

Flexus 企业搜索服务

最佳实践

文档版本 01
发布日期 2024-08-27



版权所有 © 华为云计算技术有限公司 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 配置 OCR 增强实现文档智能解析..... | 1 |
| 2 Flexus 企业搜索对接 MaaS 实现文档问答..... | 6 |

1 配置 OCR 增强实现文档智能解析

应用场景

Flexus企业搜索服务（后续简称为KooSearch），KooSearch结合OCR服务文档识别能力，通过NET网关打通网络，使KooSearch支持智能文档解析能力，对PDF/Word等格式的文档内容进行版式分析，智能识别文档中的各种版式类型，如表格、图片、公式、标题等，以及识别单栏、双栏版面结构及段落顺序，对用户上传的文档进行智能解析。本文介绍Flexus企业搜索服务如何配置OCR增强实现文档智能解析功能。

步骤 1：开通 OCR

开通OCR服务的智能文档解析功能。

1. 登录文字识别OCR管理控制台。
2. 在“总览”页面的使用说明模块选择“智能文档解析”。
3. 单击操作列的“开通管理”，弹出开通管理页面，确认是否已经为操作者开通了智能文档解析功能。
 - 是，则进行下一步。
 - 否，则单击操作列的“开通服务”为操作者开通智能文档解析功能。

如图1-1所示，表示所有人都已开通智能文档解析功能。

图 1-1 开通智能文档解析



详细操作指导请参见[文字识别OCR](#)。

步骤 2：开通 Flexus 企业搜索服务

购买智能问答版Flexus企业搜索服务（KooSearch）。

1. 登录云搜索服务[管理控制台](#)。

2. 在左侧导航栏选择“KooSearch>Flexus 企业搜索服务”，进入Flexus企业搜索服务页面。
3. 单击页面右上角的“购买Flexus企业搜索服务”按钮，进入购买页面。
4. 在购买页面，完成Flexus企业搜索服务配置。

表 1-1 购买 Flexus 企业搜索服务

| 参数 | 子参数 | 说明 |
|-------|-------|--|
| 区域 | - | 选择实例的所在区域。默认选择“西南-贵阳一”区域。 说明 仅“西南-贵阳一”区域支持开通和使用KooSearch服务。 |
| 版本规格 | - | <ul style="list-style-type: none">智能搜索版：提供开箱即用的知识库，主要用于RAG，为大语言模型（LLM）提供外部知识来源，扩大知识检索范围。智能问答版：提供内置大语言模型（LLM），一站式快速搭建企业级RAG服务。 本案例选择“智能问答版”。 |
| 实例名称 | - | 实例名称为Flexus企业搜索服务的名称。 实例名称在4位到16位之间，必须以字母开头，可以包含字母、数字、中划线或者下划线，不能包含其他的特殊字符。 本案例以“kos_ocr”为例。 |
| OCR增强 | - | OCR增强功能是引用OCR服务能力，来增强实例的搜索和问答能力。 开通OCR增强插件会创建委托访问文字识别服务，需额外计费。 |
| 网络配置 | 虚拟私有云 | 指定集群节点使用的虚拟专用网络，实现不同业务的网络隔离。 单击“查看虚拟私有云”跳转到虚拟私有云列表，查看已创建或共享至当前账号下的VPC名称和ID。 如果没有合适的VPC，建议联系CSS服务管理员新建VPC，具体请参见 创建虚拟私有云和子网 。 |
| | 子网 | 集群使用子网实现与其他网络的隔离，并独享所有网络资源，以提高网络安全。 选择当前虚拟私有云下集群需要的子网。支持选择共享VPC下的子网。 |
| | 安全组 | 安全组起着虚拟防火墙的作用，为集群提供安全的网络访问控制策略。 选择集群需要的安全组，单击“查看安全组”跳转到安全组列表，可以了解安全组详情。 |

| 参数 | 子参数 | 说明 |
|------|-----|---|
| 购买时长 | - | 支持按月购买，购买周期越长价格越优惠，根据实例购买时长，一次性支付实例费用。最短时长为1个月，最长时长为1年。如果购买时长超过9个月，建议包年购买，价格更优惠。也可以勾选自动续费复选框，到期会自动续费。 |

5. 配置完成后，单击“立即购买”，确认信息无误，单击“去支付”，跳到支付页面。
6. 根据页面提示成功付款，即可完成KooSearch的开通。

步骤 3：配置网络

1. 登录NAT网关管理控制台。
2. 在公网NAT网关列表页面，单击“购买公网NAT网关”，进入购买公网NAT网关页面。
3. 在购买公网NAT网关页面，配置NAT网关。

关键参数请参见表1-2配置，其他参数可以保持默认值或基于业务场景参考[购买公网NAT网关](#)完成配置。

表 1-2 购买公网 NAT 网关

| 参数 | 说明 |
|-------|------------------------------|
| 计费模型 | 因为KooSearch是离线服务，建议选择“按需计费”。 |
| 区域 | 和KooSearch保持一个区域，选择“西南-贵阳一”。 |
| 虚拟私有云 | 和KooSearch选择同一个VPC。 |
| 子网 | 和KooSearch选择同一个子网。 |

4. 单击“立即购买”，确认产品规格无误后，单击“提交”，开始创建公网NAT网关。当状态为“运行中”时，表示公网NAT网关创建成功。
5. 当公网NAT网关创建成功后，在公网NAT网关列表页面，单击操作列的“设置规则”，在弹窗中单击“添加规则”进入网关详情页的“SNAT规则”页面。
6. 在“SNAT规则”页面，单击“添加SNAT规则”，在弹窗中配置公网NAT网关的SNAT规则。

表 1-3 添加 SNAT 规则

| 参数 | 说明 |
|------|------------------------------|
| 使用场景 | 选择“虚拟私有云”。 |
| 网段 | 选择“使用已有”，选择和KooSearch同一个VPC。 |

| 参数 | 说明 |
|--------|---|
| 公网IP类型 | 选择“弹性公网IP”，勾选满足要求的IP。 如果没有可选的弹性公网IP，可以单击“查看弹性公网IP”跳转到弹性公网IP列表页， 申请弹性公网IP 。 建议弹性公网IP的“计费模式”为“按需计费”，“公网带宽”为“按流量计费”，“带宽大小”为“20”Mbit/s。 |

- 单击“确定”完成SNAT规则的添加。

步骤 4：验证 OCR 增强

- 登录云搜索服务[管理控制台](#)。
- 在左侧导航栏选择“KooSearch>Flexus 企业搜索服务”，进入Flexus企业搜索服务页面。
- 选择文档问答服务，单击操作列的“问答”，前往KooSearch控制台。
- 在KooSearch控制台，左侧导航栏选择“知识库管理”。
进入知识库管理页面。
- 在知识库管理页面，单击右上角“新建知识库”进入新建知识库页面。
- 在“新建知识库”页面设置知识库信息。
在“解析拆分设置”页签勾选“OCR增强”，使知识库可以调用OCR服务进行智能文档识别，如表格解析或扫描文件等。
其他参数可以保持默认值或基于业务场景参考[新建知识库](#)完成配置。

图 1-2 知识库开启 OCR 增强



- 设置好后，单击“确定”完成知识库创建。
- 在知识库管理页面，选择已创建的知识库，单击操作列的“文档管理”，进入知识库详情页面。
- 选择“文档管理”页签，单击“上传”。
- 在上传对话框中单击“选择文档”，本地选择已提前准备好的文档。重复的文档不允许上传。

图 1-3 上传文档



11. 单击“确认”。文档上传后，可在文档管理页签查看已上传的文档，当“文档状态”为“正常”，说明文档已上传成功。
12. 文档上传成功后，可以单击文档名称，查看切片效果。

2 Flexus 企业搜索对接 MaaS 实现文档问答

Flexus企业搜索对接ModelArts Studio大模型即服务平台（后续简称为MaaS）适用于多种应用场景。本文介绍Flexus企业搜索对接MaaS实现文档问答。

步骤 1：在 MaaS 部署模型服务

在MaaS服务中完成模型服务的部署，并获取模型服务的信息用于KooSearch对接MaaS时使用。

1. 登录ModelArts管理控制台。
2. 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
3. 在ModelArts Studio大模型即服务平台选择合适的模型，创建为“我的模型”。操作指导请参见[在MaaS中创建模型](#)。
4. 将创建好的模型部署为模型服务，具体操作请参见[使用MaaS部署模型服务](#)。
5. 选择使用AppCode认证方式调用部署好的模型服务，具体操作请参见[调用MaaS部署的模型服务](#)。
6. 获取并记录模型服务的“调用模型名称”、“API接口公网地址”和“AppCode”，当KooSearch对接MaaS时会使用到这些参数。

步骤 2：开通 Flexus 企业搜索服务

购买智能问答版Flexus企业搜索服务（KooSearch）。

1. 登录云搜索服务[管理控制台](#)。
2. 在左侧导航栏选择“KooSearch>Flexus 企业搜索服务”，进入Flexus企业搜索服务页面。
3. 单击页面右上角的“购买Flexus企业搜索服务”按钮，进入购买页面。
4. 在购买页面，完成Flexus企业搜索服务配置。

表 2-1 购买 Flexus 企业搜索服务

| 参数 | 子参数 | 说明 |
|-------|-------|--|
| 区域 | - | 选择实例的所在区域。默认选择“西南-贵阳一”区域。 说明 仅“西南-贵阳一”区域支持开通和使用KooSearch服务。 |
| 版本规格 | - | <ul style="list-style-type: none">智能搜索版：提供开箱即用的知识库，主要用于RAG，为大语言模型（LLM）提供外部知识来源，扩大知识检索范围。智能问答版：提供内置大语言模型（LLM），一站式快速搭建企业级RAG服务。 本案例选择“智能问答版”。 |
| 实例名称 | - | 实例名称为Flexus企业搜索服务的名称。 实例名称在4位到16位之间，必须以字母开头，可以包含字母、数字、中划线或者下划线，不能包含其他的特殊字符。 本案例以“kos_ocr”为例。 |
| OCR增强 | - | OCR增强功能是引用OCR服务能力，来增强实例的搜索和问答能力。 开通OCR增强插件会创建委托访问文字识别服务，需额外计费。 |
| 网络配置 | 虚拟私有云 | 指定集群节点使用的虚拟专用网络，实现不同业务的网络隔离。 单击“查看虚拟私有云”跳转到虚拟私有云列表，查看已创建或共享至当前账号下的VPC名称和ID。 如果没有合适的VPC，建议联系CSS服务管理员新建VPC，具体请参见 创建虚拟私有云和子网 。 |
| | 子网 | 集群使用子网实现与其他网络的隔离，并独享所有网络资源，以提高网络安全。 选择当前虚拟私有云下集群需要的子网。支持选择共享VPC下的子网。 |
| | 安全组 | 安全组起着虚拟防火墙的作用，为集群提供安全的网络访问控制策略。 选择集群需要的安全组，单击“查看安全组”跳转到安全组列表，可以了解安全组详情。 |
| 购买时长 | - | 支持按月购买，购买周期越长价格越优惠，根据实例购买时长，一次性支付实例费用。最短时长为1个月，最长时长为1年。如果购买时长超过9个月，建议包年购买，价格更优惠。也可以勾选自动续费复选框，到期会自动续费。 |

- 配置完成后，单击“立即购买”，确认信息无误，单击“去支付”，跳到支付页面。
- 根据页面提示成功付款，即可完成KooSearch的开通。

步骤 3：配置网络

1. 登录NAT网关管理控制台。
2. 在公网NAT网关列表页面，单击“购买公网NAT网关”，进入购买公网NAT网关页面。
3. 在购买公网NAT网关页面，配置NAT网关。

关键参数请参见表2-2配置，其他参数可以保持默认值或基于业务场景参考[购买公网NAT网关](#)完成配置。

表 2-2 购买公网 NAT 网关

| 参数 | 说明 |
|-------|------------------------------|
| 计费模型 | 因为KooSearch是离线服务，建议选择“按需计费”。 |
| 区域 | 和KooSearch保持一个区域，选择“西南-贵阳一”。 |
| 虚拟私有云 | 和KooSearch选择同一个VPC。 |
| 子网 | 和KooSearch选择同一个子网。 |

4. 单击“立即购买”，确认产品规格无误后，单击“提交”，开始创建公网NAT网关。当状态为“运行中”时，表示公网NAT网关创建成功。
5. 当公网NAT网关创建成功后，在公网NAT网关列表页面，单击操作列的“设置规则”，在弹窗中单击“添加规则”进入网关详情页的“SNAT规则”页面。
6. 在“SNAT规则”页面，单击“添加SNAT规则”，在弹窗中配置公网NAT网关的SNAT规则。

表 2-3 添加 SNAT 规则

| 参数 | 说明 |
|--------|---|
| 使用场景 | 选择“虚拟私有云”。 |
| 网段 | 选择“使用已有”，选择和KooSearch同一个VPC。 |
| 公网IP类型 | 选择“弹性公网IP”，勾选满足要求的IP。 如果没有可选的弹性公网IP，可以单击“查看弹性公网IP”跳转到弹性公网IP列表页， 申请弹性公网IP 。 建议弹性公网IP的“计费模式”为“按需计费”，“公网带宽”为“按流量计费”，“带宽大小”为“20”Mbit/s。 |

7. 单击“确定”完成SNAT规则的添加。

步骤 4：KooSearch 对接 MaaS

在KooSearch新建MaaS模型服务，创建知识库并绑定MaaS模型服务，并设置模型的Prompt。

1. 登录云搜索服务管理控制台。

2. 在左侧导航栏选择“KooSearch>Flexus 企业搜索服务”，进入Flexus企业搜索服务页面。
3. 选择文档问答服务，单击操作列的“问答”，前往KooSearch控制台。
4. 创建模型服务，对接MaaS的模型服务。
 - a. 在KooSearch左侧导航栏选择“模型管理”，进入“模型管理”页面。
 - b. 单击页面的“新建模型服务”，弹出新建模型服务页面。

图 2-1 新建模型服务



- c. 在新建模型服务页面。根据下表填写对应参数后，单击“确定”按钮。

表 2-4 新建模型服务

| 参数 | 说明 |
|----------|---|
| 模型服务名称 | 自定义模型服务名称，不能为空。 |
| 模型类型 | 选择“NLP模型-昇腾云”，表示通过昇腾云的MaaS服务提供的NLP大模型访问方式。如果选择此模型进行问答，建议设置模型生成最大新词数不超过512。 |
| 访问地址 | MaaS服务的模型服务的“API接口公网地址”。获取方式请参见步骤1：在MaaS部署模型服务。 |
| 模型描述 | 模型详细信息的描述。 |
| 昇腾云模型名称 | 来源于MaaS服务的模型服务的“调用模型名称”，获取方式请参见步骤1：在MaaS部署模型服务。 |
| 上下文长度（K） | NLP大模型在进行一次特定的推理时可以考虑的最大令牌数。可以扩展上下文以生成更全面的响应。 |
| 认证类型 | 勾选“自定义认证”，添加自定义请求头。“标头键”填写“X-ApiG-AppCode”，“标头值”填写MaaS服务中模型服务的“AppCode”，获取方式请参见步骤1：在MaaS部署模型服务。 |

- d. 创建成功后，可以在模型管理页面中看到创建的模型服务。
5. 创建知识库。
 - a. 在KooSearch控制台，左侧导航栏选择“知识库管理”。进入知识库管理页面。
 - b. 在知识库管理页面，单击右上角“新建知识库”进入新建知识库页面。
 - c. 在“新建知识库”页面设置知识库信息。在“模型设置”页签，设置“NLP模型服务”，选择上一步新建的模型服务。

其他参数可以保持默认值或基于业务场景参考[新建知识库](#)完成配置。

- d. 设置好后，单击“确定”完成知识库创建。
6. 设置模型的Prompt。
 - a. 在知识库管理页面，选择上一步新建的知识库，单击操作列的“文档管理”，进入知识库详情页面。
 - b. 在知识库详情页面，单击“设置”，选择“更多设置”页签。
 - c. 在更多设置页面，找到“搜索增强自定义prompt”，单击“配置”，在弹窗中设置模型的Prompt。

下面以通用问答场景的Prompt举例。

你结合对话历史和给定文档回复用户请求，遵循以下原则：

1. 严格遵从文档原文术语和阐述逻辑；
 2. 如果结合对话历史和给定文档无法回复用户请求，请按照自己的知识回答。
- 文档的基础信息如下：
{0}
问题：{1}


- d. 设置完成后，单击“配置prompt”完成设置。

步骤 5：体验文档问答

1. 登录云搜索服务[管理控制台](#)。
2. 在左侧导航栏选择“KooSearch>Flexus 企业搜索服务”，进入Flexus企业搜索服务页面。
3. 选择文档问答服务，单击操作列的“问答”，前往KooSearch控制台。
4. 上传文档。
 - a. 在KooSearch控制台，左侧导航栏选择“知识库管理”，进入知识库管理页面。
 - b. 在知识库管理页面，选择5创建的知识库，单击操作列的“文档管理”。进入知识库详情页面。
 - c. 选择“文档管理”页签，单击“上传”。
 - d. 在上传对话框中单击“选择文档”，本地选择已提前准备好的文档。重复的文档不允许上传。

图 2-2 上传文档



- e. 单击“确认”。文档上传后，可在文档管理页签查看已上传的文档，当“文档状态”为“正常”，说明文档已上传成功。
5. 选择知识库。
 - a. 在KooSearch控制台，左侧导航栏选择“体验平台”，进入体验平台页面。
 - b. 单击右上角 ，在“资料来源”对话框勾选5创建的知识库，单击“确定”。

问答体验将在所选择的知识库中进行答案搜索。


6. 体验问答。
 - a. 在“体验平台”页面右上角单击“问答”，切换至问答体验页面。
 - b. 在输入框中输入问题，单击，查看返回的答案。

图 2-3 体验问答

