Содержание

Сборка IPFIX потоков rcollector2

Введение

Утилита предназначена для дополнения данными вспомогательных потоков таких как clickstream, SIP из потока сессий (netflow). По умолчанию полученные данные сохраняются в файлы, но данные так же можно сохранять и в БД ИС СОРМ-3.

Установка и обновление

1. подключите репозиторий VAS Experts

```
rpm --import http://vasexperts.ru/centos/RPM-GPG-KEY-vasexperts.ru
rpm -Uvh
http://vasexperts.ru/centos/6/x86_64/vasexperts-repo-1-0.noarch.rpm
```

2. установите rcollector2:

```
yum install -y rcollector2
```

3. настройте файлы конфигурации в директории /etc/rcollector2 см. далее

 Внимание! При переходе с предыдущей версии rcollector на rcollector2 необходимо внести изменения в скрипты запуска.

Файлы поставки

1. примеры конфигурации:

```
/etc/rcollector2/rc2flowprocess - пример исполняемого файла для обработки
сессий (netflow)
/etc/rcollector2/rc2urlprocess - пример исполняемого файла для обработки
HTTP сессий
/etc/rcollector2/rc2sipprocess - пример исполняемого файла для обработки
SIP сессий
/etc/rcollector2/rcollector_flow.properties - пример конфигурационного
файла для режима работы flow
/etc/rcollector2/rcollector_url.properties - пример конфигурационного
файла для режима работы url
/etc/rcollector2/rcollector_sip.properties - пример конфигурационного
файла для режима работы sip
```

2. исполняемый файл:

```
/bin/rcollector2
```

Параметры запуска программы

Программа rcollector2 имеет следующие параметры запуска:

```
usage: rcollector2 ОПЦИИ
```

ОПЦИИ:

- -h, --help отобразить краткую справку.
- -fCONFIG, --config-file=CONFIG конфигурационный файл.
- -mMODE, --mode=MODE режим работы программы. Режимы: flow, urlget, sipget
- -uidUNIQUEID, --uniqueid=UNIQUEID уникальный номер точки обработки.
- -ifINFILE, --infile=INFILE входной файл.
- -ofOUTFILE, --outfile=OUTFILE выходной файл.
- -asnASN, --localasn=ASN список локальных автономных систем (разделенных запятой)
- -oufOUTFILTER, --outfilter=OUTFILTER вывод отдельных записей по dpi protocol id в файл. Формат: <outfile>,<protocol>,...<protocol>|[next filter]. Пример: telnet.dump,22,23
- -tdUSEFILTER, --tordb=USEFILTER путь к файлу с IP адресами сети TOR. По умолчанию /var/data/tor/ip.db.gz.
- -sdbDIR, --sessiondb=DIR путь к каталогу с данными о сессиях. По умолчанию /var/db/rcollector.
- -outmailFILE, --outfilemail=FILE выходной файл для данных о mail соединениях в режиме sipget.
- -outftpFILE, --outfileftp=FILE выходной файл для данных о ftp соединениях в режиме sipget.
- -outimFILE, --outfileim=FILE выходной файл для данных о im соединениях в режиме sipget.
- -dhINTEGER, --depth=INTEGER глубина поиска фалов с данными о сессиях.
- -sdrt, --session-db-read-thread использовать несколько потоков при загрузке данных о сессиях.

 Внимание! В некоторых случаях, например, при частом сбросе кеша ОС, данная опция может существенно увеличить время загрузки данных о сессиях. При использовании данной опции, по умолчанию, создаются 2 потока для чтения данных. **В примерах исполняемых файлов указана данная опция.**

- -v, --version вывести версию программы.

Конфигурация

Параметры работы программы задаются в .properties файле. По умолчанию загружается конфигурационный файл из каталога /etc/rcollector2 с именем rcollector_MODE.properties, где MODE это выбранный режим работы. Для режима flow - flow; для режима urlget - url; для режима sipget - sip.

 При вставке данных в БД, в конфигурационном файле **обязательно** должны быть заданы следующие параметры:

- db.host
- db.port
- db.user
- db.pass
- db.name
- db.telco_code
- db.bad_rows_dir
- db.validation_error_path

При вставке данных в БД выходные файлы не создаются. В случае отсутствия подключения к БД будут создаваться выходные файлы согласно параметрам командной строки. В процессе работы может создаваться файл в каталоге из параметра db.validation_error_path, который который содержит краткую информацию об отброшенных данных, не прошешдих проверку на наличие необходимых полей. Имя файла соответствует имени входного файла с добалвением расширения .err .

Параметр cachedb

Данный параметр позволяет настроить работу с файлами данных о сессиях.

- max_reader_threads - максимальное количество потоков, запускаемых одновременно, для чтения данных о сессиях из файлов. Целое число от 0 до 6.

 В некоторых случаях слишком большое количество потоков может привести к замедлению загрузки файлов, что приведет к общему замедлению обработки данных.

Параметр stats

Данный параметр устанавливает возможность отправки статистики работы программы в telegraf.

- stats.socket_path - путь к datagram socket telegraf'a.
- stat.stag - тег, выставляемый в поле rcollector_tag при отправке статистики в telegraf.

Параметр db

Данные параметр позволяет организовать вывод полученных данных в БД ИС СОРМ-3.

- db.host - адрес сервера postgresql
- db.port - порт
- db.user - имя пользователя
- db.pass - пароль пользователя
- db.name - имя БД

- db.bad_rows_dir - каталог для размещения файлов с данными в формате PGCOPY, которые были отвергнуты сервером posgresql
- db.validation_error_path - каталог для файлов с описанием причины для оброшенных входных данных
- db.copy_threads - количество потоков, которые вставляют данные в posqresql используя COPY в бинарном формате, по умолчанию 1
- db.commit_rows - количество строк в одном блоке, отправляемом на запись в бд при использовании COPY, по умолчанию 5000
- db.telco_code - идентификатор telco для записи в соответствующее поле бд
- db.llds_id - идентификатор типа источника, по умолчанию устанавливаются следующее значения:
  - для режима flow - 309
  - для режима urlget - 310
  - для режима sipget - 311
- db.ftp.ldds_ldst_id - идентификатор типа источника для режимов flow и sipget при вставке ftp данных, по умолчанию 307
- db.email.ldds_ldst_id - идентификатор типа источника для режима sipget при вставке email данных, по умолчанию 304
- db.im.ldds_ldst_id - идентификатор типа источника для режима sipget при вставке im данных, по умолчанию 306
- db.terminal.ldds_ldst_id - идентификатор типа источника для режима flow при вставке terminal данных, по умолчанию 308
- db.h323.ldds_ldst_id - идентификатор типа источника для режима flow при вставке h323 данных, по умолчанию 311
- db.ftp_proto - идентификаторы для определения данных как ftp и их занесение в БД, по умолчанию "20,21,69,115,152,349,574,662,989,990,3713,6620,6621,6622,65086". Для отключения необходимо указать none.
- db.ssh_proto - идентификаторы для определения данных как terminal и их занесение в БД, по умолчанию "22,23,3820,992,220". Для отключения необходимо указать none.
- db.h323_proto - идентификаторы для определения данных как h323 и их занесение в БД, по умолчанию "4569,49217". Для отключения необходимо указать none.
- db.require_subscriber_id - проверять наличие subscriber_id во входных данных, по умолчанию true. Если subscriber_id будет отсутствовать и параметр выставлен в true, то запись будет отброшена, о чем будет сообщено в информационном файле
- db.http.length.htrq_url - максимальное количество символов для поля htrq_url. По умолчанию 1024
- db.ftp.length.ftpc_server_name - максимальное количество символов для поля ftpc_server_name. По умолчанию 256
- db.ftp.length.ftpc_user_name - максимальное количество символов для поля ftpc_user_name. По умолчанию 64
- db.ftp.length.ftpc_user_password - максимальное количество символов для поля ftpc_user_password. По умолчанию 256
- db.email.length.emlc_sender - максимальное количество символов для поля emlc_sender. По умолчанию 256
- db.email.length.emlc_subject - максимальное количество символов для поля emlc_subject. По умолчанию 256
- db.email.length.emlc_reply_to - максимальное количество символов для поля emlc_reply_to. По умолчанию 256
- db.email.length.emcr_receiver - максимальное количество символов для поля emcr_receiver. По умолчанию 256
- db.email.length.mlcs_server - максимальное количество символов для поля mlcs_server. По

умолчанию 256
- db.im.length.imcn_user_login - максимальное количество символов для поля imcn_user_login. По умолчанию 20
- db.im.length.imcn_user_password - максимальное количество символов для поля imcn_user_password. По умолчанию 16
- db.im.length.imcn_sender_screen_name - максимальное количество символов для поля imcn_sender_screen_name. По умолчанию 32
- db.im.length.imcn_sender_uin - максимальное количество символов для поля imcn_sender_uin. По умолчанию 256
- db.im.length.imcr_receiver_screen_name - максимальное количество символов для поля imcr_receiver_screen_name. По умолчанию 32
- db.voip.length.vipc_conference_id - максимальное количество символов для поля vipc_conference_id. По умолчанию 64
- db.voip.length.vipc_originator_name - максимальное количество символов для поля vipc_originator_name. По умолчанию 64
- db.voip.length.vipc_calling_original_number - максимальное количество символов для поля vipc_calling_original_number. По умолчанию 128
- db.voip.length.vipc_called_original_number - максимальное количество символов для поля vipc_called_original_number. По умолчанию 128
- db.raw.length.rawf_sni_cn - максимальное количество символов для поля rawf_sni_cn. По умолчанию 128
- db.do_content_id - флаг, позволяющий сохранять dpi session_id в полях data_content_id бд. По умолчанию false
- db.raw.sni_protocol - если указан протокол dpi (например 443 для ssl), то при наличии данных host_cn они будут добавлены в поле rawf_sni_cn таблицы raw_flows. По умолчанию выключено

Параметр csv

- csv.url.extra_data - если значение данного флага true, то в выходном файле для url будут содержаться дополнительные поля: user_agent, cookie, referal. По умолчанию выключено
- csv.raw.sni_protocol - если указан протокол dpi (например 443 для ssl), то при наличии данных host_cn они будут добавлены в выходной файл. По умолчанию выключено

Параметр nat

Данные параметры позволяют дополнить flow данными о трансляциях адресов в случае их отсутствия во входном файле.

- nat.sessions_dir - каталог для поиска файлов трансляций NAT. Для обработки берутся последние по времени создания файлы. Маска поиска файлов url_*.dump, url_*.dump.gz.
- nat.files_cnt - количество файлов, которые будут использованы для обработки. По умолчанию 1 файл.

Файл трансляций должен быть в формате csv с символом разделения табуляция и иметь следующий формат полей:

№ поля	**Описание**
1	Время трансляции (timestamp)

№ поля	Описание
2	Протокол
3	Тип события NAT
4	IP адрес источника
5	Порт источника
6	IP адрес источника после NAT
7	Порт источника после NAT

Параметр logging

Данный параметр отвечает за настройку логирования программы.

- logging.loggers.root.level - уровень логирования
- logging.loggers.root.channel - канал для вывода сообщений
- logging.channels.fileChannel.class - класс канала вывода
- logging.channels.fileChannel.path - путь к лог-файлу
- logging.channels.fileChannel.rotation - параметр ротации
- logging.channels.fileChannel.archive - параметр имени архивных файлов
- logging.channels.fileChannel.purgeCount - количество архивных файлов
- logging.channels.fileChannel.formatter.class - класс форматировщика
- logging.channels.fileChannel.formatter.pattern - шаблон для форматировщика
- logging.channels.fileChannel.formatter.times - время

 Более подробно ознакомиться с параметрами логирования можно по ссылке Class FileChannel.

Статистика работы программы

Типы полей статистических данных о работе программы.

Режим sip

- read_lines - количество прочитанных строк входного файла
- sip_bye - количество записей SIP BYE
- sip_invite - количество записей SIP INVITE
- sip_miss - количество записей, не имеющих информацию в кэше соединений
- count_ftp - количество записей ftp
- bad_ftp - количество ftp записей не сохраненных в файл
- out_ftp - количество ftp записей успешно сохраненных в файл
- dup_ftp - количество дублированных ftp записей
- count_mail - количество записей mail
- bad_mail - количество mail записей не сохраненных в файл
- out_mail - количество mail записей успешно сохраненных в файл
- dup_mail - количество дублированных mail записей
- count_im - количество записей im
- bad_im - количество im записей не сохраненных в файл

- out_im - количество im записей успешно сохраненных в файл
- bad_sip - количество sip записей не сохраненных в файл
- out_sip - количество sip записей успешно сохраненных в файл
- dup_sip - количество дублированных sip записей
- work_time - время работы программы в миллисекундах

Режим url

- read_lines - количество прочитанных строк входного файла
- sess_miss - количество записей для которых нет информации в данных о сессиях
- resp_miss - количество записей для которых нет информации в данных об ответах
- resp_skip - количество отброшенных записей (эти записи ответы от серверов)
- out_lines - количество сохраненных строк в выходном файле
- work_time - время работы программы в миллисекундах

Режим flow

- read_lines - количество прочитанных строк входного файла
- marked_as_tor - количество записей, промаркированных как TOR
- out_lines - количество сохраненных строк в выходном файле
- work_time - время работы программы в миллисекундах