Security Analytics 8.0.1
Best Searching Practices

Updated: Tuesday, November 20, 2018

Symantec

Symantec Security Analytics 8.0.1

Copyrights, Trademarks, and Intellectual Property

These help files are intended to help you use the web browser and console interfaces for Security Analytics to perform network traffic capture, filtering, and playback, as well as general administration. It is not intended as a guide to policies or procedures for network security or network forensics.

Symantec Support

Your serial number is visible in *About*.

- Contact Information: support.symantec.com/en_US/contact-support.html
- Symantec Customer Care, Network Protection: **NP_customercare@symantec.com**
- Security Analytics Documentation: support.symantec.com
- Documentation Feedback: **documentation_inbox@symantec.com**

This document describes best practices to help analysts find specific data ("target data") using the multiple filtering and data-extraction functions that are provided by Security Analytics. This document presumes that the reader has an understanding of networking and the corresponding relevant OSI models.

Search Procedure Summary

Follow these steps when searching for a particular piece of information. Click the link for a detailed explanation of the bullet points for each step.

1. Select the metadata that you want to index on *Menu*  *> Settings > Metadata*.
    - Remember that changing the *Metadata Settings* page requires an appliance reboot.
    - Metadata that is not selected cannot be indexed and is therefore not available for reports.
    - As desired, you can reindex selected timespans after adding new metadata attributes.
2. **Optional** — Create a "sparse" default *Summary* page:
    - Use as few widgets as possible: 18 or fewer.
    - Omit higher-cost widgets such as *HTTP URI* and *File Name*.
    - *Application Group* and *Application ID* are good starting filters for large timespans.
    - All widgets are from the same namespace.
    - Open different reports/views in different tabs.
3. Identify the target data's characteristics:
    - Located in metadata, packet data, artifacts, or reputation.
    - Email, binary, image, IM, script, or web page.
    - Unique or common.
    - Temporally restricted or ubiquitous.
    - Number of users.
4. Consider factors that contribute to report costs:
    - Natively indexed metadata vs. metadata produced by a post-DPI process.
    - Cached reports are faster.
    - Large numbers of results (+100K) vs. limited possible results.
    - Stored as a bitmask or a string.

- Hashes: whether enabled, written to the Indexing DB, or extracted only.
- CMC: number of sensors plus amount of data.

5. Apply timespan filters:
   - Fastest way to narrow the amount of data to search.
   - Use histograms to select smaller timespans (*Application Group over Time* on the *Summary* page or *[X] over Time* on the *Reports* page).
   - Avoid searching timespans where target data is not present.
   - Always triggers a new search/extraction.
6. Apply primary filters:
   - Searches only the Indexing DB (indexed metadata).
   - Use filter sequence to quickly narrow the results.
   - Think in terms of exclusion: **protocol_id**, **application_id**, **mime_type** can eliminate large amounts of data quickly.
   - Primary filters can be saved as indicators; they persist across *Analyze* pages.
   - The Indexing DB data usually persists longer than packet data.
   - Use wildcards judiciously: usually as the last filtering element.
   - Always triggers a new search/extraction.
7. Apply advanced filters:
   - Applies only to filter results—no new search/extraction triggered.
   - Easy-to-apply complex AND/OR series.
   - Only way to find hashes that were not written to the Indexing DB.
8. **Perform extractions last!**
   - Not until search parameters are as narrow as possible.
   - Use a primary filter to exclude protocols that Security Analytics cannot extract, such as **SSL** and **TLS**.
   - Only keywords in plaintext can be detected; use third-party resources for application-encoded, compressed, and encrypted files.
9. Use **ngrep** to search across packet data.

Identify the Target Data

As obvious as it may seem, taking the time to identify the precise nature of the target data is a step that many analysts neglect. Without a precise understanding of what to find, analysts can waste time using the wrong tools or methods.

Establish Data Characteristics

Ask questions such as these:

- Which data can be excluded right away?
- Is the data linked to a particular user or group of users?
- Is the data likely to be found during a particular time span?
- Is the data contained in a particular file type: email, instant messages, web sites, binary files?
- Is the data likely to be in a particular flow type: encrypted or non-encrypted, UDP or TCP?
- Is the data unique or unusual: infrequent country of origin, low-traffic web site, an unusually large or small size, a rare port number, high frequency of data transport?
- Is the data an MD5, SHA1, SHA256, or fuzzy hash?

Determine the Data Location

Data are located in various Security Analytics subsystems:

- **Indexed Metadata** — Indexed metadata (on the indexing array or Indexing DB) contains data that is extracted by the deep-packet inspection (DPI) engine from packet-header contents such as transport-layer or HTTP headers. Other attributes are provided by the system at the time that the data is being indexed.
- **Packet Payloads** — The capture drive contains the raw packet data of all traffic that was mirrored to the capture interfaces. The data on the capture drive can be downloaded in standards-based PCAP and PCAPNG formats.
- **Artifacts** — Artifacts are files that are reconstructed from packets by the Security Analytics extraction process. Artifact types include but are not limited to PDFs, archives, config files, binaries, email, images, multimedia, documents, executables, libraries, web pages, and JavaScript files.
- **Enriched Data and Reputation Services** — Intelligence Services, reputation providers and analytics resources can provide additional data about an artifact such as file or URL reputation, file scanning and analysis, and URL categorization.

Search-Method Characteristics

Each search method provides its own advantages to help you provide the most efficient and most accurate results.

Timespan Filters

Timespan filters should be the first or second filter that you apply in nearly all circumstances. Because narrower time spans usually equate to smaller data sets, they produce faster results than larger timespans.

Advantages

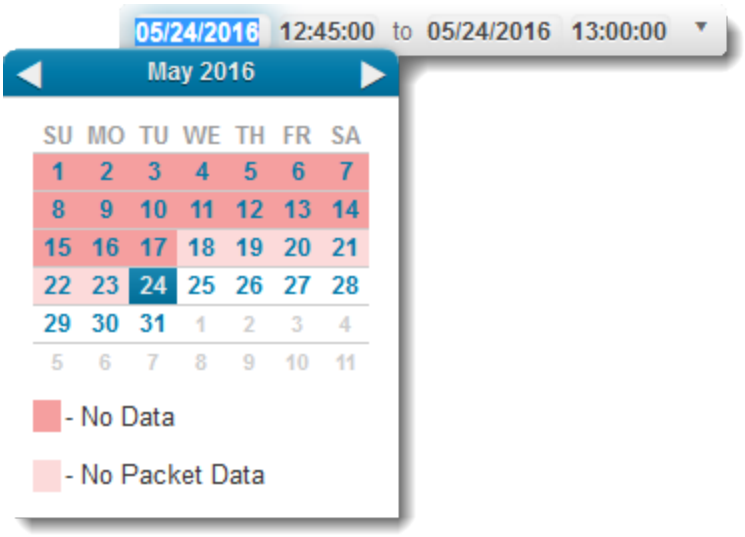
- Fastest way to constrain the amount of data to search.

Considerations

- Updating a timespan filter triggers the creation of a new report.
- All of the data in the entire timespan is searched regardless of whether the target data is present. For example, if the target IP address is present only in the last 15 minutes of traffic, but the timespan filter specifies the last 24 hours, all 24 hours of data will be examined. If possible, after identifying where data exists, narrow the window of time to reduce the amount of data searched.
    - If the target data is not present during the specified timespan, the report will generally build faster than when data is present; however, the data in the report is unlikely to be useful.
    - Searching timespans where data is present — which generates data that is useful to your investigation — will build a report more slowly than running a report for the same timespan where data is not present. For example, if you search a 15-minute timespan for an IP address that is present, the report will be built more slowly than if you search that same timespan for an IP address that is not present. The same holds true for the frequency of data in the report. As data increases in frequency, report time also increases. For example, a report that matches an IP address 10,000 times will build slower than a report where the IP address matches only 100 times.

Primary Timespan Filter

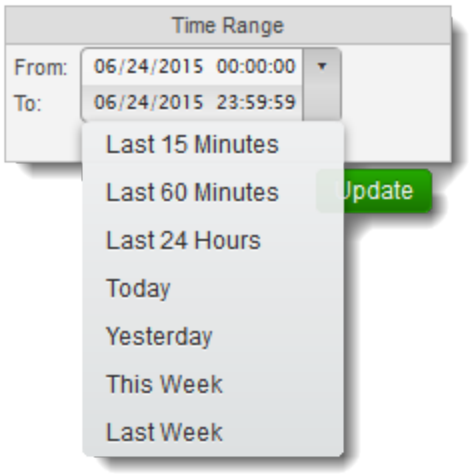
The primary timespan filter is available on all *Analyze* sub-pages — *Summary*, *Reports*, *Extractions*, *Geolocation*.



Primary Timespan Filter

Time-Range Widget

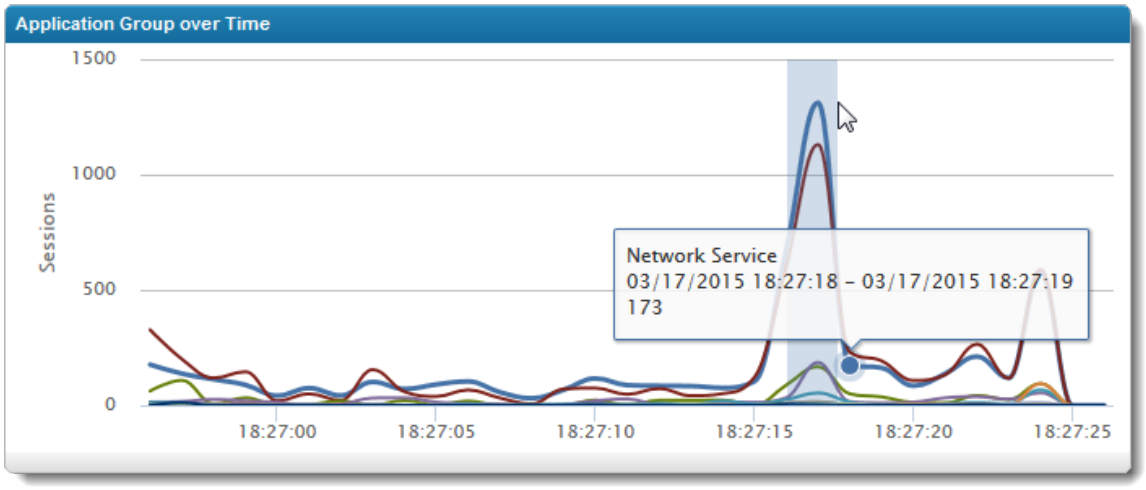
The *Time Range* widget on the *Alerts* and *Anomalies* pages has its own **Update** button and therefore operates separately from the primary timespan filter. Updating this widget does not trigger the creation of a new report.



Time Range Widget

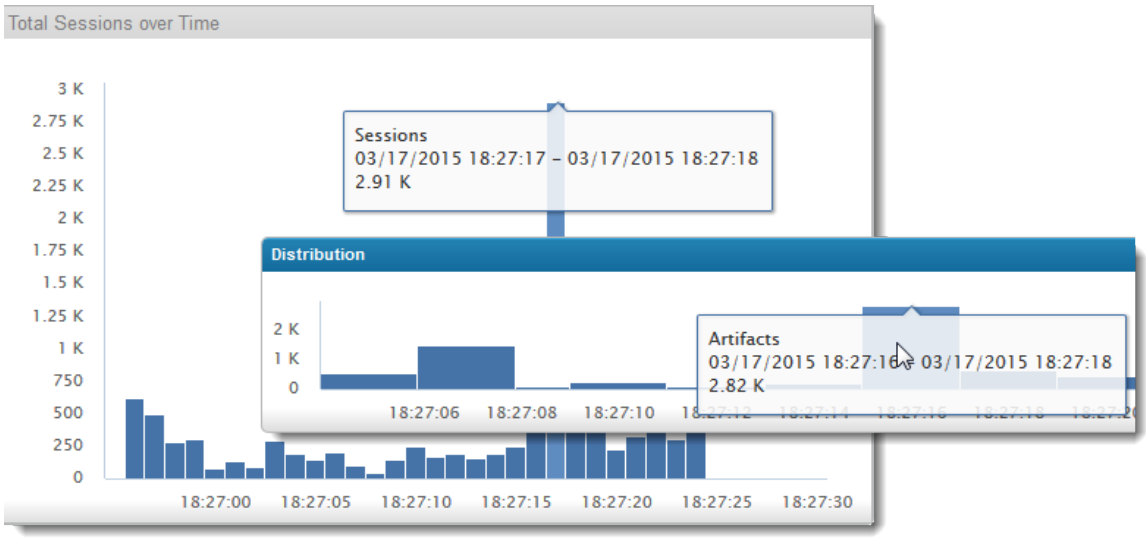
Histograms

Set primary timespans using histograms that are present on the UI.



Application Group over Time Histogram

On the *Application Group over Time* widget, select a single point or a span. The timespan that is displayed in the balloon is transferred to the primary timespan filter when you release the mouse button.



Reports and Extractions Histograms

Most reports present a *Total Sessions over Time* histogram on the *Reports* page, and the *Extractions* pages display the distribution of artifacts across a timeline. The vertical bars in each histogram can be selected individually or in a group to populate the primary timespan.

The *Artifacts Timeline* extraction page provides further details about artifact distribution.

Primary Filter

Primary filter attributes search all network flows for the presence of indexed metadata.

Primary Filter Bar

Advantages

- The indexing DB contains only metadata, so searches are faster than searching against extracted artifacts or across all packets.
- Attribute order controls how broad or narrow the search is. Primary filters are applied from left to right. If you know a highly specific attribute of the target data, input that attribute first.
- You can use **AND** and **OR** operators to create more complex filters.
- Primary filters persist across all *Analyze* tabs: *Summary*, *Reports*, *Extractions*, *Geolocation*.
- Filters can be saved as indicators, which can then be used in rules, subsequent filters, or reports.
- The indexing drive usually has a longer recycle interval, so metadata often persists after the packet data has been overwritten.

Considerations

- Updating a primary filter triggers the creation of a new search or extraction.
- Only attributes that are indexed with the DPI engine are available for search and reporting at the metadata level.
- Wildcards in the filter can considerably add to the cost.
- You cannot combine filters from different namespaces in complex filters.

Primary Filter Hierarchy

Primary filters consist of one or more attribute/value pairs.

> See the list of the Primary Filter Attributes in the *Security Analytics 8.0.1 Help Files* on **support.symantec.com**.

The filters are applied from left to right, such that for the filter

Example Primary Filter

the data is filtered first on the **application_id** value and then on the **ipv4_initator** value, which results in all entries where the application is HTTP and the initiator IP is not **10.10.2.123**.

Avoid using a primary filter with multiple indicators including wildcards as your first attempt to find the target data. For example, inputting the top filter will not generate results as fast as the second filter.

Example of Less-Efficient Primary Filter

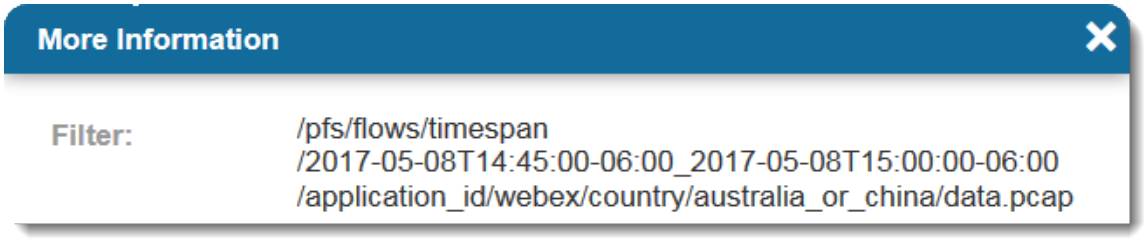
Example of More-Efficient Primary Filter

It is better to determine which filters will most quickly exclude unwanted data and apply them first, before using wildcards or other indicators that force the query handler to read far more records than needed. If you need to include a wildcard in a filter, add that filter last, after you have narrowed down the data set as much as possible with other filters.

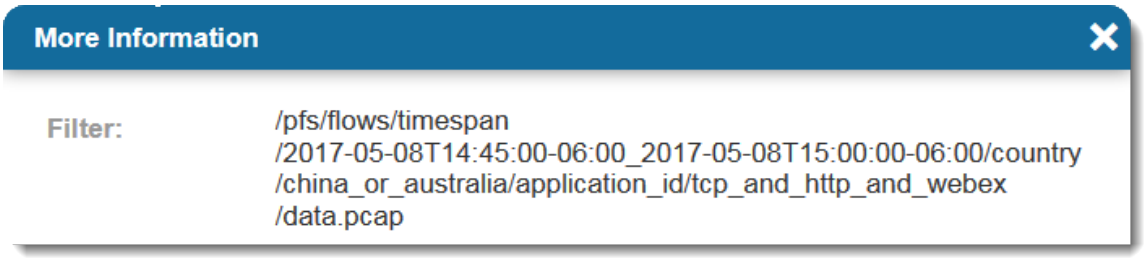
Primary Filter Logic

When you enter filter definitions in the primary filter bar, the logical equivalent is displayed below the graphical display.

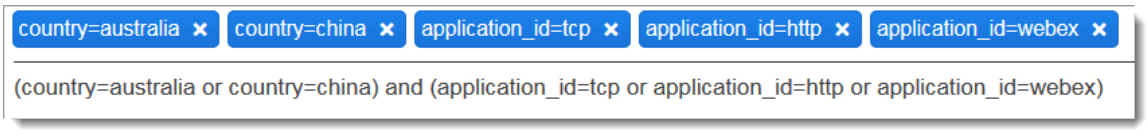
The logical display shows how Boolean **AND** joins filters with different attributes, whereas filters with the same attribute are joined with **OR**. This is the query path:

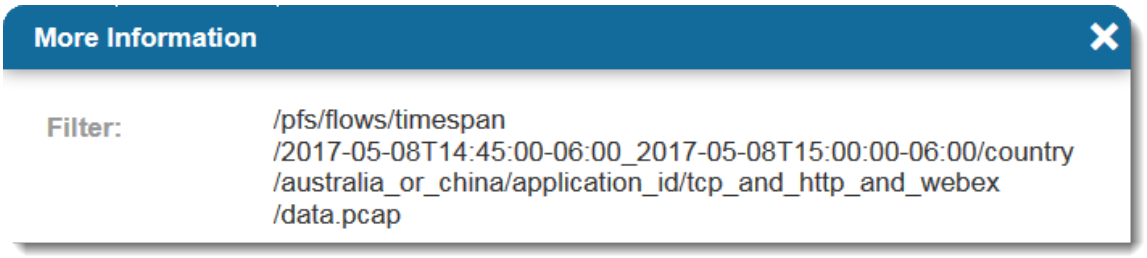


The logical display also shows how filters that contain multiple, comma-delimited values for the same attribute are joined by **AND**.



If the **application_id** values were entered as individual attributes, they would be joined by **OR**.





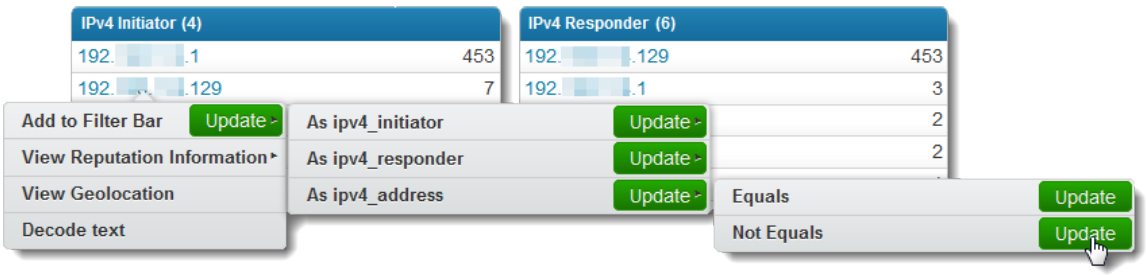
> Also see Creating Complex Filters in the *Security Analytics 8.0.1 Help Files* on **support.symantec.com**.

Primary Filter Usage

If you have determined that the target data is likely to be found in the metadata, consult the **Attribute** column in the metadata settings tables for the syntax of the most restrictive known characteristic of the target data. Before entering the filter, determine which other attributes you want to see for that data and configure either the *Summary* view (multiple attributes) or the *Reports* page (one attribute) to see those attributes. (The *Reports* page usually produces faster results).

For example, you know that the target of an exploit is in the **10.1.7.0/24** network and you suspect that the exploit involves **HTTP**.

1. On the *Report* page, select **HTTP_URI.**
2. In the primary filter bar enter **ipv4_responder=10.0.7.*** and click **Update**.
3. The histogram helps identify where the target data is present in the timespan. If possible, select a narrower timespan before continuing the investigation.
4. Pivot to other reports or switch to the *Summary* view to find other attributes to add to the primary filter.



Creating a Primary Filter from the Web UI

5. Click a value in a report or report widget to add it to the primary filter bar.
6. As desired, view other *Summary* views or select other reports to view.
7. As applicable, continue to adjust the timespan to narrow the amount of data to search.

> Do not click the **Extractions** tab until the narrowest possible filter is present in the primary filter bar.

Intelligent Search-Filter Tricks

Some examples of "interesting" filters are included with the standard set of indicators, such as a filter to find SSH traffic on ports other than 22 (Non Standard SSH Port) or traffic on port 22 that is not SSH (Non Standard SSH). Included below are filter definitions that you can create, which may provide value or spark ideas for investigating or hunting for things in new ways.

Query	Filter	Sample Returned Data
Unusually short user agent	`len(user_agent)<=10`	*User Agent* report **Connection**
Potential session hijack: User Agent changed during flow	`num(user_agent)>=2`	*User Agent* report **Mozilla/5.0 (Windows NT 6.1; WOW64; rv=0~xa** **Mozilla/5.0 (Windows NT 6.1; WOW64; rv:41.0**

Query	Filter	Sample Returned Data
Large number of files in a flow	**num(filename)>50**	*File Name* report **4fa776c524e20468351e916aa87d3442.10.jpg** **4fa776c524e20468351e916aa87d3442.11.jpg** **4fa776c524e20468351e916aa87d3442.12.jpg**
Large number of email recipients in a flow	**num(email_recipient)>50**	*Email Recipient* report **<redacted>@gmail.com <redacted2>@gmail.com**
Traffic without an application classification. Useful for finding unclassifiable traffic.	**application_id=unknown**	*Application* report **tcp > unknown** **udp > unknown**
Report for any traffic destined to a particular subnet that does not use TCP port 3389	Create an indicator for **tcp_port=3389** called **TCPPORT**. Use the inverse of that indicator: **!TCPPORT** in the primary filter bar:	*IPv4 Port Conversation* report **10.1.1.1:2055-10.1.1.27:2055**

Advanced Filters

Advanced filters are present on the *Reports*, *Extractions*, *Geolocation*, and *Alerts* pages.

Advantages

- Searches only the results of the report.
- Applying and updating filters does not trigger a new report or extraction.
- Complex filters can be created by groups of attributes that are linked by Boolean operators:

```
[a OR b OR C] AND [d OR e]
```

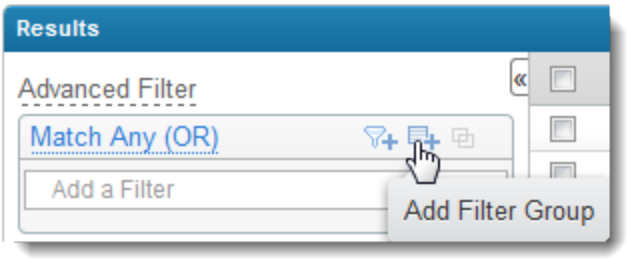
Considerations

- Advanced filters do not persist across pages and reports.
- Filters cannot be saved for later use.
- Keyword searches are case-sensitive, are valid only for cleartext strings (not encoded or compressed), and available only on the **Extractions** tab.

> See the Advanced-Filter Attributes in the *Security Analytics 8.0.1 Help Files* on **support.symantec.com**.

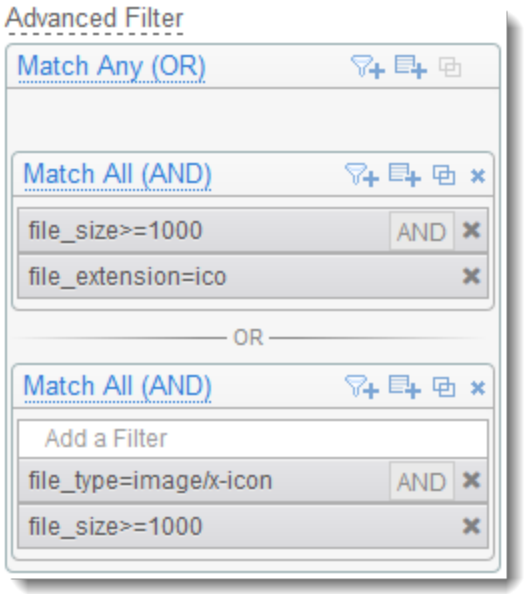
Advanced Filter Usage

1. To create complex filters, select the Boolean that links the first group with the second (AND, OR) and then click **Add Filter Group**.



Adding an Advanced Filter Group

2. Select the Boolean that links the terms within the first group and then add filters.
3. To add the second group, click the same **Add Filter Group** icon as for the first group and then add the filters for the second group.



Nested Advanced Filters

RAW.TSV File

Advantages

- Retrieves data from multiple fields relative to a single data item.

Considerations

- A spreadsheet application such as Excel is required to view the contents in a human-friendly format.
- Data is "raw" — not formatted.
- Not all report attributes are available in the **raw.tsv**; for example, hashes and data-enrichment verdicts.

Download the RAW.TSV File

On any *Analyze* page *[Summary, Reports, Extractions, Geolocation]*:

1. Select **Actions > Download Raw TSV**.
2. Select the fields to include in the **TSV** file and click **Download Raw TSV**.
3. Follow your browser prompts to download **raw.tsv**.

> Consult **Reference > RAW.TSV Fields** in the *Security Analytics 8.0.1 Help Files* on **support.symantec.com** for a list of **valid RAW.TSV** attributes.

ngrep

ngrep is similar to **grep** except that it searches the data in captured packets.

Advantages

- Permits GNU regex searches

Considerations

- Searches across packet data only (capture drive); searches may be slower than metadata queries.
- Only cleartext data is searched. Compressed or encoded data is not inflated or normalized on the capture drive.
- If the packet data is saved in an encrypted format, you cannot search it with **ngrep**. Apply a proper primary filter to exclude encrypted transports such as SSL, SSH, RDP.
- Searches from the CLI only.

- **ngrep** supports searching with regular expressions and hexadecimal values. The main use for **ngrep** is to search a **data.pcap** file that is generated by the capture system in **/pfs/flows**.

Type **ngrep -help** to see the syntax and options.

```
usage: ngrep <-hnxviwqpevxldttrm> <-io pcap_dump=""> <-n num=""> <-d dev=""> <-a num=""> <-s
snaplen=""> <-s limitlen=""> <-w normal|byline|single|none=""> <-c cols=""> <-p char=""> <-f
file=""> <match expression=""> <bpf filter="">
```

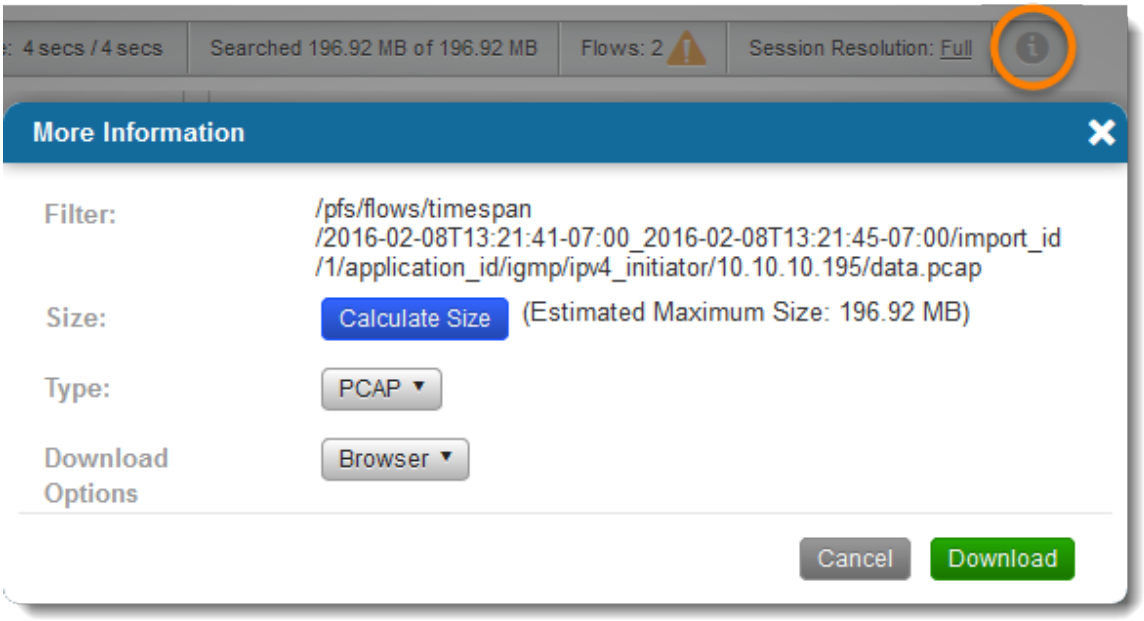
ngrep Usage

Follow these steps to run **ngrep**:

1. Log in as **root**.
2. Navigate to the timespan to search:

```
cd /pfs/flows/timespan/<YYYY-MM-DD>T<hh:ii:ss>[+|-]<zz>:00_<YYYY-MM-DD>T<hh:ii:ss>
[+|-]<zz>:00
```

> If you are unsure of the **/pfs/flows** syntax, build the query in the web UI and then click the **More Info** icon to see and copy the proper path.



***More Information* Dialog, Showing /pfs/flows Syntax**

3. At the timespan prompt, enter an **ngrep** command such as the following, which searches for MasterCard credit card numbers:

```
ngrep -q -w '5[0-9]{3}[ -]?[0-9]{4}[ -]?[0-9{4}[ -]?[0-9]{4}' -l -I data.pcap
```

> In versions 7.3.3 and later, you can create open parser rules to apply regex searches to incoming traffic.

4. The results are formatted as follows:

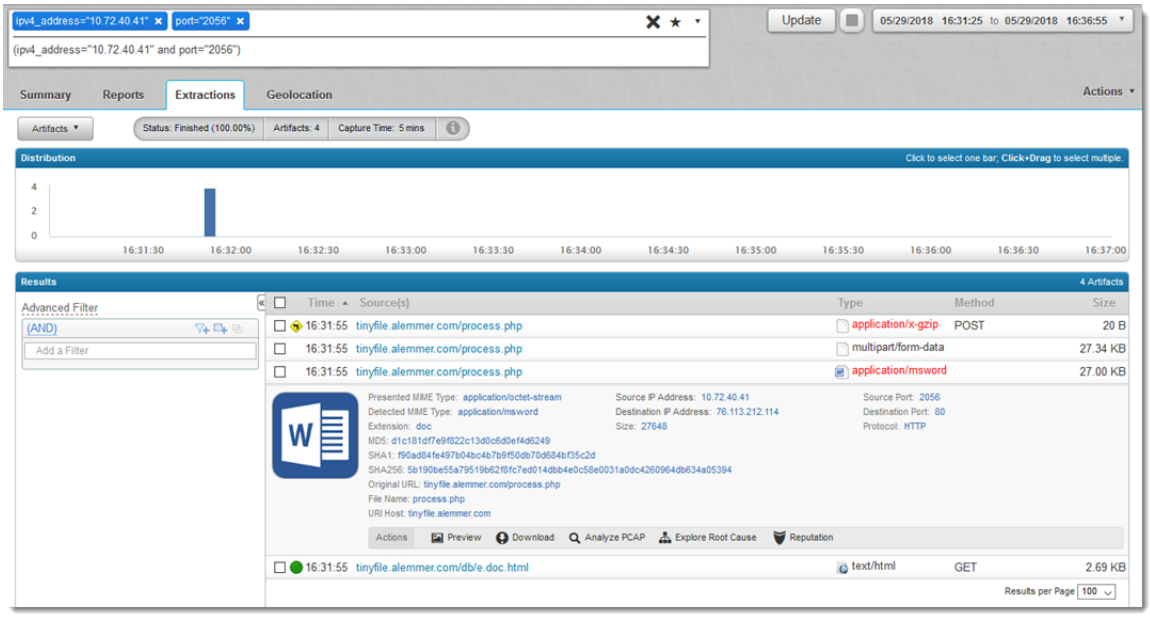
```
T <ip>:<port> -> <ip>:<port>
<raw text>
```

example 1

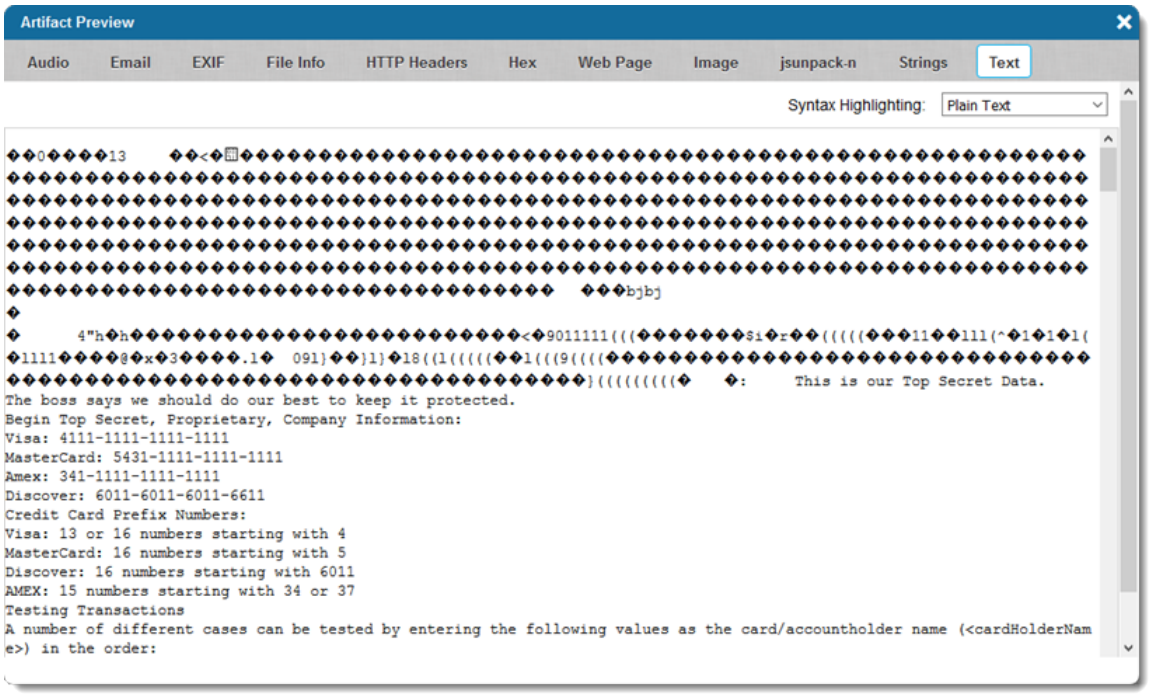
Find MasterCard numbers in all PCAPs for the timespan:

```
[root@hostname ~]# cd /pfs/flows/timespan/2018-05-29T16:31:25-06:00_2018-05-29T16:36:55-06:00
[root@hostname 2018-05-29T16:31:25-06:00_2018-05-29T16:36:55-06:00]# ngrep -q -w '5[0-9]{3}[ -]?[0-9]
{4}[ -]?[0-9{4}[ -]?[0-9]{4}' -l -I data.pcap
T 10.72.40.41:2056 -> 76.113.212.114:80 [A]
ur best to keep it protected..Begin Top Secret, Proprietary, Company Information:.Visa: 4111-1111-
1111-1111.MasterCard: 5431-1111-1111-1111.Amex: 341-1111-1111-1111.Discover: 6011-6011-6011-
6611.Credit Card Prefix Numbers:.Visa: 13 or 16 numbers starting with 4.MasterCard: 16 numbers
starting with 5.Discover: 16 numbers starting with 6011.AMEX: 15 numbers starting with 34 or
37.Testing Transactions.A number of different cases can be tested by entering the foll owing values as
the card/accountholder name (<cardHolderName>) in the order:. REFUSED will simulate a refused payment.
REFERRED - will simulate a refusal with the refusal reason referred'. FRAUD - w ill simulate a refusal
with the refusal reason fraud suspicion'. ERROR - will simulate a payment that ends in error.All other
card/accountholder names will simulate an authorised payment..For test purposes we have provi ded a
set of test credit and debit card numbers, these are listed below in the Test Card Numbers
section..Captures and refunds can be simulated through the Merchant Interface. Use the "Capture" or
"Refund" button in the Payment and Order Details page. Alternatively, you can send an XML capture or
refund order modification to the Test environment..Test Card Numbers..The following card numbers can
be used when you make test transactions in Test environments onl y - do not use them in live,
Production environments:..Card Type....
```

To extract this file in the UI, input the IP addresses and port to the Primary Filter, configure the timespan filter and run the extraction.



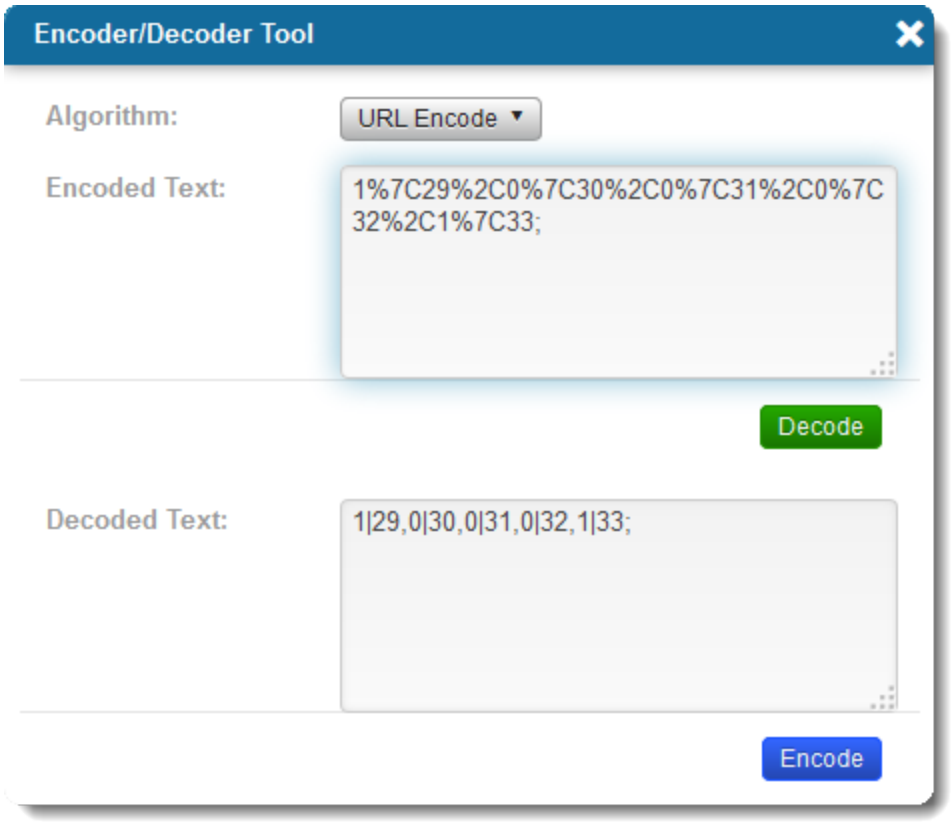
- To narrow the results to one artifact, apply an advanced filter on a unique string in the URL.
- For an HTTP **GET** artifact, select **Preview > Text** to see the same results that were returned by **ngrep**.



Text Preview with ngrep Results String

- Because the number is rendered in plaintext, **ngrep** was able to find it, even though it appears that this is not a real MasterCard number. Had the credit-card number been URL-encoded, program-encoded, or otherwise obfuscated, the number would not have been returned by **ngrep**.

- As desired, you can paste short snippets of text into ***[Account Name]*** > **Encoder/Decoder Tool** to encode or decode it in URL, Punycode, Rot13, or Base64 encoding.



Encoder/Decoder Tool

example 2

Find MasterCard numbers in all PCAPs except SSL and SSH traffic:

```
[root@hostname ~]# cd /pfs/flows/timespan/2017-06-24T21:23:17-06:00_2017-06-24T21:23:19-06:00/
application_id/_not_ssl/application_id/_not_ssh/
[root@hostname 2017-06-24T21:23:17-06:00_2017-06-24T21:23:19-06:00]# ngrep -q -w '5[0-9]{3}[ -]?[0-9]
{4}[ -]?[0-9{4}[ -]?[0-9]{4}' -I data.pcap
```

More documentation for **ngrep** can be found online:

- http://ngrep.sourceforge.net/
- https://sickbits.net/mining-networks-for-pii-with-ngrep/
- http://packetlife.net/blog/2010/may/14/grepping-packets-ngrep/

Report Costs

All investigations require that you run one or more reports, so you should take the "cost" to generate a report into account; otherwise, you might expend too much time running costly reports instead of efficient ones. The cost (time or system resources) to run a report is affected by the following factors:

- Number of Records to Return
- Metadata Source
- Post-Indexing Calculations
- Bitmask Storage
- System Load

Number of Records to Return

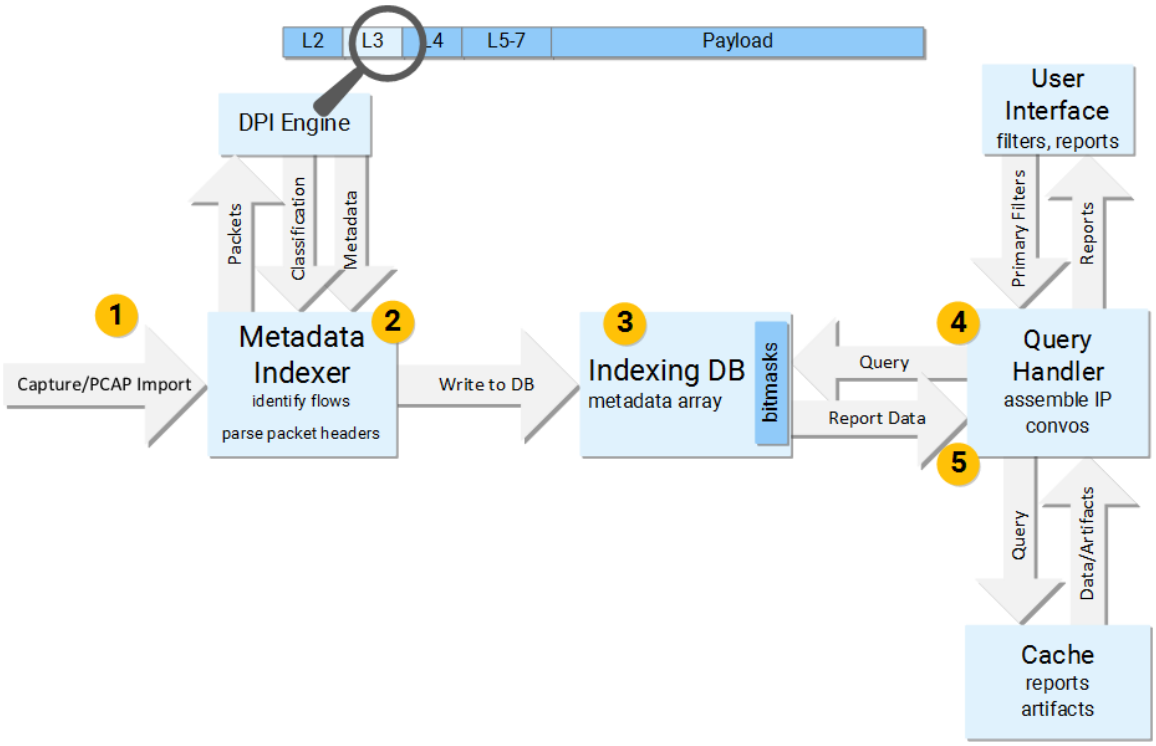
Any report that matches 100,000 entries or more will take longer to produce, because the query handler performs a series of calculations to determine which 100,000 entries to return. For example, if your query matches 2 million records in the Indexing DB, the query handler will retrieve all 2 million records, sort them by session, and then return the most pertinent 100,000 — which is roughly the top and bottom halves of the sorted table so that you get the outliers as well as the top results. A query that matches fewer than 100,000 records, however, is written directly to the report results without delay and furthermore includes all matches for the timespan instead of a selection.

In many cases, you can predict which reports are at risk for producing more than 100K matches by considering the following:

- **Unique Values** — For some reports, the number of unique results is theoretically infinite, such as file names, IP addresses, URLs, or hashes. Other reports are limited by their own definitions. For example, the *Country*-related reports are limited by the number of recognized countries in the world, and *IP Protocol* is limited by the number of Layer-4 protocols in use.
- **Type of Data Captured** — If your network data contains a high concentration of unique values, any report that searches for those values is likely to exceed 100,000 entries, especially if your timespan is broad. For example, if your traffic contains a high number of unique IP addresses, the IPv4 reports are much costlier than those same reports on a LAN that primarily serves users on internal networks.
- **Amount of Data Queried** — Usually related to the query timespan, if the query would exceed 100,000 values for a report, run the same report multiple times using smaller timespans.

Metadata Source

Most metadata is written to the Indexing DB according to the process shown below. Such metadata is considered to be "natively indexed," and is written to the Indexing DB more quickly than other types of metadata:



Report-Generation Process

1	Packet data arrives from either the capture interfaces or imported PCAPs.
2	The metadata indexer sends the packets to the DPI engine, which classifies the contents of the packets. The metadata indexer also identifies which packets belong to the same "flow" — a complete session between an initiating device and its responder.
3	The metadata indexer writes the metadata with its corresponding flow ID to the Indexing DB (indexing array). Some of the metadata is stored with a bitmask (IP addresses, ports), which greatly reduces the time it takes to search for and retrieve the metadata.
4	When an analyst initiates queries from the user interface, the APIs, or the CLI, the query is sent to the query handler. The query handler reads the requested data from the Indexing DB and returns it to the user interface. (If the query requests the *IPv4* or *IPv6 Conversation* report, the query handler must perform additional calculations to return the initiator/responder pairs. The conversations are not written to the Indexing DB but must be calculated each time they are requested.)
5	If the report data has been cached, the query handler returns results more quickly. Only reports with the same parameters as a cached report can be retrieved from the cache. New query parameters generate new reports.

Post-Indexing Calculations

The query handler must assemble *IPv4/IPv6 Conversation* reports upon request, by mapping the initiator IPs with their corresponding responder IPs from the same session. Every time the *Conversation* reports are requested, the query handler must re-map the IPs unless the request is exactly the same as a cached report.

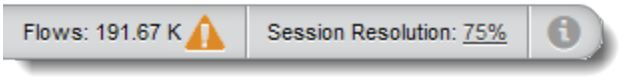
Bitmask Storage

Values that are stored with bitmasks are returned more quickly than those that are not: purely numerical values such as IP addresses and port numbers are stored in bitmask form, whereas strings such as web queries, HTTP URIs, and HTTP content disposition cannot be stored as bitmasks. See Report Costs for a list of reports that store content as bitmasks.

When you include wildcards in a query, the query handler cannot use the bitmasks to exclude non-matching entries, so the retrieval speed decreases considerably.

System Load

If the system is recording traffic in excess of its recommended capture-rate guidelines, it's possible that some packets may not be indexed immediately following capture. On the *Summary* page, an alert icon warns you when packets have not been indexed. Security Analytics automatically returns to the unindexed packets during periods of lower system usage and indexes them. You can also force the reindexing of selected data using the following methods:



Classification Discard Alert

- On the default *Summary* page, select the timespan to re-index, click the alert icon and click **Give Priority to This Timespan** to move the unindexed data in the current view to the top of the reindexing queue.
- On the *Capture Summary* page, select **Actions > Reprocessing** to see the progress of the reindexing jobs. On the capture summary graph, **View > Classification Discards** shows the rate at which packets are not being indexed.

Central Manager Considerations

When viewing reports on a Central Manager Console (CMC), an additional set of considerations come into play when determining which reports are the most resource-intensive. The CMC sends report queries to the individual sensors (appliances), which generate the reports locally and then send the results to the CMC.

- **Number of Records** — As with reports on a single appliance that contain more than 100K records, when the CMC receives more than 100K records from multiple sensors, it sorts the records by session, identifies the duplicate records, and returns the 100K most-significant records to the query handler.
- **Connection Speed** — A report that a sensor sends over a LAN connection arrives much faster than a report sent over a slower WAN connection.

Data Enrichment

Metadata that is produced by Data Enrichment is written to the Indexing DB after additional processes such as on-box evaluations, hash submissions to a cloud (plus a delay for the verdict to return), and off-box file analysis. See Data-Enrichment Process for an explanation of how verdict data is written to the Indexing DB.

Hash Calculation

The hash reports are not populated by the DPI engine nor the metadata indexer. Hashes are calculated by the extractor under the following circumstances:

- At least one data-enrichment rule is activated — and that rule sends either a file or a file hash to one of these enrichment providers:
  - File Reputation Service
  - ICAP
  - Malware Analysis
  - Calculate and Store Hashes
  - ClamAV
  - jsunpack-n
  - YARA
  - Cuckoo
  - FireEye AX-series
  - Lastline File or Hash
  - TitaniumScale
  - VirusTotal File or Hash
- **Fuzzy Hash Only** — Fuzzy-hash reports are not populated until after you edit **/etc/solera/extractor/extractord.conf** as shown and then run **systemctl restart solera-extractord**:

```
# Flag to calculate the fuzzy hash
calc_fuzzy_hash=1   <== Uncomment this line and set the value to 1
```

- Because the hash reports contain data that is calculated after the flows are sent through the rules engine, you cannot use hash attributes as valid indicators for rules. For example, **md5_hash~93fd02e** cannot trigger a rule; however, it can be a valid primary or advanced filter.

> Enable hash calculation for *manual* extractions on *Settings > System*. (Those settings do not affect hash-related reports.)

Real-time extraction produces incomplete hash-related reports

Because the RTE is performed on matching artifacts only, the Indexing DB has hash data only on artifacts that matched a rule, regardless of whether a verdict was returned. Other artifacts have no hash-related data in the Indexing DB, so a hash-related report for a given timespan will provide the analyst with little visibility instead of full visibility.

How to calculate hashes for all artifacts

- Create a data-enrichment rule with an indicator that includes all artifacts: for example, **filename=***.
- For the data enrichment provider, select **Calculate and Store Hashes**. This provider will calculate the **MD5**, **SHA1**, **SHA256**, and fuzzy hashes (if configured) for all matching artifacts and write them to the Indexing DB.
- If you need to calculate hashes for data that has already been captured, follow these steps:
  - Go to *Capture* > *Summary*, select **Actions > Reprocess**, click **New**, and enter the timespan.
  - Click **Save** to run all of the traffic in the timespan through the rules engine and indexer again. Remember that a manual reprocessing job has a lower priority than real-time capture, so it will take longer to re-index this data than the first time it was captured.
  - After the reprocessing job is complete, run any hash-related report for the timespan. All of the missing hashes are returned by the report.

Summary Views

Because the *Summary* views permit you to see report results for several attributes at once, they can provide a fast way to see multiple data points that are associated with a single item. However, generating multiple reports simultaneously can also be costly, especially at the beginning of an investigation. Consult Report Costs for per-report information.

Number of Report Widgets per View

The report widgets on a single *Summary* view are run in parallel as soon as the view is launched, so it is recommended that you reduce the number of widgets on each view to only the ones you need for the type of investigation you are performing.

- Reports for up to 12 widgets are run simultaneously. Reports for additional widgets are queued.
- To add a widget to a view, select **Actions > Add/Edit Widgets**. Click any of the **Available Reports** to move them to the **Selected Reports** list. (Press and hold **Ctrl** to select multiple reports; click the single arrow (**>**) to move all of them at the same time.) You can also remove **Selected Reports** by moving them back to the **Available Reports** list.
- To directly delete a report widget from a view, place your cursor over the widget's header and click the **X**. Deleting a widget from one view does not affect its presence on another view.

Namespace

When the report widgets on a single *Summary* view come from different namespaces, results are produced more slowly. Consult Report Costs for the namespace of each report.

Sparse Default View

The *Default View* that arrives on a new system displays five commonly used report widgets. If this view loads too slowly because your system has a high capture volume, you should consider creating a "sparse" default view so that you do not waste time and resources generating reports that you do not want or need.

Create this sparse default view either by modifying the existing *Default View* or by creating a new view and designating it as the default. On this view, use the smallest possible number of widgets, being careful to select widgets that tend to produce the quickest results on your system.

- *Application Group* (the histogram is automatically included with the list) and *Application ID* produce results fairly quickly under most circumstances.
- If your data contains a high concentration of unique IP addresses (over 100K for most reports), avoid using the IPv4/IPv6 reports, or reduce the time range to avoid 100K entries.
- For any data enrichment provider that you do not have, delete all of those widgets from all of the views, such as from the *Threat Intel View*. For those enrichment providers that you do have, place each widget on its own view. Consult that view only after applying several other filters to reduce the amount of data to search.
- Remember to take into account each report's cost, including the namespace.
- It is entirely valid to place only one widget in the default view; however, if you need to run a report on only one attribute, Symantec recommends that you use the *Reports* page so that you can also see the histogram and full report data.

Report Costs Table

The table below shows which reports have characteristics that affect the speed at which they are produced.

- **Report Group** — Report group, as shown in the selection list on the *Menu* > *Analyze* > *Reports* page
- **Report Name** — Name of the report
- **Attribute** — Filter attribute that corresponds to the report
- **Namespace** — Namespace of the report; avoid combining reports from different namespaces in the same view.
- **Finite** — The possible unique values for this report are finite.
- **Rule** — Data is written to the Indexing DB by a data-enrichment rule.
- **Calculation** — Data is produced by a calculation after the data is retrieved from the Indexing DB.
- **External** — Data is returned by an external resource.
- **Bitmask** — Data is stored with a complete bitmap mask (8- and 16-bit values only).

> Only the default reports are displayed in this table. The user-selectable metadata on *Settings > Metadata*, have relatively low costs because all of them are extracted from packet headers by the DPI engine.

Report Group	Report Name	Attribute	Namespace	Finite	Rule	Calculation	External	Bitmask
Application	Application	**application_id**	*flows*	**X**				
	Application Group	**application_group**	*flows*	**X**				
DNS	DNS Answer Count	**dns_ancount**	*flows*	**X**				
	DNS Answer Name	**dns_name**	*flows*	**X**				
	DNS Autogenerated Domain	**autogenerated_ domain**	*flows*					
	DNS Autogenerated Domain Score	**autogenerated_ domain_score**	*flows*	**X**				
	DNS IPv4 Answer	**dns_host_ipv4_addr**	*flows*					
	DNS IPv6 Answer	**dns_host_ipv6_addr**	*flows*					
	DNS Query	**dns_query**	*flows*					
	DNS Time-to-Live	**dns_ttl**	*flows*					
	DNS Web Application Info	**dns_web_ application_info**	*flows*					
Email	Email Recipient	**email_recipient**	*flows*					
	Email Sender	**email_sender**	*flows*					
	Email Subject	**subject**	*flows*					
	Email URI	**mail_uri**	*flows*		**X**			
Encryption	SSL Certificate Serial Number	**ssl_serial_number**	*flows*					
	SSL Cipher Suite	**ssl_cipher_suite**	*flows*	**X**				
	SSL Common Name	**ssl_common_name**	*flows*					
	SSL Protocol	**ssl_protcol**	*flows*	**X**				
	TLS Heartbeat Attack Attempted	**tls_heartbeat_ attack_attempt**	*flows*					
	TLS Heartbeat Mismatch	**tls_heartbeat_ mismatch**	*flows*					

Report Group	Report Name	Attribute	Namespace	Finite	Rule	Calculation	External	Bitmask
File	Detected File Type	**file_type**	*flows*	**X**				
	File Extension	**file_extension**	*flows*	**X**				
	File Name	**filename**	*flows*					
	Fuzzy Hash	**fuzzy_hash**	*groups*		**X**			
	MD5 Hash	**md5_hash**	*groups*		**X**			
	Presented MIME Type	**mime_type**	*flows*	**X**				
	SHA1 Hash	**sha1_hash**	*groups*		**X**			
	SHA256 Hash	**sha256_hash**	*groups*		**X**			
	VoIP ID	**voip_id**	*flows*	**X**				
Geographical	Country Initiator	**country_initiator**	*flows*	**X**				
	Country Responder	**country_responder**	*flows*	**X**				

Report Group	Report Name	Attribute	Namespace	Finite	Rule	Calculation	External	Bitmask
Network Layer	Ethernet Initiator	`ethernet_initiator`	*flows*					

Report Group	Report Name	Attribute	Namespace	Finite	Rule	Calculation	External	Bitmask
	Ethernet Initiator Vendors	**ethernet_initiator_vendors**	*flows*					
	Ethernet Protocol	**ethernet_protocol**	*packets*					
	Ethernet Responder	**ethernet_responder**	*flows*					
	Ethernet Responder Vendors	**ethernet_responder_vendors**	*flows*					
	Flow Duration	**flow_duration**	*flows*					
	Flow ID	**flow_id**	*flows*	**X**				**X**
	Interface	**interface**	*flows*	**X**				
	IP Bad Checksums	**ip_bad_csums**	*flows*					
	IP Fragments	**ip_fragments**	*flows*					
	IP Protocol	**ip_protocol**	*flows*	**X**				
	IPv4 Conversation*		*flows*			**X**		
	IPv4 Initiator	**ipv4_initiator**	*flows*					
	IPv4 Port Conversation*		*flows*			**X**		
	IPv4 Responder	**ipv4_responder**	*flows*					
	IPv6 Conversation*		*flows*			**X**		
	IPv6 Initiator	**ipv6_initiator**	*flows*					
	IPv6 Port Conversation*		*flows*			**X**		
	IPv6 Responder	**ipv6_responder**	*flows*					
	Machine ID	**machine_id**	*flows*					
	Packet Length	**packet_length**	*packets*					
	Port Initiator	**port_initiator**	*flows*	**X**				
	Port Responder	**port_responder**	*flows*	**X**				
	Size in Bytes	**bytes**	*flows*					
	Size in Packets	**packets**	*flows*					
	TCP Initiator	**tcp_initiator**	*flows*					
	TCP Responder	**tcp_responder**	*flows*					

Report Group	Report Name	Attribute	Namespace	Finite	Rule	Calculation	External	Bitmask
	Tunnel Initiator	**tunnel_initiator_ip**	*flows*					
	Tunnel Responder	**tunnel_responder_ip**	*flows*					
	UDP Initiator	**udp_initiator**	*flows*	**X**				
	UDP Responder	**udp_responder**	*flows*	**X**				
	VLAN ID	**vlan_id**	*flows*					
Social Persona	Password	**password**	*flows*					
	Social Persona	**social_persona**	*flows*					
	User Name	**user_name**	*flows*				**X**	
Threat Intel	File Signature Verdict	**file_signature_ verdict**	*verdicts*	**X**	**X**		**X**	
	Local File Analysis	**local_file_ analysis_verdict**	*verdicts*	**X**	**X**			
	Malware Analysis Verdict	**malware_analysis_ verdict**	*verdicts*	**X**	**X**		**X**	
	Third-Party Verdict	**third_party_ integration_verdict**	*verdicts*	**X**	**X**		**X**	
	Threat Category	**threat_category**	*verdicts*	**X**	**X**		**X**	
	Threat Description	**threat_description**	*verdicts*		**X**		**X**	
	Threat Severity	**threat_severity**	*verdicts*	**X**	**X**		**X**	
	URL Categories	**url_categories**	*verdicts*	**X**	**X**			
	URL Risk Verdict	**url_risk_verdict**	*verdicts*	**X**	**X**			

Report Group	Report Name	Attribute	Namespace	Finite	Rule	Calculation	External	Bitmask
Web	Database Query	**database_query**	*flows*					
	HTTP and Email URIs	**uri**	*flows*		**X**			
	HTTP Code	**http_code**	*flows*	**X**				
	HTTP Content Disposition	**http_content_ disposition**	*flows*					
	HTTP Content Length	**http_content_len**	*flows*					
	HTTP Forward Address	**http_forward_addr**	*flows*					
	HTTP Location	**http_location**	*flows*					
	HTTP Method	**http_method**	*flows*	**X**				
	HTTP Server	**http_server**	*flows*					
	HTTP URI	**http_uri**	*flows*					
	Referrer	**referer**	*flows*					
	User Agent	**user_agent**	*flows*	**X**				
	Web Query	**web_query**	*flows*					
	Web Server Type	**web_server**	*flows*					

**The IPv[X] Conversations* are not written to the Indexing DB but are assembled by the query handler; therefore, the report must be newly generated for each query, unless the identical report is in the cache.

Example Searches

Consult these examples to see the best practices in finding the target data.

- Find all web sites accessed by a user
- Find a hash
  - Hash is present in the Indexing DB
  - Hash is not present in the Indexing DB
  - Finding a percentage match of a fuzzy hash
- Understand an event that occurs on a regular basis
- Find a keyword in extracted files
- Find a file name in a broad timespan
- Find a string across a data set, apply the results elsewhere

Find all web sites accessed by a user

You want to know which web sites (**domain.tld**) a user accessed during a 6-hour period. You know the user's IPv4 address.

Known Data	Timespan, user IP
Target Data	List of web sites accessed by that user during the timespan
Factors	Target data is contained in the *HTTP Server* and *HTTP URI* reports. Because *HTTP URI* is likely to have a larger number of unique results, *HTTP Server* is a less-costly report to run.

Which web sites were accessed?

With **ipv4_initiator=<user_ip>** in the primary filter bar and the six-hour timespan set, go to *Analyze* > *Summary* > *Reports* and select the *Web* > *HTTP Server* report. The list is your target data.

Further Investigation

To derive more information from your target data, you can:

- Click a web site in the list and select **View Reputation Information > [provider]** to see any known information on that site.

> Additional pivot providers can be added to the reputation list on *Settings* > *Data Enrichment* > *Third-Party Integration Providers* or by running **scm pivot_only_provider** from the CLI.(Consult the *Security Analytics8.0.x Reference Guide* on **support.symantec.com**) for instructions and for a list of suggested providers.

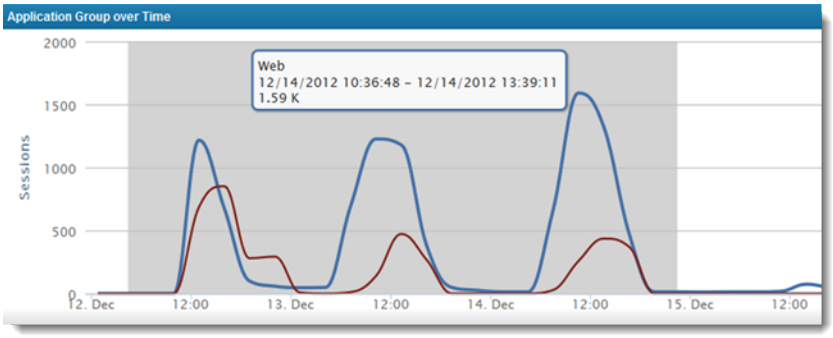
- If the Web Reputation Service was enabled during capture, you may refine the results for a particular site in question — click a web site in the list and add it to the primary filter bar as **http_server**, and then select the *Threat Intel* > *URL Categories* and *URL Risk Verdict* reports.

Understand an event that occurs on a regular basis

While viewing the *Application Group over Time* widget, you see three activity spikes in the histogram in the Web application group. You want to understand what is behind the seemingly repetitive pattern.

Known Data	Repeating traffic spikes in the *Application Group* graph
Target Data	Underlying cause of each spike
Factors	None

1. On the *Menu* > *Analyze* > *Summary* page *Application Group* histogram, click on the repetitive spikes from the same *Application Group* family. Add the *Application Group* to the primary filter.



***Application Group over Time* Histogram, Showing Three Activity Spikes**

2. Select the first spike in the *Total Sessions over Time* histogram to transfer the timespan to the timespan filter and click **Update**. Open two more browser tabs and select the second and third spikes, respectively. Then open a fourth tab and select a non-spike timespan as a baseline. Each spike spans one second.
3. By comparing the application IDs of the four timespans, you see that the **qqdownload** application drastically increases its share of the report-summary pie chart in the spikes but not in the baseline sample.
4. Add that **application_id** to the primary filter and then select the *Network Layer* > *IPv4 Initiator* report. One IP address dominates the results, so you apply the same **application_id** filter and *IPv4 Initiator* report to the other two tabs.
5. For each of the other two tabs, a different IP address dominates the report results, so you know that the event is likely not being initiated by the same user.
6. On the first tab you select the *Web* > *HTTP Server* report and see that one domain dominates the results. When you select the *HTTP Server* report on the other two tabs, the same domain dominates the results.
7. Click the HTTP server name and select **View Reputation Information > Google Search.** The Google results page opens in a new tab with the URL in the search field. The top result identifies the site as a BitTorrent repository.
8. Add the **http_server** for that server to the primary filter and select **File > File Name**. Repeat for the other two tabs.
9. The same file name is displayed at the top of the results for all three tabs.
10. In one tab, add the **filename** attribute to the primary filter bar and then select a timespan that encompasses the timespan for all three tabs.
11. When you click **Update**, the same original three spikes are displayed in the histogram. The filename is therefore your target data: the underlying cause of each spike.

12. Select one of the spikes to reset the timespan to one of the file instances, click **Extractions**, and then click **Update**.
13. On the *Extractions* page you see that the BitTorrent is about 5GB. Three users downloaded it from the same site, which explains the traffic spike.

Find a keyword in extracted files

Confidential information was leaked to the press. You want to see whether it was transported off your network, by whom, and to whom.

Known Data	Keyword
Target Data	Specific artifacts that contain the keyword Identities of sender/receiver of artifacts
Factors	Your timespan consists of the 18 hours between the time the file was created and the time it appeared on the Internet. The leaked file is an MS Word **DOCX** file, but it could have been converted to **TXT**, **PDF**, **DOC**, or **ZIP** before transport.

1. On the *Menu*  > *Summary* page, enter the 18-hour timespan. Add the following pre-loaded indicators to the primary filter bar to limit the search to all PDF, Office, and compressed files that were sent from the private network space to external networks:

   PDF - Presented MIME Type | **Archives - Presented MIME Type** | **Office Docs - Presented MIME Type** | **Archives - Detected File Type** | **RFC1918 IPv4 Initiators**

2. Click **Update**. Remember that Security Analytics will return all of the files in the flows that contain filter matches, so some of the files that are returned will not be of the specified type.
3. Review the respective *File Name*, *Application ID*, and *HTTP* reports (if applicable) to see if any files stand out for investigation. Where applicable, add the discovered filename(s) to the primary filter to further refine results.
4. Click the **Extractions** tab and wait for the extraction to complete.
5. In the Advanced Filter enter **keyword_utf8=<keyword>** and press **Enter**. When searching for multiple keywords simultaneously, click **Match All (AND)** to change it to **Match Any (OR)**, then enter **keyword_utf8=<keyword>** and press **Enter**.
6. The resulting list shows every file that contains the plaintext keyword. If the file has been compressed or encoded, as in the case of XML-based Office documents, the keyword cannot be detected because it is not available in a "cleartext" string; however, in many cases, a compressed or encoded file still contains a few plaintext strings that can be detected by Security Analytics. Furthermore, you can use application- and content-aware third-party tools to further process the data. External tools such as Google Search appliances, Agent Ransack®, and Copernic® perform optical character recognition (OCR) and can parse flat-file content. Download and pass the artifacts for investigation to the tool(s).

7. Expand an artifact entry and click **Preview**. Click the **Text** tab and use your browser's **Find** feature to see whether the keyword occurrence is significant. The keyword can also occur in the **HTTP Headers**, **File Info**, or **Strings** tabs.
8. If **Preview** does not provide enough detail into the content, download the artifact and open it with its native application, for example, MS Word or Adobe® Reader.
9. When you find the artifact that contains the leaked material, expand the artifact's entry and use the IP addresses to identify the sender and receiver.
10. If you suspect that the keyword is contained in a compressed or encoded file, create an Advanced Filter that returns the artifacts in question, for example, **file_extension=pdf** OR **file_extension=docx** OR **file_type=compressed**. Click the check box at the top of the list to select all of the artifacts and then click **Download Artifacts**. Use the tools at your disposal to parse the files for the keyword.

Find a hash

Three examples are provided for locating hashes under differing circumstances.

Scenario A

You have received an alert from an outside source that a particular **PDF** file contains a virus. You have the MD5 hash for that file, and you know that it is transported over HTTP. You want to know whether any users on your network have downloaded that **PDF** during the last 8 hours.

Known Data	The MD5 hash, the transport protocol
Target Data	Users that downloaded the infected file
Factors	The **md5_hash** attribute has been populated in the Indexing DB for all PDFs because a rule that detects PDFs uses the **Calculate and Store Hashes** provider.

1. On the *Menu* > *Summary* page, input **application_id=http** and **md5_hash=<hash>** (or **md5_hash~<part of hash>**) to the primary filter bar and set the timespan.
2. Click the **Reports** tab and click **Update**.
3. Select the *Network Layer* > *IPv4 Initiator* report. The list of IP addresses is your target data.

Scenario B

Your third-party malware-detection system alerted on a file that appeared to have malicious characteristics. You have the SHA1 hash for that file and a timestamp from the malware-detection system but no other information. You want to know whether the file is actually malicious and where it came from.

Known Data	The SHA1 hash

Target Data	The file name, file type, file reputation, and point of ingress
Factors	You do not have any active data-enrichment rules that would populate the **sha1_hash** attribute in the Indexing DB.

1. On the *Menu*  > *Summary* view, input a timespan to cover a short time before and after the time of the alert; use your judgment to estimate the difference between your malware-detection system alerting and Security Analytics capturing the same file. A span of a few seconds is ideal.
    - **Recommended** — Apply a primary filter that limits the search to a known application-transfer mechanism such as HTTP or FTP.
2. Click the **Extractions** tab and wait for the extraction to complete.
3. Input **sha1=<hash>** to the Advanced Filter. If no result is produced, expand the timespan by a few minutes in both directions and run the extraction again.
    - If required, you can expand the time frame to cover a much larger window. Try to limit the artifacts extracted to a manageable number, usually less than 100,000.
4. When you get a result, expand the artifact entry. The entry displays the filename, file type, URI hostname, and the IP address that accessed it.
5. Click **Reputation**. The **Reputation Information** dialog displays any known information on that file.
6. If the artifact was downloaded over HTTP, add the **Original URL** value to the primary filter bar as **http_uri**. If the transfer mechanism was something other than HTTP, put the source port number as the primary filter.
7. Click the **Reports** tab, broaden the timespan, select the **Network Layer > IPv4 Initiator** report, and click **Update**.
8. The total sessions over time chart shows all of the other times when the file was accessed, and the results list shows the IP address of the user who accessed the file.

Scenario C

After identifying the malware file in Scenario B, you want to know whether variants of the same file exist on the network.

Known Data	An extracted artifact on Security Analytics of the original malware file.
Target Data	Any files that match the original file with 80% confidence.
Factors	Fuzzy-hash calculations and data-enrichment rules were disabled at the time of capture.

1. To enable fuzzy hashes for reports, edit **etc/solera/extractor/extractord.conf** as shown and then run**systemctl restart solera-extractord**:

```
# Flag to calculate the fuzzy hash
calc_fuzzy_hash=1  <== Uncomment this line and set the value to 1
```

2. Locate the original malware file on the *Extractions* page and expand the artifact entry. The **Fuzzy Hash** attribute is missing. Make a note of the detected MIME type, date/time, and source port for the artifact.
3. Open a second tab and enable fuzzy-hash calculation on **Settings > System**. This setting affects only the *Extractions* page.
4. Create a data-enrichment rule that triggers hash calculations:
   - Select **Menu**  **> Settings > Data Enrichment** and under *Local File Analysis Providers*, select **Calculate and Store Hashes**.
   - Select **Menu**  **> Analyze > Indicators** and click **New**. Create a new indicator with **file_type=<detected_file_type>** or **port=<source port>** or both as the filter.
   - Select **Menu**  **> Analyze > Rules** and click **New**. Create a new rule with the new indicator; select **Data Enrichment** for **Type**, and for **Send to**, select **Calculate and Store Hashes**.
5. Open a third tab and select **Menu**  **> Capture > Summary**.
6. Select **Actions > Reprocess** and click **New**. Select a reasonable time range that includes the original artifact in the investigation. For example, if you are concerned only with files that arrived in the past 6 hours, do not specify the last 24 hours of traffic.
7. When the job reaches **100% Percent Complete** return to the tab with the *Extractions* page.
8. Adjust the timespan by a second and click **Update** to run another extraction.
9. Locate the malware artifact entry and expand it. The **Fuzzy Hash** attribute is displayed.
10. Add the **Fuzzy Hash** to the filter bar. Change the filter to **fuzzy_hash>=<fuzzy hash>%80**.
11. Click the **Reports** tab and select the **File > File Name** report.
12. The results are your target data: a list of files with a 80% match or greater.

Find a filename in a broad timespan

A malware-tracking website alerts you to a file that was just discovered but that has been in the wild, undetected, for three weeks. You need to know whether that file has crossed your network.

Known Data	The name of the file: **gotchernose.js**

Target Data	Any instance of the file.
Factors	■ The timespan is three weeks. ■ You have six sensors in four remote locations: ◦ During those three weeks they've captured a total of 200 terabytes. ◦ The WAN links on three of the sensors may add latency to report retrieval. ■ Efficiencies in data retrieval can be had with smart filtering versus running a simple **filename=<filename>** query on the three-week timespan over six sensors.

1. Select one sensor to test the query before applying it to all sensors at once.
2. If possible, identify transfer mechanisms of interest for the file, such as HTTP or SMB. Apply as many filters as possible to reduce the amount of data to search without missing possible matches. For example, if you have reason to believe it was transferred over port 80 with HTTP only to certain network segments, apply a filter such as **tcp_port=80**, **application_id=http**, **ipv4_address!=<subnet_of_non-interest>**, **filename=gotchernose.js**
3. On the test sensor, click the **Reports** tab and select the *Application Reports* > *Application Group* report, because it has a limited number of possible values.
4. Apply the filter, set a one-hour timespan, and click **Update**. Note how long it takes to produce the report, even if there is no data in it. Use this information to estimate how long a query with a broader timespan, more sensors, and target data might take.
   - If the query returns the target file, you can explore the file's metadata using other reports to see whether anything is unique enough to exclude more data, such as a unique HTTP server. When you are satisfied that you have created a query that excludes as much data as possible, return to the CMC and run the query on all six sensors.
   - If desired, divide the three-week query into six-hour timespans or smaller, depending on how quickly you can produce relevant data each time to further narrow the search.

Find a string across a data set, apply results elsewhere

You have received an alert about a JavaScript exploit that downloads malware from the Internet. The script contains a unique string, **12fwo08n79**, but you suspect that the file name and extension are deceptive. You want to find the bad **JS** file's name, find out how it got into the system, and see whether it has downloaded anything.

Known Data	String, timespan
Target Data	**JS** file containing string, downloaders of the **JS** file, instances of downloads.
Factors	■ JavaScript code is not obscured. ■ You need to locate the string without extracting all of the artifacts from the entire timespan.

1. Initiate an SSH session with the appliance and log in as **root**.
2. Navigate to the timespan using **pfs/flows/**:

   ```
   cd /pfs/flows/timespan/2017-07-13T20:38:33-06:00_2017-07-13T20:39:05-06:00
   ```

3. Perform an **ngrep** query for the string:

   ```
   [root@hostname 2017-07-13T20:38:33-06:00_2017-07-13T20:39:05-06:00]# ngrep -q '12fwo08n79'
   -l -I data.pcap
   ```

4. The results show four pairs of IP addresses that are associated with the **JS** file. The internal addresses are the **JS** file's downloaders — the entrance point to your network.
5. To locate and extract the target **JS** file, select one of the IP pairs and make a note of the IP address and initiator port number, which is likely unique.
6. On the *Menu*  > *Analyze* > *Summary* page of the web UI, input **port=<initiator_port>**, **ipv4_address=<ip_address>**, click the **Reports** tab, select the **File Reports > File Name** report, set the original timespan, and click **Update**.
7. If the list contains too many extraneous files, use different **mime_type** filters to remove extraneous content — for example **mime_type!~pdf –** to the primary filter. Recall that filters return flows that contain a matched indicator and it may not be possible to completely remove all extraneous content without removing the corresponding files of interest as well.
8. When the list of files is manageable, click the **Extractions** tab and wait for the extraction to finish.
9. In the **Advanced Filter**, type **keyword_utf8=12fwo08n79**. The result should be your target **JS** file.
10. Expand the **JS** artifact's entry, click **Preview** and then select the **Text** tab. For **Syntax Highlighting**, select **JavaScript Formatted**. (If the script is fairly long, the **Syntax Highlighting** list may not appear for a few seconds.)
11. As necessary, use your browser's **Find** function to locate the target string, or parse the code to determine which file(s) it downloads from the Internet. You can also click the **jsunpack-n** tab to see if any malicious properties were detected.