

Модели данных

В1. Корпоративное управления данными



Московский государственный технический университет
имени Н.Э. Баумана

Факультет ИБМ

Окт 2025 года

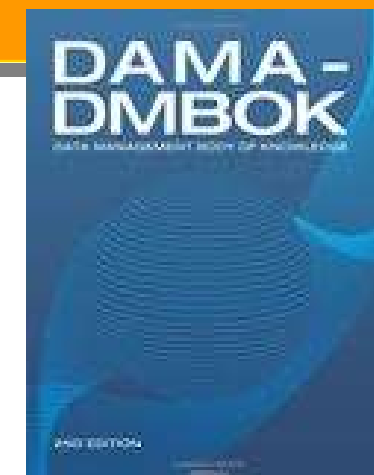
Москва

Артемьев Валерий Иванович © 2025

Что такое управление корпоративными данными (Data Management)?

Управление корпоративными данными – управление данными как активами в масштабе предприятия с точки зрения бизнеса

Ещё не включены Big Data и Data Science. В книге им посвящена отдельная глава



Свод знаний по управлению данными

Data Management Framework

Датацентричный подход

Сейчас: ориентация на приложения	В будущем: ориентация на данные (дата-центричность)
Непомерно высокие затраты на изменения в бизнес-приложениях	Разумная стоимость изменений
Данные привязаны к приложениям, потому что приложения владеют данными	Данные — открытый ресурс, который переживает любое приложение
Каждый новый проект сопровождается большими усилиями по преобразованию данных	Каждый новый проект использует существующие хранилища данных
Данные существуют в широком разнообразии разнородных форматов, структур, значений и терминологии	Данные глобально интегрированы, имеют общее значение и экспортируются из общего источника в любой необходимый формат
Интеграция данных потребляет 35–65% ИТ-бюджета	Интеграция данных всегда будет бесплатной
Трудно или невозможно интегрировать внешние данные с внутренними данными	Внутренние и внешние данные легко интегрируются

Принципы управления данными

Требования к управлению данными являются бизнес-требованиями

- Управление данными подразумевает управление качеством данных
- Для управления данными необходимы метаданные
- Для управления данными требуется планирование
- Требования к управлению данными должны оказывать определяющее влияние на решения в области информационных технологий

Данные имеют ценность

- Данные — актив с уникальными свойствами
- Ценность данных может и должна выражаться в экономических терминах

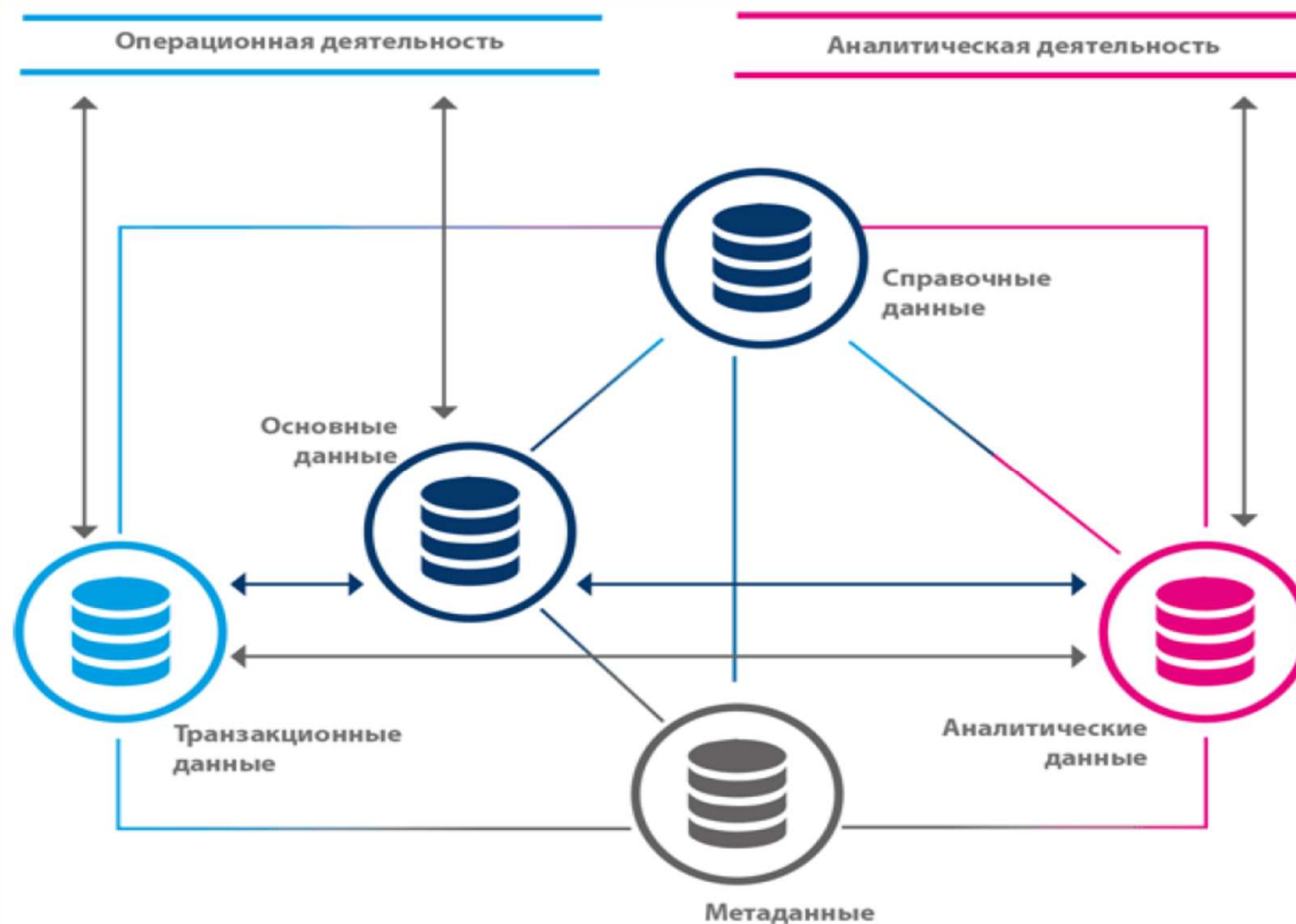
Управление данными требует разнообразных навыков

- Управление данными — кросс-функциональный процесс
- Управление данными требует целостного взгляда на функционирование организации
- Управление данными должно осуществляться с учетом разноплановых перспектив

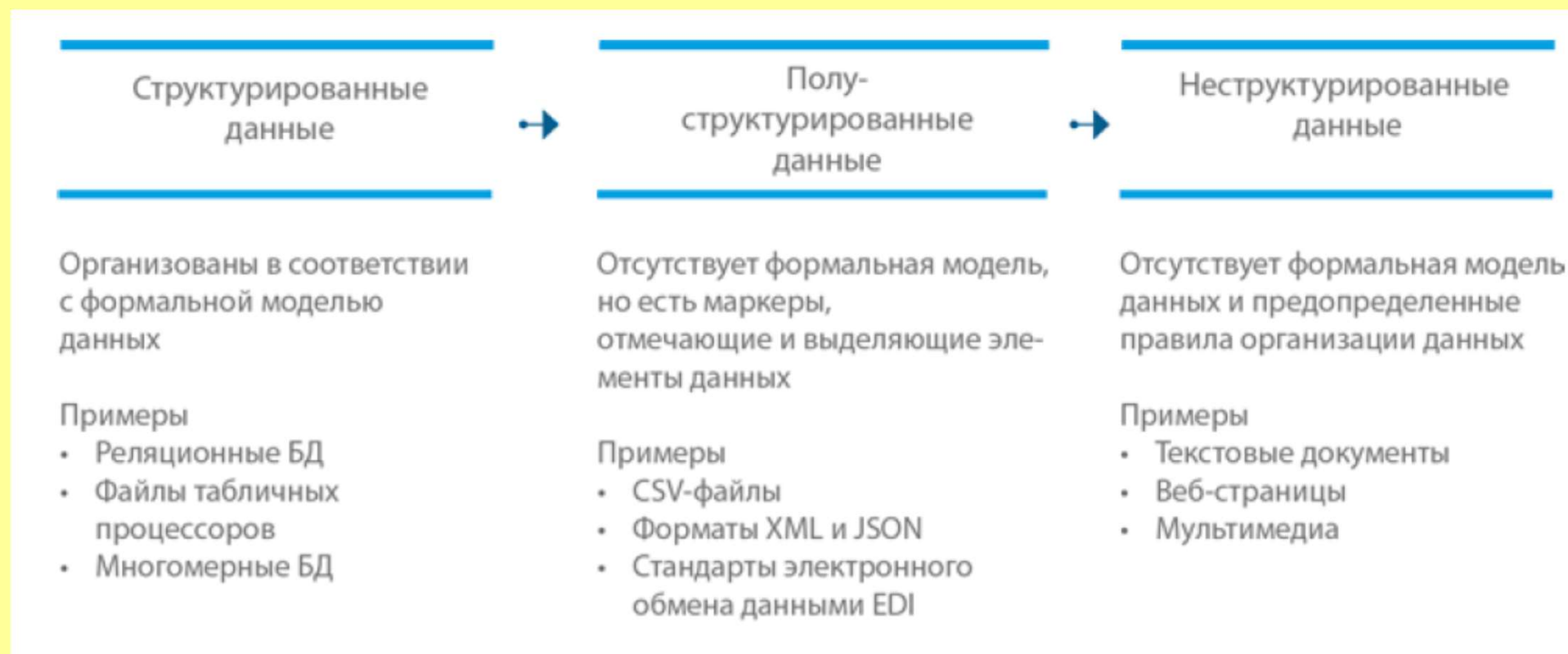
Управление данными — это управление их жизненным циклом

- Данные различного вида имеют различные характеристики жизненного цикла
- Управление данными включает управление рисками, связанными с данными

Взаимоотношение основных категорий данных



Форматы хранения и передачи данных с разной степенью структурированности



Соотношения между категориями данных



Данные верхних уровней лежат в основе формирования данных нижних уровней

Жизненный цикл данных



Управление метаданными (Metadata Management)

Детальность и связность метаданных



Архитектура и модели данных

Разработка и ведение

- потоков данных
- корпоративной модели EDM (HLDM, CDM, LDM предметных областей)
- канонической модели (XML-схемы, XBRL-таксономии)



Ведение частных метаданных

Разработка и ведение

- прикладных моделей данных и форматов
- mapping моделей
- правил контроля данных
- семантических слоёв пользователя



Каталог данных и бизнес-гlossарий

- Выявление и учёт информационных активов
- Создание и ведение бизнес-гlossария
- Публикация каталога и бизнес-гlossария
- Навигация и поиск



Управление репозиторием

- Импорт и экспорт метаданных
- Связывание метаданных
- Анализ зависимости
- Анализ происхождения данных
- Навигация и поиск



Разновидности информационных активов

«Сырые» данные

Собираемая отчётность и микроданные, события, сообщения, измерения с датчиков

Информация

Фактографические данные, реестры, классификаторы/справочники, результаты обработки, документы, контент, архивы

Метаданные

Описания информационных активов

Знания

Семантика предметных областей, описательные и прогнозные модели, имитация, рекомендации, отношения/связи, извлечение фактов и выявление тональности текста

Артефакты искусственного интеллекта

Предписывающие модели, распознавание/ генерация письменной и устной речи, распознавание образов, поиск и запросы на естественном языке



**Критичность,
важность,
ценность,
качество актива,
глубина истории,
частота наблюдений**

Метаданные – основа управления информационными активами

Каталог данных

Назначение данных/ информации, уровень доступа, источники, владельцы/ кураторы, поставщики, потребители, ресурс/ приложение, ссылка на прикладную модель данных, качество данных

Корпоративная модель данных

Верхнеуровневая модель данных, концептуальные и логические модели предметных областей, логические и физические прикладные модели, ссылки на бизнес-гlossарий. Каноническая модель данных

Бизнес-гlossарий данных

Предметные области данных, терминология, связи с нормативной базой и тезаурусом, сущности и атрибуты, области допустимых значений, бизнес-правила

Каталог аналитических моделей

Назначение модели, структура модели, метод, алгоритм, приложение/ инструмент, обучающий и тестовый наборы, параметры, вход/ выход, качество моделей

Каталог данных – учёт информационных ресурсов



Моделирование и проектирование данных (Data Modeling & Design)

- **Модель данных** – описание структуры и содержания данных для представления реального объекта, процесса или концепции
- **Понятность и повторная используемость данных**
- **Операции**
 - ведение моделей
 - версионирование моделей
 - генерация моделей
 - связывание моделей
 - навигация и поиск моделей
 - импорт/ экспорт моделей

Владельцы/ кураторы данных

Бизнес-аналитики

Модельеры, архитекторы

Аналитики и исследователи
данных

Спецы по качеству данных

Инженеры по данным

Проектировщики, программисты

Администраторы БД

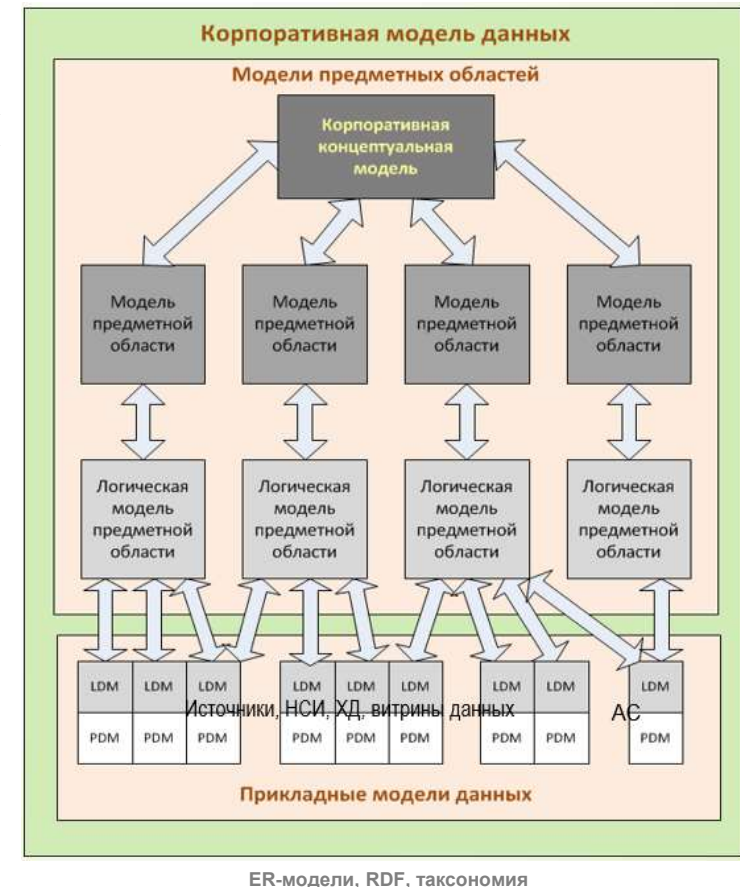
Концептуальный
обзор всех
предметных
областей
корпорации

Концептуальное
представление
сущностей и связей
для каждой области

Логическое
представление
для каждой
предметной области

Каноническая
модель данных:
форматы сообщений
и web-сервисов
(XML / JSON)

ЛМД и ФМД,
специфичные для
реализации
приложений или
проектов

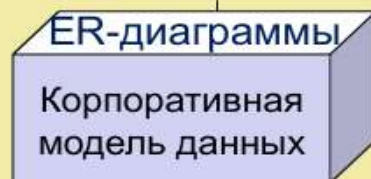


Бизнес-гlossарий данных – учёт семантики данных

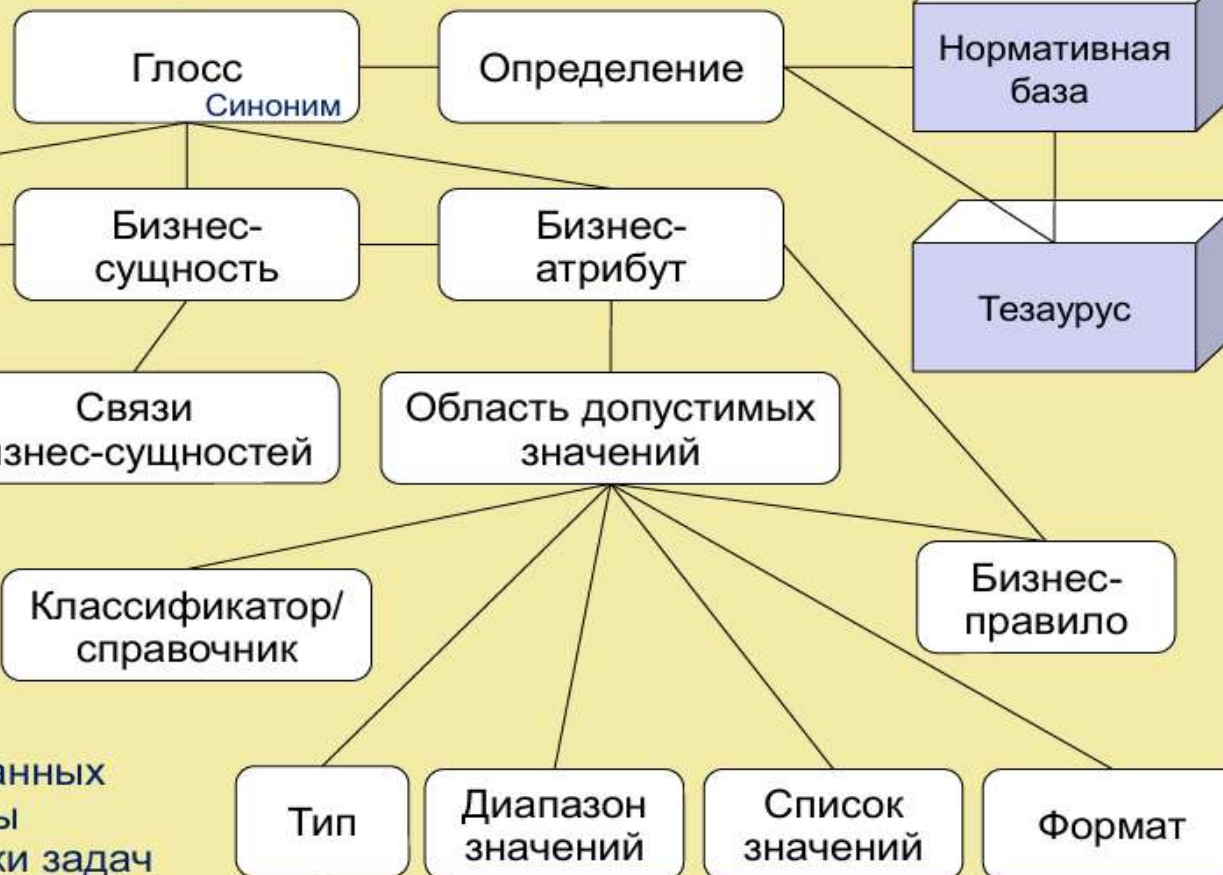
Понятность данных в бизнес-терминах

Описание логических
моделей данных
предметных областей

Навигация
и нечёткий
поиск,
связывание

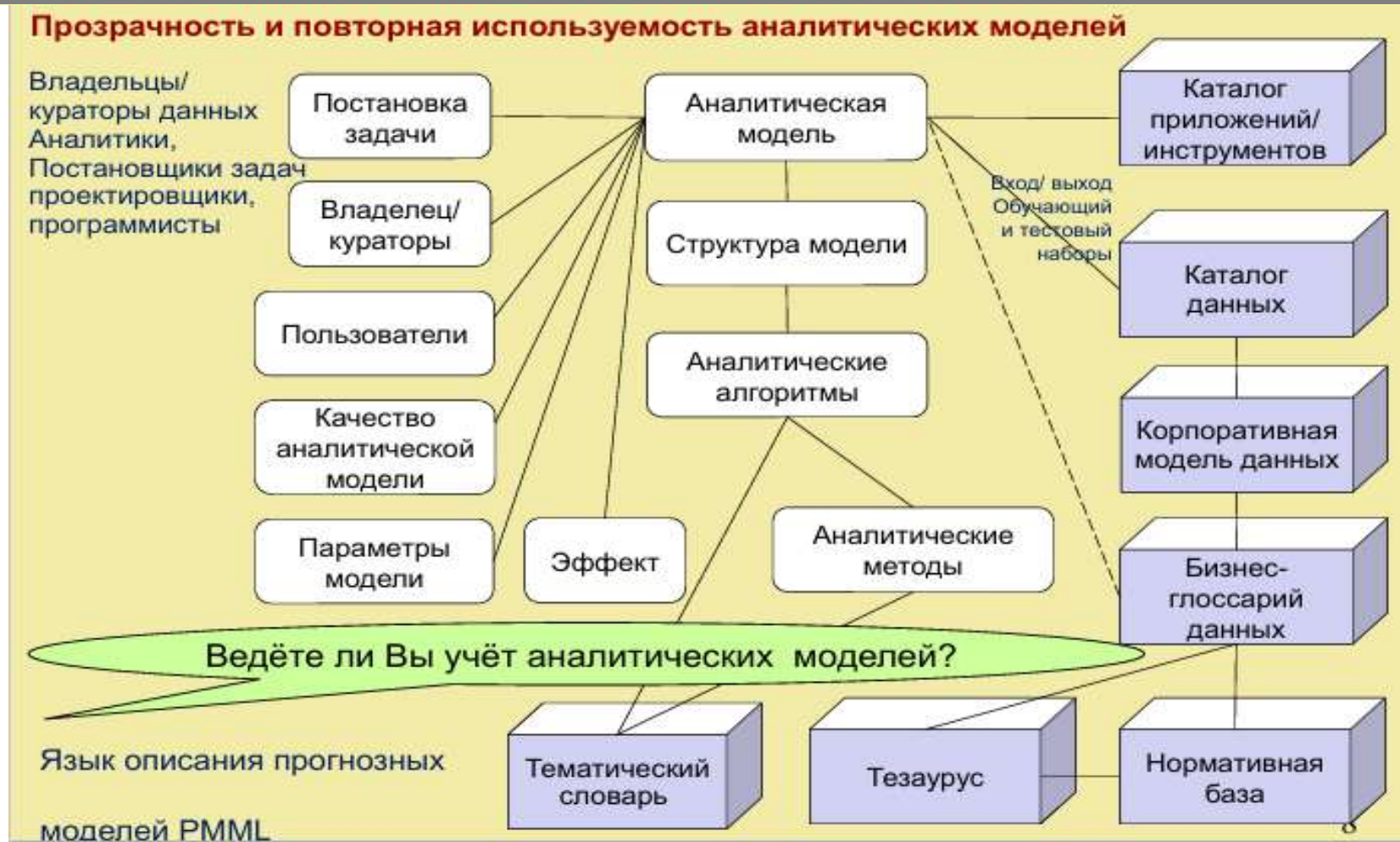


Руководители
Владельцы/ кураторы данных
Модельеры, архитекторы
Аналитики, постановщики задач
Специалисты по качеству данных
Проектировщики, программисты



Wiki / RDF / таксономия / онтология
Может быть частью Корпоративной модели

Каталог аналитических моделей – учёт знаний



Управление справочниками и реестрами (Reference & Master Data Management)

Ведение, интеграция и обеспечение качества справочников и реестров – критических и важных информационных ресурсов

■ **Реестры (мастер-данные)** – данные о ключевых бизнес-сущностях

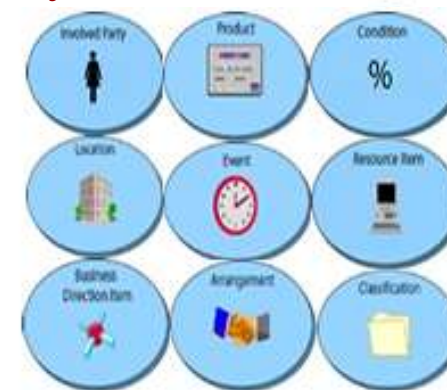
- Влияют на операционные и репутационные риски
- Важно отсутствие дублирования экземпляров данных
- Авторитетное ведение реестров

■ **Классификаторы** – аналитическая ценность данных

- Влияют на управленческие и репутационные риски
- Важна устойчивость и полнота схем классификации
- Обобщение и клонирование классификаторов
- Авторитетное ведение классификаторов

■ **Обеспечение качества реестров и справочников**

- Унификация идентификации субъектов
- Формирование «золотых» записей (история изменений, кластеризация, оценка достоверности, приоритеты заполнения)
- Стандартизация текстовых атрибутов (парсинг, правила, шаблоны, словари, matching, определения вида субъекта)
- Ведение истории изменений



Интеграция данных и интероперабельность (Data Integration & Interoperability)

Степень интеграции и качество данных



Сбор данных,
обработка событий
и потоков

- XML, JSON, AVRO
- XBRL, SDMX
- eForm
- EAI, ESB
- Pub / sub
- Web-сервисы
- REST API
- Microservice
- Streams
- CEP/EDA



Консолидация
и согласование
данных

- ODS
- Data Warehouse
- Data Marts
- Data Lake
- Date Vault / 6NF
- LDW
- Data Hub
- ETL / ELT
- CDC
- Autoscripts



Управление
справочниками
и реестрами

- Ведение реестров и справочников
- Интеграция мастер-данных
- Устранение дублей
- Стандартизация
- Проверки качества данных



Управление
качеством данных

- Профилирование
- Проверки полноты, допустимости и целостности
- Тест бизнес-правил
- Очистка данных
- Мониторинг, извещения и раскрытие

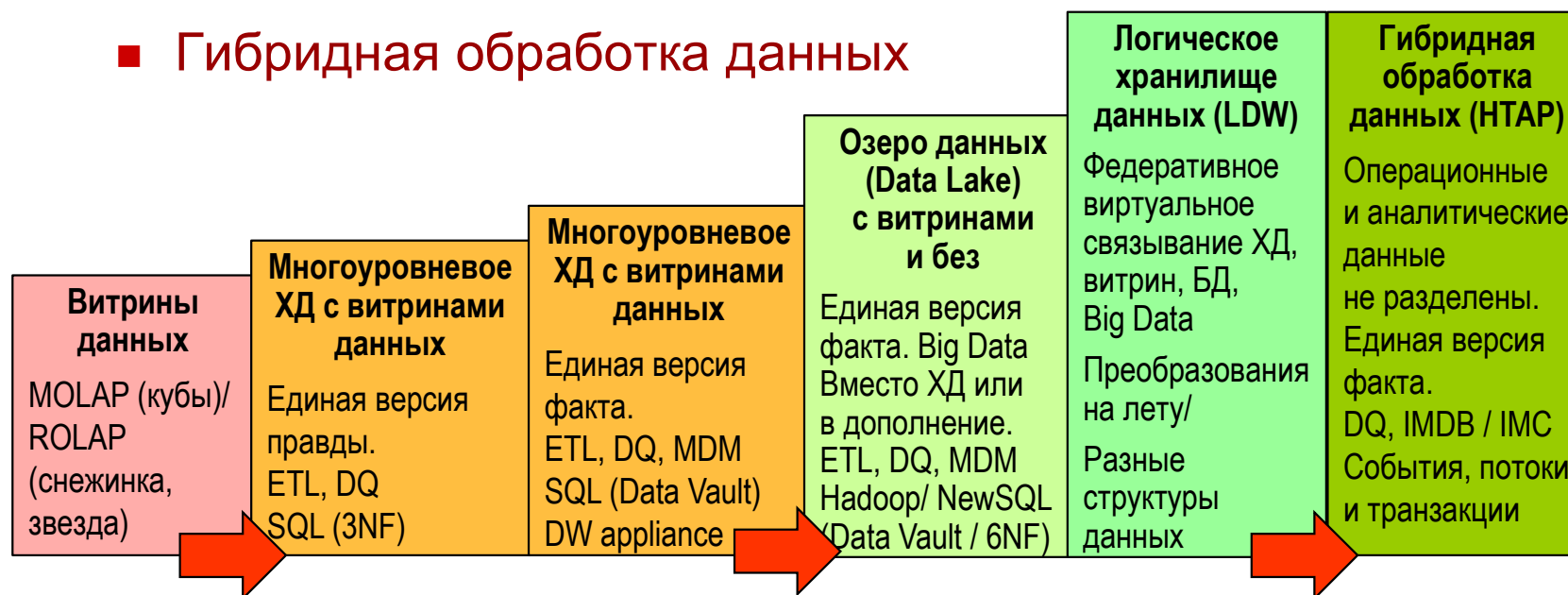


Развитие интеграции данных

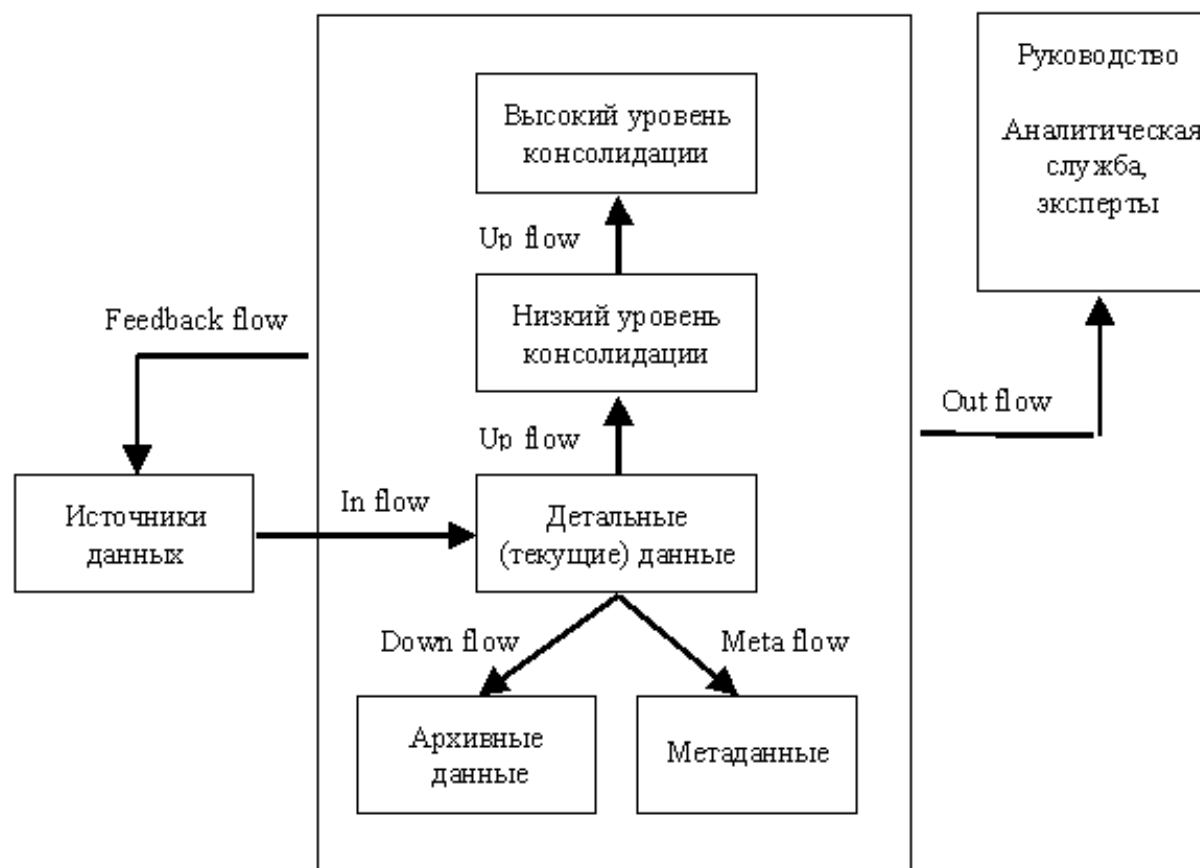


Стремление меньше перемещать данные привело к двум перспективным способам интеграции данных:

- Логическое хранилище данных
- Гибридная обработка данных



Потоки данных в аналитических БД



Управление качеством данных (Data Quality Management)

Управление качеством данных –

обеспечение соответствия состояния данных согласованным требованиям пользователей данных

Непрерывный и распределённый процесс на всём ЖЦ данных

- подготовка данных
- сбор данных
- преобразование и загрузка данных в ХД
- преобразование и загрузка данных в витрину/ песочницу
- обработка данных

Развитие не новой для нас темы

Не просто проверка данных при их сборе и загрузке, но также:

- *профилирование* – начальное ознакомление с источниками данных
- *измерение качества данных* путём контроля данных и расчёта метрик
- *наблюдение и реагирование* на превышение установленных уровней ошибок
- *очистка*: корректировка и преобразование данных
- *раскрытие сведений*: публикация отчётов, визуализация на информационных панелях, доступ к данным с отметками об ошибках



Принципы, важные для качества данных

Не воспроизводить мусор GIGO (англ.: Garbage In – Garbage Out)

«Грязные» данные влияют

- на исполнение операций (операционные риски)
- на качество аналитических моделей и принимаемых решений (управленческие риски)
- на качество публикаций в Интернет/печати (репутационные риски)

Однозначность элементов данных

Для представления (хранения) значений каждой величины должен отводиться отдельный элемент данных.

Исключение: строчные композиты (например: ФИО, адреса и т.п.)

Историчность данных

- явная привязка фактов ко времени свершения (наблюдения)
- указание периода актуальности для записей справочников/ реестров
- данные имеют определённую историческую глубину

«Неопределённость Гейзенберга» при оценке качества данных

- либо получаем достоверные, но не совсем актуальные данные
- либо получаем актуальные данные, но не совсем достоверные

Тихоокеанский
«мусороворот»



Мусоро-сжигательный
завод Вены

Процедуры управления качеством данных



Характеристики и показатели качества данных

ПОЛНОТА – наличие содержимого и структурных связей

- **Полнота содержания** – наличие непустых значений для элементов данных (композиций), всех необходимых (или достаточных) записей или наборов данных
- **Целостность структуры** – наличие связей наборов данных

ДОСТОВЕРНОСТЬ – соответствие содержания и структуры данных реали

- **Допустимость** – соответствие отдельных элементов данных (или групп элементов данных в записи) *области допустимых значений*
- **Согласованность** – соответствие данных бизнес-правилам на уровне элементов данных, записей (композиций), наборов данных, информационных ресурсов

СВОЕВРЕМЕННОСТЬ – соответствие данных временным ограничениям


- **Срочность** – необходимые частота и сроки получение данных
- **Актуальность** – сохранение ценности данных до момента использования

РЕЛЕВАНТНОСТЬ – отражение определённых понятий в данных

- уникальность экземпляров данных моделируемого субъекта/объекта
- уникальность бизнес-сущностей
- учёт и отражение имен сущностей, связей и атрибутов в моделях
- соответствие правилам именования сущностей, связей и атрибутов
- согласованность имен сущностей, связей и атрибутов, нет дублирования

ПОНЯТНОСТЬ – наличие описания данных в терминах предметной об

- пояснения к данным в терминах предметной области
- области допустимых значений данных
- описание связей и бизнес-правил



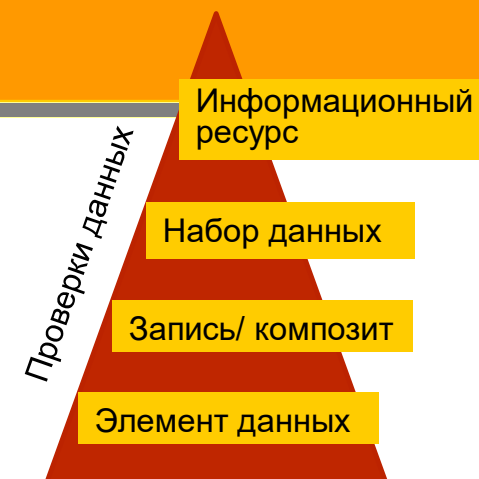
- тип данных
- разрядность и точность числа
- длина строки/ кода
- формат (шаблон)
- диапазон/ перечень допустимых значений
- контрольный разряд
- единица измерения

Проверки и метрики качества данных

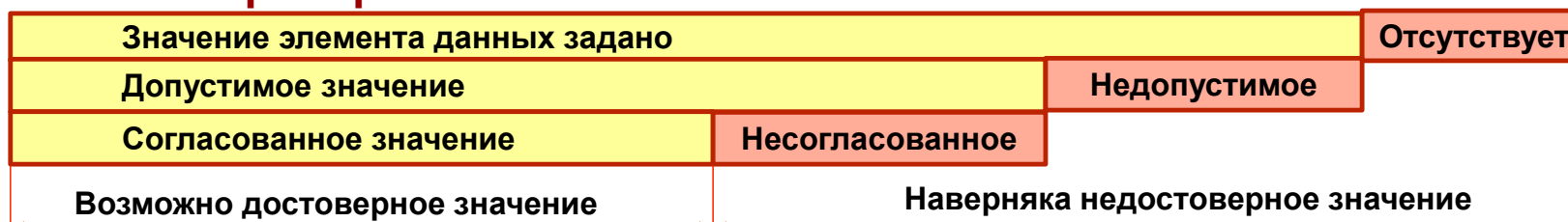
Проверки данных

Для каждого критичного/ важного уровня данных разрабатывают подходящие проверки по полноте, целостности структуры, допустимости и согласованности.

Важно чётко формулировать суть проверки и относить к определённому показателю качества и уровню данных



Зависимость проверок

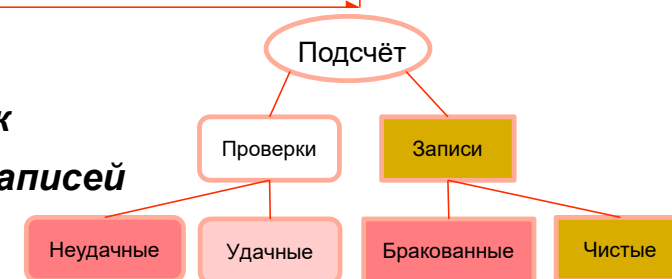


Метрики показателей качества

Доля неудачных проверок = Число ошибок / Общее кол-во проверок

Доля бракованных записей = Число бракованных записей / Число записей

Доля бракованных записей \neq Доля неудачных проверок



Пример. Оценка качества Витрины данных торгового репозитория и депозитария НРД

Динамика доли ошибок витрины данных по показателям качества

Год	Методологическая определенность	Согласованность	Достоверность	Полнота	Целостность	Общая доля ошибок
2015	293,2	19,4	18,3	1,5	0,9	15,4
2016	14,7	35,7	3,5	6,8	5,3	9,5
2017	0,4	12,5	4,3	4,6	4,5	5,1

Нечёткие границы определения проверок показателей качества

Сравнение двух интегральных метрик наборов данных и витрины

Тип ПФИ	Число ошибок	Число проверок	Число записей	Число бракованных записей		Доля бракованных записей		Доля ошибок
				НГ	ВГ	НГ	ВГ	
Форвард на фондовые активы	7 397	43 266	7 211	7 211	7 211	100,00%	100,00%	17,10%
РЕПО	416 271	34 799 466	2 676 882	174 930	416 271	6,53%	15,55%	1,20%
Товарный форвард	1 246	137 520	22 920	929	1 246	4,05%	5,44%	0,91%
Валютный спот или форвард	90 602	18 739 568	2 342 446	86 351	90 602	3,69%	3,87%	0,48%
Валютный (конверсионный) своп	10 857	9 499 980	791 665	5 486	10 857	0,69%	1,37%	0,11%
ИТОГО	526 373	63 219 800	5 841 124	214 829	526 187	4,71%	9,01%	0,83%

Метрика доли ошибок по проверкам завышает качество данных

НГ (Доля бракованных записей) = $\text{МАКС} \{ \text{Кол-во ошибок по } i\text{-ой проверке} \} / \text{Кол-во записей}$

ВГ (Доля бракованных записей) = $\text{МИН} (\text{Кол-во ошибок}; \text{Кол-во записей}) / \text{Кол-во записей}$

Оценка качества данных ЕГРЮЛ

Интегральные метрики

Характеристика	Доля ошибок	НГ брака	ВГ брака
Согласованность	4,83%	28,55%	28,96%
Полнота	3,24%	19,38%	45,35%
Целостность	1,32%	1,32%	1,32%
Допустимость	0,04%	0,12%	0,12%
Общий итог	3,16%	28,55%	

Доля ошибок =
Число ошибок / Число проверок

Доля брака =
Число записей с ошибками /
Общее число записей

Проверки и детальная метрика

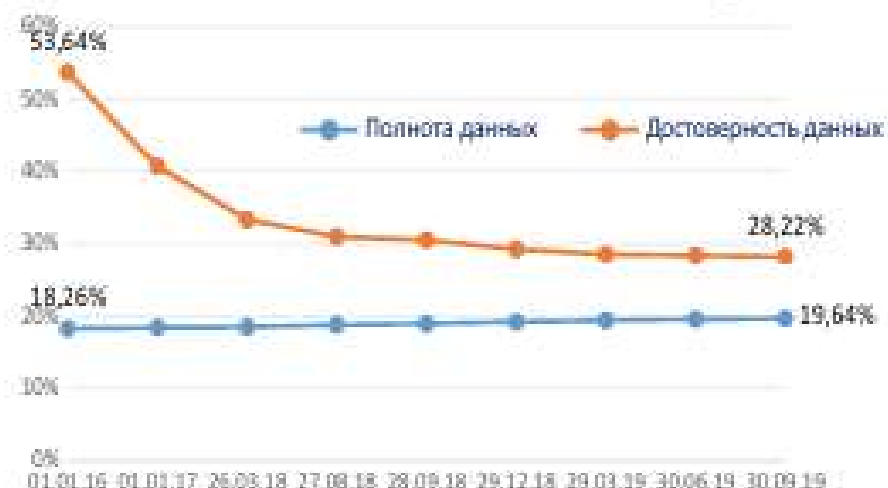
Характеристика и проверки	Число ошибок	Число проверок	Доля ошибок
Согласованность	1 056 826	21 898 128	4,83%
Сумма долей в УК в точности равна 100%	1 042 056	3 649 688	28,55%
Дублирующиеся лицензии отсутствуют	10 343	3 649 688	0,28%
Основной код ОКВЭД ровно один	4 293	3 649 688	0,12%
Совпадает КПП в записи об учете в налоговом органе	102	3 649 688	0,0028%
Совпадает ИНН в записи об учете в налоговом органе	16	3 649 688	0,0005%
Ровно одна запись по сочетанию ИНН и ОГРН	16	3 649 688	0,0005%

Динамика метрик качества данных ЕГРЮЛ

Динамика доли ошибок по проверкам ЕГРЮЛ



Динамика мин. доли бракованных записей ЕГРЮЛ



Важно оценивать динамику метрик качества по времени.

Обратите внимание, что доля ошибок и минимальная доля бракованных записей отличаются на порядок.

Оценка качества Реестра производителей «ювелирки»

Статистика бракованных записей (на 01.072018)

Реестр	Брак	Записей	Процент	Интервал
ЮЛ	198	903	21,9%	4,6
ИП	593	3214	18,5%	5,4
Производители	791	4117	19,2%	5,2

Примеры перегруженных элементов данных (от 2 до 6)

Уч.номер - 0080007782 ИНН 5403172040
№ 0140008560 (ЮЛ7801401687 с 18.01.2018 от 02.02.2018 №30-14-02-03/117) ИНН 7806104390

Курьёзы наименования типов организаций:

- «...общество с ограниченной **возможностью**»
- «...общество с ограниченной **общественностью**»
- «...общество с ограниченной **отнесенностью**»
- «...общество с ограниченной **отвлечённостью**»

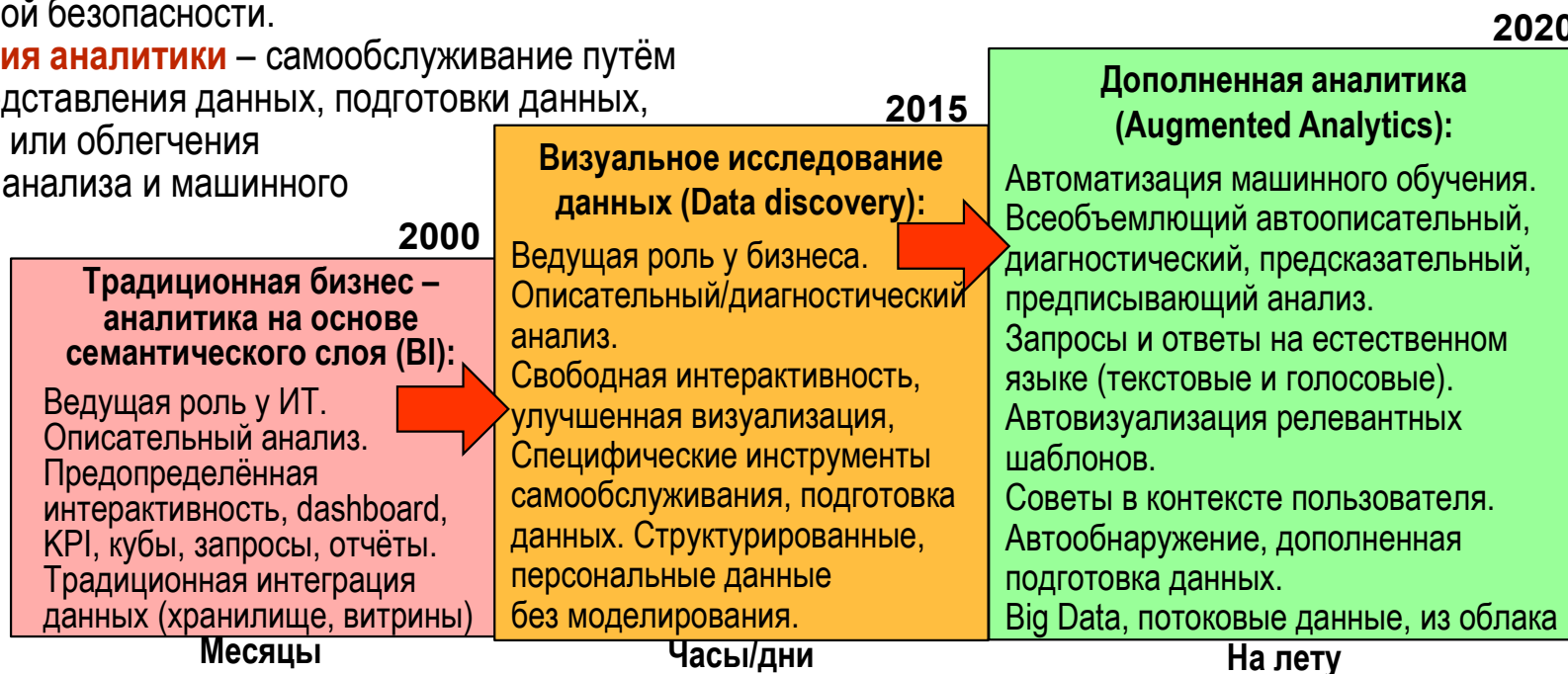
Анализ данных (Data Analysis)



Демократизация данных и аналитики vs Data Science



- **Способы самообслуживания:**
 - новые средства бизнес-аналитики, простые в применении для бизнес-пользователей
 - продвинутые средства анализа, но сложные в использовании для data scientist
- **Демократизация данных** – устранение оргструктурных барьеров, упрощение процедуры доступа к данным, исключение надуманных ограничений, однако с соблюдением необходимой информационной безопасности.
- **Демократизация аналитики** – самообслуживание путём упрощения представления данных, подготовки данных, автоматизации или облегчения исследования, анализа и машинного обучения.



Промежуток времени от постановки задачи до получения результата

Хранение и обработка данных (Data Storage & Operations)

Системы управления базами данных

- Реляционные СУБД SQL
- Многомерные СУБД MDDDB
- Системы управления объектно-ориентированными БД OODB
- Системы управления XML-данных XDB
- СУБД в оперативной памяти IMDB,
- Нереляционные СУБД NOSQL
- СУБД NewSQL
- Облачные СУБД

Файловые системы

- Файлы разных форматов
- Распределённая файловая система HDFS
- Blockchain

Системы хранения данных

- Дисковые массивы
- SSD
- NAS, SAN и облачные хранилища

Хранение и обработка данных (Data Storage & Operations)

Обработка данных

- Оперативная обработка транзакций OLTP
- Интерактивная аналитическая обработка OLAP
- Массивно-параллельная обработка MPP
- Распределённая пакетная обработка MapReduce
- Распределённые вычисления в памяти Spark
- Обработка информации с датчиков IoT& Edge computing
- Поточковая обработка данных Streams
- Программно-аппаратные комплексы бизнес-аналитики BI appliances
- Облачные сервисы Cloud Service
- Графические процессоры GPU

Управление документами и контентом (Document & Content Management)

Документы и контент

- Управленческие документы
- Информационные фонды
- Архивы
- Техническая документация (тексты, схемы и чертежи)
- Мультимедиа (картинки, фото, видео, аудиозаписи)

Управление документами и контентом

- Офисные пакеты и электронная почта
- Системы управления документами и документооборотом
- Системы архивирования
- Системы управления знаниями
- Мультимедийные системы
- Web-сайты и порталы

Защита данных (Data Security)

- Получение доступа
 - Идентификация, авторизация и аутентификация пользователей
 - Права доступа к приложениям
 - Права доступа к файлам
 - Права доступа к БД
-
- Ролевое управление доступом к данным RBAC
 - Атрибутивное управление доступом ABAC
 - Уровни доступа
 - Матрица доступа
 - Шифрование
 - Маскирование данных
 - Обезличивание данных

Управление качеством данных (Data Quality Management)

- Нужен широкий взгляд на управление данными в корпоративном масштабе
- Учитывайте информационные активы, чтобы обеспечить их прозрачность, понятность и повторное использование
- Не должно быть бесхозных информационных активов
- Управление бизнес–метаданными, реестрами и справочниками – залог высокого качества данных
- Измеряйте, отслеживайте, разрешайте инциденты, очищайте и раскрывайте сведения о качестве данных
- Повышайте вашу осведомлённость и ваши компетенции по корпоративному управлению данными

**Терпения и удачи всем, кто связан
с управлением данных**

Спасибо за внимание!

Валерий Иванович Артемьев

МГТУ имени Н.Э. Баумана, кафедра ИУ-5

Банк России

Департамент данных, проектов и процессов

Тел.: +7(495) 753-96-25

e-mail: viart@bmstu.ru