数据集成

产品简介

产品简介

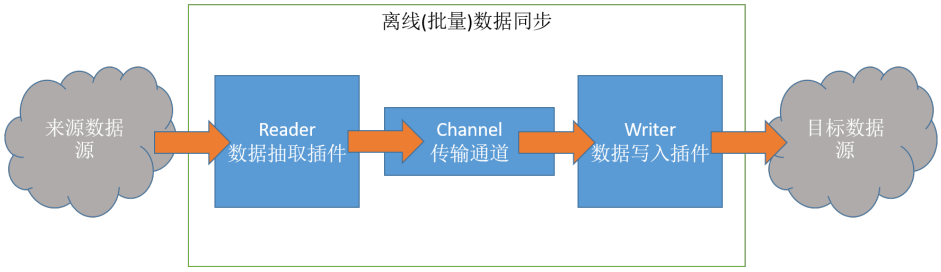
数据集成概述

数加·数据集成，是阿里集团对外提供的稳定高效、弹性伸缩的数据同步平台。 致力于提供复杂网络环境下、丰富的异构数据源之间数据高速稳定的数据移动及同步能力。丰富的数据源支持:文本存储(FTP/SFTP/OSS/多媒体文件 等)、数据库(RDS/DRDS/MySQL/PostgreSQL 等)、NoSQL(Memcache/Redis/MongoDB/HBase 等)、大数据(MaxCompute/ AnalyticDB/HDFS 等) 、MPP数据库（ HybridDB for MySQL等 ）

数据集成简介

离线(批量)的数据通道主要通过定义数据来源和去向的数据源和数据集，提供一套抽象化的数据抽取插件(称之为 Reader)、数据写入插件(称之为 Writer)，并基于此框架设计一套简化版的中间数据传输格式，从而达到任意结构化、半结构化数据源之间数据传输之目的 。

可以参考下图：



支持数据源类型

支持的数据源类型情况请参见：**支持数据源类型**。
由于每个数据源的配置信息差距较大，需要根据使用情况详细查询参数配置信息 。在数据源配置、作业配置页面提供了详细描述，请根据自己情况进行查询使用 。

同步开发说明

同步开发提供两种开发模式：向导模式、脚本模式 。

- 向导模式：提供向导式的开发引导，通过可视化的填写和下一步的引导，帮助快速完成数据同步任务的配置工作 。向导模式的学习成本低，但无法享受到一些高级功能 。

- 脚本模式：用户可以通过直接编写数据同步的 JSON 脚本来完成数据同步开发，适合高级用户，学习成本较高 。脚本模式可以提供更丰富灵活的能力，做精细化的配置管理 。

注意：

向导模式生成的代码可以转换为脚本模式，此转换为单向操作，转换完成后无法恢复到向导模式。因为脚本模式能力是向导模式的超集 。
代码编写前需要完成数据源的配置和目标表的创建 。

网络类型说明

网络类型分为：经典网络、专有网络(VPC)、本地 IDC 网络（规划中）。

- 经典网络：统一部署在阿里云的公共基础网络内，网络的规划和管理由阿里云负责，更适合对网络易用性要求比较高的客户 。
- 专有网络：基于阿里云构建出一个隔离的网络环境 。您可以完全掌控自己的虚拟网络，包括选择自有的 IP 地址范围，划分网段，以及配置路由表和网关 。
- 本地 IDC 网络：用户自身构建机房的网络环境，与阿里云网络是隔离不可用的 。

经典网络和专有网络相关问题请参见： 经典网络和 VPC 常见问题 FAQ 。

补充说明：

网络连接可以支持公网连接，网络类型选择经典网络即可 。需要注意公网带宽的速度和相关网络费用消耗。无特殊情况不建议使用 。
规划中的网络连接，进行数据同步，可以使用本地新增运行资源+脚本模式的方案进行数据同步传输。或者使用 SHELL+DataX 方案，此方案请参见：https://help.aliyun.com/document_detail/45055.html 。

专有网络 VPC 是构建一个隔离的网络环境，可以自定义 IP 地址范围、网段、网关等随着专有网络安全性提高，专有网络运用越来越广，所以数据集成提供了 RDS-MySQL、RDS-SQL Server、RDS-PostgreSQL 在专有网络下不需要购买一台跟 VPC 同网络的 ECS，系统通过反向代理会自动检测从而网络能够互通。对于阿里云其他的数据库 PPAS、OceanBase、Redis、MongoDB、Memcache、TableStore、HBase 在不久的将来会支持。所以非 RDS 的数据源在专有网络下配置数据集成的同步任务需要购买同网络的 ECS,这样可以通过 ECS 连通网络 。

约束与限制

支持且仅支持结构化(例如 RDS、DRDS 等)、半结构化、无结构化(OSS、TXT 等, 要求具体同步数据必须抽象为结构化数据)的数据同步 。换言之，Data Integration 支持传输能够抽象为逻辑二维表的数据同步，其他完全非结构化数据，例如 OSS 中存放的一段 MP3，Data Integration 暂未支持将其同步到 MaxCompute，这个功能会在后期实现 。

支持单个和部分跨 region 地域内数据存储相互同步、交换的数据同步需求 。

部分地域通过经典网络是可以传输的，不能保证 。如果必须使用且测试经典网络不通，可以考虑使用公网方式连接 。

仅完成数据同步(传输)，本身不提供数据流的消费方式 。

参考文档：

数据同步任务配置的详细介绍请参见： 快速开始->创建数据同步任务 。

若处理像 OSS 等非结构化数据的详细介绍请参见：MaxCompute 访问 OSS 数据 。

支持的数据源类型

数据集成是阿里集团对外提供的稳定高效、弹性伸缩的数据同步平台，为阿里云大数据计算引擎（包括 MaxCompute、AnalyticDB、OSS）提供离线（批量）的数据进出通道。

数据同步支持的数据源类型如下表所示：

数据源分类	数据源类型	抽取（Reader）	导入（Writer）	支持方式	支持类型
关系型数据库	MySQL	支持	支持	向导/脚本	阿里云/自建
关系型数据库	SQL Server	支持	支持	向导/脚本	阿里云/自建
关系型数据库	PostgreSQL	支持	支持	向导/脚本	阿里云/自建
关系型数据库	Oracle	支持	支持	向导/脚本	自建
关系型数据库	DRDS	支持	支持	向导/脚本	阿里云
关系型数据库	DB2	支持	支持	脚本	自建
关系型数据库	达梦（对应数据源名称是 dm）	支持	支持	脚本	自建
关系型数据库	RDS for PPAS	支持	支持	脚本	阿里云
MPP	HybridDB for MySQL	支持	支持	向导/脚本	阿里云

MPP	HybridDB for PostgreSQL	支持	支持	向导/脚本	阿里云
大数据存储	MaxCompute（对应数据源名称 odps）	支持	支持	向导/脚本	阿里云
大数据存储	DataHub	不支持	支持	脚本	阿里云
大数据存储	ElasticSearch	不支持	支持	脚本	阿里云
大数据存储	AnalyticDB（对应数据源名称 ADS）	不支持	支持	向导/脚本	阿里云
非结构化存储	OSS	支持	支持	向导/脚本	阿里云
非结构化存储	HDFS	支持	支持	脚本	自建
非结构化存储	FTP	支持	支持	向导/脚本	自建
NoSQL	HBase	支持	支持	脚本	阿里云/自建
NoSQL	MongoDB	支持	支持	脚本	阿里云/自建
NoSQL	Memcache	不支持	支持	脚本	阿里云/自建 Memcached
NoSQL	Table Store（对应数据源名称 OTS）	支持	支持	脚本	阿里云
NoSQL	LogHub	不支持	支持	脚本	阿里云
NoSQL	OpenSearch	不支持	支持	脚本	阿里云
NoSQL	Redis	不支持	支持	脚本	阿里云/自建
性能测试	Stream	支持	支持	脚本	

各数据源测试连通性支持情况

数据源	数据源类型	网络类型	是否支持测试连通性	是否添加自定义资源组

MySQL	云数据库	经典网络	支持	-
		专有网络	支持	-
	有公网IP		支持	-
	无公网IP		不支持	添加自定义资源组
	ECS自建	经典网络	支持	-
		专有网络	不支持	添加自定义资源组
SQL Server	云数据库	经典网络	支持	-
		专有网络	支持	-
	有公网IP		支持	-
	无公网IP		不支持	添加自定义资源组
	ECS自建	经典网络	支持	-
		专有网络	不支持	添加自定义资源组
PostgreSQL	云数据库	经典网络	支持	-
		专有网络	支持	-
	有公网IP		支持	-
	无公网IP		不支持	添加自定义资源组
	ECS自建	经典网络	支持	-
		专有网络	不支持	添加自定义资源组
Oracle	有公网IP		支持	-
	无公网IP		不支持	添加自定义资源组
	ECS自建	经典网络	支持	-
		专有网络	不支持	添加自定义资源组
DRDS	云数据库	经典网络	支持	-
		专有网络	排期中	添加自定义资源组
HybridDB for MySQL	云数据库	经典网络	支持	-
		专有网络	排期中	添加自定义资源组
HybridDB for PostgreSQL	云数据库	经典网络	支持	-

		专有网络	排期中	添加自定义资源组
MaxCompute（对应数据源名称是 odps）	云数据库	经典网络	支持	-
		专有网络	支持	-
AnalyticDB（对应数据源名称 ADS）	云数据库	经典网络	支持	-
		专有网络	排期中	添加自定义资源组
OSS	云数据库	经典网络	支持	-
		专有网络	支持	-
Hdfs	有公网IP		支持	-
	ECS自建	经典网络	支持	-
		专有网络	不支持	添加自定义资源组
FTP	有公网IP		支持	-
	无公网IP		不支持	添加自定义资源组
	ECS自建	经典网络	支持	-
		专有网络	不支持	添加自定义资源组
MongoDB	云数据库	经典网络	支持	-
		专有网络	排期中	添加自定义资源组
	有公网IP		支持	-
	ECS自建	经典网络	支持	-
		专有网络	不支持	添加自定义资源组
Memcache	云数据库	经典网络	支持	-
		专有网络	排期中	添加自定义资源组
Redis	云数据库	经典网络	支持	-
		专有网络	排期中	添加自定义资源组
	有公网IP		支持	-
	ECS自建	经典网络	支持	-
		专有网络	不支持	添加自定义资源组
Table Store（对应数据源名称是	云数据库	经典网络	支持	-
		专有网络	排期中	添加自定义资源

OTS ）				组

对上面的几种情况进行说明：

上述表格中的 - 表示没有此种说法，**不支持** 并不代表不能配置同步任务，只是 **单击测试连通性无效**，需要添加自定义资源组。

VPC 环境数据源：

VPC 环境的 RDS 数据源支持测试连通性。

其他数据源 VPC 网络正在排期。

金融云网络暂时不支持测试连通性。

ECS 自建数据源：

经典网络支持 JDBC 的格式测试连通性，一般是走公网。

VPC 环境暂时不支持测试连通性。

跨区域暂时不支持测试连通性。

金融云网络暂时不支持测试连通性。

目前要实现数据同步都是添加自定义资源组的方法，详情请参见 VPC 环境数据同步配置（金融云）。

关于 ECS 自建的数据源，需要特别注意安全组的添加，在 ECS 安全组中入/出方向添加调度集群的 IP（公网和经典网络都要在对应的入/出方向添加），如果没有添加相应的安全组同步会出现相应的连接不上的问题。详情请参见 如何添加安全组。大的端口范围无法在 ECS 安全组界面添加，请使用 ECS 的安全组 API 进行添加，详情请参见 AuthorizeSecurityGroup。

没有公网 IP 本地 IDC 机房或 ECS 搭建的数据源：

不支持测试连通性。

配置同步任务要添加自定义资源组。

更多详情请参见 同步数据库的数据（无公网IP）。

有公网 IP 本地 IDC 机房或 ECS 搭建的数据源：

一律走公网 JDBC 格式，如果测试连通性失败，则检查您本地网络的限制或者数据库本身的限制。

关于调度集群情况

调度集群目前在华东 2、华南 1、香港、新加坡均有部署，目前根据经验，以调度集群在 **华东 2** 为准和用户数据源做对比。假设用户的 MongoDB 数据源在华北经典网络以我们调度集群在华东 2 经典网络为准，跨区域是不通的。

OXS 集群和 ECS 集群，内网不通。

RDS 的调度集群是 OXS，OXS 集群是和内网大陆所有区域的 RDS 是互通的。其他数据源的调度集群是另外一套 ECS 经典网络的调度集群。

比如 RDS 同步到自建数据库测试的时候，RDS 和自建数据库数据源测试连通性都能成功。但实际调度的时候，RDS 会下发到 OXS 调度集群，自建的会跑到 ECS 集群，RDS 和 ECS 集群不通，所以会失败。一般建议您将 RDS 改为 MySQL > JDBC 方式，这样都会跑到 ECS 集群上连接是可以通的。

怎么查看任务下发执行集群

出现 RDS 作为数据源的时候，任务会到 OXS 集群同步。日志如下图所示：

当数据源为其他数据源在是 ECS 调度集群，如下图所示：

当调度集群为自定义调度资源时，日志如下图所示（非常重要，用于判断用户是否是自定义资源组）：

进入数据集成测试页面，直接单击 **运行**，统一走的是 ECS 调度集群，所以会有用户反馈，RDS 相关任务手动运行成功，调度失败。因为 RDS 作为数据源跨区域时候，需要在 OXS 调度集群执行。所以需要您进行 **调度运维 > 测试运行**。

注意：

关于数据源的测试连通性，可能大家最想了解的走公网的收费情况。下面以 RDS 同步到 MaxCompute 收费问题为例：

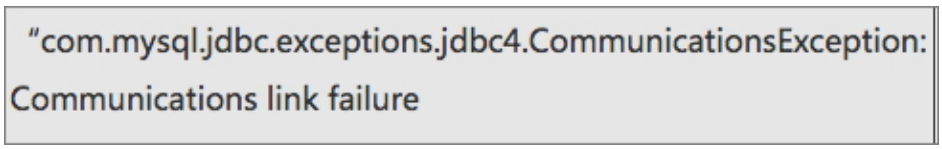
数据集成目前是不收费的，但有可能涉及一些收费的产品，这个可以参考涉及的收费产品来判断，DataWorks 配置关于 MaxCompute 的数据同步，都是不收费的。除非您自己在脚本模式自己加参数配置 MaxCompute 的 Tunnel 的公网地址，会收费（实际上我们在脚本模式生成的模板中没有提供这个参数）。

总结

当遇到测试连通性失败，一般有下面两种情况：

常见错误实例，填写信息错误，核实数据源数据库名称，用户名、密码是否正确：

没有连接上数据库，核实数据源区域，网络类型，白名单是否添加完整，实例 ID 等相关信息：



同步过程中出现网络断开等。

首先要完整日志，看下调度资源是哪个，是否是自定义资源。

如果是自定义资源，核实自定义资源组的 IP 是否添加到数据源比如 RDS 白名单（MongoDB也是有白名单限制的，也需要添加）。

核实两端数据源连通性是否通过，核实 RDS，MongoDB 白名单是否会添加完整（如果不完整，有时候会成功有的时候会失败，如果任务下发到已添加的调度服务器上会成功，没添加的会失败）。

任务显示成功，但是日志出现 8000 断开报错。

出现上述报错，是因为用户使用的自定义调度资源组，没有对 10.116.134.123，访问 8000 端口在安全组内网入方向放行，添加后重新运行即可。

关于测试连通性失败的示例

示例一

问题现象：

测试连接失败，测试数据源连通性失败。连接数据库失败，数据库连接串：jdbc:mysql://xx.xx.xx.x:xxxx/t_uoer_bradef， 用户名：xxxx_test，异常消息：Access denied for user ‘xxxx_test’ @’ %’ to database ‘yyyy_demo’ 。

排查思路：

确认其添加的信息有没有问题。

密码、白名单或者用户的账号有没有对应数据库的权限，RDS 管控台可以添加授权的。

示例二

问题现象：

测试连接失败，测试数据源连通性失败。报错如下：

```
error message: Timed out after 5000 ms while waiting for a server that matches
ReadPreferenceServerSelector{readPreference=primary}. Client view of cluster state is {type=UNKNOWN,
servers=[(xxxxxxxxxx), type=UNKNOWN, state=CONNECTING,
exception={com.mongodb.MongoSocketReadException: Prematurely reached end of stream}}]
```

排查思路：
非 VPC 的 Mongodb，添加 Mongodb 数据源测试连通性要添加相应的白名单，详情请参见 如何添加白名单
。

基本概念

数据同步

数据同步的定义

广义的数据同步是指为保持两端数据一致性而进行的数据传输过程。一般来讲，数据集成的数据同步是为保证源宿两端数据逻辑的一致性，将数据从数据源端移到数据目的端，并伴随一定的数据转换或者清洗的过程。在数据集成的功能边界中，数据同步定义为云上各种存储产品之间进行的数据转移过程。

数据同步的要素

数据集成同步核心概念主要由三个要素构成：

- 数据源：指数据同步的数据源存储，包括寻址信息（IP地址、库等信息，用以同步寻址）、同步内容

（同步的表、字段信息等）、控制信息（编码清洗等）。
- 数据目的端：指数据同步的数据目的端存储，包括寻址信息（IP地址、库等信息，用以同步寻址）以及同步内容（同步的表）、控制信息（脏数据处理等）。

- 数据转换过程：指数据同步过程中存在的数据转换过程，泛指数据的计算、清洗等过程，该过程不是必要条件。

数据同步的种类

离线数据同步

离线数据同步指的是数据周期性（例如每天、每周、每月等）、成批量地从源端系统传输到目标端系统。对于离线数据同步系统，数据以读取Snapshot（快照）的方式从源端传输到目的端。离线同步存在生命周期，一个离线同步的任务有开始状态同样也有结束状态。数据集成中是使用Job概念来描述和定义离线同步任务。

流式数据同步

数据以实时或者准实时将变化的变更日志从源端系统传输到目标端系统。对于流式数据同步系统，数据以Stream（变更流水）的方式从源端传输到目的端。实时同步不存在任务自动结束，而将数据的变化日志同步一直持续下去。

无论是数据流式同步还是离线同步（批处理数据同步），同步的过程都包含上述同步核心要素，也即提取E（Extract）、转换T（Transform）、加载L（Load）。

Job（作业）

Job是数据集成进行数据批量同步的基本业务单位，数据集成的Job面向表级别数据同步，Job描述了一个数据同步作业完成一次数据同步任务所需要的信息，包括E（Extract）、T（Transform）、L（Load）等用户描述信息，也包括作业的运行信息，例如同步数据量、同步速率、当前进度等计量信息，还包括生命周期等，Job运行完毕即完成了一次数据同步工作。

作用

作业模型

数据集成本身不保存作业信息，数据集成对用户提交每一次作业都生成一个Job对象，并为其分配了

唯一的Job ID。对于用户多次提交同一个作业，数据集成识别为多次提交，并分配多个Job ID。即对于数据集成同步任务（批处理同步和流式同步）而言，数据基层提供触发式任务服务能力。 类似于Hadoop的作业概念模型，数据集成将提交的一个实例化作业抽象为Job，运行一次即是一个独立的Job。

调度模型

作业速率上限是指数据同步作业可能达到的最高速率，其最终实际速率受网络环境、数据库配置等影响。

单并发同步作业：作业并发数 * 单并发的传输速率 = 作业传输总速率。

在作业速率上限已选定的情况下，应该如何选择作业并发数？

如果您的数据源是线上的业务库，建议您不要将并发数设置过大，以防对线上的业务库造成影响。

如果您特别在意数据同步速率，建议您选择最大作业速率上限和较大的作业并发数。作业速率上限和作业并发数在json里的表现形式。

mbps：表示作业并发的速率上限，例如： “mbps”：“1” ，表示作业速率上限是1MB/S。

concurrent：表示并发的数目，例如：” concurrent”：“1” ，表示作业并发的数目为1。

约束限制

数据集成暂未能实现对数据源schema信息同步功能，因此用户需要提前在目的端数据源进行建表操作，并且最好做到目标表的字段个数、类型与源端大致一致。

数据集成按照源宿两端Column的进行传输，而不是依靠Column名称或者类型进行，是根据相关的映射情况进行传输，例如源端Column为a，b，c三列，目标端为x，y，z三列。数据集成将源端数据a，b，c按照目标端数据x，y，z顺序导入。

数据集成本身存在字段类型隐式转换规则，支持常见的转换规则例如整形、浮点型可以自动转为字符串类型。

权限和安全

用户角色

为最简化底层权限模型，用户角色分为：项目管理员、开发、运维、部署、访客。复杂的权限模型体系，只提

供最基本的权限模型，才能更好的管理用户的账号。一般主账号默认是项目管理员角色，主账号可以给子账号赋予相应的权限。对应项目的角色，权限概述如下：

角色	平台权限特征
项目管理员	指项目空间的管理者，可对该项目空间的基本属性、数据源、当前项目空间计算引擎配置和项目成员等 进行管理，并为项目成员赋予项目管理员、开发、运维、部署、访客角色。
开发	开发角色的用户能够创建工作流、脚本文件、资源和UDF，新建/删除表，同时可以创建发布包，但不能执行发布操作。
运维	运维角色的用户由项目管理员分配运维权限；拥有发布及线上运维的操作权限，没有数据开发的操作权限。
部署	部署角色与运维角色相似，但是它没有线上运维的操作权限。
访客	访客角色的用户只具备查看权限，没有权限进行编辑工作流和代码等操作。

系统隔离

数据集成支持系统**多租户隔离**（一个主账号就是一个租户，两个租户间任务不会互相影响），类似MaxCompute，数据集成使用多租户隔离概念做系统权限和运行资源的隔离。系统权限隔离指不同租户下的用户相互之间是无法管控对方的租户及下属所有对象信息，包括Job配置信息、Job传输数据流信息。运行资源隔离指不同租户下的用户相互之间环境完全隔离，保证不同租户下的Job运行环境不再相互干扰。

数据源鉴权

数据集成系统仅能负责用户对数据集成请求API鉴权，但无法负责对用户请求数据同步的源端和宿端的权限进行鉴别。目前数据集成使用的策略是数据集成不参与数据源鉴权，让用户在数据集成Job提交过程传递鉴权信息，数据集成透传该鉴权信息到两端数据源进行鉴权。

同时，为了避免用户鉴权信息（例如AccessId、AccessKey等敏感字段）泄露，数集成本身提供了一套安全非对称加密方式，保证用户敏感信息不会存在泄密风险。

数据集成SDK向数据集成的服务端使AK签名式请求认证，数据集成认证通过后，给SDK颁发非对称加密的公钥。

数据集成SDK使用该公钥对用户需要传递的敏感信息进公钥加密，随后使该密文进作业启动请求。

数据集成接收到该作业启动请求后，为保证链路的安全性，其数据集成本不保存解密私钥，将该信息直接透传下发到执集群。

执行集群利用第三方安全系统提供的私钥信息，对该敏感信息进行解密操作。该解密操作完全在内存进行，不输出不落地，保证用户敏感信息不外泄。

产品与技术

产品与概念

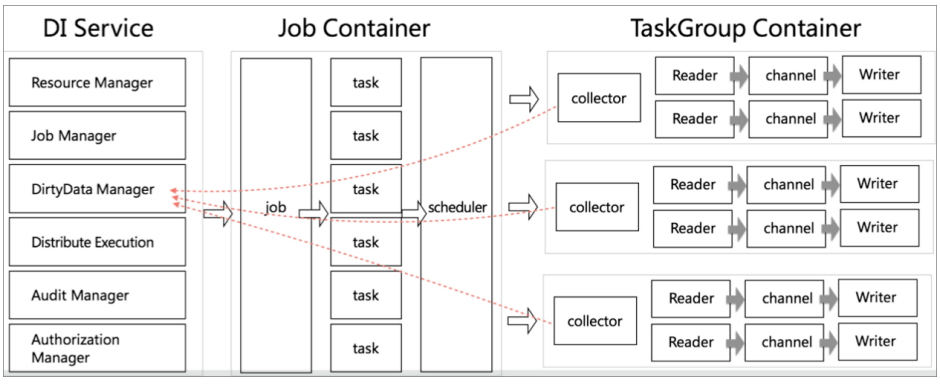
数据集成定义只完成数据同步/传输过程，并且整体数据传输过程完全控制于数据集成的同步集群模型下，同步的通道以及同步数据流对用户完全隔离。同时，数据集成本身不提供传输同步数据流的消费功能，即您不能直接通过数据集成的的API消费数据流，所有针对数据操作，您必须在同步数据流两端存储端操作。

以RDS通过数据集成同步到MaxCompute为例，如下图所示，表格中是数据集成支持的数据类型。

数据集成支持的类型	备注与说明
Long	定点数 Int、Short、Long、BigInteger
Double	浮点数 Float、Double、BigDecimal(无限精度)
String	字符串类型，底层不限长，使用通用字符集(Unicode)
Date	时间类型，包括日期、时间、日期时间类型
Bool	布尔值
Bytes	二进制，可以存放诸如MP3等非结构化数据

技术与原理

数据集成在阿里云上提供一套分布式离线数据同步平台，同时提供一套抽象化的数据抽取插件（称之为Reader）、数据写入插件（称之为Writer），并基于此框架设计一套简化版的中间数据传输格式，从而达到任意结构化、半结构化数据源之间数据传输之目的。从用户角度来看，一个数据集成同步任务运行Job示意图如下所示：



上述中，红色虚箭头是代表通过collector状态收集器监控数据返回到脏数据管理服务器进行数据分析，灰色方向箭头代表数据流向。DI Service主要是包含资源管理器、Job管理器、脏数据管理器、分布式服务、鉴权服务等。Job Container主要是将数据集成运行任务分成若干个task，然后通过scheduler调度管理。TaskGroup Container主要是数据抽取通过数据通道（channel）将数据写入。

使用数据集成Job启动API，向数据集成服务端发起调用，提交一个离线数据同步Job。

数据集成收到Job API请求后，将负责做必要的安全和权限校验，待校验通过后，数据集成会下发相应的Job到执行集群节点启动离线数据同步任务。

Job启动后，根据您提供的源端（Reader）、目的端（Writer）的配置信息，加载并初始化相关插件，连接两端数据源，开始数据同步工作。

Job运行过程中，将随心跳向数据集成汇报当前进度、速度、数据量等关键运行指标，您可根据Job的状态API实时获取该Job运行状态，直至Job运行结束（成功或者失败）。

数据集成术语表

A

- **安全组**

安全组是一个逻辑上的分组，是一种虚拟防火墙，是由同一个地域（Region）内具有相同安全保护需求并相互信任的实例组成，可用于设置单台或多台ECS实例的网络访问控制，是重要的网络安全隔离手段。每个实例至少属于一个安全组，在创建时就需要指定。同一安全组内的实例之间网络互通，不同安全组的实例之间默认内网不通。可以授权两个安全组之间互访。

B

- **白名单**

数据集成连接RDS（MySQL）同步数据需要使用MySQL标准协议连接。RDS默认允许所有IP连接，但如果用户在RDS配置指定了IP白名单，则用户需要添加数据集成执行节点到IP白名单。用户没有指定RDS白名单情况下，不需要给数据集成提供白名单。

C

插件

分为读插件（reader）和写插件（writer），读插件负责将数据从源端存储系统抽取出来并转化为中

间格式，写插件负责将中间格式的数据写入到目标端存储系统。

Console

数据集成提供的基于命令交互式的操作管理具。

错误记录数

错误记录数，表示脏数据的最大容忍条数。示例如下：

如果您配置为0，表示严格不允许脏数据存在。

如果您不填此项，则代表允许存在脏数据，即如果出现脏数据，数据集成会记录并打印部分脏数据，方便您进行排查。

常量

常量是固定值，在程序执行期间不会改变。常量可以是任何的基本数据类型，比如整数常量、浮点常量、字符常量，或字符串字面值。

重跑与幂等

数据集成定位在为各类数据存储提供数据传输通道功能，在定期自动化运行数据同步的场景，如数仓ETL流程，要求所有的数据同步任务能够做到多次同步和单次同步最终结果一致。例如当一次数据同步任务出现Fail，您可以直接重启任务而无需到目的数据端进行线上数据清理操作。这在数仓领域属于作业幂等性要求。数据集成作业的幂等性是通过Writer插件的前置条件来实现的，例如您在MaxCompute配置中提供数据写入前的清理动作，保证每次数据导入前都会先清除当前表或者分区的现有数据，这样能够保证数据多次写入的结果和一次性写入结果一致。

F

分区

分区表是指在创建表时指定分区空间，即指定表内的某几个字段作为分区列。大多数情况下，用户可以将分区类比为文件系统下的目录。

MaxCompute将分区列的每个值作为一个分区（目录）。您可以指定多级分区，即将表的多个字段作为表的分区，分区之间正如多级目录的关系。

J

结构化数据

结构化数据（即行数据，存储在数据库里，可以用二维表结构来逻辑表达实现的数据）。

非结构化数据，包括所有格式的办公文档、文本、图片、图像和音频/视频信息等等。

所谓半结构化数据，就是介于结构化数据和非结构化数据之间的数据，HTML文档就属于半结构化数据。它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。

Job（作业）

Job是同步的基本业务单元，描述了完成一次数据同步所需要的全部配置信息，包括源端配置，目的端配置，出错限制等。

经典网络的IP

目前经典网络IP地址由阿里云统一分配，分为公网IP和私网IP。

每个实例会分配一块私网网卡，并绑定一个私网IP。私网IP是必选的且无法修改。

您购买了公网带宽（即公网带宽不为0Mbps），阿里云会为您的实例分配一块公网网卡，并为网卡配置一个公网IP地址。

L

离线同步

指数据周期性（例如每天、每周、每月等）、成批量地从源端系统传输到目标端系统。对于离线数据同步系统，数据以读取Snapshot（快照）的方式从源端传输到目的端。

流式同步

数据以实时或者准实时的时延，将变化的变更日志从源端系统传输到目标端系统。对于流式数据同步系统，数据以Stream（变更流水）的方式从源端传输到目的端。实时同步不存在任务结束，将数据的变化日志同步一直持续下去。数据集成暂不支持流式数据同步模型。

流量控制

支持对通道流量控制，即用户可以对单个Job分配带宽最大限制。注意流量度量值是数据集成本身的度量值，不代表实际网卡流量。

T

通道

指支持的数据存储类型，如MySQL、MaxCompute等。

同步

一般来讲，数据同步是为保证源宿两端数据逻辑的一致性，将数据从数据源移动到数据目的端，并伴随一定的数据转换或者清洗的过程。

Task

数据集成在进行数据同步过程中，为了提升数据传输吞吐能力，通常对传输数据集进行细粒度切分（称之为Task），并启动多线程乃至于多进程容器运行Task进行数据传输服务。

V

VPC

专有网络VPC构建逻辑隔离网络, 增强不同环境的隔离性、减少共享网络带来的卡顿、以及尽量避免业务规模发展后可能会遇到的安全性问题。

Z

增量同步

数据集成通过使用where过滤条件做增量抽取，具体来讲，在源表上增加个时间戳字段，系统中更新修改表数据的时候，同时修改时间戳字段的值。当进行数据抽取时，通过在where条件中放置类似于gmt_modified>sysdate - 1来决定增量抽取哪些数据。

最高速率上限

作业速率上限是指数据同步作业可能达到的最高速率，其最终实际速率受网络环境、数据库配置等影响。

脏数据

数据同步通常会对接源宿两端数据存储，需要根据源宿两端数据源的具体信息适配和转换相应的数据内容。在传输过程中，可能存在由于两端元数据不匹配或者本身的业务数据传输转换失败（例如OSS上一个定义为Integer的类型存放了“abc”字符串），数据集成将自动识别上述异常情况，并提供自动记录和容错机制,最大限度保证数据传输的可靠性和健壮性。数据集成基于自动识别脏数据功能上，还提供数据传输容错上限。例如，由于历史遗留问题， 若您知晓脏数据影响情况并且对于源端脏数据有一定容忍度，则可以配置单个Job最大脏数据条数阈值。

计量计费

计量计费

阿里云数据集成如何计量和收费？

数据集成的基本计量单位为DMU(Data Migration Unit，即数据移动单位)，代表单个单位在数据集成中的能力（包含 CPU、内存、网络资源分配）。

一个DMU描述了一个数据集成作业最小运行能力，即在限定的CPU、内存、网络资源情况下对于数据同步的处理的能力。一个数据集成作业可以指定在1个或者多个DMU上运行。

- 如果您的任务运行在系统资源组上，则同步任务的计费公式为：

```
一次同步任务消耗费用=任务配置的DMU数量*DMU单价*任务运行时长
```

价格如下：

计费项	价格
DMU	0.35元/小时

- 如果您的任务运行在自定义资源组上，则同步任务的计费公式为：

```
一次同步任务消耗费用=小时单价*任务运行时长
```

价格如下：

计费项	价格
运行时长	0.14元/小时

注：以上任务运行时长精确至分钟级别，且向上取整

数据集成商业化后，所有区域的数据集成产品，将提供为期3个月的0折（0元）的优惠服务，2018年7月2日将正式收取您的任务消耗费用。优惠期间，您可以到阿里云控制台费用中心查看您的消费明细和使用记录，以此帮助您预估正式收费时的消耗账单。
请保持余额充足以免影响业务使用；如不需继续使用，请至数据集成控制台删除所有已配置同步任务，以免产生账单费用。

阿里云数据集成是否带来其他费用？

阿里云数据集成独立于其从中读取数据的源端和向其中写入数据的目标端。因此您需要分别为与输入和输出数据源相关的上下游付费，例如您向OSS写入数据，需要提供相应的存储费用。请查看所涉及到对应存储产品的收费细则，在此不再赘述。另外，可能存在因数据传输产生的公网流量费，这部分费用也不包含在数据集成收费中。