Lake House в Магните и наш путь к self-service

DATA

Платформа данных Магнит в числах

>1 ПБ

Данных

10 000

Пользователей отчетов и ad-hoc аналитики

~60

Систем источников, которые поставляют данные в платформу

Из них в текущем КХД:

~500 ТБ

Объем продуктивных данных (без учета репликаций) + песочницы

203 Млн

Среднее количество запросов в месяц

2479

Пользователей ежемесячно

>3 000

Боевых витрин

84

Песочницы предоставлено бизнесу для прототипирования решения своих задач

TECH

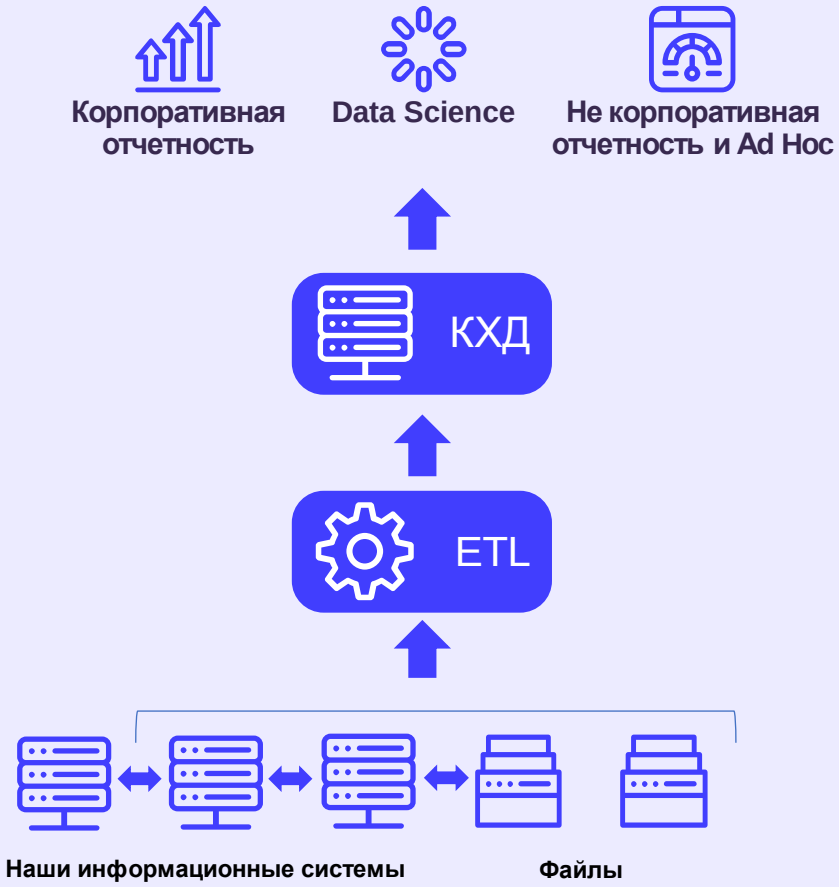
Предпосылки к появлению Озера

Прототипирование ограничено	Только команда КХД имеет доступ к сырым данным в текущей платформе, остальные ходят в источники
Сначала витрина, потом анализ	Результат разработки не всегда соответствует ожиданиям заказчика, а средний Cycle Time на витрину 50 дней
Разбор инцидентов затруднен	Храним только несколько последних порций данных ИС и не видим историю их изменения
Описание данных и их качество	Данные ИС плохо описаны, их качество хромает, источник меняется без предупреждения

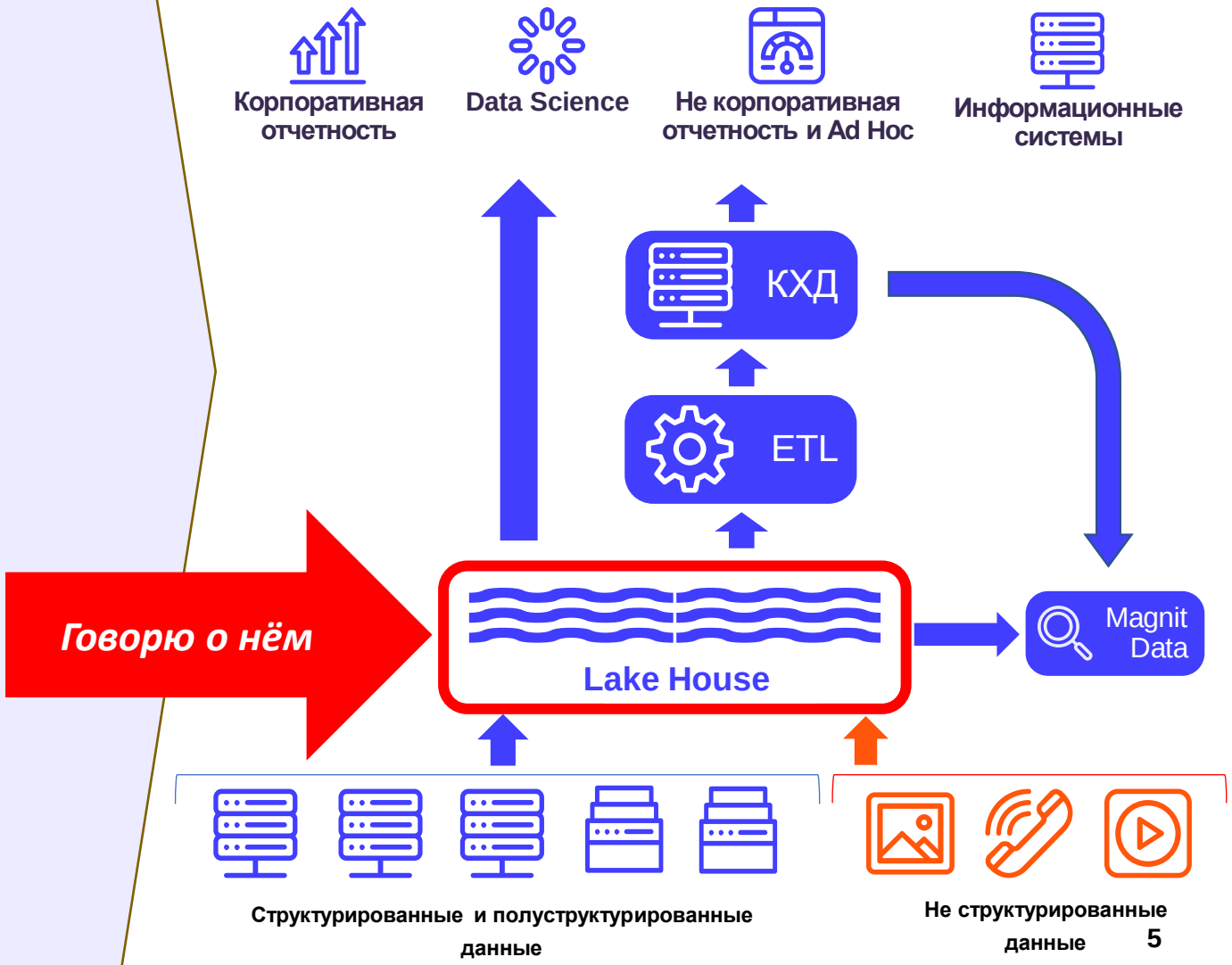
Концепт

1. Подключаем быстро
2. Унифицируем сценарии подключения
3. Унифицируем доступ к данным
4. Грузим данные под конкретные задачи
5. Описываем все загружаемые данные
6. Строим и публикуем Data Lineage

AS IS



TO BE



TECH

А сколько вообще бизнес-данных есть?

~60

Информационных систем, которые поставляют данные в текущую платформу

>36 000

Объектов всего

>9 000

Объектов, содержащих бизнес-данные

>100 000

Атрибутов объектов, содержащих бизнес-данные

200+

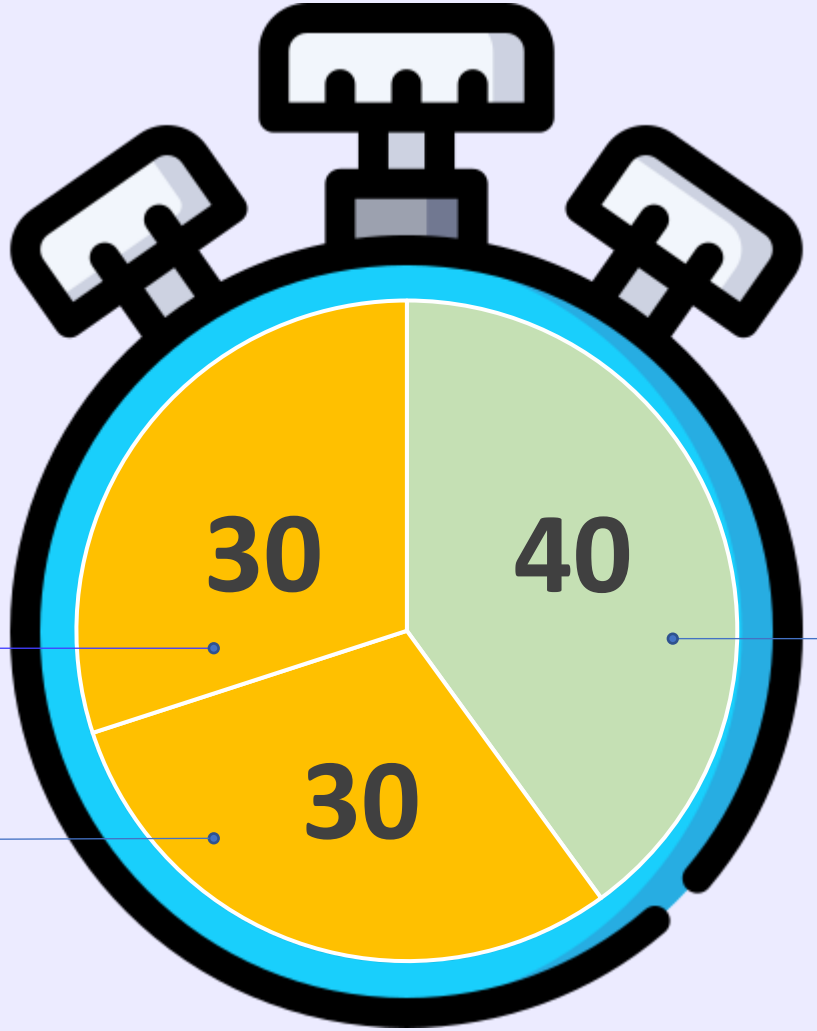
Информационных систем, которые производят бизнес-данные

~30%

Файловые интеграции

TECH

Время на интеграцию в конце 2024



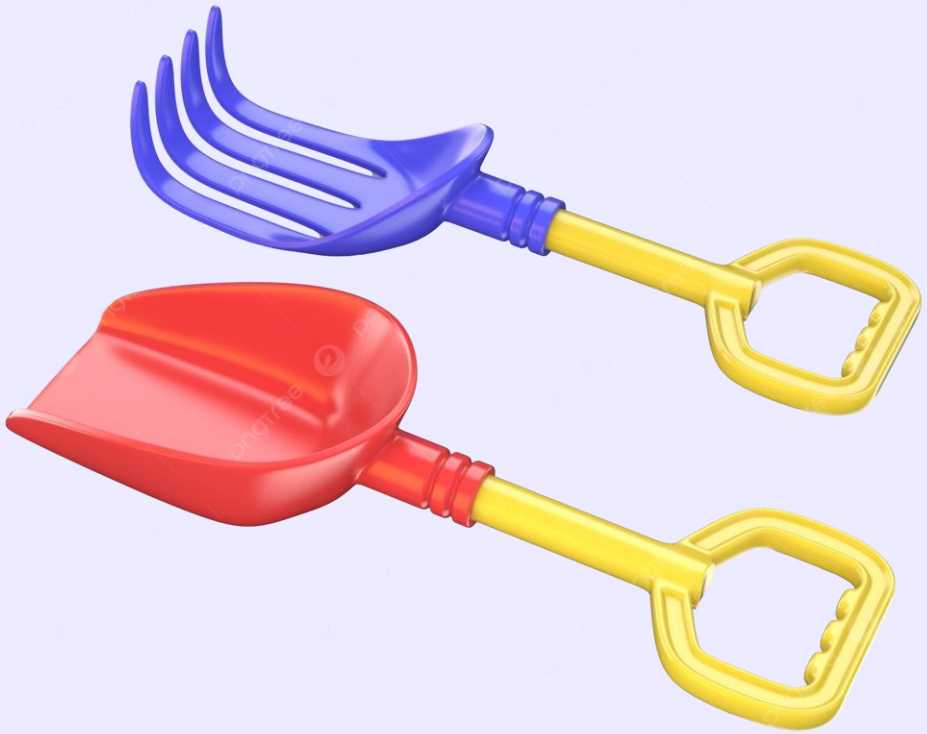
Настройка потока на стороне Lake House

Сверка данных в Lake House с источником

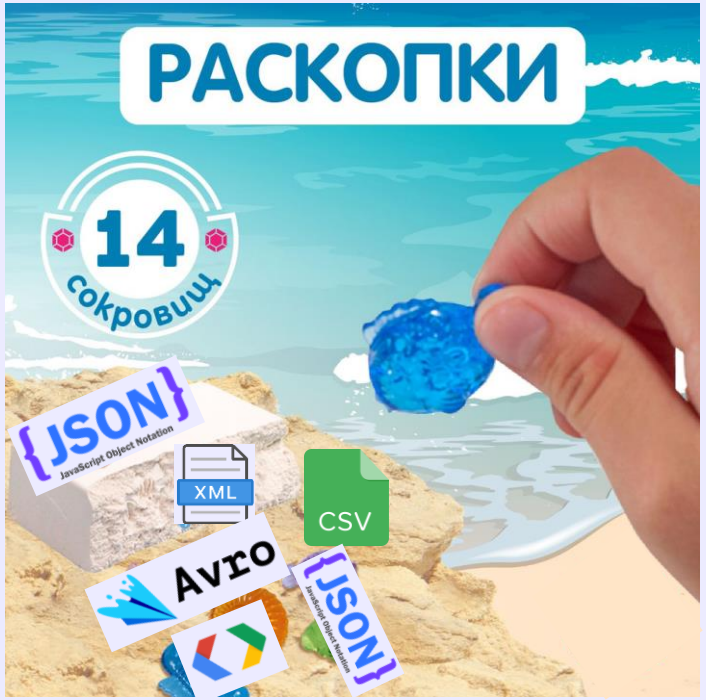
Работы на стороне ИС:

1. Описание
2. Категорирование
3. Определение Владельцев данных
4. Открытие доступов
5. Подготовка источника / выгрузки

Как сократить работы на стороне LH?



Дать Self–Service



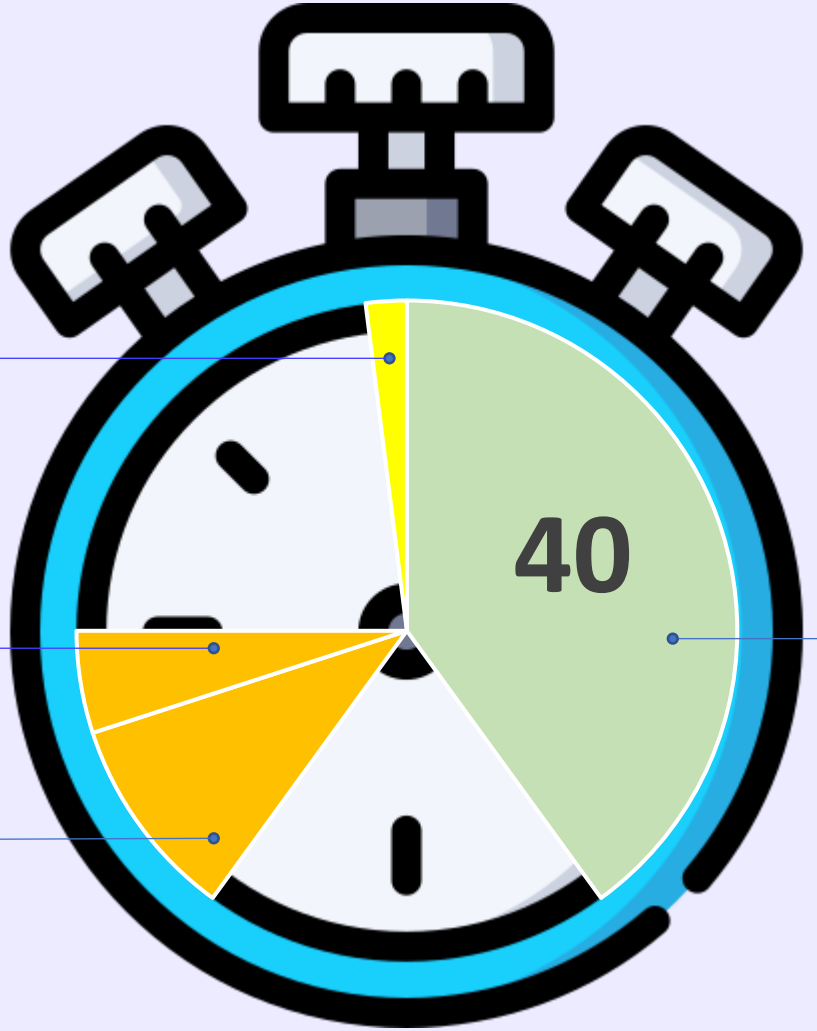
Дать доступ к файлам

Время на интеграцию с SS

Публикация файловых данных в сыром виде

Настройка потока на стороне Lake House

Сверка данных в Lake House с источником



Работы на стороне ИС:

1. Описание
2. Категорирование
3. Определение Владельцев данных
4. Открытие доступов
5. Подготовка источника / выгрузки

TECH

Что сделали?

Сейчас		*В конце Q2*	
50+	→	**80+**	Подключили ИС
500+	→	**1200+**	Объектов загрузили и обновляем
600ТБ	→	**1,5ПБ**	Данных загрузили

Проблемы

Организационные

1. Содержание проекта
2. Новый процесс, много участников
3. Ответственность сторон
4. Определение «правильных» источников данных

Технические

1. Производительность
2. «Грязные» данные в ИС
3. Особенности источников / приемников

Кейсы использования

№1 **Отчетность**

Товародвижение использует данные для построения отчётности о доступности товаров на полках магазинов, товарном запасе и уровне сервиса

№2 **Аналитика**

Онлайн использует данные о чеках для выявления злоумышленников в начислении бонусов среди пользователей программы лояльности

№3 **Бизнес-процессы**

Финансы выполняют сверку данных LH с данными из ФНС и, в случае выявления расхождений, формируют новые чеки - чеки «коррекции», тем самым выполняя требование 54ФЗ

Магазины берут из LH фотографии товаров, которые клиенты Магнита видят на кассах самообслуживания и весах

TECH

Планы развития

Коннекторы для новых типов источников (в том числе делаем свой загрузчик)

Интерфейсы доступа к данным в Lake House

Data Contract и контроль за их исполнением

Хранение неструктурированных данных

и т.д.

Спасибо за внимание!

DATA