目录

产品概述

金山云智•离线计算提供多种编程形式，支持在线进行可视化的流程编排和代码编写，高效完成大数据离线计算的核心业务开发，通过离线计算服务，可进行离线计算的任务调度管理、运行情况监控和告警等在线运维管理。

名词解释

-**作业流** 是指一个由作业节点组成的图。每个作业节点按照配置完成一定的处理逻辑。作业节点之间通过有向边进行依赖关联，但关联时不能形成环路。一个画布中的全部作业节点及其依赖称为一个作业流。一般来说，在作业流调度模型中，作业流为调度单元，而其中的作业节点为最小粒度的执行单元。

-**作业** 作业流中的一个节点，即由用户定义的完成一定工作的逻辑单元。在任务调度模型中，作业（或任务）是最小执行单元。

-**插件** 一个作业配置模板，它包含了作业类型和该种类型作业的必要参数，通过插件创建作业时，只需要填写作业类型和必要的参数就可以完成作业的创建，可以极大的节省创建作业的时间。

-**算子** 一段可被高度提炼的逻辑，比如一段被高频率使用的SQL，算子必须依赖于插件存在，并最终可被插件解释和执行。

-**依赖包** 被作业依赖的外部资源,比如一个jar文件。

-**在线测试** 作业流提交到测试环境执行，通过ENV_ID区分，在线测试不需要作业流是发布状态，任何状态都可以测试。

-**作业测试** 同在线测试，但仅运行单个节点作业。

-**立即执行** 将作业流提交到生产环境运行，作业流状态必须是已发布状态。

-**提交调度** 将作业流提交到生产环境并按指定频率运行,作业流状态必须是已发布状态。

-**项目管理员** 项目管理员具有项目下的所有权限，可以添加或删除项目成员，项目成员又分为开发人员和运维人员。

-**运维人员** 主要负责作业流的执行、调度及审批等。

-**开发人员** 负责作业流的开发，资源维护、UDF开发等。

行业场景与资源

对大数据平台上的数据进行数据的代码编辑开发。灵活的使用多种编辑方式，包括：Shell，Python，Spark SQL等。快速进行用户业务逻辑开发。满足数据加工需求。

产品优势

丰富的任务模型

支持数据同步、SQL开发、Spark代码片段、Shell脚本、自定义插件等多种任务类型。

可视化流程编排

支持在Web界面上以可视化的方式进行批计算流程编排，创建、修改大数据处理流程

在线代码编写

支持Web在线编写SQL、Scala、Python、Shell脚本等代码。

超大规模数据处理

支持日超千亿级数据处理量承载验证和PB级数仓管理。

产品功能

多种任务流程编排

支持大数据类离线计算集群任务、非大数据类集群任务的混合编排，并保证非集群任务运行时的高可用。

高性能调度管理

提供离线计算任务调度管理，可以快速将各类依赖任务关联到相关工作流程中进行可靠执行，告别各系统之间繁琐的Crontab依赖，使工作流程可管理、可监控、可调度和高可用。

丰富的插件支持

支持SQL、Python/Scala代码片段、Shell脚本等，支持用户进行自定义插件开发。

多人协作线上开发

多人协作，在线开发，共享开发任务，同时支持开发，测试等不同环境充分隔离。

SLA监控告警

SLA监控告警包括任务执行和数据质量监控和告警。

作业开发

1. 新建文件夹 选择【任务开发】，点击如图图标，下拉菜单中找到【新建文件夹】，输入文件夹名称，选择放置文件夹的目录
2. 新建作业流 点击【新建作业流】，输入名称，在选择文件夹处选择刚刚为作业流创建的文件夹或根目录。
3. 新建解决方案 点击【新建解决方案】，输入名称，选择创建的工作流完成解决方案的创建，新建的解决方案会存放在名为【解决方案】的文件夹下。
4. 数据开发（以Spark SQL为例）

（1） 将Spark SQL拖拽至右侧，输入节点名称添加节点

（2） 双击打开，在深色区域输入指令

（3） 点击【格式化】进行自动的换行和缩进

（4） 点击右侧包依赖添加依赖包

（5） 点击【运行】进行测试，完成后点击【保存】。偷锁相关功能与数据加工处相同。

（6） 回到作业流页面，对各节点进行连线建立依赖关系，完成后点击【保存】。

作业设置

在离线计算开发过程中不同作业支持不同的作业设置。所有的作业设置有：基本信息、参数设置、调度设置、前后命令行、告警设置、包依赖、函数依赖、版本。

基本信息

基本信息功能仅仅在作业流显示，显示作业流的ID、名称、责任人和描述等。

参数设置

在数据开发过程中，可以使用参数，参数分为系统参数和用户自定义参数。对于系统参数，由系统进行赋值，具体支持的系统参数值参照数据调度产品说明文档【变量设置】部分。用户自定的参数需要用户在参数设置中进行赋值。

调度配置（作业流）

在离线计算开发过程，作业/作业流需要进行周期性的执行。离线开发与调度系统进行了无缝的整合，用户在作业/作业流开发的过程中就可以进行周期设置。调度的周期设置分为，作业和作业流两部分。其中作业流上设置调度的首次生效日期，运行周期日历，执行策略，调度时间依赖和作业流上的外部事件依赖、外部作业依赖以及外部作业流依赖等。

调度配置（作业）

作业上设置调度的信息包括频度设置，优先级设置，失败重试，逻辑资源组以及在作业上的外部事件依赖、外部作业依赖和外部作业流依赖。

前后命令行

作业设置前后处理之后将会在作业主程序之前和之后增加相应的前后处理命名。目前前后处理支持shell和Python 3.6两种命令格式。在选择开启之后，用户可以进行前后处理代码编辑。

告警设置

针对每个作业用户可以进行告警设置。告警设置目前分为两类：任务失败告警、运行时间延时告警。当作业出现相应的作业运行异常时，可以按照作业的重要程度设置告警级别是通知还是严重。通知和严重仅仅作业告警信息通知文本体现。 告警设置中支持用户进行作业返回码映射功能。返回码对应的映射对应功能包括：成功、失败（重试）、失败（不重试）。目前返回码映射功能仅支持容器类作业，且最大返回码为255。YARN类型作业暂不支持返回码设置。

包依赖

开发作业过程中如果需要引用外部资源可以在包依赖中进行相关设置。引用的包是在【资源管理】中进行上传的。 点击【增加包】弹出包选择弹窗，进行资源的选择。

函数依赖

作业开发过程中可以引用函数。函数分为系统函数和用户自定义函数。用户使用系统函数系统会自动替换不需要进行函数依赖设置。当用户引入用户自定义函数时，需要在函数依赖中添加依赖。

版本

数据开发可以进行作业和作业流的版本管理。当作业流进行提交操作的时候，作业流及其提交时包含的作业会生成一个新的版本。作业的版本生成只能依赖作业流，作业无法独立提交生成新的版本。

1. 作业流版本 用户点击作业流上【版本】可以查看当前作业流有提交过哪些版本。 点击【查看】查看当前版本作业流编辑情况。页面会跳转到一个独立的页面，页面上显示作业流版本内容。 仅可以进行查看操作，不能进行编辑再开发。 作业流版本当勾选两个版本后，可以进行作业流版本信息对比。对比两个作业流DAG图。 跳转到新页面之后，显示选中两个版本的DAG图区别。

2. 作业版本 点击作业上【版本】显示作业的版本列表。 点击【查看】跳转到新页面，查看当前版本的内容。 新页面也是只允许查看，不允许进行编辑的。 当选中两个版本之后可以进行两个版本之间的内容对比。 点击【版本对比】之后，跳转到新页面。版本对比分为两部分一部分是配置对比。此时页面上勾选的配置会议json的形式进行文本对比。 另一部分是作业编辑内容的对比。会高亮显示出来两个版本中有区别的地方。 作业上的版本允许进行回滚操作，当点击某个版本后面的代码回滚后，当前版本代码就将回滚到编辑区，可以在当前版本上进行再编辑。

开发辅助功能

任务开发

任务开发组件为用户提供了任务管理的界面。目前任务管理为三个层级的管理：作业、作业流、解决方案。作业是用户创建的单个任务。作业流是一些相关作业的集合。解决方案为作业流的集合。作业必须添加到作业流中。作业流可以放到解决方案里。

同时，还提供了【我的文件】功能，快速过滤出项目中哪些文件是我创建的。【定位】能够将右侧开发界面中任务在左侧树列表中定位出来。

库表信息

在开发过程中用户需要及时参照库表结构进行库表字段的确认。库表管理模块为用户提供及时查询字段信息和基本数据预览的功能。 选择需要的数据源、库表之后预览表信息，显示表里面有哪些字段。 切换到【数据预览】进行表中基础数据的预览。

临时查询

在数据开发过程中执行的SQL是不能直接在开发界面返回结果的，如果需要实时的将SQL运行结果显示出来，可以在临时查询中执行需要运行的SQL语句。 选择【临时查询】点击【新建临时查询】，输入查询脚本文件名称和存放路径。 新建临时查询后，自动打开临时查询交互界面。选择查询的数据源及库信息。

资源管理

开发界面上可以进行资源的管理。目前脚本资源分为两类：脚本类和jar包类。 其中脚本类包含：TXT、SQL、JSON、SHELL、PYTHON。 Jar包类包含：JAR、GZ。 脚本类作业可以进行在线的编辑开发。Jar包类只能进行上传更新。

函数管理

函数管理可以用于查看平台支持的系统函数和用户自定义函数。 系统函数包括了：数学函数、功能性函数、日期函数、条件函数、字符函数、类型转换函数、字符转换函数。每个函数说明包含了格式说明、用户、参数说明及返回值。 自定义函数可以查看用户创建了哪些自定义的，以及这些函数的变更历史情况。

作业模板管理

作业模板是一个带有参数的脚本类作业代码。用户在开发作业的过程中可以引用模板并将作业中参数在【参数设置】中进行赋值，快速创建作业。

回收站

回收站中存放着当前项目所删除的作业，用户可以选择将作业彻底删除，彻底删除的作业不可恢复，或者将作业还原到特定作业流中进行再次编辑。

常见问题

1. **什么是数据开发？数据开发包含哪些功能?**

数据开发是大数据云提供的一套离线数据脚本处理方法，全称为离线数据开发。提供了Shell，Spark SQL,Spark SQL,Python 2.7,Python 3.6,Perl以及作业模板等插件功能。帮助用户在线进行脚本开发、测试、提交、发布上线等一整套流程。

2. **目前数据开发脚本插件支持哪些数据源？**

特定类型的脚本插件支持特定类型的数据源。目前不同脚本的数据源支持情况如下：

- Spark SQL：default HIVE数据源
- Spark Shell:MySQL、Oracle数据源
- Shell：不支持选择数据源
- Python2.7\Python3.6：MySQL、Oracle数据源
- Perl：MySQL、Oracle数据源

3. **数据开发支持的插件哪些是大数据类，哪些不是大数据类？**

在插件中，大数据类插件包括：Spark SQL，Spark Shell。非大数据类插件包括：Shell，Python2.7\Python 3.6，Perl。运行大数据类作业主要使用CU资源，运行非大数据作业主要使用DCU资源

4. **数据开发对于作业的版本是如何管理？**

数据开发过程中提交/发布操作等都是以作业流的粒度进行操作，用户提交作业流时选择作业流内需要提交的作业，被提交的作业就会生成一个新的版本。可以选择两个版本进行版本间差异对比。也可以进行历史版本的回退操作。