ДЕПАРТАМЕНТ ОБРАЗОВАНИЯ И НАУКИ ГОРОДА МОСКВЫ
Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»
(ГАОУ ВО МГПУ)
Институт цифрового образования
Департамент информатики, управления и технологий

Лабораторная работа № 3.1

по дисциплине «Платформы Data Engineering»

Выполнил:
студент группы БД251м

Направление подготовки/Специальность
38.04.05 - Бизнес-информатика
St_62
(Ф.И.О.)

Проверил:
Кандидат технических наук, доцент
(ученая степень, звание)
Босенко Тимур Муртазович
(Ф.И.О.)

Москва 2025

Введение

Цель работы — пройти полный цикл data-driven исследования: от формулировки вопросов и сбора данных до построения интерактивного дашборда в Yandex DataLens и получения интерпретируемых выводов по теме “Хранилища данных: современные подходы и архитектуры”.

Процесс разработки

Данные и сбор: Инструмент опроса: Google Forms

Исследовательские вопросы:

- Какие архитектуры DWH (On-prem, Cloud, Hybrid) встречаются и как они связаны с подходом интеграции (ETL/ELT/смешанный)?
- Как размещение и архитектура соотносятся с латентностью обновления данных?
- Какие субъективные оценки (производительность, масштабируемость, стоимость, простота сопровождения) получают текущие решения?
- Какие инструменты/практики применяются и какие планы развития у команд?

Ответов: 7
Примечание: опрос искусственный, заполнен одним респондентом 7 раз для демонстрации полного цикла — от опроса до дашборда
Причины: ограничение по доступу к респондентам и дедлайн
Экспорт: CSV → предобработка в Excel → загрузка в DataLens (xlsx)

Создание чартов:

Линейчатая диаграм...

Y

architecture

X

respondents_count

Цвета

architecture

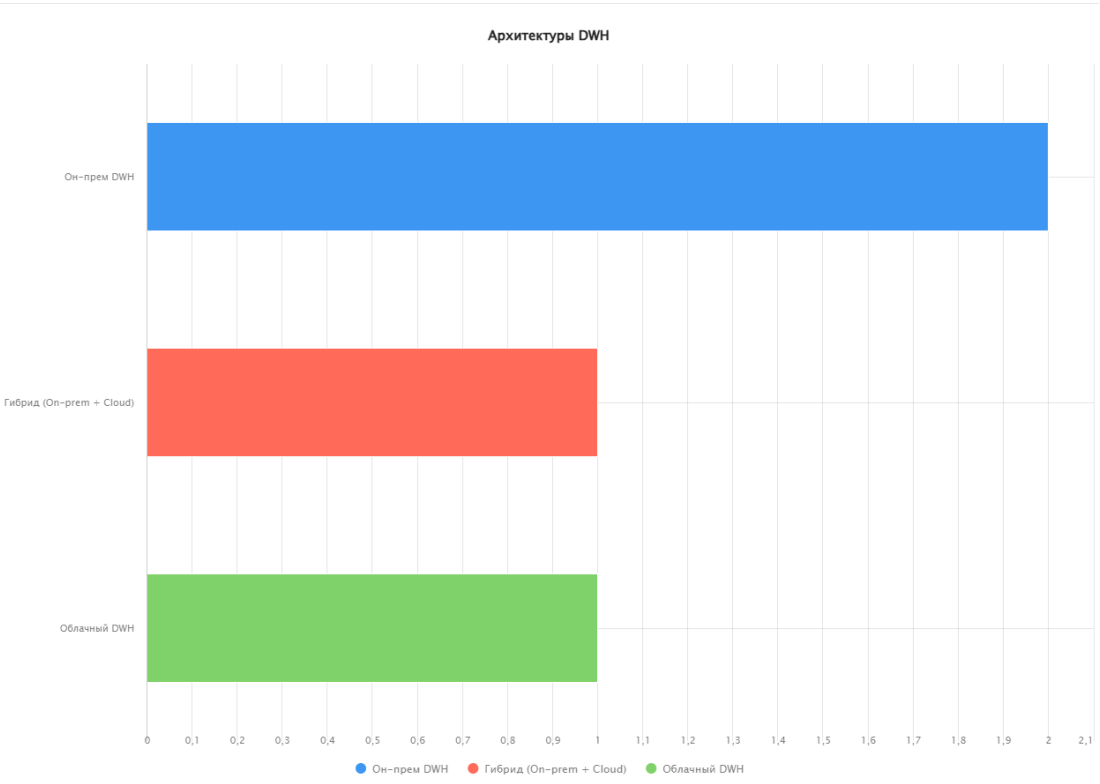
Сортировка

Order

Подписи

Фильтры

is_study: 0



Интеграции по типу размещения

На весь экран

Сохранено

Нормированная стол...

X

deployment

Y

respondents_count

Цвета

integration

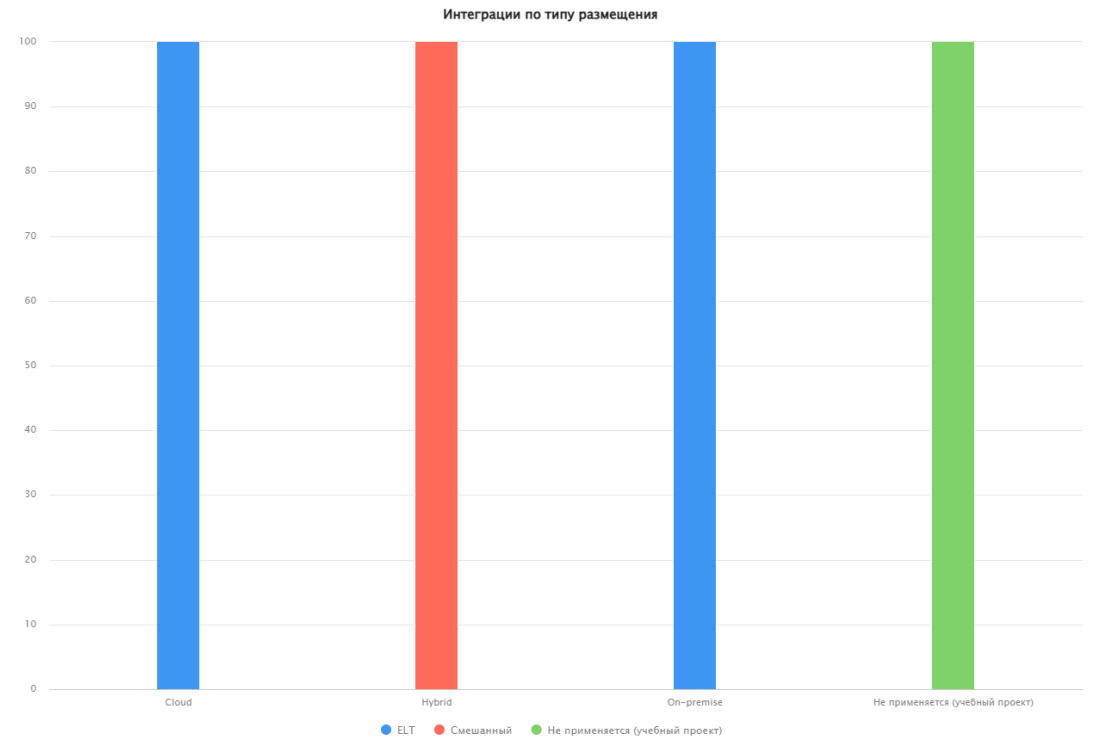
Сортировка

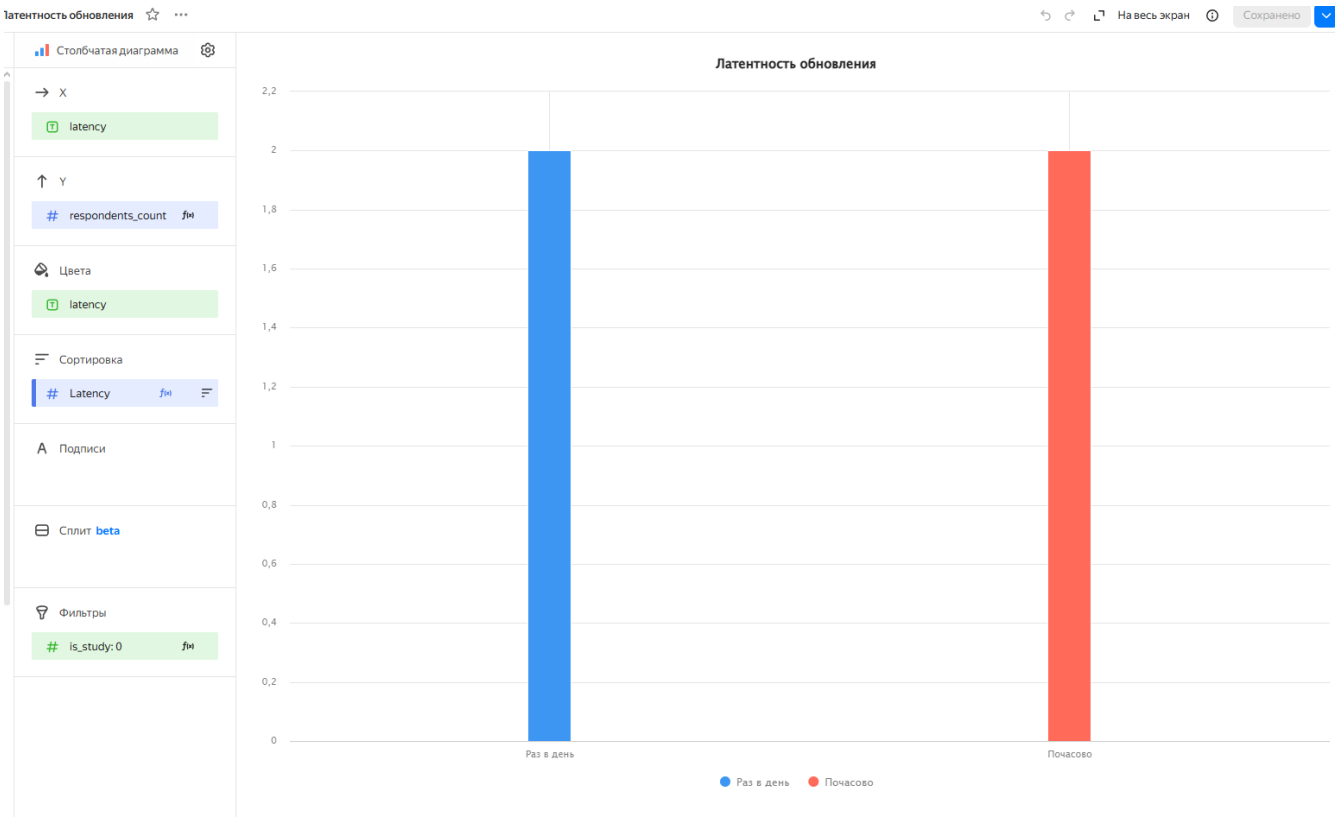
Deployment

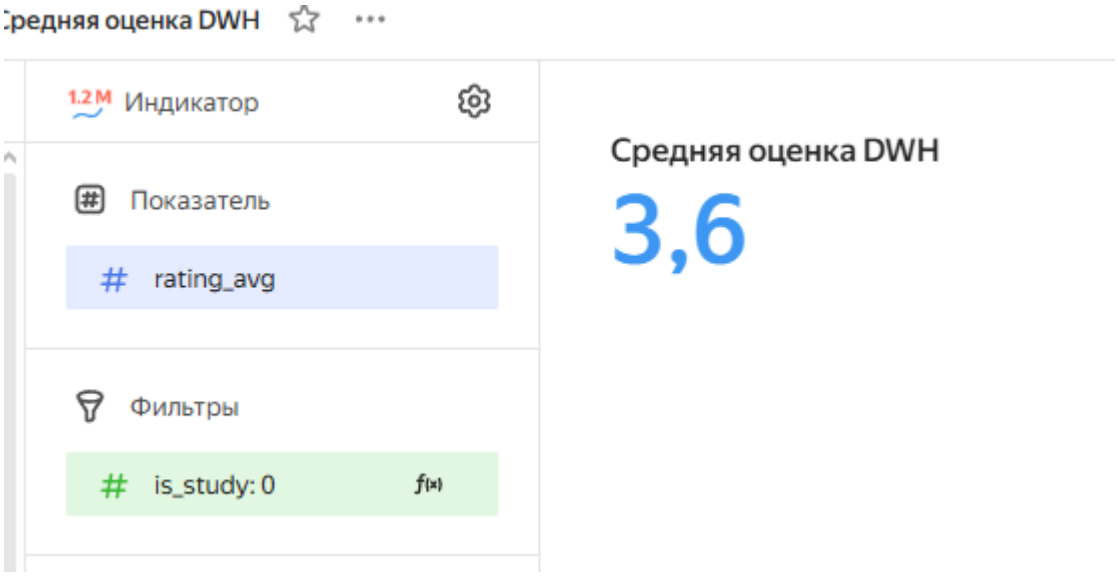
Подписи

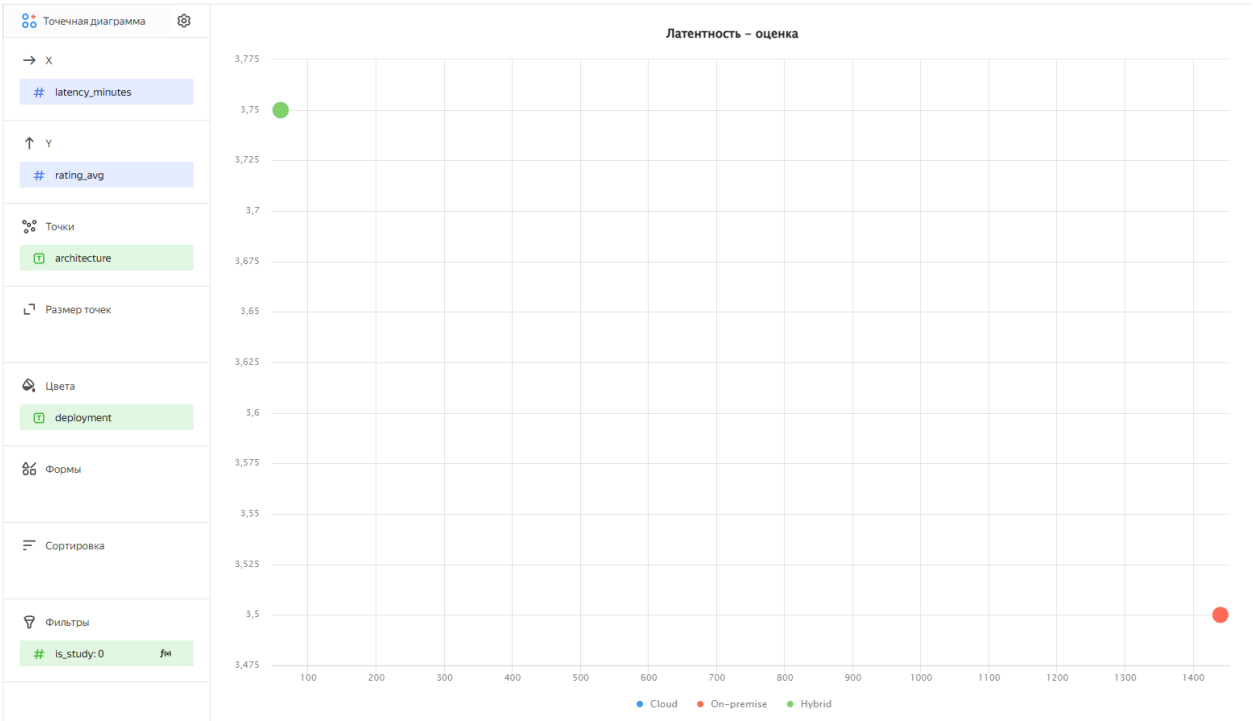
Сплит beta

Фильтры

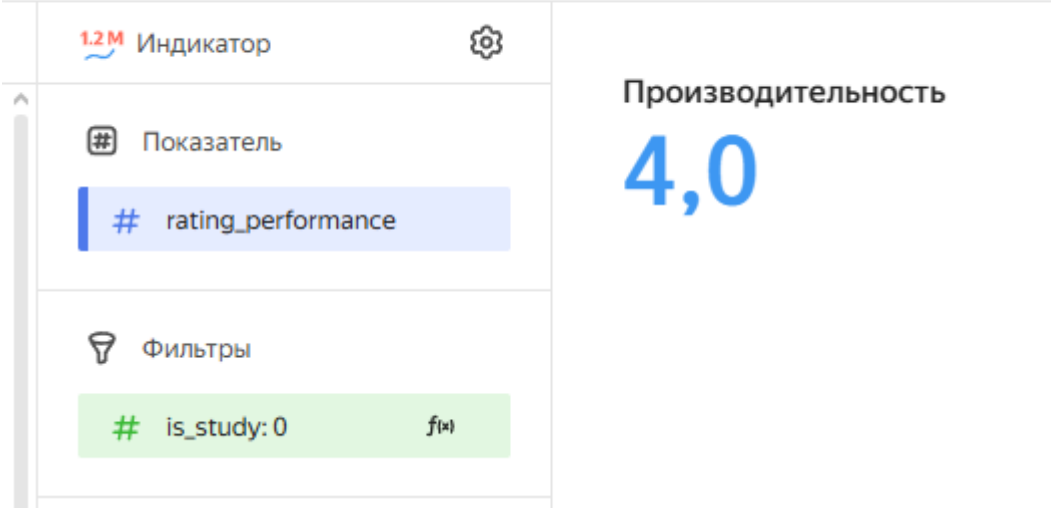



Латентность обновления
На весь экран
Сохранено
Столбчатая диаграмма
X
latency
Y
respondents_count
Цвета
latency
Сортировка
Latency
Подписи
Сплит beta
Фильтры
is_study: 0
Латентность обновления
2,2
2
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
Раз в день
Почасово
Раз в день
Почасово

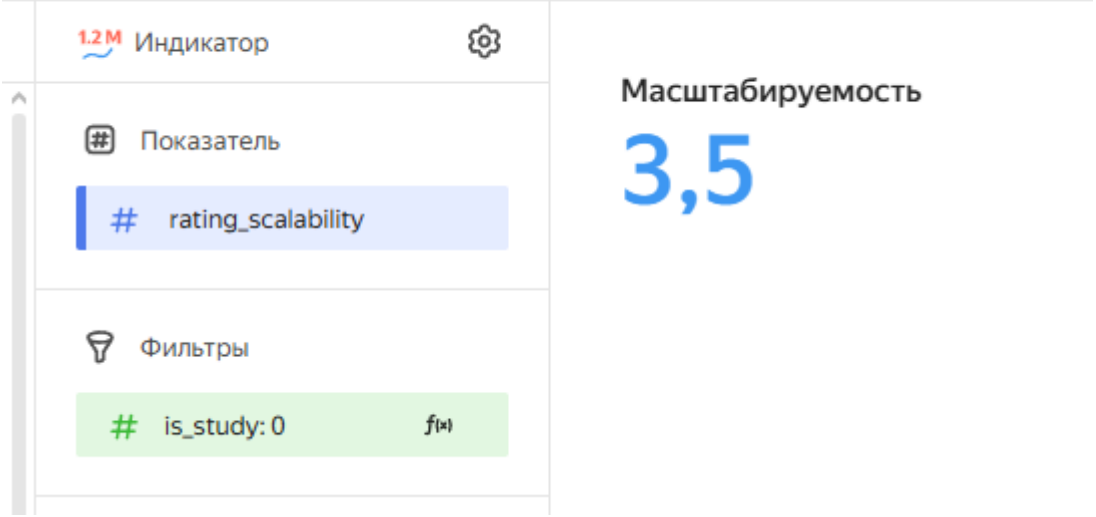

Средняя оценка DWH
1.2 M
Индикатор
Показатель
rating_avg
Фильтры
is_study: 0
Средняя оценка DWH
3,6

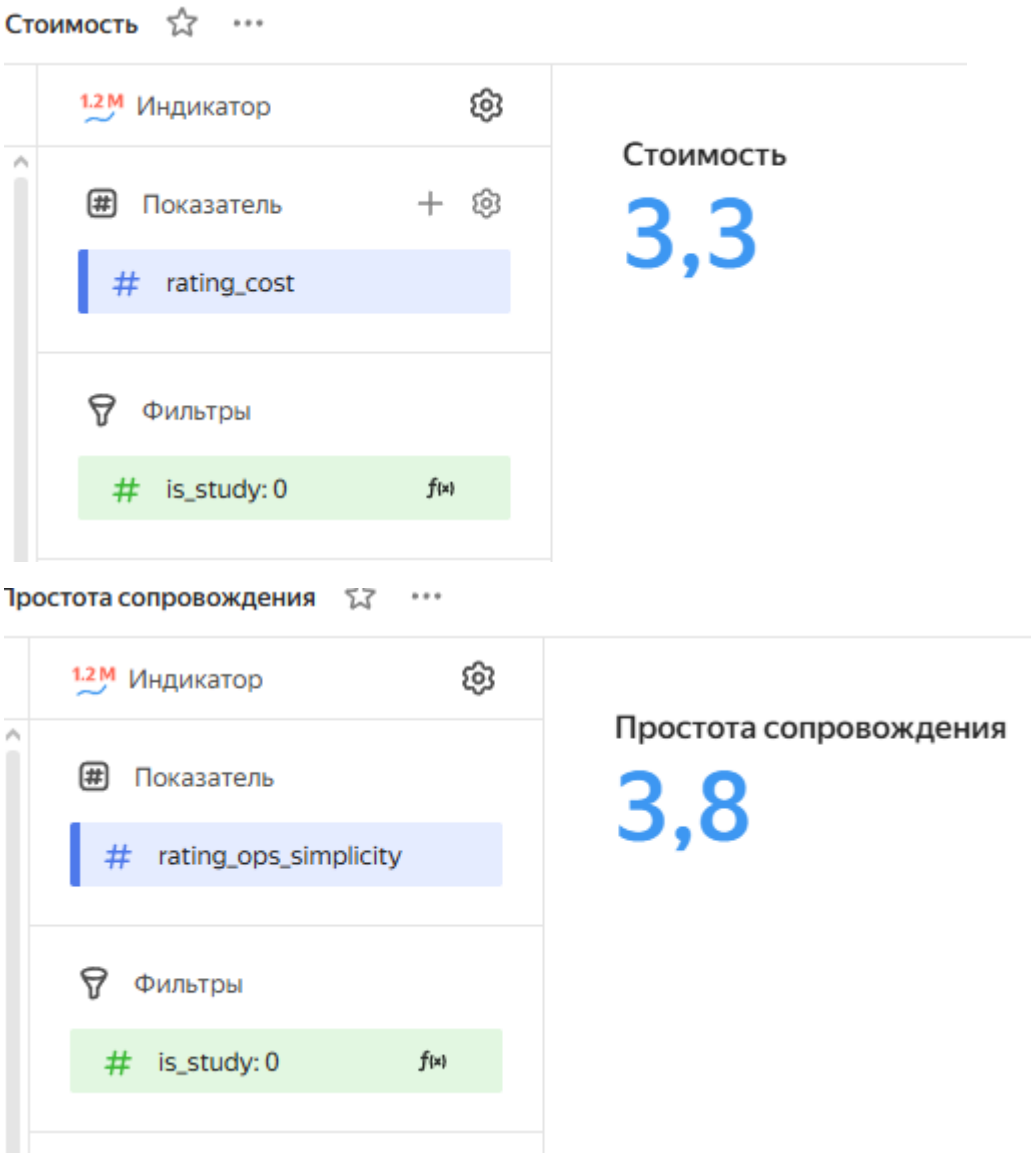


Производительность



Масштабируемость





Анализ результатов

- Архитектуры: в выборке представлены On-prem (2 ответа, MS SQL Server), Cloud (1, BigQuery) и Hybrid (1, Snowflake), On-prem встречается чуть чаще
- Интеграции: ELT является основным подходом (Cloud и On-prem), для Hybrid используется смешанный (ELT + доп. трансформации)
- Латентность: Cloud/Hybrid - “Почасово”; On-prem - “Раз в день”
- Оценки: средняя по системе ≈ 3.6, из метрик выше всего оценки “Простота сопровождения” (~3.8) и “Производительность” (~4.0), ниже — “Масштабируемость” (~3.5) и “Стоимость” (~3.3)
- Инструменты/практики/планы: Airflow используют все “рабочие” ответы, встречаются dbt/скрипты (on-prem), Fivetran/Stitch (облако), Spark/Dataproc (гибрид). В планах доминирует “Внедрение Data Catalog/Lineage” , оптимизация стоимости и переход к облаку/Lakehouse

Ограничения:

- Малый объем выборки (n=7), часть ответов — учебные, данные синтетические; результат не репрезентативен
- Оценки субъективны

Заключение

Поставленная цель достигнута: получены и нормализованы данные, построен датасет и дашборд в Yandex DataLens, сформулированы выводы. Несмотря на условность данных видно, что ELT в облаке/гибриде, меньшая латентность в Cloud/Hybrid, более высокие оценки производительности и простоты сопровождения.