

知识图谱

最佳实践

文档版本 01
发布日期 2024-04-24



版权所有 © 华为技术有限公司 2024。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编： 518129

网址： <https://www.huawei.com>

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：

<https://www.huawei.com/cn/psirt/vul-response-process>

如企业客户须获取漏洞信息，请参见如下网址：

<https://securitybulletin.huawei.com/enterprise/cn/security-advisory>

目录

1 结构化数据创建图谱.....	1
2 非结构化数据创建图谱.....	17
2.1 创建图谱简介.....	17
2.2 创建信息抽取模型.....	17
2.3 使用自定义抽取模型创建图谱.....	28
3 使用公有资产库中的图谱流水线套件构建图谱.....	42

1 结构化数据创建图谱

数据是知识图谱的基础，在知识图谱服务创建知识图谱时，需要从对象存储服务（Object Storage Service，简称OBS）导入用于构建知识图谱的基础数据。知识图谱服务支持的数据包括XLSX、CSV、JSON、多行单句文本格式，详情请见[数据格式要求](#)，其中XLSX、CSV、JSON格式的数据属于结构化数据。

本章节提供一个与电影有关的结构化数据样例，帮助您快速熟悉知识图谱的创建过程。此样例通过在控制台创建本体，完成信息抽取、信息映射、信息融合、图谱质检的配置，创建一个与电影有关的知识图谱。创建完成后，您可以搜索预览相关实体的知识图谱，也可以在创建的图谱基础上全量更新或增量更新图谱。

开始使用样例前，请仔细阅读[准备工作](#)罗列的要求，提前完成准备工作。在控制台上创建知识图谱的步骤如下所示：

- [步骤1：准备数据](#)
- [步骤2：创建本体](#)
- [步骤3：选择图谱规格](#)
- [步骤4：配置数据源](#)
- [步骤5：配置图谱本体](#)
- [步骤6：配置信息抽取](#)
- [步骤7：配置知识映射](#)
- [步骤8：配置知识融合](#)
- [步骤9：配置图谱质检](#)
- [步骤10：生成图谱](#)
- [步骤11：融合验证](#)

准备工作

已[注册](#)华为账号，并完成实名认证，具体申请信息请见[账号注册](#)和[实名认证](#)。且在使用知识图谱服务前检查账号状态，账号不能处于欠费或冻结状态。

步骤 1：准备数据

在创建图谱之前，您需要将用于创建图谱的基础数据上传至OBS桶及文件夹。

1. 登录OBS服务控制台，创建桶和文件夹，用于存放样例数据集，创建名称为“kg-test”的OBS桶，名称为“demo-data”的文件夹。

创建OBS桶和文件夹的操作指导请参见[创建桶](#)和[新建文件夹](#)。为保证数据能正常访问，请务必保证创建的OBS桶与KG服务在同一区域，桶的存储类别为“标准存储”。

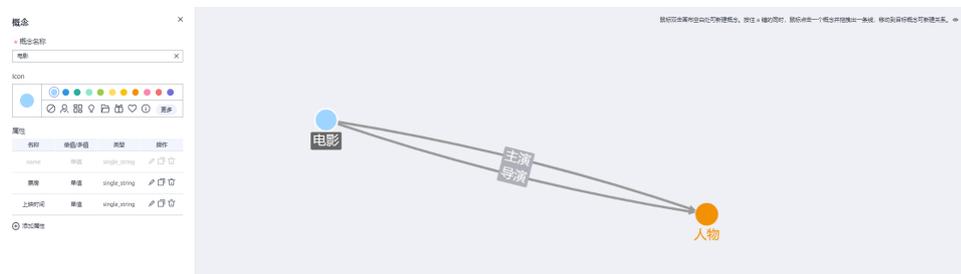
当前KG服务仅支持在“北京四”区域，建议创建OBS桶时，选择“北京四”区域。

2. 单击[数据集下载链接](#)，将数据集下载至本地，并命名为“kg_demo.xlsx”。
3. 参考[上传文件](#)，使用单个文件上传方式将“kg_demo.xlsx”文件上传至OBS路径“kg-test/demo-data”下。

步骤 2：创建本体

在创建图谱之前，您还需要提前创建本体，根据样例数据，此处本体创建2个概念和2种关系，2个概念为“电影”和“人物”，2种关系为“导演”和“主演”。

图 1-1 创建本体



1. 登录KG服务管理控制台，在左侧菜单栏中单击“我的图谱资产库 > 我的本体”，进入“我的本体”页面。

图 1-2 本体管理



2. 单击左上角“创建本体”，弹出“创建本体”对话框。
3. 在对话框中，“本体名称”文本框中填写本体名称“person_film”，在“本体描述”文本框中填写本体描述“包括电影和人物两个概念”，单击“确定”。

图 1-3 创建本体

创建本体

* 本体名称

本体描述

12/255

说明

在创建本体时，所有概念都默认有一个属性为“name”。

4. 创建两个概念节点“电影”和“人物”。
具体操作如下：
 - a. 双击界面空白处，新建概念节点。
 - b. 右键单击概念图标，选择“编辑”或双击概念图标，左侧弹出概念编辑框。
 - c. 在概念编辑框中，按表1-1编辑“概念名称”、“Icon”和“属性”，单击“保存”。

图 1-4 编辑概念

概念
×

* 概念名称

电影
×

Icon

🚫
👤
📏
💡
📁
🎁
❤️
ℹ️
更多

属性

名称	单值/多值	类型	操作
name	单值	single_string	✎ 📄 🗑️
上映时间	单值	single_string	✎ 📄 🗑️
票房	单值	single_string	✎ 📄 🗑️

⊕ 添加属性

保存
取消

表 1-1 编辑概念

参数	说明	推荐填写
概念名称	概念名称只能由大小写字母、数字、中文、下划线组成，长度为1-63位。	本样例需要创建两个概念，分别可命名为“电影”和“人物”。

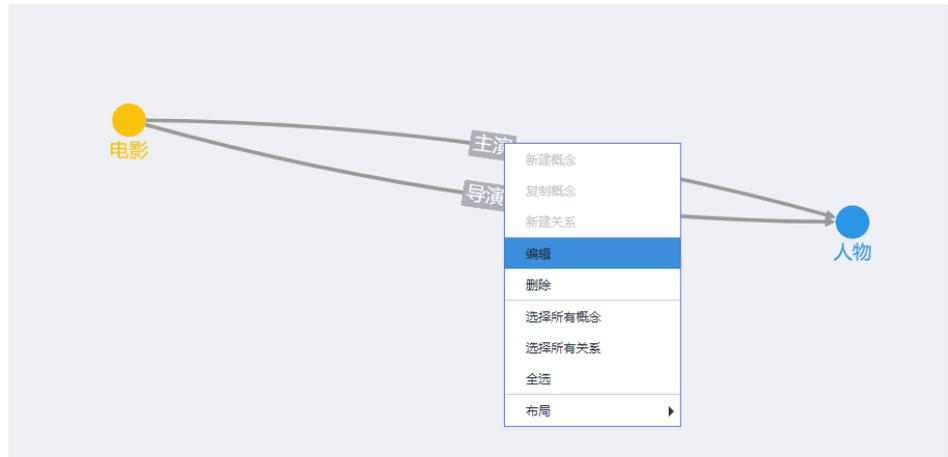
参数	说明	推荐填写
Icon	单击选择概念图标的颜色和样式，单击右侧的“更多”，可选择更多图标样式。	-
属性	<p>编辑当前概念的属性。所有概念都默认有一个属性为“name”，且默认属性类型为“single_string”。</p> <ul style="list-style-type: none"> ● 添加属性 单击“添加属性”可添加属性文本框，在文本框中填写属性的“名称”，选择“单值/多值”、“类型”、“操作”。 <ul style="list-style-type: none"> - 名称：属性名称只能由大小写字母、数字、中文、下划线组成，长度为1-63位。 - 单值/多值：选择当前属性是单值还是多值。 - 类型：选择属性的类型。 单值的可选类型有“single_string”、“single_int”、“single_double”、“single_bool”。 多值的可选类型有“set_string”、“set_int”、“set_double”。 <p>填写完概念的属性后，单击“操作”列的 ，保存信息。</p> <ul style="list-style-type: none"> ● 复制属性 单击已有属性“操作”列的 ，复制当前属性。 ● 删除属性 单击已有属性“操作”列的 ，删除当前属性。 	<ul style="list-style-type: none"> ● 概念“电影”新增属性“票房”和“上映时间”。 ● 概念“人物”新增属性“国籍”、“职业”和“出生日期”。

5. 创建“电影”与“人物”之间的关系。

具体操作如下：

- 按住键盘a键，同时鼠标单击概念节点“电影”，拖拽出一条线，移动至目标概念“人物”。
或鼠标右键单击一个概念节点“电影”，选择“新建关系”，然后拖拽出一条线，移动至目标概念“人物”。
- 右键单击关系曲线，选择“编辑”。
右侧弹出关系编辑框。

图 1-5 编辑关系



- c. 在关系编辑框中，填写“关系名称”，只能由大小写字母、数字、中文、下划线组成，长度为1-63位。
本样例需创建两次“电影”与“人物”之间的关系，关系名称分别为“主演”和“导演”。

图 1-6 编辑关系

The screenshot shows a form titled '关系' (Relationship) with a close button (X). Below the title is a red asterisk and the label '* 关系名称' (Relationship Name). A text input field contains the text '主演' (Main Actor) and has a close button (X) on its right side.

- d. 单击“保存”。

步骤 3：选择图谱规格

1. 登录KG服务管理控制台，默认进入“我的图谱”页面。
2. 单击+，然后在创建图谱页面填写信息。

图 1-7 服务选型

The screenshot shows the '创建图谱' (Create Graph) page. At the top, there are three steps: '1 服务选型' (Service Selection), '2 版本确认' (Version Confirmation), and '3 完成' (Complete). The '1 服务选型' step is active. Below the steps, there are three fields: '* 图谱名称' (Graph Name) with an empty text input; '* 图谱规格' (Graph Specification) with four buttons: '体验版 一万边' (Experience Edition 10,000 Edges), '标准版 百万边' (Standard Edition 1,000,000 Edges), '标准版 千万边' (Standard Edition 10,000,000 Edges), and '高级版 千万边' (Advanced Edition 10,000,000 Edges); and '是否购买套餐包' (Purchase Package) with a toggle switch.

表 1-2 服务选型参数说明

参数	说明
“图谱名称”	待创建图谱的名称，名称只能由大小写字母、数字、中文、下划线组成，长度为1-63位。
“图谱规格”	待创建图谱的规格大小。当前仅支持在控制台选择“体验版 一万边”，如果根据业务需要，选择标准版 百万边、标准版 千万边、高级版 千万边，请通过官网产品页“服务咨询”联系我们进行咨询购买。
“是否购买套餐包”	体验版图谱默认不能开通套餐包。如果需要选择业务版图谱，请通过官网产品页“服务咨询”联系我们进行咨询购买。

- 单击右下角“下一步”，在“版本确认”页签确认版本信息。
- 单击右下角“确认创建”。

页面提示“知识图谱创建任务提交成功”。

单击“返回”，返回至“我的图谱”页面，您会看到新创建的图谱卡片，“运行状态”为“创建中”。等待十几分钟后，运行状态变为“初始化”状态。

图 1-8 创建图谱



步骤 4：配置数据源

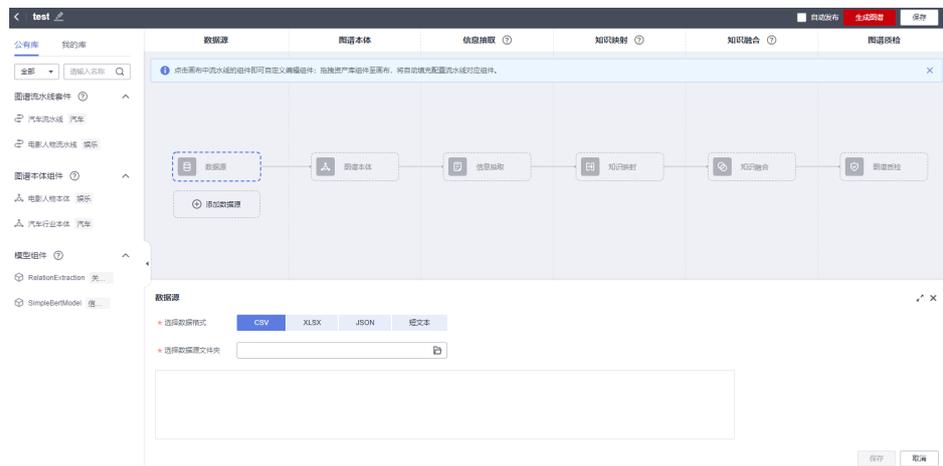
- 在“我的图谱”页面鼠标移至新创建的图谱卡片，单击“图谱构建 Workspace”。
- 进入图谱流水线构建页面。

图 1-9 图谱构建 WorkSpace



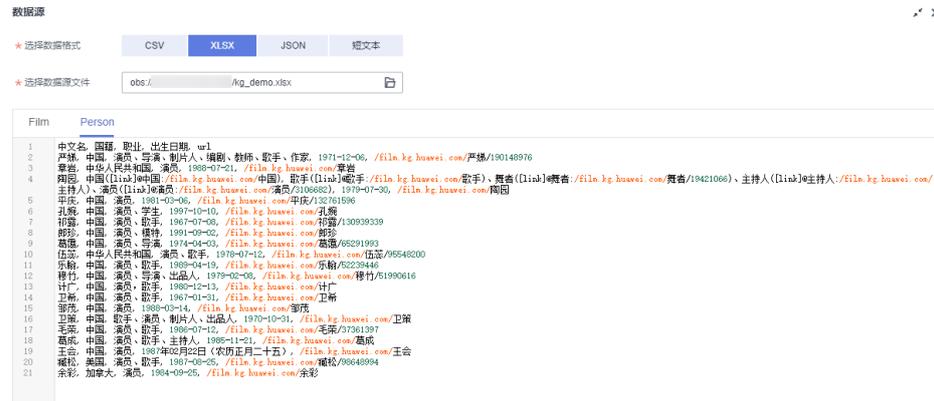
2. 在流水线构建页面，单击“数据源”，页面下方弹出数据源配置对话框，单击右侧  按钮可以放大对话框。

图 1-10 配置数据源



3. 在数据源配置对话框，填写相关信息。
 - “选择数据格式”：此样例选择“XLSX”。
 - “选择数据源文件”：单击 ，弹出“选择数据源文件”对话框，选择数据源存放在OBS的路径：
 - i. “OBS桶”：选择数据源文件存放的OBS桶“kg-test”。确保您使用的OBS桶与KG服务在同一区域，桶的存储类别为“标准存储”。
 - ii. “存储路径”：选择数据源文件存放在OBS桶的文件路径“demo-data/kg_demo.xlsx”。
 选择的OBS路径不能是加密路径，否则可能会访问失败。
 - iii. “OBS桶授权”：如果OBS桶未授权，请勾选“确认授权”。
 选择完成后，单击“确定”。

图 1-11 选择数据源文件



4. 填写完信息后，单击“保存”，完成数据源的配置。

步骤 5：配置图谱本体

1. 在流水线构建页面，单击“图谱本体”，页面下方弹出图谱本体配置对话框，单击右侧 按钮可以放大对话框。
2. 选择左侧“我的库”页签，在“图谱本体组件”中拖拽 **步骤2：创建本体** 中创建的 **本体“person_film”** 到虚线框中，单击“保存”，完成本体选择。

说明

“我的库”页签下“图谱本体组件”呈现的是“我的图谱资产库”中“我的本体”页面创建或OBS导入的所有本体。

可单击页面右下角的“编辑”，进入本体修改页面，修改后的本体将保存至“我的库”。

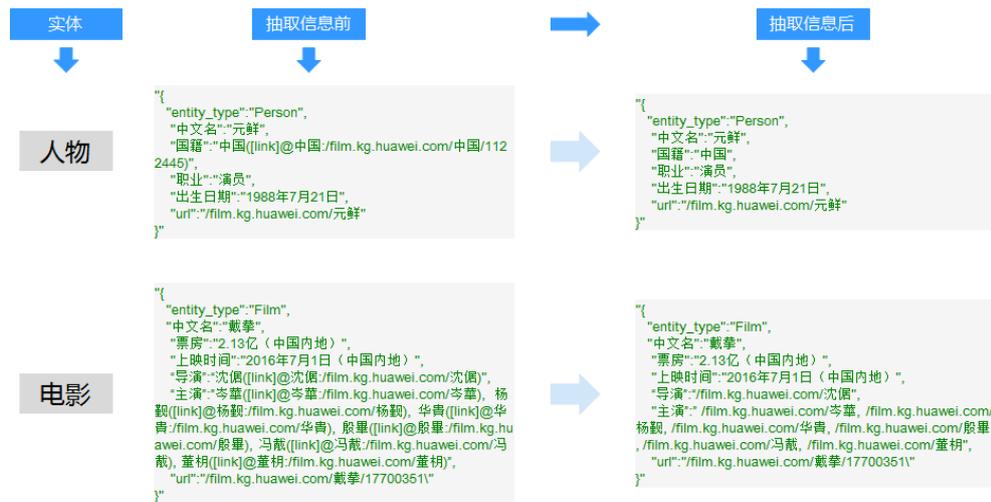
图 1-12 图谱本体



步骤 6：配置信息抽取

以两条数据为例，抽取信息前后的实体信息如图1-13所示，展示如何配置信息抽取。

图 1-13 信息抽取



1. 在流水线构建页面，单击“信息抽取”，页面下方弹出“信息抽取”对话框，单击右侧 按钮可以放大对话框。
2. 在“信息抽取”对话框中，“抽取方式”默认为“结构化抽取”，“编辑方式”选择“交互界面”，关闭“默认抽取”开关。
3. 在信息抽取页面填写信息抽取项，按表1-3填写。

表 1-3 信息抽取项示例

实体类型	数据字段	字段函数
Person	identifier	`\${url}`
	中文名	<code>regexp_replace(`\${中文名}`, '\\([link]@.*?:/film.*?)\\', '')</code>
	国籍	<code>regexp_replace(`\${国籍}`, '\\([link]@.*?:/film.*?)\\', '')</code>
	职业	`\${职业}`
	出生日期	`\${出生日期}`
	url	`\${url}`
Film	identifier	`\${url}`
	中文名	`\${中文名}`
	票房	`\${票房}`
	上映时间	`\${上映时间}`
	导演	<code>regexp_extract_all(`\${导演}`, '\\([link]@.*?:/film.*?)\\', 1)</code>
	主演	<code>regexp_extract_all(`\${主演}`, '\\([link]@.*?:/film.*?)\\', 1)</code>

实体类型	数据字段	字段函数
	url	\${url}

字段函数主要有三类，示例如下：

- `${职业}`
直接引用“职业”字段的值，不做任何操作。
 - `regexp_replace(${国籍}, '\\([link]@.*?:/film.*?)\\)', '')`
将“国籍”字段中符合正则模式的“`\\([link]@.*?:/film.*?)\\)`”替换成空字符串，即删除符合这个pattern的字符串。例如从字段“`中国([link]@中国:/film.kg.huawei.com/中国/1122445)`”中抽出信息“中国”。
 - `regexp_extract_all(${导演}, '\\([link]@.*?:/film.*?)\\)', 1)`
抽出导演字段中，与模式“`\\([link]@.*?:/film.*?)\\)`”匹配的数据值，返回的类型是一个列表。例如“`导演：“沈倨([link]@沈倨:/film.kg.huawei.com/沈倨)”`”，抽取信息后为“`导演：“/film.kg.huawei.com/沈倨”`”。
- 填写完“信息抽取项”，单击“确定”，完成信息抽取配置。

4. 填写完信息后，单击“保存”。

步骤 7：配置知识映射

以其中两条数据为例，知识与实体之间的映射关系如图1-14所示，展示如何配置知识映射。

图 1-14 知识与实体的映射关系



1. 在流水线构建页面，单击“知识映射”，页面下方弹出“知识映射”对话框，单击右侧 按钮可以放大对话框。
2. 在“知识映射”对话框下方，填写知识映射的相关信息。
 - 单击实体“电影”对应的 图标。
 - 填写实体“电影”映射前的抽取项：“Film”。
 - 填写映射到实体“电影”的抽取项，如图1-15和表1-4所示。

图 1-15 实体电影的抽取项



表 1-4 实体电影的抽取项

属性/关系名称	抽取项
name	中文名
票房	票房
上映时间	上映时间
电影:主演:人物	主演
电影:导演:人物	导演

- 单击实体“人物”对应的图标。
- 填写实体“人物”映射前的抽取项：“Person”。
- 填写映射到实体“人物”抽取项，如图1-16和表1-5所示。

图 1-16 实体人物的抽取项

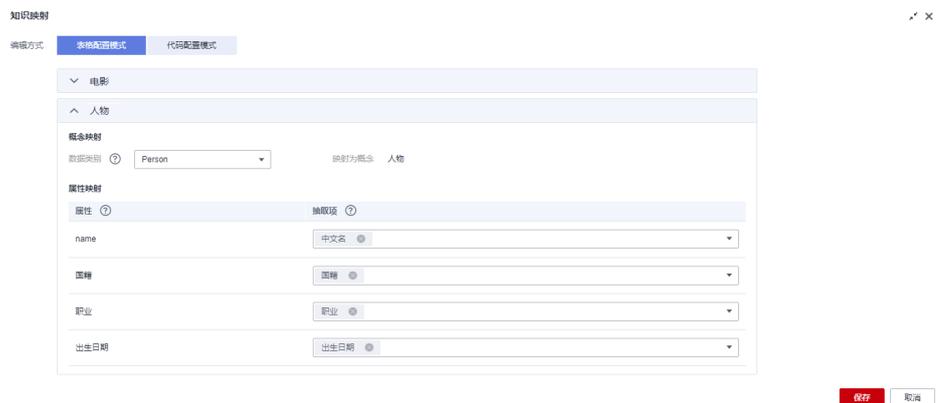


表 1-5 实体人物的抽取项

属性	抽取项
name	中文名
国籍	国籍
职业	职业
出生日期	出生日期

3. 单击“保存”，完成知识映射配置。

步骤 8：配置知识融合

以如下融合条件为例：

对电影的属性“name”相似的数据进行知识融合判断，当属性“name”的相似度大于等于0.90，且属性“上映时间”的相似度大于等于0.90时，数据进行融合。对人物的属性“name”相似的数据进行知识融合判断，当属性“name”的相似度大于等于0.90，且属性“职业”的相似度大于等于0.90时，数据进行融合。相似度函数均为编辑距离函数“EditSim”。

展示如何配置知识融合。

1. 在流水线构建页面，单击“知识融合”，页面下方弹出“知识融合”对话框，单击右侧  按钮可以放大对话框。
2. 在“知识融合”对话框，打开“知识融合”开关。
3. 在左侧本体预览图中单击实体“电影”图标。
4. 在实体“电影”对应的“融合标识符”栏单击 ，选择标识知识融合的本体属性“name”。
5. 填写实体“电影”的属性配置项信息。

可添加多组属性配置项，每组属性配置项可添加多个判断融合的属性配置。至少存在一组属性配置项，组内所有判断融合的属性相似度均大于相似度阈值时，即可融合这两条数据。

本样例只需添加一个属性配置分组，一组里填写两个属性配置项信息，如图1-17所示。

单击“添加分组”，即可添加一组新的属性配置项。

单击“增加属性配置项”，即可在组内添加新的属性配置项。

实体配置完成后，单击“确定”。

图 1-17 实体电影的知识融合属性



6. 在左侧本体预览图中单击实体“人物”图标。
7. 在实体“人物”对应的“融合标识符”栏单击 ▾，选择标识知识融合的本体属性“name”。
8. 填写实体“人物”的属性配置项信息。

可添加多组属性配置项，每组属性配置项可添加多个判断融合的属性配置。至少存在一组属性配置项，组内所有判断融合的属性相似度均大于相似度阈值时，即可融合这两条数据。

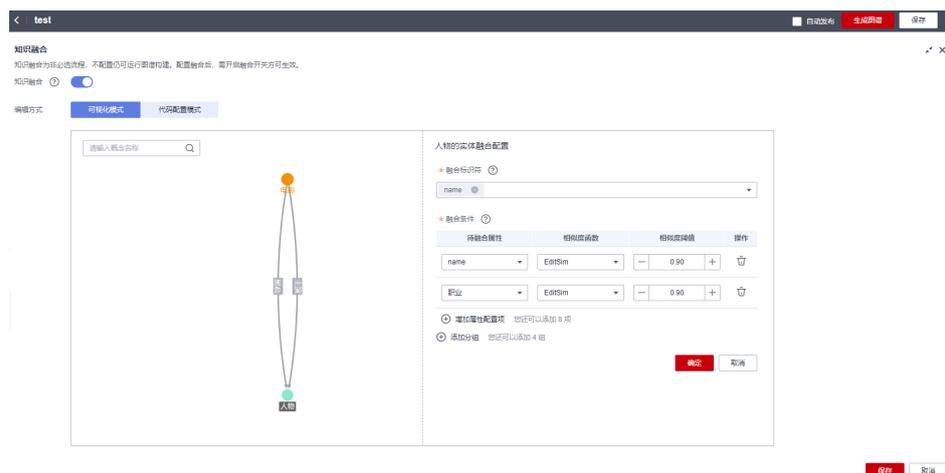
本样例只需添加一个属性配置分组，一组里填写两个属性配置项信息，如图1-17所示。

单击“添加分组”，即可添加一组新的属性配置项。

单击“增加属性配置项”，即可在组内添加新的属性配置项。

实体配置完成后，单击“确定”。

图 1-18 实体人物的知识融合属性



9. 单击“保存”。

步骤 9：配置图谱质检

1. 在流水线构建页面，单击“图谱质检”，页面下方弹出“图谱质检”对话框，单击右侧 🔍 按钮可以放大对话框。

图 1-19 图谱质检



2. 在“图谱质检”对话框，打开“图谱质检”开关，单击“保存”。

说明

图谱质检为非必选流程，关闭和开启状态均不影响图谱构建。开启质检任务后，后台会对当前版本的知识图谱的知识质量和流水线构建进行评估，并生成质检报告，详见[查看质检报告](#)。

步骤 10：生成图谱

在图谱流水线页面，单击右上角“生成图谱”，完成图谱的创建。

图 1-20 生成图谱



跳至新建图谱“test”的详情页，查看图谱的状态，当图谱的“状态”由“运行中”变为“可用”后，即可单击“图谱预览”，预览图谱。

步骤 11：融合验证

1. 登录KG服务管理控制台，单击左侧导航栏“我的图谱”，在“我的图谱”页面单击图谱“test”，进入图谱详情页。
2. 在“图谱版本”页签，等待图谱运行完成，单击“操作”列的“验证”，进入“随机验证”页面。

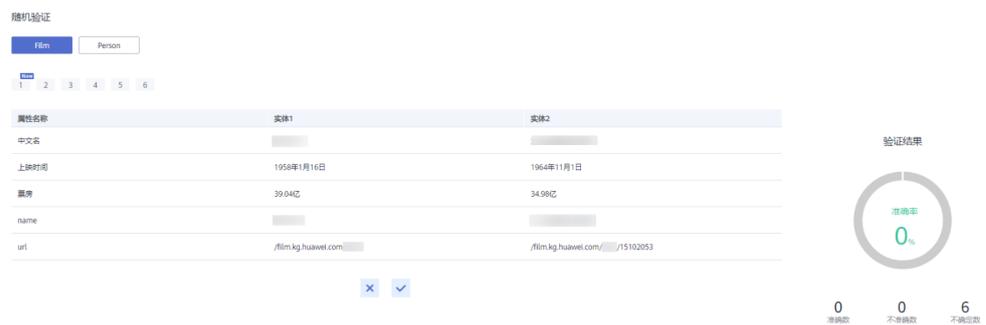
- 在随机验证页面，有随机采样的部分融合样例，您如果认可融合的结果，单击 ，若不认可，单击 。

右侧可查看融合验证的结果，包含“准确率”、“准确数”、“不准确数”、“不确定数”。

说明

- 如果随机验证页面为空页面，说明没有需要验证的融合样例。
- 如果首次构建图谱，版本会自动发布，无需手动发布。

图 1-21 融合验证



2 非结构化数据创建图谱

2.1 创建图谱简介

数据是知识图谱的基础，在知识图谱服务创建知识图谱时，需要从OBS导入用于构建知识图谱的基础数据。知识图谱服务支持的数据包括XLSX、CSV、JSON、多行单句文本格式，详情请见[数据格式要求](#)，其中多行单句文本格式的数据属于非结构化数据。

本章节提供一个与人物、电影有关的非结构化数据自定义信息抽取模型并使用自定义抽取模型创建知识图谱的流程，帮助您快速熟悉知识图谱自定义信息抽取模型创建过程和使用非结构化数据创建图谱的过程。步骤如下：

1. **创建信息抽取模型**：以在ModelArts控制台上标注数据为例，介绍信息抽取模型的构建流程。
2. **使用自定义抽取模型创建图谱**：介绍通过自定义的信息抽取模型创建知识图谱的流程。

2.2 创建信息抽取模型

如果您在创建知识图谱时，选择使用自定义模型进行信息抽取，您需要在创建知识图谱之前，创建信息抽取模型。

本章节提供一个与人物、电影有关的非结构化数据构建信息抽取模型流程，帮助您快速熟悉知识图谱自定义信息抽取模型创建过程，通过在控制台进行数据标注、模型训练，构建一个与人物、电影有关的信息抽取模型。本样例以在ModelArts控制台上标注数据为例，介绍信息抽取模型构建流程。

在创建模型之前，请仔细阅读[准备工作](#)罗列的要求，提前完成准备工作。在控制台上创建信息抽取模型的步骤如下所示：

- **步骤1：准备数据**
- **步骤2：创建数据集**
- **步骤3：标注数据**
- **步骤4：发布数据集**
- **步骤5：创建模型**
- **步骤6：创建模型版本**

- [步骤7: 发布版本](#)
- [步骤8: 创建图谱](#)

准备工作

已[注册](#)华为账号，并完成实名认证，具体申请信息请见[账号注册](#)和[实名认证](#)。且在使用知识图谱服务前检查账号状态，账号不能处于欠费或冻结状态。

步骤 1: 准备数据

在创建信息抽取模型之前，您需要自行准备基础数据并上传至OBS桶及文件夹中。

由于数据安全原因，本样例不提供具体的样例数据，仅提供样例数据要求，要求如下：

数据格式

数据格式为短文本txt格式，文件内容示例如下：

张三的生日是1990年1月1日，身高175cm，出生于北京。

李四，著名导演，毕业于电影学院，代表作有《电影1》、《电影2》。

... ..

数据划分

训练数据：80%数据用于训练信息抽取模型。

验证数据：20%数据用于验证信息抽取模型性能（即F1、P、R性能指标）。

训练数据量

训练一个基本可用的模型，标注数据量总数需要大于模型版本配置中的“batch_size”，且需大于20*三元组类型数量。

训练一个效果较好的模型，建议提供2万条以上的短句数据作为训练数据。

三元组类型

本样例构建的模型适用于人物、电影领域的信息抽取，限定抽取的三元组类型如[表2-1](#)所示。

表 2-1 三元组类型

subject_type	predicate	object_type
人物	毕业院校	学校
影视作品	改编自	作品
影视作品	制片人	人物
人物	身高	Number
人物	丈夫	人物
影视作品	主演	人物

subject_type	predicate	object_type
人物	出生日期	Date
影视作品	编剧	人物
影视作品	导演	人物
人物	祖籍	地点
人物	父亲	人物
人物	出生地	地点
影视作品	出品公司	企业
人物	妻子	人物
人物	母亲	人物
人物	国籍	国家
影视作品	上映时间	Date

- 按上文要求准备待标注的训练数据和模型数据，并保存至本地，训练数据命名为“annotation.txt”，模型数据命名为“corpus.txt”。
- 登录OBS服务，创建桶和文件夹，用于存放数据集。
创建名称为“kg-model”的OBS桶。
创建名称为“model-data”的文件夹用于存放未标注的数据集。
创建名称为“model-data-out”的文件夹用于存放已标记并发布的数据集。
创建名称为“graph-data”的文件夹用于存放创建图谱的数据集。
创建OBS桶和文件夹的操作指导请参见[创建桶](#)和[新建文件夹](#)。为保证数据能正常访问，请务必保证创建的OBS桶与KG服务在同一区域，桶的存储类别为“标准存储”。
- 参考[上传文件](#)，使用单个文件上传方式将本地准备的训练数据“annotation.txt”文件上传至OBS路径“kg-model/model-data”下，将本地准备的创建模型的数据文件“corpus.txt”上传至OBS路径“kg-model/graph-data”。

步骤 2：创建数据集

- 登录[ModelArts管理控制台](#)，在左侧菜单栏中单击“数据管理 > 数据集”，进入数据集页面，单击页面上方“前往新版”。

图 2-1 数据集



说明

- 如果数据集页面上方显示“返回旧版”，则说明当前页面已经是新版页面。
- 单击“创建数据集”，在“基本信息”页签，参考图2-2填写信息。

图 2-2 创建数据集

“名称”：填写“RelationExtraction”。

“描述”：填写简单的数据集描述，可不填。

“数据格式”：选择“文件型”。

“数据类型”：选择“文本”。

“数据集输出位置”：单击 选择输出数据集的OBS路径“/kg-model/model-data-out/”。

- 单击“下一步”，在“数据输入”页签，参考图2-3填写信息，然后单击“创建”。

数据集创建完成后，系统自动跳至数据集页面，您可以查看数据集详情。

“数据源”：默认为“OBS”。

“导入方式”：默认为“目录”。

“导入路径”：单击 选择存放待标记数据集的OBS路径“/kg-model/model-data/”。

“数据标注状态”：选择“未标注”。

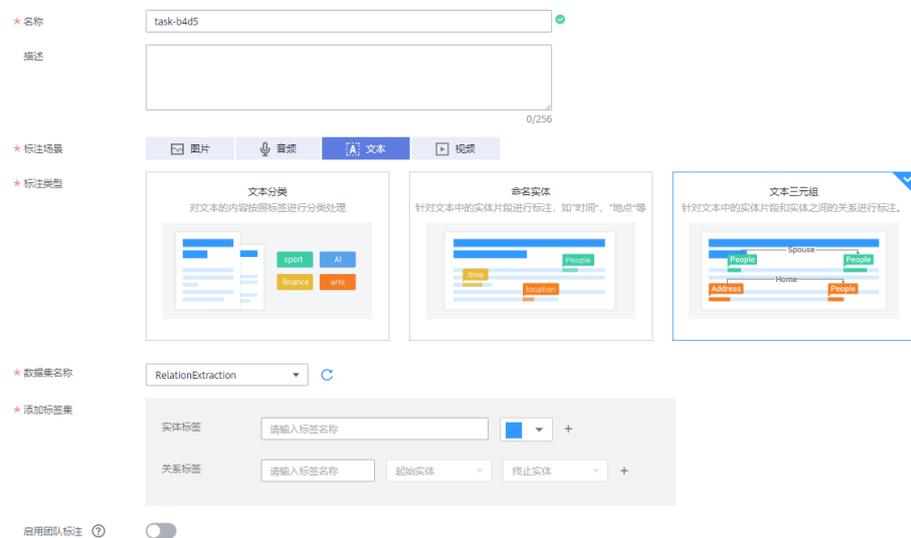
图 2-3 创建数据集

步骤 3：标注数据

- 在ModelArts管理控制台的左侧菜单栏中单击“数据管理 > 数据标注”，进入数据标注页面。

2. 单击“创建标注作业”，进入创建标注作业页面。
3. 参考图2-4填写信息，然后单击“创建”。
 - “名称”：自动生成作业名称，也可以自定义修改。
 - “描述”：填写简单的数据集描述，可不填。
 - “标注场景”：选择“文本”。
 - “标注类型”：选择“文本三元组”。
 - “数据集名称”：选择“RelationExtraction”。如果选不到，可以单击图标刷新后再选。
 - “添加标签集”：按表2-1添加实体标签和关系标签，实体标签需要添加表2-1中的“subject_type”和“object_type”，关系标签需要添加表2-1中的“predicate”，如图2-5和图2-6所示。
 - “启用团队标注”：打开开关会启动团队标注，团队成员可以进行该数据集的标注任务，可不操作。

图 2-4 创建标注作业



The screenshot shows a web form for creating an annotation task. It has the following sections:

- Name:** A text input field containing "task-b4d5".
- Description:** A large text area, currently empty.
- Annotation Scenario:** A set of tabs for "Image", "Audio", "Text", and "Video". The "Text" tab is selected.
- Annotation Type:** Three preview cards:
 - Text Classification:** Shows a document icon and labels like "sport", "AI", "finance", "arts".
 - Named Entity:** Shows a document icon and labels like "time", "location", "people".
 - Text Triplets:** Shows a diagram with nodes like "People", "Spouse", "Address", "Home", "People" connected by lines.
- Dataset Name:** A dropdown menu showing "RelationExtraction" with a refresh icon.
- Add Tag Set:** Two rows of input fields:
 - Entity Tag:** A text input field with a dropdown and a plus sign.
 - Relationship Tag:** A text input field, a dropdown menu with "Start Entity" and "End Entity" options, and a plus sign.
- Enable Team Annotation:** A toggle switch that is currently turned off.

图 2-5 实体标签



图 2-6 关系标签

关系标签	源类型	目标类型	操作
毕业院校	人物	学校	+ 删除
改编自	影视作品	作品	+ 删除
制片人	影视作品	人物	+ 删除
身高	人物	Number	+ 删除
丈夫	人物	人物	+ 删除
主演	影视作品	人物	+ 删除
出生日期	人物	Date	+ 删除
编剧	影视作品	人物	+ 删除
导演	影视作品	人物	+ 删除
祖籍	人物	地点	+ 删除
父亲	人物	人物	+ 删除
出生地	人物	地点	+ 删除
出品公司	影视作品	企业	+ 删除
妻子	人物	人物	+ 删除
母亲	人物	人物	+ 删除
国籍	人物	国家	+ 删除
上映时间	影视作品	Date	+ 删除

4. 单击“创建”后自动返回数据标注页面。在标注作业列表中单击作业名称，进入作业标注详情页，默认显示“未标注”的文本列表。
单击“同步新数据”，快速将数据集中的数据添加到标注作业中。
5. 按下文步骤标注文本数据集。
 - a. 在“未标注”页签文本列表中，页面左侧罗列“标注对象列表”。在左侧列表中单击需标注的文本对象，在右侧文本框选中相应文本内容，在页面呈现的实体类型列表中选择实体名称，完成实体标注。

图 2-7 实体标注



- b. 在完成多个实体标注后，鼠标左键依次单击起始实体和终止实体，在呈现的关系类型列表选择一个对应的关系类型，完成关系标注。

图 2-8 关系标注



- c. 单击下方“保存当前页”。

步骤 4：发布数据集

1. 在 **ModelArts 管理控制台** 的左侧菜单栏中单击“数据管理>数据集”，进入数据集页面。
2. 在数据集列表中，单击数据集“RelationExtraction”操作列的“发布”。或者，您可以单击数据集名称，进入数据集概览页，在页面右上角单击“发布”。

3. 在“发布新版本”的弹出框中，填写“版本名称”、“标注类型”、“描述”，然后单击“确定”。
“版本名称”：默认输入“V001”。
“标注类型”：选择“文本三元组”。
“描述”：可不填。

图 2-9 发布新版本



发布新版本

* 版本名称

⚠️ 启用数据切分功能时，每种标签已标注的样本数不少于5个；若有多标签样本，多标签样本数需不少于2个。详细信息可在数据集概览页查看。

* 标注类型

描述

0/256

4. 自动跳转至版本管理页面，记录V001.manifest文件的存储路径，如图2-10红框所示。

图 2-10 存储路径



步骤 5: 创建模型

1. 登录KG服务管理控制台，在左侧菜单栏中选择“我的图谱资产库 > 我的模型”，进入模型管理页面。
2. 在模型列表左上方，单击“创建模型”。
弹出“创建模型”对话框。

图 2-11 创建模型

创建模型

* 模型名称

* 模型模板

* 数据类型

* OBS桶

* 文件夹 kg-mode... / model-data-o... / RelationExtraction-b1CzzY... / annot... /
.. 返回上一级

文件
<input checked="" type="radio"/> <input type="checkbox"/> V001/

描述

3. 在“创建模型”对话框中填写相关信息，如图2-11所示，单击“确定”，完成模型创建。
 - “模型名称”：在“模型名称”文本框中填写模型名称“Person_Film_Model”。
 - “模型模板”：单击 ▾ 选择模型模板“MRC-BM-v2”，MRC-BM-v2的基本概念请参见[训练模型框架介绍](#)。
 - “数据类型”：单击 ▾ 选择“ModelArts”，训练数据类型详情请参见[训练数据类型介绍](#)。
 - “OBS桶”：选择在[步骤1：准备数据](#)新建的OBS桶“kg-model”。
 - “文件夹”：选择在4中记录的V001.manifest文件的存储路径，并勾选存放的文件夹。

步骤 6：创建模型版本

1. 在KG服务管理控制台的左侧菜单栏中选择“我的图谱资产库 > 我的模型”，进入“模型管理”页面。
2. 在模型列表中，单击在[步骤5：创建模型](#)所创建的模型“Person_Film_Model”，进入模型详情页。
3. 在版本管理区域单击“创建新版本”。
弹出“创建新版本”对话框。
4. 在“创建新版本”对话框中如[图2-12](#)填写模型参数的参数值。
一般18类三元组、2000条训练数据，训练迭代次数（epochs）在30次左右可以正确的抽取得到部分三元组，训练迭代次数（epochs）在100次左右模型达到最优解。

图 2-12 创建新版本

创建新版本

* 模型参数

参数	参数值
batch_size ?	32
epochs ?	1
max_len ?	192
learning_rate ?	0.00005

描述

请输入描述

0/255

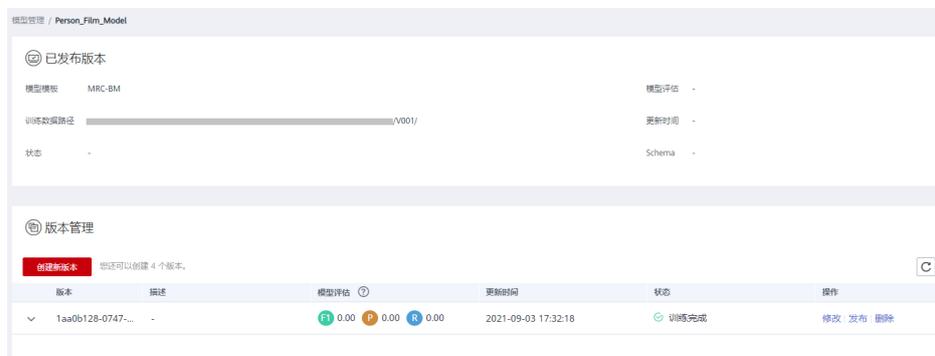
确定 取消

- 单击“确定”，完成模型版本的创建。
在模型详情页，您可以看到模型版本的“状态”处于“训练中”，等到“状态”变为“训练完成”，就可以执行下一步[步骤7：发布版本](#)。

步骤 7：发布版本

- 在KG服务管理控制台的左侧菜单栏中选择“我的图谱资产库 > 我的模型”，进入“模型管理”页面。
- 在模型列表中，单击在[步骤5：创建模型](#)所创建的模型“Person_Film_Model”，进入模型详情页。
- 在版本管理列表中的“操作”栏，单击“发布”。

图 2-13 发布版本



待状态栏处于“已发布”状态，该版本的模型即可在创建图谱过程中对文本进行信息抽取，即可执行下一步[步骤8：创建图谱](#)。

步骤 8：创建图谱

详细步骤请参见[使用自定义抽取模型创建图谱](#)。

2.3 使用自定义抽取模型创建图谱

本章节以有关于人物和电影的非结构化数据为例，提供一个使用自定义的信息抽取模型创建知识图谱的流程，帮助您快速熟悉使用非结构化数据和自定义信息抽取模型创建图谱的过程。

首先，请仔细阅读[准备工作](#)罗列的要求，提前完成准备工作。然后在控制台上创建信息抽取模型和知识图谱，步骤如下：

- [步骤1：创建本体](#)
- [步骤2：选择图谱规格](#)
- [步骤3：配置数据源](#)
- [步骤4：配置图谱本体](#)
- [步骤5：配置信息抽取](#)
- [步骤6：配置知识映射](#)
- [步骤7：配置知识融合](#)
- [步骤8：配置图谱质检](#)
- [步骤9：生成图谱](#)

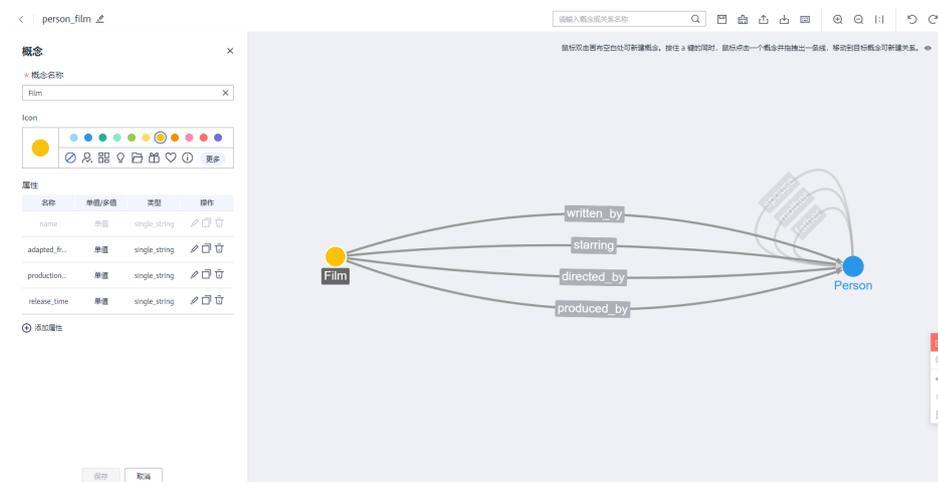
准备工作

- 已[注册](#)华为账号，并完成实名认证，具体申请信息请见[账号注册](#)和[实名认证](#)。且在使用知识图谱服务前检查账号状态，账号不能处于欠费或冻结状态。
- 已自定义信息抽取模型，详情请见[创建信息抽取模型](#)。

步骤 1：创建本体

在[创建信息抽取模型](#)中自定义的信息抽取模型需要从本章节待创建的图谱数据中抽取三元组信息，因此我们在创建本体时可以参考[表2-1](#)中的三元组类型。

图 2-14 创建本体



1. 登录KG服务管理控制台，在左侧菜单栏中选择“我的图谱资产库 > 我的本体”，进入“本体管理”页面。
2. 单击左上角“创建本体”，弹出“创建本体”对话框。
3. 在对话框中，“本体名称”文本框中填写本体名称“person_film”，在“本体描述”文本框中填写本体描述“包括电影和人物两个概念”，单击“确定”。

图 2-15 创建本体

创建本体

* 本体名称

本体描述

12/255

确定 取消

4. 创建两个概念节点“Film”和“Person”。
具体操作如下：
 - a. 双击界面空白处，新建概念节点。
 - b. 右键单击概念图标，选择“编辑”或双击概念图标，左侧弹出概念编辑框。
 - c. 在概念编辑框中，按表2-2编辑“概念名称”、“Icon”和“属性”，单击“保存”。

图 2-16 编辑概念



表 2-2 编辑概念

参数	说明	推荐填写
概念名称	概念名称只能由大小写字母、数字、中文、下划线组成，长度为1-63位。	本样例需要创建两个概念，分别可命名为“Film”和“Person”。
Icon	单击选择概念图标的颜色和样式，单击右侧的“更多”，可选择更多图标样式。	-

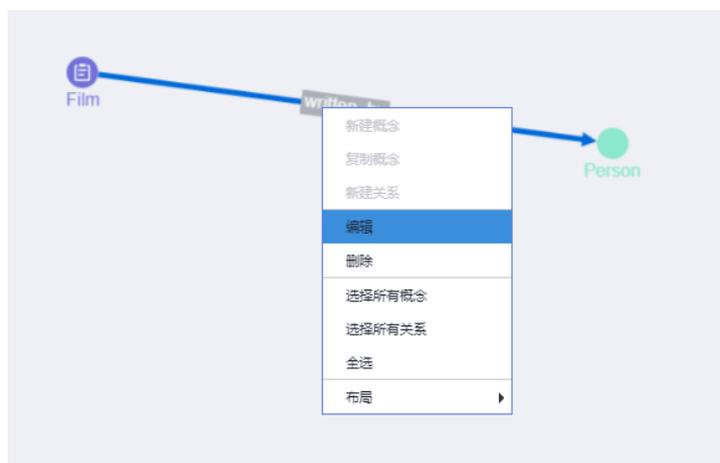
参数	说明	推荐填写
属性	<p>编辑当前概念的属性。所有概念都默认有一个属性为“name”，且默认属性类型为“single_string”。</p> <ul style="list-style-type: none"> ● 新增属性 单击“新增属性”可添加属性文本框，在文本框中填写属性的“名称”，选择“单值/多值”、“类型”、“操作”。 <ul style="list-style-type: none"> - 名称：属性名称只能由大小写字母、数字、中文、下划线组成，长度为1-63位。 - 单值/多值：选择当前属性是单值还是多值。 - 类型：选择属性的类型，可选类型有“single_string”、“single_int”、“single_double”、“single_bool”。 <p>填写完概念的属性后，单击“操作”列的 <input type="checkbox"/>，保存信息。</p> <ul style="list-style-type: none"> ● 复制属性 单击已有属性“操作”列的 <input type="checkbox"/>，复制当前属性。 ● 删除属性 单击已有属性“操作”列的 <input type="checkbox"/>，删除当前属性。 	<ul style="list-style-type: none"> ● 概念“Film”新增属性“release_time”、“production_company”和“adapted_from”。 ● 概念“Person”新增属性“ancestral_home”、“data_of_birth”、“graduated_school”、“height”、“nationality”、“nation”、“place_of_birth”。

5. 创建“Film”和“Person”之间的关系。

具体操作如下：

- a. 按住键盘a键，同时鼠标单击概念节点“Film”，拖拽出一条线，移动至目标概念“Person”。
或鼠标右键单击一个概念节点“Film”，选择“新建关系”，然后拖拽出一条线，移动至目标概念“Person”。
- b. 右键单击关系曲线，选择“编辑”。
右侧弹出关系编辑框。

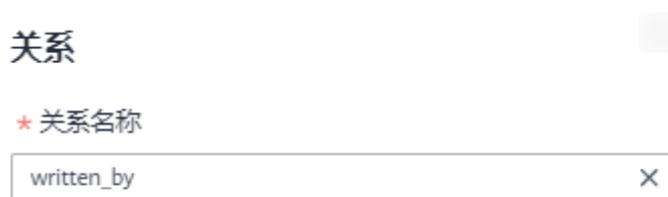
图 2-17 编辑关系



- c. 在关系编辑框中，填写“关系名称”，只能由大小写字母、数字、中文、下划线组成，长度为1-63位。

本样例需创建“Film”和“Person”之间的关系，关系名称为“written_by”、“starring”、“directed_by”、“produced_by”。

图 2-18 编辑关系



- d. 单击“保存”。
6. 创建“Person”和“Person”之间的关系。
具体操作如下。

- a. 按住键盘a键，同时鼠标单击概念节点“Person”，拖拽出一条线，移动至目标概念“Person”。

或鼠标右键单击一个概念节点“Person”，选择“新建关系”，然后拖拽出一条线，移动至目标概念“Person”。

- b. 右键单击关系曲线，选择“编辑”。

右侧弹出关系编辑框。

- c. 在关系编辑框中，填写“关系名称”，只能由大小写字母、数字、中文、下划线组成，长度为1-63位。

本样例需创建三个“Person”和“Person”之间的关系，关系名称分别为“father”、“spouse”、“mother”。

图 2-19 编辑关系

d. 单击“保存”。

步骤 2：选择图谱规格

1. 登录KG服务管理控制台，默认进入“我的图谱”页面。
2. 单击+，在创建图谱页面按表2-3填写信息。

图 2-20 服务选型

表 2-3 服务选型参数说明

参数	说明
图谱名称	待创建图谱的名称，名称只能由大小写字母、数字、中文、下划线组成，长度为1-63位。
图谱规格	待创建图谱的规格大小。当前仅支持在控制台选择“体验版 一万边”，如果根据业务需要，选择标准版 百万边、标准版 千万边、高级版 千万边，请通过官网产品页“服务咨询”联系我们进行咨询购买。
是否购买套餐包	体验版图谱默认不能开通套餐包。如果需要选择业务版图谱，请通过官网产品页“服务咨询”联系我们进行咨询购买。

3. 单击右下角“下一步”，在“版本确认”页签确认版本信息。
4. 单击右下角“确认创建”。

页面提示“知识图谱创建任务提交成功”。

单击“返回”，返回至“我的图谱”页面，您会看到新创建的图谱卡片，“运行状态”为“创建中”。等待十几分钟后，运行状态变为“初始化”状态。

图 2-21 创建图谱



步骤 3：配置数据源

创建图谱的数据使用在**步骤1：准备数据**中提前准备并上传至OBS的图谱数据。

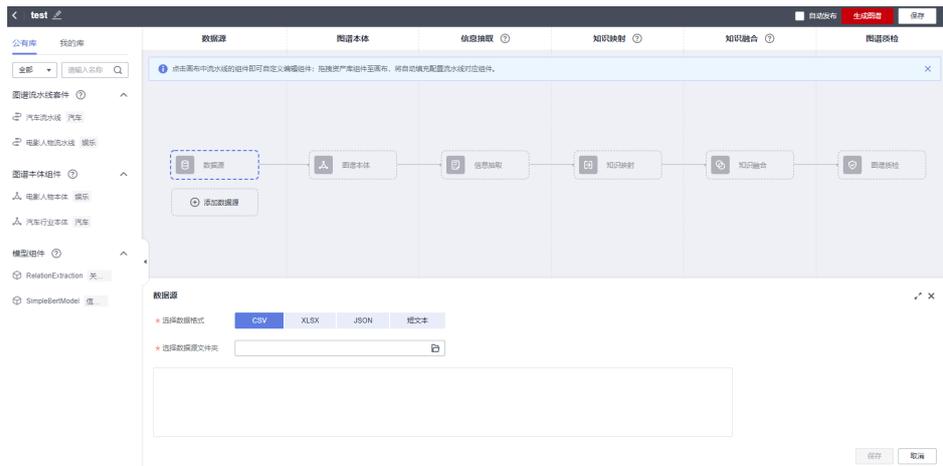
1. 在“我的图谱”页面鼠标移至新创建的图谱卡片，单击“图谱构建 Workspace”。
进入图谱流水线构建页面。

图 2-22 图谱构建 Workspace



2. 在流水线构建页面，单击“数据源”，页面下方弹出数据源配置对话框，单击右侧  按钮可以放大对话框。

图 2-23 配置数据源



3. 在“数据源配置”对话框，填写相关信息。
 - “选择数据格式”：此样例选择“短文本”。
 - “选择数据源文件”：单击 ，弹出“选择数据源文件”对话框，选择数据源存放在OBS的路径：
 - i. “OBS桶”：选择数据源文件存放的OBS桶“kg-model”。确保您使用的OBS桶与KG服务在同一区域，桶的存储类别为“标准存储”。
 - ii. “存储路径”：选择数据源文件存放在OBS桶的文件路径“kg-model/graph-data/corpus.txt”。
选择的OBS路径不能是加密路径，否则可能会访问失败。
 - iii. “OBS桶授权”：如果OBS桶未授权，请勾选“确认授权”。
 选择完成后，单击“确定”。

图 2-24 选择数据源文件



4. 填写完信息后，单击“保存”，完成数据源的配置。

步骤 4：配置图谱本体

1. 在流水线构建页面，单击“图谱本体”，页面下方弹出图谱本体配置对话框，单击右侧  按钮可以放大对话框。

- 选择左侧“我的库”页签，在“图谱本体组件”中拖拽**步骤1：创建本体**中创建的本体“person_film”到虚线框中，单击“保存”，完成本体选择。

说明

“我的库”页签下的“图谱本体组件”呈现的是“我的图谱资产库”中“我的本体”页面创建或OBS导入的所有本体。

可单击页面右下角的“编辑”，进入本体修改页面，修改后的本体将保存至“我的库”。

图 2-25 图谱本体

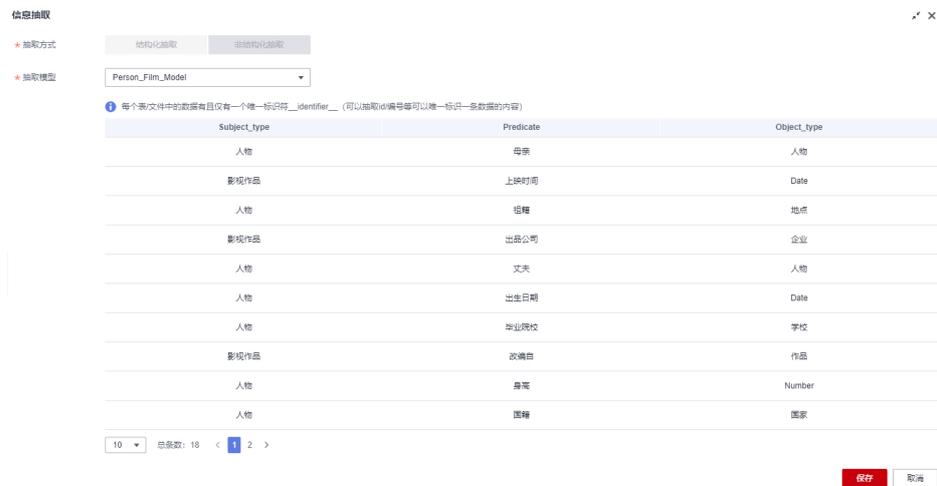


步骤 5：配置信息抽取

本样例使用在**创建信息抽取模型**中自定义的抽取模型进行信息抽取。

- 在流水线构建页面，单击“信息抽取”，页面下方弹出“信息抽取”对话框，单击右侧  按钮可以放大对话框。
- 填写信息抽取配置信息，如**图2-26**所示。
 - “抽取方式”：默认为“非结构化抽取”。
 - “抽取模型”：选择**创建信息抽取模型**用户自定义的模型“Person_Film_Model”。

图 2-26 配置信息抽取



3. 填写完信息后，单击“保存”。

步骤 6：配置知识映射

配置信息抽取后，需要配置知识映射，建立[步骤6：配置知识映射](#)中抽取出的三元组信息与本体的映射关系。

1. 在流水线构建页面，单击“知识映射”。
页面下方弹出“知识映射”对话框，单击右侧  按钮可以放大对话框。
2. 在“知识映射”对话框，选择编辑方式为“表格配置模式”，填写知识映射的相关信息。
 - 单击实体“Person”对应的  图标。
 - 填写实体“Person”映射前的抽取项：“人物”。
 - 填写映射到实体“Person”的抽取项，如[图2-27](#)和[表2-4](#)所示。

图 2-27 实体 Person 的抽取项

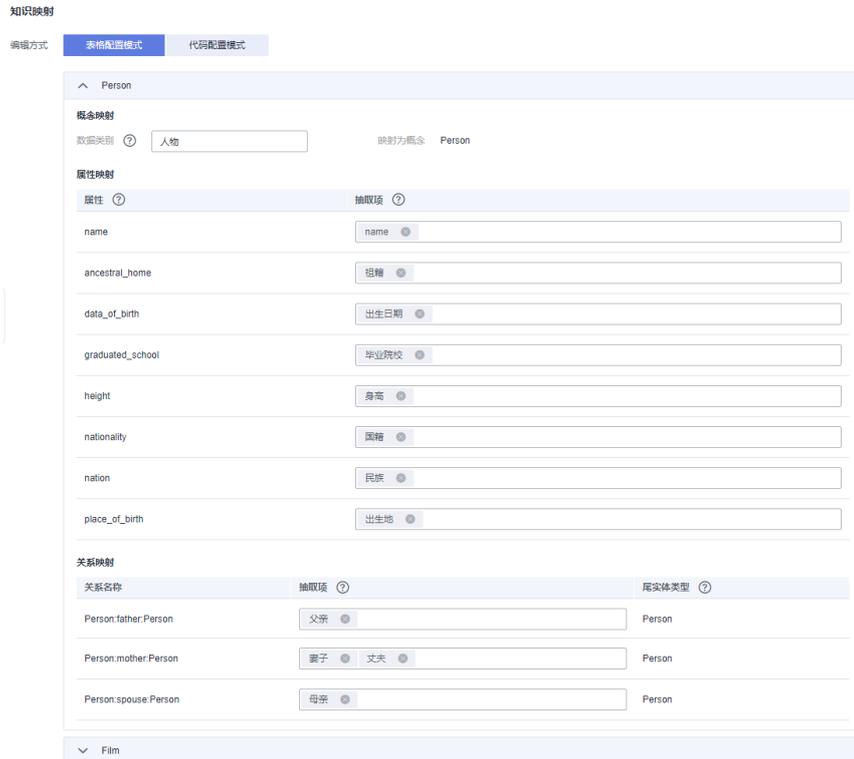


表 2-4 实体 Person 的抽取项

属性/关系名称	抽取项
name	name
place_of_birth	出生地
nation	民族
nationality	国籍
height	身高
graduated_school	毕业院校
data_of_birth	出生日期
ancestral_home	祖籍
father	父亲
spouse	妻子、丈夫
mother	母亲

- 单击实体“Film”对应的▼图标。
- 填写实体“Film”映射前的抽取项：“影视作品”。
- 填写映射到实体“Film”的抽取项，如图2-28和表2-5所示。

图 2-28 实体 Film 的抽取项

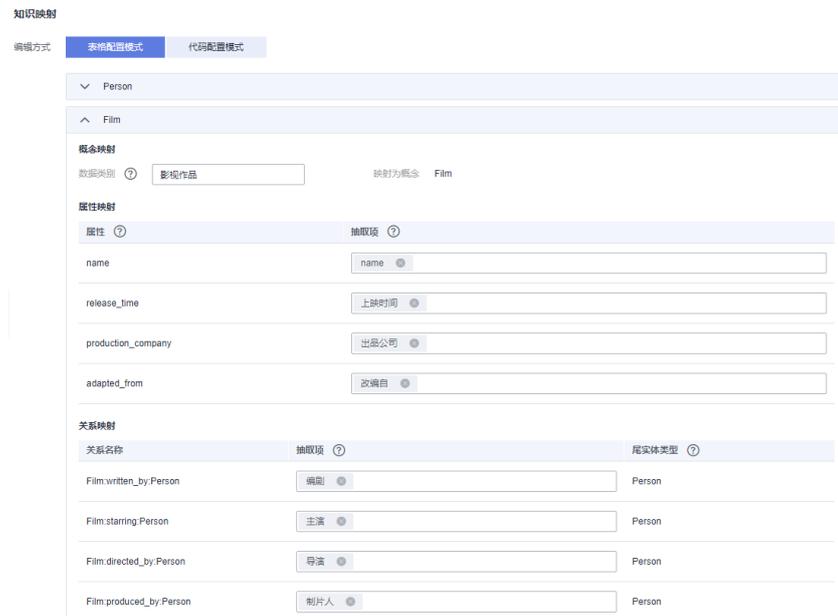


表 2-5 实体 Film 的抽取项

属性/关系名称	抽取项
name	name
adapted_from	改编自
production_company	出品公司
release_time	上映时间
written_by	编剧
starring	主演
directed_by	导演
produced_by	制片人

3. 单击“保存”，完成知识映射配置。

步骤 7：配置知识融合

本样例不需要配置知识融合，因此关闭融合开关即可。

1. 在流水线构建页面，单击“知识融合”，页面下方弹出“知识融合”对话框，单击右侧 按钮可以放大对话框。
2. 在“知识融合”对话框中，关闭“知识融合”开关，并单击“保存”。返回流水线构建页面。

图 2-29 关闭知识融合

知识融合

知识融合为非必选流程，不配置仍可运行图谱构建。配置融合后，需开启融合开关方可生效。

知识融合  

步骤 8：配置图谱质检

1. 在流水线构建页面，单击“图谱质检”，页面下方弹出“图谱质检”对话框，单击右侧  按钮可以放大对话框。

图 2-30 图谱质检



2. 在“图谱质检”对话框，打开“图谱质检”开关，单击“保存”。

说明

图谱质检为非必选流程，关闭和开启状态均不影响图谱构建。开启质检任务后，后台会对当前版本的知识图谱的知识质量和流水线构建进行评估，并生成质检报告，详见[查看质检报告](#)。

步骤 9：生成图谱

- 在流水线构建页面单击右上角“生成图谱”，完成图谱的创作。

图 2-31 生成图谱



跳至新建图谱“test”的详情页，查看图谱的状态，当图谱的“状态”由“运行中”变为“已完成”后，即可单击“图谱预览”，预览图谱。

3 使用公有资产库中的图谱流水线套件构建图谱

本章节提供一个与电影有关的非结构化数据样例，帮助您快速熟悉使用流水线套件构建知识图谱的过程，以及如何使用公有资产库中的图谱本体组件、模型组件。创建完成后，您可以搜索预览相关实体的知识图谱，也可以在创建的图谱基础上全量更新或增量更新图谱。

开始使用样例前，请仔细阅读[准备工作](#)罗列的要求，提前完成准备工作。在控制台上使用图谱流水线套件创建知识图谱的步骤如下所示：

- [步骤1：准备数据](#)
- [步骤2：选择图谱规格](#)
- [步骤3：使用图谱流水线套件](#)
- [步骤4：配置数据源](#)
- [步骤5：配置图谱本体（可选）](#)
- [步骤6：配置信息抽取](#)
- [步骤7：配置知识映射](#)
- [步骤8：配置知识融合](#)
- [步骤9：配置图谱质检](#)
- [步骤10：生成图谱](#)

准备工作

已[注册](#)华为账号，并完成实名认证，具体申请信息请见[账号注册](#)和[实名认证](#)。且在使用知识图谱服务前检查账号状态，账号不能处于欠费或冻结状态。

步骤 1：准备数据

在创建信息抽取模型之前，您需要自行准备基础数据并上传至OBS桶及文件夹中。

由于数据安全原因，本样例不提供具体的样例数据，仅提供样例数据要求，要求如下：

数据格式

数据格式为短文本txt格式，文件内容示例如下：

张三的生日是1990年1月1日，身高175cm，出生于北京。

李四，著名导演，毕业于电影学院，代表作有《电影1》、《电影2》。

... ..

数据划分

训练数据：80%数据用于训练信息抽取模型。

验证数据：20%数据用于验证信息抽取模型性能（即F1、P、R性能指标）。

训练数据量

训练一个基本可用的模型，标注数据量总数需要大于模型版本配置中的“batch_size”，且需大于20*三元组类型数量。

训练一个效果较好的模型，建议提供2万条以上的短句数据作为训练数据。

三元组类型

本样例构建的模型适用于人物、电影领域的信息抽取，限定抽取的三元组类型如表3-1所示。

表 3-1 三元组类型

subject_type	predicate	object_type
人物	毕业院校	学校
影视作品	改编自	作品
影视作品	制片人	人物
人物	身高	Number
人物	丈夫	人物
影视作品	主演	人物
人物	出生日期	Date
影视作品	编剧	人物
影视作品	导演	人物
人物	祖籍	地点
人物	父亲	人物
人物	出生地	地点
影视作品	出品公司	企业
人物	妻子	人物
人物	母亲	人物
人物	国籍	国家
影视作品	上映时间	Date

1. 按上文要求准备待标注的训练数据和模型数据，并保存至本地，训练数据命名为“annotation.txt”，模型数据命名为“corpus.txt”。
2. 登录OBS服务，创建桶和文件夹，用于存放数据集。
创建名称为“kg-model”的OBS桶。
创建名称为“model-data”的文件夹用于存放未标注的数据集。
创建名称为“model-data-out”的文件夹用于存放已标记并发布的数据集。
创建名称为“graph-data”的文件夹用于存放创建图谱的数据集。
创建OBS桶和文件夹的操作指导请参见[创建桶](#)和[新建文件夹](#)。为保证数据能正常访问，请务必保证创建的OBS桶与KG服务在同一区域，桶的存储类别为“标准存储”。
3. 参考[上传文件](#)，使用单个文件上传方式将本地准备的训练数据“annotation.txt”文件上传至OBS路径“kg-model/model-data”下，将本地准备的创建模型的数据文件“corpus.txt”上传至OBS路径“kg-model/graph-data”。

步骤 2：选择图谱规格

1. 登录KG服务管理控制台，默认进入“我的图谱”页面。
2. 单击+，然后在创建图谱页面填写信息。

图 3-1 服务选型

表 3-2 服务选型参数说明

参数	说明
“图谱名称”	待创建图谱的名称，名称只能由大小写字母、数字、中文、下划线组成，长度为1-63位。
“图谱规格”	待创建图谱的规格大小。当前仅支持在控制台选择“体验版 一万边”，如果根据业务需要，选择标准版 百万边、标准版 千万边、高级版 千万边，请通过官网产品页“服务咨询”联系我们进行咨询购买。
“是否购买套餐包”	体验版图谱默认不能开通套餐包。如果需要选择业务版图谱，请通过官网产品页“服务咨询”联系我们进行咨询购买。

- 单击右下角“下一步”，在“版本确认”页签确认版本信息。
- 单击右下角“确认创建”。
页面提示“知识图谱创建任务提交成功”。
单击“返回”，返回至“我的图谱”页面，您会看到新创建的图谱卡片，“运行状态”为“创建中”。等待十几分钟后，运行状态变为“初始化”状态。

图 3-2 创建图谱



步骤 3：使用图谱流水线套件

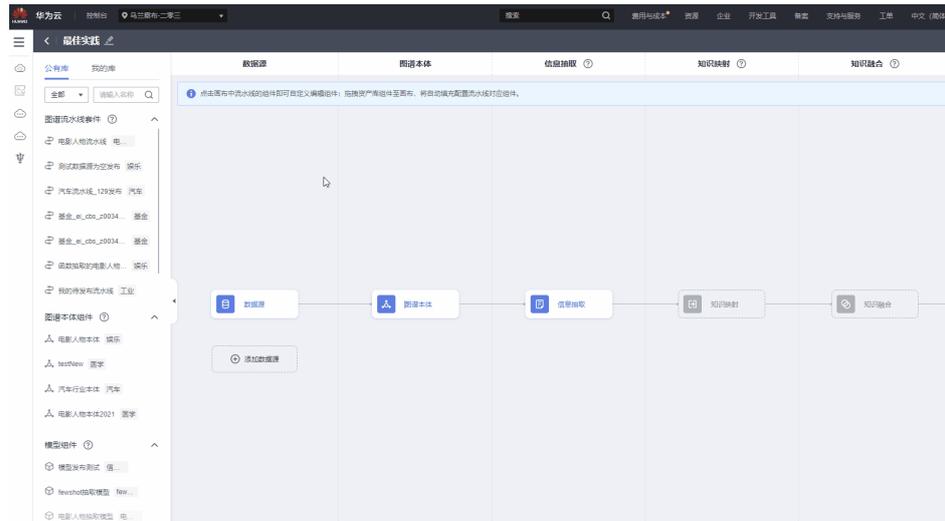
- 在“我的图谱”页面鼠标移至新创建的图谱卡片，单击“图谱构建 WorkSpace”。
进入流水线构建页面。

图 3-3 图谱构建 WorkSpace



- 在流水线构建页面，从左侧“公有库 > 图谱流水线套件”页签中拖拽“电影人物流水线”套件至右侧流水线画布中。
- 在弹出的对话框中，单击“确定”。
页面提示“图谱流水线套件应用成功”。

图 3-4 应用图谱流水线套件

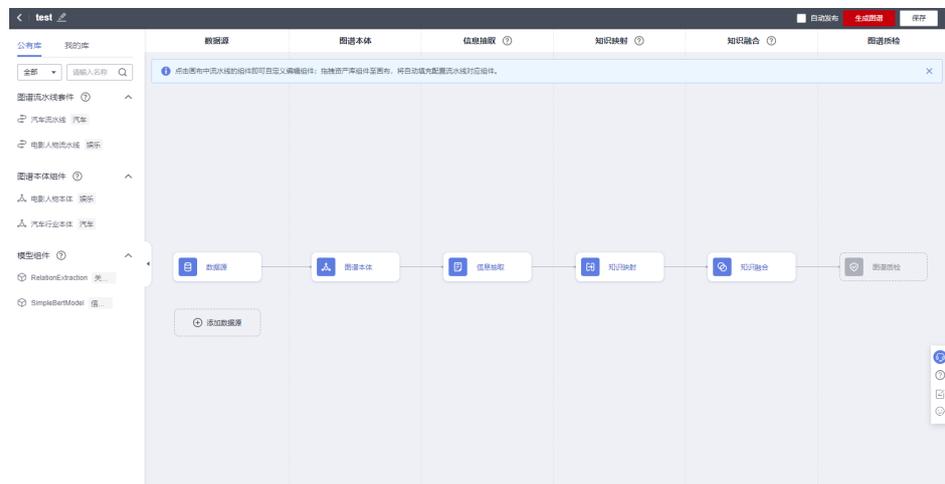


4. 单击流水线中的蓝色组件，可以查看配置样例。

步骤 4：配置数据源

1. 在流水线构建页面，单击“数据源”，页面下方弹出数据源配置对话框，单击右侧 按钮可以放大对话框。

图 3-5 配置数据源



2. 在“数据源配置”对话框，填写相关信息。
 - “选择数据格式”：此样例选择“短文本”。
 - “选择数据源文件”：单击 ，弹出“选择数据源文件”对话框，选择数据源存放在OBS的路径：
 - i. “OBS桶”：选择数据源文件存放的OBS桶“kg-model”。确保您使用的OBS桶与KG服务在同一区域，桶的存储类别为“标准存储”。
 - ii. “存储路径”：选择数据源文件存放在OBS桶的文件路径“kg-model/graph-data/corpus.txt”。
 选择的OBS路径不能是加密路径，否则可能会访问失败。

- iii. “OBS桶授权”：如果OBS桶未授权，请勾选“确认授权”。
选择完成后，单击“确定”。

图 3-6 选择数据源文件



3. 填写完信息后，单击“保存”，完成数据源的配置。

步骤 5: 配置图谱本体 (可选)

📖 说明

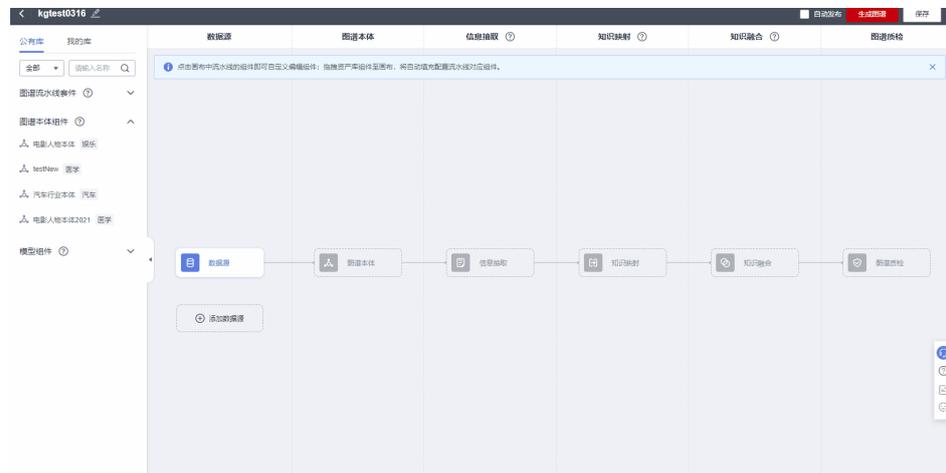
本步骤以使用国有资产库图谱本体组件为例，介绍图谱本体配置操作。

使用图谱流水线套件时，流水线套件中内置了图谱本体。如果内置本体满足您的需求，则不用执行配置图谱本体的步骤。

1. 在流水线构建页面，单击“图谱本体”，页面下方弹出图谱本体配置对话框，单击右侧 按钮可以放大对话框。
2. 在左侧“公有库”页签，拖拽“图谱本体组件”到虚线框中，单击“保存”，完成本体选择。

可单击页面右下角的“编辑”，进入本体修改页面，修改后的本体将保存至“我的库”。

图 3-7 图谱本体



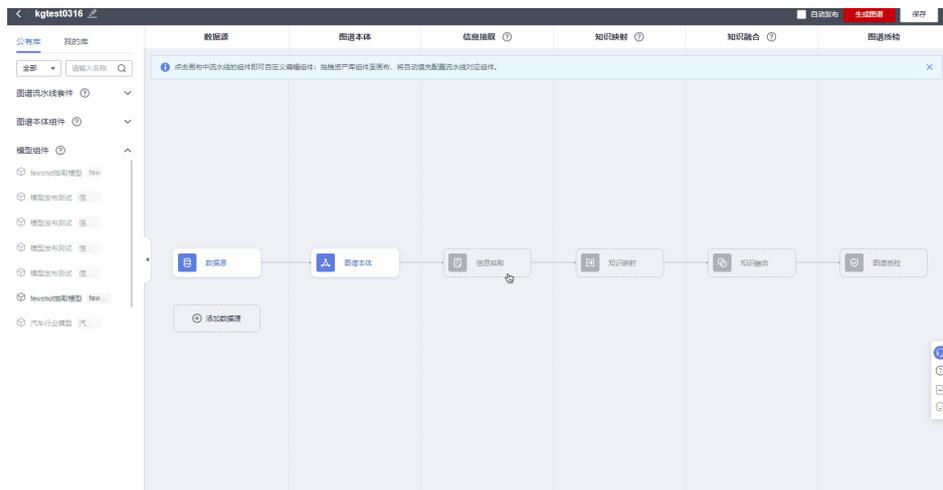
步骤 6：配置信息抽取

📖 说明

本步骤以使用国有资产库模型组件为例，介绍信息抽取的配置操作。

1. 在流水线构建页面，单击“信息抽取”，页面下方弹出“信息抽取”对话框，单击右侧  按钮可以放大对话框。
2. 填写信息抽取配置信息，如图3-8所示。
 - “抽取方式”：默认为“非结构化抽取”。
 - “抽取模型”：从左侧“公有库 > 模型组件”页签中拖拽至右侧虚线框。

图 3-8 配置信息抽取



3. 填写完信息后，单击“保存”。

步骤 7：配置知识映射

配置信息抽取后，需要配置知识映射，建立抽取出的三元组信息与本体的映射关系。

1. 在流水线构建页面，单击“知识映射”。

页面下方弹出“知识映射”对话框，单击右侧  按钮可以放大对话框。
2. 在“知识映射”对话框，选择编辑方式为“表格配置模式”，填写知识映射的相关信息。
 - 单击实体“Person”对应的  图标。
 - 填写实体“Person”映射前的抽取项：“人物”。
 - 填写映射到实体“Person”的抽取项，如图3-9和表3-3所示。

图 3-9 实体 Person 的抽取项

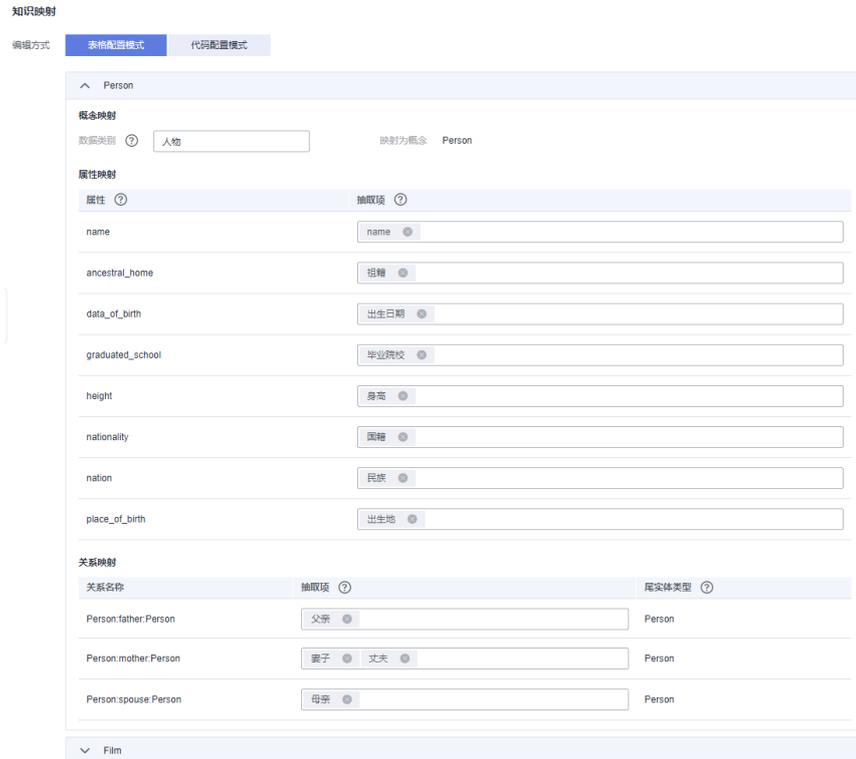


表 3-3 实体 Person 的抽取项

属性/关系名称	抽取项
name	name
place_of_birth	出生地
nation	民族
nationality	国籍
height	身高
graduated_school	毕业院校
data_of_birth	出生日期
ancestral_home	祖籍
father	父亲
spouse	妻子、丈夫
mother	母亲

- 单击实体“Film”对应的▼图标。
- 填写实体“Film”映射前的抽取项：“影视作品”。
- 填写映射到实体“Film”的抽取项，如图3-10和表3-4所示。

图 3-10 实体 Film 的抽取项

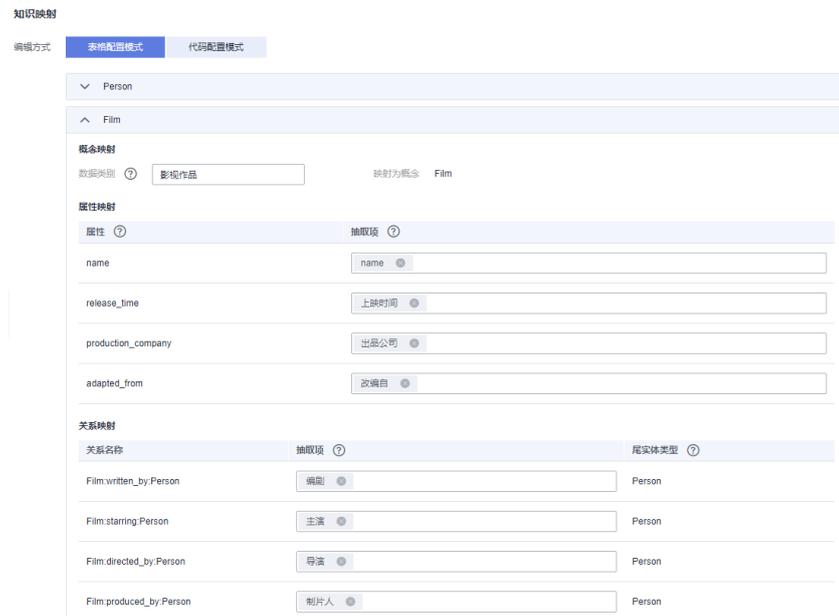


表 3-4 实体 Film 的抽取项

属性/关系名称	抽取项
name	name
adapted_from	改编自
production_company	出品公司
release_time	上映时间
written_by	编剧
starring	主演
directed_by	导演
produced_by	制片人

3. 单击“保存”，完成知识映射配置。

步骤 8：配置知识融合

本样例不需要配置知识融合，因此关闭融合开关即可。

1. 在流水线构建页面，单击“知识融合”，页面下方弹出“知识融合”对话框，单击右侧 按钮可以放大对话框。
2. 在“知识融合”对话框中，关闭“知识融合”开关，并单击“保存”。返回流水线构建页面。

图 3-11 关闭知识融合

知识融合

知识融合为非必选流程，不配置仍可运行图谱构建。配置融合后，需开启融合开关方可生效。



步骤 9：配置图谱质检

1. 在流水线构建页面，单击“图谱质检”，页面下方弹出“图谱质检”对话框，单击右侧  按钮可以放大对话框。

图 3-12 图谱质检



2. 在“图谱质检”对话框，打开“图谱质检”开关，单击“保存”。

说明

图谱质检为非必选流程，关闭和开启状态均不影响图谱构建。开启质检任务后，后台会对当前版本的知识图谱的知识质量和流水线构建进行评估，并生成质检报告，详见[查看质检报告](#)。

步骤 10：生成图谱

- 在图谱流水线页面，单击右上角“生成图谱”，完成图谱的创建。

图 3-13 生成图谱



跳至新建图谱“test”的详情页，查看图谱的状态，当图谱的“状态”由“运行中”变为“可用”后，即可单击“图谱预览”，预览图谱。