

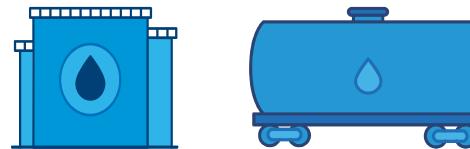
Практика создания «умного» озера данных

15.10.2019

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



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



- 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**

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





> 1 100
пользователей



>90 систем-источников
>50 внешних и несистемных
источников данных



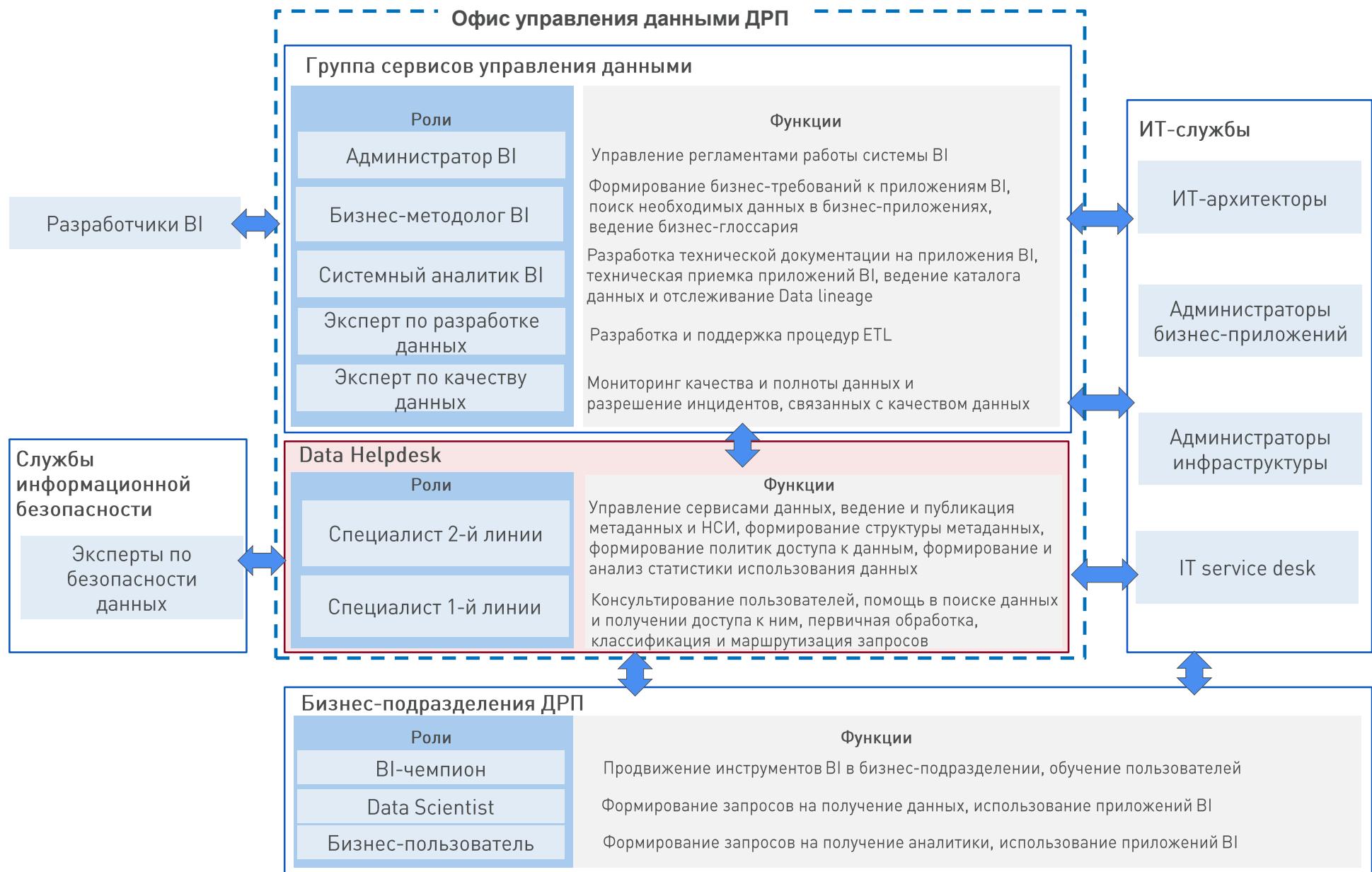
> 250
аналитических
приложений

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



Making Your Landscape Less Uncertain — Focus Areas



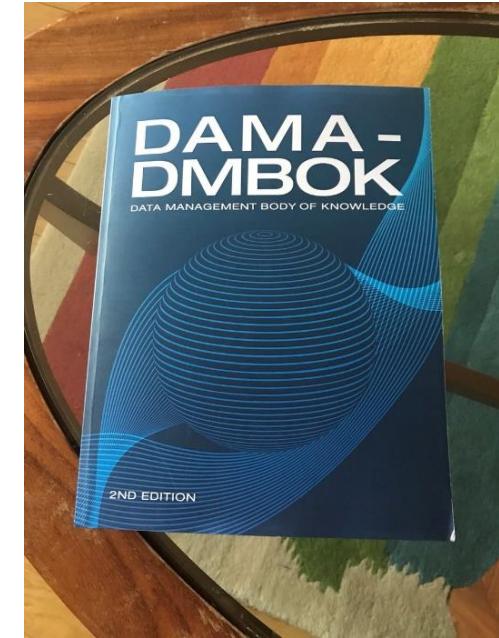
Component —
not core



Operational and
analytics skills



Security and
metadata

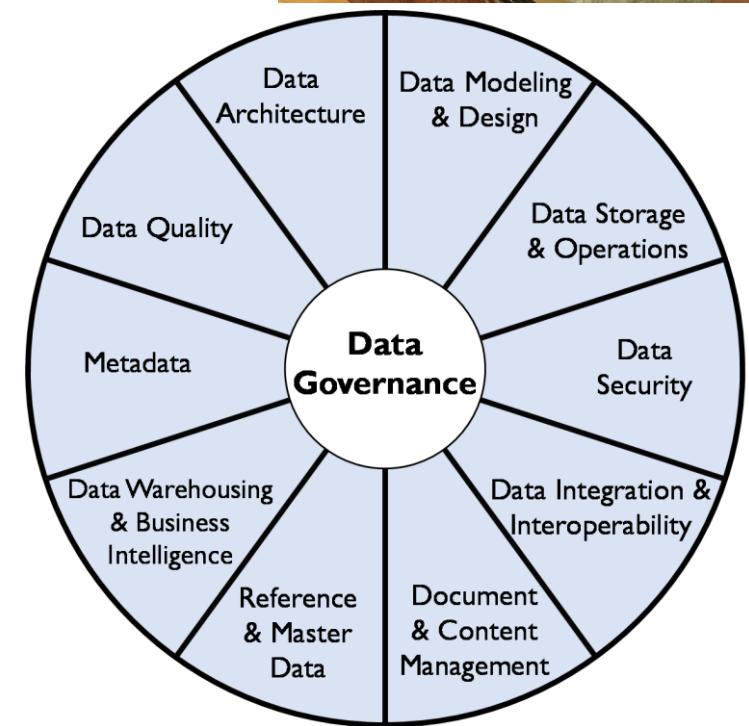
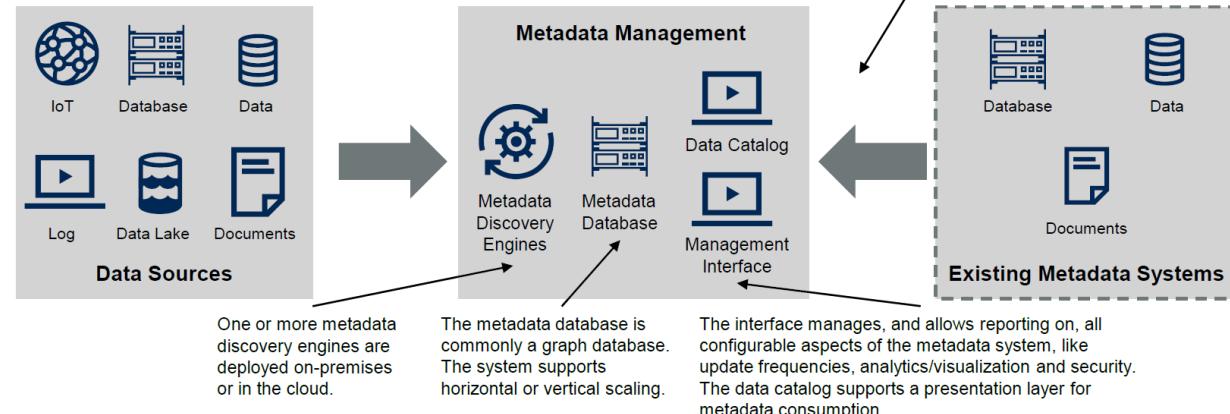


Data Catalogs Are the New Black in Data Management and Analytics

Published: 13 December 2017 ID: G00338777

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

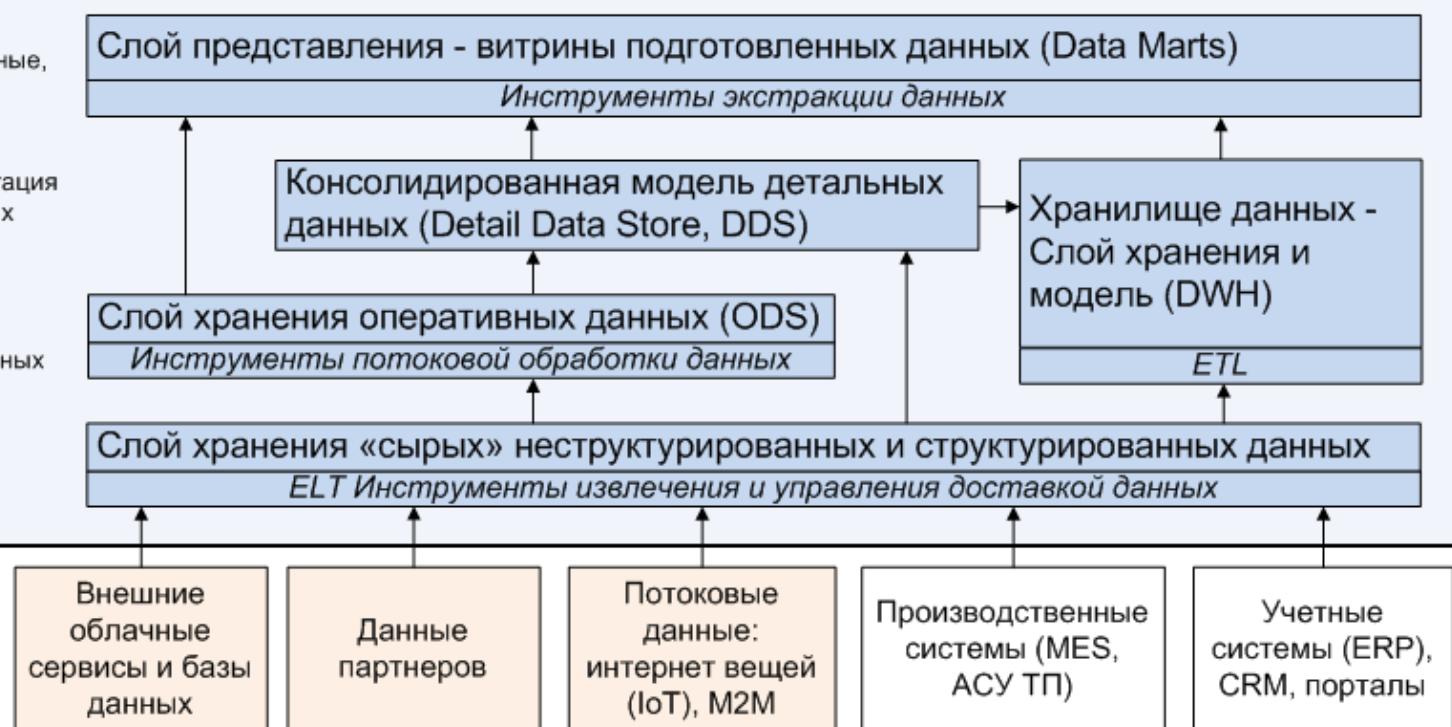
Metadata Management Landscape



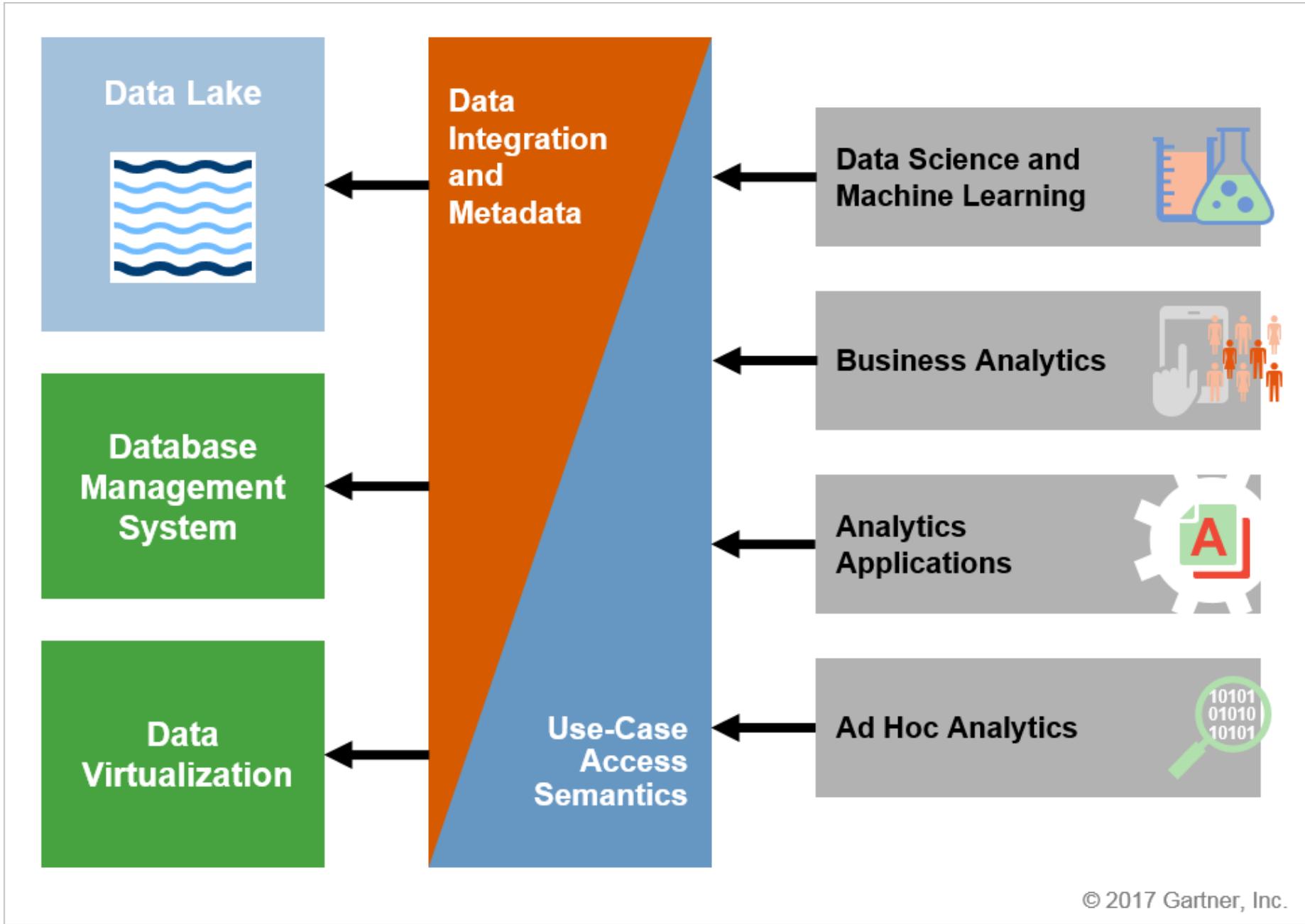
Озеро данных (Big Data Lake) - программно-аппаратный комплекс как совокупность различных массивов данных, объединенных в единую логическую среду, для обеспечения доступа, управления и преобразования данных. Цель создания озер данных – дать высококвалифицированным аналитикам возможность изучать неочищенные,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

«Озеро данных» (Big Data Lake) + аналитическое Хранилище данных

- Необходимые для бизнес-пользователей данные, очищенные, агрегированные и в легком для понимания формате
- Бизнес-логика, расчеты и агрегация
- Объединение данных из разных источников
- Расширенные компоненты Корпоративного Хранилища данных
- Полные копии источников данных в едином формате, включая исторические данные



- 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.



© 2017 Gartner, Inc.



- 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

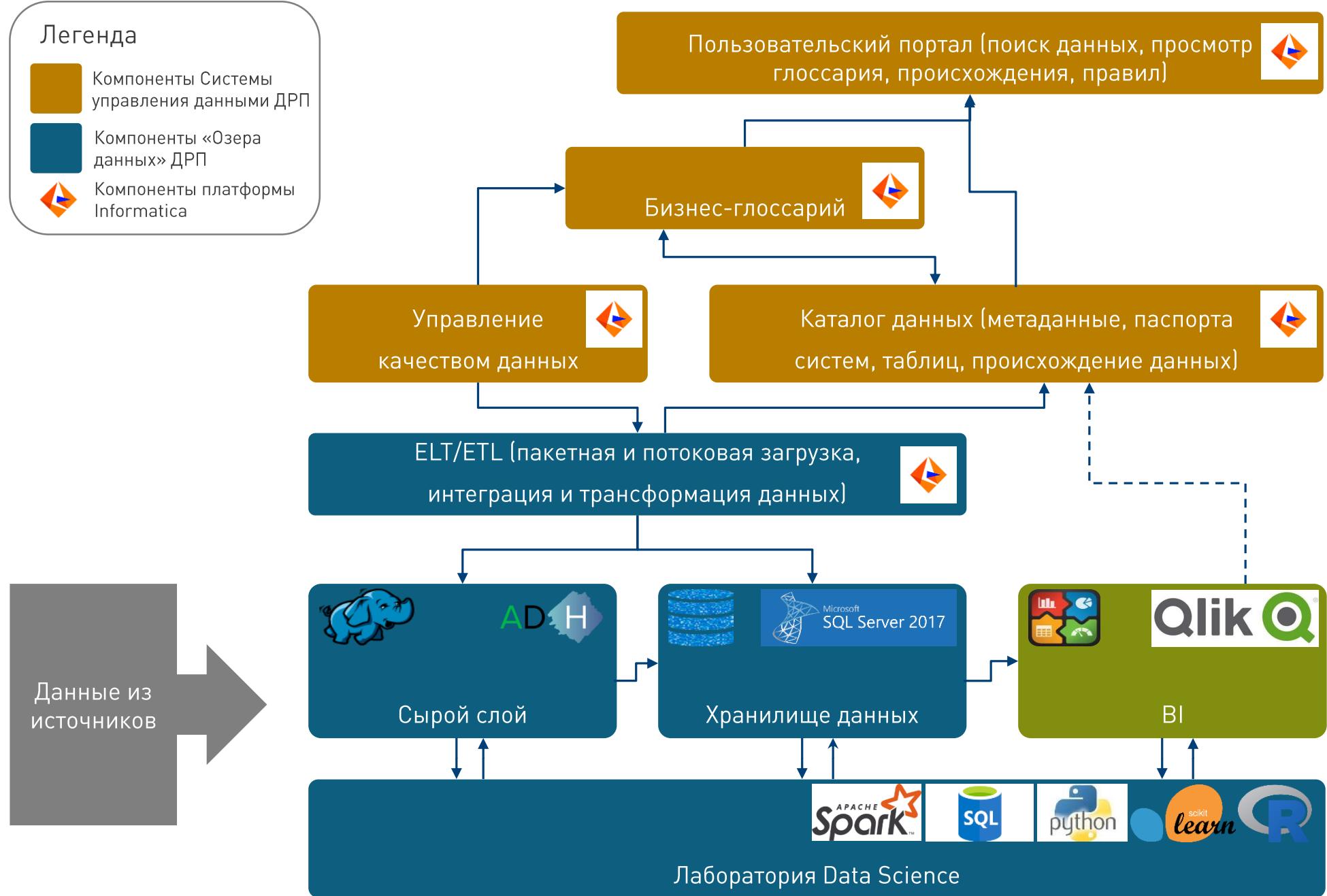
- Единый каталог правил по качеству данных
- Подсистема управления качеством данных, интегрированная с «озером данных»
- Измерение, отображение и аудит всех проверок по качеству данных в контуре «озера данных» в разрезе потоков обработки данных

- 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

Легенда

- Компоненты Системы управления данными ДРП
- Компоненты «Озера данных» ДРП
- Компоненты платформы Informatica



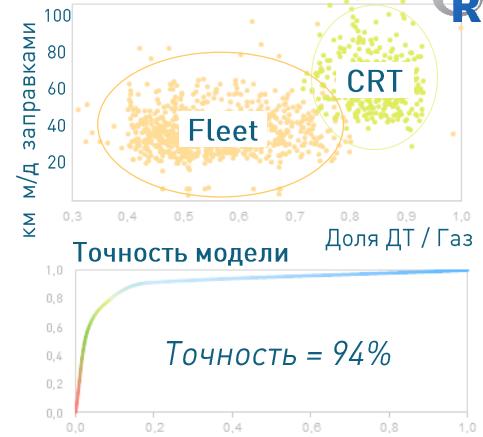


СЕГМЕНТЫ Клиентского контура

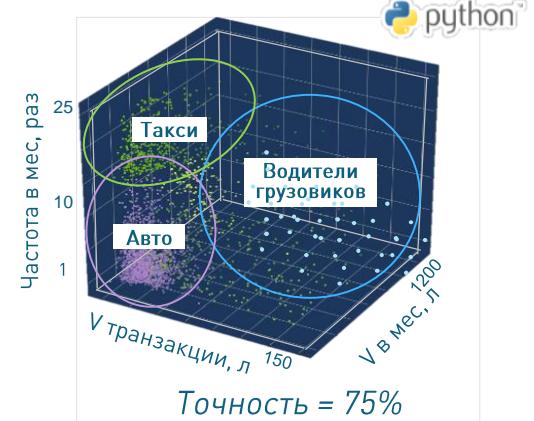


Сегментация клиентов на основе методов машинного обучения и продвинутой аналитики

1) Подсегменты КП



2) Подсегменты ПЛ



- ✓ 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



Особенности:

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

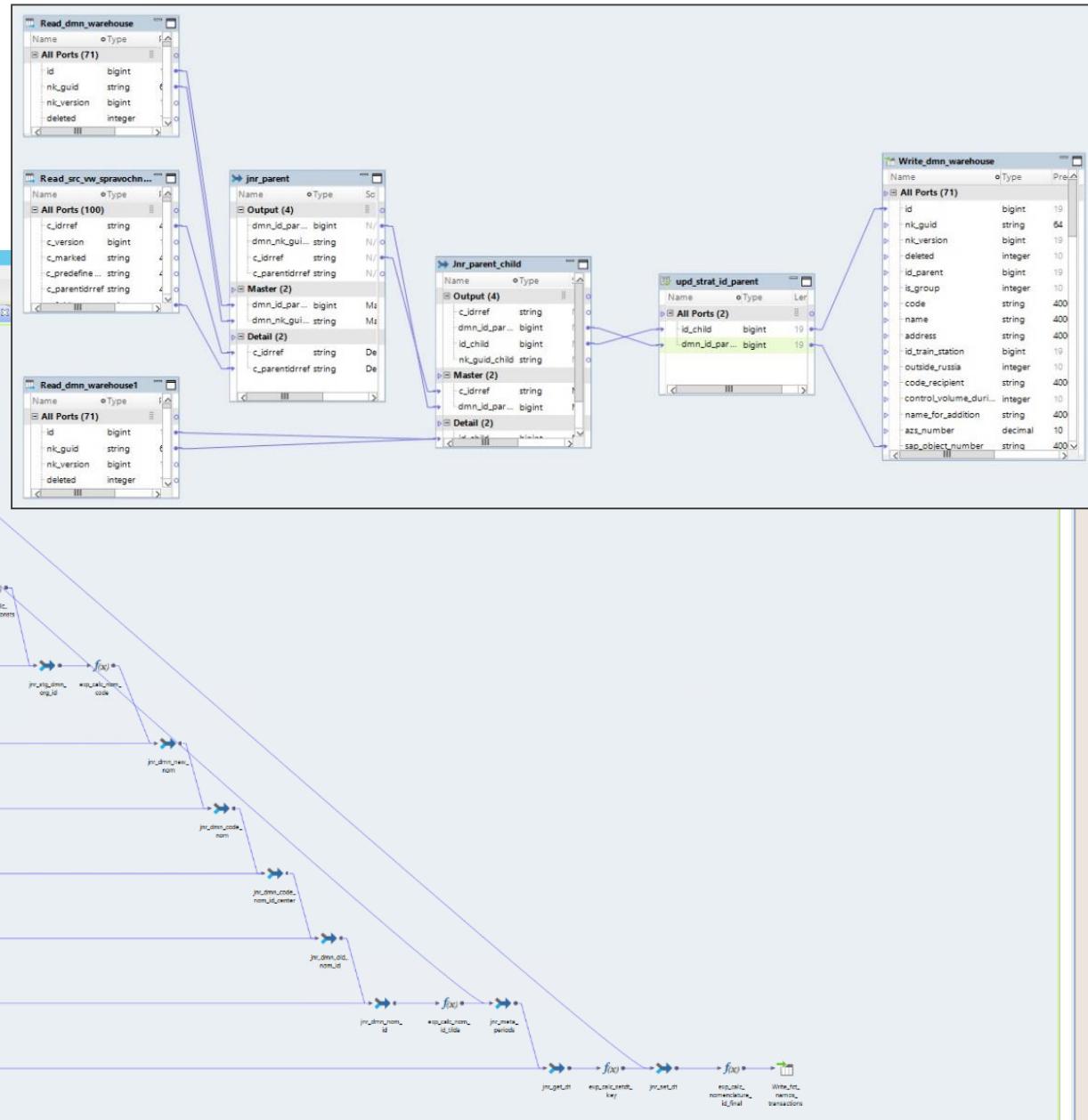


! Требуется значительное количество соглашений по практике разработки для автоматической работы инструментов управления данными

The screenshot shows the Talend Data Integration environment. On the left, the Object Explorer lists various projects and components, including mappings for mobile card transactions and FCT tables. The main workspace displays a complex ETL job flow with numerous components connected by blue arrows. Key components visible include:

- Readers:** m_STG_Mobile_Card_Transactions_Add_L, m_FCT_MK_transactions, m_FCT_Namos_transactions.
- Processors:** Address Value, Aggregator, Association, Case Conversion, Classifier, Comparison, Consolidation, Data Masking, Data Processor, Decision, Exception, Expression, Filter, Hierarchical, Java, Joiner, Key Generator, Labeler, Lookup, Match, Merge, Normalizer, Parser, Rank, Read, Relational Table, REST Web Service, Router, Sequence Generator, Sorter, SQL, Standardizer, Union, Update Strategy, Web Service, Weighted Average, Window, Write.
- Writers:** Write_FCT_Namos_transactions.

The flow involves multiple parallel paths of data processing, with many steps involving joins and calculations (e.g., f(x) functions). The overall complexity highlights the need for extensive schema mapping and agreement between developers.



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

DQ_NAMOS_LAYER0

1 Specify General Properties | 2 Select Source | 3 Specify Settings | 4 Specify Rules

7 Applied

Name

- DQ_Namos_PLOTNOST_00001
- DQ_Namos_VREMYAOPERACII_00001**

Name: DQ_Namos_VREMYAOPERACII_00001
Rule Condition: IIF ((DATE_COMPARE(DATA_I_VREMYA_NACHALA_SMENY, DATA_I_VREMYA_REGISTRACII_PRODAZHI)=1) , '0', '1')
Output Column: DQ_Namos_VREMYAOPERACII_00001

DQ_Namos_VREMYAOPERACII_00002

Name: DQ_Namos_VREMYAOPERACII_00002
Rule Condition: IIF ((DATE_COMPARE(DATA_I_VREMYA_REGISTRACII_PRODAZHI, DATA_I_VREMYA_OKONCHANIYA_SMENY)=1) , '0', '1')
Output Column: DQ_Namos_VREMYAOPERACII_00002

DQ_NAMOS_RNN_00001

Name: DQ_NAMOS_RNN_00001
Rule Condition: IIF(RRN != RRN_MARKETINGOVOJ_TRANZAKCII, '1', '0')
Output Column: DQ_NAMOS_RNN_00001

DQ_NAMOS_LAYER0

Back to Overview | Profile Run 5 of 5 | 59 of 62 Columns | 7 of 7 Rules | 100000 Rows | 08/14/2019 11:34:06 AM

General

Total rows 100000

Null	0	0%
Distinct	2	0.01%
Non-distinct	99998	99.99%

Values

2 distinct values (2 Non-unique, 0 Unique) Sort By: Frequency Descending

Value	Frequency	Length	Percentage
1	81389	1	81.39%
0	18611	1	18.61%

Datatype

Datatype	Documented
string(10)	100.00%
Decimal(1)	100.00%
Fixed Length String(1)	100.00%
Integer(1)	100.00%
String(1)	100.00%

Patterns

Pattern	Percentage
9	100.00%

Data Preview

[DQ_NAMOS_RNN_00001 : "0"]
First 100 rows only

IDENTIFIKATO...	NOMER_SMENY	DATA_I_VREMY...	DATA_I_VREMY...	DATA_I_VREMY...
1 '0000025061	920	10/07/2019 04:0...	10/08/2019 03:5...	10/07/2019 04:0...
2 '0000025061	920	10/07/2019 04:0...	10/08/2019 03:5...	10/07/2019 04:0...
3 '0000025061	920	10/07/2019 04:0...	10/08/2019 03:5...	10/07/2019 04:0...

Доступно управление типами (аспектами) и уровнем критичности правил

Informatica | New | Open | **Discovery** | Design | Scorecards | Library

DQ_NAMOS_LAYER0

Back to Overview

Columns and Rules (20) | DQ_NAMOS_RNN_00001 |

General

Total rows 1000000

- Null 0 0%
- Distinct 2 0.01%
- Non-distinct 99998 99.99%

2 distinct values (2 Non-unique, 0 Unique)

Value	Frequency	Length	Percentage
1	81369	1	81.39%
0	18611	1	18.61%

Rule used: DQ_NAMOS_RNN_00001
Input columns: RRN, RRN_MARKETINGOVVOJ_TRANZAKCII

Add Tag | Add Comment

Datatype

string(10) Documented

Decimal(1) 100.00%
Fixed Length String(1) 100.00%
Integer(1) 100.00%
String(1) 100.00%

Length: (min → max) 1 → 1
Value (min → max) 0 → 1
Average: 0 | Sum: | Standard Deviation: 0

Patterns

9 100.00%
IDENTIFIKATO... NOMER_SMENY... 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...

Data Domain

No data domain is inferred

Business Term

DS-159: fct_stock_index
Data Set in DW_SPB_SPB99_DBT01_DWH_DEV03

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

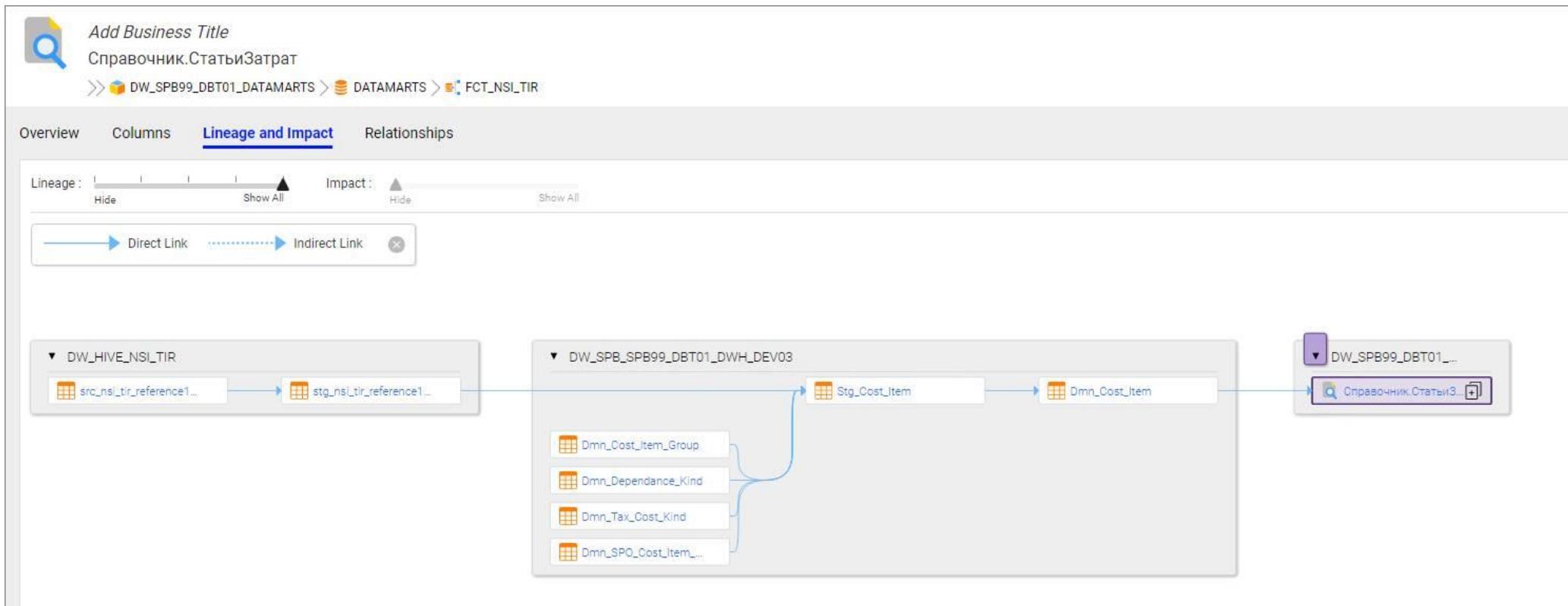
Dashboard Rules

DETAILS (8 records)

Cont...	Rule Type	Ref.	Description	A...	Attributes	System	Data Set	Target	Last ...	Last Result Date
	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_StockIndexValue	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_StockIndexValue	Числовое значение не отрицательное	Active	volume	DW_SPB_SPB99_DB	fct_stock_index	90%	100%	26-Jul-2019

8 records

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



Цепочки происхождения данных в каталоге разворачиваются до полей таблиц



Add Business Title

Справочник.СтатьиЗатрат

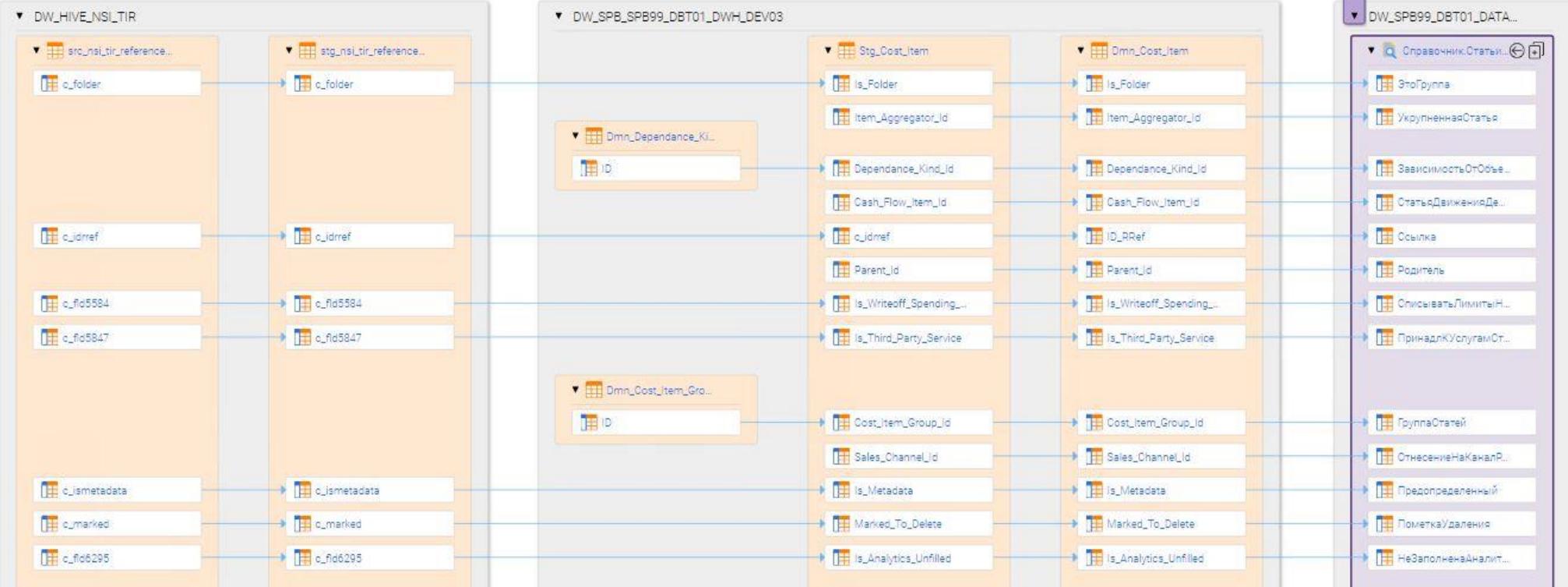
>> DW_SPB99_DBT01_DATAMARTS > DATAMARTS > FCT_NS1_TIR

Overview Columns Lineage and Impact Relationships

Lineage : Hide Show All

Impact : Hide Show All

Direct Link Indirect Link



Add Business Title PRECALC_FULL

DW_SPB99_DBT02_DWH > DWH > MK

Overview Columns Lineage and Impact Relationships

PRECALC_FULL > app_mk_transaction_test

Direct Link Indirect Link

Properties

RESOURCE NAME: app_mk_transaction_test

RESOURCE TYPE: BDMScanner

PRECISION: 18

INPUT: false

OUTPUT: true

SCALE: 7

UPPER: 1

GROUP NAME: Group

LOWER: 1

EXPRESSION: IF[mapНом2Скидка_Номенклатура_ЕКТУ_Не_Учитывать_Скидку = 'Да', b_Транзакции_Розница_Цена_Реализ_Нетто_С_Налогом_Руб_Литр, b_Транзакции_Розница_Цен

VARIABLE: false

LAST MODIFIED: 2019-10-01T10: 9.023Z

↑

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

Add Business Title Справочник.СтатьиЗатрат

DW_SPB99_DBT01_DATAMARTS > DATAMARTS > FCT_NSL_TIR

Overview Columns Lineage and Impact Relationships

Справочник.СтатьиЗатрат > app_load_cost_item_full

Direct Link Indirect Link

App_Load_Cost_Item_Full

DW_HIVE_NSLTIR

stg_nsl_tir_reference...

c_fld5582

m_Load_DIM_Cost_Item

Read_stg_nsl_tir_reference1...

Group

c_fld5582

Exp_calc_hash_string_and_boolean_dft_vals

Exp_calc_hash_string_and_boolean_dft_vals

c_fld5582_pass

Is_Affect_To_Tax_Base

Exp_calc_hash_string_and_boolean_dft_vals

Group

Is_Affect_To_Tax_Base

Properties

RESOURCE NAME: app_load_cost_item_full

RESOURCE TYPE: BDMScanner

UPPER: 0

LOWER: 0

INPUT: false

OUTPUT: true

VARIABLE: false

PRECISION: 1

SCALE: 0

GROUP NAME: Exp_calc_hash_string_and_boolean_dft_vals

EXPRESSION: SUBSTR(SUBSTR(C_FLD5582_PASS, 2, 1), 0, 1)

LAST MODIFIED: 2019-06-11T15:38:46.355Z

→

Статья затрат

src_nsi_tir_reference1131

>> DW_HIVE_NSITIR > 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	ВидСтатиЗатратМСФОКР		3.16	0.81	96.03	string (N/A)	Decimal(32) 3.86% +3 more
5 c_fld2827rref	ВидСтатиЗатратМСФОАР		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	ВидСтатиЗатратМСФОР		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	МожетБытьРБЛ		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

Статья затрат

» EDC_AXON

Overview Relationships

Source Description

Группировки затрат, отражающих потребление производственных ресурсов по их видам, образующих себестоимость продукции (работ, услуг)

Business Logic

Статьи затрат используются для понимания, на что были израсходованы средства предприятия, не обязательно только денежные, но и материалы, товарные или продуктовые запасы, износ оборудования и прочее.

Related Glossary Assets (0) ?

No details found.

Classified Assets (0) ?

Related Technical Assets (4) ?

Asset Name	Path
Stg_Cost_Item	DW_SPB_SPB99_DBT01_DWH_DEV03 / DWH_DEV03 / FCT_NSI_TIR
src_nsi_tir_reference1131	DW_HIVE_NSI_TIR / Hive Metastore / nsi_tir
Dmn_Cost_Item	DW_SPB_SPB99_DBT01_DWH_DEV03 / DWH_DEV03 / FCT_NSI_TIR
stg_nsi_tir_reference1131	DW_HIVE_NSI_TIR / Hive Metastore / nsi_tir

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

The screenshot displays the EDC interface for the 'Код' (Code) column. At the top, the navigation path shows: DW_HIVE_NSI_TIR > Hive Metastore > nsi_tir > src_nsi_tir_reference1131. Below the navigation, there are tabs for Overview, Lineage and Impact, and Relationships. The Overview tab is selected.

Description: Код статьи затрат в системе НСИ-ТИР ДРП.

Value Frequency: Total Rows: 6037. A horizontal bar chart shows the distribution of values. The most frequent value is '00' at 0.02%. Other values like '01', '020000000000', etc., also appear at 0.02% frequency. A legend indicates Null, Distinct, and Non-Distinct categories.

Value	Frequency	Percentage
00	1	0.02
01	1	0.02
020000000000	1	0.02
030000000000	1	0.02
040000000000	1	0.02
050000000000	1	0.02
050000000001	1	0.02
050000000002	1	0.02

Similar Columns: Three columns are shown: .../stg_nsi_tir_reference1131 (c_code), .../Dmn_Cost_Item (Код Статьи затрат НСИ-ТИР), and .../Vw_Dmn_Cost_Item (Code). Each has a 99% confidence level and includes links for Distinct Values, Pattern, and Data.

People: Data steward (not assigned), Information Owner (not assigned).

Data Domains: EDC Status, Описание на согласование.

Custom Attributes: EDC Status, Описание на согласование.

System Attributes: None listed.

Pattern: 99.999.999.9(4).999 (5326 rows, 88.22%), Others (711 rows, 11.78%).

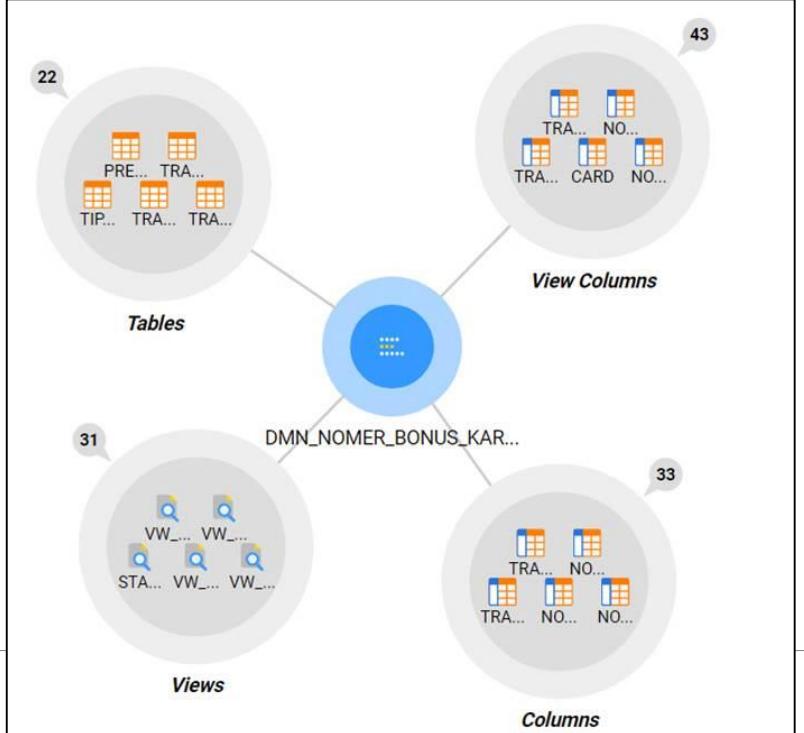
Inferred Data Types: Decimal(14) (8 rows, 0.13%), Integer(14) (8 rows, 0.13%), String(19) (6037 rows, 100.00%).

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



The diagram illustrates the structure of the DMN_NOMER_BONUS_KARTY resource. At the center is a blue circle labeled "DMN_NOMER_BONUS_KARTY". Four arrows point from this central node to four surrounding circles, each representing a different type of data entity:

- Tables**: Represented by a grey circle containing icons for PRE..., TRA..., TIP..., TRA..., TRA... (with a count of 22).
- Views**: Represented by a grey circle containing icons for VW..., VW..., VW... (with a count of 31).
- Columns**: Represented by a grey circle containing icons for TRA..., NO..., TRA..., NO..., NO... (with a count of 33).
- View Columns**: Represented by a grey circle containing icons for TRA..., NO..., TRA..., CARD, NO... (with a count of 43).

Below the diagram, the "DataDomain" resource page is shown:

- Add Business Title**: A button to add a title.
- PDN**: A button to add a business term.
- Overview**: The active tab, showing a brief description of the composite domain.
- Lineage and Impact**: A tab for lineage analysis.
- Relationships**: A tab for relationship mapping.
- Description**: A section describing the domain's purpose: "Композитный домен проверяет таблицы на вхождение персональных данных о клиенте: ФИО, дата рождения, номер телефона, E-mail".
- Source Description**: A section describing the source of the domain: "Композитный домен с содержанием пунктов КТ-040 (ПДН)".
- Found in**: A table showing instances where the domain is used:

Name	Resource Type	Business Term
DW_SPB99_DBT02_DWH	JDBC	
- Instance 1 :** A list of tables under this instance:

 VLADELCY_BONUSNYX_KART_OLD
--

Вид выбытия TERMIN

Glossary

Edit FOLLOW

SUMMARY RELATIONSHIPS DATA QUALITY STAKEHOLDERS IMPACT DATA ENTERPRISE CATALOG HISTORY CHANGE

DEFINITION

Вид выбытия нефтепродуктов и сопутствующих товаров и услуг с АЗС

Ref.: GLOS-130

Alias Names: Not specified

Format: Другое - Not specified

LDM Reference: Not specified

Business Logic:

Вид выбытия используется для выделения оборотов в коммерческом учете.

Источники транзакционных данных являются различными АСУ АЗС: Namos, Мобильная карта (МК) и др. Данные, приходящие из этих систем имеют различные форматы и значения характеристики "Вид выбытия". Для выделения общих групп оборотов в коммерческом учете производится приведение данных к стандартизованным значениям, которые хранятся в системе НСИ-ТИР ДРП.

Алгоритм определения вида выбытия для данных СТ и НП из АСУ АЗС Namos:

CLASSIFICATIONS

Axon Status: Active

Lifecycle: НА СОГЛАСОВАНИИ

Axon Viewing: Public

OTHER INFORMATION

Security Classification: Public

Type: TERMIN

Last Updated: 30-Jul-2019

Updated By: --

CIA Rating: --

KDE: --

SCORECARD

COMMUNITY 67%	CONNECTION 0%
CONTROL 0%	ACTIVITY 0%

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

Карта ПЛ TERMIN

Glossary

Edit FOLLOW

SUMMARY RELATIONSHIPS DATA QUALITY STAKEHOLDERS IMPACT DATA ENTERPRISE CATALOG HISTORY CHANGE DIFFERENCE REPORT

HISTORY

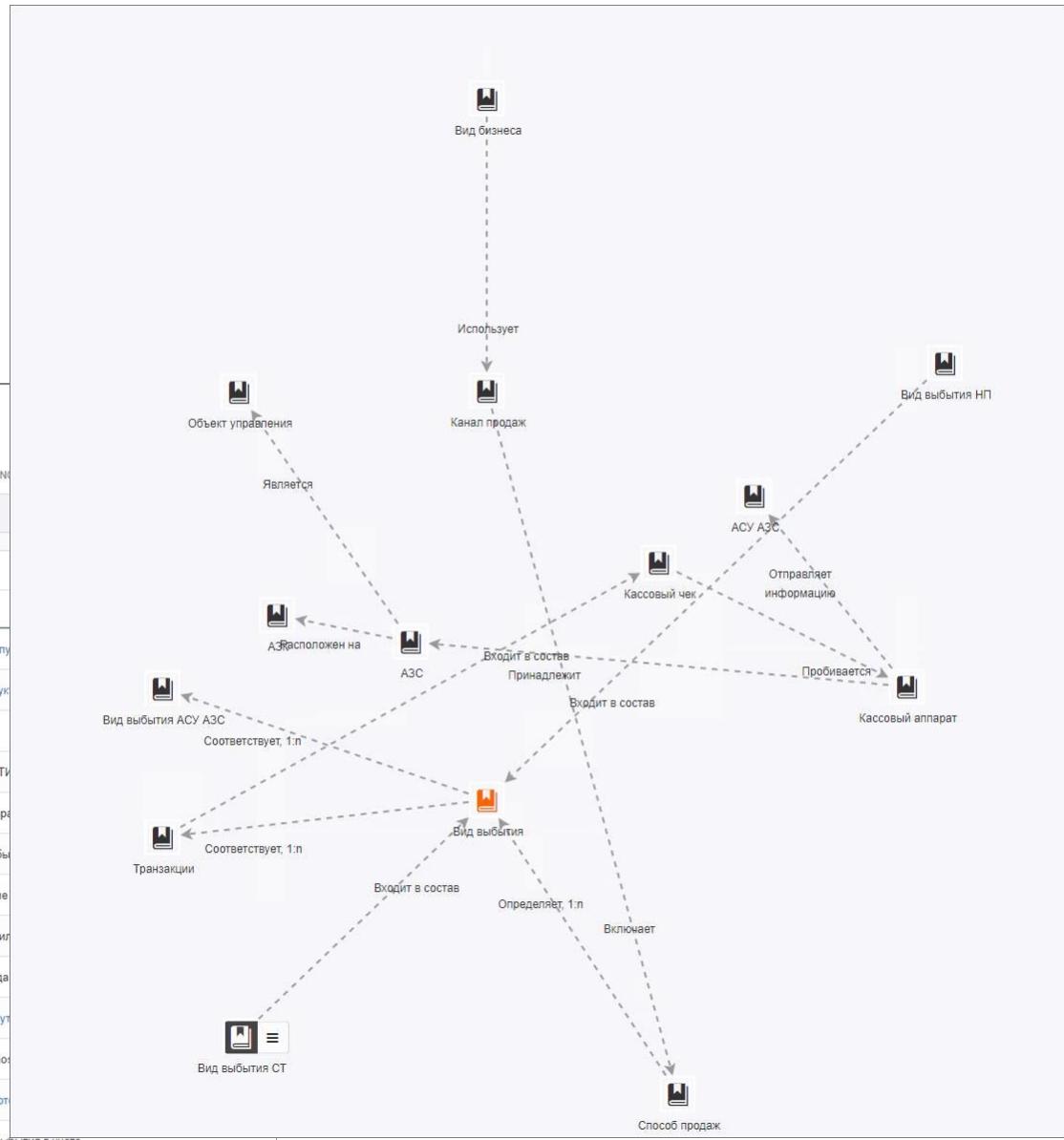
From Date: * 01-Apr-2019 To Date: * 08/10/2019 Compare

Object	Update Type	Field	From	To	Author	Date
Glossary	Updated	Description	Карта ПЛ "Нам по пути": Карта, выдаваемая Участнику ПЛ, которая пред.	Карта ПЛ "Нам по пути": Карта, выдаваемая Участнику ПЛ, которая пред.		12-Apr-2019
Glossary	Status Change	Lifecycle	На согласование	Согласован		12-Apr-2019
Glossary	Updated	Description	Карта ПЛ "Нам по пути": Карта, выдаваемая Участнику ПЛ, которая пред.	Карта ПЛ "Нам по пути": Карта, выдаваемая Участнику ПЛ, которая пред.		10-Apr-2019
Stakeholders	Updated	reltype	39			12-Apr-2019
Stakeholders	Added	Name		Стюард объекта глоссария		12-Apr-2019
Stakeholders	Added	Role				12-Apr-2019
Stakeholders	Status Change	Accepted				12-Apr-2019
Stakeholders	Added	Axon Status		Active		12-Apr-2019

8 records

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

Вид выбытия ТЕРМИН			
	SUMMARY	RELATIONSHIPS	DATA QUALITY
Hierarchy	Relationships		
GLOSSARY HIERARCHY			
Glossary	Type	Relationship Type	Description
Вид выбытия	ТЕРМИН	Вид выбытия нефтепродуктов и сопутствующих видов выбытия	
Признак Вида выбытия нефтепродуктов	АНАЛИТИКА	Признак вида выбытия нефтепродуктов	
Наименование Вида выбытия	АНАЛИТИКА	Наименование вида выбытия	
Код Вида выбытия НСИ-ТИР	АНАЛИТИКА	Код вида выбытия в системе НСИ-ТИР	
Идентификатор Вида выбытия НСИ-ТИР	АНАЛИТИКА	Идентификатор вида выбытия в справочнике НСИ-ТИР	
Наименование Вида выбытия (расчетное)	АНАЛИТИКА	Расчетное наименование вида выбытия	
Наименование Вида выбытия СТиУ (расчетное)	АНАЛИТИКА	Расчетная аналитика наименование	
Код Вида выбытия МК	АНАЛИТИКА	Код видов выбытия в АСУ АЗС Мобил	
Дата окончания Вида выбытия	АНАЛИТИКА	Дата окончания использования вида выбытия	
Идентификатор Вида выбытия СТиУ АСКУ	АНАЛИТИКА	Идентификатор видов выбытия сопутствующих видов выбытия	
Код Вида выбытия Namos	АНАЛИТИКА	Код видов выбытия в АСУ АЗС Namos	
Идентификатор Вида выбытия НП АСКУ	АНАЛИТИКА	Идентификатор видов выбытия нефтепродуктов	
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	



DS-166: fctRegional_index
Data Set in DW_SPB_SPB99_DBT01_DWH_DEV03

Edit

SUMMARY ATTRIBUTES RELATIONSHIPS STAKEHOLDERS VALUES DATA QUALITY IMPACT HISTORY CHANGE FOLLOW

MAP

Map type: Layout: Overlay: Filters:

Data Set Lineage Left-To-Right Data Quality All selected (2)

Data Quality

Ref	Name	Полнота	Своевременность	Уникальность	Точность	Зависимость	Допустимость
AT-2686	index_value	-	-	-	-	-	100%
AT-2689	dt	-	-	-	-	-	100%
AT-2690	volume	-	-	-	-	-	100%
AT-2694	value_per_liter	-	-	-	-	-	100%
AT-2710	value	-	-	-	-	-	100%

W_Hive_internet_sources → DW_Hive_internet_sources → DW_SPB_SPB99_DBT01_160:vwRegional_index

DS-68:srcRegional_index_ext → DS-163:stgRegional_index_pqt → DW_SPB_SPB99_DBT01_160:vwRegional_index

INBOUND RELATIONSHIPS

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



DS-159: fct_stock_index

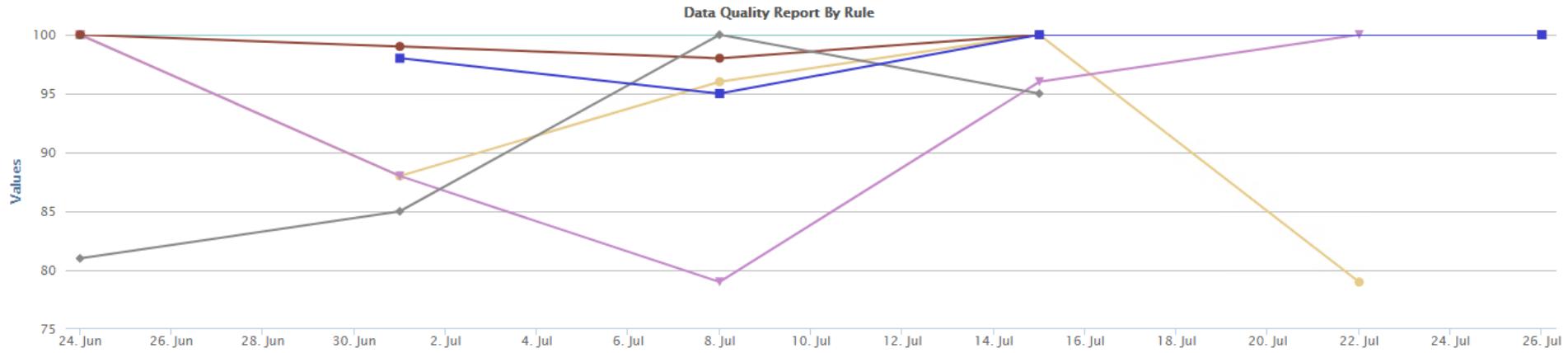
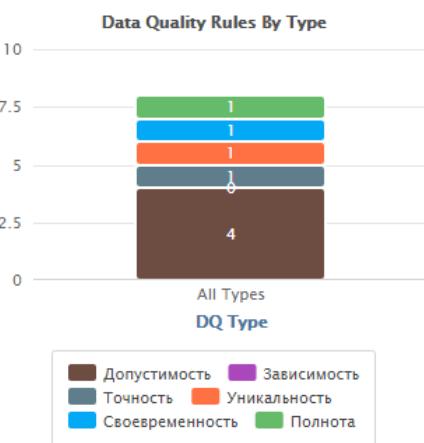
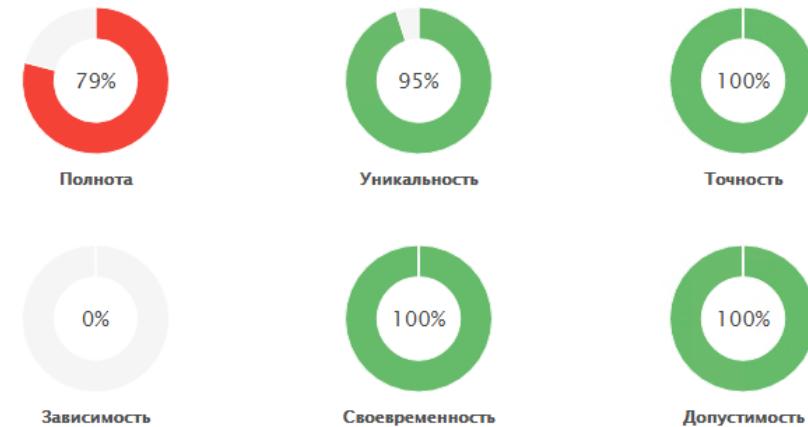
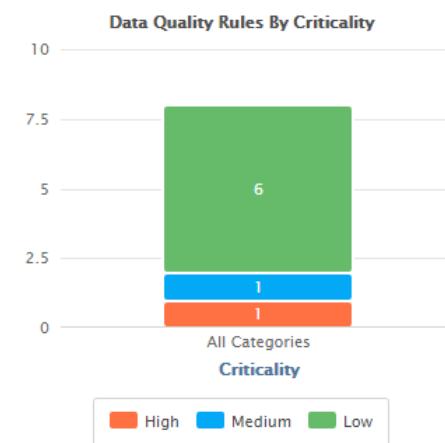
Data Set in DW_SPB_SPB99_DBT01_DWH_DEV03

Edit

[SUMMARY](#) [ATTRIBUTES](#) [RELATIONSHIPS](#) [STAKEHOLDERS](#) [VALUES](#) [DATA QUALITY](#) [IMPACT](#) [HISTORY](#) [CHANGE](#) [FOLLOW ▾](#)

Dashboard

DATA QUALITY



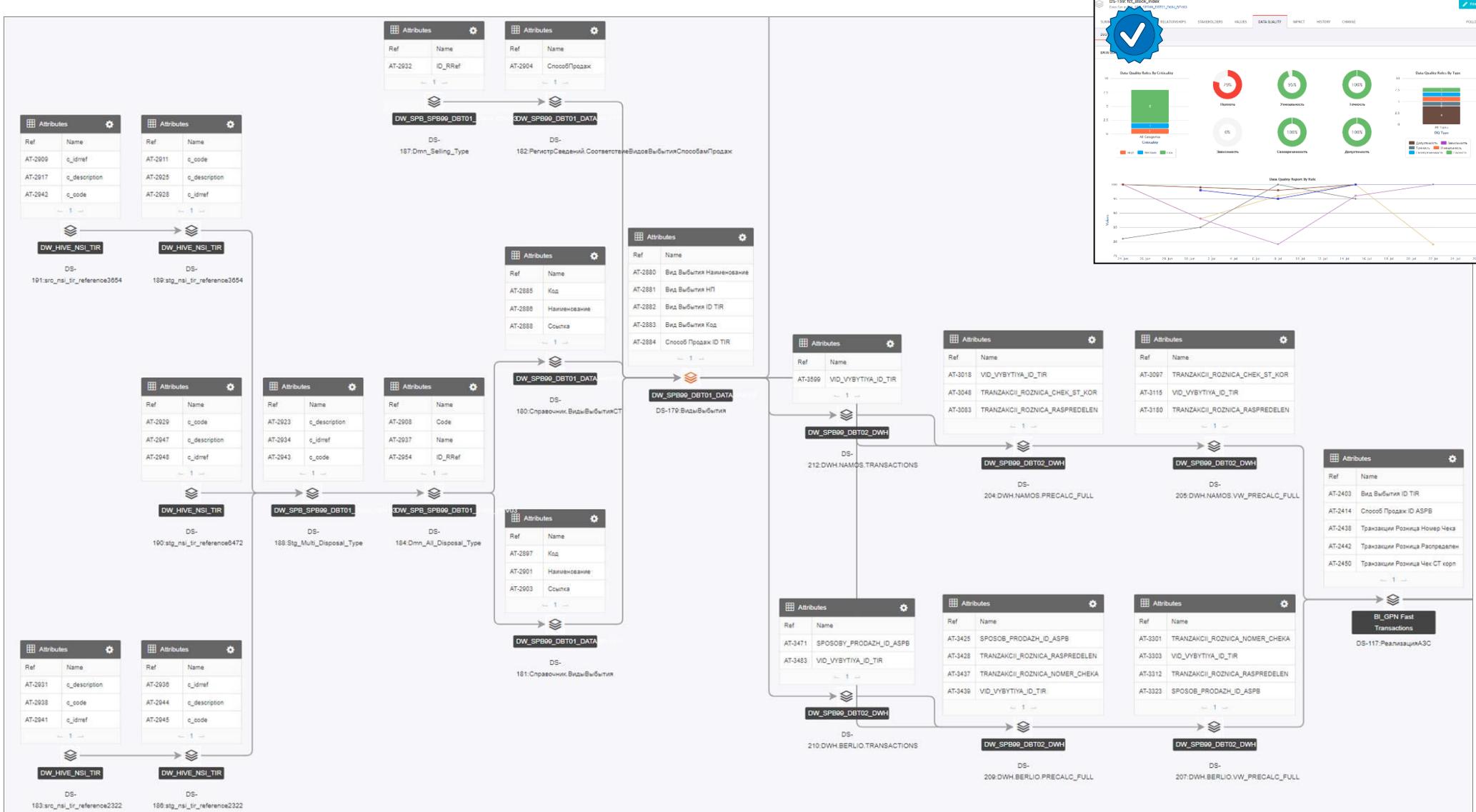
До мая 2019 года разработка «озера данных» велась без использования ETL-средств Informatica (ручной код, MS SSIS)

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

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)*

Data Lake Zones

Data Ingestion and Landing Zone

Data Processing and ETL Zone

Data Exploration and Data Science Zone

Queryable Archive Zone

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

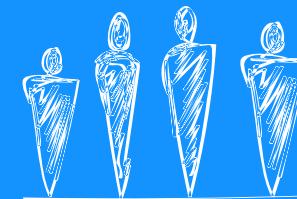
- ✓ Снижение затрат на интеграцию данных для всех связанных с аналитикой проектов (*KPI – не менее 50% всех проектов ДРП на конец 2019 года используют инфраструктуру Озера данных*)
- ✓ Доступность данных на всех слоях конвейера для пользователей
- ✓ Интеграция с инструментами Data Science (Лаборатория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

Задачи



Интеграция компонент

Интеграция с DevOps

Производительность и
стабильность

18

сотрудников
в команде,
без внешних
подрядчиков

Фокусы

Только актуальные
бизнес-задачиУправляем только
важными даннымиСтандартизация практик
и процессовХранилище данных
(январь - сентябрь 2019)

10 ТБ

объем данных

1350

таблиц и view

2000

Заданий в сутки



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: <ul style="list-style-type: none"> «Архитектор BI» «Разработчик BI» «Системный аналитик BI» 	4
		Новая роль <ul style="list-style-type: none"> «Эксперт по качеству данных» «Администратор BI» 	10
2017 год	Центр компетенций BI	[Рост внутреннего ресурса, отказ от внешних подрядчиков] Новая роль <ul style="list-style-type: none"> «Архитектор данных» 	30
2018 год	Офис управления данными	[Старт создания системы управления данными, сразу ставка на внутренние ресурсы] Новые роли: <ul style="list-style-type: none"> «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: <ul style="list-style-type: none"> «Тренер по аналитике» «Разработчик Озера данных» «Администратор систем управления данными» 	60

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

6. Обучения и тесты по методам и инструментам анализа данных

Python, R, SQL, BI
Методы анализа данных
Машинное обучение и т.п.

3. Центр компетенций Data Science

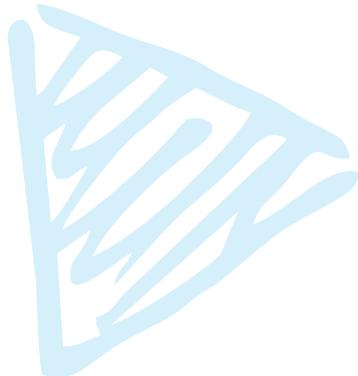
Модели и прототипы на данных
Участие в проектах с аналитической составляющей
Центр знаний и ресурсов

5. Управление централизованной аналитической инфраструктурой

Аналитические песочницы и инструменты
Инжиниринг и окружение
Стандартизация: данные как сервис

4. Бизнес-партнерство в части новых проектов

Оценка новых инициатив с точки зрения данных
Участие в проектных командах



Иван Черницын

CHERNITSYN.IG@gazprom-neft.ru

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

