# 数据工场

## 产品简介

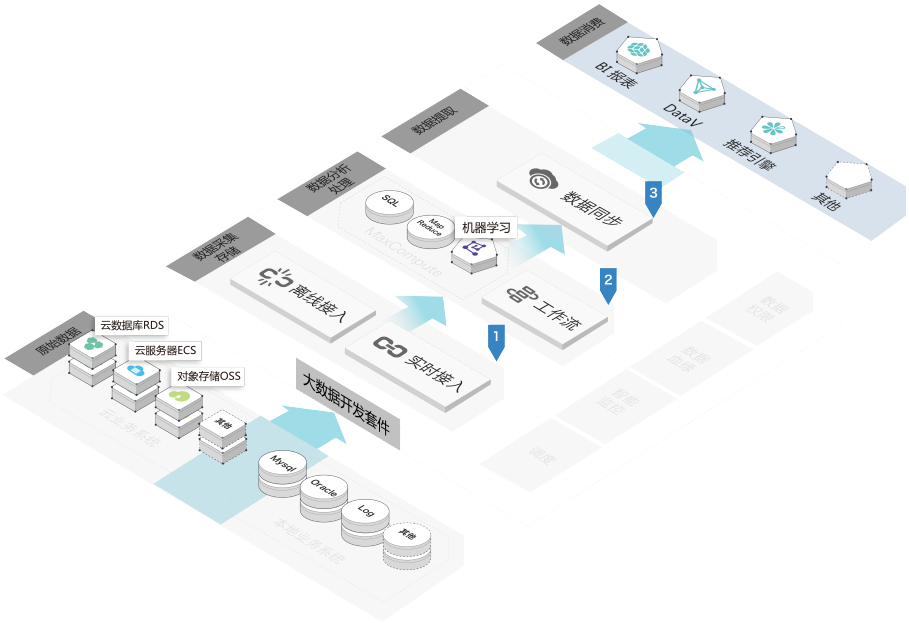
# 产品简介

DataWorks（数据工场，原大数据开发套件）是阿里云数加重要的 PaaS 平台产品，它提供全面托管的工作流服务，一站式开发管理的界面，帮助企业专注于数据价值的挖掘和探索。

DataWorks（数据工场）基于 MaxCompute 作为核心的计算、存储引擎，提供了海量数据的离线加工分析、数据挖掘的能力，详情请参见 MaxCompute 简介。

> **DataWorks（数据工场）** 是阿里云推出的大数据领域 PaaS 平台，是一站式的 DW 能力平台，提供数据集成、数据开发、数据管理、数据治理等全方位的产品服务。

使用 DataWorks（数据工场），可对数据进行数据传输、数据转换等相关操作，从不同的数据存储引入数据，对数据进行转化处理，最后将数据提取到其他数据系统。完成整个数据的分析流程，如下图所示：




## 功能概述

### 全面托管的调度

提供强大的调度能力，支持按照时间、依赖关系的任务触发机制，支持每日 **千万级别** 的任务按照 DAG 关系准确、准时运行。支持分钟、小时、天、周和月多种调度周期配置。

完全托管的服务，无需关心调度服务器资源问题。租户之间提供隔离，保证不同租户之间的任务不会相互影响。

## 支持多种任务类型

支持 **数据同步、SHELL、MaxCompute SQL、MaxCompute MR** 等多种任务类型，通过任务之间的相互依赖完成复杂的数据分析处理。

> 数据转化能力依托 MaxCompute 强大的能力，保证了大数据的分析处理性能。更多详情请参见 MaxCompute 简介。

> 数据同步能够依托 **DataWorks（数据工场）> 数据集成** 的强力支撑，支持多达 20+ 数据源，提供稳定高效的数据传输。更多详情请参见 数据集成简介。

## 可视化开发

提供可视化的代码开发、工作流设计器页面，无需搭配任何开发工具，简单的拖拽和开发就可以完成复杂的数据分析任务。只要有浏览器有网络，便可随时随地进行开发工作。

## 监控告警

运维中心提供可视化的任务监控管理工具，支持以 DAG 图的形式展示任务运行时的全局情况。

可方便地配置短信报警，任务发生错误可及时通知相关同学，保证业务正常运行。

# 约束与限制

- 仅支持 Chrome 浏览器 54 以上版本。
- 目前无法支持 SQL 运行在阿里云云数据库、阿里云分析型数据库等产品，仅支持 MaxCompute。

# 搭建新能源产业互联网大数据应用服务云平台

**能够实现：**

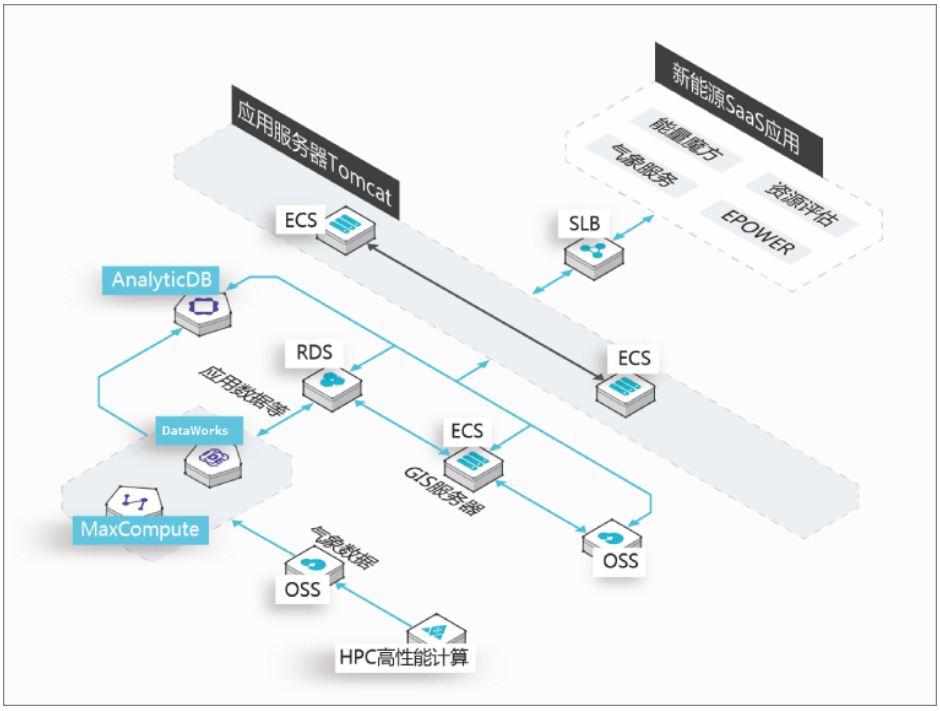
> 让企业更专注于业务
>
> 可在短时间内，将业务全面的交付云端，让云端的海量资源真正为业务服务。阿里云成熟的业务扩展方案可让企业在业务无缝扩展等具体事务上无需操心太多。
>
> 降低投资、运维成本
>
> 可极大减少自建大数据平台的物力投入、人力运维投入和研发投入。

安全稳定

全方位服务能力及其稳定安全的表现可确保数据上云万无一失。

**推荐搭配使用：**

DataWorks（数据工场，原大数据开发套件）+ AnalyticDB + MaxCompute




# 天气查询业务和广告业务日志分析

**能够实现：**

提高工作效率

日志数据全部通过 SQL 进行分析，工作效率可提升 5 倍以上。

提升存储利用率

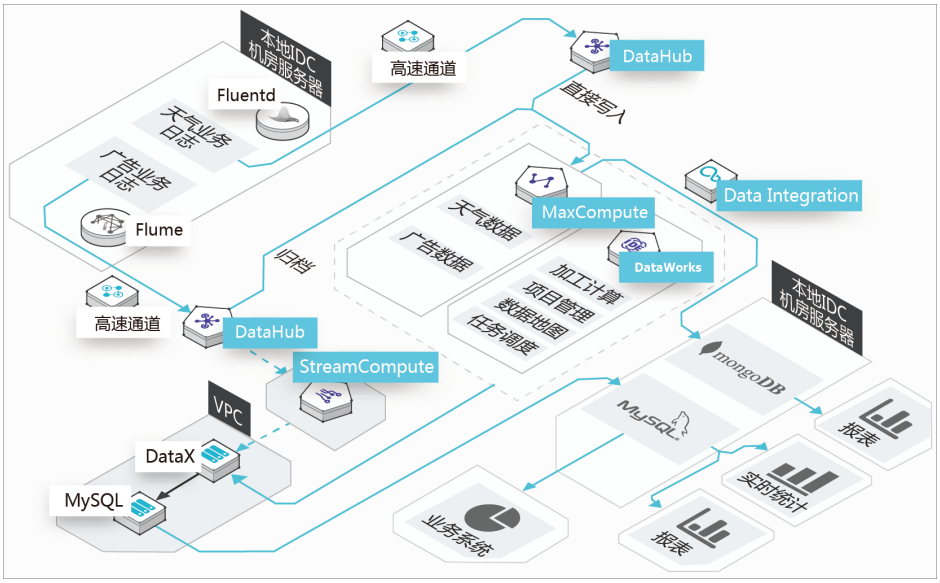
整体存储和计算的费用可节省 70%，性能和稳定性均可提升。

降低大数据使用门槛

MaxCompute 提供多种开源软件的插件，可轻松完成数据上云。

**推荐搭配使用：**

DataWorks（数据工场，原大数据开发套件）+ 数据集成 + AnalyticDB + Quick BI + MaxCompute




# 精细化运营

**能够实现：**

提升业务洞察能力

通过 MaxCompute 计算能力可实现针对百万用户的精细化运营。

业务数据化

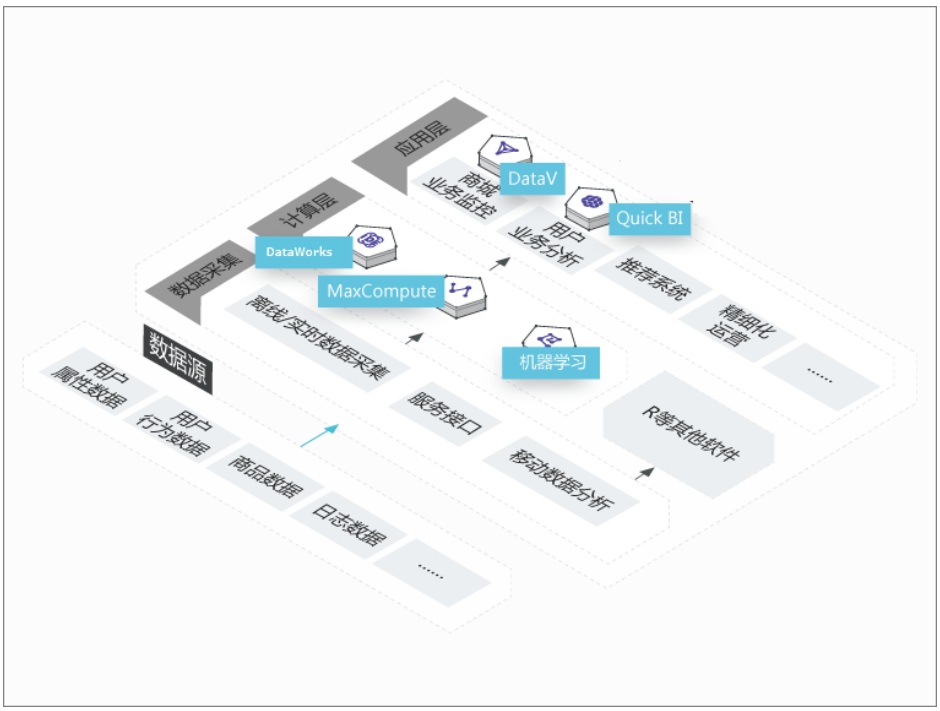
可提升对业务数据的分析能力并进行有效监控，更好的业务赋能。

快速响应业务需求

数加生态满足新业务数据分析需求的随机应变能力。

**推荐搭配使用：**

DataWorks（数据工场，原大数据开发套件）+ 数据集成 + Quick BI + MaxCompute




关于更多应用场景，请参见 客户案例。

# 任务（Task）

任务是指定义对数据执行的操作。示例如下：

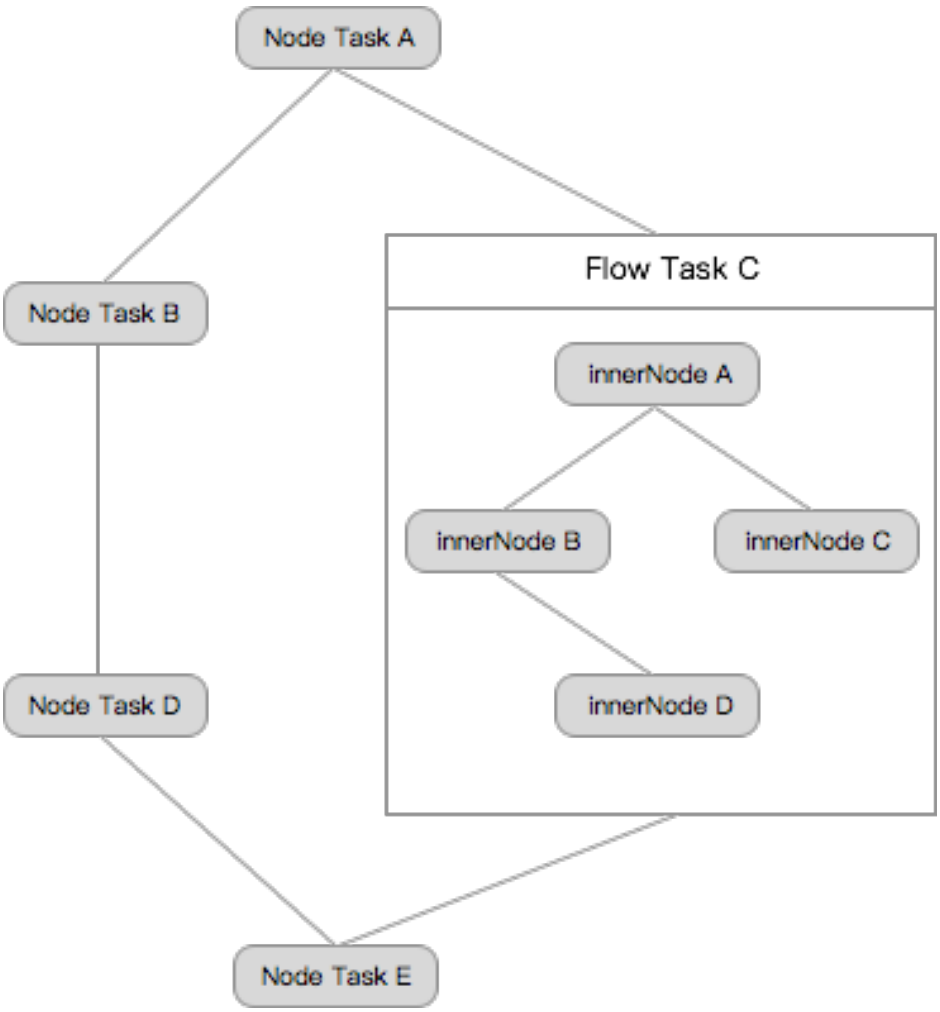
使用数据同步节点任务将数据从 RDS 复制到 MaxCompute。

使用 MaxCompute SQL 节点任务运行 MaxCompute SQL 来进行数据的转换。

使用工作流任务，通过内部几个不同的 SQL 内部节点完成一系列的数据转换。

每个任务使用 0 或 0 个以上的数据表（数据集）作为输入，生成一个或多个数据表（数据集）作为输出。

任务主要分为三种：节点任务（node task），工作流任务（flow task）和内部节点（inner node）。三种类型的关系如下图所示：




节点任务（node task）：一个数据执行的操作。可以与其他节点任务、工作流任务配置依赖关系，组成 DAG 图。

工作流任务（flow task）：解决一个小业务的一组内部节点组成一个工作流任务。工作流任务数量建议小于 10 个。工作流任务内部节点，无法被其他工作流任务、节点任务依赖，工作流任务可以与其他工作流任务、节点任务配置依赖关系，组成 DAG 图。

内部节点（inner node）：工作流任务内部的节点，与节点任务能力基本相同。其调度周期会继承工作流任务的调度周期，无法进行单独配置，依赖关系也按照拖拽关系。

数据执行可以选择的操作类型，请参见 任务类型介绍。

任务的调度参数配置，请参见 调度配置介绍。

# 实例（Instance）

在调度系统中的任务经过调度系统、手动触发运行后会生成一个实例，实例代表了某个任务在某时某刻执行的

一个快照，实例中会有任务的运行时间、运行状态、运行日志等信息。示例如下：

假如设置每天 2:00 运行 task1 任务，调度系统会在每天 23：30 根据周期节点任务定义好的时间自动生成一个快照，也就是 task1 的一个第二天 2:00 运行的实例，到第二天 2:00 时，同时判断上游任务已经完成，task1 实例便会如期启动运行。

> **注意**：
>
> 您可以在 **运维中心 > 任务运维** 页面查询任务实例的相关信息。

# 提交（Submit）

提交是指开发的节点任务、工作流任务从开发 DataWorks 环境发布到调度系统的过程。完成提交以后，相应的代码、调度配置全部合并到调度系统中，调度系统按照相关配置进行调度操作。

> **注意**：
>
> 未提交的节点任务、工作流任务不会进入到调度系统。

# 脚本开发（Script）

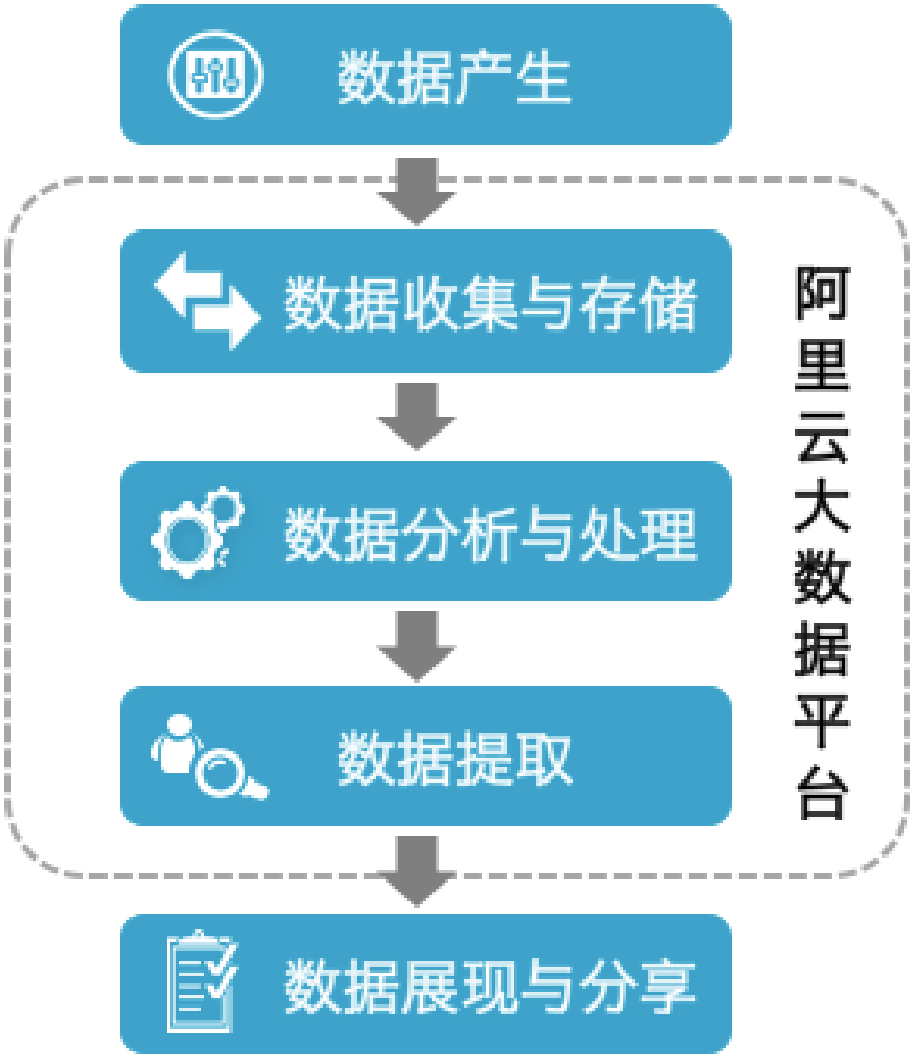
脚本开发是提供给数据分析使用的一个代码存储空间，脚本开发的代码无法发布到调度系统，无法进行调度参数配置，仅可以进行一些数据查询分析的工作。

# 资源、函数

资源、函数均为 MaxCompute 的概念，详情请参见 MaxCompute 资源 和 MaxCompute 函数。

在 DataWorks（数据工场，原大数据开发套件）中，可以通过界面管理资源、函数。如果通过 MaxCompute 的其他方式进行资源、函数管理，则无法在 DataWorks 中进行相关的查询。

通常情况下，数据开发的总体流程包括数据产生、数据收集与存储、数据分析与处理、数据提取和数据展现与分享，如下图所示：




**注意：**

上图中，虚线框内的开发流程都可基于阿里云大数据平台来完成。

数据开发流程说明如下：

**数据产生**

业务系统每天会产生大量结构化的数据，这些数据都存储在业务系统所对应的数据库中，包括 MySQL、Oracle、RDS 等类型。

**数据收集与存储**

若想利用 MaxCompute 的海量数据存储与处理能力来分析这些已有的数据，首先需要将不同业务系统的数据同步至 MaxCompute 中。

DataWorks 提供数据集成服务，可支持多种数据源类型将业务系统数据按照预设的调度周期同步到

MaxCompute。

**数据分析与处理**

随之可对 MaxCompute 上的数据进行加工（MaxCompute SQL、MaxCompute MR）、分析与挖掘（数据分析、数据挖掘）等处理，从而发现其价值。

**数据提取**

分析与处理后的结果数据，需同步导出至业务系统，以供业务人员使用其分析的价值。

**数据展现和分享**

最后可通过报表、地理信息系统等多种展现方式来展示与分享大数据分析、处理后的成果。