快手大数据架构介绍&databend潜在应用场景

房孝敬

快手大数据架构师

调度&计算研发工作

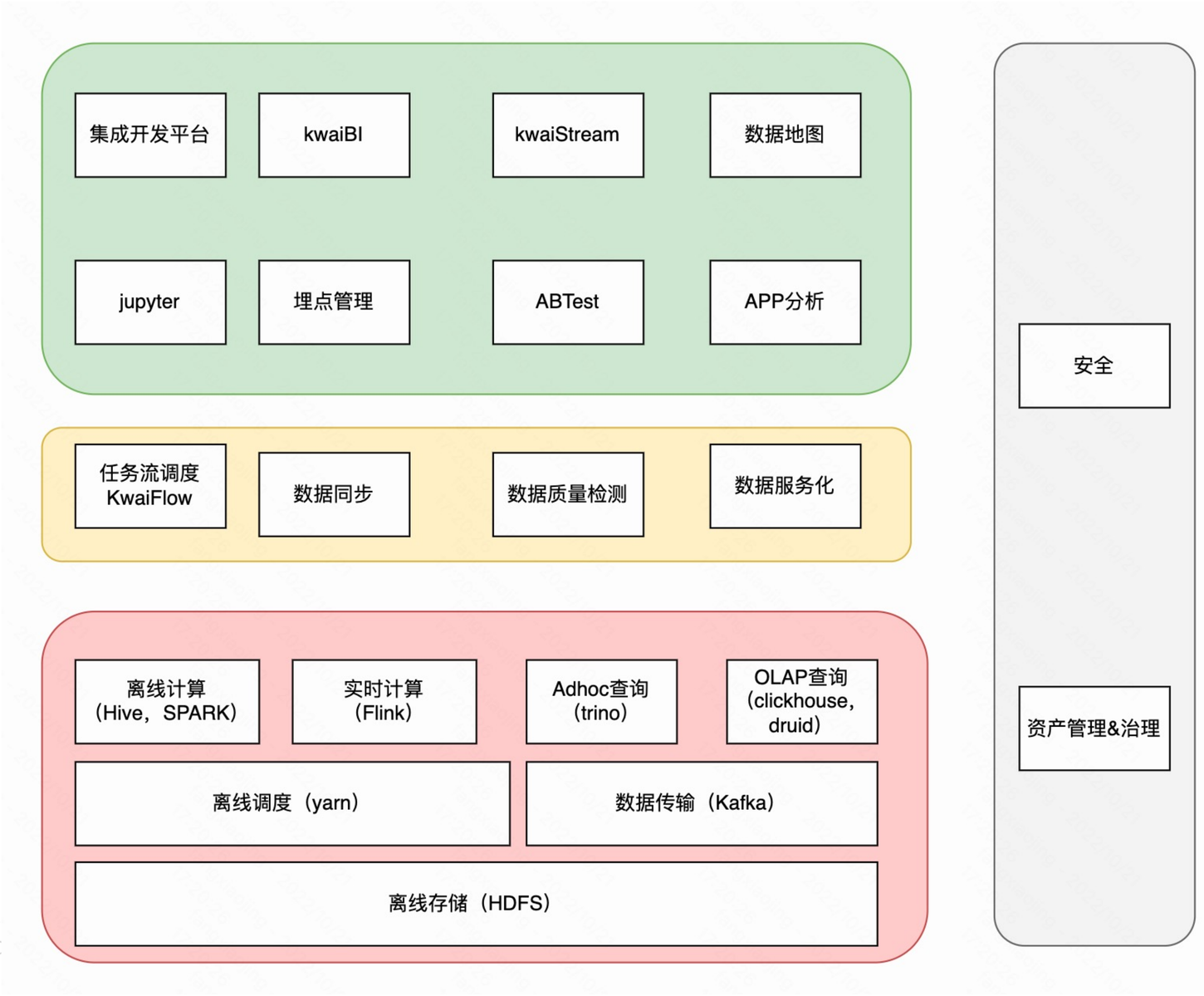
快手

提纲

- 大数据整体流程
- 实践经验
- why databend

快手

快手大数据平台



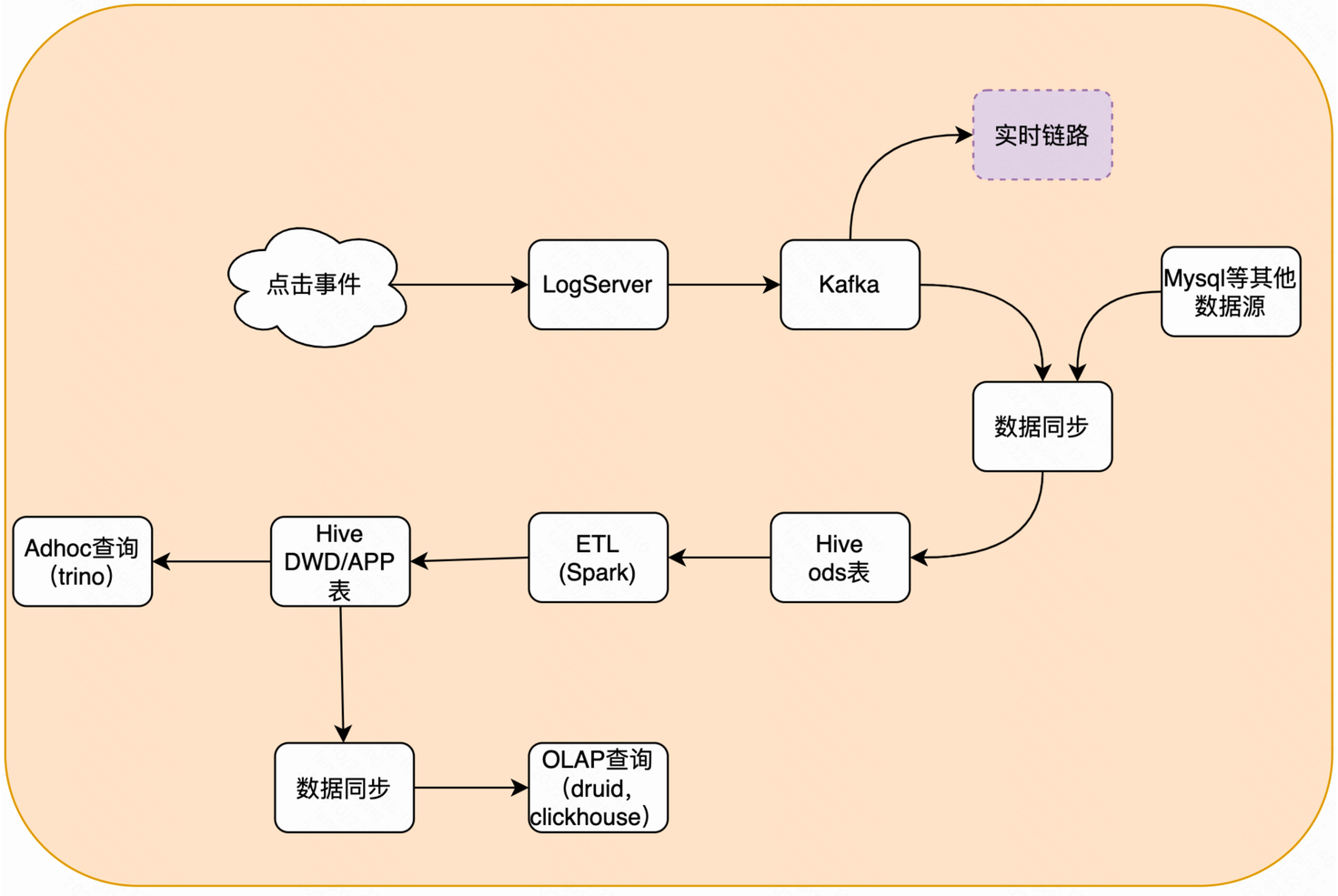
数据应用

数据工具链

数据架构

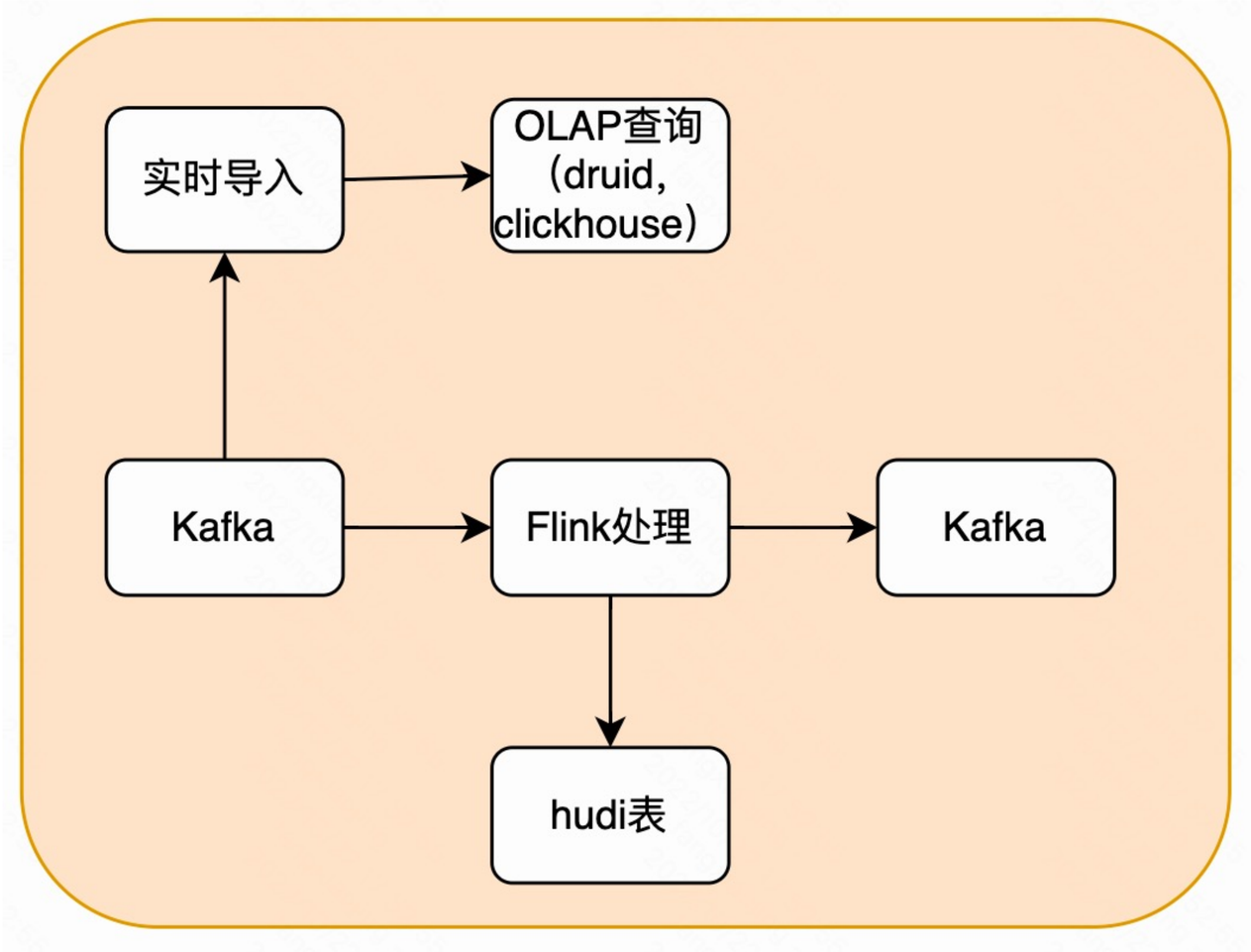
快手

数据奇遇记 一条日志离线处理之路



快手

实时链路



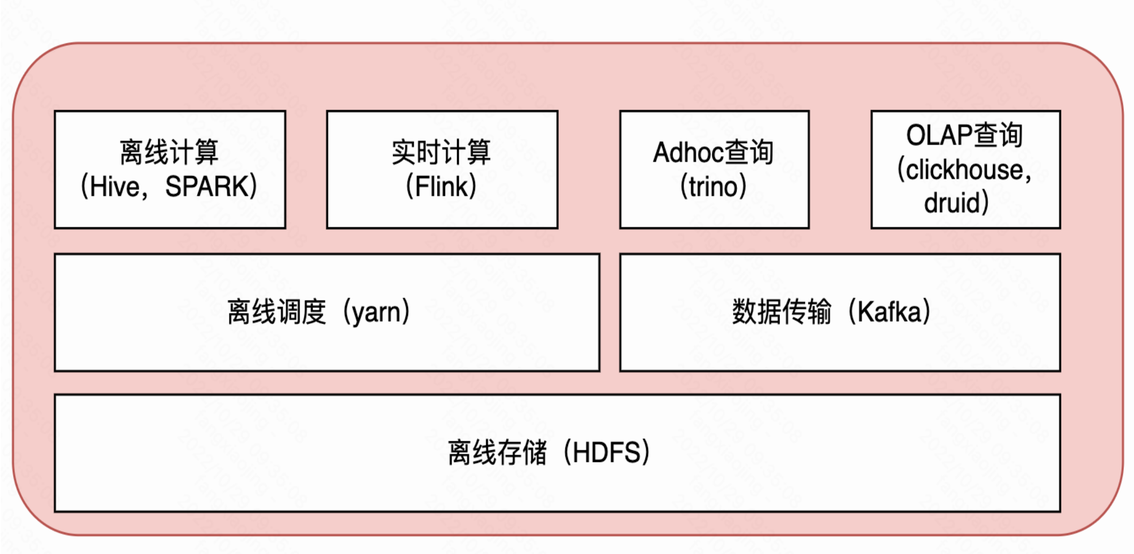
快手

提纲

- 大数据整体流程
- 实践经验
- why databend

快手

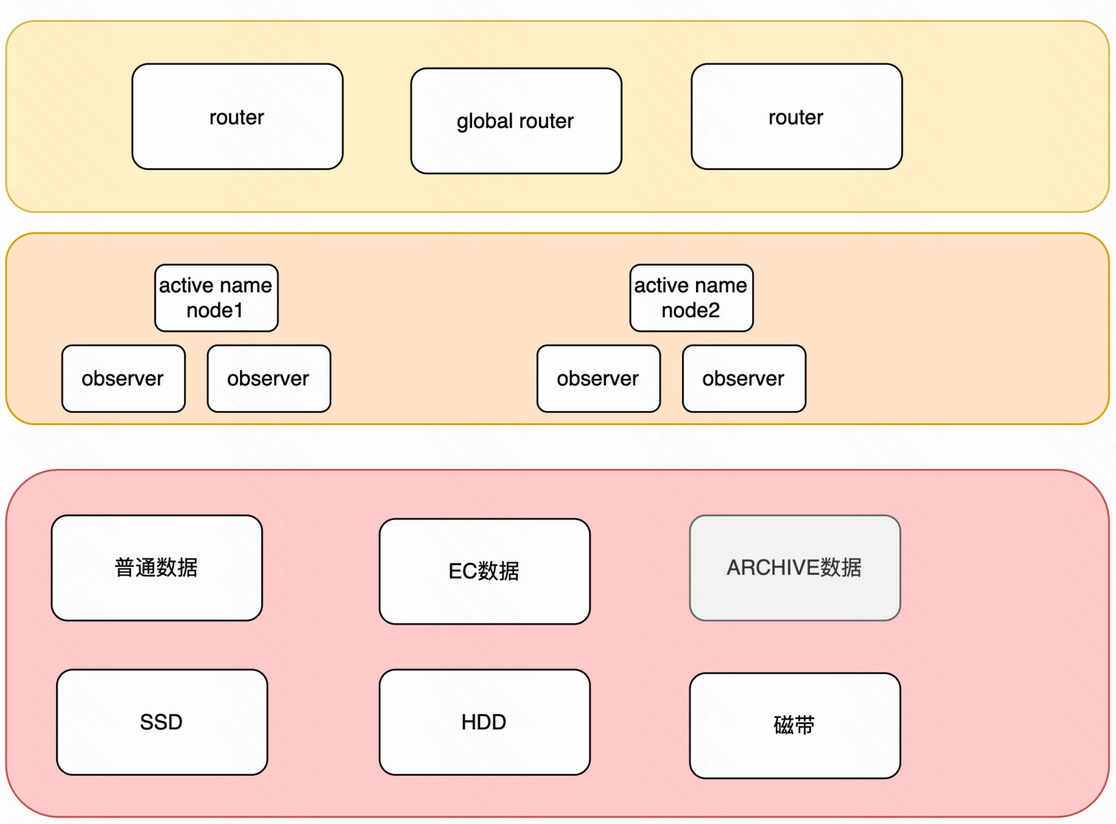
大数据架构规模



- 数据量：数EB
- 单集群规模： 数万台
- 跨AZ统一集群
- 每天作业量： 百万

快手

大数据底座 HDFS



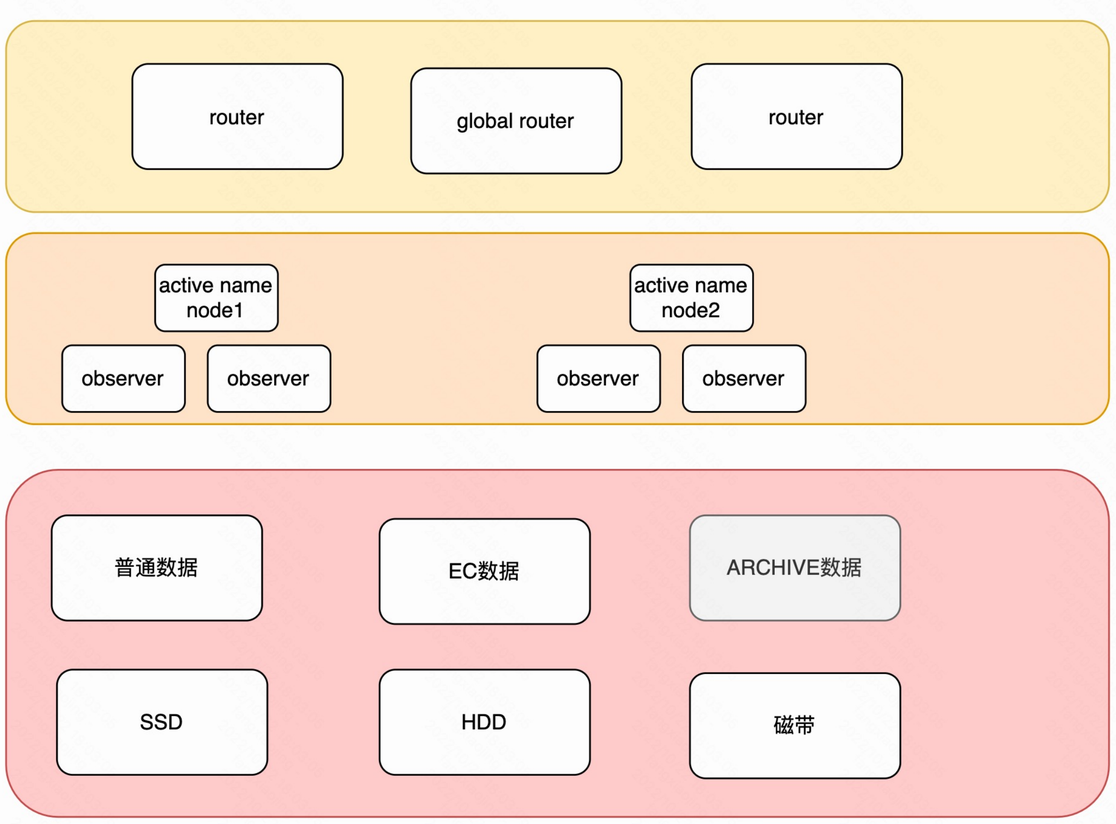
基于廉价设备提供的高可用性，高扩展性文件存储解决方案

问题：

- 小文件过多, NN内存压力大
- NameNode性能
- 3副本存储

快手

大数据底座 HDFS



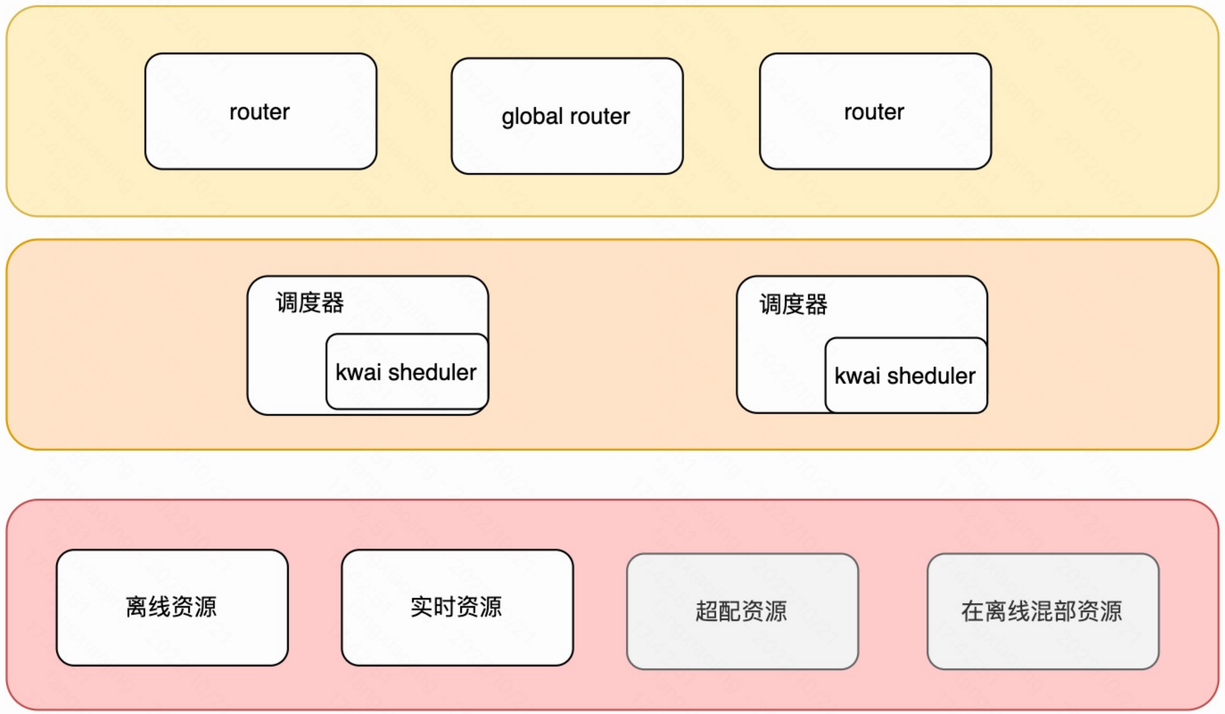
元数据管理：raid, 元数据本地磁盘存储， router架构

性能： nn拆锁，observer node，分级保障

成本： ec， 冷热数据分层存储

快手

大数据搬运工 YARN



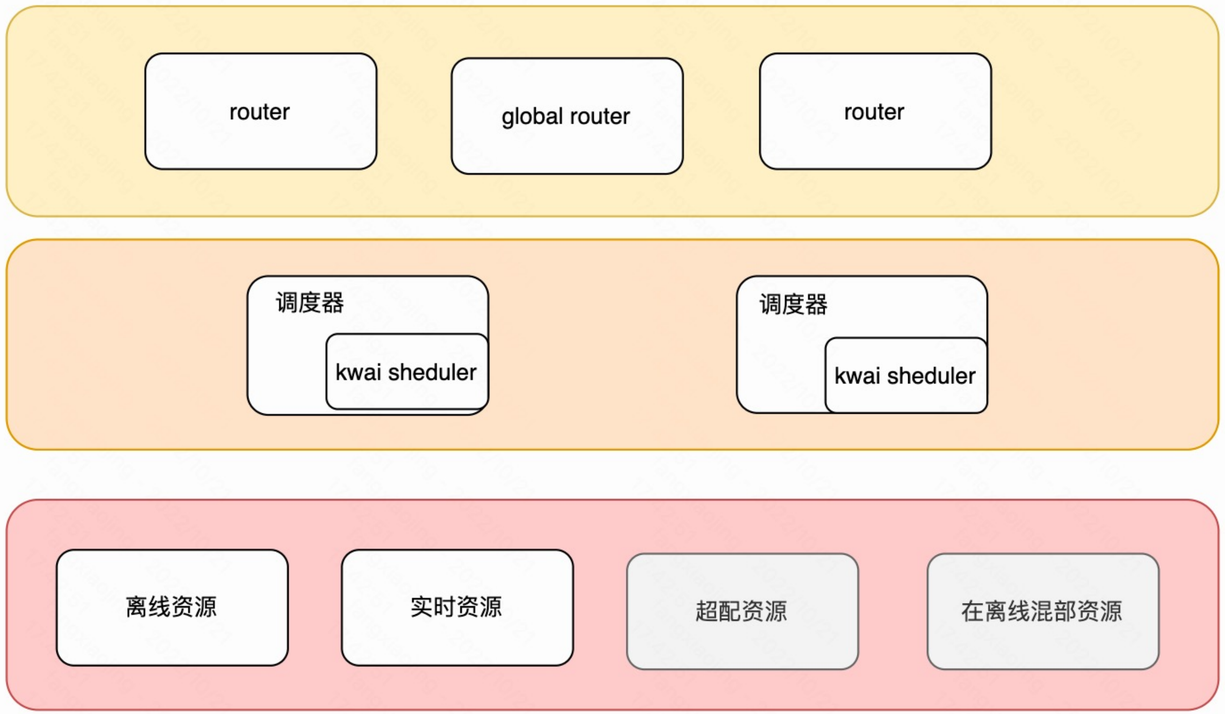
资源池化，资源强保障的解决方案

问题：

- RM调度瓶颈
- 机器资源利用率

快手

大数据搬运工 YARN



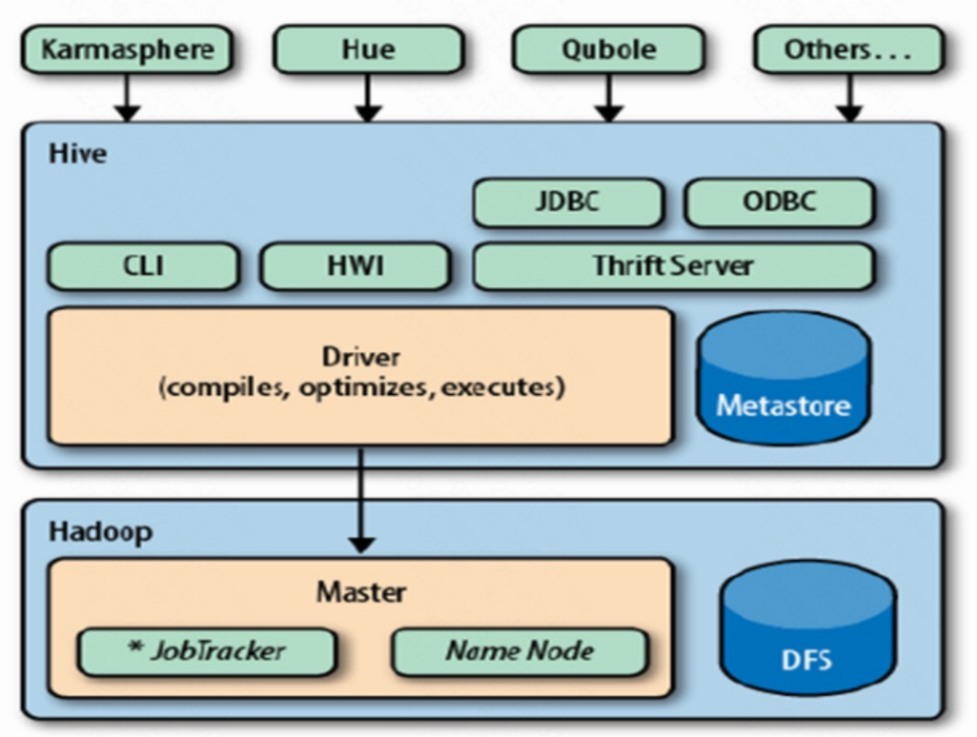
性能： kwai scheduler

扩展性： router架构

成本： 离线资源超配，
在离线资源混合部署

快手

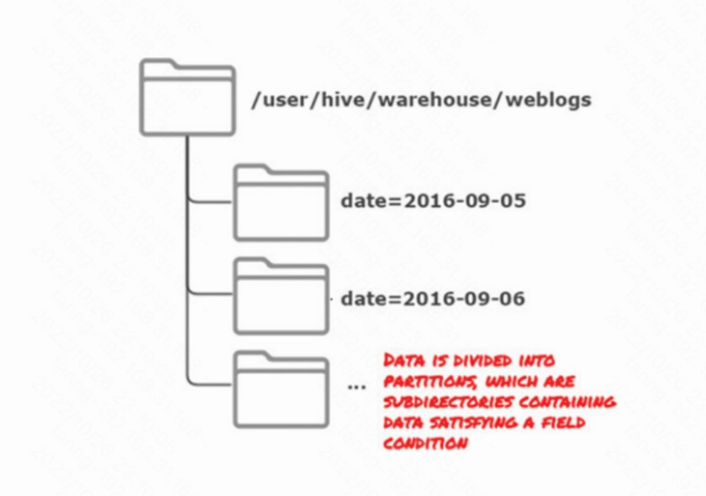
Hive原始架构介绍



基于hdfs存储能力，
hadoop计算能力，提供的
大数据数仓解决方案

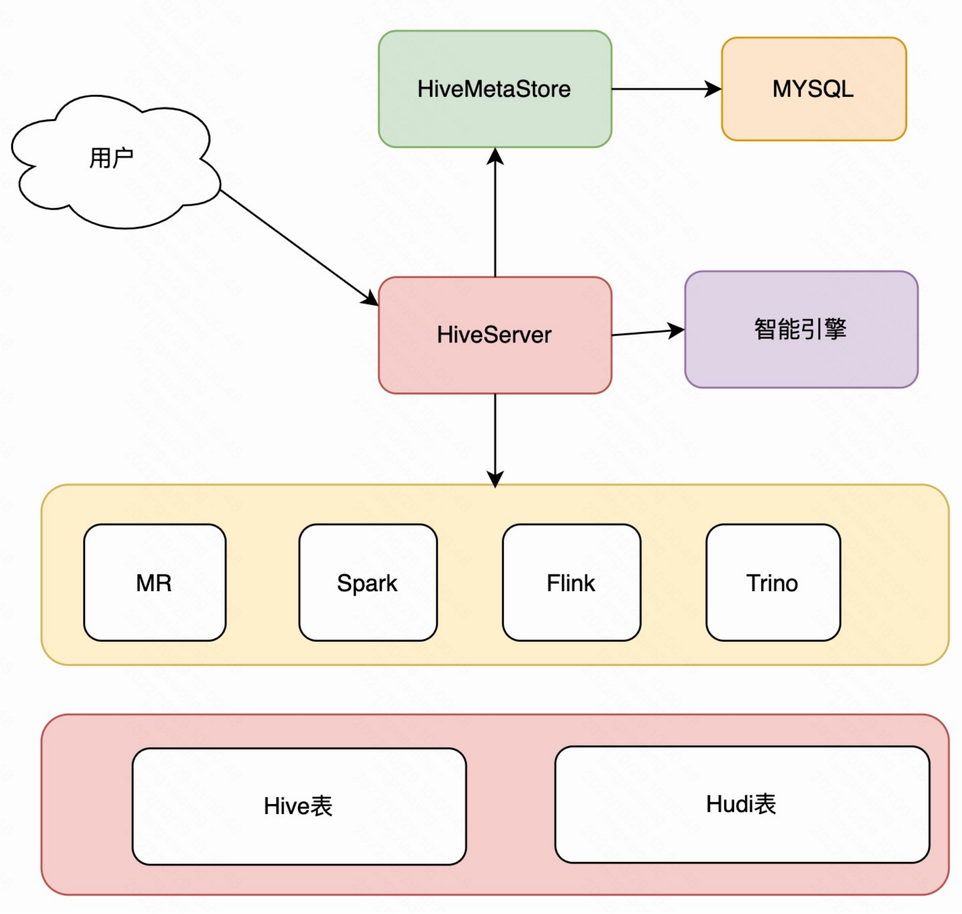
特点：

- 元信息 分层存储
- 分区&分桶（动态分区）
- 不支持数据修改&删除



快手

大数据魔法师 计算

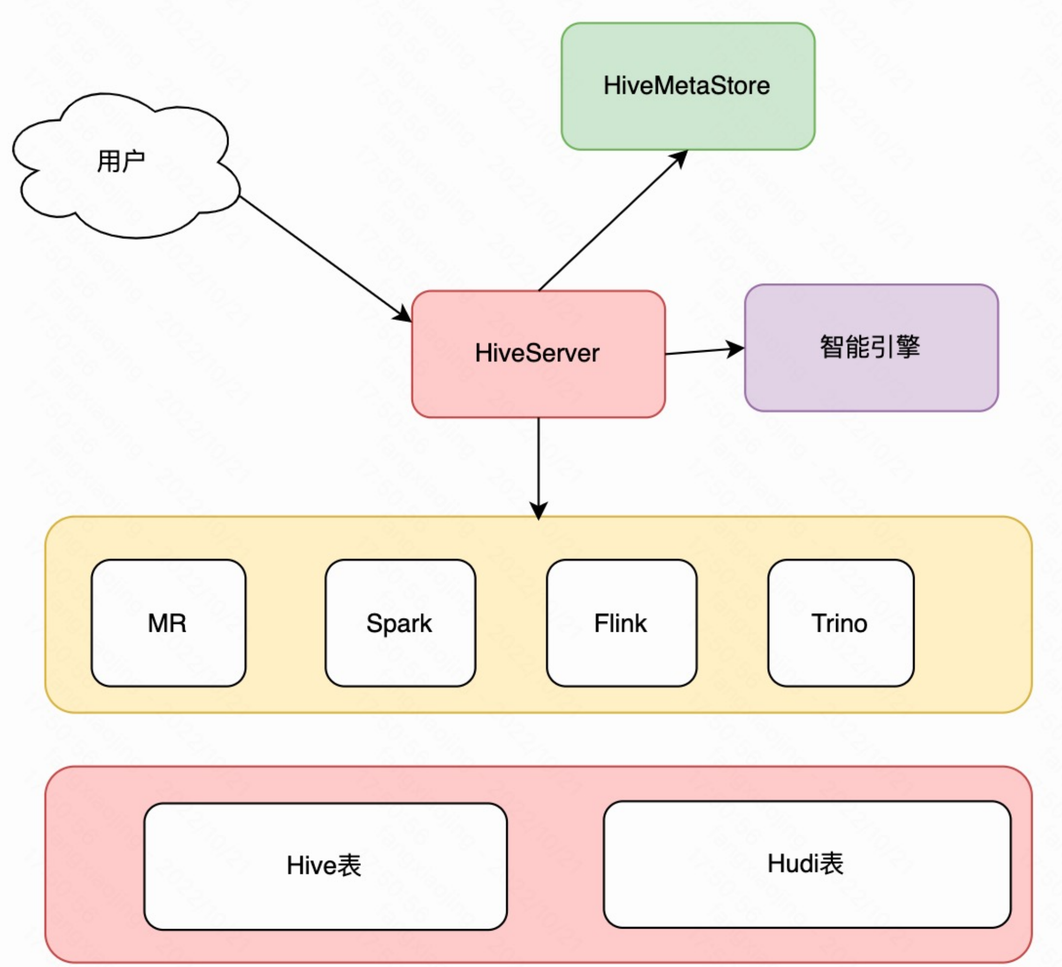


问题：

- Metastore 容量&性能瓶颈
- Spark 大作业shuffle稳定性
- 计算引擎算力

快手

大数据魔法师 计算



扩展性： hive metastore读写分离，federation改造

稳定性： **spark shuffle** 优化，

RSS

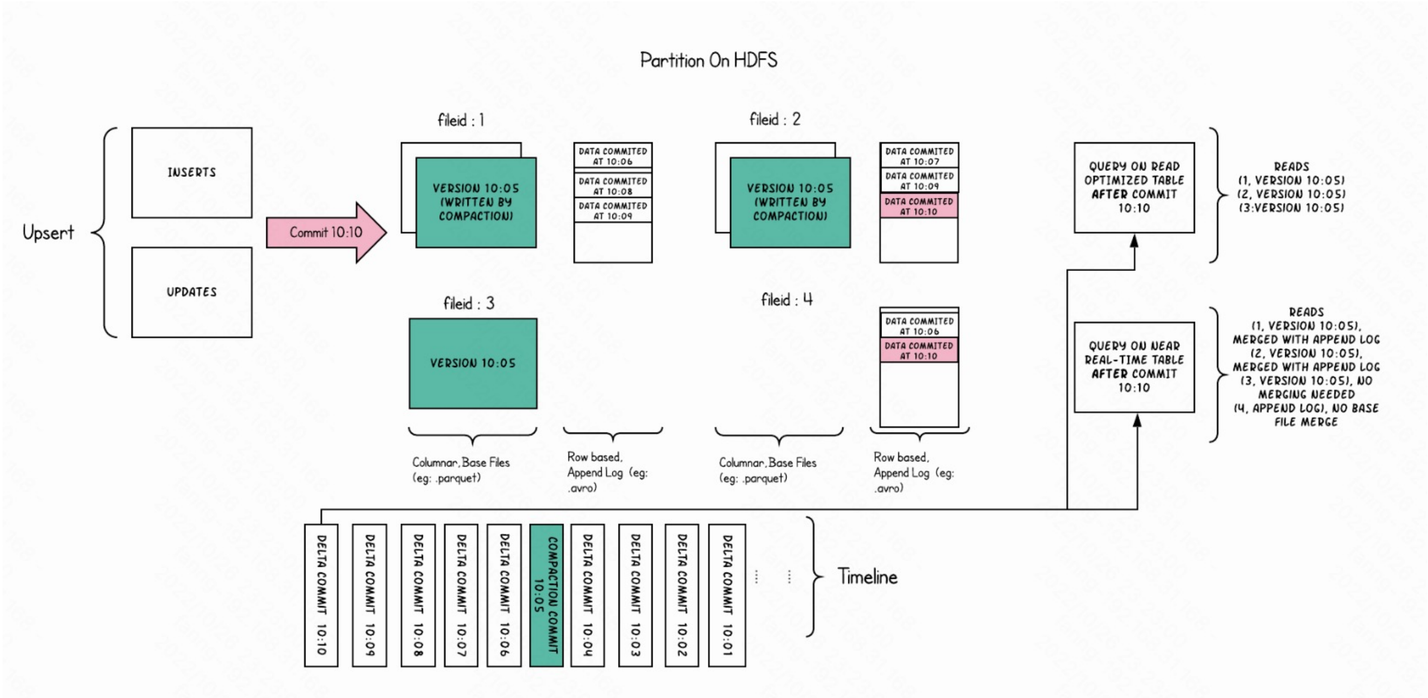
成本： mr->spark，blaze项目

实时性： **hudi**

效率：智能路由

快手

Hudi介绍



新的数据组织方式

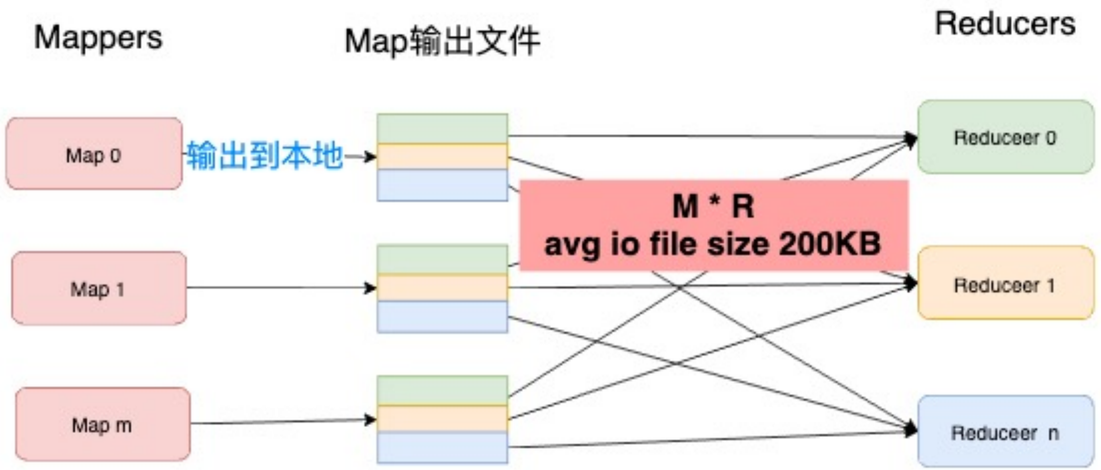
Delta写入，支持数据快速upsert

数据分区&分桶分布

支持主键

快手

MR&Spark 原有shuffle模式



stage调度 vs MPP

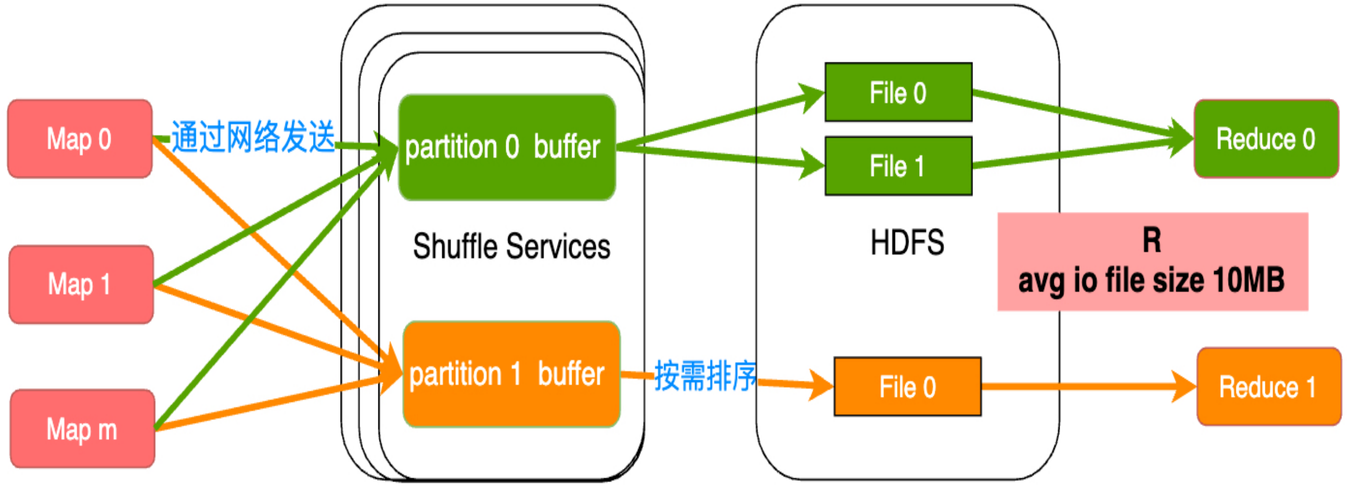
小IO

单盘热点，作业稳定性差

存算一体，不利于在离线混部

快手

Remote shuffle service



顺序大IO

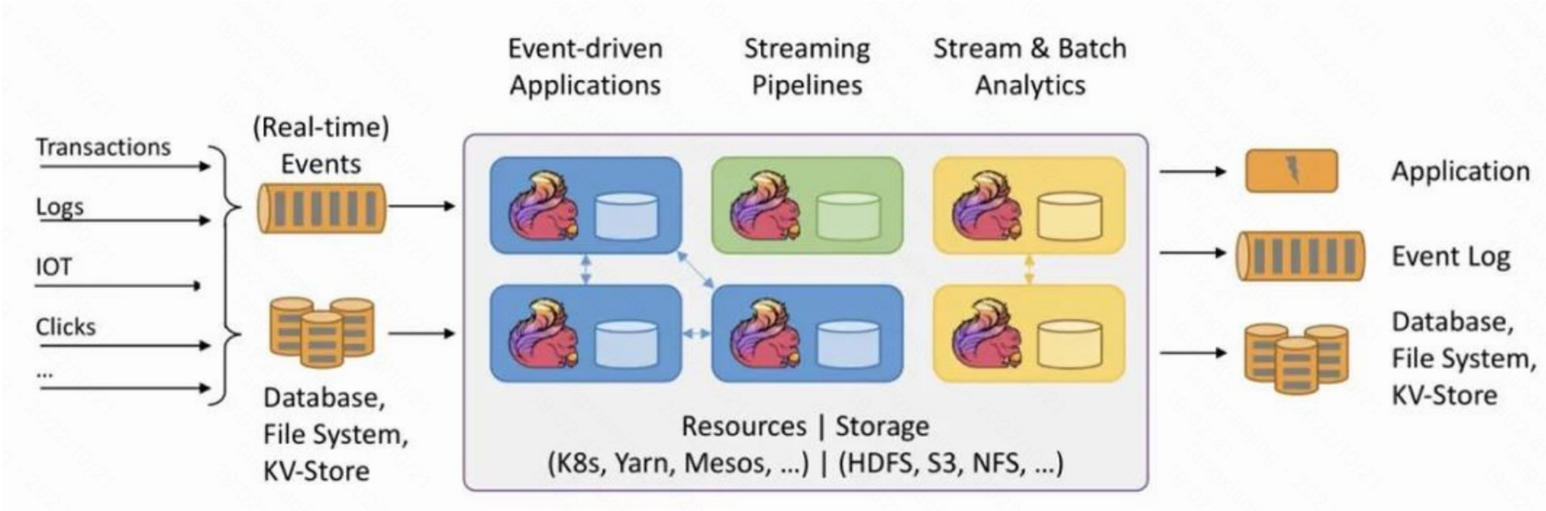
提升大shuffle作业稳定性

计算存算分离

资源混部

快手

实时引擎（Flink）



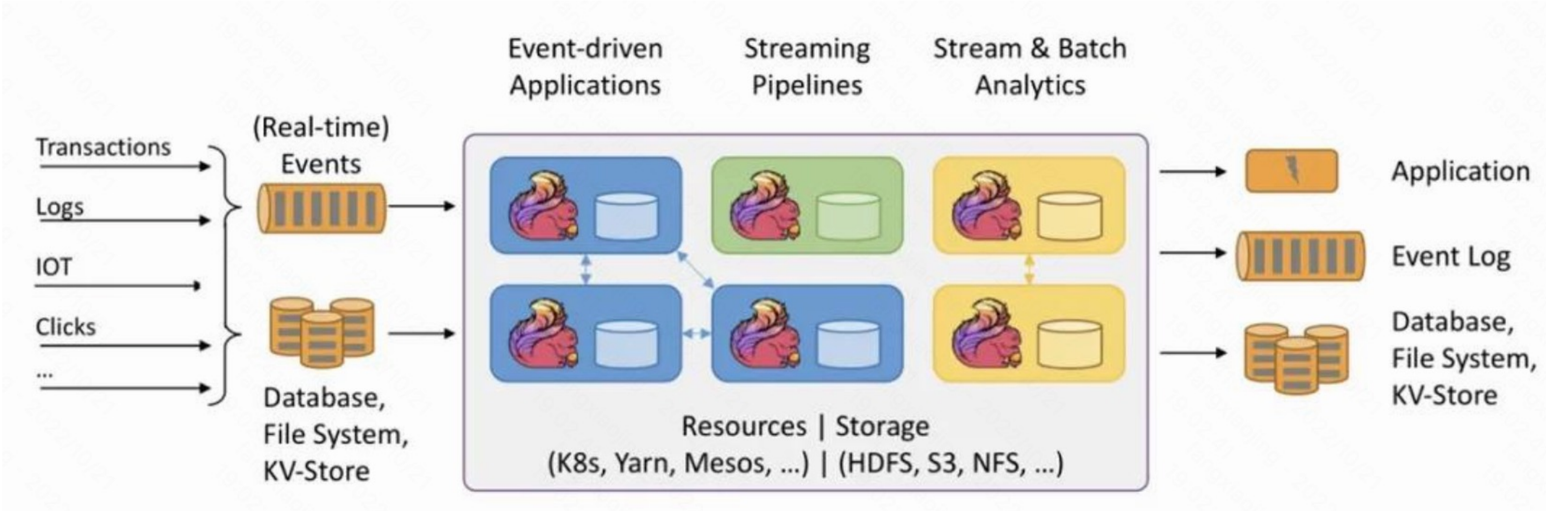
有状态的，提供精准一致保障的实时计算引擎

问题：

- 稳定性
- 开发效率

快手

实时引擎（Flink）



稳定性

慢节点&坏节点发现，规避，快速恢复

双链路跨机房容灾

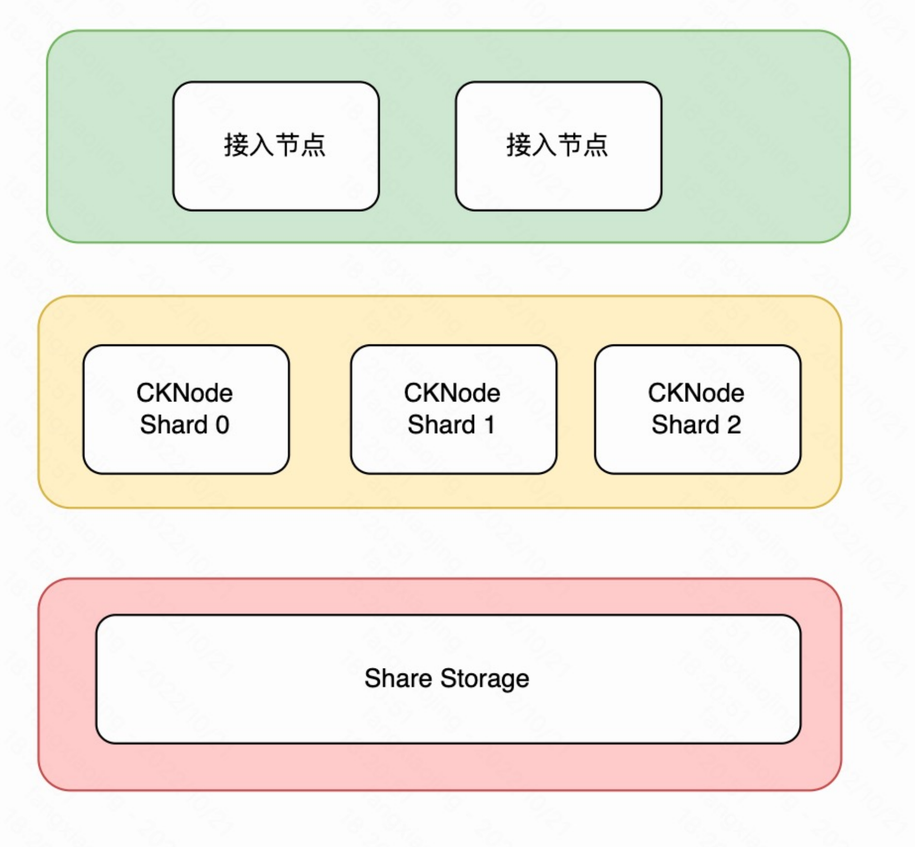
Flink存算分离

FlinkSQL

流批一体

快手

OLAP查询引擎（CK）



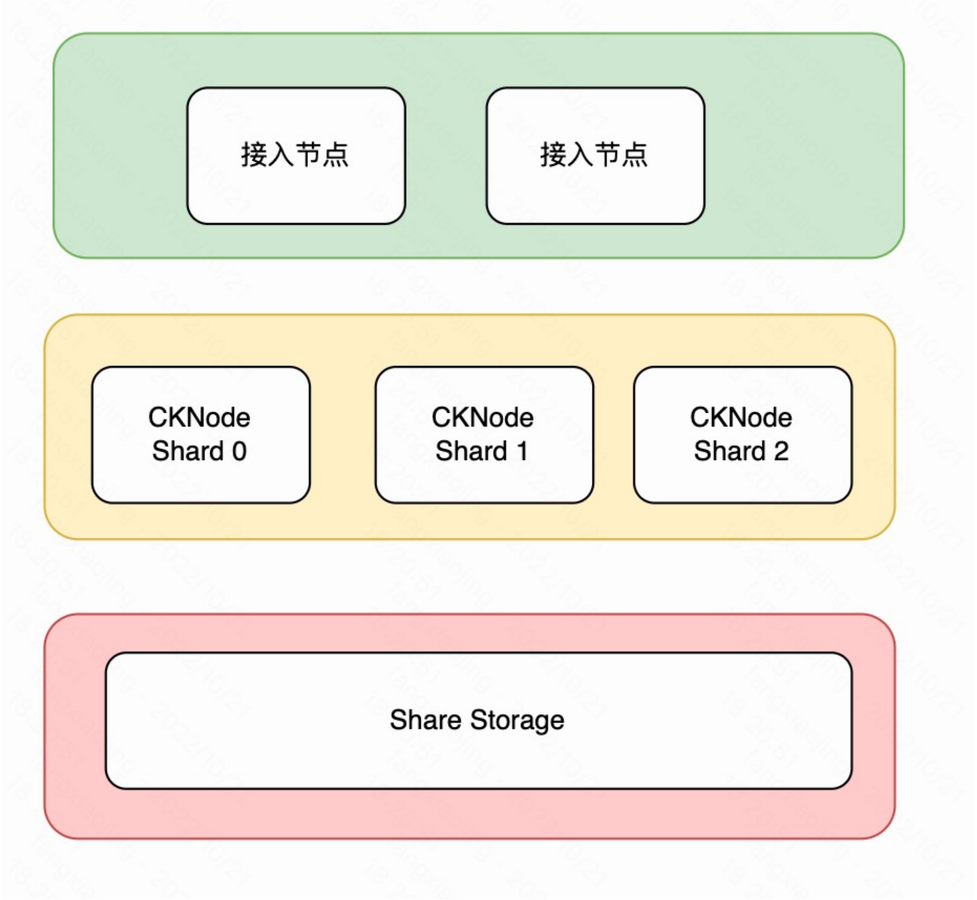
通过索引，物化视图等功能提供高效的多维数据分析洞察能力

问题：

- 数据导入一致性
- 可运维性
- JOIN支持差

快手

OLAP查询引擎（CK）



离线和实时数据导入一致性保障

可运维性： ck 存算分离

性能：ck projection

快手

提纲

- 大数据整体流程
- 实践经验
- why databend

快手

Adhoc 查询

速度可以更快？

成本更低？

快手

OLAP查询需求

可运维性

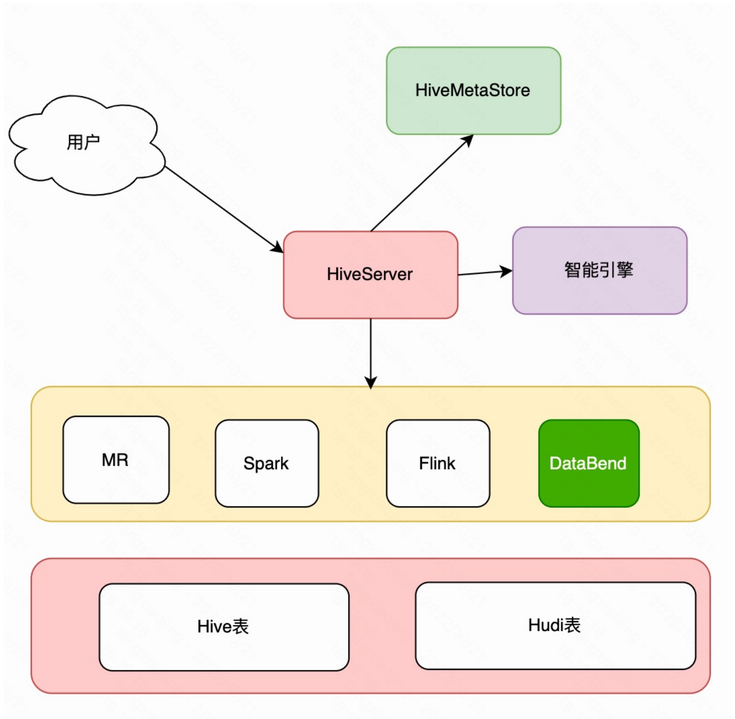
存算分离

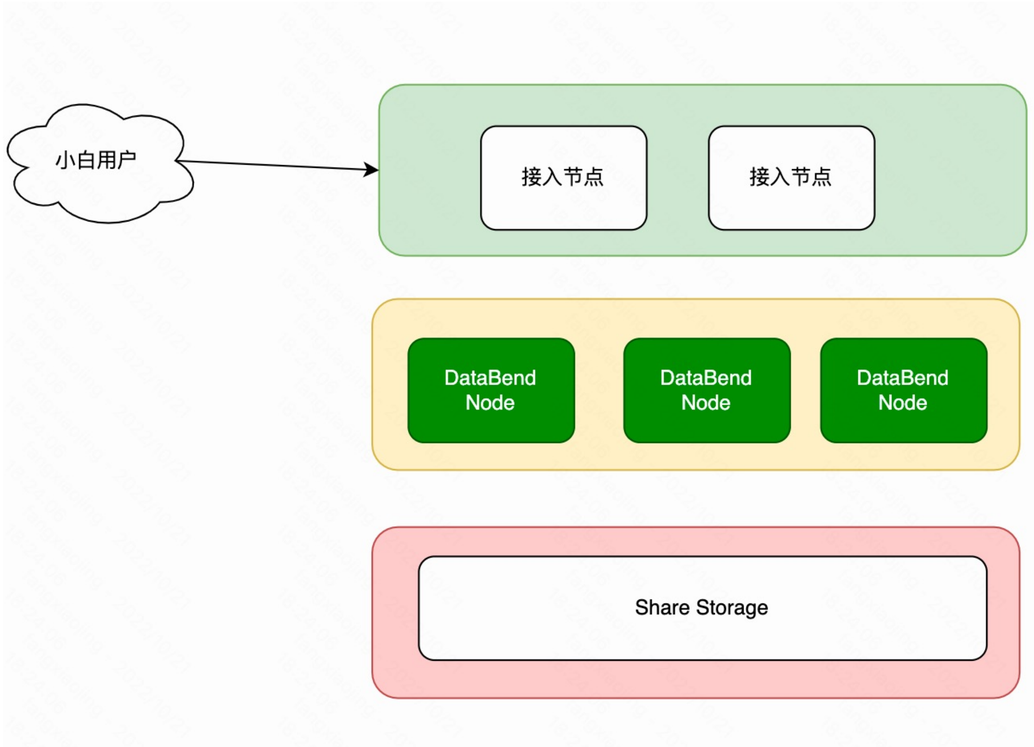
JOIN支持

数据导入， 数据导入一致性， 时间成本，资源成本

快手

Let it be





极致性能

存算分离

JOIN支持

To be done

数据一致性问题

Hive常用函数&用户java UDF支持

资源管控能力

JOIN性能

快手

THANKS

快手