

数据集成 实践教程





【版权声明】

©2013-2025 腾讯云版权所有

本文档(含所有文字、数据、图片等内容)完整的著作权归腾讯云计算(北京)有限责任公司单独所有,未经腾讯云事先明确书面许可,任何主体不得以任何形式 复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯,腾讯云将依法采取措施追究法律责任。

【商标声明】

🔗 腾讯云

及其它腾讯云服务相关的商标均为腾讯云计算(北京)有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标,依法由权利人所有。未经腾讯云及有关 权利人书面许可,任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为,否则将构成对腾讯云及有关权利人商标权的侵犯,腾讯云将依 法采取措施追究法律责任。

【服务声明】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况,部分产品、服务的内容可能不时有所调整。 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默示的承 诺或保证。

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务,及相应的技术售后服务,任何问题请联系 4009100100或95716。





文档目录

实践教程

DLC 数据实时导入与小文件合并 MySQL 分库分表同步至 Hive SQL Server 数据离线同步至 Hive 分区 MySQL + gh-ost 实时同步至 Kafka 将 MySQL 数据同步至 Doris 概述 源端 MySQL 准备 目标端 Doris 准备 配置 DataInlong 项目空间及集成资源 配置单表实时同步任务 配置整库实时迁移任务 任务运维 常见问题

实践教程 DLC 数据实时导入与小文件合并

最近更新时间:2025-06-0410:24:12

业务场景

通过 DataInLong 数据集成将业务数据源实时导入至 DLC iceberg 表的过程中,伴随着实时同步过程的推进,目标系统端会不断生成小文件。对于目标系统 内已生成的小文件,基于周期合并的方式可避免由于小文件的累积造成目标系统 DLC 引擎查询效率恶化。

操作场景

本文以 MySQL 实时同步至 DLC iceberg 表为例,介绍实时任务配置及小文件合并操作实践。

操作步骤

创建目标表

进入 DLC 控制台,根据以下语句创建 DLC 原生表(内表), DLC 内表默认为 iceberg 表。详细创建 DLC 原生表属性及数据优化,详情请参见 数据优化 。



配置项目空间

```
① 说明:
若您使用的是 WeData 产品,配置项目空间操作请参见 项目列表。
```

1. 进入 DataInLong 控制台,单击**项目列表 > 新建**,新建项目空间。

DataInLong	项目列表 🕲 北京 🔻	新建项目空间	×
器 概 览	新建	项目名称•	
白 集成资源		項目标识・	
■ 项目列表	未配置集成资源	描述	
	項目标识:		
	进入空间		
	绑定集成资源	▲高级设置	
		項目成员 ① 请选择	*
	未配置集成资源	成员角色 ① □ 项目管理员 □ 数据工程师 □ 运维工程师 □ 访客	
	项目标识:		
	进入空间 绑定集成资源		

2. 您可以参考下表配置项目空间信息。

参数	说明
项目名称/标识	项目命名与唯一标识,其中唯一标识创建后不可修改。



高级设置 – 项目成员	为此创建的项目中添加其他项目成员,创建者默认加入项目空间。
成员角色	批量为项目成员配置角色(此处默认为前面添加的成员添加统一的角色,后续可项目管理模块修改)。

配置集成资源组

1. 进入 DataInLong 控制台选择集成资源并单击创建,进入集成资源组购买页。

DataInLong	集成资源 ◎北京 ·							
言 概范	0538						请输入资料	8416.RF Q Q
集成資源	集成资源组名称/ID	1016	网络	- 绑定項目 ¥	状态	资源包规格/数量	到期时间	操作
10 -9:0994X		ŕж			基础的: • 运行中	8C 16G / 1	2022-11-27 23:47:42	关联项目 解除关联 调整配置 成数 初级
		<i>г</i> #			基础包:④ 运行中	8C 16G / 1	2022-10-27 22:47:27	关联项目 新除关联 调整配置 续费 明报

🕛 说明

若您使用的是 WeData 产品,请点击进入 WeData 控制台。



2. 购买集成资源组。

<text> W Description Set in the set in</text>	集成资源	原组购买。这些产品评商	自产品文档 ⑦计费说明 器产品控制台
 INTER INTERSION INTERSIO	使用说明 数装	R集成资源为紊线同步、实时同步任务独厚扶行资源。配置详情请参见集成资源说明 td	
NHARE NL NUTRIPIO (2.5.17)	资源配置		_1. 远择资源万案
 e. Status (S) (S) (S) (S) (S) (S) (S) (S) (S) (S)	应用场景	离线数据同步 高线+实时同步 (无队列)	q线+实时同步(含队列)
NHE NHE 2. 配置离线/实时包 NHE NHE NHE NHE NHE NHE AND AND AND AND AND AND AND AND AND AND AND AND <		 Q交持裏域同步任务 这用于素或数仓构建、全/增量批量数据迁移等业务场景 集成资源组可仅包含离线资源 集成资源组可仅包含离线资源 集成资源组它仅包含离线资源 集成资源组它仅包含离线资源 集成资源组它公司法 集成资源组它公司法 集成资源组它公司法 集成资源组它公司法 第一次时资源包 ● ● ● <!--</td--><td>交持AgentigSDK实时主动上报日志文件数据、CDC/Binlog实时同步、 离线同步 运用于系统日志买集与投递、百万亿级则海量数据传输等业务场景 集成资源给包含离线资源包、实时资源包、海量队列</td>	交持AgentigSDK实时主动上报日志文件数据、CDC/Binlog实时同步、 离线同步 运用于系统日志买集与投递、百万亿级则海量数据传输等业务场景 集成资源给包含离线资源包、实时资源包、海量队列
<pre>kit is is is is it is it</pre>	规格配置	◎ 真核資源信 主要用于基础管理及当时真线同步任务	า
Note Image: Imag		资源包积格 8C160 8C32G 采用与过量资则评求风格与过剩限制 2	2. 配置离线/实时包 规格
		○ ★内共通告 130m1 #11341891891 (1412) (141	
All I I I I I I I I I I I I I I I I I I		2時後藤堂市林市美麗市市地震地である	
新福 東京 京京 <		资源包款量 - 1 +	
R8 	地域	广州 上海 北京 成都 美国社会 南京 東京道道の心地は、投手不同地域の比小道局用局格不正通、创建成心影不可能描述。 彼る道境高井 建议力学品版本选择用一地域 可降低服务活用时间	3. 设置地域/网络
京都部 「予始集選選組-Stacymb 日初 市政工業 「日井安正	网络	✓ ○ нессолутяр, 就会可用ессилут 重成表示意义未准正的转量并在于同一VPCH,或并这VPC需要量已将公开能力。如同和的时间不会试。如可以在的相位的描述有用的名词称并有用名	
計工 日本 計工 日本 日本 日本 <td>资源组名称</td> <td>广州集成资源组-51symth</td> <td>_</td>	资源组名称	广州集成资源组-51symth	_
I 计数型 计数型 加速期 加速期 加速期 加速期 加速期 加速期 Marrier Marrier <	描述	诸输入集成资源规则还	
I 竹貝 本口 S × D × D × D × D × D × D × D × D × D ×			
N SAR E-Exp BRXHK 170 271 373 473 573 673 773 873 973 174 24 34 47 BRXHK 170 271 373 473 573 673 673 773 873 973 174 24 34 47 BRXHK * BlakkT BX*980.007, GRXHKBEBRIBLIND GRHK SKRIPE 4. SKRIPEGIC (OT 3.) SKRIPE 1900 100 000, GRXBBBBBBBBBBBBBB, END GRHKBEBRIBLIND GHKB SKRIPE 1900 100 000, GRXBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB	计费类型	(A)(#4) B	
Ky fit		G+ G7	
除了 ● 品格灯 HP 今都和見時, 设数形版目用品域数	购买时长	11月 2个月 3个月 4个月 5个月 6个月 7个月 8个月 9个月 1年 2年 3年 4年	
★联項目空 ・ ・ ・ ・ ・ ・ ・ ・ ・ ・ ・ ・ ・ ・ ・	续订	☑ 自动统订 账户余额先见时,设备到期后批月自动成员	
文联项目空间 ····································			4. 关联项目(可诜)
XXXY118 XXXXX YXXX YXXX YXXXX YXXXX YXXXX YXXXXX YXXXXX YXXXXXX YXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	关联项目空		
項目空间 北京 ・ terry ・ 忠忠家 北京 ・ ・ 忠忠家 北京 ・ ・	大联项目	14.45大联 采取项目空间后,优先项目内任务可提用率资源通道行。后该可在产品控制台中推动。	
かび条款 ズ 我已阅读并同意 (編秀协议) は		項目空间 北京 V terry V 参記商業 北京 地域的 ⁺¹ 品級券, 加新集成資源級仅可其基本地域下的項目空间。	
	协议条款	※ 我已阅读并同意 (服务协议) 22	

🕛 说明

- 离线资源包与实时资源包可根据实际数据情况配置规格、以及数量。
- 资源组网络建议选择 MySQL 和 DLC 所在网络;若 MySQL 和 DLC 不在一个 VPC 环境,可为 VPC 配置开通公网,详细操作参见 资源组配置公网 。

3. 购买完成后,返回控制台并关联资源组与项目空间。

🕛 说明

若在购买页面内已经关联资源组与项目空间,可忽略此步骤。



集成资源 ◎北京 ▼							
018							
集成资源组名称/ID							
				基础包:④ 运行中	8C 16G / 1		
				基础包: 🕞 运行中	8C 16G / 1	201 0-21 22:47 27	
					16C 64G / 1		
	广州	关联项目		×	8016G/1	2022-10-23 15:10:58	
		资源名称			16C 64G / 1		
	-itur	資源地域 广州 項目名称 请选择 选择 动脉 动动	而日交詞	¥	8C 16G / 1		
		2257年 C2 03388 8 9 4	稳定 取制		16C 64G / 1		
and the second	北京	vpc-d3rmf9qk/subnet- 9pt7ol5l	emr_test(110225083950397 0304)	基础包:③ 运行中	8C 16G / 1		

配置数据源

1. 配置 MySQL 数据源。

进入**项目管理**模块,选择**数据源管理 > 新建数据源 > 选择 MySQL**。以 MySQL 数据源为例,数据连通性测试成功后,单击**保存**。

数据源管理							数据源 ×
新建数据源 批量授权	批量移交					所属项目•	test
助据源名称 数据源类	型▼ 类型▼	显示名	描述	所属项目 🍸	0.	数据源名称 ★	
MYSQL	自定义源				dI	显示名	inlong_5_7
DLC 25	自定义源				di	描述	unouă"2"1,
MYSQL	自定义源				dI	数据源权限	• 项目共享 ① 仅个人与管理员
cos	自定义源				dI	获取实例 ◆	北京 * 🗘
СКАҒКА	自定义源				dI	或這座否称* 用户名*	root
cos	自定义源				dI	密码★	
DLC	系统源					数据连通性	开始调试 🥥 這通成功
MYSQL	自定义源				dl	确认	1023/j

2. 配置 DLC 数据源。

进入 项目管理模块,选择数据源管理 > 新建数据源 > 选择 DLC,配置数据源参数,并在连通性测试成功后,即可单击保存。



所属项目 *	test 🔻
数据源名称 *	
显示名	and control and control of the contr
描述	test
数据源权限	○ 项目共享 ○ 仅个人与管理员
JDBC URL *	jdbc:dlc:c task_typ∈ at aLakeCat
secretId 🛈 *	
secretKey 🛈 *	
数据连通性	开始测试
确认	取消 2/2

参数	说明
JDBC URL	格式参考: jdbc:dlc:dlc.internal.tencentcloudapi.com? task_type=SparkSQLTask&database_name=&datasource_connection_name=DataLakeCatalog®ion=ap- beijing&data_engine_name=test_engine 若需要使用小文件合并,数据源的访问域名必须使用 dlc.internal.tencentcloudapi.com, task_type 必须使用 SparkSQLTask, data_engine_name 指定的引擎会用于实时同步后的小文件合并。
	介 注意: 小文件合并会使用此处配置的 DLC 数据引擎并在合并的时候占用部分资源,请合理配置资源,如果不启动小文件合并,该 DLC 引擎不会被使用。
secretId/secre tKey	填写拥有引擎和 SQL 执行权限的账号或者子账号的密钥。可在 API 密钥管理 中查看。

配置实时同步任务

1. 创建任务。

进入数据集成模块,创建**实时同步任务**,在弹出的提示框中输入任务名称和备注,选择 **画布模式**或**表单模式**,并单击<mark>完成</mark>。本介绍以画布模式为例。



实时同步								
新建实时同步					请输入任务名称		Q	φ
任务名称	创建方式	描述	创建人 ▼	创建时间 \$	1	操作		
	表单模式	-		2022-09-27 20:55:2	3	扁辑 删除		
	画布模式	-		2022-09-19 16:51:0	5	扁辑 删除		
	画布模式	-	-	2022-08-01 14:30:2	4 1	扁辑 删除		
-	画布模式	-	-	2022-08-01 14:22:3	9 1	扁辑 删除		
	画布模式	-		2022-07-26 11:14:2	21 1	扁辑 删除		
	画布模式		1001010-0000	2022-07-22 14:24:1	5	扁辑 删除		

2. 编辑任务。

单击新建的实时同步任务名称,进入任务编辑界面,通过拖拽分别新建读取数据源和写入数据源,并通过连线指定数据流向。



3. 配置 MySQL 节点。

双击画布中的 MySQL 节点,对读取数据源进行配置。如下图选择需要同步的数据库表,读取模式选择全量模式,完成后单击保存。



	读取节点		×
	基本信息		
	节点类型	MySQL读取节点 操作指引	
mysql_in_1	节点名称		
	数据来源		
_	数据源	•	新建数据源
dlc_out_1	库	8	
	表 🕄	8	
		添加分库分表	(
	格式 🛈	utf-8 👻	
	读取模式	●全量	

4. 配置 DLC 节点。

双击画布中的 DLC 节点,对 DLC 写入数据源进行配置。如下图选中需要写入的库表,根据业务需求选择写入模式,并指定唯一键。例子中指定唯一键为 ID 和 MySQL 的主键保持一致。

	写入节点		×
	基本信息		
	节点类型	DLC写入节点	
mysql_in_1	节点名称		
	数据来源		
¥	数据源	· · · · · · · · · · · · · · · · · · ·	新建数据源
dlc_out_1	库	8	
	表	8	
	写入模式 🛈	Oupsert oppend	
	唯一键 🛈	id 🕄	
	参数	请输入参数名称及值(格式为:parameter=value),多个参数使用换行符 分割	
	取消	保存	

下拉至底部,配置 MySQL 与 DLC 表字段映射,完成后单击**保存**。

来源表字段名	类型		目标表字段名	类型	同名映射
database_wedata_di	string 💮	••	traceid	string	同行映射
table_wedata_di	string 💮	••	os	string	清除映射
type_wedata_di	string 💮	••	terminaltype	string	排 序
ts_wedata_di	timestamp 💮	••	devicetype	string	
processing_time_wedata_di	timestamp 💮	••	appversion	string	
name	varchar 💮	• •	sdkversion	string	

5. 任务保存与提交。

○ 配置完节点后,单击任务数据配置集成资源组。此资源组为 配置集成资源组 步骤3中已关联至本空间的资源组。



○ 完成后,单击**提交**按钮,并在弹窗口中勾选**立即启动**。

提交	×
提交当前任务配置。当前任务线上无运行中或停止状态的作业版本,可立即启动运行或 前往运维中心手动启动 <mark>-</mark> 立即启动	
确认 取消	

6. 查看并运维实时任务。



○ 提交任务后,可进入**实时运维**页面查看并监控任务状态。

◎ 同步链路 ^	实时任务运维						
实时同步	這行 暂停 應成 (₱止 更多操作 ▼					请输入任务名称
高致问题	任务名称	君任人 ▼	运行状态 平	最近启动时间 •	结束时间 •	最近操作时间 •	過作
山 任务运输 ^			◎法行由	2022-05-05 15:39:29	2022-05-24 19:29:43	2022-06-05 15:39-29	1005-1020 1007 1010 1008 48.0
実时运推	Contrast.		C AB13 T	2022-00-00-10-00-20	1012-00-04 1922-00	1011-00-00 10:0010	MATINE MATI NERK MATY TAL
离线运维							

○ 单击运行监控,可查看当前任务数据指标统计、以及配置监控告警等。

存量任务处理

1. 如果存量实时同步任务需要添加小文件合并功能,首先需按照步骤一修改表属性。

① 说明 ALTER TABLE db_name . new_table_name SET TBLPROPERTIES ('write.compact.enable' = 'true', 'write.compact.snapshot.interval' = '20');其中合并周期参数 'write.compact.snapshot.interval' 需要根据业务需求进行调整。

• 如果存量表已经存在大量的小文件,推荐手动将小文件合并到一定数量之下后,再启动定时合并功能。

2. 将实时同步任务**停止**再运行即可。

③ 同步链路 ^	实时任务运维							
实时同步 高线同步	317 N# N# ##	更多接作 *					请输入任务名称	٩ ٥
山 任务运维 ~	□ 任务名称	責任人 〒	追行状态 ▼	最近启动时间 \$	结束时间 \$	最近操作时间 \$	18:17	1. 点击停止
实时运维			•运行中	2022-06-06 15:39:29	2022-05-24 19:29:43	2022-06-06 15:39:29	通行监控 运行 目標 醫师	停止
离线运维							2. 停止成功	后,点击运行

MySQL 分库分表同步至 Hive

最近更新时间: 2024-07-09 22:01:41

目前离线及实时数据同步任务支持 MySQL 分库分表同步至目标数据源中。本实践介绍 MySQL 分库分表离线同步至 Hive 数据源。

适用场景

业务层对基础同结构的业务数据使用分库分表的方式存储在不同的 MySQL 数据库实例或同一实例的不同表内。应用层需要对分布在不同数据库内表同时同步到 数仓 ODS 层中统一存储。

操作步骤

- 1. 创建离线同步任务,并从离线同步列表中点击对应任务名称进入画布配置界面。
- 拖拽 MySQL 读取节点,默认 MySQL 读取节点支持一个数据源,单个数据源(库)内可选择多张 MySQL 表。若存在分库情况,可单击添加分库分表,即 可创建多个数据源并添加对应的表对象。



() 说明:

分库分表情况下选择的多个表对象需保证 Schema 信息一致(包括字段名称、字段类型)。**数据字段**模块内系统默认展示第一个数据源的第一张表 的元数据字段信息,若多表间字段不一致可能会导致运行失败。



3. 拖拽 Hive 写入节点,并配置读写字段映射。



4. 保存任务信息,进行测试运行或提交至运维中心。

SQL Server 数据离线同步至 Hive 分区

最近更新时间: 2024-07-09 22:01:41

离线周期同步任务支持周期性更新表分区数据,本实践介绍 SQL Server 表数据定时更新至 Hive 分区。

适用场景

周期性创建或更新 Hive 分区表数据,将数据表中内容按照周期任务实例计划调度时间写入到对应一级或多级分区。

操作步骤

- 1. 创建离线同步任务,并从离线同步列表中单击对应任务名称进入画布配置界面。
- 2. 拖拽 SQL Server 读取节点,配置数据来源信息并在数据字段模块内配置时间参数。

← demo1	I ◎				
请输入节点类型	Q		读取节点		×
读取	*		其太信息		
写入	Ŧ		节点类型	SQL Server读取节点	
关系数据库	a		节点名称	input-SQLServer-6	
MySQL	TDSQL-C Mysql	[]			
PostgreSQL	SQL Server	input-SQLServer-6	数据来源		
Oracle	DB2		数据源	sqlserver v	新建数据源
			库	wedata_dev 📀	
大数据	UNI TIANA		模式	dbo 🕲	
😵 Hive	HBase	output-Hive-7	表	myuser 🕲	
Clickhouse	DLC		切割键 🛈	请手动输入或选择	
Kudu	(B) HDFS		筛选条件 🛈	选填。请根据数据源类型对应的SQL语法填写SQL	
Greenplum	SaussDB				
Gbase	B TBase		数据字段		
半结构化			子段	EP.	
© COS	FTP		id	int	
G SFTP			\$(nany-MM-dd)	varchar	
NOSQL			\$(yyyy-wiw-dd)	2 XA	
Elasticsearch	Redis		\${HH:mm:ss}	参数	
				+ 字段配置 在读取节点端配置时间4	参数
			取消	采存	



() 说明:

- 时间参数定义方式请参见 时间参数说明。
- 时间参数类型需配置为"参数"。





4. 保存任务信息,进行测试运行或提交至运维中心。

MySQL + gh-ost 实时同步至 Kafka

最近更新时间: 2023-12-28 08:54:41

业务场景

- gh-ost 的业务场景是在 MySQL 中进行在线表结构变更,即 Online DDL,而不影响业务的正常运行。它可以解决传统的 alter table 或 create index 等命令导致的表锁、性能下降、同步延迟等问题。它适用于需要对表进行修改的场景,例如增加新列、添加索引、修改字段类型等。
- 本实践介绍使用 gh-ost 变更 MySQL 的表后,实时同步它的 DDL 变更记录到 Kafka。

gh-ost 使用前提

MySQL 环境准备:数据集成 MySQL 环境准备与数据库配置。

- 1. gh-ost 必须能访问 MySQL。
- 2. 如果 MySQL 是腾讯云的 CDB,在 gh-ost 执行命令的参数里面需要添加 `--aliyun-rds`。
- 3. gh−ost 工具执行过程中,生成的临时表规则为 ^_(.*)_(gho|ghc|del)\$,其中 (.*) 是变更表的名称,不支持自定义临时表的名称。
- 4. 其他 gh-ost 限制,可以参考: gh-ost/requirements-and-limitations.md at master · github/gh-ost · GitHub。

操作步骤

1. 创建 MySQL 实时同步到 Kafka 的整库同步任务。

🗾 WeData 🔰 🕀 数	居集成 ~				200	
○ 集成概览	← demo1				任务提交	皆南 亿
配置中心	提交 🖻	5 - 2				
🕒 实时同步						
🖻 离线同步	1 链路选择	> (2) 数据来源设置 > (3) 数据目标	前设置 > (4)运行设置			
运维中心	链路类型	MySQL	⊸ → к	afka	v	
🐼 实时运维		· ▲ 快速选择				
🖧 离线运维		全部链路 同步至Kafka 同步至DLC	同步至StarBocks 同步至Doris 同步至Hive	軍名		
◎ 告警事件						
设置管理		📉 Mysql 🛛 🛶 🔕 Doris	Mysql → 🔯 DLC	📉 Mysql 🛛 🛶 😵 Iceberg		
四 采集器管理	Г (支持MySOL实假组验检 	支持MySQL实例级监控	• 支持自动创建Iceberg库表		
@ 数据源管理 12		📉 Mysql 🛛 —> 🐉 Kafka	📆 TDSQL-C 🛛 —→ 👂 Doris	🚼 TDSQL-C 國 DLC		
白 集成资源管理 I2	L L	 支持Topic自动创建 	• 支持MySQL实例级监控	• 支持MySQL实例级监控		
		 중 TDSQL-C → 당 Iceberg 支持自动创雄Iceberg库表 	중 TDSQL-C → 紧 Kaika 支持Topic自动创建 	 餐 Kafka → @ DLC 支持消息内容动态匹配目标表 		
		 Kafka → 貸 leeberg 支持消息内容动态匹配目标表 	 	 Oracle → 愛 Iceberg 支持自动创建Iceberg库表 		
		PostgreSQL → 器 Kafka • 支持Topic自动的建	PostgreSQL 👂 Doris	StarRocks		
		S Mongo → 器 Kafka • 支持Topic自动的建	S Mongo → S PostgreSQL	Mongo → 図 DLC ● 支持自动创建DLC 座表		
		 Q Oracle → 餐 Hive ● 支持源表新增列同步 				
E	取消下一	- 5				

2. 在任务配置中的数据来源设置,高级设置开启同步 gh-ost 临时表功能。



🗾 WeData 🔰 🕀 🕸	据集成 ~		
● 集成概览	← demo1		任务提交指南 团
配置中心	提交	a	
🕒 实时同步			
Ⅲ 离线同步	💛 链路选择	> 2 数線未課设置 > 3 数据目标设置 > 4 运行设置 > 5 配置預定	Mysql实时整库同步至Kafka
 道中心 今 突时运進 △ 素线运進 ④ 告警事件 ③音響事件 ③音響 四 采集器管理 ④ 数据源管理 ビ ● 集成资源管理 ビ 		Srydgeranded ry Gang, Srydgeranded Standing Q ↓ Srydgeranded ry Gang, Srydgeranded Standing Bold Srydgeranded ry Gang, Srydgeranded Standing Bold Srydgeranded Standing <	
	 · (東東模式 · (武道操作 ① · (国際) · (· (· (金羅・理羅	

3. 在任务配置中的**运行设置,消息处理策略**选择需要的 DDL 消息处理策略,这里选择下发新增列,删除列,重命名列和修改列类型。配置完其余选项后,提交 任务。

🗾 WeData 🛛 🕣 黝	日 集成 - 日 集成	•			
○ 集成概览	← demo1				任务提交指南 IZ
配置中心	提交 🔒	Ð			
🕒 实时同步	_				
■ 离线同步	🕑 链路选择	> 🛛 😼 数据来	「源设置 〉 🧹 数据目标设置 〉	> 🤞 运行设置 > 👍 配置預算	Mysql实时整库同步至Kafka
运维中心	JobManager规格	1		v	
💮 实时运维	TaskManager規格	1		×	
🕰 离线运维	并行度 🚯	- 1 +			
◎ 告警事件					
设置管理	消息处理策略				
🖂 采集器管理	DDL消息处理策略 🛈	新建表	忽略信息 ▼		
@1 数据源管理 13		删除表	忽略信息		
集成资源管理 12		重命名表	忽略信息 マ		
		新增列	不忽略消息,传递变更消息至下游 🔻		
		删除列	不忽略消息,传递变更消息至下游 ▼		
		重命名列	不忽略消息,传递变更消息至下游 ▼		
		修改列类型	不忽略消息, 传递变更消息至下游 ▼		
		清空表	忽略信息		
	checkpoint间隔	- 1 +	分钟 🔻		
	斯数据策略	cos归档 不	归档		
Ξ	上一步下	-#			

4. 使用 gh-ost 变更原始表。

4.1 查看表结构。

Field Type Null Key Default Extra
id int NO PRI NULL name varchar(255) YES NULL



2 rows in set (0.00 sec)

```
4.2 使用 gh-ost 工具在 源表 上增加一个字段。
gh-ost 工具下载: <u>Releases</u> · <u>github/gh-ost</u> · <u>GitHub</u>
下载后解压,执行下面命令,为 `tb1` 表添加 `c` 字段。
```

./gh-ost `

- -max-load=Threads_running=25
- --critical-load=Threads_running=1000
- -chunk-size=1000 \setminus
- -throttle-control-replicas="" \
- --max-lag-millis=1500 \
- --user="root" \ // 用尸名
- -password="test" \ // 🎦 🖪
- --host=12/.0.0.1 \ // mysql **Hy**i}
- -allow-on-master \
- -database="databaseName" \ // **需要变更表所在的数据库名称**
- -table="tb1" \ // **需要变更的表**
- -verbose \
- -alter="engine=innodb" \
- --switch-to-rbr \
- --allow-master-master \
- --exact-rowcount \
- --concurrent-rowcount `
- --default-retries=120 \
- --panic-flag-file=/tmp/ghost.panic.flag \
- --default-retries=120 \
- --alter="add column c varchar(255);" \ // alter **变更语句**
- --approve-renamed-columns \
- --initially-drop-ghost-table
- --initially-drop-old-table \
- --ok-to-drop-table '
- ___aliyun_rds \ // 使用腾讯云CDB,需要添加这个参数
- --execute

4.3 查看变更后的表结构。

Field			Extra
id		PRI	
name	YES		
	YES		

5. 在 Kafka 上查看结果。

{"data":[],"es":1683519346931,"table":"tb1","type":"INSERT","database":"databaseName","ts":0,"sql":"alter table `databaseName`.`tb1` add column c varchar(255)","mysqlType": {"id":"INT(-1)","name":"VARCHAR(255)","c":"VARCHAR(255)"},"sqlType": {"id":4,"name":12,"c":12},"isDdl":true,"pkNames":["id"],"operation": {"type":"alterOperation","alterColumns":[{"alterType":"ADD_COLUMN","newColumn":{"name":"c","definition": ["255"],"jdbcType":12,"isNullable":true,"defaultValue":"","comment":"","nullable":true}}],"operationType": "ALTER"},"incremental":true,"dataSourceName":"databaseName","ddl":true}

示例如下:



∕⊘膨	新田云
-----	-----

消息详情		×
() 当	前查询的消息已经被强制转换为String类型,如出现乱码,请分析您消息的序列化格式以及编码格式	
Headers	暂无数据	
Кеу	暂无数据	
Value	{"data":[],"es":1683519346931,"table":"tb1","type":"INSERT","database":"menghuiyu","ts":0,"sql":"alter table `i ``````````````````````````````````	
	确定	

将 MySQL 数据同步至 Doris 概述

最近更新时间: 2024-06-18 14:47:41

本文将为您介绍如何通过 DataInLong 数据集成将 MySQL 中的数据实时导入 Doris 中,此处以腾讯云数据库 MySQL 为例,介绍实时同步任务的配置及操 作实践。

△ 注意:

建议提前合理分配及配置网络环境,确保 MySQL、Doris、数据集成资源之间网络互通。

- 若 MySQL、Doris、DataInLong 集成资源处于同一个 VPC 内:此时网络联通,可直接使用(推荐方式)。
- 若与集成资源位于不同 VPC: 需购买对等连接打通集成与数据源所在 VPC。
- 若数据源位于 IDC 或其他经典网络环境下:可购买 VPN 连接或 专线网关打通集成资源与 MySQL/Doris 集群所在 VPC。
- 若数据源可开通公网:可购买 NAT 网关,允许集成资源通过网关连通数据源所在 VPC。NAT 网关配置流程请参见 集成资源组配置公网。

步骤1:登录注册

登录腾讯云官网。如果没有账号,请参考 账号注册教程 。

步骤2:提前准备好 MySQL 数据库实例

数据集成目前支持 MySQL 数据库版本为: 5.6, 5.7, 8.0.x。具体操作请参见 源端 MySQL 准备。

步骤3:提前开通好云数据仓库 TCHouse-D 集群

建议选择内核版本:1.1、1.2或以上。具体操作请参见 目标端 Doris 准备 。

步骤4: 购买 DataInlong,并配置项目空间和集成资源

建议选择内核版本: 1.1、1.2或以上。具体操作请参见 配置 DataInlong 项目空间及集成资源。

步骤5:在 WeData 或 DataInlong 配置数据实时同步任务

- 需要进行单表同步时,可配置单表实时同步任务,具体操作请参见 配置单表实时同步任务。
- 需要进行整库迁移时,可配置整库实时迁移任务,具体操作请参见 配置整库实时迁移任务。

步骤6:进入运维中心进行任务运维

具体操作请参见 任务运维 。

源端 MySQL 准备

最近更新时间: 2024-08-16 16:27:21

新建 MySQL 数据源实例

! 说明

目前数据集成支持 MySQL 数据库版本为:5.6,5.7,8.0.x。

使用腾讯云 MySQL 时

- 1. 您可登录 云数据库 TencentDB 控制台,进入 MySQL 实例列表。
- 2. 单击新建购买指定数据库版本的云实例。

	, <mark>注意</mark> : 购买 MySQL 云	实例时,建议配置	MySQL、I	Doris 处于同一个	VPC.			
MySC	QL - 实例列表 🛛 🛇	北京 7 其他地域 37 ▼					新功能速速 72	🌲 异常告警 用户指南 🖸
G	MySQL数据库代理(Prov	y)火热公测中,数据库代码	理可实现 自动读写 ?	分离 ,将读请求转发至只读:	实例,降低主库的负:	载。	羊犢	• • • ×
新建	L 一键诊断	对比监控重启	续费	更多操作 ▼			多个关键字用竖线 " " 分隔,多个过滤标签用回车键分隔	Q \$\$\$\$ \$\$
	实例 ID / 名称 🍸	监控 / 状态 / 任务 🍸	可用区 🔻	配置 ▼	数据库版本 🔻	引擎 ▼	内网地址 ③ 计费模式 ▼ ◆ 所属项目 ▼	操作
	10 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	山 ● 运行中	北京一区	双节点(本地盘) 通用型-4核8000MB/… 网络:Default-VPC - 测试子网	MySQL8.0	InnoDB	按量计费 默认项目	登录 管理 更多 ▼

使用非腾讯云 MySQL 时

您可以在 MySQL 数据库中通过如下语句查看当前 MySQL 数据库版本,检查当前待同步的 MySQL 是否符合版本要求。

```
select version();
创建账号并赋权
```

▲ 注意:

为保证实时数据同步顺利进行,您必须定义一个对 Debezium MySQL 连接器监控的所有数据库具有适当权限的 MySQL 用户。该 MySQL 账号必须 拥有数据库的 SELECT、REPLICATION SLAVE 和 REPLICATION CLIENT 权限。

使用腾讯云 MySQL 时

- 1. 您可登录 云数据库 TencentDB 控制台,单击 实例 ID/名称 进入实例详情页。
- 2. 进入数据库管理 > 账号管理 页面,单击创建账号来新增账号,修改权限来配置账号权限。



实例详情	实例监控	数据库管理	安全组	备份恢复	操作日志	只读实例	数据安全	连接检查			
数据库列表	参数设置	帐号管理									
创建帐号	导出帐号	密码复杂度: [关]	使用动态的	凭据()						请输入账号名	Q
帐号名		主机		ì	连接数限制		备注		操作		
I man		%		-	-				修改权限克	隆帐号 更多 ▼	
root		%			-				重置密码 重	置权限	
共2项									10 ▼ 条 / 页 🛛 🛛 🖌 🗸	1 /1页	

使用非腾讯云 MySQL 时

您需要通过 SQL 语句授予并刷新账号权限。

mysql> ('passwor mysql>FI	GRANT SELECT, rd'; LUSH PRIVILEGI	SHOW ES;	DATABASES,	REPLICATION	SLAVE,	REPLICATION	CLIENT	ON *.* TO	'user'	IDENTIFIED B	Υ

- 🕛 说明:
 - 1. 启用 scan.incremental.snapshot.enabled 时不再需要 RELOAD 权限(默认启用)。
 - 2. 查看更多关于权限说明。

创建数据库

使用腾讯云 MySQL 时

- 1. 您可登录 云数据库 TencentDB 控制台,单击 实例 ID/名称 进入实例详情页。
- 2. 进入数据库管理 > 数据库列表 页面,单击 创建数据库 来创建 MySQL 数据库。

← cdb39	93546								登录	NEW 一键诊断	重启	手动备份
实例详情	实例监控	数据库管理	安全组	备份恢复	操作日志	只读实例	数据库代理	数据安全	连接检查			
数据库列	 表 参数设置	帐号管理										
数据导入	创建数据库	使用云访问安全	全代理 🛈								请输入数	据库名
数据库名						状态		数据	居库字符集		服务器字	符集
- 1						运行中		UTF	-8		UTF8	
共1项											10 ▼ 条/页	

3. 在数据库登录跳转页面输入账号管理中已创建好的账号名和密码。



>数据库管理		
数据库管理(DMC)是一个高效、可靠的一站式数据库管理平台,帮您 更加便捷、规范地管理多种数据库实例。	类型 MySQL ▼ 地域 华南地区 (广州) ▼	
 新建库表、视图、存储过程等 数据导入导出 SQL 执行及安全审计 权限管控、数据变更审批 了解更多 >>> 	实例 账号 数据库账号 在"账号管理"中创建的账号 密码 数据库密码 账号名对应的密码	名
更多数据库 SaaS 服务 数据传输服务 DTS 数据库智能管家 DBbrain 数据库备份服务 DBS	登录	

使用非腾讯云 MySQL 时

您可以通过 MySQL Client 等客户端进行建库操作。

开启 Binlog 并确认 Binlog 格式

数据集成基于 MySQL binlog 进行数据同步时,要求 MySQL 服务器开启 binlog,并将 binlog 格式符配置为 ROW、将 binlog_row_image 配置格式为 FULL。

使用腾讯云 MySQL 时

腾讯云 MySQL 默认已开启 Binlog,可进入对应 MySQL 数据库管理--参数设置 页面设置并管理对应 binlog 参数:

() 说明:

MySQL 8.X 默认 binlog_format 为 ROW,无需额外配置。

云数据库 TencentDB	÷				NEW 重启 手动备份 回档
〓 概览	实例详情 实例监控 数据库管理 安全组	备份恢复	操作日志 只读实例	数据安全 连接检查	
关系型数据库 長 TDSQL-C ・	数据库列表 参数设置 帐号管理		N	EWI	
दी MySQL ^	批量修改参数 默认模板 自定义模板 导入参数	导出参	数 另存为模板 智能参数调优		请输入参数名 Q
实例列表	参数名	是否重启	参数默认值①	参数运行值	参数可修改值
 任务列表 	avoid_temporal_upgrade①	否	OFF	OFF	[ON OFF]
 参数模板 ・ 回收站 	back_log ③	是	3000	3000	[1-65535]
・诊断优化	binlog_cache_size①	否	2097152	2097152	[4096-16777216],且必须为 4096 的倍数
・ 数据迁移	binlog_checksum	否	CRC32	CRC32	[NONE CRC32]
 ・ 灾备同步 	binlog_format①	否	ROW	ROW	[ROW]
 数据库备份 数据库审计 	binlog_order_commits①	否	ON	ON	[ON OFF]
MariaDB	binlog_row_image①	否	FULL	FULL	[FULL MINIMAL]
I SQL Server 👻	binlog_rows_query_log_events	否	OFF	OFF 🖋	[ON OFF]
₽ PostgreSQL ·	binlog_stmt_cache_size①	否	32768	32768	[4096-16777216]
企业级分布式数据库	block_encryption_mode	否	AES-128-ECB	aes-128-ecb	[AES-128-ECB AES-192-ECB AES-256-ECB AES-128-CBC AES-192-CBC AES-256-CE
III TDSQL-H	bulk_insert_buffer_size①	否	8388608	8388608	[0-4294967295]
NoSQL 数据库	cdb_enable_offset_pushdown	否	OFF	OFF	[ON OFF]
三 给产品打个分 🧿					

使用非腾讯云 MySQL 时



可通过以下命令查看并配置 binlog 格式。

variables	
variables	

创建数据表

使用腾讯云 MySQL 时

1. 您可进入 数据库管理,登录腾讯云 MySQL 。

>数据库管理		
数据库管理(DMC)是一个高效、可靠的一站式数据库管理平台,帮您 更加便捷、规范地管理多种数据库实例。	类型	MySQL ▼ 华南地区 (广州)
 ・新建库表、视图、存储过程等 ・数据导入导出 	实例	
 SQL 执行及安全审计 权限管控、数据变更审批 了解更多 >> 	帐号 密码	数据库帐号 在"账号管理"中创建的账号名 数据库密码 账号名对应的密码
更多数据库 SaaS 服务		登录
数据传输服务 DTS 数据库智能管家 DBbrain 数据库备份服务 DBS		

2. 在数据库管理页面新建数据表。

2	数据库管理	新建 ~	库管理	实例监控	实例会话	SQL 窗口	导入	、导出	数据库智能管家 🖸	腾讯云图数据可视化 🗹
-	1.1.1.1.1	•	φ	首页 库	管理 新建	表				
田 表	模糊匹配表名	Q	+ 新建表	所属库:	n. n. nf. ff					
				基本信息	列信息	索引	外键	分区		
闰视	 ▶ I 列 ▶ II 列 			基本信息						
密	▶ itest			表名 *	请输入表名					
日 存储过程				备注						
				存储引擎	InnoDB		•			
fx K				字符集	请选择		•			
数				校验规则	请选择		•			
€				更多选项	1214-12					
器				行格式	请选择		•			



使用非腾讯云 MySQL 时

您可以通过 MySQL Client 等客户端进行建表操作。



最近更新时间: 2024-06-14 16:36:51



购买 Doris 集群

- 购买方式一:进入腾讯云官网,登录后在官网首页单击**立即选购**。
- 购买方式二:登录 腾讯云数据仓库 TCHouse-D 控制台,在集群列表页新建集群。

🔗 腾讯云 🛛 🕮	云产品 √						搜索产品、文档	Q	⑦ 小程序		集团账号 >	备案	工具~	支持 ~	费用 ~
云数据仓库 for Apache Doris	集群列表 🕲 北京 2	● 其它地域 19 ▼													採作
晋 概览	8532:02.21						每个搜索项用回车	目键分隔;	单个搜索项中用	用竖线"["分	啊: 集群标签	り关键字用]*key:value	形式,若仅	填 C
■ 集群列表 □ CDW Studio IZ	ID/名称	状态	FE节点	BE节点	内核版本	可用区	网络/子网	付费到	(型	1	创建时间		操作		
	consume utilitispz1 līg rili∎_ III. ∕	运行中	标准型, 1个节点 4核16G, 200G	标准型, 3个节点 4枝16G, 200G	1.2	北京六区	vp= h0x=F=00	按量;	费	:	2023-05-25 14:	42:24	水平扩容	垂直变配 等	搬 更多 ▼
		运行中	标准型。1个节点 4核16G, 200G	标准型, 3个节点 4核16G, 200G	1.2	北京六区	vpc g subnation of the	按量计	费	1	2023-03-31 15:	33:37	水平扩容	垂直变配 訇	服 更多 ▼
	共 2 条										10	▼ 条/页	н -	1	/1页)

• 进入集群购买页面,按需选购 FE、BE 的集群资源规格,详细操作步骤可参见 新建集群。

· · · · · · · · · · · · · · · · · · ·
● 建议生产集群配置如下:
○ FE:开启高可用,3节点,每个节点16核64G。
○ BE: 3节点,每个节点16核64G。
● 购买 Doris 集群时,建议配置 MySQL、Doris 处于同一个 VPC 内。

创建账户并赋权

1. 集群购买完毕后,可进入腾讯云数据仓库 TCHouse-D 控制台集群列表,单击集群 ID/名称进行集群管理。

🔗 腾讯云 🛛 🖧	│ 云产品 ~	▲ d± b¥ bkr.st		₩ 新日本 # HC 第二	+ 4		搜索产品、文档…	Q @ 小程序
云数据仓库 for Apache Doris	集群列表 🕓 北京 2	2 ● 其它地域 19 ▼						
晋 概览	新建集群						每个搜索项用回车	建 分隔; 单个搜索项中用;
集群列表	ID/名称	状态	FE节点	BE节点	内核版本	可用区	网络/子网	付费类型
뎚 CDW Studio IZ	test 🖍	运行中	标准型, 1个节点 4核16G, 200G	标准型, 3个节点 4核16G, 200G	1.2	北京六区	vr. 1	按量计费
	Concorns Jyju1ut5 līj	运行中	标准型, 1个节点 4核16G, 200G	标准型, 3个节点 4核16G, 200G	1.2	北京六区	vp. 0 1010111g state of	按量计费
	共 2 条							

- 2. 可在"集群管理"模块查看集群具体信息、查看 BE/FE 监控指标、配置备份恢复策略等,详细操作说明可参见 集群常规操作。
- 3. 为更方便的进行 doris 库表操作,需进入集群详情页,单击**账户管理**,创建库表并进行用户授权。单击**新增账户**,输入账户、名称后即可完成账号创建。

cdwdoris-u3i15pz1 rizon_test	账户管理		水平扩容 垂直变配 转包年包月
集群信息	① 用户可创建doris账户,并对账户赋予数据库、表粒度的管理、查询、写入等权限。		
集群监控	新規版户		请输入账户名称 Q 🗘
数据管理 ^	SEP	描述	操作
・ 备份恢复	admin	系统用户,权限不能修改,也不能删除	修改密码
・ 数据库审计			(A)110回 蚕果家口 曲(A)
本次許理	-	-	1912年2月2日 1915年1月1日 1915年1月11日 1915年1月1日 1915年1月11日 1915年1月11日 1月11日 1月11年1月11日 1月11日 1月111日 1月111日 1月1110月110101 1月1101010000000000
查询管理 配置管理 ~	-	•	PPLATARA MEMICIPE MORT
 査询管理 ・ 修改配置 			PRATATA BEECH
 査询管理 配置管理 修改配置 修改历史 			
查询管理 配置管理 · 修改配置 · 修改历史 角色管理		·	



4. 账号创建完毕后,在账户列表单击**修改权限**按钮,即可为账号赋予相应的数据权限,支持为账户授予全部库表权限或指定表的权限,详细操作请参见 账户管理 和权限管理 。

admin			系统用户、权限不能修改、也	不能删除	
	权限修改			×	修改权限重置密码
	数据库账户•	rizon	v		
	授权访问范围•	所有数据库和表 部分数据库和表		1	
	普通权限•	🔽 査询 🔽 设置① 🔽 插入①			
	高危权限🛆	✓ 创建库表 ✓ 删除库表			
		确定取消			

▲ 注意:

授权所有数据库和表后,也将获取外部数据源的权限,但外部数据源仅支持查询,不支持插入、增删库表等。

进入 CDW Studio

云数据仓库 for Apache Doris	集群列表 🕓 北京	京 2 ● 其它地域 19 ▼			
₩ 概览	新建集群				
回集群列表 R CDW Studio R	ID/名称	状态	FE节点	BE节点	内核版本
		运行中	标准型, 1个节点 4核16G, 200G	标准型, 3个节点 4核16G, 200G	1.2
	con	运行中	标准型, 1个节点 4核16G, 200G	标准型, 3个节点 4核16G, 200G	1.2
	共 2 条				

• 进入 CDW Studio 时,需指定集群和登录账号。

- Admin 用户登录后,默认有所有库表的权限。
- 其他账号登录后,只能操作自己有权限的库表(可参见上文账号管理)。

🕛 说明:

若忘记密码:

- 普通账户可在**集群详情页 > 账户管理**中直接重置密。
- Admin 账户可通过 提交工单 联系我们重置密码。



	· MERCAL II · MAN	х н _т у о н		EI • • • • • • • • • • • • • • • • • • •	
٩	 运行 	合格式化	打开查询分析Profile		
doris_audit_tbl	1 select * '	连接集群		×	
		 每个 	集群只支持登录一个用户。		
		地域	\$ 北京 3 其它地域 18	•	
		集群	∋st	•	
		账户	rilan: v	¢	
			账户智	理区	
		密码			
		数据连通性	測试 ⊘ 数据连通正常		
	执行记录 运		確定 取消		
	执行SQL: select * fr				

创建 Doris 数据库

▲ 注意:

集群创建成功后,会初始化内置catalog(Internal)和2个系统数据库,请勿对系统数据库进行增删改等操作:

- 1. 系统库 doris_audit_db___: 审计日志数据库,用于记录 Doris 系统的操作日志和安全事件。通过审计日志,可以追踪系统的操作历史和安全事件,保证系统的安全性和可靠性。
- 2. 系统库 information_schema:即 Doris的元数据数据库,可通过查询 information_schema 来获取系统的元数据信息,了解系统的结构和属性,方便进行数据管理和查询操作。

在任意库下的 SQL 编译框中输入建库的 SQL 语句,并单击运行后,即可完成建库操作:

CREATE DATABASE Sales_DB;

建库完成后,单击数据源管理中的刷新,即可在目录树中回显新建的数据库。

数据源管理 新建 🖸	欢迎页
搜索Catalog/Database 9	表 视国
▼ 🖻 internal	搜索表 Q 运行
Sales_DB(default_cluster)	▶ ⊞ doris_audit_tbl1
⊜ doris_audit_db(default_cl	<pre>rizon_table 3 CREATE DATABASE Sales_DB; 4</pre>
information_schema(default	▶ ⊞ test
😂 db(default_clus	
⊜ ∵ໍູ _≣ s_db1(default_clu	

创建 Doris 数据表

双击左侧库名,即可切换至此数据库下,然后可在弹出的 SQL 编译框中进行建表操作。



rizon_test - 北京 ▼]						
数据源管理	新建 🕄	欢迎页	Console doris_aud	lit ×	😑 Console test	Console Sales_DB	×
搜索Catalog/Database	Q	表 视图	¢	🗐 SQL 1	×		
🔻 🗬 internal		搜索表		⊙ 运行	 格式化	打开查询分析Profile	
Sales_DB(default_c	cluster)			1			
🛢 doris_audit_db(d	lefault_cl						
e information_schem	a(default						
😂 🖬 🖬 🖬 🖨 🖨 🖨	ault_clus	双击库名	,切换至此数据	库下			
e in db1(de	fault_clu						

Doris 表模型选择

Doris 支持 Aggregate、Unique 和 Duplicate三种表模型,数据模型在建表时就已经确定,且无法修改。所以,选择一个合适的数据模型非常重要。

() 说明:

具体表模型说明可参考文档: Doris 数据表和数据模型。MySQL 表同步 Doris 时的选型建议如下:

- 源端有主键的:建议创建 Unique 表模型接入数据。
- 源端没有主键的:建议创建 Duplicate 表模型接入数据。

△ 注意:

- key 列在表中的顺序需要和在 key 定义的顺序保持一致。
- String/text 类型的列不适合作为 key 列,可以转为 varchar 类型。
- 1. Aggregate 模型:可以通过预聚合,极大地降低聚合查询时所需扫描的数据量和查询的计算量,非常适合有固定模式的报表类查询场景,但该模型对 count(*)查询很不友好。

建表 Demo 如下:

```
CREATE TABLE site_visit
(
siteid INT,
city SMALLINT,
username VARCHAR(32),
pv BIGINT SUM DEFAULT '0'
)
AGGREGATE KEY(siteid, city, username)
DISTRIBUTED BY HASH(siteid) BUCKETS 10;
```

Unique 模型:针对需要唯一主键约束的场景,Unique key 相同时,新记录覆盖旧记录,可以保证主键唯一性约束,适用于有更新需求的分析业务。
 建表 Demo 如下:





RIBUTED BY HASH(orderid) BUCKETS 10;

3. Duplicate 模型:相同的行不会合并,适合任意维度的 Ad-hoc 查询。虽然无法利用预聚合的特性,但是不受聚合模型的约束,可以发挥列存模型的优势 (列裁剪、向量执行等)。

建表 Demo 如下:

CREATE TABLE se	ession_data
(
visitorid	
sessionid	
visittime	
city	
province	
ip	
brower	
url	
)	
DUPLICATE KEY(1	visitorid, sessionid)
DISTRIBUTED BY	HASH(sessionid, visitorid) BUCKETS 10;

Doris 表分区分桶建议

Doris 支持两层的数据划分,第一层是 Partition(分区),支持 Range 和 List 的划分方式。第二层是 Bucket(分桶),仅支持 Hash 的划分方式。一个 表可以不分区,但必须分桶。

() 说明:

- 具体分区、分桶说明可参考文档: 数据分区和分桶,简而言之:
- 数据量较小的场景:推荐使用一级分片表(即指不分区的分桶表)。
- 数据量较大或数据有明确日期归属的场景(如事实表):推荐使用两级分片表,即既分区又分桶的表。

```
1. 创建一级分片表的示例:
```

```
Create Table `user_login_new`
(
 `loginId` bigint NOT NULL COMMENT '登录Id',
 `userAcount` varchar(255)NOT NULL COMMENT '用户账号',
 `onlineTime` decimal(10, 2) NOT NULL DEFAULT '0.00' COMMENT '玩家在线时长,单位: 小时',
 `androidVersion` varchar
)
UNIQUE KEY (`loginid`)
COMMENT '用户归因登录表'
DISTRIBUTED BY HASH(`loginid`) buckets 8;
```

△ 注意:

创建分桶的语法: DISTRIBUTED BY HASH(`loginid`) buckets 8; 其中两个关键点: 1. 分桶列:分桶列有一定要求:必须是 key 列,类型不能是 string,text。

2. 分桶数量:根据数据大小确定,最好保证分桶后每个桶的大小在1-10G。

```
2. 创建两级分片表的示例
```

```
Create Table `user_login_beifen`
(
`loginId` bigint NOT NULL COMMENT '登录Id',
`loginTime` date NOT NULL COMMENT '登录时间',
`loginIp` varchar(255)NOT NULL COMMENT '登录ip',
```



```
`loginProvince` varchar(255)COMMENT '登录地区',
 `onlineTime` decimal(10, 2) NOT NULL DEFAULT '0.00' COMMENT '玩家在线时长, 单位: 小时'
)
UNIQUE KEY (`loginId`,`loginTime`)
COMMENT '用户归因登录表'
PARTITION BY RANGE(`loginTime`)
(
        PARTITION `p201701` VALUES LESS THAN ("2017-02-01"),
        PARTITION `p201702` VALUES LESS THAN ("2017-03-01"),
        PARTITION `p201703` VALUES LESS THAN ("2017-04-01")
)
DISTRIBUTED BY HASH(`loginId`) buckets 8;
```

△ 注意:

上为创建 Range 类型分区的语法示例,有几个关键点:

- 1. 分区列可以是一列或多列,但都需要是key列。
- 2. 创建分区时不可添加范围重叠的分区。
- 3. 数据写入前数据归属的分区要提前创建好。

```
以 Doris 也支持预先创建分区、自动创建分区,即动态分区特性:例子如下:
```

```
Create Table user_login_beifen2

(
    loginId bigint NOT NULL COMMENT '登录时间',

    loginIp varchar(255)NOT NULL COMMENT '登录时间',

    loginProvince varchar(255)COMMENT '登录地区',

    onlineTime decimal(10, 2) NOT NULL DEFAULT '0.00' COMMENT '玩家在线时长, 单位: 小时'

)

UNIQUE KEY (loginId, loginTime)

COMMENT '用户归因登录表'

PARTITION BY RANGE(loginTime)()

DISTRIBUTED BY HASH(loginId) buckets 8

PROPERTIES

(
    "dynamic_partition.time_unit" = "DAY",

    "dynamic_partition.end" = "3",

    "dynamic_partition.create_history_partition" = "true",

    "dynamic_partition.prefix" = "p"

);

--L述建表demo预先创建好今天之前10天至后3天的分区表,按loginTime列分区,分区名以'p'开头,每个分区内分成8个桶
```

Doris 索引使用建议

如果经常对某列进行精确匹配过滤并且列的基数比较高,建议在此列上创建 bloom filter 索引。

建表时尽量避免的操作

```
特别注意:截止当前最近 Doris 版本(1.2.4.2),以下表相关功能还不完善,不建议生产使用。
```

△ 注意:

- 不建议使用 "Merge-on-Write" 功能。
- 不建议使用 "auto bucket" 功能。



• 不建议使用"动态Schema表"功能。

集群配置建议

创建集群时,会初始化以下5个配置文件,说明如下:

← ● cdwdoris-u3i15pz1	修改配置			水平扩容 垂直变配	转包年包月
1201_1651					
集群信息	配置文件 上传	应用到集群 重置所有修改 重启服务			
集群监控	apache hdfs broker conf	参数名	当前运行值		操作
账户管理					
数据管理 ^	be.cont	broker_lpc_port (j)	8000		-
・ 备份恢复	fe.conf	client_expire_seconds (j)	300		-
・ 数据库审计	hdfs-site.xml	XMX ①	2g 🧨		
查询管理	odbeinst ini	新增参数			
配置管理 ^	ou bon bunn				
・ 修改配置					
・ 修改历史					
角色管理					
操作记录					

配置文件	配置建议
apache_hdfs _broker.conf	建议保持默认配置不变
be.conf	大多数配置保持默认配置不变,需特殊注意的参数如下: compaction_task_num_per_disk:每个磁盘可并发执行的 compaction 任务数量,默认值是2,如果想要提高导入速度可适当调大。具体可参考社区文档 BE 配置项 。 disable_auto_compaction=true:建议不要调整。
fe.conf	大多数配置保持默认配置不变,需特殊注意的参数如下: • max_running_txn_num_per_db: 控制同一个 DB 的并发导入个数,默认值100,当集群中有过多的导入任务正在运行时, 可适当调大。
core-site.xml	建议保持默认配置不变。
hdfs-site.xml	建议保持默认配置不变。
odbcinst.ini	建议保持默认配置不变。



配置 DataInlong 项目空间及集成资源

最近更新时间: 2024-08-15 21:18:52

创建项目空间

进入 DataInlong 控制台,单击项目列表 > 新建,创建并配置项目及所包含成员。

DataInLong	项目列表	新建项目空间		×
₽₽ 概览 ■ 集成资源	新建	项目名称•	demo_workspace	
□ 菜瓜贝湯 ■ 项目列表	dlc_test_rickywei	项目标识・	mysql2Doris	
- 芸 告警配置 ・	项目标识: 创建人/时 间: 20:20:46	描述		
	集成资源:未关联,添加关联资源 项目成员:2	▲ 高级设置		
	任务配置项目管理	项目成员 🛈	And the second	
	dsdsdsds	成员角色 🚯	☑ 项目管理员 数据工程师 运维工程师 访客	
	项目标识: 创建人/时 ²⁰²³⁻⁰⁵⁻¹⁹ 间: 17:22:59			
	集成资源:未关联,添加关联资源 项目成员:2			
	任务配置项目管理			
	test_0515			
	项目标识: 创建人时 2023-05-15 间: 10:27:39 每时姿调: 主关联 (Sein 44 PE 2021)	确定	RC/H	

购买集成资源组并关联项目

1. 进入 DataInlong 控制台,选择集成资源并单击创建。

DataInLong	集成资源 ⑤北京 ▼										
₩ 概 览	创建							请输入资源组名	日称	Q ¢	5
■ 集成资源 回 项目列表	集成资源组名称/ID	地域	网络	绑定项目 ▼	状态	资源包规格/数量	到期时间		操作		
一 		北京			离线包: 🕟 运行中	8C 16G / 1	2023-06-24 17:33:	35	关联项目 解除关联 认 续费 销毁	调整配置	
		100			实时包: 🕟 运行中	16C 64G / 1	2023-06-24 17:39:	36	关联项目 解除关联 认 续费 销毁	调整配置	
	h	北京	w. A		离线包: 🕟 运行中	8C 16G / 2	2023-06-19 04:39:4	49	关联项目 解除关联 ` 读费 销毁	调整配置	
	共 2 条						10 👻 🖞	条/页 🖂 🖣	1 /1页	► H	

2. 购买并配置集成资源。



	集成资源为离线同步、实时同步任务指审执行资源。配置详情语参见集成资源识明 46	
资源配置		1. 选择资源方案
应用场景	高线数据同步 高线+实时同步 (无队列)	
	 (反克持來紙同步任务) (成克持來紙同步任务) (成用于東京社会会院後、全/復進社童家版任務符会旁培展) (集成資源場合で保合資格経資源) (集成資源場合) (集成資源場合) (集成資源場合) (集成資源場合) (集成資源場合) (第一) 	在英國政等企务场景 发升Apport语SCH采取于总让上版目态文件数据。CDC/Binlog实时同步, 派标同步 运用于系统已正式集制 标准,在75亿级的用量数组的场场等企务场景 集成资源组合合高档资源台、实时资源化、海最认列
风格配置	◎ 真城實源包 主東州宇華協智校及這行與地別步任务	
	R注意記載	2. 配置离线/实时 ⁻ 规格
	① 案財務務員 1世界手段行気的目的任務、可適請DevoorCOC方式目前数据	
	表示包括卷 19094g	
	10.45 · 1 · 11.67.07.07.17.07.07.07.07.07.07.07.07.07.07.07.07.07	
	克波信款最 — 1 ÷	
地域		3. 设置地域/网络
网络	 О леккотуре, веляекотуре, калакистрование, калакистро Калакистрование, калакистрование, калакистрование, калакистрование, калакистрование, калакистрование, калакистро Калакистрование, калакистрование, калакистрование, калакистрование, калакистрование, калакистрование, калакистро	a
资源结名称	广州集成资源结-Stoymh	
描述	研除入用成的时间 在这	
计费类型		
计费模式	包车包月	
购买时长	1个月 2个月 3个月 4个月 5个月 6个月 7个月 8个月 9个月 1年 2年 3年 4年	
续订	※自該銀订 1000年1月1日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日	
	ne-ranocay, seasonson/HMER	
关联项目空间	1	4. 天联坝目(可辽
A0038	立想文現 對不失 に	
	第回空间 <u>北京 〜 16mp 〜</u> 2005年1月2日、地域2015年4月3日日の11日1日 11日1日1日1日1日日日11日日日11日日日11日日日11	
ι ι		

() 说明:

- 离线资源包与实时资源包可根据实际数据情况配置规格以及数量。
- 资源组网络建议和 MySQL 及 Doris 在同一个 VPC 下,若不在一个 VPC 下,可为 VPC 配置开通公网,可参见 资源组配置公网。
- 购买完成后,返回控制台并关联资源组与项目空间。
- 3. 关联资源组和项目空间

说明: 若在购买页面内已经关联资源组与项目空间,可忽略此步骤。

MANAGE AND CONTRACT							
808							
集成资源组名称/ID			協定項目 マ				38/11
					8C 16G / 1	2022-11-27 23:47:42	
					8C 16G / 1	202 10-27 22:47:27	
					16C 64G / 1		
	<i>17</i> #	关联项目		×	8C 18G / 1	2022-10-23 15:10:58	
		资源名称 资源始级 广州			16C 64G / 1		关联项目 解除关联 调整配置 续费 钢铁
	北京	項目名称 请选择 法 经 已 创建 约1	商日安间	×	8C 16G / 1		关联项目 新除关联 调整配置 续员 制限
		225年 (二) 253年 637	映日 王 同 時定 取消		16C 64G / 1		关联项目 解除关联 调整配置 续费 删除
-	北京	vpc-d3rmf9qk/subnet- 9pt7ol5l	emr_test(110225083950397 0304)	基础包: 🕟 运行中	8C 16G / 1	2022-10-16 11:37:53	

进入项目空间并注册数据源

配置 MySQL 数据源



支持注册腾讯云 MySQL 或本地自建 MySQL 数据源。进入项目管理模块,选择数据源管理 > 新建数据源 > 选择 MySQL。

• 使用腾讯云 MySQL 时:可通过配置"云实例"直接关联已购买地域数据库云实例构建数据源。

個 基本信息配置	数据源管理	数据源管理								新達Mysql数据源 ×			
2 成员管理 2 数据源管理	新建数据源 批量	世授权 批量移交						🗸 选择	类型 >	2 配置数据源			
」 项目执行资源组	数据源名称	数据源类型 ▼	类型 ▼	显示名	描述	所属项目 ▼	创建人 🍸	连接类型•	◯ 云实例	○ 连接串			
⑦ 存算引擎配置 ☑ 恋景沿景		DORIS	自定义源	-		集成项目		所属项目・	集成	Ť			
		POSTGRE	自定义源	-		集成项目		数据源名称•	MySQL_dat	asource			
								显示名	选填, 请输,	入显示名,不填默认显示数据源名称			
		MYSQL 自定义源 - wedata_dev123 集成							选填, 谓轴.	入遥述内容			
		MYSQL	自定义源	演示专用		集成							
		MYSQL	自定义源	-		集成项目		数据源权限	 项目共享 	○ 仅个人与管理员			
		MYSOI	白宗议道			集成项目		数据库名称•	л <i>и</i> л.	*			
		WIGGE	DEAM	-	-	****		用户名•	root				
		DLC	自定义源			集成		密码•					
	共 108 条							数据连通性	开始测试	⊘ 连通成功			
Ξ								上一步	保存				
参数	参数说	印											
连接类型	腾讯云	€ MySQL	数据库通过云	实例方式添	加,此方式下可	「直接获取当	当前账号)	⋝的 My	SQL 数	女据 。			
数据源名称	新建的	的数据源的名	称,由用户	自定义且不可	可为空。命名以	字母开头,	可包含字	母、数型	之、下均	l线。长度在20字符以内。			
显示名	数据源	原在产品中使	用时的显示	名称,不填黑	状认显示数据源	名称。							
描述	选填,	对本数据源	的描述。										
数据源权限	项目共	毕享表示当前	数据源项目用	所有成员均可	可使用 , 仅个人	和管理员表	示该数据	源仅创	建人和工	页目管理员可用。 			
获取实例	选择账	《户下云数据	库实例所在的	的地域、实例	列名称及 ID 信息	急。							
数据库名	填入云	实例下数据	库名称;此	数据库后续将	将作为后续数据	源的默认数	据库。						
用户	连接数	的居库的用户	名称。										
密码	连接数	如据库的密码	0										
	测试是	合能够连通	所配置的数据	居库。									
连通性		<mark>注意:</mark> • 若连通† • 如果连道 安全组,	生测试不通过 鱼性测试不通 ,	,可以继续 过,可能是	创建数据源,低 因为 WeData	3后续数据读 被数据库所	续写时会排 i在网络防	發错。 5火墙禁	止,请参	参见 添加腾讯云 MySQL 数据库			

• 使用非腾讯云 MySQL 时:通过"连接串" JDBC 方式添加自建数据库作为数据源。

E 基本信息配置	数据源管理					新建Mysql数据源				
2 成员管理	新建数据源 批量授权 批	還移交				✓ 选择类型 > 2 配置数据源				
2 数据源管理 1 项目执行资源组	数据源名称 数据源类型 ▼	类型 ▼	显示名	描述	所属项目 ▼					
 ⑦ 存算引擎配置 □ → □	DORIS	自定义源	-	-	集成项目	11日末間・ 11日本町 - 11日本町 所蔵項目・ 集成				
包 变量设置	10.101					数据源名称 * MySQL_datasource				
	POSTGRE	自定义源	-	-	集成项目	显示名 选填,请输入显示名,不填默认显示数据源名称				
	MYSQL	自定义源		wedata_dev123	集成	建 建 送頃,请给入描述内容				
	MYSQL	自定义源	演示专用	-	集成					
	MYSQL	自定义源		-	集成项目	乾癬源校開 ・				
	MYSQL	自定义源		-	集成项目	区域和网络• 广州 ▼ 数据集成研发托管账户专用 (v, ■■ ■) ▼ (
	DLC	自定义源		-	集成	JDBC URL* jdbc:mysql://host.port/database				
						東京市合約* ter ー 田戸冬・ mot				
	共 108 条									
						数据连通性 开始测试				
=										
						上一步 保存				
参数	说明									
连接类型	非腾讯云数据库到	实例可通过连	接串方式连	接。						
数据源名称	新建的数据源的得	名称,由用户	自定义且不可	可为空。命名以	以字母开头,	可包含字母、数字、下划线。长度在20字符以内。				
显示名	数据源在产品中的	吏用时的显示	名称,不填	默认显示数据》	原名称。					
描述	选填,对本数据》	原的描述。								
数据源权限	项目共享表示当前	前数据源项目	所有成员均	可使用,仅个	人和管理员表	示该数据源仅创建人和项目管理员可用。				
部署方式	 CDB: 仅适用 自建实例: 适 	用于使用腾讯 油于自建且升	云数据库。 F通 VPC 环	。 這方的 MyS(QL 集群。					
	• 公网头例: 迫	1用于升通了2	21时 IVIY2	QL 集群。						
数据库名	填入数据库名称;	此数据库后	续将作为后续	续数据源的默认	人数据库。					
用户	连接数据库的用所	白名称。								
密码	连接数据库的密码	马。								
	测试是否能够连证	通所配置的数	据库。							
连通性	▲ 注意: 若连通性器	注意: 若连通性测试不通过,可以继续创建数据源,但后续数据读写时会报错。								

配置 Doris 数据源

进入项目管理模块,选择数据源管理 > 新建数据源 > 选择 Doris,配置数据源参数并在连通性测试成功后即可保存。



数据源管理							编辑DORIS数	居源 ×
新建数据源 批量	夏权 批量移交						连接类型•	● 连接串
数据源名称	数据源类型 ▼	类型 ▼	显示名	描述	所属项目 ▼	创建人 🍸	所属项目*	Rizon测试 🔻
-	MYSQL	自定义源	-	-		-	数据源名称•	Internal
	DORIS	自定义源	-	-		-	显示名	选填,请输入显示名,不填默认显示数据源名称
-	MYSQL	自定义源	-	-	- 100		描述	选填,请输入描述内容
in insensit	DLC	系统源	-	系统源,系统自动生成				
1000 01 10 Th	HBASE	系统源	-	系统源,系统自动生成	1461		数据源权限	 项目共享 仅个人与管理员
10.000	HIVE	系统源	-	系统源,系统自动生成	1400		部署方式 *	● 自建实例 ○ 公网实例
共 6 条							区域和网络★	
							JDBC URL 🛈 *	jdbc:mysql://1 9030/de
							FE URL 🚯 •	加速をついた。
							用户名 •	安大3/6/平口1/9
							密码•	
							数据连通性	开始测试
							确认	ER:30
参数	ì	兑明						

参数	说明
数据源名称	新建的数据源的名称,由用户自定义且不可为空。命名以字母开头,可包含字母、数字、下划线。长度在20字符以内。
描述	选填,对本数据源的描述。
数据源权限	项目共享表示当前数据源项目所有成员均可使用,仅个人和管理员表示该数据源仅创建人和项目管理员可用。
部署方式	 自建实例:位于 VPC 环境内 Doris 实例。 公网实例:可使用公网访问的实例。
区域和网络	选择账户下云数据库实例所在的地域、实例名称及 ID 信息。
JDBC URL	用于连接 Doris 数据源的连接串信息: • 端口若为自建实例请填入内网IP地址和端口,多个地址间逗号(,)分隔,例如:jdbc:mysql://内网IP:port/参数。 • 若为公网实例请填入公网ip地址和端口,例如:jdbc:mysql://公网IP:port/参数。 注:上述"参数"填写数据源下的任一"数据库名称"即可,用于校验连通性。
FE URL	输入 fe http 地址,格式为:IP 地址:http 端口(无需 https:// 或 http://前缀),多个地址之间使用逗号(,)分隔, 例如:172.17.16.3:8030,172.17.16.4:8030。
用户名	连接数据源的用户名称。
密码	连接数据源的密码。
数据连通性	测试是否能够连通所配置的数据库。

配置单表实时同步任务

最近更新时间: 2024-08-15 21:18:52

数据集成支持单表同步、整库迁移两种数据同步方式,下文将介绍单表实时同步的操作方法。

Շ DatalnLong	分数据集成 → □□□										0
()集成概览	实时同步										使用指
🕒 实时同步	Bit (#1) 1 50			故主同止]	日本双條			
🖻 离线同步	整库迁移支持来源	嗜的数据及结构监控,可同步源端所有	库表 🔮 🥥	单表同步采用固定sc	chema同步的方式,仅I	同步与目标表建立		通过Agent、SDK方式	:主动上报CVM云实例、自建服务器或	Log	2
	的全量或增量数据。	,支持自动建表、字段变更同步等特性	•	映射的字段。可通过	:画布、表单两种配置方	可完成数据读取		TKE内的日志文件数据	§至外部目标端。		
	新建			新建				新建			
△ 离线运维											
◎ 告警事件	任务列表							高级	搜索 选择属性进行过滤	Q	¢
回 采集器管理	任务类型	请选择任务类型	是否已提交	是否已提交 🔻	来源类型	请选择来源类型	v	来源数据源名称 请输入来源	数据源名称		
@ 数据源管理 ピ	目标类型	请选择目标类型 ▼	目标数据源名称	请输入来源数据源名称	数据集成资源组	请选择数据集成资	■ 原組 ▼	確定 推開			
白 集成资源管理 C											
	任务名称	类型		描述		创建人		创建时间	操作		
	MySQL-Doris-画布	单表同步		-		-		2023-05-30 14:34:50	前往這维 删除		
	整库-MySQL-Doris	整库迁移						2023-05-30 14:32:33	前往這維 删除		
	MySQL-DLC	单表同步						2023-05-25 11:29:30	前往远维 删除		
	MySQL-Doris	单表同步				-		2023-05-24 16:58:55	前往运维 删除		
	共 4 条								20 * 条/页 🛛 🔺 1	/1页	F F

创建单表同步任务

1. 进入数据集成 > 实时同步页面,单击新建创建单表同步任务。

○ 集成概览	实时同步
配置中心	
🕒 实时同步	
□ 离线同步	整体正移 整体正移支持来激端的数据及结构监控,可同步激端所有库表的全量 の必要数据 支持自动建美 生的变面而决策结性 の で、 の の か の か か か か か か か か か か か か か か か
运维中心	
\bigotimes 实时运维	新建 新建
🕰 离线运维	

2. 在弹出的提示框中输入任务名称和备注,选择表单模式或画布模式后(此处 demo 选择"单表模式"),单击"创建并配置"或"仅创建"完成任务创建。

实时	同步					
	整库迁移 整库迁移支持来源的 的全量或增量数据。 新建	ll的数据及结构监控。可同步源端所有库 支持自动建表、字段变更同步等特性。		单表同步 单表同步采用固定schema同步的方式,仅同步与目标表建立 缺射的字段。可通过画布、表单两种配置方式完成数据读取		日志采集 通过Agent、SDK方式主动上报GVM云实所 TKE内的日志文件数据至外部目标端。 所注
f	E务列表		任务类型任务名称	单表同步 请输入		高级提案 这样是
			是否已常配置模式	● 表单模式 ◎ 画布模式	-	来源数据源名称 请输入来源数据源名称
			目标数: 描述	选填	-	确定 重宜
						[1] 編[1] [1]
				创建并配置仅创建		2023-05-30 14:34:50
		整库迁移		X1/K2		2023-05-30 14:32:33

编辑任务

创建完成任务后,可在任务列表页面单击新建的实时同步任务名称进入任务编辑界面,本文以单表模式为例,进行作业配置。

任务列表					高約	选择属性进行过滤				Q Ø
任务类型 目标类型	请选择任务类型 请选择目标类型	v	是否已提交 目标数据源名称	是否已提交 请输入目标数据源名称	▼ 来源类型 数据集成资源	请选择来源类型 请选择数据集成资源组	▼ ¥	来源数据源名称 请输入来) 确定 重置	源数据源名称	
任务名称		类型		描述		创建人		创建时间	操作	
ii						-		2023-06-08 15:21:04 6 days ago	前往远维删除	
MySQL-Doris-画布	MySQL-Doris-画布		步			-		2023-05-30 14:34:50 15 days ago	前往运维删除	
整库-MySQL-Doris		整库迁移						2023-05-30 14:32:33 15 days ago	前往运维删除	
MySQL-DLC	MySQL-DLC 单表同步		-		*		2023-05-25 11:29:30 20 days ago	前往运维删除		
MySQL-Doris	MySQL-Doris 单表同步		-			*		2023-05-24 16:58:55 21 days ago	前往运维删除	
共 5 条									20 ♥ 条/页 🛛 🖛 🔺 1 /	1页 🕨 🕨

配置 MySQL 信息

🔗 腾讯云

÷	MySQL	L-Doris										数据节点配置指南	j 🗹
6	⊿ 8	Ð											
1、配	置数据源			数据来源					数据	目标			
ſ	数据源	原类型	MySQL		v]	н	H	数据源类型	Doris		v	
	数据源	京	-		~				数据源	-		v	
					新建数据源							新建数据源	
	库		₽ -	Annana 194	Ŧ				库	⊳ ∗	1. 1.1.		
	表 🛈		₽ *	全选	Ŧ	共2个			表	⊳ ∗	10	v	
					添加分库分表(Ð						一键建立目标表	
	表主键	1 (i)	2 -	id					▶ 高级设置				
	格式 (i)	utf-8		v								
	读取模	莫式	• 全量	○増量									
	过滤探	影作 🚯	插入	更新删除									
	▲ 高編	级设置											
	参数		请输入者 parame	参数名称及值(格式为: iter=value),多个参数使用换行符分	搜索参数、参数说明	Q							
			。 割 如splitF	actor=xxx	▶ scan.incremental.snapshot.chunk.siz 0000 添加	e=2							
					▶ split-key.even- distribution.factor.upper-bound=10.0 添加	d							
					▶ scan.newly-added-table.enabled=tru 添加	e							
-							,						

参数	说明
数据源类型	可支持多种,此处选择 MySQL。
数据源	可选择上文"数据源管理"中已注册好的 MySQL 数据源。
库	单选,支持选择值、输入值/表达式两种方式: • 选择值时:可通过下拉列表选择在"数据源管理"中注册的数据库。 • 输入值/表达式时:可手动输入选定数据源下的已有库。



表	可多选,支持选择值、输入值/表达式两种方式: • 选择值时:可通过下拉列表选择。 • 输入值/表达式时:可手动输入。
	⚠ 注意: 选择多个表时需保证多表 schema一致,当选定的多个表的 schema 不一致时,将以选中的第一个表的 schema 为准。
≠ → 63	 支持选择值、输入值/表达式两种方式: ●选择值时:可通过下拉列表选择。 ● 输入值/表达式时:可手动输入。
表王键	⚠ 注意: 分库分表模式下默认表 schema 一致,当选定的多个表的 schema 不一致时,系统将使用拉取的第一张表的主键。
读取模式	 全量:将同步库内历史数据,全量同步结束后,会继续同步增量数据。 增量:仅从任务启动后的 binlog cdc 位点开始同步数据。
过滤操作	支持插入、更新和删除三种操作,设置后将不同步指定操作类型的数据。
高级设置	当前算子的运行参数,具体参数说明请参见 实时节点高级参数 。

配置 Doris 信息

- MySQL-Doris							数据节点配	置指南 ピ (
6 ⊿ ⊖ ∋								
1、配置数据源	数据来源				数据	目标		
数据源类型	MySQL	v	н	M	数据源类型	Doris		¥
数据源	<	Ŧ			数据源	8		×
		新建数据源					新建数排	言源
库	Þ *	v			库	Ø *		Ŧ
表 🛈	▶ ▼ 全选		共2个		表			¥
		添加分库分表()					•
表主键 🛈	🛃 🔻 id				▲ 高级设置			
格式 ③	uff_8				參数	请输入参数名称及值(格式为: parameter=value),多个参数使用换行符分	搜索参数、参数说明	Q
	di-0					割	▶ sink.properties.*=xxx 添加	
读取模式	● 全量 ◎ 増量						▶ sink.properties.columns=xxx 添加	חו
过滤操作 ① ▶ 高级设置	插入 更新 删除						▶ sink.batch.size = 500000 sink.batch.bytes= 209715200 sink.batch.interval= 10s 添加	

参数	说明
数据源类型	可支持多种,此处选择 Doris。
数据源	可选择上文"数据源管理"中已注册好的 Doris 数据源。
库	单选,支持选择值、输入值/表达式两种方式: 选择值时:可通过下拉列表选择在"数据源管理"中注册的数据库。 输入值/表达式时:可手动输入选定数据源下的已有库。
表	单选,支持选择值、输入值/表达式两种方式: 选择值时:可通过下拉列表选择。 输入值/表达式时:可手动输入。
过滤操作	支持插入、更新和删除三种操作,设置后将不同步指定操作类型的数据。



高级设置

配置表字段映射

当前算子的运行参数,具体参数说明请参见 实时节点高级参数。

此处可设置来源和目标端数据对应关系,后续任务仅同步具有映射关系的字段之间的数据。



参数	说明
同名映射	即建立源端表、目标表同名字段的映射关系。
同行映射	即根据相同的行号建立源端表、目标表字段的映射关系。
手动映射	除同名映射、同行映射的快捷方式外,还支持手动连线的方式进行字段映射。
清除映射	即清除当前已创建好的映射关系。
排序	对当前字段映射进行格式化显示,点击后具有连线关系字段将显示为一行。 说明:此排序并不变更实际表内字段顺序。
	可单击 字段配置 手动添加字段名称及类型。
字段配置	 说明: 1. MySQL / Doris 已提供直接获取数据表结构能力,您可以直接使用或查看界面内已展示出的字段。 2. 源端提供 flink 函数对字段进行转换,可在添加字段内选择"函数"类型增加转换字段写入结果中。

▲ 注意:

• 未配置映射关系的目标字段内容将为空。

• 若来源字段类型与目标字段类型间无法转换时,可能会导致任务失败。

MySQL > Doris字段映射说明如下:

类型	MySQL 数据类型	建议转化 Doris 数据类型	补充说明
	BOOLEAN	BOOLEAN	-
数值类型	TINYINT	TINYINT	-
	SMALLINT	SMALLINT	-
	MEDIUMINT	INT	-
	INT	INT	-



	BIGINT	BIGINT	-
	UNSIGNED TINYINT	SMALLINT	-
	UNSIGNED MEDIUMINT	INT	-
	UNSIGNED INT	BIGINT	-
	UNSIGNED BIGINT	LARGEINT	-
	FLOAT	FLOAT	-
	DOUBLE	DOUBLE	-
	DECIMAL	DECIMALV3	-
	YEAR	SMALLINT	-
	TIME	STRING	-
日期时间类型	DATE	DATEV2	-
	DATETIME	DATETIMEV2	-
	TIMESTAMP	DATETIMEV2	TIMESTAMP 字段数据会随着系统时区而改变但 DATETIME 字段数据不会,建议根据业务场景进行时区转化
	CHAR	CHAR	-
	VARCHAR	VARCHAR	如果 MySQL 字段长度超过65533,建议转化为 string
字符串类型	TINYTEXT、TEXT	STRING	-
	MEDIUMTEXT、 LONGTEXT	STRING	MySQL 字段长度超过1048576 字节时可能精度丢失
	TINYBLOB、BLOB	STRING	-
二进制字符串	MEDIUMBLOB、 LONGBLOB	STRING	MySQL 字段长度超过1048576 字节时可能精度丢失
	BINARY、VARBINARY	STRING	-
	JSON	STRING	MySQL 字段大小超过1M时可能精度丢失
其他	SET、BIT	STRING	-
	ENUM	UNSUPPORTED	暂不支持

配置任务属性



资源配置		
集成资源组	北京集成资源组-	т Ф
		资源联通性说明 🖸 新建集成资源组
版本	v13	
ManagerUrl		
JobManager规格	1	•
TaskManager规格	1	v
并发度 🛈	- 1 +	
运行策略		
	- 1 + 分钟	T
checkpoint间隔		
checkpoint间隔 最大重启次数 (j	1 + 次	
checkpoint间隔 最大重启次数 (〕 ▲ 高级设置	1 + 次	
checkpoint间隔 最大重启次数 ① ▲ 高级设置 参数 ③	1 + 次	搜索参数、参数说明
checkpoint间隔 最大重启次数 ④ 高级设置 参数 ④ 请输入参数名称 parameter=valu 入text 	- -1 + 次 及值(格式为: e),多个参数使用换行符	搜索参数、参数说明 (▶ taskmanager.memory.managed.frae on=0.1 添加
checkpoint间隔 最大重启次数 ③ ▲ 高级设置 参数 ③ 请输入参数名称 parameter=valu 分割	- -1 + 次 及值(格式为: 。),多个参数使用换行符	搜索参数、参数说明 (▶ taskmanager.memory.managed.frac on=0.1 添加 ▶ table.exec.sink.upsert- materialize=NONE 添加

参数	说明
集成资源组	可选择绑定此项目的集成资源组,一个任务仅可绑定一个资源组。 若未购买资源组或未绑定资源组,请先进行绑定操作。
JobManager 规 格	支持0.25、0.5、1、2CU,设置后任务将默认占用此规格。 CU 任务实际占用 CU 数= JobManager 规格 + TaskManager 规格 × 并行度。
TaskManager 规格	支持0.25、0.5、1、2CU,设置后任务将默认占用此规格 。 CU 任务实际占用 CU 数= JobManager 规格 + TaskManager 规格 × 并行度.
并发度	每个算子的默认并行度,默认1。 任务实际占用CU数= JobManager 规格 + TaskManager 规格 × 并行度.
checkpoint 间隔	当前任务提交的最大 checkpoint 间隔.
最大重启次数	设置在执行过程中发生故障时任务最大的重启阈值,若运行中重启次数超过此阈值,任务状态将置为"失败"。设置范围为 [−1,100], 阈值为0表示不重启,−1 表示不限制最大重启次数.
高级设置	设置任务级别运行参数,具体参数说明请参见 实时节点高级参数 。

任务保存与提交

1. 配置完成后,单击页面左上角的保存按钮完成配置保存,再单击提交按钮完成作业启动。



a 4 7							
1、配置数据源	数据来源			数据目	目标		
数据源类型	MySQL	▼ H	▶ 数据	源类型	Doris	v	-
数据源	1961 - 1987	~	数据	源	lates.	•	
	新建数据	源				新建数据源	
库		*	库		₽ *	Y	-
表 🛈	▶ * 全选	▼ 共2个	表		Γ> ▼ tε `	v	-
	添加分库分	ŧ.				一键建立目标表	
表主键 访	🛃 🔻 id		▲高	易级设置			
格式 ①	utf-8	*	参数		请输入参数名称及值(格式为: parameter=value),多个参数使用换行符分	搜索参数、参数说明	Q,
					割	▶ sink.properties.*=xxx 添加	
读取模式	○ 全量 当量					▶ sink.properties.columns=xxx 添加	
过滤操作 🛈	插入 更新 删除					▶ sink.batch.size = 500000 sink.batch.bytes= 209715200 sink.batch.interval= 10s 添加	
▶ 高级设置							

2. 作业启动前,会对必要配置进行校验,请确认无误后再提交。

检测 任务	到1个告警, 失败	建议您修缮	夏后再提交;	若直接提交,	可能造成	再次检测提交	直接提交
⊘ 任	务配置检测						
来	源配置	检测完成					
目	标配置	检测完成					
映	射关系配置	检测完成					
资	源组配置	检测完成					
① 资	源监测						
资	源状态检测	检测完成					
资	源余量检测	检测完成					
资	源连通性检测	警告	当前资源北京集展 致,可能会造成低 公网 C	成资源组 -kyluxpwv [壬务运行失败。请调整	2与 数据源:rizon 整数据源与资源使	_ mysql 	

配置整库实时迁移任务

最近更新时间: 2024-08-06 18:01:21

数据集成支持单表同步、整库迁移两种数据同步方式,下文将介绍**整库实时迁移**的操作方法。

C DataInLong						φ 🔋				
()集成概览	实时同步									使用指
配置中心										
 实时同步 国 离线同步 	整库迁移	罐的数据及结构监控,可同步源量	所有成表 🧯 🌍	单表同步 单表同步采用固定。	ichema同步的方式,仅同	步后日标表建立 。		日志采集 通过Agent, SDK方式:	丰动上报CVM云军例、自建服务器成	
运输中心	的全量或增量数据	。支持自动建表、字段变更同步等	特性。	取射的字段。可通过 新建	过画布、表单两种配置方式	式完成数据读取		TKE内的日志文件数据 新建	至外部目标端。	
 (5) 头时运维 (2) 离线运维 								_		
○ 告警事件	任务列表							高级	發索 选择属性进行过滤	Q Ø
回 采集器管理	任务类型	请选择任务类型	▼ 是否已提交	是否已提交	• 来源类型	请选择来源类型	▼ 来源	数据源名称 请输入来源于	数据源名称	
12 ⁵ 数据源管理 12 白 集成资源管理 12	目标类型	请选择目标类型	▼ 目标数据源名称	请输入来源数据源名称	数据集成资源组	请选择数据集成资源	e •	幕定 重置		
	任务名称	类型		描述		创建人		创建时间	操作	
	MySQL-Doris-画布	单表	司步					2023-05-30 14:34:50	前往远维 删除	
	整库-MySQL-Doris	整库;	EB	-		-		2023-05-30 14:32:33	前往运维 删除	
	MySQL-DLC	单表	9步					2023-05-25 11:29:30	前往远维 删除	
	MySQL-Doris	单表	司步	-				2023-05-24 16:58:55	前往运维 删除	
	共 4 条								20 * 条/页 H 4 1	/1页 ▶ н

创建整库同步任务

1. 进入数据集成 > 实时同步页面,单击新建创建整库迁移任务。

C DataInLong	⊕ 数据集成 ~ 〔□□ Rizon測试 ~ 〕
● 集成概览	实时同步
🕒 实时同步	数十计20 从主向上
🗉 离线同步	整件工作 整度迁移支持来源端的数据及结构监控。可同步源端所有库表的全量 或增量制张、支持自动接美、实践变更同步源端所有库表的全量
⊗ 实时运维	新建新建

2. 创建完毕后,单击任务列表中的任务名称,即可进行具体配置。

任务列表						高级搜索 选择属性进行	过滤			Q	φ
任务类型 目标类型	请选择任务类型	v	是否已提交 目标数据源名称	是否已提交 请输入目标数据源名称	▼ 来源类型 数据集成资源级	请选择来源类型	 来源数据源名積 确定 	请输入来源数据源名称 重置			
任务名称		类型		描述		创建人		创建时间	操作		
	警戒江谷						前往运维 删除				
MySQL-Doris-画布		单表同步		-		-		2023-05-30 14:34:50 15 days ago	前往远维制政		
整库-MySQL-Doris	ć	整库迁移						2023-05-30 14:32:33 15 days ago	前往远维 删除		
MySQL-DLC		单表同步						2023-05-25 11:29:30 20 days ago	前往远维 删除		
MySQL-Doris		单表同步						2023-05-24 16:58:55 21 days ago	前往远维 删除		
共 5 条									20 * 条/页	/1页 ▶	H

选择同步至 Doris 目标端的链路



1 链路选择	> ② 数据来源设置 > ③ 数据目标设置 > ④ 运行设置
链路类型	请选择来源失型 ▼ → Doris
	▲ 快速选择
	全部链路 同步至Kafka 同步至DLC 同步至StarRocks 同步至Doris 更多
	Nysql → Doris 🔂 TDSQL-C → Doris 😵 Kafka → Doris
	•支持MySQL实例级监控 •支持MySQL实例级监控 •支持MySQL实例级监控
	🔂 PostgreSQL 🛛 🔶 Doris

配置 MySQL 源端读取多张表

() 集成概览	
配置中心	
🕒 实时同步	
🖃 离线同步	✓ 链路选择 > ② 数据未源设置 > ③ 数据目标设置 > ④ 运行设置 > ⑤ 配置预览
运维中心	来源类型 MySOL
🛇 实时运维	数 (注)
🕰 离线运维	
◎ 告警事件	
设置管理	(A) POP (BLOCK)
🖂 采集器管理	选择激端库表 分别选择1个库中的2个表
@「数据源管理 12	多个关键字用竖线 1° 分隔,多个过滤标差用回车键分隔 Q wwedsta_dev
白 集成资源管理 ビ	v wedata_dev v wedata_dev
	Information_schema test_name1
	▶ mysql test_name2
	Performance_schema
	v wedata dev
	✓ test_name1
	✓ test_name2
	童歌 _ 量歌+量全 ● 法财源派
	过滤操作 ① 描入 更新 删除
	时还 Asia/Shanghai v
E	<u>L-9</u> <u>T-9</u>

参数	说明
数据源	选择需要同步的已配置好的 MySQL 数据源。
来源表	 所有库表:监控数据源下所有库。任务运行期间新增库、表默认将同步至目标端。 指定表:此选项下需指定到具体表名称,设置后任务仅同步指定表。 指定库:此选项下需指定具体库名、以表名正则表达式,设置后,任务运行期间符合表名表达式的新增表默认将同步至目标端。示例如下: 多表使用","多表匹配 ".*"表示全表匹配 "(a b)*"表示a_或者b_开头的表
读取模式	 全量 + 增量:数据同步分为全量和增量同步阶段,全量阶段完成后任务进入增量阶段。全量阶段将同步库内历史数据,增量阶段从任务 启动后 binlog cdc 的位点开始同步。 增量:仅从任务启动后的 binlog cdc 位点开始同步数据。
过滤操作	支持插入、更新和删除三种操作,设置后将不同步指定操作类型的数据。
时区	设置日志时间所属时区,默认上海。



设置 Doris 目标写入方式

💙 链路选择	> < 数据来源设置 > 3 数据目标设置 > 4 运行设置 > 5 配置预览
目标类型	Doris
数据源	sr_as_doris ▼ 新建数据源
库匹配策略	与来源库同名 自定义
表匹配策略	与来源表同名 自定义
▲ 高级设置 参数	请输入参数名称及值(格式为:parameter=value),多个参数使用换行符 分割。
上一步下	一步

参数	说明
数据源	选择已经创建的 Doris 数据源。
库/表匹配策略	设置任务运行时 Doris 中数据库以及数据表对象的名称匹配规则: 与来源库/表同名:任务运行时系统将默认在目标数据源内匹配与来源库/表同名对象。 自定义:自定义规则支持设置来源与目标之间特殊关系,例如,统一将源端库名或表名加上统一固定前缀或者后缀在写入目标 库或表任务运行时。此策略下,任务运行时系统将默认根据命名规则匹配目标对象。
高级设置 – 参数	设置 Doris 写入端的运行参数,此参数可根据业务需求配置。 Doris 端已支持参数详情请参见实时节点高级参数。

配置运行资源和策略

• 集成资源配置

为当前任务关联的集成资源组,同时设定运行时 JM、TM 规格以及任务运行并行度。其中,当前任务实际运行时实际占用 CU 数= JobManager 规格 + TaskManager 规格 × 并行度。

• 消息处理策略

参数	策略名	策略说明
DDL 消息处理	新建表	 自动建表:当来源端被监控的库中出现新建表时,Doris 端将自动创建同结构的表及字段: 若来源端表包含主键,任务默认创建 Unique key 模型表。 若来源端表包含主键,任务默认创建 Duplicate 模型表。 名略变更:目标端忽略来源端的产生的 DDL 变更消息,Doris 端及日志不做任何响应或消息提醒。 日志告警:目标端仅接收 DDL 变更消息,并在日志内打印消息内容,不触发新建表操作。 任务出错:目标端接收 DDL 变更消息并持续重启任务,重启过程中任务日志报错并出现数据写入异常。
	新增列	 新增列:当来源端被监控的库中出现表增加字段时,Doris端将自动同步新增同名字段。 忽略变更:目标端忽略来源端的产生的DDL变更消息,Doris端及日志不做任何响应或消息提醒。 日志告警:目标端仅接收DDL变更消息,并在日志内打印消息内容。此策略并不触发新增列操作。 任务出错:目标端接收DDL变更消息并持续重启任务,重启过程中任务日志报错并出现数据写入异常。
	删除表	除新建表、新增字段外其他 DDL 变更消息不支持自动响应,目前提供忽略变更、日志告警、任务出错
	重命名表	二种束略选择: 1. 忽略变更:目标端忽略来源端的产生的 DDL 变更消息,Doris 端及日志不做任何响应或消息提醒。
	删除列	2. 日志告警:目标端仅接收 DDL 变更消息,并在日志内打印消息内容。此策略并不触发新建表操作。

	重命名列	 任务出错:目标端接收 DDL 变更消息并持续重启任务,重启过程中任务日志报错并出现数据写入异常。 								
	修改列									
	删除列									
写入异常	部分停止	数据无法写入目标表时丢弃数据,后续该异常表对应的数据自动丢弃不再同步。								
	异常重启	任意表数据写入异常后任务将异常退出并自动重启。重启后任务将持续尝试写入,直到所有表均可正常 同步。重启期间可能导致部分表数据重复写入。								
	忽略异常	忽略表内无法写入的异常数据并标记为脏数据,任务继续读取并写入剩下的数据。								
阳白米尔甘己	COS 归档	写入异常策略配置为 忽略异常 时,将未写入至目标端的数据同步写入到指定的 COS 桶及文件内。								
加工女人工厅	不归档	不归档保存未写入的异常的数据。								

配置预览及提交

配置完成后可进行预览,确认后单击**保存**。

数据来源设置 编辑	
数据源	100.0
来源表	指定表
	100.0100.010.010.011.000.010.010.010
卖取模式	全量+增量
过滤操作 🛈	none
时区	Asia/Shanghai
数据目标设置 编辑	損
数据源	1.0.00
车匹配策略	与来源库同名
表匹配策略	与米源表问名
运行设置编辑	
集成资源组	1000000
JobManager规格	1
TaskManager规格	1
并行度 🚯	1
DDL消息处理策略 🛈) 新速表 忽略变更
	删除表 忽略变更
	重命名表 忽略变更
	新增列 忽略变更
	· 删除列 忽略变更



任务运维

最近更新时间:2024-07-25 11:24:21

提交任务以后,可进入**数据集成 > 实时运维**页面查看并监控当前任务状态、读写指标统计、日志及配置当前任务监控规则。

C DataInLong	会 数据集成 → □									ф 💽 📫 -
● 集成概览	实时任务运维									使用指南 IZ
配置中心 〇 实时同步	运行 暂停	¥ 继续 停止	更多操作 ▼			高级搜索	选择属性进行过	Link		Q ¢ \$
国 离线同步	任务类型	请选择任务类型	任务状态	请选择任务状态	来源类型	请选择来源类型 ▼	来源数据源名称	请输入来源数据源名称		
运维中心	目标类型	请选择目标类型	目标数据源名称	请输入目标数据源名称	数据集成资源组	请选择数据集成资源组 ▼	最近启动时间	选择时间 选择	hi 🖬	
🛇 实时运维	B 17 48 (km) 19									
🛆 离线运维	最近操作时间			第定 重重						
△ 告警事件										
设置管理	任务名称/ID	责任人 🍸	类型	同步方向	运行状态 🔻	累计读取(条) 🛈	成功写入(条) 🛈	写入延时(ms) 🛈	近一: 操作	
回 采集器管理	MySQL-Doris	ר ב	单表同步	MySQL-Doris	▶ 运行中	0	0		437 暂停 停止 到	13 •
@ 数据源管理 12										
E1 集成资源管理 C	共 1 条							10	▼条/页 H 4	1 /1页 ▶ ▶

统计指标

统计指标页面展示了任务内读写及资源运行情况。

读写指标

展示当前任务读写整体条数、全增量同步阶段、以及读写速度等。

[0]27 【	按间隔滚动 ▼ ◀ 2023-05-25 21:11:20 ~	2023-06-01 15:2	0:17 🛅 →	• 统计间隔:	一小时,	1 自动展	08fi: Ø	30秒 平																
展计读取③ 95条 元/\\`āo.13 KB	累计成功写入③ 33条 大小: 11.66 KB)						■ 11 年数編 ① 62 条 大小 3.96 KB								第1日1991年③ 6天18时9分6秒 2023-05-28 21:11:20 - 2023-06-01 15:20:18								
80																								
40 20 0 05/25 05/28 05/28 05/28 05/26 05/2 21:00 02:00 07:00 12:00 17:00 22:00	3 05/27 05/27 05/27 05/27 05/2 03:00 08:00 13:00 18:00 23:0	27 05/28 0 0 04:00 0	05/28 05/28 09:00 14:00	3 05/28 19:00	05/29 00:00	05/29 05:00	05/29 10:00	05/29 15:00	05/29 20:00	05/30 01:00	05/30 06:00	05/30 11:00	05/30 16:00	05/30 21:00	05/31 02:00	05/31 07:00	05/31 12:00	05/31 17:00	05/31 22:00	06/01 03:00	06/01 08:00	06/		
任务全增量同步进度 ▼ 数据资名称 节点D ①: mysq_1 28507_1	•				读取条数 🗕	— 写入想	fitty — 8	註数据																
#计读取# 1 个							#iii‡¤ 2.↑	以表																
				全量同	步阶段 ②	已完成	•					共 2 三 ④启动 ⑤结束	私表 1时间 2023 1时间 2023	/05/25 21 /05/25 21	:13:00 :13:52									
任务前	置环境准备⊘	>	i	協量同	歩節段 の																			

	指标参数	说明
同步概览	累计读取	本次任务运行期间,从来源端实际读取数据条数。此指标不包含筛选过滤等方式剔除的数据总量。
	累计成功写入	本次任务运行期间,已读取的数据中成功写入到目标端的数据总量。
	累计脏数据	本次任务运行期间,已读取的数据中异常写入失败的数据总量。此指标不包含任务配置中主动忽略/过滤而导

		致未写入的数据,包括指定部分停止、异常重启等运行策略,以及数据过滤等。
	累计运行时长	本次任务启动后,累计总运行时长(包含暂停时间)。
	累计读取库	本次任务运行期间,从来源端实际读取数据库数量。
	累计读取表	本次任务运行期间,从来源端实际读取数据表数量,并且分别全量同步阶段和增量同步阶段数量。
	全量/增量状态	提供未启动、进行中和已完成三种状态。
	全量同步阶段	读取源端库表中的所有记录,本阶段内仅统计读取成功且有存量业务数据的表,并且同步展示增量启动时 间、统计时间、全量结束时间。
全增量同步进度	增量同步阶段	从 binlog 消费变更数据,本阶段内仅统计读取成功且有新增业务数据的表,并且同步展示增量启动时间。
	读取速度	读取速度 = 统计间隔内总读取条数 / 统计间隔。
	读取吞吐	读取吞吐 = 统计间隔内总读取总量 / 统计间隔。
读写详情	写入速度	写入速度 = 统计间隔内成功写入条数 / 统计间隔。
	写入吞吐	写入吞吐 = 统计间隔内成功写入总量 / 统计间隔。
	写入延时	来源 Source 端至写入 Sink 端之间的链路延迟,写入延时 = 系统时间 – 记录读取时间(读取端 LatencyMarker 时间戳)。
	作业重启次数	统计间隔内当前任务重启次数。

资源指标

展示当前任务使用资源情况。



	指标参数	说明				
	上次 Checkpoint 大小	当前作业最近一次的 Checkpoint 大小。				
CheckPoin t	Checkpoint 耗时	当前作业的 Checkpoint 耗时。				
	Checkpoint 失败总次数	当前作业的 Checkpoint 的失败总次数。				
TaskMana	TaskManager CPU 使用率	当前作业 TaskManager 的 CPU 使用率。				
ger	TaskManager 堆内存使用量	当前作业 TaskManager 堆内存的用量。				
	TaskManager 老年代总 GC 次	当前作业 TaskManager 老年代 GC 次数。				



	数						
	TaskManager 老年代总 GC 时 间	当前作业 TaskManager 老年代 GC 时间。					
	TaskManager 物理内存用量	当前作业 TaskManager 所在的 JVM 的物理内存用量(RSS),包括堆内、堆外、Native 等所有区域的总内存用量。					
JobManag er	JM CPU Load	TaskManager 维度的 JVM 最近 CPU 利用率。					
	JM Head Memory	TaskManager 维度的堆内存使用情况。					
	JM GC Count	TaskManager 维度的 Status.JVM.GarbageCollector. <garbagecollector>.Count,GC(垃圾回收)次数。</garbagecollector>					
	JM GC Time	TaskManager 维度的 Status.JVM.GarbageCollector. <garbagecollector>.Time,GC(垃圾回收)时间。</garbagecollector>					

链路详情

展示整库任务下每张表的读写情况(仅整库同步时会展示此页面)。

统计指标	运行日志	配置告警								
读写指标	来源表	目标表							Q	Φ
资源指标 链路详情	数据源名称		数据库名称	表名称	成功读取条数	成功读取字节(MB)	读取速度(条/s)	读取吞吐(MB/s)	操作	
	mysql_1		di_test	test_table1	80	25.23 KB	趋势图	趋势图	查看更多	
	mysql_1		di_test	test_table2	4	1.26 KB	趋势图	趋势图	查看更多	
	mysql_2		di_test_02	test_table1	11	3.64 KB	趋势图	趋势图	查看更多	
	共3条							20 ▼ 条/页 网	(◀ 1 /1页 ▶)	M

运行日志

展示运行日志。

TaskManager	▼ cql-eijpkmc	r-761354-taskmana 🔻	近1小时 近24小	时 近7天	2023-05-31 15:3):46 至 2023-06-01 15	30:46 🖬 🕅	新		
 最近一次启动时间 	司: 2023-05-25 21:11:20,	结束时间:								
inlong-metric-s inlong-metric-s (INCREASE_PHASE {groupId=acf1fe readPhaseTimest OperatorMetricG nodeId=28507_1, numBytesIn=1292 readPhaseMetric 2023-06-01 15:3 base.util.Metri internalList=[M sinkMetricData: groupId=acf1fe numRecordSoutPa numRecordSoutPa lumRecordSoutPa 1000000000000000000000000000000000000	tates snapshot so =ReadPhaseMetricD c29-1047-462c-904 amp=1685020432118 readPhase=1}, re readPhase=1}, re sadPhase=1}, re sadPhase=1}, re sadPhaseMetricD DatAMape[}, subSO 0:28.359 [Sink: S CStateUt11s - sna etric-states', as etric-states', as etric-State{subtas} SinkTableMetricDb 3, numByteSOut=11 rSecond=0.0, numB 0:20.360 [Source: e -> Calc(select= sates SnapsMot so c29-1047-462c-904 amp=1685020432023 roup@Sc65021ab, la , readPhase=1), r 1, numByteSIn=373 S]:	<pre>urceMetricData:Sou ata(metricOata:Sou ata(metricGroup=or, 7-db7c9353b07, st), SNAPSHOT_PHASE= bels={groupId=acf1 adPhaseTimestamp=1 DataMap={}, subplace ink(table=[default pshotMetricStateFo signmentMode=UNION kIndex=0, metrics= ta{SinkMetricData{ rdh7c93553b07, st 941, numRecordSoUt PteSOutPerSecond=0 TableSourceScan(tr [CAST(meta.data_can tacMetricData'Sou ata(met</pre>	ccrEableMetricD , apache.flink. reamId=b_acfife ReadPhaseMetricD Sassource and a second ccatalog.defaul catalog.defaul catalog.defaul catalog.defaul catalog.defaul catalog.defaul catalog.defaul catalog.defaul (dirtyRecordsOu metriGroup=org reamId=b_acfife ForMeter=33, nu 0),subSinkMetricD so.OperatorMetre=33, nu 0),subSinkMetricD so.OperatorMetre=33, nu 0),subSinkMetricD so.OperatorMetre=33, nu 0),subSinkMetricD so.OperatorMetre=33, nu catalog.default_ so.OperatorMetre=34, nu so.OperatorMetre=34, nu catalog.default_ so.OperatorMetre=34, nu so.Sassource and a second catalog.default_ so.OperatorMetre=34, nu so.Sassource and a second catalog.default_ so.OperatorMetre=34, nu so.Sassource and a second catalog.default_ so.OperatorMetre=34, nu so.Sassource and a second catalog.default_ so.OperatorMetre=34, nu so.OperatorMetre=34, nu so.OperatorMet	ata (numReco runtime.met c29-1047-46 Data{metric -9047-db72 subSourceM : Partitiona Serializer t=62, numBy apache.fli c39-1047-46 BytesOutFo icMap={weda icGroup@647 database=we BytesOutFo icMap={}}} catalog, de (1/1)#0] IN ata{numReco runtime.met c29-1047-46 Data{metric -9047-db72 , subSourceMetri	<pre>set = set = s</pre>	<pre>BytesIn=27126, peratorMetricGr SS53b07, nodeId the.flink.runti mId=b_acf1fec2 rest.test_table urceTableMetri fields=[data] tateMetaInTosR ink.api.java.t numRecordsOut= rrics.groups.Op dirtyRecordsOu sink_realtime= (groupId=acf1fec dirtyRecordsOu sink_reatime= (groupId=acf1fec2 dirtyRecordsOu reatorMetricGr SS53b07, nodeId the.flink.runti reatorMetricGr SS53b07, nodeId the.flink.runti reatorMetricGr SS53b07, nodeId the.flink.runti the.fli</pre>	readPhaseMetri oup@16dccl5e, =28507_1, read me.metrics.gro =01407-462c-9b 2=SourceTableM CData(numRecor) (1/1)#0] INF rggisteredOpera yeutils.runti 33, dirtyBytes reatorMetri2Gr =948_2,3, audi t=62, dirtyByt :c29-1047-462c- =948_2,3, audi t=62, dirtyByt :c29-1047-462c- interesting :eartics.gro =1047-462c-9b able1=SourceTa	cDataMap= Labels= Phase=2}, ups. 47-db7C93553b07, etricData{numRecc dsIn=80, numBytes 0 org.apache.inl torBackendStateMw e.PojoSerializer Dut=4052}}], oupg647ce69, labe toperator=null, esOut=4052, cData{sinkMetricI 9b47-db7C93553b07; he.reader.MySqlSot DataMap= Labels= dPhase=2}, ups.	ordsIn=4, SIn=25834, long.sort. etaInfo r@65bb829d}, els= Data 7,) -> urceReader -	



配置告警

配置告警页面支持对实时任务创建监控规则及告警渠道。

统计指标	运行日志	配置告警						
创建告警规则							请知	输入规则名称 Q 🗘
规则ID \$		規则名称	规则状态 👅	告警级别 🔻	告警指标 ▼	指标阈值	告警方式	操作
		a	(1) 关闭	● 普通	1、脏数据条数 2、脏数据字节数	1、脏数据条数大于1条后触发告 警 2、脏数据字节数大于1字节后触 发告警	邮件	查看告董事件 编辑 删除
共 1 条							10 ▼ 条/页	



常见问题

最近更新时间: 2024-07-25 11:24:21

Doris 规格如何选型及调优?

请参考 Doris 资源规格选型及调优建议 。

导入任务过多,新导入任务提交报错 "current running txns on db xxx is xx, larger than limit xx"?

调整 fe 参数: max_running_txn_num_per_db,默认100,可适当调大,建议控制在500以内。

导入频率太快出现 err=[E-235] 错误?

- 参数调优建议:可通过适当调大 max_tablet_version_num 参数暂时解决,此参数默认200,建议控制在2000以内。
- 业务调优建议:降低导入频率才能根本解决这个问题。

导入文件过大,被参数限制。报错"The size of this batch exceed the max size"?

调整 be 参数: streaming_load_max_mb,建议超过需要导入的文件大小。

导入数据报错: "[-238]"?

- 原因:-238错误通常出现在同一批导入数据量过大的情况,从而导致某一个 tablet 的 Segment 文件过多(由)。
- 参数调优建议:可适当调大 BE 参数 max_segment_num_per_rowset ,此参数默认值200,可按倍数调大(如400、800),建议控制在2000以内;
- 业务调优建议:建议减少一批次导入的数据量。

导入失败,报错: "too many filtered rows xxx, "ErrorURL":"或 Insert has filtered data in strict mode, tracking url=xxxx."?

原因:表的 schema、分区等与导入的数据不匹配。可在 CDW Studio 或客户端执行 doris 命令查看具体原因: show load warnings on <tracking url> , <tracking url> 即为报错信息中返回的 error url。