ALIBABA CLOUD

阿里云

大数据计算服务

产品简介

文档版本：20220711

阿里云

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. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，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
危险	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。	危险 重置操作将丢失用户配置数据。
警告	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。
注意	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。
说明	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	命令或代码。	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MaxCompute

MaxCompute（ODPS）是适用于数据分析场景的企业级SaaS（Software as a Service）模式云数据仓库，以Serverless架构提供快速、全托管的在线数据仓库服务，消除了传统数据平台在资源扩展性和弹性方面的限制，最小化用户运维投入，使您可以经济并高效地分析处理海量数据。

随着数据收集手段不断丰富，行业数据大量积累，数据规模已增长到了传统软件行业无法承载的海量数据（TB、PB、EB）级别。MaxCompute提供离线和流式数据的接入，支持大规模数据计算及查询加速能力，为您提供面向多种计算场景的数据仓库解决方案及分析建模服务。MaxCompute还为您提供完善的数据导入方案以及多种经典的分布式计算模型，您可以不必关心分布式计算和维护细节，便可轻松完成大数据分析。

MaxCompute适用于100 GB以上规模的存储及计算需求，最大可达EB级别，并且MaxCompute已经在阿里巴巴集团内部得到大规模应用。MaxCompute适用于大型互联网企业的数据仓库和BI分析、网站的日志分析、电子商务网站的交易分析、用户特征和兴趣挖掘等。详细发展历程、产品荣誉及客户案例请参见发展历程和客户案例。

MaxCompute还深度融合了阿里云如下产品：

- DataWorks

  基于DataWorks实现一站式的数据同步、业务流程设计、数据开发、管理和运维功能。

- 机器学习PAI

  基于机器学习平台的算法组件实现对MaxCompute数据进行模型训练等操作。

- Quick BI

  基于Quick BI对MaxCompute数据进行报表制作，实现数据可视化分析。

MaxCompute融合的更多阿里云产品信息，请参见支持的云服务。

视频简介

学习路径

您可以通过MaxCompute学习路径快速了解MaxCompute的相关概念、基础操作、进阶操作等。

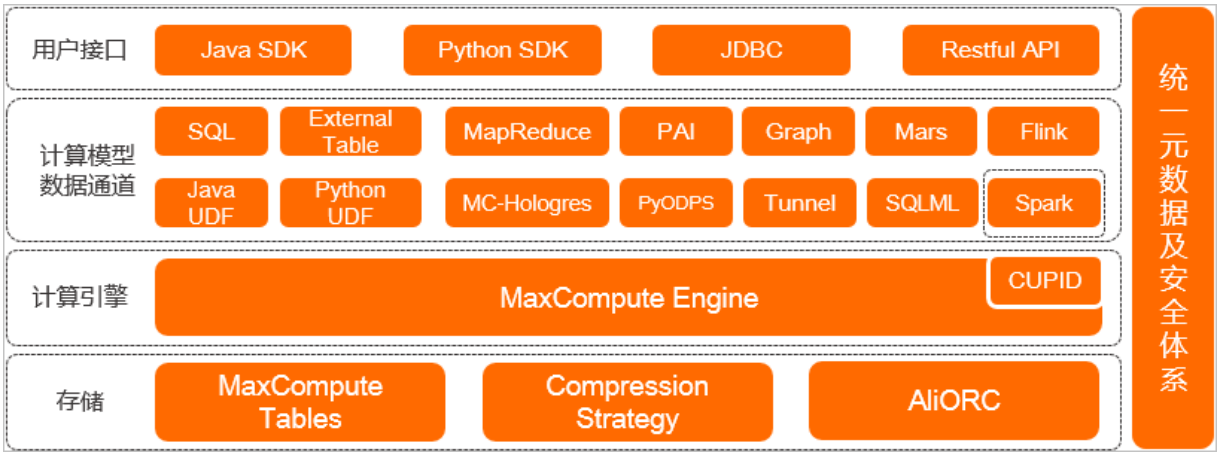
核心功能

功能分类	功能描述
全托管的Serverless在线服务	• 对外以API方式访问的在线服务，开箱即用。 • 预铺设大规模集群资源，可以按需使用、按量计费。 • 无需平台运维，最小化运维投入。
弹性能力与扩展性	• 存储和计算独立扩展，支持企业将全部数据资产在一个平台上进行联动分析，消除数据孤岛。 • 支持实时根据业务峰谷变化分配资源。
统一丰富的计算和存储能力	• MaxCompute支持多种计算模型和丰富的UDF。 • 采用列压缩存储格式，通常情况下具备5倍压缩能力，可以大幅节省存储成本。
与DataWorks深度集成	一站式数据开发与治理平台DataWorks，可实现全域数据汇聚、融合加工和治理。DataWorks支持对MaxCompute项目进行管理以及Web端查询编辑。

功能分类	功能描述
集成AI能力	• 与机器学习平台PAI无缝集成，提供强大的机器学习处理能力。 • 您可以使用熟悉的Spark-ML开展智能分析。 • 使用Python机器学习三方库。
深度集成Spark引擎	• 内建Apache Spark引擎，提供完整的Spark功能。 • 与MaxCompute计算资源、数据和权限体系深度集成。
湖仓一体	• 集成对数据湖（OSS或Hadoop HDFS）的访问分析，支持通过外部表映射、Spark直接访问方式开展数据湖分析。 • 在一套数据仓库服务和用户接口下，实现数据湖与数据仓库的关联分析。 详细信息，请参见MaxCompute湖仓一体。
支持流式采集和近实时分析	• 支持流式数据实时写入并在数据仓库中开展分析。 • 与云上主要流式服务深度集成，轻松接入各种来源的流式数据。 • 支持高性能秒级弹性并发查询，满足近实时分析场景需求。
提供持续的SaaS化云上数据保护	为云上企业提供基础设施、数据中心、网络、供电、平台安全能力、用户权限管理、隐私保护等三级超20项安全功能，兼具开源大数据与托管数据库的安全能力。

产品架构

MaxCompute的产品架构如下。



模块名称	功能说明
存储	• MaxCompute Tables：表是MaxCompute的数据存储单元。MaxCompute中不同类型作业的操作对象（输入、输出）都是表。 • Compression Strategy：MaxCompute采用列压缩存储格式，通常情况下具备5倍压缩能力。 • AliORC：MaxCompute数据存储格式全面升级为AliORC，具备更高存储性能。

模块名称	功能说明
计算引擎	MaxCompute本身具备计算引擎能力。在处理Spark作业时，MaxCompute运行在阿里云自研的CUPID平台之上，可以原生支持开源社区Yarn所支持的计算框架。
计算模型数据通道	MaxCompute支持多种数据通道满足多场景需求： • SQL：MaxCompute对外提供SQL功能。您可以将MaxCompute作为传统的数据库软件操作，但其却能处理EB级别的海量数据。 说明 ◦ MaxCompute SQL不支持事务、索引。 ◦ MaxCompute的SQL语法与Oracle、MySQL有一定差别，您无法将其他数据库中的SQL语句无缝迁移至MaxCompute中。详情请参见与其他SQL语法的差异。 ◦ MaxCompute主要用于100 GB以上规模的数据计算，因此MaxCompute SQL最快支持在分钟或秒钟级别完成查询返回结果，但无法在毫秒级别返回结果。 ◦ MaxCompute SQL的优点是学习成本低，您不需要了解复杂的分布式计算概念。如果您具备数据库操作经验，便可快速熟悉MaxCompute SQL的使用。 • External Table：提供处理除MaxCompute内部表以外的其他数据的能力。您可以通过一条简单的DDL语句，在MaxCompute上创建一张外部表，通过外部表关联外部数据源。 • Java UDF：当MaxCompute的内建函数无法满足计算需求时，您可以通过Java构建自定义函数。 • Python UDF：当MaxCompute的内建函数无法满足计算需求时，您可以通过Python构建自定义函数。 • MapReduce：MapReduce是MaxCompute提供的Java MapReduce编程模型，它可以简化开发流程，更为高效。 • Hologres：Hologres与MaxCompute在底层无缝连接，您无须移动数据，即可使用标准的PostgreSQL语句查询分析MaxCompute中的海量数据，快速获取查询结果。 • PAI：PAI是基于MaxCompute的一款机器学习算法平台。它实现了数据无需搬迁，便可进行从数据处理、模型训练、服务部署到预测的一站式机器学习。 • PyODPS：PyODPS是MaxCompute的Python版本的SDK，提供简单方便的Python编程接口。 • Graph：Graph是一套面向迭代的图计算处理框架。 • Tunnel：提供高并发的数据上传下载服务。 • Mars：Mars是一个基于张量的统一分布式计算框架。Mars能利用并行和分布式技术，为Python数据科学栈加速。 • SQLML：SQLML功能依赖MaxCompute和机器学习PAI。您可以通过客户端开发MaxCompute SQLML作业，基于机器学习PAI对MaxCompute上的数据进行学习，并利用机器学习模型对数据进行预测，进而为业务规划提供指导。 • Flink：Flink为MaxCompute提供实时数据处理能力。 • Spark：Spark是MaxCompute提供的兼容开源Spark的计算服务。它在统一的计算资源和数据集权限体系之上，提供Spark计算框架，支持您以熟悉的开发使用方式提交运行Spark作业，满足更丰富的数据处理分析需求。

模块名称	功能说明
用户接口	MaxCompute提供如下用户接口： • Java SDK • Python SDK • JDBC • Restful API
统一元数据及安全体系	MaxCompute的Information Schema提供项目元数据及使用历史数据等信息，您可以对作业的运行情况，例如资源消耗、运行时长、数据处理量等指标进行分析，用于优化作业或规划资源容量。 MaxCompute还提供了完善的安全管理体系，例如访问控制、数据加密、动态脱敏等为数据安全性提供保障。更多安全相关信息，请参见安全管理。

产品优势

MaxCompute的主要优势如下：

- 简单易用
  - 面向数据仓库实现高性能存储、计算。
  - 预集成多种服务，标准SQL开发简单。
  - 内建完善的管理和安全能力。
  - 免运维，按量付费，不使用不产生费用。
- 匹配业务发展的弹性扩展能力

  存储和计算独立扩展，动态扩缩容，按需弹性扩展，无需提前规划容量，满足突发业务增长。
- 支持多种分析场景

  支持开放数据生态，以统一平台满足数据仓库、BI、近实时分析、数据湖分析、机器学习等多种场景。
- 开放的平台
  - 支持开放接口和生态，为数据、应用迁移、二次开发提供灵活性。
  - 支持与Airflow、Tableau等开源和商业产品灵活组合，构建丰富的数据应用。

联系我们

如果您在使用MaxCompute的过程中有任何疑问或建议，欢迎填写钉钉群申请表单加入钉钉群进行反馈。

2.基本概念
2.1. 术语表

在开始使用MaxCompute产品前，您可以提前查阅MaxCompute所涉及的术语及其含义，为了解产品及快速上手提供帮助。本文为您介绍MaxCompute涉及的术语及其概念。

A

- AccessKey

  简称AK，包括AccessKey ID和AccessKey Secret，是访问阿里云API的密钥。在阿里云官网注册云账号后，可以在AccessKey管理页面生成该信息，用于标识用户，为访问MaxCompute、其他阿里云产品或连接第三方工具做签名验证。请妥善保管AccessKey Secret，必须保密，如果存在泄露风险，请及时禁用或更新AccessKey。

- 安全

  MaxCompute提供多租户数据安全体系，主要包括用户认证、项目的用户与授权管理、跨项目的资源分享以及项目的数据保护。更多MaxCompute安全操作信息，请参见权限概述。

C

- Console

  即MaxCompute客户端，是运行在Window或Linux下的工具，您可以在MaxCompute客户端通过运行命令的方式完成项目管理、DDL、DML等操作。MaxCompute客户端的操作指导，请参见使用客户端（odpscmd）连接。

D

- Data Type

  MaxCompute表中列的数据类型。MaxCompute支持的数据类型版本及各版本的数据类型列表，请参见数据类型版本说明。

- DDL

  Data Definition Language，数据定义语言。例如创建表、创建视图等操作。更多DDL语法信息，请参见DDL语句。

- DML

  Data Manipulation Language，数据操作语言。例如INSERT、UPDATE、DELETE操作。更多DML语法信息，请参见DML操作。

F

- Function（函数）

  MaxCompute提供函数功能，包括内建函数和UDF。更多函数信息，请参见函数。

- fuxi（伏羲）

  伏羲是飞天平台内核中负责资源管理和任务调度的模块，同时也为应用开发提供了一套编程基础框架。MaxCompute的底层任务调度模块为fuxi的调度模块。

I

- Instance（实例）

  即实际运行作业的一个具体实例，类同Hadoop中Job的概念。详情请参见任务实例。

M

- MapReduce

  MapReduce是处理数据的一种编程模型，通常用于大规模数据集的并行运算。您可以使用MapReduce提供的接口（Java API）编写MapReduce程序，来处理MaxCompute中的数据。编程思想是将数据的处理方式分为Map（映射）和Reduce（规约）。

  在正式执行Map前，需要将输入的数据进行分片。所谓分片，就是将输入数据切分为大小相等的数据块，每一块作为单个Map Worker的输入被处理，以便于多个Map Worker同时工作。每个Map Worker在读入各自的数据后，进行计算处理，最终通过Reduce函数整合中间结果，从而得到最终计算结果。详情请参见MapReduce。

N

- Networklink（网络连接）

  当您使用外部表、UDF或湖仓一体功能时，MaxCompute默认未建立与外网或VPC网络间的网络连接，您需要开通网络连接以访问外网或VPC中的目标服务（例如HBase、RDS、Hadoop等）。更多开通网络连接信息，请参见网络开通流程。

O

- ODPS

  Open Data Processing Service，MaxCompute的原名。

P

- Partition（分区）

  分区Partition是指一张表下，根据分区字段（一个或多个字段的组合）对数据存储进行划分。如果表没有分区，数据是直接放在表所在的目录下。如果表有分区，每个分区对应表下的一个目录，数据是分别存储在不同的分区目录下。更多分区信息，请参见分区。

- Project（项目）

  项目是MaxCompute的基本组织单元，类似于传统数据库的Database或Schema的概念，是进行多用户隔离和访问控制的主要边界。更多项目信息，请参见项目。

Q

- Quota（配额）

  配额是MaxCompute的计算资源池，提供作业运行所需计算资源。更多配额信息，请参见配额。

R

- Role（角色）

  角色是MaxCompute安全功能中的概念，可以理解为拥有相同权限的用户的集合。多个用户可以同时存在于一个角色下，一个用户也可以隶属于多个角色。给角色授权后，该角色下的所有用户拥有相同的权限。更多角色管理信息，请参见角色规划与管理。

- Resource（资源）

资源是MaxCompute中特有的概念。当您使用MaxCompute的自定义函数（UDF）或MapReduce功能时，需要依赖资源来完成。更多资源信息，请参见资源。

S

- SDK

  Software Development Kit，软件开发工具包。一般都是一些被软件工程师用于为特定的软件包、软件实例、软件框架、硬件平台、操作系统、文档包等建立应用软件的开发工具的集合。MaxCompute支持Java SDK和Python SDK。

- 授权

  项目管理员或者项目Owner可以授予其他角色对MaxCompute中的对象（例如表、任务、资源等）进行某种操作的权限，包括读、写、查看等。更多授权信息，请参见用户规划与管理。

- 沙箱（Sandboxie）

  沙箱是一种按照安全策略限制程序行为的执行环境。沙箱机制是一种安全机制，将Java代码限定在特定的运行范围中，并且严格限制代码对本地系统资源访问，通过这样的措施来保证对代码的有效隔离，防止对本地系统造成破坏。MaxCompute MapReduce及UDF程序在分布式环境中运行时受到Java沙箱的限制。

T

- Table（表）

  表是MaxCompute的数据存储单元。更多表信息，请参见表。

- Tunnel

  MaxCompute的数据通道，提供高并发的离线数据上传下载服务。您可以使用Tunnel服务向MaxCompute批量上传数据或者向本地进行批量数据下载。相关命令请参见Tunnel命令或批量数据通道SDK。

U

- UDF

  User Defined Function，用户自定义函数。

  广义的UDF代表了自定义标量函数、自定义聚合函数及自定义表函数三种类型。MaxCompute支持通过Java、Python编程接口开发自定义函数，详情请参见MaxCompute UDF。

  狭义的UDF指用户自定义标量值函数（User Defined Scalar Function），它的输入与输出是一对一的关系，即读入一行数据，写出一条输出值。

- UDAF

  User Defined Aggregation Function，自定义聚合函数。它的输入与输出是多对一的关系， 即将多条输入记录聚合成一条输出值。可以与SQL中的GROUP BY语句联用。详情请参见UDAF。

- UDTF

  User Defined Table Valued Function，自定义表值函数。它是唯一能返回多个字段的自定义函数。详情请参见UDTF。

- User（用户）

  用户是MaxCompute安全功能中的概念，MaxCompute支持您通过阿里云账号、RAM用户或RAM角色访问MaxCompute。非MaxCompute项目所有者（Project Owner）的用户必须被加入MaxCompute项目中，且被授予相应的权限，才能操作MaxCompute项目中的数据、作业、资源及函数。更多用户管理信息，请参见用户规划与管理。

V

- View（视图）

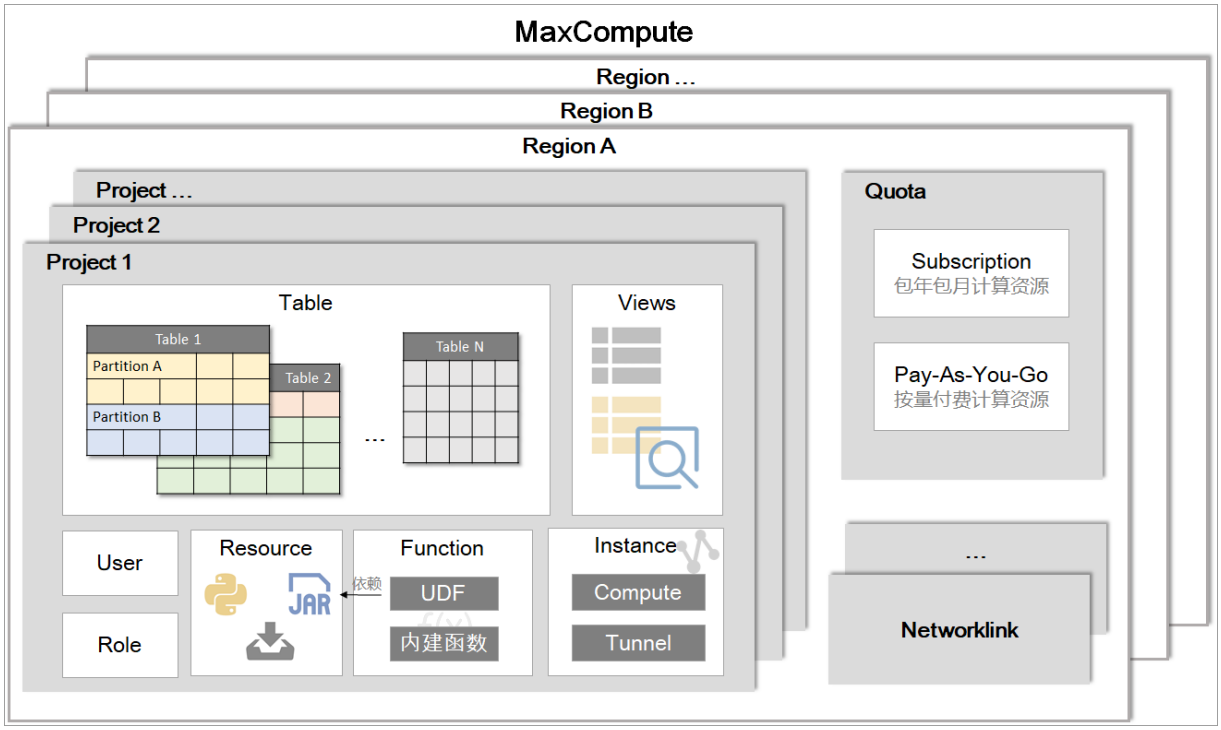
  视图是在表之上建立的虚拟表，它的结构和内容都来自表。一个视图可以对应一个表或多个表。如果您想保留查询结果，但不想创建表占用存储，可以通过视图实现。更多视图信息，请参见视图操作。

2.2. 核心概念

2.2.1. 核心概念的层次结构

MaxCompute具有层次结构，您可以通过了解其结构，为后期项目规划、安全管理等提供思路。本文为您介绍MaxCompute中核心概念的层次结构及简要含义。

MaxCompute核心概念的层次结构如下。



核心概念	说明
Project（项目）	项目是MaxCompute的基本组织单元，类似于传统数据库的Database或Schema的概念，是进行多用户隔离和访问控制的主要边界。更多项目信息，请参见项目。
Table（表）	表是MaxCompute的数据存储单元。更多表信息，请参见表。
Partition（分区）	分区Partition是指一张表下，根据分区字段（一个或多个字段的组合）对数据存储进行划分。如果表没有分区，数据是直接放在表所在的目录下。如果表有分区，每个分区对应表下的一个目录，数据是分别存储在不同的分区目录下。更多分区信息，请参见分区。

核心概念	说明
View（视图）	视图是在表之上建立的虚拟表，它的结构和内容都来自表。一个视图可以对应一个表或多个表。如果您想保留查询结果，但不想创建表占用存储，可以通过视图实现。更多视图信息，请参见视图操作。
User（用户）	用户是MaxCompute安全功能中的概念，MaxCompute支持您通过阿里云账号、RAM用户或RAM角色访问MaxCompute。非MaxCompute项目所有者（Project Owner）的用户必须被加入MaxCompute项目中，且被授予相应的权限，才能操作MaxCompute项目中的数据、作业、资源及函数。更多用户管理信息，请参见用户规划与管理。
Role（角色）	角色是MaxCompute安全功能中的概念，可以理解为拥有相同权限的用户的集合。多个用户可以同时存在于一个角色下，一个用户也可以隶属于多个角色。给角色授权后，该角色下的所有用户拥有相同的权限。更多角色管理信息，请参见角色规划与管理。
Resource（资源）	资源是MaxCompute中特有的概念。当您使用MaxCompute的自定义函数（UDF）或MapReduce功能时，需要依赖资源来完成。更多资源信息，请参见资源。
Function（函数）	MaxCompute提供函数功能，包括内建函数和UDF。更多函数信息，请参见函数。
Instance（实例）	即实际运行作业的一个具体实例，类同Hadoop中Job的概念。详情请参见任务实例。
Quota（配额）	配额是MaxCompute的计算资源池，提供作业运行所需计算资源。更多配额信息，请参见配额。
Networklink（网络连接）	当您使用外部表、UDF或湖仓一体功能时，MaxCompute默认未建立与外网或VPC网络间的网络连接，您需要开通网络连接以访问外网或VPC中的目标服务（例如HBase、RDS、Hadoop等）。更多开通网络连接信息，请参见网络开通流程。

2.2.2. 项目

项目（Project）是MaxCompute的基本组织单元，它类似于传统数据库的Database或Schema的概念，是进行多用户隔离和访问控制的主要边界。项目中包含多个对象，例如表（Table）、资源（Resource）、函数（Function）和实例（Instance）等。

MaxCompute为您提供方便的项目操作与管理。

- 开通MaxCompute服务后，需要通过项目使用MaxCompute，如何创建MaxCompute项目，详情请参见创建MaxCompute项目。

- 创建MaxCompute项目后，您需要进入项目才可以执行后续开发、分析、运维等一系列操作。详情请参见项目空间操作。
- MaxCompute提供项目数据保护机制，为数据安全提供保障。详情请参见安全操作。
- MaxCompute提供跨项目的资源访问。

  一个用户可以同时拥有多个项目的权限。通过安全授权，可以在一个项目中访问另一个项目中的对象，详情请参见基于Package跨项目访问资源。

> **说明** MaxCompute项目即DataWorks的工作空间。详情请参见DataWorks简单模式与标准模式工作空间。

MaxCompute支持一种特殊类型的项目，即外部项目（External Project）。

- 外部项目无法被独立创建和使用，需要配合数据湖集成，用以实现访问和管理Hadoop集群Hive数据库中的表数据，或数据湖构建DLF中的表数据。详情参见MaxCompute湖仓一体。
- 外部项目本身没有执行作业的权限，需要到关联MaxCompute项目，通过<external_project_name>.<table_name>的方式访问外部项目中的表数据。详情请参见使用SQL管理外部项目。
- 外部项目本身不产生计费，查询所用的计算资源归属为关联的MaxCompute内部项目。

2.2.3. 配额

配额（Quota）是MaxCompute的计算资源池，为MaxCompute SQL、MapReduce、Spark、Mars、PAI等计算作业提供所需计算资源（CPU及内存）。

MaxCompute计算资源单位为CU，1 CU包含1 CPU及4 GB内存。您可购买的Quota分为包年包月计算资源和按量计费计算资源两种，分别对应包年包月规格类型和按量计费规格类型，更多规格信息，请参见规格类型。

如果您购买的Quota为包年包月计算资源，可进一步通过MaxCompute管家进行如下更细粒度的管理：

- 设置配额组

  支持新建、修改或删除配额组，也支持设置配额组的分时时间段，满足不同业务项目在不同时间段对计算资源的需求。
- 修改项目配额组

  支持修改MaxCompute项目关联的配额组。
- 包年包月项目支持按量计费配额

  支持在MaxCompute包年包月项目中对指定SQL使用按量计费配额组。

您可以通过如下方式关联MaxCompute项目及配额组，项目关联配额组后，在MaxCompute项目中提交的计算作业默认使用所关联的配额组进行计算：

- 在创建MaxCompute项目时，您可以通过**配额组**参数选择需要关联的配额组。
- 对于存量项目，您可以通过MaxCompute控制台的**切换配额组**功能修改项目关联的配额组。或通过MaxCompute管家的**项目**页签修改项目关联的配额组，更多操作信息，请参见修改项目配额组。

> **说明** 建议您根据项目业务情况为不同的MaxCompute项目关联不同的配额组。

2.2.4. 表

表是MaxCompute的数据存储单元。它在逻辑上是由行和列组成的二维结构，每行代表一条记录，每列表示相同数据类型的一个字段，一条记录可以包含一个或多个列，表的结构由各个列的名称和类型构成。

MaxCompute中不同类型计算任务的操作对象（输入、输出）都是表。您可以创建表、删除表以及向表中导入数据。

> **说明** DataWorks的数据管理模块可以对MaxCompute表进行新建、收藏、修改数据生命周期管理、修改表结构和数据表/资源/函数权限管理审批等操作。

MaxCompute的表格有两种类型：内部表和外部表（MaxCompute 2.0版本开始支持外部表）。

- 对于内部表，所有的数据都被存储在MaxCompute中，表中列的数据类型可以是MaxCompute支持的任意一种数据类型版本说明。
- 对于外部表，MaxCompute并不真正持有数据，表格的数据可以存放在OSS或OTS中 。MaxCompute仅会记录表格的Meta信息，您可以通过MaxCompute的外部表机制处理OSS或OTS上的非结构化数据，例如，视频、音频、基因、气象、地理信息等。

2.2.5. 分区

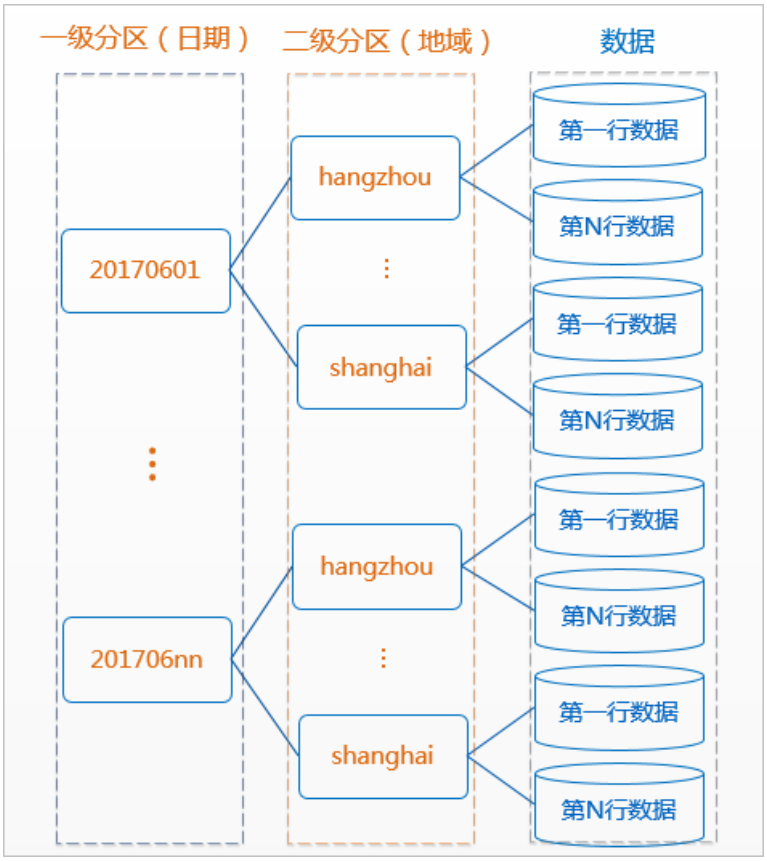
分区表是指拥有分区空间的表，即在创建表时指定表内的一个或者某几个字段作为分区列。分区表实际就是对应分布式文件系统上的独立的文件夹，一个分区对应一个文件夹，文件夹下是对应分区所有的数据文件。

概述

分区可以理解为分类，通过分类把不同类型的数据放到不同的目录下。分类的标准就是分区字段，可以是一个，也可以是多个。

MaxCompute将分区列的每个值作为一个分区（目录），您可以指定多级分区，即将表的多个字段作为表的分区，分区之间类似多级目录的关系。

分区表的意义在于优化查询。查询表时通过WHERE子句查询指定所需查询的分区，避免全表扫描，提高处理效率，降低计算费用。使用数据时，如果指定需要访问的分区名称，则只会读取相应的分区。



部分对分区操作的SQL的运行效率较低，会给您带来较高的费用，例如插入或覆写动态分区数据（DYNAMIC PARTITION）。

对于部分操作MaxCompute的命令，处理分区表和非分区表时语法有差别，详情请参见表操作和INSERT操作。

使用限制

- 单表分区层级最多为6级。
- 单表分区数最大值为60000个。
- 单次查询允许查询最多的分区个数为10000个。
- STRING分区类型的分区值不支持使用中文。

分区列的数据类型

MaxCompute 2.0数据类型版本支持的分区字段为TINYINT、SMALLINT、INT、BIGINT、VARCHAR、STRING。

MaxCompute 1.0数据类型版本支持的分区字段仅有STRING。虽然可以指定分区列的类型为BIGINT，但是除了表的字段显示为BIGINT类型，任何其他情况（例如，字段的计算和比较）下都当作STRING类型处理。执行如下语句后，返回的结果只有一行。

```
---创建表parttest。
create table parttest (a bigint) partitioned by (pt bigint);
---向表中插入数据。
insert into parttest partition(pt)(a,pt) values (1, 1);
insert into parttest partition(pt)(a,pt) values (1, 10);
---查询表中字段pt大于等于2的行。
select * from parttest where pt >= '2';
```

示例

- 创建分区。

```
--创建一个二级分区表，以日期为一级分区，地域为二级分区
CREATE TABLE src (shop_name string, customer_id bigint) PARTITIONED BY (pt string,region
string);
```

- 使用分区列作为过滤条件查询数据。

```
--正确使用方式。MaxCompute在生成查询计划时只会将'20170601'分区下region为'hangzhou'二级分区的数据
纳入输入中。
select * from src where pt='20170601'and region='hangzhou';
--错误的使用方式。在这样的使用方式下，MaxCompute并不能保障分区过滤机制的有效性。pt是STRING类型，当S
TRING类型与BIGINT（20170601）比较时，MaxCompute会将二者转换为DOUBLE类型，此时有可能会有精度损失。
select * from src where pt = 20170601;
```

2.2.6. 生命周期

本文为您介绍MaxCompute表的生命周期概念。

MaxCompute表的生命周期（Lifecycle），指表（分区）数据从最后一次更新的时间算起，在经过指定的时间后没有变动，则此表（分区）将被MaxCompute自动回收。这个指定的时间就是生命周期。

- 生命周期单位为天，取值为正整数。
- 对于非分区表，如果表数据在生命周期内没有被修改，经过指定天数后此表将会被MaxCompute自动回收（类似DROP TABLE操作）。生命周期从最后一次表数据被修改的时间（LastDataModifiedTime）起开始计算。
- 对于分区表，每个分区可以分别被回收。在生命周期内未被修改数据的分区，经过指定的天数后此分区将会被回收，否则会被保留。每个分区的生命周期是从最后一次分区数据被修改的时间（LastDataModifiedTime）起开始计算。不同于非分区表，分区表的最后一个分区被回收后，该表不会被删除。

> **说明**
>
> - 生命周期回收为每天定时启动，扫描全量分区。LastDataModifiedTime需要超过生命周期指定的时间才回收。
>
> 假设某个分区表生命周期为1天，该分区数据最后一次被修改的时间是2020年02月17日15时。如果在2020年02月18日15时之前扫描此表（不到一天），则不会回收表分区。如果2020年02月19日回收扫描时发现表分区LastDataModifiedTime超过生命周期指定的时间，则上述分区会被回收。
> - 生命周期主要提供定期回收表或分区的功能，每天根据服务的繁忙程度，不定时回收。不能确保表或分区的生命周期到期后，立刻被回收。
> - 删除表后，表的所有属性信息全部会删除，包括生命周期。新建同名表后，表的生命周期以新设置的属性为准。

- 只能在表级别设置生命周期，不能在分区级设置生命周期。为分区表指定的生命周期，适用于该表所有的分区。创建表时即可指定生命周期。
- 如果您没有为表指定生命周期，则表（分区）不会根据生命周期规则被MaxCompute自动回收。

关于建表时如何指定、修改表生命周期、修改表`LastDataModifiedTime`等操作，请参见表操作。

2.2.7. 资源

本文为您介绍MaxCompute的资源（Resource）概念，可为MaxCompute特定操作提供资源依赖。

概念

资源（Resource）是MaxCompute的特有概念，如果您想使用MaxCompute的自定义函数（UDF）或MapReduce功能需要依赖资源来完成，如下所示：

- SQL UDF：您编写UDF后，需要将编译好的JAR包以资源的形式上传到MaxCompute。运行此UDF时，MaxCompute会自动下载这个JAR包，获取您的代码来运行UDF，无需您干预。上传JAR包的过程就是在MaxCompute上创建资源的过程，这个JAR包是MaxCompute资源的一种。
- MapReduce：您编写MapReduce程序后，将编译好的JAR包作为一种资源上传到MaxCompute。运行MapReduce作业时，MapReduce框架会自动下载这个JAR资源，获取您的代码。

您同样可以将文本文件以及MaxCompute中的表作为不同类型的资源上传到MaxCompute，您可以在UDF及MapReduce的运行过程中读取、使用这些资源。MaxCompute提供了读取、使用资源的接口。详情请参见资源使用示例及UDF使用说明。

> **说明** MaxCompute的自定义函数（UDF）或MapReduce对资源的读取有一定的限制，详情请参见使用限制。

资源类型

MaxCompute支持上传的单个资源大小上限为500 MB，资源包括以下几种类型：

- File类型：仅支持.zip、.so和.jar类型的File资源。
- Table类型：MaxCompute中的表。

> **说明** MapReduce引用的Table类型资源中，Table字段类型目前只支持BIGINT、DOUBLE、STRING、DATETIME、BOOLEAN，其他类型暂未支持。

- JAR类型：编译好的Java JAR包。
- Archive类型：通过资源名称中的后缀识别压缩类型，支持的压缩文件类型包括.zip、.tgz、.tar.gz、.tar、.jar。
- Python类型：您编写的Python代码，用于注册Python UDF函数。

资源的相关操作请参见资源操作或MaxCompute资源。

2.2.8. 函数

本文为您介绍MaxCompute提供的函数功能，包括内建函数和UDF。

MaxCompute为您提供了SQL计算功能，您可以在MaxCompute SQL中使用系统的内建函数完成一定的计算和计数功能。但当内建函数无法满足要求时，您可以使用MaxCompute提供的Java编程接口开发自定义函数（User Defined Function，以下简称UDF）。

自定义函数（UDF）可以进一步分为标量值函数（UDF），自定义聚合函数（UDAF）和自定义表值函数（UDTF）三种类型。

您在开发完成UDF代码后，需要将代码编译成Jar包，并将此Jar包以Jar资源的形式上传到MaxCompute，最后在MaxCompute中注册此UDF。

> **说明** 使用UDF时，只需在SQL中指明UDF的函数名及输入参数即可，使用方式与MaxCompute提供的内建函数相同。

函数的相关操作请参见创建函数， 删除函数及查看函数清单。

2.2.9. 任务

任务（Task）是MaxCompute的基本计算单元，SQL及MapReduce功能都是通过任务完成的。

对于您提交的大多数任务，特别是计算型任务，例如SQL DML语句、MapReduce，MaxCompute会对其进行解析，得到任务的执行计划。执行计划由具有依赖关系的多个执行阶段（Stage）构成。

目前，执行计划逻辑上可以被看做一个有向图，图中的点是执行阶段，各个执行阶段的依赖关系是图的边。MaxCompute会依照图（执行计划）中的依赖关系执行各个阶段。在同一个执行阶段内，会有多个进程，也称之为Worker，共同完成该执行阶段的计算工作。同一个执行阶段的不同Worker只是处理的数据不同，执行逻辑完全相同。计算型任务在执行时，会被实例化，您可以对这个实例（Instance）进行操作，例如获取实例状态（Status Instance）、终止实例运行（Kill Instance）等。

部分MaxCompute任务并不是计算型的任务，例如SQL中的DDL语句，这些任务本质上仅需要读取、修改MaxCompute中的元数据信息。因此，这些任务无法被解析出执行计划。

> **说明** 在MaxCompute中，并不是所有的请求都会被转化为任务（Task），例如项目空间（Project）、资源（Resource）、自定义函数（UDF）及实例（Instance）的操作均不需要通过MaxCompute的任务来完成。

2.2.10. 任务实例

本文向您介绍MaxCompute任务实例及实例状态。

在MaxCompute中，SQL、Spark和Mapreduce任务在执行时会被实例化，以MaxCompute实例（下文简称为实例或Instance）的形式存在。实例会经历运行（Running）和结束（Terminated）两个阶段。

运行阶段的实例状态为Running（运行中），而结束阶段则会有Success（成功）、Failed（失败）和Canceled（被取消）三种状态。您可以根据运行任务时MaxCompute给出的实例ID进行查询、改变任务的状态等操作，示例如下。

```
--查看某实例的状态。
status instance_id;
--停止某实例，将其状态设置为Canceled。
kill instance_id;
--查看某实例的运行日志。
wait instance_id;
```

其中，instance_id为需要查询的实例ID号，请您根据实际的ID号进行替换。

2.3. ACID语义

本文为您介绍MaxCompute在作业并发情况下ACID的语义及Transactional表的ACID语义。

相关术语

- 操作：指在MaxCompute上提交的单个作业。
- 数据对象：指持有实际数据的对象，例如非分区表、分区。
- INTO类作业：指INSERT INTO、DYNAMIC INSERT INTO等包含关键字INTO的SQL作业。
- OVERWRITE类作业：指INSERT OVERWRITE、DYNAMIC INSERT OVERWRITE等包含关键字OVERWRITE的SQL作业。
- Tunnel数据上传：可以归结为INTO类或OVERWRITE类作业。

ACID语义描述

- 原子性（Atomicity）：一个操作或是全部完成，或是全部不完成，不会结束在中间某个环节。
- 一致性（Consistency）：从操作开始至结束的期间，数据对象的完整性没有被破坏。
- 隔离性（Isolation）：操作独立于其他并发操作完成。
- 持久性（Durability）：操作处理结束后，对数据的修改将永久有效，即使出现系统故障，该修改也不会丢失。

MaxCompute并发写操作的ACID特性

- 原子性（Atomicity）
  - 任何时候MaxCompute会保证在冲突时只有一个作业执行成功，其他冲突作业执行失败。
  - 对于单个表或分区的CREATE、OVERWRITE、DROP操作，可以保证其原子性。
  - 跨表操作时不支持原子性（例如MULTI-INSERT）。
  - 在极端情况下，以下操作可能不保证原子性：
    - `DYNAMIC INSERT OVERWRITE` 多于一万个分区，不支持原子性。
    - INTO类操作：这类操作失败的原因是事务回滚时数据清理失败，但不会造成原始数据丢失。
- 一致性（Consistency）
  - OVERWRITE类作业可保证一致性。
  - INTO类作业在冲突失败后可能存在失败作业的数据残留。
- 隔离性（Isolation）

  - 非INTO类操作保证读已提交。
  - INTO类操作存在读未提交的场景。
- 持久性（Durability）
  - MaxCompute保证数据的持久性。

Transactional表的ACID特性

Transactional表的ACID特性在MaxCompute并发写操作的ACID特性基础上，支持如下新特性：

- INTO类操作保证读已提交，作业冲突执行失败后无数据残留。
- 对于单个非分区表或单个分区的UPDATE、DELETE、MERGE小文件操作，可以保证其原子性。

  例如，当两个UPDATE操作并发修改同一分区时，只会有一个UPDATE操作执行成功。不会存在一个UPDATE操作部分执行成功，也不会存在两个UPDATE操作分别执行成功的情况。

操作并发冲突说明

当作业并发运行且写入相同目标表时，可能出现冲突。产生冲突时，先结束的作业会执行成功，后结束的作业可能会因冲突而报错。

下表为作业并发提交场景下，对同一个非分区表或分区的并发操作先后结束的冲突说明。

作业类型	INSERT OVERWRITE/TRUNCATE作业（后结束）	INSERT INTO作业（后结束）	UPDATE/DELETE作业（后结束）	MERGE小文件作业（后结束）
INSERT OVERWRITE/TRUNCATE作业（先结束）	• 先、后结束的作业都会执行成功。 • INSERT OVERWRITE/TRUNCATE作业会覆盖先结束的INSERT OVERWRITE/TRUNCATE作业的数据。	• 先、后结束的作业都会执行成功。 • INSERT INTO作业会在先结束的INSERT OVERWRITE/TRUNCATE作业的数据上追加数据。	• 后结束的UPDATE/DELETE作业会报错。 • 后结束的UPDATE/DELETE作业对应的非分区表或分区被先结束的INSERT OVERWRITE/TRUNCATE作业修改。	• 后结束的MERGE小文件作业会报错。 • 后结束的MERGE小文件作业对应的非分区表或分区被先结束的INSERT OVERWRITE/TRUNCATE作业修改。
INSERT INTO作业（先结束）	• 先、后结束的作业都会执行成功。 • 后结束的INSERT OVERWRITE/TRUNCATE作业会覆盖先结束的INSERT INTO作业的数据。	• 先、后结束的作业都会执行成功。 • 后结束的INSERT INTO作业会在先结束的INSERT INTO作业的数据上追加数据。	• 后结束的UPDATE/DELETE作业会报错。 • 后结束的UPDATE/DELETE作业对应的非分区表或分区被先结束的INSERT INTO作业修改。	• 后结束的MERGE小文件作业会报错。 • 后结束的MERGE小文件作业对应的非分区表或分区被先结束的INSERT INTO作业修改。

作业类型	INSERT OVERWRITE/TRUNCATE作业（后结束）	INSERT INTO作业（后结束）	UPDATE/DELETE作业（后结束）	MERGE小文件作业（后结束）
UPDATE/DELETE作业（先结束）	• 先、后结束的作业都会执行成功。 • 后结束的INSERT OVERWRITE/TRUNCATE作业会覆盖先结束的UPDATE/DELETE作业的数据。	• 先、后结束的作业都会执行成功。 • 后结束的INSERT INTO作业会在先结束的UPDATE/DELETE作业的数据上追加数据。	• 后结束的UPDATE/DELETE作业会报错。 • 后结束的UPDATE/DELETE作业对应的非分区表或分区被先结束的UPDATE/DELETE作业修改。	• 后结束的MERGE小文件作业会报错。 • 后结束的MERGE小文件作业对应的非分区表或分区被先结束的INSERT INTO作业修改。
MERGE小文件作业（先结束）	• 先、后结束的作业都会执行成功。 • 后结束的INSERT OVERWRITE/TRUNCATE作业会覆盖先结束的MERGE小文件作业的数据。	• 先、后结束的作业都会执行成功。 • 后结束的INSERT INTO作业会在先结束的MERGE小文件作业的数据上追加数据。	• 后结束的UPDATE/DELETE作业会报错。 • 后结束的UPDATE/DELETE作业对应的非分区表或分区被先结束的MERGE小文件作业修改。	• 后结束的MERGE小文件作业会报错。 • 后结束的MERGE小文件作业对应的非分区表或分区被先结束的MERGE小文件作业修改。

综上所述，冲突报错规则概括如下：

- INSERT类操作不会因为数据变化而产生冲突报错。
- UPDATE、DELETE、MERGE小文件操作会因为目标非分区表或分区数据变化而产生冲突报错。

> **说明** 需要注意的是，在极端情况下，如果多个作业并发且元数据正处于更新阶段，可能因元数据更新而产生冲突报错。

3.使用须知

本文根据您的角色推荐不同的文档阅读顺序。

如果您是MaxCompute初学者

如果您是初学者，建议先熟悉如下模块，然后再有针对性地对深入学习其他模块。

模块	说明
产品简介	介绍MaxCompute产品的概况、主要功能、应用场景、使用限制及基本概念。通过阅读该章节，您会对MaxCompute有一个总体的认知。
准备工作	通过示例指导您如何准备账号、准备环境、创建表、导入数据、运行SQL及导出结果数据。
快速入门	
常用命令列表	介绍MaxCompute的常用命令。您可以进一步熟悉如何操作MaxCompute。
工具	您需要在分析数据前掌握MaxCompute涉及的客户端、查询编辑器、MaxCompute Studio等工具。
Endpoint	介绍MaxCompute各地域支持的连接方式及Endpoint信息，并对您在与其他云产品（ECS、Tablestore或OSS）互访场景中遇到的网络连通性和下载数据收费等问题进行说明。

如果您是数据分析师

如果您是数据分析师，建议熟读SQL模块的内容。您可以查询并分析存储在MaxCompute上的大规模数据。MaxCompute SQL支持如下主要功能。

功能项	说明
DDL操作	支持管理表、分区、列、生命周期及视图。
DML操作	支持插入或更新表、分区数据。
DQL操作	支持SELECT、子查询等多种查询操作。
增强操作	支持通过命令导入导出MaxCompute表中的数据、复制表数据等SQL增强操作。
内建函数	支持通过内建数学函数、窗口函数、日期函数、聚合函数、字符串函数等处理数据。
UDF	支持通过创建自定义函数来满足更多的计算需求。

如果您拥有一定开发经验

如果您拥有一定的开发经验，了解分布式概念，且希望解决某些无法用SQL实现的数据分析问题，推荐您学习MaxCompute更高级的功能模块。

模块	说明
MapReduce	MaxCompute提供Java MapReduce编程模型。您可以使用MapReduce提供的接口（Java API）编写MapReduce程序，处理MaxCompute中的数据。
Graph	一套面向迭代的图计算处理框架。使用图进行建模，图由点（Vertex）和边（Edge）组成，点和边包含权值（Value）。通过迭代对图进行编辑、演化，最终得出结果。
Tunnel	您可以使用Tunnel服务向MaxCompute批量上传离线数据或从MaxCompute下载离线数据。
Java SDK	向开发者提供的Java接口。
Python SDK	向开发者提供的Python接口。

如果您是项目Owner或管理员

如果您是一个项目的Owner（创建和使用项目）或管理员（管理项目、安全和费用）需要熟知如下模块。

模块	子模块	说明
	创建项目前期工作	项目（Project）是MaxCompute的基本组织单元，它类似于传统数据库的Database或Schema的概念，是进行多用户隔离和访问控制的主要边界。一个用户可以同时拥有多个项目的权限，通过安全授权，可以在一个项目中访问另一个项目中的对象，例如表（Table）、资源（Resource）、函数（Function）和实例（Instance）。使用MaxCompute，实际是操作项目中的各种对象。前期准备工作如下： • 资源预算 MaxCompute收费资源主要包含存储、计算和公网下载流量。 ○ 存储资源：按量阶梯计费。您可以按照数据量套用公式预估费用。由于数据不是当天全部存储在MaxCompute，且每时每刻都会存在数据导入导出，所以预算结果不是绝对值。 ○ 计算资源：计算资源分为按量计费和包年包月模式。由于使用初期不容易评估计算资源使用量，建议您先使用按量计费模式，测试一段时间后根据使用量再决定是否使用包年包月模式。 ○ 外网下载流量：按量计费，只有通过外网下载才会收费。 详细计费说明请参见计费项与计费方式概述。 • 准备账号并开通服务 创建MaxCompute项目前，必须先开通MaxCompute服务，且只能将阿里云账号作为主账号，同时该账号为计费主体。确定账号后，在开通MaxCompute服务时，您需要根据资源预算结论选择按量计费或包年包月模式。
	创建项目	创建项目具体操作，请参见创建MaxCompute项目。 创建项目时，需要从业务角度考虑选择标准模式或简单模式项目，从安全角度考虑使用个人账号或计算引擎指定账号，详情请参见MaxCompute访问身份和MaxCompute数据安全管理指南。

模块	子模块	说明
项目管理	项目成员管理	成员管理主要考虑成员的职责和安全问题，如果通过DataWorks使用MaxCompute，您需要考虑两个产品之间的关联权限，详情请参见MaxCompute和DataWorks的权限关系。
	RAM用户管理	MaxCompute项目支持阿里云账号和RAM用户两种账号体系。您可以将阿里云账号下的任意RAM用户加入MaxCompute的某一个项目中。RAM用户详情，请参见准备RAM用户。 通过DataWorks使用MaxCompute和DataWorks的工作空间，仅支持添加阿里云账号下的RAM用户为成员。因此，需要阿里云账号通过RAM系统创建RAM用户，并对RAM用户进行维护管理。 **说明** • MaxCompute不提供RAM中的权限定义，因此RAM用户的权限还需通过MaxCompute命令行或者DataWorks的项目管理功能实现，详情请参见MaxCompute数据安全管理指南。 • 建议一个RAM用户对应一个项目成员，禁止多个成员共用一个RAM用户。 • 离职或转岗的成员，需要及时清理对应RAM用户账号。若RAM用户在DataWorks中被加为项目成员，请先清除项目成员再到RAM系统中删除RAM用户。
	调度资源管理	• 调度资源 即DataWorks上的调度资源，调度资源用于执行或分发调度系统下发的任务。DataWorks的调度资源分为如下两种模式，详情请参见查看资源组列表。 ○ 默认调度资源。指DataWorks的公共资源池。当DataWorks节点并发量很高，调度资源紧张时会进入等待调度状态。直到占用到资源，节点才开始执行下发任务。 ○ 自定义调度资源。指将您自助购买的ECS配置为可以执行分发任务的调度服务器。阿里云账号可以新建自定义调度资源，调度资源包括若干台物理机或ECS，主要用于执行数据同步或其他任务。自定义资源组可以有效避免默认调度资源组的限制，当前新建自定义资源组功能需要您提交工单申请，已有的自定义资源组不受影响。 • 独享资源 DataWorks提供独享调度、独享数据集成和DataStudio运行空间（生产环境）资源供您选择，以获取最优性能保障。详情请参见DataWorks独享资源。
	项目设置	在项目开发过程中，部分项目的设置操作需要项目Owner来执行。例如，设置项目是否允许全表扫描、设置项目默认打开2.0新类型等。详情请参见项目操作。

模块	子模块	说明
安全管理	人员管理 角色管理 权限管理	安全管理包括人员管理、角色管理、权限管理等。通过DataWorks使用MaxCompute时，由于DataWorks和MaxCompute有各种权限模型，因此您需要理清楚两个产品之间的权限关系，再从业务需求出发进行权限管理。安全管理过程中，您需要了解如何进行用户授权、跨项目的资源共享、设置项目的数据保护功能、Policy授权等操作： • 权限管理和安全参数列表。 • 安全管理案例：提供从真实客户业务需求转化的安全案例。
费用管理	无	资源预算是在使用之前进行成本预估。基于MaxCompute的计费方式，很多业务无法更准确地预估成本，因此在整个业务开发过程中需要进行费用管理，主要需要关注： • 产品的计费定价，详情请参见计费方式。 • 产品欠费预警与停机策略，详情请参见欠费预警与停机策略。 • 包年包月升级和降配操作，详情请参见升级和降配。 • 包年包月续费，详情请参见续费管理。 • 产品支持按量计费和包年包月转换，详情请参见计费方式转换。 • 查看MaxCompute账单信息，详情请参见账单查看和MaxCompute账单分析最佳实践。

4.使用限制

在使用MaxCompute前，建议您先了解产品相关使用限制，确保业务可顺利开展。本文为您介绍使用MaxCompute过程中的操作限制。

数据上传下载限制

在MaxCompute中上传下载数据时的使用限制如下：

更多数据上传下载信息，请参见数据上传下载。

SQL限制

在MaxCompute中开发SQL作业时的使用限制如下。

限制项	最大值/限制条件	分类	说明
表名长度	128字节	长度限制	表名、列名中不能有特殊字符，以字母开头，且只能用英文小写字母（a-z）、英文大写字母（A-Z）、数字和下划线（_）。
注释长度	1024字节	长度限制	长度不超过1024字节的有效字符串。
表的列定义	1200个	数量限制	单表的列定义个数最多为1200个。
单表分区数	60000个	数量限制	单表的分区个数最多为60000个。
表的分区层级	6级	数量限制	在表中创建的分区层次不能超过6级。
屏显	10000行	数量限制	SELECT语句屏显最多输出10000行。
`INSERT` 目标个数	256个	数量限制	`MULTI-INSERT` 场景，目标表的数量限制为256个。
`UNION ALL`	256个	数量限制	`UNION ALL` 场景，最多允许合并256个表。
`MAPJOIN`	128个	数量限制	`MAPJOIN` 场景，最多允许连接128个小表。
`MAPJOIN` 内存限制	512 MB	数量限制	`MAPJOIN` 场景，所有小表的内存不能超过512 MB。
`ptinsubq`	1000行	数量限制	子查询中存在分区列时，子查询的返回结果不能超过1000行。
SQL语句长度	2 MB	长度限制	SQL语句的最大长度为2 MB。包括您使用SDK调用SQL的场景。
`WHERE` 子句条件个数	256个	数量限制	`WHERE` 子句中的条件个数最大为256个。
列记录长度	8 MB	数量限制	表中单个单元的最大长度为8 MB。

限制项	最大值/限制条件	分类	说明
IN的参数个数	1024	数量限制	IN的最大参数限制，例如 `in (1,2,3….,1024)` 。如果 `in(…)` 的参数过多，会影响编译性能。1024为建议值，不是限制值。
`jobconf.json`	1 MB	长度限制	`jobconf.json` 的大小为1 MB。当表包含的分区数量较多时，大小可能超过 `jobconf.json` ，超过1 MB。
视图	不可写	操作限制	视图不支持写入，不支持 `INSERT` 操作。
列的数据类型	不可修改	操作限制	不允许修改列的数据类型及列位置。
Java UDF函数	不允许为 `ABSTRACT` 或者 `STATIC` 。	操作限制	Java UDF函数不能为 `ABSTRACT` 或 `STATIC` 。
最多查询分区个数	10000个	数量限制	最多查询分区个数不能超过10000个。
SQL执行计划长度	1 MB	长度限制	MaxCompute SQL生成的执行计划不能超过1 MB，否则会触发 `FAILED: ODPS-0010000:System internal error - The Size of Plan is too large` 报错。

更多SQL信息，请参见SQL。

MapReduce限制

在MaxCompute中开发MapReduce作业时的使用限制如下。

边界名	边界值	分类	配置项名称	默认值	是否可配置	说明
Instance内存占用	[256 MB,12 GB]	内存限制	`odps.stage.mapper(reducer).mem` 和 `odps.stage.mapper(reducer).jvm.mem`	2048 MB + 1024 MB	是	单个Map Instance或Reduce Instance占用Memory，有框架Memory（默认2048 MB）和JVM的Heap Memory（默认1024 MB）两部分。
Resource数量	256个	数量限制	-	无	否	单个Job引用的Resource数量不超过256个，Table、Archive按照一个单位计算。
输入路数和输出路数	1024个和256个	数量限制	-	无	否	单个Job的输入路数不能超过1024（同一个表的一个分区算一路输入，总的不同表个数不能超过64个），单个Job的输出路数不能超过256。

边界名	边界值	分类	配置项名称	默认值	是否可配置	说明
Counter数量	64个	数量限制	-	无	否	单个Job中自定义Counter的数量不能超过64，Counter的Group Name和Counter Name中不能带有井号（#），两者长度和不能超过100。
Map Instance	[1,100000]	数量限制	odps.stage.mapper.num	无	是	单个Job的Map Instance个数由框架根据Split Size计算得出，如果没有输入表，可以通过odps.stage.mapper.num直接设置，最终个数范围[1,100000]。
Reduce Instance	[0,2000]	数量限制	odps.stage.reducer.num	无	是	单个Job默认Reduce Instance个数为Map Instance个数的1/4，用户设置作为最终的Reduce Instance个数，范围[0,2000]。可能出现这样的情形：Reduce处理的数据量会比Map大很多倍，导致Reduce阶段比较慢，而Reduce只能最多2000。
重试次数	3	数量限制	-	无	否	单个Map Instance或Reduce Instance失败重试次数为3，一些不可重试的异常会直接导致作业失败。
Local Debug模式	Instance个数不超100	数量限制	-	无	否	Local Debug模式下： • 默认Map Instance个数为2，不能超过100。 • 默认Reduce Instance个数为1，不能超过100。 • 默认一路输入下载记录数100，不能超过10000。
重复读取Resource次数	64次	数量限制	-	无	否	单个Map Instance或Reduce Instance重复读一个Resource次数限制<=64次。
Resource字节数	2 GB	长度限制	-	无	否	单个Job引用的Resource总计字节数大小不超过2 GB。
Split Size	大于等于1	长度限制	odps.stage.mapper.split.size	256 MB	是	框架会参考设置的Split Size值来划分Map，决定Map的个数。
STRING列内容长度	8 MB	长度限制	-	无	否	MaxCompute表STRING列内容长度不允许超出限制。

边界名	边界值	分类	配置项名称	默认值	是否可配置	说明
Worker运行超时时间	[1,3600]	时间限制	odps.function.timeout	600	是	Map或者Reduce Worker在无数据读写且没有通过 `context.progress()` 主动发送心跳的情况下的超时时间，默认值是600s。
MapReduce引用Table资源支持的字段类型	BIGINT、DOUBLE、STRING、DATETIME、BOOLEAN	数据类型限制	-	无	否	MapReduce任务引用表资源时，若表字段有其他类型字段执行报错。
MapReduce是否支持读取OSS数据	-	功能限制	-	无	否	MapReduce不支持读取OSS数据。
MapReduce是否支持MaxCompute 2.0新类型	-	功能限制	-	无	否	MapReduce不支持MaxCompute 2.0新类型。

更多MapReduce信息，请参见MapReduce。

PyODPS限制

在MaxCompute中基于DataWorks开发PyODPS作业时的使用限制如下：

- PyODPS节点获取本地处理的数据不能超过50 MB，节点运行时占用内存不能超过1 GB，否则节点任务会被系统中止。请避免在PyODPS任务中写额外的Python数据处理代码。
- 在DataWorks上编写代码并进行调试效率较低，为提升运行效率，建议本地安装IDE进行代码开发。
- 在DataWorks上使用PyODPS时，为了防止对DataWorks的Gate Way造成压力，对内存和CPU都有限制，该限制由DataWorks统一管理。如果您发现有**Got killed**报错，即表明内存使用超限，进程被中止。因此，请尽量避免本地的数据操作。通过PyODPS发起的SQL和DataFrame任务（除to_pandas外）不受此限制。
- 由于缺少matplotlib等包，如下功能可能受限：
  - DataFrame的plot函数。
  - DataFrame自定义函数需要提交到MaxCompute执行。由于Python沙箱限制，第三方库只支持所有的纯粹Python库以及Numpy，因此不能直接使用Pandas。
  - DataWorks中执行的非自定义函数代码可以使用平台预装的Numpy和Pandas。不支持其他带有二进制代码的三方包。
- 由于兼容性原因，在DataWorks中，options.tunnel.use_instance_tunnel默认设置为False。如果需要全局开启instance tunnel，需要手动将该值设置为True。
- 由于实现的原因，Python的atexit包不被支持，请使用try-finally结构实现相关功能。

更多PyODPS信息，请参见PyODPS。

Graph限制

在MaxCompute中开发Graph作业时的使用限制如下：

- 单个Job引用的Resource数量不超过256个，Table、Archive按照一个单位计算。
- 单个Job引用的Resource总计字节数大小不超过512 MB。
- 单个Job的输入路数不能超过1024（输入表的个数不能超过64）。单个Job的输出路数不能超过256。
- 多路输出中指定的Label不能为NULL或者空字符串，长度不能超过256个字符串，只能包括A-Z、a-z、0-9、下划线（_）、井号（#）、英文句点（.）和短划线（-）。
- 单个Job中自定义Counter的数量不能超过64个。Counter的 `group name` 和 `counter name` 中不能带有井号（#），两者长度和不能超过100。
- 单个Job的Worker数由框架计算得出，最大为1000个，超过抛异常。
- 单个Worker占用CPU默认为200个，范围为[50,800]。
- 单个Worker占用Memory默认为4096 MB，范围为[256 MB,12 GB]。
- 单个Worker重复读一个Resource次数限制不大于64次。
- `split_size` 默认为64 MB，您可自行设置，范围为0< `split_size` ≤（9223372036854775807>>20）。
- MaxCompute Graph程序中的GraphLoader、Vertex、Aggregator等在集群运行时，受到Java沙箱的限制（Graph作业的主程序则不受此限制），具体限制请参见Java沙箱。

更多Graph信息，请参见Graph。

其他限制

各地域下的单个MaxCompute项目支持同时提交的作业数量限制如下。

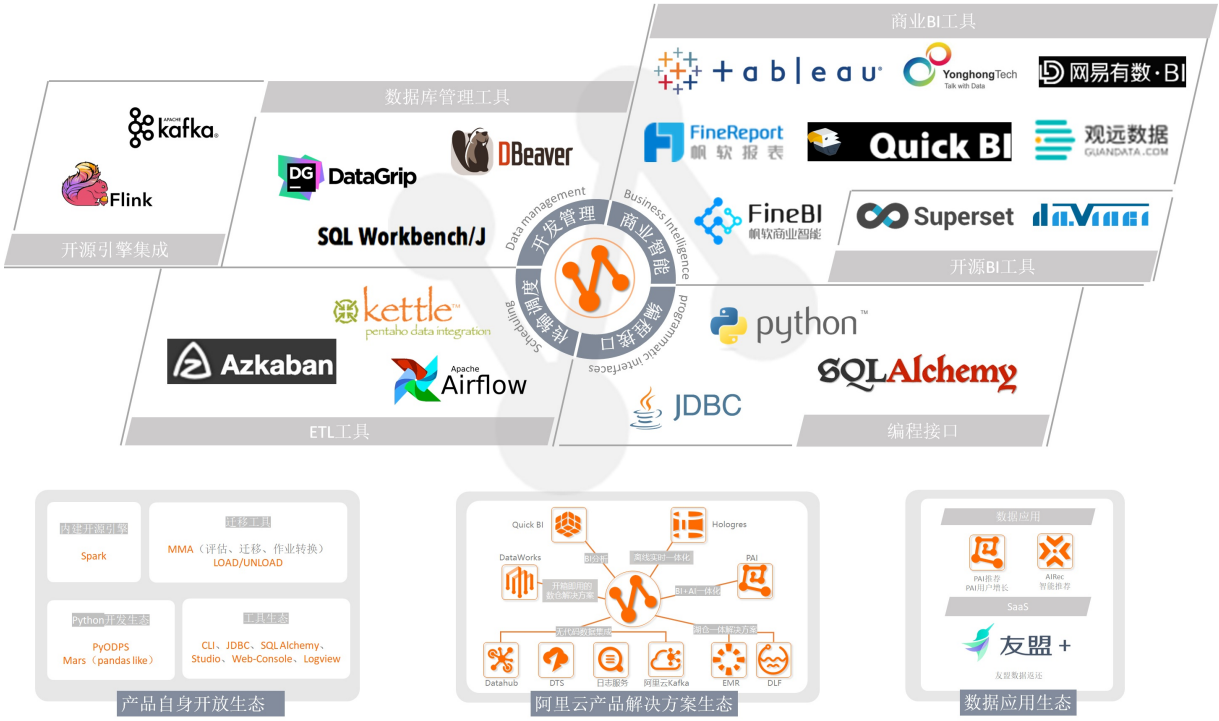
地域	单个MaxCompute项目作业并发上限
华东1（杭州）、华东2（上海）、华北2（北京）、华北3（张家口）、华南1（深圳）、西南1（成都）	2500
华东2金融云（上海）、华南1金融云（深圳）	200
华北2政务云（北京）、中国香港、新加坡、澳大利亚（悉尼）、马来西亚（吉隆坡）、印度尼西亚（雅加达）、日本（东京）、德国（法兰克福）、美国（硅谷）、美国（弗吉尼亚）、英国（伦敦）、印度（孟买）、阿联酋（迪拜）	300

如果MaxCompute项目提交的作业已经达到上限，继续提交作业会返回报错。错误信息示例： `com.aliyun.odps.OdpsException: Request rejected by flow control. You have exceeded the limit for the number of tasks you can run concurrently in this project. Please try later` 。

5.生态对接

本文为您介绍MaxCompute支持连接的商业智能BI工具、数据库管理工具及ETL工具。

MaxCompute的生态架构如下图所示。



商业智能（BI）工具

商业智能（BI）工具支持将计算引擎得到的数据通过仪表板、图表或其他图形输出实现数据可视化，以直观的形式展示给决策者，帮助高层管理者做出更明智的业务决策。

MaxCompute支持的BI工具如下。

- **商业BI工具**

工具	版本要求	接入方法	参考资源
Tableau	○ Tableau：Desktop 2019.4及以上版本 ○ MaxCompute：JDBC驱动v3.0.1及以上版本	Tableau连接MaxCompute	○ How to connect Tableau to Alibaba MaxCompute ○ Manage Data Sources
FineBI	○ FineBI：v5.1.9及以上版本 ○ MaxCompute：JDBC驱动v3.2.8及以上版本	FineBI连接MaxCompute	阿里云MaxCompute数据连接（FineBI）

工具	版本要求	接入方法	参考资源
FineReport	○ FineReport：v10.0及以上版本 ○ MaxCompute：JDBC驱动v3.2.8及以上版本	FineReport连接MaxCompute	阿里云MaxCompute数据连接（FineReport）
Yonghong BI	Yonghong Desktop：v8.6及以上版本	Yonghong BI连接MaxCompute	添加MaxCompute数据源
Quick BI	无特殊要求	Quick BI连接MaxCompute	云数据源MaxCompute
观远BI	无特殊要求	观远BI连接MaxCompute	观远BI
网易有数BI	无特殊要求	网易有数BI连接MaxCompute	网易有数BI

- **开源BI工具**

工具	版本要求	接入方法	参考资源
Davinci	○ Davinci：无特殊要求 ○ MaxCompute：JDBC驱动v3.2.8及以上版本	Davinci连接MaxCompute	Davinci数据源配置
Superset	○ Superset：无特殊要求 ○ PyODPS：v0.10.7及以上版本	Superset连接MaxCompute	Extra databases settings

数据库管理工具

数据库管理工具，即数据库图形化工具，是数据库人员必需的工具之一，根据这种工具，可以形象化、方便快捷地查询数据信息。

MaxCompute支持的数据库管理工具如下。

工具	版本要求	接入方法	参考资源
DBeaver	• DBeaver：无特殊要求 • MaxCompute：JDBC驱动v3.2.8及以上版本	DBeaver连接MaxCompute	Create Connection
DataGrip	• DataGrip：无特殊要求 • MaxCompute：JDBC驱动v3.2.8及以上版本	DataGrip连接MaxCompute	Database Connection

工具	版本要求	接入方法	参考资源
SQL Workbench/J	• SQL Workbench/J：无特殊要求 • MaxCompute：JDBC驱动v3.0.1及以上版本	SQL Workbench/J连接MaxCompute	JDBC驱动程序

ETL工具

ETL（Extract-Transform-Load）用来描述将数据从来源端经过抽取（Extract）、转换（Transform）、加载（Load）至目的端的过程。

MaxCompute支持的ETL工具如下。

工具	接入方法
Kettle	使用Kettle调度MaxCompute
Apache Airflow	使用Apache Airflow调度MaxCompute
Azkaban	使用Azkaban调度MaxCompute

6.支持的云服务

基于MaxCompute的数据仓库能力，您可以与阿里云其他产品集成，实现可视化开发、数据存储、数据迁移、机器学习、业务决策等能力，构建满足实际业务需求的解决方案。本文为您介绍支持与MaxCompute集成的各阿里云产品信息。

MaxCompute支持集成的阿里云产品如下。

阿里云产品	说明
DataWorks	DataWorks是基于MaxCompute计算和存储，提供工作流可视化开发、调度运维托管的一站式海量数据离线加工分析平台。您可以将DataWorks理解为MaxCompute的一种Web客户端，MaxCompute是DataWorks的一种计算引擎。 MaxCompute和DataWorks提供完善的ETL、数据分析、数据地图、数据治理和数据仓库管理能力，并支持SQL、MapReduce、Graph等多种经典的分布式计算模型，能够更快速地解决用户海量数据计算问题，有效降低企业成本，保障数据安全。 更多DataWorks信息，请参见DataWorks。
数据集成	MaxCompute可以通过数据集成功能加载不同数据源（例如MySQL数据库）的数据，也可以通过数据集成把MaxCompute的数据导出到各种业务数据库。 数据集成功能已集成在DataWorks上，您可以直接在DataWorks上配置MaxCompute数据源并读写MaxCompute表。更多操作请参见配置MaxCompute数据源、读取MaxCompute表和写入MaxCompute表。 更多数据集成信息，请参见数据集成。
机器学习PAI	机器学习PAI是基于MaxCompute的一款机器学习算法平台，实现了数据无需搬迁，便可进行从数据处理、模型训练、服务部署到预测的一站式机器学习。您创建MaxCompute项目并开通机器学习服务后，即可通过机器学习平台的算法组件对MaxCompute数据进行模型训练等操作。 更多机器学习PAI信息，请参见机器学习PAI。
Quick BI	Quick BI是一个专为云上用户量身打造的新一代智能BI服务平台。在MaxCompute上对数据进行加工处理后，您可以将MaxCompute项目添加为Quick BI数据源，即可在Quick BI页面制作报表，对MaxCompute表数据进行可视化分析。 更多Quick BI信息，请参见Quick BI。

阿里云产品	说明
AnalyticDB MySQL版	AnalyticDB MySQL版是海量数据实时高并发在线分析（Realtime OLAP）云计算服务，与MaxCompute结合应用于大数据驱动业务系统的场景。通过MaxCompute离线计算挖掘，产出高质量数据后，导入分析型数据库，供业务系统调用分析。 将MaxCompute数据导入AnalyticDB MySQL版，有如下两种方式： • 通过DMS for AnalyticDB for MySQL的通过外表将MaxCompute数据导入至AnalyticDB MySQL和通过外表导出AnalyticDB MySQL数据至MaxCompute功能进行配置。 • 通过DataWorks配置数据同步任务，请参见通过DataWorks同步数据。 更多AnalyticDB MySQL版信息，请参见云原生数据仓库AnalyticDB MySQL版。
表格存储	表格存储是构建在阿里云飞天分布式系统之上的分布式NoSQL数据存储服务，MaxCompute 2.0支持直接通过外部表方式访问表格存储中的表数据并进行处理，详情请参见OTS外部表。 更多表格存储信息，请参见表格存储。
对象存储OSS	对象存储OSS是海量、安全、低成本、高可靠的云存储服务，MaxCompute 2.0支持直接通过外部表方式访问对象存储中的表数据并进行处理，详情请参见OSS外部表。 更多OSS信息，请参见对象存储OSS。
开放搜索OpenSearch	开放搜索OpenSearch是一款阿里云自主研发的大规模分布式搜索引擎平台。您通过MaxCompute对数据进行计算处理后，可以在OpenSearch平台上通过添加数据源的方式将MaxCompute数据接入，详情请参见MaxCompute数据源配置。 更多OpenSearch信息，请参见开放搜索OpenSearch。
移动数据分析Quick A+ Digital Analytics	移动数据分析Quick A+ Digital Analytics是阿里云推出的一款移动App数据统计分析产品，为开发者提供一站式数据化运营服务。当移动数据分析自带的基础分析报表不能满足App开发者的个性化需求时，App开发者可以将数据一键同步至MaxCompute，结合自己的业务需求来进一步加工、分析数据。 更多Quick A+ Digital Analytics信息，请参见移动数据分析Quick A+ Digital Analytics。
日志服务SLS	日志服务SLS能快速完成日志类数据采集、消费、投递以及查询分析等操作。日志数据采集后，需要更多的个性化分析、挖掘，您可以通过DataWorks的数据集成将日志服务数据同步到MaxCompute，通过MaxCompute对日志数据进行个性化、深层次的数据分析、挖掘。 更多SLS信息，请参见日志服务SLS。

阿里云产品	说明
访问控制RAM	RAM是阿里云为客户提供的用户身份管理与资源访问控制服务。MaxCompute与RAM的集成使用主要有两个场景： • 通过DataWorks使用MaxCompute时，管理RAM用户。 阿里云账号开通并创建MaxCompute项目后，若需要通过DataWorks使用MaxCompute且多个账户协同开发，必须由阿里云账号到RAM服务中创建RAM用户，将RAM用户添加为项目成员从而进行协同开发，详情请参见准备RAM用户和添加工作空间成员和角色。 • MaxCompute处理非结构化数据时，通过RAM对非结构化数据进行授权。 MaxCompute支持直接处理非结构化数据（包含OSS和表格存储），但是需要提前在RAM中授予MaxCompute访问OSS或表格存储的权限，详情请参见OSS外部表和OTS外部表。

阿里云产品支持的字符集

在同时使用MaxCompute和阿里云其他产品过程中，需要关注字符集格式，确保满足MaxCompute字符集格式要求，避免因使用了不支持的字符集导致产品无法正常使用。

产品名称	支持的字符集
MaxCompute	UTF-8
DataWorks	在DataStudio中进行数据上传，支持UTF-8、GBK、CP936、ISO-8859，但在DataWorks中会统一为UTF-8。数据下载支持UTF-8、GBK。
数据集成	
机器学习PAI	UTF-8
Quick BI	UTF-8
AnalyticDB MySQL版	UTF-8
表格存储	UTF-8
对象存储OSS	UTF-8
开放搜索OpenSearch	UTF-8
移动数据分析Quick A+ Digital Analytics	UTF-8
日志服务SLS	UTF-8
访问控制RAM	不涉及

7.计算模型的开通地域

本文为您介绍MaxCompute支持的每种计算模型在各地域的开通情况。

地域	SQL	MapReduce	Hologres	Spark	PyODPS	Mars
华北2（北京）	✓	✓	✓	✓	✓	✓
华东1（杭州）	✓	✓	✓	✓	✓	✓
华东2（上海）	✓	✓	✓	✓	✓	✓
华南1（深圳）	✓	✓	✓	✓	✓	✓
西南1（成都）	✓	✓	✗	✓	✓	✓
华北3（张家口）	✓	✓	✓	✓	✓	✓
中国（香港）	✓	✓	✓	✓	✓	✓
新加坡（新加坡）	✓	✓	✓	✓	✓	✗
马来西亚（吉隆坡）	✓	✓	✓	✓	✓	✗
印度尼西亚（雅加达）	✓	✓	✓	✓	✓	✗
澳大利亚（悉尼）	✓	✓	✗	✓	✓	✗
日本（东京）	✓	✓	✓	✓	✓	✗
美国（硅谷）	✓	✓	✓	✓	✓	✗
美国（弗吉尼亚）	✓	✓	✗	✓	✓	✗
英国（伦敦）	✓	✓	✗	✓	✓	✗
德国（法兰克福）	✓	✓	✓	✓	✓	✗

地域	SQL	MapReduce	Hologres	Spark	PyODPS	Mars
印度（孟买）	✅	✅	✅	✅	✅	❌
阿联酋（迪拜）	✅	✅	❌	✅	✅	❌

> **说明** ✅表示已开通，❌表示未开通。

8.客户案例

MaxCompute已被广泛应用于各大领域处理云上大数据，帮助众多企业解决了海量数据分析问题，同时降低企业运维成本，企业人员可更专注于业务开发。本文为您介绍MaxCompute的精选客户案例。

MaxCompute的全量客户案例信息，请参见行业客户案例。

友盟+

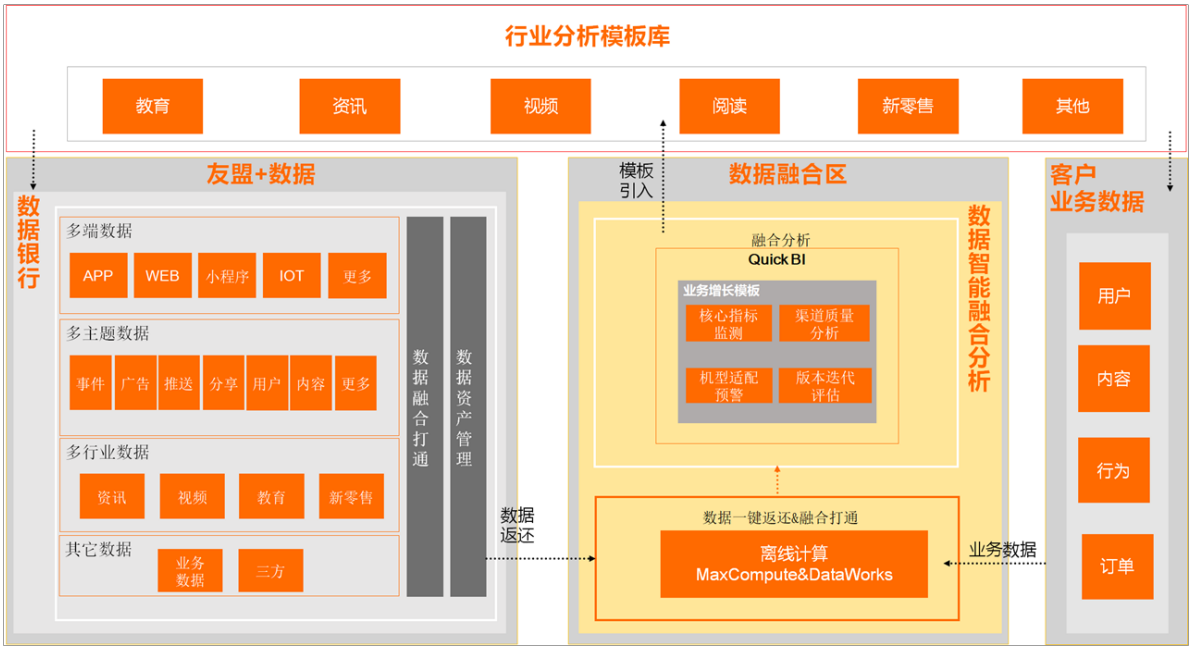
- **客户简介**

  友盟+是独立的第三方全域数据智能服务商，基于技术与算法能力，结合全域数据资源，挖掘标签及分析指标，帮助企业实现深度用户洞察、实时业务决策和持续业务增长。

- **客户诉求**
  - 帮助企业和开发者解决数据系统独立，无法融合分析的问题。
  - 帮助企业和开发者解决BI分析系统灵活性与业务可用性难以平衡的问题。

- **解决方案**

  友盟+联合MaxCompute构建开发者数据银行，为企业提供面向分析的、实现友盟域数据与企业私域数据全面融合的自助分析服务“U-DOP数据开放”。该服务通过订阅数据包返还数据到MaxCompute，预置分析模板并结合可视化分析BI工具来快速完成数据分析工作，为企业提供更加灵活的一站式数据分析能力。解决方案架构如下。



详细案例信息，请参见友盟+案例。

电商案例：玩物得志

- **客户简介**

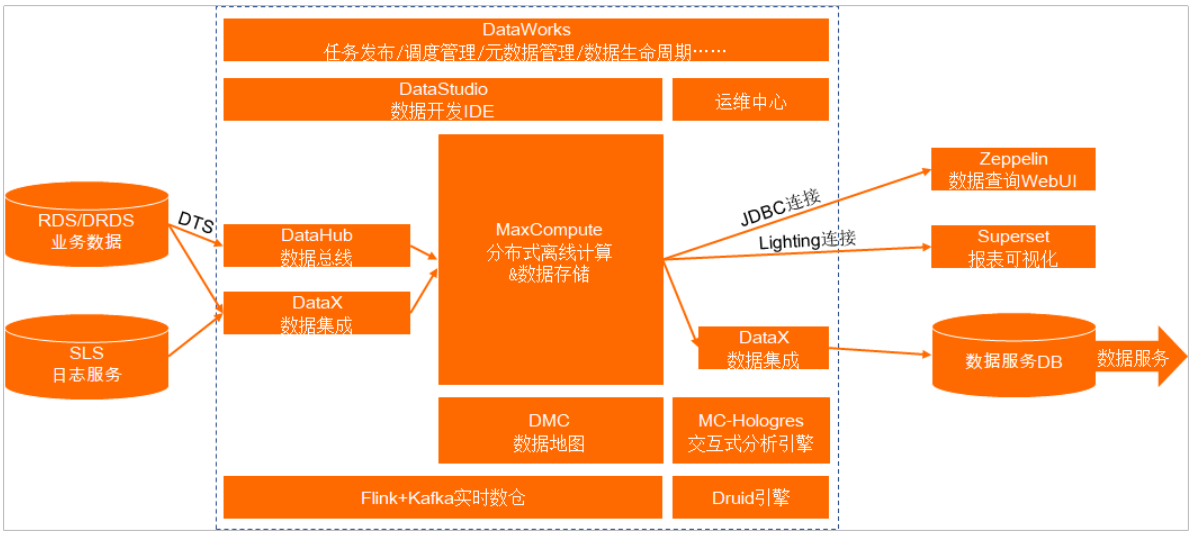
  玩物得志是一家涵盖文玩工艺品交易、行家在线鉴宝和国风生活社区的综合移动互联网平台，平台月活跃用户数高达600万。提供直播、拍卖和一口价三大核心服务。通过“先鉴别，后发货”的服务模式，为用户打造透明、健康、安全的文化消费环境，让普通消费者敢于购买国风好物，感受国风文化的魅力。

- **客户诉求**

随着电商业务的快速发展，期望能采用云平台提供的SaaS、PaaS服务搭建研发系统，助力提升开发效率，同时还要节省人力成本，并能高效地将原MySQL体系迁移上云。

- **解决方案**

  玩物得志基于阿里云DataWorks+MaxCompute框架搭建大数据平台，使用其核心存储、计算等组件、上层可视化及业务查询能力，在开源方案的基础上进行了二次开发。解决方案架构如下。



详细案例信息，请参见玩物得志案例。

互联网社交案例：小打卡

- **客户简介**

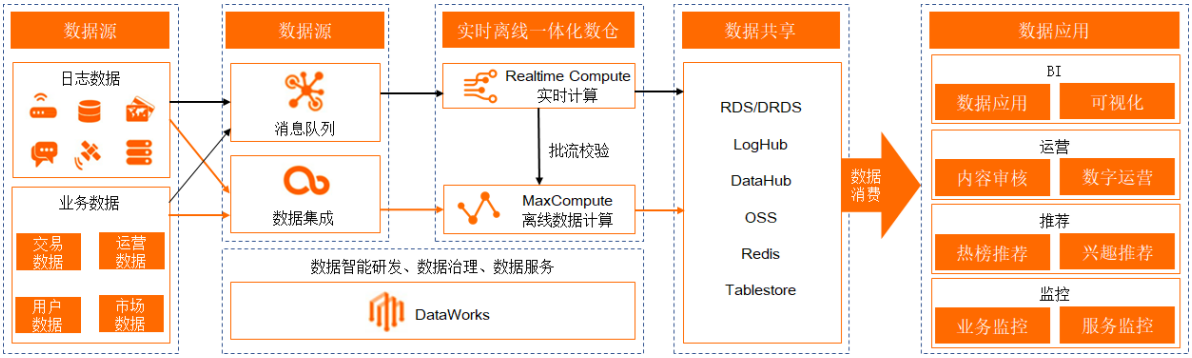
  小打卡是一个兴趣社区程序，聚集绘画、瑜伽、健身、摄影、亲子、阅读、潮玩等兴趣圈，可以帮助用户快速发现感兴趣的圈子，并一起分享、交流、成长。小打卡每日会有数百万活跃用户，可以产生TB级的数据。

- **客户诉求**

  在满足低费用成本、低运维成本的基础上构建具备敏捷开发、高扩展性的数据仓库。

- **解决方案**

  基于阿里云MaxCompute构建数据仓库，在开发人员成本、软硬件成本上具有明显优势。从初期至今，基于MaxCompute构建的数据仓库有极高的消费比。初期只有一个开发人员的情况下，可以快速地搭建数据仓库，且费用成本极低。解决方案架构如下。



详细案例信息，请参见小打卡案例。

互联网金融案例：天弘基金

- **客户简介**

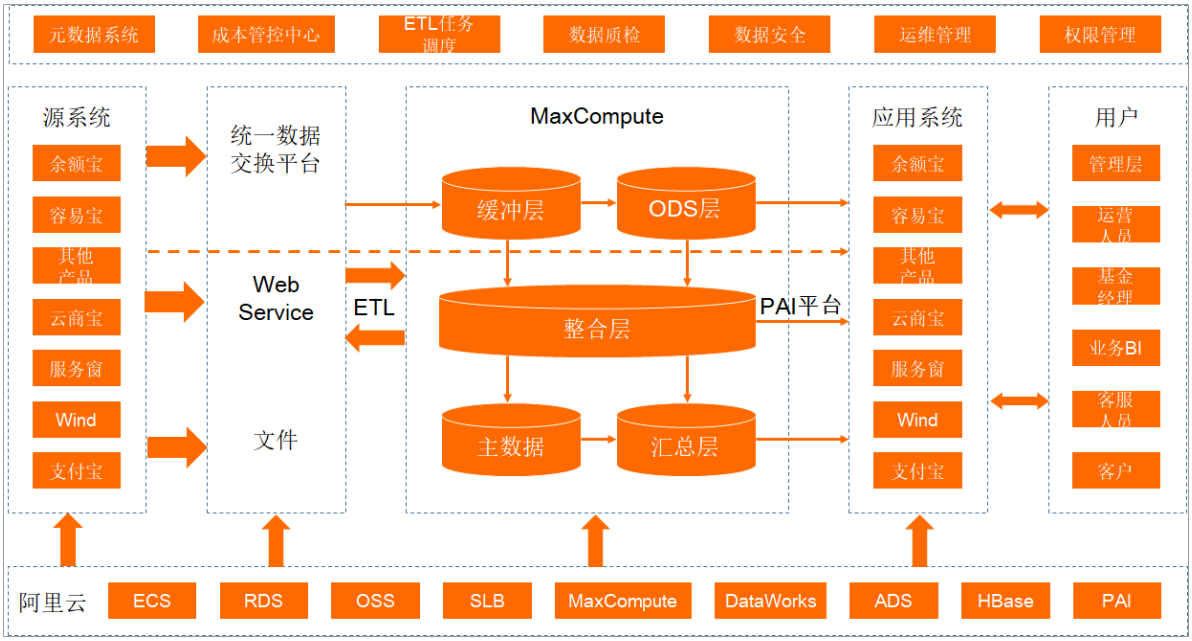
  天弘基金是中国总规模最大的公募基金，以变成最大的指数基金服务商为目标，基于客户底层需求进行创新，以指数基金为底层工具，为客户提供全方面、一站式的服务。天弘基金与支付宝合作推出了余额宝。

- **客户诉求**

  在余额宝用户数持续呈指数级增长，数据量也成倍增长的情况下，已经无法通过简单的Hadoop集群管理数据，同时业务端需要通过数据了解用户、分析行为进而对业务决策和用户行为进行精准预测。

- **解决方案**

  天弘基金基于阿里云MaxCompute构建了企业级一站式大数据解决方案。MaxCompute对于海量数据的存储、运维、计算能力强大且安全稳定。MaxCompute服务将原本需要清算8小时的用户交易数据缩短至1.5小时，同时减少了本地服务器部署压力，在显著提升工作效率的同时减少了大量开发成本和人力成本，使开发人员能更专注于业务发展，为用户提供高品质、高价值的金融服务。解决方案架构如下。



详细案例信息，请参见天弘基金案例。

智慧物流案例：千寻位置

- **客户简介**

  千寻位置是全球领先的精准位置服务公司，提供高达动态厘米级和静态毫米级的定位能力，是IoT时代重要的基础设施之一。千寻位置基于北斗卫星系统（兼容GPS、GLONASS、Galileo）定位数据，利用遍及全国的超过2400个地基增强站及自主研发的定位算法，通过互联网技术进行大数据运算，为遍布全国的用户提供精准定位及延展服务。

- **客户诉求**

  提升计算精准度及速度，满足用户基于实时精准位置的多种应用需求。

- **解决方案**

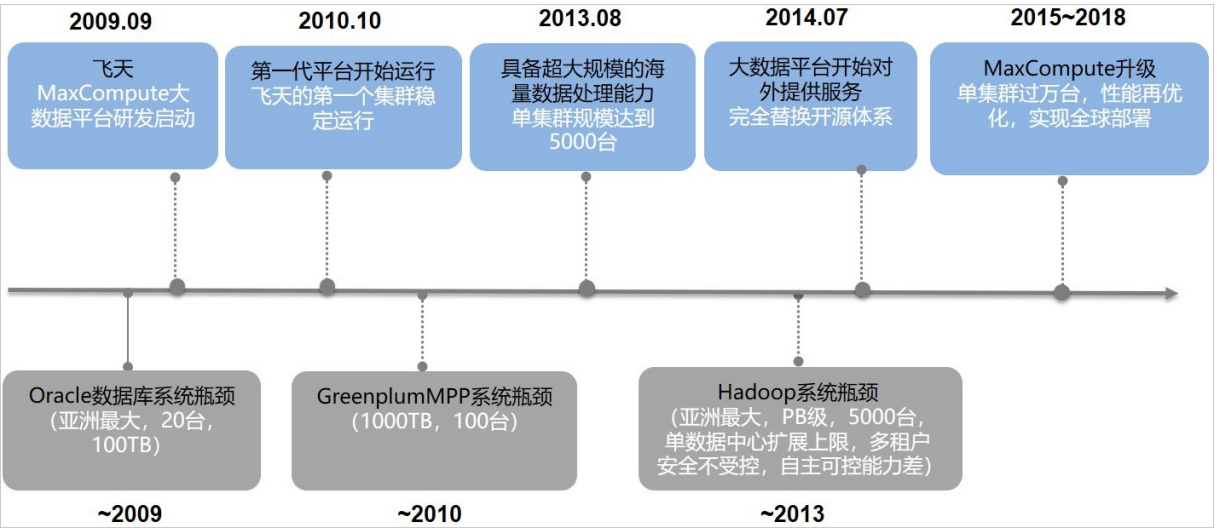
  千寻位置全程使用阿里云方案不再建设自己的集群。在混合云架构下，机密数据在专有云内完成，云端的大规模数据的计算则通过MaxCompute完成，定位数据的播发在公共云上完成。

详细案例信息，请参见千寻位置案例。

9.发展历程

本文为您介绍了MaxCompute从诞生到成熟的发展历程。

关键性里程碑



- 2009年9月，ODPS（即现在的MaxCompute）大数据平台飞天项目正式启动。
- 2010年10月，阿里巴巴集团自主研发的第一代云计算平台稳定运行。
- 2013年8月，平台的单集群规模已达到5000台。
- 2014年7月，平台开始对外提供服务，完全替换开源体系。
- 2015~2018年，平台开始日趋成熟，ODPS更名为MaxCompute。单集群已过万台，性能再优化，实现了全球部署。

产品荣誉

- 2018年11月，MaxCompute，DataWorks和AnalyticDB代表阿里云入选Forrester Wave™ Q4 2018云数据仓库报告。
- 2018年9月，基于公共云的BigBench在100 TB规模上，MaxCompute的性能指标较2017年10月提升了一倍，达到18176.71 QPM（Queries Per Minute）。此外，在超小型10 TB规模的指标上，MaxCompute的性能是其他开源竞品性能的3倍。
- 2018年4月，MaxCompute的多个客户案例荣获“2017大数据优秀产品和应用解决方案案例”奖。
- 2018年3月，MaxCompute登上Forrester *2018年一季度云端数据仓库*大数据服务榜单。
- 2018年3月，Gartner发布了*2017年分析型数据管理解决方案（DMSA）魔力象限*报告，阿里云作为云服务商成功冲进Gartner魔力象限。
- 2017年10月，TPC的benchmark适配MaxCompute，进行了全球首次基于公共云的BigBench大数据基准测试，数据规模拓展到100 TB，成为首个突破7000分的引擎，性能达到7830 QPM。
- 2017年6月， MaxCompute获得中国国际软件博览会金奖。
- 2016年11月，在CloudSort竞赛中，MaxCompute以$0.82/TB的成绩获得Indy（专用目的排序）和Daytona（通用目的排序）两个子项的世界冠军，打破了AWS（Amazon Web Services）在2014年保持的纪录$4.51/TB。
- 2015年10月，在GraySort竞赛中，MaxCompute用377秒完成了100 TB的数据排序，打破了此前Apache Spark创造的1406秒纪录。

产品认证

- 中国大陆首家工信部单集群万台扩展能力认证。
- 工信部信通院和中电标准化研究院认证。
- MaxCompute通过了独立的第三方审计师针对阿里云对AICPA可信服务标准中关于安全性、可用性和机密性原则符合性描述的审计。审计报告请参见SOC 3报告。

深度参与和推动全球大数据领域标准化建设

- MaxCompute代表阿里巴巴计算平台，成为国际TPC（Transaction Processing Performance Council）委员会大数据评测标准BigBench的委员会委员，是中国担任此国际性能标准化测试组织委员的唯一企业。
- 全球两大热门计算存储标准化开源体系ORC（Optimized Row Columnar）社区的PMC（Production Material Control），MaxCompute成为近两年贡献代码量最多的贡献者，引导存储标准化。
- MaxCompute积极投入全球热门的优化器项目Calcite，拥有一个专委席位，是中国大陆前两家具备该领域影响力的公司。

10.常见问题

本文列举了MaxCompute的用户经常咨询和关注的一些问题，帮助您快速了解MaxCompute。

MaxCompute的用户经常咨询和关注的一些问题如下：

- 使用MaxCompute需要具备什么专业技能？
- 如何理解开源与云原生的大数据技术与产品？
- MaxCompute作为大数据平台，对业务数据是否有好的监控手段？
- MaxCompute的项目发挥什么作用？
- 如何获取MaxCompute中的Accesskey_ID和AccessKey_Secret？
- 现有账号的AccessKey被禁用，创建一个新的AccessKey，会对之前AccessKey创建的周期性任务有影响吗？
- MaxCompute建表默认有压缩功能吗？可以指定压缩格式和存储格式吗？
- MaxCompute的表格类型有几种，分别是什么？
- 如果想使用MaxCompute的自定义函数（UDF）或MapReduce功能需要依赖什么资源来完成？
- MaxCompute常见错误信息如何理解，怎么定位问题？

使用MaxCompute需要具备什么专业技能？

MaxCompute支持多种计算模型数据通道，满足多场景需求。所以您只需要会使用SQL、Python、Java等开发语言就可以使用MaxCompute进行数据分析。

如何理解开源与云原生的大数据技术与产品？

推荐您阅读从开源到云原生，你不得不知的大数据实战。

MaxCompute作为大数据平台，对业务数据是否有好的监控手段？

MaxCompute仅支持通过DataWorks的数据质量功能配置数据监控规则。无法监控外部数据源的字段变化。

MaxCompute的项目发挥什么作用？

项目（Project）是MaxCompute的基本组织单元，类似于传统数据库的Database或Schema的概念，是进行多用户隔离和访问控制的主要边界。项目中包含多个对象，例如表（Table）、资源（Resource）、函数（Function）和实例（Instance）等。一个用户可以同时拥有多个项目的权限。通过安全授权，可以在一个项目访问另一个项目中的对象。

如何获取MaxCompute中的Accesskey_ID和AccessKey_Secret？

您可以进入AccessKey管理页面，创建或查询AccessKey。

现有账号的AccessKey被禁用，创建一个新的AccessKey，会对之前AccessKey创建的周期性任务有影响吗？

有影响。如果AccessKey被禁用或删除，将直接影响您的DataWorks中各类任务的正常运行。请谨慎操作。

MaxCompute建表默认有压缩功能吗？可以指定压缩格式和存储格式吗？

MaxCompute默认自动压缩3~5倍，默认存储格式为AliORC，不支持自定义。

MaxCompute的表格类型有几种，分别是什么？

MaxCompute的表格有两种类型：内部表和外部表（MaxCompute 2.0版本开始支持外部表）。

- 对于内部表，所有的数据都存储在MaxCompute中，表中列的数据类型可以是MaxCompute支持的任意一种数据类型。
- 对于外部表，MaxCompute并不真正持有数据，表格的数据可以存放在OSS或OTS中。MaxCompute仅会记录表格的Meta信息，您可以通过MaxCompute的外部表机制处理OSS或OTS上的非结构化数据，例如视频、音频、基因、气象、地理信息等。

如果想使用MaxCompute的自定义函数（UDF）或MapReduce功能需要依赖什么资源来完成？

- UDF：您编写UDF后，需要将编译好的JAR包以资源的形式上传到MaxCompute。运行此UDF时，MaxCompute会自动下载这个JAR包，获取您的代码来运行UDF，无需您干预。上传JAR包的过程就是在MaxCompute上创建资源的过程，JAR包是MaxCompute的一种资源。
- MapReduce：您编写MapReduce程序后，需要将编译好的JAR包作为一种资源上传到MaxCompute。运行MapReduce作业时，MapReduce框架会自动下载这个JAR包，获取您的代码。

您同样可以将文本文件以及MaxCompute中的表作为不同类型的资源上传到MaxCompute，您可以在UDF及MapReduce的运行过程中读取、使用这些资源。

MaxCompute常见错误信息如何理解，怎么定位问题？

MaxCompute的常见报错信息编号有规范定义，格式为：`异常编号：通用描述 - 上下文相关说明`。其中SQL、MapReduce、Tunnel的错误信息是不一样的。更多错误信息，请参见错误码概述。