数据集成

用户指南

用户指南

作业配置说明

在数据集成提供的配置格式中，所有的Job配置规则都是按照统一的数据集成Job定义，Job配置的格式如下所示：

```
{
"type": "job",
"version": "1.0",
"configuration": {
"setting": {},
"reader": {
"plugin": "",
"parameter": {}
},
"writer": {
"plugin": "",
"parameter": {}
}
}
}
```

配置说明：

type=job，代表底层使用批量离线的数据同步。

version，版本号，公测版本数据集成目前仅支持1.0版本格式。

configuration={}，是DataX本身进行数据传输的描述。具体而言，描述的是{源、目的、转换器（暂不支持）}。

所有的配置信息定义，均按照驼峰命名法进行配置key命名，且大小写敏感 。用户必须保证配置必须严格按照文档定义格式进行填写。

由于JSON本身不支持注释， 您在参考数据集成提供的样例配置必须过滤注释信息 ，否则JSON解析将会报错。

Job基本配置

Job基本配置定义了一个Job基础的、框架级别的配置信息，包括：

```
{
"type": "job", //用户提交同步类型，包括Job/Stream
"version": "1.0", //版本号，公测版本CDP目前仅支持1.0版本格式。
"configuration": {
"settting": {
"key": "value"
},
"reader": {
"plugin": "mysql", //这里填写源头数据存储类型的名称
"parameter": {
"key": "value"
}
},
"writer": {
"plugin": "odps", //这里填写目标端数据存储类型的名称
"parameter": {
"key": "value"
}
}
}
}
```

其中：

type

指定本次提交同步任务是Job、Stream。

version

目前所有Job仅支持版本号”**1.0**“，您只能填写版本号为”1.0”。

Job Setting配置

Job的setting域描述的是Job配置参数中，排除源端、目的端外有关Job全局信息的配置参数，比如Job流控、Job类型转换。总体如下：

```
{
"type": "job",
"version": "1.0",
"configuration": {
"setting": {
```

```
"errorLimit": {},
"speed": {},
"column": {}
}
}
}
```

configuration.setting.column（类型转换）

数据集成支持最基本的类型转换，用户可以自定义设定类型转换配置，用以描述一些基本的字段类型转换信息，包括：

datetimeFormat：设定datetime类型和string类型的转换format。

timeFormat：设定time类型和string类型的转换format。

dateFormat：设定date类型和string类型的转换format。

encoding：设定byte类型和string类型的转换编码格式。

目前支持的类型转换格式配置如下：

```
{
"type": "job",
"version": "1.0",
"configuration": {
"setting": {
"column": {
"datetimeFormat": "yyyy-MM-dd HH:mm:ss",
"timeFormat": "HH:mm:ss",
"dateFormat": "yyyy-MM-dd",
"encoding": "utf-8"
}
}
}
}
```

configuration.setting.errorLimit（脏数据控制）

数据集成Job支持用户对于脏数据的自定义监控和告警，包括对脏数据最大记录数阈值或者脏数据占比阈值，当Job传输过程出现的脏数据大于用户指定的数量/百分比，数据集成Job报错退出。如下：

```
{
"type": "job",
"version": "1.0",
```

```
"configuration": {
"setting": {
"errorLimit": {
"record": 1024
}
}
}
}
```

上述配置用户指定了errorLimit 上限为1024条record，当Job在传输过程中出现脏数据记录数大于1024，Job报错退出。同样，用户可以指定脏数据占总体数据传输记录数的百分比，如果大于该值，Job报错退出:

configuration.setting.speed（流量控制）

数集成Job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。数据集成公测期间，最大Job流量阈值为5MB/s，数据集成将直接报错。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。通常情况下，网卡流量往往是数据集成通道流量膨胀到1至2倍左右，实际流量膨胀看具体的数据存储系统传输序列化情况。配置如下：

```
{
"type": "job",
"configuration": {
"setting": {
"speed": {
"mbps": 1 //代表1MB/s的传输带宽
}
}
}
}
```

创建向导模式任务

以开发者身份进入 阿里云数加平台>大数据开发套件>管理控制台，点击“项目列表”下对应项目操作栏中的 **进入工作区**。

点击顶部菜单栏中的 **数据集成** 中左侧导航栏的 **同步任务** 。

点击界面中的**向导模式**。

向导模式：向导模式是可视化界面配置同步任务， 一共涉及到五步，选择来源，选择目标，字段映射，通道控制，预览保存。在每个不同的数据源之间，这几步的界面可能有不同的内容，向导模式可以转换成脚本模式。
脚本模式：进入脚本界面你可以选择相应的模板，此模板包含了同步任务的主要参数，将相关的信息填写完整，但是脚本模式不能转化成向导模式。

1. 下面以 MaxCompute (原ODPS) 同步到 MySQL 为例介绍一下向导模式的五个步骤，不同数据源可能页面的展现会有所不同。

选择来源：

数据源：数据来源与数据源名保持一致，支持模糊查询请输入更多检索条件得到精确结果；
表：数据来源里的表，搜索结果只展示Top匹配到的25张表，请输入更多检索条件得到精确结果；
分区信息：MaxCompute数据源同步的表有分区则会展现分区信息，没有则显示无分区信息；
数据预览：数据浏览默认是收起的

选择目标：

数据源：目标数据源与数据源名保持一致，支持模糊查询请输入更多检索条件得到精确结果；
表：目标表，搜索结果只展示Top匹配到的25张表，请输入更多检索条件得到精确结果；
导入前准备语句：执行数据同步任务之前率先执行的 SQL 语句，目前向导模式只允许执行一条 SQL 语句，脚本模式可以支持多条SQL语句，例如清除旧数据。
导入后准备语句：执行数据同步任务之后执行的 SQL 语句，目前向导模式只允许执行一条 SQL 语句，脚本模式可以支持多条 SQL 语句，例如加上某一个时间戳。

主键冲突：选择导入模式，可以支持 insert/replace/insert ignore 方式， insert 指当主键/唯一性索引冲突，数据集成视为脏数据进行处理。 replace 指没有遇到主键/唯一性索引冲突时，与 insert 行为一致，当主键/唯一性索引冲突时会用新行替换原有行所有字段。 insert ignore 指当主键/唯一性索引冲突，数据集成将直接忽略更新丢弃，并且不记录！

字段映射：点击下一步，选择字段的映射关系。需对字段映射关系进行配置，左侧“源头表字段”和右侧“目标表字段”为一一对应的关系。

同行映射：单击同行映射能将同行的源表列和目标表列映射关系连接起来。
自动排版：调整源端表和目标表排版。
增加一行：添加源表没有的列（不同的数据源有不同的规范，可以参考“增加一行”按钮后面的提示）。

通道控制：

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、数据库配置等的影响。
作业并发数:作业速率上限=作业并发数*单并发的传输速率 当作业速率上限已选定的情况下，应该如何选择作业并发数？

① 如果你的数据源是线上的业务库，建议您不要将并发数设置过大，以防对线上库造成影响；
② 如果您对数据同步速率特别在意，建议您选择最大作业速率上限和较大的作业并发数；

预览保存:展现上面几步配置的信息，这边可以修改相关的配置信息，确认信息无误点击**保存**。

其他向导模式配置同步任务请参考下面文档：

- MySQL通过数据集成导入导出
- SQL Server通过数据集成导入/导出
- PostgreSQL通过数据集成导入/导出
- Oracle通过数据集成导入/导出
- DRDS通过数据集成导入/导出
- HybridDB for MySQL通过数据集成导入/导出
- HybridDB for PostgreSQL通过数据集成导入/导出
- MaxCompute通过数据集成导入/导出
- ADS通过数据集成导入数据
- OSS通过数据集成导入/导出
- Hdfs通过数据集成导入/导出
- SFTP通过数据集成导入/导出
- MongDB通过数据集成导入/导出
- TableStore在数据集成里数据导入/导出
- Redis通过数据集成导入数据

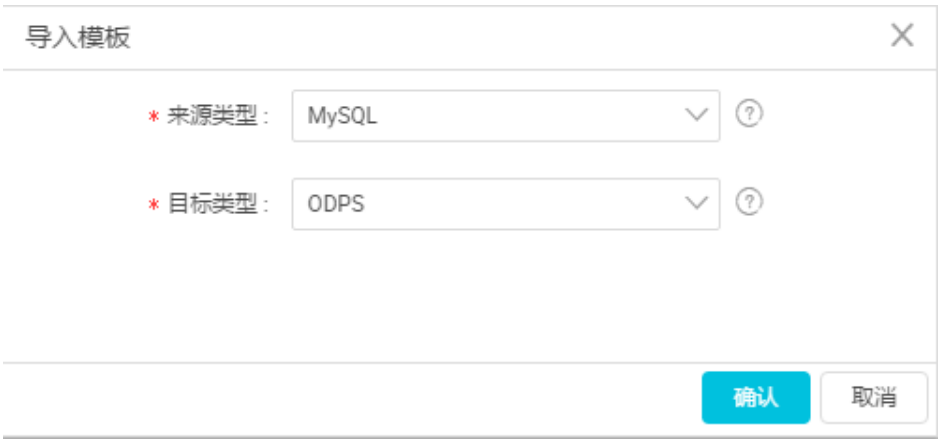
创建脚本模式任务

以开发者身份进入 阿里云数加平台>大数据开发套件>管理控制台，点击“项目列表”下对应项目操作栏中的 **进入工作区**；

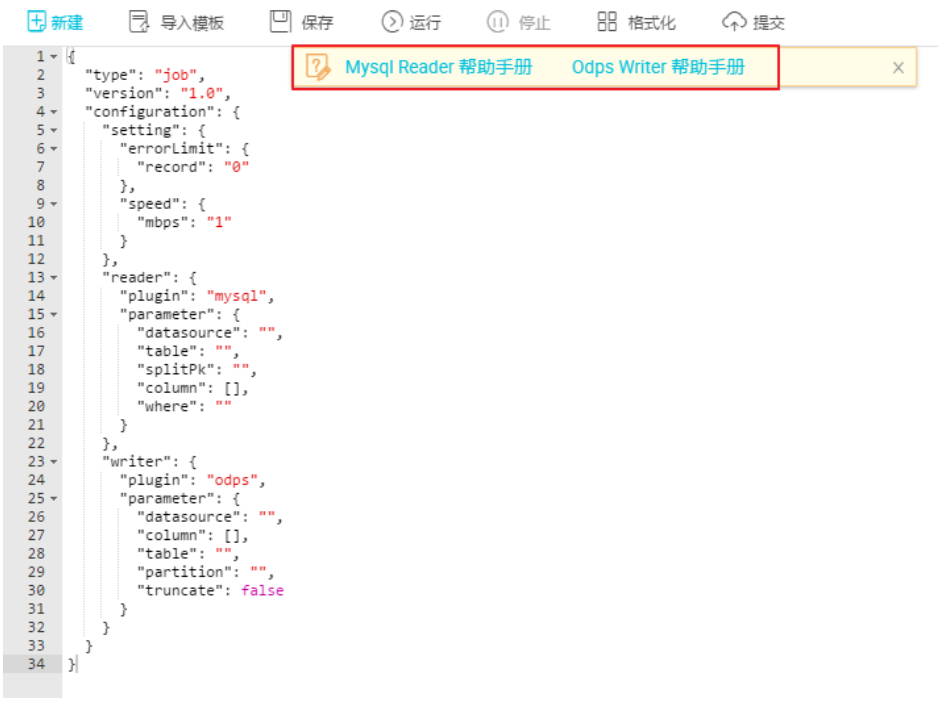
点击顶部菜单栏中的 **数据集成** 中左侧导航栏的 **同步任务** ；

点击界面中的**脚本模式**；

在弹出的“导入模板”中选择自己需要的“来源类型”和“目标类型”，如下图所示：



点击确认后即进入脚本模式配置页面，可根据自身情况进行配置（详情见下文），如有问题可点击右上方的帮助手册进行查看，如下图所示：



完成后点击“保存” 。

备注：若想选择新模板，可点击工具栏中的“导入模板” ，但一旦导入新模板，原有内容将会被全部覆盖；同时您也可在建好的向导模式中点击工具栏中的“转换脚本” ，将其转换为脚本模式 。

脚本模式基本配置

数据集成 JSON 框架级别的配置信息，包括：

```
{
"type": "job",
"version": "1.0",
"configuration": {
"settting": {
"key": "value"
},
"reader": {
"plugin": "填写源头数据存储类型的名称",
"parameter": {
"key": "value"
}
},
"writer": {
"plugin": "填写目标端数据存储类型的名称",
"parameter": {
"key": "value"
}
```

```
}
}
}
```

其中：

• **type**

指定本次提交的同步任务，仅支持 Job 参数。用户只能填写为 “Job”。

• **version**

目前所有 Job 仅支持版本号 “1.0”，用户只能填写版本号为 “1.0”。

系统调优配置

Job 的 setting 域描述的是 Job 配置参数中除源端、目的端外有关 Job 全局信息的配置参数，比如 Job 流控、Job 类型转换 。总体如下：

```
{
"type": "job",
"version": "1.0",
"configuration": {
"setting": {
"errorLimit": {},
"speed": {}
}
}
}
```

• **configuration.setting.errorLimit (脏数据控制)**

支持用户对于脏数据的自定义监控和告警，包括对脏数据最大记录数阈值，当 Job 传输过程出现的脏数据大于用户指定的数量，则报错退出 。如下所示：

```
{
"type": "job",
"version": "1.0",
"configuration": {
"setting": {
"errorLimit": {
"record": 1024
}
}
}
}
```

上述配置中用户指定了 errorLimit 上限为 1024 条 record，当 Job 在传输过程中出现脏数据记录数大于 1024，则 Job 报错退出 。

• **configuration.setting.speed (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 。

配置如下，代表 1MB/s 的传输带宽：

```
{
"type": "job",
"configuration": {
"setting": {
"speed": {
"mbps": 1
}
}
}
}
```

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 。通常情况下，网卡流量往往是通道流量膨胀到 1 至 2 倍左右，实际流量膨胀看具体的数据存储系统传输序列化情况 。

半结构化的单个文件没有切分键的概念，多个文件可以设置“作业速率上限”来提高同步的速度，但“作业速率上限”跟文件的个数有关，比如有 n 个文件，“作业速率上限”设置最多设置为 n MB/s，如果设置 n+1 MB/s 还是以 n MB/s 速度同步，如果设置为 n-1 MB/s，则以 n-1 MB/s 速度同步 。

关系型数据库设置“作业速率上限”和“切分键”才能根据“作业速率上限”将表进行切分，关系型数据库只支持数值型作为切分键，但 oracle 数据库是支持数值型和字符串类型的作为切分键 。

其他脚本模式详细配置信息请参考下面文档：

Reader插件配置

- MySQL Reader 配置
- SqlServer Reader配置
- PostgreSql Reader配置
- MaxCompute Reader 配置
- Drds Reader 配置
- Oracle Reader 配置
- OSS Reader 配置
- FTP Reader配置
- OTS Reader配置
- Stream Reader配置
- DB2 Reader配置
- Hdfs Reader 配置
- MongoDB Reader配置
- HbaseReader 配置

- OTSReader-Internal 配置

Writer插件配置

- MySQL Writer配置
- SqlServer Writer配置
- PostgreSql Writer配置
- MaxCompute Writer 配置
- Drds Writer配置
- Oracle Writer配置
- OSS Writer 配置
- FTP Writer配置
- OTS Writer配置
- Stream Writer配置
- DB2 Writer配置
- Hdfs Writer 配置
- MongoDB Writer配置
- HbaseWriter 配置
- AnalyticDB Writer 配置
- OCS Writer配置
- LogHub Writer配置
- OpenSearch Writer配置
- Redis Writer配置
- OTSWriter-Internal 配置

如何设置同步任务的通道控制参数

数据集成同步任务的通道控制，主要是控制整个同步任务的同步速率和并发数，各概念含义解释如下：

作业速率上限

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、数据库配置等的影响。

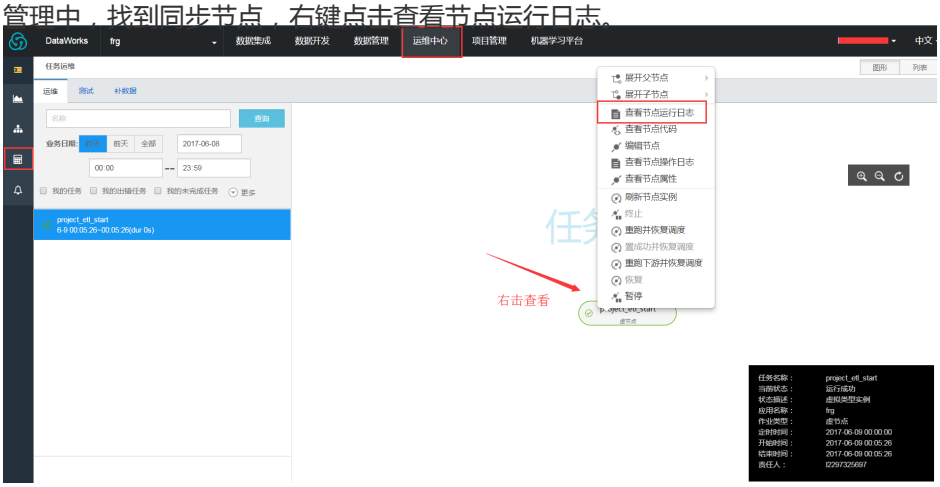
作业并发数

从单同步作业来看：作业并发数*单并发的传输速率=作业传输总速率；

当作业速率上限已选定的情况下，应该如何选择作业并发数？
① 如果你的数据源是线上的业务库，建议您不要将并发数设置过大，以防对线上库造成影响；
② 如果您对数据同步速率特别在意，建议您选择最大作业速率上限和较大的作业并发数

错误记录数

表示脏数据的最大容忍条数，如果您配置0，则表示严格不允许脏数据存在；如果不填则代表容忍脏数据，即如果出现脏数据，数据集成会记录并打印部分脏数据，方便用户排查。查看脏数据日志的方法：在运维中心-任务管理中，找到同步节点，右键点击查看节点运行日志。



运维中心

任务列表

周期任务是指：在数据开发模块设置任务调度类型为周期调度，提交后，调度系统按调度配置自动定时执行的任务，如下图所示：

默认展示规则：1、展示工作流任务2、展示当前登录账号名下的任务

> 注：任务提交后，将会在第二天23:30自动生成实例来运行任务，若是在23:30以后提交的任务，那么第三天才会开始生成实例来自动运行任务。

危险操作

请勿操作**project_etl_start**节点，此节点为项目根节点，周期任务的实例均依赖于此节点，若将此节点冻结，那么周期任务实例将不会运行。

周期任务列表

展示已提交的任务信息，可以对这些任务进行测试运行、补数据、添加报警、修改责任人、修改资源组、冻结/解冻等操作，具体操作页面如下：

筛选功能：如上图①部分，筛选条件过滤出要查询的任务；根据任务类型、任务名称、责任人、今日修改的任务、冻结状态等条件精确筛选。

- 任务类型在筛选时有三种，分别是工作流、节点任务、内部节点(内部节点是指工作流内的节点任务)。

> 注意：任务名搜索的结果，会受到其他筛选条件的影响，只有同时满足所有筛选条件的结果才会展示出来。

操作：如上图②部分，您可以对任务进行测试、补数据、冻结、解冻、查看实例、添加报警、修改责任人等操作。

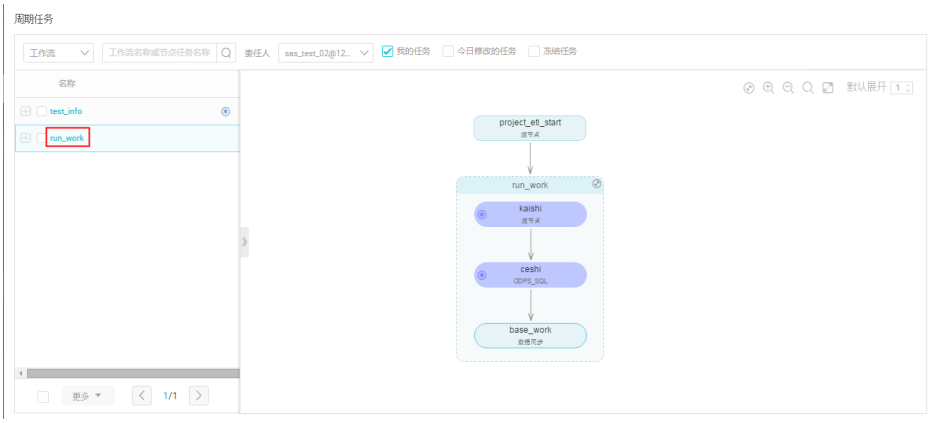
> 冻结：冻结状态的任务，生成的实例也是冻结状态，不会直接运行，必须将实例解冻以后再点击**重跑**，才会运行。

批量操作：如上图③部分，您可以批量选择任务，进行添加报警、修改责任人、修改资源组、冻结、解冻等操作。

> 注意：工作流任务无法修改资源组；冻结的任务生成的测试/补数据实例也会是冻结状态，只有解冻以后任务才会开始运行。

任务DAG图

若想看某个任务的详细信息，则可直接在任务列表中点击任务名/工作流名，会弹出一个任务信息的DAG图，如下图所示：



在DAG图中可以选择具体的任务/工作流，来查看任务属性和操作日志，也可以对该任务的属性进行修改。

注：工作流属性和内部节点属性是不同的，内部节点是可以查看代码内容的。

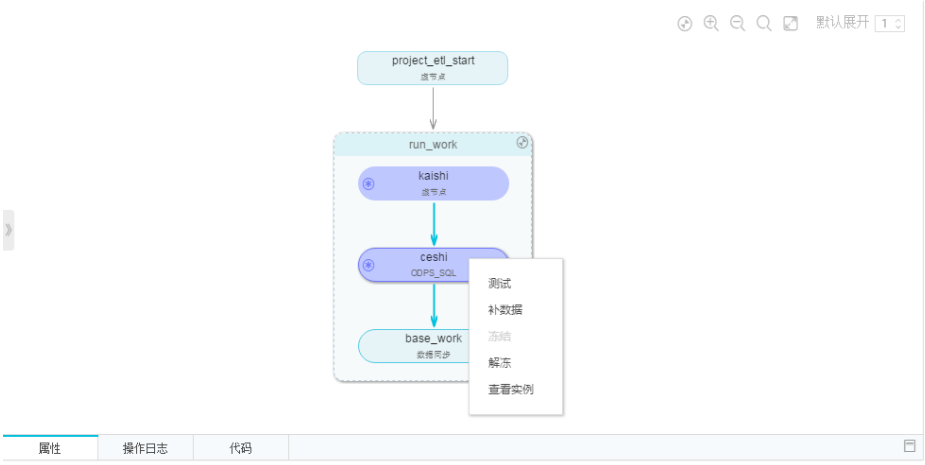
- 打开工作流后的内容如下：

工作流任务只能修改责任人和设置监控，无法修改资源组；可以查看操作日志，操作日志会记录您对该任务进行过的操作，比如说更新工作流内容、冻结实例调度、补数据、冒烟测试等等。

- 打开工作流中的内部节点后，内容如下：

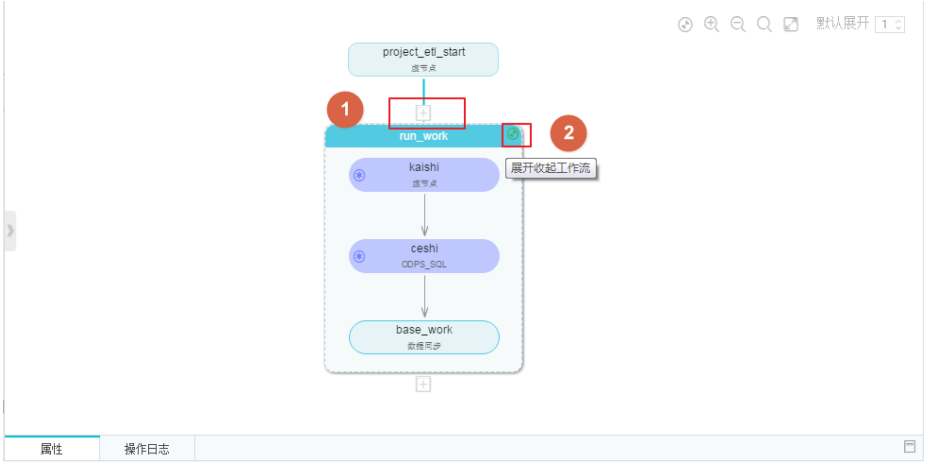
工作流中的内部节点可以修改责任人、修改资源组和设置监控；可以查看操作日志，操作日志会记录您对该任务进行过的操作，比如说更新工作流内容、冻结实例调度、补数据、冒烟测试等等；还可以查看节点代码，但是不能修改。

- 右键单击工作流/内部节点，内容如下:



右键单击工作流/内部节点都会出现上图所示的操作框，可以选择测试、补数据、冻结、解冻、查看实例等操作。

- 查看工作流/节点上下游依赖，以及展开/关闭工作流，内容如下：



如图①所示，您可以展开上下游依赖，如图②所示，您单击以后，可以展开/关闭工作流。

手动任务是指：新建任务时，调度类型选择**手动任务**后，提交到调度系统的任务。提交到调度系统后，手动任务不会自动运行，只有手动触发才会运行，如下图所示：

默认展示规则：1、优先展示工作流任务2、默认展示当前登录账号名下的任务

> 注：手动任务不会自动运行，只能手动触发。

手动任务列表

以列表的形式，展示已提交的手动任务，可以在手动任务列表中，单击操作栏中的运行按钮，来触发手动任务实例，具体操作页面如下：

筛选功能：如上图①部分，我们提供了丰富的筛选条件，来帮助您过滤出您想要的任务；您可以根据任务类型、任务名称、责任人、今日修改的任务等条件来进行精确筛选。

> 注意：任务名搜索的结果，会受到其他筛选条件的影响，只有同时满足所有筛选条件的结果才会展示出来。

操作：如上图②部分，您可以对任务进行运行、查看实例、修改责任人等操作。

> 查看实例：会跳转到手动实例列表中去，将该任务所有的实例都展示出来。

批量操作：如上图③部分，您可以批量选择任务来修改责任人。

注：手动任务无法选择资源组运行。

任务DAG图

单击手动任务的任务名，会弹出该任务信息的DAG图，如下图所示：

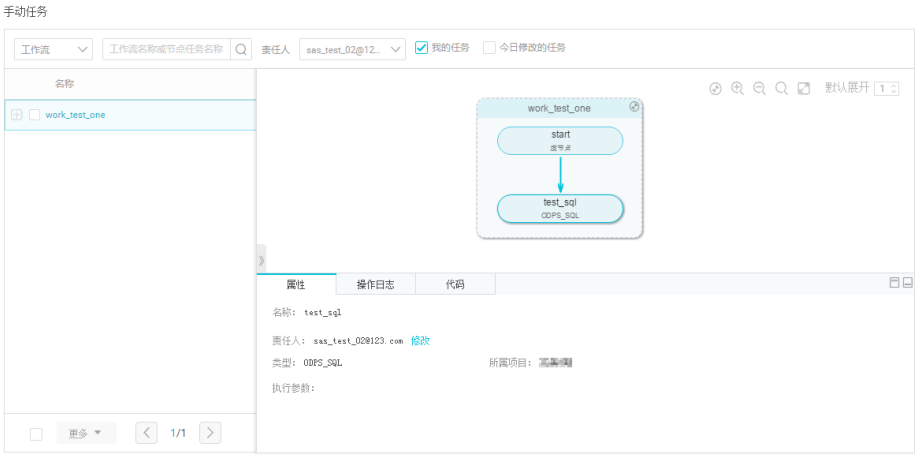
在DAG图中可以选择任务/工作流，来查看任务属性和操作日志，也可以对该任务的属性进行修改，但手动任务能修改的属性只有任务责任人。

注：工作流属性和内部节点属性是不同的，内部节点是可以查看代码内容的。

- 打开工作流后的内容如下：

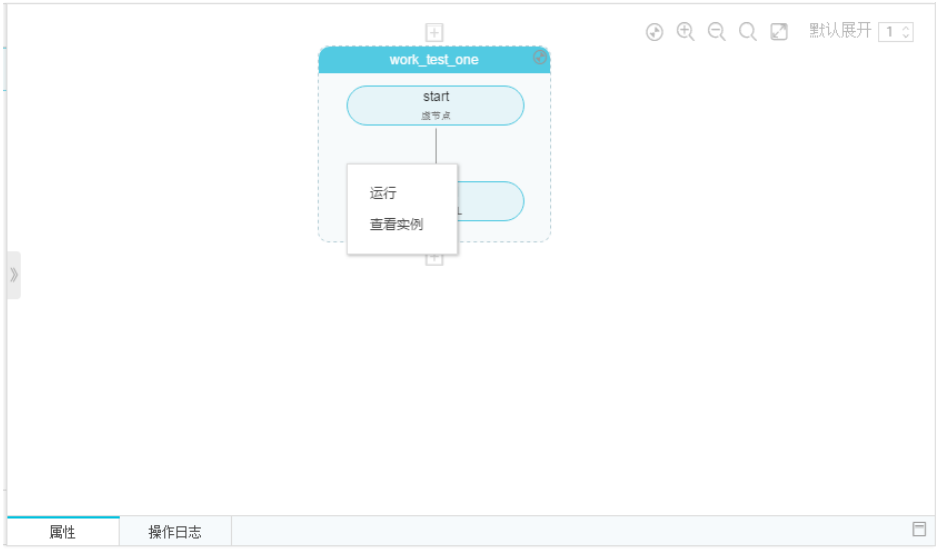
工作流任务只能修改责任人和查看操作日志，操作日志是记录您对该任务进行过的操作，比如说：更新工作流内容，重跑任务实例，运行任务等。

- 打开工作流中的内部节点后，内容如下：



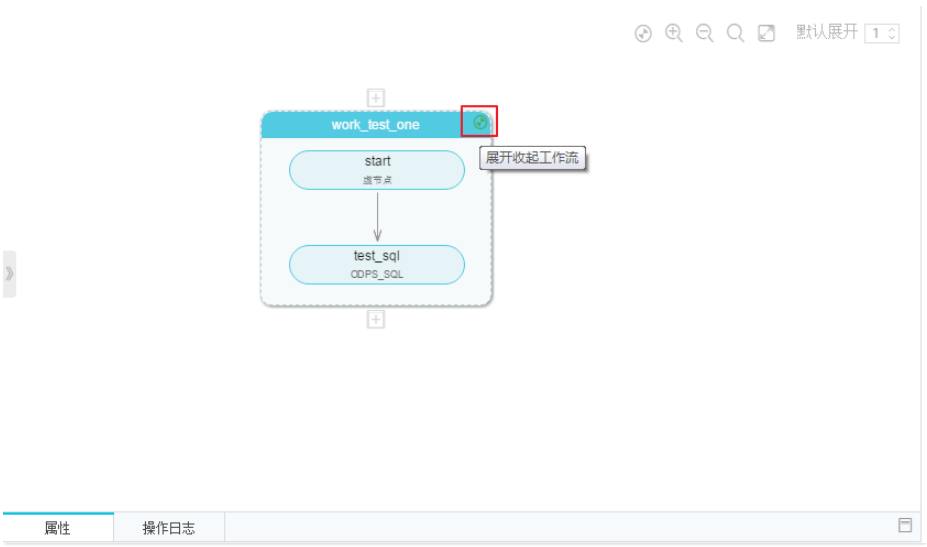
工作流中的内部节点可以修改责任人；可以查看操作日志，操作日志会记录您对该任务进行过的操作，比如说：更新工作流内容，重跑任务实例，运行任务等等；还可以查看节点代码，但是不能修改。

- 右键单击工作流/内部节点，内容如下：



右键单击工作流/内部节点都会出现上图所示的操作框，可以选择运行、查看实例等操作。

- 展开/关闭工作流，内容如下：



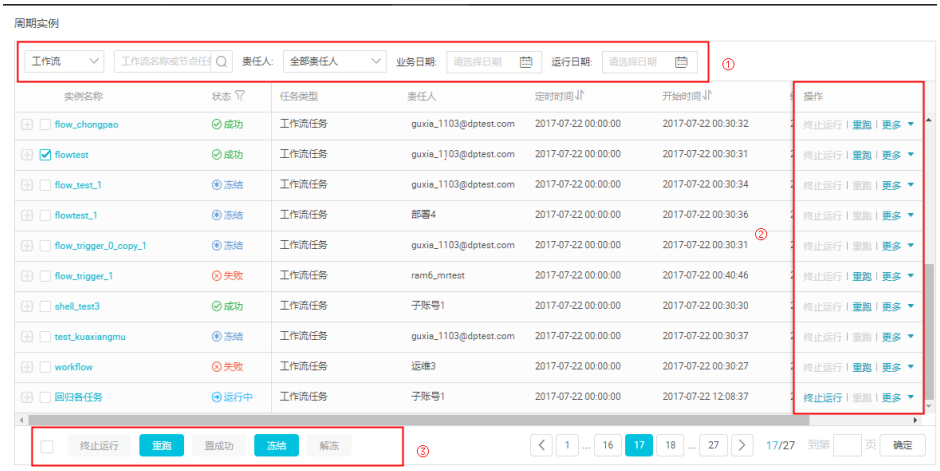
单击红色框框中的按钮，可以展开/关闭工作流。

任务运维

周期实例是指当周期任务达到启用调度所配置的周期性运行时间时被自动调度起来的实例快照，每调度一次，则生成一个实例工作流，可对已调度起的实例任务进行日常的运维管理，如对查看运行状态，对任务进行终止、重跑、解冻等操作。

实例列表

以列表形式对被调度起来的任务运维及管理。包括检查运行日志、重跑任务、终止正在运行的任务等，具体功能如下：



筛选功能：如上图中①部分，有丰富的筛选条件，默认筛选条件为业务日期是当前时间前一天的工作流任务

，用户可添加任务名称、运行时间、责任人等条件进行更精确的筛选。

终止运行：只可对等待运行、运行中状态的实例进行终止运行操作，进行此操作后，该实例将为失败状态。

重跑：可以重跑某任务，任务执行成功后可以触发下游未运行状态任务的调度。常用于处理出错节点和漏跑节点。

> 前置条件：只能重跑未运行、成功、失败状态的任务。

重跑下游：可以重跑某任务及其下游任务，需要用户自定义勾选，勾选的任务将被重跑，任务执行成功后可以触发下游未运行状态任务的调度。常用于处理数据修复。

> 前置条件：只能勾选未运行、完成、失败状态的任务，如果勾选了其他状态的任务，页面会提示“已选节点中包含不符合运行条件的节点”，并禁止提交运行。

置成功：将当前节点状态改为成功，并运行下游未运行状态的任务。常用于处理出错节点。

> 前置条件：只能失败状态的任务能被置成功。

冻结：冻结状态的任务会生成实例，但是不会运行；若需要运行冻结的实例，可以解冻实例，单击重跑，实例才会开始运行。

解冻：可以将冻结状态的实例解冻；若该实例还未运行，则上游任务运行完毕后，会自动运行；若上游任务都运行完毕，则该任务会直接被置为失败，需要手动重跑后，实例才会正常运行。

批量操作：如上图中③部分，批量操作包括，终止运行、重跑、置成功、冻结、解冻5个功能。

实例dag图

在实例dag视图中，右键单击实例，可以查看该实例的依赖关系和详细信息并进行终止运行、重跑等具体操作。如下图：

> 刷新节点实例：若在实例生成后，有修改了代码或调度参数的话，可以点击此按钮，来使用最新的代码及

参数（暂不支持批量操作）。

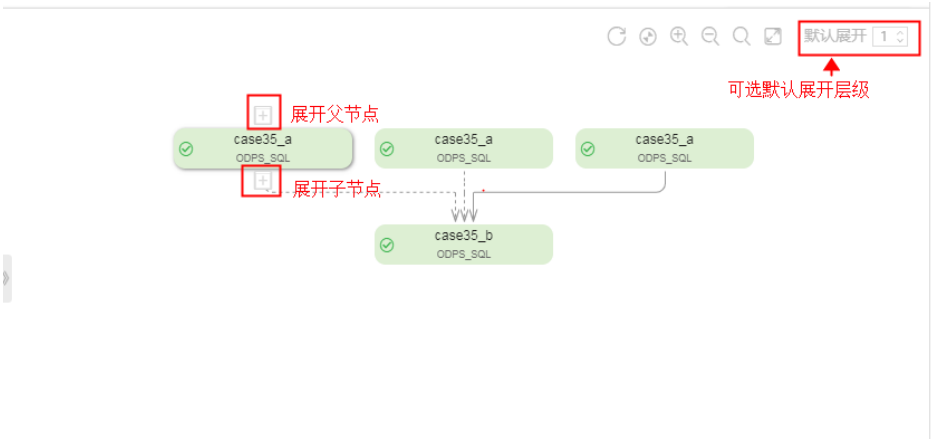
属性：查看实例属性，包括实例运行的各种时间信息、运行状态等。

运行日志：节点正在运行、成功、失败等状态时查看任务运行的日志。

操作日志：记录用户给该实例的进行的操作，例如终止运行，重跑等。

代码：可查看该实例任务的代码。

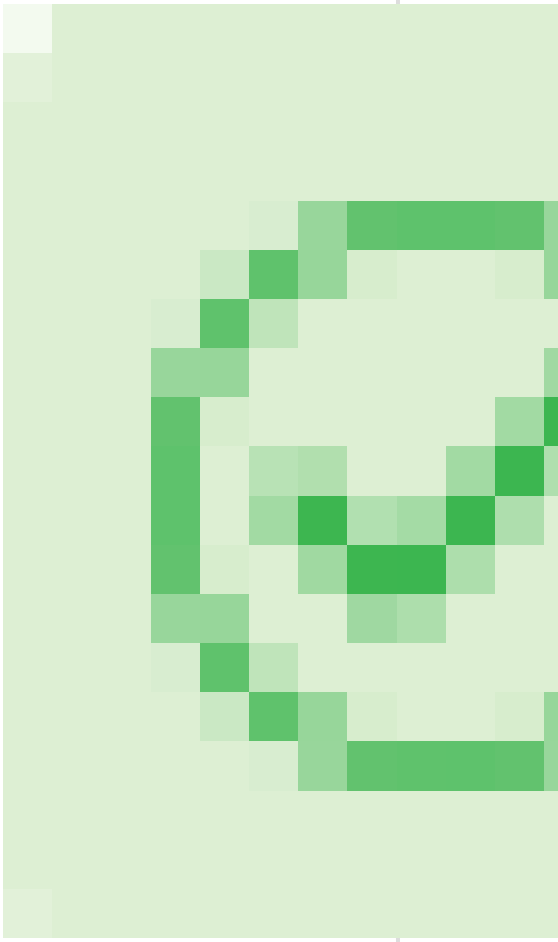
展开父节点/子节点：当一个工作流有3个节点及以上时，运维中心展示任务时会自动隐藏节点，用户可通过展开父子层级，来看到全部节点的内容。如下图所示：



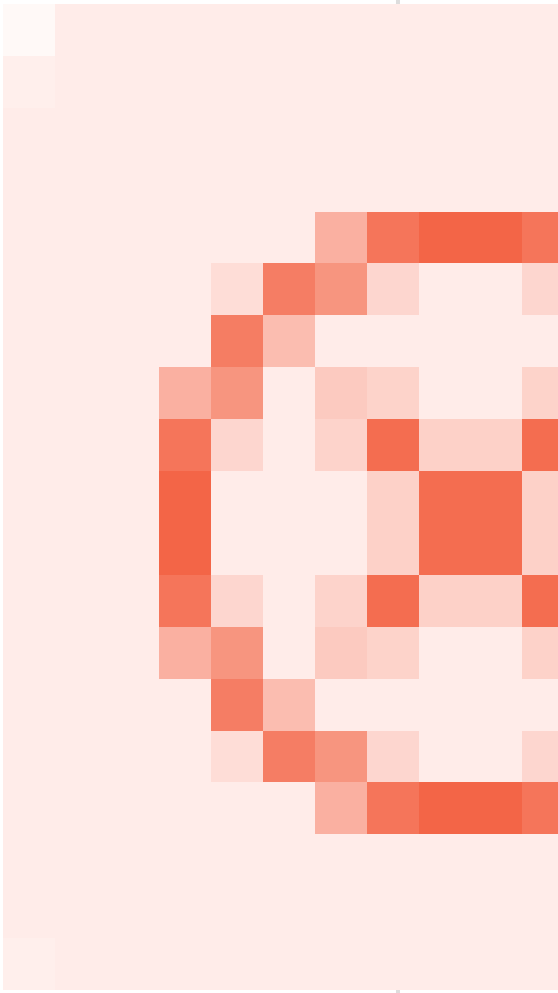
展开/关闭工作流：有工作流任务时，可以展开工作流任务，查看内部节点任务的运行状态。如下图：

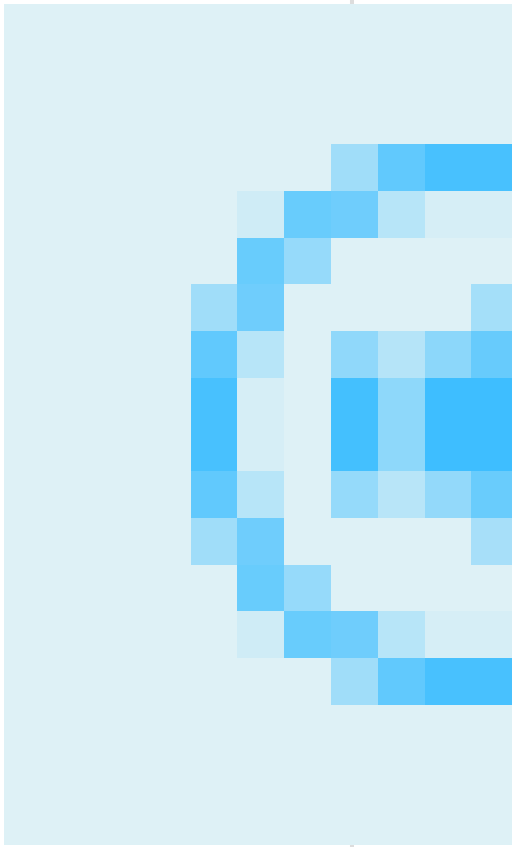
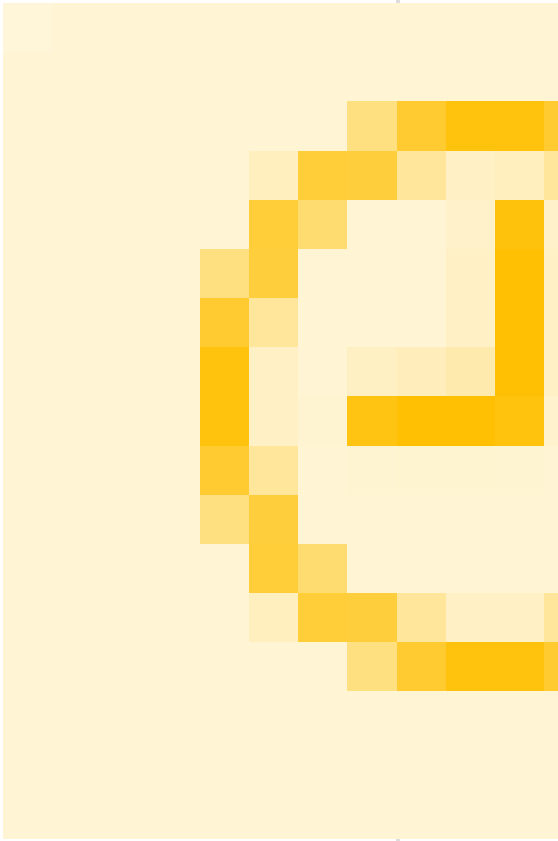
实例状态说明

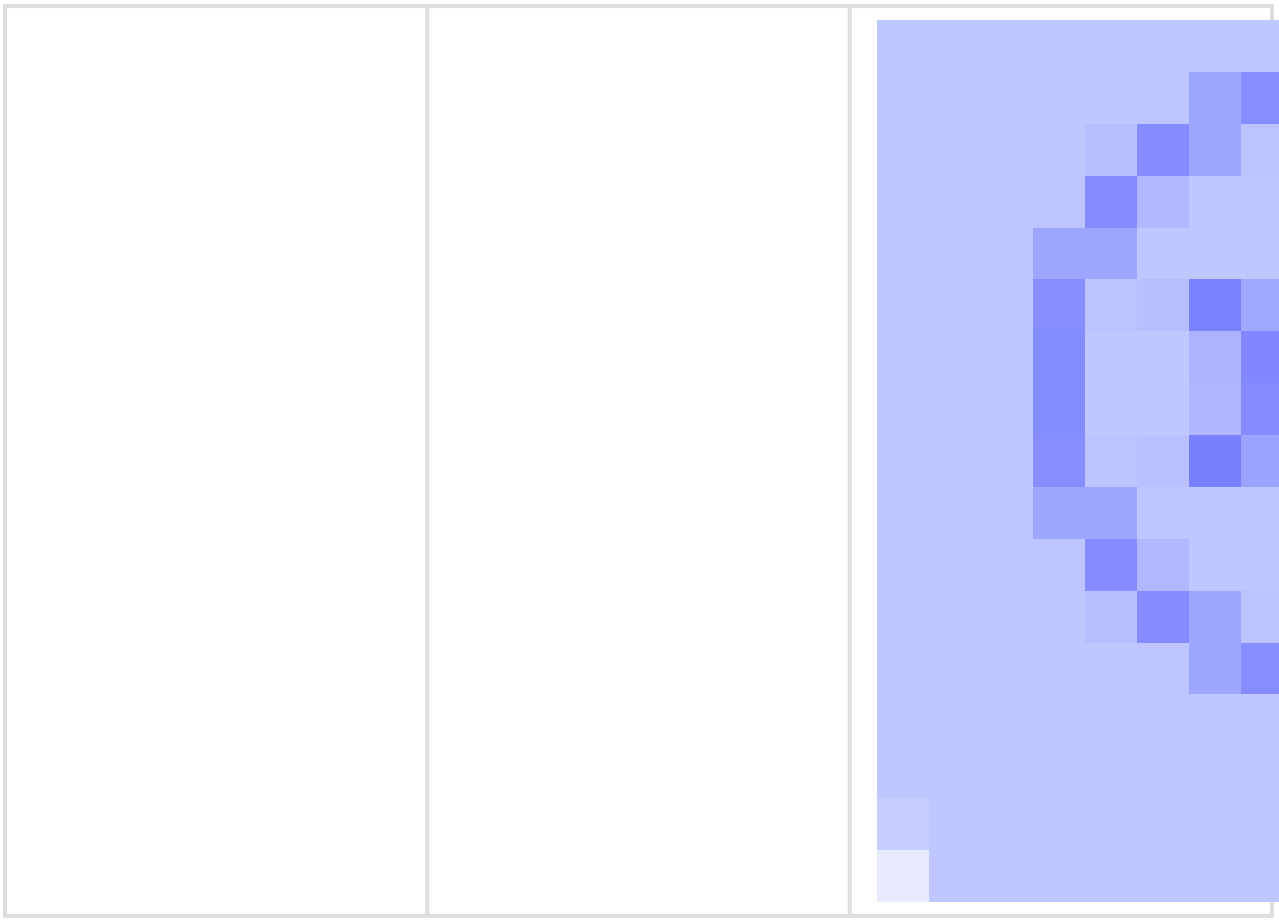
序号	状态类型	状态标识
1	运行成功状态	

		
2	未运行状态	

		
3	运行失败状态	

		
4	正在运行状态	

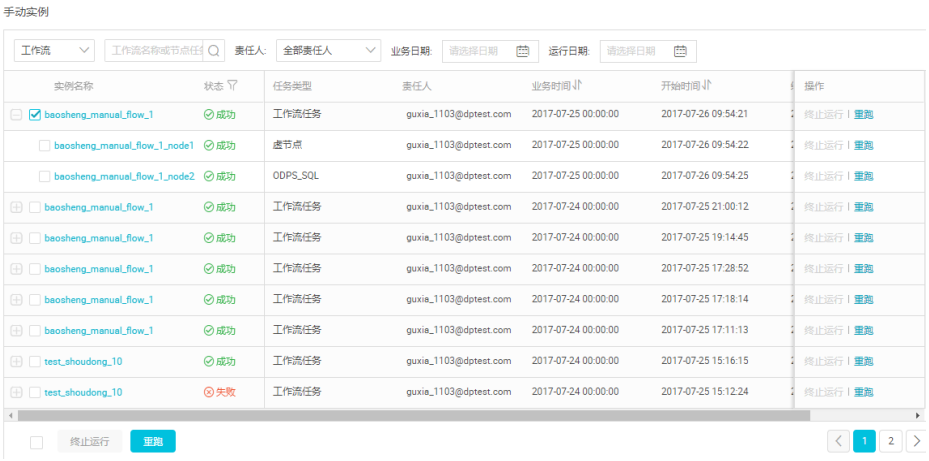
		
5	等待状态	
6	冻结状态	



手动实例是指当手动任务触发运行后产生的实例，可对已调度起的实例任务进行运维管理，如对查看运行状态，对任务进行终止、重跑等操作。

实例列表

以列表形式对手动起调的任务运维及管理。包括检查运行日志、重跑任务、终止正在运行的任务等，具体功能如下：



筛选功能：有丰富的筛选条件，默认筛选条件为业务日期是当前时间前一天的工作流任务，用户可添加任务名称、运行时间、责任人等条件进行更精确的筛选。**终止运行**：只可对等待运行、运行中状态的实例进行终止运行操作，进行此操作后，该实例将为失败状态。**重跑**：可以重跑某任务，任务执行成功后可以触发下游未运行状态任务的调度。常用于处理出错节点和漏跑节点。

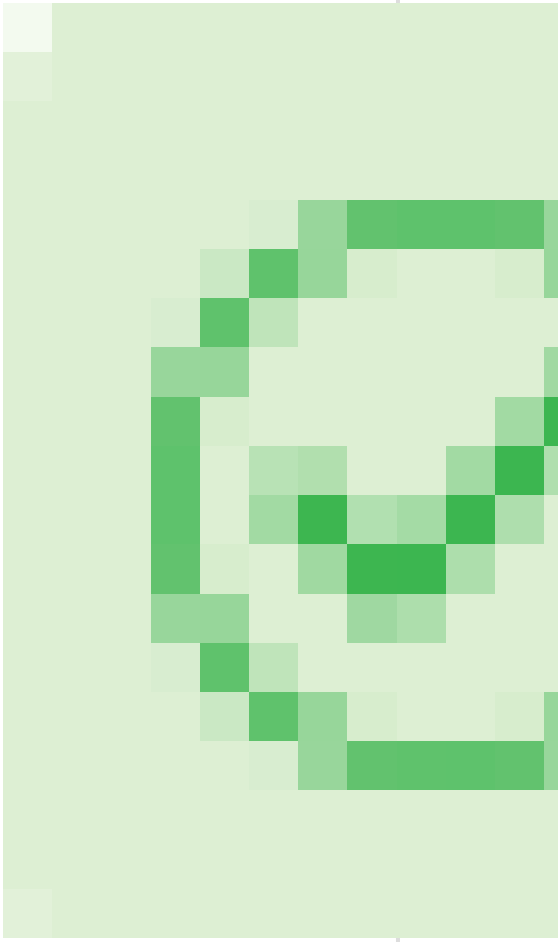
前置条件：只能重跑未运行、成功、失败状态的任务。

实例dag图

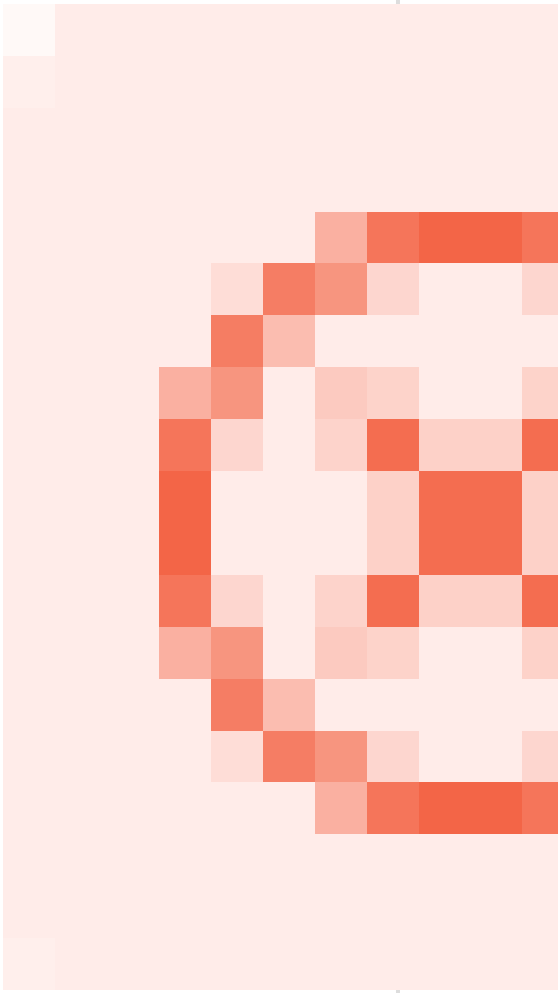
在实例dag视图中，可以查看该实例的详细信息并进行终止运行、重跑等具体操作。如下图：

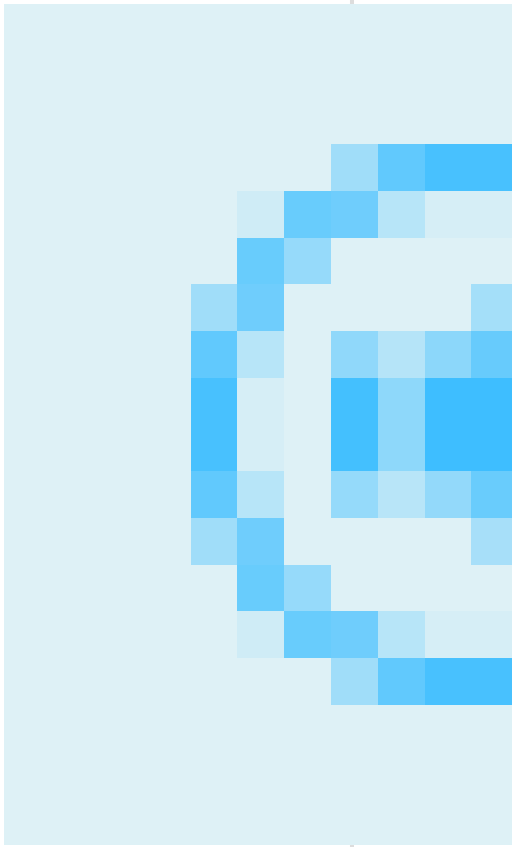
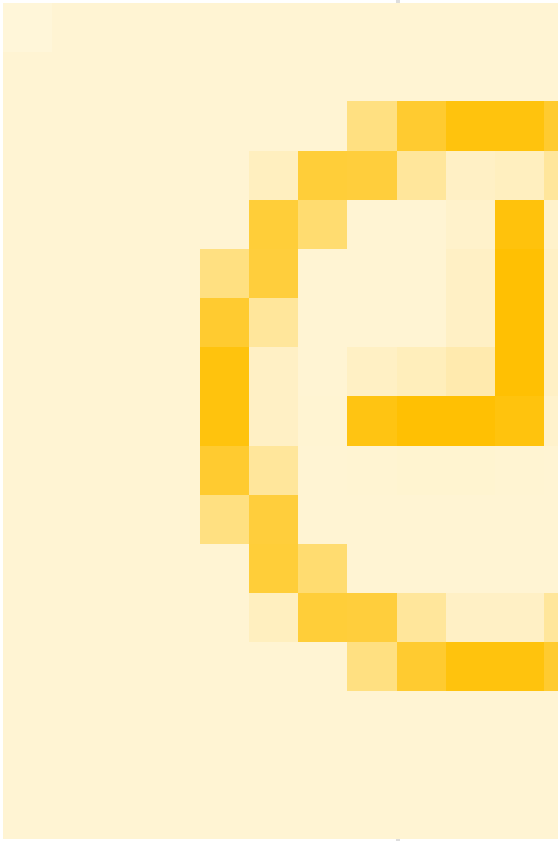
属性：查看实例属性，包括实例运行的各种时间信息、运行状态等。**运行日志**：节点正在运行、成功、失败等状态时查看任务运行的日志。**操作日志**：记录用户给该实例的进行的操作，例如终止运行，重跑等。**代码**：可查看该实例任务的代码。**展开父节点/子节点**：当一个工作流有3个节点及以上时，运维中心展示任务时会自动隐藏节点，用户可通过展开父子层级，来看到全部节点的内容。**展开/关闭工作流**：有工作流任务时，可以展开工作流任务，查看内部节点任务的运行状态。如下图：

实例状态说明

序号	状态类型	状态标识
1	运行成功状态	
2	未运行状态	

		
3	运行失败状态	

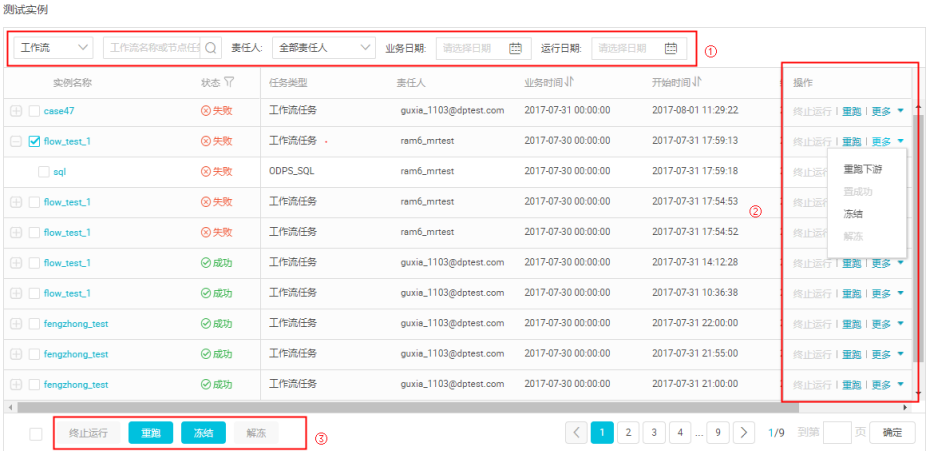
		
4	正在运行状态	

		
5	等待状态	
6	冻结状态	

测试实例是指对周期任务测试运行时产生的实例，可对已调度起的实例任务进行运维管理，如查看运行状态，对任务进行终止、重跑、解冻等操作。

实例列表

以列表形式对测试任务的运维及管理。包括检查运行日志、重跑任务、终止正在运行的任务等，具体功能如下：



筛选功能：如上图中①部分，有丰富的筛选条件，默认筛选条件为业务日期是当前时间前一天的工作流任务，用户可添加任务名称、运行时间、责任人等条件进行更精确的筛选。

终止运行：只可对等待运行、运行中状态的实例进行终止运行操作，进行此操作后，该实例将为失败状态。

重跑：可以重跑某任务，任务执行成功后可以触发下游未运行状态任务的调度。常用于处理出错节点和漏跑节点。

> 前置条件：只能重跑未运行、成功、失败状态的任务。

重跑下游：可以重跑某任务及其下游任务，需要用户自定义勾选，勾选的任务将被重跑，任务执行成功后可以触发下游未运行状态任务的调度。常用于处理数据修复。

> 前置条件：只能勾选未运行、完成、失败状态的任务，如果勾选了其他状态的任务，页面会提示“已选节点中包含不符合运行条件的节点”，并禁止提交运行。

置成功：将当前节点状态改为成功，并运行下游未运行状态的任务。常用于处理出错节点。

> 前置条件：只能失败状态的任务能被置成功。

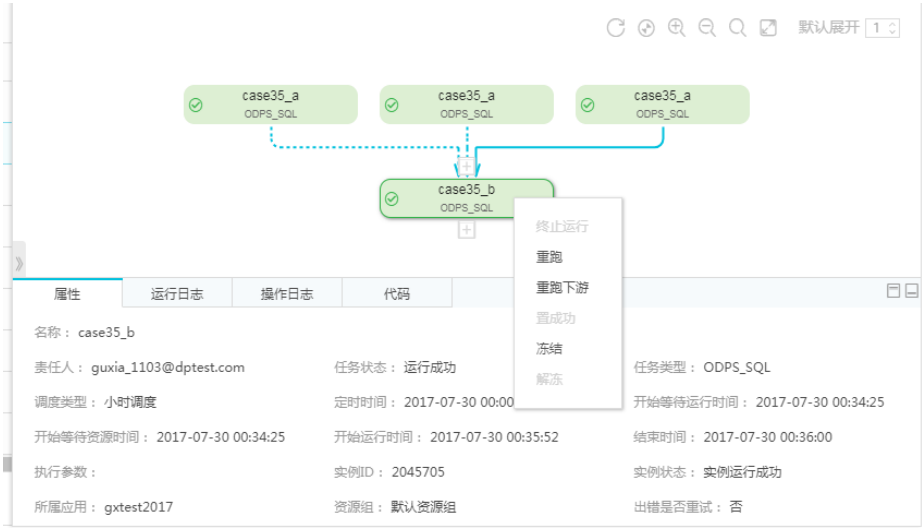
冻结：冻结状态的任务会生成实例，但是不会运行；若需要运行冻结的实例，可以解冻实例，单击**重跑**，实例才会开始运行。

解冻：可以将冻结状态的实例解冻；若该实例还未运行，则上游任务运行完毕后，会自动运行；若上游任务都运行完毕，则该任务会直接被置为失败，需要手动重跑后，实例才会正常运行。

批量操作：如上图中③部分，批量操作包括，终止运行、重跑、置成功、冻结、解冻5个功能。

实例dag图

在实例dag视图中，可以查看该实例的详细信息并进行终止运行、重跑等具体操作。如下图：



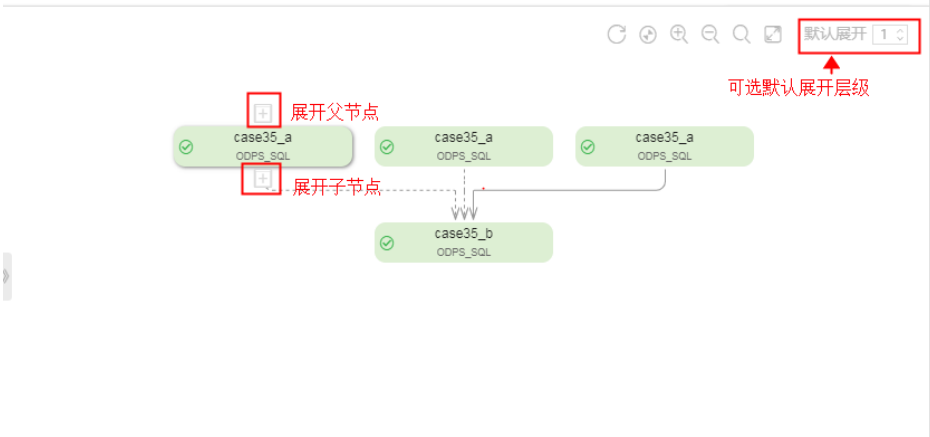
属性：查看实例属性，包括实例运行的各种时间信息、运行状态等。

运行日志：节点正在运行、成功、失败等状态时查看任务运行的日志。

操作日志：记录用户给该实例的进行的操作，例如终止运行，重跑等。

代码：可查看该实例任务的代码。

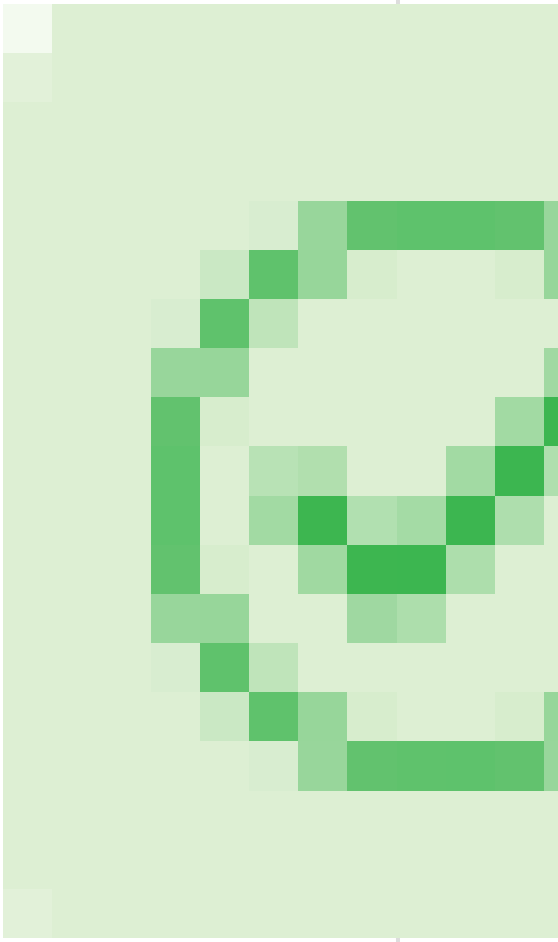
展开父节点/子节点：当一个工作流有3个节点及以上时，运维中心展示任务时会自动隐藏节点，用户可通过展开父子层级，来看到全部节点的内容。如下图所示：



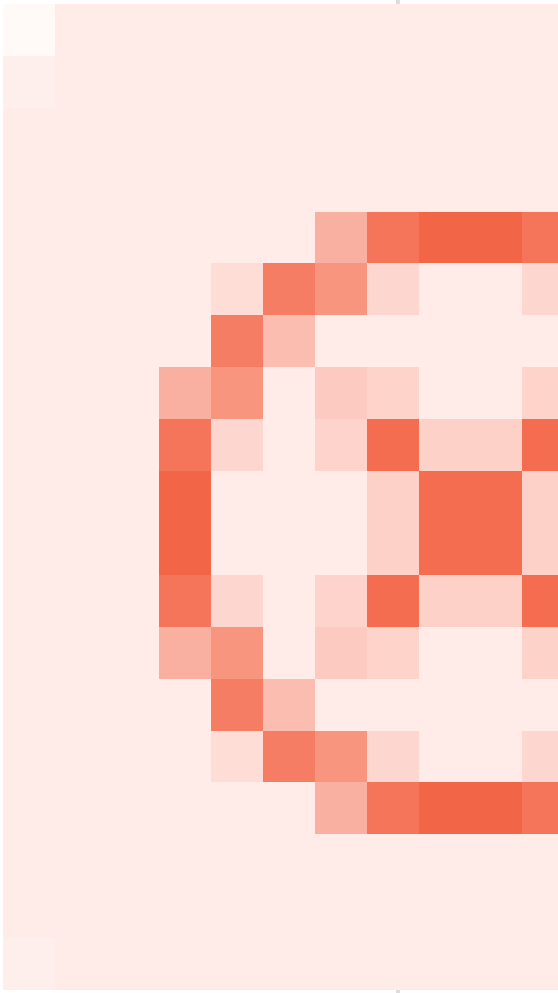
展开/关闭工作流：有工作流任务时，可以展开工作流任务，查看内部节点任务的运行状态。如下图：

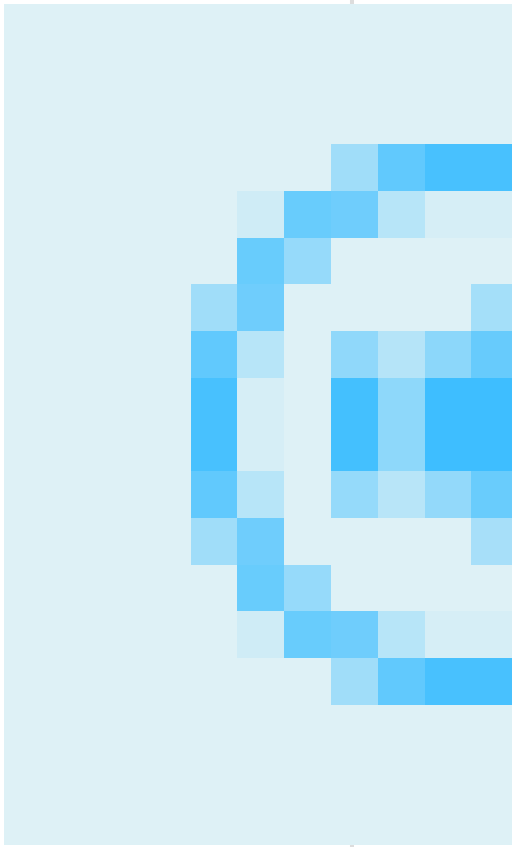
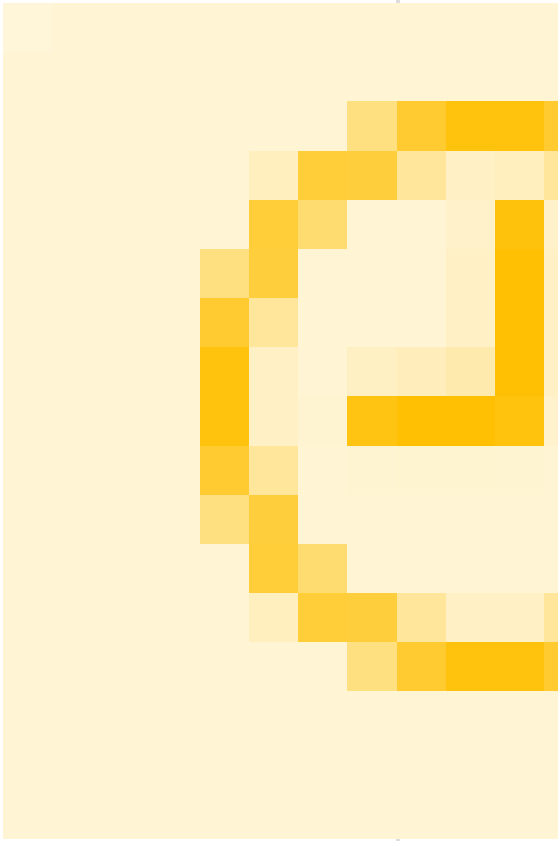
实例状态说明

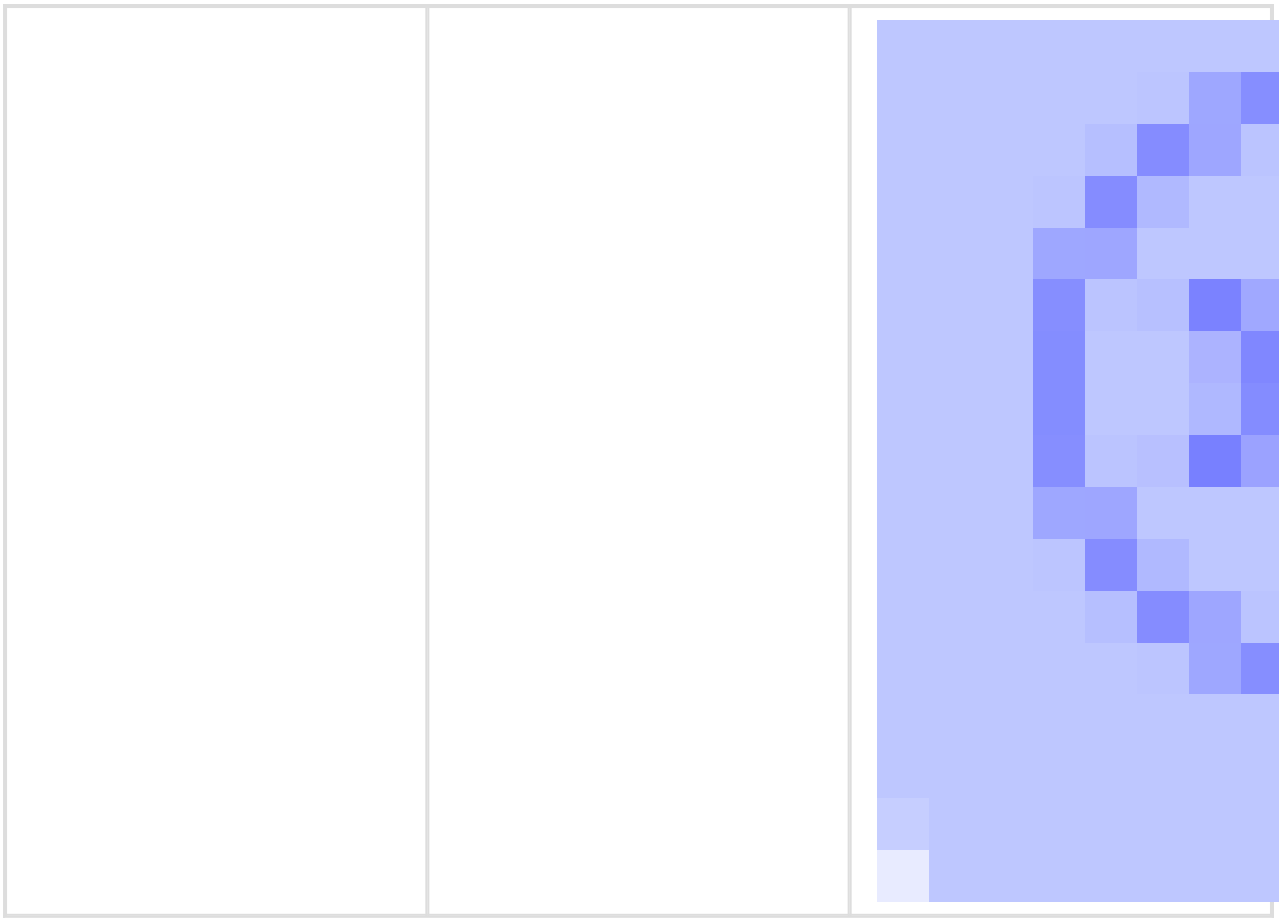
序号	状态类型	状态标识
1	运行成功状态	

		
2	未运行状态	

		
3	运行失败状态	

		
4	正在运行状态	

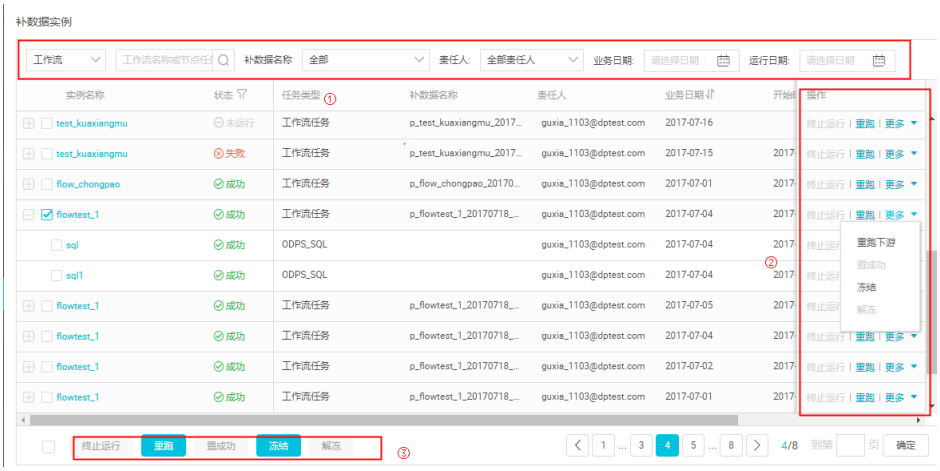
		
5	等待状态	
6	冻结状态	



补数据实例是对周期任务进行补数据时产生的实例，可对已调度起的实例任务进行运维管理，如对查看运行状态，对任务进行终止、重跑、解冻等操作。

实例列表

以列表形式对被补数据触发的任务运维及管理。包括检查运行日志、重跑任务、终止正在运行的任务等，具体功能如下：



筛选功能：如上图中①部分，有丰富的筛选条件，默认筛选条件为业务日期是当前时间前一天的工作流任务，用户可添加任务名称、运行时间、责任人等条件进行更精确的筛选。

终止运行：只可对等待运行、运行中状态的实例进行终止运行操作，进行此操作后，该实例将为失败状态。

重跑：可以重跑某任务，任务执行成功后可以触发下游未运行状态任务的调度。常用于处理出错节点和漏跑节点。

> 前置条件：只能重跑未运行、成功、失败状态的任务。

重跑下游：可以重跑某任务及其下游任务，需要用户自定义勾选，勾选的任务将被重跑，任务执行成功后可以触发下游未运行状态任务的调度。常用于处理数据修复。

> 前置条件：只能勾选未运行、完成、失败状态的任务，如果勾选了其他状态的任务，页面会提示“已选节点中包含不符合运行条件的节点”，并禁止提交运行。

置成功：将当前节点状态改为成功，并运行下游未运行状态的任务。常用于处理出错节点。

> 前置条件：只能失败状态的任务能被置成功。

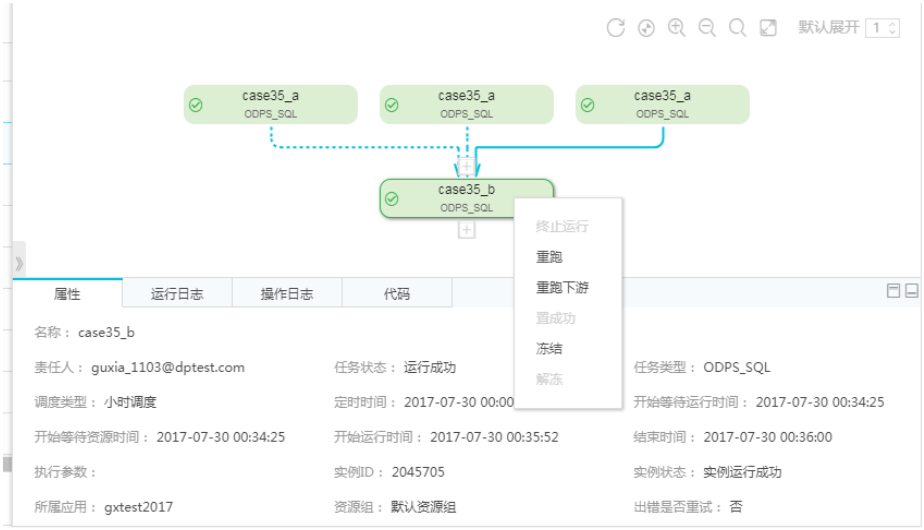
冻结：冻结状态的任务会生成实例，但是不会运行；若需要运行冻结的实例，可以解冻实例，单击重跑，实例才会开始运行。

解冻：可以将冻结状态的实例解冻；若该实例还未运行，则上游任务运行完毕后，会自动运行；若上游任务都运行完毕，则该任务会直接被置为失败，需要手动**重跑**后，实例才会正常运行。

批量操作：如上图中③部分，批量操作包括，终止运行、重跑、置成功、冻结、解冻5个功能。

实例dag图

在实例dag视图中，可以查看该实例的依赖关系和详细信息并进行终止运行、重跑等具体操作。如下图：



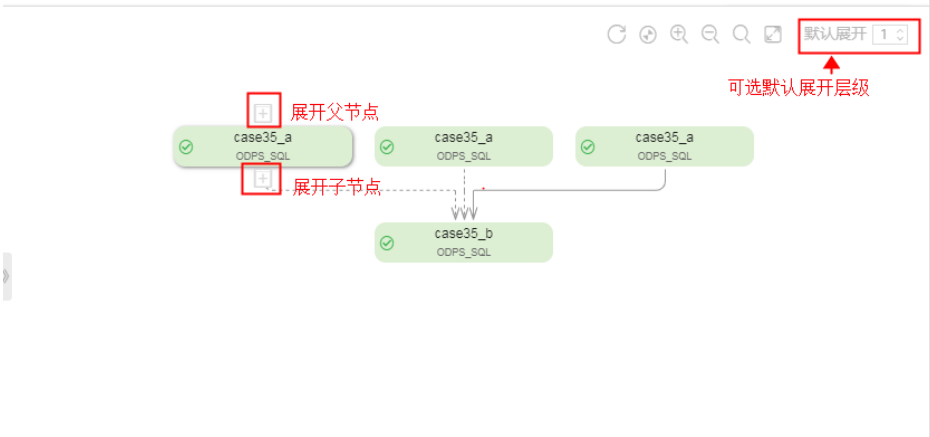
属性：查看实例属性，包括实例运行的各种时间信息、运行状态等。

运行日志：节点正在运行、成功、失败等状态时查看任务运行的日志。

操作日志：记录用户给该实例的进行的操作，例如终止运行，重跑等。

代码：可查看该实例任务的代码。

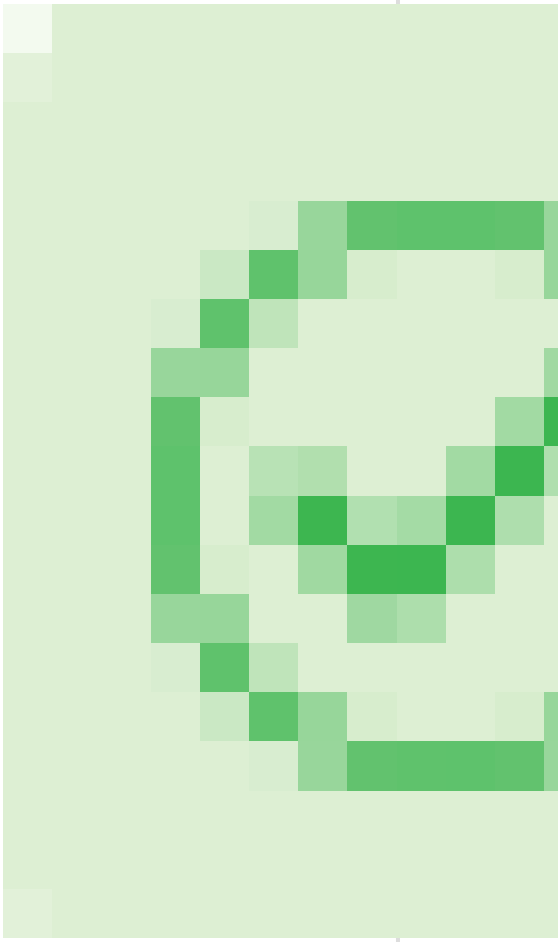
展开父节点/子节点：当一个工作流有3个节点及以上时，运维中心展示任务时会自动隐藏节点，用户可通过展开父子层级，来看到全部节点的内容。



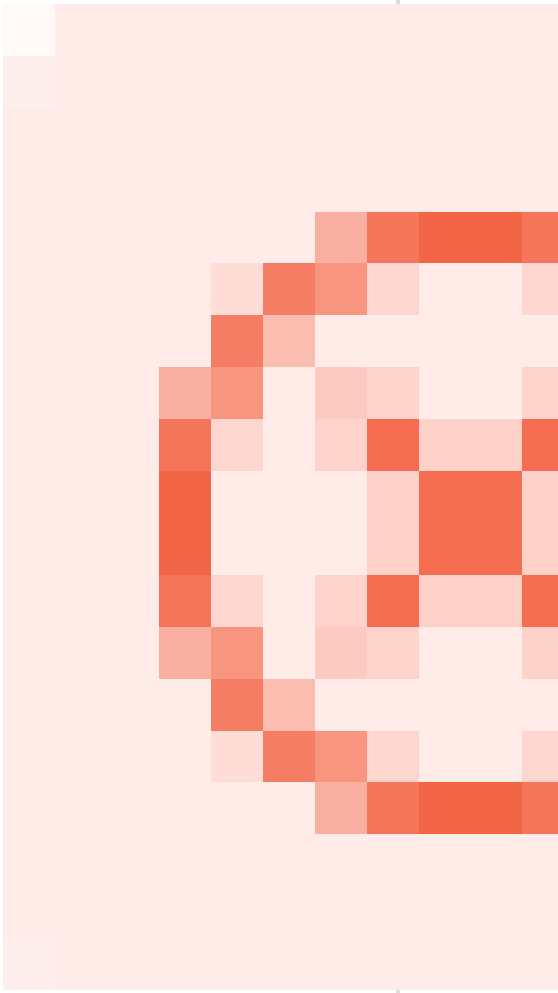
展开/关闭工作流：有工作流任务时，可以展开工作流任务，查看内部节点任务的运行状态。如下图：

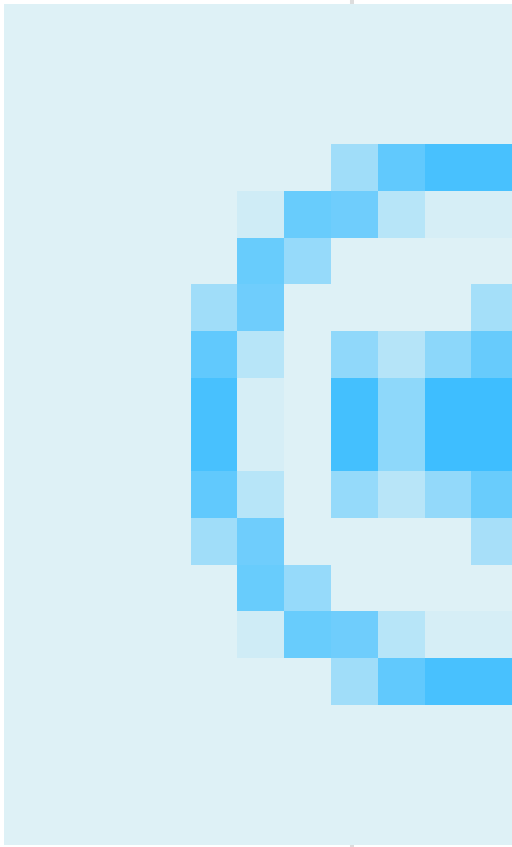
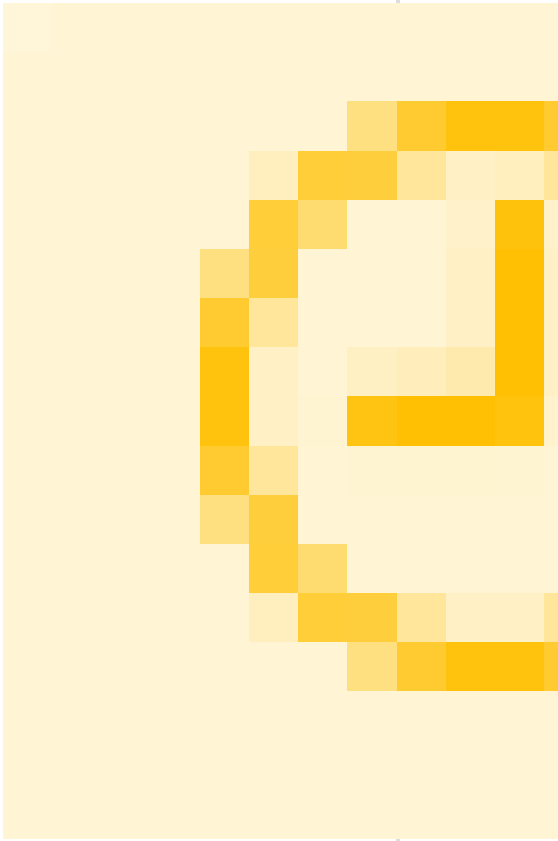
实例状态说明

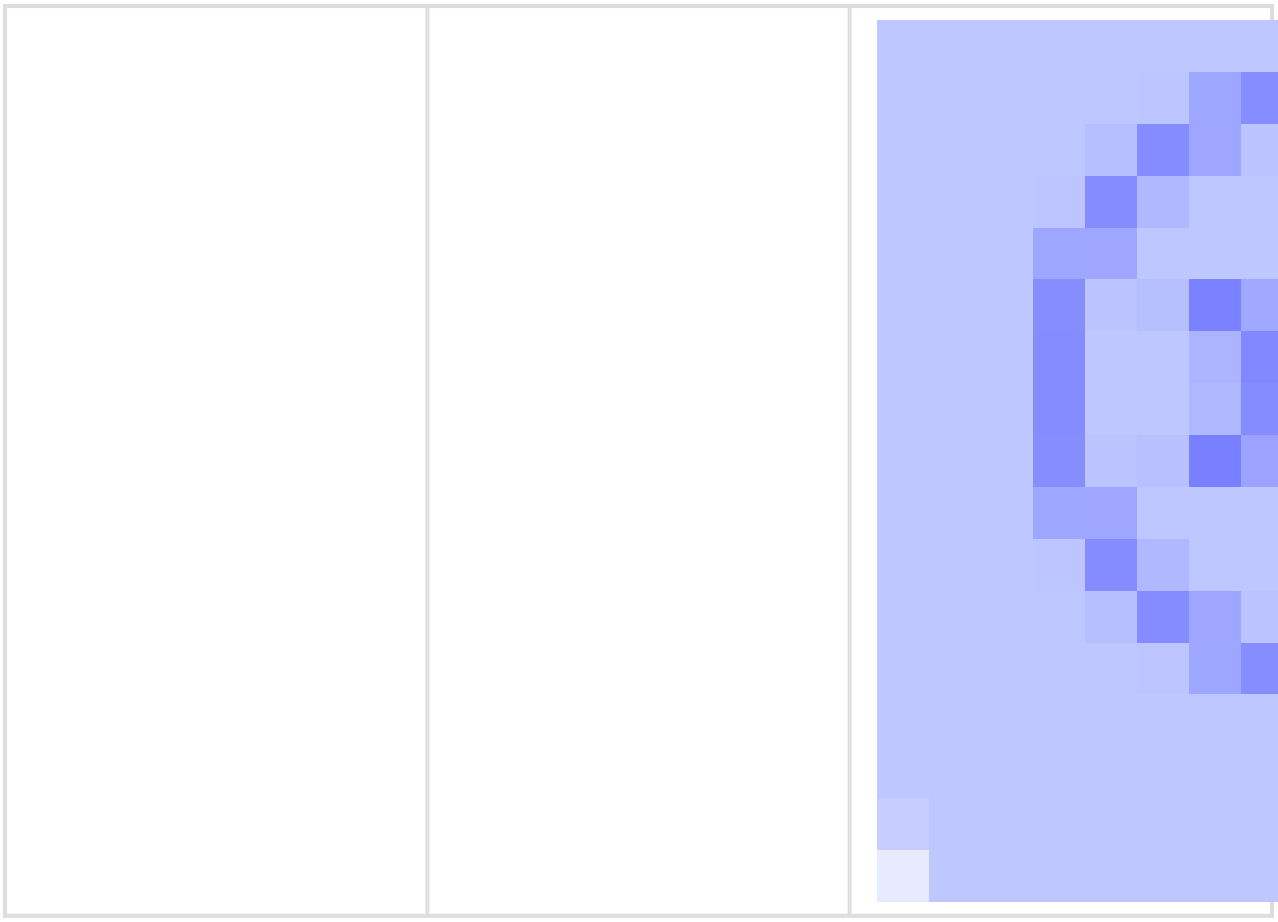
序号	状态类型	状态标识
1	运行成功状态	

		
2	未运行状态	

		
3	运行失败状态	

		
4	正在运行状态	

		
5	等待状态	
6	冻结状态	



针对补数据实例的展示，实例操作解释等。

概览是您进入运维中心后最先看到的页面，此页面将以图表的方式展示当前项目空间下周期性调度任务的整体运行情况 。

运维中心概览主要包含如下几个模块：

- 实例执行概览
- 任务节点执行时长排行
- 近一个月出错排行

注：实例执行概览的数据和实时数据有五分钟的时间差，其他两个模块的数据是离线的，延迟一天。

实例执行概览

本模块主要针对正常周期性调度今天、昨天与历史平均水平的任务完成情况进行对比统计，如果 3 条曲线偏移过多，则表示在某个时间段内有异常情况出现，需进行进一步的检查与分析 。

注：此处只统计已完成状态的任务实例

如上述折线统计图所示，分别以三种不同颜色折线显示对当天 00:00 ~24:00 时间段内当前项目空间中所有类型任务完成进度的统计，包括今天的任务完成情况、昨天的任务完成情况和历史平均水平的完成情况。主要对实例执行的情况做了一个报表展示，支持根据任务类型过滤数据。

任务节点执行时长排行

本模块展示的是：业务日期前一天，实例状态为正常或失败状态的周期任务执行时长排行，只展示TOP5。

任务节点执行时长排行

节点名称	负责人	执行时长
test_shell	base_2oxs_1	3分35秒
open_mr	base_2oxs_1	1分20秒
sql	base_2oxs_1	1分19秒
sql	base_2oxs_1	1分18秒
cdp	base_2oxs_1	1分12秒

注：单击任务名，可直接跳转到实例页面。业务日期前一天：比如今天是2017-08-22号，业务日期是2017-08-21号，那么这里展示的是2017-08-20号的数据。

近一个月出错排行

本模块展示的是：当前业务日期的近一个月的出错任务排行榜，只展示TOP5。

近一个月出错排行

节点名称	负责人	出错次数
分钟任务	base_2oxs_1	384
gxodpssql	base_2oxs_1	72
sql	base_2oxs_1	48
sql	base_2oxs_1	45
test_0a	base_2oxs_1	32

注：单击任务名，可直接跳转到任务列表页面。

整库迁移

整库迁移 是帮助提升用户效率、降低用户使用成本的一种快捷工具，它可以快速把一个 Mysql DB 库内所有表一并上传到 MaxCompute 的工作，节省大量初始化数据上云的批量任务创建时间。

假设 DB 有 100 张表，您原本可能需要配置 100 次数据同步任务，但有了整库上传便可以一次性完成。同时，由于数据库的表设计规范性的问题，此工具并无法保证一定可以一次性完成所有表按照业务需求进行同步的工作，即它有一定的约束性，本文将主要从功能性和约束性对整库上传进行介绍。

任务生成规则

完成配置后，根据选择的需要同步的表，依次创建 MaxCompute 表，生成数据同步任务。

MaxCompute 表的表名、字段名和字段类型根据高级配置生成，如果没有填写高级配置，则与 Mysql 表的结构完全相同。表的分区为 pt，格式为 yyyymmdd。

生成的数据同步任务是按天调度的周期任务，会在第二天凌晨自动运行，传输速率为 1M/s，它在细节上会因为同步的方式、并发配置等有所不同，您可以在同步任务目录树的 **clone_database > 数据源名称 > mysql2odps_表名** 中找到生成的任务，然后对其进行更加个性化的编辑操作。

备注：

建议您当天对数据同步任务进行冒烟测试，相关任务节点可以在 **运维中心-任务管理** 中的 **project_etl_start > 整库上传 > 数据源名称** 下找到所有此数据源生成的同步任务，然后右键单击测试相应的节点即可。

约束限制

由于数据库的表设计规范性的问题，整库上传具有一定的约束性，具体如下：

目前仅提供 Mysql 数据源的整库上传到 MaxCompute，后续 Hadoop/Hive 数据源、Oracle 数据源功能会逐渐开放。

仅提供每日增量、每日全量的上传方式

如果您需要一次性同步历史数据，则此功能无法满足您的需求，故给出以下建议：

建议您配置为每日任务，而非一次性同步历史数据。您可以通过调度提供的补数据，来对历史数据进行追溯，这样可避免全量同步历史数据后，还需要做临时的 SQL 任务来拆分数据。

如果您需要一次性同步历史数据，可以在任务开发页面进行任务的配置，然后单击 **运行**，完成后通过 SQL 语句进行数据的转换，因为这两个操作均为一次性行为。

如果您每日增量上传有特殊业务逻辑，而非一个单纯的日期字段可以标识增量，则此功能无法满足您的需求，故给出以下建议：

数据库数据的增量上传有两种方式：通过 binlog（DTS 产品可提供）和数据库提供数据变更的日期字段来实现。目前数据集成支持的为后者，所以要求您的数据库有数据变更的日期字段，通过日期字段，系统会识别您的数据是否为业务日期当天变更，即可同步所有的变更数据。

为了更方便地增量上传，建议您在创建所有数据库表的时候都有：gmt_create, gmt_modify 字段，同时为了效率更高，建议增加 id 为主键。

整库上传提供分批和整批上传的方式

分批上传为时间间隔。目前不提供数据源的连接池保护功能，此功能正在规划中。

为了保障对数据库的压力负载，整库上传提供了分批上传的方式，您可以按照时间间隔把表拆分为几批运行，避免对数据库的负载过大，影响正常的业务能力。以下有两点建议：

如果您有主、备库，建议同步任务全部同步备库数据。

批量任务中每张表都会有 1 个数据库连接，上限速度为 1M/s。如果您同时运行 100 张表的同步任务，就会有 100 个数据库进行连接，建议您根据自己的业务情况谨慎选择并发数。

- 如果您对任务传输效率有自己特定的要求，此功能无法实现您的需求。所有生成任务的上限速度均为 1M/s。

仅提供整体的表名、字段名及字段类型映射

整库上传会自动创建 MaxCompute 表，分区字段为 pt，类型为字符串 string，格式为 yyyymmdd。

注意：

选择表时必须同步所有字段，它不能对字段进行编辑。

整库迁移是为了提升用户效率、降低用户使用成本的一种快捷工具，它可以快速完成把MySQL DB库内所有表一并上传到MaxCompute的工作，关于整库迁移的详细介绍请参见整库迁移概述。

本文将通过实践操作，为您介绍如何使用整库迁移功能，完成MySQL数据整库迁移到MaxCompute。

操作步骤

登录到DataWorks>数据集成控制台，单击左侧的**离线同步>数据源**，进入数据源管理页面。

单击右上角的**新增数据源**，添加一个面向整库迁移的MySQL数据源clone_databae，如下图所示：

单击**测试连通性**验证数据源访问正确无误后，确认并保存此数据源。

新增数据源成功后，即可在数据源列表中看到新增的MySQL数据源clone_databae。单击对应MySQL数据源后的**整库迁移**，即可进入对应数据源的整库迁移功能界面，如下图所示：

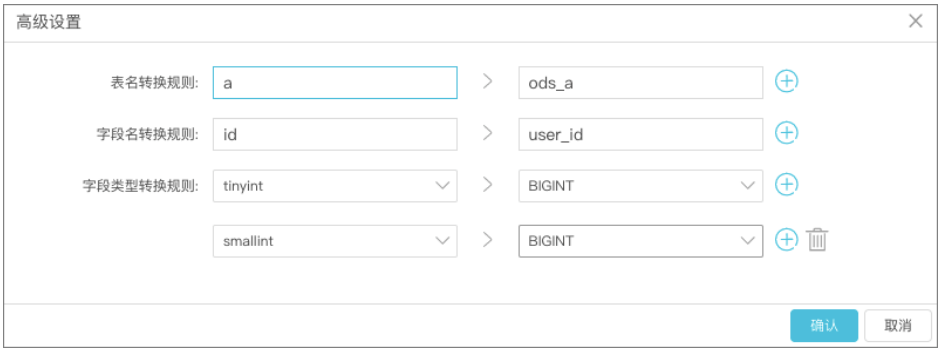
整库迁移界面主要分为3块功能区域，如下图所示：

待迁移表筛选区，此处将MySQL数据源clone_databae下的所有数据库表以表格的形式展现出来，您可以根据实际需要批量选择待迁移的数据库表。

高级设置，此处提供了MySQL数据表和MaxCompute数据表的表名称、列名称、列类型的映射转换规则。

迁移模式、并发控制区，此处可以控制整库迁移的模式（全量、增量）、并发度配置（分批上次、整批上传）、提交迁移任务进度状态信息等。

单击**高级设置**按钮，您可以根据您具体需求选择转换规则。比如MaxCompute端建表时统一增加了ods_这一前缀，如下图所示：



在迁移模式、并发控制区中，选择同步方式为**每日增量**，并配置增量字段为gmt_modified，数据集成默认会根据您选择的增量字段生成具体每个任务的增量抽取where条件，并配合DataWorks调度参数比如${bdp.system.bizdate}形成针对每天的数据抽取条件。如下图所示：

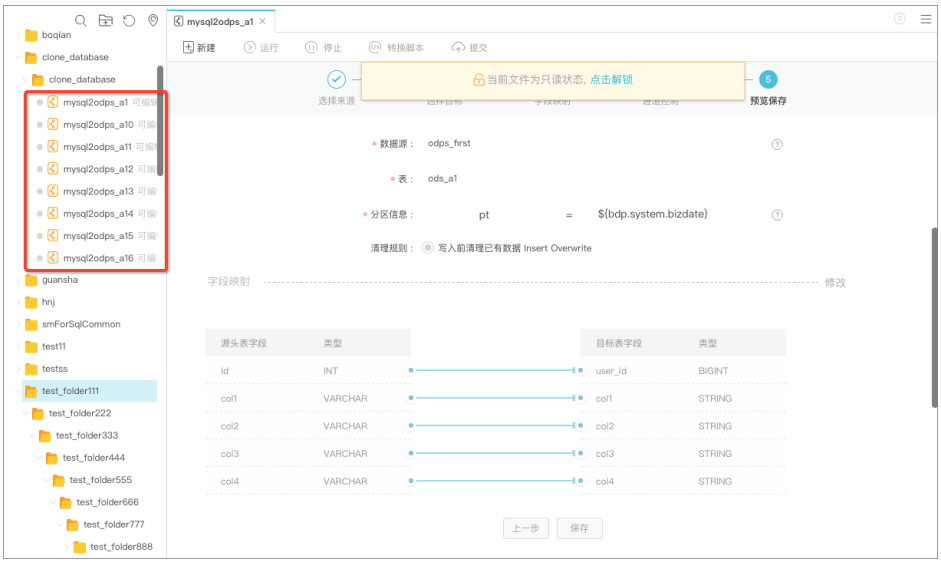
数据集成抽取MySQL库表的数据是通过JDBC连接远程MySQL数据库，并执行相应的SQL语句，将数据从MySQL库中Select出来。由于是标准的SQL抽取语句，可以配置Where子句控制数据范围。此处您可以查看到增量抽取的Where条件如下所示：

```
STR_TO_DATE('${bdp.system.bizdate}', '%Y%m%d') <= gmt_modified AND gmt_modified <
DATE_ADD(STR_TO_DATE('${bdp.system.bizdate}', '%Y%m%d'), interval 1 day)
```

为了对源头MySQL数据源进行保护，避免同一时间点启动大量数据同步作业带来数据库压力过大，此处选择分批上传模式，并配置从每日0点开始，每1小时启动3个数据库表同步。

最后，单击**提交任务**按钮，这里可以看到迁移进度信息，以及每一个表的迁移任务状态。

单击a1表对应的迁移任务，会跳转到数据集成的任务开发界面。如下图所示：

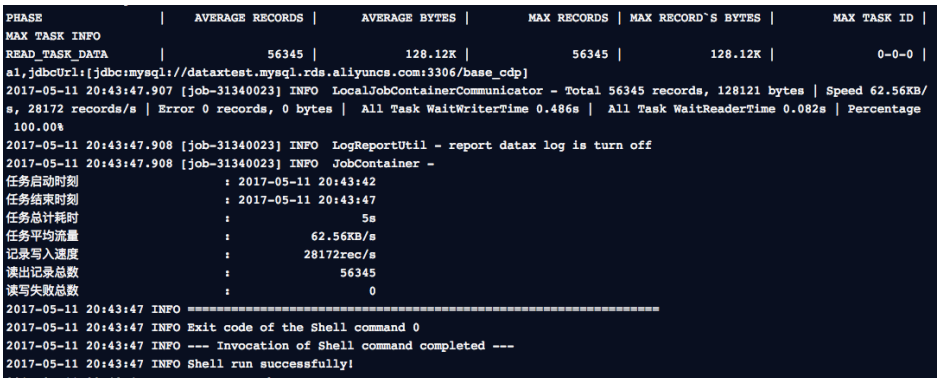


由上图可以看到源头a1表对应的MaxCompute表ods*a1创建成功，列的名字和类型也符合之前映射*

转换配置。在左侧目录树clone_database目录下，会有对应的所有整库迁移任务，任务命名规则是：mysql2odps源表名，如上图红框部分所示。

此时便完成了将一个MySQL数据源clone_databae整库迁移到MaxCompute的工作。这些任务会根据配置的调度周期（默认天调度）被调度执行，您也可以使用DataWorks调度补数据功能完成历史数据的传输。通过**数据集成>整库迁移**功能可以极大减少您初始化上云的配置、迁移成本。

整库迁移a1表任务执行成功的日志如下图所示：



为保证数据库的安全稳定，在开始使用某些数据库时实例前，您需要将访问数据库的 IP 地址或者 IP 段加到目标实例的 白名单 或安全组中。本文将主要介绍您选择不同 region 的 DataWorks（数据工场，原大数据开发套件）时，如何添加需要的安全组。

添加安全组

如果您的 ECS 上的自建数据源同步任务运行在自定资源组上，要给自定资源组机器授权，将自定义机器内/外网 IP 和端口添加到 ECS 安全组上。

如果您的 ECS 上的自建数据源运行默认的资源组上，要给默认的机器授权，根据您选择 DataWorks 的 region来填写您的安全组内容，如下表所示：

Region	授权对象	账号ID
华东2（上海）	sg-bp13y8iuj33uqpqvgqw2	1156529087455811
华南1（深圳）	sg-wz9ar9o9jgok5tajj7ll	1156529087455811
亚太东南1（新加坡）	sg-t4n222njci99ik5y6dag	1156529087455811
香港	sg-j6c28uqpqb27yc3tjmb6	1156529087455811
美国西部 1 (硅谷)	sg-rj9bowpmdvhyl53lza2j	1156529087455811
华北2（北京）	sg-2ze3236e8pcbxw61o9y0	1156529087455811

ECS 添加安全组

操作步骤

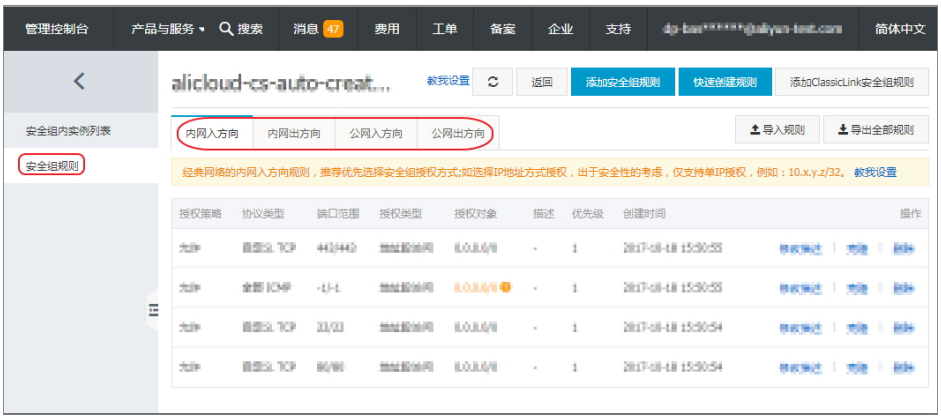
登录云服务器 ECS 管理控制台。

在左侧导航栏中，选择 **网络和安全 > 安全组**。

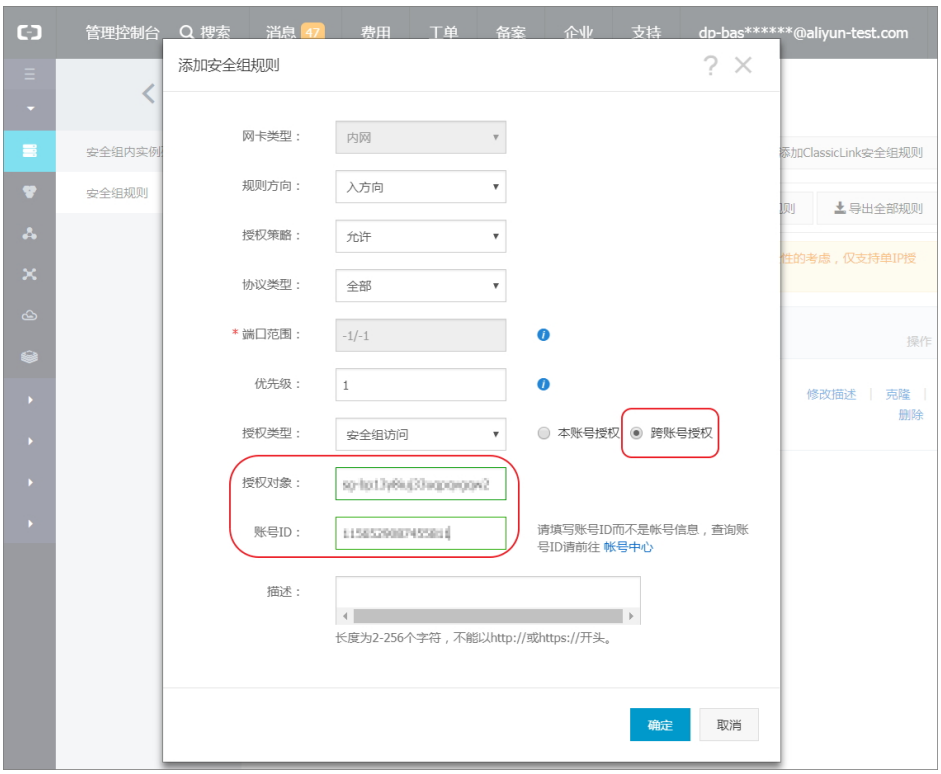
选择目标地域。

找到要配置授权规则的安全组，在 **操作** 列中，单击 **配置规则**。

进入 **安全组规则** 页面，单击 **添加安全组规则**。



在弹出的对话框中，设置以下参数：


管理控制台
搜索
消息 47
费用
工单
备案
企业
支持
dp-bas******@aliyun-test.com
添加安全组规则
网卡类型：
内网
规则方向：
入方向
授权策略：
允许
协议类型：
全部
* 端口范围：
-1/-1
优先级：
1
授权类型：
安全组访问
本账号授权
跨账号授权
授权对象：
账号ID：
请填写账号ID而不是帐号信息，查询账号ID请前往 帐号中心
描述：
长度为2-256个字符，不能以http://或https://开头。
确定
取消
安全组内实例
安全组规则
添加ClassicLink安全组规则
导出全部规则
操作
修改描述
克隆
删除

单击 **确认**。