目录

产品概述

金山云智•数据集成是金山云提供的稳定高效、弹性伸缩的批量数据ETL工具。致力于提供复杂网络环境下、异构数据源的批量可拖拽的数据同步。降低了用户数据上云以及数据开发的门槛。

名称解释

- **数据集成** 数据集成提供了一整套包括数据同步，数据加工以及业务检核的数据处理工具集合。满足多种业务场景，快速上手。

- **数据同步** 稳定高效的数据同步工具。能够在复杂的网络情况下进行异构数据源之间数据高效稳定的同步迁移。

- **数据加工** 可视化拖拽式的数据加工工具，满足不同数据源在数据加工过程中的整合，转换，聚合等。降低数据加工门槛，快速获得数据加工处理能力。

- **业务检核** 和数据质量中的业务规则无缝衔接，对数据进行全方位的规则检核。

- **数据源** 数据集成所处理的数据来源，支持多种不同类型的数据来源，且支持不同数据源之间的转换。

- **脏数据** 脏数据是指数据格式本身不符合规范，或者不满足用户定义的格式的数据。脏数据会影响干扰后续的数据处理，造成数据偏差，数据错误等。

- **插件** 插件指用户在开发界面上可操作的最小单元。一个插件相当于一个作业类型，当用户拖拽一个插件后生成一个具体的作业。

- **算子** 算子指在以拖拽形式开发的插件内部用户可进行操作的最小单元。单个算子无法进行运行，需组合成一个处理逻辑后作为一个作业整体运行。

行业场景和资源

本地数据迁移上云

使用数据集成中的数据同步服务，用户可以快速、低成本的创建面向对象存储、标准数据接口服务（JDBC适配的数据库）、NoSQL等多种数据源的数据同步任务，通过调度的周期性任务设置，企业可轻松实现不同数据源的周期性数据接入，大大降低企业本地数据上云门槛。

可拖拽式的SQL开发

使用数据集成中的数据加工功能，对于不熟悉SQL的业务人员，可以使用拖拽的形式进行可视化的SQL开发，满足日常数据分析的基本需求。

产品优势

数据源类型丰富

多种不同类型数据源传输，有效整合分散的数据资产，解决数据孤岛问题

拖拽式便捷开发

向导式，拖拽式的开发方式实现数据计算逻辑设计，零代码开发，降低使用门槛，提升开发效率

提升数据质量

对无效数据，异常数据等脏数据进行清洗，规范化等，有效提升数据质产量

保障数据安全

丰富的数据脱敏，加密等转换，加强数据安全合规

产品功能

数据同步

稳定高效的数据同步工具。能够在复杂的网络情况下进行异构数据源之间数据高效稳定的同步迁移。同步过程中同步进行数据转换，数据标准化等。

数据加工

可视化拖拽式的数据加工工具，对接多种数据源，可实现连接，过滤，采样，聚合等多种SQL操作。

业务检核

在数据管理中配置多种业务检核规则后在数据集成中周期性运行，保证上云数据质量，确保数据的可用性。

数据同步

数据同步工具不仅能够满足传统数据集成服务在复杂网络环境下进行多种异构数据源的导入导出需求，同时在数据导入导出的过程中的进行数据清洗、去重、规范化等提高数据质量。防止脏数据、垃圾数据的传播。

1. 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。

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。

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：

- 第一步选择数据源表 选择数据源的过程中可以在【数据过滤】中添加过滤语句，进行数据的增量同步。
- 第二步选择数据目标表
- 第三步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。

用户可以在源表字段上进行字段的行级信息转换：进行字段格式转换、对字段应用系统函数、常量设置等。也可以新增字段进行字段转换。

在目标表字段中可以设置默认值，如有上游有数据传输下来使用上游字段，如果上游数据为空，使用默认值设置。

源和目标之间的连线设置表示数据的流向关系。

在数据同步开发过程中可以进行变量数设置，变量设置格式为${}。 其中系统支持变量数如下：

- 日期变量

yyyyMMdd

yyyy-MM-dd

yyyy/MM/dd

yyyy

MM

dd HH

mm

ss

- 作业运行批次变量

job.batch.no

- 作业名称变量

job.name

用户在使用的时系统将自动进行变量替换。例如${yyyyMMdd}系统将替换为作业运行时的业务日期。 除了系统变量外，还可以使用自定义变量，自定义变量需要在作业【参数设置】中进行变量赋值。

数据加工

数据加工工具采用可视化拖拽的方式进行数据开发，降低开发门槛，使没有SQL经验的业务人员也能够进行快速的数据逻辑开发。

1. 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。

双击作业流，进入作业流开发面板，拖拽数据加工插件，输入节点名称。生成一个数据加工作业节点。

双击打开新建的数据加工任务，进入数据加工的开发界面。数据加工是拖拽式的开发过程，左侧显示了用户可拖拽的开发算子。

双击进入加工任务，拖动添加源表和目标表

3. 依次选择源类型-数据源-数据库-数据表,拖动添加转换算子，双击图标进行添加字段和填写功能备注,拖动连线确定关系。
4. 点击上方【运行】按钮进行测试，点击【停止】停止运行，点击【运行实例】进行查看
5. 完成后点击【保存】保存当前编辑，如果选择了【偷锁编辑】，那么在同一时间其他用户不能进行修改，点击【保存解锁】可以解除锁定。

业务检核

和数据质量中的业务规则无缝衔接，对数据进行全方位的规则检核。

1. 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
2. 点击进入【数据开发】->【离线作业开发】。
3. 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 双击作业流，进入作业流开发面板，拖拽业务检核插件，输入节点名称。 双击打开新建的业务检核作业，显示业务检核操作界面。
4. 选择一种数据源后，确定表，表上面的字段信息就会展开。如果某个字段上配置了业务检核则会在标签字段上显示检核图标。
5. 点击检核图标，弹出字段上的业务检核设置，可以看到字段上配置了哪些检核规则。也可以进行勾选确定是否在作业运行是应用某个具体的检核规则。应用业务检核后，对检核结果查看可以去【数据管理-数据质量】中查询结果。

常见问题

1. **数据集成是什么？**

数据集成是大数据云提供的一套离线数据处理加工检核等功能的开发套件，套件中包含数据同步、数据加工、数据整合、业务检核。

2. **数据集成和数据同步之间的关系是什么？**

数据集成是一整套离线数据开发工具的总称，其中包含数据同步工具。数据同步是数据集成中的一个向导式的数据迁移工具，用户可以快速进行跨源异构的数据同步操作。

3. **业务检核可以应用在哪些数据源上？**

目前支持业务检核的数据源包括：HIVE，Oracle和MySQL等关系型数据库。