AI 开发平台 ModelArts
7.5.1
数据准备

文档版本	01
发布日期	2026-02-05

华为云计算技术有限公司

商标声明


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云计算技术有限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，华为云计算技术有限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。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 数据准备功能说明 1
2 数据集格式要求 4
2.1 文本类数据集格式要求 4
2.2 图片类数据集格式要求 9
2.3 视频类数据集格式要求 10
2.4 音频类数据集格式要求 11
2.5 其他类数据集格式要求 12
2.5.1 Alpaca 数据集格式要求 12
3 数据连接 14
3.1 快速实现数据连接 14
3.2 创建数据连接 17
3.3 管理数据连接 21
4 数据资产管理 24

1 数据准备功能说明

功能介绍

数据决定了大模型的能力上限。ModelArts**数据准备**功能提供了一站式、全流程的数据处理和管理服务，致力于解决大模型开发中“数据获取难、质量参差不齐、处理效率低”的核心痛点。通过内置的行业级数据处理算子与自动化流水线，系统化的处理数据获取、加工、发布等过程，帮助您将海量、多模态的原始数据，高效转化为高可用、高纯度的训练数据集，提高数据质量和处理效率，显著降低模型训练成本，提升模型泛化能力。

数据工程开发流程

ModelArts平台（下称“平台”）提供了快捷的数据开发流程，您通过**数据连接**即可完成模型数据集的开发。

- **数据连接**：数据获取是数据工程的第一步，支持将不同来源和格式的数据导入平台，并生成“原始数据集”。通过这些功能，用户可以轻松将大量数据导入平台，为后续的数据加工和模型训练等操作做好准备。详见**数据连接**章节。

数据资产管理

在集成了数据连接，数据精炼功能外，平台还支持对原始数据集、加工数据集、发布数据集、数据合成指令进行一站式管理。在大规模数据集的构建过程中，平台的数据工程功能为用户提供了极大的灵活性和高效性，确保了数据处理的各个环节都能紧密协作，快速响应不断变化的业务需求和技术要求。管理各类型数据集以及通过数据质量评估确保数据满足大模型训练的多样性、平衡性和代表性需求，并促进数据的高效流通与应用。详见**数据资产管理**章节。

ModelArts 平台支持的数据类型

ModelArts平台（下称“平台”）提供了业界最全面的数据处理功能。包括对文本类、图片类、音频类、视频类常规数据集处理，还提供了自己定义数据集功能，支持业界使用广泛的Alpaca等数据集格式，能够灵活处理多样化的数据。

平台多样化的数据精炼和管理能力，为您提供丰富而全面的数据集，是您开发大模型的利器。

平台支持的数据类型见**表1-1**，各类型数据格式详细要求请参考**数据集格式要求**。

表 1-1 平台支持的数据类型

数据类型	数据内容	支持的文件格式	数据集要求
文本	单轮问答	jsonl、csv	文本类数据集格式要求
	单轮问答（人设）	jsonl、csv	
	多轮问答	jsonl	
	多轮问答（人设）	jsonl	
	问答排序	jsonl、csv	
	偏好优化DPO	jsonl	
	偏好优化DPO（人设）	jsonl	
图片类	图片	• 图片+jsonl（可选） – 图片格式支持：jpg、jpeg、png、bmp。 – jsonl为非必须文件类型。当存在jsonl文件时，图片文本保存为一份jsonl文件，jsonl文件中图片名称必须要与tar包中的图片名称一致。**注意：**jsonl文件仅支持UTF-8编码。 • tar+jsonl（可选）：所有图片保存为tar包。 – 图片格式支持：jpg、jpeg、png、bmp。 – jsonl为非必须文件类型。当存在jsonl文件时，图片文本保存为一份jsonl文件，jsonl文件中图片名称必须要与tar包中的图片名称一致。**注意：**jsonl文件仅支持UTF-8编码。	图片类数据集格式要求

数据类型	数据内容	支持的文件格式	数据集要求
	图片+Caption	• tar+jsonl：所有图片保存为tar包。 – 图片格式支持：jpg、jpeg、png、bmp。 – 图片文本保存为一份jsonl文件，jsonl文件中图片名称必须要与tar包中的图片名称一致。**注意：**jsonl文件仅支持UTF-8编码。	
视频类	视频	mp4、avi	**视频类数据集格式要求**
	视频+标注	• 视频+jsonl – 视频格式支持：mp4、avi。 – 标注文件格式：jsonl，jsonl文件仅支持UTF-8编码。	
音频类	音频	• 音频+jsonl – 音频文件：支持mp3、flac、wav、opus、aac、m4a格式，允许放在根目录或下层目录中。 – 标注文件格式：可选，格式为UTF-8编码的jsonl文件，每一行描述一个音频文件在数据集中的相对路径以及其它信息。	**音频类数据集格式要求**
其他类	自定义	支持构建用户自定义场景下所需的数据集类型。支持Alpaca格式数据集。	**其他类数据集格式要求**

2 数据集格式要求

2.1 文本类数据集格式要求

ModelArts Studio大模型开发平台支持创建文本类数据集，创建时可导入多种形式的数据，具体格式要求详见**表2-1**。

表 2-1 文本类数据集格式要求

文件内容	文件格式	文件要求
单轮问答	jsonl、csv	• jsonl盘古格式-非思维链：数据由问答对构成，context、target分别表示问题、答案，具体格式示例如下： {"context": ["你好，请介绍自己"], "target": "我是盘古大模型"} • jsonl盘古格式-非思维链（用于RFT）：context为问题的描述，target为问题的标答。具体格式示例如下： {"context": "你是一位经验丰富的医生。请基于患者的症状信息，对候选疾病进行可能性排序。\n\n输入信息包括：\n1.显性症状列表：[['咳嗽', 'True'], ['鼻塞', 'True'], ['黄鼻涕', 'True'], ['绿鼻涕', 'True']]\n2.隐性症状列表：[['出生后20天左右着凉', 'True'], ['清鼻涕', 'True'], ['哭时喉咙里有痰', 'True'], ['右边眼睛发痒', 'True'], ['有黄眼屎', 'True']]\n3.候选疾病列表：['呼吸道感染', '腺样体肥大', '上呼吸道感染']\n\n请按照以下步骤进行分析：\n1.综合分析所有症状信息\n2.对每个候选疾病评估其与症状的匹配程度\n3.基于症状表现的典型性和特异性进行排序\n4.将疾病按可能性从高到低排序，并以json格式输出最终的排序列表\n\n示例输出：\n{\n \"possible_diseases\": [\"疾病1\", \"疾病2\", \"疾病3\"]\n}\n\n请基于以上标准对该患者的候选疾病进行分析并给出排序结果。", "target": "上呼吸道感染"} • jsonl盘古格式-非思维链（用于GRPO）：目前仅支持数学类数据，context和target分别代表问题和可验证的回答。target的内容为问题的标准答案，不包含任何推理流程，只需要最终结果。具体格式示例如下： {"context": "Let $P(x)$ be a polynomial of degree $3n$ such that\n\\begin{align*} P(0) = P(3) = \\dots = P(3n) &= 2, \\\\ P(1) = P(4) = \\dots = P(3n+1-2) &= 1, \\\\ P(2) = P(5) = \\dots = P(3n+2-2) &= 0. \\end{align*}\nAlso, $P(3n+1) = 730$. Determine n.", "target": "1"} • jsonl盘古格式-思维链：数据由问答对构成，context、target分别表示问题、答案，并且target必须包含think标签对表示思考过程，具体格式示例如下： {"context": ["你好，请介绍自己"], "target": "<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>我是盘古大模型"} • csv盘古格式-非思维链：csv文件的第一列对应context，第二列对应target，具体格式示例如下： "你好，请介绍自己","我是盘古大模型" • csv盘古格式-思维链：csv文件的第一列对应context，第二列对应target，并且target必须包含think标签对，具体格式示例如下： "你好，请介绍自己","<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>我是盘古大模型" • 从OBS导入：单个文件/压缩包大小不超过20GB，文件数量不限制。

文件内容	文件格式	文件要求
单轮问答（人设）	jsonl、csv	• jsonl盘古格式-非思维链：system表示人设，context、target分别表示问题、答案。具体格式示例如下： {"system":"你是一个机智幽默问答助手","context":["你好，请介绍自己"],"target":"哈哈，你好呀，我是你的聪明助手。"} • jsonl盘古格式-非思维链（用于RFT）：system代表人设，context为问题的描述，target为问题的标答。具体格式示例如下： {"system": "你是一个擅长于逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。", "context": "你是一位经验丰富的医生。请基于患者的症状信息，对候选疾病进行可能性排序。\n\n输入信息包括：\n1.显性症状列表：[['咳嗽', 'True'], ['鼻塞', 'True'], ['黄鼻涕', 'True'], ['绿鼻涕', 'True']]\n2.隐性症状列表：[['出生后20天左右着凉', 'True'], ['清鼻涕', 'True'], ['哭时喉咙里有痰', 'True'], ['右边眼睛发痒', 'True'], ['有黄眼屎', 'True']]\n3.候选疾病列表：['呼吸道感染', '腺样体肥大', '上呼吸道感染']\n\n请按照以下步骤进行分析：\n1.综合分析所有症状信息\n2.对每个候选疾病评估其与症状的匹配程度\n3.基于症状表现的典型性和特异性进行排序\n4.将疾病按可能性从高到低排序，并以json格式输出最终的排序列表\n\n示例输出：\n{\n \"possible_diseases\": [\"疾病1\", \"疾病2\", \"疾病3\"]\n}\n\n请基于以上标准对该患者的候选疾病进行分析并给出排序结果。", "target": "上呼吸道感染"} • jsonl盘古格式-非思维链（用于GRPO）：目前仅支持数学类数据，system代表人设，context和target分别代表问题和可验证的回答。target的内容为问题的标准答案，不包含任何推理流程，只需要最终结果。具体格式示例如下： {"system":"数学专家", "context": "Let $P(x)$ be a polynomial of degree $3n$ such that\n\\begin{align*} P(0) = P(3) = \\dots = P(3n) &= 2, \\\\ P(1) = P(4) = \\dots = P(3n+1-2) &= 1, \\\\ P(2) = P(5) = \\dots = P(3n+2-2) &= 0. \\end{align*}\nAlso, $P(3n+1) = 730$. Determine n.", "target": "1"} • jsonl盘古格式-思维链：system表示人设，context、target分别表示问题、答案， 并且target必须包含think标签对表示思考过程，具体格式示例如下： {"system":"你是一个机智幽默问答助手","context":["你好，请介绍自己"],"target":"<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>哈哈，你好呀，我是你的聪明助手。"} • csv盘古格式-非思维链：csv文件的第一列对应system，第二三列分别对应context、target。具体格式示例如下： "你是一个机智幽默问答助手","你好，请介绍自己","哈哈，你好呀，我是你的聪明助手。" • csv盘古格式-思维链：csv文件的第一列对应system，第二三列分别对应context、target，并且target必须包含think标签对表示思考过程，具体格式示例如下： "你是一个机智幽默问答助手","你好，请介绍自己","<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>哈哈，你好呀，我是你的聪明助手。" • 从OBS导入：单个文件/压缩包大小不超过20GB，文件数量不限制。

文件内容	文件格式	文件要求
多轮问答	jsonl	• jsonl盘古格式-非思维链：数组格式，由一轮或多轮问答对构成。context、target分别表示问题、答案，具体格式示例如下： [{"context":["你好"],"target":"你好，请问有什么可以帮助你的？"},{"context":["请介绍一下华为云的产品。"],"target":"华为云提供包括但不限于计算、存储、网络等产品服务。"}] • jsonl盘古格式-思维链：数组格式，由一轮或多轮问答对构成，其中context、target分别表示问题、答案，并且至少有一轮问答的target包含think标签对表示思考过程，具体格式示例如下： [{"context":["你好"],"target":"<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>你好，请问有什么可以帮助你的？"},{"context":["请介绍一下华为云的产品。"],"target":"华为云提供包括但不限于计算、存储、网络等产品服务。"}] • 从OBS导入：单个文件/压缩包大小不超过20GB，文件数量不限制。
多轮问答（人设）	jsonl	• jsonl盘古格式-非思维链：数组格式，由一轮或多轮问答对构成。system表示人设，context、target分别表示问题、答案。具体格式示例如下： [{"system":"你是一位书籍推荐专家"},{"context":["你好"],"target":"嗨！你好，需要点什么帮助吗？"},{"context":["能给我推荐点书吗？"],"target":"当然可以，基于你的兴趣，我推荐你阅读《自动驾驶的未来》。"}] • jsonl盘古格式-思维链：数组格式，由人设一轮或多轮问答对构成。system表示人设，context、target分别表示问题、答案，并且至少有一轮问答的target包含think标签对表示思考过程，具体格式示例如下： [{"system":"你是一位书籍推荐专家"},{"context":["你好"],"target":"<think>用户在打招呼，需要回复以及询问</think>嗨！你好，需要点什么帮助吗？"},{"context":["能给我推荐点书吗？"],"target":"<think>我需要以专家的身份给客户推荐书籍</think>当然可以，基于你的兴趣，我推荐你阅读《自动驾驶的未来》。"}] • 从OBS导入：单个文件/压缩包大小不超过20GB，文件数量不限制。
问答排序	jsonl、csv	• jsonl格式：context表示问题，targets答案1、2、3表示答案的优劣顺序，最好的答案排在最前面。 { "context":"context内容","targets":["回答1","回答2","回答3"]} • csv格式：csv文件的第一列对应context，其余列为答案。 "问题","回答1","回答2","回答3" • 从OBS导入：单个文件/压缩包大小不超过20GB，文件数量不限制。

文件内容	文件格式	文件要求
偏好优化DPO	jsonl	• jsonl盘古格式-非思维链：context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案。具体格式示例如下： 单轮问答 {"context": ["你好，请介绍自己"], "target":"我是盘古大模型", "bad_target":"我不会回答"} 多轮问答 {"context": ["你好，请介绍自己", "我是盘古大模型", "请介绍一下有哪些产品。"], "target":"提供包括但不限于计算、存储、网络等产品服务。", "bad_target":"我不会回答"} • jsonl盘古格式-思维链：context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案，答案中至少有一个包含think标签对表示思考过程，具体格式示例如下： 单轮问答 {"context": ["你好，请介绍自己"], "target":"我是盘古大模型", "bad_target":"我不会回答"} 多轮问答 {"context": ["你好，请介绍自己", "我是盘古大模型", "请介绍一下有哪些产品。"], "target":"<think>客户想要了解产品</think>提供包括但不限于计算、存储、网络等产品服务。", "bad_target":"我不会回答"} • 从OBS导入：单个文件/压缩包大小不超过20GB，文件数量不限制。
偏好优化DPO（人设）	jsonl	• jsonl盘古格式-非思维链：system表示人设，context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案。具体格式示例如下： 带人设单轮问答 {"system": "你是一位机智幽默的问答助手", "context": ["你好，请介绍自己"], "target":"哈哈，你好呀，我是你的聪明助手，怎么帮到你？ ", "bad_target":"我不会回答"} 带人设多轮问答 {"system": "你是一位机智幽默的问答助手", "context": ["你好，请介绍自己", "哈哈，你好呀，我是你的聪明助手，怎么帮到你？ ", "请介绍一下有哪些产品。"], "target":"我们产品种类繁多，不仅涵盖计算、存储和网络，还有更多选择哦！ ", "bad_target":"我不会回答"} • jsonl盘古格式-思维链：system表示人设，context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案，答案中至少有一个包含think标签对表示思考过程，具体格式示例如下： 带人设单轮问答 {"system": "你是一位机智幽默的问答助手", "context": ["你好，请介绍自己"], "target":"哈哈，你好呀，我是你的聪明助手，怎么帮到你？ ", "bad_target":"我不会回答"} 带人设多轮问答 {"system": "你是一位机智幽默的问答助手", "context": ["你好，请介绍自己", "哈哈，你好呀，我是你的聪明助手，怎么帮到你？ ", "请介绍一下有哪些产品。"], "target":"<think>客户想要了解产品</think>我们产品种类繁多，不仅涵盖计算、存储和网络，还有更多选择哦！ ", "bad_target":"我不会回答"} • 从OBS导入：单个文件/压缩包大小不超过20GB，文件数量不限制。

2.2 图片类数据集格式要求

ModelArts Studio大模型开发平台支持创建图片类数据集，创建时可导入多种形式的数据，具体格式要求详见表2-2。

表 2-2 图片类数据集格式要求

文件内容	文件格式	文件要求
图片	图片+jsonl tar+jsonl	• 图片：支持jpg、jpeg、png、bmp类型。  • tar：tar包内图片支持jpg、jpeg、png、bmp图片类型。  • 根目录下可存在单个annotation.jsonl文件，image_name字段必选。 `{"image_name":"图片名称（abc.jpg）","tar_name":"tar包名称（1.tar）"}` • 从OBS导入：单个压缩包大小不超过20GB(只支持tar类型的压缩包)，单个文件大小不超过20GB，文件数量不限制。

文件内容	文件格式	文件要求
图片+Caption	tar+jsonl	• 图片：图片以tar包格式存储，可以多个tar包。tar包存储原始的图片，每张图片命名要求唯一（如abc.jpg）。图片支持jpg、jpeg、png、bmp格式。 • jsonl：图片描述jsonl文件放在最外层目录，一个tar包对应一个jsonl文件，文件内容中每一行代表一段文本，形式为： {"image_name":"图片名称（abc.jpg）","tar_name":"tar包名称（1.tar）","caption":"图片对应的文本描述"} • 从OBS导入：单个压缩包大小不超过20GB(只支持tar类型的压缩包)，单个文件大小不超过20GB，文件数量不限制。 

2.3 视频类数据集格式要求

ModelArts Studio大模型开发平台支持创建视频类数据集，创建时可导入多种形式的数据，具体格式要求详见表2-3。

表 2-3 视频类数据集格式要求

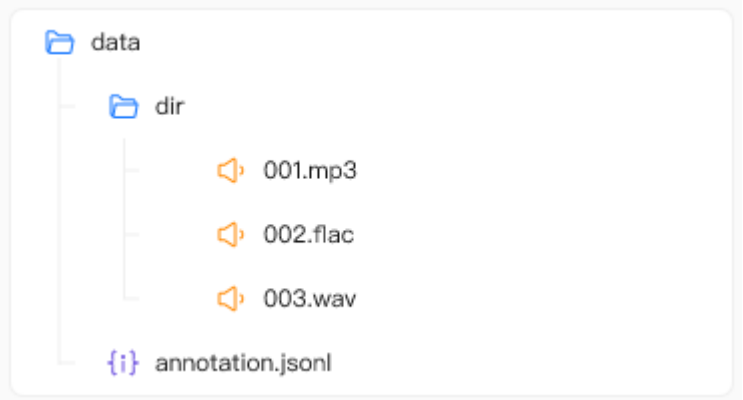
文件内容	文件格式	文件要求
视频	mp4或avi	• 支持mp4、avi视频格式上传，所有视频可以放在多个文件夹下，每个文件夹下可以同时包含mp4或avi格式的视频。 • 从OBS导入：单个文件/压缩包大小不超过20GB，文件数量不限制。

文件内容	文件格式	文件要求
视频+标注	视频+jsonl	• 视频格式支持：mp4、avi • 标注文件格式：jsonl，jsonl文件仅支持UTF-8编码。 示例如下所示：  具体的jsonl标注文件参考： {"video_fn": "13/ad098173-af09-48fe-95c3-e72fd629688e.mp4"视频相对路径， "prompt": "A person pours a clear liquid from a bottle into a shot glass, then lifts the glass to their mouth and drinks the shot. The background includes a red coat and other indistinct background elements."视频摘要生成（简略）， "long_prompt": "A person is seen pouring a clear liquid from a green glass bottle into a small glass. The individual is wearing a white shirt with a lace collar and a beige cardigan. The background appears to be a cozy indoor setting, possibly a cafe or a restaurant, with red and white elements visible, such as a red coat hanging on the wall and a white table. The person carefully pours the liquid, ensuring it is filled to the brim of the glass. The liquid is clear and has some green leaves floating in it. The person then holds the glass up, possibly to show the contents or to prepare for a drink.",视频摘要生成（详细） }

2.4 音频类数据集格式要求

ModelArts Studio大模型开发平台支持创建音频类数据集，具体格式要求详见表2-4

表 2-4 音频类数据集格式要求

文件内容	文件格式	文件要求
音频	音频+jsonl（可选）	• 音频格式支持：mp3、flac、wav、opus、aac、m4a格式，允许放在根目录或下层目录中。 标注文件格式：可选，格式为UTF-8编码的jsonl文件，每一行描述一个音频文件在数据集中的相对路径以及其它信息。 • 从OBS导入：单个文件/压缩包大小不超过20GB，文件数量不限制。 示例如下所示：  具体的jsonl标注文件参考： {"audio_name":"dir/16k_16bit_1channel_2s.flac","caption":"1"} {"audio_name":"dir/16k_16bit_1channel_2s.mp3","caption":"2"} {"audio_name":"dir/16k_16bit_1channel_2s.opus","caption":"3"} {"audio_name":"dir/16k_16bit_1channel_2s.wav","caption":"4"}

2.5 其他类数据集格式要求

除文本、图片、视频、音频数据集外，平台还支持导入其他类数据集，即用户训练模型时使用的自定义数据集，比如常用的开源Alpaca格式数据集。

从OBS导入：单个文件/压缩包大小不超过20GB，文件数量不限制。

本章将会对常见的开源数据集介绍其格式要求。

2.5.1 Alpaca 数据集格式要求

Alpaca是开源模型（如DeepSeek系列、Qwen系列等）常用的数据集格式，是开源模型数据微调使用的主要数据集格式。特别用于instruction-tuning，即指令微调。其数据格式的特点是提供了一个明确的任务描述（instruction）、输入（input）和输出（output）三部分。

典型的Alpaca数据集格式：

```
[
  {
    "instruction": "人类指令（必填）",
```

```
    "input": "人类输入（选填）",
    "output": "模型回答（必填）",
    "system": "系统提示词（选填）",
    "history": [
      [
        "第一轮指令（选填）",
        "第一轮回答（选填）"
      ],
      [
        "第二轮指令（选填）",
        "第二轮回答（选填）"
      ]
    ]
  }
]
```

字段说明：

- instruction: 任务的指令，告诉模型需要完成什么操作。
- input: 任务所需的输入。如果任务是开放式的或者不需要明确的输入，这一字段可以为空字符串。
- output: 任务的期望输出，也就是模型在给定指令和输入情况下需要生成的内容。如果想训练带思考模式的模型，需要加<think></think>标签，或者引导思考的prompt，比如“Let's think step by step”。
- system：系统提示词（如什么风格、什么角色），该字段可选。
- history：是由多个字符串二元组构成的列表，分别代表历史消息中每轮对话的指令和回答。在指令监督微调时，历史消息中的回答内容也会被用于模型学习，该字段可选。

特点：

- Alpaca的数据格式结构非常简单，开发者能够快速搞清楚各字段含义及快速上手构建数据集。
- 任务指令和输入内容是分离的，适合各种自然语言处理任务，如文本生成、翻译、总结等。

3 数据连接

3.1 快速实现数据连接

ModelArts平台提供了方便的数据连接功能，您可以将自有数据集导入ModelArts后直接做训练模型。

本文将通过以下假设场景介绍如何使用ModelArts数据连接功能。

业务场景

ModelArts平台提供了最新的Qwen3大模型，您希望通过本地准备好的训练数据集对Qwen3模型做微调。

数据集为**Alpaca格式**，可以直接使用该数据集对Qwen3做微调。

您需要将本地数据集导入到ModelArts后做模型微调。

前提条件

1. 已注册华为账号并开通华为云，进行了实名认证，且在使用ModelArts前检查账号状态，账号不能处于欠费或冻结状态。
   - **注册华为账号并开通华为云**
   - **进行实名认证**
2. 配置委托访问授权

   ModelArts使用过程中涉及到OBS等服务交互，首次使用ModelArts需要用户配置委托授权，允许访问这些依赖服务。
3. 您本地要有能够训练Qwen3的训练数据集，数据集为Alpaca格式，格式说明参见**Alpaca数据集格式要求**。

计费说明

数据连接计费涉及到数据存储OBS计费和数据转换计费。计费说明如下：

1. 数据连接在上传数据时涉及到计费，具体可参考**数据管理计费项**。
2. 数据连接如果勾选“转换成Alpaca格式”开关，涉及计算资源使用，当前版本限时免费。

步骤一：本地数据上传至 OBS

参考**OBS桶上传**操作上传数据。

步骤二：修改数据连接配置任务

1. 前往**ModelArts管理控制台**。
2. 在控制台左侧导航栏选择“数据准备 > 数据连接”，选择后右侧展开“数据连接”工作区，如图**图3-1**所示。

图 3-1 数据连接工作区

3. 在“数据连接”工作区右上方单击“创建数据连接”按钮，打开“创建数据连接”配置页面。输入数据连接任务名称和描述。

图 3-2 “创建数据连接”配置任务

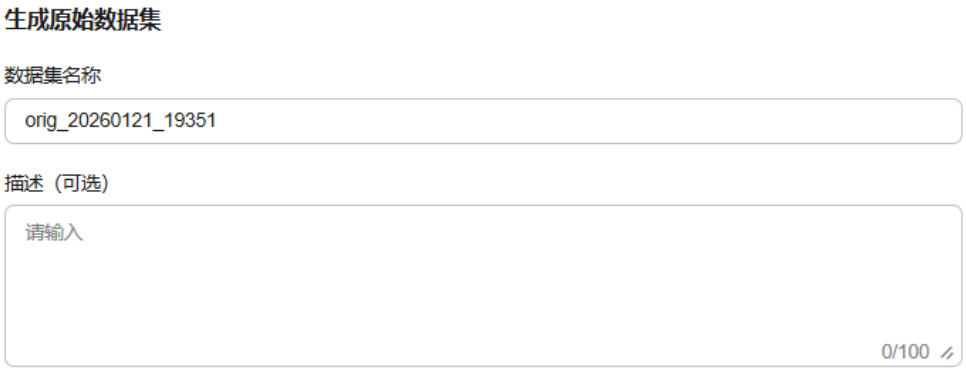
任务名称为必选，描述信息为可选，任务名称命名格式要求：长度为2~63字符。以中文、字母开头，以中文、字母、数字结尾，只允许输入中文、字母、数字、中划线、下划线等字符，具体参见**创建数据连接任务**中任务命名要求。

4. 导入本地数据。在“数据导入”配置项选择数据集类型为“其他 > 自定义”。如**图3-3**所示。选择“导入来源”为OBS，将步骤一导入到OBS的数据作为本次数据集的来源。

图 3-3 数据导入

5. 将OBS导入数据作为一个数据集，需要给数据集重新命名。输入数据集名称及描述信息（可选），此时本地数据才算是ModelArts上的一个数据集。

图 3-4 填写生成数据集信息



6. 生成数据集还有一些扩展信息可以选填，说明数据集的属性和版权信息，本文示例不填该信息。
7. 数据集填写完成后，勾选“生成后自动上线数据集”，勾选后数据集才能作为数据集资产，后续训练模型才能选到该数据集。

图 3-5 勾选“生成后自动上线数据集”

8. 所有配置都已经完成，单击工作区右下角“立即创建”按钮，开始启动本次数据连接任务。待任务完成后，就可使用该数据集做Qwen3的微调工作了。

相关参考

1. 数据集相关格式问题，请参见**数据集格式要求**。
2. 开源数据集说明请参见**Alpaca数据集格式要求**。

3.2 创建数据连接

使用场景

在数据处理和模型训练的场景中，用户需要将多种类型的数据集高效、准确地导入到ModelArts数据平台中，以支持后续的数据精炼和模型训练任务。然而，传统的数据导入方式存在诸多限制，如不支持自定义任务名称、数据格式转换功能有限等，导致用户在导入数据时面临操作不便和数据处理效率低下的问题。如何在新的平台中实现更加灵活和高效的数据导入功能，成为用户亟待解决的问题。为此，ModelArts平台提供了增强的数据导入功能，支持多种基础数据类型的导入，允许用户在创建导入任务时编辑任务名称和描述，同时支持数据格式转换和数据集的直接发布，从而显著提升了数据处理的灵活性和效率，满足了用户在数据准备阶段的多样化需求。

前提条件

1. 已注册华为账号并开通华为云，进行了实名认证，且在使用ModelArts前检查账号状态，账号不能处于欠费或冻结状态。
   - **注册华为账号并开通华为云**
   - **进行实名认证**
2. 配置委托访问授权

   ModelArts使用过程中涉及到OBS等服务交互，首次使用ModelArts需要用户配置委托授权，允许访问这些依赖服务。
3. 创建导入任务前，请先按照**数据集格式要求**提前准备数据。

约束限制

- 数据连接导入数据文件或压缩包不超过20GB。

计费说明

数据连接计费涉及到数据存储OBS计费和数据转换计费。计费说明如下：

1. 数据连接在上传数据时涉及到计费，具体可参考**数据管理计费项**。

2. 数据连接如果勾选“转换成Alpaca格式”开关，涉及计算资源使用，当前版本限时免费。

创建数据连接任务

数据连接操作步骤如下：

1. 前往**ModelArts管理控制台**。
2. 在控制台左侧导航栏选择“数据准备 > 数据连接”，选择后右侧展开“数据连接”工作区，如图**图3-6**所示。

图 3-6 数据连接

3. 在“数据连接”工作区右上方单击“创建数据连接”按钮，打开“创建数据连接”配置页面。输入数据连接任务名称和描述。任务名称为必选，描述信息为可选。

图 3-7 “创建数据连接”配置任务

说明

任务名称：命名默认为data-connect-年月日-时分秒，如：data-connect-20260116-233300，也可自定义名称，命名要求如下：

- 命名长度：2~63字符。
- 格式要求：以中文、字母**开头**，以中文、字母、数字**结尾**。只允许输入中文、字母、数字、中划线、下划线等字符。

描述：无格式要求，长度不超过200字符，内容可选填。

4. 从OBS导入数据至ModelArts平台，当前支持导入的数据类型参见**ModelArts平台支持的数据类型**，如**图3-8**所示。请根据不同的作用，导入不同类型的数据，“导入来源”为OBS。

图 3-8 数据导入

5. OBS上的数据可以被重复使用，故OBS上的数据还不能称之为数据集，只有通过数据导入功能将数据导入到ModelArts平台，这些数据才算是一个数据集。数据集可以有很多种，比如模型训练的数据集，微调的数据集，评测的数据集。而这些数据集的源头数据也有可能是同一份数据。数据集代表了一类有相同使用属性的数据集合。

   导入的数据形成新的数据集，需要给数据集重新命名。输入数据集名称及描述信息（可选）。

   说明

   数据集名称：

   - 命名长度：2~63字符。
   - 格式要求：以中文、字母**开头**，以中文、字母、数字**结尾**。只允许输入中文、字母、数字、中划线、下划线等字符。
   - **描述：**无格式要求，长度不超过200字符，内容可选填。

6. 部分文本类数据支持数据格式转换，支持将华为ModelArts内部支持的格式转换成Alpaca格式。便于使用训练过华为内部模型的数据方便的转为开源格式数据，训练开源模型。

   当前只有**单轮问答、单轮问答（人设）、多轮问答、多轮问答（人设）**四种数据集类型支持格式转换。

图 3-9 生成原始数据集

7. 生成数据集还有一些扩展信息可以选填，扩展信息包括“数据集属性”与“数据集版权”，具体说明如下：

- 数据集属性。可以给数据集添加行业、语言和自定义信息。
- 数据集版权。训练模型的数据集除用户自行构建外，也可能会使用开源的数据集。数据集版权功能主要用于记录和管理数据集的版权信息，确保数据的使用合法合规，并清晰地了解数据集的来源和相关的版权授权。通过填写这些信息，可以追溯数据的来源，明确数据使用的限制和许可，从而保护数据版权并避免版权纠纷。

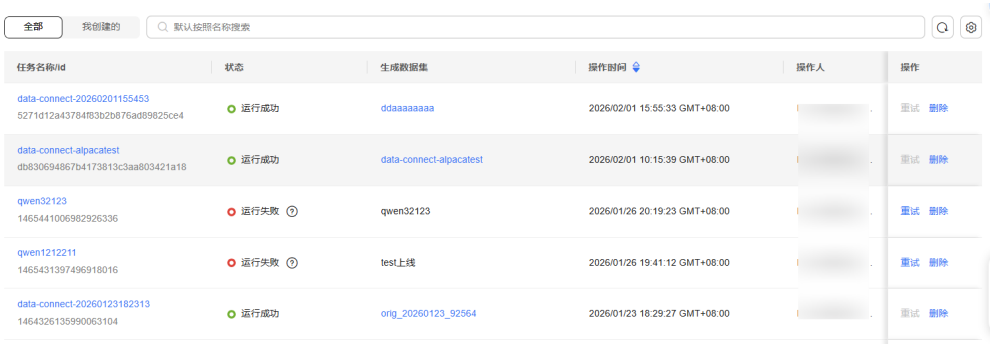
图 3-10 扩展信息

8. 数据集填写完成后，配置“生成后自动上线数据集”。默认该配置不勾选，此时数据集成功后，不会直接作为数据资产，模型无法直接使用该数据集做训练。勾选该配置后，数据集才能作为数据集资产直接上线至“资产管理”，后续训练模型才能选到该数据集。如需了解数据资产相关内容，请参考**数据资产管理**章节。

图 3-11 勾选“生成后自动上线数据集”

9. 单击页面右下角“立即创建”，返回至“数据连接”页面，在该页面可以查看数据集的任务状态，若状态为“运行成功”，则数据导入成功。

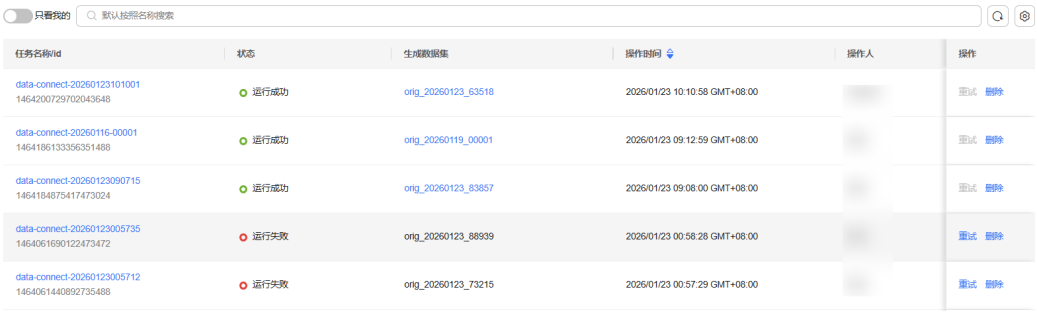
图 3-12 查看数据连接任务列表



3.3 管理数据连接

数据连接需要对所有任务做统一管理。在数据连接任务工作区可以查看数据连接任务创建的状态，数据集名称、创建时间、创建人以及支持的一些操作。如图3-13所示。本文将详细介绍如何管理数据连接任务。

图 3-13 查看数据连接任务列表



展示/隐藏功能简介

打开数据连接工作区，最上方默认会展示数据连接的功能简介，用户能快速感知数据连接的功能亮点及业务范围，快速上手无难度。

图 3-14 打开新手引导

对于经验丰富的用户，也可单击功能简介，关闭该功能。如图3-15所示。

图 3-15 关闭新手引导

数据连接任务管理

数据连接任务支持过滤、搜索、删除、重试等操作。以下分别讲解如何操作。

- **数据任务过滤和搜索。**在数据任务列表比较多的情况下，支持通过不同维度过滤，当前支持按照**任务id、任务名称、状态、生成数据集、操作人**等维度过滤想要的任务，便于快速找到目标任务。也可勾选“只看我的”开关，只列举当前登录用户创建的数据连接任务。

图 3-16 数据集过滤

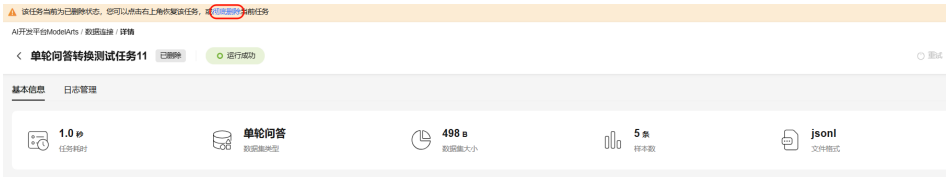
- **数据任务删除。**

  对于列表中创建的数据连接任务，支持“删除”操作。可以在“操作”列单击“删除”按钮，在弹出删除对话框选择“确定”后，该任务将被删除。删除后的任务不是彻底删除，为避免误删，如果还想再继续使用，可以恢复任务。已删除的任务在任务清单有“已删除”标签，如图3-18所示。对于已删除的任务，可以选择“彻底删除”，如图3-19所示。彻底删除后的任务不可恢复。

图 3-17 删除任务

图 3-18 已删除任务的“已删除”标签

图 3-19 任务“彻底删除”



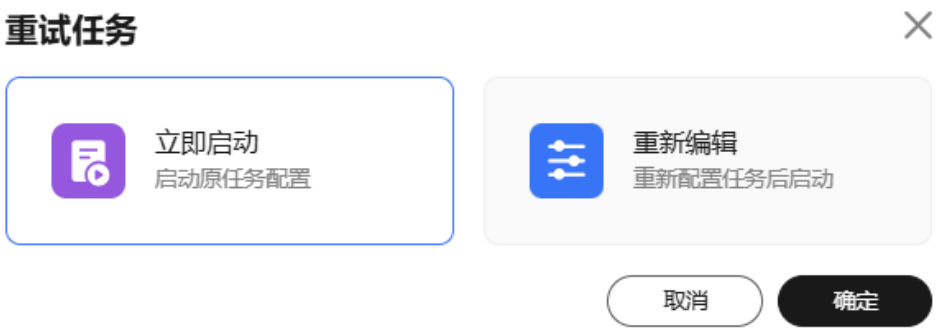
- **数据任务重试。**

  对于运行失败的任务，支持“重试”操作重新运行该任务。可以在“操作”列单击“重试”按钮，在弹出的对话框中选择“直接重启”或“重新编辑”，如图3-20所示。两种重试任务的区别如下：

  直接重启：不修改连接任务的任何配置，直接重新启动任务。

  重新编辑：重新进入数据连接配置界面，修改配置后重新启动任务。

图 3-20 重试任务



选择重启任务类型后，单击“确认”可重启任务。

数据连接任务详情管理

数据任务详情页面展示了当前任务详细信息。在数据连接工作区，单击任意任务名称，就进入该任务的任务详情页面。该页面有基本信息和日志管理两个子页面，以下分别说明两个页面的作用和涉及的操作。

- **基本信息**。该页面列举了该任务的数据集信息、数据配置详情以及生成数据集的链接。在该页面右上角可以删除任务。对于失败的任务，可以单击“重试”按钮重新启动。
- **日志管理。**在该界面可以查看“操作记录”和“运行日志”。操作记录会记录当前任务做过的所有操作。运行日志则记录运行过程中的日志。两者都有助于定位在创建任务过程中出现的问题。

4 数据资产管理

数据资产介绍

数据资产是指在ModelArts平台中被纳入管理、存储并可供使用的数据集。用户可以通过“数据准备 > 数据连接 ”功能创建对应任务，勾选“发布后自动上线数据集”开关，数据集将自动上线至资产。也可手动上线数据集资产。数据资产只有上线才能被用于后续模型的训练和评测。

发布的数据集支持查看详细信息、删除、上线和下线。以下将分别介绍数据集资产支持的管理操作。

数据资产上下线和删除

在“数据连接”任务如果配置“发布后自动上线数据集”，生成的数据集将自动上线为资产。如果没有勾选，则在资产清单是未上线的状态。具体上下线操作如下：

1. 在左侧导航栏中选择“资产管理 > 数据”，在右侧“数据”工作区能够查看所有数据集和前用户自己创建的资产列表，也可以按照**数据集名称、数据模态、数据集类型、上线状态、创建者**维度过滤数据集资产。

   图 4-1 过滤数据集资产

2. 选择一个的数据集，在“操作”列支持如下操作：
   - **上线。**未上线的数据集支持上线。单击“上线”，在弹出的对话框确认后，数据集完成上线。上线后的数据集能够作为模型开发的数据。
   - **下线。**已上线的数据集支持下线。单击“下线”，在弹出的对话框确认后，数据集完成下线。上下线后的数据集不能作为模型开发的数据。
   - **删除。**数据集可被删除。删除后的数据集不是彻底删除，为避免误删，如果还想再继续使用，可以恢复数据集。对于已删除的数据，可以选择彻底删除，彻底删除后的数据集不可恢复。

图 4-2 已删除数据集

图 4-3 已删除数据集可恢复或彻底删除

- **恢复。**对于已经删除的数据集，可以通过该选项恢复数据集。

数据资产详情管理

数据资产详情页面展示了当前数据集详细信息。在数据集工作区，单击任意数据集名称，就进入该数据集的详情页面。该页面有基本信息、数据预览和操作记录三个子页面，以下分别说明页面的作用和涉及的操作。在该页面右上角可以删除数据集。单击删除后，该数据集将彻底删除，请谨慎操作。

- **基本信息**。该页面列举了该数据集信息、数据配置详情以及生成数据集的链接。
- **数据预览**。该页面可以查看数据集详情，对于部分文本类型数据集，支持通markdown格式查看。
- **操作记录。**在该界面可以查看“操作记录”。操作记录会记录当前数据集做过的所有操作。