DataWorks

Product Introduction

Product Introduction

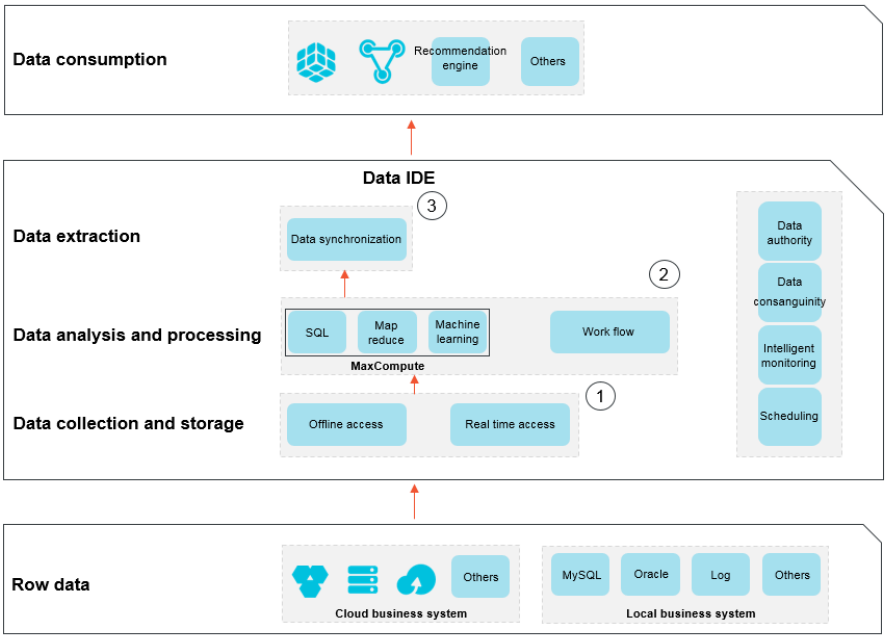
What is DataWorks

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

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

> **DataWorks** is a big data PaaS platform released by Alibaba Cloud. As a one-stop DW capability platform, it offers a wide-range of products and services, including data integration, data development, data management, and data governance.

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



Function overview

Fully-hosted scheduling

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

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

Supports various task types

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

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

Visual development

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

Monitoring and alarms

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

The alarm service provides the monitoring alarm capability allowing you to obtain up-to-date metric data for troubleshooting any cloud product abnormality in a timely manner. You can create alarm rules and add an alarm contact and alarm contact group.

You must provide the contact and contact group information, which is a prerequisite for the alarm rule function. This is because when any exception occurs, an alarm is triggered and the alarm notification is sent to the alarm contact and the alarm contact group. The alarm notifications can be

sent through text message, an email or TradeManager. Hence, it is required to create a contact and a contact group when you begin to use the alarm rules function for the first time.

Constraints and limitations

DataWorks only supports Chrome 54 or later.

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

Scenarios

DataWorks provides an extensive array of scenarios. The various scenarios and advantages of DataWorks are explained as follows:

Builds a cloud platform for Internet big data application services

- Allows enterprises to focus more on the core business

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

- Reduces investment and O&M costs

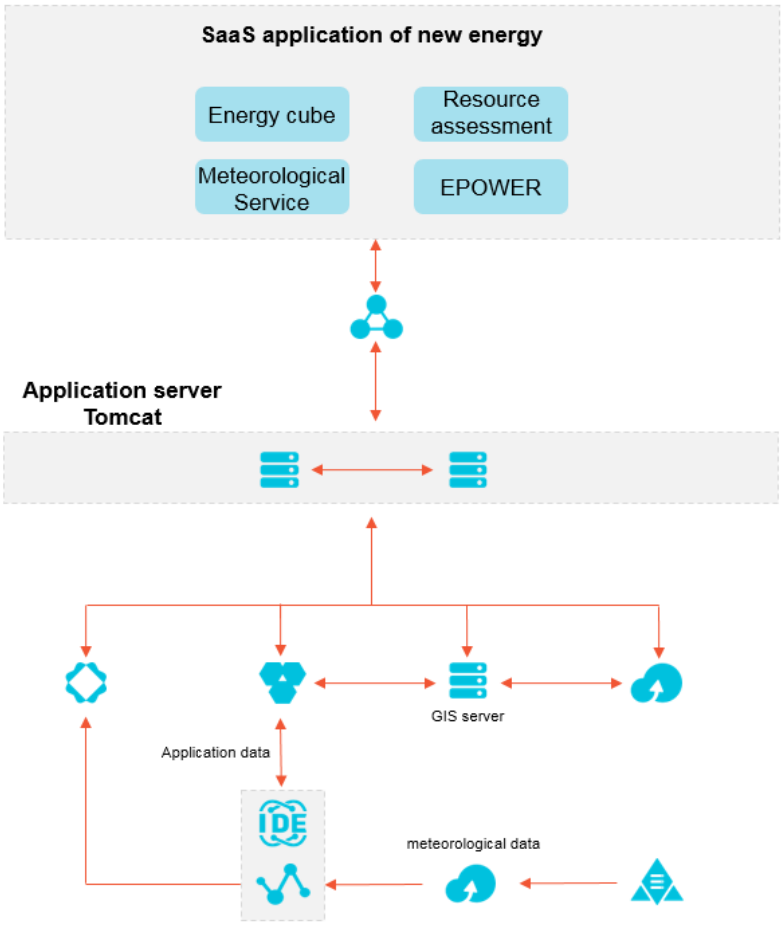
It can greatly reduce the material resources, labor, and R&D investment required for any self-built big data platforms.

Security and stability

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

Recommended combination:

DataWorks + AnalyticDB + MaxCompute



Weather queries and advertisement business log analysis

Advantages:

Improves work efficiency

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

Improves storage utilization

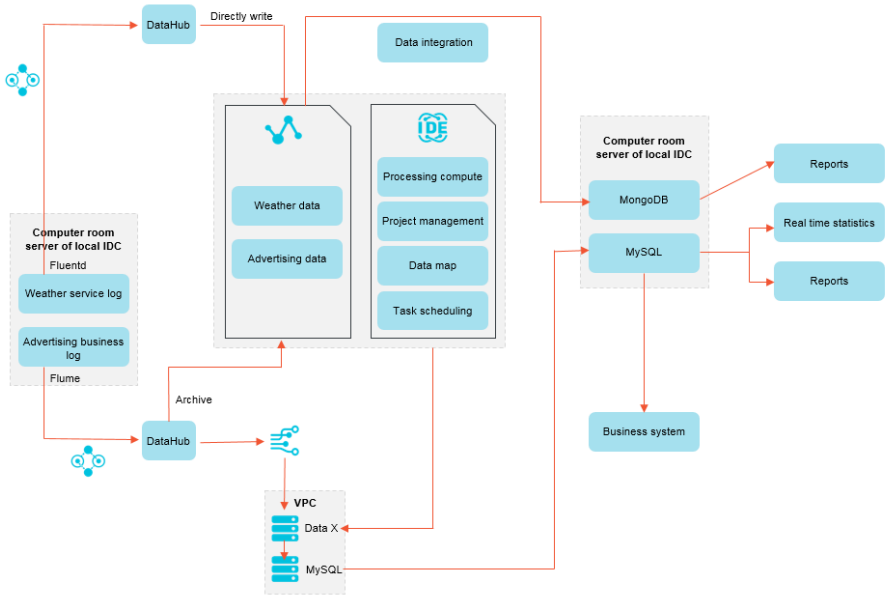
DataWorks can reduce overall storage and computing costs by 70%, improving both performance and stability.

Makes big data products easy to use

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

Recommended combination:

DataWorks + Data Integration + AnalyticDB + Quick BI + MaxCompute



Detail-oriented operations

Improves business insights

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

Data-driven businesses

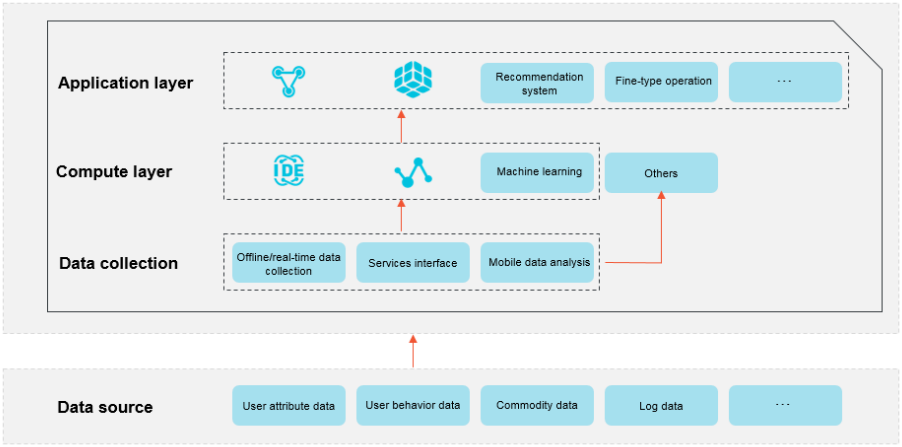
DataWorks empowers businesses by providing enhanced data analysis capabilities and effective monitoring functions.

Quick response to business needs

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

Recommended combination:

DataWorks + Data integration + Quick BI + MaxCompute



Basic terms

Task

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

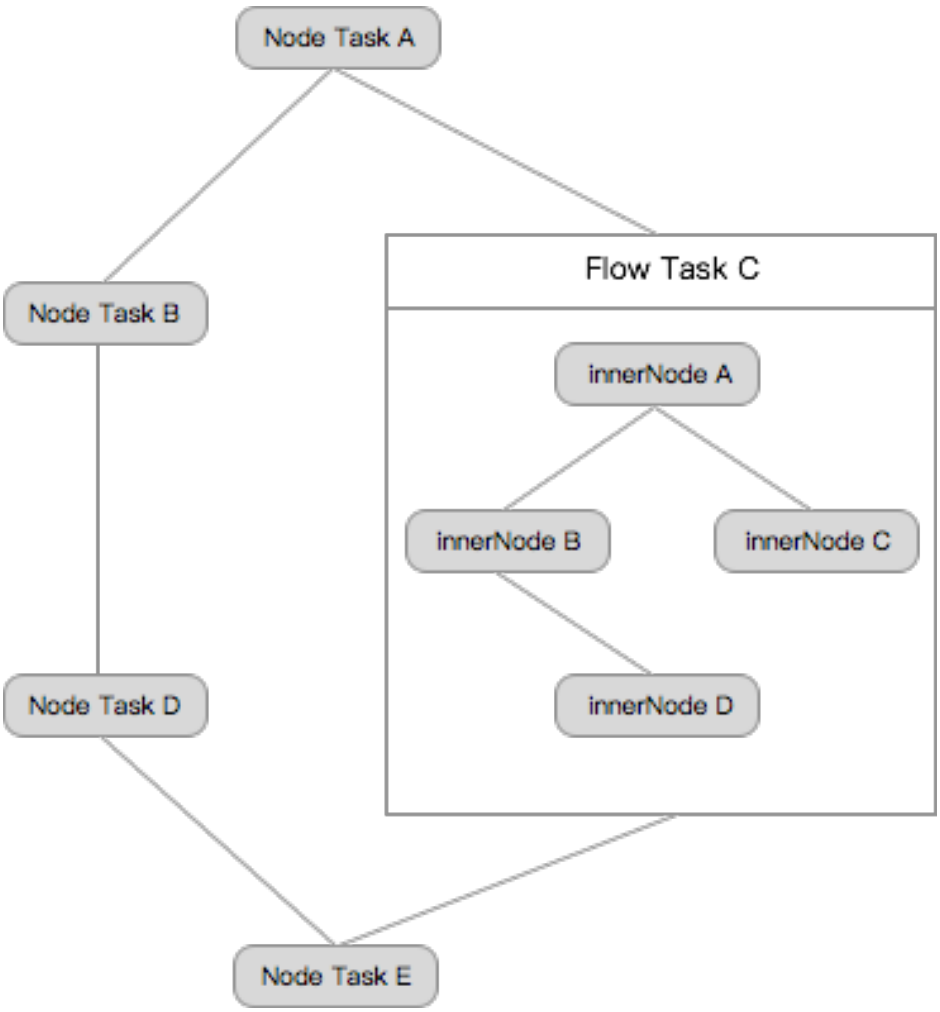
A data synchronization node task is used to copy data from RDS to MaxCompute.

A MaxCompute SQL node task is used to run MaxCompute SQL for data conversion.

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

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

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



A node task is an operation performed on data. It can be configured to be dependent on other node tasks and flow tasks to form a Directed Acyclic Graph (DAG).

A flow task is formed by a group of inner nodes that are processing a small business. We recommend using less than 10 flow tasks. Inner nodes of a flow task cannot depend on by other flow or node tasks. A flow task can be configured to be dependent on other flow and node tasks to form a DAG.

An inner node is a node inside a flow task. It basically provides the same capabilities as a node task. Its scheduling frequency is inherited from the scheduling frequency of the flow task, and cannot be configured independently. The dependency can only be dragged.

For more infomation about data operation types, see Task type description.

Instance

When a task is scheduled by the system or triggered manually, an instance is generated. An instance

is a snapshot that runs by a task at a certain moment. The instance contains the task operating time, operating status, operating logs, and other information. For example:

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

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

Submit

Submit is a process by which the developed node task or flow task is released from the development environment to the scheduling system. After a task is submitted, its code and scheduling configuration are synchronized to the scheduling system, which schedules the task according to the configuration.

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

Script

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

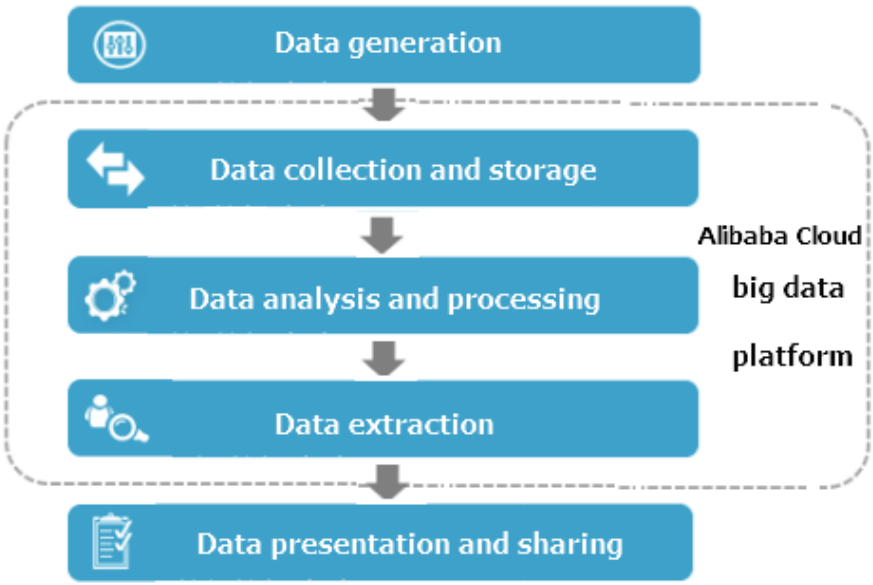
Resources and functions

Both resources and functions are concepts of MaxCompute. For more information, see MaxCompute resources and MaxCompute functions.

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

Data development processs

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



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

The data development process is explained as follows:

Data generation

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

Data collection and storage

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

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

Data analysis and processing

Next, you can start to process (ODPS_SQL and OPEN_MR), analyze, and mine (data analysis and data mining) the data on MaxCompute to find valuable information.

Data extraction

The data after analysis and processing must be synchronized to your business system for

further use.

Data presentation and sharing

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

Functions

Common actions

- Update account information
- Create a project
- Connectivity test of data sources
- Add member and authorize
- Perform periodic O&M and view log troubleshooting results
- Create tables
- Upload a local file
- Create a task
- Create UDF
- OPEN MR
- Cyclic tasks
- System scheduling parameters