云数据迁移

快速入门

文档版本 01

发布日期 2023-07-12

华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目 录

1 场景介绍

本章节介绍云数据迁移（Cloud Data Migration，以下简称CDM）的基础使用方法，通过CDM迁移RDS for MySQL数据到数据仓库服务DWS的具体操作，帮助您了解、熟悉CDM服务，具体场景如图1-1所示。

图 1-1 MySQL 迁移到 DWS

CDM的基本使用流程如下：

1. **创建CDM集群**
2. **创建连接**
3. **创建并执行作业**
4. **查看作业运行结果**

2 步骤 1：创建集群

操作场景

创建CDM集群，用于执行MySQL数据同步到DWS的任务。

说明

- 当CDM集群与其他云服务所在的区域、VPC、子网、安全组一致时，可保证CDM集群与其他云服务内网互通，无需专门打通网络。
- 当CDM集群与其他云服务所在的区域和VPC一致、但子网或安全组不一致时，需配置路由规则及安全组规则以打通网络。配置路由规则请参见**如何配置路由规则**章节，配置安全组规则请参见**如何配置安全组规则**章节。
- 当CDM集群与其他云服务所在的区域一致、但VPC不一致时，可以通过对等连接打通网络。配置对等连接请参见**如何配置对等连接**章节。

  注：如果配置了VPC对等连接，可能会出现对端VPC子网与CDM管理网重叠，从而无法访问对端VPC中数据源的情况。推荐使用公网做跨VPC数据迁移，或联系管理员在CDM后台为VPC对等连接添加特定路由。
- 当CDM集群与其他云服务所在的区域不一致时，则需要通过公网或者专线打通网络。通过公网互通时，需确保CDM集群已绑定EIP、CDM云上安全组出方向放通云下数据源所在的主机、数据源所在的主机可以访问公网且防火墙规则已开放连接端口。
- 另外，如果创建了企业项目，则企业项目也会影响CDM集群与其他云服务的网络互通，只有企业项目一致的云服务才能打通网络。

前提条件

- 已创建RDS for MySQL实例，且所在的区域、VPC、子网、安全组与CDM集群一致，如果有企业项目也必须一致。
- 已创建DWS集群，且所在的区域、VPC、子网、安全组与CDM集群一致，如果有企业项目也必须一致。
- 如果RDS for MySQL实例或DWS集群所在的区域、VPC、子网、安全组与CDM集群不一致，则需要通过网络配置、EIP或专线等方式打通与CDM集群之间的网络。

操作步骤

步骤1 登录CDM管理控制台。

图 2-1 CDM 管理控制台

说明

“创建来源”列仅通过DataArts Studio服务进入数据集成界面可以看到。

步骤2 单击“购买云数据迁移服务”，进入创建CDM集群的界面，集群配置样例如下：

- 当前区域：选择CDM集群的区域，不同区域的资源之间内网不互通，这里必须选择与MySQL实例或DWS集群所在区域一致。
- 可用区：指在同一区域下，电力、网络隔离的物理区域，可用区之内内网互通，不同可用区之间物理隔离。这里任选一个即可。
- 集群名称：集群名称在4位到64位之间，必须以字母开头，可以包含字母、数字、中划线或者下划线，不能包含其他的特殊字符，例如：“cdm-aff1”。
- 实例类型：用户按实际业务数据量选择实例规格。
  - cdm.large：8核CPU、16G内存的虚拟机，最大带宽/基准带宽为3/0.8 Gbps，集群作业并发数上限为16。
  - cdm.xlarge：16核CPU、32G内存的虚拟机，最大带宽/基准带宽为10/4 Gbps，集群作业并发数上限为32，适合使用10GE高速带宽进行TB级以上的数据量迁移。
  - cdm.4xlarge：64核CPU、128G内存的虚拟机，最大带宽/基准带宽为40/36 Gbps，集群作业并发数上限为128。
- 虚拟私有云：即VPC（Virtual Private Cloud），这里选择与MySQL实例或DWS集群相同的VPC。
- 子网：推荐与MySQL实例或DWS集群的子网一致。
- 安全组：推荐与MySQL实例或DWS集群的安全组一致。
- 企业项目：如果已经创建了企业项目，这里才可以选择。必须与MySQL实例或DWS集群的企业项目一致。
- 其它参数保持默认即可。

图 2-2 创建集群 1

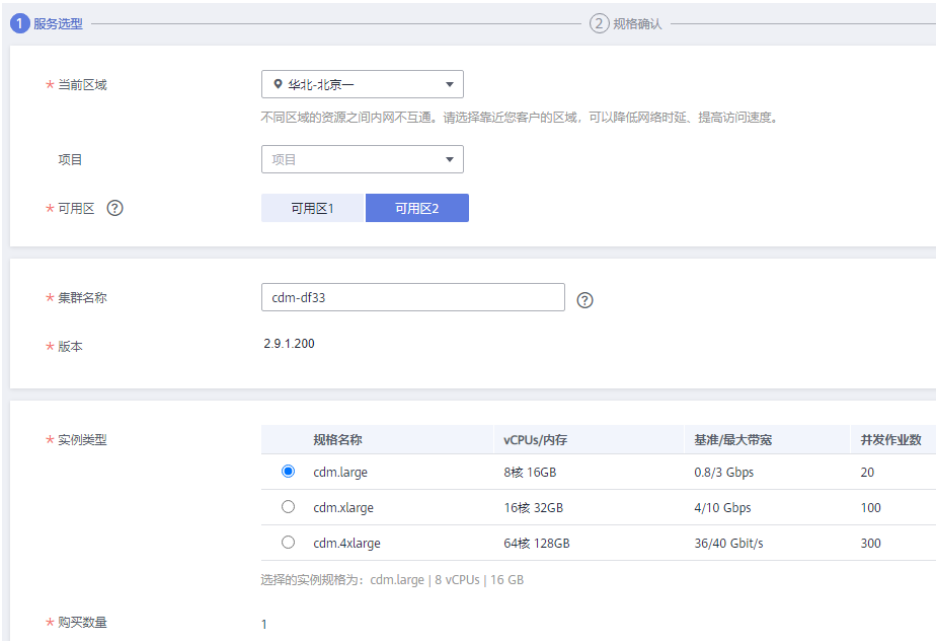


图 2-3 创建集群 2

步骤3 查看当前配置，确认无误后单击“立即购买”进入规格确认界面。

说明

集群创建好以后不支持修改规格，如果需要使用更高规格，需要重新创建。

步骤4 单击“提交”，系统开始自动创建CDM集群，在“集群管理”界面可查看创建进度。

----结束

3 步骤 2：创建连接

操作场景

迁移MySQL数据库到数据仓库服务DWS前，需要创建两个连接：

- MySQL连接：CDM连接RDS for MySQL实例。
- DWS连接：CDM连接DWS集群。

前提条件

- 已拥有RDS for MySQL实例，已获取连接MySQL数据库的数据库名称、用户名、密码，且拥有MySQL数据库的读、写和删除权限。
- 已拥有DWS集群，已获取连接DWS数据库的数据库名称、用户名、密码，且拥有DWS数据库的读、写和删除权限。
- 已参考**管理驱动**，上传了MySQL数据库驱动。

创建 MySQL 连接

步骤1 进入CDM主界面，单击左侧导航上的“集群管理”，找到**步骤1：创建集群**章节创建的集群“cdm-aff1”。

步骤2 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如**图3-1**所示。

图 3-1 选择连接器类型

步骤3 选择“云数据库 MySQL”后单击“下一步”，配置云数据库 MySQL连接的参数。

图 3-2 创建 MySQL 连接

单击“显示高级属性”可查看更多可选参数，具体请参见**配置云数据库MySQL/MySQL数据库连接**。这里保持默认，必填参数如**表3-1**所示。

表 3-1 MySQL 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	mysqllink
数据库服务器	MySQL数据库的IP地址或域名。	-
端口	MySQL数据库的端口。	3306

参数名	说明	取值样例
数据库名称	MySQL数据库的名称。	sqoop
用户名	拥有MySQL数据库的读、写和删除权限的用户。	admin
密码	用户的密码。	-
使用本地API	使用数据库本地API加速（系统会尝试启用MySQL数据库的local_infile系统变量）。	是
使用Agent	是否选择通过Agent从源端提取数据。	否
local_infile字符集	MySQL通过local_infile导入数据时，可配置编码格式。	utf8
驱动版本	CDM连接关系数据库前，需要先上传所需关系数据库的JDK8版本.jar格式驱动。MySQL的驱动请从**https://downloads.mysql.com/archives/c-j/**选择5.1.48版本下载，从中获取mysql-connector-java-5.1.48.jar，然后进行上传。	-

步骤4 单击“测试”测试参数是否配置无误，“测试”成功后单击“保存”创建该连接，并回到连接管理界面。

图 3-3 创建 MySQL 连接成功

----结束

创建 DWS 连接

步骤1 单击CDM集群后的“作业管理”，进入作业管理界面，再选择“连接管理 > 新建连接”，进入选择连接器类型的界面，如图3-4所示。

图 3-4 选择连接器类型

步骤2 连接器类型选择“数据仓库服务（DWS）”后单击“下一步”配置DWS连接参数。

图 3-5 创建 DWS 连接

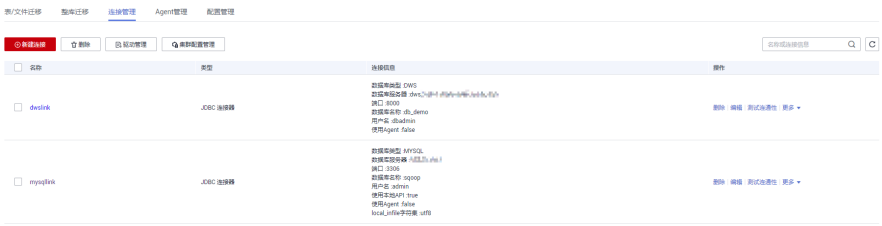
单击“显示高级属性”可查看更多可选参数，具体请参见**配置关系数据库连接**。必填参数如**表3-2**所示，可选参数保持默认即可。

表 3-2 DWS 连接参数

参数名	说明	取值样例
名称	输入便于记忆和区分的连接名称。	dwslink
数据库服务器	DWS数据库的IP地址或域名。	-
端口	DWS数据库的端口。	8000
数据库名称	DWS数据库的名称。	db_demo
用户名	拥有DWS数据库的读、写和删除权限的用户。	dbadmin
密码	用户的密码。	-
使用Agent	是否选择通过Agent从源端提取数据。	否

步骤3 单击“测试”测试参数是否配置无误，“测试”成功后单击“保存”创建该连接，并回到连接管理界面。

图 3-6 创建 DWS 连接成功



----结束

4 步骤 3：创建并执行作业

操作场景

创建CDM迁移数据表的作业，执行从MySQL数据库迁移表到DWS的任务。

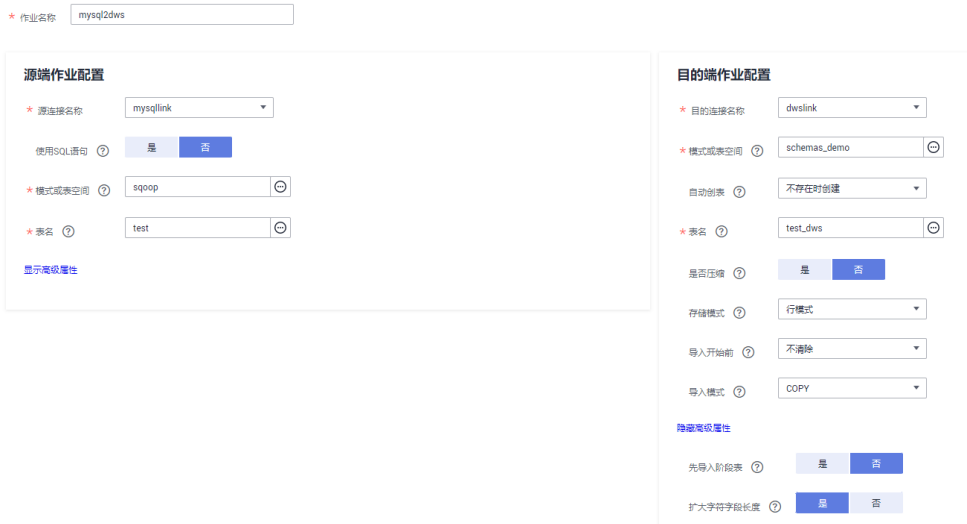
操作步骤

步骤1 在集群管理界面，找到**步骤1：创建集群**章节创建的集群“cdm-aff1”。

步骤2 单击该CDM集群后的“作业管理”，进入作业管理界面。

步骤3 选择“表/文件迁移 > 新建作业”，配置作业基本信息。

图 4-1 新建作业



- 作业名称：输入便于记忆、区分的作业名称，例如：“mysql2dws”。
- 源端作业配置
  - 源连接名称：选择**步骤2：创建连接**章节中创建的MySQL连接“mysqllink”。
  - 使用SQL语句：选择“否”。

  - 模式或表空间：选择从MySQL的哪个数据库导出表。
  - 表名：选择导出哪张表。
  - 其他可选参数保持默认即可，详细说明可参见**配置MySQL源端参数**。
- 目的端作业配置
  - 目的连接名称：选择**步骤2：创建连接**章节中创建的DWS连接“dwslink”。
  - 模式或表空间：选择导入到DWS的哪个模式。
  - 自动创表：这里选择“不存在时创建”，当下面“表名”参数中配置的表不存在时，CDM会自动在DWS数据库中创建该表。
  - 表名：选择导入到DWS数据库的哪张表。
  - 高级属性参数-“扩大字符字段长度”：这里选择“是”。由于MySQL和DWS存储中文时编码不一样，所需的长度也不一样，一个中文字符在UTF-8编码下可能要占3个字节。该参数选择为“是”后，在DWS中自动创表时，会将字符类型的字段长度设置为原表的3倍，避免出现DWS表的字符字段长度不够的报错。
  - 其他可选参数保持默认即可，详细说明可参见**配置DWS目的端参数**。

步骤4 单击“下一步”进入字段映射界面，CDM会自动匹配源端和目的端的数据表字段，需用户检查字段映射关系是否正确。

- 如果字段映射关系不正确，用户单击字段所在行选中后，按住鼠标左键可拖拽字段来调整映射关系。
- 导入到DWS时需要手动选择DWS的分布列，建议按如下顺序选取：
  a. 有主键可以使用主键作为分布列。
  b. 多个数据段联合做主键的场景，建议设置所有主键作为分布列。
  c. 在没有主键的场景下，如果没有选择分布列，DWS会默认第一列作为分布列，可能会有数据倾斜风险。
- 如果需要转换源端字段内容，可在该步骤配置，详细请参见**字段转换**，这里选择不进行字段转换。

图 4-2 字段映射

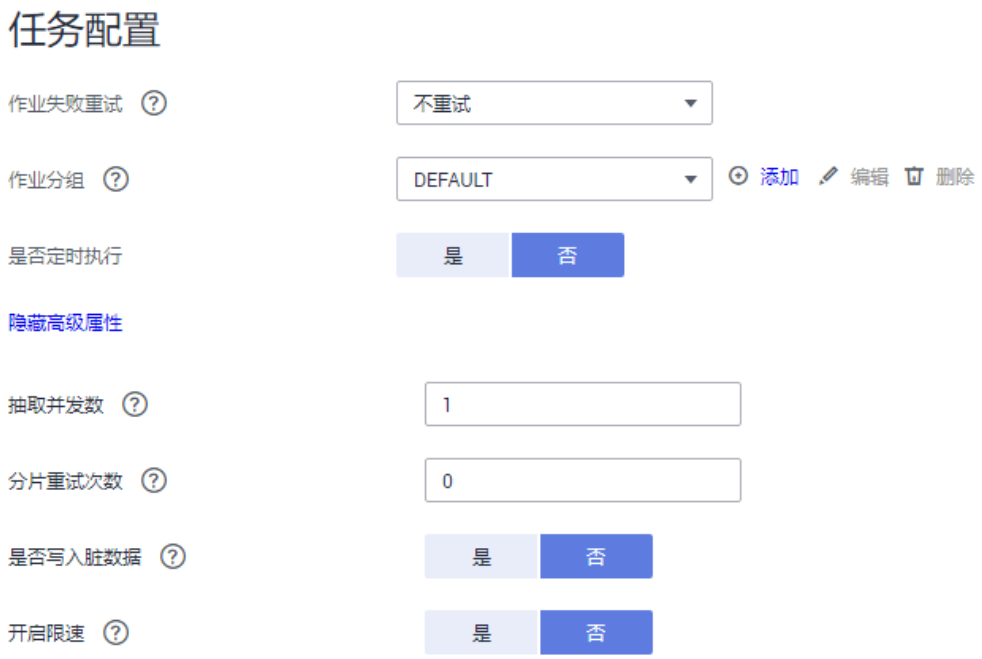
步骤5 单击“下一步”配置任务参数，一般情况下全部保持默认即可。

该步骤用户可以配置如下可选功能：

- 作业失败重试：如果作业执行失败，可选择是否自动重试，这里保持默认值“不重试”。
- 作业分组：选择作业所属的分组，默认分组为“DEFAULT”。在CDM“作业管理”界面，支持作业分组显示、按组批量启动作业、按分组导出作业等操作。
- 是否定时执行：如果需要配置作业定时自动执行可开启。这里保持默认值“否”。
- 抽取并发数：设置同时执行的抽取任务数，适当的抽取并发数可以提升迁移效率，配置原则请参见**性能调优**。这里保持默认值“1”。

- 是否写入脏数据：如果需要将作业执行过程中处理失败的数据、或者被清洗过滤掉的数据写入OBS中，以便后面查看，可通过该参数配置，写入脏数据前需要在CDM先配置好OBS连接。这里保持默认值“否”即可，不记录脏数据。

图 4-3 任务配置



步骤6 单击“保存并运行”，CDM立即开始执行作业。

图 4-4 作业执行

----结束

5 步骤 4：查看作业运行结果

操作场景

作业完成后，可以查看作业执行结果及最近90天内的历史信息，包括写入行数、读取行数、写入字节、写入文件数和日志等信息。

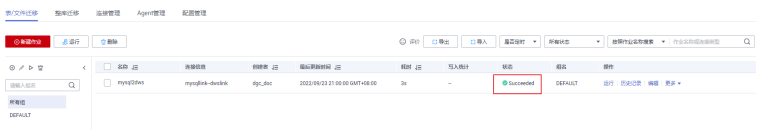
操作步骤

步骤1 在集群管理界面，找到**步骤1：创建集群**章节创建的集群“cdm-aff1”。

步骤2 单击该CDM集群后的“作业管理”，进入作业管理界面。

步骤3 找到**步骤3：创建并执行作业**章节创建的作业“mysql_dws”，查看该作业的执行状态。作业状态为Succeeded即迁移成功。

图 5-1 作业状态



说明

作业状态有New，Pending，Booting，Running，Failed，Succeeded，stopped。

其中“Pending”表示正在等待系统调度该作业，“Booting”表示正在分析待迁移的数据。

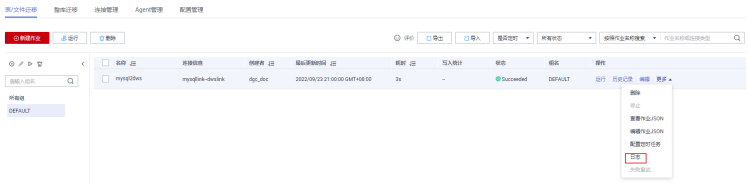
步骤4 单击作业后面的“历史记录”，可查看作业的写入行数、读取行数、写入字节或写入文件数。

图 5-2 查看历史记录

步骤5 在历史记录界面，再单击“日志”可查看作业执行的日志信息。

也可以在作业列表界面，选择“更多 > 日志”来查看该作业最近的一次日志。

图 5-3 查看作业日志



----结束

6 入门实践

当您参考**创建集群**、**创建连接**等一系列操作后，可以根据自身的业务需求使用CDM提供的一系列常用实践。

表 6-1 常用最佳实践

实践		描述
使用教程	**创建MRS Hive连接器**	MRS Hive连接适用于MapReduce服务，本最佳实践为您介绍如何创建MRS Hive连接器。
	MySQL数据迁移到OBS	CDM支持表到OBS的迁移，本最佳实践介绍如何通过CDM将MySQL表数据迁移到OBS中。
参数传递	**通过数据开发使用参数传递灵活调度CDM作业**	如果CDM作业接收来自数据开发作业配置的参数，则在数据开发模块可以使用诸如EL表达式传递动态参数来调度CDM作业。本最佳实践介绍通过数据开发使用参数传递功能灵活调度CDM作业。
增量迁移	**文件增量迁移**	CDM支持对文件类数据源进行增量迁移，全量迁移完成之后，第二次运行作业时可以导出全部新增的文件，或者只导出特定的目录/文件。
	关系数据库增量迁移	CDM支持对关系型数据库进行增量迁移，全量迁移完成之后，可以增量迁移指定时间段内的数据（例如每天晚上0点导出前一天新增的数据）。
案例	**案例：贸易数据统计与分析**	使用云数据迁移（Cloud Data Migration，简称CDM）将本地贸易统计数据导入到OBS，再使用数据湖探索（Data Lake Insight，简称DLI）进行贸易统计分析，帮助H咨询公司以极简、极低成本构建其大数据分析平台，使得该公司更好地聚焦业务，持续创新。

实践		描述
	案例：车联网大数据业务上云	为搭建H公司车联网业务集团级的云管理平台，统一管理、部署硬件资源和通用类软件资源，实现IT应用全面服务化、云化，CDM（Cloud Data Migration，简称CDM）助力H公司做到代码“0”改动、数据“0”丢失迁移上云。