НАУЧНЫЕ ВЫЗОВЫ ТЕХНОЛОГИЯМ СУБД

О. Бартунов

Введение

Повсеместное распространение интернет, ускорение и унификация доступа к информации и т.п. привело к формулированию концепции кибероббщества (информационного общества) как реалистичного сценария постиндустриального общества - новой исторической фазы развития цивилизации, в которой главными продуктами производства являются информация и знания.

Составной частью информационного общества является так называемая e-Science - синтез науки и информатики, наступающий когда роль информации и ее обработки в научных исследованиях становится превалирующей. Переход на е-стадию (информационную стадию) развития - реальная ситуация, затронувшая на сегодняшний день ряд естественных наук, оперирующих огромными объемами информации: физика (в первую очередь исследование элементарных частиц и физика высоких энергий), геофизика и геология, астрономия, биология, экономика, медицина. В этих науках происходит процесс лавинного поступления информации, в первую очередь связанный с успехами в технологии создания приемных устройств (сенсорно-ориентированная наука). Также, приходится работать с гигантскими объемами синтетических данных, полученными при численном моделировании. В современных крупных экспериментах (включая и численное моделирование) анализ терабайтов и даже петабайтов научных данных становится повседневной задачей.

Причины "информационного взрыва" в астрономии обусловлены следующими факторами:

- Астрономия стала всеволновой. Начиная с 70-х годов прошлого века наблюдения астрономических объектов ведутся не только в видимом свете, как раньше, а во всем диапазоне электромагнитного спектра, также регистрируются другие виды частиц и сигналов.
- Астрономические данные хранятся бесконечно долго. Так как данные астрономических наблюдений привязываются к конкретным объектам, то их необходимо хранить пока эти объекты существуют. Времена эволюции астрономических объектов очень велики, в обыденном смысле с хорошей точностью могут считаться бесконечными.
- Астрономия снова стала широкопольной. До конца XIX века наблюдения велись визуальным способом и давали информацию об отдельных объектах: одно наблюдение - один объект. Ситуация изменилась с появлением фотографии, на фотопластинках одновременно фиксировалось большое количество объектов. Ценность этой информации была ясна с самого начала, астрономические фотопластинки, снятые с последней четверти XIX до конца XX века хранятся в так называемых "стеклянных библиотеках". Во второй половине прошлого века широкопольную астрофотографию потеснили гораздо более точные электронные методы фотометрии в которых, однако, одновременно можно было наблюдать только один объект (или небольшое количество объектов). Возвращение к "широкопольности" произошло после появления ПЗС-матриц большого размера. Сегодня одно наблюдение, длящееся от нескольких секунд до нескольких минут, дает от нескольких мегабайт до нескольких гигабайт информации.
- Политика доступа к информации. Данные всех космических и наземных экспериментов NASA, ESA и частично ESO становятся публично доступными спустя 1 год после их получения. Последние 10 лет КТБТ (Комитет по Тематике Больших Телескопов - занимается распределением наблюдательного времени на крупнейших оптических инструментах России) требует открытия данных через 2 года после их получения.
- Этому же способствует так называемая "Early Science"("Быстрая наука"): необходимость исследовать и публиковать полученные данные в строго определенный срок для того, чтобы успеть подготовить и подать успешную заявку на следующий цикл исследований. Это приводит к предельной интенсификации изучения полученных данных (возможно, делает его существенно более поверхностным).

Доступ к информации осложнен тем, что результаты наблюдений хранятся в неоднородных распределенных архивах. Разнородность архивов определяется тем, что они создавались независимо и были ориентированы на различные эксперименты со своими целями. Распределенность информации связана со следующими причинами:

- На сегодняшний день нет (и скорее всего, не будет и в будущем) технических возможностей содержать всю астрономическую информацию в одном хранилище из-за слишком больших объемов информации.
- Создание нескольких копий информации в различных местах повышает надежность хранения информации.
- Распределенное хранение и наличие копий снижает нагрузку на сеть и повышает скорость доступа к информации.
- Необходимость обработки данных определенного эксперимента требует их локализации на достаточно длинный срок для быстрого доступа.
- В некоторых случаях распределенное хранение информации возникает по "физическим" причинам: Например, в эксперименте SNAP (орбитальный телескоп) большой поток информации и отсутствие

существенных объемов памяти на борту обсерватории приводит к построению распределенной системы центров по приему и дальнейшему хранению данных.

Специфика научных данных состоит в необходимости "вечного" хранения сырых данных (raw data, - это данные, полученные непосредственно с приемника и не подвергшиеся никакой обработки) , что накладывает повышенные требования к масштабируемости и защищенности систем хранения. Отметим общие проблемы и особенности современной науки, связанные с увеличивающимся потоком данных (сейчас это сотни терабайтов, в ближайшие 5-10 лет - это десятки-сотни петабайт):

- количество "сырых" данных только увеличивается и их необходимо хранить вечно, так как может потребоваться их переобработка.
- очень сложная процедура получения научных данных из "сырых" данных. Развитие сенсоров только увеличивает разрыв между "сырыми" данными и научными данным, и зачастую требуются большие вычислительные ресурсы для получения научных данных. Задача усложняется тем, что современные научные эксперименты представляют собой сложный комплекс уникальных приборов, требующих специализированных методов обработки получаемых "сырых" данных, практически всегда несовместимых друг с другом.
- Еще одна особенность современных научных экспериментов - это сочетание распределенного хранилища данных с необходимостью доступа к высокопроизводительным вычислительным комплексам для получения научных данных и их анализа. Такие вычислительные комплексы в настоящее время в основном используются для решения расчетных задач, не требующих работы со сверхбольшими данными. Однако, гигантские объемы данных полностью исключили традиционный раннее способ работы - загрузка данных из хранилища на сервер для обработки. Причем, основная проблема состоит в стоимости каналов связи, а не хранилища. Все работы по обработке данных эксперимента требуется вести в самом хранилище с использованием вычислительных кластеров. С другой стороны, и в задачах численного моделирования появились требования к возможности сохранения текущего состояния в СУБД, например, расчет космологической эволюции Вселенной требует сотни гигабайт для сохранения одного "слепка" Вселенной. Подобные расчеты ведутся на распределенных кластерах с тысячами процессоров и возможность работы с такими данными в СУБД позволяет проследить историю эволюции отдельных объектов Вселенной (частицы, звезды, галактики, скопления галактик…).
- Обычно, из-за низкой производительности современных систем, исходные данные научных наблюдений хранятся вне каких-либо СУБД, и только метаданные индексируются в базе данных. Для доступа и обработки исходных данных научным коллективам приходится разрабатывать свои программные системы под каждую конкретную задачу. При таком подходе очень трудно поддерживать целостность данных, версионность данных, историю их изменений, получение научных результатов из "сырых" данных, что затрудняет поддерживание одного из основных принципов науки - повторяемости научных результатов.
- сложная организация проектов - много участников, разные источники финансирования, что определяет необходимость поддержки определенной политики доступа к данным. С другой стороны, в науке ценят доступность данных, лицензионные ограничения на использование СУБД могут мешать свободному обмену данными. Кроме того, закрытая лицензия может мешать развитию программных средств.
- распределенность данных - данные хранятся в разных научных центрах для локализации трафика, по физическим причинам, резервирование данных, масштабирование нагрузок;
- очень трудно отслеживать изменяемость данных, например, изменилась процедура обработки "сырых" данных, добавились новые данные, итд; Данные должны сопровождаться информацией о происхождении (источник, автор, качество,..). Это очень важно, так как в науке нередки запросы, в которых участвуют данные из разных архивов и надо быть уверенным, что, например, устраивает качество данных. Это называется data provenance, lineage, pedigree. Очень важный аспект data provenance - это query inversion. Представим, что у вас на сайте публикуется автоматически сгенеренный по базе данных график распределения какой-то величины, и в одно прекрасное утро вы замечаете на нем важные изменения и естественное желание ученого узнать из-за чего это произошло заставляет его рыться в базе, программах, разного рода логах поступления данных, работы коллег итд. Это безумно тяжелая работа ! Более строго можно сказать так: Найти какие записи в БД (изменения в каких записях) повлияли на результат работы запроса, т.е. - это обратная задача к обычному запросу.
- Аннотирование данных - это возможность хранить пометки разной степени детализации - на уровне таблицы, на уровне конкретного значения. Требуется эффективное хранение аннотаций и доступ к ним для баз данных петабайтного размера;
- популярные задачи анализа данных, поиска зависимостей в сверхбольших базах данных являются крайне неэффективными в силу немасштабированности архитектуры классических СУБД;
- очень большое разнообразие типов данных и запросов - трехмерные объекты, временные ряды, треки элементарных частиц и т.д.;
- нет поддержки работы с данными,которые имеют погрешность измерений, пропущенными данными;
- требование получения "быстрых" результатов ("Early Science");

- Машины стали основными производителями информации и ее потребителями, поэтому требуется обеспечить прежде всего не интерактивную работу с данными, а программный доступ к ним, чтобы можно было автоматизировать рутинные работы обработки наблюдений, поиска данных. Прежде всего это относится к проблеме эффективного хранения и доступа семантической информации в базах данных.

Эти проблемы необходимо срочно решать в ближайшее время, так как технологии производства приемных устройств (сенсоров) непрерывно улучшается, что приводит к дальнейшему росту данных, а следовательно, к усугублению описанных проблем.

Что такое очень большие базы данных ?

Следует различать базы данных как хранилища метаданных, которые содержат очень большое количество записей с активным доступом и базы данных, ориентированные на архивное хранение очень больших бинарных объектов (их также может быть очень много).

- На сегодня официально анонсирована самая большая в мире база данных с активным доступом - Yahoo Everest, которая на май 2008 года имела хранилище размером более 2 Pb, несколько триллионов записей, с ежедневным поступлением около 24 млрд событий и более 1/2 миллиарда пользователей в месяц. В 2009 году база данных доросла до 10Pb. Интересно отметить, что Yahoo Everest - это свободная СУБД PostgreSQL с распределенным вертикально-ориентированным хранилищем и поддержкой кластеризации. В 2010 году стало известно, что Yahoo рассматривает переход на Hadoop. Из планируемых научных экспериментов выделяются
- Большой Адронный Коллайдер (LHC, http://lhc.web.cern.ch/lhc/), который ежегодно будет производить около 15 Pb данных, распределенное хранилище будет состоять из примерно 200 центров данных по всему миру
- большого телескопа для обзора неба (LSST, http://www.lsst.org), с диаметром зеркала 8.4 метра и матрицей размером 3.2 Гп (гига-пикселей). Ожидается наполнение БД в 49 миллиардов объектов (256 атрибутов), 2.8 триллиона источников (56 атрибутов). К 2025 году ожидается накопить 14 Pb данных !
- Российский Космический Эксперимент "Лира" (КЭ Лира), который разрабатывается в ГАИШ-МГУ совместно с РосКосмос, планирует получение около 400 терабайтов сырых данных для получения многополосной высокоточной фотометрии звезд всего неба, в результате которого будет проведен большой ряд однородных наблюдений более 400 миллионов звезд.

Что же изменилось в компьютерных технологиях ?

- Порог петабайтных БД преодолен. Количество данных растет быстрее (http://www.b-eye-network.com/view/7188) чем закон Мура (http://en.wikipedia.org/wiki/Moore_Law).
- Данные стали разными, новые запросы - многомерные данные, запросы не ограничиваются операциями сравнения, например, найти 10 самых похожих изображений.
- Много запросов, другие требования к производительности и расширяемости - новые технологии (AJAX), динамические документы, увеличилось кол-во запросов, требование выполнение десятые доли секунды.
- Клиенты стали другими - раньше были операторы, сейчас в основном это бездушные клиенты, большей частью через http, большой уровень конкурентности.

Не удивительно, что сейчас насчитывается около сотни различных СУБД, начиная от класических реляционных баз данных (Oracle, SQL Server, PostgreSQL, MySQL, Firebird, Ingres,…), которые обладают богатым набором возможностей, но их архитектура закладывалась во времена одного (не сетевого) большого и дорого компьютера с маленькой памятью и одноядерным процессором, и кончая специализированными хранилищами, оптимизированных для решения определенных задач (Vertica, H-Store, StreamDB..). Посередине находятся СУБД, для которых самым важным является масштабирование и ограниченный набор возможностей. Эти СУБД ориентированы на современную многоядерную архитектуру дешевых серверов с большой памятью, организованных в кластера. Поскольку один сервер уже не справляется с нагрузкой, то имеется два способа масштабирования:

- Использовать реляционные СУБД с шардингом по большому количеству узлов. При этом многие свойства реляционной модели уже не поддерживаются (соединения,агрегаты, …);
- Использовать масштабируемое (ключ,значение) хранилище - это Project Voldemort, Scalaris, Dymomite, MemcacheDB, CouchDB, Cassandra, HBase, Hypertable, SimpleDB (Список NoSQL баз данных сейчас насчитывает около 40 баз данных). Для этих (ключ,значение) СУБД характерен уход от принципа целостности данных ACID к BASE, который более мягок и говорит о целостности базы данных "в конце концов" (eventually consistent).

Довольно часто вертикально-ориентированные базы данных отождествляют с нереляционными и NoSQL СУБД. На самом деле это не так, например,Vertica (C-Store), MonetDB - это реляционные СУБД с поатрибутным хранением и SQL. Более подробно о двух типах вертикально-ориентированных хранилищах.

Какой же класс СУБД годится для науки ? Очевидно, что богатые возможности реляционных СУБД крайне интересны для науки, но также очевидно, что строгая целостность и изоляция данных (CI в ACID) не важны, так как данные в науке в основном WORM (Write Once Read Many) и вполне достаточна eventual consistency. Кроме того, реляционной модели не присуща внутренняя упорядоченность, в то время как для

"сенсорно-ориентированной" науки, естественно хранить данные в массивах, которым присуща упорядоченность ! В реляционной модели реализация массивов очень неэффективна. Например, для сенсорных данных характерны задачи по поиску ближайших соседей, которая в реляционной модели может выглядеть очень логично и прозрачно

```
CREATE TABLE Observation (I integer NOT NULL, J integer NOT NULL, V float NOT NULL);

SELECT A1.I, A1.J, AVG(A2.V)
  FROM Observation A1, Observation A2
 WHERE A2.I BETWEEN A1.I – 1 AND A1.I + 1
   AND A2.J BETWEEN A1.J – 1 AND A1.J + 1
 GROUP BY A1.I, A1.J;
```

Однако, практическая реализация будет очень неэффективна.

Масштабируемость нужна по объему данных, но не нужна большая конкурентность и ориентированность на фиксированное время ожидания результата. В то же время, науке требуется более богатая модель данных нежели (ключ, значение). Многие науки согласились с тем, что наиболее важная структура данных - это многомерный вложенный массив с неровными краями и оптимизацией для разреженных данных. Если добавить сюда требования, специфические для науки, такие как поддержка версионности, происхождения, аннотирования данных, данных с ошибками итд, то приходим к выводу, что на сегодняшний момент нет СУБД, ориентированной на науку.

Майк Стоунбрейкер считает, что надо перестать "латать" устаревшие СУБД, что требуются кардинальные изменения в технологии СУБД, а именно — изменение принципа хранения данных. Он считает, что эра обычных больших СУБД общего назначения прошла (http://www.databasecolumn.com/2007/09/one-size-fits-all.html) и требуется совершенно новые подходы для создания современной БД, которая с самого начала будет ориентирована на распределенность, параллельное исполнение запросов, компрессию, ориентацию на хранение по атрибутам, высокую доступность, линейное масштабирование с использование кластеров независимых серверов.

Сложившаяся ситуация в больших научных проектах была оценена ведущими учеными из разных наук,представителями коммерческих компаний и разработчиками в области СУБД (систем управления баз данных) на серии конференций XLDB 2007,2008,2009 гг, в результате чего возник проект SciDB под руководством профессора MIT Майка Стоунбрейкера и его коллег из крупнейших университетов США. Основная цель проекта - разработка в кратчайшие сроки СУБД для нужд больших научных и промышленных проектов, в которых требуется анализ сверхбольших объемов данных (сотни и тысячи петабайт), масштабируемая на тысячи серверов.

Новая СУБД для больших объемов научных данных.

Система SciDB разрабатывается в первую очередь исходя из требований больших научных проектов и имеет ряд принципиальных отличий от существующих СУБД. SciDB разрабатывается как система для хранения и анализа сырых и производных научных данных. Некоторые основные функции традиционных баз данных не поддерживаются в SciDB, позволяя системе более эффективно обрабатывать аналитические запросы.

Например, так как исходные данные фактически не обновляются, в SciDB не предусмотрена эффективная поддержка больших объемов транзакций, что позволяет избежать серьезных накладных расходов. Наконец, SciDB – проект с открытым исходным кодом и бесплатной лицензией на использование, что отвечает требованиям большинства заказчиков. Открытый код позволяет экономить средства заказчиков на масштабные внедрения системы, а открытый процесс разработки обеспечивает высокое качество технических решений. Кроме того, открытость СУБД обеспечивает технологическую независимость и возможность обмена данными между разными научными коллективами.

Кроме привычных функций систем управления базами данных, в SciDB присутствуют новые механизмы работы с данными, специально разработанные для анализа научных данных. Модель данных SciDB представляет из себя многомерные вложенные массивы, таким образом ученым не надо моделировать свои данные как таблицы записей, что в свою очередь ведет к более простой формулировке аналитических запросов и на порядки увеличивает производительность системы. Так как в SciDB будут хранится данные полученные с приборов, SciDB поддерживает погрешность измерений на уровне модели данных и языка запросов. Наконец, SciDB изначально разрабатывается для работы на большом спектре вычислительных систем, от переносного ПК до больших кластеров и суперкомпьютеров. Таким образом, ученые смогут работать с данными в одной среде, например отлаживая аналитические алгоритмы на персональных компьютерах используя небольшую выборку данных, а отлаженные запросы без изменений запускать на высоко-производительных кластерах. Также, SciDB интегрируется с популярными вычислительными пакетами программного обеспечения, такими как R, Matlab и другие, что позволит ученым использовать уже готовые алгоритмы обработки данных при переходе на SciDB.

Основные характеристики разрабатываемой СУБД

- хранение "сырых" данных, их обработка происходит в самой СУБД с помощью пользовательских процедур для обеспечения версионности и истории изменения данных (новая идея) - полноценная поддержка полного цикла работы с научными данными;
- Модель описания научных данных - это многомерный вложенный массив (новая идея);
- Вертикальное (поатрибутное) хранение данных для компрессии и уменьшения операций ввода-вывода;
- Сохранность данных за счет репликации части данных на разных узлах системы;
- Масштабируемость СУБД от ноутбука до x1000 серверов для хранения x10 петабайтов;
- Расширяемость типов данных и запросов;
- Отказ от поддержки транзакций, которые не нужны для научных данных (WORM - Write Once Read Many), и которые сильно усложняют архитектуру СУБД и вносят существенные расходы на их поддержание. Вместо ACID будет использоваться модель BASE (eventual consistency), что вполне достаточно для научных данных.

Полноценная поддержка полного цикла работы с научными данными

Как упоминалось раньше, из-за недостатков существующих СУБД, большинство научных проектов, в которых встает задача анализа больших объемов данных, осуществляют обработку и анализ исходных данных вне системы управления базами данных. SciDB решает эту проблему, обеспечивая эффективное и удобное хранилище исходных данных и широкий набор инструментов для обработки и анализа данных. Версионное хранилище и учет всех преобразований данных позволяет пользователям SciDB получить точную информацию о версиях данных и о всех вычислениях, произведенных над исходными данными. Это позволяет эффективно устранять ошибки в алгоритмах переработки данных, отслеживать процесс переработки исходных данных при получении подозрительных результатов, и в точности повторять вычисления над исходными данными. При этом SciDB работает без каких-либо ограничений, как на суперкомпьютерном кластере, так и на персональном компьютере, что позволит ученым работать в одной и той же среде со своими данными. После переработки исходных данных, SciDB позволяет делиться полученными результатами, осуществлять выборки и выполнять аналитические запросы широкому кругу коллег, при этом соблюдая произвольный политику доступа как к данным, так и полученным результатам.

Сотрудничество с ведущими научными проектами

SciDB разрабатывается в тесном сотрудничестве с ведущими научными проектами – потенциальными заказчиками системы. В научный совет SciDB входят ученые от различных направлений науки, включая: астрономию, нанотехнологии, генетику, сейсмологию, ядерную физику, метеорологию, и др. При этом, два проекта, LSST (Large Synaptic Survey Telescope) и российский космический проект ЛИРА (многоцветный фотометрический обзор всего неба до 16-17 звездной величины), предоставили детальные требования для использования SciDB в своих системах и часть исходных данных. Следовательно, система SciDB разрабатывается прямо под требования заказчиков и проходит апробирование на реальных задачах уже в процессе разработки.

Космический Эксперимент "Лира"

КЭ "Лира" - это первый российский высокоточный многоцветный фотометрический обзор звезд всего неба до 16-17 звездной величины, над которым работают ГАИШ-МГУ (Государственный Астрономический институт им. П.К. Штернберга, Московский Государственный Университет им. М.В.Ломоносова) и ОАО РКК "Энергия" контракт No.351-8623/07 от 05.06.2007 г.. В обзор войдут около 400 млн. звезд. Уникальная методика наблюдений позволит получить точность измерения блеска для звезд предельной величины около 1%, а для ярких звезд (ярче 12 зв. Величины) - 0.1%. Измерения будут вестись в 10 спектральных полосах от 0.2 до 1.0 мкм (т. е. в оптическом и близком УФ и ИК диапазонах) с борта Российского сегмента МКС. Ожидаемый старт проекта - 2013 год. На протяжении 5 лет ожидается получить около 400 Тб данных, для хранения и обработки которых потребуется масштабируемое распределенное хранилище и мощный вычислительный кластер для получения научных данных их наблюдений и решения задач поиска закономерностей (data mining).

Текущий статус SciDB и планы развития

- Сформированы международные команды исследователей и разработчиков под руководством крупнейших авторитетов в области баз данных (Стоунбрейкер,ДеВитт,Майер и другие)
- Разработан прототип системы, который был представлен на крупнейших международных конференциях SIGMOD 2009 (Providence, USA) , VLDB 2009 (Lyon, France)
- На основе прототипа в первом квартале 2010 года планируется первая публичная версия SciDB для ознакомления научной общественностью.
- Ведутся периодические телефонные конференции для выработки совместных планов работы над следующей версией SciDB
- Американские исследователи и разработчики получили частичное финансирование от американских научных фондов
- К 2012 году планируется начало тестирование SciDB в проекте LSST

Российская команда разработчиков SciDB

Российские разработчики (НИИСИ РАН) приняли участие уже на самом раннем этапеработы над SciDB и заняли лидирующие позиции среди сновных разработчиков.

В дальнейшем к команде присоединились ведущие российские разработчики (ГАИШ МГУ) крупнейшей СУБД PostgreSQL, имеющие опыт не только в разработке СУБД, но и участия в крупных научных проектах и работе с очень большими базами данных. ГАИШ МГУ в рамках подготовки КЭ "Лира" работает над списком научных задач, выработкой требований, а также располагает серьезной инфраструктурой, необходимой для разработки и тестирования СУБД.

О проекте Лира

Уникальность проекта "Лира":

- Единственный в мире обзор всего неба в 10 полосах от ультрафиолета до ближнего ИК-диапазона
- Высокая фотометрическая точность наблюдений
- Высокая однородность наблюдений

В результате выполнения проекта будут получены важнейшие научные результаты:

- Высокоточный многоцветный каталог фотометрических стандартов для атмосферных и внеатмосферных наблюдений
- Многоцветный каталог переменных звезд объемом больше 30 миллионов звезд
- Построена трехмерная структура нашей Галактики (по межзвездному поглощению в ультрафиолете)
- Получены данные о физических характеристиках поверхности астероидов