

数据实时同步到Elasticsearch 解决方案分享

阿里云 王明亚(云时) 2021.07.10

目录

CONTENTS

1

阿里云数据集成简介

2

离线实时同步技术原理与架构

3

数据同步到Elasticsearch解决方案

4

典型案例分享

为什么需要数据集成？

数据集成、同步需求场景

对接业务数据库初始化数仓

用户需要把自己的云下数据快速安全的迁移到云上存储并做进一步的业务分析，如线下mysql、oracle到云上Maxcompute
传统数仓迁移上云
一次性全量和持续增量



流式数据汇聚

批流数据汇聚
IOT数据汇总分析
各种Log数据流收集

云上各个产品之间数据同步

从AWS、GCP到阿里云
阿里云内部RDS、MaxCompute、Hologres、PolarDB、AnalyticDB、Elasticsearch等之间的数据同步



云上计算数据回流存储数据库

用户的原始数据需要转移存储系统存储，做容灾或者长期备份，比如MaxCompute上的数据可以写入OSS为文件，并可使用压缩节省存储空间

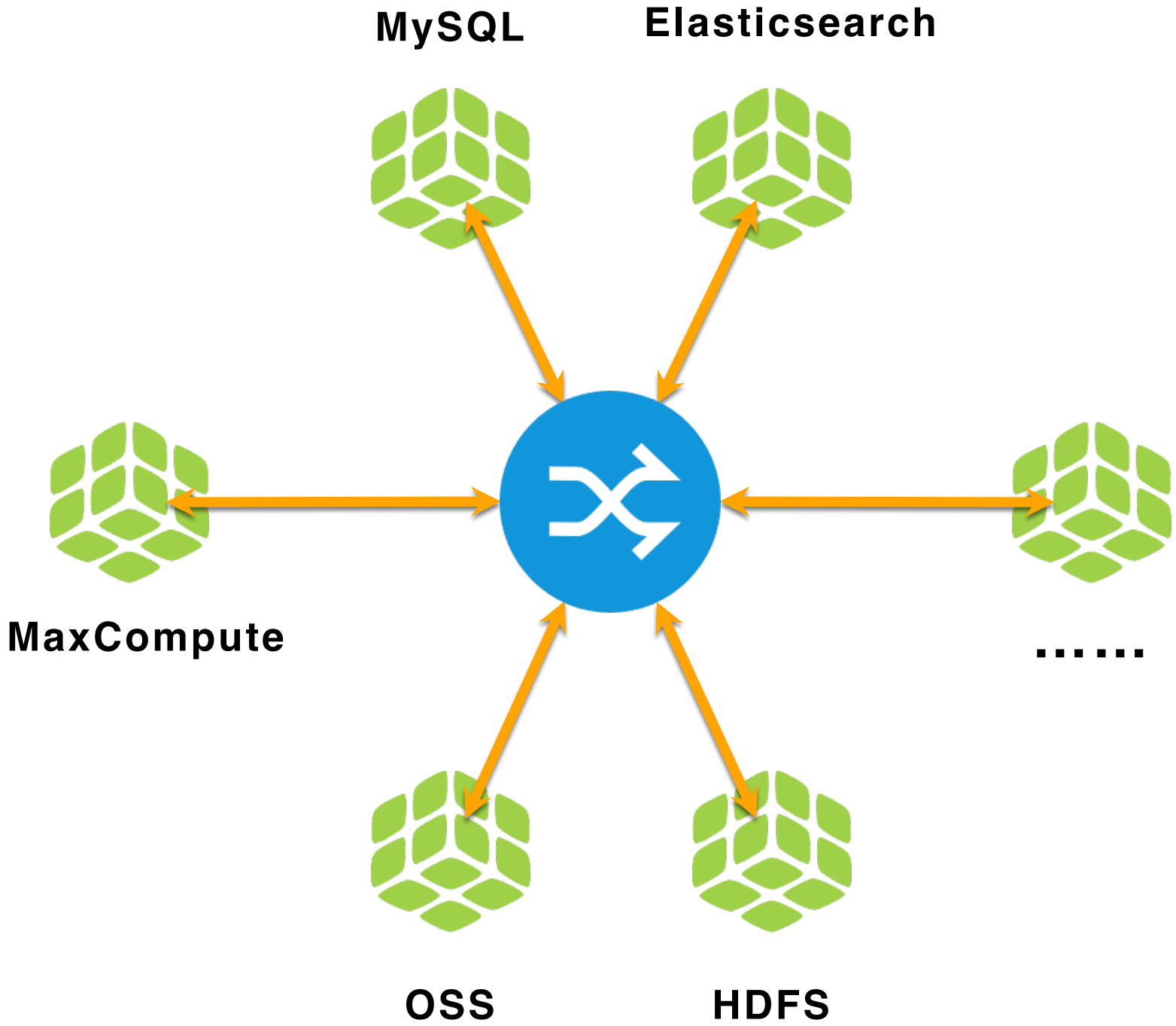
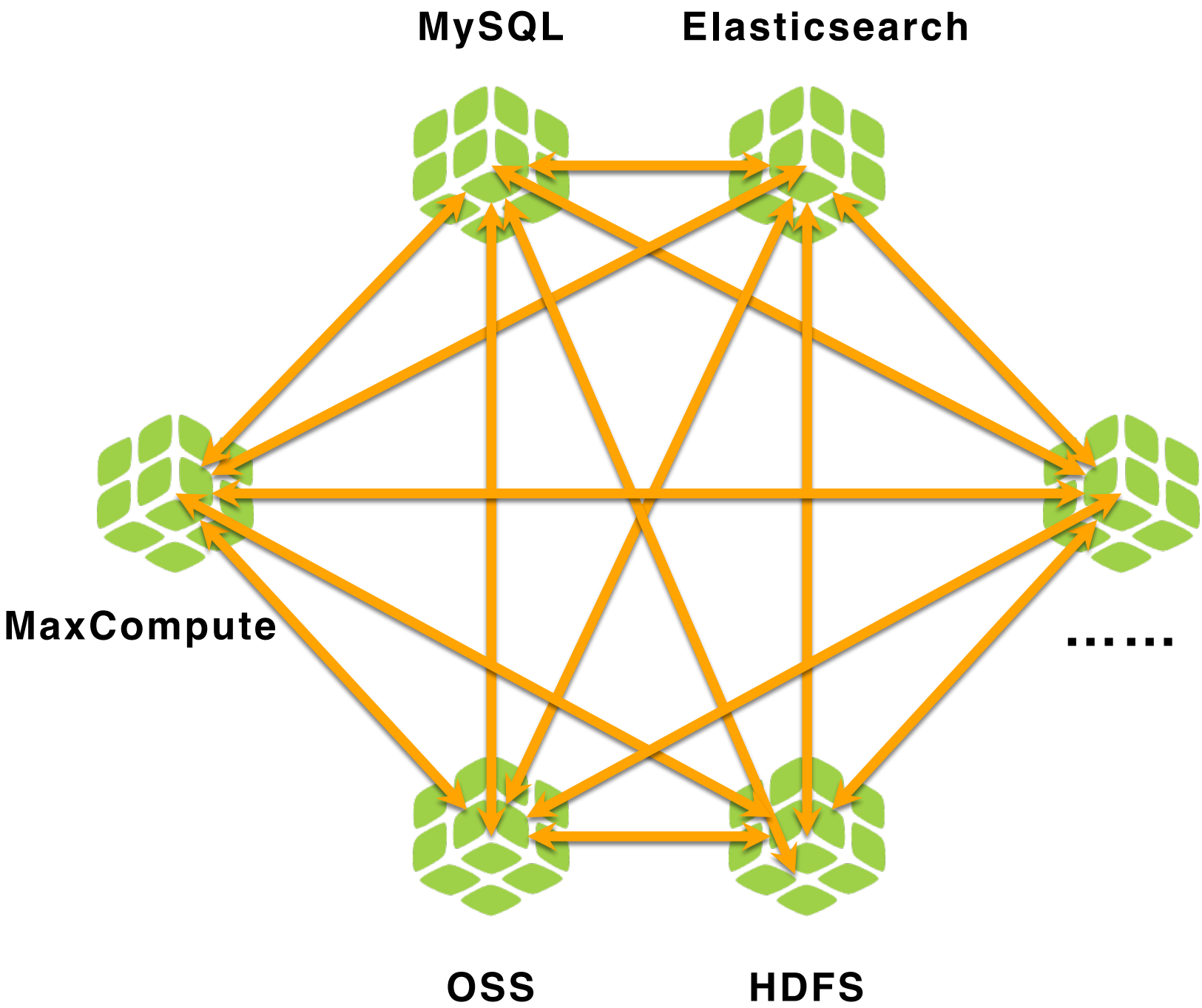
阿里云数据集成产品定位

数据上云的核心枢纽：异构数据存储、可靠、安全、低成本、可弹性扩展的数据同步平台



阿里云数据集成DataX设计核心思路

极具可扩展性

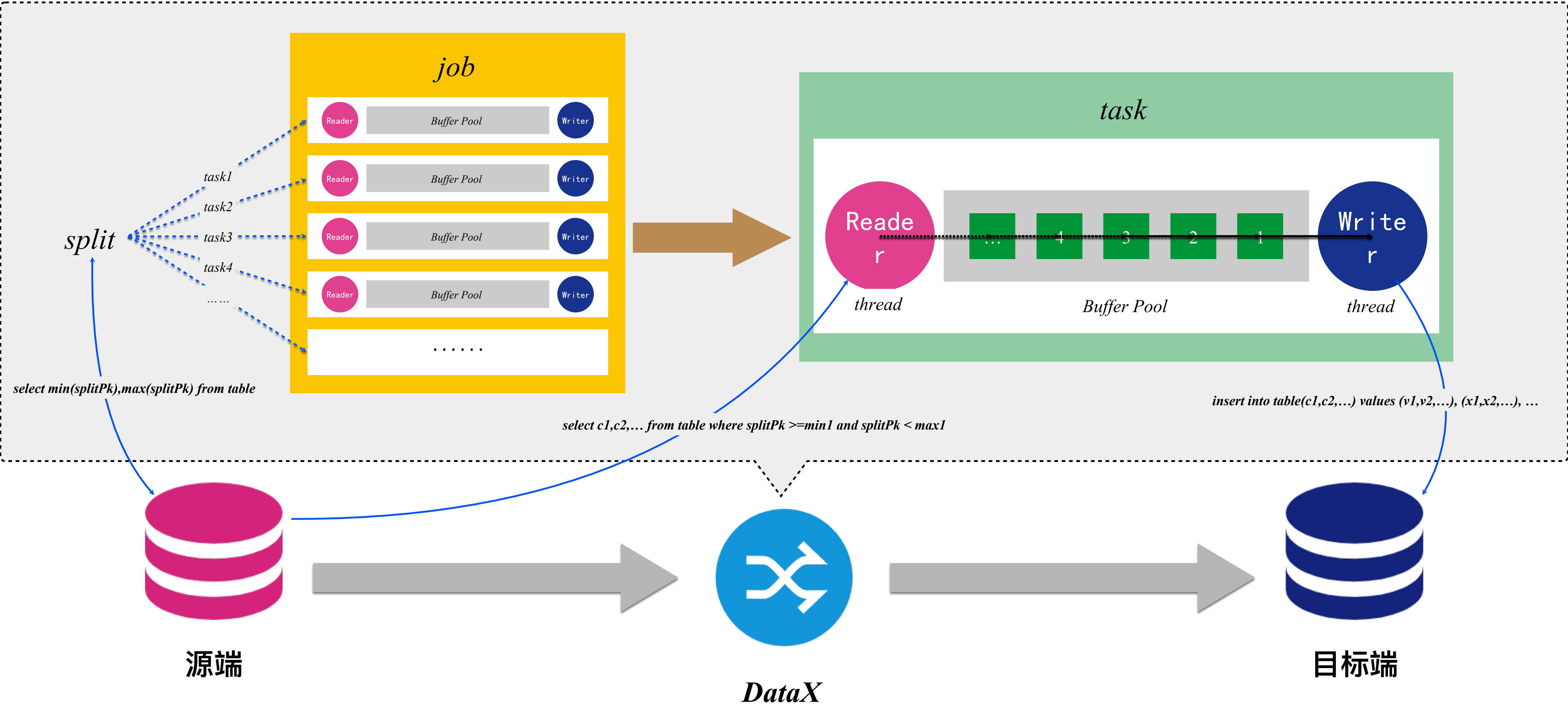


数据集成同步引擎

开源部分 <https://github.com/alibaba/DataX>

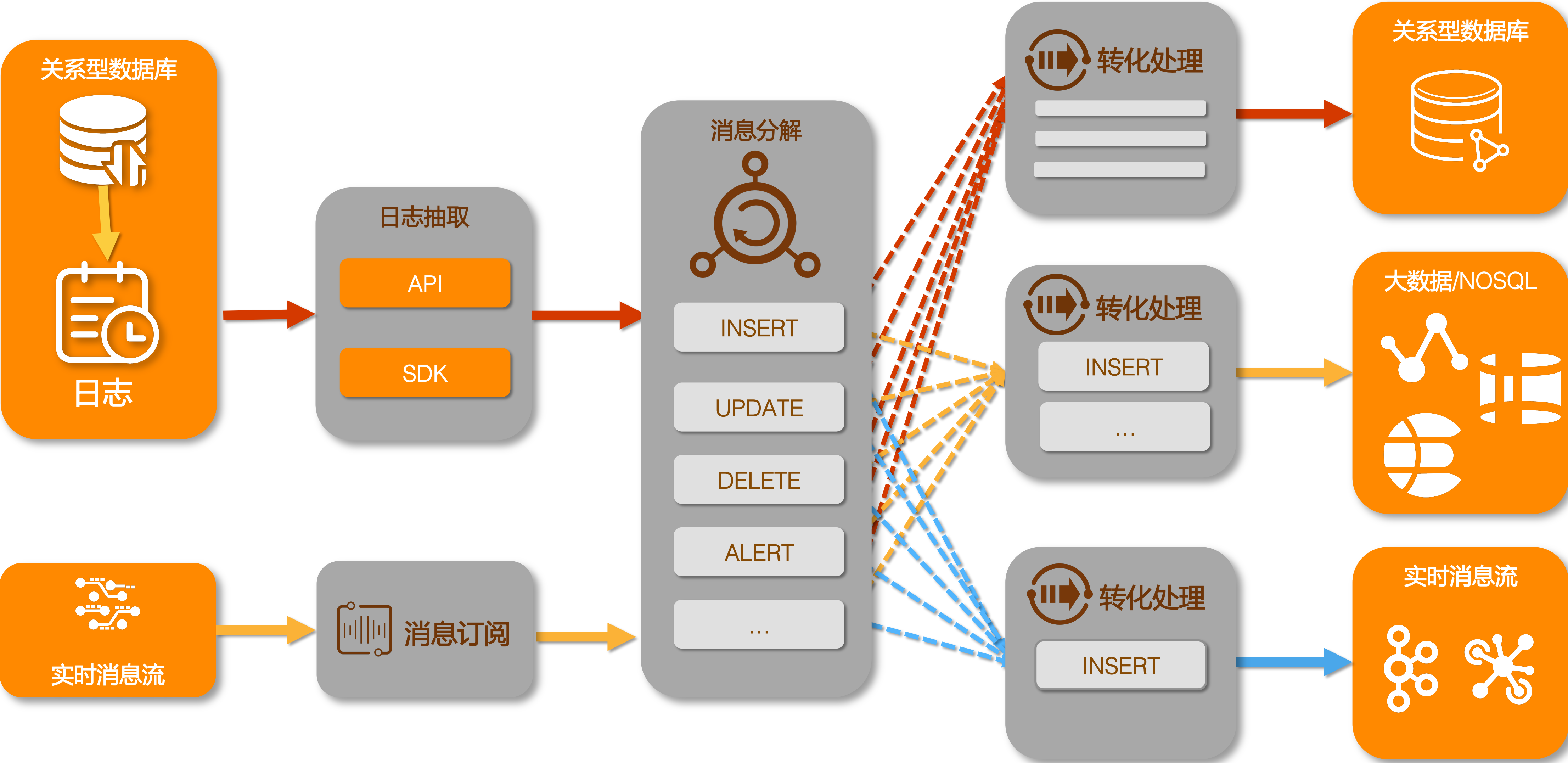
离线数据同步原理

基于数据库JDBC协议或数据源SDK进行读写



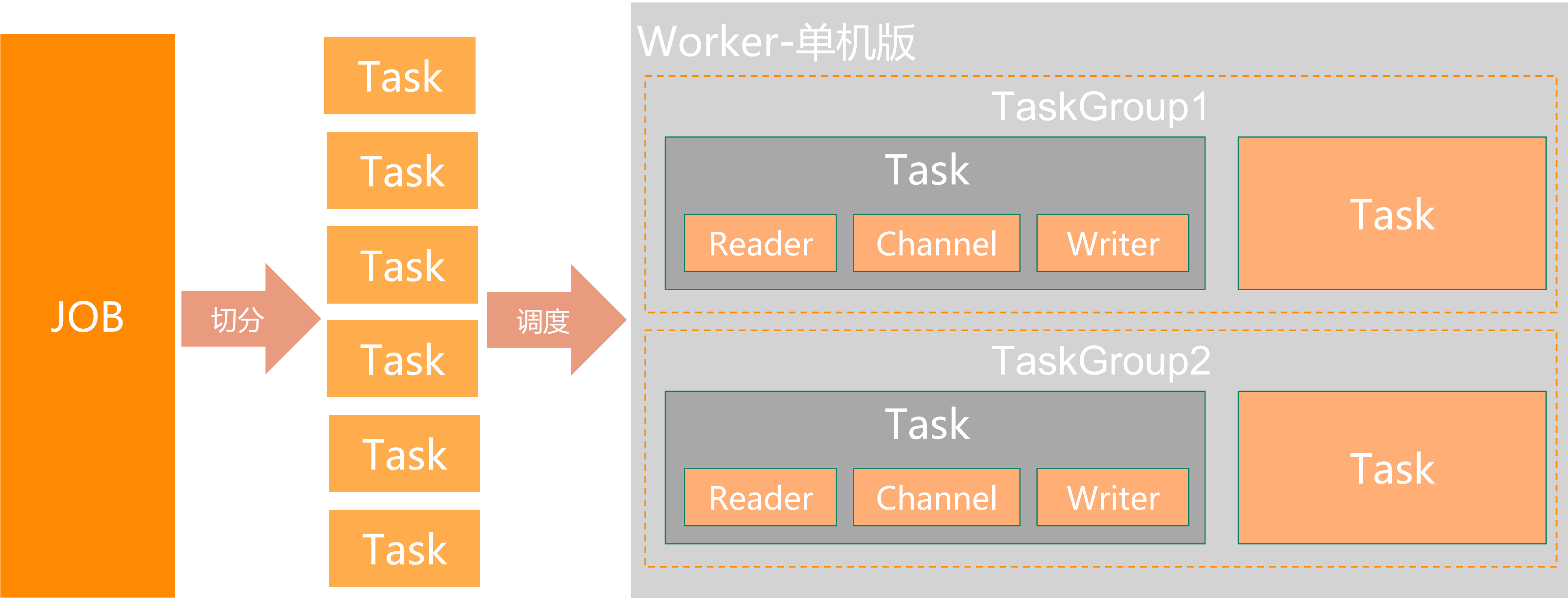
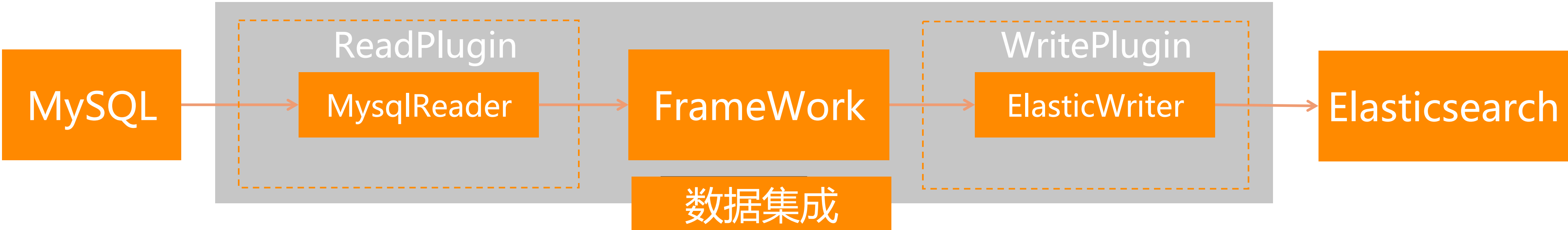
实时数据同步原理

基于数据库日志、cdc、或数据源SDK进行读写



阿里云数据集成DataX架构

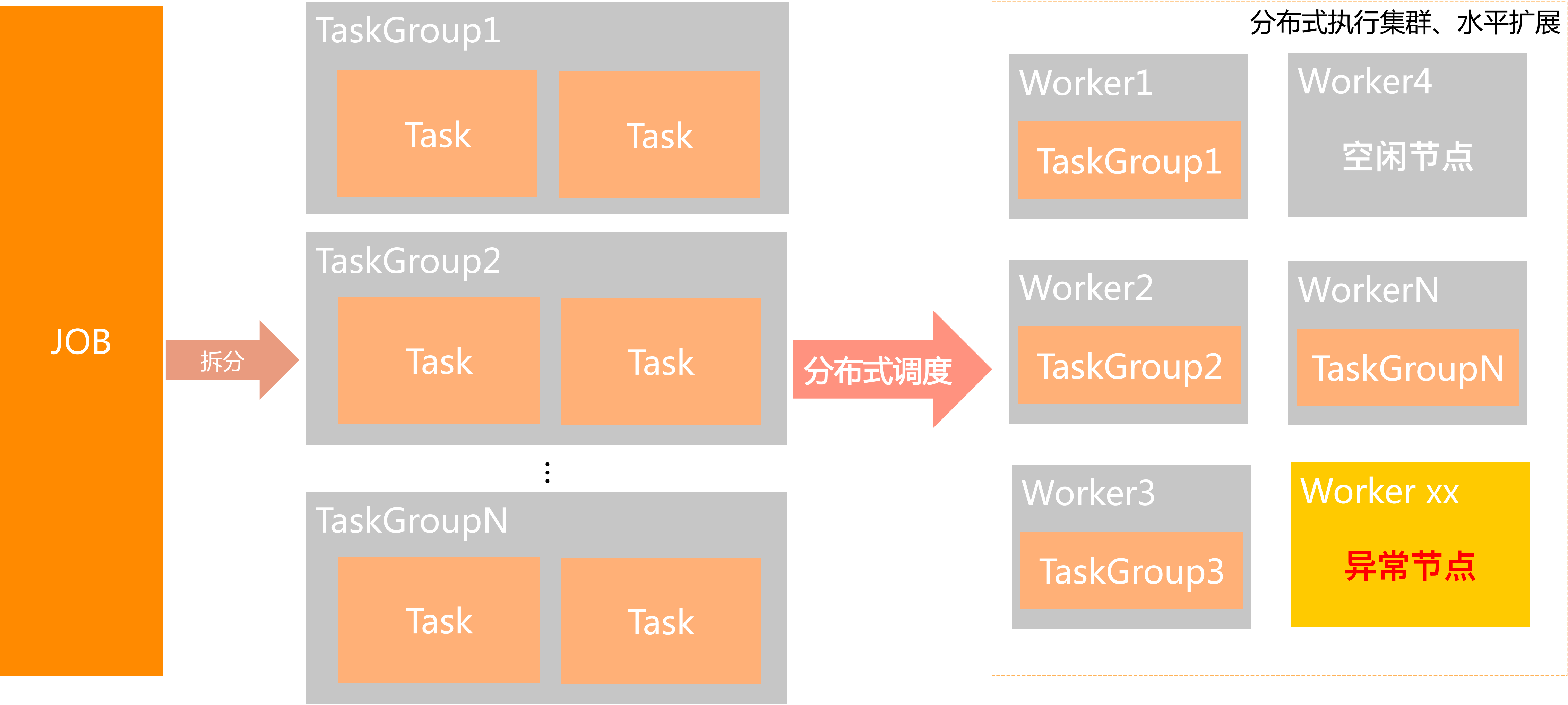
Standalone 模式



阿里云数据集成DataX架构

Distribute模式（商业化版本能力）

处理海量数据同步

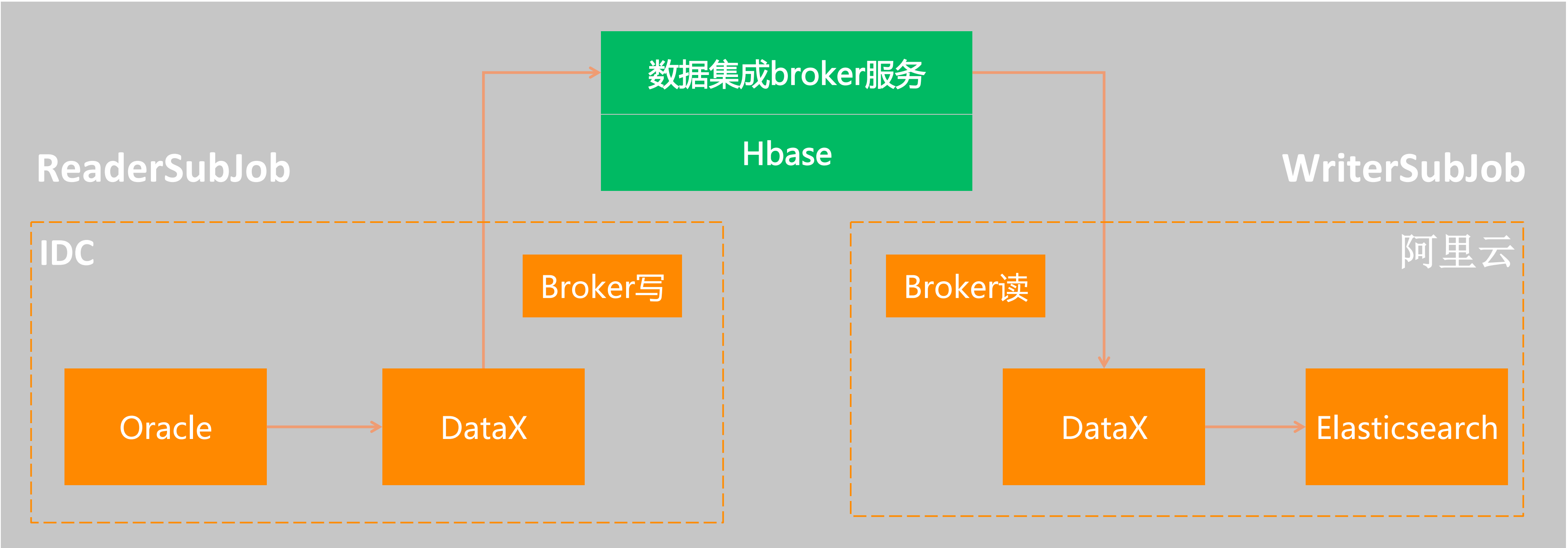


阿里云数据集成DataX架构

Broker模式（商业化版本能力）

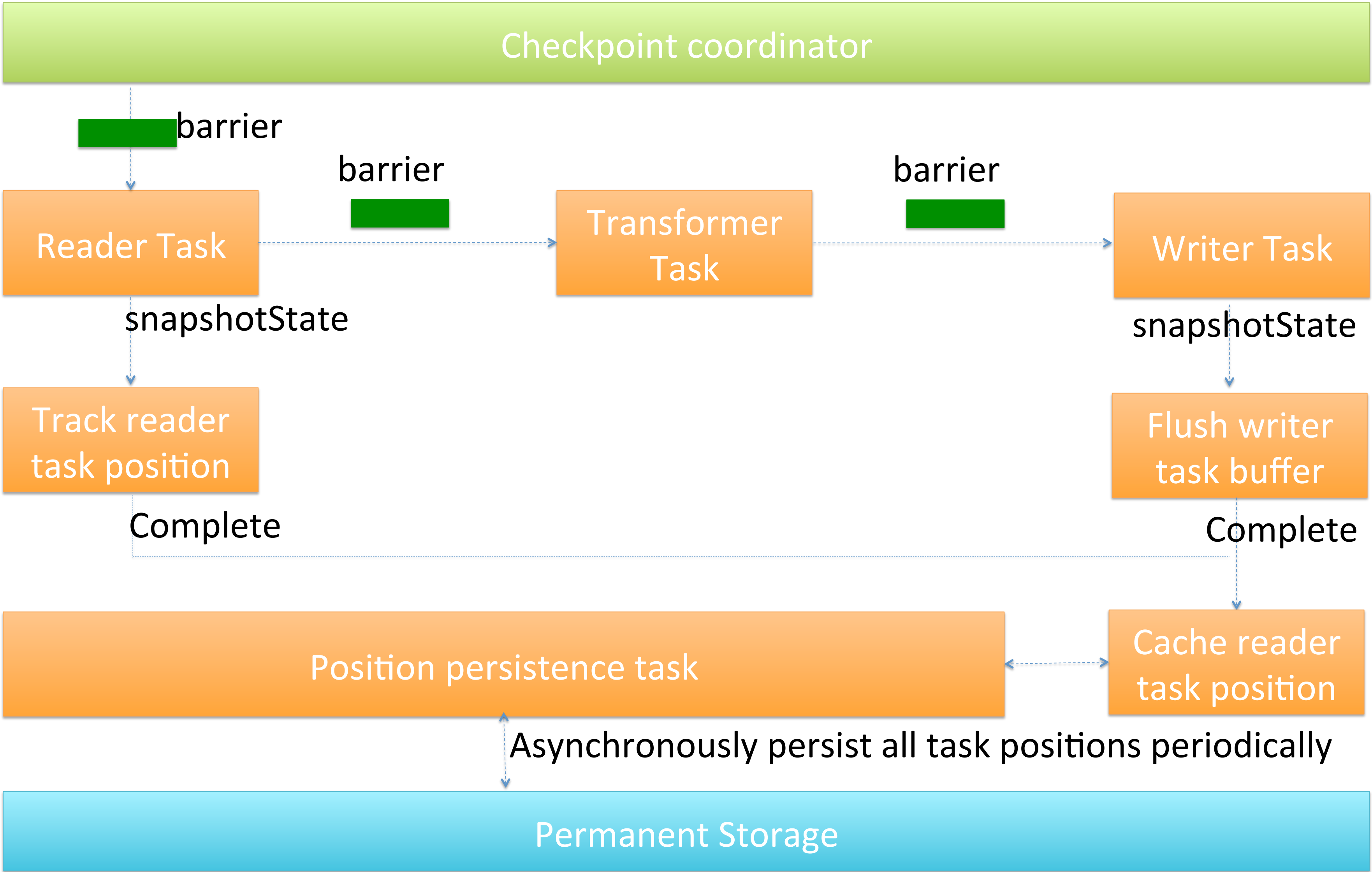


- 加密、压缩
- 对上层用户透明
- 复用已有DataX功能体系，DRY



阿里云数据集成DataX架构

实时同步 Checkpoint 机制（商业化版本能力）



阿里云数据集成实时同步-核心亮点

低成本

- 数据集成独享资源组成本优势

云原生基因

- 与阿里云产品(MaxCompute, Hologres, DataHub, RDS, PolarDB, Oceanbase等)融合度较高, 有天然的云原生基因

无侵入&无插件

- 对数据库本身无侵入, 无需在数据库侧安装Agent, 不需要提供堡垒机。

丰富多样的数据源

- 支持星型链路组合, 任何一种输入源都可以和任何一种输出源搭配组成同步链路。
 - 支持一个任务一读多写多路输出

支持断点续传

读取

MySQL、Oracle、SQLSever
OceanBase、Kafka、LogHub
DataHub、PolarDB

MaxCompute
Hologres、Datahub、Kafka、
ElasticSearch等

写入

插件化机制

- 插件化机制, 对新的数据源支持扩展能力强

轻松监控运维告警

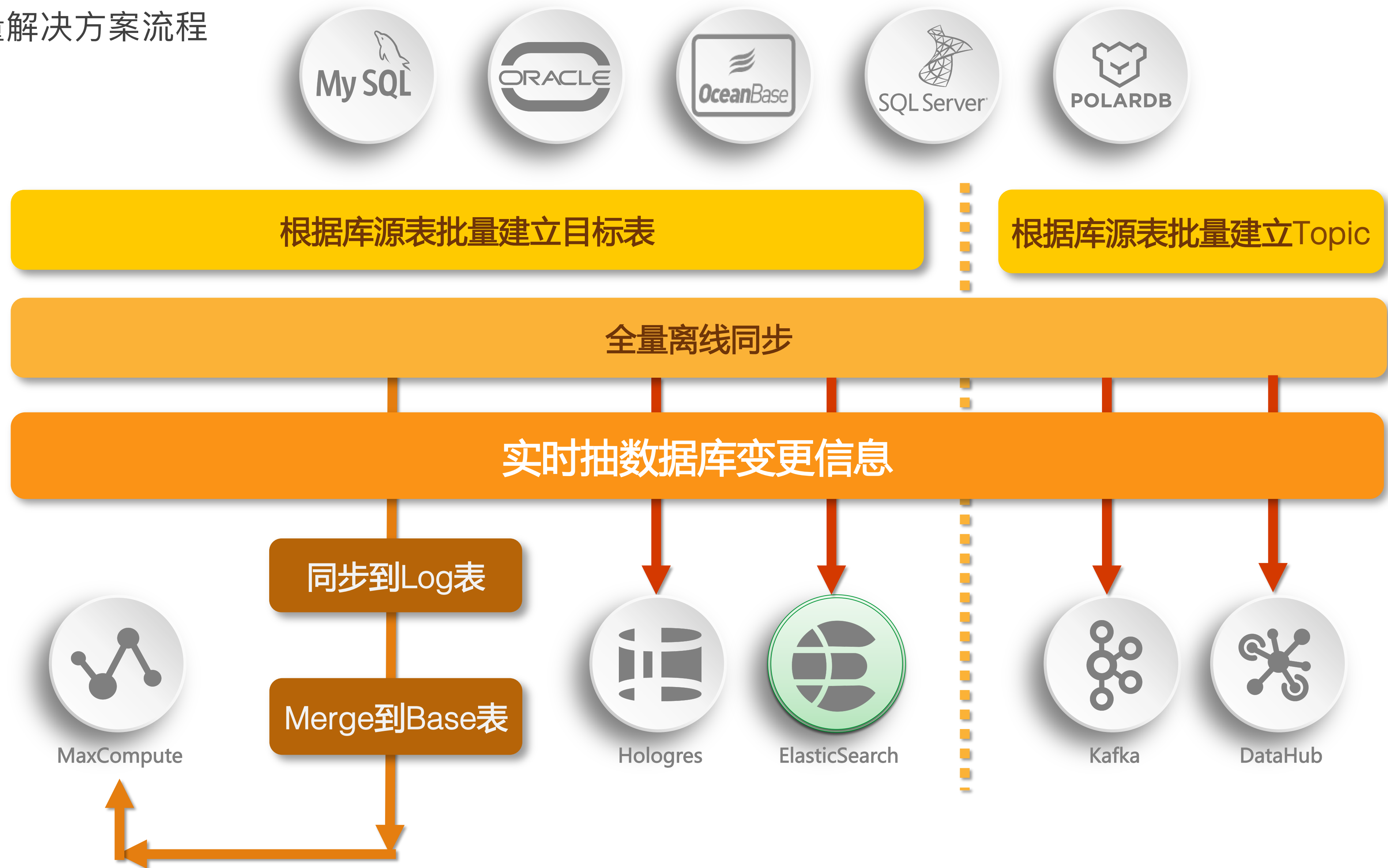
- 提供运维大盘、监控报警、FailOver等运维能力
- 可以监控业务延迟、Failover、脏数据、心跳检查、失败信息
- 支持邮件、电话、钉钉通知

一站式解决方案系统

- 常见数据源整库全增量到MaxCompute、Hologres、Elasticsearch、Datahub等
- 支持分库分表、单表、整库多表、DDL消息

阿里云数据集成解决方案系统

全增量解决方案流程

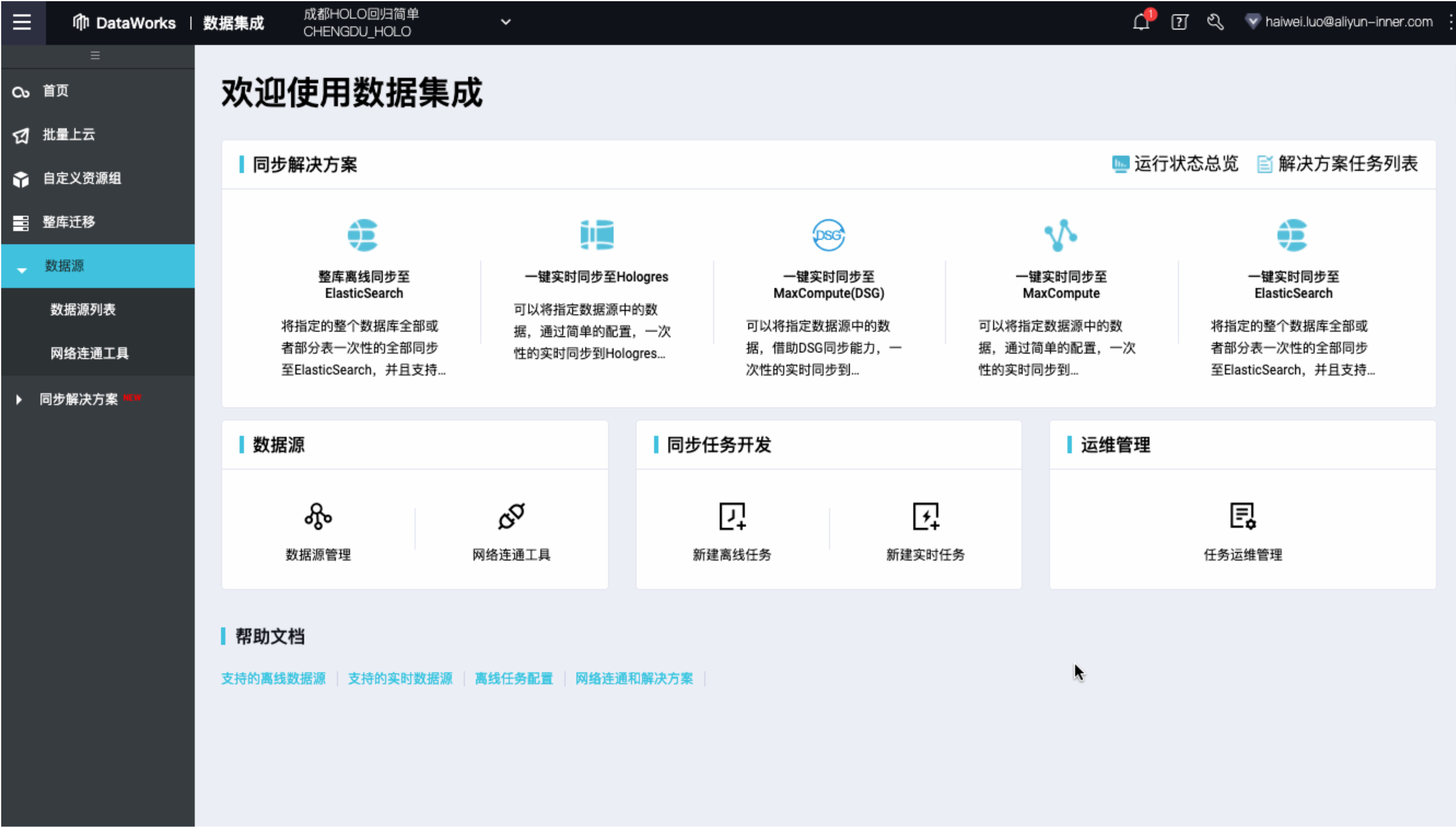


阿里云数据集成解决方案系统

拖拽化开发模式

界面演示

•支持Mysql、Oracle、Datahub、Loghub、Kafka 实时同步到Elasticsearch、Datahub、Kafka等目的端，通过拖拽即可完成开发。



阿里云数据集成解决方案系统

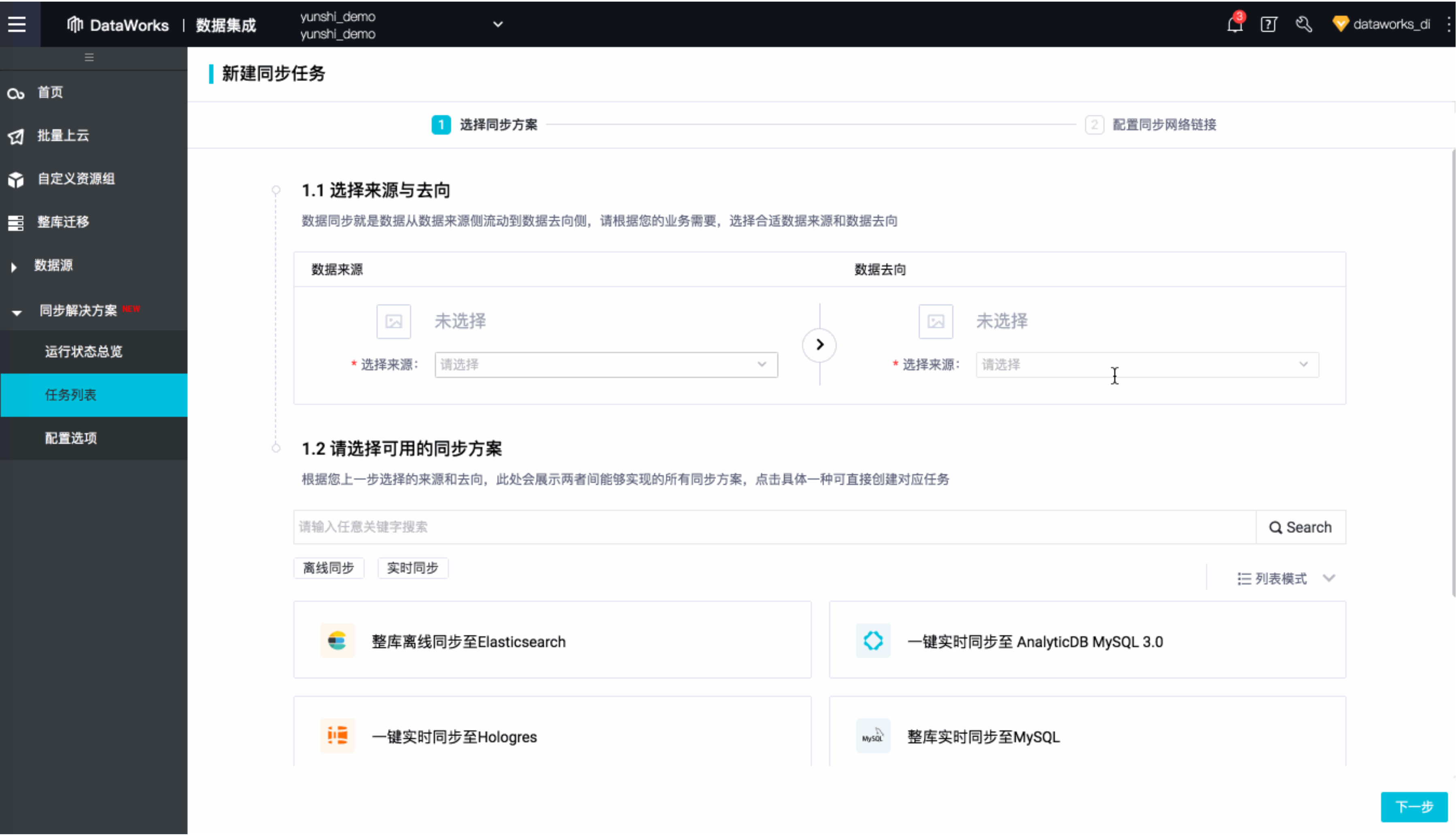
全增量解决方案
以增全量实时到Elasticsearch为例

界面演示

可以非常方便的将现有数据库通过简单的配置后，完成存量的全量迁移，以及后续增量的实时同步。

支持：

- 自动在Elasticsearch建立索引
- 自动建立离线同步任务
- 自动建立实时任务
- 自动启动离线任务
- 自动启动实时任务
- 全流程分布监控和展示，支持分步重试

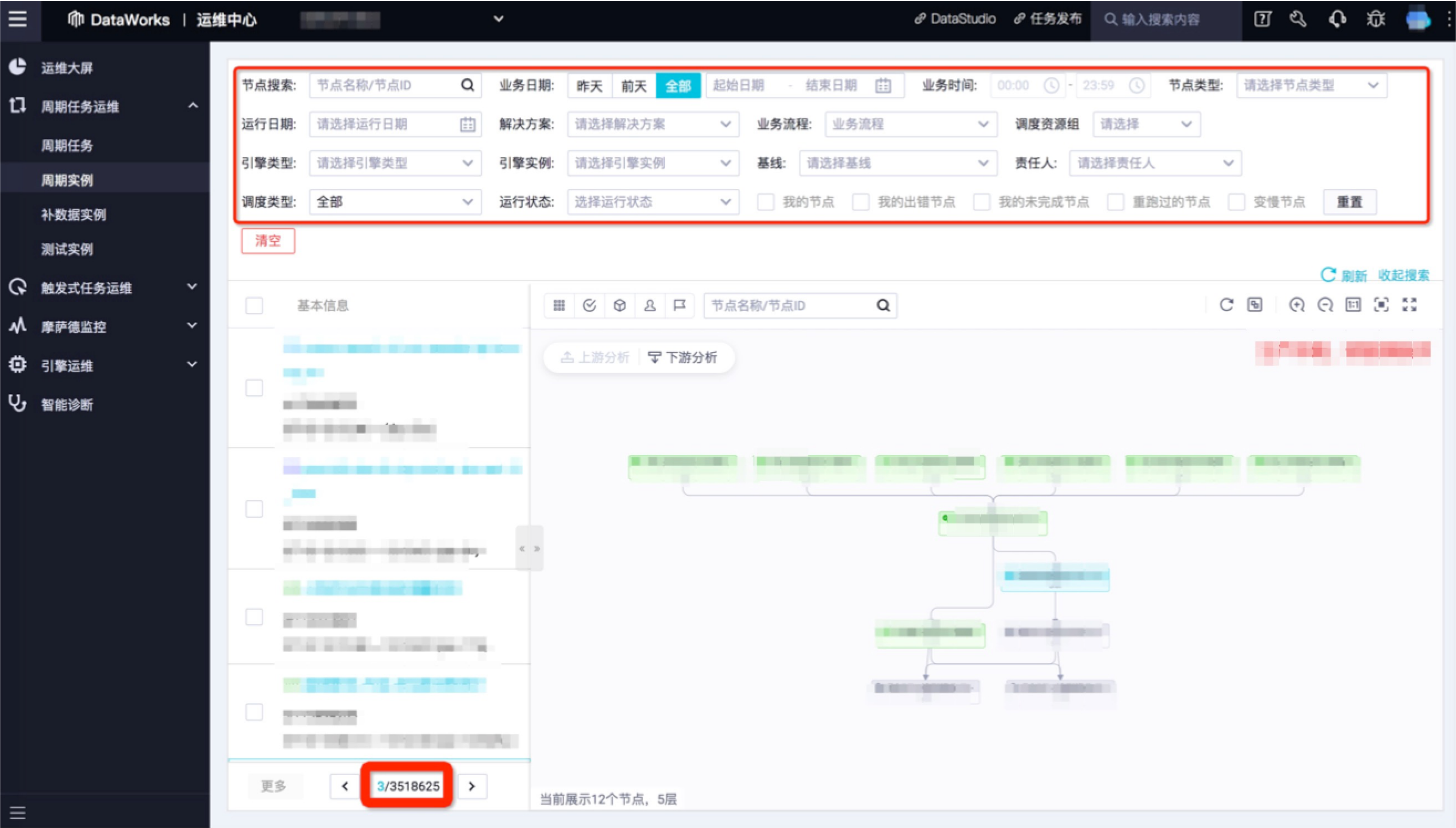


典型案例分享(一)

大规模数据加速模糊查询

当数据库中的数据量级达到百万/千万级别，复杂条件的模糊查询性能会急速下降。

为了解决大规模数据加速模糊查询问题，将数据库数据实时同步到Elasticsearch，提升查询性能和用户体验。



典型案例分享(二)



原始的Metric日志采集信息



某数据实时交换平台大屏实时展示图

DataWorks产品交流群

DataWorks 交流群 | ... 部门



该群属于“大数据+AI”部门群，仅组织内部成员可以加入，如果组织外部人员收到此分享，需要先申请加入该组织。

- 重点技术支持（非工作时间可通过工单）
 - 产品最新活动
 - 产品优先内测通道
 - 产品日常答疑
 - 产品最新直播
 - 好友邀请入群
 - 产品团队支持
 - 大数据专家支持
- DataWorks 招聘
 - 招聘窗口长期开放
 - 涉及技术、产品、测试
 - 逸农 gang.lvg@antgroup.com
 - 枕水 cuiliang.cl@alibaba-inc.com



奥运会全球指定云服务商