Alibaba Cloud MaxCompute

Product Introduction

Issue: 20191021

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. **You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.**
2. **No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company, or individual in any form or by any means without the prior written consent of Alibaba Cloud.**
3. **The content of this document may be changed due to product version upgrades, adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.**
4. **This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults" and "as available" basis. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequent**

ial, exemplary, incidental, special, or punitive damages, including lost profits arising from the use or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.

5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.
6. Please contact Alibaba Cloud directly if you discover any errors in this document.

Document conventions

Style	Description	Example
	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	 **Danger:** Resetting will result in the loss of user configuration data.
	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	 **Warning:** Restarting will cause business interruption. About 10 minutes are required to restart an instance.
	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	 **Notice:** If the weight is set to 0, the server no longer receives new requests.
	A note indicates supplemental instructions, best practices, tips, and other content.	 **Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click Settings > Network > Set network type.
Bold	Bold formatting is used for buttons, menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands.	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *`Instance_ID`*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`

Style	Description	Example
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Contents

1 What is MaxCompute?

The big data computing service (MaxCompute, formerly called ODPS) is a fast and fully hosted GB/TB/PB level data warehouse solution.

MaxCompute supports a variety of classic distributed computing models that enable you to solve massive data calculation problems while reducing business costs, and maintaining data security.

MaxCompute seamlessly integrates with DataWorks, which provides one-stop data synchronization, task development, data workflow development, data operation and maintenance, and data management for MaxCompute. For more information, see *#unique_4*.

MaxCompute is mainly used to store and compute batches of structured data. It provides a massive range of data warehouse solutions as well as big data analysis and modeling services. As data collection techniques are becoming increasingly diverse and comprehensive,industries are amassing larger and larger volumes of data. The scale of data has increased to the level of massive data (100 GB, TB and even PB) that traditional software industry can not carry.

Given these massive data volumes, the limited processing capacity of a single server has prompted analysts to move towards distributed computing. However, distributed computing models are not easy to maintain and demand highly-qualified data analysts. When using a distributed model,data analysts not only need to understand their business needs, but also must be familiar with the underlying computing model. The purpose of MaxCompute is to provide you with a convenient way of analyzing and processing mass data, and you can achieve the purpose of analyzing large data without having to care about the details of distributed computing.

 **Note:**
MaxCompute is widely used by Alibaba Group in scenarios such as data warehouse and BI analysis, web log analysis, transaction analysis of e-commerce sites, and customer behavior analysis.

MaxCompute learning path

You can quickly learn about MaxCompute's related concepts, basic operations and advanced operations through *MaxCompute learning path*.

Product advantage

- **Large-scale computing and storage**

  MaxCompute is suitable for the storage and processing of large volumes of data (up to PB-level).
- **Multiple computational models**

  MaxCompute supports data processing methods based on SQL, MapReduce, Graph, MPI iteration algorithm, and other programming models.
- **Strong data security**

  MaxCompute has stabilized alloffline analysis for all Alibaba Group's business for more than seven years, providing multilayer sandbox protection and monitoring.
- **Cost-effective**

  MaxCompute can help reduce procurement costs by 20%-30% compared with on-premises private cloud models.

Function

- **Data tunnel**
  - **Supports large volumes of historical data channels**

    TUNNEL **provides high concurrency data upload and download services. This service supports the import and export of terabytes or petabytes of data on a daily basis, which is particularly useful for the batch import of full or historical data. Tunnel Provides you with a Java programming interface, and in the MaxCompute client tool, there are corresponding commands for local file and service data interchange.**
  - **Real-time and incremental data channels**

    For real-time data upload scenarios, MaxCompute provides DataHub services with low latency and convenient usage. It is especially suitable for incrementa l data imports. DataHub also supports a variety of data transmission plug-ins, such as Logstash, Flume, Fluentd, and Sqoop, it supports Log. Service's

delivery log to MaxCompute, and then use DataWorks to do log analysis and mining.

- Computing and analysis tasks

  MaxCompute provides multiple computing models.

  - SQL: In MaxCompute, data is stored in tables. MaxCompute provides an SQL query function for the external interface. You can operate MaxCompute similarly to a traditional database software but with the ability to process PB-level data.

    

    Note:

    ■ MaxCompute SQL does not support transactions, index, or Update/Delete operations.

    ■ MaxCompute SQL syntax differs from Oracle and MySQL, notably, you cannot seamlessly migrate SQL statements of other databases into MaxCompute.

    ■ In terms of usage, MaxCompute SQL can complete queries at the second- to millisecond-level,and can not return results at milliseconds.

    ■ The advantage of MaxCompute SQL is low learning cost. You don't need to understand the concept of complex distributed computing. If you have experience in database operations, you can familiarize yourself with MaxCompute SQL quickly.

  - *UDF*: A user-defined function.

    MaxCompute provides numerous *built-in functions*to meet your computing needs, while also supporting the creation of custom functions.

  - *MapReduce*: MapReduce is a Java MapReduce programming model provided by MaxCompute. It uses the Java programming interface and is designed to simplify the development process. However, users are recommended to have a basic understanding of the concept of distribution, and relevant programming experience before using MapReduce. MaxCompute MapReduce provides you with Java programming interface.

  - *Graph*: Graph in MaxCompute is a processing framework designed for iterative graph computing. Graph computing jobs use graphs to build models. Graphs are composed of vertices and edges. Vertices and edges contain values.

After performing iterative graph editing and evolution, you can get the final result. Typical applications include PageRank, SSSP algorithm, and K-Means algorithm.The graph is edited and evolved through an iteration, and the results are finally solved. Typical applications: *PageRank*, *single source shortest distance algorithm*, *K-means clustering algorithm*, **and so on.**

- **SDK**

  A convenient toolkit provided for developers. For more information, see *MaxCompute SDK*.

- **Secure**

  Maxcompute offers powerful security services to protect your data, for more information, see the *security guide*.

2 What you must know

This topic highlights important key features that developers, and project owners or administrators must be aware of before using MaxCompute.

For beginners

If you are a beginner, we recommend that you start from the following topics:

- *MaxCompute Summary*: **Introduces MaxCompute, including its main function modules. By reading this chapter, you can have a general knowledge of MaxCompute.**
- *Quick Start*: **Provides a step-by-step guide including how to apply for an account, install the client, create a table, authorize a user, export/import data, run SQL tasks, run UDF, and run MapReduce programs.**
- *MaxCompute glossary*and *#unique_18*: **Detail key terms and frequently used commands of MaxCompute. You can be further familiar with how to operate MaxCompute.**
- *Tools*: **Before analyzing the data, you may need to master how to download, configure and use the frequently used tools.**

  #unique_20:**You can operate MaxCompute through this tool.**
- *Endpoints and Data Centers*: **MaxCompute Region opens and connects to answer network connectivity and download data charges that you encounter in other cloud products (ECS, Table Store, OSS) interchange scenarios.**

After you are familiar with those modules that mentioned preceding, you are recommended to perform a further study on other modules.

For data analysts

If you are a data analyst, we recommend that you read the following topic:

- *MaxCompute SQL*: Query and analyze massive volumes of data that are stored on MaxCompute. It includes the following functions:
  - Use DDL statements CREATE, DROP, and ALERT to manage tables and partitions.
  - Use a SELECT statement to select records in a table, and use a WHERE clause to view the records meeting the filter condition.
  - Associate two tables through an Equijoin operation.
  - Aggregate columns using a GROUP BY statement.
  - You can insert the result records into another table through Insert overwrite/into syntax.
  - You can use built-in functions and user-defined functions (UDF) to complete a variety of computations.

For developers

If you have a certain level of development experience, understand the concept of distribution, and know that some data analysis functions may not be possible with SQL, then we recommend that you learn more about the following advanced functional modules of MaxCompute:

- *MapReduce*: Explains the MapReduce programming interface. You can use the Java API, which is provided by MapReduce, to write MapReduce program for processing data in MaxCompute.
- *Graph*: Provides a set of frameworks for iterative graph computing. This function uses graphs to build models. Graphs are composed of vertices and edges. Vertices and edges contain values. This process outputs a result after performing iterative graph editing and evolution.
- *Eclipse Plugin*: Facilitates you to use the Java SDK of MapReduce, UDF, and Graph for development work.
- *Tunnel*: Facilitates users to use the Tunnel service to upload batch offline data to MaxCompute, or download batch offline data from MaxCompute.
- SDK:
  - *Java SDK*: Provides developers with Java interfaces.
  - *Python SDK*: Provides developers with Python interfaces.

 **Note:**

> *MapReduce* and *Graph* are still in open beta, and if you want to use this feature, applications can be submitted through the job system. Please specify the name of your project when you apply, and we will process it within 7 working days.

For project owners or administrators

- **Project management**

  Projects are the smallest units of MaxCompute. They are similar to traditiona l databases or schemas and serve to isolate users and manage users' access permissions. Each user can be granted the permissions for multiple projects. This allows the same user to access such objects as tables, resources, functions , and instances in different projects. Operating MaxCompute means operating various objects in projects.

  - **Prepare for creating a project.**

    ■ **Estimate the resources you need.**

    MaxCompute charges fees for the following three types of resources:

    1. **Storage resources: charged by using the Pay-As-You-Go billing method. The prices are divided into multiple levels. You can estimate the fees you need to pay based on the data volume. However, the data is not stored in MaxCompute all at once within a single day. Instead, the data may be read from or written into MaxCompute at any time on every day. Therefore, the estimated resources and fees may differ from your bill.**
    2. **Computation resources: charged by using the Subscription or Pay-As-You -Go billing method. Computation resources are used for SQL, MapReduce , Spark, and Lightning tasks. Estimating the fees for computation resources is difficult at the very beginning. Therefore, we recommend that you start from the Pay-As-You-Go billing method and then decide**

whether to switch to the Subscription billing method after a period of testing.

3. Internet download traffic: charged by using the Pay-As-You-Go billing method. You are billed only when you consume traffic for downloading resources through the Internet.

For more information about metering and pricing, see *#unique_25*,*#unique_26*,*#unique_27*.

■ Register with Alibaba Cloud and activate the MaxCompute services.

Before you create a project, you must register with Alibaba Cloud to obtain an Alibaba Cloud account. Then, determine whether you want to use the Subscription or Pay-As-You-Go billing method based on your resource estimation, and use your Alibaba Cloud account to activate the MaxCompute services. MaxCompute will deduct fees from your Alibaba Cloud account.

- Create a project.

  For information about how to create a project, see *#unique_28*.

- Manage the members in a project.

  You need to assign roles and permissions to the project members. If you use MaxCompute through DataWorks, you also need to consider the mapping between MaxCompute permissions and DataWorks permissions.

- Manage RAM users.

  MaxCompute projects support two types of accounts: Alibaba Cloud accounts and RAM user accounts. You can add any RAM user under your Alibaba Cloud account to a MaxCompute project, but MaxCompute does not consider how the permissions of the RAM user are defined when it verifies the RAM user. For more information, see *#unique_29*.

  When you operate MaxCompute through DataWorks, you can only use your Alibaba Cloud account to create RAM users under your Alibaba Cloud account , add the RAM users as members to a DataWorks workspace, and manage the RAM users as needed.

   Note:

  ■ Each project member must have a unique RAM user account.

■ Once a project member has left the company or has been transferred to another job position, you must delete the RAM user account of the project member immediately.

Note:
If the RAM user is a project member in DataWorks, you must delete the project member from DataWorks and then delete the RAM user from the RAM user management system.

- Manage scheduling resources.

  ■ Scheduling resources

  The scheduling resources in DataWorks are categorized as default scheduling resources and custom scheduling resources. They are used to distribute or run tasks. For more information, see *#unique_30*.

  1. Default scheduling resources are public resources in DataWorks. When a large number of DataWorks nodes are running concurrently, the DataWorks nodes that cannot occupy scheduling resources enter the waiting state. These DataWorks nodes start to distribute tasks immediatel y after they occupy scheduling resources.
  2. Custom scheduling resources are used to distribute or run data synchronization or other tasks. You can use your Alibaba Cloud account to configure a physical device or an ECS instance as a scheduling server that can distribute tasks. With custom scheduling resources , MaxCompute can properly distribute and run tasks even when the default scheduling resources are running out. If you want to create custom scheduling resources in a custom resource group, you need to open a ticket. If you want to create custom scheduling resources in an existing custom resource group, you do not need to open a ticket.

- Set a project.

  If you are the project owner, then you need to set the project such as specifying whether to enable full table scan and whether to enable MaxCompute 2.0 by default. For more information, see *#unique_31*.

- **Security management**

  You need to manage users, roles, and permissions. MaxCompute and DataWorks each have a unique permission model. When you operate MaxCompute through DataWorks, you must know the mapping between MaxCompute permissions and DataWorks permissions so that you can manage permissions based on your service requirements. Specifically, you can grant permissions to users, share resources among projects, and enable data protection and set policies for projects. For more information, see *#unique_32*.

- **Cost management**

  You need manage your costs based on the pricing and billing of MaxCompute.

3 Definitions

3.1 MaxCompute glossary

This article lists the common concepts and terminologies of MaxCompute. For detailed description, please refer to the link in this article.

A

- **AccessKey**

  Access Key (AK for short, including Access Key Id and Access Key Secret) is the key to access the Aliyun API. After registering cloud account on Ali cloud official website,it can be generated on the Accesskeys management page to identify users and do signature verification for accessing MaxComputer or other cloud products. Access Key Secret must be kept secret.

- **Security**

  MaxCompute multi-tenant data security system mainly includes user authentication, user and authorization management in project space, resource sharing across project space and data protection in project space. For more details on MaxCompute security operation, see *Safety Guide*.

C

- **Console**

  MaxCompute Console is a client tool running under Windows/Linux. It can submit commands to complete project management, DDL, DML and other operations through Console. For tool installation and common parameters, See the *Client* **for tool installations and common parameters.**

D

- **Data Type**

  The data type corresponding to all columns in the MaxCompute table. For data types currently supported, see *Basic Concepts>Data types*.

- **DDL**

  Data Definition Language.Like creating tables, creating views, and so on, MaxCompute Div syntax see *User's Guide>DDL*.

- **DML**

  Data Manipulation Language. For example, INSERT operations, MaxCompute DML syntax, please see *Insert Operation* .

F

- **Fuxi**

  Fuxi is the module responsible for resource management and task scheduling in the core of Flying Platform. It also provides a basic programming framework for application development. The bottom task scheduling module of MaxCompute is the scheduling module of Fuxi.

I

- **Instance (Instance)**

  A specific instance of a job that represents a job that actually runs, similar to the concept of a job in hadoop. See *Basic Concepts>Task Instance* for details.

M

- **MapReduce**

  MaxCompute a programming model for processing data, usually used for parallel operation of large data sets. You can use the interface provided by MapReduce (Java API) to write MapReduce programs to process data in MaxCompute. The idea of programming is to divide data processing methods into Map (mapping) and Reduce (Protocol).

  Before formally executing Map, a partition is required. A slice is a cut of input data into equal-size blocks, each of which acts as a single map. The input of the worker is processed so that multiple Map Worker can work together.Each Map Worker reads in its own data, calculates and processes them, and finally integrates the intermediate results through the Reduce function to get the final results. For details, refer to the *User's guide>MapReduce*.

O

- **ODPS**

  ODPS is the original name of MaxCompute.

P

- **Partition (partition)**

  Partition partition refers to a table, based on the partition field (one or more combinations) divide the data store. That is, if the table does not have a partition, the data is placed directly under the directory where the table is located. If a table has a partition, each partition corresponds to one of the directories in the table, the data is stored separately in a different partition directory. For more information about partitions, see *Basic Concepts>Partitions*.
- **Project (Project)**

  Project is the basic organizational unit of MaxCompute. It is similar to the concept of database or Scheme in traditional database, and is the main boundary of multi-user isolation and access control. For details, please see *Basic Concepts>Project* .

R

- **Role (role)**

  Roles are concepts used in MaxCompute security functions and can be seen as a collection of users with the same privileges. Multiple users can appear at one role at the same time, and a user can also belong to multiple roles. After all roles are authorized, all users under the role have the same permissions. For more information about role management, please see *User's guide>Role management* .
- **Resource (resources)**

  Resource (Resource) is a unique concept in MaxCompute. If you want to use MaxCompute's custom function (UDF) or MapReduce function, you need to rely on resources to complete it. For details, please see *Basic Concepts>Resource*.

S

- **SDK**

  Software Development Kits. Generally, it is a collection of development tools used by software engineers to build application software for specific software

packages, software instances, software frameworks, hardware platforms, operating systems, document packages, etc. MaxCompute currently supports *#unique_44* and Python SDK.

- **Authorization**

  The project space administrator or project owner gives you permission to perform certain operations on Objects (or objects, such as tables, tasks, resources, etc.) in MaxCompute, including reading, writing, viewing, etc. For the specific operation of authorization, see *User management*.

- **Sandbox**

  Security restrictions: MaxCompute MapReduce and UDF programs are restricted by Java Sandbox while running in a distributed environment.

T

- **Table (table)**

  The table is the data storage unit of MaxCompute. See *Basic Concepts>Table*.

- **Tunnel**

  MaxCompute's data channels provide high concurrency offline data upload and download service. You can use the tunnel service to bulk upload or download data to MaxCompute. Please refer to the tunnel command operation or bulk data channel SDK for relevant commands.

U

- **UDF**

  Generalized UDF, user defined Function, the Java programming interface provided by MaxCompute develops custom functions, for more information, refer to *User 's guide>UDF*.

  In a narrow sense, UDF refers to user-defined scalar function, whose input and output are one-to-one, that is, reading in a row of data and writing out an output value.

- **UDAF**

  User Defined Aggregation Function, Custom aggregation function, whose input and output are many-to-one, aggregates multiple input records into one output

value. It can be combined with the Group By statement in SQL. For details, please see *Java UDF>UDAF*.

- UDTF

  User Defined Table Valued Function,customictablevaluedFunction, the userDefinedTablevaluedFunction, the customictablevaluedFunction, the function of the function called to the function, the set to the the function to the function, the function to the set to the value to the returned the function, the the function of the function of the the function of the function call to the value of the *Java UDF>UDAF*.

3.2 Table

A table is the data storage unit in MaxCompute. A table is a two-dimensional data structure composed of rows and columns. Each row represents a record, and each column represents a field with the same data type. One record can contain one or more columns. The column name and data type comprise the schema of a table.

The operating objects (input, output) of various computing tasks in MaxCompute are tables. You can create a table, delete a table, and import data into a table. For more information, see *Table operations*.

Note:
The data management module of DataWorks allows you to create, organize, and modify data lifecycles for MaxCompute tables and grant management permissions. For more information, see *#unique_50*.

MaxCompute 2.0 supports internal tables and external tables.

- Data of internal tables is stored in MaxCompute. The columns in external tables can be of any *data types* supported by MaxCompute.
- Data of external tables is not stored in MaxCompute. Instead, this data can be stored in *OSS* or *OTS*. MaxCompute only records metadata of the external tables. You can use MaxCompute to process unstructured data of external tables, such as video, audio, or meteorological data.

Note:
Use of DUAL tables:

- **Unlike databases such as Oracle, MaxCompute does not automatically create DUAL tables.**
- **If you are using DUAL tables for testing, you can run the `CREATE TABLE IF NOT EXISTS DUAL (DUMMY VARCHAR(1));` command to create a table named DUAL with only one field for testing.**
- **DUAL tables are used in the same way as Oracle. For example, you can run the `select getdate() from dual;` to use a DUAL table.**

4 MaxCompute features in different regions

MaxCompute is a big data computing service that provides multiple built-in computing models to meet a wide range of data analysis requirements. This topic lists the enabling status of these computing models in different regions.

Region	SQL	MapReduce	Spark
China (Beijing)	Enabled	Enabled	Enabled
China (Hangzhou)	Enabled	Enabled	Enabled
China (Shanghai)	Enabled	Enabled	Enabled
China (Shenzhen)	Enabled	Enabled	Enabled
China (Chengdu)	Enabled	Enabled	Enabled
China (Hong Kong)	Enabled	Enabled	Enabled
Singapore	Enabled	Enabled	Enabled
Malaysia (Kuala Lumpur)	Enabled	Enabled	Enabled
Indonesia (Jakarta)	Enabled	Enabled	Enabled
Australia (Sydney)	Enabled	Enabled	Enabled
Japan (Tokyo)	Enabled	Enabled	Enabled
US (Silicon Valley)	Enabled	Enabled	Enabled
US (Virginia)	Enabled	Enabled	Enabled
Germany (Frankfurt)	Enabled	Enabled	Enabled
India (Mumbai)	Enabled	Enabled	Enabled
UK (London)	Enabled	Enabled	Enabled