阿里云
通用解决方案

迁移解决方案

文档版本：20190723

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。
2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。
3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。
4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的”现状“、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。
5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含”阿里云”、Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。
6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	 **禁止:** 重置操作将丢失用户配置数据。
	该类警示信息可能导致系统重大变更甚至故障，或者导致人身伤害等结果。	 **警告:** 重启操作将导致业务中断，恢复业务所需时间约10分钟。
	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	 **说明:** 您也可以通过按Ctrl + A选中全部文件。
>	多级菜单递进。	设置 > 网络 > 设置网络类型
粗体	表示按键、菜单、页面名称等UI元素。	单击 确定。
`courier` 字体	命令。	执行 `cd /d C:/windows` 命令，进入Windows系统文件夹。
*## *	表示参数、变量。	`bae log list --instanceid` *`Instance_ID`*
[]或者[a|b]	表示可选项，至多选择一个。	`ipconfig` *`[-all|-t]`*
{}或者{a|b}	表示必选项，至多选择一个。	`swich` *`{stand | slave}`*

目录

1 迁移背景信息

随着大数据、云计算的到来并逐渐普及，很多企业从要不要上云的转为关注业务系统和数据如何上云的问题。阿里云针对不同规模不同类型的企业，提供丰富的迁移解决方案，满足各种迁移目的的需求。

各个企业的迁移目的，迁移的需求场景均不同。

迁移目的分类

从各个企业迁移上云的目的来看，迁移的场景一般有以下几种：

- 提高高可用性：

  企业的生产环境仍在本地机房，通过不同的迁移策略将生产环境的业务数据、业务系统备份迁移至云上，作为本地机房的容灾备份，以防出现故障时，可将云上数据迅速恢复至本地，或将业务直接切换至云上，提高整个系统的高可用性。
- 节约成本：

  企业将生产或测试环境直接部署于云上，借助云上按量使用、弹性伸缩的特点，且企业无需投入构建机房、服务器等硬件设备，节约整体的业务建设成本，并保障业务流量高峰期时资源可迅速扩容。
- 提效优化：

  企业借鉴阿里云的“中台”概念，希望通过迁移上云来整改优化已经老旧的业务系统，最终通过迁移并重新优化业务、IT架构来激活企业的创新，打开企业的新局面。

源系统类型分类

从企业迁移上云的源头来看，迁移的场景可分为：

- P2V：即从本地机房的服务器、数据库服务器等硬件中将应用系统或数据迁移至阿里云上。
- V2V：即从虚拟环境或云厂商中将应用系统或数据迁移至阿里云上，包括：
  - 企业自建的虚拟化环境：例如VMware、KVM、Ren等虚拟化环境。
  - 第三方云厂商环境：例如AWS、Azure、Google Cloud等。
  - 阿里云不同地域间的迁移：例如从华北地域将系统或数据迁移至华东地域。

迁移方案需求分类

从企业迁移上云的方案需求来看，迁移的场景可分为：

- 平滑迁移：企业的迁移方案要求应用、组织不变，迁移上云后业务体验一致，是完全源系统、数据的云上复制。

- 优化改造：企业的迁移方案需要结合企业新型业务需求，对上云后的业务、IT系统均进行不同程度的改造优化，云上系统与源系统不完全一致。

迁移难点

面对繁复的迁移场景，企业在制定迁移方案并实施迁移上云时，通常有以下难点：

- 传统企业IT人员较少且对云产品不熟悉，不完全具备设计迁移方案、实施迁移任务的能力。
- 安全性要求较高的行业，例如金融、政企等，迁移上云的安全性、合规性难以自主评估并保障。
- 企业原先业务、IT架构太过复杂，迁移方案难以自主制定。

阿里云针对不同规模不同类型的企业，提供丰富的迁移解决方案及便捷的迁移工具，满足各种迁移目的的需求。详细的方案介绍请参考 阿里云迁移解决方案 章节。

2 阿里云迁移解决方案

2.1 阿里云迁移服务

阿里云为满足不同企业的多种迁移上云需求，提供便捷的迁移工具及专属的迁移服务，提供一站式迁移解决方案。

阿里云自成立之初即成立迁移团队，为各企业供专业的迁移服务，包括迁移咨询服务、迁移实施服务、云化咨询服务。

- 迁移咨询服务：阿里云迁移团队结合多年迁移实战经验和企业业务、IT系统现状，针对企业的迁移目的可提供专业的迁移咨询服务，为企业迁移上云提供靠谱的方案建议与问题解决方案。例如评估业务系统迁移阿里云平台的可行性，设计业务系统的产品选型和应用架构，以及应用系统、存储、数据库等迁移方案。
- 迁移实施服务：阿里云迁移团队可为企业提供实施迁移任务的服务，助力企业高效、安全的完成上云迁移。例如通过技术支持或协助实现客户的在线业务系统、数据库及存储等内容迁移到阿里云，并顺利完成业务系统的割接。
- 云化咨询服务：阿里云迁移团队结合企业的战略目标，为企业三五年甚至更久的云化规划提供战略咨询服务，以专业的云专家助力企业在云计算时代进一步创新发展。例如为IT系统运行在阿里云的客户，提供云计算与新技术应用规划、架构、容器及微服务设计等全方位咨询、架构最佳实践指导。

企业可联系阿里云客户经理、地域经理或直接在线申请，以获取专属的迁移服务。

2.2 阿里云迁移常用工具

阿里云提供诸多便捷的迁移工具，帮助企业完成迁移任务。

迁移工具包括但不限于：

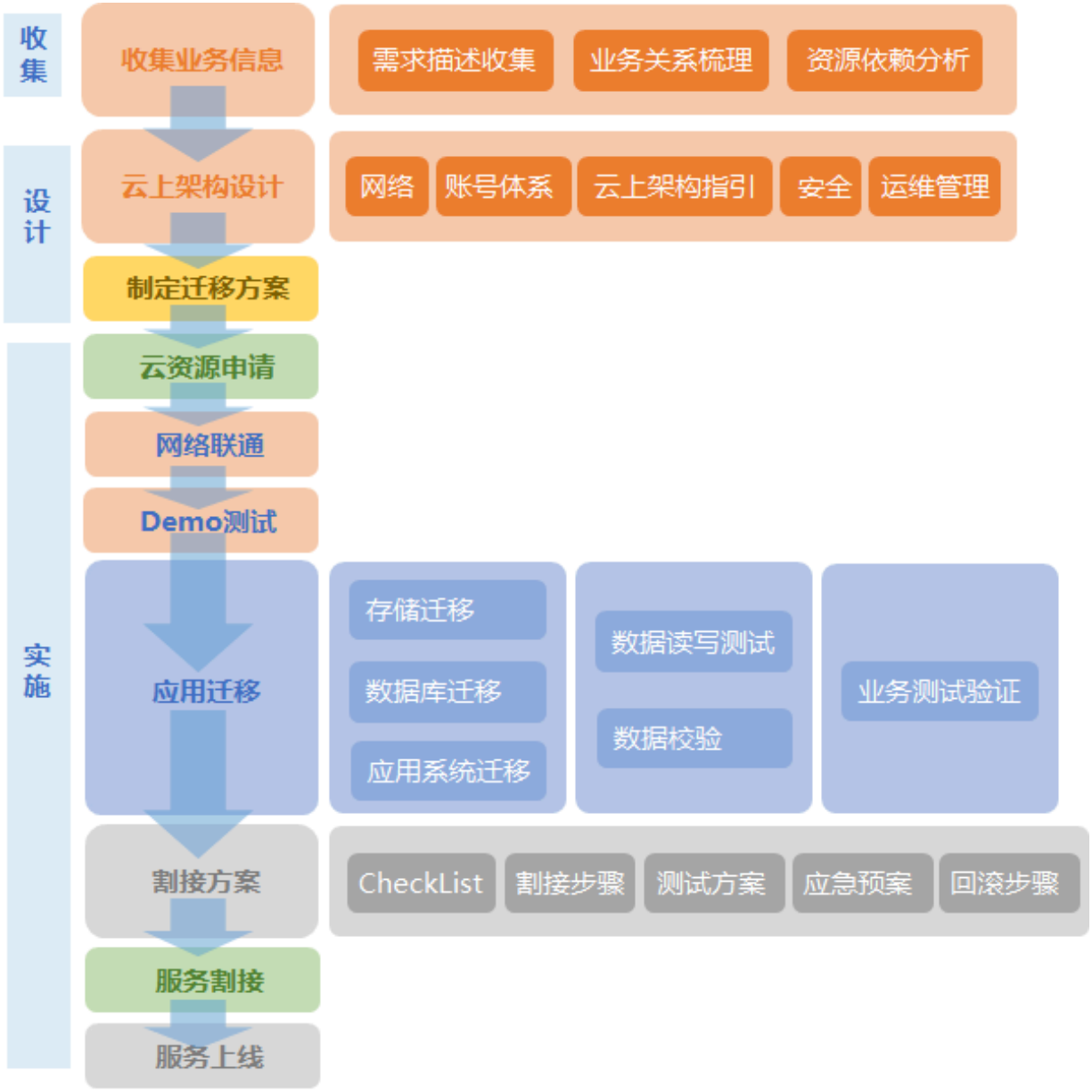
迁移类型	迁移说明	工具
应用系统迁移	将业务系统迁移至阿里云ECS	· 迁移工具 · ADAM
存储迁移	将业务系统迁移至阿里云OSS	· 在线迁移服务 · 闪电立方

迁移类型	迁移说明	工具
数据库迁移	将业务系统迁移至阿里云RDS	· 数据传输 DTS · Rump · redis-port · ADAM
大数据迁移	将业务系统迁移至阿里云MaxCompute	DataWorks

企业可根据迁移类型和方案选择合适的迁移工具，自主或联系阿里云迁移团队完成迁移任务。

2.3 迁移推荐流程

阿里云迁移解决方案建议按照分析、设计、实施、优化这四个步骤进行迁移任务。



1. 分析：对应用系统、存储、数据库等进行专业评估，明确迁移收益和投入风险，制定初步业务迁移规划。
2. 设计：确定合适的迁移策略，结合业务需求设计云上架构，设计并验证典型应用的平滑高效迁移。
3. 实施：搭建云上基础设施方案：云产品配置、网络接入、安全策略配置等，并批量迁移数据和应用系统。
4. 优化：借助阿里云的平台和工具来管理应用程序，并根据业务需求和管理数据持续改进、优化架构。

3 迁移案例 典型架构平滑上云

3.1 迁移需求与流程

以下以一个传统企业的典型IT架构平滑迁移的迁移场景，示例介绍阿里云迁移解决方案。根据客户迁移的需求，并结合客户的业务、IT系统现状，制定迁移的流程。

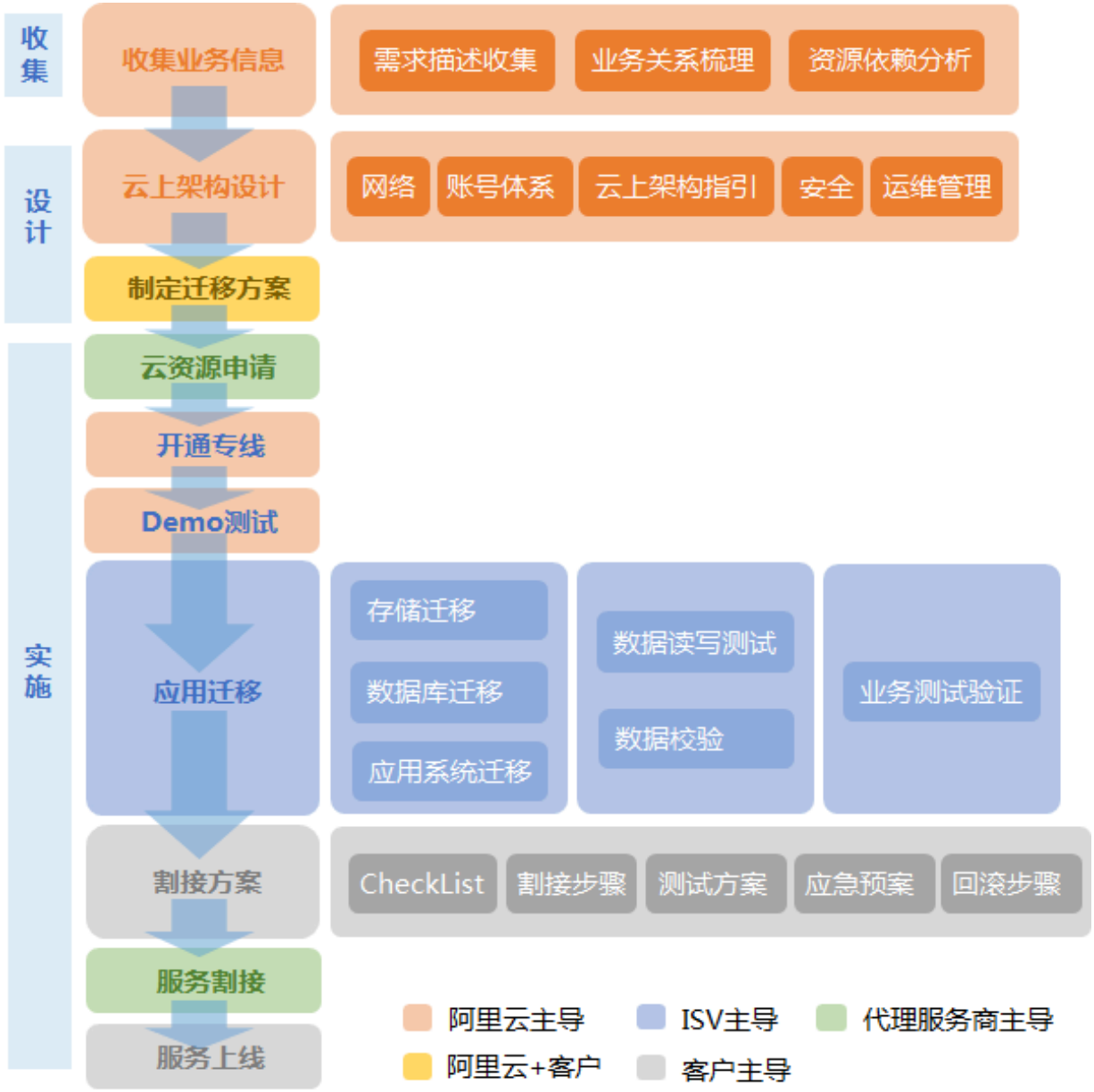
针对传统企业将IT系统迁移上云的场景，阿里迁移团队结合客户的迁移诉求，为客户拟定迁移实施方案、云上网络架构、云上安全架构、业务账号体系方案、业务访问方案、运维访问方案，并与客户一起实施完成完整的迁移任务。

迁移需求

本方案的迁移基于并满足以下的客户迁移需求：

- 应用不改
- 组织不变
- 体验一致

迁移流程



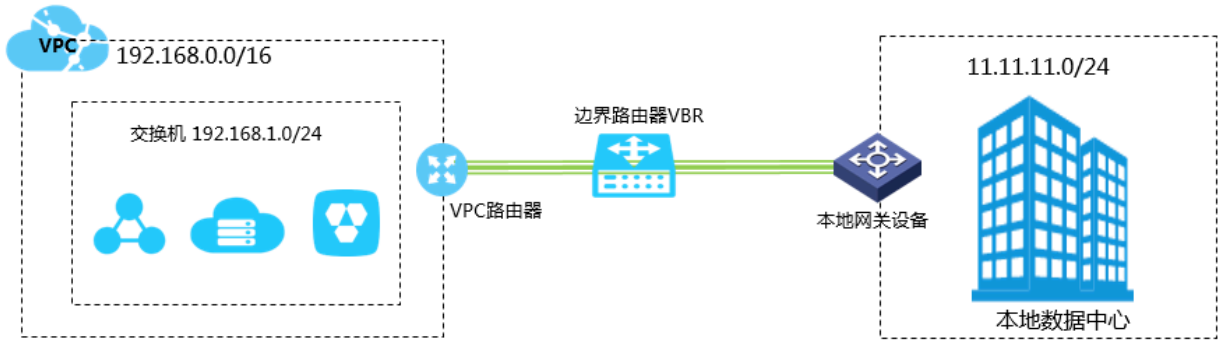
1. 分析：对应用系统、存储、数据库等进行专业评估，明确迁移收益和投入风险，制定初步业务迁移规划。
2. 设计：确定合适的迁移策略，结合业务需求设计云上架构，设计并验证典型应用的平滑高效迁移。本示例中设计完成上云后，网络架构、安全架构等参见方案设计章节。
3. 实施：搭建云上基础设施方案：云产品配置、网络接入、安全策略配置等，并批量迁移数据和应用系统。
4. 优化：借助阿里云的平台和工具来管理应用程序，并根据业务需求和管理数据持续改进、优化架构。

3.2 方案设计

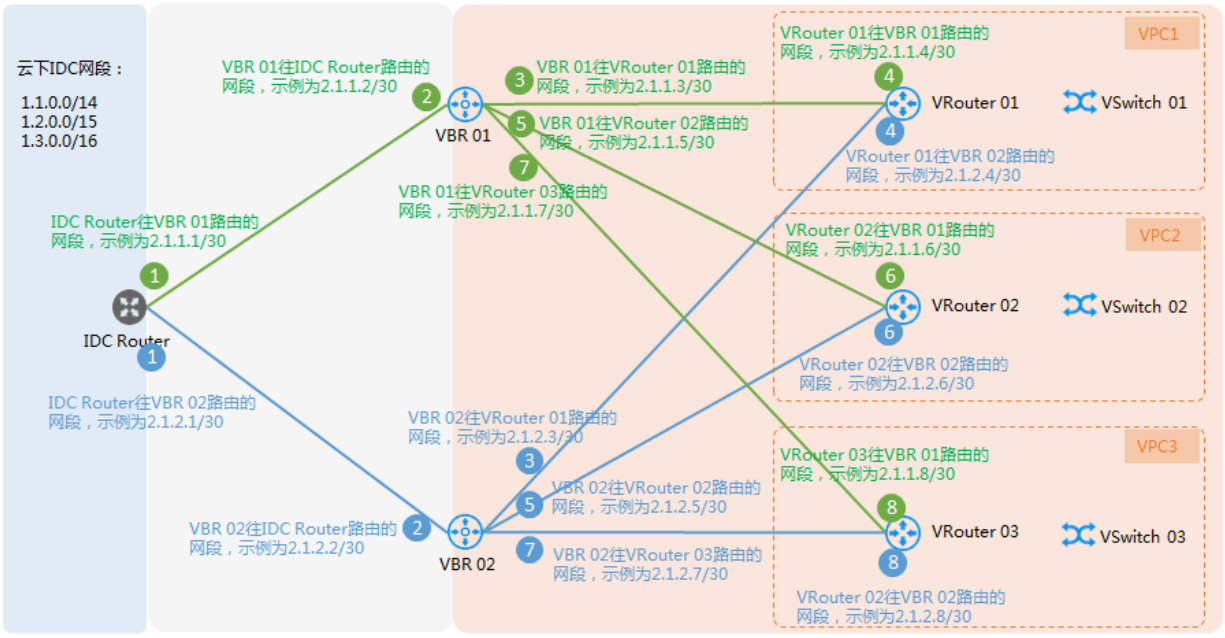
3.2.1 云上网络架构方案

迁移上云后，客户机房的线下业务与云上业务构成混合云架构，云上云下可通过专线或VPN等方式联通网络。

本示例以使用物理专线联通云上云下，构成混合云架构为例，如下图所示。



其中，云下IDC通过专线与云上网络联通后，需根据新的网络架构再次规划您的业务网段，以下网段为样例示意，需根据实际的业务重新规划划分。



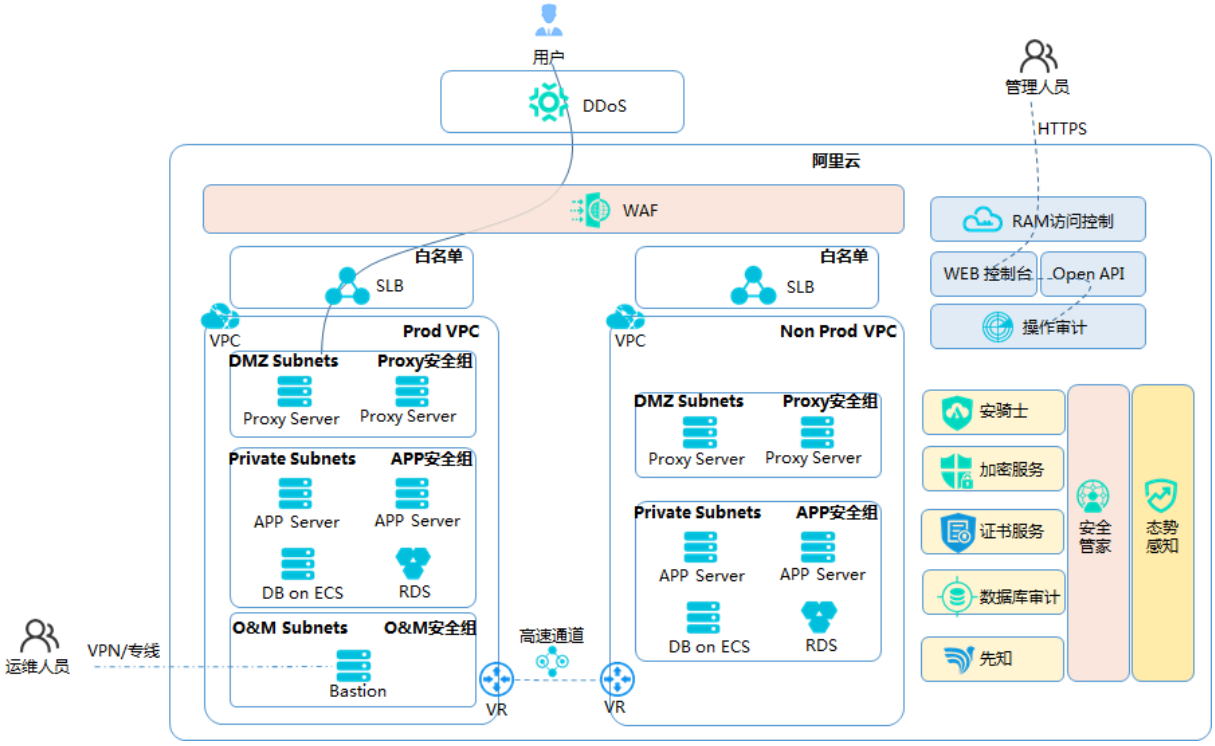
架构说明：

- 安通融合线上线下业务时，可以使用物理专线或VPN服务，构成混合云组网。
- 使用物理专线是，可采用BGP双线高速通道，保障线上线下的专线高可用。
- 云上可划分多个VPC，将不同业务网络隔离开，例如生产主备环境、开发测试环境彼此处于不同VPC。

3.2.2 云上安全架构方案

迁移上云后，通过使用阿里云丰富的安全产品及安全配置，可从“系统-业务-内容”全方位防护，以完善的云端纵深防御，提高整体的安全性。

迁移上云后，使用的安全产品及安全机构如下图所示。



- 抗DDoS攻击：阿里云为应对不同的DDoS攻击场景提供不同防护产品，例如基础防护、防护包、高防IP、游戏盾等，可助力完善各种DDoS攻击场景的安全防护。更多DDoS防护解决方案可参考 DDoS防护解决方案 章节。
- Web应用防护：阿里云提供Web应用防火墙（WAF），网站所有的公网流量都会先经过WAF，恶意攻击流量在WAF上被检测过滤，而正常流量返回给源站IP，从而确保源站IP安全、稳定、可用。
- 白名单、安全组：阿里云的ECS等产品提供白名单、安全组的安全加固配置，尽量避免将非业务必须的服务端口暴露在公网上，缩小暴露面，隔离资源和不相关的业务，降低被攻击的风险。
- 专有网络隔离：内部通过专有网络（VPC）实现网络内部逻辑隔离，防止来自内网肉鸡的攻击。
- 授权与访问控制：阿里云提供访问控制（RAM）功能，当存在多用户协同操作资源时，使用RAM 可以让您避免与其他用户共享云账号密钥，按需为用户分配最小权限，从而降低您的企业信息安全风险。
- 审计日志：阿里云提供操作审计（ActionTrail）的功能，并可将审计记录以日志的形式存储在制定路径中。通过ActionTrail保存的所有操作记录，您可以实现安全分析、资源变更追踪以及合规性审计。

- 威胁检测：阿里云提供态势感知服务用于云上安全监控和诊断服务，面向云上资产提供安全事件检测、漏洞扫描、基线配置核查等服务。
- 安全管家：阿里云安全管家服务是阿里云安全专家基于阿里云多年安全最佳实践经验为云上用户提供的全方位安全技术和咨询服务，为云上用户建立和持续优化云安全防御体系，保障用户业务安全。
- 其他安全防护：通过安骑士、证书服务等多维安全产品，全方位提高主机、业务的安全性。

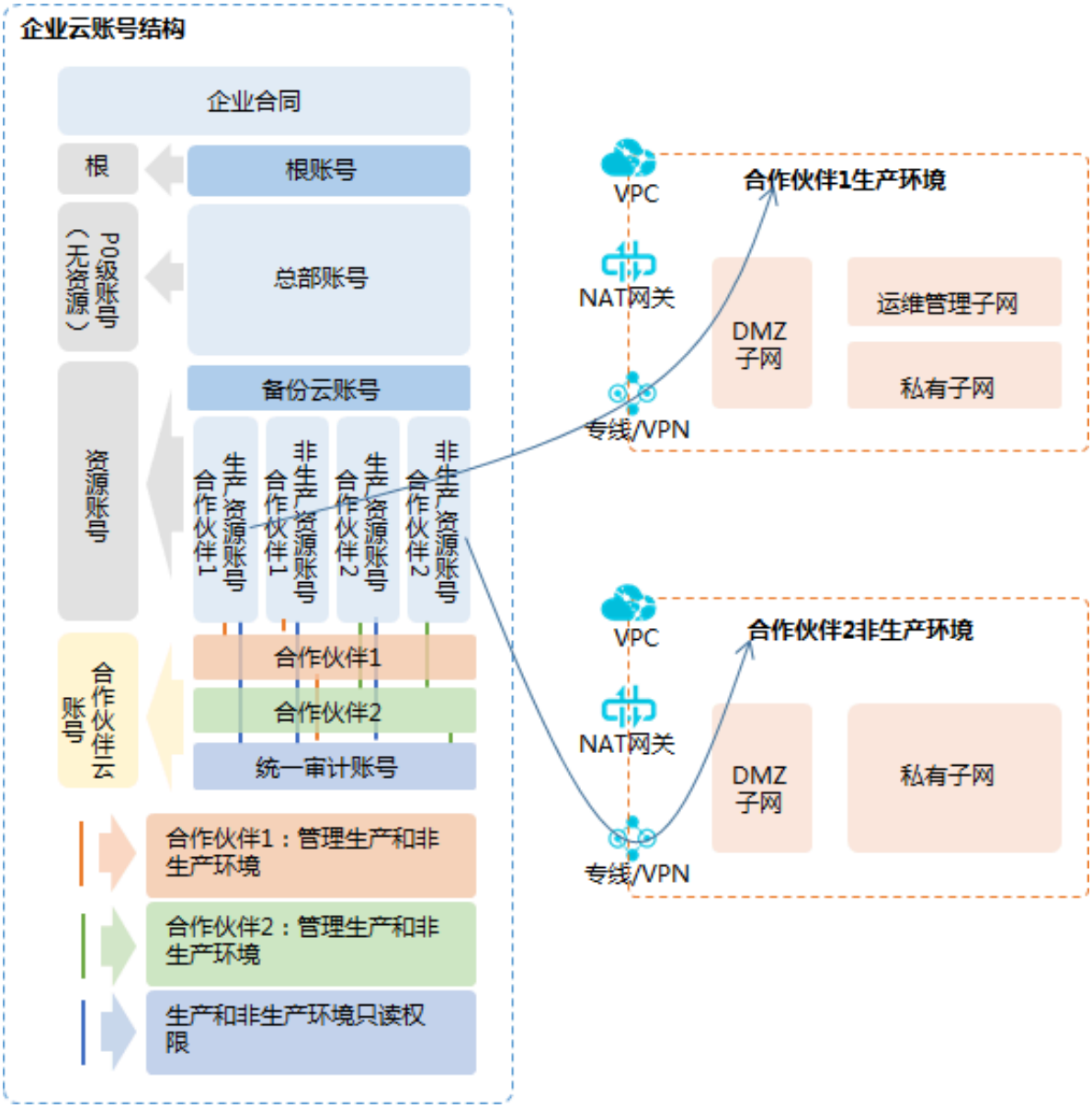
3.2.3 业务账号体系方案

迁移上云后，面对客户内部多个部门、不同和合作伙伴需要设计安全完善的账号体系。

本示例中，迁移上云后客户对账号体系的需求入下：

- 内部有多个部门，各部门之间账号彼此独立且权限最小分配。
- 外部大量的供应商与合作伙伴，各合作伙伴需拥有独立的子账号。
- 审计账号必须只读。

结合用户账号体系要求，上云后的账号体系方案如下图所示。



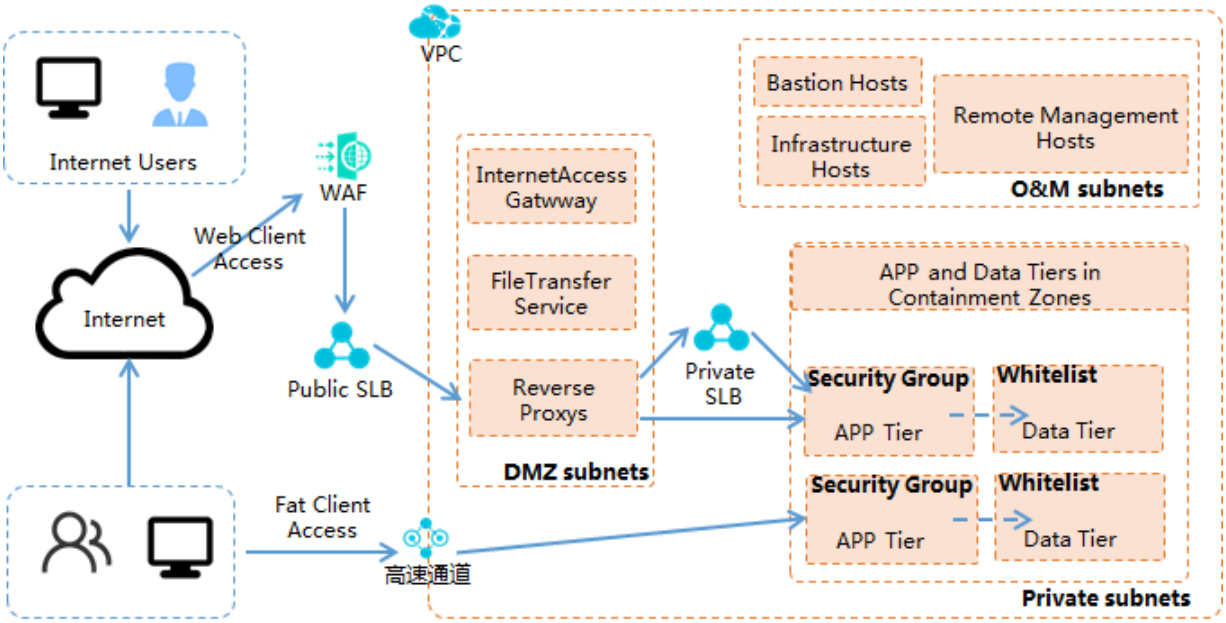
架构说明：

- **每个部门合作伙伴都有独立的子账号。**
- **资源账号：管理云资源，合作伙伴使用。**
- **备份账号：备份数据存储使用账号。**
- **审计账号：赋予只读权限。**

3.2.4 业务访问方案

迁移上云后，通过将业务区域逻辑上隔离划分为多个“安全域”，通过DMZ区连接内外，提高整体业务访问的安全性。

详细的业务访问方案如下图所示。



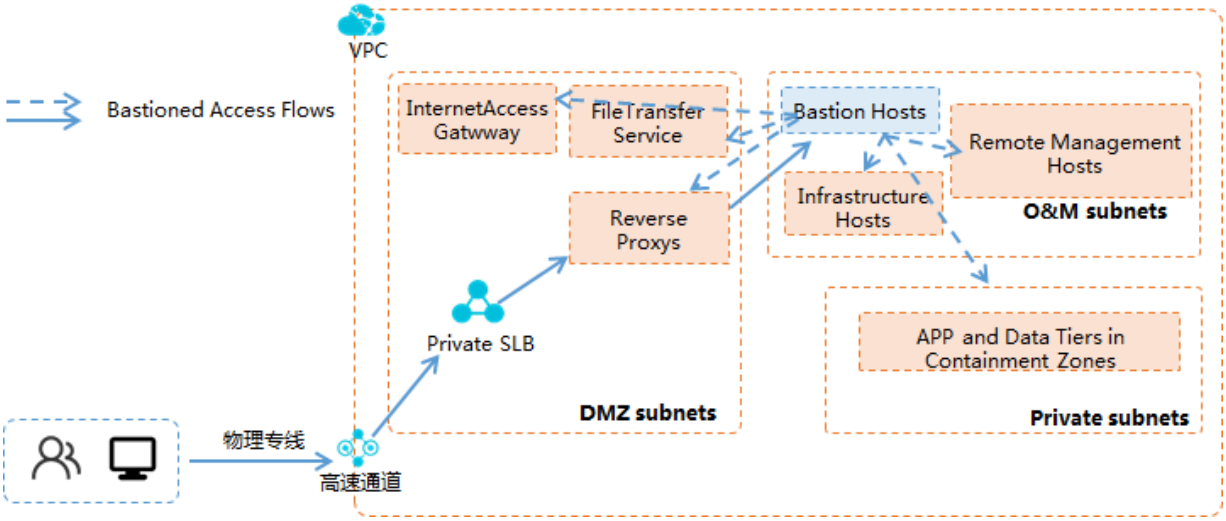
访问路径说明：

- **在访问源与内部应用直接划分DMZ区，作为互联网访问的接入区域，隔离互联网访问与内部应用，提高整体的安全性。**
- **针对不同的访问源，设置不同的访问路径：**
  - **Web类应用：互联网 > 公网SLB > DMZ代理服务器 > 内网SLB > 内部应用**
  - **胖客户端类应用：内网 > VPN/专线 > 内部应用**

3.2.5 运维访问方案

与业务访问方案相似，运维访问方案中，也划分“DMZ”安全域，隔离内外系统，以提高整体的安全性。

详细的运维访问方案如下图所示。



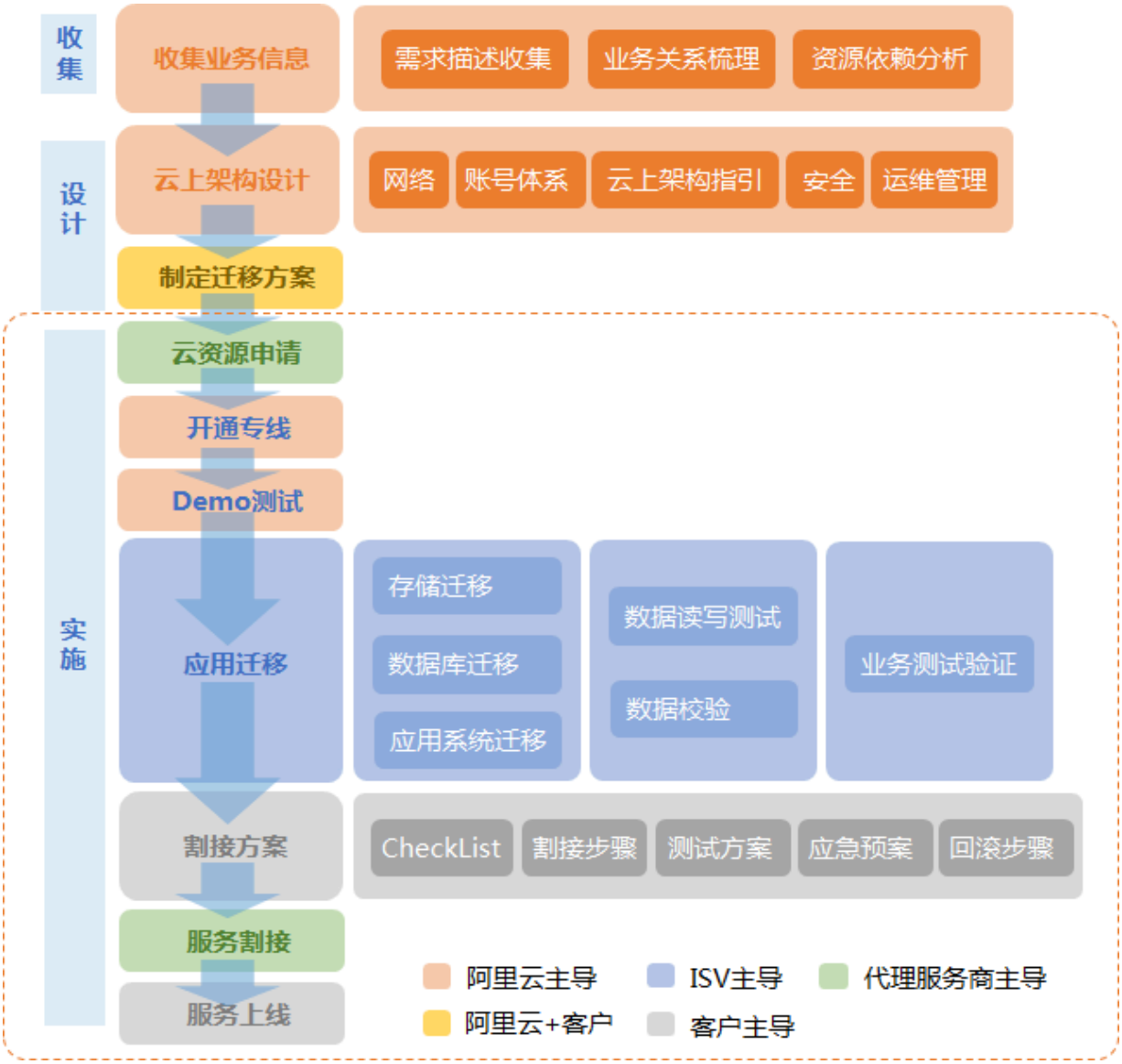
路径说明：

- 在运维访问源与内部运维系统、内部应用系统间设置DMZ区，隔离内外系统。
- 运维系统入口通过堡垒机统一收口，由堡垒机作为所有运维访问的第一跳，通过堡垒机再转至运维系统的其他子系统，或转至内部应用系统。
- 运维的路径顺序为：物理专线 > DMZ > 公网SLB > 反向代理 > 内网堡垒机 > 内部云资源。

3.3 迁移实施

完成云上架构设计后，需结合当前现状及云上方案制定迁移方案，并根据迁移方案实施迁移操作。

本示例中，整个迁移实施包含：云资源申请、开通专线、Demo测试、应用迁移、割接方案、服务割接、服务上线等步骤。其中各迁移步骤分别由阿里云、客户、代理服务商、ISV分别主导实施，分工如下图所示。



您使用阿里云迁移咨询服务和阿里云迁移实施服务，根据您的迁移场景定制您的迁移流程及实施步骤。

其中关于应用系统迁移、文件存储迁移、数据库迁移、大数据迁移可参考本文档借鉴。

4 迁移网络改造

4.1 网络改造概述

迁移上云的过程中，需要根据迁移后的网络架构进行网络的迁移改造。

根据迁移前的网络架构现状及上云后网络架构的方案不同，迁移中的网络迁移改造流程、操作也不一致。

- 网络架构方案如果是混合云组网，您需要通过专线、VPN等方式打通云上云下网络环境，并根据网络架构方案中的网段规划配置好网段。
- 迁移上阿里云后如果您的网络环境均为专有网络，建议您先根据云上网络架构方案，创建好VPC网络环境及交换机配置等操作，后续云产品的开通直接根据架构方案在各对应的VPC中创建。
- 如果您是从经典网络将环境迁移至专有网络，网络改造可参考迁经典网络迁移至专有网络进行迁移。
- 如果您在网络改造过程中，需同步将SNAT网关变更为使用云上SNAT网关，可参考自建SNAT网关平滑迁移到NAT网关章节进行迁移操作。

4.2 迁移方案概述

您可以将部署在经典网络中的资源迁移到专有网络中。专有网络是隔离的网络环境，安全性更高。

为什么要迁移至VPC?

专有网络VPC（Virtual Private Cloud）是您自己独有的的云上私有网络。您可以在自己定义的VPC中使用阿里云资源。VPC有如下优势：

- 安全的网络环境

  VPC基于隧道技术，实现数据链路层的隔离，为每个租户提供一张独立、隔离的安全网络。不同专有网络之间网络完全隔离。

- 可控的网络配置

  您可以完全掌控自己的虚拟网络，例如选择自己的IP地址范围、配置路由表和网关等，从而可以轻松地实现内网的网络资源规划以及路由表的路径选择。此外，您也可以通过专线或VPN等连接方式将您的专有网络与传统数据中心相连，形成一个按需定制的网络环境，实现应用的平滑迁移上云和对数据中心的扩展。

如何迁移?

目前，阿里云提供以下两种将经典网络迁移到VPC的方案。这些方案可以独立使用，也可以组合使用，以满足不同的迁移场景：

- 混访混挂方案

  如果您的服务依赖RDS、SLB等云产品，建议您选择混访混挂的迁移方案。该方案可以平滑地将系统迁移至专有网络环境中，保证服务的稳定性。

  搭配使用ClassicLink功能，以满足未迁移的经典网络ECS实例访问VPC中云资源的需求。详情参见ClassicLink概述。

- 单ECS迁移方案

  如果您的应用部署在了ECS实例上，且ECS实例重启对系统没有影响，可以选择单ECS迁移方案。

混挂和混访方案

混挂和混访方案是一种系统平滑迁移方案，即用户通过在VPC中新建ECS等云产品实例，然后将系统平滑迁移到VPC。当所有系统都迁移到VPC后，再将经典网络内的资源释放，从而完成经典网络到VPC的迁移。详情参见混访混挂迁移示例。

- 混挂

  混挂指一个负载均衡实例可以同时添加经典网络和VPC网络的ECS作为后端服务器接收监听转发的请求，且支持虚拟服务器组形式的混挂。

  公网负载均衡实例和私网负载均衡实例都可开通混挂。

  > 说明:
  > VPC私网负载均衡实例同时挂载经典网络和专有网络ECS时，如果使用四层（TCP和UDP协议）监听，目前无法在经典网络ECS上获取客户端的真实IP，但在专有网络ECS上还可以正常获取客户端的真实IP。对七层监听（HTTP和HTTPS协议）没有影响，可以正常获取客户端的真实IP。

- 混访

  云数据库RDS和对象存储OSS等云产品支持混访，即支持同时被经典网络和专有网络中的ECS访问。通常该类产品都提供两个访问域名，一个是经典网络访问域名，另外一个是专有网络访问域名。

在使用本方案时，请注意：

- 本方案可以满足绝大部分系统的迁移要求。但如果系统中的专有网络ECS和经典网络ECS有内网通信的需求，可通过ClassicLink功能实现。
- 本方案仅用于经典网络迁移到VPC。

单ECS迁移方案

单ECS迁移方案，即无需通过创建镜像、重新购买等步骤就能把经典网络的ECS实例迁移到专有网络。

在控制台上完成迁移预约后，阿里云会根据您设置的迁移时间进行迁移，迁移完成后，您将收到迁移成功的短信消息提醒。

在使用单ECS迁移方案时，注意：

- 迁移过程中ECS需要进行重启，请关注对系统的影响。
- 迁移后，不需要进行任何特殊配置，ECS实例的公网IP都不变。
  - 虽然公网IP没有变化，但无法在ECS的操作系统中查看到这个公网IP（称之为VPC类型的ECS的固定公网IP）。您可以将按流量计费的ECS实例的固定公网IP转换为EIP，方便管理，详情参见ECS固定公网IP转换为EIP。
  - 如果您的个别应用对ECS操作系统上可见的公网IP有依赖，迁移后会有影响，请谨慎评估。
- 迁移后，所有地域的ECS实例的私网IP都会变化。
- 迁移到的目标VPC的交换机的可用区必须和待迁移的ECS的可用区相同。
- 迁移过程中实例ID及登录信息不变。
- 包年包月购买方式的实例迁移过程中不需要额外付费。从新的计费周期开始，按照同规格专有网络的价格计算。且迁移到VPC后，ECS的使用费用会降低。
- 迁移前如有续费变配未生效订单或未支付订单，迁移后该订单将被取消且不能恢复，您需要重新下单。
- 迁移到VPC后，若ECS有使用其它云服务，需将访问方式调整到VPC访问方式(云产品混访方案)。

4.3 自建SNAT网关平滑迁移到NAT网关

通过使用路由表的最长匹配原则，您可以将搭建在ECS实例的SNAT网关平滑迁移至阿里云NAT网关。

背景信息

如果您已经在VPC中基于ECS搭建了SNAT网关，又想将架构切换为基于NAT网关实现的SNAT，您可以将原有自建SNAT网关拆除，再进行NAT网关的创建和配置。但该操作会导致SNAT功能中断一段时间。

本教程的迁移方法利用路由表的一些特性（主要是“最长匹配原则”），实现从自建SNAT网关到阿里云NAT网关的无缝切换。切换过程中，不会出现SNAT功能不可用，仅在切换的一瞬间发生已有TCP连接的断开，应用进行重连即可。

本操作中作为示例的VPC和ECS配置如下：

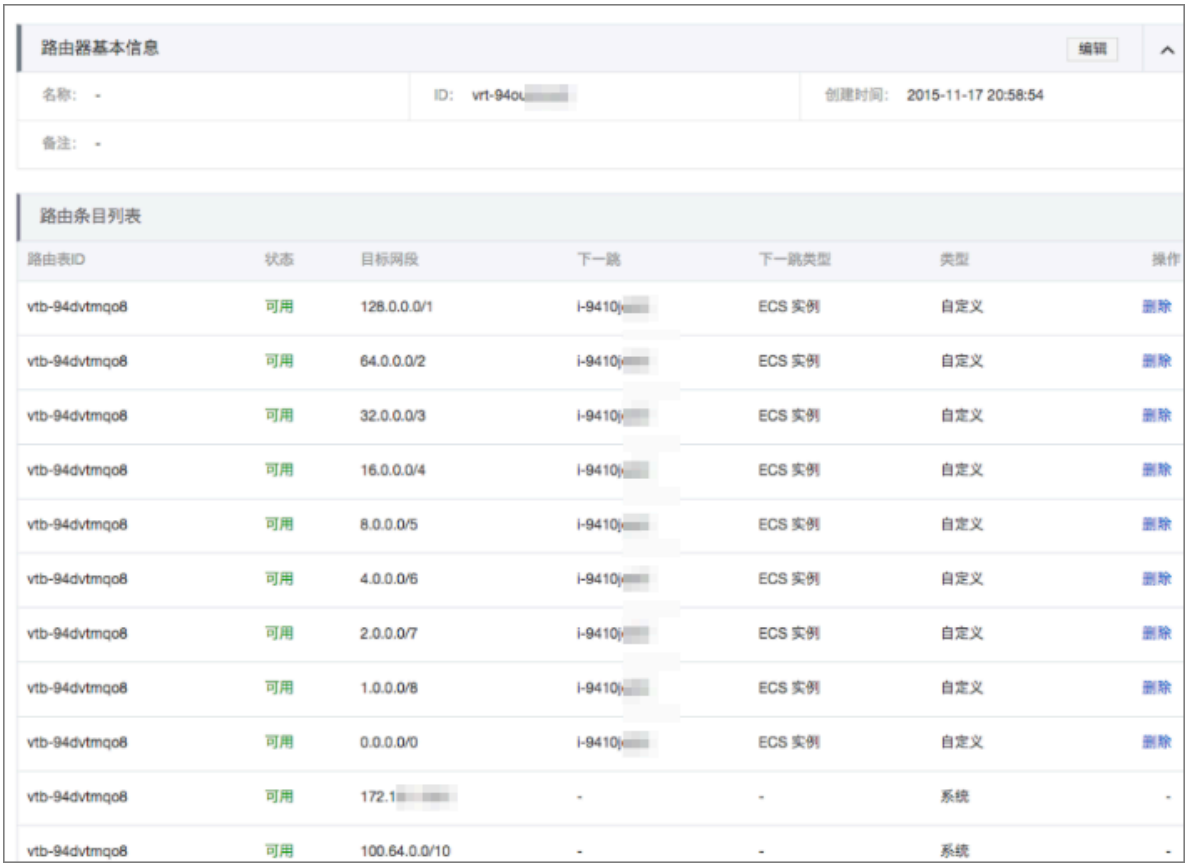
- VPC中有两个ECS实例：
  - i-9410jxxxx配置了自建的SNAT网关。这台ECS上绑定了一个EIP，并且开启了转发服务、配置了iptables规则以实现SNAT转发。
  - i-94kjwxxxx为需要SNAT功能来访问互联网的服务器。
- VPC的路由器上，添加了一条自定义路由（目标网段为0.0.0.0/0），将公网访问请求转发给i-9410jxxxx。

操作步骤

1. 在VPC中添加8条路由条目，对原有路由进行覆盖。

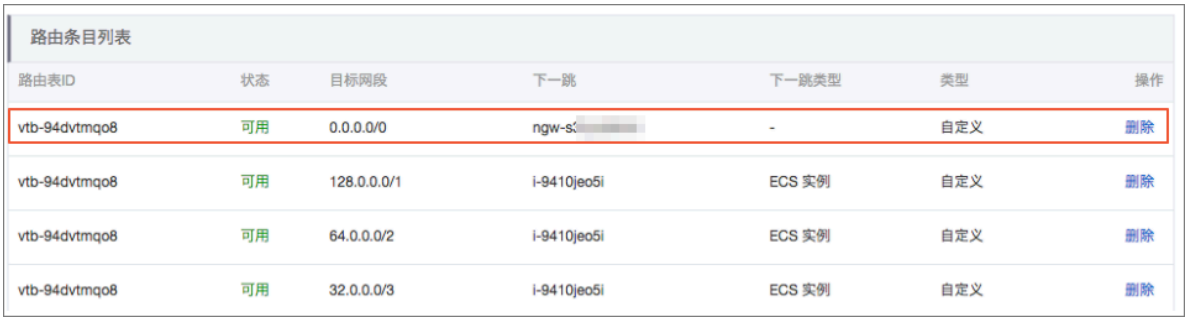
   路由条目的目标网段分别为1.0.0.0/8、2.0.0.0/7、4.0.0.0/6、8.0.0.0/5、16.0.0.0/4、32.0.0.0/3、64.0.0.0/2、128.0.0.0/1，下一跳均为i-9410jxxxx。

   由于路由表按照最长匹配原则，会优先匹配子网掩码最长的路由条目；而去往任意IP地址的数据包，都会匹配到这8条中的一条；因此，0.0.0.0/0这条路由实际上已经不再有用了。



2. 删除目标网段为0.0.0.0/0的路由条目。
3. 创建NAT网关。

   创建NAT网关后，系统会自动添加一条0.0.0.0/0的路由，指向NAT网关。



4. 绑定弹性公网IP。

   注意:

   确保EIP的带宽和自建NAT的带宽一致。因为只要在NAT网关添加了SNAT规则，SNAT规则中的ECS的出公网方向的流量就会受EIP带宽的限速。

5. 配置SNAT规则。
6. 删除VPC中添加的8条路由路由条目，使路由器把公网访问请求不再转发给自建SNAT，而是转发给NAT网关。

   至此，已经完成了从自建SNAT网关到使用官方NAT网关的SNAT功能的全部切换流程。

5 应用系统迁移

5.1 迁云概述

本文档围绕如何将您的服务迁移到阿里云，提供了多个场景的迁云方案。

背景信息

在云计算服务高速发展的今天，如何方便快捷地将已有的服务器系统迁移上云，有着非常重要的意义。阿里云服务器迁移服务方案，即迁云服务，正是在这个需求背景下应运而生。它极大地简化了服务器系统迁移工具的使用条件、降低了使用成本，使用户的系统一键迁移到阿里云成为可能。

使用迁云服务来进行系统迁移比较便捷，您可以先参考迁云工具帮助文档了解使用条件及基本操作。

迁云流程

1. 熟悉迁云工具使用方法，提前做好测试演练。
2. 评估迁移时间/成本，制定迁移计划。
3. 正式迁移，可咨询阿里云团队支持。
4. 创建按量实例，进行系统业务联调。
5. 切换到云端系统，将实例升级为包年包月。

5.2 应用场景

不同的操作系统、源平台和迁移规模，有着不同的迁移方式。本文介绍阿里云迁云服务支持的场景。

支持的主流操作系统

主流服务器系统迁云痛点难点：

- 操作配置麻烦：需要虚拟化驱动配置、系统引导配置、磁盘配置等；专业知识基础要求高
- 迁移成本高：需要等量本地存储空间做中转，迁移周期长、易中断
- 迁移方式不统一：主流Windows/Linux操作系统种类繁多，没有统一的方式兼容各种系统版本

针对上述问题，迁云服务能够提供：

一键自动处理系统配置，不需要占用本地存储空间，支持断点续传、压缩传输，提供统一的迁移操作流程方式，并支持以下主流Windows/Linux操作系统：

- Windows Server 200/2008/2012/ 2016

- CentOS 5/6/7
- Ubuntu 10/12/14/16/17/18
- Debian 7/8/9
- Red Hat 5/6/7
- SUSE 11.4/12.1/12.2
- Amazon Linux 2014及以上
- Oracle Linux 5/6/7

支持的主流服务器平台

主流服务器平台迁云痛点难点：

- 各平台底层环境不兼容：物理机、虚拟机
- 各平台文件系统不兼容：文件格式、分区类型、磁盘类型
- 各平台系统服务不兼容：Cloud-Init、SELINUX服务等

针对上述问题，迁云服务能够提供：

它不依赖底层环境，支持P2V/V2V（物理机/虚拟机迁移），兼容多文件格式、多分区类型、多磁盘类型，也支持自动处理指定系统服务，达到兼容各个平台的目的，目前支持以下主流平台服务器迁移：

- 自建IDC机房
- 本地虚拟机（VMware/Vitrual Box/XEN/KVM）
- 阿里云（不同账号或地域之间）
- AWS EC2
- AZURE VM
- GOOGLE VM
- HUAWEI ECS
- 腾讯云 CVM
- 其他主流厂商云（例如UCloud、电信云、青云等）

支持的迁云规模

如果迁云工作无法自动化，人力、物力和时间成本都会随着迁移数量的增加而成指数倍的增长。主要困难有以下几点：

- 需要大批量操作和部署准备
- 需要大批量迁移资源消耗
- 需要大量迁移周期

针对上述问题，迁云服务能够提供：

- 利用阿里云平台弹性计算资源的优势，能自动按需申请迁移资源进行迁移规模弹性扩充
- 支持大批量迁移任务并发进行
- 迁云服务工具本身体量小便于批量分发，支持命令行调用，客户只需要编写简单的自动化脚本配合迁云工具即可支持大批量迁移

迁云服务支持以下规模的迁移：

- 1-10 微小规模数量迁移
- 10-100 中小规模数量迁移
- 100-500 中大规模数量迁移

5.3 评估设计

在开始迁云之前，您需要先评估业务内容、迁移时间和成本等因素，制定迁移计划。本文介绍需要评估的几大因素。

操作系统

- 内核版本：要求CentOS/RedHat 5及以上、Ubuntu 10及以上、Windows Server 2003及以上等。对于低版本系统内核，需要先升级内核。
- 虚拟化驱动：必须安装KVM virtio驱动。
- 需要服务软件：Linux系统必需安装rsync，建议安装curl；Windows系统需确保VSS（Volume Shadow Services）服务正常。
- GRUB引导程序：部分低内核系统如CentOS/Red Hat 5、Debian 7需要升级GRUB至1.99及以上。
- 磁盘大小：系统盘40-500GiB；数据盘20-32768GiB。

应用业务

- 业务暂停问题：如果有数据库等大型服务应用，如Oracle、SQLServer、MongoDB、MySQL和Docker，可以考虑暂停服务应用迁移。如果不能暂停业务，迁移时可以先将服务应用数据目录排除，待服务器迁移完成后，再同步数据库的数据。
- 大数据量问题：如果有大量或海量数据文件，可以先使用迁云服务只迁移服务器应用环境，同时评估是否需要使用专线或闪电立方等专用大数据量传输方案来迁移以获得更好的传输速度。
- 软件授权问题：评估源系统需要授权的软件在迁移后是否需要重新授权。
- 网络配置问题：迁移后公网IP会发生变化，需评估是否会影响原业务。

网络模式

您需要评估待迁移的服务器系统所需网络传输模式。

迁移服务器分为3个阶段：

1. 迁移资源准备
2. 数据传输
3. 迁移收尾

其中，1、2、3阶段都默认使用公网，默认情况下需要您的待迁移服务器能够访问以下阿里云服务地址和端口：

- 阶段1、3
  - ECS服务：https://ecs.aliyuncs.com 443 端口。更多接入地址视区域而定
  - VPC服务：https://vpc.aliyuncs.com 443端口
  - STS服务：https://sts.aliyuncs.com 443端口
- 阶段2：临时中转实例的（默认公网）IP地址，8080 和 8703端口

此外，迁云服务针对您的实际网络环境需求提供了多种网络传输模式：

- 默认公网传输：阶段1、2、3都默认使用公网。
- 手动内网传输：阶段1、3使用公网，阶段2使用VPC内网IP；适合不能访问上述阿里云服务地址、已打通指定VPC内网的服务器系统，但是需要额外需要准备一台可以访问上述阿里云服务地址的同类型系统来配合操作，详情请参见VPC内网迁移。
- 自动内网传输：阶段1、3使用公网，阶段2使用VPC内网。适合能访问上述阿里云服务地址，已经打通指定VPC内网，并且希望数据传输（阶段2）走VPC内网的服务器系统。此模式相较于手动内网传输模式的操作更简单，详情请参见VPC内网迁移。

迁移数量

如果您需要做批量服务器迁移，还需要注意以下问题：

1. 迁移前：
   - 本地网络运营商流量限制，建议与网络运营商协调确认，或者在迁云工具中配置传输带宽上限。
   - 阿里云镜像数量及按量资源（如vCPU）的额度限制，您可以提交工单申请放开限制。
2. 迁移中：
   - 服务器系统是否支持自动化批量运维，来批量下发和运行迁云工具。
   - 是否需要进行批量迁移进度日志统计分析。

3. 迁移后：

- 迁移后系统如何批量创建和配置。
- 迁移后系统批量验证。

迁移周期

迁移周期与迁移服务器数量和实际数据量成正比，建议您根据实际迁移测试演练进行评估。

迁移周期主要分为迁移前、迁移中、迁移后3部分：

- 迁移前时间 = 迁移条件准备时间

  迁移条件准备时间视实际情况而定
- 迁移中时间 = 数据传输时间 + 镜像制作时间（可选）

  数据传输时间 = 实际数据量 / 实际网速

  镜像制作时间 = 实际数据量 / 快照服务速度

  说明：
  迁云工具传输数据时默认是打开了压缩传输选项的，对于传输速度会有30%-40%的提升；镜像制作时间是依赖阿里云快照服务，目前速度在10-30MB/s左右。
- 迁移后时间 = 迁移后系统增量同步时间（可选） + 系统配置验证时间

  系统增量同步时间 = 实际增量数据量 / 实际网速

  系统配置验证时间视实际情况而定

说明：
迁云服务默认迁移结果是生成全量镜像，如果需要尽量缩短迁移周期，也可以选择直接迁移到目标实例，来达到缩短迁移周期的目的，更多支持可以联系迁云服务技术人员。

迁移成本

- 迁云工具是免费工具，不收取额外的费用。但是，在迁云过程中会涉及少量资源计费：
  - 迁云过程中，会创建快照以生成自定义镜像，该快照会按照实际占用容量收取少部分费用。详情请参见快照计费方式。
  - 迁云时，系统默认在您的阿里云账号下创建一个默认名为INSTANCE_FOR_GOTOALIYUN 的ECS实例做中转站。该中转实例付费类型为按量付费，您需要确保账号余额大于等于 100 元。按量付费实例产生的资源耗费及计费说明请参见按量付费。

说明:

迁云完成后，中转实例（包含云盘）资源会自动释放。如果迁云失败，中转实例会保留在ECS控制台，以便于重新迁云。如果您不再继续迁移，需要自行手动释放实例，以免产生不必要的费用。

5.4 迁移实施

5.4.1 迁移场景概览

迁移场景	实施指导
首次从线下IDC或者静态的应用环境中迁移到阿里云时，您需要先将当前所有的数据做一次全量迁移。全量的迁移不需要您停止当前的业务，但迁移过程中的增量数据需要在后续做增量迁移。	全量迁移
在启动全量迁移之后，如果您的数据有变化，建议在全量迁移结束后暂停业务，并在源服务器系统和目标ECS实例之间再做一次增量同步。	增量迁移
一次性迁移较多数量的服务器时，使用单台全量迁移的方式较耗时。服务器数量为十台以上，建议您制作脚本进行批量迁移。	批量迁移
如果您能直接从自建机房（Integrated Data Center，IDC）、虚拟机环境或者云主机访问某一阿里云地域下的专有网络VPC，建议您使用源服务器与VPC内网互连的迁云方案。VPC内网迁云能获得比通过公网更快速更稳定的数据传输效果，提高迁云工作效率。	VPC内网迁移
迁云工具常规使用的是创建快照制作自定义镜像的迁移方式。如果您已经创建了相对应数量的目标ECS实例，可以使用直接迁移到目标实例的方式。这种方式能够提升迁移速度，您无需创建快照，也不必生成自定义镜像。	迁移到目标实例
如果您使用的服务器系统较早，或者不在适用列表中，可以联系技术人员咨询迁云方案。	其他迁移方案

5.4.2 全量迁移

首次从线下IDC或者静态的应用环境中迁移到阿里云时，您需要先将当前所有的数据做一次全量迁移。全量的迁移不需要您停止当前的业务，但迁移过程中的增量数据需要在后续做增量迁移。

Windows系统全量迁移

准备工作

1. 检查并确保Windows系统VSS服务为启动状态 。
2. 检查是否安装qemu-agent工具。如果安装了此工具，您需要先卸载。卸载步骤请参见迁云工具FAQ。

操作步骤

1. 下载并安装迁云工具到待迁移的服务器。具体步骤请参见下载并安装迁云工具。
2. 配置*user_config.json*。

   *user_config.json*配置文件的主要配置项包括：

   - 阿里云账号AccessKey信息
   - 迁移目标区域、目标镜像名称
   - （可选）目标系统盘大小、目标数据盘配置
   - 迁移源系统平台、架构

   各配置项的详细配置方法，请参见配置迁移源和迁移目标。
3. （可选）配置无需迁移的目录或文件。具体配置方法，请参见（可选）排除不迁移的文件或目录。
4. 运行迁云工具主程序。

   以管理员身份运行go2aliyun_client.exe或go2aliyun_gui.exe。如果是GUI版本，则需要单击start按钮开始迁移。

Linux系统全量迁移

我们以CentOS 7.6系统为例，为您介绍Linux系统全量迁移的操作步骤。其它Linux系统的迁移步骤相同，具体操作命令可能稍有差别。

准备工作

1. 运行以下命令将迁云工具下载到待迁移的服务器。

```
wget http://p2v-tools.oss-cn-hangzhou.aliyuncs.com/Alibaba_Cl
oud_Migration_Tool.zip
```

2. 运行以下命令解压缩迁云工具。

```
unzip Alibaba_Cloud_Migration_Tool.zip
```

3. 运行以下命令查看待迁移Linux系统的型号，并将适用于该系统型号的迁云工具包解压缩。

```
uname -a
unzip <适用于待迁移系统型号的迁云工具包>
```

本示例中，Linux型号为`x86_64`，因此，适用于该系统型号的迁云工具包为*go2aliyun_client1.3.2.3_linux_x86_64.zip*，如下图所示。

4. 运行以下命令进入解压后的迁云工具目录。

```
cd <解压后的迁云工具目录>
```

本示例中，该命令为`cd go2aliyun_client1.3.2.3_linux_x86_64`。

5. 运行以下命令检测Linux服务器是否迁移条件。

```
chmod +x ./Check/client_check
./Check/client_check --check
```

如果所有检测项的结果都为OK，表示该服务器满足迁移条件。您可以继续后面的迁移操作。

操作步骤

1. 配置*user_config.json*。

   *user_config.json*配置文件的主要配置项包括：

   - 阿里云账号AccessKey信息
   - 迁移目标区域、目标镜像名称
   - （可选）目标系统盘大小、目标数据盘配置
   - 迁移源系统平台、架构

   各配置项的详细配置方法，请参见配置迁移源和迁移目标。

2. （可选）配置无需迁移的目录或文件。具体配置方法，请参见（可选）排除不迁移的文件或目录。

3. 使用root权限运行以下命令，为迁云工具主程序添加可执行权限并执行该程序。

```
chmod +x go2aliyun_client
./go2aliyun_client
```

4. 等待迁云工具运行完成。当运行迁云工具的界面上提示Go to Aliyun Finished!时，表示迁移完成。如下图所示。

下一步

前往ECS控制台的镜像详情页查看结果。您的源服务器中的操作系统、应用程序以及应用数据将以自定义镜像的形式出现在相应地域的ECS控制台上。

对于全量迁移期间产生的增量数据，需要做增量迁移。

5.4.3 增量迁移

在启动全量迁移之后，如果您的数据有变化，建议在全量迁移结束后暂停业务，并在源服务器系统和目标ECS实例之间再做一次增量同步。

如果您想要在在线增量同步数据库的数据，推荐使用 阿里云DTS服务。

前提条件

已完成 全量迁移，该迁移在ECS控制台上成功生成自定义镜像（即全量镜像）。

操作步骤

1. 暂停您的业务。
2. 使用全量镜像 创建一个按量收费的ECS实例，并配置网络与源系统连通。
3. 使用增量同步工具在源系统和目标ECS实例之间做增量数据同步，减少业务暂停时间。

   同步工具推荐您使用rsync，goodsync等。此处以rsync工具为例，说明如何在源系统和目标ECS实例之间进行数据同步。假设您的目标ECS实例的IP是10.0.0.11，需要同步的目录路径是/disk1，rsync命令的示例代码为 `rsync -azvASX --partial --progress -e "ssh " /disk1/ root@10.0.0.11:/disk1/`。更多rsync使用介绍，请参见 rsync官网参数说明。

   说明:

   对于数据库增量同步，您可以考虑使用阿里云DTS服务。

5.4.4 批量迁移

一次性迁移较多数量的服务器时，使用单台全量迁移的方式较耗时。服务器数量为十台以上，建议您制作脚本进行批量迁移。

背景信息

对于大批量的服务器系统，一般都会配备自动化运维工具来统一管理，例如较常用的Ansible。使用Ansible可以方便地完成一些需要重复操作的工作，例如，向100台服务器拷贝同一个文件，或者同时在100台服务器上安装Apache服务并启动。

自动化运维工具可以批量下发并执行脚本。迁云工具是一个客户端工具程序，无需安装或复杂配置即可使用。

操作步骤

1. 准备自动化批量运维工具。
2. 使用迁云工具命令行进行调用。

   迁云工具提供一系列的命令行参数，适用于命令行调用的场景。例如：

   - --noenterkey：禁用交互
   - --nocheckversion：禁用提示版本更新
   - --progressfile：设置进度日志文件
3. 编写批量迁移任务脚本。

   根据实际迁移任务的需要来编写自动化批量迁移任务脚本，脚本中主要包括以下几项：

   a. 批量下发迁云工具到待迁移服务器

   b. 批量配置迁云工具，如目标镜像名等信息

   c. 批量执行迁云工具，同时获取迁移任务结果

示例脚本

```
#首先向所有服务器发送迁云工具程序

ansible -f 6 -i host.file all -m copy -a
"src=go2aliyun_client1.2.9.1_linux_x86_64.zip dest=/temp"

#然后解压缩程序

ansible -f 6 -i host.file all -m shell -a "cd /temp &&
unzip \

go2aliyun_client1.2.9.1_linux_x86_64.zip"

#再执行修改配置文件脚本

ansible -f 6 -i host.file all -m shell -a "cd
/temp/go2aliyun_client1.2.9.1_linux_x86_64 && ./config.sh"
```

```
sleep 120

# 配置文件脚本./config.sh工作是配置目标镜像名，主要根据子网IP来配置。（其他配置如AK，区域、磁盘信息等都是一致已配置好的）

#!/bin/bash

image_name=`ip a | grep inet | grep eth0 | grep brd | awk '{print $2}' | awk -F '/' '{print $1}'| awk -F '.' '{print "move_"$1"_"$2"_"$3"_"$4}'`

sed -i "s/IMAGE_NANE/${image_ame}/" user_config.json



#最后执行迁移脚，同时运行并发量是6个

ansible -f 6 -i host.file all -m shell -a "cd /temp/go2aliyun_client1.2.9.1_linux_x86_64 && chmod +x go2aliyun_client
&&./go2aliyun_client --nocheckversion --noenterkey"

#获取迁云结果，从client_data中获取生成的镜像Id以及完成状态

#判断client_data里的status自带，如果是Finished则表示迁云完成，同时image_id字段就是最终生成的镜像Id。
```

5.4.5 VPC内网迁移

如果您能直接从自建机房（Integrated Data Center，IDC）、虚拟机环境或者云主机访问某一阿里云地域下的专有网络VPC，建议您使用源服务器与VPC内网互连的迁云方案。VPC内网迁云能获得比通过公网更快速更稳定的数据传输效果，提高迁云工作效率。

您可以使用高速通道和VPN打通VPC内网的方案，然后使用迁云工具来进行VPC内网迁移。

背景信息

迁云工具1.2.8以上版本支持VPC内网迁移。要完成VPC内网迁移，需要将client_data的net_mode字段配置为1或2。

net_mode的参数说明如下：

- 0：默认为0，表示公网迁移，需要待迁移系统支持公网，数据传输阶段会走公网。
- 1：表示待迁移支持访问指定VPC内网；迁移过程分阶段1、2、3，阶段2数据传输在当前系统中进行，同时阶段1和3需要在其他公网环境系统中进行。
- 2：表示待迁移系统同时支持公网和支持访问特定VPC内网；跟一般操作过程相同，但数据传输阶段会走内网。

不同的参数设置有不同的迁移方式。

方式一

当net_mode设置为1时，参考以下步骤。

1. 在外网环境中创建中转实例。
   a. 登录某个有外网的系统A，并下载迁云工具。
   b. 配置*user_config.json*文件。
   c. 配置*client_data*文件的指定目标vpc_id、vswitch_id、zone_id等信息。详情请参见配置client_data文件到指定的VPC环境。
   d. 运行迁云工具，直到提示`Stage 1 Is Done!`。

2. 在VPC内网环境中传输系统数据。
   a. 登录需要迁移的VPC内网环境系统B。
   b. 将系统A的迁云工具复制到系统B。

   说明:
   系统B中的*user_config.json*和*client_data*文件必须要与系统A迁云工具中的一致。

   c. 运行迁云工具，直到提示`Stage 2 Is Done!`。

3. 在外网环境中创建镜像。
   a. 回到系统A，将系统B的迁云工具复制到系统A。

   说明:
   *user_config.json*和*client_data*文件必须与系统A迁云工具中的一致。

   b. 运行迁云工具，直到提示`Stage 3 Is Done!`以及迁云完成。

方式二

当net_mode设置为2时，参考以下步骤。

1. 登录待迁移的系统中，并下载迁云工具。
2. 配置*user_config.json*文件。
3. 配置*client_data*文件的指定目标vpc_id、vswitch_id、zone_id等信息。详情请参见配置client_data文件到指定的VPC环境。
4. 运行迁云工具直至迁云完成。

说明:

迁移过程中，迁移数据阶段通过VPC内网传输，其他阶段通过公网传输。

配置client_data文件

按照以下步骤配置client_data文件到指定的VPC环境。

1. 配置vpc_id为指定VPC的ID。

2. 配置vswitch_id为指定交换机的ID。

3. 配置zone_id为指定可用区相关的ID。

4. （可选）配置security_group_id为指定安全组的ID。如果您没有手动配置，则会自动创建。

说明:

指定安全组需要开放入方向的8080、8703端口。

5.4.6 迁移到目标实例

迁云工具常规使用的是创建快照制作自定义镜像的迁移方式。如果您已经创建了相对应数量的目标ECS实例，可以使用直接迁移到目标实例的方式。这种方式能够提升迁移速度，您无需创建快照，也不必生成自定义镜像。

准备工作

- 联系技术支持人员申请权限。
- 在阿里云目标区域准备一个或多个目标ECS实例，数量与源服务器对应。并且，ECS实例必须处于已停止状态。
- 迁移完成后，目标实例的系统盘会被替换，请提前做好备份工作。

操作步骤

参考以下步骤迁移到一个目标ECS实例。如果需要迁移多个目标ECS实例，重复以下步骤即可。

1. 下载迁云工具，版本为1.2.9.7或以上。

2. 配置*client_data*文件，在extra中配置target_instance_id项，表示目标实例ID。

> 说明:
> 目标实例的磁盘类型默认为高效云盘。如果目标实例的磁盘类型是SSD类型，则需要配置*client_data*文件，将instance_disk_cloud_ssd项设置为true。

3. 配置*user_config.json*。详情可参考全量迁移。
4. 执行迁云工具迁直至云完成。

后续操作

迁移完成后不会生成镜像，您可以启动目标ECS实例验证系统是否正常。

5.4.7 其他迁移方案

如果您使用的服务器系统较早，或者不在适用列表中，可以联系技术人员咨询迁云方案。

低版本系统上云

部分较早的版本系统，例如早于CentOS和Red Hat 5.5的版本，因为内核没有支持virtio等必要的虚拟化驱动，无法直接迁移到阿里云。这里我们以旧系统版本为CentOS 5.1（内核版本为2.6.18-53.el5），新系统版本为CentOS 5.5（内核版本为2.6.18-194.el5）为例，为您提供一种升级内核版本并迁移上云的方案。

操作步骤

1. 运行下列命令，确认系统版本为CentOS 5.1，内核版本为2.6.18-53.el5。

```
cat /etc/redhat-release
uname -r
```

2. 运行下列命令，下载并安装CentOS 5.5内核安装包。

```
wget http://vault.centos.org/5.5/os/x86_64/CentOS/kernel-2.6.18-194.
el5.x86_64.rpm
rpm -ivh ./kernel-2.6.18-194.el5.x86_64.rpm
```

> 说明:
> 如果新版本内核安装过程中报错，您需要检查报错日志。如果错误是由现有软件与新内核冲突引起的，您需要先手动卸载现有软件再重新安装新内核。新内核安装成功之后，再重装之前的软件即可。

3. 升级系统的GRUB引导程序至1.99版本。具体操作步骤请参阅 如何为 Linux 服务器安装 GRUB 。

   说明:

   您需要屏蔽旧版GRUB 0.97程序，以免新旧版本混淆影响使用。

4. 使用GRUB1.99版本重做引导：

   a. 运行 `grub-mkconfig -o /boot/grub/grub.cfg`命令更新GRUB配置文件。

   b. 运行 `cat /boot/grub/grub.cfg`命令检查该配置文件中是否包含旧版内核（2.6.18-53.el5）和新版内核（2.6.18-194.el5）。

   c. 运行 `fdisk -l` 命令找出系统盘设备。

   d. 假设您的系统盘设备为`/dev/sda`，运行 `grub-install --no-floppy --modules=part_msdos --boot-directory=/boot /dev/sda` 命令。

   e. 将新内核设置为默认启动项：

      A. 运行 `cat /boot/grub/grub.cfg |grep menuentry` 命令，查看内核启动项列表。

      B. 找到新内核启动项对应的标号，运行下列命令将新内核设置为默认启动项。

      ```
      mkdir /usr/local/etc/default/ -p
      echo "GRUB_DEFAULT=<新内核的启动项对应的标号>" >> /usr/local/etc/default/grub
      grub-mkconfig -o /boot/grub/grub.cfg
      ```

      例如，新内核为GNU/Linux, with Linux 2.6.18-194.el5，对应的标号为2，则命令为

      ```
      mkdir /usr/local/etc/default/ -p
      echo "GRUB_DEFAULT=2" >> /usr/local/etc/default/grub
      grub-mkconfig -o /boot/grub/grub.cfg
      ```

5. 重启操作系统。系统应正常启动并进入GRUB菜单页面，默认使用新内核2.6.18-194.el5进入操作系统。

6. 上述过程成功完成后，您可以 下载并安装迁云工具 进行迁移。

其他系统上云

如果您的系统不在应用场景的适用列表中，例如Oracle Linux，Amazon Linux、XenServer等，请联系技术支持人员，可以根据您的实际需求来进行相关系统测试，并提供相关迁云方案。

5.5 后续工作

完成迁移后，您需要测试确保系统能够正常运行。

创建实例

服务器迁移之后得到的是对应数量的自定义镜像，后续还要将这些镜像创建成实例，来进行系统验证测试等。

- 创建少量实例

  如果创建少量实例，建议在ECS售卖页面使用自定义镜像创建实例。创建时，计费方式选择按量付费，并指定VPC、交换机、安全组等网络配置，创建后可以手动修改为指定的内网IP。

- 创建大批量实例

  如果创建大批量实例，需要满足以下需求：

  - 创建按量收费的实例来做验证，验证完成后再升为包年包月的；
  - 保留跟原来系统相同的子网IP，因为涉及原业务相关；
  - 少量实例可以去控制台创建实例页面去购买，但大批量实例操作是不可能的，需要有工具调用来做。

  本文提供的方案使用阿里云命令行工具 CLI调用API配合脚本。主要步骤如下：

1. 下载阿里云CLI并配置AccessKeyId和AccessKeySecret。
2. 使用#unique_60接口创建实例。

   假设创建实例的目标区域是cn-qingdao，镜像ID是m-xxxxxxxxx， VSwitch是vsw-xxxxxxx，子网IP是10.0.0.10，实例规格是ecs.n1.samll，则做如下调用即可：

   ```
   aliyun ecs CreateInstance --RegionId 'cn-qingdao' --ImageId 'm-
   xxxxxxxxx' --VSWitchId 'vsw-xxxxxxx' --PrivateIP '10.0.0.10' --
   InstanceType 'ecs.n1.samll'
   ```

3. 将迁云工具所生成的镜像ID信息和对应的子网IP等信息做成配置，然后编写脚本调用CLI来自动读取进行批量执行创建实例。

> 说明：
> 批量创建实例并启动实例之后，您可以使用阿里云自动化批量运维工具云助手来批量管理和配置实例。

检查迁移后的Linux服务器

1. 检查系统盘数据是否完整。
2. 如果有数据盘，您需要自行 挂载数据盘。

3. 检查网络服务是否正常。
4. 然后检查其他系统服务是否正常。

检查迁移后的Windows服务器

1. 检查系统盘数据是否完整。
2. 如果有数据盘缺失，进入磁盘管理检查盘符是否丢失。
3. 等待文件系统权限修复过程完成后，选择是否重启实例。

> **说明:**
> 初次启动ECS实例后，如果文件系统权限修复程序未自启动，您可以运行`C:\go2aliyun_prepare\go2aliyun_restore.exe`手动修复。执行前要确保实例上的磁盘数量和盘符路径与源系统保持一致。

4. 检查网络服务是否正常。
5. 检查其他系统应用服务是否正常。

6 文件存储类迁移

6.1 迁移概述

文件类存储迁移至阿里云时，建议存储至阿里云OSS，迁移时可以使用在线迁移服务或闪电立方等迁移工具。

使用在线迁移服务

阿里云在线迁移服务是阿里云提供的存储产品数据通道。使用在线迁移服务，您可以将第三方数据轻松迁移至阿里云对象存储 OSS，也可以在对象存储 OSS 之间进行灵活的数据迁移。

使用在线迁移服务时，不同源迁移至阿里云OSS的操作教程如下：

- 阿里云 OSS 之间迁移
- HTTP/HTTPS 源迁移
- 腾讯云 COS 迁移
- AWS S3 迁移
- 七牛云迁移
- Azure Blob 迁移
- 又拍云迁移
- 百度云 BOS 迁移
- 金山云 KS3 迁移
- NAS 迁移至 OSS
- NAS 之间迁移
- ECS 数据迁移至 OSS

使用闪电立方

闪电立方Lightning Cube（海外叫数据迁移Data Transport）是阿里云为TB乃至PB级数据迁移提供的服务。它使用定制设备，安全，高效，低成本的帮助您把海量数据从本地机房迁移到云端或者是从云端迁移会本地机房，极大节省上云成本。

闪电立方目前提供三种类型的设备，分别适用到不同的数据量的迁移场景，可多套叠加使用：

- 闪电立方-Mini：适用40TB数据量的迁移
- 闪电立方-II：适用100TB数据量的迁移
- 闪电立方-III：适用480TB及其以上的数据量迁移

7 数据库迁移

7.1 数据迁移方案概览

RDS提供了多种数据迁移方案，可满足不同上云或迁云的业务需求，使您可以在不影响业务的情况下平滑将数据库迁移至阿里云云数据库RDS上面。通过使用阿里云数据传输服务（DTS），您可以实现MySQL数据库的结构迁移、全量迁移和增量迁移。另外，云数据库MySQL版还支持通过物理备份文件和逻辑备份文件两种途径，将云上数据迁移到本地数据库。

下表列出了RDS支持的上云、迁云、数据导出场景以及相关的操作链接：

使用场景	引擎类型	相关操作
将本地数据库迁移到云数据库MySQL	MySQL	· 使用 DTS 迁移 MySQL 数据 · 使用 mysqldump 迁移 MySQL 数据
	SQL Server	· 全量备份数据上云SQL Server 2008 R2版（推荐） · 使用 DTS 迁移 SQL Server 数据 · SQL Server 不停机迁移
	PostgreSQL	· 本地PostgreSQL迁移至RDS for PostgreSQL · 使用 psql 命令迁移 PostgreSQL 数据
	PPAS	Oracle到PPAS不停机数据迁移
将ECS上的自建库迁移到云数据库MySQL	· MySQL · SQL Server · PostgreSQL · PPAS	· 将ECS上的自建MySQL数据库迁移到RDS · 将ECS上的自建MySQL数据库迁移到其它阿里云账号下的RDS
将其它品牌的云数据库迁移到阿里云云数据库MySQL	MySQL	· 从AWS RDS迁移MySQL到阿里云RDS · 从腾讯云云数据库迁移MySQL到阿里云RDS
RDS实例间的数据库迁移	· MySQL · SQL Server · PostgreSQL · PPAS	· 不同RDS实例下库名不同的数据库之间的数据迁移 · 将云数据库MySQL迁移到其它阿里云账号的RDS

使用场景	引擎类型	相关操作
单个RDS实例内的数据迁移	· MySQL · SQL Server · PostgreSQL · PPAS	RDS实例内不同数据库之间的数据迁移
将RDS数据迁移到本地MySQL数据库	MySQL	迁移 RDS for MySQL 数据到本地 MySQL
	SQL Server	迁移 RDS for SQL Server 数据到本地 SQL Server
	PostgreSQL	迁移 RDS for PostgreSQL 数据到本地 PostgreSQL
	PPAS	· 迁移 RDS for PPAS 数据到本地 Oracle · 迁移 RDS for PPAS 数据到本地 PPAS

7.2 数据迁移方案概览

HybridDB for PostgreSQL提供了多种数据迁移方案，可满足不同的上云或迁云的业务需求，使您可以在不影响业务的情况下平滑地在其他数据库和HybridDB for PostgreSQL之间进行数据迁移。

HybridDB for PostgreSQL支持的各种数据迁移应用场景及相关操作如下：

操作	场景
OSS 外部表的使用	通过OSS外部表将数据在HybridDB for PostgreSQL和OSS之间进行导入或者导出。
使用数据集成迁移及批量同步数据	通过数据集成（Data Integration）在HybridDB for PostgreSQL中进行数据的导入或者导出。
使用 rds_dbsync 迁移/同步 MySQL 数据到AnalyticDB for PostgreSQL	通过mysql2pgsql工具将本地MySQL中的表导入到HybridDB for PostgreSQL中。
使用 rds_dbsync 迁移/同步PostgreSQL数据到AnalyticDB for PostgreSQL	通过pgsql2pgsql工具将HybridDB for PostgreSQL/Greenplum Database/PostgreSQL/PPAS中的表导入到HybridDB for PostgreSQL中。
COPY 命令的使用	通过`\COPY`命令，将本地的文本文件的数据导入到HybridDB for PostgreSQL中。

操作	场景
Amazon Redshift迁移数据到AnalyticDB for PostgreSQL	通过Amazon S3和阿里云OSS将Amazon Redshift的数据导入到HybridDB for PostgreSQL中。

7.3 数据迁移方案概览

HybridDB for MySQL提供了多种数据迁移方案，可满足不同的上云或迁云的业务需求，使您可以在不影响业务的情况下平滑地在其他数据库和HybridDB for MySQL之间进行数据迁移。

HybridDB for MySQL支持的数据迁移应用场景及操作如下：

操作	适用的引擎类型	场景
“数据集成”导入	事务引擎/分析引擎	使用数据集成（Data Integration）将HybridDB for MySQL导入或导出到其他阿里云数据库产品。
“数据传输”导入	事务引擎/分析引擎	使用数据传输（Data Transmission Service）将自建MySQL数据库或者RDS for MySQL数据库迁移到HybridDB for MySQL中。
从MySQL批量导入导出	分析引擎	HybridDB for MySQL支持从自建MySQL中全量导入和导出数据。
从MaxCompute批量导入导出	分析引擎	HybridDB for MySQL支持从MaxCompute中导入和导出数据。
从OSS批量导入导出	分析引擎	HybridDB for MySQL支持从OSS中导入和导出数据。
数据导出到Redis	分析引擎	HybridDB for MySQL支持将数据从HybridDB for MySQL导出到Redis。
数据导出到MongoDB	分析引擎	HybridDB for MySQL支持将数据从HybridDB for MySQL导出到MongoDB。

操作	适用的引擎类型	场景
从RDS原生实时同步	分析引擎	HybridDB for MySQL原生支持直接从RDS实时同步数据，您可以快速地构建起RDS到HybridDB for MySQL的同步关系，轻松实现数据流转和复杂查询加速。
从OSS导入JSON数据文件	分析引擎	HybridDB for MySQL支持从OSS中导入JSON数据文件。

7.4 迁移方案概览

您可以在不影响业务的情况下将Redis数据在本地数据库与云数据库Redis版之间进行平滑迁移。

云数据库Redis版提供了多种数据迁移方案，能满足您在不同场景中的上云或者迁移需求。您可以：

- 使用阿里云数据传输服务DTS，实现Redis数据库的全量迁移和增量迁移；
- 使用RDB文件或者AOF文件进行迁移；
- 使用redis-shake等工具进行迁移。

请参考下表中的场景和操作，快速发现满足您需求的文档。

表 7-1: 迁移方案

场景	操作
从本地数据库迁移到云数据库Redis版	Codis/Redis集群版通过redis-shake迁移上云
	使用redis-shake进行迁移
	使用DTS进行迁移
	使用redis-shake迁移RDB文件内的数据
	使用AOF文件进行迁移
从云数据库Redis版迁移到本地数据库	备份集迁移
云数据库Redis版之间的迁移	使用redis-shake在云数据库Redis版实例之间迁移
从第三方数据库迁移到Redis	将AWS ElastiCache for Redis数据库迁移到阿里云
	将SSDB数据库迁移到云数据库Redis版
	将Google Cloud Platform Memorystore数据库迁移到阿里云Redis

迁移完成后，为了确保目的端和源端数据一致，您可以使用redis-full-check进行数据校验，详细步骤请参见此文档。

8 大数据迁移

8.1 大数据迁移概述

本文整理总结将数据移迁移到MaxCompute的最佳实践相关文档。

当前很多用户的数据存放在传统的关系型数据库（RDS，做业务读写操作）中，当业务数据量庞大的时候，传统关系型数据库会显得有些吃力，此时经常会将数据迁移到大数据计算服务MaxCompute上。MaxCompute为您提供了完善的数据导入方案以及多种经典的分布式计算模型，能够更快速的解决海量数据存储和计算问题，有效降低企业成本。DataWorks（MaxCompute开发套件）为MaxCompute提供了一站式的数据同步、任务开发、数据工作流开发、数据管理和数据运维等功能。数据集成概述为您介绍阿里集团对外提供的稳定高效、弹性伸缩的数据同步平台。

最佳实践合集

- 通过使用DataWorks数据同步功能，将Hadoop数据迁移到阿里云MaxCompute大数据计算服务上，请参见Hadoop数据迁移MaxCompute最佳实践。详细的视频介绍，请参见Hadoop数据迁移到MaxCompute最佳实践（视频）。自建Hadoop迁移阿里云MaxCompute实践定期整理一些数据迁移和脚本迁移遇到的问题及解决方案，帮助企业快速拥有阿里巴巴同款数据仓库，构建自己的数据中台，并开展数据业务。
- 使用DataWorks数据集成同步功能，自动创建分区，动态的将RDS中的数据，迁移到MaxCompute大数据计算服务上，请参见RDS迁移到MaxCompute实现动态分区。
- 利用DataWorks数据集成将JSON数据从OSS迁移到MaxCompute，并使用MaxCompute内置字符串函数GET_JSON_OBJECT提取JSON信息，详细描述请参见JSON数据从OSS迁移到MaxCompute最佳实践。
- 利用DataWorks数据集成直接从MongoDB提取JSON字段到MaxCompute，请参见JSON数据从MongoDB迁移到MaxCompute最佳实践。

8.2 Hadoop数据迁移MaxCompute最佳实践

本文向您详细介绍如何通过使用DataWorks数据同步功能，将Hadoop数据迁移到阿里云MaxCompute大数据计算服务上。

环境准备

1. Hadoop集群搭建

进行数据迁移前，您需要保证自己的Hadoop集群环境正常。本文使用阿里云EMR服务自动化搭建Hadoop集群，详细过程请参见：创建集群。

本文使用的EMR Hadoop版本信息如下：

EMR版本: EMR-3.11.0

集群类型: HADOOP

软件信息: HDFS2.7.2 / YARN2.7.2 / Hive2.3.3 / Ganglia3.7.2 / Spark2.2.1 / HUE4.1.0 / Zeppelin0.7.3 / Tez0.9.1 / Sqoop1.4.6 / Pig0.14.0 / ApacheDS2.0.0 / Knox0.13.0

Hadoop集群使用经典网络，区域为华东1（杭州），主实例组ECS计算资源配置公网及内网IP，高可用选择为否（非HA模式），具体配置如下所示。

2. MaxCompute

请参考：开通MaxCompute。

开通MaxCompute服务并创建好项目，本文中在华东1（杭州）区域创建项目bigdata_DOC，同时启动DataWorks相关服务，如下所示。

数据准备

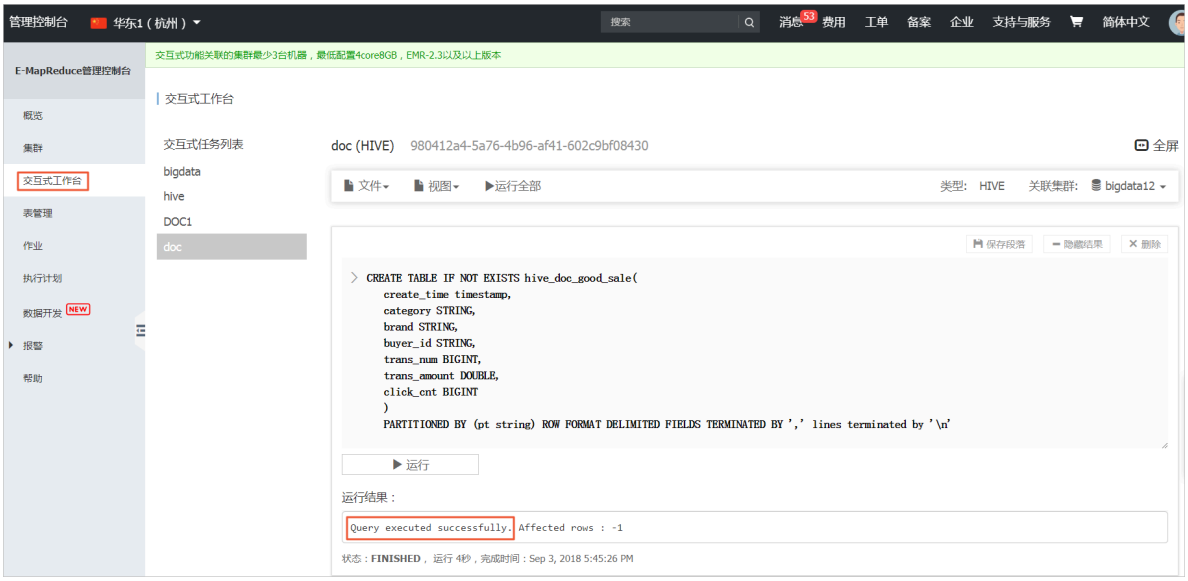
1. Hadoop集群创建测试数据

进入EMR Hadoop集群控制台界面，使用交互式工作台，新建交互式任务doc。本例中HIVE建表语句：

```
CREATE TABLE IF NOT
EXISTS hive_doc_good_sale(

   create_time timestamp,

   category STRING,

   brand STRING,

   buyer_id STRING,

   trans_num BIGINT,

   trans_amount DOUBLE,

   click_cnt BIGINT

   )

   PARTITIONED BY (pt string) ROW FORMAT
```

```
DELIMITED FIELDS TERMINATED BY ',' lines terminated by '\n'
```

选择运行，观察到Query executed successfully提示则说明成功在EMR Hadoop集群上创建了测试用表格hive_doc_good_sale，如下图所示。

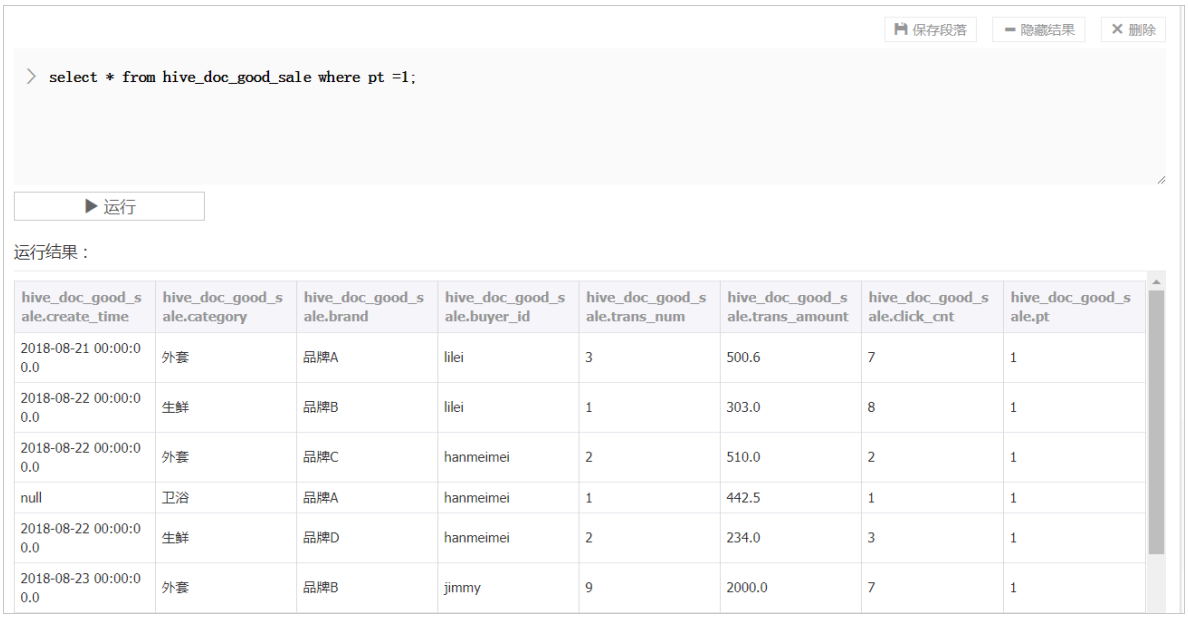


插入测试数据，您可以选择从OSS或其他数据源导入测试数据，也可以手动插入少量的测试数据。本文中手动插入数据如下：

```
insert into
hive_doc_good_sale PARTITION(pt =1 ) values('2018-08-21','外套','品
牌A','lilei',3,500.6,7),('2018-08-22','生鲜','品牌B','lilei',1,303
,8),('2018-08-22','外套','品牌C','hanmeimei',2,510,2),(2018-08-22
,'卫浴','品牌A','hanmeimei',1,442.5,1),('2018-08-22','生鲜','品牌D','
hanmeimei',2,234,3),('2018-08-23','外套','品牌B','jimmy',9,2000,7),('
2018-08-23','生鲜','品牌A','jimmy',5,45.1,5),('2018-08-23','外套','品
牌E','jimmy',5,100.2,4),('2018-08-24','生鲜','品牌G','peiqi',10,5560
,7),('2018-08-24','卫浴','品牌F','peiqi',1,445.6,2),('2018-08-24','外
```

```
套','品牌A','ray',3,777,3),('2018-08-24','卫浴','品牌G','ray',3,122,3
),('2018-08-24','外套','品牌C','ray',1,62,7) ;
```

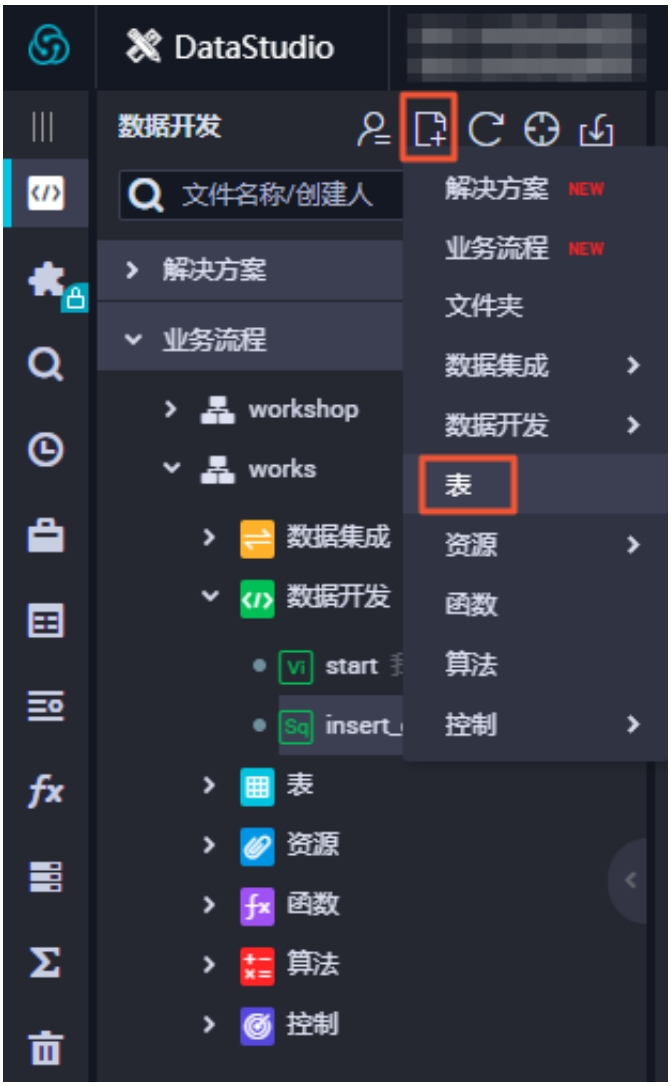
完成插入数据后，您可以使用select * from hive_doc_good_sale where pt =1;语句检查Hadoop集群表中是否已存在数据可用于迁移：



hive_doc_good_sale.create_time	hive_doc_good_sale.category	hive_doc_good_sale.brand	hive_doc_good_sale.buyer_id	hive_doc_good_sale.trans_num	hive_doc_good_sale.trans_amount	hive_doc_good_sale.click_cnt	hive_doc_good_sale.pt
2018-08-21 00:00:00.0	外套	品牌A	lilei	3	500.6	7	1
2018-08-22 00:00:00.0	生鲜	品牌B	lilei	1	303.0	8	1
2018-08-22 00:00:00.0	外套	品牌C	hanmeimei	2	510.0	2	1
null	卫浴	品牌A	hanmeimei	1	442.5	1	1
2018-08-22 00:00:00.0	生鲜	品牌D	hanmeimei	2	234.0	3	1
2018-08-23 00:00:00.0	外套	品牌B	jimmy	9	2000.0	7	1

2. 利用DataWorks新建目标表

在管理控制台，选择对应的MaxCompute项目，点击进入数据开发页面，点击新建表，如下所示。



在弹框中输入SQL建表语句，本例中使用的建表语句如下：

```
CREATE TABLE IF NOT EXISTS hive_doc_good_sale(
   create_time string,
   category STRING,
   brand STRING,
   buyer_id STRING,
   trans_num BIGINT,
   trans_amount DOUBLE,
   click_cnt BIGINT
   )
```

```
    PARTITIONED BY (pt string) ;
```

在建表过程中，需要考虑到HIVE数据类型与MaxCompute数据类型的映射，当前数据映射关系可参见：与Hive数据类型映射表。

由于本文使用DataWorks进行数据迁移，而DataWorks数据同步功能当前暂不支持timestamp类型数据，因此在DataWorks建表语句中，将create_time设置为string类型。上述步骤同样可通过odpscmd命令行工具完成，命令行工具安装和配置请参考：安装并配置客户端。执行过程如下所示：

```
odps@ bigdata_DOC>CREATE TABLE IF NOT EXISTS hive_doc_good_sale(
   create_time timestamp,
   category STRING,
   brand STRING,
   buyer_id STRING,
   trans_num BIGINT,
   trans_amount DOUBLE,
   click_cnt BIGINT
   )
   PARTITIONED BY (pt string) ;
                >               >               >               >
        >               >               >               >
 >
ID = 20180906110540873gev1bpim
OK
odps@ bigdata_DOC>drop table hive_doc_good_sale;
Confirm to "drop table hive_doc_good_sale;" (yes/no)? yes

ID = 20180906110825180gxh66292
```

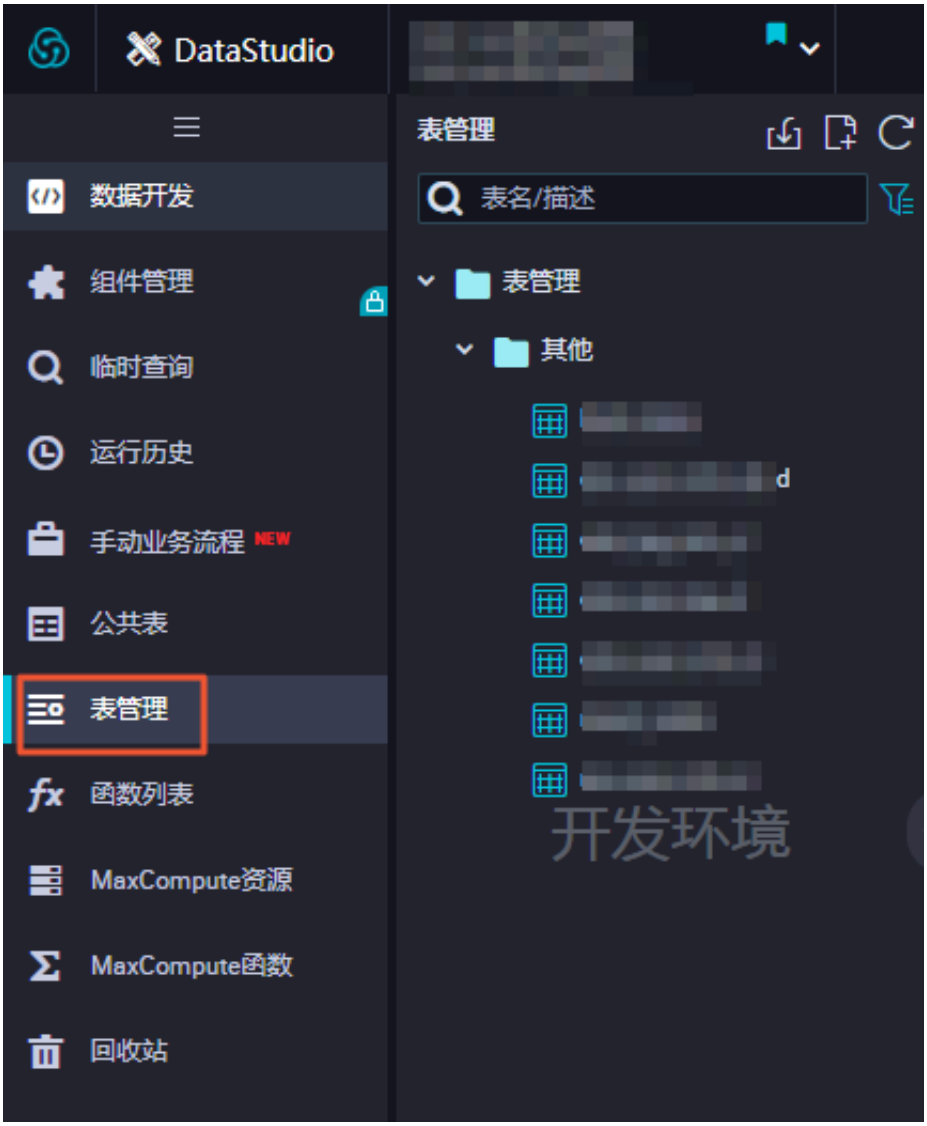
说明:

考虑到部分HIVE与MaxCompute数据类型的兼容问题，建议在odpscmd客户端上执行以下命令。

```
 set odps.sql.type.system.odps2=true;set
```

```
odps.sql.hive.compatible=true;
```

完成建表后，可在DataWorks数据开发>表查询一栏查看到当前创建的MaxCompute上的表，如下所示。



数据同步

1. 新建自定义资源组

由于MaxCompute项目所处的网络环境与Hadoop集群中的数据节点（data node）网络通常不可达，我们可通过自定义资源组的方式，将DataWorks的同步任务运行在Hadoop集群的Master节点上（Hadoop集群内Master节点和数据节点通常可达）。

a. 查看Hadoop集群datanode

在EMR控制台上首页/集群管理/集群/主机列表页查看，如下图所示，通常非HA模式的EMR上Hadoop集群的master节点主机名为 emr-header-1，datanode主机名为emr-worker-X。

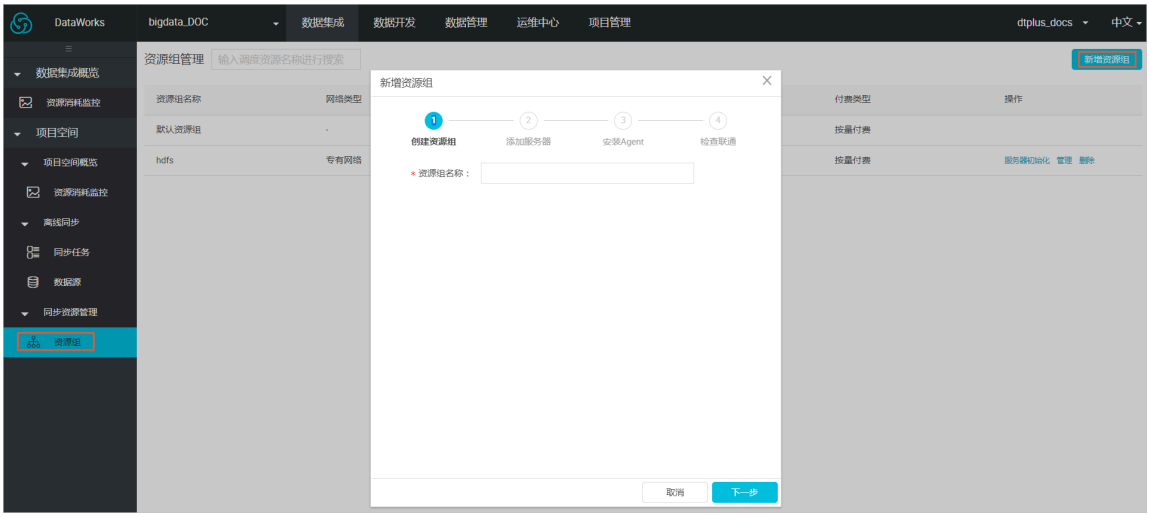
您也可以通过点击上图中Master节点的ECS ID，进入ECS实例详情页，通过点击远程连接进入ECS，通过 hadoop dfsadmin –report命令查看datenode，如下图所示。

```
DFS Used%: 0.05%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0

-------------------------------------------------
Live datanodes (2):

Name: 10.31.122.189:50010 (emr-worker-1.cluster-74503)
Hostname: emr-worker-1.cluster-74503
Decommission Status : Normal
Configured Capacity: 333373341696 (310.48 GB)
DFS Used: 155725824 (148.51 MB)
Non DFS Used: 325541888 (310.46 MB)
DFS Remaining: 332892073984 (310.03 GB)
DFS Used%: 0.05%
DFS Remaining%: 99.86%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu Sep 06 19:41:01 CST 2018


Name: 10.81.78.209:50010 (emr-worker-2.cluster-74503)
Hostname: emr-worker-2.cluster-74503
Decommission Status : Normal
Configured Capacity: 333373341696 (310.48 GB)
DFS Used: 155725824 (148.51 MB)
Non DFS Used: 325451776 (310.38 MB)
DFS Remaining: 332892164096 (310.03 GB)
DFS Used%: 0.05%
DFS Remaining%: 99.86%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu Sep 06 19:41:02 CST 2018
```

由上图可以看到，在本例中，datanode只具有内网地址，很难与DataWorks默认资源组互通，所以我们需要设置自定义资源组，将master node设置为执行DataWorks数据同步任务的节点。

b. 新建自定义资源组

进入DataWorks数据集成页面，选择资源组，点击新增资源组，如下图所示。关于自定义资源组的详细信息请参考新增调度资源。



在添加服务器步骤中，需要输入ECS UUID和机器IP等信息（对于经典网络类型，需输入服务器名称，对于专有网络类型，需输入服务器UUID。目前仅DataWorks V2.0 华东2区支持经典网络类型的调度资源添加，对于其他区域，无论您使用的是经典网络还是专有网络类型，在添加调度资源组时都请选择专有网络类型），机器IP需填写master node公网IP（内网IP有可能不可达）。ECS的UUID需要进入master node管理终端，通过命令dmidecode

| grep UUID获取（如果您的hadoop集群并非搭建在EMR环境上，也可以通过该命令获取），如下所示：

```
[root@emr-header-1 logs]# dmidecode | grep UUID
        UUID: F631D86C-
```

完成添加服务器后，需保证master node与DataWorks网络可达，如果您使用的是ECS服务器，需设置服务器安全组。如果您使用的内网IP互通，可参考添加安全组设置。

如果您使用的是公网IP，可直接设置安全组公网出入方向规则，本文中设置公网入方向放通所有端口（实际应用场景中，为了您的数据安全，强烈建议设置详细的放通规则），如下图所示。

完成上述步骤后，按照提示安装自定义资源组agent，观察到当前状态为可用，说明新增自定义资源组成功：

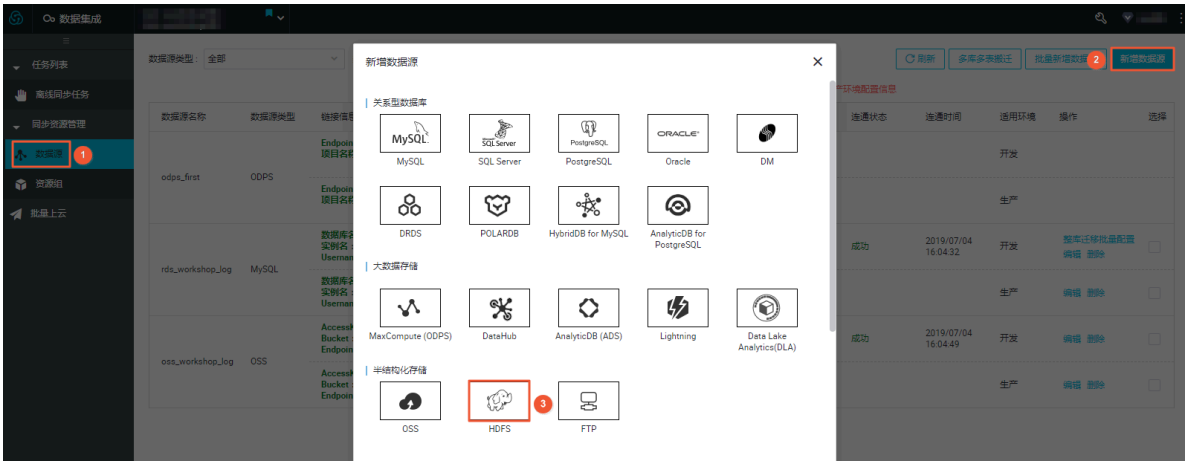
如果状态为不可用，您可以登录master node，使用`tail -f/home/admin/alisataskn ode/logs/heartbeat.log`命令查看DataWorks与master node之间心跳报文是否超时，如下图所示。

```
[root@emr-header-1 logs]# hdfs dfs -ls /user/hive/warehouse/hive_doc_good_sale/
Found 1 items
drwxr-x--x   - hive hadoop          0 2018-09-03 17:46 /user/hive/warehouse/hive_doc_good_sale/pt=1
[root@emr-header-1 logs]# tail -f /home/admin/alisatasknode/logs/heartbeat.log
2018-09-06 21:47:34,440 INFO  [pool-6-thread-1] [HeartbeatReporter.java:104] [] - heartbeat start, current status:2
2018-09-06 21:47:34,465 INFO  [pool-6-thread-1] [HeartbeatReporter.java:133] [] - heartbeat end■ cost time:0.025s
2018-09-06 21:47:39,465 INFO  [pool-6-thread-1] [HeartbeatReporter.java:104] [] - heartbeat start, current status:2
2018-09-06 21:47:39,491 INFO  [pool-6-thread-1] [HeartbeatReporter.java:133] [] - heartbeat end■ cost time:0.026s
2018-09-06 21:47:44,491 INFO  [pool-6-thread-1] [HeartbeatReporter.java:104] [] - heartbeat start, current status:2
2018-09-06 21:47:44,515 INFO  [pool-6-thread-1] [HeartbeatReporter.java:133] [] - heartbeat end■ cost time:0.024s
2018-09-06 21:47:49,516 INFO  [pool-6-thread-1] [HeartbeatReporter.java:104] [] - heartbeat start, current status:2
2018-09-06 21:47:49,538 INFO  [pool-6-thread-1] [HeartbeatReporter.java:133] [] - heartbeat end■ cost time:0.022s
2018-09-06 21:47:54,539 INFO  [pool-6-thread-1] [HeartbeatReporter.java:104] [] - heartbeat start, current status:2
2018-09-06 21:47:54,555 INFO  [pool-6-thread-1] [HeartbeatReporter.java:133] [] - heartbeat end■ cost time:0.016s
^C
```

2. 新建数据源

关于DataWorks新建数据源详细步骤，请参见：数据源配置。

DataWorks新建项目后，默认设置自己为数据源odps_first。因此我们只需添加Hadoop集群数据源:在DataWorks数据集成页面，点击数据源>新增数据源，在弹框中选择HDFS类型的数据源：



在弹出窗口中填写数据源名称及defaultFS。对于EMR Hadoop集群而言，如果Hadoop集群为HA集群，则此处地址为hdfs://emr-header-1的IP:8020，如果Hadoop集群为非HA集群，则此处地址为hdfs://emr-header-1的IP:9000。在本文中，emr-header-1与DataWorks通过公网连接，因此此处填写公网IP并放通安全组。

完成配置后，点击测试连通性，如果提示“测试连通性成功”，则说明数据源添加正常。

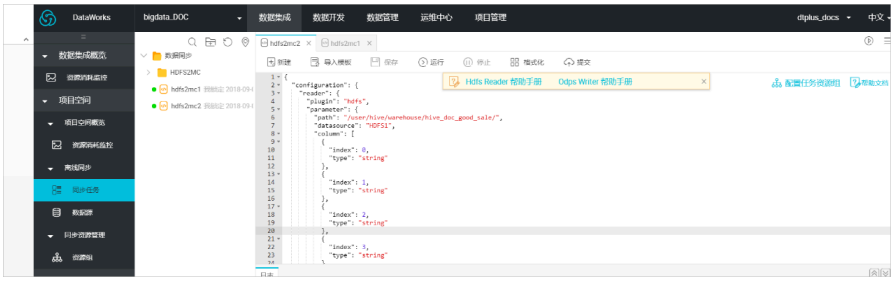
说明:

如果EMR Hadoop集群设置网络类型为专有网络，则不支持连通性测试。

3. 配置数据同步任务

在DataWorks数据集成页面点击同步任务，选择新建>脚本模式，在导入模板弹窗选择数据源类型如下：

完成导入模板后，同步任务会转入脚本模式，本文中配置脚本如下，相关解释请参见：脚本模式配置。



在配置数据同步任务脚本时，需注意DataWorks同步任务和HIVE表中数据类型的转换如下：

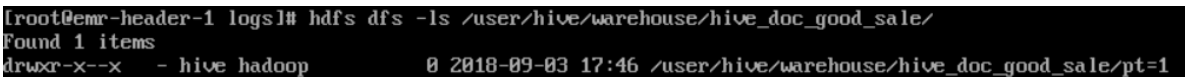
在Hive表中的数据类型	DataX/DataWorks 内部类型
TINYINT,SMALLINT,INT,BIGINT	Long
FLOAT,DOUBLE,DECIMAL	Double
String,CHAR,VARCHAR	String
BOOLEAN	Boolean
Date,TIMESTAMP	Date
Binary	Binary

详细代码如下：

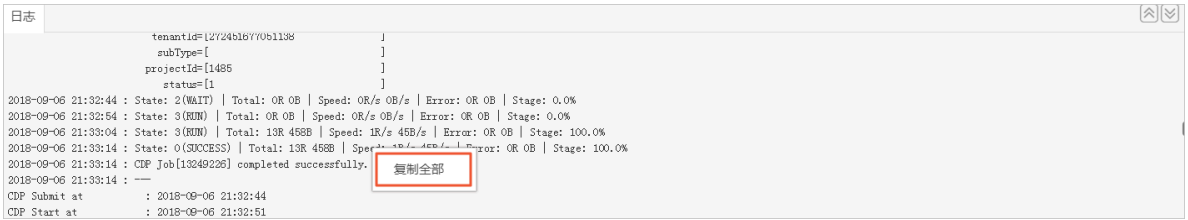
```
{
  "configuration": {
    "reader": {
      "plugin": "hdfs",
      "parameter": {
        "path": "/user/hive/warehouse/hive_doc_good_sale/",
        "datasource": "HDFS1",
        "column": [
          {
            "index": 0,
            "type": "string"
          },
          {
            "index": 1,
            "type": "string"
          },
          {
            "index": 2,
            "type": "string"
          },
          {
            "index": 3,
            "type": "string"
          },
          {
            "index": 4,
            "type": "long"
          },
          {
            "index": 5,
            "type": "double"
          },
          {
            "index": 6,
            "type": "long"
          }
        ],
        "defaultFS": "hdfs://121.199.11.138:9000",
        "fieldDelimiter": ",",
        "encoding": "UTF-8",
        "fileType": "text"
      }
    },
    "writer": {
      "plugin": "odps",
```

```
        "parameter": {
          "partition": "pt=1",
          "truncate": false,
          "datasource": "odps_first",
          "column": [
            "create_time",
            "category",
            "brand",
            "buyer_id",
            "trans_num",
            "trans_amount",
            "click_cnt"
          ],
          "table": "hive_doc_good_sale"
        }
      },
      "setting": {
        "errorLimit": {
          "record": "1000"
        },
        "speed": {
          "throttle": false,
          "concurrent": 1,
          "mbps": "1",
          "dmu": 1
        }
      }
    },
    "type": "job",
    "version": "1.0"
}
```

其中，path参数为数据在Hadoop集群中存放的位置，您可以在登录master node后，使用hdfs dfs –ls /user/hive/warehouse/hive_doc_good_sale命令确认。对于分区表，您可以不指定分区，DataWorks数据同步会自动递归到分区路径，如下图所示。



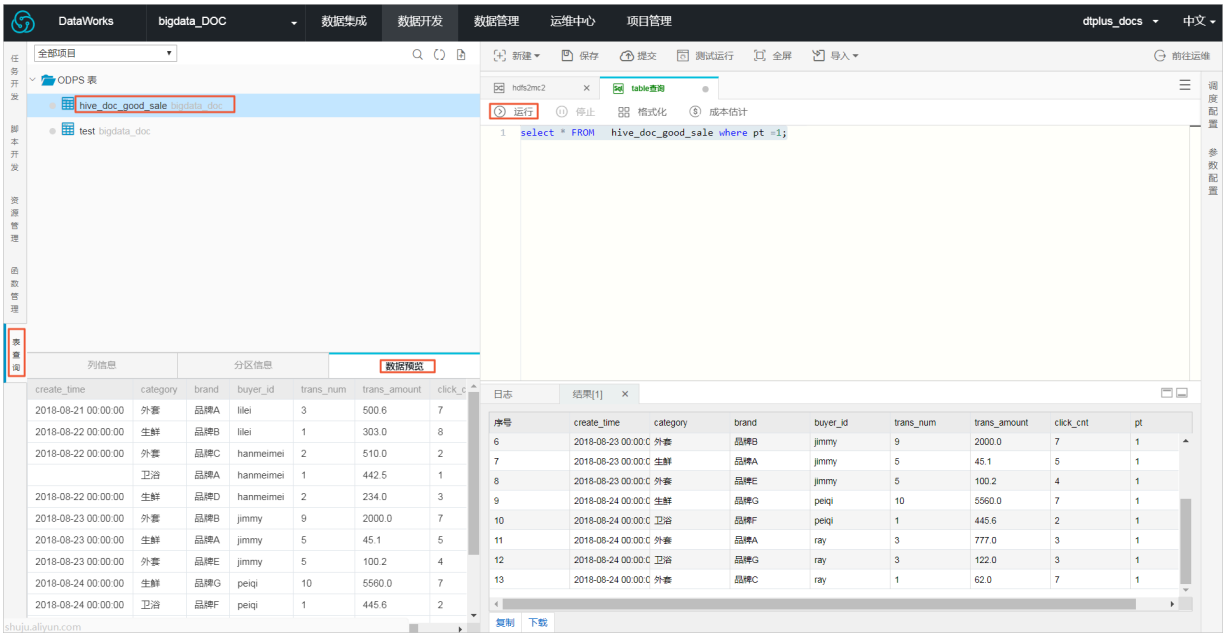
完成配置后，点击运行。如果提示任务运行成功，则说明同步任务已完成。如果运行失败，可通过复制日志进行进一步排查。



验证结果

在DataWorks数据开发/表查询页面，选择表hive_doc_good_sale后，点击数据预览可查看HIVE数据是否已同步到MaxCompute。您也可以通过新建一个table查询任务，在任务中输入

脚本`select * FROM hive_doc_good_sale where pt =1;`后，点击运行来查看表结果，如下图所示。



当然，您也可以通过在odpscmd命令行工具中输入`select * FROM hive_doc_good_sale where pt =1;`查询表结果。

MaxCompute数据迁移到Hadoop

如果您想实现MaxCompute数据迁移到Hadoop。步骤与上述步骤类似，不同的是同步脚本内的reader和writer对象需要对调，具体实现脚本举例如下。

```
{
  "configuration": {
    "reader": {
      "plugin": "odps",
      "parameter": {
      "partition": "pt=1",
      "isCompress": false,
      "datasource": "odps_first",
      "column": [
        "create_time",
        "category",
        "brand",
      "buyer_id",
      "trans_num",
      "trans_amount",
      "click_cnt"
    ],
    "table": "hive_doc_good_sale"
    }
  },
  "writer": {
    "plugin": "hdfs",
    "parameter": {
    "path": "/user/hive/warehouse/hive_doc_good_sale",
```

```
    "fileName": "pt=1",
    "datasource": "HDFS_data_source",
    "column": [
      {
        "name": "create_time",
        "type": "string"
      },
      {
        "name": "category",
        "type": "string"
      },
      {
        "name": "brand",
        "type": "string"
      },
      {
        "name": "buyer_id",
        "type": "string"
      },
      {
        "name": "trans_num",
        "type": "BIGINT"
      },
      {
        "name": "trans_amount",
        "type": "DOUBLE"
      },
      {
        "name": "click_cnt",
        "type": "BIGINT"
      }
    ],
    "defaultFS": "hdfs://47.99.162.100:9000",
    "writeMode": "append",
    "fieldDelimiter": ",",
    "encoding": "UTF-8",
    "fileType": "text"
    }
  },
  "setting": {
    "errorLimit": {
      "record": "1000"
  },
  "speed": {
    "throttle": false,
    "concurrent": 1,
    "mbps": "1",
    "dmu": 1
  }
  }
},
"type": "job",
"version": "1.0"
}
```

您需要参考配置HDFS Writer在运行上述同步任务前对Hadoop集群进行设置，在运行同步任务后手动拷贝同步过去的文件。

8.3 RDS迁移到MaxCompute实现动态分区

本文向您详细介绍如何使用DataWorks数据同步功能，自动创建分区，动态的将RDS中的数据，迁移到Maxcompute大数据计算服务上。

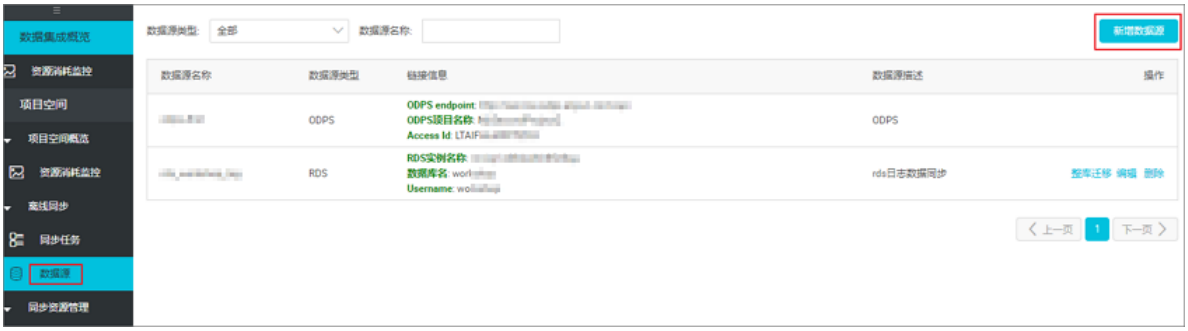
准备工作

1. MaxCompute

   开通MaxCompute服务并创建好项目，本文中在华北2（北京）区域创建项目，同时启动DataWorks相关服务，如下所示。

> **说明:**
> 如果您是第一次使用DataWorks，请确认已经根据准备工作模块的操作，准备好账号和项目角色、项目空间等内容，开通MaxCompute请参见请参见开通MaxCompute。然后进入DataWorks管理控制台，单击对应项目后的进入数据开发，即可开始数据开发操作。

2. 准备数据源，如下图。



新增ODPS数据源和MYSQL数据源，详细描述请参考文档配置MaxCompute数据源和配置MySQL数据源。

3. 确认作为目标的ODPS数据库中有表

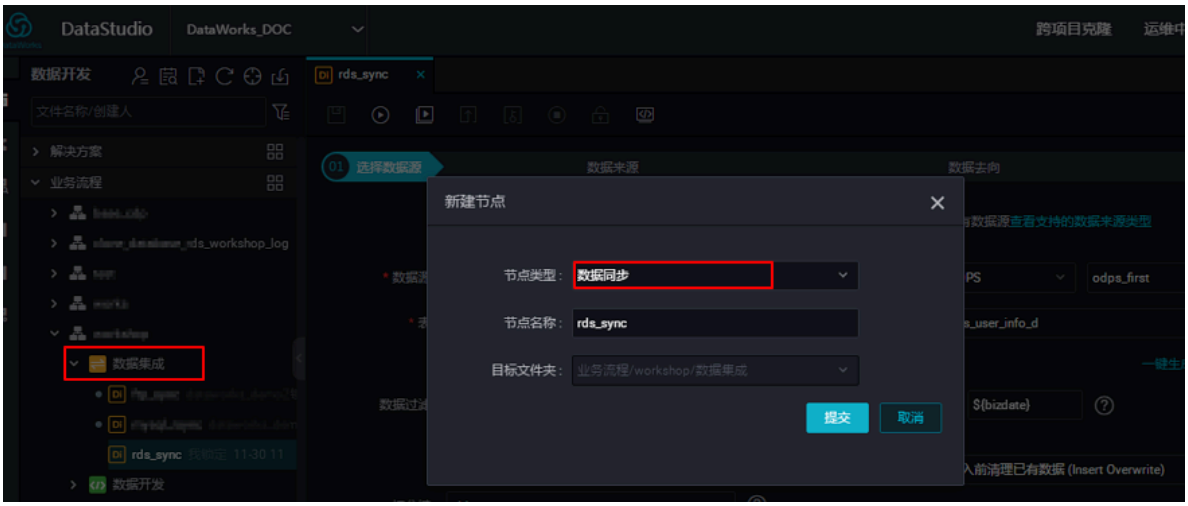
在ODPS数据库中创建RDS对应目标表emp_test_new。

建表完成后，可执行`desc emp_test_new;`语句查看表详情。

详细的信息请参见建表并上传数据。

4. 新建并配置同步任务节点

本节将新建一个同步节点sdtest并进行配置，将MySQL中的表emp_test_new中的数据写入到自己的ODPS数据库中。如下图。

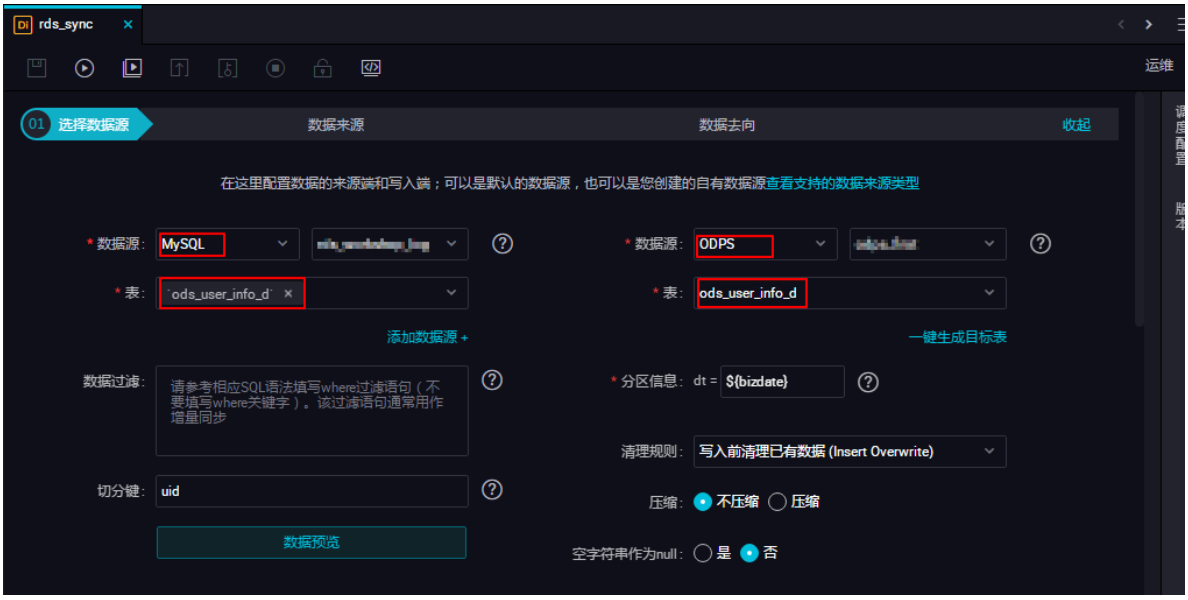


详细的数据同步任务的操作和配置请参见DataWorks数据开发和运维。

自动创建按天日期的分区

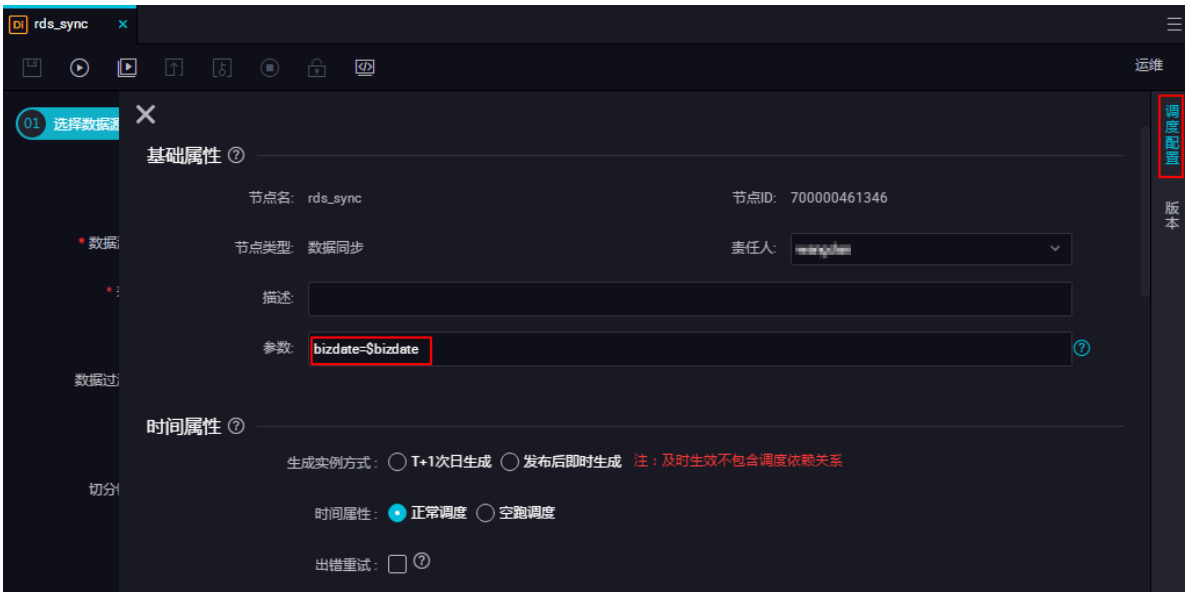
将RDS中的数据定时每天同步到MaxCompute中，自动创建按天日期的分区。

1. 选择来源，如下图。



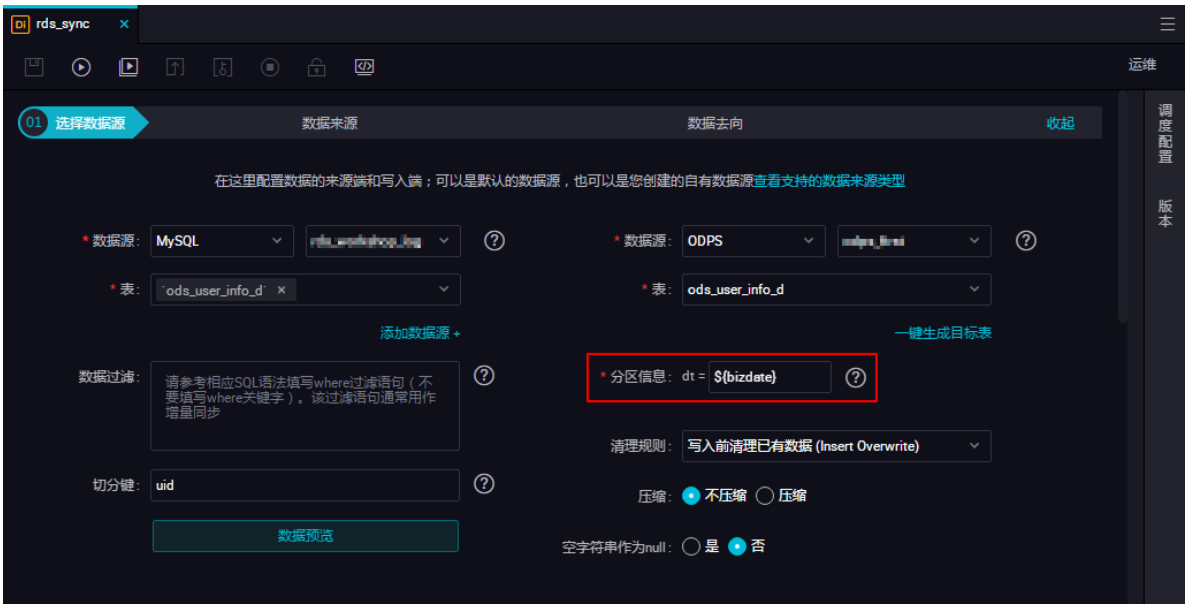
2. 选择目标，如下图。

3. 参数配置，如下图。



一般配置到这个地方的时候，默认是系统自带的时间参数：`${bdp.system.bizdate}`，格式为yyyymmdd。也就是说在调度执行这个任务的时候，这个分区会被自动替换为任务执行日

期的前一天，一般用户会在当前跑前一天的业务数据，这个日期也叫业务日期。如果用户要使用当天任务运行的日期作为分区值，需要自定义这个参数。如下图。



自定义参数设置，格式非常灵活，日期是当天日期，用户可以自由选择哪一天以及格式。可供参考的变量参数配置方式如下：

后N年：`$[add_months(yyyymmdd,12*N)]`

前N年：`$[add_months(yyyymmdd,-12*N)]`

后N月：`$[add_months(yyyymmdd,N)]`

前N月：`$[add_months(yyyymmdd,-N)]`

后N周：`$[yyyymmdd+7*N]`

前N周：`$[yyyymmdd-7*N]`

后N天：`$[yyyymmdd+N]`

前N天：`$[yyyymmdd-N]`

后N小时：`$[hh24miss+N/24]`

前N小时：`$[hh24miss-N/24]`

后N分钟：`$[hh24miss+N/24/60]`

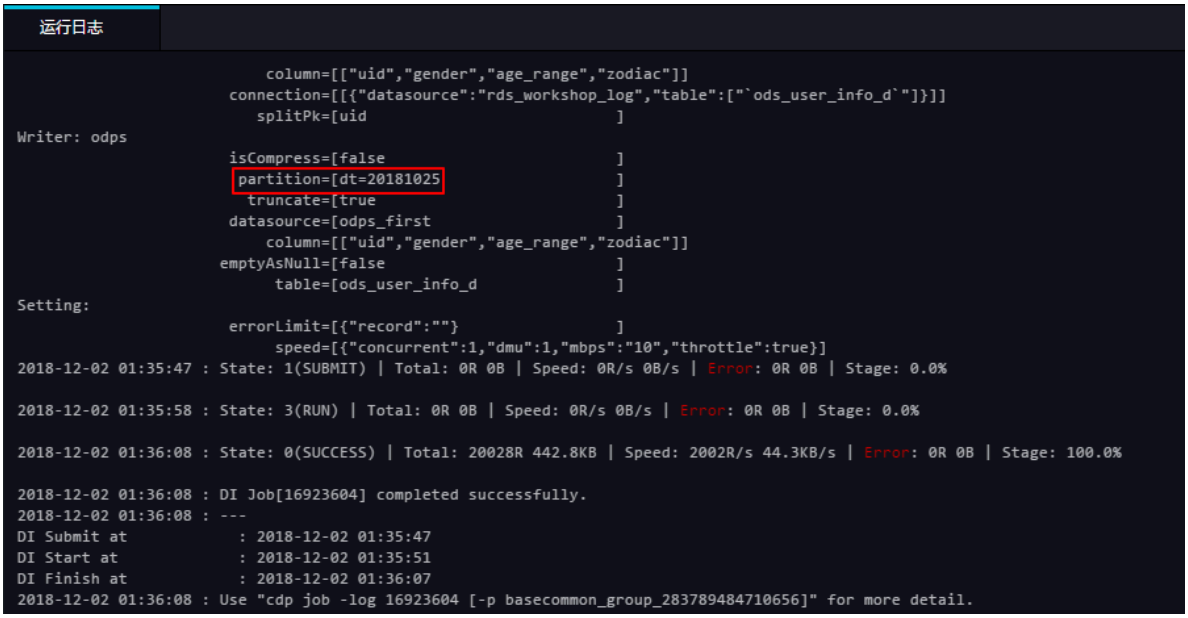
前N分钟：`$[hh24miss-N/24/60]`

说明：

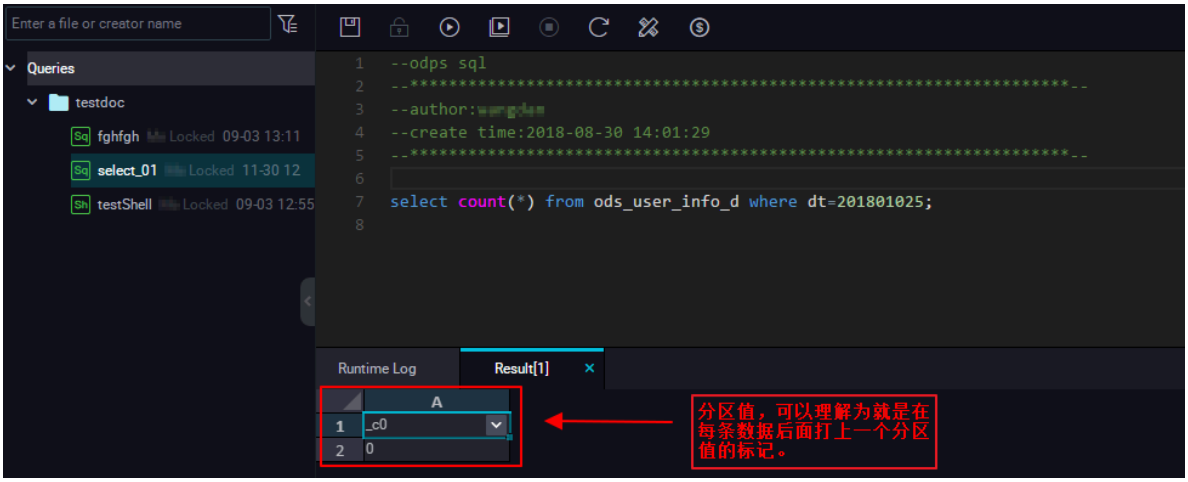
- 请以中括号 [] 编辑自定义变量参数的取值计算公式，例如 `key1=$[yyyy-mm-dd]`。
- 默认情况下，自定义变量参数的计算单位为天。例如 `$[hh24miss-N/24/60]` 表示 `(yyyymmddhh24miss-(N/24/60 * 1天))` 的计算结果，然后按 hh24miss 的格式取时分秒。
- 使用 add_months 的计算单位为月。例如 `$[add_months(yyyymmdd,12 N)-M/24/60]` 表示 `(yyyymmddhh24miss-(12 N 1月))-(M/24/60 1天)` 的结果，然后按 `yyyymmdd` 的格式取年月日。

参数设置请参见参数配置。

4. 测试运行，如下图。



可以看到日志中，MaxCompute（日志中打印原名ODPS）的信息中partition分区，date_test=20170829，自动替换成功。检查下实际的数据有没有转移到ODPS表中，如下图。



说明：

在maxcompute2.0中分区表查询需要添加分区筛选，SQL语句如下。其中分区列需要更新为业务日期，如任务运行的日期为20180717，那么业务日期为20180716。

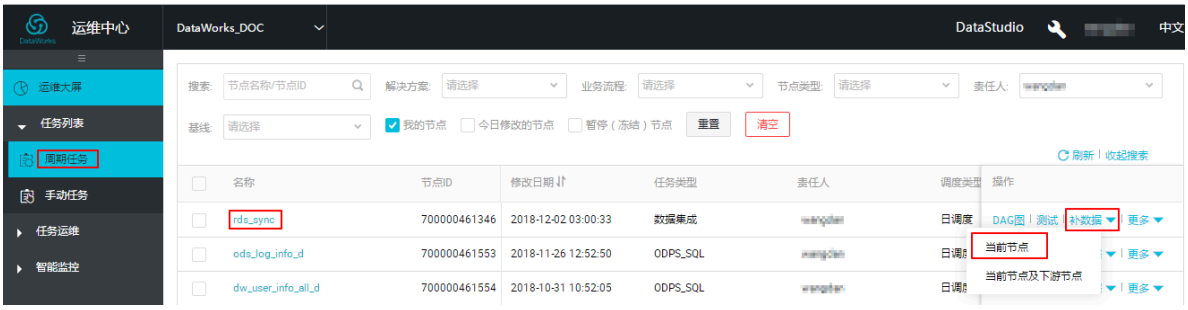
```
--查看是否成功写入MaxCompute
select count(*) from ods_raw_log_d where dt=业务日期;
select count(*)from ods_user_info_d where dt=业务日期;
```

此时看到数据已经迁移到ODPS表中，并且成功地创建了一个分区值。那么这个任务在执行定时调度的时候，会每天将RDS中的数据同步到MaxCompute中按照日期自动创建的分区中。

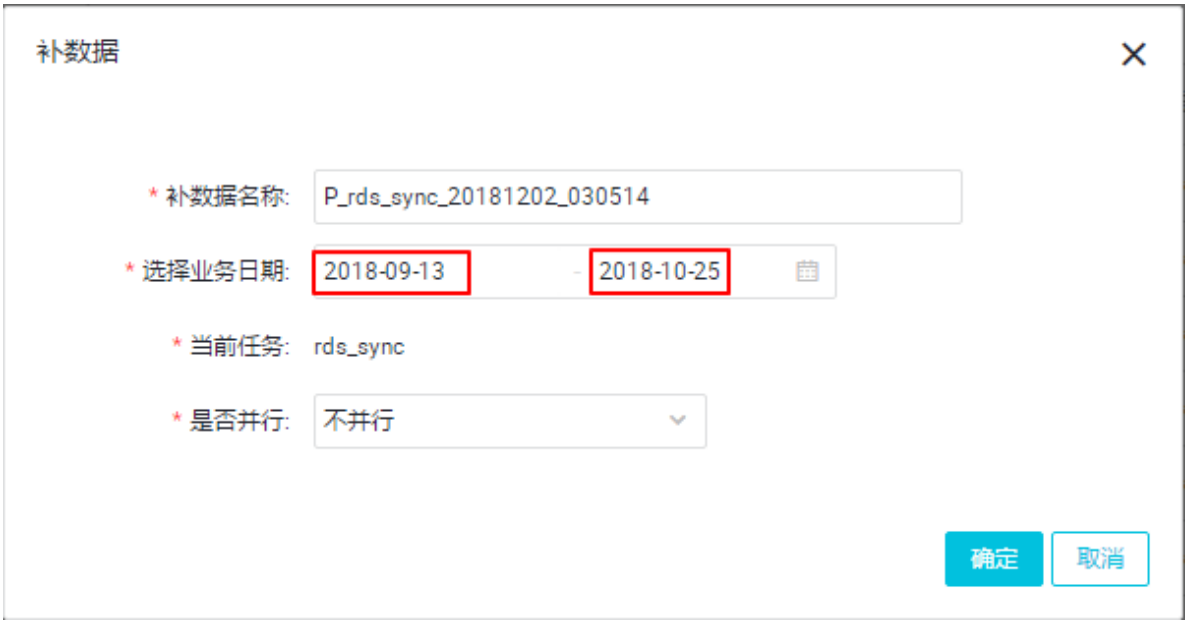
补数据功能实现历史数据自动同步

1. 首先，我们需要在RDS端把历史数据按照日期筛选出来，比如历史数据2017-08-25这天的数据，我们要自动同步到MaxCompute的20170825的分区中。在RDS阶段可以设置where过滤条件，如图。

2. 补数据操作。然后保存 > 提交。提交后到运维中心 > 任务管理 > 图形模式右键单击补数据节点。如下图。

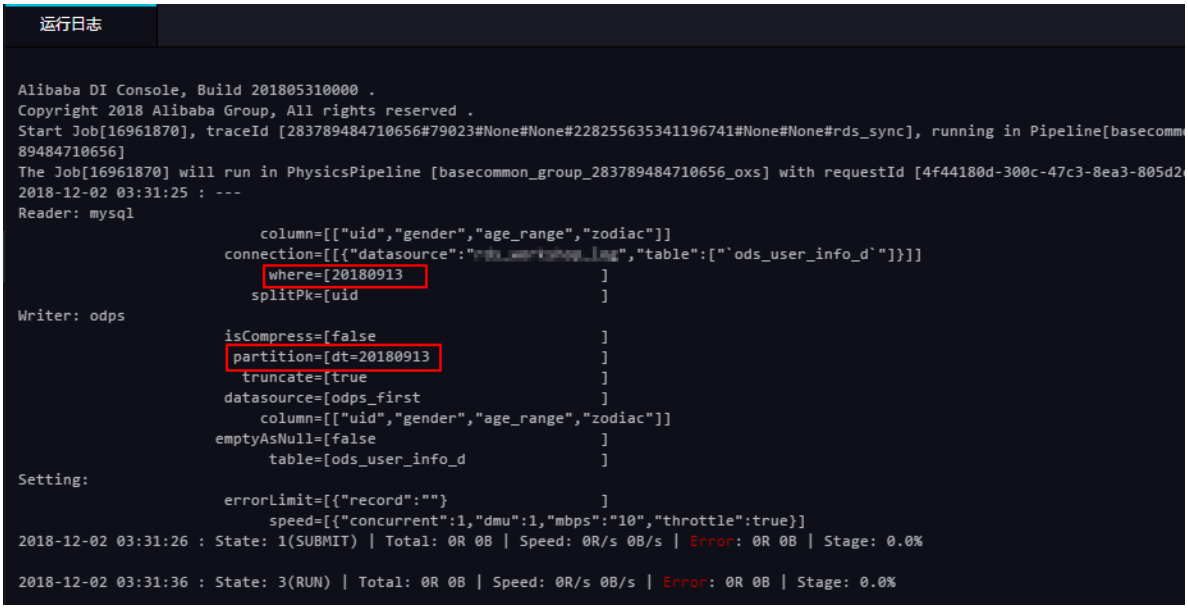


3. 跳转至补数据节点页面。选择日期区间，如下图。

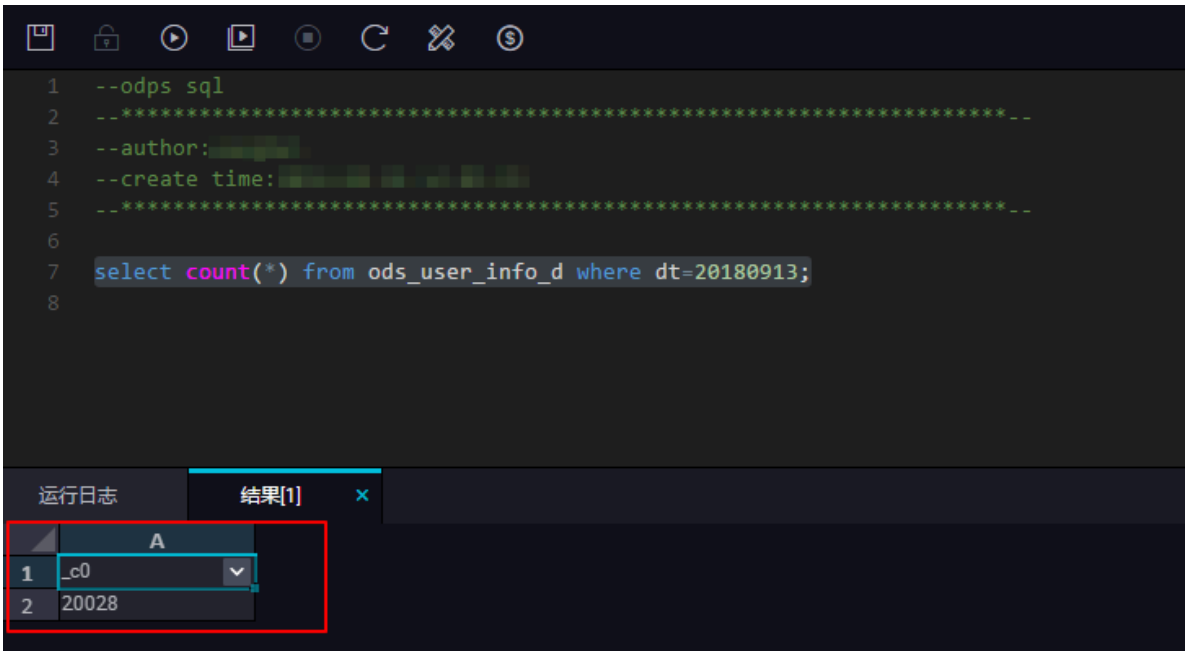


4. 单击提交 > 运行。此时会同时生成多个同步的任务实例按顺序执行。如下图。

5. 查看运行的日志，可以看到运行过程中对RDS数据的抽取。此时MaxCompute已自动创建分区，如图。



查看运行结果。数据写入的情况，以及是否自动创建了分区，数据是否已同步到分区表中，如下图。



说明：

在maxcompute2.0中分区表查询需要添加分区筛选，SQL语句如下。其中分区列需要更新为业务日期，如任务运行的日期为20180717，那么业务日期为20180716。

```
--查看是否成功写入MaxCompute
select count(*) from ods_raw_log_d where dt=业务日期;
```

```
select count(*)from ods_user_info_d where dt=业务日期;
```

Hash实现非日期字段分区

如果用户数据量比较巨大，或者第一次全量的数据并不是按照日期字段进行分区，而是按照省份等非日期字段分区，那么此时数据集成操作就不能做到自动分区了。也就是说，可以按照RDS中某个字段进行hash，将相同的字段值自动存放到这个字段对应值的MaxCompute分区中。

流程如下：

1. 首先我们需要把数据全量同步到MaxCompute的一个临时表。创建一个SQL脚本节点。命令如下。

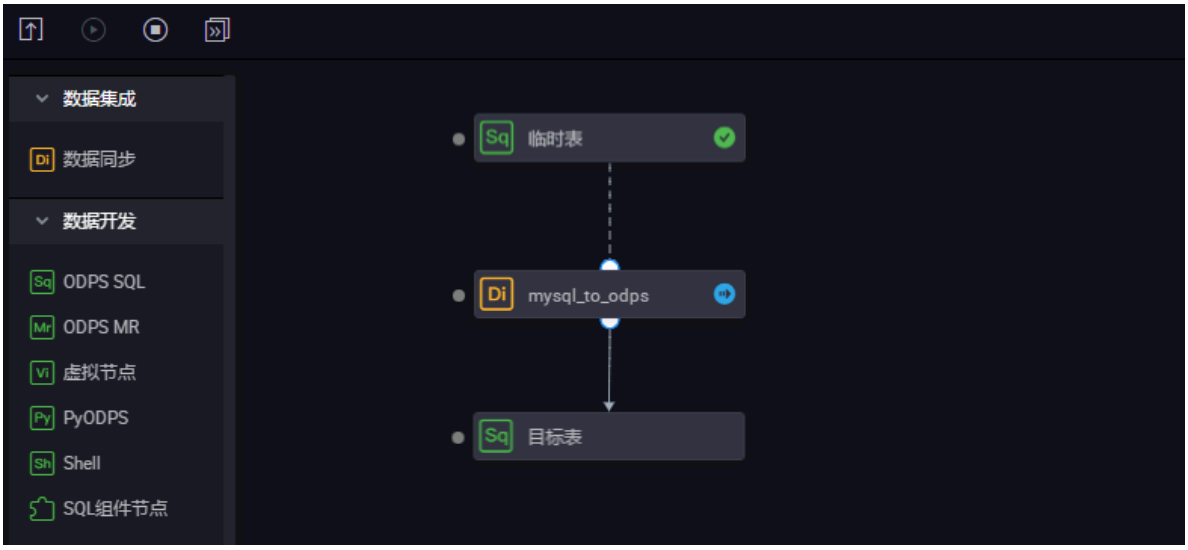
```
drop table if exists emp_test_new_temp;
CREATE TABLE emp_test_new_temp
(date_time STRING,
 name STRING,
 age BIGINT,
 sal DOUBLE);
```

2. 创建同步任务的节点，就是简单的同步任务，将RDS数据全量同步到MaxCompute，不需要设置分区。
3. 使用sql语句进行动态分区到目标表。命令如下：

```
drop table if exists emp_test_new;
--创建一个ODPS分区表（最终目的表）
 CREATE TABLE emp_test_new (
 date_time STRING,
 name STRING,
 age BIGINT,
 sal DOUBLE
)
PARTITIONED BY (
 date_test STRING
);
--执行动态分区sql，按照临时表的字段date_time自动分区，date_time字段中相同的数据值，会按照这个数据值自动创建一个分区值
--例如date_time中有些数据是2017-08-25，会自动在ODPS分区表中创建一个分区，date=2017-08-25
--动态分区sql如下
--可以注意到sql中select的字段多写了一个date_time，就是指定按照这个字段自动创建分区
insert overwrite table emp_test_new partition(date_test)select date_time,name,age,sal,date_time from emp_test_new_temp
--导入完成后，可以把临时表删除，节约存储成本
drop table if exists emp_test_new_temp;
```

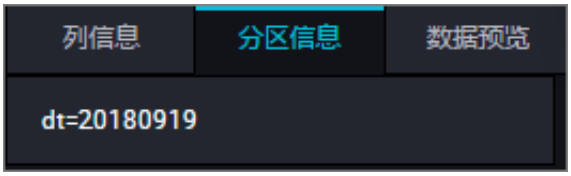
在MaxCompute中我们通过SQL语句来完成同步。详细的SQL语句介绍请参见阿里云大数据利器MaxCompute学习之--分区表的使用。

4. 最后将三个节点配置成一个工作流，按顺序执行。如下图。



5. 查看执行过程。我们可以重点观察最后一个节点的动态分区过程，如下图。

查看数据。动态的自动化分区完成。相同的日期数据迁移到了同一个分区中。如下图。



如果是以省份字段命名分区，执行步骤参请考上述内容。

DataWorks数据同步功能可以完成大部分自动化作业，尤其是数据的同步迁移，调度等，了解更多的调度配置请参见调度配置时间属性。

8.4 JSON数据从OSS迁移到MaxCompute最佳实践

本文为您介绍如何利用DataWorks的数据集成功能将JSON数据从OSS迁移到MaxCompute，并使用MaxCompute内置字符串函数GET_JSON_OBJECT提取JSON信息的最佳实践。

准备工作

- 数据上传OSS

  将您的JSON文件重命名后缀为TXT文件，并上传到OSS。本文中使用的JSON文件示例如下。

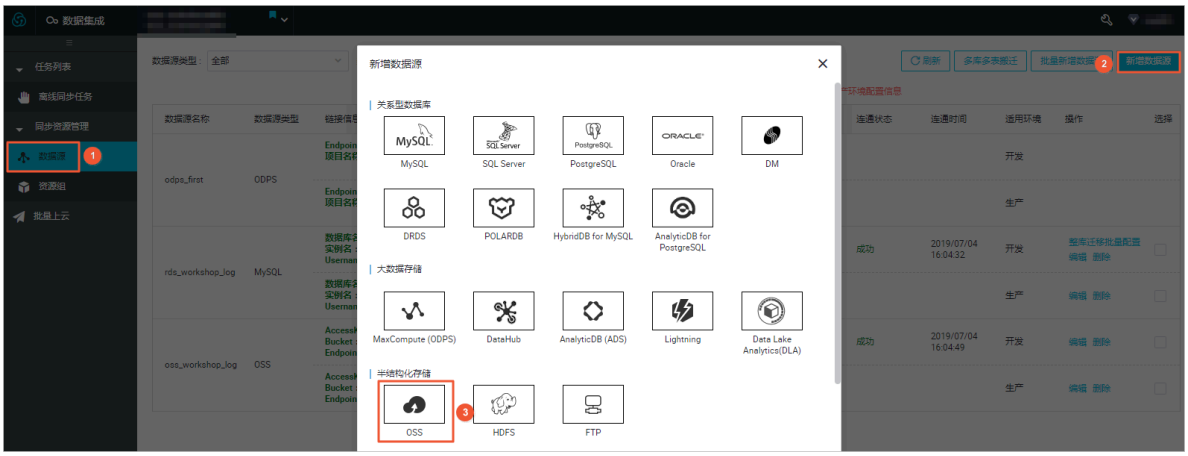
```
{
    "store": {
        "book": [
            {
               "category": "reference",
               "author": "Nigel Rees",
               "title": "Sayings of the Century",
               "price": 8.95
            },
            {
               "category": "fiction",
               "author": "Evelyn Waugh",
               "title": "Sword of Honour",
               "price": 12.99
            },
            {
                "category": "fiction",
                "author": "J. R. R. Tolkien",
                "title": "The Lord of the Rings",
                "isbn": "0-395-19395-8",
                "price": 22.99
            }
        ],
        "bicycle": {
            "color": "red",
            "price": 19.95
        }
    },
    "expensive": 10
}
```

  将*applog.txt*文件上传到OSS，本文中OSS Bucket位于华东2区。

使用DataWorks导入数据到MaxCompute

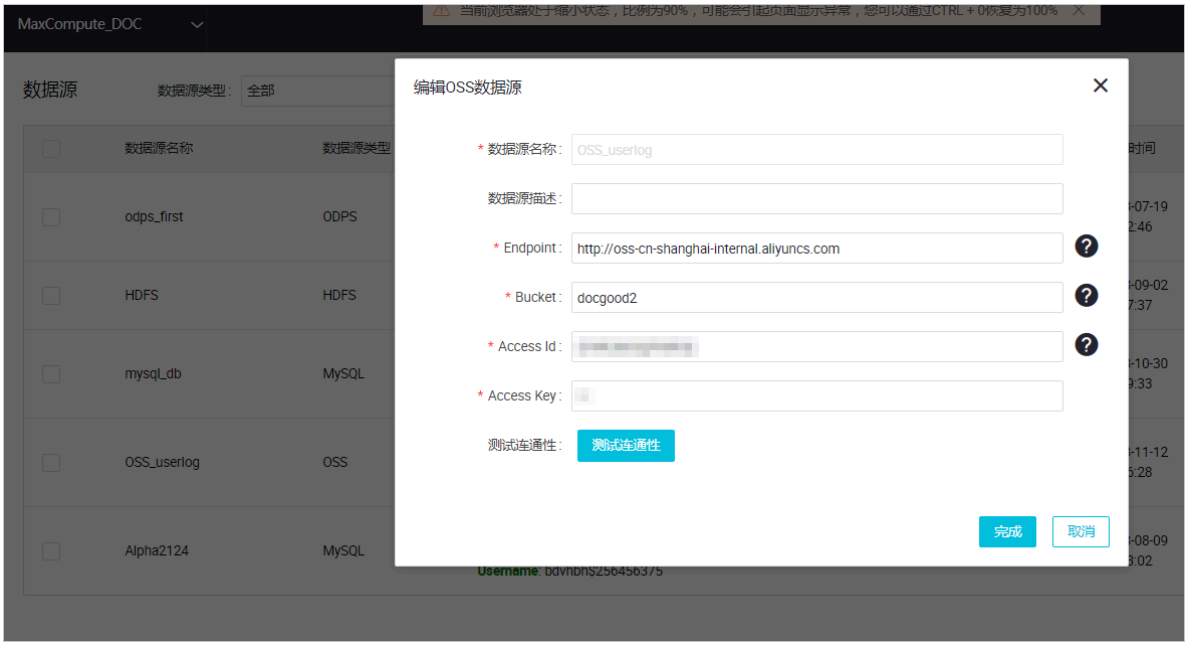
- **步骤1：新增OSS数据源**

  进入DataWorks数据集成控制台，新增OSS类型数据源。



具体参数如下所示，测试数据源连通性通过即可点击完成。Endpoint地址请参见OSS各区域的外网、内网地址，本例中为*http://oss-cn-shanghai.aliyuncs.com*或*http://oss-cn-shanghai-internal.aliyuncs.com*

> **说明:**
> 由于本文中OSS和DataWorks项目处于同一个region中，本文选用后者，通过内网连接。

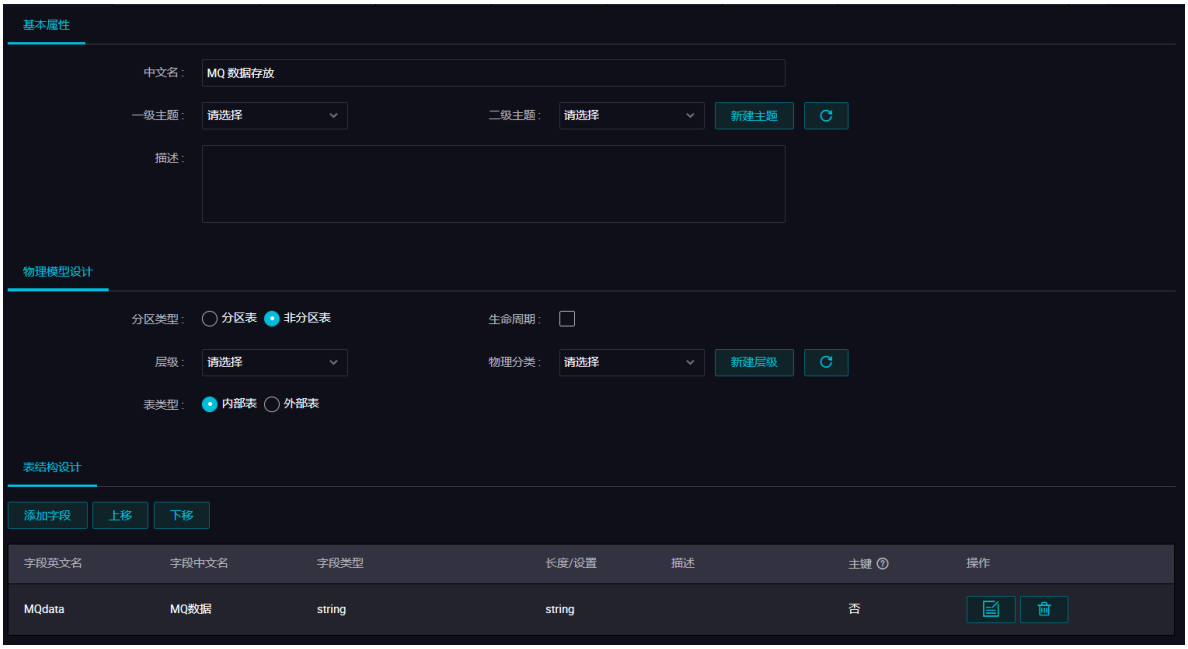


- 步骤2：新建数据同步任务

  在DataWorks上新建数据同步类型节点。

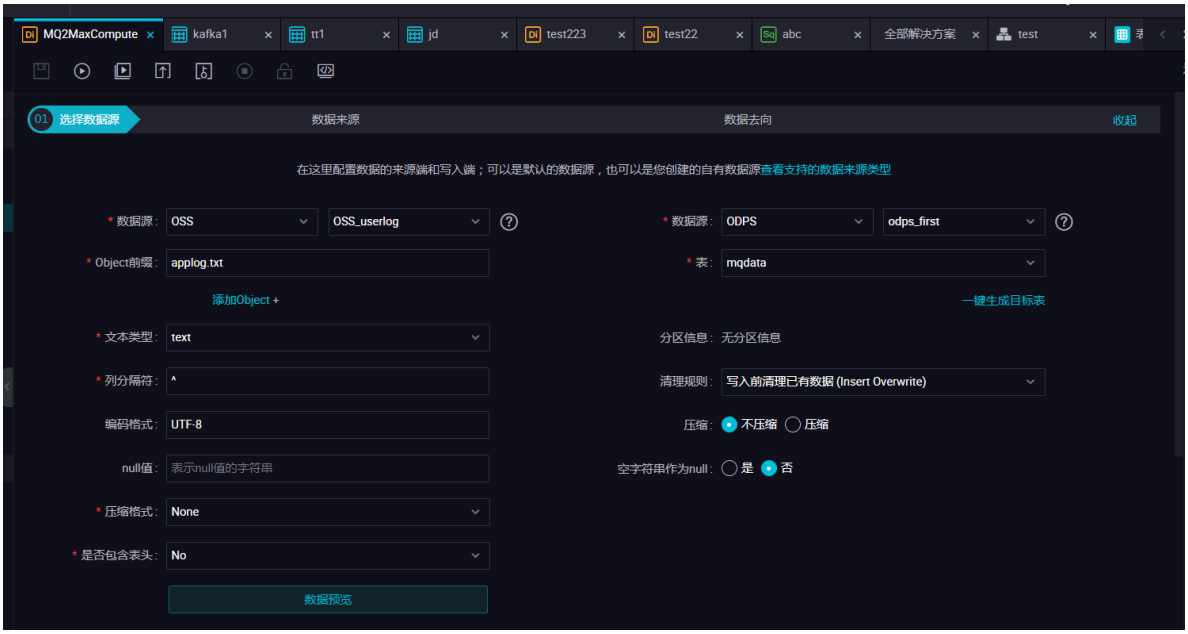
  新建的同时，在DataWorks新建一个建表任务，用于存放JSON数据，本例中新建表名为mqdata。

  表参数可以通过图形化界面完成。本例中mqdata表仅有一列，类型为string，列名为MQ data。



- 步骤3：配置同步任务参数

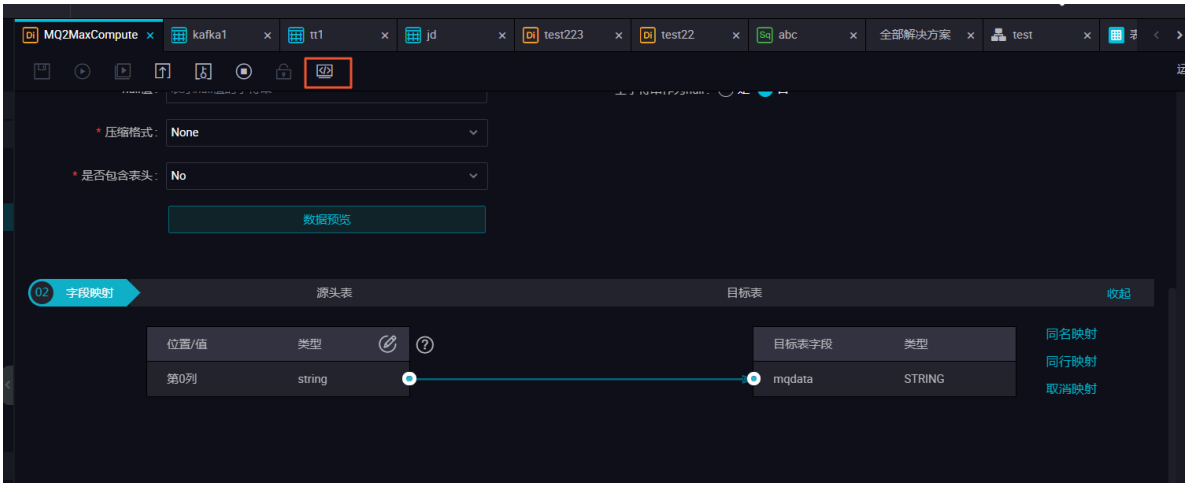
  完成上述新建后，您可以在图形化界面配置数据同步任务参数，如下图所示。选择目标数据源名称为odps_first，选择目标表为刚建立的mqdata。数据来源类型为OSS，Object前缀可填写文件路径及名称。如下图。



说明：

列分隔符使用TXT文件中不存在的字符即可，本文中使用^（对于OSS中的TXT格式数据源，Dataworks支持多字符分隔符，所以您可以使用例如%&%#^$$^%这样很难出现的字符作为列分隔符，保证分割为一列）。

映射方式选择默认的同行映射即可，如下图。



点击左上方的切换脚本按钮，切换为脚本模式。修改fileFormat参数为：`"fileFormat":"binary"`。脚本模式代码示例如下。

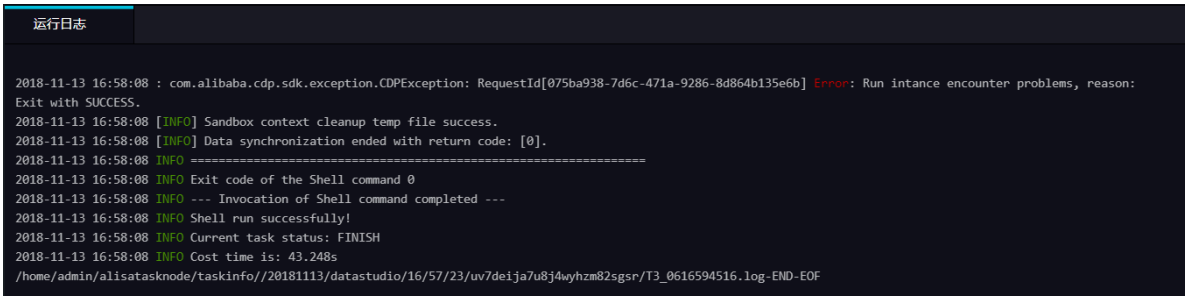
```
{
    "type": "job",
    "steps": [
        {
            "stepType": "oss",
            "parameter": {
                "fieldDelimiterOrigin": "^",
                "nullFormat": "",
                "compress": "",
                "datasource": "OSS_userlog",
                "column": [
                    {
                        "name": 0,
                        "type": "string",
                        "index": 0
                    }
                ],
                "skipHeader": "false",
                "encoding": "UTF-8",
                "fieldDelimiter": "^",
                "fileFormat": "binary",
                "object": [
                    "applog.txt"
                ]
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "odps",
            "parameter": {
                "partition": "",
                "isCompress": false,
```

```
                "truncate": true,
                "datasource": "odps_first",
                "column": [
                    "mqdata"
                ],
                "emptyAsNull": false,
                "table": "mqdata"
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": ""
        },
        "speed": {
            "concurrent": 2,
            "throttle": false,
            "dmu": 1
        }
    }
}
```

说明:

该步骤可以保证OSS中的JSON文件同步到MaxCompute之后存在同一行数据中，即为一个字段，其他参数保持不变。

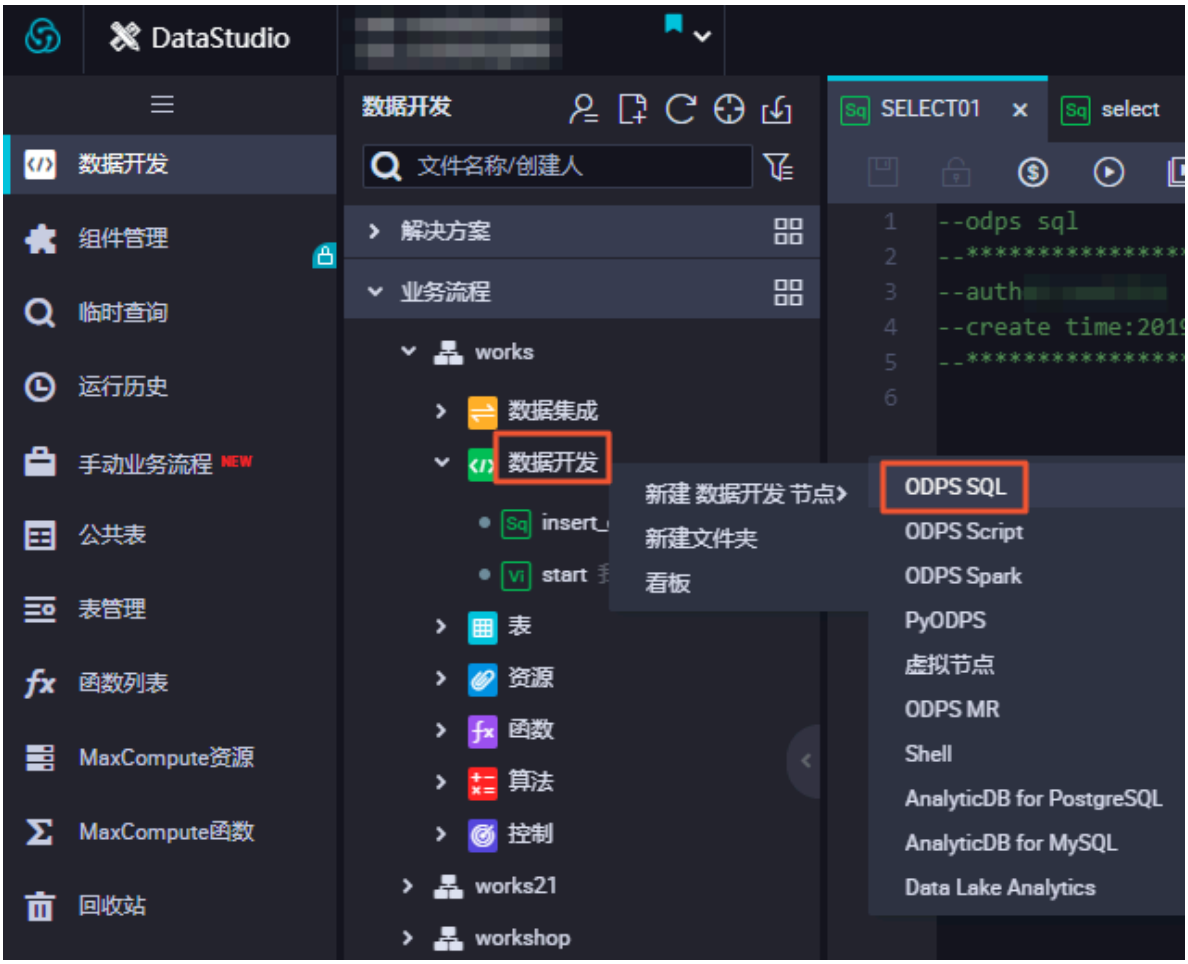
完成上述配置后，点击运行接即可。运行成功日志示例如下所示。



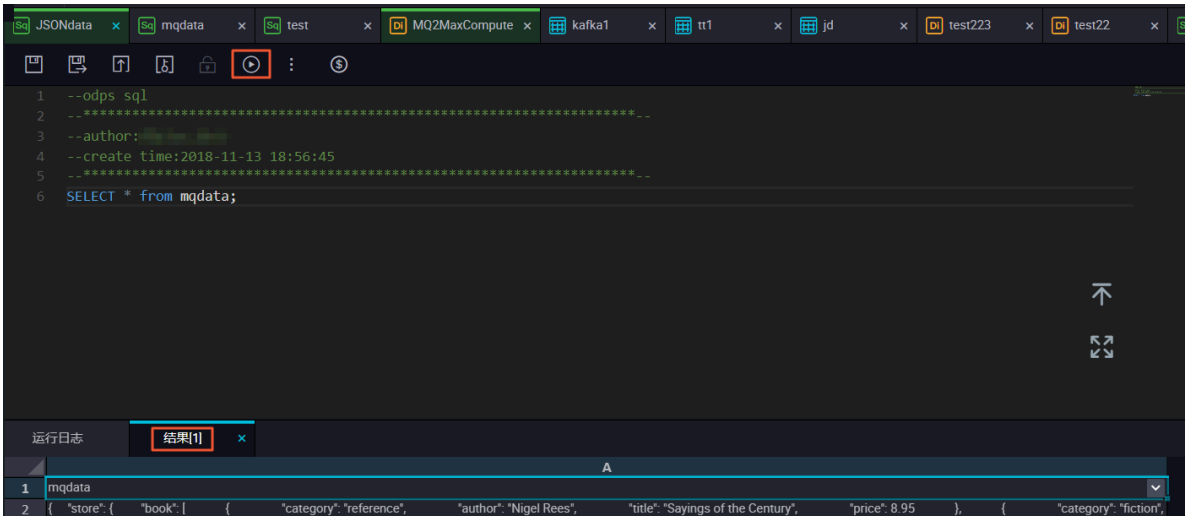
结果验证

- 获取JSON字段信息

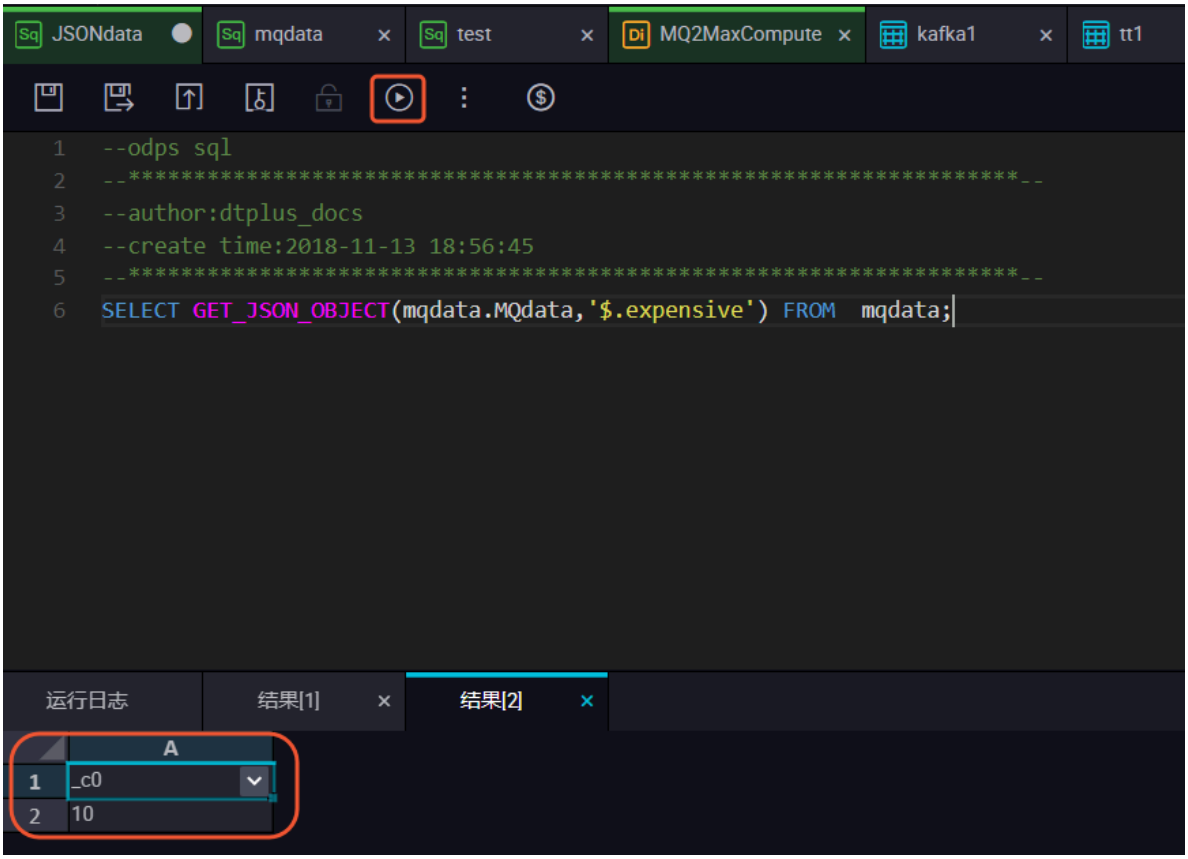
  在您的业务流程中新建一个ODPS SQL节点。



  您可以首先输入`SELECT*from mqdata;`语句，查看当前mqdata表中数据。这一步及后续步骤，也可以直接在MaxCompute客户端中输入命令运行，如下图。



确认导入表中的数据结果无误后，使用`SELECT GET_JSON_OBJECT(mqdata.MQdata,'$.expensive') FROM mqdata;`获取JSON文件中的`expensive`值，如下图所示。



更多信息

在进行迁移后结果验证时，您可以使用MaxCompute内建字符串函数GET_JSON_OBJECT获取您想要的JSON数据。

8.5 JSON数据从MongoDB迁移到MaxCompute最佳实践

本文为您介绍如何利用DataWorks的数据集成功能，将从MongoDB提取的JSON字段迁移到MaxCompute的最佳实践。

准备工作

1. 账号准备

在数据库内新建用户，用于DataWorks添加数据源。本例中使用命令`db.createUser({user:"bookuser",pwd:"123456",roles:["root"]})`，新建用户名为`bookuser`，密码为`123456`，权限为`root`。

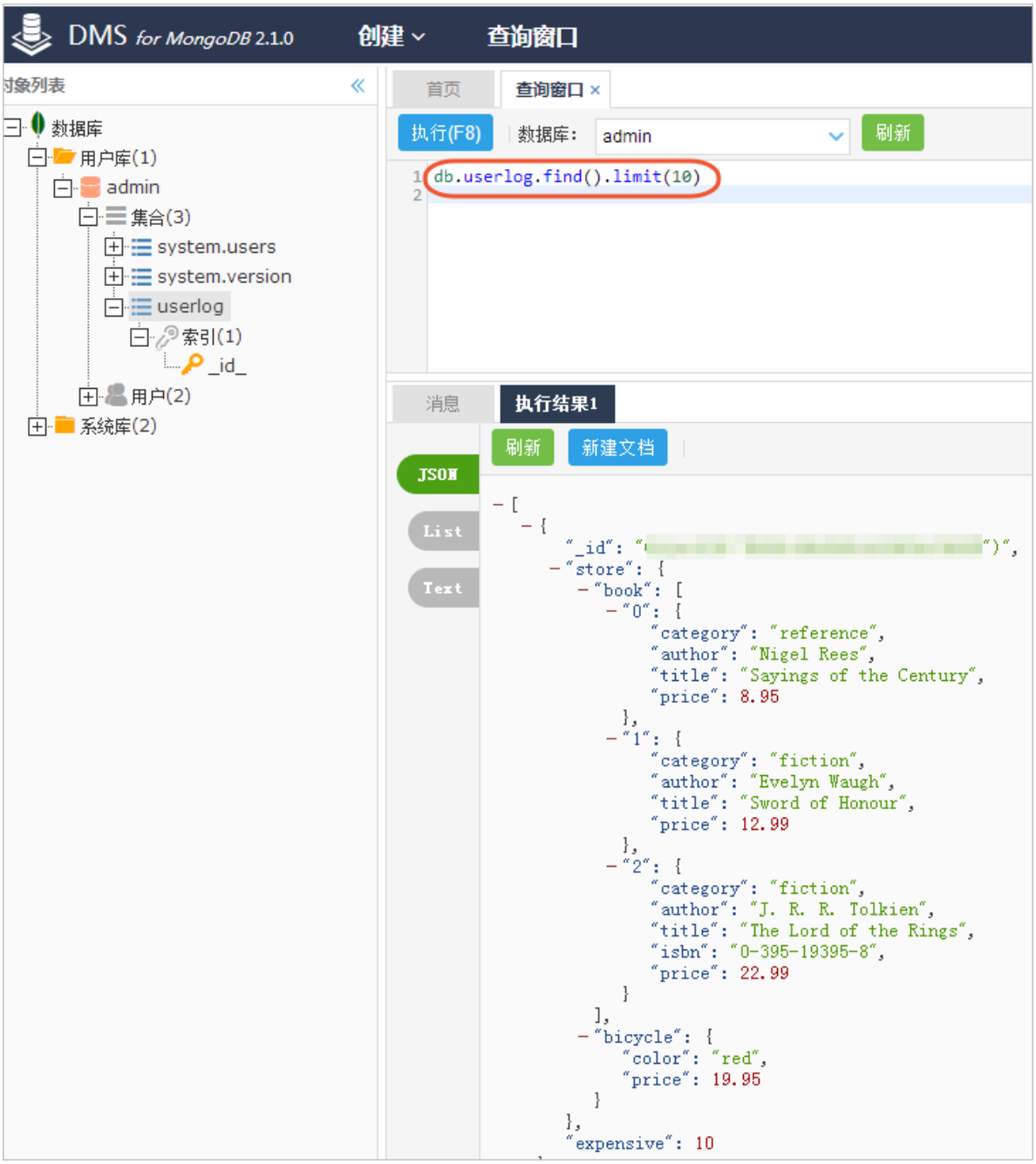
2. 数据准备

首先您需要将数据上传至您的MongoDB数据库。本例中使用阿里云的云数据库 MongoDB版，网络类型为VPC（需申请公网地址，否则无法与DataWorks默认资源组互通），测试数据如下。

```
{
    "store": {
        "book": [
             {
                "category": "reference",
                "author": "Nigel Rees",
                "title": "Sayings of the Century",
                "price": 8.95
             },
             {
                "category": "fiction",
                "author": "Evelyn Waugh",
                "title": "Sword of Honour",
                "price": 12.99
             },
             {
                 "category": "fiction",
                 "author": "J. R. R. Tolkien",
                 "title": "The Lord of the Rings",
                 "isbn": "0-395-19395-8",
                 "price": 22.99
             }
         ],
         "bicycle": {
             "color": "red",
             "price": 19.95
         }
    },
    "expensive": 10
```

```
}
```

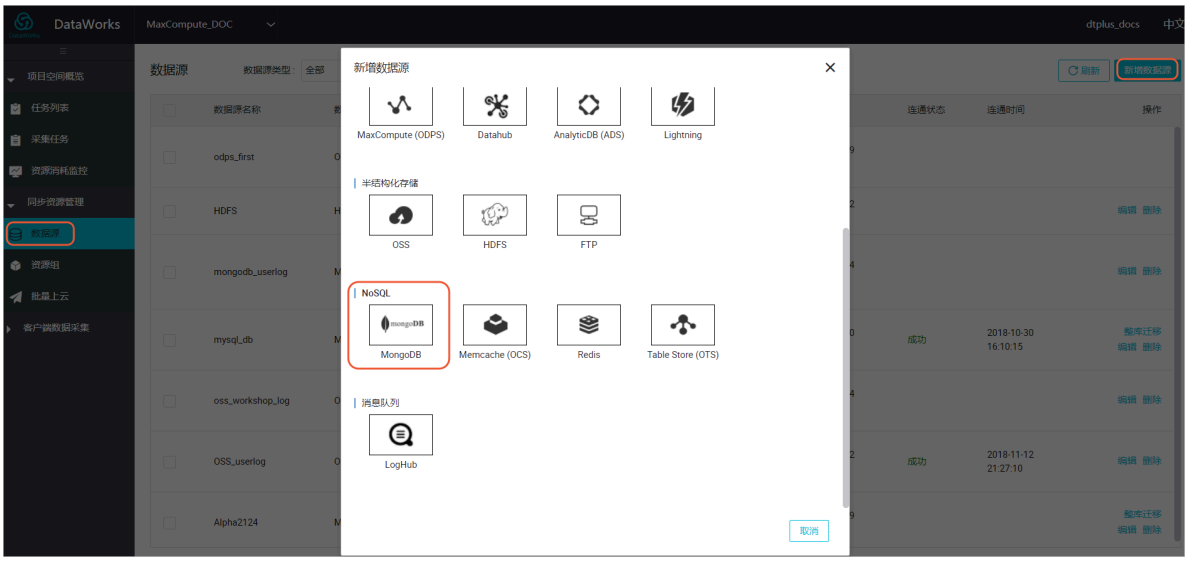
登录MongoDB的DMS控制台，本例中使用的数据库为admin，集合为userlog，您可以在查询窗口使用db.userlog.find().limit(10)命令查看已上传好的数据，如下图所示。



使用DataWorks提取数据到MaxCompute

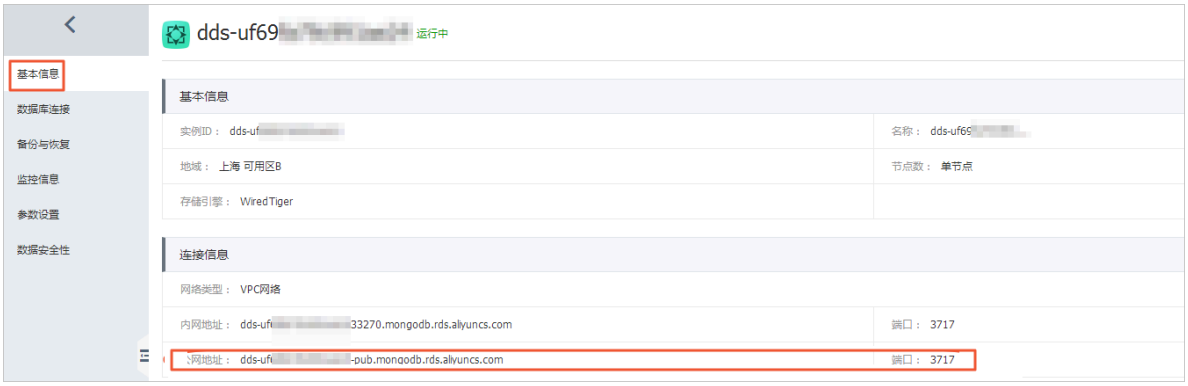
- 步骤1：新增MongoDB数据源

  进入DataWorks数据集成控制台，新增MongoDB类型数据源。



  具体参数如下所示，测试数据源连通性通过即可点击完成。由于本文中MongoDB处于VPC环境下，因此数据源类型需选择有公网IP。

访问地址及端口号可通过在MongoDB管理控制台点击实例名称获取，如下图所示。

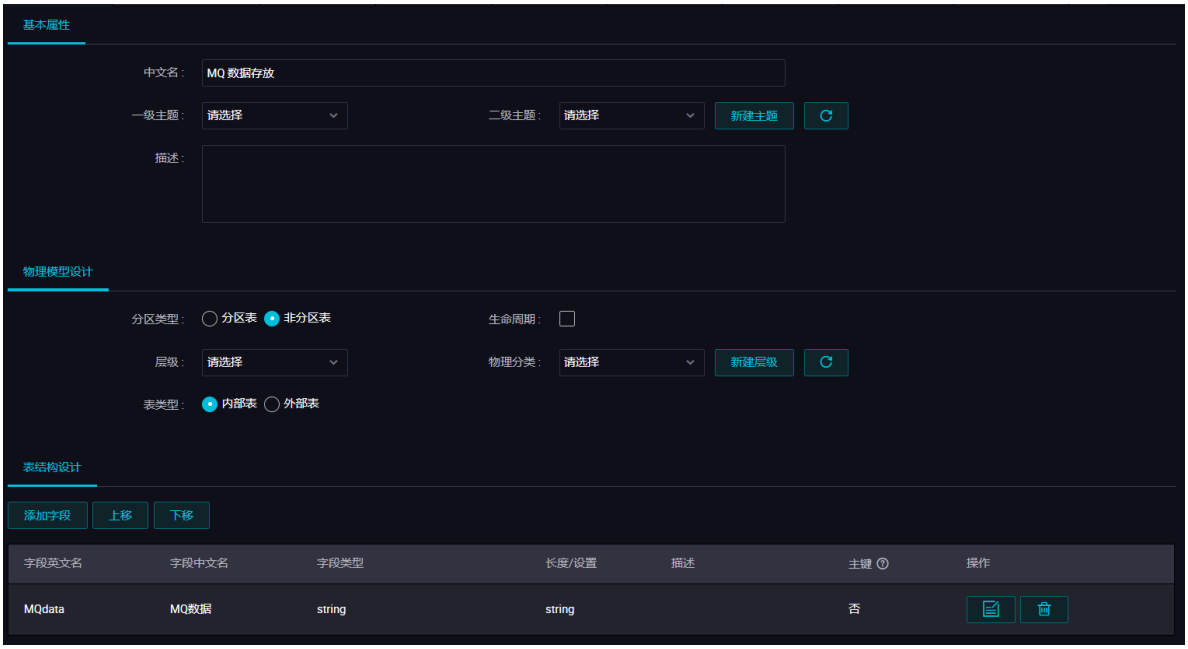


- 步骤2：新建数据同步任务

  在DataWorks上新建数据同步类型节点。

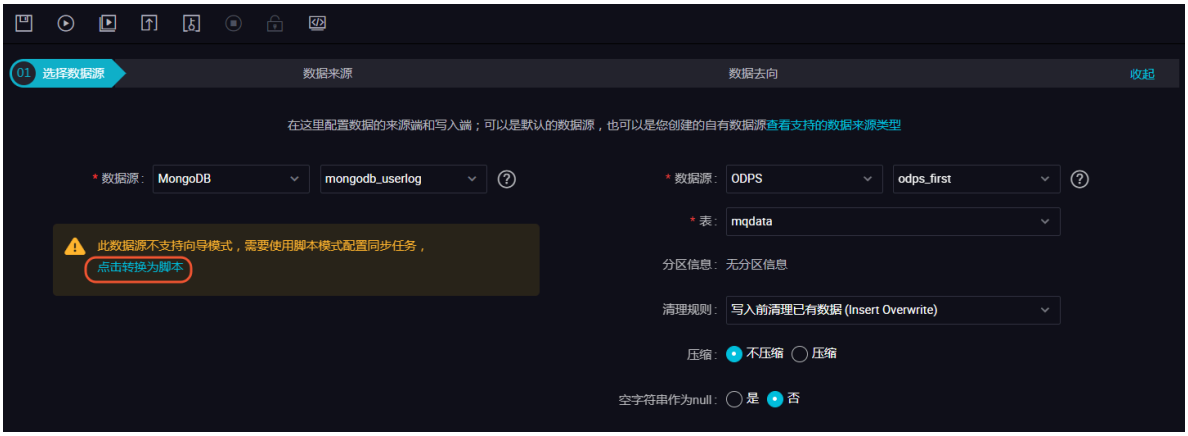
  新建的同时，在DataWorks新建一个建表任务，用于存放JSON数据，本例中新建表名为mqdata。

  表参数可以通过图形化界面完成。本例中mqdata表仅有一列，类型为string，列名为MQ data。



- 步骤3：参数配置

  完成上述新建后，您可以在图形化界面进行数据同步任务参数的初步配置。选择目标数据源名称为odps_first，选择目标表为新建的mqdata。数据来源类型为MongoDB，选择新建立的据源mongodb_userlog。完成上述配置后，点击转换为脚本，跳转到脚本模式，如下图。

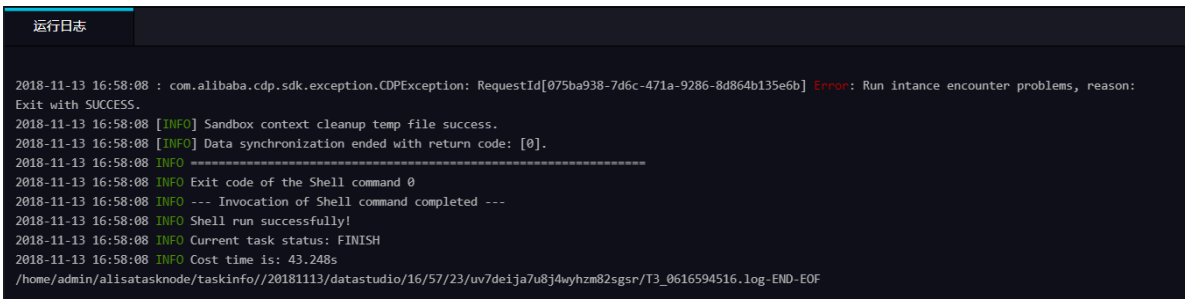


  脚本模式代码示例如下。

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "mongodb",
            "parameter": {
                "datasource": "mongodb_userlog",
 //数据源名称
                "column": [
                    {
                        "name": "store.bicycle.color", //JSON字段路
径，本例中提取color值
```

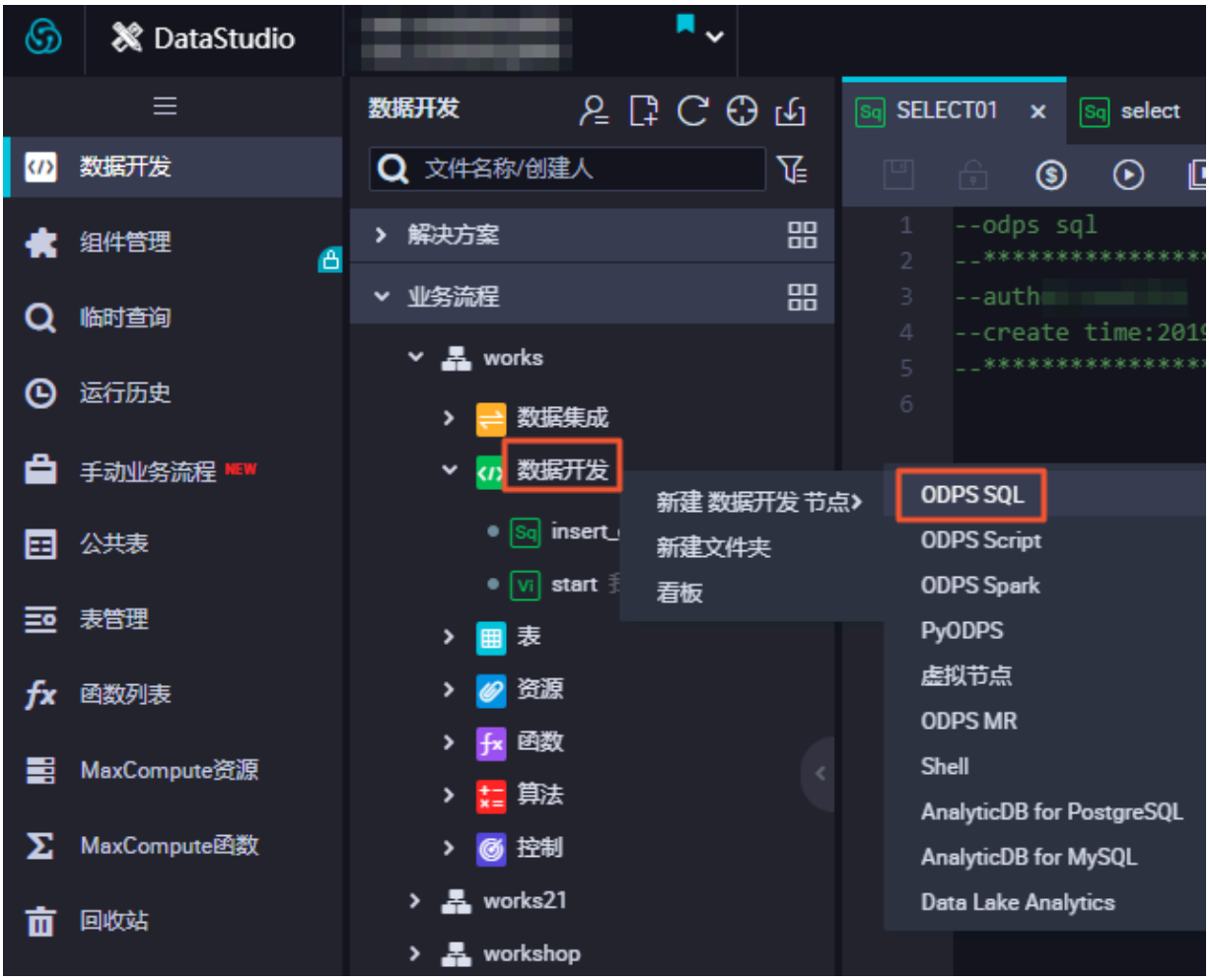
```
                                "type": "document.document.string" //本栏目的
字段数需和name一致。假如您选取的JSON字段为一级字段，如本例中的expensive，则直
接填写string即可。
                            }
                        ],
                        "collectionName //集合名称": "userlog"
                    },
                    "name": "Reader",
                    "category": "reader"
                },
                {
                    "stepType": "odps",
                    "parameter": {
                        "partition": "",
                        "isCompress": false,
                        "truncate": true,
                        "datasource": "odps_first",
                        "column": [
                                    "mqdata"  //MaxCompute表列名
                        ],
                        "emptyAsNull": false,
                        "table": "mqdata"
                    },
                    "name": "Writer",
                    "category": "writer"
                }
            ],
            "version": "2.0",
            "order": {
                "hops": [
                    {
                        "from": "Reader",
                        "to": "Writer"
                    }
                ]
            },
            "setting": {
                "errorLimit": {
                    "record": ""
                },
                "speed": {
                    "concurrent": 2,
                    "throttle": false,
                    "dmu": 1
                }
            }
        }
```

完成上述配置后，点击运行接即可。运行成功日志示例如下所示。

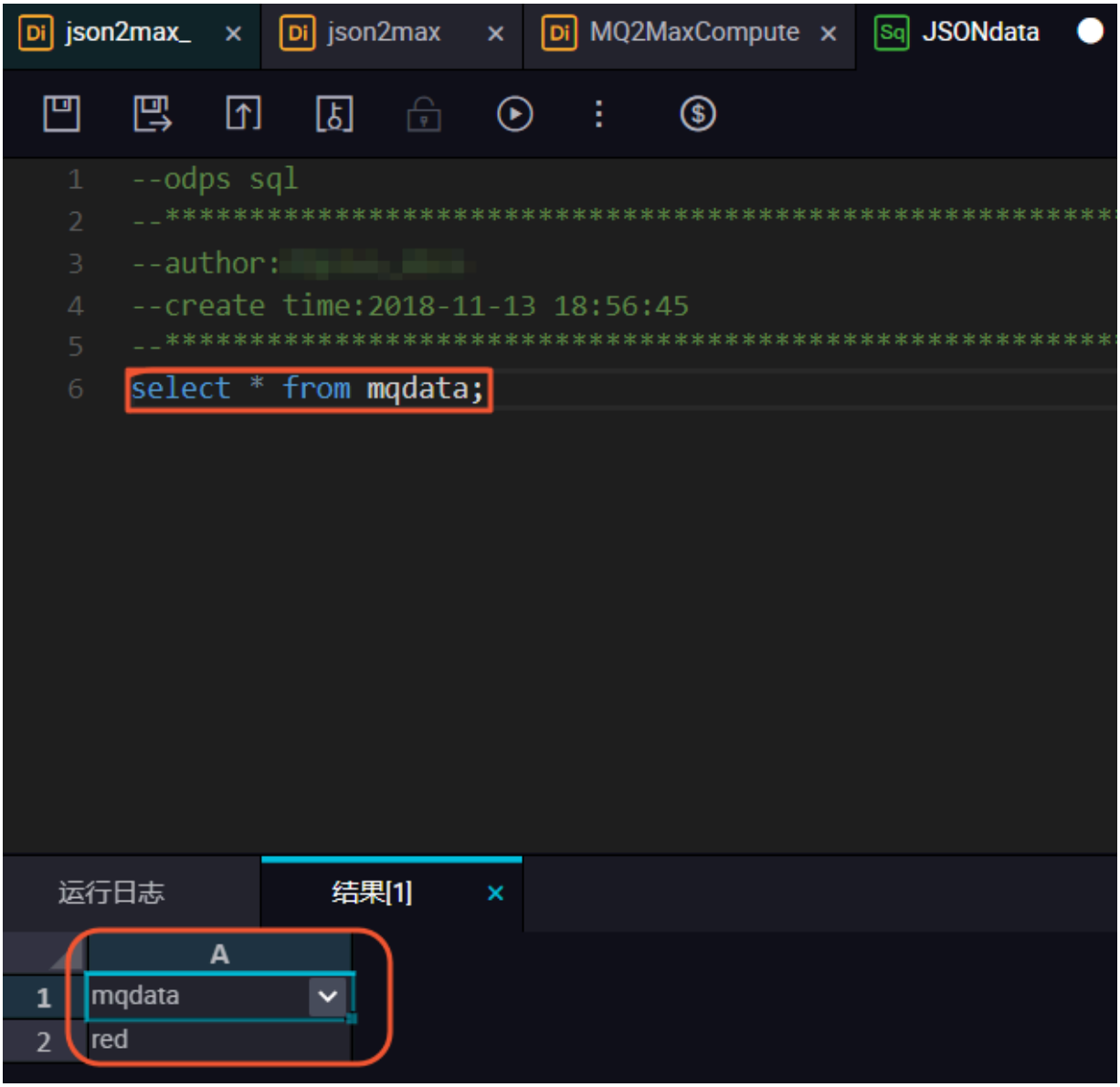


结果验证

在您的业务流程中新建一个ODPS SQL节点，如下图。



您可以输入`SELECT * from mqdata;`语句，查看当前mqdata表中数据。这一步您也可以直接在MaxCompute客户端中输入命令运行，如下图。


json2max_
json2max
MQ2MaxCompute
JSONdata
--odps sql
--author:
--create time:2018-11-13 18:56:45
select * from mqdata;
运行日志
结果[1]
A
1
mqdata
2
red