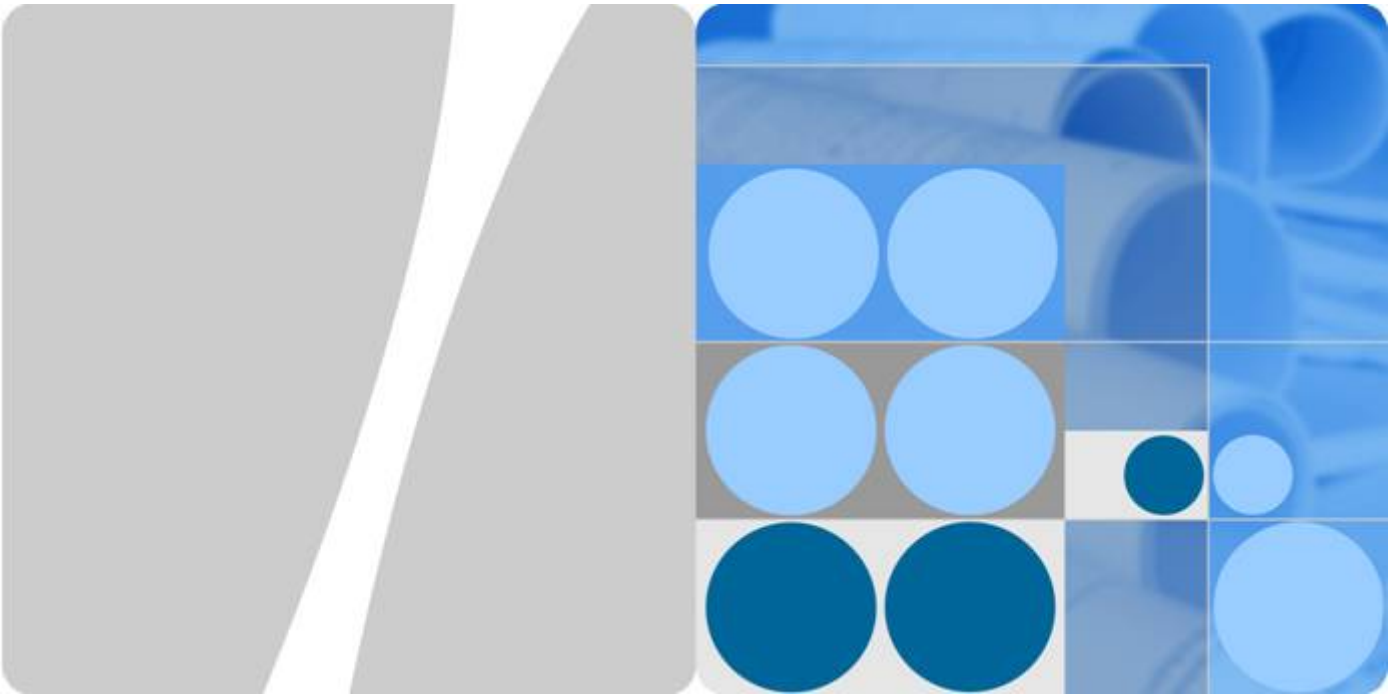


数据接入服务

常见问题

文档版本　01

发布日期　2024-10-25

华为技术有限公司

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 一般性问题

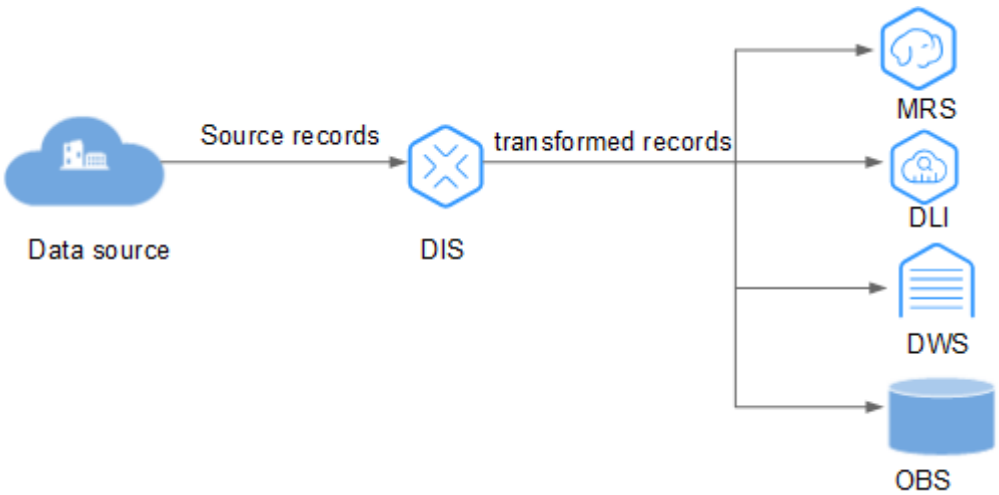
1.1 什么是 DIS?

数据接入服务（Data Ingestion Service）为处理或分析流数据的自定义应用程序构建数据流管道，主要解决云服务外的数据实时传输到云服务内的问题。数据接入服务每小时可从数十万种数据源（如IoT数据采集、日志和定位追踪事件、网站点击流、社交媒体源等）中连续捕获、传送和存储数TB数据。

数据流向

- DIS实时从多种数据源采集数据。
- DIS连续传输数据，自动将数据传输至MRS，DLI，DWS和OBS等服务做计算，分析和存储。

图 1-1 数据流向



1.2 什么是分区?

分区（Partition）是DIS数据通道的基本吞吐量单位。创建通道时，将指定所需的分区数量。

- 普通通道单分区容量：最高发送速度可达1MB/秒或1000条记录/秒（达到任意一种速度上限才会被限流），最高提取速度可达 2MB/秒，单次请求的记录总大小不能超过1MB（不包含partitionKey数据大小）。
- 高级通道单分区容量：最高发送速度可达 5MB/秒或2000条记录/秒（达到任意一种速度上限才会被限流），最高提取速度可达 10MB/秒，单次请求的记录总大小不能超过5MB（不包含partitionKey数据大小）

目前每个租户默认Partition配额范围为1～50个，租户可以根据需要配置Partition个数。

若需扩大配额，请**提交工单**增加配额，具体上限需要根据集群的实际负载情况进行计算。

1.3 DIS 主要应用于哪些场景?

DIS对于从数据生产者快速移出数据，然后进行持续处理非常有用。以下是使用DIS的典型场景：

- 加速日志和数据传送获取：您无需等待批量处理数据，而是让数据生产者在生成数据后立即输入DIS数据通道，防止因数据生产者出现故障导致的数据损失。例如，系统和应用程序日志可以持续添加到数据通道并可在数秒内进行处理。
- 实时指标和报告：实时从DIS数据通道数据提取指标并生成报告。例如，数据接入服务应用程序可以处理系统和应用程序日志的指标和报告，因为数据被流入而不是等待收到批量数据。
- 实时数据分析：通过数据接入服务，可以运行实时通道数据分析。例如，可以通过API把数据实时添加到DIS数据通道中，并让您的DIS应用程序实时运行分析，从而在数分钟内从数据中获得重要见解，而无需数小时或数天时间。
- 复杂的数据通道处理：您可以创建DIS应用程序和数据通道的Directed Acyclic Graphs（DAG）。在这一情景中，一个或多个DIS应用程序可将数据添加到一个DIS数据通道进行进一步处理，以便于进行通道处理器的后续阶段。

1.4 DIS 有哪些特点和优势?

- 无限扩展：DIS数据通道的吞吐量每小时可从数MB扩展到数TB，PUT记录每秒钟可从数千次扩展到数百万。
- 易于使用：您可以在几秒钟内创建DIS数据通道，轻松地将数据放入通道中，并构建用于数据处理的应用程序。
- 成本低廉：DIS没有前期成本，您只需要为实际使用的资源付费即可。
- 并行处理：DIS可让您用多个应用程序同时处理同一个数据通道。例如，您可以让一个应用程序运行实时分析，让其他应用程序从同一个DIS数据通道中将数据发送至对象存储服务（Object Storage Service，简称OBS）。
- 安全可靠：DIS可将数据保留24小时，N的取值为1~7的整数，以防数据在应用程序故障、个别机器故障或设施故障时丢失。

1.5 DIS 有哪些模块及各模块功能?

- 服务控制面
  - 完成服务的开通、删除、配置操作，并将用户信息同步到数据面。

  - 完成数据面资源的申请与自动部署。
- 服务数据面
  - 接收用户发送数据的请求，对已鉴权的数据接收并存储。
  - 接收用户获取数据的请求，在鉴权后输出对应的用户数据。
  - 按时老化存储在系统中的用户数据。
  - 根据用户配置，将用户数据存储到对象存储服务（Object Storage Service，简称OBS）。
- 服务维护
  - 负责服务的安装、升级。
  - 负责服务的配置、巡检、日志收集与分析、运行监控。
  - 负责服务工单处理。
- 用户SDK
  - 提供Java接口，供用户上传与下载数据。
  - 提供数据加密功能。

1.6 如何开通 DIS 通道?

步骤1 使用注册账户登录**DIS控制台**。

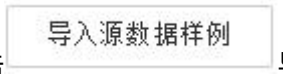
步骤2 单击管理控制台左上角的 ，选择区域和项目。

步骤3 单击“购买接入通道”配置相关参数。

表 1-1 接入通道参数说明

参数	参数解释	参数示例
计费模式	按需计费	按需计费
区域	指的是云服务所在的物理位置。您可以在下拉框中选择并切换区域。	-
基本信息		
通道名称	用户发送或者接收数据时，需要指定通道名称，通道名称不可重复。通道名称由英文字母、数字、中划线和下划线组成。长度为1～64个字符。	dis-Tido
通道类型	• 普通通道单分区容量：最高发送速度可达1MB/秒或1000条记录/秒（达到任意一种速度上限才会被限流），最高提取速度可达 2MB/秒，单次请求的记录总大小不能超过1MB（不包含partitionKey数据大小）。 • 高级通道单分区容量：最高发送速度可达5MB/秒或2000条记录/秒（达到任意一种速度上限才会被限流），最高提取速度可达 10MB/秒，单次请求的记录总大小不能超过5MB（不包含partitionKey数据大小）	-

参数	参数解释	参数示例
分区数量	分区是DIS数据通道的基本吞吐量单位。	5
分区计算	用户可以根据实际需求通过系统计算得到一个建议的分区数量值。 1. 单击“分区计算”，弹出“计算所需分区数量”对话框。 2. 根据实际需求填写“平均记录大小”、“最大写入记录数”和“消费程序数量”，“预估所需分区数量”选项框中将显示所需的分区数量，此值不可修改。 **说明** 所需分区计算公式： – 按流量计算所需写分区数：（所得数值需向上取整后作为分区数） 普通通道：平均记录大小*（1+分区预留比例20%）*最大写入记录数/（1*1024KB） 高级通道：平均记录大小*（1+分区预留比例20%）*最大写入记录数/（5*1024KB） – 按消费程序数量计算读分区数：（消费程序数量/2后的数值需要保留两位小数，然后乘以“按流量计算所需写分区数”，最终取值需向上取整） （消费程序数量/2）*按流量计算所需的写分区数 获取“按流量计算所需写分区数”、“按消费程序数量计算读分区数”中的最大值作为预估所需分区数量。 3. 单击“使用计算值”将系统计算出的建议值应用于“分区数量”。	-
生命周期（小时）	存储在DIS中的数据保留的最长时间，超过此时长数据将被清除。 取值范围：24~72的整数。	24
源数据类型	● BLOB：存储在数据库管理系统中的一组二进制数据。“源数据类型”选择“BLOB”，则支持的“转储服务类型”为“OBS”。 ● JSON：一种开放的文件格式，以易读的文字为基础，用来传输由属性值或者序列性的值组成的数据对象。“源数据类型”选择“JSON”，则支持的“转储服务类型”为“OBS”、“MRS”、“DLI”和“DWS”。 ● CSV：纯文本形式存储的表格数据，分隔符默认采用逗号。 “源数据类型”选择“CSV”，则支持的“转储服务类型”为“OBS”、“DLI”、“DWS”。	JSON

参数	参数解释	参数示例
自动扩缩容	创建通道的同时是否开启自动扩缩容功能。 通过单击通过单击  或  来关闭或开启自动扩缩容开关。	**说明** 用户可在创建通道时定义是否自动扩缩容，也可对已创建的通道修改自动扩缩容属性。
自动缩容最小分区数	设置自动缩容的分区下限，自动缩容的目标分区数不小于下限值。	-
自动扩容最大分区数	设置自动扩容的分区上限，自动扩容的目标分区数不超过上限值。	-
源数据分隔符	源数据为CSV格式时的数据分隔符。	-
Schema开关	创建通道的同时是否为其创建数据Schema。源数据类型为JSON或CSV时可配置该参数。 通过单击  或  来关闭或开启Schema配置开关。 **说明** 若创建通道时，没有同时创建数据Schema,可待通道创建成功后。到通道的管理页面创建数据Schema，详情请参见**管理源数据Schema**。	“源数据类型”为“JSON”和“CSV”时，可选择创建数据Schema。
源数据Schema	支持输入和导入源数据样例，源数据样例格式为JSON或者CSV，详细操作请参见**管理源数据Schema**。 1. 在左侧文本框中输入JSON或者CSV格式的源数据样例，也可单击  导入源数据样例。 2. 在左侧文本框中单击  ，可删除左侧文本框中已输入或导入的源数据样例。 3. 在左侧文本框中单击  ，可在右侧文本框中根据源数据样例生成Avro schema。 4. 在右侧文本框中单击  ，可删除已生成的Avro schema。 5. 在右侧文本框中单击  ，可修改已生成的Avro schema。	仅当“Schema配置开关”配置为“开启”：时需要配置此参数。

参数	参数解释	参数示例
企业项目	配置通道所属的企业项目。已开通企业项目管理服务的用户才可以配置该参数。默认值为default。 企业项目是一种云资源管理方式，企业项目管理服务提供统一的云资源按项目管理，以及项目内的资源管理、成员管理。 您可以选择默认的企业项目“default”或其他已有的企业项目。如果要创建新的企业项目，请登录企业管理控制台进行创建，详细操作请参考《企业管理用户指南》。	-
现在配置	单击“现在配置”，呈现添加标签。 添加标签具体请参考**管理通道标签**。	-
暂不配置	暂不配置任何信息。	-
标签	标签是通道的标识。为通道添加标签，可以方便用户识别和管理拥有的通道资源。	-

步骤4 单击“立即购买”，弹出“规格确认”页面。

步骤5 单击“提交”，完成通道接入。

----结束

1.7 数据存储在 DIS 和转储其他资源有什么区别?

开通DIS通道时需要选择“转储服务类型”。具体区别如**表1-2**所示。

- 选择“OBS”表示存储在DIS中，并周期性导入对象存储服务（Object Storage Service，简称OBS）。
- 选择“MRS”表示存储在DIS中，并周期性导入MapReduce服务（MRS）集群的HDFS中。
- 选择“DLI”表示存储在DIS中，并周期性导入DLI。
- 选择“DWS”表示存储在DIS中，并周期性导入数据仓库服务（DWS）中。
- 选择“CloudTable”表示存储在DIS中，并实时导入CloudTable集群的HBase表或OpenTSDB表中。

表 1-2 DIS 和转储其他资源区别

DIS存储	OBS存储	MRS存储	DLI存储	DWS存储	CloudTable存储
DIS服务自带。	需要另外申请。	需要另外申请。	需要另外申请。	需要另外申请。	需要另外申请。

DIS存储	OBS存储	MRS存储	DLI存储	DWS存储	CloudTable存储
无需另外付费。	需要根据OBS收费标准另外付费。	需要根据MRS和OBS收费标准另外付费。	需要根据DLI和OBS收费标准另外付费。	需要根据DWS和OBS收费标准另外付费。	需要根据CloudTable收费标准另外付费。
临时存储（最长保留168小时）。	数据可长期存储在OBS中，具体保存时长根据用户购买的OBS服务时长决定。	数据可长期存储在MRS中，具体保存时长根据用户购买的MRS服务时长决定。	数据可长期存储在DLI中，具体保存时长根据用户购买的DLI服务时长决定。	数据可长期存储在DWS中，具体保存时长根据用户购买的DWS服务时长决定。	数据可长期存储在CloudTable中，具体保存时长根据用户购买的CloudTable服务时长决定。
只存储在DIS中。	存储在DIS中，并周期性导入OBS。	存储在DIS中，并周期性导入MRS集群的HDFS中。 **说明** 导入MRS集群前临时存储在OBS，待转储MRS完成后删除OBS上的临时存储文件。	存储在DIS中，并周期性导入DLI。 **说明** 导入DLI前临时存储在OBS，待转储DLI完成后删除OBS上的临时存储文件。	存储在DIS中，并周期性导入DWS。 **说明** 导入DWS前临时存储在OBS，待转储DWS完成后删除OBS上的临时存储文件。	存储在DIS中，实时导入CloudTable集群的HBase表或OpenTSDB表中。

1.8 如何校验软件包完整性?

获取DIS SDK软件包及校验文件后，可以在Linux系统上按如下步骤对软件包的完整性进行校验。

前提条件

- 已获取“PuTTY”工具。
- 已获取“WinSCP”工具。

操作步骤

步骤1 使用“WinSCP”工具将“huaweicloud-sdk-dis-x.x.x.zip”上传至Linux系统任一目录。

说明

x.x.x表示DIS SDK包的版本号。

步骤2 使用“PuTTY”工具登录Linux系统，进入到“huaweicloud-sdk-dis-x.x.x.zip”所在目录，执行如下命令，获取DIS SDK压缩包的校验码。

sha256sum huaweicloud-sdk-dis-x.x.x.zip

显示类似如下校验码：

```
# sha256sum dis-sdk-x.x.x.zip
8be2c937e8d78b1a9b99777cee4e7131f8bf231de3f839cf214e7c5b5ba3c088 huaweicloud-sdk-dis-x.x.x.zip
```

步骤3 打开DIS SDK的校验文件“huaweicloud-sdk-dis-x.x.x.zip.sha256sum”与上一步骤中获取的校验码进行对比。

- 一致，说明从获取的DIS SDK压缩包没被篡改。
- 不一致，说明DIS SDK压缩包被篡改，需要重新获取。

----结束

1.9 DIS 如何发送和接收数据?

步骤1 开通DIS通道，在IAM（用户认证中心）中获取账号的AK/SK。

步骤2 在**这里**中下载“dis-sdk-*X.X.X*.zip”压缩包并解压缩。

步骤3 建立工程，配置用户AK/SK、endpoint、projectId、region、通道名称、分区数量等。

步骤4 配置完成后运行程序即可发送数据。

步骤5 建立工程，配置用户AK/SK、endpoint、project、region、通道名称、partitionId和startingSequenceNumber。

步骤6 配置完成后运行程序即可接收数据。

----结束

1.10 什么是流控?

流控就是超过通道内分区的最大吞吐量开始限流，对资费和数据没有影响。

2 转储相关问题

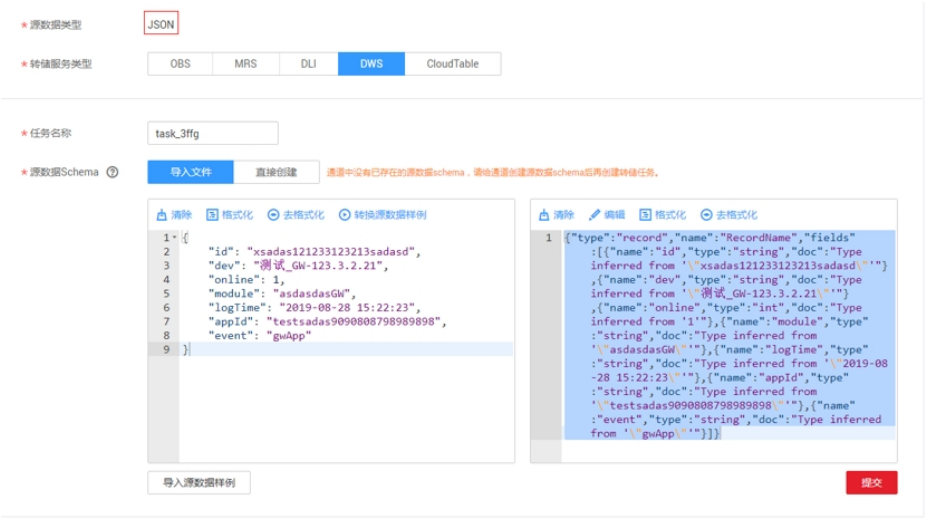
2.1 DIS 如何实现转储数据至 DWS 的特定列

DIS支持将源数据类型为JSON格式的数据转储至DWS。转储前，需要配置源数据Schema。

源数据Schema，即用户的JSON数据样例，用于描述JSON数据格式。DIS可以根据此JSON数据样例生成Avro schema, 将通道内上传的JSON数据转换为Parquet或CarbonData格式。

1. 参考**创建源数据Schema**，创建源数据Schema。如下以添加转储任务时创建源数据Schema为例进行说明。
   a. 选择源数据类型是Json的通道。
   b. 在通道详情页面的“转储任务”页签，单击“添加转储任务”。
   c. 转储服务类型选择DWS，通过导入文件的方式配置源数据Schema。
   d. 输入源数据样例，单击“转换源数据样例”并提交，生成源数据Schema。

   图 2-1 创建源数据 Schema



2. 配置Schema属性过滤功能。

说明

schema过滤功能，只针对源数据schema根节点或一级子节点非array类型，才有效。即**管理源数据Schema**创建的源数据schema，满足根节点或一级子节点非array类型，界面才呈现此配置。

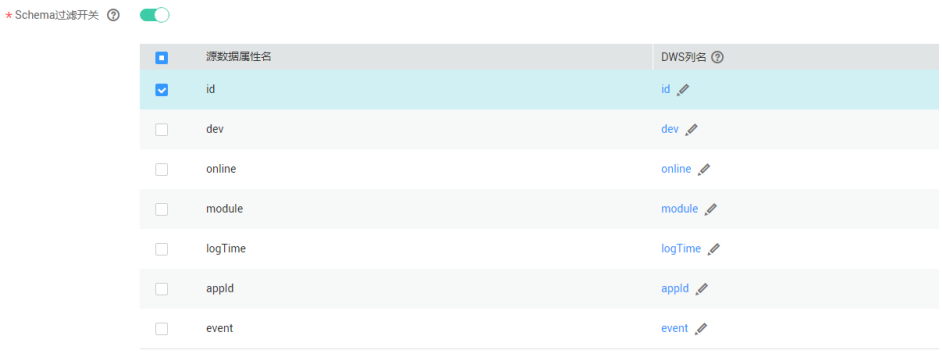
a. 打开Schema过滤开关。

b. 在源数据属性名列表中，勾选对应的属性名，完成DWS表中指定列的映射。

说明

源数据属性名列表中的属性由源数据Schema的name字段生成，匹配DWS的列名称。

图 2-2 配置 Schema 属性



c. 如**图2-2**所示，源数据属性名只选择id，即少于对应表的总字段。

DWS侧创建集群，并执行如下命令创建表。

CREATE TABLE dis_test3(id TEXT,dev TEXT,online BIGINT,module TEXT default 'a',logTime TEXT,appId TEXT,event TEXT);

d. DIS侧转储数据至DWS成功后，登录集群数据库查询dis_test3表格数据，可看到仅id列和module列插入数据，其中module列是默认数据。如**图2-3**所示。

图 2-3 Schema 属性过滤结果

```
postgres=> select * from dis_test3;
            id            | dev | online | module | logtime | appid | event
--------------------------+-----+--------+--------+---------+-------+-------
 xsadas1212331233213sadasd |     |        | a      |         |       |
 xsadas1212331233213sadasd |     |        | a      |         |       |
(2 rows)
```

2.2 Schema 如何支持字段缺省或者为 NULL

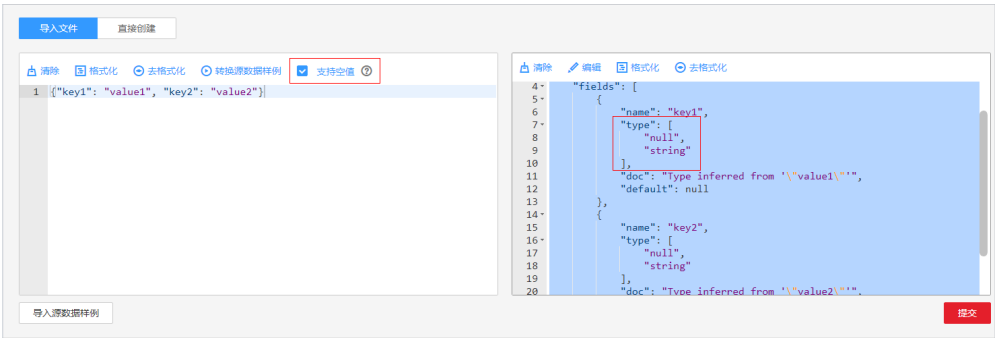
源数据Schema，即用户的JSON数据样例，用于描述JSON数据格式。DIS可以根据此JSON数据样例生成Avro schema，默认情况下不支持字段缺省或者为NULL，如**图2-4**。

图 2-4 不支持字段缺省样例

"key1"字段对应的类型为**"String"**（Avro Schema中体现为："type": "string"），这时如果源数据中"key1"不传或者传的值为NULL，那么转储任务会报错。

如果需要根据JSON数据样例生成的Schema可以支持缺省或者NULL，则需要勾选**"支持空值"**复选框，再单击"转换源数据样例"，如图2-5所示。

图 2-5 支持字段缺省样例



这时，**"key1"**字段对应的类型为**"Union"**复合类型（Avro Schema中体现为："type": ["null", "string"]），如果源数据中"key1"不传或者传的值为NULL，那么会自动填补NULL为默认值，转储任务可以正常进行格式转换。

2.3 如何专线接入 DIS

步骤1 参见**自助开通云专线**开通云专线，然后使用开通专线的云账号访问虚拟私有云VPC。

步骤2 选择“VPC终端节点 > 终端节点”，单击“购买终端节点”。

步骤3 “服务类别”选择“云服务”，并在服务中选择DIS服务的终端节点。然后选择专线所在的VPC和子网即可。

步骤4 创建终端节点成功之后，会自动分配节点IP，使用此节点IP访问DIS服务即可。

----结束

2.4 读取通道数据时，如何区分不同类型数据?

- 不同类型的消息使用不同的通道；

- 使用同一个通道的不同分区。上传消息时，不同类型的消息指定不同的partition_key，消费时根据partition_key来区分不同类型消息。

3 DIS Agent 相关问题

3.1 Agent 如何配置监听多目录或文件?

DIS Agent支持配置监听多个目录或文件，例如想收集"**/home/folder1/file1**"和"**/home/folder2/file2**"这两个文件的日志，可以通过配置多个DISStream来实现：

```
---
region: REGION
ak: YOUR_AK
sk: YOUR_SK
projectId: YOUR_PROJECTID
endpoint: ENDPOINT
flows:
  - DISStream: YOUR_STREAM
    filePattern: /home/folder1/file1
    initialPosition: START_OF_FILE
    maxBufferAgeMillis: 5000

  - DISStream: YOUR_STREAM
    filePattern: /home/folder2/file2
    initialPosition: START_OF_FILE
    maxBufferAgeMillis: 5000
```

3.2 Agent 如何配置递归监听一个目录?

DIS Agent支持配置递归监听，将配置项"**directoryRecursionEnabled**"的值配置为"**true**"即可支持，例如以下配置可以匹配到"**/home/one.log**"，"**/home/child/two.log**"，"**/home/child/child/three.log**"：

```
---
region: REGION
ak: YOUR_AK
sk: YOUR_SK
projectId: YOUR_PROJECTID
endpoint: ENDPOINT
flows:
  - DISStream: YOUR_STREAM
    filePattern: /home/*.log
    directoryRecursionEnabled: true
    initialPosition: START_OF_FILE
    maxBufferAgeMillis: 5000
```


3.3 Agent 如何配置代理?

DIS Agent支持通过配置代理上传数据到DIS，需要配置**"PROXY_HOST"**，**"PROXY_PORT"**，**"PROXY_USERNAME"**，**"PROXY_PASSWORD"**，这几个配置项介绍可以查看**Agent配置文件**说明。

```
---
region: REGION
ak: YOUR_AK
sk: YOUR_SK
projectId: YOUR_PROJECTID
endpoint: ENDPOINT
PROXY_HOST: YOUR_PROXY_HOST
PROXY_PORT: YOUR_PROXY_PORT
PROXY_USERNAME: YOUR_PROXY_USERNAME
PROXY_PASSWORD: YOUR_PROXY_PASSWORD
flows:
  - DISStream: YOUR_STREAM
    filePattern: /home/*.log
    initialPosition: START_OF_FILE
    maxBufferAgeMillis: 5000
```

3.4 Agent 如何配置 AK/SK 加密?

在配置项中，需要配置用户的SK，这属于敏感信息，如需加密，可以按如下步骤：

步骤1 进入bin/目录

cd /opt/dis-agent-X.X.X/bin

步骤2 执行加密脚本，输入密码后回车

bash dis-encrypt.sh

步骤3 控制台打印的“Encrypt result:”后面的字符串即为加密后的结果。通过这种方式分别加密MySQL密码和用户SK，并将密文配置到配置文件中即可。

----结束

