ALIBABA CLOUD

Alibaba Cloud

DataWorks

Product Introduction

Document Version: 20220707

Alibaba Cloud

Legal disclaimer

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

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.

2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.

3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.

4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).

5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.

6. Please directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
⚠ Danger	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	⚠ **Danger:** Resetting will result in the loss of user configuration data.
🔔 **Warning**	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	🔔 **Warning:** Restarting will cause business interruption. About 10 minutes are required to restart an instance.
🔈 **Notice**	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	🔈 **Notice:** If the weight is set to 0, the server no longer receives new requests.
❓ **Note**	A note indicates supplemental instructions, best practices, tips, and other content.	❓ **Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

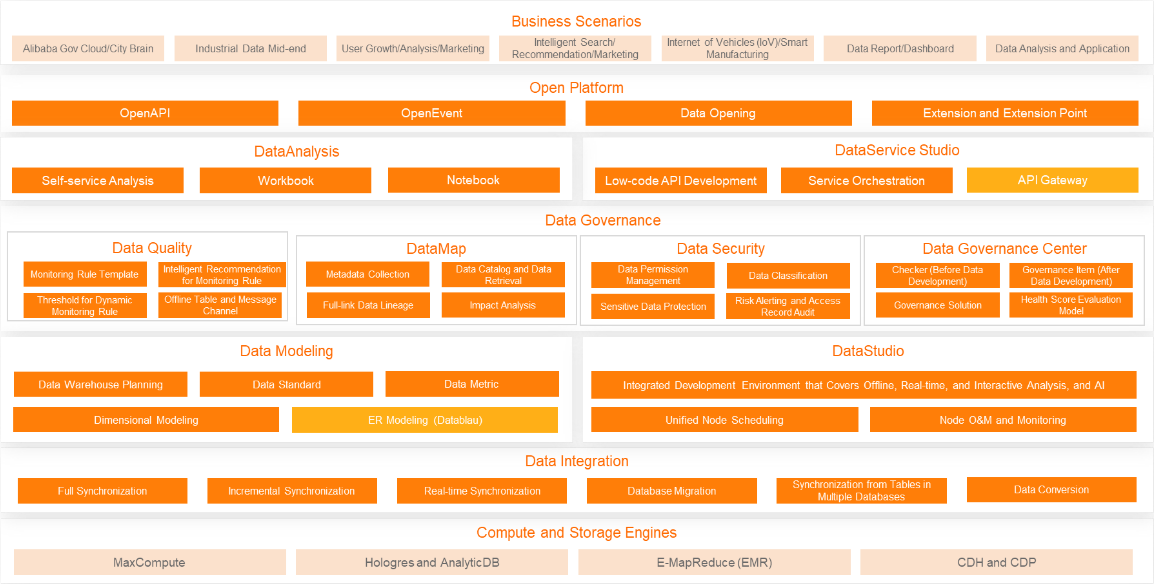
Table of Contents

1.What is DataWorks?

DataWorks is an end-to-end big data development and governance platform that provides data warehousing, data lake, and data lakehouse solutions based on big data compute engines, such as MaxCompute, Hologres, E-MapReduce (EMR), AnalyticDB, and CDH. Since 2009, DataWorks has been summarizing and improving the big data development methodology of Alibaba to support data mid-end building. DataWorks collaborates with public service sectors, state-owned enterprises, and customers in various industries, such as finance, retail, Internet, energy, and manufacturing, to improve data application efficiency and facilitate the digital transformation of industries.

Service architecture

DataWorks has developed and accumulated hundreds of core capabilities for more than ten years. DataWorks provides data modeling, data integration, data development, data governance, data security, and data analysis services. These services provide end-to-end data governance capabilities to help enterprises reduce data processing costs, increase data value, and improve data productivity. For more information about the data modeling, data integration, and data development services, see Overview, Overview, Overview.

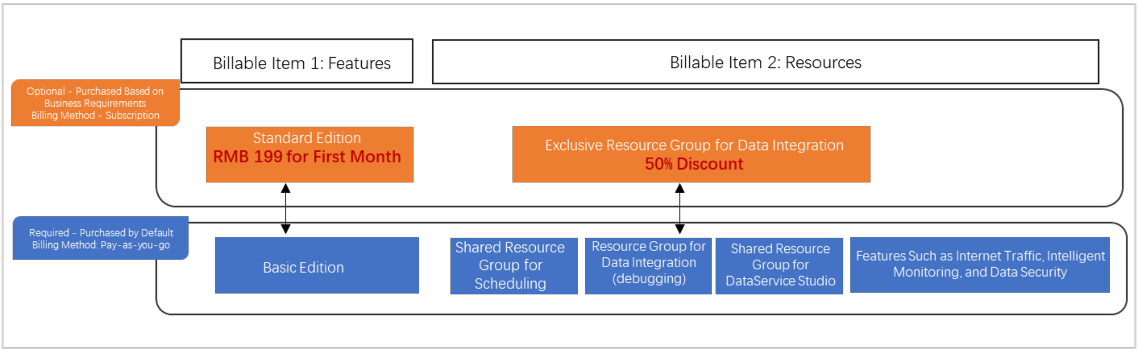

Business Scenarios
Alibaba Gov Cloud/City Brain
Industrial Data Mid-end
User Growth/Analysis/Marketing
Intelligent Search/ Recommendation/Marketing
Internet of Vehicles (IoV)/Smart Manufacturing
Data Report/Dashboard
Data Analysis and Application
Open Platform
OpenAPI
OpenEvent
Data Opening
Extension and Extension Point
DataAnalysis
Self-service Analysis
Workbook
Notebook
DataService Studio
Low-code API Development
Service Orchestration
API Gateway
Data Governance
Data Quality
Monitoring Rule Template
Intelligent Recommendation for Monitoring Rule
Threshold for Dynamic Monitoring Rule
Offline Table and Message Channel
DataMap
Metadata Collection
Data Catalog and Data Retrieval
Full-link Data Lineage
Impact Analysis
Data Security
Data Permission Management
Data Classification
Sensitive Data Protection
Risk Alerting and Access Record Audit
Data Governance Center
Checker (Before Data Development)
Governance Item (After Data Development)
Governance Solution
Health Score Evaluation Model
Data Modeling
Data Warehouse Planning
Data Standard
Data Metric
Dimensional Modeling
ER Modeling (Datablau)
DataStudio
Integrated Development Environment that Covers Offline, Real-time, and Interactive Analysis, and AI
Unified Node Scheduling
Node O&M and Monitoring
Data Integration
Full Synchronization
Incremental Synchronization
Real-time Synchronization
Database Migration
Synchronization from Tables in Multiple Databases
Data Conversion
Compute and Storage Engines
MaxCompute
Hologres and AnalyticDB
E-MapReduce (EMR)
CDH and CDP

Purchase guide

> ? **Note** DataWorks supports only Google Chrome 69 and later and the new Microsoft Edge that is based on Chromium.

The first time you use DataWorks, we recommend that you follow the instructions provided in this section to purchase features and resources. For more information, see Purchase guide.

- Recommended configurations


Billable Item 1: Features
Billable Item 2: Resources
Optional - Purchased Based on Business Requirements Billing Method - Subscription
Standard Edition
RMB 199 for First Month
Exclusive Resource Group for Data Integration
50% Discount
Required - Purchased by Default Billing Method: Pay-as-you-go
Basic Edition
Shared Resource Group for Scheduling
Resource Group for Data Integration (debugging)
Shared Resource Group for DataService Studio
Features Such as Internet Traffic, Intelligent Monitoring, and Data Security

- Recommendation reasons
  - Features: We recommend that you activate **DataWorks Professional Edition**. This edition provides features such as DataStudio, Operation Center, DataMap, and Data Quality and can meet your requirements for standard data warehouse building.
  - Resources: We recommend that you purchase **exclusive resource groups for Data Integration**. You can use this type of resource group in synchronization solutions such as batch synchronization, real-time synchronization, full synchronization, and incremental synchronization.

Customer use cases

- Big Data Center of State Grid Corporation of China (SGCC): DataWorks helped achieve centralized management of petabytes of data for SGCC and 27 subordinate provincial and municipal corporations. DataWorks also helped SGCC accelerate the digital transformation and upgrade of business by using the end-to-end governance and monitoring systems for data mid-ends.
- Mondelēz International (Fortune Global 500): Mondelēz China used DataWorks Data Modeling to perform end-to-end data model governance. This helped Mondelēz China significantly improve the self-service capability of data mid-ends, delegate data-related decision making, and unleash the digital power of the new retail industry.
- iDreamSky (a listed company): iDreamSky replaced the self-developed scheduling system with DataWorks based on open source EMR, which enabled technical personnel in the company to focus more on business and facilitated digital operations of the gaming industry.

For more information about customer use cases, see 客户案例.

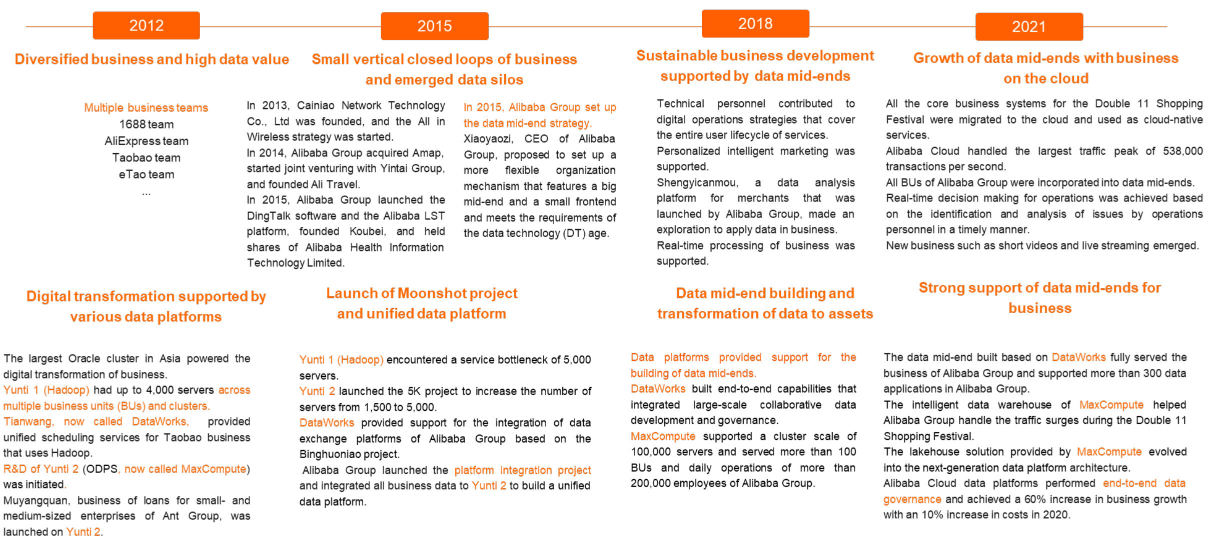
Development history

Development history of DataWorks within Alibaba Group

Since 2009, DataWorks has been used to build data mid-ends and data governance capabilities within Alibaba Group over multiple technology phases based on big data compute engines such as MaxCompute and Hologres. DataWorks has more than 50,000 daily active users within Alibaba Group. This indicates that one out of three employees in Alibaba Group use DataWorks on average. DataWorks supports over 300 data applications and serves more than 100 business units within Alibaba Group.

Four Development Stages of Data Platforms in Alibaba Group

Mutual Development of Business and Data Platform over Past 12 Years



Diversified business and high data value

Multiple business teams
1688 team
AliExpress team
Taobao team
eTao team
...

Small vertical closed loops of business and emerged data silos

In 2013, Cainiao Network Technology Co., Ltd was founded, and the All in Wireless strategy was started.
In 2014, Alibaba Group acquired Amap, started joint venturing with Yintai Group, and founded Ali Travel.
In 2015, Alibaba Group launched the DingTalk software and the Alibaba LST platform, founded Koubei, and held shares of Alibaba Health Information Technology Limited.

In 2015, Alibaba Group set up the data mid-end strategy.
Xiaoyaozi, CEO of Alibaba Group, proposed to set up a more flexible organization mechanism that features a big mid-end and a small frontend and meets the requirements of the data technology (DT) age.

Sustainable business development supported by data mid-ends

Technical personnel contributed to digital operations strategies that cover the entire user lifecycle of services.
Personalized intelligent marketing was supported.
Shengyicanmou, a data analysis platform for merchants that was launched by Alibaba Group, made an exploration to apply data in business.
Real-time processing of business was supported.

Growth of data mid-ends with business on the cloud

All the core business systems for the Double 11 Shopping Festival were migrated to the cloud and used as cloud-native services.
Alibaba Cloud handled the largest traffic peak of 538,000 transactions per second.
All BUs of Alibaba Group were incorporated into data mid-ends.
Real-time decision making for operations was achieved based on the identification and analysis of issues by operations personnel in a timely manner.
New business such as short videos and live streaming emerged.

Digital transformation supported by various data platforms

The largest Oracle cluster in Asia powered the digital transformation of business.
Yunti 1 (Hadoop) had up to 4,000 servers across multiple business units (BUs) and clusters.
Tianwang, now called DataWorks, provided unified scheduling services for Taobao business that uses Hadoop.
R&D of Yunti 2 (ODPS, now called MaxCompute) was initiated.
Muyangquan, business of loans for small- and medium-sized enterprises of Ant Group, was launched on Yunti 2.

Launch of Moonshot project and unified data platform

Yunti 1 (Hadoop) encountered a service bottleneck of 5,000 servers.
Yunti 2 launched the 5K project to increase the number of servers from 1,500 to 5,000.
DataWorks provided support for the integration of data exchange platforms of Alibaba Group based on the Binghuoniao project.
Alibaba Group launched the platform integration project and integrated all business data to Yunti 2 to build a unified data platform.

Data mid-end building and transformation of data to assets

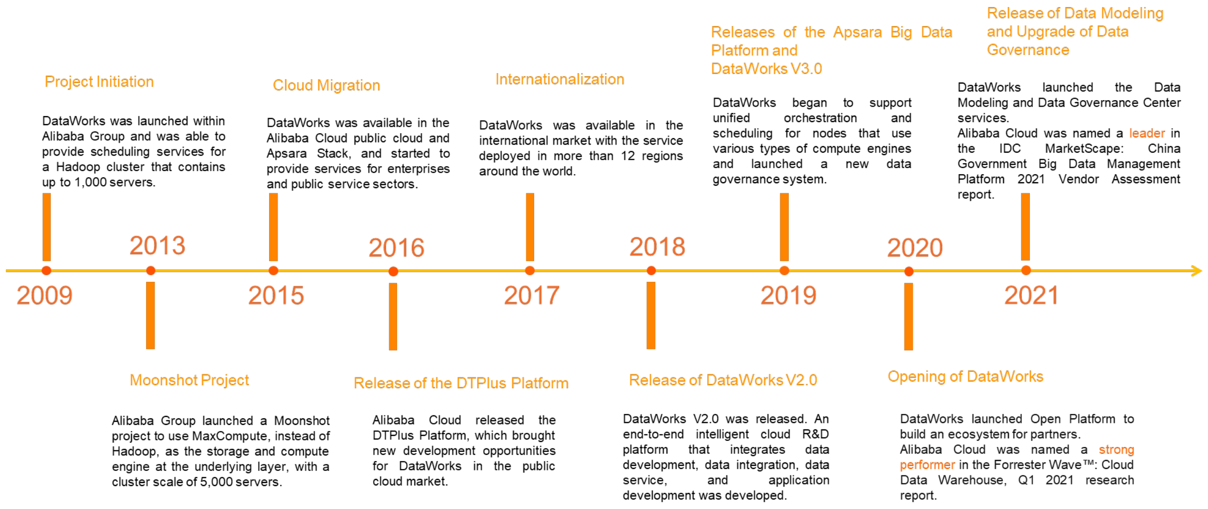
Data platforms provided support for the building of data mid-ends.
DataWorks built end-to-end capabilities that integrated large-scale collaborative data development and governance.
MaxCompute supported a cluster scale of 100,000 servers and served more than 100 BUs and daily operations of more than 200,000 employees of Alibaba Group.

Strong support of data mid-ends for business

The data mid-end built based on DataWorks fully served the business of Alibaba Group and supported more than 300 data applications in Alibaba Group.
The intelligent data warehouse of MaxCompute helped Alibaba Group handle the traffic surges during the Double 11 Shopping Festival.
The lakehouse solution provided by MaxCompute evolved into the next-generation data platform architecture.
Alibaba Cloud data platforms performed end-to-end data governance and achieved a 60% increase in business growth with an 10% increase in costs in 2020.

Development history of DataWorks on the cloud

DataWorks was migrated to the cloud in 2015. Since this year, DataWorks has launched services for Alibaba Cloud users based on the big data building methodology accumulated over the years. DataWorks constantly promotes the iteration of service capabilities of end-to-end data governance and is committed to improving data management and enhancing data value by collaborating with customers and partners from various industries and fields.



Project Initiation

DataWorks was launched within Alibaba Group and was able to provide scheduling services for a Hadoop cluster that contains up to 1,000 servers.

Moonshot Project

Alibaba Group launched a Moonshot project to use MaxCompute, instead of Hadoop, as the storage and compute engine at the underlying layer, with a cluster scale of 5,000 servers.

Cloud Migration

DataWorks was available in the Alibaba Cloud public cloud and Apsara Stack, and started to provide services for enterprises and public service sectors.

Release of the DTPlus Platform

Alibaba Cloud released the DTPlus Platform, which brought new development opportunities for DataWorks in the public cloud market.

Internationalization

DataWorks was available in the international market with the service deployed in more than 12 regions around the world.

Release of DataWorks V2.0

DataWorks V2.0 was released. An end-to-end intelligent cloud R&D platform that integrates data development, data integration, data service, and application development was developed.

Releases of the Apsara Big Data Platform and DataWorks V3.0

DataWorks began to support unified orchestration and scheduling for nodes that use various types of compute engines and launched a new data governance system.

Opening of DataWorks

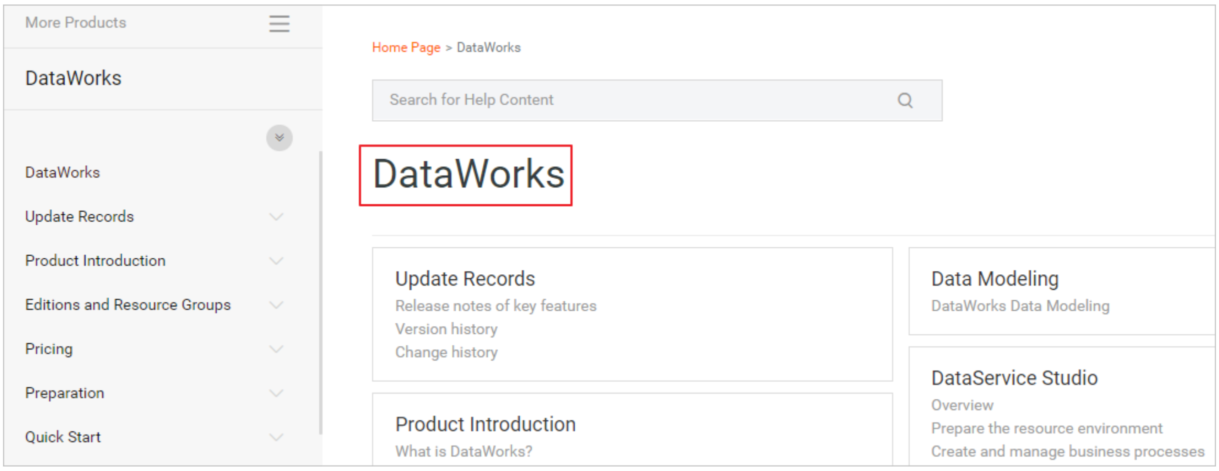
DataWorks launched Open Platform to build an ecosystem for partners.
Alibaba Cloud was named a strong performer in the Forrester Wave™: Cloud Data Warehouse, Q1 2021 research report.

Release of Data Modeling and Upgrade of Data Governance

DataWorks launched the Data Modeling and Data Governance Center services.
Alibaba Cloud was named a leader in the IDC MarketScape: China Government Big Data Management Platform 2021 Vendor Assessment report.

Learning path

You can quickly learn the concepts, basic operations, and advanced operations of DataWorks from the documentation homepage of DataWorks. For more information, see Documentation homepage.



More Products

DataWorks

- DataWorks
- Update Records
- Product Introduction
- Editions and Resource Groups
- Pricing
- Preparation
- Quick Start

Home Page > DataWorks

Search for Help Content

DataWorks

Update Records
Release notes of key features
Version history
Change history

Data Modeling
DataWorks Data Modeling

DataService Studio
Overview
Prepare the resource environment
Create and manage business processes

Product Introduction
What is DataWorks?

Support for DataWorks

You can visit Alibaba Cloud Support and Services to submit a ticket for after-sales services or join the DataWorks DingTalk group for pre-sales or after-sales services. If you join the DingTalk group, you can directly contact the DingTalk chatbot or contact on-duty technical personnel.

2.Functions and features

2.1. Data Integration: Integration of data from various data sources

Data Integration is a stable, efficient, and scalable data synchronization service. It can be used to migrate and synchronize data among a wide range of heterogeneous data sources that reside in complex network environments in a fast and stable manner.

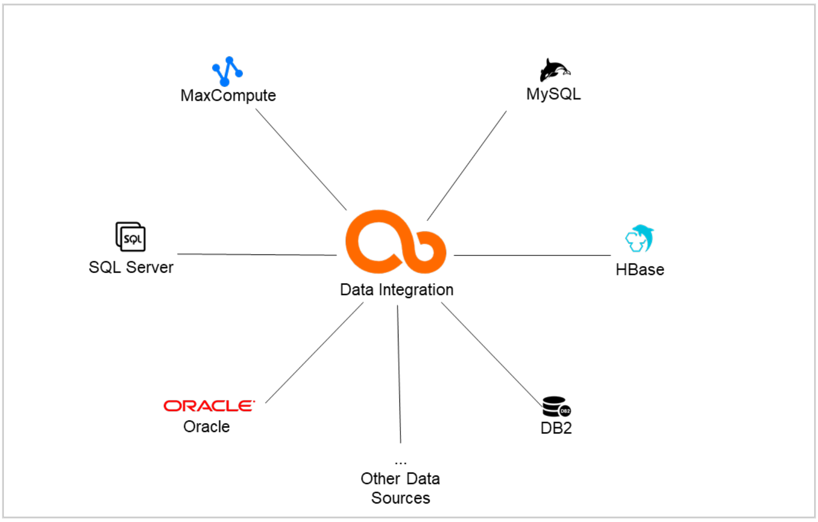
Overview

DataWorks Data Integration supports batch synchronization, real-time synchronization, and full or incremental synchronization that combines batch and real-time synchronization.

- You can configure a scheduling cycle for a batch synchronization node.
- Data synchronization among more than 50 types of heterogeneous data sources such as relational databases, data warehouses, file storage systems, and message queues is supported.
- Network connectivity solutions for data source connections in complex network environments are provided. You can use Data Integration to connect data sources that reside on the Internet, or in data centers or virtual private clouds (VPCs).
- Security control and O&M monitoring are supported to ensure that the data synchronization process is secure and controllable.

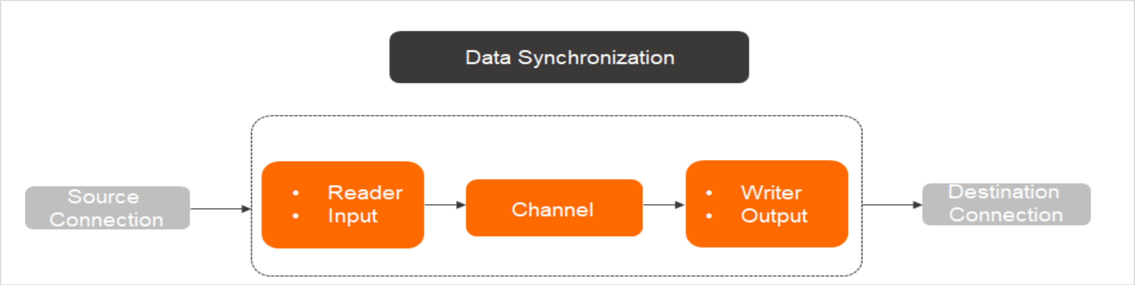
Core technology and architecture

- Engine architecture



A star-shaped engine architecture is provided. After a data source is added to Data Integration, the data source can be connected to another data source in Data Integration to form a data synchronization link. Then, data can be synchronized between the data sources. For more information about the supported data sources, see Supported data source types, readers, and writers and Plug-ins for data sources that support real-time synchronization.

- Resource groups for Data Integration and network connectivity



Before you synchronize data, you must connect the data sources to your resource group for Data Integration, as shown in the preceding figure. DataWorks allows you to use exclusive or custom resource groups for Data Integration to synchronize data. You can select a resource group type based on your business scenario. For more information about network connectivity solutions that can be used, see Select a network connectivity solution.

Use scenarios

DataWorks Data Integration is suitable for data transmission scenarios such as data ingestion into data warehouses or data lakes, sharding, real-time data archiving, and data forwarding between clouds.

Billing

You may be charged the following fees for running data synchronization nodes in Data Integration:

- Fees for using **resource groups for Data Integration**

  For more information about the billing details of resource groups for Data Integration, see Billing of exclusive resource groups for Data Integration (subscription) and Billing of the shared resource group for Data Integration (pay-as-you-go).

- Fees for using **resource groups for scheduling**

  For more information about the billing details of resource groups for scheduling, see Billing of exclusive resource groups for scheduling (subscription) and Shared resource group for scheduling.

- (Optional) Fees for **Internet traffic** generated during data synchronization

  If you run a data synchronization node to synchronize data over the Internet, Internet traffic is generated and you are charged for the generated Internet traffic. For more information about the billing details of Internet traffic, see Billing of Internet traffic.

> **Note** When you run data synchronization nodes in Data Integration, fees for node configurations may be generated. The fees are not charged by DataWorks, and the bills for the fees are not generated in DataWorks. For example, when you run a data synchronization node, fees for using data sources, computing and storage features of the related compute engine instance, and network services in the node may be generated. These fees are not charged by DataWorks. Network service fees include fees for using Express Connect, EIP Bandwidth Plan, and Elastic IP Address (EIP).

Activate DataWorks

After you activate DataWorks of a specific edition, you can purchase a resource group for Data Integration based on your business requirements and select an appropriate network connectivity solution to develop a data synchronization node in Data Integration. For more information about how to use Data Integration, see Overview.

2.2. DataStudio and Operation Center: Data processing

The **DataStudio** and **Operation Center** services of DataWorks allow you to develop data, create workflows, and perform intelligent O&M on workflows in an efficient and standard manner.

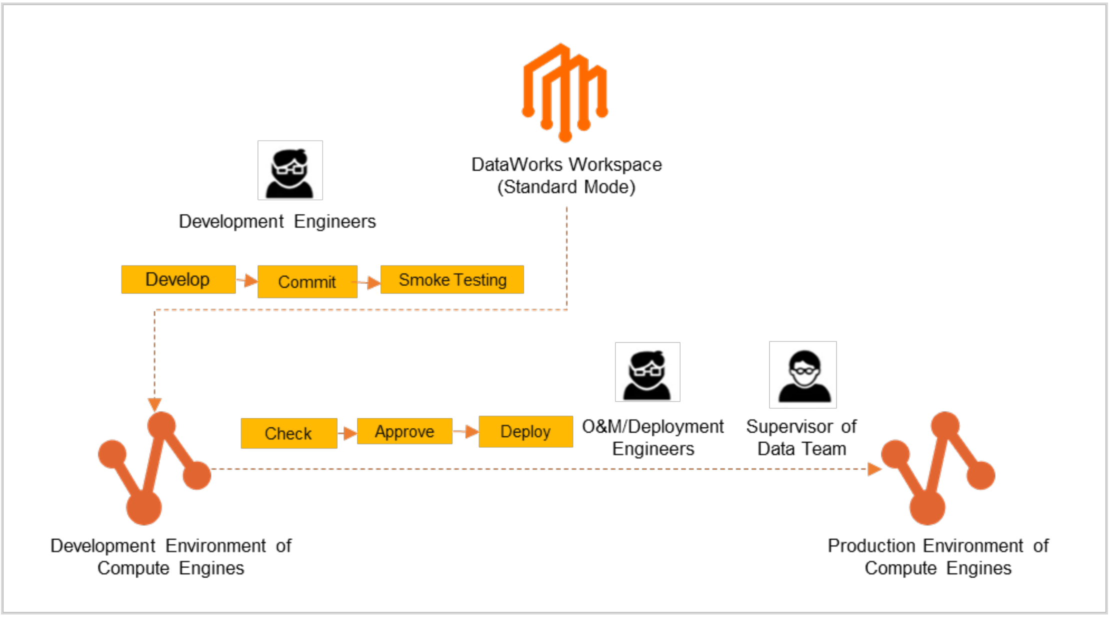
Overview

DataStudio and Operation Center support the following features:

- DataStudio supports various compute engines, such as MaxCompute, E-MapReduce (EMR), Cloudera's Distribution including Apache Hadoop (CDH), Hologres, AnalyticDB, and ClickHouse. You can create, test, deploy, and perform O&M operations on nodes of the preceding compute engines in DataStudio.
- DataWorks provides an intelligent code editor and the scheduling capability and allows you to configure scheduling dependencies in a visualized manner. The scheduling capability is verified by tens of millions of scheduling nodes and complex business dependencies in Alibaba Group.
- DataStudio isolates the development environment from the production environment and provides features such as code version management, code review, smoke testing, and release control. In addition, DataStudio works together with ActionTrail. This way, enterprises can develop data in a standard manner.
- Operation Center supports features such as data timeliness assurance, intelligent diagnosis, impact analysis, automatic O&M, and mobile O&M.

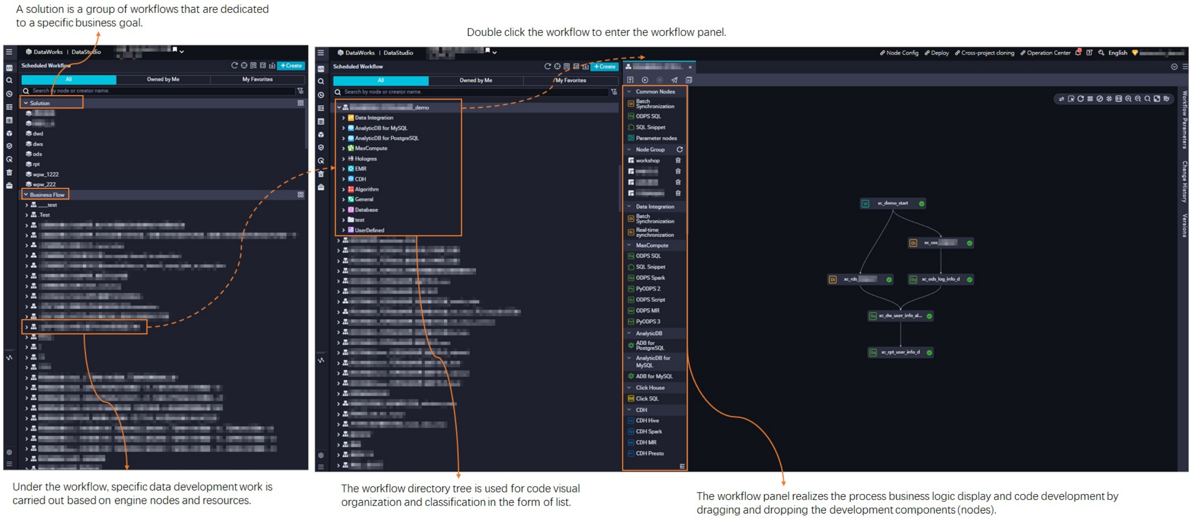
Core technology and architecture

- Efficient and standard development process



Note DataWorks workspaces in standard mode isolate the development environment from the production environment. For more information, see 简单模式和标准模式的区别.

- Visualized development user interface (UI)


A solution is a group of workflows that are dedicated to a specific business goal.

Double click the workflow to enter the workflow panel.

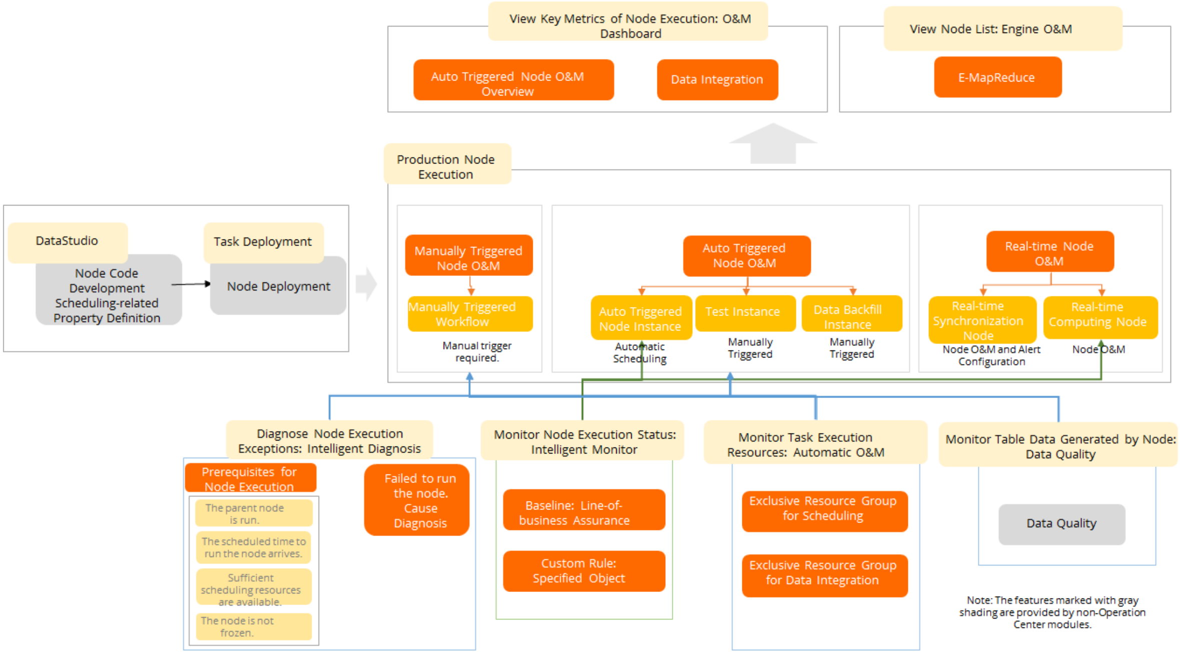
Under the workflow, specific data development work is carried out based on engine nodes and resources.

The workflow directory tree is used for code visual organization and classification in the form of list.

The workflow panel realizes the process business logic display and code development by dragging and dropping the development components (nodes).

You can create workflows by performing simple drag-and-drop operations. You can also configure scheduling parameters and develop data in the UI.

- Node monitoring and troubleshooting


View Key Metrics of Node Execution: O&M Dashboard
Auto Triggered Node O&M Overview
Data Integration
View Node List: Engine O&M
E-MapReduce
Production Node Execution
DataStudio
Node Code Development Scheduling-related Property Definition
Task Deployment
Node Deployment
Manually Triggered Node O&M
Manually Triggered Workflow
Manual trigger required.
Auto Triggered Node O&M
Auto Triggered Node Instance
Automatic Scheduling
Test Instance
Manually Triggered
Data Backfill Instance
Manually Triggered
Real-time Node O&M
Real-time Synchronization Node
Node O&M and Alert Configuration
Real-time Computing Node
Node O&M
Diagnose Node Execution Exceptions: Intelligent Diagnosis
Prerequisites for Node Execution
The parent node is run.
The scheduled time to run the node arrives.
Sufficient scheduling resources are available.
The node is not frozen.
Failed to run the node. Cause Diagnosis
Monitor Node Execution Status: Intelligent Monitor
Baseline: Line-of-business Assurance
Custom Rule: Specified Object
Monitor Task Execution Resources: Automatic O&M
Exclusive Resource Group for Scheduling
Exclusive Resource Group for Data Integration
Monitor Table Data Generated by Node: Data Quality
Data Quality

Note: The features marked with gray shading are provided by non-Operation Center modules.

Billing

When you use DataStudio and Operation Center, you may be charged for the following items:

- **Resource groups for scheduling** that are used to run nodes

  For more information, see Billing of exclusive resource groups for scheduling (subscription) and Shared resource group for scheduling.

- Baseline instances that are used to scan alert rules.

  For more information, see Baseline instances.

> Note When you develop data in DataStudio, you may be charged when you use the computing and storage features of the related compute engine instance. The fees are not charged by DataWorks, and the bills for the fees are not generated in DataWorks.

Activate DataWorks and use DataStudio and Operation Center

After you activate DataWorks, you can purchase resource groups for scheduling and associate a compute engine instance with your workspace based on your business requirements. Then, you can start to develop data. For more information, see DataStudio Overview and Operation Center Overview.

2.3. Data Modeling: Intelligent data modeling

Data Modeling is an intelligent data modeling service that is developed by Alibaba Cloud DataWorks. Data Modeling has accumulated the best practices of the methodology for data warehouse modeling of Alibaba Group over the past ten years. Data Modeling provides the Data Warehouse Planning, Data Standard, Dimensional Modeling, and Data Metric modules. Data Modeling helps enterprises strengthen modeling capabilities including reverse modeling capabilities in the process of data mid-end and data mart building and quickly build data assets.

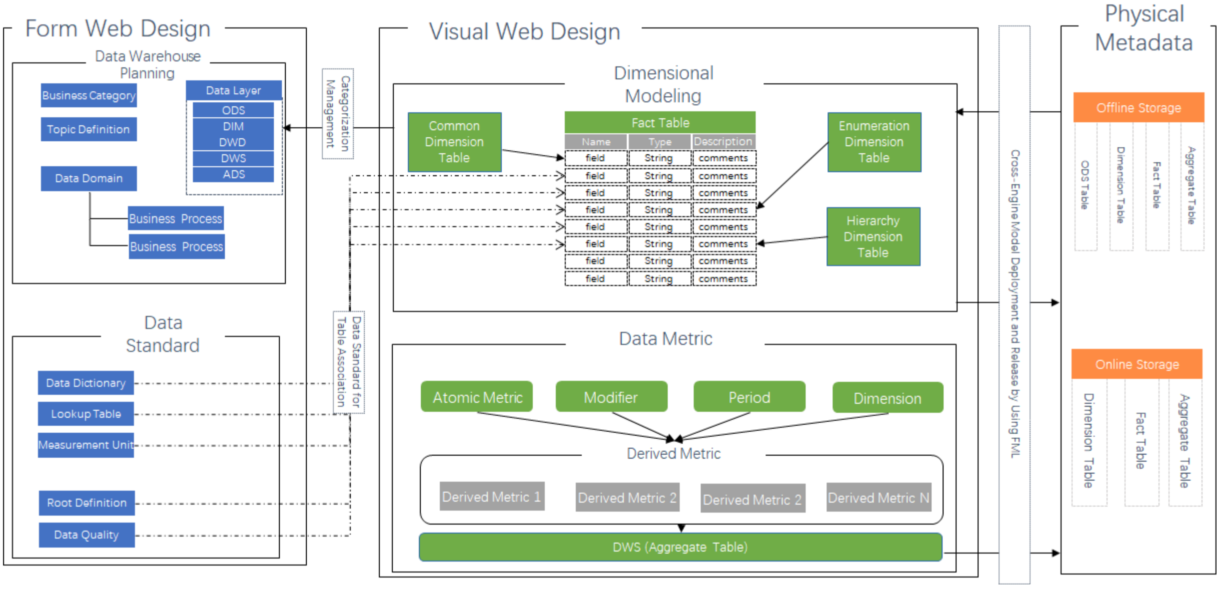
Overview

Data Modeling consists of the following modules: **Data Warehouse Planning**, **Data Standard**, **Dimensional Modeling**, and **Data Metric**.

- **Data Warehouse Planning**: allows you to plan data layers, data domains, and data marts, and configure model design workspaces. Different units can share the same data standards and the same data model.
- **Data Standard**: allows you to define data standards, lookup tables, measurement units, and naming dictionaries. This module also allows the system to generate quality rules based on lookup tables. Checks that are based on the generated quality rules are simple.
- **Dimensional Modeling**: supports reverse modeling, which helps resolve the issue of the cold start of modeling based on existing data warehouses. This module also supports visualized dimensional modeling based on data warehouses and allows you to import data by using Excel files and quickly build data models by using FML statements, a type of domain-specific language (DSL) similar to SQL statements. You can seamlessly integrate this module with DataStudio to enable the system to generate extract, transform, and load (ETL) code.
- **Data Metric**: allows you to create atomic metrics and derived metrics. You can create a single derived metric or multiple derived metrics at a time based on the same atomic metric and different periods and modifiers. This module is seamlessly integrated with Dimensional Modeling.

Core technology and architecture

The following figure shows the architecture of each module of Data Modeling.



Use scenarios

Data Modeling helps enterprises build modeling capabilities and mine the value of data assets. You can use Data Modeling in the following scenarios:

- **Standardize management of massive data**

  Larger enterprises have more complex data structures. How to manage and store data in a structured and orderly manner is a challenge that every large enterprise faces.

- **Break information barriers by interconnecting business data**

  If the data of each business or department in an enterprise is isolated from one another, the decision-makers cannot clearly or fully understand the data. How to break data silos between departments or business domains is a requirement for business data management.

- **Integrate data standards to achieve unified and flexible data interconnection**

  Inconsistent descriptions of the same data result in duplicate data, incorrect calculation results, and difficulties in business data management. How to formulate a unified data standard without changing the original system architecture and realize flexible interconnection between upstream and downstream business is one of the core focuses of standardized management.

- **Maximize data value to maximize profit**

  Make the most of various types of enterprise data to maximize the data value to deliver a more efficient data service for enterprises.

Activate DataWorks and use Data Modeling

After you activate DataWorks of an advanced edition, you can activate Data Modeling. Then, you can use the **Data Warehouse Planning**, **Data Standard**, **Dimensional Modeling**, and **Data Metric** modules provided by Data Modeling. For more information, see Overview.

2.4. DataAnalysis

DataAnalysis provides easy-to-use tools for non-professional data developers such as product staff, operations staff, and data analysts to obtain and use data. These tools help improve the efficiency of data analysis.

Overview

DataAnalysis allows you to upload data, use public datasets, query tables, add tables to favorites, execute SQL statements to obtain data, share SQL files, download SQL query results, and view workbooks.

Scenarios

DataAnalysis is suitable for non-professional data developers such as product staff, operations staff, and data analysts because of the following benefits:

- High capacity: DataAnalysis uses compute engines to analyze large amounts of data in an efficient manner.
- Data sharing: DataAnalysis can analyze data obtained from the databases of different business systems. DataAnalysis allows you to export data to MaxCompute tables. DataAnalysis also allows you to share result data with specific members and grant the members the permissions to access the data. This way, data can be shared among different systems and different users.
- High security: Operations such as SQL queries and downloads of SQL query results can be audited.

Billing rules

The DataAnalysis service of DataWorks is free of charge. You can use DataAnalysis after you activate DataWorks. The features provided by DataAnalysis vary based on the edition of DataWorks. For more information, see Differences among DataWorks editions.

Activate DataWorks and use DataAnalysis

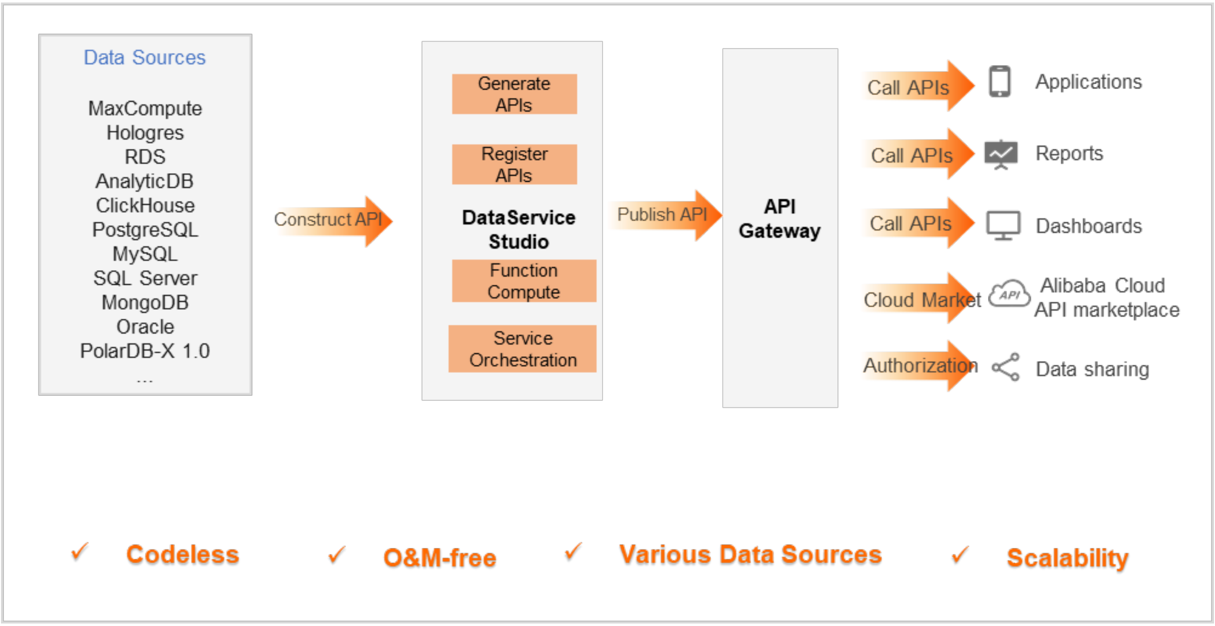
After you activate DataWorks, you can log on to the DataWorks console to use DataAnalysis. For more information, see DataAnalysis Overview.

2.5. DataService Studio: Fast publishing of APIs at low costs

DataService Studio is a flexible, secure, stable, and cost-effective data platform that allows you to create and publish APIs. DataService Studio provides enterprises with comprehensive data sharing capabilities. You can share data and explore the value of data in terms of publishing, approval, and authorization of APIs, statistics on API calls, and resource isolation.

Overview

DataService Studio serves as a bridge between data warehouses and upper-layer applications. DataService Studio builds a service bus to help enterprises create and manage private and public APIs in a centralized manner. DataService Studio also provides a solution to the last mile issue among data warehouses, databases, and data applications, and facilitates data sharing.



- DataService Studio allows you to create APIs based on tables in various data sources without the need to write code. You can also create APIs by specifying custom SQL statements. DataService Studio allows you to use functions to process the request parameters and returned results of APIs.
- DataService Studio is built based on a serverless architecture. You can publish APIs to API Gateway by performing simple operations. During the publishing, you do not need to focus on the infrastructure such as the runtime environment.

Core technology and architecture

DataService Studio is built based on a serverless architecture. You need to focus only on the query logic of APIs instead of the infrastructure such as the runtime environment. DataService Studio prepares the computing resources for you, supports elastic scaling, and requires zero O&M cost.

Activate and use DataService Studio

After you activate DataWorks, you can log on to the DataWorks console to use DataService Studio. For more information, see DataService Studio overview.

2.6. Open Platform: Comprehensive capability openness

DataWorks Open Platform provides the OpenAPI, OpenEvent, and Extensions modules. You can use the modules to integrate DataWorks with your applications and subscribe to event messages. These modules facilitate process management of data processing, data governance, and data O&M, and allow you to identify important changes in DataWorks and respond to the changes at the earliest opportunity.

Modules

DataWorks Open Platform provides the **OpenAPI**, **OpenEvent**, and **Extensions** modules.

- **OpenAPI**

  The OpenAPI module allows you to integrate DataWorks with your applications. For example, you can use this module to create, deploy, and perform O&M operations on multiple nodes at the same time, which helps improve big data processing efficiency and reduce costs of manual operations.

  For more information about the OpenAPI module, see OpenAPI.

- **OpenEvent**

  The OpenEvent module allows you to subscribe to system events in DataWorks to identify and respond to event changes in real time. For example, you can subscribe to table change events so that you can be notified of changes to core tables in real time. You can also subscribe to node change events to customize a dashboard to display the status of real-time synchronization nodes.

  For more information about the OpenEvent module, see OpenEvent.

- **Extensions**

  The Extensions module is a service-level plug-in that combines the capabilities of the OpenAPI and OpenEvent modules. You can use the Extensions module to create a custom data processing process. For example, you can create a custom node deployment management plug-in to intercept nodes that do not meet the specifications and requirements of deployment.

  For more information about the Extensions module, see Extensions.

Scenarios

DataWorks Open Platform provides comprehensive openness capabilities that allow you to integrate DataWorks with your applications, enable automated operations, define processes, and monitor business data. Users and partners are welcomed to use DataWorks Open Platform to develop industrial and scenario-specific data applications and plug-ins.

Billing

- Only users of DataWorks Enterprise Edition and DataWorks Ultimate Edition can use the OpenAPI module. The number of times that you can call API operations in this module free of charge varies based on the DataWorks edition that you use. After the free quota of API operation calls is exhausted, you are charged for the number of API operation calls. For more information about the billing details, see Billing of DataWorks API operations.
- The OpenEvent and Extensions modules are in invitational preview. Only users of DataWorks Enterprise Edition can join the invitational preview. After you apply for the invitational preview, you are not charged additional fees during the invitational preview.

Activate and use the modules

After you purchase DataWorks Enterprise Edition or DataWorks Ultimate Edition, you can use the OpenAPI module. For more information, see OpenAPI overview. If you want to use the OpenEvent and Extensions modules, submit a ticket to apply for the invitational preview and trial use. For more information, see OpenEvent overview and Extensions overview.

2.7. Migration Assistant: Migration to or within DataWorks

The Migration Assistant service of DataWorks allows you to migrate the tasks of open source scheduling engines to DataWorks. Migration Assistant also allows you to migrate data objects within DataWorks across clouds, regions, or accounts. This way, you can quickly clone and deploy tasks in DataWorks. To quickly migrate data and tasks to the cloud, you can also obtain help from the DataWorks team and the big data service team of Alibaba Cloud.

Overview

Migration Assistant supports the following operations:

- Migrate the tasks of open source scheduling engines to DataWorks.
- Migrate data objects within DataWorks.

Scenarios

- Migration of tasks to the cloud: You can use Migration Assistant to migrate the tasks of open source scheduling engines to DataWorks.
- Back up node code: You can use Migration Assistant to periodically back up your node code to prevent data from being accidentally deleted.
- Replicate a common workflow: You can use Migration Assistant to replicate a common workflow by performing import and export operations.
- Build a test environment: You can use Migration Assistant to copy all node code and replace production data with test data to build a test environment.
- Develop data across clouds: You can use Migration Assistant to migrate node code between DataWorks of the Alibaba Cloud public cloud and DataWorks of Alibaba Cloud Apsara Stack to collaboratively develop data across clouds.

Billing

After you activate DataWorks, you can use Migration Assistant. The features that are provided by Migration Assistant vary based on the DataWorks edition. For more information, see Differences among DataWorks editions.

Activate DataWorks and use Migration Assistant

After you activate DataWorks, you can log on to the DataWorks console and use Migration Assistant. For more information, see Migration Assistant Overview.

3.Benefits

DataWorks provides powerful basic capabilities that help improve work efficiency, ensure the timely generation of data, facilitate data governance, and allow you to construct data services at minimum costs.

Low learning costs

Common users other than technical personnel can have a good command of data development and governance procedures within 1 to 2 hours and no longer need to use traditional command line tools to perform development operations. This greatly reduces learning costs.

DataWorks allows you to organize nodes that are run by using various heterogeneous compute engines and to configure dependencies between the nodes in the same directed acyclic graph (DAG). This way, you do not need to separately maintain different technology stacks, and node organization efficiency is improved. The nodes include data synchronization nodes, SQL nodes, MR nodes, ODPS Spark nodes, real-time computing nodes, and Machine Learning Platform for AI (PAI) nodes.

Reduced labor costs

You can activate the DataWorks service by performing only simple configurations. After the service is activated, you can use the out-of-the-box features provided by this service to build data warehouses. This frees you from heavy development, deployment, and maintenance work and significantly reduces O&M costs.

Comprehensive features

DataWorks provides comprehensive features that can be used in data transmission, data development, data production, data governance, and data security scenarios. The whole lifecycle of big data in each scenario is covered by the related features. This helps address issues encountered by enterprises in data warehouse building, data mid-end building, and digital transformation.

- Data synchronization from data sources that reside in complex network environments, and real-time and batch synchronization of full and incremental data are supported.
- Scheduling for tens of millions of nodes for a single user is supported. Data processing is more fluent.

4.Common scenarios

4.1. Audiences and core capabilities

This topic describes the service positioning, audiences, and core capabilities of DataWorks.

Service positioning

DataWorks provides an end-to-end, standard, visualized, transparent, and intelligent cloud R&D platform that covers the full lifecycle of big data and provides individual development capabilities and full-stack data R&D capabilities for data developers, data analysts, and data asset managers. DataWorks allows you to use a single platform to perform operations in complex scenarios in which data transmission, data computing, data governance, and data sharing are required.

DataWorks is committed to developing features that meet the requirements of enterprises for building data warehouses and data mid-ends. DataWorks also provides support for the digital transformation of enterprise business.

Audiences

- Technical personnel such as data developers and algorithm developers
- Business personnel such as sales and operations personnel and business intelligence analysts
- Administrators who are engaged in data security and data compliance
- Data application developers
- Managers who manage the core data assets of enterprises

Key features

DataWorks provides the following key features:

- **Data integration**: supports data transmission and data migration to the cloud between various data sources that reside in complex network environments.
- **Data development**: provides a data development mode in which the development environment and production environment are isolated. This feature allows you to develop nodes that use different compute engines and to configure complex scheduling dependencies for the nodes. The nodes include batch processing nodes, stream processing nodes, and machine learning nodes.
- **Data analysis (available in DataWorks only on the Alibaba Cloud public cloud)**: allows you to perform quick and flexible ad hoc queries based on workbooks.
- **Data service**: allows you to quickly generate serverless APIs without the need to use code.
- **Data quality**: allows you to configure table-level or field-level monitoring rules to monitor data quality and helps you identify dirty data at the earliest opportunity.
- **Monitoring and alerting**: allows you to easily configure monitoring and alerting settings for complex workflows.
- **Data map (available in DataWorks only on the Alibaba Cloud public cloud)** or **Data management (available only in Apsara Stack DataWorks)**: provides powerful capabilities such as data search, data categorization, and data lineage.
- **Data asset management (available only in Apsara Stack DataWorks)**: allows you to manage data assets such as data tables and APIs in DataWorks in a centralized manner.
- **Data security**: provides capabilities such as data audit, data masking, and permission control.

- **Application development (available in DataWorks only on the Alibaba Cloud public cloud):** allows you to easily build data applications by dragging components on the DataStudio page.
- **Workspace management (available in DataWorks only on the Alibaba Cloud public cloud)** or **Platform management (available only in Apsara Stack DataWorks):** provides capabilities of managing the permissions of DataWorks users or members and the configurations of underlying compute engines for administrators at the system level.

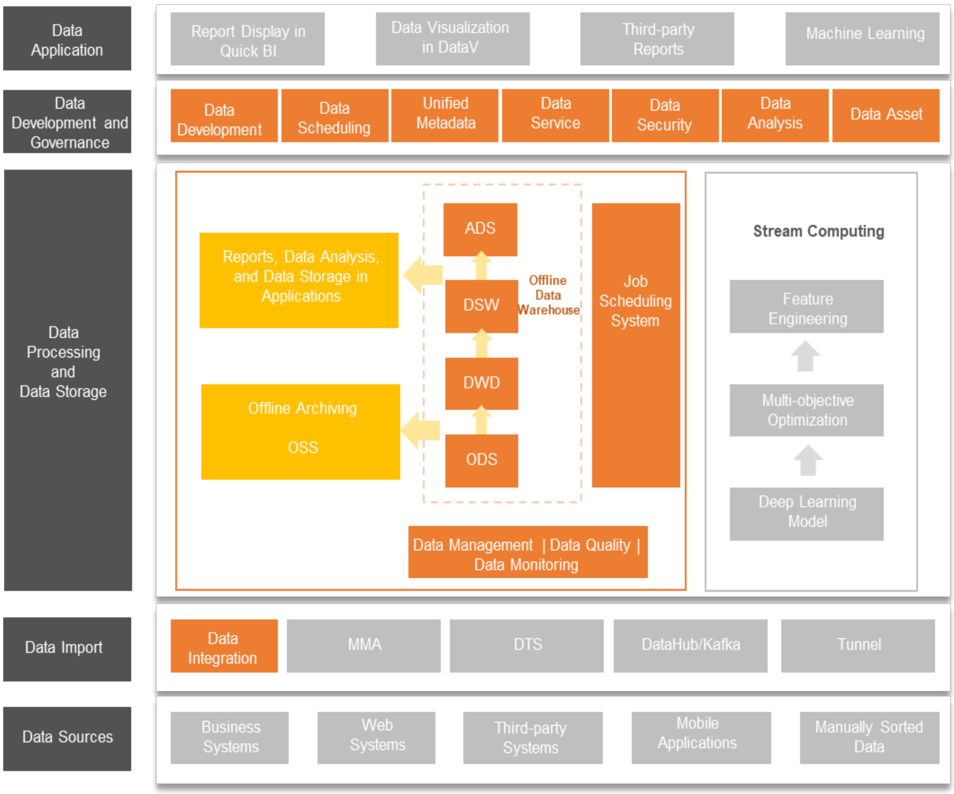
You can use DataWorks to process and analyze massive data in offline mode. You can also use DataWorks to complete the best practices that cover the full lifecycle of big data. The best practices include data aggregation and integration, development, scheduling and O&M, online and offline analysis of data, data quality governance and asset management, security audit, data sharing and services, machine learning, and application building. DataWorks provides an end-to-end solution from data collection to data display and from data analysis to application driving and helps users apply data in business and present business status by using data.

4.2. Data warehouse solutions

DataWorks allows you to develop and implement data governance in a visualized manner. This topic describes how to build a big data warehouse on the cloud and a real-time data warehouse.

Big data warehousing solution

We recommend that you build a big data warehouse on the cloud based on the following architecture:



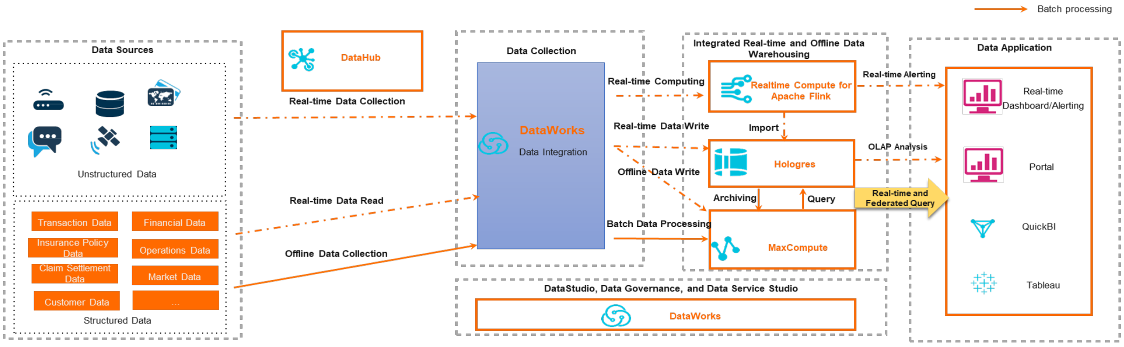
- **Customers:** This solution is suitable for customers from all industries.
- **Benefits:** The big data warehouse solution is the best practice of Alibaba Cloud and features high performance, low cost, and a serverless architecture. You can use this solution to build O&M-free and fully managed big data warehouses. This way, big data developers of enterprises can focus on the

development, production, and governance of business data.

- **Product portfolio** :MaxCompute, Realtime Compute for Apache Flink, and DataWorks.
- **Use scenarios:**
  - User data comes from various sources, such as the cloud and external data sources. Data from different sources is integrated into a data warehouse in a unified manner for data cleansing and data modeling.
  - The application scenarios are complex. You can use a big data warehouse to perform speech recognition, semantic analysis, and sentiment analysis for unstructured speech and natural language text. You can also build an enterprise-class data management platform to process structured data. This helps reduce computing and storage costs.
  - A data warehouse that supports abundant applications is required. You can use machine learning algorithms for complex data analysis, BI reports for chart display, products for data display on the dashboard, and other custom methods for data consumption.

Real-time data warehousing solution

We recommend that you build a real-time data warehouse based on the following architecture:

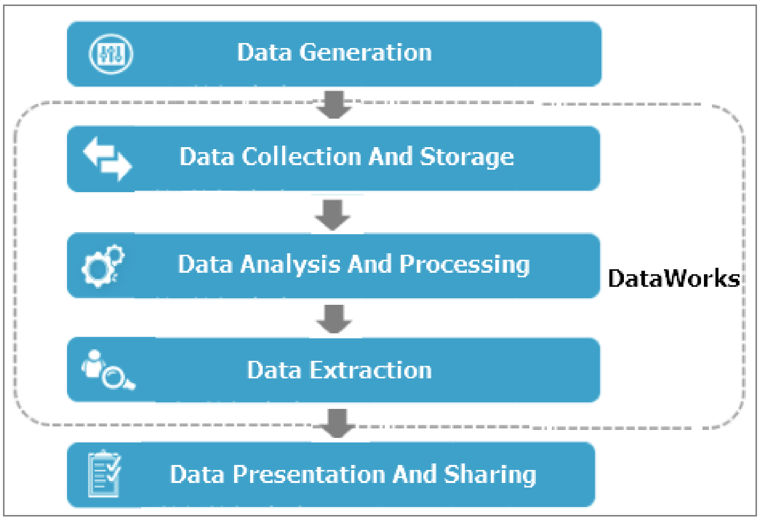


- **Customers**: This solution is suitable for scenarios in which large amounts of data need to be queried in real time in Internet industries, such as e-commerce, gaming, and social networking.
- **Benefits**:
  - Alibaba Cloud real-time data warehouses can be seamlessly integrated with offline data warehouses.
  - A single cost-effective storage system can satisfy the requirements for real-time and offline computing.
- **Product portfolio** :DataHub, Realtime Compute for Apache Flink, Hologres, MaxCompute, DataWorks, and Quick BI or DataV.
- **Use scenarios:**
  - Data collection: You can use DataWorks to collect batch data and DataHub to collect real-time data.
  - Data development: You can use DataWorks to complete end-to-end data development. The data development process includes data integration, extract, transform, and load (ETL), data computing, and scheduling, monitoring, and alerting of nodes. DataWorks provides security control capabilities to eliminate security risks in the data development process. DataWorks also provides unified DataService Studio APIs based on the DataStudio service.

- Real-time data processing: You can use Realtime Compute for Apache Flink to perform real-time ETL and import the results to databases. Then, you can use Hologres to build real-time data warehouses and application marts and perform real-time interactive query and analysis of large amounts of data.
- Interactive analysis: You can use a real-time data warehouse to perform real-time, offline, and federated queries. Historical offline data is stored in MaxCompute, and real-time analysis data is stored in Hologres. You can use Alibaba Cloud Quick BI or a third-party data analysis tool, such as Tableau, to visualize data and build data applications.

4.3. Data development process

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


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

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

The data development process involves the following steps:

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

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

5.Services that work with DataWorks

DataWorks can work with compute engines to support end-to-end big data development and governance. DataWorks allows you to add data sources to Data Integration and then use Data Integration to transmit data between the data sources. This topic provides the services that can work with DataWorks in typical scenarios.

Supported compute engines

DataWorks allows you to associate compute engine instances with your DataWorks workspaces. After you associate a compute engine instance with a DataWorks workspace, you can create nodes of the same type as the compute engine instance in the DataWorks console and then enable the system to periodically schedule the nodes. DataWorks supports the following compute engines:

- MaxCompute
- E-MapReduce
- Hologres
- ADB for Posgre
- ADB for Mysql
- CDH
- ClickHouse

For more information about how to associate a compute engine instance with a workspace, see Configure a workspace.

Supported data sources

DataWorks can synchronize batch data or real-time data between different data sources. You can configure clusters or instances in the following services as the data sources of DataWorks: Alibaba Cloud services and self-managed services that are related to databases, unstructured storage, big data, and message queues. You can use DataWorks to integrate data only after you configure the data source.

- For more information about data sources that support batch synchronization, see Supported data source types, readers, and writers.
- For more information about data sources that support real-time synchronization, see Plug-ins for data sources that support real-time synchronization.

6.Terms

This topic describes the terms that are related to DataWorks, including workspace, workflow, solution, SQL script template, node, instance, commit operation, script, resource, function, and output name.

workspace

A workspace is a basic unit for managing nodes, members, roles, and permissions in DataWorks. The administrator of a workspace can add users to the workspace as members and assign the Workspace Manager, Development, O&M, Deploy, Safety Manager, or Visitor role to each member. This way, workspace members to which different roles are assigned can collaborate with each other.

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

You can associate compute engine instances such as MaxCompute, E-MapReduce (EMR), and Realtime Compute for Apache Flink compute engine instances with a workspace. After you associate compute engine instances with a workspace, you can configure and schedule nodes in the workspace.

workflow

A workflow is abstracted from business to help you manage and develop code based on your business requirements and improve the efficiency of node management.

> **Note** A workflow can be added to multiple solutions.

Workflows help you manage and develop code based on your business requirements. A workflow has the following features:

- Allows you to develop and manage code by node type.
- Supports a hierarchical directory structure. We recommend that you create a maximum of four levels of subdirectories for a workflow.
- Allows you to view and optimize a workflow from the business perspective.
- Allows you to deploy and manage nodes in a workflow as a whole.
- Provides a dashboard for you to develop code with improved efficiency.

solution

A solution contains one or more workflows.

Solutions have the following benefits:

- A solution can contain multiple workflows.
- Multiple solutions can use the same workflow.
- An organizational solution can contain various types of nodes. This improves user experience.

SQL script template

SQL script templates are general logic chunks that are abstracted from SQL scripts and can help reuse code.

Each SQL script template involves one or more source tables. You can filter source table data, join source tables, and aggregate source tables to generate a result table based on your business requirements. An SQL script template contains multiple input and output parameters.

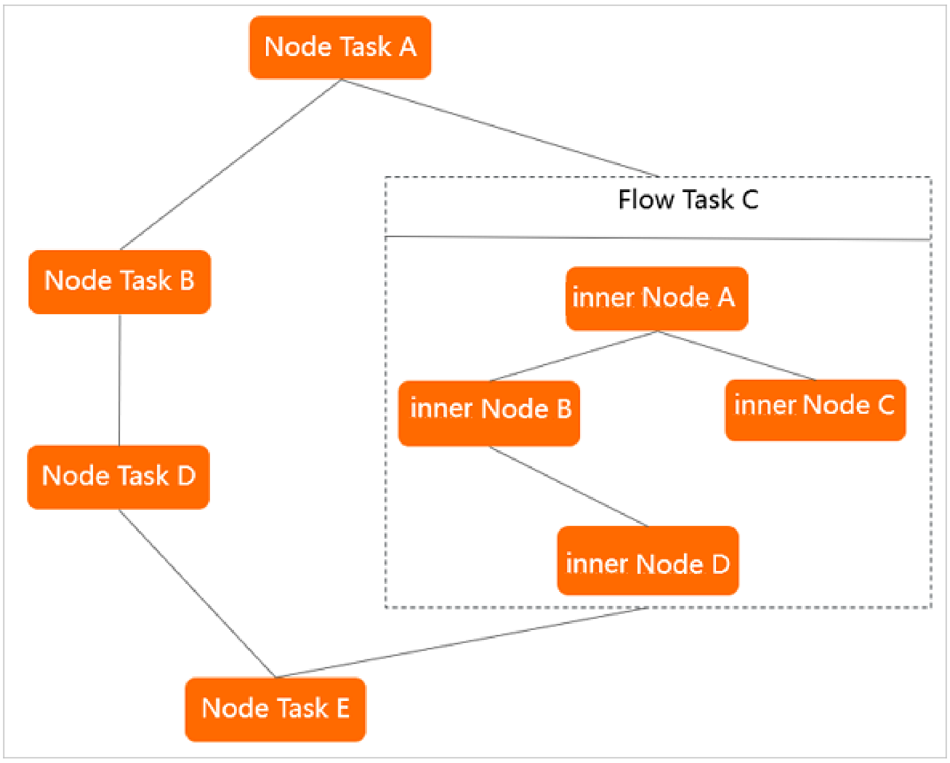
node

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

- A synchronization node is used to synchronize data from a source to a destination, such as data synchronization from ApsaraDB RDS to MaxCompute.
- An ODPS SQL node is used to convert data by executing SQL statements that are supported by MaxCompute.

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

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



Node type	Description
Node task	A node task is used to perform a data operation. You can configure dependencies between a node task and flow tasks or other node tasks to form a directed acyclic graph (DAG).

Node type	Description
Flow task	A flow task contains a group of inner nodes that process a workflow. We recommend that you create less than 10 flow tasks in a workspace. Inner nodes in a flow task cannot be configured as the dependencies of node tasks or other flow tasks. You can configure dependencies between a flow task and node tasks or other flow tasks to form a DAG. **Note** In DataWorks V2.0 and later, the flow tasks that are created in DataWorks V1.0 are retained, but you cannot create flow tasks. Instead, you can create workflows to perform similar operations.
Inner node	An inner node is a node within a flow task. The features of an inner node are basically the same as those of a node task. You can drag lines between inner nodes in a flow task to configure dependencies. However, you cannot configure scheduling properties for inner nodes because these nodes use the scheduling configurations of the flow task.

instance

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

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

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

commit

You can commit nodes and workflows from the development environment to the scheduling system. The scheduling system runs the code of the committed nodes and workflows based on the related configurations.

Note The scheduling system runs only committed nodes and workflows.

script

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

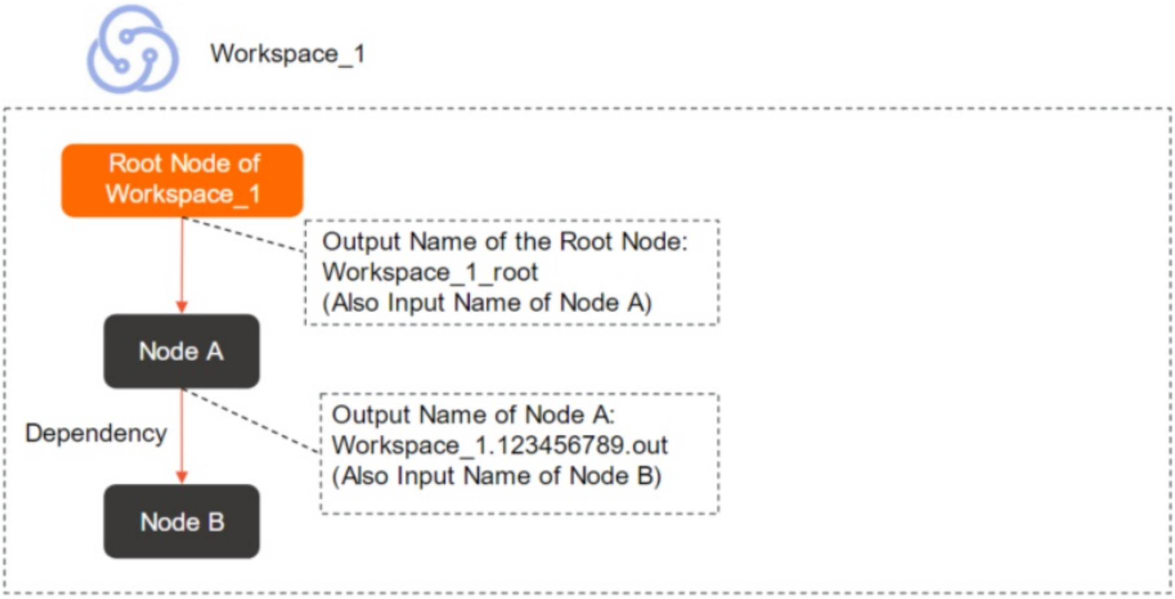
resource and function

Resources and functions are terms in MaxCompute and refer to resources and functions that are used by the MaxCompute compute engine. For more information, see Resource and Function.

output name

Each node has an output name. When you configure dependencies between nodes within an Alibaba Cloud account, the output name of a node is used to connect to its descendant nodes.

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


Workspace_1
Root Node of Workspace_1
Output Name of the Root Node: Workspace_1_root (Also Input Name of Node A)
Node A
Dependency
Output Name of Node A: Workspace_1.123456789.out (Also Input Name of Node B)
Node B

Note The output name of a node distinguishes the node from other nodes within the same Alibaba Cloud account. By default, the output name of a node is in the following format: Workspace name.Randomly generated nine-digit number.out. You can customize the output name for a node. You must make sure that the output name of the node is unique within your Alibaba Cloud account.

metadata

Metadata describes data attributes, data structures, and other relevant information. Data attributes include the name, size, and data type, data structures include the field, type, and length, and other relevant information includes the location, owner, output node, and access permissions. In DataWorks, metadata refers to information about tables or databases. DataMap is the main service used to manage metadata.

data backfill

After an auto triggered node is developed, and committed and deployed to the scheduling system, the scheduling system runs the node as scheduled. If you want to perform computing on data that is generated in a historical period of time, you can backfill data for the node. The generated data backfill instance is run based on the specified data timestamp.