ALIBABA CLOUD

阿里云

数据湖分析

DMS

文档版本：20211227

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

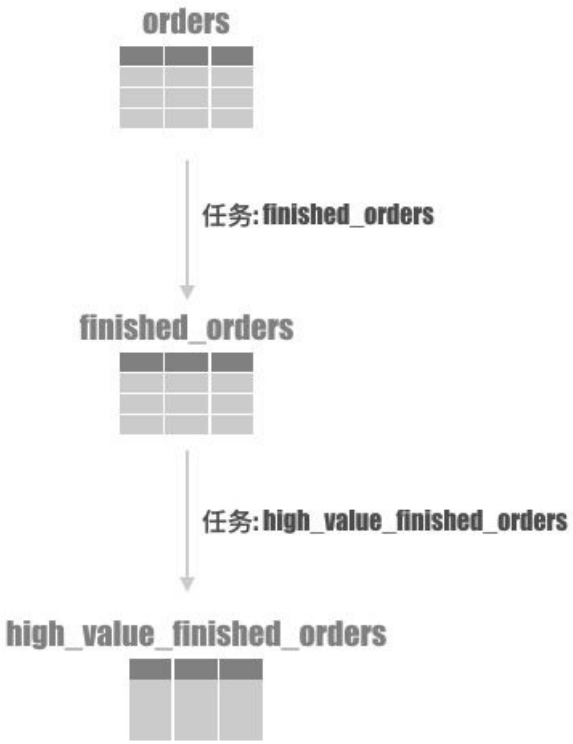
目录

1.调度DLA Presto任务 ---------- 05
2.调度DLA Spark任务 ---------- 10

1.调度DLA Presto任务

您可以在数据管理DMS中创建和调度DLA任务流实现数仓开发，任务流只需定义一次，即可周期性地自动被调度执行，大大减少人工操作成本。同时执行DLA任务流时产生的中间结果可以被复用，例如用于其他数据分析任务，从而最大化利用DLA的资源。

任务调度中一个重要的功能是任务之间的依赖，为演示该功能，本文在DMS中创建两个DLA任务，表、任务之间的关系如下图所示。



- **任务一**

  从orders表查询出已经完成的订单：`o_orderstatus = 'F'`，并将其写入finished_orders表。

- **任务二**

  从finished_orders中查询出总价大于10000的订单：`o_totalprice > 10000`，并将其写入high_value_finished_orders表。

前提条件

- 您需要开通DLA服务，然后创建一个Schema（database）以及在Schema中创建测试表。
  i. 开通DLA服务，请参见开通云原生数据湖分析服务。
  ii. 初始化数据库主账号密码，请参见重置数据库账号密码。
  iii. 创建Schema和表。

  创建Schema

```
  CREATE SCHEMA gmall with DBPROPERTIES(
    LOCATION = 'oss://oss-bucket-name/gmall/',
    catalog='oss'
  );
```

创建orders表

```
CREATE EXTERNAL TABLE orders (
  O_ORDERKEY INT,
  O_CUSTKEY INT,
  O_ORDERSTATUS STRING,
  O_TOTALPRICE DOUBLE,
  O_ORDERDATE DATE,
  O_ORDERPRIORITY STRING,
  O_CLERK STRING,
  O_SHIPPRIORITY INT,
  O_COMMENT STRING
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
STORED AS TEXTFILE
LOCATION 'oss://bucket-name/gmall/orders';
```

finished_orders表

```
CREATE EXTERNAL TABLE IF NOT EXISTS finished_orders (
    O_ORDERKEY INT,
    O_TOTALPRICE DOUBLE
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
STORED AS TEXTFILE
LOCATION 'oss://bucket-name/gmall/finished_orders/';
```

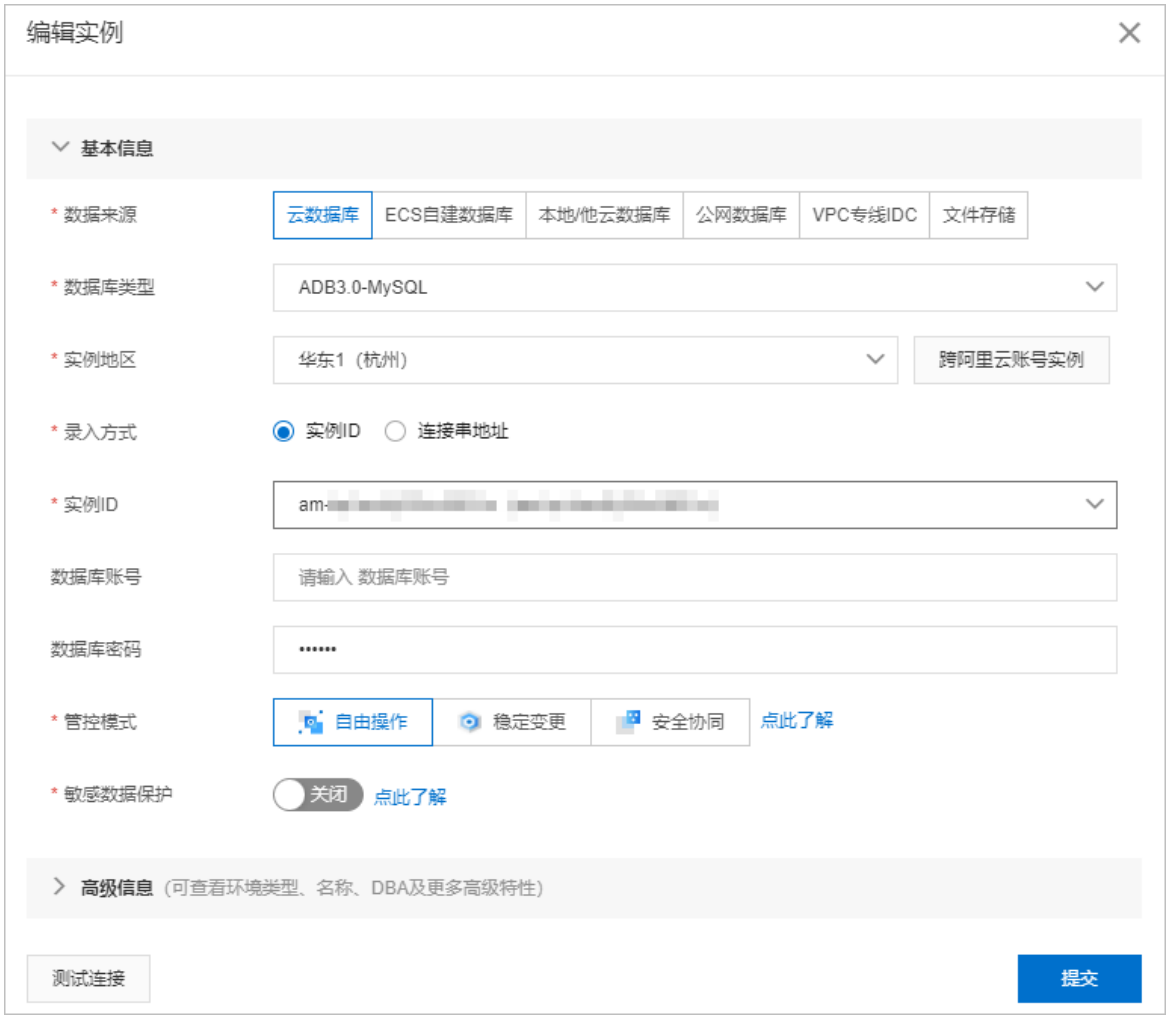
high_value_finished_orders表

```
CREATE EXTERNAL TABLE IF NOT EXISTS high_value_finished_orders (
  O_ORDERKEY INT,
  O_TOTALPRICE DOUBLE
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
STORED AS TEXTFILE
LOCATION 'oss://bucket-name/gmall/high_value_finished_orders/';
```

- 在DMS产品详情页单击**立即购买**，开通DMS服务。

步骤一：新增DLA数据源

1. 登录数据管理DMS控制台。
2. 在左侧导航栏选择**实例管理**，单击**实例列表**。
3. 单击目标实例地址所在行**更多**中的**编辑实例**。
4. 在**编辑实例**页面，根据页面提示进行参数配置。



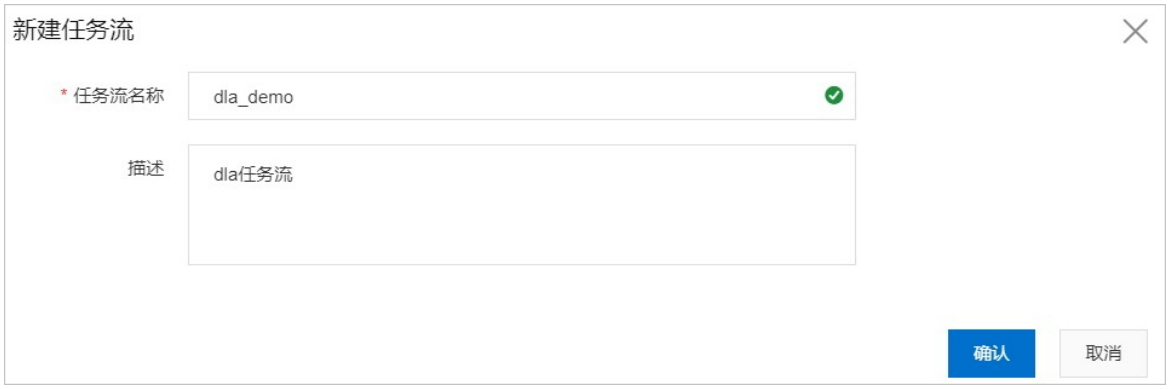
参数类别	参数名称	参数说明
基本信息	**数据来源**	选择**云数据库**。
	数据库类型	选择**DLA-数据湖分析**。
	实例地区	选择DLA所在地域。
	录入方式	您可以通过两种方式录入DLA信息： ○ **实例ID**，从实例列表中选择DLA实例ID。 ○ **连接串地址**，手动输入DLA连接地址。
	实例ID	○ 录入方式为实例ID时，从实例列表中选择DLA实例ID。 ○ 录入方式为连接串地址时，手动输入DLA连接地址。
	数据库账号	DLA中的数据库账号。

参数类别	参数名称	参数说明
	数据库密码	数据库账号对应的密码。
	管控模式	请参见管控模式。
	安全规则	如果您选择的管控模式为**安全协同**，需要填写**安全规则**，选择DLA默认规则。
高级信息	**环境类型**	本例选择**生产**。
	实例名称	DLA的实例ID或者名称。
	开启跨库查询	默认选中**开启跨库查询**，且不支持用户修改。
	无锁结构变更	默认选中**关闭**，且不支持用户修改。
	开启SSL	您可以选择**默认（DMS与数据库服务自动协商是否开启）**、**开启**或**关闭**。
	实例DBA	您可以根据实际情况选择一个账号作为**实例DBA**。
	查询超时时间（s）	本例设置为60。
	导出超时时间（s）	本例设置为600。

5. 完成参数配置后单击**测试连接**，测试成功后单击**提交**。

步骤二：编排任务流

1. 登录数据管理DMS控制台。
2. 在上方导航栏，单击**传输与加工（DTS）**。
3. 单击左侧导航栏中的**任务编排**，然后单击**新增任务流**。
4. 在**新建任务流**对话框，设置**任务流名称**和**描述**，然后单击**确认**。



5. 拖拽两个DLA节点到中间的空白区域中，然后单击任务1的下边框连接点，引出一条到达任务2的带箭头直线关联任务1和任务2。

6. 分别单击任务1和任务2节点，在**内容设置**中输入或者选择DLA Schema的名字，然后单击**保存**。
7. 单击任务1节点，在**内容设置**中输入SQL实现任务一，即从orders表查询出已经完成的订单`o_orderstatus = 'F'`，并将其写入finished_orders表，单击**保存**。

```
insert into finished_orders
  select O_ORDERKEY, O_TOTALPRICE
  from orders
  where O_ORDERSTATUS = 'F';
```

8. 单击任务2节点，在**内容设置**中输入SQL实现任务二，即从finished_orders中查询出总价大于10000的订单`o_totalprice > 10000`，并将其写入high_value_finished_orders表，单击**保存**。

```
insert into high_value_finished_orders
select * from finished_orders
where O_TOTALPRICE > 10000;
```

步骤三：运行任务

任务流编排完成后，您可以通过如下方式运行任务。

- 单击**试运行**，系统弹出提示窗口，单击**确认**运行任务。
- 开启**调度配置**为任务设置周期运行策略。

步骤四：查看任务

任务运行结束后，您可以通过如下方式查看任务。

- 单击**运维中心**查看任务运行状态。
- 登录Data Lake Analytics管理控制台，在DLA中查询任务运行结果。
  - 任务一：从orders表查询出已经完成的订单`o_orderstatus = 'F'`。
  - 任务二：从finished_orders中查询出总价大于10000的订单`o_totalprice > 10000`。

2.调度DLA Spark任务

DLA Serverless Spark目前支持DataWorks和DMS编排调度任务，同时也提供自定义SDK和Spark-Submit工具包供用户自定义编排调度。本文将介绍如何使用DMS来编排调度Spark任务。

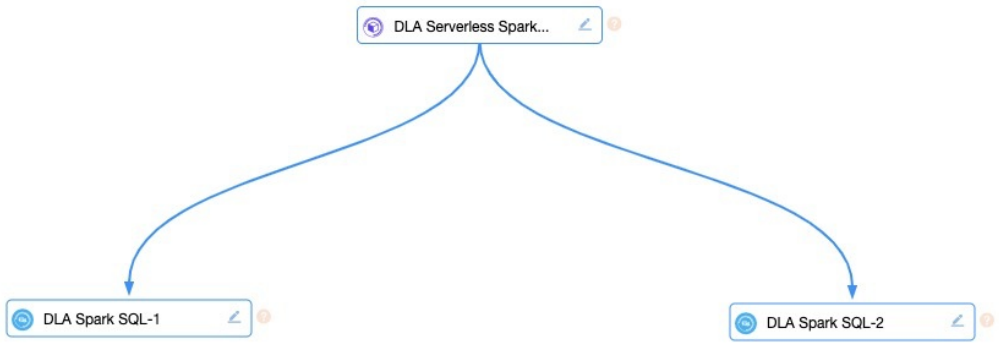
前提条件

- 您已开通DMS服务。
- 您已开通云原生数据湖分析服务，详情请参见开通云原生数据湖分析服务
- 您已开通OSS服务，详情请参见开通OSS服务。
- 您已创建DLA Spark虚拟集群，详情请参见创建虚拟集群。

> **说明** 如果您想用RAM用户提交Spark作业，且之前未使用过RAM用户在DLA控制台提交作业，您可以参见细粒度配置RAM子账号权限进行子账号提交作业配置。

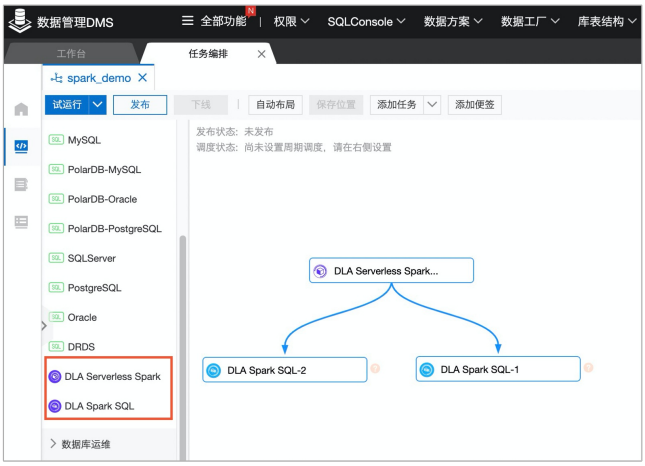
操作步骤

任务调度中一个重要的功能是任务之间的依赖，为演示这个功能，本文会在DMS中创建一个DLA Serverless Spark任务和两个 DLA Spark SQL任务，任务之间的依赖关系如下图所示，任务**DLA Spark SQL-1**和**DLA Spark SQL-2** 依赖上游任务完成之后，才能执行。

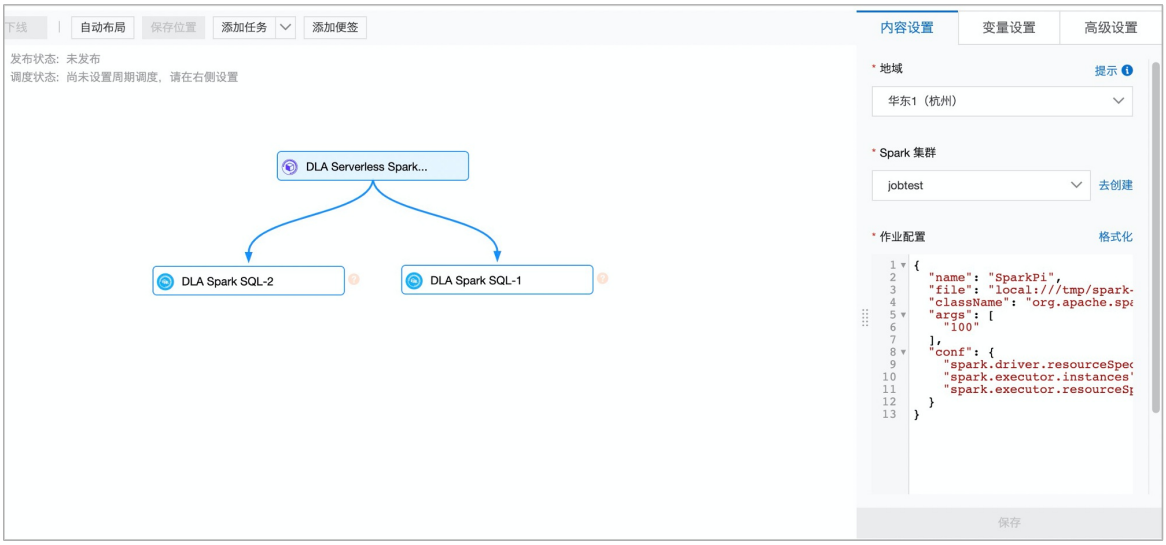


1. 登录DMS控制台。
2. 在顶部菜单栏中，单击**数据工厂 > 任务编排**。
3. 在**任务编排**页面的**自由编排任务**区域，单击**新增任务流**。

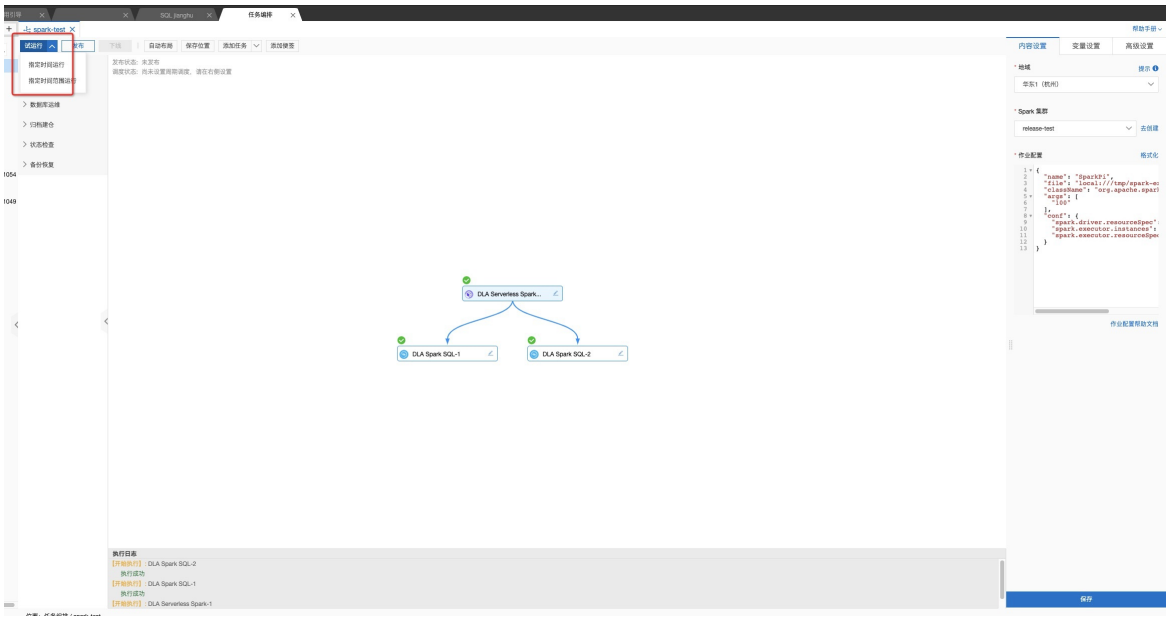
4. 在**新建任务流**对话框，将**任务流名称**设置为spark_demo，将**描述**设置为spark demo，完成后单击**确认**。
5. 在**任务编排**页面，从左侧任务类型中拖拽一个**DLA Serverless Spark**任务和两个**DLA Spark SQL**任务，并按照下图中的依赖关系进行连线。



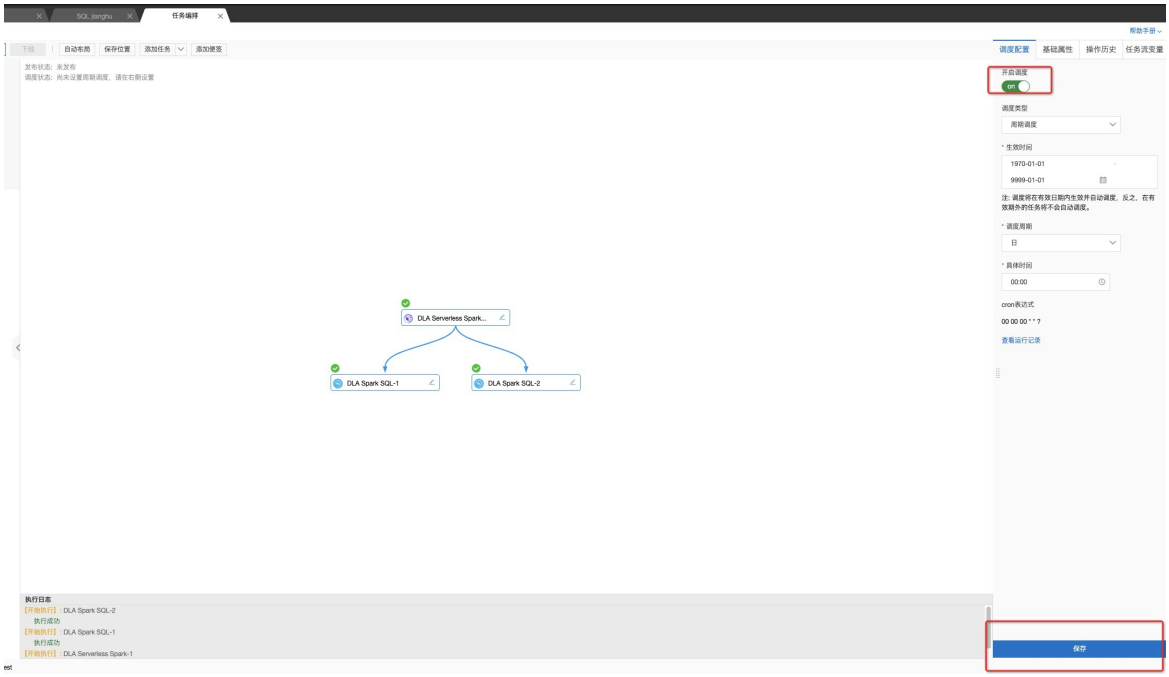
6. 依次单击3个节点，在每个节点的右侧面板中选中**内容设置**页签，并配置以下信息：
   - 在**地域**列表中，选择目标Spark集群所在的地域。
   - 在**Spark 集群**列表中，选择目标Spark集群。
   - 在**作业配置**文本框中已有运行内置Sparkpi的配置参数或者`show databases`，如果您需要运行自定义作业，可根据您的需求进行修改。



7. 完成以上配置后，单击**保存**按钮，然后单击页面左上方的**试运行**或**指定时间运行**或**指定时间范围运行**任务。



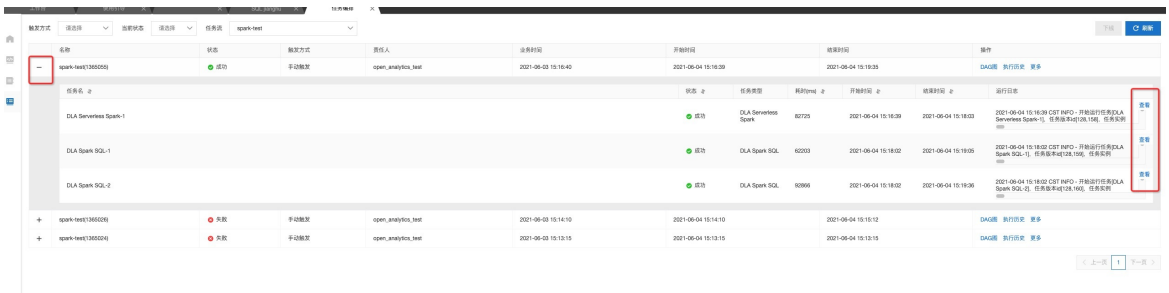
8. 您也可以在点击任务编排界面的空白界面，进行整个任务的调度设置。



查看任务日志

1. 点击任务编排界面，左侧的**运维中心**。

2. 单击对应任务左侧的加号按钮，即可看到任务流的所有任务，点击子任务的**查看**按钮，即可查看运行日志。



3. 从日志中可以查看该作业的JobId和对应的SparkUI， 如果任务出错，无法排查，请记下JobId和SparkUI联系Spark值班。

自定义任务编排调度

DLA Serverless Spark除了上述调度集成之外，还提供了SDK以及Spark-Submit工具用于提交Spark作业、查询作业状态、获取作业日志等功能，详情请参见：

- SDK安装与使用
- Spark-Submit命令行工具

您可以利用上述工具，使用第三方任务编排调度系统（例如Apache Airflow）来打造自己的工作流。