Alibaba Cloud

DataWorks

Update Records

Document Version: 20220106

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.Version history
1.1. DataWorks V2.0

This topic describes the release notes and new features of DataWorks V2.0.

Release information

Version: DataWorks V2.0

- Date: July 25, 2018
- Region: China (Shanghai)
- Content: Based on DataWorks V1.0, DataWorks V2.0 adds some concepts including the workflow and SQL script template, and also improves the data development process. DataWorks V2.0 supports workspaces in basic mode and standard mode. It can isolate the development environment from the production environment to help you develop data in a standard manner and reduce errors in code.

Available regions

Currently, all regions support DataWorks V2.0.

Operation updates

- DataWorks V2.0 does not display retroactive instances created in earlier versions. To check whether a retroactive instance is created, choose **Data Management > Tables > Partitions** to check whether the corresponding time partition is generated. If the time partition does not appear as expected, we recommend that you create a retroactive instance.

  If you fail to determine whether the retroactive instance is created, submit a ticket.

Feature updates

DataWorks V2.0 improves the overall visual interaction and user experience of the **DataStudio** service. In addition, it adds four services: Monitor, Data Security Guard, Data Quality, and DataService Studio.

Service	Feature	Item	DataWorks V1.0	DataWorks V2.0	Description
MaxCompute project	Project management	Project management method	A DataWorks workspace is associated with one MaxCompute project.	The standard mode is added for workspaces. In standard mode, a DataWorks workspace is associated with two MaxCompute projects to isolate the development environment from the production environment. For more information, see Basic mode and standard mode.	The development environment can be isolated from the production environment to guarantee the stability of code in the production environment.

Service	Feature	Item	DataWorks V1.0	DataWorks V2.0	Description
		Overall feature	You can develop code and configure scheduling properties for a single node or a single flow. Then, you can commit the node or flow to Operation Center for automatic scheduling and running.	• The service name is changed to DataStudio. • The concepts of a solution and a workflow are added. • The concept of a flow is deleted. • The SQL editor is more intelligent. More permissions are granted on node scheduling and dependency configurations.	
		SQL developm ent	You can develop and test code for a single node or a single flow in the SQL editor.	The SQL editor provides more intelligent and user-friendly features, including **syntax highlight**, **code formatting**, **intelligent code completion**, **error prompts**, and **schemas**. The Code Structure tab on the right of the SQL editor displays the SQL code structure in an intuitive manner.	

Service	Feature development	Item	DataWorks V1.0	DataWorks V2.0	Description
		Node configuration	You can develop your business only by associating nodes and flows.	**Flows** are replaced by **workflows**. You can develop your business by associating nodes in a workflow and adding the required tables, resources, and functions to the workflow. You can also integrate correlated workflows into a solution for unified management.	
		Recurrence configuration	The recurrence configuration of a node is subject to the flow recurrence.	You cannot configure the recurrence for associated nodes as a whole. Instead, you must configure the recurrence for each node separately. The recurrence type of a node can be the same as or different from that of its ancestor or descendant nodes.	
		Dependency	You can only configure dependencies between flows.	You can configure dependencies between nodes of different workflows.	

Service	Feature	Item	DataWorks V1.0	DataWorks V2.0	Description
DataStudio	Script	Overall feature	You can use scripts to process acylic temporary data. For example, you can use a script to add, delete, or modify a temporary table. This feature is a supplement to auto triggered nodes and does not require recurrence and dependency configurations.	The feature name is changed to **Ad-Hoc Query**.	• The SQL editor is optimized to provide more user-friendly and immersive SQL development experience. • Node development and management are simplified by using **workflows** and **solutions**. • More permissions are granted in the scheduling system to handle complex business demands. • Some other features are added to
	Manually triggered node	Overall feature	In a manually triggered flow task, all nodes must be manually triggered and cannot be automatically scheduled by DataWorks.	The feature name is changed to **Manually Triggered Workflows**.	
	Resource management	Overall feature	Resource management is an independent feature used to manage all resources in a MaxCompute project, including the JAR, file, and archive resources.	Resource management is changed to a feature of workflows. You can add the required resources to a workflow and create multi-level folders to manage these resources.	
	Function management	Overall feature	Function management is an independent feature used to manage the built-in functions and custom functions required by an ODPS SQL node.	Function management is an independent feature used to manage all the functions. It is also a feature of workflows and allows you to manage the functions of a workflow.	

Service	Feature	Item	DataWorks V1.0	DataWorks V2.0	Description
					resolve pain points and improve user experience.
	Table query	Overall feature	You can find, preview, and reference all tables of a MaxCompute project.	No update.	
	Table management	Overall feature	Not supported.	Table management is added. It allows developers to manage their own tables. For example, developers can set the time-to-live (TTL) and modify the category, description, fields, and partitions for tables, and hide, show, and delete tables.	
	Ad-hoc query	Overall feature	Not supported.	The ad-hoc query feature is added to test code in the development environment. You do not need to commit and deploy ad-hoc query nodes or set scheduling parameters for ad-hoc query nodes.	

Service	Feature	Item	DataWorks V1.0	DataWorks V2.0	Description
	SQL script template management	Overall feature	Not supported.	SQL script template management is added. It allows you to abstract SQL code as an SQL script template to reuse the SQL code. You can select SQL script templates and configure input and output parameters for these templates based on your business requirements.	
	Runtime logs	Overall feature	Not supported.	The Runtime Logs tab is added to show the records of all nodes that have been run in the development environment in the last three days. You can view and filter the running results.	
	Filtering of SQL results	Overall feature	Not supported.	An Excel component is integrated into the SQL editor so that you can filter, screen, and sort the running results of SQL statements.	
	Recycle bin	Overall feature	Not supported.	A recycle bin is added to prevent misoperation. You can view the nodes deleted from the current workspace in the recycle bin and restore them as needed.	

Service	Feature	Item	DataWorks V1.0	DataWorks V2.0	Description
	Global code search	Overall feature	Not supported.	Global code search is supported. You can search for an ODPS SQL, Shell, or sync node by entering a part of a string included in the node to quickly find the node that you want to view or manage.	
	Node deployment	Overall feature	DataWorks V1.0 supports node deployment.	The feature name is changed to **Cross-Project Cloning**. You can clone a node only between workspaces in basic mode.	
Operation Center	Node list	Feature	You can search for a node by node type, node name, and owner in the node list.	More filters are added to help you search for nodes, including **Workflow**, **Solution**, and **Baseline**.	Nodes can be managed based on the business. More features are added for node development.
	Node management	Feature	You can search for a node by node type, node name, owner, data timestamp, and running date in the node list.	More filters are added to help you search for nodes, including **Workflow**, **Solution**, and **Baseline**.	
	Alerting	Feature	You can configure alerts based on the completion status and running status of nodes.	The **Baseline Management**, **Event Management**, and **Rule Management** features are added to build a more intelligent and comprehensive alerting system.	

Service	Feature	Item	DataWorks V1.0	DataWorks V2.0	Description
Monitor	The Monitor service is added as a node monitoring and analysis system of DataWorks. It monitors the running status of nodes and sends alerts based on the intervals, notification methods, and recipients specified in alert triggers. It can automatically select the most appropriate alerting time, notification methods, and recipients.				A one-stop cloud platform is built to provide data development, security, governance, and sharing services.
Data Quality	The Data Quality service is added for you to control the data quality of heterogeneous connections. In Data Quality, you can check data quality, configure alert notifications, and manage connections. Data Quality monitors data in datasets. Currently, it allows you to monitor MaxCompute tables and DataHub topics. When offline MaxCompute data changes, Data Quality checks data and blocks nodes if it detects exceptions. This prevents nodes from being affected. In addition, Data Quality allows you to manage the check result history so that you can analyze and evaluate the data quality.				
DataService Studio	The DataService Studio service is added. You can quickly create API operations in DataService Studio based on tables or register existing API operations with DataService Studio. DataService Studio and API Gateway are interconnected. This allows you to deploy API operations to API Gateway with ease. DataService Studio works together with API Gateway to provide a secure, stable, low-cost, and easy-to-use data sharing service.				
Data Security Guard	The Data Security Guard service is added to identify data assets, detect sensitive data, classify data, de-identify data, monitor data access behavior, report alerts, and audit risks.				

1.2. DataWorks V3.0

This topic describes the release notes and new features of DataWorks V3.0.

Release information

Version: DataWorks V3.0

- Date: December 18, 2019
- Region: all regions that support DataWorks
- Content: In addition to the MaxCompute engine supported by DataWorks V2.0, DataWorks V3.0 integrates many other computing engines such as E-MapReduce, Realtime Compute, Hologres, and Graph Compute to build a multi-engine architecture.

  You can add multiple types of engine instances to manage your workflows, nodes, and tables in a workspace. You can also configure dependencies between nodes in different regions to flexibly schedule nodes and facilitate cross-region collaboration.

Key features

- Multiple types of computing engines

DataWorks V3.0 supports plug-ins of multiple computing engines. In addition to the MaxCompute engine supported by DataWorks V2.0, DataWorks V3.0 integrates many other computing engines such as E-MapReduce, Realtime Compute, Hologres, and Graph Compute.

- MaxCompute: MaxCompute is a fast and fully-managed computing platform for large-scale data warehousing. It supports processing exabytes of data. MaxCompute is the first and maturest computing engine supported by DataWorks. Almost all of its features have been seamlessly integrated into DataWorks. For more information, see What is MaxCompute?

- E-MapReduce: E-MapReduce is a big data engine that runs on Alibaba Cloud Elastic Compute Service (ECS) based on open-source Apache Hadoop and Apache Spark. You can analyze and process your data by using peripheral systems such as Apache Hive in the Hadoop and Spark ecosystems.

  All services and features of DataWorks V3.0 support E-MapReduce, such as metadata management, Data Map, data lineage, DataStudio, scheduling, node management and monitoring, and Data Quality. Currently, only DataWorks V3.0 Professional Edition and higher support E-MapReduce. To use E-MapReduce in DataWorks, you must bind the target workspace ID to an E-MapReduce cluster and add the endpoint of the E-MapReduce cluster to the whitelist of DataWorks. For more information, see What is E-MapReduce?

- Realtime Compute: Built on Apache Flink, Realtime Compute is a one-stop, high-performance platform used to process big data in real time. All services and features of DataWorks V3.0 support Realtime Compute. DataWorks V3.0 provides Stream Studio for you to create real-time computing nodes by using drag-and-drop operations in directed acyclic graphs (DAGs). DAGs and Flink SQL statements can be converted into each other to facilitate the development of real-time computing nodes with intelligent management and diagnosis.

- Hologres: Hologres is a real-time interactive data analysis service that is fully compatible with PostgreSQL and seamlessly integrated with Alibaba Cloud big data services.

  Using Hologres, you can gain an analytical insight into thousands of billions of concurrent data records from multiple dimensions with low latency and explore for new business opportunities. You can also use your business intelligence (BI) tools with Hologres.

  DataWorks V3.0 provides a one-stop online analytical processing (OLAP) service Holo Studio to facilitate standard development management and help you build real-time data warehouses. This contributes to effective and simple development.

- Graph Compute: Graph Compute is a new-generation, one-stop platform used to manage and analyze graph data. It supports quick data loading, auto scaling, millisecond-level query latency, hybrid computing engines for online and offline graph computing, and shared data storage.

  DataWorks V3.0 provides Graph Studio based on Graph Compute, which supports one-stop development services, including instance modeling, data import, data query by using Gremlin, and visualized data analysis.

- Custom wrappers

  In addition to a variety of computing engines, DataWorks V3.0 Enterprise Edition also supports custom wrappers. For more information, see Overview of custom node types. You can use wrappers to access specified computing services or database query services.

  You can also use the intelligent SQL editor to configure custom nodes and schedule, orchestrate, manage, and monitor these custom nodes as required.

- Multiple engine instances in a workspace

In DataWorks V2.0, you can configure only one engine instance for a workspace. For example, if the computing engine is MaxCompute, you can create only one MaxCompute project for a workspace. DataWorks V3.0 Professional Edition and higher allow you to create or bind multiple engine instances for or to a workspace. You can manage the computing engines, computing nodes, and tables required by your business in a more flexible manner.

- Cross-region dependencies

  In DataWorks V2.0, you can only configure dependencies between nodes in the same region. DataWorks V3.0 Enterprise Edition and higher allow you to configure dependencies between nodes in different regions in Mainland China under an Alibaba Cloud account. You can use this feature to schedule nodes for your business across different regions.

- Resource group orchestration (available soon)

  DataWorks V3.0 will support resource group orchestration. With this feature, you can quickly configure and change the resource group for multiple nodes at a time. For example, you can change multiple nodes from the default shared resource group to an exclusive resource group.

- Workspace import and export (available soon)

  DataWorks V2.0 supports the data backup and recovery of workspaces. DataWorks V3.0 upgrades this feature and makes it more flexible. You can import or export nodes, table data definition language (DDL) statements, resources, functions, and connections to or from workspaces. This facilitates workspace migration and initialization.

2.Change history

This topic describes the change history of DataWorks documentation. You can learn the new features and feature changes of DataWorks.

> ⑦ **Note**   DataWorks can be automatically updated. The update has no impact on existing users.

Changes in November 2021

Date	Item	Type	Description	Topic
2021-11-24	Updated feature	Data Integration	A topic is updated to describe how to configure HDFS Reader and HDFS Writer by using the codeless user interface (UI).	HDFS Reader and HDFS Writer
2021-11-20	New feature	API operation	A topic is added to describe how to call the ListDags operation to obtain the details of directed acyclic graphs (DAGs) for a single data backfill instance based on OpSeq.	ListDags
2021-11-14	New feature	DataStudio	A topic is added to describe how to perform operations on multiple DataWorks objects at a time. DataWorks allows you to modify the configurations such as the owner of multiple nodes, resources, or functions at a time. After the modification, you can commit and deploy the nodes, resources, or functions to the production environment at a time to make the modification take effect.	Perform operations on multiple DataWorks objects at a time
2021-11-08	New feature	DataStudio	A topic is added to describe the resource group orchestration feature. This feature allows you to change resource groups for scheduling for multiple nodes in a workflow at the same time. If multiple resource groups for scheduling exist in your workspace, you can change resource groups for scheduling for nodes in the workspace based on your business requirements. This helps you improve resource usage.	Change the resource groups for scheduling for one or more nodes

Changes in October 2021

Date	Item	Type	Description	Topic
2021-10-26	New feature	Data Modeling	• A topic is added to describe the naming dictionary feature. A naming dictionary can be used to manage the roots and morphemes of business terms, physical tables, and fields and the standardized translation of the roots and morphemes. DataWorks Data Standard allows you to create naming dictionaries and export existing naming dictionaries. • A topic is added to describe the reverse modeling feature. If you use a modeling tool to generate models and you want to use DataWorks Dimensional Modeling for subsequent modeling operations, you can use the reverse modeling feature provided by DataWorks Dimensional Modeling. The reverse modeling feature allows you to import the models that are generated by using the modeling tool into a compute engine instance. The system creates models based on the imported models. This way, you do not need to manually create models in DataWorks Dimensional Modeling, and your time costs are reduced.	• Naming dictionary • Reverse modeling

Date	Item	Type	Description	Topic
2021-10-22	Update d feature	Data Securit y Guard	• A topic is updated to describe how to manage data sensitivity levels in a more efficient manner. DataWorks allows you to classify data based on the value, sensitivity level, impact, and distribution range of the data. The management policy and development requirements vary for data of different sensitivity levels. • A topic is updated to describe how to identify sensitive field types in an efficient manner. DataWorks allows you to identify sensitive data in your workspace by using the data identification rules configured for built-in and custom sensitive field types.	• Mange data sensitivity levels • Manage sensitive field types
2021-10-15	New feature	API operati on	• A topic is added to describe how to call the ListDeployments operation to query the information about deployment tasks. • A topic is added to describe how to call the UpdateIDEEventResult operation to query the check results of events that have specific extension points to DataWorks DataStudio when an extension point is triggered during data development. • A topic is added to describe how to call the GetIDEEventDetail operation to query the data snapshot of an extension point based on the ID of a DataWorks open message when an event that has the extension point is triggered.	• ListDeployments • UpdateIDEEventResult • GetIDEEventDetail
2021-10-14	New feature	API operati on	A topic is added to describe how to call API operations to create, configure, and manage a data synchronization node in Data Integration.	Use API operations to create, configure, and manage a data synchronization node

Date	Item	Type	Description	Topic
2021-10-11	New feature	DataSt udio	A topic is added to describe the code search feature. The code search feature allows you to query code snippets in nodes by keyword. The search results show the details of each code snippet and the nodes that contain the code snippets. You can use this feature to trace the node that causes changes in a table.	Code search

Changes in September 2021

Date	Item	Type	Description	Topic
2021-09-30	New feature	Schedu ling setting s in DataSt udio	A topic is added to describe the configurations of scheduling parameters. Scheduling parameters are used during the running of DataWorks nodes. The values of scheduling parameters are automatically replaced with specific values based on the data timestamps of the nodes and the value formats of the scheduling parameters. This implements dynamic parameter settings during the running of nodes.	Configure scheduling parameters
2021-09-30	New feature	Schedu ling setting s in DataSt udio	A topic is added to describe how to configure cross-cycle dependencies between nodes and the types of cross-cycle dependencies supported by DataWorks. If you configure cross-cycle dependencies for a node, the instances of this node can be run only if the last-cycle instances of the dependent node are run as expected.	Configure cross-cycle scheduling dependencies
2021-09-26	New feature	Data Map	Topics are added to describe the new features of the Data Map service. Data Map allows you to query APIs in all workspaces that are owned by the current tenant and view details of the APIs. This facilitates quick queries. On the details page of an API, you can view the basic information, parameters, and sample responses of the API.	Query an API View the details of an API

Date	Item	Type	Description	Topic
2021-09-15	New feature	DataAnalysis	A topic is added to describe how to use SQL statements to query and analyze data of the added data sources in DataAnalysis.	SQL queries
2021-09-02	New feature	Operation Center	A topic is added to describe the advanced mode of generating data backfill node instances for auto triggered nodes. The advanced mode is used to generate data backfill node instances for multiple nodes at a time. You can select nodes that may not have dependencies with each other. You can select nodes for which you want to generate data backfill node instances in the directed acyclic graph (DAG) or in the node list on the Cycle Task page.	Generate retroactive data and manage retroactive instances

Changes in August 2021

Date	Item	Type	Description	Topic
2021-08-29	New feature	Data Integration	A topic is added to describe how to use the data masking feature. This feature can mask sensitive data in a single table that is synchronized in real time and store the data in a specified database.	Configure data de-identification
2021-08-22	New feature	Data Integration	A topic is added to describe how to synchronize data to Kafka in real time by using the Data Integration service of DataWorks.	Plan and configure resources

Date	Item	Type	Description	Topic
2021-08-11	New feature	SSL-based authentication	Topics are updated to describe the SSL-based authentication for data sources. If you configure data sources such as MySQL, SQL Server, and PostgreSQL databases, you can enable SSL-based authentication for the data sources. After the configuration is complete, only trusted applications and services can access the data resources. In DataWorks, SSL-based authentication is provided as a third-party identity authentication mechanism. Third-party identity authentication mechanisms are used to perform strict identity authentication on users and services. These mechanisms prevent untrusted applications or services from accessing data and improve the stability of data access during data synchronization.	• Add a MySQL data source • Add an SQL Server data source • Add a PostgreSQL data source
2021-08-07	Updated feature	Permission management system	A topic is added to describe the permission management system of DataWorks. The permission management system of DataWorks consists of two parts: *permissions controlled by using RAM* and *permissions controlled by DataWorks*.	Overview of roles and permissions

Date	Item	Type	Description	Topic
2021-08-01	New feature	Migration Assistant	A topic is updated to describe the Migration Assistant service of DataWorks. Migration Assistant was officially commercialized on August 1, 2021. Migration Assistant allows you to migrate data objects across different DataWorks versions, Alibaba Cloud accounts, regions, and workspaces. You can export the data objects in your workspace, including auto triggered nodes, manually triggered nodes, resources, functions, data sources, table metadata, ad hoc queries, and script templates. You can also create full export tasks, incremental export tasks, or custom export tasks to export your data objects in DataWorks based on your business requirements.	Overview

Changes in July 2021

Date	Item	Type	Description	Topic
2021-07-22	New feature	API operation	A topic is added to describe how to call the CreateDISyncTask operation to create a batch synchronization node.	CreateDISyncTask
2021-07-14	New feature	Configurations in the console	You can log on to the DataWorks console and add a RAM user or a RAM role as an alert contact on the Alert Contacts page. If an error occurs during the running of a node, DataWorks sends alert notifications to the alert contact that you specified. This allows you to handle exceptions at the earliest opportunity.	Configure and view alert contacts
2021-07-09	New feature	Billing method	A topic is updated to describe the billing method for different editions that are used in the China East 2 Finance and China South 1 Finance regions.	DataWorks advanced editions

Date	Item	Type	Description	Topic
2021-07-03	New feature	Data Security Guard	A topic is added to describe how to use the data traceability feature provided by DataWorks. This feature allows you to extract the watermark information of the data in a leaked data file. This helps you trace the users who may cause data leaks.	Trace leak sources
2021-07-02	New feature	Data Security Guard	A topic is added to describe how to generate and manage sample libraries based on the sample files that you provide. You can associate a sample library with a data identification rule to identify data. If the data to be identified contains the data in the sample library, the data to be identified matches the data identification rule. You can use sample libraries to identify enumerated values, such as employee names and user addresses.	Create and manage sample libraries
2021-07-02	New feature	Data Security Guard	A topic is added to describe how to use sample fields to train models. DataWorks extracts the characteristics of these fields and generates a rule model. You can use this rule model to identify the data that has similar characteristics in your data assets.	Generate a custom data identification model

Changes in June 2021

Date	Item	Type	Description	Topic
2021-06-11	New feature	DataStudio	A topic is added to describe how to create and run an E-MapReduce (EMR) Streaming SQL node. EMR Streaming SQL nodes allow you to use SQL statements to develop streaming analytics jobs.	Create and use an EMR Streaming SQL node

Date	Item	Type	Description	Topic
2021-06-11	New feature	DataStudio	A topic is added to describe how to create and run an EMR Spark Streaming node. EMR Spark Streaming nodes are used to process real-time streaming data with high throughput. This type of node supports fault tolerance, which helps you restore data streams on which errors occur.	Create and use an EMR Spark Streaming node
2021-06-09	New feature	Operation Center	A topic is added to describe how to view the basic information and status of real-time computing nodes on the Stream Task page in the DataWorks console. This allows you to monitor the status of the nodes. In addition, you can configure alert rules for the nodes that you want to monitor. This way, you can identify and handle exceptions at the earliest opportunity.	Manage real-time computing nodes

Changes in May 2021

Date	Item	Type	Description	Topic
2021-05-20	New feature	Operation Center	A topic is added to describe how to create and manage a shift schedule in DataWorks. If you set the Recipient parameter to Varies According to Shift Schedule and select a shift schedule when you create a custom alert rule, DataWorks can send alert notifications to the on-duty engineers that you specify for the shift schedule. After the engineers receive the alert notifications, they can identify and handle exceptions at the earliest opportunity.	Create and manage a shift schedule
2021-05-17	New feature	DataStudio	A topic is added to describe how to create and run ClickHouse SQL nodes. A ClickHouse SQL node allows you to use a distributed SQL query engine to process structured data. This improves task efficiency.	Create and use a ClickHouse SQL node

Date	Item	Type	Description	Topic
2021-05-15	New feature	Data Integration	Topics are added to describe how to synchronize data to AnalyticDB for MySQL 3.0 by using the Data Integration service of DataWorks.	- Plan and configure resources - Configure data sources for data synchronization from PolarDB - Configure data sources for data synchronization from MySQL - Configure a data source (OceanBase) - Add data sources - Create and configure a synchronization solution - FAQ

Changes in April 2021

Date	Item	Type	Description	Topic
2021-04-29	New tutorial	Quick start	A topic is added to provide practical AI tutorials that help you learn how to develop nodes in DataWorks.	AI tutorials
2021-04-28	New feature	Data Integration	A topic is added to describe how to add or remove source tables after you run a node to synchronize data to Hologres.	Add or remove source tables to or from a synchronization solution that is running
2021-04-22	New feature	DataStudio	A topic is added to describe how to create an FTP Check node that can be used to periodically detect whether a specific file exists based on FTP. If the FTP Check node detects that the file exists, the scheduling system starts to run the descendant node of the FTP Check node. Otherwise, the FTP Check node retries the detection based on the configured detection interval. The FTP Check node stops the retry until the condition for stopping the detection is met. In most cases, FTP Check nodes are used for communications between the DataWorks scheduling system and external scheduling systems.	Create an FTP Check node

Date	Item	Type	Description	Topic
2021-04-06	API operati on	API operati on	A topic is added to describe how to call the GetPermissionApplyOrderDetail operation that is provided by Security Center.	GetPermissionApplyOrderDet ail
2021-04-05	New feature	Data Integra tion	A topic is added to describe how to synchronize data to Kafka in real time by using the Data Integration service of DataWorks.	Plan and configure resources

Changes in March 2021

Date	Item	Type	Description	Topic
2021-03-19	New feature	Custo m role	A topic is added to describe how to customize a role in a DataWorks workspace.	Manage members and roles
2021-03-11	New engine	Open source engine s from which tasks are export ed or migrat ed	Topics are updated to describe how to import tasks that are exported from the open source scheduling engines into DataWorks or migrate tasks from these scheduling engines to DataWorks.	• Export tasks from open source engines • Import tasks of open source engines
2021-03-11	New feature	Engine O&M	A topic is added to describe how to use the engine O&M feature of DataWorks to view the details of each EMR node and identify and remove the nodes that fail to run. This prevents failed nodes from affecting the performance of descendant nodes.	Use the engine O&M feature
2021-03-09	New feature	Aggreg ate analysi s of auto trigger ed nodes in a DAG	Topics are added to describe the aggregate view and the downstream and upstream analysis features of a DAG. You can view the details of an auto triggered node from the DAG and perform operations based on your business requirements.	• View auto triggered nodes • View auto triggered node instances • Generate retroactive data and manage retroactive instances

Date	Item	Type	Description	Topic
2021-03-03	New feature	API operati on	Topics are added to describe API operations that are added in the Operation Center, Data Security Guard, and Migration Assistant services of DataWorks.	• GetNodeParents • GetNodeChildren • GetOpSensitiveData • CreateImportMigration

Changes in February 2021

Date	Item	Type	Description	Topic
2021-02-24	New feature	Viewin g of the status inform ation about synchr onizati on nodes	A topic is added to describe how to view the distribution and execution details of synchronization nodes that are run and troubleshoot abnormal nodes. This way, the O&M efficiency of synchronization nodes can be improved.	View the status information about sync solutions
2021-02-09	New feature	Real- time synchr onizati on node	A topic is updated to describe how to create a real-time synchronization node to synchronize data from a specified table. After the real-time synchronization node is created, you can view the status of the node.	Synchronize data in a single table
2021-02-06	New feature	Real- time synchr onizati on node	Topics are added to describe how to create a real-time synchronization node to synchronize data from specified or all tables in a database to MaxCompute, Hologres, or DataHub. After the real-time synchronization node is created, you can view the status of the node.	• Synchronize data to MaxCompute in real time • Synchronize data to Hologres in real time • Synchronize data to DataHub in real time
2021-02-05	New feature	Data source type	A topic is added to describe how to add an ApsaraDB for OceanBase data source. You can use the ApsaraDB for OceanBase data source to create a synchronization node.	Add an ApsaraDB for OceanBase data source

Changes in January 2021

Date	Item	Type	Description	Topic
2021-01-28	New feature	New node types supported by DataStudio	Topics are added to describe how to create and run a MySQL node and an AnalyticDB for MySQL node. You can use SQL statements to develop data for a MySQL data source and an AnalyticDB for MySQL data source.	• Create and use a MySQL node • Create and use an AnalyticDB for MySQL node
2021-01-20	New feature	Synchronization solution	Topics are added to describe how to create a batch synchronization node and a real-time synchronization node to synchronize data from specified or all tables in a database to Elasticsearch. After the batch synchronization node and real-time synchronization node are created, you can view the status of the nodes.	• Configure and view a batch synchronization solution used to synchronize all data in a database • Configure and view a real-time sync solution used to synchronize all data in a database
2021-01-19	New feature	Whitelist configuration and category management in Data Map	A topic is added to describe how to configure whitelists and attach the AliyunDataWorksFullAccess policy to the specified RAM user. After the configuration, you can collect metadata and manage categories in Data Map.	Configure IP address whitelists for metadata collection
2021-01-13	New feature	Integration with ActionTrail	A topic is added to describe how to query DataWorks behavior events in ActionTrail. You can use the queried event details to perform behavior analysis, security analysis, resource change tracking, and compliance auditing.	Use ActionTrail to query behavior events
2021-01-13	New feature	Billing method for using the Datablau feature of DataWorks	A topic is added to describe the billing method for using the Datablau feature of DataWorks.	

Date	Item	Type	Description	Topic
2021-01-07	New feature	Data synchronization	A topic is added to describe how to synchronize data from a MySQL data source to Elasticsearch. You can learn how to prepare the required resource groups and data sources, create a synchronization node, and view the status of the synchronization node.	Synchronize data to Elasticsearch

Changes in December 2020

Date	Item	Type	Description	Topic
2020-12-24	New feature	Synchronization solution	Topics are added to describe how to synchronize data from the PolarDB, Oracle, or MySQL data source to Hologres or MaxCompute. You can learn how to prepare the required resource groups and data sources, create a synchronization node, and view the status of the synchronization node and the answers to the FAQ.	• Synchronize data to Hologres • Synchronize data to MaxCompute
2020-12-14	New feature	Metadata collection	A topic is added to describe how to create a crawler and collect metadata from a Tablestore data source to DataWorks. You can view collected metadata on the Data Map page.	Collect metadata from a Tablestore data source

Changes in November 2020

Date	Item	Type	Description	Topic
2020-11-18	New feature	API operation	A topic is added to describe how to call the CreateManualDag operation to trigger the running of a manually triggered workflow.	CreateManualDag
2020-11-18	New feature	API operation	A topic is added to describe how to call the GetManualDagInstances operation to query the information about instances in a manually triggered workflow.	GetManualDagInstances
2020-11-18	New feature	API operation	A topic is added to describe how to call the GetDag operation to query the details of the DAG based on the DAG ID.	GetDag

Date	Item	Type	Description	Topic
2020-11-18	New feature	API operation	A topic is added to describe how to call the SearchNodesByOutput operation to query a node based on the output.	SearchNodesByOutput
2020-11-10	New FAQ	User experience optimization	A topic is added to provide answers to frequently asked questions about Operation Center.	Overview
2020-11-02	New feature	Code review configuration	A topic is updated to describe the code review configuration in DataWorks. If you enable forcible code reviews, you must commit each node for the specified reviewer to review the code of the node. You can deploy the node only after the reviewer approves the code.	Code review

Changes in October 2020

Date	Item	Type	Description	Topic
2020-10-30	API operation overview	User experience optimization	A topic is added to describe the applicable scopes, billing methods, and call limits of API operations in DataWorks.	Overview
2020-10-28	New feature	Table type	A topic is added to describe how to create an EMR table.	Create an EMR table
2020-10-28	New feature	Association with EMR clusters	A topic is updated to describe how to associate an EMR cluster with a DataWorks workspace. In DataWorks, you can create nodes such as Hive, MapReduce, Presto, and Spark SQL nodes based on an EMR compute engine and configure EMR workflows. You can also schedule the nodes and manage metadata. This improves your data output.	Associate an EMR cluster with a DataWorks workspace

Changes in September 2020

Date	Item	Type	Description	Topic

Date	Item	Type	Description	Topic
2020-09-03	Updates of billing method	Pricing	A topic is updated to describe the pay-as-you-go billing method that allows you to use all the basic features of DataWorks in a cost-efficient manner.	Pay-as-you-go
2020-09-03	Updates	User experience optimization	A topic is updated to describe the basic information of DataWorks, features of DataWorks, and limits.	What is DataWorks?
2020-09-02	New tutorial	User experience optimization	A topic is added to describe how to use DataWorks together with Machine Learning Platform for AI (PAI) to automatically identify users who steal electricity. This ensures that users use electricity in a safe manner.	Overview

Changes in August 2020

Date	Item	Type	Description	Topic
2020-08-07	Added resource groups	User experience optimization	DataWorks provides custom resource groups for scheduling and custom resource groups for Data Integration to ensure that nodes are flexibly scheduled and data is synchronized as early as possible. A topic is added to describe how to create a custom resource group for scheduling and change the resource group for a node to the created custom resource group for scheduling.	Create custom resource groups for scheduling
2020-08-07	New feature	Data source type	A topic is added to describe how to configure a Hive data source. A Hive data source allows you to read data from and write data to Hive. You can configure synchronization nodes by using the codeless UI or code editor.	Add a Hive data source
2020-08-07	New feature	Data source type	A topic is updated to describe how to configure a GBase8a data source. A GBase8a data source allows you to read data from and write data to GBase8a by using GBase8a Reader and Writer. You can use the codeless UI or code editor to configure synchronization nodes for GBase8a.	Add a GBase8a data source

Date	Item	Type	Description	Topic
2020-08-07	New feature	Data source type	A topic is updated to describe how to configure a Hologres data source. A Hologres data source allows you to read data from and write data to Hologres by using Hologres Reader and Writer. You can use the codeless UI or code editor to configure synchronization nodes for Hologres.	Add a Hologres data source
2020-08-07	New feature	Data source type	A topic is updated to describe how to configure an HBase data source. An HBase data source allows you to read data from and write data to HBase by using HBase Reader and Writer. You can use the code editor to configure synchronization nodes for HBase.	Add an HBase data source
2020-08-07	New feature	Data source type	A topic is updated to describe how to configure an Elasticsearch data source. An Elasticsearch data source allows you to read data from and write data to Elasticsearch. You can use the code editor to configure synchronization nodes for Elasticsearch.	Add an Elasticsearch data source
2020-08-07	New FAQ	User experience optimization	A topic is added to describe how to troubleshoot issues related to connectivity, parameters, and permissions when you add data sources in DataWorks.	Troubleshooting for connections
2020-08-07	New feature	Node type	A topic is added to describe how to create an EMR Presto node. EMR Presto nodes allow you to perform interactive analysis and queries on large-scale structured and unstructured data.	Create an EMR Presto node
2020-08-05	New release notes of features	User experience optimization	A topic is added to describe the release notes of key features of DataWorks.	Release notes of key features

Changes in June 2020

Date	Item	Type	Description	Topic

Date	Item	Type	Description	Topic
2020-06-30	New FAQ	User experience optimization	Topics are added to provide answers to commonly asked questions about DataWorks services and features. Such services and features include Data Integration, DataStudio, custom resource groups, exclusive resource groups, dependencies, Alarm, and DataService Studio.	Overview
2020-06-28	New feature	Route management	A topic is added to describe how to add a route to a VPC or a data center.	Add a route
2020-06-28	New best practice	User experience optimization	A best practice is added to describe how to use exclusive resource groups for Data Integration to migrate data from a self-managed MySQL database hosted on Elastic Compute Service (ECS) to MaxCompute.	Migrate data from a user-created MySQL database on an ECS instance to MaxCompute
2020-06-28	New best practice	User experience optimization	A best practice is added to describe how to use AIRec. AIRec is developed based on cutting-edge big data, AI technologies, and years of experience of Alibaba in the e-commerce industry. It provides a personalized recommendation service for developers to increase the customer purchase rate and order conversion rate.	Intelligently recommend items on e-commerce websites
2020-06-28	New best practice	User experience optimization	A best practice is added to describe how to grant specified users access permissions on specified resources such as tables and user-defined functions (UDFs). This best practice involves data encryption and decryption algorithms and relates to data security.	Grant access to a specific UDF to a specified user
2020-06-28	New best practice	User experience optimization	A best practice is added to describe how to build a data warehouse for an enterprise based on AnalyticDB for MySQL and use the data warehouse for O&M and metadata management.	Build a data warehouse for an enterprise based on AnalyticDB for MySQL

Date	Item	Type	Description	Topic
2020-06-28	New best practice	User experience optimization	A best practice is added to describe how to use a PyODPS node in DataWorks to segment Chinese text based on Jieba, an open source segmentation tool. The topic also describes how to write the segmented words and phrases to a new table and use closure functions to segment Chinese text based on a custom dictionary.	Use a PyODPS node to segment Chinese text based on Jieba
2020-06-28	New best practice	User experience optimization	A best practice is added to describe how to use a PyODPS node that is running on an exclusive resource group to send emails.	Use a PyODPS node to send emails
2020-06-28	New best practice	User experience optimization	A best practice is added to describe how to connect DataV to DataWorks DataService Studio. You can create APIs in DataService Studio and call the APIs in DataV. DataV then presents analysis results of the MaxCompute data.	Connect DataV to DataWorks DataService Studio
2020-06-28	New best practice	User experience optimization	A best practice is added to describe how to use a PyODPS node in DataWorks to reference a third-party package. In this topic, a PyODPS 2 node is used as an example.	Use a PyODPS node to reference a third-party package
2020-06-28	New best practice	User experience optimization	A best practice is added to describe how to implement automatic synchronization of IoT data to the cloud. IoT is a network that carries data based on the Internet and traditional telecommunication networks. It enables data sources among all physical objects that are independently addressable.	Automatically synchronize IoT data to the cloud
2020-06-16	New tutorial	User experience optimization	Data quality is the basis for effective and accurate data analysis. A tutorial is added to describe the application scenarios and the standards for assessing the data quality.	Overview
2020-06-16	New tutorial	User experience optimization	A tutorial is added to help you understand and use the big data services of Alibaba Cloud to build an online operation analysis platform.	Business scenarios and development process

Date	Item	Type	Description	Topic
2020-06-15	New feature	Data source type	A topic is added to describe how to configure an ApsaraDB for OceanBase data source. An ApsaraDB for OceanBase data source allows you to read data from and write data to ApsaraDB for OceanBase. You can use the code editor to configure synchronization nodes.	Add an ApsaraDB for OceanBase data source
2020-06-15	New feature	Data source type	A topic is added to describe how to configure a Vertica data source. A Vertica data source allows you to read data from and write data to Vertica. You can use the code editor to configure synchronization nodes.	Add a Vertica data source
2020-06-15	New feature	Reader type	A topic is added to describe the data types that are supported by GBase8a Reader and the parameters that you can use to configure GBase8a Reader. For example, you can specify the data source and configure field mapping for GBase8a Reader. This topic also provides an example for configuring GBase8a Reader.	Gbase8a Reader
2020-06-15	New feature	Reader type	A topic is added to describe Hologres Reader. Hologres Reader allows you to export data from Hologres data warehouses. You can read data from Hologres tables and then write the data to other data sources based on the standard protocol of Data Integration.	Hologres Reader
2020-06-15	New feature	New feature	A topic is updated to describe Time Series Database (TSDB) Reader. TSDB Reader allows you to read data from TSDB. This topic describes the data types that are supported by TSDB Reader and the parameters that you can use to configure TSDB Reader. For example, you can specify the data source and configure field mapping for TSDB Reader. This topic also provides an example for configuring TSDB Reader.	TSDB Reader
2020-06-15	New feature	Writer type	A topic is added to describe Hologres Writer. Hologres Writer allows you to import data from multiple data sources to Hologres for real-time data analysis.	Hologres Writer

Date	Item	Type	Description	Topic
2020-06-15	New feature	Config uration	A topic is added to describe how to configure a resource group for node scheduling. You can select the required resource group for node scheduling in the Resource Group section.	Configure the resource group
2020-06-15	New descrip tion	User experie nce optimi zation	A topic is added to describe the logic of scheduling dependencies. Correct node dependencies are fundamental to effective workflows. This ensures that business data is generated in an effective and timely manner and implements standardized data development by using node dependencies.	Logic of same-cycle scheduling dependencies
2020-06-15	New feature	Resour ce group type	A topic is added to describe how to create and use exclusive resource groups for scheduling. DataWorks allows you to associate an exclusive resource group for scheduling with a VPC so that the resource group can connect to data sources in the VPC.	Create and use an exclusive resource group for scheduling

Changes in May 2020

Date	Item	Type	Description	Topic
2020-05-27	New usage notes	User experie nce optimi zation	A topic is added to describe the scenarios and methods of using shared resource groups, exclusive resource groups, and custom resource groups in DataWorks.	Overview
2020-05-27	New feature	Report templa te manag ement	A topic is added to describe how to manage report templates. You can create a template of data quality reports on the Report Template Management page. DataWorks Data Quality can periodically generate and send data quality reports based on the template.	Create and manage report templates
2020-05-27	New feature	Rule templa te manag ement	A topic is added to describe how to manage rule templates. DataWorks Data Quality allows you to manage a set of custom rules and create a rule template library to configure rules more efficiently.	Create, manage, and use rule templates

Date	Item	Type	Description	Topic
2020-05-27	New feature	Rule template management	A topic is added to describe the verification logic of Data Quality and the built-in rule templates that DataWorks provides for monitoring offline data.	Built-in rule templates for offline data

Changes in April 2020

Date	Item	Type	Description	Topic
2020-04-19	Service upgrade	DataWorks V3.0	Topics are updated to describe how to use Operation Center. In Operation Center, you can view the dashboard, manage auto triggered nodes and manually triggered nodes, and monitor nodes.	Operation Center
2020-04-18	Service upgrade	DataWorks V3.0	Topics are updated to describe the overall process of building a MaxCompute data warehouse.	Build and optimize a data warehouse
2020-04-18	Service upgrade	DataWorks V3.0	Topics are added to describe the Data Integration service of DataWorks. Data Integration is a stable, efficient, and scalable data synchronization service. It is designed to migrate and synchronize data between a wide range of heterogeneous data sources fast and stably in complex network environments.	Data Integration
2020-04-08	Service upgrade	DataWorks V3.0	A topic is updated to guide you through a complete process of data development and O&M.	Overview
2020-04-08	Service upgrade	DataWorks V3.0	Topics are updated to provide an overview of DataWorks, including the basic concepts, scenarios, and data development processes.	What is DataWorks?

Changes in March 2020

Date	Item	Type	Description	Topic
2020-03-26	New tutorial	User experience optimization	A tutorial is added to describe the complete operations in the DataWorks for EMR workshop.	Prepare the environment

Date	Item	Type	Description	Topic
2020-03-17	Service upgrade	DataWorks V3.0	A topic is updated to describe the upgraded data development mode. In the upgraded data development mode, you can group multiple workflows under a solution in a workspace. The previous hierarchical structure is abandoned.	Create a solution
2020-03-17	Service upgrade	DataWorks V3.0	Topics are updated to describe various types of nodes in DataWorks, such as batch synchronization nodes, MaxCompute nodes, EMR nodes, general nodes, and custom nodes.	Create a batch synchronization node
2020-03-17	Service upgrade	DataWorks V3.0	Topics are updated to describe how to perform configuration for DataStudio on the Setup page. For example, you can add and delete modules, and configure code templates, folders, and table levels.	Setup
2020-03-02	Service upgrade	DataWorks V3.0	Topics are updated to provide an overview of the DataWorks console. You can view the workspaces, resource groups, and compute engines in the DataWorks console.	Overview of the DataWorks console

Changes in February 2020

Date	Item	Type	Description	Topic
2020-02-29	New best practice	User experience optimization	A best practice is added to describe how to use the data synchronization feature of DataWorks to migrate data from Oracle to MaxCompute.	Best practice to migrate data from Oracle to MaxCompute
2020-02-28	New feature	Project management	A topic is added to describe how to create and manage projects on the Projects page in App Studio.	Projects
2020-02-28	New feature	Application management	A topic is added to describe the Apps page in App Studio. The Apps page displays applications that are created by you and shared by you, and third-party applications.	Apps
2020-02-28	New feature	Template management	A topic is added to describe the Templates page in App Studio. The Templates page displays all templates that are created based on projects.	Templates

Date	Item	Type	Description	Topic
2020-02-28	New feature	Application deployment	A topic is added to describe how to create an application in App Studio and deploy it in the production environment so that the application can be accessed over the Internet.	App deployment
2020-02-02	New feature	DataAnalysis	Topics are added to describe how to use the DataAnalysis service. DataAnalysis allows you to collaboratively edit and analyze workbooks, manage MaxCompute tables in tabular mode, and generate and share visual reports.	Overview

Changes in December 2019