Table of Contents

IPFIX flows assembly performed by rcollector2 3
- ***Introduction*** 3
- ***Installing and upgrading*** 3
- ***Supplied configuration files*** 3
- ***rcollector2 command line arguments*** 4
- ***Конфигурация*** 4
- ***Статистика работы программы*** 8

IPFIX flows assembly performed by rcollector2

Introduction

The utility is designed to complement data with auxiliary streams such as clickstream, SIP from the session stream (netflow). By default, the received data is saved to files, but it can also be saved in the SORM-3 information system database.

Installing and upgrading

1. add the VAS Experts repository as it is shown in n.1 of DPI installation manual.
2. install rcollector2:

   ```
   yum install -y rcollector2
   ```

3. configure the configuration files in the directory/etc/rcollector2 см. далее

Attention! It is needed to modify some startup scripts when upgrading from previous version of rcollector to rcollector2.

Supplied configuration files

1. configuration examples:

   ```
   /etc/rcollector2/rc2flowprocess - example of executable file designed
   to handle sessions (netflow)
   /etc/rcollector2/rc2urlprocess - example of executable file designed to
   handle HTTP sessions
   /etc/rcollector2/rc2sipprocess - example of executable file designed to
   handle SIP sessions
   /etc/rcollector2/rcollector_flow.properties - example of configuration
   file used in flow mode
   /etc/rcollector2/rcollector_url.properties - example of configuration
   file used in url mode
   /etc/rcollector2/rcollector_sip.properties - example of configuration
   file used in sip mode
   ```

2. executable file:

   ```
   /bin/rcollector2
   ```

rcollector2 command line arguments

rcollector2 uses the following command line arguments:

```
usage: rcollector2 OPTIONS
```

OPTIONS:

- -h, --help display brief help.
- -fCONFIG, --config-file=CONFIG configuration file.
- -mMODE, --mode=MODE mode of operation. Available modes: flow, urlget, sipget
- -uidUNIQUEID, --uniqueid=UNIQUEID unique processing point number.
- -ifINFILE, --infile=INFILE input file.
- -ofOUTFILE, --outfile=OUTFILE output to a file.
- -asnASN, --localasn=ASN comma separated list of local autonomous systems
- -oufOUTFILTER, --outfilter=OUTFILTER output records according to the dpi protocol id to a file. Format to use: <outfile>,<protocol>,...<protocol>|[next filter]. Example: telnet.dump,22,23
- -tdUSEFILTER, --tordb=USEFILTER path to the file containing IP addresses of the TOR network. By default /var/data/tor/ip.db.gz is used.
- -sdbDIR, --sessiondb=DIR path to the session data directory. By default /var/db/rcollector is used.
- -outmailFILE, --outfilemail=FILE output file intended to store data about mail connections in sipget mode.
- -outftpFILE, --outfileftp=FILE output file intended to store data about ftp connections in sipget mode.
- -outimFILE, --outfileim=FILE output file intended to store data about im connections in sipget mode.
- -dhINTEGER, --depth=INTEGER search depth to find files containing session data.
- -sdrt, --session-db-read-thread use multiple threads when loading session data.

Attention! In some cases, for example, when OS cache being flushed frequently, this option can significantly increase the load time of session data. When using this option, 2 streams are created for reading data by default. **Examples of executable files presented above use this option.**

- -v, --version display the program version.

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)
2	Протокол
3	Тип события NAT
4	IP адрес источника
5	Порт источника
6	IP адрес источника после NAT

№ поля	Описание
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