阿里云
内容安全

产品简介

文档版本：20200413

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。
斜体	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 什么是内容安全

内容安全是一款多媒体内容智能识别服务，支持对图片、视频、文本、语音等对象进行多样化场景检测，帮助您有效降低内容违规风险。

观看以下视频，快速了解云盾内容安全服务。

背景

随着互联网、智能设备及各种新生业务的飞速发展，互联网上的数据呈现爆炸式增长，图片、视频、发文、聊天等互动内容已经成为人们表达感情、记录事件和日常工作不可或缺的部分。每天，通过互联网上传的视频、图片数量超过10亿，通过各种社交网络、媒体平台的发文数量超过5亿，而且这种趋势还是继续快速增长。

这些日益增长的内容中也充斥着各种不可控的风险因素，例如色情视频和图片、涉政暴恐内容、各种垃圾广告等等。随着政府监管的日渐严格，这些都是各网站及平台亟待认真对待和管理的工作。而另一方面，人们对这些非结构化内容的认识和解读和也处于初级阶段，需要更加智能的技术和系统来帮助大家深度发掘这其中蕴藏的巨大商业价值。

产品定位

云盾内容安全是内容安全领域的先行者，源自阿里巴巴多年安全技术积累，依托阿里云、淘宝、支付宝等平台的管控经验，为企业用户提供成熟的、轻量化接入的内容安全解决方案，帮助企业、开发者在复杂多变的互联网环境下快速发现文本、图片、视频、语音的各类风险，保障应用的信息内容安全。

目前，内容安全包括三部分：内容检测API，OSS违规检测，站点检测。

- 内容检测API

  内容检测API主要能对包含色情、涉政、暴恐、广告、垃圾信息的文本、图片、视频、语音进行检测和识别，通过系统化的方式提供审核、打标、自定义配置等能力来保障您接入的效果和个性化需求落地，您需要一定的开发工作量，通过调用阿里云提供的接口来进行内容检测。

  针对的用户包括但不限于：视频网站、直播平台、社交平台、媒体平台、垂直社区／论坛、电商网站、存储平台、CDN平台等一切UGC（用户生成内容）平台和一切需要对网站内容进行安全管控的平台。

- OSS违规检测

  针对使用阿里云OSS存储文件的用户，提供一键式的图像鉴黄、涉政暴恐检测服务、您可以将保存在OSS中的图片、视频等内容进行鉴黄、涉政暴恐风险检测，并且提供删除和冻结文件的功能，该功能无需您开发，只需要少许页面配置即可接入使用。

- 站点检测

  针对拥有网站的用户，提供首页检测服务和全站网页内容检测服务，帮助您检查您的网站首页或全站内容是否具有违规内容风险。当您的站点中网页疑似有违规信息时，会通知您并提供违规网页地址及快照查看功能，方便您及时对网页内容进行整改。该功能无需您开发，只需要少许页面配置即可接入使用。

2 本地化部署方案

内容安全提供本地化部署版本。您可以将内容安全部署在本地，并对接本地数据中心，直接调用本地数据执行内容检测。本地化部署方案帮助您省却数据上传的工作，满足数据中心利旧和数据本地化需求。

内容安全本地化部署以软件方式提供，购买后由阿里云安全工程师到现场完成部署。通过本地化部署，您可以在自有数据中心（如自有IDC、物理机/私有云、混合云等）内获取阿里云云上环境同等量级的内容安全检测能力，也可以无缝获取公共云的弹性扩展能力，适应您的定制化要求和生态建设需要。

功能特性

内容安全本地化部署版本支持对本地文本、图片、视频内容进行特定场景的违规内容检测或特定内容识别，并提供管理控制台方便您进行相关配置和查看检测结果。

检测服务

以HTTP/HTTPS API接口形式提供指定场景的检测服务，便于您将检测环节集成到整体业务流程。支持的检测场景包括：

- 文本反垃圾：采用NLP自然语言理解算法识别色情、暴恐涉政、广告、辱骂等文本垃圾，并且能够结合行为策略有效管控灌水、刷屏等恶意行为。
- 图片/视频色情识别：对图片和视频进行色情内容识别以及色情程度量化。
- 图片/视频涉政暴恐识别：识别暴恐旗帜、人物和场景以及敏感政治人物等风险信息。
- 图片/视频敏感人脸识别：提供包括政治人物、敏感人物、以及名人明星等人物的面部识别，能够避免业务的违规和侵权风险。
- OCR图文识别：识别图片中的文字，精准定位图片中文字位置，准确识别斜排字、艺术字等字体。

管理控制台

管理控制台帮助您执行以下操作：

- 内容安全私有化管理，包括用户管理、系统配置（算法业务策略配置）、运维信息管理、调用报表查询等。
- 内容安全检测配置，包括自定义图库、词库管理，检测结果查看、反馈等。

购买及部署方案

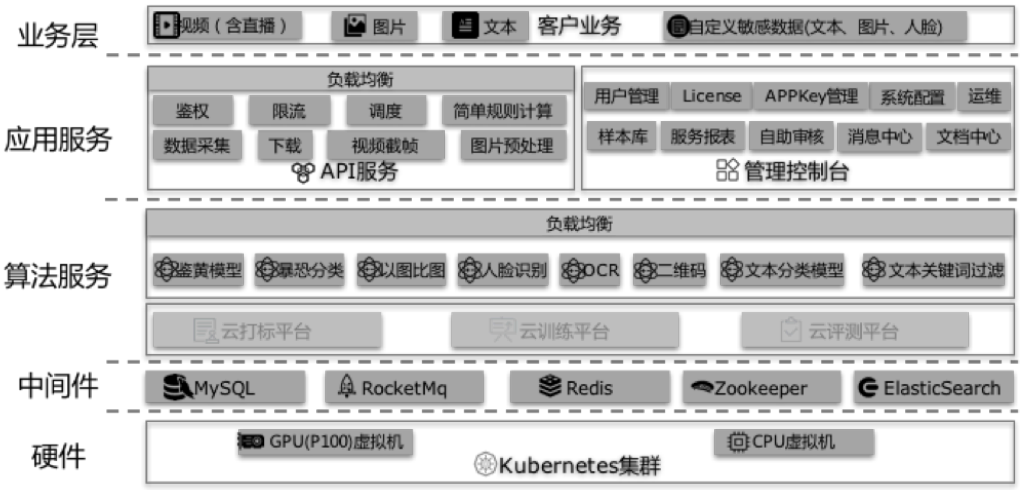
如何购买和收费

内容安全本地化部署版本目前开放线下购买途径，如有需求，您可以联系阿里云客户经理进行咨询和开通（您也可以提工单联系我们）。

本地化部署版本按照算法模型数量、服务量级及服务时间等收取相应的软件及服务授权费用，详情请咨询阿里云客户经理。

如何部署

下图阐释了内容安全的本地化部署架构。



在您购买内容安全本地化部署版本，并根据推荐的硬件配置要求准备好相应的硬件之后，阿里云安全工程师将到现场帮助您完成部署。

3 功能特性

内容安全提供内容检测API、OSS违规检测和站点检测功能，每个功能适用于多种场景。

内容检测API

应用场景	描述
图片违规内容检测	检测图片违规或识别图片中的不良信息。具体支持以下场景： • 智能鉴黄 • 暴恐涉政检测 • 图文违规检测 • 二维码检测 • 不良场景检测 • logo检测
视频违规内容检测	检测视频中的违规内容或不良信息。具体支持以下场景： • 智能鉴黄 • 涉政暴恐检测 • 图文违规检测 • 不良场景检测 • logo检测
文本垃圾内容检测	检测文本中的违规或不良内容，具体包括以下场景： • 广告内容检测 • 涉政暴恐检测 • 辱骂内容检测 • 色情内容检测 • 灌水内容检测 • 无意义内容检测 • 违禁品内容检测 • 自定义关键词检测

应用场景	描述
语音垃圾内容检测	检测语音中的违规或不良内容，具体包括以下场景： • 广告内容检测 • 涉政暴恐检测 • 辱骂内容检测 • 色情内容检测 • 灌水内容检测 • 无意义内容检测 • 违禁品内容检测 • 自定义关键词检测
图文OCR识别	识别图片中的各种文字信息（结构化或非结构化信息）。支持识别的结构化卡证对象包括： • 身份证 • 护照 • 银行卡 • 营业执照 • 增值税发票 • 行驶证 • 驾驶证 • 车牌 • 车辆Vin码 **说明：** 图文OCR支持您自定义模板进行识别。
人脸识别	人脸识别包括以下能力： • 人脸属性检测 • 人脸比对 • 人脸检索 • 翻拍检测 • 端上活体检测 **说明：** 端上活体检测提供离线安卓SDK，翻拍检测指服务端API。
相似图检索	根据给定的图片到用户自定义图库检索相似的topN张图片。
图片标签	识别图片中的主体，并输出对应的标签。
视频指纹	根据给定的视频，从视频库中检索同源视频。
视频标签	识别视频中的主体、场景、行为等内容，并输出标签以及出现的时间点。

OSS违规检测

应用场景	描述
图片或视频涉黄、涉政暴恐检测	检测OSS存储空间中的违规内容，支持增量内容自动检测或存量内容手动扫描。

站点检测

应用场景	描述
站点内容风险检测	定期自动检查网站首页或全站中存在的违规内容，支持检测的风险场景如下： • 首页篡改 • 色情低俗 • 涉政暴恐 • 垃圾广告 • 用户自定义关键词 • 用户自定义相似图

4 产品优势

内容安全支持对海量多媒体内容进行快速检测，接入便捷、成本低，且经历实战检验，有效帮助您节省90%以上的人工审核成本。

内容安全具备以下优势：

- **性价比高**

  在节省90%以上的人力成本的同时，支持秒级返回结果，达到最高99%以上的准确率。

- **经历实战检验**

  支撑阿里系淘宝、支付宝等核心业务，经历双11实战检验，拥有海量的特征样本及丰富的数据模型分析经验。

- **接入成本低**

  一次接入即可提供音视频、图片、文字等形式内容检测，覆盖暴恐、鉴黄、涉政、广告等风险防范。

- **服务方式灵活**

  既与OSS、ECS等云产品无缝对接，又可以通过API方式与用户审核系统集成。

- **对海量数据快速检测**

  基于云计算平台，对海量数据进行快速检测。

5 名词解释

本文解释了内容安全中用到的术语。

内容检测API

- AK信息：阿里云AccessId和AccessSecret。
- 自助审核：通过云盾控制台将您认为机器审核不准的结果进行人工矫正。

OSS违规检测

- 违规分值：代表属于某一识别结果的置信度。
- 冻结：冻结OSS中文件的方式。支持以下方式：
  - 将您的OSS文件访问权限设置为私有。如果文件本身就是私有模式访问，则不会对文件造成影响。
  - 将您的OSS文件移动到您bucket中的其他目录下。
- 自动冻结：系统自动帮您把超过阈值的内容进行冻结，无需手动操作。
- 存量扫描：对OSS bucket中的文件进行一次性扫描，需要手动触发。
- 增量扫描：对OSS bucket中变动过的文件进行增量扫描，开启后自动进行，无需手动操作。

站点检测

- 首页监测：帮助您定期检查您的网站首页是否具有内容篡改、色情低俗、涉政暴恐、垃圾广告等风险内容。
- 全站内容检测：定时帮助您检测网站全站的内容是否有风险隐患。当您站点中的网页疑似有违规信息时会通知您，并提供违规网页地址及快照查看功能，方便您对网页内容进行整改。