Управление данными экспериментов с использованием современных web-технологий♣

А.П. Григорюк, Л.П. Брагинская

Институт вычислительной математики и математической геофизики СО РАН
and@opg.sscc.ru, ludmila@opg.sscc.ru

Аннотация

В докладе рассматриваются концептуальные основы, архитектура и программное обеспечение web-ориентированной информационно-вычислительной системы для управления структурированными и неструктурированными данными натурных и вычислительных экспериментов. Система обеспечивает пользователей поисковым, вычислительно-аналитическим и ГИС сервисами для эффективной работы с данными.

В качестве примера практической реализации предложенной архитектуры представлена разработанная авторами доклада информационно-вычислительная система «Вибросейсмическое просвечивание Земли» для управления данными, полученными в ходе экспериментов по активному вибросейсмическому мониторингу.

1. Введение

Согласно определению международной организации DAMA (The Data Management Association International) управление данными – набор процессов, обеспечивающих накопление, организацию, запоминание, обновление, хранение, обработку данных и поиск информации. В настоящее время роль информации и ее обработки в научных исследованиях становится доминирующей. Это связано с тем, что современные высокоточные научные инструменты, применяемые при проведении экспериментов, порождают большие объемы данных. Также, приходится работать с большими объемами синтетических данных, полученными при численном моделировании. Считается, что каждый год объем этих данных почти удваивается, достигая во многих научных областях терабайтных размеров на исследовательскую группу, работающую над конкретным проектом [5]. С другой стороны, для обнаружения «тонких» эффектов во вновь получаемых высокоточных данных требуются все более ресурсоемкие алгоритмы анализа. Следует также отметить коллективность и узкую специализацию современных научных исследований при широкой географии участников конкретного проекта.

В этих условиях представляется целесообразным централизованное управление научными данными с переносом собственно данных и программ для их анализа и визуализации на сервер и организацией интерактивного доступа пользователей через Интернет. Далее рассматриваются основные вопросы, связанные с разработкой подобных систем [1].

2. Модель данных

Организация структурированных табличных (реляционных) данных или метаданных обычно затруднений не вызывает. Для этого идеально подходят реляционная модель и язык SQL, лежащие в основе современных СУБД. Сложнее обстоит дело с данными, получаемыми от сенсоров и численных моделей в процессе эксперимента, представляющими собой n-мерные, в общем случае, числовые массивы, которые не могут быть структурированы и поэтому не поддерживаются реляционными СУБД непосредственно.

В настоящее время для работы одновременно с реляционными и нереляционными данными, в основном используют одну из двух архитектур:

– как реляционные, так и нереляционные данные находятся в базе данных;
– реляционные данные находятся в базе данных, а нереляционные данные – в файловых системах или на файловых серверах.

Каждый из этих двух подходов имеет свои преимущества и недостатки. В первом случае одна база данных становится удобным централизованным хранилищем для обоих типов данных. Однако нереляционные данные хранятся в формате больших двоичных объектов (BLOB), скорость доступа к этим объектам существенно уступает скорости доступа к файлам. Во втором случае обеспечивается высокая скорость доступа, но усложняется разработка приложений и управление ими, так как приложения должны поддерживать согласованность между записями в базе данных и файлами, связанными с этими записями. Данную проблему можно частично или полностью решить за счет модели данных, обеспечивающей эффективную индексацию файловой системы из базы данных.

Интернет и современное общество: сборник научных статей XVI Всероссийской объединенной конференции IMS-2013, Санкт-Петербург, 9 - 11 октября 2013 г.

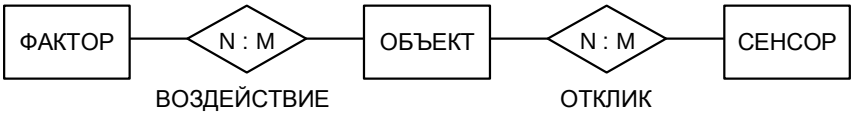

Рис. 1. Концептуальная модель эксперимента

При построении концептуальной модели предметной области мы исходили из того, что экспериментально изучаемому объекту может быть приписан определенный набор параметров, соответствующий представлениям исследователей о состоянии и поведении этого объекта. Параметры это то, что можно измерять, наблюдать и изменять в процессе исследований. В процессе экспериментов на изучаемый объект воздействуют некоторые факторы с контролируемыми параметрами и, с помощью сенсоров, регистрируется ряд параметров объекта при фиксированных других параметрах. Тогда концептуальную модель эксперимента можно представить с помощью приведенной на Рис.1 ER-диаграммы (диаграмма «сущность-связь») [2].

Диаграмма содержит три класса сущностей: ОБЪЕКТ, ФАКТОР и СЕНСОР. Каждый класс с набором атрибутов, определяемых конкретной областью исследований. Взаимоотношения сущностей выражаются двумя классами связей: ВОЗДЕЙСТВИЕ и ОТКЛИК. В случае пассивного эксперимента или наблюдения класс сущностей ФАКТОР может отсутствовать.

Для перехода к реляционной модели данных заменим сущности и связи ER-диаграммы на соответствующие отношения R с первичными ключами K и атрибутами A:

ОБЪЕКТ – *R1(K1, A11, A12, … , A1n);*
ФАКТОР – *R2(K2, A21, A22, … , A2n);*
СЕНСОР – *R3(K3, A31, A32, … , A3n);*
ВОЗДЕЙСТВИЕ – *R4(K1, K2);*
ОТКЛИК – *R5(K1, K3).*

Вспомогательные отношения *R4* и *R5* служат для организации связи типа M:N (многие-ко-многим) между отношениями *R1*, *R2* и *R1*, *R3* соответственно. Первичными ключами K могут служить, например, порядковые номера кортежей соответствующих отношений. В общем случае ключевые атрибуты должны содержать значения из конечных множеств P:

P1={ K11, K12, … , K1m };
P2={ K21, K22, … , K2m };
P3={ K31, K32, … , K3m }.

Такая модель позволяет организовать адресацию файлового архива, имеющего следующую иерархическая структуру:

/ data / P1 / P2 / P3 / P1 P2 P3 N (1).

где строка «*P1 P2 P3 N*» - имя файла данных, образованное конкатенацией атрибутов *P1*, *P2*, *P3* и номера канала сенсора N для многоканальных сенсоров. Данная структура соответствует естественной древовидной структуре файловой системы.

Предложенная модель данных в сочетании со способом адресации неструктурированных данных обеспечивает естественную однозначную связь между записями в базе данных и соответствующими файлами. В то же время пользователи могут полностью абстрагироваться от имен или шаблонов имен файлов и каталогов, работая только с атрибутами, каталогизирующими свойства и происхождение каждого файла.

Как и модель эксперимента, модель данных является обобщенной, ее необходимо адаптировать для каждой конкретной научной области и вида экспериментов. В большинстве случаев может потребоваться декомпозиция отношений ОБЪЕКТ, ФАКТОР и СЕНСОР с учетом функциональных зависимостей между атрибутами.

2. Структурная схема

Структурная схема информационно-вычислительной системы (ИВС), реализующей изложенную выше концепцию управления данными научных экспериментов представлена на Рис. 2. Пользователи взаимодействуют с системой с помощью стандартного web-браузера, посылая запросы на поиск, анализ и визуализацию данных. В запросе на поиск указываются интересующие пользователя параметры объекта, параметры воздействующих на объект факторов и параметры сенсоров, регистрирующих данные. Запрос на анализ должен содержать перечень процедур анализа, которые будут применены к найденным данным и параметры этих процедур.

В результате выполнения запроса на поиск из базы данных извлекаются необходимые для обращения к файловому архиву атрибутивные данные. На основе этих данных web-приложение формирует адреса файлов в соответствии с (1) и передает их модулю анализа. Модуль анализа представляет собой приложение, выполняющее анализ данных в соответствии с алгоритмами, применяемыми в конкретной области экспериментальных исследований. В большинстве случаев это классические и специальные математико-статистические процедуры анализа многомерных числовых массивов.

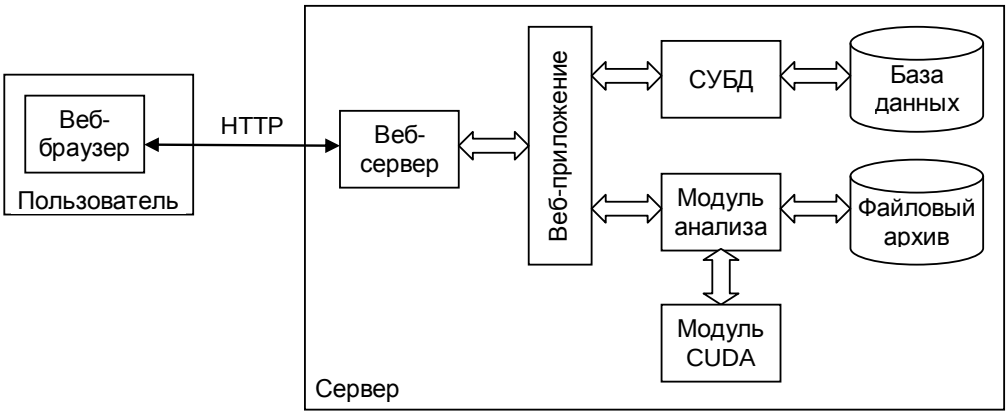

Рис. 2. Структурная схема информационно-вычислительной системы

Для обеспечения необходимого для on-line режима быстродействия при обработке данных ресурсоемкие вычислительные процедуры могут выполняться программно-аппаратным модулем на основе архитектуры CUDA. В модуле применяются графические процессоры (GPU) и математические библиотеки компании NVIDIA [3]. За счет многоядерной параллельной архитектуры GPU превосходят по быстродействию процессоры общего назначения (CPU) на 1–3 порядка при выполнении большинства вычислительных процедур.

Полученные в результате анализа числовые массивы возвращаются web-приложению, которое «на лету» формирует графики, таблицы, текст и отправляет всё это пользователю в виде готовой web-страницы. Конечное представление информации реализуется при помощи клиентских технологий JavaScript, HTML и стилевых таблиц CSS.

Конкретная аппаратно-программная реализация структурной схемы Рис. 2 определяется масштабами системы, сложностью применяемых алгоритмов анализа, количеством пользователей и т.д.

3. Управление геопространственными данными

Во многих научных областях (геофизика, экология и т.д.) исследователи имеют дело с пространственно обусловленными данными или геопространственными данными. Поэтому архитектура ИВС должна предусматривать подсистему управления геоданными и картографическую подсистему.

Большинство современных СУБД, как коммерческих, так и свободно распространяемых, поддерживают класс пространственных данных непосредственно или с помощью специальных расширений. Картографические сервисы, в частности web-сервисы до недавнего времени строились преимущественно на основе специализированного серверного ПО, позволяющего публиковать в сети Интернет карты, сопровождаемые базовым ГИС-инструментарием. Однако, в последние годы в Интернете все большее распространение получают гибридные ГИС. В таких системах геоданные из прикладной базы данных интегрируются с картографическим сервисом, предоставляемым специализированным web-сервером. На сегодняшний день наиболее развитым картографическим web - сервисом является Google Maps компании Google [4]. Сервис базируется на данных дистанционного зондирования (спектрозональные снимки со спутников Landsat, SPOT, Quickbird с разрешением до 0.68м) совмещенных с топографическими картами в проекции Меркатора. Компания Google предоставляет пользователям интерфейс Google Maps API в виде классов объектов JavaScript для генерации карт и нанесения на них собственных маркеров, контуров, а также готовых слоев в формате KML. Данные для отображения могут находиться как непосредственно в коде web-страниц, так и во внешних XML и KML файлах. Схема взаимодействия сервера ИВС, сервера Google Maps и клиентского браузера показана на Рис. 3.

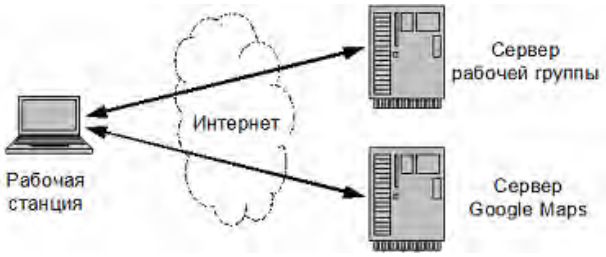

Рис. 3. Структура гибридной ГИС

4. ИВС «Вибросейсмическое просвечивание Земли»

На основе изложенной выше концепции нами была разработана база данных и web-ориентированная информационно-вычислительная система «Вибросейсмическое просвечивание Земли». Система предназначена для управления данными экспериментов по активному вибросейсмическому мониторингу литосферы, которые проводились на протяжении 1995–2012 годов институтами Сибирского отделения Российской академии наук

совместно с другими отечественными и зарубежными научными организациями.

ИВС реализует следующие основные функции:

- получение подробной информации по любому из проведенных экспериментов (метаданные);
- поиск в базе данных одновременно по 18 параметрам вибропросвечивания (типы источников, вид и параметры сигналов, географические координаты и др.);
- интерактивный on-line анализ (корреляционный, спектральный, спектрально-временной и т.д.) найденных сейсмотрасс с отображением результатов непосредственно в web-браузере пользователя
- построение по результатам поиска интерактивных карт и спутниковых снимков с обозначенными источниками и регистраторами сейсмических волн.

В настоящее время система доступна по адресу http://opg.sscc.ru [6].

Литература

[1] Григорюк А.П., Брагинская Л.П. Интернет-ресурс по вибросейсмическому мониторингу. Современное состояние и перспективы развития // Горный информационно-аналитический бюллетень. 2009. Т. 18, № 12. С. 200—204.

[2] Дейт К.Д. Введение в системы баз данных. 7-е изд. М.: Вильямс, 2001. 702 с.

[3] CUDA Zone. URL: http://www.nvidia.ru/object/cuda_home_new_ru.html.

[4] Google Maps API Documentation. URL: http://www.google.com/apis/maps/documentation/.

[5] Gray J., Liu D.T., Nieto-Santisteban M., Szalay A., DeWitt D., Heber G. Scientific Data Management in the Coming Decade // SIGMOD Record. 2005. Vol. 34. No. 4. Dec.

Experimental Data Magament Using Modern Web Ttechnologies

A.P. Grigoriuk, L.P. Braginskaya

Conceptual basis, architecture and software of web-oriented data-processing system for managing structured and unstructured data of natural and computational experiments are considered. The system provides search, computational analytical and GIS services to work effectively with the data.

As an example of the practical implementation of the proposed architecture is presented developed by the authors of the report data-processing system "Vibroseismic Earth Sounding" for the management data obtained during the experiments on active vibroseismic monitoring.

♣Работа выполнена при поддержке Российского фонда фундаментальных исследований, гранты № 05-07-90081 и № 07-07-00106.