ЛЕКЦИЯ 9

Ростелеком

DIS GROUP

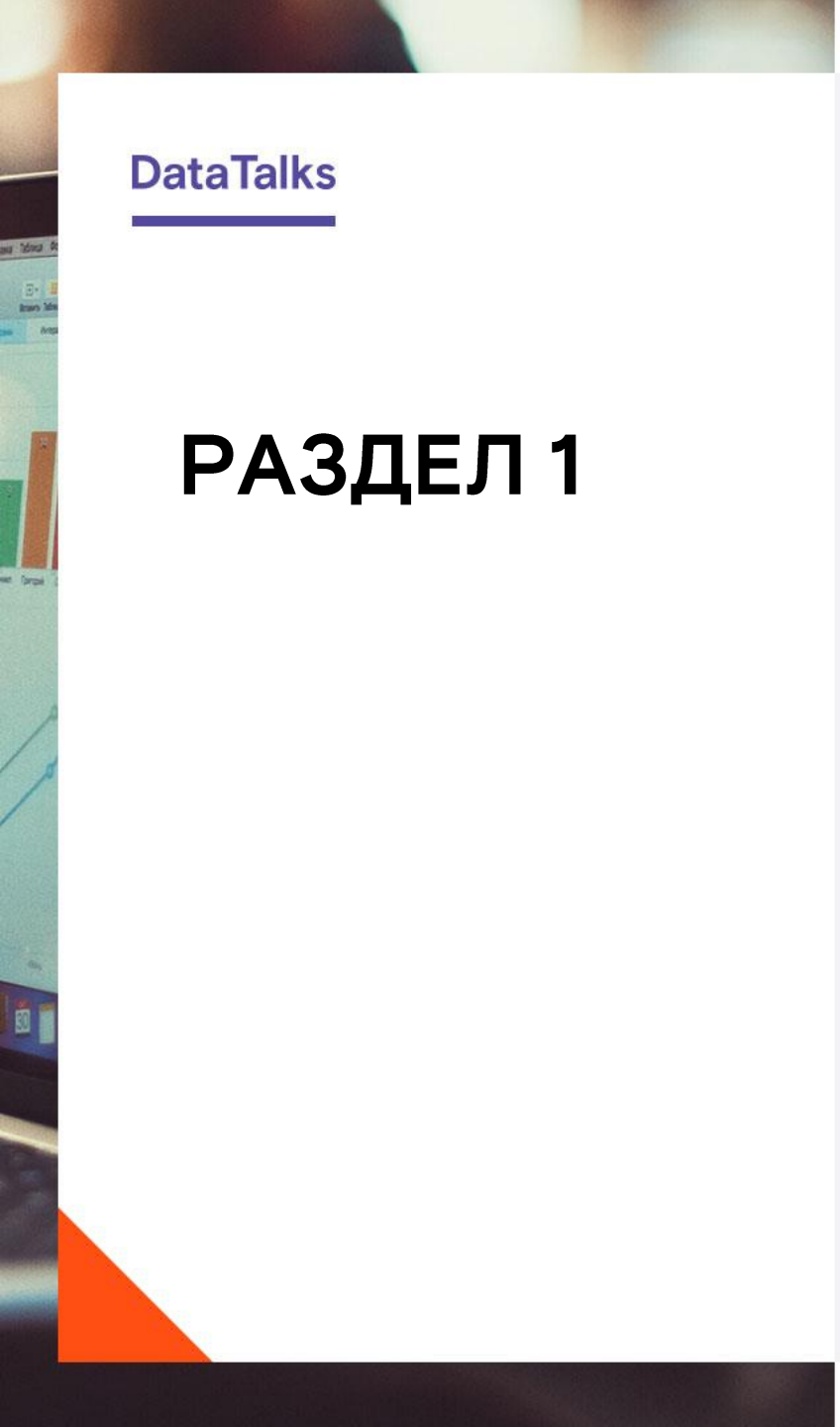
Олег Гиацинтов

Технический директор,

DIS Group

Работа с качеством данных. Профилирование, очистка и DQ-мониторинг

DataTalks



DataTalks

РАЗДЕЛ 1

Качество данных. Общие вопросы

- Что такое качество данных
- Метрики качества данных
- Основные виды проблем качества данных

Ростелеком

Что такое качество данных

Качество данных – совокупность свойств и характеристик данных, уровень или вариант которых формируется при создании или использовании с целью удовлетворения существующих потребностей

Стандарты по качеству данных:

- ГОСТ Р – 201X/ISO/TS 8000-150:2011 – Качество данных. Часть 150. Основные данные. Структура управления качеством.
- ИСО 8000-100/ISO 8000-100:2012 – Качество данных. Часть 100. Основные данные: обзор.

Бизнес- и техническое качество данных

Ростелеком

- **Техническое качество** исходных данных – соответствие загружаемых данных ограничениям, которые определяются моделью хранения данных в хранилище данных (полнота, ссылочная целостность, уникальность полей и т.п.)

- Процедуры проверки технического качества исходных данных - «очистка данных» или «технический контроль качества данных»

- **Бизнес-качество** данных - соответствие данных, уже загруженных в структуры хранения условиям и логике решаемых бизнес-пользователями задач (методики, эмпирические зависимости и т.п. априорные положения)

- Процедуры проверки бизнес-качества данных - «аудит данных, загруженных в хранилище»

Метрики качества данных

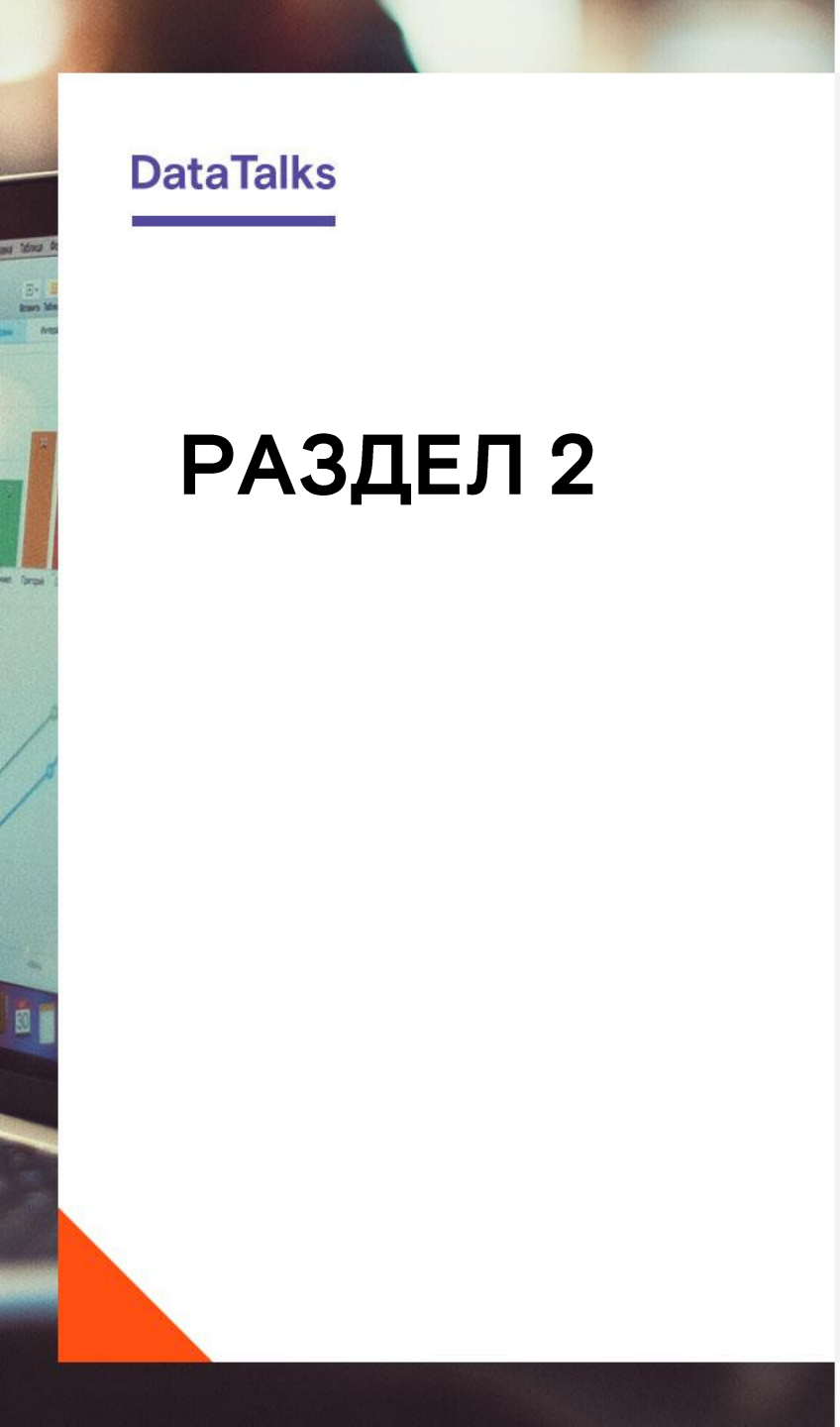
Полнота	Достаточность объема данных, глубины данных и широты данных
Своевременность	Обязательность фиксирования и использования моделей количественной оценки рисков данных с требуемой актуальностью
Согласованность	Непротиворечивость данных, отражающих различные аспекты свойств и характеристик, а также целостность соответствующих идентификационных ссылок
Доступность	Обеспеченность пользователей прямым доступом к данным в необходимых разрезах на всех этапах разработки и количественной оценки рисков
Контролируемость	Осуществление контроля качества и происхождения данных, в том числе посредством отражения структуры и зависимостей данных
Восстанавливаемость	Способность данных сохранять установленный уровень функциональности и качества после их утраты, повреждения или изменения

Основные проблемы качества данных

Код	Наименование	Страна	Валюта	Дата погашения	Рейтинг	Вероятность дефолта (PD)
34598	Bank of Scotland	GB	GBP	01.12.2012	AA**	0,3
65656	Химмашимпэкс	RUS	RUB	21.02.2010	B	4,0
54335	ООО "Издательство "ВОКРУГ СВЕТА"	RUS	RUB	17.06.2011	AA	0,3
45667	Волгоэлектромонтаж	RUS	RUB	18.01.2010	AA	0,3
32345	ЗАО "Воскресенские тепловые сети"	RUS	RUB	19.01.2010	2A	0,3
8468	ООО"Технострой" (филиал в Воронеже)	RUS	RUB	18.03.2010	B	4,0
75435	ГУП "Мосзеленхоз"	RUS	RUB	02.01.2011	AA	0,3
	ООО "Навигатор"	RUS	RUB	15.01.2011	CC	0,3
23411	Delaware Bay Company	US	USD	23.01.2007	B	4,0
64721	First Allied Securities	US	EUR	24.02.2010	B	104,0
98723	ЗАО "Первоуральский торговый дом"	RUS	RUB	26.08.2010	B	4,0
ASD43	ООО "Печатный Мир"	RUS	RUB	27.01.2010	B	4,0
83256	ОАО "Плутон"	RUS	RUB	28.01.2010		4,0
5185	ОАО "Полимербыт"	RUS	RUB	29.01.2010	B	4,0
51623	#INPUT ERROR	RUS	RUB	29.12.2010	B	-4,0
45	ООО "Провинция-2000"	RUS	RUB	30.01.2010	B	4,0
278	ООО"Технострой" (филиал в Москве)	RUS	RUB	14.01.2010	B	4,0
167	ЗАО "Топливно-бункерная компания"	RUS	RUB	01.02.2010	B	4,0
2086	ТОВ компанія УКРТОРГСЕРВІС	UKR	RUB	02.02.2010	B	4,0
52457	НПО Укроргсинтез	UKR	UAH	05.05.2010	1B	4,0
43344	ООО "Фудстар"	RUS	RUB	05.02.2010	B	4,0
23323	ООО "Химкомплект" (НЕ ИСПОЛЬЗОВАТЬ!)	RUS	RUB	27.07.2010	B	4,0
54545	ОАО "Химконверс"	RUS		07.02.2014	B	4,0
87434	ЗАО "Химмашимпэкс"	RUS	RUB	21.02.2010	B	4,0
76788	ОАО "Хлебный Дом"	RUS	RUB	12.02.2010	B	4,0

- Полнота
- Соответствие стандартам
- Взаимное соответствие
- Дублирование
- Связность и целостность
- Корректность

DataTalks



DataTalks

РАЗДЕЛ 2

Принципы управления качеством данных

- Основные принципы управления качеством данных
- Компоненты решения
- Цикличный процесс управления качеством

Ростелеком

Основные принципы управления качеством данных

Ростелеком

Для успешного управления качеством основных данных, организация должна выполнять следующие основные условия согласно ГОСТ Р – 201X/ISO/TS 8000-150:2011:

- Активное вовлечение в работу сотрудников на всех уровнях управления качеством данных
- Вовлечение в работу менеджеров самого высокого уровня управления качеством данных очень важно для оптимизации и повышения авторитета организации и всех происходящих в ней процессов
- Концентрация на точности данных измерений и на внесении поправок – это не достаточная мера для повышения качества данных организации; желаемое качество данных достигается в случае, если процесс управления качеством влияет и на сам источник данных

Основные принципы управления качеством данных

Ростелеком

Основные условия согласно ГОСТ (продолжение):

- Качество данных постоянно совершенствуется благодаря обработке, проверке измерений и постоянной корректировке данных
- Организации должны совершенствовать не только процессы управления данными, но и бизнес-процессы, в которых напрямую используются информационные данные
- Все процессы управления качеством основных данных совпадают или соответствуют автоматически контролируемым и проверяемым требованиям, обеспечивающим постоянный обмен основными данными между организациями и системами

Основные принципы управления качеством данных

Ростелеком

	Операции над данными	Непрерывный контроль качества	Усовершенст-вование качества
Управляющий данными	**Управление структурой данных**	**Планирование качества данных**	**Управление потоком данных**
Администратор данных	**Разработка и конструирование данных**	**Подготовка критериев данных**	**Анализ причин ошибок**
Управляющий данными	**Обработка данных**	**Оценка качества данных**	**Исправление данных**

DataTalks

Компоненты решения по управлению качеством

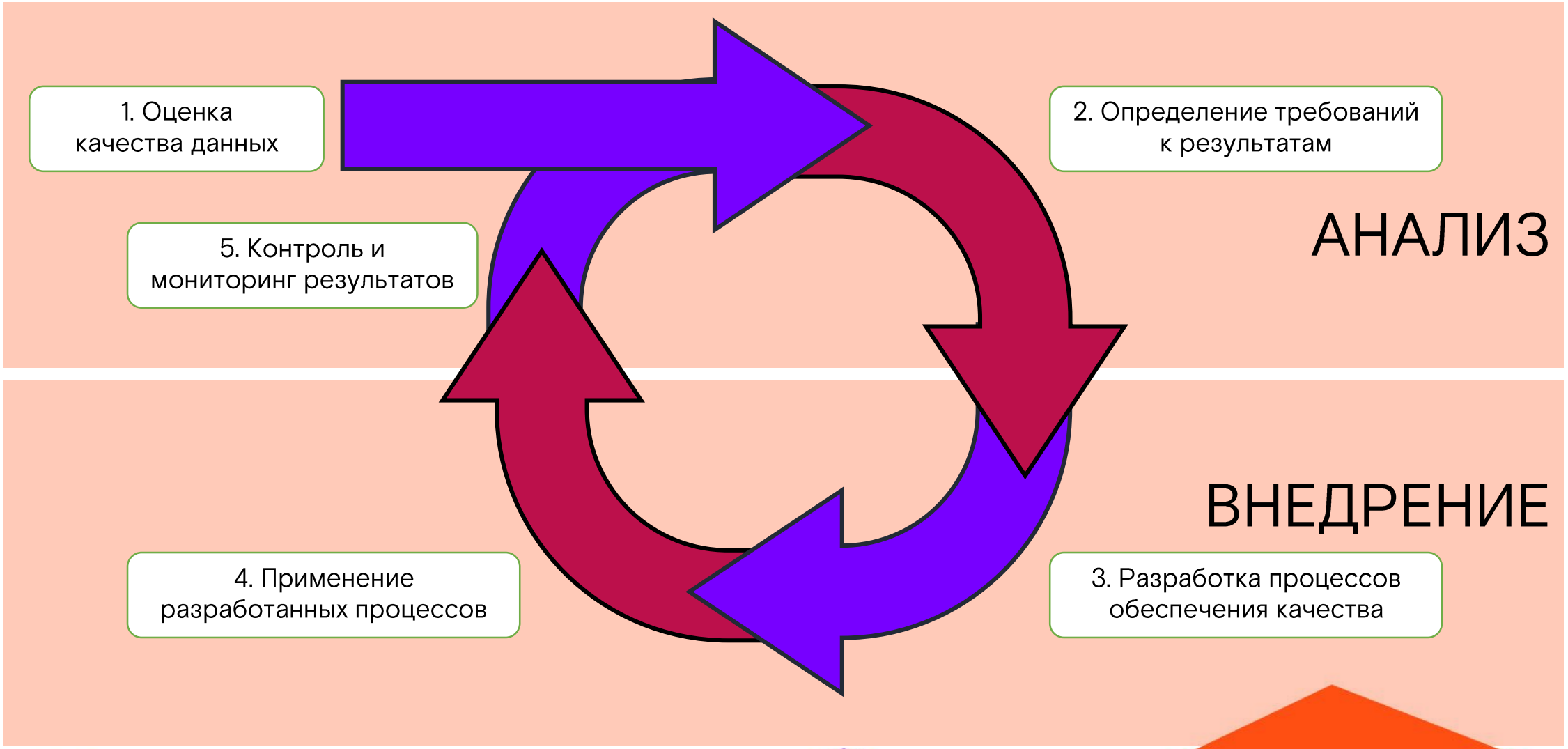
Ростелеком

Регламенты и стандарты	Распоряжения, приказы, порядки ведения работ, организационная структура
Сотрудники и исполнители	Внедрение и поддержка процессов обеспечения качества, учет бизнес-требований, мониторинг
Бизнес- и технические процессы	Построение бизнес-процессов, ведение справочников, данные в корпоративных системах
Платформы управления качеством	Обеспечение и мониторинг качества данных по всей организации на основе механизмов платформы

DataTalks

Цикличность процесса управления качеством

Ростелеком



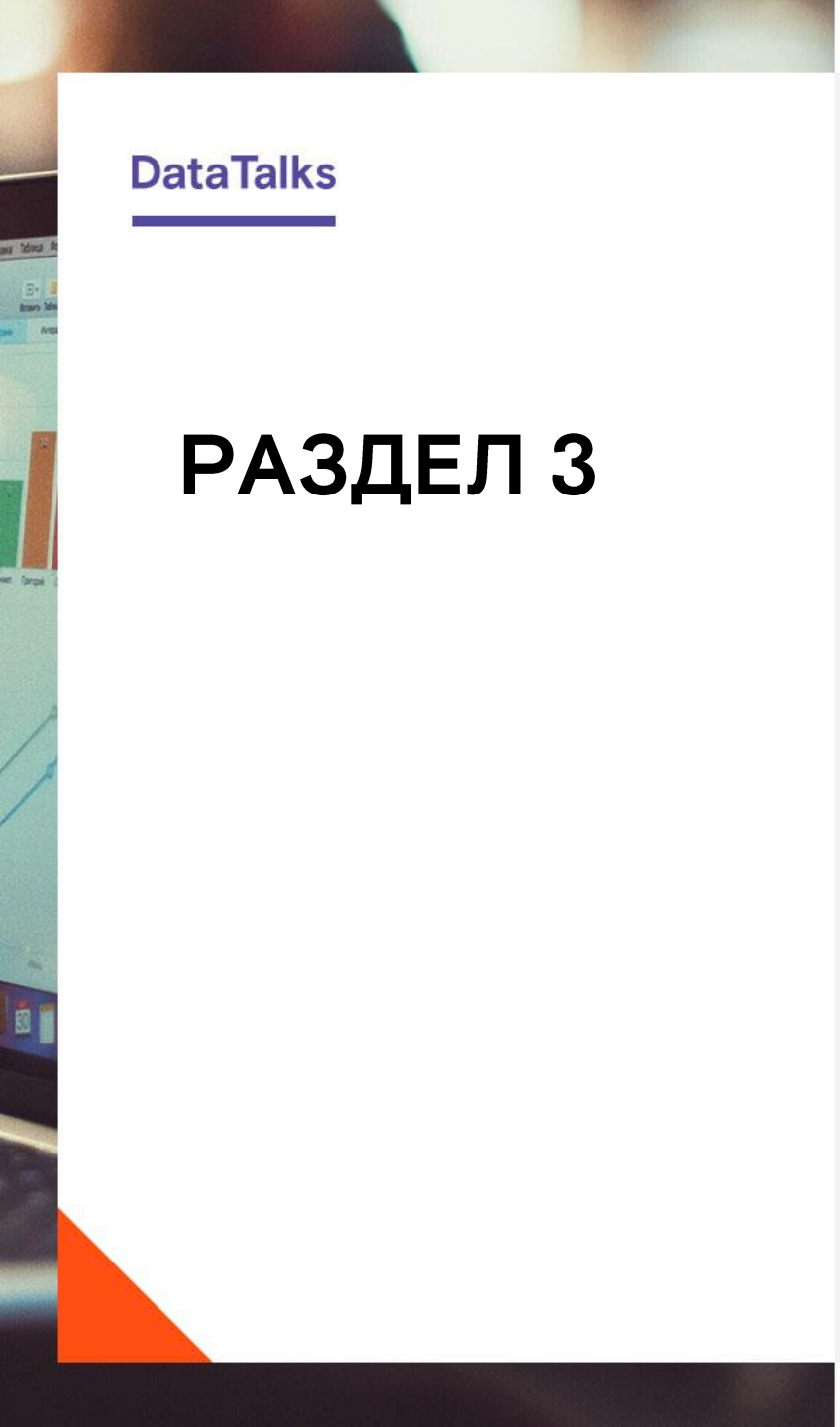
DataTalks

Процесс управления качеством

Ростелеком

Изучение и планирование	Разработка проекта и спецификаций	Внедрение
Детальное профилирование и изучение данных	Запуск проектов по очистке, консолидации и стандартизации данных	Встраивание процессов качества в ИТ-инфраструктуру и управление бизнес-процессами
• Оценка полноты и избыточности данных • Определение основных проблем и приоритетов • Разработка стратегии обеспечения качества данных	• Анализ для определения приемников очищенных данных и построения спецификаций • Выработка правил для автоматизированных процессов • Внедрение процессов для исключения ручных операций	• Проверка и очистка данных в реальном времени • ET_Q_L-процесс • Управление НСИ (MDM) • Интеграция с BI для построения отчетов • Управление правилами качества данных

DataTalks



DataTalks

РАЗДЕЛ 3

Инструменты управления качеством данных

- Функции средств обеспечения качества данных
- Профилирование данных
- Стандартизация данных
- Методы выявления дубликатов записей
- Мониторинг качества данных

Ростелеком

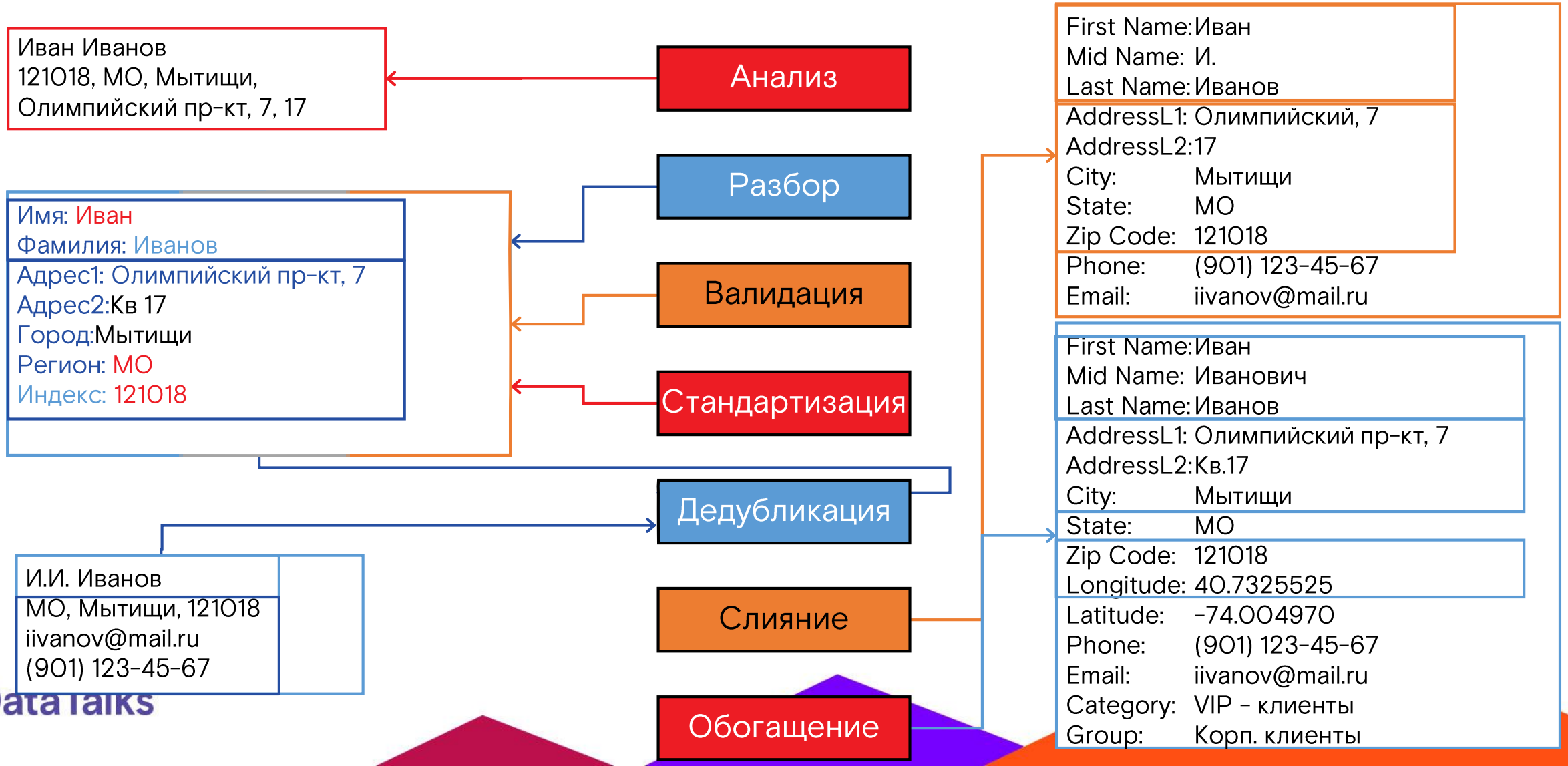
Основные инструменты по управлению качеством данных

Ростелеком

- Специализированные промышленные средства обеспечения качества данных
- Средства интеграции данных (ETL, ELT)
- Ручные операции

DataTalks

Основные функции решений по управлению качеством данных



DataTalks

Основные функции решений по управлению качеством данных

Ростелеком

Исходные данные

Данные низкого качества

Анализ

Профилирование	Оценка полноты, соответствия стандартам, взаимного соответствия
Сопоставление (matching)	Выявление дубликатов, поиск взаимозависимостей

Извлеченные данные

Обогащение данных

Стандартизация	Улучшение полноты, приведение к стандартам и взаимному соответствию
Консолидация	Удаление дубликатов, поиск зависимостей замена неточных данных

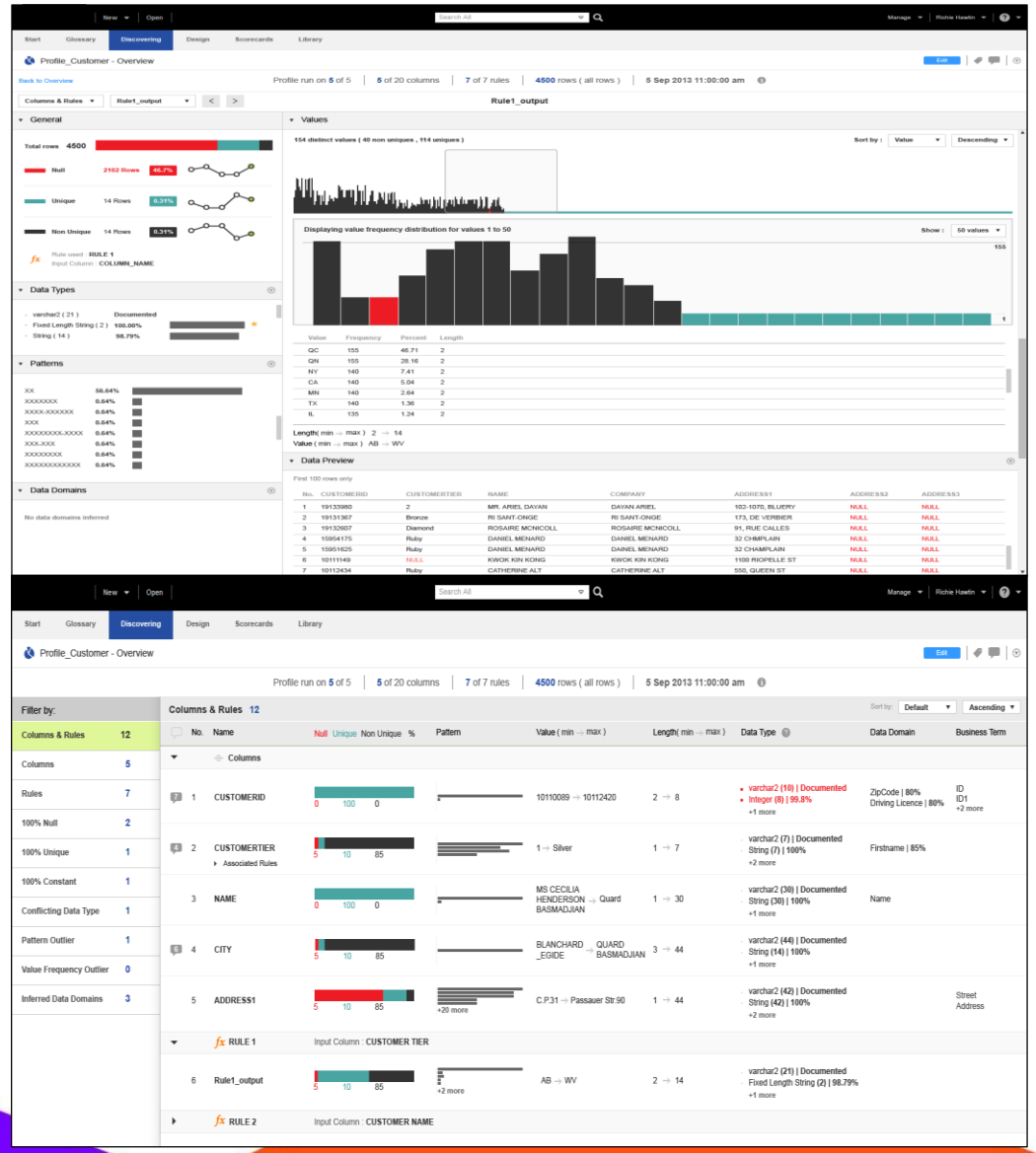
Отчеты по качеству

Выходные данные

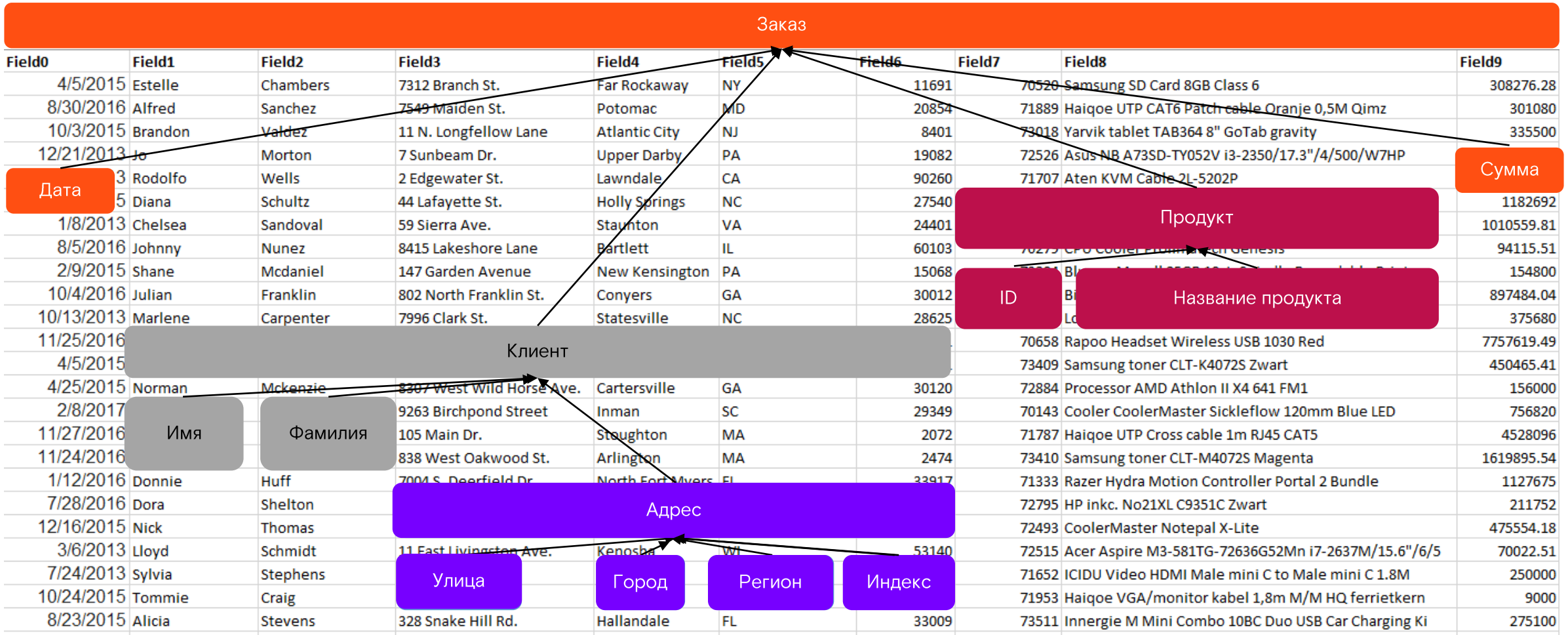
DataTalks

Профилирование данных

- Определение параметров данных
  - Статистика по полям, форматам
  - Определение уникальности, полноты, дубликатов значений, соответствия форматам и т.д.
- Структурный и функциональный анализ
  - Функциональные зависимости полей и записей
  - Расширенный анализ структур данных источников
  - Проверка целостности ссылочных данных
- Визуализация результатов профилирования
- Использование собственной логики для анализа качества данных



Понимание сути данных



Field0	Field1	Field2	Field3	Field4	Field5	Field6	Field7	Field8	Field9
4/5/2015	Estelle	Chambers	7312 Branch St.	Far Rockaway	NY	11691	70520	Samsung SD Card 8GB Class 6	308276.28
8/30/2016	Alfred	Sanchez	7549 Maiden St.	Potomac	MD	20854	71889	Haiqoe UTP CAT6 Patch cable Oranje 0,5M Qimz	301080
10/3/2015	Brandon	Valdez	11 N. Longfellow Lane	Atlantic City	NJ	8401	73018	Yarvik tablet TAB364 8" GoTab gravity	335500
12/21/2013	Jo	Morton	7 Sunbeam Dr.	Upper Darby	PA	19082	72526	Asus NB A73SD-TY052V i3-2350/17.3"/4/500/W7HP	
3	Rodolfo	Wells	2 Edgewater St.	Lawndale	CA	90260	71707	Aten KVM Cable 2L-5202P	
5	Diana	Schultz	44 Lafayette St.	Holly Springs	NC	27540			1182692
1/8/2013	Chelsea	Sandoval	59 Sierra Ave.	Staunton	VA	24401			1010559.81
8/5/2016	Johnny	Nunez	8415 Lakeshore Lane	Bartlett	IL	60103			94115.51
2/9/2015	Shane	Mcdaniel	147 Garden Avenue	New Kensington	PA	15068			154800
10/4/2016	Julian	Franklin	802 North Franklin St.	Conyers	GA	30012			897484.04
10/13/2013	Marlene	Carpenter	7996 Clark St.	Statesville	NC	28625			375680
11/25/2016							70658	Rapoo Headset Wireless USB 1030 Red	7757619.49
4/5/2015							73409	Samsung toner CLT-K4072S Zwart	450465.41
4/25/2015	Norman	Mckenzie	8307 West Wild Horse Ave.	Cartersville	GA	30120	72884	Processor AMD Athlon II X4 641 FM1	156000
2/8/2017			9263 Birchpond Street	Inman	SC	29349	70143	Cooler CoolerMaster Sickleflow 120mm Blue LED	756820
11/27/2016			105 Main Dr.	Stoughton	MA	2072	71787	Haiqoe UTP Cross cable 1m RJ45 CAT5	4528096
11/24/2016			838 West Oakwood St.	Arlington	MA	2474	73410	Samsung toner CLT-M4072S Magenta	1619895.54
1/12/2016	Donnie	Huff	7004 S. Deerfield Dr.	North Fort Myers	FL	33917	71333	Razer Hydra Motion Controller Portal 2 Bundle	1127675
7/28/2016	Dora	Shelton					72795	HP inkc. No21XL C9351C Zwart	211752
12/16/2015	Nick	Thomas					72493	CoolerMaster Notepal X-Lite	475554.18
3/6/2013	Lloyd	Schmidt	11 East Livingston Ave.	Kenosha	WI	53140	72515	Acer Aspire M3-581TG-72636G52Mn i7-2637M/15.6"/6/5	70022.51
7/24/2013	Sylvia	Stephens					71652	ICIDU Video HDMI Male mini C to Male mini C 1.8M	250000
10/24/2015	Tommie	Craig					71953	Haiqoe VGA/monitor kabel 1,8m M/M HQ ferrietkern	9000
8/23/2015	Alicia	Stevens	328 Snake Hill Rd.	Hallandale	FL	33009	73511	Innergie M Mini Combo 10BC Duo USB Car Charging Ki	275100

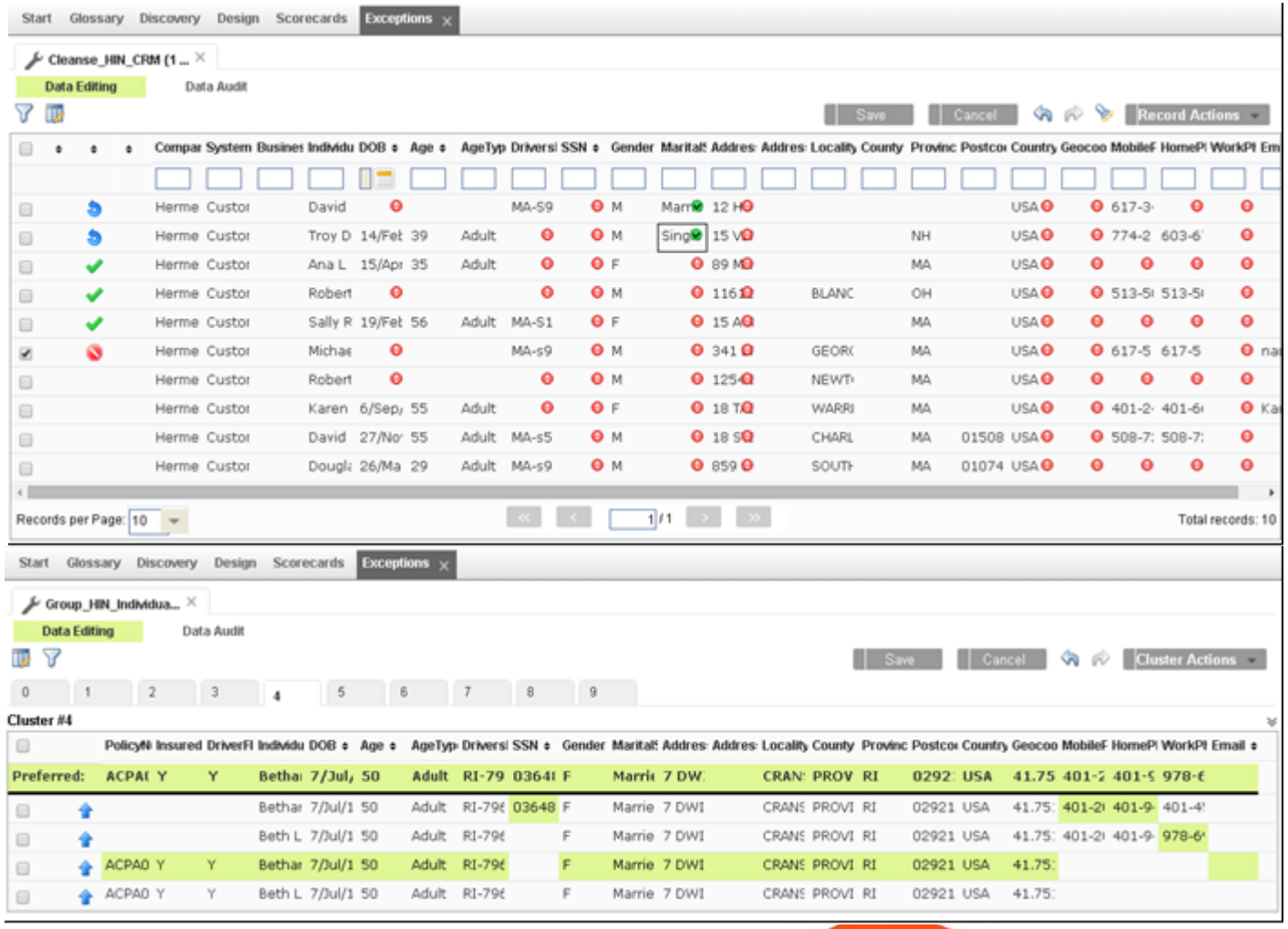
Основные методы очистки данных

Ростелеком

- Обогащение данных – дополнение несуществующих данных
- Стандартизация (нормализация) данных – приведение данных к виду и формату согласно требованиям
- Выявление дубликатов записей
- Слияние потенциальных дубликатов в одну запись – формирование «золотой» записи
- Разделение записей, ошибочно принятых за дубликаты

Обогащение данных

- Дополнение данных, отсутствующих в одном источнике, данными из другой системы
- Фактически, является функцией средства интеграции данных (ETL)
- Еще один вариант обогащения данных – внесение их вручную при отсутствии систем, где эти данные присутствуют



Стандартизация/нормализация данных

Ростелеком

Реализация:

- Разбор выражения на составные части (например, по словарям)
- Поиск замен по каждой найденной части в случае отсутствия прямого совпадения с правильным значением
- Составление нового выражения в нужном порядке из найденных замен

Контактное имя	Префикс1	Имя1	Фамилия1	Пре-фикс2	Имя2	Фамилия2	Обращение
Иванов Иван	Г-н	Иван	Иванов				Г-н Иванов
Иванов Иван	Г-н	Иван	Иванов				Г-н Иванов
Петров Петр, Федоров Федор	Г-н	Петр	Петров	Г-н	Федор	Федоров	Г-н Петров
Господин Иван Иванов	Г-н	Иван	Иванов				Г-н Иванов
Васильева__Вера	Г-жа	Вера	Васильева				Г-жа Васильева
Сидоров Иван Иванович	Г-н	Иван	Сидоров				Г-н Сидоров
Г-жа Петрова Елена	Г-жа	Елена	Петрова				Г-жа Петрова

До очистки

После очистки

DataTalks

Стандартизация/нормализация данных

Ростелеком

Полный адрес	Индекс	Город_ преф	Город	Ул_преф	Улица	Дом
111111, город Москва, Красная пл.,д.1	111111	г	Москва	пл	Красная	1
г.С.-Петербург,Невский проспект,10	211222	г	Санкт-Петербург	пр-кт	Невский	10
Москва Тверская 20	123456	г	Москва	ул	Тверская	20
222222,деревня Иваново, д.10	222222	д	Иваново	-	-	10
Москва г., Тверская ул,д.34,321456	321456	г	Москва	ул	Тверская	34

До очистки

После очистки

DataTalks

Методы выявления дубликатов данных

- Три основных метода:
  - Строгая логика – значения записей совпадают
  - Вероятностная логика (Jaro Distance, Edit Distance, Bigram, Soundex и многие другие) – значения похожи с учетом потенциального наличия ошибок в 1-2 символах, перестановки букв, слогов или слов
  - Нечеткая логика – механизмы нейронных сетей для выявления дубликатов на основе семантики и других механизмов
- Результаты сравнения можно интерпретировать как:
  - Высокая степень совпадения значений записей позволяет сделать вывод о том, что записи являются дубликатами
  - Низкая степень совпадения значений записей позволяет сделать вывод о том, что записи уникальны
  - В зоне от 70% до 85% совпадения значений (всегда зависит от состава данных) крайне высок процент ошибок при автоматическом принятии решения – это «серая зона»

«Нечеткая» логика выявления дубликатов

- Нечёткая логика — набор нестрогих правил, в которых для достижения поставленной цели могут использоваться радикальные идеи, интуитивные догадки, а также опыт специалистов, накопленный в соответствующей области
- Интеллектуальное индексирование и построение ключей и диапазонов поиска на основе многочисленных видов правил
- Технология сравнения данных (матчинга) на основе полученных индексов
- Главную роль в правильной работе нечеткой логики для процессов обеспечения качества данных играет Популяция – набор правил и показателей, влияющих на создание индексов
- Нечеткая логика позволяет иногда искать дубликаты записей без их предварительной очистки, если данные присутствуют

Код	Описание	Размер	Цена
AP-2199	Sailors' Desk Lamp	12 in	27.99
AP2199	Nautical Lamp	12 inch	27.99
PA-2119	Sailors Lamp	12 дюймов	34.99

Имя	Дата	Адрес	Город	Индекс
Карманов Сергей	03.01.1967	ул. Советская, д. 16, кор. 4, кв. 32	Ростов-на-Дону	352658
Карманов С.А.	03.01.1967	Советская, 16-4-32	Ростов	352 658
Серега Карманов	9/9/99	Донской район, Советская, 16/3	Р-н-Д	
Карманов Сергей Александрович	13/01/1967	Ростов-Дон, Советская, 16, кв. 32	Ростов-Дон	352-658
Sergey Karmanoff		16-3, app. 32 Sovetskaya st.	Rostov-on-Don	352658

Когда и какую логику сравнения использовать

Ростелеком

- Нечеткая логика дает прекрасный результат на ФИО, названиях чего-либо (продуктов, брендов и т.п.), но практически неприменима на данных типа номеров телефонов, email и части адресов (дома, корпуса, квартиры)
- Но, если популяция для работы нечеткой логики построена неверно, то даже явно разные записи могут оказаться похожими (например, Даниленко = Данильчук, что неверно)
- Вероятностная логика применяется для чисто строчных данных – названия, ФИО и т.п.
- Результаты работы вероятностной логики обычно ниже, чем у нечеткой, но нет зависимости от популяций
- Строгую логику использовать нужно только для строго форматированных значений, дат, времени, чисел
- Самая высокая производительность у механизмов строгой логики (если не сравниваются строки), а самые низкие – у вероятностной логики
- Чаще всего используются все типы логики последовательно или параллельно в зависимости от типов данных

Слияние записей потенциальных дубликатов

Ростелеком

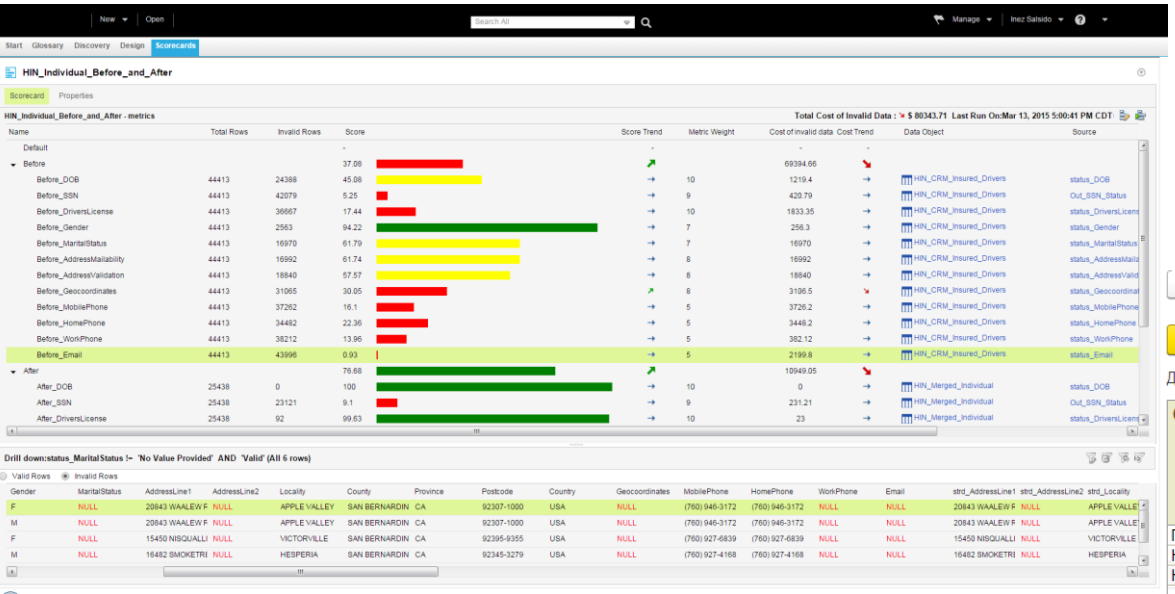
- Автоматическое слияние дубликатов зависит целиком от приоритетов:
  - Систем-источников исходных записей
  - Степени доверия к системам-источникам
  - Даты изменения данных
- Иногда по каждому атрибуту строится отдельное правило слияния
- Для работы с данными «серой зоны» крайне рекомендуется работа пользователя вручную

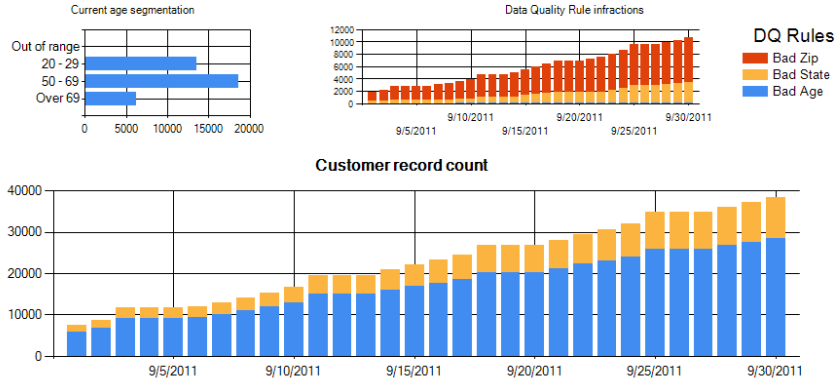
	Клиент	1. Клиент Оценка совпадения: 100	Просмотр слияния
Общее	текущая запись	Main_1	Кандидаты: 1
BVT для даты вступления в силу	18/авг/2016 00:00		18/авг/2016 00:00
Последний расчет BVT	18/авг/2016 19:17	24/май/2016 18:03	18/авг/2016 19:17
Rowid Object	20010	240001	20010
Фамилия	Иванов	Иванов	Иванов
Имя	Иван	Иван	Иван
Отчество	Иванович	Иванович	Иванович
Дата рождения	30/июл/1981	30/июл/1981	30/июл/1981
ИНН	810730312341	810730312341	810730312341
▸ Контакт	3	0	3

DataTalks

Мониторинг и отчетность по качеству данных

- Мониторинг качества данных – одна из основных функций процесса управления качеством
- Не должен зависеть от типа проверок качества
- Обычно ориентируются на качество исходных данных и после обработки средствами обеспечения качества
- Для построения полноценного процесса управления данными (Data Governance) нужна передача результатов проверок в средства работы с бизнес-глоссариями и техническими метаданными



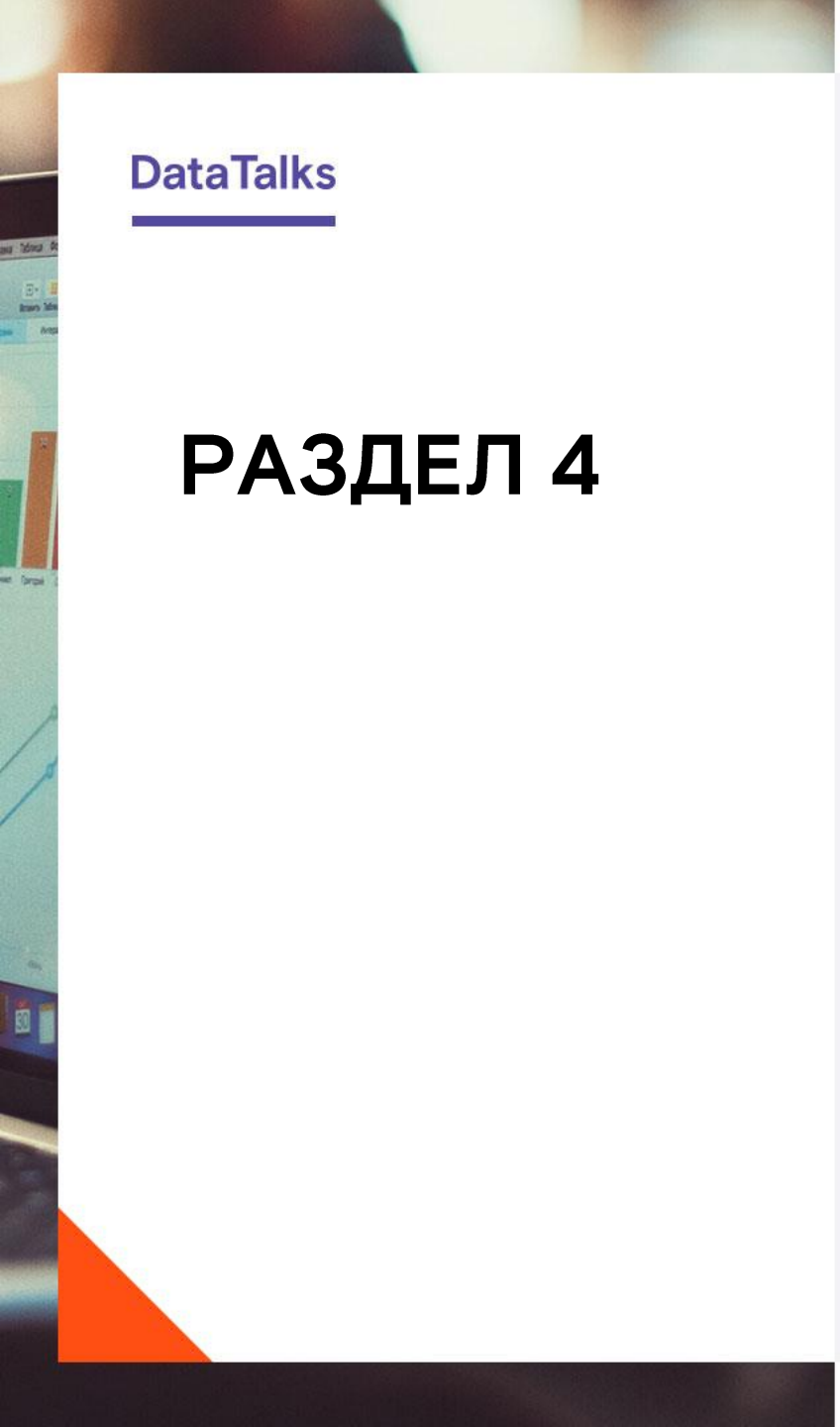


Отчет по качеству данных

Сформировать | Выбрать вариант... | Настройки...

Детализация:

Справочник			Не уникальные значения реквизитов		Проверка по ЕГРЮЛ		Прочие проверки		Пустые обязательные реквизиты	
	Общее количество записей	Проверено	Ошибки	% записей с ошибками	Ошибки	% записей с ошибками	Ошибки	% записей с ошибками	Ошибки	% записей с ошибками
Графики оплаты	1	1							1	100,00
Контрагенты	5	5	2	40,00	1	20,00			1	20,00
Номенклатура	4	4							4	100,00



DataTalks

РАЗДЕЛ 4

Основные виды проектов с применением технологий обеспечения качества данных

- Обеспечение качества данных для проектов построения хранилищ данных
- Качество данных при построении единых справочников
- Качество данных в проектах слияний и поглощений
- Проекты Data Governance и требования к качеству данных

Ростелеком

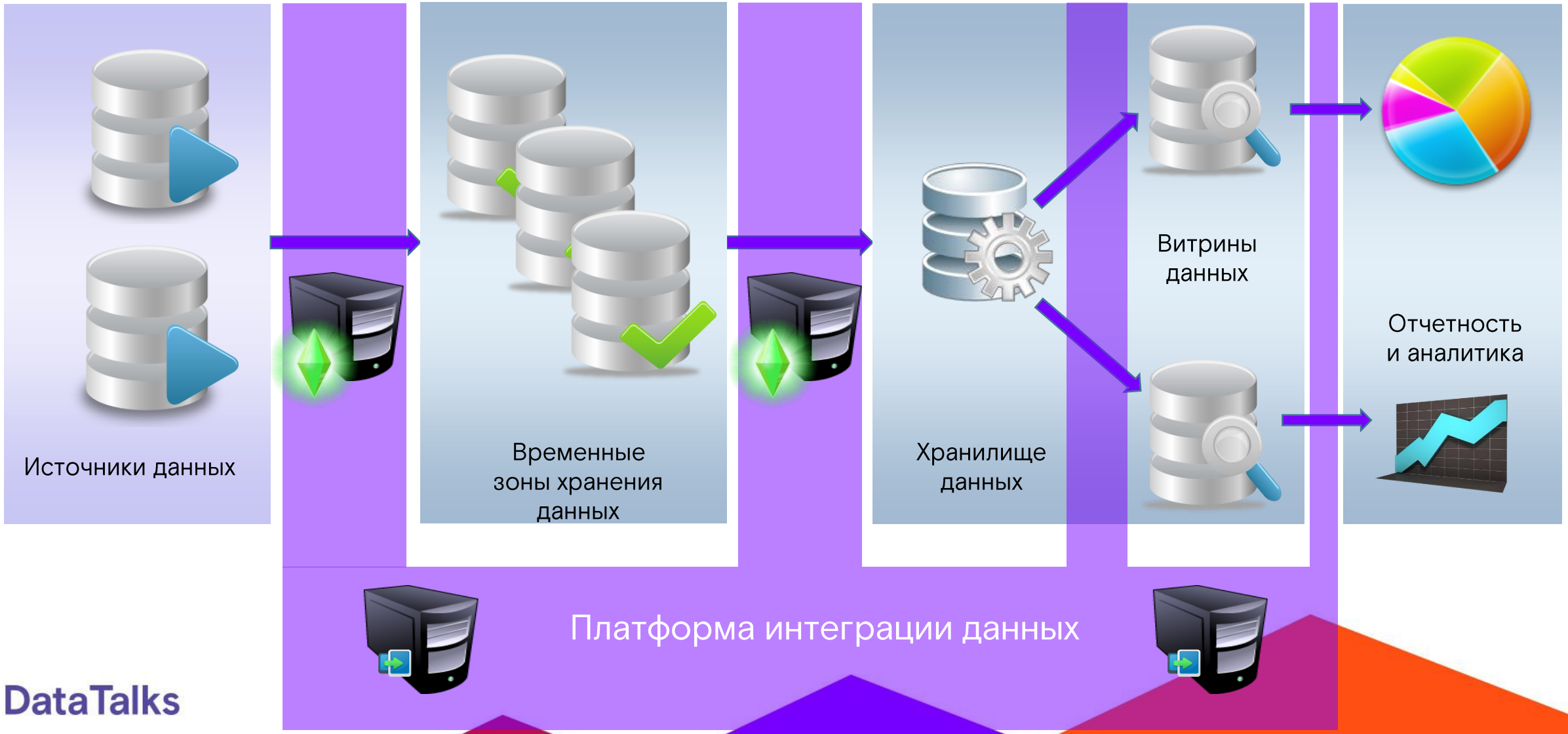
Типы проектов с обеспечением качества данных

Ростелеком

- Построение хранилищ данных
- Миграция данных при внедрении новых систем, новых версий систем или замене приложений
- Построение единых справочников (Master Data Management)
- Разовые проекты по очистке данных
- Онлайн-сервисы улучшения бизнес-качества (например, как часть процессов другого приложения)
- Проекты Data Governance

Качество данных в интеграционных проектах

Ростелеком

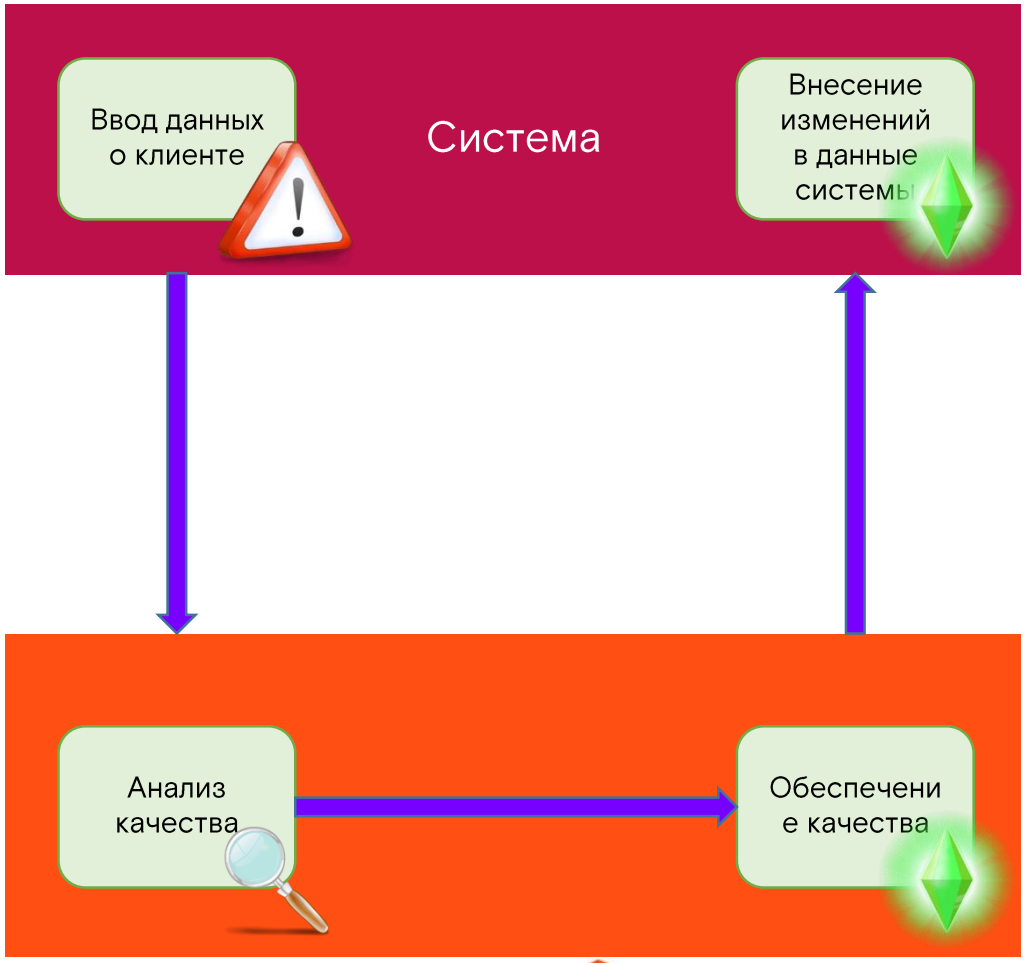


DataTalks

Качество данных в интеграционных проектах

Ростелеком

- При построении хранилища данных процесс обеспечения качества встраивается после выгрузки данных из источников
- Строится процесс ET-Q-L
- Формируется отдельная область качественных данных как источник для загрузки в хранилище данных
- Для онлайн-проектов проверка качества может осуществляться непосредственно при появлении данных



DataTalks

Пример проекта с обеспечением качества данных для хранилища и других систем

Ростелеком

ФИНАНСОВЫЙ СЕКТОР

Качественные данные для задач растущего Банка

DataTalks

БИЗНЕС-ЗАДАЧИ

- Быстрое слияние с поглощаемыми банками (объединение клиентских данных)
- Построение центральной управленческой отчетности на базе качественных данных
- Внедрение AML-системы (Anti-Money Laundering – против легализации доходов, полученных преступным путем)

РЕШЕНИЕ

- Использование решений по интеграции и обеспечению качества данных для построения хранилищ качественных данных для Казначейства, Корпоративного и Инвестиционного банков
- Внедрение мониторинга качества данных: очистка, стандартизация, обогащение и преобразование данных к единому формату

РЕЗУЛЬТАТ

- Объединение клиентов поглощенных банков
- Консолидации данных из всех систем, унификация справочной информации
- Успешное внедрение AML-системы благодаря наполнению качественными данными из хранилища

Качество данных при создании единых справочников

Ростелеком

Приложения

SAP
ORACLE
SIEBEL eBusiness
salesforce.com Success On Demand.

Унаследованные приложения

Legacy Systems

Внешние данные

D&B
ims
REUTERS

1 Интеграция данных

2 Анализ данных

3 Обеспечение качества

4 Система Master Data Management

Операционные

Приложения

Унаследованные приложения

Аналитические

Portal/ Dashboard

Витрины данных

Хранилище данных

Business Intelligence

DataTalks

Создание единых справочников – модель зрелости

Ростелеком

Источник: Forrester

DataTalks

Пример проекта построения мастер-справочников

Ростелеком

ФИНАНСОВЫЙ СЕКТОР

Повышение эффективности продаж

DataTalks

БИЗНЕС-ЗАДАЧИ

- Реализация новой системы управления эффективностью продаж (Sales Performance Management)
- Создание в новой системе актуальных мастер-данных по основным бизнес-объектам, регулярно используемым большим количеством бизнес-процессов
- Минимизация ошибок сотрудников и дополнительных затрат

РЕШЕНИЕ

- Сбор изменений мастер-данных из разных систем Банка с помощью средств интеграции данных
- Очистка, дедубликация и унификация данных с помощью средств обеспечения качества данных
- Создание унифицированной «золотой записи» мастер-данных и передача ее назад в системы-источники с помощью системы класса Master Data Management

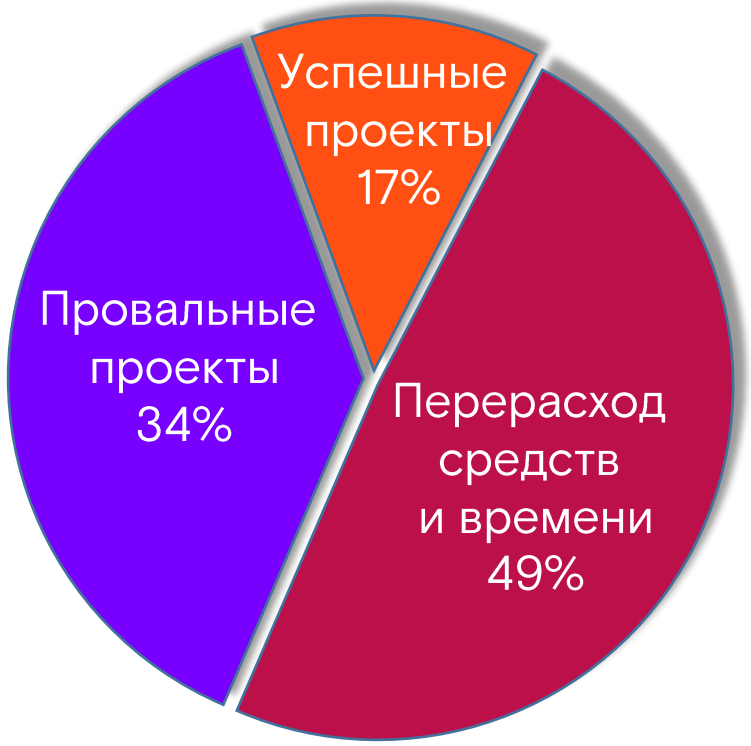
РЕЗУЛЬТАТ

- Сокращение издержек, связанных с неточностями в данных
- Уверенность пользователей данных в верной и актуальной информации
- Поддержка текущих бизнес-инициатив Банка
- Готовность масштабирования решения на другие системы Банка

Качество данных при слияниях и поглощениях

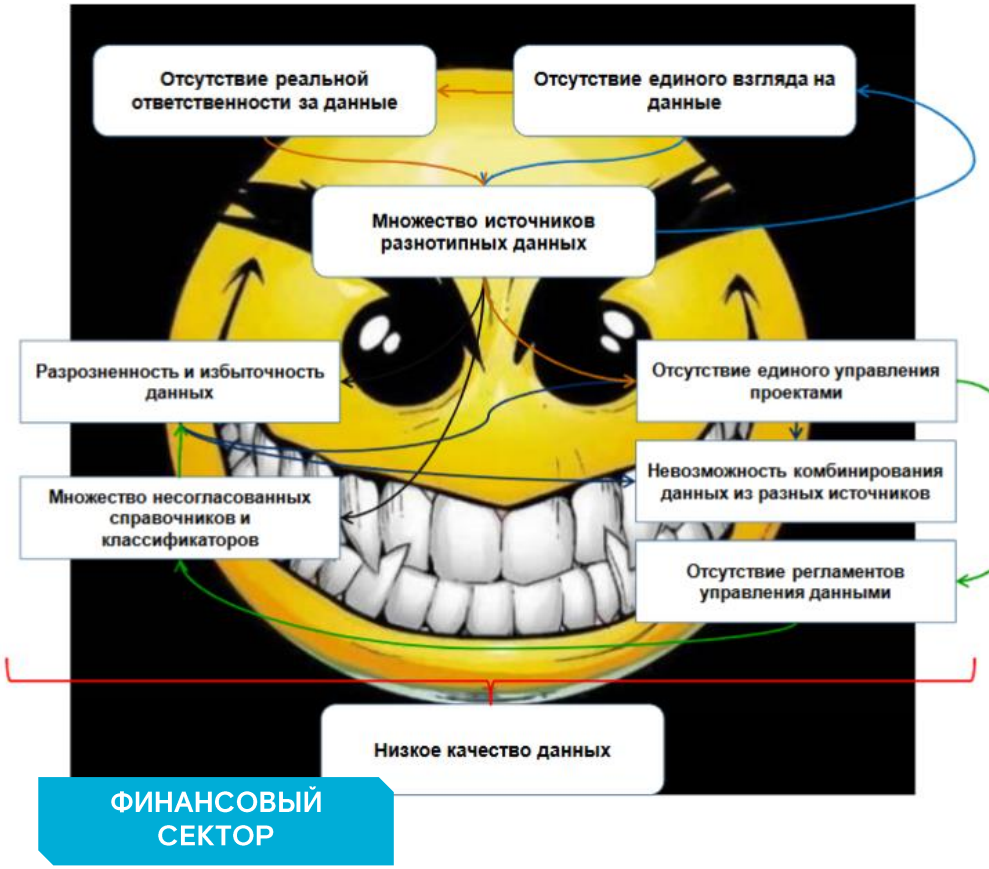
Ростелеком

- Самый частый вид проекта – миграция данных
- Для успешного перемещения данных в системы нужно:
  - Подготовить справочные данные согласно требованиям собственных систем
  - Выстроить логику преобразования данных с учетом особенностей форматов и представлений данных в старых системах
  - Найти все исключения, которые приведут к падению процессов миграции
  - Построить процессы валидации загруженных данных в новые системы



Источник: *Standish Group*

Пример проекта подготовки данных для миграции



ФИНАНСОВЫЙ СЕКТОР

БИЗНЕС-ЗАДАЧИ

- Подготовить клиентские данные к миграции при объединении банков

РЕШЕНИЕ

- Анализ систем на наличие и местонахождение клиентских данных
- Очистка, дедубликация и унификация данных
- Создание единой клиентской записи в составе справочника физических лиц

РЕЗУЛЬТАТ

- За несколько месяцев полностью создан единый справочник клиентов
- Банк затем успешно влился в состав другого банка

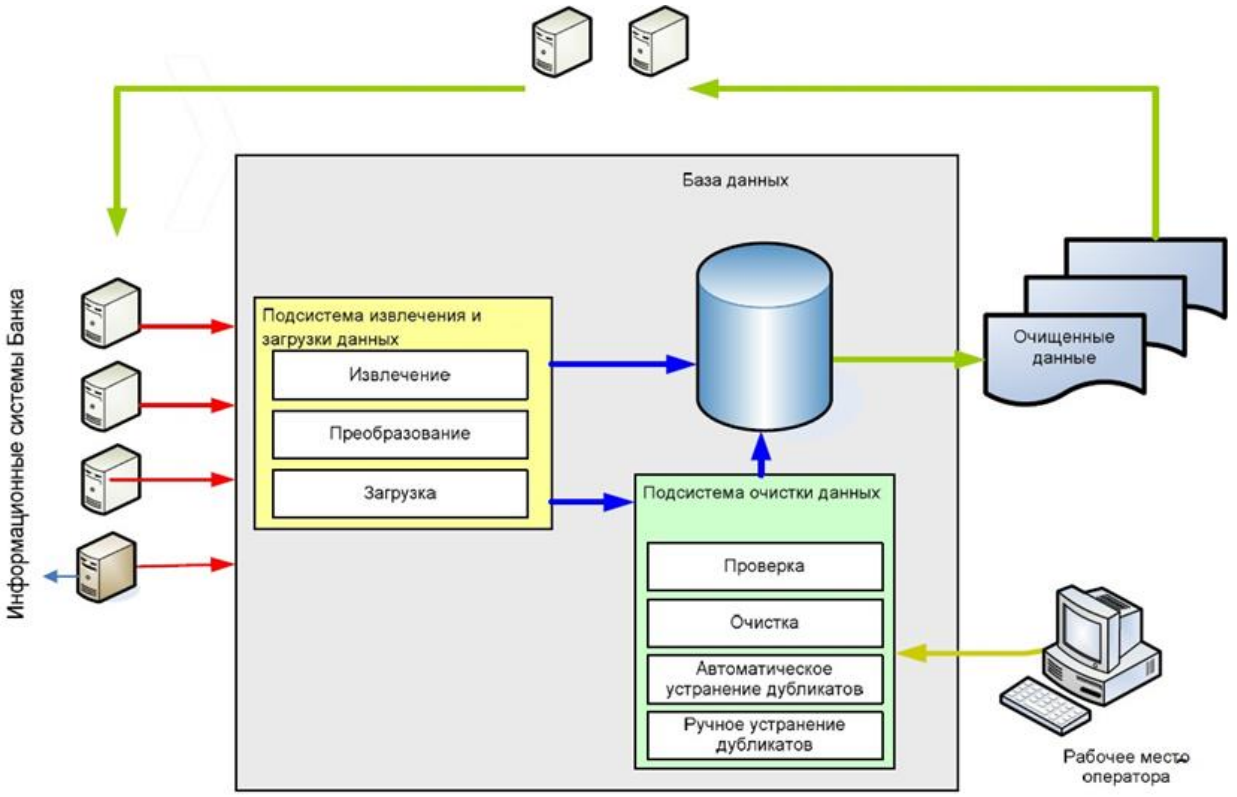
Подготовка к успешному поглощению другим банком

DataTalks

Качество данных для выявления мошенничества

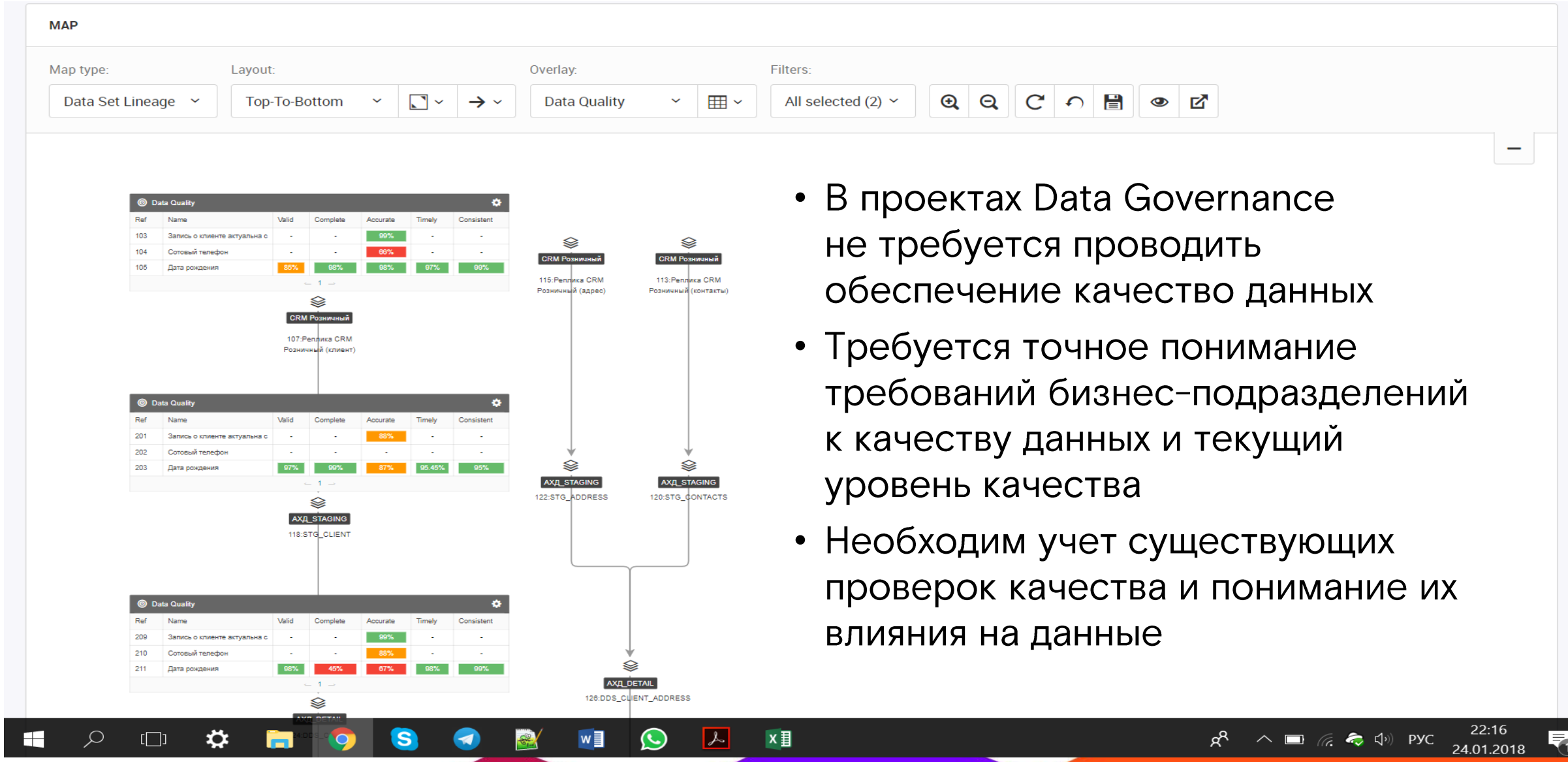
Ростелеком

- В проектах выявления отмывания денежных средств и финансирования терроризма обязательно требуется функциональность Know Your Customer – качество клиентских данных должно быть максимальным
- Для выявления случаев финансового фрода крайне важны механизмы вероятностной и нечеткой логики для поиска контактной информации по физическим и юридическим лицам, которые прибегают к незначительному изменению фамилий или наименований, чтобы избежать последствий
- Большинство подобных проектов работают в режиме реального времени



DataTalks

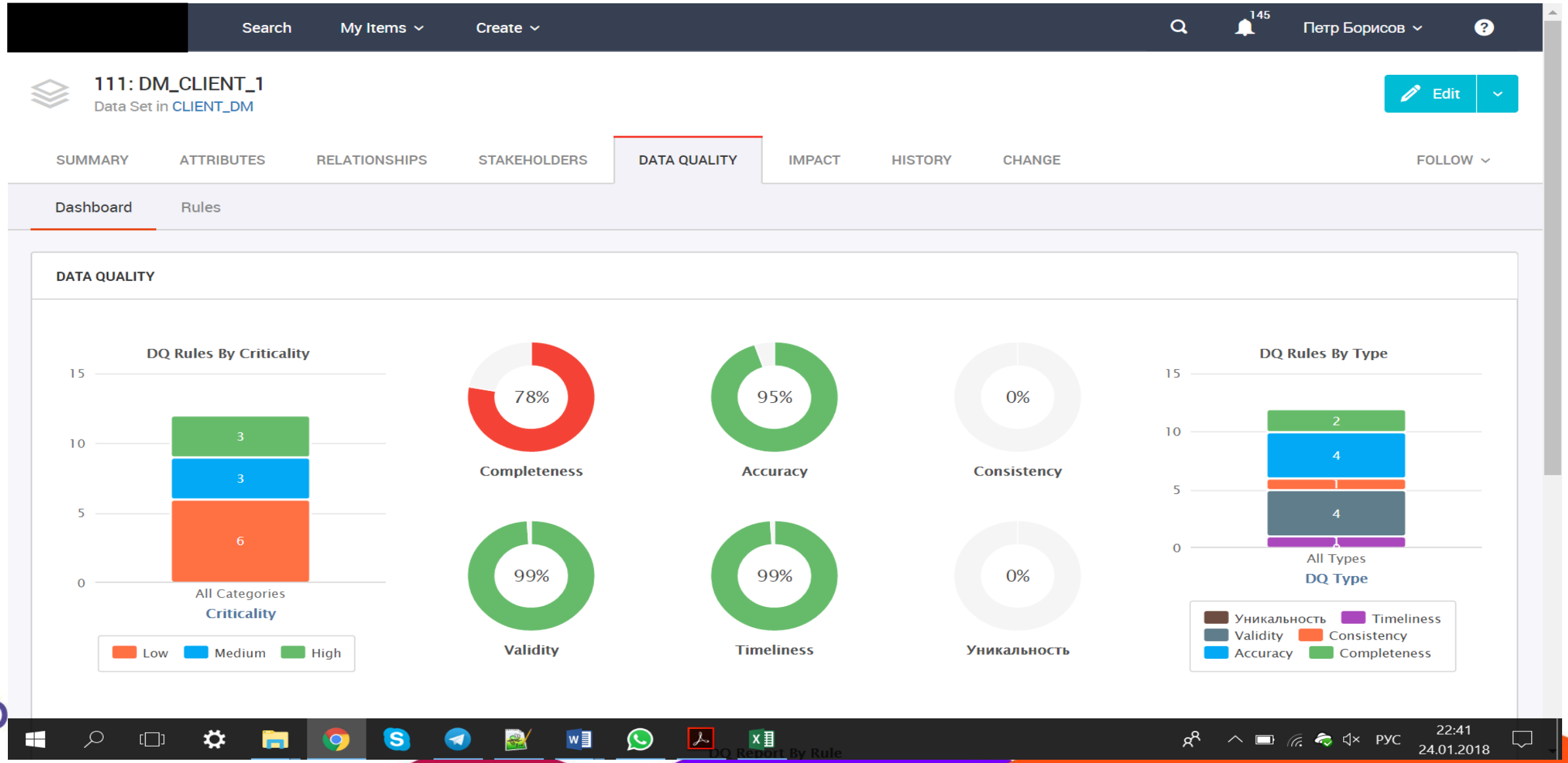
Качество данных для проектов Data Governance



- В проектах Data Governance не требуется проводить обеспечение качество данных
- Требуется точное понимание требований бизнес-подразделений к качеству данных и текущий уровень качества
- Необходим учет существующих проверок качества и понимание их влияния на данные

Качество данных для проектов Data Governance

Ростелеком



Пример проекта Data Governance с учетом качества данных

Ростелеком

ФИНАНСОВЫЙ СЕКТОР

Повышение качества и управляемости данных

БИЗНЕС-ЗАДАЧИ

- Повышение качества данных и их управляемости для использования информации отделами рисков и финансов
- Устранение несогласованности данных
- Расширение возможности совместного использования данных и управление изменениями

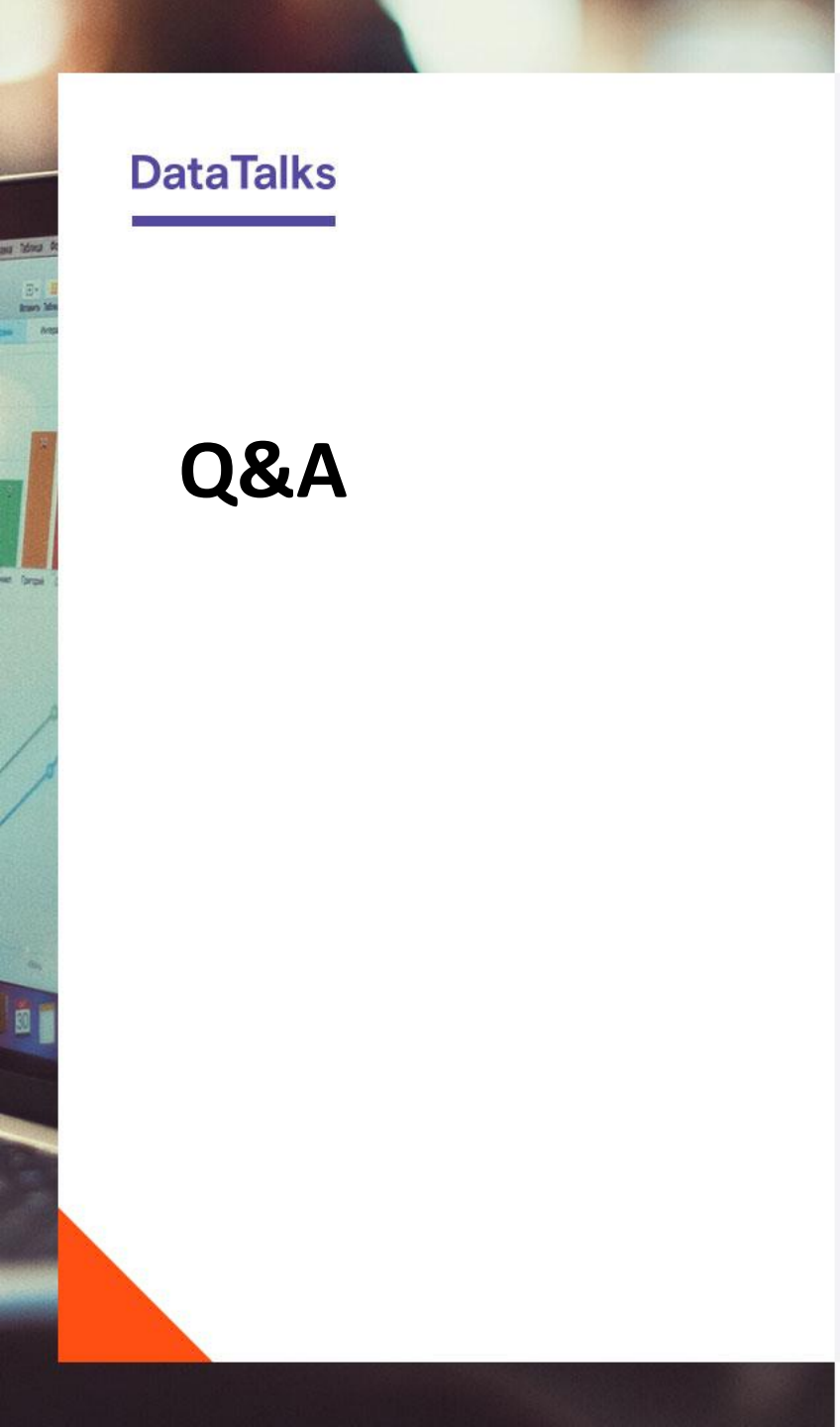
РЕШЕНИЕ

- Установлена и визуализирована связь между показателями форм регуляторной отчетности и отчетами МСФО
- Составлен бизнес-глоссарий терминов по показателям отчетов, проведен разбор технических метаданных и построен data lineage по указанным отчетам
- Выполнена проверка качества данных по показателям

РЕЗУЛЬТАТ

- Создано рабочее решение, позволяющее проводить оценку качества информации, согласование терминологии, построить взаимосвязи метаданных по всем системам Банка

DataTalks



DataTalks

Q&A

Вопросы и ответы

- Основные термины
- Проблемы качества данных
- Принципы управления качеством данных
- Основные методы обеспечения качества данных
- Проекты с обеспечением качества данных

Ростелеком