УДК 004.4:528.9

Разработка технологии организации каталогов спутниковых данных*

А.А. Донцов, Н.В. Волков, А.А. Лагутин

Алтайский государственный университет (Барнаул, Россия)

Technology Development for Satellite Data Warehouse Design

A.A. Dontsov, N.V. Volkov, A.A. Lagutin

Altai State University (Barnaul, Russia)

Предложена обобщенная модель построения хранилища данных большого объема, основанная на использовании NoSQL-технологии. Данная технология не предъявляет строгих требований к организации структуры хранилища данных, обладает высокой скоростью обработки запросов, возможностью проводить масштабирование в широких пределах, а также предоставляет допустимость репликации баз на нескольких серверах с поддержкой автоматической синхронизации данных.

Организация каталогов данных с использованием технологии NoSQL позволяет проектировать базы данных, отличные от традиционных реляционных систем управления базами данных (СУБД) с доступом к данным средствами языка запросов SQL. Для реализации системы управления каталогами спутниковых данных используется документо-ориентированная СУБД MongoDB.

Информационной основой разрабатываемой системы являются данные прибора спутникового базирования MODIS, установленного на платформах Terra и Aqua. Для извлечения информации из файлов спутниковых данных используется утилита `gdalinfo` библиотеки GDAL.

Тестовый анализ, проведенный при построении каталога данных объемом более 20 гигабайт на сервере с двухъядерным процессором с тактовой частотой 2,1 ГГц и объемом оперативной памяти 2 Гб, позволил получить оценку времени — около ~ 5 минут, требуемого для выполнения всей последовательности операций.

Ключевые слова: геопортал, хранилища и каталоги данных, системы управления базами данных, приборы спутникового базирования.

DOI 10.14258/izvasu(2014)1.2-29

A generalized model for a big data warehouse based on the NoSQL technology is proposed. This technology does not impose strict requirements to a database structure. It is scalable in large extend with a high-speed processing of queries and also provides the ability of database multiple servers replication with automatic data synchronization support.

Data warehousing with NoSQL technology allows the design of databases to be different from the traditional relational database management systems (DBMS) based on SQL query tools for data access. The MongoDB database system is used for satellite metadata management. The source of data for the developed system is satellite-based spectroradiometer MODIS data (Terra and Aqua platforms). Gdalinfo utility from GDAL library is used for satellite data files processing.

A test analysis is performed for the warehouse of more than 20 Gigabytes. The dual-core processor server with a CPU clock speed of 2.1 GHz and RAM 2 GB is utilized. The time for completing the entire sequence of operations is estimated in about 5 minutes.

Key words: document preparation system, scientific paper, manual.

*Работа выполнена при поддержке Минобрнауки РФ (государственное задание на проведение фундаментальных и прикладных научных исследований, выполняемых в АлтГУ).

Введение. Повсеместное активное использование данных дистанционного зондирования Земли приводит к тому, что к комплексам приема, хранения и обработки спутниковых данных предъявляются более жесткие требования по оперативности и технологичности их работы. Для внедрения таких систем в алгоритмы решения прикладных задач к ним предъявляется дополнительное требование — системы должны быть максимально автоматизированы.

Одним из ключевых элементов разрабатываемых комплексов, независимо от способа реализации, является автоматизированная система хранения спутниковых данных, предназначенная для решения последовательности задач, связанных с архивацией данных и предоставлением к ним доступа. Среди таких задач выделяются: аннотация данных, входной контроль целостности данных, каталогизация и обеспечение оперативного доступа к данным.

Целью работы является представление результатов разработки технологии организации каталогов спутниковых данных для решения задач оперативного мониторинга параметров атмосферы и подстилающей поверхности. На этапе проектирования комплекса основное внимание уделяется гибкости и масштабируемости предлагаемых программных решений.

1. Технологическая основа. В разрабатываемой системе для организации каталогов данных используется технология баз данных NoSQL (англ. *not only SQL*, *не только SQL*) [1]. Термин NoSQL обозначает ряд подходов, направленных на проектирование баз данных, имеющих существенные отличия от моделей, использующихся в традиционных реляционных системах управления базами данных (СУБД) с доступом к данным средствами языка запросов SQL.

Основными отличительными особенностями NoSQL-решений являются:

- возможность совместного использования хранилищ разных типов и архитектуры;
- возможность разработки базы данных без проектирования предварительной схемы и, как следствие, сокращение времени разработки;
- полноценная поддержка многопроцессорных технологий;
- линейная масштабируемость (увеличение вычислительных мощностей линейным образом влияет на производительность);
- использование «не только SQL» существенно расширяет возможности для хранения и обработки данных.

Главным преимуществом NoSQL технологии над реляционными СУБД является горизонтальная масштабируемость [1]. В данном контексте горизонтальное масштабирование обозначает возможность произвольного изменения таблиц без перестройки структуры всей базы. Горизонтальное масштабирование существующих традиционных СУБД обычно является трудоемкой, дорогостоящей и эффективной только до определенного уровня задачей. В то же время многие NoSQL-решения проектируются именно исходя из необходимости оперативной горизонтальной масштабируемости.

В NoSQL базах, в отличие от реляционных, структура данных строго не регламентирована. В отдельной строке или документе можно добавить произвольное поле без предварительного изменения структуры всей таблицы. Таким образом, если возникает необходимость изменить поля записей в строке, которые описывают какие-либо характеристики файлов спутниковых данных, то достаточно отразить изменение в программном коде. Это позволяет записывать информацию о файлах, полученных с различных спутниковых сенсоров, имеющих разные поля описания, в одну таблицу каталога данных.

Для реализации комплекса управления каталогами спутниковых данных была выбрана документо-ориентированная СУБД MongoDB [2]. Основные особенности MongoDB заключаются в следующем:

- документо-ориентированная система хранения (JSON/BSON-подобная схема обмена данными, основанная на JavaScript и обычно используемая именно с этим языком [3, 4]);
- гибкий синтаксис языка для формирования запросов;
- поддержка динамических запросов;
- поддержка индексов;
- профилирование запросов;
- журналирование операций, модифицирующих данные в базе;
- поддержка отказоустойчивости и масштабируемости: асинхронная репликация, набор реплик и распределения базы данных на узлы;
- может работать в соответствии с парадигмой модели распределенных вычислений MapReduce, представленной компанией Google [5]. Данная модель используется для параллельных вычислений над очень большими (до нескольких петабайт) наборами данных в компьютерных кластерах;
- полнотекстовый поиск, в том числе на русском языке, с поддержкой морфологии поискового запроса.

2. Принцип работы. Информационной основой разрабатываемой системы являются данные, полученные прибором MODIS (MODerate-

resolution Imaging Spectroradiometer — сканирующий спектрорадиометр среднего разрешения), установленным на борту спутников Terra и Aqua [6–8]. Работы по приему и обработке этих данных проводятся в отделе космического мониторинга АлтГУ уже более десяти лет. За это время накоплен обширный архив файлов формата HDF (Hierarchical Data Format) [9, 10]. Этот формат разработан для хранения больших объемов цифровой информации и является основным форматом для хранения данных, полученных со спутниковых приборов.

В процессе инициализации каталога данных используется утилита `gdalinfo` библиотеки GDAL [11], которая позволяет получать подробную информацию о файлах географических данных. На основе этой утилиты с использованием средств языка программирования Python был разработан специальный модуль. Этот модуль на вход получает информацию о файловой системе хранилища спутниковых данных, в частности, — таблицу путей к сохраненным файлам. Затем рекурсивно заносит полученную информацию в каталог базы данных. На этапе работы с СУБД MongoDB используется библиотека PyMongo [12] языка Python. Для программного доступа к функциям библиотеки GDAL используются инструменты GDAL Python API [11].

Для записи данных в базу используется функция `insert` библиотеки PyMongo [12]. Эта функция в качестве входных данных получает Python-словарь, содержащий атрибуты HDF-файлов, полученные с помощью утилиты `gdalinfo`.

По окончании этапа инициализации коллекция полей, описывающих информацию о HDF-файлах, имеет структуру, подобную приведенной ниже:

```
{"_id":id записи каталога БД,
 "File":путь к файлу,
 "GRINGPOINTLATITUDE":широта,
 "GRINGPOINTLONGITUDE":долгота,
 "RANGEBEGINNINGDATE":"2012-12-05" }
```

В примере представлена лишь часть полей описания файлов в каталоге. Таким образом, для хранения данных используется JSON/BSON-подобная схема данных по принципу «ключ — значение» [3, 4]. У каждой коллекции полей описаний есть свой индивидуальный идентификатор — `id`, который можно устанавливать программно. По умолчанию СУБД MongoDB генерирует `id` автоматически.

Для выборки из каталога по какому либо критерию используется функция `find` библиотеки PyMongo [12]. Ниже представлен пример выборки из документа с именем `products` всех файлов за 12 декабря 2013 г.

```
db.products.find({
"RANGEBEGINNINGDATE":"2013-12-01"
})
```

Вызов функции `find` без указания параметров позволяет получить весь список коллекций полей. Также существует возможность выборки, используя поиск по строковым полям. Ниже представлен пример такого запроса:

```
db.product.find({
"RANGEBEGINNINGDATE": {$regex: /2013/}})
```

В данном случае поиск осуществляется по части строки, содержащей последовательность символов `2013`. Такой запрос может быть полезен в случае, когда не заданы строгие критерии поиска.

Помимо функций, приведенных выше, библиотека PyMongo [12] позволяет обновлять информацию в каталоге, удалять записи, искать дубликаты описаний файлов. Например, если в хранилище есть два файла с разными именами, но содержащие одну и ту же информацию, то при условии совпадения всех полей описания можно удалить дублирующий файл и описание в каталоге.

3. Тестирование системы. Разрабатываемая система каталогов может применяться как в виде самостоятельной программы каталогизации файлов спутниковых данных, так и в качестве модуля любой автоматизированной системы хранения и визуализации спутниковых данных. На начальном этапе проверка работы системы осуществлялась в виде консольного приложения. После отладки система была интегрирована в качестве модуля геопортала, разрабатываемого в отделе космического мониторинга АлтГУ. Предварительные результаты разработки геопортальной системы дистанционного зондирования Земли обсуждались в наших работах [13, 14].

Результаты тестирования системы, построенной с использованием обсуждаемых в данной работе технологических решений, указывают на перспективность предлагаемого подхода. Качественный тестовый анализ, проведенный при построении каталога HDF-файлов объемом более 20 гигабайт на сервере с двухъядерным процессором с тактовой частотой 2,1 ГГц и объемом оперативной памяти 2 Гб, позволил получить оценку времени — около ~ 5 минут, требуемого для выполнения всей последовательности операций. Следует отметить, что данная процедура выполняется однократно. В процессе работы системы, например в качестве модуля геопортала, после поступления новых данных в хранилище атрибуты новых файлов добавляются автоматически без повторной индексации всего хранилища. Передача параметров для выборки файлов в геопортальной системе происходит при формировании задачи по обработке данных после заполнения и отправки пользователем формы запроса данных [13, 14].

Результаты и выводы. В работе представлена технология организации каталогов спутниковых данных. Далее приведены результаты и программные решения, полученные в процессе разработки.

1. Используя утилиту `gdalinfo` библиотеки GDAL [11], разработан вспомогательный модуль, позволяющий извлекать информацию из файлов спутниковых данных, сохраненных в сетевых хранилищах.
2. Разработан основной модуль системы, который сохраняет полученную информацию в таблицы СУБД MongoDB [2] и позволяет проводить все операции с таблицами. Базы данных MongoDB являются центральным ядром системы. Основной особенностью проектируемых баз является использование парадигмы NoSQL, позволяющей оперативно проводить масштабирование баз данных в широких пределах без привлечения существенных вычислительных ресурсов.
3. Проведена интеграция разработанных модулей в геопортальную систему дистанционного зондирования Земли, разрабатываемую в отделе космического мониторинга АлтГУ [13, 14]. Посредством геопортальной системы спутниковые данные могут быть запрошены из хранилища, обработаны и представлены конечному пользователю.
4. Наконец, отметим, что разработанный модуль управления каталогами данных может быть использован в качестве самостоятельного программного обеспечения при решении любых задач, связанных с каталогизацией и обработкой больших массивов данных.

Библиографический список

1. Фаулер М., Садаладж П.Дж. NoSQL: новая методология разработки нереляционных баз данных = NoSQL Distilled. — М., 2013.

2. Official site MongoDB project [Electronic resource]. — URL: https://www.mongodb.org/

3. Введение в JSON [Electronic resource]. — URL: http://json.org/

4. Спецификация BSON [Electronic resource]. — URL: http://bsonspec.org/

5. Dean J., Ghemawat S. (Google, Inc.) MapReduce: Simplified Data Processing on Large Clusters // OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA. — 2004. [Electronic resource]. — URL: http://research.google.com/archive/mapreduce.html

6. Salomonson V.V., Barnes W.L., Maymon P.W. et al. MODIS: Advanced Facility Instrument for Studies of the Earth as a System // IEEE Trans. Geosci. Remote Sens. — 1989. — V. 27, № 2.

7. Barnes W.L., Xiong X.L., Salomonson V.V. Terra MODIS and Aqua MODIS // Adv. Space Res. — 2003. — № 32.

8. Лагутин А.А., Никулин Ю.А., Жуков А.П. и др. Математические технологии оперативного регионального спутникового мониторинга характеристик атмосферы и подстилающей поверхности. Ч. 1. MODIS // Выч. технол. — 2007. — Т. 12, № 2.

9. HDF — Hierarchical Data Format [Electronic resource]. — URL: http://www.hdfgroup.org/

10. Yang W. A review of remote sensing data formats for earth system observations // In book J.J. Qu, W. Gao, M. Kafatos et al. Earth Science Satellite Remote Sensing. Vol. 2: Data, Computational Processing, and Tools, Tsinghua University Press, Beijing and Springer-Verlag. — 2006. — 335 P.

11. GDAL — Geospatial Data Abstraction Library [Electronic resource]. — URL: http://www.gdal.org/

12. PyMongo 2.7rc0 Documentation [Electronic resource]. — URL: http://api.mongodb.org/python/2.7rc0

13. Волков Н.В., Донцов А.А., Лагутин А.А. Разработка геопортальной системы для решения задач регионального космического мониторинга // Известия Алт. гос. ун-та — 2013. — № 1/2(77). DOI:10.14258/izvasu(2013)1.2-30.

14. Донцов А.А., Волков Н.В. Геопортальная система регионального космического мониторинга // Дистанционное зондирование Земли из космоса: алгоритмы, технологии, данные: Матер. молодежной школы-семинара. — Барнаул, 2013.