Разработка распределенных алгоритмов и высокопроизводительной программной системы для облачного хранения, потоковой обработки и сбора в реальном времени сверхбольших наборов научных данных*

М.Н. Жижин, А.Н. Поляков, А.А. Пойда, Д.П. Медведев

НИЦ «Курчатовский институт»

В статье описываются предварительные результаты, полученные в ходе первого этапа выполнения научно-исследовательской работы по разработке программных систем, объединяющих высокопроизводительные технологии и параллельные алгоритмы управления сверхбольшими наборами научных данных, адаптируемых для решения широкого круга многодисциплинарных научных задач в различных областях приложения, включая: физику высоких энергий и астрофизику, науки о Земле и глобальные изменения климата, дистанционное зондирование и обработку многомасштабных изображений, биоинформатику, нанотехнологии и т.д.

1. Введение

Целью исследований является разработка программных систем, объединяющих высокопроизводительные технологии и параллельные алгоритмы управления сверхбольшими наборами научных данных, адаптируемых для решения широкого круга многодисциплинарных научных задач в различных областях приложения, включая: физику высоких энергий и астрофизику, науки о Земле и глобальные изменения климата, дистанционное зондирование и обработку многомасштабных изображений, биоинформатику, нанотехнологии и т.д.

Необходимость автоматизации управления научными данными вызвана прежде всего лавинообразным нарастанием объемов собираемых экспериментальных и данных и увеличением их сложности, связанным с ростом числа и разрешающей способности научных сенсоров, а также с экспоненциальным ростом вычислительных возможностей и объемов результатов вычислений, требующих хранения для повторного анализа.

Еще одним важным направлением исследований является адаптация существующий сервисов доступа и управления данными, работающими преимущественно с локальными файловыми системами и отчасти с удаленными наборами данных, к работе с облачным хранилищем данных. За счет высокого параллелизма и множественности точек доступа, в системах облачных вычислений есть возможность превысить производительность, показываемую при работе с локальными данными.

Специализированные системы хранения, такие как RasDaMan [1] и SciDB [2, 3], представляют большой интерес для научных приложений, так как они используют модель данных, специально адаптированную под научные приложения, при этом к их недостаткам можно отнести относительную сложность настройки и использования, в частности сложности при масштабировании на большое число машин.

Распределенные системы хранения общего назначения, такие как Cassandra [4] и Swift [5], удобны тем, что они позволяют пользователю максимально абстрагироваться от аппаратного обеспечения и обеспечивают хорошие показатели отказоустойчивости и масштабируемости. При этом используемая модель данных слишком проста для многих научных задач.

Наиболее перспективным представляется направление исследований, сочетающее особенности специализированных систем хранения научных данных с гибкостью и надежностью современных распределенных систем хранения общего назначения NoSQL, подобных Cassandra и Swift. Возможным вариантом реализации такой гибридной системы может стать разработка промежуточного программного обеспечения, реализующего

* Государственные контракты № 07.514.11.4045 от 26.09.2011 г. и № 07.514.11.4022 от 23.09.2011

специализированную модель данных поверх хранилища более общего назначения, подобно тому как RasDaMan опирается на традиционную реляционную СУБД для хранения данных.

Для проведения научных исследований чистых данных мало, необходимо иметь технологию для их анализа и обработки. В работе с большими наборами данных, хранящимися на нескольких удаленных серверах, возникает задача о том, чтобы системы анализа и обработки работали распределенно на удаленных серверах, ближе к месту хранения данных. Но распределенная обработка решает задачу обработки и анализа сверхбольшого объема данных. Какими бы характеристиками не обладал сервер, размер данных может быть настолько велик, что не поместится в оперативную память целиком. В этом случае придется делать свопинг на жесткий диск, что приведет к потере скорости. При этом в конвейере распределенных обработчиков возникнет простой: следующий обработчик не сможет начать работу пока не закончит предыдущий.

Одним из решений, устраняющих вышеперечисленные недостатки, является использование потоковой обработка данных, для которой объектом обработки становится не набор данных или его законченный фрагмент, а непрерывный поток данных. Обработчики могут использовать лишь небольшое окно в этом потоке для вычисления производных величин.

Проведя анализ существующих программных систем для потоковой обработки, исполнители проекта остановились на двух: OGSA-DAI [7,8] и Twitter Storm [9]. Сравнивая OGSA-DAI и Twitter Storm, надо отметить следующие особенности:

1. Twitter Storm изначально создавался для обработки потоков, тогда как система OGSA-DAI позволяет как передачу целых наборов данных, так и поблочную передачу с последовательной передачей.

2. В обеих системах пользователю потребуется реализовывать свои элементы обработки на языках программирования. В системе Twitter Storm поддерживается большее число языков программирования, чем для OGSA-DAI. В системе Twitter Storm достаточно написать jar-файл с кодом, который система автоматически подключит, в то время как в системе OGSA-DAI требуется вручную изменять конфигурационные файлы и перезагружать систему.

3. Storm автоматически распараллеливает работу между узлами, в то время как при организации потока в OGSA-DAI требуется явно прописывать хосты обработчиков.

4. Обработчики в OGSA-DAI могут накапливать (кэшировать) данные, в то время как в системе Twitter Storm нет встроенных функций для организации хранилища. Это может затруднить организацию блочно-потоковых алгоритмов (т.е. алгоритмов, в которых используется не только текущее значение потока, но и некоторое число предшествующих элементов).

Перечисленные особенности, кроме пункта 4, склоняют выбор технологии потоковой обработки в пользу системы Twitter Storm, но невозможность организовать поблочную схему обработки может стать существенным препятствием к реализации важных алгоритмов.

2. Постановка задачи

Мировая практика работы со сверхбольшими наборами данных и потоками данных с сенсорных сетей в реальном времени накопила положительный опыт использования данных технологий в различных областях человеческой деятельности. Существуют частные решения как проприетарные, так и открытые для внедрения параллельных и распределенных облачных технологий хранения и обработки данных в различных предметных областях. В России имеются собственные технологии хранения и доступа к большим массивам данных, например, созданный авторами настоящего отчета Интерактивный ресурс данных по солнечно-земной физики SPIDR. Но на сегодня пока не существует достаточно универсальной междисциплинарной платформы для интеграции распределенных массивов, потоков и сервисов данных.

Первая задача состоит в создании эффективного способа распределенного хранения и параллельного чтения-записи данных с учетом слабо структурированного характера научных данных (многие коллекции данных представляют собой просто многомерные числовые массивы). При этом метаданные должны обеспечивать эффективный поиск источников (сервисов) данных, а также содержать всю необходимую информацию о содержании,

происхождении (data provenance [6]) и различных характеристиках того или иного набора данных. Также необходимо предусмотреть возможность помечать и комментировать отдельные элементы данных или более крупные фрагменты наборов данных, например, выделение отдельных событий во временных рядах.

Вторая задача состоит в создании гибкого и достаточно простого в использовании языка запросов, учитывающего гетерогенный характер данных. При этом необходимо исходить из ограниченного набора моделей и форматов данных, общепринятых в конкретных предметных областях.

Третья задача состоит в создании индексов для быстрого поиска и обработки распределенных данных для создания сервисов и языков управления рабочим потоком распределенных вычислений (locality), которые оптимизированы по использованию памяти и ресурсов локального диска и отказоустойчивы при работе с большими объемами данных и позволяют кэшировать промежуточные результаты.

3. Реализация

Решение перечисленных задач может лежать в следующем:

1. Для эффективного извлечения новых знаний из сверхбольших объемов данных, ученые все чаще обращаются к распределению параллельных вычислений в облаке (data cloud). Вычислительное облако представляет собой абстракцию для удаленных, неограниченно масштабируемых вычислений и объемов хранения. На практике оно базируется в больших ЦОД, содержащих тысячи серверов и дисковых носителей. Для того чтобы можно было надежно хранить в прямом доступе сверхбольшие наборы научных данных, они должны быть распределены и тиражированы на тысячах серверов.

2. Необходимо разработать методы и алгоритмы, позволяющие создавать высокопроизводительные распределенные облачные хранилища сверхбольших массивов научных данных на основе общей модели данных и метаданных с возможностью отслеживания происхождения и цикла изменения данных и реализующие общий язык запросов для распределенного выполнения и управления рабочим потоком параллельных вычислительных задач по их обработке, анализу и визуализации.

3. Разработанные методы и алгоритмы должны позволить интеграцию и совместный анализ в следующих предметных областях:

- в экспериментах, обсерваторских наблюдениях и вычислительных моделях физики высоких энергий и астрофизики;
- в вычислительных моделях и сенсорных сетях в области метеорологии, экологии, глобального изменения климата;
- для сбора в реальном времени, распределенного хранения и интерактивного многомасштабного анализа данных дистанционного зондирования Земли из космоса;
- в вычислительных моделях и сетях сенсоров и обсерваторий для фундаментальной и прикладной геофизики.

4. Для реализации сервисов распределенной потоковой обработки данных необходимо:

- обеспечить конечному пользователю возможность создания, отслеживания, управления рабочим потоком на множестве реализованных обработчиков в соответствии с требованиями базовой системы (OGSA-DAI или Twitter Storm);
- обеспечить взаимодействие с сервисами распределенной потоковой обработки данных по технологии REST;
- реализовать библиотеку потоковых алгоритмов для научного и ниженерного анализа данных; в первую очередь будут реализованы алгоритмы: детектирования сейсмических P-волн [10], быстрого преобразования Фурье, алгоритмы модального анализа колебаний по методам «peak-picking» и «frequency domain decomposition» [11].

5. Для проверки эффективности разработанных методов и алгоритмов необходимо создать экспериментальный образец программного комплекса, осуществляющий решение основных

прикладных задач, возникающих при создании центров данных и облачных сервисов для работы со сверхбольшими наборами научных данных, источниками которых являются сети сенсоров и вычислительные модели в предметных областях, указанных в разделе.

6. Должны быть разработаны эталонные наборы данных и тестовые примеры и методики, обеспечивающие измерение и сравнение эффективности различных способов хранения и типов обработки и анализа сверхбольших массивов данных, источником которых являются вычислительные модели и сенсорные сети, включая оценки надежности, скорость доступа и масштабируемости хранилища, оценки функционала и производительности алгоритмов и программ обработки и анализа данных; критерии эффективности созданных систем хранения/передачи, управления и визуализации масштабными потоками и многомерными массивами научных данных и т.п. Список создаваемых эталонных наборов данных должен включать:

1) в области астрофизики и физики высоких энергий временные ряды наблюдений Мировой системы центров данных и обсерваторий космических лучей, геомагнитных обсерваторий INTERMAGNET и солнечных радиотелескопов Radio Solar Telescope Network (RSTN);

2) в области метеорологии и изменений климата архивы наблюдений Всемирной метеорологической организации и реанализ климата NCEP/NCAR Reanalysis, а также загрузку потоков данных в реальном времени для глобального оперативного прогноза NWS HOAA США;

3) в области дистанционного зондирования архив и обработка в реальном времени потока данных с американских метеорологических спутников DMSP (Defence Meteorological Satellite Program);

4) в области фундаментальной и прикладной геофизики архивы наблюдений и обработка в реальном времени потоков данных с сети сейсмических станций и GPS обсерваторий ДВО РАН.

Разработчики видят следующие направления исследований в области процесса обработки данных:

- Разработка RESTful сервиса создания и управления рабочим потоком, поддерживающая возможность распределенной параллельной обработки, а также REST-языка запросов к нему. Система управления рабочим потоком должна позволять использовать как существующие сервисы обработки (в том числе локальные программы и удаленные REST сервисы), так и позволять создавать собственные.
- Разработка отдельных RESTful сервисов обработки данных, которые можно использовать в рабочем потоке.

В качестве реализации первого направления может быть выбрана одна из существующих систем управления создания и управления рабочим потоком.

Из рассмотренных ранее систем управления рабочим потоком в системах LONI, TAVERNA и KEPLER плохо организована функциональность организации распределенных вычислений на нескольких машинах. Однако из этих трех систем TAVERNA в наибольшей степени подходит для использования в распределенной среде, так как имеет серверное ядро и веб-API с поддержкой REST-интерфейса.

Система OGSA-DAI имеет наиболее мощный функционал для организации вычислений в распределенной среде, но не имеет REST-интерфейса. Также OGSA-DAI не имеет графического интерфейса, что может существенно осложнить понимание синтаксиса запросов и в конечном итоге привести к снижению его использования со стороны конечного пользователя.

Поэтому разработчики в данном направлении иследования видят два варианта: использовать TAVERNA и дооснащать ее функциональность либо использовать OGSA-DAI и писать для него REST-оболочку.

Набор сервисов во втором направлении исследования должен быть определен в ходе дальнейшей работы над проектом, однако среди обязательных сервисов (разработка каждого из них – отдельное направление исследования) должны быть:

- сервис поиска данных на основе нечеткой логики;

- сервис подготовки пирамиды сверхбольших изображений по технологии DeepZoom [12] и позволяющий визуализацию подготовленных изображений;
- сервис обработки данных средствами пакета прикладных математических программ Scilab;
- сервис анализа и визуализации геологических данных на основе системы CoreWall [13, 14];
- сервисы конвертации форматов данных, поддерживающие стандартные промежуточные форматы и сетевые протоколы представления данных и метаданных: FITS, GRIB, NetCDF/HDF, OPeNDAP, CDF.

Методы, по которым будут проводиться исследования по представленным выше направлениям, включают следующую последовательность действий:

1. Разработка архитектуры основополагающих механизмов сервиса создания и управления рабочим потоком.
2. Разработка архитектуры и спецификации базовых сервисов обработки данных (см. выше) в соответствии с требованиями, накладываемыми архитектурой, разработанной в пункте 1. Параллельно - уточнение архитектуры сервиса управления рабочим потоком.
3. Разработка спецификации расширенного набора сервисов обработки данных (исходя из конкретных предметных задач) в соответствии с требованиями, накладываемыми архитектурой, разработанной в пункте 1 и доработанной в пункте 2. Параллельно – доработка архитектуры сервиса управления рабочим потоком.
4. Создание прототипа системы управления рабочим потоком. При необходимости – доработка спецификаций и архитектур, разработанных в пунктах 1-3.
5. Реализация прототипов базовых сервисов обработки данных, их соединение через систему управления рабочим потоком в экспериментальный образец.
6. Проведение профилирования и анализ полученных результатов. Оптимизация и доработка экспериментального образца.
7. Реализация прототипов расширенного набора сервисов из пункта 3, их внедрение в экспериментальный образец.
8. Анализ и доработка системы, выводы.

4. Заключение

Создаваемые методы и программы могут быть использованы как в фундаментальных приложениях, таких как изучение влияния солнечной активности на климат, так и в чисто практических целях, например для экологического мониторинга, в частности прогноза распространения вулканического пепла или радиоактивных загрязнений. Методы, применяемые в настоящее время для этой цели, в значительной степени определяются конкретными задачами и объектами, для которых они разрабатывались, и неэффективны для сравнения и анализа сврехбольших наборов данных, возникающих на стыке естественнонаучных дисциплин. Областью приложения алгоритмов автоматизированного управления научными данными явятся задачи, возникающие в области сбора и оперативного анализа потоков данных с большого числа сенсоров или детального численного моделирования динамики астро-, гео- и биофизических явлений . Кроме того, мы предполагаем, что часть создаваемых методов найдет применение в анализе телеметрии и изображений для медицинскими приложений.

Накопленные в результате работ по проекту сверхбольшие наборы данных и созданные информационные технологии должны упрощать междисциплинарную интеграцию и позволять масштабируемое и надежное хранение научных данных с общими моделью данных и языком запросов, ускорять запись и чтение данных, позволять распределенный поиск метаданных и отслеживание происхождения данных (data provenance), облегчать интерактивный анализ и визуальный поиск закономерностей, а также повторное использование данных и гибкое управление доступом.

Коммерциализация полученных в рамках проекта научно-технических результатов возможна в виде заключения договоров с лабораториями, занимающимися прикладными и

фундаментальными исследованиями, связанными со сбором и анализом сверхбольших наборов данных, по оказанию услуг в области доступа к уже имеющимся «эталонным» сверхбольшим базам данных, хостинга наборов данных для заказчика и организации сервисов по коммерческому доступу, аренде вычислительных ресурсов для анализа и аппаратно-программных средств визуализации этих данных.

Литература

1. Peter Baumann, Andreas Dehmel, Paula Furtado, Roland Ritsch, Norbert Widmann: *The Multidimensional Database System RasDaMan.* Proceedings ACM SIGMOD'98, Seattle, Washington, USA. June 1998. http://www.faculty.jacobs-university.de/pbaumann/iu-bremen.de_pbaumann//Papers/sigmod98.zip
2. Олег Бартунов, Павел Велихов и др., SciDB – новая СУБД для больших объемов научных данных, 2011, http://supercomputers.ru/index.php?option=com_k2&view=item&id=167:scidb
3. Сергей Кузнецов, Год эпохи перемен в технологии баз данных, 2009, http://citforum.ru/database/articles/epoch/
4. Eben Hewitt, Cassandra: The Definitive Guide, O'Reilly Media, 2010
5. Ken Pepple, Deploying OpenStack, O'Reilly Media, 2011
6. Efficient provenance storage over nested data collections. Anand, Manish Kumar and Bowers, Shawn and McPhillips, Timothy and Ludascher, Bertram (2009) . Pages: 958--969. url: http://portal.acm.org/citation.cfm?id=1516470
7. Jackson, M., Antonioletti, M., Dobrzelecki, B., Chue Hong, N. Distributed data management with OGSA-DAI. Grid and Cloud Database Management (eds. Fiore, S. and Aloisio, G.), Springer-Verlag, July 2011, pp63-86. ISBN 978-3-642-20044-1. DOI 10.1007/978-3-642-20045-8.
8. Dobrzelecki, B., Krause, A., Hume, A., Grant, A., Antonioletti, M., Alemu, T., Atkinson, M., Jackson, M. and Theocharopoulos, E. Integrating distributed data sources with OGSA–DAI DQP and Views. Phil. Trans. R. Soc. A 13 September 2010 vol. 368 no. 1926 4133-4145 DOI: 10.1098/rsta.2010.0166.
9. Распределенная вычислеительная система Twitter Storm. Сайт проекта, url: https://github.com/nathanmarz/storm/wiki
10. M.N. Zhizhin, D. Rouland, J. Bonnin, A.D. Gvishiani, A. Burtsev. Rapid Estimation of Earthquake Source Parameters from Pattern Analysis of Waveforms Recorded at a Single Three-Component Broadband Station, Port Vila, Vanuatu. Seismological Society of America, 2006. Vol. 96, no. 6.
11. Zimmerman, A. T., Shiraishi, M., Swartz, R. A. and Lynch, J. P., "Automated modal parameter estimation by parallel processing within wireless monitoring systems," *Journal of Infrastructure Systems*, 14(1), 102-113, 2008.
12. Технология построение пирамид изображений DeepZoom. Сайт проекта, url: http://msdn.microsoft.com/ru-ru/library/cc645050%28v=vs.95%29.aspx
13. Chen, Y., Hur, H., Lee, S., Leigh, J., Johnson, A., Renambot, L. Case Study - Designing An Advanced Visualization System for Geological Core Drilling Expeditions. Proceedings of the ACM Conference on Human Factors in Computing Systems 2010 (CHI 2010), Atlanta, GA, 04/10/2010 - 04/15/2010
14. Conze,R., Krysiak, F., Reed, J., Chen, Y., Wallrabe-Adams, H., Graham, C. and the New Jersey Shallow Shelf Science Team, Wennrich, V. and the Lake El'gygytgyn Science Team. New Integrated Data Analyses Software Components. Scientific Drilling Journal, ISSN: 1816-8957, 04/01/2010 - 04/01/2010