

Формирование единого информационного пространства организации

SAS Data Management Platform

Ларичев Глеб, 25.09.2014



70% ресурсов при внедрении аналитических систем тратятся на подготовку данных необходимой полноты и качества



Каким бы не был
Совершенным
аналитический
механизм,
плохие данные
не позволят
добиться
требуемого
качества анализа



- Отсутствие стандартизованного представления данных в системах
- Неоднозначность данных (Наличие дубликатов)
- Неполнота данных (Необходимость обогащения)



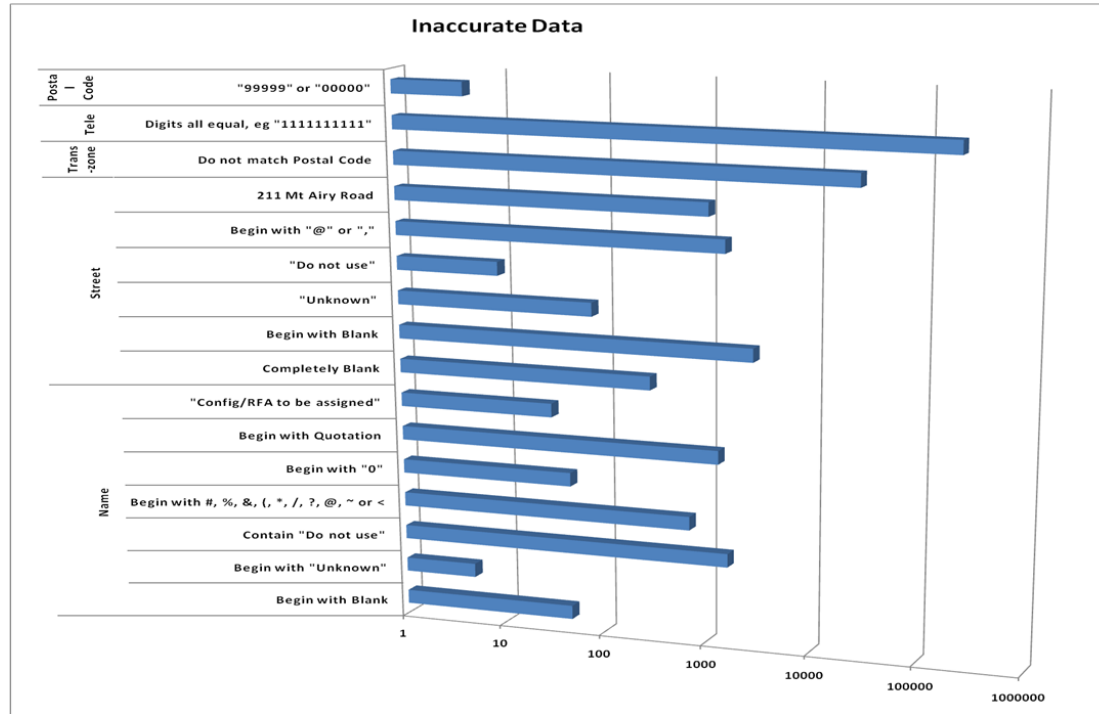
- Нет единых стандартов регламентирующих требования к дан
- Отсутствие единого представления о качестве данных в организации
- Путаница в том, кто несет ответственность за данные и их изменение



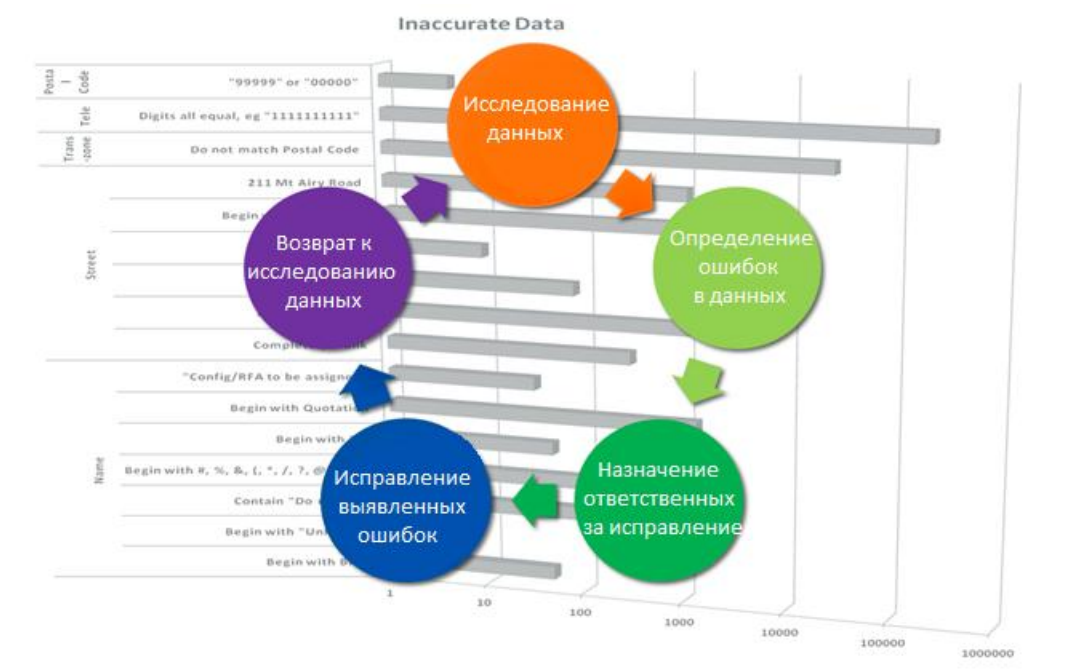
Обеспечение качества данных



ТРАДИЦИОННЫЙ ПОДХОД К РЕШЕНИЮ ЗАДАЧ КАЧЕСТВА ДАННЫХ



ИТЕРАЦИОННЫЙ ПОДХОД К РЕШЕНИЮ ЗАДАЧ КАЧЕСТВА ДАННЫХ





Анализ и исследование данных:

- Инструменты анализа числовых данных (статистика, выбросы, пропущенные значения и др.);
- Инструменты анализа текстовых данных (профилирование, анализ взаимосвязей, нечёткий поиск связей и др.).

Автоматизированная очистка “грязных” данных:

- Локализация алгоритмов для России;
- Мощные, настраиваемые инструменты очистки текстовых данных (грамматики, словари, фонетика, дедупликация и др.);
- Детерминированные алгоритмы – повторяемый результат.

Обогащение данных и аналитика:

- Обогащение данных из внешних источников – для России поддержка КЛАДР-ФИАС, база телефонных кодов, СПАРК, возможность подключения других источников;
- Обогащение данных на основе углублённого анализа (определение пола, родственных связей и др.).

Настраиваемые бизнес-процессы ручной очистки и согласования исправлений в данных.

Мониторинг и контроль качества данных:

- Настраиваемые правила, визуальные инструменты;
- Возможность проверки на этапе ввода данных в транзакционные системы.

The screenshot displays the SAS Data Quality interface. In the background, a large circular network diagram shows numerous blue nodes connected by light blue lines, with a few green lines highlighting specific connections. Overlaid on this is the 'Identification Analysis' window. On the left of this window is a tree view of 'All Identification Definitions' with categories like Address/City, Date, and Individual/Organization. The 'Province' category is selected. The main area of the window shows 'Identity matches' for the 'Province' field. A blue text box with white text is overlaid on the table, and another blue text box is at the bottom right of the window.

Поиск полей с одними Бизнес-данными

Field Name	Table	Databa...	Length	Type	Percen...	Method
PROV	CAMPIONE	SAS Feder	2	Character	96 %	Sample dat
PROV	anagrafica_normaliz	MySQL	2	Character	98 %	Sample dat
Province	CONTACTS	Contacts	255	Character	98 %	Sample dat
Province	FINDATA	Findata	255	Character	97 %	Sample dat
PROVINCIA_DI_NASCITA	anagrafica_normaliz	MySQL	255	Character	100 %	Sample dat
ref_province	anagrafica_normaliz	MySQL	2	Character	100 %	Sample dat

Identity match count: 6

Анализ связей на основе контекста, включая нечёткий поиск

SAS Data Quality ПРОФИЛИРОВАНИЕ ДАННЫХ

e_mail Table: STG_CUSTOMERS_OPN Schema: dbo Data Source: sas_cdi

Column Profiling	Frequency Distribution	Pattern Frequency Distribution	Percentiles	Outliers	Primary Key/Foreign Key Analysis	Notes
Pattern	Alternate	Count	Percentage			
9	9	73460	93,74			
Other		2337	2,98			
9999999999	9(11)	1729	2,21			
AAA	A(3)	451	0,58			
9999-999-99-99	9(4)-9(3)-9(2)-9(2)	256	0,33			

По результатам профилирования были найдены телефонные номер в полях для хранения e-mail

9999999999	9(12)	
aaaaaaaaaaaa	a(9) a(4).	
9999999 A	9(7) A	
aaaaaaaaaaaaaaa	a(11) a(3)	
999999999999, aa	9(11), a(2)	
9-999-99-99-999	9-9(3)-9(2)-9(2)-9(3)	
a	a	
AAAAAA AAAA.AA	A(6) A(4).A(2)	
+9999-99999999	+9(4)-9(7)	
+9-999-999-99-9	+9-9(3)-9(3)-9(2)-9	
9 999 99 99 999	9 9(3) 9(2) 9(2) 9(3)	
99999999 aaaaaaa	9(7) a(7)	
aaaaaaaa aaaaaaa	a(7) a(7)	
999999999999,aaa	9(11),a(3)	
aaaaaa9 aaaaaa.	a(6)9 a(6).	
aaaaaa aaaa.aa	a(5) a(4).a(2)	
A.999-99-99	A.9(3)-9(2)-9(2)	
+999999999	+9(8)	
aaaaaa aaaaaaaa.a	a(5) a(7).a	
aaaa-99 aaaa.aa	a(4)-9(2) a(4).a(2)	
a-aaaaaaaa aa.aa	a-a(7) a(2).a(2)	
AAAAA	A(5)	

e_mail Table: STG_CUSTOMERS_OPN Schema: dbo Data Source: sas_cdi

Column Profiling	Frequency Distribution	Pattern Frequency Distribution	Percentiles	Outliers
Metric Name	Metric Value			
Ordinal Position	46			
Count	2140868			
Null Count	2062500			
Percent Null	96.3			
Blank Count	0			
Minimum Value	(632)2906959			
Maximum Value	энерг			
Mode	0			
Pattern Count	313			
Unique Count	4298			
Uniqueness	5.48			
Primary Key Candidate	no			
Data Type	varchar			
Data Length	250 chars			
Actual Type	integer (97%)			
Minimum Length	1			
Maximum Length	15			

Неверный формат e-mail

Профилирование – анализ существующих данных с целью определения их пригодности для использования в планируемом бизнес процессе

Определение корректности ИНН

Описание	Анализируя длину и контрольное число, определяет корректность ИНН	
Примеры	Входные Данные	Результат
	96661014540	Ошибка ИНН
	504901209231	ИНН ИП

Определение корректности адреса электронной почты

Описание	Анализируя шаблон электронной почты, определяет её корректность	
Примеры	Входные Данные	Результат
	maria.v.ermakova@gmail.com	ОК
	Ptich@	Ошибка
	7717032288	Ошибка

Адрес			
Описание	Результат Анализа		
Пример	Входные данные	Результат	
	г. Заволжье, ул. Пономарева, д. 1, кв. 12	Почтовый Индекс	606524
		Область	Нижегородская
		Район	Городецкий
		Город	Заволжье

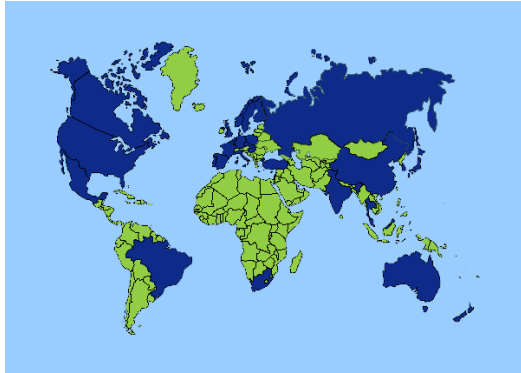
Телефон						
Описание	Определение типа телефона, стандартизация номера, определение кода КЛАДР					
Пример	Входные данные	Результат				
	Телефонный номер	Код страны	Код города	Номер	Код КЛАДР	Тип
	7-343-3710589	8	343	3710589	66000000000000	Дополнительный
	8-985-1238825	8	485	1238825	77000000000017	Мобильный
	7-495-1354184	8	499	1354184	77000000000012	Дополнительный

Определение пола		
Описание	Анализируя имя, определяет пол	
Примеры	Входные Данные	Результат
	Пафнутий Львович Чебышев	М
	Софья Ковалевская	Ж

Человек / Организация		
Описание	Определяет, представляет ли запись человека или организацию	
Примеры	Входные Данные	Результат
	Федор Достоевский	ЧЕЛОВЕК
	ОАО У Швейка	ОРГАНИЗАЦИЯ
	ООО Промкнигторг	ОРГАНИЗАЦИЯ

Match-код – уникальный идентификатор записи обладающий регулируемой чувствительностью и используемый для поиска дубликатов записей, формирования кластеров похожих записей, а также для нечёткого поиска.

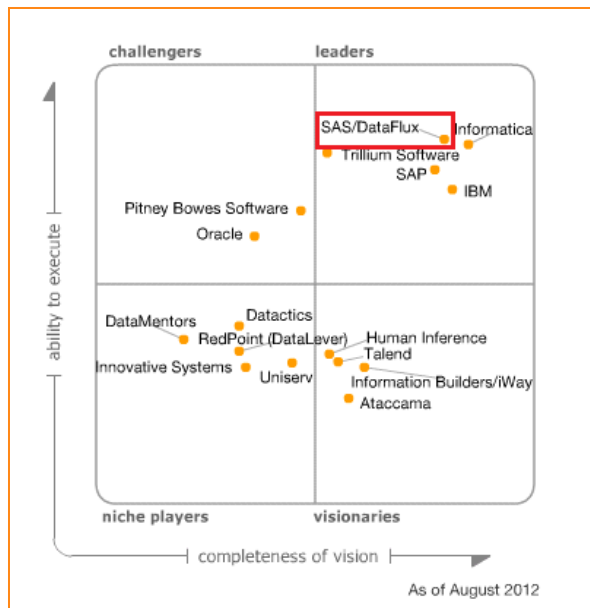
ФИО			
Описание	Генерация match-кода для ФИО		
Пример	Входные данные	Результат (чувств. 75)	Результат (чувств. 85)
	Ермакова Мария Владимировна	ЙРМХФА\$\$\$\$\$МША\$\$\$\$\$\$\$\$\$	ЙРМКБА\$\$\$\$\$МША\$\$\$\$\$ВЛДМ
	Ермакова Маша	ЙРМХФА\$\$\$\$\$МША\$\$\$\$\$\$\$\$\$	ЙРМКБА\$\$\$\$\$МША\$\$\$\$\$\$\$\$\$
	Ермакова Мария Владимир	ЙРМХФА\$\$\$\$\$МША\$\$\$\$\$\$\$\$\$	ЙРМКБА\$\$\$\$\$МША\$\$\$\$\$\$\$\$\$



Quality Knowledge Base (QKB) – предоставляет **правила, словари, грамматики, фонетику и процессы** для автоматизации процесса обеспечения качества данных:

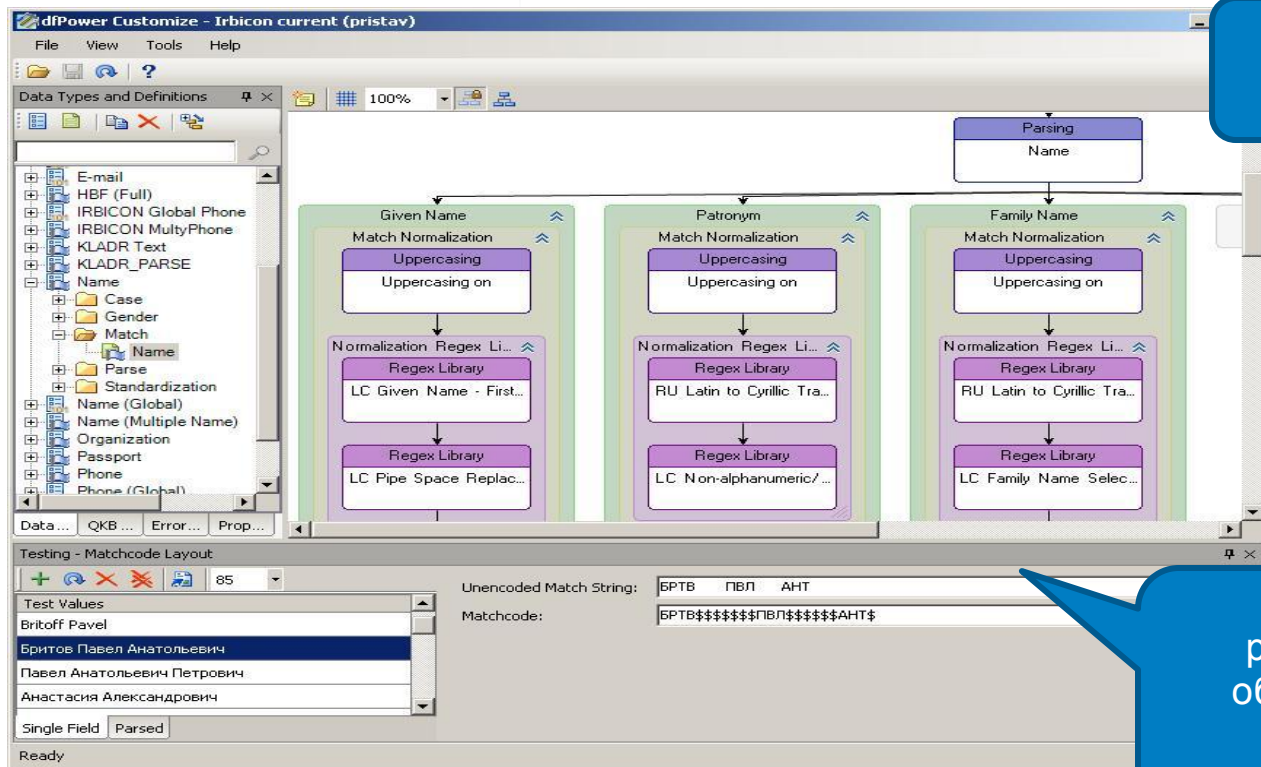
- Стандартизация
- Грамматический разбор (парсинг)
- Проверка шаблонов написания
- Построение кодов нечеткого поиска
- Идентификационный анализ (Пол, Юрилицо/Физлицо, тип ДУЛ и пр.)
- Выделение значимой части из строки текста

QKB является **открытой, расширяемой кастомизируемой** базой знаний и предоставляет **готовое решение** для информации о клиенте или контрагенте, закупках и пр., включает локализацию для множества языков и более 30 стран.



GartnerDQ Tools 2008 - 2012

- С 2007 года существует и развивается локализация базы знаний QKB для русского языка
- База содержит более 1/2 миллиона имен и фамилий, около 150 тысяч географических названий в РФ (по КЛАДР), десятки правил грамматического и фонетического разбора и пр.
- Реализованы категории: ФИО, паспорт, адрес (короткий, полный), телефоны, название организации и пр.
- Обогащение по базам адресов, номерных ёмкостей операторов, замены номерных ёмкостей и пр.
- Реализован анализ и корректировка отдельных латинских букв в русских словах (C→С, A→А, и пр.)
- Реализованы проверки кодов ИНН, ОГРН, ОГРНИП, и др., кросс-проверки адреса, телефонов и пр.

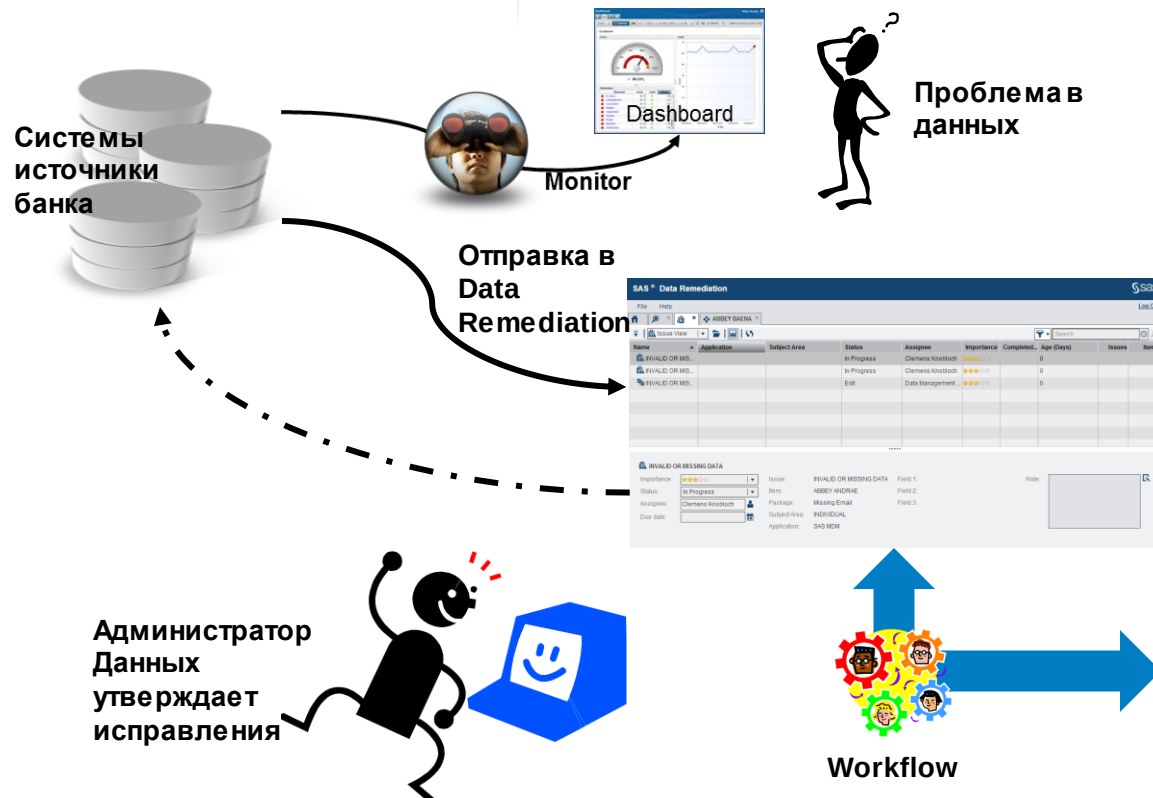


База знаний не является
«черным ящиком» !

Встроенные инструменты
редактирования алгоритмов
обработки данных и создания
методов для новых типов
обрабатываемых данных

Бизнес-процесс ручной корректировки и согласования данных

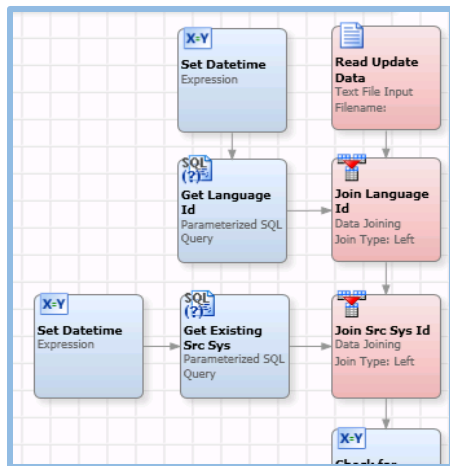




- ✓ Улучшенное взаимодействие бизнеса и IT
- ✓ Быстрое разрешение проблем
- ✓ Центральный репозиторий для фиксации проблем в данных
- ✓ Контроль через цепочки согласования и утверждения

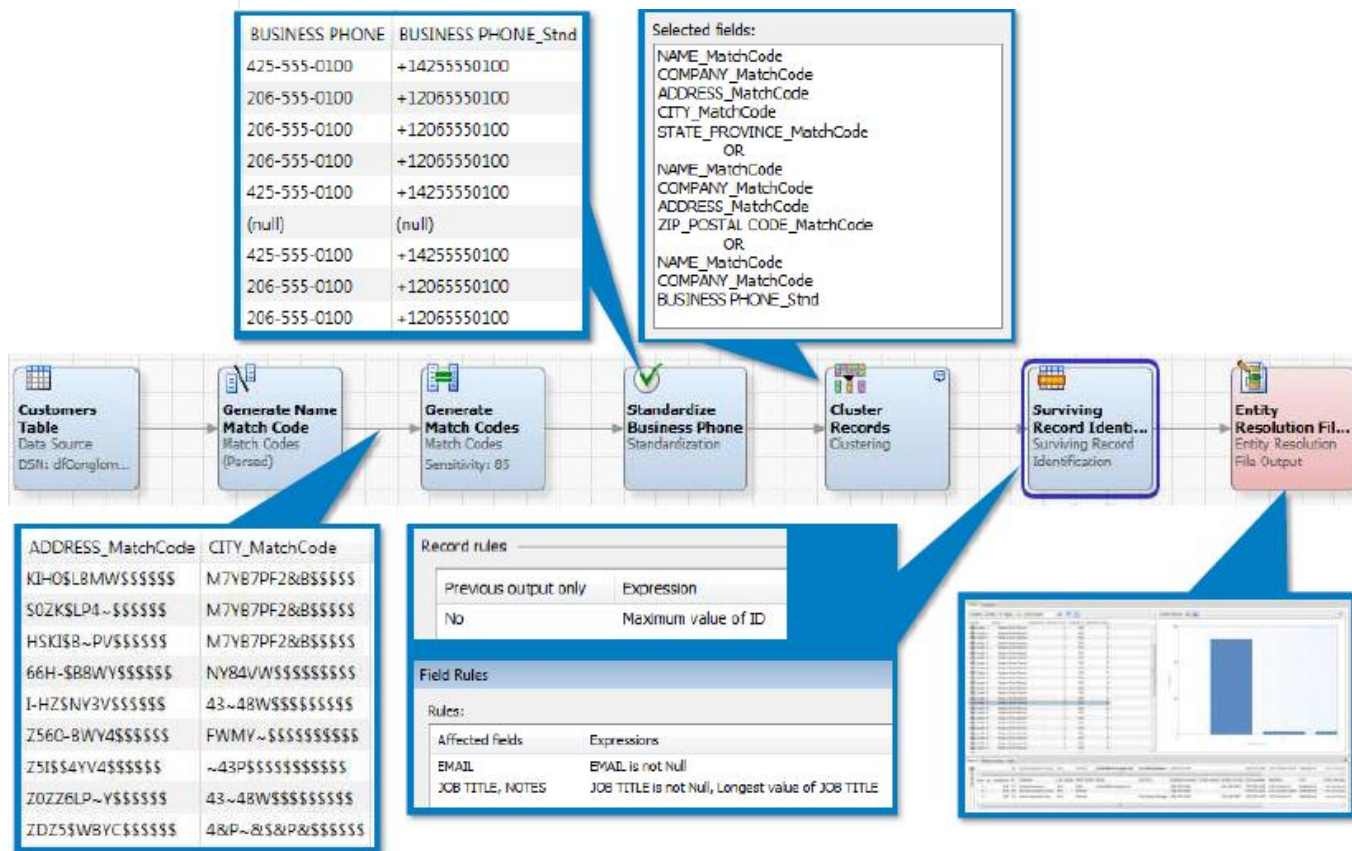
Интеграционная платформа





- **Исполняет в промышленной среде** процессы, правила и сервисы, созданные в Data Management Studio
- Работает как под Windows, так и под Unix/Linux
- Все процессы исполняются как в **пакетном режиме**, так и в **режиме реального времени**.
- **Интеграция** с широким спектром приложений и сред через **SOA** (web сервисы) или через **API** (C, C#, Java, etc.)
- Необходимый уровень **безопасности** обеспечивается через SAS Metadata Server





Построение процесса мониторинга качества данных организации



SAS Data Governance

Управление единым глоссарием бизнес- терминов

Описание ограничений существования бизнес-терминов

Привязка бизнес-терминов к физической модели данных организации

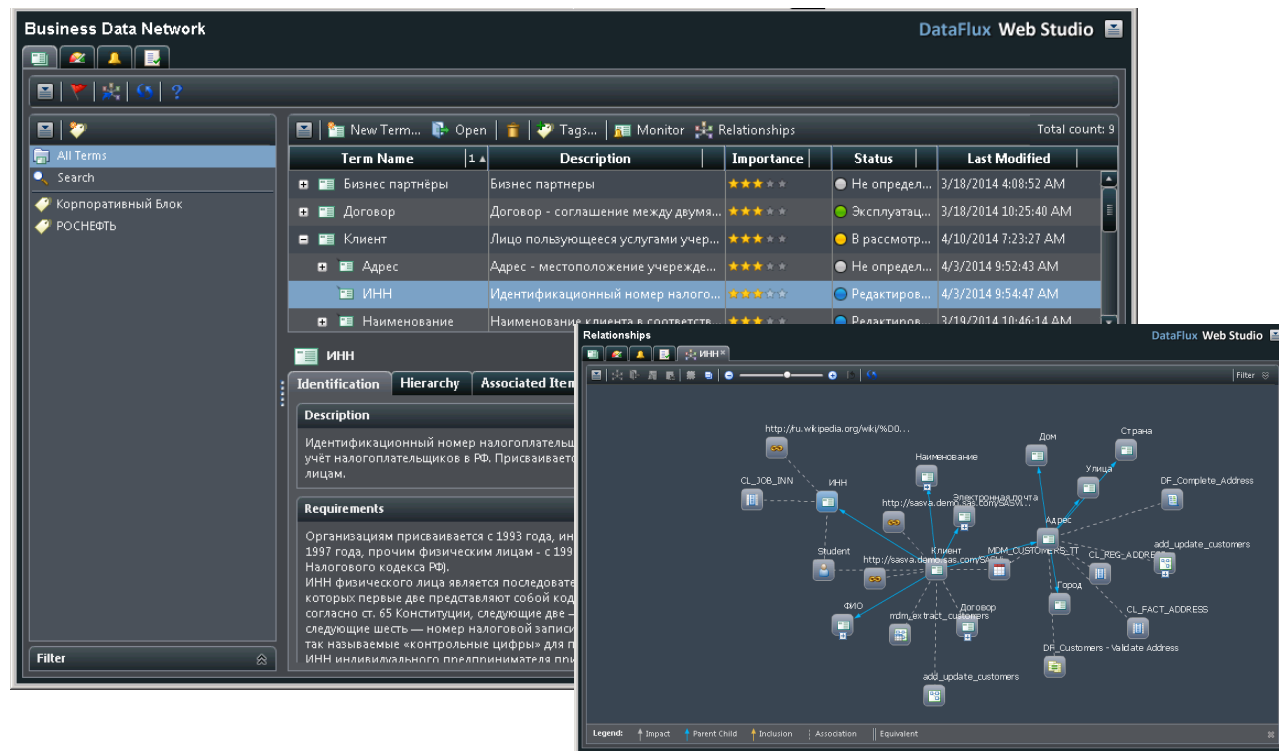
Создание бизнес-правил контроля на основе ограничений

Назначение ответственных за конкретные бизнес-термины

Управление версиями и возврат к предыдущим

Визуализация отношений объектов

БИЗНЕС ГЛОССАРИЙ – ФОРМАЛИЗАЦИЯ ЕДИНОГО ПРЕДСТАВЛЕНИЯ КАЧЕСТВА ДАННЫХ В РАМКАХ ВСЕГО ПРЕДПРИЯТИЯ



SAS Data Governance

Data monitoring – позволяет контролировать выполнение конкретных ограничений заложенных в бизнес-правилах над данными описанными в глоссарии, отслеживать события и создавать оповещения владельцам бизнес-терминов.

МОНИТОРИНГ ОТДЕЛЬНЫХ МЕТРИК КАЧЕСТВА ДАННЫХ

DataFlux Web Studio - Windows Internet Explorer

http://sasbi.demo.sas.com:21079/webstudio/#module=Mor

DataFlux Web Studio

Monitor

Filter: Repository: 'all' | Date: 'Last 90 Days'

Apply Reset

Repository: (all)

Date: Last 90 Days

General

Rule: (all)

Rule label: (all)

Task: (all)

Task label: (all)

Source: (all)

Description: (all)

Dashboard: (all)

Triggered rules:

Filter: Repository: 'all' | Date: 'Last 90 Days'

Total count: 22

Rule Label	Task Label	Date	% Triggers
Проверка адреса по КЛАДР	Проверка корректности адреса	4/4/2014 5:31:25 AM	14
Проверка адреса по КЛАДР	Проверка корректности адреса	4/4/2014 5:31:25 AM	0
Проверка адреса по КЛАДР	Проверка корректности адреса	4/4/2014 5:31:25 AM	29
Проверка адреса по КЛАДР	Проверка корректности адреса	4/4/2014 5:31:25 AM	0
Проверка адреса по КЛАДР	Проверка корректности адреса	4/4/2014 5:31:25 AM	0
Проверка адреса по КЛАДР	Проверка корректности адреса	4/4/2014 5:31:25 AM	0

Summary Trigger Values Rule Code Status History Graph

4/4/2014 10:52:31 AM

Rule: DF_Completeness_Customers

Task: 4/4/2014

Date: 10:52:31 AM

Time:

Trigger percent: 0% →

Trigger count: 0

Rows processed: 0

Importance: *****

Trigger status:

Trigger reason:

Rule label: Клиенты - основные поля

Task label:

Run status: Monitor Exec - Required field 'CUSTOMER_PHONE' was not provided.

Run owner: sas

Repository: MDM

Data source: Клиенты

Description: Заполненность обязательных полей

Dashboard: Default

Dimension: Полнота

SAS Data Governance

Dashboards - панели, отображающие основные показатели качества данных в целом.

К каждому отдельному бизнес-правилу привязывается ряд макро-метрик и весов к ним. Данная информационная панель позволяет показать состояние качества данных в организации по макрометрикам собрав информацию со всех бизнес-правил.

КОНТРОЛЬ КАЧЕСТВА ДАННЫХ ОРГАНИЗАЦИИ ВЦЕЛОМ



**Владелец:
Бизнес
пользователь**

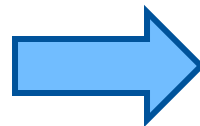
📋 Бизнес-термин и ограничения

Описание

Ценовые и количественные правила для сделок по торговле деривативами

Условия

Сумма сделки не должна превышать \$ 1 млн.
Сумма сделок в день не должна превышать 10% всех фьючерсных сделок. Значения должны быть рассчитаны в долларах или преобразованы из других валют.



**Единый Глоссарий
Бизнес-данных**



📋 Бизнес-правила

Формула расчёта правила: Trade Out of Range

```
if ('Trade' < 1 or 'Trade' > 1000000)
then
    return true
else
    return false
```



🔔 Мониторинг и улучшение KPI Бизнесом и ИТ

Имя правила	Дата	% Triggers	# Triggers	Кол-во строк
Trade Out of Range	6/10/2010 1:14:13 AM	3	563	5002
Incorrect ZIP	6/9/2010 4:5:27 AM	6	786	4789
Incomplete Phone	6/8/2010 11:23:7 AM	2	42	4378
Trade Out of Range	5/29/2010 1:1:1 AM	3	453	3689

**Владелец:
Администратор
данных**

ERGO
Insurance Group

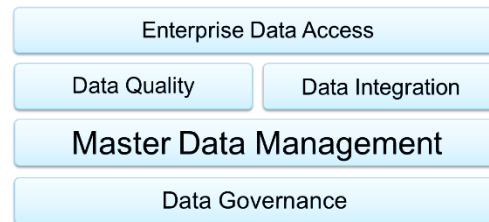
TURKCELL

GRUPO POPULAR

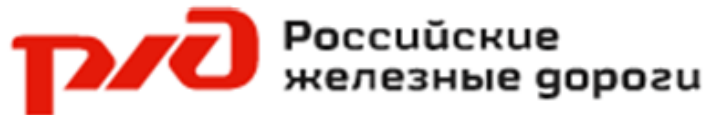
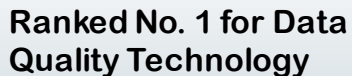
«Оркестровка» людей, процессов и технологий для использования данных как части активов



Роли и обязанности
Владельцы данных, администраторы данных,
ответственные за данные



SAS - ПРИЗНАННЫЙ МИРОВОЙ ЛИДЕР В ТЕХНОЛОГИЯХ РАБОТЫ С ДАННЫМИ

**ГАЗПРОМБАНК**

Тинькофф. Кредитные Системы





- Сквозное решение обеспечения качества данных от интеллектуального поиска ошибок до выстраивания процесса выверки любой сложности
- Интеграция любых технологий систем источников, включая неструктурированные БД
- Онлайн интеграция любой сложности
- Самая эффективная очистка “грязных” данных на рынке
- Локализация для русского языка
- Открытая, расширяемая база знаний, настраиваемые под Ваши потребности
- Выверенная методология повышения качества данных
- Низкая стоимость внедрения и владения
- Встроенные в платформу технологии контроля качества и полноты информации бизнес-заказчиками (Data Governance)

СПАСИБО ЗА ВНИМАНИЕ!



**THE
POWER
TO KNOW®**