阿里云
阿里云Elasticsearch

ElasticFlow

文档版本：20191125

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. **阿里云文档中所有内容，包括但不限于图片、架构设计、页面布局、文字描述，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
	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。	 **禁止:** 重置操作将丢失用户配置数据。
	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。
	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。
	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。	执行`cd /d C:/window`命令，进入Windows系统文件夹。
*## *	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 数据源概述

ElasticFlow通过读取数据源中的信息来拉取对应数据，再进行数据处理发送至下游。数据源与数据处理任务的配置是分开的，即可以先配置好数据源，再在配置数据处理任务时，指定配置好的数据源，完成数据的同步和处理。

数据源类型

ElasticFlow支持以下四种类型的阿里云数据源，部分类型的数据源需要进行一次访问授权，在创建数据源时可以根据控制台提示引导完成。

- **RDS MySQL**

  您可以将阿里云上的RDS实例ID、数据库名、以及具备读权限的账号密码配置到ElasticFlow中，作为一个数据源。再在创建数据同步任务时，指定该数据源中需要同步的数据表名。完成以上步骤并启动任务后，ElasticFlow将从配置的数据源表中拉取数据。

注意:

当前不支持同步视图增量数据。

- **MaxCompute**

  您可以将阿里云上的MaxCompute项目名，以及具备读权限的Access key/secret配置到ElasticFlow中，作为一个数据源。再在创建数据同步任务时，指定该数据源中需要拉取数据的表名。完成以上步骤并启动任务后，ElasticFlow将从配置的数据源表中拉取数据。

- **LogService**

  您可以将阿里云上的LogService项目名配置到ElasticFlow中，作为一个数据源。再在创建数据同步任务时，指定该数据源中需要拉取数据的表名。完成以上步骤并启动任务后，ElasticFlow将从配置的数据源表中拉取数据。

  说明:

ElasticFlow不支持LogService的Shard扩容和缩容，如果分裂或者合并了某个数据同步任务正在使用的Shard，会导致数据丢失或者同步任务出错。有关Shard的具体操作，请参见*#unique_4*。

- **Elasticsearch**

  您可以将阿里云上的Elasticsearch实例ID，以及具备读权限的Access key/secret配置到ElasticFlow中，作为一个数据源。再在创建数据同步任务时，指定该数据源中需要拉取数据的表名。完成以上步骤并启动任务后，ElasticFlow将从配置的数据源表中拉取数据。

数据源管理

在数据源列表页面，您可以创建、搜索、编辑、删除数据源。创建数据源详情请参见创建数据源（跨云服务授权）。

2 加工函数及参数说明

2.1 加工函数概述

本文档为您介绍ElasticFlow相关的加工函数，以及各函数的功能。

相关加工函数及对应功能概述列表如下。

函数名	功能概述	参考文档
CONCAT	连接两个或多个字符串值，组成一个新的字符串。	*CONCAT*
TO_TIMESTAMP	将VARCHAR类型的日期转换成TIMESTAMP类型。	*TO_TIMESTAMP*
REPLACE	字符串替换函数。	*REPLACE*
CURRENT_TIMESTAMP	返回当前UTC时间（GMT+0）的时间戳，小于北京时间8小时。	*CURRENT_TIMESTAMP*
SUBSTRING	获取字符串子串。例如从位置start开始截取长度为len的子串。	*SUBSTRING*
SPLIT_INDEX	以sep作为分隔符，将字符串str分隔成若干段，取其中的第index段。	*SPLIT_INDEX*
PARSE_URL	解析url，获取partToExtract的值。例如，当`partToExtract='QUERY'`时，可以获取url参数key的值。	*PARSE_URL*
REGEXP_REPLACE	正则匹配替换。用字符串replacement替换字符串str中正则模式为pattern的子串，返回新的字符串。	*REGEXP_REPLACE*
REGEXP_EXTRACT	使用正则模式pattern，匹配抽取字符串str中的第index个子串。	*REGEXP_EXTRACT*

2.2 CONCAT

本文为您介绍如何使用实时计算字符串函数CONCAT。

功能描述

连接两个或多个字符串值，从而组成一个新的字符串。当参数为NULL时，则跳过该参数。

语法

```
VARCHAR CONCAT(VARCHAR var1, VARCHAR var2, ...)
```

输入参数

参数	数据类型	说明
var1	VARCHAR	普通字符串值
var2	VARCHAR	普通字符串值

示例

测试数据

var1(VARCHAR)	var2(VARCHAR)	var3(VARCHAR)
Hello	My	World
Hello	null	World
null	null	World
null	null	null

测试语句

```
CONCAT(var1, var2, var3)
```

测试结果

var(VARCHAR)
HelloMyWorld
HelloWorld
World
N/A

2.3 CURRENT_TIMESTAMP

本文为您介绍如何使用日期函数CURRENT_TIMESTAMP。

功能描述

返回当前UTC时间（GMT+0）的时间戳，小于北京时间8小时。

语法

```
TIMESTAMP CURRENT_TIMESTAMP
```

示例

测试语句

```
CURRENT_TIMESTAMP
```

测试结果

```
result(TIMESTAMP)
10:02.0
```

2.4 PARSE_URL

本文为您介绍如何使用字符串函数PARSE_URL。

功能描述

解析url，获取partToExtract的值。例如，当`partToExtract ='QUERY'`时，可以获取url参数key的值。partToExtract可取HOST、PATH、QUERY、REF、PROTOCOL、FILE、AUTHORITY、USERINFO。

说明:
参数为null，则返回null。

语法

```
VARCHAR PARSE_URL(VARCHAR urlStr, VARCHAR partToExtract [, VARCHAR key
])
```

输入参数

参数	数据类型	说明
urlStr	VARCHAR	url的字符串。
partToExtract	VARCHAR	解析后获取的值。

参数	数据类型	说明
key	VARCHAR	参数名。

示例

测试数据

url1(VARCHAR)	nullstr(VARCHAR)
http://facebook.com/path/p1.php?query=1	null

测试语句

```
PARSE_URL(url1, 'QUERY', 'query') // var1,
PARSE_URL(url1, 'QUERY') // var2,
PARSE_URL(url1, 'HOST') // var3,
PARSE_URL(url1, 'PATH') // var4,
PARSE_URL(url1, 'REF') // var5,
PARSE_URL(url1, 'PROTOCOL') // var6,
PARSE_URL(url1, 'FILE') // var7,
PARSE_URL(url1, 'AUTHORITY') // var8,
PARSE_URL(nullstr, 'QUERY') // var9,
PARSE_URL(url1, 'USERINFO') // var10,
PARSE_URL(nullstr, 'QUERY', 'query') // var11
```

测试结果

var1(VARCHAR)	var2(VARCHAR)	var3(VARCHAR)	var4(VARCHAR)	var5(VARCHAR)	var6(VARCHAR)	var7(VARCHAR)	var8(VARCHAR)	var9(VARCHAR)	var10(VARCHAR)	var11(VARCHAR)
1	query=1	facebook.com	/path/p1.php	null	http	/path/p1.php?query=1	facebook.com	null	null	null

2.5 REGEXP_EXTRACT

本文为您介绍如何使用字符串函数REGEXP_EXTRACT。

功能描述

使用正则模式pattern，匹配抽取字符串str中的第index个子串。index从1开始，正则匹配提取。当参数为null或者正则不合法时，会返回null。

语法

```
VARCHAR REGEXP_EXTRACT(VARCHAR str, VARCHAR pattern, INT index)
```

输入参数

参数	数据类型	说明
str	VARCHAR	指定的字符串。
pattern	VARCHAR	匹配的字符串。
index	INT	第几个被匹配的字符串。

说明:

正则常量请按照Java代码来写，codegen会自动将常量字符串转化成Java代码。如果要描述一个数字（\d），则需要写成`'\d'`，类似于Java中的正则表达式。

示例

测试数据

str1 (VARCHAR)	pattern1(VARCHAR)	index1 (INT)
foothebar	foo(.*?)(bar)	2
100-200	(\d+)-(\d+)	1
null	foo(.*?)(bar)	2
foothebar	null	2
foothebar	无	2
foothebar	(	2

测试语句

```
REGEXP_EXTRACT(str1, pattern1, index1)
```

测试结果

result(VARCHAR)
bar
100
null
null
null

result(VARCHAR)
null

2.6 REGEXP_REPLACE

本文为您介绍如何使用字符串函数REGEXP_REPLACE。

功能描述

用字符串replacement替换字符串str中，正则模式为pattern的子串，返回新的字符串。正则匹配替换。

说明:

当参数为null或者正则不合法时，返回null。

语法

```
VARCHAR REGEXP_REPLACE(VARCHAR str, VARCHAR pattern, VARCHAR
replacement)
```

输入参数

参数	数据类型	说明
str	VARCHAR	指定的字符串。
pattern	VARCHAR	被替换的字符串。
replacement	VARCHAR	用于替换的字符串。

说明:

正则常量请按照Java代码来写，codegen会自动将SQL常量字符串转化成Java代码。如果要描述一个字符串（\d），则需要写成`'\d'`，类似于Java中的正则表达式。

示例

测试数据

str1(VARCHAR)	pattern1(VARCHAR)	replace1(VARCHAR)
2014/3/13	-	空
null	-	空
2014/3/13	-	null
2014/3/13	空	s

str1(VARCHAR)	pattern1(VARCHAR)	replace1(VARCHAR)
2014/3/13	(	s
100-200	(\d+)	num

测试语句

```
REGEXP_REPLACE(str1, pattern1, replace1)
```

测试结果

result(VARCHAR)
20140313
null
null
2014/3/13
null
num-num

2.7 REPLACE

本文为您介绍如何使用字符串函数REPLACE。

功能描述

字符串替换函数。

语法

```
VARCHAR REPLACE(str1, str2, str3)
```

输入参数

参数	数据类型	说明
str1	VARCHAR	原字符
str2	VARCHAR	目标字符
str3	VARCHAR	替换字符

示例

测试数据

str1(INT)	str2(INT)	str3(INT)
alibaba es	es	eflow

测试语句

```
REPLACE(str1, str2, str3)
```

测试结果

```
result(VARCHAR)
alibaba eflow
```

2.8 SPLIT_INDEX

本文为您介绍如何使用字符串函数SPLIT_INDEX。

功能描述

以sep作为分隔符，将字符串str分隔成若干段，取其中的第index段。index从0开始，取不到字段，则返回null。

如果任一参数为NULL，则返回null。

语法

```
VARCHAR SPLIT_INDEX(VARCHAR str, VARCHAR sep, INT index)
```

输入参数

参数	数据类型	说明
str	VARCHAR	被分隔的字符串。
sep	VARCHAR	分隔符的字符串。
index	INT	截取的字段位置。

示例

测试数据

str(VARCHAR)	sep(VARCHAR)	index(INT)
Jack,John,Mary	,	2
Jack,John,Mary	,	3
Jack,John,Mary	null	0
null	,	0

测试语句

```
SPLIT_INDEX(str, sep, index)
```

测试结果

Value(VARCHAR)
Mary
null
null
null

2.9 SUBSTRING

本文为您介绍如何使用字符串函数SUBSTRING。

功能描述

获取字符串子串。截取从位置start开始，长度为len的子串。若未指定len则截取到字符串结尾。start从1开始，start为0当1看待，为负数时表示从字符串末尾倒序计算位置。

语法

```
VARCHAR SUBSTRING(VARCHAR a, INT start)
VARCHAR SUBSTRING(VARCHAR a, INT start, INT len)
```

输入参数

参数	数据类型	说明
a	VARCHAR	指定的字符串。
start	INT	在字符串a中开始截取的位置。
len	INT	类截取的长度。

示例

测试数据

str(VARCHAR)	nullstr(VARCHAR)
k1=v1;k2=v2	null

测试语句

```
sql SUBSTRING('', 222222222) // var1, SUBSTRING(str, 2) // var2,
SUBSTRING(str, -2) // var3, SUBSTRING(str, -2, 1) // var4, SUBSTRING
```

```
(str, 2, 1) // var5, SUBSTRING(str, 22) // var6, SUBSTRING(str, -
22) // var7, SUBSTRING(str, 1) // var8, SUBSTRING(str, 0) // var9,
SUBSTRING(nullstr, 0) // var10
```

测试结果

var1(VARCHAR)	var2(VARCHAR)	var3(VARCHAR)	var4(VARCHAR)	var5(VARCHAR)	var6(VARCHAR)	var7(VARCHAR)	var8(VARCHAR)	var9(VARCHAR)	var10(VARCHAR)
无	1=v1;k2=v2	v2	v	1	无	无	k1=v1;k2=v2	k1=v1;k2=v2	null

2.10 TO_TIMESTAMP

本文为您介绍如何使用日期函数TO_TIMESTAMP。

功能描述

将VARCHAR类型的日期转换成TIMESTAMP类型。

语法

```
TIMESTAMP TO_TIMESTAMP(VARCHAR date)
TIMESTAMP TO_TIMESTAMP(VARCHAR date, VARCHAR format)
```

输入参数

参数	数据类型	说明
date	VARCHAR	无。
format	VARCHAR	默认格式为yyyy-MM-dd HH:mm:ss[.SSS]。

示例

测试数据

timestamp1(VARCHAR)	timestamp2(VARCHAR)
2017/9/15 0:00	20170915000000

测试用例

```
TO_TIMESTAMP(timestamp1) // result1
TO_TIMESTAMP(timestamp2, 'yyyyMMddHHmmss') // result2
```

测试结果

result1(TIMESTAMP)	result2(TIMESTAMP)
2017-09-15 00:00:00.0	2017-09-15 00:00:00.0

3 算子

3.1 算子类型与逻辑表

本章节为您介绍与算子相关的基本概念。

算子

任务由多个算子构成，各算子通过逻辑表串联，算子类型包含数据导入算子、数据加工算子、数据目标算子三种。

> **注意:**
> **使用算子与逻辑表时请注意以下几点：**
>
> - **一个任务只能有一个数据导入算子和数据目标算子。**
> - **流计算任务可以有多个数据加工算子，索引迁移任务不支持数据加工算子。**
> - **产出的所有逻辑表都必须使用到。**

数据导入算子

配置拉取数据的数据源，产出是逻辑表，详情请参见数据导入算子。

数据加工算子（非必须）

对逻辑表进行加工处理，例如对逻辑表进行数据过滤、对逻辑表中的某些字段通过加工函数进行转换，产出是逻辑表，详情请参见数据加工算子。

数据目标算子

设置导入到目标Elasticsearch实例中的逻辑表，详情请参见数据目标算子。

逻辑表

逻辑表由字段构成，是一个虚拟的概念，类似数据库视图。

3.2 数据导入算子

数据导入算子用于配置拉取数据的数据源，产出为逻辑表。

> **说明:**
> **数据源相关说明请参见数据源；逻辑表相关说明请参见逻辑表。**

支持数据源

数据导入算子支持RDS MySQL、MaxCompute、LogService以及Elasticsearch数据源。

在添加数据导入算子时，需要注意：

- **需要为产出的逻辑表设置主键。**
- **RDS MySQL需要在VPC网络下，数据库版本需要为5.6，暂时不支持5.7。**
- **MaxCompute必须有分区，设置数据导入算子时需填写全部分区信息，多个分区间以英文逗号分割。**
- **LogService需要用户自己手动添加逻辑表字段，并为字段设置类型。**
- **Elasticsearch只需要设置产出逻辑表名称，不需要设置逻辑表结构。**

数据导入类型

数据导入算子支持以下两种配置：

- **全量**

  从数据源中，将当前时刻已有的数据导入到目标Elasticsearch中，导入完成即任务完成。

- **增量**

  从任务运行起，一直监控数据源中的数据信息变化，同步到目标Elasticsearch。任务不会停止，除非在控制台手动停止任务。

在配置导入算子时，根据数据源类型可选组合如下。

表 3-1: 全量与增量的组合

全量	增量
MySQL	无
无	MySQL
MySQL	MySQL
MaxCompute	无
MaxCompute	LogService
无	LogService
Elasticsearch	无

3.3 数据过滤

数据过滤是根据过滤条件，从逻辑表中筛选出符合条件的数据。

过滤条件

过滤条件格式支持类似SQL WHERE的booleanExpression表达式。

连接符支持And和OR；操作符支持的类型如下。

操作符	描述
=	等于
<>	不等于
>	大于
>=	大于等于
<	小于
<=	小于等于

说明:

- 如果有数值型变量比较，需要显式的指定常数的类型。例如`x<= 5`，如果x是浮点类型，需要显式指定常数为浮点类型，例如`x<=5.0`。
- 如果有数值型变量比较，那么操作符的两侧不能都是变量。

示例

筛选条件为`City='Beijing'`，并且逻辑表数据如下所示。

Address	City
Oxford Street	Beijing
Fifth Avenue	Beijing
Changan Street	shanghai

根据以上的过滤条件和逻辑表数据，过滤出符合条件的数据，如下所示。

Address	City
Oxford Street	Beijing
Fifth Avenue	Beijing

3.4 数据加工算子

本文为您介绍，如何对数据导入算子产出的逻辑表进行加工处理，例如对逻辑表进行数据过滤、使用加工函数对逻辑表中的某些字段进行转换，最终产出为逻辑表。

说明:
数据加工算子非必须，根据实际需求进行创建。

任务详情

任务详情主要展示任务ID、任务创建时间、任务名称（支持修改）、以及任务当前状态等信息。

任务配置

任务配置主要用于添加、编辑和删除数据加工算子，也支持根据算子名称进行搜索。

添加数据加工算子

添加数据加工算子功能，可用于创建数据加工算子操作流程，详情请参见创建数据加工算子。

编辑数据加工算子

单击数据加工算子操作栏的编辑，可对数据加工算子进行编辑。该操作与创建数据加工算子流程类似，详情请参见创建数据加工算子。

删除数据加工算子

1. **选择要删除的数据加工算子。**
2. **单击操作栏下的删除。**

> **警告：**
> **数据加工算子删除后不可恢复，请谨慎操作。**

3. **在操作提示提示框中，单击确认。**

3.5 数据目标算子

数据目标算子用于设置导入到目标Elasticsearch实例中的逻辑表。目标数据源仅支持Elasticsearch。任务类型支持流计算资源和索引迁移两类。

添加流计算数据目标算子时，需要输入以下参数：

- **目标算子名称**

  为数据目标算子命名。

- 数据目标类型

  目标数据源仅支持Elasticsearch。

- 实例ID

  阿里云Elasticsearch产品对应目标实例 ID。

- 用户名

  访问阿里云Elasticsearch实例用户名（要有对应索引读写权限，默认用户名为elastic）。

- 密码

  访问阿里云Elasticsearch实例的用户对应密码。

- 引用逻辑表

  逻辑表为数据导入算子任务或数据加工算子任务产出的逻辑表。

- 目标索引

  支持现有索引和自动生成索引两种方式。

- **索引名称**

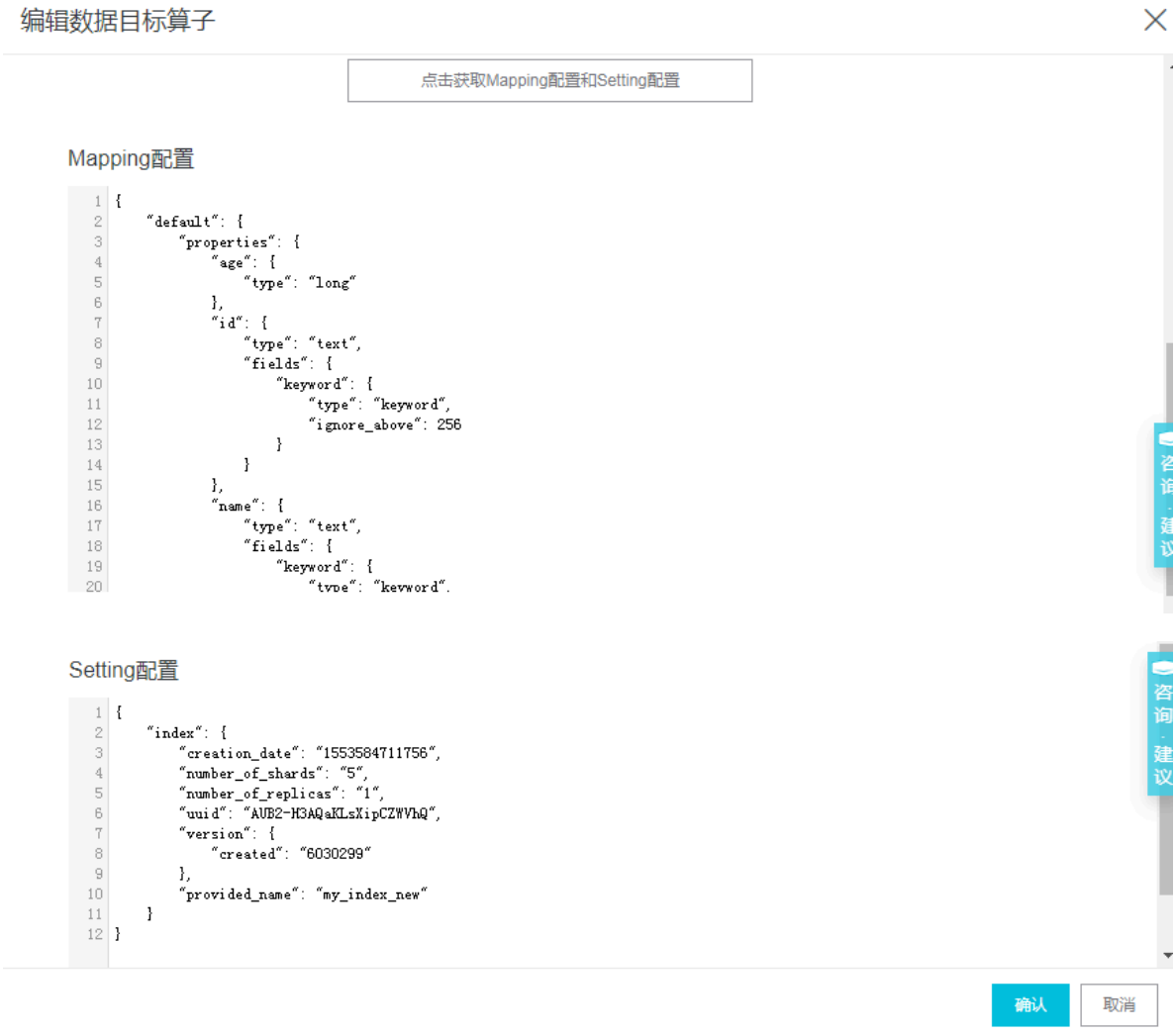
  可选择目标阿里云Elasticsearch实例中现有索引或自动生成索引。

  - **现有索引**

    选择当前索引已存在的type，继续使用已选择的索引的mapping和setting配置（不支持修改）。

  - **自动生成索引**

    根据用户选择逻辑表生成新索引的mapping配置（支持修改），新索引的setting配置默认使用系统默认配置（支持修改）。



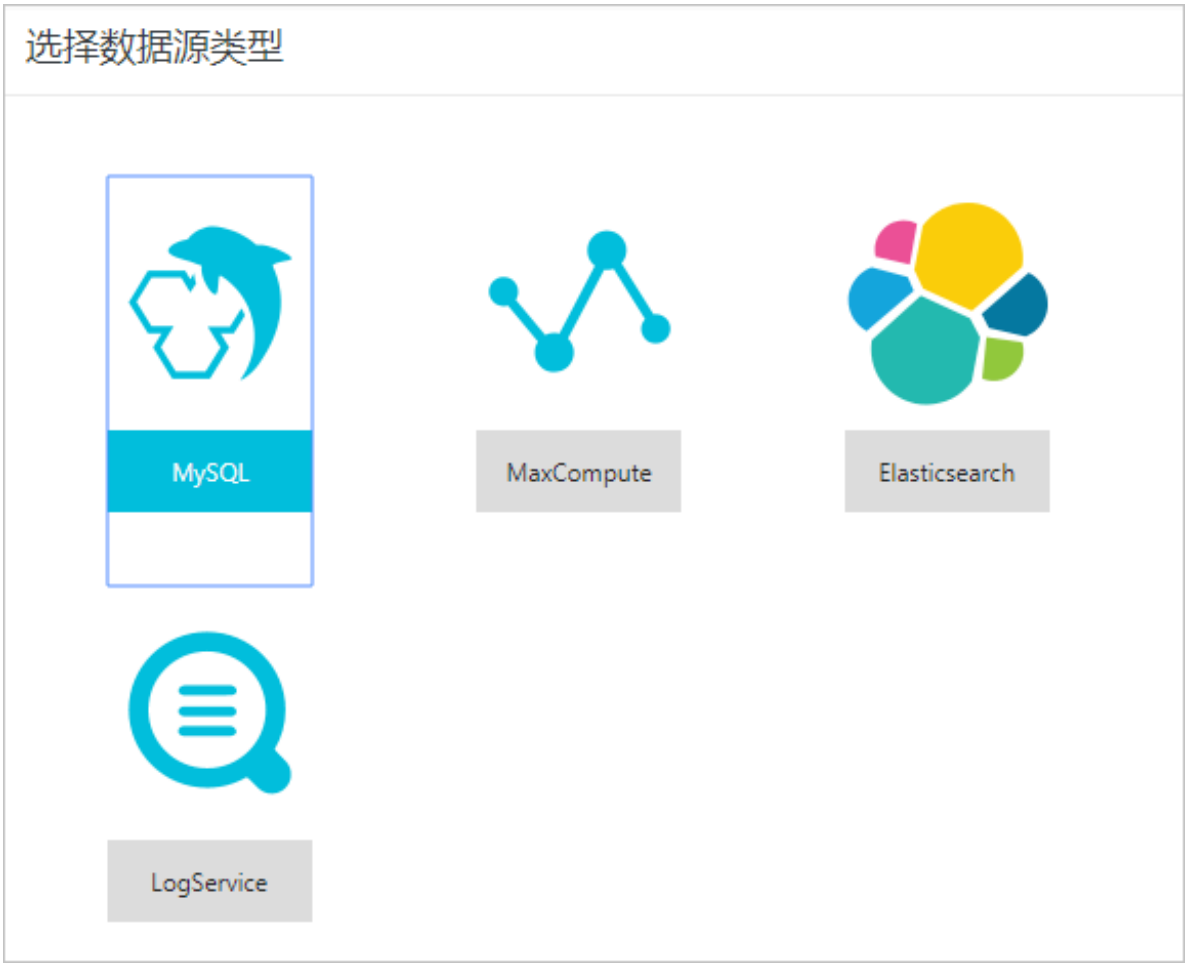
4 快速开始

4.1 创建数据源（跨云服务授权）

在数据源管理页面您可以创建数据源，并且对创建的数据源进行管理。

创建数据源

1. **进入 ElasticFlow 控制台。**
2. **单击左侧数据源列表标签页。**
3. **在数据源列表管理页面，单击创建。**
4. **选择对应数据源类型（例如MySQL），单击下一步。**



5. 在新建数据源页面，输入相关参数并单击确认，验证配置信息，通过后完成创建。

管理数据源

数据源创建完成后，您可以在数据源列表页面对数据源进行以下操作：

- 在搜索框中输入数据源ID或数据源名称搜索对应数据源。
- 创建数据源。
- 编辑数据源（对应数据源不能被任务调用，否则需要先去掉调用再编辑）。
- 删除数据源（对应数据源不能被任务调用，否则需要先去掉调用再删除）。

4.2 创建项目和任务

数据源创建完成后，您可以创建项目和任务。

创建项目

1. **在ElasticFlow控制台，单击项目列表。**
2. **在项目列表管理页面，单击创建。**
3. **在购买项目页面，选择对应区域和资源类型（例如流计算资源），并为对应项目后续将要创建的增量或全量任务分配适当的处理资源，详情请参见任务资源及分配策略。**

4. **单击创建，系统自动跳转至项目列表管理页面。**

创建任务

项目创建完成后，您可以在该项目中创建任务，任务详情请参见任务类型。

1. **在项目列表管理页面，选择要创建任务的项目，单击项目名称或任务列表。**
2. **在任务列表管理页面，单击创建。**
3. **在创建任务页面，输入任务名称单击确认**

> **说明:**
> **任务名称长度为1-30个字符，以大小字母，数字或中文开头，可以包含_或-符号。**

4.3 创建数据导入算子

任务创建完成后，您可以为对应任务配置数据导入算子。

1. **在任务列表管理页面，单击任务名称或者编辑按钮。**
2. **在任务基础信息页面，单击任务配置标签页下的添加数据导入算子。**
3. **在添加数据导入算子页面，输入以下参数：**

- **导入算子名称**
- **选择数据源**

  如果选择 RDS 或 MaxCompute 数据源，则需要配置数据导入算子，即需要通过全量还是增量方式导入数据。

  说明：

  如果选择了RDS、MaxCompute数据源，请注意以下几点：

  - **MaxCompute：除了设置对应表，还需要填写对应分区（必需）。**
  - **LogService：单击选择Logstore，选择对应Logstore。**
  - **Elasticsearch：单击选择索引，选择对应索引。**
- **选择数据表**
- **产出逻辑表表名**

  如果选择了 RDS、MaxCompute、LogService数据源，可以单击侧边栏左下角增加字段，添加需要的字段到逻辑表。单击逻辑表中的主键 列设置逻辑表主键。

  说明：

设置逻辑表时请注意以下几点：

- **Elasticsearch数据源不需设置逻辑表主键。**
- **逻辑表表名称长度为1-30个字符，支持大小写字母、数字、_。**
- **逻辑表表名不能以_开头。**
- **不能与现有逻辑表表名重复。**

4. 单击确认保存数据导入算子。

4.4 创建数据加工算子

数据导入算子创建完成后，您可以对数据导入算子产出的逻辑表进行加工处理，例如对逻辑表过滤数据、对逻辑表中的某些字段用加工函数做转换，最终产出为逻辑表。此数据加工算子非必须，请根据实际需要进行创建。

说明:

在数据加工标签中，可以添加多个数据加工算子，算子之间目前仅支持线性关系，如下图所示，一个算子的应用逻辑表为其上游算子的产出逻辑表，多个算子被逻辑表连接，形成一个链式结构。因此每个逻辑表都必须被引用，否则任务无法顺利执行。

1. 在任务列表管理页面，单击任务名称或者编辑按钮。

2. 在任务基础信息页面，单击任务配置下的数据加工标签页。

3. 单击添加数据加工算子。

4. 在添加数据加工算子页面，输入以下参数：

- 加工算子名称
- 加工算子类型

  当前版本仅支持数据加工。
- 引用逻辑表

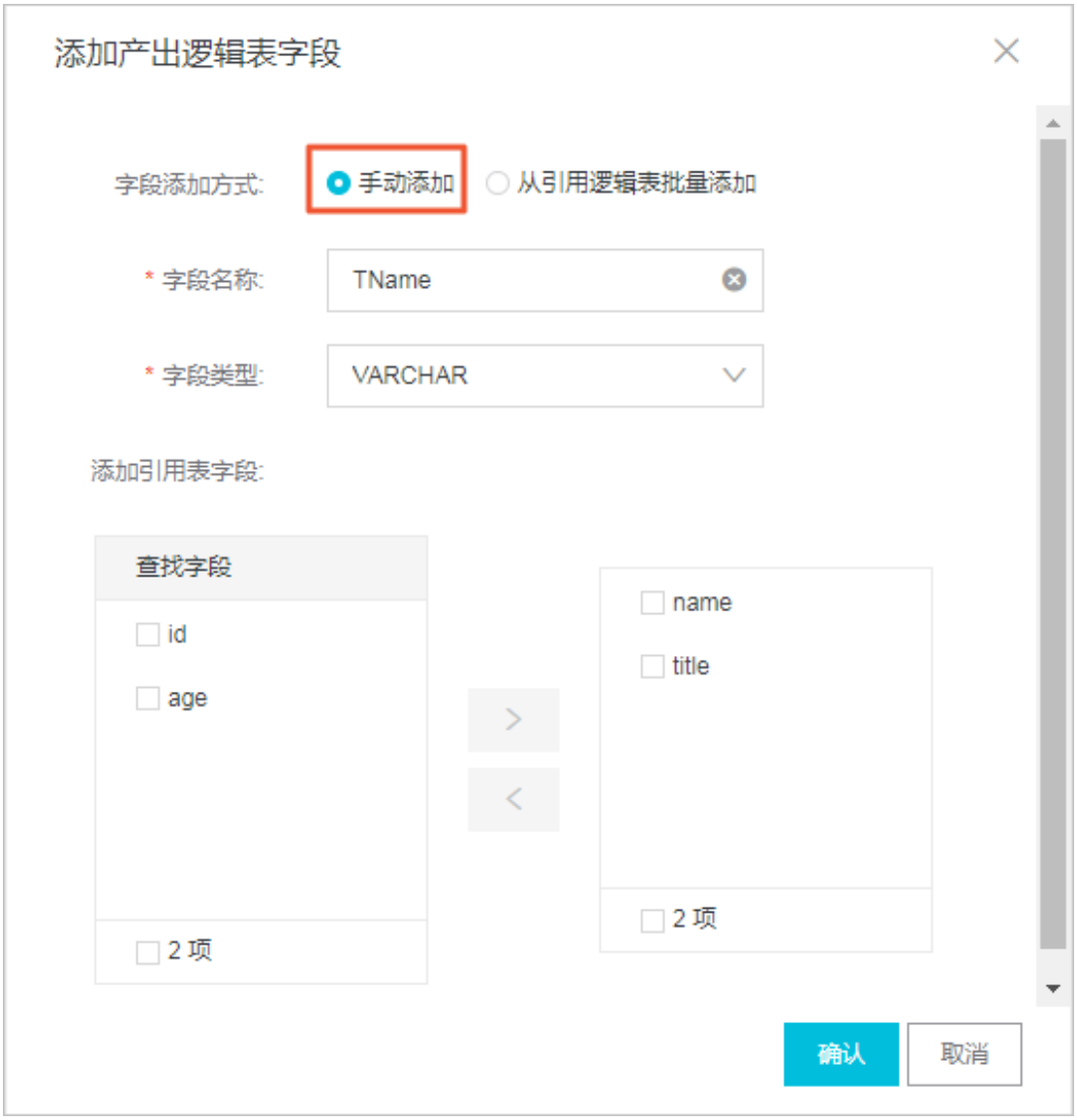
  选择一张逻辑表，可以是数据导入算子产出的逻辑表，或其它算子产出的逻辑表。
- 产出逻辑表

  为产出的逻辑表起一个名字，不可与当前已存在逻辑表重名。
- 添加数据过滤条件

  此处配置过滤出id>1的文档，相当于SQL语句中的where子句，是作用在引用逻辑表上的一个过滤条件，详情请参见数据过滤。
- 增加字段

  有两种方式向算子中添加字段，分别是“手动添加”和“从引用逻辑表批量添加”：

  - 手动添加：一次只能添加一个字段到当前算子，适用于较为复杂的数据加工场景，可以修改目标字段名称，也可以选择多个源字段。如果对应字段的数据加工配置函数中需要多个输入参数，也必须进行手动添加。



单击确认后，在算子字段列表中显示如下：

产出表字段	主键	类型	引用表字段	加工函数	操作
id	●	INT	id	—	数据加工配置 | 删除
name		VARCHAR	name	—	数据加工配置 | 删除
age		INT	age	—	数据加工配置 | 删除
title		VARCHAR	title	—	数据加工配置 | 删除
TName		VARCHAR	name title	—	数据加工配置 | 删除

+ 增加字段

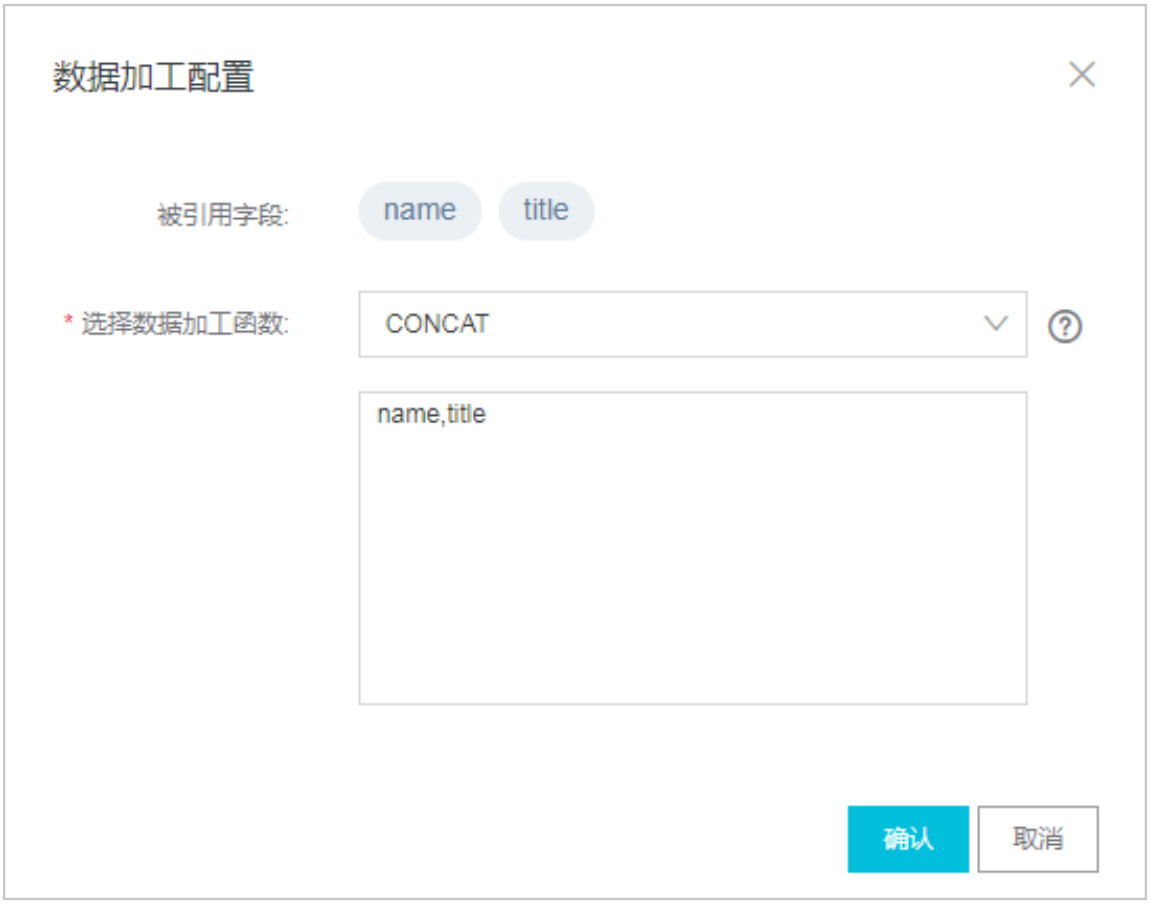
- **从引用逻辑表批量添加：可以一次性从引用逻辑表中选择多个字段添加到当前算子，其字段名称不可修改。适用于较为简单的数据加工流程，例如不需要数据加工函数，或数据加工函数仅需要一个参数的场景。**

在查找字段列勾选需要的字段，单击向右箭头，再单击确认，所选择的所有字段被添加到算子中。每个字段自动生成一个同名的目标字段，如下图所示：

数据加工配置

在添加字段完成后，单击数据加工配置开始配置。以TName为例，选择CONCAT函数，该函数的功能是对两个字段值进行拼接。

在下拉列表中选择函数CONCAT，在下方输入框中，填写该函数所需要的输入参数（即需要进行连接的 name 和 title 这两个字段名称），各参数之间用英文逗号隔开。配置完成后单击确认，如下图所示：



确认提交后，在算子字段列表中显示如下：

5. 当算子中的所有字段配置完成后，再点击右下方确认按钮保存配置。

4.5 创建数据目标算子

数据导入或数据加工算子创建完成后，您可以创建对应数据目标算子。

1. 在任务列表管理页面，单击任务名称或者编辑按钮。

2. 在任务基础信息页面，单击任务配置下的数据目标标签页。

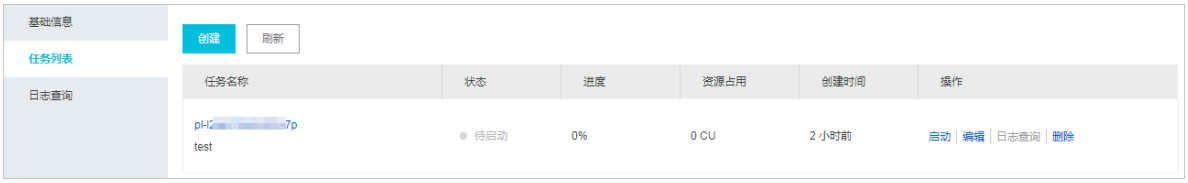
3. 单击添加数据目标算子，在添加数据目标算子页面，输入相关参数，详情见数据目标算子。

4. 单击确认。

4.6 启动任务

数据目标算子创建完成后，您可以启动对应任务运行处理。

1. 进入任务列表管理页面。



2. 选择要启动的任务，单击启动。

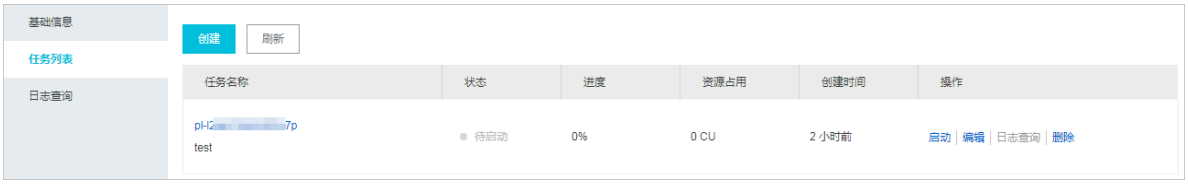
> **说明:**
> 确认当前项目中的任务资源分配是合理的（项目资源分配可参考任务资源及分配策略），以及各算子的配置信息正确，任务启动后后台会做完整性检查。

3. 单击日志查询查看已启动任务的处理进度或报错提示信息。

4.7 删除任务

当确认不再需要某个项目下的某个同步任务时，可以在对应任务列表中删除该同步任务。

1. **进入任务列表管理页面。**



2. **单击对应任务右侧操作栏下的删除。**

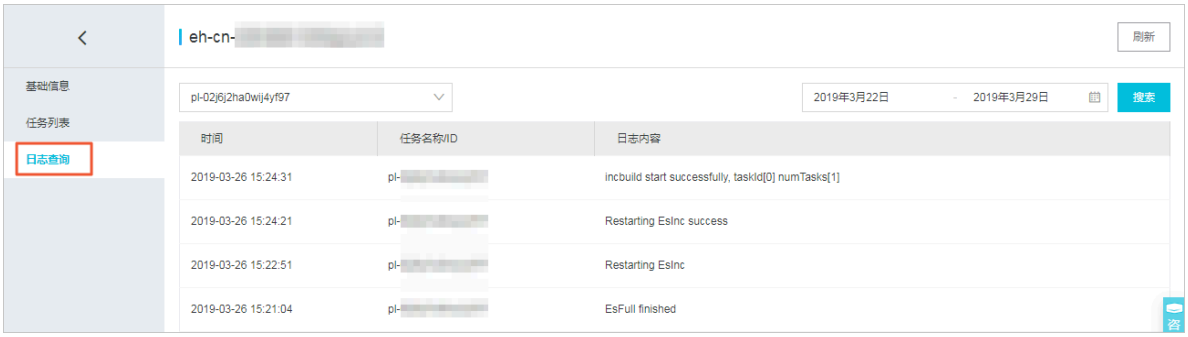
> **警告：**
> **任务删除后，相关算子配置也会同时被删除，且不可恢复，请谨慎操作。**

3. **在删除任务提示框中，单击确认。**

5 日志查询

通过日志查询，您可以查看任务运行时间、任务名称/ID、日志内容等信息。

1. 进入阿里云*Elasticsearch*控制台。
2. 单击左侧导航栏的ElasticFlow > 项目列表。
3. 在项目列表页面，单击项目名称或右侧操作栏下的项目配置。
4. 在项目配置页面，单击左侧导航栏的日志查询。



5. 在日志查询页面，选择任务ID和日期范围，单击搜索。

搜索成功后，系统会展示所选任务下，所选时间范围内的日志信息。

6 任务

6.1 任务概述

本文档为您介绍ElasticFlow任务列表页面概览，以及创建任务和操作任务的方法。

任务列表

1. **进入阿里云*Elasticsearch*控制台。**
2. **单击左侧导航栏的ElasticFlow > 项目列表。**
3. **在项目列表页面，单击右侧操作栏下的任务列表。**
4. **在任务列表管理页面，可以查看任务名称、状态、进度、资源占用、创建时间等信息。**

创建任务

在任务列表管理页面，单击创建可进行任务创建，详情请参见创建项目和任务。

操作任务

在任务列表管理页面，单击右侧操作栏下的操作项，可完成对应的操作。

- **启动：单击启动，运行对应任务。**
- **编辑：单击编辑，修改对应任务。**
- **日志查询：单击日志查询，查看对应任务运行进度和报错提示信息。**
- **删除：单击删除，删除对应任务。**

6.2 任务类型

本文档为您介绍ElasticFlow支持的任务类型，以及各任务类型的差异对比。

任务类型

ElasticFlow支持如下任务类型：

- **全量导入数据（全量）：针对RDS MySQL或者MaxComputer数据，全量拉取所有数据并建立索引，有完整的开始和结束时间。资源分配评估详情请参见任务资源及分配策略。**

- 增量导入数据（增量）：针对LogService或者DTS数据，持续的向Elasticsearch发送更新数据，具有无界（Unbounded）的特点。资源分配评估详情请参见任务资源及分配策略。

类型对比

全量	增量
支持的数据源：MaxComputer、RDS MySQL、Elasticsearch。	支持的数据源：DTS、LogService。
计算类型：批处理。	计算类型：流处理。
生命周期：有明确开始和结束时间。	生命周期：有明确的开始时间。任务不会停止，除非在控制台手动停止任务。
所需CU资源：相对较多。	所需CU资源：相对较少。

6.3 任务资源及分配策略

ElasticFlow任务拓扑结构是典型的Source->Processor->Sink模型，计费单位是CU，一个CU对应虚拟的1核CPU和4G内存。您对增量任务和全量任务指定的计算资源，会全部分配给相应的任务。为了让Source、Processor和Sink能够达到最大吞吐量（没有瓶颈节点，不会反压），可以根据本文中的策略来平衡资源分配。

增量任务分配策略

- 资源分配：各个节点和并发均匀分配（通常情况下5CU能够满足1000QPS）。
- 在Source节点：DTS采用单并发策略，即只有一个并发。LogService会根据Shard来控制并发。
- 在Sink节点：Sink节点单并发向目标Elasticsearch发送数据。

全量任务分配策略

- 资源分配：根据数据源和Elasticsearch的index动态计算。通常当资源量在20~200CU时，能够在2小时内建立100G数据的index。
- 在Source节点：MaxCompute数据源中的数据会被水平切分成多个并发。RDS考虑到数据一致性采用单并发策略，即只有一个并发。
- 在Sink节点：最大限度满足资源需求，同时考虑index的partition数量。