云数据迁移

产品介绍

文档版本 1.0

发布日期 2023-06-21

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。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 图解云数据迁移

2 什么是云数据迁移

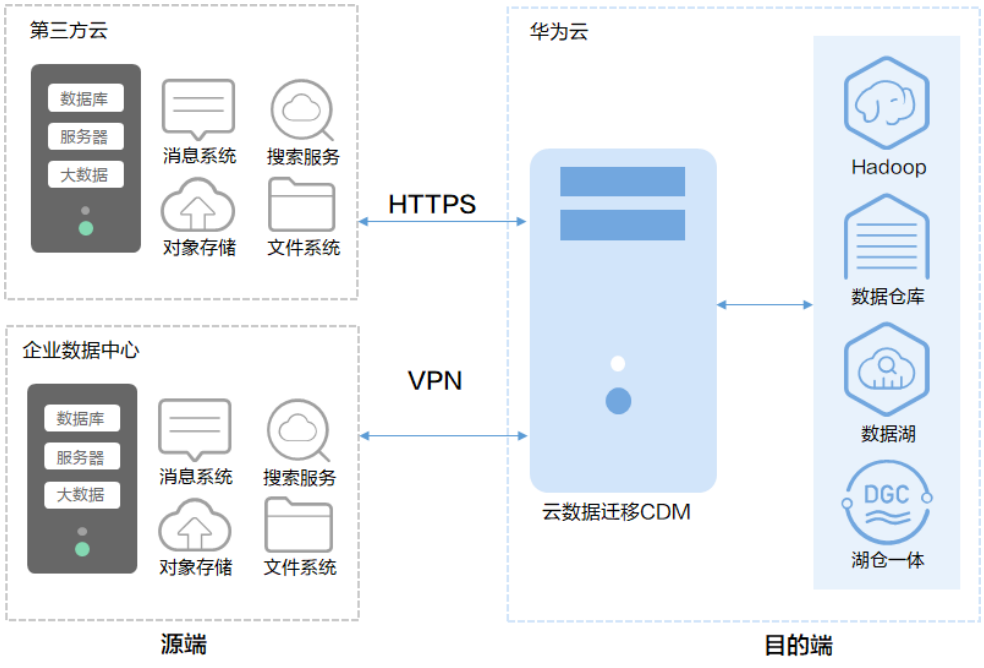
产品定义

云数据迁移（Cloud Data Migration, 简称CDM），是一种高效、易用的数据集成服务。 CDM围绕大数据迁移上云和智能数据湖解决方案，提供了简单易用的迁移能力和多种数据源到数据湖的集成能力，降低了客户数据源迁移和集成的复杂性，有效的提高您数据迁移和集成的效率。

在数据治理中心（DataArts Studio）服务中，CDM作为其中的“数据集成”组件使用，产品能力与独立的CDM服务保持一致。因此，后文中的“云数据迁移”、“数据集成”均指CDM服务。

CDM服务基于分布式计算框架，利用并行化处理技术，支持用户稳定高效地对海量数据进行移动，实现不停服数据迁移，快速构建所需的数据架构。

图 2-1 CDM 定位



产品功能

- **表/文件/整库迁移**

  支持批量迁移表或者文件，还支持同构/异构数据库之间整库迁移，一个作业即可迁移几百张表。

- **增量数据迁移**

  支持文件增量迁移、关系型数据库增量迁移、HBase/CloudTable增量迁移，以及使用Where条件配合时间变量函数实现增量数据迁移。

- **事务模式迁移**

  支持当CDM作业执行失败时，将数据回滚到作业开始之前的状态，自动清理目的表中的数据。

- **字段转换**

  支持去隐私、字符串操作、日期操作等常用字段的数据转换功能。

- **文件加密**

  在迁移文件到文件系统时，CDM支持对写入云端的文件进行加密。

- **MD5校验一致性**

  支持使用MD5校验，检查端到端文件的一致性，并输出校验结果。

- **脏数据归档**

  支持将迁移过程中处理失败的、被清洗过滤掉的、不符合字段转换或者不符合清洗规则的数据单独归档到脏数据日志中，便于用户查看。并支持设置脏数据比例阈值，来决定任务是否成功。

3 产品优势

用户在云上进行数据集成、数据备份、新应用开发时，经常会涉及到数据迁移。通常情况下用户要进行数据迁移，会开发一些数据迁移脚本，从源端读取数据再写入目的端，相对这样传统的做法，CDM的优势如表3-1所示。

表 3-1 CDM 优势

优势项	用户自行开发	CDM
易使用	自行准备服务器资源，安装配置必要的软件并进行配置，等待时间长。 程序在读写两端会根据数据源类型，使用不同的访问接口。一般是数据源提供的对外接口，例如JDBC、原生API等，因此在开发脚本时需要依赖大量的库、SDK等，开发管理成本较高。	CDM提供了Web化的管理控制台，通过Web页实时开通服务。 用户只需要通过可视化界面对数据源和迁移任务进行配置，服务会对数据源和任务进行全面的管理和维护。用户只需关注数据迁移的具体逻辑，而不用关心环境等问题，极大降低了开发维护成本。 CDM还提供了REST API，支持第三方系统调用和集成。
实时监控	需要自行选型开发。	您可以使用云监控服务监控您的CDM集群，执行自动实时监控、告警和通知操作，帮助您更好地了解CDM集群的各项性能指标。
免运维	需要自行开发完善运维功能，自行保证系统可用性，尤其是告警及通知功能，否则只能人工值守。	使用CDM服务，用户不需要维护服务器、虚拟机等资源。CDM的日志，监控和告警功能，有异常可以及时通知相关人员，避免7X24小时人工值守。
高效率	在迁移过程中，数据读写过程都是由一个单一任务完成的，受限于资源，整体性能较低，对于海量数据场景通常不能满足要求。	CDM任务基于分布式计算框架，自动将任务切分为独立的子任务并行执行，能够极大提高数据迁移的效率。针对Hive、HBase、MySQL、DWS（数据仓库服务）数据源，使用高效的数据导入接口导入数据。

优势项	用户自行开发	CDM
多种数据源支持	数据源类型繁杂，针对不同数据源开发不同的任务，脚本数量成千上万。	支持数据库、Hadoop、NoSQL、数据仓库、文件等多种类型的数据源，具体数据类型请参见**支持的数据源**。
多种网络环境支持	随着云计算技术的发展，用户数据可能存在于各种环境中，例如公有云、自建/托管IDC（Internet Data Center，互联网数据中心）、混合场景等。在异构环境中进行数据迁移需要考虑网络连通性等因素，给开发和维护都带来较大难度。	无论数据是在用户本地自建的IDC中、云服务中、第三方云中，或者使用弹性云服务器（Elastic Cloud Server，ECS）自建的数据库或文件系统中，CDM均可帮助用户轻松应对各种数据迁移场景，包括数据上云，云上数据交换，以及云上数据回流本地业务系统。

4 支持的数据源

4.1 支持的数据源（2.9.3.300）

数据集成有两种迁移方式，支持的数据源有所不同：

- 表/文件迁移：适用于数据入湖和数据上云场景下，表或文件级别的数据迁移，请参见**表/文件迁移支持的数据源类型**。
- 整库迁移：适用于数据入湖和数据上云场景下，离线或自建数据库整体迁移场景，请参见**整库迁移支持的数据源类型**。

说明

本文介绍2.9.3.300版本CDM集群所支持的数据源。因各版本集群支持的数据源有所差异，其他版本支持的数据源仅做参考。

表/文件迁移支持的数据源类型

表/文件迁移可以实现表或文件级别的数据迁移。

表/文件迁移时支持的数据源如**表4-1**所示。

表 4-1 表/文件迁移支持的数据源

数据源分类	源端数据源	对应的目的端数据源	说明
数据仓库	数据仓库服务（DWS）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS ClickHouse • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	不支持DWS物理机纳管模式。
	数据湖探索（DLI）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS ClickHouse • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable），MongoDB • 搜索：Elasticsearch，云搜索服务（CSS）	MongoDB建议使用的版本：4.2。
	MRS ClickHouse	数据仓库：MRS ClickHouse，数据湖探索（DLI）	MRS ClickHouse建议使用的版本：21.3.4.X。

数据源分类	源端数据源	对应的目的端数据源	说明
Hadoop	MRS HDFS MRS HBase	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 支持本地存储，仅MRS Hive、MRS Hudi支持存算分离场景。 • 仅MRS Hive支持Ranger场景。 • 不支持ZK开启SSL场景。 • MRS HDFS建议使用的版本： – 2.8.X – 3.1.X • MRS HBase建议使用的版本： – 2.1.X – 1.3.X • MRS Hive、MRS Hudi暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X
	MRS Hive	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS Clickhouse • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	MRS Hudi	• 数据仓库：数据仓库服务（DWS） • Hadoop：MRS HBase	

数据源分类	源端数据源	对应的目的端数据源	说明
	FusionInsight HDFS FusionInsight HBase FusionInsight Hive	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• FusionInsight数据源不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。 • FusionInsight HDFS建议使用的版本： – 2.8.X – 3.1.X • FusionInsight HBase建议使用的版本： – 2.1.X – 1.3.X • FusionInsight Hive建议使用的版本： – 1.2.X – 3.1.X

数据源分类	源端数据源	对应的目的端数据源	说明
	Apache HBase Apache Hive Apache HDFS	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• Apache数据源不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。 • Apache HBase建议使用的版本： – 2.1.X – 1.3.X • Apache Hive暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X • Apache HDFS建议使用的版本： – 2.8.X – 3.1.X
对象存储	对象存储服务（OBS）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 对象存储服务之间的迁移，推荐使用对象存储迁移服务OMS。 • 不支持二进制文件导入到数据库或NoSQL。

数据源分类	源端数据源	对应的目的端数据源	说明
文件系统	FTP SFTP	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS） • 对象存储：对象存储服务（OBS）	• 文件系统不支持作为目的端。 • FTP/SFTP到搜索的迁移仅支持如CSV等文本文件，不支持二进制文件。 • FTP/SFTP到OBS的迁移仅支持二进制文件。 • HTTP到OBS的迁移推荐使用obsutil工具，请参见**obsutil简介**。
	HTTP	Hadoop：MRS HDFS	
关系型数据库	云数据库MySQL	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive，MRS Hudi • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 关系型数据库：云数据库MySQL，云数据库 PostgreSQL，云数据库 SQL Server • 搜索：Elasticsearch，云搜索服务（CSS）	• OLTP数据库之间的迁移推荐通过数据复制服务DRS进行迁移。 • 云数据库 MySQL不支持SSL模式。 • Microsoft SQL Server建议使用的版本：2005以上。 • 金仓和GaussDB数据源可通过PostgreSQL连接器进行连接，支持的迁移作业的源端、目的端情况与PostgreSQL数据源一致。
	云数据库SQL Server 云数据库PostgreSQL	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 关系型数据库：云数据库MySQL，云数据库 PostgreSQL，云数据库 SQL Server • 搜索：Elasticsearch，云搜索服务（CSS）	

数据源分类	源端数据源	对应的目的端数据源	说明
	MySQL PostgreSQL Oracle	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive，MRS Hudi • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	Microsoft SQL Server	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	

数据源分类	源端数据源	对应的目的端数据源	说明
	SAP HANA	● 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） ● Hadoop：MRS Hive	SAP HANA数据源存在如下约束： ● SAP HANA不支持作为目的端。 ● 仅支持2.00.050.00.1592305219版本。 ● 仅支持Generic Edition。 ● 不支持BW/4 FOR HANA。 ● 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 ● 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 ● 迁移时不支持目的端自动建表。
	分库	● 数据仓库：数据湖探索（DLI） ● Hadoop：MRS HBase，MRS Hive ● 搜索：Elasticsearch，云搜索服务（CSS） ● 对象存储：对象存储服务（OBS）	分库数据源不支持作为目的端。
	神通（ST）	● Hadoop：MRS Hive，MRS Hudi	-

数据源分类	源端数据源	对应的目的端数据源	说明
NoSQL	分布式缓存服务（DCS）	Hadoop：MRS HDFS，MRS HBase，MRS Hive	除了表格存储服务（CloudTable）外，其他NoSQL数据源不支持作为目的端。 Redis到DCS的迁移，可以通过其他方式进行，请参见**自建Redis迁移至DCS**。
	Redis		
	文档数据库服务（DDS）		
	MongoDB		
	表格存储服务（CloudTable HBase）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	Cassandra	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
消息系统	数据接入服务（DIS）	搜索：云搜索服务（CSS）	消息系统不支持作为目的端。
	Apache Kafka		
	DMS Kafka		

数据源分类	源端数据源	对应的目的端数据源	说明
	MRS Kafka	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• MRS Kafka不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。
搜索	Elasticsearch	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	Elasticsearch仅支持非安全模式。
	云搜索服务（CSS）		导入数据到CSS推荐使用Logstash，请参见**使用Logstash导入数据到Elasticsearch**。

说明

上表中非云服务的数据源，例如MySQL，既可以支持用户本地数据中心自建的MySQL，也可以是用户在ECS上自建的MySQL，还可以是第三方云的MySQL服务。

整库迁移支持的数据源类型

整库迁移适用于将本地数据中心或在ECS上自建的数据库，同步到云上的数据库服务或大数据服务中，适用于数据库离线迁移场景，不适用于在线实时迁移。

数据集成支持整库迁移的数据源如**表4-2**所示。

表 4-2 整库迁移支持的数据源

数据源分类	数据源	读取	写入	说明
数据仓库	数据仓库服务（DWS）	支持	支持	-
Hadoop （仅支持本地存储，不支持存算分离场景，不支持Ranger场景，不支持ZK开启SSL场景）	MRS HBase	支持	支持	整库迁移仅支持导出到MRS HBase。 建议使用的版本： ● 2.1.X ● 1.3.X
	MRS Hive	支持	支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
	FusionInsight HBase	支持	不支持	建议使用的版本： ● 2.1.X ● 1.3.X
	FusionInsight Hive	支持	不支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
	Apache HBase	支持	不支持	建议使用的版本： ● 2.1.X ● 1.3.X
	Apache Hive	支持	不支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X

数据源分类	数据源	读取	写入	说明
	MRS Hudi	支持	支持	支持本地存储、存算分离场景。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
关系数据库	云数据库 MySQL	支持	支持	不支持OLTP到OLTP迁移，此场景推荐通过数据复制服务DRS进行迁移。
	云数据库 PostgreSQL	支持	支持	
	云数据库 SQL Server	支持	支持	
	MySQL	支持	不支持	
	PostgreSQL	支持	不支持	
	Microsoft SQL Server	支持	不支持	
	Oracle	支持	不支持	
	SAP HANA	支持	不支持	● 仅支持2.00.050.00.1592305219版本。 ● 仅支持Generic Edition。 ● 不支持BW/4 FOR HANA。 ● 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 ● 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 ● 迁移时不支持目的端自动建表。

数据源分类	数据源	读取	写入	说明
	达梦数据库 DM	支持	不支持	仅支持导出到DWS、Hive
NoSQL	分布式缓存服务（DCS）	不支持	支持	仅支持MRS到DCS迁移。
	文档数据库服务（DDS）	支持	支持	仅支持DDS和MRS之间迁移。
	表格存储服务（CloudTable）	支持	支持	-

4.2 支持的数据源（2.9.2.200）

数据集成有两种迁移方式，支持的数据源有所不同：

- 表/文件迁移：适用于数据入湖和数据上云场景下，表或文件级别的数据迁移，请参见**表/文件迁移支持的数据源类型**。
- 整库迁移：适用于数据入湖和数据上云场景下，离线或自建数据库整体迁移场景，请参见**整库迁移支持的数据源类型**。

说明

本文介绍2.9.2.200版本CDM集群所支持的数据源。因各版本集群支持的数据源有所差异，其他版本支持的数据源仅做参考。

表/文件迁移支持的数据源类型

表/文件迁移可以实现表或文件级别的数据迁移。

表/文件迁移时支持的数据源如**表4-3**所示。

表 4-3 表/文件迁移支持的数据源

数据源分类	源端数据源	对应的目的端数据源	说明
数据仓库	数据仓库服务（DWS）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS ClickHouse • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	不支持DWS物理机纳管模式。
	数据湖探索（DLI）		-
	MRS ClickHouse	数据仓库：MRS ClickHouse，数据湖探索（DLI）	MRS ClickHouse建议使用的版本：21.3.4.X。
Hadoop	MRS HDFS	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 支持本地存储，仅MRS Hive、MRS Hudi支持存算分离场景。 • 仅MRS Hive支持Ranger场景。 • 不支持ZK开启SSL场景。 • MRS HDFS建议使用的版本： – 2.8.X – 3.1.X • MRS HBase建议使用的版本： – 2.1.X – 1.3.X • MRS Hive、MRS Hudi暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X
	MRS HBase		

数据源分类	源端数据源	对应的目的端数据源	说明
	MRS Hive	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI），MRS Clickhouse • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	MRS Hudi	数据仓库：数据仓库服务（DWS）	
	FusionInsight HDFS FusionInsight HBase FusionInsight Hive	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• FusionInsight数据源不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。 • FusionInsight HDFS建议使用的版本： – 2.8.X – 3.1.X • FusionInsight HBase建议使用的版本： – 2.1.X – 1.3.X • FusionInsight Hive建议使用的版本： – 1.2.X – 3.1.X

数据源分类	源端数据源	对应的目的端数据源	说明
	Apache HBase Apache Hive Apache HDFS	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• Apache数据源不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。 • Apache HBase建议使用的版本： – 2.1.X – 1.3.X • Apache Hive暂不支持2.x版本，建议使用的版本： – 1.2.X – 3.1.X • Apache HDFS建议使用的版本： – 2.8.X – 3.1.X
对象存储	对象存储服务（OBS）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 对象存储服务之间的迁移，推荐使用对象存储迁移服务OMS。 • 不支持二进制文件导入到数据库或NoSQL。

数据源分类	源端数据源	对应的目的端数据源	说明
文件系统	FTP SFTP	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• 文件系统不支持作为目的端。 • FTP/SFTP到搜索的迁移仅支持如CSV等文本文件，不支持二进制文件。 • HTTP到OBS的迁移推荐使用obsutil工具，请参见**obsutil简介**。
	HTTP	Hadoop：MRS HDFS	
关系型数据库	云数据库MySQL	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive，MRS Hudi • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 关系型数据库：云数据库MySQL，云数据库 PostgreSQL，云数据库 SQL Server • 搜索：Elasticsearch，云搜索服务（CSS）	• OLTP数据库之间的迁移推荐通过数据复制服务DRS进行迁移。 • 云数据库 MySQL不支持SSL模式。 • Microsoft SQL Server建议使用的版本：2005以上。 • 金仓和GaussDB数据源可通过PostgreSQL连接器进行连接，支持的迁移作业的源端、目的端情况与PostgreSQL数据源一致。
	云数据库SQL Server 云数据库PostgreSQL	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 关系型数据库：云数据库MySQL，云数据库 PostgreSQL，云数据库 SQL Server • 搜索：Elasticsearch，云搜索服务（CSS）	

数据源分类	源端数据源	对应的目的端数据源	说明
	MySQL PostgreSQL Oracle	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive，MRS Hudi • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	Microsoft SQL Server	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	

数据源分类	源端数据源	对应的目的端数据源	说明
	SAP HANA	• 数据仓库：数据湖探索（DLI） • Hadoop：MRS Hive	SAP HANA数据源存在如下约束： • SAP HANA不支持作为目的端。 • 仅支持2.00.050.00.1592305219版本。 • 仅支持Generic Edition。 • 不支持BW/4 FOR HANA。 • 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 • 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 • 迁移时不支持目的端自动建表。
	分库	• 数据仓库：数据湖探索（DLI） • Hadoop：MRS HBase，MRS Hive • 搜索：Elasticsearch，云搜索服务（CSS） • 对象存储：对象存储服务（OBS）	分库数据源不支持作为目的端。 分库指的是同时连接多个后端数据源，该连接可作为作业源端，将多个数据源的数据合一迁移到其他数据源上。
NoSQL	Redis	Hadoop：MRS HDFS，MRS HBase，MRS Hive	除了表格存储服务（CloudTable）外，其他NoSQL数据源不支持作为目的端。
	文档数据库服务（DDS）		
	MongoDB		

数据源分类	源端数据源	对应的目的端数据源	说明
	表格存储服务（CloudTable HBase）	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server，MySQL，PostgreSQL，Microsoft SQL Server，Oracle • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
	Cassandra	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	
消息系统	数据接入服务（DIS）	搜索：云搜索服务（CSS）	消息系统不支持作为目的端。
	Apache Kafka		
	DMS Kafka		

数据源分类	源端数据源	对应的目的端数据源	说明
	MRS Kafka	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	• MRS Kafka不支持作为目的端。 • 仅支持本地存储，不支持存算分离场景。 • 不支持Ranger场景。 • 不支持ZK开启SSL场景。
搜索	Elasticsearch	• 数据仓库：数据仓库服务（DWS），数据湖探索（DLI） • Hadoop：MRS HDFS，MRS HBase，MRS Hive • 对象存储：对象存储服务（OBS） • 关系型数据库：云数据库 MySQL，云数据库 PostgreSQL，云数据库 SQL Server • NoSQL：表格存储服务（CloudTable） • 搜索：Elasticsearch，云搜索服务（CSS）	Elasticsearch仅支持非安全模式。
	云搜索服务（CSS）		导入数据到CSS推荐使用Logstash，请参见**使用Logstash导入数据到Elasticsearch**。

说明

上表中非云服务的数据源，例如MySQL，既可以支持用户本地数据中心自建的MySQL，也可以是用户在ECS上自建的MySQL，还可以是第三方云的MySQL服务。

整库迁移支持的数据源类型

整库迁移适用于将本地数据中心或在ECS上自建的数据库，同步到云上的数据库服务或大数据服务中，适用于数据库离线迁移场景，不适用于在线实时迁移。

数据集成支持整库迁移的数据源如**表4-4**所示。

表 4-4 整库迁移支持的数据源

数据源分类	数据源	读取	写入	说明
数据仓库	数据仓库服务（DWS）	支持	支持	-
Hadoop （仅支持本地存储，不支持存算分离场景，不支持Ranger场景，不支持ZK开启SSL场景）	MRS HBase	支持	支持	整库迁移仅支持导出到MRS HBase。 建议使用的版本： ● 2.1.X ● 1.3.X
	MRS Hive	支持	支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
	FusionInsight HBase	支持	不支持	建议使用的版本： ● 2.1.X ● 1.3.X
	FusionInsight Hive	支持	不支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X
	Apache HBase	支持	不支持	建议使用的版本： ● 2.1.X ● 1.3.X
	Apache Hive	支持	不支持	整库迁移仅支持导出到关系型数据库。 暂不支持2.x版本，建议使用的版本： ● 1.2.X ● 3.1.X

数据源分类	数据源	读取	写入	说明
关系数据库	云数据库 MySQL	支持	支持	不支持OLTP到OLTP迁移，此场景推荐通过数据复制服务DRS进行迁移。
	云数据库 PostgreSQL	支持	支持	
	云数据库 SQL Server	支持	支持	
	MySQL	支持	不支持	
	PostgreSQL	支持	不支持	
	Microsoft SQL Server	支持	不支持	
	Oracle	支持	不支持	
	SAP HANA	支持	不支持	• 仅支持2.00.050.00.1592305219版本。 • 仅支持Generic Edition。 • 不支持BW/4 FOR HANA。 • 仅支持英文字母的数据库名、表名与列名，不支持存在空格、符号等特殊字符。 • 仅支持日期、数字、布尔、字符（除SHORTTEXT）类型的数据类型，不支持二进制类型等其他数据类型。 • 迁移时不支持目的端自动建表。
	达梦数据库 DM	支持	不支持	仅支持导出到DWS、Hive
NoSQL	Redis	支持	支持	-
	文档数据库服务（DDS）	支持	支持	仅支持DDS和MRS之间迁移。
	表格存储服务（CloudTable）	支持	支持	-

4.3 支持的数据类型

配置字段映射时，数据源支持的数据类型请参见表4-5，以确保数据完整导入到目的端。

表 4-5 支持的数据类型

数据连接类型	数据类型说明
MySQL	请参见**MySQL数据库迁移时支持的数据类型**。
SQL Server	请参见**SQL Server数据库迁移时支持的数据类型**。
Oracle	请参见**Oracle数据库迁移时支持的数据类型**。
PostgreSQL	请参见**PostgreSQL数据库迁移时支持的数据类型**。
神通（ST）	请参见**神通（ST）数据库迁移时支持的数据类型**。
SAP HANA	请参见**SAP HANA数据库迁移时支持的数据类型**。
DWS	请参见**DWS数据库迁移时支持的数据类型**。
达梦	请参见**达梦数据库迁移时支持的数据类型**。
DLI	请参见**DLI数据库迁移时支持的数据类型**。
Elasticsearch/云搜索服务（CSS）	请参见**Elasticsearch/云搜索服务（CSS）数据库迁移时支持的数据类型**。

MySQL 数据库迁移时支持的数据类型

源端为MySQL数据库，目的端为Hive、DWS时，支持的数据类型如下：

表 4-6 开源 MySQL 数据库作为源端时支持的数据类型

类别	类型	简要释义	存储格式示例	Hive	DWS
字符串	CHAR（M）	固定长度的字符串是以长度为1到255之间个字符长度（例如：CHAR（5）），存储右空格填充到指定的长度。 限定长度不是必需的，它会默认为1。	‘a’或‘aaaaa’	CHAR	CHAR

类别	类型	简要释义	存储格式示例	Hive	DWS
	VARCHAR（M）	可变长度的字符串是以长度为1到255之间字符数（高版本的MySQL超过255）；例如：VARCHAR（25）. 创建VARCHAR类型字段时，必须定义长度。	‘a’或‘aaaaa’	VARCHAR	VARCHAR
数值	DECIMAL（M，D）	非压缩浮点数不能是无符号的。在解包小数，每个小数对应于一个字节。 定义显示长度（M）和小数（D）的数量是必需的。 NUMERIC是DECIMAL的同义词。	52.36	DECIMAL	D为0时对应BIGINT D不为0时对应NUMBERIC
	NUMBERIC	与 DECIMAL 相同。	-	DECIMAL	NUMBERIC
	INTEGER	一个正常大小的整数，可以带符号。如果是有符号的，它允许的范围是从-2147483648到2147483647。 如果是无符号，允许的范围是从0到4294967295。 可以指定多达11位的宽度。	5236	INT	INTEGER
	INTEGER UNSIGNED	INTEGER 的无符号形式。	-	BIGINT	INTEGER
	INT	与INTEGER相同。	5236	INT	INTEGER
	INT UNSIGNED	与INTEGER UNSIGNED相同。	-	BIGINT	INTEGER

类别	类型	简要释义	存储格式示例	Hive	DWS
	BIGINT	一个大的整数，可以带符号。如果有符号，允许范围为-9223372036854775808到92233720368547758 07。如果无符号，允许的范围是从0到184467440737095516 15. 可以指定最多20位的宽度。	5236	BIGINT	BIGINT
	BIGINT UNSIGNED	BIGINT的无符号形式。	-	BIGINT	BIGINT
	MEDIUMINT	一个中等大小的整数，可以带符号。如果有符号，允许范围为-8388608至8388607。 如果无符号，允许的范围是从0到16777215，可以指定最多9位的宽度。	-128，127	INT	INTEGER
	MEDIUMINT UNSIGNED	MEDIUMINT的无符号形式。	-	BIGINT	INTEGER
	TINYINT	一个非常小的整数，可以带符号。如果是有符号，它允许的范围是从-128到127。 如果是无符号，允许的范围是从0到255，可以指定多达4位数的宽度。	100	TINYINT	SMALLINT
	TINYINT UNSIGNED	TINYINT的无符号形式。	-	TINYINT	SMALLINT
	BOOL	MySQL的bool实际上就是tinyint（1）。	-128、127	SMALLINT	BYTEA

类别	类型	简要释义	存储格式示例	Hive	DWS
	SMALLINT	一个小的整数，可以带符号。如果有符号，允许范围为-32768至32767。 如果无符号，允许的范围是从0到65535，可以指定最多5位的宽度。	9999	SMALLINT	SMALLINT
	SMALLINT UNSIGNED	SMALLINT的无符号形式。	-	INT	SMALLINT
	REAL	同DOUBLE。	-	DOUBLE	-
	FLOAT（M，D）	不能使用无符号的浮点数字。可以定义显示长度（M）和小数位数（D）。这不是必需的，并且默认为10，2。其中2是小数的位数，10是数字（包括小数）的总数。小数精度可以到24个浮点。	52.36	FLOAT	FLOAT4
	DOUBLE（M，D）	不能使用无符号的双精度浮点数。可以定义显示长度（M）和小数位数（D）。 这不是必需的， 默认为16，4，其中4是小数的位数。小数精度可以达到53位的DOUBLE。 REAL是DOUBLE同义词。	52.36	DOUBLE	FLOAT8
	DOUBLE PRECISION	与DOUBLE相似。	52.3	DOUBLE	FLOAT8
位	BIT（M）	存储位值的BIT类型。BIT（M）可以存储多达M位的值，M的范围在1到64之间。	B'1111100' B'1100'	TINYINT	BYTEA

类别	类型	简要释义	存储格式示例	Hive	DWS
日期时间	DATE	以YYYY-MM-DD格式的日期，在1000-01-01和9999-12-31之间。 例如，1973年12月30日将被存储为1973-12-30。	1999-10-01	DATE	TIMESTAMP
	TIME	用于存储时、分、秒信息。	'09:10:21'或'9:10:21'	不支持（String）	TIME
	DATETIME	日期和时间组合以YYYY-MM-DD HH:MM:SS格式，在1000-01-01 00:00:00到9999-12-31 23:59:59之间。例如，1973年12月30日下午3:30，会被存储为1973-12-30 15:30:00。	'1973-12-30 15:30:00'	TIMESTAMP	TIMESTAMP
	TIMESTAMP	1970年1月1日午夜之间的时间戳，到2037的某个时候。这看起来像前面的DATETIME格式，无需只是数字之间的连字符; 1973年12月30日下午3点30分将被存储为19731230153000（YYYYMMDDHHMMSS）。	19731230153000	TIMESTAMP	TIMESTAMP
	YEAR（M）	以2位或4位数字格式来存储年份。如果长度指定为2（例如YEAR（2）），年份就可以为1970至2069（70～69）。如果长度指定为4，年份范围是1901-2155，默认长度为4。	2000	不支持（String）	不支持
多媒体（二进制）	BINARY（M）	字节数为M，允许长度为0-M的变长二进制字符串，字节数为值得长度加1。	0x2A3B4058（二进制数据）	不支持	BYTEA
	VARBINARY（M）	字节数为M，允许长度为0-M的定长二进制字符串。	0x2A3B4059（二进制数据）	不支持	BYTEA

类别	类型	简要释义	存储格式示例	Hive	DWS
	TEXT	字段的最大长度是65535个字符。 TEXT是“二进制大对象”，并用来存储大的二进制数据，如图像或其他类型的文件。	0x5236（二进制数据）	不支持	不支持
	TINYTEXT	0-255字节短文本二进制字符串。	-	-	不支持
	MEDIUMTEXT	0-167772154字节中等长度文本二进制字符串。	-	-	不支持
	LONGTEXT	0-4294967295字节极大长度文本二进制字符串。	-	-	不支持
	BLOB	字段的最大长度是65535个字符。 BLOB是“二进制大对象”，并用来存储大的二进制数据，如图像或其他类型的文件。BLOB大小写敏感。	0x5236（二进制数据）	不支持	不支持
	TINYBLOB	0-255字节短文本二进制字符串。	-	不支持	不支持
	MEDIUMBLOB	0-167772154字节中等长度文本二进制字符串。	-	不支持	不支持
	LONGBLOB	0-4294967295字节极大长度文本二进制字符串。	0x5236（二进制数据）	不支持	不支持
特殊类型	SET	SET是一个字符串对象，可以有零或多个值，其值来自表创建时规定的允许的一列值。指定包括多个SET成员的SET列值时各成员之间用逗号（ ‘，’ ）间隔开。这样SET成员值本身不能包含逗号。	-	-	不支持
	JSON	-	-	不支持	不支持（TEXT）

类别	类型	简要释义	存储格式示例	Hive	DWS
	ENUM	当定义一个ENUM，要创建它的值的列表，这些是必须用于选择的项（也可以是NULL）。例如，如果想要字段包含“A”或“B”或“C”，那么可以定义为ENUM为 ENUM（“A”，“B”，“C”）也只有这些值（或NULL）才能用来填充这个字段。	-	不支持	不支持

Oracle 数据库迁移时支持的数据类型

源端为Oracle数据库，目的端为Hive、DWS时，支持的数据类型如下：

表 4-7 Oracle 数据库作为源端时支持的数据类型

类别	类型	简要释义	Hive	DWS
字符串	char	定长字符串，会用空格填充来达到最大长度。	CHAR	CHAR
	nchar	包含unicode格式数据的定长字符串。	CHAR	CHAR
	varchar2	是VARCHAR的同义词。这是一个变长字符串，与CHAR类型不同，它不会用空格将字段或变量填充至最大长度。	VARCHAR	VARCHAR
	nvarchar2	包含unicode格式数据的变长字符串。	VARCHAR	VARCHAR
数值	number	能存储精度最多高达38位的数字。	DECIMAL	NUMERIC
	binary_float	2位单精度浮点数。	FLOAT	FLOAT8
	binary_double	64位双精度浮点数。	DOUBLE	FLOAT8
	long	能存储最多2GB的字符数据。	不支持	不支持
日期时间	date	7字节的定宽日期/时间数据类型，其中包含7个属性：世纪、世纪中的哪一年、月份、月中的哪一天、小时、分钟、秒。	DATE	TIMESTAMP

类别	类型	简要释义	Hive	DWS
	timestamp	7字节或11字节的定宽日期/时间数据类型，它包含小数秒。	TIMESTAMP	TIMESTAMP
	timestamp with time zone	3字节的timestamp，提供了时区支持。	TIMESTAMP	TIME WITH TIME ZONE
	timestamp with local time zone	7字节或11字节的定宽日期/时间数据类型，在数据的插入和读取时会发生时区转换。	TIMESTAMP	不支持（TEXT）
	interval year to month	5字节的定宽数据类型，用于存储一个时段。	不支持	不支持（TEXT）
	interval day to second	11字节的定宽数据类型，用于存储一个时段。将时段存储为天/小时/分钟/秒数，还可以有9位小数秒。	不支持	不支持（TEXT）
多媒体（二进制）	raw	一种变长二进制数据类型，采用这种数据类型存储的数据不会发生字符集转换。	不支持	不支持
	long raw	能存储多达2GB的二进制信息。	不支持	不支持
	blob	能够存储最多4GB的数据。	不支持	不支持
	clob	在Oracle 10g及以后的版本中允许存储最多（4GB）×（数据库块大小）字节的数据。CLOB包含要进行字符集转换的信息。这种数据类型很适合存储纯文本信息。	String	不支持
	nclob	这种类型能够存储最多4GB的数据。当字符集发生转换时，这种类型会受到影响。	不支持	不支持
	bfile	可以在数据库列中存储一个oracle目录对象和一个文件名，用户可以通过它来读取这个文件。	不支持	不支持
其他类型	rowid	实际上是数据库表中行的地址，它有10字节长。	不支持	不支持
	urowid	是一个通用的rowid，没有固定的rowid的表。	不支持	不支持

SQL Server 数据库迁移时支持的数据类型

源端为SQL Server数据库，目的端为Hive、DWS、Oracle时，支持的数据类型如下：

表 4-8 SQL Server 数据库作为源端时支持的数据类型

类别	类型	简要释义	Hive	DWS	Oracle
字符串数据类型	char	定长字符串，会用空格填充来达到最大长度。	CHAR	CHAR	CHAR
	nchar	包含unicode格式数据的定长字符串。	CHAR	CHAR	CHAR
	varchar	可变长度的字符串是以长度为1到255之间字符数（高版本的MySQL超过255）；例如： VARCHAR（25）； 创建VARCHAR类型字段时，必须定义长度。	VARCHAR	VARCHAR	VARCHAR
	nvarchar	与varchar类似，存储可变长度Unicode字符数据。	VARCHAR	VARCHAR	VARCHAR
数值数据类型	int	int存储在4个字节中，其中一个二进制位表示符号位，其它31个二进制位表示长度和大小，可以表示-2的31次方~2的31次方-1范围内的所有整数。	INT	INTEGER	INT
	bigint	bigint存储在8个字节中，其中一个二进制位表示符号位，其它63个二进制位表示长度和大小，可以表示-2的63次方~2的63次方-1范围内的所有整数。	BIGINT	BIGINT	NUMBER
	smallint	smallint类型的数据占用了两个字节的存储空间，其中一个二进制位表示整数值的正负号，其它15个二进制位表示长度和大小，可以表示-2的15次方~2的15次方-1范围内的所有整数。	SMALLINT	SMALLINT	NUMBER
	tinyint	tinyint类型的数据占用了一个字节的存储空间，可以表示0~255范围内的所有整数。	TINYINT	TINYINT	NUMBER
	real	可以存储正的或者负的十进制数值。	DOUBLE	FLOAT4	NUMBER
	float	其中为用于存储float数值尾数的位数（以科学计数法表示），因此可以确定精度和存储大小。	FLOAT	FLOAT8	binary_float

类别	类型	简要释义	Hive	DWS	Oracle
	decimal	带固定精度和小数位数的数值数据类型。	DECIMAL	NUMERIC	NUMBER
	numeric	用于存储零、正负定点数。	DECIMAL	NUMERIC	NUMBER
日期时间数据类型	date	存储用字符串表示的日期数据。	DATE	TIMESTAMP	DATE
	time	以字符串形式记录一天的某个时间。	不支持（String）	TIME	不支持
	datetime	用于存储时间和日期数据。	TIMESTAMP	TIMESTAMP	不支持
	datetime2	datetime的扩展类型，其数据范围更大，默认的最小精度最高，并具有可选的用户定义的精度。	TIMESTAMP	TIMESTAMP	不支持
	smalldatetime	smalldatetime类型与datetime类型相似，只是其存储范围是从1900年1月1日到2079年6月6日，当日期时间精度较小时，可以使用smalldatetime，该类型数据占用4个字节的存储空间。	TIMESTAMP	TIMESTAMP	不支持
	datetimeoffset	用于定义一个采用24小时制与日期相组合并可识别时区的时间。	不支持（String）	TIMESTAMP	不支持
多媒体数据类型（二进制）	text	用于存储文本数据。	不支持（String）	不支持（String）	不支持
	netxt	与text类型作用相同，为长度可变的非Unicode数据。	不支持（String）	不支持（String）	不支持
	image	长度可变的二进制数据，用于存储照片、目录图片或者图画。	不支持（String）	不支持（String）	不支持
	binary	长度为n个字节的固定长度二进制数据，其中n是从1~8000的值。	不支持（String）	不支持（String）	不支持
	varbinary	可变长度二进制数据。	不支持（String）	不支持（String）	不支持
货币数据类型	money	用于存储货币值。	不支持（String）	不支持（String）	不支持

类别	类型	简要释义	Hive	DWS	Oracle
	smallmoney	与money类型相似，输入数据时在前面加上一个货币符号，如人民币为￥或其它定义的货币符号。	不支持（String）	不支持（String）	不支持
位数据类型	bit	位数据类型，只取0或1为值，长度1字节。bit值经常当作逻辑值用于判断true（1）或false（0），输入非0值时系统将其替换为1。	不支持	不支持	不支持
其他数据类型	rowversion	每个数据都有一个计数器，当对数据库中包含rowversion列的表执行插入或者更新操作时，该计数器数值就会增加。	不支持	不支持	不支持
	uniqueidentifier	16字节的GUID（Globally Unique Identifier，全球唯一标识符），是Sql Server根据网络适配器地址和主机CPU时钟产生的唯一号码，其中，每个为都是0~9或a~f范围内的十六进制数字。	不支持	不支持	不支持
	cursor	游标数据类型。	不支持	不支持	不支持
	sql_variant	用于存储除文本，图形数据和timestamp数据外的其它任何合法的Sql Server数据，可以方便Sql Server的开发工作。	不支持	不支持	不支持
	table	用于存储对表或视图处理后的结果集。	不支持	不支持	不支持
	xml	存储xml数据的数据类型。可以在列中或者xml类型的变量中存储xml实例。存储的xml数据类型表示实例大小不能超过2GB。	不支持	不支持	不支持

PostgreSQL 数据库迁移时支持的数据类型

源端为PostgreSQL数据库，目的端为Hive、DWS、DLI时，支持的数据类型如下：

表 4-9 PostgreSQL 数据库作为源端时支持的数据类型

类别	类型	简要释义	Hive	DWS	DLI
字符	char	定长字符串，存储右空格填充到指定的长度。	CHAR	CHAR	不支持（String）
	varchar	变长字符串，不会用空格将字段或变量填充至最大长度。	CARCHAR	CARCHAR	不支持（String）

类别	类型	简要释义	Hive	DWS	DLI
数值	smallint	拓展名 int2，存储在2个字节中，它允许的范围是从-32768到32767。	SMALLINT	SMALLINT	SMALLINT
	int	拓展名 int4，存储在4个字节中，它允许的范围是从-2147483648到2147483647。	INTEGER	INT	INT
	bigint	拓展名 int8，存储在8个字节中，允许范围为-9223372036854775808到9223372036854775807。	BIGINT	BIGINT	BIGINT
	decimal（p，s）	精度p表示为值存储的有效位数，刻度s表示可以在小数点后存储的位数。p最大位数是1000。	DECIMAL（P，S）	DECIMAL（P，S）	DECIMAL（P，S）
	float	4字节或8字节存储。float（n）：n取值在1-24内，精度有效位数为6 位数，长度4 个字节，是单精度，n取值在25-53内，精度有效位数为15 位数，长度8 字节，是双精度。	FLOAT/DOUBLE	FLOAT/DOUBLE	FLOAT/DOUBLE
	smallserial	序列数据类型，以smallint格式存储。	SMALLINT	SMALLINT	SMALLINT
	serial	序列数据类型，以int格式存储。	INTEGER	INT	INT
	bigserial	序列数据类型，以bigint格式存储。	BIGINT	BIGINT	BIGINT
日期时间	date	存储日期数据。	DATE	DATE	DATE
	timestamp	存储日期和时间数据，无时区。	TIMESTAMP	TIMESTAMP	不支持（String）
	timestamptz	存储日期和时间数据，有时区。	TIMESTAMP	TIMESTAMPZ	不支持（String）

类别	类型	简要释义	Hive	DWS	DLI
	time	只用于一日内时间，无时区。	不支持（String）	TIME	不支持（String）
	timez	只用于一日内时间，有时区。	不支持（String）	TIMEZ	不支持（String）
	interval	时间间隔。	不支持（String）	不支持（String）	不支持（String）
位串类型	bit	定长位串，例如：b'000101'。	不支持（String）	不支持（String）	不支持（String）
	varbit	可变长位串，例如：b'101'。	不支持（String）	不支持（String）	不支持（String）
货币类型	money	存储在8个字节中，它允许的范围是从-92233720368547.5808到922337203685477.5807。	DOUBLE	MONEY	DECIMAL（P，S）
布尔类型	boolean	存储在1个字节中，可以取值为 1、0 或 NULL。	BOOLEAN	BOOLEAN	BOOLEAN
文本类型	text	变长文本，无长度限制。	不支持（String）	不支持（String）	不支持（String）

DWS 数据库迁移时支持的数据类型

源端为DWS数据库时，支持的数据类型如下：

表 4-10 DWS 数据库作为源端时支持的数据类型

类别	类型	简要释义
字符	char	定长字符串，存储右空格填充到指定的长度。
	varchar	变长字符串，不会用空格将字段或变量填充至最大长度。
数值	double	用于存储指明双精度的浮点数。
	decimal（p，s）	精度p表示为值存储的有效位数，刻度s表示可以在小数点后存储的位数。p最大位数是1000。

类别	类型	简要释义
	numeric	用于存储零、正负定点数。
	real	与double相同。
	int	int存储在4个字节中，其中一个二进制位表示符号位，其它31个二进制位表示长度和大小，可以表示-2的31次方~2的31次方-1范围内的所有整数。
	bigint	bigint存储在8个字节中，其中一个二进制位表示符号位，其它63个二进制位表示长度和大小，可以表示-2的63次方~2的63次方-1范围内的所有整数。
	smallint	smallint类型的数据占用了两个字节的存储空间，其中一个二进制位表示整数值的正负号，其它15个二进制位表示长度和大小，可以表示-2的15次方~2的15次方-1范围内的所有整数。
	tinyint	tinyint类型的数据占用了一个字节的存储空间，可以表示0~255范围内的所有整数。
日期时间	date	存储日期数据。
	timestamp	存储日期和时间数据，无时区。
	time	只用于一日内时间，无时区。
位串类型	bit	定长位串，例如： b'000101'。
布尔类型	boolean	存储在1个字节中，可以取值为 1、0 或 NULL。
文本类型	text	变长文本，无长度限制。

神通（ST）数据库迁移时支持的数据类型

源端为神通（ST）数据库，目的端为MRS Hive、MRS Hudi时，支持的数据类型如下：

表 4-11 神通（ST）数据库作为源端时支持的数据类型

类别	类型	简要释义	存储格式示例	MRS Hive	MRS Hudi
字符	VARCHAR	用于存储指定定长字符串。	‘a’或‘aaaaa’	VARCHAR（765）	STRING
	BPCHAR	用于存储指定变长字符串。	‘a’或‘aaaaa’	VARCHAR（765）	STRING

类别	类型	简要释义	存储格式示例	MRS Hive	MRS Hudi
数值	NUMERIC	用于存储零、正负定点数。	52.36	DECIMAL（10，0）	DECIMAL（18，0）
	INT	用于存储零、正负定点数。	5236	INT	INT
	BIGINT	用于存储有符号整数，精度为19，标度为0。	5236	BIGINT	BIGINT
	TINYINT	用于存储有符号整数，精度为3，标度为0。	100	SMALLINT	INT
	BINARY	用于存储定长二进制数据。	0x2A3B4058	不支持	FLOAT
	VARBINARY	用于存储可变长二进制数据。	0x2A3B4058	不支持	BINARY
	FLOAT	用于存储带二进制精度的浮点数。	52.36	FLOAT	FLOAT
	DOUBLE	用于存储指明双精度的浮点数。	52.3	DOUBLE	DOUBLE
日期时间	DATE	用于存储年、月、日信息。	'1999-10-01'、'1999/10/01'或'1999.10.01'	DATE	DATE
	TIME	用于存储时、分、秒信息。	'09:10:21'或'9:10:21'	STRING	STRING
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。	'2002-12-12 09:10:21'、'2002-12-12 9:10:21'、'2002/12/12 09:10:21' 或'2002.12.12 09:10:21'	TIMESTAMP	TIMESTAMP
多媒体	CLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。	0x5236（二进制数据）	STRING	STRING
	BLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。	0x5236（二进制数据）	不支持	BINARY

类别	类型	简要释义	存储格式示例	MRS Hive	MRS Hudi
布尔类型	BOOLEAN	存储在1个字节中，可以取值为 1、0 或 NULL。	1	BOOLEAN	BOOLEAN

SAP HANA 数据库迁移时支持的数据类型

源端为SAP HANA数据库时，支持的数据类型如下：

表 4-12 SAP HANA 数据库作为源端时支持的数据类型

类别	类型	简要释义
字符	VARCHAR	用于存储指定定长字符串。
	NVARCHAR	包含unicode格式数据的变长字符串。
	TEXT	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。
数值	BIGINT	用于存储有符号整数，精度为19，标度为0。
	TINYINT	用于存储有符号整数，精度为3，标度为0。
	SMALLINT	SMALLINT类型的数据占用了两个字节的存储空间，其中一个二进制位表示整数值的正负号，其它15个二进制位表示长度和大小，可以表示-2的15次方~2的15次方-1范围内的所有整数。
	REAL	可以存储正的或者负的十进制数值。
	DECIMAL	带固定精度和小数位数的数值数据类型。
	FLOAT	用于存储带二进制精度的浮点数。
	DOUBLE	用于存储指明双精度的浮点数。
日期时间	DATE	用于存储年、月、日信息。
	TIME	用于存储时、分、秒信息。
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。
多媒体	CLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。
	NCLOB	这种类型能够存储最多4GB的数据。当字符集发生转换时，这种类型会受到影响。
布尔类型	BOOLEAN	存储在1个字节中，可以取值为 1、0 或 NULL。

DLI 数据库迁移时支持的数据类型

源端为DLI数据库时，支持的数据类型如下：

表 4-13 DLI 数据库作为源端时支持的数据类型

类别	类型	简要释义
字符	CHAR	用于存储指定定长字符串。
	VARCHAR	与CHAR相同。
	STRING	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。
数值	BIGINT	用于存储有符号整数，精度为19，标度为0。
	TINYINT	用于存储有符号整数，精度为3，标度为0。
	SMALLINT	SMALLINT类型的数据占用了两个字节的存储空间，其中一个二进制位表示整数值的正负号，其它15个二进制位表示长度和大小，可以表示-2的15次方~2的15次方-1范围内的所有整数。
	INT	用于存储有符号整数，精度为10，标度为0。
	DECIMAL	带固定精度和小数位数的数值数据类型。
	FLOAT	用于存储带二进制精度的浮点数。
	DOUBLE	用于存储指明双精度的浮点数。
日期时间	DATE	用于存储年、月、日信息。
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。
布尔类型	BOOLEAN	存储在1个字节中，可以取值为 1、0 或 NULL。

Elasticsearch/云搜索服务（CSS）数据库迁移时支持的数据类型

源端为Elasticsearch/云搜索服务（CSS）数据库时，支持的数据类型如下：

表 4-14 Elasticsearch/云搜索服务（CSS）数据库作为源端时支持的数据类型

类别	类型	简要释义	存储格式示例	MySQL
字符	keyword	用于存储字符串。	“keyword”	String
	text	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。	“long string”	TEXT

类别	类型	简要释义	存储格式示例	MySQL
	string	用于存储长字符串类型，其字符串的长度最大为 2G-1，存储长的文本串。	“a string”	String
整数	short	用于存储16位有符号整数，取值范围为-32768至32767。	32765	smallInt
	integer	用于存储32位有符号整数，取值范围为-2^{31}至2^{31}-1。	3276566	int
	long	用于存储64位有符号整数，取值范围为-2^{63}至2^{63}-1。	3276566666	BIGINT
数值	double	64位双精度IEEE 754浮点类型。	21.333	double
	float	32位单精度IEEE 754浮点类型 。	21.333	double
布尔类型	boolean	存储在1个字节中，可以取值为 1、0 或 NULL。	1	Boolean
对象	object	扁平化存储对象的字符串。	{"users.name": ["John", "Smith"], users.age": [26, 28], "users.sex" :[1，2]}	TEXT
嵌套	nested	嵌套存储对象的字符串。	{"users.name" : "John" ， "users.age" : 26， "users.sex" : 1} { "users.name" : "Smith", "users.age" : 28， "users.sex" : 2}	TEXT

类别	类型	简要释义	存储格式示例	MySQL
日期	date	日期格式的字符串。	“2018-01-13”或“2018-01-13 12:10:30”	DATE或timeStamp
特殊	ip	Ip地址格式的字符串。	“192.168.127.100”	String
数组	string_array	全部是字符串的数组。	[“str”，“str”]	TEXT
	short_array	全部是16位整数的数组。	[1，1，1]	TEXT
	integer_array	全部是32位整数的数组。	[1，1，1]	TEXT
	long_array	全部是64位整数的数组。	[1，1，1]	TEXT
	float_array	全部是32位浮点数的数组。	[1.0，1.0，1.0]	TEXT
	double_array	全部是64位浮点数的数组。	[1.0，1.0，1.0]	TEXT
范围	completion	自动补全的字符串。	“string”	TEXT

达梦数据库迁移时支持的数据类型

源端为达梦数据库，目的端为Hive、DWS时，支持的数据类型如下：

表 4-15 达梦数据库作为源端时支持的数据类型

类别	类型	简要释义	存储格式示例	Hive	DWS
字符	CHAR	用于存储指定定长字符串。	‘a’或‘aaaaa’	CHAR	CHAR
	CHARACTER	与 CHAR 相同。	‘a’或‘aaaaa’	CHAR	CHAR
	VARCHAR	用于存储指定变长字符串.	‘a’或‘aaaaa’	VARCHAR	VARCHAR
	VARCHAR2	与 VARCHAR 相同。	‘a’或‘aaaaa’	VARCHAR	VARCHAR

类别	类型	简要释义	存储格式示例	Hive	DWS
数值	NUMERIC	用于存储零、正负定点数。	52.36	DECIMAL	NUMERIC
	DECIMAL	与 NUMERIC 相似。	52.36	DECIMAL	NUMERIC
	DEC	与 DECIMAL 相同。	52.36	DECIMAL	NUMERIC
	NUMBER	与 NUMERIC 相同。	52.36	DECIMAL	NUMERIC
	INTEGER	用于存储有符号整数，精度为10，标度为0。	5236	INT	INTEGER
	INT	与 INTEGER 相同。	5236	INT	INTEGER
	BIGINT	用于存储有符号整数，精度为19，标度为0。	5236	BIGINT	BIGINT
	TINYINT	用于存储有符号整数，精度为3，标度为0。	100	TINYINT	SMALLINT
	SMALLINT	用于存储有符号整数，精度为5，标度为0。	9999	SMALLINT	SMALLINT
	BYTE	与 TINYINT 相似，精度为3，标度为0。	100	TINYINT	SMALLINT
	BINARY	用于存储定长二进制数据。	0x2A3B4058	BINARY（NULL）	BYTEA（NULL）
	VARBINARY	用于存储可变长二进制数据。	0x2A3B4058	BINARY（NULL）	BYTEA（NULL）
	FLOAT	用于存储带二进制精度的浮点数。	52.36	FLOAT	FLOAT8
	DOUBLE	与FLOAT类似。	52.36	DOUBLE	FLOAT8
	REAL	用于存储带二进制精度的浮点数，但它不能由用户指定使用的精度。	52.3	FLOAT	FLOAT4
	DOUBLE PRECISION	用于存储指明双精度的浮点数。	52.3	DOUBLE	FLOAT8

类别	类型	简要释义	存储格式示例	Hive	DWS
位串	BIT	用于存储整数数据1、0 或 NULL。	1、0 或 NULL	TINYINT（1 0 NULL）	BOOLEAN（true false NULL）
日期时间	DATE	用于存储年、月、日信息。	1999-10-01'、'1999/10/01'或'1999.10.01'	DATE	TIMESTAMP
	TIME	用于存储时、分、秒信息。	'09:10:21'或'9:10:21'	不支持（String）	TIME
	TIMESTAMP	用于存储年、月、日、时、分、秒信息。	2002-12-12 09:10:21'，'2002-12-12 9:10:21' '2002/12/12 09:10:21' 或'2002.12.12 09:10:21'	TIMESTAMP	TIMESTAMP
	TIME WITH TIME ZONE	用于存储一个带时区的 TIME 值，其定义是在 TIME 类型的后面加上时区信息。	'09:10:21 +8:00'，'09:10:21+8:00'或'9:10:21+8:00'	不支持（String）	TIME WITH TIME ZONE
	TIMESTAMP WITH TIME ZONE	用于存储一个带时区的 TIMESTAMP 值，其定义是TIMESTAMP类型的后面加上时区信息。	2002-12-12 09:10:21 +8:00'，'2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00'或'2002.12.12 09:10:21 +8:00'	TIMESTAMP	TIMESTAMP WITH TIME ZONE

类别	类型	简要释义	存储格式示例	Hive	DWS
	TIMESTAMP WITH LOCAL TIME ZONE	用于存储一个本地时区的 TIMESTAMP 值，能够将标准时区类型 TIMESTAMP WITH TIME ZONE 类型转化为本地时区类型。	2002-12-12 09:10:21 +8:00'，'2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00'或'2002.12.12 09:10:21 +8:00'	不支持（String）	不支持（TEXT）
	DATETIME WITH TIME ZONE	同TIMESTAMP WITH TIME ZONE。	2002-12-12 09:10:21 +8:00'，'2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00'或'2002.12.12 09:10:21 +8:00'	TIMESTAMP	TIMESTAMP WITH TIME ZONE
	INTERVAL YEAR	描述一个若干年的间隔，引导精度规定了年的取值范围。	INTERVAL '0015' YEAR	不支持（String）	不支持（VARCHAR）
	INTERVAL YEAR TO MONTH	描述一个若干年若干月的间隔，引导精度规定了年的取值范围。	INTERVAL '0015-08' YEAR TO MONTH	不支持（String）	不支持（VARCHAR）
	INTERVAL MONTH	描述一个若干月的间隔，引导精度规定了月的取值范围。	INTERVAL '0015' MONTH	不支持（String）	不支持（VARCHAR）
	INTERVAL DAY	描述一个若干日的间隔，引导精度规定了日的取值范围。	INTERVAL '150' DAY	不支持（String）	不支持（VARCHAR）
	INTERVAL DAY TO HOUR	描述一个若干日若干小时的间隔，引导精度规定了日的取值范围。	INTERVAL '923' DAY TO HOUR	不支持（String）	不支持（VARCHAR）

类别	类型	简要释义	存储格式示例	Hive	DWS
	INTERVAL DAY TO MINUTE	描述一个若干日若干小时若干分钟的间隔，引导精度规定了日的取值范围。	INTERVAL '09 23:12' DAY TO MINUTE	不支持（String）	不支持（VARCHAR）
	INTERVAL DAY TO SECOND	描述一个若干日若干小时若干分钟若干秒的间隔，引导精度规定了日的取值范围。	INTERVAL '09 23:12:01.1' DAY TO SECOND	不支持（String）	不支持（VARCHAR）
	INTERVAL HOUR	描述一个若干小时的间隔，引导精度规定了小时的取值范围。	INTERVAL '150' HOUR	不支持（String）	不支持（VARCHAR）
	INTERVAL HOUR TO MINUTE	描述一个若干小时若干分钟的间隔，引导精度规定了小时的取值范围。	INTERVAL '23:12' HOUR TO MINUTE	不支持（String）	不支持（VARCHAR）
	INTERVAL HOUR TO SECOND	描述一个若干小时若干分钟若干秒的间隔，引导精度规定了小时的取值范围。	INTERVAL '23:12:01.1' HOUR TO SECOND	不支持（String）	不支持（VARCHAR）
	INTERVAL MINUTE	描述一个若干分钟的间隔，引导精度规定了分钟的取值范围。	INTERVAL '150' MINUTE	不支持（String）	不支持（VARCHAR）
	INTERVAL MINUTE TO SECOND	描述一个若干分钟若干秒的间隔，引导精度规定了分钟的取值范围。	INTERVAL '12:01.1' MINUTE TO SECOND	不支持（String）	不支持（VARCHAR）
	INTERVAL SECOND	描述一个若干秒的间隔，引导精度规定了秒整数部分的取值范围。	INTERVAL '51.1' SECOND	不支持（String）	不支持（VARCHAR）

类别	类型	简要释义	存储格式示例	Hive	DWS
多媒体	IMAGE	IMAGE 用于指明多媒体信息中的图像类型。 图像由不定长的像素点阵组成，长度最大为 2G-1 字节。该类型除了存储图像数据之外，还可用于存储任何其它二进制数据。	0x2A3B4058（二进制数据）	不支持	不支持
	LONGVARBINARY	与IMAGE相同。	0x2A3B4059（二进制数据）	不支持	不支持
	TEXT	用于存储长字符串类型，其字符串的长度 最大为 2G-1，存储长的文本串。	0x5236（二进制数据）	不支持	不支持
	LONGVARCHAR	与 TEXT 相似。	0x5236（二进制数据）	不支持	不支持
	BLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。	0x5236（二进制数据）	不支持	不支持
	CLOB	用于存储变长的二进制大对象，长度最大为2G-1字节。	0x5236（二进制数据）	不支持	不支持
	BFILE	用于指明存储在操作系统中的二进制文件， 文件存储在操作系统而非数 据库中，仅能进行只读访问。	-	不支持	不支持

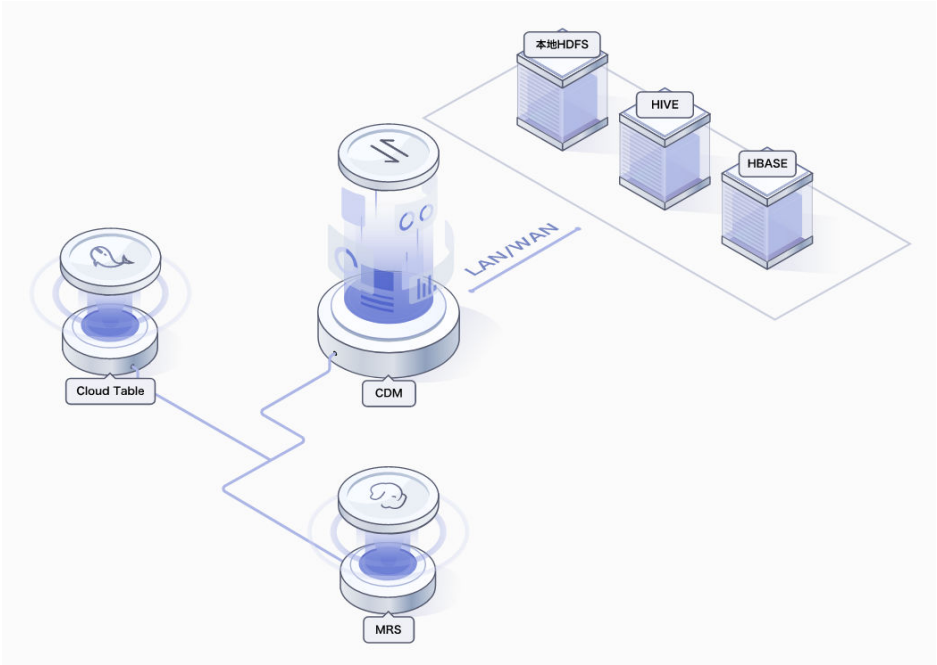
5 应用场景

大数据迁移上云

本地数据是指存储在用户自建或者租用的IDC中的数据，或者第三方云环境中的数据，包括关系型数据库、NoSQL数据库、OLAP数据库、文件系统等。

这个场景是用户希望利用云上的计算和存储资源，需要先将本地数据迁移上云。该场景下，需要保证本地网络与云上网络是连通的。

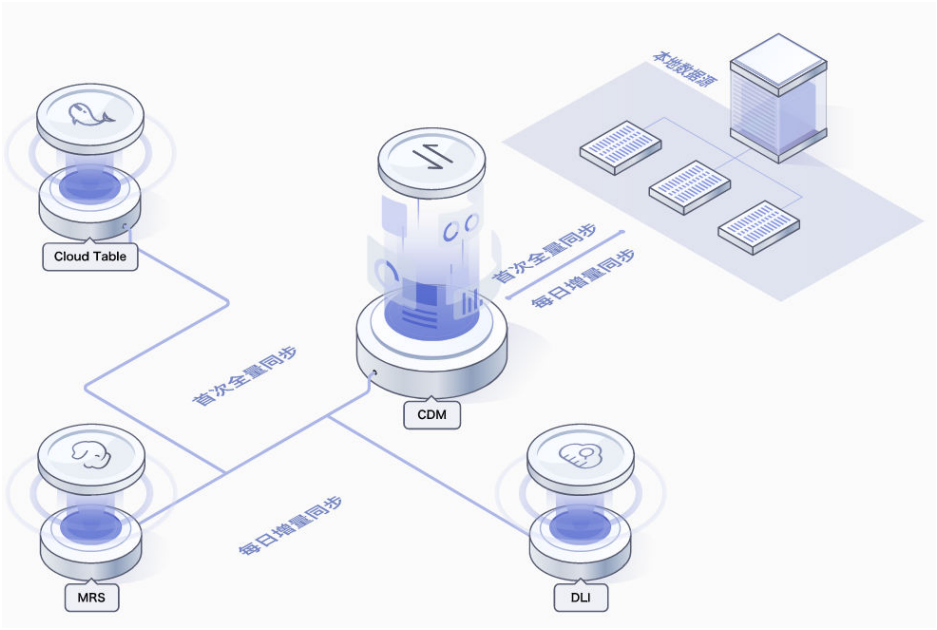
图 5-1 大数据迁移上云



数据批量入湖

这个场景支持用户本地数据全量和T+1增量入湖。

图 5-2 数据批量入湖



6 基本概念

CDM 集群

CDM集群是指用户拥有的CDM实例，一个CDM集群由1个或多个虚拟机组成。一个用户可以创建多个CDM集群，例如为财务部门和采购部门各创建一个CDM实例，实现数据访问权限的隔离。

本地环境

本地环境是指用户自建或者租用的IDC中的数据存储系统，或者第三方云环境中的数据存储系统，包括关系型数据库以及文件系统。

本地数据

本地数据是指存储在用户自建或者租用的IDC中的数据，或者第三方云环境中的数据，包括关系型数据库、NoSQL数据库、OLAP数据库、文件系统等。

连接器

连接器是CDM内置的连接某种数据源所需的对象模板，目前CDM支持连接OBS、MRS、数据库等多种连接器，并可扩展增加新的连接器。

连接

连接是用户基于连接器创建的用于连接某个具体数据源的对象。

创建连接时需要指定连接的名称、连接器类型，以及数据源的地址、鉴权信息，例如连接到MySQL数据库需要主机地址、端口号、用户名、密码等配置信息。

一个连接被创建后可以被多个作业引用，且既可以作为源连接，也可以作为目的连接。

作业

作业是指用户创建的数据迁移任务，迁移特定数据源的数据到另一特定数据源。创建时需要指定源连接、目的连接、数据映射规则。

源端作业配置

在创建作业的过程中，由源连接指定抽取哪个数据源的数据，不同源连接对应的源端作业参数不同，例如从哪个表或哪个目录导出数据，这些信息在源端作业配置中指定。

目的端作业配置

在创建作业的过程中，由目的连接指定加载数据到哪个数据源，不同目的连接对应的目的端作业参数不同，例如将数据导入到哪个表或哪个目录，这些信息在目的端作业配置中指定。

字段映射

在创建作业的过程中，尤其是异构数据源之间的迁移作业，一般需要配置源端和目的端数据源之间的对应关系，例如字段对应、字段类型对应，这个过程在CDM中称为字段映射。

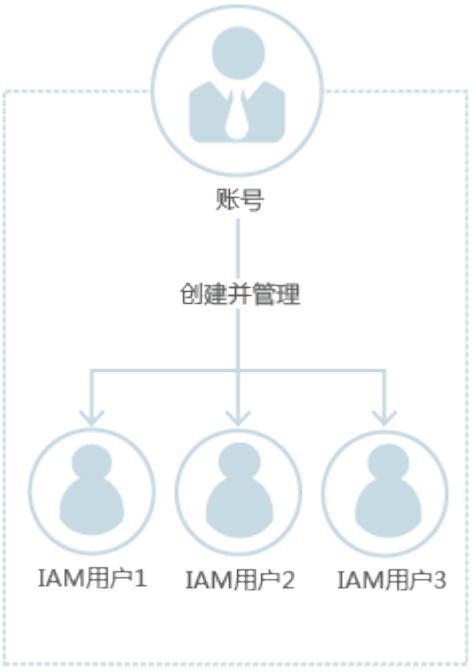
账号

当您首次登录云时注册的账号，该账号是您的云资源归属、资源使用计费的主体，对其所拥有的资源及云服务具有完全的访问权限，可以重置用户密码、创建IAM用户、分配IAM用户权限等。账号统一接收所有IAM用户进行资源操作时产生的费用账单。账号在登录控制台时，使用“账号登录”方式登录。

IAM 用户

由账号在IAM中创建的用户，类似于子账号，具有身份凭证（密码和访问密钥），可以使用自己单独的用户名和密码通过控制台或者API访问云服务。IAM用户根据账号授予的权限，帮助账号管理资源。IAM用户不拥有资源，不进行独立的计费，这些IAM用户的权限和资源由所属账号统一控制和付费。IAM用户在登录控制台登录时，使用“IAM用户登录”方式登录。

图 6-1 账号与 IAM 用户的关系



用户组

用户组是IAM用户的集合，IAM用户需要加入特定用户组后，才具备对应的权限，否则无法访问您账号中的任何资源或是云服务。一个IAM用户可以加入多个用户组，以获得不同的权限。

“admin”为系统缺省提供的用户组，具有所有云服务资源的操作权限，将IAM用户加入该用户组后，可以操作并使用所有云资源，包括但不限于创建用户组及IAM用户、修改用户组权限、管理云资源等。

策略与授权

策略是以JSON格式描述一组权限集的语言，它可以精确地描述被授权的资源集和操作集。包括系统策略和自定义策略两种：

- 系统策略是IAM预置的策略，您可以使用但不能修改。
- 若系统策略不满足授权要求，您可以创建自定义策略，自由搭配需要授予的权限集。

通过给用户组授予策略（包括系统策略和自定义策略），用户组中的IAM用户就能获得策略中定义的权限，这一过程称为授权。

例如拥有CDM所有权限的策略（CDM Administrator）内容如下：

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Action": [
                "cdm:*:*",
                "ecs:*:*",
                "vpc:*:*",
                "evs:*:*",
```

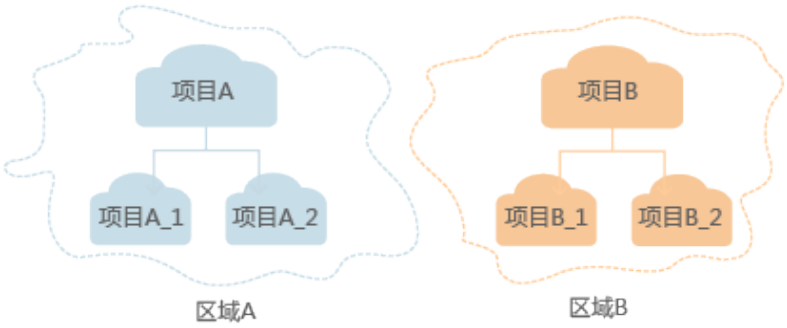
```
                    "bss:*:*",
                    "CTS:*:*",
                    "eps:*:*",
                    "obs:*:*",
                    "CES:*:*"
                ],
                "Effect": "Allow"
            }
        ]
}
```

项目

云服务所属的区域默认对应一个项目，这个项目由系统预置，用来隔离物理区域间的资源（计算资源、存储资源和网络资源）。

- 以默认项目为单位进行授权时，IAM用户可以访问您账号中该区域的所有资源。
- 如果您希望进行更加精细的权限控制，可以在区域默认的项目中创建子项目，并在子项目中创建资源，然后以子项目为单位进行授权，这样IAM用户仅能访问特定子项目中资源，使得资源的权限控制更加精确。

图 6-2 项目隔离模型



身份凭证

身份凭证是识别用户身份的依据，您通过控制台或者API访问云服务时，需要使用身份凭证来通过系统的鉴权认证。身份凭证包括密码和访问密钥，您可以在IAM中管理账号以及IAM用户的身份凭证。

- 密码：常见的身份凭证，密码可以用来登录控制台，还可以调用云服务的API。
- 访问密钥：即AK/SK（Access Key ID/Secret Access Key），调用API的身份凭证，不能用来登录控制台。访问密钥中具有验证身份的签名，通过加密签名验证可以确保机密性、完整性和请求双方身份的正确性。

7 区域和可用区

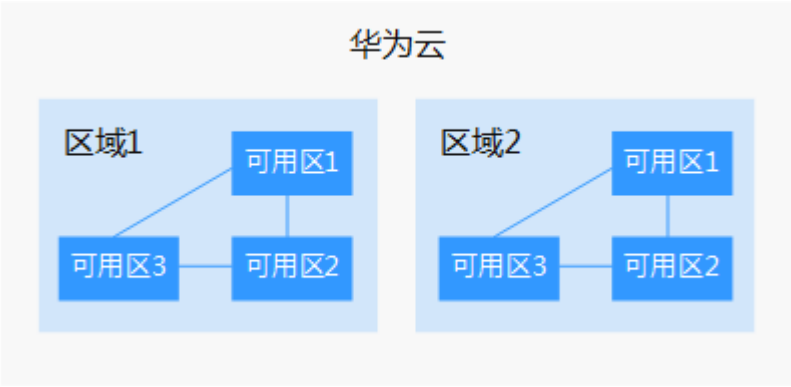
什么是区域、可用区?

我们用区域和可用区来描述数据中心的位置，您可以在特定的区域、可用区创建资源。

- 区域（Region）：从地理位置和网络时延维度划分，同一个Region内共享弹性计算、块存储、对象存储、VPC网络、弹性公网IP、镜像等公共服务。Region分为通用Region和专属Region，通用Region指面向公共租户提供通用云服务的Region；专属Region指只承载同一类业务或只面向特定租户提供业务服务的专用Region。
- 可用区（AZ，Availability Zone）：一个AZ是一个或多个物理数据中心的集合，有独立的风火水电，AZ内逻辑上再将计算、网络、存储等资源划分成多个集群。一个Region中的多个AZ间通过高速光纤相连，以满足用户跨AZ构建高可用性系统的需求。

图7-1阐明了区域和可用区之间的关系。

图 7-1 区域和可用区



目前，华为云已在全球多个地域开放云服务，您可以根据需求选择适合自己的区域和可用区。更多信息请参见华为云全球站点。

如何选择区域?

选择区域时，您需要考虑以下几个因素：

- 地理位置

  一般情况下，建议就近选择靠近您或者您的目标用户的区域，这样可以减少网络时延，提高访问速度。不过，在基础设施、BGP网络品质、资源的操作与配置等方面，中国大陆各个区域间区别不大，如果您或者您的目标用户在中国大陆，可以不用考虑不同区域造成的网络时延问题。

  香港、曼谷等其他地区和国家提供国际带宽，主要面向非中国大陆地区的用户。如果您或者您的目标用户在中国大陆，使用这些区域会有较长的访问时延，不建议使用。

  - 在除中国大陆以外的亚太地区有业务的用户，可以选择“亚太-曼谷”或“亚太-新加坡”区域。
  - 在非洲地区有业务的用户，可以选择“南非-约翰内斯堡”区域。
  - 在欧洲地区有业务的用户，可以选择“欧洲-巴黎”区域。

- 资源的价格

  不同区域的资源价格可能有差异，请参见**华为云服务价格详情**。

如何选择可用区?

是否将资源放在同一可用区内，主要取决于您对容灾能力和网络时延的要求。

- 如果您的应用需要较高的容灾能力，建议您将资源部署在同一区域的不同可用区内。
- 如果您的应用要求实例之间的网络延时较低，则建议您将资源创建在同一可用区内。

区域和终端节点

终端节点（Endpoint）即调用API的**请求地址**，不同服务不同区域的终端节点不同。本服务的Endpoint可从**终端节点Endpoint**获取。

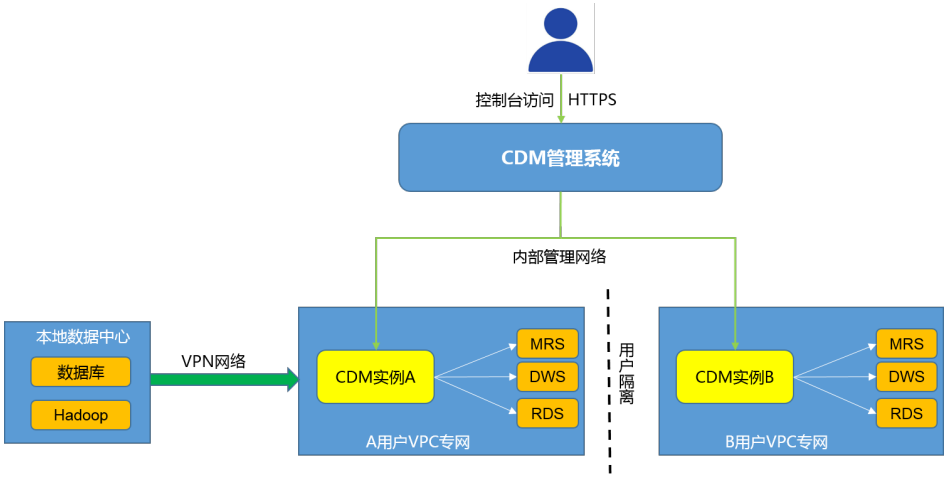
8 迁移原理

CDM 迁移原理

用户使用CDM服务时，CDM管理系统在用户VPC中发放全托管的CDM实例。此实例仅提供控制台和Rest API访问权限，用户无法通过其他接口（如SSH）访问实例。这种方式保证了CDM用户间的隔离，避免数据泄漏，同时保证VPC内不同云服务间数据迁移时的传输安全。用户还可以使用VPN网络将本地数据中心的数据迁移到云服务，具有高度的安全性。

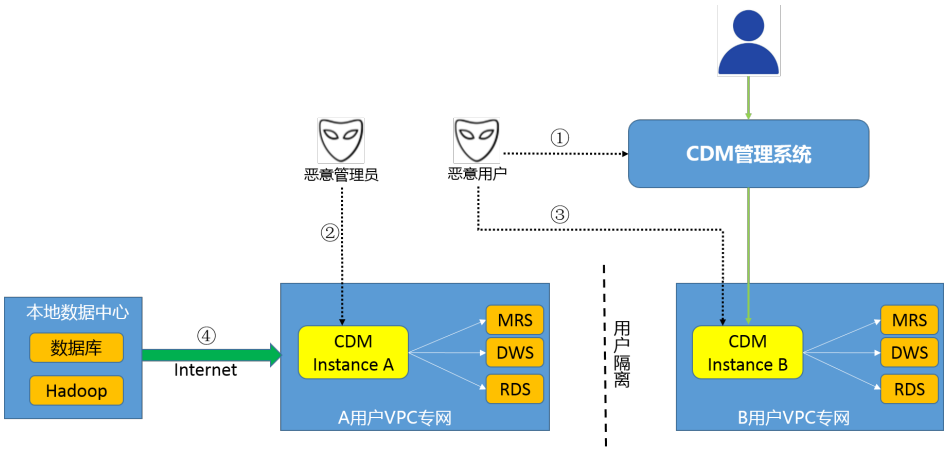
CDM数据迁移以抽取-写入模式进行。CDM首先从源端抽取数据然后将数据写入到目的端，数据访问操作均由CDM主动发起，对于数据源（如RDS数据源）支持SSL时，会使用SSL加密传输。迁移过程要求用户提供源端和目的端数据源的用户名和密码，这些信息将存储在CDM实例的数据库中。保护这些信息对于CDM安全至关重要。

图 8-1 CDM 迁移原理



安全边界和风险规避

图 8-2 风险规避



如图8-2所示，CDM可能存在以下威胁：

1. 互联网威胁：恶意用户可能通过CDM控制台攻击CDM。
2. 数据中心威胁：恶意CDM管理员获取用户的数据源访问信息（用户名和密码）。
3. 恶意用户威胁：恶意用户窃取其他用户的数据。
4. 数据暴露公网：从公网迁移数据时暴露数据的威胁。

对于这些潜在的威胁，CDM提供以下机制来规避终端用户的风险：

1. 针对互联网威胁：用户不能直接通过公网登录CDM控制台。CDM提供双层安全保障机制。
   a. 云控制台框架要求用户访问任何控制台页面都要进行用户认证。
   b. Web应用防火墙（Web Application Firewall，简称WAF）过滤所有控制台的请求内容并停止请求攻击代码/内容。
2. 针对数据中心威胁：用户必须向CDM系统提供迁移源端和目的端的访问用户名和密码信息，才能完成数据迁移。避免CDM管理员获取此类信息并攻击用户的重要数据源对于CDM非常重要，CDM为此类信息提供三级保护机制。
   a. CDM在本地数据库中存储经过AES-256加密的密码，确保用户隔离。本地数据库使用用户Ruby运行，数据库仅侦听127.0.0.1，用户没有远程访问数据库的权限。
   b. 用户实例发放完毕后，CDM将虚拟机的root和Ruby用户密码更改为随机密码且不会保存在任何地方，以阻止CDM管理员访问用户实例和含有密码信息的数据库。
   c. CDM实例迁移以推拉模式进行，因此CDM实例在VPC上没有侦听端口，用户无法从VPC访问本地数据库或操作系统。
3. 针对恶意用户的威胁：CDM对每个用户，使用单独的虚拟机来运行各自的CDM实例，用户之间的实例是完全隔离和安全的。恶意用户无法访问其他用户的实例。
4. 针对数据暴露公网的威胁：CDM的抽取-写入模型下，即使CDM绑定了弹性IP，也不会开放端口到弹性IP，攻击者无法通过弹性IP来访问和攻击CDM。不过从公网迁移数据的方式下，由于用户数据源也会暴露在公网，存在被第三方攻击的威

胁，推荐用户在数据源服务器上通过ACL或防火墙对源端进行防护，例如仅放通来自CDM绑定的弹性IP的访问请求。

9 与其他云服务的关系

统一身份认证服务

您注册的云账号对其所拥有的资源及云服务具有完全的访问权限，如果您需要给企业中的员工设置不同的CDM访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。IAM提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制云资源的访问。

虚拟私有云

CDM的集群创建在虚拟私有云（Virtual Private Cloud，简称VPC）的子网内，VPC通过逻辑方式进行网络隔离，为用户的集群提供安全、隔离的网络环境。

MapReduce 服务

CDM支持从MapReduce服务（MapReduce Service，简称MRS）导入、导出数据。

对象存储服务

CDM支持从对象存储服务（Object Storage Service，简称OBS）导入、导出数据，同时CDM还利用OBS存储集群数据备份文件和日志。

云监控

CDM服务使用云监控（Cloud Eye）监控CDM服务集群中的多项性能指标，从而集中高效地呈现状态信息，具体如表9-1所示。

表 9-1 CDM 的监控指标

指标名称	指标含义	取值范围	测量对象
网络流入速率	该指标用于统计每秒流入测量对象的网络流量。 单位：字节/秒。	≥ 0 bytes/s	CDM集群实例

指标名称	指标含义	取值范围	测量对象
网络流出速率	该指标用于统计每秒流出测量对象的网络流量。 单位：字节/秒。	≥ 0 bytes/s	CDM集群实例
CPU使用率	该指标用于统计测量对象的CPU使用率。 单位：%。	0%～100%	CDM集群实例
内存使用率	该指标用于统计测量对象的内存使用率。 单位：%。	0%～100%	CDM集群实例

云审计服务

CDM使用云审计服务（Cloud Trace Service，以下简称CTS）记录CDM相关的操作事件，便于日后的查询、审计和回溯，具体如表9-2所示。

表 9-2 支持云审计的关键操作列表

操作名称	资源类型	事件名称
创建集群	cluster	createCluster
删除集群	cluster	deleteCluster
修改集群配置	cluster	modifyCluster
开机	cluster	startCluster
重启	cluster	restartCluster
导入作业	cluster	clusterImportJob
绑定弹性IP	cluster	bindEip
解绑弹性IP	cluster	unbindEip
创建连接	link	createLink
修改连接	link	modifyLink
测试连接	link	verifyLink
删除连接	link	deleteLink
创建任务	job	createJob
修改任务	job	modifyJob
删除任务	job	deleteJob
启动任务	job	startJob

操作名称	资源类型	事件名称
停止任务	job	stopJob

数据仓库服务

CDM支持从数据仓库服务（Data Warehouse Service，简称DWS）导入、导出数据。

关系型数据库服务

CDM支持从关系型数据库服务（Relational Database Service，简称RDS）导入、导出数据，包括云数据库 MySQL、云数据库 PostgreSQL、云数据库 SQL Server。

文档数据库服务

CDM支持从文档数据库服务（Document Database Service，简称DDS）导出数据，暂不支持导入数据到DDS。

云搜索服务

CDM支持从云搜索服务（Cloud Search Service，简称CSS）导入、导出数据。

数据接入服务

CDM支持导入数据到数据接入服务（Data Ingestion Service，简称DIS），从DIS导出时，目前只支持导出到云搜索服务CSS。

表格存储服务

CDM支持从表格存储服务（CloudTable Service，简称CloudTable）导入、导出数据。

数据湖探索服务

CDM支持导入数据到数据湖探索服务（Data Lake Insight，简称DLI），暂不支持从DLI导出数据。

弹性文件服务

CDM支持从弹性文件服务（Scalable File Service，简称SFS）导入、导出数据。

分布式消息服务

CDM支持导入数据到分布式消息服务（Distributed Message Service，简称DMS），从DMS导出时，目前只支持导出到CSS。

数据治理中心

CDM服务可以作为数据治理中心（DataArts Studio）服务的数据集成组件，可与DataArts Studio其他组件配合完成数据迁移和周期调度等任务。

10 约束与限制

CDM 系统级限制和约束

1. DataArts Studio实例赠送的CDM集群，推荐作为DataArts Studio管理中心数据连接的Agent代理使用，不建议同时作为Agent代理和运行数据迁移作业使用。
2. 用于运行数据迁移作业的其他规格CDM集群可以在DataArts Studio控制台以增量包的形式购买，也可以在云数据迁移CDM服务控制台直接购买。二者差异体现在如下方面：
   a. 套餐计费：在DataArts Studio控制台购买的CDM集群，套餐计费时仅支持在DataArts Studio控制台购买的套餐包；在CDM控制台购买的CDM集群，套餐计费仅支持在云数据迁移CDM服务控制台购买的折扣套餐。
   b. 权限控制：在DataArts Studio控制台购买的CDM集群，按照DataArts Studio的权限体系进行权限管理；在CDM控制台购买的CDM集群，按照云数据迁移CDM服务的权限体系进行权限管理。
   c. 使用场景：在DataArts Studio控制台购买的CDM集群按工作空间隔离，需要在关联的工作空间使用；在CDM控制台购买的CDM集群，不支持DataArts Studio工作空间级别的资源隔离，所有DataArts Studio工作空间均可使用。
3. 集群创建好以后不支持修改规格，如果需要使用更高规格的，需要重新创建一个集群。
4. CDM集群为ARM或X86版本，依赖于底层资源的架构。
5. CDM暂不支持控制迁移数据的速度，请避免在业务高峰期执行迁移数据的任务。
6. 在迁移数据时CDM会对数据源端产生压力。建议创建新的数据库账号用于数据迁移，并配置账号策略用于以限制迁移对数据源端的资源消耗。例如可配置当CPU使用率超30%就清理该账号下的连接，从而避免影响业务。
7. 当前CDM集群cdm.large实例规格网卡的基准/最大带宽为0.8/3 Gbps，单个实例一天传输数据量的理论极限值在8TB左右。同理，cdm.xlarge实例规格网卡的基准/最大带宽为4/10 Gbps，理论极限值在40TB左右；cdm.4xlarge实例规格网卡的基准/最大带宽为36/40 Gbps，理论极限值在360TB左右。对传输速度有要求的情况下可以使用多个数据集成实例实现。

   上述数据量为理论极限值，实际传输数据量受数据源类型、源和目的数据源读写性能、带宽等多方面因素制约，实测cdm.large规格最大可达到约8TB每天（大文件迁移到OBS场景）。推荐用户在正式迁移前先用小数据量实测进行速度摸底。
8. 迁移文件或对象时支持文件级增量迁移（通过配置跳过重复文件实现），但不支持断点续传。

例如要迁移3个文件，第2个文件迁移到一半时由于网络原因失败，再次启动迁移任务时，会跳过第1个文件，从第2个文件开始重新传，但不能从第2个文件失败的位置重新传。

9. 文件迁移时，单个任务支持千万数量的文件，如果待迁移目录下文件过多，建议拆分到不同目录并创建多个任务。
10. 用户在CDM上配置的连接和作业支持导出到本地保存，考虑到密码的安全性，CDM不会将对应数据源的连接密码导出。因此在将作业配置重新导入到CDM前，需要手工编辑导出的JSON文件补充密码或在导入窗口配置密码。
11. 不支持集群自动升级到新版本，需要用户通过作业的导出和导入功能，实现升级到新版本。
12. 在无OBS的场景下，CDM系统不会自动备份用户的作业配置，需要用户通过作业的导出功能进行备份。
13. 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。
14. CDM迁移，当目的端为DWS和NewSQL的时候，不支持将源端的主键和唯一索引等约束一起迁移过去。
15. CDM迁移作业时，需确保两个集群版本的JSON文件格式保持一致，才可以从将源集群的作业导入到目标集群。
16. 作业运行过程中，任务异常中断，目标端已写入的部分数据不会清理，需手动清理。
17. 单文件传输大小不超过1TB。

数据库迁移通用限制和约束

1. CDM以批量迁移为主，仅支持有限的数据库增量迁移，不支持数据库实时增量迁移，推荐使用数据复制服务（DRS）来实现数据库增量迁移到RDS。
2. CDM支持的数据库整库迁移，仅支持数据表迁移，不支持存储过程、触发器、函数、视图等数据库对象迁移。

   CDM仅适用于一次性将数据库迁移到云上的场景，包括同构数据库迁移和异构数据库迁移，不适合数据同步场景，比如容灾、实时同步。
3. CDM迁移数据库整库或数据表失败时，已经导入到目标表中的数据不会自动回滚，对于需要事务模式迁移的用户，可以配置“先导入到阶段表”参数，实现迁移失败时数据回滚。

   极端情况下，可能存在创建的阶段表或临时表无法自动删除，也需要用户手工清理（阶段表的表名以“_cdm_stage”结尾，例如：cdmtet_cdm_stage）。
4. CDM访问用户本地数据中心数据源时（例如本地自建的MySQL数据库），需要用户的数据源可支持Internet公网访问，并为CDM集群实例绑定弹性IP。这种方式下安全实践是：本地数据源通过防火墙或安全策略仅允许CDM弹性IP访问。
5. 仅支持常用的数据类型，字符串、数字、日期，对象类型有限支持，如果对象过大会出现无法迁移的问题。
6. 仅支持数据库字符集为GBK和UTF-8。
7. 字段名不可包含&和%。
8. jdbc2hive，hive2jdbc整库迁移的实现机制就是按字段名称映射的，不支持字段名称不一致的迁移场景。

关系数据库迁移权限配置

常见关系数据库迁移需要的最小权限级：

- MySQL：INFORMATION_SCHEMA库的读权限，以及对数据表的读权限。
- Oracle：需要该用户有resource角色，并在tablespace下有数据表的select权限。
- 达梦：具有该schema下select any table的权限。
- DWS：需要表的schema usage权限和数据表的查询权限。
- SQL Server：用户需要有sysadmin权限。
- PostgreSQL：角色拥有数据库下schema下表的select权限。

FusionInsight HD 和 Apache Hadoop 数据源约束

FusionInsight HD和Apache Hadoop数据源在用户本地数据中心部署时，由于读写Hadoop文件需要访问集群的所有节点，需要为每个节点都放通网络访问。

推荐使用**云专线服务**，解决网络访问的同时，还可以提升迁移速度。

数据仓库服务(DWS)数据源约束

1. DWS主键或表只有一个字段时，要求字段类型必须是如下常用的字符串、数值、日期类型。从其他数据库迁移到DWS时，如果选择自动建表，主键必须为以下类型，未设置主键的情况下至少要有一个字段是以下类型，否则会无法创建表导致CDM作业失败。
   - INTEGER TYPES：TINYINT，SMALLINT，INT，BIGINT，NUMERIC/DECIMAL
   - CHARACTER TYPES：CHAR，BPCHAR，VARCHAR，VARCHAR2，NVARCHAR2，TEXT
   - DATA/TIME TYPES：DATE，TIME，TIMETZ，TIMESTAMP，TIMESTAMPTZ，INTERVAL，SMALLDATETIME

   说明

   2.9.1.200及之前版本的集群，DWS源端暂不支持NVARCHAR2数据类型。
2. DWS字符类型字段认为空字符串（''）是空值，有非空约束的字段无法插入空字符串（''），这点与MySQL行为不一致，MySQL不认为空字符串（''）是空值。从MySQL迁移到DWS时，可能会因为上述原因导致迁移失败。
3. 使用GDS模式快速导入数据到DWS时，需要配置相关安全组或防火墙策略，允许DWS/LibrA的数据节点访问CDM IP地址的25000端口。
4. 使用GDS模式导入数据到DWS时，CDM会自动创建外表（foreign table）用于数据导入，表名以UUID结尾（例如：cdmtest_aecf3f8n0z73dsl72d0d1dk4lcir8cd），作业失败正常会自动删除，极端情况下可能需要用户手工清理。

对象存储服务（OBS）数据源约束

1. 迁移文件时系统会自动并发，任务配置中的“抽取并发数”无效。
2. 不支持断点续传。CDM传文件失败会产生OBS碎片，需要用户到OBS控制台清理碎片文件避免空间占用。
3. 不支持对象多版本的迁移。

4. 增量迁移时，单个作业的源端目录下的文件数量或对象数量，根据CDM集群规格分别有如下限制：大规格集群30万、中规格集群20万、小规格集群10万。

   如果单目录下文件或对象数量超过限制，需要按照子目录来拆分成多个迁移作业。

DLI 数据源约束

- 使用CDM服务迁移数据到DLI时，当前用户需拥有OBS的读取权限。
- 目的端为DLI数据源时，抽取并发数建议配置为1，否则可能会导致写入失败。

Oracle 数据源约束

不支持Oracle实时增量数据同步。

分布式缓存服务（DCS）和 Redis 数据源约束

1. 第三方云的Redis服务无法支持作为源端。如果是用户在本地数据中心或ECS上自行搭建的Redis支持作为源端或目的端。
2. 仅支持Hash和String两种数据格式。

文档数据库服务（DDS）和 MongoDB 数据源约束

从MongoDB、DDS迁移数据时，CDM会读取集合的首行数据作为字段列表样例，如果首行数据未包含该集合的所有字段，用户需要自己手工添加字段。

云搜索服务和 Elasticsearch 数据源约束

1. CDM支持自动创建索引和类型，索引和类型名称只能全部小写，不能有大写。
2. 索引下的字段类型创建后不能修改，只能创建新字段。

   如果一定要修改字段类型，需要创建新索引或到Kibana上用Elasticsearch命令删除当前索引重新创建（数据也会删除）。
3. CDM自动创建的索引，字段类型为date时，要求数据格式为“yyyy-MM-dd HH:mm:ss.SSS Z”，即“2018-08-08 08:08:08.888 +08:00”。

   迁移数据到云搜索服务时如果date字段的原始数据不满足格式要求，可以通过CDM的**字段转换**功能转换为上述格式。

数据接入服务（DIS）和 Kafka 数据源约束

- 消息体中的数据是一条类似CSV格式的记录，可以支持多种分隔符。不支持二进制格式或其他格式的消息内容解析。
- 设置为长久运行的任务，如果DIS系统发生中断，任务也会失败结束。
- 迁移作业源端为MRS Kafka时，字段映射不支持自定义字段。
- 迁移作业源端为DMS kafka时，字段映射支持自定义字段。

表格存储服务（CloudTable）和 HBase 数据源约束

1. CloudTable或HBase作为源端时，CDM会读取表的首行数据作为字段列表样例，如果首行数据未包含该表的所有字段，用户需要自己手工添加字段。
2. 由于HBase的无Schema技术特点，CDM无法获知数据类型，如果数据内容是使用二进制格式存储的，CDM会无法解析。

Hive 数据源约束

- Hive中使用Parquet格式存储时间戳数据时，时间戳的精度为纳秒级别（即精确到毫微秒），即2023-03-27 00:00:00.000。当源端数据精度大于纳秒级别时，字段映射时会对数据进行截取。例如源端数据为2023-03-27 00:00:00.12345，目的端数据会被截取为2023-03-27 00:00:00.123。
- Hive作为迁移的目的时，如果存储格式为Textfile，在Hive创建表的语句中需要显式指定分隔符。例如：

```
CREATE TABLE csv_tbl(
smallint_value smallint,
tinyint_value tinyint,
int_value int,
bigint_value bigint,
float_value float,
double_value double,
decimal_value decimal(9, 7),
timestmamp_value timestamp,
date_value date,
varchar_value varchar(100),
string_value string,
char_value char(20),
boolean_value boolean,
binary_value binary,
varchar_null varchar(100),
string_null string,
char_null char(20),
int_null int
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
"separatorChar" = "\t",
"quoteChar"     = "'",
"escapeChar"    = "\\"
)
STORED AS TEXTFILE;
```

11 计费说明

本小节主要介绍云数据迁移的计费说明，包括计费项、计费方式等。

须知

在您使用CDM的过程中，可能还会产生以下相关服务的费用，敬请知悉：

- OBS服务：数据迁移时，CDM可能会将脏数据写入到OBS服务中，则会产生对象存储服务费用，收费标准请参见**OBS价格详情**。
- EIP服务：如果您为CDM集群开通了公网IP，则会产生弹性公网IP服务费用，收费标准请参见**EIP价格详情**。

计费项

计费项	计费说明
CDM集群实例	• 对您选择的实例规格计费。 • 针对CDM集群实例，提供两种计费方式：按需（小时）计费、折扣套餐（按需资源包）。

计费方式

CDM计费仅包含集群实例费用，提供按需计费和折扣套餐方式供您灵活选择，使用越久越便宜。

- 按需计费

  这种购买方式比较灵活，可以即开即停，按实际使用时长计费。

  - 按需计费只包含集群实例费用，不包含公网流量费用。
  - 集群和具体的区域绑定，购买的集群只能在绑定的区域使用。
  - 购买集群后会自动创建CDM集群，如果需要绑定EIP，用户需要前往集群管理界面自行为CDM集群绑定EIP。

  按需计费方式下各实例的具体价格，请参见**产品价格详情**。
- 折扣套餐（按需资源包）

这种购买方式建立在按需计费的基础之上，是通过预付费购买一定时间期限内的使用量套餐。相对于按需计费更优惠，对于长期使用者，推荐该方式。

- 折扣套餐（按需资源包）购买后，系统不自动创建CDM集群，而是和具体的区域和实例规格绑定，在生效期内的每个计费月内按月提供745小时/月的使用时长，在绑定区域给对应实例规格的CDM集群使用。
- 如果当前绑定区域有1个或多个对应实例规格的CDM集群，则扣费方式是先扣除已购买资源包内的时长额度，超出部分以按需计费的方式进行结算（资源包对应多个集群时，会出现每月订购周期内可使用时长不足的情况）。

  例如购买了1个月的折扣套餐（745小时/月），按区域和实例规格匹配到两个CDM集群后，从当前开始的1个月订购有效期内，两个集群同时使用只能使用745/2=372.5小时，约15.5天，剩余时间内两个集群按照按需计费的方式结算费用。
- 如果当前绑定区域没有对应实例规格的CDM集群，购买折扣套餐后不会消耗所购买的时长；但在生效期内，若未使用CDM集群，折扣套餐也不会延期。建议您先安排好服务使用计划，再购买折扣套餐。
- 如果您希望享受折扣套餐的优惠价格，需要先购买一个“折扣套餐”，再购买一个和“折扣套餐”具有相同区域和规格的“按需计费”集群。
- 如果您先购买一个“按需计费”集群，再购买一个相同区域和规格的“折扣套餐”，则在购买折扣套餐之前已经产生的费用按“按需计费”计费，购买折扣套餐之后的费用按“折扣套餐”计费。

各折扣套餐包在不同规格下的具体价格，请参见**产品价格详情**。

变更配置

在开通CDM时有4种集群规格供您选择，您可根据业务需要选择合适的实例规格。

当集群创建成功后，无法对集群进行规格变更，不过您可以通过删除集群后重建集群，实现变更。

续费

资源包到期后，您可以进行续费以延长资源包的有效期，也可以设置到期自动续费。

到期与欠费

折扣套餐资源包到期后，自动转为按需计费。转按需后如果账号欠费，会根据“客户等级”和“订购方式”定义不同的宽限期和保留期时长，宽限保留期内资源处理和费用详见**宽限期保留期**。

退订

CDM服务运行期间，删除集群则不再按需计费；折扣套餐当前不支持退订。

- 删除CDM集群后无法恢复，一旦删除则不再按需计费或扣除折扣套餐时长，详情请参见**删除集群**。
- 折扣套餐为按需资源包，当前不支持退订，具体详情可查看**不可退订**。

另外在删除集群或退订CDM后，对于在CDM使用过程中可能会产生费用的以下相关服务，请分别退订其资源，避免其依然计费。

- OBS服务：数据迁移时，CDM可能会将脏数据写入到OBS服务中，则会产生对象存储服务费用，收费标准请参见**OBS价格详情**。

- EIP服务：如果您为CDM集群开通了公网IP，则会产生弹性公网IP服务费用，收费标准请参见**EIP价格详情**。

12 安全

12.1 责任共担

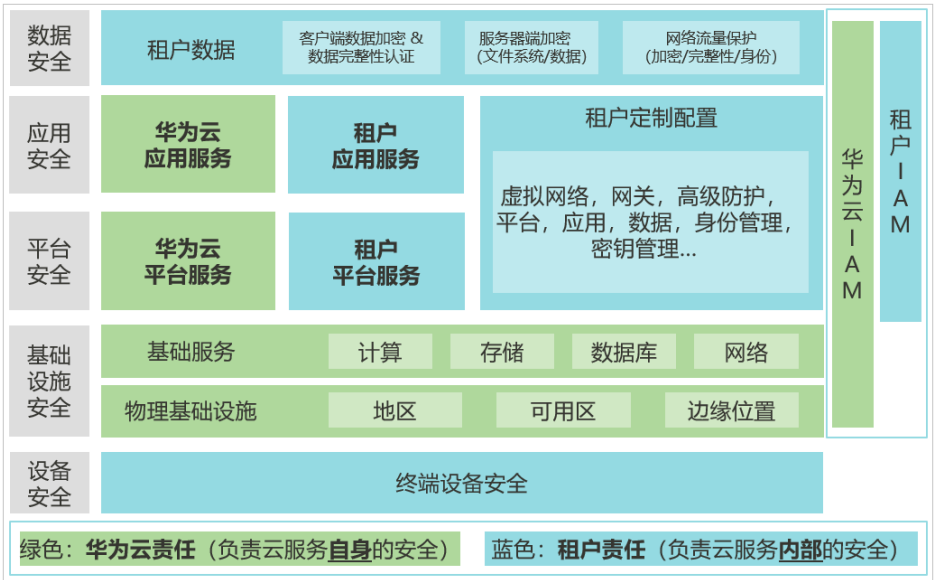
华为云秉承“将公司对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

安全性是华为云与您的共同责任，如**图12-1**所示。

- **华为云**：负责云服务**自身**的安全，提供安全的云。华为云的安全责任在于保障其所提供的 IaaS、PaaS 和 SaaS 类云服务自身的安全，涵盖华为云数据中心的物理环境设施和运行其上的基础服务、平台服务、应用服务等。这不仅包括华为云基础设施和各项云服务技术的安全功能和性能本身，也包括运维运营安全，以及更广义的安全合规遵从。
- **租户**：负责云服务**内部**的安全，安全地使用云。华为云租户的安全责任在于对使用的 IaaS、PaaS 和 SaaS 类云服务内部的安全以及对租户定制配置进行安全有效的管理，包括但不限于虚拟网络、虚拟主机和访客虚拟机的操作系统，虚拟防火墙、API 网关和高级安全服务，各项云服务，租户数据，以及身份账号和密钥管理等方面的安全配置。

《华为云安全白皮书》详细介绍华为云安全性的构建思路与措施，包括云安全战略、责任共担模型、合规与隐私、安全组织与人员、基础设施安全、租户服务与租户安全、工程安全、运维运营安全、生态安全。

图 12-1 华为云安全责任共担模型



12.2 资产识别与管理

云资源的标识与管理可通过标签实现。

使用场景

通常您的业务系统可能使用了华为云的多种云服务，您可以为这些云服务下不同的资源实例分别设置标签（对于CDM而言，标签作用于其集群上），各服务产生的计费详单中都会体现这些资源实例和实例上设置的标签。如果您的业务系统是由多个不同的应用构成，为同一种应用拥有的资源实例设置统一的标签将很容易帮助您对不同的应用进行使用量分析和成本核算。

对CDM来说，标签用于标识购买的集群，以此来达到对购买的CDM集群进行分类的目的。当为集群添加标签时，该集群上所有请求产生的计费话单里都会带上这些标签，从而可以针对话单报表做分类筛选，进行更详细的成本分析。例如：某个集群作用于A部门，我们可以用该部门名称作为标签，设置到被使用的集群上。在分析话单时，就可以通过该部门名称的标签来分析此部门的开发使用成本。

CDM以键值对的形式来描述标签。一个集群默认最大拥有10个标签。每个标签有且只有一对键值。键和值可以任意顺序出现在标签中。同一个集群标签的键不能重复，但是值可以重复，并且可以为空。

使用方式

CDM支持通过控制台方式创建集群标签，详情请参见**创建集群标签**。

12.3 身份认证与访问控制

身份认证

用户访问CDM的方式主要有两种，包括CDM Console界面、Open API等，其本质都是通过CDM提供的REST API接口进行请求。

CDM的接口均需要通过认证鉴权才能访问，控制台发送的请求与调用API接口的请求均支持Token认证鉴权。

访问控制

您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制华为云资源的访问。关于IAM的详细介绍，请参见**IAM产品介绍**。

权限根据授权精细程度分为角色和策略。

- 角色：IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于华为云各服务之间存在业务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。
- 策略：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。例如：不允许某用户组删除集群，仅允许CDM基本操作（如创建、查询作业等）。CDM支持的授权项请参见**权限策略及授权项**。

如**表1**所示，包括了CDM的所有系统权限。

表 12-1 CDM 系统权限

系统角色/策略名称	描述	策略类别
CDM Administrator	操作权限： • CDM管理员权限，可以对CDM资源执行任意操作，拥有该权限的用户必须同时拥有Tenant Guest和Server Administrator权限。 • 拥有VPC Administrator权限的CDM用户可以创建VPC或子网。 • 拥有云监控Administrator权限的CDM用户，可以查看CDM集群的监控指标信息。	系统角色
CDM FullAccess	CDM管理员权限，拥有CDM服务所有权限。	系统策略
CDM FullAccessExceptEIPUpdating	拥有除绑定/解绑EIP外的所有CDM服务权限。	系统策略
CDM CommonOperations	拥有CDM作业和连接的操作权限。	系统策略
CDM ReadOnlyAccess	CDM服务只读权限，拥有该权限的用户仅能查看CDM集群、连接、作业。	系统策略

12.4 数据保护技术

数据存储安全

为了确保您的**个人敏感数据**（例如用户名、密码、手机号码等）不被未经过认证、授权的实体或者个人获取，CDM对用户数据的存储和传输进行加密保护，以防止个人数据泄露，保证您的个人数据安全。

数据销毁机制

用户删除CDM集群后，存储在集群上的用户个人敏感数据会随之删除。

用户在控制台上删去填写的手机号、邮箱，并关闭消息通知功能后，数据库中会同步删除用户的手机号、邮箱信息。

数据传输安全

用户个人敏感数据将通过TLS 1.2、TLS1.3进行传输中加密，所有华为云CDM的API调用都支持 HTTPS 来对传输中的数据进行加密。

12.5 审计与日志

云审计服务（Cloud Trace Service，CTS），是华为云安全解决方案中专业的日志审计服务，提供对各种云资源操作记录的收集、存储和查询功能，可用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。

CTS可记录的CDM操作列表详见**支持云审计的关键操作**。用户开通**开通云审计服务**并创建和配置追踪器后，CTS开始记录操作事件用于审计，用户可查看CTS保存最近7天的审计日志。

CTS支持**配置关键操作通知**。用户可将与IAM相关的高危敏感操作，作为关键操作加入到CTS的实时监控列表中进行监控跟踪。当用户使用CDM服务时，如果触发了监控列表中的关键操作，那么CTS会在记录操作日志的同时，向相关订阅者实时发送通知。

12.6 服务韧性

CDM通过流量限制、备份恢复等技术方案，保障数据的持久性和可靠性。

12.7 监控安全风险

CDM提供基于云监控服务CES的资源监控能力，帮助用户监控账号下的CDM集群，执行自动实时监控、告警和通知操作。用户可以实时掌握集群运行中所产生的网络流入速率、网络流出速率、CPU使用率、内存使用率、磁盘利用率、失败作业率等信息

关于CDM支持的监控指标，以及如何创建监控告警规则等内容，请参见**查看监控指标**。

12.8 故障恢复

CDM集群支持**定时备份**功能，开启后可以将作业定时备份到OBS上，当服务故障后，可以通过作业导入功能恢复作业。

12.9 更新管理

更新漏洞

CDM云服务通过华为云安全公告密切关注漏洞，如Apache Log4j2 远程代码执行漏洞（CVE-2021-44228）、Fastjson存在反序列化漏洞（CNVD-2022-40233）等，如发现服务模块涉及漏洞影响，会迅速通过官方解决方案升级现网更新漏洞。

更新配置

CDM云服务通过版本更新升级更新配置。

12.10 认证证书

合规证书

华为云服务及平台通过了多项国内外权威机构（ISO/SOC/PCI等）的安全合规认证，用户可自行**申请下载**合规资质证书。

图 12-2 合规证书下载

资源中心

华为云还提供以下资源来帮助用户满足合规性要求，具体请查看**资源中心**。

图 12-3 资源中心

销售许可证&软件著作权证书

另外，华为云还提供了以下销售许可证及软件著作权证书，供用户下载和参考。具体请查看**合规资质证书**。

图 12-4 销售许可证&软件著作权证书

13 配额说明

CDM服务应用的基础设施如下：

- 弹性云服务器
- 虚拟私有云
- 弹性公网IP
- 消息通知服务
- 统一身份认证服务

其配额查看及修改请参见**关于配额**。

14 权限管理

如果您需要对CDM集群，给企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制云资源的访问。

通过IAM，您可以在云账号中给员工创建IAM用户，并使用策略来控制其对云资源的访问范围。例如您的员工拥有CDM的使用权限，但是不希望其拥有删除CDM集群等高危操作的权限，那么您可以使用IAM为员工创建IAM用户，通过授予仅能使用CDM服务，但是不允许删除CDM集群的权限策略，实现控制其对CDM的使用范围。

如果云账号已经能满足您的要求，不需要创建独立的IAM用户进行权限管理，您可以跳过本章节，不影响您使用CDM的其它功能。

IAM是华为云提供权限管理的基础服务，无需付费即可使用，您只需要为您账号中的资源进行付费。关于IAM的详细介绍，请参见**IAM产品介绍**。

CDM 权限

默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略或角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

CDM部署时通过物理区域划分，为项目级服务，需要在各区域（如华北-北京1）对应的项目（cn-north-1）中设置相关权限，并且该权限仅对此项目生效。如果需要所有区域都生效，则需要在所有项目都设置权限。访问CDM时，需要先切换至授权区域。

权限根据授权精细程度分为角色和策略。

- 角色：IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于华为云各服务之间存在业务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。
- 策略：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。例如：不允许某用户组删除集群，仅允许CDM基本操作（如创建、查询作业等）。CDM支持的授权项请参见**权限策略及授权项**。

如**表1**所示，包括了CDM的所有系统权限。

表 14-1 CDM 系统权限

系统角色/策略名称	描述	策略类别
CDM Administrator	操作权限： • CDM管理员权限，可以对CDM资源执行任意操作，拥有该权限的用户必须同时拥有Tenant Guest和Server Administrator权限。 • 拥有VPC Administrator权限的CDM用户可以创建VPC或子网。 • 拥有云监控Administrator权限的CDM用户，可以查看CDM集群的监控指标信息。	系统角色
CDM FullAccess	CDM管理员权限，拥有CDM服务所有权限。	系统策略
CDM FullAccessExceptEIPUpdating	拥有除绑定/解绑EIP外的所有CDM服务权限。	系统策略
CDM CommonOperations	拥有CDM作业和连接的操作权限。	系统策略
CDM ReadOnlyAccess	CDM服务只读权限，拥有该权限的用户仅能查看CDM集群、连接、作业。	系统策略

表14-2列出了CDM常用操作与系统权限的授权关系，您可以参照该表选择合适的系统权限。

表 14-2 常用操作与系统权限的关系

操作	CDM FullAccess	CDM FullAccessExceptEIPUpdating	CDM CommonOperations	CDM ReadOnlyAccess
创建集群	√	√	×	×
集群绑定/解绑EIP	√	×	×	×
查询集群列表	√	√	√	√
查询集群详情	√	√	√	√
重启集群	√	√	×	×
修改集群配置	√	√	×	×

操作	CDM FullAccess	CDM FullAccessExceptEIPUpdating	CDM CommonOperations	CDM ReadOnlyAccess
删除集群	√	√	×	×
创建连接	√	√	√	×
查询连接	√	√	√	√
修改连接	√	√	√	×
删除连接	√	√	√	×
创建作业	√	√	√	×
查询作业	√	√	√	√
修改作业	√	√	√	×
启动作业	√	√	√	×
停止作业	√	√	√	×
查询作业状态	√	√	√	√
查询作业执行历史	√	√	√	√
删除作业	√	√	√	×