DataWorks

Product Introduction

MORE THAN JUST CLOUD | Alibaba Cloud

Product Introduction

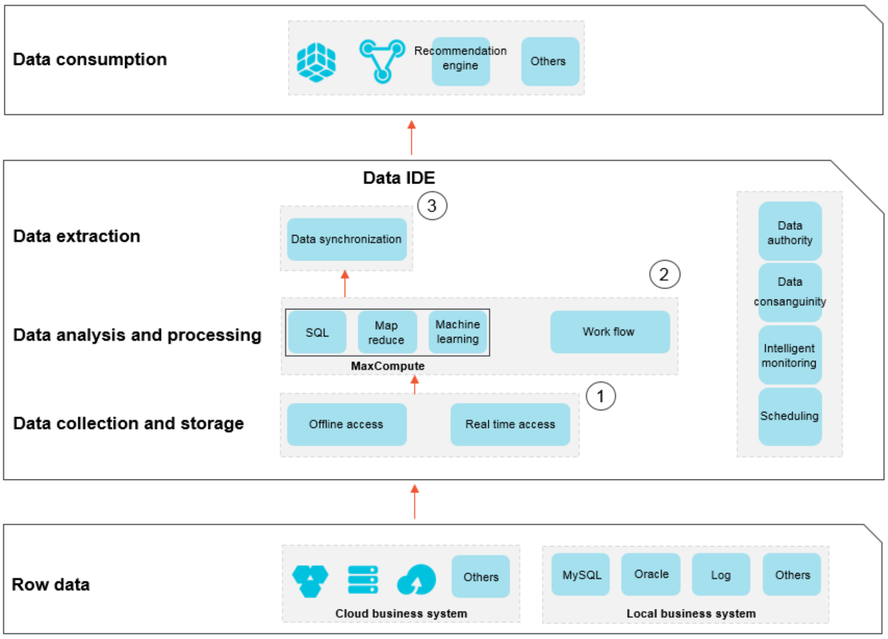
What is DataWorks

The DataWorks is an important PaaS product in the Alibaba Cloud. It provides fully hosted workflow services and a one-stop development and management interface to help enterprises mine and explore the full 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 provides a comprehensive range of products and services, including data integration, data development, data management, and data governance.

With DataWorks, you can transmit, convert and other work with data. This allows you to import data from different storage services, and convert and ultimately extract the data to other data systems. A complete data analysis process is shown in the following figure:



Function overview

Fully-hosted scheduling

DataWorks provides powerful scheduling capabilities including time-based or dependency-based task trigger mechanisms to perform **tens of millions** of tasks accurately and punctually each day based on DAG relationships. It supports multiple scheduling frequency configurations, by minute, hour, day, week, and month.

The fully-hosted service removes your worry about scheduling server resources. The system isolates different tenants, ensuring that their tasks do not interfere with each other.

Supports various task types

DataWorks supports multiple task types, including **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 ensure the 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 highly-efficient data transmission. For more details, see Data integration overview.

Visual development

This product provides 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 enables you to carry out development tasks wherever you are.

Monitoring and alarms

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

You can easily configure text message alarms, so that the relevant staff will be notified as soon as a task error occurs. This ensures the smooth operation of your business.

Constraints and limitations

- Only supports Chrome 54 or later.

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

Application scenarios

Build a cloud platform for Internet big data application services of new energy industry

Advantages:

Allows enterprises to be more focused on their actual businesses

Complete businesses can be delivered to the cloud within a short period of time, so the massive resources on the cloud can truly provide services to the business. With Alibaba Cloud’ s mature business scaling solutions, enterprises do not need to focus too much on seamless service expansion and other specific matters.

Reduces investment and O&M costs

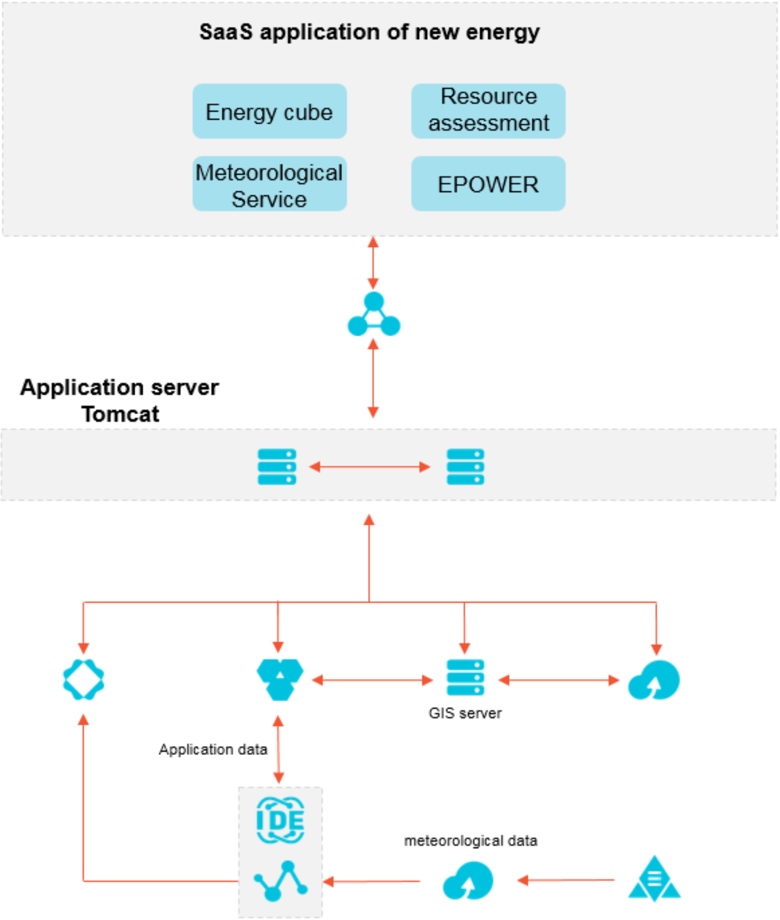
It can greatly reduce the material resources, labor, and R&D investment required for self-built big data platforms.

Security and stability

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

Recommended combination:

DataWorks + AnalyticDB + MaxCompute



Weather query and advertisement business log analysis

Advantages:

Improves work efficience

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

Improves storage utilization

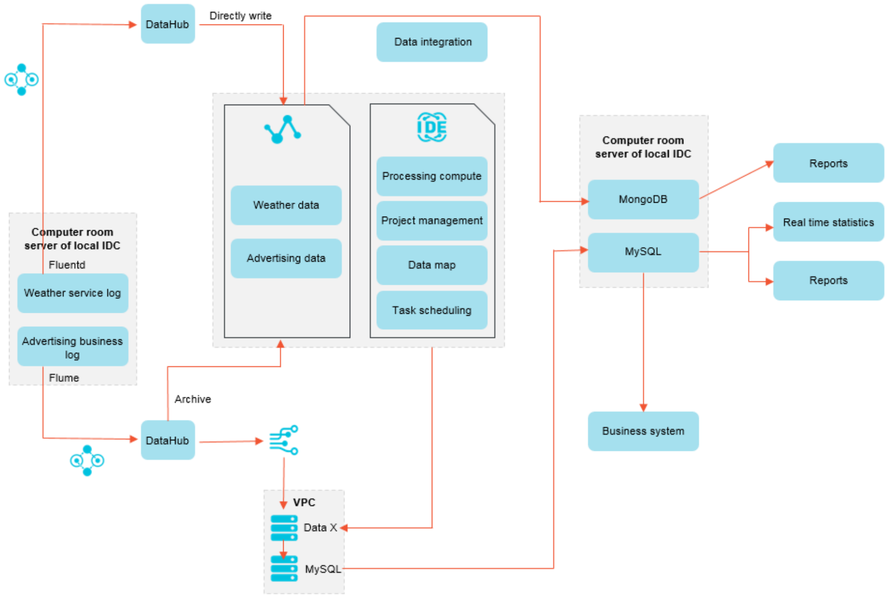
DataWorks can reduce overall storage and computing costs by 70%, improving both performance and stability.

Makes big data product is easier 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

Advantages:

Improves business insights

MaxCompute’ s computing capability can achieve detailed operation for millions of users.

Data-driven businesses

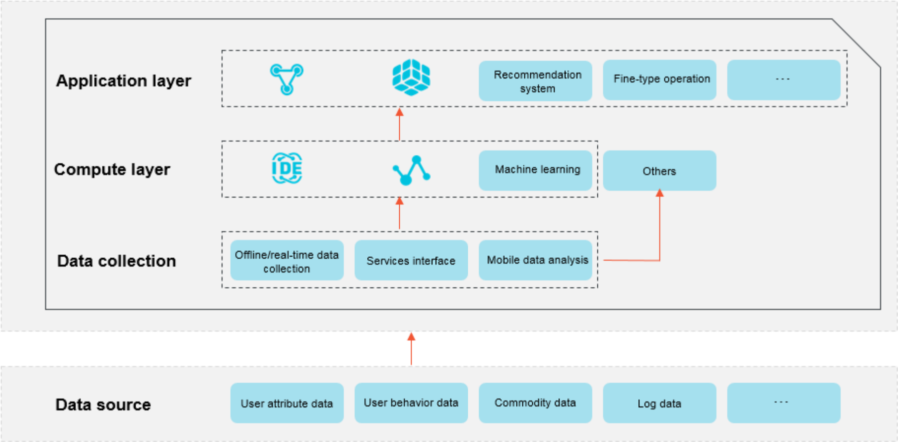
DataWorks empowers businesses by providing better data analysis capabilities and effective monitoring functions.

Rapid response to business needs

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

Recommended combination:

DataWorks + Data integration + Quick BI + MaxCompute



Terms

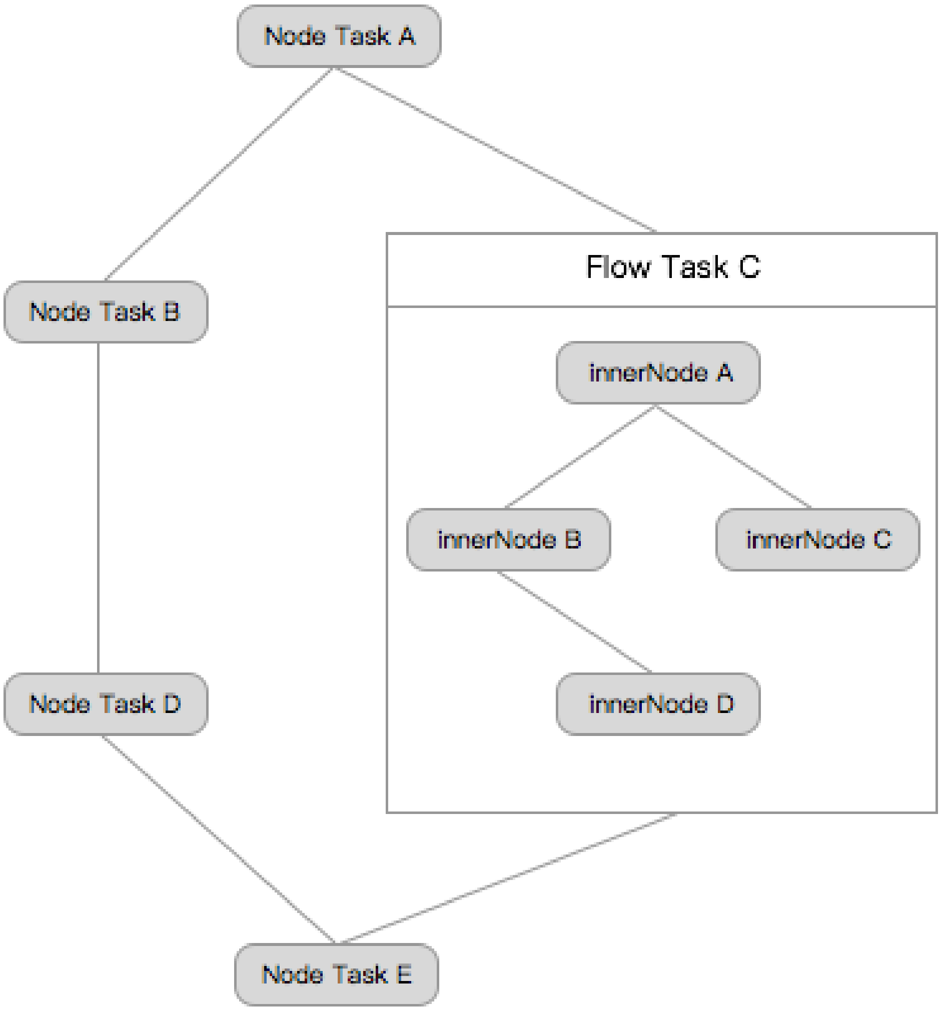
Tasks

A task is an operation performed on data. For example:

- Use a data synchronization node task to copy data from RDS to MaxCompute.
- Use a MaxCompute SQL node task to run MaxCompute SQL for data conversion.
- Use a flow task to perform a series of data conversions among several inner SQL nodes.

Each task uses zero or more data tables (data sets) as inputs, and generates one or more data tables (data sets) as outputs.

Tasks are divided into node tasks, flow tasks, and inner nodes. The relationships between these tasks are shown 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 a task 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 be depended 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 provides basically 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.

Instances

When a task is scheduled by the system or triggered manually, an instance is generated. An instance is a snapshot that is executed 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 2:00 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 2:00 on the next day. When detecting that the upstream task is completed, the system automatically runs the task 1 instance at 2:00 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.

Scripts

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

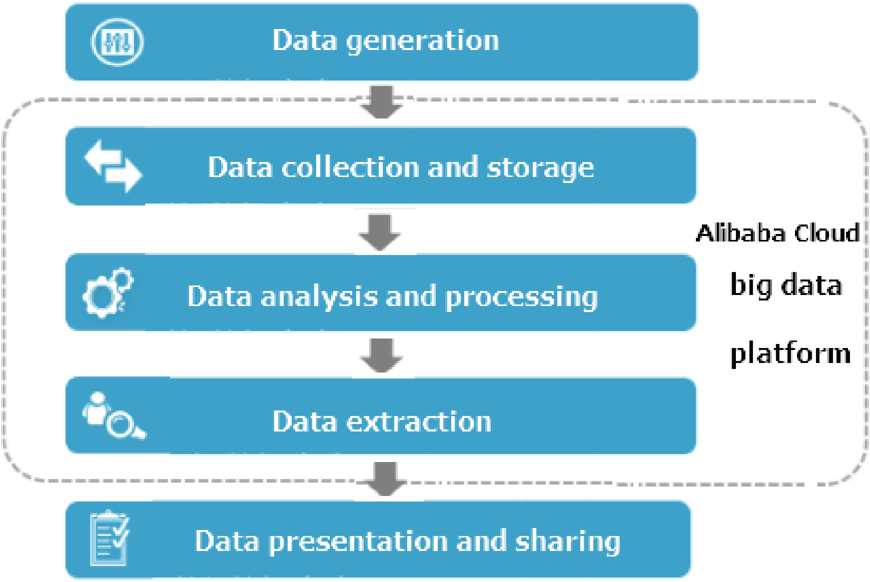
Resources and functions

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

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

Data development processs

Normally, the overall data development process includes data generation, data collection and storage, data analysis and processing, data extraction, and data presentation and sharing, as shown in the following figure:


Data generation
Data collection and storage
Data analysis and processing
Data extraction
Data presentation and sharing
Alibaba Cloud
big data
platform

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 described 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 use by staff.

Data presentation and sharing

Finally, the result of big data analysis and processing are presented and shared as reports, geographic information systems, and in various of other formats.

Functions

Common actions

- New Project
- New Data Source
- Add a Project Member
- Tabulation with Script
- Visualized Tabulation
- Import Data
- New Wordflow
- New Job
- New UDF
- Use OPEN MR
- Workflow O&M
- Scheduling Parameter Usage