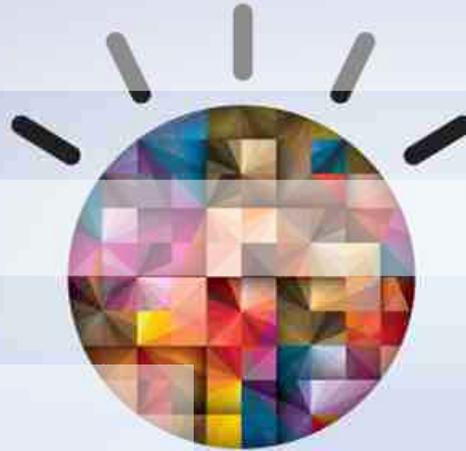


Стратегия и решения IBM в области обработки и анализа Больших Данных

Сергей Лихарев, IBM Information Management Solutions
SLikharev@ru.ibm.com



Наш инструментальный и взаимосвязанный мир

12+ TB
данных из ТВИТОВ
каждый день



? TBs данных
каждый день



25+ TB
данных
логов
каждый
день



30 млрд RFID
меток сегодня
(1.3 млрд в
2005)



4.6 млрд
мобильных
с камерой в
мире



100и млн
устройств
с GPS
каждый год



76 млн умных
счетчиков в 2009...
200М до 2014

2+ млрд
людей в
интернет
в конце
2011



Вызов: объединить вместе большие объемы разнородных данных и получить новое понимание



Понимание и анализ взаимодействия с клиентами



Раннее обнаружение угрозы здоровью в больницах



Предсказание погодных моделей для оптимального размещения ветряных турбин



Решения по принятию рисков на основе данных реального времени



Идентификация угроз безопасности на основе потоков видео и аудио данных

Что такое большие данные?

Извлечение знаний из огромных объемов разнородных данных, в контексте решаемой задачи



Variety: Управление сложностью разнородных и неструктурированных данных

Velocity: Большие объемы потоковых данных

Volume: От терабайт к петабайтам (1,000 TB) и к зетабайтам (1,000,000 TB)

Дилемма больших данных

- Процентное соотношение данных, которые компания может анализировать уменьшается по отношению с доступным данным
- Проще говоря, как организация, мы становимся “более наивными” со временем
- Мы не знаем что мы могли бы знать....

Данные доступные
для компании

Сигналы
и шум



11001000100100010010001100100

Данные, которые
компания может
обработать



Объединяя традиционный подход и большие данные

Традиционный подход
Структурный и повторяемый анализ

Большие данные
Итеративность и исследование

Бизнес

Определяет что спросить



ИТ

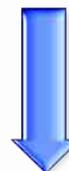
Структурирует данные для ответа на вопрос



Месячная отчетность
 Анализ прибыльности
 Анализ анкет

ИТ

Обеспечивает платформу для креативного анализа



Бизнес

Исследует что можно спросить



Отношение к бренду
 Стратегия продуктов
 Оптимизация ресурсов



Почему большие данные не появились раньше: разве у нас не было инструментов?



Инструмент – платформа IBM Big Data

Любой источник, с любой скоростью



- Анализ разнородных данных в огромных объемах
- Анализ потоковых данных
- Анализ структурированных данных

Платформа IBM Big Data

- Variety
- Velocity
- Volume



Платформа решает задачи корпоративных клиентов

Что необходимо

Обеспечить
гибкость

Фокус на конечном
результате

Сокращение и
управление
сложностью

Сокращение затрат
на разработку

Что дает платформа



- Платформа для V³
- Аналитика для V³
- Легкость использования для разработчиков и пользователей
- Масштаб предприятия
- Интеграция в инфраструктуру



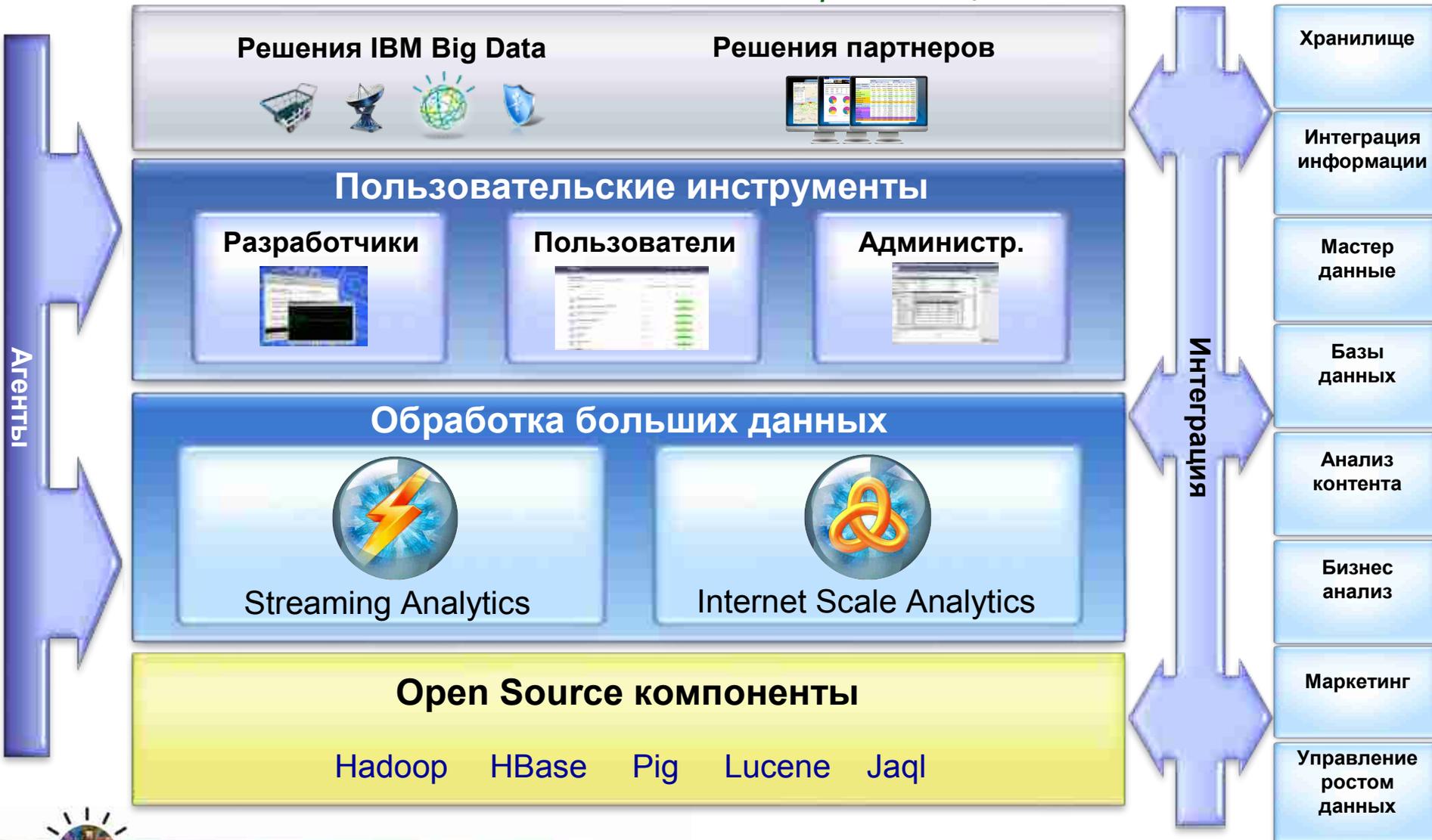
Большие данные – часть информационной архитектуры организации

- Важный компонент информационной архитектуры
- Не может быть отдельно стоящим решением
- Должен быть легко развертываемым и интегрируемым компонентом



Платформа IBM Big Data

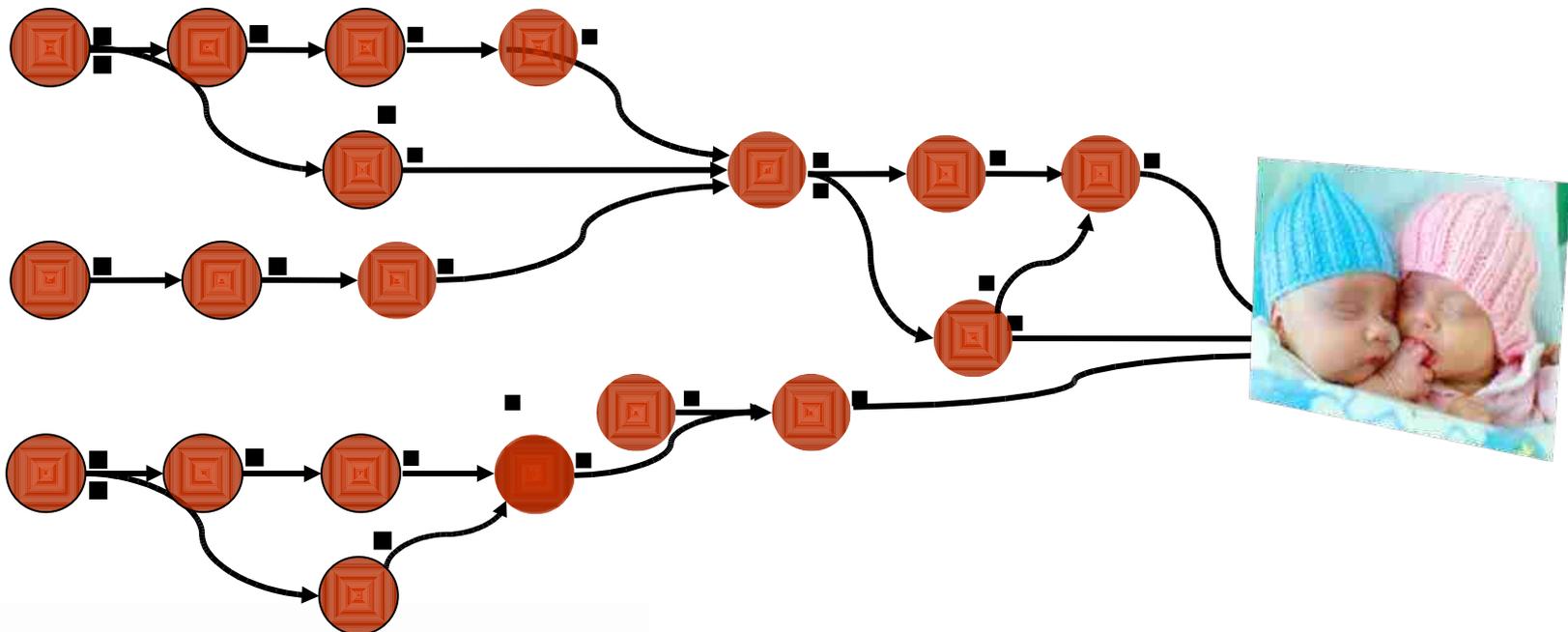
Большие данные в масштабах организации



Пример: данные «в движении»



- Раньше: сбор и анализ информации каждые 30-60 минут и удаление через 72 часа
- Сейчас: анализ 1000 элементов уникальной диагностической информации в секунду и сохранение в динамической модели
- Результат: снижение на 20% смертности в контрольной группе
 - 120 детей под контролем: 120К сообщений/сек... миллиарды в день



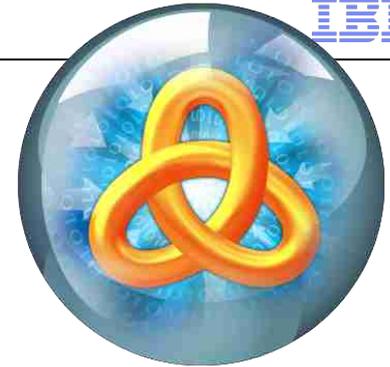
Что такое Streams?



- Платформа для разработки приложений, которые **быстро принимают, анализируют и сопоставляют** информацию, поступающую из сотен и тысяч источников в **реального времени**.
- **Время отклика порядка микросекунд** позволяет реагировать на события и тенденции по мере их возникновения.
- **Хорошо масштабируемая** и гибкая инфраструктура может поддерживать широкий спектр структурированных и неструктурированных типов данных.
- **Высокая доступность** и быстрое восстановление после сбоев.
- **Интеграция** с корпоративными СУБД/ХД: DB2, Informix, InfoSphere Warehouse, IBM Smart Analytics System, Netezza, Microsoft SQL Server, MySQL, solidDB, Oracle.



Что такое BigInsights?



- Позволяет хранить, анализировать и исследовать огромные объемы реляционных и неструктурированных данных (например: текст, аудио, видео, логи).
- Построено на Apache Hadoop, технологии с открытым кодом для управления большими объемами сложных данных.
- BigInsights расширяет возможности базовой технологии для удовлетворения корпоративным требованиям, добавляя средства администрирования, безопасности, развертывания, а также возможности аналитики от IBM Research.
- Интеграция с корпоративными СУБД/ХД: DB2, InfoSphere Warehouse, IBM Smart Analytics System и Netezza с помощью разработанного IBM языка Jaql.



Редакции nfoSphere BigInsights

Платформа для V³

- На базе Hadoop
- Индексирование

Аналитика для V³

- Интегрированный анализ текста

Удобство

- Development Studio
- Консоль администратора

Уровень предприятия

- Хранение и безопасность

Интеграция

- Интегрированная установка
- Связь с DB2, InfoSphere Warehouse и IBM Smart Analytics System.



IBM Text Analytics Toolkit позволяет вам...

- Разрабатывать приложения для анализа текстов до **50% быстрее** чем вручную
- **Выполнять быстрее** анализ текстов
- Получать **более точные и правильные ответы**



Применение анализа больших данных

Финансы

- Решения по рискам
- Анализ мнения клиентов
- Борьба с отмыванием денег

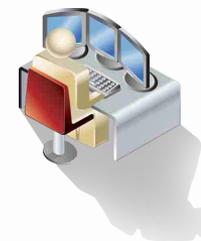


Энергетика

- Влияние погоды на генерацию энергии
- Анализ данных от умных счетчиков

Транспорт

- Влияние погоды и трафика на доставку и потребление топлива



ИТ

- Анализ логов от разных транзакционных систем

Колл центр

- Анализ расшифровок разговоров для понимания поведения клиентов

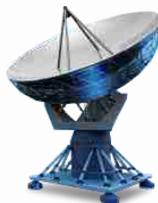


E Commerce

- Анализ поведения и покупательских моделей

Телко

- Анализ операций и сбоях сети



Интеграция каналов взаимодействия

- Моделирование поведения клиентов



Пример анализа потенциальных проектов в страховой компании

5 проектов

Голос клиента

- Что клиенты и эксперты говорят

Оптимизация внутренних операций

- Оптимизация распределения ресурсов на основе анкет, логов службы поддержки, документов..

Предсказательная аналитика

- Выявление подозрительных действий, возможных судебных исков, автоматизация маршрутизации работ

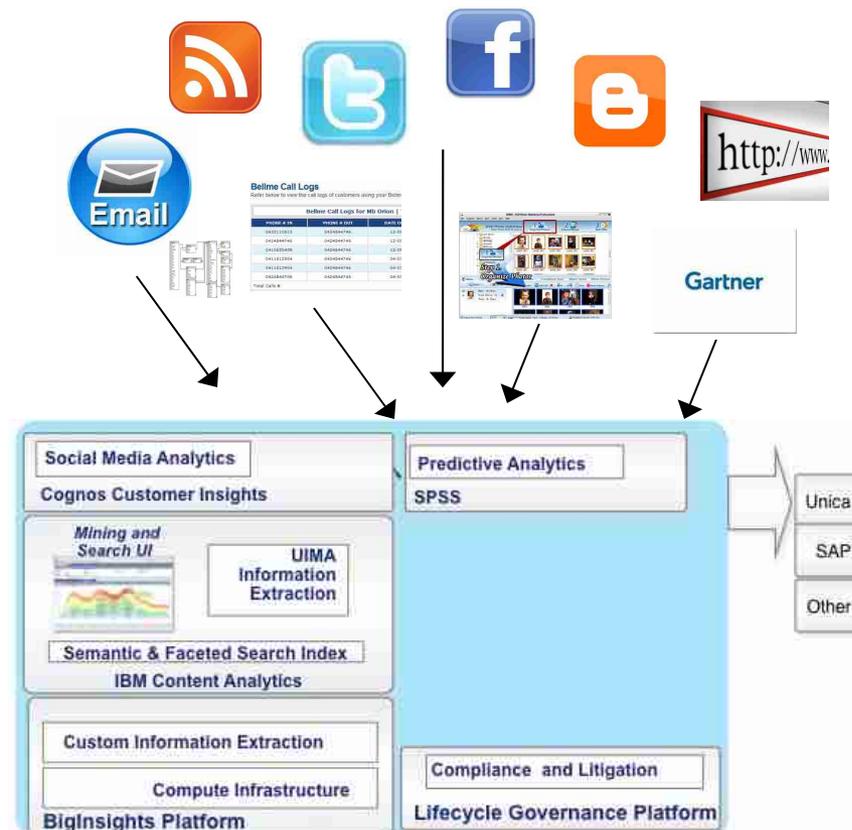
Конкурентный анализ

- Поиск и анализ информации из сети

Корпоративный поиск

- Поиск и организация документов по смыслу

Структурированные и неструктурированные данные



Какое наше следующее лучшее предложение?



BigInsights в облаке

Изучаем Hadoop!

- **Гибкое** дистанционное обучение на рабочем месте в удобное время
- **Бесплатные** курсы, **бесплатные** учебные материалы
- Область для экспериментов в облаке – **ничего не нужно устанавливать и настраивать**
- 8500+ студентов
- Конкурсы разработки на Hadoop

The screenshot displays the BigDataUniversity website interface. At the top, the logo reads "BigDataUniversity BETA" with the tagline "Learn from the industry's best". Navigation links include HOME, LEARN, DOWNLOAD, RESOURCES, JOBS, and LEARN Hadoop. A search bar is present with the text "search courses".

Key content areas include:

- Hadoop Fundamentals:** A prominent orange banner for a "FREE course" that allows users to "Start your career on Hadoop with this FREE course" and "Learn with hands-on exercises on a Hadoop cluster". An "Enroll now!" button is visible.
- Why register?:** A section with three bullet points:
 - Easy and Affordable:** "Learning Hadoop and other Big Data technologies has never been more affordable! Many courses are FREE!"
 - Latest industry trends:** "Acquire valuable skills and get updated about industry's latest trends right here. Today!"
 - Learn from the Experts!** "Big Data University offers education about Hadoop and other technologies by the industry's best!"
 - Learn at your Own Pace!** "Find everything right here when you need it and from wherever you are."
- Sign up now for a chance to receive: FREE Books:** A call-to-action with a "SIGN UP" button.
- Study Made Easy!** A section header.
- Student Testimonials:** A section titled "Balázs (USA)" with a "go to sign up" link. The testimonial describes the training material as having short, easy-to-digest videos with transcripts that can be cut and pasted for reproduction. It also mentions online support on course forums.

At the bottom right, there are links for "about us", "legal", "contact", and "bug reports".



THINK

BIG

BIG

Вопросы?

