Минцифры России

Региональные витрины данных: Моделирование и обеспечение качества данных

НИИ «Восход», Департамент управления данными
Дмитрий Кашко

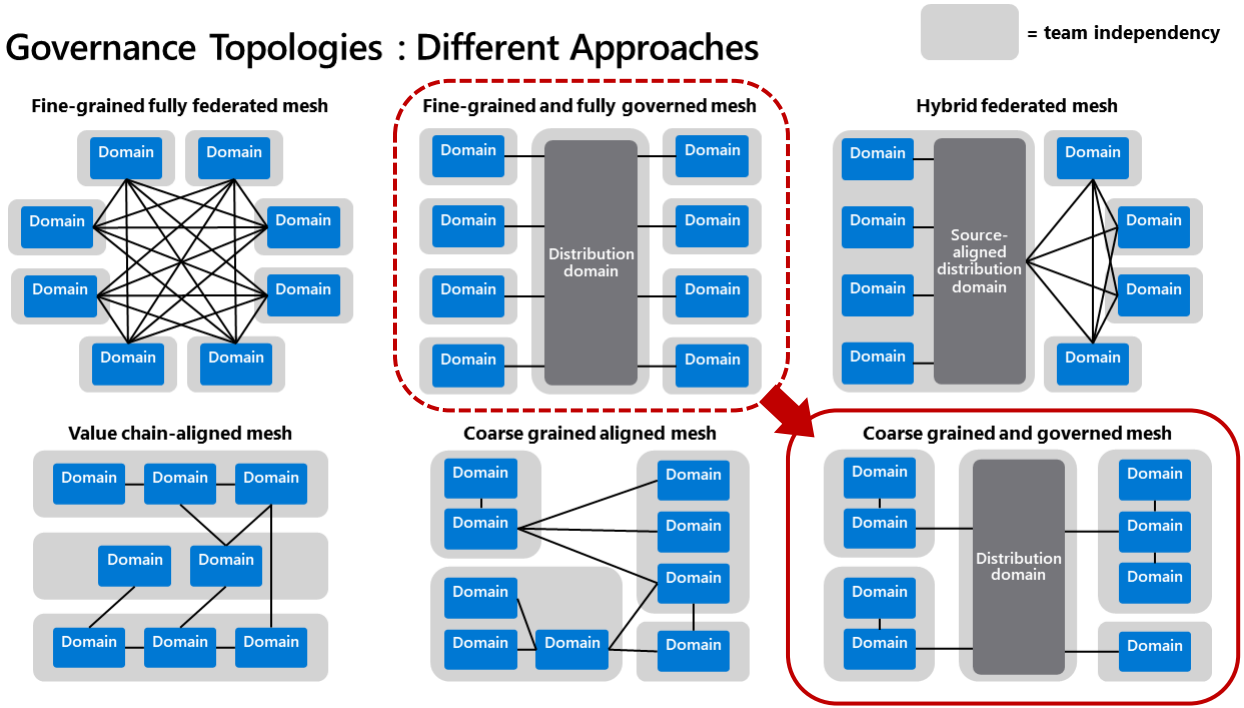
18.08.2022

Data Mesh – Сетка данных

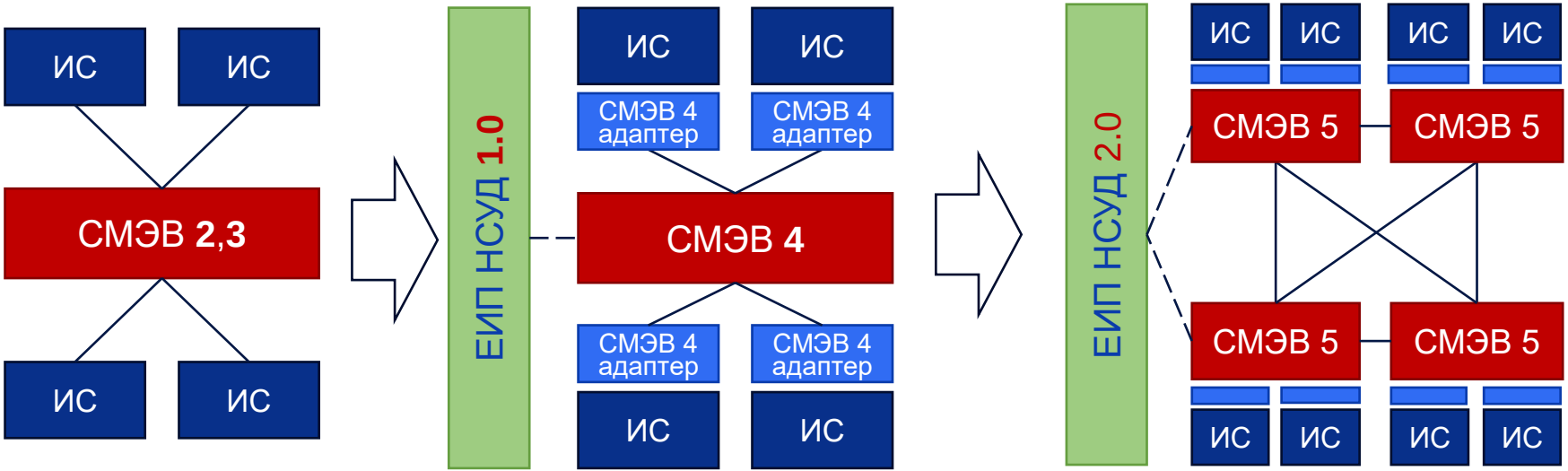
Из распространенных архитектурных моделей СМЭВ4 + НСУД реализует Data Mesh

Основные принципы Data Mesh:

- ✓ децентрализованное владение данными;
- ✓ данные как продукт;
- ✓ инфраструктура данных как платформа;
- ✓ федеративное управление вычислениями



Развитие архитектурной парадигмы СМЭВ



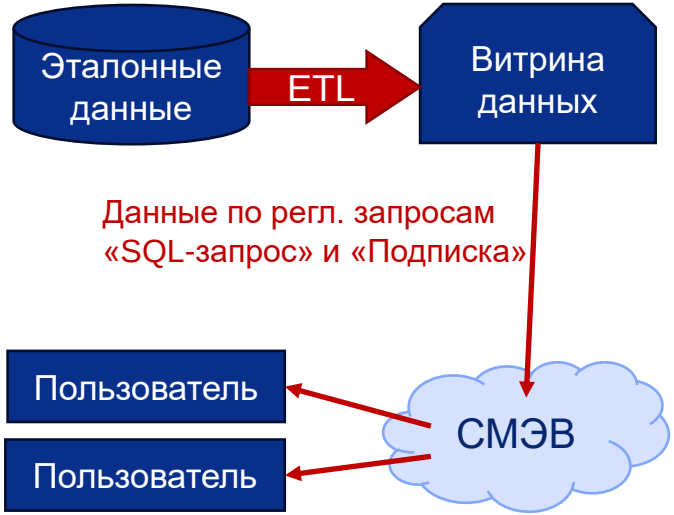
- Ограничено масштабирование;
- Сложное подключение;
- Сложная поддержка обратной совместимости

- Ограничено масштабирование;
- Простое подключение;
- Легкое развитие;
- Познаваемая модель данных

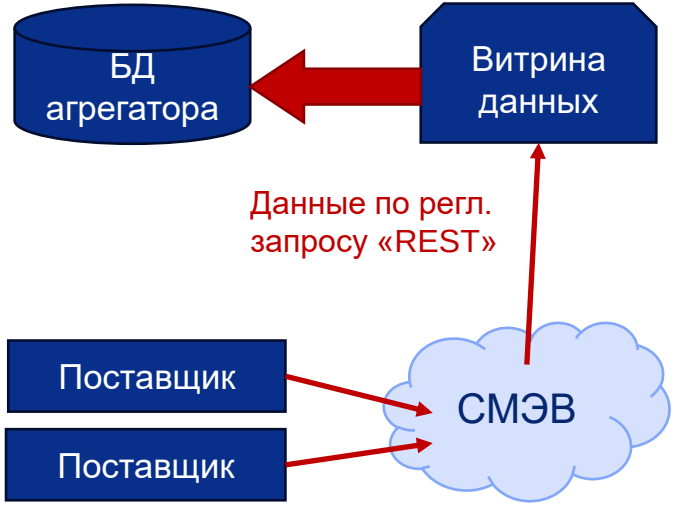
- Масштабирование неограничено;
- Простое подключение;
- Легкое развитие;
- Доменная топология, настраиваемые правила

Два подхода к использованию витрин

Представление данных



Сбор данных



Важное звено ETL

ETL (Extract – Transrorm - Load) может быть куда более мощным инструментом, чем традиционная выгрузка-загрузка текстовых файлов

Извлечь

- Выбрать нужные данные;
- Разместить в сериализованном виде в промежуточном хранилище.

Преобразовать

- Привести к необходимым типу и формату;
- Проключевать;
- Агрегировать, обогатить;
- Очистить.

Загрузить

- Определить и разрешить конфликты;
- Записать в БД витрины

Моделирование витрины данных

Витрина – не аналитическая БД, высокая нормализация ей не к чему

№	Название поля		Тип данных
1	**Серия**	PK	Строка
2	**Номер**	PK	Число
3	**Дата выдачи**		Дата
4	**Кем выдан**	FK	Число
5	**Статус документа**	FK	Число
6	**Причина аннулирования**	FK	Число
7	**Фамилия**		Строка
8	**Имя**		Строка
9	**Отчество**		Строка
10	**СНИЛС**		Строка

Связь с уникальным справочником организаций: предусмотреть гармонизацию с ЕГРЮЛ

Справочник статусов излишен, описать возможные статусы в комментарии к полю

Справочник причин излишен, приводить явно текст причины

Типичные ошибки моделирования:

- Названия (включая комментарии) не раскрывают суть таблиц, атрибутов
- Излишняя нормализация (НФ3 и выше)
- Одни и те же данные в разных таблицах
- Неверные типы полей
- Неверные первичные ключи

* Статьи и документы на nsud.gosuslugi.ru содержат более детальные рекомендации

Проверки качества данных

Проверки качества обязательны для каждой таблицы витрины и для всех индексов

- **Заполненность** (для строковых полей – наличие хотя бы одного печатного знака)
- **ФЛК** (допустимые символы для строк, допустимые диапазоны для дат и чисел)
- **Отсутствие дублей** (значения справочников, потенциальные ключи)
- **Связанность** (соответствие эталонным справочникам)

Возможности проверок витрин ограничены (например, сложно оценить полноту данных)

Обеспечение качества данных

Как качество продукции обеспечивается на заводе, так и качество данных обеспечивается в ведомственных информационных системах

Определите ответственных

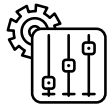
CDO, подразделения, роли и должностные обязанности, цели и KPI

Распланируйте деятельность

Выявление источников ошибок, автоматические проверки и регулярные аудиты, PDCA-цикл

Наладьте обратную связь с пользователями

Виджеты и чат-боты для пользователей, обработка инцидентов СЦ и ПОС, анализ тенденций

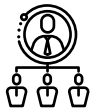


Устраните источники проблем

Ручной ввод, дублирующие данные, отсутствие контроля…

Вопрос: Централизация или федерализация?

Взаимодействие через единый хаб



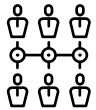
Плюсы:

- Единое управление и контроль на уровне субъекта;
- Во многих регионах уже есть соответствующие оргструктуры и технические решения

Сложности:

- Управление правами доступа к данным («хаб» не имеет прав доступа к многим передаваемым данным);

Каждый участник самостоятелен



Плюсы:

- Технологически простое решение;
- Обычно меньше затрат на внедрение

Сложности:

- Усложнены контроль и управление на уровне региона
- Обычно больше затрат на сопровождение

Спасибо за внимание