应用平台

快速入门

文档版本 06

发布日期 2024-12-17

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。

目 录

1 初始化 AppStage 基础信息

在使用AppStage各中心功能前，需要组织管理员对组织、人员、服务等进行初始化配置，为后续的应用开发、运维及运营等应用生命周期管理活动进行必要的环境和人员配置。

前提条件

- 已**购买AppStage**
- 已**为AppStage关联组织**
- 已**配置AppStage各中心服务授权**
- 登录用户为组织管理员。

配置基础信息

步骤1 登录**AppStage**，进入AppStge首页。

步骤2 创建部门。

1. 在AppStge首页右上角，选择“组织 > 部门管理”，进入“成员部门管理 > 部门管理”页面。
2. 单击“添加部门”，设置部门信息，如**图1-1**所示，单击“确认”。

图 1-1 创建部门

3. 在已添加的部门对应的“操作”列下，单击“添加子部门”，设置子部门信息，单击“确认”。
4. 重复执行步骤2.2或步骤2.3，可创建一个完整的部门。创建的部门信息展示在部门列表中。

步骤3 创建成员。

1. 在“成员部门管理 > 部门管理”页面左侧导航，选择“成员管理”。
2. 单击“创建成员”，设置成员信息，如图1-2所示，单击“保存”。

图 1-2 创建成员

步骤4 创建产品。

1. 在AppStge首页右上角，选择“产品与服务 > 产品管理”，进入应用基础信息管理“产品管理”页面。
2. 单击“创建产品”，设置产品信息，如图1-3所示，单击“创建”。创建的产品显示在产品列表中。

图 1-3 创建产品

3. 在已创建的产品对应的“操作”列下，单击“发布”，在弹框中单击“确定”。
   在产品列表中，新建的产品“状态”由“草稿”变为“已发布”。

步骤5 创建服务。

1. 在应用基础信息管理左侧导航，选择“服务管理”。
2. 单击“创建服务”，设置服务信息，如图1-4所示，单击“创建”。创建的服务显示在服务列表中。

图 1-4 创建服务

3. 在已创建的服务对应的“操作”列下，单击“发布”，在弹框中单击“确定”。
   在服务列表中，新建的服务“状态”由“草稿”变为“已发布”。

步骤6 创建微服务。

1. 在应用基础信息管理左侧导航，选择“微服务管理”。
2. 单击“创建微服务”，设置微服务信息，如图1-5所示，单击“创建”。创建的微服务显示在微服务列表中。

图 1-5 创建微服务

3. 在已创建的微服务对应的“操作”列下，单击“发布”，在弹框中单击“确定”。

   在微服务列表中，新建的微服务“状态”由“草稿”变为“已发布”。

----结束

2 使用开发中心进行版本管理

开发中心的重要功能是通过版本管理来管理和跟踪应用开发过程中的代码变更，是对软件、文档、代码等进行版本控制和管理的过程。它可以帮助团队协作开发，保证代码的稳定性和可靠性，同时也可以追踪历史版本，方便回溯和修复问题，进而确保团队成员之间的协作和代码的稳定性。同时版本管理是持续集成、持续交付的基础，对自动化的研发流程起到支撑作用，也对交付团队内部的协同工作起到巨大的促进作用。本章节以Scrum类型的项目为例，介绍如何使用开发中心进行版本管理。

前提条件

- 已完成**基础信息配置**。
- 已完成**团队管理**。
- 已安装Git客户端并配置SSH密钥，具体操作请参见**环境和个人配置**。
- 需要具备项目经理角色权限，权限申请方法请参见**申请权限**。具体角色权限说明请参考**用户角色和权限说明**。

创建并规划版本

步骤1 进入已创建的团队空间。

1. 登录**AppStage业务控制台**。
2. 选择“开发中心”，进入AppStage开发中心。
3. 在开发中心首页下方的“我的团队”区域，单击需要操作的团队卡片，进入该团队空间。

步骤2 创建版本。

1. 在左侧导航栏选择“版本管理”，在“版本管理”页面，单击右上角“创建版本”。
2. 在“创建版本”页面，设置版本的基本及配置信息，如**图2-1**所示，然后单击“提交”。

图 2-1 创建版本

步骤3 规划版本交付件。

1. 在“版本管理”页面的版本列表中，单击已创建版本的版本号“24.12.0.1”，进入版本详情页面。
2. 在版本详情页面的“版本持续交付”区域，选择“持续规划 > 交付件规划”，进入“交付件规划”页面。
3. 单击右上角“添加交付件”，选择交付件类型，设置交付件名称，单击“确定”。

步骤4 创建版本需求。

1. 在左侧导航栏选择“需求管理”，进入“需求管理”页面。
2. 单击“创建需求”，配置如表2-1所示参数，然后单击“确定”。

表 2-1 创建需求

参数名称	参数说明	取值示例
需求标题	输入需求标题。	【智慧语音】语音控制手表遥控拍照
归属版本	选择需求归属的版本，即已创建的版本号。	24.12.0.1
需求描述（可选）	输入需求的详细描述。	通过语音控制智能手表遥控拍照
当前处理人	指定该需求由谁处理。	Chris

----结束

开发版本

步骤1 上传版本交付件。

1. 在“版本管理”页面的版本列表中，单击已创建版本的版本号“24.12.0.1”，进入版本详情页面。
2. 在版本详情页面的“版本持续交付”区域，选择“持续开发 > 交付件管理”，进入“交付件管理”页面。
3. 在交付件列表中，单击交付件所在行“操作”列的“创建文档”。
4. 在“上传文档”对话框，选择“交付方式”为“离线文档”，单击“添加文件”将本地已准备好的文档进行上传，并单击“确定”。

步骤2 创建代码仓。

1. 在左侧导航栏选择“代码仓管理”，进入“代码仓管理”页面。
2. 单击页面右上角“创建仓库”，输入仓库名称及仓库描述，单击“确定”。

   创建后仓库列表显示该仓库，状态为“创建中”，待状态变为“使用中”，可以单击该仓库所在行“操作”列的“详情”，进入仓库详情页面，使用该仓库。

步骤3 下载代码仓到本地。

1. 在“代码仓管理”页面，单击代码仓列表中已创建的代码仓名称，查看代码仓地址。
2. 打开Git Bash客户端。

   在本地计算机上新建一个文件夹用于存放代码仓库，在空白处单击鼠标右键，打开Git Bash客户端。
3. 输入如下命令，克隆代码托管仓库。

   ```
   git clone 代码仓地址
   ```

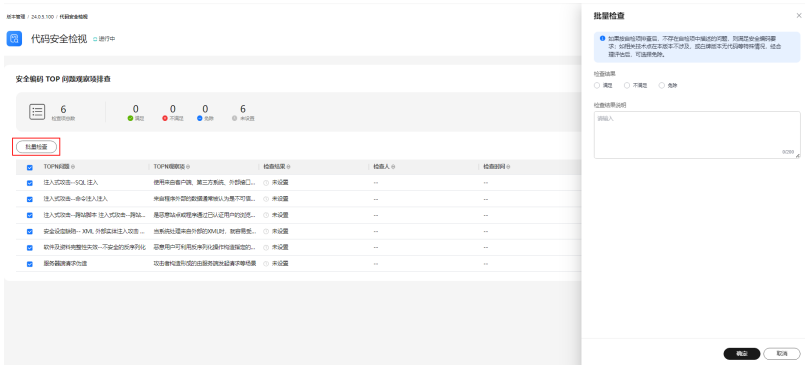
步骤4 开发代码。开发人员需要根据需求分析和设计文档，使用编程语言和开发工具编写程序代码，然后进行测试和调试，最终交付使用，详细指导请参见**《开发指南》**。

步骤5 提交代码到代码仓。在本地完成业务代码和IaC脚本后，需要提交代码文件至代码仓库，详细指导请参见**提交代码到代码托管仓库**。

步骤6 代码安全检视。

1. 在“版本管理”页面的版本列表中，单击已创建版本的版本号“24.12.0.1”，进入版本详情页面。
2. 在版本详情页面的“版本持续交付”区域，选择“持续开发 > 代码安全检视”。
3. 在“代码安全检视”页面勾选多条检查项并单击“批量检查”，在“批量检查”页面，设置检查结果并输入检查结果说明，如图2-2所示，然后单击“确定”。

图 2-2 批量检查



----结束

构建版本

步骤1 在左侧导航栏选择“流水线管理”，进入“流水线管理”页面。

步骤2 在“流水线管理”页面，选择已创建版本的版本号“24.12.0.1”，单击“关联流水线”，在“关联流水线”页面，根据界面提示单击“立即前往”新建流水线，如图2-3所示。

图 2-3 新建流水线

步骤3 在“流水线管理”页面，单击“新建流水线”。

步骤4 配置基本信息参数，参数说明如表2-2所示，单击“下一步”，选择“空模板”，单击“确定”，进入“任务编排”页面。

表 2-2 参数说明

参数名称	参数说明	取值示例
名称	输入流水线名称。	Pipeline01
代码源	选择代码源。	Repo
代码仓	选择已创建的代码仓库。	Repo01
默认分支	选择默认分支。	master

步骤5 根据需要**配置流水线**，然后单击“保存”。

步骤6 进入“流水线管理”页面，单击右上角“关联流水线”。

步骤7 在“关联流水线”页面，勾选已新建和配置完成的流水线，单击“确定”。

步骤8 在流水线列表中，单击已关联的流水线所在行“操作”列的“执行”，单击“确定”，流水线开始构建版本的发布软件包。

----结束

测试版本

在产品研发过程中，存在各团队、各项目各自为战，产品质量难管控、缺陷修复进度难追踪的问题，严重影响产品交付效率。产品特性和功能在测试验证阶段发现的问题，可以使用缺陷单进行跟踪，对于发现的缺陷进行记录、跟踪、分析和解决，确保产品质量。

步骤1 创建版本缺陷。

1. 在左侧导航栏选择“缺陷管理”，进入“缺陷管理”页面。
2. 单击“创建缺陷”，配置如**表2-3**所示参数，然后单击“确定”。

表 2-3 创建缺陷

参数名称	参数说明	取值示例
缺陷标题	输入缺陷标题。	【智慧语音】语音控制手表遥控拍照后，未获取到照片
归属版本	选择缺陷归属的版本，即已创建的版本号。	24.12.0.1
缺陷描述（可选）	输入缺陷的详细描述。	通过语音控制智能手表遥控拍照，照片无法存储或语音控制未生效
当前处理人	指定该缺陷由谁处理。	Chris

步骤2 完成测试评估。

1. 在左侧导航栏选择“版本管理”。
2. 在“版本管理”页面的版本列表中，单击已创建版本的版本号“24.12.0.1”，进入版本详情页面。
3. 在版本详情页面的“版本持续交付”区域，选择“持续开发 > 测试评估”，进入“测试评估”页面。
4. 在“测试评估”页面完成测试评估
   a. 在“总体测试结论”区域，单击右侧“编辑”，添加测试报告文件并编辑评估说明。
   b. 在“测试结论”页签，分别编辑遗留DI值、功能评估、性能评估和安全评估测试类型，给出单项测试结论。

----结束

发布版本

步骤1 版本基线化。

1. 在左侧导航栏选择“版本管理”。
2. 在“版本管理”页面的版本列表中，单击已创建版本的版本号“24.12.0.1”，进入版本详情页面。
3. 在版本详情页面的“版本持续交付”区域，选择“持续开发 > 版本基线化”。
4. 单击“添加基线化软件包”，在“选择流水线”下拉列表选择流水线，确认最近一次发布构建信息。
5. 单击右下角“基线化”。
6. 在版本详情页面的“版本持续交付”区域，选择“持续开发 > 内建质量”。
7. 在“内建质量”页面，查看到当前版本基线化后的流水线代码检查执行结果。

步骤2 检查标准发布准入信息。

1. 在版本详情页面的“版本持续交付”区域，选择“持续部署发布 > 标准发布”。
2. 在“标准发布检查结果”列表中，查看相应的检查项的检查规则、检查结论、检查结果。
3. 检查均通过后，单击“下一步”，进入“标准发布”页面。

步骤3 申请标准发布。

1. 在“标准发布”页面的“发布信息”区域，选择运维中心站点，填写发布内容。
2. 在“审核信息”区域的“项目经理”下拉列表中选择审批发布的项目经理。
3. 单击“提交”。该版本发布申请提交将生成一条待办信息至项目经理的AppStage首页的“我的待办”中，由其单击待办链接跳转至标准发布审批页面完成审批操作。

审批通过后，发布成功的版本软件包将发布到对应的部署平台。

----结束

相关信息

开发中心深度集成CodeArts的需求管理、代码仓管理、流水线管理等功能，如果不想使用CodeArts相关能力，可以配置工具链，集成并使用其他产品能力，具体操作请参见**管理AppStage开发中心系统配置**。

3 使用运维中心统一管理资源

应用平台运维中心提供了一站式智能化运维平台，助力企业提升运维质量、效率与可靠性。您可以将公有云已创建的资源纳管至运维中心进行统一管理。

前提条件

- 已创建VPC和子网，具体操作请参见**创建虚拟私有云和子网**。
- 已购买主机，具体操作请参见**购买弹性云服务器ECS**或**购买裸金属服务器BMS**。
- 已购买数据库实例，具体操作请参见**购买GaussDB(for MySQL)实例**、**购买GaussDB实例**、**购买RDS for PostgreSQL实例**、**购买GeminiDB Cassandra实例**或**购买RDS for MySQL实例**。
- 已购买CCE容器集群，具体操作请参见**购买集群**。
- 已完成**基础信息配置**。
- 已获取基础运维岗位权限或运维管理员权限，权限申请操作请参见**申请权限**。
- 待纳管主机的服务已**规划业务账号**。

约束限制

当前仅部分区域的主机支持接入AppStage运维中心，包括华北-北京四、华南-广州、华东-上海一、华东-上海二和华北-乌兰察布一 ，如需接入其他区域的主机，请联系技术支持工程师。

资源接入运维中心

步骤1 **进入AppStage运维中心**。

步骤2 在顶部导航栏选择需要接入资源的服务。

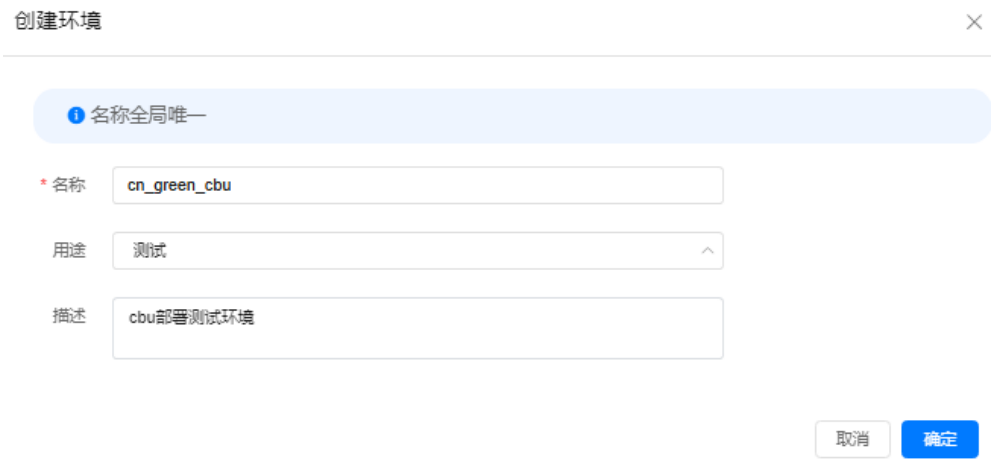
步骤3 单击“运维接入一站式地图”后的“接入引导”，如**图3-1**所示，进入“运维中心一站式接入流程”页面。

图 3-1 接入引导

步骤4 配置环境。

1. 在公共配置区域，选择需要接入资源所属的账号及Region。
2. 在配置环境区域，单击“创建环境”。
3. 输入环境名称，选择用途并输入环境描述，然后单击“确定”，如图3-2所示。创建并启用环境，同时将环境与所选的公共配置关联。

图 3-2 创建环境



4. 单击“下一步：纳管VPC”。

步骤5 纳管VPC。

1. 在纳管VPC区域，单击“创建纳管”。
2. 选择需要纳管的虚拟私有云（VPC），并选择终端节点子网，然后单击“确定”，纳管VPC并将VPC与所选环境关联。

说明

纳管VPC时运维中心会自动创建的两个终端节点，终端节点会产生费用，按终端节点实例的实际使用时长计费，如需查看费用账单请参见**费用账单**。

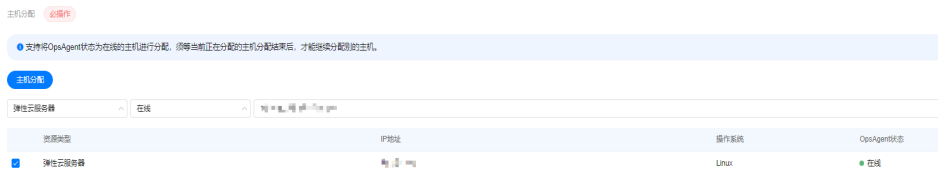
3. 单击“下一步：纳管主机”。

步骤6 纳管主机。

以Linux主机为例，介绍如何根据纳管主机指引完成首次纳管Linux主机。

1. 为Linux主机手动安装OpsAgent。
   a. 单击CURL命令后的，复制安装命令。
   b. 使用root账号远程登录主机后，执行安装命令安装OpsAgent。
2. 主机分配。勾选待纳管的主机，单击“主机分配”完成主机纳管，如图3-3所示。

图 3-3 主机分配

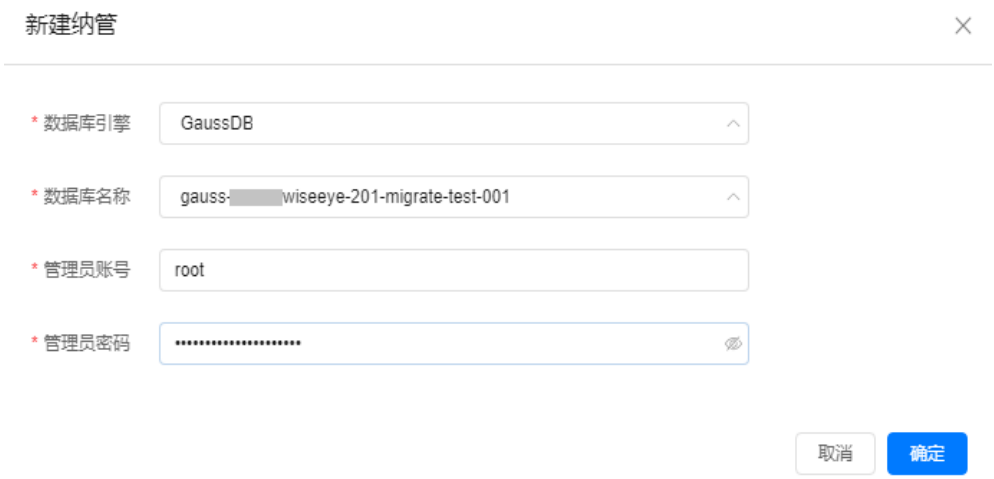


3. 设置执行机。单击已纳管的主机所在行后的“设置执行机”。
4. 单击“下一步：纳管数据库”。

步骤7 纳管数据库。

1. 在纳管数据库区域，单击“新建纳管”。
2. 配置纳管数据库相关参数，如图3-4所示，然后单击“确定”，完成数据库纳管。

图 3-4 纳管数据库



3. 单击“下一步：纳管容器集群”。

步骤8 纳管容器集群。

1. 在纳管容器集群区域，单击“新建纳管”。
2. 选择需要纳管的华为云CCE集群，并输入集群别名。然后单击“确定”，完成容器集群纳管。
3. 单击“完成”，完成所有资源接入。

----结束

相关信息

资源接入后，您还可以进行如表3-1所示操作。

表 3-1 相关操作

资源类型	相关操作
终端节点	删除终端节点：在华为云控制台选择“网络 > VPC终端节点”，进入“终端节点”页面删除，具体操作请参见删除终端节点。
VPC	取消纳管VPC：在弹性网络服务“网络规划 > VPC纳管”页面的VPC列表中，单击VPC所在行“操作”列的“取消纳管”。

资源类型	相关操作
OpsAgent	• 升级OpsAgent：在主机管理服务“云服务器管理 > 弹性云服务器”页面的主机列表中，勾选主机，单击列表上方的“升级OpsAgent”。 • 卸载OpsAgent：在主机管理服务“云服务器管理 > 未纳管主机”页面的主机列表中，勾选主机，单击列表上方的“卸载OpsAgent”。 **说明** 卸载OpsAgent前需要先取消纳管主机。
主机	• 管理主机：具体操作请参见**管理已纳管的主机**。 • 取消纳管主机：具体操作请参见**取消弹性云服务器主机纳管**。
数据库	• 管理数据库：具体操作请参见**在WiseDBA中纳管和管理数据库实例**。 • 取消纳管数据库：在数据库治理“实例管理 > 实例列表”页面的数据库实例列表，单击数据库实例所在行“操作”列的“更多 > 取消纳管”。
容器集群	• 管理容器集群：具体操作请参见**在ERS管理已纳管的容器集群**。 • 取消纳管容器集群：在弹性资源服务“集群列表”页面，单击集群所在行“操作”列的“更多 > 删除纳管”。

4 使用监控服务进行主机运维监控

AppStage运维中心支持将华为云主机接入主机管理服务（VMS）进行统一管理，并使用监控服务（ServiceInsight）的日志、监控和告警功能对主机进行运维监控。

本章节以Linux主机为例介绍如何使用监控服务进行主机运维监控的操作。

- 日志：完成**主机日志接入**，将日志接入监控服务中，接入后可以在“日志检索”页面查看已接入日志，具体操作请参见**查看已接入日志**。
- 监控：完成**主机监控接入**，为主机绑定监控模板，根据监控模板定义的插件采集主机监控数据，监控数据接入后可以在“虚机报表”页面查看，具体操作请参见**查看虚机报表**。
- 告警：完成**主机告警配置**，定义告警上报内容，然后可以在“告警列表”页面查看并处理已上报告警，具体操作请参见**查看并处理告警**。

前提条件

- 已完成**Linux主机接入**。
- 日志接入或者告警配置需要获取服务运维岗位权限或运维管理员权限，权限申请操作请参见**申请权限**。

主机日志接入

步骤1 创建日志项目。

1. **进入AppStage运维中心**。
2. 在顶部导航栏选择服务。
3. 单击☰，选择“运维 > 监控服务（ServiceInsight）”。
4. 选择左侧导航栏的“日志 > 日志接入”。
5. 在“日志接入”页面，选择左侧导航栏的“日志项目”，单击“创建日志项目”。
6. 自定义日志项目名称并输入日志项目描述，单击“创建”。

步骤2 创建日志空间

1. 在“日志接入”页面，选择左侧导航栏的“日志空间”。
2. 单击“申请实时日志空间”。

3. 配置实时日志空间信息，参数说明如表4-1所示，配置完成后，单击“下一步”。

表 4-1 实时日志空间信息参数说明

参数名称	参数说明
空间名称	自定义日志空间名称，建议包含日志类型语义。
空间描述	输入日志空间描述，非必填项。
日志类型	选择需接入的日志类型。
日志大小	预计一天的日志量，默认为1GB。
开启日志检索	如果需要使用日志检索功能，可以打开该开关。
检索空间类型	选择ClickHouse。
原索引名称(ClickHouse)	可选择现有的ClickHouse，如果不填会自动生成。
TTL	日志索引的生命周期，即可以检索的日志时间范围。

4. 配置实时日志字段信息，参数说明如表4-2所示，配置完成后，单击“下一步”。

表 4-2 实时日志字段信息参数说明

参数名称	参数说明
自定义字段	勾选需要接入的日志字段，包括通用字段、容器字段和虚机字段。
新增自定义环境变量	如需添加自定义环境变量，请选择环境变量名，然后单击“添加”。虚机暂无可选自定义环境变量。
清洗规则	选择日志清洗规则。 请优先使用算子清洗模式采样，通过配置解析脚本将原始日志清洗为业务需要的日志字段，算子清洗功能及使用样例请参见**算子清洗功能介绍**。原始日志采样清洗只适用于单纯采样，不需要清洗的场景。
日志样例	输入日志样例。
解析脚本	配置解析脚本，将日志样例清洗为字段显示。 **说明** 配置解析脚本时字段命名不支持使用中划线“-”，支持使用下划线“_”。
清洗字段	配置解析脚本后单击“配置解析脚本”，自动生成清洗字段，查看字段是否符合预期。

参数名称	参数说明
开启汇聚	选择是否开启日志汇集，如果日志量较大且不需要关注原始日志时可以进行日志汇集。 开启后需要配置汇集相关参数。
汇聚粒度	开启汇聚后，需要设置汇集粒度。支持分钟级和秒级数据汇聚。选择分钟级，每一分钟会生成一个统计点，选择秒级，每一秒会生成一个统计点。
汇聚时间戳	仅支持时间戳格式字段timestamp，获取当前计算的日志的时间。
时间戳格式	选择时间戳格式。支持秒、毫秒、纳秒级时间戳，获取当前计算的日志的时间格式。
汇聚维度	结合业务场景需要，选择日志是以哪些日志字段进行日志汇聚，支持多选。
汇聚度量	设置对日志字段以COUNT、SUM、MAX、MIN进行度量。 原始字段是日志中的字段，用来获取原始值；度量字段是用户自定义字段名，计算后，度量的值会赋值给该字段。
输出原始日志	选择是否需要输出原始日志。如果打开输出原始日志，原始日志也会上报。

5. 日志字段确认，确认日志字段配置是否达到预期，达到预期后可单击“下一步”。
6. 申请日志空间共享，如果需要其他自有服务共用这个空间进行日志下发和日志检索，可以添加共享服务。配置完成后，单击“保存”。

步骤3 创建日志采集配置。

1. 在“日志接入”页面，选择左侧导航栏的“日志采集配置”。
2. 单击“创建日志采集配置”。
3. 配置日志采集参数，配置完成后，单击“确定”。

表 4-3 日志采集配置参数说明

参数名称	参数说明
日志项目	选择已创建的日志项目，相同服务的不同日志使用同一个日志项目。
日志空间	选择已创建的日志空间。选择日志空间时日志提取规则会展示日志空间定义的日志格式，采集的日志须满足对应格式。
配置名称	自定义日志采集配置名称。
配置类型	选择日志采集配置类型，建议选择“FILEBEAT”。
日志类型	输入采集日志类型。

参数名称	参数说明
日志路径	填写实际日志路径，可使用通配符进行匹配。 **说明** – 接入容器日志需要根据通配符匹配完成。 – 注意避免同一台主机上下发的多个采集任务重复采集相同的日志文件，会导致filebeat进程异常。
日志TPS	TPS表示单实例每秒日志条数，请准确填写，用于推荐资源自动计算。 – 如果采集路径是单个日志，则按照单个日志单台机器（pod）的TPS值填写，且按照高峰期计算。 – 如果采集路径配置了通配符，则将采集的日志TPS累加，累加计算高峰期单台机器（pod）的TPS，建议按近期业务增长预期填写。
日志模式	选择日志采集模式，是单行模式还是多行模式。
是否支持软连接	当填写的日志路径为链接路径时，需要开启支持软连接。
首行正则表达式	日志模式选择多行模式时，需要输入首行正则表达式。
日志提取规则	根据填写的配置参数会自动生成提取规则。

步骤4 创建日志配置下发任务。

1. 在“日志接入”页面，选择左侧导航栏的“任务管理”。
2. 单击页面右上角的“新建任务”。
3. 配置任务参数，参数说明如表4-4所示，配置完成后，单击“确定”。

表 4-4 日志配置下发任务参数说明

参数名称	参数说明
日志项目	选择已创建的日志项目。
任务名称	自定义任务名称。
任务类型	选择任务类型。
配置类型	选择日志采集配置类型。
配置列表	选择需要下发的配置。
用户名称	选择已规划并拥有日志读取权限的业务账号。
选择主机	选择需要下发配置的主机。
已选主机	显示已选主机。

4. 在任务列表中查看已创建的任务，单击任务所在行“操作”列的“执行”。
   执行完成后，状态为成功即表示日志配置内容已下发成功，即会按照配置将日志接入AppStage。

----结束

主机监控接入

步骤1 进入AppStage运维中心。

步骤2 在顶部导航栏选择服务。

步骤3 单击，选择“运维 > 监控服务（ServiceInsight）”。

步骤4 选择左侧导航栏的“运维数据采集 > 模板管理”。

步骤5 单击“新建”，进入“新建模板”页面。

步骤6 输入模板名称、选择模板类型、输入模板版本，也可为模板添加说明。

步骤7 单击“已选中插件的具体详情”后的。

步骤8 在“选择插件”页面单击需选择插件后的或。表示可以选择多次，表示只能选择一次。

步骤9 关闭“选择插件”页面，在“新建模板”页面可以对已选择的插件参数进行编辑，单击已选插件名称后的，如图4-1所示。

图 4-1 编辑插件

步骤10 在“配置参数”页面编辑插件参数，编辑完成后单击“确定”。

步骤11 配置完成后，在“新建模板”页面单击“确定”。

步骤12 选择左侧导航栏的“运维数据采集 > 绑定管理”。

步骤13 在主机列表，单击待绑定主机所在行“操作”列的“配置监控”。

步骤14 勾选模板后单击  ，单击“确定”。

----结束

主机告警配置

步骤1 进入AppStage运维中心。

步骤2 在顶部导航栏选择服务。

步骤3 单击  ，选择“运维 > 监控服务（ServiceInsight）”。

步骤4 选择左侧导航栏的“告警 > 策略配置”。

步骤5 单击“统一告警定义”，进入告警定义页面。

步骤6 单击“创建”。

步骤7 配置AIOps规则参数，参数说明如表4-5所示，配置完成后，单击“确定”。

表 4-5 AIOps 规则参数说明

参数名称	参数说明
指标来源	选择告警的指标来源“AIOps”。
告警定义名称	自定义告警定义的名称。
级别	选择该规则生成告警的级别。
告警类型	选择告警类型，上报的告警会显示类型信息，可根据类型筛选查看告警。
指标	选择在指标仓库已创建的指标，创建指标请参见在运维中心指标仓库创建指标。
维度列表	来自于指标的逻辑实体上的维度，选择异常检测需要对哪些维度做检测。
ALL维度列表	选择需要过滤的维度。
维度过滤设置	只关注维度部分取值时，可以设置该参数对维度取值进行过滤。
指标类型	选择指标类型。
算法类型	选择固定阈值或动态阈值，固定类型还需要设置阈值的上限、下限和预估维度数。

----结束

5 使用应用平台进行应用运营

本文旨在帮助您对运营中心的入门操作有初步的认识，有助于您快速掌握运营中心的基本功能。

使用运营中心的步骤如下所示：

步骤一：新建数据源

步骤二：新建数据接入

步骤三：应用指标模板

步骤四：自定义运营看板

步骤五：看板

前提条件

需要具备AppStage指标开发者或运营管理员权限，权限申请操作请参见**申请权限**。

步骤一：新建 OBS 数据源

数据源是数据分析的基础，首先要将数据源接入运营中心后再分析数据。

1. 登录**AppStage业务控制台**。
2. 在快捷入口选择“运营中心”，进入运营中心。
3. 在左侧导航栏选择“数据管理 > 数据源管理”，进入数据源页面。
4. 单击“新建数据源”，进入新建数据源页面。
5. 配置数据源基础信息，参数说明如**表5-1**所示。

表 5-1 基础配置参数说明

参数	说明
数据源名称	必填。数据源的名称或标识符。 长度不超过20字符。
数据源类型	必填。数据源的类型。 取值：OBS数据源

参数	说明
OBS终端节点	OBS（Object Storage Service）的终端节点（Endpoint），用于连接到OBS服务并访问存储在OBS中的数据。 **说明** OBS终端节点的获取方法如下： 1. 获取已授予OBS桶读写权限的账号。对账号授予桶的读写权限的方法，请参见**对其他账号授予桶的读写权限**。 2. 登录控制台，选择“存储 > 对象存储服务 OBS”，进入OBS控制台。 3. 鼠标移动到指定桶所在行，在弹出的基本信息中，“Endpoint”参数值即为OBS终端节点。
端口号	必填。用来标识一个网络设备上的应用程序或服务，以便其他设备可以通过端口号来访问这个应用程序或服务。 该参数默认值为443。
访问标识（AK）	Access Key，是一种身份验证凭证，用于标识访问者的身份。访问者需要提供正确的AK才能访问数据源。
密钥（SK）	Secret Key，与AK配对使用，用于加密和解密访问请求和响应。只有持有正确的AK和SK才能访问数据源。

6. 配置完成后，单击“测试连接”。

   说明

   为保证配置成功，推荐测试连接。如果测试连接不通，说明数据源配置有问题，需重新配置。

7. 测试连接提示连接成功后，单击“保存”。

步骤二：新建 OBS 数据接入

运营中心提供通用数据接入能力，通过与数据源连接，可以建立数据源中数据表的迁移任务，将源端数据迁移到运营中心，为后续业务做数据准备。

1. 在运营中心左侧导航栏选择“数据管理 > 通用数据接入”，进入通用数据接入页面。
2. 单击“新建数据接入 > OBS接入”，进入新建数据接入页面。
3. 配置数据接入基础信息和任务信息，参数说明如**表5-2**和**表5-3**所示。

表 5-2 基础信息参数说明

参数	说明
数据接入名称	必填。数据的名称或标识符。 长度不小于3，且不大于20。
数据源	必填。参数枚举值在“数据源”页面配置。

参数	说明
OBS数据路径	必填。数据在OBS桶中的路径。当配置“数据源”后，该参数显示。 **说明** • OBS数据路径可选择文件夹与文件。如果选择文件夹要确保该文件夹下有csv或json格式的文件。 • OBS数据路径中，如果同个文件夹里包含多个文件，请确保其文件格式一致。 • OBS数据路径中，只能包含中文、英文、数字、下划线。
数据格式	必填。数据的格式。 • JSON：表示扩展名为.json的文件。 • CSV：表示扩展名为.csv的文件。 **说明** • CSV文件支持UTF-8格式。 • JSON文件、CSV文件不能包含pt_d字段，且需要包含时间格式的字段。

表 5-3 任务配置参数说明

参数	说明
任务类型	指定配置的任务类型。 • 周期性任务 • 一次性任务
是否存在历史数据	根据实际情况选择“是”或“否”。周期性任务数据不会导入以前的数据，如果需要查看以前的数据，该参数需配置为“是”。 仅当“任务类型”为“周期性任务”时，该参数为必填参数。
运行周期	指定配置任务的运行周期。仅当“任务类型”为“周期性任务”时，该参数为必填参数。 默认值：天

4. 配置完成后，单击“下一步”。
5. 在数据表结构页面，选择“分区字段”。

   说明

   选择时间字段为分区字段。

6. 请校验数据表结构中的字段类型，如果不匹配可能会导致后续的数据建模及分析出现异常或错误。

   说明

   数据表结构中的“字段名”只能包含字母、数字、下划线。

7. 校验数据表结构完成后，单击“保存并启用”。

步骤三：应用指标模板

运营中心提供指标模板应用能力，模板中内置数据模型、指标定义、图表卡片等经验内容。运营中心通过提供多场景模板，以场景驱动一键应用，赋能用户低门槛、高效率地构建指标体系。

1. 在运营中心左侧导航栏选择“运营模板 > 指标模板库”。
2. 从指标模板库查找所需指标模板，单击选中的模板。
3. 在“模板详情”页面，选中模板某指标，进入“应用模板”页面。
4. 在“应用模板”页面，完成字段映射，如表5-4所示。

表 5-4 字段映射参数说明

参数	说明
模型显示名	在数据分析或监控系统中显示的模型名称，通常是一个更易于理解和识别的字符串。
源表	指需要从中抽取数据的源数据表。选择源表后，单击源表后面的 ⊕，可添加引用表，如表5-5所示。
字段映射	指将源表和引用表中的字段，与模板字段进行映射。

表 5-5 引用表参数说明

参数	说明
源表	指需要从中抽取数据的源数据表。
引用表	指需要与源表进行关联的参考数据表。
字段关系	指源表和引用表之间的字段关系。 **说明** 最少需输入1个字段关系。

5. 单击“确定”，完成指标模板应用配置。

步骤四：自定义运营看板

通过自定义运营看板，可以把产品运营中的关键数据统一呈现出来，可按人员权限和业务类型展示不同的数据看板，可视化展现产品运营现状。

1. 使用在线构建图表卡片。
   a. 在运营中心左侧导航栏选择“看板管理 > 我的卡片”。
   b. 在“我的卡片”页面中，选择卡片分类，单击“新建卡片”。
   c. 选择卡片创建方式。在“创建方式”下单击“在线构建”。
   d. 在左侧“组件库”下，拖拽组件至中间画布中。
   e. 在左侧“图层”下，为组件设置显示效果。
   f. 在右侧设置数据服务，数据服务设置请参考表5-6。

表 5-6 图表卡片数据服务参数说明

参数	说明
数据源类型	选择数据源类型。支持静态数据、Restful。 静态数据：在JSON编辑器中，编辑静态数据。 ●  ：数据校验，校验JSON文件数据格式正确性。 ●  ：复制，复制JSON文件。
类型	当“数据源类型”值为“Restful”时，才有此参数。 ● 默认 ● 指标
数据服务URL	当“数据源类型”值为“Restful”且“类型”值为“默认”时，才有此参数。 数据服务的URL。
请求类型	当“数据源类型”值为“Restful”且“类型”值为“默认”时，才有此参数。 在下拉框中选择请求类型。支持GET、POST。
认证	当“数据源类型”值为“Restful”且“类型”值为“默认”时，才有此参数。 在下拉框中选择认证方式。支持No auth、BasicAuth、Token。 ● No auth：Restful API不对SVE进行鉴权。 ● BasicAuth：Restful API会对SVE进行BasicAuth认证，此时，还需要填写鉴权用户名和密码。 ● Token：Token方式，会自动获取当前登录运营中心的用户对应的Token。 ● AK/SK：Restful API会对SVE进行AK/SK方式认证。
请求头	当“数据源类型”值为“Restful”且“类型”值为“默认”时，才有此参数。
指标列表	当“数据源类型”值为“Restful”且“类型”值为“指标”时，才有此参数。 在下拉框中选择关联指标。指标在“指标模板库”或“指标管理”界面配置。 ●  ：数据格式化，校验JSON文件数据格式正确性。 ●  ：复制，复制JSON文件。

参数	说明
请求参数	当“数据源类型”值为“Restful”时，才有此参数。 单击⊕，填写Key和Value的值。
请求间隔（秒）	当“数据源类型”值为“Restful”时，才有此参数。 设置请求时间间隔。
查看服务返回结果	单击“查看服务返回结果”，进入“数据响应结果”页面，进行数据影射，配置完成后自动保存。

g. 在右侧设置组件属性。不同组件显示的组件属性不同，具体以界面显示为准。这里以“图表组件 > 折线图”为例，请参考表5-7。

表 5-7 图表卡片组件属性参数说明

参数		说明
基础属性		• 绑定数据：绑定数据源，在下拉框中设置。数据源在5的“查看数据返回结果”中进行设置。 • 组件样式：设置组件的宽高及XY轴的值。
特有属性	常规配置	设置组件上下左右的边距。
	轴线配置	• 折线属性：设置折线类型、粗细，根据需求开启曲线、折叠、阶梯型开关。 –  ：表示关闭。 –  ：表示开启。 • 标记：设置标记点的类型、大小。 • 标线/最大值/最小值/区域：根据需求开启并设置。
	数值标签	根据需求开启并设置。
	坐标轴	设置是否显示X/Y/Z轴，以及X/Y/Z轴的文本样式、位置、轴线样式、刻度样式、分割线样式、指示器颜色及透明度。
	系列	当“坐标轴”设置显示Z轴后，选择折线展示的基准轴线。
样式属性	图例	根据需求开启图例，设置文本样式、尺寸、标记、间距、排列方式、位置、是否滚动。
	提示框	根据需求开启提示框，设置文本样式、背景颜色、边距。

参数		说明
	主题	选择主题信息，根据需求开启自定义颜色。

h. 在右侧设置卡片信息，请参考表5-8。

表 5-8 卡片信息参数说明

参数	说明
名称	必填项。 自定义卡片的名称。由1~64个字符组成，包含中文、字母、数字及下划线。
版本	卡片的版本信息，不可修改。
样式分类	卡片的分类，不可修改。
卡片分类	必选项。 在下拉框中选择卡片的分类名称。默认显示第一个分类名称。
描述	必填项。 自定义卡片的描述信息。
封面图	卡片的封面图。 • 自动截图：单击“截屏”，自动获取画布中的图片。 • 上传封面：单击“添加图片”，在弹出界面，单击  ，选择本地准备好的图片，裁剪出符合要求的尺寸后，单击  。

i. 在界面右上方单击“保存”，系统自动弹出“保存”页面，确认卡片信息，单击“保存”。

j. 在界面右上角单击“发布”。

k. 返回“我的卡片”界面，发布后的卡片，默认显示在“我的卡片”所选择的卡片分类列表下。

2. 快捷构建屏幕。

a. 在运营中心左侧导航栏选择“看板管理 > 我的屏幕”。

b. 在“我的屏幕”界面，选择屏幕分组，单击“新建屏幕”。

c. 在“选择构建方式”页面，在“快捷构建”下单击“屏幕构建”。

d. 配置新建屏幕基本信息，参数说明请参考表5-9。

表 5-9 参数说明

参数	说明
屏幕名称	屏幕的名称，由1~64个字符组成，包含中文、字母、数字及下划线。屏幕名称满足唯一性。
屏幕分组	选择相应的分组。可以在“我的屏幕”页签，选择项目后添加分组。
屏幕描述	屏幕相关描述，由1~400个字符组成，包含中文、字母、数字及下划线。

e. 单击“确定”，进入“运营中心公用模板”界面。

f. 在右侧配置“屏幕属性”，具体的参数说明如**表2 参数说明**所示。

表 5-10 参数说明

参数	说明
基础属性	• 屏幕名称/屏幕分组/屏幕描述：可修改。 • 屏幕尺寸：可对屏幕尺寸进行修改，包含默认、2K屏、4K屏和自定义。 • 编辑状态：可以通过切换编辑状态设置屏幕是否可编辑。 • 主题背景：可以单击“切换主题背景”切换主题风格。 • 渐进渲染/屏幕水印/版权信息/悬浮按钮：根据需求开启。
版面属性	• 屏幕全屏/平台渲染/微前端加载：根据需求开启或关闭。 • 缩放方式：分为全屏铺满、固定高宽、等比缩放、等比缩放宽度铺满、等比缩放高度铺满。 • 可对上下左右边距进行自定义调整。
其他	屏幕封面包含以下两种： • 截取封面：单击“截取封面”，即可自动获取封面。 • 上传封面：单击“上传封面”后，将本地准备好的封面进行上传。

g. 在右侧配置“卡片属性”，包括“基本信息”、“交互”和“属性”，具体的参数说明如**表3 参数说明**所示。

表 5-11 参数说明

参数		说明
基本信息	基础信息	卡片的基本信息，不可修改。

参数		说明
交互	交互方式	● 无：无交互方式。 ● 联动：一个区域可有多张卡片，可设置切换。 – 局部事件：当打开局部事件时，仅对选中卡片的选中区域实现交互效果；反之，则是对选中卡片的全局产生效果。 – 显示目标：选择相应卡片，与主体卡片产生联动交互效果。 – 隐藏目标：选择相应卡片，隐藏卡片。 ● 弹出：可设置屏幕内卡片的弹出方式。 – 触发方式：可选值为“点击”。 点击：任意点击卡片即可触发弹出效果。 – 弹出方式：可选值为“弹窗”“侧边抽屉”。 – 局部事件：当“触发方式”为“点击”时，才有此参数。当打开局部事件时，仅对选中卡片的选中区域实现交互效果；反之，则是对选中卡片的全局产生效果。 – 弹出卡片：选择相应卡片，与主体卡片产生弹出交互效果。选择卡片时支持搜索选择。 – 弹窗宽高：设置弹窗尺寸。 ● 下钻： – 局部事件：请选择已设置发送消息event1的卡片，否则不会生效。 发送消息event1在“屏幕模板 > 消息联动 > 自定义事件”进行设置。 当开启局部事件时，可设置多个下钻屏幕。 – 当值为：当开启局部事件时，才有此参数。卡片局部事件传出来的值。当配置为“*”时，适用于所有局部事件。 – 下钻屏幕：选择下钻的屏幕。 – 链接打开：下钻的屏幕显示窗口。 – 参数：设置参数。 – 效果：可选项“无”“滑窗”“翻页”“缩放”。
	卡片特效	可根据需要对卡片配置“无”“飞入”、“浮入”、“滑入”效果。 **说明** Windows操作系统，需要在“我的电脑 > 高级系统设置 > 高级 > 性能 > 设置 > 视觉效果 > 自定义”中勾选“窗口内的动画控件和元素”，才能显示卡片特效。电脑重启后，需要再次设置。
	扩展交互	支持卡片全屏、卡片显隐、卡片固定位置。

参数	说明
属性	可对卡片的内边距、标题、布局、数值及柱状样式等进行自定义配置。

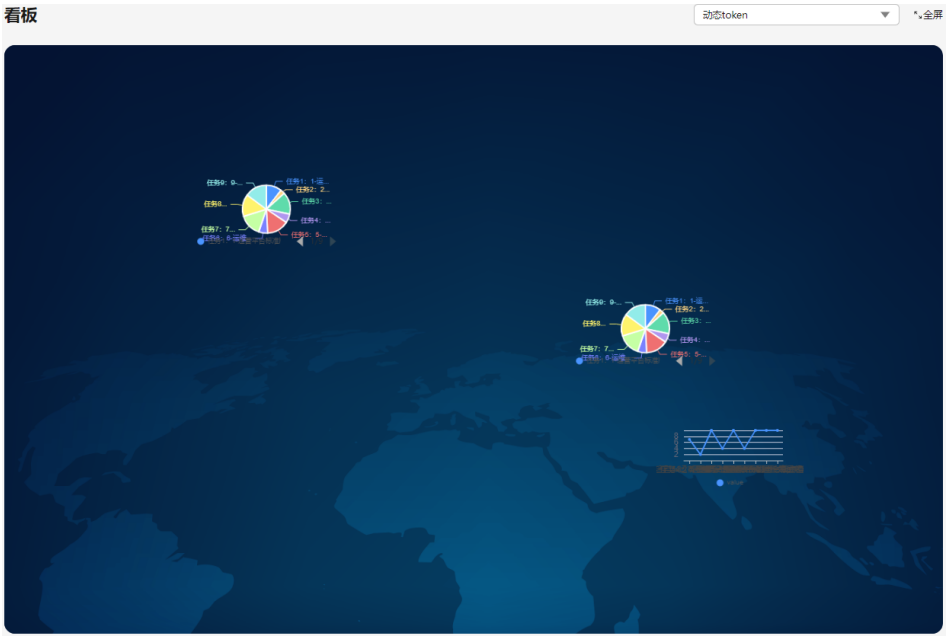
h. 在右侧配置“屏幕卡片”，可以对卡片进行升级、删除、隐藏、展示等操作。
i. 配置完成后，单击“完成”。
j. 在“我的屏幕”界面，选择创建的屏幕，单击“编辑”，可以再次返回屏幕构建页面对相关配置进行修改。
k. 单击发布，在“发布”界面，根据需求开启分享状态，单击“确定”。

步骤五：看板

产品运营人员能直观查看关键数据，分析产品运营过程中取得的成效和潜在问题。

在左侧导航栏选择“看板”，如图5-1所示。

图 5-1 看板



6 使用 AI 原生应用引擎完成模型调优

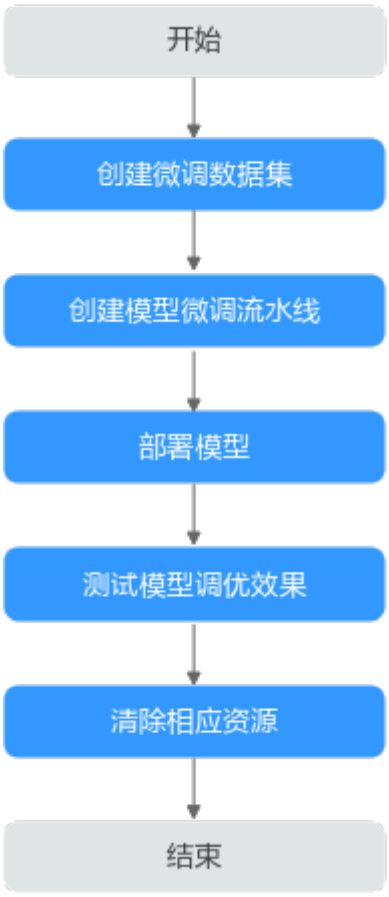
模型调优是一种通过对模型进行微调来适应特定任务或数据集的技术，适用于需要个性化定制或在特定任务上追求更高性能表现的场景。在模型调优过程中，用户需要构建一个符合业务场景任务的训练集，这个训练集通常由业务数据和业务逻辑构成。然后，用户需要调整模型的参数，以便模型可以更好地学习这个训练集。最后，用户需要训练模型，让模型可以在业务场景中更好地表现。模型调优的目标是提高模型在特定任务或数据集上的性能。通过模型调优，用户可以让模型更好地适应业务场景，从而提高模型在实际应用中的效果。

本文以AI原生应用引擎预置的baichuan-13B为基座模型，通过微调提升模型的数据分析和指标计算能力，帮助您快速掌握使用AI原生应用引擎完成模型调优并测试效果。

使用流程

通过图6-1您可以了解如何在AI原生应用引擎创建数据集、创建微调任务、部署推理模型以及在线测试模型效果。

图 6-1 模型调优流程



1. **创建微调数据集**
   数据集是模型微调的基础，首先需要创建用于模型训练的数据集。
2. **创建模型微调流水线**
   通过模型微调任务进行模型训练，微调任务结束后，将生成改进后的新模型。
3. **部署模型**
   模型部署是通过为基座模型（即原模型）和微调后的新模型创建用于预测的模型服务的过程实现。
4. **测试模型调优效果**
   在线测试微调后的模型（输入问题发起请求获取数据分析结果），可检验模型的准确性、可靠性及反应效果。
5. **清除相应资源**
   对于不再需要使用的微调任务和模型服务，及时清除相应资源，避免不必要的扣费。

准备工作

- 已**注册华为账号并开通华为云**。
- 已**实名认证华为账号**。
- 使用AI原生应用引擎前请检查账号状态，确保账号未处于欠费或冻结状态。
- 已订购大模型微调-SFT局部调优资源，订购方法请参见**购买AI原生应用引擎按需计费资源**。

步骤一：创建微调数据集

数据集是模型微调的基础，AI原生应用引擎统一纳管训练模型的数据集，将分散的数据进行集中式管理，从而节省了数据收集和管理的成本。

步骤1 在AI原生应用引擎的左侧导航栏选择“知识中心 > 微调数据集”。

步骤2 在“微调数据集”页面，单击右上角“创建微调数据集”。

步骤3 在“创建微调数据集”页面，参照表6-1进行相关参数的配置。

表 6-1 数据集基础配置参数说明

参数名称		参数说明
基础配置	数据集名称	自定义数据集名称。支持中英文、数字、下划线（ _ ），长度2-50个字符，以中英文、数字开头。 本文以创建名称为“智能分析数据集”为例。
	数据集描述	根据自己的需要对数据集进行描述，例如介绍数据集的用途、样例数据等。
	标签	根据自己需要选择标识该适用数据集的模型、领域、行业等，例如大语言模型。
	任务领域	选择“自然语言处理”。
	数据集格式	选择“对话文本”。 文件内容要求为标准json数组，例如： [{"instruction": "aaa", "input": "aaa", "output": "aaa"}, {"instruction": "bbb", "input": "bbb", "output": "bbb"}]
数据接入	数据来源	选择数据集的数据来源。支持以下两种来源： • 本地上传 • OBS接入
	本地上传	当“数据来源”选择“本地上传”时，需配置此参数。 单击“文件上传”选择本地JSON格式的文件进行上传（仅支持JSON格式）。
	OBS桶名	当“数据集来源”选择“OBS接入”时，需配置此参数。 在下拉列表中选择数据所在的OBS桶名。
	OBS路径	当“数据集来源”选择“OBS接入”时，需配置此参数。 在下拉列表中选择数据所在的具体OBS路径。
	调度类型	可选如下两种类型，其中本地文件上传仅支持一次性调度，OBS接入支持一次性调度或定时调度两种类型。 • 一次性调度 • 定时调度

参数名称		参数说明
	版本模式	当“调度类型”选择“定时调度”时，需配置此参数。 ● 覆盖模式：每次调度成功，会覆盖唯一的版本。 ● 多版本模式：每次调度成功，会生成一个新版本。
	执行时间	当“调度类型”选择“定时调度”时，需配置此参数。 设置每日执行时间。
	立即执行	当“调度类型”选择“定时调度”时，需配置此参数。 选择是否立即执行。

步骤4 单击“提交”。创建的数据集显示在“我创建的”页签的数据集列表中，创建数据集完成。

----结束

步骤二：创建模型微调流水线

模型微调任务是指调整大型语言模型的参数以适应特定任务的过程，通过在与任务相关的数据集上训练模型来完成。所需的微调量取决于任务的复杂性和数据集的大小。在深度学习中，微调用于改进预训练模型的性能。操作本步骤前请确保以下两点：

- 已订购大模型微调服务API在线调用-SFT局部调优，订购方法请参见**购买AI原生应用引擎按需计费资源**。
- 已创建格式为“对话文本”的微调数据集。

步骤1 在AI原生应用引擎的左侧导航栏选择“模型中心 > 模型微调流水线”。

步骤2 在模型微调流水线页面，单击“创建微调任务”，选择“通用能力增强微调”。

步骤3 参照**表6-2**配置基础信息、模型、数据及任务。

表 6-2 创建微调任务参数说明

参数名称		参数说明
基础信息	任务名称	自定义任务名称，例如“baichuan-13b-chat-sft-nl2sql”。
	任务描述(可选)	自定义任务相关的描述。
模型配置	微调前模型	选择微调的模型，本文案例选择“baichuan”。
	训练模式	默认选择“LoRA”。 LoRA（Low-Rank Adaptation，低秩适应），是一种将预训练模型权重冻结，并将可训练的秩分解矩阵注入Transformer架构每一层的技术，该技术可减少下游任务的可训练参数数量。

参数名称		参数说明
	微调后名称	自定义模型微调后的新名称。支持英文、数字、中划线（-）、下划线（_），长度1-64个字符，仅支持字母或下划线开头。 例如“baichuan-13b-chat-sft-nl2sql”。
数据配置	数据集	在下拉列表中选择**步骤一：创建微调数据集**创建的“智能分析数据集”。
	数据集版本	在下拉列表中选择数据集版本。
	训练数据比例	填写训练数据比例，如果填为0，则任务不执行训练阶段。 训练数据比例是指用于训练模型的数据在完整数据集中所占的比例。 在实际应用中，训练数据比例的选择取决于许多因素，例如可用数据量、模型复杂度和数据的特征等。通常情况下，会选择较大的训练数据比例，以便训练出更准确的模型。一般来说，训练数据比例在70%到90%之间是比较常见的选择。
	验证数据比例	填写验证数据比例，如果填为0，则任务不执行验证阶段。 验证数据比例是指模型训练过程中，用于验证模型当前训练效果的数据在完整数据集中所占的比例。 验证集的比例对于机器学习模型的性能评估非常重要。如果验证集的比例过小，可能导致模型在验证集上表现不够稳定，无法准确评估模型的性能。如果验证集的比例过大，可能会导致训练集的样本量不足，影响模型的训练效果。因此，在选择验证集的比例时，需要根据具体情况进行调整，以保证模型的性能评估和训练效果的准确性。
	测试数据比例	填写测试数据比例。如果填为0，则任务不执行测试阶段。 测试数据比例是指模型训练结束之后，用于测试模型训练效果的数据在完整数据集中所占的比例。 通常，测试数据比例在20%到30%之间较为常见，但具体比例取决于数据集的大小和质量，以及模型的复杂度和训练时间等因素。较小的测试数据比例可能导致过拟合，而过大的比例则可能导致欠拟合。因此，选择适当的测试数据比例对于训练出准确可靠的机器学习模型非常重要。
任务配置	资源池	选择执行任务的资源池，在下拉列表可以看到各资源池的可用卡数，根据实际情况选择。

步骤4 单击“下一步”，参照**表6-3**和**表6-4**以及模型、数据集等实际情况配置模型微调任务的基础参数、LoRA参数。

表 6-3 基础参数配置说明

参数英文名	参数中文名	参数说明
global_bs	各设备batch size综合	表示多个设备上使用的总样本数量。
num_train_epochs	训练epoch数	优化算法在完整训练数据集上的工作轮数。
learning_rate	学习率	学习率是每一次迭代中梯度向损失函数最优解移动的步长。
weight_decay	权重衰减因子	对模型参数进行正则化的一种因子，可以缓解模型过拟合现象。
warmup_ratio	学习率热启动比例	学习率热启动参数，一开始以较小的学习率去更新参数，然后再使用预设学习率，有效避免模型震荡。

表 6-4 LoRA 参数配置说明

参数英文名	参数中文名	参数说明
lora_rank	秩	LoRA微调中的秩。
lora_alpha	缩放系数	LoRA微调中的缩放系数。
target_modules	LoRA微调层	LoRA微调的layer名关键字。 baichuan系列：down_proj,gate_proj,up_proj,W_pack,o_proj chatglm系列：dense_4h_to_h,dense_h_to_4h,dense,query_key_value

步骤5 配置完成后，单击“创建”。新创建的微调任务显示在“微调任务”页面的任务列表中。

步骤6 在任务列表中“操作”列单击“启用”开始训练模型。

单击任务名称，进入任务详情界面查看运行日志和Loss曲线。

- 观察日志是否出现Error信息，如果有则表示训练失败，请根据日志提示定位原因并解决。
- 根据Loss曲线观察损失值的下降情况，据此调整训练的超参数来寻找一组最优参数。

步骤7 微调任务执行完成后，单击“操作”列的“发布”，发布模型。

----结束

步骤三：部署模型

部署模型是将模型部署为在线服务，通过创建部署服务实现，创建成功后，可以对在线服务可以进行预测和调用。本文需要为基座模型（原模型）和微调后的新模型分别创建模型服务。由于在线运行需消耗资源，请确保账户未欠费。

步骤1 在AI原生应用引擎的左侧导航栏选择“模型中心 > 我的模型服务”。

步骤2 在“我的模型服务”页面右上角单击“部署模型服务”。

步骤3 在“创建部署服务”页面，参照**表6-5**配置模型信息。

表 6-5 模型信息参数说明

参数名称	参数说明
模型来源	选择模型来源，本文以选择“微调的模型”为例。
选择模型	在下拉列表选择相应来源的具体模型。本文选择在**步骤二：创建模型微调流水线**中微调后的新模型。
服务名称	自定义服务名称，支持中英文、数字、中划线（-）、下划线（_）、点（.），长度2-36个字符，仅支持以中英文开头。此处以创建“baichuan-13b-chat-sft-nl2sql-s”为例。
模型服务描述	根据需要填写服务相关描述。
标签	为模型服务选择标签分类。可从以下几个维度选择（支持多选）： • 行业 • 适用领域 • 通用

步骤4 配置部署模型参数，参数说明如**表6-6**所示。

表 6-6 微调的模型部署参数说明

参数名称	参数说明
实例个数	设置模型服务部署的实例个数。 不同的模型部署1个实例需要的推理单元个数不同，比如，ChatGLM3-6B需要2个实例。 不同的模型因为模型参数量不同，模型参数量越多，需要消耗的资源越多，因此需要的推理单元个数越多。
推理单元资源	在下拉列表可以查看已购买的推理单元的可用个数，根据实际情况选择。 如果推理单元个数不足以满足实例个数，则需减少实例个数以使推理单元资源满足需求。 **说明** 在推理单元到期后，部署的模型将被下架，可通过购买推理单元资源恢复。

参数名称	参数说明
流控配置	超出流控值，则触发限流，用户的请求会因为流控而失败。 • 无限制 • 10次/秒 • 50次/秒 • 100次/秒 • 200次/秒

步骤5 单击“保存”，部署模型服务，新部署的服务显示在“我部署的”页签中以查看模型服务的部署情况。

步骤6 参照**步骤2**~**步骤5**，为基座模型创建名称为“baichuan-13b-chat-s”的模型服务。

----结束

步骤四：测试模型调优效果

通过在线测试模型服务，可检验模型的准确性、可靠性及反应效果，发现模型中存在的问题和局限性，确保模型能够在实际应用中正常运行，并且能够准确地预测和处理数据。调测模型前请确保部署模型服务已成功，即模型服务状态处于“运行中”。

步骤1 在AI原生应用引擎的左侧导航栏选择“模型中心 > 模型调测”，进入“模型调测”页面。

步骤2 在“模型类型”下选择“文本对话”并配置**表6-7**所示参数。

表 6-7 调测文本对话类型模型参数说明

参数名称	参数说明
模型服务	选择要调测的模型服务，在下拉列表选择**步骤三：部署模型**中部署的模型服务。
输出方式	可选非流式、流式。二者区别如下： • 非流式：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，大语言模型完整生成回答后一次性返回。 • 流式：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，逐个字词的快速返回模式，不需等待大语言模型生成完成。
输出最大token数	表示模型输出的最大token数。
温度	较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。建议该参数和“多样性”只设置1个。
多样性	影响输出文本的多样性，取值越大，生成文本的多样性越强。建议该参数和“温度”只设置1个。
存在惩罚	介于-2.0和2.0之间的数字。正值会尽量避免重复已经使用过的词语，更倾向于生成新词语。

参数名称	参数说明
频率惩罚	介于-2.0和2.0之间的数字。正值会尽量避免使用常见的单词和短语，更倾向于生成较少见的单词。
内容安全监测配置	当“输出方式”为“非流式”时，显示此参数。 选择是否打开开关，开启后，可对返回内容中的文本和图片进行安全监测。

步骤3 在右侧“效果预览”区域，可通过以下两种方式进行模型测试。

- 在对话输入框输入测试语句后按Enter键或单击进行模型测试。
- 单击“引用已有提示语模板”，弹出“选择模板”面板，可通过分类筛选我创建的、我收藏的或平台预置的提示语模板，然后按Enter键或单击进行模型测试。

输入相同的问题，可以看到微调前模型返回结果是错误的，而微调后模型返回了正确的计算指标数据的关键要素。

----结束

步骤五：清除相应资源

对于不再需要使用的微调任务和模型服务，建议及时清除相应资源，避免产生不必要的费用。

- 删除不需要的微调任务
  a. 在AI原生应用引擎的左侧导航栏选择“模型中心 > 模型微调流水线”。
  b. 在“模型微调流水线”页面的任务列表中，单击任务状态为“已完成”的微调任务所在行的“操作”列“更多 > 删除”，删除不需要的微调任务。
- 停用或删除不需要的模型服务
  a. 在AI原生应用引擎的左侧导航栏选择“模型中心 > 我的模型服务”。
  b. 在“我的模型服务 > 我部署的”页面的服务列表中，单击目标模型服务所在行的“操作”列的“停用”或“更多 > 删除”，停用或删除不需要的模型服务。

7 入门实践

当您购买AppStage后，可以根据自身的业务需求使用AppStage提供的一系列常用实践。

表 7-1 常用最佳实践

实践	描述
一站式应用开发、应用托管以及应用运维	介绍如何使用应用平台AppStage一站式功能，完成基于应用维度提供的开发、测试、版本发布、托管部署、运维监控的全场景全生命周期管理。
基于运维数仓的数据开发与应用	介绍如何通过AppStage运维中心完成对业务实时数据的接入、处理、开发与应用。
基于Spring Cloud框架进行应用上云	以Spring Cloud Demo项目为例，带您体验使用AppStage的开发中心、运维中心及运行时引擎进行工程创建、代码开发、打包发布、部署上线的全过程。