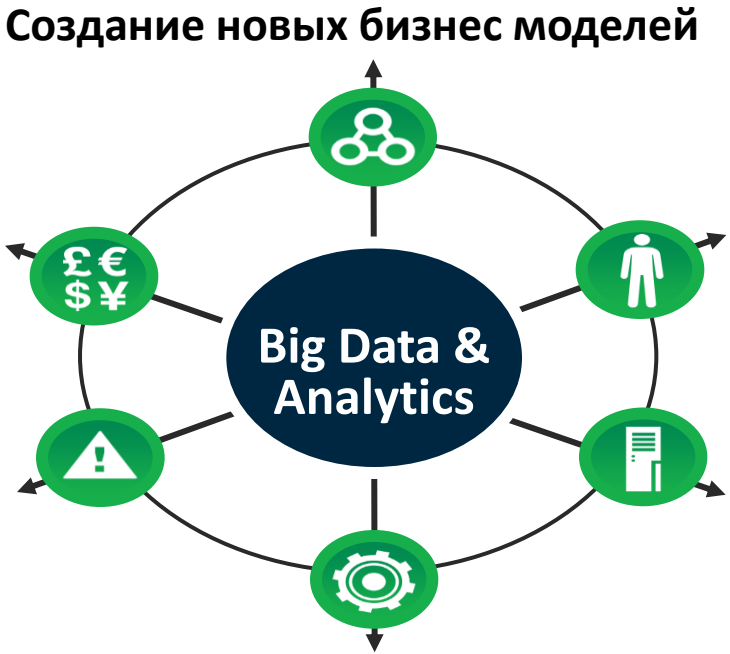
BusinessConnect
соединяя бизнес и технологии

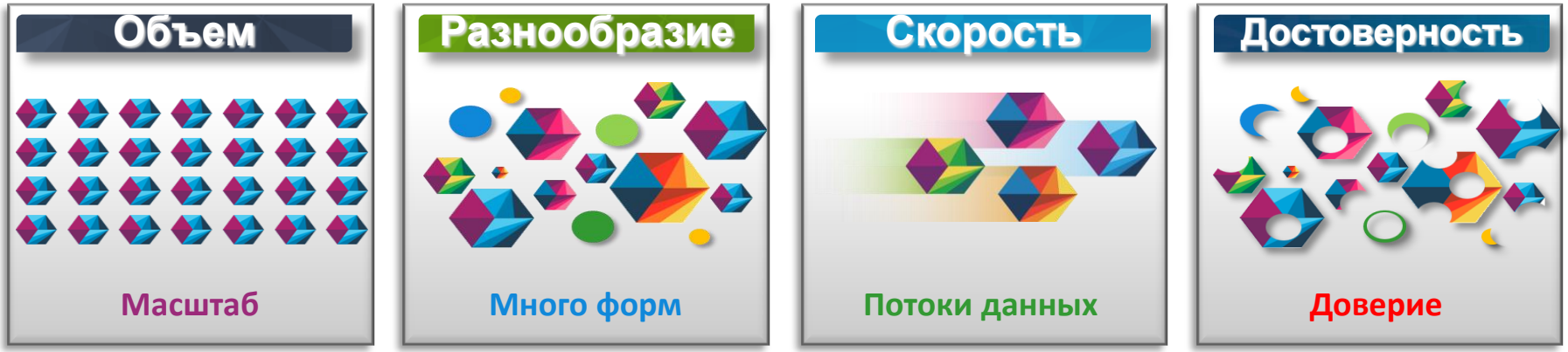
Большие данные: новые возможности и конкурентное преимущество для бизнеса

Сергей Лихарев
IBM Analytics

Аналитика помогает решить ключевые бизнес задачи предприятия



Большие данные – все данные



Необходима новая архитектура работы с данными

IBM Watson Analytics – инновации в аналитике

Аналитика для большого круга пользователей
Упрощенный доступ и очистка данных
Доставка через облако для гибкости и скорости

Понять бизнес

Автоматическое исследование данных позволяет задать правильные вопросы

Рассказать историю

Интеллектуальная визуализация поддерживает процесс принятия решений и презентации результатов

Думать вперед

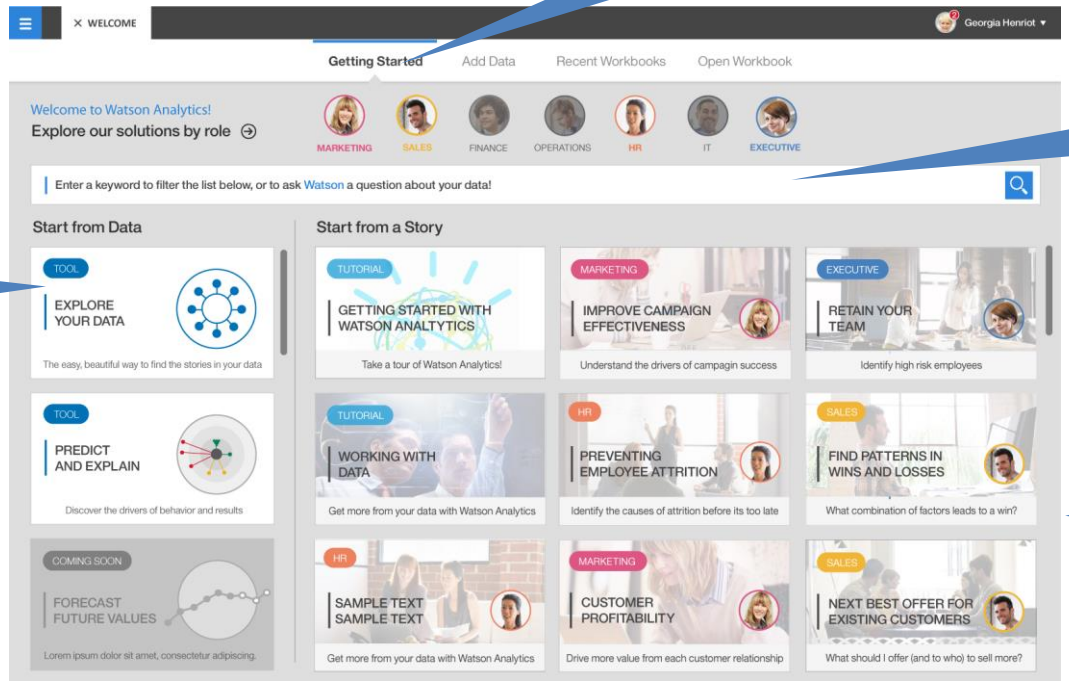
Предиктивная аналитика открывает новые возможности и скрытые знания

Получить более качественные данные

Встроенные средства очистки и улучшения качества данных

IBM Watson Analytics

Быстрый старт
Понятный интерфейс



Диалог с системой

Исследование данных

Доступно для мобильных устройств

Гибкость облачной среды

IBM Watson Analytics

Доступ и очистка данных

Взаимодействие

Интеллект без настройки

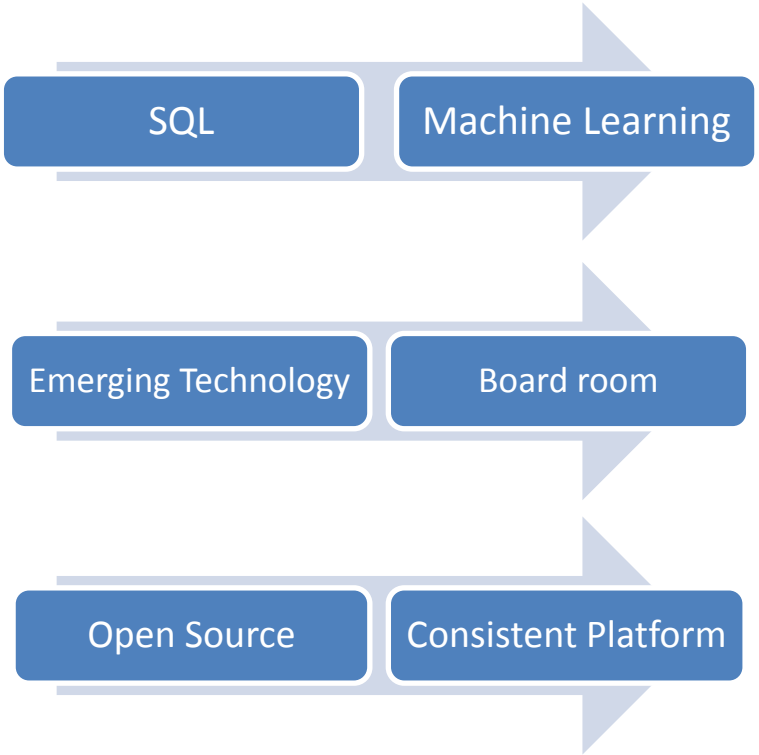
Отчеты и информационные панели

Подсказки в исследовании

Связывание элементов в историю

Использование всех доступных форм данных

Рынок Hadoop развивается



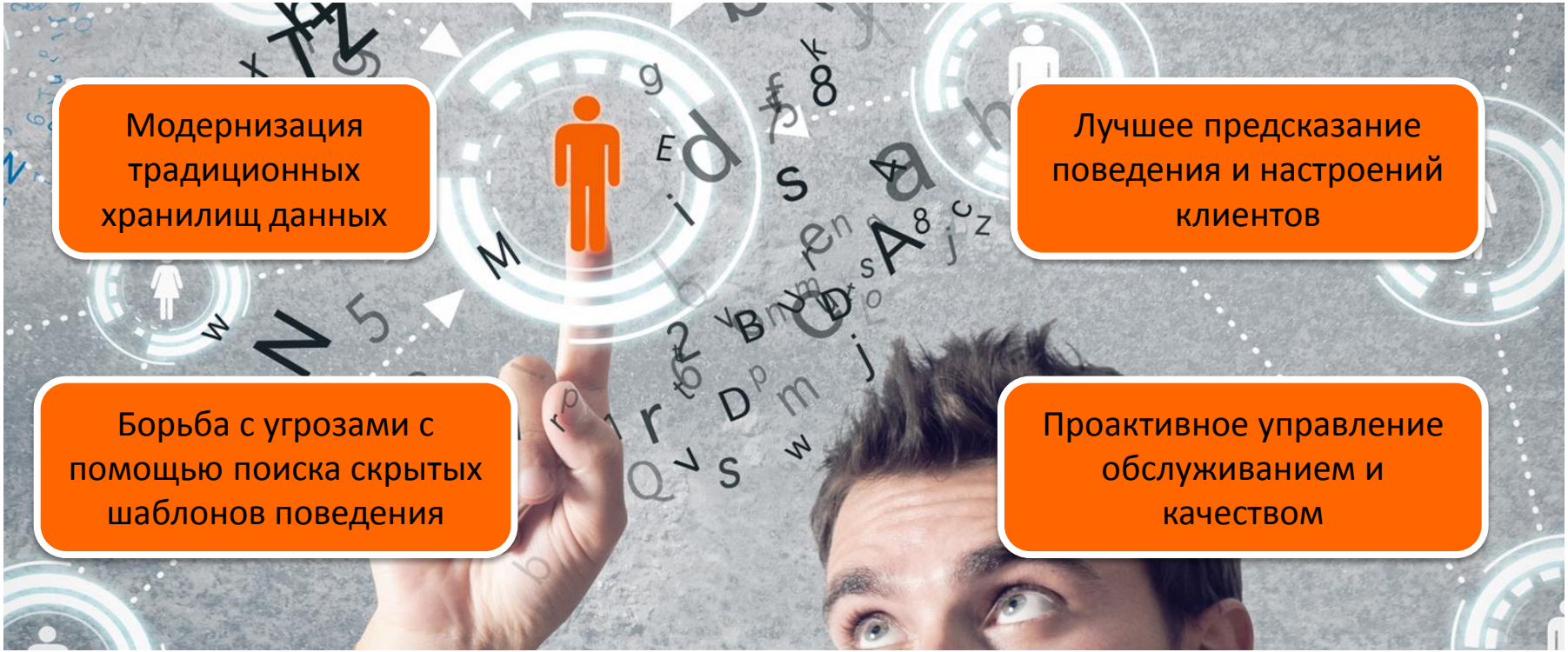
От доступ к данным для разработчика к исследованию данных для data scientist

От перспективной технологии к основе для изменения бизнес стратегии

Индустрия идет к открытой и целостной платформе с доступом к инновациям для всех

Использование данных для новых знаний о бизнесе и лучших решений

IBM BigInsights for Apache Hadoop



IBM BigInsights for Hadoop: 100% Open Source Hadoop, и все что нужно для предприятия

Дополнительные возможности

SQL on Hadoop
Big SQL – optimized ANSI compliant SQL

Шаблоны приложений
Toolkits and accelerators

Поиск
BigIndex and Data Explorer

Исследование данных
BigSheets “schema-on-read”

Предиктивное моделирование
Big R – scalable data mining

Анализ текстов
Advanced text processing with AQL

Аналитика реального времени
InfoSphere Streams

Управление данными и безопасность
Data Click, LDAP, Secure cluster

Интеграция с системами хранения
GPFS - POSIX Distributed Filesystem

Производительность и надёжность
Adaptive MapReduce, Recoverable jobs

100% Standard Apache Open-Source компоненты

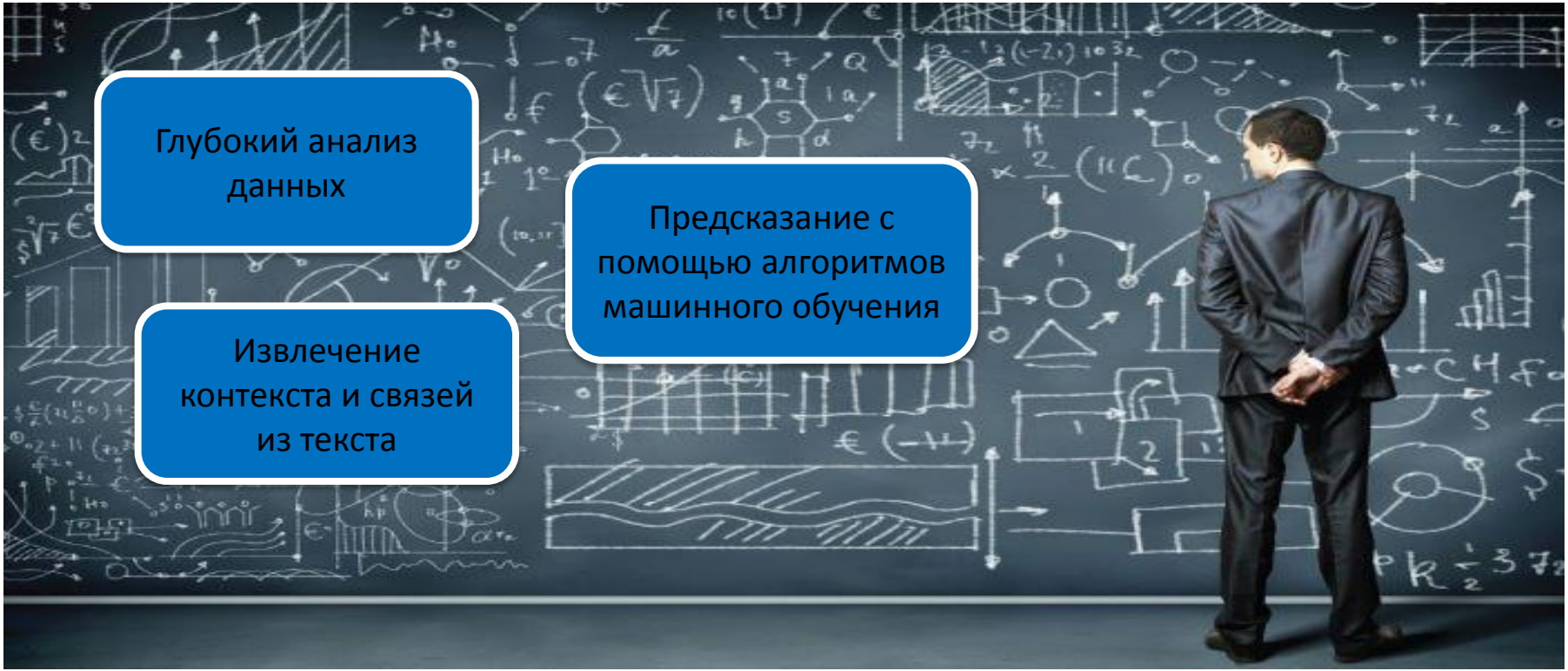
Oozie	Jaql	Zookeeper	Hive	HCatalog
HDFS	MapReduce	HBase	Flume	Sqoop
YARN	Spark	Avro	Pig	Solr/Lucene

Возможности для специалистов

Роль	*Business Analyst*	*Data Scientist*	*Administrator*
			
Потребность	• Выявление данных для анализа • Визуализация данных для действий • Использование существующих навыков (SQL, spreadsheets)	• Выявление шаблонов, трендов, результаты алгоритмов машинного обучения • Статистические модели на больших объемах данных	• Управление нагрузкой и обеспечение уровня производительности • Реализация политик безопасности для снижения рисков

Data science на Hadoop

IBM BigInsights Data Scientist module



3 ключевых возможности Big R

Использование популярного языка R на Hadoop

- Запуск native R функций
- Использование существующих ресурсов R (code & CRAN)

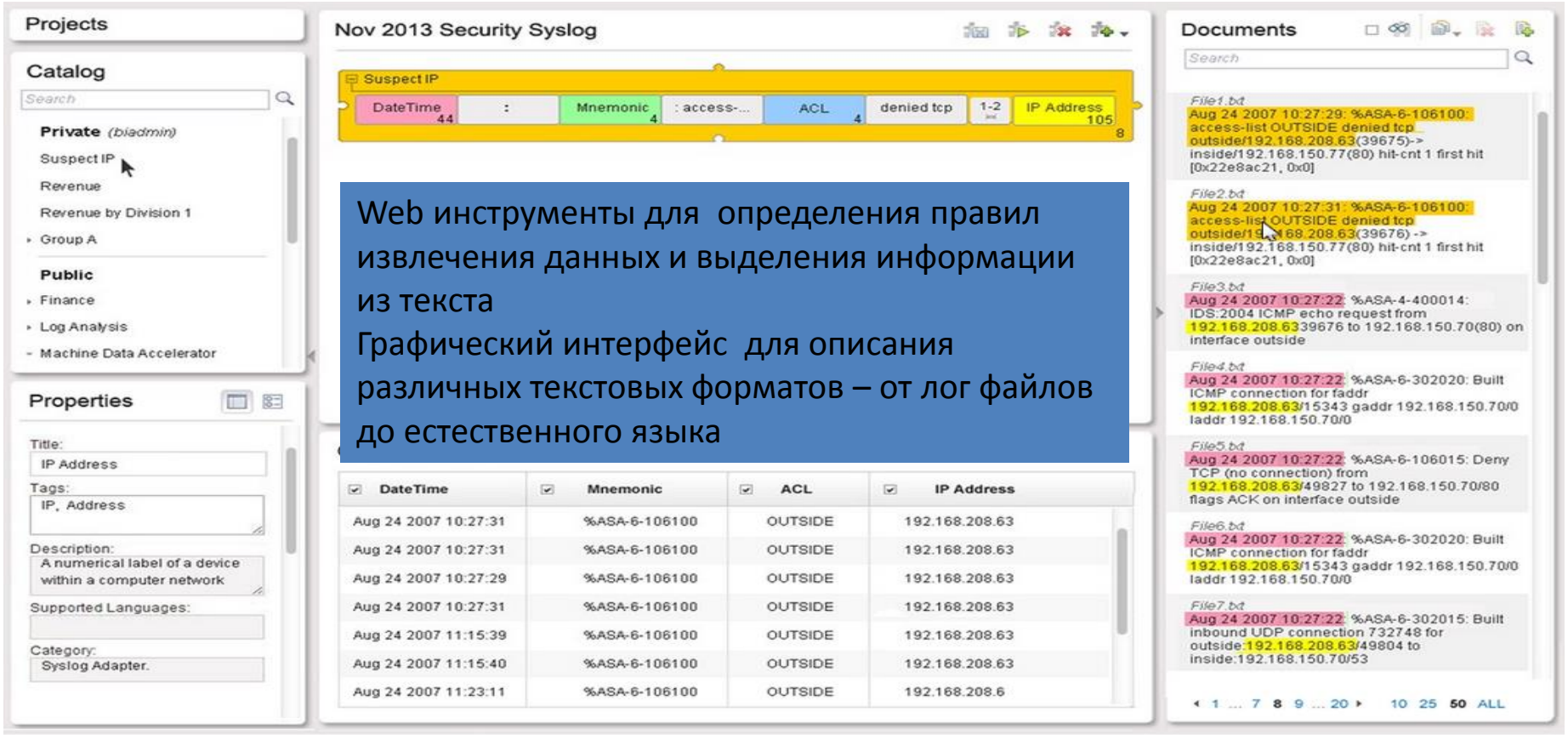
Новое: Запуск масштабируемых machine learning алгоритмов за пределами R на Hadoop

- Широкий набор алгоритмов и список растет
- R-like синтаксис для новых алгоритмов и настройки существующих

Новое: Использование масштабируемости Hadoop для ускорения анализа

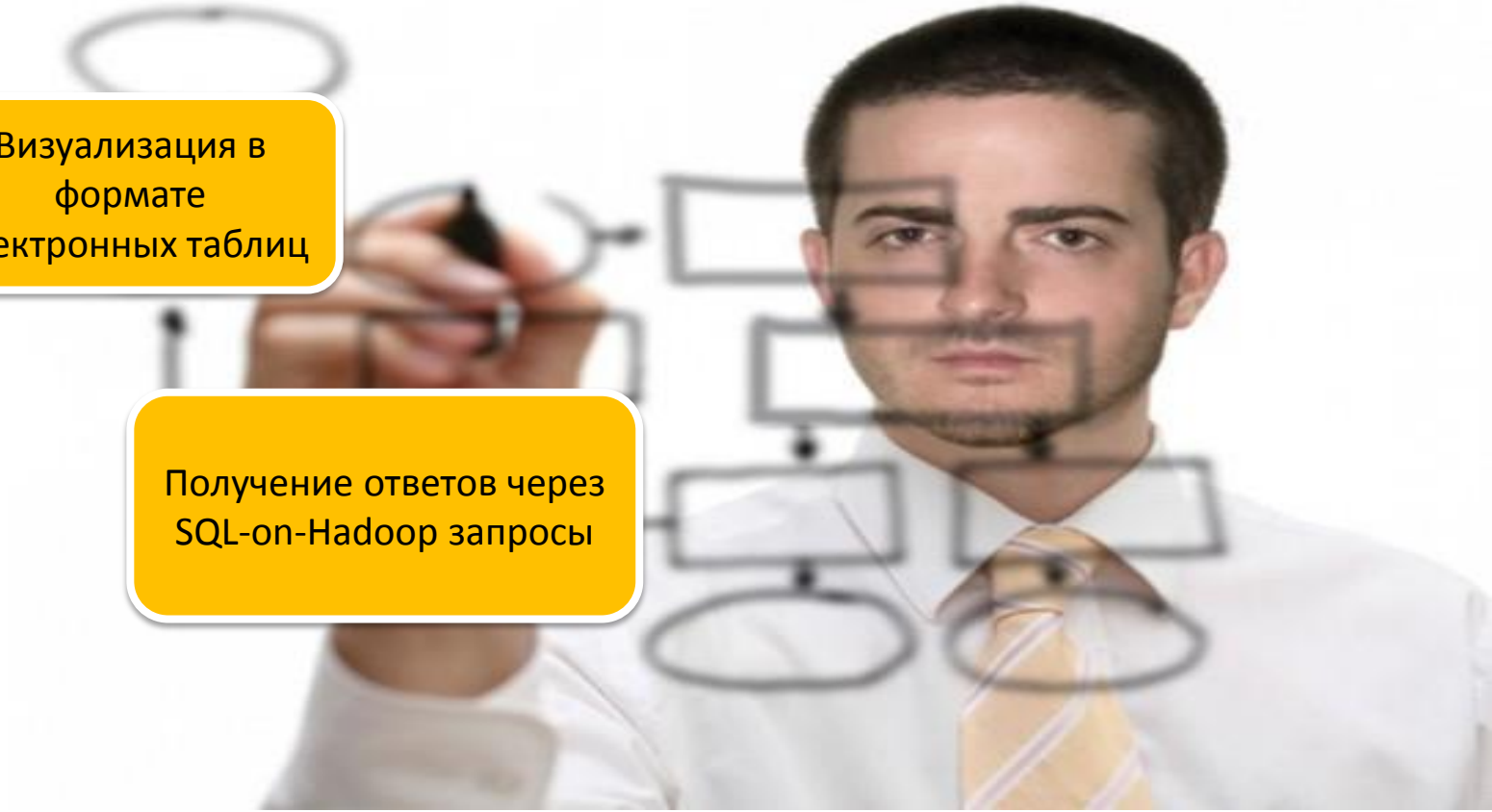
- Только IBM сейчас может использовать всю память кластера
- Только IBM сейчас может запускать тысячи моделей параллельно

Инструменты текстовой аналитики



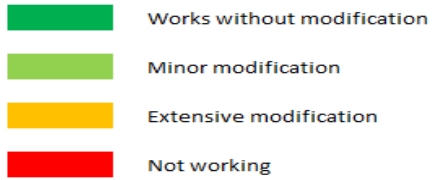
Дать бизнес аналитику знакомые инструменты

IBM BigInsights Analyst module



Отличия IBM BigInsights: BigSQL

IBM Big SQL – 100% TPC запросов



IBM

Query 01	Query 34	Query 67
Query 02	Query 35	Query 68
Query 03	Query 36	Query 69
Query 04	Query 37	Query 70
Query 05	Query 38	Query 71
Query 06	Query 39	Query 72
Query 07	Query 40	Query 73
Query 08	Query 41	Query 74
Query 09	Query 42	Query 75
Query 10	Query 43	Query 76
Query 11	Query 44	Query 77
Query 12	Query 45	Query 78
Query 13	Query 46	Query 79
Query 14	Query 47	Query 80
Query 15	Query 48	Query 81
Query 16	Query 49	Query 82
Query 17	Query 50	Query 83
Query 18	Query 51	Query 84
Query 19	Query 52	Query 85
Query 20	Query 53	Query 86
Query 21	Query 54	Query 87
Query 22	Query 55	Query 88
Query 23	Query 56	Query 89
Query 24	Query 57	Query 90
Query 25	Query 58	Query 91
Query 26	Query 59	Query 92
Query 27	Query 60	Query 93
Query 28	Query 61	Query 94
Query 29	Query 62	Query 95
Query 30	Query 63	Query 96
Query 31	Query 64	Query 97
Query 32	Query 65	Query 98
Query 33	Query 66	Query 99

cloudera IMPALA

Query 01	Query 34	Query 67
Query 02	Query 35	Query 68
Query 03	Query 36	Query 69
Query 04	Query 37	Query 70
Query 05	Query 38	Query 71
Query 06	Query 39	Query 72
Query 07	Query 40	Query 73
Query 08	Query 41	Query 74
Query 09	Query 42	Query 75
Query 10	Query 43	Query 76
Query 11	Query 44	Query 77
Query 12	Query 45	Query 78
Query 13	Query 46	Query 79
Query 14	Query 47	Query 80
Query 15	Query 48	Query 81
Query 16	Query 49	Query 82
Query 17	Query 50	Query 83
Query 18	Query 51	Query 84
Query 19	Query 52	Query 85
Query 20	Query 53	Query 86
Query 21	Query 54	Query 87
Query 22	Query 55	Query 88
Query 23	Query 56	Query 89
Query 24	Query 57	Query 90
Query 25	Query 58	Query 91
Query 26	Query 59	Query 92
Query 27	Query 60	Query 93
Query 28	Query 61	Query 94
Query 29	Query 62	Query 95
Query 30	Query 63	Query 96
Query 31	Query 64	Query 97
Query 32	Query 65	Query 98
Query 33	Query 66	Query 99

HIVE

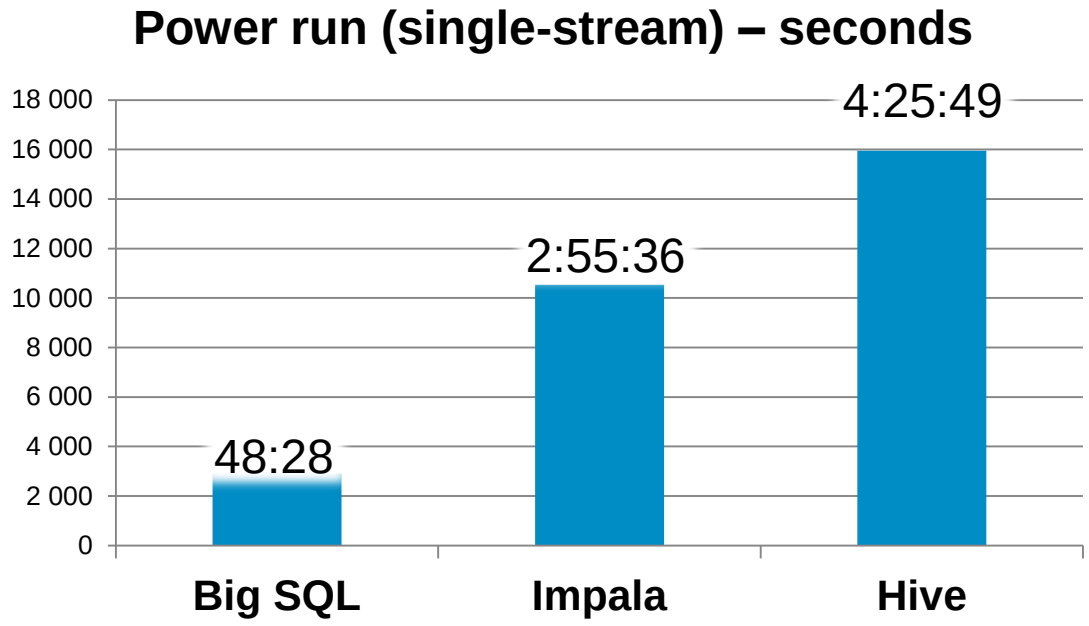
Query 01	Query 34	Query 67
Query 02	Query 35	Query 68
Query 03	Query 36	Query 69
Query 04	Query 37	Query 70
Query 05	Query 38	Query 71
Query 06	Query 39	Query 72
Query 07	Query 40	Query 73
Query 08	Query 41	Query 74
Query 09	Query 42	Query 75
Query 10	Query 43	Query 76
Query 11	Query 44	Query 77
Query 12	Query 45	Query 78
Query 13	Query 46	Query 79
Query 14	Query 47	Query 80
Query 15	Query 48	Query 81
Query 16	Query 49	Query 82
Query 17	Query 50	Query 83
Query 18	Query 51	Query 84
Query 19	Query 52	Query 85
Query 20	Query 53	Query 86
Query 21	Query 54	Query 87
Query 22	Query 55	Query 88
Query 23	Query 56	Query 89
Query 24	Query 57	Query 90
Query 25	Query 58	Query 91
Query 26	Query 59	Query 92
Query 27	Query 60	Query 93
Query 28	Query 61	Query 94
Query 29	Query 62	Query 95
Query 30	Query 63	Query 96
Query 31	Query 64	Query 97
Query 32	Query 65	Query 98
Query 33	Query 66	Query 99

Results for 10TB scale shown here

Ключевые моменты

- В конкурентных средах много запросов нужно переписать, некоторые значительно
- Из-за различных ограничений некоторые запросы не могут быть переписаны или не выполнились
- Переписывание запросов в текстах где результат известен заранее и реальная система – разные вещи

Отличия IBM BigInsights: производительность BigSQL

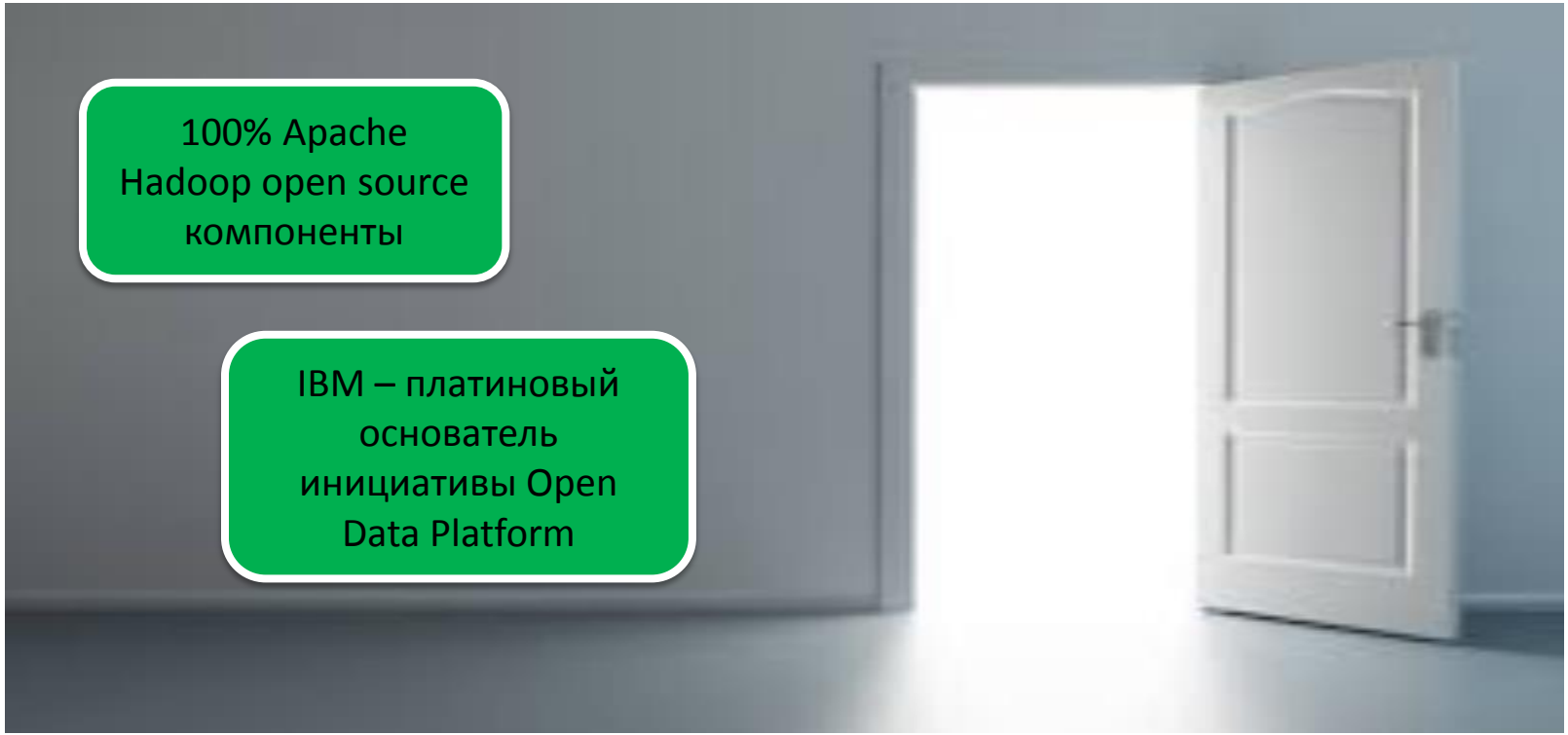


- Big SQL существенно быстрее Impala и Hive – **3.6x раза быстрее** Impala в подтвержденных результатах тестов
- Big SQL – единственное решение, выполнившее все запросы на тестах 10TB и 30TB

Letter of attestation available from InfoSizing, TPC certified auditors

Поддержка открытых стандартов и переносимость

IBM Open Platform with Apache Hadoop



Поддержка Open Source

IBM теперь поддерживает все последние версии компонентов платформы Hadoop

Component Name	Version
Ambari	1.7.0
Avro	1.7.7
Flume	1.5.2
Hadoop	2.6
HBase	0.98.8
Hive	0.14.0
Knox	0.5.0
Oozie	4.0.1
Pig	0.14.0
Parquet (hadoop)	1.5.0
Parquet (format)	2.1.0
Spark	1.2.1
Snappy	1.0.5
Sqoop	1.4.5
Solr	4.10.3
Slider	0.6.0
Zookeeper	3.4.5

Мы будем поддерживать актуальность в <u>каждом</u> новом релизе

IBM теперь использует установщик Apache Ambari – часть Open Data Platform Initiative

Не нужно больше больших загрузок образов
Загрузка небольшого пакета и последующая загрузка только необходимых компонентов

Необходима новая архитектура работы с данными

BusinessConnect
соединяя бизнес и технологии

Спасибо!