ALIBABA CLOUD

阿里云

数据湖分析

CDN

文档版本：20210611

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.读取CDN日志数据 ---------- 05

1.读取CDN日志数据

阿里云内容分发网络（Content Delivery Network，CDN）是建立并覆盖在承载网之上，由分布在不同区域的边缘节点服务器群组成的分布式网络。CDN分担源站压力，避免网络拥塞，确保在不同区域、不同场景下加速网站内容的分发，提高资源访问速度。您可以将CDN日志转存到OSS，或者将CDN日志配置实时日志推送，然后进行日志分析。接下来您就可以在不移动日志文件的情况下通过DLA分析OSS中的CDN日志数据，及时发现问题，并有针对性的解决问题，提升CDN服务质量。

前提条件

通过DLA分析CDN日志数据之前，您需要先完成以下准备工作：

- **函数计算**

  什么是函数计算

- OSS
  i. 开始使用OSS
  ii. 创建存储空间
  iii. 上传文件

  > **说明** 根据业务需求，判断是否需要新建文件夹，将CDN转存过来的日志存储在新建文件夹中。

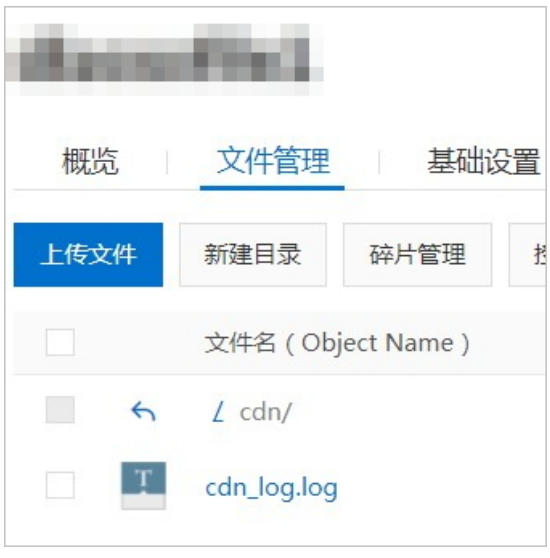
- DLA
  i. 开通云原生数据湖分析服务
  ii. 重置数据库账号密码

步骤一：将CDN日志转存到OSS

将CDN日志转存到OSS，详情请参见日志转存。

步骤二：创建OSS Schema

假设OSS Bucket中存储了以下CDN日志文件。



您可以在DMS页面编写SQL创建OSS Schema、创建日志文件表以及读取日志文件数据，也可以通过MySQL客户端或者MySQL命令行工具连接DLA，然后编写SQL创建OSS Schema、创建日志文件表以及读取日志文件数据。

```
CREATE SCHEMA cdn_log_schema with DBPROPERTIES(
  catalog='oss',
 location = 'oss://bucket-name/cdn/'
 );
```

步骤三：创建日志文件表

CDN日志文件数据如下所示。

```
[9/Jun/2015:01:58:09 +0800] 188.165.15.75 - 1542 "-" "GET http://www.aliyun.com/index.html" 200 191 2830 MISS "Mozilla/5.0 (compatible; AhrefsBot/5.0; +http://ahrefs.com/robot/)" "text/html"
```

日志字段含义如下表所示。

字段	说明
[9/Jun/2015:01:58:09 +0800]	时间
188.165.15.75	访问IP地址
-	代理IP
1542	responsetime（单位：ms）
"-"	referer
"GET http://www.aliyun.com/index.html"	request
200	httpcode
191	requestsize（单位：byte）
2830	responsesize（单位：byte）
MISS	cache命中状态
"Mozilla/5.0 (compatible; AhrefsBot/5.0; +http://ahrefs.com/robot/)"	UA头
"text/html"	文件类型

在DLA中建表时，需要采用正则表达式对CDN日志进行解析，每个字段和对应的正则表达式如下所示。

字段	说明
[9/Jun/2015:01:58:09 +0800]	(-|\\[[^\\]]*\\])
188.165.15.75	([^]*)

-	([^]*)
1542	([^]*)
"-"	([^ \"]*|\"[^\"]*\")
"GET http://www.aliyun.com/index.html"	([^ \"]*|\"[^\"]*\")
200	(-|[0-9]*)
191	(-|[0-9]*)
2830	(-|[0-9]*)
MISS	([^]*)
"Mozilla/5.0 (compatible; AhrefsBot/5.0; +http://ahrefs.com/robot/)"	([^ \"]*|\"[^\"]*\")
"text/html"	([^ \"]*|\"[^\"]*\")

完整的正则表达式为：`(-|\\[[^\\]]*\\]) ([^ ]*) ([^ ]*) ([^ ]*) ([^ \"]*|\"[^\"]*\") ([^ \"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*) (-|[0-9]*) ([^ ]*) ([^ \"]*|\"[^\"]*\") ([^ \"]*|\"[^\"]*\")`。

DLA中通过以下SQL创建日志文件表。

```
CREATE EXTERNAL TABLE cdn_log (
 log_timestamp       VARCHAR,
 access_ip           VARCHAR,
 proxy_ip            VARCHAR,
 response_time       VARCHAR,
 referer             VARCHAR,
 request             VARCHAR,
 httpcode            SMALLINT,
 request_size        BIGINT,
 response_size       BIGINT,
 cache_hit_status    VARCHAR,
 ua_header           VARCHAR,
 file_type           VARCHAR
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
"input.regex" = "(-|\\[[^\\]]*\\]) ([^ ]*) ([^ ]*) ([^ ]*) ([^ \"]*|\"[^\"]*\") ([^ \"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*) (-|[0-9]*) ([^ ]*) ([^ \"]*|\"[^\"]*\") ([^ \"]*|\"[^\"]*\")"
)
STORED AS TEXTFILE
LOCATION 'oss://bucket-name/cdn/'
TBLPROPERTIES ('recursive.directories' = 'true');
```

步骤四：查询、分析CDN日志

```
SELECT * FROM cdn_log;
+-----------------------------+--------------+----------+--------------+---------+--------------------------------+----------+----------
----+--------------+-----------------+-------------------------------------------------------------------+-------------------------+
| log_timestamp               | access_ip    | proxy_ip | response_time | referer | request                        | httpcode | req
uest_size | response_size | cache_hit_status | ua_header                                  | file_type               |
+-----------------------------+--------------+----------+--------------+---------+--------------------------------+----------+----------
----+--------------+-----------------+-------------------------------------------------------------------+-------------------------+
| [18/Jun/2019:05:08:33 +0800] | 47.92.115.203 | -      | 777          | "-"     |                                |    200 |      201 |      7159
| MISS             | "Mozilla/5.0 (Linux; Android 4.1.1; Nexus 7 Build/JRO03D))" | "text/html;charset=UTF-8" |
+-----------------------------+--------------+----------+--------------+---------+--------------------------------+----------+----------
----+--------------+-----------------+-------------------------------------------------------------------+-------------------------+
```

可以使用DLA系统函数，对数据进行分析。

```
SELECT date_parse(substring(log_timestamp, 2, length(log_timestamp) - 8), '%d/%b/%Y:%H:%i:%s') as log_
timestamp,
    access_ip,
    proxy_ip,
    response_time,
    substring(referer, 2, length(referer) - 2) as referer,
    substring(request, 2, length(request) - 2) as request,
    httpcode,
    request_size,
    response_size,
    cache_hit_status,
    substring(ua_header, 2, length(ua_header) - 2) as ua_header,
    substring(file_type, 2, length(file_type) - 2) as file_type
FROM cdn_log;
+-------------------------+--------------+----------+--------------+---------+--------------------------------+----------+-------------+--
-------------+-----------------+-------------------------------------------------------------------+-------------------------+
| log_timestamp           | access_ip    | proxy_ip | response_time | referer | request                        | httpcode | reques
t_size | response_size | cache_hit_status | ua_header                                  | file_type               |
+-------------------------+--------------+----------+--------------+---------+--------------------------------+----------+-------------+--
-------------+-----------------+-------------------------------------------------------------------+-------------------------+
| 2019-06-18 05:08:33.000 | 47.92.115.203 | -      | 777          | -       |                                |    200 |      201 |      7159 | MISS
| Mozilla/5.0 (Linux; Android 4.1.1; Nexus 7 Build/JRO03D)) | text/html;charset=UTF-8 |
+-------------------------+--------------+----------+--------------+---------+--------------------------------+----------+-------------+--
-------------+-----------------+-------------------------------------------------------------------+-------------------------+
```