

ООО «АИДАТех»

119021, Г.Москва, ул Льва Толстого, д. 2/22 стр. 6

ИНН 9704261020

КПП 770401001



ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ «KAGECORE ML PLATFORM»

Руководство пользователя

Листов 74



Содержание

Обозначения и сокращения	4
Термины и определения	5
1 Компоненты Системы	Ошибка! Закладка не определена.
1.1 Основные компоненты Системы	Ошибка! Закладка не определена.
2 Начало работы	8
2.1 Регистрация новых пользователей	8
2.2 Авторизация пользователей	8
2.3 Личный кабинет.....	9
2.4 Главная страница.....	Ошибка! Закладка не определена.
2.5 Заказы	12
2.6 Базовые вычисления	13
2.6.1 Заказать VM с unix-подобной ОС	14
2.6.2 Заказать VM с Windows ОС.....	15
2.7 Среда разработки.....	16
2.7.1 File Browser	16
2.7.2 VS Code.....	18
2.7.3 Jupyter Lab	21
2.8 Контейнеры	24
2.8.1 Заказать Kubernetes cluster.....	25
2.9 Базы данных.....	27
2.10 Брокеры сообщений	30
2.11 Веб-приложения	32
2.11.1 Заказать NGINX.....	33
2.11.2 Заказать Wildfly	34
2.12 Список VM	36
2.12.1 Управление доступом	37
2.12.1 Организация	37
2.12.2 Настройка организационной структуры	38
2.12.2.1 Создание папки/проекта.....	38
2.12.2.2 Удаление папки/проекта	40
2.13 Управление ресурсными квотами	41
2.13.1 Добавление ресурсных квот для папки/проекта	41

2.13.2 Редактирование ресурсных квот для папки/проекта	41
2.13.3 Удаление ресурсных квот для папки/проекта	42
2.14 Управление счетами.....	42
2.15 Сервисные аккаунты	43
2.15.1 Создать сервисный аккаунт	43
2.15.2 Изменить сервисный аккаунт	43
2.15.3 Удалить сервисный аккаунт	43
2.16 Учетные записи	44
2.16.1 Добавить пользователя к организации/папке/проекту	44
2.16.2 Редактировать права пользователя	44
2.16.3 Отозвать права пользователя	44
2.17 Роли.....	45
2.17.1 Создать новую роль.....	45
2.17.2 Удалить роль	45
2.18 SSH-ключи	46
2.18.1 Добавить SSH-ключ	46
2.19 Мониторинг	47
2.19.1 Метрики VM.....	47
2.19.1.1 Метрики VM на ОС Linux.....	48
2.19.1.2 Метрики VM на ОС Windows.....	50
2.19.1.3 Импорт графиков	54
2.19.2 Метрики подов.....	55
2.19.3 Мониторинг платформы	57
2.19.3.1 Описание преднастроенных дашбордов системы визуализации метрик.....	57
2.20 Inference	62

Обозначения и сокращения

В настоящем документе применяют следующие сокращения и обозначения:

API	-	Application Programming Interface, программный интерфейс взаимодействия
CPU	-	Central Processing Unit, процессор
GB	-	Gigabyte, гигабайт
GPU	-	Graphics Processing Unit, графический процессор
ID	-	Identifier, идентификатор
IP	-	Internet Protocol, межсетевой протокол
JSON	-	JavaScript Object Notation, текстовый формат обмена данными, основанный на JavaScript
ML	-	Machine Learning, машинное обучение
NGINX	-	engine X, веб-сервер и почтовый прокси-сервер, работающий на Unix-подобных операционных системах
NTP	-	Network Time Protocol, протокол сетевого времени
OS	-	Operating System, операционная система
RAM	-	Random Access Memory, оперативная память
RBAC	-	Role-Based Access Control, управление доступом на основе ролей
RSA	-	Rivest–Shamir–Adleman, криптографический алгоритм с открытым ключом, используемый для безопасной передачи данных и цифровой подписи
SSH	-	Secure Shell, безопасная оболочка
TCP	-	Transmission Control Protocol, протокол управления передачей
БД	-	база данных
ВМ	-	виртуальная машина
ГБ	-	гигабайт
ОС	-	операционная система
ПО	-	программное обеспечение
ЦП	-	центральный процессор

Термины и определения

В настоящем документе применяют следующие термины с соответствующими определениями:

- | | | |
|-----------|---|--|
| кластер | - | это логическая группа хостов с общими доменами хранения и ЦП одного типа (Intel или AMD). Если модели ЦП хостов относятся к разным поколениям, то используются только те функции, которые присутствуют во всех моделях. Виртуальные машины динамически распределяются между хостами кластера и могут перемещаться между ними в соответствии с политиками, заданными в кластере, и настройками виртуальных машин. Кластер является самым высоким уровнем, на котором могут определяться политики электропитания и разделения нагрузки |
| контейнер | - | легковесные запускаемые образы, в состав которых входит некоторое ПО и его зависимости. Поскольку в контейнерах виртуализируется операционная система, вы можете запускать контейнеры одинаково в любом совместимом окружении |
-

Введение

Платформа KageCore ML Platform, включая модули KageCore ML Platform. Модуль тарификации, KageCore ML Platform. Модуль витрины сервисов и KageCore ML Platform. Модуль пользовательского мониторинга (далее – ПО, Система), предназначена для предоставления высокопроизводительных вычислительных ресурсов и совокупности сервисов (IaaS/PaaS/SaaS) в интересах одной или нескольких организаций и проектов, позволяя эффективнее обучать и эксплуатировать модели искусственного интеллекта, оптимизировать производственные и научно-исследовательские процессы и тем самым способствовать ускоренному развитию технологий ИИ.

ПО предназначено для решения следующих задач:

- формирование единого пула аппаратных средств (серверы с CPU и GPU, высокоскоростные сети, системы хранения), доступного пользователям на основе квотирования, ролевой модели и механизма биллинга;
 - автоматизация процесса заказа ресурсов (vCPU, RAM, GPU, объём дисков) через портал самообслуживания и программный интерфейс (API);
 - использование готовых шаблонов (маркетплейса) с преднастроенными библиотеками и фреймворками (TensorFlow, PyTorch, Scikit-learn, Keras), а также средствами разработки JupyterLab, VSCode;
 - использование механизмов контейнеризации и виртуализации для гибкой оркестрации и оперативного масштабирования ML-заданий;
 - предоставление инфраструктурных сервисов (виртуальные машины) и платформенных (среды разработки, БД, аналитические инструменты) в унифицированном виде;
 - разграничение прав и ресурсов на уровне отдельных организаций, проектов и групп, что позволяет параллельно вести несколько сценариев инференса и обучения;
 - внедрение роли и квот (RBAC), позволяющих ограничивать доступ и лимитировать объём ресурсов (CPU, GPU, память, хранилище);
 - учёт и детальный биллинг (включая CPU, GPU, хранение данных), обеспечивающие прозрачность и справедливое распределение затрат между участниками;
 - поддержка экспериментов и трассировки (логирование метрик, параметров, артефактов) с целью воспроизводимости и контроля качества обучаемых моделей;
 - предоставление пользователям возможности оперативного выбора и автоконфигурации необходимых сервисов из унифицированного каталога (marketplace), содержащего преднастроенные модули, которые охватывают базовые и
-

прикладные сервисы, а также автоматизированного развертывания в контейнерной или виртуальной среде без ручной настройки.

1 Начало работы

1.1 Регистрация новых пользователей

Для получения доступа к платформе KageCore ML Platform необходимо направить запрос на регистрацию новой учётной записи администратору системы или в службу технической поддержки, указав контактную электронную почту и, при необходимости, информацию о принадлежности к проекту или команде.

Регистрация выполняется вручную, и после её завершения на указанную почту автоматически отправляется уведомление с данными для входа: ссылка на веб-интерфейс платформы, уникальный логин пользователя (включающий персональный идентификатор) и временный пароль, который необходимо сменить при первом входе. Учётная запись пользователя наделяется правами, соответствующими его роли и месту размещения в структуре платформы — в частности, пользователь может создавать и управлять виртуальными ресурсами в рамках назначенной папки или проекта. Доступ к ресурсам за пределами назначенной папки запрещён. Платформа применяет политики безопасности, соответствующие внутренним стандартам организации, включая требования к сложности пароля, обязательную смену временного пароля, возможность регулярной ротации паролей и блокировку учётной записи после нескольких неудачных попыток входа.

1.2 Авторизация пользователей

После регистрации и получения письма с данными для входа на KageCore ML Platform пользователь может авторизоваться на портале.

Для авторизации на KageCore ML Platform:

1. В поле **Логин или e-mail** введите логин, созданный ранее.
 2. В поле **Пароль** введите пароль, созданный ранее.
 3. Нажмите **Войти**.
-

Вход

Логин или email

Manage Passwords



[Забыли пароль?](#)

Войти

После успешной авторизации вы попадете на главную страницу портала.

Если авторизоваться не получилось, проверьте введенные значения в поле **Логин или e-mail** и **Пароль**. Если проблема сохраняется, обратитесь в техническую поддержку.

1.3 Личный кабинет

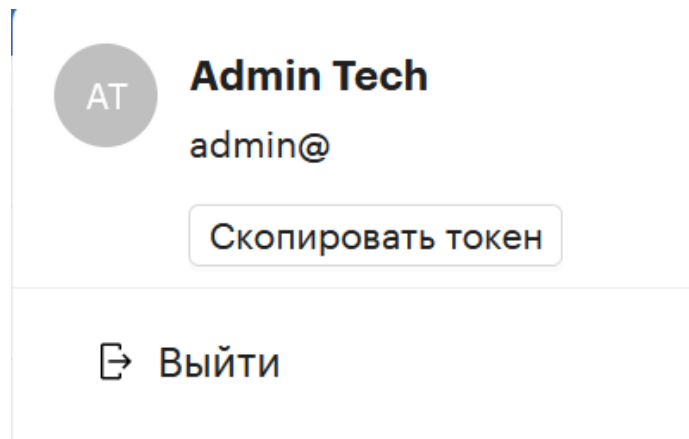
Чтобы открыть информацию об аккаунте, нажмите в правом верхнем углу на круг с инициалами логина.

Продукт	Платформа	Имя	CPU, RAM, Диск	Создан
Ubuntu Базовые вычисления		ratel-dgraph	4; 8Гб; 20Гб	04.08.2025, 12:44 admin@
Ubuntu Базовые вычисления		dgraph-alpha-zero-3	4; 8Гб; 20Гб	04.08.2025, 12:40 admin@
Ubuntu Базовые вычисления		dgraph-alpha-zero-2	4; 8Гб; 20Гб	04.08.2025, 12:35 admin@
Ubuntu Базовые вычисления		Dgraph-alpha-zero-1	4; 8Гб; 20Гб	04.08.2025, 12:25 admin@
MilvusCluster Базы данных		milvus-clus-0121	8; 32Гб; 0Гб	04.08.2025, 09:50 admin@

В информации об аккаунте отображается следующая информация:

- имя и фамилия текущего пользователя
- логин на KageCore ML Platform
- электронная почта
- актуальная версия KageCore ML Platform

Также здесь можно скопировать токен.



1.4 Web-интерфейс

В web-интерфейсе KageCore ML Platform отображаются следующие компоненты:

- Текущий контекст** проекта с информацией о сумме списаний в сутки. При изменении набора заказов эта сумма обновляется автоматически.
- Баланс счета** с информацией о доступной сумме, которая обновляется автоматически.
- Информационная панель **История действий**. В ней содержится информация о времени выполнения действия, последнем изменении, статусе выполнения действия и инициаторе.

 proj-develop-dev

 Заказы

 Управление доступом

 Мониторинг

>

 Скрыть меню

На главной странице расположены:

- главное окно

Информация в главном окне зависит от выбора в боковом меню:

- боковое меню (сворачивается кнопкой < **Скрыть меню** , разворачивается ≡)

Пункты бокового меню:

- Заказы
 - Все ресурсы
 - Базовые вычисления
 - Контейнеры
 - Базы данных
 - Брокеры сообщений
 - Веб-приложения
 - Список ВМ
 - События
 - Управление доступом
 - Организация
 - Сервисные аккаунты
 - Учетные записи
 - Роли
 - SSH-ключи
-

- Мониторинг ВМ
- Мониторинг подов

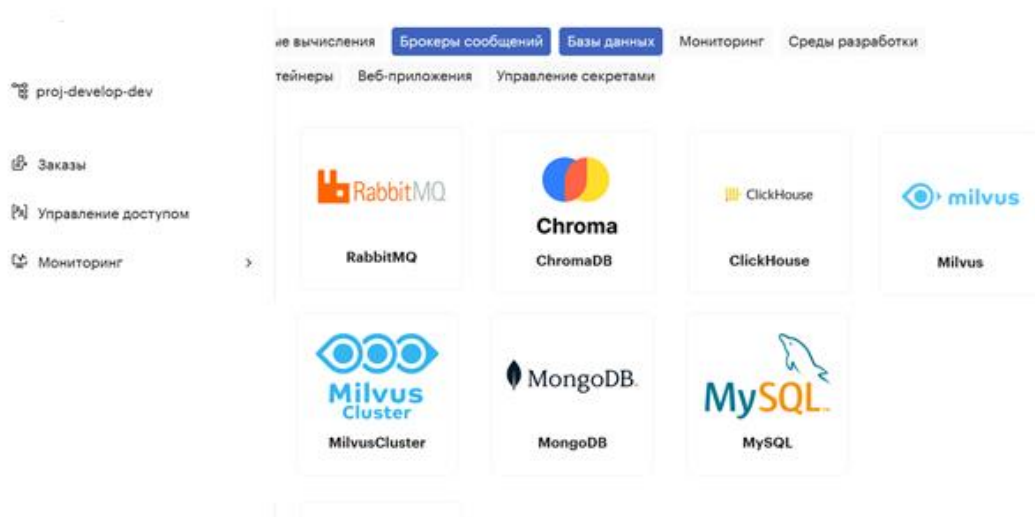
1.5 Заказы

В разделе **Заказы** отображается список всех созданных ресурсов в проекте. К ним можно применить сортировку по параметрам «Платформа», «Имя», «CPU, RAM, Диск», "Дата создания".

☐	Продукт	Платформа	Имя	CPU, RAM, Диск	Создан		
☐	Ubuntu Базовые вычисления		ratel-dgraph	4; 8ГБ; 20ГБ	04.08.2025, 12:44 admin@	✓	⋮
☐	Ubuntu Базовые вычисления		dgraph-alpha-zero-3	4; 8ГБ; 20ГБ	04.08.2025, 12:40 admin@	✓	⋮
☐	Ubuntu Базовые вычисления		dgraph-alpha-zero-2	4; 8ГБ; 20ГБ	04.08.2025, 12:35 admin@	✓	⋮
☐	Ubuntu Базовые вычисления		Dgraph-alpha-zero-1	4; 8ГБ; 20ГБ	04.08.2025, 12:25 admin@	✓	⋮
☐	MilvusCluster Базы данных		milvus-clus-0121	8; 32ГБ; 0ГБ	04.08.2025, 09:50 admin@	✓	⋮

Здесь можно заказать любой из ресурсов. Для этого:

1. В правом верхнем углу нажмите кнопку **+ Добавить ресурс**. Отобразятся ресурсы всех категорий.
2. Выберите ресурс и укажите параметры для его создания. Подробное описание параметров в разделах:
 - Базовые вычисления
 - Контейнеры
 - Базы данных
 - Брокеры сообщений
 - Веб-приложения



1.6 Базовые вычисления

В разделе **Заказы** на вкладке **Базовые вычисления** можно заказать услуги этой категории и просмотреть информацию о состоянии инстансов виртуальных машин. К созданным ресурсам можно применить сортировку по параметрам «Продукт», «Платформа», «Имя» и «CPU, RAM, Диск».

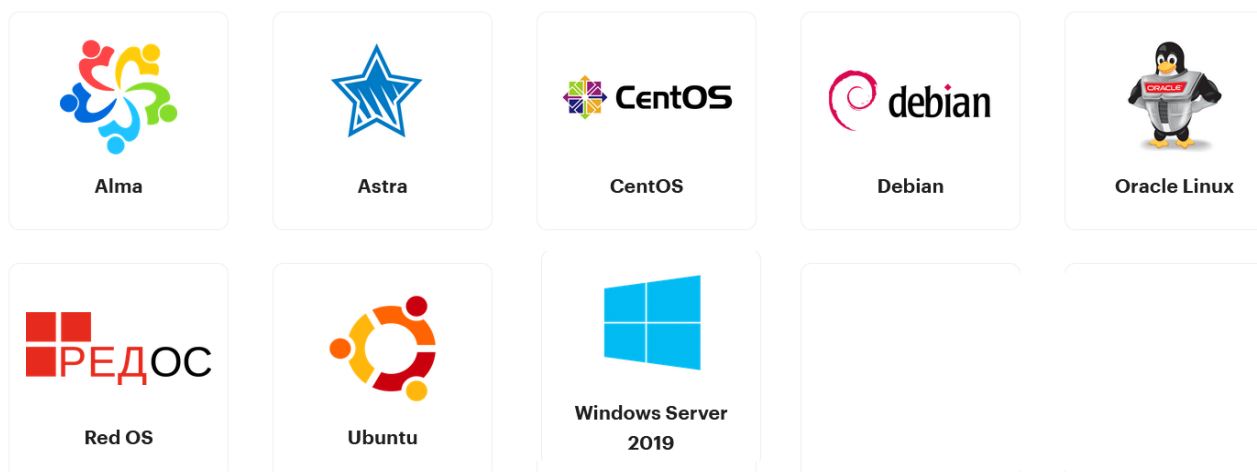
Для заказа услуги:

1. Перейдите в раздел **Базовые вычисления** и нажмите **Заказать**.
2. Выберите услугу:
 - Unix-подобные операционные системы:
 - Alma
 - Astra
 - CentOS
 - Debian
 - Red OS
 - Ubuntu
 - Операционная система Windows:
 - Windows Server 2019

proj-develop-dev | Папка fold-develop (951 524,83 ₽)

[Все продукты](#)
[Базовые вычисления](#)
[Брокеры сообщений](#)
[Базы данных](#)
[Мониторинг](#)
[Среды разработки](#)

[RAG-платформы](#)
[Контейнеры](#)
[Веб-приложения](#)
[Управление секретами](#)



Также заказать услуги всех категорий можно в разделе [Все ресурсы](#).

1.6.1 Заказать VM с unix-подобной ОС

Для заказа одной из виртуальных машин с операционными системами **Alma**, **Astra**, **CentOS**, **Debian**, **Red OS**, **Ubuntu** выполните следующие действия.

1. Нажмите на карточку услуги и заполните параметры:
 - a. **Имя виртуальной машины** — укажите название виртуальной машины.
 - b. **Сетевой сегмент** — выберите название логической портгруппы, из которой будет присвоение статического IP-адреса для виртуальной машины.
 - c. **Дата-центр** — выберите дата-центр.
 - d. **Платформа** — выберите платформу виртуализации, на которой будет размещена виртуальная машина.
 - e. **CPU и RAM** — выберите спецификацию VM, где CPU — размер vCPU, а RAM — размер оперативной памяти в ГБ

Примечание

- Значения в полях «Сетевой сегмент», «Дата-центр», «Платформа», «CPU и RAM» создаются на основании справочников, которые модерируются пользователями с правами администратора.
 - От выбранных конфигураций могут зависеть стоимость услуг
- f. **GPU** — выбрать из выпадающего списка "Целая или Mig", после выпадет GPU HOST.
 - g. **GPU HOST** — выберите Host из выпадающего списка.

- h. **Графические карты** — нажмите добавить и в выпадающем списке выберите нужные настройки.
 - **Количество** — введите количество GPU.
 - **Модель GPU** — выберите модель GPU используемую в поде или оставьте по умолчанию.
- i. **Загрузочный диск** — укажите размер загрузочного диска в ГБ или оставьте по умолчанию.
- j. **Точка монтирования** (опционально) — заполните параметры для подключения дополнительных дисков:
 - **размер, Гб** — укажите размер дополнительного диска.
 - **Путь монтирования** — укажите каталог для монтирования дополнительного диска.
 - **Тип файловой системы** — по умолчанию файловая система xfs.
- k. **SSH-ключи** — выберите имеющиеся ключи доступа или создайте новый ключ, нажав + **Добавить SSH-ключ**.

С созданной виртуальной машиной можно выполнить следующие действия:

- Изменить конфигурацию VM
- Выключить VM
- Удалить VM

1.6.2 Заказать VM с Windows ОС

Для заказа виртуальной машины на базе ОС Windows Server 2019 выполните следующие действия:

1. Нажмите на карточку услуги и заполните параметры:
 - a. **Имя виртуальной машины** — укажите название виртуальной машины.
 - b. **Сетевой сегмент** — выберите название логической портгруппы, из которой будет присвоение статического IP-адреса для виртуальной машины.
 - c. **Дата-центр** — выберите дата-центр.
 - d. **Платформа** — выберите платформу виртуализации, на которой будет размещена виртуальная машина.
 - e. **CPU и RAM** — выберите спецификацию VM, где core — размер vCPU, а RAM — размер оперативной памяти в ГБ.

- Значения в полях «Сетевой сегмент», «Дата-центр», «Платформа», «CPU и RAM» создаются на основании справочников, которые модерируются пользователями с правами администратора.
- От выбранной конфигурации может зависеть стоимость услуги.

- f. **Загрузочный диск** — укажите размер загрузочного диска в ГБ или оставьте по умолчанию.
- g. **Дополнительный диск** (опционально) --- укажите:
 - **Размер, Гб** — передвиньте ползунок слайдера, чтобы задать размер дополнительного диска.
 - **Local Disk** — укажите название для дополнительного диска. Название может быть любым.

- **Буква тома** — выберите, какая буква будет присвоена этому диску.
 - h. **Учетная запись** — задайте логин для пользователя ВМ.
 - i. **Пароль пользователя** — задайте пароль пользователя. Он должен состоять из 10-20 символов, включать цифры, символы, прописные и строчные латинские буквы.
2. Нажмите **Заказать**. Если не заполнены все обязательные поля, отмеченные знаком *, кнопка **Заказать** будет неактивна.

← Новый заказ
Windows Server 2019 — Базовые вычисления

Имя виртуальной машины: win19-0001

Основной сетевой интерфейс и ресурсы

Сетевой сегмент: DEV_DEV (Обезличенные данные)

Дата-центр: cod-a

Платформа:

Образ ОС: win19-template

CPU и RAM:

2 CPU	4 CPU	8 CPU	16 CPU
2 GB	4 GB	8 GB	12 GB
32 GB	48 GB	64 GB	

Загрузочный диск, Гб: 40

Дополнительный диск

Вы можете подключить к виртуальной машине дополнительные диски. Сочетание букв и имени тома не должно повторяться в разных дисках. Можно менять тип файловой системы, если он не установлен по умолчанию.

Добавить

Доступ

Придумайте имя и пароль пользователя. Длина пароля от 10 до 20 символов. Обязательно использовать цифры, прописные и строчные латинские буквы. Можно добавить несколько пользователей с уникальными именами.

Учетная запись: administrator

Пароль: [маска]

Добавить запись

Заказать

Итого в день: 12,62 Р

Оперативная память	4,91 Р
Процессор	3,56 Р
Жесткий диск	4,16 Р
Итого	12,62 Р

1.7 Среды разработки

1.7.1 File Browser

В разделе **Среды разработки** можно заказать услуги этой категории и просмотреть информацию о состоянии инстансов сред разработки. К созданным ресурсам можно применить сортировку по параметрам «Продукт», «Платформа», «Имя» и «CPU, RAM, Диск».

Для заказа услуги:

1. Перейдите в раздел **Среды разработки** и нажмите **Заказать**.

2. Выберите среду разработки:

- File Browser

Новый заказ

proj-develop-dev | Папка fold-develop (999 989 462,51 ₽)

Все продукты

Базовые вычисления

Среды разработки

Базы данных



File Browser



Jupyter Lab



VS Code

3. Укажите следующие параметры:

- Наименование пода** — укажите название пода.
- Образ ОС** — выберите образ операционной системы
- Платформа** — выберите платформу виртуализации, на которой будет размещен под.
- CPU и RAM** — выберите спецификацию VM, где CPU — размер vCPU, а RAM — размер оперативной памяти в ГБ

← **Новый заказ**

File Browser | Среды разработки

File Browser

Наименование пода

filebrowser-0012

Образ ОС

filebrowser:v2.31.2

Платформа

CPU и RAM

2 CPU 4 CPU 8 CPU 16 CPU

2 GB 4 GB 8 GB 12 GB 16 GB 24 GB

32 GB 48 GB 64 GB

Оперативная память для
filebrowser**4,91 ₽**

4 X 4,91 ₽

Процессор filebrowser

3,56 ₽

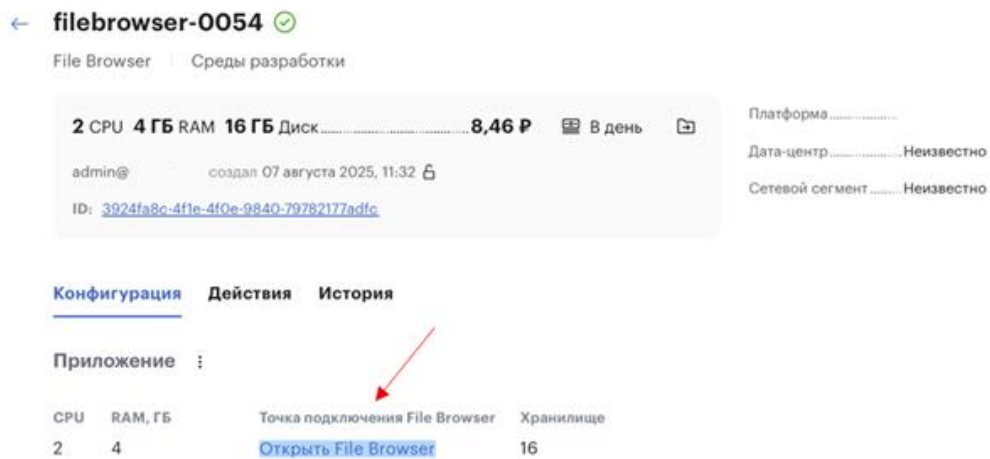
2 X 3,56 ₽

Итого в день

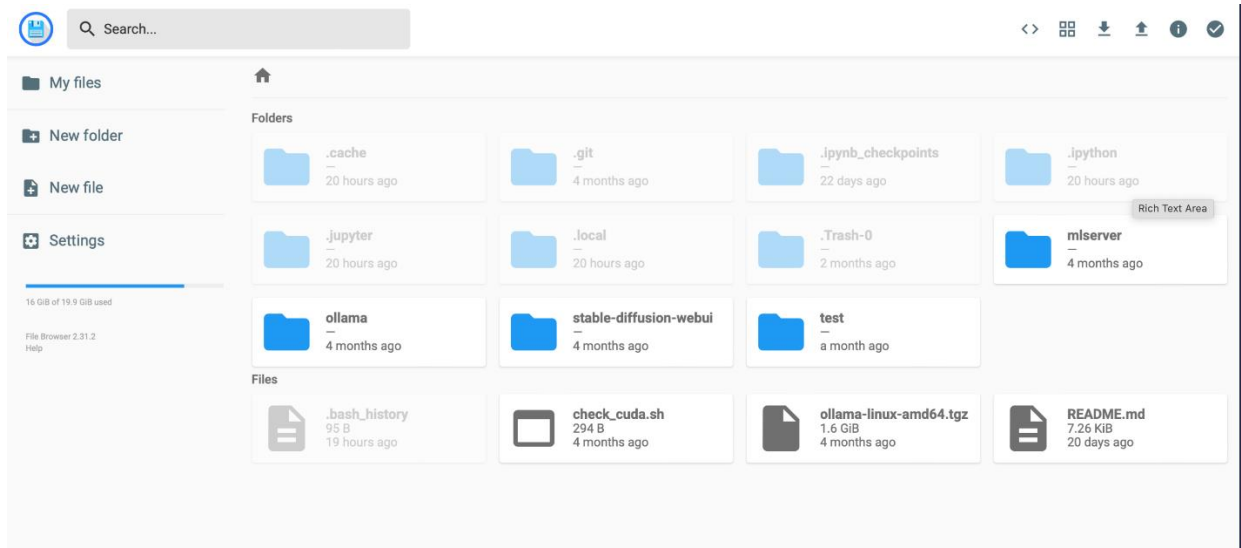
8,46 ₽

Заказать

4. Нажмите кнопку "Заказать" и после "Перейти к схеме выполнения". После создания заказа вы можете перейти в file browser, нажав на ссылку в поле "Точка подключения".



5. После этого у вас откроется file browser.



1.7.2 VS Code

В разделе **Среды разработки** можно заказать услуги этой категории и просмотреть информацию о состоянии инстансов сред разработки. К созданным ресурсам можно применить сортировку по параметрам «Продукт», «Платформа», «Имя» и «CPU, RAM, Диск».

Для заказа услуги:

1. Перейдите в раздел **Среды разработки** и нажмите **Заказать**.
2. Выберите среду разработки:
 - VS Code

Новый заказ

proj-develop-dev | Папка fold-develop (999 989 462,51 ₽)

Все продукты

Базовые вычисления

Среды разработки

Базы данных



File Browser



Jupyter Lab



VS Code

3. Укажите следующие параметры:
 - a. **Наименование пода** — кажите название пода.
 - b. **CPU и RAM** — выберите спецификацию VM, где CPU — размер vCPU, а RAM — размер оперативной памяти в ГБ
 - c. **Образ ОС** — выберите образ операционной системы
 - d. **Платформа** — выберите платформу виртуализации, на которой будет размещен под.
 - e. **Графические карты** — нажмите добавить и в выпадающем списке выберите нужные настройки.
 - Количество — введите количество GPU.
 - Модель GPU — выберите модель GPU используемую в поде или оставьте по умолчанию.
 - Значения в полях «Количество», «Модель GPU» создаются на основании справочников, которые модерируются пользователями с правами администратора.
 - От выбранной конфигурации может зависеть стоимость услуги.
 - f. **Временное хранилище** — введите размер временного хранилища в ГБ или оставьте по умолчанию.

← **Новый заказ**
VS Code | Среда разработки

VSCode

Наименование пода

CPU и RAM

2 CPU **4 CPU** 8 CPU 16 CPU

2 GB **4 GB** 8 GB 12 GB 16 GB 24 GB 32 GB

48 GB 64 GB

Образ ОС

Платформа

Графические карты

Модель GPU

Количество

Модель GPU

Временное хранилище

размер, Гб

Итого в день **16,90 Р**

Оперативная память для vscode **4,91 Р**
4 X 4,91 Р

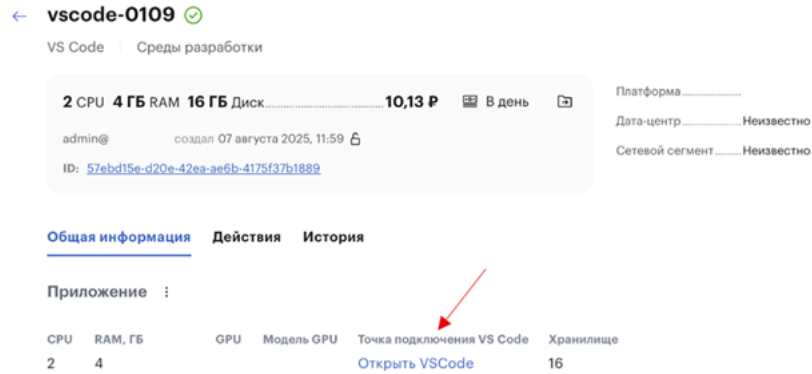
Процессор vscode **7,11 Р**
4 X 7,11 Р

Хранилище для vscode **1,66 Р**
16 X 1,66 Р

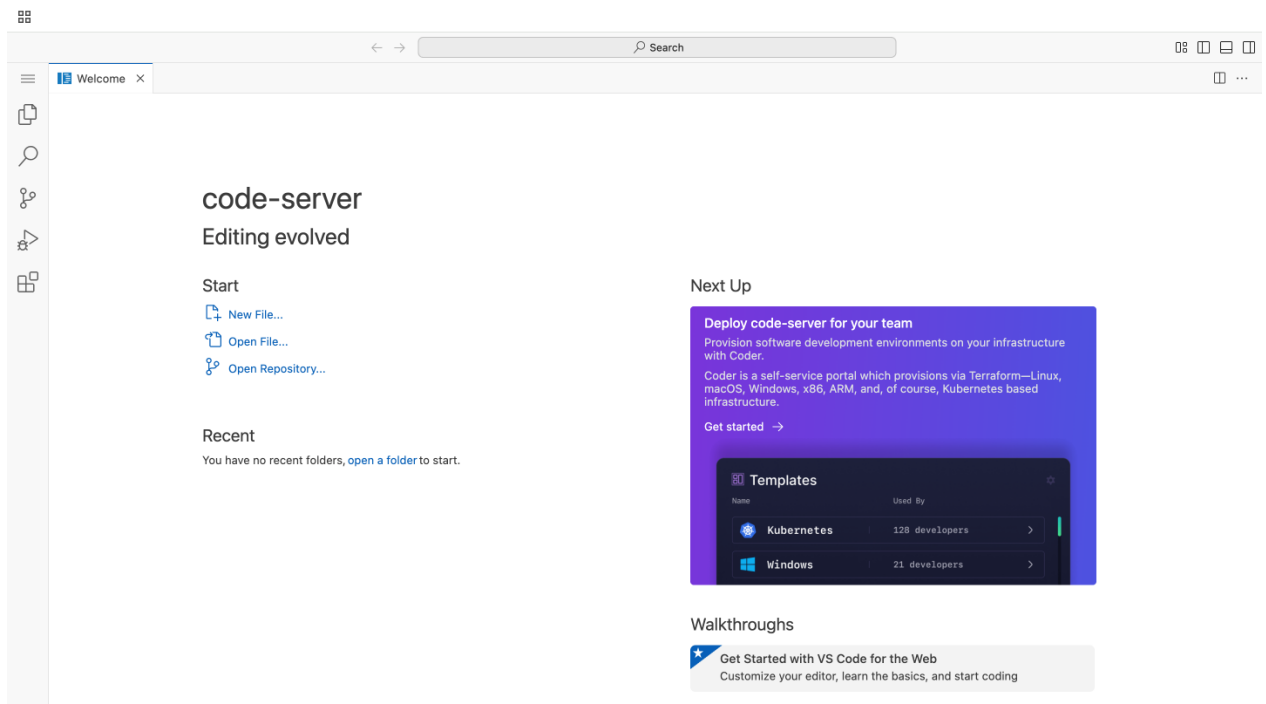
Графические карты NVIDIA-H100-80GB-HBM3-MIG-1g.10gb-SHARED для vscode **3,22 Р**
1 X 3,22 Р

Заказать

- Нажмите кнопку "Заказать" и после "Перейти к схеме выполнения". После создания заказа вы можете перейти в VS Code, нажав на ссылку в поле "Точка подключения".



5. После этого у вас откроется VS Code.



1.7.3 Jupyter Lab

В разделе **Среды разработки** можно заказать услуги этой категории и просмотреть информацию о состоянии инстансов сред разработки. К созданным ресурсам можно применить сортировку по параметрам «Продукт», «Платформа», «Имя» и «CPU, RAM, Диск».

Для заказа услуги:

6. Перейдите в раздел **Среды разработки** и нажмите **Заказать**.
7. Выберите среду разработки:
 - Jupyter Lab

Новый заказ

proj-develop-dev | Папка fold-develop (999 989 462,51 ₽)

Все продукты

Базовые вычисления

Среды разработки

Базы данных



File Browser



Jupyter Lab



VS Code

8. Укажите следующие параметры:
 - g. **Наименование пода** — кажите название пода.
 - h. **CPU и RAM** — выберите спецификацию VM, где CPU — размер vCPU, а RAM — размер оперативной памяти в ГБ
 - i. **Образ ОС** — выберите образ операционной системы
 - j. **Платформа** — выберите платформу виртуализации, на которой будет размещен под.
 - k. **Графические карты** — нажмите добавить и в выпадающем списке выберите нужные настройки.
 - Количество — введите количество GPU.
 - Модель GPU — выберите модель GPU используемую в поде или оставьте по умолчанию.
 - Значения в полях «Количество», «Модель GPU» создаются на основании справочников, которые модерируются пользователями с правами администратора.
 - От выбранной конфигурации может зависеть стоимость услуги.
 - l. **Временное хранилище** — введите размер временного хранилища в ГБ или оставьте по умолчанию.

← **Новый заказ**

Jupyter Lab | Среды разработки

Jupyter Lab

Наименование пода

CPU и RAM

2 CPU 4 CPU 8 CPU 16 CPU

2 GB 4 GB 8 GB 12 GB 16 GB 24 GB

32 GB 48 GB 64 GB

Образ ОС

Платформа

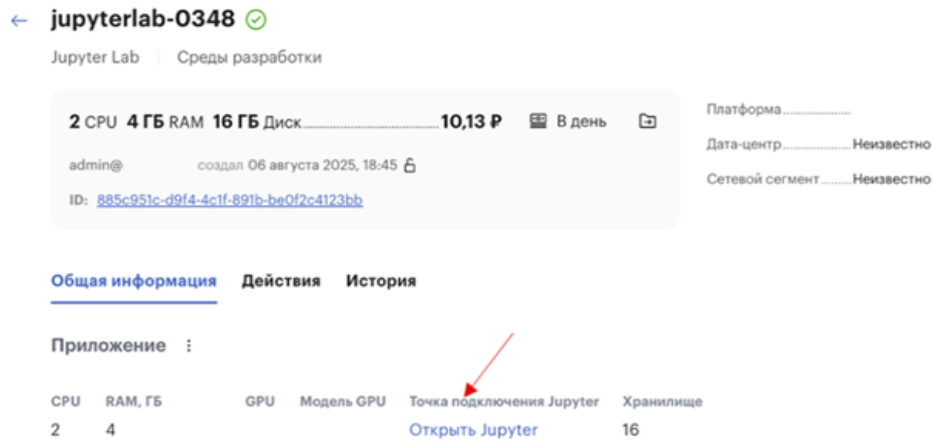
Оперативная память для jupyterlab **4,91 Р**
За 4 шт

Хранилище для jupyterlab **1,66 Р**
За 16 ГБ

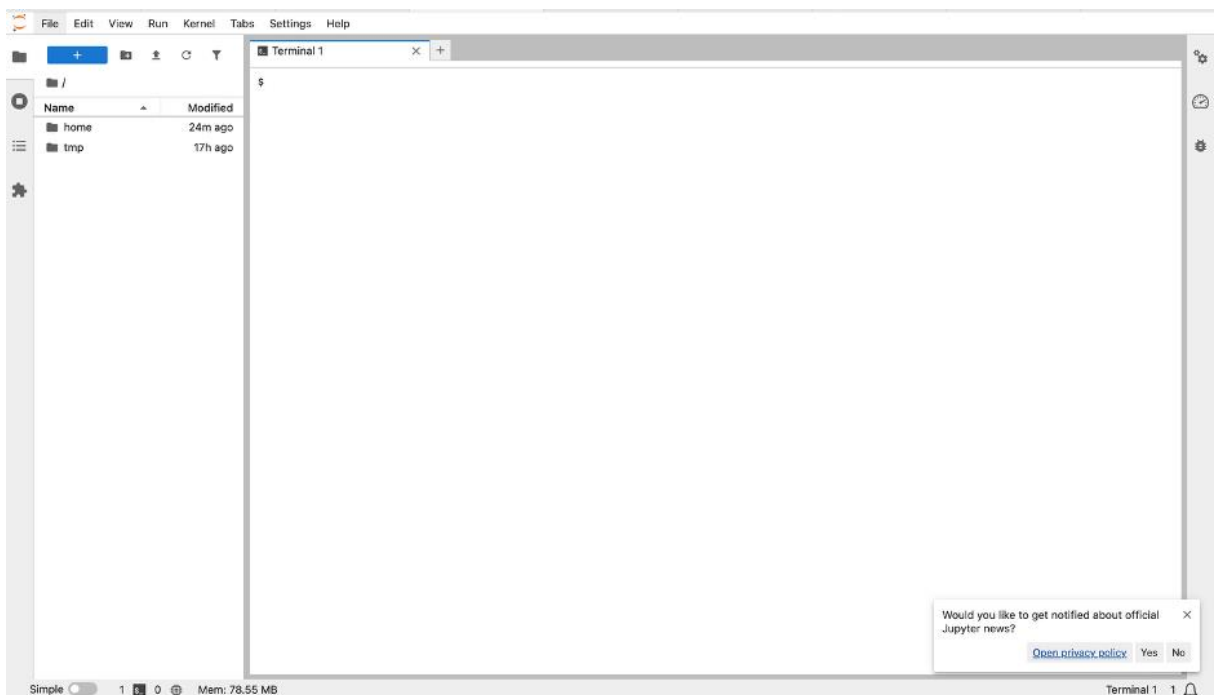
Процессор jupyterlab **3,56 Р**
За 2 шт

Итого в день **10,13 Р**

9. Нажмите кнопку "Заказать" и после "Перейти к схеме выполнения". После создания заказа вы можете перейти в Jupyter Lab, нажав на ссылку в поле "Точка подключения".



10. После этого у вас откроется Jupyter Lab.



1.8 Контейнеры

В разделе **Контейнеры** можно заказать услуги этой категории и просмотреть информацию о состоянии инстансов контейнеров. К созданным ресурсам можно применить сортировку по параметрам «Продукт», «Платформа», «Имя» и «CPU, RAM, Диск».

Для заказа услуги:

1. Перейдите в раздел **Контейнеры** и нажмите **Заказать**.
2. Выберите тип системы контейнеризации:
 - Kubernetes cluster

Новый заказ

Project1 | Организация Облачные решения (Демо) (14 489,47 ₽)

Все продукты

Базовые вычисления

Брокеры сообщений

Базы данных

Контейнеры

Веб-приложения



Также заказать услуги всех категорий можно в разделе [Все ресурсы](#).

1.8.1 Заказать Kubernetes cluster

1. Нажмите на карточку Kubernetes cluster и укажите параметры:

- a. **Имя кластера** — укажите уникальное название вашего кластера.
- b. **Сетевой сегмент** — выберите название логической портгруппы, из которой будет присвоение статического IP-адреса для виртуальной машины.
- c. **Дата-центр** — выберите центр обработки данных.
- d. **Платформа** — выберите платформу виртуализации.

- Значения в полях «Сетевой сегмент», «Дата-центр», «Платформа», «CPU и RAM» модерируются пользователями с правами администратора.
- От выбранной конфигурации может зависеть стоимость услуги.

e. **Образ ОС** — выберите выберите один из образов операционной системы.

f. **Master:**

- Количество — укажите количество создаваемых master-нод.
- Конфигурации CPU/RAM — выберите спецификацию, где core — количество vCPU, RAM — размер оперативной памяти в ГБ.
- Загрузочный диск — укажите количество ГБ для загрузочного диска.
- Диск для контейнеров — укажите от 5 до 100 ГБ памяти для использования контейнерами.

g. **Worker:**

- Количество — укажите количество создаваемых worker-нод.
- Конфигурации CPU/RAM — выберите спецификацию worker-нод.
- Загрузочный диск — укажите количество ГБ для загрузочного диска.
- Диск для контейнеров — укажите от 5 до 100 ГБ.

- h. **SSH-ключи** — выберите один или несколько ключей доступа или создайте новый, нажав + **Добавить SSH-ключ**.
 - 2. Проверьте параметры и нажмите кнопку **Заказать**. Если не заполнены все обязательные поля, отмеченные знаком *, кнопка **Заказать** будет неактивна.
-

← Новый заказ

Kubernetes cluster | Базовые вычисления

Имя кластера

k8s-test

Сетевой сегмент

DEV_DEV (Обезличенные данные)

Дата-центр

cod-a

Платформа

Образ ОС

ubuntu-template

Master

Количество

1

CPU и RAM

2 CPU 4 CPU 8 CPU 16 CPU

2 GB 4 GB 8 GB 12 GB 16 GB 24 GB

32 GB 48 GB 64 GB

Загрузочный диск, Гб

20

Диск для контейнеров, Гб

15

Worker

Количество ⓘ

1

10

1

CPU и RAM

2 CPU 4 CPU 8 CPU 16 CPU

2 GB 4 GB 8 GB 12 GB 16 GB 24 GB

32 GB 48 GB 64 GB

Загрузочный диск, Гб

20

Диск для контейнеров, Гб

15

Доступ

Выберите SSH-ключ из уже имеющихся или создайте новый

key-demo (key-demo количество: 1) ×

Процессор vsphere в типе среды PROD	7,11 Р
Оперативная память для vsphere	9,81 Р
Жесткий диск для vsphere в типе среды PROD	4,16 Р
Жесткий диск для vsphere в типе среды PROD	3,12 Р
Итого в день	24,21 Р

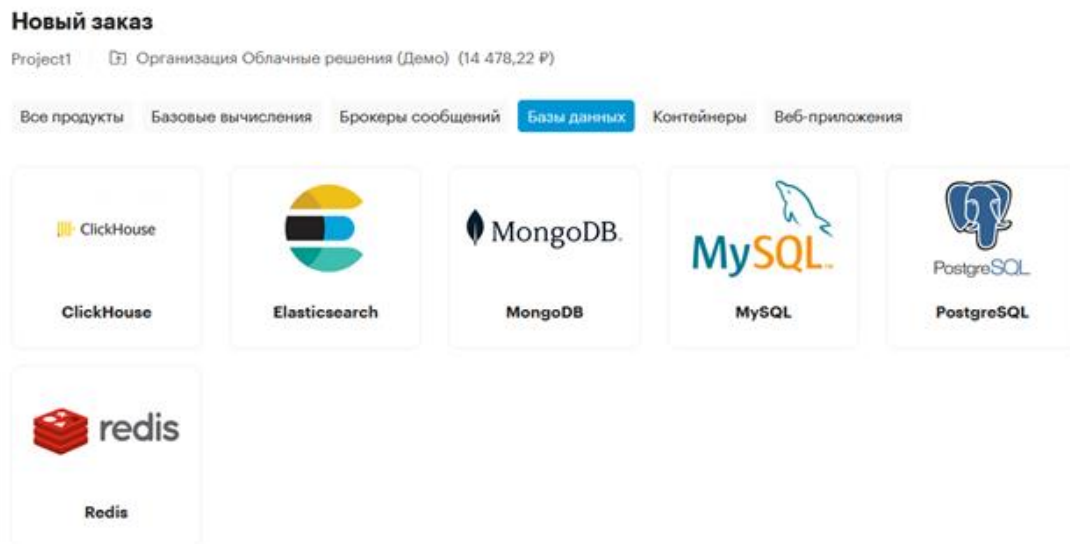
1.9 Базы данных

В разделе **Базы данных** можно заказать услуги этой категории и просмотреть информацию о состоянии инстансов баз данных. К созданным ресурсам можно применить сортировку по параметрам «Продукт», «Платформа», «Имя» и «CPU, RAM, Диск».

Для заказа услуги:

1. Перейдите в раздел **Базы данных** и нажмите **Заказать**.
2. Выберите базу данных:
 - Clickhouse
 - Elasticsearch
 - MongoDB
 - MySQL
 - PostgreSQL
 - Redis

Также заказать услуги всех категорий можно в разделе [Все ресурсы](#).



3. Укажите следующие параметры:
 - **Версия БД** — укажите версию БД или оставьте значение по умолчанию.
 - **Пароль пользователя БД** — введите пароль пользователя БД. Он должен состоять из 20-64 символов, включать цифры, прописные и строчные латинские буквы.
 - **Имя виртуальной машины** — укажите уникальное название ВМ.
 - **Сетевой сегмент** — выберите название логической портгруппы, из которой будет присвоение статического IP-адреса для виртуальной машины.
 - **Дата-центр** — выберите центр обработки данных.
 - **Платформа** — выберите платформу виртуализации.
 - **Образ ОС** — выберите один из образов операционной системы.

- Значения в полях «Сетевой сегмент», «Дата-центр», «Платформа», «CPU и RAM» создаются на основании справочников, которые модерируются пользователями с правами администратора.
 - От выбранной конфигурации может зависеть стоимость услуги.
 - **Точка монтирования** (опционально) — заполните параметры для подключения дополнительных дисков:
 - **размер, Гб** — укажите размер дополнительного диска.
 - **Путь монтирования** — укажите каталог для монтирования дополнительного диска.
 - **Тип файловой системы** — по умолчанию файловая система xfs.
 - **SSH-ключи** — выберите имеющиеся ключи доступа или создайте новый ключ, нажав + **Добавить SSH-ключ**.
4. Проверьте параметры и нажмите кнопку **Заказать**. Если не заполнены все обязательные поля, отмеченные знаком *, кнопка **Заказать** будет неактивна.
-

← Новый заказ

ClickHouse | Базовые вычисления

Сервер ClickHouse

Имя виртуальной машины

clickhouse-0263

Имя пользователя
clickhouse

clickhouse_admin

Версия Clickhouse

23.5.3.24

В пароле обязательно использовать цифры, прописные и строчные латинские буквы. Длина от 20 до 64 символов.

Пароль пользователя
clickhouse

••••••••••••••••••••

Основной сетевой интерфейс

Сетевой сегмент

DEV_DEV (Обезличенные данные)

Дата-центр

cod-a

Платформа

Образ ОС

ubuntu-template

CPU и RAM

2 CPU

4 CPU

8 CPU

16 CPU

2 GB

4 GB

8 GB

12 GB

16 GB

24 GB

32 GB

48 GB

64 GB

Загрузочный диск, Гб

30

Процессор vsphere в типе
среды PROD**3,56 P**
2 X 3,56 PОперативная память для
vsphere**4,91 P**
4 X 4,91 PЖесткий диск для vsphere в
типе среды PROD**3,12 P**
30 X 3,12 P

Итого в день

11,58 P

Точка монтирования на диске

Создайте каталог, который будет использоваться в качестве точки монтирования для диска виртуальной машины, и укажите его размер. Адреса каталогов не должны повторяться. Можно менять тип файловой системы, если он не установлен по умолчанию.

Добавить

Доступ

Выберите SSH-ключ из уже имеющихся или создайте новый

key-demo (key-demo количество: 1) ×

⊕

1.10 Брокеры сообщений

В разделе **Брокеры сообщений** можно заказать услуги этой категории и просмотреть информацию о состоянии инстансов для обмена сообщениями между различными компонентами и приложениями. К созданным ресурсам можно применить сортировку по параметрам «Продукт», «Платформа», «Имя» и «CPU, RAM, Диск».

Для заказа услуги:

1. Перейдите в раздел **Брокеры сообщений** и нажмите **Заказать**.
2. Выберите брокер сообщений:
 - Kafka
 - RabbitMQ

Новый заказ

Project1 |  Организация Облачные решения (Демо) (14 472,60 ₽)

Все продукты

Базовые вычисления

Брокеры сообщений

Базы данных

Контейнеры

Веб-приложения



Apache Kafka



RabbitMQ

Также заказать услуги всех категорий можно в разделе [Все ресурсы](#).

3. Укажите следующие параметры:
 - **Версия брокера сообщений** — выберите подходящую версию или оставьте по умолчанию.
 - **Пароль пользователя** — введите пароль пользователя. Он должен состоять из 20-64 символов, включать цифры, прописные и строчные латинские буквы.
 - **Имя виртуальной машины** — укажите уникальное название VM.
 - **Сетевой сегмент** — выберите название логической портгруппы, из которой будет присвоение статического IP-адреса для виртуальной машины.
 - **Дата-центр** — выберите центр обработки данных.
 - **Платформа** — выберите платформу виртуализации.
 - **Образ ОС** — выберите один из образов операционной системы.
- Значения в полях «Сетевой сегмент», «Дата-центр», «Платформа», «CPU и RAM» создаются на основании справочников, которые модерируются пользователями с правами администратора.
 - От выбранной конфигурации может зависеть стоимость услуги.
 - **Точка монтирования** (опционально) — заполните параметры для подключения дополнительных дисков:
 - **размер, Гб** — укажите размер дополнительного диска.
 - **Путь монтирования** — укажите каталог для монтирования дополнительного диска.
 - **Тип файловой системы** — по умолчанию файловая система xfs.
 - **SSH-ключи** — выберите имеющиеся ключи доступа или создайте новый ключ, нажав + **Добавить SSH-ключ**.

4. Проверьте параметры и нажмите кнопку **Заказать**. Если не заполнены все обязательные поля (отмеченные знаком *), кнопка **Заказать** будет неактивна.

← **Новый заказ**

Apache Kafka
Базовые вычисления

Сервер Apache Kafka

Имя виртуальной машины

Имя пользователя kafka

Версия Apache Kafka

Пароль пользователя kafka

kafka-test

kafka_admin

3.4.0

В пароле обязательно использовать цифры, прописные и строчные латинские буквы. Длина от 20 до 64 символов.

••••••••••••••••••••

🔑

🔄

📋

Процессор vsphere в типе среды PROD

Оперативная память для vsphere

Жесткий диск для vsphere в типе среды PROD

Итого в день

3,56 Р

4,91 Р

3,12 Р

11,58 Р

2 X 3,56 Р

4 X 4,91 Р

30 X 3,12 Р

▼

Основной сетевой интерфейс

Сетевой сегмент

Дата-центр

Платформа

Образ ОС

CPU и RAM

Загрузочный диск, Гб

DEV_DEV (Обезличенные данные)

cod-a

ubuntu-template

2 CPU

4 CPU

8 CPU

16 CPU

2 GB

4 GB

8 GB

12 GB

16 GB

24 GB

32 GB

48 GB

64 GB

30

Точка монтирования на диске

Создайте каталог, который будет использоваться в качестве точки монтирования для диска виртуальной машины, и укажите его размер. Адреса каталогов не должны повторяться. Можно менять тип файловой системы, если он не установлен по умолчанию.

Добавить

Доступ

Выберите SSH-ключ из уже имеющихся или создайте новый

key-demo (key-demo количество: 1) ×

⊕

1.11 Веб-приложения

В разделе **Веб-приложения** можно заказать услуги этой категории и просмотреть информацию о состоянии инстансов веб-приложений. К созданным ресурсам можно применить сортировку по параметрам «Продукт», «Платформа», «Имя» и «CPU, RAM, Диск».

Для заказа услуги:

1. Перейдите в раздел **Веб-приложения** и нажмите **Заказать**.

2. Выберите веб-приложение:

- Nginx
- WildFly

Новый заказ

Project1 | 📁 Организация Облачные решения (Демо) (14 471,19 ₽)

Все продукты

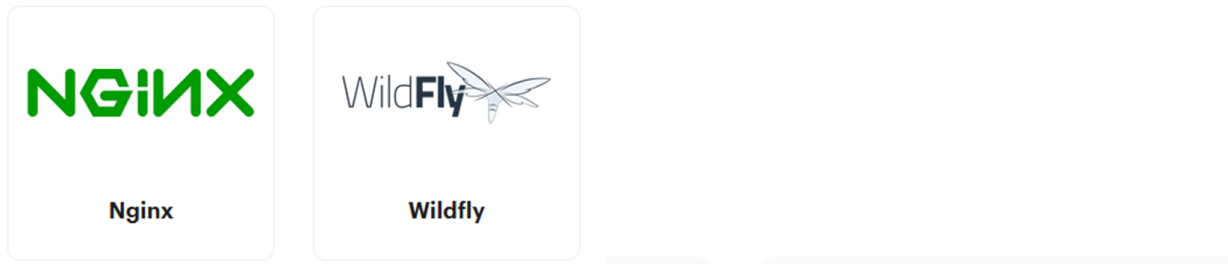
Базовые вычисления

Брокеры сообщений

Базы данных

Контейнеры

Веб-приложения



1.11.1 Заказать NGINX

NGINX — это веб-сервер и почтовый прокси, который работает под управлением операционных систем семейства Linux/Unix и Microsoft.

Для заказа:

1. Нажмите на карточку услуги NGINX и укажите параметры:
 - a. **Версия nginx** — выберите версию веб-приложения.
 - b. **Имя виртуальной машины** — задайте уникальное название виртуальной машины.
 - c. **Сетевой сегмент** — оставьте по умолчанию.
 - d. **Дата-центр** — выберите центр обработки данных.
 - e. **Платформа** — выберите платформу виртуализации.
 - f. **CPU и RAM** — выберите спецификацию VM, где core — размер vCPU, а RAM — размер оперативной памяти (в ГБ).
 - g. **Образ ОС** — выберите образ операционной системы.
- Значения в полях «Сетевой сегмент», «Дата-центр», «Платформа», «CPU и RAM» создаются на основании справочников, которые модерируются пользователями с правами администратора.
 - От выбранной конфигурации может зависеть стоимость услуги.
- h. **Загрузочный диск** — укажите размер загрузочного диска в ГБ или оставьте по умолчанию.
 - i. **Точка монтирования** (опционально) — заполните параметры для подключения дополнительных дисков:
 - **размер, Гб** — укажите размер дополнительного диска.
 - **Путь монтирования** — укажите каталог для монтирования дополнительного диска.
 - **Тип файловой системы** — по умолчанию файловая система xfs.

- j. **SSH-ключи** — выберите имеющиеся ключи доступа или создайте новый ключ, нажав + **Добавить SSH-ключ**.
2. Проверьте параметры и нажмите кнопку **Заказать**. Если не заполнены все обязательные поля, отмеченные знаком *, кнопка **Заказать** будет неактивна.

Сервер Nginx

Имя виртуальной машины nginx-test

Версия nginx 1.23.4

Основной сетевой интерфейс

Сетевой сегмент DEV_DEV (Обезличенные данные)

Дата-центр cod-a

Платформа

Образ ОС ubuntu-template

CPU и RAM

2 CPU 4 CPU 8 CPU 16 CPU

2 GB 4 GB 8 GB 12 GB 16 GB 24 GB

32 GB 48 GB 64 GB

Загрузочный диск, Гб 30

Процессор vsphere в типе среды PROD 3,56 Р 2 X 3,56 Р

Оперативная память для vsphere 4,91 Р 4 X 4,91 Р

Жесткий диск для vsphere в типе среды PROD 3,12 Р 30 X 3,12 Р

Итого в день 11,58 Р

Точка монтирования на диске

Создайте каталог, который будет использоваться в качестве точки монтирования для диска виртуальной машины, и укажите его размер. Адреса каталогов не должны повторяться. Можно менять тип файловой системы, если он не установлен по умолчанию.

Добавить

Доступ

Выберите SSH-ключ из уже имеющихся или создайте новый

key-demo (key-demo количество: 1) ×

+

Заказать

1.11.2 Заказать Wildfly

WildFly (ранее JBoss Application Server или JBoss AS) — это кроссплатформенный сервер приложений с открытым исходным кодом, разработанный компанией RedHat.

Для заказа:

1. Нажмите на карточку услуги Wildfly и укажите параметры:
 - a. **Версия Wildfly** — выберите версию веб-приложения.
 - b. **Имя пользователя Wildfly** — введите имя пользователя. Оно должно содержать только латинские буквы, символы не поддерживаются.
 - c. **Пароль пользователя Wildfly** — введите пароль пользователя. Он должен состоять из 20-64 символов, включать цифры, прописные и строчные латинские буквы.
 - d. **Разрешить внешний доступ к консоли администрирования** (опционально) — активируйте эту опцию, если нужен доступ к консоли администратора другим пользователям. Включение этого параметра может увеличить риски безопасности, так как позволяет удаленный доступ к панели управления WildFly.
 - e. **Имя виртуальной машины** — задайте уникальное название для виртуальной машины.
 - f. **Сетевой сегмент** — выберите название логической портгруппы, из которой будет присвоение статического IP-адреса для виртуальной машины.
 - g. **Дата-центр** — выберите центр обработки данных.
 - h. **Платформа** — выберите платформу виртуализации.
 - i. **CPU и RAM** — выберите спецификацию ВМ, где core — размер vCPU, а RAM — размер оперативной памяти (в ГБ).
 - j. **Образ ОС** — выберите образ операционной системы.

- Значения в полях «Сетевой сегмент», «Дата-центр», «Платформа», «CPU и RAM» создаются на основании справочников, которые модерируются пользователями с правами администратора.
 - От выбранной конфигурации может зависеть стоимость услуги.

 - k. **Загрузочный диск** — укажите размер загрузочного диска в ГБ или оставьте по умолчанию.
 - l. **Точка монтирования** (опционально) — заполните параметры для подключения дополнительных дисков:
 - **размер, Гб** — укажите размер дополнительного диска.
 - **Путь монтирования** — укажите каталог для монтирования дополнительного диска.
 - **Тип файловой системы** — по умолчанию файловая система xfs.
 - m. **SSH-ключи** — выберите имеющиеся ключи доступа или создайте новый ключ, нажав + **Добавить SSH-ключ**.
2. Проверьте параметры и нажмите кнопку **Заказать**. Если не заполнены все обязательные поля, отмеченные знаком *, кнопка **Заказать** будет неактивна.

← Новый заказ

Wildfly | Базовые вычисления

Сервер Wildfly

Версия Wildfly

27.0.0

Имя пользователя Wildfly

admin

Пароль пользователя Wildfly

В пароле обязательно использовать цифры, прописные и строчные латинские буквы. Длина от 20 до 64 символов.

••••••••••••••••••••

☐ Разрешить внешний доступ к консоли администрирования

Пожалуйста, имейте в виду, что включение этого параметра может увеличить риски безопасности, позволяя удаленный доступ к панели управления Wildfly.

Имя виртуальной машины

wildfly-0227

Основной сетевой интерфейс

Сетевой сегмент

DEV_DEV (Обезличенные данные)

Дата-центр

cod-a

Платформа

CPU и RAM

2 CPU 4 CPU 8 CPU 16 CPU

2 GB 4 GB 8 GB 12 GB 16 GB 24 GB

32 GB 48 GB 64 GB

Образ ОС

ubuntu-template

Загрузочный диск, Гб

30

Процессор vsphere в типе среды PROD

3,56 ₽

2 X 3,56 ₽

Оперативная память для vsphere

4,91 ₽

4 X 4,91 ₽

Жесткий диск для vsphere в типе среды PROD

3,12 ₽

30 X 3,12 ₽

Итого в день

11,58 ₽

Точка монтирования на диске

Создайте каталог, который будет использоваться в качестве точки монтирования для диска виртуальной машины, и укажите его размер. Адреса каталогов не должны повторяться. Можно менять тип файловой системы, если он не установлен по умолчанию.

Добавить

Доступ

Выберите SSH-ключ из уже имеющихся или создайте новый

key-demo (key-demo количество: 1) ×

1.12 Список VM

В этом разделе можно просмотреть сведения по всем созданным виртуальным машинам.

Здесь нет сведений о заказанных услугах в целом, как в разделе **Все ресурсы**. В разделе **Список всех ВМ** можно отследить состояние каждой виртуальной машины, даже если их несколько в рамках одной услуги (например, для сервиса Kubernetes создается как минимум 2 виртуальных машины: одна — для master-ноды, другая — для worker-ноды).

Также в этом разделе можно выполнить следующие действия с ВМ:

- **Включить/Выключить**
- **Выключить принудительно**
- **Изменение конфигурации ВМ**
- **Снять ВМ с мониторинга/Поставить ВМ на мониторинг**
- **Скопировать ID**

Для выполнения

1. Выберите виртуальную машину.
2. Нажмите на кнопку **:** и выберите действие.

2.1.12 Управление доступом

1.12.1 Организация

Организация — это иерархический список каталогов (папок/проектов). Используется для разграничения проектных и линейных активностей

В **Организации** пользователь может создавать папки и проекты.

Папка — логическое объединение папок и проектов (например, папка «Департамент розничного бизнеса»). Папка обладает свойством бюджетных лимитов, на эту сущность можно заложить сумму, за пределы которой владельцы пространства не могут выйти.

Проект — это сущность, которая содержит в себе инфраструктурные объекты (виртуальные машины, платформенные сервисы и т.д.). Заказать услуги можно только на уровне проекта.

Создание папок в структуре **Организации** нужно для разграничения прав доступа пользователей к проектам. Рекомендуется создавать папки разного уровня вложенности в соответствии с активностями организационных подразделений или групп пользователей из разных подразделений, работающих в рамках одного проекта.

Структура **Организации** имеет следующие ограничения:

- уровней вложенных папок — не более **8**;
- папок на одном уровне — не более **50**;
- проектов на одном уровне — не более **15**.

Выбрать организацию, папку или проект:

1. Перейдите в раздел **Управление доступом** → **Организация**.
2. Напротив папки/проекта нажмите на кнопку **:**.
3. Нажмите **Выбрать контекст**.

Управление доступом

proj-develop-dev | Папка fold-develop

Организация | Сервисные аккаунты | Учётные записи | Роли | SSH ключи

- ▼ develop 13098757516.20 ₺
 - develop-default-folder 13098757516.20 ₺
 - ▼ fold-bb-develop 0.00 ₺
 - ▼ fold-dept-develop -73.30 ₺
 - ▼ fold-develop 947620.85 ₺
 - Тест 0.00 ₺
 - proj-develop-dev
 - proj-develop-ift
 - proj-develop-prod
 - 1234
 - integrations 13098757516.20 ₺

1.12.2 Настройка организационной структуры

Создание и настройка организационной структуры доступны только пользователю с соответствующими правами.

1.12.2.1 Создание папки/проекта

1. Перейдите в раздел **Управление доступом** → **Организация**.
2. Напротив организации или папки нажмите **:**.
3. Нажмите **Создать папку** или **Создать проект**.

Управление доступом

proj-develop-dev | Папка fold-develop

Организация | Сервисные аккаунты | Учётные записи | Роли | SSH ключи

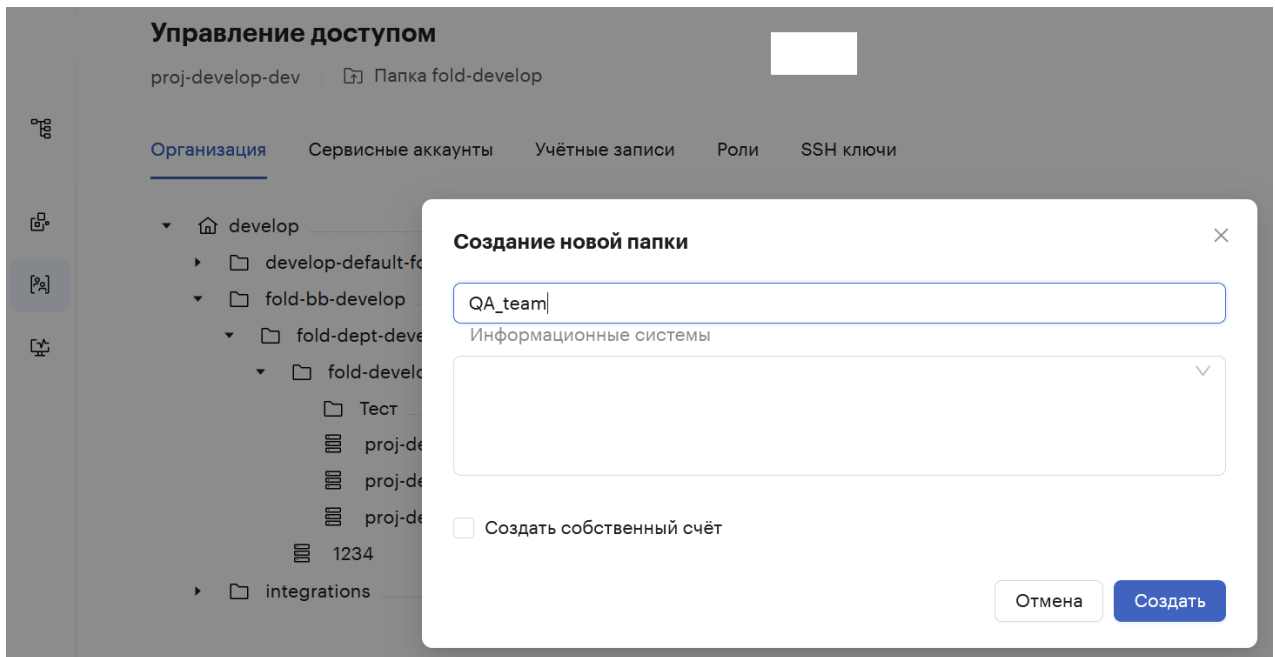
- ▼ develop 13098757516.20 ₽
 - ▶ develop-default-folder 13098757516.20 ₽
 - ▼ fold-bb-develop 0.00 ₽
 - ▼ fold-dept-develop -73.30 ₽
 - ▼ fold-develop 947620.85 ₽
 - Тест 0.00 ₽
 - proj-1
 - proj-2
 - proj-3
 - 1234
 - ▶ integrations 13098757516.20 ₽

Перевести
 Создать папку
 Создать проект
 История переводов
 Переименовать
 Удалить

4. Введите параметры:

Для новой папки:

- Введите **Название**.
- При необходимости отключите опцию **Создать собственный счет**.
- Нажмите **Сохранить и продолжить**.
- Выберите одну или более **Информационные системы** из списка.
- Нажмите **Создать**.



- Для нового проекта:
 - f. Введите **Наименование**.
 - g. Из списка выберите **Информационную систему**.
 - h. При создании проекта из списка выберите **Среду**.
 - i. Нажмите **Создать**.

Созданная папка/проект отобразится в разделе **Организация**.

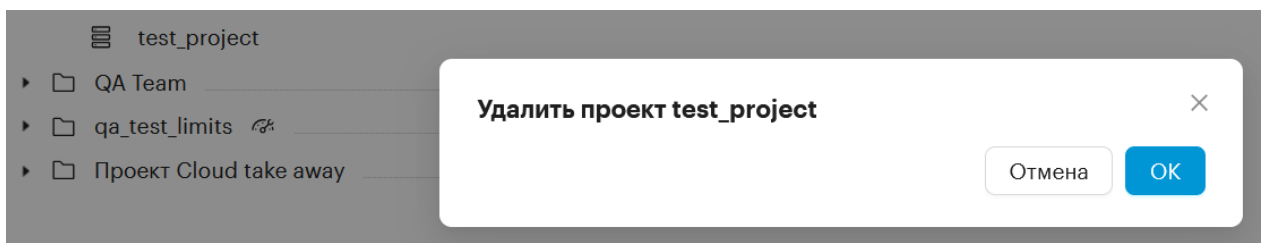
1.12.2.2 Удаление папки/проекта

Пользователь с соответствующими правами может удалить папку или проект.

Перед удалением папки с дочерними элементами (папки/проекты), нужно предварительно удалить эти дочерние элементы.

Для удаления папки/проекта:

1. Перейдите в раздел **Управление доступом** → **Организация**.
2. Нажмите на кнопку **:** напротив выбранной папки/проекта.
3. В списке выберите пункт **Удалить**.
4. Подтвердите операцию удаления, нажав **ОК**.



Вместе с папкой безвозвратно удалится всё содержимое. Средства перераспределяются в родительскую папку. image::delete-folder1.png[role=center]

1.13 Управление ресурсными квотами

Создание ресурсных квот доступно при создании организации и на странице уже существующей организации на портале Control-panel. Добавление ресурсных квот для конкретного проекта осуществляется через пользовательский KageCore ML Platform.

1.13.1 Добавление ресурсных квот для папки/проекта

Добавление ресурсных квот для конкретного проекта осуществляется в пользовательском KageCore ML Platform через учётную запись пользователя с ролью **Администратор Портала**.

Необходимо учесть, что:

- При добавлении квоты на вышестоящий элемент оргструктуры, выполняется проверка на наличие других квот нижестоящих элементов.
То есть если есть квота на **Проект** имеет значение CPU 100, то для **Папки**, где расположен данный проект, можно будет создать квоту ≥ 100 .
- При добавлении квоты на нижестоящий элемент оргструктуры, выполняется проверка на наличие других квот вышестоящих элементов.
То есть если есть квота на **Папку** имеет значение RAM 200, то для **Проекта**, расположенного в данной папке, можно будет создать квоту ≤ 200 .
- Наличие уже созданных квот для других элементов.
Если на **Папку** стоит квота со значением CPU 300, внутри находится 2 **Проекта** и в одном Проекте указана квота CPU имеет значение 100, то во втором Проекте можно указать значение CPU ≤ 200 .
- При выставлении квот учитываются уже созданные заказы.

Порядок действий

1. Перейдите в раздел **Управление доступом** → **Организация**.
2. Напротив папки или проекта нажмите **:** и выберите пункт **Создать квоту**
3. В окне **Назначение ресурсных квот**
 - a. Выберите платформу
 - b. Настройте значения ресурсов
4. Нажмите **Сохранить**
5. Установленные ресурсные квоты можно отобразить, нажав на значок **Ресурсные квоты** напротив Организации, Папки или проекта.

1.13.2 Редактирование ресурсных квот для папки/проекта

Порядок действий

1. Перейдите в раздел **Управление доступом** → **Организация**.
2. Нажмите на значок **Ресурсные квоты** напротив Папки или Проекта.
3. В открывшемся боковом меню нажмите на  напротив имени ресурсной квоты.

4. В окне **Изменение ресурсных квот** настройте необходимые параметры CPU, RAM и Дисков.
5. Нажмите **Сохранить**.

1.13.3 Удаление ресурсных квот для папки/проекта

Порядок действий

1. Перейдите в раздел **Управление доступом** → **Организация**.
2. Нажмите на значок **Ресурсные квоты** напротив Папки или Проекта.

В пользовательском KageCore ML Platform ресурсные квоты Организации удалить нельзя.

3. В открывшемся боковом меню нажмите на  напротив имени ресурсной квоты.
4. В окне **Удалить квоты** нажмите на **Да,удалить**.

1.14 Управление счетами

Пользователи с соответствующими правами на KageCore ML Platform могут управлять счетами папок: переводить денежные средства со счетов родительских папок на дочерние счета.

Перевести средства можно только из **организации** или **папки**. Из **проектов** перевод недоступен.

Порядок действий

1. Перейдите в раздел **Управление доступом** → **Организация**.
2. Нажмите на иконку  напротив папки или организации.
3. В списке выберите пункт **Перевести**.
4. В открывшемся окне заполните поля:
 - **Счет-отправитель** — по умолчанию указан тот счет, с которого отправляются средства.
 - **Счет-получатель** — выберите организацию или папку, в которую нужно отправить средства.
 - **Сумма** — введите сумму для перевода, не превышающую значения поля **Счет-отправитель**.
 - **Описание операции** — введите текст, характеризующий этот перевод средств.
5. Нажмите **Подтвердить действие**.

Для просмотра истории переводов средств нажмите  напротив необходимой папки и выберите — **История переводов**.

1.15 Сервисные аккаунты

Для управления пользователями на KageCore ML Platform применяется ролевая политика по модели RBAC (Role Based Access Control). Пользователь может выполнять те или иные действия в зависимости от присвоенной ему роли.

В этом разделе можно создать, изменить, удалить сервисные аккаунты с определенными ролями, а также просмотреть список созданных сервисных аккаунтов.

1.15.1 Создать сервисный аккаунт

Для создания нового сервисного аккаунта:

1. Перейдите в раздел **Управление доступом** → **Сервисные аккаунты**.
2. Нажмите кнопку + (Добавить).
3. В появившемся окне заполните параметры:
 - **Название**
 - **Роли в каталоге** — выберите одну или несколько ролей.
4. Нажмите **Создать**

После успешного выполнения действия отображается уведомление о создании сервисного аккаунта, содержащее его идентификатор и API-ключ.

Идентификатор сервисного аккаунта и API-ключ необходимо скопировать, после закрытия окна уведомления данные будут недоступны.

1.15.2 Изменить сервисный аккаунт

Для изменения параметров сервисного аккаунта:

1. Перейдите в раздел **Управление доступом** → **Сервисные аккаунты**.
2. Напротив сервисного нажмите **:** и выберите **Редактировать**.
3. Измените параметры:
 - **Название** — введите новое название.
 - **Роли в каталоге** - можно удалить существующие или добавить новые из списка.
4. Нажмите **Применить**.

1.15.3 Удалить сервисный аккаунт

Для удаления сервисного аккаунта:

1. Перейдите в раздел **Управление доступом** → **Сервисные аккаунты**.
 2. Напротив сервисного аккаунта нажмите **:** и выберите **Удалить аккаунт**.
 3. Нажмите **Да,удалить**.
-

1.16 Учетные записи

В разделе **Учетные записи** можно просмотреть существующих пользователей KageCore ML Platform и добавить новых.

Можно фильтровать список пользователей при работе с фильтром, заполнив поля **Поиск** и/или **Роли**.

Также в выпадающем списке **Показать пользователей** можно выбрать:

- **Текущего уровня** — отобразятся записи, входящие в текущий проект.
- **Текущего уровня и ниже** — отобразятся записи, входящие в текущую папку и её дочерние папки и проекты.
- **Текущего уровня и выше** — отобразятся записи, входящие в текущую папку/проект и родительские папки и организацию.

1.16.1 Добавить пользователя к организации/папке/проекту

В разделе **Учетные записи** можно добавить пользователя к нужному контексту. Для этого:

1. Перейдите в раздел **Управление доступом** → **Учетные записи**
2. Нажмите кнопку **+Добавить**.
3. В появившемся окне:
 - Выберите одного или нескольких пользователей в поле **Пользователь**. Для поиска пользователя начните набирать его имя.
 - В поле **Роли** из выпадающего меню выберите роли, которые хотите назначить пользователю для этой организации/папки/проекта.

1.16.2 Редактировать права пользователя

Права пользователя в контексте можно изменить. Для их редактирования:

1. Активируйте необходимый контекст соответствующей организации.
2. Перейдите на экран **Управление доступом** → **Учетные записи**.
3. Напротив нужного пользователя в списке нажмите **:** и выберите **Редактировать**.
4. В появившемся окне добавьте и/или удалите необходимые роли.
5. Нажмите **Применить**.

1.16.3 Отозвать права пользователя

1. Активируйте необходимый контекст соответствующей организации
2. Перейдите в раздел **Управление доступом** → **Учетные записи**.
3. В строке с нужным пользователем нажмите **:** и выберите **Отозвать права**.

После выполнения операции пользователь будет удалён из списка пользователей организационной единицы.

1.17 Роли

Раздел **Роли** позволяет просматривать роли и соответствующий им набор разрешений, а также создавать новые роли в рамках организации

С помощью ролей можно гибко регулировать доступы пользователей к ресурсам KageCore ML Platform, что обеспечивает защиту корпоративных данных.

В этом разделе можно увидеть роли, созданные системой (тип — "service" или "basic"), так и созданные пользователями (тип — "custom").

Просмотр и создание ролей в проекте доступно только пользователям с соответствующими правами.

1.17.1 Создать новую роль

Для создания новой роли:

1. Активируйте необходимый контекст соответствующей организации.
2. Перейдите в раздел **Управление доступом** → **Роли**.
3. В правом верхнем углу нажмите кнопку + (Создать).
 - а. В окне **Создание роли** заполните следующие параметры:
 - ***Название** — введите название роли. Это название может быть любым, но в нем должны использоваться только латинские буквы и цифры.
 - **Код роли** — введите код роли. В качестве кода для сервисных ролей используется ключевое слово **service**, для базовых ролей используется ключевое слово **basic**, все остальные роли в системе будут обозначены ключевым словом **custom**.
 - **Описание** — опишите, для чего эта роль применяется.
 - **Выберите права из списка** — выберите одно или несколько разрешений для этой роли.

Выбранные разрешения для роли переместятся из левой колонки в правую колонку (Добавленные права).

- б. Нажмите **Создать роль** вверху страницы.

Новая роль появится в общем списке. Ее можно будет отредактировать, удалить или создать на ее основе другие роли.

О том, как применить роль к пользователю — в инструкции [Редактировать права пользователя](#).

1.17.2 Удалить роль

Удалить можно только те роли, которые создали сами пользователи (тип — "custom"). Роли с типом "service" или "basic" удалить нельзя.

Для удаления роли:

1. Перейдите в раздел **Управление доступом** → **Роли**.
2. Напротив роли нажмите **:** и выберите **Удалить**.

1.18 SSH-ключи

В этом разделе можно просмотреть уже существующие SSH-ключи и добавить новые.

SSH-ключи используются для идентификации клиента при подключении к серверу по зашифрованному сетевому протоколу SSH. Они представляют собой пару: публичный ключ — размещается на виртуальной машине, приватный ключ — сохраняется локально.

Публичный ключ можно предоставить любому пользователю, но только у вас должен быть доступ к вашему приватному ключу.

1.18.1 Добавить SSH-ключ

Для добавления SSH-ключа:

1. Перейдите в раздел **SSH-ключи**.
2. В правом верхнем углу нажмите кнопку **+ Добавить**.
3. Выберите тип ключа.
4. В поле **Название** введите название добавляемого ключа.
5. В поле **Логин** введите имя пользователя, для которого добавляется ключ.
6. Сгенерируйте пару ключей на клиентской машине:
 - o Linux/macOS
 - a. Откройте терминал и выполните команду:


```
ssh-keygen -t rsa
```

Если команда `ssh-keygen` не найдена, установите пакет `openssh`.
 - b. Укажите путь до файла, в который нужно сохранить ssh-ключ или оставьте по умолчанию:


```
Enter file in which to save the key (/home/user/.ssh/id_rsa):
```
 - c. Нажмите **Enter**.
 - d. Введите кодовую фразу (если фраза введена, она будет запрашиваться при обращении на сервер) или оставьте строку без нее и нажмите **Enter**.


```
Enter passphrase (empty for no passphrase):
```

В папке `.ssh` (по умолчанию) будут созданы два файла: `id_rsa` (приватный ключ) и `id_rsa.pub` (публичный ключ). В терминале отобразится отпечаток сгенерированного ключа:

```
Your identification has been saved in /home/user/.ssh/id_rsa.
Your public key has been saved in /home/user/.ssh/id_rsa.pub.
The key fingerprint is:
476:b2:a8:7f:08:b4:c0:af:81:25:7e:21:48:01:0e:98
user@localhost
The key's randomart image is:
+--[ RSA 2048 ]-----+
```

```

|+.o.      |
|ooE       |
|oo        |
|o.+..     |
|.+.+.. S . |
|...+ o +   |
|.o ....   |
|. . .     |
|...       |
+-----+

```

e. В терминале выполните следующую команду:

f. `cat ~/.ssh/id_rsa.pub`

В консоли отобразится сгенерированный SSH-ключ.

g. Скопируйте значение ключа и вставьте в поле ключа в окне **Добавить SSH-ключ** на портале.

o Windows

a. Загрузите клиент Putty [с официального сайта](#).

b. Распакуйте загруженный архив и запустите файл **puttygen.exe**.

c. Выберите тип ключа **SSH-2 RSA** и длину **2048** бит.

d. Нажмите кнопку **Generate**. Во время генерации водите курсором в пустой области окна (это нужно для создания псевдослучайности).

e. Сохраните сгенерированную пару ключей с помощью кнопок **Save public key** и **Save private key**.

f. Скопируйте сгенерированный публичный ключ (расширение pub) и вставьте его в поле ключа в окне **Добавить SSH-ключ** на портале.

7. В окне **Добавить SSH-ключ** на портале нажмите **Добавить**. Этот ключ сохранится в списке всех ssh-ключей на портале.

8. После добавления ключа выполните в терминале команду:

```
$ ssh root@[IP-адрес сервера]
```

После выполнения всех действий будет установлено соединение с сервером. Вводить пароль при этом не потребуется.

4. Нажмите **Добавить**

На странице **Учетные записи** отобразится новая запись.

Роли в каталоге создаются пользователями с правами администратора согласно ролевой модели организации.

1.19 Мониторинг

1.19.1 Метрики VM

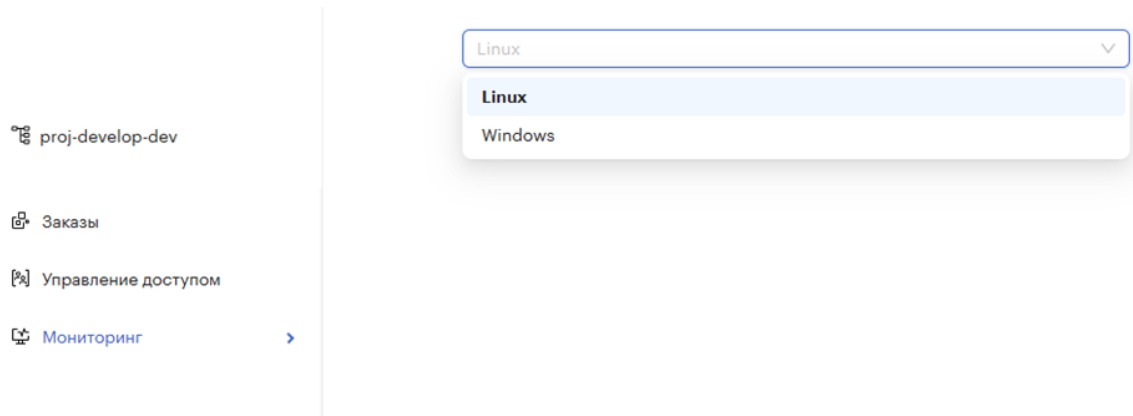
В разделе **Метрики VM** можно просмотреть графики потребления ресурсов по каждой виртуальной машине, основанные на VictoriaMetrics и визуализации в дашбордах Grafana.

Данные метрик ВМ обновляются автоматически каждые 30 минут по умолчанию (частоту обновлений метрик можно изменить в интерфейсе Grafana).

Для просмотра метрик ВМ:

1. Выберите операционную систему:

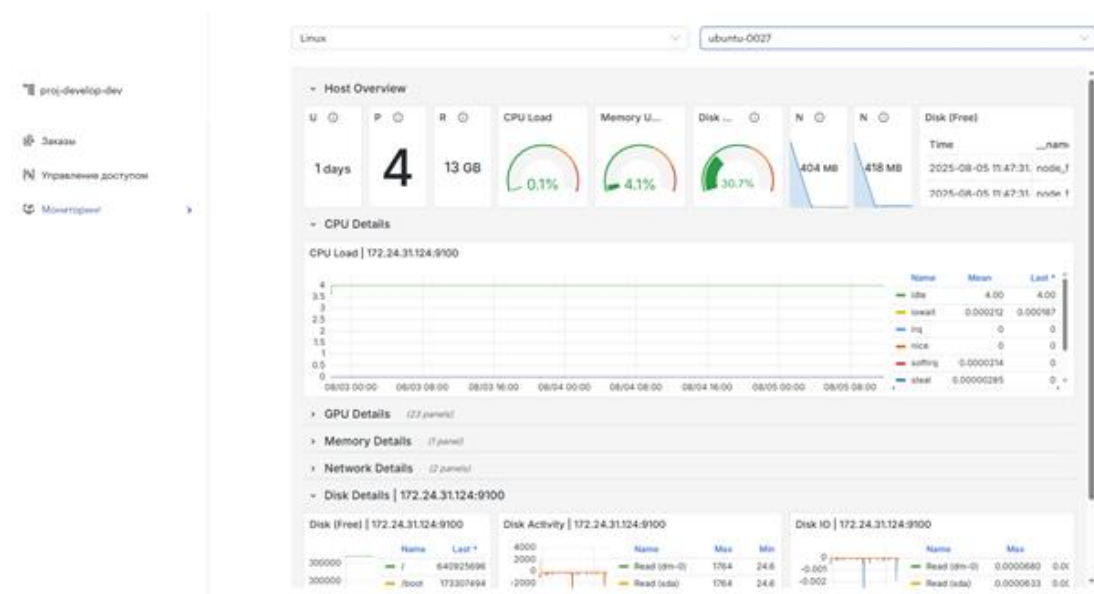
- Linux
- Windows



2. Выберите название виртуальной машины (**хостнейм**), для которой нужно отобразить данные мониторинга.

Набор дашбордов в мониторинге зависит от выбора операционной системы. Для ОС Windows и Linux этот набор отличается и учитывает особенности систем.

1.19.1.1 Метрики ВМ на ОС Linux



На странице доступны следующие дашборды:

1. **Host Overview:**

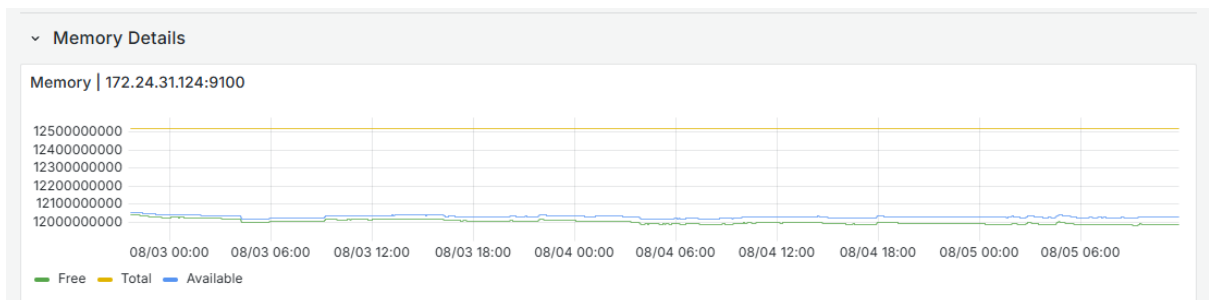
- **Uptime** — время нахождения виртуальной машины в активном состоянии с момента последнего запуска. Если ВМ была однажды перезапущена, то отсчет времени Uptime начнется заново.
- **Processors** — отображает количество процессоров на виртуальной машине.
- **RAM** — количество занятой оперативной памяти ВМ.
- **CPU Load** — загрузка процессора в процентном соотношении (%).
- **Memory use** — количество занятой виртуальной памяти в процентном соотношении (%).
- **Disk Free (Total)** — общее свободное место на дисках ВМ в процентном соотношении (%).
- **Net IN** — график количества принятых пакетов данных в ГБ (GB).
- **Net OUT** — график количества отправленных пакетов данных в ГБ (GB).

2. CPU Details

- **CPU Load** — отображение графика загрузки по системе по часам.

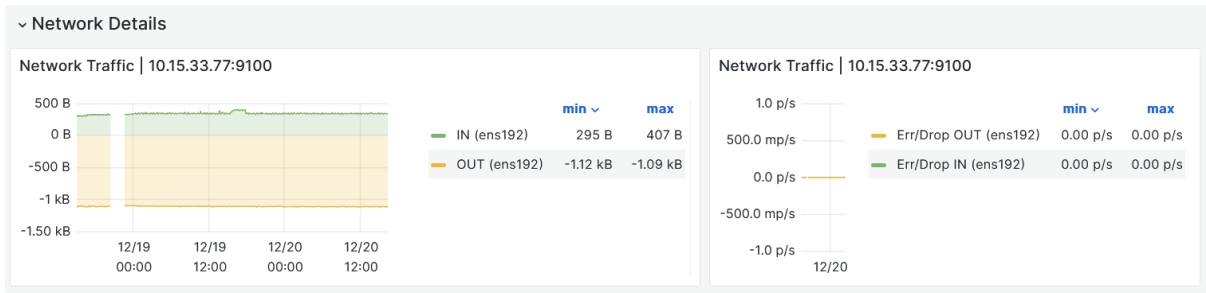
3. Memory details

- **Memory** — график использования памяти по ГБ/час.



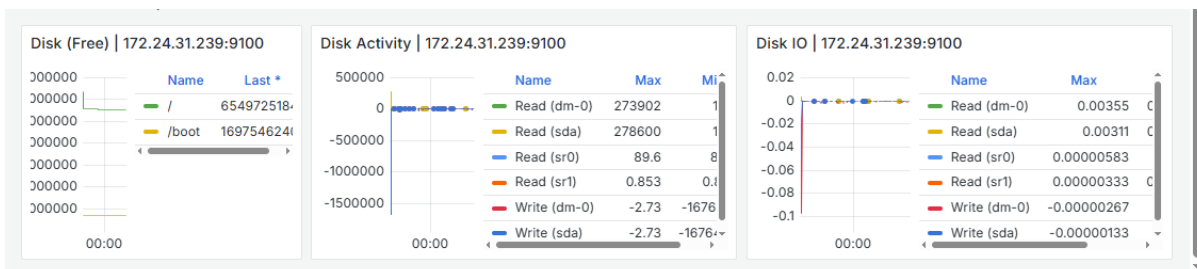
4. Network Details

- **Network Traffic** — почасовой график передачи и приема пакетов данных через интернет.
- **Network Traffic Err** — почасовой график возникновения ошибок при передаче и приеме данных через интернет.

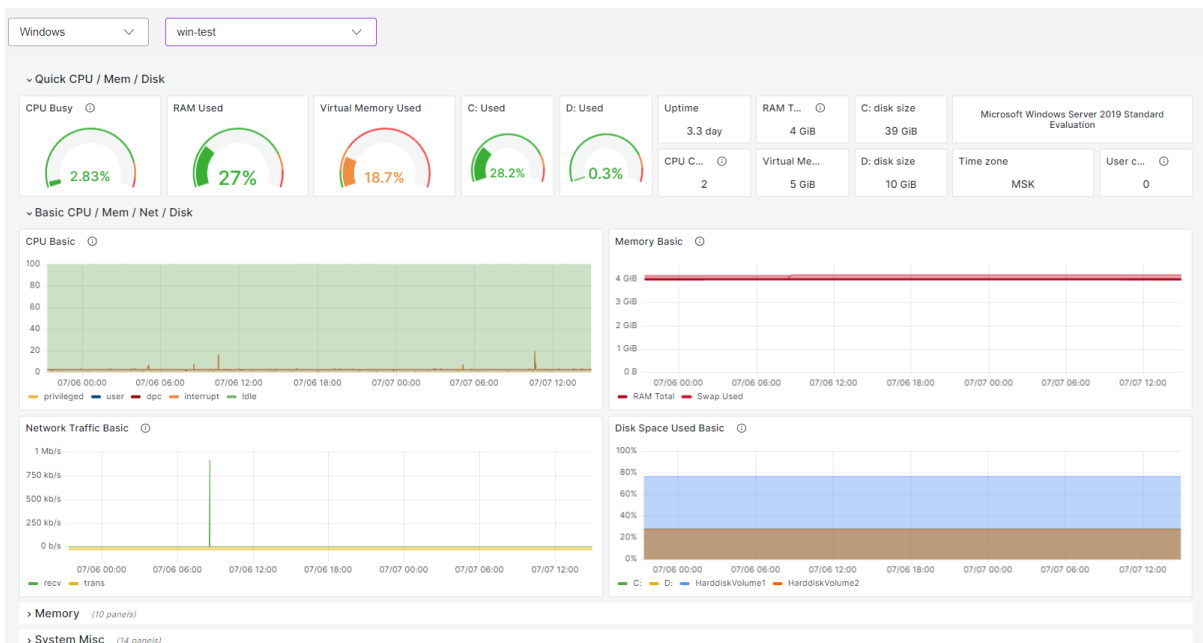


5. Disk Details

- **Disk Free** — количество памяти на диске.
- **Disk Activity** — количество операций чтения/записи.
- **Disk IO** — количество операций ввода/вывода.



1.19.1.2 Метрики VM на ОС Windows



На странице доступны следующие дашборды:

1. Quick CPU / Mem / Disk

- **CPU Busy** — загрузка процессора в процентном соотношении (%).

- **RAM Used** — количество занятой оперативной памяти ВМ в процентном соотношении (%).
 - **Virtual Memory Used** — количество занятой виртуальной памяти в процентном соотношении (%).
 - **C:Used** — отображает использование жесткого диска "C" за определенный период времени. Чтобы отобразить значения для этого графика, выделите промежуток времени на шкале **Disk Space Used Basic**.
 - **D:Used** — отображает использование жесткого диска "D" за определенный период времени. Чтобы отобразить значения для этого графика, выделите промежуток времени на шкале **Disk Space Used Basic**.
 - **Uptime** — время нахождения виртуальной машины в активном состоянии с момента последнего запуска. Если ВМ была однажды перезапущена, то отсчет времени Uptime начнется заново. Чтобы отобразить значения для этого графика, выделите промежуток времени на шкале **Disk Space Used Basic**.
 - **CPU Cores** — количество ядер CPU на виртуальной машине.
 - **RAM Total** — всего оперативной памяти на ВМ.
 - **Virtual Memory** — всего виртуальной памяти.
 - **C: disk size** — размер диска "C" (ГБ).
 - **D: disk size** — размер диска "D" (ГБ).
2. **Basic CPU / Mem / Net / Disk** — графики загрузки процессора, памяти, места на диске и использования сети за определенный промежуток времени. Промежуток времени можно выделить на шкале, чтобы посмотреть подробные данные.

Дашборд состоит из графиков:

- Pugins Files
 - Memory Basic
 - Standby Page List
 - Memory Committed
 - Paged/Non-Paged Pool
 - File System Cache
 - Page Fault Rates
-

- Swap page read/write
- Pageable OS system code + device drivers
- Paged/Non-Paged Pool - Allocation Calls

3. **System Misc** — группа показателей ошибок системы и системных вызовов.

Дашборд состоит из графиков:

- Context Switches / Interrupts
- Interrupts Detail
- Clock Tick Interrupts
- Idle breaks
- Received and serviced deferred procedure calls (DPCs)
- Time spent in low-power idle states
- Parking Status
- Core Frequency
- Exceptions
- Processor Queue
- File Descriptors
- CPU time spent in user and system contexts
- Processes & Threads
- System Calls

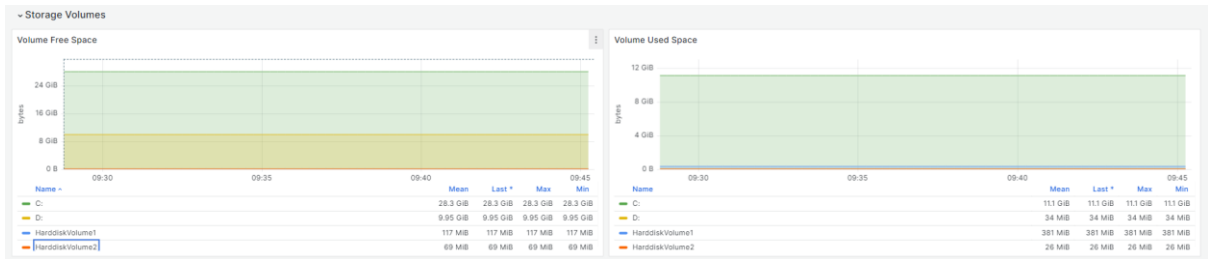
4. **System Timesync** — группа метрик, отражающих состояние синхронизации времени в системе

- Time Synchronized Drift
- Time Adjustment
- NTP Server Requests
- Time Misc

5. **Storage Volumes** — группа метрик заполненности хранилища. Для просмотра одного из дисков нажмите на его название в легенде под графиком.

Дашборд состоит из графиков:

- Volume Free Space
 - Volume Used Space
-



6. **Storage IO** — группа метрик производительности хранилища и дисковой системы, расшифровывается «Input/output Operations Per Second» — количество операций ввода/вывода в секунду.

Дашборд состоит из графиков:

- Disk IOps
- I/O Usage Read / Write Bytes
- I/O Utilization
- Average Queue Size
- Time Spent Doing I/Os
- Idle I/O
- IO Splits

7. **Network Traffic** — группа метрик скорости интернет-трафика, приема и передачи пакетов данных.

Дашборд состоит из графиков:

- Network Traffic
- Network Traffic by Packets
- Network Traffic Drop*
- Network Traffic Errors
- Speed
- Unknown Packets Received

8. **Network Netstat** — группа сетевых метрик, связанных с состоянием и качеством TCP-соединений.

Дашборд состоит из графиков:

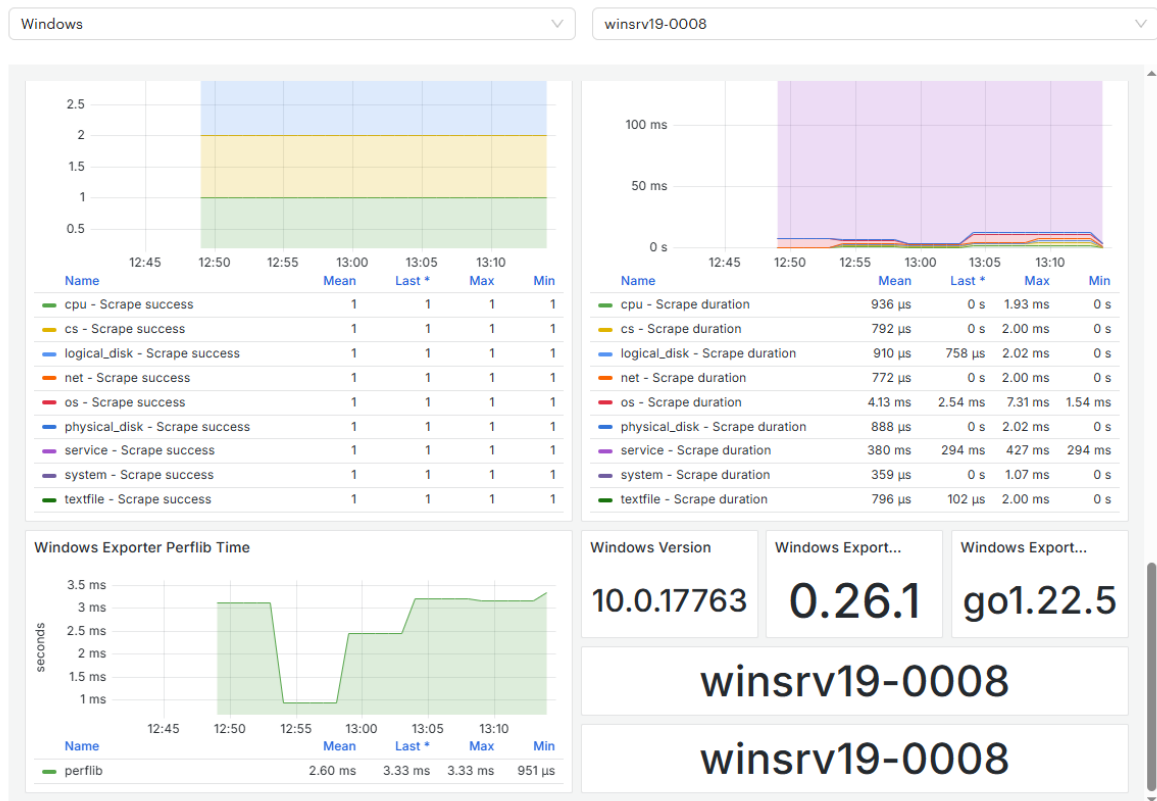
- TCP Connections
- TCP Retransmission

- TCP IPv4
- TCP IPv6
- TCP Direct Transition IPv4
- TCP Direct Transition IPv6
- TCP Connection Failures

9. **Windows Exporter** — группа метрик с серверов Windows-инфраструктуры.

Дашборд состоит из графиков:

- Exporter Scrape
- Windows Exporter Scrape Time
- Windows Exporter Perflib Time. Также здесь можно увидеть версию Windows, версию Exporter.



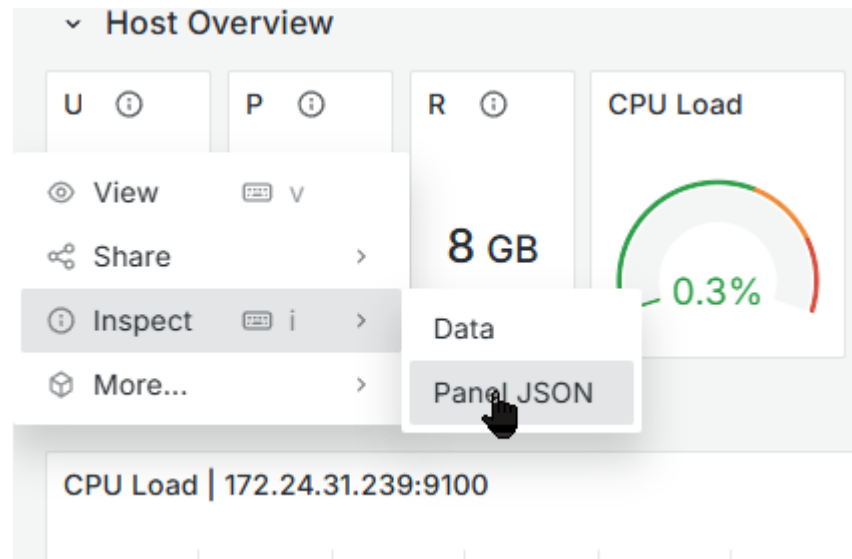
1.19.1.3 Импорт графиков

Каждый дашборд в **Метриках ВМ** можно импортировать, так как он основан на языке программирования JSON.

Для импорта:

1. Нажмите сверху нужного графика на ⋮ (Menu).

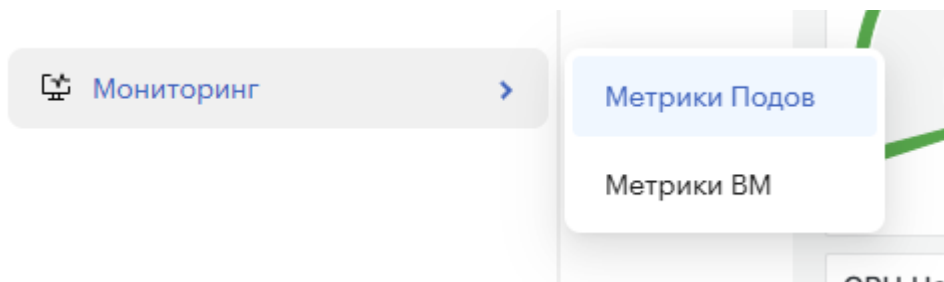
2. Выберите **Inspect** → **Panel JSON**.



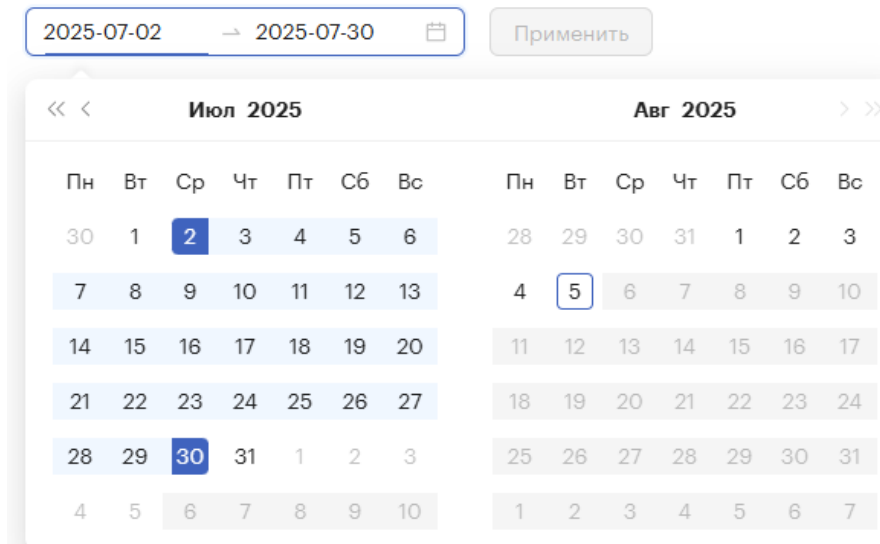
1.19.2 Метрики подов

В разделе **Метрики подов** можно просмотреть график потребления ресурсов в кластере Kubernetes подами, за определенный промежуток времени, основанный на мониторинге внутри кластера и визуализации в дашбордах Grafana.

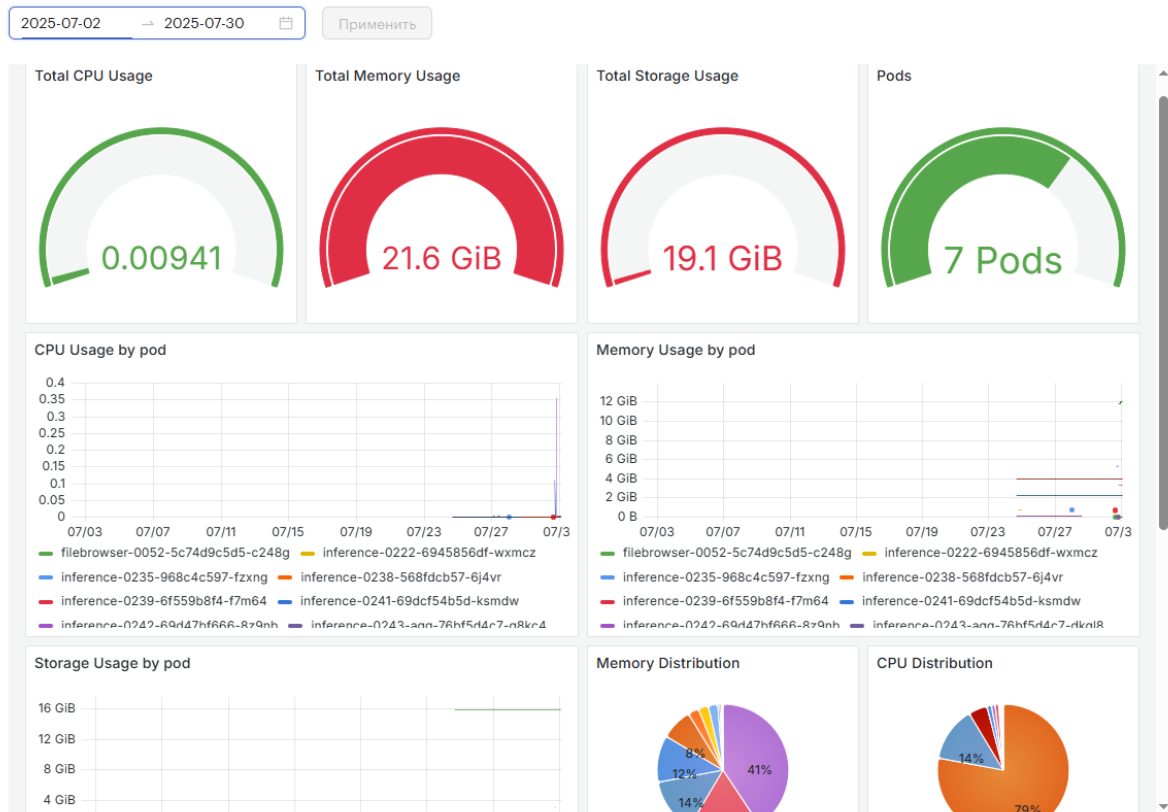
Для просмотра метрик подов:



1. Выберите промежуток времени:



2. После этого будет отображаться данные показатели:



На странице представлены следующие дашборды:

1. Общая информация:

- **Total CPU Usage:** Совокупная загрузка CPU всех подов. Отображается в виде числового индикатора и круговой шкалы
- **Total Memory Usage:** Общее использование оперативной памяти (21 ГиБ), что указывает на высокую нагрузку

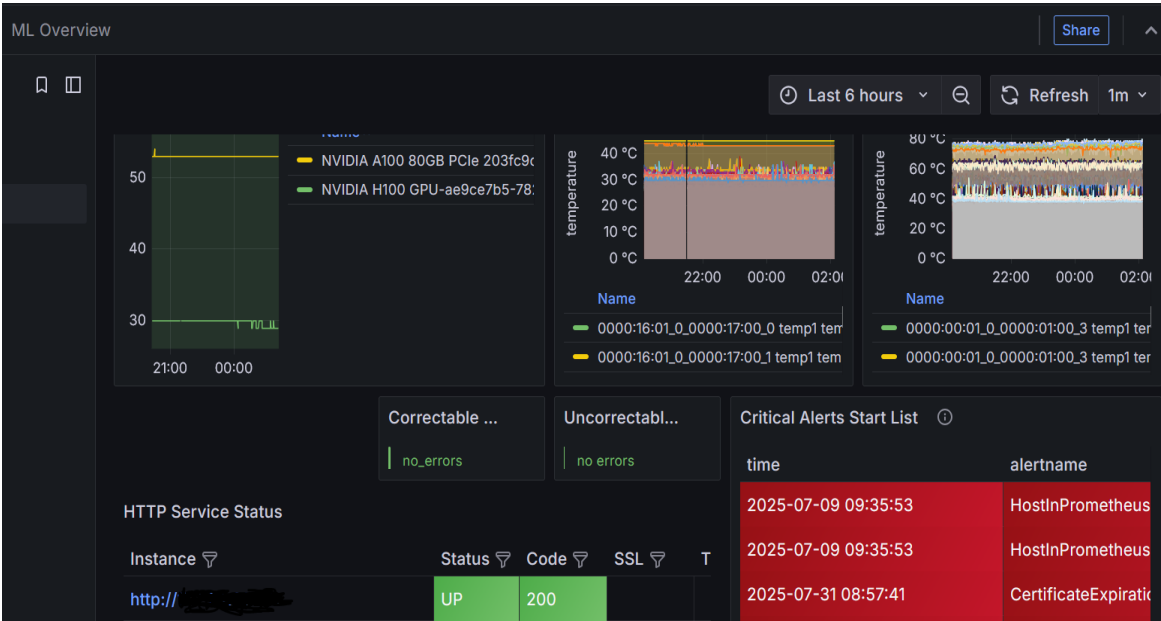
- **Total Storage Usage:** Общий объем используемого хранилища (19.1 ГиБ)
- **Pods:** Текущее количество активных подов
- 2. Детализация по подам
 - **CPU Usage by pod:** График загрузки CPU по каждому поду во времени
 - **Memory Usage by pod:** График использования оперативной памяти; один из подов потребляет более 12 ГиБ
 - **Storage Usage by pod:** Использование дискового пространства; один под использует более 16 ГиБ
- 3. Распределение ресурсов:
 - **Memory Distribution:** Круговая диаграмма, показывающая процент памяти, потребляемой каждым подом. Самый ресурсоемкий под использует 41% всей памяти
 - **CPU Distribution:** Круговая диаграмма распределения нагрузки CPU между подами. Один под занимает 79% всей загрузки CPU, что может указывать на высокую активность или необходимость оптимизации
- 4. Состояние подов и оповещения:
 - **Pod Uptime** — визуализация времени непрерывной работы подов с момента последнего запуска или перезапуска. Отображается в неделях или днях:
 - jupyterlab-0310 работает **3.86 недели**;
 - jupyterlab-0313 — **2.85 недели**;
 - vscode-allysenko-01... — **4.03 дня**.
 - **Alerts** — панель оповещений. В текущий момент активных тревог или уведомлений **не зафиксировано** (отображается сообщение "No alerts matching filters"), что свидетельствует об отсутствии критических инцидентов в кластере на момент мониторинга.

1.19.3 Мониторинг платформы

1.19.3.1 Описание преднастроенных дашбордов системы визуализации метрик

1.19.3.1.1 ML Overview

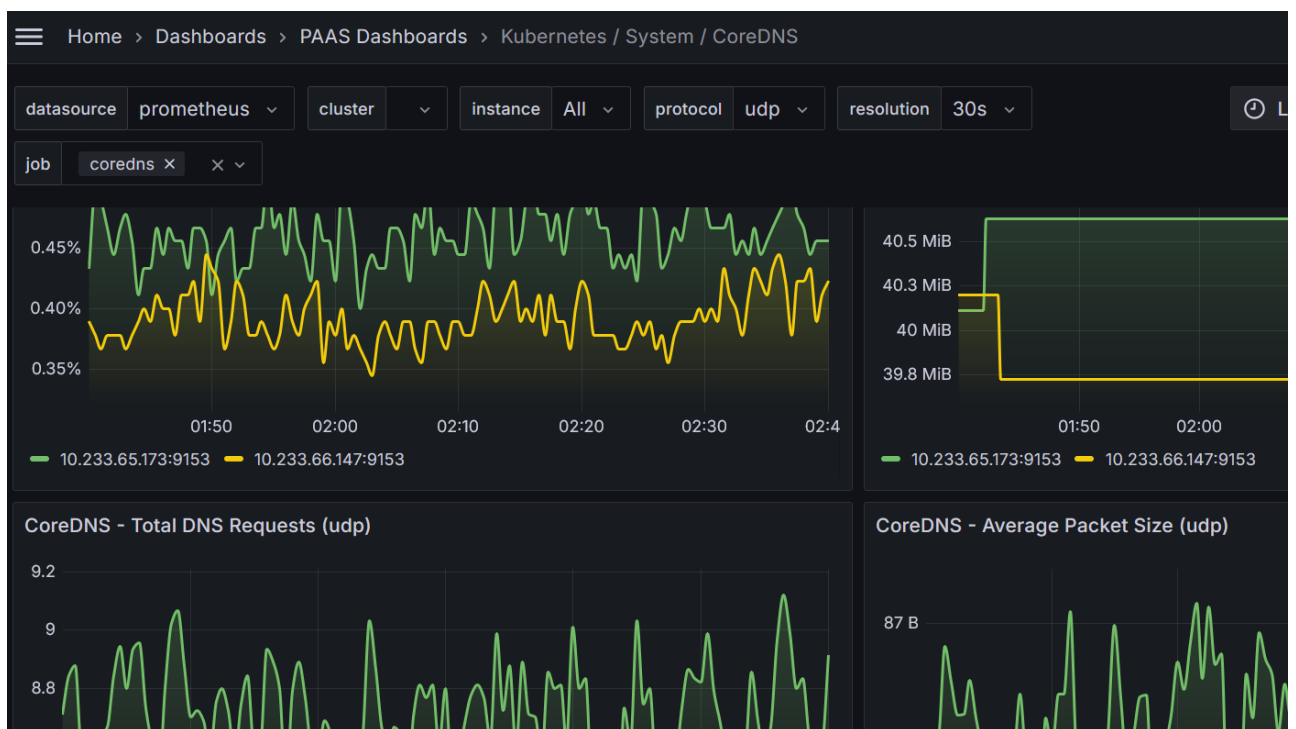
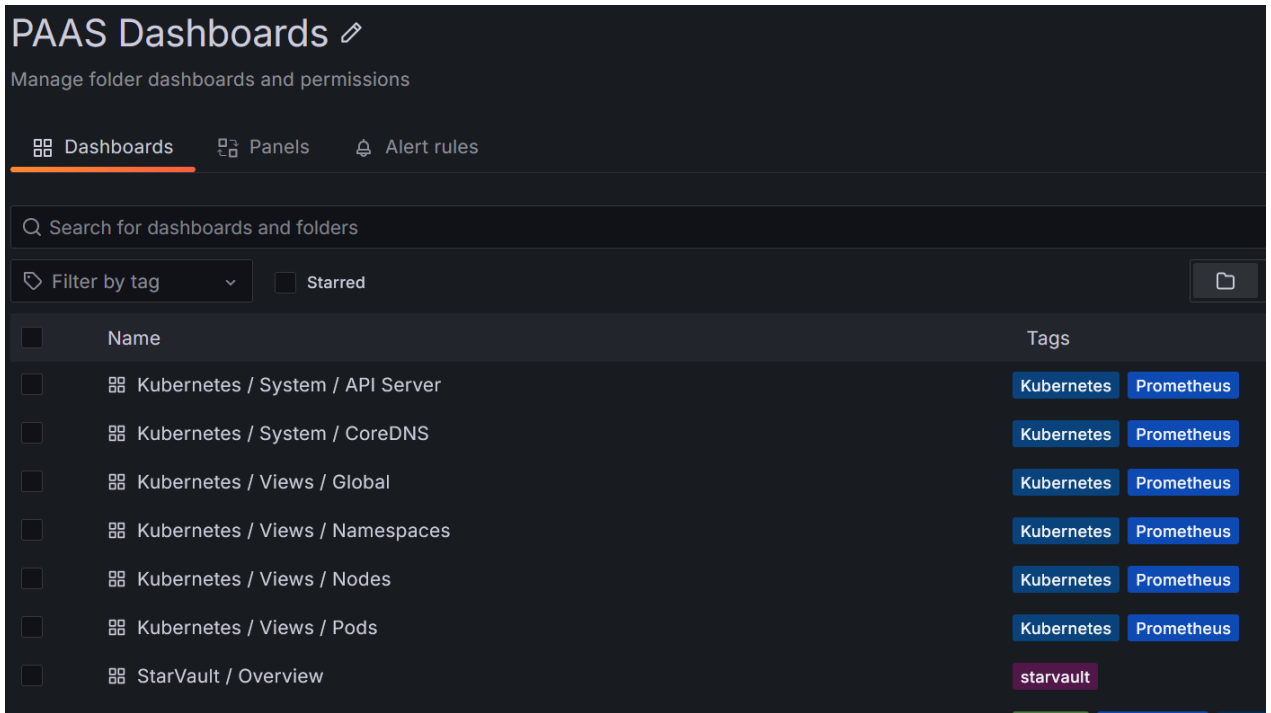
Статус основных API Платформы, температура доступных датчиков на гипервизорах, некорректируемые/корректируемые ошибки памяти (количество) а также температура GPU и лист активных оповещений.



1.19.3.1.2 Каталог дашбордов PAAS Dashboards

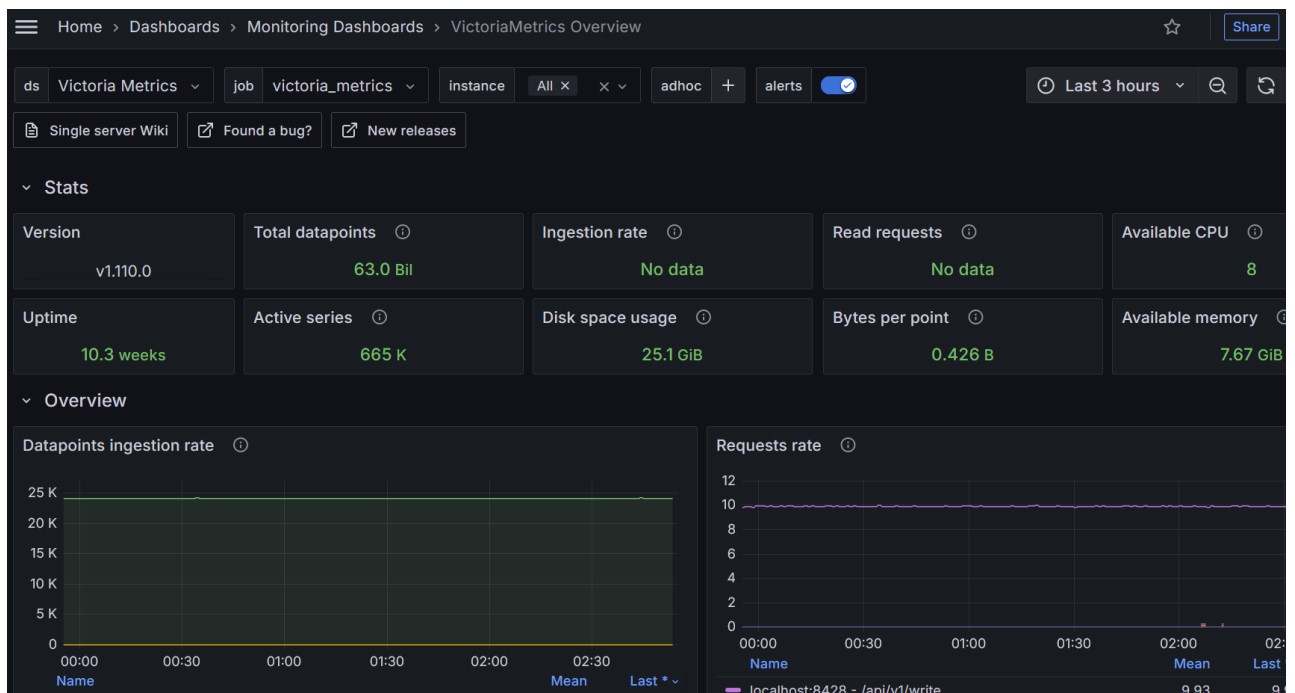
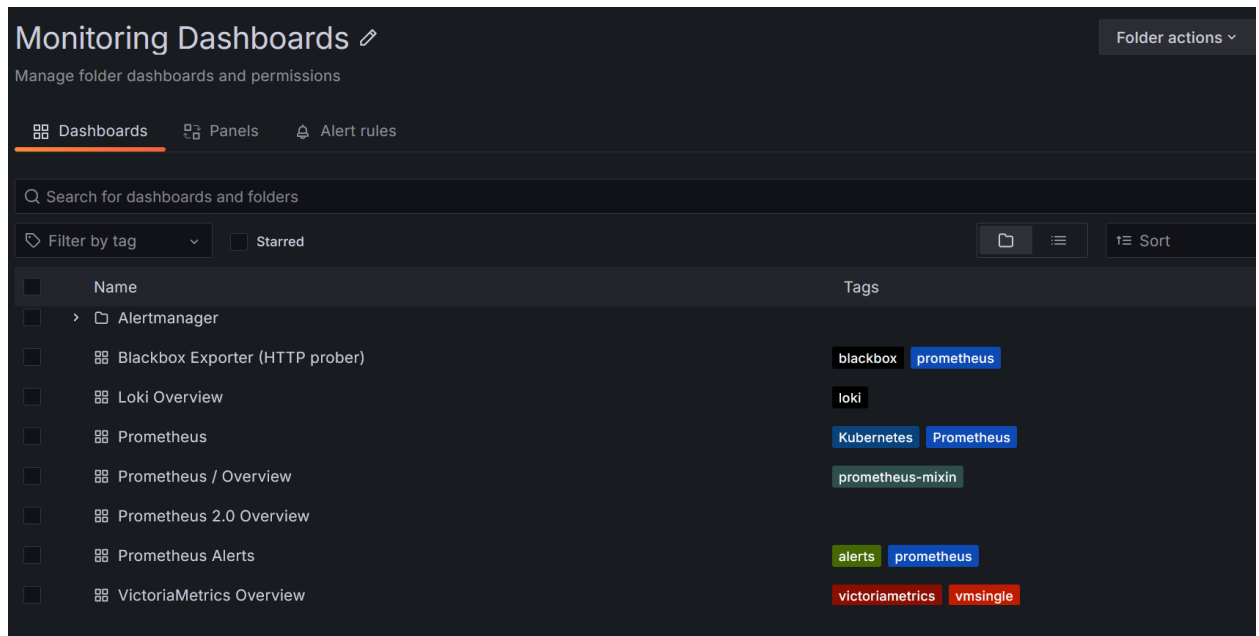
Состояние основных сервисов, используемых в платформе контейнеризации, а также подов, неймспейсов и пользовательской нагрузки:

Core Dns, Api, flux, cilium, nginx ingress.



1.19.3.1.3 Каталог дашбордов Monitoring dashboards

Состояние и метрики всех основных сервисов, входящих в состав платформы: VictoriaMetrics, Prometheus, Loki, Prometheus Alert manager.



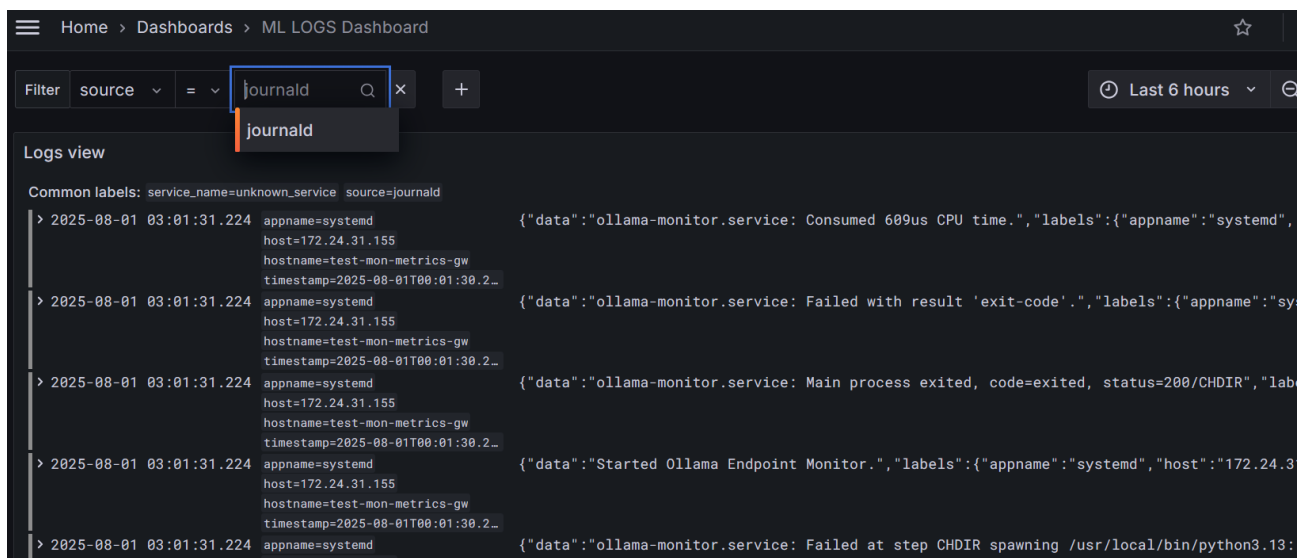
1.19.3.1.4 Каталог дашбордов платформы виртуализации

Основные метрики по гипервизорам и VM: uptime, активные виртуальные машины, аллоцированные ядра, память, диски.



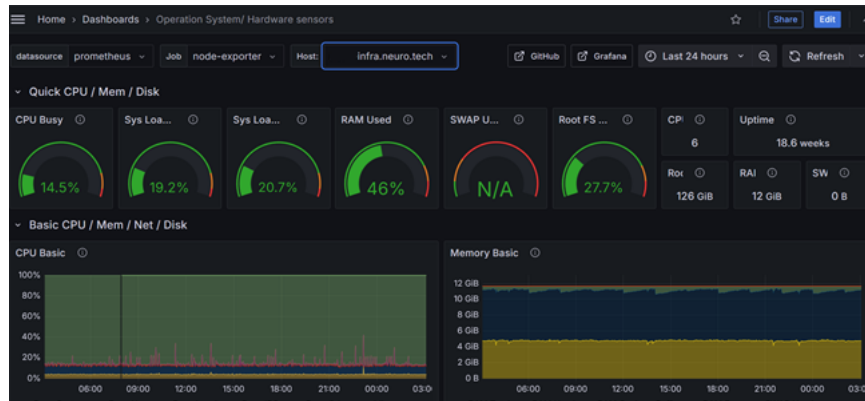
1.19.3.1.5 Дашборд ML LOGS Dashboard

Фильтрация и поиск по всем журналам, которые собираются с помощью vector-agent и хранятся в Loki.



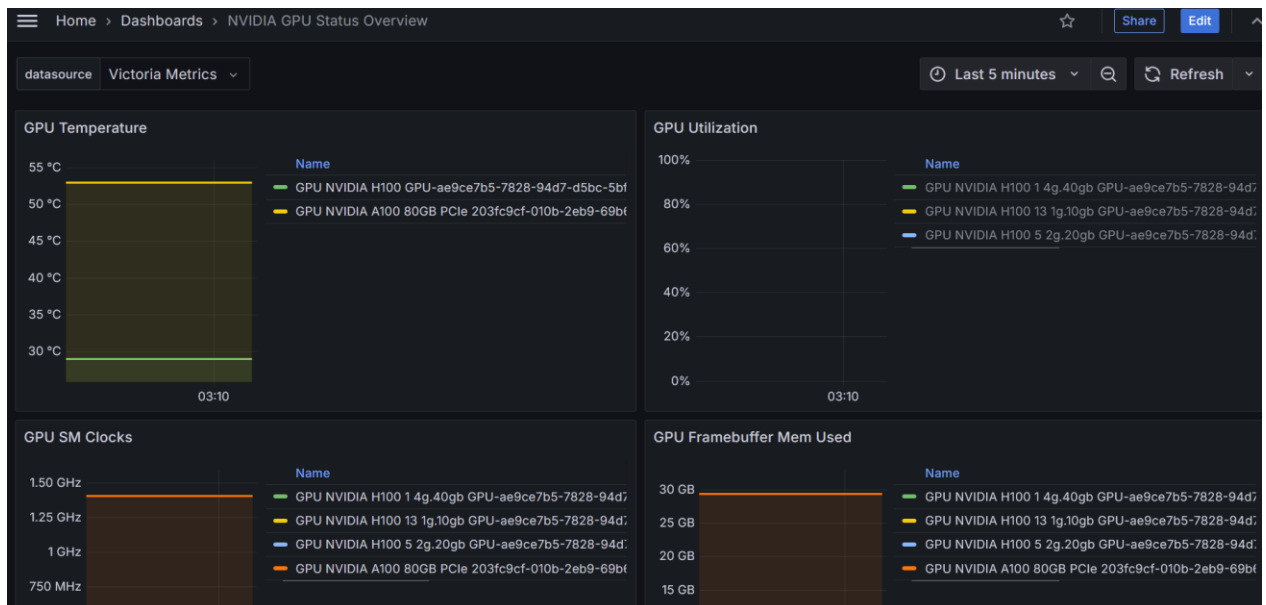
1.19.3.1.6 Дашборд operation System/Hardware sensors

Отображение метрик, собираемых с помощью node-exporter, по всем ВМ и серверам в платформе: процессор, память, диск, своп, метрики сетевых интерфейсов, systemd сервисов, сенсоров.



1.19.3.1.7 Дашборд Nvidia Gpu Status Overview

Статистика по всем GPU в платформе: температура, загрузка ядер, потребление питания.

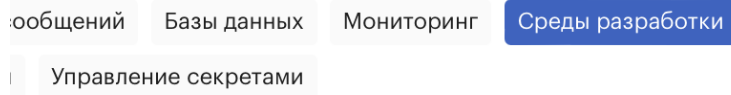


1.20 Inference

В разделе **Среды разработки** можно заказать услуги этой категории и просмотреть информацию о состоянии инстансов сред разработки. К созданным ресурсам можно применить сортировку по параметрам «Продукт», «Платформа», «Имя» и «CPU, RAM, Диск».

Для заказа услуги:

1. Перейдите в раздел **Среды разработки** и нажмите **Заказать**.
2. Выберите среду разработки:
 - Inference



3. Укажите следующие параметры:

- a. **Наименование пода** — кажите название пода.
- b. **Платформа** — выберите платформу виртуализации, на которой будет размещен под.
- c. **Runtime** — выберите фреймворк для запуска (vLLM, Dynamo, Transformers v4.51.3).
 - i. **NVIDIA Dynamo** — это высокопроизводительная платформа для вывода данных с низкой задержкой. рекомендуется для развертывания моделей с (LLM) с миллиардами параметров (от 64b) на одном или нескольких графических ускорителях для задач класса generate.
 - ii. **vLLM** - Движок для запуска больших языковых моделей vLLM, часть экосистемы PyTorch. Рекомендуется для запуска инференса небольших и средних моделей (до 64b), поддерживает запуск в режиме classify, score, generate, embed, reward.
 - iii. **Transformers** — библиотека Python с открытым исходным кодом, и унифицированный интерфейс для работы трансформерными моделями в области обработки естественного языка (NLP). Библиотека поддерживает широкий спектр архитектур нейронных сетей, включая BERT, GPT, T5, RoBERTa, DistilBERT, ELECTRA, DeBERTa, и др.
- d. **Модель** — в списке моделей выберите предварительно загруженную модель из списка, например, mtsair-Kodify-Nano-GPTQ.
- e. **Токен доступа** — сгенерируйте токен для доступа.
- f. **CPU и RAM** — выберите спецификацию ВМ, где CPU — размер vCPU, а RAM — размер оперативной памяти в ГБ.
- g. **Хранилище** — укажите размер хранилища
- h. **GPU** — выберите из списка тип GPU "Целая" или "Mig".
- i. **Графические карты** — нажмите добавить и в выпадающем списке выберите нужные настройки.
 - i. Количество — введите количество GPU.
 - ii. Модель GPU — выберите модель GPU используемую в поде или оставьте по умолчанию.

← Новый заказ
Inference | Среда разработки

Inference

Название:

Платформа:

Runtime:

Модель:

Обязательно использовать цифры, прописные и строчные латинские буквы. Длина от 10 до 20 символов.

Токен доступа:

CPU и RAM:

Хранилище:

GPU:

Графические карты

Модель GPU:

Количество:

Модель GPU:

Процессор Inference 2,45 Р
За 2 шт

Оперативная память для Inference 4,91 Р
За 4 шт

Хранилище для Inference 24,54 Р
За 20 Гб

Итого в день 31,90 Р

Заказать

4. Для Runtime Dynamo укажите следующие параметры:

- Конфигурация etcd** — по умолчанию указаны оптимальные параметры, для изменения отредактируйте спецификацию VM, где CPU — размер vCPU, а RAM — размер оперативной памяти в Гб и размер хранилища в Гб.
- Конфигурация nats** — по умолчанию указаны оптимальные параметры, для изменения отредактируйте спецификацию VM, где CPU — размер vCPU, а RAM — размер оперативной памяти в Гб и размер хранилища в Гб.

← inference-0247
Inference | Среда разработки

4 CPU 16 GB RAM 20 GB Диск 49,07 Р в день

адрес: <https://inference-0247.prod.us-east-1.amazonaws.com/>

Имя: <https://inference-0247.prod.us-east-1.amazonaws.com/>

Платформа:

Длина имени:

Ссылка на документ:

Общая информация | Действия | История

Конфигурация

CPU	RAM, Гб	Хранилище	GPU	Модель GPU
4	16	20	1	NVIDIA-H100-MIG-2g.20gb

Модель

Runtime	Модель	Тип кода
vLLM	YandexGPT-5-Lite-8B-instruct	generate

Подключение

Токен доступа: <https://inference-0247.prod.us-east-1.amazonaws.com/>

Адрес API для подключения: <https://inference-0247.prod.us-east-1.amazonaws.com/>

Пример запроса

```
curl -X POST https://inference-0247.prod.us-east-1.amazonaws.com/v1/chat/completions -H "Authorization: Bearer OQOJ7M7D" -H "Content-Type: application/json" -d '{"model": "YandexGPT-5-Lite-8B-instruct", "messages": [{"role": "user", "content": "Расскажите об интеграции исключений"}], "top_p": 0.1}'
```

5. Нажмите кнопку "Заказать".

6. Вы можете обратиться к модели из кода или приложения по адресу из поля "Адрес API для подключения". Вы можете начать работать с моделью, используя данные с формы заказа