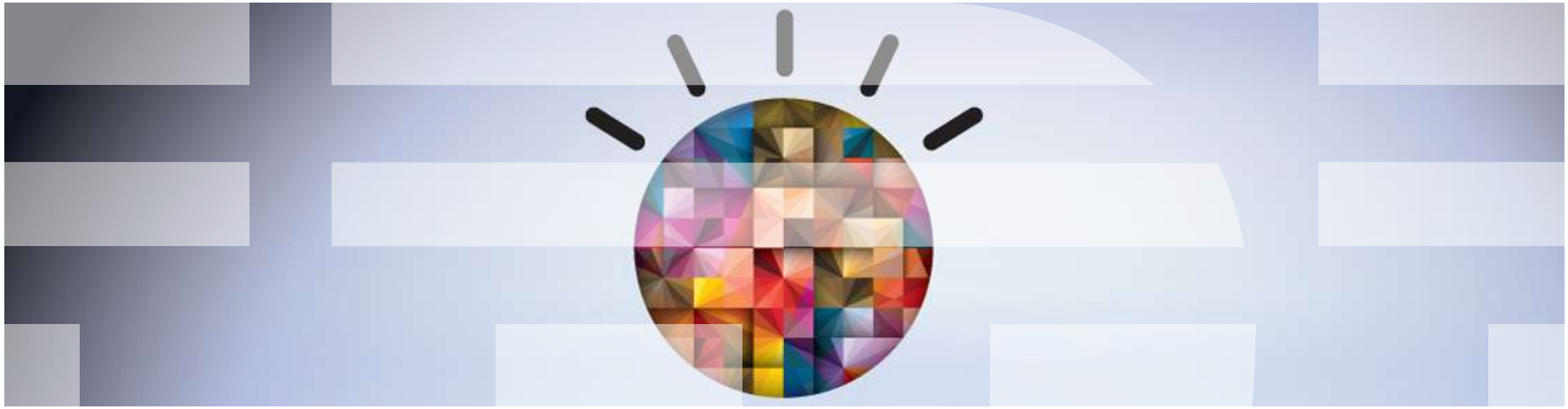
ЕС-лизинг

IBM

Современные тенденции Больших Данных.
Взгляд технолога

Б.А. Позин
Д.т.н., профессор
Технический директор ЗАО «ЕС-лизинг»



ЕС-лизинг

Роль технологии больших данных в национальной экономике

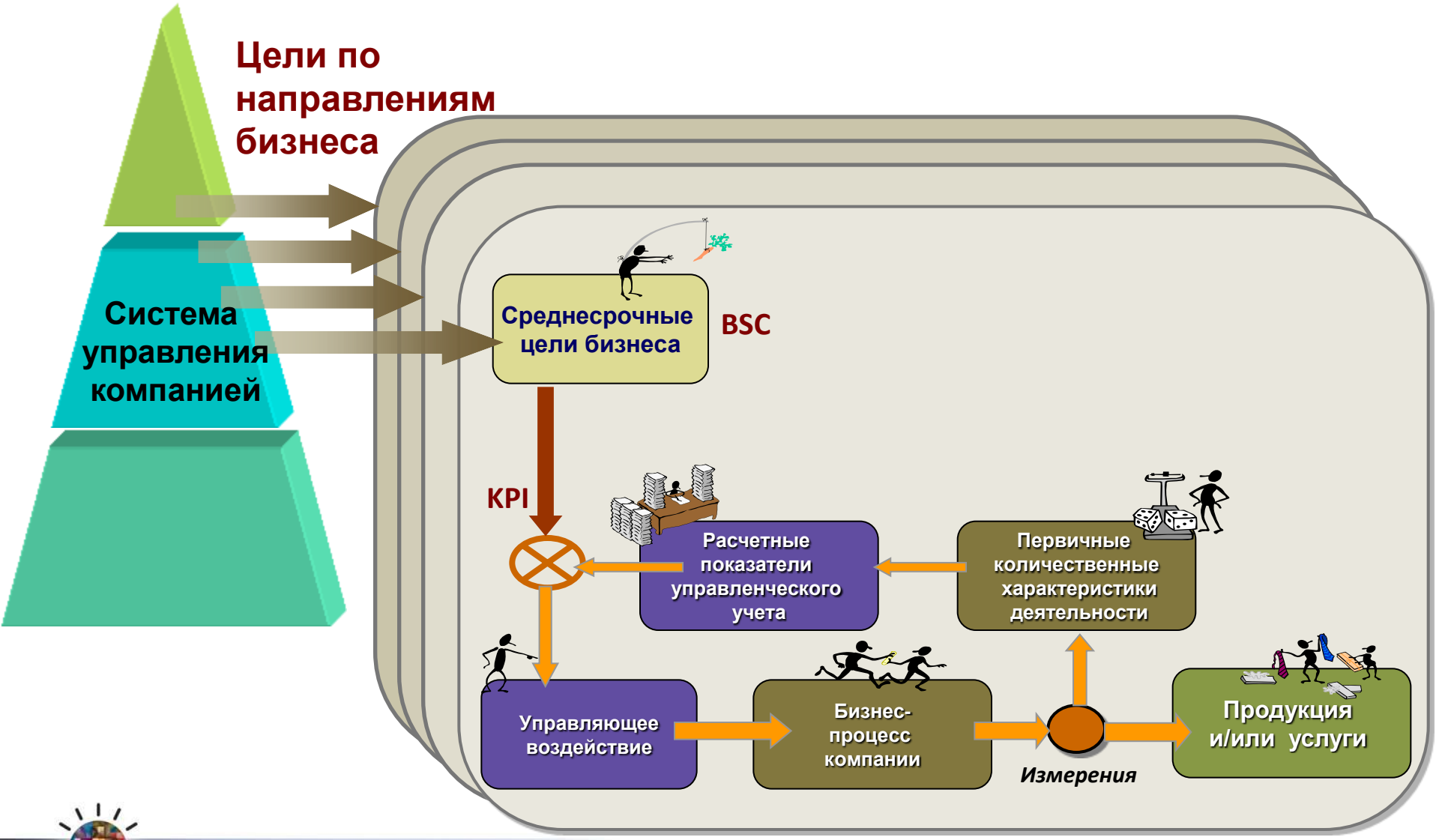
- Роль технологии Больших данных в национальной экономике будет повышаться по мере того и в той степени, как и в которой технология Больших данных будет использоваться для решения практических задач компаний
- Объемы производства и в этом секторе, как и ранее, будут определяться как **ЧИСЛОМ** работающих, так и **ПРОИЗВОДИТЕЛЬНОСТЬЮ ТРУДА** работающих
- Важно как можно быстрее начать работать на готовых платформах Больших данных, чтобы обучить специалистов по информационно-аналитическим системам, создать работающие в компаниях технологии, а затем совершенствовать как технологии, так и используемые в них методы

ЕС-лизинг

Факторы и процессы, влияющие на управление компанией

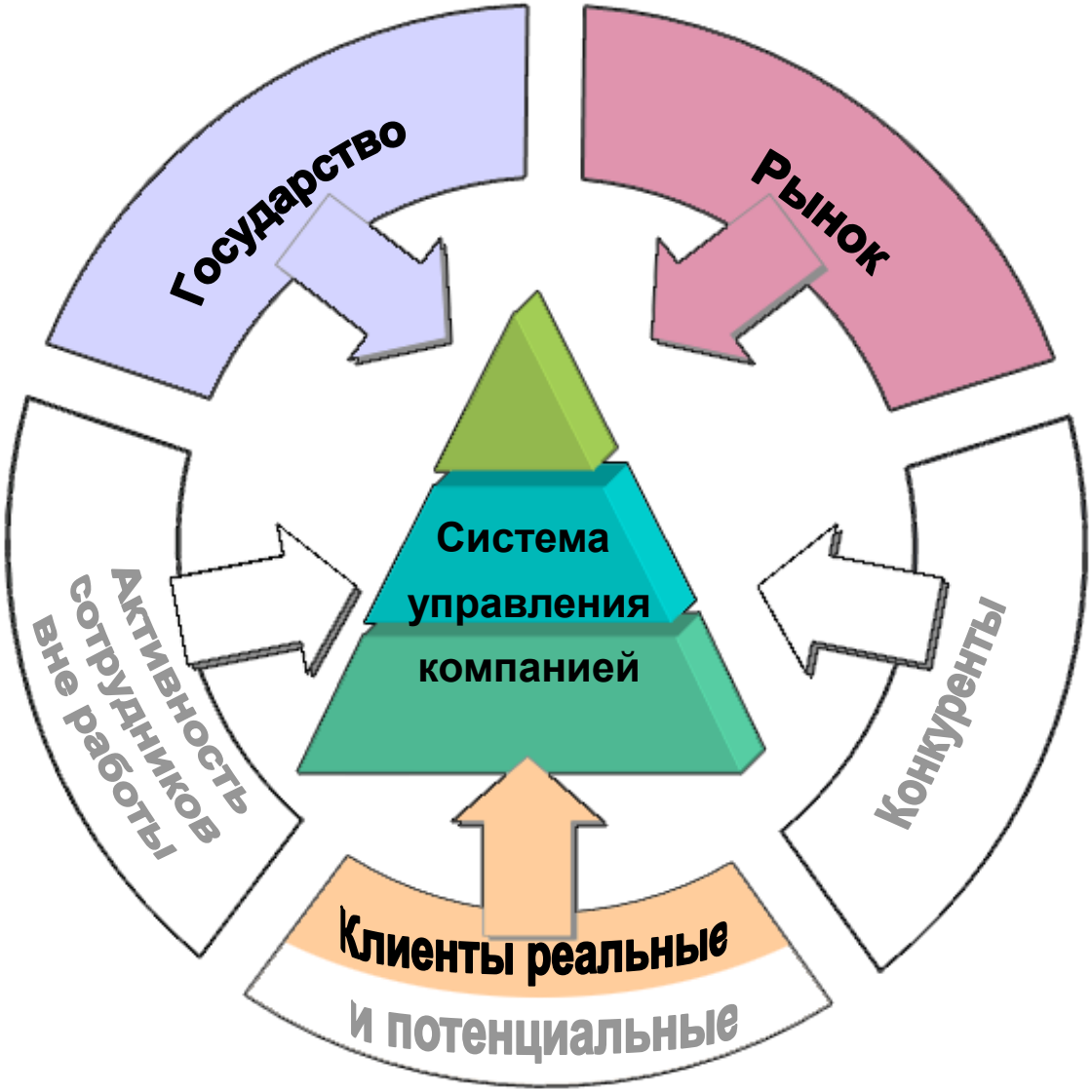
ЕС-лизинг

Направления бизнеса компании



ЕС-лизинг

Использование структурированных данных для управления компанией

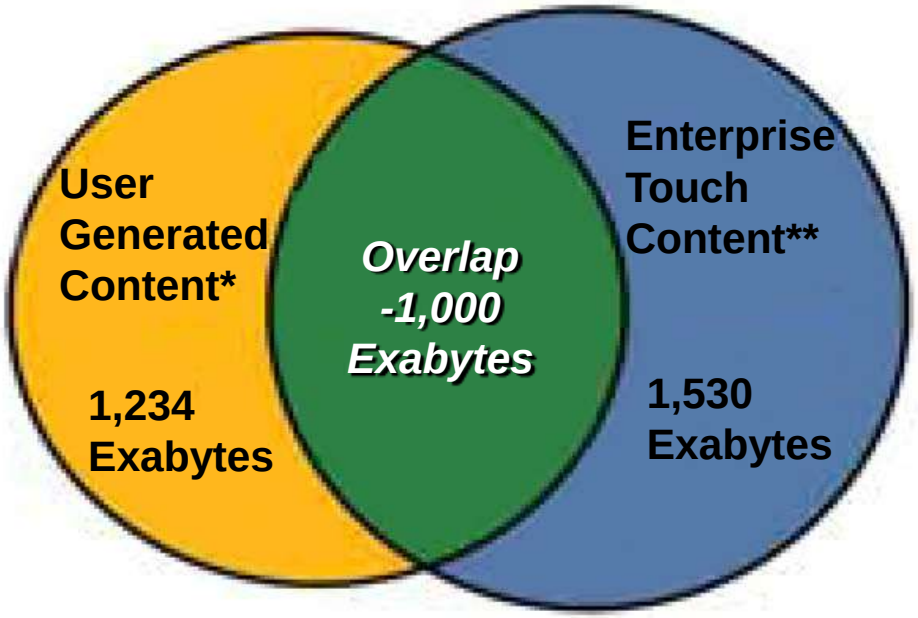


Каких данных сколько?

Figure 5

User Creation: Enterprise Worries

*Обсуждение личных проблем, планов и событий



**Текущая информация о действиях конкурентов, интервью конкурентов, светская хроника, аналитические материалы и пр.

Size of Digital Universe in 2011

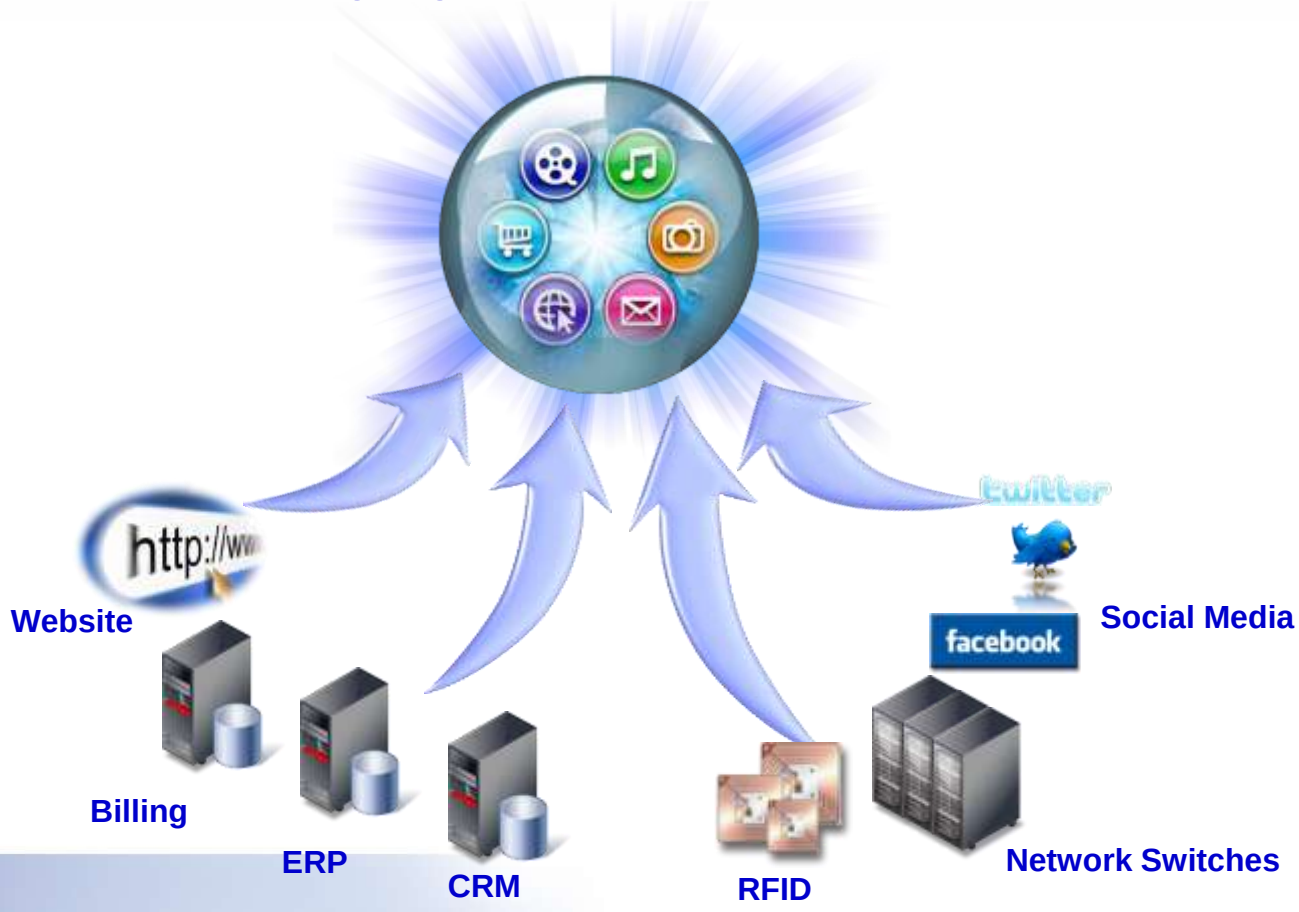
1,773 Exabytes

Source: IDC, 2008

Привлечение новых данных для управления компанией

Большие данные – горячая тема, потому что технологии сделали возможным анализ ВСЕХ доступных данных

Эффективно, с точки зрения затрат, управлять и анализировать все доступные данные в их первозданном виде – структурированные, неструктурированные, потоковые



Наиболее эффективный подход к реализации ИАС

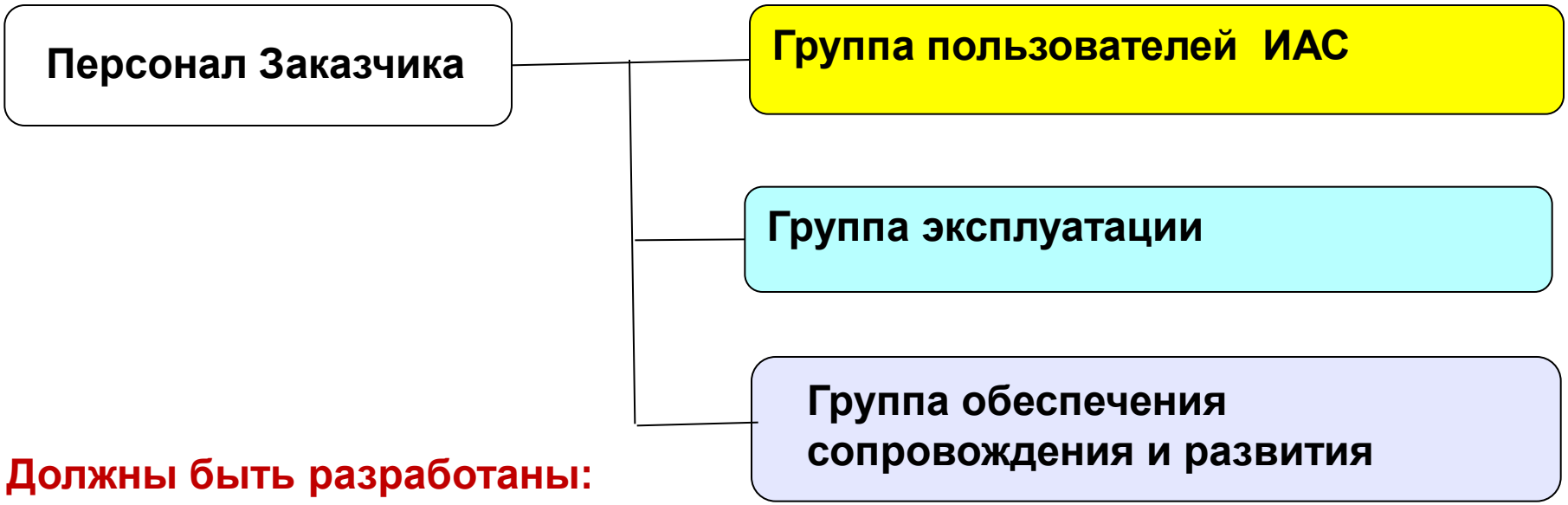
- Как можно быстрее выйти на реальные источники данных
- Как можно быстрее выйти на создание прототипа информационно - аналитической системы, чтобы совместно с конечным пользователем определить цели и методики анализа
- Развивать прототип средствами, доступными для понимания предметников, их руками и таким образом, чтобы выделить метаданные для интеграции и настройки отдельных инструментов, получения инструмента для комплексного анализа информации, для ручного обучения ИАС и в последующем самообучения ИАС
- Отрабатывать методики и инструменты функционального развития ИАС в процессе освоения предметной области

Вывод: нужен аналог «крупноблочного строительства» ИАС

Особенности ИАС, построенной на базе платформы

Система = Люди + Инфраструктура + программное обеспечение

Люди



Должны быть разработаны:

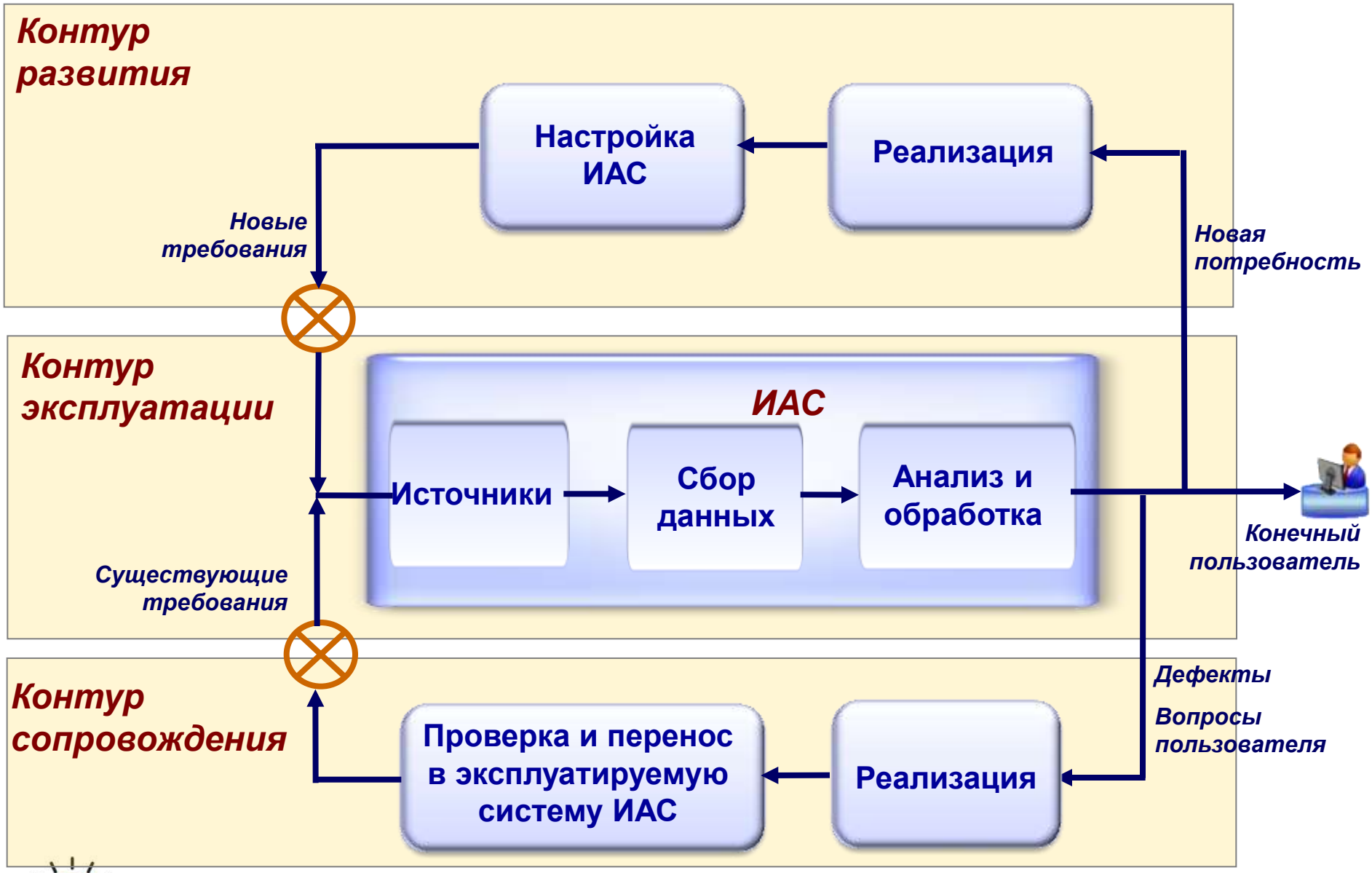
- **оргструктура,**
- **бизнес-процессы деятельности персонала ИАС,**
- **регламенты деятельности**

Инфраструктура : линейка системотехнических платформ (IBM) для создания ИАС

Программное обеспечение: платформа (IBM Big Data), настроенная на решение задач конечного пользователя

ЕС-лизинг

Обеспечение жизненного цикла ИАС



ЕС-лизинг

«Хирургическая бригада», или креативная команда

Роль	Компетенция
Аналитик	•Определение цели (что ищем/делаем), детализация информационной потребности
Ассистенты – предметники	•Представление о предметной области на сегодня, предметная постановка задачи, •где ищем, критерии оценки результата, •контроль состава предметной области, •изменения потребности и соответствующих областей поиска
Специалист по данным (Data Scientist)	•Свойства данных, • как хранить разнородные данные и информацию, • как обрабатывать данные и формировать информацию
Специалист – технолог (Data Scientist)	Какими методами и инструментами ищем

ЕС-лизинг

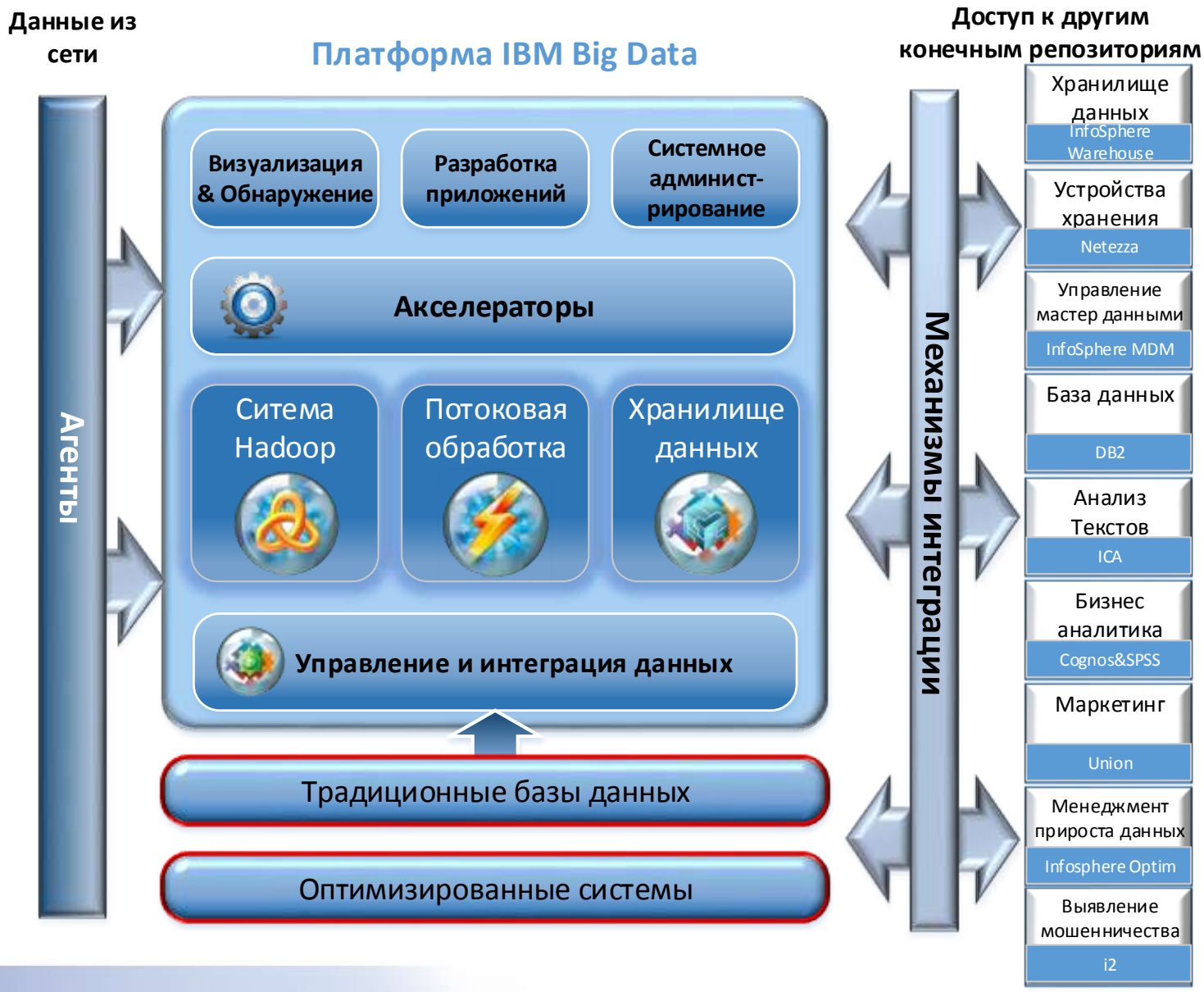
Требования к оснащению креативной команды

Роль	Потребность в языках	Потребность в инструментах
Аналитик	Языки описания требований и критериев	Средства визуализации информации Средства обобщения и сопоставления информации, добытой членами бригады
Предметник	Высокоуровневое описание задачи	Поддержка языка описания задачи
Data Scientist (данные)	Языки описания данных	Инструменты хранения разнородных видов данных и информации, обработки данных и формирования информации
Data Scientist (методы и инструменты)	Языки описания методов обработки и оптимизации расчетов	Инструменты преобразования данных, их агрегирования, добычи, разных видов обработки с целью получения информации, средства презентации информации предметникам и ЛПР

Платформа (комплекс инструментов) должна быть

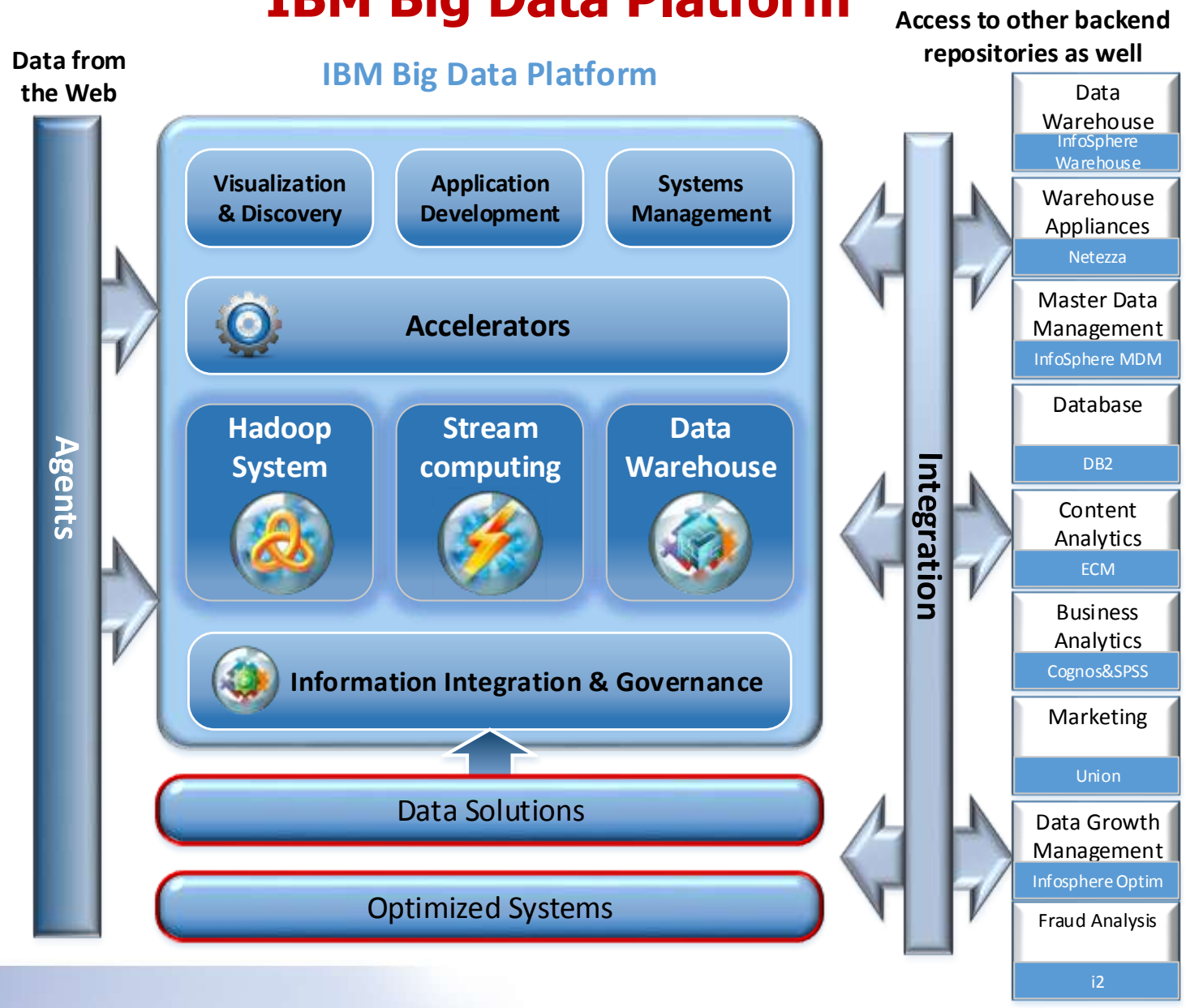
- **функционально полна**
- **должна обеспечивать максимальную производительность труда при сборке ИАС**

Платформа IBM Big Data



ЕС-лизинг

Подход к построению информационно-аналитических систем на базе платформы IBM Big Data Platform



ЕС-лизинг

Технологии IBM для использования в проектах Big Data

- **IBM Big Data platform**
  - InfoSphere Streams
  - InfoSphere BigInsights
  - InfoSphere Data Explorer
  - PureData for Analytics (Netezza)

- **Акселераторы разработки**
  - Анализ текстов
  - Акустика
  - Гео-данные
  - Видео
  - Интеллектуальный анализ
  - Предсказательные модели
  - Статистика

- **Аналитические пакеты**
  - IBM Cognos
  - IBM SPSS
  - IBM Content Analyzer
  - i2

- **Интеграция данных**
  - IBM InfoSphere Information Server
  - IBM Change Data Capture

- **Мастер-данные**
  - IBM InfoSphere Master Data Management Server

- **Защита баз данных**
  - InfoSphere Guardium

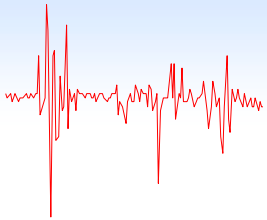
ЕС-лизинг

Акселераторы : ускорители разработки прикладных задач

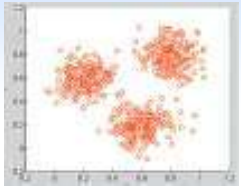
Более умная аналитика!!!

Текст (слушать, глагол), (радио, существительное)

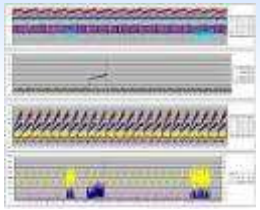
Простой & Комплексный текст



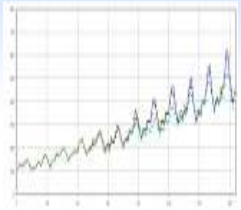
Звук



Добыча в микросекундах



Комплексные математические модели



Прогнозирование

$$\sum_{population} R(s_t, a_t)$$

Статистика

Геопространство

Фото & Видео

Готовые аналитические функции

Статистика

- Descriptive Statistics+
- Distance Measures*
- Hypothesis Testing*
- Chi-Square & Contingency Tables*
- Univariate & Multivariate Distributions+
- Monte Carlo Simulation*

Преобразование

- Data Profiling / Descriptive Statistics+
- General Diagnostics
- Statistics+
- Sampling
- Data prep

Временные ряды

- Autoregressive+
- Forecasting*

Математические методы

- Basic Math*
- Permutation and Combination*
- Greatest Common Divisor and Least Common Multiple*
- Conversion of Values*
- Exponential and Logarithm*
- Gamma and Beta Functions
- Matrix Algebra+
- Area Under Curve*
- Interpolation Methods*

Анализ данных

- Association Rules+
- Clustering+
- Feature Extraction+
- Discriminant Analysis*

Прогнозирование

- Linear Regression+
- Logistic Regression+
- Classification
- Bayesian
- Sampling
- Model Testing

Гео пространство

- Geospatial Data Type
- Geometric Functions
- Geometric Analysis

* Fuzzy Logix DB Lytix capabilities

\+ Netezza Analytics and Fuzzy Logix DB Lytix capabilities

Общая схема компонентов платформы Big Data

Инструменты Streams

Обработка потоковой информации

Streams

Декларативный язык: Stream Processing Language (SPL)

Готовые средства разработки (акселераторы разработки):

Анализ текстов
Телекоммуникационные данные
Гео-данные
Видео
Интеллектуальный анализ
Предсказательные модели
Статистика
Анализ машинных журналов (СПО)
Анализ данных из сетей (СПО)

Инструменты:

Standard Toolkit
Internet Toolkit
Database Toolkit
Financial Toolkit
Data Mining Toolkit
Big Data toolkit
Text Toolkit

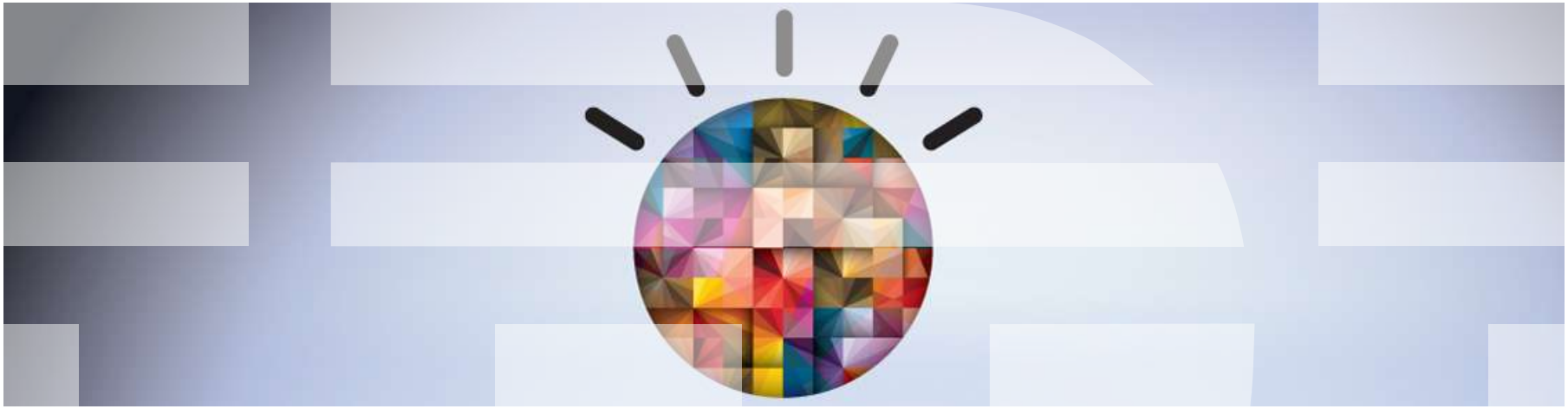
Коннекторы:

Netezza Connector
Hadoop Connector

Языки программирования 3-го поколения:

Java, C/C++, Python, Perl, JavaScript, Ruby и т.д.

Спасибо за внимание!



Вопросы?