УДК 004.822

Обработка и импорт больших XML–данных партнерских сайтов

Малышев Владислав Андреевич
Приамурский государственный университет имени Шолом-Алейхема
Студент

Голубь Илья Сергеевич
Приамурский государственный университет имени Шолом-Алейхема
Студент

Глаголев Владимир Александрович
Приамурский государственный университет имени Шолом-Алейхема
к.г.н., доцент кафедры информационных систем, математики и правовой информатики

Аннотация
В данной статье рассмотрено загрузка больших XML файлов и добавление их в базу данных на вебсервере компонентами языка PHP (SimpleXML, XMLReader).
Ключевые слова: XML, SimpleXML, XMLReader, Mysql, PHP.

Processing and import of large XML data partner sites

Malyshev Vladislav Andreevich
Sholom-Aleichem Priamursky State University
Student

Golub Iliya Sergeevich
Sholom-Aleichem Priamursky State University
Student

Glagolev Vladimir Alexandrovich
Sholom-Aleichem Priamursky State University
candidate of geographical Sciences, associate professor of the Department of information systems, mathematics and law informatics

Abstract
This article describes the download of the XML files and add them to the database on the webserver using two tools PHP: SimpleXML and XMLReader.
Keywords: XML, SimpleXML, XMLReader, Mysql, PHP.

В настоящее время появляется всё больше интернет-организаций, которые хотят расширить предлагаемый ассортимент товаров и услуг.

Однако перед ними появляется проблема обработки и выдачи многоструктурной информации большого объема, а также ее дальнейшего импорта

Цель исследования является провести сравнение способов обработки и импорта больших XML-данных на примере партнерских сведений издательских организаций.

Исследованиями в данной теме занимались многие авторы, такие как Д.С.Яковлев рассмотрел вопрос использования XML - документов в качестве баз данных для веб-приложений [1]. А.А. Горбунов, М.Б. Хорошко в своей статье рассмотрели расширение возможностей языка программирования PHP с использованием набора инструментов SimpleXML [2]. А.А. Горбунов рассмотрел подсистему разработанной Amazon для взаимодействия со своей торговой площадкой [3]. F.Wang, J.Li, H.Homayounfar в своём исследовании разработали свой DOM parse [4].

SimpleXML – это набор инструментов который служит для преобразования XML в объектах баз данных [6, 8, 9]. Этот набор работает на версии PHP 5 и выше.

Расширение XMLReader - синтаксический анализатор XML. Класс-читатель выступает в качестве курсора, следует по потоку документа и останавливается на каждом узле на этом пути [7]. Это расширение было включено по умолчанию с версии PHP 5.1.2.

Перед нами поставлена задача: загрузить в существующую базу данных, взятую с издательского сайта в партнёрских целях XML – документ объемом 307 мегабайт [5], в котором насчитывалось 174 тысячи записей.

Пример одной записи из этого файла:

<offer id="476215" type="book" available="true"><url>https://www.xxxx.ru/ivan-mirolubov/vosem-let-na-sahaline/</url><price>0</price><currencyId>RUR</currencyId><categoryId>40400</categoryId><picture>https://www.xxxx.ru/static/bookimages/01/09/70/01097095.bin.dir/01097095.cover.jpg</picture><author>И.П. Миролюбов</author><name>Восемь лет на Сахалине</name><publisher>Библиотечный фонд</publisher><series></series><year>1901</year><ISBN></ISBN><description></description><downloadable>true</downloadable><age>0</age><param name="Форматы"></param><genres_list>5214</genres_list></offer>.

Структура XML-файла была определена бесплатно-распространяемой по лицензии GNU программой Notepad++.

Рис1.Структура XML документа

В первом случае был создан код на основе набора инструментов SimpleXML.

```
<?php
set_time_limit(72000); $start = microtime(true);
$mysqli = new mysqli("localhost", "root","", "xxxx" );
$sxml = simplexml_load_file($_FILES['xml']['tmp_name']);
echo
'id;type;available;url;price;currencyId;categoryId;picture;author;name;publisher;series;year;
ISBN;description;downloadable;age;litres_isbm;genres_list<br>';
foreach($sxml->shop as $shop) {
   foreach($sxml->shop->offers as $offers) {
    foreach($sxml->shop->offers->offer as $offer) {
      $id=$offer->attributes()['id'];      $type=$offer->attributes()['type'];      $available=$offer-
>attributes()['available'];      $url=$offer->url;      $price=$offer->price;      $currencyId=$offer-
>currencyId;      $categoryId=$offer->categoryId;      $picture=$offer->picture;
$author=str_replace("'","''",$offer->author);      $name=str_replace("'","''",$offer->name);
$publisher=str_replace("'","''",$offer->publisher);;      $series=str_replace("'","''",$offer-
>series);      $year=$offer->year;      $ISBN=$offer->ISBN;
$description=str_replace("'","''",$offer->description);      $downloadable=$offer-
>downloadable;      $age=$offer->age;      $litres_isbn=$offer->litres_isbn;
$genres_list=$offer->genres_list;
      $sql = "INSERT INTO `offers` (`id`,`type`,`available`,`url`,`price`,
        `currencyId`,`categoryId`,`picture`,`author`,`name`,`publisher`,`series`,
        `year`,`ISBN`,`description`,`downloadable`,`age`,`litres_isbn`,`genres_list`)   VALUES
('$id','$type','$available','$url','$price','$currencyId','$categoryId','$picture','$author','$name','$
publisher','$series','$year','$ISBN',
         '$description','$downloadable','$age','$litres_isbn','$genres_list')";
      $result = $mysqli->query($sql);
   }
}
}
echo  'Время выполнения скрипта: '.(microtime(true) - $start).' сек.';
?>
```

Во втором случае был создан код на основе расширения XMLReader.

```
<?php
$start = microtime(true);
$mysqli = new mysqli("localhost", "root","", "litres" );
$reader = new XMLReader();
$reader->open($_FILES['xml']['tmp_name']);
while ($reader->read()) {
  switch ($reader->nodeType) {
  case (XMLREADER::ELEMENT):
   if ($reader->name == "offer")
     {
      $id = $reader->getAttribute("id");     $type = $reader->getAttribute("type");     $price =
$reader->getAttribute("available");
}
   if ($reader->name == "url")
```

```
    {      $reader->read();     $url = $reader->value;  }
  if ($reader->name == "price")
    {      $reader->read();      $price = $reader->value;    }
   if ($reader->name == "currencyId")
       {         $reader->read();         $currencyId = $reader->value;       }
      if ($reader->name == "categoryId")
       {         $reader->read();         $categoryId = $reader->value;       }
      if ($reader->name == "picture")
       {          $reader->read();          $picture = $reader->value;         }
      if ($reader->name == "author")
       {          $reader->read();
        $author = $reader->value;          }
       if ($reader->name == "name")
        {       $reader->read();            $name = $reader->value;          }
       if ($reader->name == "publisher")
        {            $reader->read();            $publisher = $reader->value;           }
       if ($reader->name == "series")
        {             $reader->read();             $series = $reader->value;           }
        if ($reader->name == "year")
         {             $reader->read();             $year = $reader->value;            }
        if ($reader->name == "isbn")
         {              $reader->read();              $isbn = $reader->value;             }
        if ($reader->name == "description")
         {             $reader->read();             $description = $reader->value;            }
        if ($reader->name == "downloadable")
         {              $reader->read();              $downloadable = $reader->value;             }
         if ($reader->name == "age")
          {               $reader->read();               $age = $reader->value;              }
         if ($reader->name == "litres_isbn")
          {               $reader->read();               $litres_isbn = $reader->value;              }
         if ($reader->name == "genres_list")
          {        $reader->read();
           $genres_list = $reader->value;
           $sql = "INSERT INTO `offers` (`id`,`type`,`available`,`url`,`price`,
            `currencyId`,`categoryId`,`picture`,`author`,`name`,`publisher`,`series`,
            `year`,`ISBN`,`description`,`downloadable`,`age`,`litres_isbn`,`genres_list`)
             VALUES ('$id','$type','$available','$url','$price','$currencyId','$categoryId',
              '$picture','$author','$name','$publisher','$series','$year','$ISBN',
              '$description','$downloadable','$age','$litres_isbn','$genres_list')";
           $result = $mysqli->query($sql);
           break;
          }  } }
echo  'Время выполнения скрипта: '.(microtime(true) - $start).' сек.';
?>
```

Для тестирования этих скриптов был использован компьютер с основными характеристиками: Процессор: AMD fx – 6300; оперативная

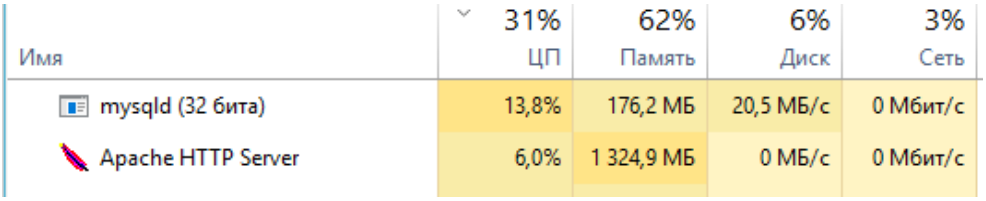
память: 16 гигабайт; память: SSD 120 гигабайт; apache v2.4 64x; PHP v7.0 64x; MySQL v5.6

В результате использования кода с SimpleXML можно видеть сколько было всего затрачено времени на выполнение скрипта.

Время выполнения скрипта: 52.613950967789 сек.

Рис2. Результат выполнения скрипта

Так же можно увидеть сколько ресурсов компьютера было затрачено на выполнение скрипта.



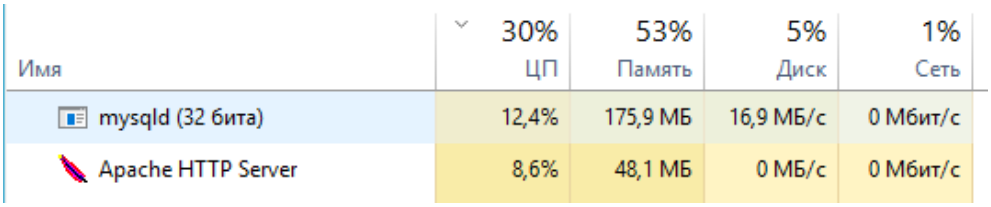
Имя	31% ЦП	62% Память	6% Диск	3% Сеть
mysqld (32 бита)	13,8%	176,2 МБ	20,5 МБ/с	0 Мбит/с
Apache HTTP Server	6,0%	1 324,9 МБ	0 МБ/с	0 Мбит/с

Рис3. Диспечер задач

В результате использования кода с XMLReader был получен следующий результат.

Время выполнения скрипта: 56.806421995163 сек.

Рис4. Результат выполнения скрипта



Имя	30% ЦП	53% Память	5% Диск	1% Сеть
mysqld (32 бита)	12,4%	175,9 МБ	16,9 МБ/с	0 Мбит/с
Apache HTTP Server	8,6%	48,1 МБ	0 МБ/с	0 Мбит/с

Рис5. Диспечер задач

Отсюда можно сделать выводы: SimpleXML очень ресурсозатратный и поэтому он лучше всего подходит для обработки небольших XML - документов, так как предостовляемая мощность сервера может быть крайне мала. XMLReader отличное решения для обработки больших XML – документов, так как он практически не нагружает систему.

Библиографический список

1. Яковлев Д.С. Использование xml-документов в качестве баз данных для вэб-приложений // Вестник магистратуры. 2015. № 4-1 (43). С. 8-10.
2. Горбунов А.А., Хорошко М.Б. Использование расширения simplexml на основе языка программирования php // В сборнике: Информационные и

измерительные системы и технологии Сборник научных статей по материалам Международной научно-технической конференции. 2016. С. 34-36.
3. Горбунов А.А. Подсистема информационного взаимодействия с Amazon // В сборнике: Информационные и измерительные системы и технологии сборник научных статей по материалам еженедельного научно-технического семинара. 2016. С. 63-66.
4. Wang F., Li J., Homayounfar H. A space efficient XML DOM parser // Data & Knowledge Engineering. 2007. Т.60. №1. С. 185-207
5. База книг ЛитРес // URL: http://www.litres.ru/static/ds/partners.yml.gz (дата обращения: 13.01.2018).
6. SimpleXML // URL: http://php.net/manual/ru/book.simplexml.php (дата обращения: 13.01.2018).
7. XMLReader // URL: http://php.net/manual/ru/book.xmlreader.php (дата обращения: 13.01.2018).
8. Глаголев В.А Разработка банка данных метеорологических параметров для анализа пожарной опасности территории // Региональные проблемы. 2007. № 8. С. 152-155.
9. Глаголев В.А. Создание баз данных для оценки и прогноза пожарной опасности растительности по природно-антропогенным условиям // Региональные проблемы. 2014. Т. 17. № 2. С. 78-83.