

Alibaba Cloud MaxCompute

Product Introduction

Issue: 20190114

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, consequential, 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 content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectu

al property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of the Alibaba Cloud website, product programs, or content 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.

Generic conventions

Table -1: Style conventions

Style	Description	Example
	This warning information 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.
	This warning information 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 restore business.
	This indicates warning information, supplementary instructions, and other content that the user must understand.	 Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	 Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	It is used for buttons, menus, page names, and other UI elements.	Click OK .
Courier font	It is used for commands.	Run the <code>cd /d C:/windows</code> command to enter the Windows system folder.
<i>Italics</i>	It is used for parameters and variables.	<code>bae log list --instanceid Instance_ID</code>
[] or [a b]	It indicates that it is a optional value, and only one item can be selected.	<code>ipconfig [-all -t]</code>
{ } or {a b}	It indicates that it is a required value, and only one item can be selected.	<code>swich {stand slave}</code>

Contents

Legal disclaimer.....	I
Generic conventions.....	I
1 What is MaxCompute.....	1
2 Announcement.....	5
3 Reading guidance.....	7
4 History.....	9

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 [DataWorks](#).

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 has been widely used in the Alibaba group, such as data warehouse and BI analysis of large Internet enterprises, web log analysis, transaction analysis of e-commerce sites, user characteristics and interest mining.

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 incremental 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](#).

What to do next

Now, you have learned about MaxCompute's product advantages, functional features and other related profiles, you can continue to learn the next tutorial. In this tutorial, you will understand the related charges of MaxCompute. For more information, see [Product Pricing](#).

2 Announcement

This article provides you with updated records of the use of MaxCompute.

January 15th,2019 HongKong Region Underlying Structure Optimization

MaxCompute will optimize the underlying Meta Data Warehouse structure during **16:00 - 20:00 January 15th 2019** to improve the stability and performance. During the release window , **HongKong Region** users could encounter task submission delay about **one minute** or task failure. Under very extremely scenario , application would be **not available for 30 minutes**. Please avoid to submit task during the release window. There is NO affect to other regions. If you have any concerns, please contact us by DingTalk or ticket .

Dec 24 ,2018 MaxCompute Supports Timezone Configuration

The default timezone of MaxCompute is UTC+8. It computes base on UTC+8 time when you use datetime/timestamp/date data types or time related built-in functions. Dec 24 ,2018 , MaxCompute is starting to support timezone configuration functionality. There are two ways to configure timezone in MaxCompute:

- Session level: `set odps.sql.timezone=<timezoneid>;` This statement needs to be submitted with the calculation statement, such:

```
set odps.sql.timezone=Asia/Tokyo;
select getdate();
Results:
output:
+-----+
| _c0      |
+-----+
| 2018-10-30 23:49:50 |
+-----+
```

- Project level: `setProject odps.sql.timezone=<timezoneid>;` This query has to be executed with command line by Project owner. Once timezone configuration submitted in Project level, all time related computing would use new timezone immediately and there would be affection to data of existing tasks. We suggest to configure timezone for new projects rather than existing projects.

Note:

- Timezone configuration supports:
 1. SQL built-in date functions GETDATE, TO_DATE, TO_CHAR
 2. The conversion between datetime and string.

- Timezone configuration does not support:
 1. All other tasks except SQL, like MR.
 2. UDF, UDJ, UDT, select transform function in SQL.
- Timezone configuration format is like Asia/Shanghai (daylight saving is considered) rather than GMT+9.
- When there is difference between SDK and project timezone , you need to configure GMT timezone for executing `date to string` command.
- After timezone configuration, there would be discrepancy between real time and output time . The difference is 5 minutes 52 seconds when real time is between 1900 and 1928. The difference is 9 seconds when real time is before 1928.
- MaxCompute service, Java SDK and related console will release new version (Java SDK/ Console version with postfix `-over sea`) in near future to ensure the accuracy of MaxCompute datetime type data in multiple regions. The display of existing MaxCompute datetime data which is older than 1928 will be affected in new version.
- Please also upgrade Java SDK/Console(if the local timezone is not UTC+8) when you do MaxCompute upgrading, to ensure the accuracy and consistence between SQL computing result and tunnel transforming data after `1900-01-01`. For the datetime data before `1900-01-01`, there would be 343 seconds difference between SQL computing output and tunnel transforming data. For the existing datetime data before `1928-01-01`, the output time will reduce 352 seconds in new version.
- There would be data discrepancy risk between SQL output and tunnel transforming data if SDK/Console version without postfix of `-over sea` used continually. The time difference is 9 seconds if real time is before `1900-01-01`. The time difference is 352 seconds if real time is in the range of 1900-01-01 ~ 1928-01-01.

**Note:**

The time zone update or configuration of Java SDK/Console version does not affect the time zone configuration of DataWorks, so there will be differences in time zone. You need to evaluate the impact of timed task scheduling in DataWorks. In Japan Region, DataWorks server is GMT+9, in Singapore Region, DataWorks server is GMT+8.

- When using third-party clients connected by JDBC, you need to set time zones on the client side to ensure consistency with time zones on the server side.

3 Reading guidance

This article recommends different document reading orders for you according to your roles, and introduces you the limitations of MaxCompute product modules.

For first time users

The following sections give recommended reading for users of differing expertise.

- [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.
- [Basic Introduction](#): Details 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.

We provide the following [tool](#): 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, it is recommended that you read the contents of the SQL module.

- [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, understanding the concept of distribution, and some data analysis may not be possible with SQL, you are advised to learn more about MaxCompute's advanced functional modules. As shown in the following:

- [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/administrators

If you are a project owner or administrator, you are recommended to read:

- [Security](#): Explains how to grant privileges to a user, share resource span projects, set project protection, and grant privilege by policy, and so on.
- [Billing](#): Details the pricing of MaxCompute.
- Commands that only the project owner can use. For example, the **SetProject** operation in [Others](#) of [Common Commands](#).

4 History

When Alibaba Cloud was founded in September 2009, our vision was to be the first platform for data processing and sharing. In April 2010, that vision came a little closer to reality with the Open Data Processing Service (ODPS) supporting the newly launched loan business area of Ant Financial. From there, things went from strength to strength. In 2012, a Unified Data Platform was established. By the end of 2013, we had the ability to process massive volumes of data on a large scale. From 2014 to 2015, the big data platform was refined, and in 2016, MaxCompute 2.0 was born, providing massive data warehousing solutions and big data modeling capabilities.

Key milestones

Time	Development
2010.04	Named ODPS, the service is released as an operational component of Alibaba Group's Ant Financial
2013.05	ODPS is released for beta testing
2013.07	ODPS v1.0 is released as a commercially available service. A single cluster contains 5,000 servers, with support for multi-level clusters available.
2016.09	Renamed MaxCompute, v2.0 is released as a commercially available service.