Alibaba Cloud DataWorks

Product Introduction

Issue: 20190516

Legal disclaimer

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

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

or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.

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

Generic conventions

Table -1: Style conventions

Style	Description	Example
⊖	This warning information indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	⊖ Danger: Resetting will result in the loss of user configuration data.
⚠	This warning information indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	⚠ Warning: Restarting will cause business interruption. About 10 minutes are required to restore business.
📋	This indicates warning information, supplementary instructions, and other content that the user must understand.	ⓘ Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	📋 Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	It is used for buttons, menus, page names, and other UI elements.	Click OK.
`Courier font`	It is used for commands.	Run the `cd / d C :/ windows` command to enter the Windows system folder.
Italics	It is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	It indicates that it is a optional value, and only one item can be selected.	`ipconfig` *[-all|-t]*

Style	Description	Example
{} or {a|b}	It indicates that it is a required value, and only one item can be selected.	`swich` *{stand | slave}*

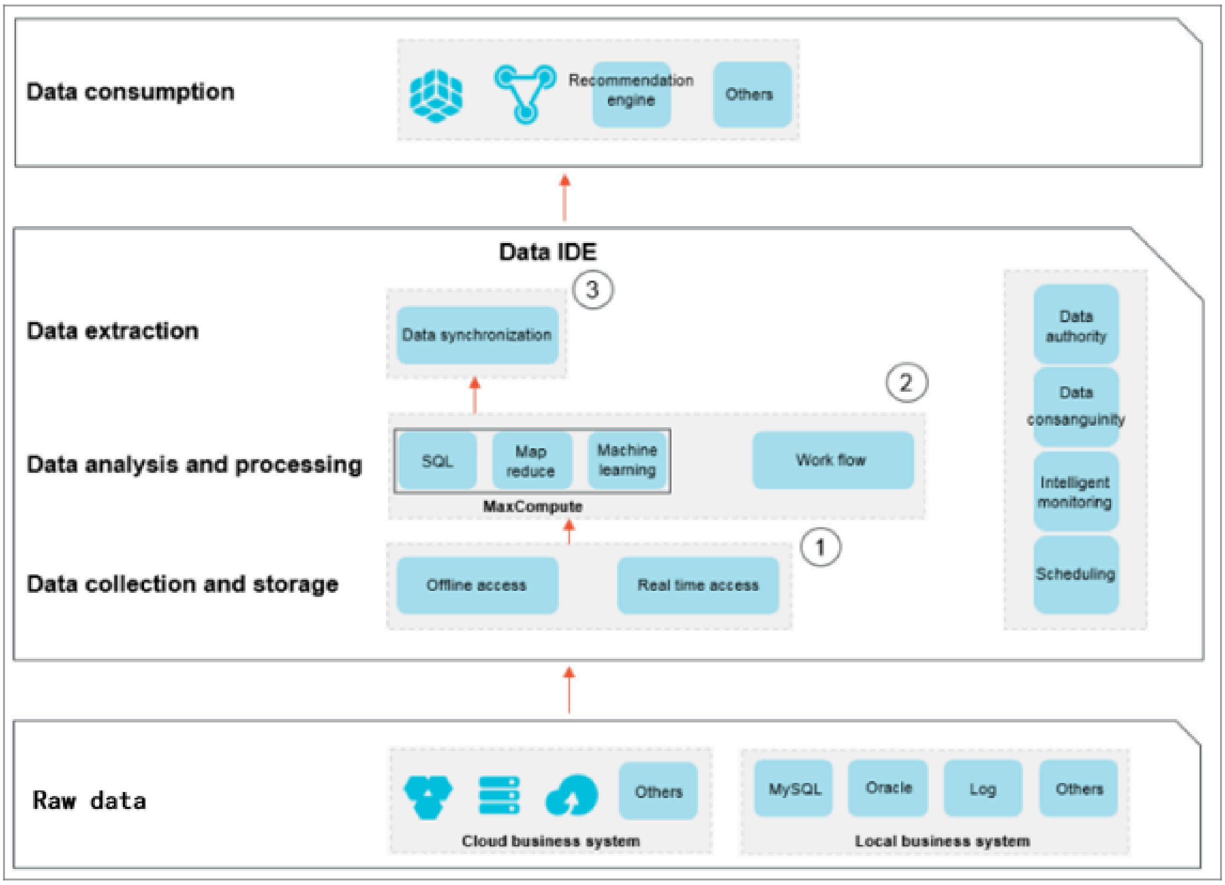
Contents

1 What is DataWorks?

DataWorks is an important Alibaba Cloud Platform as a service (PaaS) product . It offers fully hosted workflow services and a one-stop data development and management interface to help enterprises mine and comprehensively explore data value.

DataWorks uses MaxCompute as the core computing and storage engine to provide massive offline data processing, analysis, and mining capabilities. For more information, see MaxCompute overview.

DataWorks simplifies data transmission and conversion .You can import data from different storage services, convert and ultimately extract data that is transmitted to other data systems. See the following figure for a comprehensive overview of DataWorks data analysis process.



Features

- Fully-hosted scheduling

  DataWorks provides powerful scheduling capabilities. Based on Directed Acyclic Graph (DAG) relationships, the time-based or dependency-based tasks trigger configurations to perform tens of millions of tasks punctually and precisely daily . The multiple scheduling frequency configurations are supported by minute, hourly, daily, weekly, and monthly basis.

  The fully-hosted service eliminates all server resource scheduling concerns. The system isolates different tenants to guarantee tasks run independently.
- Supports various task types

  DataWorks supports multiple task types, such as data synchronization, SHELL, MaxCompute SQL, and MaxCompute MR tasks. Complex data analysis processes are based on dependencies between tasks.

  - Powered by MaxCompute, DataWorks provides powerful data conversion capabilities to guarantee high performance of big data analysis.
  - For data synchronization, DataWorks relies on powerful data integration capabilities to support more than 20 types of data sources and provide stable and highly-efficient data transmission. For more information, see *Data integration overview*.
- Data visualization development

  This product offers visualization code development and workflow designer pages . No additional development tools are required to drag and drop components to develop complex data analysis tasks. Development tasks can be performed from anywhere in the globe through Internet connection and web browsers.
- Monitoring and alarms

  The O&M center provides visual task monitoring and management tools, and displays global conditions in DAG format when tasks are running.

  SMS alarms can be easily configured to notify the relevant alarm contact of task errors for immediate troubleshooting.

Constraints and limits

- DataWorks only supports Chrome 54 or later versions.

- **Currently, DataWorks only supports SQL operations on Alibaba Cloud's MaxCompute.**

2 Concepts

Business flow

This topic describes DataWorks business flows, solutions, components, tasks, instances, submissions, script development, resources, functions, and output name concepts.

Advantages:

- Helps organize data codes from business perspectives supports code organization based on task types and multi-level sub-directories (Alibaba Cloud recommends no more than four levels).
- Provides work flow overview from business perspectives to facilitate optimization.
- Provides business flow dashboards for efficient development.
- Organizes release and maintenance based on business flows.

Solution

DataWorks offers customizable and integrable business flow solutions.

Advantages:

- Multiple business flows
- Reusable business flows for different solutions
- Comprehensive solutions for immersive development

Component

A component is a SQL code procedure template with multiple input and output parameters, SQL code procedures are generally handled by introducing one or more data table sources through filtering, connect, aggregate, and other operations to process target tables for new business needs. The common logic in SQL can be abstract components to enhance code reuse.

Task

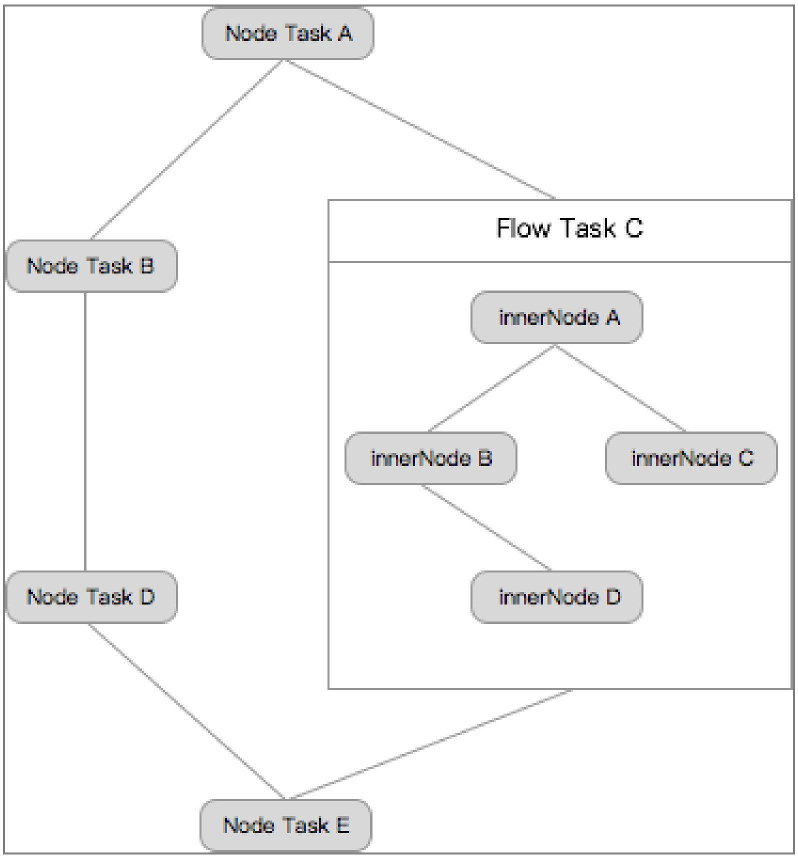
A task performs various data operations . The following describes various task applications:

- A data synchronization node task is used to copy data from RDS to MaxCompute.
- A MaxCompute SQL node task is used to run MaxCompute SQL for data conversion.

- A flow task is used to perform a series of data conversions among several inner SQL nodes.

Each task uses zero or more data tables (data sets) as an input, and generates one or more data tables (data sets) as the output.

Tasks are divided into node tasks, flow tasks, and inner nodes. See the relationships between these tasks in the following figure:



- A node task is a data operation. It can be configured to be dependent on other node tasks and flow tasks to form a Directed Acyclic Graph (DAG).
- A flow task is formed by a group of inner nodes that process a a work flow task. We recommend using less than 10 flow tasks. Inner nodes of a flow task cannot be dependencies of other flow or node tasks. A flow task can be configured to be a dependency of other flow and node tasks to form a DAG.
- An inner node is a node within a flow task. It basically has the same capabilities as a node task. Its scheduling cycle is inherited from the flow task scheduling frequency and cannot be configured independently. The dependency can only be dragged.

Data execution can be selected from an operation type, see *Node types overview*.

For details about task scheduling parameter configurations, see *Scheduling configuration*.

Instance

An instance is generated when a task is scheduled by the system or triggered manually. An instance is a snapshot that runs by a task at a certain time. The instance contains the task operating time, operating status, operating logs, and other information. For example:

Assume that Task 1 is configured to run at 02:00 each day. In this case, the scheduling system automatically generates a snapshot at the time predefined by the periodic node task at 23:30 each day. That is, the instance of Task 1 will run at 02:00 the next day. If the system detects the upstream task is complete, the system automatically runs the Task 1 instance at 02:00 the next day.

Note:

You can query task instance information on the O&M Center > Task O&M page.

Submit

Submit refers to the node task development process, and the work flow tasks from developing environments that are published to the scheduling systems. After a task is submitted, its code and scheduling configuration are synchronized to the scheduling system, which schedules the task according to the configuration.

Note:

Unsubmitted node tasks and flow tasks do not enter the scheduling system.

Script

A script is a code storage space for data analysis. The script code cannot be released to the scheduling system, and its scheduling parameters cannot be configured. It can only be used for data query and analysis.

Resources and functions

Resources and functions are both MaxCompute concepts. For details, see *MaxCompute resources* and *MaxCompute functions*.

In DataWorks, interfaces are used for resource and function management. Resources and functions managed through other MaxCompute methods cannot be queried in DataWorks.

Output name

The output name is the name of each task's output point. If users set dependencies within a Alibaba Cloud account single tenant, a virtual entity that connects upstream and downstream tasks. .

If a task is set to form upstream and downstream dependencies with other tasks, the setting must be based on the output name. The task output name is also the input name for the downstream node.

Note:

Similar to a task ID, the output name can be a unique conceptual object for a task that is different from other tasks in the same tenant. You can also add custom output names to a task, however, the output node name must be unique within the tenant.

3 Scenarios

Building a cloud platform for Internet big data application services

Features:

- Allows enterprises to focus on core businesses

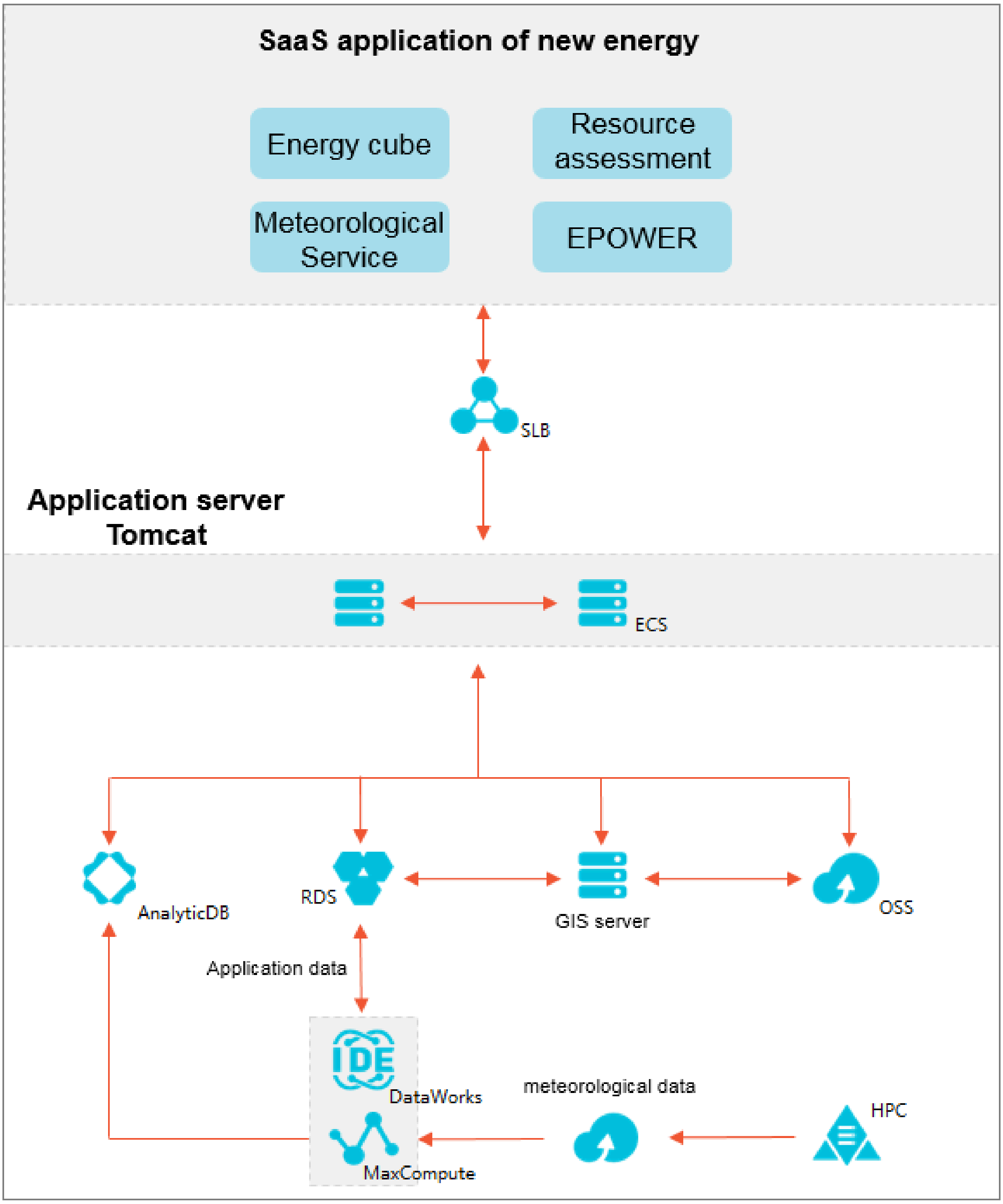
  The entire business infrastructure can be quickly migrated to Alibaba Cloud to optimize business productivity with available massive resources. Alibaba Cloud’s mature enterprise scaling solutions removes the need for enterprises to focus on scaling seamlessly and other related matters.
- Reduces investment and O&M costs

  Greatly reduces material resources, labor, and R&D investment required for on-premises big data platforms.
- Security and stability

  DataWorks comprehensive service capabilities foolproof data migration to the cloud and provides stable and assured performance.

Recommended combination:

DataWorks + AnalyticDB + MaxCompute



Weather queries and advertisement business log analysis

Features:

- Improves work efficiency

  All log data is parsed through SQLs, increasing work efficiency more than five times.

- Improves storage utilization

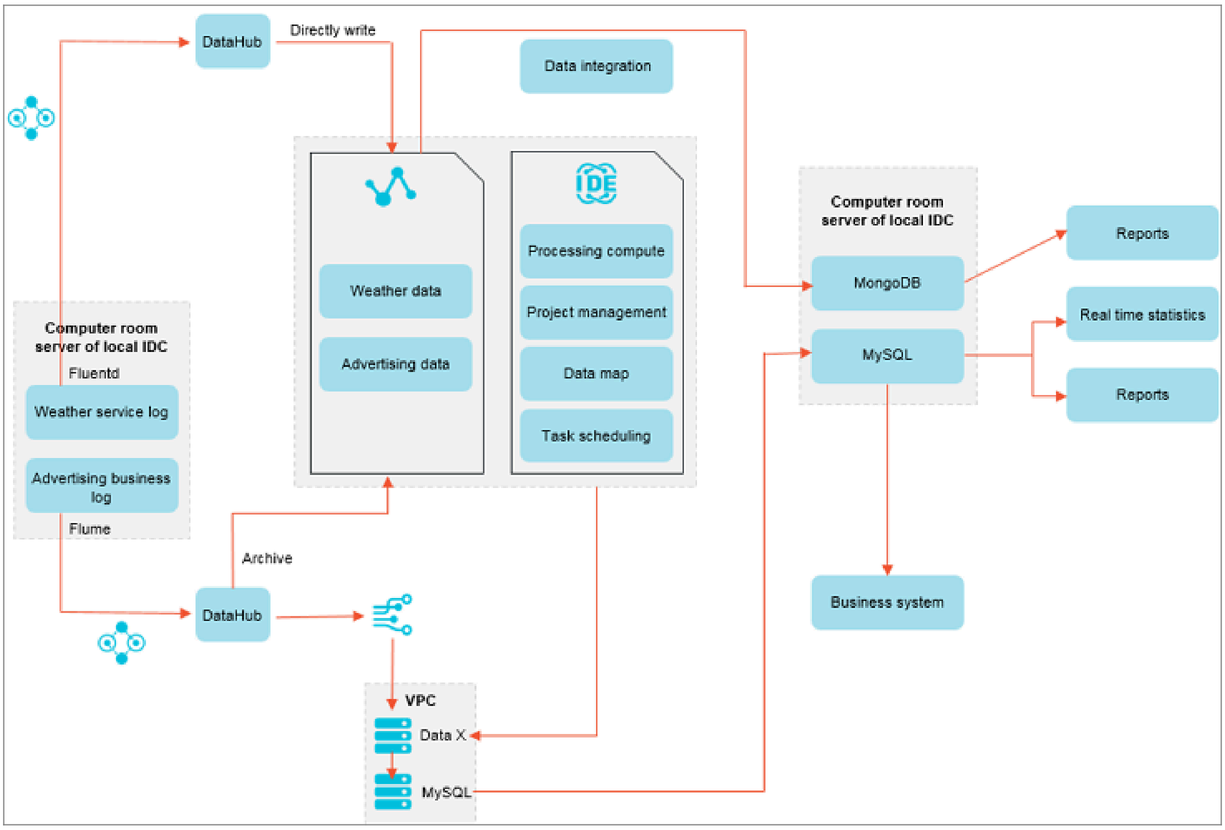
  DataWorks reduces overall storage and computing costs by 70%, and improves both performance and stability.

- Makes big data products easy to use

  MaxCompute provides plugins for multiple open-source softwares, allowing you to easily migrate data to the cloud.

Recommended combination:

DataWorks + Data Integration + AnalyticDB + Quick BI + MaxCompute



Delicacy management operations

- Improves business insights

  MaxCompute’ s computing capability can realize delicacy management operations for millions of users.

- Data-driven businesses

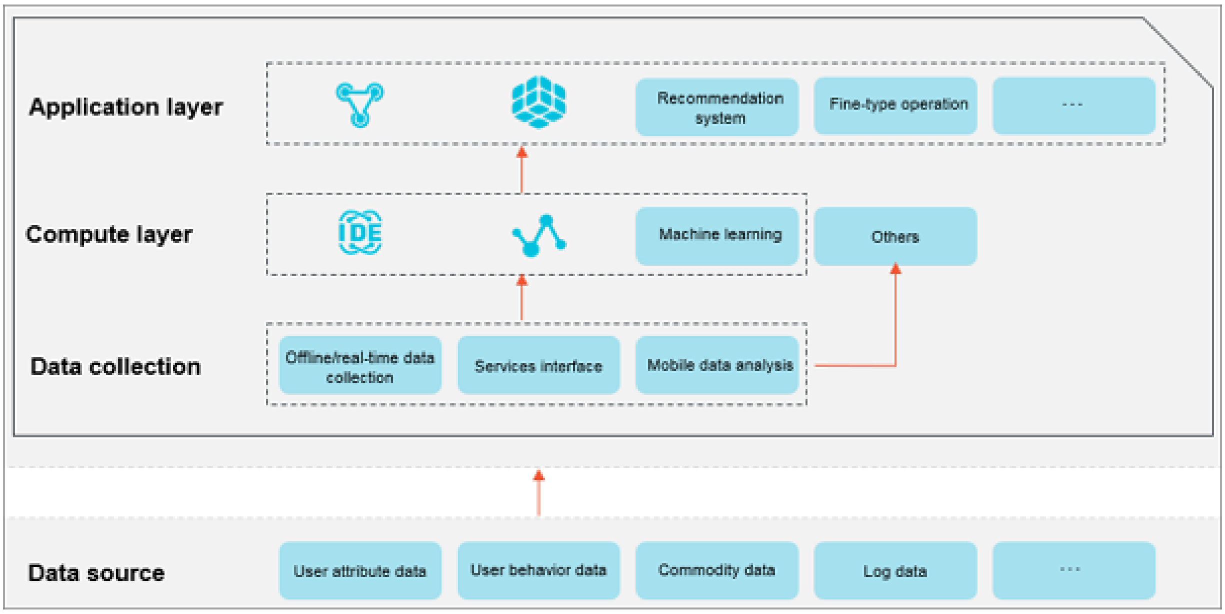
  DataWorks empowers businesses by providing enhanced data analysis capabilities and effective monitoring functions.

- Quick response to business requirements

  The DTplus ecosystem quickly responds to new business data analysis requirements.

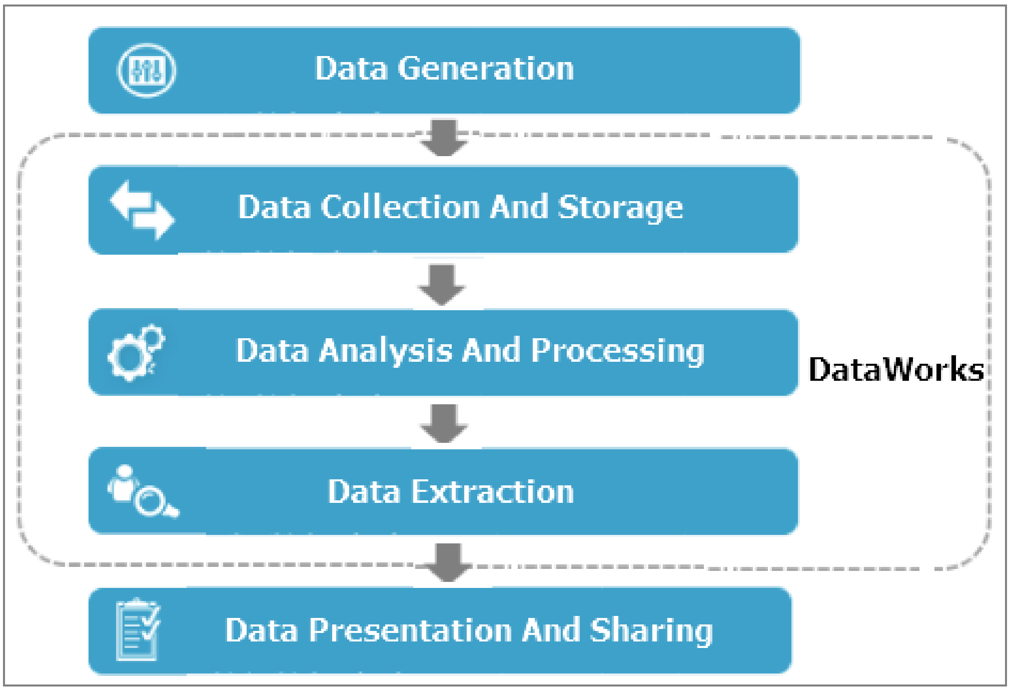
Recommended combination:

DataWorks + Data integration + Quick BI + MaxCompute



4 Data development process

The data development process comprises data generation, data collection and storage, data analysis and processing, data extraction, and data presentation and sharing. See the following graphical process representation .


Data Generation
Data Collection And Storage
Data Analysis And Processing
Data Extraction
Data Presentation And Sharing
DataWorks

Note:

In the preceding figure, data development processes within the dotted box are completed on the Alibaba Cloud Big Data Platform.

The data development process is as follows:

- Data generation

  A business system generates a large amount of structured data every day. The data is stored in business system databases, such as MySQL, Oracle, and RDS.

- Data collection and storage

  To use MaxCompute’s massive data storage and processing capabilities for data analysis, you must synchronize data from different business systems to MaxCompute.

  DataWorks provides data integration services so you can synchronize various data types from business systems to MaxCompute according to predefined scheduling periods.
- Data analysis and processing

  Next, you can process (MaxCompute_SQL and OPEN_MR), analyze, and mine (data analysis and data mining) the data on MaxCompute to find valuable information.
- Data extraction

  The data after analysis and processing must be synchronized to your business system for further use.
- Data presentation and sharing

  Finally, the results of big data analysis and processing are presented and shared as reports, geographical information systems, and in different accessible formats.

5 Simple mode and standard mode

The new version of DataWorks introduces both simple and standard modes, this article introduces you to the differences between simple and standard modes.

Simple Mode

A simple mode refers to a DataWorks project that corresponds to a MaxCompute project and cannot set up a development and Production Environment, you can only do simple data development without strong control over the data development process and table permissions.

The advantage of the simple mode is that the iteration is fast, and the code is submitted without publishing, it will take effect.

The risk of a simple mode is that the development role is too privileged to delete the tables under this project, there is a risk of table permissions.

Standard Mode

Standard mode refers to a DataWorks project corresponding to two MaxCompute projects, which can be set up to develop and produce dual environments, improve code development specifications and be able to strictly control table permissions, the operation of tables in Production Environments is prohibited, and the data security of production tables is guaranteed.

- All Task edits can be performed only in the Development Environment, and the Production Environment Code cannot be directly modified, reduce the Production Environment code modification entry, as much as possible to ensure the Production Environment code stability.
- The Development Environment does not turn on task scheduling by default, avoid the development of environmental project cycle operation and production of environmental projects to seize resources, the stability of the operation of Production Environment tasks is better guaranteed.
- The Production Environment runs with a default production account, all the tables produced by the production account belong to the main account, you need to use production tables during the development process, all of which need to be applied separately, better control of table permissions.

When creating a project, select project mode as the standard mode, fill in the project name and project description, the remaining configuration item select the default value.

Note:

The MaxCompute access identity of the Production Environment cannot be modified to a personal account, otherwise, the data security of the Production Environment cannot be guaranteed.

6 Version history

DataWorks V2.0 release

Release Version: DataWorks V2.0

- Release time July 25, 2018
- Release scope: East China 2 deployment only
- Release: DataWorks V2.0 adds business processes and components on the basis of DataWorks V2.0, it also improved the data R&D system, supports dual projects , isolated development and production, and ensures data development specificat ions to reduce error codes.
- You can watch videos to learn more about DataWorks V2.0:

Regions that support Dataworks 2.0

Regions that support Dataworks 2.0:

- East China 1
- East China 2
- North China 2
- South China 1

DataWorks V2.0 update list

DataWorks V2.0 update list					
DataWorks V2.0 has upgraded the overall visual interaction and data development module's usage experience. Furthermore, DataWorks V2.0 provides four new modules, including intelligent monitoring, data protection, data quality and data service. For a smooth transition to the newest version, the following is a list of all DataWorks V2.0 updates. .					
Module name	Sub-Module	Comparison	DataWorks V1. 0	DataWorks V2.0	Improved effects

DataWorks V2.0 update list					
MaxCompute ProjectManagement	**Project** Management Mode	methods	A DataWorks project corresponds to a MaxCompute project.	Introducing DataWorks "Standard Mode" concept. Under this mode, a project corresponds to two MaxCompute projects, including development environment and production environment. (See *Simple mode and standard mode*)	Isolate risks to protect code stability in production environments.

DataWorks V2.0 update list					
Data developmen t	Task developmen t	Overall function	Performs a single task, workflow code writing, cycle scheduling configuration. After completion , it can be submitted to the operation center for automatic scheduling.	· Renamed: Data Development · New: solutions, business process concepts · Deleted: workflow (concept) · Optimization: More intelligent SQL editor , task cycle configurat ion, more open dependency configurat ion.	1. SQL Editor : provides a more user-friendly and immersive SQL development experience. 2. Task Management: business processes, solutions make it easy to manage complex development tasks. 3. Task Scheduling: a more open scheduling system that can easily handle complex business scenarios. 4. Other features : optimized new features to take care of user pain points in detail.
		SQL R&D	Write SQL code on the page in the form of a single task or WorkFlow and test run it.	Provides a more intelligent SQL editor with code highlighting, formatting, intelligent supplement, error tips, table structure display and other user-friendly functions. At the same time, you can see the SQL internal structure visually in the graphical form.	
		Node configuration	Combine business code through single nodes and workflow modes.	Introduces the business process concept of a a workflow. You can combine tasks in a business process, and manage different resources in business processes based on their needs (all tasks, tables, resources, and functions must belong to a business process). You can also consolidate business processes in one step through a solution, unifying business process management with strong relevance.	

DataWorks V2.0 update list					
		Cycle configuration	The workflow overall cycle configuration affects the periodic configuration of individual tasks.	All nodes can be configured separately and the scheduling cycle type is not affected by upstream and downstream nodes.	
		Dependency attributes	The dependencies between workflows are limited.	Task nodes in different business processes can be dependencies, and do not need to be dependencies of the business processes.	
	Script development	Overall function	A periodic task supplement usually used for non-periodic temporary data processing.	Same function, renamed as manual business process.	
	Resource management	Overall function	Manage all resources in the MaxCompute project as a separate tab, including jar, file , and archive.	As a sub-label in the business process, users can join resources involved in the business process on demand, while creating multi-tiered folders for management.	
	Function management	Overall function	A separate tag that manages the system and custom functions required for the MaxCompute SQL edit.	Can manage all functions as a separate tag or as a subtab in a work process that manages required functions.	

DataWorks V2.0 update list						
	Table query	Overall function	Shows alltables under the MaxCompute project, with the ability to preview the content, reference, and representation.	Same	
	Table Management (new)	Overall function	None	For developers to manage their own tables, life cycle settings, and table management. Supports table management features include modifying the category , description, field, partition, hide or show table , delete table, and more.	
	Temporary query (new)	Overall function	None	Used to test if the code matches expectations. Does not contain the following features : submit, publish, set schedule, and parameters function.	
	Component Management (new)	Overall function	None	Abstracts a large number of similar and reusable SQL code in SQL code blocks or node tasks, you can configure input and output arguments, and apply it to a variety of practical businesses.	

DataWorks V2.0 update list					
	Run history (new)	Overall function	None	Displays all task records that were run locally in the last three days, you can also view the task run results and provide simple filtering capabilities.	
	Results filter (new)	Overall function	None	Provides SQL results integrating the Excel component, allows users to obtain expected results by filtering, and ordering after the page prints the results.	
	Recycle Bin (new)	Overall function	None	Prevents business losses caused by mis-deleting user tasks, you can see all deleted nodes under the current item in the recycle bin and provide recovery capabilities.	
	Code global search	Overall function	None	You can input an incomplete string to find the MaxCompute SQL, Shell, Data Synchronization tasks, and quickly locate tasks you need to view or operate.	
	Release function	Overall function	Keeps the publishing function under DataWorksV1 standard mode project.	Rename: Project clone. Simple schema projects have the ability to clone tasks to other projects automatically.	

DataWorks V2.0 update list					
O&M center	Task list	Feature	Search for tasks based on the node type, name, and owner.	Adds the ability to search for tasks through business processes, solutions, and baseline names.	From a business perspective, the task O&M matches new features on the task development interface.
	Task O&amp; M	Feature	Search for tasks based on node type, name, owner, business date, and run date.	Adds the ability to search for tasks through business processes, solutions, and baseline names.	
	Alert	Feature	The monitoring alarm is based on the following: errors, completion, incomplete events and more.	Integrate Baseline Monitoring, incident alarm, custom alarm three functions to build a more intelligent, and complete alarm system.	
Intelligent Monitoring (new)	Alarm is a monitoring and analysis system for running DataWorks tasks. Based on monitoring rules and task operation, Intelligent Monitor decides whether to alert, when to alert, how to alert, and who to alert. Intelligent Monitor automatically selects the most appropriate alert time, alert method, and alert object.				Give you a one-stop access to data development and data on the cloud (secure) and a closed-loop experience of governance and data sharing.
Data quality (new)	Data quality is a one-stop platform that supports quality verification, notification, and management services for a wide range of heterogeneous data sources. DQC uses a dataset as a monitoring object, supports monitoring MaxCompute data table, and DataHub real-time data stream. When changes are made to offline MaxCompute data changes, DQC verifies the data and blocks the production link to avoid diffusion of problematic data pollution. Furthermore, DQC provides verification of historical results. Thus, you can analyze and quantify data quality.				

DataWorks V2.0 update list		
Data Service (new)	DataService studio provides the ability to quickly generate data APIs from data tables, enabling users to quickly register existing APIs to a data service platform for unified management and publishing. In addition, Data Service is connected to API Gateway. You can deploy APIs to API Gateway with one-click. Data Service is compatible with API Gateway to provide a secure, stable, low-cost, and easy-to-use data sharing service.	
Data umbrella (new)	Data Security Guard provides data asset identification, sensitive data discovery, data classification, desensitization, monitor access ability, identify, alert and audit.	