QGrep Documentation

Created by Joris Bouwsma [X], last updated by Michael Rus on Jan 30, 2026 • 15 minute read

This page provides an overview of how to use the QGrep Self Service Interface, and some important things to keep in mind when working with QGrep.

-

What is QGrep?

QGrep is a tool that allows internal Akamai users to query raw log data from the DDC log archive. It provides access to FreeFlow and ESSL log data received by DDC from ghost. This allows internal Akamai users to retrieve log data for troubleshooting, research, and customer requests.

To initiate a retrieval request, the user fills out the Make a new retrieval request form. They specify the start time and duration for the log lines of interest and other filter criteria such as a list of cp codes. Clicking submit will create a Qgrep batch retrieval job that will find the relevant ghost log files for the query and filter them to return the log lines that match the query.

Qgrep Self Service Interface

View request results | Make a new retrieval request | Help

Enter the filter criteria for the log records desired.

Note: hover over the row names for help

Query Description	mrus's qgrep query
Network	FF (includes ESSL)
Log Start Time	Start date should be after GMT Mar 15 2023 23:21 (logs are available for 14 days back in time) Mar 28 2023 Time: 00 : 00 GMT, for the next 1 hours and 0 minutes
CP Codes	
Ghost IP's	
Country	All / Non-EU only / EU only / Specific country codes
Prepend Ghost IP	Prepend ghost ip to each log line / Don't prepend
Max logs to filter	Your query limits are 200000 logs per query, with up to 200000 per 2 hour time window. You have filtered 0 logs in the last 2 hour period and have a remaining quota of 200000. 100000 ← You can increase this default value up to 200000 with your remaining quota
Result mode	Retrieve data (but abort if the query would exceed max logs) Retrieve data (and allow result truncation if the query exceeds max logs) Calculate the number of logs only
Restrict by record type	v w r f i l p d m t b s a h k R S F
Log filter clauses	
Field names	
Cluster name	NSS1 3978 (Paris, FR)

Submit Reset

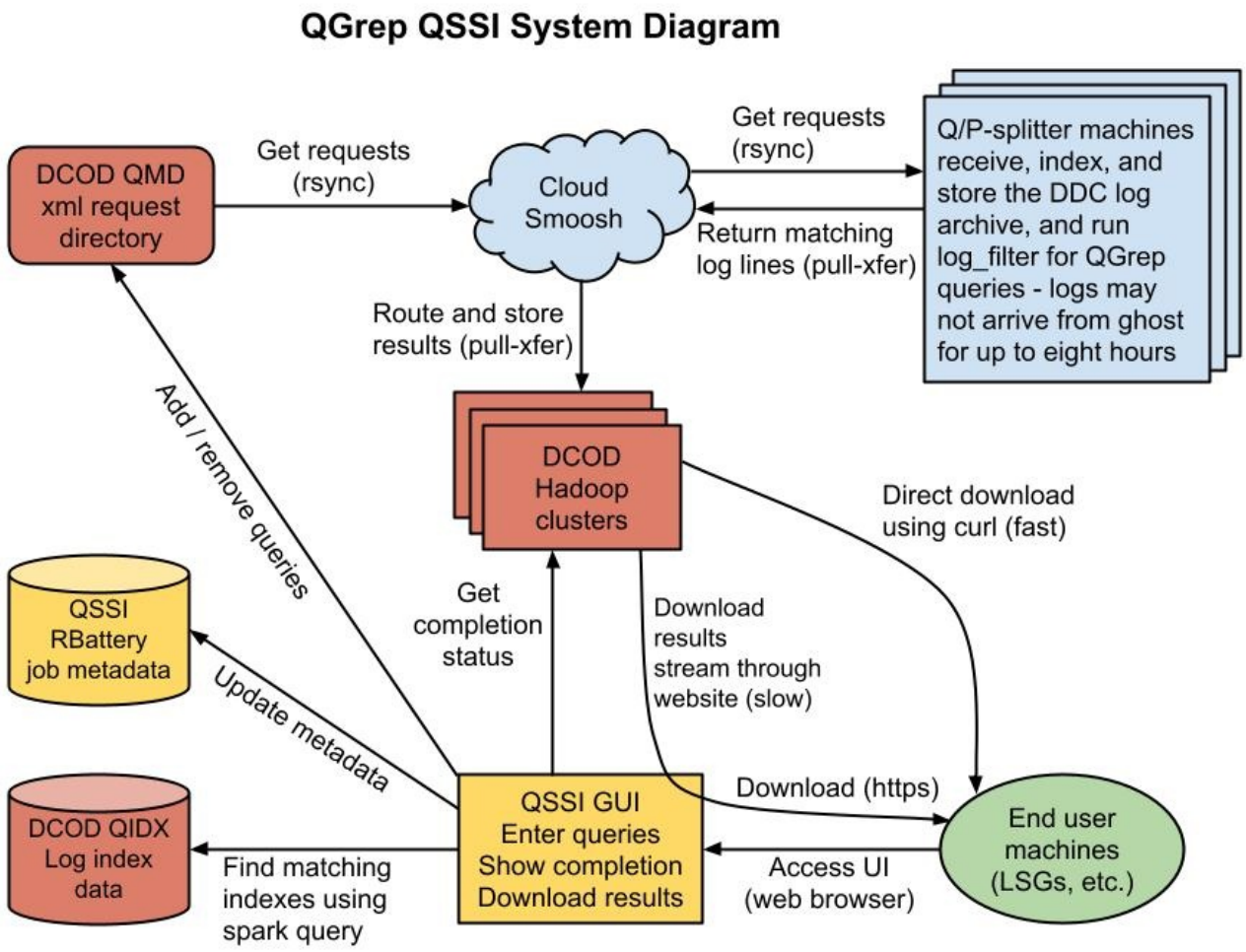
QGrep Log Archive Retrieval Overview

The DDC (Distributed Data Collection) network receives edge log files for processing into billing and other pipelines. A copy of the received logs is preserved in a distributed archive, with the individual logs spread over thousands of DDC machines. The QGrep interface provides access to this log data via ad hoc queries. See the system diagram below.

QGrep queries are run as distributed batch processes on the DDC network. The first step uses an indexing system to identify log files that contain data that match the specified criteria, such as cp code, time range, and ghost IP. The machines hosting the identified logs are then directed to run the log_filter utility to extract matching log lines. The extracted data is collected and transported to a Hadoop cluster, where it can be further analyzed using map/reduce subqueries, or downloaded by a user to their local machine or into Pliny.

Typical retrieval time for QGrep requests is about 12 minutes, depending on the current DDC network load and the size of the results.

The archive currently uses a replication factor of 1, so log files hosted on DDC machines which are out of service, will not be accessible. Also some machines may be too busy with other work and not return results. Retrieval rates of matching log data will typically exceed 90%, with the percentage slowly degrading the further removed the time range is from the current day.



What Networks does QGrep Support?

QGrep can fetch data from FF and ESSL access logs, i.e. ghost.ddc.log

How to Access the QGrep Interface

In a web browser running within the Akamai network, go to https://qssi.akamai.com

Along the top of the screen are three main links:

View request results | Make a new retrieval request | Help

The "**View request results**" link will bring up a list of submitted queries, with the newest queries shown first. Each entry shows the progress and details of the query, along with links for available actions such as previewing results or creating a subquery. If you only want to see your own queries, put your username in the "Filter username:" box.

The "**Make a new retrieval request**" link brings up a form for submitting a new QGrep query by filling out the filter criteria.

The "**Help**" link brings up the support page. This page also shows the current limits on queries, which are adjusted from time to time.

Note that **access to cache.log and hats.log** has recently been added. For information on that, go to here. The number of ghost IPs you can query at a time has been increased to 1024.

How to Enter a QGrep Query

Click on the "Make a new retrieval request" link and enter the filter criteria for the log lines you want to retrieve. Use the tab key to move between fields or click on the fields you wish to edit. When ready, click on the submit button to add the query to the request queue (or you can press the "enter" key when editing a text field to submit the form).

Most of the filter criteria fields are self-explanatory, but below are some points to consider for completing the fields. **Note that if you hover your mouse pointer over the row names in the form, you will see flyover help for filling out the field.**

- **Query Description:** A generic default description is filled in, but you can change it to something more meaningful so you can identify what the query was for.
- **Log Start Time:** Enter the start UTC time for the log lines of interest. The end time is specified as the duration from the start time. The duration is currently limited to a 24 hours, so if you need a longer duration, you can enter multiple queries. To query for less than one hour, set the hours to 0 and set the minutes to the number of minutes you want to scan. Data should be available in QGrep about 15 minutes after being received by the DDC network. If you want to retrieve the most recent data you can set the duration to extend into the future and you will get as much recent data as possible. Note that some ghosts are slow to send data to DDC, so when full data is needed, it is best to wait several hours after the target time range before querying it.
- **CP Codes:** if you want data for multiple CP Codes in one result set and for the same time range, then you can enter multiple CP Codes. Otherwise you could split the cp codes into separate queries.
- **GHost IP's (optional):** enter the numeric dotted IP address of the ghosts rather than machine names. If you are looking for a region or a virtual ghost IP, you will need to look up the list of actual IPs that correspond to the region or virtual IP in the network configuration file (i.e. //sysgrp/main/configuration/configuration) and then enter that list of actual IPs.
- **Country:** If you want to filter based on the country the ghost is located in you can select the desired filtering. If you want to scan only logs from non-EU countries, select Non-EU, for EU only, select EU only, and if you have specific countries in mind, enter their country code numbers in the "Specific country codes" box. Clicking on "codes" will open a pop-up window showing the country code numbers you can enter.
- **Prepend GHost IP:** This controls is the log lines will have the ghost ip address pre-pended to each log line or not. If you select the option to prepend the ghost IP to log lines, bear in mind that subqueries you run on the results will have to be designed to deal with the ghost IP. For example, if you try to run a subquery using log_filter on the results, it will be confused by the prepended ghost IP. This can be handled by supplying '-a ghost_ip' to log_filter. Equivalently, it modifies log_filter config by adding ghost_ip to the front of the existing log format. For instance, log_filter -c /a/etc/netmgmt.log_filter.conf -f ddc -a ghost_ip "rt,unix_date,cpcode,c_ip". Note: The "Log filter clauses" input field in QGrep's GUI **won't** be confused by the prepended ghost IP.
- **Max logs to filter:** this option allows you to limit the number of ghost log files you want to scan for your query. Each log file contains up to one hour's worth of log lines, with busy ghosts having only minutes per log file. Note that "max logs to filter" is **not** a limit on the number of ***log lines*** you can retrieve, but only on the number of ***log files*** that can be filtered. If you only need partial results rather than all matching data, you could set this value to a smaller number than the default and use the "allow truncate" retrieval mode. This would make your query run somewhat faster, the result set would be smaller, and you would use less of your log limit quota. You can also increase the default value up to the maximum logs allowed per query, provided you have sufficient quota available. Some tips to reduce the number of logs that need to be filtered for a query are to reduce the time range of the query, reduce the number of cp codes in the query, filter by country, and/or add ghost ip restrictions to the query. See Qgrep Max Logs to filter limits for more details.
- **Result mode:**
  - *Retrieve data (but abort if the query would exceed max logs)* - This is the default retrieval mode. It protects against running a query where the results would be truncated. When the query is submitted, QGrep will check if the number of logs that have relevant data exceeds the value for "Max logs to filter". If it does, then the query will be aborted and you can then resubmit it with a larger value for "Max logs to filter" so that you will get complete results. The aborted query will not count against your quota usage.
  - *Retrieve data (and allow result truncation if the query exceeds max logs)* - In this retrieval mode, QGrep will not abort a query where the number of relevant logs exceeds the "Max logs to filter". It will truncate the number of logs to scan to the "Max logs to filter" value and return the partial query results.
  - *Calculate the number of logs only* - this option lets you preview the number of ghost logs (i.e. indexes) that would be filtered for your query without actually running the query. It is sometimes a good idea to check this if you are unsure of the volume for a cp code before you run your actual retrieval so that you have an idea of how large the results may be, or if you will use up all of your quota with the query. The preview will also show you how many of the matching logs are from the EU countries, and give you a rough estimate of the number of log lines that match your query.
- **Restrict by record type:** The checkboxes for record types can be useful for reducing the size of your results. If you only care about 'r' records for example, then you would check the 'r' box and this will save data transport time and storage space in the Hadoop cluster. If you're not sure which records you care about, then you can leave all the checkboxes blank and you will get all record types. You can later run subqueries on the results to reduce the data if needed. For a description of record types, see the Ghost Log Format Specification.
- **Log filter clauses:** The "Log filter clauses" field is for advanced users that are familiar with the syntax for the log_filter utility (see log_filter). You can enter log filter clauses here to reduce the size of your results, which may be useful for high volume cp codes, or if you are looking for something specific such as log lines that match a certain URL pattern. For example, you might enter "arl_full =~ /akamai\.com/i" without the quotes in the text box and your results would only contain log lines that matched the specified arl. You can enter multiple clauses (one per line) and they will be ANDed together. If you want to specify an "or" condition, you must enclose the expression in parenthesis like this: (c_status eq '200' or obj_size > 1000). Use caution when filling out the log filter clauses and field name filters since if you make a syntax mistake or misspell a field name, you may get an empty result set since no log records will match. If you have an LSG account, you can run log_filter there to test your filter clauses.
- **Field names:** The "field names" field allows you to limit the results to certain log line fields, such as entering "c_ip, arl_full" without the quotes. Use caution when filling out the log filter clauses and field name filters since if you make a syntax mistake or misspell a field name, you may get an empty result set since no log records will match. The field names used by QGrep can be found on LSGs and production machines in "/a/etc/netmgmt.log_filter.conf". The netmgmt.log_filter.conf is pushed from UMP, and can be downloaded here. All LSGs have the log_filter program available in /a/bin. See the Ghost Log Format Specification for a detailed description of the fields in FF/ESSL log files.
- **Cluster name:** This specifies which destination Hadoop cluster your results will be sent to. This value will be set for you automatically, but you can override it if for some reason you wanted your results to go to a specific cluster. If you download results using curl, you may get faster download speeds by selecting a cluster that is nearest to where you are downloading to.

EU Data Localization

Due to GDPR data privacy laws, queries that contain EU data must be sent to a storage cluster located in the EU. Most QGrep queries contain some data from the EU, so most queries will be treated as an EU query. The primary QGrep storage clusters have been migrated to the EU to comply with the GDPR rules. The data shown in the results preview screen for queries containing EU data will have the low byte of IPV4 addresses anonymized by masking them to 0 (but note that the data will not be anonymized when results are downloaded). In the unlikely event that there are no EU storage clusters available at the time a query is submitted, QGrep will skip scanning the EU logs in the query and only scan logs located outside of the EU and send the U.S. results to a U.S. based storage cluster. You can later get the missing EU data by re-running the query once an EU cluster becomes available, and selecting "EU only" for the ghost country option on the query form.

How to retrieve log lines for ChinaCDN cpcodes

China CDN queries should be done using the global cp codes directly in the "CP Codes" box. Previously QGrep indexed logs based on the bill_5mi pipeline, so it indexed the ChinaCDN cp codes, and the global cp codes needed to be put in the "Log Filter Clauses" box. That method will no longer work since the billing ChinaCDN codes are no longer indexed. So ChinaCDN queries should be done using only the global cp codes that appear in the raw log lines. You can optionally enter the China country code of 47 in the "Specific country codes" filter box.

Limitations

QGrep is only able to retrieve log lines that have been fully processed and indexed by DDC. QGrep indexes only log lines that contain a cp code.

It can only retrieve log lines that log_filter is able to extract. For example, no 'W' websocket log lines can be retrieved.

Log types are limited to FF and ESSL access logs, i.e. ghost.ddc.log. Other log types are not supported.

Quota limitations are discussed below in Understanding Quotas and Retention Limits.

Tracking the Progress of Queries

Once you submit a query you will automatically be taken to the results page, or you can click on the "View request results" link at the top of the page. Here you will see a list of queries, both active and completed.

The "Status" column will show the state of each query, such as indexing, queued, active, or done. The "Logs completed" column will give an estimate of the percentage of logs that have been filtered for the query so far. Note that the percentage is only for the logs that QGrep found when the query was submitted. Additional delayed logs may arrive from ghost after that time that are not included in the query.

It will take several minutes to get the first results from your query once it becomes active, with the bulk of the results arriving over the next several minutes and then gradually trailing off. You can preview and download the current accumulated results at any time using the links in the "Actions Available" column. Once you have enough data for your needs, you can click on the "Abort request" link to stop the query. This will reduce the load on the DDC network and allow other requests to proceed more quickly. The query will also stop automatically after a time limit is reached.

The query results will not always reach 100% completion since machines hosting some of the logs may be unavailable, but completion rates over 95% are typical.

How to Access Results

If you click on the "Preview results" link in the "Actions Available" column, you will see up to the first 1000 lines of results. You can copy and paste these results into a log parser (such as the engr log parser or the gss log parser) or other application. There is a "Copy to clipboard" button to simplify this.

If you click on the "Download archive" link in the "Actions Available" column, you can download the entire results received to that point and save them in a .gz compressed archive. To access the contents you can use zcat or any utility capable of decompressing Linux gzip files. The download archive link streams the data through app-battery, so may be inefficient and slow for large result sets.

For large file downloads, you can use the "Download via curl" link. This option allows you to download the data directly from the Hadoop cluster, and may be considerably faster. It is also an easy way to download the data to an LSG machine or Pliny.

How to Re-use a Previous Query

On the results page you will see a "Description" column which has a link next to the "ID #" for the query. If you click on that link, it will show the details of the query. Below the details will be a button "Edit as a new query". If you click on that it will pre-populate a new retrieval request form with the values originally entered for the query. You can then edit the description and change the values as desired and submit it as a new query.

Understanding Quotas and Retention Limits

Since the volume of data per hour can be massive for high volume cp codes, queries to the archive will be limited by time range and the number of ghost log files that can be filtered. This is to minimize the impact on normal DDC processing of billing data. The limits are are adjusted from time to time based on capacity and other considerations. The current limits are shown on the QGrep Help page.

Each user has a quota on how many log files they can filter during a sliding time window. This is intended to spread out the QGrep requests during the day to prevent spikes in resource utilization. On the "Make a new retrieval request" form, it will show you the current amount of quota you have remaining in the "Max logs to filter" section. This section is also color coded to indicate when your quota is running low.

You will usually not know how many log files a query will need to scan until you try running it. The system will try to protect you against accidentally using up all of your quota in a single query. See "Result mode" and "Max logs to filter" in the How to Enter a QGrep Query section. If you submit a query and see a status message saying "**Query aborted, Needed to filter N logs**", it means the default value (or number you specified) on the query form for "Max logs to filter" was too small to filter all the matching log files for your query. If you want all the available data and have sufficient quota available (i.e. more than N), resubmit the query and specify a value for "Max logs to filter" that is larger than N. If you prefer to get partial results and not use up as much of your quota, then resubmit the query using "Result mode" *Retrieve data (and allow result truncation if the query exceeds max logs)*.

The limits on the QGrep Help page have the following meanings:

- **Maximum logs filtered per query N (default M)** - N is the maximum number of log files that can be filtered by one query. You can set the request field "Max logs to filter" up to this value provided you have enough remaining quota. The default value M for "Max logs to filter" is lower than the maximum so that users don't use up all their quota by accident.
- **Maximum logs filtered per X hour period: Y** - The value for Y is the limit on how many logs you are allowed to filter using a sliding window over the last X hours (i.e. your allowed quota). If you do not have enough quota left to run a new query, you can come back and retry after some of your recent queries have become older than X hours.
- **Maximum concurrent indexing queries** - This is the limit on how many queries you can have in the indexing state at the same time.
- **Maximum results size per query** - This is the limit on how much data will be stored for your query results. Data beyond this size will be discarded. If you hit this limit but need all the data, try splitting your query into multiple queries with different time ranges. Usually when you hit this limit it is not practical to actually download it, so you may want to add more filters to your query to reduce the result size.
- **Maximum cp codes per query** - This is the limit on how many cp codes you can specify for a single query.
- **Maximum ghost ip's per query** - This is the approximate limit on the number of ghost ips you can specify in one query.
- **QGrep Archive retention** - This controls how far back in time the start time for your query can be. It is related to how long ghost logs are retained in the DDC log archive.
- **Results retention** - This is how long your results will be downloadable from QGrep before they are purged.

If you cancel a query, only the logs that had completed filtering prior to cancelling will be counted against your quota.

If you are in a situation where you are querying a high volume cp code, and are unable to get all the data needed due to your quota not being large enough, it is possible to request a temporary increase in your quota limit. See QGrep Support for details.

Need Help?

Please see the QGrep Support page for Frequently Asked Questions, and instructions on how to get support.

Links

Ghost Log Format Specification

log_filter

QGrep UI

QGrep Support

Data Service Catalog - Log Archive

No labels

11 Comments

Christian Mortensen

Jul 08, 2019

I want to test out my log_filter script on my own machine as proposed on this page. But how do I get the log_filter script? It is not installed on my Akamai Linux Desktop.

Maurice Cinquini
Mar 24, 2022

All LSG's (e.g. lsg-east.akamai.com) have the log_filter executable in your $PATH, and the updated /a/etc/netmgmt.log_filter.conf config file.

Jake Soder
Sep 23, 2021

I have talked to a number of other Qgrep users about how to reduce their load on the system and/or stay within the max logs to filter limits on individual qgrep queries. I have drafted a proposed addition/amendment to this documentation related to this. It is located in this google doc:

Monica Loria
Mar 22, 2022

Hi! I noticed an error message when trying to extract logs prior to 14 days. Aren't the logs available for 45 days?

Thanks! cc @Bryan Leong [X]

Maurice Cinquini
Mar 24, 2022

The QGrep self-service UI has been limited to the prior 14 days since Friday, March 4, 2022. This is in preparation for the actual reduction of the DDC archive to that period to save $7.5 M per year. BTW, before Qgrep the retention period used be only 10 days with LDS troubleshooting log spinning - and then only if the cpcode was requested in advance.

Suyeon Kim
May 13, 2022

Hi team,

Regarding the limitations below, how much time does it typically takes for DDC to fully process and index log lines?

Limitations
QGrep is only able to retrieve log lines that have been fully processed and indexed by DDC. QGrep indexes only data for traffic that has passed Access Control (i.e. billable traffic).

Michael Rus
Mar 29, 2023

With the new Breeze-based indexing pipeline, log data will be queryable in QGrep about 15 minutes after being received by the DDC network. Some distant ghosts may be slow to send log data successfully to DDC, so it can take several hours to receive log data from some ghosts.

So when maximum data completeness is needed, it is best to wait several hours before querying a time period.

If the newest data is desired, you can set your end query time in the future and this will maximize the amount of recent data that QGrep can retrieve, but bear in mind not all ghosts will have sent their data to DDC for very recent time periods, so you will be getting partial results from a subset of ghosts.

Anna Blasiak
Jul 08, 2025

There are many more columns in the 'r' and 'S' lines than listed in netmgmt.log_filter.conf

How would we get the later columns added?

Maurice Cinquini
Jul 08, 2025

I believe the ghost team is responsible for updating this when new fields are added (needed for nsh as well as qgrep) and then pushing it out via UMP. https://ump.akamai.com/ump/#ump_sidebar=netmgmt.log_filter.conf%2Crequest%2Chistory shows when it was last updated (2021-05-14!) and by who. I found more (perhaps outdated) info in Making changes to netmgmt.log_filter.conf

Zhanhao Zhang [X]

Jul 08, 2025

Are we able to retrieve columns added in Ghost 9 (e.g. Connection ID in r80 and S63) by entering it in the "Field names" in QGrep? The Connection ID column doesn't seem to be provided in https://docs.akamai.com/ops/metadata/p4-1681/metadata/ump/netmgmt.log_filter.conf

Maurice Cinquini

Jul 09, 2025

Yes, you will get all fields ghost logs by default.
If you only want to get some fields including ones that aren't in netmgmt.log_filter.conf please ask the ghost team to add those new fields as I mentioned to Anna above. The Ghost team needs to be doing that whenever they add new fields.

ATLASSIAN