ALIBABA CLOUD

Alibaba Cloud

MaxCompute

Product Introduction

Document Version: 20220511

Alibaba Cloud

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 because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should 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 this document based on the "status quo", "being defective", and "existing functions" of its products and services. 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 take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, 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 directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
⚠ Danger	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.
🔔 Warning	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.
📢 Notice	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.
? Note	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]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.What is MaxCompute?

MaxCompute (formerly known as ODPS) is an enterprise-level cloud data warehouse that uses the software as a service (SaaS) model. MaxCompute is suitable for scenarios that require data analysis. It provides a fast, fully managed online data warehousing service in a serverless architecture. MaxCompute eliminates the constraints of traditional data platforms in terms of resource extensibility and elasticity, minimizes operations and maintenance (O&M) costs, and allows you to efficiently process and analyze large amounts of data at low costs.

As data collection techniques continue to diversify, enterprises in various industries accumulate terabytes, petabytes, or even exabytes of data. The rapid increase in the data amount exceeds the processing capacity of the traditional software industry. MaxCompute provides offline and streaming data access, supports large-scale data computing and query acceleration, and provides data warehousing solutions and analysis and modeling services for a variety of computing scenarios. MaxCompute also provides comprehensive data import solutions and various typical distributed computing models. It allows you to complete big data analytics without knowledge about distributed computing and maintenance.

MaxCompute is suitable for scenarios in which more than 100 GB of data needs to be stored or computed. MaxCompute can process up to exabytes of data and is widely used in Alibaba Group. MaxCompute is suitable for various big data processing scenarios, such as data warehousing and business intelligence (BI) analysis for large Internet enterprises, website log analysis, e-commerce transaction analysis, and exploration of user characteristics and interests.

MaxCompute is deeply integrated with the following Alibaba Cloud services:

- DataWorks

  DataWorks provides a variety of features, such as end-to-end data synchronization, workflow design, data development, data management, and O&M for MaxCompute.

- Machine Learning Platform for AI

  The algorithm components of PAI can be used to train models based on data in MaxCompute.

- Quick BI

  Quick BI allows you to create reports for data in MaxCompute and analyze the data in a visualized manner.

Learning path

For more information about the concepts, basic operations, and advanced operations of MaxCompute, see MaxCompute Learning Path.

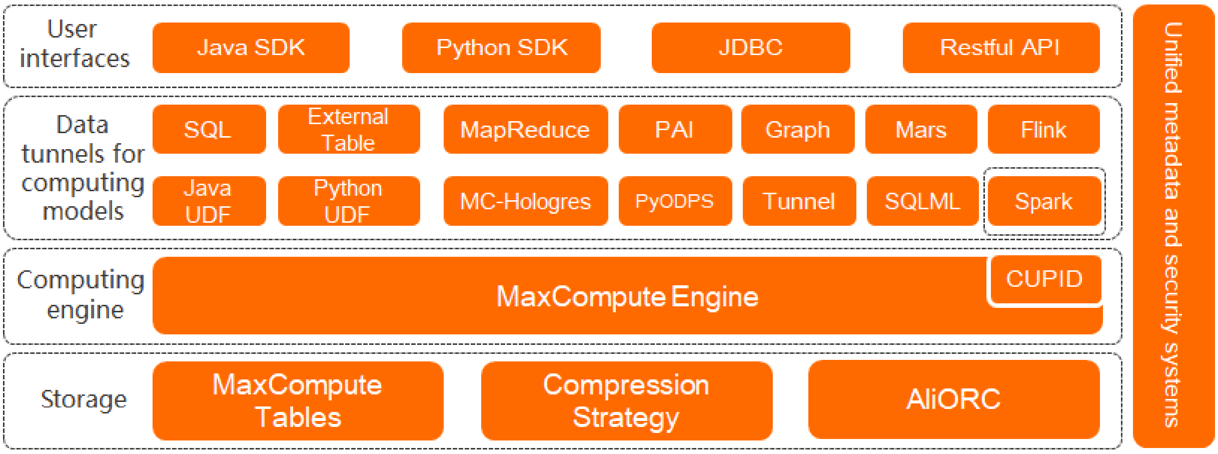
Core features

Feature	Description
Fully managed online data warehousing service in a serverless architecture	• Supports access over an API. The online service is an out-of-the-box service. • Provides a large number of cluster resources. You can purchase resources on demand by using the pay-as-you-go billing method. • Is O&M-free. The O&M cost is minimized.

Feature	Description
High elasticity and extensibility	- Separately extends storage and computing capabilities. MaxCompute allows you to analyze all data assets on the same platform. This way, data silos are eliminated. - Allocates resources based on the peaks and valleys of your business in real time.
Centralized, rich computing and storage capabilities	- Supports multiple computing models and a variety of user-defined functions (UDFs). - Supports column compression, which reduces the data size to 20% of the original size in most cases. This way, storage costs are significantly reduced.
Deep integration with DataWorks	Integrates with DataWorks, an end-to-end data development and data governance platform. DataWorks enables global data aggregation, processing, and governance. DataWorks can be used to manage MaxCompute projects and edit web-side query code.
Integrated AI capability	- Seamlessly integrates with PAI, which provides powerful machine learning capabilities. - Allows you to use Spark ML for BI analysis. - Uses third-party Python libraries for machine learning.
Deep integration with a Spark engine	- Provides a built-in Apache Spark engine, which supports all Spark features. - Deeply integrates the computing resources, data, and permission systems of MaxCompute into the Spark engine.
Lakehouse	- Integrates with data lakes such as Object Storage Service (OSS) and Hadoop Distributed File System (HDFS). MaxCompute allows you to analyze data in data lakes by using external tables. You can also use Spark to directly access data lakes and analyze data in the data lakes. - Supports association analysis between a data lake and a data warehouse based on a set of data warehousing services and user interfaces. For more information, see Lakehouse of MaxCompute.
Streaming data collection and near real-time analysis	- Allows you to write streaming data in real time and analyze the data in a data warehouse. - Deeply integrates with major streaming services in the cloud to read streaming data from various sources. - Supports elastic, parallel queries in the scale of seconds to meet requirements for near real-time analysis.
Continuous SaaS-based data protection in the cloud	Provides enterprises with three levels of more than 20 security features, such as infrastructure, data center, network, power supply, and platform security capabilities, user permission management, and privacy protection. MaxCompute also provides the same security capabilities as open source big data services and managed databases.

Architecture

The following figure shows the architecture of MaxCompute.



Module	Description
Storage	• MaxCompute tables: Table is the data storage unit of MaxCompute. Tables are the input and output objects of all types of jobs in MaxCompute. • Compression strategy: MaxCompute supports column compression, which reduces the data size to 20% of the original size in most cases. • AliORC: The data storage format of MaxCompute is upgraded to AliORC for higher storage performance.
Compute engine	MaxCompute supports various compute engines. MaxCompute runs Spark jobs on the Cupid platform developed by Alibaba Cloud. The Cupid platform is fully compatible with the computing framework that is supported by open source YARN.
	MaxCompute supports various data tunnels, which can meet your requirements in different scenarios: • SQL: MaxCompute supports SQL queries. You can use MaxCompute as traditional database software. However, MaxCompute is far more powerful than traditional database software and is capable of processing up to exabytes of data. Note ○ MaxCompute SQL does not support transactions or indexes. ○ The SQL syntax of MaxCompute is different from the SQL syntax of Oracle or MySQL. You cannot seamlessly migrate SQL statements from other databases to MaxCompute. ○ You can use MaxCompute to compute more than 100 GB of data. MaxCompute SQL can return query results in minutes or seconds, but not in milliseconds. ○ MaxCompute SQL is easy to use. To use MaxCompute SQL, you do not need to understand complex distributed computing concepts. If you have experience in database operations, you can familiarize yourself with MaxCompute SQL within a short period of time.

Module	Description
Data tunnels for computing models	● External Table: You can use external tables to process data that is stored outside MaxCompute. You can execute a simple DDL statement to create an external table in MaxCompute. This external table is associated with an external data source. ● Java UDFs: If the built-in functions of MaxCompute cannot meet your computing requirements, you can use Java to build UDFs. ● Python UDF: If the built-in functions of MaxCompute cannot meet your computing requirements, you can use Python to build UDFs. ● Overview: MaxCompute provides a Java MapReduce programming model, which can simplify the development process and improve development efficiency. ● Hologres: Hologres seamlessly integrates with MaxCompute at the underlying layer. This allows you to use standard PostgreSQL statements to query and analyze large amounts of data in MaxCompute without the need to migrate data. This way, the amount of time that is required to obtain query results is reduced. ● PAI: a machine learning algorithm platform based on MaxCompute. PAI provides an end-to-end machine learning platform for data processing, model training, service deployment, and prediction without the need for data migration. ● PyODPS: MaxCompute SDK for Python. It provides easy-to-use Python programming interfaces. ● Graph: an iterative graph computing and processing framework. ● Tunnel: a service that supports highly concurrent data uploads and downloads. ● Mars: a tensor-based unified distributed computing framework. Mars can use parallel and distributed computing technologies to accelerate data processing for Python data science stacks. ● SQLML: SQLML depends on MaxCompute and PAI. You can develop MaxCompute SQLML jobs on a client, learn MaxCompute data by using PAI, and then use machine learning models to make predictions. Then, use these results to guide your business planning. ● Flink: Flink provides real-time data processing capabilities for MaxCompute. ● Spark on MaxCompute: a computing service that is provided by MaxCompute and compatible with open source Spark. This service provides a Spark computing framework based on unified computing resource and dataset permission systems. The service allows you to use your preferred development method to submit and run Spark jobs. Spark on MaxCompute can fulfill diverse data processing and analysis requirements.
User interfaces	MaxCompute provides the following user interfaces: ● Java SDK ● Python SDK ● JDBC ● Restful API

Module	Description
Unified metadata and security systems	The Information Schema service of MaxCompute provides information such as project metadata and historical data. You can analyze job metrics such as the resource usage, job execution duration, and size of processed data to optimize jobs or plan resource capacity. MaxCompute also provides comprehensive security management systems, such as access control, data encryption, and dynamic data masking systems, to ensure data security. For more information about security, see Security features.

Benefits

MaxCompute has the following benefits:

- Ease-of-use
  - Helps you build a data warehouse that delivers high-performance storage and computing.
  - Pre-integrates multiple services, which simplifies standard SQL development.
  - Provides comprehensive management and security capabilities.
  - Is O&M-free and supports the pay-as-you-go billing method. You are charged only for the resources that you use.
- High scalability to meet business requirements

  Supports separate extension of storage and computing capabilities. The dynamic scaling feature frees you from planning capacity in advance and can meet the storage and computing requirements of rapid business growth.
- Various analysis scenarios

  Uses an open, unified platform to meet business requirements in various scenarios, such as data warehousing, BI, near real-time analysis, data lake analysis, and machine learning.
- Open platform
  - Supports open interfaces and data ecosystems, which ensures flexible data migration, application migration, and custom software development.
  - Supports flexible combination with commercial or open source services, such as Airflow and Tableau, to build various data applications.

Contact us

If you have questions or suggestions about MaxCompute, you can fill in the DingTalk group application form to join the DingTalk group for feedback.

2.FAQ

This topic provides answers to some frequently asked questions about MaxCompute.

What professional skills are required to use MaxCompute?

MaxCompute supports various data tunnels of computing models to meet your business requirements in different scenarios. To use MaxCompute for data analysis, you need only to be capable of using programming languages, such as SQL, Python, and Java.

Does MaxCompute provide an effective method to monitor business data?

MaxCompute allows you to configure data monitoring rules only by using the Data Quality feature of DataWorks. MaxCompute cannot monitor the changes in the fields of external data sources.

What is the role of a MaxCompute project?

A project is a basic organizational unit of MaxCompute. Similar to a database or schema in a traditional database system, a project is used to isolate users and control access requests. A project contains multiple objects, such as tables, resources, functions, and instances. You can have permissions to manage multiple projects. You can access objects of another project from your project after security authorization.

How do I obtain an AccessKey pair in MaxCompute?

You can go to the AccessKey Pair page to create or query an AccessKey pair.

If the AccessKey pair of the current account is disabled and a new AccessKey pair is created, are the auto triggered nodes that are created by using the previous AccessKey pair affected?

Yes, the auto triggered nodes that are created by using the previous AccessKey pair are affected. If you disable or delete an AccessKey pair, nodes in your DataWorks workspace fail to be run. Proceed with caution.

Does MaxCompute automatically compress data when I create a MaxCompute table? Can I specify the compression format and storage format?

Yes, MaxCompute automatically compresses data at a ratio of 3:1 to 5:1. The default storage format is AliORC and cannot be changed.

Which types of tables are supported by MaxCompute?

Internal tables and external tables are supported by MaxCompute. MaxCompute V2.0 and later support external tables.

- Data of internal tables is stored in MaxCompute. Data types of columns in internal tables can be any data types that are supported by MaxCompute.
- Data of external tables is not stored in MaxCompute. The data can be stored in Object Storage Service (OSS) or Tablestore. MaxCompute records only metadata of external tables. You can use external tables of MaxCompute to process unstructured data that is stored in OSS or Tablestore. The unstructured data includes video, audio, genetic, meteorological, or geographic data.

To accomplish tasks by using user-defined functions (UDFs) or MapReduce, what resources do I need to use?

- UDF: After you write a UDF, you must package it into a JAR file and upload the file to MaxCompute as a resource. When you run the UDF, MaxCompute automatically downloads the JAR file and obtains the code to run the UDF. JAR files are a type of MaxCompute resource. When you upload a JAR file, a resource is created in MaxCompute.
- MapReduce: After you write a MapReduce program, you must package it into a JAR file and upload the file to MaxCompute as a resource. When you run the MapReduce program, the MapReduce framework automatically downloads the JAR file and obtains the code to run the MapReduce program.

You can upload text files and MaxCompute tables to MaxCompute as different types of resources. Then, you can read or use these resources when you run UDFs or MapReduce programs.

How do I understand common error messages in MaxCompute and troubleshoot issues based on the messages?

Common error messages in MaxCompute are defined in the following standard format: `Error code: General description - Context-related description`. Common error messages of MaxCompute SQL, MapReduce, and Tunnel jobs are different. For more information about error messages, see Error code overview.

3.Definitions

3.1. Terms

This topic describes the terms and concepts used in MaxCompute. This helps you better understand MaxCompute before you use MaxCompute.

A

- AccessKey

  An AccessKey pair is a credential for accessing Alibaba Cloud APIs. An AccessKey pair consists of an AccessKey ID and an AccessKey secret. After you create an Alibaba Cloud account on the International site (alibabacloud.com), an AccessKey pair is generated on the AccessKey Management page. AccessKey pairs are used to identify users and verify the signature of requests for accessing MaxCompute or other Alibaba Cloud services, or connecting to third-party tools. Keep your AccessKey secret confidential to prevent credential leaks. If the AccessKey secret is accidentally leaked, disable or update your AccessKey secret immediately.

- authorization

  Authorization allows a project administrator or project owner to grant permissions on MaxCompute objects to other users. After authorization, these users can perform specific operations on MaxCompute objects. For example, these users can read, write, and view objects, such as tables, tasks, and resources. For more information about authorization, see Permission overview.

C

- console

  A MaxCompute client that runs on Windows or Linux. The MaxCompute client allows you to run commands to perform operations, such as project management operations, DDL operations, and DML operations. For more information about how to use the MaxCompute client, see MaxCompute client.

D

- data type

  The types of data in the columns of a MaxCompute table. For more information about MaxCompute data type editions and the data types supported by each edition, see Data type editions.

- DDL

  DDL operations, such as create a table or view. For more information about DDL syntax, see DDL statements.

- DML

  DML operations, such as INSERT, UPDATE, and DELETE operations. For more information about DML syntax, see DML statements.

F

- function

  Functions provided by MaxCompute include built-in functions and user-defined functions (UDFs). For more information about functions, see Function.

I

- instance

  Instances are used to run jobs. For more information, see Task instance.

J

- Job Scheduler

  Job Scheduler is a module in the kernel of the Apsara distributed operating system. Job Scheduler is used to manage resources and schedule jobs. Job Scheduler also provides a basic programming framework for application development. Job Scheduler serves as the underlying task scheduling module of MaxCompute.

M

- MapReduce

  MapReduce is a programming model for data processing. MapReduce is used for parallel operations on large datasets. You can use the Java API provided by MapReduce to write MapReduce programs and process MaxCompute data. The idea of MapReduce is to classify data processing methods as Map and Reduce. The Map method is used for the mapping of data and the Reduce method is used for the combination of data.

  Before you perform the Map operation, make sure that the input data is sliced into data blocks of equal size. Each data block is processed as the input to a single Map worker node. This way, multiple Map worker nodes can work at the same time. Each Map worker node processes an input data block and generates the intermediate result to a Reduce worker node. Then, the Reduce worker node combines the outputs of multiple Map worker nodes to obtain the final result. For more information, see MapReduce.

O

- ODPS

  ODPS is the original name of MaxCompute.

P

- partition

  A partition is a division of a table based on the partition key, which consists of one or more partition key columns. Partitions are used to divide the data stored in a table. If a table is not partitioned, data in the table is stored in the directory that stores the table. If a table is partitioned, each partition corresponds to a subdirectory in the directory that stores the table. In this case, data is stored in separate subdirectories. For more information about partitions, see Partition.

- project

  A project is a basic organizational unit of MaxCompute. Similar to a database or schema in a traditional database system, a project is used to isolate users and control access requests. For more information about projects, see Project.

Q

- quota

Quota serves as a computing resource pool of MaxCompute. Quotas provide computing resources that are required for running jobs. For more information about quotas, see Quota.

R

- role

  Role is a concept in the MaxCompute security feature. A role can be considered a set of users who have the same permissions. One user can assume multiple roles, and multiple users can assume the same role. After you grant permissions to a role, all users who are assigned this role are granted the same permissions. For more information about how to manage roles, see Role planning and management.

- resource

  Resource is a special concept of MaxCompute. You must have the required resources to implement UDFs and MapReduce operations in MaxCompute. For more information about resources, see Resource.

S

- SDK

  A Software Development Kit (SDK) 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, or document packages. MaxCompute supports SDK for Java and SDK for Python.

- sandbox

  A sandbox is an isolated environment to restrict program actions based on security policies. A sandbox serves as a security mechanism to isolate Java code execution in a separate environment and restrict malicious code from accessing local system resources. This prevents damage to the local system. MaxCompute MapReduce and UDFs that run in a distributed environment are restricted by Java sandbox.

- security

  The MaxCompute multi-tenant data security system provides features, such as user authentication, user and permission management, resource sharing across projects, and project data protection. For more information about the security management operations of MaxCompute, see Permission overview.

T

- table

  In MaxCompute, tables are used to store data. For more information about tables, see Table.

- Tunnel

  Tunnel is a data channel in MaxCompute. Tunnel provides highly concurrent offline data uploads and downloads. You can use MaxCompute Tunnel to upload data in batches to MaxCompute or download data in batches to your on-premises machine. For more information about related commands, see Tunnel commands or MaxCompute Tunnel SDK.

U

- UDF

In a broad sense, UDFs include user-defined scalar functions, user-defined aggregate functions (UDAFs), and user-defined table-valued functions (UDTFs). MaxCompute allows you to develop UDFs in Java or Python. For more information, see MaxCompute UDF.

In a narrow sense, UDFs refer to only user-defined scalar functions. The input and output data of a UDF have a one-to-one mapping relationship, which indicates that one value is returned every time a UDF reads one row of data.

- UDAF

  The input and output data of a UDAF have a many-to-one mapping relationship. Multiple input records are aggregated to generate one output value. UDAFs can be used with the GROUP BY clause of SQL statements. For more information about UDAFs, see UDAF.

- UDTF

  Only UDTFs can return multiple fields. For more information about UDTFs, see UDTF.

- user

  User is a concept in the MaxCompute security feature. You can access MaxCompute by using an Alibaba Cloud account, a RAM user, or a user who is assigned a RAM role. All users, except the project owner, must be added to a MaxCompute project and granted the related permissions to manage data, jobs, resources, and functions in MaxCompute. For more information about how to manage users, see User planning and management.

V

- view

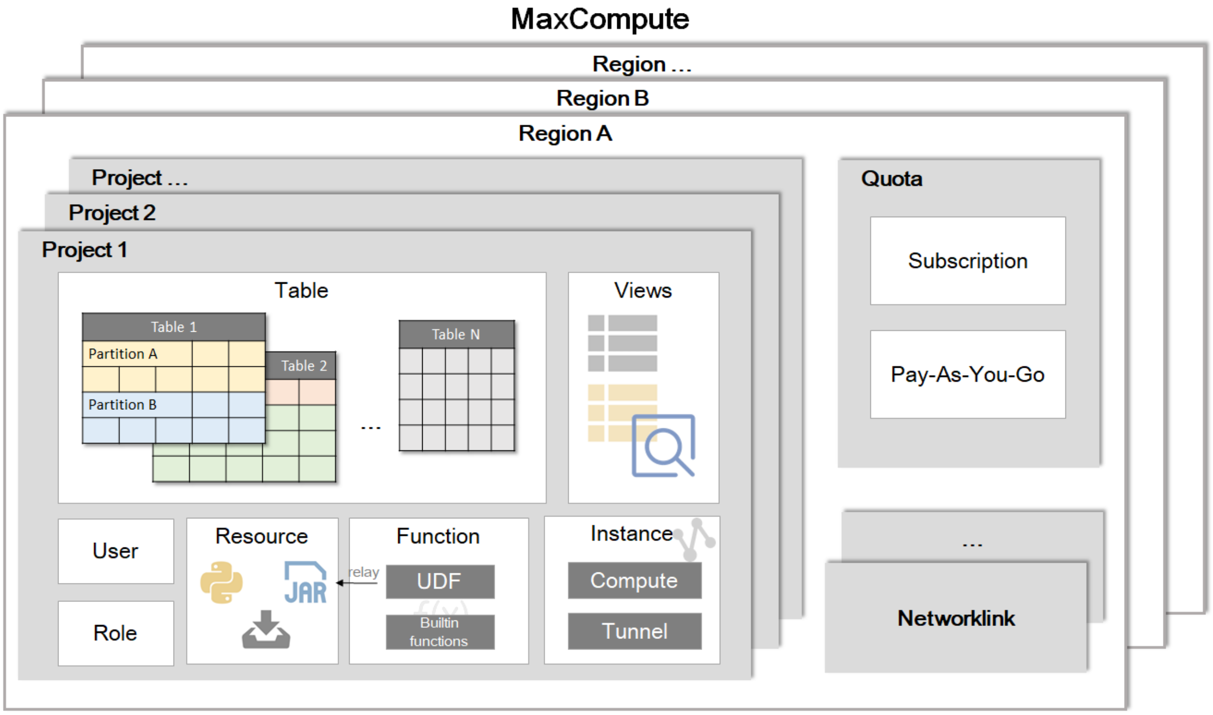
  A view is a virtual table that is created based on existing tables. Its schema and content are derived from these tables. A view corresponds to one or more tables. You can use views if you want to retain query results without the need to create additional tables. For more information about views, see View-related operations.

3.2. Core concepts

3.2.1. Concept hierarchy

MaxCompute introduces a concept hierarchy that can inspire you to put forward ideas for subsequent project planning and security management. This topic introduces the hierarchy and brief definitions of core concepts in MaxCompute.

The following figure shows the hierarchy of core concepts in MaxCompute.



Core concept	Description
Project	A project is a basic organizational unit of MaxCompute. Similar to a database or schema in a traditional database system, a project is used to isolate users and control access requests. For more information about projects, see Project.
Table	In MaxCompute, tables are used to store data. For more information about tables, see Table.
Partition	A partition is a division of a table based on the partition key, which consists of one or more partition key columns. Partitions are used to divide the data stored in a table. If a table is not partitioned, data in the table is stored in the directory that stores the table. If a table is partitioned, each partition corresponds to a subdirectory in the directory that stores the table. In this case, data is stored in separate subdirectories. For more information about partitions, see Partition.
View	A view is a virtual table that is created based on existing tables. Its schema and content are derived from these tables. A view corresponds to one or more tables. You can use views if you want to retain query results without the need to create additional tables. For more information about views, see View-related operations.

Core concept	Description
User	User is a concept in the MaxCompute security feature. You can access MaxCompute by using an Alibaba Cloud account, a RAM user, or a user who is assigned a RAM role. All users, except the project owner, must be added to a MaxCompute project and granted the related permissions to manage data, jobs, resources, and functions in MaxCompute. For more information about how to manage users, see User planning and management.
Role	Role is a concept in the MaxCompute security feature. A role can be considered a set of users who have the same permissions. One user can assume multiple roles, and multiple users can assume the same role. After you grant permissions to a role, all users who are assigned this role are granted the same permissions. For more information about how to manage roles, see Role planning and management.
Resource	Resource is a special concept of MaxCompute. You must have the required resources to implement UDFs and MapReduce operations in MaxCompute. For more information about resources, see Resource.
Function	Functions provided by MaxCompute include built-in functions and user-defined functions (UDFs). For more information about functions, see Function.
Instance	Instances are used to run jobs. For more information, see Task instance.
Quota	Quota serves as a computing resource pool of MaxCompute. Quotas provide computing resources that are required for running jobs. For more information about quotas, see Quota.
Network link	Before you use external tables, UDFs, or the lakehouse solution, you must establish network links between MaxCompute and other services in a virtual private cloud (VPC) or over the Internet. This way, MaxCompute can access services, such as HBase, RDS, and Hadoop in a VPC or over the Internet. For more information about network links, see Network connection process.

3.2.2. Project

A project is a basic organizational unit of MaxCompute. A project in MaxCompute is similar to a database or schema in a traditional database management system. Projects are used to isolate users and manage access requests. A project contains multiple objects, such as tables, resources, functions, and instances.

You can have permissions to manage multiple projects. You can access objects of another project from your project after relevant security authorization. For more information, see Cross-project resource access based on packages.

You can run the **use project;** command to enter a project. The following command shows an example:

```
-- Enter a project named my_project.
use my_project;
```

After you run the preceding command, you can enter a project named my_project and manage objects such as tables, resources, functions, and instances in the project. The **use project;** command is provided by the MaxCompute client. For more information about other commands provided by MaxCompute, see Common MaxCompute commands.

Note A project in MaxCompute is associated with a workspace in DataWorks. For more information, see Basic mode and standard mode.

3.2.3. Quota

Quota serves as a computing resource pool of MaxCompute. It provides the CPU and memory resources that are required for computing jobs, such as MaxCompute SQL jobs, MapReduce jobs, Spark jobs, Mars jobs, and Machine Learning Platform for AI (PAI) jobs.

The unit of MaxCompute computing resources is compute unit (CU). One CU equals 1 CPU core and 4 GB of memory. Quotas are classified into subscription resource quotas and pay-as-you-go resource quotas based on the billing methods of resources. For more information about the billing methods, see Subscription标准版.

If you purchase subscription resource quotas, you can perform fine-grained management for these quotas. For more information, see Use MaxCompute Management.

- Configure quota groups

  You can add, change, or remove quota groups. You can also configure resource scheduling periods for quota groups. This ensures that different projects can schedule computing resources in different periods of time.

- Change the quota group of a project

  You can change the quota groups that are associated with a MaxCompute project.

You can associate MaxCompute projects with quota groups by using the following methods. After projects are associated with quota groups, the computing jobs that you submit in the projects use the quota groups that are associated with the projects to compute data.

- When you create a MaxCompute project, you can configure the **quota group** parameter to specify the quota group that you want to associate with the project.
- To change the quota group that is associated with an existing MaxCompute project, you can go to the MaxCompute console, find the project, and then click **Switch quota group** in the Actions column. You can also change the quota group that is associated with the project on the **Projects** tab of MaxCompute Management. For more information, see Change the quota group of a project.

Note We recommend that you associate different quota groups with different MaxCompute projects based on your business requirements.

3.2.4. Table

Tables are the units that are used to store data in MaxCompute. Logically, a table is a two-dimensional structure that consists of rows and columns. Each row represents a record. Each column represents a field whose values are of the same data type. One record can contain one or more columns. The column names and data types constitute the schema of the table.

Tables are the input and output objects of all computing tasks in MaxCompute. You can create a table, delete a table, and import data to a table. For more information, see Table operations.

> **Note** The Data Map module of DataWorks allows you to create and organize MaxCompute tables, manage data lifecycles, modify table schemas, and manage permissions on tables, resources, or functions.

MaxCompute V2.0 or later supports internal tables and foreign tables.

- Data of internal tables is stored in MaxCompute. Columns in internal tables can be of any data types that are supported by MaxCompute.
- Data of foreign tables is not stored in MaxCompute. Instead, the data can be stored in Object Storage Service (OSS) or Tablestore. MaxCompute records only metadata of foreign tables. You can use foreign tables of MaxCompute to process unstructured data that is stored in OSS or Tablestore, such as video, audio, genetic, meteorological, or geographic data.

3.2.5. Partition

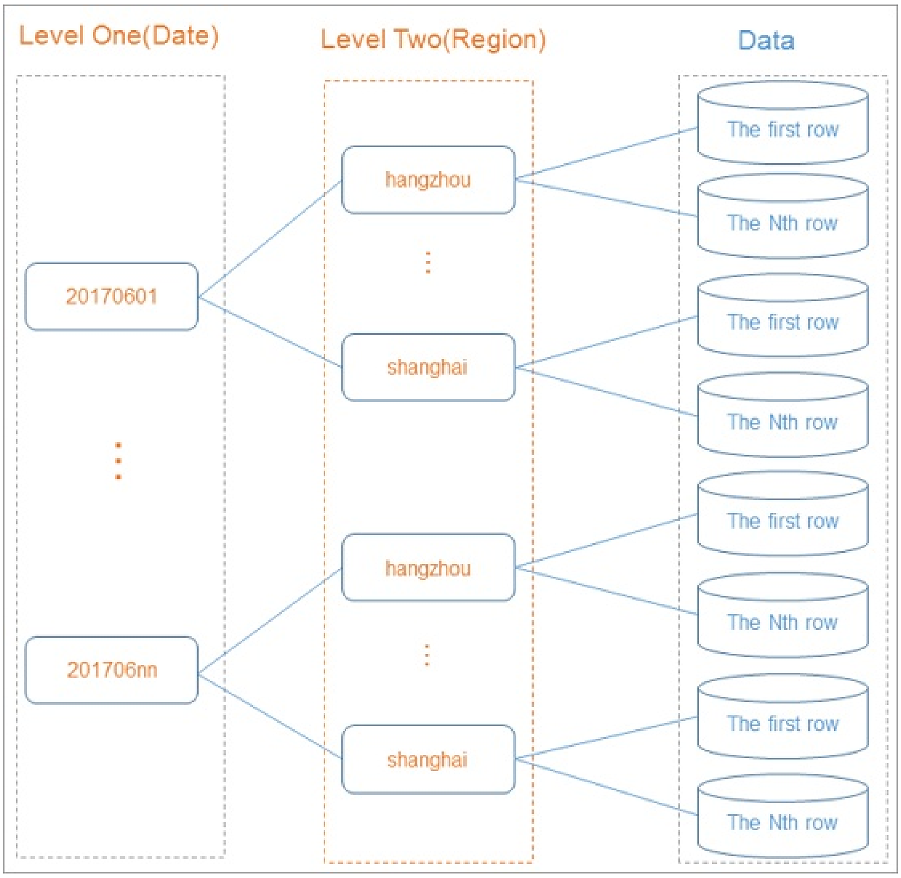
A partitioned table is a table with partitions. You can specify one or more columns as partition key columns to create a partitioned table. Partitioned tables are similar to individual directories in a distributed file system. A partition is similar to a directory and all data in the partition is similar to all data files under the directory.

Overview

To partition a table is to classify data of the same category into the same partition. The classification is based on the partition key, which can consist of one or more primary key columns in the table.

In MaxCompute, each value in a partition key column is specified as a partition. You can specify multi-level partitions with multiple partition key columns. Multi-level partitions are similar to multi-level directories in structure.

Partitioned tables improve query efficiency. You can specify the name of the partition that you want to query by using the WHERE clause. This way, MaxCompute scans only the specified partition, which improves processing efficiency and reduces cost. If you specify the name of the partition that you want to access when you query the table, only the specified partition is read.



The execution of some SQL jobs for operations on partitions is less efficient and may incur higher costs. For more information, see Insert or overwrite data into dynamic partitions (DYNAMIC PARTITION).

The statements that are used to process partitioned and non-partitioned tables are different in MaxCompute. For more information, see Table operations and INSERT OVERWRITE and INSERT INTO.

Limits

- A table can contain up to six levels of partitions.
- A table can contain up to 60,000 partitions.
- Up to 10,000 partitions can be queried at a time.
- The values in a partition key column of the STRING type cannot contain Chinese characters.

Data types of partition key columns

MaxCompute V2.0 supports partition key columns of the TINYINT, SMALLINT, INT, BIGINT, VARCHAR, and STRING types.

MaxCompute V1.0 supports only partition key columns of the STRING type. You can specify the data type of a partition key column as BIGINT. However, only the partition key column is of the BIGINT type. All the data in the partition key column is processed as a string in operations, such as the calculation and comparison of data in partition key columns. In the following example, the return result of the statements contains only one row. This is because 10 is treated as a string to be compared with 2 and the row where the value of pt is 10 is not returned.

```
--- Create a table named parttest.
CREATE TABLE parttest (a bigint) PARTITIONED BY (pt bigint);
--- Insert data into the parttest table.
INSERT INTO parttest partition(pt) (a,pt) values (1, 1);
INSERT INTO parttest partition(pt) (a,pt) values (1, 10);
--- Query the rows where the value of pt is greater than or equal to 2.
SELECT * FROM parttest WHERE pt >= '2';
```

Examples

- Create a partition.

```
-- Create a partitioned table that contains two levels of partitions. In the partitioned
table, pt is used as a level-1 partition key column and region is used as a level-2 parti
tion key column.
CREATE TABLE src (shop_name string, customer_id bigint) PARTITIONED BY (pt string,region
string);
```

- Use the values in partition key columns as filter conditions to query a table.

```
-- The following example shows a correct usage. When MaxCompute generates a query plan, o
nly the data whose region is 'hangzhou' in the '20170601' partition is used as input data
.
select * from src where pt='20170601'and region='hangzhou';
-- The following example shows an incorrect usage. In this example, the effectiveness of
the partition filtering cannot be ensured. Data in the pt partition key column is conside
red as a string. When a value of the STRING type is compared with a value of the BIGINT t
ype, 20170601 in this example, MaxCompute converts both data types to DOUBLE, which cause
s a loss in precision.
select * from src where pt = 20170601;
```

3.2.6. Lifecycle

This topic describes the lifecycle of a MaxCompute table.

The lifecycle of a MaxCompute table or a partition starts at the time when the data in the table or partition was last updated. If the data remains unchanged for a specific period of time, MaxCompute automatically reclaims the table or partition. The period of time for which the data remains unchanged is the lifecycle of the table or the partition.

- The value of a lifecycle is a positive integer. Unit: days.
- If the data in a non-partitioned table remains unchanged within the lifecycle of the table, MaxCompute automatically executes a statement that is similar to DROP TABLE to reclaim the table. The lifecycle of a non-partitioned table starts from the time specified by LastDataModifiedTime. LastDataModifiedTime specifies the time when the data in a non-partitioned table is last modified.
- The partitions of a table can be separately reclaimed. MaxCompute automatically reclaims partitions

whose data remains unchanged within the lifecycle. The lifecycle of a partition starts from the time specified by LastDataModifiedTime. LastDataModifiedTime specifies the time when the data in a partition is last modified. Unlike non-partitioned tables, a partitioned table is not deleted even if all of its partitions have been reclaimed.

> ⓘ Note
>
> - A lifecycle-based table scan is performed at a scheduled time each day to scan all the partitions of a table. A partition can be reclaimed only if the period after the time specified by LastDataModifiedTime exceeds the lifecycle.
>
>   For example, the lifecycle of a partitioned table is one day and the data in one of the table partitions was last modified at 15:00 on February 17, 2020. If MaxCompute scans this table before 15:00 of February 18, 2020, the table partition is not reclaimed because the period after the time specified by LastDataModifiedTime is less than the one-day lifecycle. If MaxCompute scans this table on February 19, 2020, the table partition is reclaimed because the period after the time specified by LastDataModifiedTime exceeds the one-day lifecycle.
>
> - The lifecycle feature allows MaxCompute to periodically reclaim a table or a partition. The availability of the system determines whether MaxCompute can immediately reclaim a table or a partition when the period after the time specified by LastDataModifiedTime exceeds the lifecycle of the table or the partition. Therefore, MaxCompute cannot always reclaim a table or a partition immediately after the lifecycle elapsed.
> - After a table is dropped, all properties of the table are dropped, including the lifecycle. If you create another table that has the same name as the dropped table, the lifecycle properties of the table that you created prevail.

- You can specify a lifecycle for tables. You cannot specify a lifecycle for partitions. The lifecycle that you specify for a partitioned table applies to all partitions of this table. You can specify a lifecycle when you create a table.
- If you do not specify a lifecycle for a table, the table will not be automatically reclaimed based on lifecycle rules.

For more information about how to specify and modify the lifecycle of a table or how to configure the `LastDataModifiedTime` parameter of a table, see Table operations.

3.2.7. Resource

This topic describes the concept of resource and the types of resources in MaxCompute. Resources are used when you perform some specific operations.

Concept

Resource is a concept specific to MaxCompute. To accomplish tasks by using user-defined functions (UDFs) or MapReduce, you must upload files as MaxCompute resources.

- SQL UDF: After you write a UDF, you must compress the code of the UDF into a JAR package and upload the package to MaxCompute as a resource. When you execute the UDF, MaxCompute automatically downloads the JAR package and obtains the code in the package to execute the UDF. JAR files are a type of MaxCompute resource. When you upload a JAR file, a resource is created in MaxCompute.
- MapReduce: After you write a MapReduce program, you must compress the program into a JAR

package and upload the package to MaxCompute as a resource. When you run a MapReduce job, MapReduce automatically downloads the JAR package and obtains the code in the package to run the MapReduce job.

You can upload text files and MaxCompute tables to MaxCompute as resources. Then, you can access or use the resources when you execute UDFs or run MapReduce jobs. MaxCompute provides APIs that you can call to access and use resources. For more information, see Resource samples and Java UDFs.

> ? **Note** Some limits are imposed on how MaxCompute UDFs and MapReduce access resources. For more information, see Limits.

Resource types

You can upload an object whose size is no more than 500 MB to MaxCompute as a resource. MaxCompute supports the following types of resources:

- File: files in the .zip, .so, or .jar format.
- Table: tables in MaxCompute.

  > ? **Note** Only BIGINT, DOUBLE, STRING, DATETIME, and BOOLEAN fields are supported in tables that are referenced by MapReduce.

- JAR: compiled JAR packages.
- Archive: compressed files that are identified by the resource name extension. The following file types are supported: .zip, .tgz, .tar.gz, .tar, and .jar.
- Python: the Python code that you write. You can use Python code to register Python UDFs.

For more information about resource-related operations, see Resource operations or MaxCompute resources.

3.2.8. Function

This topic describes the built-in functions and user-defined functions (UDFs) that MaxCompute provides.

MaxCompute provides SQL computing capabilities. You can use the built-in functions in MaxCompute SQL statements to complete some computing and counting tasks. If the built-in functions do not meet your requirements, you can use the Java APIs that MaxCompute provides to develop UDFs.

UDFs can be classified into scalar-valued functions, user-defined aggregate functions (UDAFs), and user-defined table functions (UDTFs).

After you develop a UDF, you must compile the UDF code to a JAR package, upload the package to MaxCompute as a resource, and then register the UDF in MaxCompute.

> ? **Note** To use a UDF in MaxCompute, you only need to specify its name and parameters in an SQL statement as you do when you use the built-in functions of MaxCompute.

For more information about how to manage functions, see Create a function, Delete a function, and List functions.

3.2.9. Task

A task is the basic computing unit of MaxCompute. Computing jobs such as those involving SQL and MapReduce are completed by using tasks.

MaxCompute parses most of the tasks that you submit, especially computing tasks such as SQL DML statements and MapReduce tasks. Then, MaxCompute generates task execution plans based on the parsing results. An execution plan consists of several mutually dependent stages.

The execution plan can be logically defined as a directed graph. In this graph, vertices represent stages, and edges represent dependencies between the stages. MaxCompute executes the stages based on the dependencies in the graph (execution plan). A single stage comprises multiple threads, also known as workers. The workers in each stage work together to complete computing for the stage. Different workers in the same stage process different data but run on the same execution logic. A computing task is instantiated as an instance when it is executed. You can perform operations on the instance. For example, you can run the status command to query the instance status and the kill command to terminate the instance.

Some MaxCompute tasks, such as SQL DDL statements, are not computing tasks. These tasks only need to read and modify metadata in MaxCompute. MaxCompute does not generate execution plans for these tasks.

Note MaxCompute does not convert all the requests to tasks. For example, operations on projects, resources, user-defined functions (UDFs), and instances are not processed as tasks.

3.2.10. Task instance

This topic describes task instances in MaxCompute and the status that a task instance may have throughout its lifecycle.

When you start to run SQL, Spark, or MapReduce tasks in MaxCompute, the tasks are instantiated as instances. The lifecycle of task instances includes two stages: Running and Terminated.

Task instances that are in the Running stage are in the Running state. Task instances that are in the Terminated stage can be in the Success, Failed, or Canceled state. You can query or change the status of a task instance based on the ID that MaxCompute assigns to the task instance. Examples:

```
-- Query the status of a task instance.
status instance_id;
-- Terminate a task instance. After you run the kill command, the status of the task instan
ce changes to Canceled.
kill instance_id;
-- View the operational logs of a task instance.
wait instance_id;
```

In the preceding commands, instance_id indicates the ID of the task instance that you want to manage. Replace this parameter with the actual instance ID.

3.3. ACID semantics

This topic describes the atomicity, consistency, isolation, and durability (ACID) semantics for concurrent jobs in MaxCompute, and the ACID semantics for transactional tables.

Terms

- Operation: a single job submitted in MaxCompute.
- Data object: an object that stores data, such as a non-partitioned table or a partition.
- INTO job: an SQL job that contains the INTO keyword, such as INSERT INTO or DYNAMIC INSERT INTO.
- OVERWRITE job: an SQL job that contains the OVERWRITE keyword, such as INSERT OVERWRITE or DYNAMIC INSERT OVERWRITE.
- Data upload by using Tunnel: an INTO or OVERWRITE job.

Description of ACID semantics

- Atomicity: An operation is fully complete or not performed at all. That is, an operation is not partially performed.
- Consistency: The integrity of data objects is maintained when an operation is performed.
- Isolation: An operation can be performed independent of other concurrent operations.
- Durability: After an operation is complete, modified data is permanently valid and is not lost even if a system failure occurs.

ACID semantics for concurrent write jobs in MaxCompute

- Atomicity
  - If multiple jobs conflict with each other, MaxCompute ensures that only one job succeeds.
  - The atomicity of the CREATE, OVERWRITE, and DROP operations on a single table or partition can be ensured.
  - The atomicity of cross-table operations such as MULTI-INSERT cannot be ensured.
  - In extreme cases, the following operations may not be atomic:
    - A `DYNAMIC INSERT OVERWRITE` operation that is performed on more than 10,000 partitions.
    - An INTO operation. The atomicity of INTO operations cannot be ensured because data cleansing fails during a transaction rollback. However, the data cleansing failure does not cause loss of original data.
- Consistency
  - The consistency can be ensured for OVERWRITE jobs.
  - If an INTO job fails due to a conflict, data from the failed job may remain.
- Isolation
  - For non-INTO operations, MaxCompute ensures that read operations are submitted.
  - For INTO operations, some read operations may not be submitted.
- Durability
  - MaxCompute ensures data durability.

ACID semantics for transactional tables

In addition to the ACID semantics for concurrent write jobs, MaxCompute supports the following ACID semantics for transactional tables:

- For INTO operations, MaxCompute ensures that read operations are submitted. If an INTO job fails due to a conflict, data from the failed job does not remain.
- The atomicity of the UPDATE, DELETE, and small file MERGE operations on a non-partitioned table or a partition can be ensured.

For example, if two UPDATE operations are performed on a partition at the same time, only one UPDATE operation succeeds. The following cases do not exist: 1. An UPDATE operation is partially performed. 2. Both UPDATE operations succeed.

Conflict of concurrent operations

When jobs are concurrently performed on the same destination table, a conflict may occur. In the event of a conflict, the job that ends earlier succeeds, and the job that ends later may fail due to the conflict.

The following table describes the results of jobs that are submitted at the same time on a non-partitioned table or a partition.

Job type	INSERT OVERWRITE or TRUNCATE job that ends later	INSERT INTO job that ends later	UPDATE or DELETE job that ends later	Small file MERGE job that ends later
INSERT OVERWRITE or TRUNCATE job that ends earlier	• Both jobs succeed. • Data from the INSERT OVERWRITE or TRUNCATE job that ends later overwrites data from the INSERT OVERWRITE or TRUNCATE job that ends earlier.	• Both jobs succeed. • The INSERT INTO job that ends later appends its data to data from the INSERT OVERWRITE or TRUNCATE job that ends earlier.	• The UPDATE or DELETE job that ends later reports an error. • The INSERT OVERWRITE or TRUNCATE job that ends earlier modifies the data of the non-partitioned table or partition on which the UPDATE or DELETE job that ends later is performed.	• The small file MERGE job that ends later reports an error. • The INSERT OVERWRITE or TRUNCATE job that ends earlier modifies the data of the non-partitioned table or partition on which the small file MERGE job that ends later is performed.
INSERT INTO job that ends earlier	• Both jobs succeed. • Data from the INSERT OVERWRITE or TRUNCATE job that ends later overwrites data from the INSERT INTO job that ends earlier.	• Both jobs succeed. • The INSERT INTO job that ends later appends its data to data from the INSERT INTO job that ends earlier.	• The UPDATE or DELETE job that ends later reports an error. • The INSERT INTO job that ends earlier modifies the data of the non-partitioned table or partition on which the UPDATE or DELETE job that ends later is performed.	• The small file MERGE job that ends later reports an error. • The INSERT INTO job that ends earlier modifies the data of the non-partitioned or partition on which the small file MERGE job that ends later is performed.

Job type	INSERT OVERWRITE or TRUNCATE job that ends later	INSERT INTO job that ends later	UPDATE or DELETE job that ends later	Small file MERGE job that ends later
UPDATE or DELETE job that ends earlier	• Both jobs succeed. • Data from the INSERT OVERWRITE or TRUNCATE job that ends later overwrites data from the UPDATE or DELETE job that ends earlier.	• Both jobs succeed. • The INSERT INTO job that ends later appends its data to data from the UPDATE or DELETE job that ends earlier.	• The UPDATE or DELETE job that ends later reports an error. • The UPDATE or DELETE job that ends earlier modifies the data of the non-partitioned table or partition on which the UPDATE or DELETE job that ends later is performed.	• The small file MERGE job that ends later reports an error. • The INSERT INTO job that ends earlier modifies the data of the non-partitioned or partition on which the small file MERGE job that ends later is performed.
Small file MERGE job that ends earlier	• Both jobs succeed. • Data from the INSERT OVERWRITE or TRUNCATE job that ends later overwrites data from the small file MERGE job that ends earlier.	• Both jobs succeed. • The INSERT INTO job that ends later appends its data to data from the small file MERGE job that ends earlier.	• The UPDATE or DELETE job that ends later reports an error. • The small file MERGE job that ends earlier modifies the data of the non-partitioned table or partition on which the UPDATE or DELETE job that ends later is performed.	• The small file MERGE job that ends later reports an error. • The small file MERGE job that ends earlier modifies the data of the non-partitioned or partition on which the small file MERGE job that ends later is performed.

In conclusion, conflicting jobs succeed or report errors based on the following rules:

- INSERT operations do not report errors due to conflicts when data changes.
- The UPDATE, DELETE, and small file MERGE operations report errors due to conflicts when data in the destination non-partitioned table or partition changes.

> **Note** In extreme cases, if multiple jobs are concurrently performed when metadata is updating, the jobs may report errors due to conflicts caused by metadata changes.

4.Limits

Before you use MaxCompute, we recommend that you learn the limits on the use of MaxCompute. This topic describes the limits on the use of MaxCompute.

Limits on data upload and download

Before you upload or download data in MaxCompute, take note of the following limits:

-

For more information about data upload and download, see Data upload and download.

Limits on SQL

The following table describes the limits on the development of SQL jobs in MaxCompute.

Item	Maximum value/Limit	Category	Description
Table name length	128 bytes	Length	A table or column name can contain only letters, digits, and underscores (_). It must start with a letter. Special characters are not supported.
Comment length	1,024 bytes	Length	A comment is a valid string that cannot exceed 1,024 bytes in length.
Column definitions in a table	1,200	Quantity	A table can contain a maximum of 1,200 column definitions.
Partitions in a table	60,000	Quantity	A table can contain a maximum of 60,000 partitions.
Partition levels of a table	6	Quantity	A table can contain a maximum of six levels of partitions.
Output display	10,000 rows	Quantity	A SELECT statement can return a maximum of 10,000 rows.
Number of destination tables for `INSERT` operations	256	Quantity	The `MULTI-INSERT` statement allows you to insert data into a maximum of 256 tables at the same time.
`UNION ALL`	256	Quantity	The `UNION ALL` statement allows you to combine a maximum of 256 tables.
`MAPJOIN`	128	Quantity	A `MAPJOIN` hint allows you to join a maximum of 128 small tables.

Item	Maximum value/Limit	Category	Description
`MAPJOIN` memory	512 MB	Size	The memory size for all small tables cannot exceed 512 MB when you specify a `MAPJOIN` hint in SQL statements.
`ptinsubq`	1,000 rows	Quantity	A PT IN SUBQUERY statement can generate a maximum of 1,000 rows.
Length of an SQL statement	2 MB	Length	An SQL statement cannot exceed 2 MB in length. This limit is suitable for the scenarios in which you use an SDK to call SQL statements.
Conditions of a `WHERE` clause	256	Quantity	A `WHERE` clause can contain a maximum of 256 conditions.
Length of a column record	8 MB	Length	The maximum length of a column record in a table is 8 MB.
Parameters in an IN clause	1,024	Quantity	This item specifies the maximum number of parameters in an IN clause, such as `IN (1,2,3….,1024)`. If the number of parameters in an `IN` clause is too large, the compilation performance is affected. We recommend that you use a maximum of 1,024 parameters, but this is not a fixed upper limit.
`jobconf.json`	1 MB	Size	The maximum size of the `jobconf.json` file is 1 MB. If a table contains a large number of partitions, the size of the `jobconf.json` file may exceed 1 MB.
View	Not writable	Operation	A view is not writable and does not support the `INSERT` statements.
Data type and position of a column	Unmodifiable	Operation	The data type and position of a column cannot be modified.
Java user-defined functions (UDFs)	Not allowed to be `abstract` or `static`	Operation	Java UDFs cannot be `abstract` or `static`.
Partitions that can be queried	10,000	Quantity	A maximum of 10,000 partitions can be queried.

Item	Maximum value/Limit	Category	Description
SQL execution plans	1 MB	Size	The size of an execution plan that is generated by using MaxCompute SQL statements cannot exceed 1 MB. Otherwise, the error message `FAILED: ODPS-0010000:System internal error - The Size of Plan is too large` is reported.

For more information about SQL, see SQL.

Limits on MapReduce

The following table describes the limits on the development of MapReduce jobs in MaxCompute.

Item	Value range	Classification	Configuration item	Default value	Configurable	Description
Memory occupied by an instance	[256 MB,12 GB]	Memory	`odps.stage.mapper(reducer).mem` and `odps.stage.mapper(reducer).jvm.mem`	2,048 MB and 1,024 MB	Yes	The memory occupied by a single map or reduce instance. The memory consists of two parts: the framework memory, which is 2,048 MB by default, and Java Virtual Machine (JVM) heap memory, which is 1,024 MB by default.
Number of resources	256	Quantity	-	N/A	No	Each job can reference up to 256 resources. Each table or archive is considered as one resource.
Numbers of inputs and outputs	1,024 and 256	Quantity	-	N/A	No	The number of the inputs of a job cannot exceed 1,024, and that of the outputs of a job cannot exceed 256. A partition of a table is regarded as one input. The number of tables cannot exceed 64.
Number of counters	64	Quantity	-	N/A	No	The number of custom counters in a job cannot exceed 64. The counter group name and counter name cannot contain number signs (#). The total length of the two names cannot exceed 100 characters.

Item	Value range	Classification	Configuration item	Default value	Configurable	Description
Number of map instances	[1,100000]	Quantity	odps.stage.mapper.num	N/A	Yes	The number of map instances in a job is calculated by the framework based on the split size. If no input table is specified, you can set the odps.stage.mapper.num parameter to specify the number of map instances. The value ranges from 1 to 100,000.
Number of reduce instances	[0,2000]	Quantity	odps.stage.reducer.num	N/A	Yes	By default, the number of reduce instances in a job is 25% of the number of map instances. You can set the number to a value that ranges from 0 to 2,000. Reduce instances process much more data than map instances, which may result in long processing time in the reduce stage. A job can have 2,000 reduce instances at most.
Number of retries	3	Quantity	-	N/A	No	The maximum number of retries that are allowed for a map or reduce instance is 3. Exceptions that do not allow retries may cause jobs to fail.
Local debug mode	A maximum of 100 instances	Quantity	-	N/A	No	In local debug mode: • The number of map instances is 2 by default and cannot exceed 100. • The number of reduce instances is 1 by default and cannot exceed 100. • The number of download records for one input is 100 by default and cannot exceed 10,000.

Item	Value range	Classification	Configuration item	Default value	Configurable	Description
Number of times a resource is read repeatedly	64	Quantity	-	N/A	No	The number of times that a map or reduce instance repeatedly reads a resource cannot exceed 64.
Resource bytes	2 GB	Length	-	N/A	No	The total bytes of resources that are referenced by a job cannot exceed 2 GB.
Split size	Greater than or equal to 1	Length	odps.stage.mapper.split.size	256 MB	Yes	The framework determines the number of map instances based on the split size.
Length of a string in a column	8 MB	Length	-	N/A	No	A string in a column cannot exceed 8 MB in length.
Worker timeout period	[1,3600]	Time	odps.function.timeout	600	Yes	The timeout period of a map or reduce worker when the worker does not read or write data, or stops sending heartbeats by using `context.progress()` . The default value is 600 seconds.
Field types supported by tables that are referenced by MapReduce	BIGINT, DOUBLE, STRING, DATETIME, and BOOLEAN	Data type	-	N/A	No	When a MapReduce task references a table, an error is returned if the table has field types that are not supported.
Object Storage Service (OSS) data read	-	Feature	-	N/A	No	MapReduce cannot read OSS data.

Item	Value range	Classification	Configuration item	Default value	Configurable	Description
New data types in MaxCompute V2.0	-	Feature	-	N/A	No	MapReduce does not support the new data types in MaxCompute V2.0.

For more information about MapReduce, see Overview.

Limits on PyODPS

Before you use DataWorks to develop PyODPS jobs in MaxCompute, take note of the following limits:

- Each PyODPS node can process a maximum of 50 MB of data and can occupy a maximum of 1 GB of memory. Otherwise, DataWorks terminates the PyODPS node. Do not write unnecessary Python data processing code in PyODPS tasks.
- The efficiency of writing and debugging code in DataWorks is low. We recommend that you install an integrated development environment (IDE) on your machine to write code.
- To avoid excess pressure on the gateway of DataWorks, DataWorks limits the CPU utilization and memory usage. If the system displays **Got killed**, the memory usage exceeds the limit and the system terminates the related processes. Therefore, we recommend that you do not perform local data operations. However, the limits on the memory usage and CPU utilization do not apply to SQL or DataFrame nodes, except to_pandas, that are initiated by PyODPS.
- Functions may be limited in the following aspects due to the lack of packages such as matplotlib:
  - The use of the plot function of DataFrame is affected.
  - DataFrame user-defined functions (UDFs) can be used only after they are submitted to MaxCompute. As required by the Python sandbox, you can use only pure Python libraries and the NumPy library to run UDFs. Other third-party libraries such as pandas cannot be used.
  - However, you can use the NumPy and pandas libraries that are pre-installed in DataWorks to run non-UDFs. Third-party packages that contain binary code are not supported.
- For compatibility reasons, options.tunnel.use_instance_tunnel is set to False in DataWorks by default. If you want to enable InstanceTunnel globally, you must set this parameter to True.
- For implementation reasons, the Python atexit package is not supported. You must use try-finally to implement relevant features.

For more information about PyODPS, see PyODPS.

Limits on Graph

Before you develop Graph jobs in MaxCompute, take note of the following limits:

- Each job can reference up to 256 resources. Each table or archive is considered as one unit.
- The total bytes of resources referenced by a job cannot exceed 512 MB.
- The number of the inputs of a job cannot exceed 1,024, and that of the outputs of a job cannot exceed 256. The number of input tables cannot exceed 64.
- Labels that are specified for multiple outputs cannot be null or empty strings. A label cannot exceed 256 strings in length and can contain only letters, digits, underscores (_), number signs (#), periods (.), and hyphens (-).

- The number of custom counters in a job cannot exceed 64. The counter `group name` and `counter name` cannot contain number signs (#). The total length of the two names cannot exceed 100 characters.
- The number of workers for a job is calculated by the framework. The maximum number of workers is 1,000. An exception is thrown if the number of workers exceeds this value.
- A worker consumes 200 units of CPU resources by default. The range of resources consumed is 50 to 800.
- A worker consumes 4,096 MB memory by default. The range of memory consumed is 256 MB to 12 GB.
- A worker can repeatedly read a resource up to 64 times.
- The default value of `split_size` is 64 MB. You can set the value as needed. The value of `split_size` must be greater than 0 and smaller than or equal to the result of the 9223372036854775807>>20 operation.
- GraphLoader, Vertex, and Aggregator in MaxCompute Graph are restricted by the Java sandbox when they are run in a cluster. However, the main program of Graph jobs is not restricted by the Java sandbox. For more information, see Java Sandbox.

For more information about Graph, see Graph.

Other limits

The following table describes the maximum parallelism of jobs that you can submit in a MaxCompute project in different regions.

Region	Maximum job parallelism for a MaxCompute project
China (Hangzhou), China (Shanghai), China (Beijing), China (Zhangjiakou), China (Shenzhen), and China (Chengdu)	2500
China (Hong Kong), Singapore (Singapore), Australia (Sydney), Malaysia (Kuala Lumpur), Indonesia (Jakarta), Japan (Tokyo), Germany (Frankfurt), US (Silicon Valley), US (Virginia), UK (London), India (Mumbai), and UAE (Dubai)	300

If you continue to submit jobs when the parallelism of jobs that you submit in a MaxCompute project reaches the maximum, an error is returned. The following error message shows an example: `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.Customer stories

MaxCompute is widely used in various fields to process big data in the cloud. MaxCompute solves issues that are related to the analysis of large amounts of data and reduces O&M costs of enterprises. This way, enterprises can concentrate on business development. This topic describes featured customer stories of MaxCompute.

Umeng

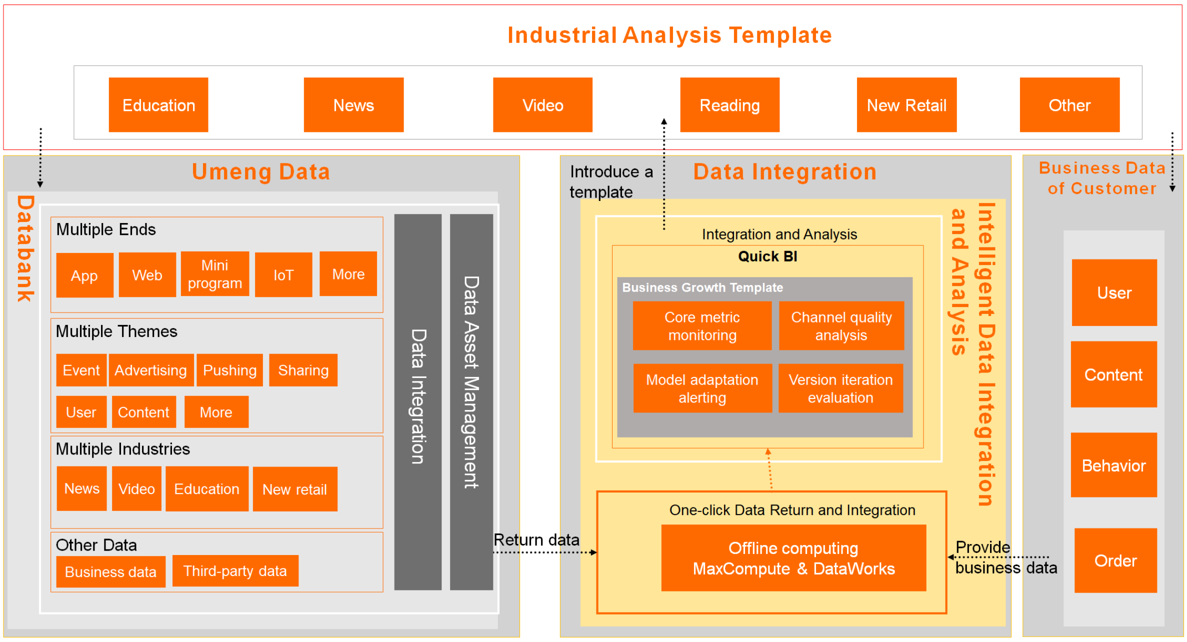
- Customer profile

  Umeng is an independent third-party global data intelligence service provider that uses technology and algorithm capabilities to mine labels and analyze metrics based on global data resources. Umeng helps enterprises gain deep user insights to drive real-time business decision-making and continuous business growth.

- Customer demands
  - Help enterprises and developers integrate and analyze data of independent data systems.
  - Help enterprises and developers achieve a balance between the flexibility of business intelligence (BI) systems and business availability.

- Solution

  Umeng and MaxCompute jointly build a developer databank to provide the self-service analysis service Data Open Platform (U-DOP). U-DOP fully integrates the data of Umeng with the data of enterprises. U-DOP returns data to MaxCompute by subscribing to data packages, presets analysis templates, and analyzes data by using BI tools. This way, U-DOP provides enterprises with flexible end-to-end data analysis capabilities. The following figure shows the architecture of the solution.



E-commerce: Wanwudezhi

- Customer profile

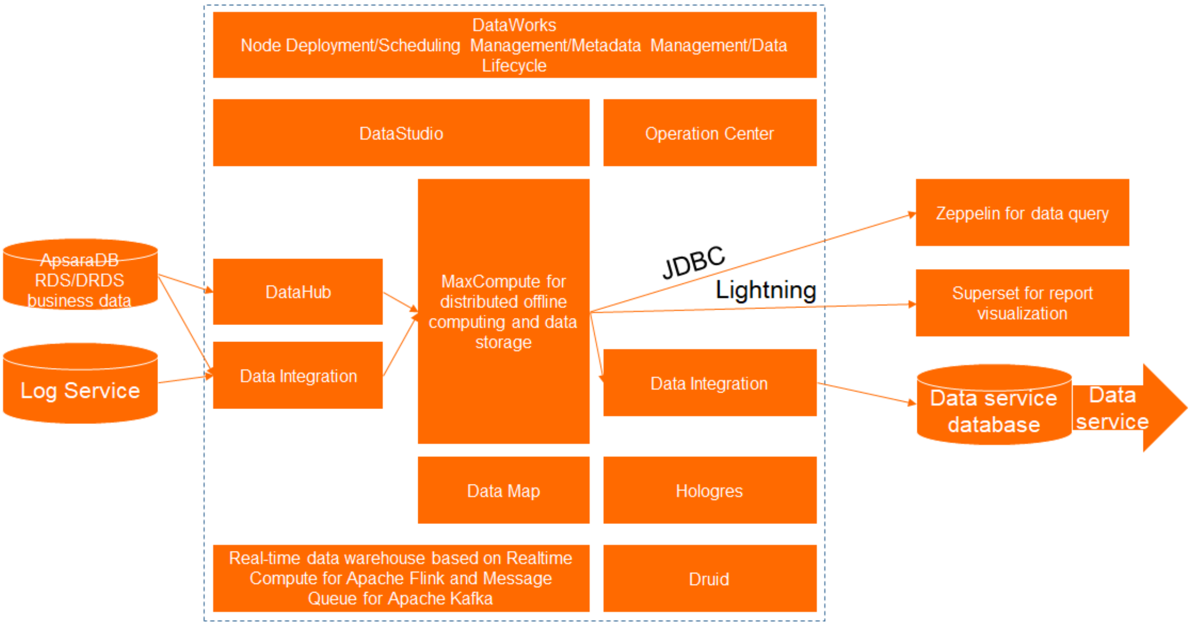
Wanwudezhi is an online platform for trading Chinese handicrafts and antiques. This platform provides an online antique appraisal service and an online space for the traditional Chinese culture community. The number of monthly active users on this platform reaches 6,000,000. Wanwudezhi provides three core services: live broadcast, auction, and Buy It Now. Wanwudezhi follows the "appraisal first, delivery later" principle and builds a transparent, healthy, and secure cultural consumption environment for users. This way, consumers can buy traditional Chinese items with high quality and feel the charm of traditional Chinese culture.

- Customer demands

  Wanwudezhi wants to use the SaaS and PaaS services provided by the cloud platform to build a research and development system that can help improve development efficiency and reduce labor costs. Wanwudezhi also wants to efficiently migrate the original MySQL system to the cloud.

- Solution

  A big data platform is built for Wanwudezhi based on the Alibaba Cloud DataWorks and MaxCompute frameworks. This platform provides core storage and computing components, and upper-level visualization and business query capabilities for Wanwudezhi. Custom software development is performed based on the original open source solution. The following figure shows the architecture of the solution.


DataWorks
Node Deployment/Scheduling Management/Metadata Management/Data Lifecycle
DataStudio
Operation Center
ApsaraDB RDS/DRDS business data
Log Service
DataHub
Data Integration
MaxCompute for distributed offline computing and data storage
JDBC
Lightning
Zeppelin for data query
Superset for report visualization
Data Integration
Data service database
Data service
Data Map
Hologres
Real-time data warehouse based on Realtime Compute for Apache Flink and Message Queue for Apache Kafka
Druid

Internet-based social media: Xiaodaka

- Customer profile

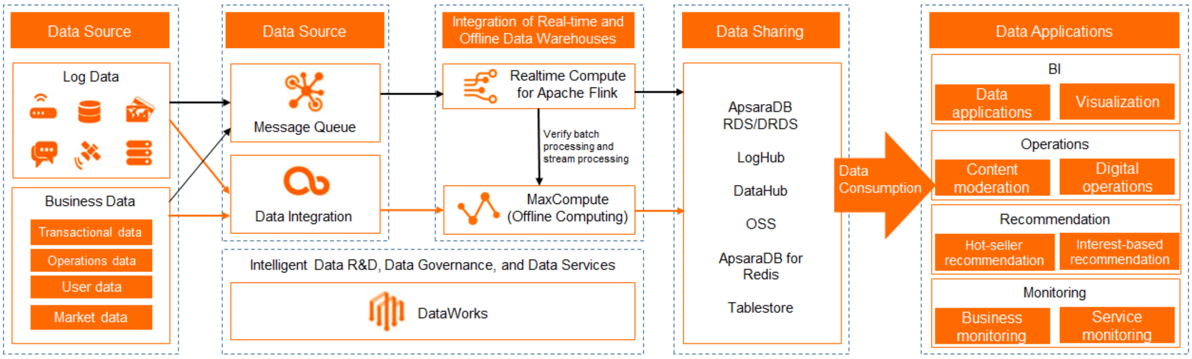
  Xiaodaka is an interest community program that offers interest circles such as painting, yoga, fitness, photography, parenting, reading, and designer toys. Xiaodaka allows users to share things, communicate, and grow together in the circles that they are interested in. Xiaodaka has millions of daily active users and can generate terabytes of data every day.

- Customer demands

  Xiaodaka wants to build a data warehouse that features agile development and high extensibility at low development costs and O&M costs.

- Solution

A MaxCompute data warehouse is built at low labor costs and software and hardware costs. The data warehouse ensures high resource utilization. The deployment process of the data warehouse is easy and can be complete by only one developer. The following figure shows the architecture of the solution.



Internet finance: Tianhong Asset Management

- Customer profile

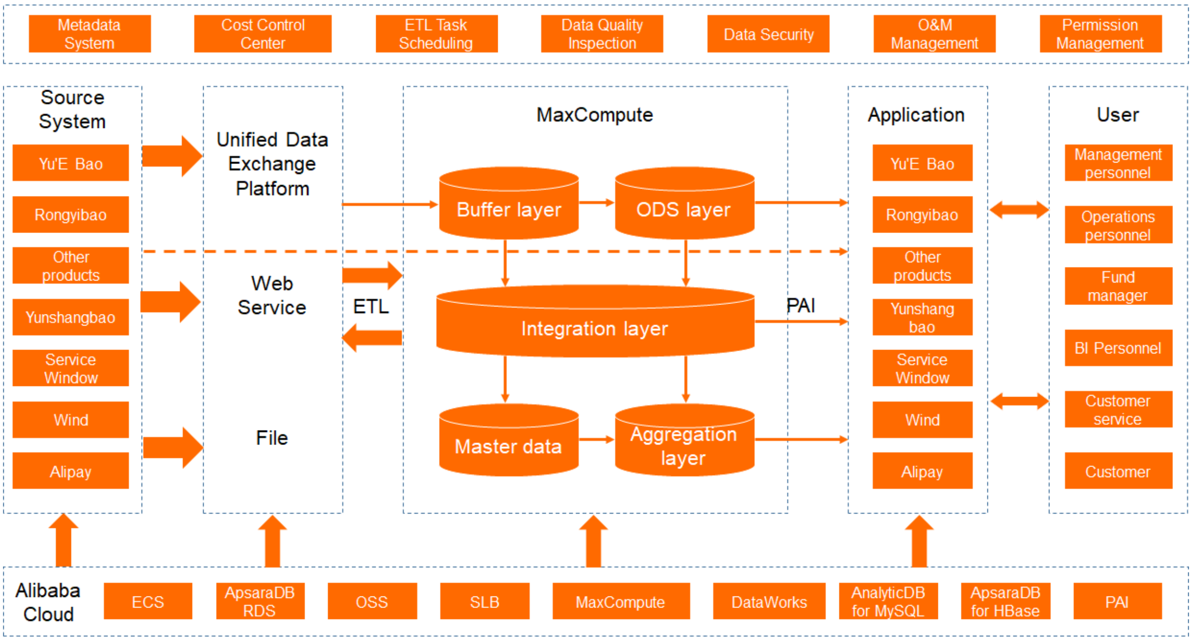
  Tianhong Asset Management is the largest public fund manager in China and aims to become the largest index fund service provider. Tianhong Asset Management innovates its services based on the needs of customers and uses index funds as the underlying tool to provide customers with end-to-end services. Tianhong Asset Management and Alipay jointly launched Yu'E Bao.

- Customer demands

  As the number of Yu'E Bao users and the amount of data exponentially increase, Tianhong Asset Management can no longer manage data only by using Hadoop clusters. In addition, to make business decisions and predict user behavior, business personnel require accurate user analysis results.

- Solution

  Tianhong Asset Management builds an enterprise-level end-to-end big data solution based on MaxCompute. MaxCompute has powerful, secure, and stable storage, O&M, and computing capabilities that can be used to store and process large amounts of data. MaxCompute significantly improves the data processing efficiency of Tianhong Asset Management. For example, MaxCompute reduces the data processing time from 8 hours to 1.5 hours. MaxCompute also reduces the workloads of on-premises servers, development costs, and labor costs, which allows developers to focus only on business development. This solution enables Tianhong Asset Management to provide better financial services for users. The following figure shows the architecture of the solution.



Smart delivery: Qianxun Spatial Intelligence

- **Customer profile**

  Qianxun Spatial Intelligence (Qianxun SI) is a leading high-precision positioning service provider in the world. Qianxun SI offers centimeter-level dynamic positioning services and millimeter-level static positioning services. Qianxun SI relies on BeiDou Navigation Satellite System and is compatible with the Global Positioning System (GPS), Global Navigation Satellite System (GLONASS), and Galileo. Qianxun SI has deployed more than 2,400 ground-based augmentation stations across China. Qianxun SI provides users across China with high-precision positioning services and extended services by using self-developed positioning algorithms and Internet-based big data computing technologies.

- **Customer demands**

  Qianxun SI wants to improve computing precision and speed to meet various requirements for real-time high-precision positioning.

- **Solution**

  A solution based on MaxCompute is developed for Qianxun SI to eliminate the need to build self-managed clusters. In the hybrid cloud architecture, confidential data is processed on the Alibaba Cloud hybrid cloud, large-scale data in the cloud is computed in MaxCompute, and positioning data is broadcast on the Alibaba Cloud public cloud.

6.Computing models enabled in each region

This topic lists the MaxCompute computing models that are enabled in each region.

Region	SQL	MapReduce	Hologres	Spark	PyODPS	Mars
China (Beijing)	✓	✓	✓	✓	✓	✓
China (Hangzhou)	✓	✓	✓	✓	✓	✓
China (Shanghai)	✓	✓	✓	✓	✓	✓
China (Shenzhen)	✓	✓	✓	✓	✓	✓
China (Chengdu)	✓	✓	✗	✓	✓	✓
China (Zhangjiakou)	✓	✓	✓	✓	✓	✓
China (Hong Kong)	✓	✓	✓	✓	✓	✓
Singapore (Singapore)	✓	✓	✓	✓	✓	✗
Malaysia (Kuala Lumpur)	✓	✓	✓	✓	✓	✗
Indonesia (Jakarta)	✓	✓	✓	✓	✓	✗
Australia (Sydney)	✓	✓	✗	✓	✓	✗
Japan (Tokyo)	✓	✓	✓	✓	✓	✗
US (Silicon Valley)	✓	✓	✓	✓	✓	✗
US (Virginia)	✓	✓	✗	✓	✓	✗
UK (London)	✓	✓	✗	✓	✓	✗

Region	SQL	MapReduce	Hologres	Spark	PyODPS	Mars
Germany (Frankfurt)	✓	✓	✓	✓	✓	✕
India (Mumbai)	✓	✓	✓	✓	✓	✕
UAE (Dubai)	✓	✓	✕	✓	✓	✕

Note ✓ indicates that a computing model is enabled. ✕ indicates that a computing model is disabled.

7.MaxCompute development history

This topic provides an overview of the key milestones and achievements of Alibaba Cloud MaxCompute.

Key milestones

- In September 2009, the ODPS big data platform, which is the predecessor of MaxCompute, was officially launched.
- In October 2010, ODPS ran stably as the first-generation cloud computing platform developed by Alibaba Group.
- As of August 2013, a maximum of 5,000 servers can be deployed in a single cluster.
- In July 2014, ODPS was released for commercial use to completely replace open source systems.
- Between the years 2015 and 2018, the platform grew to maturity with optimized performance and global presence. More than 10,000 servers can be deployed in a single cluster. ODPS was renamed MaxCompute.

Awards

- In November 2018, Alibaba Cloud, represented by MaxCompute, DataWorks, and AnalyticDB, was positioned as a contender in Cloud Data Warehouse (CDW) in Forrester Wave™ Cloud Data Warehouse, Q4 2018.
- In the public cloud-based BigBench test in September 2018, the performance of MaxCompute in processing 100 TB of data reached 18,176.71 queries per minute (QPM), which doubled that from October 2017. In addition, the performance of MaxCompute in processing ultra-small 10 TB of data was three times that of other competitors.
- In April 2018, several customer cases related to MaxCompute won the "2017 Big Data Outstanding Product and Application Solution Cases Award."
- In March 2018, MaxCompute made it onto the big data service list of Forrester Wave™ *Cloud Data Warehouse, Q1 2018*.
- In March 2018, Gartner released its "*2017 Magic Quadrant for Data Management Solutions for Analytics (DMSA) Magic Quadrant (MQ)*" report. Alibaba Cloud entered the Magic Quadrant as a cloud service provider.
- In October 2017, the first public cloud-based BigBench test was conducted on MaxCompute. MaxCompute processed more than 100 TB of data with a performance of 7,830 QPM, making it the first engine in the world to exceed the data processing performance of 7,000 QPM.
- In June 2017, MaxCompute won the Gold Medal at the China International Software Expo.
- In the CloudSort competition in November 2016, MaxCompute won the world championships of Indy (special-purpose sort) and Daytona (general-purpose sort) with a record of USD 0.82/TB. This breaks the USD 4.51/TB record set by Amazon Web Services (AWS) in 2014.
- In the GraySort benchmark competition in October 2015, MaxCompute completed the sorting of 100 TB data in 377 seconds, which broke the 1,406-second record set by Apache Spark.

Certifications

- MaxCompute was certified by the Ministry of Industry and Information Technology (MIIT) as the first big data platform in mainland China that can support up to 10,000 servers in one cluster.
- MaxCompute attained certifications from China Academy of Information and Communications Technology (CAICT) and China Electronics Standardization Institute (CESI).

- MaxCompute has passed an independent third-party audit on compliance with the trust services criteria for security, availability, and confidentiality established by American Institute of Certified Public Accountants (AICPA). For more information about the audit report, see SOC 3 Report.

Global recognition for advancements in the formulation of big data standards

- Alibaba, represented by MaxCompute, is the only company in mainland China to become a special member of the Transaction Processing Performance Council (TPC) for TPCx-BB.
- Optimized Row Columnar (ORC) is one of the top two open source systems for computing and storage standardization. As a Production Material Control (PMC) member of the ORC community, MaxCompute has contributed the most code to the community in the past two years. MaxCompute has become the forerunner for storage standardization.
- MaxCompute is an active player in Apache Calcite, the most famous optimizer project in the world. It has become a special member of this project, making Alibaba one of the two leading companies in the optimizer field in mainland China.

8.Usage notes

This topic provides reading recommendations based on your roles.

MaxCompute beginners

If you are a beginner in MaxCompute, we recommend that you first familiarize yourself with the modules described in the following table.

Module	Description
Product Introduction	Provides an overview of MaxCompute and describes the features, scenarios, limits, and basic concepts of MaxCompute. This module helps you obtain a general knowledge of MaxCompute.
Preparations	Describes how to create an account, prepare an environment, create a table, import data, run SQL jobs, and export returned data.
Quick Start	
Common SQL statements	Describes the commonly used commands in MaxCompute. This module helps you familiarize yourself with operations on MaxCompute.
Tools	Describes the common tools in MaxCompute, such as the MaxCompute client, query editor, and MaxCompute Studio. Before you analyze data, you must familiarize yourself with the tools.
Endpoints	Describes the network connection modes supported in different regions and the endpoints that correspond to each region. This module also describes the issues that may occur when MaxCompute is connected to other Alibaba Cloud services, such as Elastic Compute Service (ECS), Tablestore, and Object Storage Service (OSS). These issues include network connectivity issues and issues related to data download charges.

Data analysts

If you are a data analyst, we recommend that you familiarize yourself with the SQL topics. You can query and analyze large volumes of data stored in MaxCompute. The following table describes the features that are provided by MaxCompute SQL.

Feature	Description
DDL statements	Allows you to manage tables, partitions, columns, lifecycles, and views.
DML statements	Allows you to insert data into or update data in tables or partitions.
DQL statements	Allows you to perform various query operations, such as SELECT and subqueries.
SQL enhancement operations	Allows you to perform SQL enhancement operations, such as importing and exporting data from MaxCompute tables and cloning table data, by using commands.

Feature	Description
Built-in functions	Allows you to process data by using MaxCompute built-in functions, such as the mathematical functions, window functions, date functions, aggregate functions, and string functions.
UDF	Allows you to create user-defined functions (UDFs) to meet your computing requirements.

Users with development experience

If you have development experience, understand the distributed architecture, and want to obtain data analytics capabilities that SQL cannot deliver, we recommend that you familiarize yourself with advanced functional modules of MaxCompute.

Module	Description
MapReduce	MaxCompute provides the MapReduce programming model in Java. You can use the Java API provided by MapReduce to write MapReduce programs and process data in MaxCompute.
Graph	Graph is a processing framework for iterative graph computing. A graph consists of vertices and edges, both of which contain values. MaxCompute Graph iteratively edits and evolves graphs to obtain analysis results.
Tunnel	MaxCompute Tunnel enables you to upload or download large amounts of data to or from MaxCompute at a time.
Java SDK	MaxCompute provides an SDK for Java for developers.
Python SDK	MaxCompute provides an SDK for Python for developers.

Project owners or administrators

If you are a project owner or administrator, we recommend that you familiarize yourself with the modules described in the following table. A project owner can create and use projects, and a project administrator can manage projects, security operations, and costs.

Module	Feature	Description

Module	Feature	Description
	Prepare for project creation	A project is a basic organizational unit of MaxCompute. Similar to a database or schema in a traditional database system, a project is used to isolate users and control access requests. A user can have permissions on multiple projects. After a user is granted the related permissions, the user can access objects, such as tables, resources, functions, and instances, across projects. MaxCompute is used to manage various objects in projects. You must make the following preparations before you create a project: ● Prepare your budget for resources You are charged for storage resources, computing resources, and resources for Internet-based data downloads. ○ Storage resources: You are charged for these resources based on the pay-as-you-go billing method and tiered unit prices. You can estimate their costs based on the volume of data stored. Data stored in MaxCompute changes all the time. As a result, the costs also change. ○ Computing resources: You are charged for these resources based on the pay-as-you-go and subscription billing methods. It is difficult to estimate the number of required computing resources at the beginning of your project. We recommend that you use the pay-as-you-go billing method and then decide whether to switch to the subscription billing method based on the number of computing resources used. ○ Resources for Internet-based data downloads: You are charged for these resources based on the pay-as-you-go billing method. For more information, see Storage pricing (pay-as-you-go), Computing pricing, and Download pricing (pay-as-you-go). ● Create an account and activate the service Before you create a MaxCompute project, you must create an Alibaba Cloud account and activate MaxCompute. Bills are issued to the Alibaba Cloud account. After the account is created, you must choose the pay-as-you-go or subscription billing method based on your budget for the resources you require.
	Create a project	For more information, see Create a MaxCompute project.
	Manage project members	Members are managed based on member responsibilities and security requirements. If you use MaxCompute in the DataWorks console, you must understand the permission relationships between the two services.

Module	Feature	Description
Project management	Manage RAM users	You can manage MaxCompute projects by using your Alibaba Cloud account or the credentials of a RAM user. You can add RAM users of your Alibaba Cloud account to a MaxCompute project. For more information about RAM users, see Prepare a RAM user. If you manage MaxCompute projects and DataWorks workspaces in the DataWorks console, you can add only RAM users of your Alibaba Cloud account as members. Therefore, you must use your Alibaba Cloud account to create RAM users and manage these RAM users in the Resource Access Management (RAM) console. Note • We recommend that you do not allow multiple project members share one RAM user. • When a project member is transferred to a new position or resigns, you must delete the RAM user of the project member at the earliest opportunity. If a RAM user is added as a project member in the DataWorks console, delete the project member in the DataWorks console and then delete the RAM user in the RAM console.
	Manage scheduling resources	You are required to manage the scheduling resources of DataWorks. These resources are used to execute or distribute the tasks that are delivered by the scheduling system. Scheduling resources of DataWorks are categorized into the following types: • Default scheduling resources. Default scheduling resources are the resources in the public resource pool of DataWorks. If the parallelism of DataWorks nodes is high and the scheduling resources are insufficient, the nodes wait for resources. After resources are allocated to the nodes, the nodes run the delivered tasks. • Custom scheduling resources. You can configure your ECS instance as a scheduling server to run or distribute delivered tasks. You can use your Alibaba Cloud account to create custom scheduling resources. Scheduling resources include physical machines or ECS instances that are used to run tasks, such as data synchronization tasks. You can submit a ticket to create a scheduling resource group. Custom scheduling resource groups that exist are not affected. This eliminates the limits of the default scheduling resource group.
	Configure projects	Only the owner of a project has the permissions to configure the project. For example, the project owner can specify whether to enable full table scan and whether to enable the MaxCompute V2.0 data type edition. For more information, see Project operations.

Module	Feature	Description
Cost management	None	Budgets for resources help you estimate costs before you use the resources. It is difficult to estimate the precise costs due to the different billing methods of MaxCompute. You must manage costs during the entire business development process. ● For more information about pricing, see Billing method. ● You can switch between the pay-as-you-go and subscription billing methods. For more information, see Switch billing methods.