INFOWATCH

Руководство по установке, конфигурированию и администрированию

19/03/2024

© АО “ИнфоВотч”

Тел./Факс +7 (495) 229-00-22

http://www.infowatch.ru

СОДЕРЖАНИЕ

1 Аудитория 4

2 Комплект документов 5

2.1 Техническая поддержка пользователей 5

3 Обзор 6

4 Аппаратные и программные требования Data Analysis Service 7

4.1 Аппаратные требования 7

4.2 Программные требования 7

5 Подготовка сервера к установке/обновлению Data Analysis Service 9

6 Установка Data Analysis Service 11

6.1 Чтобы установить Data Analysis Service: 11

6.2 Чтобы установить Data Analysis Service в режиме филиала: 13

6.3 Установка в тихом режиме 14

7 Обновление Data Analysis Service 17

7.1 Чтобы обновить Data Analysis Service: 17

8 Настройка Data Analysis Service 19

8.1 Настройка производительности 20

8.2 Настройка использования ресурсов сервера 20

8.3 Настройка модуля транскрибации 21

9 Удаление Data Analysis Service 24

Настоящее руководство содержит сведения по установке, удалению и настройке компонентов сервиса InfoWatch Data Analysis Service (далее также Сервис или Data Analysis Service).

1 Аудитория

Информация, содержащаяся в Руководстве, предназначена для пользователей, работающих с Сервисом (выполняющих установку, удаление и т. п.).

Руководство рассчитано на пользователей, знакомых с основами работы в среде операционных систем Linux и СУБД PostgreSQL.

2 Комплект документов

В документацию по InfoWatch Data Analysis Service входит:

- «InfoWatch Data Analysis Service. Руководство по установке, конфигурированию и администрированию». Содержит описание установки, обновления, удаления и настройки Сервиса.

Сопутствующая документация по продукту InfoWatch Traffic Monitor включает в себя:

- «InfoWatch Traffic Monitor. Руководство по установке». Содержит описание установки, настройки и удаления системы InfoWatch Traffic Monitor (далее Traffic Monitor);
- «InfoWatch Traffic Monitor. Руководство администратора». Содержит информацию по администрированию Traffic Monitor (база данных, серверная часть);
- «InfoWatch Traffic Monitor. Руководство пользователя». Содержит описание работы с Traffic Monitor (настройка конфигурации, экспорт/импорт данных, составление политик для обработки объектов);
- «InfoWatch Traffic Monitor. Справочник по конфигурационным файлам». Содержит пояснения к часто используемым конфигурационным файлам.

2.1 Техническая поддержка пользователей

При возникновении вопросов, связанных с работой Сервиса, вы можете обратиться в службу технической поддержки:

- если вы приобрели Data Analysis Service у партнера компании InfoWatch, то обратитесь в службу технической поддержки партнера;
- если продукт приобретен у компании InfoWatch напрямую, то обратитесь в службу технической поддержки компании InfoWatch по адресу support@infowatch.com.

Часы работы службы технической поддержки – с 7:00 до 21:00 по московскому времени с понедельника по пятницу, исключая официальные выходные и праздничные дни РФ. Вы также можете посетить раздел технической поддержки на нашем сайте: https://www.infowatch.ru/services/support.

3 Обзор

Data Analysis Service предназначен для потоковой постобработки и анализа данных. Data Analysis Service обрабатывает данные, предоставляемые продуктами компании InfoWatch.

Функциональные возможности Data Analysis Service

- Протокластеризация текстовых данных.
  Данная функциональность позволяет анализировать текстовые документы и разбивать их на тематические группы. Протокластеризация используется для работы функции *Поиск аналогов* InfoWatch Traffic Monitor. Подробнее см. "InfoWatch Traffic Monitor. Руководство пользователя", статья "Поиск аналогов".
- Транскрибация.
  Данная функциональность позволяет конвертировать аудиозапись в текст. Транскрибация используется в работе с аудиозаписями в InfoWatch Activity Monitor. Подробнее см. "InfoWatch Activity Monitor. Руководство пользователя", статья "Работа с аудиозаписями".
- Детектирование непристойного контента.
  Данная функциональность позволяет распознавать непристойные изображения на снимках экрана. Детектирование непристойного контента реализовано для InfoWatch Activity Monitor.

4 Аппаратные и программные требования Data Analysis Service

Для того чтобы Data Analysis Service мог получать данные, необходима интеграция с каким-либо другим продуктом InfoWatch. Data Analysis Service поддерживает интеграцию со следующими продуктами:

- Traffic Monitor начиная с версии 7.8
- Activity Monitor 2.2

> **Примечание:**
>
> Data Analysis Service может работать с одним или несколькими серверами Traffic Monitor.

4.1 Аппаратные требования

Минимальные системные требования, необходимые для установки:

Процессор	ОЗУ	Жесткий диск
Количество ядер – 4; Частота – 3.1 ГГц; Разрядность – 64 bit; Поддержка инструкций SSE 4.2.	16 Гбайт	от 150 Гбайт

> **Примечание:**
>
> Чтобы точно рассчитать требования для конкретного проекта, обратитесь в службу технической поддержки компании InfoWatch.

4.2 Программные требования

Для установки и работы Data Analysis Service требуется использовать следующее программное обеспечение:

Тип ПО	Варианты
Операционная система	• Astra Linux Special Edition 1.6 "Смоленск": • Update 10; • Astra Linux Special Edition 1.7 "Смоленск": • версии 1.7.x (начиная с 1.7.0-11.06.2021_12.40); • Red Hat Enterprise Linux: • версии 7.x (начиная с 7.7); • версии 8.x (начиная с 8.0); • Oracle Linux: • версии 7.x (начиная с 7.9);

Тип ПО	Варианты
	• версии 8.x (начиная с 8.4); • РЕД ОС: • версии 7.x.x (начиная с версии 7.3.1).
СУБД	PostgreSQL версии 12 и выше

Примечание:

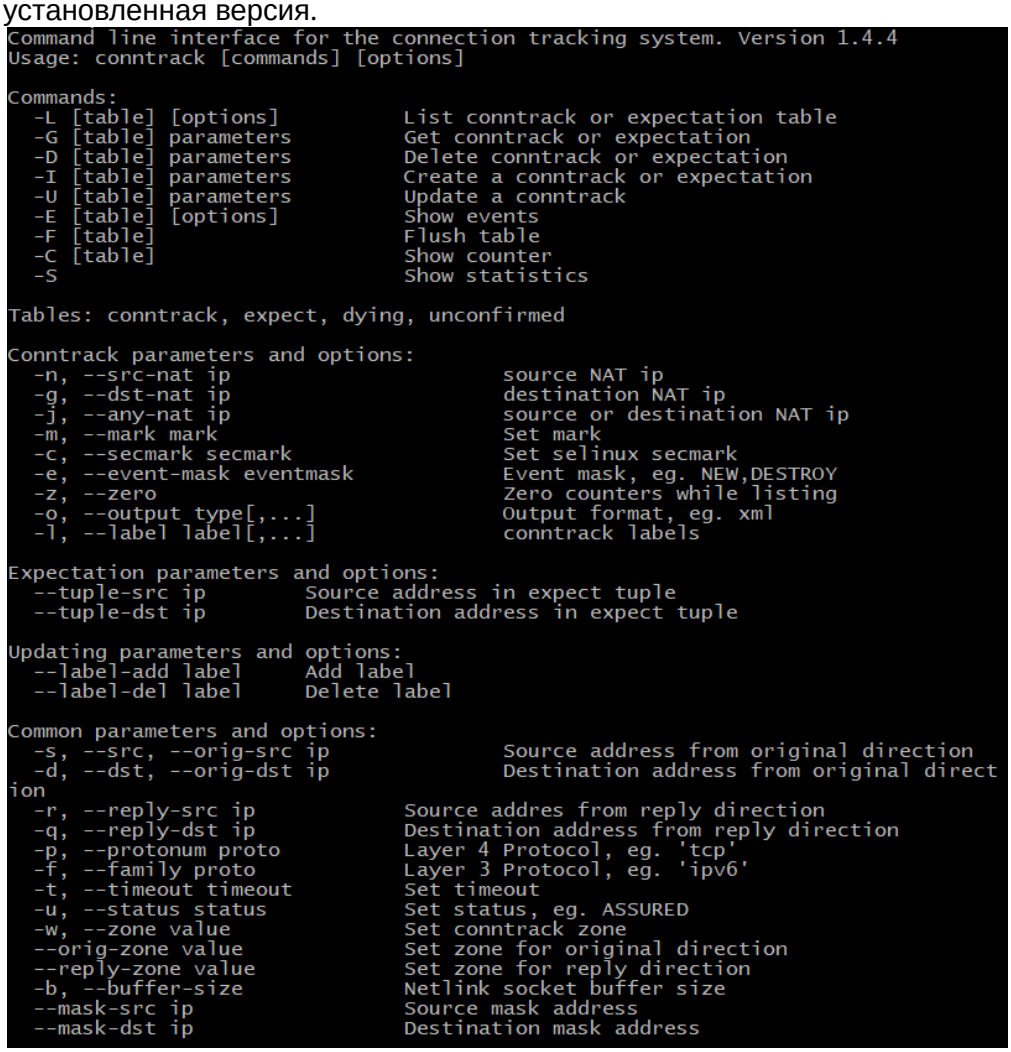
Аппаратные и программные требования для использования графического процессора (GPU) при транскрибации приведены в статье Настройка модуля транскрибации.

5 Подготовка сервера к установке/обновлению Data Analysis Service

Перед установкой или обновлением Data Analysis Service обязательно подготовьте сервер:

1. Программа установки написана на языке Python версии 2.7, поэтому для установки Сервиса установите интерпретатор этого языка. Пакет с интерпретатором не входит в состав дистрибутива Сервиса.
2. В окружении Red Hat Enterprise Linux и РЕД ОС обновите политики SELinux. Для этого скачайте и установите пакет `container-selinux`.
   Вы можете установить пакет, выполнив команду:
   `yum install container-selinux`
3. На сервере под управлением ОС Red Hat Enterprise Linux и РЕД ОС отключите систему контроля доступа SELinux. Для этого:
   a. Выполните команды от имени суперпользователя `root`:
      ```
      setenforce 0
      sed -i 's/SELINUX=enforcing/SELINUX=disabled/g' /etc/selinux/config
      ```
   b. Перезагрузите сервер с помощью команды:
      `reboot`
   c. Проверьте статус системы SELinux с помощью команды:
      `sestatus`
4. Настройте правила POD сети или отключите межсетевой экран (подробнее см. статьи Базы знаний "Конфликты при взаимодействии службы firewalld и Kubernetes", "Полное отключение межсетевого экрана"). В противном случае установка Сервиса будет невозможна.
5. Для установки и корректной работы Data Analysis Service в ОС Astra Linux Special Edition убедитесь, что настройка безопасности *astra-nochmodx-lock* отключена.
   Вы можете отключить эту настройку с помощью команды:
   `astra-nochmodx-lock disable`
6. Если вы используете антивирус, настройте его для совместной работы с Data Analysis Service (см. статью Базы знаний "Настройка антивируса для продуктов Платформы").
7. Data Analysis Service использует компонент Kubernetes. Для корректной установки Сервиса и работы Kubernetes на сервере должен быть установлен пакет *conntrack*. Убедитесь, что пакет установлен. Для этого введите в командной строке:
   `conntrack`
   В результате будут выведены сведения о работе с модулем *conntrack* и его

установленная версия.



```
Command line interface for the connection tracking system. Version 1.4.4
Usage: conntrack [commands] [options]

Commands:
  -L [table] [options]          List conntrack or expectation table
  -G [table] parameters         Get conntrack or expectation
  -D [table] parameters         Delete conntrack or expectation
  -I [table] parameters         Create a conntrack or expectation
  -U [table] parameters         Update a conntrack
  -E [table] [options]          Show events
  -F [table]                    Flush table
  -C [table]                    Show counter
  -S                            Show statistics

Tables: conntrack, expect, dying, unconfirmed

Conntrack parameters and options:
  -n, --src-nat ip                      source NAT ip
  -g, --dst-nat ip                      destination NAT ip
  -j, --any-nat ip                      source or destination NAT ip
  -m, --mark mark                       Set mark
  -c, --secmark secmark                 Set selinux secmark
  -e, --event-mask eventmask            Event mask, eg. NEW,DESTROY
  -z, --zero                            Zero counters while listing
  -o, --output type[,...]               Output format, eg. xml
  -l, --label label[,...]               conntrack labels

Expectation parameters and options:
  --tuple-src ip        Source address in expect tuple
  --tuple-dst ip        Destination address in expect tuple

Updating parameters and options:
  --label-add label     Add label
  --label-del label     Delete label

Common parameters and options:
  -s, --src, --orig-src ip              Source address from original direction
  -d, --dst, --orig-dst ip              Destination address from original direct
ion
  -r, --reply-src ip            Source addres from reply direction
  -q, --reply-dst ip            Destination address from reply direction
  -p, --protonum proto          Layer 4 Protocol, eg. 'tcp'
  -f, --family proto            Layer 3 Protocol, eg. 'ipv6'
  -t, --timeout timeout         Set timeout
  -u, --status status           Set status, eg. ASSURED
  -w, --zone value              Set conntrack zone
  --orig-zone value             Set zone for original direction
  --reply-zone value            Set zone for reply direction
  -b, --buffer-size             Netlink socket buffer size
  --mask-src ip                 Source mask address
  --mask-dst ip                 Destination mask address
```

Если модуль не установлен, установите его с помощью команды:

- `yum install conntrack` – для ОС Red Hat Enterpise Linux, Oracle Linux и РЕД ОС;
- `apt-get install conntrack` – для ОС Astra Linux.

Если пакет недоступен в репозитории, скачайте его, перейдите в директорию с пакетом и установите его вручную с помощью команды вида:

- `rpm -Uhv ./<имя_пакета>` – для ОС Red Hat Enterpise Linux, Oracle Linux и РЕД ОС;
- `dpkg -i ./<имя_пакета>` – для ОС Astra Linux.

6 Установка Data Analysis Service

> **Важно!**
>
> Перед установкой **обязательно** подготовьте сервер (см. Подготовка сервера к установке/обновлению Data Analysis Service).

Установка Data Analysis Service производится вручную, с помощью командной строки.

Для установки Сервиса используется архив `iw_das_setup_x.x.x.xx.tar.xz`, где x.x.x.xx – номер версии. Архив входит в состав дистрибутива. В нашем примере будет использована версия `1.1.0.41`, ей соответствует архив `iw_das_setup_1.1.0.41.tar.xz`. Архив содержит полный набор бинарных модулей образов контейнеров, необходимых для развертывания Сервиса без доступа к сети Интернет. Также архив содержит программу установки.

> **Примечание:**
>
> Не рекомендуется устанавливать Data Analysis Service на один сервер с Traffic Monitor.
>
> Возможна установка Data Analysis Service на один сервер с продуктами Платформы.

Data Analysis Service может быть установлен в режиме центрального офиса или филиала.

Установка в режиме центрального офиса

6.1 Чтобы установить Data Analysis Service:

1. Создайте новую директорию на диске (например, DAS):
   `mkdir DAS`
2. Скопируйте архив **`iw_das_setup_ 1.1.0.41 .tar.xz`** в созданную директорию.
3. Перейдите в директорию, в которую был скопирован архив.
4. Распакуйте архив с дистрибутивом Data Analysis Service в эту директорию:
   `tar -xvf iw_das_setup_ 1.1.0.41 .tar.xz`
5. Запустите программу установки Сервиса:
   `./setup.py install`
6. Ознакомьтесь с условиями лицензионного соглашения. Лицензионное соглашение содержит несколько страниц. Для перехода на следующую страницу используйте клавишу **Enter**.
7. Введите "**y**", чтобы принять лицензионное соглашение, и нажмите **Enter**.
8. Введите IP-адрес сетевого интерфейса для взаимодействия с кластером в формате IPv4: "xxx.xxx.xxx.xxx" и нажмите **Enter**.

   > **Примечание:**
   >
   > Если указать `0.0.0.0`, будут использованы все доступные сетевые интерфейсы.

Здесь и далее для использования значений, предложенных по умолчанию, нажмите **Enter** без ввода значений.

9. Если на данном сервере не установлен ни один продукт Платформы:
   a. Выделите объем оперативной памяти для размещения данных Clickhouse (по умолчанию: `80%`) и нажмите **Enter**.
   b. Укажите путь для размещения данных Clickhouse (по умолчанию: `/mnt/chdata`) и нажмите **Enter**.
   c. Укажите путь для размещения данных PostgreSQL (по умолчанию: `/mnt/pgdata`) и нажмите **Enter**.
   d. Укажите путь для размещения данных NATS (по умолчанию: `/mnt/natsdata`) и нажмите **Enter**.
   e. Укажите путь для размещения бинарных данных (по умолчанию: `/mnt/dsdata`) и нажмите **Enter**.
   f. Укажите порт подключения к веб-интерфейсу Платформы (по умолчанию: `443`) и нажмите **Enter**.
   g. Дождитесь окончания процесса установки продуктовой платформы.
   h. Ознакомьтесь с отчетом об установке Платформы. В графе **`web ui`** указаны адрес и порт для подключения к веб-интерфейсу Платформы:

```
###Result###
Install product: Infowatch Platform(platform) Version: 1.10.3.573
Install node mode: central
Install node label: central
installed 38 components
updated   0 components
add ref   0 components
web ui:
https://10.60.20.154:443
```

10. Введите "**y**", чтобы включить модуль транскрибации, или "**n**", чтобы оставить его выключенным, (по умолчанию: "**n**") и нажмите **Enter**.
11. Введите "**y**", чтобы включить модуль детектирования непристойного контента, или "**n**", чтобы оставить его выключенным, (по умолчанию: "**n**") и нажмите **Enter**.
12. Оставьте модуль кластеризации выключенным, так как он не используется в данной версии Data Analysis Service. Для этого нажмите **Enter**.
13. Дождитесь окончания процесса установки Data Analysis Service.
14. Ознакомьтесь с отчетом об установке Сервиса:

```
###Result###
Install product: Infowatch DAS(das) Version: 1.1.0.41
Install node mode: central
Install node label: central
installed 5 components
updated   0 components
add ref   0 components
```

15. Чтобы вывести на экран информацию о Платформе и установленных продуктах, выполните команду:
    `./setup.py showproducts`
16. Убедитесь, что все сервисы запущены, выполнив команду:
    `kubectl get pods -n infowatch`
    Команда выведет список сервисов. Каждый сервис должен иметь статус `Running`.

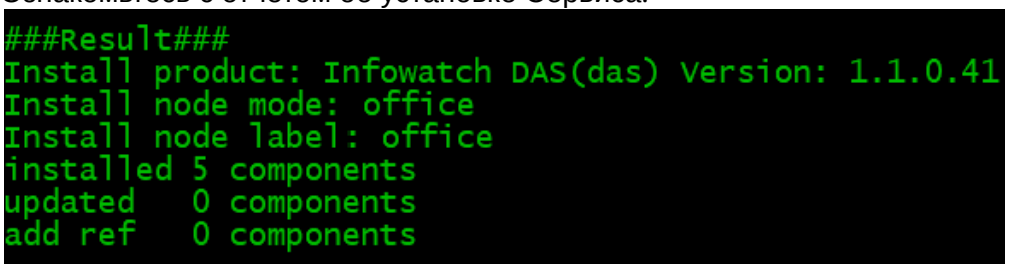
Установка в режиме филиала

6.2 Чтобы установить Data Analysis Service в режиме филиала:

1. Создайте новую директорию на диске (например, DAS):
   `mkdir DAS`
2. Скопируйте архив **`iw_das_setup_1.1.0.41.tar.xz`** в созданную директорию.
3. Перейдите в директорию, в которую был скопирован архив.
4. Распакуйте архив с дистрибутивом Data Analysis Service в эту директорию:
   `tar -xvf iw_das_setup_1.1.0.41.tar.xz`
5. Запустите программу установки Сервиса :
   `./setup.py install --nodemode=office`
6. Ознакомьтесь с условиями лицензионного соглашения. Лицензионное соглашение содержит несколько страниц. Для перехода на следующую страницу используйте клавишу **Enter**.
7. Введите "**y**", чтобы принять лицензионное соглашение, и нажмите **Enter**.
8. Если на данном сервере не установлен ни один продукт Платформы:
   a. Укажите адрес центрального офиса в формате IP-адреса или доменного имени (например: `192.0.2.0` или `host.example.com`).
   b. Укажите токен для подключения к центральному офису.
      Чтобы получить токен:
      i. Подключитесь к серверу центрального офиса.
      ii. В командной строке сервера выполните команду для получения токена:
      `kubeadm token list`
      На экран будет выведен токен:
      ```
      TOKEN

      3wcvoz.zh2ccxzil1hiz7iy
      ```
      Токен в нашем примере: `3wcvoz.zh2ccxzil1hiz7iy`
   c. Укажите лейбл филиала (по умолчанию: `office`) и нажмите **Enter**.
      Лейбл может состоять из букв латинского алфавита и/или цифр. Длина лейбла не должна превышать 32 символа.
   d. Выделите объем оперативной памяти для размещения данных Clickhouse (по умолчанию: `80%`) и нажмите **Enter**.
   e. Укажите путь для размещения данных Clickhouse (по умолчанию: `/mnt/chdata`) и нажмите **Enter**.
   f. Укажите путь для размещения данных PostgreSQL (по умолчанию: `/mnt/pgdata`) и нажмите **Enter**.
   g. Укажите путь для размещения данных NATS (по умолчанию: `/mnt/natsdata`) и нажмите **Enter**.
   h. Укажите путь для размещения бинарных данных (по умолчанию: `/mnt/dsdata`) и нажмите **Enter**.
   i. Дождитесь окончания процесса установки продуктовой платформы.
   j. Ознакомьтесь с отчетом об установке Платформы.
   ```
   ###Result###
   Install product: Infowatch Platform(platform) Version: 1.10.3.573
   Install node mode: office
   Install node label: office
   installed 12 components
   updated   0 components
   add ref   0 components
   ```

9. Введите "**y**", чтобы включить модуль транскрибации, или "**n**", чтобы оставить его выключенным, (по умолчанию: "**n**") и нажмите **Enter**.
10. Введите "**y**", чтобы включить модуль детектирования непристойного контента, или "**n**", чтобы оставить его выключенным, (по умолчанию: "**n**") и нажмите **Enter**.
11. Оставьте модуль кластеризации выключенным, так как он не используется в данной версии Data Analysis Service. Для этого нажмите **Enter**.
12. Дождитесь окончания процесса установки Data Analysis Service.
13. Ознакомьтесь с отчетом об установке Сервиса:



14. Чтобы вывести на экран информацию о Платформе и установленных продуктах, выполните команду:
`./setup.py showproducts`
15. Убедитесь, что все сервисы запущены, выполнив команду:
`kubectl get pods -n infowatch`
Команда выведет список сервисов. Каждый сервис должен иметь статус `Running`.

Чтобы начать использование Data Analysis Service, настройте его взаимодействие с InfoWatch Traffic Monitor (подробнее см. "InfoWatch Traffic Monitor. Руководство администратора", статья "Настройка интеграции с Data Analysis Service") или продуктом Платформы (подробнее см. в документации на соответствующий продукт Платформы).

См. также в этом разделе:

- Установка в тихом режиме

6.3 Установка в тихом режиме

Для минимизации участия в процессе установки Data Analysis Service, в инсталляторе предусмотрен тихий режим установки.

Чтобы запустить установку в тихом режиме:

1. Создайте новую директорию на диске (например, `DAS`):
`mkdir DAS`
2. Скопируйте архив с дистрибутивом Data Analysis Service `iw_das_setup_1.1.0.41.tar.xz` в созданную директорию.
3. Перейдите в директорию, в которую был скопирован архив.
4. Распакуйте архив в эту директорию:
`tar -xvf iw_das_setup_1.1.0.41.tar.xz`
5. Выполните команду `./setup.py install`, указав параметры для установки.

Вы можете указать следующие параметры:

Параметр	Описание
`--nodemode`	Режим установки: центральный офис или филиал. Чтобы установить Data Analysis Service в режиме филиала, укажите этот параметр со значением `office`. В режиме центрального офиса параметр можно не указывать
`--loc`	Локализация. Значение по умолчанию: `ru`. Также можно задать значение `en`
`--licagree`	Принятие лицензионного соглашения. Чтобы принять, укажите значение `yes`
`--advertise`	IP-адрес сетевого интерфейса для взаимодействия с кластером. Значение по умолчанию: `0.0.0.0` Не указывается в случае установки в режиме филиала.
`--chmemsize`	Объем оперативной памяти, используемой для данных ClickHouse. Значение по умолчанию: `80%`
`--chdbpath`	Путь для размещения данных ClickHouse. Значение по умолчанию: `/mnt/chdata`
`--pgdbpath`	Путь для размещения данных PostgreSQL. Значение по умолчанию: `/mnt/pgdata`
`--natspath`	Путь для размещения данных NATS. Значение по умолчанию: `/mnt/natsdata`
`--dsdatapath`	Путь для размещения бинарных данных. Значение по умолчанию: `/mnt/dsdata`
`--guiport`	Порт, по которому будет осуществляться подключение к веб-интерфейсу. значение по умолчанию: `443` Не указывается в случае установки в режиме филиала.
`--pclust-enable`	Включение модуля протокластеризации. Значение по умолчанию: `y`
`--transcribation-enable`	Включение модуля транскрибации. Значение по умолчанию: `n`

Параметр	Описание
`--nsfw-enable`	Включение модуля детектирования непристойного контента. Значение по умолчанию: `n`
`--cluster`	Адрес подключения к центральному офису. Указывается в случае установки в режиме филиала.
`--clustertoken`	Токен подключения к центральному офису. Указывается в случае установки в режиме филиала. В центральном офисе можно: • получить токен, выполнив команду: `kubeadm token list` • создать токен с помощью команды: `kubeadm token create`
`--nodelabel`	Метка удаленного офиса. Указывается в случае установки в режиме филиала. В тексте метки можно использовать латинские буквы и цифры. Длина метки не должна превышать 32 символа.

Пример:

```
./setup.py install \
--loc=ru \
--licagree=yes \
--advertise=0.0.0.0 \
--chmemsize=60% \
--chdbpath=/mnt/chdata \
--pgdbpath=/mnt/pgdata \
--natspath=/mnt/natsdata \
--dsdatapath=/mnt/dsdata \
--guiport=443 \
--pclust-enable=y \
--transcribation-enable=y \
--nsfw-enable=y \
--clusterization-enable=y
```

После установки убедитесь, что все сервисы запущены, выполнив команду:
`kubectl get pods -n infowatch`

Команда выведет список сервисов. Каждый сервис должен иметь статус `Running`.

7 Обновление Data Analysis Service

Обновление Data Analysis Service производится вручную, с помощью командной строки.

Для обновления Сервиса используется архив `iw_das_setup_x.x.x.xx.tar.xz`, где x.x.x.xx – номер версии. Архив входит в состав дистрибутива. Архив содержит полный набор бинарных модулей образов контейнеров, необходимых для развертывания Сервиса без доступа к сети Интернет.

Важно!

Перед обновлением **обязательно** подготовьте сервер (см. Подготовка сервера к установке/обновлению Data Analysis Service).

Важно!

Для успешного обновления Data Analysis Service необходимо не менее 150 Гбайт свободного места на жестком диске.

Примечание:

Во время обновления текущая версия Сервиса будет удалена. Все задачи, активные на момент начала удаления, не будут завершены. Контенты не будут обработаны.
При обновлении с версии 1.0, чтобы проверить, все ли задачи завершены, выполните команду:
`ps -aux|grep cwolf_processor`
Если процесс **cwolf_processor** не использует ресурсы сервера, то активных задач нет.

Если необходимы результаты обработки конкретных контентов, дождитесь, когда эти результаты будут отображены в продукте - источнике контента, например, результаты транскрибации – в InfoWatch Activity Monitor. И только после этого начинайте процедуру обновления.

В версии 1.1 имя процесса **cwolf_processor** изменилось на **das_processor**. При обновлении с версии 1.1 используйте в командах новое имя процесса.

7.1 Чтобы обновить Data Analysis Service:

1. Удалите Data Analysis Service текущей версии. Подробнее см. Удаление Data Analysis Service.
2. Установите новую версию Сервиса. Подробнее см. Установку Data Analysis Service.
3. Для работы Поиска аналогов в Traffic Monitor:
   a. Получите у специалистов InfoWatch лицензию для новой версии Data Analysis Service.
   b. Загрузите лицензию в Консоли Управления Traffic Monitor.

      Примечание:

Другие настройки взаимодействия Data Analysis Service с Traffic Monitor или продуктами Платформы не требуются.

Важно!

После обновления Data Analysis Service функция поиска аналогов в Traffic Monitor будет работать отдельно для событий до и для событий после обновления. При поиске аналогов в новых событиях старые не будут учтены. Аналогично для старых событий не будут учтены новые.

8 Настройка Data Analysis Service

В состав Data Analysis Service входят:

- модуль протокластеризации;
- модуль транскрибации;
- модуль детектирования непристойного контента.

Включение/выключение модулей Data Analysis Service

Модуль протокластеризации включен по умолчанию. Модуль транскрибации и модуль детектирования непристойного контента можно включить во время установки Сервиса.

Чтобы включить или выключить какой-либо компонент Data Analysis Service после установки, создайте переменную окружения для необходимого компонента и присвойте ей значение. Для этого на сервере Data Analysis Service выполните следующую команду :

```
kubectl set env deploy -n infowatch das-processor-central <name>="<value>"
```

где `<name>` – это имя переменной, которое может быть:

- `PROTO_CLUSTERIZATION`
- `TRANSCRIBATION`
- `NSFW_DETECTION`

`<value>` – это значение переменной. Оно может быть равно `true` или `false`.

Пример:

Для включения модуля транскрибации выполните команду:

```
kubectl set env deploy -n infowatch das-processor-central
 TRANSCRIBATION="true"
```

Чтобы вывести все переменные окружения Data Analysis Service и их значения, выполните команду:

```
kubectl get deployment -n infowatch das-processor-central -o
jsonpath='{range .spec.template.spec.containers[0].env[*]}{.name}={.value}{"\n"}
{end}'
```

Примечание:

Чтобы включить или выключить модуль транскрибации или модуль детектирования непристойного контента, вы также можете переустановить Data Analysis Service и при установке выбрать необходимую опцию.

Примечание:

Рекомендуется выключать неиспользуемые функции в целях экономии ресурсов.

Также в этом разделе:

- Настройка производительности

- Настройка использования ресурсов сервера
- Настройка модуля транскрибации

8.1 Настройка производительности

По умолчанию на сервере функционирует один POD Data Analysis Service. Для анализа данных один POD может использовать до 5 ГБ оперативной памяти, если ее использование не ограничено. Для протокластеризации текстовых данных рекомендуется использовать один POD не более чем при 20 000 событий в час. При этом предполагается, что каждое событие в среднем содержит два вложения, из которых извлечен текст.

Если ресурсов модуля протокластеризации одного POD недостаточно для обработки текстовых данных, то в Консоли Управления Traffic Monitor кнопка **Найти аналоги** может не отображаться для новых событий.

Чтобы оценить производительность модуля протокластеризации, проанализируйте количество необработанных объектов в базе данных. Для этого на сервере Data Analysis Service в течение 2-3 часов каждые 30-60 минут выполняйте команду:

```
kubectl exec -ti -n infowatch $(kubectl get pods -n infowatch | grep postgres | awk '{print $1}') -c postgresql-central -- psql -U postgres -d das_adapter -c 'select count(*) from content_handling_progress;'
```

При росте числа необработанных объектов и нехватке производительности модуля протокластеризации рекомендовано ее увеличение. Для этого создайте необходимое число реплик POD в зависимости от количества поступающих событий и доступной оперативной памяти. Сделайте это с помощью команды:

```
kubectl scale deploy --replicas=<количество реплик> -n infowatch das-processor-central
```

8.2 Настройка использования ресурсов сервера

В Data Analysis Service вы можете ограничить использование Сервисом ресурсов сервера, например, если на одном сервере с Data Analysis Service установлены другие продукты Платформы. Данные ограничения распространяются на каждый POD Сервиса.

> **Примечание:**
>
> Получить информацию об использовании ресурсов сервера можно с помощью утилиты **htop**.

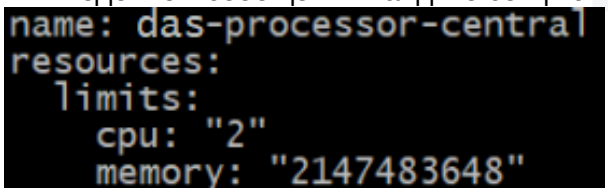
По умолчанию Data Analysis Service использует все доступные ядра центрального процессора и всю доступную оперативную память. Ограничить использование ресурсов можно с помощью скрипта `resources_setter.sh`. Для этого:

1. Перейдите в директорию `<SETUP_FOLDER>/tools/resource_management`, где `<SETUP_FOLDER>` – это директория, куда был распакован Data Analysis Service.
2. Запустите скрипт с помощью команды:
3. `./resources_setter.sh -c <CPU_LIMIT> -m <MEMORY_LIMIT>`
   где `CPU_LIMIT` – количество ядер центрального процессора. Может принимать одно из следующих значений: **0.5**, **1**, **2**, **3** и т.д.
   `MEMORY_LIMIT` – объем оперативной памяти. Указывается в байтах в формате `<value>` или в гигабайтах в формате `<value>G`.
   Параметры `-c` и `-m` можно указать, как оба сразу, так и один из них.

Пример команды запуска скрипта:

Чтобы Data Analysis Service использовал не более 4 ядер центрального процессора и не более 10 гигабайтов оперативной памяти, выполните команду: `./resources_setter.sh -c 4 -m 10G`

Чтобы проверить, какие ограничения использования ресурсов установлены в текущий момент:

1. Выполните команду:
   `kubectl get deploy -o yaml -n infowatch das-processor-central`
2. В выведенном сообщении найдите секцию `resources`:
   

   Объем оперативной памяти в сообщении указан в байтах.

8.3 Настройка модуля транскрибации

Модуль транскрибации включает в себя две модели обработки: малую и большую. Использование большой модели существенно повышает качество транскрибации, но уменьшает ее скорость. По умолчанию используется малая модель.

Вы можете увеличить производительность модуля транскрибации, используя вместо центрального процессора (CPU) графический (GPU):

- Чтобы ускорить обработку аудио-данных, включите использование GPU.
- Чтобы повысить качество транскрибации:
  - Включите использование GPU;
  - Выберите большую модель обработки.

Аппаратные и программные требования

Для использования GPU вам необходима видеокарта NVIDIA с поддержкой Compute Unified Device Architecture (CUDA).

На сервере должны быть установлены:

- Драйвер NVIDIA с версией не ниже 384.81
- NVIDIA Container Toolkit с версией не ниже 1.7.0 (для использовании встроенного GPU в системе на базе NVIDIA Tegra – не ниже 1.11.0). С инструкцией по установке можно ознакомиться на официальном сайте (страница доступна на английском языке)
- Kubernetes с версией не ниже 1.10
- NVIDIA Kubernetes Device Plugin версии 0.14.1. Его можно скачать с официального сайта (страница доступна на английском языке)

Ресурсы GPU должны быть доступны в кластере. Инструкция по настройке доступна на официальном сайте (страница доступна на английском языке).

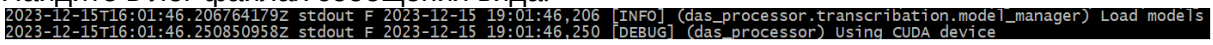
Чтобы включить использование GPU:

1. Убедитесь, что окружение удовлетворяет указанным выше требованиям:
   - Чтобы узнать версию установленного драйвера NVIDIA, используйте команду:
     `cat /sys/module/nvidia/version`

- Чтобы узнать версию установленного NVIDIA Container Toolkit, используйте команду:
  `nvidia-ctk --version`
- Чтобы узнать версию установленного Kubernetes, используйте команду:
  `kubectl version --short`
- Чтобы проверить доступность GPU на сервере, используйте команду:
  `gpu_count=$(kubectl describe -n infowatch node | grep nvidia.com | wc -l); if [[ ${gpu_count} > 0 ]]; then echo "Ok"; else echo "No GPU support"; fi`

2. Перейдите в директорию `<SETUP_FOLDER>/tools/k8s/gpu_plugin`, где `<SETUP_FOLDER>` – это директория, куда был распакован Data Analysis Service.
3. Сделайте файл скрипта настройки использования GPU исполняемым:
   `sudo chmod u+x ./enable_gpu.sh`
4. Запустите скрипт настройки использования GPU, выполнив команду:
   `sudo ./enable_gpu.sh`
5. После успешного завершения работы скрипта убедитесь, что все POD запущены, выполнив команду:
   `sudo kubectl get pods -n infowatch`
   Каждый POD должен иметь статус Running.

Чтобы убедиться, что Data Analysis Service использует GPU:

1. Установите для Сервиса режим логирования DEBUG:
   `sudo kubectl -n infowatch set env deploy das-processor-central SRV_LOG_LEVEL=DEBUG`
2. Обеспечьте передачу в Data Analysis Service аудио-данных для обработки.
3. Перейдите в директорию `/var/log/pods/infowatch_das-processor-central-…/das-processor-central`, где находятся лог-файлы.
4. Найдите в лог-файлах сообщения вида:
   

   Такие сообщения появляются при загрузке моделей модуля транскрибации. В момент загрузки моделей в лог-файл записывается сообщение `Load models`. Сообщение `Using CUDA device` означает, что Сервис использует GPU.
   Модели загружаются для обработки аудио-данных и уничтожаются, если неактивны более часа. При уничтожении в лог-файл записывается сообщение `Drop models`.

> **Примечание:**
>
> За использование GPU в Сервисе отвечает переменная окружения `USE_GPU`. Чтобы переключиться обратно на использование CPU, выполните команду:
>
> `sudo kubectl -n infowatch set env deploy das-processor-central USE_GPU=FALSE`

Чтобы выбрать модель обработки, выполните команду:

- для использования малой модели:
  `sudo kubectl set env deployment -n infowatch das-processor-central TRANSCRIBER=whisper_small_transcriber`

- для использования большой модели:
  `sudo kubectl set env deployment -n infowatch das-processor-central TRANSCRIBER=whisper_large_v1_transcriber`

> **Примечание:**
>
> Большая модель может работать только на GPU.

9 Удаление Data Analysis Service

Data Analysis Service удаляется с помощью командной строки.

> **Примечание:**
>
> Все задачи, активные на момент начала удаления, не будут завершены. Контенты не будут обработаны.
> Чтобы проверить, все ли задачи завершены, выполните команду:
> `ps -aux|grep das_processor`
> Если процесс **das_processor** не использует ресурсы сервера, то активных задач нет.
>
> Если необходимы результаты обработки конкретных контентов, дождитесь, когда эти результаты будут отображены в продукте - источнике контента, например, результаты транскрибации – в InfoWatch Activity Monitor. И только после этого начинайте процедуру удаления.

Чтобы удалить Data Analysis Service:

1. Перейдите в папку, в которую был распакован инсталлятор с последней установленной версией Сервиса.
2. Выполните команду:
   `./setup.py remove`
3. Дождитесь завершения процесса.

> **Примечание:**
>
> Инсталлятор удалит компоненты Data Analysis Service. После удаления на сервере останутся результаты работы Сервиса и база данных. Данные останутся в директориях, которые были указаны при установке Data Analysis Service.