

目录

目录	1
入门概述	2
创建项目，购买服务	2
子账号授权	3
添加数据源	4
新建库表	5
数据同步作业	6
配置调度	6
作业流发布	6

入门概述

本模块将指引您快速完成一个完整的数据同步和运维操作。

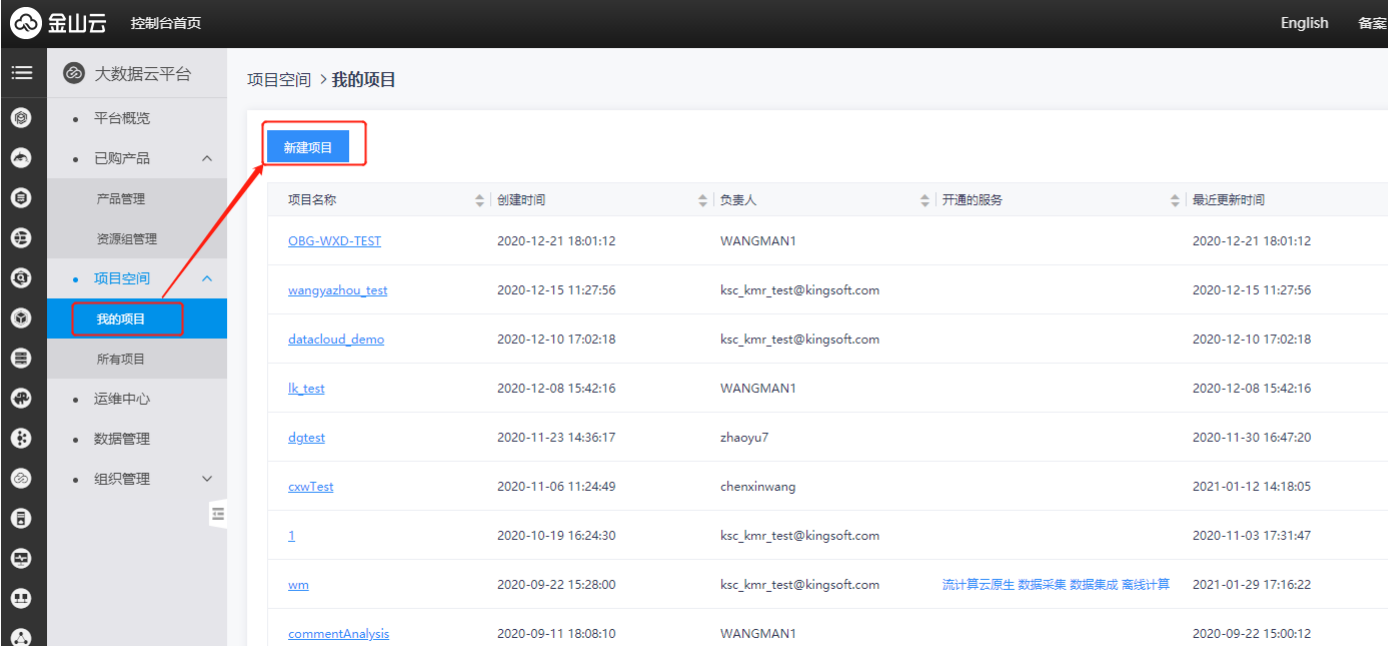
使用场景：将业务数据库MySQL的数据同步到数据仓库Hive的表中。

主要包括以下步骤：

1. [准备工作](#)
2. [元数据管理](#)
3. [创建作业](#)
4. [配置调度及发布](#)

创建项目，购买服务

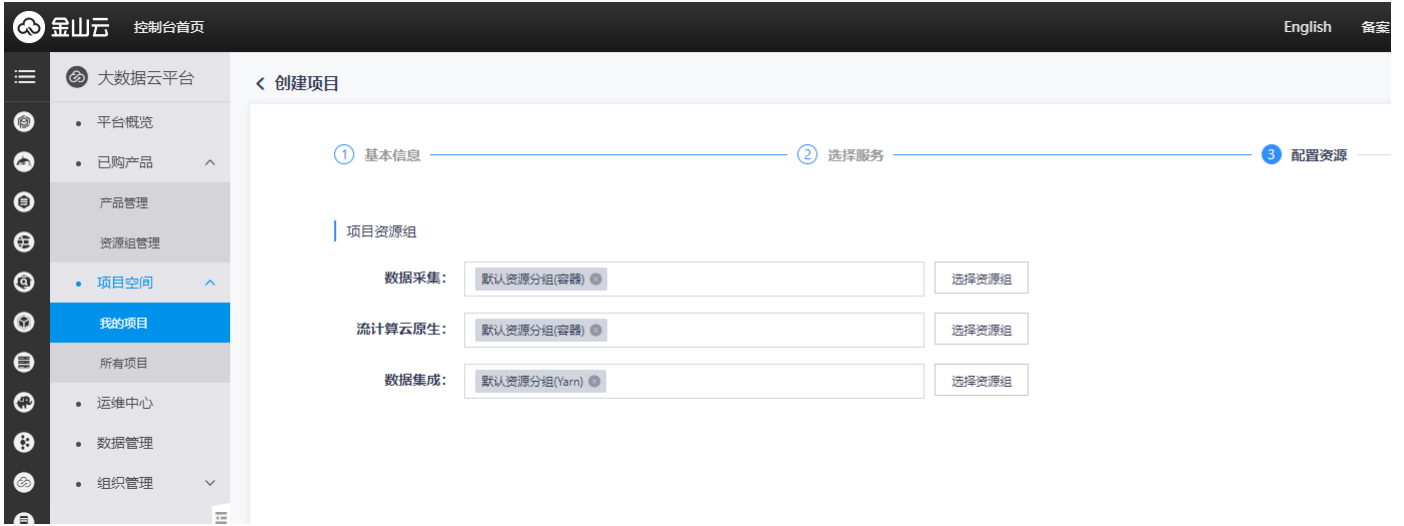
1. 用主账号登录产品控制台，左侧导航栏点击项目空间 > 我的项目，点击新建项目，填写项目基本信息。



2. 点击去购买，按需购买所需服务。公共基础服务必须购买，其他服务按需选择购买。



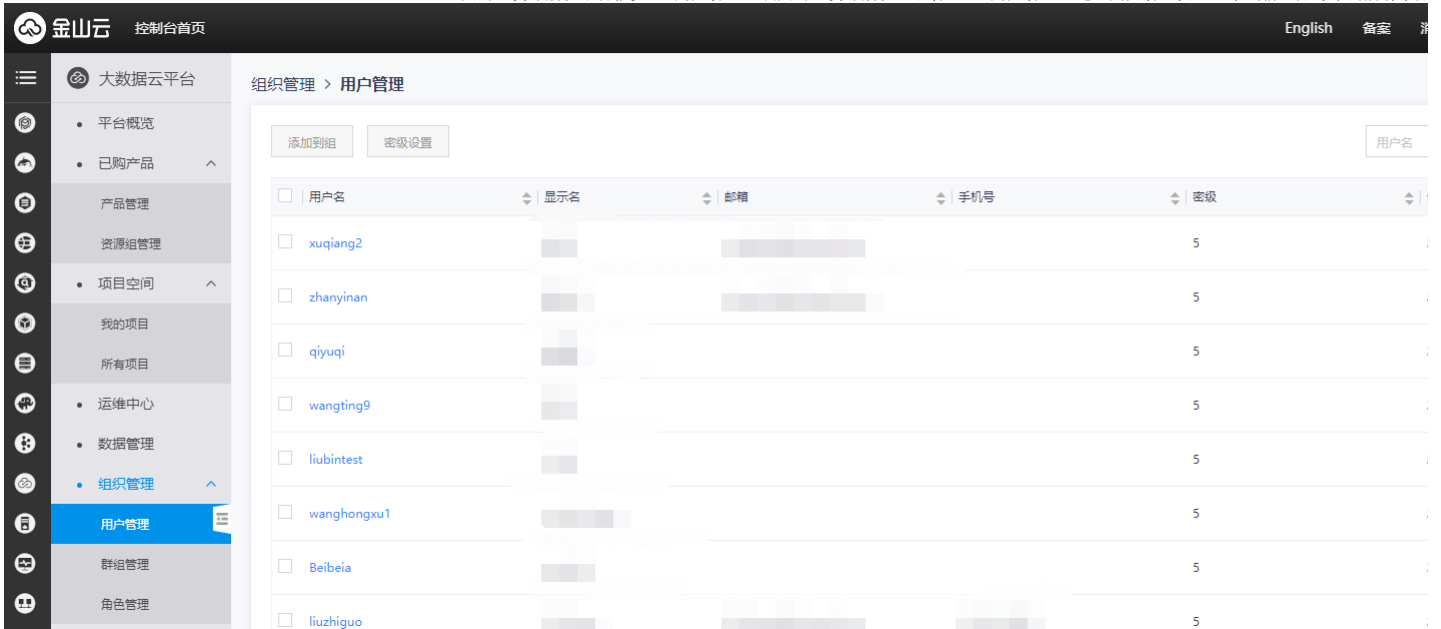
3. 为已选的服务配置资源组。您可选默认资源组，或到资源组管理中新建资源组。



4. 确认信息后，完成服务购买。

子账号授权

1. 点击左侧导航栏平台概览 > 用户管理，或点击左侧导航栏组织管理 > 用户管理，进入用户管理页，显示当前主账号下的所有子账号



2. 角色授权，可以为单个子账号授予角色，也可以将子账号添加到群组，为群组授予角色。
产品内置的角色如下：

数据资产管理员

系统管理员

全局数据管理员

项目管理员

开发人员

运维人员

业务人员

项目数据管理员

3. 为子账号授予项目的权限，在**我的项目**中，选择需要授权的项目，点击**项目名称**，进入项目管理页面，点击**成员管理** > **添加成员**，选择需要授予该项目权限的子账号，授权后，子账号拥有使用该项目下服务的权限。



添加数据源

1. 新增起始及目标数据源

(1) 点击左侧导航栏**数据管理** > **数据源管理**。



(2) 点击**新增数据源**，选择**MySQL**作为起始数据源。选择测试环境和测试环境，托管模式，填写相关信息，测试连通性，确定完成添加。

数据管理

数据源管理

元数据管理

数据地图

数据搜索

数据目录

数据权限

发布管理

数据管理 > 数据源管理

新增数据源
批量删除

数据源名称	数据源类型
<input type="checkbox"/>	ftp
<input type="checkbox"/>	outhive
<input type="checkbox"/>	redis
<input type="checkbox"/>	dgkafkainner
<input type="checkbox"/>	es

请选择您要添加的数据源

关系型数据库

oracle

mysql

大数据存储

hive

hbase

elasticsearch

半结构化存储

ftp

ks3

消息队列

kafka

其他

redis

http

(3) 点击新增数据源，选择Hive作为目标数据源。选择测试环境和测试环境，托管模式，填写相关信息，测试连通性，确定完成添加。

2. 同步元数据

在数据源管理页面，选择刚才创建的MySQL数据源，点击同步，选择要同步的数据库，所属项目选择之前创建的项目，确定。将元数据同步到库表管理中。

数据管理

数据源管理

元数据管理

数据地图

数据搜索

数据目录

数据权限

发布管理

数据源名称	数据源类型	连接模式	连接模式	托管模式	创建时间	创建人	
<input type="checkbox"/>	outhive	hive	端口号: 10000 地址: 120.92.94.43 用户名: hi**	端口号: 10000 地址: 120.92.94.4 用户名: hi**	完全托管	2021-03-02 17:48:13	自建
<input type="checkbox"/>	redis	redis	连接模式: masterSlave 主节点地址: 120.92.113.29:59771	连接模式: masterSlave 主节点地址: 120.92.113.117:22748	非托管读写	2021-03-02 15:39:43	自建
<input type="checkbox"/>	dgkafkainner	kafka	kafka集群地址: 10.100.5.79:6667	kafka集群地址: 10.100.4.3:6667	完全托管	2021-03-02 15:14:47	自建
<input type="checkbox"/>	es	elasticsearch	http访问地址: http://120.131.13.254:9200	http访问地址: http://120.131.13.47:9200	完全托管	2021-03-02 15:06:30	自建
<input type="checkbox"/>	mysql	mysql	端口号: 53920 地址: 120.92.20.55 用户名: ad***	端口号: 20648 地址: 120.92.10.155 用户名: ad***	完全托管	2021-03-02 14:33:55	自建
<input type="checkbox"/>	dgkafka	kafka	kafka集群地址: 120.92.33.79:6667	kafka集群地址: 120.92.10.42:6667	完全托管	2021-03-02 10:42:38	自建

新建库表

新建库表的具体步骤如下：

1. 新建数据库

- (1) 点击数据管理 > 元数据管理，进入库表管理页面。
- (2) 选择关系型数据库，点击新建数据库。
- (3) 填写相关信息，所属项目选择刚才创建的项目，确认创建成功。
- (4) 在库表管理页面，选择刚才创建的数据库，点击发布到测试。数据源类型选择Hive，数据源名称选择刚才创建的Hive数据源。
- (5) 确认，发布到测试环境。审核通过后，从测试环境发布到生产环境。

2. 新建数据表

- (1) 选择**关系型数据库**，点击**数据表**Tab，进入数据表管理页面。
- (2) 点击**新建数据表**，填写相关信息：
 - 基本信息：所属项目选择刚才创建的项目。
 - 字段设置：参照与MySQL中要同步的表。
 - 设置分区
- (3) 确认创建成功。
- (4) 选择刚才创建的数据表，点击**发布到测试**。数据源类型选择**Hive**，数据库选择刚才发布到测试的数据库。
- (5) 确认，发布到测试环境。审核通过后，从测试环境发布到生产环境。

数据同步作业

数据同步作业的具体操作步骤如下：

1. 创建作业

- (1) 点击左侧导航栏**项目空间 > 我的项目**，或点击刚才创建的项目名称进入**项目空间 > 我的项目**，点击项目名称进入大数据开发套件。点击进入**数据开发 > 离线作业开发**。
- (2) 选择**任务开发**，在左侧目录点击创建的作业流，新建一个作业流。

2. 编辑作业

- (1) 双击作业流，进入作业流开发面板，拖拽数据同步插件，输入节点名称。
- (2) 双击打开新建的同步任务，打开同步任务页面后整个同步任务分成三步：
 - 选择数据源表 选择刚才创建的MySQL数据源中要同步的表。
 - 选择数据目标表 选择刚才创建的Hive数据表，写入方式选择**insert overwrite**。
 - 设置数据源表和数据目标表的映射管理 在映射过程中左边字段信息来自源表，右边字段信息来自目标表。
3. 点击上方**运行**进行测试，点击**停止**停止运行。
4. 点击**前往运维**，查看测试环境的运行情况。

配置调度

1. 作业流调度信息配置

- (1) 在作业流面板，点击作业流右侧**调度配置**。
- (2) 填写调度的首次生效日期，一般指定当天既可。调度周期时区选择选择北京东八区，并设置是否节假日运行状态为否表示节假日不运行作业流，以及相应的节假日日历信息。
- (3) 配置作业流的时间依赖：分为指定时间范围和指定时间点两个时间范围选项。可以指定时间点为希望运行的时间。

2. 作业调度信息配置

- (1) 双击进入创建的数据同步作业，点击作业右侧**调度配置**。
- (2) 选择调度频度为周期类，且频度选择每一天。作业就依照作业流上选定的的周期日历每日运行。

作业流发布

1. 作业流提交

- (1) 点击作业流界面**提交**按钮，选择提交的作业节点和提交备注。提交之后生成一个作业流新版本。
- (2) 点击**发布**，选择刚刚生成的作业流新版本，点击确定进行发布。

2. 发布作业流审核

- (1) 切换到**发布管理**界面查看**发布列表**查看相应的发布任务，当作业流经项目经理审批通过之后就上线了。
- (2) 切换到**运维中心 > 生产实例**页面，查看作业周期性运行的实例状态。