Alibaba Cloud DataWorks

Product Introduction

Issue: 20180911

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.	📋 **Note:** 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]*
{} or {a|b}	It indicates that it is a required value, and only one item can be selected.	`swich` *{stand | slave}*

Contents

1 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 the concept of business processes and components on the basis of dataworks V2.0, it has also improved the data research and development system, supported the development of dual projects, and isolated development and production, ensure data development specifications to reduce the emergence of error codes.

Dataworks V2.0 update list

Dataworks V2.0 update list					
Dataworks V2.0 has revolutionary improved 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. To offer you smooth transition between old and new versions, here is a Dataworks V2.0 update list.					
Module name	**Sub-Module**	**Comparison**	**Dataworks V1. 0**	**Dataworks V2.0**	**Improved effects**
Maxcompute Project	Project Management Mode	Management methods	A dataworks project corresponds to a maxcompute project.	Introducing the concept of "Standard Mode", a dataworks project corresponds to two maxcompute projects, respectively: development environment, production environment. (See: *differences between simple and standard* modes)	Isolate risks to protect code stability in production environments.
Data development	Task development	Overall function	Perform single task, workflow code writing, cycle scheduling configuration. After completion, it can be submitted to the operation center	• Renamed: Data Development • New: solutions, business process concepts • Cut: workflow (concept) • Optimization: SQL editor more intelligent, better task cycle configuration,	**1.** SQL Editor: provides a more humanized, immersed SQL development experience. **2.** Task Management: **business**

Dataworks V2.0 update list					
			for automatic scheduling.	dependency configuration more open	**processes**, **solutions** make it easy to manage complex development tasks. **3.** Task Scheduling: a more open scheduling system that can easily handle more complex business scenarios. **4.** Other features: new features to take care of the user experience in detail.
		SQL research and development	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 replenishment**, **error tips**, **table structure display** and other humanized 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.	Introducing the concept of a **business process** instead of a **workflow**. You can freely combine tasks in a business process, and manage different resources into business processes according to their needs (all tasks, tables, resources, functions must belong to a business process). You can also consolidate business processes in one step through a solution, unifying management of business processes with strong relevance.	
		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 the upstream and downstream nodes.	
		Dependence attributes	The dependencies between workflows are limited.	Task nodes in different business processes can be dependent on each other, and do not need to rely on the overall business process.	

Dataworks V2.0 update list					
	Script development	Overall function	As a supplement to periodic tasks, it is 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/archive.	As a sub-label in the business process, users can join the resources involved in the business process on demand, while creating multi-tiered folders for management.	
	Function management	Overall function	As a separate tag, manage the system and custom functions that are required for the MaxCompute SQL edit.	Can exist as a separate tag and manage all functions, can also manage only the functions that are required for that business process as a subtab in the business process.	
	Table query	Overall function	Shows all the tables 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, table management including modifying the category, description, field , partition , table hiding/unhiding, table deleting, and so on.	
	Temporary query (new)	Overall function	None	Used to test if the code matches the expectations with no commit, publish, set	

Dataworks V2.0 update list					
				schedule parameters function .	
	Component Management (new)	Overall function	None	Abstract a large number of similar and reusable SQL code into SQL code blocks or node tasks, the user is free to configure input and output parameters, and apply it to a variety of practical business.	
	Run history (new)	Overall function	None	Display all task records that have run locally in the last three days, you can also view the results of the task run and provide simple filtering capabilities.	
	Results filter (new)	Overall function	None	Provide SQL results integrating the Excel component, allow users to get the desired results by simply filtering, filtering, and ordering after the page prints the results.	
	Recycle Bin (new)	Overall function	None	Used to prevent business losses caused by misdeleting of user tasks, you can see all the 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 task, quickly locating tasks that you need to view or manipulate.	
O&M center	Task list	Feature	Search for tasks based on node	Add the ability to search for tasks through **business**	From a business perspective, the task is operated to match the new

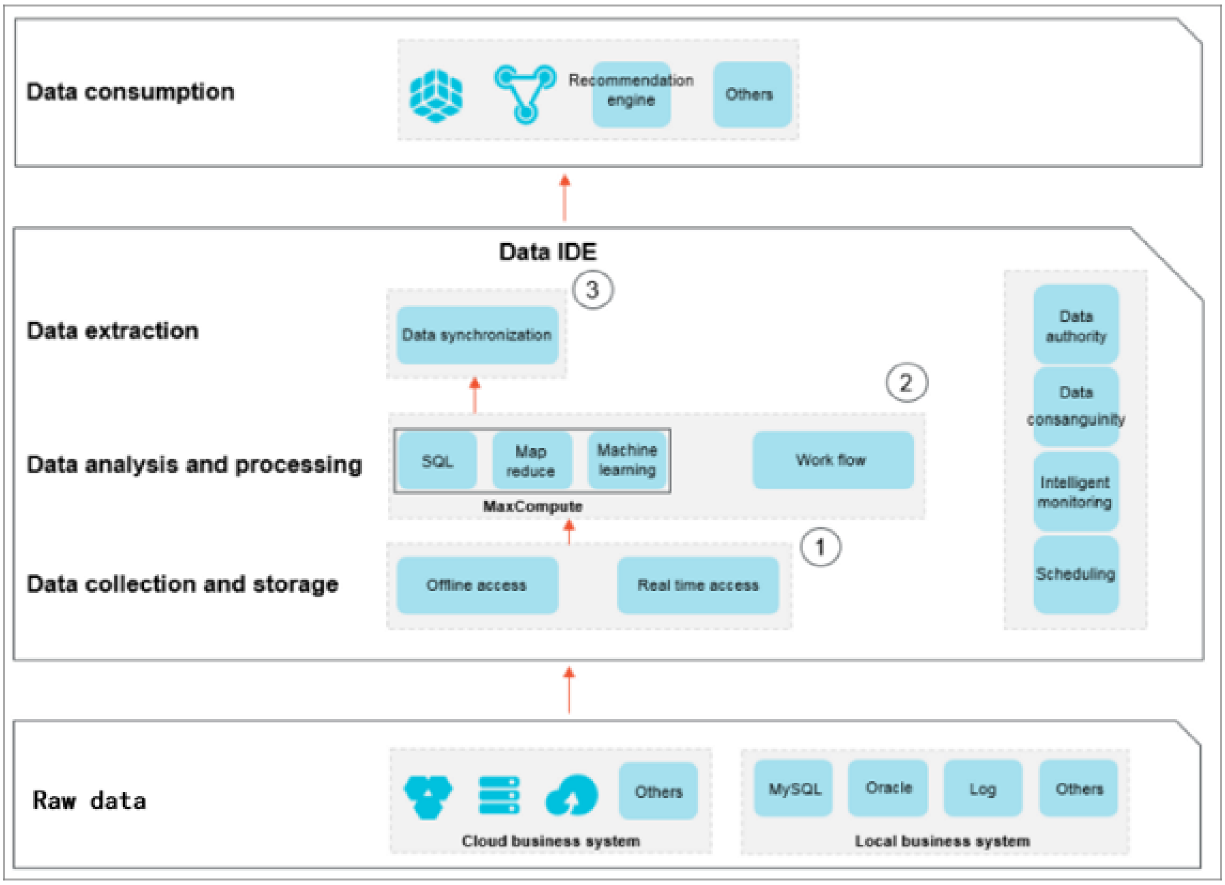
Dataworks V2.0 update list					
			type, name, and owner.	**processes**, **solutions**, **baseline names**.	features of the task development interface.
	Task O&amp; M	Feature	Search for tasks based on node type, name, owner , business date, and run date.	Add the ability to search for tasks through **business processes**, **solutions**, **baseline names**.	
	Alert	Feature	Through error , completion, incomplete events and so on as the basis of monitoring alarm.	Integrate **Baseline Monitoring**, **incident alarm**, **custom alarm** three functions to build a more intelligent, complete alarm system.	
Intelligent Monitoring (new)	*Intelligent Monitor* is a monitoring and analysis system for the running of DataWorks tasks. Based on monitoring rules and task operation, Intelligent Monitor decides whether or not to alert, when to alert, how to alert, and who to target. Intelligent Monitor automatically selects the most appropriate alert time, alert method, and alert object.				Give you 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)	*Dataworks data quality (DQC)* 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, monitoring MaxCompute data table and the DataHub real-time data stream . When the offline MaxCompute data changes, DQC verifies the data and blocks the production link to avoid the diffusion of problematic data pollution. Furthermore, DQC provides verification of historical results. Thus, you can analyze and quantify data quality.				
Data Services (new)	*Data Services* provides the ability to quickly generate data APIs from data tables, enables users to quickly register existing APIs to a data services 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 works together with API Gateway to provide a secure, stable, low-cost, and easy-to-use data sharing service.				
Data umbrella (new)	*Data protection for maxcompute* provides data asset identification, sensitive data discovery, data classification, desensitization, access ability to monitor, identify, alert and audit.				

2 What is DataWorks

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

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

DataWorks makes data transmission and conversion a lot more easier. It allows you to perform further data operations. You can import data from different storage services, and convert and ultimately extract the data to other data systems. See the following figure to have a complete insight about the data analysis.



Function overview

- **Fully-hosted scheduling**

  DataWorks provides powerful scheduling capabilities. Based on DAG relationships, the time-based or dependency-based tasks trigger configurations to perform tens of millions of tasks on time with maximum accuracy each day. The multiple scheduling frequency configurations are supported by minute-to-minute, hourly, daily, weekly, and monthly basis.

The fully-hosted service eliminates all your concerns about scheduling server resources. The system isolates different tenants that guarantees the tasks run independently.

- **Supports various task types**

  DataWorks supports multiple task types, such as data synchronization, SHELL, MaxCompute SQL, and MaxCompute MR tasks. The dependencies between tasks form complex data analysis processes.

  - Powered by MaxCompute, DataWorks provides powerful data conversion capabilities to guarantee high performance of big data analysis.
  - For data synchronization, DataWorks relies on DataWorks' powerful data integration capabilities to support over 20 data sources and provide stable and a highly-efficient data transmission. For more information, see *Overview of data integration*.
- **Visual development**

  This product offers visual code development and workflow designer pages. Without additional development tools, you can drag and drop components to develop complex data analysis tasks . A browser with Internet connection alone equips you to carry out development tasks wherever you are.
- **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.

  You can easily configure the SMS alarm. If the task is wrong, you can notify the relevant students in time to ensure the normal operation of the business.

Constraints and limits

- DataWorks only supports Chrome 54 or later.
- Currently, DataWorks only supports SQL operations on MaxCompute, instead of Alibaba Cloud ApsaraDB or Analytic DB.

3 Concepts

Business flow

Business flows: for business entities, the concept of business flows is abstract, enable users to organize data code development from a business perspective to improve task management efficiency. Business flows can be used repeatedly in different solutions.

Advantages:

- Helps organize codes from business perspectives in a clearer way. Supports the code organization based on task types and multi-level sub-directories (no more than levels).
- Views the overall work flows from business perspectives for optimization.
- Provides dashboards of business flows for efficient development.
- Organize the release and maintenance according to business flows.

Solution

Solution: You can customize and combine some business flows into a solution in a self-defining manner.

Advantages:

- Multiple business flows
- Reusable business flows in different solutions
- Complete 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 source data tables through filtering, connect, aggregate, and other operations to process target tables for new business needs . The common logic in SQL can be abstract into components to enhance code reuse.

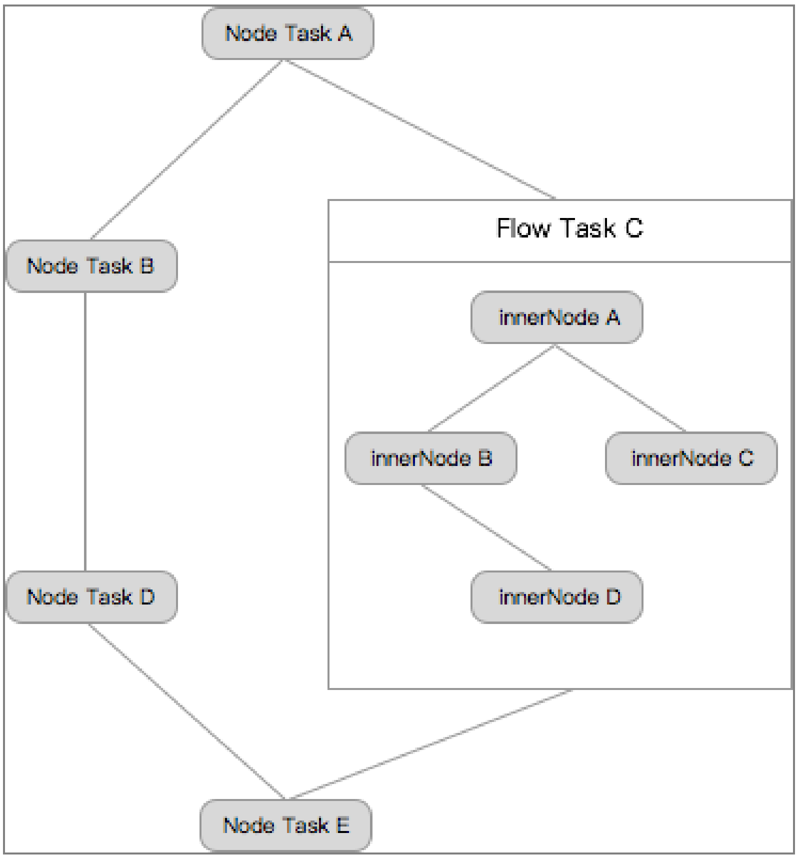
Task

A task is used to perform various operations on data. The following describes the uses of various tasks:

- 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 an operation performed on data. 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 are processing a small business. We recommend using less than 10 flow tasks. Inner nodes of a flow task cannot depend on by other flow or node tasks. A flow task can be configured to be dependent on other flow and node tasks to form a DAG.
- An inner node is a node inside a flow task. It basically provides the same capabilities as a node task. Its scheduling frequency is inherited from the scheduling frequency of the flow task, and cannot be configured independently. The dependency can only be dragged.

Data execution can be selected as an operation type, see *Introduction of Node Type*.

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

Instance

When a task is scheduled by the system or triggered manually, an instance is generated. An instance is a snapshot that runs by a task at a certain moment. 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 to be run at 02:00 the next day. When it is detected that 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

Commit refers to the process of developing node tasks, work flow tasks from developing environments to publish to 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:

Node tasks and flow tasks that are not submitted do not enter the scheduling system.

Script

A script is a code storage space that is provided 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, you can use interfaces for resource and function management. Resources and functions that are managed through other MaxCompute methods, cannot be queried in DataWorks .

Output name

The output name is the name of each task's output point, it is when a user sets dependencies within a single tenant (Alibaba Cloud account. A virtual entity that connects two tasks upstream and downstream.

When the user is setting up a task to form upstream-downstream dependencies with other tasks, the setting must be done based on the output name (not the node name or node ID). The output name of the task is also used as the input name for its downstream node.

Note:

Similar to the ID of the task, the output name can also be a unique conceptual object for a task that is different from other tasks in the same tenant. Users can also add custom output names to a task, note, however, that the output node name is not allowed to repeat within the tenant.

4 Scenarios

Builds a cloud platform for Internet big data application services

Can be achieved:

- Allows enterprises to focus more on the core business

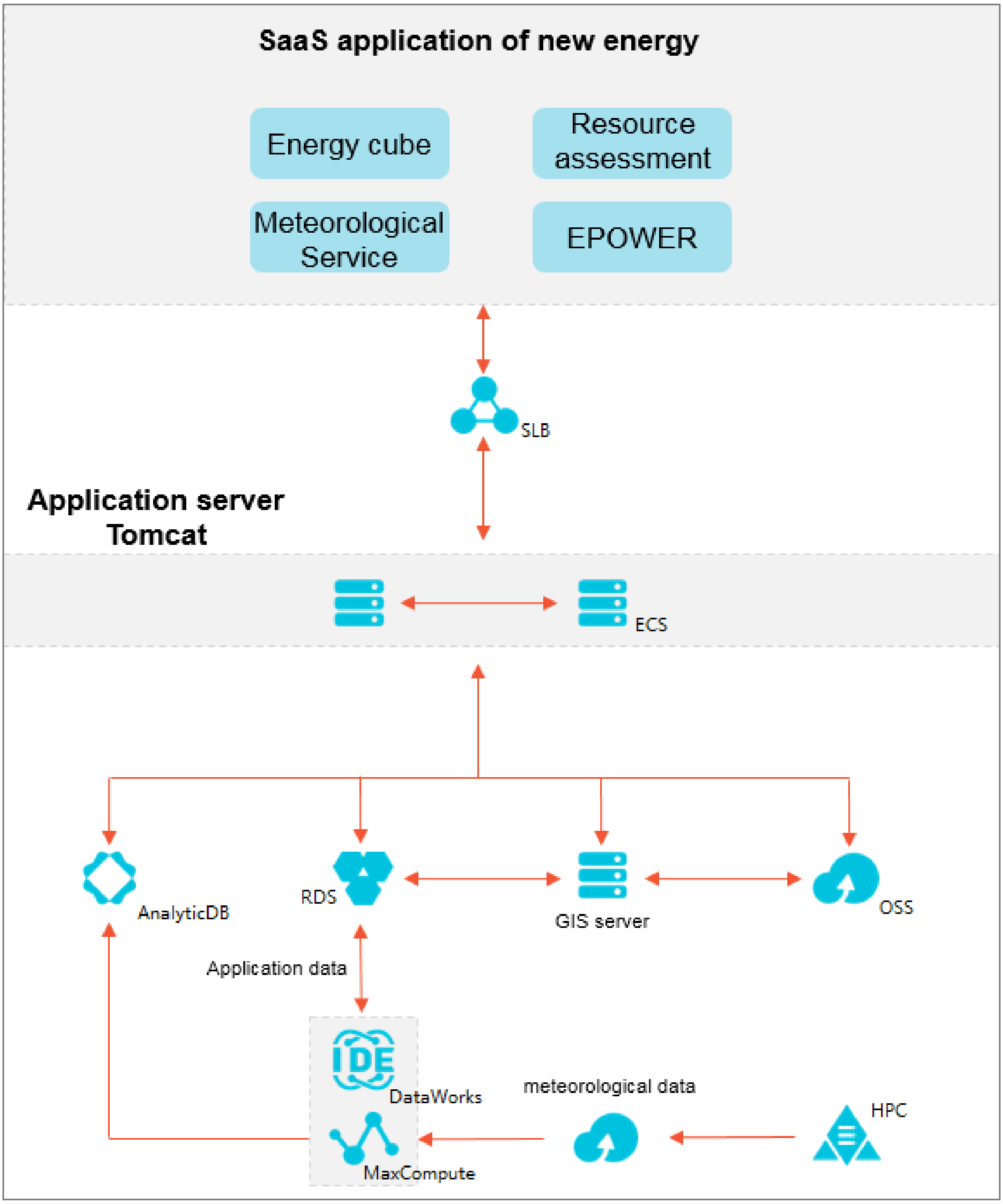
  Your entire business infrastructure can be migrated to the Alibaba Cloud much sooner than you have imagined. This way, you can make maximum use of the massive resources that Alibaba Cloud offer and optimize business productivity. With Alibaba Cloud’s mature business scaling solutions, enterprises do not need to focus too much on seamless service expansion and other allied matters.
- Reduces investment and O&M costs

  It can greatly reduce the material resources, labor, and R&D investment required for any self-built big data platforms.
- Security and stability

  Foolproof data migration to the cloud is guaranteed by DataWorks’s comprehensive service capabilities providing stable and assured performance.

Recommended combination:

DataWorks + AnalyticDB + MaxCompute



Weather queries and advertisement business log analysis

Can be achieved:

- Improves work efficiency

  All log data is analyzed based on SQLs, increasing work efficiency more than five times over.

- Improves storage utilization

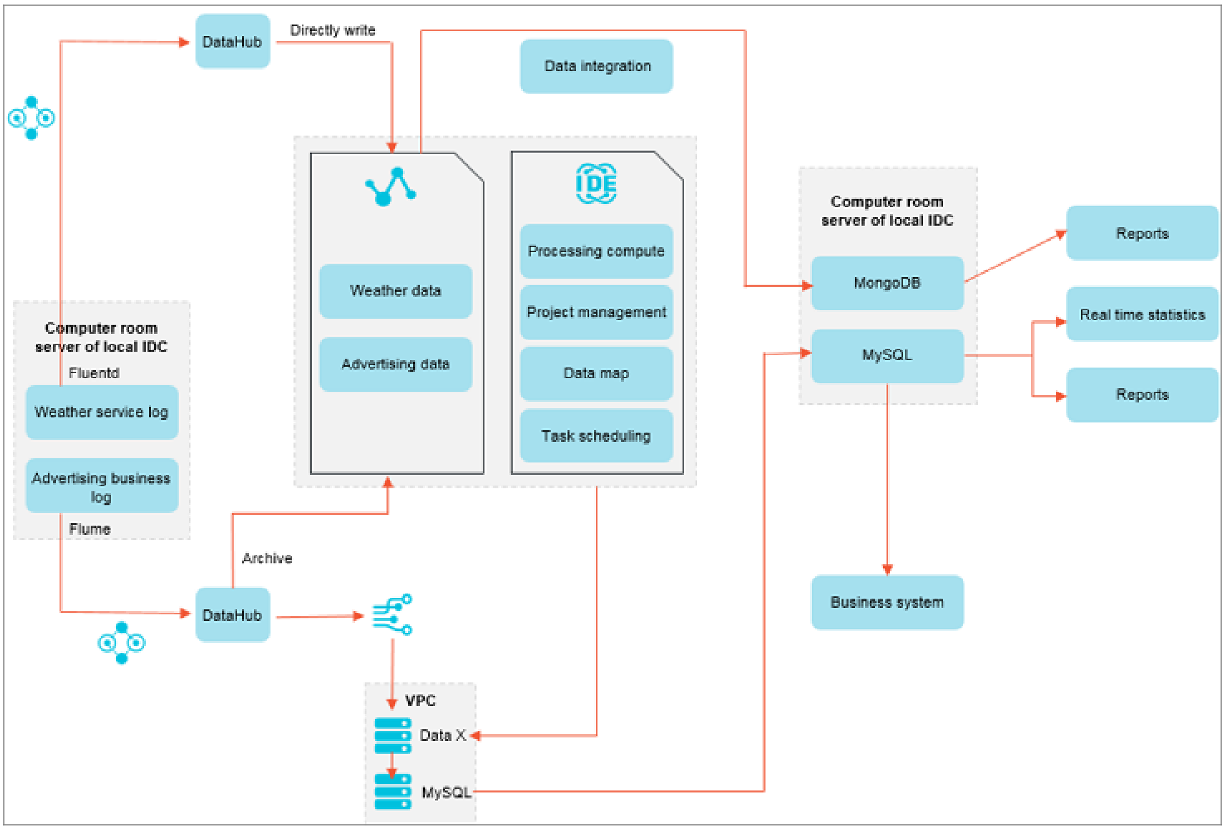
DataWorks can reduce overall storage and computing costs by 70%, improving both performance and stability.

- Makes big data products easy to use

  MaxCompute provides plugins for a wide range of open-source software, allowing you to easily migrate data to the cloud.

Recommended combination:

DataWorks + Data Integration + AnalyticDB + Quick BI + MaxCompute



Detail-oriented operations

- Improves business insights

  MaxCompute’s computing capability can achieve detailed-oriented operations for millions of users.

- Data-driven businesses

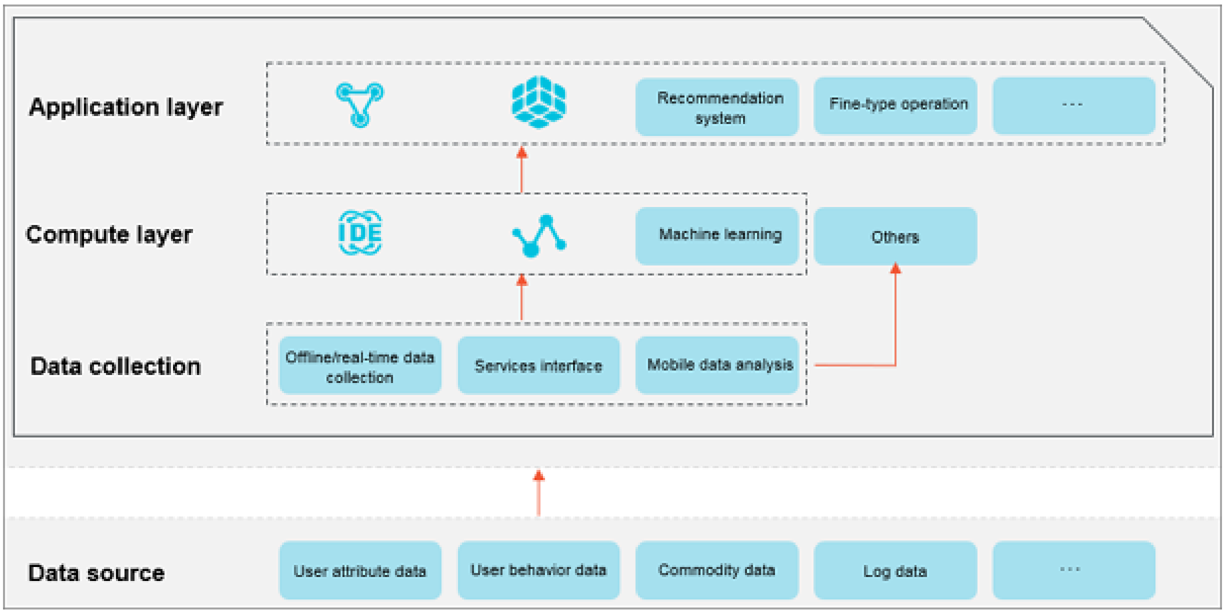
  DataWorks empowers businesses by providing enhanced data analysis capabilities and effective monitoring functions.

- Quick response to business needs

The DTplus ecosystem quickly responds to the new business data analysis needs.

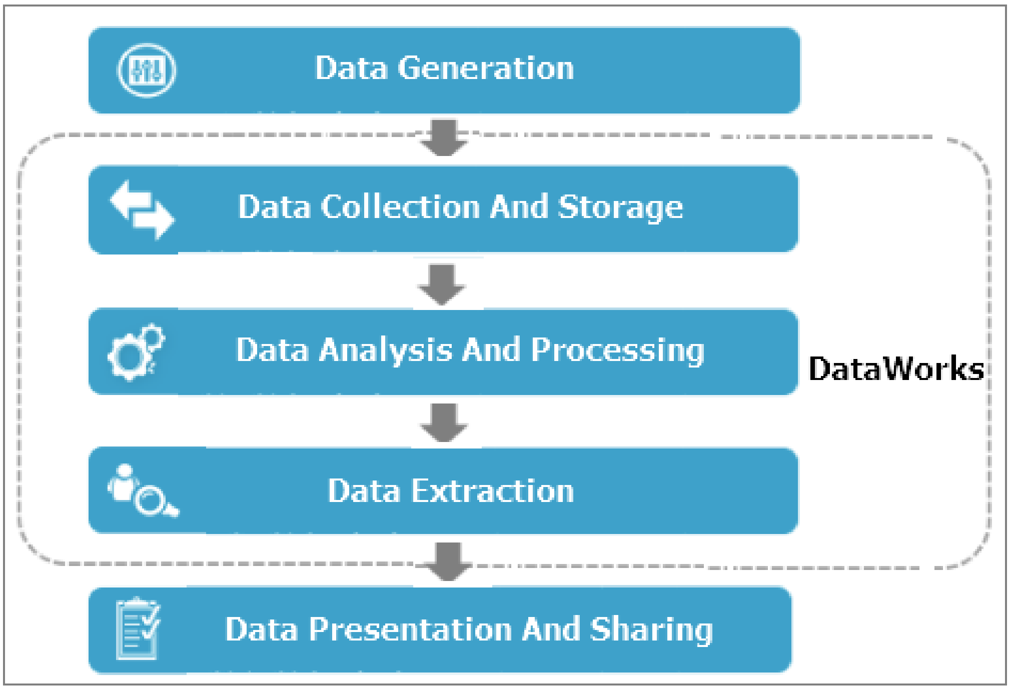
Recommended combination:

DataWorks + Data integration + Quick BI + MaxCompute



5 Data development process

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


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

Note:

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

The data development process is explained 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 the data from different business systems to MaxCompute.

  DataWorks provides data integration services for you to synchronize various types of data from business systems to MaxCompute according to predefined scheduling periods.

- **Data analysis and processing**

Next, you can start to process (ODPS_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 number of different accessible formats.