金山云

金山云大数据存算分离实践

云存储产品中心
侯雪峰

金山云

目录

CONTENTS

金山云

大数据存算分离介绍

大数据平台面临的挑战

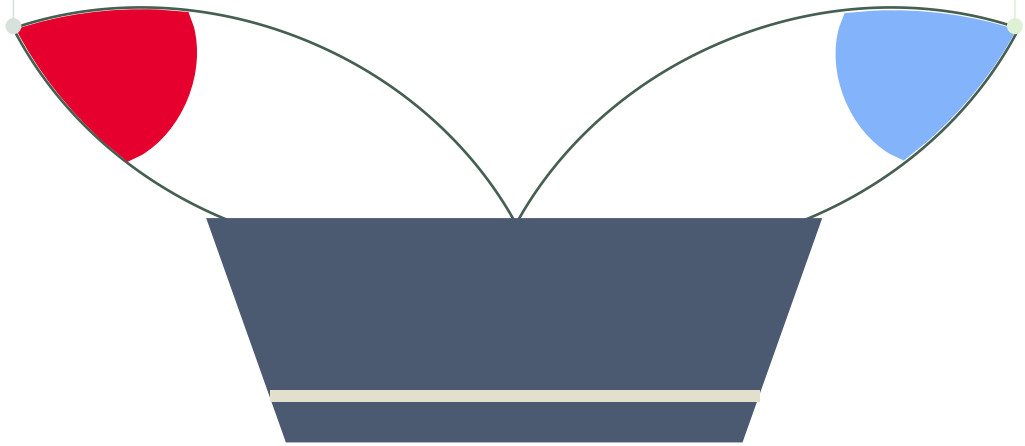
大数据平台现状

1. 计算任务越来越多
2. 计算效率越来越差
3. 存储空间占用越来越大

成本降低

计算资源　存储资源

灵活的弹性伸缩



企业的要求

效率提升

1. 通过数据治理提效
2. 提升计算任务性能

如何降低大数据平台的降本？

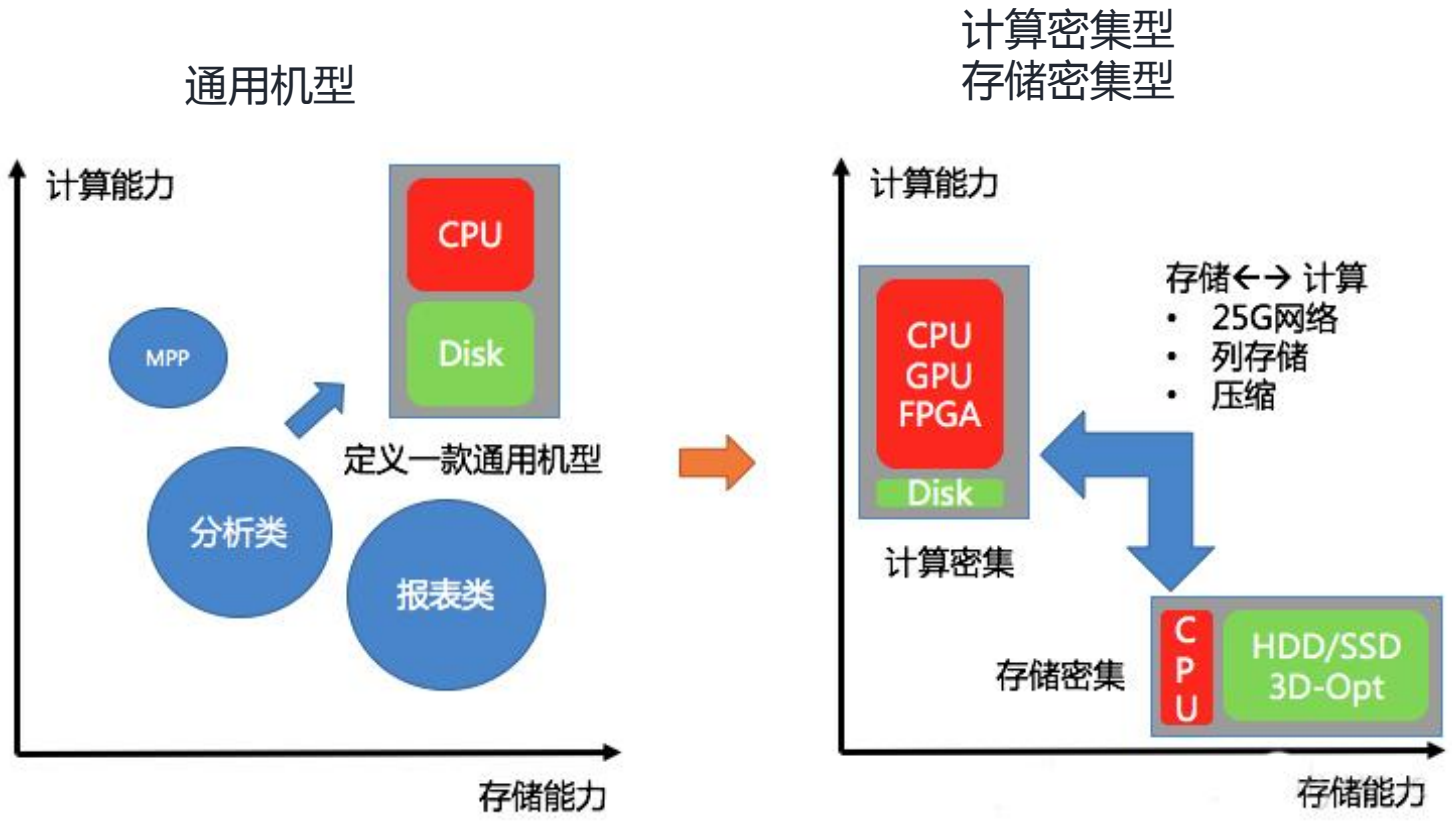
HDFS与对象存储对比

- HDFS
  - 文件存储
  - 单一节点故障
  - 规模受限
  - 所有节点暴露
  - 适合大数据分布式计算
- 对象存储
  - 对象存储，偏平的存储结构
  - Restful api
  - 规模更大
  - 适合存储混杂的数据，构建数据湖

	HDFS	对象存储	结论
价格	三副本存储	EC存储	对象存储价格较HDFS低
弹性	计算资源灵活弹性，容量规划很难做，存储资源缺乏弹性能力，扩容需要需要数据均衡	按需购买，按量付费	对象存储更具备弹性能力
SLA	依赖云存储和ECS以及业务本身，很难量化。 可用性一般在99.9%左右	可靠性：11个9 可用性：标准存储提供99.95%	对象存储的可靠性和可用性更高
性能	提供稳定高效的IO能力 元数据操作性能好	IO不太稳定 不是文件系统 list性能差 rename不是原子操作	大数据场景下对象存储比HDFS在性能和语义上有一定差距

Reference： https://sw.ksyun.com/pro/calc/#/com/37443/doc

存算融合到存算分离



- **网络吞吐大幅提升**
  - 千兆 -> 万兆 -> 25G
  - 带宽充足情况下，本地性优化对计算任务影响只有2%
- **存储介质提升不大**
- **业务混布**
  - cpu密集型、IO密集型等多种业务对资源需求不同 ……
  - 业务时效性不同
- **云时代的选择，必然趋势**

通过一款机型通吃存储+计算方案，已经演变成存储+计算各自服务化，通过高速网络进行连接的趋势

大数据架构演进

大数据架构变革

Hadoop 1.x 计算存储一体

- 大数据基本能力
  - MapReduce
  - 资源、数据管理合一
  - 三副本

面向单一MR分析业务

第一代：融合型

Hadoop 2.x 计算解耦

- 解耦的大数据平台
  - 资源管理独立
  - 多计算引擎
  - Namenode 联邦

面向复合型分析业务

第二代：扩展型

Hadoop 3.x 多样化存储

- 开放的生态
  - 服务容器化
  - 接入数据湖
  - 支持EC存储

面向数据湖，企业级大数据业务

第三代：开放型

1 计算层逐渐轻量化，逐步**与数据解耦**

2 HDFS存储层逐渐支持多种存储，逐步走向**存算分离**

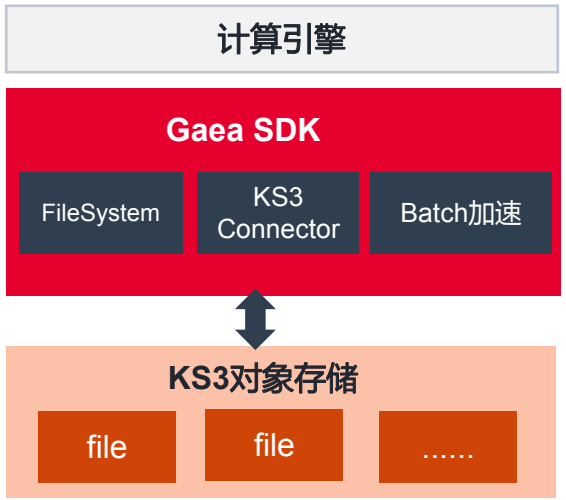
3 基于计算存储分离，逐步向**数据湖**架构演进

金山云存算分离方案

存算分离架构的三种模式

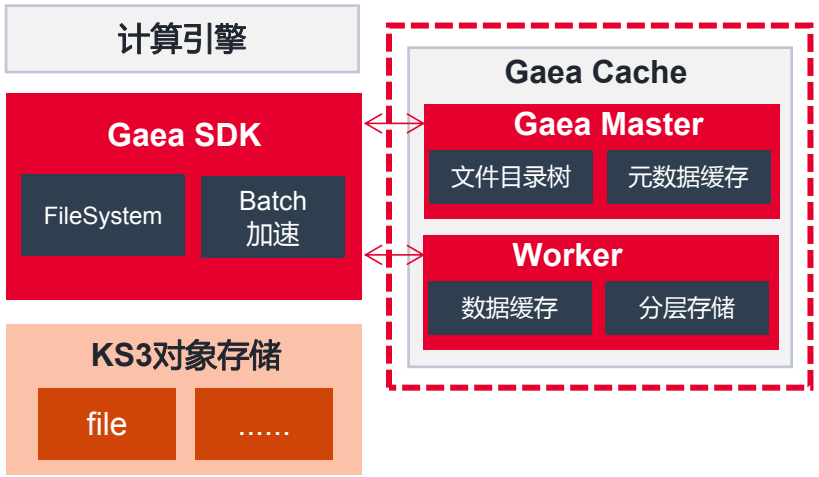
金山云

直连模式



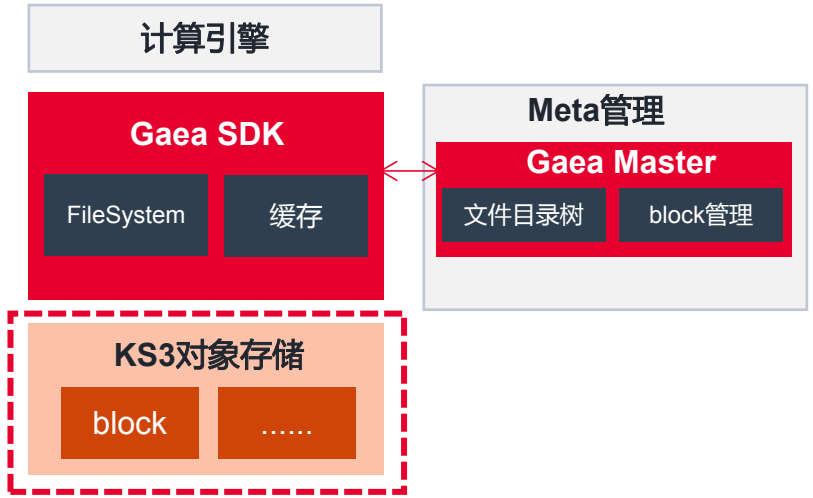
- KS3 connector
- 类HDFS文件系统接口
- 提供针对计算引擎的性能优化
- 是一个SDK，无需额外部署
- 直观、易用
- 性能较差，使用场景有限

对象模式



- 提供元数据缓存能力
- 可以缓存热数据，减少KS3访问
- 是一个可靠集群，需要额外资源部署
- 数据生命周期管理
- 提升性能明显
- 使用相对复杂一点

块模式



- 将文件分块存储在ks3
- 文件元数据完全由自身管理，带外不可见
- 细粒度的IO优化
- 完整的posix语义
- 大数据场景能力有限，有丢失元数据风险

存算分离：解耦大数据存储和计算

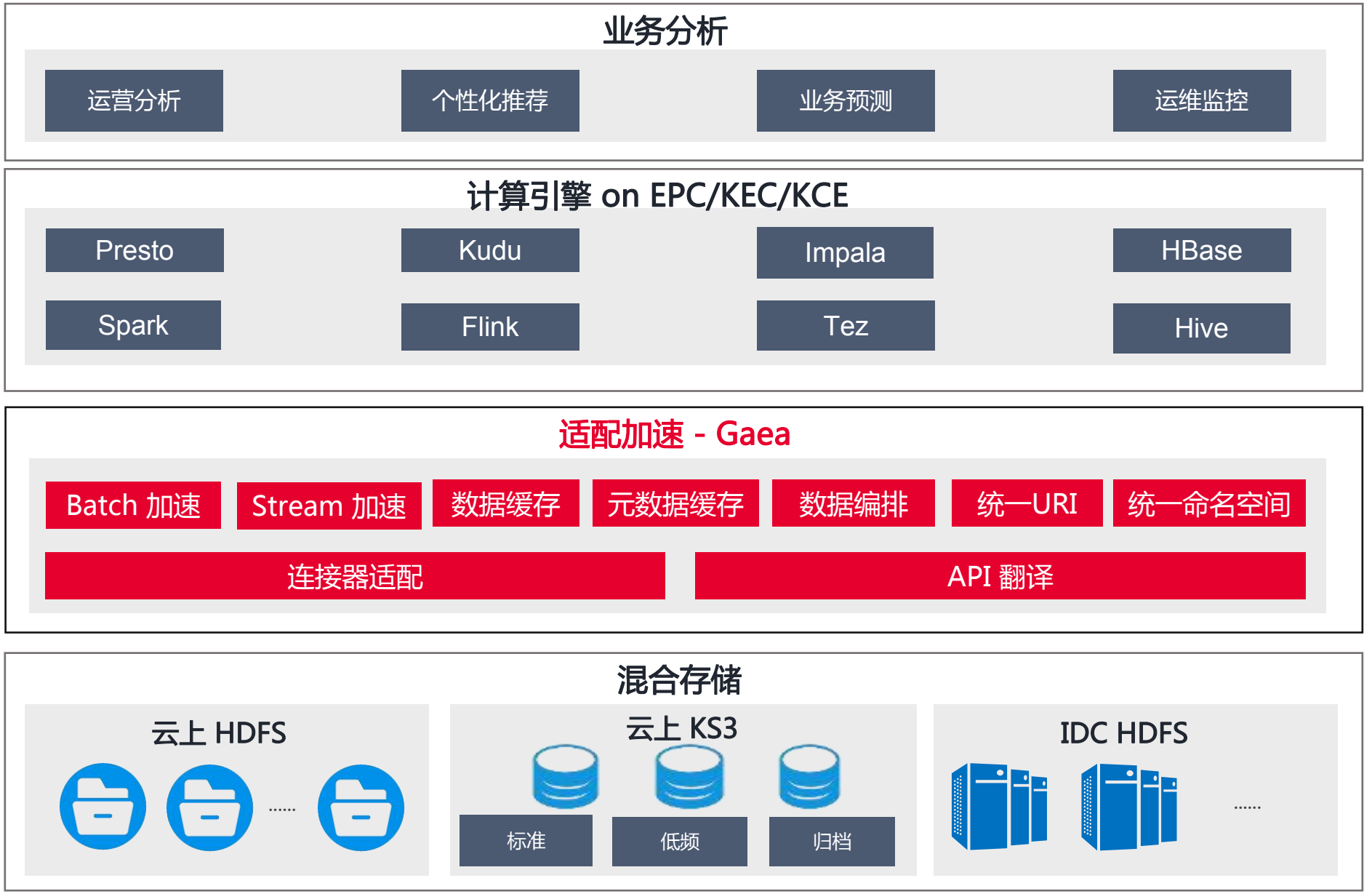
金山云

方案优势

- 计算弹性伸缩
- 存储计算分离
- 数据冷热分离
- 混合、多云模式下统一存储

收益

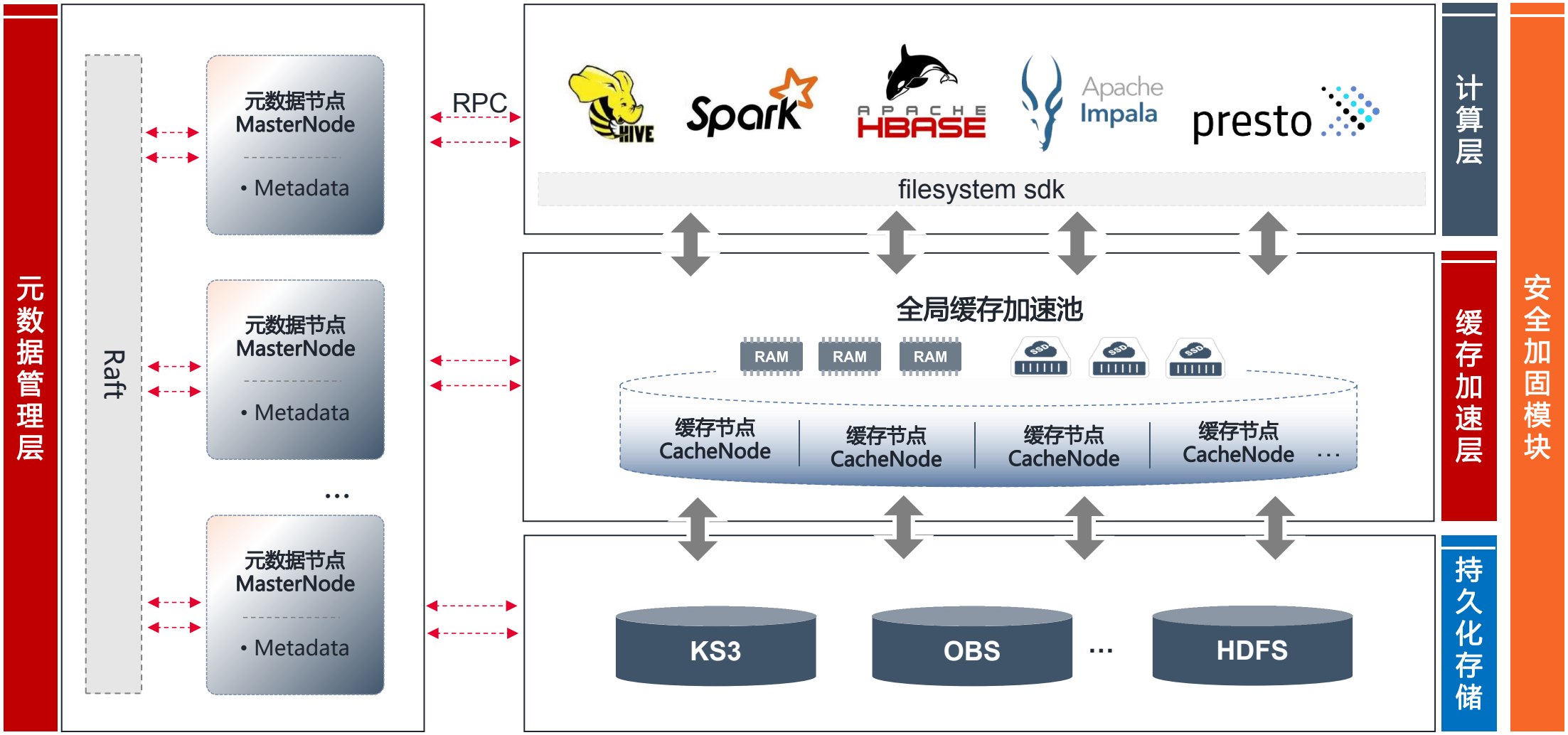
- 成本较低
- 数据高可靠
- 提升性能



金山云

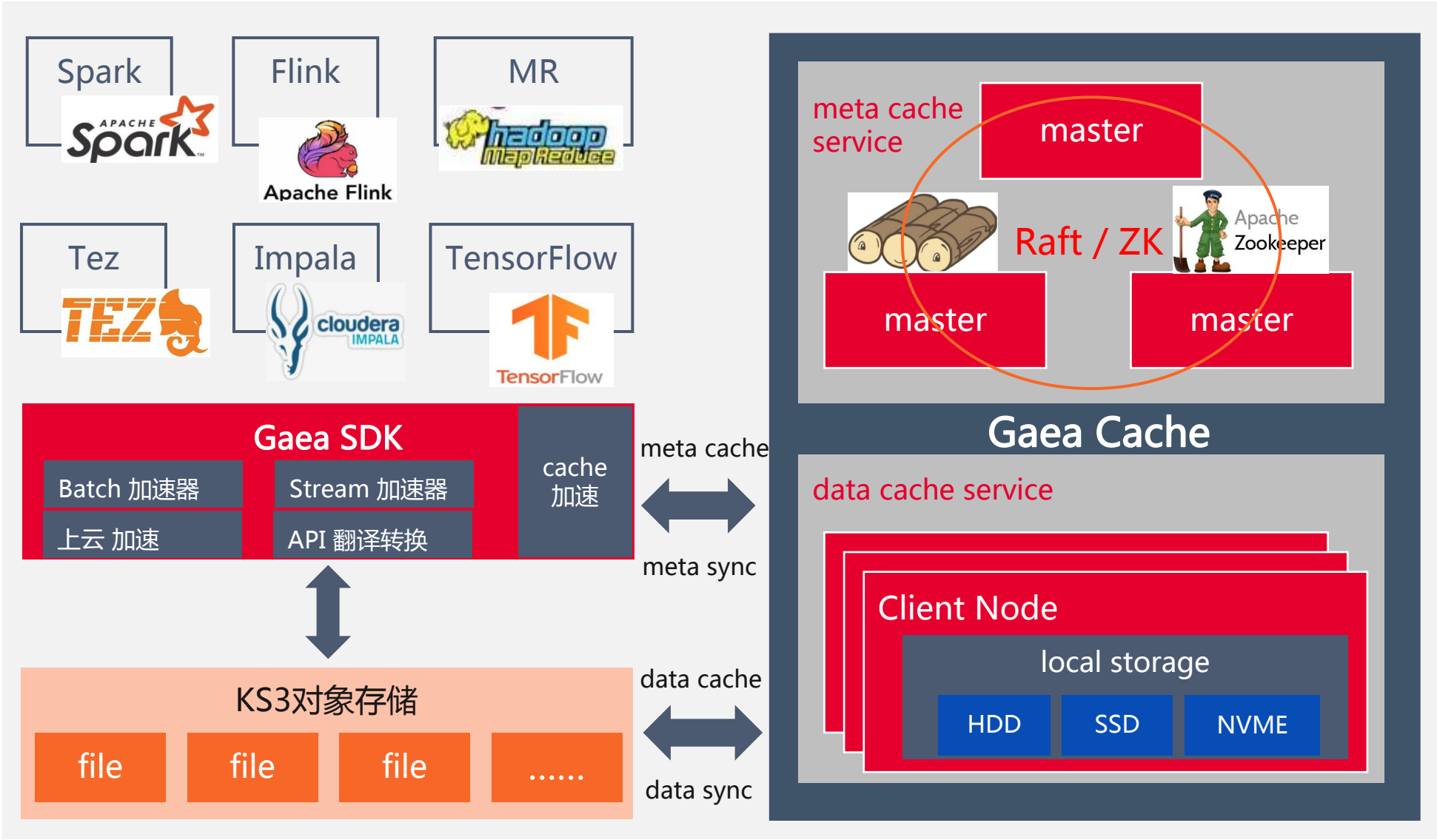
存算分离方案KS3-HDFS

存算分离逻辑架构


元数据管理层
Raft
元数据节点
MasterNode
• Metadata
元数据节点
MasterNode
• Metadata
...
元数据节点
MasterNode
• Metadata
RPC
HIVE
Spark
APACHE HBASE
Apache Impala
presto
filesystem sdk
计算层
全局缓存加速池
RAM
RAM
RAM
SSD
SSD
SSD
缓存节点 CacheNode
缓存节点 CacheNode
缓存节点 CacheNode
缓存节点 CacheNode
...
缓存加速层
KS3
OBS
...
HDFS
持久化存储
安全加固模块

金山云存算分离核心模块

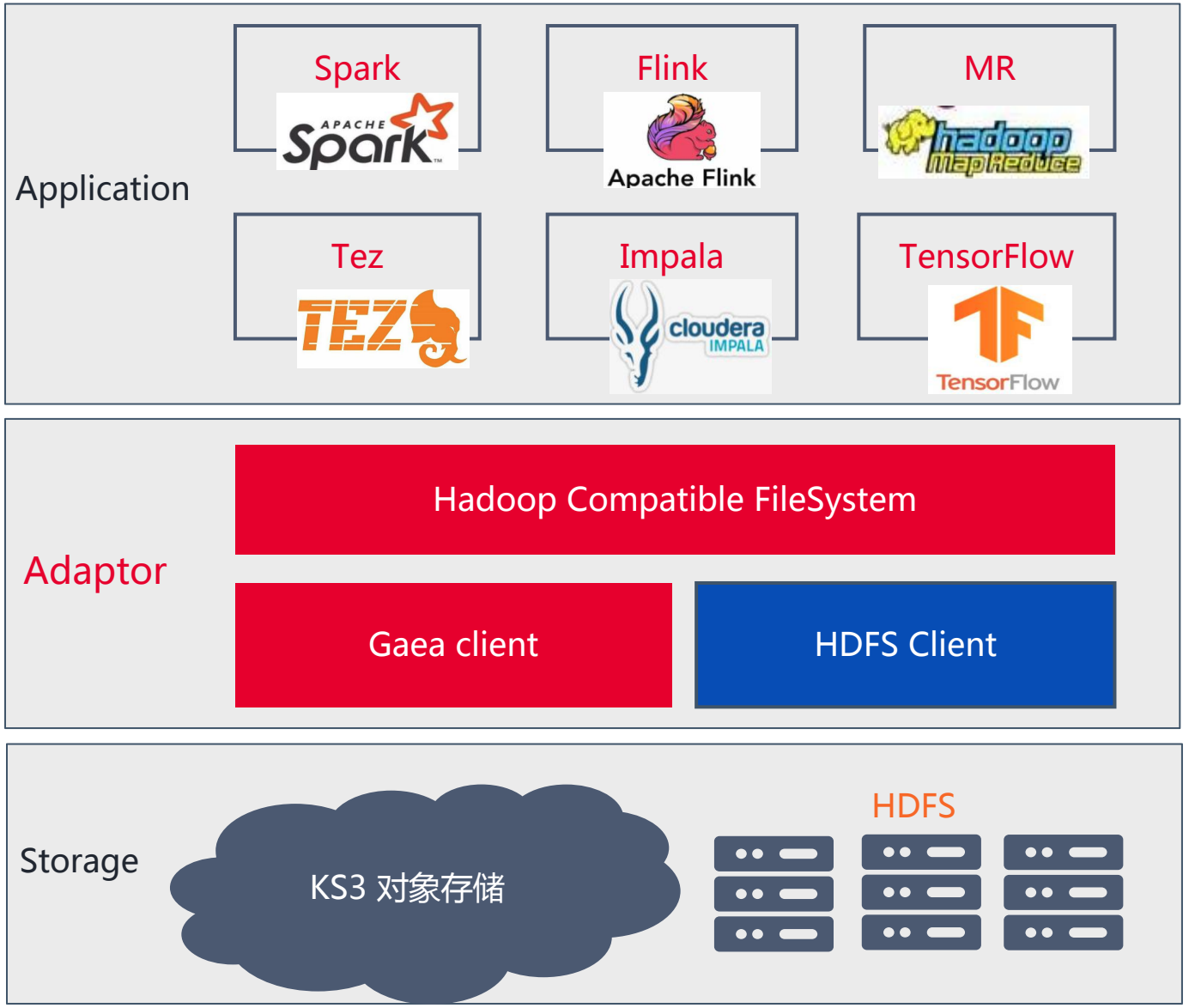
- Gaea SDK
  - ks3 Connector
  - Batch 加速
  - 完整的Streamming 语义
  - DistCp加速
- Gaea Master
  - 元数据缓存和管理
  - 高可用
  - 高性能
- Gaea Cache
  - 客户端数据缓存
  - 高性能



Gaea SDK关键特性：Hadoop生态与ks3之间的桥梁

让用户像使用HDFS一样使用ks3

- HDFS 兼容的 FileSystem
  - Ks3 connector
  - 无缝对接，业务无侵入
  - 支持客户端加密
- 优化list操作
  - 目录填充
  - 减少scan
- IO 加速



Gaea SDK关键特性：加速大数据场景批量计算

Batch on ks3 挑战

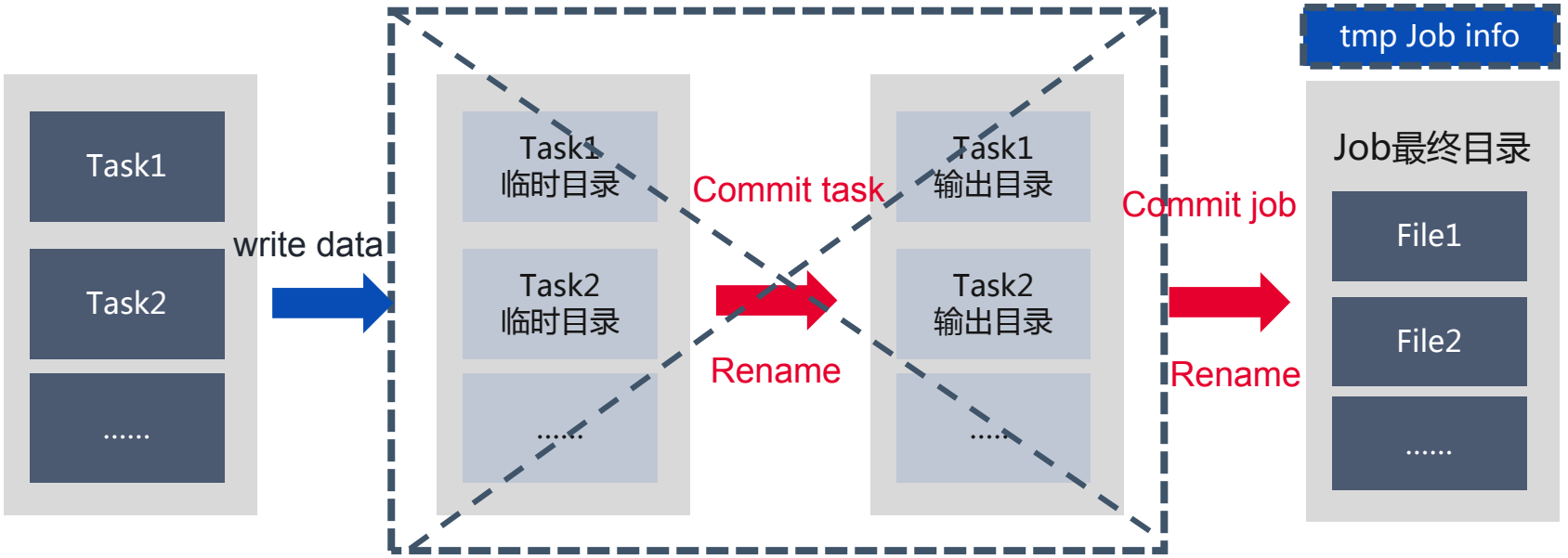
- 性能差
- Job不稳定

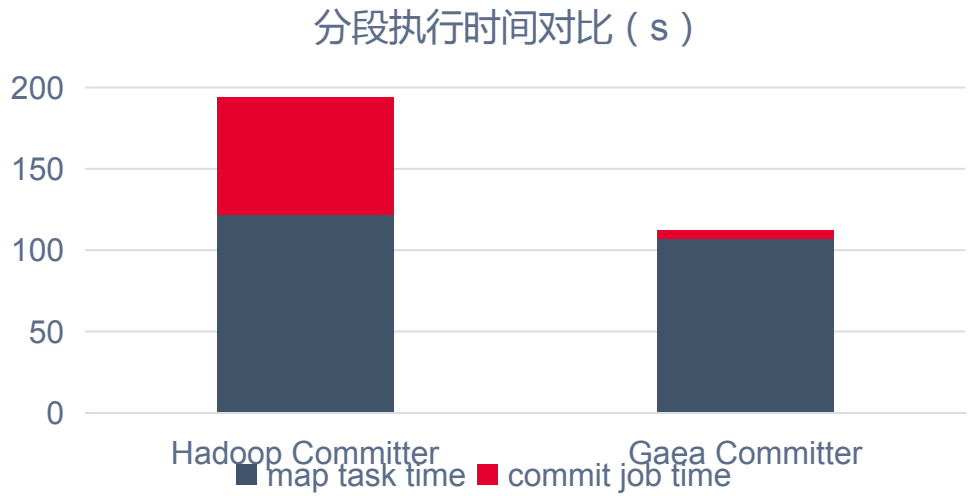
问题原因

- 计算框架的commiter机制
  - rename操作
  - list操作

解决方案

- 优化list
  - 目录进行物理填充，优化list
  - 流式处理
- 去rename
  - task rename、job rename
  - 支持推测执行
  - MPU保证数据一致性





TeraGen 测试

数据量： 50GB

集群规模：2master 3core

节点规格：8C32G 100G

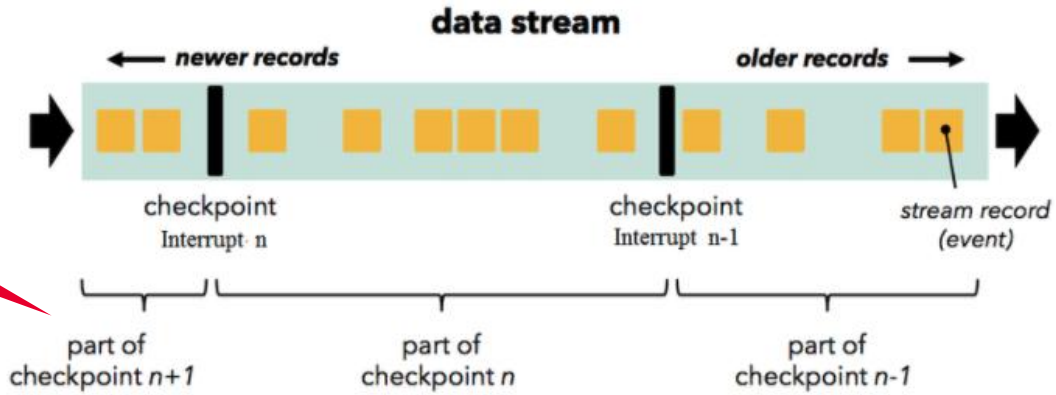
版本：kmr hadoop 2.7.3

Gaea SDK关键特性：完整支持实时计算Streamming语义

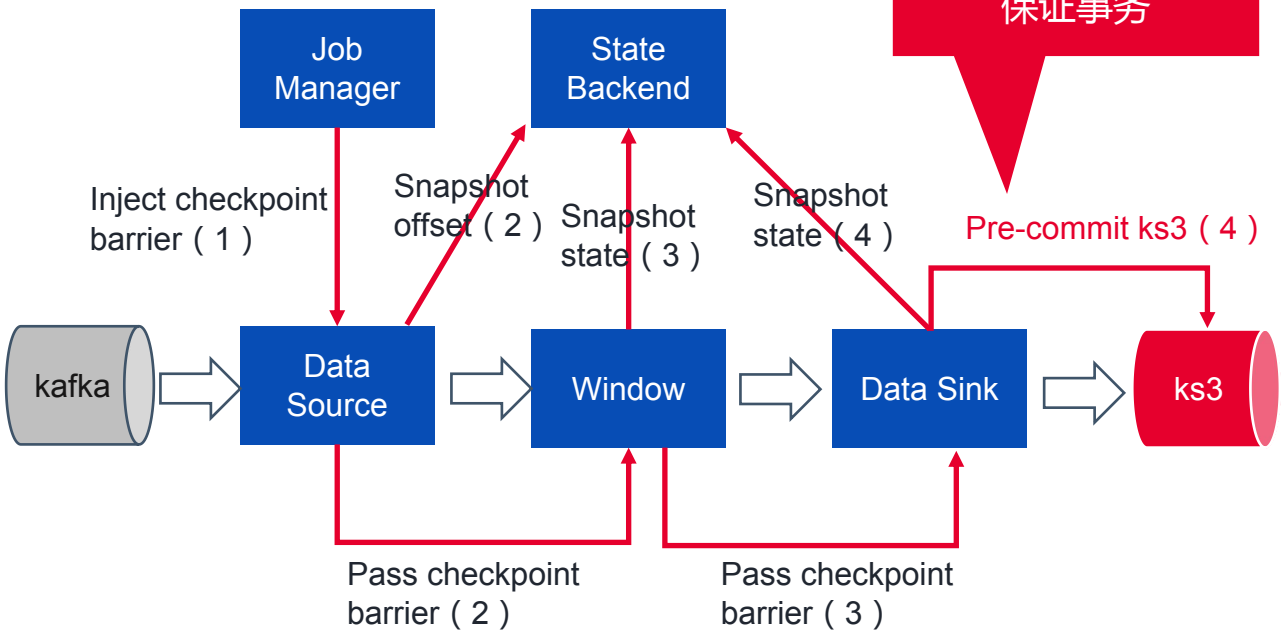
Flink Streaming on ks3挑战

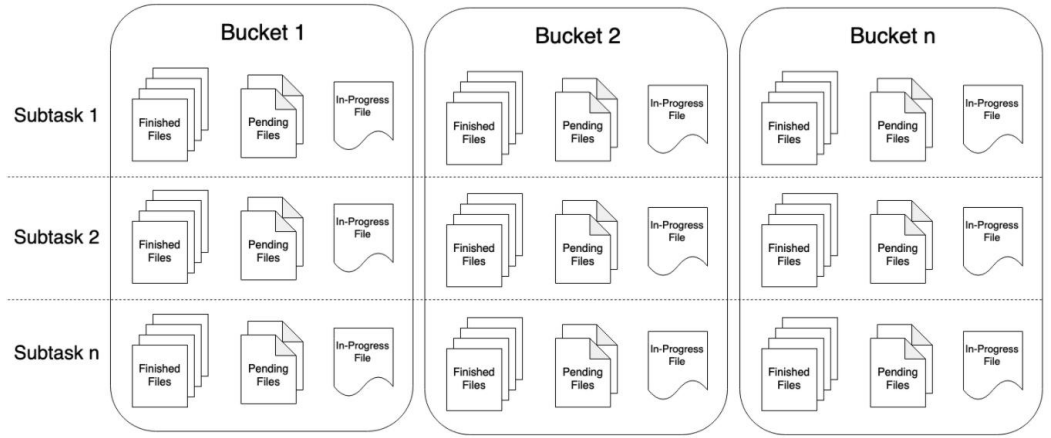
- ks3 as source
- ks3 as sink
  - ks3不是幂等的
  - 不支持 Exactly Once语义

使用checkpoint
状态恢复
分段容错



2PC + MPU
保证事务





In-progress ：当前文件正在写入中
Pending ：文件写入完成，但未提交当处于 In-progress 状态的文件关闭，成为 Pending 状态
Finished ：在成功的 Checkpoint 后，Pending 状态将变为 Finished 状态

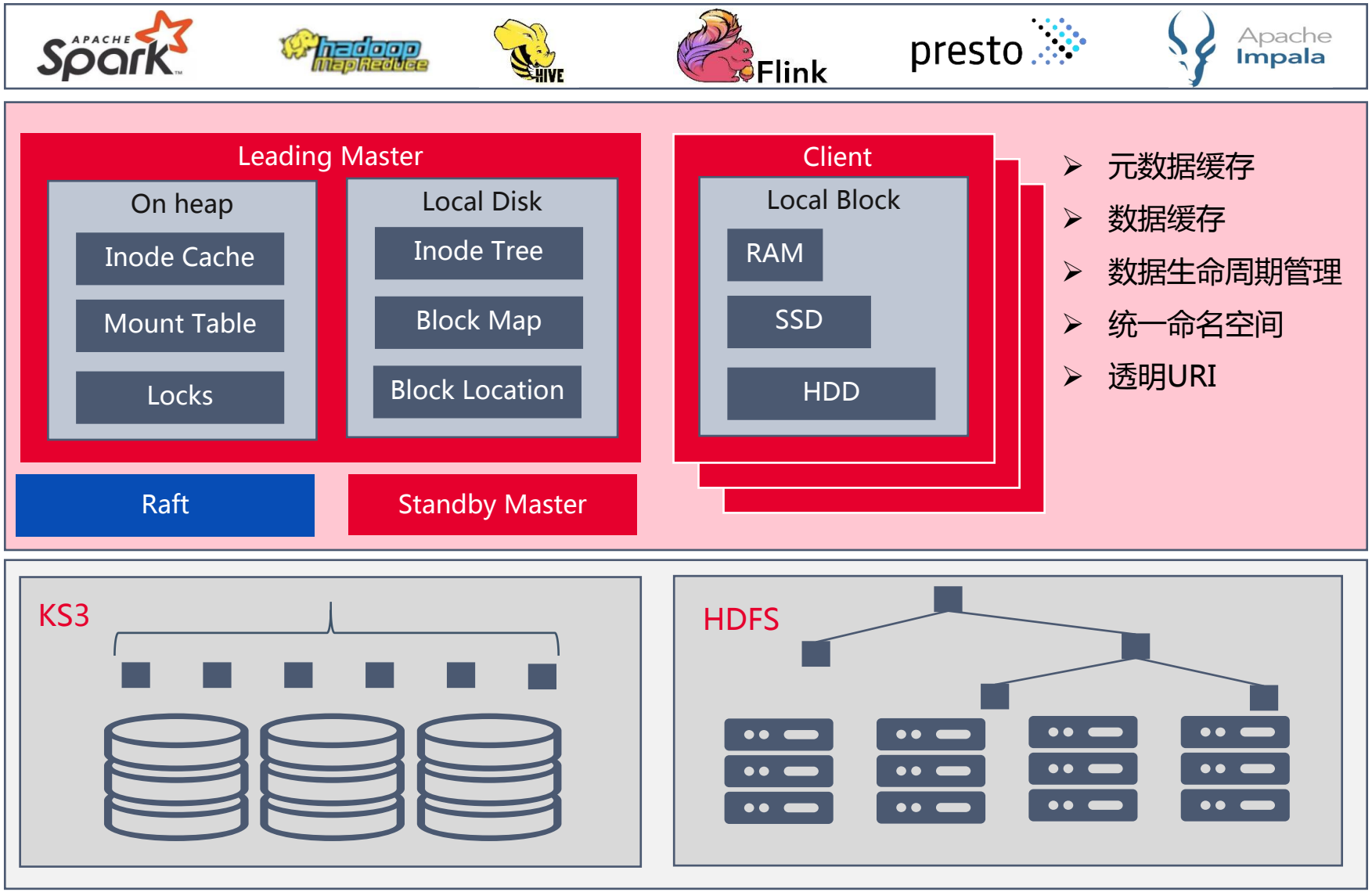
Gaea Cache：存算分离的缓存架构

挑战

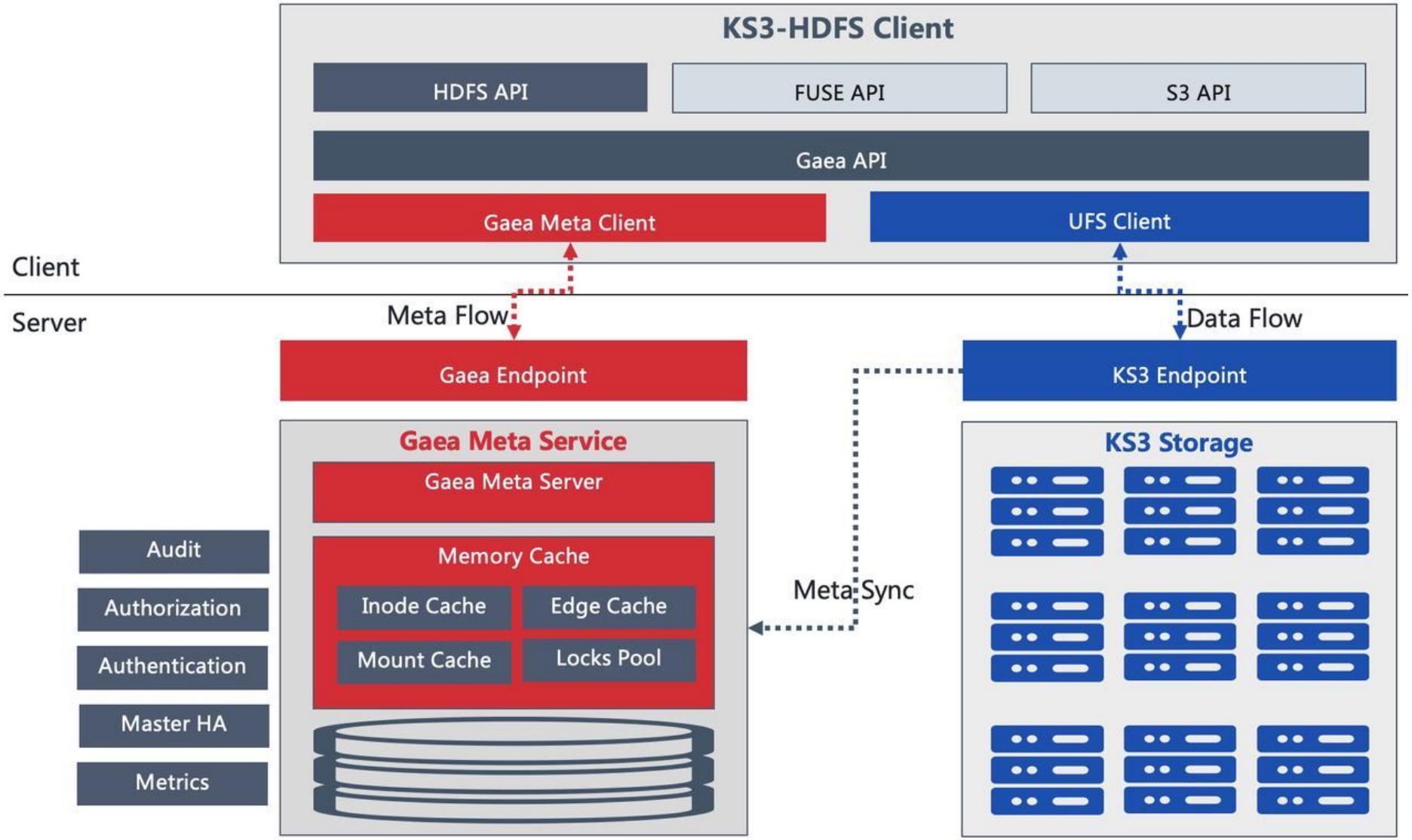
- Gaea SDK不通用
- 远程调用ks3性能不高
  - List、rename
  - 热数据反复调用
  - 远程带宽有限
- 异构复杂性
  - 多种schema并存
  - 多版本的Hadoop集群
  - 客户端配置混乱

解决方案：Gaea Cache

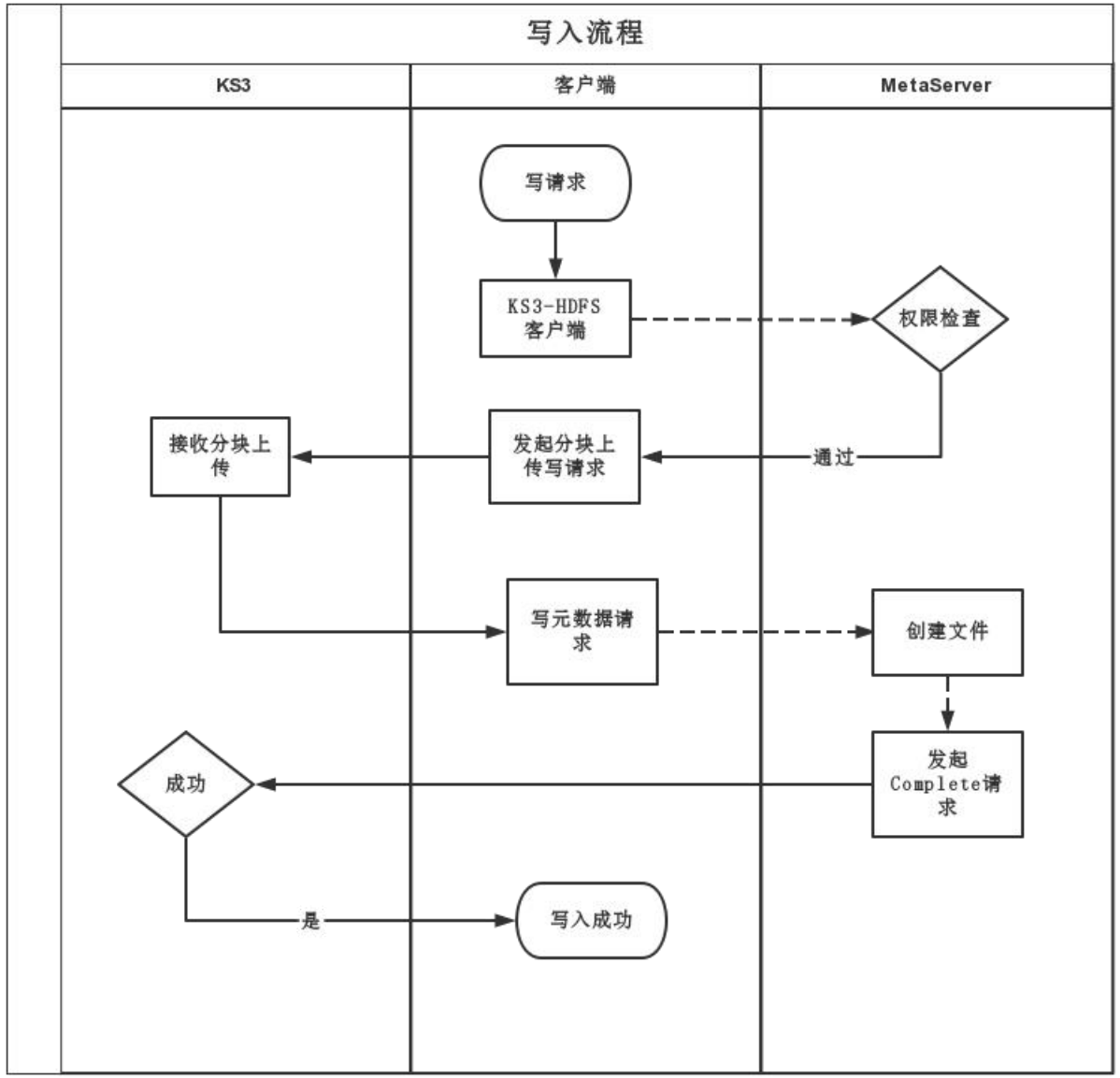
- 元数据缓存
- 数据缓存
- 数据生命周期管理
- 统一命名空间
- 透明URI



KS3-HDFS数据访问流程



如何保证数据一致性



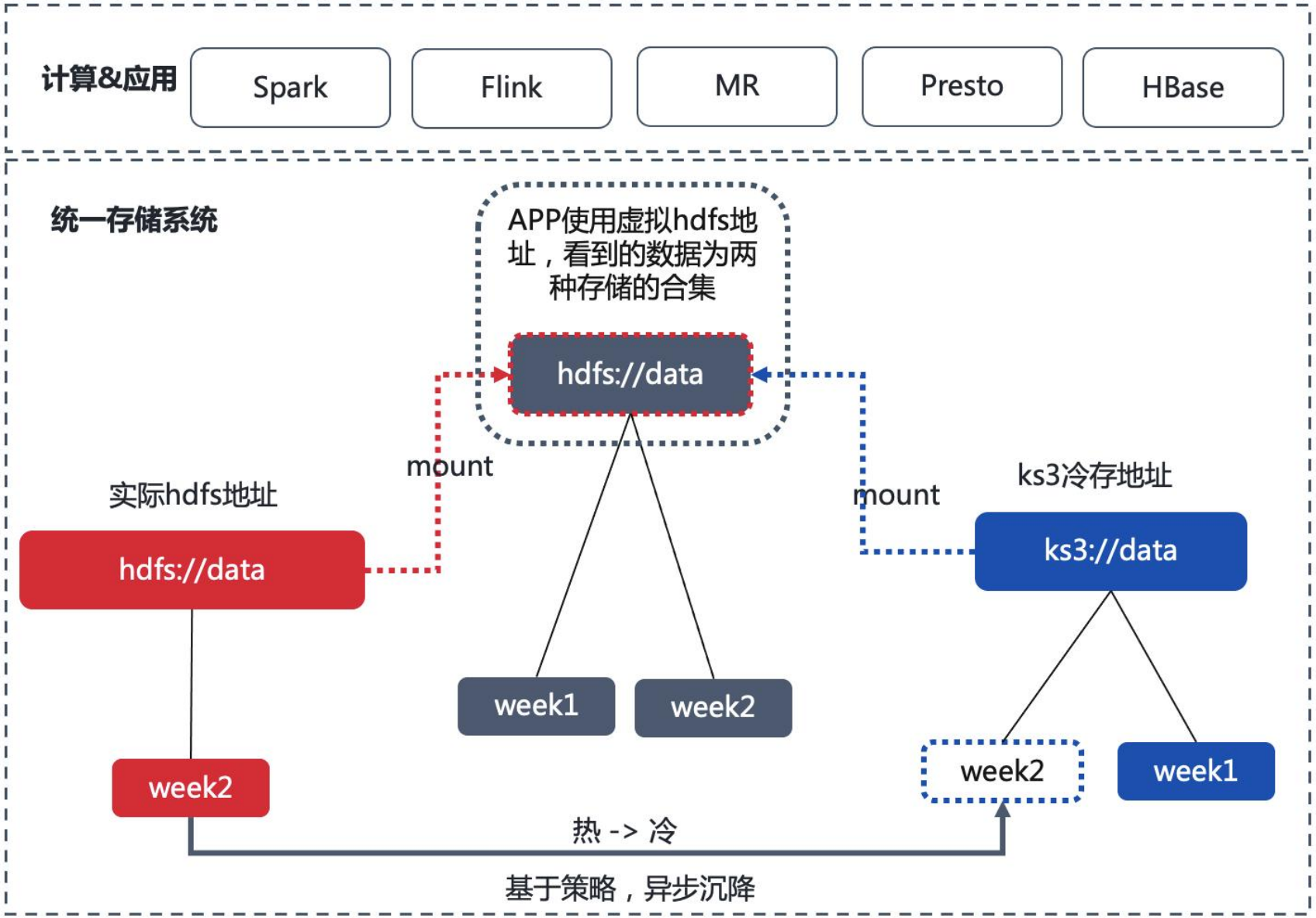
数据流

元数据流

1. 分块上传失败？
2. 写元数据到metaserver失败？
3. complete分块上传失败？
4. metaserver在任务不同阶段发生宕机？

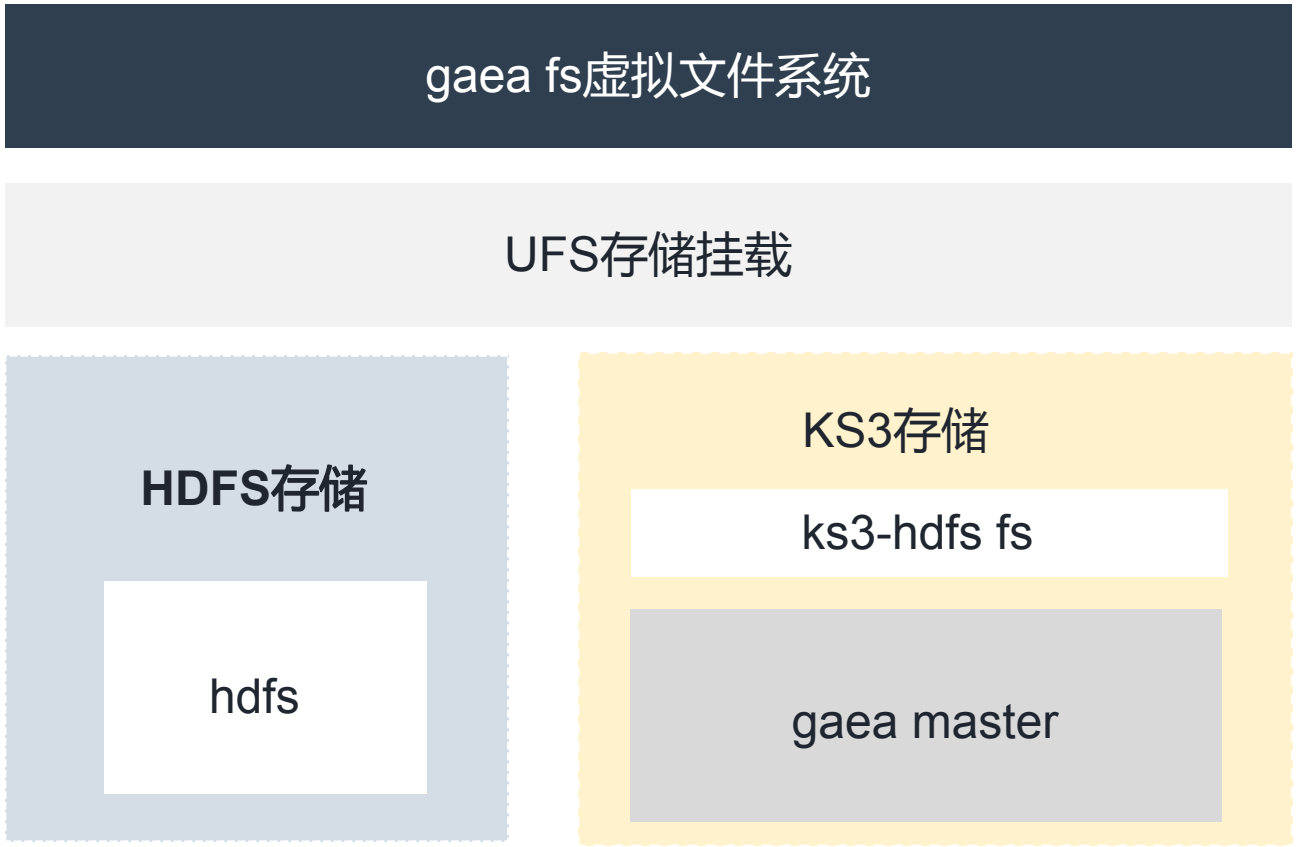
统一文件系统

怎么平滑将hdfs服务迁移到ks3？



统一文件系统逻辑结构

基于HDFS和对象存储构建的统一文件系统



gaea fs ：用户使用存储的统一入口，是一个虚拟文件系统。基于ks3-hdfs的client来实现。

-hdfs：原生hdfs文件系统

-ks3-hdfs：实际ks3-hdfs文件系统，在ks3 的基础上具备元数据加速能力

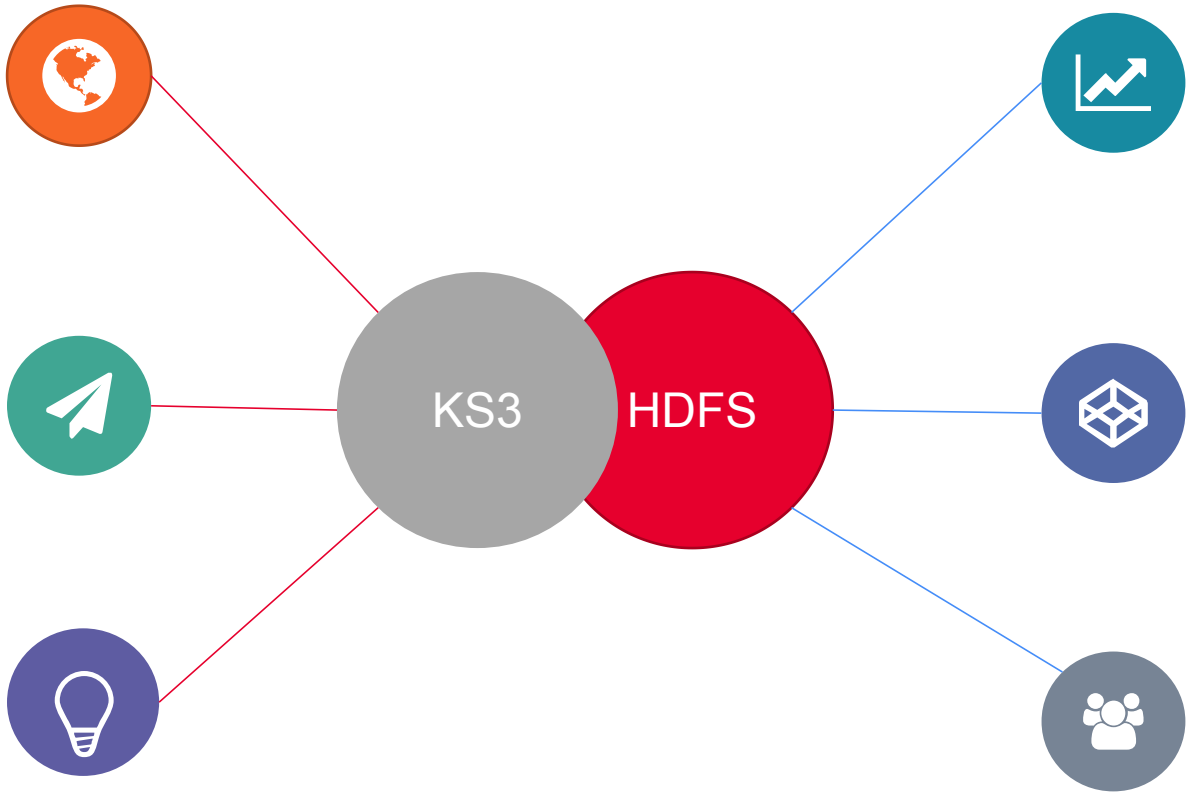
构建一个统一文件系统 gaea FS。gaeaFS根据配置路由接口到HDFS client 以及 ks3 hdfs client。

提升性能的核心点

单连接加速，
100MB/S

服务端缓存加速

支持Shadow copy
支持Batch delete
支持Batch copy



元数据缓存加速

客户端数据缓存加速

同步操作改异步

KS3-HDFS产品优势

金山云

简单易用

- 一键开通，数据自动更新
- 开发人员无需改动业务调用方式，安装SDK即可轻松使用
- 丰富的API、SDK

兼容性高

- 兼容HDFS语义
- 对象存储数据组织形式不变，数据视图一致
- 对现有服务零入侵

安全可靠

- 元数据多副本数据存储
- 完善的权限管理确保访问安全
- 支持双向数据同步，保证数据一致性
- 完善的监控告警系统

成本低廉

- Serverless化服务，按需使用
- 免去服务运维人力
- 数据存储成本低廉

存算分离性能指标

金山云

元数据服务

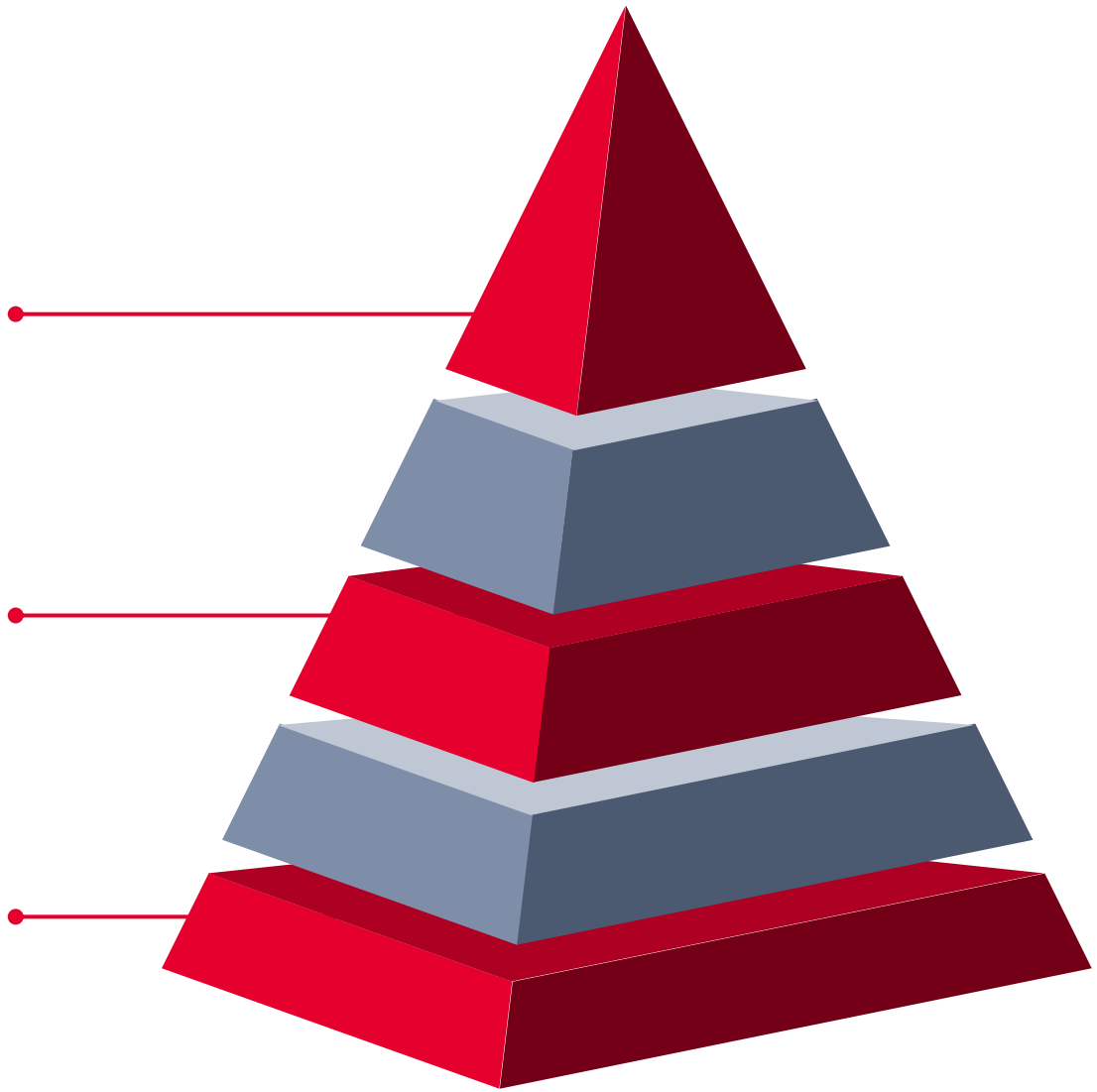
- 毫秒级文件Rename：无需Copy/Delete数据
- 单Bucket 10w QPS
- 读写混合场景下：List 10w 文件百ms

数据服务

- 单连接带宽：100+MB/s
- 百Gbps级别带宽：满足高吞吐需求
- ms级别99时延

服务稳定可靠

- 提供99.999999999%的数据可靠性保障
- 标准存储提供99.95%的访问可用性

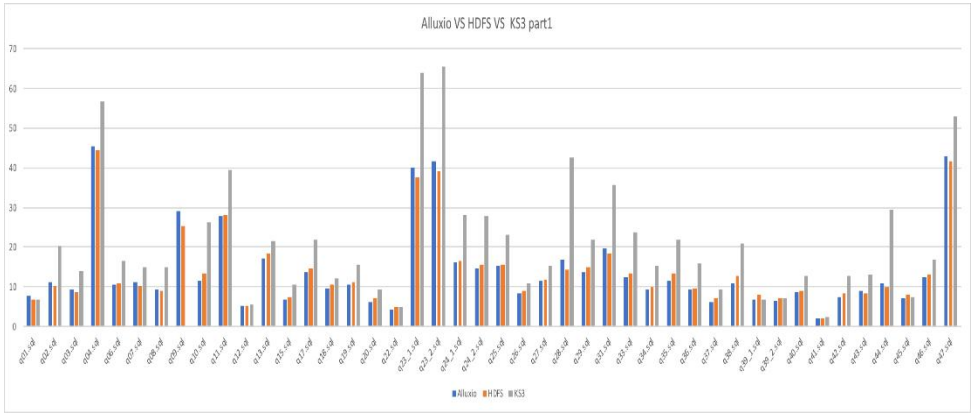


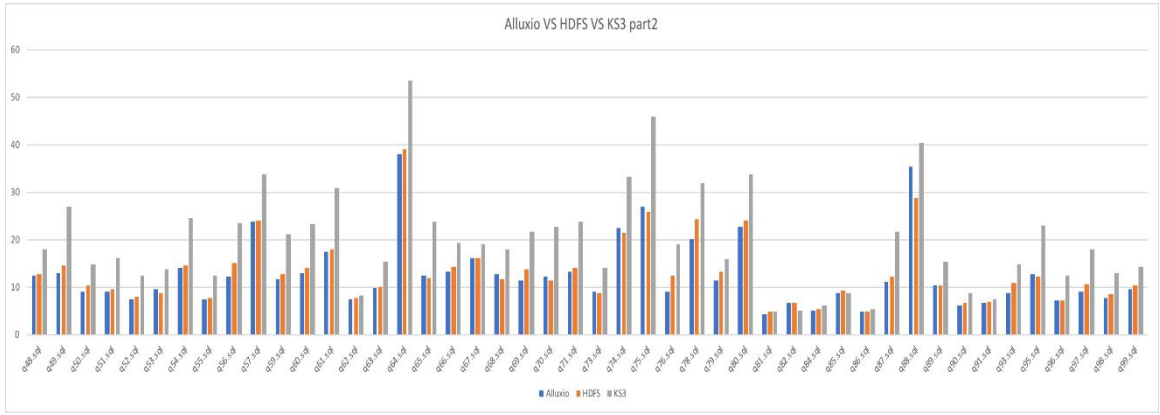
存算分离性KS3-HDFS能对比

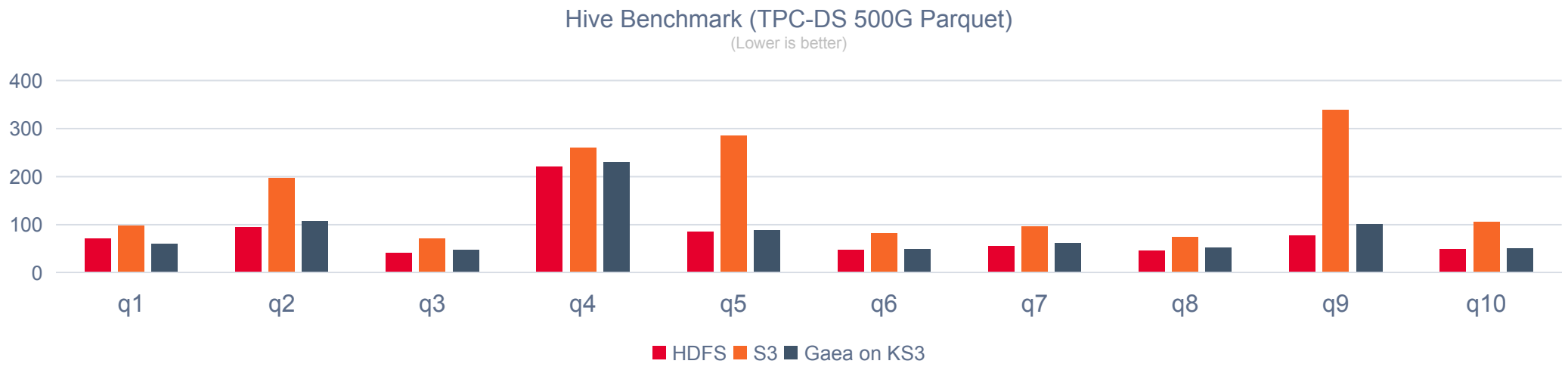
金山云

Presto 场景测试

Gaea平均总用时比ks3提高约36%，比hdfs提高约3%







参考脚本：http://www.tpc.org/tpcds/

某客户大数据存算分离方案

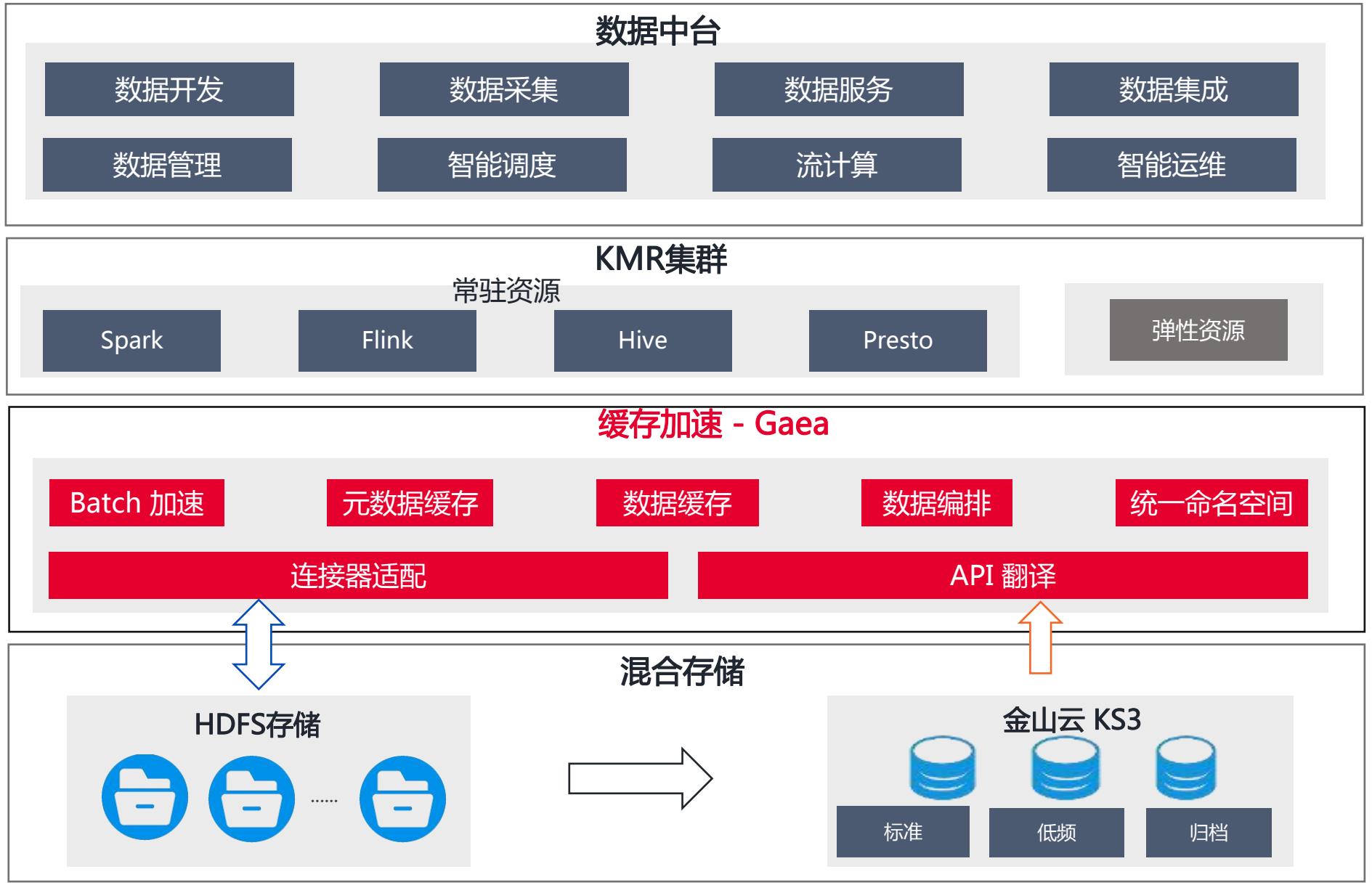
金山云

方案特点

- 计算资源灵活弹性伸缩
- 存储计算分离
- 数据冷热分离
- 数据按需预加载

收益

- 成本较低
- 数据高可靠
- 高性能，降低内网带宽消耗



金山云

感谢你的收听

Thank you for listening