущенные на разных вычислительных узлах. Однако, в силу специфики работы веб-сервиса, а именно векторного формата SHAPE его выходных параметров, возникает задача склейки результатов выполнения копий веб-сервиса в один файл. В данной работе предложен веб-сервис идентификации объектов на растровом изображении и реализован обработчик операции reduce, решающий задачу обработки векторных данных на этапе склейки в рамках программной модели MapReduce.

Веб-сервис идентификации объектов на спутниковых снимках

Веб-сервис производит поиск и идентифицирует объекты на изображении с использованием заданной пользователем совокупности характеристик. Веб-сервис принимает на вход три параметра: входное изображение; запрос пользователя; пороговое значение оценки объектов. На выходе формируется файл в формате SHAPE, в котором содержатся найденные объекты со значением оценки ниже порога. Значение оценки записывается в поле объекта `threshold`. Для формирования запроса (совокупности характеристик объекта) используется логический язык SOQL [3]. Пример описания объекта прямоугольной формы с заданной текстурой и спектральными характеристиками `def(A,B,C,D):-line(A,B), line(B,C), line(C,D), line(D,A), dist(A,B)=30, dist(B,C)=20, dist(C,D)=30, dist(D,A)=20, angle(A,B,C)=90, angle(B,C,D)=90, angle(C,D,A)=90, texture(A,B,C,D,sample).` Соответствующий запрос выглядит следующим образом `?-def(A,B,C,D).`

С целью ускорения работы веб-сервиса по времени предлагается запустить его в рамках модели MapReduce. Модель MapReduce предполагает разбиение и сборку исходных данных. Для обеспечения корректной работы веб-сервиса исходные данные необходимо разделить на пространственные ячейки таким образом, чтобы они частично перекрывали друг друга. Такой подход на этапе объединения результатов от различных узлов распределенной сети приведет к дублированию объектов [4]. Поэтому предлагается на основе разработанных спецификаций [5] реализовать возможность обработки векторных результатов выполнения распараллеленных сервисов.

Распараллеливание веб-сервиса идентификации

При распараллеливании выполнения веб-сервисов в рамках программной модели MapReduce сначала производится операция разделения входных данных. При разделении возможна ситуация, когда искомый объект может находиться на границе частей входного разделяемого файла. Для обработки такого рода граничных случаев в

спецификации, соответствующей операции map, необходимо указать соответствующий параметр перекрытия частей разделяемого растра для обеспечения нахождения всех объектов. Пример спецификации, приведенный ниже, определяет область перекрытия равной 500 метрам:

```
{map: {
    overlap: 500, // Ширина перекрытия
    units: «m» // Единицы измерения
}}
```

После выполнения операции разделения входного параметра происходит запуск распределенных копий веб-сервиса идентификации объектов, каждому из которых передается своя часть разделенного файла. Схема распараллеливания выполнения веб-сервиса идентификации объектов приведена на рис. 1. Модуль Геопортала принимает на вход параметры, передаваемые на вход веб-сервису. Далее параметры, для которых определена спецификация, разделяются в соответствии со спецификацией операции map. Затем происходит запуск копий веб-сервиса идентификации в соответствии со стандартом WPS. По мере завершения работы копий веб-сервисов данные собираются Геопорталом, вызывается соответствующий reduce-обработчик, работающий в соответствии со спецификацией и полученный файл возвращается пользователю.

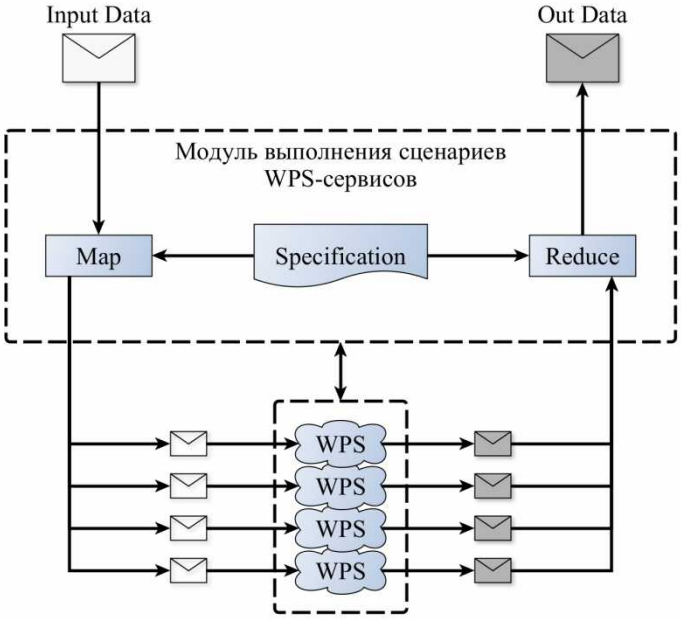


Рис. 1. Схема распараллеливания выполнения веб-сервиса

Распараллеливание выполнения веб-сервиса позволяет уменьшить общее время выполнения веб-сервиса, а также задействовать доступные аппаратные мощности для решения задачи.

Склейка результатов

При склейке результатов выполнения копий веб-сервисов возможно возникновение конфликтов. Для полигональных объектов существует определенный набор топологических правил, например, превышение кластерного доступа, недопустимость перекрытия объектов или пробелов внутри полигонов, необходимость совпадения границ и площадей объектов из разных классов и т. д. Топологические правила используются в картографии для устранения ошибок. Нарушение одного из правил приводит к конфликту. Для рассматриваемой задачи характерно перекрытие объектов.

Склейка результатов выполнения копий распараллеленного веб-сервиса идентификации объектов осуществляется в соответствии со следующей спецификацией:

```
{
  reduce: {
    conflict: { // Секция спецификации обработки конфликтов
      type: “Intersection”, // Тип конфликта
      keepAll: false, // Определяет, разрешать ли конфликты
      resolveFieldName: «threshold»,    // Сравниваемое поле
      filter: “MIN” // Условие
    }
  }
}
```

Псевдокод алгоритма склейки двух соседних файлов данных, содержащих список найденных объектов, представлен ниже. Массивы list1 и list2 содержат элементы рассматриваемых соседних SHAPE файлов. resolveFieldName, filter, keepAll — параметры алгоритма, получаемые из спецификации. mergedList — список элементов из соседних файлов, между которыми конфликты разрешены (в случае необходимости), то есть объекты готовы для записи в результирующий файл. Поле type указывает на топологическое правило, в случае нарушения которого производится исправление (в данном случае это правило недопустимости перекрытия объектов).

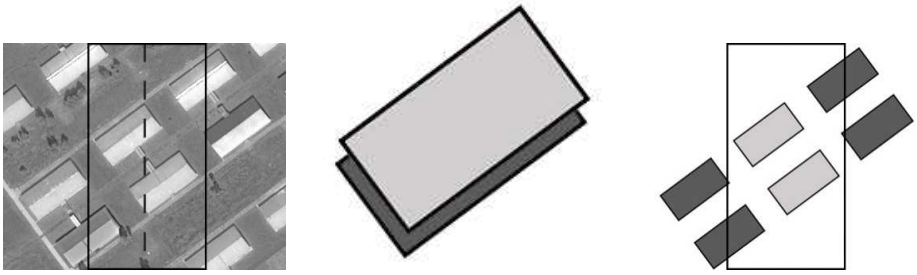
```
ВХОД: list_1, list_2, resolveFieldName, filter, keepAll.
ВЫХОД: mergedList.
Если keepAll=true тогда выполнить
  Объекты из list_1 и list_2 добавить в mergedList
Иначе выполнить
  Если type=Intersection тогда выполнить
    Сформировать списки пересекающихся объектов,
```

```
        соответствующие одному реальному lists
      Цикл для каждого списка из lists выполнить
        Взять объект, у которого значение поля resolveFieldName
        удовлетворяет условию в поле filter
        и добавить его в mergedList
        Конец цикла
      Добавить оставшиеся не пересекающиеся объекты в mergeList
    Конец условия
  Конец условия
Вернуть список объектов mergedList
```

Веб-сервис может быть распараллелен на произвольное количество вычислительных узлов, таким образом, число результирующих векторных файлов может также варьироваться. Так как алгоритм склейки работает только с двумя соседними файлами, процесс получения общего результата выполнения сервиса, содержащего в себе все полученные результирующие файлы, разбивается на несколько этапов. На каждом из этапов выбираются пары соседних файлов, которые затем склеиваются. Один и тот же файл может быть только в одной паре. На следующий этап переходят склеенные файлы и файлы, не получившие пару на текущем этапе. На последнем этапе происходит склейка двух последних файлов, и получившийся файл возвращается пользователю как результат выполнения веб-сервиса.

Апробация

Апробация работы алгоритма склейки SHAPE файлов показана на рисунке 2. В данном примере необходимо было найти объекты прямоугольной формы техногенного происхождения. Исходное изображение рисунок 2.а. было разделено на два, частично перекрывающих друг друга, относительно фактической границы раздела — пунктир, имеющих общую область — рамка. После выполнения копий веб-сервиса происходит этап склейки, в процессе которого исключаются пересекающиеся объекты, как показано на рисунке 2.б. После формирования результирующего файла в него добавляются найденные объекты, находящиеся в общей области, прошедшие проверку и оставшаяся часть объектов, что изображено на рисунке 2.в.



а б в

Рис. 2. Апробация работы алгоритма объединения двух SHAPE файлов в один

В целях определения эффективности метода рассматриваемый сервис идентификации был распараллелен и его копии были запущены на 12 вычислительных узлах, схожих по программным и аппаратным характеристикам. Время выполнения сервиса, с учетом времени разбиения, передачи и склейки данных, а также времени непосредственного выполнения копий сервиса, равно 110 секундам (взято среднее значение среди 20 запусков). Сервис идентификации объектов, запущенный без распараллеливания на одном вычислительной узле, выполняется 1 240 секунд (взято среднее значение среди 20 запусков). Распараллеливание работы веб-сервиса позволило ускорить его выполнение в 11,2 раза.

Таким образом, при распараллеливании выполнения веб-сервиса (при условии, что время выполнения сервиса линейно зависит от объема входных данных, а также при условии идентичности аппаратных и программных характеристик вычислительных узлов как при распараллеливании, так и без) в соответствии с программной моделью MapReduce ускорение его выполнения в среднем равно числу узлов, на которых происходит распараллеливание.

Заключение

В результате проделанной работы разработаны спецификации, определяющие параметры обработки векторных данных при процессе их склейки в рамках модели MapReduce. Реализованы соответствующие обработчики, осуществляющие склейку данных на основании спецификаций. Реализованный способ распараллеливания работы веб-сервисов был апробирован и показал свою работоспособность и эффективность. Применение управляемых спецификациями обработчиков позволяет упростить выполнение сервисов в рамках модели MapReduce.

Литература

1. Авраменко Ю. В., Шумилов А. С. Спецификации методов разбиения и сборки растровых изображений в рамках программной модели MapReduce // География и природные ресурсы. 2016. № 6. С. 156–159.

2. Компоненты среды wps-сервисов обработки геоданных / Р. К. Федоров, И. В. Бычков, А. С. Шумилов, Г. М. Ружников // Вестник Новосибирского Государственного Университета. Серия: информационные технологии. 2014. Т. 12, № 3. С.16–24.

3. Интерпретатор языка SOQL для обработки растровых изображений / И. В. Бычков, Р. К. Федоров, Ю. В. Авраменко, Г. М. Ружников // Новосибирск: Вычислительные технологии. 2016. Т. 21, № 1. С. 49–59.

4. Hadoop: GIS: A High Performance Spatial Data Warehousing System over MapReduce / Aji A. [et al] // The 39th International Conference on Very Large Data Bases. 2013. Vol. 6, №. 11. P. 1009–1020.

5. Авраменко Ю. В., Шумилов А. С. Метод обработки растровых изображений в рамках модели mapreduce // Информационные и математические технологии в науке и управлении. 2016. № 4–2. С. 110–115.

VECTOR DATA PROCESSING BY MEANS OF SPECIFICATIONS IN ACCORDENCE WITH THE MAPREDUCE PROGRAMMING MODEL

Roman K. Fedorov
Cand. Sci. (Engineering), Leading Researcher,
Matrosov Institute of System Dynamics and Control Theory, SB RAS,
134 Lermontova St., Irkutsk 664033, Russia
E-mail: fedorov@icc.ru

Yuriy V. Avramenko
Programmer
Matrosov Institute of System Dynamics and Control Theory, SB RAS,
134 Lermontova St., Irkutsk 664033, Russia
E-mail: avramenko@icc.ru

Aleksandr S. Shumilov
Research Assistant,
Matrosov Institute of System Dynamics and Control Theory, SB RAS,
134 Lermontova St., Irkutsk 664033, Russia
E-mail: shumilov@icc.ru

One of the approaches to accelerating data processing is the application of the distributed computing model MapReduce. In this model, input data are distributed among the computational nodes, processed, and then assembled into the resulting body of data. Existing software that implements MapReduce does not take into account the specificity of spatial data processing. In order to minimize the service execution time and to provide uniform use of available computing resources of the local cloud infrastructure of SB RAS Institute of System Dynamics and Control Theory within the framework of the MapReduce model, we have proposed the implementations of operations controlled by specifications for parallel execution of services. The article considers application of specifications for processing the results in a vector format on the example of a web service for object identification.
Keywords: MapReduce; WPS; SHAPE; spatial data; image processing.