ЖИЗНЕННЫЙ ЦИКЛ ДАННЫХ

при проектировании и разработке информационных систем

Калядин Леонид

Cluster Data lead

главный администратор данных

Experience

- → **DWH developer** — Aircraft company
- → **Senior DWH developer** — Russian Post
- → **DWH Team lead** — DOM.RF
- → **Architecture and Tech lead** — CROC
- → **Cluster Data lead** — MTS

Hard Skills

Skills and abilities

BI and Reporting

Data Warehouses and integrations

MDM

System & Business Analysis

Data Governance/ Data Quality

Жизненный цикл данных

Теоретическая часть

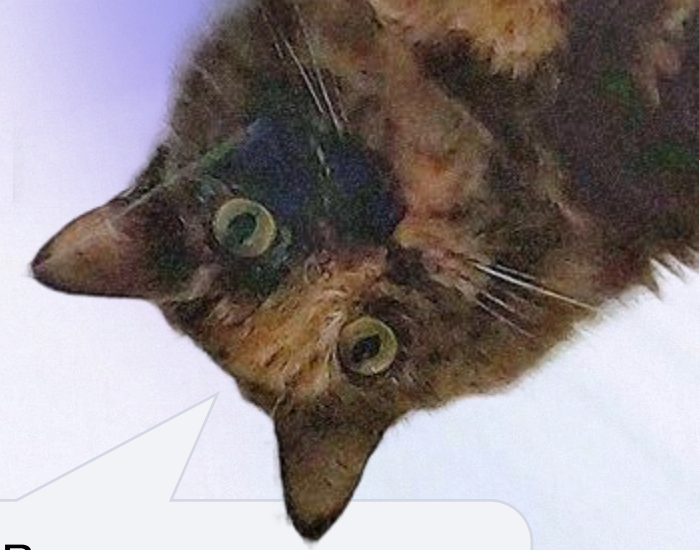
Жизненный цикл
Что это?

Жизненный цикл
Что это?

По обратной интеграции удали

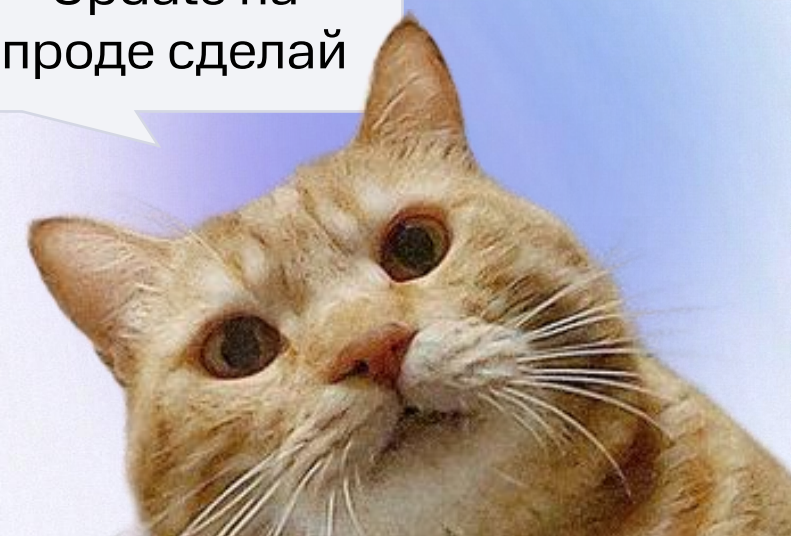
Excel’ку загрузи

- Создание
- Хранение
- Изменение
- Использование
- Удаление/ Архив



Возможность для ручного изменения оставь

Update на проде сделай



Жизненный цикл данных

Создание

01. Создание данных

Хранение

Жизненный цикл данных

02. Хранение истории данных

SCD 1

client

name	passport	Update_at
Леонид	1111 222222	17.09.2023
Леонид	1111 333333	20.09.2023

SCD 3

client

name	passport	Update_at	Prev_pass
Леонид	1111 222222	17.09.2023	null
Леонид	1111 333333	20.09.2023	1111 222222

SCD 4

client

name	passport	Update_at
Леонид	1111 222222	17.09.2023
Леонид	1111 333333	20.09.2023

client_hist

name	passport	Update_at
Леонид	1111 222222	17.09.2023
Леонид	1111 222222	17.09.2023
Леонид	1111 333333	20.09.2023

02. Хранение истории данных

SCD 2

client

name	passport	Dt_from	Dt_to
Леонид	1111 222222	17.09.2023	01.01.9999

Леонид	1111 222222	17.09.2023	20.09.2023
Леонид	1111 333333	20.09.2023	01.01.9999

SCD Hybrid

client

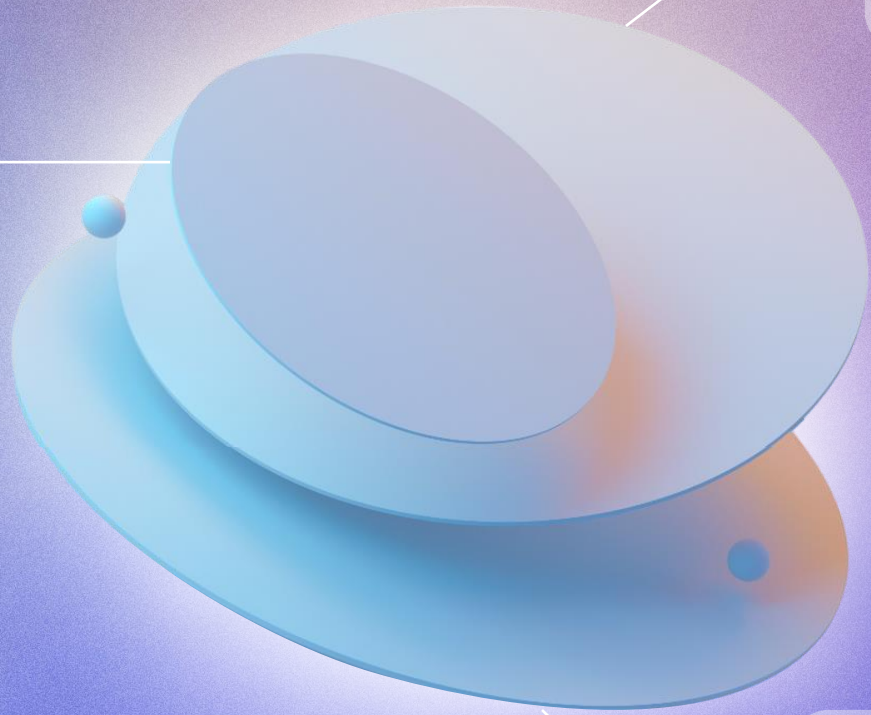
name	passport	Dt_from	Dt_to	Is_current
Леонид	1111 222222	17.09.2023	01.01.9999	True()

Леонид	1111 222222	17.09.2023	20.09.2023	False()
Леонид	1111 333333	20.09.2023	01.01.9999	True()

Сбор требований

Определить
системы-потребители данных

Определить
какой **тип системы**



Определить
бизнес-сценарии использования данных

Жизненный цикл данных

Изменение

03. Изменение данных

Атрибуты клиента в разных системах

client_sys1

name	snils	Is_married	Childs_cnt	updated_at
Леонид	1111	False()	0	18.09.2021
Леонид	1111	True()	0	20.09.2022
Леонид	1111	True()	1	21.09.2023

client_sys2

name	snils	salary	updated_at
Константин	1112	100	19.11.2021
Леонид	1111	150	21.09.2022
Леонид	1111	200	21.09.2023

Исходные данные

clients

name	snils	Is_married	Childs_ cnt	Salary	Valid_from	Valid_to
Леонид	1111	False()	0	*null*	18.09.2021	19.11.2021
Константин	1112	False()	0	100	19.11.2021	20.09.2022
Константин	1112	True()	0	100	20.09.2022	21.09.2022
Леонид	1111	True()	0	150	21.09.2022	21.09.2023
Леонид	1111	True()	1	200	21.09.2023	01.01.9999

03. Изменение данных

Атрибуты клиента в разных системах сильно отличаются и могут «противоречить» друг другу

client_sys1

Full_name	snils	Is_married	workplace	worktitle	Cars_cnt	Childs_cnt	updated_at
Леонид	1112	False()	Other	Tech-lead	1	0	18.09.2021
Леонид	1111	True()	null	null	1	0	20.09.2022
Леонид	1111	True()	MTS	Cluster Lead	2	1	21.09.2023

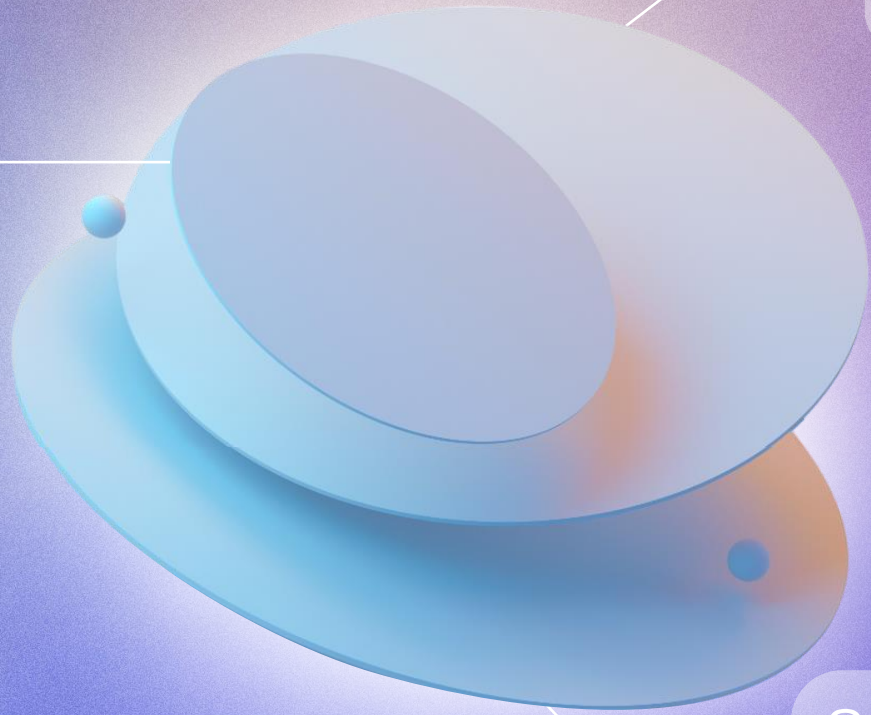
client_sys2

first_name	last_name	patronymic	Is_married	snils	salary	updated_at	first_name
Константин	Иванов	Иванович	False()	1112	100	19.11.2021	Константин
Леонид	Калядин	Валерьевич	False()	1111	150	20.09.2022	Леонид
Леонид	Калядин	Валерьевич	False()	1111	200	21.09.2023	Леонид

Сбор требований

Определить, **поле по которому можно рассчитать инкремент**

Определить **бизнес-ключ по каждому источнику**



Определить частоту изменений данных, возможность **изменения данных задним числом**

Жизненный цикл данных

Использование

04. Использование данных. Data Quality

loan

Номер кредита в системе	Дата выдачи кредита	Статус	Тип заёмщика	ФИО заёмщика	ID заёмщика	Остаток основного долга	Дата информации по остатку основного долга
19991-2221	20.09.2023	Выдан	Физ. лицо	Иванов Иван Иванович	285	100	20.09.2023

Проверки качества данных

Дата выдачи кредита не может быть больше сегодняшней даты

Номер кредита не может быть пустым

ФИО заёмщика не может быть пустым

Остаток основного долга за текущий период должен быть <= остатка основного долга за предыдущий период

04. Использование данных. Data Quality

loan

Номер кредита в системе	Дата выдачи кредита	Статус	Тип заёмщика	ФИО заёмщика	ID заёмщика	Остаток основного долга	Дата информации по остатку основного долга
19991-2221	20.09.2023	Выдан	Физ. лицо	Иванов Иван Иванович	285	100	20.09.2023

Проверки качества данных

Дата выдачи кредита не может быть больше сегодняшней даты

Если статус «Выдан»

Номер кредита не может быть пустым

ФИО заёмщика не может быть пустым

Остаток основного долга за текущий период должен быть <= остатка основного долга за предыдущий период

04. Использование данных. Data Quality

loan

Номер кредита в системе	Дата выдачи кредита	Статус	Тип заёмщика	ФИО заёмщика	ID заёмщика	Остаток основного долга	Дата информации по остатку основного долга
19991-2221	20.09.2023	Выдан	Физ. лицо	Иванов Иван Иванович	285	100	20.09.2023

Проверки качества данных

Дата выдачи кредита не может быть больше сегодняшней даты

Если статус «Выдан»

Номер кредита не может быть пустым

Это бизнес-ключ, по другому никак

ФИО заёмщика не может быть пустым

Остаток основного долга за текущий период должен быть <= остатка основного долга за предыдущий период

04. Использование данных. Data Quality

loan

Номер кредита в системе	Дата выдачи кредита	Статус	Тип заёмщика	ФИО заёмщика	ID заёмщика	Остаток основного долга	Дата информации по остатку основного долга
19991-2221	20.09.2023	Выдан	Физ. лицо	Иванов Иван Иванович	285	100	20.09.2023

Проверки качества данных

Дата выдачи кредита не может быть больше сегодняшней даты

Если статус «Выдан»

Номер кредита не может быть пустым

Это бизнес-ключ, по другому никак

ФИО заёмщика не может быть пустым

Тип заёмщика – Физ. лицо

Остаток основного долга за текущий период должен быть <= остатка основного долга за предыдущий период

04. Использование данных. Data Quality

loan

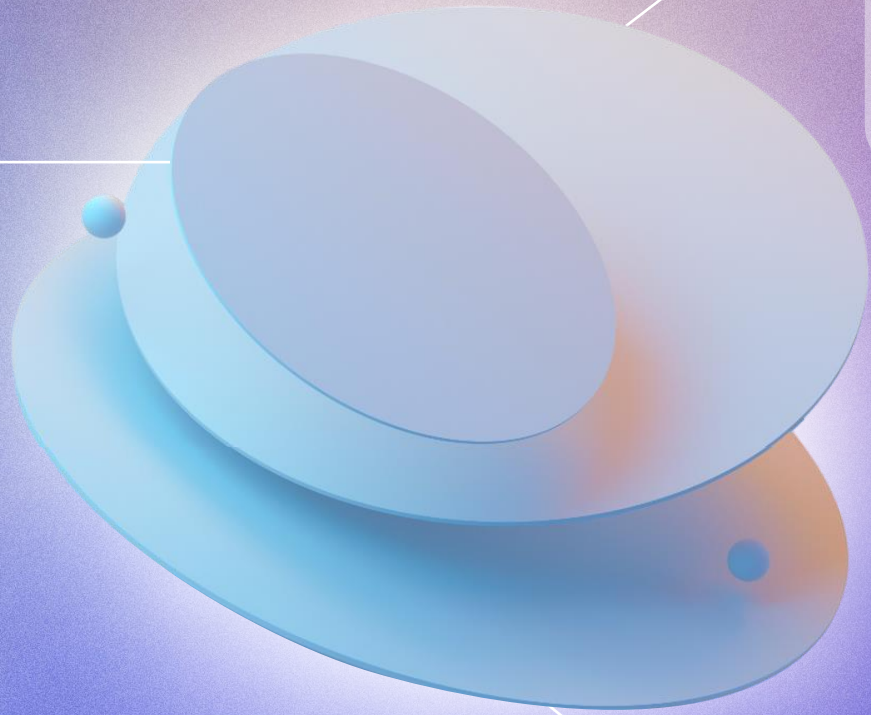
Номер кредита в системе	Дата выдачи кредита	Статус	Тип заёмщика	ФИО заёмщика	ID заёмщика	Остаток основного долга	Дата информации по остатку основного долга
19991-2221	20.09.2023	Выдан	Физ. лицо	Иванов Иван Иванович	285	100	20.09.2023

Проверки качества данных

Дата выдачи кредита не может быть больше сегодняшней даты	Номер кредита не может быть пустым	ФИО заёмщика не может быть пустым	Остаток основного долга за текущий период должен быть <= остатка основного долга за предыдущий период
Если статус «Выдан»	Это бизнес-ключ, по другому никак	Тип заёмщика – Физ. лицо	Иногда возможно

Сбор требований

Определить
в каких ситуациях данные не должны попасть в систему,
в каких допустимо



Определить, что делать с ошибками

1. храним отдельно
2. выгружаем в отчет
3. как происходит оповещение

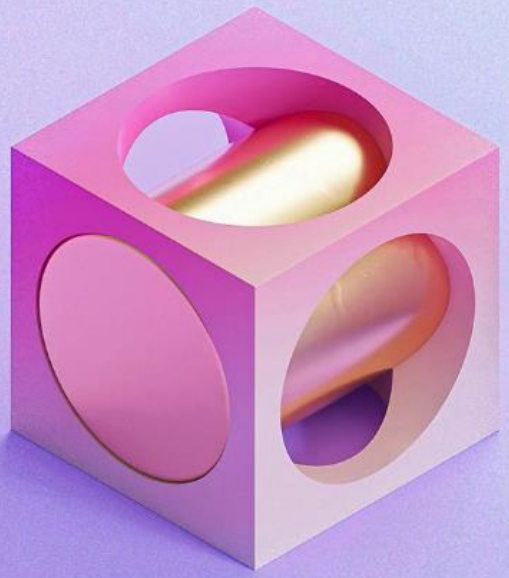
Кто **исправляет ошибки** данных

Жизненный цикл данных

Архивирование

05. Архивирование

Варианты архивирования



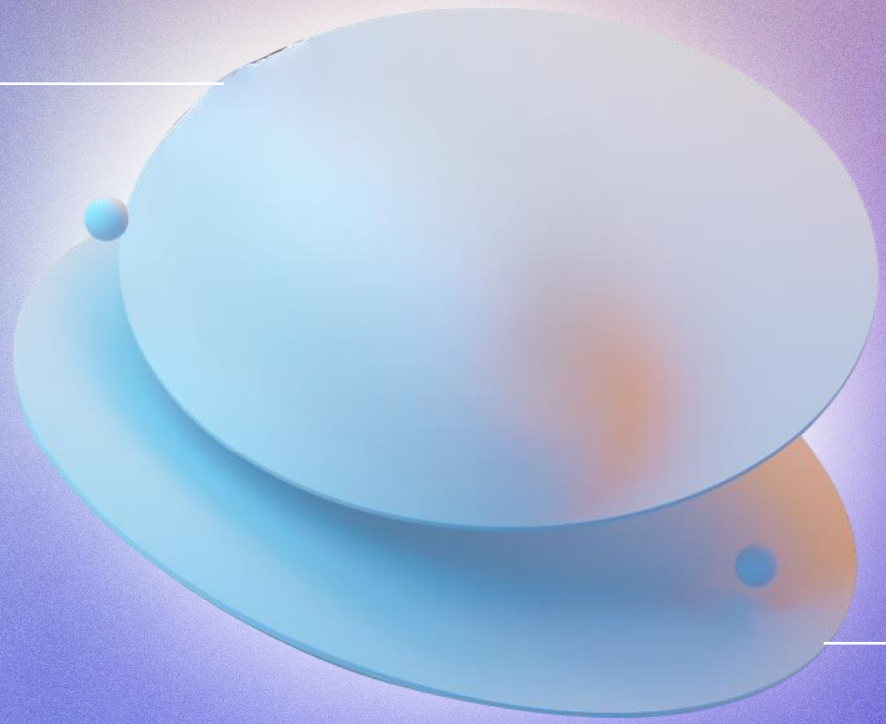
Отгрузить
куда-то и забыть

Всё удалить

Разделение на слои
или поднять стенд
для архива

Сбор требований

Определить **регуляторные требования к архивации данных**



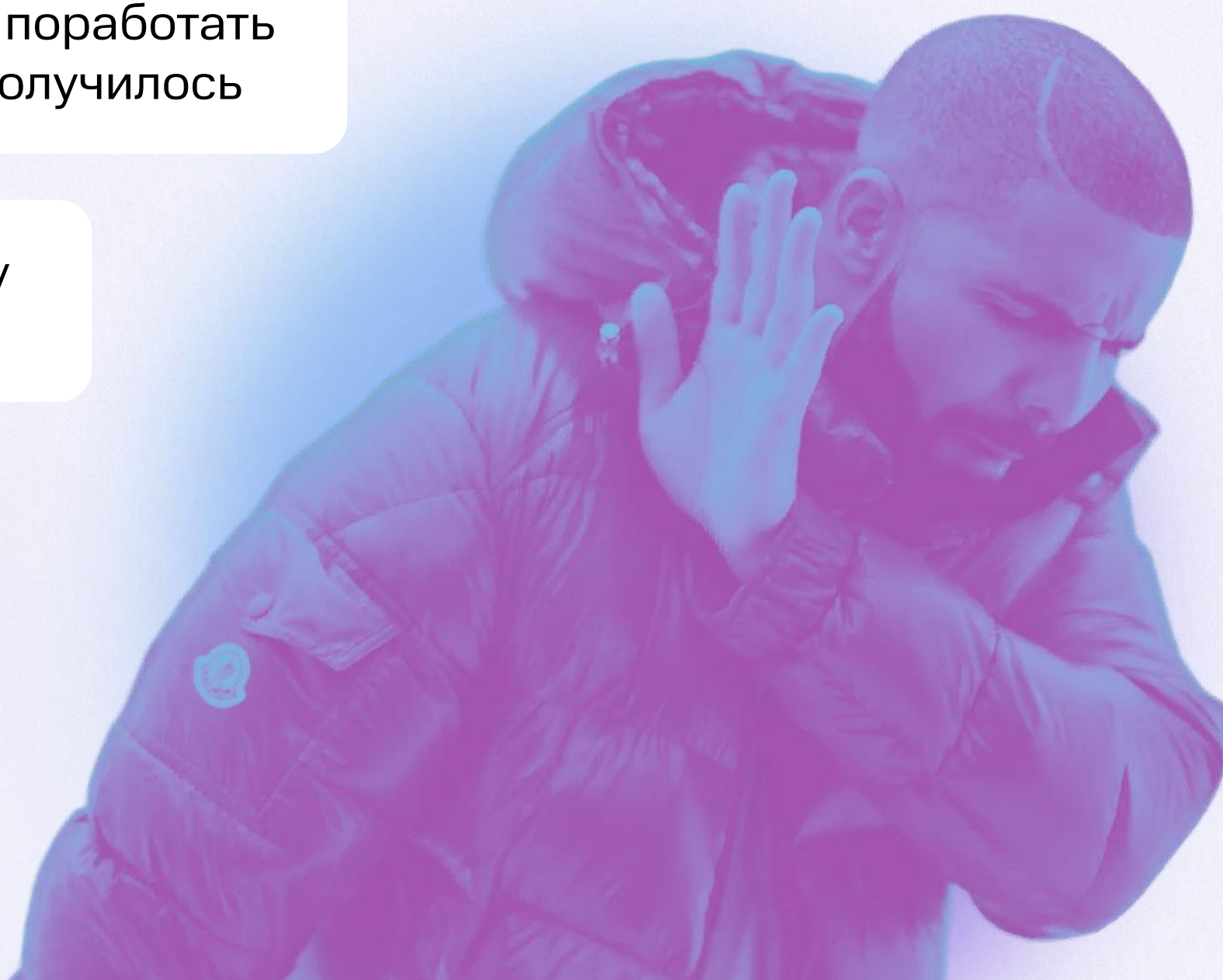
Определить, **технические возможности для создания архива**

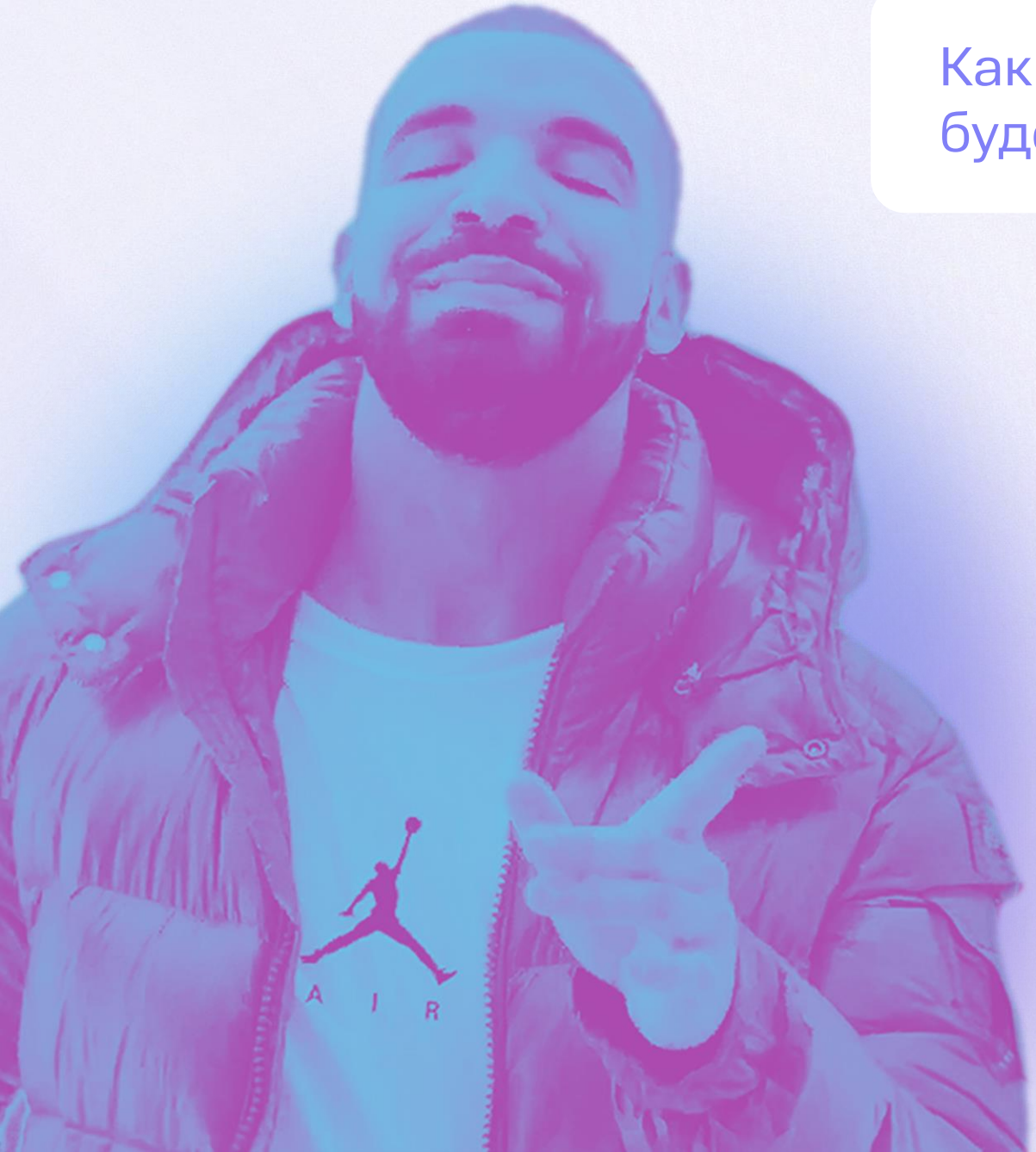
Нам не дали возможности поработать как следует и все коряво получилось

Эти ребята каждую минуту меняют требования

Я в принципе не знаком с этими технологиями

Ну теперь тут всю систему переписывать





Какие требования к истории данных, будет ли требоваться отчетность?

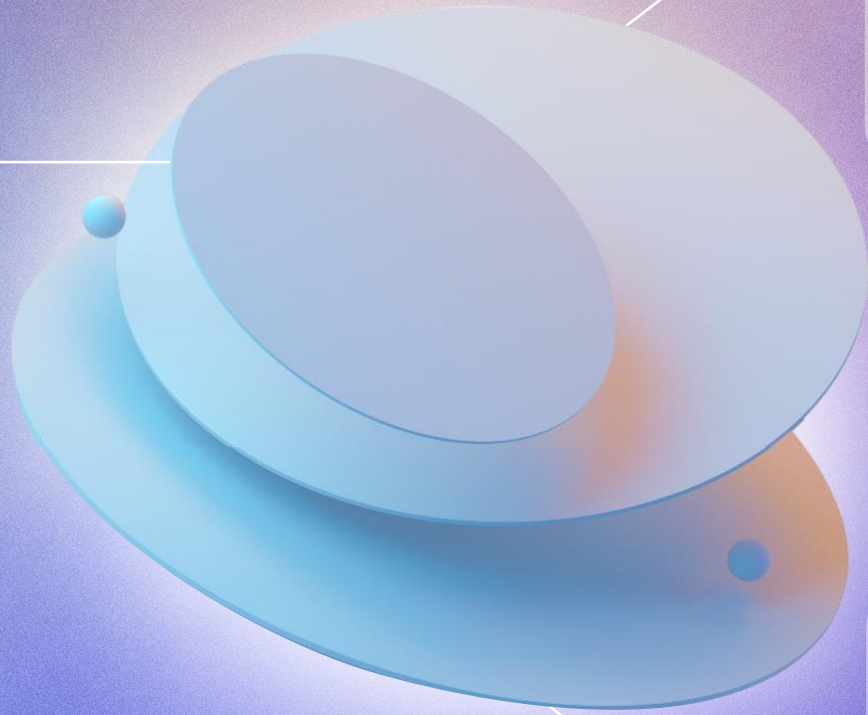
Сколько источников данных предполагается?

Где будем вести документацию?

Что на счет процессов Data Quality?

Итоги

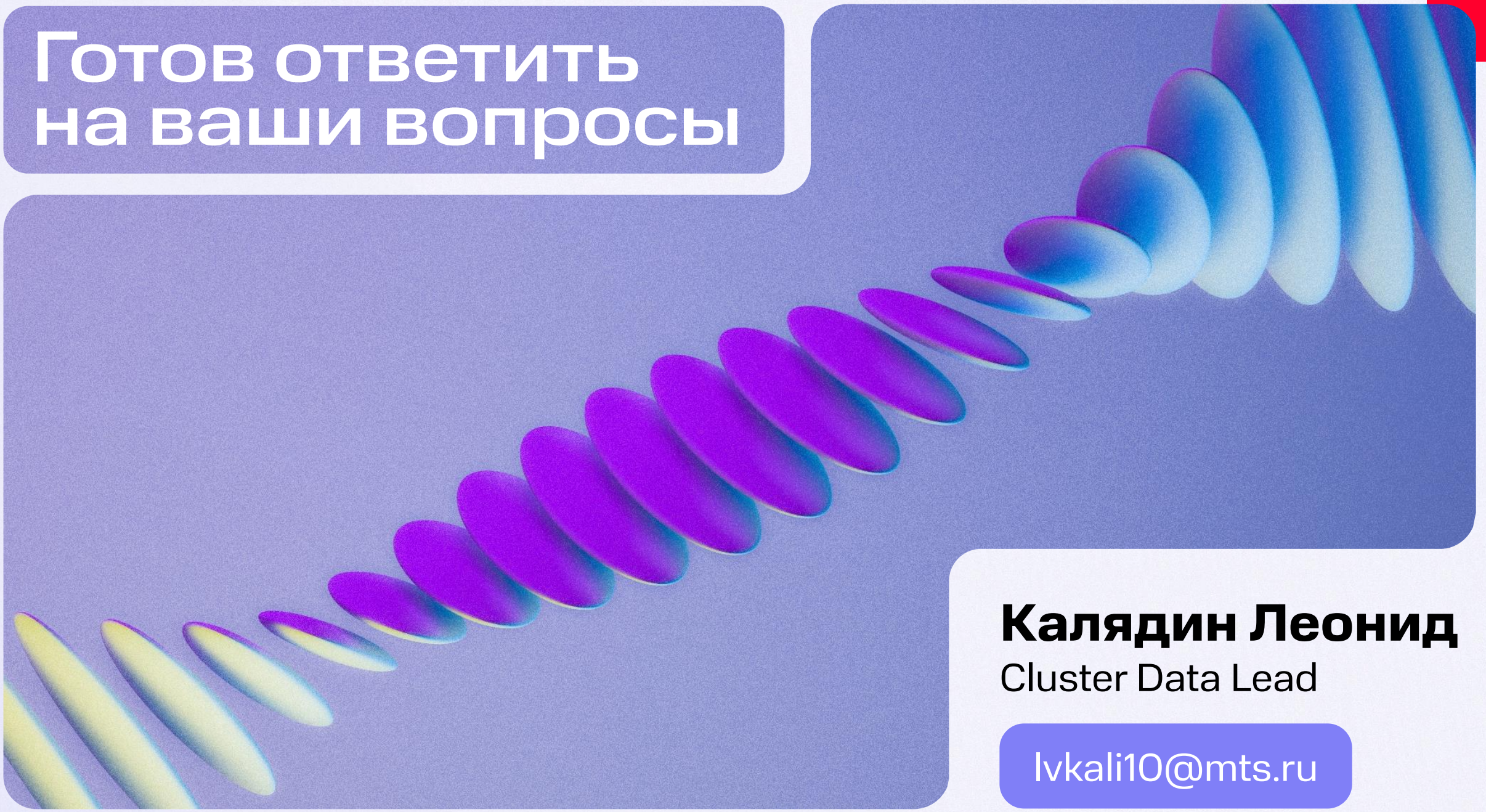
Жизненный цикл данных есть **не только у ваших** данных



Если **требований к истории хранения** у бизнеса сейчас нет, то когда появятся, может быть поздно

Качество данных определяется особенностями вашего бизнес-процесса и жизненным циклом данных

Готов ответить на ваши вопросы



Калядин Леонид
Cluster Data Lead
lvkali10@mts.ru