

DataWorks V1.0

常见问题

常见问题

准备工作常见问题

Q：大数据开发平台上的项目怎么绑原来已有的 MaxCompute project？

A：主账号在新建 DataWorks（数据工场，原大数据开发套件）项目的时候，选择已有 MaxCompute Project 对应的计费方式，项目名称填写已有 MaxCompute Project Name，并单击 **确定**，即可添加已有 MaxCompute Project 到 DataWorks（数据工场）中。

Q：公司已经创建组织，怎么把我加为成员？

A：联系组织管理员索要子账号，联系对应项目管理员将该子账号添加至对应项目中。添加成功子账号登录即可，~~相关子账号登录问题，请参见 准备 RAM 子账号。~~

Q：如何授权？

A：将子账号添加至项目中，授予不同角色身份，以获取相应的操作权限。

注意：

账号进入租户后默认没有任何权限（创建者项目除外），账号被添加至某一个项目中后，会根据相应的授权角色获取相应的权限。

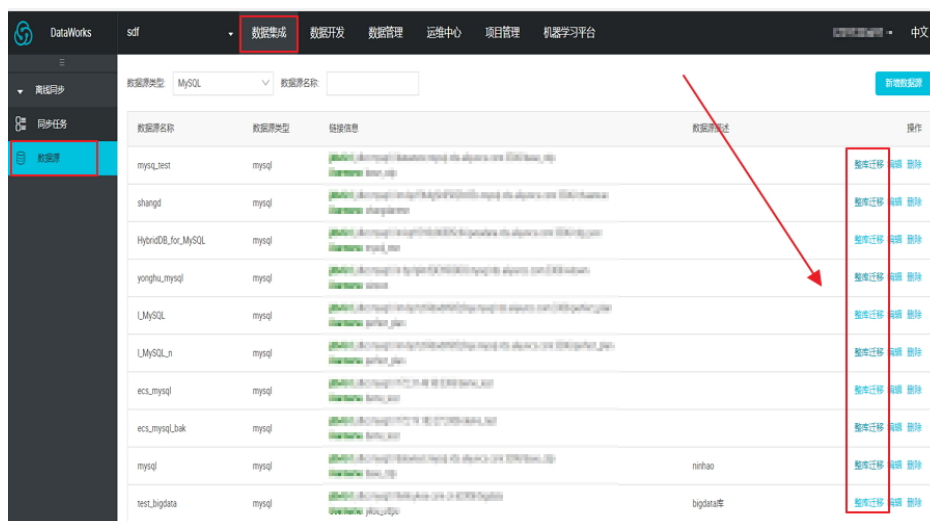
Q：修改了主账号的登录名，用修改后的登录名将子账号拉进 A 项目中，子账号使用 A 项目时，报错：非法的子账号，不是组织主账号的子账号。

A：主账号进入 用户管理 页面，单击 **刷新用户** 即可。

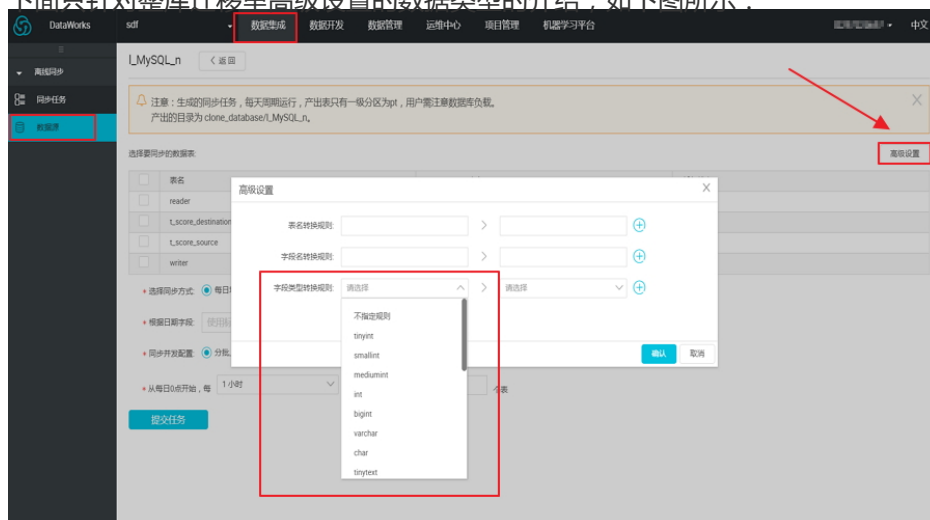
数据集成常见问题

整库迁移数据类型

整库迁移目前只提供MySQL（包括rds里的MySQL）同步到MaxCompute，可以从已经添加好的MySQL数据源里进入整库迁移界面，如下图所示：



下面只针对整库迁移里高级设置的数据类型的介绍，如下图所示：



整库迁移源端MySQL支持的数据类型：tinyint、smallint、mediumint、int、bigint、varchar、char、tinytext、text、mediumtext、longtext、year、float、double、decimal、date、datetime、timestamp、time、bool
目标端MaxCompute支持的数据源类型：bigint、string、bouble、datetime、boolean
上面MySQL支持的数据类型支持都与MaxCompute数据源类型之间的转换。

注意：mysql里的bit，如果是bit（2）以上，则目前不支持与bigint、string、double、datetime、boolean这些类型转换；如果是bit(1)，则会被转换成boolean

同步任务等待槽位

用户问题

任务未正常运行，日志提示：

目前实例还没有产生日志信息,在等待槽位

问题原因

出现这个提示的原因是因为任务是配置调度使用自定义资源。但是目前没有存在空闲可用的自定义资源

1. 用户可以在大数据开发套件->运维中心->任务运维里，找到没有如期调度的任务，右键点击“查看节点属性”，看到任务使用的资源组是什么。
2. 在项目管理->调度资源管理里找到这个调度资源，点击服务器管理，看看服务器的状态是否是停止状态，是否已使用上看到被其他任务占用了。
3. 如果以上排查无法解决问题，可以考虑重启一下服务，命令为

```
su - admin  
/home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart
```

如问题还未解决，请联系售后技术支持。

ECS上自建的数据库测试连通性失败

问题：ECS上自建的数据库测试连通性失败

解决方法：

- 1.检查是否添加了安全组；

请在ecs安全组上对,允许内网入:数据库端口

10.46.67.156

10.46.64.81

10.117.39.238

10.117.28.203

10.27.63.41

10.27.63.60

10.27.63.38

10.27.63.15

如果用户使用ecs的外网地址，需要：

允许公网入:数据库端口

120.27.160.26

120.27.160.81

121.43.110.160

121.43.112.137

118.178.84.74

118.178.56.228

118.178.59.233

118.178.142.154

如果是自定义资源组，需要自定义调度服务器和ecs数据库同区域网络互通，并把调度服务器ip加入数据库ecs安全组开放端口。

2.检查是否因为地域不同。

我们所属地域位于华东 1（杭州），其他的地域内网可能不通，如果不通请尝试使用外网，请参照下面的规则：

不同地域的云服务器 ECS、关系型数据库 RDS、开放存储服务 OSS 内网不互通。

不同地域之间的云服务器和其他云资源，如 RDS、OSS 等内网不互通。

RDS 数据源测试连通性不通

问题：新增数据源时，RDS 数据源测试连通性不通。

解决方法：添加 RDS 的白名单。

当 RDS 数据源测试连通性不通时，需要到自己的 RDS 上添加数据同步机器 IP 白名单：

11.192.97.82,11.192.98.76,10.152.69.0/24,10.153.136.0/24,10.143.32.0/24,120.27.160.26,10.46.67.156,1

20.27.160.81,10.46.64.81,121.43.110.160,10.117.39.238,121.43.112.137,10.117.28.203,118.178.84.74,10.27.63.41,118.178.56.228,10.27.63.60,118.178.59.233,10.27.63.38,118.178.142.154,10.27.63.15,100.64.0.0/8

注意：若使用自定义资源组调度 RDS 的数据同步任务，必须把自定义资源组的机器 IP 也加到 RDS 的白名单中。

- 1、网络：确保网络和端口之间是能够联通，网络连接失败，可以再检查一下 ECS 的防火墙以及安全组设置；
- 2、确保添加的信息正确：用户名\密码\jdbcUrl 中的 IP 地址和端口必须确保正确；
- 3、确保自建的数据库涉及的安全访问限制，权限的限制和能否远程登录的情况；
- 4、在 VPC 的环境下购买的 ECS 只能用脚本模式运行任务，在添加数据源时测试连通性不能成功。购买 ECS 可以添加自定义资源，将同步任务下发到相应的资源组运行。

RDS同步失败转换成jdbc格式

Q：从 RDS（MySQL/SQLServer/PostgreSQL）同步到自建 MySQL/SQLServer/PostgreSQL 时，报错为：DataX 无法连接对应的数据库。

A：以 RDS（MySQL）的数据同步到自建 SQLServer 为例，操作如下：

1. 新建一个数据源，将数据源配置为 MySQL->JDBC 格式；

配置同步任务使用新数据源，重新执行即可。

注意：若是在 RDS(MySQL)->RDS(SQLServer) 等云产品之间同步时，建议选择 RDS(MySQL)->RDS(SQLServer) 数据源来配置同步任务。

数据同步内存溢出

用户问题

数据同步报错，日志提示：Code:[Framework-13], Description:[DataX插件运行时出错, 具体原因请参看 DataX运行结束时的错误诊断信息 .]. - java.lang.OutOfMemoryError: Java heap space at

```
java.util.Arrays.copyOf(Arrays.java:3236)
```

问题原因

出现这个提示的原因是因为任务数据量稍大并且同步的速度太快，一般是接近10M每秒，导致用户的调度服务器内存溢出报错。

1. 用户可以在大数据开发套件->数据集成-查看数据同步任务的速度情况，一般出现内存溢出都是速度设置过大导致的，这里建议尝试把同步速度调小。
2. 用户可以在大数据开发套件->数据开发-看下数据同步任务的同步速度是不是太大，检查该调度资源有其他大的任务或者比较多的任务在执行
3. 先尝试把同步速度调小，看下其他任务执行状态，资源比较多时候执行同步。
4. 可以将数据任务分时间段来执行，不要在同一时间执行，另外同步速度不宜很大。如果有需求，可以考虑增加调度服务器的配置或者实例。

如问题还未解决，请联系售后技术支持。

如何排查数据集成问题？

当通过数据集成实现某操作出现问题时，首先要定位问题的相关信息，比如：运行在什么机器上，查看数据源的相关信息，查看在哪个 Region 配置的同步任务等。

运行机器

运行在阿里的机器上：

```
running in Pipeline[basecommon_group_xxxxxxxxx]
```

运行在自己的机器上：

```
running in Pipeline[basecommon_xxxxxxxxx]
```

数据源的相关信息

当出现数据集成问题时，您需要查看数据源的相关信息：

确认是什么数据源之间的同步。

确认是什么环境的数据源。

阿里云数据库，有公网 IP，无公网 IP，VPC 网络环境的数据源（RDS、其他的数据源），金融云环境（VPC 环境，经典网络）。

确认数据源测试连通性是否成功。

对照 [配置数据源 文档](#)：检查其数据源相关信息是否填错（一般填错情况，多个数据源库搞错了，在填写的信息里加了空格或特殊字符，不支持测试连通性的问题（无公网 IP，除 RDS 的 VPC 环境的数据源））。

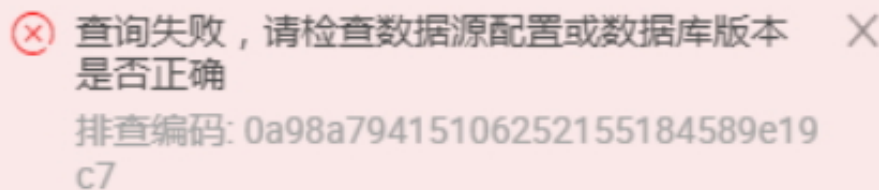
确认配置任务的 Region

进入 DataWorks 的管理控制台，可以看到相关的 Region，如：华东 2、华南 1、香港、亚太东南亚 1、欧洲中部 1、亚太东南 2 等，一般默认是华东 2。购买 MaxCompute 才能查看到相应的 region 的，如下图所示：



界面模式报错，复制排查码

如果报错，请将排查码 **复制** 出来提供给处理人员，如下图所示：



日志报异常

日志中报 SQL 语句执行失败（列包含关键字）

```
2017-05-31 14:15:20.282 [33881049-0-0-reader] ERROR ReaderRunner - Reader runner Received  
Exceptions:com.alibaba.datax.common.exception.DataXException: Code:[DBUtilErrorCode-07]
```

错误解读：

读取数据库数据失败，请检查您配置的 column/table/where/querySql 或者向 DBA 寻求帮助。

执行的 SQL 如下所示：

```
select **index**,plaid,plarm,fget,fot,havm,coer,ines,oumes from xxx
```

错误信息如下所示：

```
You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right  
syntax to use near **index**,plaid,plarm,fget,fot,havm,coer,ines,oumes from xxx
```

排查思路：

本地运行 SQL 语句select **index**,plaid,plarm,fget,fot,havm,coer,ines,oumes from xxx查看其结果，一般也会有相应的报错。

字段中有关键字 index，解决方法加单引号或修改字段。

日志中报 SQL 语句执行失败（表名带有双引号包单引号）

```
com.alibaba.datax.common.exception.DataXException: Code:[DBUtilErrorCode-07]
```

错误解读：

读取数据库数据失败，请检查您配置的 column/table/where/querySql 或者向 DBA 寻求帮助。

执行的 SQL 如下所示：

```
select /_+read_consistency(weak) query_timeout(100000000)/_ from** 'ql_ddddd_[0-31]' **where 1=2
```

错误信息如下所示：

```
You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near ' 'ql_live_speaks[0-31]' where 1=2' at line 1 -
com.mysql.jdbc.exceptions.jdbc4.MySQLSyntaxErrorException: You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near '** 'ql_ddddd_[0-31]' where 1=2' **
```

排查思路：

配置表名时，双引号包单引号，一般配置常量是双引号包单引号 “table” :[“ 'ql_ddddd[0-31]' ”]直接去掉里面的单引号。

测试数据源连通性失败（Access denied for）

连接数据库失败，数据库连接串：jdbc:mysql://xx.xx.xx.x:3306/t_demo，用户名：fn_test，异常消息：**Access denied for user 'fn_test' '@' %' to database 't_demo'** 请确定 RDS 中已经添加白名单。

排查思路：

Access denied for 一般都是填写的信息有问题，请确认您填写的信息。

白名单或者用户的账号有没有对应数据库的权限，RDS 管控台可以添加相应的白名单和授权。

路由策略有问题，运行的池子 oxs 和 ECS 集群

```
2017-08-08 15:58:55 : Start Job[xxxxxxx], traceId **running in
Pipeline[basecommon_group_xxx_cdp_oxs]**ErrorMessage:Code:[DBUtilErrorCode-10]
```

错误解读：

连接数据库失败，请检查您的账号、密码、数据库名称、IP、Port 或者向 DBA 寻求帮助（注意网络环境）。数据库连接失败，因为根据您配置的连接信息，无法从jdbc:oracle:thin:@xxx.xxxxx.x.xx:xxxx:prod 中找到可连接的 JDBCUrl。请检查您的配置并作出修改。

java.lang.Exception：DataX 无法连接对应的数据库，可能的原因如下：

配置的 ip/port/database/jdbc 错误，无法连接。

配置的 username/password 错误，鉴权失败。请和 DBA 确认该数据库的连接信息是否正确。

排查思路：**情况一：**

Oracle 同步的 RDS-PostgreSQL 直接单击 **运行**，不能在调度里运行，因为运行的池子不同。

可以在添加 RDS 的数据源时，改成添加普通 JDBC 形式的数据源，这样 Oracle 同步的 RDS-PostgreSQL 是可以的。

情况二：

- VPC 环境的 RDS-PostgreSQL 是不能运行在自定义资源组上，因为 VPC 环境的 RDS 有反向代理功能，这样跟用户自定义资源组存在网络问题，所以一般 VPC 环境的 RDS 直接运行在我们这边机器就好，如果觉得我们这边的机器不能满足您的需要，要运行在自己的机器上，你这边要将 VPC 环境的 RDS 作为 VPC 环境 jdbc 形势的数据源，购买一个同网段的 ECS。详情请参见 [VPC 环境数据同步配置](#)。
- 一般 VPC 环境的 RDS 映射出的 jdbc:mysql://100.100.70.1:4309/xxx，100，开头的 IP 是后台映射出来的，如果是一个域名的表现形式那就是非 VPC 环境的。

HBase Writer 不支持 date 类型

HBase 同步到 hbase:2017-08-15 11:19:29 : State: 4(FAIL) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0% ErrorMessage:Code:[Hbasewriter-01]

错误解读：

您填写的参数值不合法。

Hbasewriter 不支持该类型：date，目前支持的类型是：
[string, boolean, short, int, long, float, double]。

排查思路：

HBase 的 writer 不支持 date 类型，所以在 writer 中不能配置 date 类型。

直接配置 string 类型，因为 HBase 没有数据类型的概念，底层一般是 byte 数组。

JSON 格式配置错误

column 配置错误

经 DataX 智能分析，该任务最可能的错误原因如下：

com.alibaba.datax.common.exception.DataXException: Code:[Framework-02]

错误解读：

DataX 引擎运行过程出错，详情请参见 DataX 运行结束时的错误诊断信息。

java.lang.ClassCastException : com.alibaba.fastjson。JSONObject cannot be cast to java.lang.String

排查思路：

发现其 json 配置有问题。

```
writer端：
  "column" :[
  {
    "name" : "busino" ,
    "type" : "string"
  }
]
正确的写法：
  "column" :[
  {
    "busino"
  }
]
```

json List 编写少了 []

经 DataX 智能分析，该任务最可能的错误原因是：

com.alibaba.datax.common.exception.DataXException: Code:[Framework-02]

错误解读：

DataX 引擎运行过程出错，详情请参见 DataX 运行结束时的错误诊断信息。

```
java.lang.String cannot be cast to java.util.List - java.lang.String cannot be cast to java.util.List
at com.alibaba.datax.common.exception.DataXException.asDataXException(DataXException.java:41)
```

排查思路：

少了[]，list 型就变成其他的形式，找到对应的地方填上[]就能解决问题。

权限问题

权限问题（没有 delete 权限）

MaxCompute 同步到 rds-mysql , ErrorMessage:Code:[DBUtilErrorCode-07]

错误解读：

读取数据库数据失败，请检查您配置的 column/table/where/querySql 或者向 DBA 寻求帮助。

执行的 SQL 如下所示：

```
delete from fact_xxx_d where sy_date=20170903
```

具体错误信息如下所示：

```
**DELETE command denied** to user 'xxx_odps' @ '[xx.xxx.xxx.xxx](http://xx.xxx.xxx.xxx)' for table 'fact_xxx_d' - com.mysql.jdbc.exceptions.jdbc4.MySQLSyntaxErrorException: DELETE command denied to user 'xxx_odps' @ '[xx.xxx.xxx.xxx](http://xx.xxx.xxx.xxx)' for table 'fact_xxx_d'
```

排查思路：

DELETE command denied to没有删除此表的权限，到相应的数据库设置相关表的删除权限。

权限问题（没有 drop 权限）

Code:[DBUtilErrorCode-07]

错误解读：

读取数据库数据失败，请检查您配置的 column/table/where/querySql 或者向 DBA 寻求帮助。

执行的 SQL 为：truncate table be_xx_ch

具体错误信息如下所示：

```
**DROP command denied to user** 'xxx' @ '[xxx.xx.xxx.xxx](http://xxx.xx.xxx.xxx)' for table 'be_xx_ch' - com.mysql.jdbc.exceptions.jdbc4.MySQLSyntaxErrorException: DROP command denied to user 'xxx' @ '[xxx.xx.xxx.xxx](http://xxx.xx.xxx.xxx)' for table 'be_xx_ch'
```

排查思路：

mysqlwriter 配置执行前准备语句 truncate 删除表里的数据报上面错误，是因为没有 drop 的权限。

ADS 权限问题

```
2016-11-04 19:49:11.504 [job-12485292] INFO OriginalConfPretreatmentUtil - Available
```

```
jdbcUrl:jdbc:mysql://100.98.249.103:3306/ads_rdb?yearIsDateType=false&zeroDateTimeBehavior=convertToNull&tinyInt1isBit=false&rewriteBatchedStatements=true.  
2016-11-04 19:49:11.505 [job-12485292] WARN OriginalConfPretreatmentUtil
```

您的配置文件中的列配置存在一定的风险. 因为您未配置读取数据库表的列, 当您的表字段个数、类型有变动时, 可能影响任务正确性甚至会运行出错。请检查您的配置并作出修改。

2016-11-04 19:49:11.528 [job-12485292] INFO Writer\$Job

如果是 MaxCompute > ADS 的数据同步, 需要完成以下两方面的授权:

ADS 官方账号至少需要有需要同步的表的 describe 和 select 权限, 因为 ADS 系统需要获取 MaxCompute 需要同步表的结构和数据信息。

您配置的 ADS 数据源访问账号 AK, 需要拥有向指定的 ADS 数据库发起 load data 的权限, 您可以在 ADS 系统中添加授权。

2016-11-04 19:49:11.528 [job-12485292] INFO Writer\$Job

如果是 RDS (或其他非 MaxCompute 数据源) > ADS 的数据同步, 实现逻辑为先将数据装载到 MaxCompute 临时表, 再从 MaxCompute 临时表同步至 ADS, 中转 MaxCompute 项目为 cdp_ads_project, 中转项目账号为 cloud-data-pipeline@aliyun-inner.com。

权限方面:

ADS 官方账号至少需要有需要同步的表 (这里指 MaxCompute 临时表) 的 describe 和 select 权限, 因为 ADS 系统需要获取 MaxCompute 需要同步的表的结构和数据信息, 此部分部署时已经完成授权。

中转 MaxCompute 对应的账号 cloud-data-pipeline@aliyun-inner.com, 要拥有向指定的 ADS 数据库发起 load data 的权限, 您可以在 ADS 系统中添加授权。

排查思路:

出现此问题是因为没有设置 load data 权限。

中转项目账号为 cloud-data-pipeline@aliyun-inner.com, 权限方面: ADS 官方账号至少需要拥有需要同步的表 (这里指 MaxCompute 临时表) 的 describe 和 select 权限, 因为 ADS 系统需要获取 MaxCompute 需要同步的表的结构和数据信息, 此部分部署时已经完成授权, 登录 ADS 管控台给 ADS 授予 load data 的权限。

白名单问题

没有添加白名单测试连通性失败

测试连接失败，测试数据源连通性失败：

```
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=[[{address:3717=dds-bp1afbf47fc7e8e41.mongodb.rds.aliyuncs.com}(http://address:3717=dds-
bp1afbf47fc7e8e41.mongodb.rds.aliyuncs.com), type=UNKNOWN, state=CONNECTING,
exception={com.mongodb.MongoSocketReadException: Prematurely reached end of stream}},
{address:3717=dds-bp1afbf47fc7e8e42.mongodb.rds.aliyuncs.com}(http://address:3717=dds-
bp1afbf47fc7e8e42.mongodb.rds.aliyuncs.com), type=UNKNOWN, state=CONNECTING,**
exception={com.mongodb.MongoSocketReadException: Prematurely reached end of stream**}]}
```

排查思路

非 VPC 环境的 MongoDB，添加数据源时报 **Timed out after 5000**，白名单添加有问题。

注意：

如果您使用的是云数据库 MongoDB 版，MongoDB 默认会有 root 账号。出于安全策略的考虑，数据集成仅支持使用 MongoDB 数据库对应账号进行连接，您添加使用 MongoDB 数据源时，也请避免使用 root 作为访问账号。

白名单不全

for Code:[DBUtilErrorCode-10]

错误解读：

连接数据库失败，请检查您的账号、密码、数据库名称、IP、Port 或者向 DBA 寻求帮助（注意网络环境）。

错误信息如下所示：

```
java.sql.SQLException: Invalid authorization specification, message from server: "###28000ip not in
whitelist, client ip is xx.xx.xx.xx".**
2017-10-17 11:03:00.673 [job-xxxx] ERROR RetryUtil - Exception when calling callable
```

排查思路：

白名单没有添全，没有将用户自己的机器添加白名单内。

数据源信息填写错误

脚本模式配置缺少相应数据源信息（could not be blank）

```

2017-09-06 12:47:05 [INFO] Success to fetch meta data for table with **projectId [43501]** **项目ID **and
instanceId **[mongodb]数据源名.**
2017-09-06 12:47:05 [INFO] Data transport tunnel is CDP.
2017-09-06 12:47:05 [INFO] Begin to fetch alisa account info for 3DES encrypt with parameter account:
[zz_683cdbcefba143b7b709067b362d4385].
2017-09-06 12:47:05 [INFO] Begin to fetch alisa account info for 3DES encrypt with parameter account:
[zz_683cdbcefba143b7b709067b362d4385].
[Error] Exception when running task, message:** Configuration property [accessId]一般是odps数据源要填写
的信息 could not be blank!**

```

排查思路：

报错显示没有相应的 accessId 信息，一般出现这种现象是脚本模式，查看用户配置的 json 代码，是否忘记写相应的 **数据源名**。

没有填写数据源配置

```

2017-10-10 10:30:08 INFO
=====
File "/home/admin/synccenter/src/Validate.py", line 16, in notNone
raise Exception( "Configuration property [%s] could not be blank!" % (context))
**Exception: Configuration property [username] could not be blank!**

```

排查思路：

- 根据正常的打出的日志对比：

```

[56810] and instanceId(instanceName) [spfee_test_mysql]...
2017-10-09 21:09:44 [INFO] Success to fetch meta data for table with projectId [56810] and instanceId
[spfee_test_mysql].

```

- 一般反应出这样的信息说明调数据源失败成功，而且报用户为空，说明没有配置数据源或数据源的位置配置错误，这个用户就将数据源的位置配置错误。

DRDS 连接数据超时

MaxCompute 同步数据到 DRDS，经常出现下面的错误：

```
[2017-09-11 16:17:01.729 [49892464-0-0-writer] WARN CommonRdbmsWriter$Task
```

回滚此次写入，采用每次写入一行方式提交，原因如下：

```

com.mysql.jdbc.exceptions.jdbc4.CommunicationsException: **Communications link failure **
The last packet successfully received from the server was 529 milliseconds ago. The last packet sent
successfully to the server was** 528 milliseconds ago**.

```

```

2017-09-13 16:48:22.089 [50249495-1-7-writer] WARN CommonRdmsWriterTask - 回滚此次写入，采用每次写入一行方式提交。因为:Communications link failure
The last packet sent successfully to the server was 599 milliseconds ago. The last packet sent successfully to the server was 599 milliseconds
ago
2017-09-13 16:48:22.106 [50249495-1-7-writer] ERROR WriterRunner - Writer Runner Received Exceptions:
com.alibaba.datax.common.exception.DataXException: Code:[DBUtilErrorCode-05], Description:[往您配置的写入表中写入数据时失败。],
com.mysql.jdbc.exceptions.jdbc4.MySQLNonTransientConnectionException: Communications link failure during rollback(). Transaction resolution unknown.
at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)

```

排查思路：

datax 客户端的超时，在添加 DRDS 数据源的时候可以加上

?useUnicode=true&characterEncoding=utf-8&socketTimeout=3600000超时参数

示例如下：

```
jdbc:mysql://10.183.80.46:3307/ae_coupon?useUnicode=true&characterEncoding=utf-8&socketTimeout=3600000
```

系统内部问题



排查思路：

一般是在开发环境将 json 格式该错了然后直接保存，保存在后一般就会报系统内部问题。界面显示为空白，碰到这样的问题，直接将您的项目名和节点名称提供给我们这边会后台处理的。

脏数据

脏数据（String[""]不能转为 Long）

```

2017-09-21 16:25:46.125 [51659198-0-26-writer] ERROR WriterRunner - Writer Runner Received
Exceptions:
com.alibaba.datax.common.exception.DataXException: Code:[Common-01]

```

错误解读：

同步数据出现业务脏数据情况，数据类型转换错误。String[""]不能转为 Long。

排查思路：

String[""]不能转为 Long：两张表格中的建表语句一样，报上述错误是因为字段类型中的空字段不能转换成 Long 类型，直接配置成 String 类型。

脏数据 (Out of range value)

```
2017-11-07 13:58:33.897 [503-0-0-writer] ERROR StdoutPluginCollector
脏数据：
{ "exception" : "Data truncation: Out of range value for column 'id' at row
1" , "record" :[{ "byteSize" :2, "index" :0, "rawData" :-
3, "type" : "LONG" },{ "byteSize" :2, "index" :1, "rawData" :-
2, "type" : "LONG" },{ "byteSize" :2, "index" :2, "rawData" : "其他
" , "type" : "STRING" },{ "byteSize" :2, "index" :3, "rawData" : "其他
" , "type" : "STRING" }], "type" : "writer" }
```

排查思路：

mysql2mysql，源端设置的是 smallint(5)，目标端是 int(11) unsigned，因为smallint(5)范围有负数，unsigned 不允许有负数，所以产生脏数据。

脏数据 (存储 emoji)

数据表配置成了可以存储 emoji 的，同步时报脏数据。

排查思路：

数据集成默认支持的是 utf8，所以添加 JDBC 格式的数据源时，需要修改 utf8mb4 的设置，例如：
jdbc:mysql://xxx.x.x.x:3306/database?characterEncoding=utf8&com.mysql.jdbc.faultInjection.serverCharsetIndex=45，这样在数据源设置表情符可以同步成功。详情请参见 编码格式设置问题。

空字段引起的脏数据

```
{ "exception" : "Column 'xxx_id' cannot be
null" , "record" :[{ "byteSize" :0, "index" :0, "type" : "LONG" },{ "byteSize" :8, "index" :1, "rawData"
:-1, "type" : "LONG" },{ "byteSize" :8, "index" :2, "rawData" :641, "type" : "LONG" }
```

经 DataX 智能分析，该任务最可能的错误原因如下所示：

com.alibaba.datax.common.exception.DataXException: Code:[Framework-14]

错误解读：

DataX 传输脏数据超过用户预期，该错误通常是由于源端数据存在较多业务脏数据导致，请仔细检查 DataX 汇报的脏数据日志信息，或者您可以适当调大脏数据阈值。

脏数据条数检查不通过，限制是 1 条，但实际上捕获了 7 条。

排查思路：

设置 Column 'xxx_id' cannot be null 字段不能为空，但数据中用空数据导致脏数据，修改其数据或对字段进行修改。

Data too long for column 'flash' 字段设置太小引起的脏数据

```
2017-01-02 17:01:19.308 [16963484-0-0-writer] ERROR StdoutPluginCollector
脏数据：
{ "exception": "Data truncation: Data too long for column 'flash' at row
1", "record": [{ "byteSize": 8, "index": 0, "rawData": 1, "type": "LONG" }, { "byteSize": 8, "index": 3,
"rawData": 2, "type": "LONG" }, { "byteSize": 8, "index": 4, "rawData": 1, "type": "LONG" }, { "byte
Size": 8, "index": 5, "rawData": 1, "type": "LONG" }, { "byteSize": 8, "index": 6, "rawData": 1, "type
": "LONG" }
```

排查思路：

设置 Data too long for column 'flash' 字段设置太小，但数据中数据太大导致脏数据，修改其数据或对字段进行修改。

read-only 数据库权限设置问题，设置只读权限

```
2016-11-02 17:27:38.288 [12354052-0-8-writer] ERROR StdoutPluginCollector
脏数据：
{ "exception": "The MySQL server is running with the --read-only option so it cannot execute this
statement", "record": [{ "byteSize": 3, "index": 0, "rawData": 201, "type": "LONG" }, { "byteSize": 8,
"index": 1, "rawData": 1474603200000, "type": "DATE" }, { "byteSize": 8, "index": 2, "rawData": "9
月23号12点
", "type": "STRING" }, { "byteSize": 5, "index": 3, "rawData": "12:00", "type": "STRING" }
```

排查思路：

设置 read-only 模式，同步数据全为脏数据，修改其数据库模式，运行可以写入。

分区错误引起的日志

参数配置为 \${yyyyymm} 报错，日志如下所示：

```
[2016-09-13 17:00:43]2016-09-13 16:21:35.689 [job-10055875] ERROR Engine
```

经 DataX 智能分析，该任务最可能的错误原因如下：

```
com.alibaba.datax.common.exception.DataXException: Code:[OdpsWriter-13]
```

错误解读：

执行 MaxCompute SQL 时抛出异常，可重试。MaxCompute 目的表在运行 MaxCompute SQL 时抛出异常，请联系 MaxCompute 管理员处理。SQL 内容如下所示：

```
alter table db_rich_gift_record add IF NOT EXISTS partition(pt=' ${thismonth}' );
```

排查思路：

由于加了单引号，调度参数替换无效，解决方法：' \${thismonth}' 去掉引号调度参数。

column 没有配成数组形式

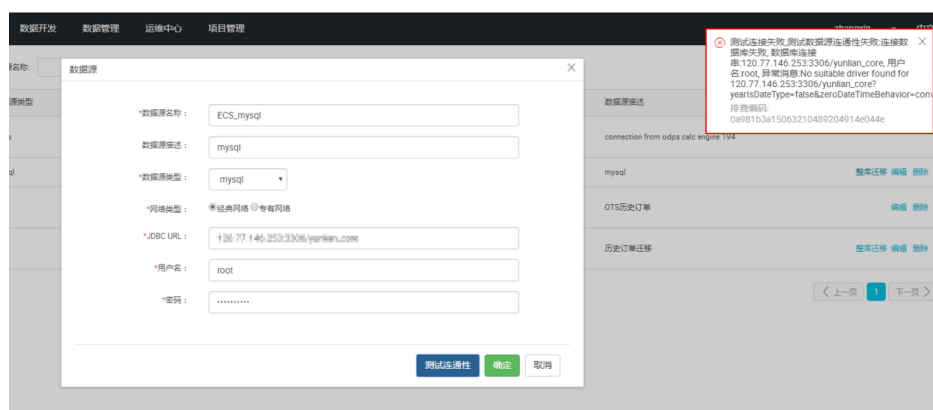
```
Run command failed.
com.alibaba.cdp.sdk.exception.CDPEException: com.alibaba.fastjson.JSONException: syntax error, **expect
{,** actual error, pos 0
at com.alibaba.cdp.sdk.exception.CDPEException.asCDPEException(CDPEException.java:23)
```

排查思路：

json 有问题，如下所示：

```
"plugin": "mysql" ,**
  "parameter" : {
    "datasource" : "xxxxx" ,
    ** "column" : "uid" , **
    "where" : "" ,
    "splitPk" : "" ,
    "table" : "xxx"
  }
  ** "column" : "uid" ,-----没有配成数组形式**
```

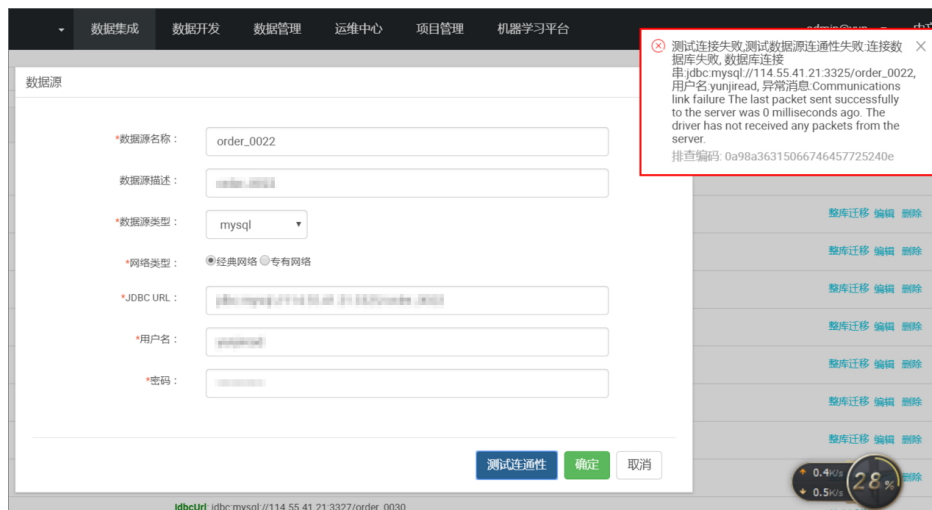
JDBC 的格式填错



排查思路：

JDBC 格式填错，正确的格式是：jdbc:mysql://ServerIP:Port/Database。

测试连通性失败



排查思路：

防火墙对这个 IP 和端口账号有没有相关的限制。

安全组的端口开发情况。

日志里报 uid[xxxxxxxx] 问题

```
Run command failed.
com.alibaba.cdp.sdk.exception.CDPException: RequestId[F9FD049B-xxxx-xxxx-xxx-xxxx] Error: CDP server encounter problems, please contact us, reason: 获取实例的网络信息发生异常,请检查RDS购买者id和RDS实例名,uid[xxxxxxxx],instance[rm-bp1cwz5886rmzio92]ServiceUnavailable : The request has failed due to a temporary failure of the server.
RequestId : F9FD049B-xxxx-xxxx-xxx-xxxx
```

排查思路：

一般 RDS 同步到 MaxCompute 时若报上述错误，您可直接将 RequestId：F9FD049B-xxxx-xxxx-xxx-xxxx 复制给 RDS 同学。

MongoDB 中的 query 参数错误

MongoDB 同步到 MySQL 报下面的问题，排查出 json 没有写好，是 json 中的 query 参数没有配置好。

```
Exception in thread "taskGroup-0" com.alibaba.datax.common.exception.DataXException:
Code:[Framework-13]
```

错误解读：

DataX 插件运行时出错，具体原因请参见 DataX 运行结束时的错误诊断信息。

```
org.bson.json.JsonParseException: Invalid JSON input. Position: 34. Character: ' ' .
```

排查思路：

错误示例：“query”：“ { ‘update_date’ :{‘ \$gte’ :new Date().valueOf()/1000}}” ，不支持 new Date() 这样的参数。

正确示例

：“query”：“ { ‘operationTime’ {‘ \$gte’ :ISODate(‘ \${last_day}T00:00:00.424+0800’)}}”

Cannot allocate memory

```
2017-10-11 20:45:46.544 [taskGroup-0] INFO TaskGroupContainer - taskGroup[0] taskId[358]
attemptCount[1] is started
Java HotSpot™ 64-Bit Server VM warning: INFO: os::commit_memory(0x00007f15ceaeb000, 12288, 0)
failed; error= ‘**Cannot allocate memory’ ** (errno=12)
```

排查思路：

内存不够，如果是用户的机器让用户加内存，如果是运行在阿里的机器上直接联系技术支持。

max_allowed_packet 参数

错误信息如下所示：

```
Packet for query is too large (70 > -1). You can change this value on the server by setting the
max_allowed_packet' variable. - **com.mysql.jdbc.PacketTooBigException: Packet for query is too large
(70 > -1). You can change this value on the server by setting the max_allowed_packet' variable.**
```

排查思路：

- max_allowed_packet 参数用来控制其通信缓冲区的最大长度。MySQL 根据配置文件会限制 server 接受的数据包大小。有时候大的插入和更新会被 max_allowed_packet 参数限制掉，导致失败。

- max_allowed_packet 参数的设置太大就改小的一点，一般 10m=10_1024_1024。

HTTP Status 500 读取日志失败

```
Unexpected Error:
Response is com.alibaba.cdp.sdk.util.http.Response@382db087[proxy=HTTP/1.1 500 Internal Server Error
[Server: Tengine, Date: Fri, 27 Oct 2017 16:43:34 GMT, Content-Type: text/html;charset=utf-8, Transfer-
```

```
Encoding: chunked, Connection: close,  
**HTTP Status 500** - Read timed out**type** Exception report**message**++Read timed  
out++**description**++The server encountered an internal error that prevented it from fulfilling this  
request.++**exception**  
java.net.SocketTimeoutException: Read timed out
```

排查思路：

调度运行报 500 的问题，若是运行在阿里的机器上，读取日志失败，请直接联系技术支持帮您解决。
如果运行在您自己的机器上，请重启 Alisa 即可。

注意：

如果刷新后还是 **停止** 状态，您可以重启 alisa 命令：切换到 admin 账号
/home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart。

hbasewriter 参数：hbase.zookeeper.quorum 配置错误

```
2017-11-08 09:29:28.173 [61401062-0-0-writer] INFO ZooKeeper - Initiating client connection,  
connectString=xxx-2:2181,xxx-4:2181,xxx-5:2181,xxxx-3:2181,xxx-6:2181 sessionTimeout=90000  
watcher=hconnection-0x528825f50x0, quorum=node-2:2181,node-4:2181,node-5:2181,node-  
3:2181,node-6:2181, baseZNode=/hbase  
Nov 08, 2017 9:29:28 AM org.apache.hadoop.hbase.zookeeper.RecoverableZooKeeper checkZk  
WARNING: **Unable to create ZooKeeper Connection**
```

排查思路：

错误示例：“hbase.zookeeper.quorum”：“xxx-2,xxx-4,xxx-5,xxxx-3,xxx-6”

正确示例：“hbase.zookeeper.quorum”：“您的zookeeperIP地址”

没有找到相应的文件

经 DataX 智能分析，该任务最可能的错误原因如下所示：

com.alibaba.datax.common.exception.DataXException: Code:[HdfsReader-08]

错误解读：

您尝试读取的文件目录为空。未能找到需要读取的文件，请确认您的配置项。

```
path : /user/hive/warehouse/tmp_test_map/*  
at com.alibaba.datax.common.exception.DataXException.asDataXException(DataXException.java:26)
```

排查思路：

按照 path 找到相应的地方，检查对应的文件，如果没有找到文件，就对文件进行处理。

表不存在

经 DataX 智能分析，该任务最可能的错误原因如下所示：

com.alibaba.datax.common.exception.DataXException: Code:[MYSQLErrCode-04]

错误解读：

表不存在，请检查表名或者联系 DBA 确认该表是否存在。

表名为：xxxx。

执行的 SQL 为：select * from xxxx where 1=2;

错误信息如下所示：

```
Table 'darkseer-test.xxxx' doesn't exist -  
com.mysql.jdbc.exceptions.jdbc4.MySQLSyntaxErrorException: Table 'darkseer-test.xxxx' doesn't exist
```

排查思路：

select * from xxxx where 1=2 判断表 xxxx 是否存在问题，如果有问题则要对表进行处理。

同步表列名是关键字任务失败

问题描述：

用户做同步任务时，同步的表的列名是关键字，导致任务失败。

处理方法：

以 MySQL 数据源为例：

新建一张表 aliyun，建表语句如下：

```
create table aliyun (`table` int,msg varchar(10));
```

因为 table 是 MySQL 的关键字，那么在数据同步的时候，拼接出来的代码就会报错，这时候我们只能建个视图，然后给这个列取个别名来绕过这个限制（不建议使用关键字作为表的列名），语句如下：

```
create view v_aliyun as select `table` as col1,msg as col2 from aliyun ;
```

上述语句就给有关键字的列取了个别名，那么在配置数据同步任务时，可以选择 v_aliyun 这个视图，来代替 aliyun 这张表。



注意：MySQL 的转义符是 ````；Oracle 和 postgresql 的转义符是 “关键字”；SQLServer 的转义符是 [关键字]。

数据同步任务如何自定义表名

数据背景：表是按天分的（如orders_20170310,orders_20170311,orders_20170312），每天一个表，表结构一模一样。

需求：创建数据同步任务将这些表导入到MaxCompute中，希望只需要创建一个同步任务，表名自定义，实现每天凌晨自动从源数据库读取昨天的表数据（如，今天是2017年3月15日，自动从源数据库中读取orders_20170314的表的数据导入，以此类推）。

实现方式：

步骤一：在大数据开发套件->数据集成通过向导模式创建数据同步任务，配置时数据来源表先选一个表名如orders_20170310。按正常配置流程配置好同步任务并保存好。

步骤二：转换脚本，把向导模式转成脚本模式。



步骤三：在脚本模式中，将来源表的表名改用变量如orders_\${tablename}。然后任务的参数配置中，给变量tablename赋值。因“数据背景”中表名是按天区分，而需求是每天读取昨天的表，所以赋值为\${yyyymmdd-1}。

或者，来源表的表名改用变量时，直接写成orders_\${bdp.system.bizdate}

步骤四：把整个任务其他配置项配置好后，保存提交，再进行后续的其他操作。

编码格式设置问题

数据集成的同步任务配置编码格式后，可能出现以下两种情况：

同步失败且产生脏数据

错误问题：

数据集成任务失败，且因编码问题产生脏数据。

报错日志如下所示：

```
016-11-18 14:50:50.766 [13350975-0-0-writer] ERROR StdoutPluginCollector - 脏数据:<br>
{"exception":"Incorrect string value: '\\xF0\\x9F\\x98\\x82\\xE8\\xA2...' for column 'introduction' at row 1", "record":{"byteSize":8, "index":0, "rawData":9642, "type":"LONG"},
{"byteSize":33, "index":1, "rawData":"优衣库兼职出来的女汉子，扛得了箱子，招待好顾客![1](http://docs-aliyun.cn-hangzhou.oss.aliyun-inc.com/assets/pic/56134/cn_zh/1498728641169/%E5%9B%BE%E7%89%877.png)被自己感动cry", "type":"STRING"},
{"byteSize":8, "index":4, "rawData":0, "type":"LONG"}, {"type":"writer"}
2016-11-18 14:50:51.265 [13350975-0-0-writer] WARN CommonRdbmsWriter$Task - 回滚此次写入, 采用每次写入一行方式提交. 因为:java.sql.BatchUpdateException: Incorrect string value: '\\xF0\\x9F\\x88\\xB6\\xEF\\xB8...' for column 'introduction' at row 1
```

错误原因：

用户要对数据库做相应的编码格式设置或在添加数据源时相应的将编码设置成utf8mb4，这样数据集成支持同步表情符。

解决方法：

添加 jdbc 格式的数据源时，需要修改 utf8mb4 的设置，例如：
jdbc:mysql://xxx.x.x.x:3306/database?com.mysql.jdbc.faultInjection.serverCharsetIndex=45，这样在数据源设置表情符可以同步成功；

将数据源编码格式改成 utf8mb4。例如：在 rds 管控台修改 rds 的数据库编码格式。

同步成功但数据乱码

错误问题：

虽然数据同步任务成功了，但是数据发生乱码。

错误原因：

发生乱码的原因有以下三种：

源端是本来就乱码；
数据库和客户端编码不一样；
浏览器编码不一样，导致预览失败或乱码。

解决方法：

原始数据乱码时，需用户处理好原始数据后再做同步任务；
数据库和客户端编码不一样，需先将编码格式进行修改；
浏览器编码不一致，需先将编码格式进行统一，然后进行数据预览。

如果问题仍未解决，请联系 售后技术支持。

数据开发常见问题

Q：如何创建 MaxCompute 表？

A：大数据平台上有两种方式建表：

在数据开发模块通过脚本文件或者工作流任务执行建表语句，详情请看 [快速开始 > 创建表](#)。以这种方式建表，owner 为项目管理中项目计算引擎设置的 **MaxCompute 访问身份** 账号，无类目管理。

是在数据管理模块可视化方式建表，详情请看 [数据管理 > 创建表](#)。以这种方式建表，owner 为当前登录账号，有类目管理。

Q：怎么把数据导入到 MaxCompute 上？

A：数据导入到 MaxCompute 基本归类为 2 种情况：

本地数据文件一次性导入到 MaxCompute 的某张表。

解决方案：直接在数据开发面板右边工作区域右上角 **新建 > 导入本地数据**，详情请参见 [数据导入](#)。

离线数据库单表导入。

解决方案：单击 **数据集成** 界面左侧导航栏的 **同步任务**，创建数据同步任务。详情请参见 [创建数据同步任务](#)。

Q：RDS 数据源连通测试不通过怎么办？

A:目前 RDS 数据源仅支持杭州地域的 RDS，北京地域暂时不支持。另外当杭州地域的 RDS 也遇到数据源测试不连通的时候，需要到自己 RDS 上添加数据同步机器 IP 白名单：

```
10.152.69.0/24,10.153.136.0/24,10.143.32.0/24,120.27.160.26,10.46.67.156,120.27.160.81,10.46.64.81,121.43.110.160,10.117.39.238,121.43.112.137,10.117.28.203,118.178.84.74,10.27.63.41,118.178.56.228,10.27.63.60,118.178.59.233,10.27.63.38,118.178.142.154,10.27.63.15,100.64.0.0/8
```

注意：

若使用自定义资源组调度 RDS 的数据同步任务，必须把自定义资源组的机器 IP 也加到 RDS 的白名单中。

Q：调度参数怎么用？

A：目前系统参数有两个，这两个系统参数只要代码里用到，不需要做额外的赋值动作，节点自动调度的时候就会自动替换。系统参数分别为：

`${bdp.system.bizdate}`格式 `yyyymmdd`，日常调度实例定时时间的前一天（年月日）。节点每天自动调度实例定时时间年月日减 1 天。

`${bdp.system.cyctime}`格式 `yyyymmddhh24miss`，日常调度实例定时时间（年月日时分秒）。
`yyyy` 表示 4 位数年份，`mm` 表示 2 位数月份，`dd` 表示 2 位数天，`hh24` 表示 24 小时制的时，`mi` 表示 2 位数分钟，`ss` 表示 2 位数秒。

使用方式：节点代码里编辑如 `pt=${bdp.system.bizdate}`，然后节点在调度上运行即可自动替换值。

若节点每天调度时代码里有变量需要赋当天减 2 的日期，即业务日期减 1 天，格式为 `yyyymmdd`，需要系统参数和 MaxCompute 内建函数一起使用，如下：

```
节点代码 dt=to_char(dateadd(TO_DATE('${bdp.system.bizdate}','yyyymmdd'), -1, 'dd'),'yyyymmdd')
```

MaxCompute 内建日期函数具体可查看帮助文档：大数据计算服务 MaxCompute > SQL > **内建函数** 定位到日期函数。

Q：本地文件数据上传功能有没有文件大小限制？

A：目前支持.csv/.txt 文件上传，一次性上传文件大小上限为 10M。

Q：我在“数据开发”模块通过脚本写建表语句建成功建表，但是为什么在“数据管理”模块“数据表管理-我拥有的表”列表里没看到？**

A：遇到此问题，首先要查看当前项目的“项目管理-MaxCompute 基本配置”，MaxCompute 访问身份配置应该是选择了“系统账号”，通过数据开发脚本提交的建表语句都是用该账号创建的表即表 owner 为该账号，而“数据管理-我拥有的表”是指当前登录账号为 owner 的表。

Q：一个工作流里的节点可以不彼此依赖吗？

A：

一个工作流里没有业务关系的节点可以不用彼此依赖。

一个工作流里只能有一个根节点，即没有父节点的节点。

一个工作流里某个节点可以依赖多个节点，也可以被多个节点依赖。

一个工作流里可以有多个节点没有被其他节点依赖。

Q：如何将 RDS 里的数据表同步到 DataWorks 中？

A：DataWorks（数据工场，原大数据开发套件）只能查看 MaxCompute 的数据，RDS 的数据表需要同步到 MaxCompute 的表才可以在 DataWorks 中操作。数据同步请参见 [创建同步任务](#)。

Q：数加平台能不能从 RDS 的只读实例同步数据？

A：支持从只读 RDS 读出数据，然后导到其他的地方。实际上，为了减少对业务库的压力，我们建议做 ETL 的时候，从只读库上导。具体数据同步配置可参考 [创建同步任务](#)。

Q:什么是配置跨周期依赖？

A：跨周期依赖分为以下几种：

- 自依赖。

场景举例：用户配置小时任务所统计，拿上一个小时的统计结果汇总新增数据，产出当前这个小时的统计数据。

用户需要针对当前小时任务配置自依赖，保证每个小时触发执行的时候上一个小时已经执行完了。

等待下游任务的上一周期执行完成。

场景举例：B 是 A 的下游任务，如果昨天的 B 任务没有执行完毕，今天的 A 任务等待昨天 B 任务的最后一个实例执行完以后再执行。

等待自定义任务的上一周期结束。

场景举例同上一个类似，只不过任务 A、任务 B 之间没有相互依赖关系。在此不再赘述。

Q：跨周期依赖的场景有哪些？

A：* 天周期任务依赖小时周期任务。

场景举例：用户配置了一个天任务，依赖小时任务，天任务对小时任务做汇总，需要每天 24 个小时任务全都跑完之后才能执行。

用户需要针对小时任务配置自依赖，保证上游的小时任务全都执行完成。

Q:同事离职后，修改任务流的责任人，却发布失败，提示节点责任人不在工作流中。

A：1、先在运维中心修改责任人，修改时展开工作流列表，全选工作流和节点，然后修改责任人。如下图所示：



2、修改责任人后，在数据开发中找到该工作流，查看责任人是否被修改，确定无误后在工作流页面随意移动一下节点(修改一下工作流)，然后提交，重新打包、发布。如下图所示：



调度运维常见问题

Q： workflow 节点出错怎么看日志？

A：进入“运维中心-任务运维-运维”，过滤条件：输入任务名称、业务日期选“接到的节点实例出错提醒的业务日期”，查询，结果列表选中任务实例，右边 DAG 图里，鼠标对实例右键点击‘查看节点运行日志’。

Q：刚提交好了一个 workflow，设置为周期调度，定时时间已经到了，怎么还不见跑起来？

A：当天新增成功的 workflow/节点，且在当天 23：30 前提交/发布成功，需要第二天才会自动调度。当天可以通过手动测试或补数据调度。

Q：创建了周期调度 workflow，且昨天就 23：30 前就发布成功了，今天运维中心里却是未运行状态？

A：以下几种原因都会导致 workflow 未运行，注意逐一排查：

1. workflow 属性-调度属性-启动调度选项没有勾选。
2. workflow 虽然勾选‘启动调度’但是今天的日期不在设置的‘生效时间’范围内。
3. workflow 设置的定时调度时间还没到。
4. workflow 设置有上游依赖，上游 workflow 未全部运行成功。
5. workflow 设置有跨周期依赖，上一周期 workflow 未运行成功。

说明：节点任务排查方式与工作流相同。

Q：工作流前段时间每天都正常调度，今天突然没有了，工作流实例都没生成。

A：请注意查看工作流是否已经删除，或者工作流属性的生效时间已经过了，若生效时间已过可以修改。

Q:数据同步任务等待资源时间过长，一直等待资源？

A：如果任务使用自定义资源组，那么：

1. ECS 上已经跑了太多的任务，导致任务排队；
2. ECS 节点无法正常工作，可以到 组织管理-》调度资源 -》资源服务器 看下 ECS 资源是否正常，也可以考虑重启一下 ECS 上的 agent，命令是:su - admin/home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart

如果任务使用默认资源组，那么就是资源组并发已满，需要等待或者可以考虑使用自定义资源组。

**Q：数据同步任务报错：获取作业配置信息失败
:/home/admin/shell_datax/T3_0028378354 - File
'/home/admin/shell_datax/T3_0028378354' does not
exist**

A：这个错误一般是通过自定义资源组执行 shell 调 datax 的任务，可以在自定义机器上 su - admin切换到 admin 账号后，执行下 datax 任务调试下，常见原因：

- 权限问题（datax 的任务是用 admin 账号执行的，如果用 root 配置了环境，可能有一些操作 admin 账号会没权限）。
- datax 配置的时候就有问题，无法正常执行。

如果还是无法解决请提交工单寻求帮助！

Q：整个工作流中间的一个节点出错，能否跳过先执行其他的节点？

A：如果节点没运行成功，下游节点是不能正常运行。若节点之间业务逻辑没有依赖关系，则应该取消其直接的调度依赖；若仅仅是本次执行可以不依赖出错节点的执行结果，则可以对出错的节点实例操作右键“置成功并恢复调度”这样可以先跳过出错节点调度下游节点。

Q：为什么按照流程部署完毕了 ECS，但界面还是显示“终止

” ？

A：

- 如果是经典网络，请确认您注册使用的机器名称是真实的机器名,请使用命令到ecs上执行：`hostname`，返回的结果即为机器名称；不支持自定义名称；
- 如果是专有网络，请确认使用的是界面提供的命令查询到的 `uuid`；
- 请确认是否有修改过 `ecs` 的主机名称，注意这里不是指实例名。如果修改过名称，请到 `ecs` 上查询 `/etc/hosts` 中是否有绑定正确；
- 如果注册信息都是正确的，而且也是按照“执行初始化”的命令复制粘贴部署的，但还是“终止”，请将 `ecs` 上的文件 `/home/admin/alisa-tasknode/logs/alisa-tasknode.log` 这份文件转为 `txt` 格式，提工单反馈给我们；

Q：配置服务器的时候，在操作“执行初始化”的时候失败？

A：目前只能支持 `centos5`、`centos6`、`centos7` 以及 `aliyunos`。如果您的 `ecs` 不是上述的 `os`，系统暂时不支持，会出现上述操作。

Q：为什么添加了调度资源，也安装 ECS 成功了，但是 shell 任务还是执行失败，报错“exec target was null”？

A：执行 `shell` 任务，需要将任务的运行资源组修改为自定义资源组；

如何新建自定义资源组请前往《项目管理》文档中查看。如果已经拥有自定义资源组，请前往运维中心修改任务执行的资源组。

Q：为什么我注册了 ECS 服务器，经常会出现自动的“暂停服务”的状态？

A：在注册 `ECS` 的时候，部署了一个 `agent`，`agent` 会有一个监听动作，发现当前 `ECS` 的内存、`cpu`、`load` 较高的时候，不适合继续接收新的任务执行，会主动将 `agent` 服务暂停。待机器指标下降后会自动置为“正常”状态；

Q：为什么安装 ECS 的时候，部署完成了，但是启动失败？

A：请检查您的 `ECS` 系统是否有 `jdk1.6` 以上的环境；如果您是 `centos5.x` 的操作系统需要您自己安装 `jdk` 的版本以及 `python2.6.5` 以上的版本；

Q：如何修改自定义 ECS 的任务并发数？

A：在配置服务器的界面中，有个最大槽位数：说明：槽位数是定义一台机器并发任务的一个量化单位，一个

shell 任务或者 MaxCompute sql 任务占用一个槽位，一个同步任务可能会占用 2-20 个不等的槽位。可以修改该值增加您 ECS 并发的大小，目前默认是 40 个槽位；

Q：为什么我的调度资源偶尔会出现“暂停服务”的状态？

A：由于在服务器上安装的 agent 会自动监控当前 ECS 的 cpu、mem、load 的使用率，监控的频率是每分钟监控 3 次，当发现使用率超过 80% 就会将 agent 的服务挂起，也就是在界面看到的“暂停服务”；

注意：该功能并不会影响正在执行的任务，但是不会在接收新的任务，待负载等指标下降后，会自动将服务器状态置为“正常”，服务器恢复正常工作。

Q：为什么跑在自定义调度资源上的任务执行成功/失败，但是日志无法获取？

A：

- 如果服务器是经典网络的，那么请确认内网入的 8000 端口是否开通了；
- 在调度资源中检查当前服务器的状态，是否“正常”；
- 如果上两步骤都是正常的，那么尝试重启下 agent 的服务，重启命令：`sudo su - admin /home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart`

Q：为什么同步任务执行成功了，但是日志中却有500错误的返回？

A：任务成功了，只是日志读取失败了，是因为 ECS 服务 agent 异常，需要重启下 agent，重启命令：`sudo su - admin /home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart`重启完毕后，可以重跑下同步任务，在检查日志是否完整；

Q：为什么任务经常处于“等待资源”？

A:

1. 如果任务是跑在**默认资源组**上：我们默认给每个租户50个槽位的并发数，可能您当前的任务已经占用超过了50个槽位，所以正在等待资源。
 2. 如果任务是跑在**自定义资源组**上：那么可以到调度资源管理界面选择对应调度资源，点击服务器管理，查看服务调度资源的服务器状态。
- **停止状态**：请登录 ECS 执行命令`sudo su - admin /home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart`恢复 agent 服务；
 - **暂停服务状态**：那么是很可能因为 ECS 上跑的任务占用了较高的 cpu 或者内存，导致了挂起。建议如果任务并发量过高，建议将 ECS 的内存 cpu 扩容下；

- **正常状态**：需要检查下当前是否任务数已经达到了 ECS 服务器的并发数上限。可以到运维中心中检查，当前调度资源下有哪些正在执行的任务。

Q：任务没有运行日志？

A：无法获取运行日志（一般出现在使用自定义调度资源的情况下），下面就说一下两种处理方案。

1. 自定义调度资源的机器在初始化时没有对我们读取日志的机器开放白名单以及8000端口，详细操作请参考新增调度资源，详细请参考“**开通 8000 端口，以便读取日志**”。
2. agent是我们去连接您机器的服务，读取日志失败可能是agent服务挂了，可以使用如下命令重启：
“`sudo su - admin /home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart`”

我们的机器只会拉取一次日志数据，如果第一次拉取日志的时候失败了，那么这个任务的日志就无法获取了，只能再次运行该任务后，才会重新拉取一次。

数据管理常见问题

Q：绑定已有MaxCompute project新建DataWorks项目后，进入数据管理模块发现MaxCompute project中原有的表没有在列表中展示

A：遇到此问题，属于正常范畴。采用这种方式构建DataWorks项目空间，MaxCompute project中已有的表需要T+1的时间才能在数据管理模块列表中显示。也就是说需要到第二天的8点-10点之后才能够显示。

Q：数据开发中可以执行建表语句成功建表，但在数据管理中创建表却没权限

A：遇到此问题，首先要查看当前项目的**项目管理>项目属性>技术引擎配置>ODPS访问身份**，此配置选择的是**计算引擎指定账号**，通过数据开发脚本提交的建表语句都是用该账号创建的表，该账号拥有一般都会有权限。而数据管理建表用的是当前登录账号，该账号需要权限必须找项目owner给开通。详情请参见数据权限点划分。

Q：表的真实存储信息和表详情页中看到的存储信息不一致

A：数据管理模块中表存储信息是离线计算得出的，默认是一天同步一次这个信息，是非实时同步，所以与真实信息会不一致。

权限相关常见问题

Q：如何申请查询别人的表的权限

A：导航至数据管理模块>查找数据>申请数据权，如下图所示：



提交申请后，等待表所属项目的管理员审批，审批入口在数据管理模块>权限管理>待我审批。

Q：被添加为项目成员后，是否拥有该项目所有表的查询权限

A：请管理员授予项目成员开发角色权限，即可拥有查询该项目下所有表的权限。

Q：数据管理模块申请权限了，但删除表又重建后权限丢失

A：数据管理模块权限申请走的是MaxCompute的ACL授权方式，表删除会把该表相关的ACL授权都清空，重建不会恢复。目前这个现象流程上还没有更好的解决方式，只能项目管理员命令行通过package方式授权解决。详情请参见MaxCompute安全文档。

Q：为何A项目的开发角色，可以操作B项目下的表

A：使用主账号查看项目管理下的Maxcompute配置，是否将Maxcompute访问身份设置为系统账号。

注意：

Maxcompute访问身份是指SQL执行时，使用的是系统账号的AK去执行，还是使用个人账号的AK去执行。

。

使用Shell/MR时的常见问题

Q：如何配置/使用 OPEN_MR 类型任务

A：需要完成以下操作：

上传 MR Jar 包：所使用的 Jar 包，需要包含 Map，Reduce，Combiner 类对应的 class 文件，可以不包含 main 方法。

上传资源：MR 代码里涉及的资源：包括 Jar 包、资源文件（资源表暂时不能支持，将在最后说明 work around 的方法）。

输入表设置：Map 端的输入表（支持多张表）。指定项目的数据表则填写格式为：`projectName.tableInfo`，如果直接写 `tableInfo` 则默认为当前 project。

注意：

如果是非分区表，直接写入表名。

如果是分区表，由于目前仅支持一级分区表，则分区表格式为：
`tablename/partitionSpace`，比如：`partition_table/ds=20160101`（此时的 `ds='20160101'` 将被系统自动识别为输入）。

- 如果是多级分区表，则只需写一级分区表的表名，而此表名下的所有分区都会被当成输入。

填写 Mapper/Reducer/Combiner 信息。

填写 Map/Reduce/Combiner 对应的类时应当注意：

默认需要写全路径（`packageName.className`），比如
`com.aliyun.odps.mr.TokenizerMapper`。

如果是内部类，需要是 `packageName.className$inerrClassName`，类似
`com.aliyun.odps.mr.WordCount$TokenizerMapper`。

Combiner 可为空，Mapper 不能为空。Reducer 只有 MapOnly（后面会提到的 NumReduceTasks 为 0）才可以为空。

输出表设置：同输入表，可以用 `project.tableInfo` 来指定 project，也可以只指定 `tableInfo`。

注意：

不支持多级分区，如果是分区表，必须指定需要写入的具体分区。如果对应的分区没创建会自动创建。

输出 Key/ Value 设置：

此处配置的是 Map 往 Reduce 发送数据的时候的格式，当 MR 任务不是 MapOnly 的时候需要配置。格式为：col1:type1，col2:type2。

其他设置：

OutputKeySortColumns 设置 Map 到 Reduce 的排序，用逗号分隔，对应到 SDK 为 `public Builder setOutputKeySortColumns(String[] sortCols)`。

OutputGroupingColumns 和 PartitionColumns 组合配置 Map 往 Reduce 发送数据的分组方式，对应 sdk 里的 `public Builder setOutputGroupingColumns(String[] cols)` 和 `public Builder setPartitionerClass(Class<? extends Partitioner> theClass)`，以逗号分隔多列。

以上 3 个设置对应的使用方法，详情请参见 示例。

SplitSize 设置 Map 读取文件的时候的单个块大小，默认是 640M，可以输入一个数字来设置（单位是 M）对应的 SDK 是 `void setSplitSize(long size)`。

NumReduceTasks 一般不需要设置 Reducer 的个数，如果是 MapOnly,则需要设置成 0。对应到 SDK 里是 `void setNumReduceTasks(int n)`。

MemoryForMapTask/MemoryForReduceTask 设置 Map 和 Reduce 的内存大小，对应的 SDK 是 `void setMemoryForMapTask(int mem)` 和 `void setMemoryForReduceTask(int mem)`。单位是 MB，默认值是 2048。

Q：怎么顺利执行 Shell/MR 类型任务

A：需要完成如下操作：

购买 ECS 云服务器。

注意：

建议使用 centos6、centos7，或者 aliyunos，其他操作系统初始化脚本暂时不支持。

如果您添加的 ECS 需要执行 MaxCompute 任务或者同步任务，需要检查当前 ECS 的 python 版本是否是 python2.6.5 以上的版本（centos5 的版本为 2.4，其余 os 自带 2.6 以上版本）。

请确保 ECS 有公网 IP。

建议 ECS 的内存大小在 8G 以上。

用户自定义添加的 ECS 只能支持执行 同步任务、ODPS_SQL、MR 和 SHELL 的任务类型，不支持其它任务类型。

您需要购买 华东2 经典网络 ECS 服务器，否则查询日志会出现网络不可达的问题，如果您是华东2的ECS建议在添加服务器选择专有网络类型。

添加安全组规则。

导航至安全组点击配置规则，进入配置规则页面，如图所示：



- 添加内网入方向规则：



配置 IP：10.116.134.123 访问 8000 端口。

添加公网出方向规则：

配置可以访问公网 IP

注意：

ECS 需开通 8000 端口，因为读取日志是通过 IP+port 方式获取的本地日志，如果不开通则无法查阅日志。若不开通公网出的 80 端口，则无法正常将该 ECS 注册到 DataWorks（数据工场，原大数据开发套件）中。

查看 ECS 主机名和 IP 地址。

购买的 ECS 主机名和 IP 如图所示：



组织管理员进入 **DataWorks > 调度资源列表**，单击 **新增调度资源**，填写新增的调度资源名称。添加后，在弹窗界面内单击新建调度资源操作栏中的 **服务器管理**，进入服务器添加页面。将购买的 ECS 云服务器添加到资源组。





单击 **增加服务器**，并在添加服务器弹出框中选择网络类型、输入 ECS 主机名和 机器 IP，然后点击 **提交**。

注意：

如果您的 ECS 未开通 VPC，请允许该服务器可访问互联网（外网出端口 80），以及可被内网以 8000 端口访问（内网入端口 8000）。

如果您的 ECS 开通了 VPC，请允许该服务器可访问互联网（外网端口 80），专有云网络填写 ECS 的 UUID，获取 UUID 的方式：

登录到 ECS 机器执行命令：`dmidecode | grep UUID`；

执行该指令后，如果系统返回结果为：713F4718-8446-4433-A8EC-6B5B62D75A24，则对应的 UID 为 713F4718-8446-4433-A8EC-6B5B62D75A24

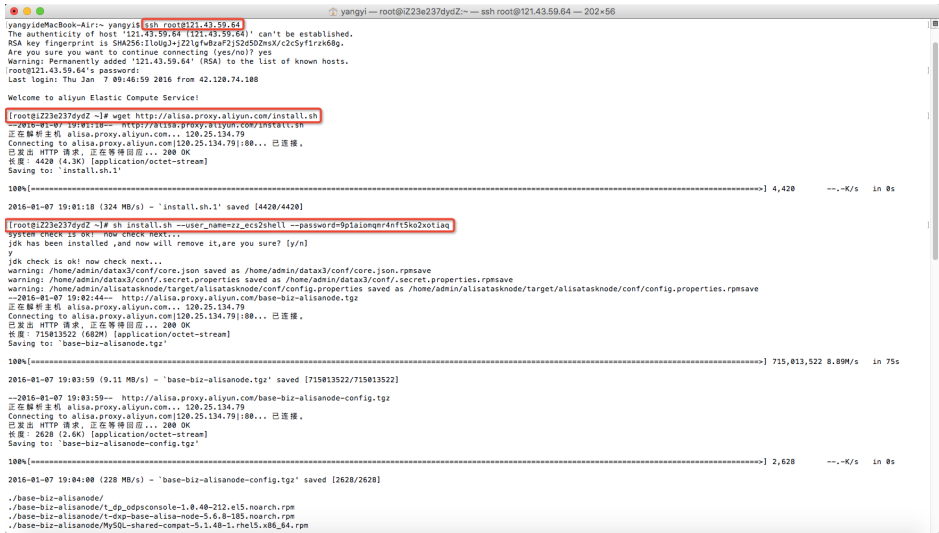
。

经过上述步骤后，已经将新购买的 ECS 信息注册到了 DataWorks 中，但是目前为止还不能服务。如果是新添加机器，请按照如下步骤操作：



注意：

如果执行 `install.sh` 过程中出错或需要重新执行，请先删除已经生成的文件，请在 `install.sh` 的同一个目录下执行：`rm -rf base-biz`；然后执行：`install.sh`。



稍后（大约 15 秒后）在添加服务器页面，单击 **刷新** 按钮，观察服务状态是否转为 **正常** 状态，若显示正常则表示新建 ECS 服务注册成功。

Q：为什么按照流程部署完 ECS，界面依旧显示终止？

A：部署完 ECS，界面依旧显示终止，原因如下：

- 如果是经典网络，请确认您注册使用的机器名称是真实的机器名,请在 ECS 上执行命令：`hostname`，返回的结果即为机器名称（不支持自定义名称）。

如果是专有网络，请确认使用的是界面提供的命令查询到的 `uuid`。

如果注册信息都是正确的，而且也是按照 **执行初始化** 的命令复制粘贴部署的，但还是 **终止**，请将 ECS 上的 `/home/admin/alisa-tasknode/logs/alisa-tasknode.log` 文件转为 `txt` 格式，提工单反馈给我们。

Q：为什么添加了调度资源，并成功安装了 ECS，但是 shell 任务还是执行失败，报错：exec target was null？

A：执行 shell 任务，需要到 **运维中心 > 任务管理 > 列表** 将任务修改资源组，修改到自定义创建的调度资源中。

Q：如何使用 shell 执行 MR 任务？

A：shell 执行 MR 任务的逻辑同您在本地使用 `odpscmd` 的执行逻辑一样，比如以下执行代码：

```
##@resource_reference{"aaa.jar"}  
/opt/taobao/tbdpapp/odpswrapper/odpsconsole/bin/odpscmd -u testid -p testkey --project=testproject --  
endpoint=http://service.odps.aliyun.com/api -e "jar -resources aaa.jar -classpath ./aaa.jar com.XXX.A"
```

注意：

##@resource_reference{"aaa.jar"} 表示您上传到 DataIDE 中的资源 Jar 包，您可以直接在编辑界面找到这个资源，然后单击 **引用** 即可，无需自己编写。

aaa.jar：表示 Jar 的名称。这个名称需要保证资源名称、classpath 的名称、resource 的名称都一致，不然会报找不到对应的 Jar。

需要自己传 accessid，accesskey，project，endpoint；其中 endpoint 固定为
endpoint=http://service.odps.aliyun.com/api。

-e 后面就是需要执行的代码，需要用双引号括起来。执行不同的 MaxCompute 任务，只需要替换双引号内部的代码即可，其余为固定的执行命令。

Q：为什么我注册了 ECS 服务器，经常会出现自动的暂停服务的状态？

A：在注册 ECS 的时候，部署了一个 agent，agent 会有一个监听动作，发现当前 ECS 的内存、cpu、load 较高，不适合继续接收新的任务执行时，会主动将 agent 服务暂停，待机器指标下降后会自动置为 **正常** 状态。

Q：为什么安装 ECS 时，部署完成了，但是启动失败？

A：请检查您的 ECS 系统是否是 jdk1.6 以上的环境。如果您是 centos5.x 的操作系统需要您自己安装 jdk 的版本以及 python2.6.5 以上的版本。

用户常见问题反馈

操作项目报错accessKeyId找不到

用户报错：FAILED: ODPS-0410051:Invalid credentials - accessKeyId not found: LTA**

原因：用户在重新生成ak后没有重新绑定账户

解决方法：登录下ak-console.aliyun.com看看这个key还在不，是不是启用状态如果key变了得重绑一次。必须要有主账号的key才能工作，注意保护好key，不要泄露。

修改AccessKey信息链接：<https://data.aliyun.com/console/akinfo>

如何同步二进制数据到MaxCompute？

目前数据同步功能暂时还不支持直接同步二进制数据，当数据源为关系型数据库时：

解决方案

如，同步mysql的二进制数据，可以将二进制数据转换成字符串，比如用base64转mysql里的二进制数据列转成字符串（即同步任务配置的时候二进制字段用to_base64(colname)作为映射的字段名字），然后用字符串来同步到MaxCompute的String列，后续在MaxCompute里，您可以用UDF/MR读取这个字符串的内容来进行计算。

其他关系型数据库的二进制数据同步也是一样的原理，同步任务配置里用合适的数据库函数转换数据。

如何跨项目配置任务依赖

问题描述

如何跨项目配置大数据开发套件的任务依赖。

处理办法

用户可以在脚本开发页面配置任务依赖的时候，设置任务依赖为其他项目的依赖：

用户可以先输入需要依赖的任务所在的项目空间，然后选择上游任务。如果任务比较多，也可以输入关键字进行过滤。

注意：目前只支持同一个组织下的不同项目才可以跨项目依赖。

调度属性 ▾

调度状态：☐ 暂停

生效日期：

1970-01-01

至

2115-12-19

*调度周期：

天

*具体时间：

00

时

00

分

依赖属性 ▾

自动推荐

所属项目：

请输入项目

上游任务：

请输入关键字查询上游任务

Q

项目名称	任务名称	责任人	操作
没有依赖上游任务			

跨周期依赖 ▾

☒ 不依赖上一调度周期

☐ 自依赖，等待上一调度周期结束，才能继续运行

☐ 等待下游任务的上一周期结束，才能继续运行

☐ 等待自定义任务的上一周期结束，才能继续运行

如果问题还未能解决，请联系售后技术支持。

同步数据报版本号异常

问题描述

用户配置的同步任务运行失败，伴有报错日志：

- ErrorMessage:Code:[OdpsWriter-10], Description:[提交 block 到 ODPS 目的表失败]. - ODPS 目的表在提交 block:[0] 时失败, uploadId=[XXXXX]. 请联系 ODPS 管理员处理. - ErrorCode=DefaultClusterNotLatestVersion, ErrorMessage=The specified table is not the latest version on default cluster and can't upload data because the data is being replicated at this time. Please initiate another upload later.

问题原因

该错误通常是由于数据同步任务涉及的MaxCompute表，在同步的过程中发生了变化导致的。

解决办法

用户可以等前面的数据导入任务结束后对这个任务再进行重试。另外因为MaxCompute里没有事物的概念，需从业务上做好任务依赖。用户可以通过大数据开发套件里的任务的依赖关系功能，保证上下游任务有序进行。

如果问题还未能解决，请联系售后技术支持。

认为未正常运行，提示exec target empty

用户问题

任务未正常运行，日志提示：

execute task failed, exception: [103:ILLEGAL_TASK]:exec target empty

问题原因

Shell任务目前需要运行在用户的ECS上，默认资源组不能运行Shell脚本。出现这个提示的原因是因为任务是Shell任务未正确配置自定义资源。

1. 用户可以在大数据开发套件->运维中心->任务运维中，找到没有如预期调度的任务，右键单击**查看节点属性**，核对是否没有正确设置资源组。
2. 在项目管理->调度资源管理中配置一个可用的调度资源。
3. 在运维中心->任务管理，单击右上角列表选择列表模式。勾选需要调整资源组的任务，然后单击修

改资源组的按钮。



如问题还未解决，请联系售后技术支持。

调度服务器ecs经常异常的常用解决办法

ecs调度服务器出现网络异常、机器负载异常等情况下，agent服务异常导致任务执行出错；

解决办法：

可以配置下agent定时自启动脚本。

部署方式：

- (1) 登陆ecs服务器，切换到admin账户 (su - admin) ;
 - (2) 在ecs的/home/admin目录下创建一个shell脚本，名称为：alisa_start.sh,
内部代码：sh /home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart
 - (3) 配置crontab:
 - a) 执行crontab -e
 - b) 在内部添加如下命令
- ```
30 18 * sh /home/admin/alisa_start.sh >> /dev/null 2>&1
```

## 数据同步中表数据能否做到增量同步？

### Q：对于不同类型的数据，如何进行增量同步？

A：如人员表、订单表一类的会发生变化的数据，根据数据仓库的4个特点里的反映历史变化的这个特点的要求，建议您每天对数据进行全量同步。也就是说每天保存的都是数据的全量数据，这样历史的数据和当前的数据都可以很方便地获得。

不过如果真实的场景下因为某些特殊情况，需要每天也只做增量同步，因为MaxCompute不支持Update语句来修改数据，只能用别的一些方法来实现。

不同的数据场景需设计不同的同步策略，详情请参见数据增量同步。

## 大数据开发套件无法登录崩溃

用户经常反馈无法登录或者登录报错，跳转页面异常，空白等。

请使用chrome最新版本浏览器，并且清空缓存，cookie，刷新或者重新登录尝试一般都可以解决。

## 未运行任务处理方案

## Q：上游有很多未运行的任务，但是我不想继续运行了，该如何处理？

A：处理此问题时，分为以下两种情况：

该任务是自依赖：

跨周期依赖 ▾

☐ 不依赖上一调度周期

☒ 自依赖，等待上一调度周期结束，才能继续运行

☐ 等待下游任务的上一周期结束，才能继续运行

☐ 等待自定义任务的上一周期结束，才能继续运行

此时您只需要将第一个未运行的任务找到，将它置为 **冻结**，下面的任务便不会再继续运行。

该任务是不依赖：

跨周期依赖 ▾

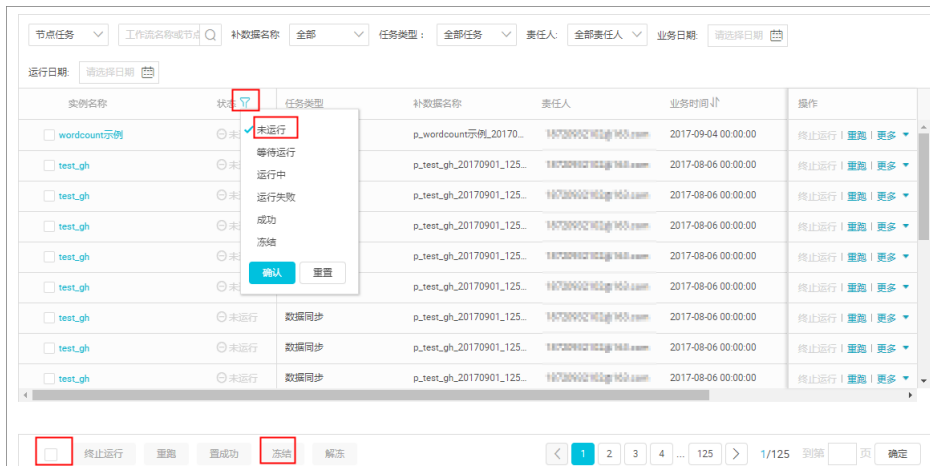
☒ 不依赖上一调度周期

☐ 自依赖，等待上一调度周期结束，才能继续运行

☐ 等待下游任务的上一周期结束，才能继续运行

☐ 等待自定义任务的上一周期结束，才能继续运行

此时您需要将所有未运行的任务都列出来，然后批量操作，将他们置为 **冻结** 状态。



**注意：**

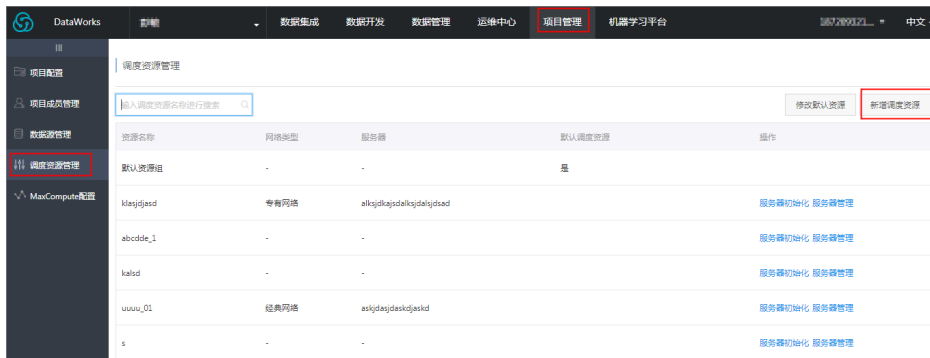
如果不想去看该任务是什么调度属性的话，建议找出未运行状态的任务，批量将任务置为冻结。

## Q：大量任务未运行，该如何处理？

A：大量任务未运行有以下两种情况：

任务数量太多，没有任务失败，但是资源不足，大量任务都是处在等待资源的状态。这种情况没有别的办法，只能在项目管理中添加资源组，购买一批 ECS 来作为调度资源，详情请参见 [新增调度资源](#)。

。



上游或上周期有任务失败了，导致下面的任务不能满足运行条件。

您可以通过以下两种方案进行排查：

- 先在 **运维中心 > 周期实例** 中，输入该任务的名称，然后过滤出失败的任务，查看是否有失败的任务。如果有失败任务的话，将该任务置为成功，下游任务就会开始运行起来了。如果没有失败的任务的话，那么请参考如下方案。

如果该任务没有失败记录，说明是上游任务失败了，那么此时先去 **数据开发** 中，找到这个

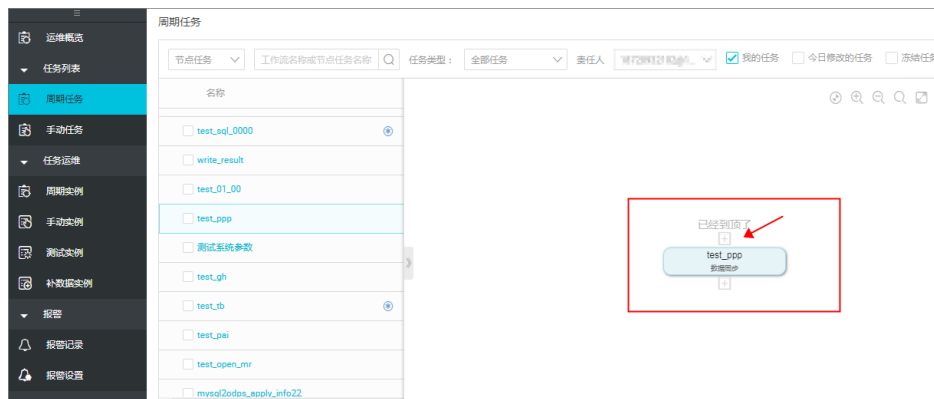
未运行任务，查看调度配置中 **依赖属性** 中，该任务的上游任务。然后进入 **周期实例**，输入该任务的上游任务名，过滤出失败的任务，查看是否有失败的上游任务。如果有失败的上游任务的话，将该上游任务置为成功，下游任务就会开始运行起来了。

| 实例名称                                            | 状态 | 任务类型     | 补数据名称                     | 责任人                 | 业务时间                | 操作                          |
|-------------------------------------------------|----|----------|---------------------------|---------------------|---------------------|-----------------------------|
| <input checked="" type="checkbox"/> wordcount实例 | 失败 | OPEN_MMR | p_wordcount实例_20170...    | 18720903102@163.com | 2017-09-03 00:00:00 | 终止运行   重跑   更多              |
| <input type="checkbox"/> test_gh                | 失败 | 数据同步     | p_test_gh_20170901_135... | 18720903102@163.com | 2017-08-31 00:00:00 | 终止运行   重跑下游   置成功           |
| <input type="checkbox"/> test_gh                | 失败 | 数据同步     | p_test_gh_20170901_135... | 18720903102@163.com | 2017-08-31 00:00:00 | 终止运行   重跑下游   置成功   冻结      |
| <input type="checkbox"/> test_gh                | 失败 | 数据同步     | p_test_gh_20170901_135... | 18720903102@163.com | 2017-08-31 00:00:00 | 终止运行   重跑下游   置成功   冻结   解冻 |

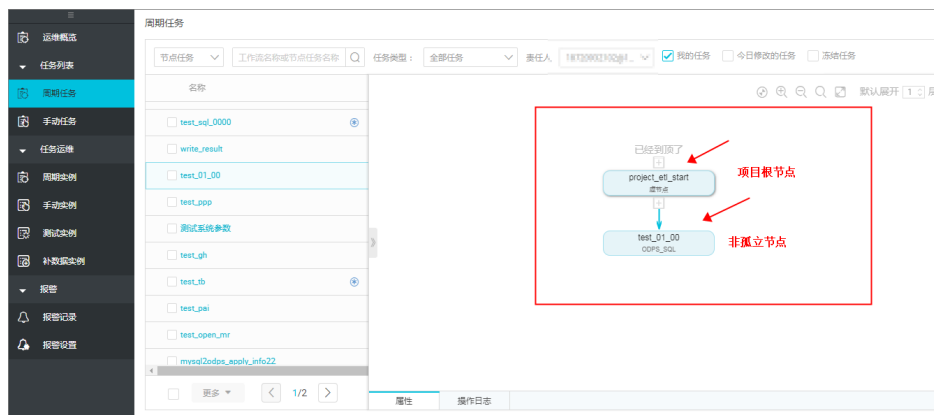
如果您不想放弃失败的任务，可以将失败的任务重跑，重跑成功后，下游任务也会开始运行

•

如果任务有实例生成，上游也没有失败的任务，但是就是运行不起来，这个可以思考一下，当时是否有删除过该任务的上游节点，导致该任务被孤立了。您可前往 **周期任务** 中查找该任务，在 DAG 环形图中，查找该任务的上游任务。如果在查找上游任务的时候，提示 **已经到顶了**，那么该节点是一个孤立节点。



非孤立节点如下：



project\_etl\_start 是项目的根节点，调度任务运行时，整体的实际是一个树状结构，所有的节点都只有挂在项目根节点下才会被调起，否则叫孤立节点。

## 数据外流选项禁止用户无法下载数据

```
odps@ nation_pol>tunnel download province_pinyin pinyin.csv;ERROR: TunnelException - RequestId=20170223134824123c9b0a00712da2, ErrorCode=NoPermission, ErrorMessage=Authorization Failed [4021], You have NO privilege 'odps:Select' on {acs:odps:projects/nation_pol/tables/province_pinyin}. project 'nation_pol' is protected. You need to contact the project owner to set the exception policy for you. Context ID:09f8cf7b-1667-42b2-8ba8-c1651c99d555. ---> Tips: Principal:p4_228321584552544298; project nation_pol is protected. ProjectOwner can set EXCEPTION POLICY for you to DOWNLOAD acs:odps:projects/nation_pol/tables/province_pinyinFAILED: error occurred while running tunnel command
```

用户在下载表数据中报错显示项目受保护，数据无法下载，是因为项目默认设置了禁止数据外流的。关闭该按钮就可以。可以找一下项目project owner 看下这个表的配置，在大数据开发套件-》项目管理-》MaxCompute配置-》项目空间数据保护 看下配置是否开开启了限制数据外流。



对应的MaxCompute命令请参考MaxCompute 安全指南中《项目空间的安全配置》。

## 任务调度500报错

任务同步报错：Response is com.alibaba.cdp.sdk.util.http.Response@58363f95[proxy=HTTP/1.1 500 Internal Server Error [Server: Tengine, Date: Fri, 17 Mar 2017 07:23:57 GMT, Content-Type: text/html;charset=utf-8

排查方法：检查自己的调度服务器状态是否正常，查看调度服务器上的安全组规则是否争取。配置内网入ip：10.116.134.123 访问8000端口。确保有公网ip。

# 使用shell执行datax任务

如果没有配置好shell环境，请参考此链接的文档：Q：怎么顺利执行Shell/MR类型任务

配置完成后，在shell节点填写代码。可以参考下面的demo代码

## - 代码说明

- 生成临时datax配置文件（您只需要修改配置文件内容即可，其他不用更改。参考文档：<https://github.com/alibaba/DataX/wiki/Quick-Start>
- 读取调度参数，分别为\$1, \$2
  - `${bdp.system.bizdate}`，`${bdp.system.cycltime}` 不需要配置，具体含义参考：系统参数
- 执行datax任务，进行数据同步
- 删除临时文件
- 判断任务成功失败，进行返回。0代表成功

```
#!/bin/bash
#*****#
##author:yixiao streamreader->streamwriter
##create time:2016-03-28 10:00:00
#*****#

shell_datax_home='/home/admin/shell_datax'
mkdir -p ${shell_datax_home}
shell_datax_config=${shell_datax_home}/${ALISA_TASK_ID}
echo ""
{
"job": {
"setting": {
"speed": {
"channel": 1
},
"errorLimit": {
"record": 0,
"percentage": 0.02
}
},
"content": [
{
"reader": {
"name": "streamreader",
"parameter": {
"column": [
{
"value": "${bdp.system.bizdate}",
"type": "string"
},

```

```

{
 "value": "${bdp.system.cyctime}",
 "type": "string"
},
{
 "value": "${params1}__${params2}",
 "type": "string"
},
{
 "value": 19890427,
 "type": "long"
},
{
 "value": "1989-06-04 00:00:00",
 "type": "date"
},
{
 "value": true,
 "type": "bool"
},
{
 "value": "test",
 "type": "bytes"
}
],
"sliceRecordCount": 10
}
},
"writer": {
 "name": "streamwriter",
 "parameter": {
 "print": true,
 "encoding": "UTF-8"
 }
}
}
]
}
}
''' > ${shell_datax_config}

params1=$1
params2=$2
datax_params='-p "-Dparams1=${params1} -Dparams2=${params2}"'

echo "`date '+%Y-%m-%d %T` shell datax config: ${shell_datax_config}"
echo "`date '+%Y-%m-%d %T` shell datax params: -p \"-Dparams1=${params1} -Dparams2=${params2}\""
/home/admin/datax3/bin/datax.py ${shell_datax_config} -p "-Dparams1=${params1} -Dparams2=${params2}"
shell_datax_run_result=$?

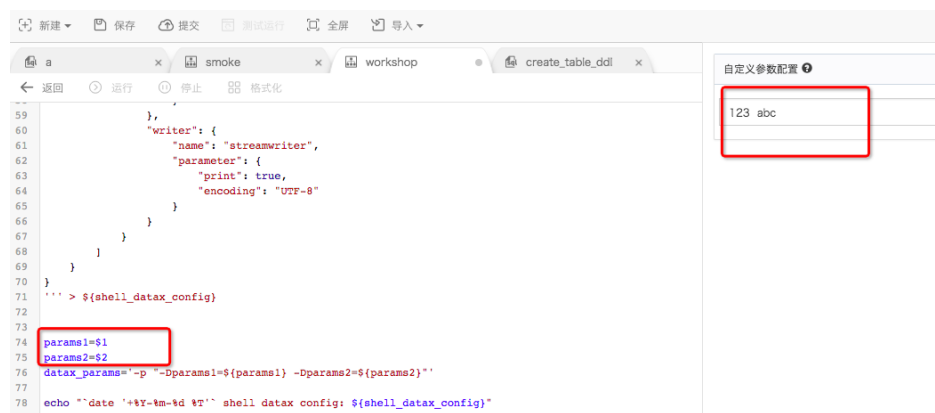
rm ${shell_datax_config}

if [${shell_datax_run_result} -ne 0]
then
echo "`date '+%Y-%m-%d %T` shell datax ended failed :("

```

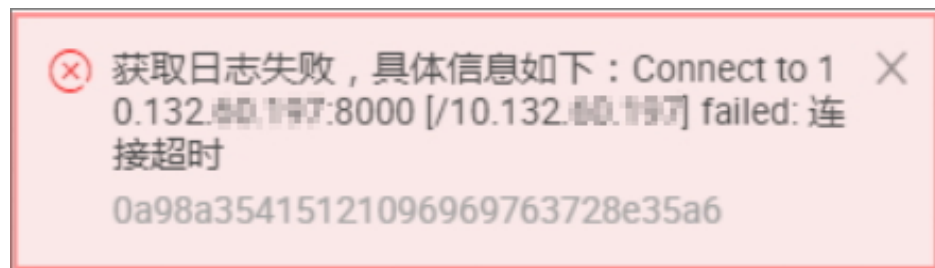
```
exit -1
fi
echo "`date '+%Y-%m-%d %T`" shell datax ended success ~ "
```

### 3. 调度系统配置参数



## 项目管理常见问题

### Q：任务执行在自定义资源组上，执行报显示连接超时



A：原因是添加自定义资源组时添加的是公网地址所以报错并获取不到相应的日志，添加自定义资源组是要添加内网IP地址，如下图：

\* 网络类型：☒ 经典网络 ☐ 专有网络 ?

\* 服务器名称： ?

\* 机器IP：

添加机器的内网IP地址

添加

取消

## Q：个账号和系统账号区别？

DataWorks sdf 数据集成 数据开发 数据管理 运维中心 项目管理

项目配置

项目成员管理

调度资源管理

MaxCompute配置

基本设置

自定义用户角色

MaxCompute基本配置

MaxCompute项目名称：dfg

MaxCompute访问身份：☐ 个人账号 ☒ 系统账号

MaxCompute Owner账号：1228973256607

**A：**系统账号权限非常高，个人账号权限比较小，跨项目的任务系统账号可以操作，然而个人账号只能操作自己所在项目的里的任务

例如：使用系统账号即使是一个子账号，都可以操作该系统账号下的所有项目的表，个人账号只能操作自己有权限的表