

Инструкция для экспертной проверки программного обеспечения

1. Подготовка к установке

1.1 Установка должна производиться на устройства отвечающие следующим минимальным требованиям к аппаратному обеспечению:

ЦП: x86_64 архитектура, от 2 ядер

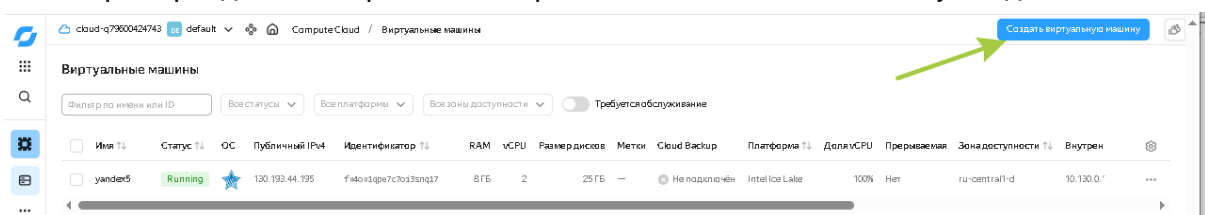
- ОЗУ: от 8 ГБ
- GPU: (Опционально), NVIDIA с поддержкой CUDA 11.8+ (Внимание! При развертывании ПО на конфигурации без GPU время обработки запросов будет значительно увеличено, до 10-15 мин\запрос)
- Дисковая подсистема: SSD от 25 ГБ .
- ОС: Astra Linux, ОСнова
- Python 3.11

1.2 Возможный вариант развертывания виртуальной машины

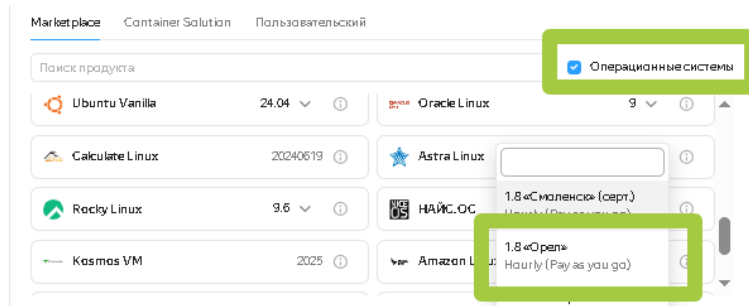
Создание VM на примере Yandex Cloud, приведенный в видеоинструкции по установке

Варианты создания инфраструктуры не ограничиваются ПО

1. Выберите раздел Все сервисы - compute cloud, нажмите на кнопку создания VM



2. В разделе выбора ОС установите чекбокс Операционные системы, найдите Astra linux и выберите в выпадающем списке версию 1.8 "Орел"



3. Установите размер HDD или SSD на менее 25 Гб

Диски и файловые хранилища

disk-ubuntu-24-04-lts-1784511213900 Загруженный

Тип [?] SSD **HDD** SSD IO Нереплицируемый SSD

Размер [?] - + CPS, bandwidth ▾

Добавить

4. В вкладке Своя конфигурация установите платформу, количество vCPU и RAM в соответствии с минимальной конфигурацией

Вычислительные ресурсы

Standard High memory ☒ HighFreq CPU Shared-core GPU **Своя конфигурация**

Платформа [?]

vCPU

Гарантированная доля vCPU [?] 20% 50% **100%**

Для решения любых задач, в том числе для высоконагруженных сервисов

RAM

Дополнительно ☐ Прерываемая [?]

5. Установите параметры доступа

Доступ

SSH-ключ [?]

Данные для аутентификации хранятся на VM и не могут быть изменены. В случае утери закрытого SSH-ключа доступ к VM станет невозможен.

Доступ по OS Login [?]

Более безопасное подключение с помощью SSH-сертификатов или SSH-ключей. Легко восстановить доступ даже в случае утери ключа.

Плагин [?]

SSH-ключ [?] ✕ ▾ или Добавить ключ

Резервное копирование [?] ☐

После создания VM, следуйте инструкции по установке

2. Установка ПО

2.1 Установка ПО на Astra linux 1.8 «Орел»

#Создание папки ПО

```
sudo mkdir -p /opt/rag
```

#Установка прав для пользователя

```
sudo chown -R $USER:$USER /opt/rag
```

```
cd /opt/rag
```

Ручная установка локального сервера

```
curl -fSLk https://app.ragsystem.ru:8111/ollama-linux-amd64.tar.zst | sudo  
tar --zstd -x -C /usr
```

```
sudo bash -c "cat << 'EOF' > /etc/systemd/system/ollama.service
```

```
[Unit]
```

```
Description=Ollama Service
```

```
After=network-online.target
```

```
[Service]
```

```
ExecStart=/usr/bin/ollama serve
```

```
User=ollama
```

```
Group=ollama
```

```
Restart=always
```

```
RestartSec=3
```

```
Environment=\"OLLAMA_HOST=0.0.0.0\"
```

```
[Install]
```

```
WantedBy=default.target
```

```
EOF"
```

```
sudo useradd -r -s /bin/false -m -d /usr/share/ollama ollama
```

```
sudo systemctl daemon-reload
```

```
sudo systemctl enable ollama
```

```
sudo systemctl start ollama
```

```
#Установка venv
```

```
sudo apt install python3.11-venv
```

```
#Создание и активация виртуального окружения
```

```
python3.11 -m venv venv_cpu
```

```
source venv_cpu/bin/activate
```

```
# Установка PyTorch 2.4+
```

```
pip install torch==2.4.0 --index-url https://download.pytorch.org/whl/cpu
```

```
sudo apt-get install -y swig libjpeg-dev zlib1g-dev libfreetype6-dev
```

Загрузите ПО (файл архива) в рабочую папку /opt/rag/ и распакуйте содержимое вручную или с помощью команд

```
bash
```

```
curl -L -o rag.zip https://oqoai.ru/rag.zip
```

```
sudo apt install -y unzip
```

```
unzip -o rag.zip
```

```
# Установка библиотек
```

```
pip install -r requirements_cpu_1.txt
```

```
#Загрузка локальных версий NLP моделей
```

```
pip install /opt/rag/models/ru_core_news_sm-3.8.0-py3-none-any.whl
```

```
/opt/rag/models/en_core_web_sm-3.8.0-py3-none-any.whl
```

```
curl -L -k --create-dirs -o /opt/rag/models/gemma-3-4b-it-UD-Q2_K_XL.gguf  
"https://oqoai.ru/gemma-3-4b-it-UD-Q2_K_XL.gguf"
```

```
ollama create gemma3:4b -f ./Modelfile
```

```
#Загрузка локальных версий эмбединг моделей
```

```
mkdir -p ~/.cache/huggingface/hub/
```

```
curl -k -O https://app.ragsystem.ru:8111/hfmodel.zip
```

```
unzip hfmodel.zip -d ~/.cache/huggingface/hub/
```

[Далее к разделу первый запуск](#)

3. Первый запуск

Для первого запуска, выполните

```
python launch.py
```

Дождитесь запуска всех серверов ПО (Первый запуск может требовать до 2-5 минут)

Откройте в браузере веб интерфейс соответствующий назначенному при запуске локальному адресу (по умолчанию при установке на локально `http://127.0.0.1:3003` или при установке на VM `http://<IP адрес VM>:3003`. Все внутренние IP адреса доступны в консоли запуска)

ПО готово к работе. Для дальнейшей эксплуатации воспользуетесь руководством пользователя ПО

4. Обзор функций

Обзор интерфейса

Интерфейс состоит из 3 частей:

Стартовый интерфейс - Сайдбар и панель запроса

Поисковый интерфейс - Сайдбар, панель источников, чат

Интерфейс просмотра - Сайдбар, Просмотр PDF, чат

Гостевой доступ. Функции без создания пользователя

В загруженном интерфейсе (по умолчанию `http://127.0.0.1:3003`) доступны запросы в демонстрационной базе знаний (Документ о RAG системе)

Пример вопросов:

- что такое RAG

- как обрабатывается документ

и т.д

Доступны запросы как к базе целиком, так и только по выбранному документу

Вход и создание пользователей

Нажмите кнопку войти в левой панели , войдите или выберете вкладку регистрация (Первый созданный пользователь получает права администратора ПО)

Внимание! В проверочной версии учетная запись администратора уже создана логин `q@q.q` пароль `qqqqqq`

Элементы управления ПО

После входа пользователь в панели навигации сайдбара получит доступ к основным функциям ПО

- Создание нового чата (Для текущей базы)
- Управление доступными пользователю базами знаний (только для администраторов ПО)
- Управление учетными записями пользователей (только для администраторов ПО)
- Выбор баз из доступных

Стартовый поисковый интерфейс содержит селектор в котором доступны для выбора базы доступные пользователю

Внизу чата доступны строка ввода и кнопка отправки запроса

В режиме Просмотра документа доступен переход в широкоэкранный режим просмотра

Функции управления

- Создание базы
- Добавление документов
- Подключение файла атрибутов
- Настройка LLM
- Настройка прав доступа к базе
- Запуск формирования базы

Пользователи

- Создание
- Редактирование
- Роли
 - Администратор
 - Пользователь

Поддержка

По вопросам, связанным установкой и эксплуатацией ПО, обращайтесь в техподдержку:

Телефон: +79600424743

E-mail: q79600424743@yandex.ru