INFOWATCH

Руководство по установке, конфигурированию и администрированию

20/01/2026

© АО “ИнфоВотч”

Тел./Факс +7 (495) 229-00-22

http://www.infowatch.ru

СОДЕРЖАНИЕ

1 Аудитория 5

2 Комплект документов 6

3 Техническая поддержка пользователей 7

4 Обзор 8

4.1 Состав Data Analysis Service: 8

5 Типы установки и обновления, совместимость с другими продуктами, аппаратные и программные требования 9

5.1 Типы установки и совместимость с другими продуктами InfoWatch 9

5.2 Типы обновления 10

5.3 Аппаратные требования 10

5.4 Программные требования 12

6 Подготовка сервера 13

7 Настройка сетевых правил доступа 20

8 Установка Data Analysis Service 21

8.1 Варианты запуска установки Data Analysis Service 22

8.2 Установка Data Analysis Service в интерактивном режиме 22

8.3 Установка Data Analysis Service в тихом режиме 27

9 Установка моделей 34

10 Обновление Data Analysis Service 38

10.1 Варианты обновления Data Analysis Service 38

10.2 Прямое обновление Data Analysis Service 39

10.3 Обновление Data Analysis Service с помощью сброса установки (reset) 43

11 Настройка Data Analysis Service 47

11.1 Настройка взаимодействия с продуктами Платформы 47

11.1.1 Включение технологий Data Analysis Service в продуктах Платформы 47

Включение транскрибации и детектирования объектов в Activity Monitor 47

Включение протокластеризации в Data Discovery 48

11.1.2 Задание нод обработки 48
11.2 Включение/выключение модулей Data Analysis Service 49
11.3 Настройка производительности 50
11.3.1 Оценка производительности модуля протокластеризации 50
11.3.2 Оценка производительности POD 51
11.3.3 Создание реплик POD 52
11.3.4 Многонодовая установка 52
11.4 Настройка использования ресурсов сервера 52
11.5 Настройка модуля транскрибации 53
11.5.1 Аппаратные и программные требования к графическому процессору 54
11.5.2 Включение использования GPU 55
11.5.3 Выбор модели транскрибации 56
11.5.4 Проверка использования GPU 57
11.6 Добавление языков для протокластеризации 57
11.7 Диагностика Data Analysis Service 58
11.7.1 Логирование 58
11.7.2 Скрипт диагностики 59
12 Удаление Data Analysis Service 60

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, обновления 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.

3 Техническая поддержка пользователей

При возникновении проблем и 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.

4 Обзор

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

4.1 Состав Data Analysis Service:

- **Модуль протокластеризации текстовых данных**
  Данный модуль позволяет анализировать текстовые документы и разбивать их на тематические группы. В ходе анализа для каждого документа и каждой группы определяется набор тегов - ключевых слов, по которым можно понять тематику и примерное содержание документа/группы документов.
  Модуль протокластеризации используется для работы следующий функций:
  - *Поиск аналогов* в InfoWatch Traffic Monitor. Подробнее см. "*InfoWatch Traffic Monitor. Руководство пользователя. Решение задач при работе в консоли Traffic Monitor. Работа с объектами перехвата. Просмотр событий. Просмотр краткой формы события. Поиск аналогов*";
  - *Поиск похожих файлов* в InfoWatch Data Discovery. Подробнее см. "*InfoWatch Data Discovery. Руководство пользователя. Раздел "Хранение файлов". Вкладка "Похожие файлы*".
- **Модуль транскрибации**
  Данный модуль позволяет конвертировать аудиозапись в текст. Модуль транскрибации используется при работе с аудиозаписями в InfoWatch Activity Monitor. Подробнее см. "*InfoWatch Activity Monitor. Руководство пользователя. Интерфейс Центра расследований. Раздел "Мониторинг". Вкладка "Аудиозаписи". Работа с аудиозаписями*".
- **Модуль детектирования объектов на изображениях**
  Данный модуль позволяет распознавать объекты на снимках экрана, в том числе непристойный контент. Детектирование объектов на изображениях используется в InfoWatch Activity Monitor. Подробнее см. "*InfoWatch Activity Monitor. Руководство пользователя. Интерфейс Центра расследований. Раздел "Мониторинг". Вкладка "Наблюдение". Таймлайн*".

5 Типы установки и обновления, совместимость с другими продуктами, аппаратные и программные требования

5.1 Типы установки и совместимость с другими продуктами InfoWatch

Data Analysis Service поддерживает интеграцию со следующими продуктами InfoWatch:

- **Traffic Monitor начиная с версии 7.8**
- **Activity Monitor 3.3**
- **Data Discovery 2.1**

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

Основой для установки Сервиса является Платформа. Инсталлятор Data Analysis Service содержит и Платформу, и компоненты Сервиса. Обе части обязательны к установке.
И компоненты Сервиса, и Платформа поддерживают установку в двух режимах:

- `central` ;
- `office` .

Режимы установки Платформы и Сервиса вы указываете при запуске инсталлятора. Если оба режима не указаны пользователем, инсталлятор по умолчанию использует значение `central` для обоих компонентов.

Платформенные продукты поддерживают **распределенную установку**: центральная (`central`) и офисные (`office`) ноды объединены в кластер.
Вы можете установить Data Analysis Service как отдельно, так и в кластере с платформенными продуктами. При установке в кластере поддерживается установка как на центральную, так и на офисную ноду.

> **Важно!**
>
> При распределенной установке примите во внимание следующие ключевые аспекты:
>
> - Data Analysis Service и платформенные продукты в кластере должны быть установлены на одной версии Платформы – в номерах версий должны совпадать первые два числа. Версия центральной ноды Платформы должна быть самой новой в кластере.
> - Если вы устанавливаете в кластер Сервис или платформенный продукт на новой версии Платформы, то после этого обновите все остальные экземпляры Сервиса и продукты.
> - В одном кластере можно развернуть только одну центральную ноду Платформы и Data Analysis Service в режиме `central` только в единственном экземпляре.

- Если в кластере уже установлен Data Analysis Serviceв режиме `central`, еще одна установка Сервиса в режиме `central` в том же кластере будет прервана ошибкой.

При установке Data Analysis Service 1.4 устанавливается Платформа версии 2.3.1.

Data Analysis Service может быть установлен на одном сервере со следующими продуктами InfoWatch:

- **Traffic Monitor начиная с версии 7.8**
- **Activity Monitor начиная с версии 3.3**
- **Data Discovery 2.1**
- **Vision 3.5**

Подробнее о процессе установки см. "Установка Data Analysis Service".

5.2 Типы обновления

Data Analysis Service поддерживает:

- прямое обновление;
- обновление с помощью сброса и переустановки на старые данные.

Подробнее о процессе обновления и поддерживаемых вариантах см. "Обновление Data Analysis Service".

5.3 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 300 Гбайт свободного места. От 15 Гбайт свободного места на диске с директорией `/tmp`. Минимальное количество последовательных операций ввода-вывода – от 50 IOPS.

Требования к аппаратному обеспечению зависят от набора используемых модулей Data Analysis Service и необходимой скорости обработки данных. Далее приведены минимальные рекомендуемые аппаратные требования для работы модулей одного POD Сервиса.

	Процессор, используемый	Модель, используемая при транск	Производительность модуля	Рекомендуемое аппаратное обеспечение

	для обработки данных	рибации		При установке отдельно от других платформенных продуктов		При установке на один сервер с другим платформенным продуктом	
			Количество обрабатываемых текстовых контентов в час [1)]	CPU, количество ядер/потоков	ОЗУ, ГБ	CPU, количество ядер/потоков	ОЗУ, ГБ
Модуль протокластеризации	CPU	–	30 000	4	16	2	4
–	–	–	Отношение времени обработки к длине аудиозаписи	–	–	–	–
Модуль транскрибации	CPU	малая	1:3	6	16	4	6
	GPU с 10 ГБ видеопамяти	малая	1:60	2	16	–	–
		большая	1:30	2	16	–	–

[1)] Для текстовых контентов со средним размером 17,5 КБ

Важно!

Аппаратные и программные требования к **графическому процессору** (**GPU**) для его использования при транскрибации приведены в статье Настройка модуля транскрибации.

При расчете аппаратных требований для совместной установки Data Analysis Service и Traffic Monitor рекомендуется суммировать их отдельные требования.

Примечание:

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

5.4 Программные требования

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

Тип ПО	Варианты
Операционная система	• Astra Linux Special Edition: • версии 1.6.10 "Смоленск"; • версии 1.7.5 "Смоленск"; • версии 1.7.5 "Орел"; • версии 1.8 "Смоленск"; • Red Hat Enterprise Linux: • версии 7.9; • версии 8.9; • Oracle Linux: • версии 7.9; • версии 8.4; • РЕД ОС 7.3.3; • ОС Альт: • Сервер 10.2; • СП 10. **Важно!** Модули **Netfilter** и **NAT** должны быть загружены в ядро ОС.

Примечание:

Data Analysis Service использует PostgreSQL 16.4.
СУБД поставляется вместе с продуктом и не требует отдельной установки.

6 Подготовка сервера

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. 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.
3. Для корректной работы Сервиса настройте сетевые правила доступа (подробнее см. "Настройка сетевых правил доступа").
4. 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:

- **Для ОС Red Hat Enterpise Linux, Oracle Linux и РЕД ОС**

```
yum install conntrack
```

- **Для ОС Astra Linux и ОС Альт**

```
apt-get install conntrack
```

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

- **Для ОС Red Hat Enterpise Linux, Oracle Linux, РЕД ОС и ОС Альт**

```
rpm -Uhv ./<имя_пакета>
```

- **Для ОС Astra Linux**

```
dpkg -i ./<имя_пакета>
```

5. Если вы используете антивирус, настройте его для совместной работы с Data Analysis Service (см. статью Базы знаний "Настройка антивируса для продуктов Платформы").
6. **На сервере под управлением Red Hat Enterprise Linux и РЕД ОС настройте SELinux.**

   Для корректной работы используемого `containerd` **обновите** политики SELinux или **отключите** систему контроля доступа:

   - **Чтобы обновить политики SELinux**, скачайте и установите пакет `container-selinux`.
     Вы можете сделать это, выполнив команду:

     ```
     dnf install container-selinux
     ```

     > **Примечание:**
     >
     > Пакет не входит в состав дистрибутива продукта.
     >
     > Пакет может отсутствовать в репозиториях. В этом случае загрузите и установите его вручную.

   - **Чтобы отключить SELinux**:
     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
        ```

7. Если вы устанавливаете Сервис на ОС Альт, установите пакеты:
   - **python-modules-json**;
   - **python-modules-distutils**;
   - **python-modules-sqlite3**.

   > **Примечание:**
   >
   > Пакеты не входят в состав дистрибутива продукта.

Пакеты могут отсутствовать в репозиториях. В этом случае загрузите и установите требуемые пакеты вручную.

8. **Если вы устанавливаете Data Analysis Service на ОС Альт СП 10, выполните следующие действия:**

   a. В файле `/etc/apt/sources.list.d/altsp.list` раскомментируйте три последних строки, указывающие репозитории.

   Пример содержимого altsp.list с раскомментированными строками

   ```
   # update.altsp.su (IVK, Moscow)

   # ALT Certified 10
   #rpm [cert8] ftp://update.altsp.su/pub/distributions/ALTLinux c10f/branch/x86_64
   classic gostcrypto
   #rpm [cert8] ftp://update.altsp.su/pub/distributions/ALTLinux c10f/branch/x86_64-
   i586 classic
   #rpm [cert8] ftp://update.altsp.su/pub/distributions/ALTLinux c10f/branch/noarch
   classic

   rpm [cert8] http://update.altsp.su/pub/distributions/ALTLinux c10f/branch/x86_64
   classic gostcrypto
   rpm [cert8] http://update.altsp.su/pub/distributions/ALTLinux c10f/branch/x86_64-
   i586 classic
   rpm [cert8] http://update.altsp.su/pub/distributions/ALTLinux c10f/branch/noarch
   classic
   ```

   b. Обновите пакеты с помощью команды:

   ```
   apt-get update
   ```

   c. Установите требуемые пакеты с помощью команд:

   ```
   apt-get install python-modules-sqlite3
   apt-get install conntrack-tools
   ```

   d. Установите менеджер контейнеров Podman с помощью команды:

   ```
   apt-get install podman
   ```

   e. С помощью Podman загрузите образы:

   ```
   podman pull registry.altlinux.org/k8s-c10f2/kube-apiserver:v1.31.1
   podman pull registry.altlinux.org/k8s-c10f2/kube-controller-manager:v1.31.1
   podman pull registry.altlinux.org/k8s-c10f2/kube-scheduler:v1.31.1
   podman pull registry.altlinux.org/k8s-c10f2/kube-proxy:v1.31.1
   podman pull registry.altlinux.org/k8s-c10f2/pause:3.10
   ```

```
podman pull registry.altlinux.org/k8s-c10f2/etcd:3.5.15-0
podman pull registry.altlinux.org/k8s-c10f2/coredns:v1.11.3
podman pull registry.k8s.io/pause:3.10
```

f. Установите Kubernetes с помощью команд:

```
apt-get install kubernetes-kubeadm kubernetes-kubelet kubernetes-crio cri-tools
systemctl enable --now crio kubelet
swapoff -a
```

g. Для корректного создания кластера отключите использование прокси-соединения:

```
unset http_proxy
unset https_proxy
```

h. Создайте кластер:

```
kubeadm init --pod-network-cidr=10.244.0.0/16 --kubernetes-version=1.31.1 --image-repository=registry.altlinux.org/k8s-c10f2
```

i. После создания кластера на экран будет выведена инструкция. Выполните предложенные действия.

> **Примечание:**
>
> **Рекомендуется** выполнить действия из инструкции.
> Если вы не будете выполнять все действия из инструкции, чтобы продолжить установку продукта, выполните команду для настройки переменной окружения:
>
> ```
> export KUBECONFIG=/etc/kubernetes/admin.conf
> ```
>
> Команда позволит использовать `kubectl` в рамках текущей сессии подключения. При следующем подключении или перезагрузке выполните команду повторно.

j. Установите для Kubernetes сетевой провайдер Flannel, для этого загрузите образы:

```
podman pull registry.altlinux.org/k8s-c10f2/flannel:v0.25.7
podman pull registry.altlinux.org/k8s-c10f2/flannel-cni-plugin:v1.4.0-flannel1
```

k. Загрузите манифест Flannel:

```
wget https://github.com/flannel-io/flannel/releases/latest/download/kube-flannel.yml
```

l. Отредактируйте файл-манифест:

i. Строку `image: docker.io/flannel/flannel-cni-plugin:v1.1.2` замените на:

Строка для замены:

```
image: registry.altlinux.org/k8s-c10f2/flannel-cni-plugin:v1.4.0-flannel1
```

ii. Строки `image: docker.io/flannel/flannel:v0.22.0` замените на:

Строка для замены:

```
image: registry.altlinux.org/k8s-c10f2/flannel:v0.25.7
```

> **Примечание:**
>
> В конфигурационном файле две одинаковых строки вида `image: docker.io/flannel/flannel:v0.22.0`, замените обе.
>
> В связи с периодическими обновлениями Flannel версии в заменяемых строках могут отличаться.

m. Откройте на редактирование конфигурацию CoreDNS:

```
kubectl edit cm -n kube-system coredns
```

n. Добавьте следующее содержимое:

```
     rewrite stop {
        name substring iwplatform cluster answer auto
     }
```

Вид конфигурационного файла после редактирования:

```
Corefile: |
  .:53 {
      log
      errors
      health {
         lameduck 5s
      }
      rewrite stop {
         name substring iwplatform cluster answer auto
      }
      ready
      kubernetes cluster.local in-addr.arpa ip6.arpa {
         pods insecure
         fallthrough in-addr.arpa ip6.arpa
```

```
            ttl 30
        }
        prometheus :9153
        forward . /etc/resolv.conf {
            max_concurrent 1000
        }
        cache 30
        loop
        reload
        loadbalance
    }
```

o. Сохраните изменения и закройте редактор.
p. Примените изменения в манифесте и запустите Flannel с помощью команды:

```
kubectl apply -f kube-flannel.yml
```

q. Дождитесь запуска всех подов.
Вы можете проверить их статусы с помощью команды:

```
kubectl get pods -A
```

Все поды должны быть в статусе `Running`.
r. Смените директорию для временных файлов:

```
export TMPDIR=/var/tmp
```

9. 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
```

10. Убедитесь, что на сервере установлены актуальные дата и время.

7 Настройка сетевых правил доступа

Для корректной работы должны быть разрешены соединения:

Соединение	Порт	Описание
Соединения с внешними системами		
Сервер Traffic Monitor → Сервер Data Analysis Service	TCP 30050	Порт сервера Data Analysis Service. Используется для обмена данными с Traffic Monitor по протоколу HTTP.
	TCP 30450	Порт сервера Data Analysis Service. Используется для обмена данными с Traffic Monitor по протоколу HTTPS.
Сервер Data Analysis Service → Сервер Traffic Monitor	TCP 443	Порт сервера Data Analysis Service. Используется для обмена данными с Traffic Monitor по протоколу HTTPS.
Работа в кластере		
Сервер Платформы в режиме `office` → Сервер Платформы в режиме `central`	TCP 6443 TCP 2379, 2380 TCP 10251, 10252, 10255	Порты сервера Платформы в режиме `central`. Используются для взаимодействия Платформы в режиме `office` с Платформой в режиме `central`
Сервер Платформы в режиме `central` → Сервер Платформы в режиме `office`	TCP 10251, 10255 TCP 30000–32767	Порты сервера Платформы в режиме `office`. Используются для взаимодействия Платформы в режиме `central` с Платформой в режиме `office`
Сервер Платформы в режиме `central` ← → Сервер Платформы в режиме `office`	UDP 8472 TCP 10250	Используются для туннелирования трафика между узлами кластера
Сервер Платформы в режиме `central` ← → Сервер Платформы в режиме `office`	TCP 10256	Порт должен быть доступен на всех нодах кластера. Используется для проверки состояния компонента `kube-proxy`

8 Установка Data Analysis Service

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 `1.4.0.38`, ей соответствует архив `iw_das_setup _1.4.0.38.tar.xz`. Архив содержит набор бинарных модулей образов контейнеров, необходимых для развертывания Сервиса без доступа к сети Интернет. Также архив содержит программу установки.
Для установки Сервиса вам также понадобятся ресурсы и модели для тех технологий, которые вы собираетесь использовать. Они входят в состав дистрибутива в виде следующих архивов:

- `proto_clusterization_resources.tar.gz` - для технологии протокластеризации;
- `transcription_resources.tar.gz`, `hf_faster_whisper_small.tar.gz` (для использования малой модели) и `hf_faster_whisper_large_v1.tar.gz` (для использования большой модели) - для технологии транскрибации;
- `image_resources.tar.gz` - для технологии детектирования объектов на изображениях.

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

Вы можете установить Data Analysis Service:

- на отдельном сервере;
- в кластере из серверов Data Analysis Service;
- на сервере совместно с продуктами InfoWatch, указанными в статье "Типы установки и обновления, совместимость с другими продуктами, аппаратные и программные требования";
- в кластере с платформенными продуктами.

Важно!

Если вы устанавливаете Data Analysis Service в одном кластере с другими продуктами InfoWatch, то убедитесь, что версия Платформы совпадает у всех продуктов на всех нодах. В противном случае продукты, установленные на старой версии Платформы, необходимо обновить.

Чтобы узнать версию Платформы устанавливаемого Сервиса:

1. Распакуйте архив с дистрибутивом Data Analysis Service. Для этого перейдите в директорию с архивом и выполните команду:

   Пример для версии 1.4.0.38:

   ```
   tar -xvf iw_das_setup_1.4.0.38.tar.xz
   ```

2. Выполните команду:

   ```
   cat platform/setup.yaml | grep version
   ```

Важно!

Необходимо, чтобы все ноды кластера были доступны друг для друга в сети. Для этого должны быть открыты необходимые порты (см. "Настройка сетевых правил доступа").

Также перед установкой убедитесь, что:

- настроено разрешение имен нод средствами DNS
  или
- в файле `/etc/hosts` каждой ноды определено соответствие между адресами и именами нод.

Для проверки доступности нод выполните команду между нодами кластера:

```
ping <имя ноды>
```

8.1 Варианты запуска установки Data Analysis Service

Вы можете запустить установку Data Analysis Service:

- **в интерактивном режиме** - инсталлятор последовательно запросит параметры установки (см. "Установка Data Analysis Service в интерактивном режиме");
- **в тихом режиме** - указав параметры установки в команде запуска инсталлятора (см. "Установка Data Analysis Service в тихом режиме").

В обоих вариантах запуска инсталлятор поддерживает и совместную установку с платформенными продуктами на одном сервере, и распределенную установку в кластере.

8.2 Установка Data Analysis Service в интерактивном режиме

Примечание:

Если вы устанавливаете Data Analysis Service на сервер, на котором уже установлены какие-либо платформенные продукты, инсталлятор не запросит часть параметров, так как эти директории и настройки уже были указаны при первичной установке Платформы.

1. Проверьте, соответствует ли сервер аппаратным и программным требованиям (см. "Типы установки и обновления, совместимость с другими продуктами, аппаратные и программные требования").

   Примечание:

   В ходе установки инсталлятор проверит наличие:

   - не менее 100 ГБ на диске для директорий размещения данных ClickHouse, PostgreSQL, NATS и бинарных данных. Эти директории необходимо указать при выполнении инструкций по установке. Если директории для хранения данных находятся на разных дисках, то проверятся наличие 100 ГБ свободного места на каждом из них;
   - не менее 20 ГБ на диске для директорий с образами;
   - не менее 8 ГБ ОЗУ;
   - не менее 2 ядер процессора;

- поддержки инструкции SSE 4.2;
- установленного пакета **`conntrack.`**

2. Выполните шаги по подготовке сервера к установке Data Analysis Service (см. "Подготовка сервера").
3. Создайте новую директорию на сервере (например, `DAS`):

```
mkdir DAS
```

4. Скопируйте архив **`iw_das_setup_1.4.x.xx.tar.xz`** в созданную директорию.
   В нашем примере будет использована версия `1.4.0.38`, соответственно, архив – `iw_das_setup _1.4.0.38.tar.xz`.
5. Перейдите в директорию, в которую был скопирован архив.
6. Распакуйте архив с дистрибутивом в эту директорию:

```
tar -xvf iw_das_setup_1.4.0.38.tar.xz
```

7. **Для установки Data Analysis Service на ОС Альт СП 10:**

   Вы должны использовать Kubernetes из репозиториев ОС.
   Для этого:

   a. Выполните действия для ОС Альт СП 10 из статьи Подготовка сервера.
      Если вы уже выполнили подготовку, пропустите этот шаг.
   b. Выполните команду:

```
kubectl taint nodes <hostname_ноды> node-role.kubernetes.io/control-plane:NoSchedule
-
```

   c. На следующем шаге при запуске инсталлятора дополнительно укажите опцию **`--externalK8S=True`**
8. Запустите программу установки с помощью команды вида:

```
./setup.py install --nodemode=<режим_установки_Платформы> --
productmode=<режим_установки_продукта>
```

> **Примечание:**
>
> Чтобы установить:
>
> - **на центральную ноду Платформы Data Analysis Service в режиме central:**
>
> ```
> ./setup.py install --nodemode=central --productmode=central
> ```

- **на центральную ноду Платформы Data Analysis Service в режиме office:**

```
./setup.py install --nodemode=central --productmode=office
```

- **на офисную ноду Платформы Data Analysis Service в режиме central:**

```
./setup.py install --nodemode=office --productmode=central
```

- **на офисную ноду Платформы Data Analysis Service в режиме office:**

```
./setup.py install --nodemode=office --productmode=office
```

Если вы не укажете режимы установки Платформы и Data Analysis Service, инсталлятор по умолчанию выберет режим `central`.

Если вы устанавливаете Data Analysis Service на сервер первым и укажете только параметр `--nodemode`, инсталлятор использует такое же значение и для параметра `--productmode`.

Важно!

Во время установки могут быть недоступны продукты Платформы, установленные на сервере.

9. Ознакомьтесь с условиями лицензионного соглашения. Лицензионное соглашение содержит несколько страниц. Для перехода на следующую страницу используйте клавишу **Enter**.
10. Чтобы принять лицензионное соглашение, введите "**y**" и нажмите **Enter**.

Примечание:

Если отклонить лицензионное соглашение, инсталлятор прервет процесс и завершит работу.

11. Если вы устанавливаете **центральную ноду Платформы** (`--nodemode=central`), введите IP-адрес сетевого интерфейса для взаимодействия с кластером в формате IPv4: "xxx.xxx.xxx.xxx" (по умолчанию: `0.0.0.0`) и нажмите **Enter**.
Если указать `0.0.0.0`, будут использованы все доступные сетевые интерфейсы.

Примечание:

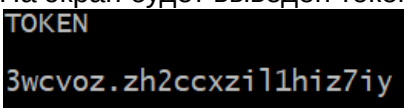
Здесь и далее для использования значений, предложенных по умолчанию, нажмите **Enter** без ввода значений.

12. Если вы устанавливаете **офисную ноду Платформы** (`--nodemode=office`):

a. Укажите адрес центральной ноды Платформы в формате IP-адреса или доменного имени (например: `192.0.2.0` или `host.example.com`).
b. Укажите токен для подключения к центральной ноде Платформы.
Чтобы получить токен для подключения к центральной ноде Платформы:
   i. Подключитесь по ssh к консоли центральной ноды Платформы;
   ii. Выполните в терминале команду для получения токена:

```
kubeadm token list
```

На экран будет выведен токен:



Токен в нашем примере: `3wcvoz.zh2ccxzil1hiz7iy`

   iii. Укажите лейбл ноды (по умолчанию: `office`) и нажмите **Enter**.
Лейбл может состоять из букв латинского алфавита и/или цифр. Длина лейбла не должна превышать 32 символа.

13. Выделите объем оперативной памяти для размещения данных Clickhouse (по умолчанию: `80%`) и нажмите **Enter**.
14. Укажите путь для размещения данных Clickhouse (по умолчанию: `/opt/chdata`) и нажмите **Enter**.
15. Укажите путь для размещения данных PostgreSQL (по умолчанию: `/opt/pgdata`) и нажмите **Enter**.
16. Укажите путь для размещения данных NATS (по умолчанию: `/opt/natsdata`) и нажмите **Enter**.
17. Укажите путь для размещения бинарных данных (по умолчанию: `/opt/dsdata`) и нажмите **Enter**.
18. Если вы устанавливаете **центральную ноду Платформы**, укажите порт подключения к веб-интерфейсу (по умолчанию: `443`) и нажмите **Enter**.

> **Важно!**
>
> При установке на один сервер с Traffic Monitor укажите порт отличный от 443, так как по умолчанию порт 443 используется в качестве порта подключения к веб-интерфейсу Traffic Monitor.

19. Дождитесь завершения установки Платформы.
20. Ознакомьтесь с отчетом об установке Платформы.



21. Введите через запятую лейблы нод для взаимодействия с платформенными продуктами (по умолчанию задается одна нода с лейблом `central`) и нажмите **Enter**.
Data Analysis Service будет обрабатывать очередь сообщений каждой указанной ноды.

Если такое взаимодействие не планируется, нажмите **Enter**.

> **Примечание:**
>
> InfoWatch Activity Monitor отправляет сообщения в очередь NATS на центральную ноду кластера.
>
> InfoWatch Data Discovery отправляет сообщения в очередь NATS на ту ноду кластера, на которой он установлен.

Чтобы узнать лейбл ноды:

Выполните команду:

```
kubectl get node -n infowatch -o jsonpath="{range .items[*]}
{.metadata.annotations.flannel\.alpha\.coreos\.com/public\-ip}
{.metadata.labels.kubernetes\.io/hostname} {.metadata.labels.iwnode}{'\n'}{end}"
```

В результате будет выведен список нод кластера с указанием их лейблов.

> **Примечание:**
>
> Изменить список лейблов нод для взаимодействия можно после установки. Подробнее см. "Задание нод обработки".

22. Укажите путь для размещения ресурсов Data Analysis Service (по умолчанию: `/opt/das_data`) и нажмите **Enter**.
23. Введите "**y**", чтобы включить модуль транскрибации, или "**n**", чтобы оставить его выключенным, (по умолчанию: "**n**") и нажмите **Enter**.
24. Введите "**y**", чтобы включить модуль детектирования объектов на изображениях, или "**n**", чтобы оставить его выключенным, (по умолчанию: "**n**") и нажмите **Enter**.
25. Дождитесь окончания процесса установки Data Analysis Service.
26. Ознакомьтесь с отчетом об установке Сервиса:



27. Чтобы вывести на экран информацию о Платформе и продуктах в кластере, выполните команду:

```
./setup.py showproducts
```

Вывод команды должен содержать строку `DbState: commited`. Например:

```
platform Title: Infowatch Platform Version: 2.3.1.424 Node: my-server(central) Mode:
central DbState: commited
```

28. Убедитесь, что все сервисы запущены, выполнив команду:

```
kubectl get pods -n infowatch
```

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

29. При установке Data Analysis Service на сервер, где уже установлен Traffic Monitor, убедитесь, что в конфигурационном файле `/opt/iw/tm5/etc/consul/consul.json` в поле "`bind_addr`" указан IP-адрес данного сервера.
30. Установите модели для технологий, которые будете использовать. Подробнее см. статью "Установка моделей".

> **Важно!**
>
> Если модели для какой-либо технологии не будут установлены, то эта технология не будет работать.

> **Важно!**
>
> Чтобы начать использование Data Analysis Service, **настройте его взаимодействие** с :
>
> - **InfoWatch Traffic Monitor** (подробнее см. "*InfoWatch Traffic Monitor. Руководство администратора. Настройка интеграции с Data Analysis Service*")
> и/или
> - продуктами Платформы, например, **InfoWatch Activity Monitor** или **InfoWatch Data Discovery** (подробнее см. в статье "Настройка взаимодействия с продуктами Платформы" и в документации на соответствующий продукт Платформы).

8.3 Установка Data Analysis Service в тихом режиме

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

Перед запуском инсталлятора Data Analysis Service:

1. Проверьте, соответствует ли сервер аппаратным и программным требованиям (см. "Типы установки и обновления, совместимость с другими продуктами, аппаратные и программные требования").
2. Выполните шаги по подготовке сервера к установке Data Analysis Service (см. "Подготовка сервера").
3. Создайте новую директорию на диске (например, `DAS`):

```
mkdir DAS
```

4. Скопируйте архив с дистрибутивом Data Analysis Service **`iw_das_setup_1.4.x.xx.tar.xz`** в созданную директорию.
   В нашем примере будет использована версия `1.4.0.38`, соответственно, архив – `iw_das_setup_1.4.0.38.tar.xz`.
5. Перейдите в директорию, в которую был скопирован архив.
6. Распакуйте архив в эту директорию:

```
tar -xvf iw_das_setup_1.4.0.38.tar.xz
```

Ознакомьтесь со списком ключей в таблице ниже, с их помощью вы можете указать инсталлятору параметры установки продукта. Указанные значения по умолчанию - значения, предлагаемые инсталлятором в интерактивном режиме.

> **Примечание:**
>
> Если вы не укажете режимы установки Платформы и продукта, инсталлятор использует значение `central`.

Ключ	Значение по умолчанию	Доступные значения	Описание
`--nodemode`	`central`	`central`, `office`	Режим установки Платформы
`--productmode`	`central`	`central`, `office`	Режим установки Data Analysis Service
`--licagree`	`n`	`y`, `yes`, `n`, `no`	Принятие лицензионного соглашения
`--chmemsize`	`80%`	1-100% или числовое значение Если не указать %, указанное значение будет принято за объем в ГБ	Объем оперативной памяти, используемой для данных Clickhouse
`--chdbpath`	`/opt/chdata`		Путь для размещения данных ClickHouse
`--pgdbpath`	`/opt/pgdata`		Путь для размещения файлов базы данных PostgreSQL

Ключ	Значение по умолчанию	Доступные значения	Описание
`--natspath`	`/opt/natsdata`		Путь для размещения данных NATS
`--dsdatapath`	`/opt/dsdata`		Путь для размещения бинарных данных.
`--disable-check-free-disk`	-	-	Отключение проверки свободного места на диске: • 20 ГБ для размещения образов; • 100 ГБ для размещения данных в указанных выше директориях **Примечание**: Если директории для хранения данных находятся на разных дисках, то проверятся наличие 100 ГБ свободного места на каждом из них
`--disable-check-total-ram`	-	-	Отключение проверки объема ОЗУ на компьютере в 8 ГБ
`--advertise`	`0.0.0.0`		IP-адрес сетевого интерфейса для взаимодействия с кластером. Указывайте при установке центральной ноды Платформы
`--guiport`	`443`		Порт подключения к веб-интерфейсу. Указывайте при установке центральной ноды Платформы. **Важно!** При установке на один сервер с Traffic Monitor указывайте порт отличный от 443, так как по умолчанию порт 443 используется в качестве порта подключения к веб-интерфейсу Traffic Monitor.
`--cluster`			Адрес центральной ноды Платформы в кластере. Указывайте при установке офисной ноды Платформы
`--clustertoken`			Токен для подключения к центральной ноде Платформы. Чтобы получить токен, в терминале центральной ноды Платформы выполните команду:

Ключ	Значение по умолчанию	Доступные значения	Описание
			`kubeadm token list` Указывайте при установке офисной ноды Платформы
`--nodelabel`	`office`		Лейбл ноды. Указывайте при установке офисной ноды Платформы
`--pclust-enable`	`y`	`y, n`	Включение модуля протокластеризации
`--transcription-enable`	`n`	`y, n`	Включение модуля транскрибации
`--imcdet-enable`	`n`	`y, n`	Включение модуля детектирования объектов на изображениях
`--subscribe-to`	`central`		Список лейблов нод для взаимодействия с платформенными продуктами. Если таких нод несколько, укажите их лейблы через запятую, взяв весь список в двойные кавычки, например, `--subscribe-to="central, office1, office2"`. Data Analysis Service будет обрабатывать очередь сообщений каждой указанной ноды. **Примечание:** InfoWatch Activity Monitor отправляет сообщения в очередь NATS на центральную ноду кластера. InfoWatch Data Discovery отправляет сообщения в очередь NATS на ту ноду кластера, на которой он установлен. **Чтобы узнать лейбл ноды:** На любой ноде кластера выполните следующую команду: `kubectl get node -n infowatch -o jsonpath="{range .items[*]}{.metadata.annotations.flannel\.alpha\.coreos\.com/public\-ip}{.metadata.labels.kubernetes\.io/hostname}{.metadata.labels.iwnode}{'\n'}{end}"`

Ключ	Значение по умолчанию	Доступные значения	Описание
			В результате будет выведен список нод кластера с указанием их лейблов. **Примечание:** Изменить список лейблов ноды для взаимодействия можно после установки. Подробнее см. "Задание нод обработки".
`--das-data-path`	`/opt/das_data`		Путь для размещения ресурсов Data Analysis Service

Чтобы установить Data Analysis Service:

1. **Для установки Data Analysis Service на ОС Альт СП 10:**

   Вы должны использовать Kubernetes из репозиториев ОС.
   Для этого:

   a. Выполните действия для ОС Альт СП 10 из статьи Подготовка сервера.
   Если вы уже выполнили подготовку, пропустите этот шаг.
   b. Выполните команду:

   ```
   kubectl taint nodes <hostname_ноды> node-role.kubernetes.io/control-plane:NoSchedule
   -
   ```

   c. На следующем шаге при запуске инсталлятора дополнительно укажите опцию `--externalK8S=True`
2. Запустите программу установки Data Analysis Service с указанием параметров.
   Пример команды установки центральной ноды Платформы и Data Analysis Service в режиме central:

   ```
   ./setup.py install \
   --nodemode=central \
   --productmode=central \
   --licagree=yes \
   --advertise=0.0.0.0 \
   --chmemsize=60% \
   --chdbpath=/opt/chdata \
   --pgdbpath=/opt/pgdata \
   --natspath=/opt/natsdata \
   --dsdatapath=/opt/dsdata \
   --guiport=443 \
   --pclust-enable=y \
   --transcription-enable=y \
   --imcdet-enable=y \
   --das-data-path=/opt/das_data
   ```

   Пример команды установки офисной ноды Платформы и Data Analysis Service в режиме office:

```
./setup.py install \
--nodemode=office \
--productmode=office \
--licagree=yes \
--cluster=10.60.25.92 \
--clustertoken=3wcvoz.zh2ccxzil1hiz7iy \
--nodelabel=office123
--chmemsize=60% \
--chdbpath=/opt/chdata \
--pgdbpath=/opt/pgdata \
--natspath=/opt/natsdata \
--dsdatapath=/opt/dsdata \
--pclust-enable=y \
--transcription-enable=n \
--imcdet-enable=n \
--subscribe-to=central \
--das-data-path=/opt/das_data
```

Пример команды установки Data Analysis Service в режиме central на сервер с установленной Платформой:

```
./setup.py install \
--productmode=central \
--licagree=yes \
--pclust-enable=y \
--transcription-enable=n \
--imcdet-enable=n \
--subscribe-to="central, office1, office2" \
--das-data-path=/opt/das_data
```

3. Если вы указали все обязательные параметры, инсталлятор запустит установку.
   Если вы не указали часть параметров, инсталлятор запустит интерактивный режим и попросит задать значения.
4. Дождитесь завершения установки Data Analysis Service.
5. Чтобы вывести на экран информацию о Платформе и продуктах в кластере, выполните команду:

```
./setup.py showproducts
```

6. Убедитесь что все сервисы запущены, выполнив команду:

```
kubectl get pods -n infowatch
```

   Команда выведет список сервисов. Каждый сервис должен иметь статус `Running`.
7. При установке Data Analysis Service на сервер, где уже установлен Traffic Monitor, убедитесь, что в конфигурационном файле `/opt/iw/tm5/etc/consul/consul.json` в поле "`bind_addr`" указан IP-адрес данного сервера.
8. Установите модели для технологий, которые будете использовать. Подробнее см. статью "Установка моделей".

Важно!

Если модели для какой-либо технологии не будут установлены, то эта технология не будет работать.

Важно!

Чтобы начать использование Data Analysis Service, **настройте его взаимодействие** с:

- **InfoWatch Traffic Monitor** (подробнее см. "InfoWatch Traffic Monitor. *Руководство администратора. Настройка интеграции с Data Analysis Service*")
  и/или
- продуктами Платформы, например, **InfoWatch Activity Monitor** или **InfoWatch Data Discovery** (подробнее см. в статье "Настройка взаимодействия с продуктами Платформы" и в документации на соответствующий продукт Платформы).

9 Установка моделей

Для работы технологий Data Analysis Service должны быть установлены модели:

- для протокластеризации – `proto_clusterization_resources.tar.gz`;
- для транскрибации с использованием малой модели:
  - `transcription_resources.tar.gz`;
  - `hf_faster_whisper_small.tar.gz`;
- для транскрибации с использованием большой модели:
  - `transcription_resources.tar.gz`;
  - `hf_faster_whisper_large_v1.tar.gz`;
- для детектирования объектов на изображениях – `image_resources.tar.gz`.

Чтобы установить модели:

1. Скопируйте из дистрибутива архивы с нужными вам моделями на сервер, например, в директорию `../DAS/tools/models_installation`.
2. Перейдите в директорию `../DAS/tools/models_installation`, где располагается скрипт распаковки архивов с моделями.
3. Убедитесь, что на диске с директорией `/tmp` есть 15 ГБ свободного места. Это необходимо для успешной распаковки архивов.
4. Запустите скрипт распаковки архивов `unpack_and_deploy_models.sh`.
   Для версии Data Analysis Service 1.4.0:

   Выполните команду:

   ```
   ./unpack_and_deploy_models.sh [-i <input_directory>] [-t <transcriber_name>] [-n
   <node_label>]
   ```

   со следующими аргументами:
   `[-i <input_directory>]` – директория, в которой находятся архивы моделей. Если аргумент не указан, то будут распакованы архивы из той директории, где находится скрипт;
   `[-t <transcriber name>]` – модель транскрибации. Возможные значения:
   `whisper_small_transcriber` - для загрузки малой модели, является значением по умолчанию;
   `whisper_large_v1_transcriber` - для загрузки большой модели;
   `all` - для загрузки обеих моделей;
   Если `-t` аргумент не указан, скрипт использует значение из POD, заданного аргументом `-n`;
   `[-n <node label>]` – лейбл ноды, необходимый для определения имени POD `das-processor-<node_label>`. По умолчанию имеет значение `central`. Указывайте этот параметр при установке на офисной ноде.

   > **Примечание:**
   >
   > Описание работы скрипта также можно найти, открыв файл скрипта с помощью текстового редактора.

   > **Примеры запуска скрипта**

Пример 1

Data Analysis Service установлен на центральной ноде кластера. Предполагается использование только технологии протокластеризации. Модель для нее скопирована в директорию `/root/das/tools/models_installation`, где располагается скрипт распаковки архивов.
Чтобы распаковать модель, выполните команду:

```
./unpack_and_deploy_models.sh
```

Пример 2

Data Analysis Service установлен на центральной ноде. Модели для протокластеризации и детектирования объектов на изображениях скопированы в директорию, отличную от той, где располагается скрипт распаковки архивов, например, в директорию `/root/models`.
Использование технологии транскрибации не предполагается.
Чтобы распаковать модели, выполните команду:

```
./unpack_and_deploy_models.sh -i /root/models
```

Пример 3

Data Analysis Service установлен на офисной ноде с лейблом "office1". Модели скопированы в директорию `/root/das/tools/models_installation`, где располагается скрипт распаковки архивов. Предполагается использование технологии транскрибации на большой модели.
Для распаковки выполните команду:

```
./unpack_and_deploy_models.sh -n office1 -t whisper_large_v1_transcriber
```

Для версии Data Analysis Service 1.4.1:

Выполните команду:

```
./unpack_and_deploy_models.sh [-n <namespace>] [-i <input_directory>] [-t
<transcriber_name>] [-l <node_label>]
```

со следующими аргументами:
`[-n <namespace>]` – пространство имен Kubernetes. Если аргумент не указан, значение берется из текущего Kubernetes context. Рекомендуется указывать аргумент, значение аргумента должно быть равным `infowatch`.
`[-i <input_directory>]` – директория, в которой находятся архивы моделей. Если аргумент не указан, то будут распакованы архивы из той директории, где находится скрипт;
`[-t <transcriber_name>]` – модель транскрибации. Возможные значения:
`whisper_small_transcriber` – для загрузки малой модели, является значением по умолчанию;
`whisper_large_v1_transcriber` – для загрузки большой модели;
`all` - для загрузки обеих моделей;
Если `-t` аргумент не указан, скрипт использует значение из POD, заданного аргументом `-l`;
`[-l <node_label>]` – лейбл ноды, необходимый для определения имени POD `das-`

`processor-<node_label>`. По умолчанию имеет значение `central`. Указывайте этот параметр при установке на офисной ноде.

> **Примечание:**
>
> Описание работы скрипта также можно найти, открыв файл скрипта с помощью текстового редактора.

> **Примеры запуска скрипта**
>
> ***Пример 1***
>
> Data Analysis Service установлен на центральной ноде кластера. Предполагается использование только технологии протокластеризации. Модель для нее скопирована в директорию `/root/das/tools/models_installation`, где располагается скрипт распаковки архивов.
> Чтобы распаковать модель, выполните команду:
>
> ```
> ./unpack_and_deploy_models.sh -n infowatch
> ```
>
> ***Пример 2***
>
> Data Analysis Service установлен на центральной ноде. Модели для протокластеризации и детектирования объектов на изображениях скопированы в директорию, отличную от той, где располагается скрипт распаковки архивов, например, в директорию `/root/models`. Использование технологии транскрибации не предполагается.
> Чтобы распаковать модели, выполните команду:
>
> ```
> ./unpack_and_deploy_models.sh -n infowatch -i /root/models
> ```
>
> ***Пример 3***
>
> Data Analysis Service установлен на офисной ноде с лейблом "office1". Модели скопированы в директорию `/root/das/tools/models_installation`, где располагается скрипт распаковки архивов. Предполагается использование технологии транскрибации на большой модели.
> Для распаковки выполните команду:
>
> ```
> ./unpack_and_deploy_models.sh -n infowatch -l office1 -t
> whisper_large_v1_transcriber
> ```

5. Если в директории назначения (по умолчанию: `/opt/das_data`) ранее уже были распакованы модели, то:
   - Чтобы заменить их на новые, подтвердите удаление старых файлов. Для этого введите **yes**:
     ```
     В /opt/das_data найдены финальные папки. Удалить их содержимое? (yes/no): yes
     ```
   - Чтобы загрузить модели, которые ранее не были установлены, а старые модели не удалять, введите **no**.

6. Дождитесь окончания процесса распаковки моделей.
   В результате будет выведено следующее сообщение:
   ```
   Все задачи успешно выполнены. Файлы перенесены в /opt/das_data.
   ```
7. Если вы установили большую модель для транскрибации, включите ее использование. Подробнее см. "Настройка модуля транскрибации".
8. Убедитесь, что модели установлены. Для этого проверьте логи с помощью команды:

   ```
   kubectl logs $(kubectl get pods -n infowatch -o name|grep das-processor) -n infowatch| grep "Waiting for model"
   ```

   При наличии в кластере нескольких экземпляров Data Analysis Service дополнительно укажите в команде лейбл ноды:

   ```
   kubectl logs $(kubectl get pods -n infowatch -o name|grep das-processor-<node_label>) -n infowatch| grep "Waiting for model"
   ```

   Если модели для включенных технологий установлены, то в результате не будет выведено ни одной строки. Если какая-либо технология включена, но модели для нее не будут найдены, то обработка данных этой технологией происходить не будет. При этом в логах появится сообщение об ошибке вида "`Waiting for model <model name>...`".

   Важно!

   Если в кластере установлено несколько экземпляров Data Analysis Service и в каком-либо из них включена транскрибация, но не установлены модели для нее, то данный экземпляр Сервиса будет забирать аудиофайлы из очереди, но не будет их обрабатывать. Таким образом, эти файлы будут недоступны для других экземпляров и не будут обработаны.

10 Обновление Data Analysis Service

Data Analysis Service поддерживает обновление с версий **1.2.x**, **1.3.x** и **1.4.0**. Если необходимо обновить Сервис с более ранних версий, то следует сначала обновиться на одну из указанных версий.

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 набор бинарных модулей образов контейнеров, необходимых для развертывания Сервиса без доступа к сети Интернет. Также архив содержит программу обновления.
Для обновления Сервиса вам также понадобятся ресурсы и модели для тех технологий, которые вы собираетесь использовать. Они входят в состав дистрибутива в виде следующих архивов:

- `proto_clusterization_resources.tar.gz` - для технологии протокластеризации;
- `transcription_resources.tar.gz`, `hf_faster_whisper_small.tar.gz` (для использования малой модели) и `hf_faster_whisper_large_v1.tar.gz` (для использования большой модели) - для технологии транскрибации;
- `image_resources.tar.gz` - для технологии детектирования объектов на изображениях.

Важно!

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

Примечание:

Все задачи, активные на момент начала обновления, 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 обновления.

10.1 Варианты обновления Data Analysis Service

Data Analysis Service 1.4.x поддерживает:

- прямое стандартное обновление с версий 1.3.1, 1.3.2 и 1.4.0;
- обновление с помощью сброса текущей установки и переустановки на старые данные.

Примечание:

Если вы обновляетесь с переходом на другой сервер или на новую ОС, используйте обновление со сбросом установки.

Если Data Analysis Service работает штатно, рекомендуется использовать стандартное обновление.

Чтобы **обновить распределенный кластер** Data Analysis Serviceи/или продуктов Платформы:

1. **Обновите центральную ноду** кластера и все установленные на ней продукты.
2. **Обновите офисные ноды** и все установленные на них продукты.

Важно!

Если вы обновляете Data Analysis Service в одном кластере с другими продуктами InfoWatch, то убедитесь, что версия Платформы совпадает у всех продуктов на всех нодах. В противном случае продукты, установленные на старой версии Платформы, необходимо обновить.

Чтобы узнать версию Платформы обновляемого Сервиса:

1. Распакуйте архив с дистрибутивом Data Analysis Service. Для этого перейдите в директорию с архивом и выполните команду:

   Пример для версии 1.4.0.38:

   ```
   tar -xvf iw_das_setup_1.4.0.38.tar.xz
   ```

2. Выполните команду:

   ```
   cat platform/setup.yaml | grep version
   ```

10.2 Прямое обновление Data Analysis Service

Важно!

Чтобы **обновить распределенный кластер** Data Analysis Service и/или продуктов Платформы:

1. **Обновите центральную ноду** кластера и все установленные на ней продукты.
2. **Обновите офисные ноды** и все установленные на них продукты.

Чтобы обновить Data Analysis Service:

1. Если вы используете технологию протокластеризации в Data Discovery, остановите все задачи сканирования в его веб-интерфейсе. Подробнее см. "*InfoWatch Data Discovery. Руководство пользователя*".
2. Выполните шаги по подготовке сервера к обновлению (см. "Подготовка сервера").

   Примечание:

В ходе обновления инсталлятор проверит наличие:

- не менее 100 ГБ на диске для директорий размещения данных ClickHouse, PostgreSQL, NATS и бинарных данных. Если директории для хранения данных находятся на разных дисках, то проверятся наличие 100 ГБ свободного места на каждом из них;
- не менее 20 ГБ на диске для директорий с образами;
- не менее 8 ГБ ОЗУ;
- не менее 2 ядер процессора;
- поддержки инструкции SSE 4.2;
- установленного пакета **`conntrack.`**

3. Создайте новую директорию на диске (например, `DAS`):

```
mkdir DAS
```

4. Скопируйте архив **`iw_das_setup_1.4.x.xx.tar.xz`** в созданную директорию.
   В нашем примере будет использована версия `1.4.0.38`, ей соответствует архив `iw_das_setup_1.4.0.38.tar.xz`.
5. Перейдите в директорию, в которую был скопирован архив.
6. Распакуйте архив с дистрибутивом Data Analysis Service в эту директорию:

```
tar -xvf iw_das_setup_1.4.0.38.tar.xz
```

7. Убедитесь, что в разделе с директорией `/var/lib/containerd` доступно не менее 150 ГБ дискового пространства.
   Чтобы проверить доступное дисковое пространство, используйте команду:

```
df -h /var/lib/containerd
```

Примечание:

Чтобы освободить дисковое пространство, вы можете удалить образы неиспользуемых контейнеров предыдущих версий. Для этого выполните следующие шаги:

a. Добавьте в файл `/etc/`**`crictl.yaml`** (при необходимости создайте его) строки:

```
runtime-endpoint: unix:///run/containerd/containerd.sock
image-endpoint: unix:///run/containerd/containerd.sock
timeout: 5
pull-image-on-create: false
disable-pull-on-run: false
```

b. Выполните команды:
   i. Для удаления контейнеров Платформы (где x.x.x.x – установленная версия Платформы):

```
crictl rmi $(crictl image | awk '($2 !~ /\<x.x.x.x\>|\<latest\>/ && $1 ~ /docker.infowatch.ru\/platform-core-release/ && $1 !~ /flannel/) {print $3}')
```

ii. Для удаления контейнеров Data Analysis Service (где x.x.x.x – установленная версия Data Analysis Service):

```
crictl rmi $(crictl image | awk '($2 !~ /\<x.x.x.x\>|\<latest\>/ && $1 ~ /docker.infowatch.ru\/clust/) {print $3}')
```

Данные команды удаляют контейнеры всех версий кроме указанной. Не удаляйте контейнеры установленной версии Data Analysis Service и Платформы!

8. Запустите программу обновления Сервиса:

```
./setup.py update
```

> **Важно!**
>
> Во время обновления продукты Платформы, уже установленные на сервере, могут быть недоступны.

9. При необходимости инсталлятор обновит Платформу. В этом случае дождитесь окончания этого процесса.
10. Укажите путь для размещения ресурсов Data Analysis Service (по умолчанию: `/opt/das_data`) и нажмите **Enter**.
11. Дождитесь окончания процесса обновления Data Analysis Service.
12. Ознакомьтесь с отчетом об обновлении Сервиса:

```
###Result###
Install product: Infowatch DAS(das)
Install manifest section: central
Install node label: central
installed 0 components
updated   4 components
add ref   0 components
```

13. Чтобы вывести на экран информацию об установленных продуктах, выполните команду:

```
./setup.py showproducts
```

14. Чтобы проверить, что обновление прошло успешно, выполните команду:

```
kubectl get pods -n infowatch
```

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

15. При обновлении с версии 1.3.x установите модели для технологий, которые будете использовать. Подробнее см. статью "Установка моделей".
    При обновлении с версии 1.4.0 обновлять уже установленные модели не требуется. Но вы можете установить модели для технологий, которые ранее не использовали.

    > **Важно!**
    >
    > Если модели для какой-либо технологии не будут установлены, то эта технология не будет работать.

16. Если до обновления было настроено взаимодействие Data Analysis Service с какими-либо нодами кластера кроме центральной, повторно задайте ноды для взаимодействия. Подробнее см. "Задание нод обработки".
    В противном случае Сервис будет обрабатывать данные, приходящие только с центральной ноды.
17. Если на сервере установлены продукты Платформы, обновите их до версий, совместимых с Data Analysis Service (см. "Типы установки и обновления, совместимость с другими продуктами, аппаратные и программные требования").
18. Если Data Analysis Service установлен в распределенном кластере, обновите другие экземпляры Сервиса и продукты Платформы на других серверах кластера до версий, совместимых с Data Analysis Service.
19. Если после обновления вам необходимо освободить место на диске, вы можете удалить неиспользуемые контейнеры аналогично п.6.
20. **Если вы используете технологию протокластеризации в Data Discovery:**

    После обновления Data Analysis Service с версии 1.3.x до версии 1.4.x функциональность тегов будет работать только для новых документов. Чтобы теги ранее обработанных документов стали доступны в Data Discovery:

    a. Убедитесь, что вы обновили Data Discovery до версии, совместимой с Data Analysis Service.
    b. На сервере Data Analysis Service перейдите в директорию `../DAS/tools/migration`. В ней располагается скрипт обновления тегов `fill_proto_tags.sh`
    c. Если Сервис установлен на офисной ноде кластера, скопируйте скрипт на центральную ноду.
    d. На центральной ноде запустите скрипт с помощью команды:

    При обновлении до версии 1.4.0:

    ```
    ./fill_proto_tags.sh -n <das_node_label>
    ```

    При обновлении до версии 1.4.1:

    ```
    ./fill_proto_tags.sh -n infowatch -l <das_node_label>
    ```

    где `<das_node_label>` – лейбл ноды, на которую установлен Data Analysis Service. Для центральной ноды аргумент можно не указывать, либо используйте лейбл `central`. Для офисной ноды лейбл задается при установке, по умолчанию: `office`.

> **Важно!**
>
> Выполнение скрипта может занимать несколько часов, вплоть до суток. В течение этого времени возможно серьёзное увеличение нагрузки на модуль протокластеризации.

После обновления уровень логирования Сервиса будет сброшен и по умолчанию иметь значение INFO. Подробнее о логировании см. "Диагностика Data Analysis Service".

10.3 Обновление Data Analysis Service с помощью сброса установки (reset)

> **Важно!**
>
> **Чтобы обновить весь распределенный кластер** Data Analysis Service и/или продуктов Центра расследований с помощью сброса (reset) установки, действуйте в следующей последовательности:
>
> 1. Выполните **сброс на офисных** нодах с продуктами в режиме `office`.
> 2. Выполните **сброс на офисных** нодах с продуктами в режиме `central`.
> 3. Выполните **сброс центральной** ноды.
> 4. **Установите центральную** ноду кластера.
> 5. **Установите офисные** ноды с продуктами в режиме `central`.
> 6. **Установите офисные** ноды с продуктами в режиме `office`.
>
> В рамках обновления кластера выполняйте сброс установки только один раз для каждой ноды.

Обновление состоит из следующих шагов:

1. Сброс установки **с помощью нового инсталлятора**. При сбросе данные Сервиса не будут удалены.
2. Установка новых версий Платформы и Сервиса с указанием уже имеющихся папок с данными.

> **Примечание:**
>
> Если вы используете удаленное подключение по протоколу SSH, чтобы избежать проблем в случае разрыва соединения с сервером, рекомендуем использовать утилиту Screen. При отключении от утилиты запущенные в ней процессы не прервутся, что позволит безопасно продолжить обновление.

Чтобы обновить Data Analysis Service:

1. Если вы используете технологию протокластеризации в Data Discovery, остановите все задачи сканирования в его веб-интерфейсе. Подробнее см. "*InfoWatch Data Discovery. Руководство пользователя*".
2. Выполните шаги по подготовке сервера к обновлению (см. "Подготовка сервера").

Примечание:

В ходе обновления инсталлятор проверит наличие:

- не менее 100 ГБ на диске для директорий размещения данных ClickHouse, PostgreSQL, NATS и бинарных данных. Если директории для хранения данных находятся на разных дисках, то проверяется наличие 100 ГБ свободного места на каждом из них;
- не менее 20 ГБ на диске для директорий с образами;
- не менее 8 ГБ ОЗУ;
- не менее 2 ядер процессора;
- поддержки инструкции SSE 4.2;
- установленного пакета **`conntrack.`**

3. Создайте новую директорию на диске (например, `DAS`):

```
mkdir DAS
```

4. Скопируйте архив **`iw_das_setup_1.4.x.xx.tar.xz`** в созданную директорию.
   В нашем примере будет использована версия `1.4.0.38`, ей соответствует архив `iw_das_setup_1.4.0.38.tar.xz`.
5. Перейдите в директорию, в которую был скопирован архив.
6. Распакуйте архив с дистрибутивом в эту директорию:

```
tar -xvf iw_das_setup_1.4.0.38.tar.xz
```

7. После сброса при установке новой версии присвойте нодам кластера те же лейблы, которые были до обновления.
   Чтобы посмотреть список лейблов в кластере, выполните команду:

```
kubectl get node -n infowatch -o jsonpath="{range .items[*]}
{.metadata.annotations.flannel\.alpha\.coreos\.com/public\-ip}
{.metadata.labels.kubernetes\.io/hostname} {.metadata.labels.iwnode}{'\n'}{end}"
```

8. До сброса установки запомните или сохраните пути к текущим директориям хранения данных. Получить список директорий можно, выполнив команду:

```
cat /var/lib/iwplatform/data/volumes.txt
```

9. Выполните сброс текущей установки **с помощью инсталлятора новой версии**:

```
./setup.py reset
```

Важно!

Если на сервере установлены продукты на базе Платформы, установки этих продуктов будут сброшены. При сбросе установки на центральной ноде продукты на офисных нодах перестанут функционировать.

После установки новой версии Data Analysis Service повторно установите продукты Платформы, указав директории с имеющимися данными.

Если на этом сервере вы **уже обновили продукт** Платформы до совместимой версии, **не выполняйте повторный сброс** при обновлении Data Analysis Service, так как требуемая версия Платформы уже установлена.

10. Дождитесь окончания процесса сброса установки Платформы и Сервиса.
11. Перезагрузите сервер с помощью команды:

```
reboot
```

12. Если вы обновляетесь с переходом на другой сервер или на новую ОС, скопируйте на другой раздел или на новый сервер директорию с данными PostgreSQL, которая была указана при установке.

Примечание:

В **Data Analysis Service 1.3.0 и более ранних** директория по умолчанию расположена по пути `/mnt`.
В **Data Analysis Service 1.3.1 и более новых** - по умолчанию по пути `/opt`.
Обязательно копируйте директории с сохранением прав и владельцев. Для копирования используйте стандартные средства, например утилиту `rsync`.

13. Чтобы освободить дисковое пространство, вы можете удалить директорию с образами контейнеров прошлых версий `/var/lib/containerd`. Для этого выполните команду:

```
rm -rf /var/lib/containerd
```

14. Установите новую версию Data Analysis Service (см. "Установка Data Analysis Service").

Важно!

Когда инсталлятор потребует указать директории для размещения данных, указывайте директории с данными прошлой версии и выбирайте «**n**» при ответе на вопрос «Очищать ли папки перед установкой?».

В версии 1.3.1 пути установки по умолчанию изменились. Если ранее вы использовали пути `/mnt`, при установке Data Analysis Service 1.4.0 и более новых по умолчанию инсталлятор не обнаружит данные прошлой версии и создаст директории по пути `/opt`. Поэтому переместите данные в `/opt` или укажите старые директории при установке новой версии.

15. Если до сброса на сервере были установлены и другие продукты, установите версии продуктов, совместимые с Data Analysis Service (см. "Типы установки и обновления, совместимость с другими продуктами, аппаратные и программные требования").

16. Если Data Analysis Service установлен в распределенном кластере, обновите другие экземпляры Сервиса и продукты Платформы на других серверах кластера до версий, совместимых с Data Analysis Service.
17. Если после обновления вам необходимо освободить место на диске, вы можете удалить неиспользуемые контейнеры. Подробнее см. п.7 статьи "Прямое обновление Data Analysis Service".
18. **Если вы используете технологию протокластеризации в Data Discovery:**

    После обновления Data Analysis Service с версии 1.3.x до версии 1.4.x функциональность тегов будет работать только для новых документов. Чтобы теги ранее обработанных документов стали доступны в Data Discovery:

    a. Убедитесь, что вы обновили Data Discovery до версии, совместимой с Data Analysis Service.
    b. На сервере Data Analysis Service перейдите в директорию `../DAS/tools/migration`. В ней располагается скрипт обновления тегов `fill_proto_tags.sh`
    c. Если Сервис установлен на офисной ноде кластера, скопируйте скрипт на центральную ноду.
    d. На центральной ноде запустите скрипт с помощью команды:

    При обновлении до версии 1.4.0:

    ```
    ./fill_proto_tags.sh -n <das_node_label>
    ```

    При обновлении до версии 1.4.1:

    ```
    ./fill_proto_tags.sh -n infowatch -l <das_node_label>
    ```

    где `<das_node_label>` – лейбл ноды, на которую установлен Data Analysis Service. Для центральной ноды аргумент можно не указывать, либо используйте лейбл `central`. Для офисной ноды лейбл задается при установке, по умолчанию: `office`.

    Важно!

    Выполнение скрипта может занимать несколько часов, вплоть до суток. В течение этого времени возможно серьёзное увеличение нагрузки на модуль протокластеризации.

После обновления уровень логирования Сервиса будет сброшен и по умолчанию иметь значение INFO. Подробнее о логировании см. "Диагностика Data Analysis Service".

11 Настройка Data Analysis Service

В этом разделе:

- Настройка взаимодействия с продуктами Платформы
- Включение/выключение модулей Data Analysis Service
- Настройка производительности
- Настройка использования ресурсов сервера
- Настройка модуля транскрибации
- Добавление языков для протокластеризации
- Диагностика Data Analysis Service

11.1 Настройка взаимодействия с продуктами Платформы

В этом разделе:

- Включение технологий Data Analysis Service в продуктах Платформы
- Задание нод обработки

11.1.1 Включение технологий Data Analysis Service в продуктах Платформы

Включение транскрибации и детектирования объектов в Activity Monitor

Data Analysis Service позволяет:

- транскрибировать аудиозаписи;
- детектировать объекты на изображениях, в том числе непристойный контент.

Чтобы использовать эти функциональные возможности в Activity Monitor:

1. Убедитесь, что Data Analysis Service настроен на обработку центральной ноды кластера. Лейбл этой ноды указывается при установке Сервиса. Также его можно задать после установки. Подробнее см. "Задание нод обработки".
2. Убедитесь, что модуль транскрибации и/или модуль детектирования объектов на изображениях включены. Подробнее см. "Включение/выключение модулей Data Analysis Service".
3. Включите транскрибацию данных в Activity Monitor, выполнив на центральной ноде кластера следующие команды:

   ```
   kubectl exec -it $(kubectl get pods -n infowatch | grep postgresql-<AM_NODE_LABEL> | awk '//{print $1;}') -n infowatch -- psql -U postgres -d activitymonitor -c "update features set status='enabled' where name='stream_transcripts'"
   kubectl rollout restart deployment datacollector-<AM_NODE_LABEL> -n infowatch
   ```

   где `<AM_NODE_LABEL>` – это лейбл ноды, на которой установлен Activity Monitor.

   Примечание:

Если Activity Monitor и Data Analysis Service установлены не в кластере, а на одном отдельном сервере, в качестве лейбла ноды укажите значение `central`.

4. Включите детектирование объектов, выполнив на центральной ноде кластера следующие команды:

```
kubectl exec -it $(kubectl get pods -n infowatch | grep postgresql-<AM_NODE_LABEL> | awk
'//{print $1;}') -n infowatch -- psql -U postgres -d activitymonitor -c "update features
set status='enabled' where name='stream_enrich'"
kubectl rollout restart deployment datacollector-<AM_NODE_LABEL> -n infowatch
```

где `<AM_NODE_LABEL>` – это лейбл ноды, на которой установлен Activity Monitor.

Примечание:

Если Activity Monitor и Data Analysis Service установлены не в кластере, а на одном отдельном сервере, в качестве лейбла ноды укажите значение `central`.

Включение протокластеризации в Data Discovery

Data Analysis Service позволяет осуществлять протокластеризацию текстовых документов для поиска похожих.

Чтобы использовать эту функциональную возможность в Data Discovery, достаточно установить Data Analysis Service в том же кластере и настроить Сервис на обработку ноды, на которой установлен Data Discovery. Лейбл этой ноды указывается при установке Сервиса. Также его можно задать после установки. Подробнее см. "Задание нод обработки".

После этого в Data Discovery вы сможете включить опцию "**Определять похожие файлы**" при создании или редактировании задачи сканирования. Подробнее о задачах сканирования см. документ "*InfoWatch Data Discovery. Руководство пользователя. Хранение файлов. Вкладка "Задачи"*".

11.1.2 Задание нод обработки

При установке Data Analysis Service в кластере с продуктами Платформы, для Сервиса необходимо задать одну или несколько нод, очередь сообщений которых он будет обрабатывать.

Чтобы изменить данный параметр после установки, задайте необходимое значение переменной окружения `SUBSCRIBE_TO`. Для этого на сервере Data Analysis Service выполните следующую команду:

```
kubectl set env deploy -n infowatch das-processor-<das_node_label>
SUBSCRIBE_TO="<message_queue_node_labels>"
```

где `<das_node_label>` – здесь и далее это лейбл ноды, на которую установлен Data Analysis Service. Для центральной ноды используйте лейбл `central`. Для офисной ноды лейбл задается при установке, по умолчанию: `office`.
`<message_queue_node_labels>` – список лейблов нод, на которые продукты Платформы отправляют сообщения для Data Analysis Service. Указывайте лейблы через запятые.

Примечание:

InfoWatch Activity Monitor отправляет сообщения в очередь NATS на центральную ноду кластера.

InfoWatch Data Discovery отправляет сообщения в очередь NATS на ту ноду кластера, на которой он установлен.

Чтобы узнать лейбл ноды, выполните команду:

```
kubectl get node -n infowatch -o jsonpath="{range .items[*]}
{.metadata.annotations.flannel\.alpha\.coreos\.com/public\-ip} {.metadata.labels.kubernetes\.io/hostname}
{.metadata.labels.iwnode}{'\n'}{end}"
```

В результате будет выведен список нод кластера с указанием их лейблов.

Чтобы узнать текущее значение переменной окружения `SUBSCRIBE_TO`, выполните команду:

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

11.2 Включение/выключение модулей Data Analysis Service

Модуль протокластеризации включен по умолчанию. Модуль транскрибации и модуль детектирования объектов на изображениях можно включить во время установки Сервиса.

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

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

где `<node_label>` – здесь и далее это лейбл ноды, на которую установлен Data Analysis Service. Для центральной ноды используйте лейбл `central`. Для офисной ноды лейбл задается при установке, по умолчанию: `office`;
`<name>` – это имя переменной. Модулям Data Analysis Service соответствуют следующие переменные окружения:

Имя переменной окружения	Компонент Data Analysis Service
`PROTO_CLUSTERIZATION`	Модуль протокластеризации
`TRANSCRIPTION`	Модуль транскрибации
`IMG_CONTENT_DETECTION`	Модуль детектирования объектов на изображениях

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

Пример:

Для включения модуля транскрибации Сервиса, установленного на центральной ноде кластера, выполните команду:

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

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

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

Примечание:

Рекомендуется выключать неиспользуемые функции в целях экономии ресурсов.

11.3 Настройка производительности

По умолчанию на сервере функционирует один POD Data Analysis Service, включающий в себя все модули Сервиса. Рекомендуемые аппаратные требования для модулей одного POD вы можете найти в статье "Типы установки, совместимость с другими продуктами, аппаратные и программные требования".

Вы можете оценить производительность Data Analysis Service следующими способами:

- Оценка производительности модуля протокластеризации
- Оценка производительности POD

По итогам оценки вы можете увеличить производительность Data Analysis Service следующими способами:

- Создание реплик POD
- Многонодовая установка

11.3.1 Оценка производительности модуля протокластеризации

Если выделенных для модуля протокластеризации ресурсов недостаточно для обработки текстовых данных, то в Консоли Управления Traffic Monitor кнопка **Найти аналоги** может не отображаться для новых событий. В таком случае оцените производительность модуля протокластеризации, проанализировав количество необработанных объектов в базе данных. Для этого на центральной ноде кластера в течение 2-3 часов каждые 30-60 минут выполняйте команду:

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

где <`das_central_node>` – это лейбл ноды, где установлен центральный экземпляр Data Analysis Service, т.е. при установке для параметра `--productmode` было задано значение `central`.

При росте числа необработанных объектов и нехватке производительности модуля протокластеризации рекомендовано ее увеличение.

11.3.2 Оценка производительности POD

Вы можете оценить производительность отдельного POD, используя информацию в лог-файлах.

Для этого:

1. Перейдите в директорию **`/var/log/pods/infowatch_das-processor-<node_label>-<POD_number>/das-processor-<node_label>`**,
   где `<node_label>` – здесь и далее это лейбл ноды, на которую установлен Data Analysis Service. Для центральной ноды используйте лейбл `central`. Для офисной ноды лейбл задается при установке, по умолчанию: `office`;
   `<POD_number>` – это уникальный номер POD.
   При обработке Сервисом каждого нового контента, в лог записывается сообщение.
   Пример сообщения в логе:

   ```
   2025-03-20T12:28:25.471557388Z stdout F 2025-03-20 15:28:25,468 (7)
   [INFO] (das_processor.handlers.transcription) Msg:
   status=Status(code=200, message='Ok') ds_node='datastorage-central'
    d
   s_object_containers=[DsContainerInfo(status=Status(code=200,
   message='Ok'), container_id='872feea0-47dc-43b8-b23d-651dd5aba506',
   ds_objects=[DsObjectId(bucket_id='epevents_2025_79_audio',
   object_id='8d968380-1ccf-41b5-82a3-6d5f44083c6e')],
   user_data={'DateTime': 1742473575, 'EndDateTime': 1742473575,
   'Count': 1, 'StreamID': '3bd67804-eff7-a7a3-2781-4e7ddde251bf',
   'TrackID':
    '00f3964e-f2c3-4726-ba1f-93e29b17ebc9', 'SourceType': 'speakers'},
   reply_bucket_id=None)] diarization=False gender_classification=False
   age_classification=False overlap_detection=False wh
   isper_detection=False scream_detection=False emotion_detection=False
   hotwords=None, Transcription service queue length: regular = 1
   ```

   В конце данного сообщения указывается количество контентов в очереди технологии, соответствующей типу контента:
   `<Название_очереди_технологии> service queue length: regular = <количество_контентов>`
   Для каждой технологии существует отдельная очередь:
   - `Transcription service queue` - для транскрибации;
   - `Universal Detector service queue` - для детектирования объектов на изображениях;
   - `Clusterization service queue` - для протокластеризации
     Очередь протокластеризации в свою очередь состоит из двух: обычной "`regular`" и приоритетной "`priority`". В логе будет указано количество контентов в каждой из них.
2. Выполните команду в зависимости от интересующей вас технологии:

   ```
   cat <имя_последнего_лог-файла> | grep "<Название_очереди_технологии> length:"|rev| cut -c 1-2
   5 -|rev
   ```

где <имя_последнего_лог_файла> - это имя лог-файла с наибольшим номером. Например, для оценки производительности транскрибации выполните команду:

```
cat 0.log | grep "Transcription service queue length:"|rev| cut -c 1-25 -|rev
```

3. Проанализируйте изменение количества контентов в очереди.
   Если это количество растет, то рекомендуется увеличить производительность Сервиса.

11.3.3 Создание реплик POD

Чтобы увеличить производительность Сервиса, вы можете создать дополнительные реплики POD. В этом случае будет увеличена производительность всех включенных модулей.

Для каждого нового POD необходимо выделить ресурсы на сервере.

Пример

Если у вас работает один POD Data Analysis Service, использующий 6 CPU и 16 GB оперативной памяти, то при создании двух реплик, вам необходимо иметь на сервере дополнительно 12 CPU и 32 GB оперативной памяти. Таким образом, всего для трех POD Сервиса потребуется 18 CPU и 48 GB оперативной памяти.

Создать реплики POD можно с помощью команды:

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

При необходимости вы можете ограничить использование ресурсов каждым POD Сервиса. Подробнее см. "Настройка использования ресурсов сервера".

11.3.4 Многонодовая установка

Для увеличения производительности Data Analysis Service вы можете установить Сервис на одной или нескольких дополнительных нодах (подробнее см. "Установка Data Analysis Service"). Приходящие в Сервис контенты будут распределяться между всеми его экземплярами в кластере и обрабатываться параллельно.

Для параллельной обработки контентов от InfoWatch Traffic Monitor, настройте взаимодействие InfoWatch Traffic Monitor с любым экземпляром Сервиса в кластере (подробнее см. документ "*InfoWatch Traffic Monitor. Руководство администратора. Настройка интеграции с Data Analysis Service*")

Для параллельной обработки контентов от продукта Платформы, настройте экземпляры Сервиса на ноду, куда этот продукт отправляет сообщения. Подробнее см. "Задание нод обработки".

11.4 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:

   ```
   ./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.

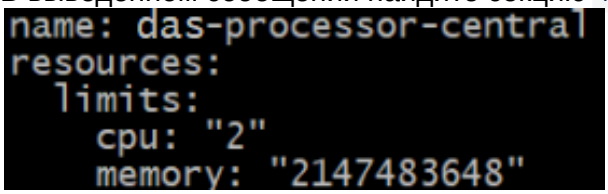
   > **Пример команды запуска скрипта:**
   >
   > Чтобы Data Analysis Service использовал не более 4 ядер центрального процессора и не более 10 гигабайтов оперативной памяти, выполните команду:
   >
   > ```
   > ./resources_setter.sh -c 4 -m 10G
   > ```

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

1. Выполните команду:

   ```
   kubectl get deploy -o yaml -n infowatch das-processor-<node_label>
   ```

   где `<node_label>` – это лейбл ноды, на которую установлен Data Analysis Service. Для центральной ноды используйте лейбл `central`. Для офисной ноды лейбл задается при установке, по умолчанию: `office`.
2. В выведенном сообщении найдите секцию `resources`:

   Объем оперативной памяти в сообщении указан в байтах.

11.5 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.

Важно!

Для использования моделей, необходимо предварительно их установить. Подробнее см. "Установка Data Analysis Service в интерактивном режиме".

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

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

Чтобы убедиться, что Сервис использует GPU, вы можете выполнить проверку.

11.5.1 Аппаратные и программные требования к графическому процессору

Для использования GPU вам необходима видеокарта NVIDIA с поддержкой Compute Unified Device Architecture (CUDA) версии 12.4.

На сервере должны быть установлены:

- Драйвер NVIDIA с версией не ниже 550.xx

  Примечание:

  Рекомендуется устанавливать наиболее свежий драйвер.
  Если вы установили драйвер недостаточной версии, то в логах процесса **das-processor** появится сообщение вида:
  `[ERROR] (das_processor.tech.transcriber) CUDA failed with error CUDA driver version is insufficient for CUDA runtime version`

- 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)

  Примечание:

  В случае установки Data Analysis Service на офисной ноде плагин NVIDIA необходимо устанавливать на центральной ноде.

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

11.5.2 Включение использования GPU

Чтобы Data Analysis Service использовал 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
     ```

     Если в результате команды будет выведено сообщение "`No GPU support`", то:
     a. Выполните команды:

        ```
        sudo systemctl restart containerd
        kubectl rollout restart ds -n kube-system nvidia-device-plugin-daemonset
        ```

     b. Повторите проверку.
        Если снова выведено сообщение "`No GPU support`", то:
        i. Выполните команды:

           ```
           sudo nvidia-ctk runtime configure --runtime=containerd && sed -i 's/\
           (default_runtime_name[^A-Za-z]\+\)\([A-Za-z]\+\)\(.*\)/\1nvidia\3/' /
           etc/containerd/config.toml
           sudo systemctl restart containerd
           ```

        ii. Повторите проверку.
2. Добавьте GPU в ресурсы процесса **das-processor**. Для этого выполните команду:

   ```
   sudo kubectl -n infowatch set resources deploy das-processor-<node_label> --
   limits=nvidia.com/gpu=1
   ```

где `<node_label>` – здесь и далее это лейбл ноды, на которую установлен Data Analysis Service. Для центральной ноды используйте лейбл `central`. Для офисной ноды лейбл задается при установке, по умолчанию: `office`.

3. Перезагрузите сервер. Загрузка кластера может занять некоторое время.
4. Убедитесь, что все POD запущены, выполнив команду:

```
sudo kubectl get pods -n infowatch
```

   Каждый POD должен иметь статус Running.
5. Укажите, что при вычислениях процесс **das-processor** должен использовать GPU. Для этого выполните команду:

```
kubectl -n infowatch set env deploy das-processor-<node_label> USE_GPU=TRUE
```

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-<node_label> USE_GPU=FALSE
```

11.5.3 Выбор модели транскрибации

Чтобы выбрать модель обработки, выполните команду:

- для использования малой модели:

```
sudo kubectl set env deployment -n infowatch das-processor-<node_label>
TRANSCRIBER=whisper_small_transcriber
```

- для использования большой модели:

```
sudo kubectl set env deployment -n infowatch das-processor-<node_label>
TRANSCRIBER=whisper_large_v1_transcriber
```

Примечание:

Большая модель может работать только на GPU.

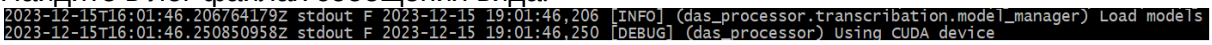
11.5.4 Проверка использования GPU

Чтобы убедиться, что Data Analysis Service использует GPU:

1. Если вы будете использовать большую модель транскрибации, выберите ее.
2. Установите для Сервиса режим логирования DEBUG:

   ```
   sudo kubectl -n infowatch set env deploy das-processor-<node_label> SRV_LOG_LEVEL=DEBUG
   ```

3. Обеспечьте передачу в Data Analysis Service аудиоданных для обработки.
4. Перейдите в директорию `/var/log/pods/infowatch_das-processor-<node_label>-…/das-processor-<node_label>` , где находятся лог-файлы.
5. 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 аудиоданных и уничтожаются, если неактивны более часа. При уничтожении в лог-файл записывается сообщение `Drop models` .

Также вы можете найти сообщения в логах с помощью команды:

```
kubectl logs $(kubectl get pods -n infowatch -o name|grep das-processor-<node_label>) -n infowatch| grep "<текст_сообщения>"
```

11.6 Добавление языков для протокластеризации

Data Analysis Service осуществляет протокластеризацию текстовых документов с учетом морфологии на тех языках, для которых добавлены словари. После установки Сервис работает с русским и английским языками. Чтобы обрабатывать с учетом морфологии документы на других языках, необходимо добавить в Data Analysis Service словари этих языков. Вы можете получить их у специалистов компании InfoWatch.

> **Примечание:**
>
> Теги для документов также создаются только для тех языков, для которых добавлены словари.

Словарь включает в себя:

1. *Морфологический словарь языка в формате JSON* – содержит различные словоформы и используется для определения того, что встречаемые в документах формы принадлежат определенному слову.
   Пример названия файла: `az.json`
2. *Файл с нормальными формами слов в формате JSON* – содержит проиндексированный список нормальных форм слов для данного языка.
   Пример названия файла: `az_tokenizer_vocabulary.json`
3. *Файл со стоп-словами в формате TXT* – содержит слова, исключаемые из анализа текста, так как те не несут смысловой нагрузки, например, предлоги.
   Пример названия файла: `stopwords_az.txt`

Чтобы добавить в Data Analysis Service словарь языка для протокластеризации:

1. Скопируйте словарь в директорию `<das_data_dir>/nlp/tokenization/dicts/<vocabulary_dir>`
   где `<das_data_dir>` – это директория для размещения ресурсов Data Analysis Service, которая задается при установке. По умолчанию: `/opt/das_data`.
   `<vocabulary_dir>` – конечная директория для файлов словаря. В качестве имени она должна иметь двухбуквенный код добавляемого языка в формате ISO 639-1.
   Например, для азербайджанского языка в Data Analysis Service 1.4, это будет директория `/opt/das_data/nlp/tokenization/dicts/az`.
   Если директория не существует, создайте ее.
2. Перезапустите процесс **das-processor** с помощью команды:

   ```
   kubectl rollout restart deploy -n infowatch das-processor-<das_node_label>
   ```

   где `<das_node_label>` – это лейбл ноды, на которую установлен Data Analysis Service. Для центральной ноды используйте лейбл `central`. Для офисной ноды лейбл задается при установке, по умолчанию: `office`.

> **Важно!**
>
> После добавления словаря не изменяйте входящие в него файлы. Если вы измените какой-либо файл, Сервис не сможет обрабатывать контенты на данном языке.

11.7 Диагностика Data Analysis Service

11.7.1 Логирование

Лог-файлы Сервиса вы можете найти в директориях:

- `/var/log/pods/infowatch_das-processor-<node_label>-…/`
- `/var/log/pods/infowatch_das-adapter-<node_label>-…/`
- `/var/log/pods/infowatch_das-proto-cl-prov-<node_label>-…/`

где `<node_label>` – здесь и далее это лейбл ноды, на которую установлен Data Analysis Service. Для центральной ноды лейбл имеет значение `central`. Для офисной ноды лейбл задается при установке, по умолчанию: `office`.

Чтобы вывести логи, выполните команду:

```
kubectl logs $(kubectl get pods -n infowatch -o name|grep das-processor) -n infowatch
```

При наличии в кластере нескольких экземпляров Data Analysis Service дополнительно укажите в команде лейбл ноды:

```
kubectl logs $(kubectl get pods -n infowatch -o name|grep das-processor-<node_label>) -n infowatch
```

Чтобы найти в логах конкретное сообщение, добавьте команду `grep` с его текстом:

```
kubectl logs $(kubectl get pods -n infowatch -o name|grep das-processor) -n infowatch| grep
"<текст_сообщения>"
kubectl logs $(kubectl get pods -n infowatch -o name|grep das-processor-<node_label>) -n infowatch| grep
"<текст_сообщения>"
```

Чтобы изменить уровень логирования, выполните команду:

```
sudo kubectl -n infowatch set env deploy das-processor-<node_label> das-adapter-<node_label> das-proto-cl-
prov-<node_label> SRV_LOG_LEVEL=<value>
```

где `<value>` – это уровень логирования, для которого вы можете установить следующие значения:

- INFO – краткая информация о работе сервиса и возникающих ошибках. Установлен по умолчанию;
- DEBUG - подробная информация о работе сервиса и возникающих ошибках.

Например, чтобы установить уровень логирования DEBUG для Сервиса на центральной ноде, выполните команду:

```
sudo kubectl -n infowatch set env deploy das-processor-central das-adapter-central das-proto-cl-prov-
central SRV_LOG_LEVEL=DEBUG
```

11.7.2 Скрипт диагностики

Начиная с версии Data Analysis Service 1.4.1 вы можете получить информацию о состоянии Сервиса, запустив скрипт диагностики `das-info.sh`. Для этого:

1. Перейдите в директорию `<DAS>/tools/diagnostics`, где `<DAS>` - это директория, в которую был распакован архив с дистрибутивом при установке.
2. Запустите скрипт с помощью команды:

   ```
   ./das-info.sh -n infowatch
   ```

В результате будет выведена следующая информация по всем экземплярам Сервиса в кластере:

- версия Data Analysis Service и Платформы;
- список и статус подов;
- список ошибок для каждого пода;
- статус Enabled/Disabled для каждой технологии;
- значения переменных окружения Data Analysis Service.

Также будет выведен список моделей, установленных на текущей ноде.

12 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 директорию, в которую был распакован дистрибутив последней установленной версии Сервиса.
2. Выполните команду:

   ```
   ./setup.py remove
   ```

3. Дождитесь завершения процесса. Инсталлятор удалит компоненты Сервиса. Данные Data Analysis Service удалены не будут. При необходимости вы можете удалить данные вручную из директорий, указанных при установке.

Примечание:

В случае распределенной установки, если Data Analysis Service установлен на офисной ноде и на ней отсутствуют другие продукты, то после удаления Data Analysis Service также рекомендуется удалить компоненты платформы. Для этого после выполнения команды `./setup.py remove`, также выполните команду:

```
./setup.py reset
```

Если Data Analysis Service установлен на центральной ноде или если на офисной ноде вместе с ним также установлены другие продукты, выполнять команду `./setup.py reset` категорически не рекомендуется.

Чтобы узнать, какие продукты установлены в кластере, вы можете перейти в директорию с инсталлятором и выполнить команду:

```
./setup.py showproducts
```

Примечание:

Если вы удалили Data Analysis Service, установленный в режиме `central`, компоненты Data Analysis Service, установленные в режиме `office` на других нодах, не будут удалены автоматически. Удалите их вручную.