GOPS2016
Beijing

2016 全球运维大会

DevOps 2.0：重塑运维价值

北京站

会议时间：12月16日 - 12月17日

会议地点：北京国际会议中心

主办单位：高效运维社区 开放运维联盟

腾讯大规模集群跨城迁移之术

方锦亮 腾讯 高级工程师

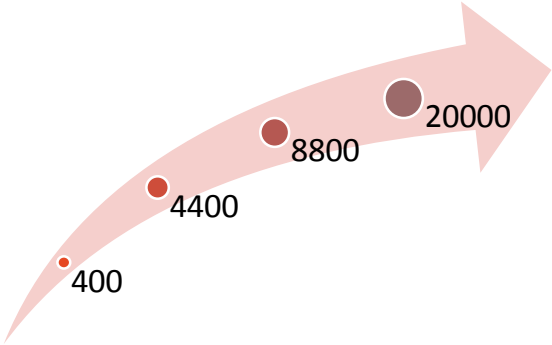
个人介绍

- Joefang(方锦亮)
  - 十年腾讯运营经验，目前负责TDW系统运营
    - 海量设备管理经验
    - 大数据系统运营经验
  - 主导多个运营支撑平台的平台建设
    - TDW运营门户
    - 发布中心
    - 迁移平台
  - 运营理念：建模解决运营难题，平台支撑模型运作

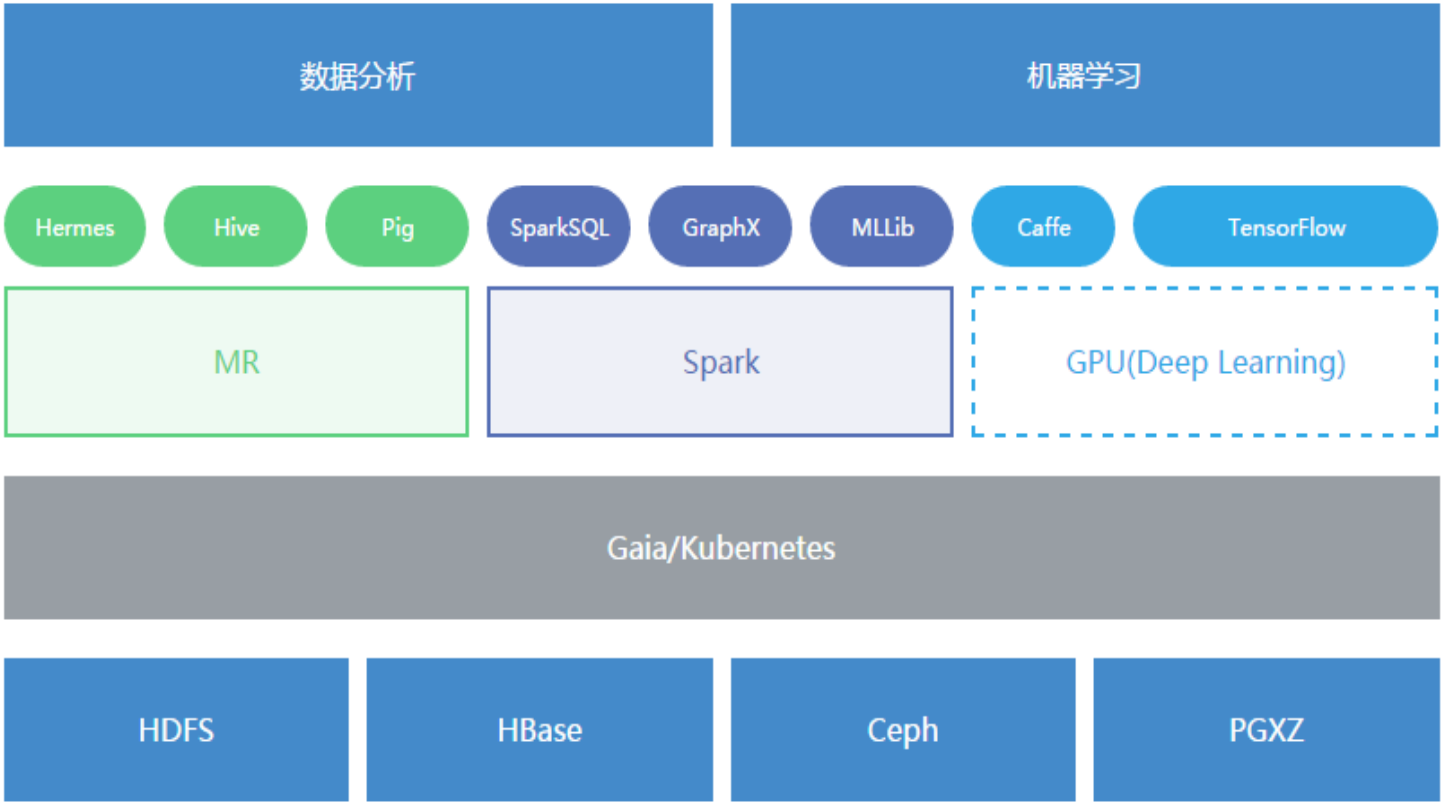
目录

腾讯大规模集群

- TDW 现状
  - 8800
  - 20PB
  - 200PB
- 大规模集群运维的挑战
- 为什么集群要进行迁移
  - 业务快速发展
  - 网络架构限制
  - IDC容量



腾讯大数据平台整体架构



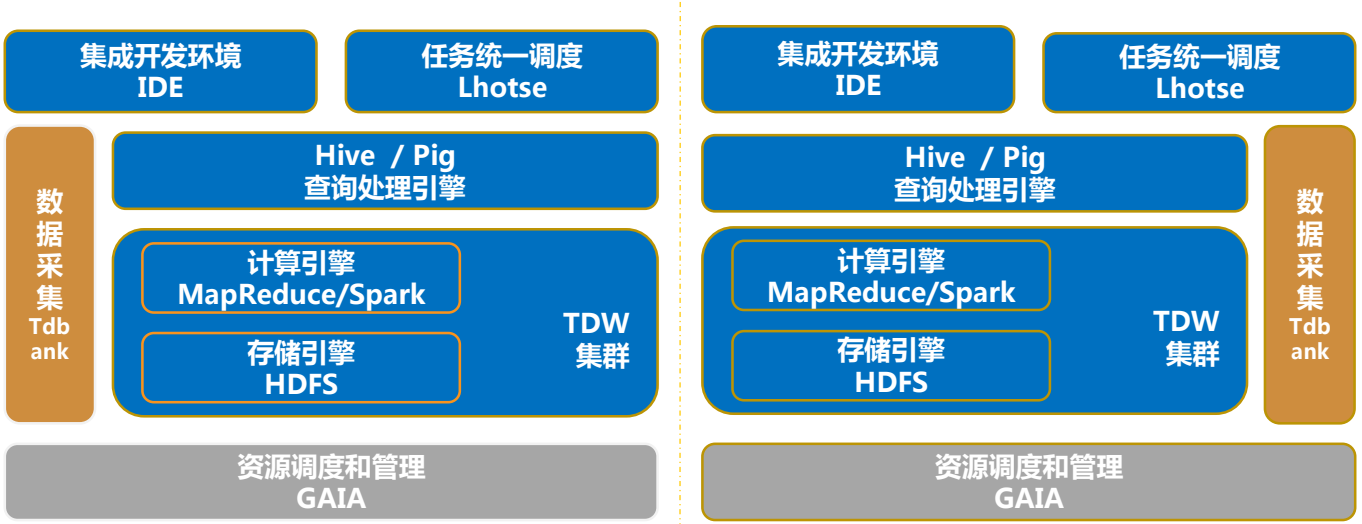
跨城迁移之痛

- 运维工作量大
  - 上百P数据腾挪
  - 几十万任务切换
  - 上万台设备的搬迁
- 业务无感知
  - 系统稳定可用
  - 数据不可丢失
  - 任务保障时效
  - 结果准确无误

- 跨城迁移最大的风险
  - 流量控制不当时影响会扩散到其他业务

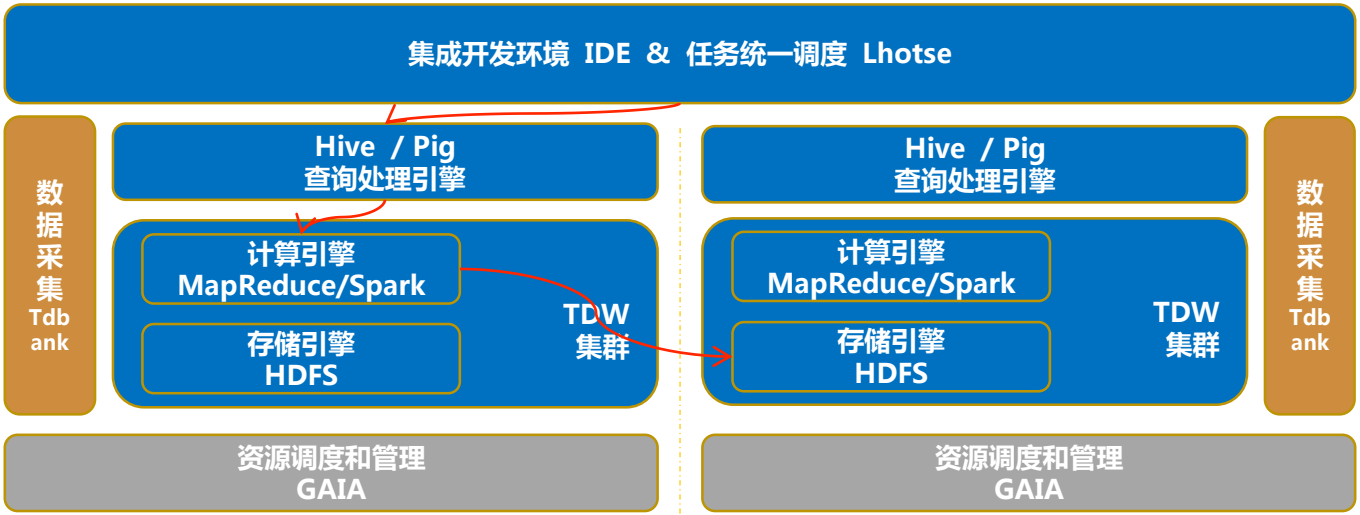
目录

跨城迁移方案——双集群方案



- 特点：系统完全独立；数据双份；任务并行
- 优点：业务完全无影响；除数据迁移外无跨城流量
- 缺点：需要大量冗余设备

跨城迁移方案——单集群方案



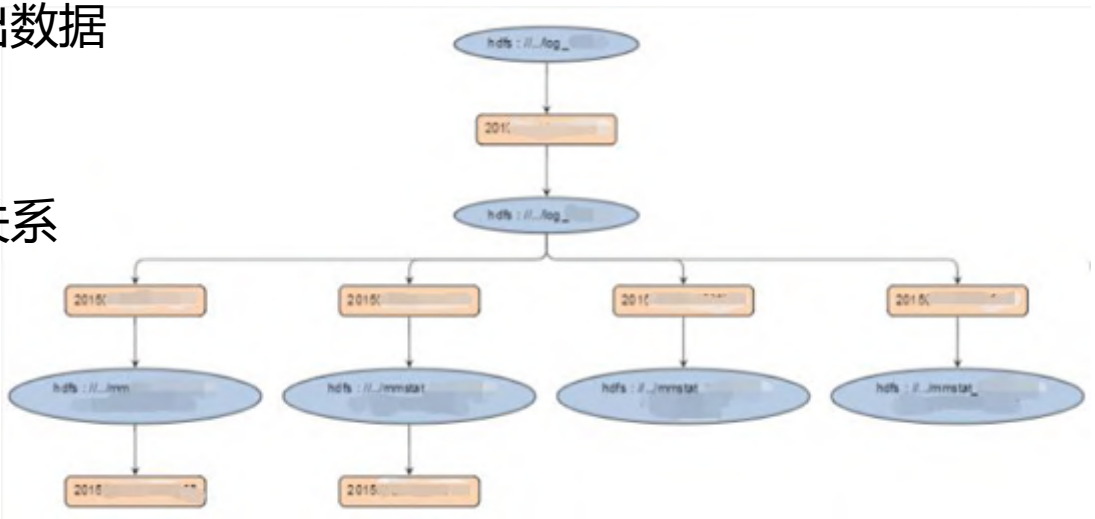
- 特点：系统耦合统一；数据单份；任务唯一
- 优点：需要少量冗余设备；统一用户接口，持续迁移
- 缺点：存在跨城流量的风险

跨城迁移模型

- 基本思路
  - 大系统考虑成本，采用单集群方案
  - 小系统成本忽略，可采用双集群方案
- 跨城流量控制
  - 数据在哪，计算在哪
  - 两边都有数据，哪边数据量大计算在哪
- 跨城迁移模型：基于关系链的迁移模型

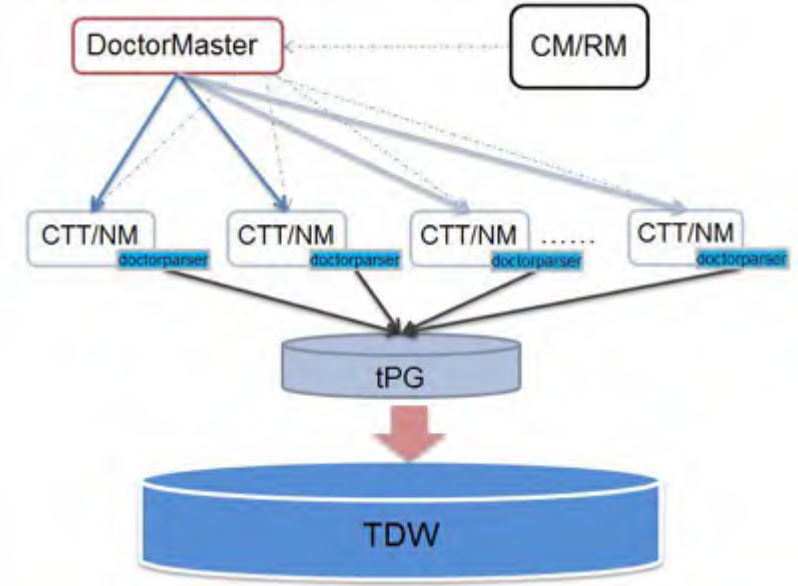
关系链

- 关系链：有关联关系的数据和任务的集合
  - 椭圆描述计算的输入输出数据
  - 矩形描述计算
  - 连线描述数据与计算的关系



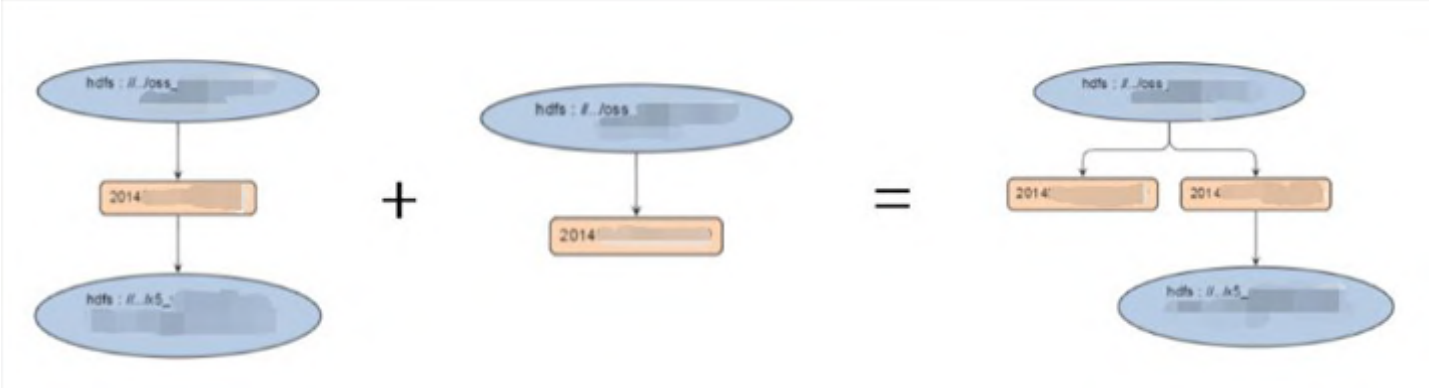
关系的采集

- Hadoopdoctor
  - 任务诊断系统
  - 任务信息采集和存储
- 关系的核心信息
  - 任务ID
  - 数据路径
  - 数据流方向
- 关系信息量大
  - 数据路径约归



关系链的生成

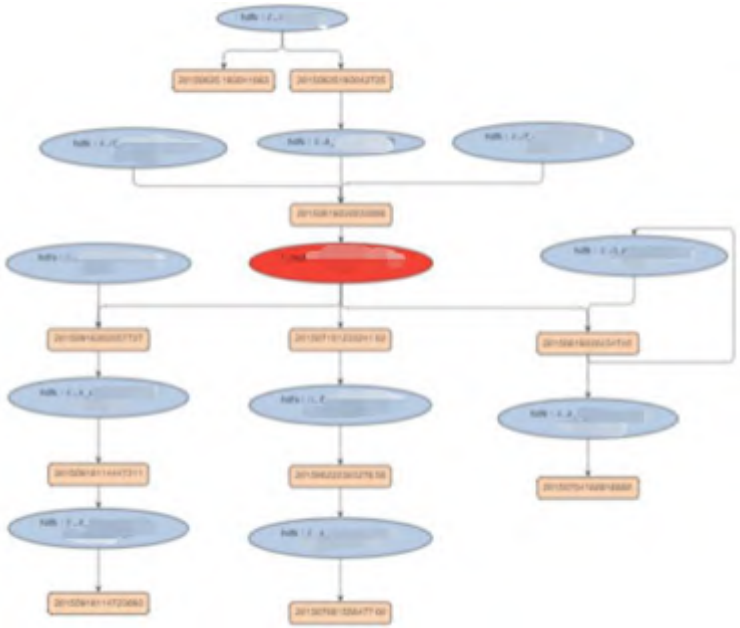
- 关系聚合成关系链



- 覆盖全面的关系链
  - 覆盖周期：日报、周报、月报
  - 任务的新增和删除

关系链的切分

- 切分的思路
  - 关键结点：存储结点
  - 从关键数据结点切开将关系链拆分成若干小关系链
- 面临的挑战：
  - 找到合适的关键结点
  - 如何迁移关系结点

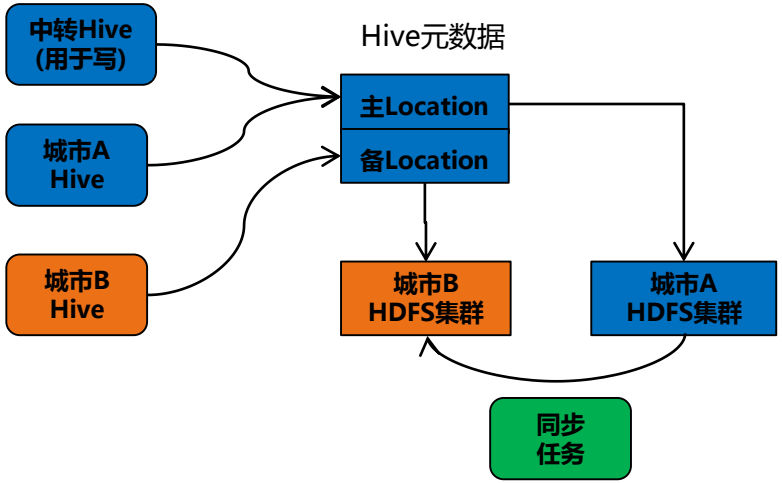


关键结点迁移

- 单份数据方案
  - 优点：容易实现
  - 缺点：可能产生大量跨城流量穿越
- 双份数据方案
  - 优点：计算就近访问数据，仅产生同步数据的流量穿越
  - 问题：
    - 如何保证就近访问
    - 数据一致性如何保证

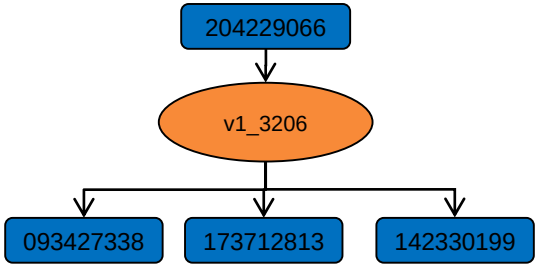
引入HIVE双写表

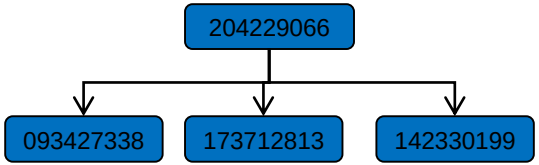
- 双写表的访问逻辑
  - 引入主备Location
  - 写操作
    - 写主Location
    - 同步任务负责数据同步
  - 读操作
    - 就近访问

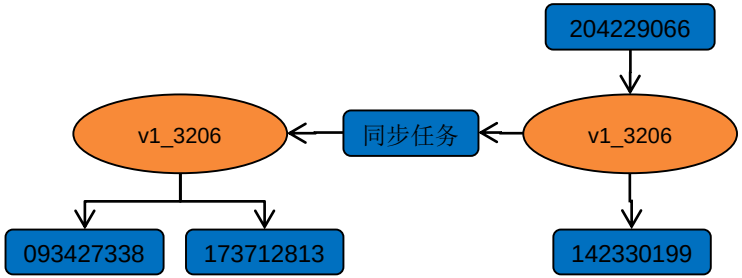


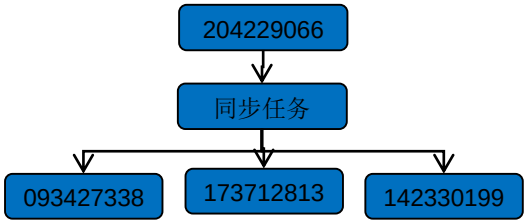
数据一致性保证

- 任务依赖
  - 增加同步任务
  - 同步任务依赖于写数据的任务
  - 读数据的任务依赖于同步任务







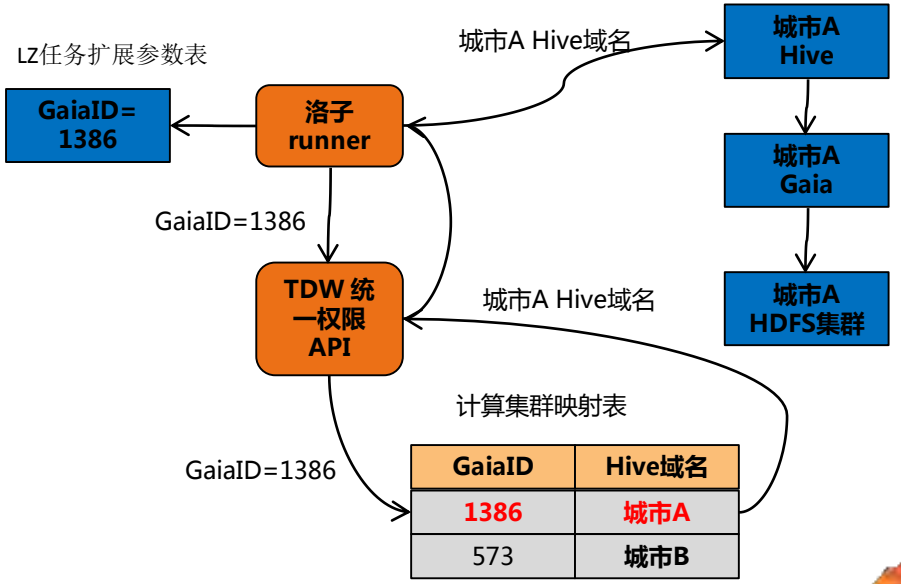


最小化切分和关系链融合

- 如何找到合适的关键结点？
  - 最优化切分：难！难！难！
  - 最小化切分：简单而优雅
- 最小化切分
  - 所有HIVE表都是双写表
  - 关系链“原子”级别切分
- 关系链融合
  - 将若干小关系链合并成大关系链

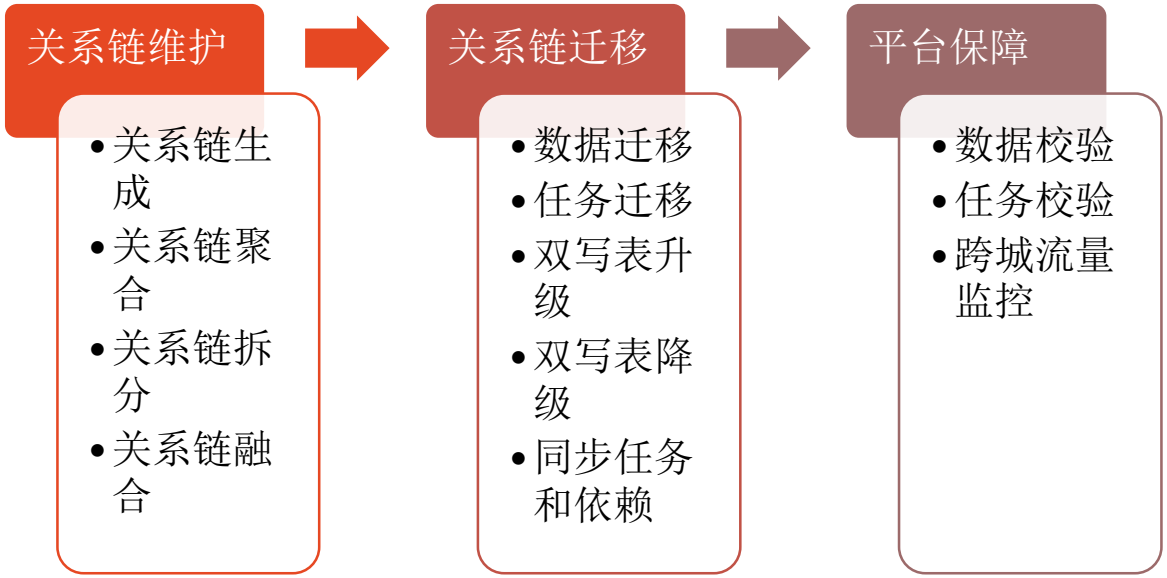
计算的迁移

- Lhotse
  - 一站式任务调度系统
  - 任务和任务参数
- 计算的迁移
  - 扩展参数决定计算集群
  - 计算集群映射表决定提交任务的路由

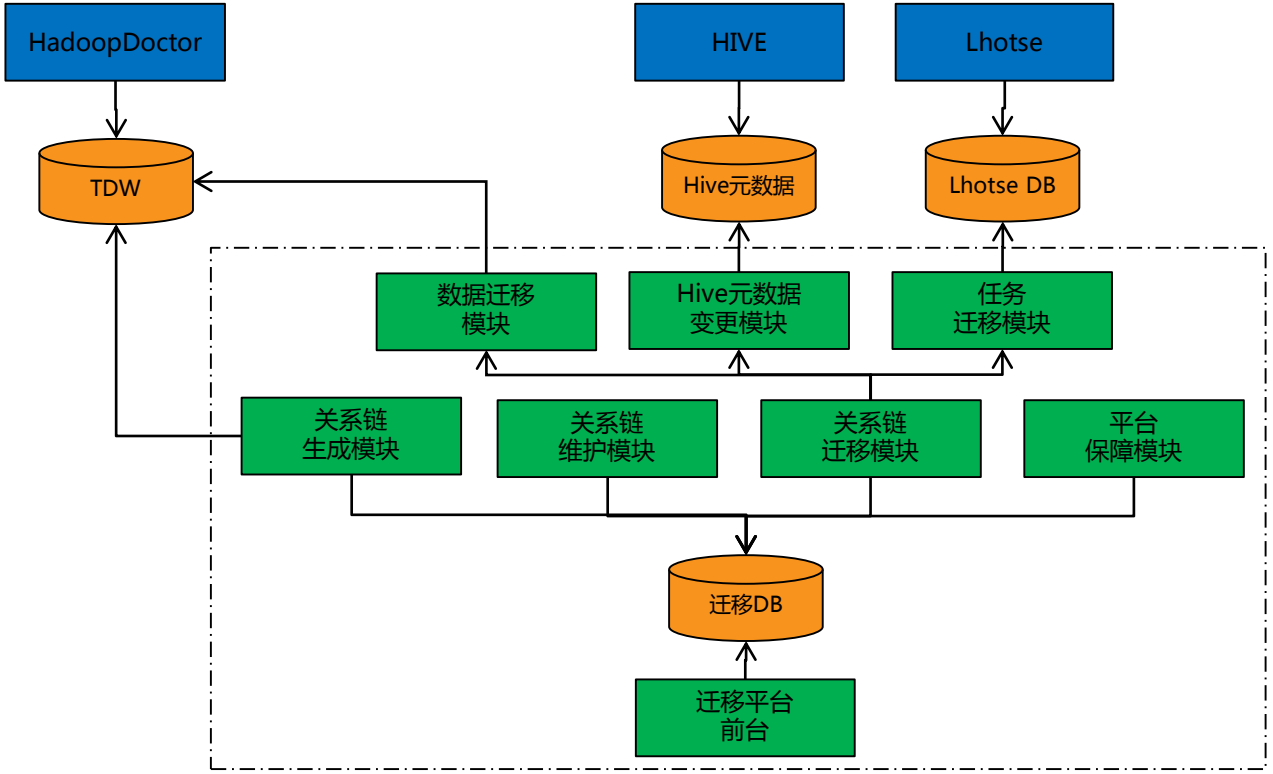


目录

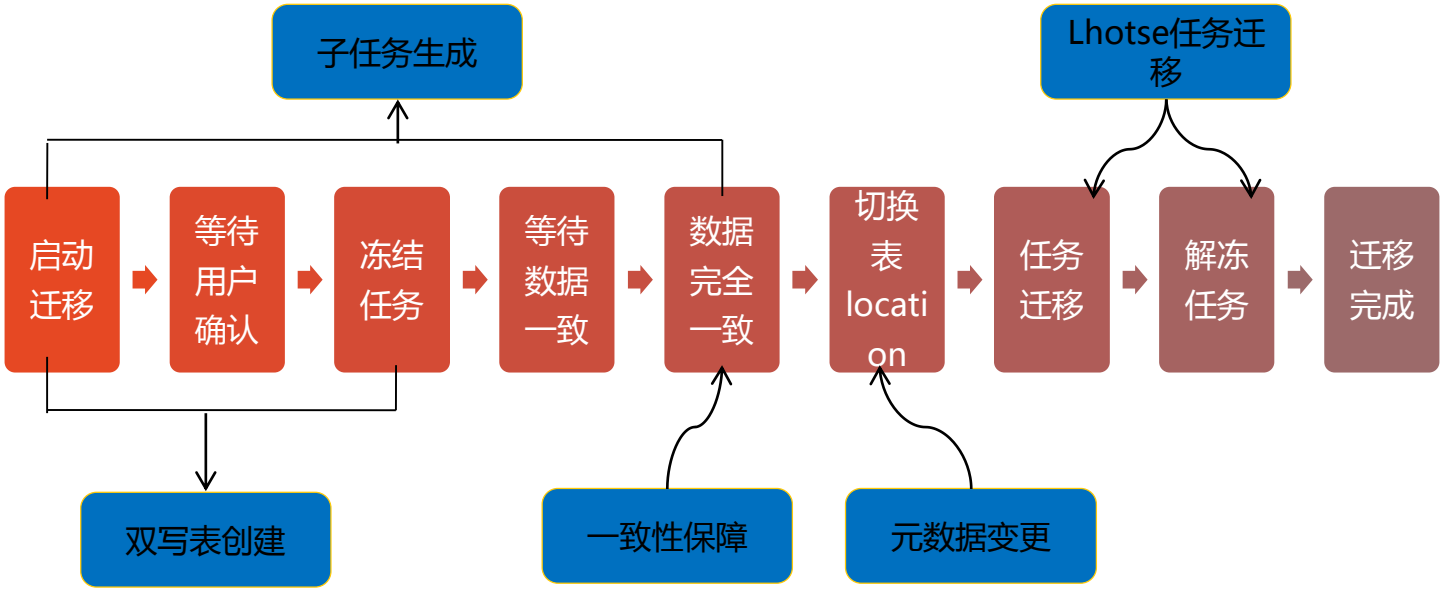
跨城迁移平台功能


关系链维护
• 关系链生成
• 关系链聚合
• 关系链拆分
• 关系链融合
关系链迁移
• 数据迁移
• 任务迁移
• 双写表升级
• 双写表降级
• 同步任务和依赖
平台保障
• 数据校验
• 任务校验
• 跨城流量监控

迁移平台整体架构


HadoopDoctor
HIVE
Lhotse
TDW
Hive元数据
Lhotse DB
数据迁移
模块
Hive元数据
变更模块
任务
迁移模块
关系链
生成模块
关系链
维护模块
关系链
迁移模块
平台
保障模块
迁移DB
迁移平台
前台

关系链迁移模块


子任务生成
Lhotse任务迁移
启动迁移
等待用户确认
冻结任务
等待数据一致
数据完全一致
切换表location
任务迁移
解冻任务
迁移完成
双写表创建
一致性保障
元数据变更

数据迁移模块

- 数据迁移
  - distcp
  - 专用迁移集群
- 多种并发策略
  - 分区级并发
  - 关系链并发
  - 混合优先级runner

平台保障模块

- 基础保障模块
  - 数据校验
  - 任务抽样重跑
- 监控保障模块
  - 数据量波动监控
  - 任务异常波动监控
  - 流量异常监控和自动切换

目录

迁移集群独立部署

- 迁移计算集群特点
  - 网络流量高消耗：万兆网络环境
  - CPU/内存低消耗：NM更高的vcores，任务更小的内存需求
- 独立部署的好处
  - 更少的设备需求
  - 迁移流量可控
  - 与业务集群隔离，不会互相影响

迁移流量控制

- 影响迁移流量的因素
  - 专线带宽
  - 迁移计算集群规模
  - 源HDFS集群规模
  - 目标HDFS集群规模

- 根据源/目标HDFS，测算迁移计算集群规模
- 动态调整迁移计算集群规模

同步任务配置策略

- 同步任务对流量影响小
  - 数据同步和迁移方向相反
- 降低同步任务对业务的影响
  - 使用独立资源池
  - 合理配置distcp map个数
  - 高优先级

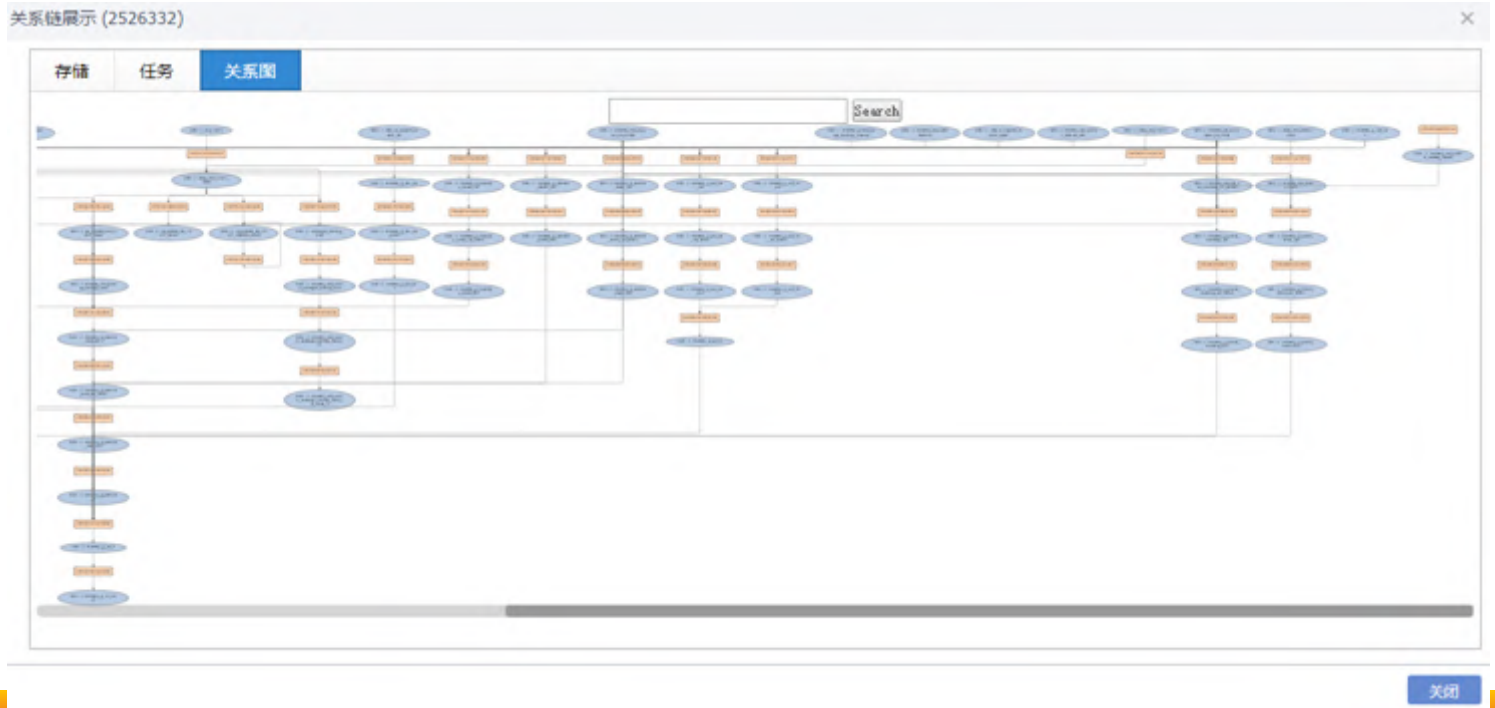
HDFS集群缩容扩容策略

- HDFS集群缩容策略
  - 优先考虑集群整体下线
  - 缩容前准备
    - 数据清理
    - 小文件合并
  - 少量结点多批次缩容
- HDFS集群扩容策略
  - Balance
  - 数据均衡前，新扩容结点不参与计算

目录

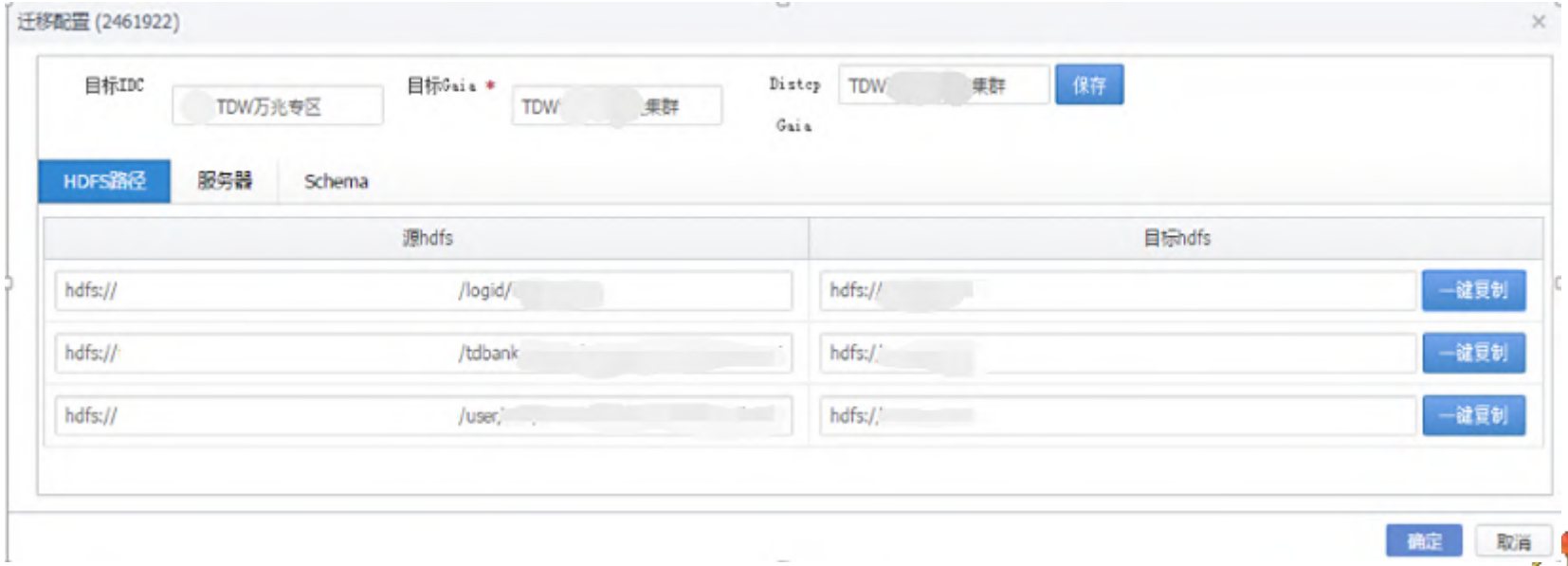
迁移平台前台

- 关系链展示



迁移平台前台

- 关系链迁移提交



迁移平台后台

- 高效
  - 每天迁移量1P
  - 累计迁移100P
  - 近10万计算任务切换
- 稳定
  - 零数据丢失或异常
  - 无运营事故

腾讯大数据
Tencent Big Data

海量数据，无限未来

腾讯大数据官网

DATA.QQ.COM

一个干货满满的网站

欢迎大数据人才加盟

kendyzhao@tencent.com

DevOpsDays 即将首次登陆中国

DevOps 之父 Patrick Debois 与您相约

DevOpsDays 北京站 2017年3月18日

门票早鸟价仅限前100名，请从速哟

http://2017-beijing.devopsdayschina.org/

想第一时间看到
高效运维社区公众号
的好文章吗？

请打开高效运维社区公众号，点击右上角小人，如右侧所示设置就好

Thanks

高效运维社区

开放运维联盟

荣誉出品