目录

产品概述

金山云大数据云平台（Cubricks）数据集成是一套稳定高效、弹性伸缩的数据接入、转换、加工、检核的可视化的数据 ETL 套件，整个套件包括数据同步、数据加工、数据整合和业务检核四大功能。极大的降低了用户数据上云以及数据开发的门槛数据集成主要包括四大功能组件：

1. **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同步进行数据清洗、去重、规范化等，提高数据质量，防止脏数据、垃圾数据的传播。
2. **数据加工工具**采用可视化拖拽的方式进行数据 ETL 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。
3. **数据整合工具**结合行业经验，沉淀丰富的贴源数据处理算法，用户只需要创建特定的表结构后通过向导式的勾选就可实现数据贴源层加工。
4. **业务检核工具**与数据质量模块相结合，对数据进行数据质量，数据波动等进行统计查询，让用户了解数据质量情况。

数据集成是大数据云服务核心组件之一，定位于为大数据云项目中有离线数据的处理，包括用户线下数据的上云迁移，可视化的 ETL 加工，已经数据同步中的检核等。是离线数据处理功能组件的一个重要部分。

名词解释

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。

数据质量监控

灵活的技术检核与业务检核配置，数据传输过程中进行数据质量全程监控并生成质量报告。

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。

数据同步支持的数据源类型

- 文件存储(ks3)
- 数据库(Oracle, MySQL)
- NoSQL(HBase)
- 大数据类(HIVE, kafka)

数据同步支持的数据目标类型

- 文件存储(ks3)
- 数据库(Oracle, MySQL)
- NoSQL(HBase, Redis)
- 大数据类(HIVE, Elasticsearch)

不同数据类型写入方式不同

insert into	insert overwrite	append	其他设置
KS3（文本类型）	每次运行是进行文件覆盖	进行数据的追加写入	1.是否写入表头 选择写入的源是否有表头，需要跳过
HIVE	每次运行进行数据追加	当表有分区时，将分区数据进行替换。当表没有分区时，直接将表清空再写入	
Oracle	每次运行时进行数据追加	每次运行时将表清空再写入	
MySQL	每次运行进行数据追加	每次运行时将表清空再写入	
HBase	1.rowkey设置 在数据管理设置rowkey，这里只进行显示		

Redis	1.KeyIndex 表明+选择多列+列间间隔 2.value type和mode string->set hah->hset、hmset list->lpush、rpush、mpush set->sadd 3.写入方式 标准模式和value转key模式 4.是否设置有效时间 5.数据有效时间 是否设置有效时间选【是】显示
ES	1.doc id生成方式 拼接列：选择多列和间隔符/特定列：选择一个列/随机UUID

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。

数据加工支持的算子

名称	说明
Source算子	数据加工的数据来源，可以选择多种数据源进行数据。
Target算子	整个数据数据加工的数据目标。
Map算子	基于行级的数据项复制、修改、计算。在同行记录中可新增、减少数据项。
Filter算子	按照条件过滤掉不符合条件的行。
Sample算子	按照一定的规律抽取数据，目前只支持按照百分比进行数据抽取。
Sorter算子	对数据按照某些字段进行升序/降序的排序。
Join算子	对两个数据源进行连接操作。只支持等值连接。Join只支持连接两个数据源，如果有多个数据源进行连接，使用多个Join。
Union算子	合并两个数据源到一个结果集。与执行“UNION ALL”SQL语句结果相似，不会删除重复行。Union只支持合并两个数据源，如果有多个数据源进行合并，使用多个Union。
Aggregator算子	对多组记录进行分组聚合计算。

业务检核

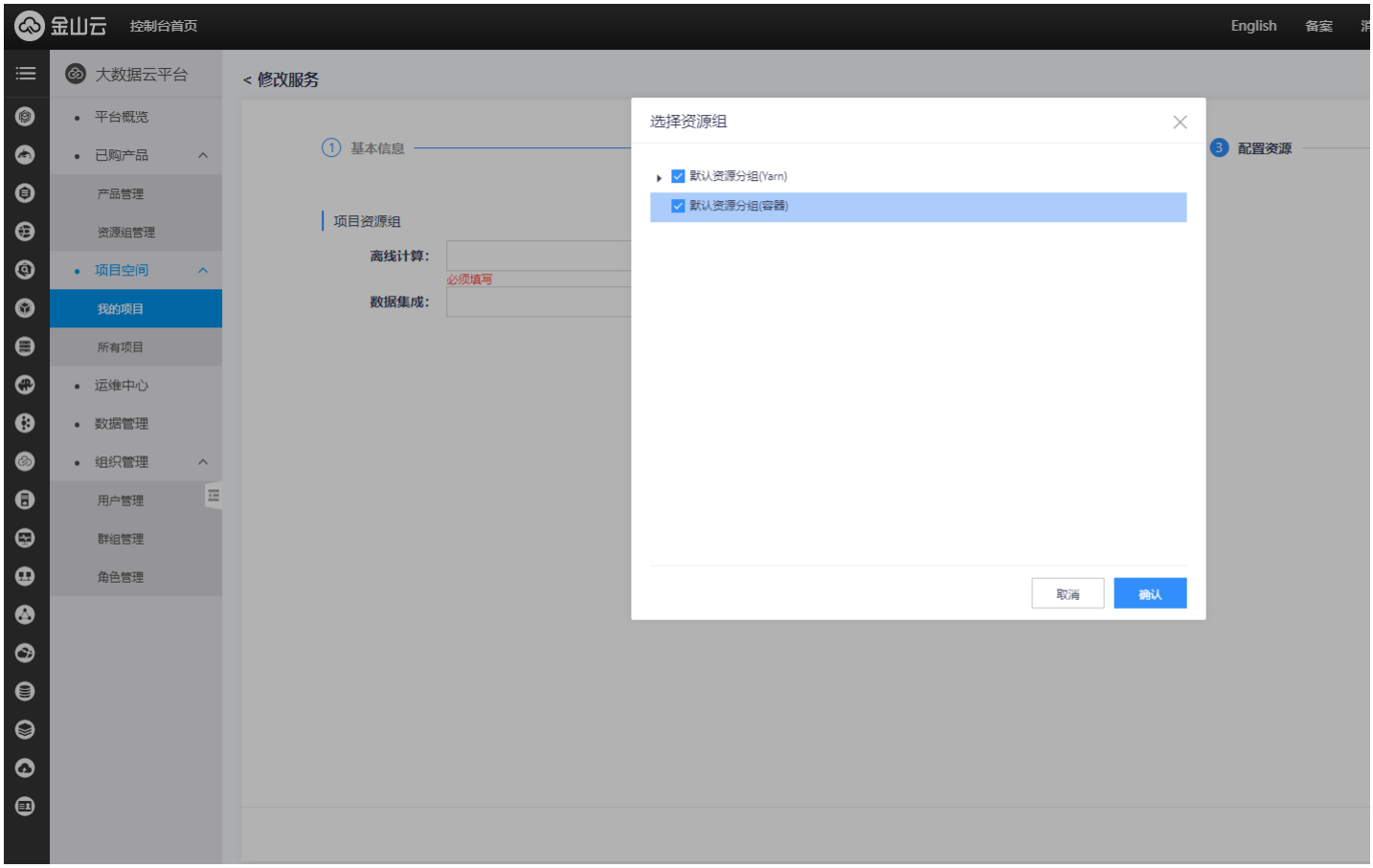
在数据管理中配置多种业务检核规则后在数据集成中周期性运行，保证上云数据质量，确保数据的可用性。

服务开通

1.登录金山云大数据云平台控制台。

2.进入**我的项目**，点击**修改服务**，购买并勾选**数据集成**服务，数据集成服务的使用依赖离线计算，请确保两个服务同时开通并勾选。

2.点击**下一步**，配置资源组，即可完成服务的开通。



3. 如果默认资源组无法满足业务需求，请到**产品管理**中进行资源的升配。

说明：

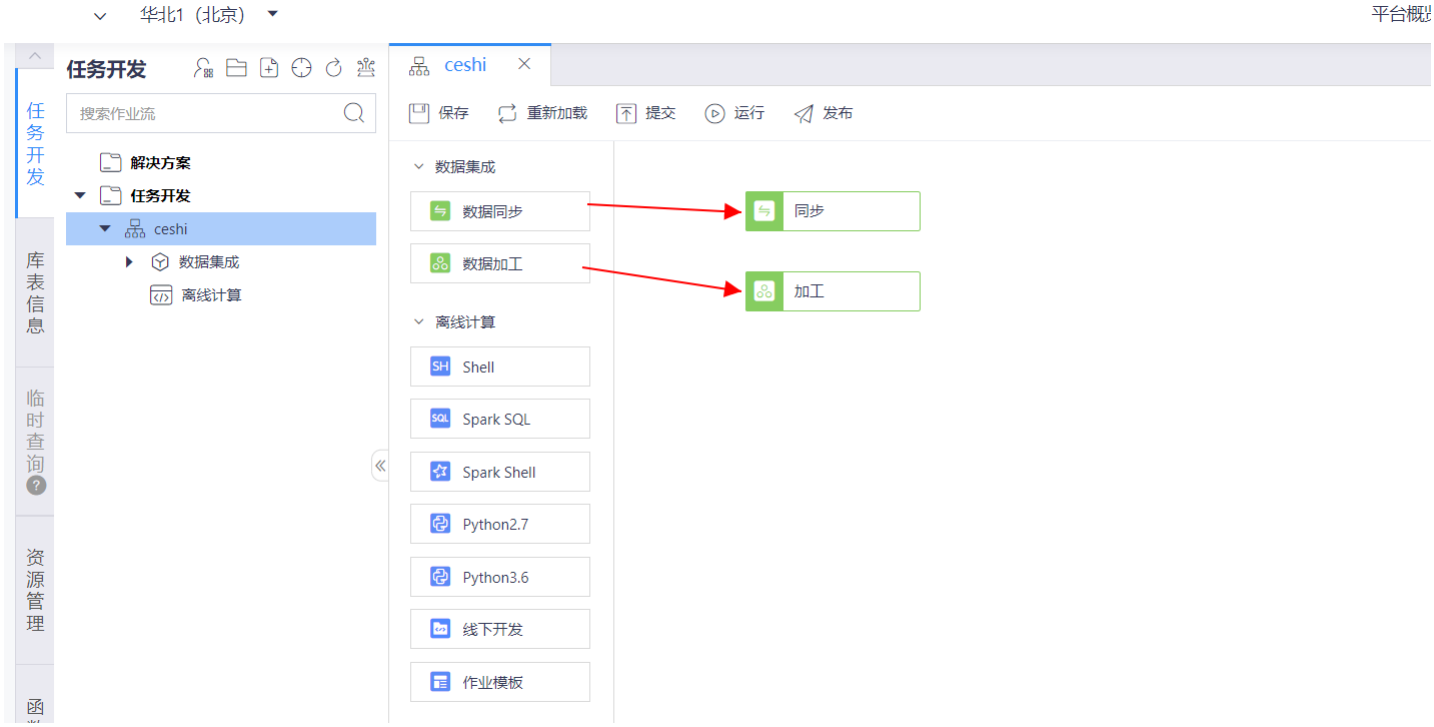
- CU对应Yarn队列资源，1CU=1C（CPU）+4GB（内存）。
- DCU对应容器资源，1DCU=1C（CPU）+4GB（内存）。
- 关于资源扩缩容：可以以CU、DCU为单位，按整数增减。
- 资源开通说明：
  （1）作业开发过程中大数据类作业使用CU资源，容器类作业使用DCU资源。
  （2）系统默认设置每个大数据作业使用3CU资源，每个容器类作业使用1DCU资源。
  （3）DCU资源最小开通值为1。按整数增减。测试DCU不小于10。生产不小于30。
  （4）CU资源开通最小为3。按整数增减。测试CU不小于15。生产不小于30。

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。
2. 点击进入**数据开发** > **离线作业开发**。
3. 选择**任务开发**，点击，新建一个作业流。

4. 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。



5. 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：

（1）选择数据源表。

- 目前源表支持的六种数据源：Oracle、MPP、HIVE、对象存储、HBASE、MySQL。选择则不同的数据源后会有各自不同的设置，详情见各不同源的设置界面。
- 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。

（2）选择数据目标表。

- 目前目标表支持：Oracle、MPP、HIVE、对象存储、Redis、HBASE、Elasticsearch、ArangoDB、MySQL。

（3）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。

数据同步-源表选择及其设置

Oracle 源

当同步源选择 Oracle 数据源时，用户可以进行数据的过滤，并且可以开启高级设置，切分设置功能。切分设置是为了加快读取 Oracle 数据进行的并行数据读取，仅支持数字类型的字段作为切分键。切分数量从 1 到 10，具体的数量用户可以根据使用运行的资源队列 CU 数的 2 至 3 倍。

HIVE 源

当数据同步选择 HIVE 数据源时，用户可以进行数据过滤。

对象存储源

当数据同步选择对象存储数据源的时候，用户可以针对文件进行多种设置。

目前数据同步针对对象存储只支持：文件结构化，JSON 半结构化两种类型。不支持非机构化的数据同步。

且因为文本中日期时间类型特殊，仅支持特定类型的日期文件，目前支持读取的日期时间格式字段为：

日期

```
yyyy-MM-dd
yyyy/MM/dd
yyyyMMdd
yyyy 年 MM 月 dd 日
```

时间

```
HH:mm:ss
HHmmss
HH:mm:ss:SSS
HHmmssSSS
```

日期时间

```
yyyy-MM-dd HH:mm:ss
yyyyMMddHHmmss
yyyy-MM-dd HH:mm:ss:SSS
yyyyMMddHHmmssSSS
yyyy/MM/dd HH:mm:ss
MM-dd HH:mm:ss
MMddHHmmss
yyyy 年 MM 月 dd 日 HH 时 mm 分 ss 秒
```

当时间日期字段出现其他类型时，同步任务读取失败。

当源文件中有表头时可以选择跳过表头设置。

如果针对的是文本结构化文件可开启技术检核。注意：除文本结构化文件可以开启检核外，其他类型文件均无开启检核功能。当开启检核后，可以设置技术检核过程中允许拒绝的数据上线。当前上线为2W，数据上线统计规则为，拒绝数据+规则为一条。也就是说如果某条数据违反了多条技术检核规则，那么算作多条数据。

HBASE 源

当同步源选择 HBASE 数据源是，用户可以进行 rowkey 高级设置和数据版本选择。

数据版本包括：版本过滤、时间戳过滤、时间段过滤

MySQL 源

当同步源选择 MySQL 数据源时，用户可以进行数据的滤，并且可以开启高级设置，切分设置功能。切分设置是为了加快读取 MySQL 数据进行的并行数据读取，仅支持数字类型的字段作为切分键。切分数量从1到10，具体的数量用户可以根据使用运行的资源队列CU数的2至3倍。

数据同步-目标表选择及其设置

Oracle 目标

当数据同步目标选择 Oracle 时，用户可以选择 insert into 和 insert overwrite 两种写入方式。	方式	说明
Insert into	每次运行进行数据追加。	
Insert overwrite	每次运行时将表清空再写入。	

HIVE 目标

当数据同步目标选择 HIVE 数据源时，用户可以进行 insert into 和 insert overwrite 写入。	方式	说明
Insert into	每次运行进行数据追加。	
Insert overwrite	当表有分区时，将分区数据进行替换。当表没有分区时，直接将表清空再写入。	

对象存储目标

当数据同步任务目标选择对象存储时，用户可以填入具体文件名称，名称可以填写变量。指定是否写入表头，确定写入文件格式等。写入方式为 append 和 overwrite。	方式	说明
Append	进行数据追加写入。	
overwrite	每次运行时进行文件覆盖。	

Redis 目标

当数据同步任务目标选择 Redis 时。

- KeyIndexs，keyIndexs 的组成方式为源表名+源表字段组合而成，选择多个源表字段后需要用分隔符进行间隔，目前支持的分割符包括:冒号(：)、逗号（，）、分号（；）、竖线（|）。
- 需要注意此处选择的字段为源表中的字段信息。
- value type 设置：支持 string、list、set、hash 四种数据类型每种数据类型对应不同的写入方式。

类型	写入方式
String	set
List	lpush、rpush
Set	sadd
Hash	hmset

写入方式：分为标准模式和 value 转 key 模式。 有效时间：用户可以对写入 redis 的数据设置有效时间。时间单位为小时。

HBase 目标

当数据同步目标选择 HBase 时，显示 HBase 的库表结构。

Elasticsearch 目标

当数据同步目标选择 Elasticsearch 时。 Doc id 生成方式。目前支持三种生成方式：拼接列、特定列、随机 UUID。	方式	说明
拼接列	选择源表的多个字段进行拼接，并选择分隔符。	
特定列	选择源表的某个字段。	
随机 UUID	使用随机数来做 doc id。	

MySQL 目标

当数据同步目标选择 MySQL 时。用户可以选择写入方式：insert into 和 insert overwrite。	方式	说明
Insert into	每次运行进行数据追加。	
Insert overwrite	每次运行时将表清空再写入。	

数据同步-源表与目标表映射

当选择完数据同步源表和目标表后，用户可以进行源表和目标表的字段映射。系统目前自动支持同名映射和同行映射。

在映射过程中用户可以对源表的字段进行字段的表达式填写，对源表字段数据进行再赋值。也可以对目标表字段进行默认值设置，当源导入为空时，使用默认值对目标字段赋值。

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**我的项目**>**数据集成**。

2. 点击 , 新建一个作业流。
3. 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。
4. 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。
5. 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。
6. 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。
7. 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。

数据加工-算子

Source算子

Source 算子支持的数据源包括：Oracle、HIVE、对象存储、HBASE、MySQL。 进入编辑态的 Source 算子会根据不同的数据源显示不同的可操作项。

操作方式

1. 拖拽 Source 算子到画板中，显示库表选择框。
2. 选择需要进行加工的库表点击确定后，Source 变为缩略态。
3. 双击 Source，显示编辑态，在编辑态中可以在过滤语句中添加过滤条件，将希望后续输出的字段**输出**进行勾选。

Target算子

Target 算子支持的数据源包括：Oracle、HIVE、对象存储、HBASE、MySQL。进入编辑态的 Target 算子会根据不同的数据源显示不同的可操作项。

操作方式

1. 拖拽 Target 算子到画板中，显示库表选择框。
2. 选择需要进行加工的库表点击确定后，target 变为缩略态。
3. 将上游算子连接到 target 算子。
4. 双击显示编辑态，在编辑态中进行上游算子字段和目标表字段的映射关系设置，并根据不同的目标源进行写入方式设置。

Aggregator算子

操作方式

1. 拖拽 Aggregator 算子到画板中，将上游算子连线到 Aggregator 算子，上游算子勾选输出的数据会同步到 Aggretator 算子中。
2. 双击 Aggregator 算子进入Aggregator 算子编辑状态。
3. 对于 Aggregator 算子需要至少有一个分组字段，增加分组字段后，再添加需要进行聚合计算的字段，下拉勾选出对字段进行 sum、avg、max、min 等聚合运算。
4. 在分组字段和聚合字段上将希望后续输出的字段**输出**进行勾选。

Filter算子

操作方式

1. 拖拽 Filter 算子到画板中，将上游算子连接到 Filter 算子，上游算子勾选输出的数据会同步到 Filter 算子中。
2. 双击 Filter 算子进入编辑状态。
3. 在 Filter 条件中添加过滤条件。将希望后续输出的字段**输出**进行勾选。

Join算子

操作方式

1. 选择 Join 算子，拖拽到工作区，生成 Join 算子缩略态。并选择两个上游算子分表连接到 Join 算子上，第一个连接的默认连线设置 0，为主表字段，第二个连接的默认连线设置为 1，为副表字段。
2. 双击 Join 算子进入编辑态。编辑主表和副表的连接关系和连接字段。

Map算子

操作方式

1. 拖拽 Map 算子到画板中，将上游算子连线到 Map 算子，上游算子勾选输出的数据会同步到 Map 算子中。
2. 双击 Map 算子进入 Map 编辑状态。
3. 可以在每行表达式中可以进行行级数据处理，如：数据类型转换，例如：`to_date(Port1,' yyyyMMdd' )`，数据项计算，例如：`(Port1+port2)/Port3`，新增变量，例如：`Port2=Port1+1` 等。将希望后续输出的字段**输出**进行勾选。

Sample算子

操作方式

4. 拖拽 Sample 算子到画板中，将上游算子连接到 Sample 算子，上游算子勾选输出的数据会同步到 Sample 算子中。
5. 双击算子进入编辑状态。
6. 在 Sample 条件中添加采样条件，按照百分比进行数据抽样。将希望后续输出的字段**输出**进行勾选。

Sorter算子

操作方式

7. 拖拽 Sorter 算子到画板中，将上游算子连接到 Sorter 算子，上游算子勾选输出的数据会同步到 Sorter 算子中。
8. 双击算子进入编辑态。
9. 在排序字段中添加需要进行排序的字段，并选择排序类型是升序还是降序。将希望后续输出的字段**输出**进行勾选。

Union算子

操作方式

10. 拖拽 Union 算子到画板中，Union 算子可以接收两个输入源。
11. 将一个上游算子拖拽到 Union 作为 Union 的第一个输入组，在选另一个上游算子拖拽到 Union 中作为 Union 的第二个输入组。
12. 第一个输入组的字段信息会显示在 Union 输出列表中，调整第一输入组，第二输入组和 Union 输出列表。需要字段类型一致。
13. 在 Union 输出列表中，将希望后续输出的字段**输出**进行勾选。

最佳实践

1. 数据同步切分键设置：当同步的数据源数据过多时，可以使用切分设置，选取某个特定字段作为切分键，并指定一定的切分数量。目前切分键字段仅支持数字类型，切分数量按照表数据量大小来指定。

2. 每日增量同步数据：在进行数据源同步时需要对源表进行每日增量数据同，此时可以在**数据过滤**中添加系统变量${BizDate},在调度周期性运行时，调度系统的业务时间会对变量赋值，从而实现了数据过滤的效果。

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。