25 лет прогнозов : что день грядущий нам готовит?

Сергей Кузнецов, ИСП РАН

25 лет прогнозов

- Future Directions in DBMS Research, 1988 г., Лагуна Бич, Калифорния
- NSF Invitational Workshop on the Future of Database Systems Research, Пало Альто, Калифорния, 1990
- NSF Workshop on the Future of Database Systems Research, Пало Альто, Калифорния, 1995
- Симпозиум "Strategic Directions in Computing Research", Кембридж, шт. Массачусетс, 1996
- The Asilomar Report on Database Research, Асиломар, неподалеку от г. Монтерей, Калифорния, 1998
- The Lowell Database Research Self Assessment, Лоуэлл, шт. Массачусетс, 2003
- The Claremont Report on Database Research, Berkeley, California, Claremont Resort, 2008
- The Beckman Report on Database Research, Irvine, California, Beckman Center of the University, 2013

Laguna Beach Report (1)

- Future Directions in DBMS Research
- 1988 г., Лагуна Бич, Калифорния
- SIGMOD Record 18(1), 1989, 17-26
- Philip A. Bernstein, Umeshwar Dayal, David J. DeWitt, Dieter Gawlick, Jim Gray, Matthias Jarke, Bruce G. Lindsay, Peter C. Lockemann, David Maier, Erich J. Neuhold, Andreas Reuter, Lawrence A. Rowe, Hans-Jörg Schek, Joachim W. Schmidt, Michael Schrefl, and Michael Stonebraker
- http://www.icsi.berkeley.edu/pubs/techreports/tr-88-001.pdf
- Будущие направления исследований в области баз данных: десять лет спустя
- http://citforum.ru/database/articles/future_01.shtml

Laguna Beach Report(2)

- Возможность широкого внедрения систем поддержки принятия решений по мере снижения цен аппаратуры
  - **Теперь аналитика повсюду и в самых разных формах**
- В области управления базами данных компьютеры общего назначения более перспективны, чем специализированная аппаратура
  - **Специализированная аппаратура почти не используется**
    - **Appliances???**
    - **Oracle RAC???**
    - **Графические процессоры (Соколинский)???**

Laguna Beach Report (3)

- Реструктуризация ОС к виду, имеющему малое ядро с набором сервисов поверх его
  - Микроядерные архитектуры ОС так и не вошли в широкий обиход
    - Minix, Таненбаум
- Важно улучшить существующий ужасный интерфейс SQL для встроенных и динамических запросов
  - Все, как было

Laguna Beach Report (4)

- Потребность в некоторых видах рекурсивных запросов, в частности, таких как запросы, выдающие транзитивные замыкания
  - **Рекурсивные запросы в SQL давно доступны, но кто ими пользуется?**
- Имеет смысл работать над стандартной моделью данных "следующего поколения«
  - **SQL, ODMG и Третий манифест**
  - **За последние 15 лет ничего больше**

The "Lagunita" Report (1)

- NSF Invitational Workshop on the Future of Database Systems Research
- Пало Альто, Калифорния, 1990
- SIGMOD Record, Vol. 19, No. 4, December 1990
- http://infolab.stanford.edu/~hector/lagi.ps
- Michael Brodie, Peter Buneman, Mike Carey, Ashok Chandra, Hector Garcia-Molina, Jim Gray, Ron Fagin, Dave Lomet, Dave Maier, Marie Ann Niemat, Avi Silberschatz, Michael Stonebraker, Irv Traiger, Jeff Ullman, Gio Wiederhold, Carlo Zaniolo, and Maria Zemankov

The "Lagunita" Report (2)

- Next-generation database applications will have little in common with today's business data processing databases
- They will
  - involve much more data,
  - require new capabilities including
    - type extensions,
    - multimedia support,
    - complex objects,
    - rule processing,
    - and archival storage,
  - and will necessitate rethinking the algorithms for almost all DBMS operations

The "Lagunita" Report (3)

- Объемы данных действительно выросли
- Расширение системы типов используется слабо
- Сложные объекты под сомнением
- Системы правил слабо используются даже в PostgreSQL
- Про архивную память разговоров нет
- Про переделку алгоритмов не слышно

The "Lagunita" Report (4)

- The cooperation between different organizations on common scientific, engineering, and commercial problems will require large-scale, heterogeneous, distributed databases
- Very difficult problems await in the areas of
  - inconsistent databases,
  - security,
  - and massive scale-up
- of distributed DBMS technology

The "Lagunita" Report (4)

- По-моему, неоднородные распределенные базы данных, по крайней мере, не находятся в mainstream
  - Сейчас и раньше
- Массивно-параллельные аналитические СУБД
- Распределенные системы NoSQL
- Масштабируемость на переднем плане
- Согласованность в рамках теоремы CAP
- Про безопасность разговоров в этом контексте не слышно

The "Lagunita" Report2 (1)

- NSF Workshop on the Future of Database Systems Research
- Пало Альто, Калифорния, 1995
- SIGMOD Record v. 25, No 1, 1996, 52-63
- http://i.stanford.edu/pub/cstr/reports/cs/tr/96/1563/CS-TR-96-1563.pdf
- Филипп Бернштейн (Phil Bernstein), Рон Брахман (Ron Brachman), Майкл Кери (Mike Carey), Рик Каттел (Rick Cattel), Гектор Гарсиа-Молина (Hector Garcia-Molina), Лаура Хаас (Laura Haas), Дейв Майер (Dave Maier), Джефф Науфтон (Jeff Naughton), Майкл Шварц (Michael Schwartz), Пат Селинджер (Pat Selinger), Ави Зильбершац (Avi Silberschatz), Майк Стоунбрейкер (Mike Stonebraker), Джефф Ульман (Jeff Ullman), Патрик Вальдурец (Patrick Valduriez), Мойше Варди (Moshe Vardi), Дженифер Вайдом (Jennifer Widom), Гио Вайдерхолд (Gio Wiederhold), Марианна Винслетт (Marianne Winslett), Мария Земанкова (Maria Zemankova)

The "Lagunita" Report2 (2)

- Поддержка мультимедийных объектов
  - Третичная память
    - **Не потребовалась**
  - Новые типы данных
    - **Не слишком много потребовалось**
  - Запросы с нечеткими критериями
    - **Исследования продолжаются, на практике не видно**
  - Поддержка пользовательских интерфейсов
    - **SQL расширен не был, других интерфейсов нет**

The "Lagunita" Report2 (3)

- Распределение информации
  - **В целом проникновение технологии баз данных в Web было переоценено**
  - Учет и расчеты
    - **Близко в идеям расчета стоимости сервисов в облаках**
  - Безопасность и конфиденциальность
    - **Снова как в облаках**
  - Репликация и согласование данных
    - **Почти в чистом виде теорема CAP**
  - Выборка и обнаружение данных
    - **Для работы с плохо структурированными данными используются «поисковики» и NoSQL**

The "Lagunita" Report2 (4)

- Новые применения баз данных
  - Интеллектуальный анализ данных
    - **Смесь OLAP и data mining, понималось еще плохо**
  - Хранилища данных
    - **Практически все сбылось (закачка данных, очистка и т.д.)**
  - Репозитарии
    - **Приложения, относящиеся к категории *репозитариев*, характеризуются тем, что они предназначаются для хранения и управления как данными, так и *метаданными*, т. е. информацией о структуре данных**
    - **Ничего подобного в практику не вошло**

Strategic Directions in Computing Research (1)

- Симпозиум "Strategic Directions in Computing Research”
- Лаборатория информатики Массачусетского технологического института (США) при поддержке ACM, National Science Foundation
- Кембридж, шт. Массачусетс, 1996
- ACM Computing Surveys, v.28, no.4, December 1996
- https://cs.arizona.edu/~rts/pubs/CompSurvDec96.pdf
- Х.Блейкли (Jose Blakeley), П.Бунеман (Peter Buneman), У.Дайал (Umesh Dayal), Т.Имилинский (Tomasz Imielinski), С.Джаджодиа (Sushil Jajodia), Х.Корт (Hank Korth), Г.Лохман (Guy Lohman), Д.Ломе (Dave Lomet), Д.Майер (Dave Maier), Ф.Манола (Frank Manola), Т.Озу (Tamer Ozsu), Р.Рамакришнан (Raghu Ramakrishnan), К.Рамамритан (Krithi Ramamritham), Х.Шек (Hans Scheck), А.Зильбершац (Avi Silberschatz), Р.Снодграсс (Rick Snodgrass), Д.Ульман (Jeff Ullman), Д.Вайдом (Jennifer Widom) и С.Здоник (Stan Zdonik)

Strategic Directions in Computing Research (2)

- Расширяемость и компонентизация
- Необходимо создать системы, которые дают возможность разработчику легко вводить новые типы данных, разработанные вне данной СУБД, которыми можно манипулировать внутри базы данных наравне с ее собственными полноправными типами
- Необходимо найти способы сделать архитектуру СУБД открытой таким образом, чтобы могли подключаться новые функциональные компоненты, и чтобы функциональные возможности системы базы данных могли конфигурироваться более гибкими способами в соответствии с потребностями приложений
- **На мой взгляд, в этом направлении продвижений не было**

Strategic Directions in Computing Research (3)

- Оптимизация запросов
- Могут измениться критерии оптимизации
- В прошлом оптимизаторы пытались сократить полное время отклика путем сокращения общего расхода ресурсов (в котором доминирует, вероятно, число доступов к диску), требуемых для обработки запроса
- Пользователи могут пожелать минимизировать свои общие расходы на информацию, используя такие источники, которые являются более дешевыми, но могут иметь гораздо большее время отклик
- **Оптимизаторы слишком сложны, чтобы реально позволить себе подобное разнообразие критериев**

Strategic Directions in Computing Research (4)

- Интеллектуальный анализ данных в базах данных
- Такие поисковые задачи, как генерация правил (ассоциаций правил), классификация и группирование, могут рассматриваться как случайные запросы, для которых необходимы новые семейства языков запросов
- К числу исследовательских задач в этой области относится разработка адекватного набора простых примитивов запросов и нового поколения методов оптимизации запросов
- **На мой взгляд, data mining с базами данных так и не интегрирован**
- **Может быть, это не нужно?**

The Asilomar Report (1)

- The Asilomar Report on Database Research
- Асиломар, неподалеку от г. Монтерей, Калифорния, 1998
- SIGMOD Record v. 27, No 4, 1998, 74-80
- http://www.sigmod.org/publications/sigmod-record/9812/asilomar.html
- Phil Bernstein, Michael Brodie, Stefano Ceri, David DeWitt, Mike Franklin, Hector Garcia-Molina, Jim Gray, Jerry Held, Joe Hellerstein, H. V. Jagadish, Michael Lesk, Dave Maier, Jeff Naughton, Hamid Pirahesh, Mike Stonebraker, and Jeff Ullman

The Asilomar Report (2)

- Системы управления базами данных в стиле "plug and play«
- Обеспечение самонастраиваемости систем баз данных, т.е. удаление мириадов параметров настройки производительности, которые должны определять пользователи в текущих продуктах
- Методы автоматического выбора индексов
- **Возились с этим много, достижения имеются, но идеала нет**
- Должно быть возможно подключить систему баз данных к сети компании или Internet и автоматически обеспечить раскрытие информации для всех других систем баз данных, доступных в сети, и взаимодействие с этими системами
- **Это осталось фантазией**

The Asilomar Report (3)

- Объединение миллионов систем баз данных
- Нужны оптимизаторы запросов, которые могут эффективно работать с федеративными системами баз данных, состоящими из 1000 и большего числа узлов
- **Нет таких оптимизаторов, трудно и не оправданно**
- Система баз данных должна быть в состоянии быстро выдать грубый ответ, а затем постепенно совершенствовать его, останавливаясь, когда пользователь решит, что ответ "достаточно хорош«
- **Наверное, сделать это можно, но опять-таки сложно и не востребовано**

The Asilomar Report (4)

- Переосмысление традиционной архитектуры систем баз данных
- Настало время переосмыслить базовые архитектурные предположения в свете появления среды, которая будет доступна в 2010-ом году
- **Среда принципиально не изменилась, по-видимому, как раз сейчас пора переосмыслять архитектуру**
  - **SSD**
  - **Энергонезависимая память**
  - **Массивная мультитредность**

The Lowell Report (1)

- The Lowell Database Research Self Assessment
- Лоуэлл, шт. Массачусетс, 2003
- Communications of the ACM, Volume 48 Issue 5, May 2005, Pages 111-118
- http://research.microsoft.com/en-us/um/people/gray/Lowell/LowellDatabaseResearchSelfAssessment.htm
- Serge Abiteboul, Rakesh Agrawal, Phil Bernstein, Mike Carey, Stefano Ceri, Bruce Croft, David DeWitt, Mike Franklin, Hector Garcia Molina, Dieter Gawlick, Jim Gray, Laura Haas, Alon Halevy, Joe Hellerstein, Yannis Ioannidis, Martin Kersten, Michael Pazzani, Mike Lesk, David Maier, Jeff Naughton, Hans Schek, Timos Sellis, Avi Silberschatz, Mike Stonebraker, Rick Snodgrass, Jeff Ullman, Gerhard Weikum, Jennifer Widom, and Stan Zdonik

The Lowell Report (2)

- Интеграция текста, данных, кода и потоков
- Пора прекратить встраивать новые конструкции в старую реляционную архитектуру
- Нужно переосмыслить базовую архитектуру СУБД с целью поддержки структурированных данных; текстовых, пространственных, темпоральных и мультимедийных данных; процедурных данных, т.е. типов данных и инкапсулирующих их методов; триггеров; потоков и очередей данных как равноправных компонентов первого сорта внутри архитектуры СУБД (как на уровне интерфейсов, так и на уровне реализации)
- Для исследовательского сообщества требуется выработка новой системы понятий
- Участники ожидают появления нескольких прототипов новой архитектуры СУБД в течение пяти лет

The Lowell Report (3)

- Такой системы понятий не видно
- Вместо всего этого в ход пошло “one size doesn’t fit all”
- Специализированные СУБД
- Архитектура коренным образом не менялась

The Lowell Report (4)

- Интеграция информации
- В Internet парадигма ETL не приемлема
- Теперь требуется производить интеграцию информации между несколькими предприятиями
- В результате потребуется интеграция, возможно, миллионов информационных источников «на лету»
- **Мне кажется, что интеграция «на лету» для анализа данных не слишком применяется**
- **По-прежнему используется ETL или анализ проводится над «сырыми» данными**

The Lowell Report (5)

- Data Mining
- Проблемой data mining в области баз данных является разработка алгоритмов и структур данных для просеивания базы данных в поисках "жемчужин“
- Такая обработка должна вестись в фоновом режиме с потреблением остаточных системных ресурсов
- Другой важной проблемой является интеграция data mining с подсистемой поддержки запросов, оптимизацией и другими средствами базы данных, такими как триггеры
- **Не очень заметно средств встраивания data mining внутрь СУБД**
- **Скорее расширяется набор средств OLAP**

The Claremont Report (1)

- The Claremont Report on Database Research
- Berkeley, California, Claremont Resort, 2008
- Communications of the ACM, Vol. 52, No. 6, 2009, Pages 56-65
- http://db.cs.berkeley.edu/claremont/claremontreport08.pdf
- Rakesh Agrawal, Anastasia Ailamaki, Philip A. Bernstein, Eric A. Brewer, Michael J. Carey, Surajit Chaudhuri, AnHai Doan, Daniela Florescu, Michael J. Franklin, Hector Garcia Molina, Johannes Gehrke, Le Gruenwald, Laura M. Haas, Alon Y. Halevy, Joseph M. Hellerstein, Yannis E. Ioannidis, Hank F. Korth, Donald Kossmann, Samuel Madden, Roger Magoulas, Beng Chin Ooi, Tim O'Reilly, Raghu Ramakrishnan, Sunita Sarawagi, Michael Stonebraker, Alexander S. Szalay, Gerhard Weikum

The Claremont Report (2)

- Факторы
- **Повышение ажиотажа вокруг больших объемов данных**
- Это приводит к быстрому росту числа пользователей традиционных систем управления базами данных (СУБД), а также стимулирует разработку новых специализированных решений управления данными на основе упрощенных компонентов
- Повсеместное использование больших объемов данных приводит и к возрастанию числа разработчиков технологий управления данными, что, несомненно, вызовет коренную реорганизацию этой области
- One size …
- NoSQL???

The Claremont Report (3)

- **Анализ данных как центр прибыли**
- Опытные аналитики привлекаются к работе во все большем числе отраслей индустрии, и все чаще их интересуют возможности анализа необработанных данных
- В то же время к аналитической работе с данными склоняется и возрастающее число лиц, принимающих решения, которые не обладают технической подготовкой
- **Первое наблюдение истинно, второе - сомнительно**

The Claremont Report (4)

- **Повсеместность структурированных и неструктурированных данных**
- Наблюдается взрывообразный рост объема структурированных данных, доступных в Web и корпоративных внутренних сетях
- Эти данные происходят из разнообразных источников, далеко не всегда из традиционных баз данных: большие объемы данных производятся при извлечении структурированной информации из текстов, источниками данных служат программные журналы и датчики, структурированные данные извлекаются при обходе сайтов Deep Web
- **Мне кажется, что сильного продвижения здесь не видно**
- **Возможно, нет реальной потребности**

The Claremont Report (5)

- **Расширяющиеся требования разработчиков**
- Некоторые разработчики не хотят спускаться на уровень SQL и считают СУБД слишком тяжеловесными для изучения и использования по сравнению с другими компонентами с открытыми кодами
- Поскольку экосистема управления базами данных развивается для поддержки далеко не только типичных пользователей СУБД, возникают благоприятные возможности разработки новых моделей программирования и системных компонентов для управления данными и манипулирования ими
- **По моим наблюдениям, новые модели (например, Model-view-controller) достаточно ограниченны**
- **Серьезных продвижений не видно**

The Claremont Report (6)

- **Архитектурные изменения в области применения компьютеров**
- На макроуровне фундаментальным изменения в архитектуре программного обеспечения сулит развитие «облачных» (cloud) компьютерных служб
- На микроуровне в компьютерных архитектурах закон Мура теперь трактуется в пользу не повышения тактовой частоты микропроцессоров, а увеличения числа процессорных ядер и потоков управления в одном кристалле
- Основные изменения в технологии хранения данных относятся к иерархии памяти в связи с доступностью бо́льшего числа кэшей увеличенного объема на одном кристалле, все более дешевой основной памяти большого объема и флэш-памяти
- **SSD остаются недоиспользованными**
- **Что будет с энергонезависимой RAM?**

The Beckman Report (1)

- The Beckman Report on Database Research
- Irvine, California, Beckman Center of the University, 2013
- ACM SIGMOD Record, Volume 43 Issue 3, September 2014, Pages 61-70
- http://beckman.cs.wisc.edu/beckman-report2013.pdf
- Daniel Abadi, Rakesh Agrawal, Anastasia Ailamaki, Magdalena Balazinska, Philip A. Bernstein, Michael J. Carey, Surajit Chaudhuri, Jeffrey Dean, AnHai Doan, Michael J. Franklin, Johannes Gehrke, Laura M. Haas, Alon Y. Halevy, Joseph M. Hellerstein, Yannis E. Ioannidis, H.V. Jagadish, Donald Kossmann, Samuel Madden, Sharad Mehrotra, Tova Milo, Jeffrey F. Naughton, Raghu Ramakrishnan, Volker Markl, Christopher Olston, Beng Chin Ooi, Christopher Re, Dan Suciu, Michael Stonebraker, Todd Walter, Jennifer Widom

The Beckman Report (2)

- Пять проблем Больших Данных:
  - масштабируемая инфраструктура больших/быстро поступающих данных;
  - преодоление разнородностей ландшафта управления данными;
  - сквозные управление и понимание данных;
  - облачные службы
  - и управление различными ролями людей в жизненном цикле данных

The Beckman Report (3)

- **Масштабируемые инфраструктуры больших/быстро поступающих данных**
- Для обработки данных сверхбольшого объема с допустимым временем отклика потребуется очень высокий уровень параллелизма
- В эффективных стратегиях обработки запросов потребуется полностью использовать возможности крупных кластеров или многоядерных процессоров, обеспечивая и вертикальное, и горизонтальное масштабирование для удовлетворения потребностей приложений
- Для адаптации к характеристикам новых данных и сокращения расходов на перемещение данных на разных стадиях их анализа потребуется интеграция подсистем обработки запросов и подсистем взятия образцов данных, интеллектуального анализа данных и вычислений на основе машинного обучения

The Beckman Report (4)

- Следует продолжать изучать способы использования специализированных процессоров, например, графических процессоров, программируемых логических интегральных схем и специализированных интегральных схем (application-specific integrated circuit, ASIC) для обработки очень крупных наборов данных
- Эти изменения технологий коммуникаций и обработки данных потребуют пересмотра параллельных и распределенных алгоритмов обработки запросов, которые традиционно ориентированы на более однородные аппаратные среды

The Beckman Report (5)

- Нужно учиться применять появляющиеся технологии основной памяти и хранения данных
- По сравнению с массово используемыми магнитными дисками твердотельные диски дороже в расчете на сохраняемый гигабайт данных, но дешевле в расчете на операцию ввода/вывода
- Разрабатываются различные технологии энергонезависимой основной памяти со случайным доступом (non-volatile random-access memory, NV-RAM)
- Все они обладают разными характеристиками скорости, энергопотребления и износостойкости
- Нужно принимать во внимание как архитектуры с подключением устройств хранения данных к серверам, так и архитектуры с сетевым подключением подобных устройств

The Beckman Report (6)

- Для обработки данных, поступающих со все возрастающими скоростями, потребуются новые масштабируемые технологии приема и обработки потоков данных
- При очень высокой скорости некоторых источников данных, часто при небольшой плотности данных потребуется обрабатывать некоторые данные в реальном времени без сохранения их в полном объеме
- Для таких данных скорее потребуется сохранение образцов или агрегатов, чтобы можно было выдавать ответы на некоторые категории запросов, поступающих, когда данные перестают быть доступными
- Для подобных данных все более важной будет непрерывная обработка запросов (progressive query processing) для обеспечения инкрементных и частных результатов с точностью, возрастающей по мере прохождения данных через конвейер

The Beckman Report (7)

- Для данных, которые сохраняются, но обрабатываются не более одного раза, мало смысла использовать средства хранения и индексации системы баз данных
- Для таких данных больший смысл может иметь использование «схемы при чтении» (schema-on-read), чем традиционной «схемы при записи» (schema-on-write), которая приводит к излишним накладным расходам во время приема данных
- В это время может быть желательно просто сбросить в систему хранения набор бит, возвращаясь к ним, когда и если кому-либо понадобиться их проинтерпретировать
- К тому же, целесообразность способа интерпретации данных, получаемых в ответ на заданный запрос, может зависеть от этого запроса и поэтому может быть неизвестна во время записи данных

The Beckman Report (8)

- Кроме существенно расширившихся требований к анализу данных, сегодняшний мир выставляет новые требования к сбору, обновлению данных и поддержке быстрого (но простого) доступа к ним
- Стремление к обеспечению высоких скоростей сбора и обновления данных, не имеющих схемы, привело к разработке систем категории NoSQL
- На сегодняшнем ландшафте платформ Больших Данных имеется ряд таких систем, почти у каждой из которых имеется своя модель транзакций и согласованности данных
- В большинстве систем обеспечиваются только базовые возможности доступа к данным, а также слабые гарантии атомарности и изолированности, что затрудняет их использование для разработки надежных приложений
- В результате развивается новый класс систем Больших Данных, поддерживающих полноценные возможности в стиле систем баз данных и опирающихся на хранилища «ключ-значение» или аналогичные подложки

The Beckman Report (9)

- Показателями масштабируемости должны служить не только объемы данных и скорость выполнения запросов, но и
  - совокупная стоимость владения (включая управление, использование энергии и т.д.),
  - скорость сквозной обработки (т.е. время от поступления сырых данных до потенциального получения аналитической информации),
  - уязвимость (например, возможность продолжать обработку данных в случае возникновения ошибок при разборе данных)
  - и простота использования (особенно для начинающих пользователей)
- Для оценки результатов при наличии этого расширенного набора показателей потребуются новые типы бенчмарков

The Beckman Report (10)

- **Разнородность ландшафта управления данными**
- Сегодняшний управляемый данными мир сталкивается с существенно *б*ольшим разнообразием типов, форм и размеров
  - В корпоративном мире данные традиционно сохраняются и анализируются в хранилище данных, которое тщательно проектируется и оптимизируется в расчете на выполнения повторяющихся и эпизодических аналитических задач
  - В сегодняшнем более открытом мире данные часто сохраняются в разных представлениях, управляемых разными программными системами с разными API, подсистемами обработки запросов и средствами анализа
- С таким разнообразием сумеет справится какая-то одна безразмерная система

The Beckman Report (11)

- Скорее будут развиваться несколько классов систем, ориентированных на удовлетворение соответствующих классов потребностей
  - исключение избыточных данных (deduplication)
  - анализ крупных графов
  - различные научные эксперименты
  - обработка потоков данных в реальном времени
- или на использование некоторого конкретного типа аппаратных платформ
  - кластеров недорогих машин
  - крупных многоядерных серверов
- Исследователи баз данных должны применять опыт сообщества для разработки средств параллельной обработки данных, ориентированной на работу с множествами, и эффективного управления наборами данных, не помещающимися в основной памяти

The Beckman Report (12)

- Потребность в сосуществовании нескольких систем Больших Данных и аналитических платформ бесспорна
- Поэтому еще одной проблемой разнообразия является потребность аналитиков в средствах объединения и анализа данных, сохраняемых в разных системах
- Для поддержки запросов к Большим Данных, пересекающих границы отдельных систем, платформы потребуется интегрировать и объединять в федерации
- Потребуется не только скрывать неоднородность форматов данных и языков доступа, но также и оптимизировать производительность обращений к данным, распространяющихся на пределы отдельных систем, и потоков, по которым данные перемещаются между системами

The Beckman Report (13)

- Требуется управлять разнообразными абстракциями программирования над крупными наборами данных
- Вместо того чтобы надеяться на появление единого языка анализа данных для Больших данных
  - возможно, путем расширения SQL или какого-то другого популярного языка,
- следует давать пользователям возможность анализировать данные в наиболее естественной для них среде
- Эта среда может опираться, например, на SQL, Pig (https://pig.apache.org/), R, Python,
- какой-либо предметно-ориентированный язык или на какую-либо низкоуровневую модель программирования с ограничениями (такую как MapReduce)
- или модель массивной синхронной обработки Лесли Валианта (Leslie Gabriel Valiant, https://en.wikipedia.org/wiki/Bulk_synchronous_parallel)

The Beckman Report (14)

- Для этого требуется разработка паттернов промежуточного уровня,
  - масштабируемого перемножения матриц,
  - операций обработки списков или парадигм итерационного выполнения,
  - с поддержкой различных языковых связываний или включений
- Потенциально полезными могут оказаться инструменты для быстрой разработки новых предметно-ориентированных языков анализа данных
  - средств, которые опрощают реализацию новых масштабируемых языков с параллелизмом по данным

The Beckman Report (15)

- Требуются модульные платформы, которые могут справится как с «сырыми», так и с «подготовленными» данными, системы, пригодные к обработки подготовленных данных во многих формах
  - таблиц,
  - матриц
  - или графов
- В таких системах будут обрабатываться цельные потоки данных или действий, сочетающие разные типы обработки данных, например, использующие SQL для запросов данных и R для их анализа
- Для обеспечения однородности систем, производящих доступ к данным такими разными способами, иногда могут оказаться полезными «ленивые» вычисления, включая
  - откладываемые разбор/преобразование/загрузку данных,
  - откладываемое построение индексов и представлений
  - и планирование запросов непосредственно перед их выполнением

The Beckman Report (16)

- Системы управления Большими данными должны становится более интероперабильными и поддающимися сборке подобно блокам Lego
- На системном уровне к такому подходу близки фреймворки
  - Mesos (http://mesos.apache.org/)
  - и YARN (https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html),
- а также системы управления потоками работ из экосистемы Hadoop и инструменты для управления потоками научных работ

The Beckman Report (17)

- **Сквозные обработка и понимание данных**
- Исследовательское сообщество баз данных должно сфокусироваться на средствах сквозных обработки и понимания данных
- Поразительно немногие инструментальные средства пригодны для сквозной обработки данных, позволяющей пройти путь от «сырых» данных до извлеченного из них знания без значительного вмешательства людей
  - Более того, эти люди должны обладать значительной компьютерной подготовкой
- Совсем мало средств относится к open source
- В основном имеются проприетарные продукты, поддерживающие отдельные шаги обработки

The Beckman Report (18)

- В результате существующим средствам мало чем помогают усилия исследователей в области интеграции данных
- Нужно сосредотачиваться не только на совершенствовании средств интеграции
  - очистка данных
  - согласование схем
  - устранение дублирующей информации
- но и на методах сбора этих разрозненных компонентов в единое сквозное решение

The Beckman Report (19)

- На что должны быть похожи такие «сквозные» инструменты?
- В своей основе это все тот же конвейер от «сырых» данных к знаниям
- Основные шаги конвейера будут состоять из
  - сбора данных (data acquisition)
  - отбора данных (selection), оценки данных (assessment), очистки данных (cleaning), преобразования данных (transformation) – *выпас данных* (data wrangling)
  - извлечения и интеграции данных
  - разного рода аналитической обработки (OLAP, mining)
  - обобщения результатов (summarization), обеспечения информации о происхождении (provenance) и разъяснения (explanation)

The Beckman Report (20)

- Существенными отличиями являются намного большее разнообразие данных и пользователей и значительно больший масштаб
- Возникают комбинации структурированных и неструктурированных данных, которые пользователи хотят совместно использовать в структурированном стиле
- Люди из разных прикладных областей создают инструменты для работы с данными, опирающиеся на обратную связь с человеком на каждом шаге конвейера
- Эти инструменты все чаще используются экспертами из прикладных областей, а не специалистами ИТ
- Получаемые аналитические результаты используются намного более разными людьми, чем раньше
- Инструменты используются во всех возможных масштабах

The Beckman Report (21)

- Нужно разрабатывать многие инструменты, каждый из которых может служить компонентом конвейера, которые можно бесшовно интегрировать и легко использовать на дилетантском и экспертном уровнях
- При возможности следует стремиться
  - использовать строительные блоки анализа данных категории open source,
  - комбинировать и повторно использовать их,
  - обеспечивать руководства по целесообразному их применению
- Инструменты должны обеспечивать возможность работы с данными разного объема
- Каждый шаг конвейера должен быть интерактивным и обеспечивать обратную связь с отдельными личностями, группами и даже краудсорсингом

The Beckman Report (22)

- Инструменты должны уметь использовать прикладные знания: словари, базы знаний и правила
- При наличии потребности анализировать данные большого объема разработчики инструментов должны прибегать с использованию средств машинного обучения для частичной автоматизации процесса настройки
- Однако правила, созданные вручную, останутся важными, поскольку во многих аналитических приложениях требуется очень высокая точность, например, в электронной коммерции
- В таких приложениях аналитики часто пишут большое число правил для обработки проблемных ситуаций, не поддающихся автоматизации
- Истинно сквозные и просто используемые инструменты должны обеспечивать поддержку написания, оценки, применения таких правил, а также управления ими

The Beckman Report (23)

- Требования к разъяснению, отслеживания происхождения, обобщению и визуализации проявляются на всех шагах конвейера
- Эти возможности критичны для обеспечения простоты использования инструментов
- Для обеспечения возможностей разъяснения, отслеживания происхождения и повторного использования ключевой является поддержка соответствующей метаинформации
- Визуализация обеспечивает важный способ взаимодействия с пользователями и особенно полезна в совокупности с методами автоматического анализа
- Визуальной аналитике уделяется взрастающее внимание в сообществах баз данных, человеко-машинных интерфейсов и визуализации
- Визуализируются запросы к базам данных, интеллектуальный анализ данных и выпас данных
- Требуется развитие методов визуализации для работы с данными большого объема

The Beckman Report (24)

- Аналитическое управление данными насыщено знаниями
- Чем больше имеется знаний о целевой области, тем лучше инструменты могут поддерживать требуемую аналитику
- В результате имеется растущая тенденция к созданию, разделению и использованию предметно-ориентированных знаний для лучшего понимания данных
- Знания часто сохраняются в базах знаний, описывающих наиболее важные сущности и связи предметной области
- Такие базы знаний используются для повышения точности аналитического конвейера, обеспечивая ответы на предметно-ориентированные запросы и поддерживая поиск прикладных экспертов
- Во многих компаниях базы знаний используются также для ответов на запросы пользователей, аннотирования текстов, поддержки электронной коммерции и анализа социальных сетей

The Beckman Report (25)

- Вероятно, возникновение «центров знаний», создаваемых, поддерживаемых и используемых онлайновыми сообществами, компаниями и т.д.
- Такие центры будут содержать базы знаний и инструменты, поддерживающие запросы, совместное использование знаний и их применение для анализа данных
- Многие инструменты будут доступны в «облаках», позволяя использовать центры данных пользователям и приложениям одной или нескольких предметных областей
- В этом направлении имеются некоторые продвижения и даже достигнуты успехи
  - Например, проект YAGO (Max Planck Institute for Computer Science, http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/)

The Beckman Report (26)

- **Облачные службы**
- С позиций платформы данных идеальной целью является обеспечение PaaS (Platform as a Service) в наиболее чистой форме
- В мире, где поддерживается PaaS для данных, у пользователей должна иметься возможность
  - загружать данные в облака,
  - запрашивать их точно так же, как запрашиваются SQL-ориентированные базы данных в интранете
  - и выборочно совместно использовать как данные, так и результаты
- Не должны возникать беспокойства относительно
  - числа абонируемых экземпляров СУБД,
  - используемой операционной системы,
  - разделения баз данных между серверами
  - и настройки систем
- Несмотря на появление служб типа Database.com, Google Big Query, Amazon Redshift и Microsoft Azure SQL Database мы все еще далеки от этой перспективы
- Ниже обсуждаются наиболее важные проблемы с позиций сообщества баз данных

The Beckman Report (27)

- Первой проблемой является эластичность
- Во многих случаях эластичностью обладают облачные вычисления, но не данные
- Если мы хотим построить Платформу Данных как Службу, то
  - Какова должна быть архитектура с учетом развития технологий хранения данных и сетей?
  - Следует ли привязывать систему хранения данных в серверам или делать ее сетевой?
  - Может ли одна облачная служба поддерживать и транзакции и аналитику?
  - Как быть с кэшированием?
- Для поддержки эластичности также требуется возможность использования дополнительных ресурсов или управления существующими ресурсами
- Серверы баз данных и аналитические платформы для Платформы Данных как Службы будут функционировать с использованием эластичных ресурсов, которые будут быстро выделяться в пиковые моменты и, возможно, перераспределяться для обслуживания пользователей, которые заказывают высший уровень обслуживания

The Beckman Report (28)

- Вторая проблема – репликация данных
- Хотя тема репликации является хорошо изученной, важно вернуться к ней в контексте облаков, имея в виду потребность
  - в высоком уровне доступности,
  - балансировки нагрузки
  - и экономичности
- Нужно иметь в виду возможность наличия географически распределенных центров данных

The Beckman Report (29)

- Третьей проблемой является системное администрирование и настройка
- Платформа данных, используемая как облачная служба, должна быть исключительно самонастраиваемой
- В мире Платформы Данных как Службы традиционные роли баз данных и системных администраторов просто не существуют
- Все задачи администратора, включая
  - планирование мощностей (capacity planning),
  - обеспечение ресурсов (resource provisioning),
  - физическое управление данными
  - и формирование политики доступности данных (admission control policy setting)
- должны быть автоматизированы
- С учетом возможных разнообразий, связанных с эластичностью ресурсов и их доступностью в облачной среде

The Beckman Report (30)

- Ключевой технической проблемой управления эластичностью служб, связанных с данными, является многоарендность (multitenancy)
- Для обеспечения конкурентноспособности провайдер Платформы Данных как Службы должен предлагать ценовую структуру, сопоставимую с той, которая обеспечивается для локальных решений
- Провайдер должен допускать совместную аренду одной службы баз данных с разделением физических ресурсов одного сервера, чтобы обеспечивать разумную ценовую политику
- Однако при этом возникают две трудности
- Во-первых, один арендатор не должны влиять на производительность службы, видимую другими арендаторами
  - Для этого требуется тщательное управление процессорами, вводом-выводом, памятью и сетевыми ресурсами
- Во-вторых, служба баз данных должна гарантировать безопасность

The Beckman Report (31)

- Соглашения об уровне обслуживания (SLA) в мире облачных служб вызывают проблемы
- В многоарендной Платформе Данных как Службе эластичность доступности глобальных ресурсов и потребность в управлении ресурсами влияет на доступность ресурсов для арендатора
- Это, в свою очередь, может влиять как качество обслуживания
- Мы только начинаем понимать связь между многоаредным распределением ресурсов и качеством обслуживания
  - Сегодняшние SLA ориентированы, прежде всего, на доступность

The Beckman Report (32)

- Еще одной проблемой является совместное использование данных
- Сообщество баз данных должно стремиться к созданию служб, использующих этот потенциал облаков
- Нужно развивать идеи в контексте аналитики
- Например,
  - как находить полезные публично доступные данные,
  - как сопоставлять собственные данные с публично доступными данными для формирования контекста,
  - как находить качественные данные в облаках
  - и как распределять затраты при разделении вычислений и данных?

The Beckman Report (33)

- **Роли людей в жизненном цикле данных**
- В мире корпоративного управления данными было ясно, кто и что делает:
  - разработчики создавали базы данных и их приложения,
  - бизнес-аналитики запрашивали данные с использованием SQL,
  - конечные пользователи генерировали данные, запрашивали и обновляли базы данных
  - и администраторы настраивали и отслеживали базы данных и рабочие нагрузки
- Сегодня мир меняется
- Один человек может играть несколько ролей в жизненном цикле данных, и многие приложения Больших Данных привлекают людей в разных ролях

The Beckman Report (34)

- Сегодня требуется рассматривать роли людей в связи с
  - пониманием и совершенствованием запросов данных,
  - выявлением заслуживающих внимания источников информации,
  - определением и совершенствованием конвейера обработки данных
  - и визуализацией получаемых данных
- Все это сочетается с микро-задачами, решаемыми экспертами предметных областей и конечными пользователями
- Предлагается классифицировать роли людей на четыре категории:
  - производители данных (producers of data),
  - кураторы данных (curators of data),
  - потребители данных (consumers of data)
  - и члены сообществ (community members)
- Далее обсуждаются эти категории и соответствующие исследовательские проблемы

The Beckman Report (35)

- Сегодня требуется рассматривать роли людей в связи с
  - пониманием и совершенствованием запросов данных,
  - выявлением заслуживающих внимания источников информации,
  - определением и совершенствованием конвейера обработки данных
  - и визуализацией получаемых данных
- Все это сочетается с микро-задачами, решаемыми экспертами предметных областей и конечными пользователями
- Предлагается классифицировать роли людей на четыре категории:
  - производители данных (producers of data),
  - кураторы данных (curators of data),
  - потребители данных (consumers of data)
  - и члены сообществ (community members)
- Далее обсуждаются эти категории и соответствующие исследовательские проблемы

The Beckman Report (36)

- Почти любой человек сегодня является производителем данных (мобильные телефоны, социальные сети и т.д.)
- Задачей сообщества баз данных является разработка алгоритмов и стимулов, поддерживающих создание и совместное использование людьми наиболее полезных данных при сохранении желательного уровня конфиденциальности
- Например, как помочь людям быстро добавить к данным метаданные?

The Beckman Report (37)

- Все больше людей становится кураторами данных
- В современном мире меньше централизованного контроля над данными
- Одним из теперешних методов кураторства данных является краудсоросинг
- Проблемой является получение высококачественных наборов данных на основе процесса с несовершенным кураторством
- Кроме того, нужны платформы для поддержки кураторства данных и соответственным образом расширенные приложения

The Beckman Report (38)

- Люди потребляют данные
- Все чаще люди хотя использовать все более запутанные данные сложными способами
- Эти приводит к многим проблемам
- В корпоративной среде потребителями данных обычно являются люди, умеющие формулировать SQL-запросы над структурированными базами данных
- Сегодняшние потребители данных вообще не умеют формулировать запросы
- Проблемой является обеспечение возможности для таких людей самим получать требуемые им ответы
- Нужны новые интерфейсы, комбинирующие средства визуализации, запросов и навигации
- Многие люди могут не знать даже о наличии или отсутствии требуемых им данных
- Если не ясно, что нужно запросить, нужны другие средства для обхода, изучения, визуализации и анализа данных

The Beckman Report (39)

- Люди являются членами сообществ
- Имеются многочисленные онлайновые сообщества, каждый день возникают новые
- Члены этих сообществ часто хотя создавать, совместно использовать данные и управлять ими
- В частности, члены сообществ могут захотеть совместно создавать базы знаний, вики и средства обработки данных
- Проблемой является создание инструментов, помогающих сообществам производить полезные данные, исследовать их, совместно использовать и анализировать

Спасибо за внимание!
С новой вас планетой!

- Замену Плутону обнаружили астрономы — сотрудники Калифорнийского технологического института Майкл Браун и Константин Батыгин