

Выберите
уровень
сложности
инференса

ОБЫЧНЫЙ



СЛОЖНЫЙ



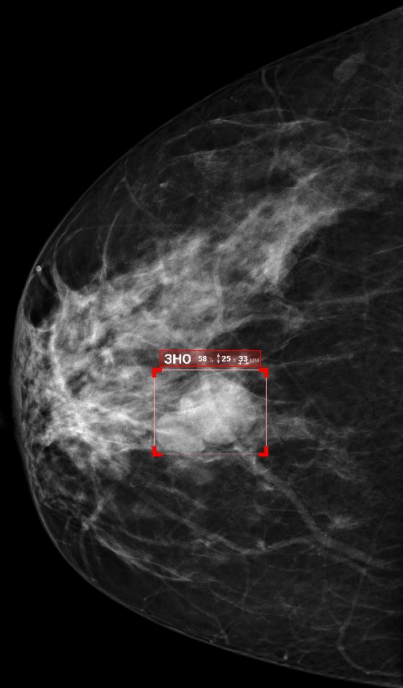
ТРЭШ



ТОЛЬКО В ИССЛЕДОВАТЕЛЬСКИХ ЦЕЛЯХ

ЦЕЛЕВАЯ ПАТОЛОГИЯ ОПРЕДЕЛЕНА CELSUS MMG®

R CC



4 Bi Rads

ACR A B C D PGMI P G M I

Версия модели 0.17.8
Качество изображения 99%

ТОЛЬКО В ИССЛЕДОВАТЕЛЬСКИХ ЦЕЛЯХ

ЦЕЛЕВАЯ ПАТОЛОГИЯ ОПРЕДЕЛЕНА CELSUS MMG®

R MLO



4 Bi Rads

ACR A B C D PGMI P G M I

Версия модели 0.17.8
Качество изображения 99%

Путешествие сквозь миры

Сегодня прыгнем в кроличью дыру медицинского ML-инференса:

- В облаках
- На удаленных защищенных bare-metal серверах
- На устройствах без доступа к интернету

Разберёмся, можно ли решить те проблемы, которые возникают вне прекрасного мира облаков

Критерии

Рассмотрим несколько важных категорий:

- **Ресурсы и масштабирование** - сколько у нас железа, легко ли получать новые, надо ли жёстко оптимизироваться

Критерии

Рассмотрим несколько важных категорий:

- **Ресурсы и масштабирование** - сколько у нас железа, легко ли получать новые, надо ли жёстко оптимизироваться
- **Обновление моделей** - легко ли деплоить новые версии моделей и небольшие фиксы

Критерии

Рассмотрим несколько важных категорий:

- **Ресурсы и масштабирование** - сколько у нас железа, легко ли получать новые, надо ли жёстко оптимизироваться
- **Обновление моделей** - легко ли деплоить новые версии моделей и небольшие фиксы
- **Железо и ОС** - есть ли особенности железа и софта

Критерии

Рассмотрим несколько важных категорий:

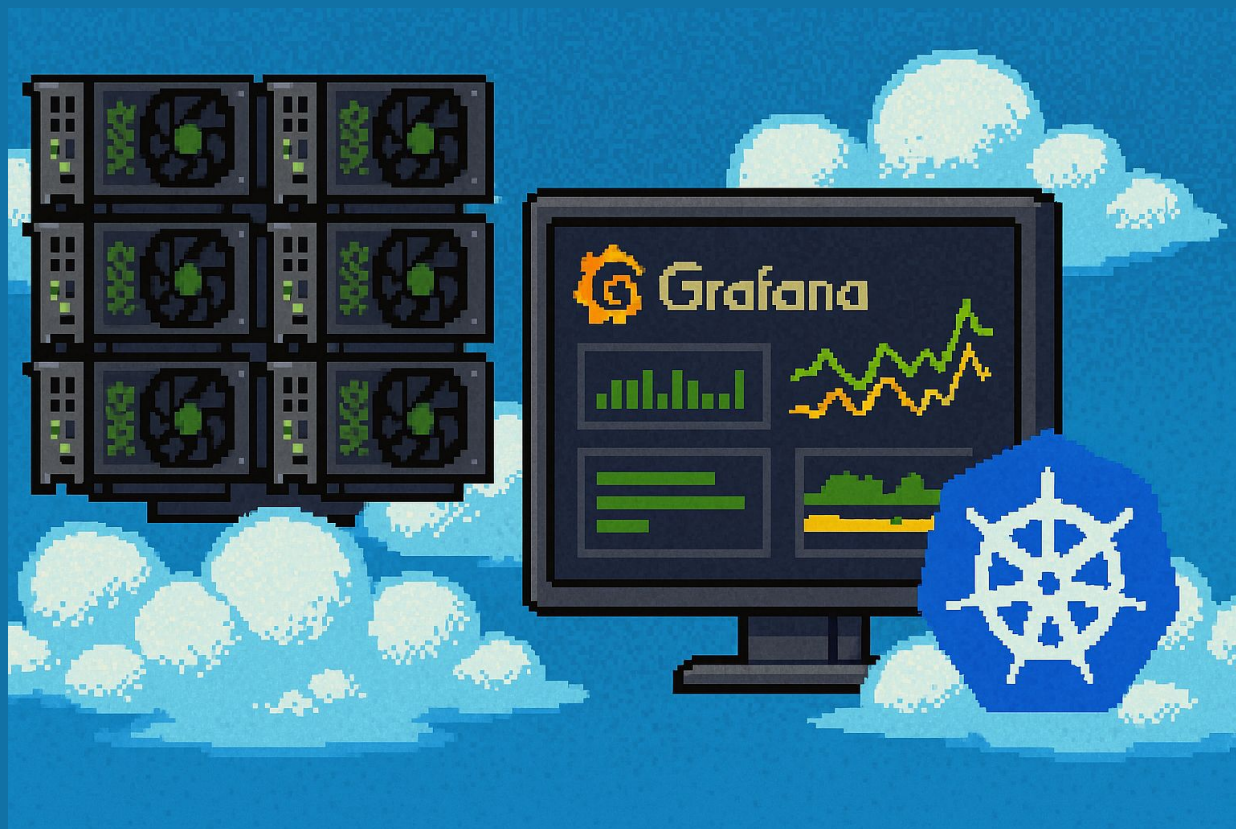
- **Ресурсы и масштабирование** - сколько у нас железа, легко ли получать новые, надо ли жёстко оптимизироваться
- **Обновление моделей** - легко ли деплоить новые версии моделей и небольшие фиксы
- **Железо и ОС** - есть ли особенности железа и софта
- **Мониторинг и алерты** - насколько прозрачна ML-система для её создателей

Критерии

Рассмотрим несколько важных категорий:

- **Ресурсы и масштабирование** - сколько у нас железа, легко ли получать новые, надо ли жёстко оптимизироваться
- **Обновление моделей** - легко ли деплоить новые версии моделей и небольшие фиксы
- **Железо и ОС** - есть ли особенности железа и софта
- **Мониторинг и алерты** - насколько прозрачна ML-система для её создателей
- **Входные данные** - какие особенности есть у входных данных?

Уровень 1



Часть клиентов живёт в облаках

- Москва
- Федеральная платформа МосМедИИ
- Часть регионов
- Часть коммерческих клиник

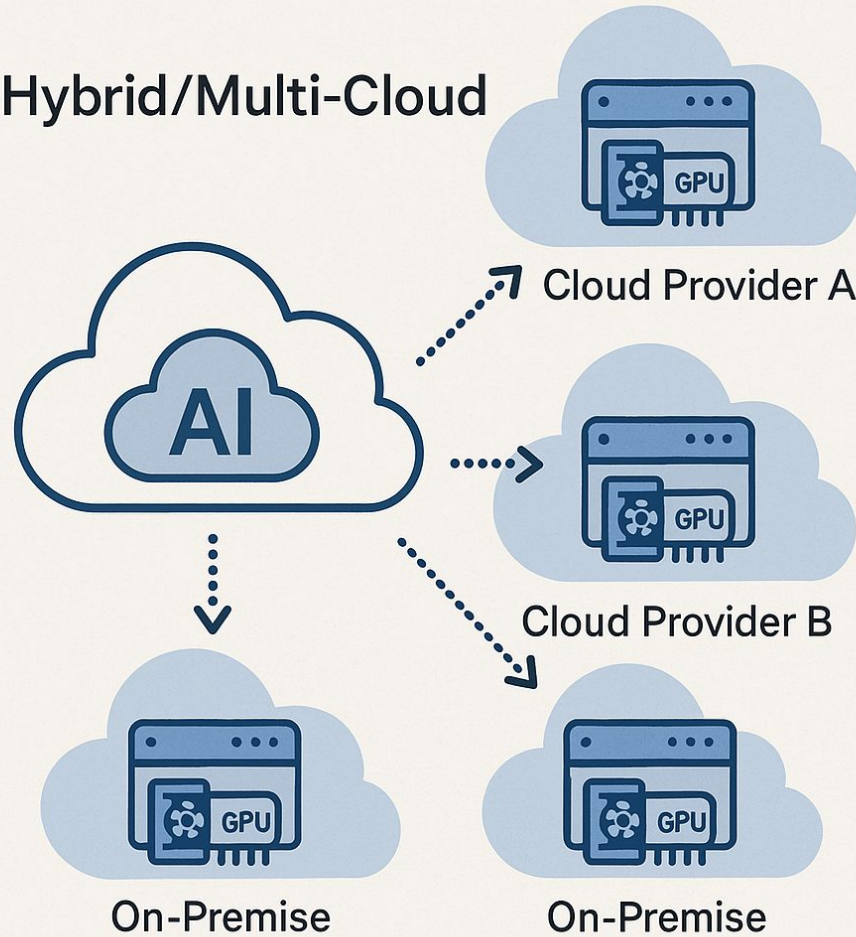
Давайте посмотрим на этот Эдем с точки зрения МЛ-инженера

Ресурсы и масштабирование

- Условно бесконечное масштабирование ресурсов

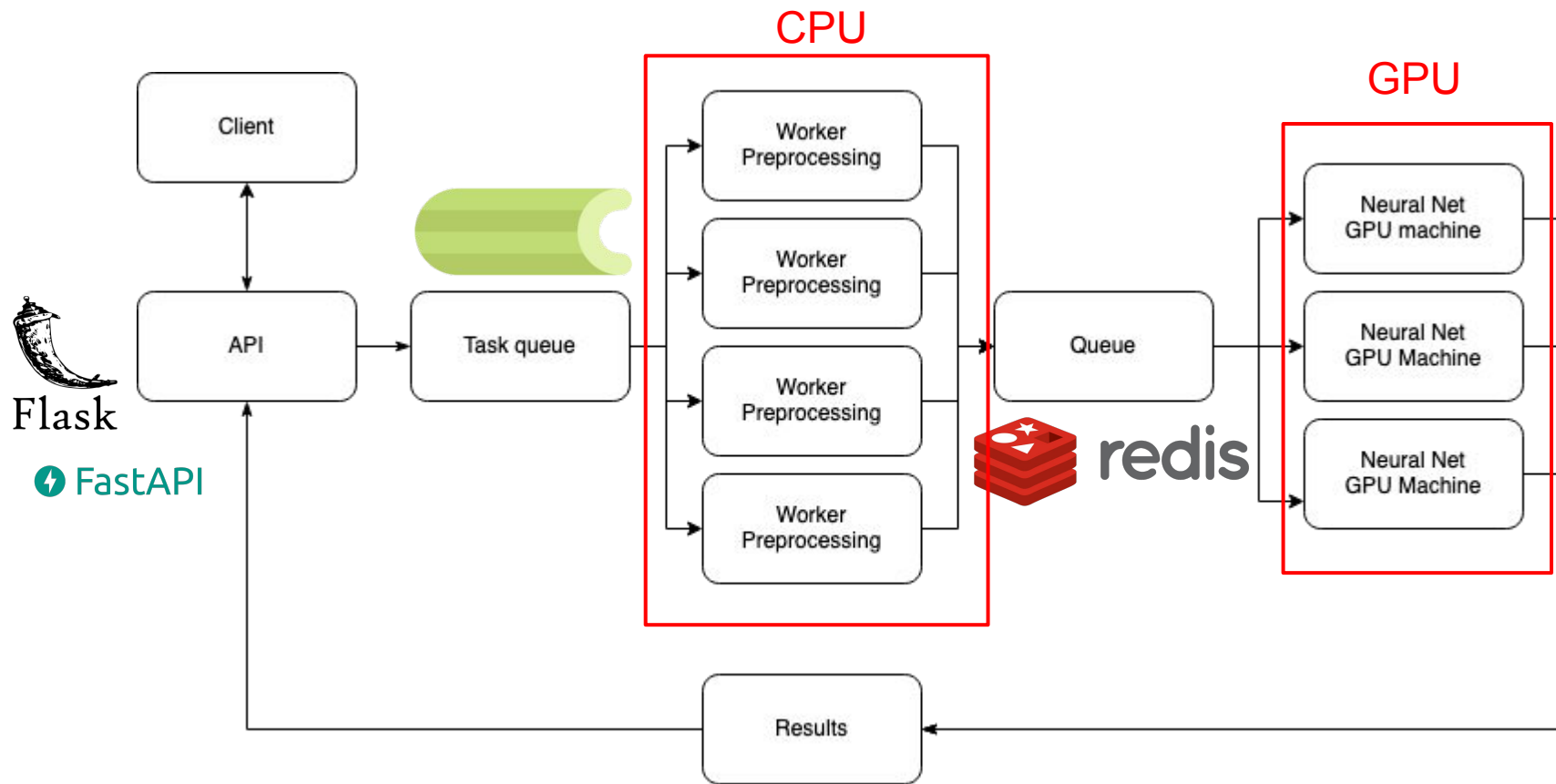


Hybrid/Multi-Cloud



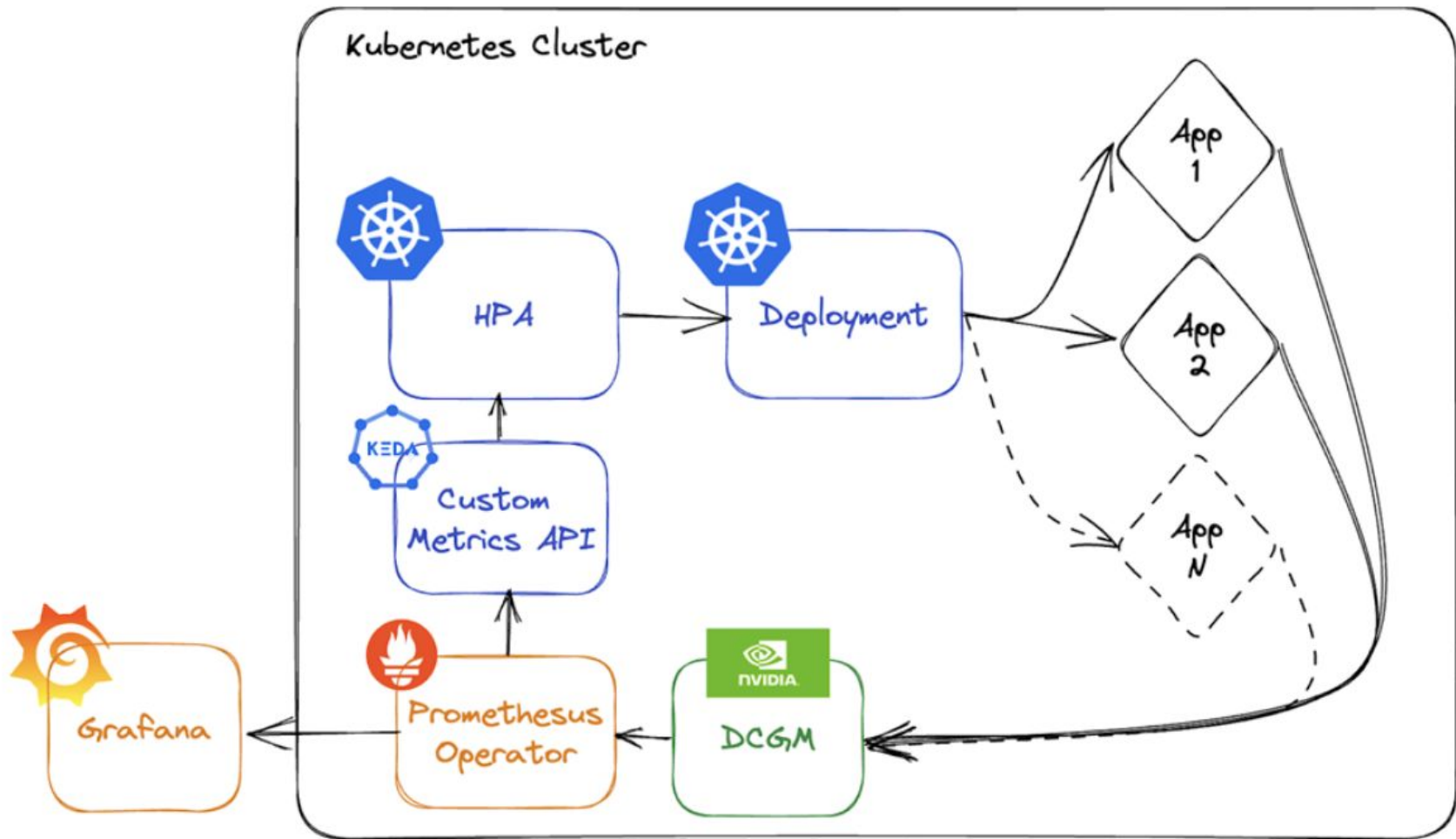
Ресурсы и масштабирование

- Условно бесконечное масштабирование ресурсов
- Возможность деплоя сервисов на разные ресурсы - GPU/CPU, GPU разного размера



Ресурсы и масштабирование

- Условно бесконечное масштабирование ресурсов
- Возможность деплоя сервисов на разные ресурсы - GPU/CPU, GPU разного размера
- Возможность автоскейлинга по GPU-метрикам или скейлинга по расписанию (в час-пик)



Картинка из книги LLMs in Production

Ресурсы и масштабирование

- Условно бесконечное масштабирование ресурсов
- Возможность деплоя сервисов на разные ресурсы - GPU/CPU, GPU разного размера
- Возможность автоскейлинга по GPU-метрикам или скейлинга по расписанию (час-пик)
- Часто не требуется GPU-оптимизация - минимальная машина в облаках часто 16гб, этого может хватить с запасом

Обновление моделей

- Деплой в Кубернетес-кластер по клику в билд-пайплайне

✓

#v0.19.7-480814 Merged PR 89937: [feature] Extra fields for LateralityDescription -> "binary"

on BreastCancer Release ↗ Retained

Run new

⋮

Summary

Releases

Environments

Associated pipelines

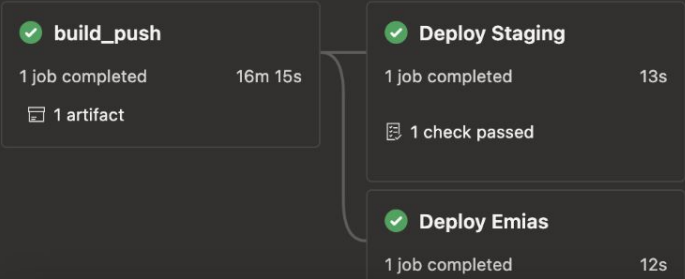
Manually run by  Лештаев Никита Дмитриевич

View 250 changes

Repository and version	Time started and elapsed	Related	Tests and coverage
 breastcancer	 Sep 9, 2024 at 10:58 AM	 0 work items	 Get started
 releases/v0.19.7  efb7209	 4d 2h 34m	 1 published	

Stages

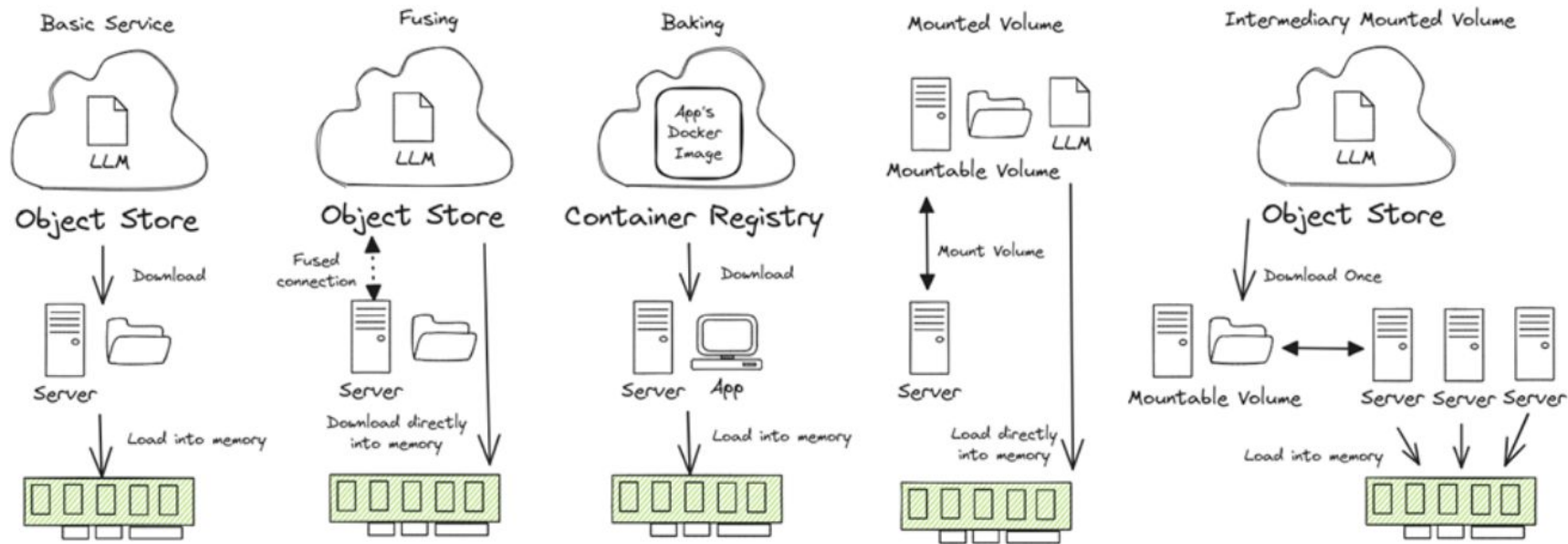
Jobs



Обновление моделей

- Деплой в Кубернетес-кластер по клику в билд-пайплайне
- Возможность нулевого даунтайма
- Возможность практически мгновенного отката

Нюансы доставки моделей до машин



Железо и ОС

- Любая ОС по выбору - в нашем случае Ubuntu
- Большой выбор видеокарт разных размеров у облачных провайдеров

Мониторинг и алерты

- Мгновенный доступ ко всем исследованиям и предсказаниям

Фильтры потока по заказчикам 🧐

Количество последних исследований за сегодня

5

-

+

Выбор даты

YYYY/MM/DD – YYYY/MM/DD

Режим отображения

Карточки



Выбери патологию



Норма?



Выбери диапазон вероятности

0.00

1.00

0.00

1.00

Время обработки исследования:

00:00

23:59

00:00

23:59

Специфичные фильтры по направлению



Поиск 🔍

⚓ Версия: 0.19.8b

💎 StudyInstanceUID:

1.2.826.0.1.3680043.8.839.1.146509357663417682717656983745878529

ID 11a3c95a-a536-416a-87dc-903fb6316f84

📅 03.04.25 12:23:02

🕒 (ML-only) 0 мин 50 сек

Норма? ❌

Вероятность: 0.2

Присутствующие патологии: lymphonodus, mass_benign, calcinates_benign

📄 Источник: Мосмед MMG RU

- 🏢 Организация из дайком тэгов - БУЗ ВО "ВОЖЕГОДСКАЯ ЦРБ"

🧠 Специфичные параметры для команд

- 🧠 *birads*: [1.0, 2.0]
- 🧠 *pgmi*: G
- 🧠 *fabric_density*: [2, 2]

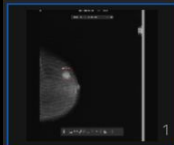
📄 скачать оригинал

📄 скачать ответ

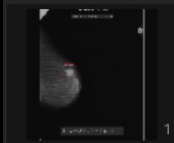
🖥️ посмотреть оригинал

🖥️ посмотреть ответ

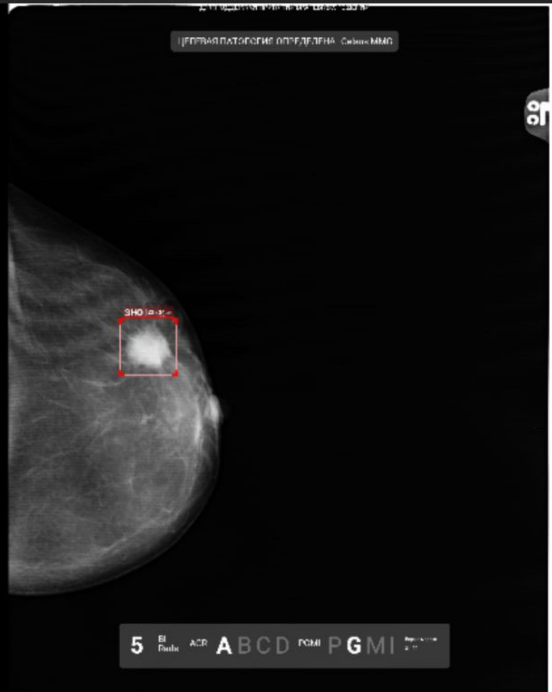
📄 Скачать логи



ANONYMIZED
4_34155606



Ser: 1002
2364 x 2964



Apr 3, 2025 08:28:56

Zoom: 18%
W: 255 L: 127
Lossless / Uncompressed

Мониторинг и алерты

- Мгновенный доступ ко всем исследованиям и предсказаниям
- Все алерты в любимом мессенджере

[FIRING:1] Процент тех дефектов. MMG. EMIAS ML HUB TEAM (mmg Мосмед MMG RU emias)

Клиент Мосмед MMG RU

Схема emias

Описание: Превышение порога по тех дефектам за 15.44291755 минут

Процент техдефектов = 57.3170731707317

Количество техдефектов = 47

▼ Show more

 Grafana v11.1.1



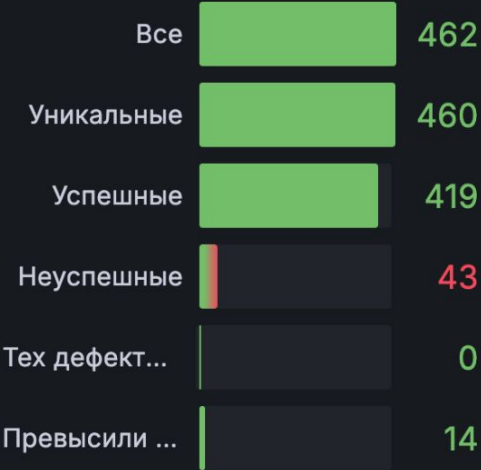
11 replies

Follow

Мониторинг и алерты

- Мгновенный доступ ко всем исследованиям и предсказаниям
- Все алерты в любимом мессенджере
- Все нужные дашборды в Графани

Количество исследований



Процент тех дефектов (пока только ...

3.26%

Время поэтапно

study_uid	start_timestamp	before_start_download	download	ml_l
1.2.392.200036.9125	2025-04-03 09:17:00	4.58 s	18.3 s	
1.2.826.0.1.3680043.8	2025-04-03 09:18:00	6.42 s	16.3 s	
1.2.840.113619.2.373.	2025-04-03 09:17:00	1.93 s	11.6 s	
1.2.840.113619.2.401.	2025-04-03 09:14:01	3.74 s	9.01 s	
1.2.840.113619.2.373.	2025-04-03 09:16:01	6.44 s	14.4 s	
1.2.840.113619.2.401.	2025-04-03 09:22:01	7.45 s	8.60 s	
1.2.826.0.1.3680043.2	2025-04-03 11:59:00	3.19 s	16.3 s	
1.2.826.0.1.3680043.8	2025-04-03 10:32:00	8.30 s	24.5 s	
1.2.826.0.1.3680043.8	2025-04-03 10:48:20	4.74 s	20.8 s	
1.2.826.0.1.3680043.2	2025-04-03 11:10:26	5.40 s	14.3 s	

Количество исследований

Все попытки обработки

783

Уникальные StudyInstanceUID

776

Успешные

724

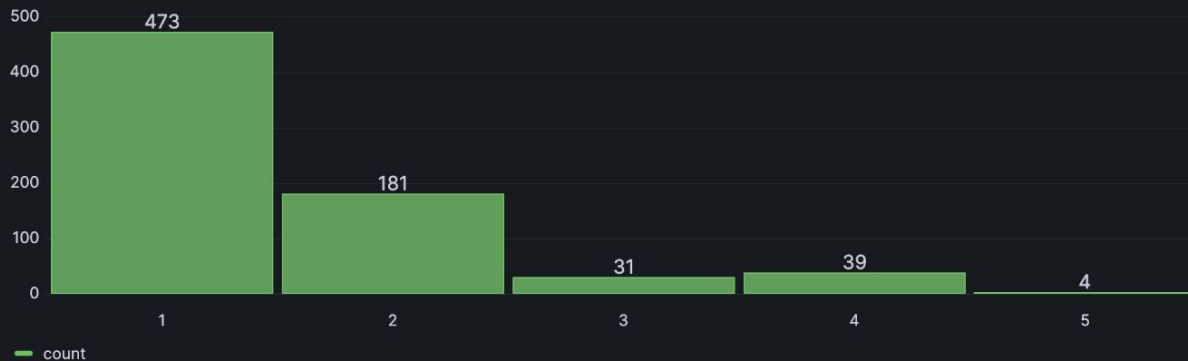
Неуспешные

52

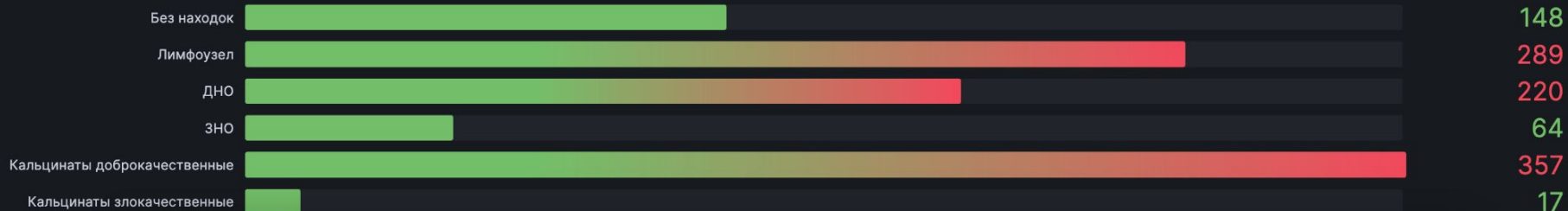
Превысили 6.5 мин

0

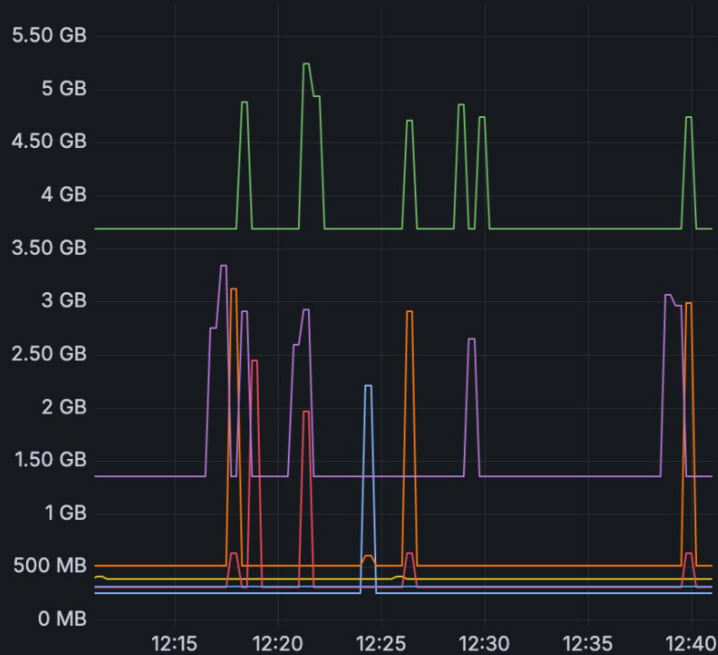
Максимальный Bi-Rads по пациентам



По патологиям



vRAM на Pod



Name	Last *、
celsus-core-kate-production-complex / detection-d5998d677-df2p2	3.68 GB
celsus-core-kate-production-complex / roiseg-78456d8ccb-r4zzl	1.35 GB
celsus-core-kate-production-complex / postprocessor-emphysema-686d6677db-j26dw	514 MB
celsus-core-kate-production-complex / lung-segmentation-787b769644-tnxcr	384 MB
celsus-core-kate-production-complex / postprocessor-ribs-fracture-69989454cd-gg7bc	316 MB
celsus-core-kate-production-complex / postprocessor-osteo-fracture-867f484fc6-lrb6d	306 MB
celsus-core-kate-production-complex / postprocessor-coronary-calcium-765d988c6b-7s6p8	250 MB

Входные данные

- Можно мониторить входные данные - случайно и по выборкам, давать фидбек заказчикам по качеству данных

Входные данные

- Можно мониторить входные данные - случайно и по выборкам, давать фидбек заказчикам по качеству данных
- Можно прикрутить автоматический мониторинг сдвигов

Appearance



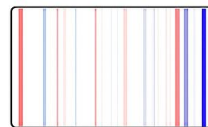
Input



Foundation Model



Foundation Model



Model Predictions



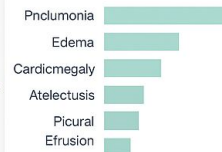
Input



Deployed Model



Predicted Probabilities



Metadata



Input



Extract Metadata



DICOM Metadata

Attributes	
Patient ID	Patient ID
Patient's Birth Date	1997-45-21
Patient's Sex	M
Number of	Framing
Modality	28
Rows	222
Columns	234
Study Date	Jan
Photometric Interpretation	TRE
Study Type	Default

Уровень 2



Спускаемся ниже

Клиенты:

- Часть регионов (централизованно на уровне Минздрава)
- Часть коммерческих клиник

Закупают выделенный сервер с GPU

Дают туда доступ по VPN

Устанавливаем нашу систему

Ресурсы и масштабирование

- Фиксированный железный сервер в ЦОДе региона

Ресурсы и масштабирование

- Фиксированный железный сервер в ЦОДе региона
- Не очень дружелюбные админы

Ресурсы и масштабирование

- Фиксированный железный сервер в ЦОДе региона
- Не очень дружелюбные админы
- Ресурсы докинуть почти невозможно, если где-то зафакапили в ТЗ

В ТЗ: видеокарта NVIDIA, архитектуры такие-то, объем VRAM не менее 8гб

В Т3: видеокарта NVIDIA, архитектуры такие-то, объем VRAM не менее 8гб

В реальности:

```
wed Nov 20 15:48:57 2024
+-----+
| NVIDIA-SMI 525.85.05      Driver Version: 525.85.05      CUDA Version: 12.0      |
+-----+-----+-----+-----+-----+-----+
| GPU  Name                Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
|                                       |                    |    MIG M. |
+-----+-----+-----+-----+-----+-----+
|   0  NVIDIA A10-6C        On         | 00000000:06:00.0 Off |   0          |
| N/A   N/A    P0     N/A /  N/A | 4956MiB / 6144MiB |   0%      Default |
|                                       |                    | Disabled |
+-----+-----+-----+-----+-----+-----+

+-----+
| Processes: |
| GPU  GI   CI           PID   Type   Process name                      GPU Memory |
|      ID   ID                                   |              Usage |
+-----+-----+-----+-----+-----+-----+
|   0   N/A  N/A         5099    C     python3                          4955MiB |
+-----+-----+-----+-----+-----+-----+
| 500MB IT 1 3 1 3 |
+-----+

```

Ресурсы и масштабирование

- Фиксированный железный сервер в ЦОДе региона
- Не очень дружелюбные админы
- Ресурсы докинуть почти невозможно, если где-то зафакапили в ТЗ
- Иногда нужно деплоить несколько систем на одном большом сервере, начинается конкуренция за ресурсы

Что делать?



Оптимизация моделей

- TensorRT, torch.compile
- Переход на новые версии торча - в одном случае снизило общее потребление на треть за счёт одного кастомного модуля
- Снижение размеров входного изображения

Что делать?



Fallback, если на конкретном исследовании падаем по памяти, то

- обрабатываем его на CPU
- уменьшаем размер
- снижаем параллелизацию

Что делать?



Дружить с админами

- Максимально облегчать им жизнь
- Не забывать про разницу в часовых поясах



Обновление моделей

Как доставлять новые модели?

- Пулл контейнеров из разрешенного реджистри
- Копирование tar.gz с образом на машину

Часто обновление только ночью или по согласованию

Что делать?



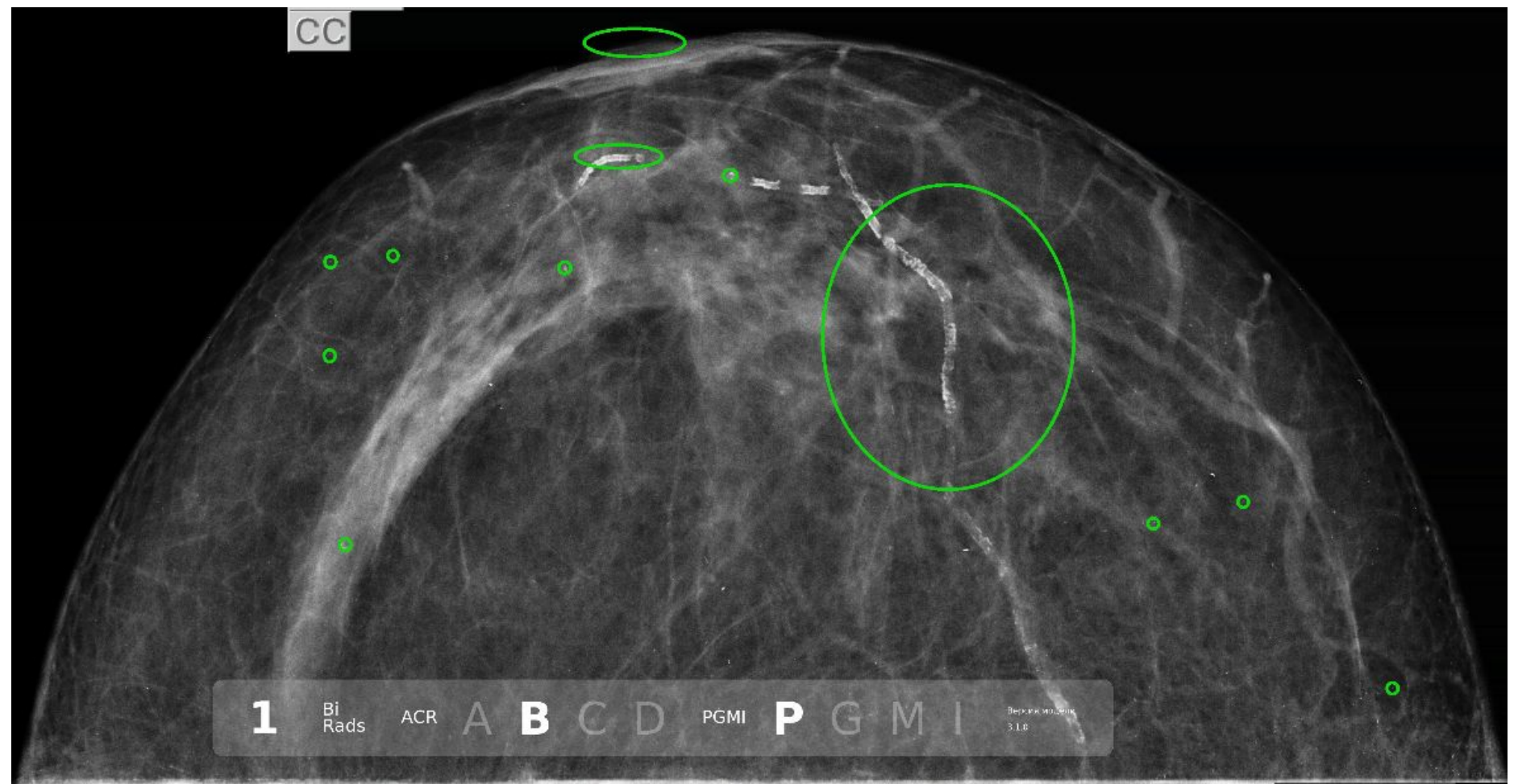
Максимальная возможная автоматизация:

- Прорываемся через бюрократию и настраиваем обновление по клику
- Автоматически собираем и пушим образы в реджистри или пакуем в архив
- Обновляем ночью по крону

Выбор железа и ОС

- Только российские ОС - АстраЛинукс, РедОС (удачи поставить на них новые драйвера NVIDIA)
- Древние видеокарты

error in DeformablePSROIPoolForward: no kernel image is available for execution on the device



Что делать?



- Требуем конкретные проверенные версии российских ОС
- Ассерты внутри сети на корректность работы всех модулей

Мониторинг и алерты

Дежурный раз в сутки

- заходит в каждый регион (помним про VPN)
- выгружает статистику из БД
- по запросу выгружает исследования (если они еще доступны)

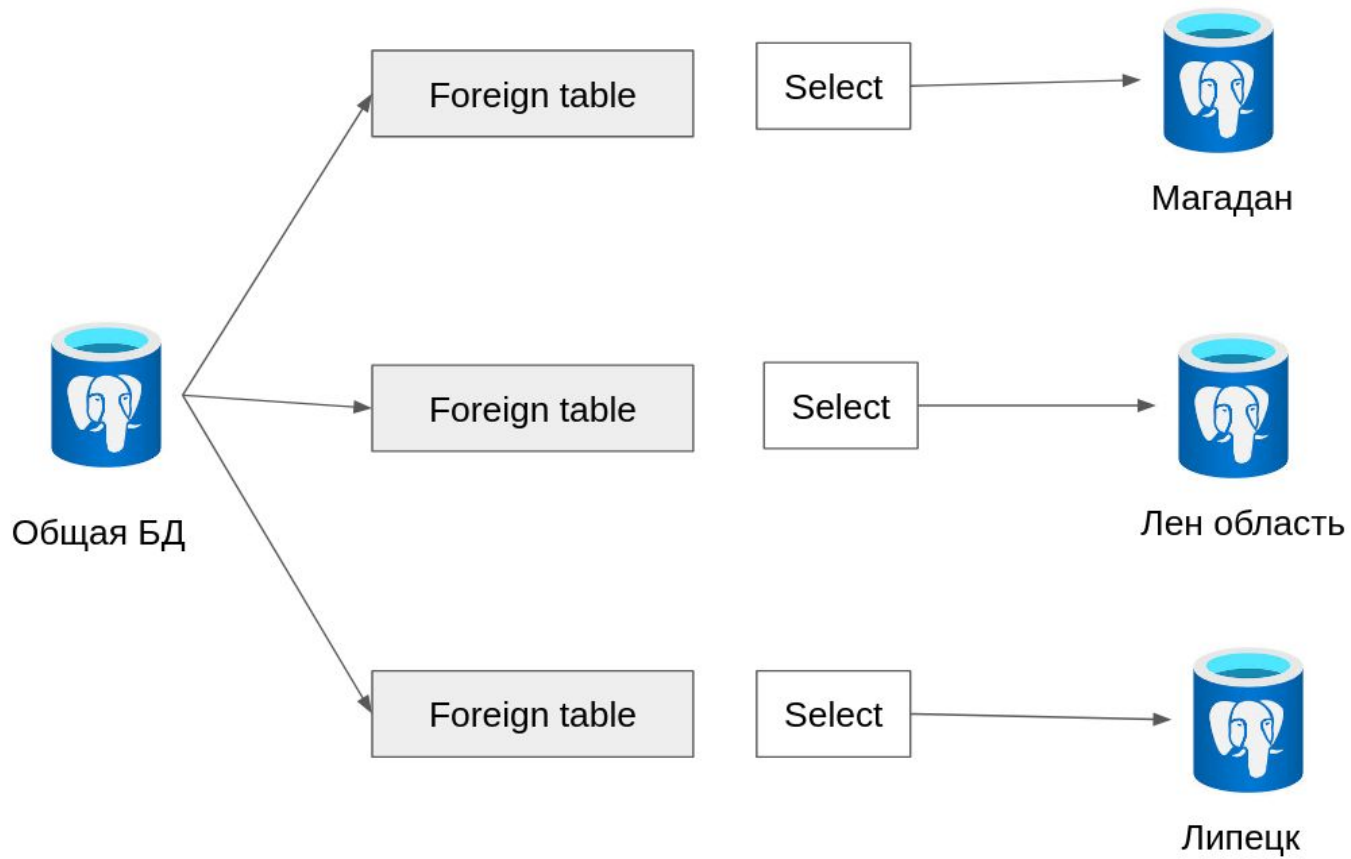


Что делать?

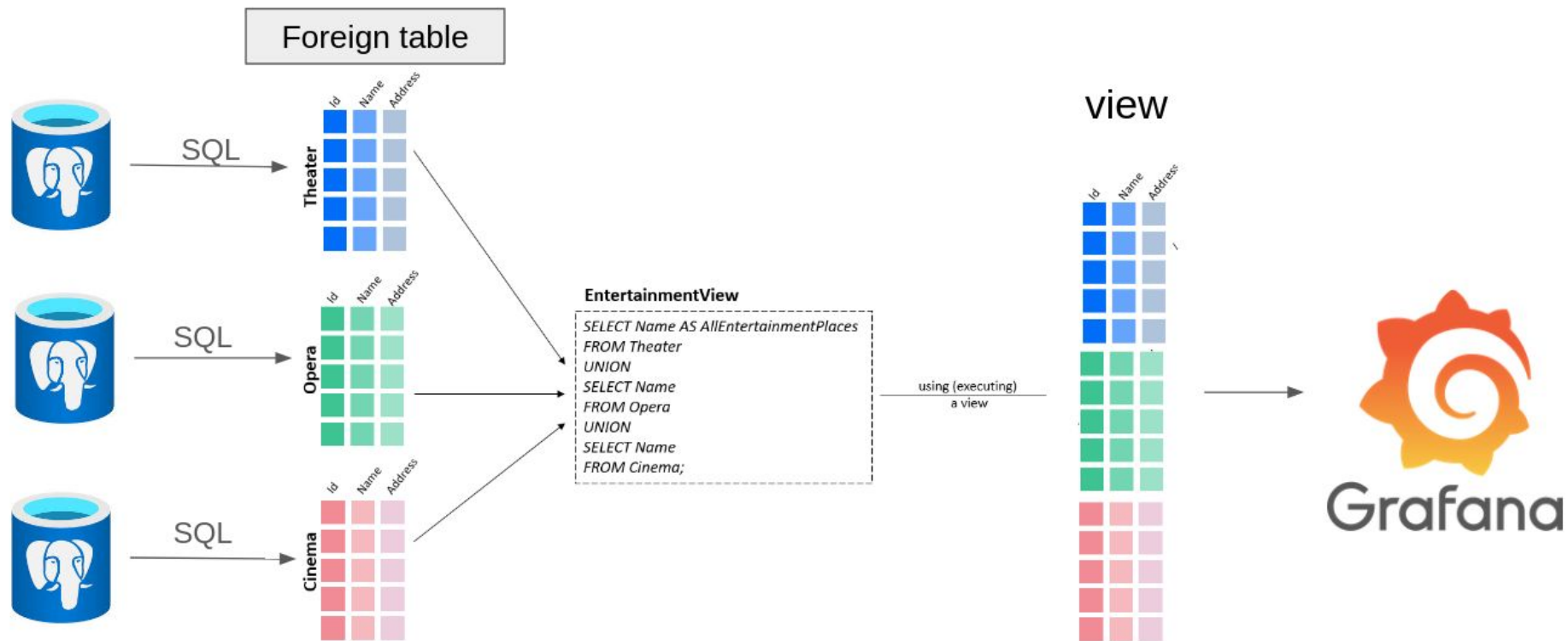
Очень хочется мониторить облачные и локальные деплои из одного места.

Используем инструменты БД:

- Foreign tables
- View с объединением данных из нужных регионов



View



Входные данные

- Очень сложно контролировать поток входных данных в реальном времени
- Очень много мусора, мы не контролируем правила маршрутизации на ИИ-сервис

ЦЕЛЕВОЙ ОРГАН СНИТ НЕ ПОЛНОСТЬЮ
ТОЧНОСТЬ ЗАКЛЮЧЕНИИ НЕ МОЖЕТ БЫТЬ ГАРАНТИРОВАНА

П

ПНЕВМОТОРАКС

ПНЕВМОТОРАКС ПЕРЕЛОМ РЕБРА

L

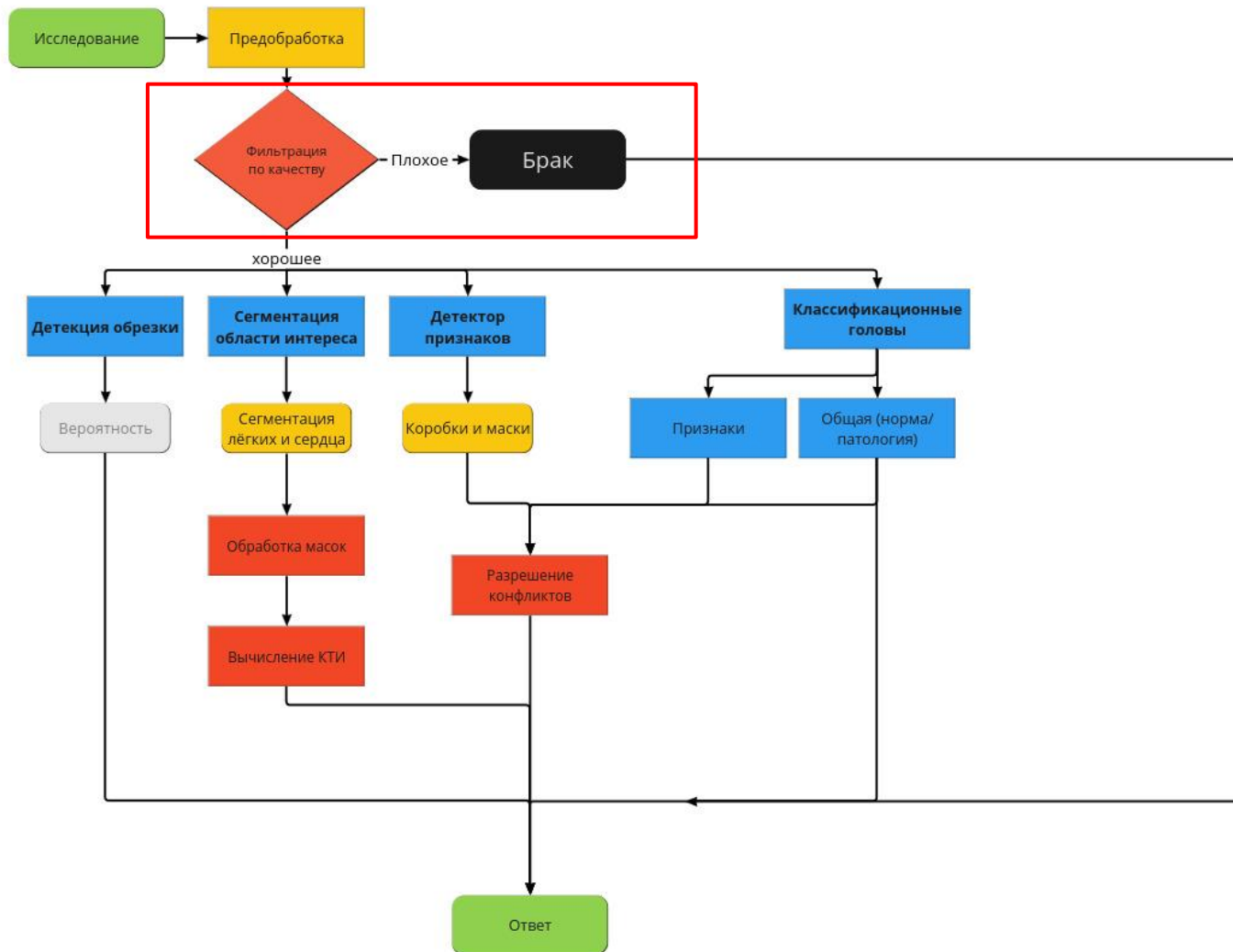
L cc

20 cm

Что делать?

Добавляем модуль фильтрации





Уровень 3



Преисподняя

- Передвижные маммографические и рентген-комплексы
- Медицинский поезд с рентгенологическим вагоном
- Как часть рентген-оборудования

Ресурсы и масштабирование

- Нулевая масштабируемость и гибкость
- Особенно внимательно пишем ТЗ или покупаем оборудование сами
- По возможности закладываемся на будущие обновления моделей

Что делать?



- **Очень внимательно** пишем ТЗ или покупаем оборудование сами
- По возможности закладываемся на будущие обновления моделей
- Не верим на слово по поводу нагрузки на систему

Обновление моделей

- Фактически вечный деплой или раз в год, когда поезд встанет на ремонт

Что делать?



- Автообновление, когда доступна связь - опасно для ML-моделей, лучше использовать только для некритических компонентов

Железо и ОС

- Может быть что угодно, но чаще всего Винда
- Оборудование чаще всего стоит кое-как, греется, из него могут выпадать провода
- Как-то раз попросили задеплоить на российской GPU
- Часто требуется инференс на CPU

~ Александр Тр...

Сейчас попробуют. Там
сейчас дизель перезапускают
и после перезапуска
запустим систему. Сообщу

10:08



~ Александр Тр...

Сервер снимаем и везем в
Москву

14:04



~ Екатерина

Ок, когда приедет примерно?

14:05



~ Илья

Меня больше интересует
когда он вернется?

14:05

Мониторинг и алерты

- Мониторинг фактически невозможен - максимум периодическая выгрузка отчётов/логов
- Доступ по SSH меняется в лучшем случае на Энидеск, в худшем - на билет в Читу



Что делать?

Критически важным становится самовосстановление системы

- Watchdog service
 - Проверяет доступность API простым health_check
 - Проверяет доступность нейронки тяжёлым health_check
 - Регулярно чистит место на диске
- Большая красная кнопка для перезагрузки всей системы
- Сохранение логов и загрузка их на удаленный сервер при наличии интернет-соединения

ИТОГИ

А что если...

Объединить лучшее из двух миров

- Сетка в облаке
- Остальное у клиента

Плюсы гибрида:

- Контроль над железом, ОС, GPU-ресурсами
- Легко мониторить входы и выходы сети
- Сетка принимает обезличенные данные

Выводы

- Каждый локальный деплой требует индивидуального подхода
- Иногда можно уговорить клиента на облака или гибрид
- Человеческие отношения решают в деплоях, где мы не контролируем инфра
- Всегда стелим инженерный пух (даже в облаках) - фоллбеки, вотчдоги, хелсчеки и другие английские слова, когда-нибудь они вас спасут



**Я вас благодарю
за внимание**

@crazyfrogspsb в ТГ

https://t.me/varim_ml - канал про ML

