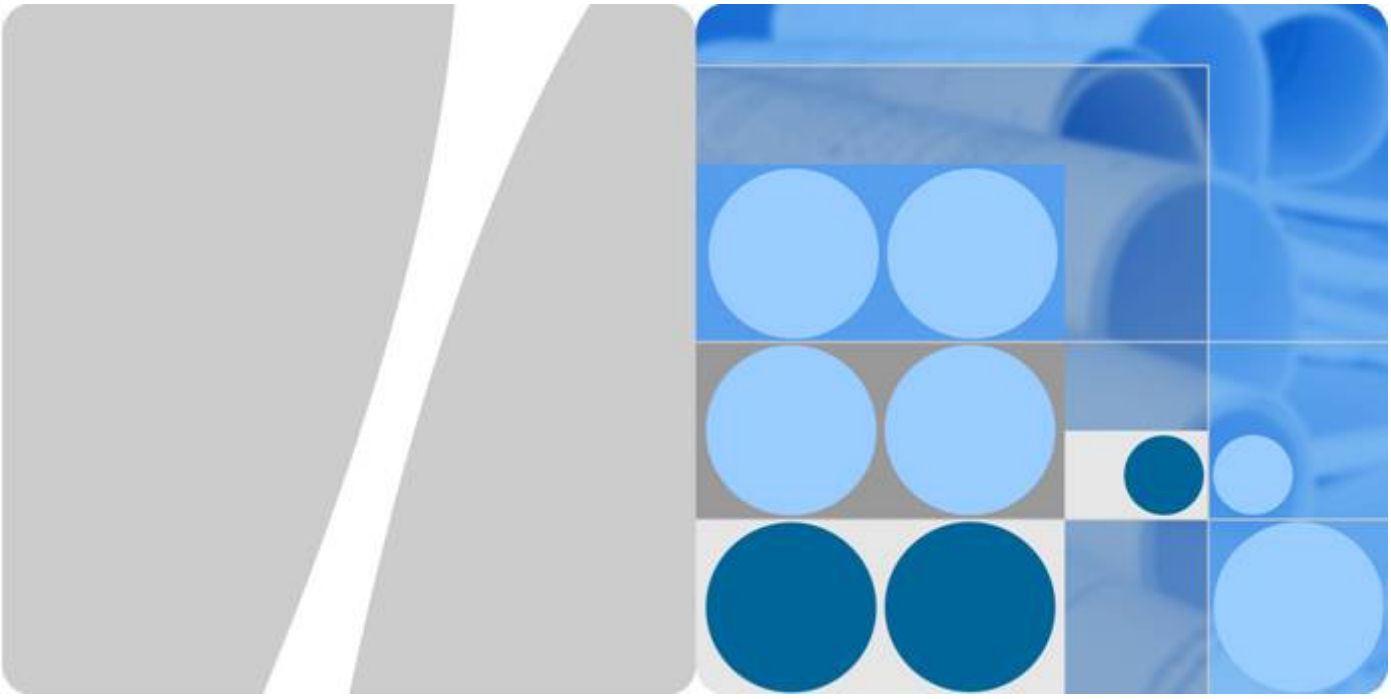


数据湖工厂

产品简介

文档版本　01

发布日期　2020-12-08

华为技术有限公司

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。

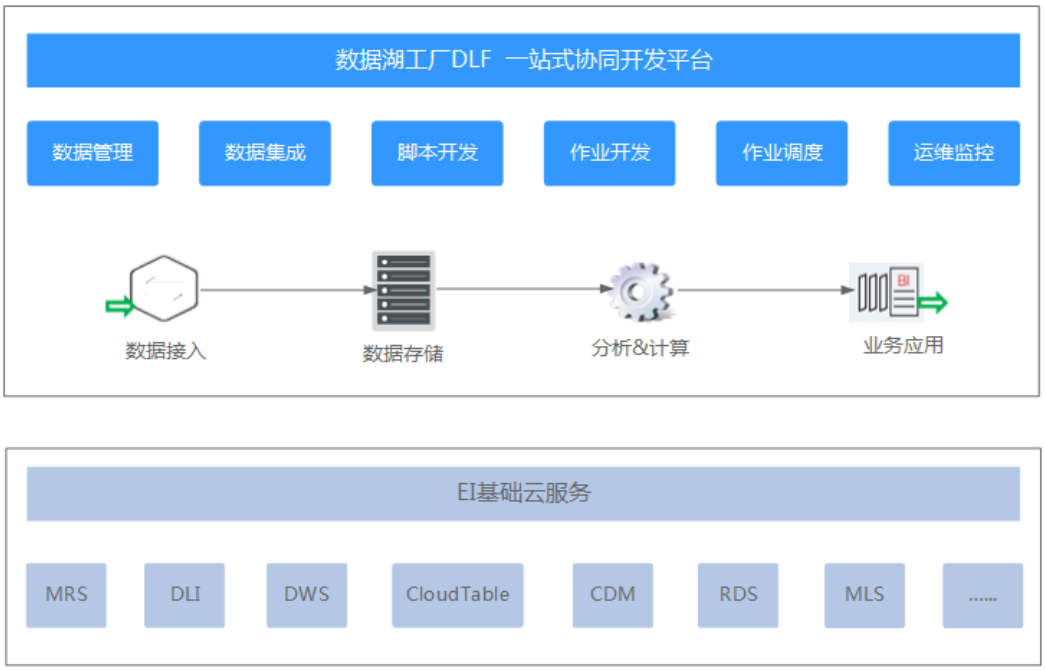
目 录

1 产品概述

数据湖工厂服务（Data Lake Factory，简称数据开发模块）是华为云大数据重要的平台产品，它可管理多种大数据服务，提供一站式的大数据开发环境、全托管的大数据调度能力，极大降低用户使用大数据的门槛，帮助用户快速构建大数据处理中心。

使用数据开发模块，用户可进行数据管理、数据集成、脚本开发、作业开发、作业调度、运维监控等操作，轻松完成整个数据的处理分析流程。

图 1-1 DLF 流程



数据管理、数据集成、脚本开发、作业开发、作业调度、运维监控的详细说明请参见**产品功能**。

2 产品优势

一站式云上数仓建设

支持一站式建设云上数仓，完成数据集成、脚本开发、作业开发、作业调度、运维监控、数据管理等操作，无须切换多个工具。

数据湖开发

支持管理DWS、DLI等多种大数据服务，在同一作业中可实现数据在不同类型数据服务中的编排与调度，实现真正的数据湖开发。

丰富的数据开发类型

支持多人在线协作开发，脚本开发可支持SQL、Shell在线编辑、实时查询；作业开发可支持CDM、SQL、MR、Shell、MLS、Spark等多种数据处理节点。

强大的作业调度能力

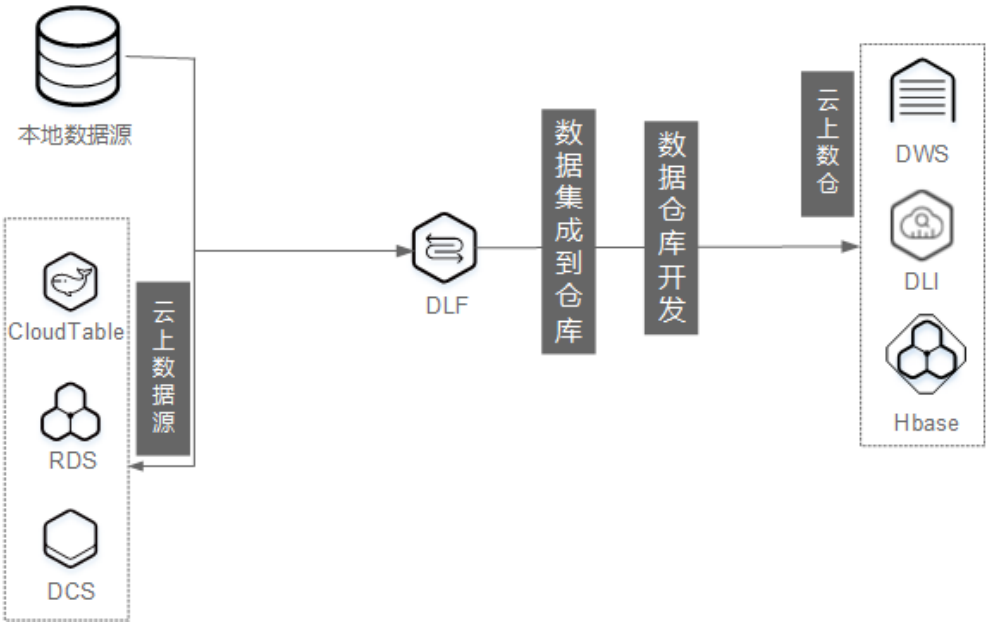
提供丰富的调度配置策略与海量的作业调度能力；支持时间周期调度、事件驱动调度、手工调度等多种调度方式。

3 应用场景

云上数仓快速搭建

通过数据开发模块将线下数据迁移到华为云上，将数据集成到华为云大数据服务中，并在数据开发模块中进行数据开发。

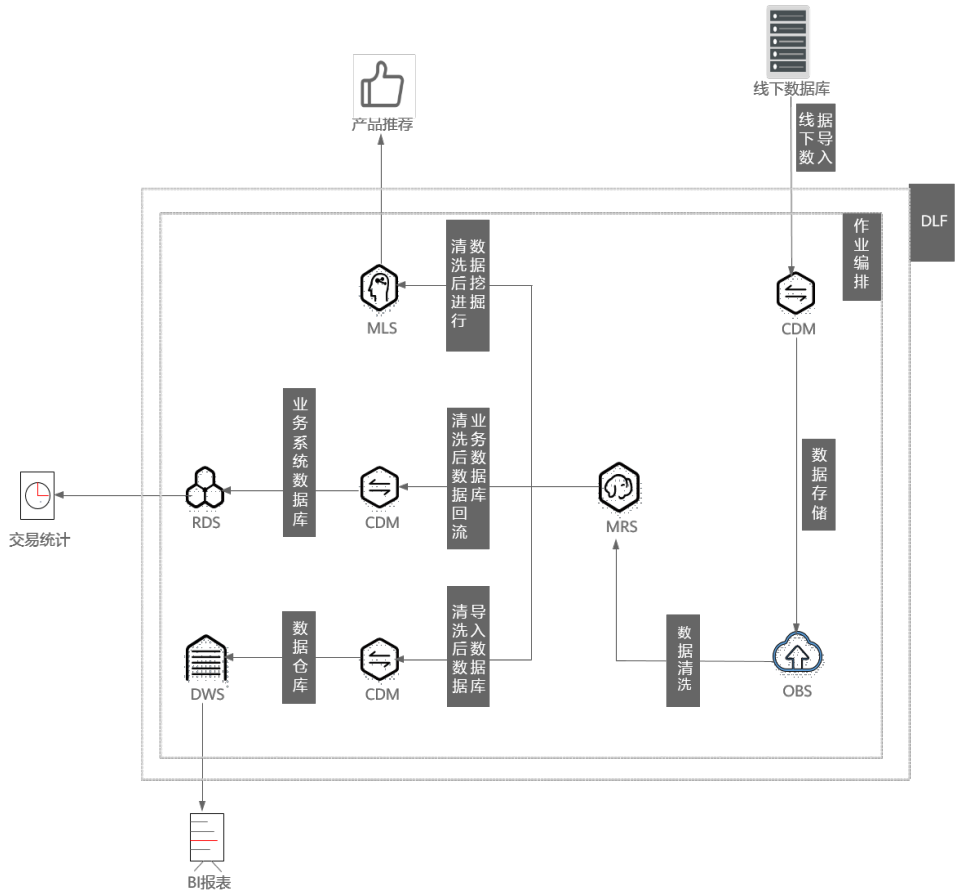
图 3-1 场景示例图



数据分析业务流自动化

通过数据开发模块实现数据导入、清洗、机器学习、数据回传、报表生成端到端流程自动化，把业务搬上自动化流水线。

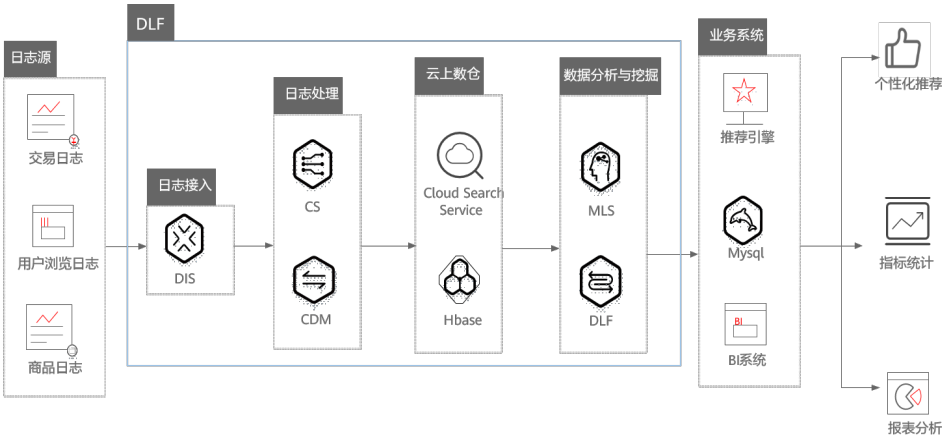
图 3-2 场景示例图



海量日志轻松分析挖掘

通过DIS将日志数据接入到OBS存储（或者Cloud Search服务），然后通过数据开发模块服务编写数据开发脚本和数据挖掘脚本，实现海量日志分析和挖掘。

图 3-3 场景示例图



4 产品功能

数据管理

- 支持管理DWS、DLI、MRS Hive等多种数据仓库。
- 支持可视化和DDL方式管理数据库表。

数据集成

与云数据迁移服务（CDM）无缝集成，依托CDM的强力支撑，支持20多种异构数据源之间可靠高效的数据传输，轻松实现多数据源集成到数据仓库。

脚本开发

- 提供在线脚本编辑器，支持多人协作进行SQL、Shell脚本在线代码开发和调测。
- 支持使用变量和函数。

作业开发

- 提供图形化设计器，支持拖拉拽方式快速构建数据处理工作流。
- 预设数据集成、SQL、MR、Spark、Shell、机器学习等多种任务类型，通过任务间依赖完成复杂数据分析处理。
- 支持导入和导出作业。

资源管理

支持统一管理在脚本开发和作业开发使用到的file、jar、archive类型的资源。

作业调度

支持单次调度、周期调度和事件驱动调度，周期调度支持分钟、小时、天、周、月多种调度周期。

运维监控

- 支持对作业进行运行、暂停、恢复、终止等多种操作。
- 支持查看作业和其内各任务节点的运行详情。

- 支持配置多种方式报警，作业和任务发生错误时可及时通知相关人，保证业务正常运行。

5 权限管理

如果您需要对华为云上购买的数据湖工厂服务 （ Data Lake Factory ）资源，给企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（ Identity and Access Management，简称IAM ）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制华为云资源的访问。

通过IAM，您可以在华为云账号中给员工创建IAM用户，并使用策略来控制他们对华为云资源的访问范围。例如您的员工中有负责软件开发的人员，您希望他们拥有数据湖工厂服务 （ Data Lake Factory ）的使用权限，但是不希望他们拥有删除数据湖工厂服务等高危操作的权限，那么您可以使用IAM为开发人员创建用户，通过授予仅能使用数据湖工厂服务，但是不允许删除数据湖工厂服务的权限策略，控制他们对数据湖工厂服务资源的使用范围。

如果华为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，不影响您使用数据湖工厂服务的其它功能。

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 》**。

DLF 权限

默认情况下，新建的IAM用户没有任何权限，您需要将其加入用户组，并给用户组授予策略或角色，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。

DLF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，则需要在所有项目都设置权限。访问DLF时，需要先切换至授权区域。

如**表5-1**所示，包括了DLF的所有系统策略权限。

权限根据授权精细程度分为角色和策略。

- 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。
- 策略：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业

对权限最小化的安全管控要求。如不允许某用户组删除作业，仅允许操作作业基本操作，如创建作业，查询作业列表等。DLF支持的API授权项请参见**权限策略和授权项**。

表 5-1 DLF 系统权限

系统角色/策略名称	描述	策略类别
DLF FullAccess	数据湖工厂服务所有权限	系统策略
DLF Development	数据湖工厂服务的开发者权限，拥有该权限的用户能使用DLF进行脚本开发与作业编排，但是不具备对工作区的增删改权限。	系统策略
DLFOperationAndMaintenanceAccess	数据湖工厂服务的运维人员权限，拥有该权限的用户可以对DLF的脚本与作业等资源进行运维，但不具备对各种资源的增删改权限。	系统策略
DLF ReadOnlyAccess	数据湖工厂服务的只读权限，拥有该权限的用户仅能查看DLF的资源。	系统策略
DLF Administrator	数据湖工厂服务管理员	系统角色

表5-2列出了DLF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 5-2 常用操作与系统权限的授权关系

操作	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
查询工作区	√	√	√	√	√
创建工作区	√	x	x	x	√
更新工作区	√	x	x	x	√
删除工作区	√	x	x	x	√
查询环境变量	√	√	√	√	√
更新环境变量	√	√	x	x	√

操作	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
导入环境变量	√	√	x	x	√
导出环境变量	√	√	x	x	√
查询表	√	√	√	√	√
创建表	√	√	x	x	√
更新表	√	√	x	x	√
删除表	√	√	x	x	√
查询数据库	√	√	√	√	√
创建数据库	√	√	x	x	√
更新数据库	√	√	x	x	√
删除数据库	√	√	x	x	√
查询模式	√	√	√	√	√
创建模式	√	√	x	x	√
更新模式	√	√	x	x	√
删除模式	√	√	x	x	√
查询目录	√	√	√	√	√
创建目录	√	√	x	x	√
更新目录	√	√	x	x	√
删除目录	√	√	x	x	√
查询解决方案	√	√	√	√	√
创建解决方案	√	√	x	x	√
更新解决方案	√	√	x	x	√

操作	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
删除解决方案	√	√	x	x	√
导入解决方案	√	√	√	x	√
导出解决方案	√	√	√	x	√
启动解决方案	√	√	√	x	√
停止解决方案	√	√	√	x	√
查询脚本列表	√	√	√	√	√
创建脚本	√	√	x	x	√
更新脚本	√	√	x	x	√
删除脚本	√	√	x	x	√
脚本语法检查	√	√	√	x	√
执行脚本	√	√	√	x	√
取消执行脚本	√	√	√	x	√
导入脚本	√	√	√	x	√
导出脚本或脚本执行结果	√	√	√	x	√
查询作业信息	√	√	√	√	√
创建作业	√	√	x	x	√
更新作业	√	√	x	x	√
删除作业	√	√	x	x	√
重命名作业	√	√	x	x	√
导入作业	√	√	√	x	√

操作	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
导出作业	√	√	√	x	√
校验作业定义的合法性	√	√	√	x	√
测试运行作业	√	√	√	x	√
启动作业	√	√	√	x	√
停止作业	√	√	√	x	√
暂停作业	√	√	√	x	√
恢复执行作业	√	√	√	x	√
查询作业实例	√	√	√	√	√
重跑作业实例	√	√	√	x	√
停止作业实例	√	√	√	x	√
强制成功作业实例	√	√	√	x	√
继续执行作业实例	√	√	√	x	√
启用作业节点	√	√	√	x	√
禁用作业节点	√	√	√	x	√
重跑作业节点	√	√	√	x	√
跳过执行作业节点	√	√	√	x	√
暂停作业节点	√	√	√	x	√
恢复执行作业节点	√	√	√	x	√

操作	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
强制成功作业节点	√	√	√	x	√
查询数据连接	√	√	√	√	√
创建数据连接	√	√	x	x	√
更新数据连接	√	√	x	x	√
删除数据连接	√	√	x	x	√
测试数据连接的连通性	√	√	x	x	√
导入数据连接	√	√	√	x	√
导出数据连接	√	√	√	x	√
查询资源列表	√	√	√	√	√
创建资源	√	√	x	x	√
更新资源	√	√	x	x	√
删除资源	√	√	x	x	√
上传资源	√	√	x	x	√
导出资源	√	√	√	x	√
导入资源	√	√	√	x	√
查询备份信息	√	√	√	√	√
启动备份	√	√	√	x	√
停止备份	√	√	√	x	√
查询通知列表	√	√	√	√	√
创建通知	√	√	x	x	√

操作	DLF FullAccesss	DLF Development	DLF OperationAndMaintenanceAccess	DLF ReadonlyAccess	DLF Administrator
更新通知	√	√	x	x	√
删除通知	√	√	x	x	√
查询补数据列表	√	√	√	√	√
创建补数据任务	√	√	x	x	√
停止补数据任务	√	√	√	x	√

6 与其他服务的关系

MapReduce 服务

数据开发模块服务的大数据类型节点（如SparkSQL）运行在MapReduce服务（MapReduce Service，MRS）中。

对象存储服务

数据开发模块服务支持从对象存储服务（Object Storage Service，简称OBS）导入数据，同时数据开发模块还利用OBS存储数据、结果、日志文件，以及用户程序。

关系型数据库

数据开发模块服务支持将数据存储在关系型数据库（Relational Database Service，RDS）中，以及执行RDS数据处理操作。

数据加密服务

数据加密服务（Data Encryption Workshop，DEW）用于加密和解密数据开发模块中数据连接的用户密码和密钥。

数据仓库服务

数据开发模块服务支持将数据存储在数据仓库服务（Data Warehouse Service，DWS）中，以及执行DWS数据处理操作。

云数据迁移

数据开发模块服务依赖云数据迁移（Cloud Data Migration，CDM）实现数据迁移相关的数据处理。

机器学习服务

数据开发模块服务依赖机器学习服务（Machine Learning Service，MLS）实现机器学习相关的数据处理。

数据湖探索

数据开发模块服务依赖数据湖探索（Data Lake Insight，DLI）实现数据探索相关的数据处理。

云搜索服务

数据开发模块服务依赖云搜索服务（Cloud Search Service）实现云搜索相关的数据处理。

消息通知服务

数据开发模块服务的通知管理功能依赖消息通知服务（Simple Message Notification，SMN）发送作业信息给用户。

实时流计算服务

数据开发模块依赖实时流计算服务（Cloud Stream Service，CS）实现实时流式大数据分析。

表格存储服务

数据开发模块支持将数据存储在表格存储服务（CloudTable Service，CloudTable）中。

数据接入服务

数据开发模块依赖数据接入服务（Data Ingestion Service，DIS）实现数据转储相关的数据处理。

统一身份认证服务

统一身份认证服务（Identity and Access Management，IAM）为数据开发模块提供了鉴权功能。

7 约束限制

使用数据开发模块前，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版本登录数据开发模块。
  - Google Chrome：54.0及更高版本

8 区域和可用区

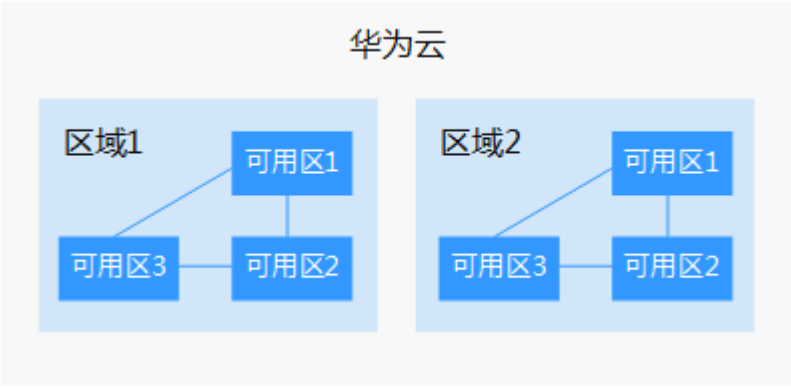
什么是区域、可用区?

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8-1阐明了区域和可用区之间的关系。

图 8-1 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。

  香港、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**。

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，请参阅**地区和终端节点**。

9 配额说明

目前默认每个用户最多可以创建1000个作业。

说明

作业使用时涉及到的其他资源服务，请参见其他服务的配额。

