приоритет

SmartMLOps – система размещения и управления сервисами искусственного интеллекта

Третий открытый онлайн семинар с командой SmartMLOps

Салех Хади Мухаммед, Департамент программной инженерии Факультета компьютерных наук НИУ ВШЭ г. Москва
hsalekh@hse.ru

https://mlops.hse.ru/

Москва 2025

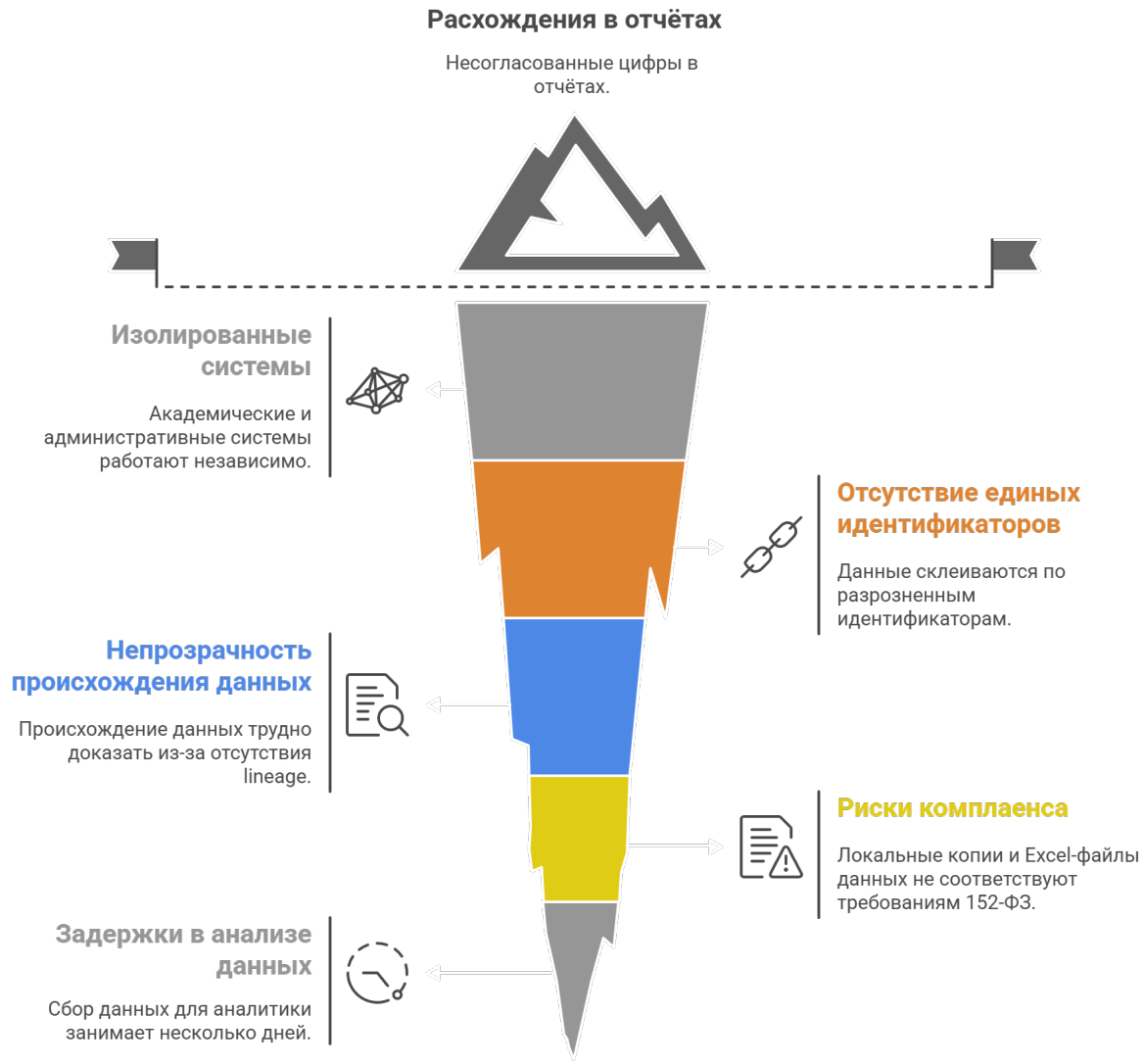
Озеро данных НИУ ВШЭ

Промежуточные результаты

Повестка семинара

1	**Определение необходимости** Определение необходимости в Озере данных для НИУ ВШЭ
2	**Обзор концепции** Обзор документа «Концепция создания и развития Озера данных»
3	**MVP-фокус** Сосредоточение на сущности «Студент» как на минимально жизнеспособном продукте
4	**Процесс сбора данных** Описание процесса сбора и обработки данных
5	**Автоописание метаданных** Внедрение автоматического описания метаданных
6	**Подключение НСИ** Интеграция с утвержденной НСИ по домену данных
7	**Прототип семантического слоя** Разработка прототипа семантического слоя
8	**Планы до октября 2025** Планирование будущих действий до октября 2025 года

Почему работа с данными сегодня – боль?



Зачем НИУ ВШЭ корпоративное Озеро данных

База для AI/ML

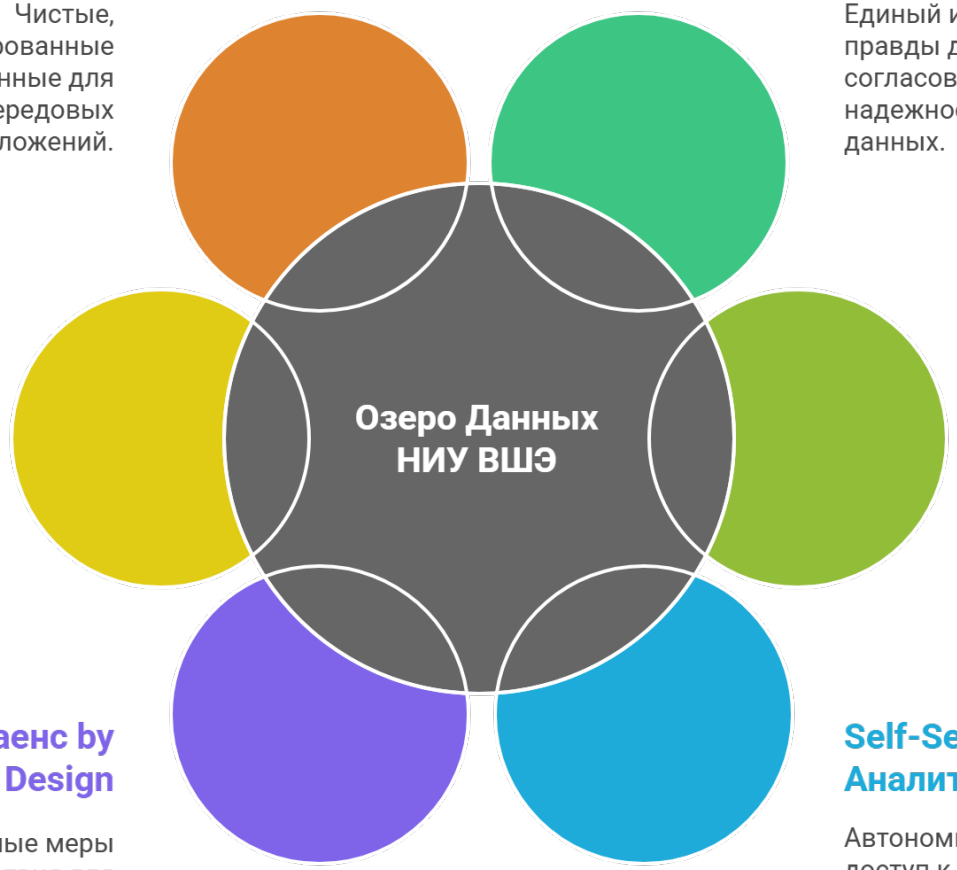
Чистые, версионированные данные для передовых приложений.

Централизация Данных

Единый источник правды для согласованности и надежности данных.

Скорость

Ускоренное время подготовки отчётов для повышения эффективности.



Каталог и Зоны Данных

Организованные зоны для эффективной навигации и контроля качества.

Комплаенс by Design

Встроенные меры соответствия для защиты данных.

Self-Service Аналитика

Автономный доступ к данным для аналитических команд.

Документ «Концепция» – наш проектный ТЗ

Цели и границы

Зафиксированы цели, границы и KPI проекта.

Описан архитектурный каркас, включая технологии, зоны и процессы.

Архитектурный каркас

Определены роли

Определены роли, такие как владелец данных, архитектор, DevOps и DPO.

Содержит оценку рисков, бюджетный коридор и квартальные вехи.

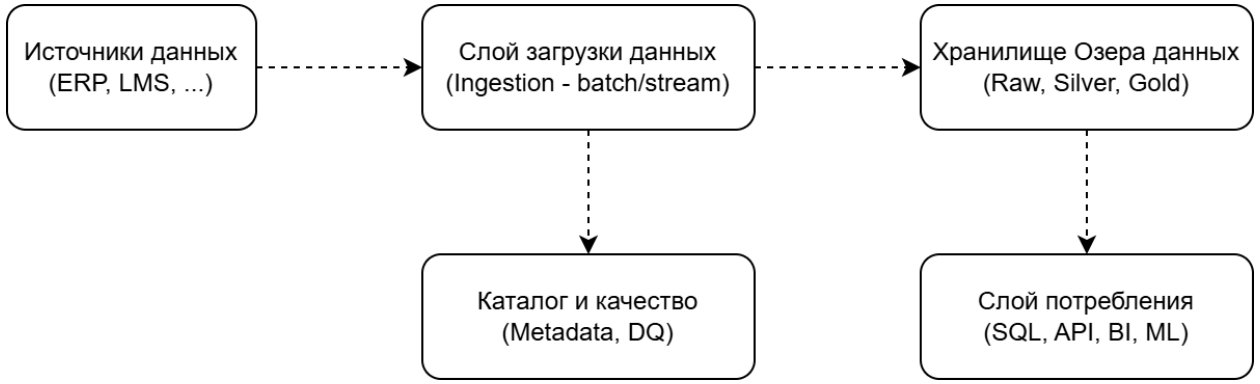
Оценка рисков

Живой документ

«Живой» документ, обновляемый по итогам спринтов и ревью стейкхолдеров.

Структура Концепции

Логическая архитектура Озера данных



- **источники данных.** Информационные системы, таблицы баз данных, файлы CSV/JSON и внешние реестры;
- **слой загрузки (Ingestion).** Коннекторы ETL/CDC принимают данные, обогащают метаданными и сохраняют в Raw;
- **хранилище Озера.** Три уровня:
    1) Raw (бронзовый) – неизменённые данные;
    2) Silver – очищенные и нормализованные;
    3) Gold – агрегированные домены данных для отчётов и ML.
- **каталог и контроль качества.** Система метаданных фиксирует происхождение данных и запускает правила Data Quality;
- **слой потребления.** Доступ через SQL, REST и BI-коннекторы; ML-лаборатории подключаются напрямую к Gold.

Data Lake ≠ DWH: зональная архитектура

Безопасность и управление данными по умолчанию

Шифрование

Защита данных в покое и в движении с помощью надежных ключей.

RBAC/ABAC

Управление доступом на основе ролей и атрибутов для контроля доступа.

Маскирование ПДн

Защита личной информации с помощью динамических методов маскирования.

Аудит & SIEM

Мониторинг и запись каждого запроса для обеспечения безопасности.

Тесты качества данных

Проведение регулярных проверок для поддержания точности данных.

Соответствие 152-ФЗ

Соблюдение нормативных требований для защиты данных.

Каталог + семантический слой = self-service аналитика

API метаданных

Доступ к lineage и описаниям для DevOps и Data Scientists

Каталог данных

Автоматическое сканирование схем и полнотекстовый поиск

BI без SQL

Аналитика, подключающаяся к доменам, а не к таблицам

Бизнес-глоссарий

Единые определения ключевых терминов

Единые метрики

Вычисление ключевых показателей один раз

Семантический слой

Модели dbt, скрывающие технические детали

Почему MVP = сущность «Студент»

Использование данных

Большинство управленческих отчётов вуза используют данные о студентах.

Данные уже доступны в трёх ключевых системах: HSE.REG, SmartPlan, MDM.

Доступность данных

Нормализация НСИ

НСИ содержит справочники, которые легко нормализовать.

Скромный объём: около сотни атрибутов, десятки таблиц, реалистично за пару месяцев.

Скромный объём

Высокий эффект

Высокий эффект: единый домен данных избавит от ручного свода.

Пригодно для ML-кейсов (выглядим “Data-Driven” уже на пилоте).

Пригодно для ML

Пятиэтапный pipeline

Выделение бизнес-сущностей

Идентификация и изоляция ключевых бизнес-сущностей.

Автоматическая инвентаризация таблиц

Автоматическая каталогизация и документирование таблиц базы данных.

Семантический анализ

Анализ и группировка полей на основе их значений.

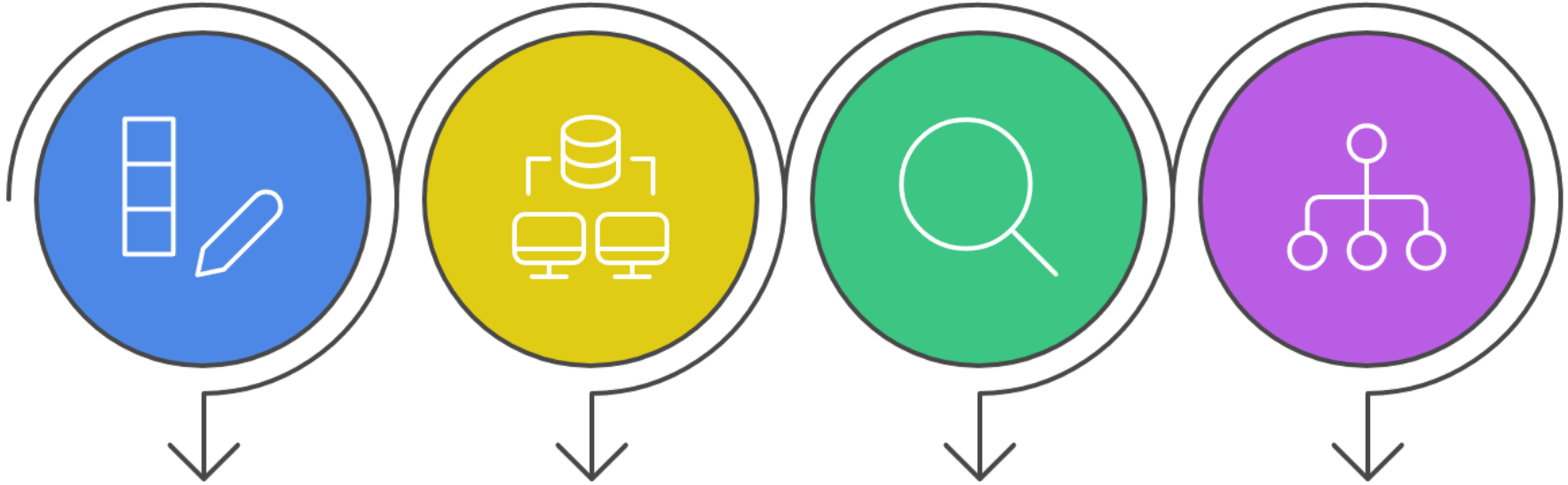
Сопоставление таблиц и сущностей

Связывание таблиц с соответствующими бизнес-сущностями.

Формирование семантического слоя

Создание представлений и семантического слоя.

Что выгружаем из таблицы



Имена таблиц и полей

Названия таблиц и полей

Типы данных

Типы данных

Примеры значений

Примеры значений

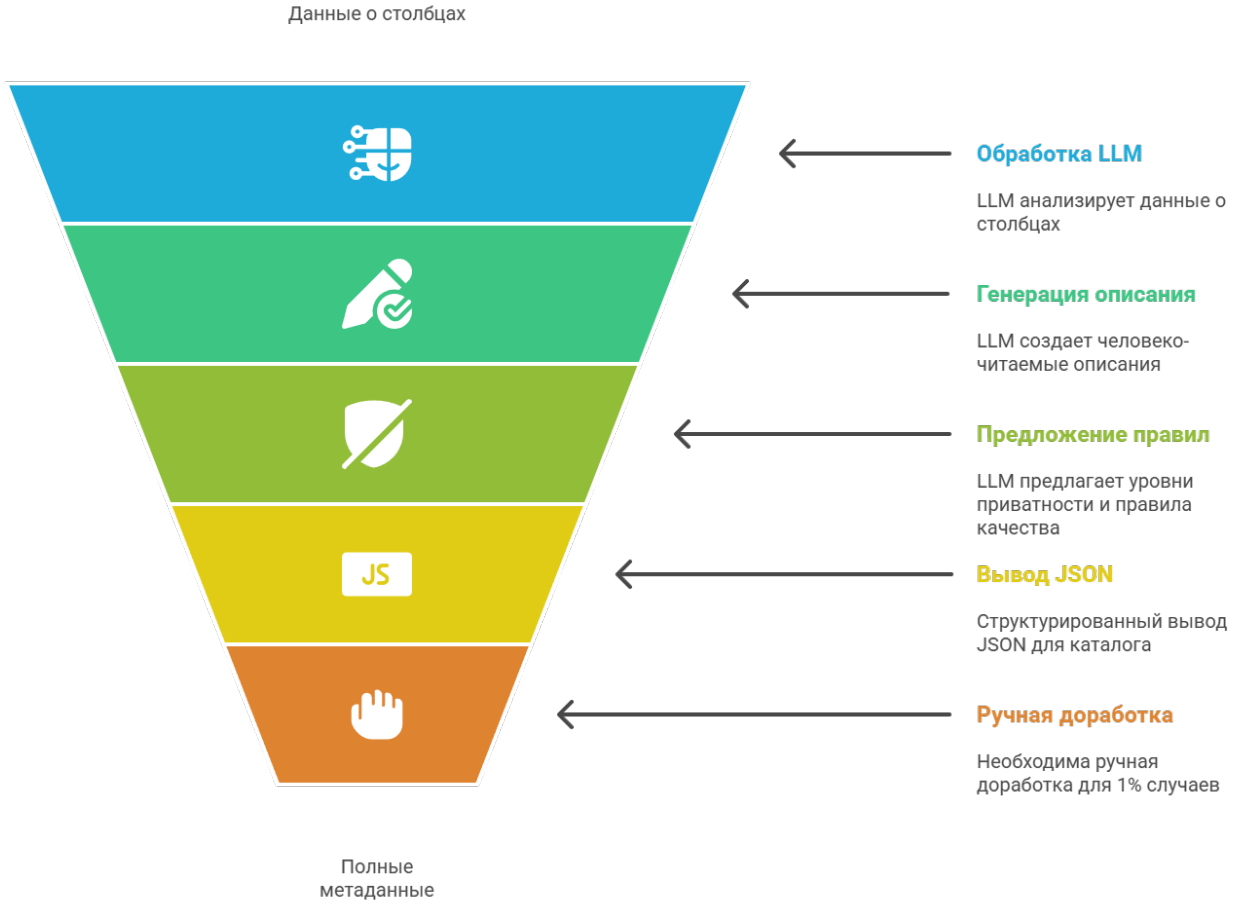
Связи и статистика

Связи и статистика

Результаты обработки метаданных

Характеристика	HSE.REG	SmartPlan	MDM	Интеграционная база
Таблицы	66	85	33	64
Столбцы	1270	1027	428	1035
Комментарии	43	10	0	11

LLM-генерация человеко-читаемых описаний колонок



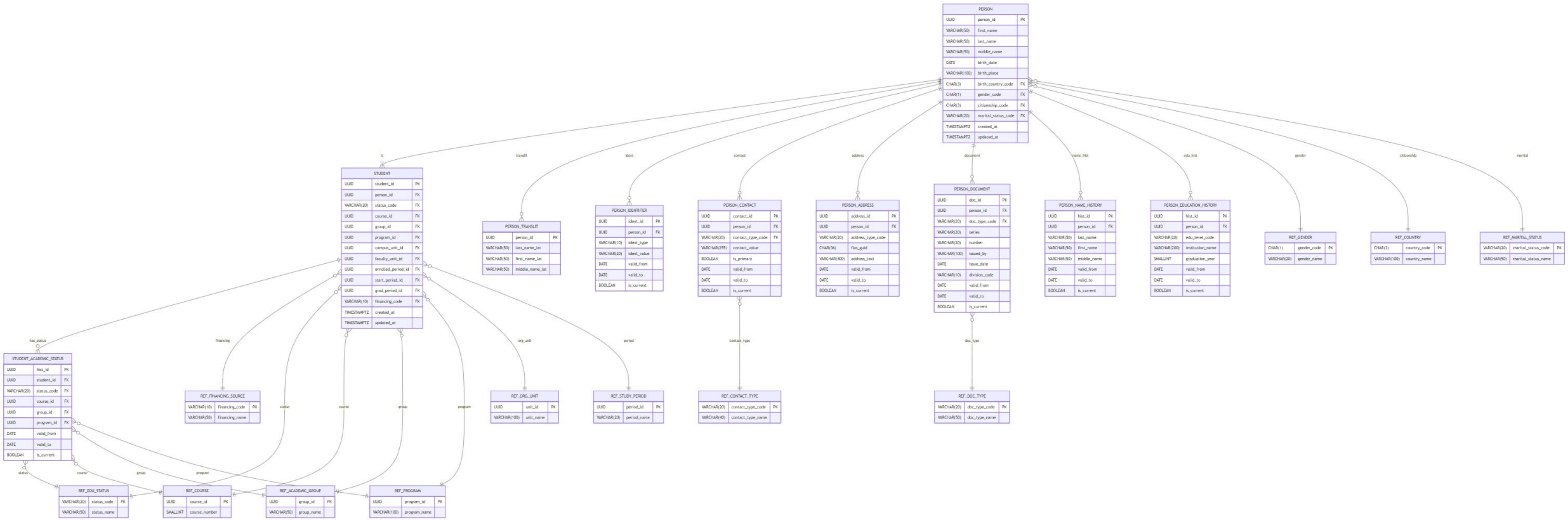
hsereg_curr.changeslearnplan	13 443	Регистр сведений "Корректировки ИУП"	Корректировка ИУП	Регистр всех изменений индивидуального учебного плана студента (добавление/удаление дисциплин, смена периода).
smart_plan_curr.cpecialitieslearnplan	2 602		Специальность индивидуального плана	Связывает запись ИУП с кодом специальности (speciality_id).
smart_plan_curr.distantsubjectincp	131 138		Дистанционные дисциплины учебного плана	Хранит связи учебного плана с онлайн-курсами (LMS/ внешние платформы): идентификаторы курса, платформы, трудоёмкость, год обучения и т.д.

Подключение НСИ: единые справочники в домене «Студент»

Справочник или реестр	Атрибут	ID атрибута НСИ НИУ ВШЭ[1]	Описание	Обязательность[2]	Система источник[3]	Подразделение, осуществляющее управление данными НСИ[4]
Справочник «Студент»	Студент	021.0013	Владелец. Значение выбирается из справочника «Физические лица». Etalon ID физического лица в МДМ.	Да	HSE.Reg	Центр сервиса «Студент» Дирекции основных образовательных программ (01.88.06)
Справочник «Студент»	Статус обучения	021.0001	Значения выбираются из вспомогательного справочника «Статусы обучения» (022.0020). Код НСИ.	Да	HSE.Reg	Центр сервиса «Студент» Дирекции основных образовательных программ (01.88.06)
Справочник «Студент»	Курс	021.0002	Значения выбираются из вспомогательного справочника «Курсы обучения» (021.0027). Etalon ID.	Нет	HSE.Reg	Центр сервиса «Студент» Дирекции основных образовательных программ (01.88.06)

[1] 006, 021, 024 – ID атрибута НСИ НИУ ВШЭ не относится к какой-либо информационной системе.
[2] Обязательность отражает необходимость заполнения соответствующих данных по каждой записи НСИ в АСУ НСИ.
[3] Информационная система, в которой осуществляется ручной ввод и первичное формирование соответствующих данных НСИ.
[4] Управление данными – право подразделения устанавливать требования к данным, право на внесение и изменение данных, а также обязанность поддержания актуальности и корректности данных в системе-источнике.

ER-диаграмма домена «Студент»



План работ: август-октябрь 2025

июл.25	Разработка прототипа семантического слоя
авг.25	Разработка сервиса анонимизации данных
сен.25	Интеграция автоматического сбора метаданных и обновление онтологии
окт.25	Проведение модульных и интеграционных тестов компонентов Разработка отчетной документации: на компоненты

Вопросы?

Силаев Юрий Владимирович

https://t.me/SilaevYV

ysilaev@hse.ru

+7-929-533-3406