# Содержание

# 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

```
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. install rcollector2:

```
yum install -y rcollector2
```

3. edit configuration files in the /etc/rcollector2/ directory (description follows)


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.

## Configuration

The program parameters are specified in the .properties file. Configuration file named rcollector_MODE.properties in /etc/rcollector2/ directory is loaded by default, where MODE is the choosen operating mode. flow mode corresponds to **flow**; urlget mode corresponds to **url**; and the sipget mode - to **sip**.

When inserting data into the database, the following parameters **have to be** specified in the configuration file:

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

When data is inserted into the database, output files are not created. If there is no connection to the database, output files will be created according to the command line arguments. During operation, a file can be created in the directory as provided by **db.validation_error_path** option, which contains brief information about the data being discarded since it fails test that is used to detect the presence of the required fields. The file name matches the name of the input file with the **.err** extension.

#### cachedb parameter

This parameter allows you to configure the way of handling files containing session data.

- max_reader_threads - the maximum number of threads starting at the same time to read session data from files. An integer from 0 to 6.

In some cases, too many threads can slow down file downloads resulting in a general data processing slowdown.

#### stats parameter

This parameter sets the ability of program statistics to be sent to telegraf.

- stats.socket_path - path to the telegraf's datagram socket.
- stat.stag - tag to be specified in the rcollector_tag field when sending statistics to the telegraf.

#### db parameter

This parameter allows you to export the data to the external database like SORM-3 (Russian lawful interception system).

- db.host - postgresql server address
- db.port - port
- db.user - login name
- db.pass - user password
- db.name - DB name
- db.bad_rows_dir - directory to store data files that were rejected by posgresql server using

PGCOPY format

- db.validation_error_path - directory to store files describing the causes of rejected input data
- db.copy_threads - the number of threads used to insert data into posqresql DB using COPY in binary format, it equals to 1 by default
- db.commit_rows - the number of rows in one block sent for writing to the database when using COPY, it equals to 5000 by default
- db.telco_code - telco identifier used to write to the corresponding db field
- db.llds_id - source type identifier, the following values are set by default:
    - for the flow mode - 309
    - for the urlget mode - 310
    - for the sipget mode - 311
- db.ftp.ldds_ldst_id - source type identifier for flow and sipget modes when inserting ftp data, default value is 307
- db.email.ldds_ldst_id - source type identifier for sipget mode when inserting email data, default value is 304
- db.im.ldds_ldst_id - source type identifier for sipget mode when inserting im data, default value is 306
- db.terminal.ldds_ldst_id - source type identifier for flow mode when inserting terminal data, default value is 308
- db.h323.ldds_ldst_id - source type identifier for flow mode when inserting h323 data, default value is 311
- db.ftp_proto - identifiers used to store data into DB as ftp, values "20,21,69,115,152,349,574,662,989,990,3713,6620,6621,6622,65086" are used by default. To disable, you should specify **none**.
- db.ssh_proto - identifiers used to store data into DB as terminal, values "22,23,3820,992,220" are used by default.
- db.h323_proto - identifiers used to store data into DB as h323, values "4569,49217" are used by default. To disable, you should specify **none**.
- db.require_subscriber_id - check for subscriber_id in the input, by default it's equal to true. If subscriber_id is absent and the parameter is set to true, then the record will be discarded, which will be reported in the information file
- db.http.length.htrq_url - maximum number of characters for htrq_url field. Default value is 1024
- db.ftp.length.ftpc_server_name - maximum number of characters for the ftpc_server_name field. Default value is 256
- db.ftp.length.ftpc_user_name - maximum number of characters for ftpc_user_name field. Default value is 64
- db.ftp.length.ftpc_user_password - maximum number of characters for ftpc_user_password field. Default value is 256
- db.email.length.emlc_sender - maximum number of characters for emlc_sender field. Default value is 256
- db.email.length.emlc_subject - maximum number of characters for emlc_subject field. Default value is 256
- db.email.length.emlc_reply_to - maximum number of characters for emlc_reply_to field. Default value is 256
- db.email.length.emcr_receiver - maximum number of characters for emcr_receiver field. Default value is 256
- db.email.length.mlcs_server - maximum number of characters for mlcs_server field. Default value is 256
- db.im.length.imcn_user_login - maximum number of characters for imcn_user_login field. Default value is 20
- db.im.length.imcn_user_password - maximum number of characters for imcn_user_password

field. Default value is 16
- db.im.length.imcn_sender_screen_name - maximum number of characters for imcn_sender_screen_name field. Default value is 32
- db.im.length.imcn_sender_uin - maximum number of characters for imcn_sender_uin field. Default value is 256
- db.im.length.imcr_receiver_screen_name - maximum number of characters for imcr_receiver_screen_name field. Default value is 32
- db.voip.length.vipc_conference_id - maximum number of characters for vipc_conference_id field. Default value is 64
- db.voip.length.vipc_originator_name - maximum number of characters for vipc_originator_name field. Default value is 64
- db.voip.length.vipc_calling_original_number - maximum number of characters for vipc_calling_original_number field. Default value is 128
- db.voip.length.vipc_called_original_number - maximum number of characters for vipc_called_original_number field. Default value is 128
- db.raw.length.rawf_sni_cn - maximum number of characters for rawf_sni_cn field. Default value is 128
- db.do_content_id - identifier allowing to store dpi session_id in data_content_id field. Default value is false
- db.raw.sni_protocol - if the dpi protocol is specified (for example, 443 for ssl), then if there is host_cn data, they will be added to the rawf_sni_cn field of the raw_flows table. It is disabled by default

**csv parameter**

- csv.url.extra_data - if the value of this flag is true, then the output file for url will contain additional fields: user_agent, cookie, referal. It is disabled by default
- csv.raw.sni_protocol - if the dpi protocol is specified (for example, 443 for ssl), then if there is host_cn data, they will be added to the output file. It is disabled by default

**nat parameter**

These parameters allow you to complement flow with data about address translations in case they are not present in the input file.

- nat.sessions_dir - directory to search for files containing NAT translations. The latest files are used for processing. The search mask for files: url_*.dump, url_*.dump.gz.
- nat.files_cnt - number of files to be used for processing. 1 file will be processed by default.

File containing NAT translations should be in csv tab-separated format and should have the following format of fields:

| **№ of field** | **Description** |
|---|---|
| 1 | Timestamp of NAT translation |
| 2 | Protocol |
| 3 | Type of NAT event |
| 4 | Source IP address |
| 5 | Source port |
| 6 | Source IP behind the NAT device |

| № of field | Description |
|---|---|
| 7 | Source port behind the NAT device |

**logging parameter**

This parameter is responsible for configuring program logging.

- logging.loggers.root.level - log level to be used
- logging.loggers.root.channel - message channel
- logging.channels.fileChannel.class - class of output channel
- logging.channels.fileChannel.path - path to the log file
- logging.channels.fileChannel.rotation - rotation parameter
- logging.channels.fileChannel.archive - parameter specifying naming of archive files
- logging.channels.fileChannel.purgeCount - number of archive files
- logging.channels.fileChannel.formatter.class - formatter class
- logging.channels.fileChannel.formatter.pattern - formatter template
- logging.channels.fileChannel.formatter.times - time

 For more information on logging options, click here: Class FileChannel.

# Statistics of the program

Types of statistics fields on program operation.

**sip mode**

- read_lines - number of input file lines that have been read
- sip_bye - number of SIP BYE entries
- sip_invite - number of SIP INVITE records
- sip_miss - number of entries that do not have information in the connection cache
- count_ftp - number of ftp entries
- bad_ftp - number of ftp entries not saved to file
- out_ftp - number of ftp records successfully saved to file
- dup_ftp - number of duplicated ftp entries
- count_mail - number of mail records
- bad_mail - number of mail records not saved to file
- out_mail - number of mail records successfully saved to file
- dup_mail - number of duplicate mail records
- count_im - number of im records
- bad_im - number of im records not saved to file
- out_im - number of im records successfully saved to file
- bad_sip - number of sip entries not saved to file
- out_sip - number of sip records successfully saved to file
- dup_sip - number of duplicated sip entries
- work_time - program run time in milliseconds

**url mode**

- read_lines - number of input file lines that have been read
- sess_miss - number of records for which there is no information in the session data
- resp_miss - number of records for which there is no information in the responses data
- resp_skip - number of skipped records (these records are responses from servers)
- out_lines - number of stored lines in the output file
- work_time - program run time in milliseconds

**flow mode**

- read_lines - number of input file lines that have been read
- marked_as_tor - number of entries marked as TOR
- out_lines - number of stored lines in the output file
- work_time - program run time in milliseconds