Как повысить ценность данных и сократить затраты на работу с ними

Михаил Комаров
Директор по развитию бизнеса,
Решения Informatica

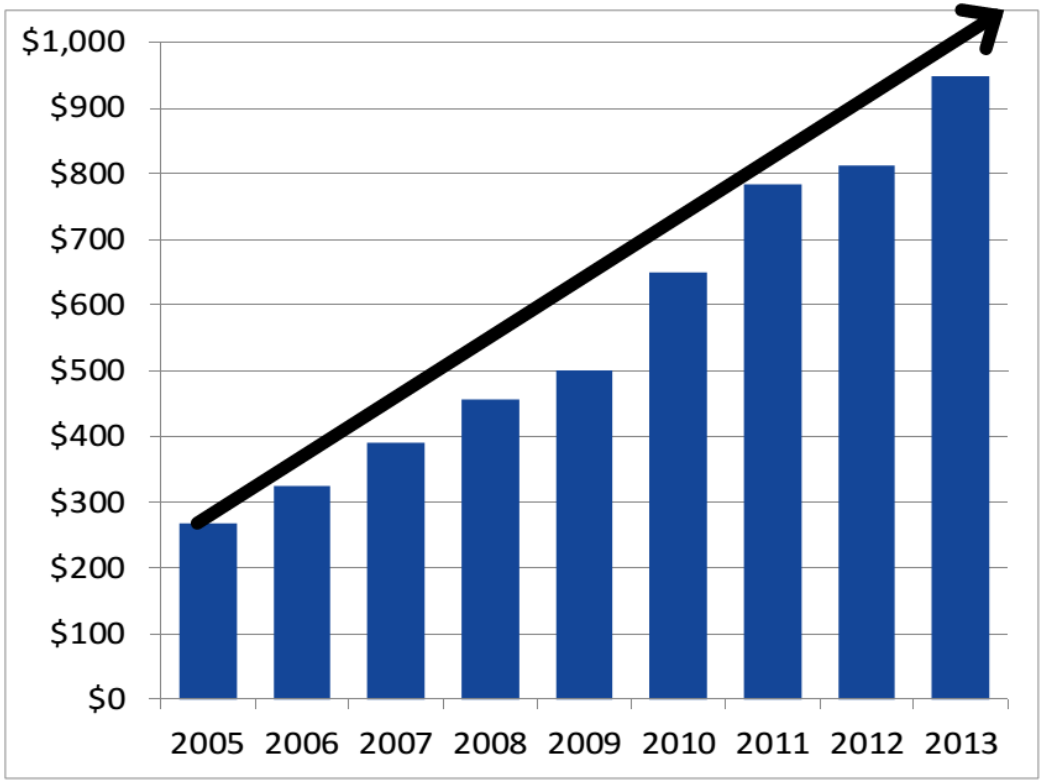
DIS Group

О компании Informatica
#1 Независимый лидер рынка интеграции и обеспечения качества данных

- **Основана** 1993 год
- **Выручка 2014** $1+ млрд.
- **Среднегодовой рост за последние 5 лет** 18% per year
- **Заказчики** 6,000+
  - 80 of Fortune 100
  - 90%+ of Dow Jones
  - 10 из 10 крупнейших банков
- **Сотрудники** 2100+

Эксклюзивный дистрибьютор в России и странах СНГ

DIS Group



#1 Gartner Magic Quadrant Data Integration Tools
#1 Gartner Magic Quadrant Data Quality Tools
#1 Customer loyalty 8 лет подряд
Лидер в кавдрантах Data Masking, MDM

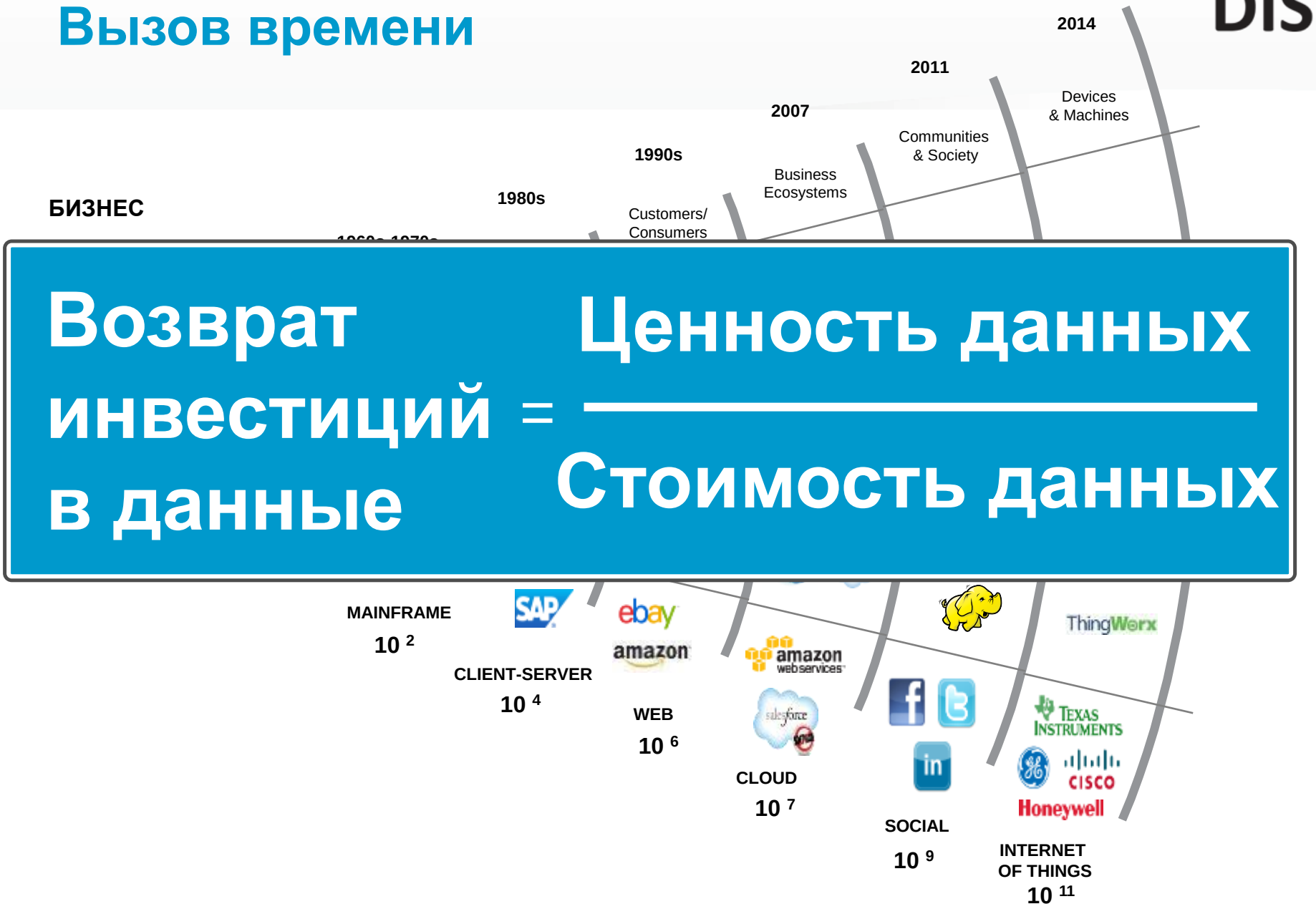
Заказчики DIS Group из ТОП 20 Банков

1	SBERBANK By your side	DIS	11	Promsvyazbank	DIS
2	VTB World Without Barriers. VTB Group	DIS	12	ROSBANK	DIS
3	GAZPROMBANK		13	Raiffeisen BANK	DIS
4	ВТБ24 Большое преимущество	DIS	14	открытие ХАНТЫ-МАНСИЙСКИЙ БАНК	
5	OTKRITIE FINANCIAL CORPORATION	DIS	15	CREDIT BANK OF MOSCOW	DIS
6	Alfa-Bank	DIS	16	БАНК РОССИЯ Банк умных решений	
7	Russian Agricultural Bank	DIS	17	BANK SAINT PETERSBURG	DIS
8	Bank of Moscow	DIS	18	AK BARS BANK	
9	NATIONAL CLEARING CENTRE MOSCOW EXCHANGE GROUP		19	B&NBANK	DIS
10	UniCredit Bank	DIS	20	RUSSIAN STANDARD BANK	DIS

Глобальные заказчики Informatica в финансовом секторе

- **31** из ТОП **35** банков по всему миру
- **37** из ТОП **50** Банков США
- **20** из ТОП **27** Банков Европы
- **9** из ТОП **10** глобальных инвестиционных банков

Вызов времени



Сокращение затрат на интеграцию данных

Что дает Informatica в интеграции данных?

DIS

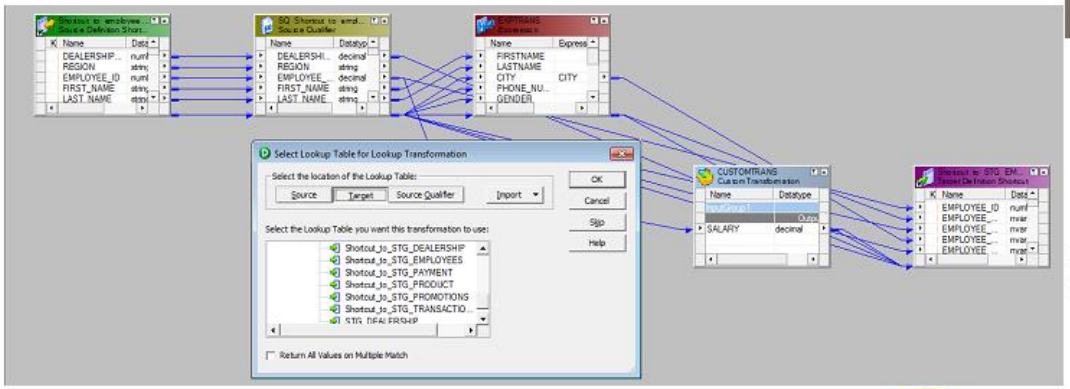
Пять аргументов за использование ETL-машины

СБЕРБАНК

Простота внесения изменений

Стиль разработки: как влияет инструмент?

Если не контролировать себя,
то обычно получается так…



…а вот в карте
Informatica PowerCenter
добавить lookup проще,
чем ветвление

Доступность данных

Уровень доступности

По результатам работы за 2010 год уровень доступности данных, передаваемых Informatica

Уровень доступности = 100%

Производительность и масштабируемость

DIS

Масштаб Informatica

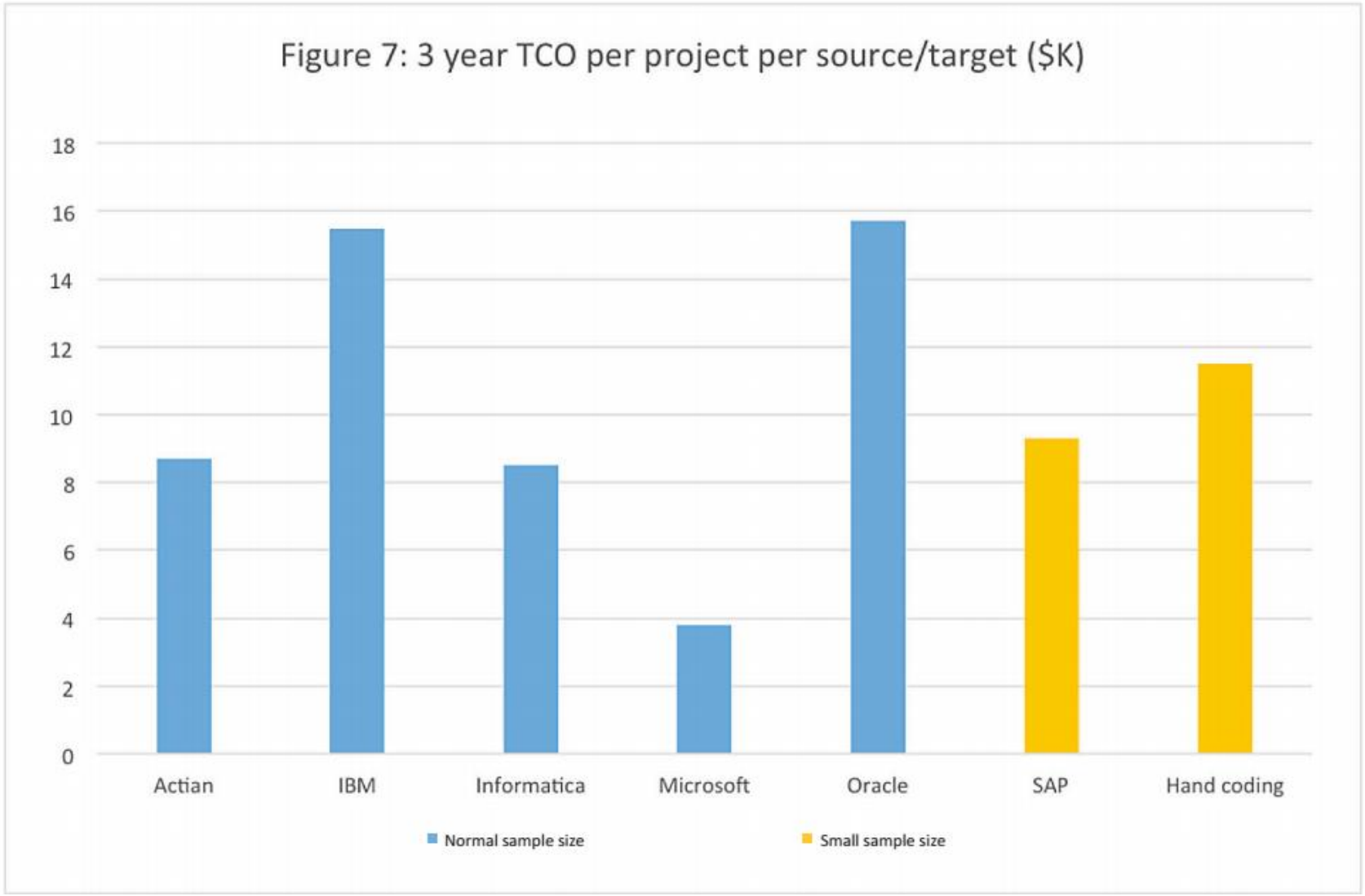
Beeline

Количество источников	61
Типы источников	Oracle, MS SQL, DB2, Informix, Text files, XML files
Ежедневный объем загружаемых данных	0,5ТБ
Количество карт загрузки	4,5 тыс.
Количество таблиц приемников	1 000+

6,7 миллиардов

ежедневно загружаемых записей

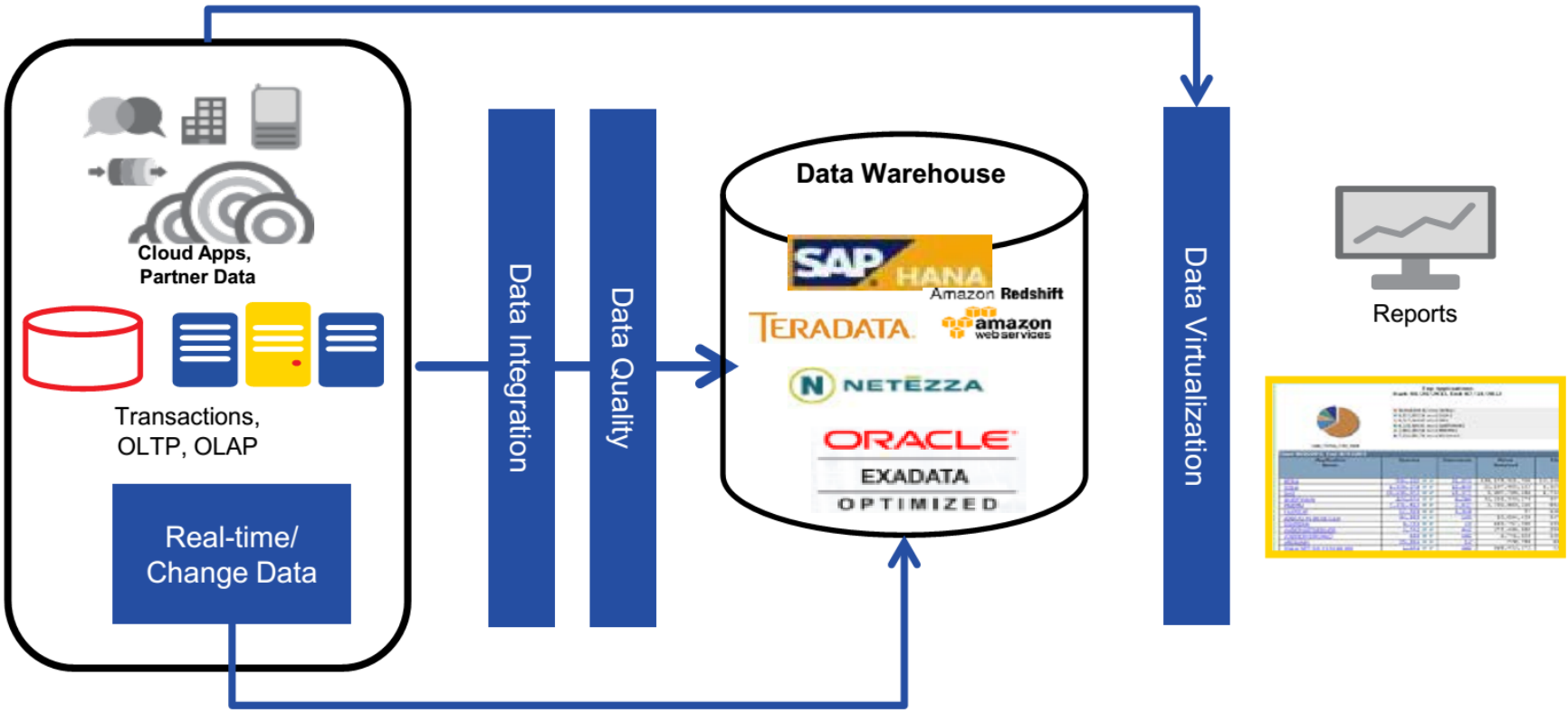
Самое низкое TCO среди лидеров Gartner



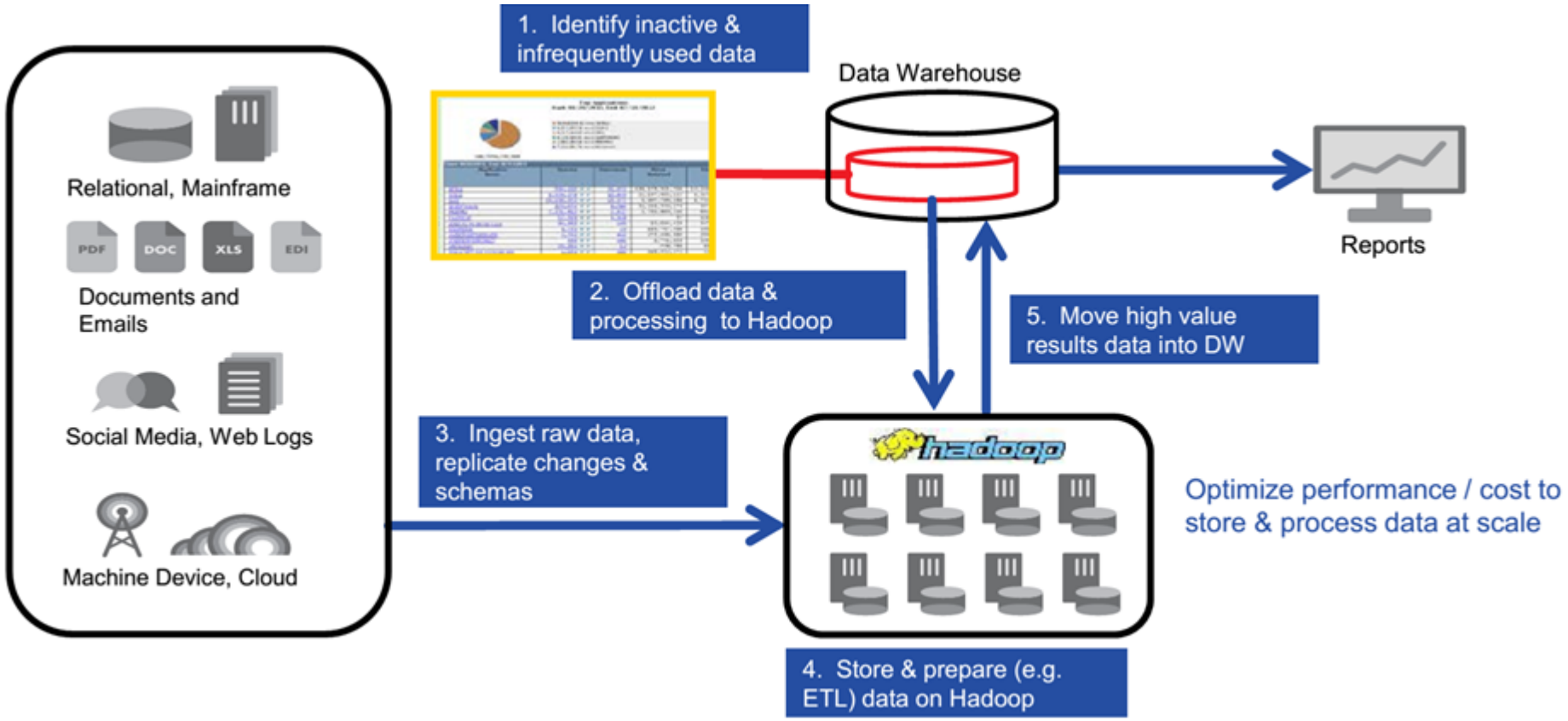
Comparative costs and uses for data integration platforms, Bloor 2014


DIS
Big Data?

Традиционное хранилище данных



Next Generation Analytics



Какие препятствия для работы с Big Data?

- Нет квалифицированных ресурсов на рынке
- Open source решения нет поддержки
- Сложная разработка процессов
- Сложная поддержка решения
- Бурный рост версий

Data Lake – Informatica ETL

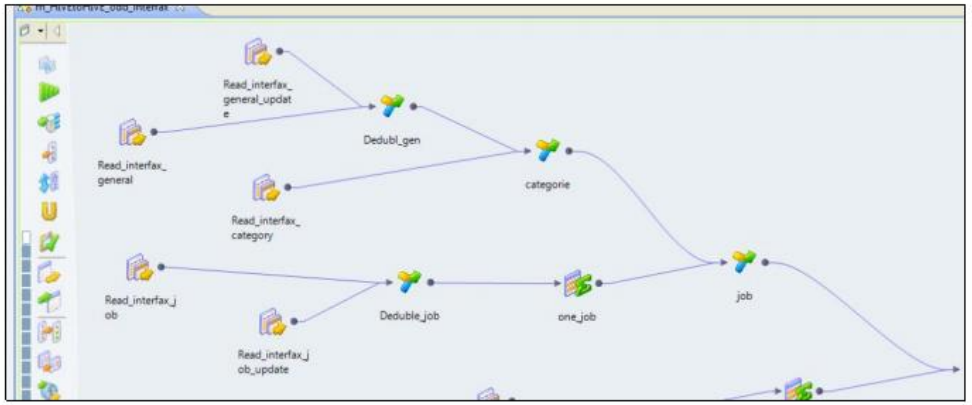
Hadoop

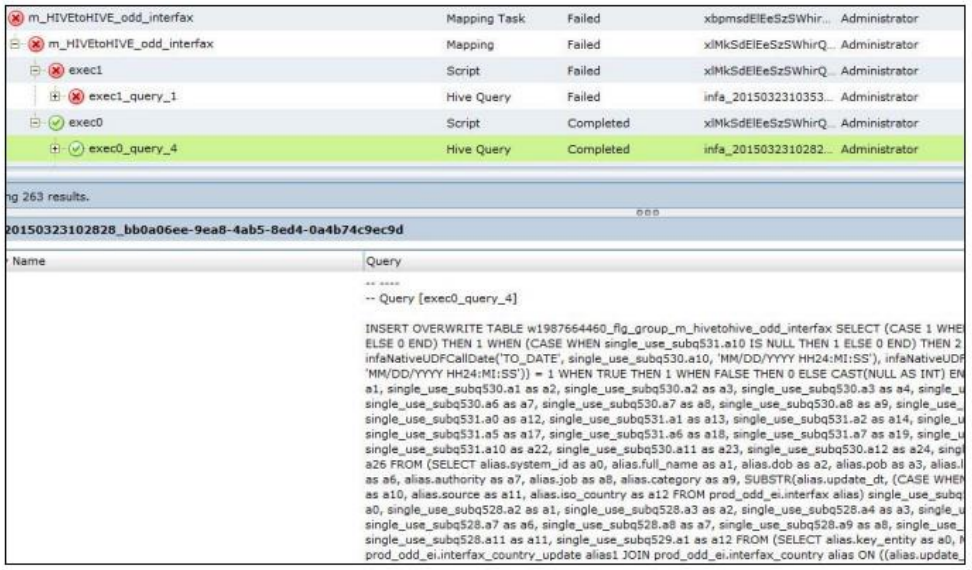
informatica

Hive

Map Reduce

Данные в HDFS





http://habrahabr.ru/company/tcsbank/blog/259173/

Задача 2008-2014 - что получили

Customer 180°

Понимаем, что клиент делает только с точки зрения нашего бизнеса

- Какие счета открывает
- Какие платежи делает
- Вовремя ли платит
- …

Задача 2014-… Что получим

Customer 360°

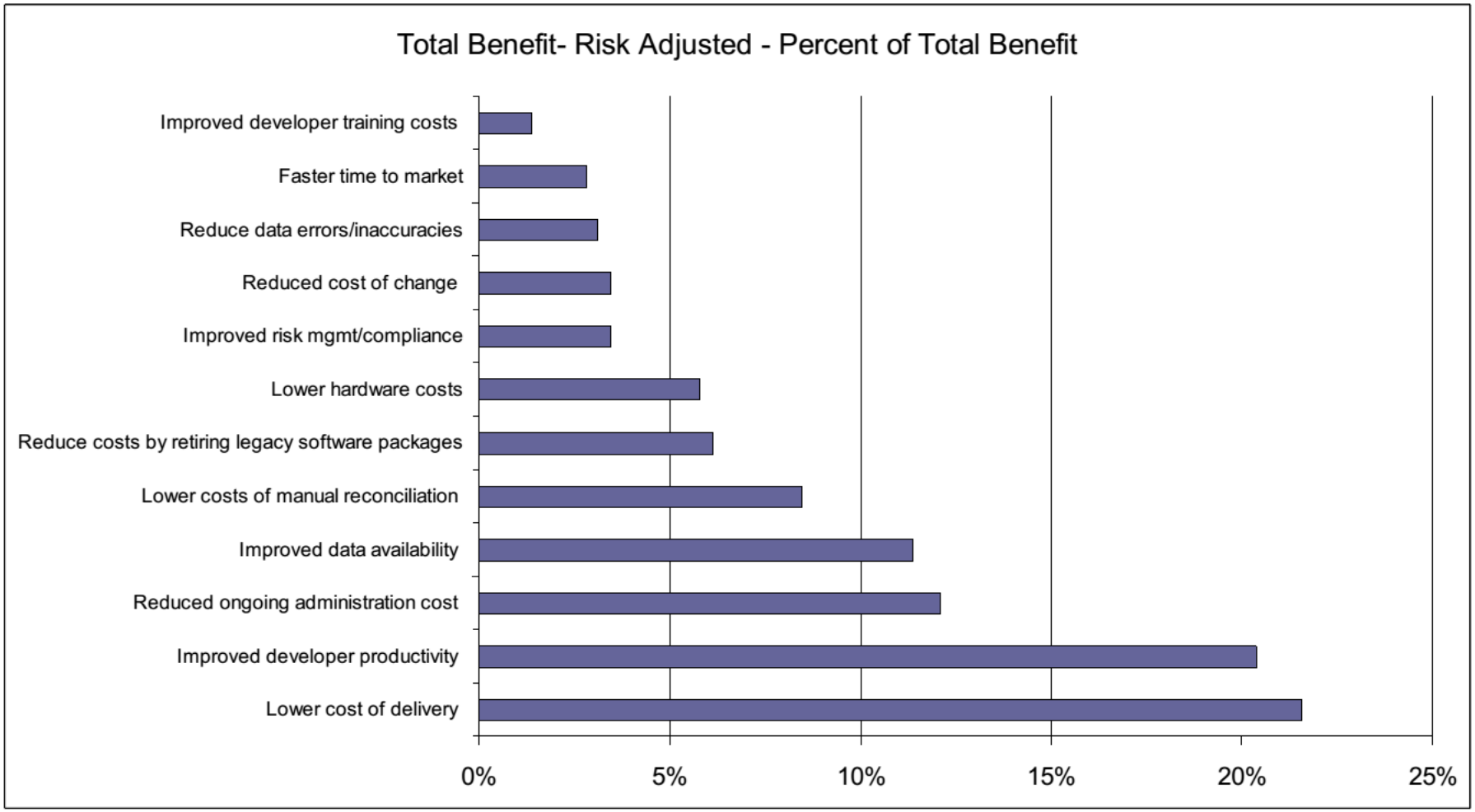
Понимаем, что происходит с клиентом

Все то же что и раньше, но плюс:

- Какие страницы открывает на нашем сайте
- Как пользуется мобильным приложением
- Как ведет себя в интернете
- …

Сокращение затрат на интеграцию данных

Figure 5 – Total Investment Benefit

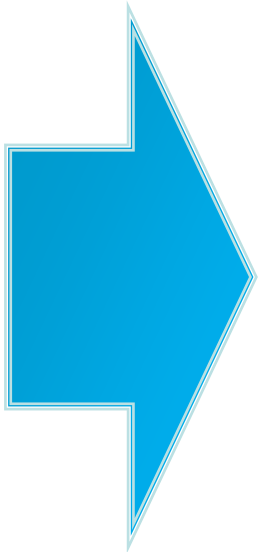


Source: Forrester Research, Inc.

Сокращение потерь из-за низкого качества данных

К чему приводят проблемы с качеством данных?

- Низкое качество данных
- Устаревшая адресная информация
- Устаревшие телефонные номера
- Дублирование клиентских данных
- Противоречивость данных



- Ошибки в отчетности
- Низкая эффективность маркетинговых компаний при высоких издержках
- Снижение эффективности коммуникаций с клиентом
- Затруднение работы службы коллекторов
- Финансовые издержки и упущенная выгода

Показатели качества данных из положения ЦБ РФ для расчета величины кредитного риска

Для каждого фактора, параметра и результата модели количественной оценки рисков (далее – компоненты):

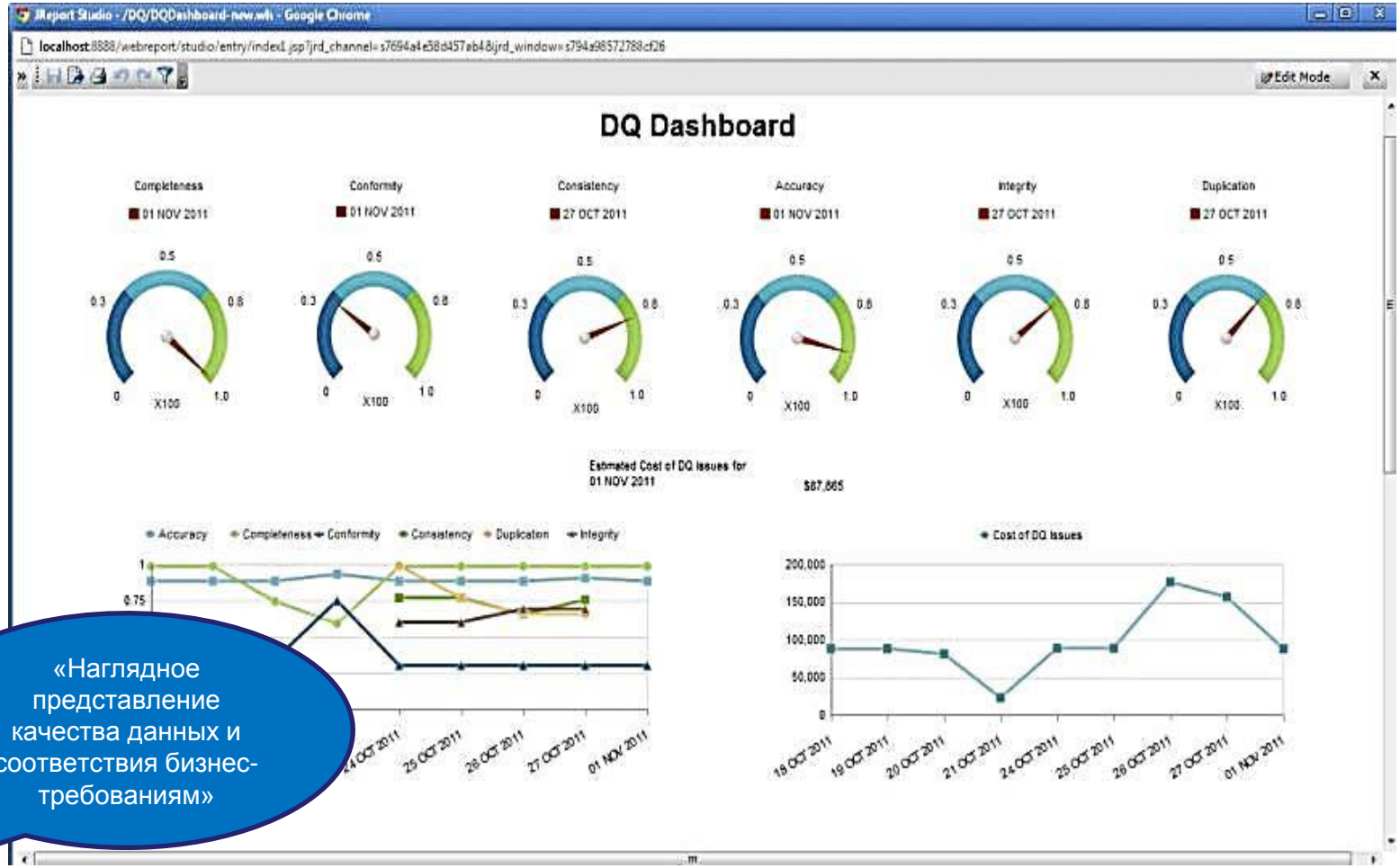
- доля записей с пропущенными (пустыми) значениями первичных данных, использованных впоследствии для расчетов значений компонент;
- доля записей с недостоверными (не соответствующими действительности) значениями первичных данных, использованных впоследствии для вычисления значений компонент;
- доля записей с аномальными (выходящими за рамки допустимого диапазона) значениями первичных данных, использованных впоследствии для вычисления значений компонент;
- доля записей с несогласующимися (противоречащими другим данным) значениями первичных данных, использованных впоследствии для вычисления значений компонент;
- доля записей с избыточными (например, полученными в результате дублирования) значениями первичных данных, использованных впоследствии для вычисления значений компонент;
- Другое

Характеристики качества данных из положения ЦБ РФ для расчета величины кредитного риска

- **точность и достоверность данных** – отсутствие синтаксических и семантических ошибок в значениях свойств и характеристик объектов (субъектов) учета, параметров моделей, используемых в моделях количественной оценки рисков, а также их соответствие реальным значениям указанных свойств, характеристик и параметров;

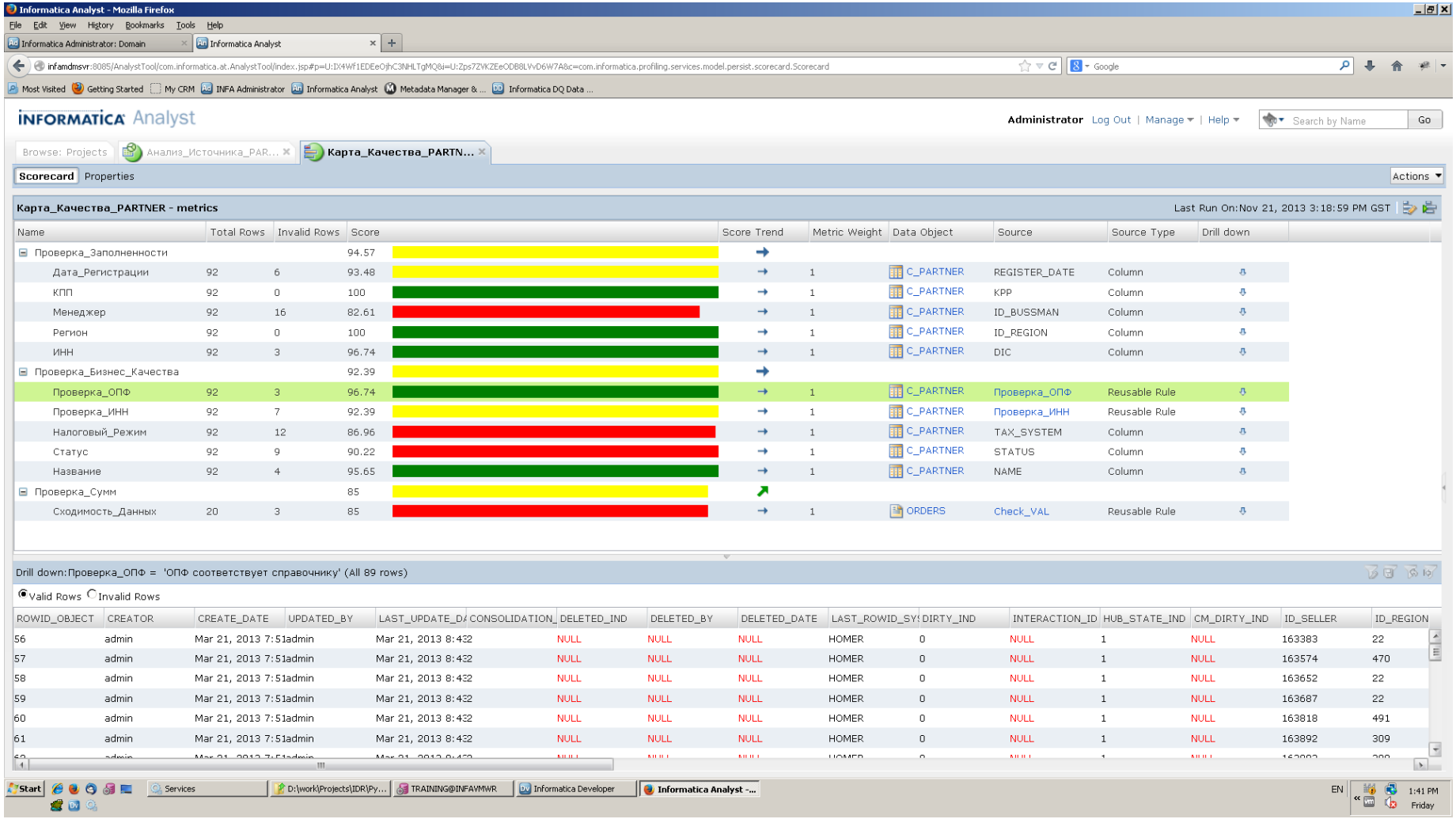
- **контролируемость данных** – возможность осуществления контроля качества и происхождения данных, в том числе посредством отражения в ИС источников данных, истории создания, изменения, преобразования, удаления, хранения и передачи данных;

Мониторинг качества для бизнес-пользователей



«Наглядное представление качества данных и соответствия бизнес-требованиям»

Профиль данных в любых системах



Готовые правила для России

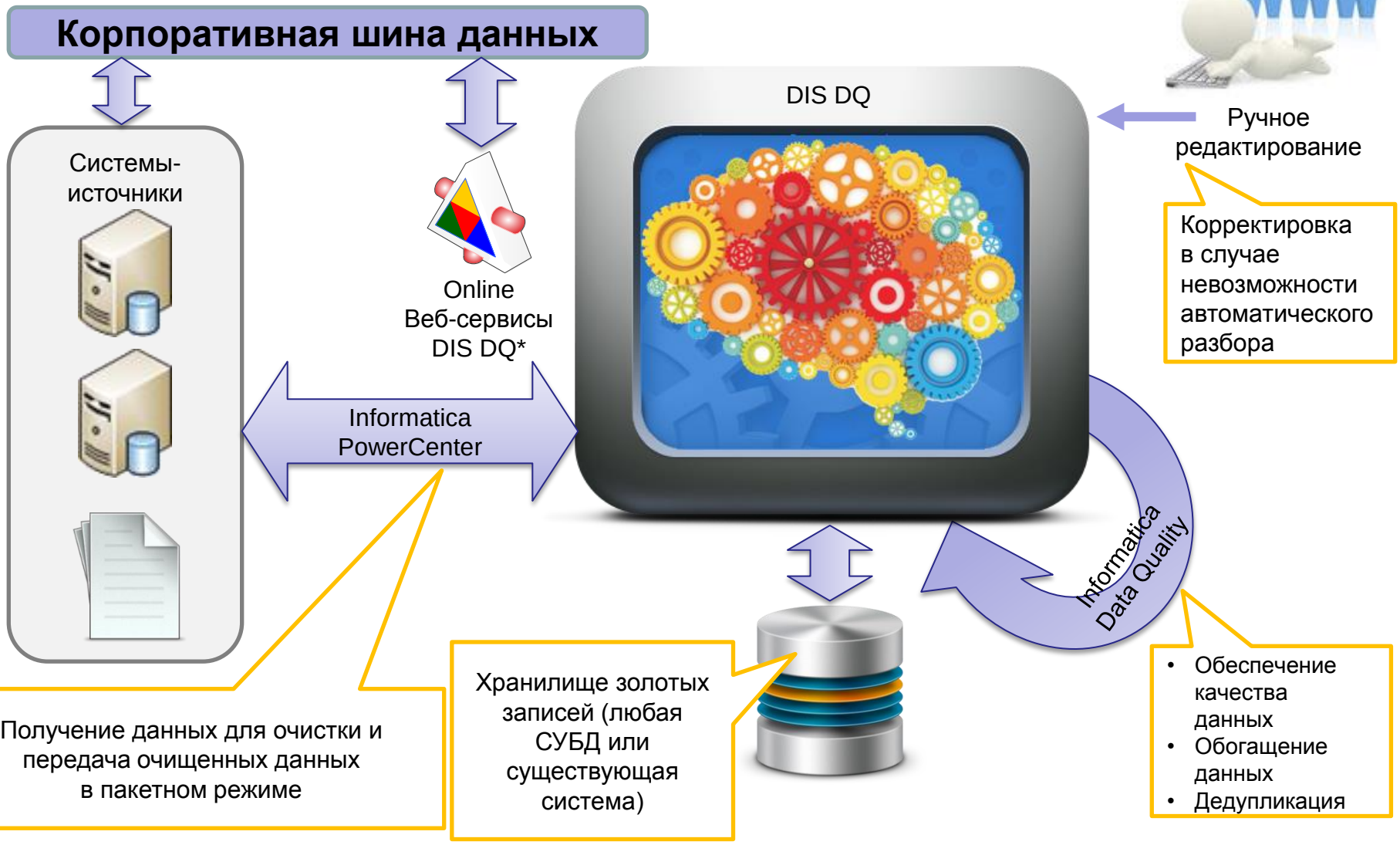
Официальные источники

- КЛАДР/ФИАС
- Почтовые индексы
- Геокоды
- Телефонные коды регионов
- Изменение телефонных кодов и номеров
- Наименования доменов
- Классификаторы

Собственные наработки

- Словарь ФИО
- Типы адресных данных
- Типы документов
- и др.

Архитектура решения



Сокращение затрат на создание и поддержку тестовых сред

Требование законодательства по обезличиванию данных

Наименование документа	Краткое описание и ссылки на необходимые разделы и пункты
Закон № 152-ФЗ «О персональных данных»	Согласно Закону № 152-ФЗ персональные данные - любая информация, относящаяся к определенному или определяемому на основании такой информации физическому лицу (субъекту персональных данных), в том числе его фамилия, имя, отчество, год, месяц, дата и место рождения, адрес, семейное, социальное, имущественное положение, образование, профессия, доходы, другая информация. Оператор при обработке персональных данных обязан принимать необходимые правовые, организационные и технические меры или обеспечивать их принятие для защиты персональных данных от неправомерного или случайного доступа к ним, уничтожения, изменения, блокирования, копирования, предоставления, распространения персональных данных, а также от иных неправомерных действий в отношении персональных данных.
Ст. 26 закона «О банках и банковской деятельности»	Согласно ст. 26 закона «О банках и банковской деятельности» к банковской тайне относится информация об операциях, счетах и вкладах клиентов и корреспондентов. По российскому законодательству кредитная организация гарантирует тайну банковского счета и банковского вклада, операций по счету и сведений о клиенте.

Типичные сложности при создании тестовых сред

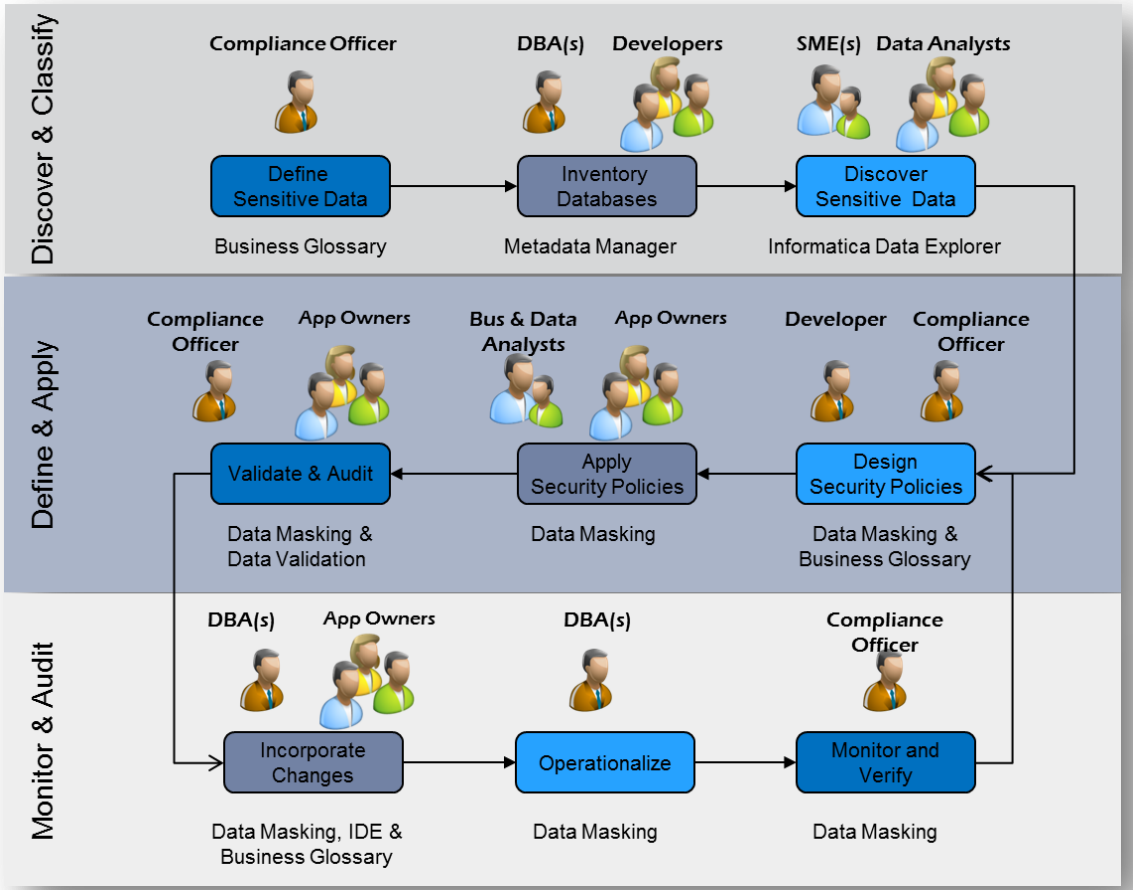
DIS

- Тестовые наборы данных повторяют продуктивную среду и занимают большое дисковое пространство
- Разные подходы для создания тестовых наборов из различных приложений
- На создание тестового набора данных требуется значительное время
- После изменения источников (например, после установки патчей) требуется значительное время на изменение процессов создания тестовой копии
- Данные не обезличиваются или обезличиваются частично
- После обезличивания теряются особенности данных
- После обезличивания теряется связанность данных
- Нет единого инструмента для управления тестовыми данными

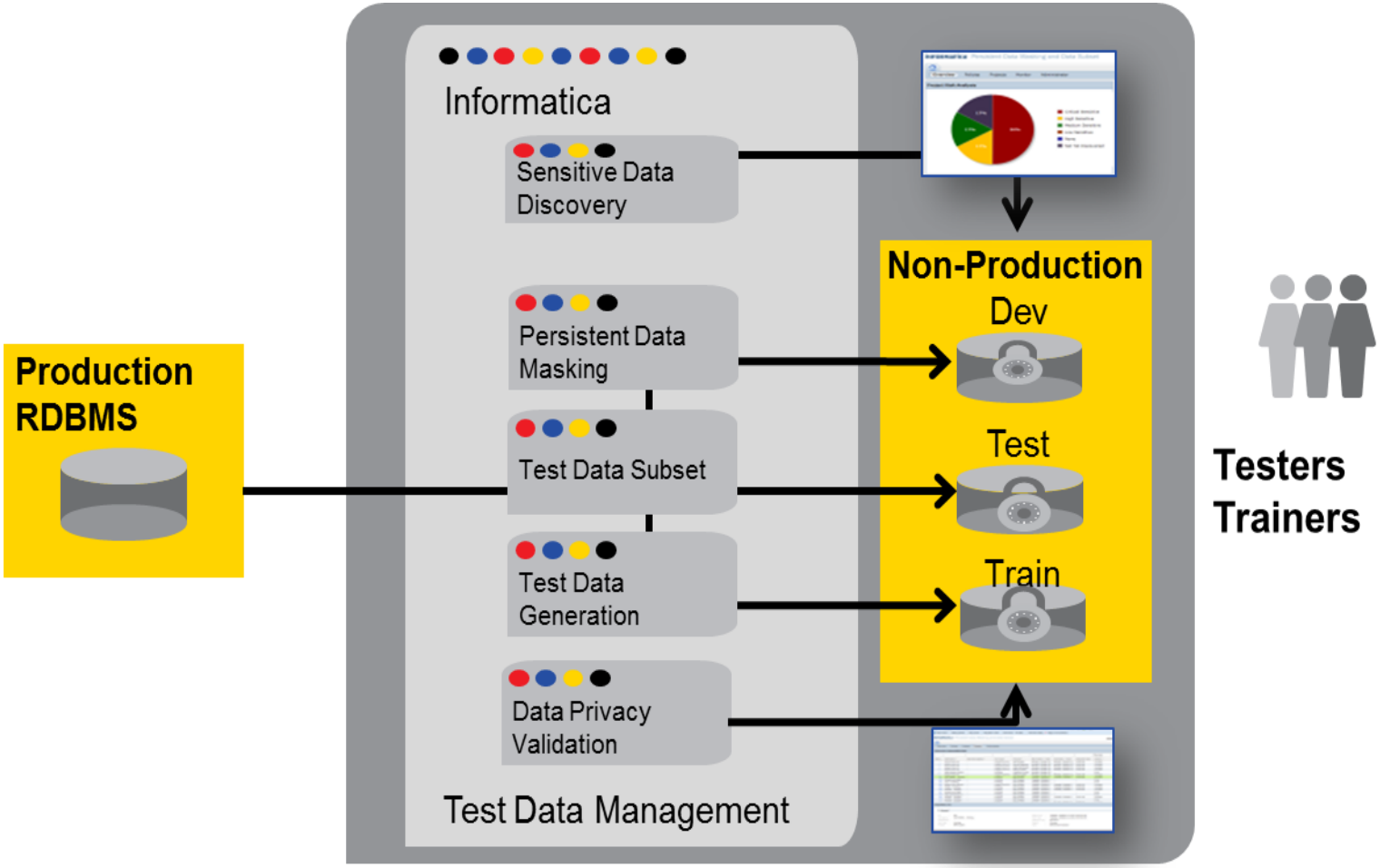
К чему это приводит:

- Высокие затраты на поддержку тестовых сред
- Долгое время создания тестовой среды
- Различные решения для различных систем сложно поддерживать
- Данные либо сильно замаскированные, либо конфиденциальная информация видна третьим сторонам
- Сложно соответствовать различным требованиям к обезличиванию данных

Методология Informatica по созданию тестовых сред

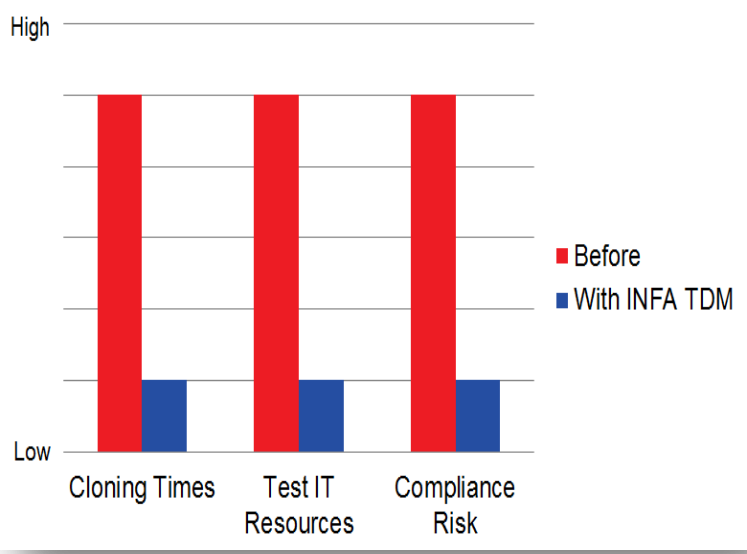


Промышленная платформа для управления тестовыми средами – Informatica TDM



Что дает Informatica?

- Единое решение для анализа, создания и обезличивания тестовых наборов данных
- Простое создание уменьшенных тестовых копий данных
- Сохранение особенностей данных при обезличивании
- Большой набор готовых правил
- Простая реализация различных политик для разных типов данных
- Быстрое внесение изменений в процессы создания тестовых сред при изменении источников
- Быстрое обучение сотрудников Банка



SBERBANK
By your side

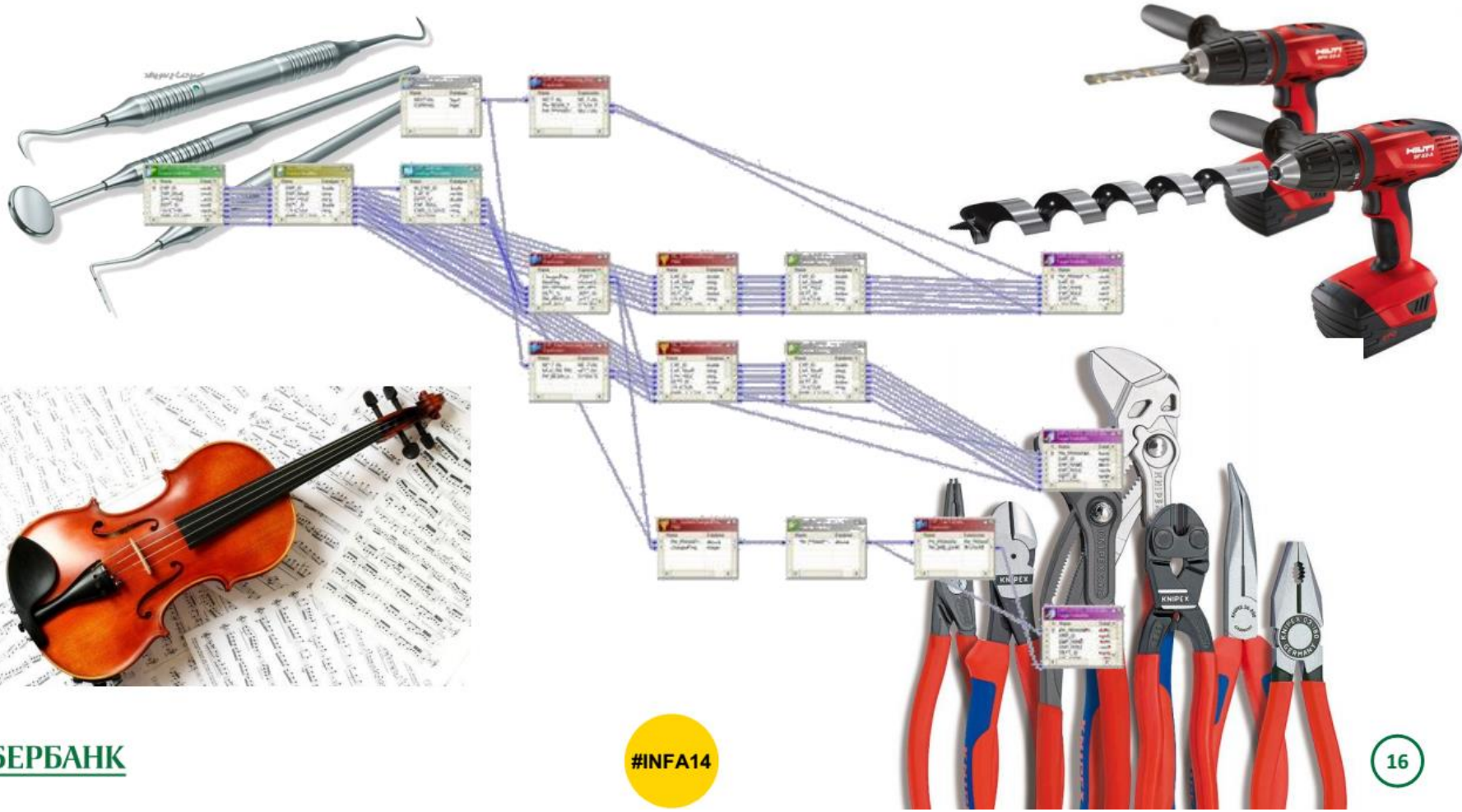
Что мы получаем в итоге?

- Сокращение затрат на интеграцию данных
- Сокращение потерь из-за низкого качества данных
- Снижение затрат на создание тестовых наборов данных
- Сокращение рисков потери конфиденциальных данных

Больше ценности данных при сокращении затрат

DIS

Пользуйтесь хорошими инструментами!



Спасибо за внимание!

+7495 645 02 01

mk@dis-group.ru

www.dis-group.ru