

GCS访问日志配置和查询

Author: eugeneyu@google.com

[概述](#)

[一、配置和使用GCS审计日志](#)

[二、配置和使用GCS Usage Logs](#)

[三、将GCS Usage Logs导入到BigQuery进行结构化存储和查询](#)

[附录](#)

概述

Google Cloud Storage (GCS) 是谷歌云的对象存储，适用于长期存储视频、图片、日志等非结构化内容文件。如果涉及到海量用户内容上传和下载，那么追踪对象访问记录，进行统计分析和错误排查，会是一个常见的需求。这时可以利用GCS的访问日志来满足需求。下面介绍开通和使用GCS访问日志的一些方法。

一、配置和使用GCS审计日志

谷歌云很多产品都提供审计日志，帮助用户查询“谁，在何地，做了什么”的问题。GCS的审计日志包含Admin Activity Logs和Data Access Logs。我们可以用Data Access Logs审计日志来查看不同用户对GCS上对象文件的访问操作。

为了节省用户费用，GCS的审计日志默认是关闭的。可以在控制台IAM菜单的Audit Logs查看其状态，并将其打开。

IAM & Admin

Audit Logs DEFAULT AUDIT CONFIG

storage

<input type="checkbox"/>	Properties	Admin Read	Data Read	Data Write	Exemptions
<input type="checkbox"/>	Google Cloud Storage	—	—	—	0
<input type="checkbox"/>	AI Platform Notebooks	—	—	—	0
<input type="checkbox"/>	Apigee	—	—	—	0
<input type="checkbox"/>	Apigee Connect API	—	—	—	0
<input type="checkbox"/>	Certificate Authority Service	—	—	—	0
<input type="checkbox"/>	Cloud AI Platform API	—	—	—	0
<input type="checkbox"/>	Cloud API Gateway API	—	—	—	0
<input type="checkbox"/>	Cloud Asset API	—	—	—	0
<input type="checkbox"/>	Cloud AutoML API	—	—	—	0
<input type="checkbox"/>	Cloud Billing API	—	—	—	0
<input type="checkbox"/>	Cloud Build API	—	—	—	0
<input type="checkbox"/>	Cloud Composer API	—	—	—	0
<input type="checkbox"/>	Cloud Data Loss Prevention (DLP) API	—	—	—	0
<input type="checkbox"/>	Cloud Dataproc API	—	—	—	0
<input type="checkbox"/>	Cloud Datastore API	—	—	—	0
<input type="checkbox"/>	Cloud Domains API	—	—	—	0
<input type="checkbox"/>	Cloud Functions API	—	—	—	0
<input type="checkbox"/>	Cloud Healthcare	—	—	—	0
<input type="checkbox"/>	Cloud Identity-Aware Proxy API	—	—	—	0
<input type="checkbox"/>	Cloud IoT API	—	—	—	0
<input type="checkbox"/>	Cloud Key Management Service (KMS) API	—	—	—	0
<input type="checkbox"/>	Cloud Life Sciences API	—	—	—	0

如果Google Cloud Storage的各个审计日志项都是“-”状态，说明没有打开。可以选中后在右侧配置面板打开各项日志，并保存配置。

IAM & Admin

Audit Logs DEFAULT AUDIT CONFIG

Google Cloud Storage Filter table

<input checked="" type="checkbox"/>	Title ↑	Admin Read	Data Read	Data Write	Exemptions
<input checked="" type="checkbox"/>	Google Cloud Storage	—	—	—	0

Google Cloud Storage

LOG TYPE EXEMPTED USERS

Turn on/off audit logging for selected services.

- ☒ Admin Read
- ☒ Admin Write
- ☒ Data Read
- ☒ Data Write

SAVE

审计日志打开后，其状态应该如下所示。

Audit Logs

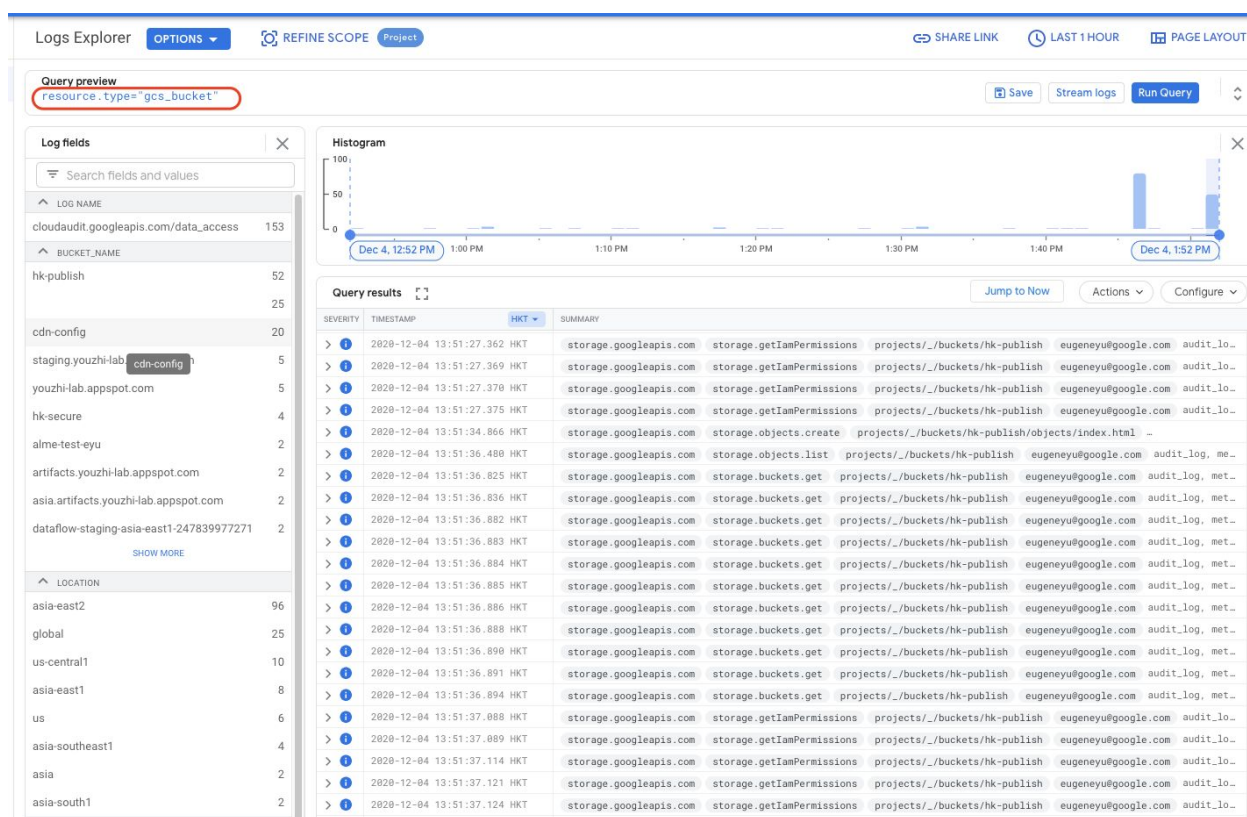
DEFAULT AUDIT CONFIG

Google Cloud Storage

Filter table

<input checked="" type="checkbox"/>	Title ↑	Admin Read	Data Read	Data Write	Exemptions
<input checked="" type="checkbox"/>	Google Cloud Storage	✓	✓	✓	0

审计日志打开后，对象的访问日志会实时导入到谷歌云日志服务。可以在日志服务的Log Explorer中进行查看。注意在查询表达式栏中输入'resource.type="gcs_bucket"'。



在列出的对象访问日志中，可以查看到每次访问的详细信息。比如展开一个访问记录，可以看到访问的操作是创建或下载，访问的用户账号，访问的对象URI，访问的结果等等。



需要注意的是，Storage Audit Log 审计日志开启后，可以看到授权实名用户访问的记录。但是**看不到**匿名用户的访问记录，和通过谷歌云CDN的访问记录。

比如用下面两种方式访问对象，不会产生日志。

1. 直接用对象链接下载

```
$ curl https://storage.googleapis.com/hk-publish/index.html
```

2. 通过CDN下载

```
$ curl http://35.244.150.103/index.html
```

如果需要查看匿名用户的访问记录，那么需要使用GCS的Usage Logs。

二、配置和使用GCS Usage Logs

GCS的Usage Logs不像审计日志一样自动导入到日志服务，而是会导出到对象存储。所以首先需要创建一个存放日志的对象存储桶。

```
gsutil mb -l ASIA-EAST2 gs://youzhi-lab-logs-bucket
```

给Usage Logs服务的服务账号提供写访问日志存储桶的权限。

```
gsutil iam ch group:cloud-storage-analytics@google.com:legacyBucketWriter
gs://youzhi-lab-logs-bucket
```

打开目标桶的Usage Logs，并将其日志输出桶设为刚建好的日志存储桶。同时也可以设置一个日志文件前缀，相当于目录。

```
gsutil logging set on -b gs://youzhi-lab-logs-bucket -o
hk-publish-access-log/ gs://hk-publish
```

确认目标桶的日志已经打开。

```
gsutil logging get gs://hk-publish
```

应该看到类似如下输出。

```
{"logBucket": "youzhi-lab-logs-bucket", "logObjectPrefix": "hk-publish-access-log/"}
```

日志打开之后，GCS日志服务会每小时一次把目标桶的访问日志输出到日志存储桶，生成日志文件。可以等1-2小时后到日志存储桶查看。

youzhi-lab-logs-bucket






OBJECTS CONFIGURATION PERMISSIONS RETENTION LIFECYCLE

Buckets > youzhi-lab-logs-bucket > hk-publish-access-log

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DOWNLOAD DELETE

Sort and filter Filter Filter objects and folders

For faster performance, select Filter by name prefix only from the filtering menu.

		Size	Type	Created time		Storage class
<input type="checkbox"/>	 _usage_2020_12_07_02_00_00_04a229c0feedd60cb4_v0					
<input type="checkbox"/>	 _usage_2020_12_07_02_00_00_04a229c0feedd60cb4	459 B	application/octet-stream	Dec 7, 2020, 11:09:...		Standard
<input type="checkbox"/>	 _usage_2020_12_07_02_00_00_06a229c0feedd60cb4	968 B	application/octet-stream	Dec 7, 2020, 11:10:...		Standard
<input type="checkbox"/>	 _usage_2020_12_07_02_00_00_01a229c0feedd60cb4	522 B	application/octet-stream	Dec 7, 2020, 11:10:...		Standard
<input type="checkbox"/>	 _usage_2020_12_07_02_00_00_07a229c0feedd60cb4	504 B	application/octet-stream	Dec 7, 2020, 11:10:...		Standard

但是手工下载和查看日志存储桶中的众多文件非常耗时，也不方便。我们建议将日志再转存到谷歌云数据仓库BigQuery，方便使用SQL进行查询，也可以接入其它数据分析工具和展示报表。

三、将GCS Usage Logs导入到BigQuery进行结构化存储和查询

首先创建一个BigQuery数据集，用来存放GCS日志。下面操作使用BigQuery的命令行工具。但也可以在控制台操作。

命令行工具的安装使用可以参考此[文档](#)。

```
bq mk --location asia-east2 storage_logs
```

下面要用官方提供的schema文件创建一个数据表，并导入一个日志文件。可以到日志文件存储桶，拷贝一个Usage Logs日志文件的URI。

←	Object details	↓ DOWNLOAD	✎ EDIT METADATA	👤 EDIT PERMISSIONS	🗑 DELETE
Buckets > youzhi-lab-logs-bucket > hk-publish-access-log > _usage_2020_12_07_02_00_00_04a229c0feedd60cb4_v0 📄					
Public access	Not public				
Type	application/octet-stream				
Size	459 B				
Created	Dec 7, 2020, 11:09:37 AM				
Last modified	Dec 7, 2020, 11:09:37 AM				
Hold status	None ✎				
Retention policy	None				
Encryption type	Google-managed key				
Custom time	—				
Public URL	Not applicable				
Authenticated URL	https://storage.cloud.google.com/youzhi-lab-logs-bucket/hk-publish-access-log/_usage_2020_12_07_02_00_00_04a229c0feedd60cb4_v0 📄				
URI	gs://youzhi-lab-logs-bucket/hk-publish-access-log/_usage_2020_12_07_02_00_00_04a229c0feedd60cb4_v0 📄				

下载Schema文件并导入一个日志文件。导入任务会根据Schema文件自动创建一个表。

```
wget http://storage.googleapis.com/pub/cloud_storage_usage_schema_v0.json

bq load --skip_leading_rows=1 storage_logs.usage \
gs://youzhi-lab-logs-bucket/hk-publish-access-log/_usage_2020_12_07_02_00_00_04a229c0feedd60cb4_v0 \
./cloud_storage_usage_schema_v0.json
```

导入完成后，可以到BigQuery控制台查看表中的数据。

① FEATURES & INFO SHORTCUT HIDE PREVIEW FEATURES

Explorer + ADD DATA

Q Type to search ?

Viewing pinned projects.

- youzhi-lab
 - youzhi_lab
 - youzhi_lab_billing
 - storage_logs
 - usage

usage QUERY TABLE ASK QUESTION SHARE TABLE COPY TABLE DELETE

Schema Details Preview

Row	time_micros	c_ip	c_ip_type	c_ip_region	cs_method	cs_uri	sc_status	cs_bytes	sc_bytes	time_taken_micros	cs_host	cs_referer	cs_user_age
1	1607309296899495	185.142.236.35	1		GET	/robots.txt	404	0	127	1127000	34.102.222.41		

上述操作成功后，删除导入的日志文件，并导入剩余的其它日志文件。

```
gsutil rm
gs://youzhi-lab-logs-bucket/hk-publish-access-log/_usage_2020_12_07_02_00_0
0_04a229c0feedd60cb4_v0

bq load --skip_leading_rows=1 storage_logs.usage \
"gs://youzhi-lab-logs-bucket/hk-publish-access-log/_usage_2020*" \
./cloud_storage_usage_schema_v0.json
```

导入完成后会看到如下提示。

```
[eugeneyu:~/Downloads]$ bq load --skip_leading_rows=1 storage_logs.usage \
"gs://youzhi-lab-logs-bucket/hk-publish-access-log/_usage_2020*" \
./cloud_storage_usage_schema_v0.json
Waiting on bqjob_r23001ebc1e29531a_0000017641d32584_1 ... (0s) Current status: DONE
```

可以在BigQuery中用对象名搜索特定对象的访问记录列表。

<div> <div>RUN</div> <div>SAVE</div> <div>SCHEDULE</div> <div>MORE</div> </div>													
1 SELECT * FROM `youzhi-lab.storage_logs.usage` WHERE cs_uri like '%index.html' LIMIT 1000													
<div> <div>Query results</div> <div>SAVE RESULTS</div> <div>EXPLORE DATA</div> </div>													
<div> <div>Query complete (0.3 sec elapsed, 49.1 KB processed)</div> <div> <div>Job information</div> <div>Results</div> <div>JSON</div> <div>Execution details</div> </div> </div>													
Row	time_micros	c_ip	c_ip_type	c_ip_region	cs_method	cs_uri	sc_status	cs_bytes	sc_bytes	time_taken_micros	cs_host	cs_referer	cs_user_agent
1	1607308171139341	203.208.61.81	1		GET	/hk-publish/index.html	200	0	6	199000	storage.googleapis.com		curl/7.64.1,gzip(gfe)
2	1607308173861389	203.208.61.81	1		GET	/hk-publish/index.html	200	0	576	16000	storage.googleapis.com		curl/7.64.1,gzip(gfe)
3	1607308172614681	203.208.61.81	1		GET	/hk-publish/index.html	200	0	576	15000	storage.googleapis.com		curl/7.64.1,gzip(gfe)

以上是用命令行对存量日志做批量导入。对增量日志，也可以用定时导入任务来自动化导入。


首先删除之前已经导入过的日志。


```


gsutil rm "gs://youzhi-lab-logs-bucket/hk-publish-access-log/_usage_2020*"


```


然后在BigQuery的Data Transfers菜单创建一个导入任务。进行如下配置。注意选择的频率是每小时一次——这是自动任务可以执行的最高频率。另外，写入模式为Append，追加写入。


 BigQuery


 Data transfers

 Scheduled queries

 Reservations

 BI Engine

 Data QnA

 BigQuery

Create transfer


Source type

Choose a data source from the list below

Source *

Google Cloud Storage

EXPLORE DATA SOURCES

 This is the Google Cloud Storage configuration. [Learn more](#)

Transfer config name

Display name *

load_gcs_usage

Schedule options

☒ Start now ☐ Start at set time

Repeats *

Custom


Custom Schedule *

every 1 hours

Start date and run time

12/9/20, 10:22 AM

HKT



Destination settings

Select the destination for the transfer data


Dataset ID *

storage_logs

Data source details

Destination table *


usage



Cloud Storage URI *


☒ youzhi-lab-logs-bucket/hk-publish-access-log/_usage_2020*

BROWSE



Write preference

APPEND



<|

BigQuery

SQL workspace

Data transfers

Scheduled queries

Reservations

BI Engine

Data QnA

BigQuery

Create transfer

Write preference
APPEND

☒ Delete source files after transfer

File format *
CSV

Transfer Options

All Formats

Number of errors allowed
0

JSON, CSV

☐ Ignore unknown values

AVRO

☐ Use avro logical types

CSV

Field delimiter
,

Header rows to skip
1

☐ Allow quoted newlines

☐ Allow jagged rows

Refresh window

Notification options

☒ Email notifications
When enabled, the transfer administrator will receive e-mail notifications on transfer run failures.

定时导入任务配置完后， 点击保存。等待几小时， 可以看到完成的任务状态。

BigQuery	BigQuery	Transfer details	RUN HISTORY	CONFIGURATION
SQL workspace	load_gcs_usage			
Data transfers	Schedule (UTC) Target date for next run			
Scheduled queries	every 1 hours December 9, 2020 at 11:37:00 AM UTC+8			
Reservations	Filter transfer runs			
BI Engine				
Data QnA				
		Run date	Schedule time	Summary
		December 9, 2020	December 9, 2020 at 1:37:00 PM UTC+8	The transfer run has completed successfully.
		December 9, 2020	December 9, 2020 at 12:37:00 PM UTC+8	The transfer run has completed successfully.
		December 9, 2020	December 9, 2020 at 11:37:00 AM UTC+8	The transfer run has completed successfully.
		December 9, 2020	December 9, 2020 at 10:37:00 AM UTC+8	The transfer run has completed successfully.

之后，可以在BigQuery里随时查看GCS的对象访问记录，也可以用程序执行查询，根据结果触发后续操作。

不过需要注意，通过gsutil工具，或者控制台做的上传、修改、下载等操作，由于属于审计日志范畴，**不会在**Usage Log里记录。这些日志需要通过第一节所介绍的审计日志查看方式来查询。

附录

- [1] [Cloud Audit Logs with Cloud Storage](#)
- [2] [Usage logs & storage logs](#)
- [3] [Using the bq command-line tool](#)