



Как повысить ценность данных и сократить затраты на работу с ними

Михаил Комаров
Директор по развитию бизнеса,
Решения Informatica

DIS Group

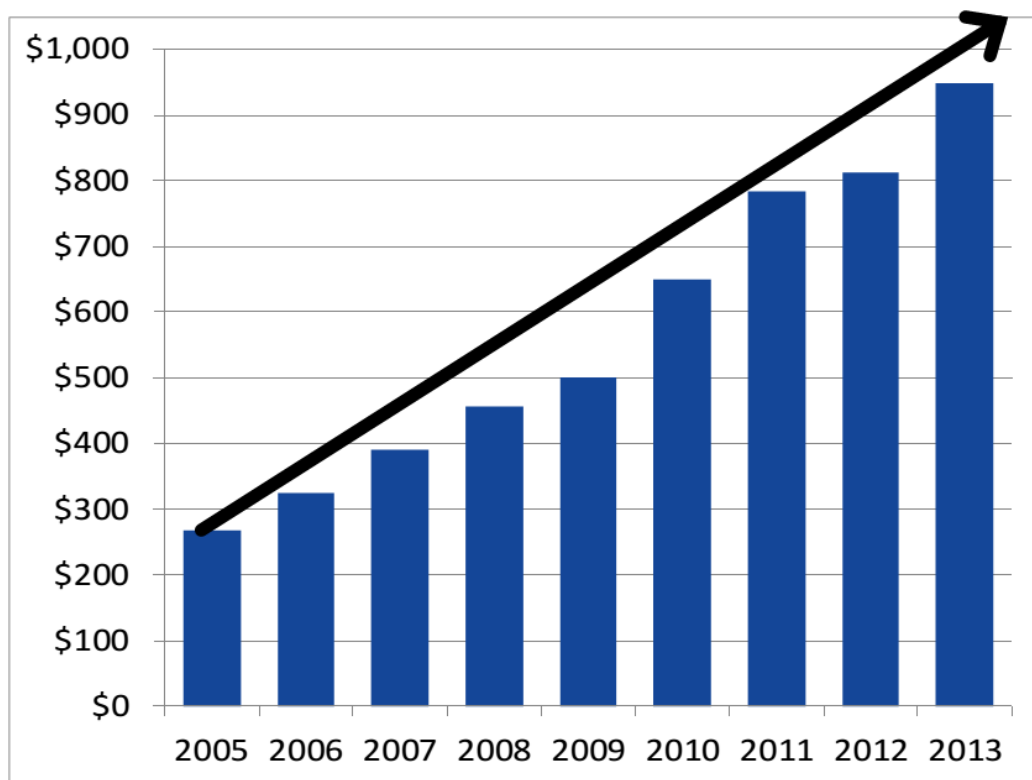
О компании Informatica

#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 <i>By your side</i>		11	 Promsvyazbank	
2	 VTB <small>World Without Barriers. VTB Group</small>		12	 ROSBANK	
3	 GAZPROMBANK		13	 Raiffeisen BANK	
4	 BTB24 <small>Большое преимущество</small>		14	 открытие <small>ХАНТЫ-МАНСКИЙ БАНК</small>	
5	 ОТКРЫТИЕ <small>FINANCIAL CORPORATION</small>		15	 CREDIT BANK OF MOSCOW	
6	 Alfa-Bank		16	 БАНК РОССИЯ <small>Банк умных решений</small>	
7	 Russian Agricultural Bank		17	 BANK SAINT PETERSBURG	
8	 Bank of Moscow		18	 AK BARS BANK	
9	 NATIONAL CLEARING CENTRE <small>MOSCOW EXCHANGE GROUP</small>		19	 B&NBANK	
10	 UniCredit Bank		20	 RUSSIAN STANDARD BANK	

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

БИЗНЕС

$$\text{Возврат инвестиций в данные} = \frac{\text{Ценность данных}}{\text{Стоимость данных}}$$

MAINFRAME

10^2

CLIENT-SERVER

10^4

WEB

10^6

CLOUD

10^7

SOCIAL

10^9

INTERNET OF THINGS

10^{11}



1990s

Customers/
Consumers

2007

Business
Ecosystems

2011

Communities
& Society

2014

Devices
& Machines

The background of the slide is a composite image. The top half shows a bright blue sky with a sunburst effect on the left and several white, fluffy clouds. The bottom half shows a flat, green field. A white curved line separates the sky from the field. The title text is centered over the sky portion.

Сокращение затрат на интеграцию данных

Что дает Informatica в интеграции данных? DIS

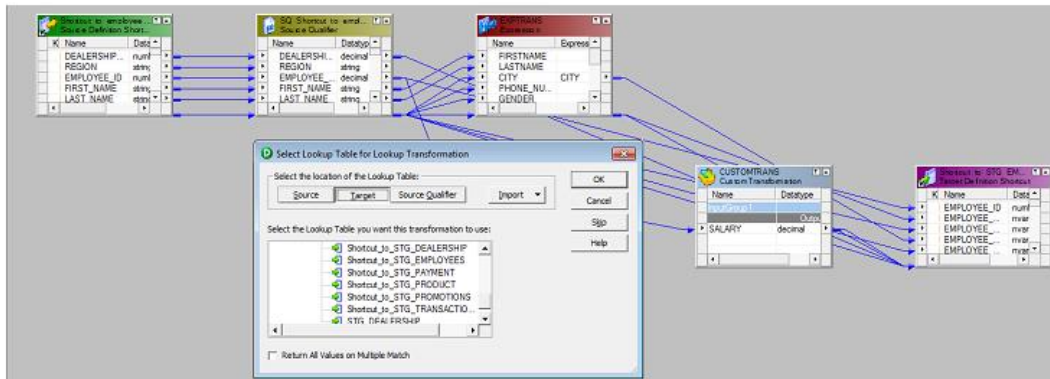
Пять аргументов за использование ETL-машины

- 
1. Понятность
 2. Производительность
 3. Поддержка правильного стиля разработки
 4. Выстраивание зависимостей
 5. Мониторинг

Простота внесения изменений

Стиль разработки: как влияет инструмент?

Если не контролировать себя,
то обычно получается так...



...а вот в карте
Informatica PowerCenter
добавить lookip проще,
чем ветвление

Доступность данных

Уровень доступности



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



Уровень доступности = 100%

Масштаб Informatica

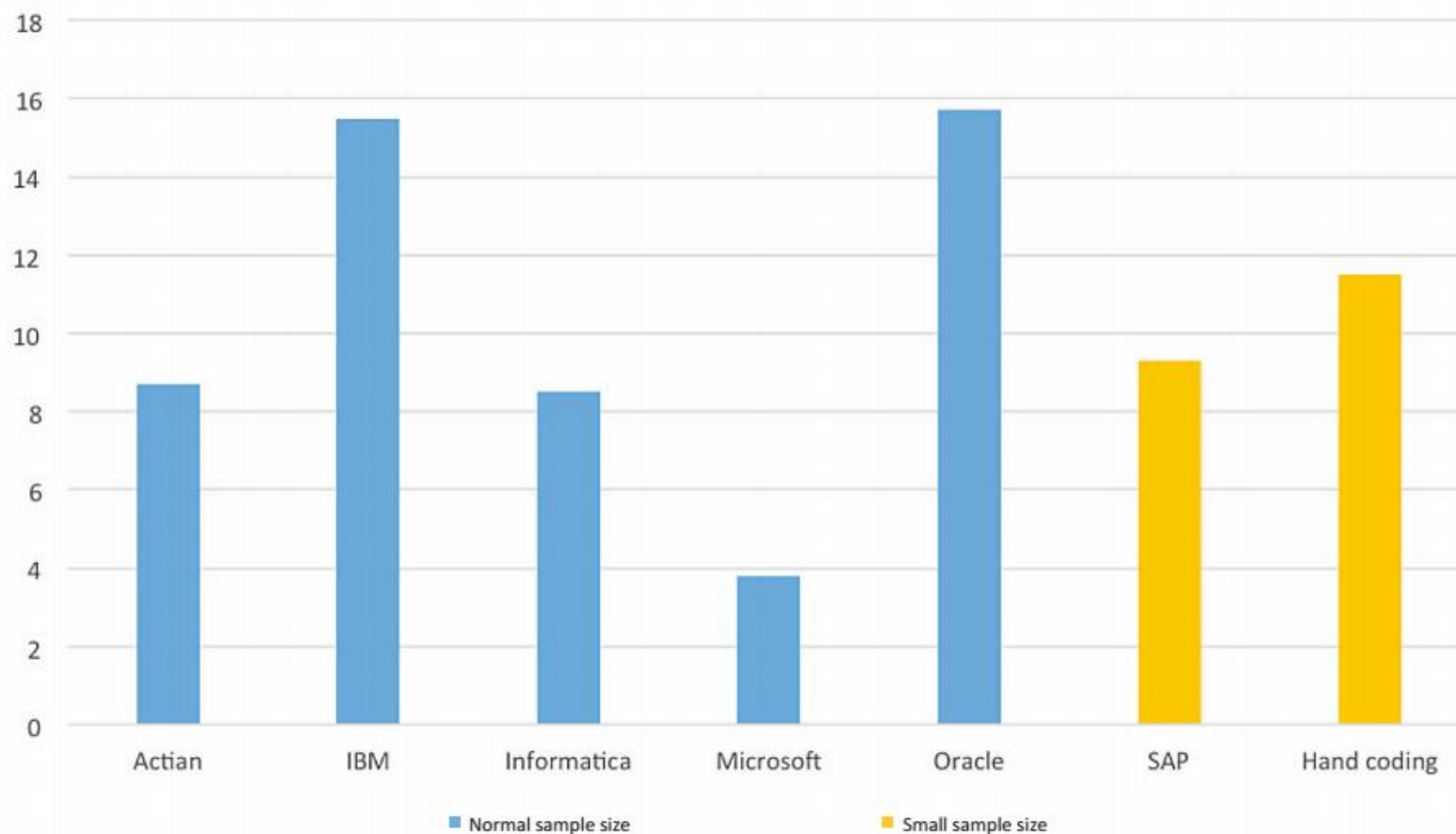


Количество источников	61
Типы источников	Oracle, MS SQL, DB2, Informix, Text files, XML files
Ежедневный объем загружаемых данных	0,5ТБ
Количество карт загрузки	4,5 тыс.
Количество таблиц приемников	1 000+

6,7 миллиардов
ежедневно загружаемых записей

Самое низкое TCO среди лидеров Gartner

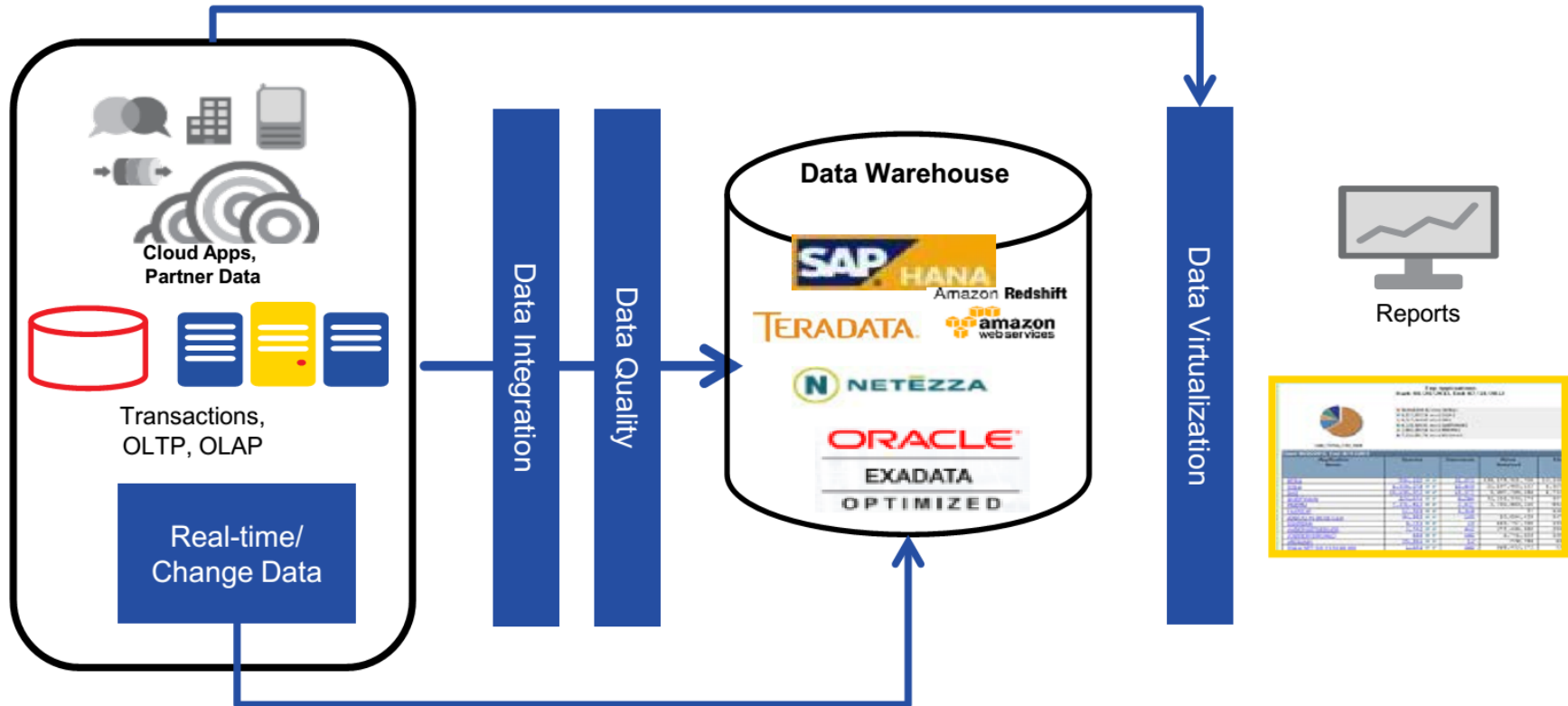
Figure 7: 3 year TCO per project per source/target (\$K)



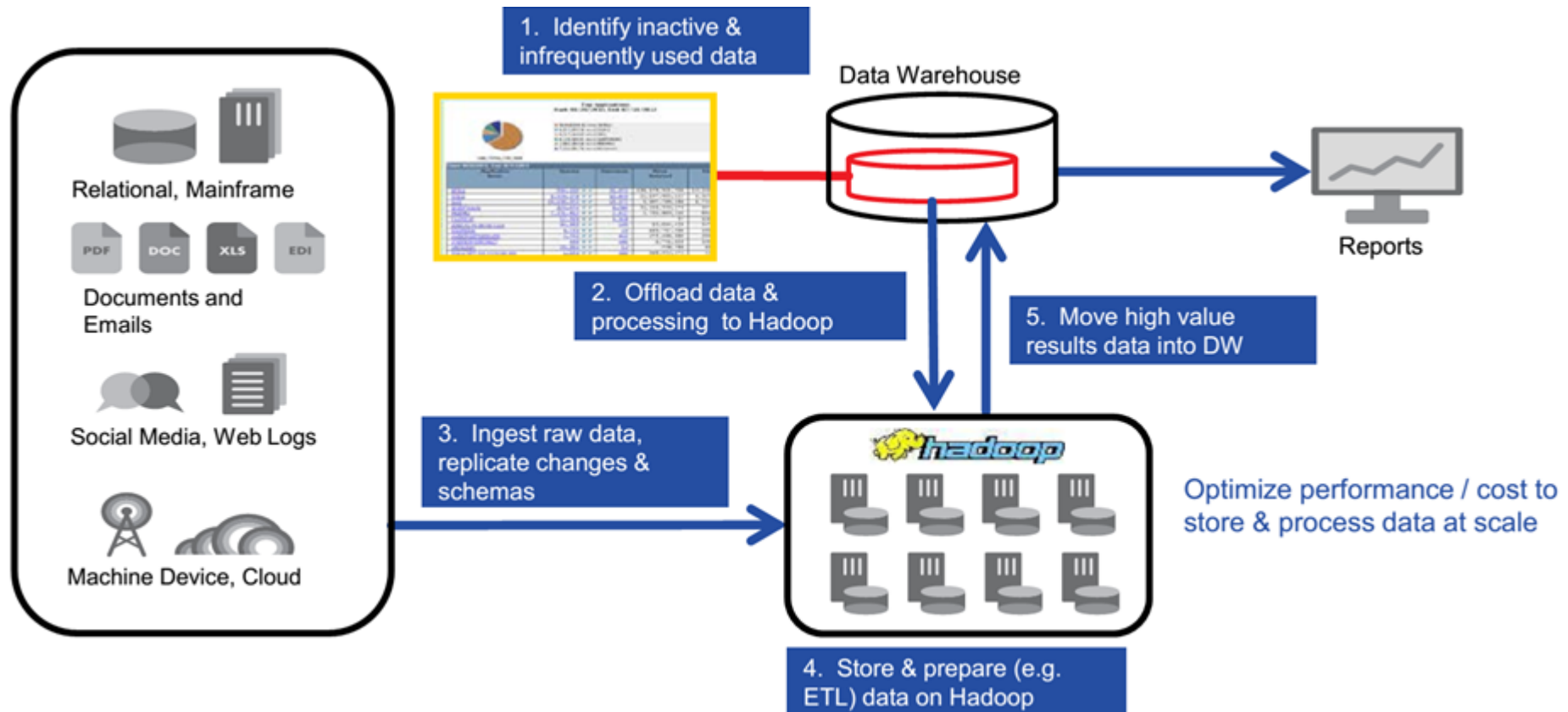
The background of the slide is a composite image. The top half shows a bright blue sky with a sunburst effect on the left and several white, fluffy clouds. The bottom half shows a flat, green field. A white curved line separates the sky from the field. The text "Big Data?" is centered in the sky area.

Big Data?

Традиционное хранилище данных



Next Generation Analytics

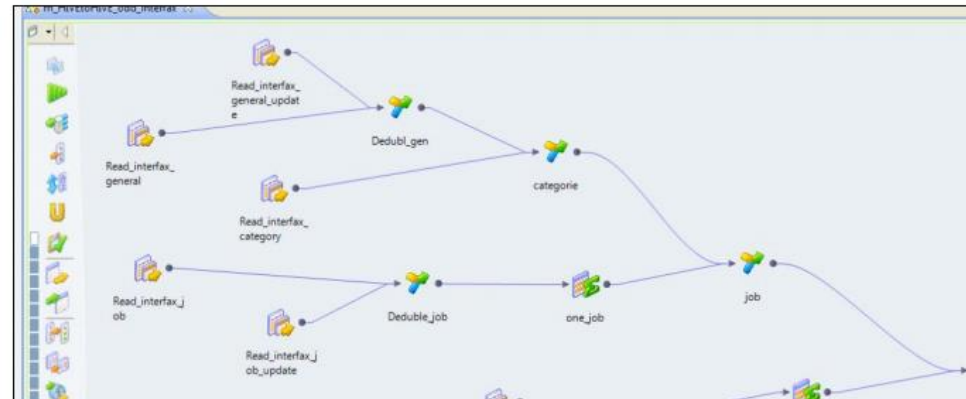
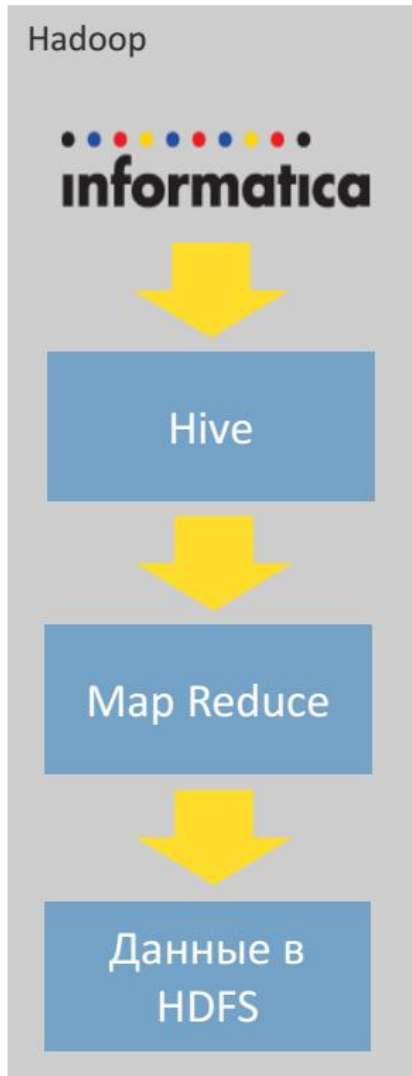


Какие препятствия для работы с Big Data?

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



Data Lake – Informatica ETL



m_HIVetoHIVE_odd_interfax	Mapping Task	Failed	xbpmsdEIEeSzSWhirQ...	Administrator
m_HIVetoHIVE_odd_interfax	Mapping	Failed	xlMkSdEIEeSzSWhirQ...	Administrator
exec1	Script	Failed	xlMkSdEIEeSzSWhirQ...	Administrator
exec1_query_1	Hive Query	Failed	infa_2015032310353...	Administrator
exec0	Script	Completed	xlMkSdEIEeSzSWhirQ...	Administrator
exec0_query_4	Hive Query	Completed	infa_2015032310282...	Administrator

hg 263 results.

20150323102828_bb0a06ee-9ea8-4ab5-8ed4-0a4b74c9ec9d

Name	Query
	-- ****
	-- Query [exec0_query_4]
	INSERT OVERWRITE TABLE w1987664460_fig_group_m_hivetohive_odd_interfax SELECT (CASE 1 WHEN ELSE 0 END) THEN 1 WHEN (CASE WHEN single_use_subq531.a10 IS NULL THEN 1 ELSE 0 END) THEN 2 infaNativeUDFCallDate('TO_DATE', single_use_subq530.a10, 'MM/DD/YYYY HH24:MI:SS'), infaNativeUDF('MM/DD/YYYY HH24:MI:SS')) = 1 WHEN TRUE THEN 1 WHEN FALSE THEN 0 ELSE CAST(NULL AS INT) EN a1, single_use_subq530.a1 as a2, single_use_subq530.a2 as a3, single_use_subq530.a3 as a4, single_u single_use_subq530.a6 as a7, single_use_subq530.a7 as a8, single_use_subq530.a8 as a9, single_use_ single_use_subq531.a0 as a12, single_use_subq531.a1 as a13, single_use_subq531.a2 as a14, single_u single_use_subq531.a5 as a17, single_use_subq531.a6 as a18, single_use_subq531.a7 as a19, single_u single_use_subq531.a10 as a22, single_use_subq530.a11 as a23, single_use_subq530.a12 as a24, sing a25 FROM (SELECT alias.system_id as a0, alias.full_name as a1, alias.dob as a2, alias.pob as a3, alias.l as a6, alias.authority as a7, alias.job as a8, alias.category as a9, SUBSTR(alias.update_dt, (CASE WHEN as a10, alias.source as a11, alias.iso_country as a12 FROM prod_odd_ei.interfax alias) single_use_subq a0, single_use_subq528.a2 as a1, single_use_subq528.a3 as a2, single_use_subq528.a4 as a3, single_u single_use_subq528.a7 as a6, single_use_subq528.a8 as a7, single_use_subq528.a9 as a8, single_use_ single_use_subq528.a11 as a11, single_use_subq529.a1 as a12 FROM (SELECT alias.key_entity as a0, h prod_odd_ei.interfax_country_update alias1 JOIN prod_odd_ei.interfax_country alias ON ((alias.update



Customer 180°



Понимаем, что клиент делает только с точки зрения нашего бизнеса

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



Customer 360°



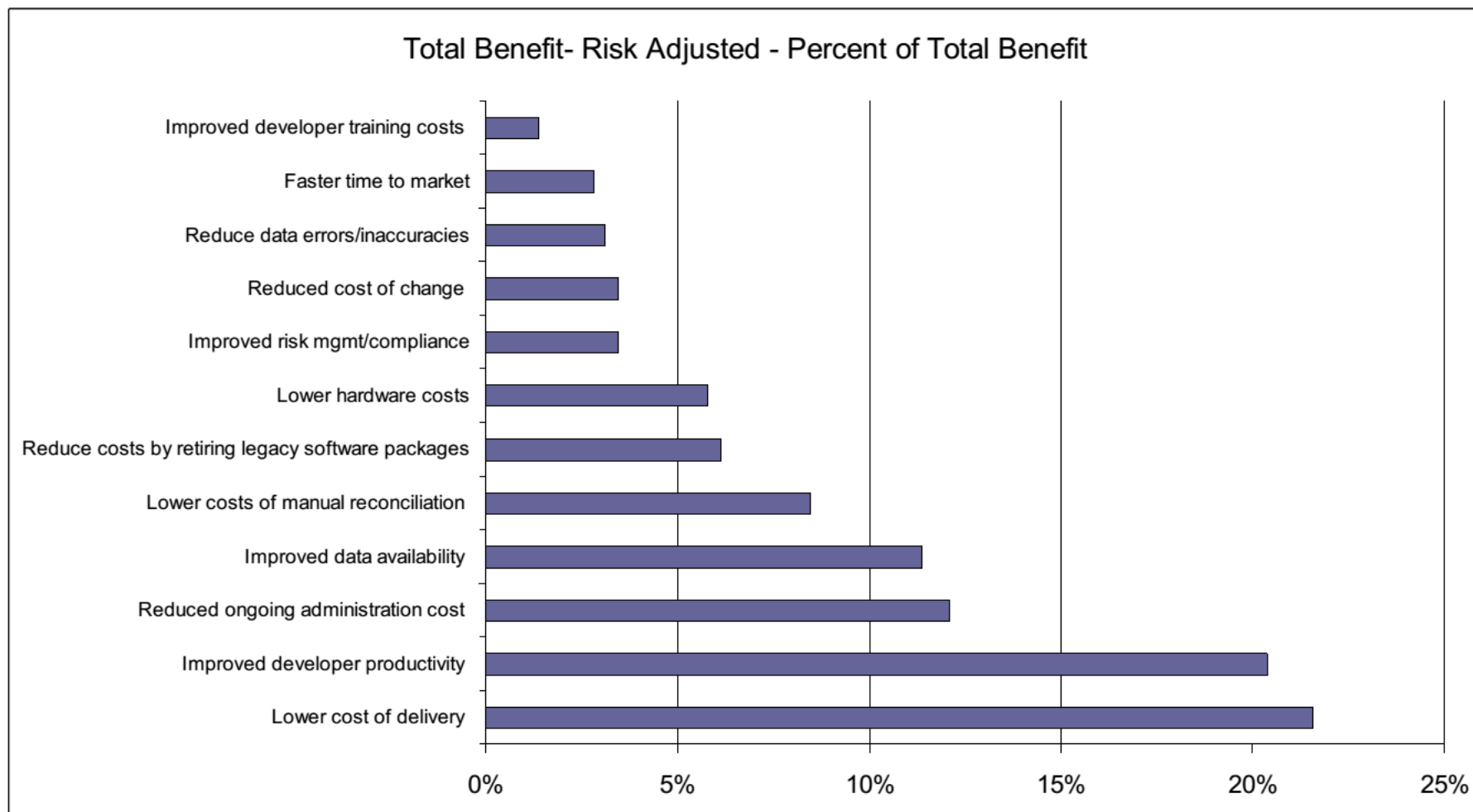
Понимаем, что происходит с клиентом

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.

The background of the slide is a composite image. The top half shows a bright blue sky with a sunburst effect on the left and several white, fluffy clouds. The bottom half shows a flat, green field. A white curved line separates the sky from the field. The title text is centered over the sky portion.

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

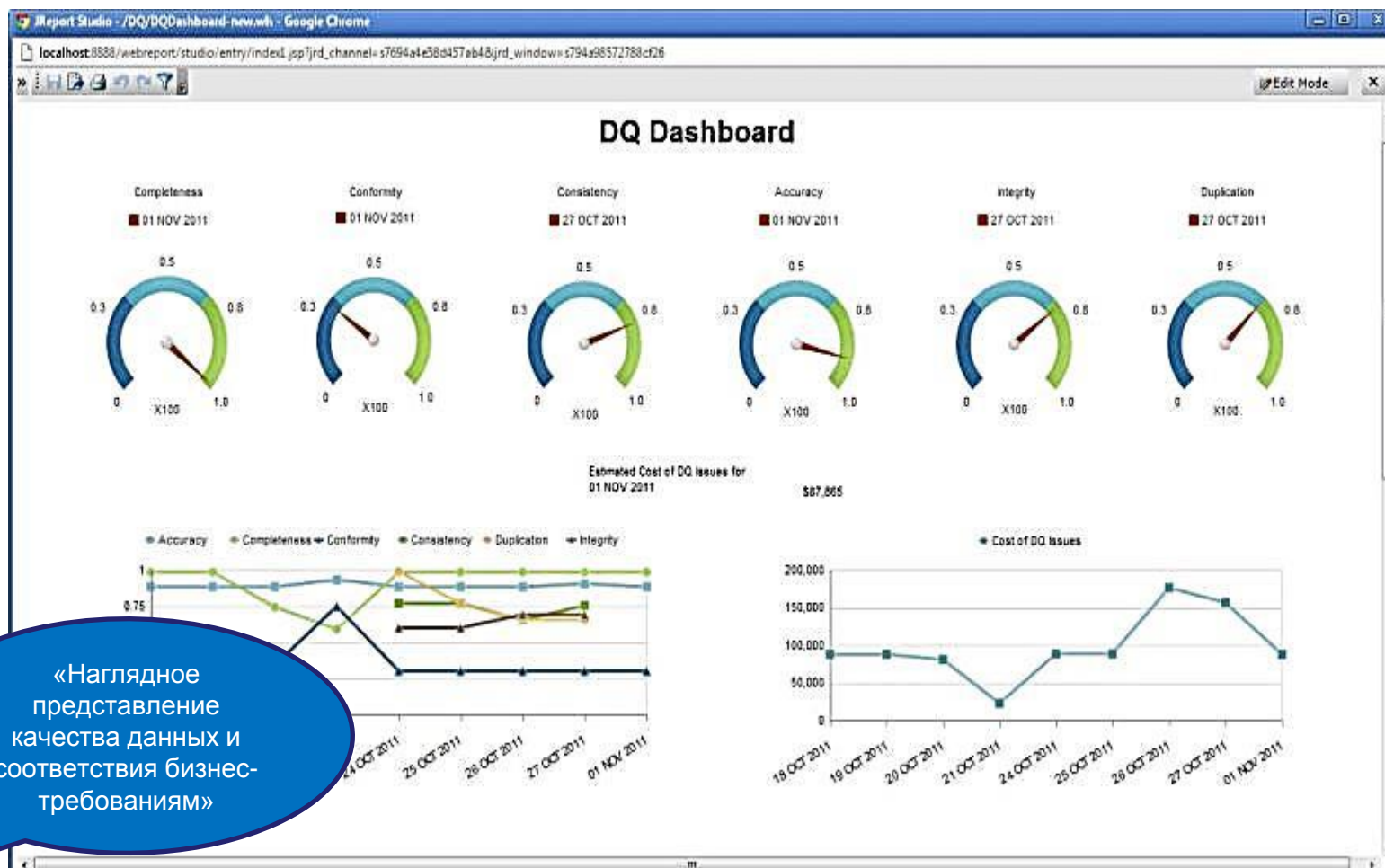
Для каждого фактора, параметра и результата модели количественной оценки рисков (далее – компоненты):

- 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



Профиль данных в любых системах



Informatica Analyst - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Informatica Administrator: Domain x Informatica Analyst x +

infandmsvr:8085/AnalystTool/com.informatica.at.AnalystTool/Index.jsp#p=U:IX4WF1EDEeOjYc3NHLTgWQ&u=U:Zps7ZKZEeOD68LVvD6W7A8c=com.informatica.profilng.services.model.persist.scorecard.Scorecard

Most Visited Getting Started My CRM INFA Administrator Informatica Analyst Metadata Manager & ... Informatica DQ Data ...

INFORMATICA Analyst Administrator Log Out Manage Help Search by Name Go

Browse: Projects Анализ_Источника_PAR... Карта_Качества_PARTN...

Scorecard Properties Actions

Карта_Качества_PARTNER - metrics Last Run On: Nov 21, 2013 3:18:59 PM GST

Name	Total Rows	Invalid Rows	Score	Score Trend	Metric Weight	Data Object	Source	Source Type	Drill down
Проверка_Заполненности			94.57						
Дата_Регистрации	92	6	93.48		1	C_PARTNER	REGISTER_DATE	Column	
КПП	92	0	100		1	C_PARTNER	KPP	Column	
Менеджер	92	16	82.61		1	C_PARTNER	ID_BUSSMAN	Column	
Регион	92	0	100		1	C_PARTNER	ID_REGION	Column	
ИНН	92	3	96.74		1	C_PARTNER	DIC	Column	
Проверка_Бизнес_Качества			92.39						
Проверка_ОПФ	92	3	96.74		1	C_PARTNER	Проверка_ОПФ	Reusable Rule	
Проверка_ИНН	92	7	92.39		1	C_PARTNER	Проверка_ИНН	Reusable Rule	
Налоговый_Режим	92	12	86.96		1	C_PARTNER	TAX_SYSTEM	Column	
Статус	92	9	90.22		1	C_PARTNER	STATUS	Column	
Название	92	4	95.65		1	C_PARTNER	NAME	Column	
Проверка_Сумм			85						
Сходимость_Данных	20	3	85		1	ORDERS	Check_VAL	Reusable Rule	

Drill down: Проверка_ОПФ = 'ОПФ соответствует справочнику' (All 89 rows)

Valid Rows Invalid Rows

ROWID_OBJECT	CREATOR	CREATE_DATE	UPDATED_BY	LAST_UPDATE_DATE	CONSOLIDATION	DELETED_IND	DELETED_BY	DELETED_DATE	LAST_ROWID_SY	DIRTY_IND	INTERACTION_ID	HUB_STATE_IND	CM_DIRTY_IND	ID_SELLER	ID_REGION
56	admin	Mar 21, 2013 7:51admin	Mar 21, 2013 8:432	NULL	NULL	NULL	HOMER	0	NULL	1	NULL	163383	22		
57	admin	Mar 21, 2013 7:51admin	Mar 21, 2013 8:432	NULL	NULL	NULL	HOMER	0	NULL	1	NULL	163574	470		
58	admin	Mar 21, 2013 7:51admin	Mar 21, 2013 8:432	NULL	NULL	NULL	HOMER	0	NULL	1	NULL	163652	22		
59	admin	Mar 21, 2013 7:51admin	Mar 21, 2013 8:432	NULL	NULL	NULL	HOMER	0	NULL	1	NULL	163687	22		
60	admin	Mar 21, 2013 7:51admin	Mar 21, 2013 8:432	NULL	NULL	NULL	HOMER	0	NULL	1	NULL	163818	491		
61	admin	Mar 21, 2013 7:51admin	Mar 21, 2013 8:432	NULL	NULL	NULL	HOMER	0	NULL	1	NULL	163892	309		

Start Services D:\work\Projects\IDR\Py... TRAINING@INFA\VMWR Informatica Developer Informatica Analyst ... EN 1:41 PM Friday

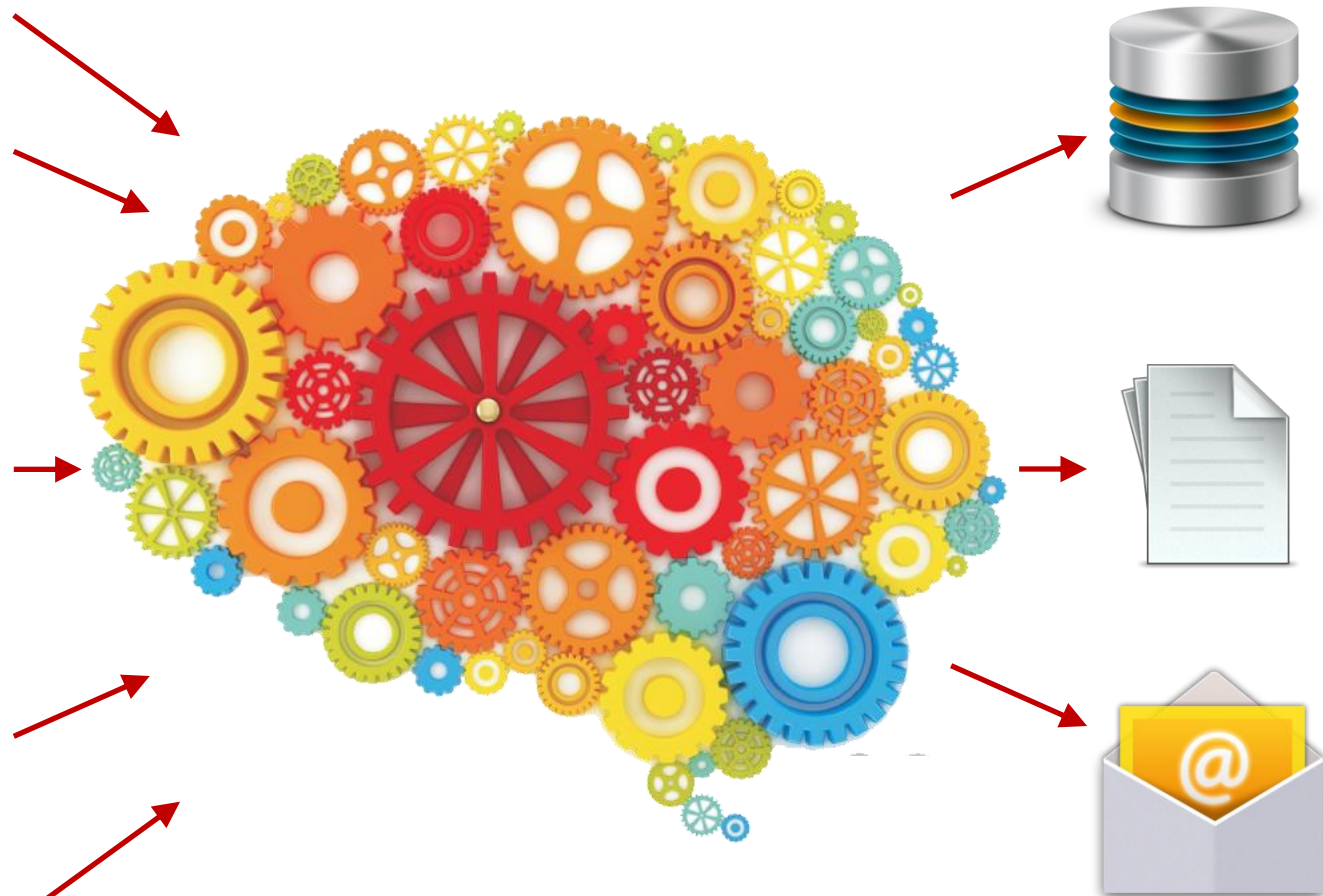
Готовые правила для России

Официальные источники

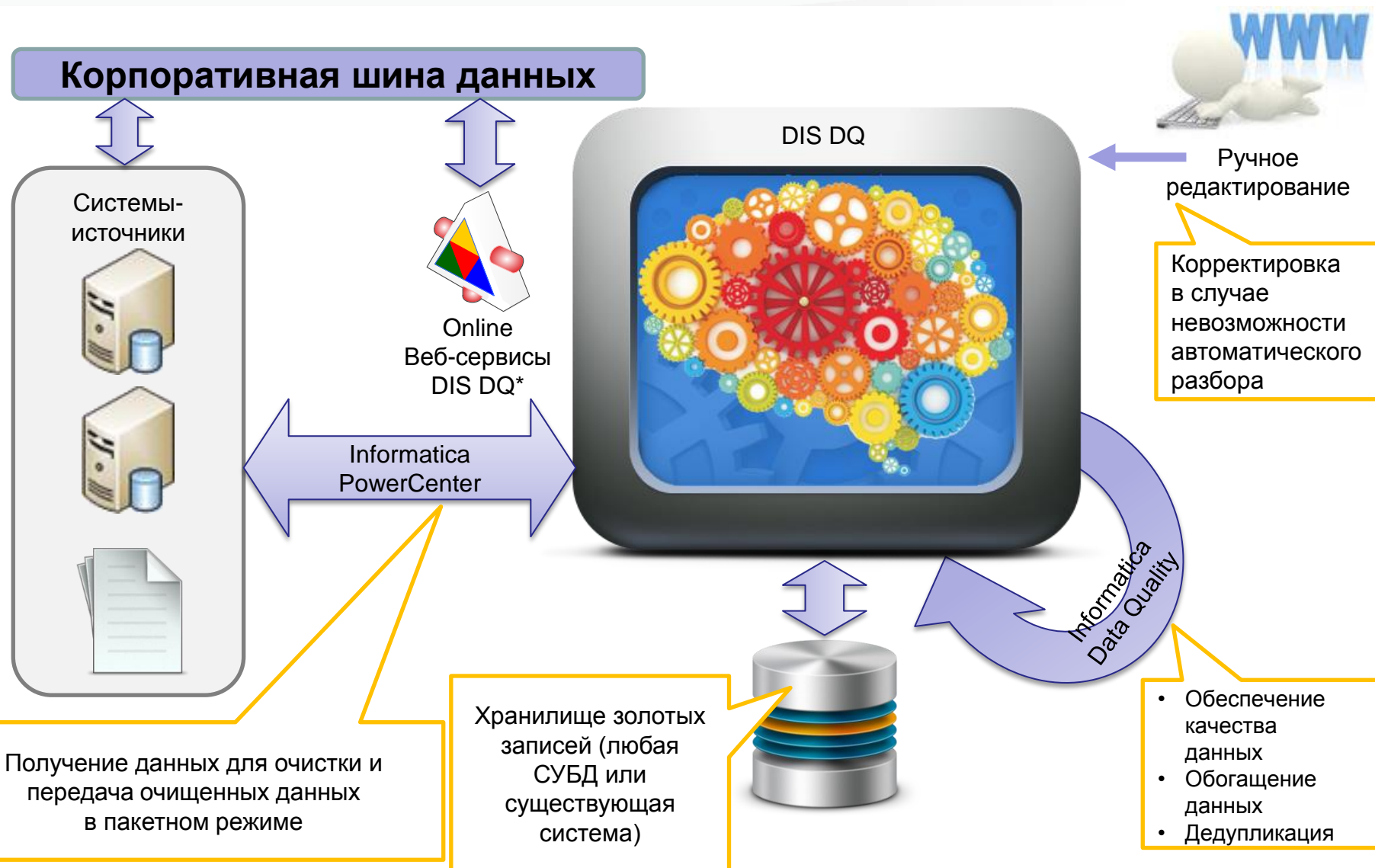
- КЛАДР/ФИАС
- Почтовые индексы
- Геокоды
- Телефонные коды регионов
- Изменение телефонных кодов и номеров
- Наименования доменов
- Классификаторы


Собственные наработки

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



Архитектура решения



The background of the slide is a composite image. The top half shows a bright blue sky with scattered white clouds and a sunburst effect on the left. The bottom half shows a flat, green field. A white curved line separates the sky from the field. The text is centered over the sky portion.

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»	<p>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.</p> <p>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.</p>
Ст. 26 закона «О банках и банковской деятельности»	<p>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.</p>

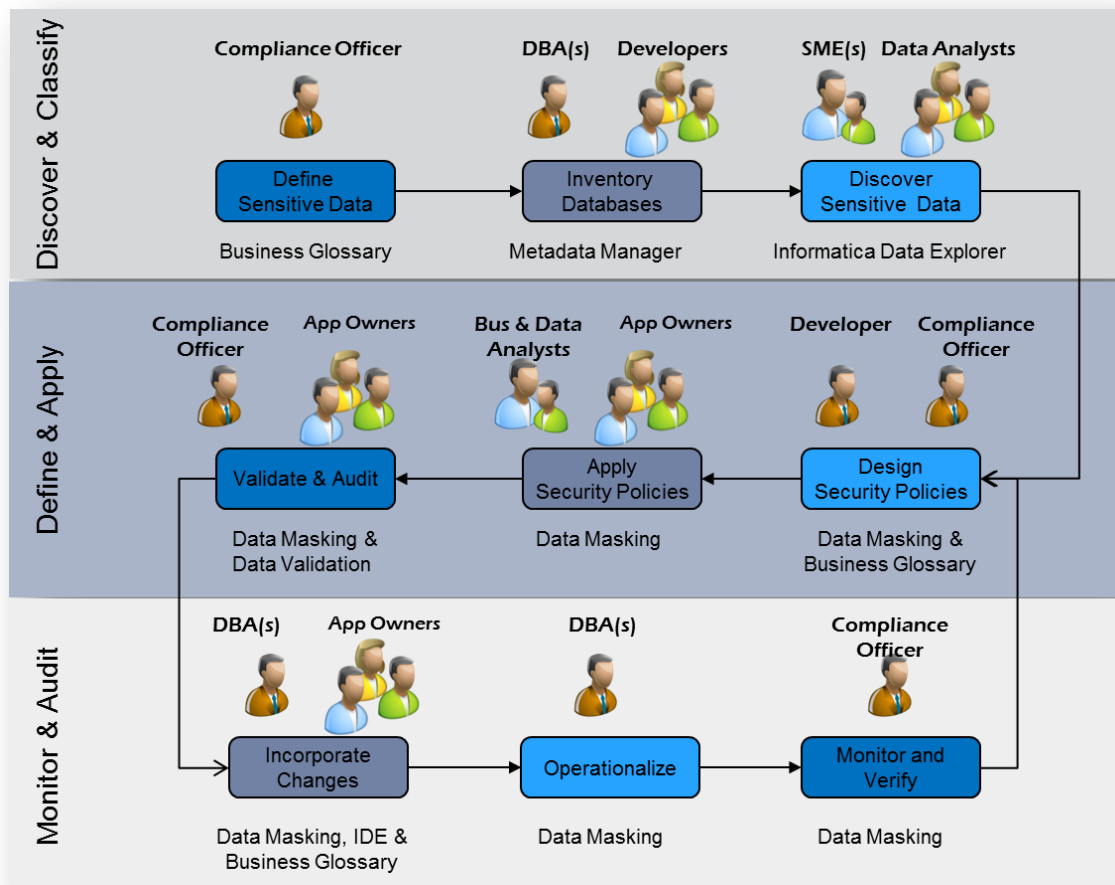
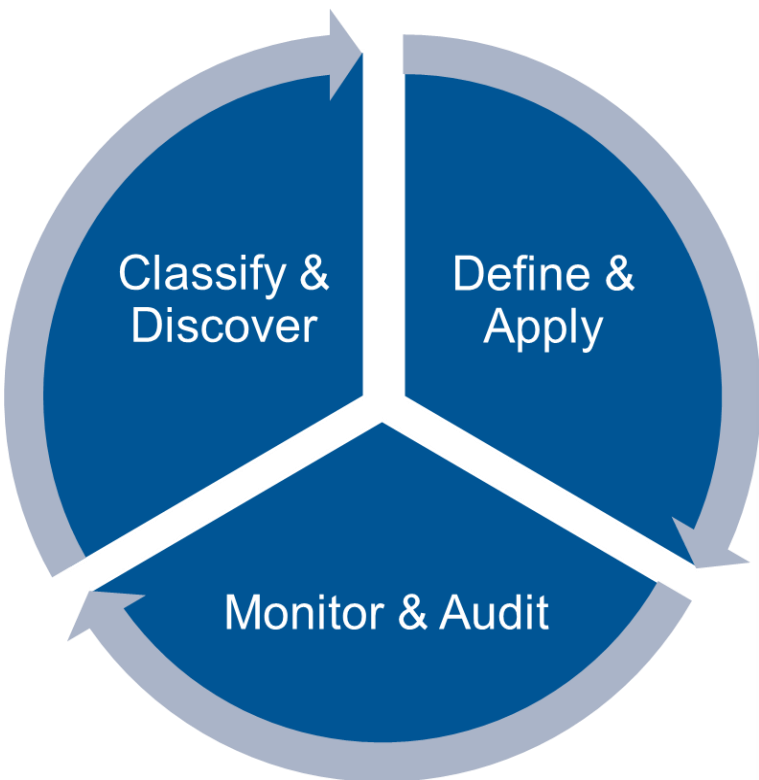
Типичные сложности при создании тестовых сред

- 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

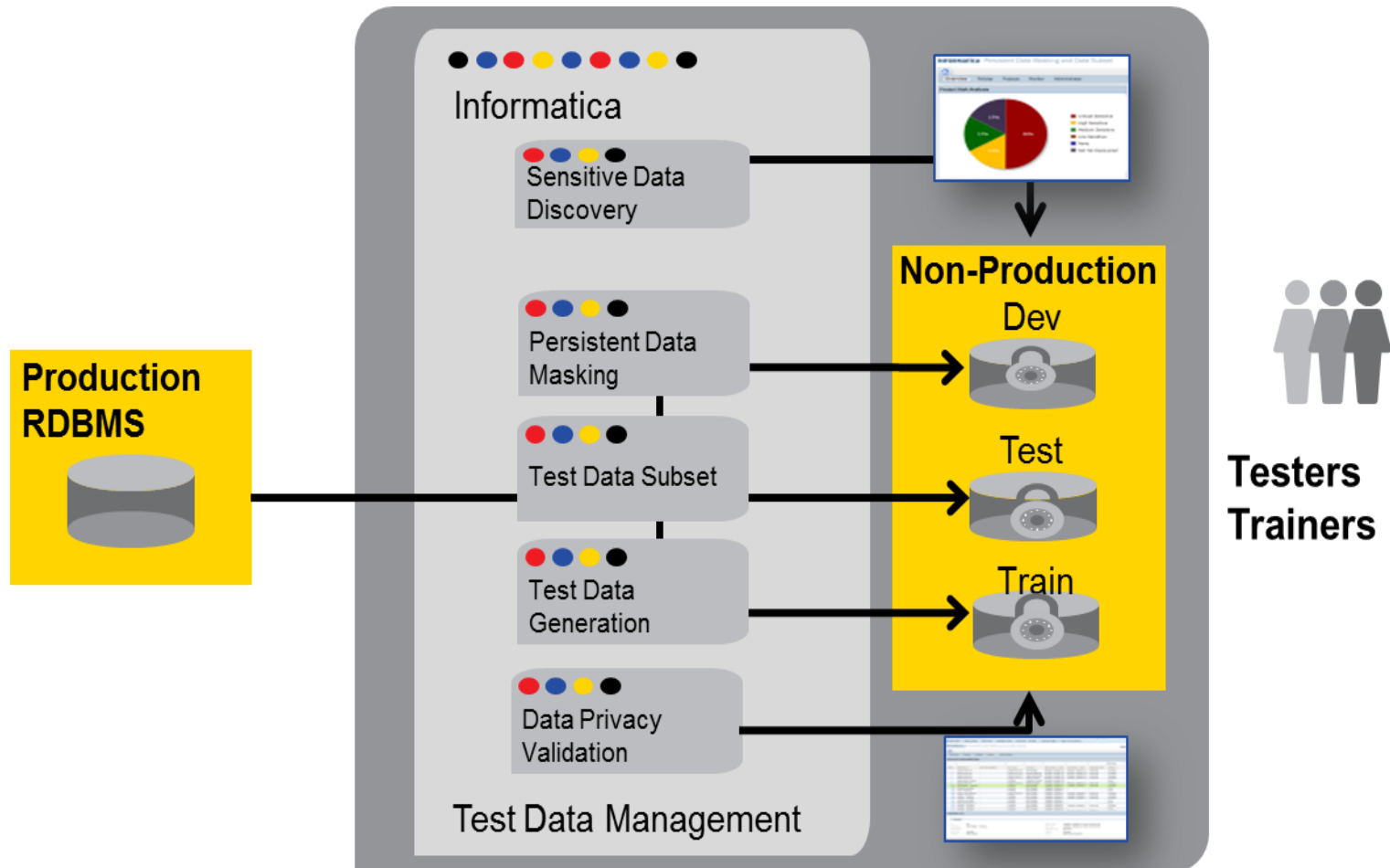
К чему это приводит:

- 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

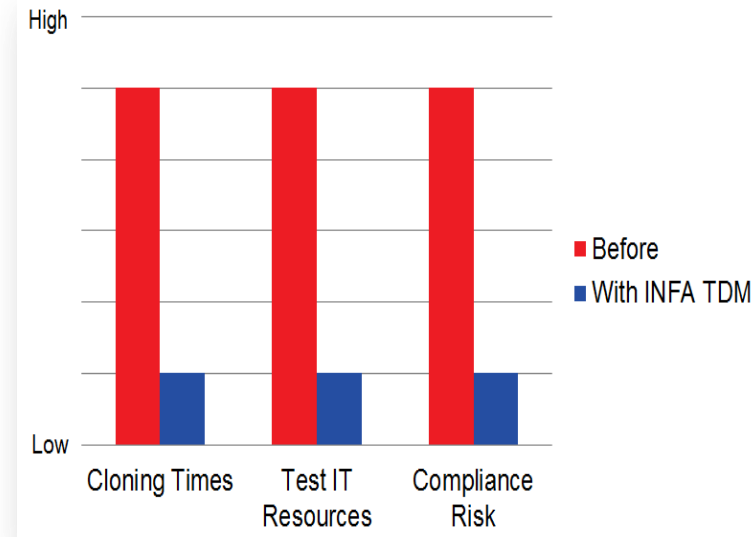


Промышленная платформа для управления тестовыми средами – Informatica TDM



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

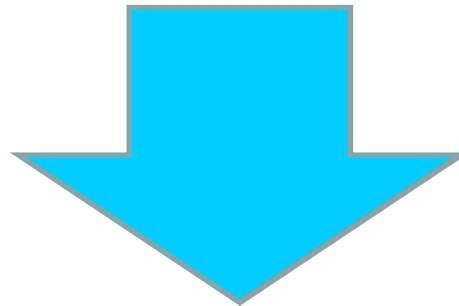


SBERBANK

By your side

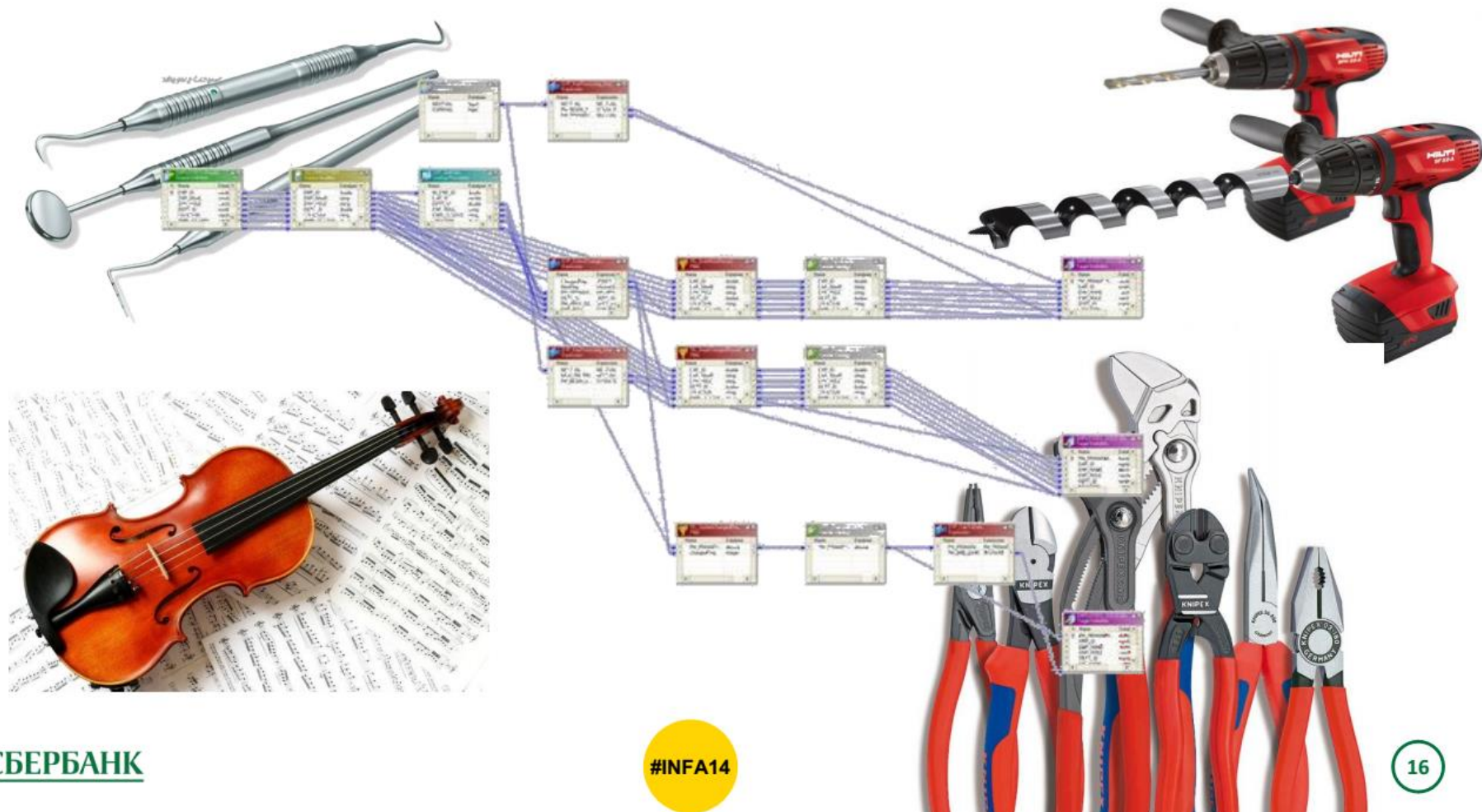
Что мы получаем в итоге?

- 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

Пользуйтесь хорошими инструментами!



Спасибо за внимание!

+7495 645 02 01

mk@dis-group.ru

www.dis-group.ru