Технологический подход к управлению данными

Олег Гиацинтов

Технический директор

- Технологический уровень управления данными
- Тенденции в управлении данными
- Концепт архитектуры
- Обзор инструментов и технологий

Операционный уровень управления данными

ПОИСК ДАННЫХ

- Анализ бизнес-проблематики
- Проектирование решения
- Инвентаризация данных
- Разработка дорожной карты
- Планирование оргизменений (назначение дата стюардов)
- Поиск/создание/приобретение данных
- Профилирование данных

МОНИТОРИНГ И ИЗМЕРЕНИЯ

- Проактивный мониторинг
- Панель управления (Governance dashboards)
- Анализ процессов и данных
- Отчетность, в том числе отчётность для комплаенс и риск-отчётность
- Управление ликвидностью актива (бизнес-ценность, TCO, ROA)

ИНТЕГРАЦИЯ ДАННЫХ

- Ведение бизнес-глоссария
- Классификатор данных и связи данных
- Корпоративная ИТ-архитектура
- Архитектура данных
- Ведение мастер-данных
- Определение / корректировка политик, процессов, правил управления данными (Data Governance)
- Технологическая интеграция данных
- Определение точек и измеряемых правил контроля

ИСПОЛЬЗОВАНИЕ ДАННЫХ

- Управление качеством данных
- Управление мастер-данными
- Управление жизненным циклом данных
- Управление защитой данных
- Управление доступностью данных

Тенденции управления данными

- Каталогизация данных, понятный глоссарий данных
- Прозрачная методика формирования показателей
- Понимание взаимосвязи данных
- Понимание технического процесса формирования показателей

- Оперативная и качественная управленческая информация
- Точность и увеличение горизонта прогнозов
- Рост бизнеса
- Новые бизнес-модели

- Скорость реализации интеграции данных
- Визуальное проектирование интеграции
- Автоматическая генерация процессов
- Переиспользование логики интеграции

- Процессы управления качеством данных
- KPI качества данных
- Ответственность за качество данных
- Сервисы управления качеством

Цифровой двойник

РЕАЛЬНОСТЬ

Что это?

УПРАВЛЕНИЕ

ПЕРСОНАЛ

БИЗНЕС-ПРОЦЕССЫ

КЛИЕНТЫ / ПОСТАВЩИКИ

ПРОИЗВОДСТВО / ПРОДУКТЫ

РЕСУРСЫ

ГЛАВНЫЙ АКТИВ - РЕАЛЬНЫЕ АКТИВЫ

Ростелеком

ЦИФРОВОЙ ДВОЙНИК

Это цифровой двойник реального актива

УПРАВЛЕНИЕ

ПЕРСОНАЛ

БИЗНЕС-ПРОЦЕССЫ

КЛИЕНТЫ / ПОСТАВЩИКИ

ПРОИЗВОДСТВО / ПРОДУКТЫ

РЕСУРСЫ

ГЛАВНЫЙ АКТИВ - ДАННЫЕ

Цифровой двойник

Ростелеком

РЕАЛЬНЫЙ МИР

ДЛИНА	100 мм
ШИРИНА	90 мм
ВЫСОТА	90 мм
ВЕС	0,1 кг
СОРТ	Макинтош
...	XXX

ЦИФРОВОЙ ДВОЙНИК

ДЛИНА	0,1 м
ШИРИНА	...
ВЫСОТА	900 мм
ВЕС	+7 (916) ...
СОРТ	Макинтош
...	XXX

БОЛЬШИЕ ДАННЫЕ

ДЛИНА	0,1 м
ШИРИНА	0,1 м
ВЫСОТА	НЕВАЖНО
ВЕС	~ 0,1 кг
СОРТ	КРЯБЛОКО
...	XXX

DIS GROUP

Технологический уровень управления данными

Ростелеком

СБОР ДАННЫХ

Внешние и внутренние источники данных

RDB
Files
Social
News
Devices
Sensors
IoT

Интеграция и сбор данных из разных систем источников

ОБРАБОТКА И ХРАНЕНИЕ ДАННЫХ

Data Lake

DWH

Консолидация и хранение полученных данных

АНАЛИТИКА И ИСПОЛЬЗОВАНИЕ ДАННЫХ

Описательная (BI) и Предиктивная (AI/ML) аналитика

Исследование данных

Искусственный интеллект

WEB

APP

API

Дашборды, отчетность, визуализация и интеграция

DIS GROUP

Концептуальная архитектура решений

Ростелеком

АВТОМАТИЧЕСКОЕ ОБНАРУЖЕНИЕ 1 12

АВТОМАТИЧЕСКОЕ СВЯЗЫВАНИЕ

РАСПРЕДЕЛЕНИЕ И КОНТРОЛЬ ОТВЕТСТВЕННОСТИ

Единый каталог бизнес-терминов и определений, управление знаниями

5

Взаимосвязи данных, людей, бизнес-процессов, проектов

7 8

Мастер-данные

5

Хранилище метаданных и формирование взаимосвязей данных на техническом уровне

Аудит и оценка качества данных

7 9

Интеграция данных

5

Файлы

Приложения

10 11

BI, аналитика

2

Хранилище

Data Lake

6

Hadoop 4

MPP 3

СУБД 3

6

ИНСТРУМЕНТЫ ДЛЯ ИТ И ПОЛЬЗОВАТЕЛЕЙ 1 12

ПАНЕЛИ МОНИТОРИНГА ДАННЫХ, ПРОЦЕССОВ, ПОЛИТИК

ПРОЦЕССЫ КОНТРОЛЯ ИЗМЕНЕНИЙ

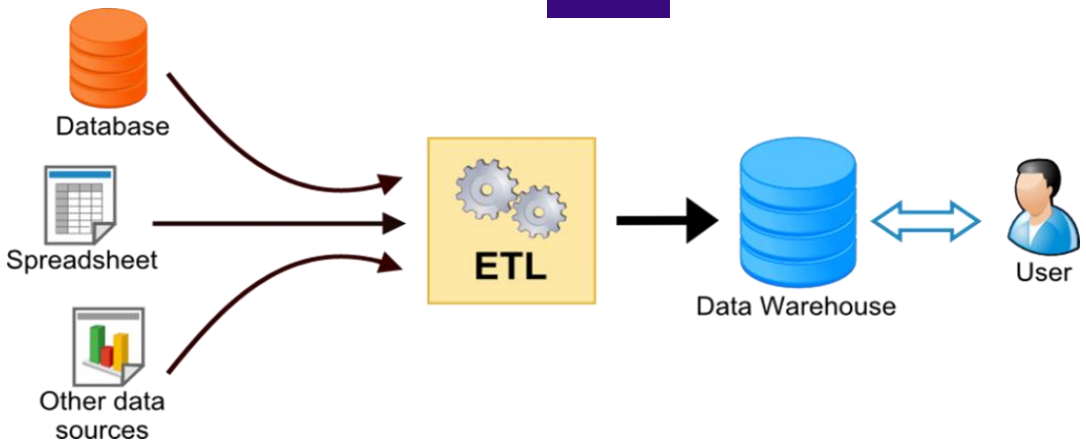
DIS GROUP

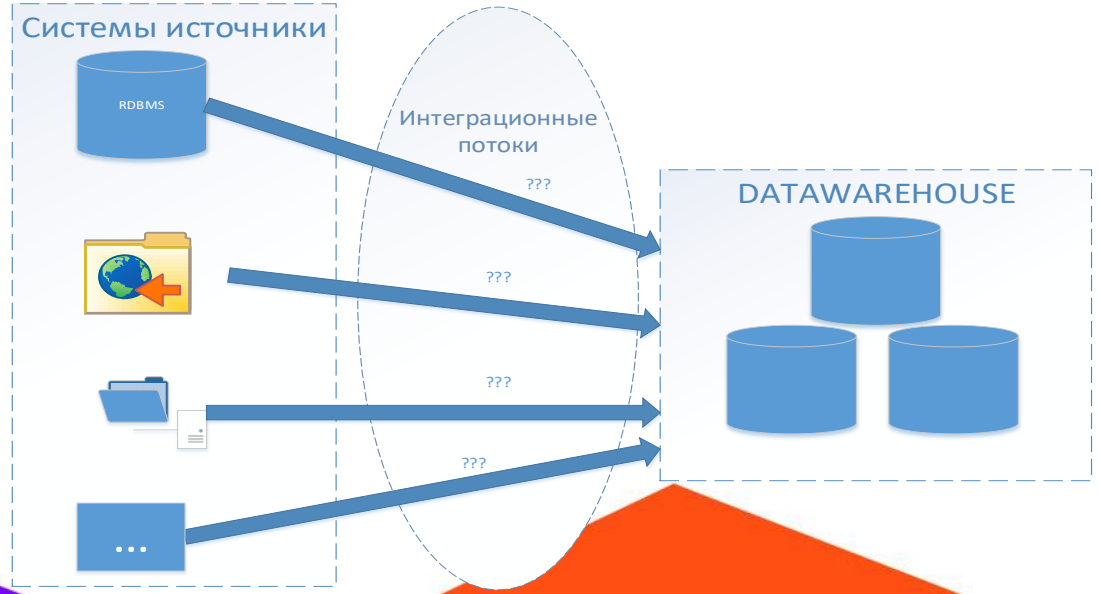
Хранилища и витрины данных

2

Ростелеком

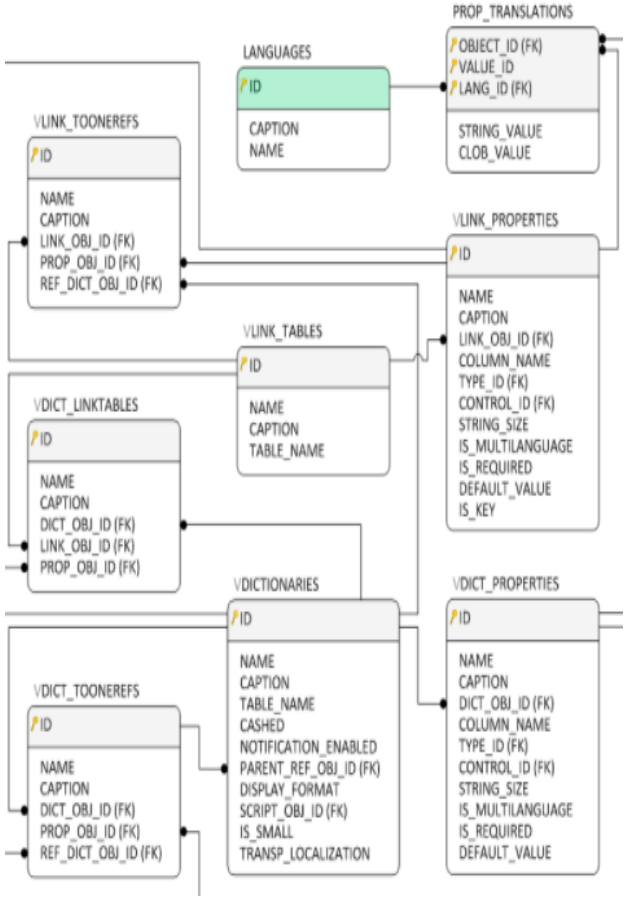
- Основной инструмент хранения данных для построения отчетности и аналитики
- Хранилище Данных не привносит новой информации и ограничено данными операционных систем-источников
- Представление данных в Хранилище Данных радикально отличается от представления в транзакционных системах
- Витрина данных – представление агрегированных данных или определенного разреза для блока отчетов или отдельного типа аналитики





DIS GROUP

Модель данных для работы с данными

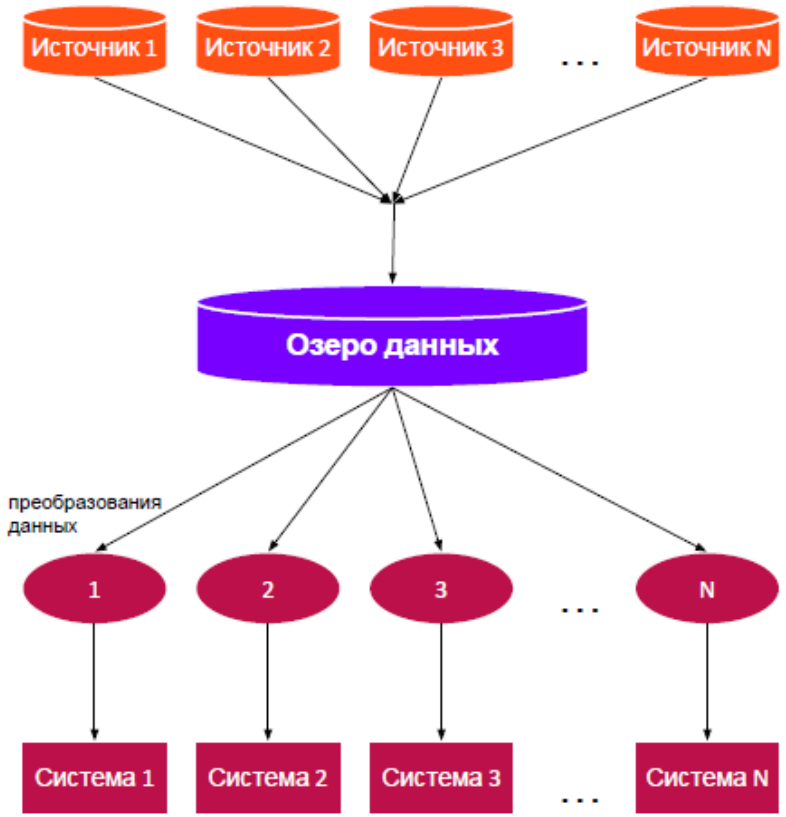


- **Модель данных** — это более высокое логическое определение объектов (абстрактное и самодостаточное), операторов и прочих элементов, в совокупности составляющих для пользователя абстрактную машину доступа к данным, с которой он взаимодействует
- Эти объекты позволяют моделировать структуру данных (реляционную, сетевую или иерархическую), а операторы — поведение данных
- **Корпоративная модель данных** – это многоуровневое, структурированное описание предметных областей организации, структур данных предметных областей, бизнес-процессов и бизнес-процедур, потоков данных, принятых в организации

Ростелеком

«Озера» данных (Data Lake)

- Data Lake – особая модель и архитектура построения хранилища данных
- Строится на основе систем, позволяющих хранить и обрабатывать как структурированные, так и полуструктурированные и неструктурированные данные
- Источники данных - в идеале **все** источники данных предприятия
- Данные извлекаются и сохраняются в Data Lake **как есть**, без преобразований и фильтрации
- Данные преобразовываются только на этапе их чтения их Data Lake непосредственно перед использованием потребителями
- Основные потребители - аналитики и системы анализа данных



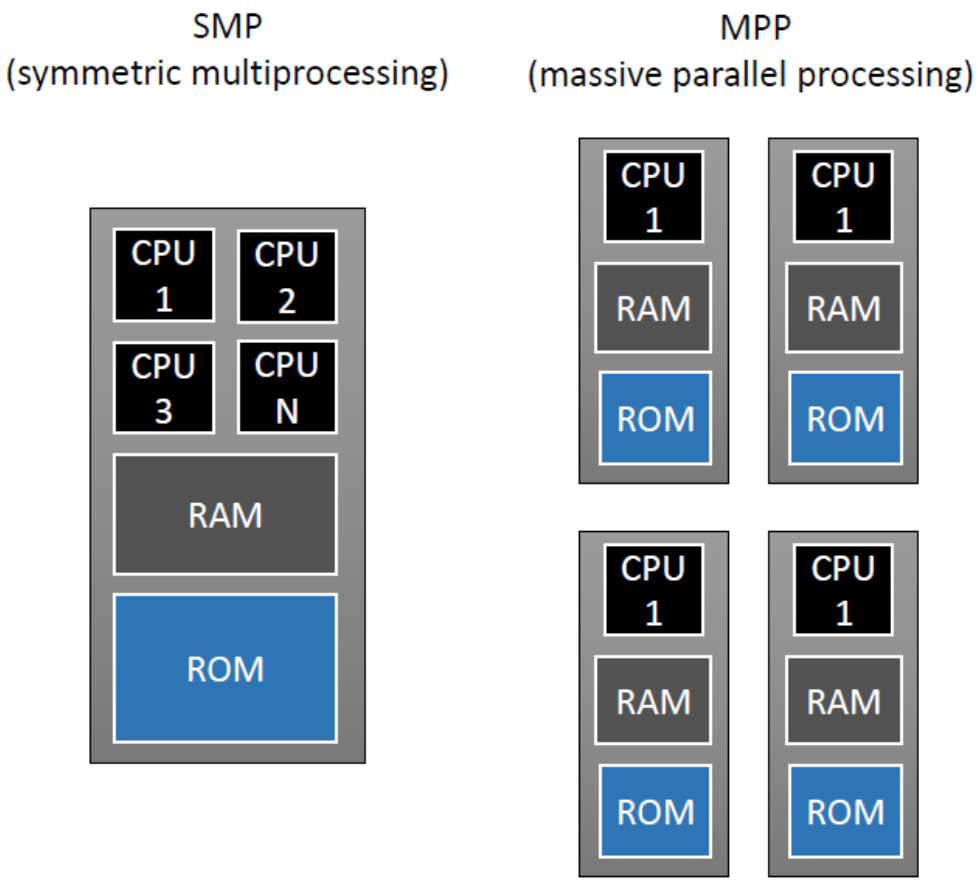
DIS GROUP

Базы данных и MPP

3

Ростелеком

- СУБД – самый распространенный инструмент для построения хранилища и витрин данных
- Чаще всего используются классические реляционные СУБД
- При больших объемах данных или логики обработки используются более мощные средства – Massive Parallel Processing

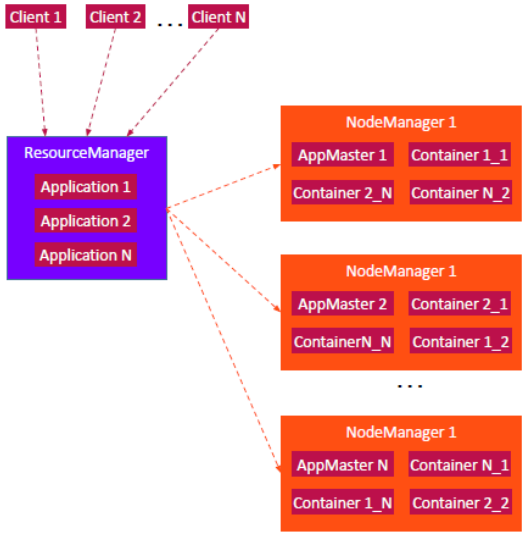


DIS GROUP

Big Data. Hadoop

Big Data – не сами данные, а концепция:

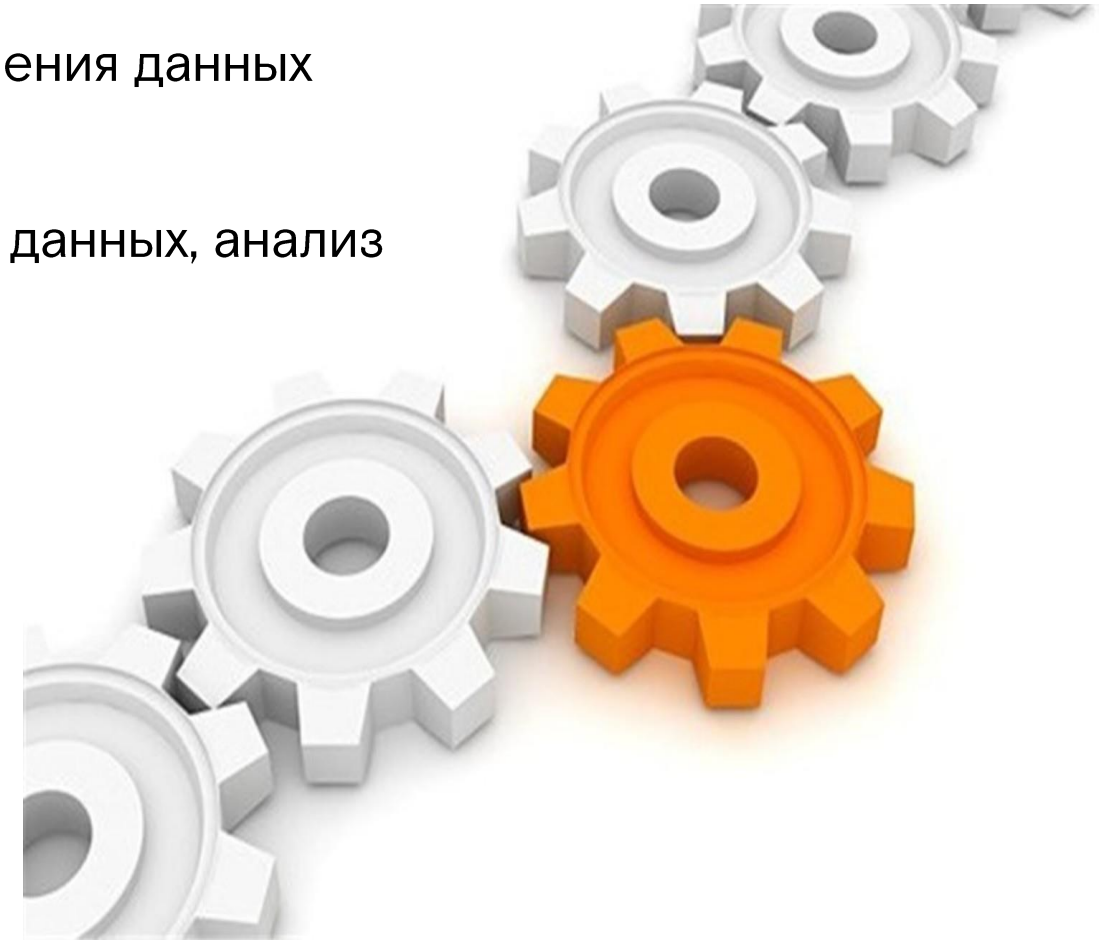
- Данные – огромных (для конкретной компании) объёмов и разнообразной структуры
- Технологии – горизонтально масштабируемые, ориентированные на параллельные вычисления
- Методы – как классические, так и современные узкоспециальные, в т.ч. алгоритмы машинного обучения и AI
- Источники – как классические, так и новые, в т.ч. открытые: социальные сети, Internet of Things, научные данные и т.д.



Apache Hadoop – open source технология, набор библиотек и утилит для распределенного, отказоустойчивого, горизонтально масштабируемого хранения больших объемов данных и их массивно-параллельной обработки

Интеграция данных

- Унифицированный доступ к источникам для извлечения данных и к приемникам – для загрузки
- Профилирование данных – исследование качества данных, анализ их зависимостей
- Трансформация данных
- Обогащение данных
- Объединение данных
- Аудит и мониторинг обрабатываемых данных
- Защита данных в процессе обработки



Ростелеком

Обеспечение качества данных

Исходные данные

Данные низкого качества

Анализ данных

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

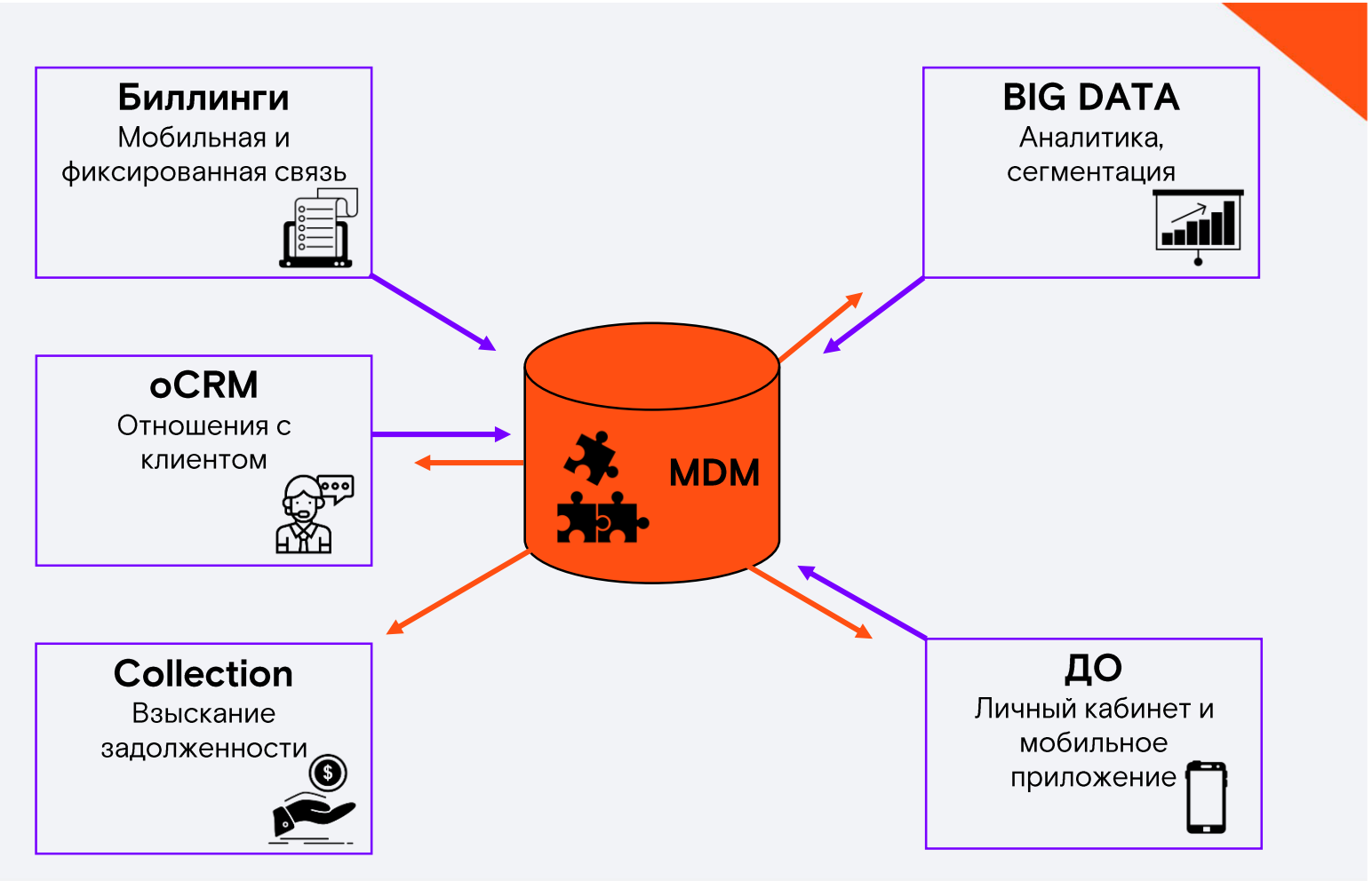
Выходные данные

DIS GROUP

Ростелеком

Построение мастер-данных

- Единый инструмент ведения и консолидации нормативно-справочной информации
- Выявление дубликатов записей
- Слияние записей и получение «золотой» (или мастер-) записи
- Связи справочных данных
- «Единая версия» справочных данных



DIS GROUP

Ростелеком

Business Intelligence и аналитика

- Основной инструмент бизнес-пользователя для работы с данными
- Целостность и интерактивность взаимодействия с данными
  - Все объекты визуализации, вне зависимости от их количества, должны обеспечивать целостность восприятия всех данных
  - Каждый объект визуализации должен обеспечивать «погружение» в данные, вне зависимости от пути исследования данных. Без нарушения целостности восприятия
- Интуитивность интерфейсов
  - Поддержка технологий drag-and-drop
  - Поддержка всех операций по работе с данными на любых мобильных устройствах, включая поддержку touch-интерфейсов (например, создание новых приложений и интерактивная работа в ранее созданных)
  - Автоматическое масштабирование на любом устройстве
  - Минимум технических особенностей в выполнении расчётов
  - Единообразие и максимальная степень свободы по взаимодействию с объектами визуализации

Чтение

Использование

Анализ

Аргументация

DIS GROUP

Два уровня управления метаданными

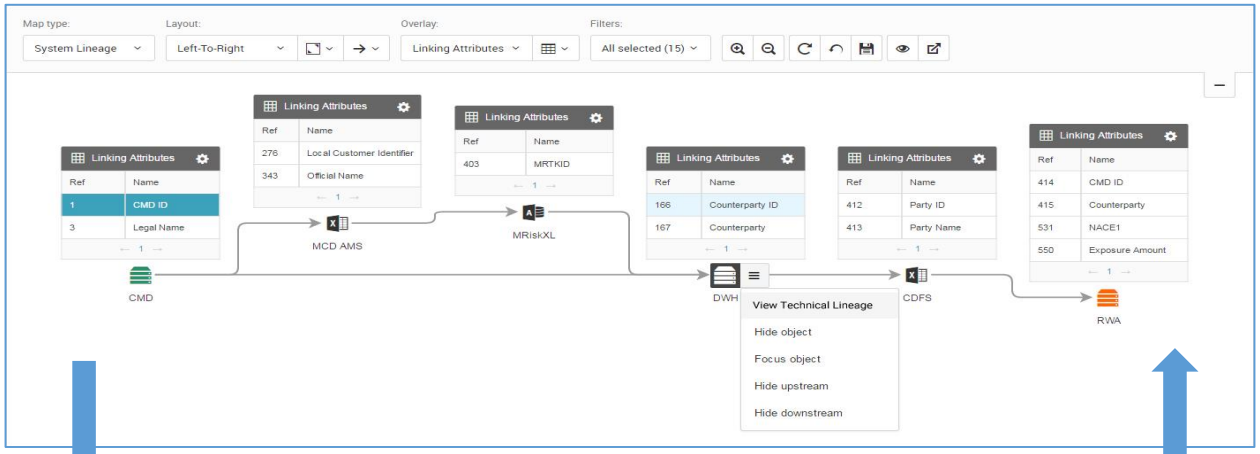
5

Ростелеком

Бизнес-метаданные:

- Концептуальная модель
- Согласованные изменения
- Связь активов данных и бизнес-окружения

Фокус на бизнес

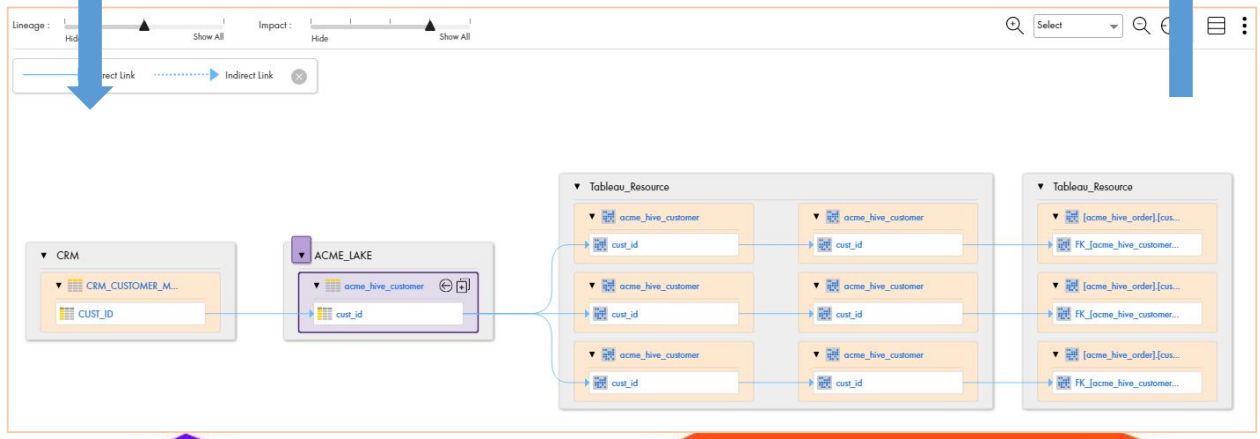


Автоматическая связь
и взаимное обогащение
технических и бизнес- метаданных

Технические метаданные:

- Физическая модель данных
- Построение Data Lineage на основе интеграционных процессов
- Профилирование и анализ данных

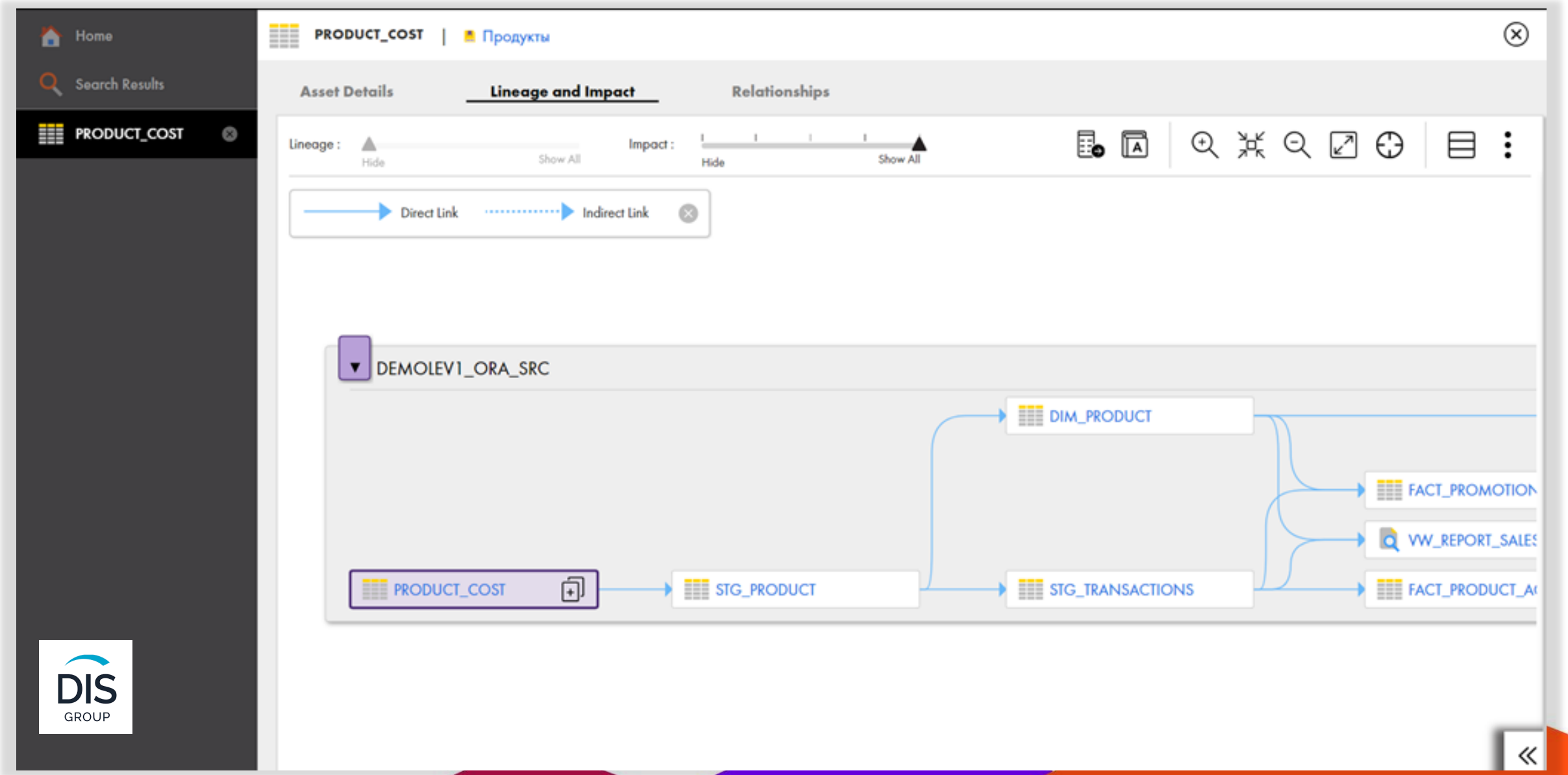
Фокус на технические аспекты



DIS GROUP

Data Lineage – происхождение и использование данных

Ростелеком



DIS GROUP

Бизнес-термины, их иерархия и история

Ростелеком

Географическая система
Glossary

Edit

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

HIERARCHY

БРД	Данные Блока Разведки и Добычи ПАО "Газпром нефть"
География	Бизнес-область "География"
Географическая сис	Область данных 'Географическая система'

CLASSIFICATIONS

BASIC CLASSIFICATIONS

DEFINITION

Область данных 'Географическая система'

Alias Names:	*Not specified*
Format:	Text - *Not specified*
LDM Reference:	*Not specified*
Business Logic:	*Not specified*

111: DM_CLIENT_1
Data Set in CLIENT_DM

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

HISTORY

Object	Field	To	Author	Date
John Admin (136)				
Data Set	Definition	Витрина данных по клиентам Детальная карта происхождения данных (Data Lineage в MM)	John Admin	21-12-2017
Data Set	Definition	Витрина данных по клиентам	John Admin	15-11-2017
Data Set	Name	DM_CLIENT_1	John Admin	03-11-2017
Data Set	RefNumber	111	John Admin	03-11-2017

HISTORY OPTIONS

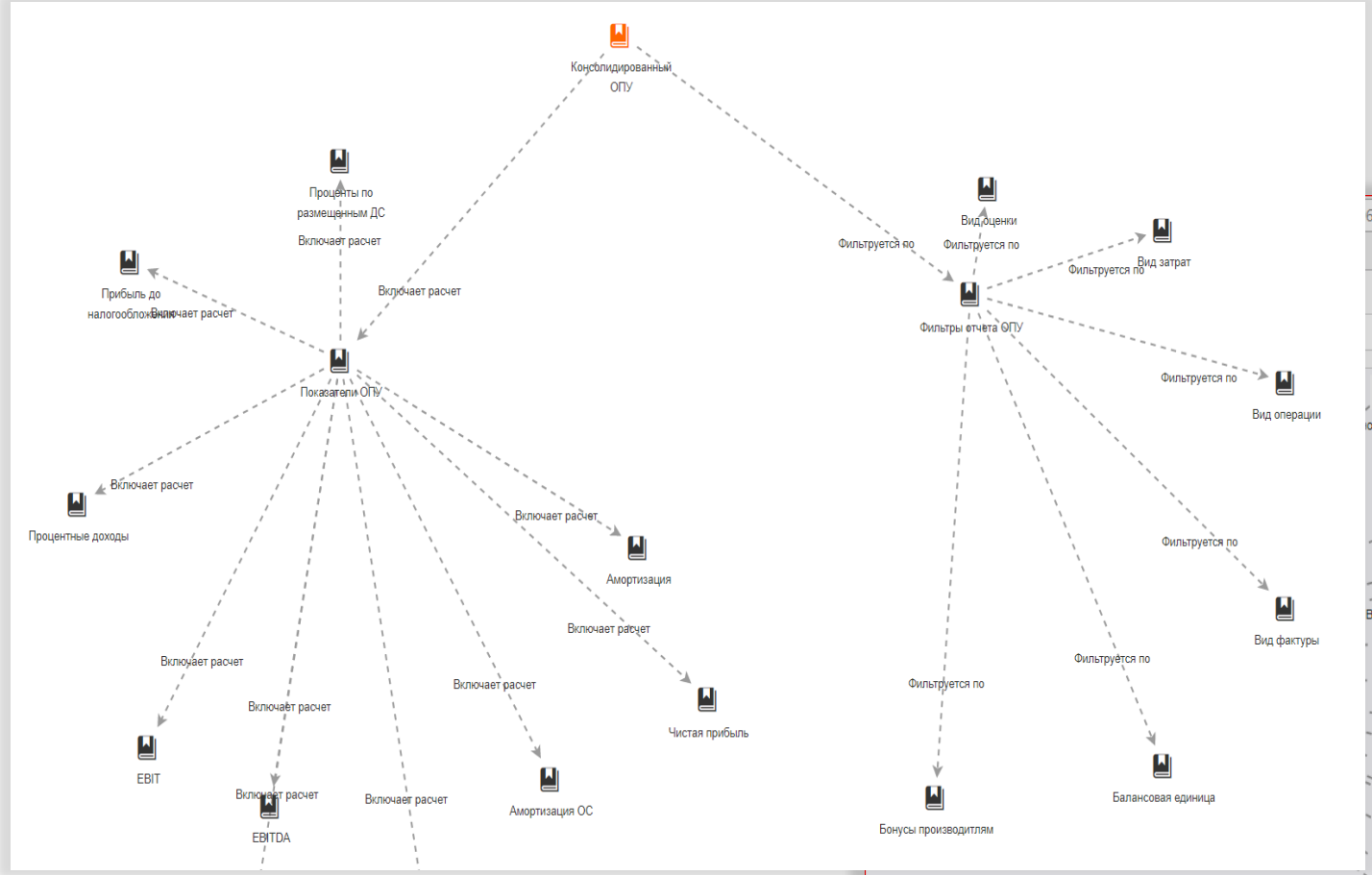
From Date *
01/12/2017

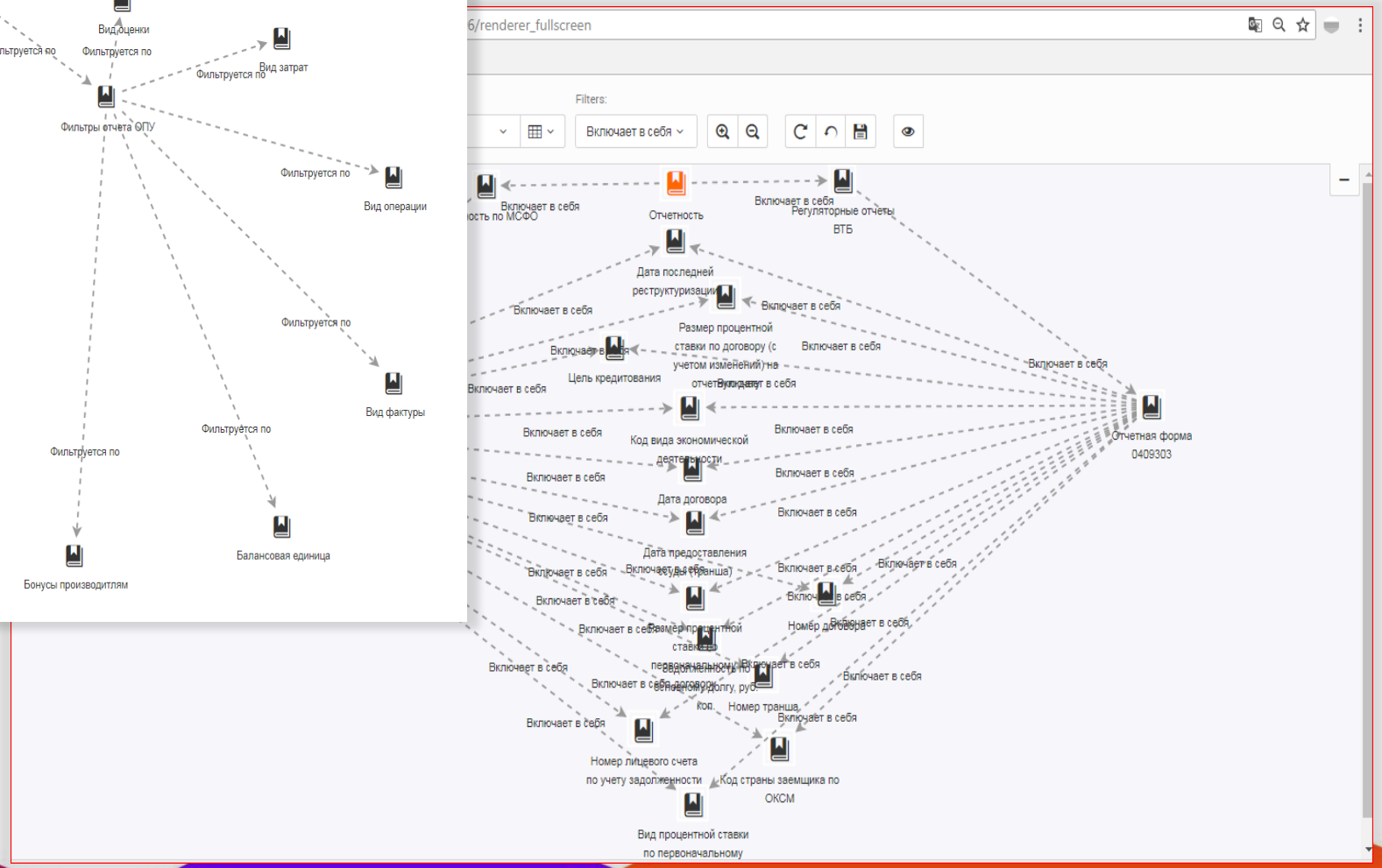
To Date *
24/01/2018

✓ Compare

DIS GROUP

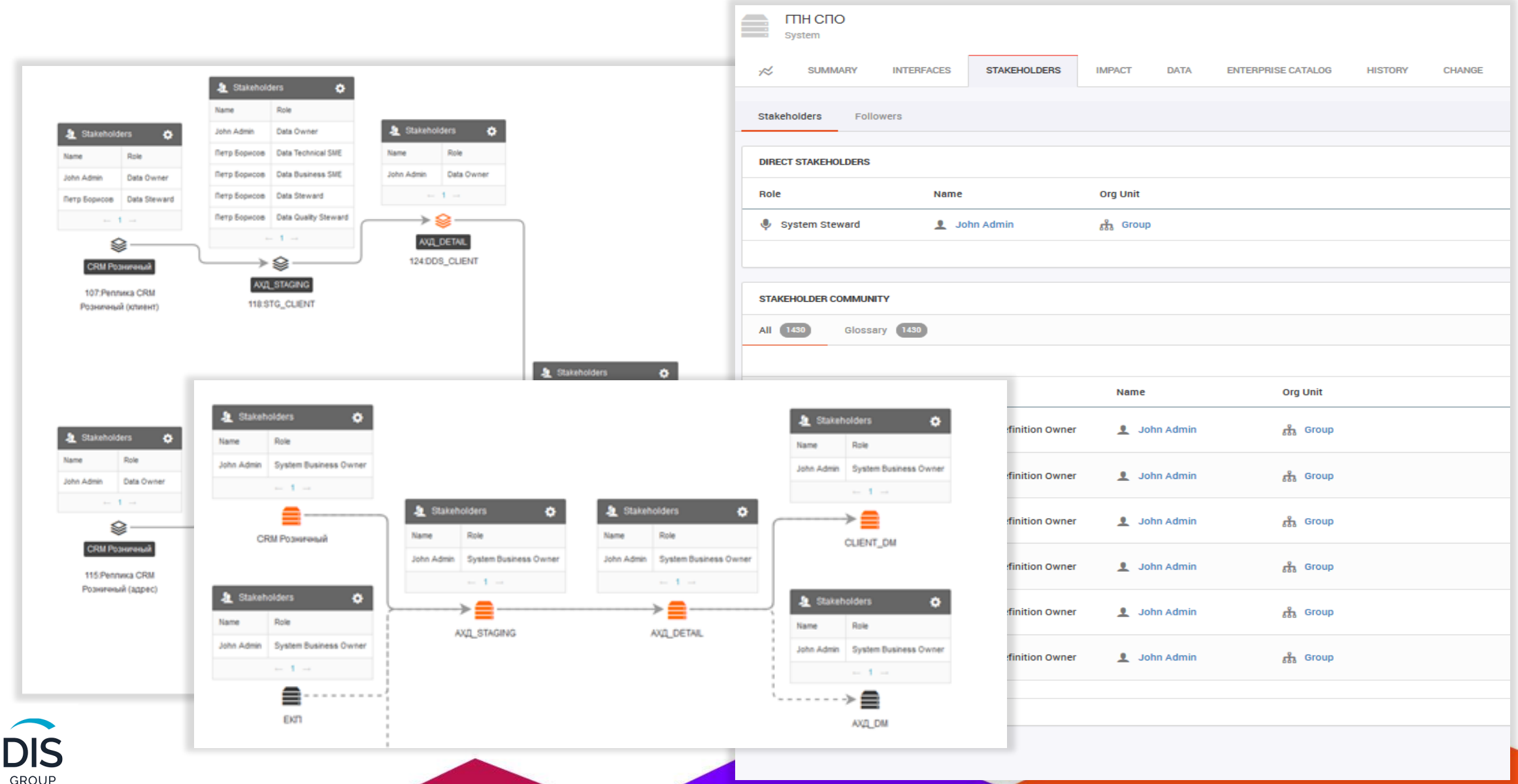
Визуализация связей бизнес-метаданных





Назначение и просмотр ответственных за данные

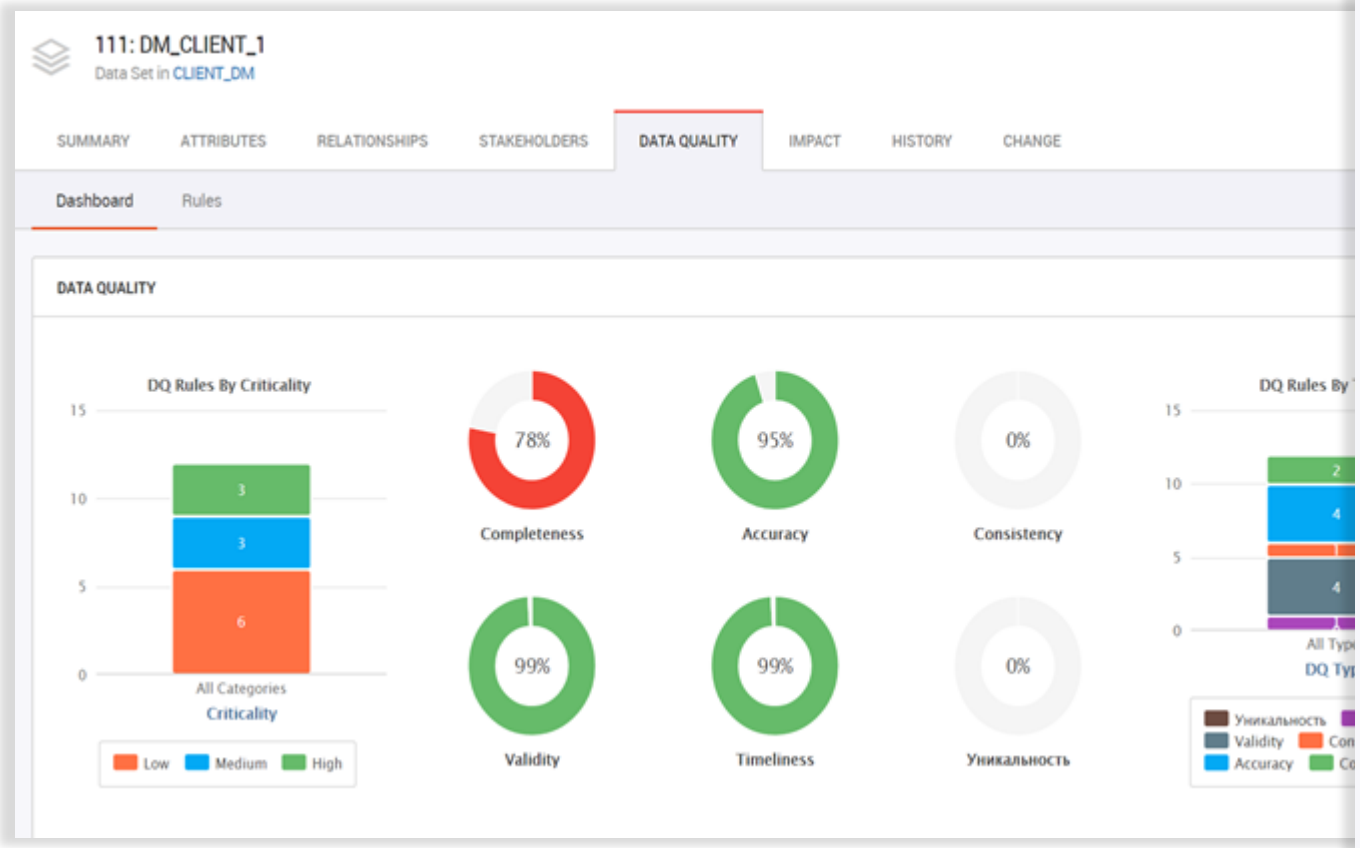
Ростелеком

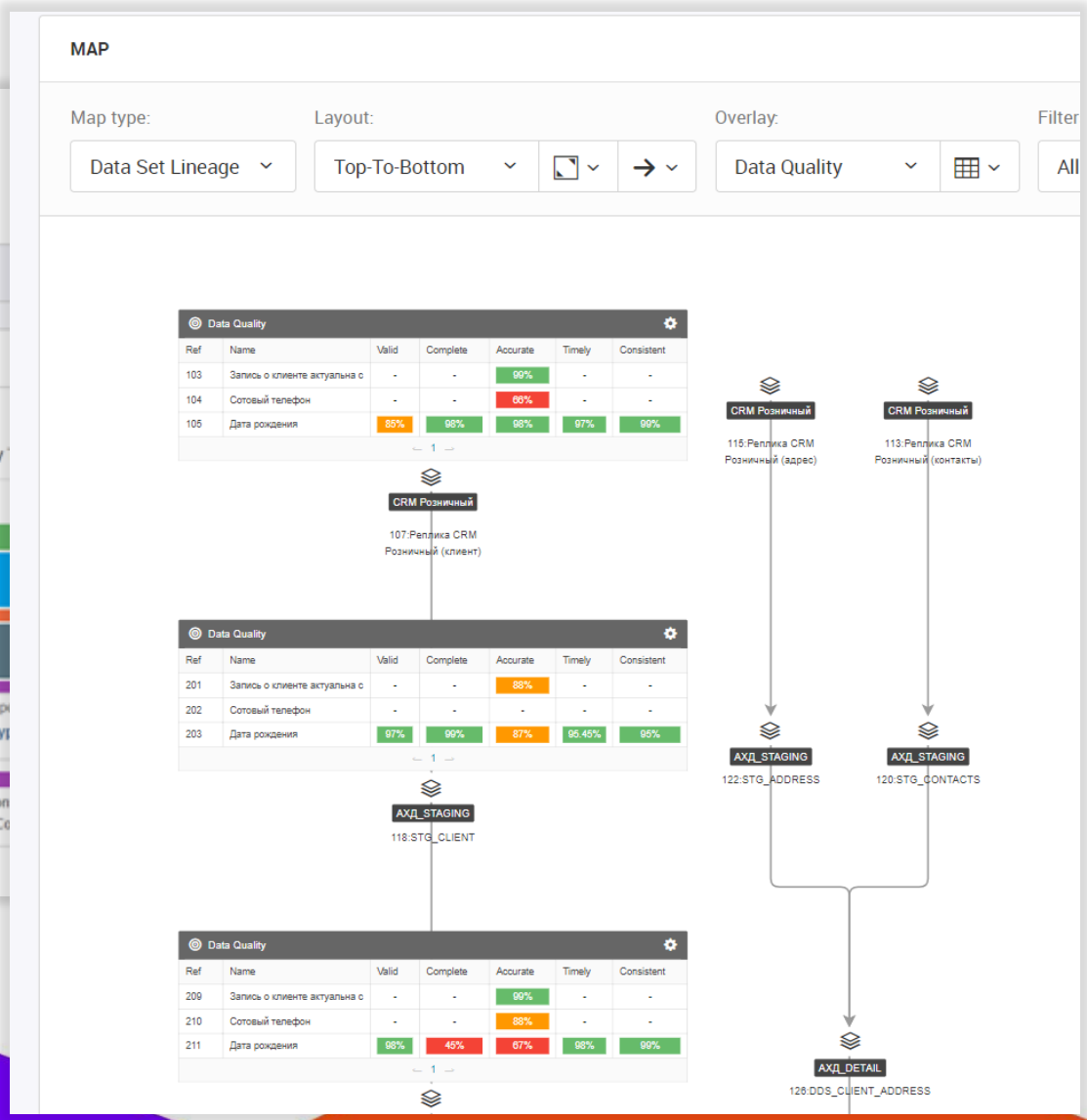


DIS GROUP

Визуализация качества отслеживаемых данных

Ростелеком





DIS GROUP

Результаты внедрения процессов управления данными

- Высокий уровень понимания данных компании
- Понимание влияния каждого домена данных на деятельность компании
- Повышение эффективности и скорости прохождения бизнес-процессов
- Работа данных на развитие компании через подход «ВЫЖМИ ИЗ ДАННЫХ ВСЕ!»

DataTalks

Q&A

Вопросы и ответы

- Подходы к управлению данными
- Стратегический подход
- Операционный подход
- Технологический подход

DIS GROUP

Ростелеком