

数据接入服务

产品简介

文档版本

01

发布日期

2023-03-30



华为技术有限公司



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。

目录

1 图解数据接入服务 DIS.....	1
2 什么是数据接入服务 DIS.....	3
3 应用场景.....	4
4 DIS 有哪些功能.....	6
5 计费说明.....	8
6 权限管理.....	10
7 与其他服务的关系.....	13
8 使用限制.....	14
9 基本概念.....	15
10 区域和可用区.....	16
11 配额说明.....	18

1 图解数据接入服务 DIS

数据接入服务DIS

轻松连接一切
让数据为您所用

01. 用户实时数据上云及流动面临的问题

1 上云门槛高

- 数据源多、数据量大、数据类型复杂、数据格式不统一
- 数据源分散、数据质量不一、数据安全风险高
- 数据源异构、数据格式复杂、数据集成难度大
- 数据源分布广、数据集成成本高、数据集成周期长

2 数据流动难度大

- 数据源分散、数据集成难度大
- 数据源异构、数据格式复杂、数据集成难度大
- 数据源分布广、数据集成成本高、数据集成周期长

3 运维成本高

- 环境维护：采购硬件资源、部署软件系统
- 资源维护：数据复制管理、手动扩容或调整资源、并定期数据备份
- 系统监控：针对自行开发的运维监控系统、需要专人值守、需要高级人工紧急响应的处理

02. 华为云DIS提供一站式数据上云服务

Data Ingestion Service (简称DIS) 是华为云提供的实时数据接入服务，它提供了实时数据接入、高效数据传输、实时数据分发能力，让您可以轻松连接于实时数据源到云上数据湖。

核心优势

- 数据源：Agent、SDK、API、IoT、网络
- 传输通道：数据通道、数据通道、数据通道
- 分析：分析、计算、连接器
- 连接器：connector、connector、connector

高效

- 海量数据、海量数据、海量数据
- 海量数据、海量数据、海量数据

安全可靠

- 数据加密、数据加密、数据加密
- 数据加密、数据加密、数据加密

易用

- 简单易学、简单易学、简单易学
- 简单易学、简单易学、简单易学

低成本

- 按需付费、按需付费、按需付费
- 按需付费、按需付费、按需付费

03. DIS应用场景

1 日志分析

应用场景：实时监控、传输、预处理、分发、实现日志处理、分析、备份全场景

数据源：日志文件、日志流、第三方工具

实时采集：DIS Agent、DIS SDK、第三方工具

实时传输和预处理：DIS、DIS、DIS

实时存储/分析：Cloud Stream、MRS、OBS、DMS、Cloud Table、Elasticsearch

客户价值：

- 丰富的数据接入方式：提供Agent、SDK、对接第三方采集工具
- 高效数据传输：高并发传输
- 数据质量提升：数据源处理支持数据格式转换、提升入口数据质量
- 多渠道数据分发：支持日志处理、分析、备份多种应用场景

2 实时监控/分析

应用场景：故障、异常、设备状态实时监控

数据源：数据库、日志流、IoT Hub

实时采集：DIS Agent、DIS SDK、第三方工具

实时传输和预处理：DIS、DIS、DIS

实时存储/分析：Cloud Stream、CloudTable、FunctionStage、Elasticsearch

客户价值：

- 支持IoT数据接入
- 高效数据传输：高并发传输
- 数据质量提升：提升入口数据质量
- 支持实时监控、报警、定位等多种业务场景

3 实时推荐/预测

应用场景：实时接入数据，对部分业务支持实时推荐/预测场景

数据源：数据库、日志流、第三方工具

实时采集：DIS Agent、DIS SDK、第三方工具

实时传输和预处理：DIS、DIS、DIS

实时存储/分析：Cloud Stream、MRS、OBS、DMS、Cloud Table、Elasticsearch

客户价值：

- 轻松构建实时推荐/预测应用

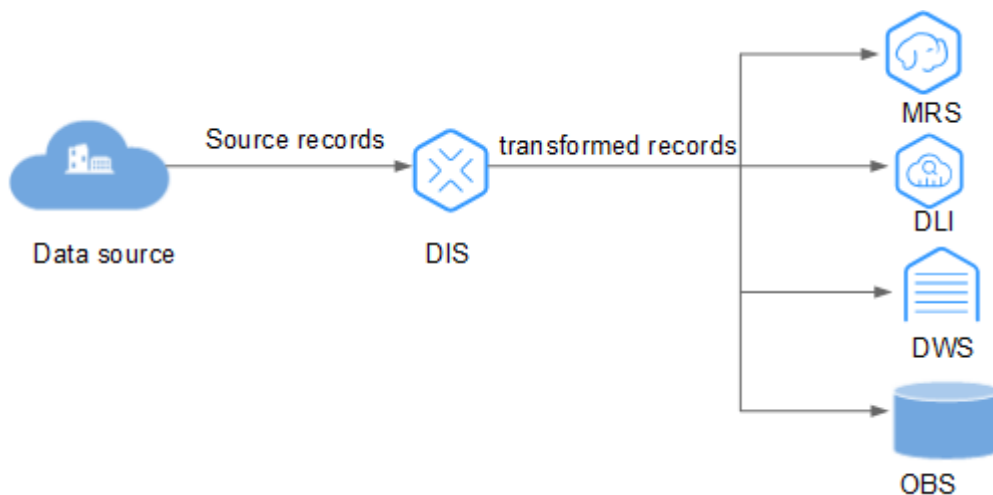
2 什么是数据接入服务 DIS

数据接入服务（Data Ingestion Service）为处理或分析流数据的自定义应用程序构建数据流管道，主要解决云服务外的数据实时传输到云服务内的问题。数据接入服务每小时可从数十万种数据源（如IoT数据采集、日志和定位追踪事件、网站点击流、社交媒体源等）中连续捕获、传送和存储数TB数据。

数据流向

- DIS实时从多种数据源采集数据。
- DIS连续传输数据，自动将数据传输至MRS，DLI，DWS和OBS等服务做计算，分析和存储。

图 2-1 数据流向

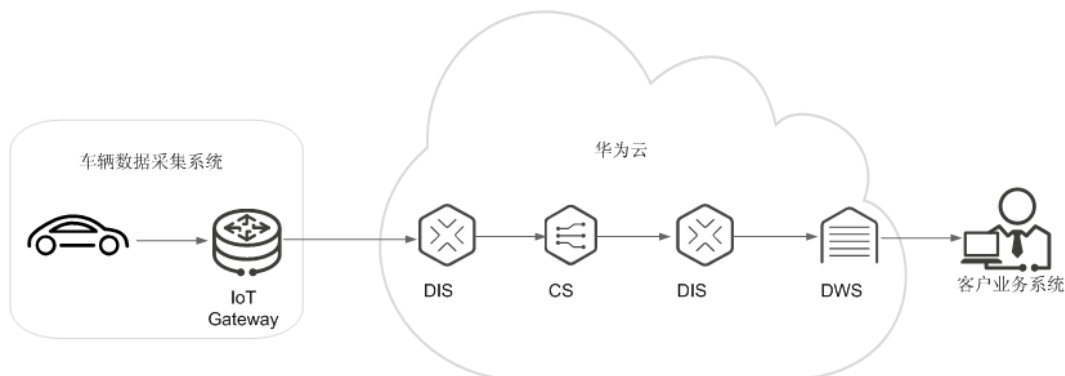


3 应用场景

实时数据处理

实时收集城市各交通枢纽的车辆通行数据，缓存在通道中，分析平台周期读取通道中的数据分析后将结果应用到调度系统，实现对停车场开放时长和交通资源的调配。

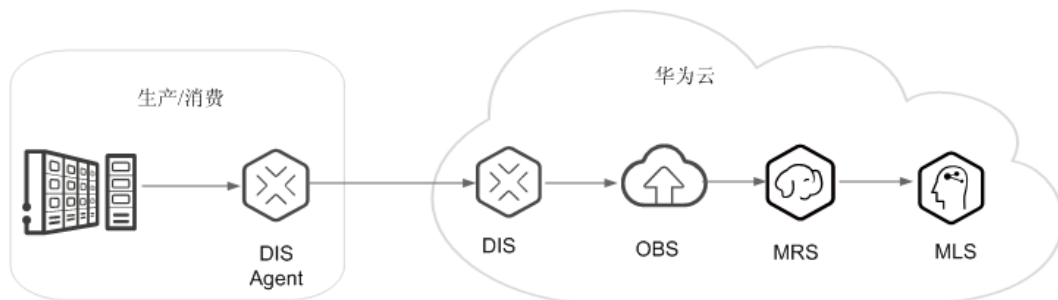
图 3-1 场景示例图



实时文件传输

实时检测客户应用系统中产生的文件，并采集上传到云上，进行离线分析、存储查询及机器学习，对客户进行分类和信息查询，识别出大型客户，加强服务，进一步提升客户满意度。

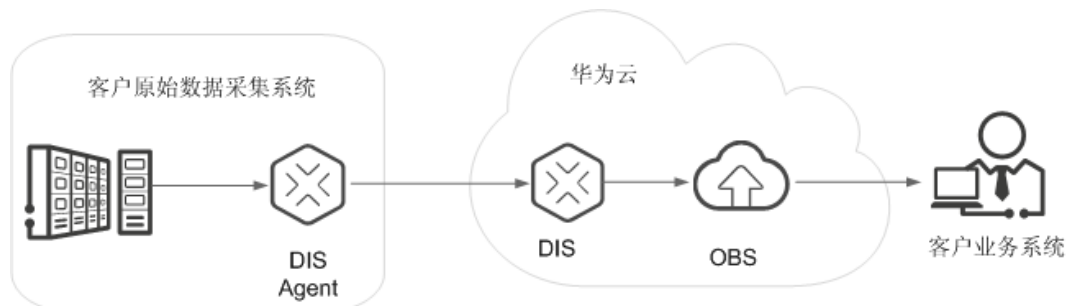
图 3-2 场景示例图



数据接入备份

将大量滚动日志文件传输到云端做备份，用于数据丢失或异常后的恢复和故障分析。同时大量小文本文件可合并转储为大文件，提高数据处理性能。

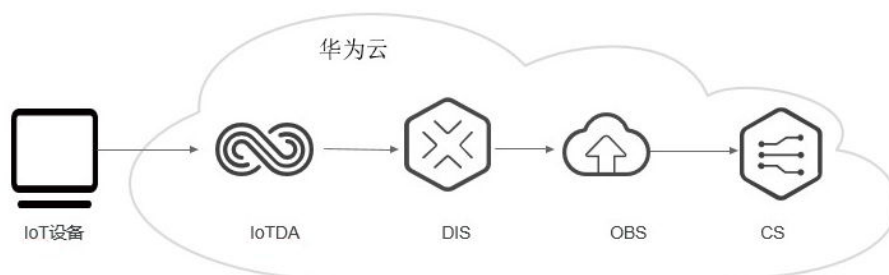
图 3-3 场景示例图



IOT 分析应用

通过在车载设备中集成IoTDA，并配置数据转发到DIS，实时上传车辆状态、监控数据，采用云上数据分析系统进行实时分析、检测，动态掌握设备状况，实现汽车故障提前感知。

图 3-4 场景示例图



4 DIS 有哪些功能

DIS对数据传输所需要的基础设置、存储、网络和配置进行管理。您无需为数据通道担心配置、部署、持续的硬件维护等。此外，DIS还可在云区域同步复制数据，为您提供数据高可用性和数据持久性。

主要模块

DIS分为4个功能模块，各模块功能如下：

- 服务控制面
 - 完成服务的开通、删除、配置操作，并将用户信息同步到数据面。
 - 完成数据面资源的申请与自动部署。
- 服务数据面
 - 接收用户发送数据的请求，对已鉴权的数据接收并存储。
 - 接收用户获取数据的请求，在鉴权后输出对应的用户数据。
 - 按时老化存储在系统中的用户数据。
 - 根据用户配置，将用户数据存储到对象存储服务（Object Storage Service，简称OBS）、MapReduce服务（MapReduce Service,简称MRS）、数据仓库服务（Data Warehouse Service，简称DWS）、数据湖探索（Data Lake Insight，简称DLI）。
- 服务维护
 - 负责服务的安装、升级。
 - 负责服务的配置、巡检、日志收集与分析、运行监控。
 - 负责服务工单处理。
- 用户SDK
 - 提供Java接口，供用户上传与下载数据。
 - 提供数据加密功能。

关键能力

- 无限扩展：DIS数据通道的吞吐量每小时可从数MB扩展到数TB，PUT记录每秒钟可从数千次扩展到数百万。
- 易于使用：您可以在几秒钟内创建DIS数据通道，轻松的将数据放入通道中，并构建用于数据处理的应用程序。

- 成本低廉：DIS没有前期成本，您只需要为实际使用的资源付费即可。
- 并行处理：DIS可让您用多个应用程序同时处理同一个数据通道。
- 安全可靠：DIS可将数据保留24~72小时，以防数据在应用程序故障、个别机器故障或设施故障时丢失。

5 计费说明

计费项

DIS根据您的消息数量和使用分区时长计费。

表 5-1 计费项信息

计费项	计费说明
消息数量	用户上传至DIS通道的消息数量，下载不计入消息数量。单条消息计量不超过25KB，即用户每次上传的单条数据超过25KB是按多条消息计算，不足25KB时，按单条消息计算。 说明 只有DIS普通通道涉及此计费因子，高级通道不按消息数量收费，只根据使用分区时长计费。
分区时间	使用分区的时长，时长从创建DIS分区开始计入。普通通道和高级通道的分区使用时长计费标准不同。 <ul style="list-style-type: none">普通通道分区使用时长计费详情请参见DIS分区计费。高级通道分区使用时长计费详情请参见DIS分区计费。

计费方式

DIS提供按需计费和包年包月计费方式供您灵活选择，使用越久越便宜。

- 按需计费

DIS默认为按需计费方式，即按实际使用的时长收费，以小时为单位，每小时整点结算，不设最低消费标准。

按需计费方式下各计费项的具体价格，请参见[产品价格详情](#)。您可以通过DIS提供的[价格计算器](#)，根据您的业务诉求，来快速计算出购买DIS的参考价格。

- 包年包月

您也可以通过包年包月（资源包）的方式，提前购买使用额度和时长，获取更多的优惠。超出资源包的部分自动转按需计费。

各资源包在不同规格及时长下的具体价格，请参见[产品价格详情](#)。您可以通过DIS提供的[价格计算器](#)，根据您的业务诉求，来快速计算出购买DIS的参考价格。

计费示例

假设用户A的上传流量为5MB/s，单条记录大小为2KB，创建5个普通分区，采用按需计费。这种情况下每小时收费如下；

1. 分区时间费用

分区时间0.1元/小时，5个分区需要： $0.1 * 5 = 0.5$ 元。

2. 消息数量费用

上传消息数量0.1元/百万条，每小时上传消息数量为： $5\text{MB/s} * 1024 / 2\text{KB} * 3600\text{秒} = 9.216$ 百万条。因此工具需要： $0.1 * 9.216 = 0.9216$ 元。

因此每小时收费总共需要： $0.5 + 0.9216 = 1.4216$ 元/小时。

6 权限管理

默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

如果云账号已经能满足您的要求，不需要创建独立的IAM用户进行权限管理，您可以跳过本章节，不影响您使用数据接入服务的其它功能。

IAM是华为云提供权限管理的基础服务，无需付费即可使用，您只需要为您账号中的资源进行付费。关于IAM的详细介绍，请参见《[IAM产品介绍](#)》。

DIS 权限

默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

DIS部署时通过物理区域划分，为项目级服务，需要在各区域（如华北-北京1）对应的项目（cn-north-1）中设置相关权限，并且该权限仅对此项目生效，如果需要所有区域都生效，则需要所有项目都设置权限。访问DIS时，需要先切换至授权区域。

角色： IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于华为云各服务之间存在业务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。

如表1所示，包括了DIS的所有系统权限。其中“依赖关系”表示该系统权限对其它权限的依赖。由于华为云各服务之间存在业务交互关系，数据接入服务的权限依赖其他服务的权限实现功能。因此给用户授予数据接入服务的权限时，需要同时授予依赖的权限，数据接入服务的权限才能生效。

表 6-1 DIS 系统权限

系统角色	策略类别	描述	依赖关系
DIS Administrator	系统角色	对数据接入服务的所有执行权限。	无

系统角色	策略类别	描述	依赖关系
DIS Operator	系统角色	通道管理权限，拥有创建删除等管理通道的权限，但不能使用通道上传下载数据。	无
DIS User	系统角色	通道使用权限，拥有使用通道上传下载数据的权限，但不能管理通道。	无

表2列出了数据接入服务（Data Ingestion Service）常用操作与系统权限的授权关系，您可以参照该表选择合适的系统权限。

表 6-2 常用操作与系统权限的关系

操作	DIS Administrator	DIS Operator	DIS User
创建通道	√	√	x
删除通道	√	√	x
查询通道列表	√	√	√
查询通道详情	√	√	√
查询通道监控信息	√	√	√
查询分区监控信息	√	√	√
获取通道消费信息	√	√	√
变更分区数量	√	√	x
上传数据	√	x	√
获取数据游标	√	x	√
下载数据	√	x	√
创建APP	√	√	√
查询APP详情	√	√	√
查询APP列表	√	√	√
删除APP	√	√	√

操作	DIS Administrator	DIS Operator	DIS User
新增 Checkpoint	√	x	√
查询 Checkpoint	√	√	√
删除 Checkpoint	√	x	√
添加转储任务	√	√	√
查询转储任务详情	√	√	√
查询转储任务列表	√	√	√
删除转储任务	√	√	√

7 与其他服务的关系

介绍DIS与其他服务的关系。

- 对象存储服务（Object Storage Service，简称OBS）
在用户购买接入通道时选择数据转储至OBS，用户上传的数据会存储至对象存储服务（OBS）。
- 统一身份认证服务（Identity and Access Management，简称IAM）
统一身份认证服务为DIS提供了用户鉴权功能。
- 云审计服务（Cloud Trace Service，简称CTS）
云审计服务记录DIS相关的操作事件，方便用户日后的查询、审计和回溯。CTS的更多信息，请参见《云审计服务用户指南》。

8 使用限制

使用DIS前，您需要认真阅读并了解以下使用限制。

- 建议使用支持的浏览器登录DIS。
 - Google Chrome: 48.0及更高版本
 - Mozilla FireFox: 43.0及更高版本
 - Internet Explorer: 9.0及更高版本

当使用Internet Explorer 9.0时可能无法登录DIS管理控制台，原因是某些Windows系统例如Win7旗舰版，默认禁止Administrator用户，Internet Explorer在安装时自动选择其他用户如System用户安装，从而导致Internet Explorer无法打开登录页面。请使用管理员身份重新安装Internet Explorer 9.0或更高版本（建议），或尝试使用管理员身份运行Internet Explorer 9.0。
- 用户使用接入通道上传数据的时候，如果数据涉及敏感信息，请使用DIS SDK提供的加密配置进行加密或者自行加密。
- 用户使用接入通道上传数据时，普通通道单分区单次请求的记录总大小不能超过1MB（不包含partitionKey数据大小），高级通道单分区单次请求的记录总大小不能超过5MB（不包含partitionKey数据大小）。

9 基本概念

- 通道名称：租户创建的逻辑单位，用以区分不同租户实时数据的集合，创建通道时指定。在用户发送或者接收实时数据时，需要指定通道名称。
- 分区数：分区（Partition）是DIS数据通道的基本吞吐量单位。创建数据通道时，您将指定所需的分区数量。
- 记录：记录（Record）是存储在DIS通道中的数据单元。记录由序列号、分区键和数据块组成。数据块是您的数据生产者添加到数据通道的重要数据。数据块的最大尺寸（Base64编码前的数据有效载荷）是1MB。
- 序列号：序列号是每个记录的唯一标识符。序列号由DIS在数据生产者调用PutRecord操作以添加数据到DIS数据通道时DIS服务自动分配的。同一分区键的序列号通常会随时间变化增加。PutRecords请求之间的时间段越长，序列号越大。
- 应用程序：一个DIS应用程序是读取和处理来自DIS数据通道的使用者。您可以使用客户端库（SDK）构建DIS应用程序。
- 客户端库：SDK是一个适用于Java的客户端库，帮助用户轻松构建DIS应用程序，用以读取和处理来自DIS数据通道的数据。
- Project：用于将OpenStack的资源（计算资源、存储资源和网络资源）进行分组和隔离。Project可以是一个部门或者一个项目组。一个账户中可以创建多个Project。账户是租户对应的账号。一个Region有n个Project，但是一个Project只关联一个Region。不同项目中的DIS通道不可以相互连通。
- Checkpoint：消费检查点。应用程序消费数据时，记录已消费数据的最新序列号作为检查点。当重新消费数据时，可根据此检查点继续消费。
- APP：应用程序标识符。当多个应用程序分别消费同一通道的数据时，为区分不同应用程序的消费检查点，使用APP作为标识。

10 区域和可用区

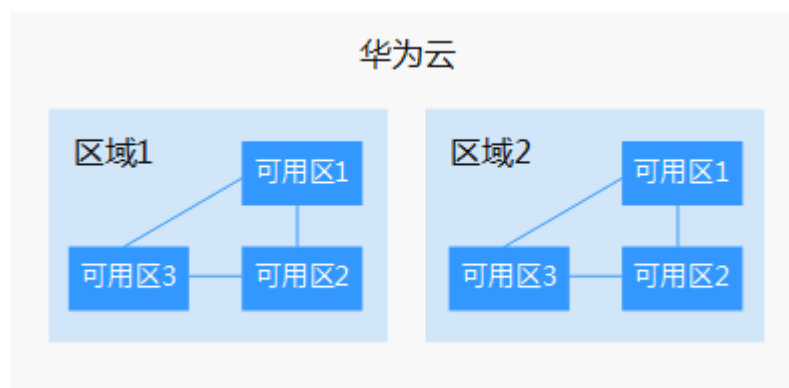
什么是区域、可用区？

我们用区域和可用区来描述数据中心的位置，您可以在特定的区域、可用区创建资源。

- 区域（Region）：从地理位置和网络时延维度划分，同一个Region内共享弹性计算、块存储、对象存储、VPC网络、弹性公网IP、镜像等公共服务。Region分为通用Region和专属Region，通用Region指面向公共租户提供通用云服务的Region；专属Region指只承载同一类业务或只面向特定租户提供业务服务的专用Region。
- 可用区（AZ，Availability Zone）：一个AZ是一个或多个物理数据中心的集合，有独立的风火水电，AZ内逻辑上再将计算、网络、存储等资源划分成多个集群。一个Region中的多个AZ间通过高速光纤相连，以满足用户跨AZ构建高可用性系统的需求。

图10-1阐明了区域和可用区之间的关系。

图 10-1 区域和可用区



目前，华为云已在全球多个地域开放云服务，您可以根据需求选择适合自己的区域和可用区。更多信息请参见[华为云全球站点](#)。

如何选择区域？

选择区域时，您需要考虑以下几个因素：

- 地理位置

一般情况下，建议就近选择靠近您或者您的目标用户的区域，这样可以减少网络时延，提高访问速度。不过，在基础设施、BGP网络品质、资源的操作与配置等方面，中国大陆各个区域间区别不大，如果您或者您的目标用户在中国大陆，可以不用考虑不同区域造成的网络时延问题。

曼谷等其他地区和国家提供国际带宽，主要面向非中国大陆地区的用户。如果您或者您的目标用户在中国大陆，使用这些区域会有较长的访问时延，不建议使用。

- 在除中国大陆以外的亚太地区有业务的用户，可以选择“亚太-曼谷”或“亚太-新加坡”区域。
- 在非洲地区有业务的用户，可以选择“南非-约翰内斯堡”区域。
- 在欧洲地区有业务的用户，可以选择“欧洲-巴黎”区域。

- 资源的价格

不同区域的资源价格可能有差异，请参见[华为云服务价格详情](#)。

如何选择可用区？

是否将资源放在同一可用区内，主要取决于您对容灾能力和网络时延的要求。

- 如果您的应用需要较高的容灾能力，建议您将资源部署在同一区域的不同可用区内。
- 如果您的应用要求实例之间的网络延时较低，则建议您将资源创建在同一可用区内。

区域和终端节点

当您通过API使用资源时，您必须指定其区域终端节点。有关华为云的区域和终端节点的更多信息，请参见[地区和终端节点](#)。

11 配额说明

- 创建通道应用的基础设施如下：
 - 普通通道分区
 - 高级通道分区
- 创建IoT实例应用的基础设施如下：
 - 弹性云服务器
 - 虚拟私有云VPC
 - 弹性负载均衡ELB
 - 弹性公网IP

其配额查看及修改请参见[关于配额](#)。