Распределенная СУБД для анализа Больших Данных в реальном времени. HP Vertica

Форум "Big Data 2012"

Кирилл Вахрамеев

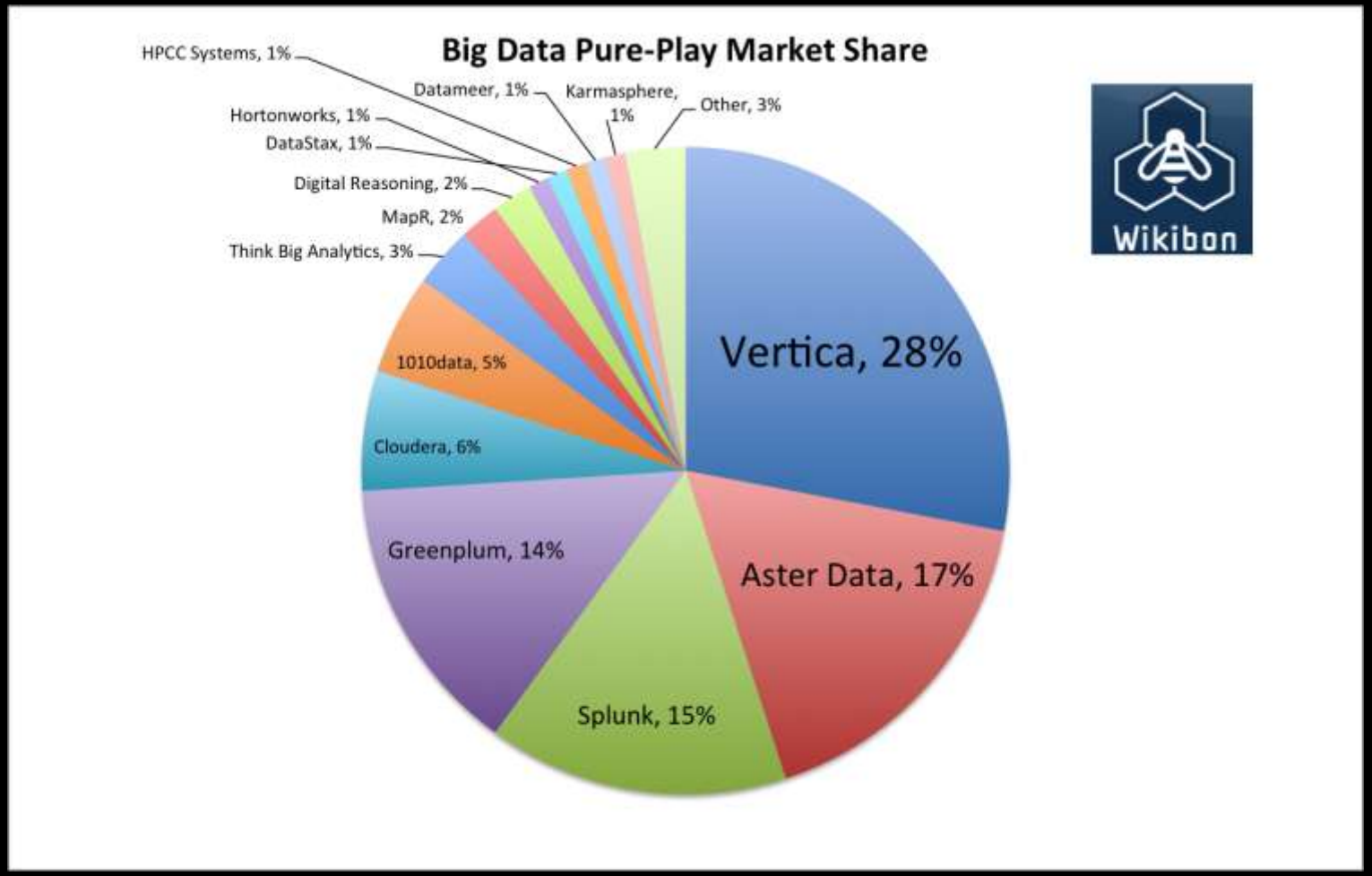
специалист по продвижению серверных решений для критически важных приложений,

OpenVMS Ambassador,

HP Россия

22 марта 2012

Vertica #1 на рынке Big Data (февраль 2012)



http://www.vertica.com/wp-content/uploads/2010/12/BigDataMarket022812rv.pdf

Вы готовы к анализу данных?

Каждый

нуждается в информации,
а не только аналитики

Объем информации растет;
IDC предсказывает рост

в

44 раза

в следующем десятилетии

100%

компаний из списка

Fortune 2000

анализируют данные

Нужно анализировать и адаптировать

РАЗНЫЕ

данные и связи между ними

Аналитические платформы активно внедряются, происходит

смещение трат

на специализированные системы

ROI

«Return on Information» - сколько денег можно получить за информацию? Новая метрика.

*IDC report 2009

Зачем нужна бизнес-аналитика в Реальном Времени?

- Снижение рисков в быстроменяющемся мире
  - Изучение и прогнозирование поведения клиентов, поставщиков и регуляторов
  - Оптимизация взаимодействия с вышеперечисленными
- Уменьшение и оптимизация операционных расходов, контроль ключевых KPI
  - Фрод-мониторинг: отслеживание подозрительных сделок
- Оценка общественных и экономических тенденций
  - Упреждающая реакция на изменения настроений заказчиков и рынка

Повышение адекватности и качества принимаемых решений!

Большие Данные это сколько?

Размер и классификация хранилища данных, сегодня

- <500ГБ – Маленькое
- 500ГБ > 20ТБ – Типовое
- 20ТБ > 50ТБ – Большое
- >50ТБ – очень Большое
- Несколько лет назад хранилище размером больше нескольких ТБ было редкостью

Большие Данные это сколько и чего?

На самом деле, размер – только один из факторов.

Мир данных трехмерен:

- размер (не помещается)
- скорость (не прокачивается)
- многообразие (не формализуется)

--- Большие Данные – это такие данные, что вышеуказанные факторы не позволяют обрабатывать их «обычными» методами

Пример Больших Данных

Даже очень больших

- 40 млн. игроков
- регистрируется каждый клик
- 3ТБ данных в день
- 200 машин в кластере

- анализ в реальном времени и мгновенное предоставление информации в виде рекомендаций
- непрерывная работа 24x7x365 – никаких «окон» на загрузку данных

Майкл Стоунбрейкер:

«Мы полагаем, что для рынка СУБД начинается эпоха очень интересных перемен.»

- Специализированные системы превосходят по производительности универсальные в 10-100 раз
- Простота использования – отказ от многочисленных настроек и даже от самой возможности настройки
- Простота интеграции – каждая система специализирована, но интерфейсы стандартные
- Vertica – коммерческий проект для аналитики «с нуля» - на базе научных разработок. Прототип - СУБД C-Store

«один размер» больше не подходит для всех

HP Vertica Analytics System

Аналитика повсюду

СКОРОСТЬ **МАСШТАБ** **ПРОСТОТА**

- Аналитика “точно вовремя”
- в 50–1000 раз быстрее среднее время обработки запросов чем в традиционных построчных системах
- До 10x прирост скорости загрузки данных
- Простота установки/использования
- Высокая масштабируемость и полный параллелизм
- Индустриально стандартная платформа x86
- Гибридная in-memory/on-disk архитектура
- Хранение данных близко к процессору
- Большие масштабы, широкие возможности

Построена вокруг 4-х «К»

Хранение и выборка данных по Колонкам

- Для уменьшения ввода-вывода и быстроты выборки

Храним и выбираем как запрошено

Кластеризация

- Колонки и узлы дублируются
  - Восстановление по логам медленно и не практично, поэтому логов нет
- Обновляемая аппаратура запрашивает остальную часть системы по мере надобности

Растут запросы? Просто добавь «железа»

Компрессия данных

- Оптимизация хранения повторяющихся данных
- Свыше 12 схем компрессии
  - Выбор определяется данными
  - Тип сжатия система выбирает сама
- Обычно 50% – 90% сжатия
- Внутренние запросы в сжатой (кодированной) форме
- Стандартный SQL-интерфейс

Больше данных – меньше аппаратуры

Круглосуточная работа

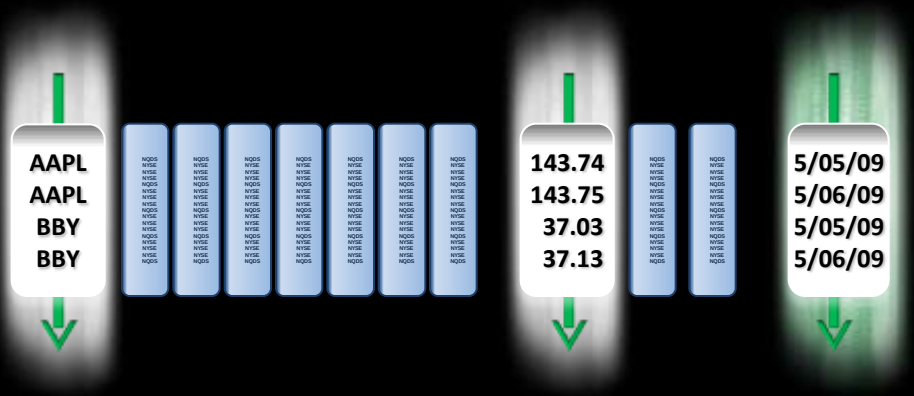
- Одновременная загрузка и обработка
  - не нужны «ночные окна»
- Изменения схем на ходу
  - Добавь колонки без остановки
  - Автоматический дизайнер баз данных
- Автоматическая репликация данных и восстановление узлов
  - Активный резерв увеличивает производительность

Запросы и загрузка 24x7 без администрирования

Хранение и выборка данных по Колонкам

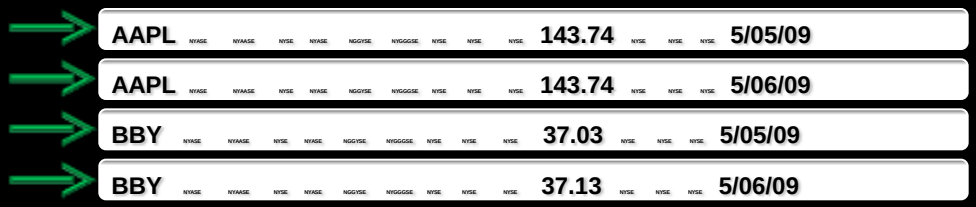
- Пример типового запроса по выборке средней стоимости акций на конкретную дату

Колоночное хранение – Читаем 3 колонки



SELECT AVG(price)
FROM tickstore
WHERE
symbol = 'AAPL" AND date = '5/06/09'

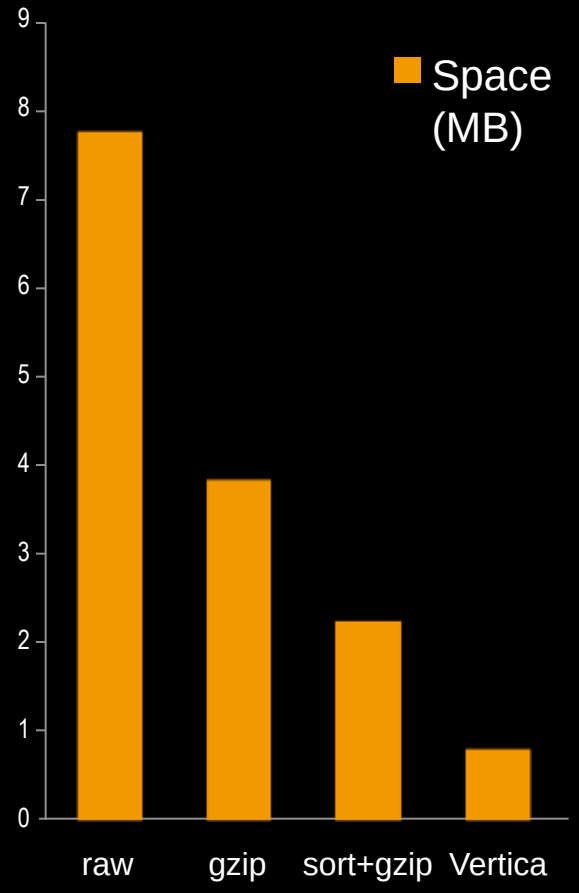
Построчное хранения – Читаем все колонки



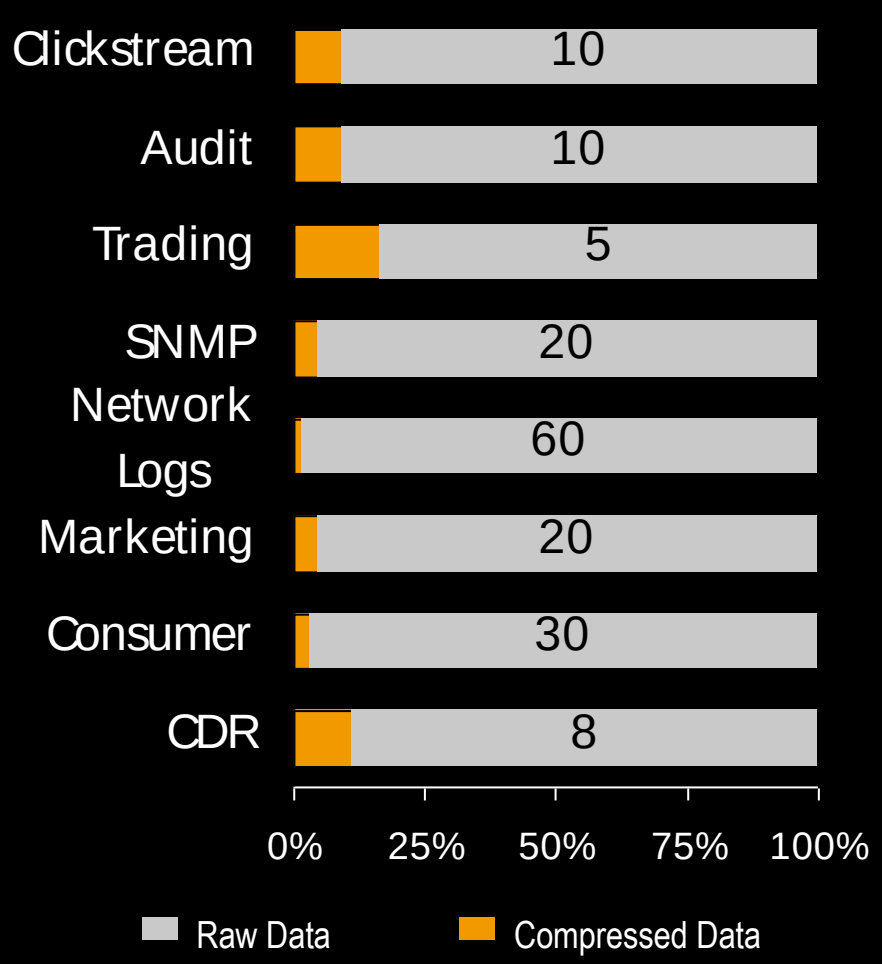
Компрессия данных

Влияние сортировки на эффективность компрессии

Алгоритм компрессии



В среднем по индустрии



Компрессия данных

Храним больше, работаем быстрее, меньше используем железа

- Сжатие 50% – 90%
- До ~1 петабайта в базе в одном шкафу
- Не требуется декомпрессия для обработки запросов
- Снижение требований к железу, сокращение капитальных затрат
- Множество типов компрессии, для сгруппированных типов данных

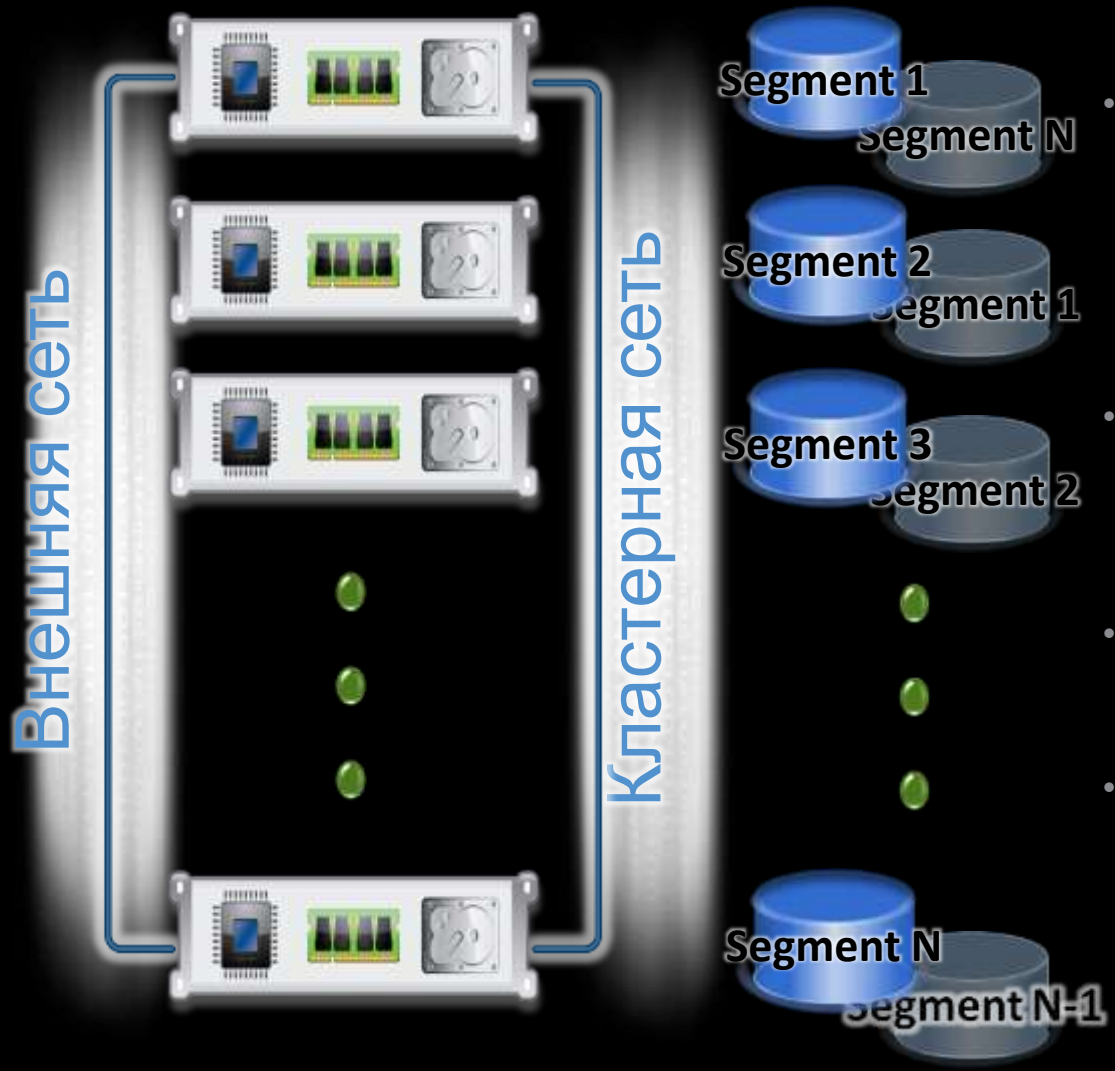
Кластеризация

Горизонтальное масштабирование, массовая параллельная обработка (MPP)

100% пиринговая сеть – нет блокировок

- Нет специализированных узлов
- Загрузка данных и запросы на любом узле
- Линейная масштабируемость
- Больше кластер = больше места для данных + выше производительность

Круглосуточная работа



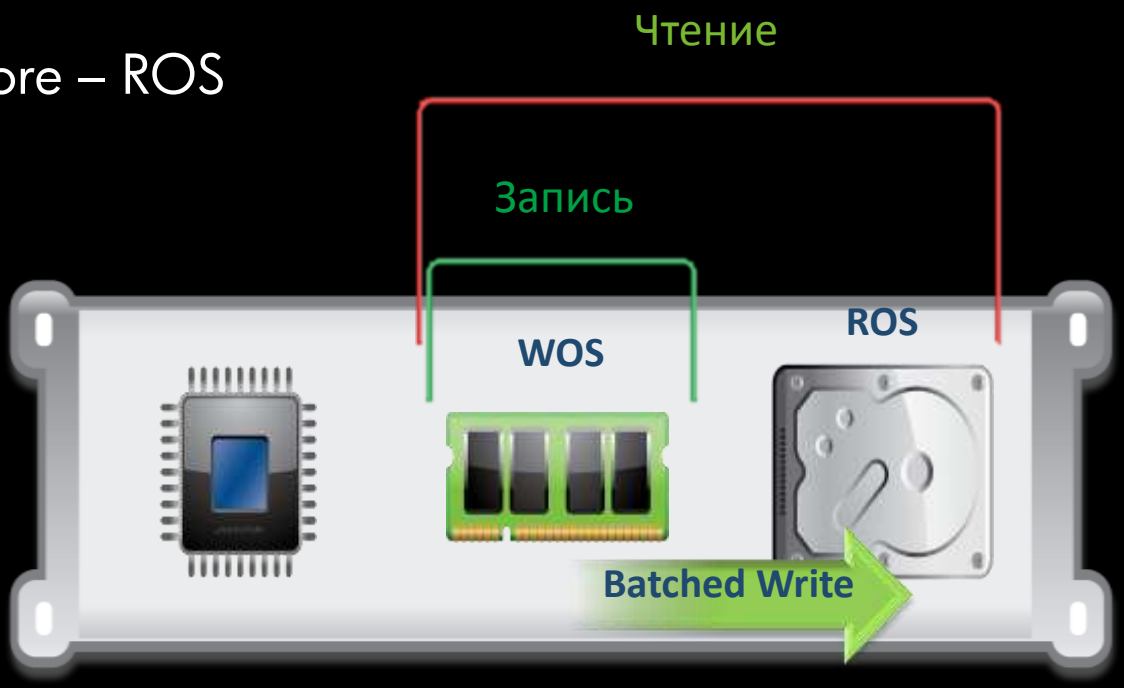
- Дублирование и даже троирование сегментов на уровне базы данных (K-safety)
- Постоянная готовность к загрузке новых данных и обработке запросов
- Избыточные копии повышают производительность
- Отсутствие единой точки отказа

Круглосуточная работа

Загрузка данных и обработка запросов одновременно в реальном времени

Гибридная in-memory/on-disk архитектура

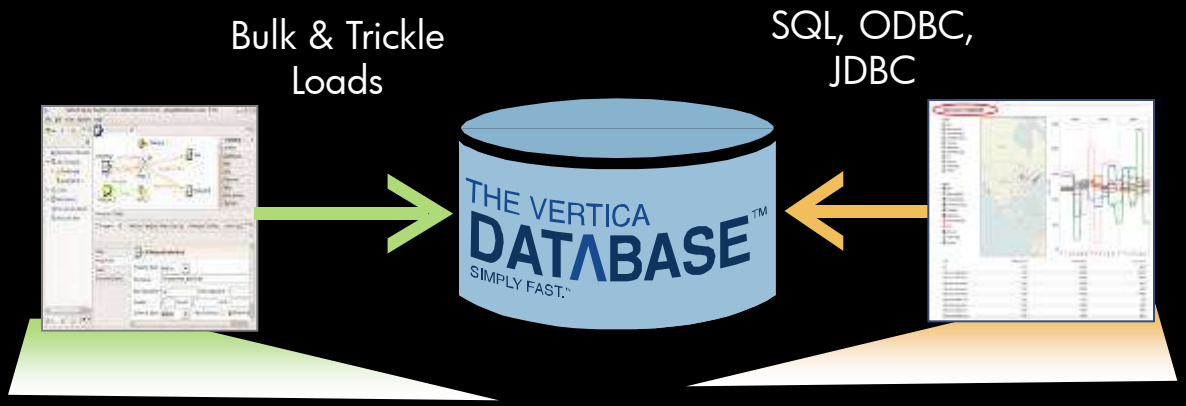
- Write-Optimized Store – WOS
- Read-Optimized Store – ROS
- Tuple Mover – TM (оптимизатор данных)



HP Vertica поддерживает стандартный SQL

Vertica поддерживает стандарт ANSI SQL-99 с расширением в части поддержки аналитики для упрощения интеграции с существующими инструментами BI и ETL

- ANSI SQL-99 +Analytics
- Простота интеграции
- Vertica's Hadoop Connector
- Database Connectors for
  - JDBC
  - ODBC
  - ADO.NET



ETL, Replication, Data Quality

IBM Ascential Software, talend open data solutions, Business Objects, pentaho open source business intelligence, INFORMATICA, GOLDEN GATE, syncsort, SAS

Analytics, Reporting

Business Objects, COGNOS, Microsoft, MicroStrategy www.microstrategy.com, pentaho open source business intelligence, JASPERSOFT, Spotfire, CORDA, SAS

Загрузи и попробуй http://myvertica.vertica.com/v-zone/download_vertica

свыше 400 заказчиков


Финансовые Сервисы


Ритейл


Телеком


Потребительский маркетинг


Здравоохранение


Интернет сетевые игры

http://www.vertica.com/customers/

Vertica – бесплатно?!

- Vertica Community Edition
  - 1ТБ сырых данных
  - 3 узла максимум.
- **Поддержка только через форум энтузиастов**
- **Машины – любые x86**

- **Открыта подписка на бета-версию**

www.vertica.com/community

Попробуй Vertica… на своих данных… в своем окружении.

www.hp.com/go/vertica

Больше данных. Больше открытий. Больше пользы. Сегодня!

- **Революционная платформа для аналитики в реальном времени** – спроектирована для решения задач завтрашнего дня, доступна уже сегодня
- **Проста в использовании** – быстрая отдача для бизнес-пользователей, DBAs, и программистов
- **Универсальная СУБД НЕ подходит для Больших Данных** – система должна быть специализирована и интегрирована
- **Производительность, гибкость** – ключевые факторы (и «кубы» строить не нужно)

Попробуй Vertica… на своих данных… в своем окружении.

www.hp.com/go/vertica

Спасибо!