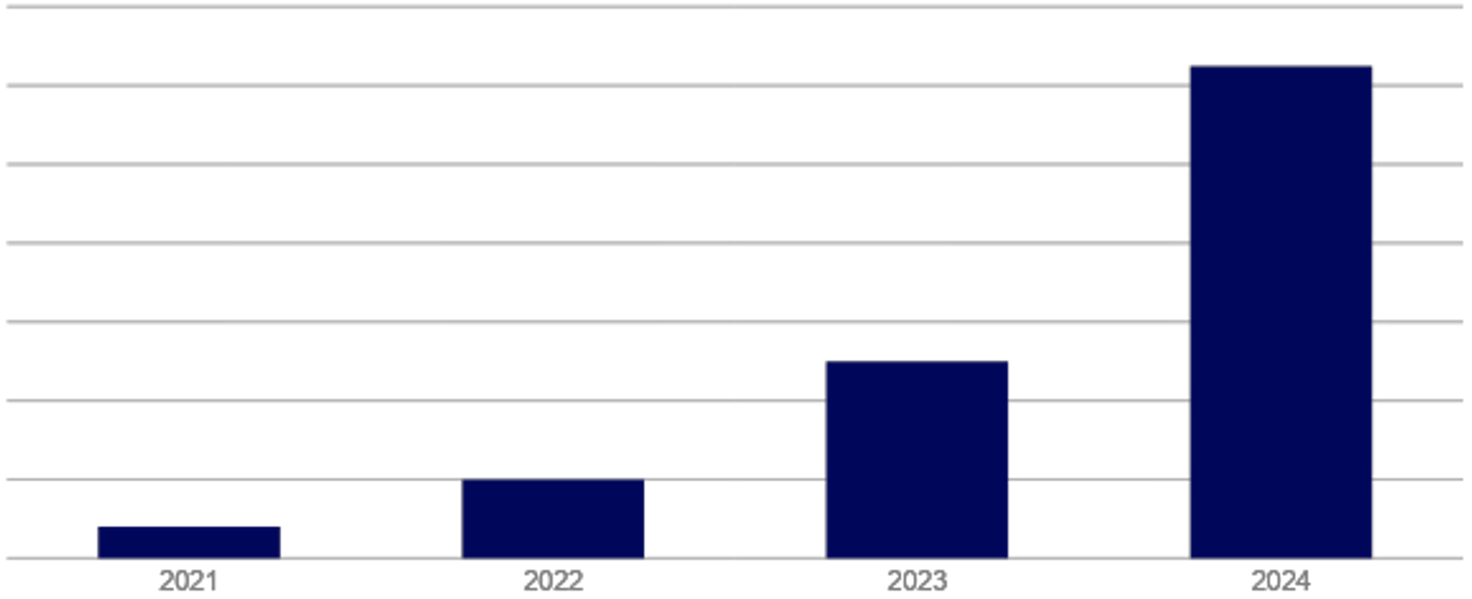
avito.tech Москва 2024

Aqueduct – open source для экономии железа в МЛ-инференсе

БУГРИМОВ ОЛЕГ Тимлид разработки Data Science

Растет потребность железа для МЛ

График роста потребности в оборудовании для МЛ



Потребность растет

Новые фреймворки и алгоритмы для МЛ-моделей всё более требовательны к потреблению железа

МЛ-сервера

Сервера включает от одного gpu-устройства. Они дороже и их тяжелее закупать

С акведуком удается уменьшить рост потребности железа

Акведук

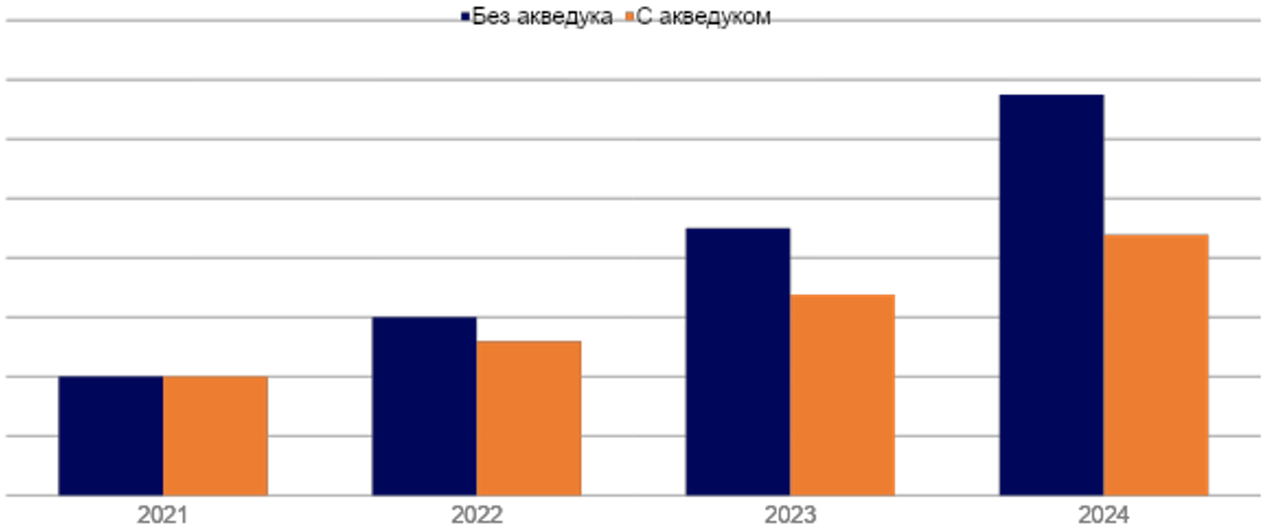
Позволяет уменьшить рост потребности в железе

Потребность растет

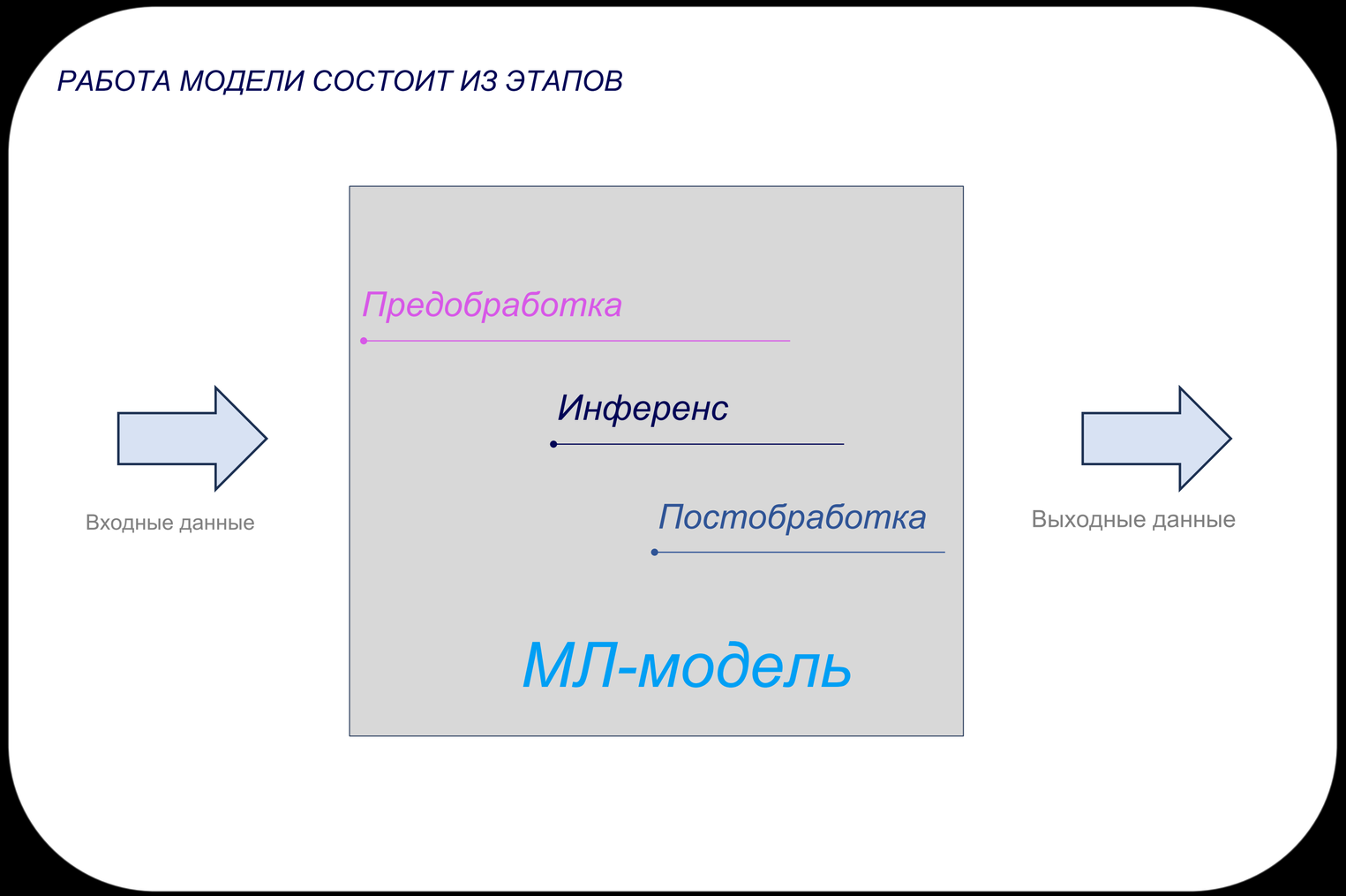
Новые фреймворки и алгоритмы для МЛ-моделей все более требовательны к потреблению железа

МЛ-сервера

Сервер включает от одного gpu-устройства. Они дороже и их тяжелее закупать



Как работает акведук



Каждая МЛ-модель — это последовательность шагов вычислений.

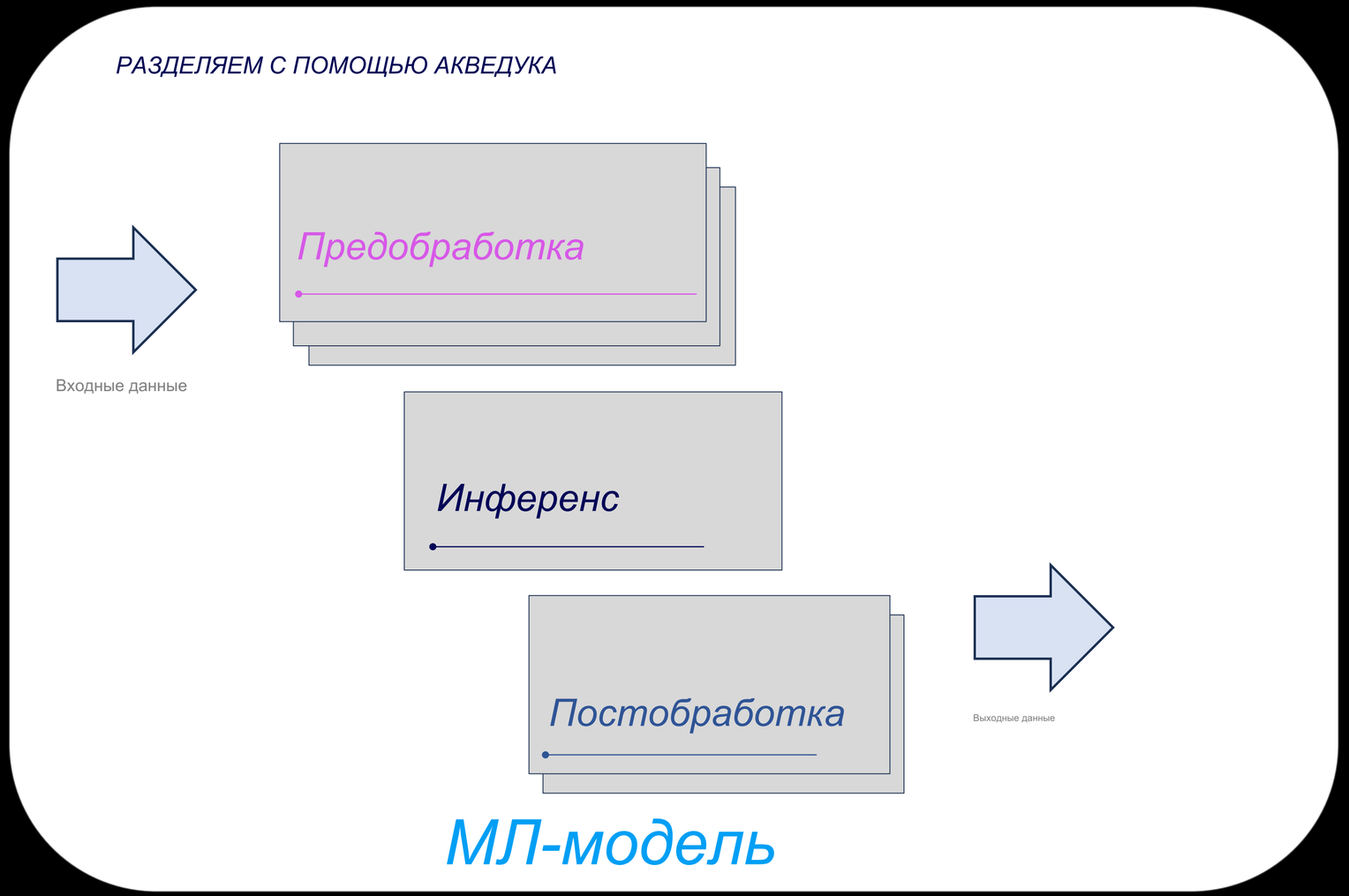
Обычно это три шага:

1)подготовка данных для отправки их в модель,

1)сам инференс, то есть работа уже самой модели,

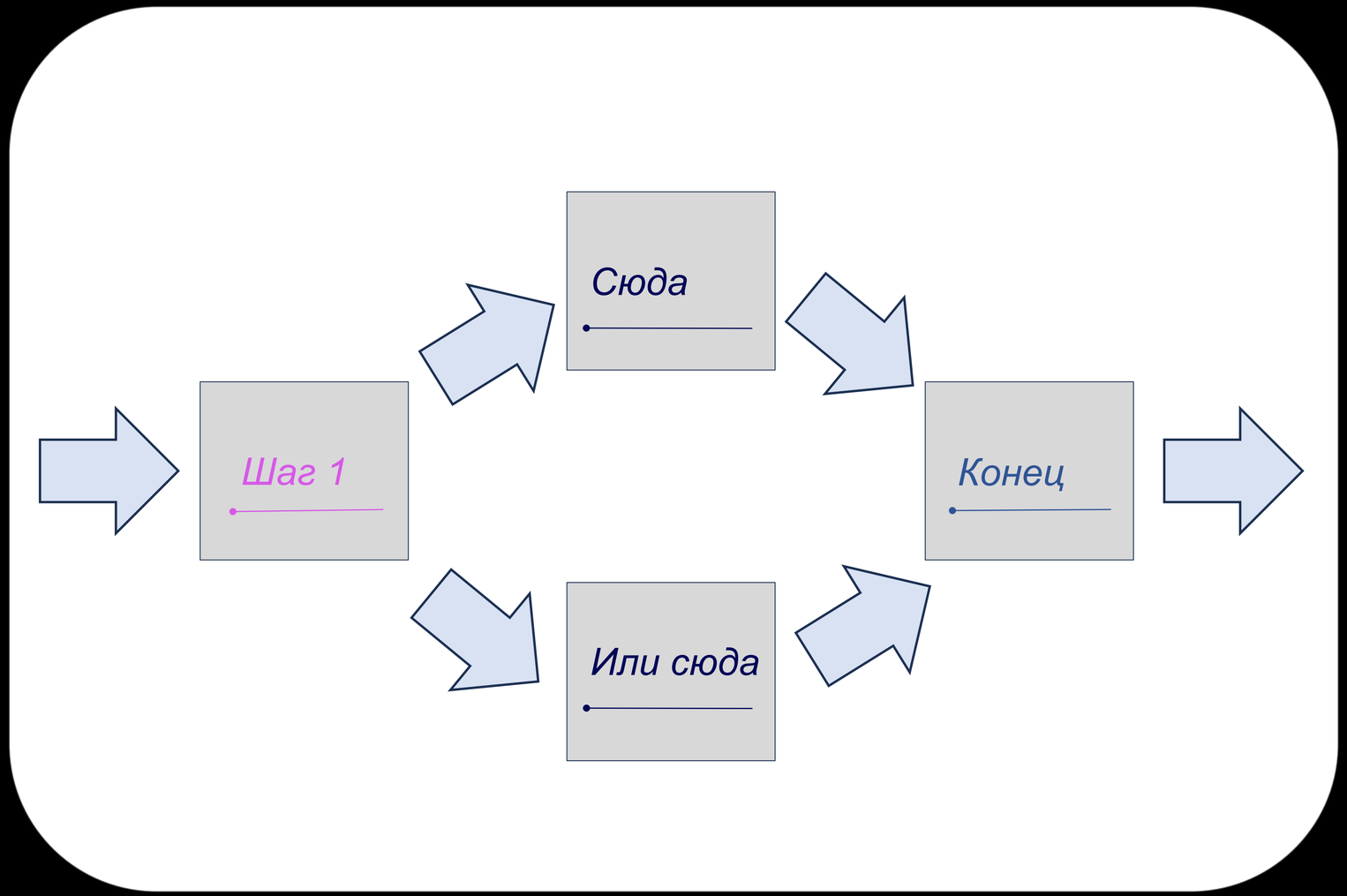
1) постобработка полученных данных.

Как работает акведук



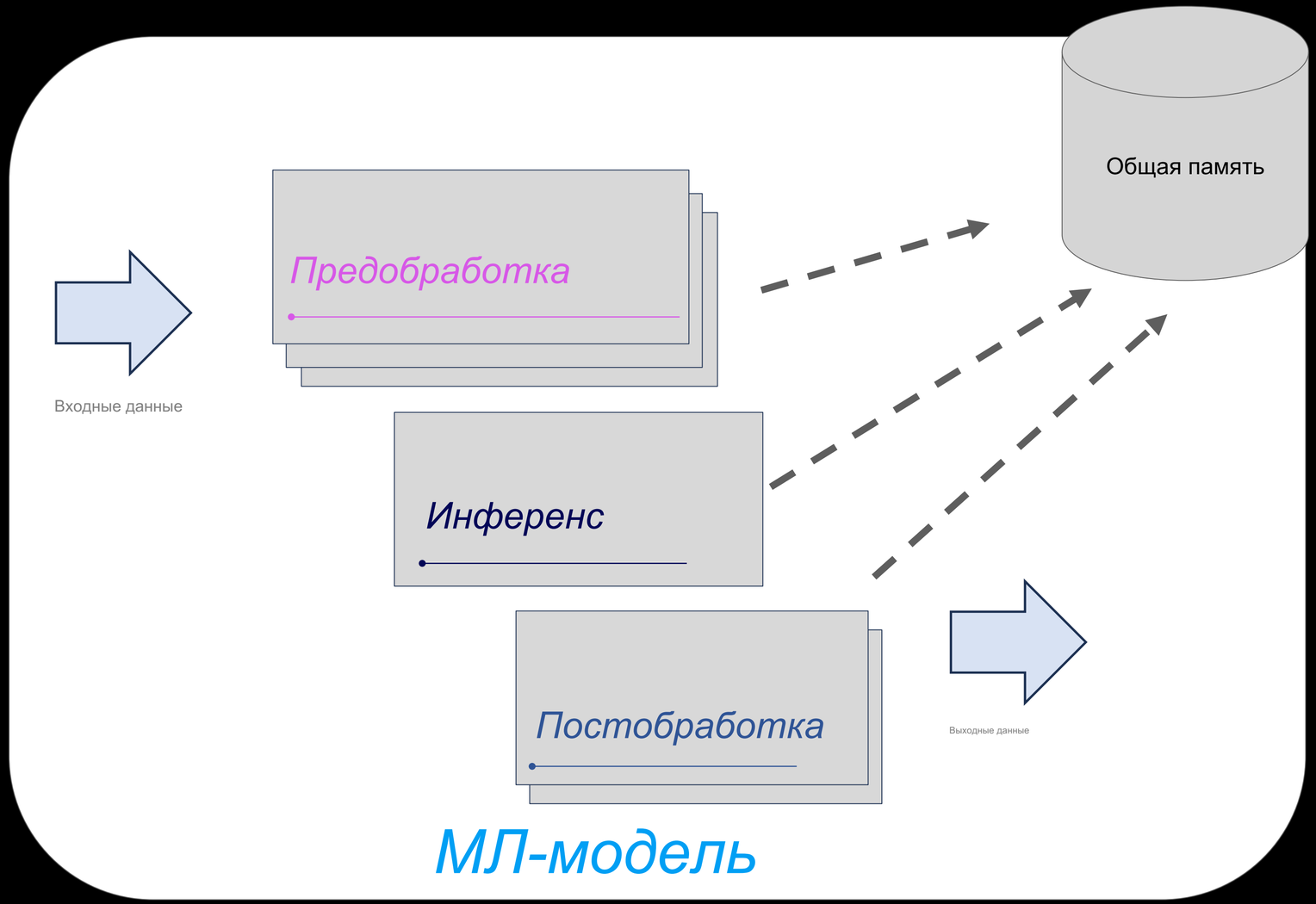
С помощью акведука можно разделить такую логику на разные этапы и сделать так, чтобы они выполнялись в разных процессах.

Условный переход между шагами



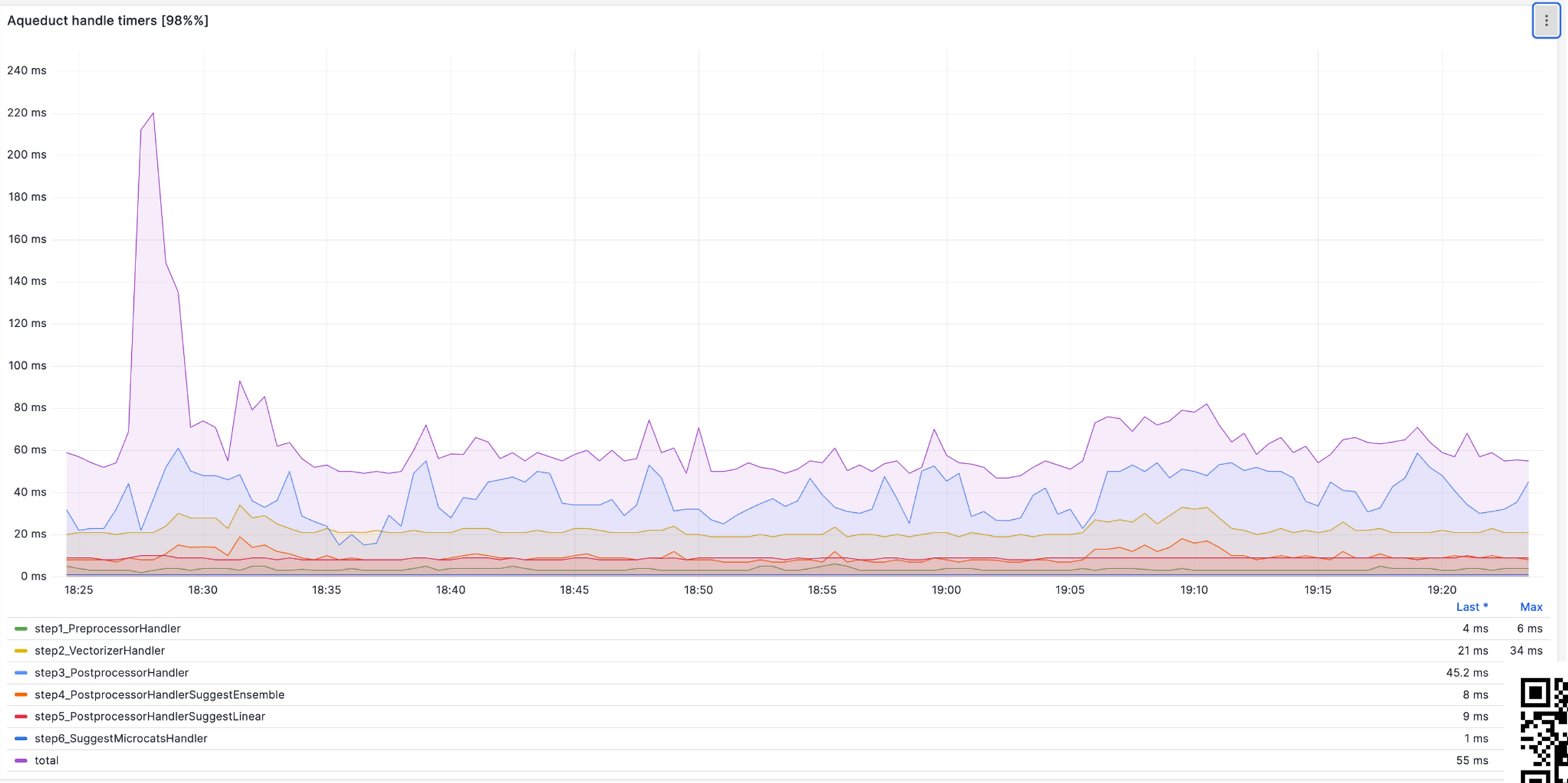
С помощью акведука можно реализовать условный переход между шагами

Шаренная память

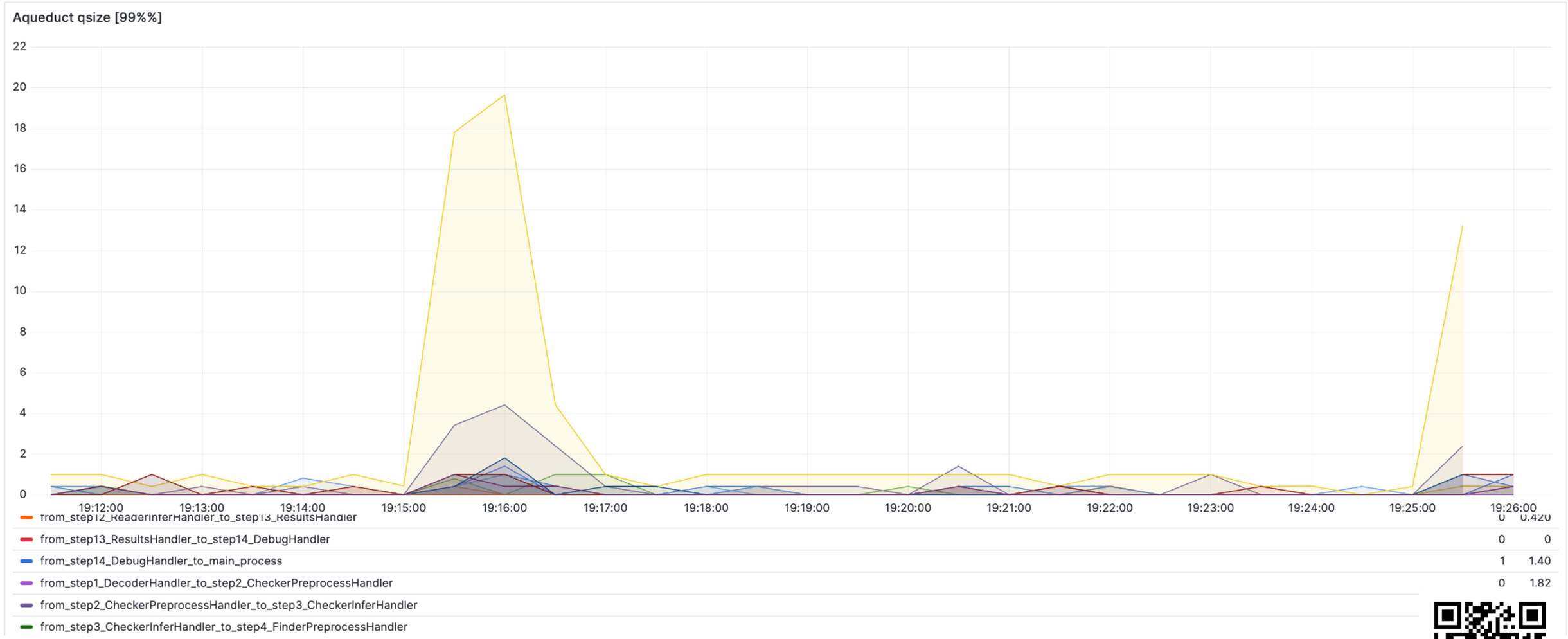


С помощью акведука можно реализовать использование общей памяти

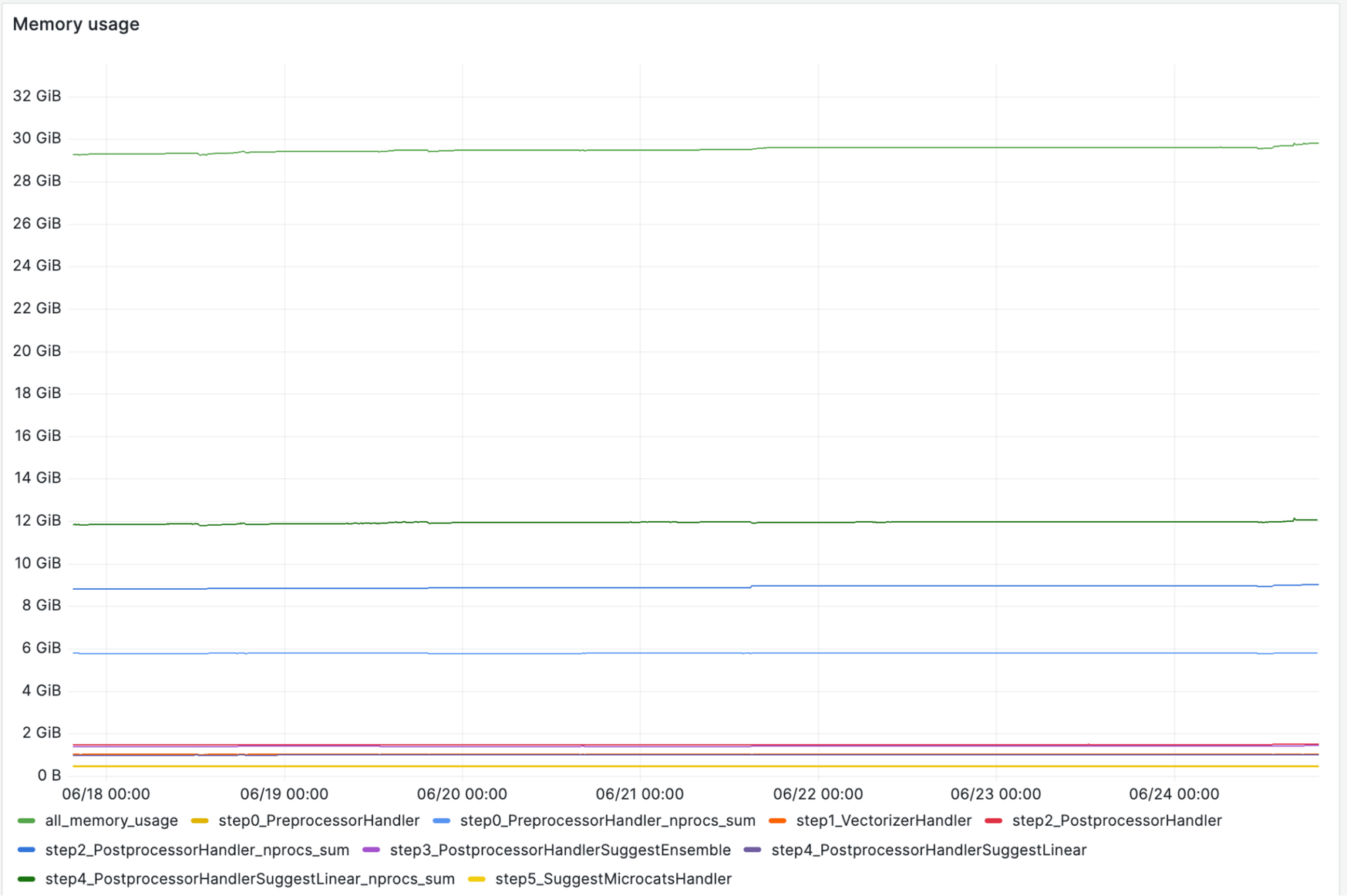
Метрики


Aqueduct handle timers [98%%]
240 ms
220 ms
200 ms
180 ms
160 ms
140 ms
120 ms
100 ms
80 ms
60 ms
40 ms
20 ms
0 ms
18:25
18:30
18:35
18:40
18:45
18:50
18:55
19:00
19:05
19:10
19:15
19:20
Last *
Max
step1_PreprocessorHandler 4 ms 6 ms
step2_VectorizerHandler 21 ms 34 ms
step3_PostprocessorHandler 45.2 ms
step4_PostprocessorHandlerSuggestEnsemble 8 ms
step5_PostprocessorHandlerSuggestLinear 9 ms
step6_SuggestMicrocatsHandler 1 ms
total 55 ms

Метрики


Aqueduct qsize [99%%]
22
20
18
16
14
12
10
8
6
4
2
0
19:12:00
19:13:00
19:14:00
19:15:00
19:16:00
19:17:00
19:18:00
19:19:00
19:20:00
19:21:00
19:22:00
19:23:00
19:24:00
19:25:00
19:26:00
from_step12_ReaderInferHandler_to_step13_ResultsHandler 0 0.420
from_step13_ResultsHandler_to_step14_DebugHandler 0 0
from_step14_DebugHandler_to_main_process 1 1.40
from_step1_DecoderHandler_to_step2_CheckerPreprocessHandler 0 1.82
from_step2_CheckerPreprocessHandler_to_step3_CheckerInferHandler
from_step3_CheckerInferHandler_to_step4_FinderPreprocessHandler

Метрики


Memory usage
32 GiB
30 GiB
28 GiB
26 GiB
24 GiB
22 GiB
20 GiB
18 GiB
16 GiB
14 GiB
12 GiB
10 GiB
8 GiB
6 GiB
4 GiB
2 GiB
0 B
06/18 00:00
06/19 00:00
06/20 00:00
06/21 00:00
06/22 00:00
06/23 00:00
06/24 00:00
all_memory_usage
step0_PreprocessorHandler
step0_PreprocessorHandler_nprocs_sum
step1_VectorizerHandler
step2_PostprocessorHandler
step2_PostprocessorHandler_nprocs_sum
step3_PostprocessorHandlerSuggestEnsemble
step4_PostprocessorHandlerSuggestLinear
step4_PostprocessorHandlerSuggestLinear_nprocs_sum
step5_SuggestMicrocatsHandler

Акведук — универсальное решение

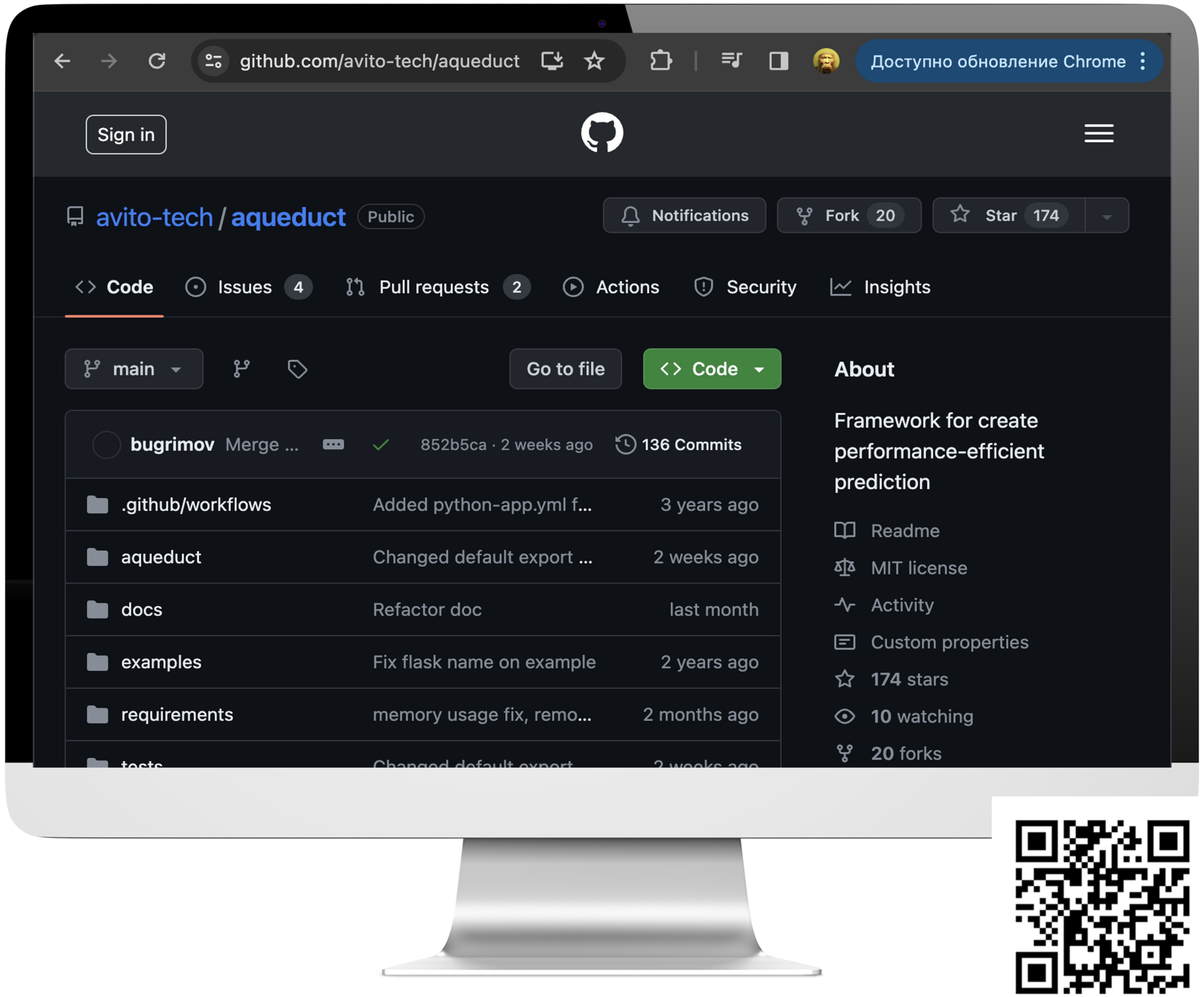
Работает для всех типов моделей.
Доступно на github, бесплатно под лицензией MIT.
Есть внедрения в других компаниях

Преимущества от выкладывания в Open Source

+ изменения/фиксы, которые мы не планировали\\ не замечали,
+ изменения от других команд внутри компании,
+ такая разработка заметно мотивирует команду
+ используют решение в своих продуктах

Недостатки от выкладывания в Open Source

- Нужно поддерживать решение, отделенное от внутренних костылей
- Больше ответственности на внесение изменений
- Больше задач на документацию



avito.tech Москва 2024

Aqueduct – open source для экономии железа в МЛ-инференсе

БУГРИМОВ ОЛЕГ Тимлид разработки Data Science