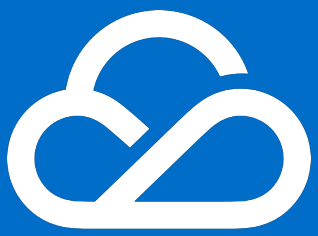
腾讯云数据工坊

快速入门

产品文档



腾讯云

【版权声明】

【商标声明】

【服务声明】

本文档意在向客户介绍腾讯云全部或部分产品、服务的当时的整体概况，部分产品、服务的内容可能有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或模式的承诺或保证。

文档目录

案例背景

Web 日志分析

某移动应用开发商在腾讯云 CDB MySQL 已经存放了大量 Web 用户日志数据。用户希望通过周期性的 Hive 工具对数据进行分析，并将分析结果，存放 CDB 数据格式已经预先经过结构化处理，按照不同字段存放。

希望通过 Hive 工具，计算日志中不同出错类型HTTP返回值（http_code）的数量。

准备工作

配置所用信息

配置所有信息如下图所示：

实体	名称	备注
CDB MySQL实例	game_analysis	数据库实例
CDB MySQL数据库	game	数据库名称
CDB MySQL日志原始表	game_log	原始日志
CDB MySQL分析结果表	game_result	内容通过导入任务，由TDF分析结果表导入
TDF工程名称	game_log	工程中包括表、工作流和任务
TDF数据源服务器标识	*****_game_log_raw	TDF中定义的逻辑实体，关联至CDB MySQL实例game_analysis；前缀*号系统自动生成
TDF数据源服务器说明	game_log原数据	同*****_game_log_raw一一对应的说明
TDF日志原始表	game_tdf_raw	为game_log原始表的子集，仅选取部分字段
TDF分析结果表	game_tdf_dest	存放分析结果，包含http_code ，count字段

说明：上图中为该示例所用到的表以及其他逻辑实体。

注册腾讯云账号并开启 TDF 服务

首先，请登录 https://www.qcloud.com/，完成注册流程。腾讯云使用 QQ 账号体系，如为企业开发商，建议使用专用非私人账户注册。如用户已经使用腾讯云产品。

在产品介绍页面 https://www.qcloud.com/product/tdf，提交 TDF 开通申请，开通后，可登录腾讯云-控制台，在云产品中选择数据工坊 TDF。

准备日志原始数据

假设日志已经存放在CDB MySQL中，实例名为 game_analysis。

示意如下图所示：

可通过数据库客户端软件，查询当前原始数据的数据库格式，包括字段等，并记录。在本场景中，日志原始数据样本有数千条，单条记录示意如下，存放在腾讯云CDB MySQL数据库实例game_analysis的game_log表中：

20160626002522 115.153.150.199 tdf.qcloud.com /tdftest_21_FotAejgyFBMfNWJu-5KNjyf7axH-.jpg?imageView2/1/w/1080/h/1920&e=1466783400&token=Q-hCY0VbL4F6NTX3TgRvE_T3vcpNEo2Gr3S9RA-b:yrrQjzCHZH2luqA-ONED8IODBjY= 185801 1465 2206 NULL 702 "AndroidDownloadManager/5.1.1 (Linux; U; Android 5.1.1; tdftest R3 A Build/LAT4AV)"

- 日志字段内容说明：

request-time、user-Agent、access_time、ip、host、request_uri、size、province、isp、http_code、referer.

其中字段内容已经做了处理，仅作示意。

您可在如下链接下载到上述测试数据至本地，并通过您自己熟悉的数据库客户端软件导入到 game_log 表项中。

另，准备空表用于存放结果，名为 game_result 在本例子中需分析不同http_code对应的记录计数。

结构如下图所示：

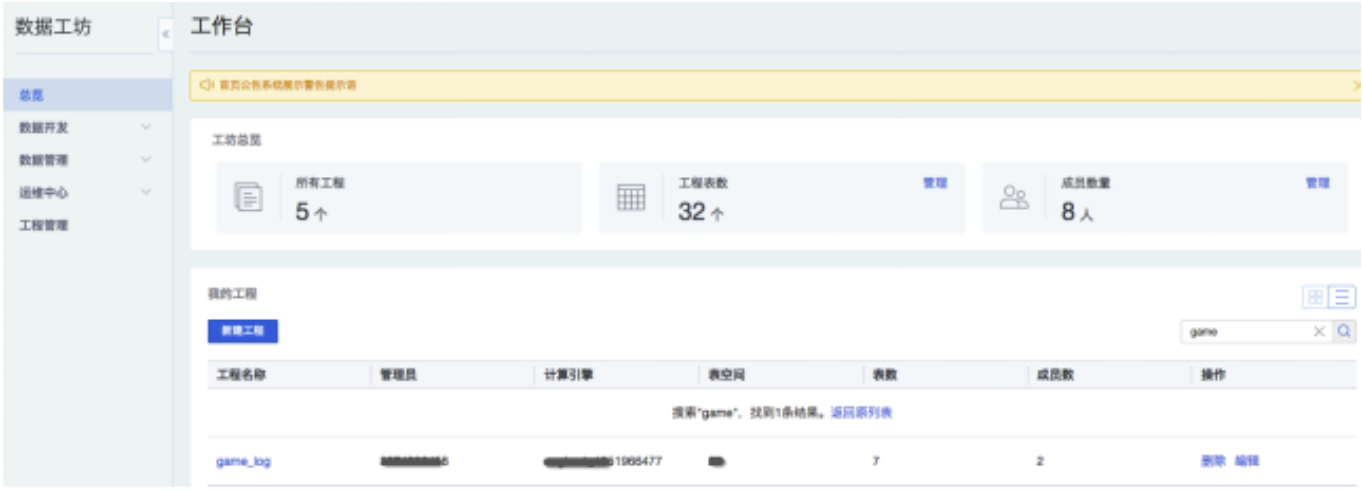
Field	Type	Null	Key	Default	Extra
http_code	int(11)	NO		NULL	
count	double	NO		NULL	

详细配置步骤

第一步：创建工程和配置基本信息

登录TDF后，首先进入总览页面。如果是初次进入，则无工程和表信息。
单击创建工程，如下图，输入工程名（例如game_log）。工程标识前缀为系统唯一标识符自动生成。工程管理员默认为用户本人；计算引擎为系统默认设置。

创建成功后，可在总览视图看到所创建的工程基本信息,如下图所示：



第二步：创建数据表

数据表可理解为数据处理过程中在 TDF 所存放的位置。

在本示例中，需要用到两个表：原始数据表（用于存放从 CDB MySQL 导入的数据，例如game_tdf_raw），以及结果表（用于存放分析后的结果，例如game_tdf_dest），过程如下：

创建原始数据表

点击管理控制台，在【数据管理–表管理】点击屏幕右侧【新建表】按钮，开始创建数据表。填写表名game_tdf_raw，选择刚刚创建的工程game_log。

在高级设置中，选择表的记录格式和对应文件格式，按照默认设置即可。

点击下一步，设置表中的字段。本例中不使用分区，在本例子中，我们只打算分析用户接入日志中不同 http_code 类型的数量，因此，仅选择如下8个字段进行分析：

access_time,host,http_code,ip,isp,province,request_time,request_uri

初始编辑界面类似如下图所示：

创建后，可在【数据管理-表管理】中，查询表结构，如下：

创建结果表

点击管理控制台，在【数据管理 –表管理】点击屏幕右侧【新建表】按钮，开始创建数据表。

填写表名 game_tdf_dest，选择刚刚创建的工程 game_log。

在高级设置中，选择表的记录格式和对应文件格式，按照默认设置即可，

点击下一步，设置表中的字段。在本例子中，我们只打算记录不
http_code类型的数量结果，因此，仅选择如下两个个字段且不使用分区：http_code，count

创建后，可在【数据管理-表管理】中，查询表结构，如下图所示：

此结构需同最终计算完毕后，拟用于存放结果的CDB MySQL表game_result结构相同。

第三步：设置数据源

在【工程管理-数据源管理】，点击【+新建配置】。

选择服务器类型：云数据库CDB-MySQL，编辑服务器标识（前缀由系统自动生成，自行输入部分例如【game_log_raw】），和服务器说明（例如【game_log原数据】），责任人默认为当前用户，下拉选择CDB数据库实例【game_analysis】（之前在CDB配置），下拉选择数据库名称【game】（之前在CDB配置），并输入对应用户名密码，连接后，即完成数据库配置。

第四步：创建工作流

在【数据开发-工作流开发】界面，下拉选择刚刚创建的工程【game_log】，然后选择新创建工作流列表。

第五步：创建数据接入任务

拖拽任务节点【数据接入-CDB MySQL导入Hive】，并单击右键编辑，基本信息配置，任务名称命名为【导入game_log任务】，类型默认为【CDB MySQL导入Hive】，责任人为默认当前用户，输入输出配置。

源数据库：选择之前创建的后缀为【game_log_raw】数据库。

源数据库查询SQL语句：例如如下语句

select access_time,host,http_code,ip,isp,province,reques
t_time,request_uri from game_log。

即仅检索以上8个字段，顺序和字段名严格按照CDB MySQL的顺序和字段名。

目标表所在工程：选择当前TDF工程game_log

目标表：选择之前建立的TDF表game_tdf_raw

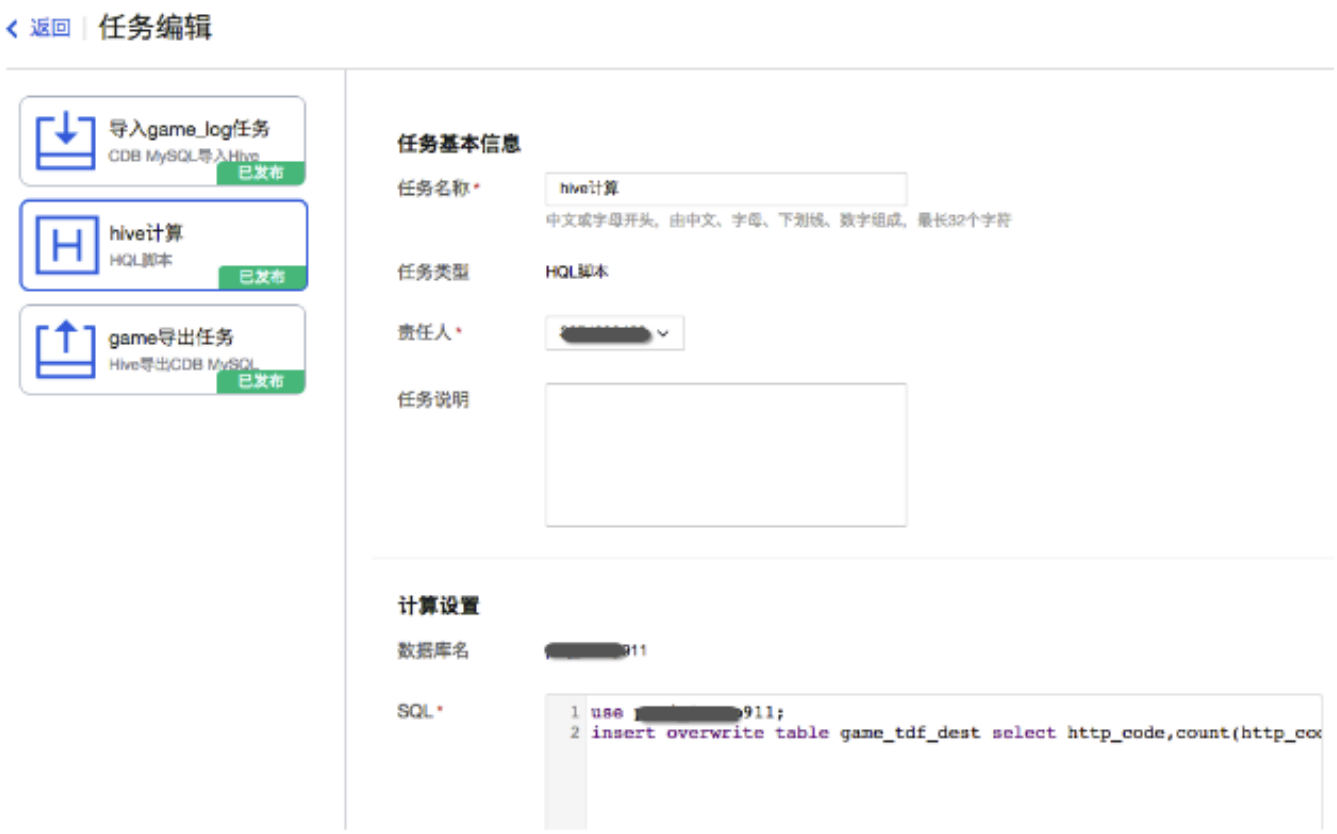
目标表列名映射：

access_time,host,http_code,ip,isp,province,request_time,request_uri

分区格式选择【无分区】，其余设置按照默认值。

第六步：创建 HQL 计算任务

该任务为计算日志中不同的 http_code 对应的记录数量。任务脚本参数如下图所示：



其中，SQL脚本代码如下：

use 数据库名称;
insert overwrite table game_tdf_dest select http_code,count(http_code) from game_tdf_raw group by http_code。

其中数据库名称需填入文本框上方的数据库名。此语句将TDF日志原始表中的记录按照http_code不同类别计数，并将结果存放回TDF分析结果表game_tdf_dest。

第七步：创建数据导出任务

拖拽任务节点【数据导出-Hive导出CDB MySQL】，并单击右键编辑。
基本信息配置，任务名称命名为【game导出任务】，类型默认为【Hive导出CDB MySQL】，责任人为默认当前用户。

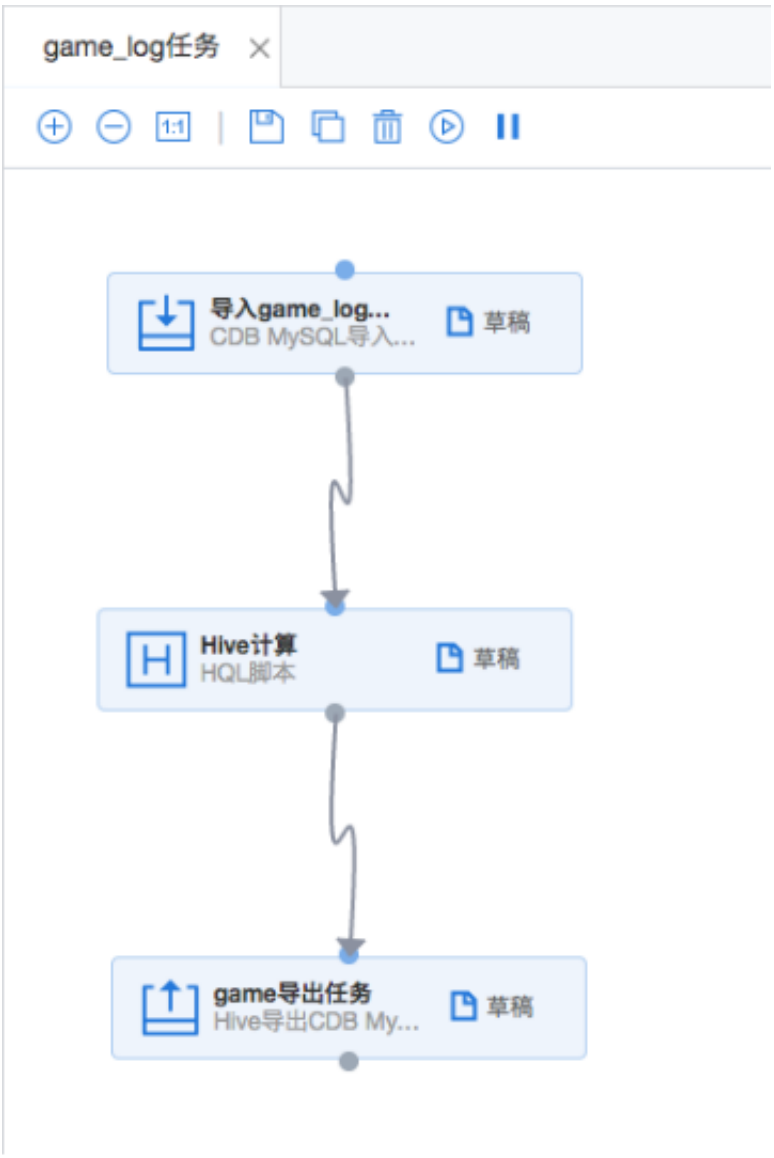
输入输出配置中，选择源数据库所在工程【game_log】，查询语句为select * from game_tdf_dest，即将上一个任务节点Hive计算得到的结果表game_tdf_dest内容全部回写到CDB

MySQL，下拉选择目标数据库以及目标表（即CDB已有数据库和目标表），目标表列名映射为：

http_code,count，其余按照默认值即可。即如下配置：

第八步：保存工作流并发布

任务节点编辑完毕后，检查节点见的依赖关系，如下图连线可见。



点击保存，和发布任务按钮，系统将提示确认调度信息，如下，选择单次执行。

发布后，在运维中心的任务列表视图可以按照工程维度查看当前的工作流的实时状态。当工作流状态切换为成功后，任务即全部运行成功（根据数据量和系统负载不同，任务运行时间将会略有区别）。

第九步：分析结果展示

按照前序设定，查询结果将首先写入TDF表，随后再写入CDB对应数据库。当工作流状态成功后，可使用数据库管理客户端，对CDB表进行查询。例如，执行完如下语句 select * from game_result 后，系统显示结果如下：

http_code	count
200	2166
206	160
304	1
405	4

即在所有日志中，http_code不同类型类型对应的计数结果如上。

也可以选择通过可视化报表软件，将最终结果进行图形化展示。