ALIBABA CLOUD

Alibaba Cloud

DataWorks

Product Introduction

Document Version: 20210120

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 DataWorks? ---- 05
2.Basic concepts ---- 07
3.Scenarios ---- 11
4.Data development process ---- 14
5.Basic mode and standard mode ---- 15

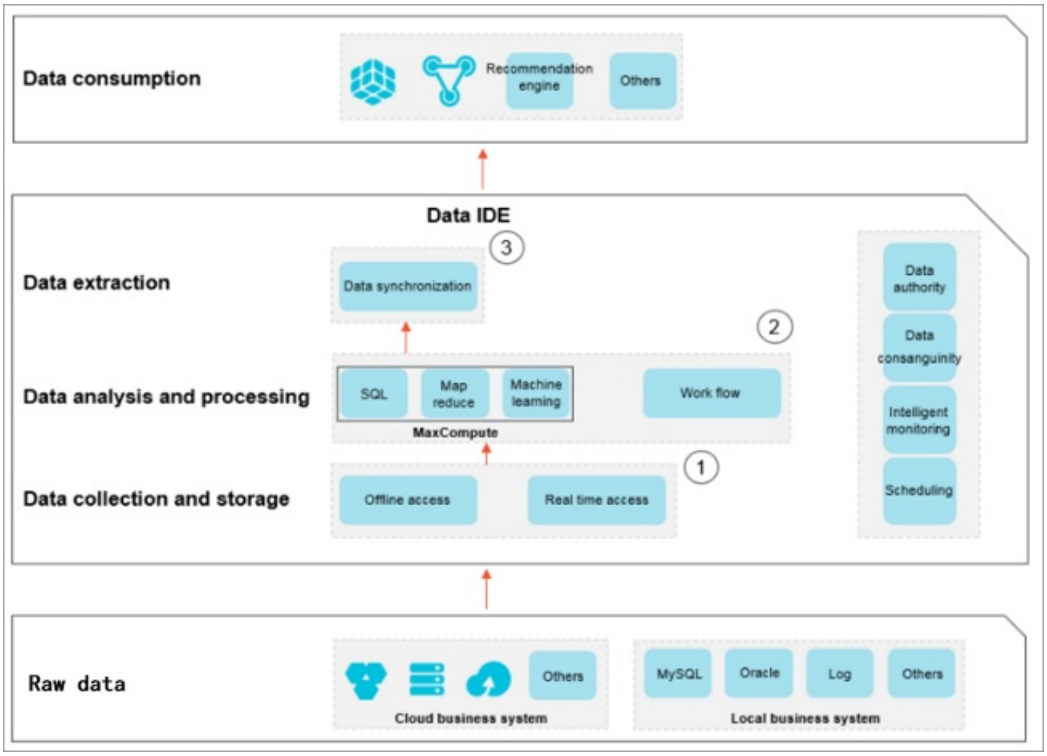
1.What is DataWorks?

This topic introduces Alibaba Cloud DataWorks, including its features and limits.

DataWorks is an important platform as a service (PaaS) of Alibaba Cloud. It offers all-around services, including Data Integration, DataStudio, Data Map, Data Quality, and DataService Studio. In addition, it provides a one-stop data development and management console to help enterprises mine and explore data value.

DataWorks supports multiple compute and storage engines, including MaxCompute, E-MapReduce, Realtime Compute for Apache Flink, Machine Learning Platform for AI, Graph Compute, and Hologres. It also allows you to use custom computing and storage services. As an all-in-one platform, DataWorks provides end-to-end big data services, artificial intelligence (AI) development, and data governance.

DataWorks simplifies data transmission, conversion, and integration. You can import data from different data stores, convert, analyze, and process the data, and then transmit the data to other data systems.



Limits

DataWorks supports only Google Chrome 54 or later.

Features

- DataWorks is hosted on the cloud.
  - DataWorks provides powerful scheduling capabilities. For more information, see Schedule.
    - In DataWorks, nodes can be triggered by time- or dependency-based scheduling configuration. For more information, see Time properties and Dependencies.
    - DataWorks enables tens of millions of nodes to run accurately and on time every day based on node relationships in directed acyclic graphs (DAGs).
    - DataWorks allows you to run nodes at custom intervals in minutes, hours, days, weeks, or months.

- DataWorks is a cloud-hosted environment that frees you from server deployment.
- DataWorks provides the isolation feature to ensure that nodes of different tenants do not affect each other.

- DataWorks supports multiple node types, including batch sync node, Shell node, ODPS SQL node, and ODPS MR node. It analyzes and processes complex data based on the dependencies between nodes.
  - Data conversion: By using the powerful computing capabilities of MaxCompute, DataWorks ensures the superior performance on analyzing and processing big data.
  - Data integration: Based on the Data Integration service, DataWorks supports more than 20 types of data stores and provides stable and efficient data transmission features. For more information, see Data Integration.
- DataWorks provides visualized code development.

  DataWorks provides a graphical user interface (GUI) for you to develop code and design workflows. You can perform simple drag-and-drop operations to create complex data analytics nodes without the need to use development tools. For more information, see GUI elements.

  A browser with Internet access enables you to develop code anytime, anywhere.
- DataWorks supports monitoring and alerting.

  Operation Center provides a visualized node monitoring and management tool and displays the overall node running status in DAGs. For more information, see Operation Center.

  You can configure various alert notification methods to promptly notify relevant staff when a node error occurs. This ensures normal business operation. For more information, see Monitor.

2.Basic concepts

This topic introduces the basic concepts in DataWorks, including workspace, workflow, solution, SQL script template, node, instance, commit operation, script, resource, function, and output name.

Workspace

Workspaces are basic units for managing nodes, members, roles, and permissions in DataWorks. A workspace administrator can add members to the workspace and assign the workspace administrator, developer, administration expert, deployment expert, security expert, or visitor role to each member. This way, workspace members with different roles can collaborate with each other.

> **Note** We recommend that you create workspaces to isolate resources by department or business unit.

You can bind instances of multiple compute engines such as MaxCompute, E-MapReduce, and Realtime Compute to a single workspace. After you bind a compute engine instance to a workspace, you can configure and schedule nodes in the workspace.

Workflow

Workflows are abstracted from business to help you manage and develop code based on business demands and improve the efficiency of node management.

> **Note** A workflow can be used in multiple solutions.

Workflows help you manage code based on business demands.

- A workflow allows you to organize nodes by type.
- A workflow supports a hierarchical directory structure. We recommend that you create a maximum of four levels of subdirectories for a workflow.
- You can view and optimize a workflow from the business perspective.
- You can deploy and manage nodes in a workflow as a whole.
- A workflow provides a dashboard for you to develop code with improved efficiency.

Solution

You can include one or more workflows in a solution.

Solutions have the following benefits:

- A solution can contain multiple workflows.
- A workflow can be used in multiple solutions.
- A solution can include various nodes. This improves user experience.

SQL script template

SQL script templates are general logic chunks that are abstracted from SQL scripts. They can improve the reusability of code.

When SQL code is processed, operations such as filter, join, and aggregate are performed on one or more source tables to generate a result table based on business requirements. An SQL script template is a template for processing SQL code and includes multiple input and output parameters.

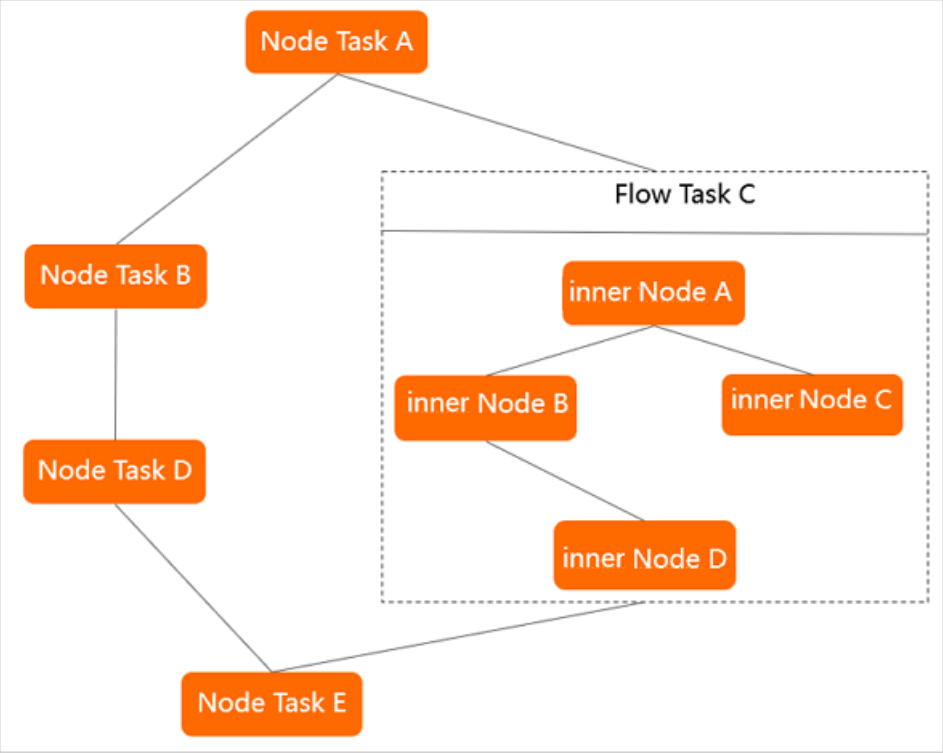
Node

Each type of node is used to perform a specific data operation. For example:

- A sync node is used to synchronize data from ApsaraDB RDS to MaxCompute.
- An ODPS SQL node is used to convert data by executing SQL statements that are supported by MaxCompute.

Each node has zero or more input tables or datasets and generates one or more output tables or datasets.

Nodes are classified into node tasks, flow tasks, and inner nodes.



Type	Description
Node task	A node task is used to perform a data operation. You can configure dependencies between a node task and other node tasks or flow tasks to form a directed acyclic graph (DAG).
Flow task	A flow task contains a group of inner nodes that process a workflow. We recommend that you create less than 10 flow tasks. Inner nodes in a flow task cannot be depended upon by other flow tasks or node tasks. You can configure dependencies between a flow task and other flow tasks or node tasks to form a DAG. **Note** In DataWorks V2.0 and later, you can find the flow tasks that are created in DataWorks V1.0 but cannot create flow tasks. Instead, you can create workflows to perform similar operations.

Type	Description
Inner node	An inner node is a node within a flow task. The features of an inner node are basically the same as those of a node task. You can configure dependencies by performing drag-and-drop operations. However, you cannot configure a recurrence for inner nodes because they follow the recurrence configuration of the flow task.

Instance

An instance is a snapshot of a node at a specific point in time. An instance is generated every time a node is run as scheduled by the scheduling system or is manually triggered. An instance contains information such as the time at which the node is run, the running status of the node, and operational logs.

Assume that Node 1 is scheduled to run at 02:00 every day. The scheduling system automatically generates a snapshot at 23:30 every day based on the time that is defined for the auto triggered node. The snapshot is an instance of Node 1 that is to run at 02:00 the next day. At 02:00 the next day, if the scheduling system verifies that all the ancestor instances are run, the system automatically runs the instance of Node 1.

> **Note** You can query the instance information on the **Cycle Instance** page of **Operation Center**.

Commit

A commit operation refers to the process of committing a node or workflow from the development environment to the scheduling system in DataWorks. When the node or workflow is committed, all of the code and scheduling configurations are also committed to the scheduling system. The scheduling system runs the node or workflow as configured.

> **Note** The scheduling system runs nodes and workflows only after they are committed.

Script

A script stores code for data analysis. The code in a script can be used only for data query and analysis. It cannot be committed to the scheduling system for scheduling.

Resource and function

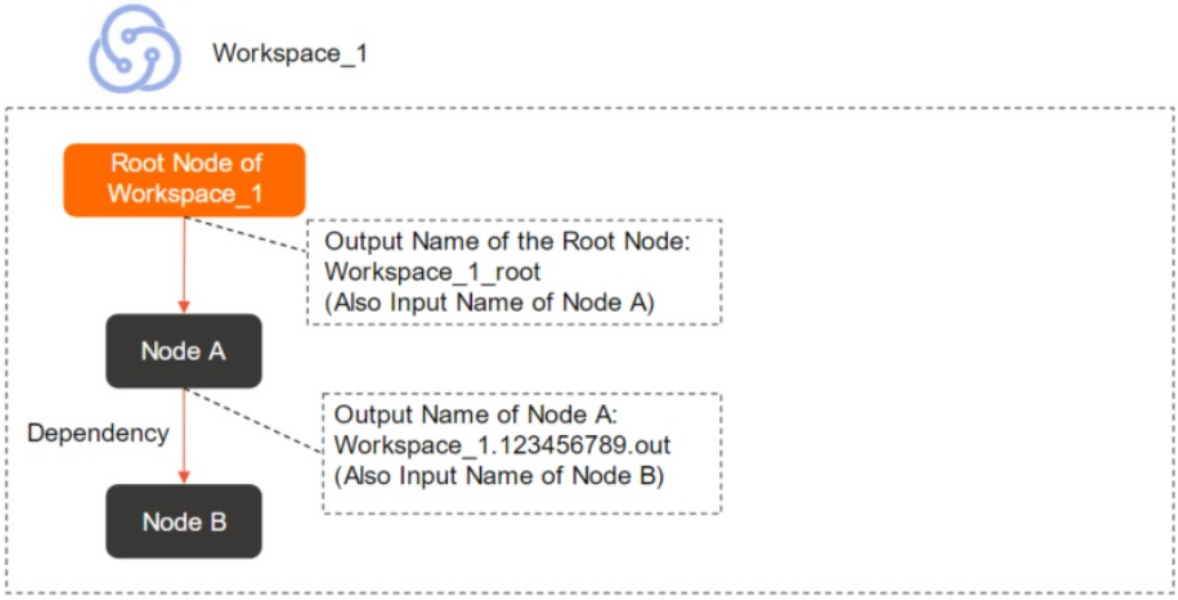
Resources and functions are concepts in MaxCompute. For more information, see Resource and Function.

The DataWorks console allows you to manage resources and functions. If resources or functions are uploaded by using other services such as MaxCompute, you cannot query them in DataWorks.

Output name

Under an Alibaba Cloud account, each node has an output name that is used to connect to its descendant nodes.

When you configure dependencies for a node, you must use its output name instead of its node name or node ID. After you configure the dependencies, the output name of the node serves as the input name of its descendant nodes.



> **Note** Each output name distinguishes a node from other nodes under the same Alibaba Cloud account. By default, the output name of each node is in the following format: Workspace name.Randomly generated nine-digit number_out. You can customize the output name for a node, but make sure that the output name of the node is unique under your Alibaba Cloud account.

Metadata

Metadata is data that provides descriptions for other data. It can describe the attributes such as the name, size, and data type, or the structure including the field, type, and length, or relevant information such as the location, owner, output node, and access permission. In DataWorks, metadata refers to information about tables or databases. Data Map is the main application for metadata management.

3.Scenarios

This topic describes typical scenarios of DataWorks.

Log and big data analysis

- Improved work efficiency

  DataWorks allows you to synchronize log data to MaxCompute and use SQL statements to analyze and process the log data. This improves your work efficiency.

- Enhanced storage efficiency

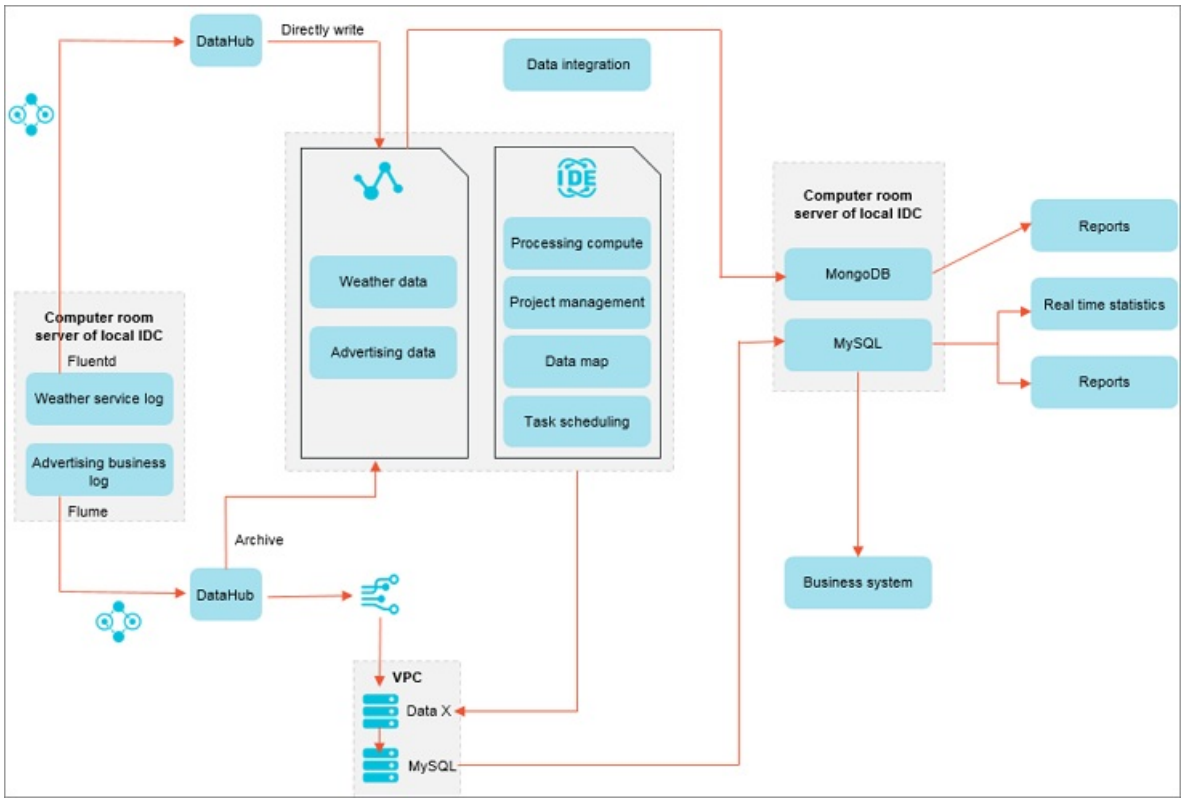
  DataWorks saves the overall costs and improves the performance and stability of storage and computing services.

- Simplified use of big data

  DataWorks supports multiple open source MaxCompute plug-ins so that you can easily migrate data to the cloud.

Related services:

DataWorks + Data Integration + AnalyticDB for MySQL + Quick BI + MaxCompute



Refined business operations

- Improved business insights

  With the help of MaxCompute, DataWorks allows you to refine business operations on millions of users.

- Data-based business

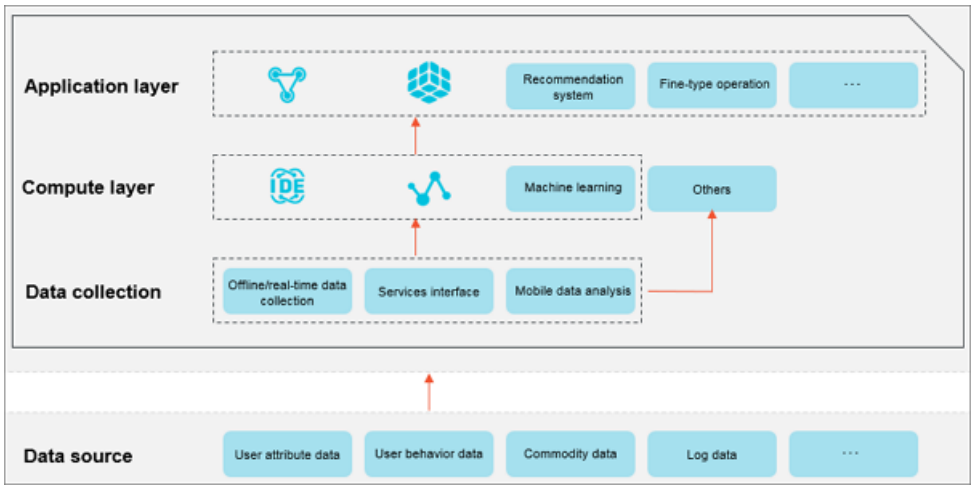
DataWorks helps you effectively analyze and monitor business data to improve your business efficiency.

- Quick response to business demands

  DataWorks supports business data analysis so that you can quickly process new business demands.

Related services:

DataWorks + Data Integration + Quick BI + MaxCompute



Data security management

- Sensitive data identification

  DataWorks can automatically identify sensitive data and use tags to classify the data based on custom rules.

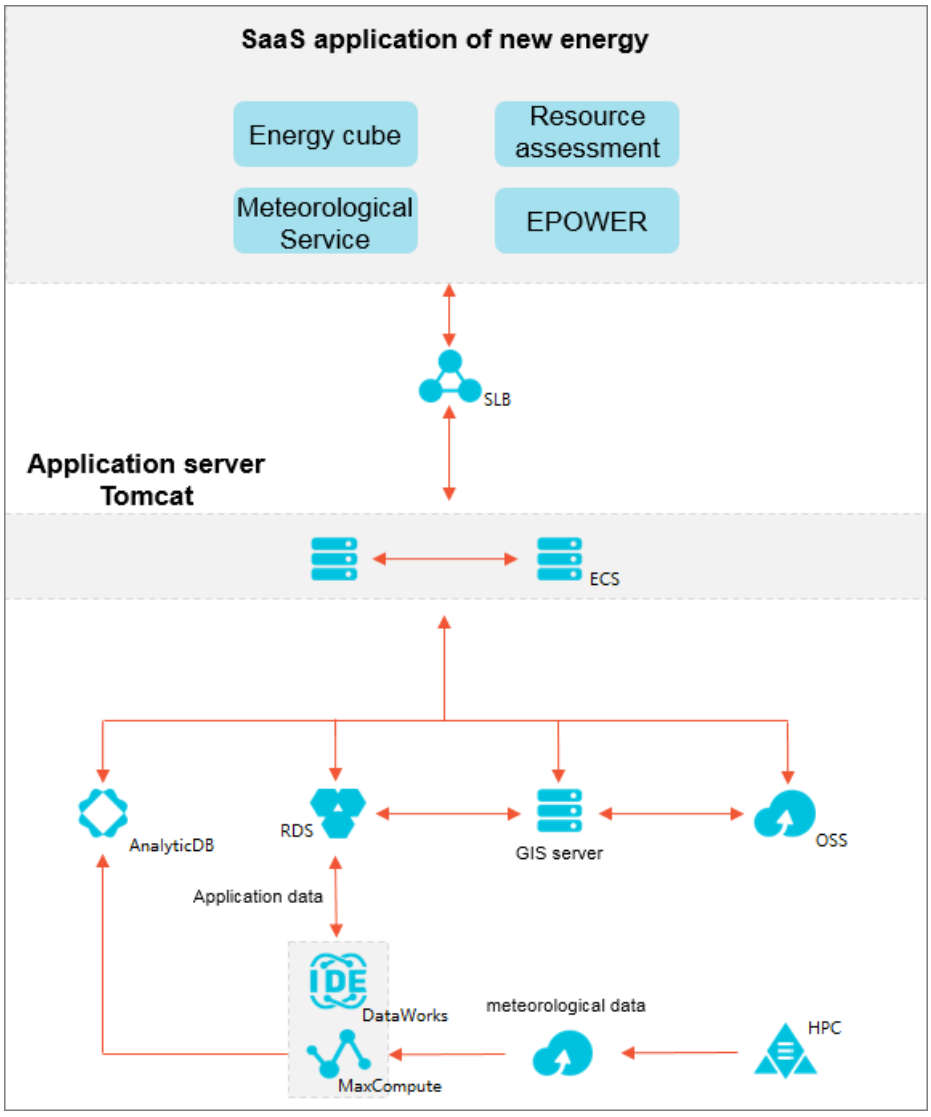
- Sensitive data de-identification and presentation

  DataWorks allows you to set data de-identification rules to de-identify the sensitive information during data presentation.

- Risk monitoring of sensitive data operations

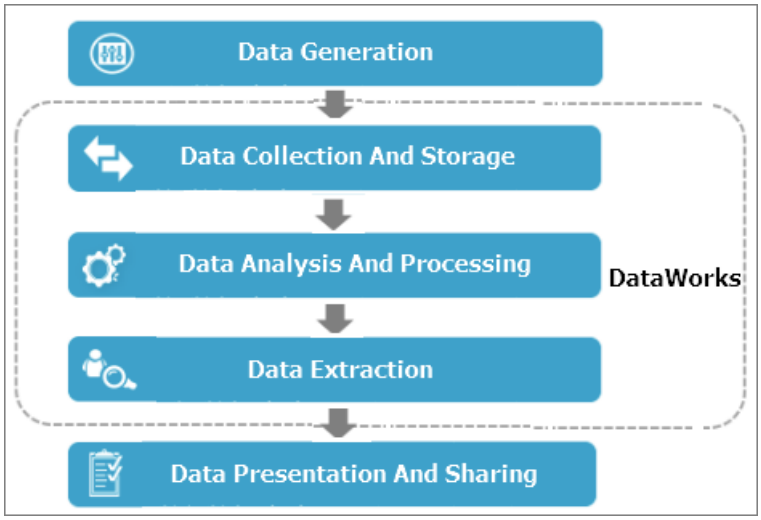
  DataWorks allows you to monitor data distribution, usage, and export in a visualized manner, and customize risk levels for auditing.

Related service: Data Security Guard of DataWorks



4.Data development process

Data development is a process of generating, collecting, storing, analyzing, computing, extracting, presenting, and sharing data.



Note As shown in the preceding figure, you can perform the steps in dashed-line boxes in DataWorks.

The data development process involves the following steps:

1. **Generate data**: Each business system generates a large amount of structured data every day and stores the data in its own databases, such as MySQL, Oracle, and RDS databases.
2. **Collect and store data**: You can synchronize data from business systems to MaxCompute and use the powerful data storage and processing capabilities of MaxCompute to analyze the data.

   The Data Integration service of DataWorks supports various connections. It allows you to synchronize data from business systems to MaxCompute based on the preset recurrence.
3. **Analyze and compute data**: After data synchronization, you can create ODPS SQL and ODPS MR nodes to process data in MaxCompute, and create other data analytics nodes to analyze and mine the data for value.
4. **Extract data**: You can export data processing and analysis results to business systems for further processing.
5. **Present and share data**: After data is extracted, you can present the big data processing and analysis results in multiple ways such as reports or a geographic information system (GIS). You can also share the results with others.

5.Basic mode and standard mode

DataWorks provides workspaces in basic mode and standard mode for you to develop data under different security control requirements. This topic describes the differences between and access accounts for workspaces in basic mode and standard mode.

Workspaces in basic mode

In basic mode, a DataWorks workspace can be bound to only one compute engine of each type, which can be a project, an instance, or a database. A workspace in basic mode does not isolate the development environment from the production environment. In this workspace, you can perform only basic data development but cannot completely control the data development process and table permissions.

A workspace in basic mode has the following benefits and risks:

- Benefits: This mode is easy to use. The scheduling system can periodically run a node to produce data immediately after you commit the node, without the need to deploy the node.
- Risks: Developers can add, modify, and commit code to the scheduling system without the need for approval. This makes the production environment unstable. In addition, if this workspace is bound to a MaxCompute compute engine, developers have the read and write permissions on all tables of the MaxCompute project by default. Developers can create, delete, and modify a table in the workspace. This puts data at risk.

Workspaces in standard mode

In standard mode, a DataWorks workspace can be bound to two compute engines of each type, which can be projects, instances, or databases. The standard mode differs from the basic mode in the following aspects:

- You can modify code only in the development environment, but cannot modify code in the production environment.
- After you commit a node, the scheduling system runs the node in the development environment as a smoke test. The node is not automatically scheduled. If you want this node to be automatically scheduled, you must deploy it to the production environment.

  You can deploy a node only after you obtain approval from a workspace administrator or an administration expert.

Access accounts for workspaces in basic mode and standard mode

You can specify the access accounts for workspaces in basic mode and standard mode in the **Computing Engine information** section of the **Workspace Management** page.

Workspace mode	Compute engine type	Environment	Access account
		Development environment	Only the current logon node owner can perform operations.

Workspace mode	MaxCompute Compute engine type	Environment	Access account
Standard mode		Production environment	The following accounts can be specified to perform operations: • Alibaba Cloud account • RAM user
	E-MapReduce	Development environment	Only the accounts with the AccessKey IDs and AccessKey secrets specified in the **New EMR cluster** dialog box can perform operations.
		Production environment	
	Hologres	Development environment	Only the current logon node owner can perform operations.
		Production environment	The following accounts can be specified to perform operations: • Alibaba Cloud account • RAM user
Basic mode	MaxCompute	Development environment, which is also the production environment	Only the current logon node owner can perform operations. The following accounts can be specified to perform operations: • Node owner • Alibaba Cloud account
	E-MapReduce	Development environment, which is also the production environment	Only the accounts with the AccessKey IDs and AccessKey secrets specified in the **New EMR cluster** dialog box can perform operations.
	Hologres	Development environment, which is also the production environment	Only the current logon node owner can perform operations. The following accounts can be specified to perform operations: • Alibaba Cloud account • RAM user