Практика создания «умного» озера данных

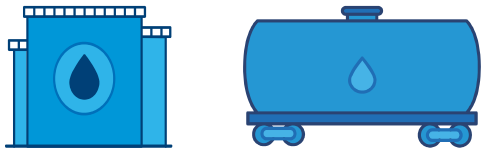
15.10.2019

Иван Черницын
Руководитель Центра аналитических решений
Дирекции региональных продаж ПАО «Газпром нефть»

BIG DATA '19 SPb

Параметры бизнеса Дирекции региональных продаж (ДРП)

- Биржевые продажи нефтепродуктов
- Реализация нефтепродуктов крупным и мелким оптом
- Хранение нефтепродуктов
- Управление >50 собств. нефтебазами



- Розничная реализация нефтепродуктов
- Продажи корпоративным клиентам
- Управление АЗС \ АЗК >1800 объектов

- >800 розничных магазинов и кафе при АЗС
- Оказание услуг моек, СТО

- Управление автотранспортом: бензовозы/газовозы
- Услуги метрологии

29 регионов присутствия в России + 4 страны СНГ

> 20 тысяч сотрудников

Клиентов-участников бонусной программы > 11,4 млн. человек

2009 – 2011 г. предпосылки: программа стандартизации и комплексной автоматизации ДРП

Стандартизация бизнес-процессов, оргструктур, внедрение унифицированных шаблонных систем породили первые практики управления данными и потребность в аналитических решениях

- Система управления нормативно-справочной информацией
- Система управления АЗС
- Система коммерческого учета
- Система бухгалтерского и управленческого учета
- Система планирования и бюджетирования
- Система управления инвестициями
- Система управления персоналом
- Система управления МТО
- Система управления ТО и ремонтами
- Система управления автотранспортом

Централизация методологии

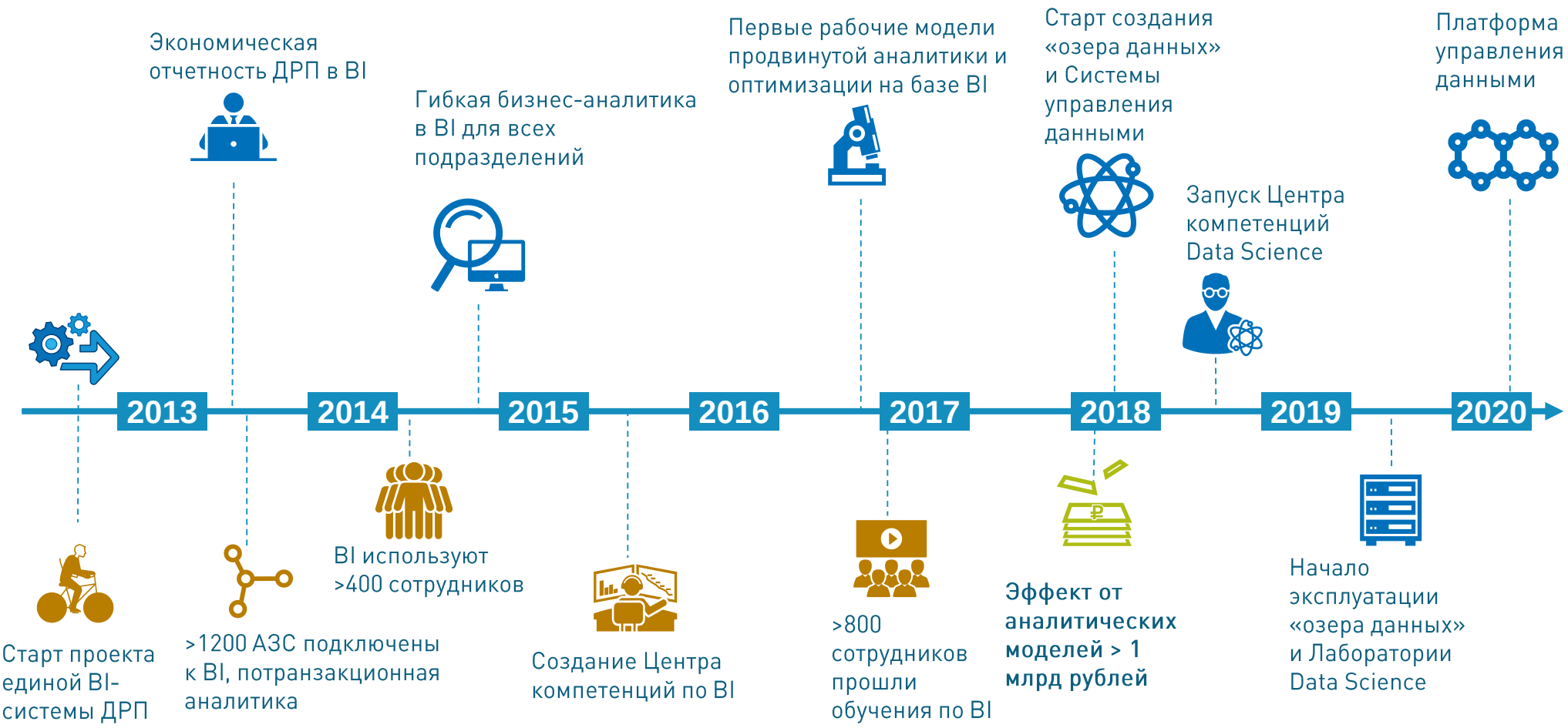
Институт владельцев систем

Единые ключевые справочники

Высокая потребность в интеграции данных

Высокий темп организационных изменений – существенная реорганизация каждые 2 года

Развитие аналитических решений в ДРП ГПН



2013 – 2016 гг.: единая BI-система ДРП на платформе Qlik интегрировала все ключевые данные

> 1 100 пользователей

>90 систем-источников

>50 внешних и несистемных источников данных

> 250 аналитических приложений

Функциональные подсистемы BI охватили ключевые направления деятельности:

- Учет и консолидация
- Экономический анализ
- Расходы
- Финансы
- Инвестиции
- Маржинальный доход
- Материальный баланс
- АРМы руководителей

- Аналитика розничных продаж
- Клиентская аналитика
- Сервисы для АЗС
- Сравнительный анализ АЗС
- Программа лояльности АЗС
- Управление ассортиментом
- Анализ работы оборудования
- Мониторинг цен

- ИТ-сервисы
- Управление проектами
- Аналитика персонала
- Эффективность бизнес-процессов
- Оптовые продажи
- Логистика
- Качество данных

2016 год: исследование «Построение службы Data Helpdesk» с Фондом «Сколково»

2017 год: источники вдохновения – материалы Gartner и DAMA DMBoK

Making Your Landscape Less Uncertain — Focus Areas

Component — not core

Operational and analytics skills

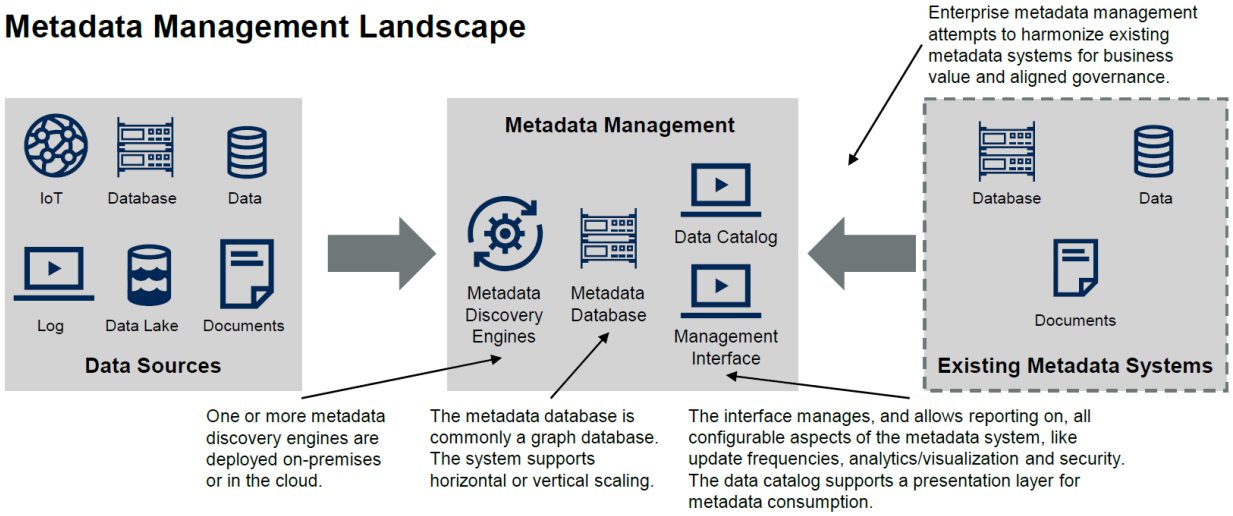
Security and metadata

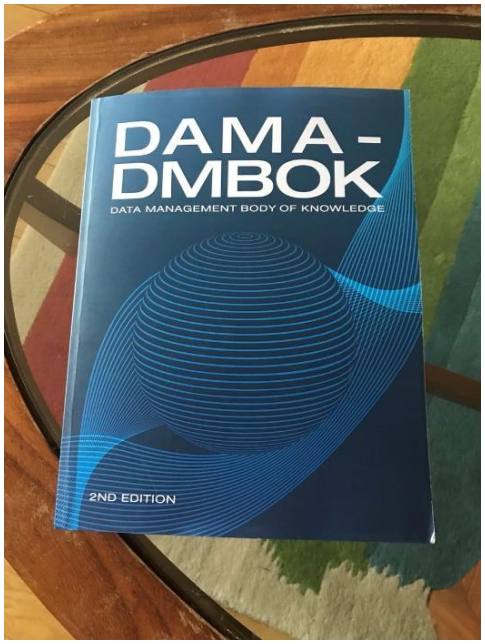
Data Catalogs Are the New Black in Data Management and Analytics

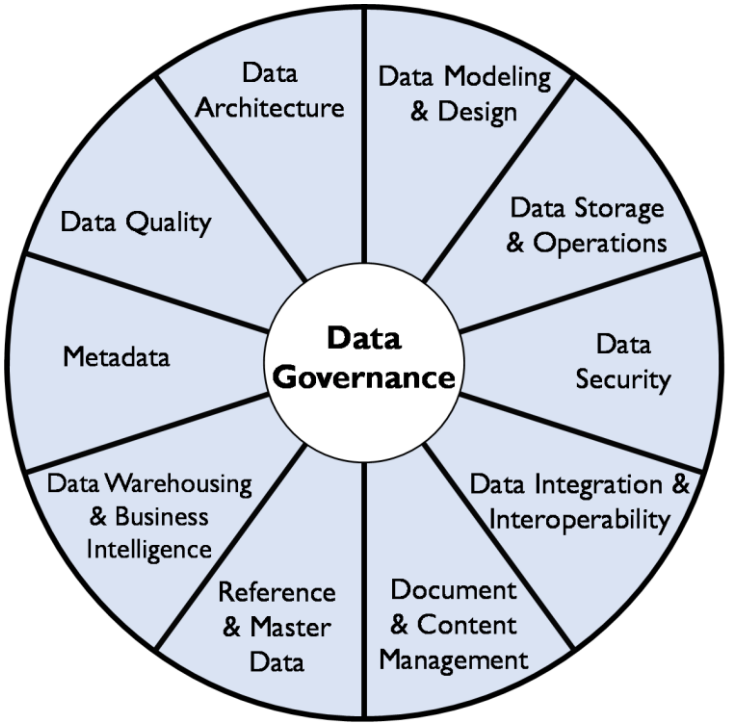
Published: 13 December 2017 **ID:** G00338777

Analyst(s): Ehtisham Zaidi, Guido De Simoni, Roxane Edjlali, Alan D. Duncan

Metadata Management Landscape

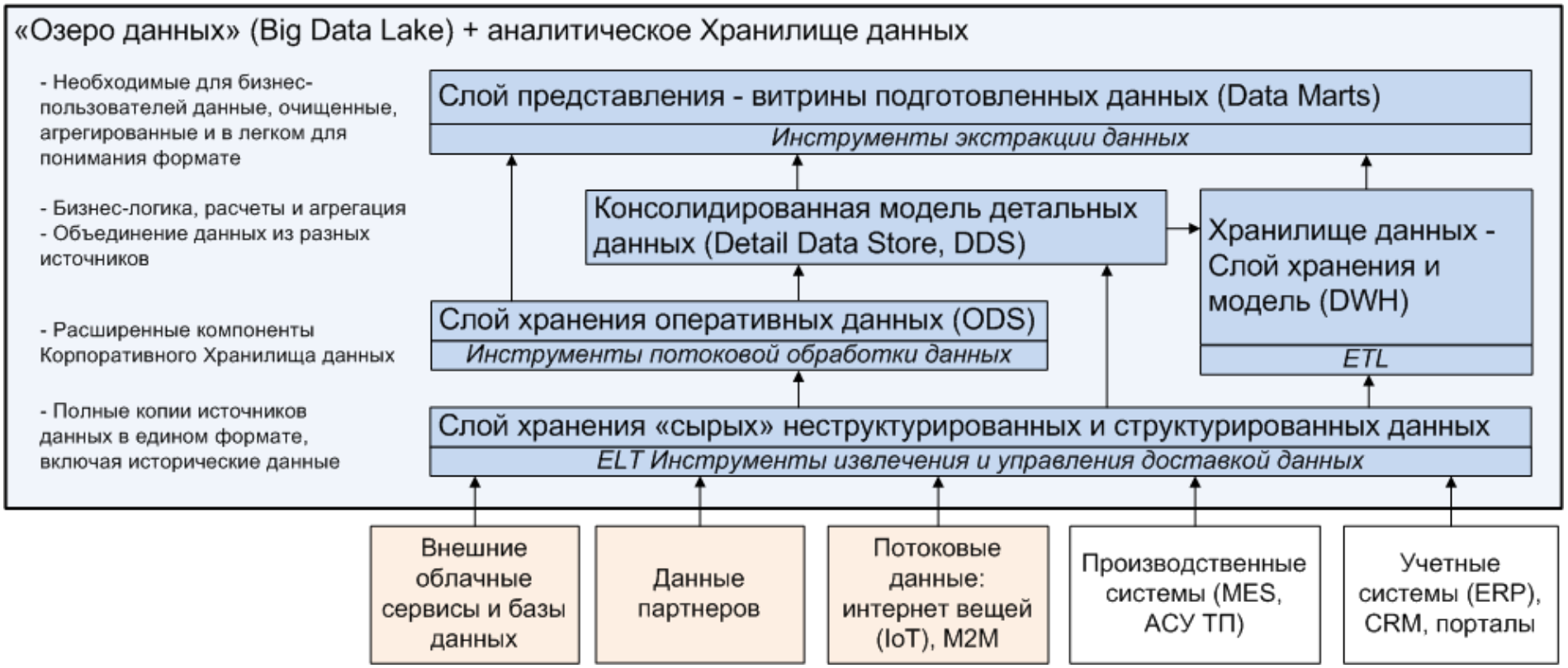






2017 год: идея «озера данных» как комплексной платформы обработки и хранения данных

Озеро данных (Big Data Lake) - программно-аппаратный комплекс как совокупность различных массивов данных, объединенных в единую логическую среду, для обеспечения доступа, управления и преобразования данных. Цель создания озер данных – дать высококвалифицированным аналитикам возможность изучать неочищенные, не агрегированные данные и применять к ним различные аналитические техники



2017 год: 4 уровня потребностей для развития аналитических решений в ДРП

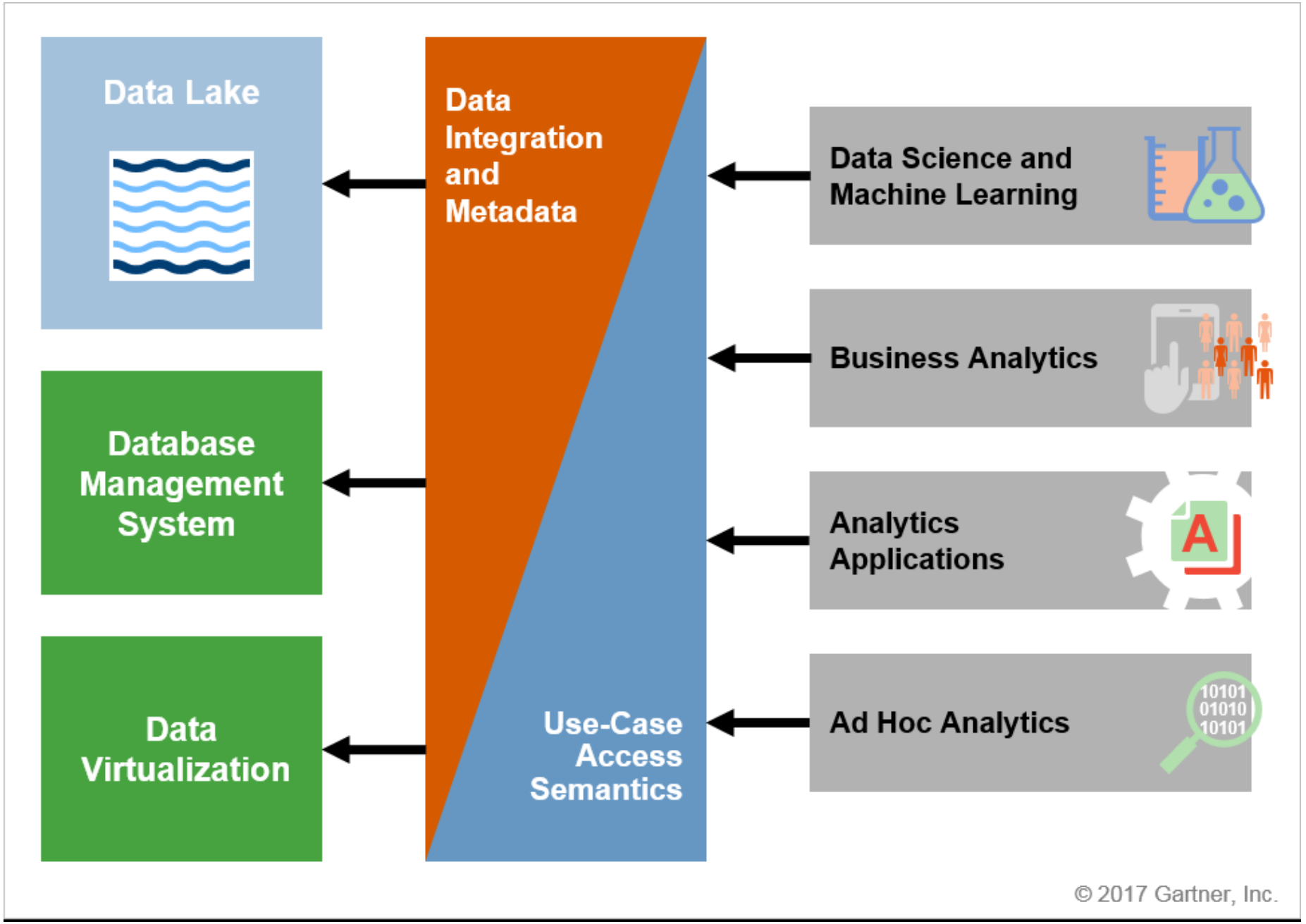
1. **«Озеро данных»** - промышленные инструменты интеграции, ETL и хранения данных

2. **Единый каталог правил по качеству данных** с управлением и мониторингом

3. **Система управления метаданными** для аналитиков: каталог доступных источников, таблиц, витрин, моделей данных и отчетов + цепочки происхождения

4. **Пользовательский портал по данным**: поиск определений данных, владельцев, ролей, отчетов и т.д.

Gartner: концепция «умного» озера данных как продвинутый архитектурный шаблон



Источник: Исследование Gartner «Derive Value From Data Lakes Using Analytics Design Patterns» (Сентябрь 2017)

2017 год: идея и защита программы проектов «умное озеро данных» ДРП

Примеры исходных требований к системе управления данными ДРП

- Ведение бизнес-метаданных и инструменты управления изменениями
- Бизнес-глоссарий с описанием бизнес-терминов, их взаимосвязей и атрибутов, связанных правил по качеству данных, процессами управления изменениями
- Бизнес-глоссарий, **интегрированный с подсистемой управления качеством данных**

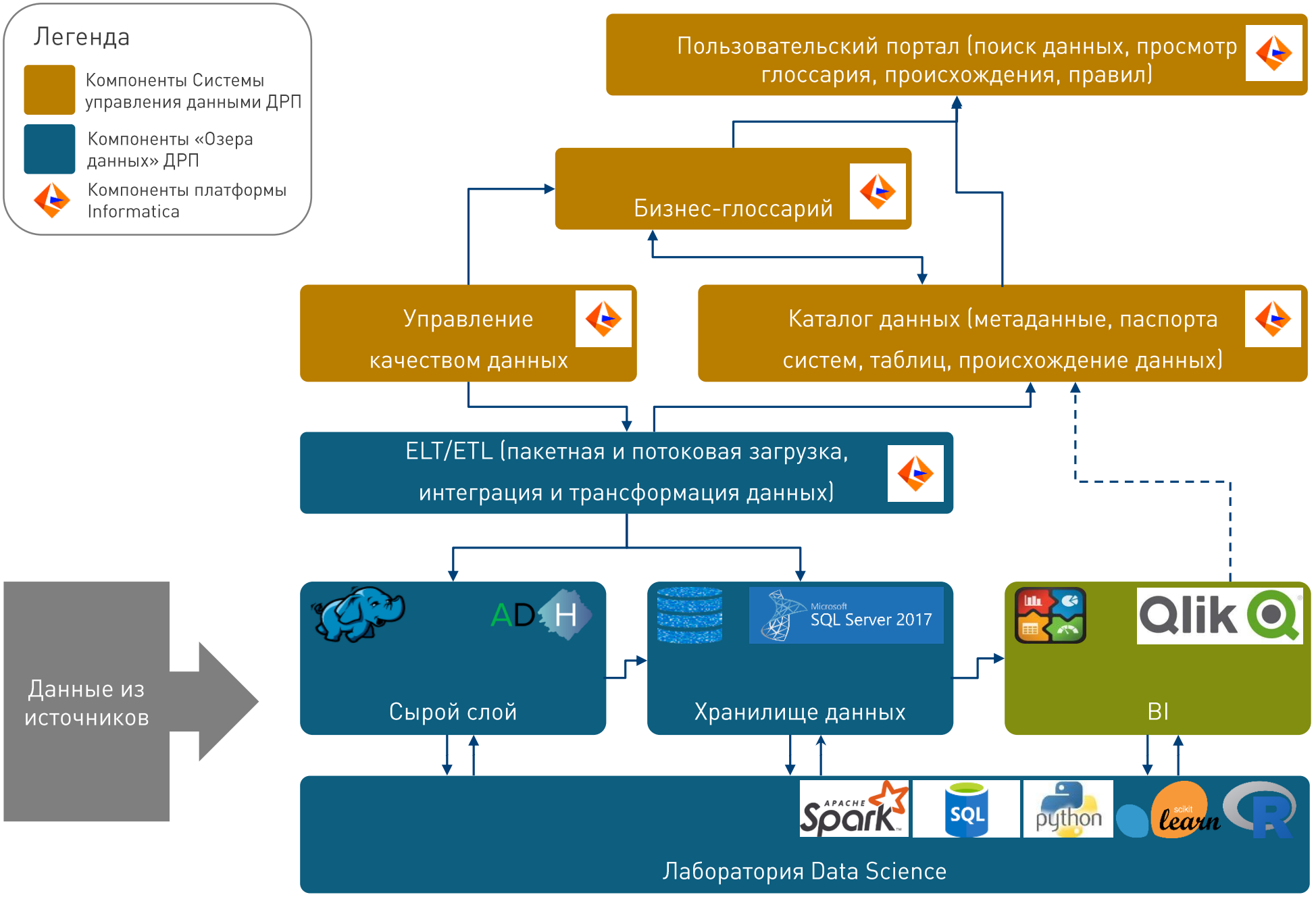
- В контуре «озера данных» **сбор потоков происхождения данных** от систем источников до конечных пользовательских представлений и BI-приложений **производится в автоматизированном режиме**
- В подсистеме управления метаданными реализован мониторинг изменений технических метаданных

- Единый каталог правил по качеству данных
- **Подсистема управления качеством данных, интегрированная** с «озером данных»
- Измерение, отображение и аудит всех проверок по качеству данных в контуре «озера данных» **в разрезе потоков обработки данных**

- Паспортизация источников и полей данных
- Функциональность профилирования данных
- Функциональность разметки данных (коммерческая тайна и персональные данные)

- Подсистема для разрешения запросов бизнес-пользователей на поиск доступных данных и ролей для доступа к данным, **интегрированная с бизнес-глоссарием и системой управления метаданными**

Компоненты решения «умное озеро данных» (упрощенная схема)



Создание «умного озера данных» ДРП: основные события

Пример бизнес-кейса в Озере данных ДРП: клиентская аналитика

СЕГМЕНТЫ Клиентского контура

Базовая сегментация:
АВТО, КОМ ТРАНС, B2G, НАЗС, ПРОМ СХ

Подсегменты: Авто, Fleet, CRT, НКП, ПРОМ, СХ

Активность: активные, неактивные, архив

Новизна клиента

Лояльность G-Drive

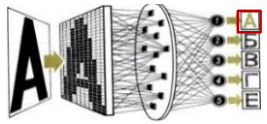
Вид деятельности по ОКВЭД

Объем потребления клиента:
массовый, средний, крупный

Активность СТиУ

Тип / класс / износ ТС

Распознавание НОМЕРА АВТО нейронная сеть



Возраст

Пол

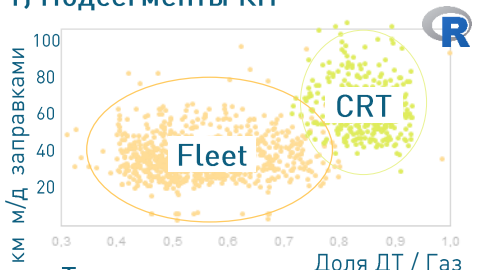
География

Используются методы для сегментаций:

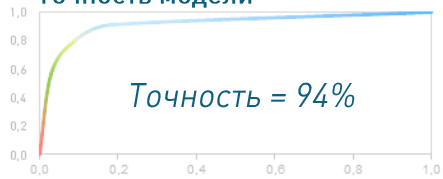
- Classification Tree
- Random Forest
- XGB
- Нейронные сети
- GBM

Сегментация клиентов на основе методов машинного обучения и продвинутой аналитики

1) Подсегменты КП



Точность модели



2) Подсегменты ПЛ

Целевой сценарий совместной работы компонент «умного озера данных» работает!

С мая 2019 года все новые кейсы ДРП реализуются по целевому сценарию:

разработка ETL на продуктах Informatica с автоматической передачей трансформаций в Систему управления данными

Далее на слайдах примеры реализованного функционала

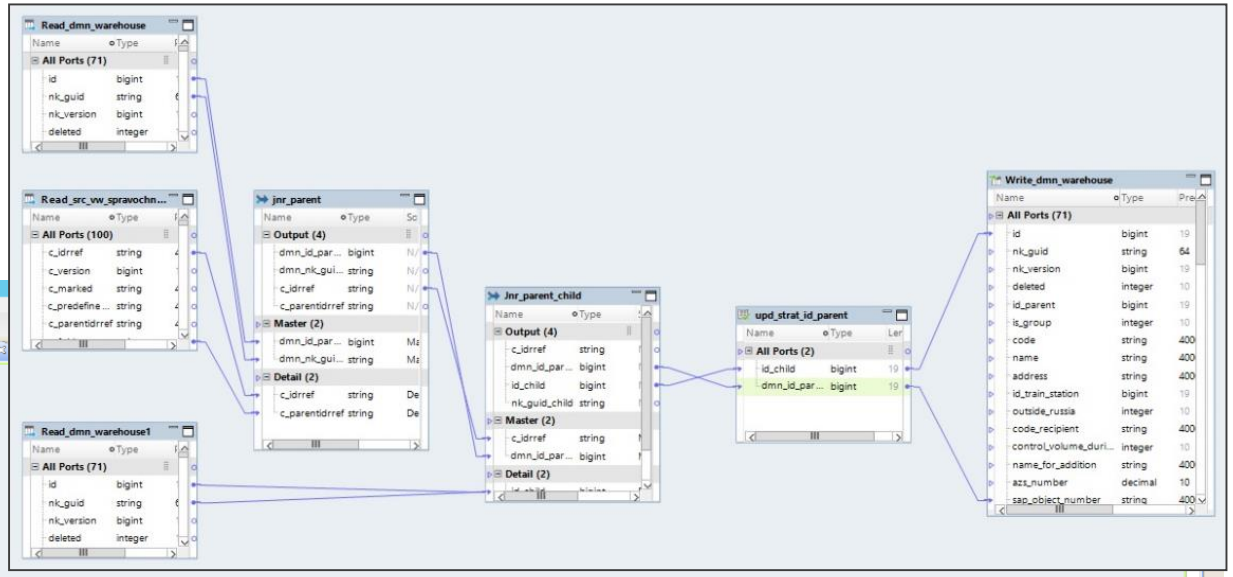
Informatica

Особенности:

- Требуется изменение культуры разработки
- Применяем итерационный подход к поиску лучших практик
- Вырабатываем шаблоны для ускорения разработки

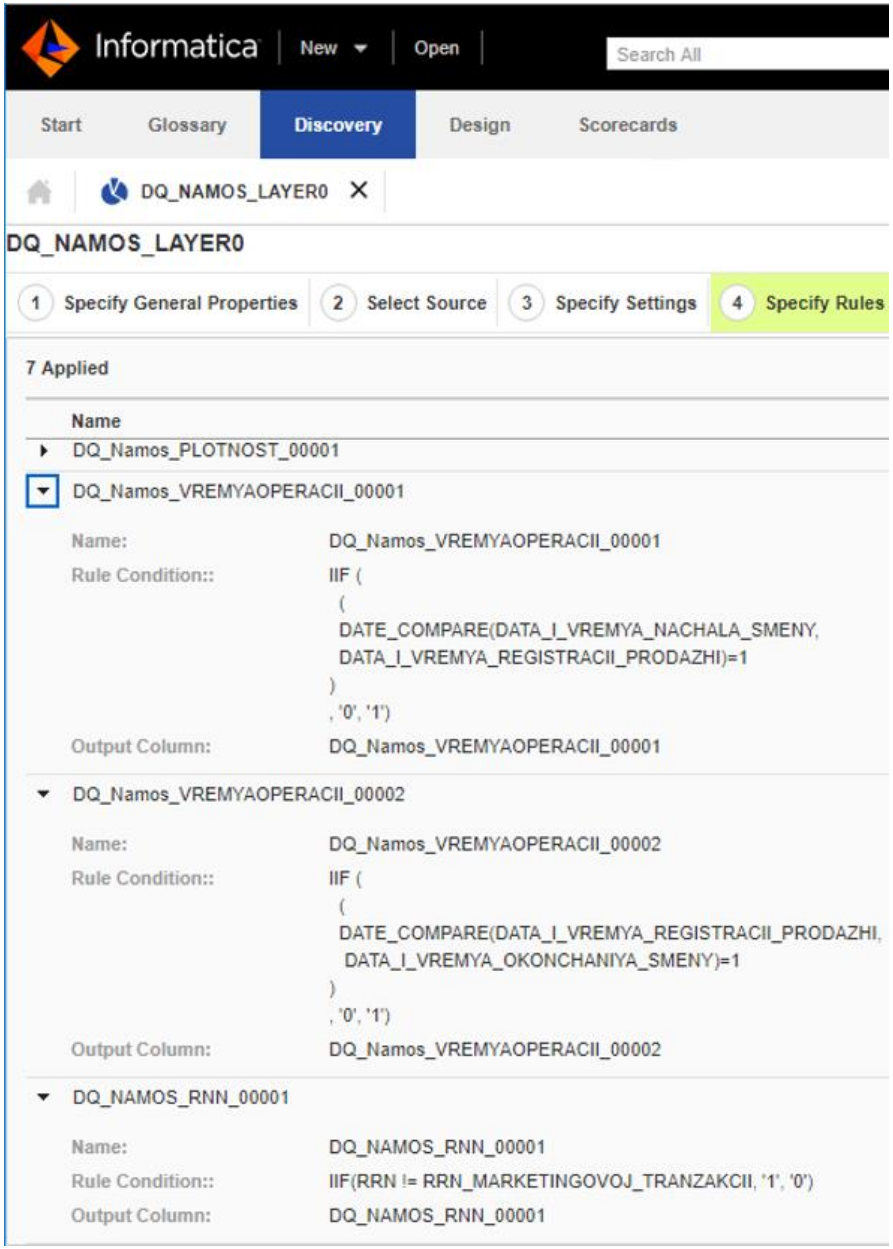
Разработка ELT/ETL-процедур (Developer)

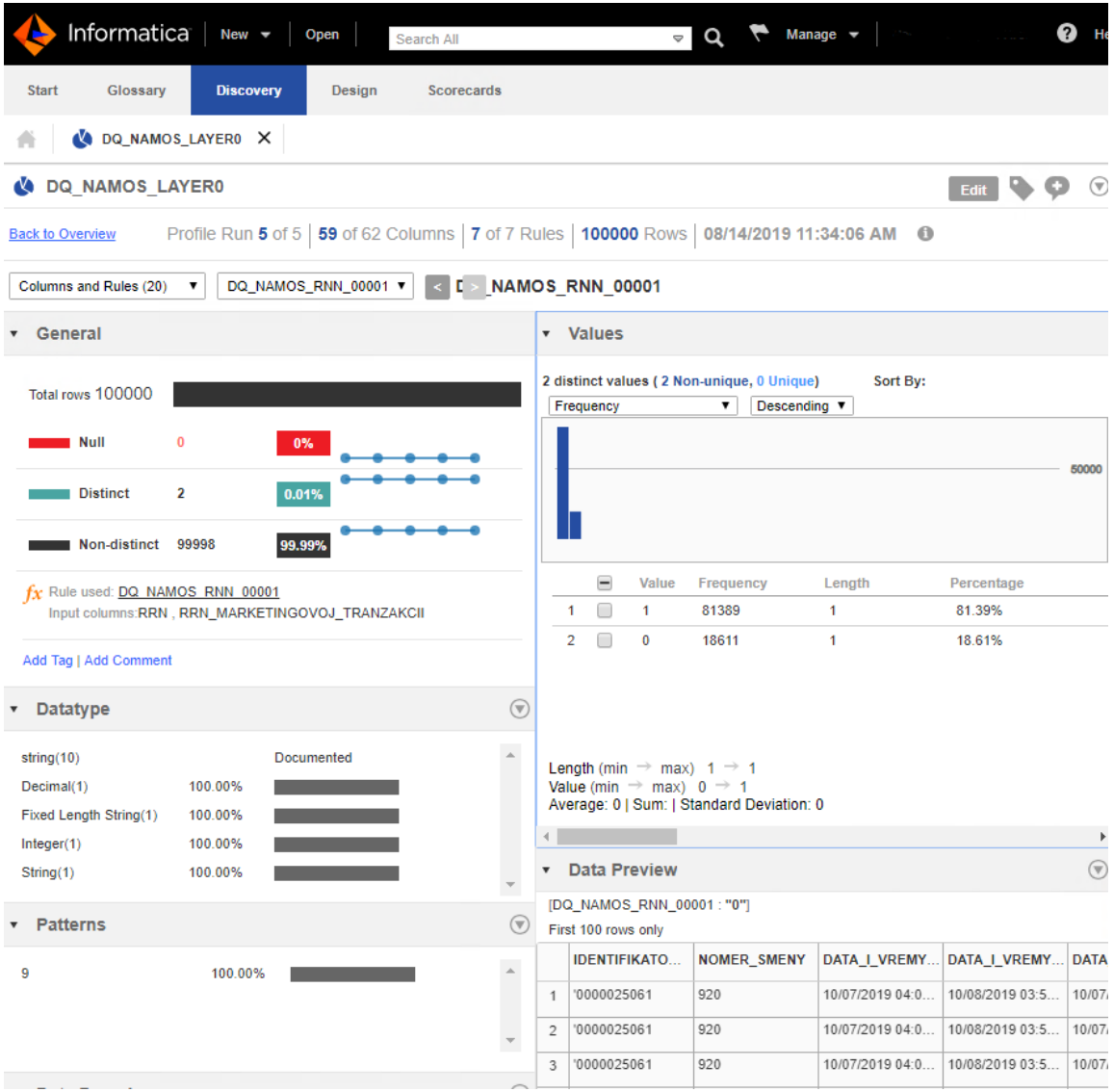
! Требуется значительное количество соглашений по практике разработки для автоматической работы инструментов управления данными



Создание и просмотр результатов правил по качеству данных (Analyst)

Подсистема качества данных интегрирована с инструментами ETL и позволяет создавать и встраивать правила по качеству данных в ETL-процессы





Единый каталог правил по качеству данных и их статусов (Analyst)

Доступно управление типами (аспектами) и уровнем критичности правил

DS-159: fct_stock_index
Data Set in DW_SPB_SPB99_DBT01_DWH_DEV03

Edit

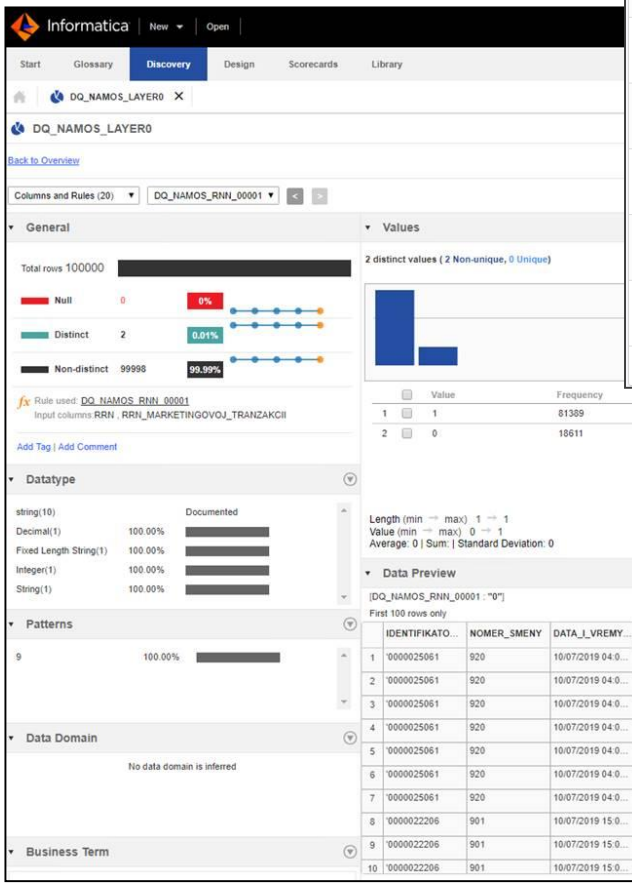
SUMMARY | ATTRIBUTES | RELATIONSHIPS | STAKEHOLDERS | VALUES | DATA QUALITY | IMPACT | HISTORY | CHANGE | FOLLOW

Dashboard | Rules

DETAILS

Cont...	Rule Type	Ref.	Description	A...	Attributes	System	Data Set	Target	Last ...	Last Result Date
(8 records)										
	Допустимость	DQ_INTERNET_StockIndexIndex	Числовое значение не отрицательное.	Active	value_wo_trar	DW_SPB_SPB99_DB	fct_stock_index	90%	100%	24-Jun-2019
	Допустимость	DQ_INTERNET_StockIndexNum	Числовое значение не отрицательное	Active	num_trades	DW_SPB_SPB99_DB	fct_stock_index	90%	100%	12-Aug-2019
	Допустимость	DQ_INTERNET_StockIndexDate_	Формат не соответствует маске DD.MM.YYYY	Active	dt	DW_SPB_SPB99_DB	fct_stock_index	90%	100%	13-Aug-2019
	Допустимость	DQ_INTERNET_StockIndexIndex	Числовое значение не отрицательное	Active	index_value	DW_SPB_SPB99_DB	fct_stock_index	90%	100%	26-Jul-2019
	Допустимость	DQ_INTERNET_StockIndexValue	Числовое значение не отрицательное	Active	value_per_lite	DW_SPB_SPB99_DB	fct_stock_index	90%	100%	30-Jul-2019
	Допустимость	DQ_INTERNET_StockIndexValue	Числовое значение не отрицательное	Active	value_per_lite	DW_SPB_SPB99_DB	fct_stock_index	90%	100%	26-Jul-2019
	Допустимость	DQ_INTERNET_StockIndexValue	Числовое значение не отрицательное	Active	value	DW_SPB_SPB99_DB	fct_stock_index	90%	100%	26-Jul-2019
	Допустимость	DQ_INTERNET_StockIndexVolur	Числовое значение не отрицательное	Active	volume	DW_SPB_SPB99_DB	fct_stock_index	90%	100%	26-Jul-2019

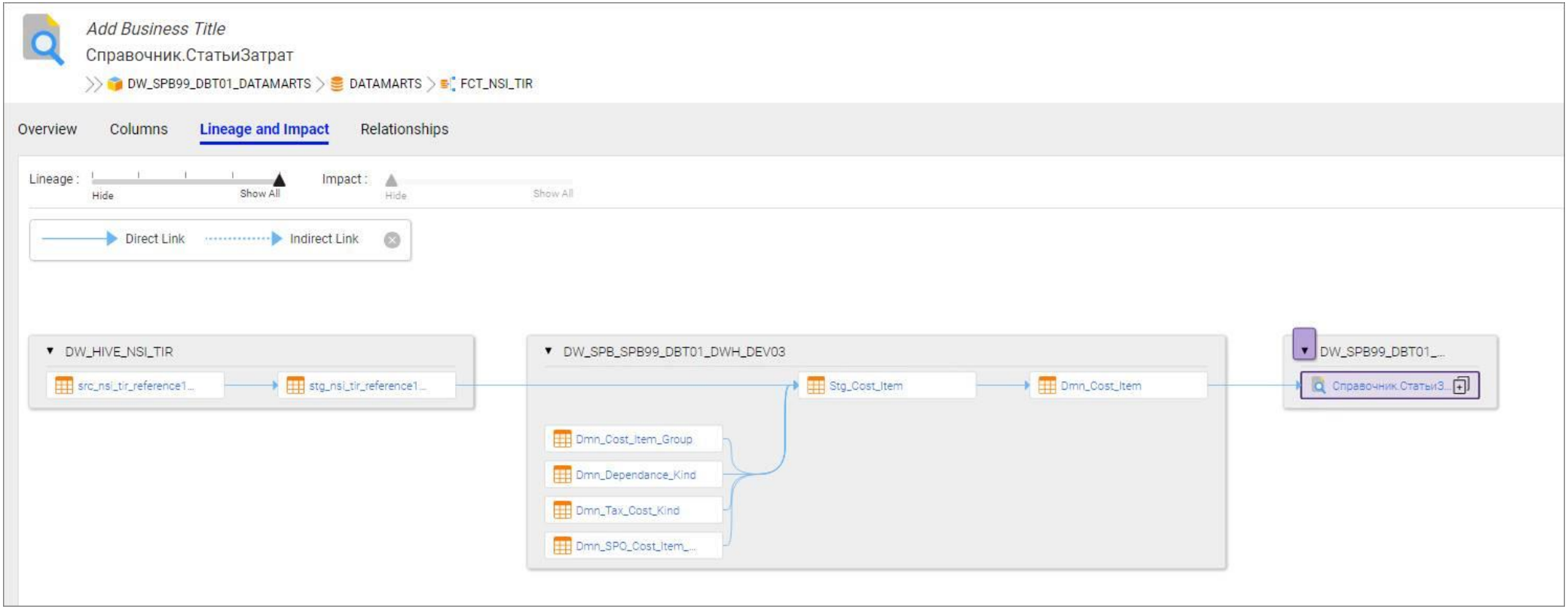
8 records



	IDENTIFIKATO...	NOMER_SMENY	DATA_I_VREMY...	DATA_I_VREMY...	DATA_I_VREMY...	KOD_PRODUKTA	SUMMA	VID_VYBYTIYA	PLOTNOST	NOMER_BONU...	RRN	SUMMA_BEZ_...	RRN_MARKETI...	DQ_Namos_PL...	DQ_Namos_VR...	DQ_Namos_VR...	DQ_Namos_NO...	DQ_NAMOS_S...	DQ_Namos_Su...	DQ_NAMOS_R...
1	'0000025061	920	10/07/2019 04:0...	10/08/2019 03:5...	10/07/2019 13:4...	4	1509.9000000	303	0.7536000		0	1509.9500000	0	1	1	1	0	1	1	0
2	'0000025061	920	10/07/2019 04:0...	10/08/2019 03:5...	10/07/2019 05:4...	9900000050060	95.0000000	0	0.0000000		0	95.0000000	0	1	1	1	0	1	1	0
3	'0000025061	920	10/07/2019 04:0...	10/08/2019 03:5...	10/07/2019 05:4...	3	1000.0000000	0	0.7300000		0	1000.1300000	0	1	1	1	0	1	1	0
4	'0000025061	920	10/07/2019 04:0...	10/08/2019 03:5...	10/07/2019 05:4...	9900000050300	106.0000000	0	0.0000000		0	106.0000000	0	1	1	1	0	1	1	0
5	'0000025061	920	10/07/2019 04:0...	10/08/2019 03:5...	10/07/2019 06:0...	13	1400.0000000	0	0.8522300		0	1400.2400000	0	1	1	1	0	1	1	0
6	'0000025061	920	10/07/2019 04:0...	10/08/2019 03:5...	10/07/2019 06:0...	9900000050061	122.0000000	0	0.0000000		0	122.0000000	0	1	1	1	0	1	1	0
7	'0000025061	920	10/07/2019 04:0...	10/08/2019 03:5...	10/07/2019 06:0...	9900000010409	90.0000000	0	0.0000000		0	90.0000000	0	1	1	1	0	1	1	0
8	'0000022206	901	10/07/2019 15:0...	10/08/2019 03:5...	10/07/2019 20:0...	9900000004296	162.0000000	0	0.0000000		0	162.0000000	0	1	1	1	0	1	1	0
9	'0000022206	901	10/07/2019 15:0...	10/08/2019 03:5...	10/07/2019 20:0...	9900000049542	118.0000000	0	0.0000000		0	118.0000000	0	1	1	1	0	1	1	0
10	'0000022206	901	10/07/2019 15:0...	10/08/2019 03:5...	10/07/2019 17:3...	4	600.0000000	0	0.7453700		0	600.2800000	0	1	1	1	0	1	1	0

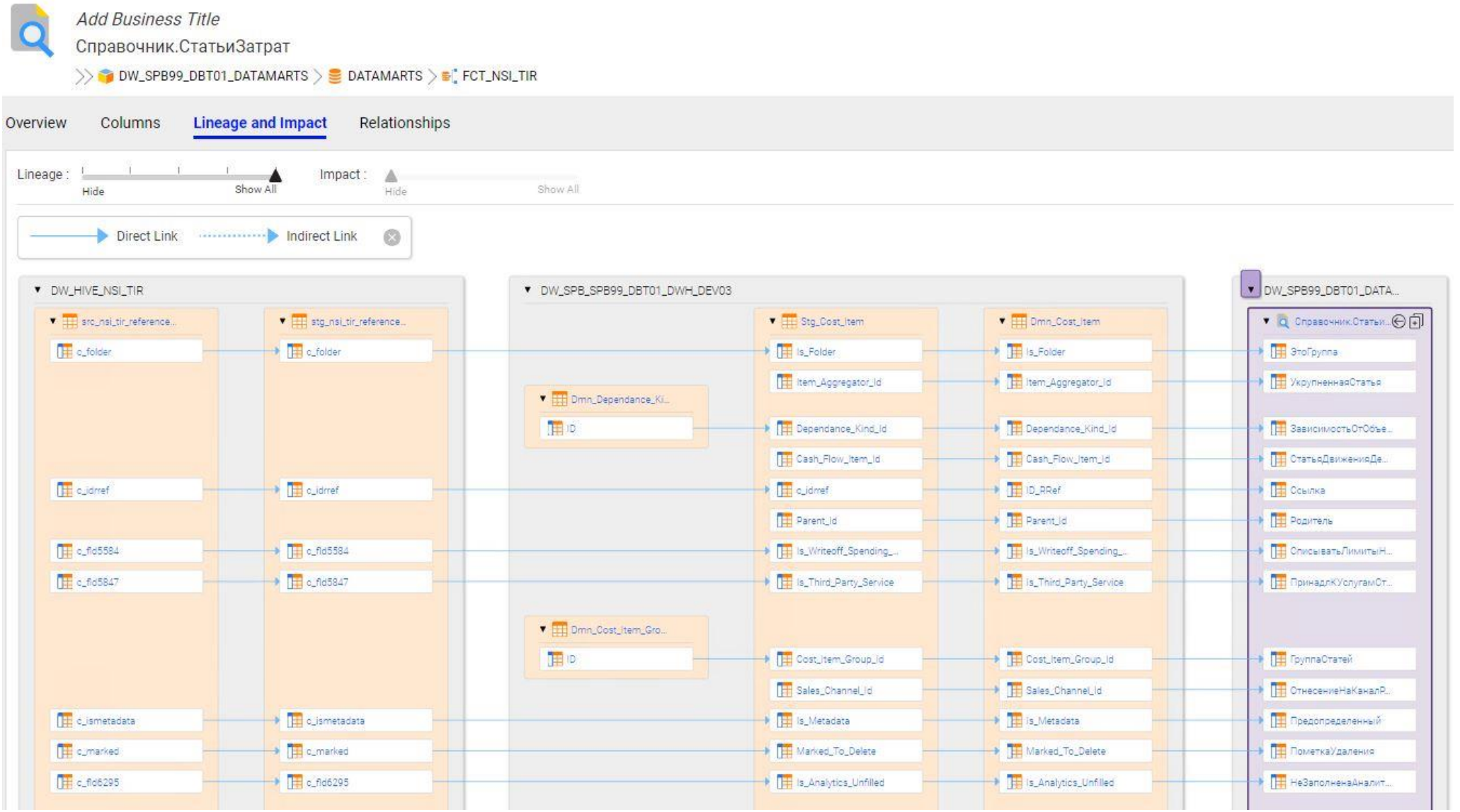
Автоматическое построение происхождения данных до уровня компонент и таблиц (EDC)

Каталог данных подключается к компонентам хранения данных (Hadoop, MS SQL Server) и исполняемым ETL-приложениям, автоматически извлекая/обновляя цепочки происхождения данных и правила/формулы трансформации

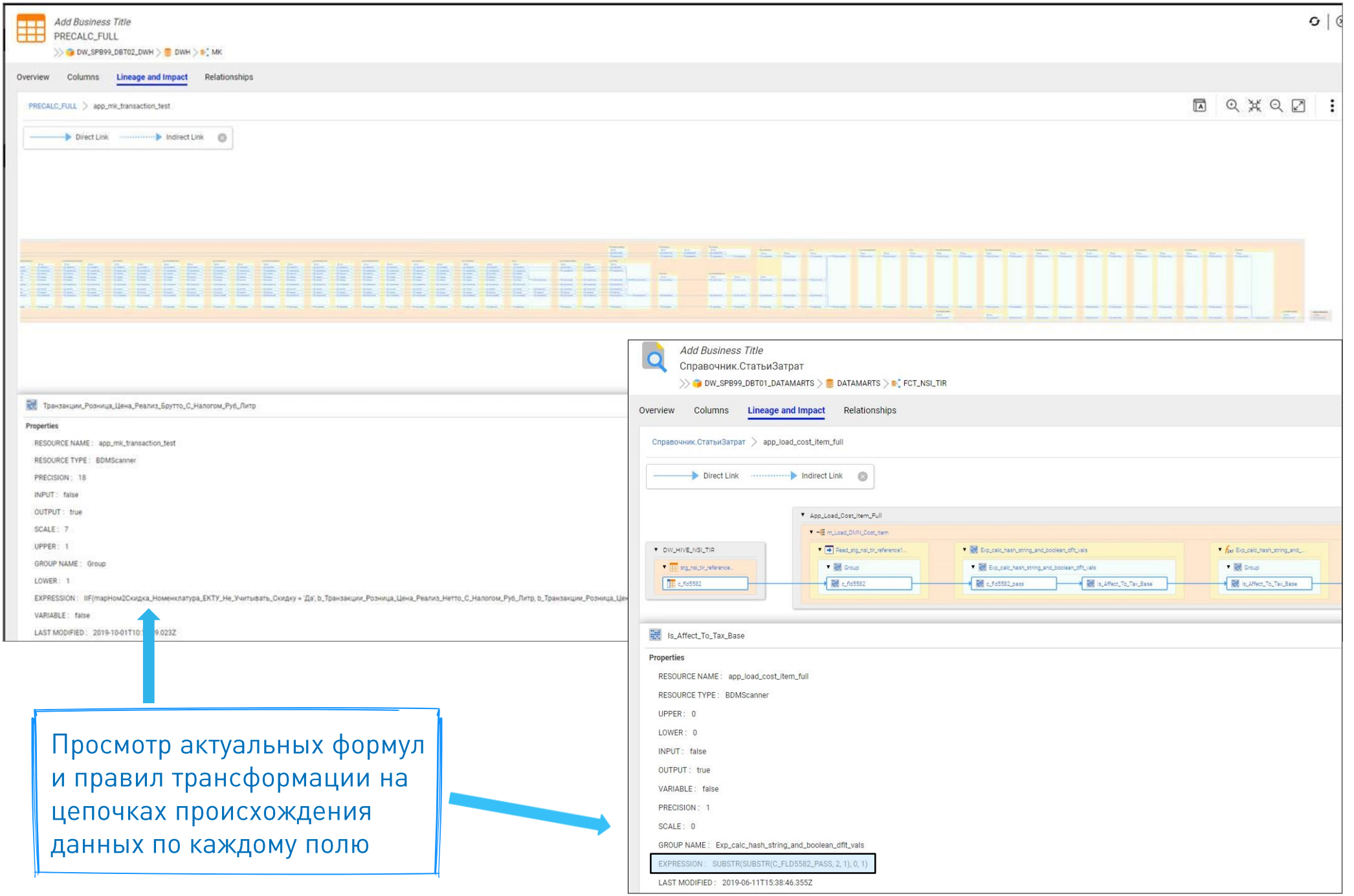


Анализ происхождения данных до уровня полей таблиц данных (EDC)

Цепочки происхождения данных в каталоге разворачиваются до полей таблиц

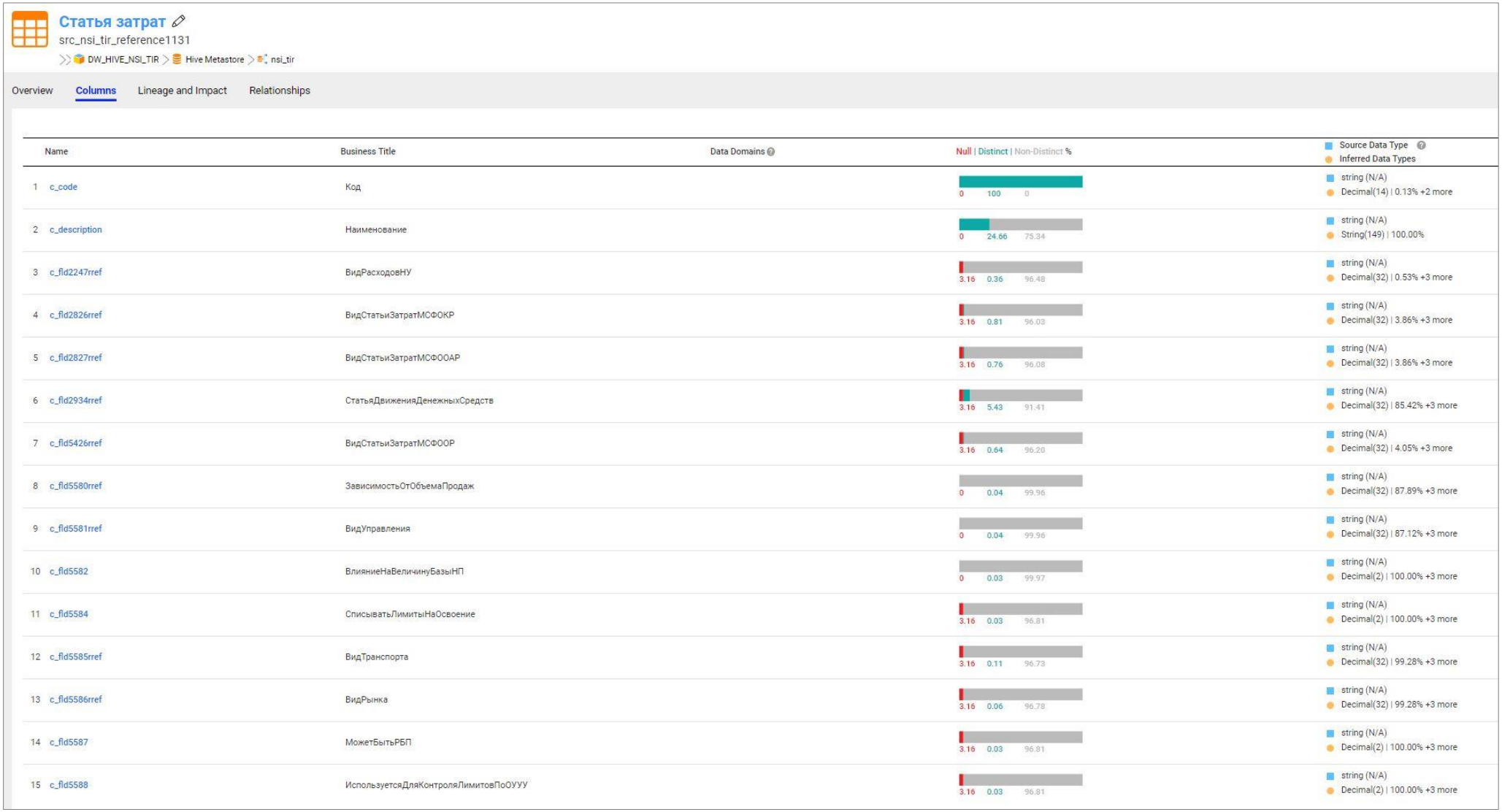


Анализ происхождения данных в каталоге (EDC)
Просмотр формул и трансформаций



Просмотр актуальных формул и правил трансформации на цепочках происхождения данных по каждому полю

Паспортизация систем, таблиц и полей данных в каталоге (EDC)

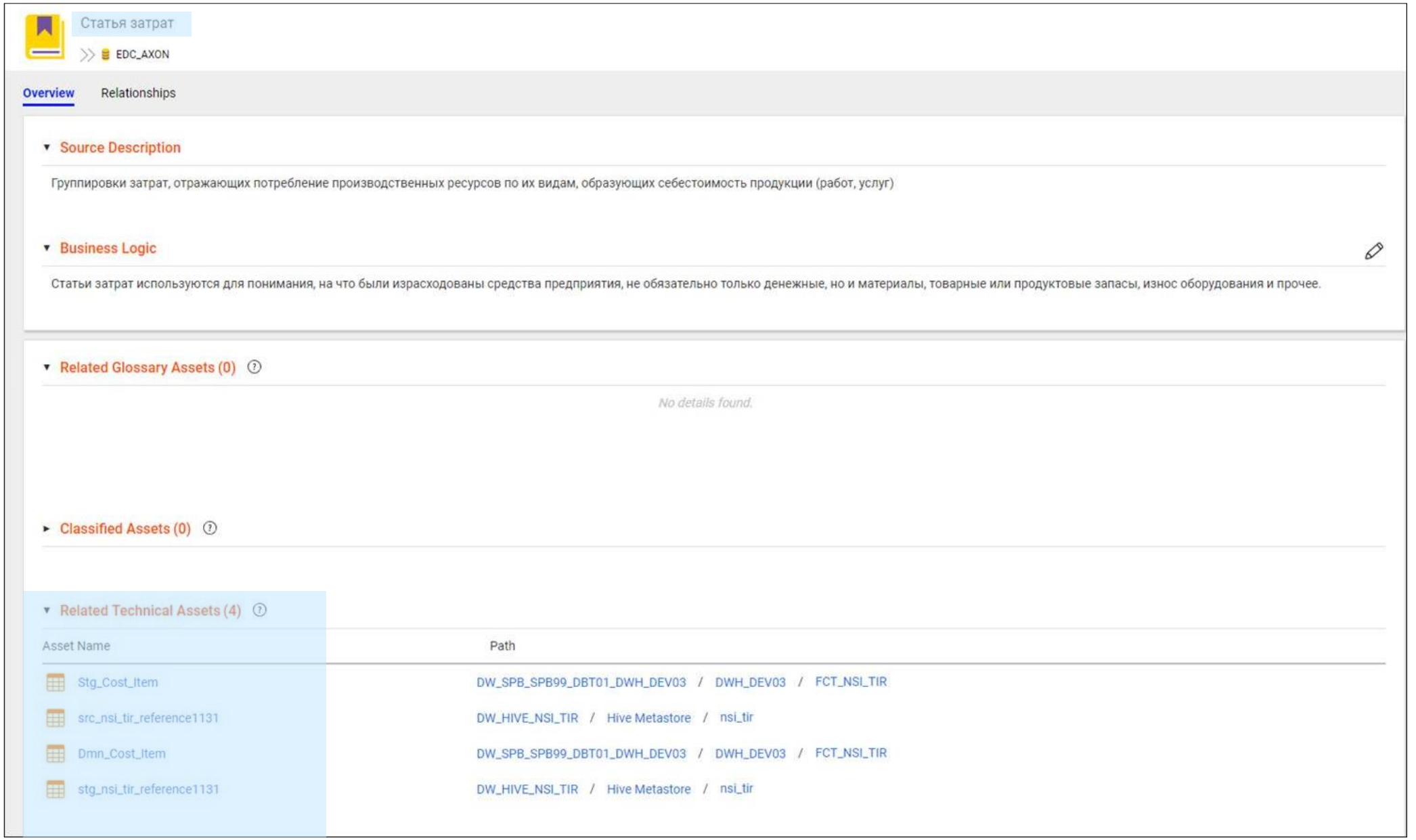


Статья затрат
src_nsi_tir_reference1131
DW_HIVE_NSI_TIR > Hive Metastore > nsi_tir

Overview | Columns | Lineage and Impact | Relationships

	Name	Business Title	Data Domains	Null | Distinct | Non-Distinct %	Source Data Type / Inferred Data Types
1	c_code	Код		0 100 0	string (N/A) / Decimal(14) | 0.13% +2 more
2	c_description	Наименование		0 24.66 75.34	string (N/A) / String(149) | 100.00%
3	c_fld2247rref	ВидРасходовНУ		3.16 0.36 96.48	string (N/A) / Decimal(32) | 0.53% +3 more
4	c_fld2826rref	ВидСтатьиЗатратМСФОКР		3.16 0.81 96.03	string (N/A) / Decimal(32) | 3.86% +3 more
5	c_fld2827rref	ВидСтатьиЗатратМСФООАР		3.16 0.76 96.08	string (N/A) / Decimal(32) | 3.86% +3 more
6	c_fld2934rref	СтатьяДвиженияДенежныхСредств		3.16 5.43 91.41	string (N/A) / Decimal(32) | 85.42% +3 more
7	c_fld5426rref	ВидСтатьиЗатратМСФООР		3.16 0.64 96.20	string (N/A) / Decimal(32) | 4.05% +3 more
8	c_fld5580rref	ЗависимостьОтОбъемаПродаж		0 0.04 99.96	string (N/A) / Decimal(32) | 87.89% +3 more
9	c_fld5581rref	ВидУправления		0 0.04 99.96	string (N/A) / Decimal(32) | 87.12% +3 more
10	c_fld5582	ВлияниеНаВеличинуБазыНП		0 0.03 99.97	string (N/A) / Decimal(2) | 100.00% +3 more
11	c_fld5584	СписыватьЛимитыНаОсвоение		3.16 0.03 96.81	string (N/A) / Decimal(2) | 100.00% +3 more
12	c_fld5585rref	ВидТранспорта		3.16 0.11 96.73	string (N/A) / Decimal(32) | 99.28% +3 more
13	c_fld5586rref	ВидРынка		3.16 0.06 96.78	string (N/A) / Decimal(32) | 99.28% +3 more
14	c_fld5587	МожетБытьРБП		3.16 0.03 96.81	string (N/A) / Decimal(2) | 100.00% +3 more
15	c_fld5588	ИспользуетсяДляКонтроляЛимитовПоОУУУ		3.16 0.03 96.81	string (N/A) / Decimal(2) | 100.00% +3 more

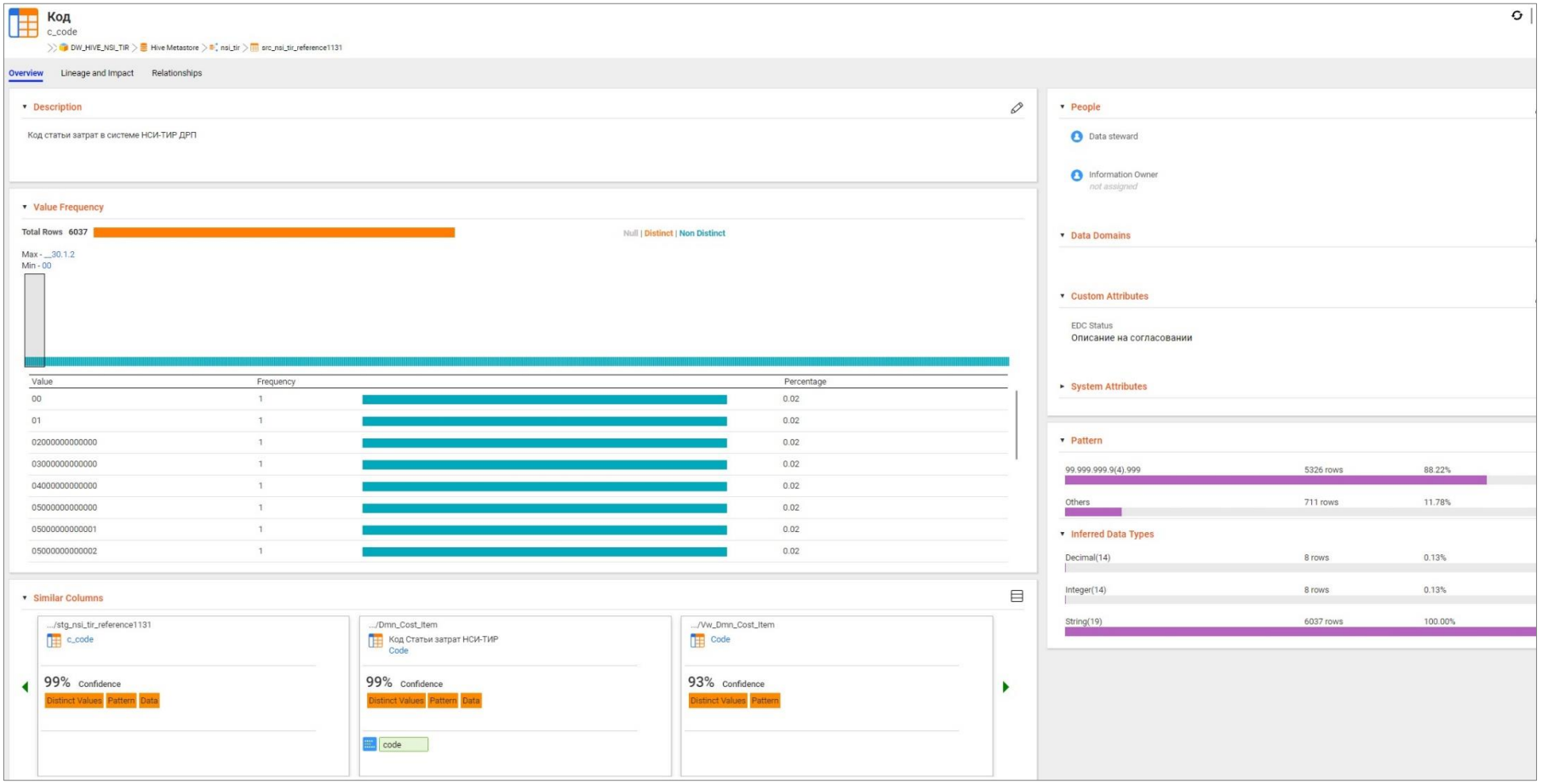
Связь бизнес-терминов (объектов глоссария) и технических полей в каталоге данных (EDC)



Профилирование данных в каталоге (EDC)

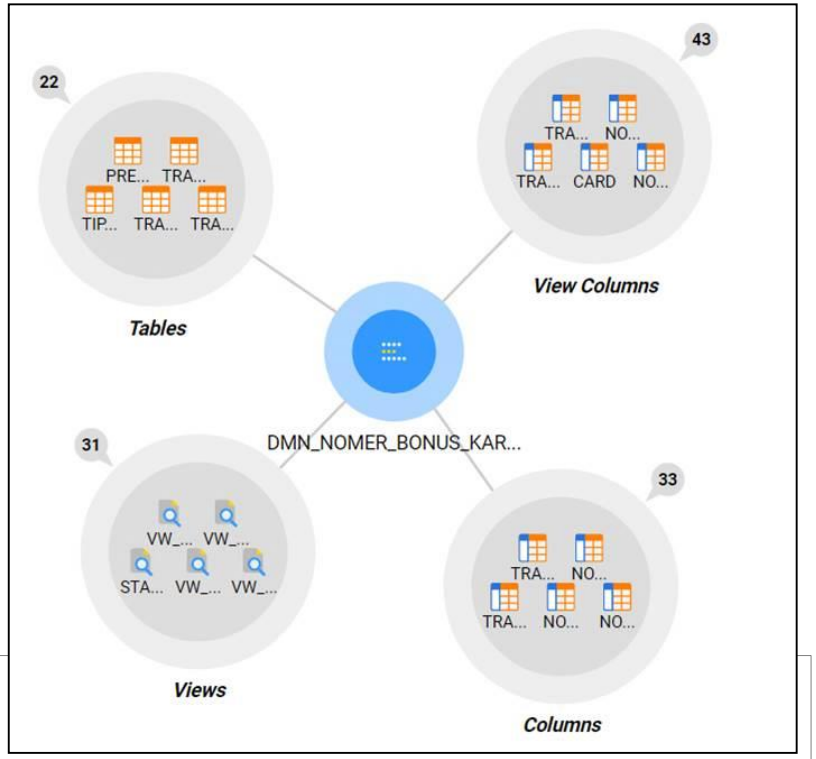
Профилирование позволяет выявить:

- типы данных
- маски
- значения и с какой частотой они встречаются в данном поле
- степень схожести с полями других таблиц по следующим параметрам: наименования, похожесть значений, похожесть типов, похожесть масок



Разметка данных в каталоге данных (EDC)

Применение разметки данных: автоматическое определение доменов, принадлежности к коммерческой тайне и персональным данным



DMN_NOMER_BONUS_KARTY

Проверяет данные по маскам бонусных карт

Resource Name: **DataDomain**

Add Business Title

PDN

Overview | Lineage and Impact | Relationships

▾ Description

Композитный домен проверяет таблицы на вхождение персональных данных о клиенте: ФИО, дата рождения, номер телефона, E-mail

▾ Source Description

Композитный домен с содержанием пунктов КТ-040 (ПДН)

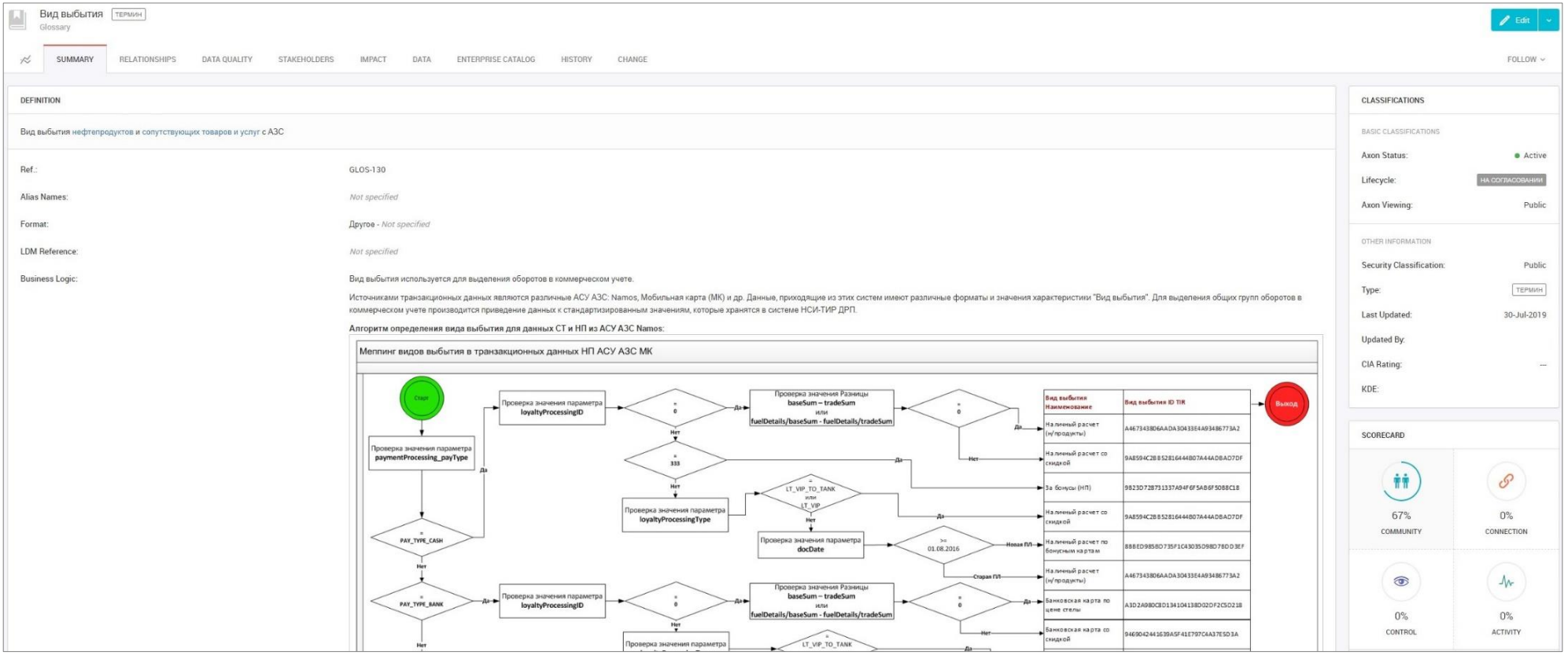
▾ Found in

Name	Resource Type	Business Term
DW_SPB99_DBT02_DWH	JDBC	

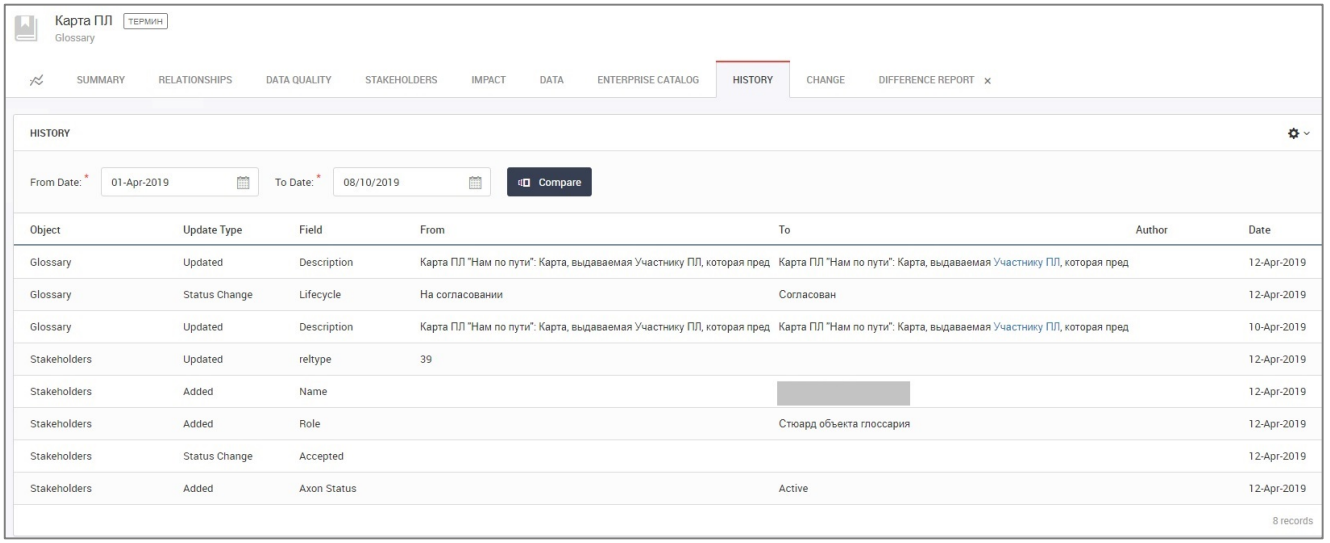
Instance 1 :

VLADELCY_BONUSNYX_KART_OLD

Бизнес-глоссарий: основной функционал (Axon)

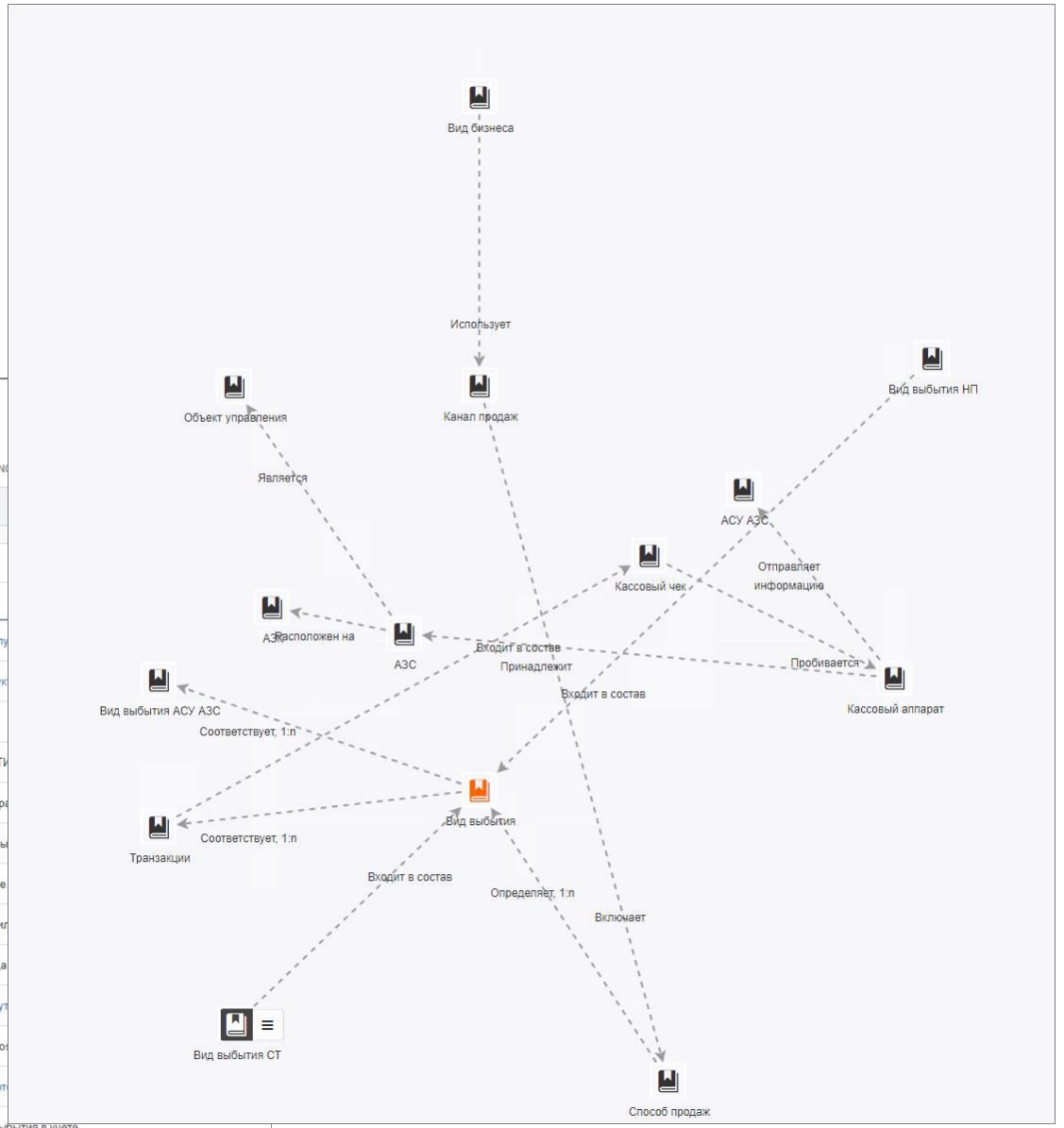


- ведение паспортов бизнес-сущностей, определений, методик расчета
- привязка к владельцам, стюардам и т.п.
- привязка к бизнес-процессам, тегам
- ведение истории изменений



Ведение связей между объектами глоссария. Граф логических связей (Axon)

- иерархии между объектами
- логические связи, которые могут характеризоваться характером связи, обязательностью и множественностью



Вид выбытия ТЕРМИН
Glossary

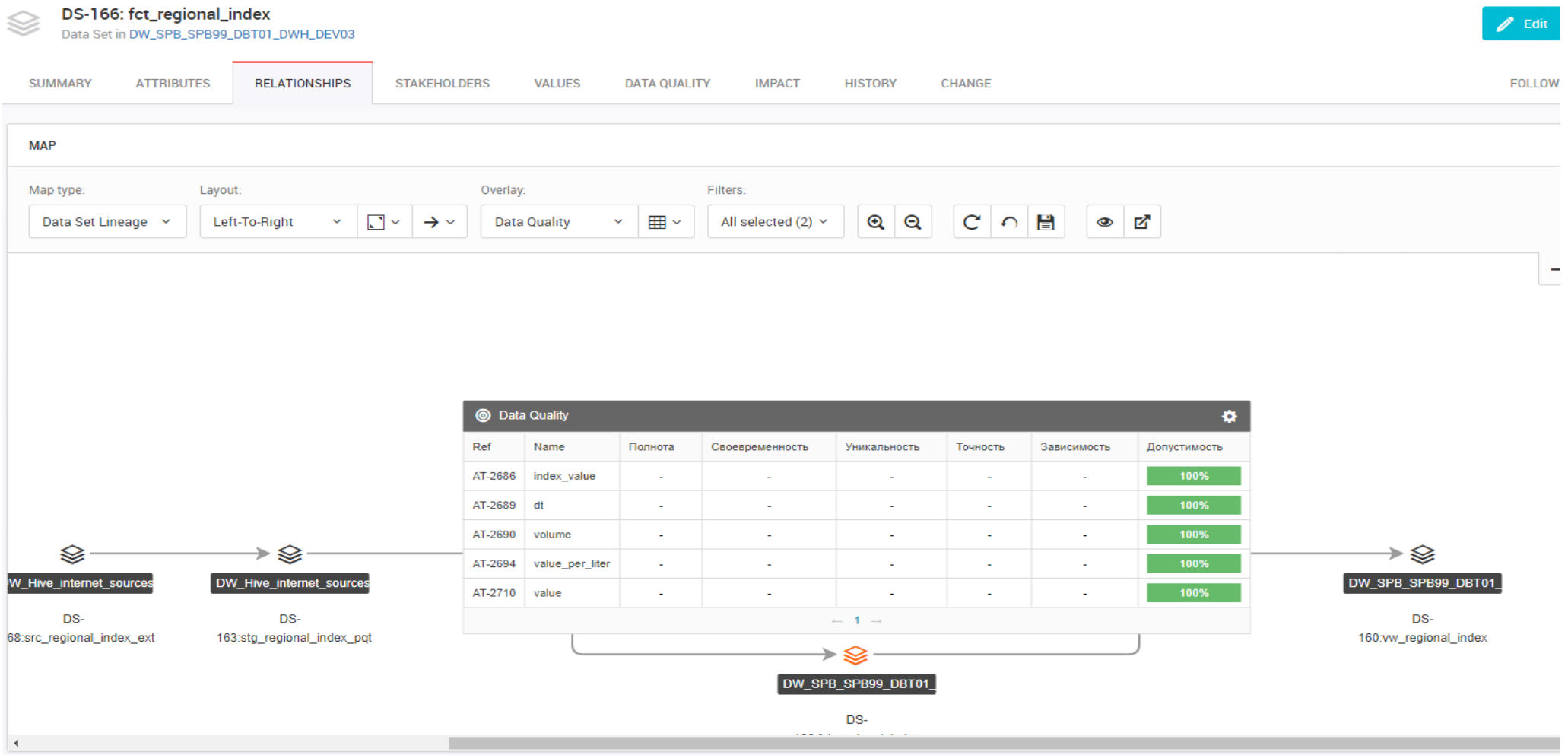
SUMMARY | RELATIONSHIPS | DATA QUALITY | STAKEHOLDERS | IMPACT | DATA | ENTERPRISE CATALOG | HISTORY | CHANG

Hierarchy | Relationships

GLOSSARY HIERARCHY

Glossary	Type	Relationship Type	Description
Вид выбытия	ТЕРМИН		Вид выбытия нефтепродуктов и сопу
Признак Вида выбытия нефтепродуктов	АНАЛИТИКА		Признак вида выбытия нефтепродук
Наименование Вида выбытия	АНАЛИТИКА		Наименование вида выбытия
Код Вида выбытия НСИ-ТИР	АНАЛИТИКА		Код вида выбытия в системе НСИ-Ти
Идентификатор Вида выбытия НСИ-ТИР	АНАЛИТИКА		Идентификатор вида выбытия в спра
Наименование Вида выбытия (расчетное)	АНАЛИТИКА		Расчетное наименование вида выбы
Наименование Вида выбытия СТиУ (расчетное)	АНАЛИТИКА		Расчетная аналитика наименование
Код Вида выбытия МК	АНАЛИТИКА		Код виды выбытия в АСУ АЗС Мобил
Дата окончания Вида выбытия	АНАЛИТИКА		Дата окончания использования вида
Идентификатор Вида выбытия СТиУ АСКУ	АНАЛИТИКА		Идентификатор виды выбытия сопут
Код Вида выбытия Namos	АНАЛИТИКА		Код виды выбытия в АСУ АЗС Namos
Идентификатор Вида выбытия НП АСКУ	АНАЛИТИКА		Идентификатор виды выбытия нефт
Дата начала Вида выбытия	АНАЛИТИКА		Дата начала использования вида выбытия в учете
Идентификатор Вида выбытия МК	АНАЛИТИКА		Идентификатор вида выбытия в АСУ АЗС Мобильная карта
Идентификатор Вида выбытия Namos	АНАЛИТИКА		Идентификатор вида выбытия в АСУ АЗС Namos

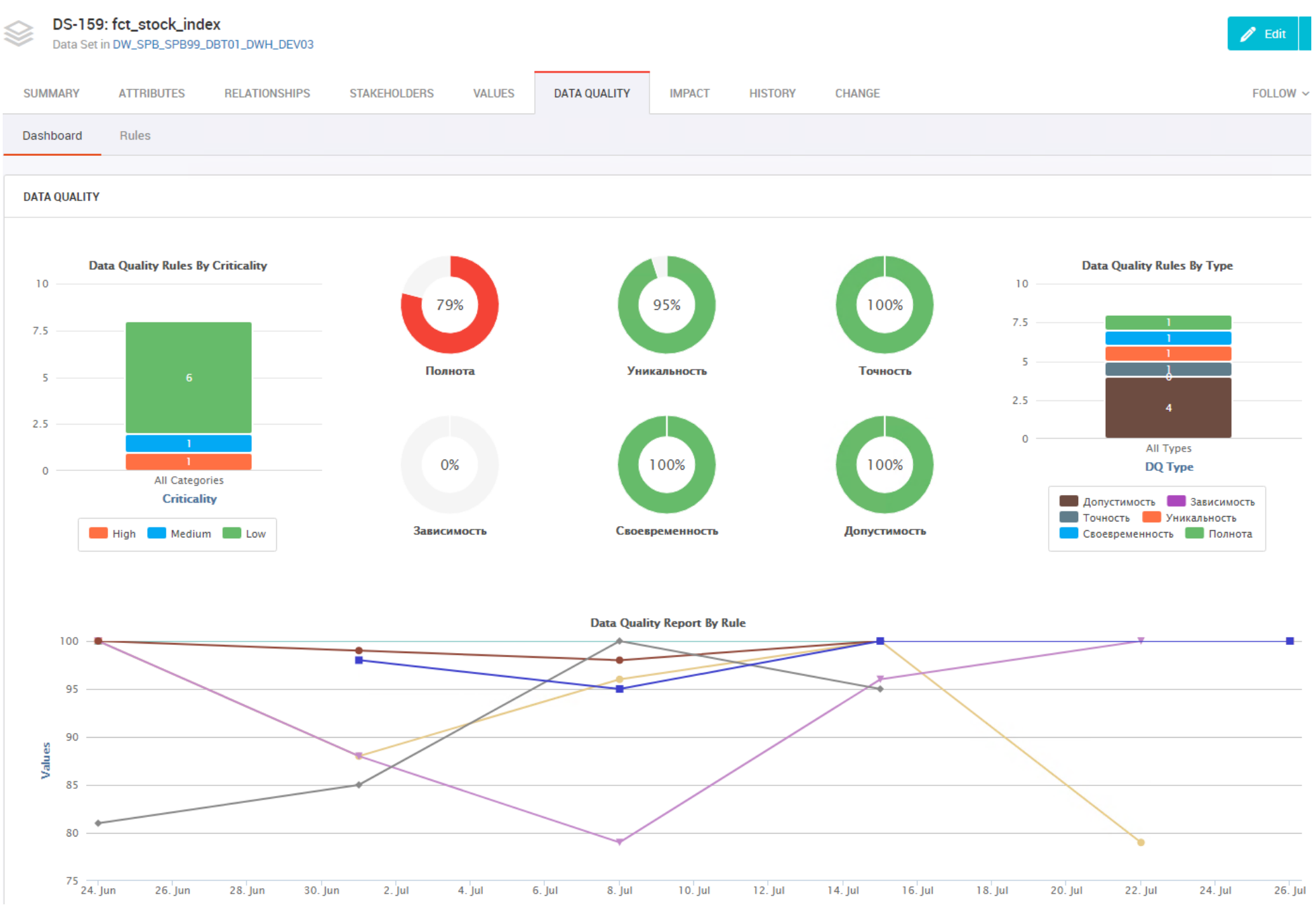
Пользовательский портал: результаты проверок качества на цепочках происхождения данных (Axon)



INBOUND RELATIONSHIPS

Attribute Name	Source System	Related Data Set	Related Attribute	Type	Scope
index_value	DW_Hive_internet_sources	DS-163: stg_regional_index_pqt	index_value	Источник	
dt	DW_Hive_internet_sources	DS-163: stg_regional_index_pqt	index_date	Источник	

Сводные дашборды по качеству данных на пользовательском портале (Axon)



Сценарий раздельной работы «озера данных» и системы управления данными

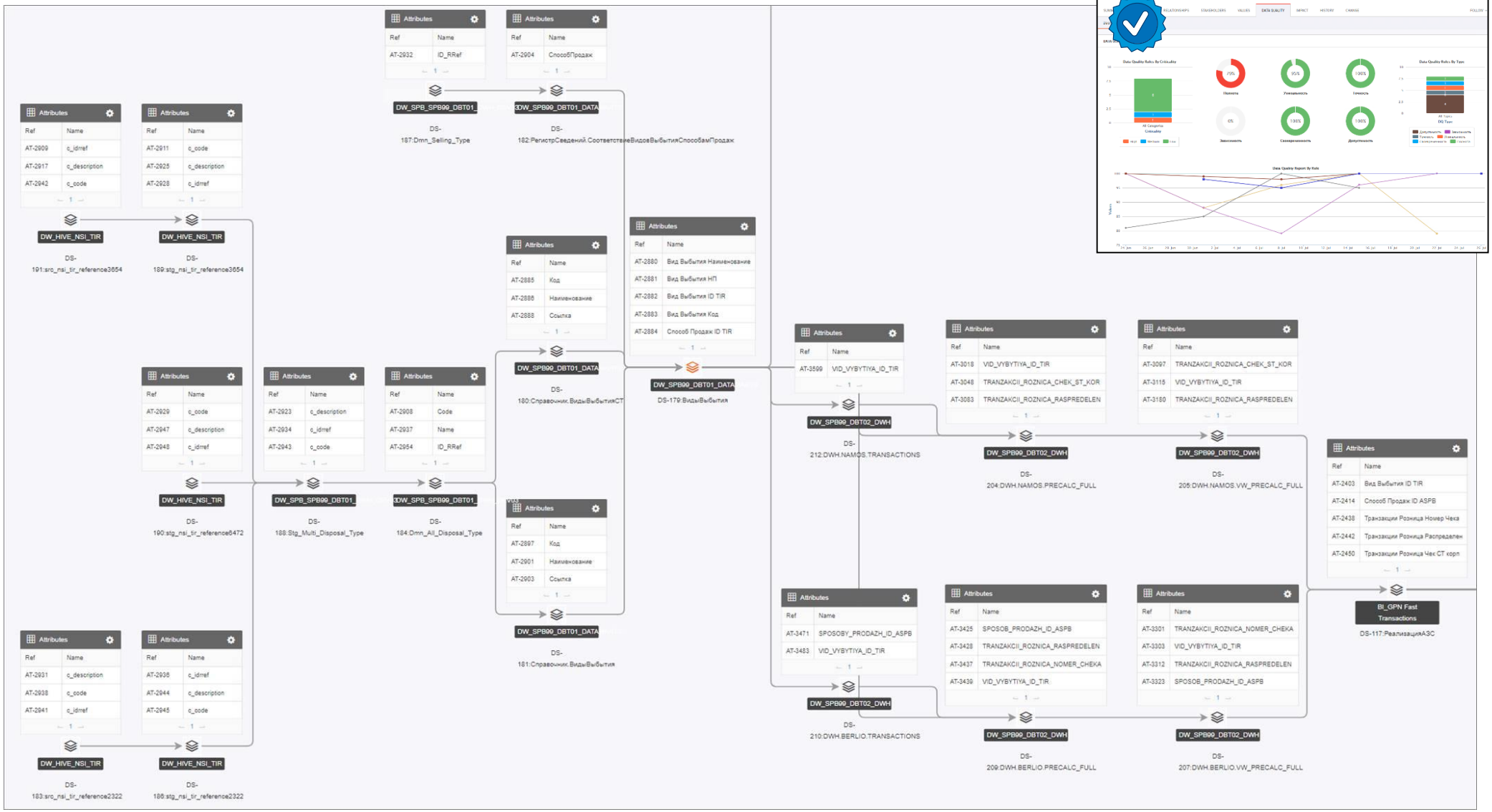
До мая 2019 года разработка «озера данных» велась без использования ETL-средств Informatica (ручной код, MS SSIS)

Для этого сценария функционал системы управления данными работает в режиме ручного заполнения

Сравнение функциональности управления данными для сценариев:

Функционал	Целевой сценарий: ETL Informatica	Сценарий: ETL другими средствами
Сбор и актуализация метаданных источников, ведение единого каталога данных	ДА в EDC в автоматическом режиме	ДА в Axon в ручном режиме
Ведение описаний в карточках БД/таблиц/полей	ДА в EDC в ручном режиме	ДА в Axon в ручном режиме
Построение Lineage на уровне таблиц/полей	ДА в EDC в автоматическом режиме	ДА в Axon в ручном режиме
Просмотр формул, скриптов трансформации данных	ДА в EDC	НЕТ
Привязка технических метаданных к бизнес-терминам	ДА в EDC и Axon в ручном режиме	ДА в Axon в ручном режиме
Автоматическое профилирование данных: уникальность, примеры значений, похожесть, типы данных, маски данных	ДА в EDC в автоматическом режиме	НЕТ
Автоматическая разметка на КТ, ПДн	ДА в EDC в автоматическом режиме	ДА в Axon в ручном режиме

Сценарий без ETL Informatica: ручное создание Lineage, все связи создаются вручную



Технически работают подключения к источникам с построением статусов, интегральных дашбордов, динамики правил по качеству данных

Зонирование «озера данных» ДРП для обеспечения гибкости в реализации запросов

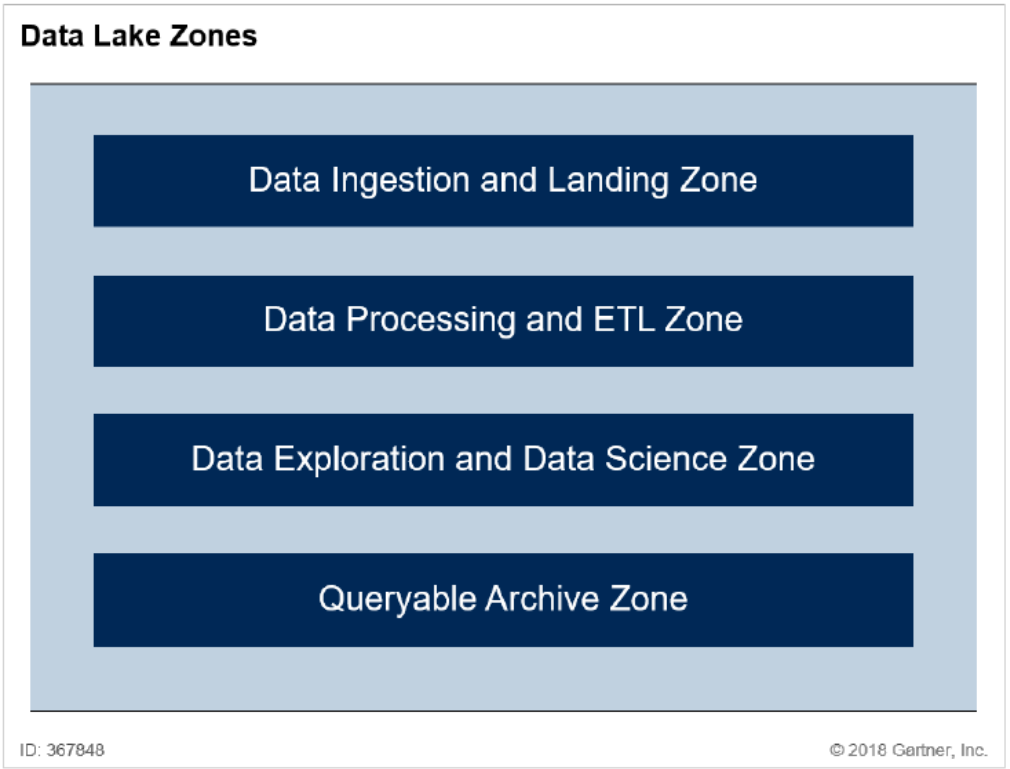
1. **Зона сырых данных** - сырой слой (копии данных из источников), где данные имеют имена как в источниках

2. **Продуктивная фабрика данных** - слой продуктивных трансформаций для аналитического хранилища с фокусом на оптимизацию, производительность и управление, объекты данных имеют целевые наименования и паспортизуются в Системе управления данными

3. **Исследовательская лаборатория** - зона «быстрой разработки» для проведения исследований и формирования прототипов решений, фокус на скорость загрузки и трансформации данных, минимум формализации, выбор инструментов не ограничивается, наименования объектов данных не регулируются, разработка может проводиться сотрудниками-аналитиками

Контур управления данными

Лучшая практика (Gartner)*



* Источник: Исследование Gartner «How to Avoid Data Lake Failures» (Август 2018)

2019 год: основные результаты программы проектов

Управление данными

- ✓ Автоматически обновляемые цепочки происхождения данных от источника до пользовательских витрин и BI-приложений
- ✓ Бизнес-глоссарий, интегрированный с уровнем представлений хранилища данных (поля отчетов и моделей расшифровываются)
- ✓ Единый управляемый каталог правил и проверок по качеству данных
- ✓ Бесшовная интеграция правил по качеству данных в ETL-цепочки
- ✓ Пользовательский портал для просмотра глоссария, происхождения данных и правил по качеству данных

Озеро данных

- ✓ Снижение затрат на интеграцию данных для всех связанных с аналитикой проектов *(KPI – не менее 50% всех проектов ДРП на конец 2019 года используют инфраструктуру Озера данных)*
- ✓ Доступность данных на всех слоях конвейера для пользователей
- ✓ Интеграция с инструментами Data Science (Лаборатория + Spark + In-database Python, R)
- ✓ Промышленные инструменты ETL и хранения для всех типов данных, потоковая и пакетная обработка
- ✓ Промышленные инструменты для создания интерфейсов интеграции

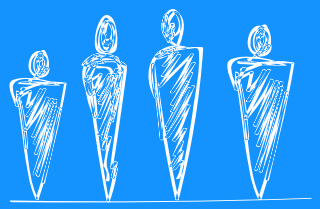
2019 год: основные характеристики программы проектов

Задачи

Интеграция компонент

Интеграция с DevOps

Производительность и стабильность



18 сотрудников в команде, без внешних подрядчиков

Хранилище данных (январь - сентябрь 2019)

10 Тб объем данных

1350 таблиц и view

2000 Заданий в сутки

Фокусы

Только актуальные бизнес-задачи

Управляем только важными данными

Стандартизация практик и процессов

Эволюция централизованного аналитического подразделения в ДРП

Период	Название подразделения	Орг. Изменение	Численность сотрудников
2012 – 2013 гг	Единое ответственное лицо за проекты BI и НСИ	[все работы выполняются внешними подрядчиками]	1
2014 год	Группа развития BI	Новые роли: • «Архитектор BI» • «Разработчик BI» • «Системный аналитик BI»	4
2015 – 2016 годы		Новая роль • «Эксперт по качеству данных» • «Администратор BI»	10
2017 год	Центр компетенций BI	[Рост внутреннего ресурса, отказ от внешних подрядчиков] Новая роль • «Архитектор данных»	30
2018 год	Офис управления данными	[Старт создания системы управления данными, сразу ставка на внутренние ресурсы] Новые роли: • «Эксперт по управлению метаданными» • «Математик-программист» (Data scientist) • «Системный аналитик Озера данных» • «Архитектор Озера данных»	40
2019 год	Центр аналитических решений	+ Центр компетенций по Data Science + Центр развития компетенций в аналитике Новые роли: • «Тренер по аналитике» • «Разработчик Озера данных» • «Администратор систем управления данными»	60

Центр аналитических решений ДРП как шаг к DataOps

1. Разработка аналитических решений

BI-приложения

Хранилища данных для аналитики

Решения для «больших данных»

2. Управление данными

Управление качеством данных

Единый каталог данных и бизнес-словарь

База знаний и консультирование по данным

3. Центр компетенций Data Science

Модели и прототипы на данных

Участие в проектах с аналитической составляющей

Центр знаний и ресурсов

4. Бизнес-партнерство в части новых проектов

Оценка новых инициатив с точки зрения данных

Участие в проектных командах

5. Управление централизованной аналитической инфраструктурой

Аналитические песочницы и инструменты

Инжиниринг и окружение

Стандартизация: данные как сервис

6. Обучения и тесты по методам и инструментам анализа данных

Python, R, SQL, BI

Методы анализа данных

Машинное обучение и т.п.

Контакты и вопросы

Иван Черницын

CHERNITSYN.IG@gazprom-neft.ru

Руководитель Центра аналитических решений
Дирекция региональных продаж
ПАО «Газпром нефть»