

Большие данные: новые возможности и конкурентное преимущество для бизнеса

Сергей Лихарев
IBM Analytics

Аналитика помогает решить ключевые бизнес задачи предприятия



Большие данные – все данные

Объем



Масштаб

Разнообразие



Много форм

Скорость



Потоки данных

Достоверность



Доверие

Необходима новая архитектура работы с данными

Все данные

Новые/расширенные приложения



IBM Watson Analytics – инновации в аналитике

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



IBM Watson Analytics

Быстрый старт
Понятный интерфейс

Диалог с
системой

Исследование
данных

Доступно для
мобильных
устройств

The screenshot shows the IBM Watson Analytics interface. At the top, there's a navigation bar with 'WELCOME' and a user profile for 'Georgia Henriot'. Below this, there are tabs for 'Getting Started', 'Add Data', 'Recent Workbooks', and 'Open Workbook'. A search bar is present with the text 'Enter a keyword to filter the list below, or to ask Watson a question about your data!'. The main content area is divided into two columns: 'Start from Data' and 'Start from a Story'. 'Start from Data' includes cards for 'EXPLORE YOUR DATA', 'PREDICT AND EXPLAIN', and 'FORECAST FUTURE VALUES'. 'Start from a Story' includes cards for 'GETTING STARTED WITH WATSON ANALYTICS', 'IMPROVE CAMPAIGN EFFECTIVENESS', 'RETAIN YOUR TEAM', 'WORKING WITH DATA', 'PREVENTING EMPLOYEE ATTRITION', 'FIND PATTERNS IN WINS AND LOSSES', 'SAMPLE TEXT', 'CUSTOMER PROFITABILITY', and 'NEXT BEST OFFER FOR EXISTING CUSTOMERS'. Each card features a small icon and a brief description. Callouts in blue speech bubbles point to various parts of the interface: 'Быстрый старт Понятный интерфейс' points to the top navigation; 'Диалог с системой' points to the search bar; 'Исследование данных' points to the 'EXPLORE YOUR DATA' card; 'Доступно для мобильных устройств' points to the 'FIND PATTERNS IN WINS AND LOSSES' card; and 'Гибкость облачной среды' points to the bottom of the dashboard.

Гибкость облачной среды

IBM Watson Analytics

Доступ и
очистка данных

Взаимодействие



Интеллект без
настройки

Подсказки в
исследовании

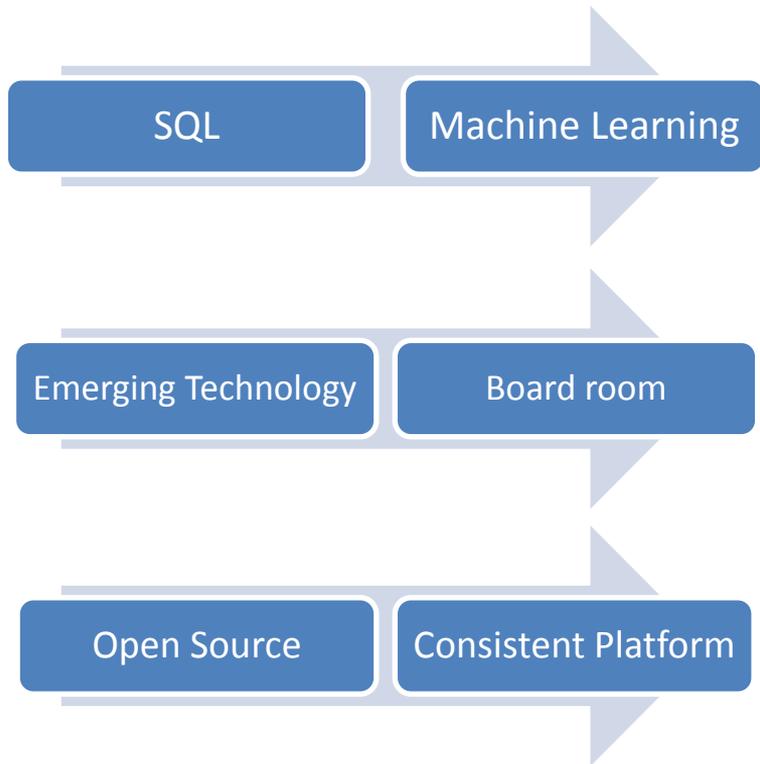
Отчеты и
информационные
панели

Связывание
элементов в
историю

Использование всех доступных форм данных



Рынок Hadoop развивается



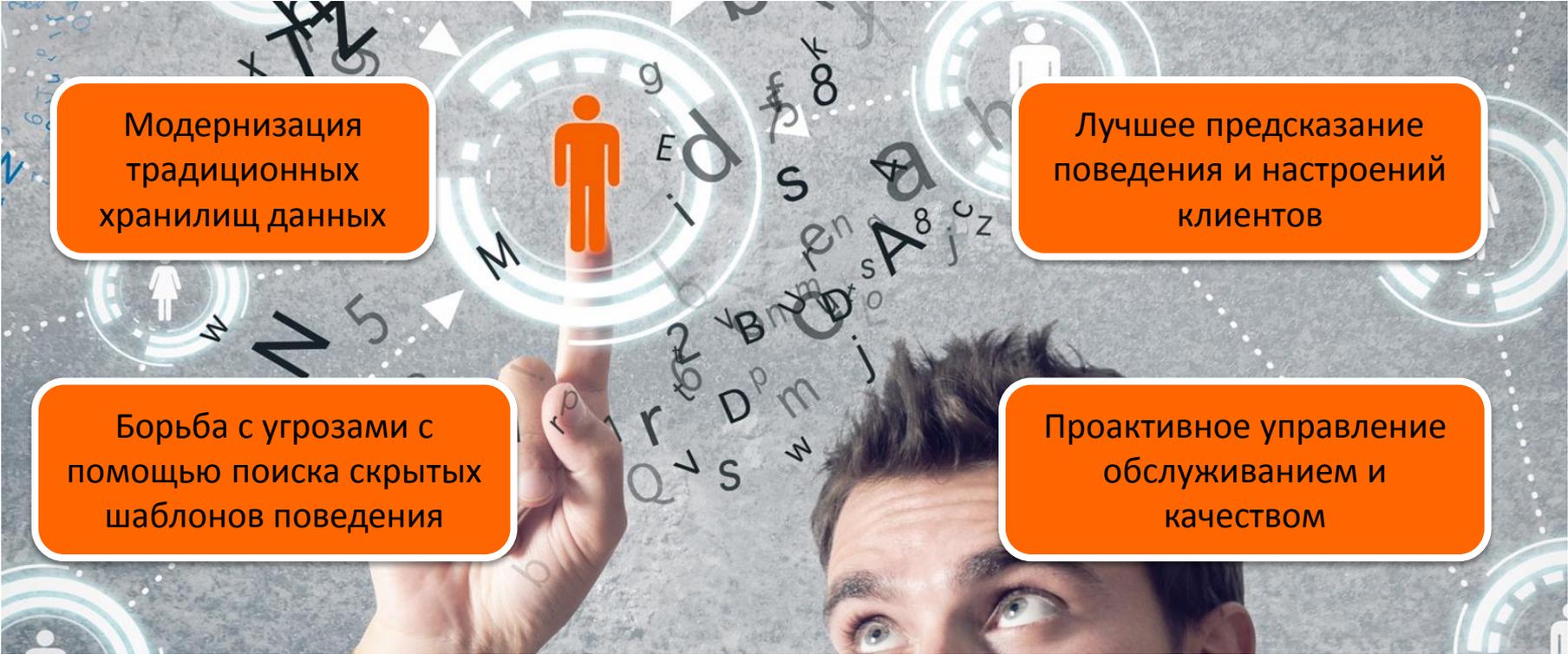
От доступ к данным для разработчика к исследованию данных для data scientist

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



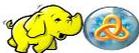
Модернизация
традиционных
хранилищ данных

Лучшее предсказание
поведения и настроений
клиентов

Борьба с угрозами с
помощью поиска скрытых
шаблонов поведения

Проактивное управление
обслуживанием и
качеством

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

- Выявление данных для анализа
- Визуализация данных для действий
- Использование существующих навыков (SQL, spreadsheets)



Data Scientist

- Выявление шаблонов, трендов, результаты алгоритмов машинного обучения
- Статистические модели на больших объемах данных



Administrator

- Управление нагрузкой и обеспечение уровня производительности
- Реализация политик безопасности для снижения рисков

Потребность

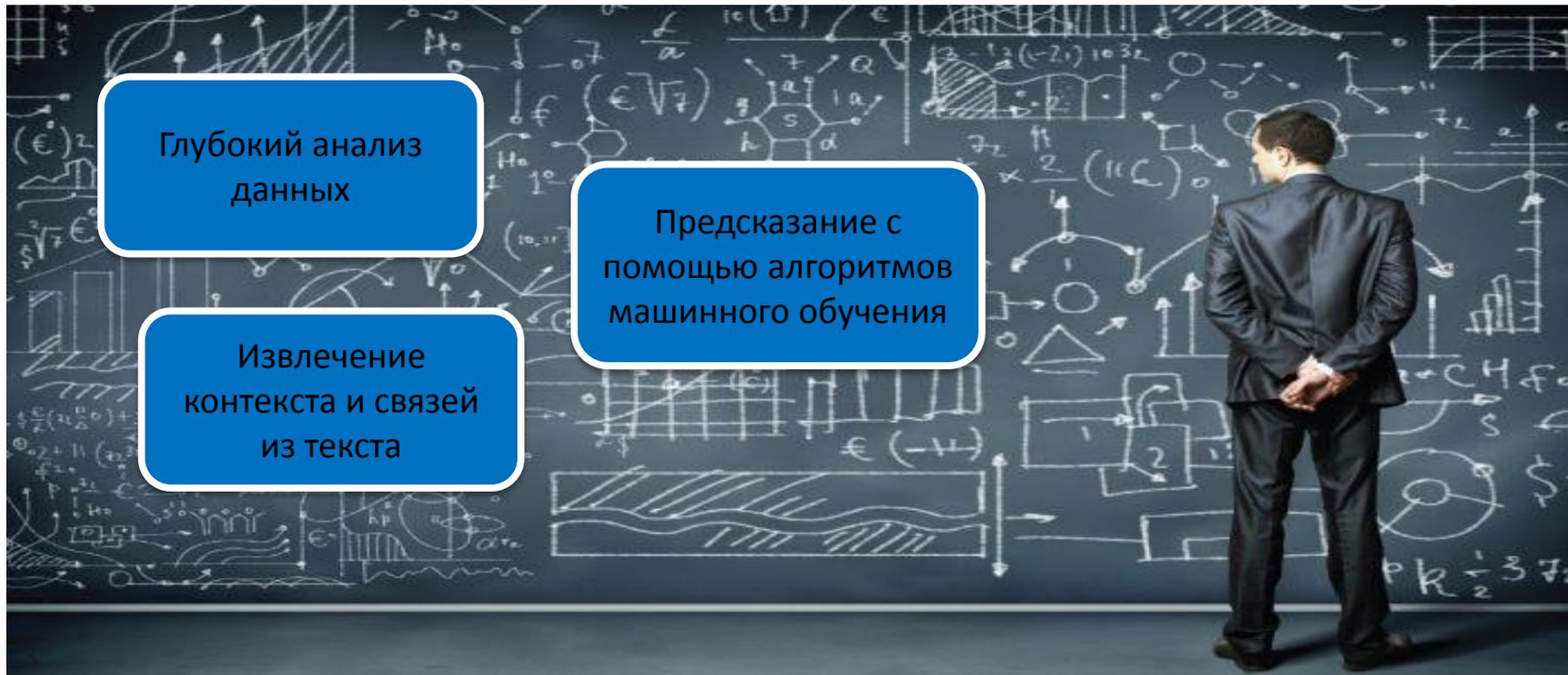
Data science на Hadoop

IBM BigInsights Data Scientist module

Глубокий анализ
данных

Извлечение
контекста и связей
из текста

Предсказание с
помощью алгоритмов
машинного обучения



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

The screenshot displays a web-based text analysis tool interface. On the left, there is a 'Projects' sidebar with a 'Catalog' section containing a search bar and a list of projects: Private (biadmin), Suspect IP, Revenue, Revenue by Division 1, Group A, Public, Finance, Log Analysis, and Machine Data Accelerator. Below the catalog is a 'Properties' section with fields for Title (IP Address), Tags (IP, Address), Description (A numerical label of a device within a computer network), Supported Languages, and Category (Syslog Adapter).

The main area is titled 'Nov 2013 Security Syslog'. A yellow highlight box labeled 'Suspect IP' is positioned over a search bar. Below the search bar, a table of log entries is shown with columns: DateTime, Mnemonic, ACL, and IP Address. The entries are as follows:

DateTime	Mnemonic	ACL	IP Address
Aug 24 2007 10:27:31	%ASA-6-106100	OUTSIDE	192.168.208.63
Aug 24 2007 10:27:31	%ASA-6-106100	OUTSIDE	192.168.208.63
Aug 24 2007 10:27:29	%ASA-6-106100	OUTSIDE	192.168.208.63
Aug 24 2007 10:27:31	%ASA-6-106100	OUTSIDE	192.168.208.63
Aug 24 2007 11:15:39	%ASA-6-106100	OUTSIDE	192.168.208.63
Aug 24 2007 11:15:40	%ASA-6-106100	OUTSIDE	192.168.208.63
Aug 24 2007 11:23:11	%ASA-6-106100	OUTSIDE	192.168.208.6

On the right side, there is a 'Documents' section with a search bar and a list of documents. The documents are:

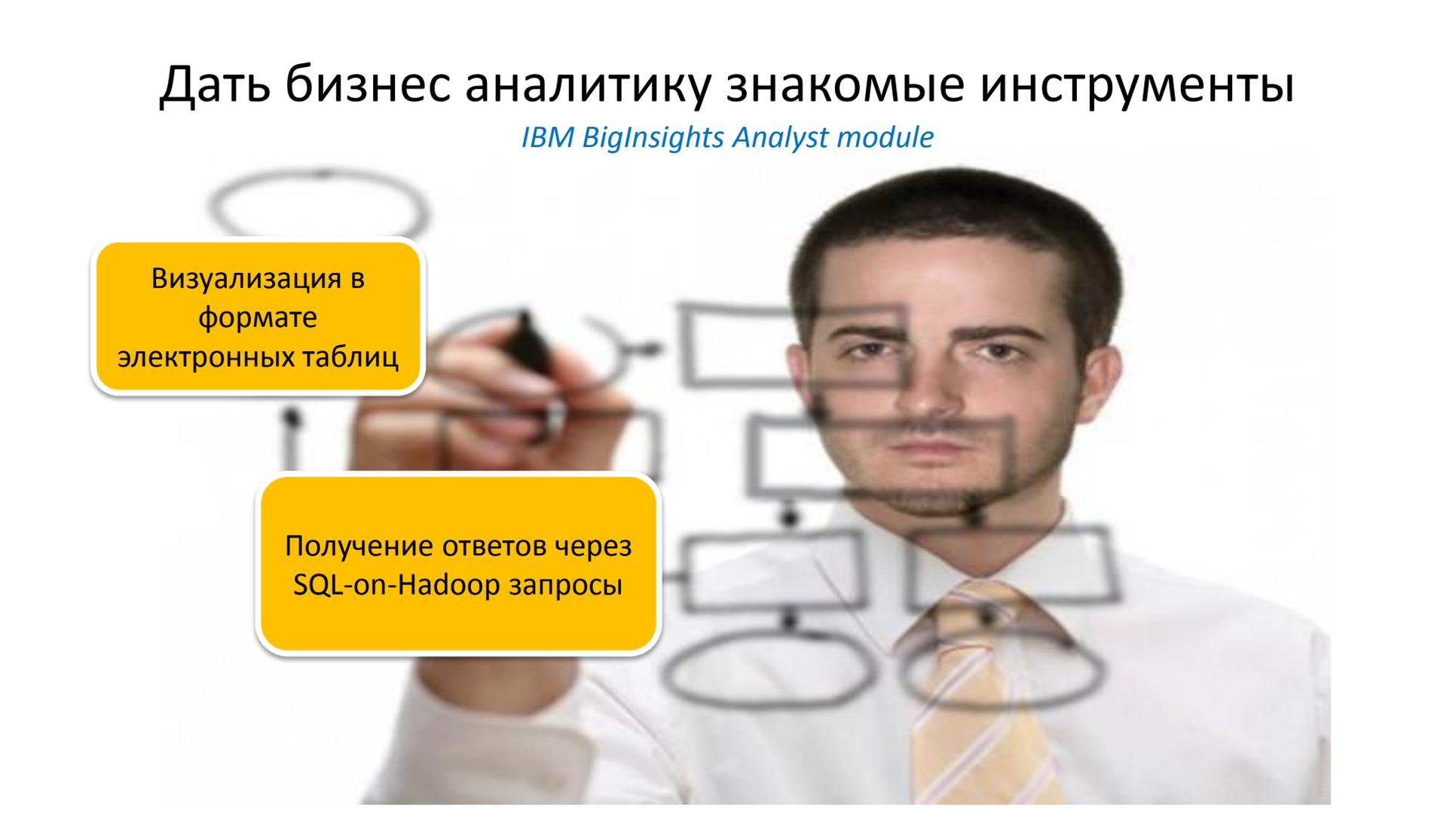
- File1.bt: Aug 24 2007 10:27:29: %ASA-6-106100: access-list OUTSIDE denied tcp outside/192.168.208.63(39675)-> inside/192.168.150.77(80) hit-cnt 1 first hit [0x22e8ac21, 0x0]
- File2.bt: Aug 24 2007 10:27:31: %ASA-6-106100: access-list OUTSIDE denied tcp outside/192.168.208.63(39676)-> inside/192.168.150.77(80) hit-cnt 1 first hit [0x22e8ac21, 0x0]
- File3.bt: Aug 24 2007 10:27:22: %ASA-4-400014: IDS:2004 ICMP echo request from 192.168.208.63/39676 to 192.168.150.70(80) on interface outside
- File4.bt: Aug 24 2007 10:27:22: %ASA-6-302020: Built ICMP connection for faddr 192.168.208.63/15343 gaddr 192.168.150.70/0 laddr 192.168.150.70/0
- File5.bt: Aug 24 2007 10:27:22: %ASA-6-106015: Deny TCP (no connection) from 192.168.208.63/49827 to 192.168.150.70/80 flags ACK on interface outside
- File6.bt: Aug 24 2007 10:27:22: %ASA-6-302020: Built ICMP connection for faddr 192.168.208.63/15343 gaddr 192.168.150.70/0 laddr 192.168.150.70/0
- File7.bt: Aug 24 2007 10:27:22: %ASA-6-302015: Built inbound UDP connection 732748 for outside:192.168.208.63/49804 to inside:192.168.150.70/53

At the bottom right, there is a pagination control showing '1 ... 7 8 9 ... 20' and '10 25 50 ALL'.

Web инструменты для определения правил извлечения данных и выделения информации из текста
Графический интерфейс для описания различных текстовых форматов – от лог файлов до естественного языка

Дать бизнес аналитику знакомые инструменты

IBM BigInsights Analyst module



Визуализация в
формате
электронных таблиц

Получение ответов через
SQL-on-Hadoop запросы

Отличия IBM BigInsights: BigSQL

IBM Big SQL – 100% TPC запросов



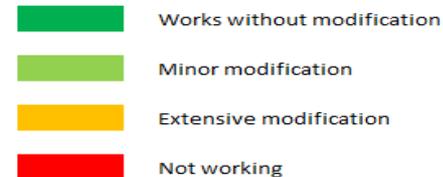
Query 01	Query 34	Query 67
Query 02	Query 35	Query 68
Query 03	Query 36	Query 69
Query 04	Query 37	Query 70
Query 05	Query 38	Query 71
Query 06	Query 39	Query 72
Query 07	Query 40	Query 73
Query 08	Query 41	Query 74
Query 09	Query 42	Query 75
Query 10	Query 43	Query 76
Query 11	Query 44	Query 77
Query 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



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



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

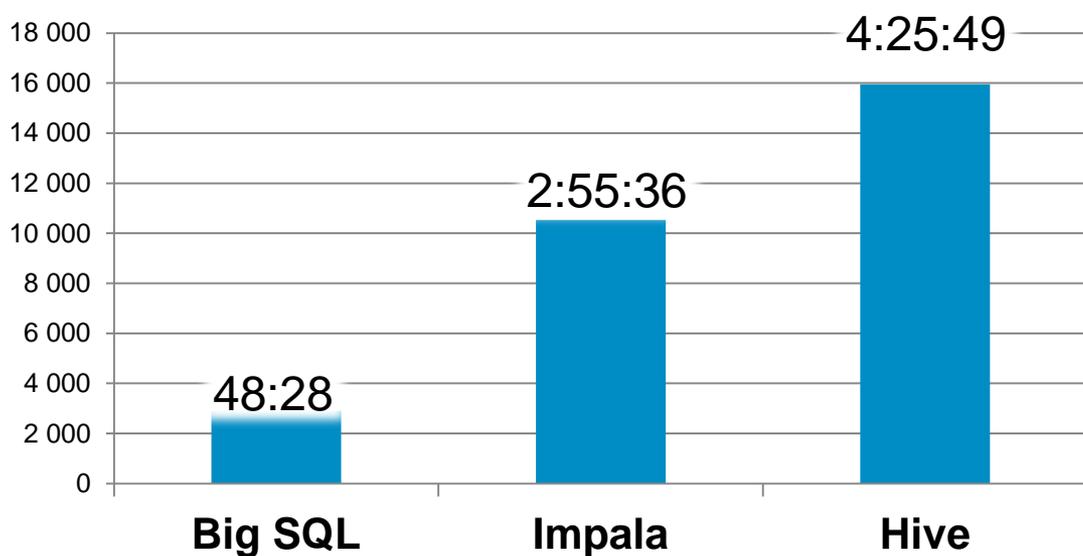


Ключевые моменты

- 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

Power run (single-stream) – seconds



- 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

Поддержка открытых стандартов и переносимость

IBM Open Platform with Apache Hadoop

100% Apache
Hadoop open source
компоненты

IBM – платиновый
основатель
инициативы Open
Data Platform



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

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

Мы будем поддерживать актуальность в каждом новом релизе

Необходима новая архитектура работы с данными

Все данные

Новые/расширенные приложения



BusinessConnect
соединяя бизнес и технологии

Спасибо!

IBM.

