MaxCompute

产品简介

为了无法计算的价值 | 阿里云

产品简介

大数据计算服务（MaxCompute，原名 ODPS）是一种快速、完全托管的 GB/TB/PB 级数据仓库解决方案。MaxCompute 为您提供了完善的数据导入方案以及多种经典的分布式计算模型，能够更快速的解决海量数据计算问题，有效降低企业成本，并保障数据安全。

同时，大数据开发套件和 MaxCompute 关系紧密，大数据开发套件为 MaxCompute 提供了一站式的数据同步，任务开发，数据工作流开发，数据管理和数据运维等功能，详情请参见 大数据开发套件。

MaxCompute 主要服务于批量结构化数据的存储和计算，可以提供海量数据仓库的解决方案以及针对大数据的分析建模服务。随着社会数据收集手段的不断丰富及完善，越来越多的行业数据被积累下来。数据规模已经增长到了传统软件行业无法承载的海量数据（百 GB、TB 乃至 PB）级别。

在分析海量数据场景下，由于单台服务器的处理能力限制，数据分析者通常采用分布式计算模式。但分布式的计算模型对数据分析人员提出了较高的要求，且不易维护。使用分布式模型，数据分析人员不仅需要了解业务需求，同时还需要熟悉底层计算模型。MaxCompute 的目的是为您提供一种便捷的分析处理海量数据的手段，您可以不必关心分布式计算细节，便可达到分析大数据的目的。

> MaxCompute 已经在阿里巴巴集团内部得到大规模应用，例如：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。

产品优势

大规模计算存储

MaxCompute 适用于 100GB 以上规模的存储及计算需求，最大可达 EB 级别。

多种计算模型

MaxCompute 支持 SQL、MapReduce、Graph 等计算类型及 MPI 迭代类算法。

强数据安全

MaxCompute 已稳定支撑阿里全部离线分析业务7年以上，提供多层沙箱防护及监控。

低成本

与企业自建私有云相比，MaxCompute 的计算存储更高效，可以降低 20%-30% 的采购成本。

功能概述

数据通道

支持批量、历史数据通道

TUNNEL 是 MaxCompute 为您提供的数据传输服务，提供高并发的离线数据上传下载服务。支持每天 TB/PB 级别的数据导入导出，特别适合于全量数据或历史数据的批量导入。Tunnel 提供 Java 编程接口供您使用，并且在 MaxCompute 的客户端工具中，有对应的命令实现本地文件与服务数据的互通。

实时、增量数据通道

针对实时数据上传的场景，MaxCompute 提供了延迟低、使用方便的 DataHub 服务，特别适用于增量数据的导入。Datahub 还支持多种数据传输插件，例如：Logstash、Flume、Fluentd、Sqoop 等，同时支持日志服务 Log Service 中的 日志数据一键投递至 MaxCompute，进而使用大数据开发套件进行日志分析和挖掘。

计算及分析任务

MaxCompute 支持多种计算模型，详情如下：

SQL：MaxCompute 只能以表的形式存储数据，并对外提供了 SQL 查询功能。您可以将 MaxCompute 作为传统的数据库软件操作，但其却能处理 TB、PB 级别的海量数据。

注意：

- MaxCompute SQL 不支持事务、索引及 Update/Delete 等操作。
- 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 语句无缝迁移到 MaxCompute 上来。
- 在使用方式上，MaxCompute SQL 最快可以在分钟，乃至秒级别完成查询，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。

UDF：即用户自定义函数。

MaxCompute 提供了很多 内建函数 来满足您的计算需求，同时您还可以通过创建自定义函数来满足不同的计算需求。

MapReduce：MaxCompute MapReduce 是 MaxCompute 提供的 Java MapReduce 编程模型

，它虽与通用的 MapReduce 有所区别，但可以简化开发流程，更为高效。您若使用 MaxCompute MapReduce，需要对分布式计算概念有基本了解，并有相对应的编程经验。MaxCompute MapReduce 为您提供 Java 编程接口。

Graph：MaxCompute 提供的 Graph 功能是一套面向迭代的图计算处理框架。图计算作业使用图进行建模，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，最终求解出结果，典型应用：PageRank，单源最短距离算法 ，K-均值聚类算法 等。

SDK

SDK 是 MaxCompute 提供给开发者的工具包，详情请参见 SDK 介绍。

安全

MaxCompute 提供了功能强大的安全服务，为您的数据安全提供保护，详情请参见 安全参考手册。

后续步骤

现在，您已经学习了 MaxCompute 的产品优势、功能特性等相关简介，您可以继续学习下一个教程。在该教程中您将快速了解如何使用 MaxCompute，详情请参见 快速开始。

从 2009 年 9 月阿里云成立，愿景就是做运算/分享数据的第一平台。2010 年 4 月，伴随阿里金融的贷款业务上线，ODPS 正式投入生产运行，2012 年建立统一数据平台，2013 年具备超大规模海量数据处理能力，2014~2015 年大数据平台开始日趋成熟，2016 年 MaxCompute 2.0 诞生，成立之初的愿景正在逐步实现。

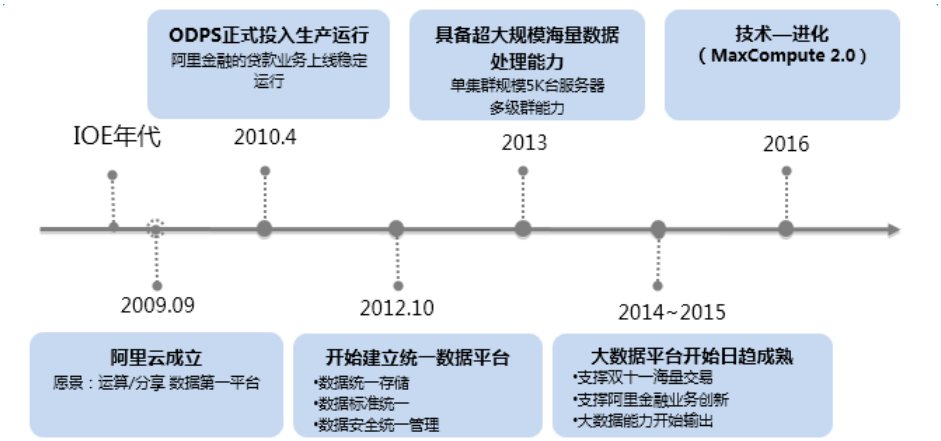
关键性里程碑

2010.04 ODPS 正式投入生产运行。阿里金融的贷款业务上线稳定运行。

2013.05 ODPS 公测。

2013.07 ODPS 正式提供商业化服务，单集群规模 5K 台服务器多级群能力。

2016.09 ODPS 正式更名为 MaxCompute，并推出 MaxCompute 2.0，实现高性能，新功能，富生态。



基本概念

A

AccessKey

AccessKey（简称 AK，包括 Access Key Id 和 Access Key Secret），是访问阿里云 API 的密钥，在阿里云官网注册云账号后，可在 accesskeys 管理 页面生成，用于标识用户，为访问 MaxCompute 或者其他云产品做签名验证。**Access Key Secret 必须保密。**

安全

MaxCompute 多租户数据安全体系，主要包括：用户认证、项目空间的用户与授权管理、跨项目空间的资源分享以及项目空间的数据保护。关于 MaxCompute 安全操作的更多详情请参见 **安全指南**。

B

Table（表）

表是 MaxCompute 的数据存储单元，详情请参见 基本概念 > 表。

C

Console

MaxCompute 运行在 Window/Linux 下的客户端工具，通过 Console 可以提交命令完成 Project 管理、DDL、DML 等操作。对应的工具安装和常用参数请参见 客户端。

D

Data type

MaxCompute 表中所有列对应的数据类型。目前支持的数据类型详情请参见 基本概念 > 数据类型。

DDL

Data Definition Language（数据定义语言）。比如创建表、创建视图等操作，MaxCompute DDL 语法请参见 用户指南 > DDL 语句。

DML

Data Manipulation Language（数据操作语言）。比如 INSERT 操作，MaxCompute DML 语法请参见 用户指南 > INSERT 操作。

F

Partition（分区）

分区 Partition 是指一张表下，根据分区字段（一个或多个组合）对数据存储进行划分。也就是说，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 Partition，每个 Partition 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。关于分区的更多介绍请参见 基本概念 > 分区。

fuxi

伏羲（fuxi）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 底层任务调度模块即 fuxi 的调度模块。

J

Role（角色）

角色是 MaxCompute 安全功能里使用的概念，可以看成是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。关于角色管理的更多介绍请参见 **安全指南** > **角色管理**。

M

MapReduce

MaxCompute 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 **用户指南** > MapReduce。

O

ODPS

ODPS 是 MaxCompute 的原名。

P

Project（项目）

项目空间（Project）是 MaxCompute 的基本组织单元，它类似于传统数据库的 Database 或 Scheme 的概念，是进行多用户隔离和访问控制的主要边界。详情请参见 **基本概念** > **项目空间**。

S

SDK

Software Development Kits 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 目前支持 JAVA SDK 和 Python SDK。

授权

项目空间管理员或者 project owner 授予您对 MaxCompute 中的 Object（或称之为客体，例如：表，任务，资源等）进行某种操作的权限，包括：读、写、查看等。授权的具体操作请参见 安全指南 > 用户及授权管理。

沙箱

MaxCompute MapReduce 及 UDF 程序在分布式环境中运行时受到 Java 沙箱 的限制。

Instance（实例）

作业的一个具体实例，表示实际运行的 Job，类同 Hadoop 中 Job 的概念。详情请参见 基本概念 > 任务实例。

T

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 批量上传数据或者将数据下载。相关命令请参见 Tunnel 命令操作 或 批量数据通道 SDK。

U

UDF

广义的 UDF，即 User Defined Function，MaxCompute 提供的 Java 编程接口开发自定义函数，详情请参见 用户指南 > 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 Java UDF > UDAF。

UDTF

User Defined Table Valued Function，自定义表值函数，用来解决一次函数调用输出多行数据的场景，也是唯一能返回多个字段的自定义函数。而 UDF 只能一次计算输出一条返回值。详情请参见 Java UDF > UDAF。

Z

Resource（资源）

资源（Resource）是 MaxCompute 中特有的概念。您如果想使用 MaxCompute 的 自定义函数（UDF）或 MapReduce 功能，都需要依赖资源来完成。详情请参见 基本概念 > 资源。

项目空间（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空间的权限，通过安全授权，可以在一个项目空间中访问另一个项目空间中的对象，例如：表（Table），资源（Resource），函数（Function）和 实例（Instance）。

您可以通过 Use Project 命令进入一个项目空间，如下所示：

```
use my_project   -- 进入一个名为 my_project 的项目空间
```

运行此命令后，您会进入一个名为 my_project 的项目空间，您可操作该项目空间下的对象，例如：表，资源，函数和实例等，不需关心操作对象所在的项目空间。Use Project 是 MaxCompute 客户端提供的命令。在详细介绍这部分内容之前，文档会对常用的命令做简短介绍，详情请参见 MaxCompute 常用命令。

表是 MaxCompute 的数据存储单元，它在逻辑上也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，各个列的名称和类型构成这张表的 Schema。

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。

大数据开发套件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，详情请参见 数据管理概述。

MaxCompute 的表格分两种类型：内部表及外部表（MaxCompute2.0 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，表中的列可以是 MaxCompute 支持的任意一种 数据类型。

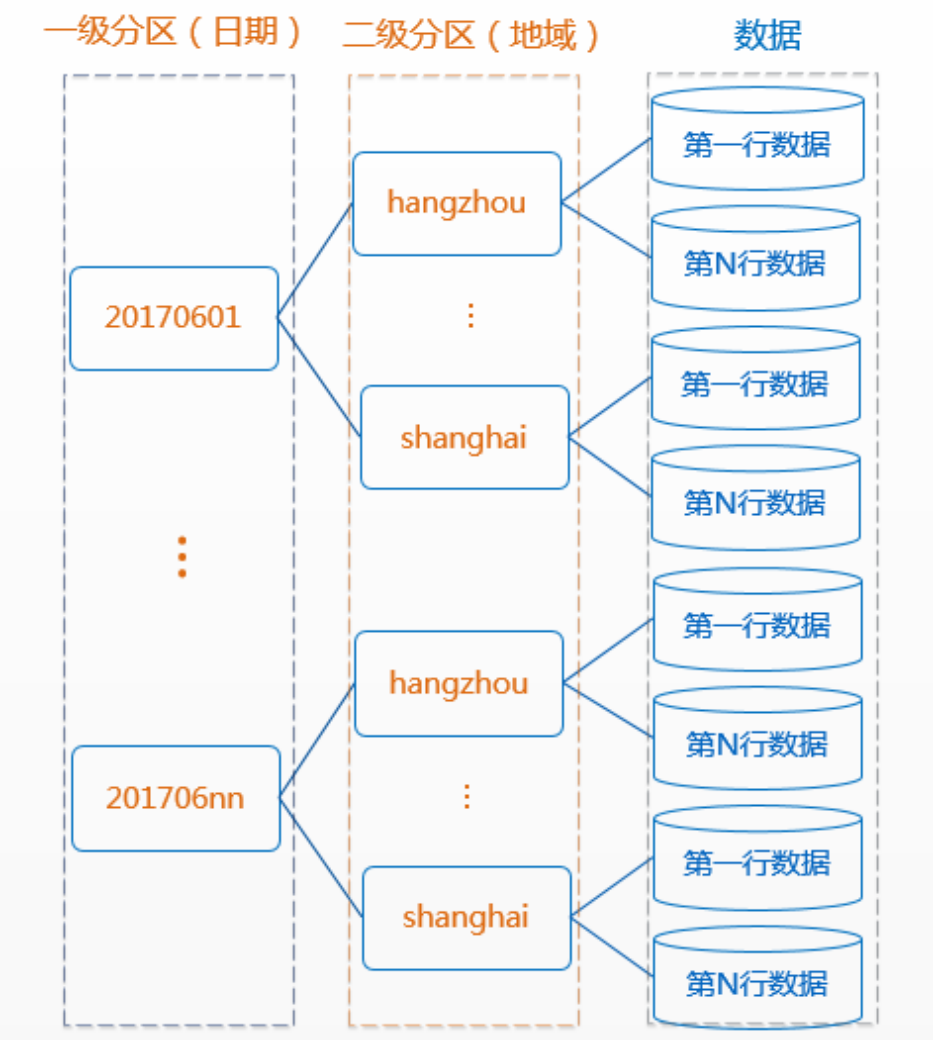
对于外部表，MaxCompute 并不真正持有数据，表格的数据可以存放在 OSS 或 OTS 中 。MaxCompute 仅会记录表格的 Meta 信息，您可以通过 MaxCompute 的外部表机制处理 OSS 或

OTS 上的非结构化数据，例如：视频、音频、基因、气象、地理信息等。外部表的处理请参见 处理非结构化数据。

分区表是指在创建表时指定分区空间，即指定表内的某几个字段作为分区列。大多数情况下，您可以将分区类比为文件系统下的目录。

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，分区之间正如多级目录的关系。

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，可避免全表扫描，提高处理效率，降低费用。



分区类型

MaxCompute2.0 对分区类型的支持进行了扩充，目前 MaxCompute 支持：TINYINT、 SMALLINT、INT、BIGINT、VARCHAR 和 STRING 分区类型。

注意：

在旧版 MaxCompute 中，仅承诺 STRING 类型分区。因为历史原因，虽然可以指定分区类型为 BIGINT，但是除了表的 schema 表示其为 BIGINT 外，任何其他情况实际都被处理为 STRING。

示例如下：

```
create table parttest (a bigint) partitioned by (pt bigint);
insert into parttest partition(pt) select 1, 2 from dual;
insert into parttest partition(pt) select 1, 10 from dual;
select * from parttest where pt >= 2;
```

执行上述语句后，返回的结果只有一行，因为 10 被当作字符串和 2 比较，没能返回。

分区使用限制

分区有以下使用限制：

- 单表分区层级最多 6 级。

  单表分区数最多允许 60000 个分区。

  一次查询最多查询分区数为 10000 个分区。

示例如下：

```
--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 (key string, value bigint) partitioned by (pt string,region string);
```

查询时，Where 条件过滤中使用分区列作为过滤条件：

```
select * from src where pt='20170601' and region='hangzhou';      -- 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; -- 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 String 类型与 Bigint（20151201）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。
```

部分对分区操作的 SQL 的运行效率则较低，会给您带来较高的计费，例如：使用动态分区。

对于部分 MaxCompute 的操作命令，处理分区表和非分区表时的语法有差别，详情请参见 DDL 语句 和 INSERT 操作 。

基本数据类型

MaxCompute2.0 支持的基本数据类型如下表，新增类型有：TINYINT、SMALLINT、 INT、 FLOAT、VARCHAR、TIMESTAMP 和 BINARY，MaxCompute 表中的列必须是下列描述的任意一种类型，详情如下：

注意：

- 若想使用新数据类型，需在 SQL 语句前加语句：set odps.sql.type.system.odps2=true;

类型	是否新增	常量定义	描述
TINYINT	是	1Y，-127Y	8 位有符号整形，范围 -128 到 127
SMALLINT	是	32767S， -100S	16 位有符号整形， 范围 -32768 到 32767
INT	是	1000，-15645787（注释1）	32位有符号整形，范围-231到231 - 1
BIGINT	否	100000000000L， -1L	64位有符号整形, 范围 -263 + 1到263 - 1
FLOAT	是	无	32位二进制浮点型
DOUBLE	否	3.1415926 1E+7	64位二进制浮点型
DECIMAL	否	3.5BD，99999999999.999999BD	10 进制精确数字类型，整形部分范围-1036+1到1036-1, 小数部分精确到 10-18
VARCHAR	是	无（注释2）	变长字符类型，n为长度，取值范围 1 到 65535
STRING	否	“abc”，’bcd’，”alibaba” ‘inc’（注释3）	字符串类型，目前长度限制为 8M
BINARY	是	无	二进制数据类型，目前长度限制为 8M
DATETIME	否	DATETIME ‘2017-11-11 00:00:00’	日期时间类型，范围从0000年1月1日到9999年12月31日，精确到毫秒
TIMESTAMP	是	TIMESTAMP ‘2017-11-11 00:00:00.123456789’	与时区无关的时间戳类型，范围从0000年1月1日到9999年12月31日 23.59:59.999999999, 精确到纳秒
BOOLEAN	否	TRUE，FALSE	boolean 类型, 取值 TRUE 或 FALSE

上述的各种数据类型均可为 NULL。

注释：

注释1：对于 INT 常量，如果超过 INT 取值范围，会转为 BIGINT；如果超过 BIGINT 取值范围，会转为 DOUBLE。

在旧版 MaxCompute 中，因为历史原因，SQL 脚本中的所有 INT 类型都被转换为 BIGINT，例如：

```
create table a_bigint_table(a int); -- 这里的int实际当作bigint处理
select cast(id as int) from mytable; -- 这里的int实际当作bigint处理
```

为与 MaxCompute 原有模式兼容，MaxCompute2.0 在未设定 odps.sql.type.system.odps2 为 true 的情况下，仍保留此转换，但会报告一个警告，提示 INT 被当作 BIGINT 处理了，如果您的脚本有此种情况，建议全部改写为 BIGINT，避免混淆。

注释2：VARCHAR 类型常量可通过 STRING 常量的隐式转换表示。

注释3：STRING 常量支持连接，例如 'abc' 'xyz' 会解析为 'abcxyz' ，不同部分可以写在不同行上。

复杂数据类型

MaxCompute2.0 支持的复杂类型见下表。

注意：

- 若想使用新数据类型，需在 SQL 语句前加语句：set odps.sql.type.system.odps2=true;

类型	定义方法	构造方法

ARRAY	array< int >;
	array< struct< a:int, b:string >>
		array(1, 2, 3);
		array(array(1, 2);
		array(3, 4))

MAP	map< string, string >; map< smallint, array< string>>	map("k1" , "v1" , "k2" , "v2"); map(1S, array('a' , 'b'), 2S, array('x' , 'y'))
STRUCT	struct< x:int, y:int>; struct< field1:bigint, field2:array< int>, field3:map< int, int>>	named_struct('x' , 1, 'y' , 2); named_struct('field1' , 100L, 'field2' , array(1, 2), 'field3' , map(1, 100, 2, 200)

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。

生命授权单位：days（天），只接受正整数。

非分区表若指定生命周期，自最后一次数据被修改的时间（LastDataModifiedTime）开始计算，经过 days 天后数据仍未被改动，则此表无需您干预，将会被 MaxCompute 自动回收（类似 drop table 操作）。

分区表若指定生命周期，则根据各个分区的 LastDataModifiedTime 判断该分区是否该被回收。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。

生命周期只能设定到表级别，不能再分区级设置生命周期。创建表时即可指定生命周期。

- 表若不指定生命周期，则表（分区）不会根据生命周期规则被 MaxCompute 自动回收。

关于建表时怎么指定/修改表生命周期、修改表 LastDataModifiedTime 等操作请参见 DDL 文档。

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 UDTF 使用说明 。

注意：

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 应用限制。

MaxCompute 资源包括以下几种类型：

File 类型。

Table 类型：MaxCompute 中的表。

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。

资源的相关操作请参见 创建资源 、删除资源 、查看资源清单 及 查看资源信息。

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。

注意：

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。

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 等，MaxCompute 会对其进行解析，得出任务的执行计划。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，您可以操作这个实例（Instance）的信息，例如：获取实例状态（Status Instance），终止实例运行（Kill Instance）等。

另一方面，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。

注意：

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。

在 MaxCompute 中，部分 任务（Task） 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。

运行阶段的状态为 Running（运行中），而结束阶段则会有 Success（成功），Failed（失败）或 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：

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

如果您是 MaxCompute 初学者

如果您是初学者，建议您从如下模块开始读起：

简介： MaxCompute 产品的总体介绍以及包含的主要功能。通过阅读该章节，您会对 MaxCompute 有一个总体的认识。

快速开始：通过示例，指导您如何进行申请账号、安装客户端、创建表、授权、导入导出数据、运行 SQL 任务、运行 UDF/Mapreduce 程序等操作。

基本介绍：MaxCompute 的基本概念及常用命令介绍。您可以进一步熟悉如何操作 MaxCompute。

工具：在分析数据之前，您需要掌握 MaxCompute 常用工具的下载，配置以及使用方法。我们提供以下客户端工具：

- Client：您可以通过此工具对 MaxCompute 进行操作。

建议您熟悉以上的模块后，再有针对性地对其他模块进行深入学习。

如果您是数据分析师

如果您是数据分析师，建议您熟读 SQL 模块：

SQL：您可以查询并分析存储在 MaxCompute 上的大规模数据。包含的主要功能如下：

支持 DDL 语句，您可以通过 Create、Drop 和 Alter 对表和分区进行管理。

您可以通过 Select 选择表中的某几条记录；通过 Where 语句查看满足条件的记录，实现过

滤功能。

您可以通过等值连接 Join 实现两张表的关联。

您可以通过对某些列 Group By，实现聚合操作。

您可以通过 Insert overwrite/into 把结果记录插入到另一张表中。

你可以通过内置函数和自定义函数（UDF）来实现一系列的计算。

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，并且某些数据分析可能无法用 SQL 来实现，此时推荐您学习 MaxCompute 更高级的功能模块。如下所示：

MapReduce：MaxCompute 提供的 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。

Eclipse Plugin：方便您使用 MapReduce，UDF 以及 Graph 的 Java SDK 进行开发工作。

Tunnel：您可以使用 Tunnel 服务向 MaxCompute 批量上传离线数据或者从 MaxCompute 下载离线数据。

SDK：

Java SDK：向开发者提供 Java 接口。

Python SDK：向开发者提供 Python 接口。

注意：

目前 MapReduce 以及 Graph 功能仍处于公测中，若您想使用这部分功能，可以通过工单系统提交申请。申请时请指明您的项目空间名称，我们会在 7 个工作日内处理。

如果您是项目 Owner 或者管理员

如果您是一个项目空间的 Owner 或者管理员，您需要熟知以下模块：

安全指南：您可以通过阅读该章节，了解如何进行给用户授权、跨项目空间的资源共享、设置项目空间的数据保护功能、policy 授权等操作。

MaxCompute 收费指南：介绍 MaxCompute 的收费模式。

- 以及部分只有项目空间 Owner 才能使用的命令，例如：常用命令 中 其他操作 的 SetProject 操作。

2017年9月19日—香港Region开服

2017年9月19日，阿里云数加 > MaxCompute 香港（香港）数据中心将正式开服售卖。

售价与华东2、华南1一致，可参考 计量计费文档。

开通时，选择什么Region您需要考虑的最主要因素是 MaxCompute 与其他阿里云产品之间的关系。

不同区域数据是不能互通，仅针对 MaxCompute 而言，假如您既购买华东 2 又购买香港的服务，那么 MaxCompute 之间的数据不互通。

如果您的云服务器 ECS 不在香港，若想访问连接香港的 MaxCompute，需跨 region 进行访问。跨 region 访问的各种服务连接请参见 访问域名和数据中心。

2017年9月7日—华南1（深圳）Region 开服

2017年9月7日，阿里云数加 > MaxCompute 华南1（深圳）数据中心将正式开服售卖，这是数加 > MaxCompute 在国内开服的第二个区域。

关于售价

华南 1 区域价格与华东 2 一致，主要分3部分进行收费：存储、计算、下载，其中计算（指 SQL 和 MR 计算任务）分预付费、按量后付费两种模式，存储和下载都是按量后付费。做预算的具体的售价信息请参见 官网定价页 或 计量计费文档。

关于开通

请您确保云账号是实名认证的账号，在开通购买页面进行区域选择时，注意选择华南1。那么什么场景适合选择区域是华南 1 呢？

选择地域时，您需要考虑的最主要因素是 MaxCompute 与其他阿里云产品之间的关系，如：

ECS 在华南 1 区域。ECS 访问、下载华南 1 的 MaxCompute 数据都可以通过内网，快速又避免跨 region 走公网下载，从而产生费用。

> 数据在华南 1 区域。如 RDS、OSS、TableStore 等在华南 1 区，那么在和华南 1 的 MaxCompute 进行数据传输时可以走内网，避免跨 Region 配置阿里云内网或 VP C 网络时不保证连通性的情况。

关于不同区域数据是否互通

仅针对 MaxCompute 而言，假如您既购买华东 2 又购买华南 1 的服务，那么 MaxCompute 之间的数据不互通。

如果您的云服务器 ECS 不在华南 1，若想访问连接华南1的 MaxCompute，需跨 region 进行访问。跨 region 访问的各种服务连接请参见 访问域名和数据中心。

关于数据跨 Region 迁移

若您原来已经开通购买华东 2 区域，目前想使用华南 1 区域，所以希望将华东 2 的所有数据迁移到华南 1。建议您通过以下方式进行迁移：

通过配置数据集成任务进行数据同步，如在华南 1 区域创建好的项目中，配置数据集成任务，来读取华东 2 区域的表，写入到华南 1 区域的表中。如果遇到问题可以提工单进行咨询。

> 若您对新服购买有疑惑，可通过工单咨询或加入 MaxCompute 购买钉钉群（群号： 11782920）进行咨询。

2017年8月16日-MR 开启计费

MaxCompute MapReduce 任务开启按量计费。计费标准如下：

```
MR 任务当日计算费用=当日总计算时 *  0.46元（人民币）
```

一个 MR 任务一次执行成功的计算时=任务运行时间（小时）*任务调用的 core 数量。

如一个 MR 任务一次执行成功是调用了 100core 并花费 0.5 小时，那么本次执行计算时为：0.5 小时× 100core= 50 个计算时。

详情请参见 计量计费说明。

2017年7月——开始分批升级 MaxCompute2.0

MaxCompute2.0 已经正式开始，升级后的 MaxCompute2.0 完全拥抱开源生态，支持更多的语言功能，带来更快的运行速度，同时新版本会执行更严格的语法检测。

> 具体升级时间会在分批升级时另行通知。

升级后需要对部分 SQL 语法做出调整，详情请参见 修改不兼容 SQL 实战 进行修改。

本次 MaxCompute2.0 产品升级以下功能：

全新的 SQL 引擎

更快的 SQL 执行引擎：降低企业大数据分析成本，SQL 执行效率更高。

生态兼容

MaxCompute 显著提升了针对 Hive 等开源产品的兼容性，直接兼容 Hive 的 UDF/UDAF/UDTF，为 Hive 开发的大部分 UDF/UDAF/UDTF 可不经修改，直接在 MaxCompute 中运行。

详情请参见 INSERT操作、SELECT操作、数据类型、日期函数、数学函数、字符串函数、其他函数和 JAVA UDF等相关文档。

非结构化处理能力

通过 MaxCompute SQL 直接处理 OSS/TableStore（OTS）数据，从而处理音频，视频，图像，气象等非结构化数据以及 K-V 类型的数据。

详情请参见前言、访问 OTS 非结构化数据 和 访问 OSS 非结构化数据。

- **MaxCompute Studio**

MaxCompute 编译器基于 MaxCompute2.0 全新自主研发的 SQL 引擎，尤其配合使用 MaxCompute Studio，提供了丰富的错误提示、警告的功能及作业排队展示。

详情请参见巧用 MaxCompute 编译器的错误和警告。

2017年06月01日

尊敬的阿里云客户：

MaxCompute 目前只针对 SQL 作业进行收费，MapReduce 作业近期将开启收费。

MaxCompute 近期将进行 2.0 版本升级，升级过程中会逐批进行，本次版本变更将带来一些新特性：

SQL 执行引擎

降低大数据分析成本。

SQL 执行效率更高。

非结构化处理能力

通过 MaxCompute SQL 直接处理 OSS/TableStore（OTS）数据，从而处理音频，视频，图像，气象等非结构化数据以及 K-V 类型数据。

生态兼容

向开源 Hadoop MapReduce 使用场景高度兼容的 MaxCompute MapReduce SDK。

基于 InteliJ 的本地开发插件

本地对 MaxCompute 作业进行开发，调试。

实时数据上传

新增 DataHub 模块，帮助企业数据实时上传。

新型算法平台

基于 GPU 的深度学习算法框架，在线预测服务等。

如果您在系统升级过程中遇到问题，请及时提交工单联系我们。

2016年09月16日

尊敬的阿里云客户：

MaxCompute 在 11 月 1 日进行了计费变更。本次变更升级存储及计算两个方面：

升级后，修复部分 SQL 作业无法计费的问题。

升级后，解决部分场景下 MaxCompute 存储费用存在的问题.升级后，存储将按照实际情况收费。

以上两处之前均不存在向您多收取费用的情况，对于历史作业不做费用补收。您可以通过计量计费说明中获取账单详情的相关指引，从阿里云官网获取每天详细的计量计费信息。

目前 MaxCompute 只有 SQL 作业收费（不包括 UDF），UDF/MapReduce/Graph/PAI 等作业仅是公测状态。具体收费计划请关注官网公告。

感谢您长期以来对 MaxCompute 产品的支持。

MaxCompute（原 ODPS）是一种大数据计算服务，能提供快速、完全托管的 PB 级数据仓库解决方案，已经与阿里云部分产品集成，可以快速实现很多业务场景。

MaxCompute 与大数据开发套件

大数据开发套件 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。在数加中，大数据开发套件控制台即为 MaxCompute 控制台。

通过大数据开发套件，您既可直接编写并运行 MaxCompute SQL，又能可视化配置工作流并定时调度运行 MaxCompute SQL、MR 等任务。更多使用说明请参考 大数据开发套件帮助文档。

> 您可以将大数据开发套件理解成 MaxCompute 的 web 客户端。

MaxCompute 与数据集成

MaxCompute 可以通过数据集成加载不同数据源数据，同样也可以通过数据集成把 MaxCompute 的数据导出到各种业务数据库。

数据集成已经集成到大数据开发套件作为 **数据同步** 任务进行配置、运行。您可直接在大数据开发套件上 配置 MaxCompute 数据源，再配置 读取 MaxCompute 表或者 写入 MaxCompute 表任务，整个过程只需在一个平台上进行操作。

MaxCompute 与机器学习

机器学习 是基于 MaxCompute 的一款机器学习算法平台。数加上创建好 MaxCompute 项目，开通好机器学习，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。详情请参见 机器学习操作文档。

MaxCompute 与 QuickBI

数据在 MaxCompute 进行加工处理后，将 Project 添加为 QuickBI 数据源，即可在 QuickBI 页面对 MaxCompute 表数据进行 报表制作，实现数据可视化分析。

MaxCompute 与 AnalyticDB

AnalyticDB 是海量数据实时高并发在线分析（Realtime OLAP）的云计算服务，与 MaxCompute 双剑合璧实现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 数据导入到 AnalyticDB，有以下两种方式：

通过 DMS for AnalyticDB 的 导入导出 功能进行配置。

通过大数据开发套件配置数据同步任务，读 MaxCompute 和 写 AnalyticDB。

MaxCompute 与推荐引擎

推荐引擎 是在阿里云计算环境下建立的一套推荐服务框架，推荐服务通常由三部分组成：日志采集，推荐计算和产品对接，而推荐计算的离线计算输入和输出都是 MaxCompute（原 ODPS）表。

在推荐引擎控制台的资源管理页面，通过 添加云计算资源 的方式，将 MaxCompute 项目添加为推荐引擎的计算资源。

MaxCompute 与表格存储

表格存储（Table Store） 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，MaxCompute2.0 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 访问 OTS 非结构化数据。

MaxCompute 与 OSS

对象存储 OSS 是海量、安全、低成本、高可靠的云存储服务，MaxCompute2.0 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 访问 OSS 非结构化数据。

MaxCompute 与 OpenSearch

阿里云 开放搜索 OpenSearch 是一款阿里巴巴自主研发的大规模分布式搜索引擎平台。数据通过 MaxCompoute 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。

MaxCompute 与移动数据分析

移动数据分析（Mobile 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的分析报表不能满足 APP 开发者的个性化需求时，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、分析自己的数据。

MaxCompute 与日志服务

日志服务 能快速完成数据采集、消费、投递以及查询分析等功能。日志数据采集后，需要更多的个性化分析、挖掘，您可以在日志服务上 投递日志到 MaxCompute，通过 MaxCompute 对日志数据进行个性化、深层次的数据分析、挖掘。

序号	文档名称	变更内容	变更时间
1	INSERT操作、SELECT操作	DML相关文档中添加MaxCompute2.0版本扩展支持的新功能，包括values、CTE、SEMIJOIN等；	2017.9
2	日期函数、数学函数、字符串函数、其他函数	相关函数文档添加新支持的函数说明。	2017.9
3	数据类型、 JAVA UDF	数据类型文档添加新支持的数据类型说明包括基本数据类型和复杂数据类型； JAVA UDF文档添加新数据类型使用方法、返回类型、复杂数据类型示例。	2017.9
4	类型转换	从SQL概要中拆分出来，同时添加新数据类型隐式转换对照表。	2017.9
5	访问域名和数据中心	添加华南1Tunnel Endpoint；添加香港、亚太东南1Region相关endpoint说明。	2017.9
6	长周期指标的计算优化方案	新增最佳实践文章。	2017.9
6	Logview、巧用MaxCompute 编译器的错误和警告	将这两篇所在目录《JOB运行信息查看》从“工具及下载”移动到“用户指南”一级目录下。	2017.9
7	处理非结构化数据	《将处理非结构化数据》相关内容移动到“用户指南”一级目录下； OSS和OTS自定义授权内容添加AliyunODPSRolePolicy的策略内容。	2017.9
8	产品简介	产品简介目录下文档内容格式规范化	2017.9