
GOTC 2023
全球开源技术峰会
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE
OPEN SOURCE, INTO THE FUTURE
「基础设施与软件架构」专场
本期议题：Apache HugeGraph 分布式存储与计算开源演进之路
张世鸣 2023年05月28日

目录

- 图数据库 HugeGraph 的 Apache 开源之路
- HugeGraph 1.0 架构方案和瓶颈
- HugeGraph 2.0 分布式架构和原理详解
- 基于 HugeGraph 的图平台建设以及应用场景
- 未来的展望 & QA

什么是图数据库？

SQL

1. user

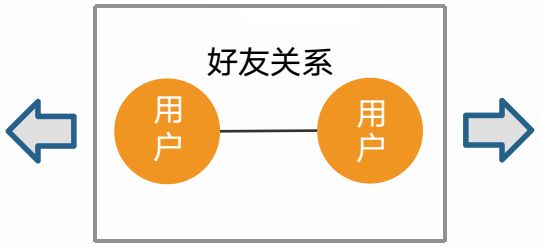
id	name	age	phone
1	Tom	22	188****1111
2	Mike	23	158****2222
...	...	...	...

2. friend

id	user1	user2	since
1	1	2	2016-01-02
2	...	...	...

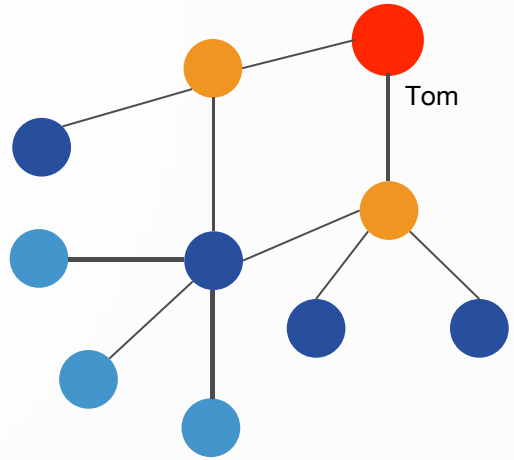
```
select * from friend a
  join friend b on b.user1 = a.user2
  join user on id=b.user2
where a.user1 in (select id from
  user where name='Tom')
```

表、SQL、Join



多关联关系的社交场景

图数据库




```
g.V().has('name', 'Tom')
  .out('friend')
  .out('friend')
```

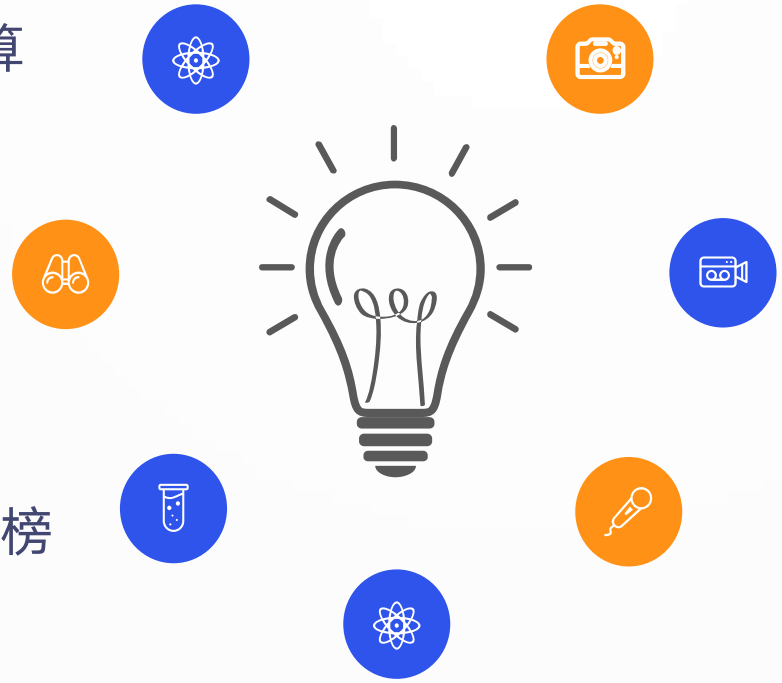
点、边迭代

Apache HugeGraph 简介

HugeGraph

大规模分布式图存储与图计算

生态完备：图存储、图计算、图可视化、图工具链（客户端、流批写入、备份、监控）



OLTP + OLAP

Gremlin + OpenCypher

国内最早开源的图数据库，
2023 “科创中国”开源创新榜

Apache 基金会唯一的图数据库&图计算项目

稳定、易用、可扩展

为什么加入 Apache？

开放治理，厂商中立

需求由社区驱动，由来自不同公司的开发者共同治理社区，而不是被某一家公司所把控，自由度会更高

协作和运营规范化

Apache 有一套成熟的协作和运营规范，并且会有专业的导师帮助项目的规范化

面向全球市场宣传

ASF 全球性的组织，20年的品牌积累，与良好的口碑。HG 希望吸引到更多的用户和贡献者来共建社区

Apache Way

Apache 是非常强调社区协作的一个组织，Apache Way 的核心就是社区高于代码，这也是 Apache 总结出来的可持续开源的成功经验

HugeGraph 开源发展历程

2016
立项研发

2018-08
正式开源

2021-12
第7个开源版本v0.12.0发布

2022-05
正式加入apache孵化

2023-02
第一个apache版本v1.0.0发布

now
分布式存储内部版本
与开源版本的适配和研发

stars
2.3k+

社区用户
100+

commits
10000+

8个
正式版本

fork
500+

下载量
2w+

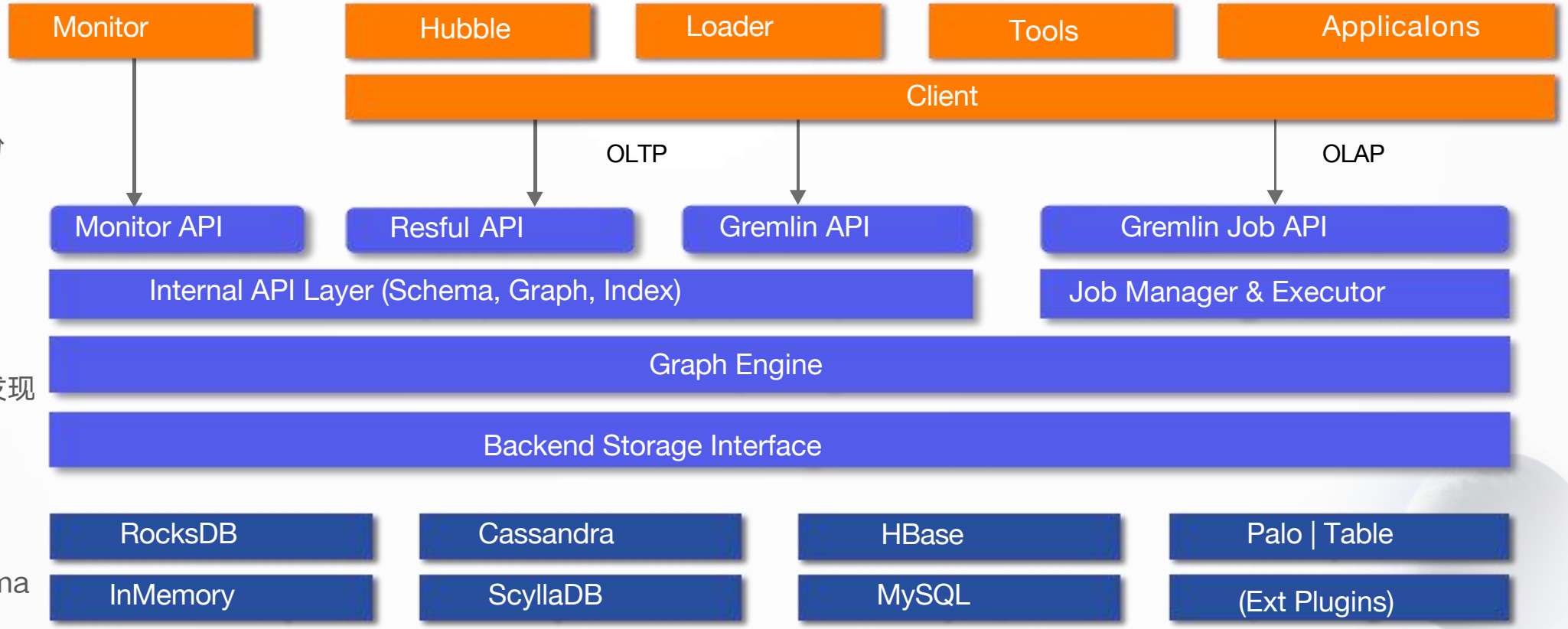
issues
1.6k+

HugeGraph 1.0 架构

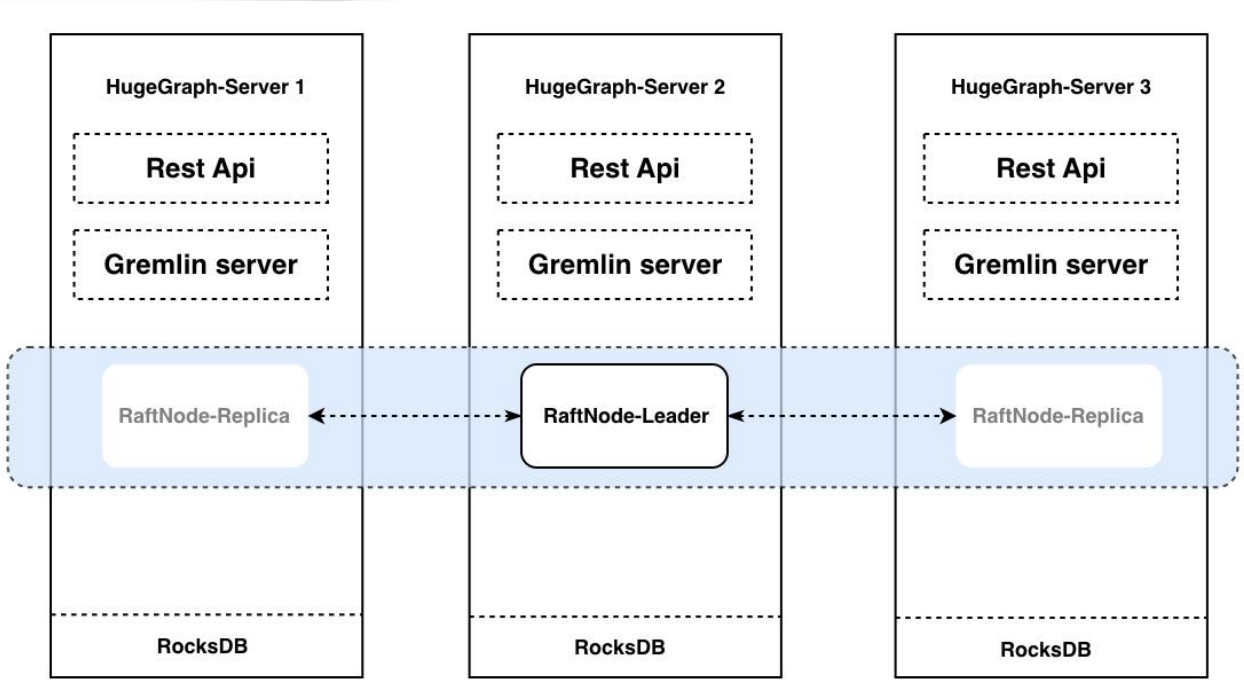
- 应用层
客户端
可视化、Loader、备份
监控

- 计算层
olap：环路检测/社区发现
oltp：readcommitted

- 存储层：
Graph/System/Schema
插件化



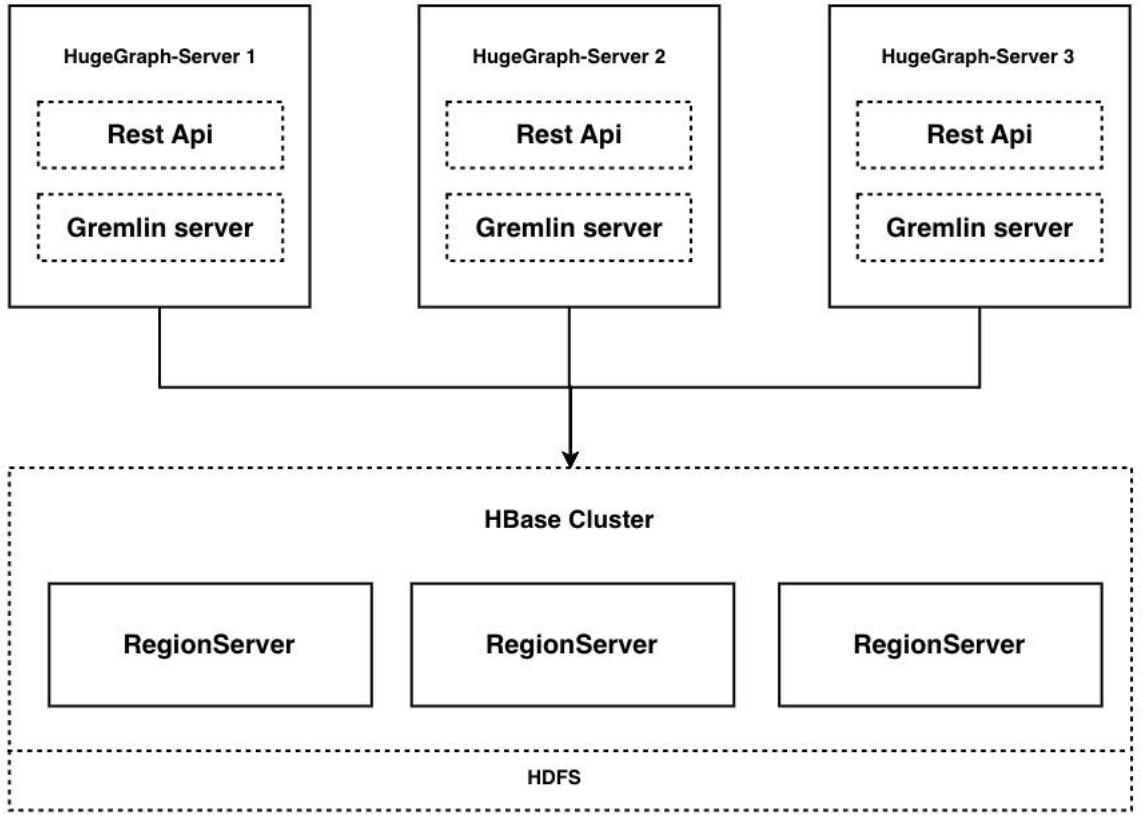
目前支持的分布式存储架构与瓶颈



基于 RocksDB 的 Replication 模式

优化：分片 + Multi Raftgroup

基于第三方的分布式存储（HBase/Cassandra...）



内存调优、集群管理、算子下沉

HugeGraph 2.0 的重大变化

01 万亿图数据 万张图存储

02 高性能读写

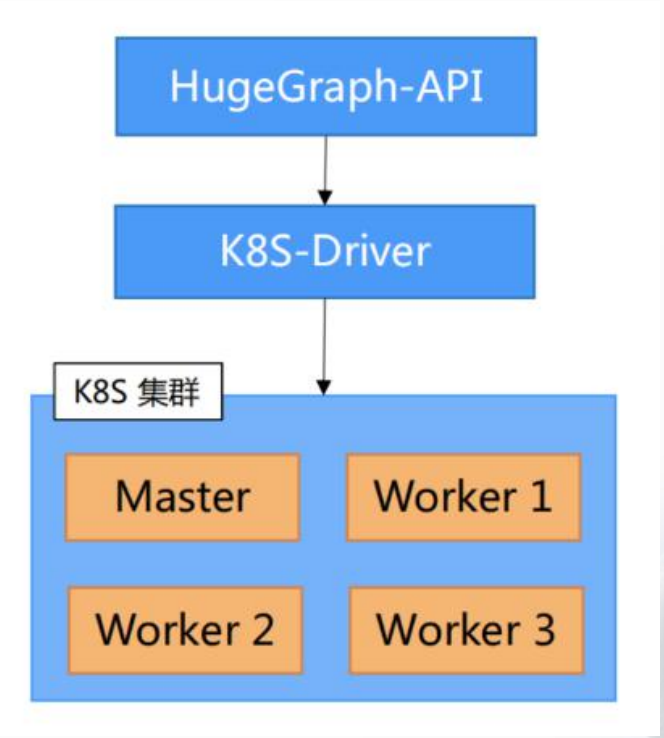
03 多活、高可用、动态伸缩，自动运维

04 使用便捷性

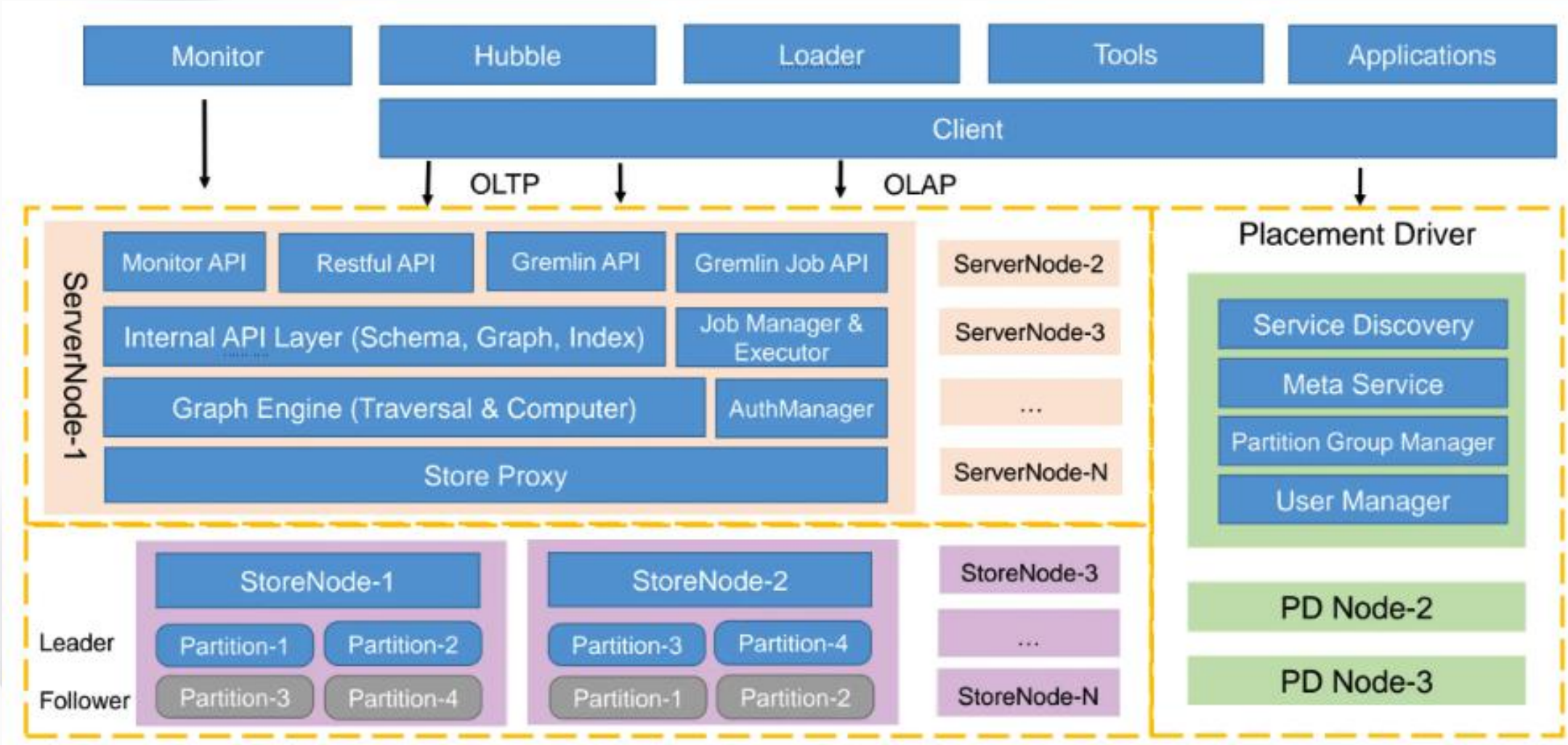
- 架构
  - 分布式架构，支持图数据分区+数据副本
  - 计算存储分离框架，便于计算及存储灵活伸缩
- 集群管理
  - 高可用设计，支持容灾及故障恢复
  - raft 相关的管理接口
  - 更多监控指标
- 查询
  - 算子下沉
  - gremlin 并行化
  - 细粒度的内存管理
- 使用（接口优化+功能增强）
  - 接口增加统计信息（遍历的顶点、边和耗时）
  - 支持动态创建图
  - unique索引可以查询
  - .....

HugeGraph 2.0 的重大变化

- 图计算 HugeGraph-computer
  - 易用性：与 hugeraph-server 做集成，内置 k8s driver，用户可以通过 API 方式调用算法和获取结果
  - 性能优化
    - 中间数据做 snapshot，（同一张图的不通图算法执行）
    - 计算过程中清理旧文件，减少磁盘占用
    - 有向图转无向图功能（加载数据时增加反向边）
    - 加载数据时过滤属性功能
  - 更多的图算法支持



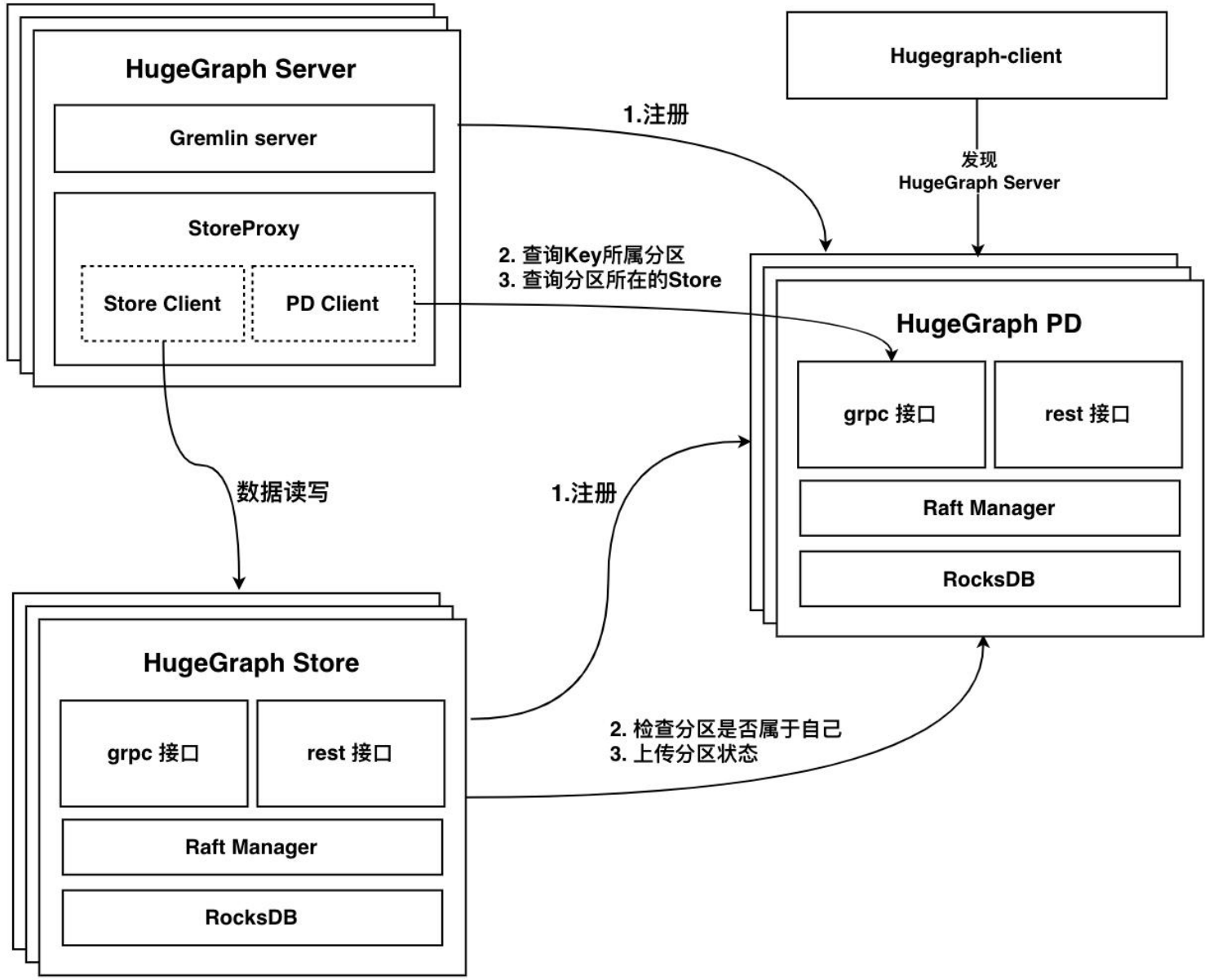
分布式存储架构设计



- leader 选举和日志同步
- 基于 rocksdb
- 逻辑分区

分布式架构交互逻辑

- 通信接口
- storeproxy
  - 缓存
  - 查询并行化
  - 网络迭代器，屏蔽底层store的差异



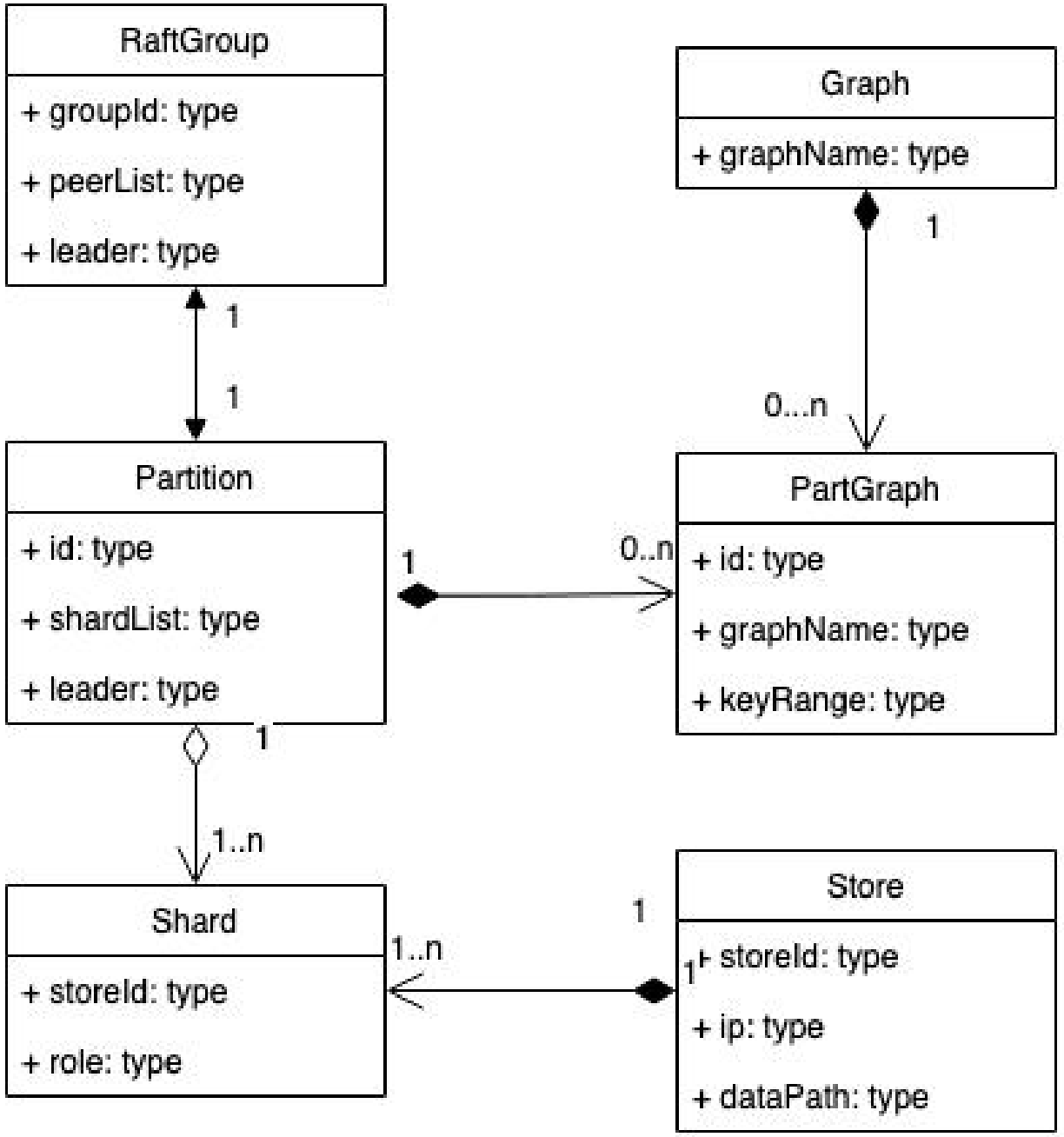
数据模型、分区、副本策略

分区

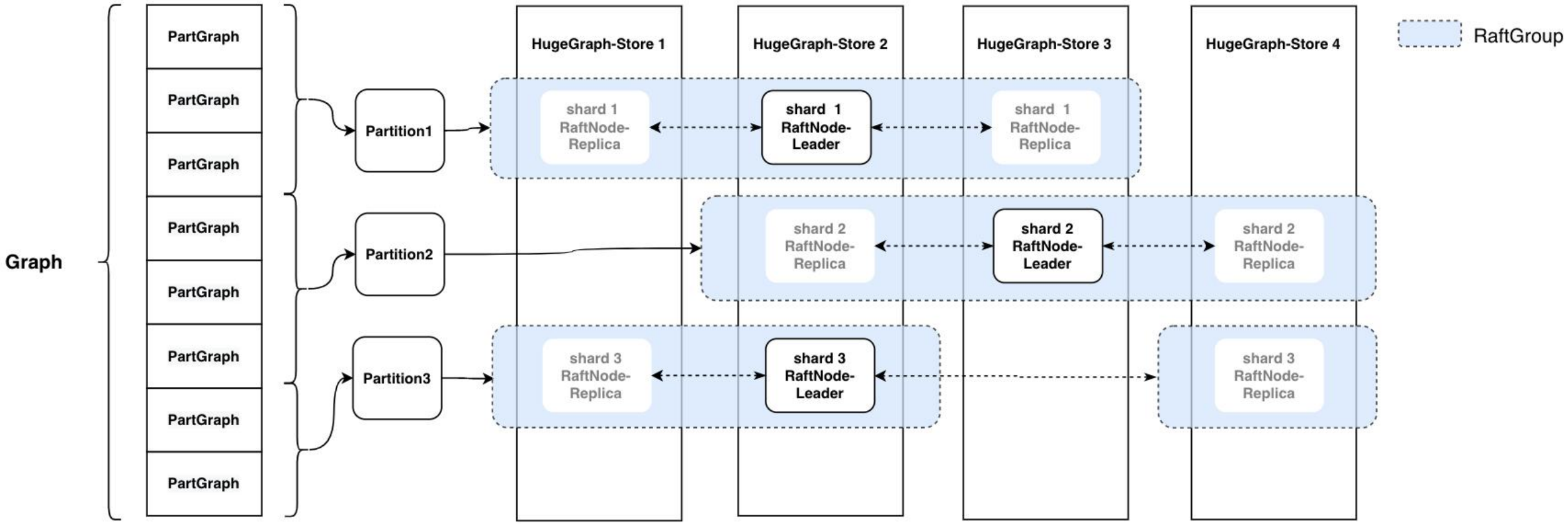
- 分区算法：基于点ID的哈希值进行范围分区

多副本策略

- 容灾能力
- raft 保证多副本一致性
- RaftGroup 横跨多个服务器节点，每个节点就是一个 shard



图结构与存储

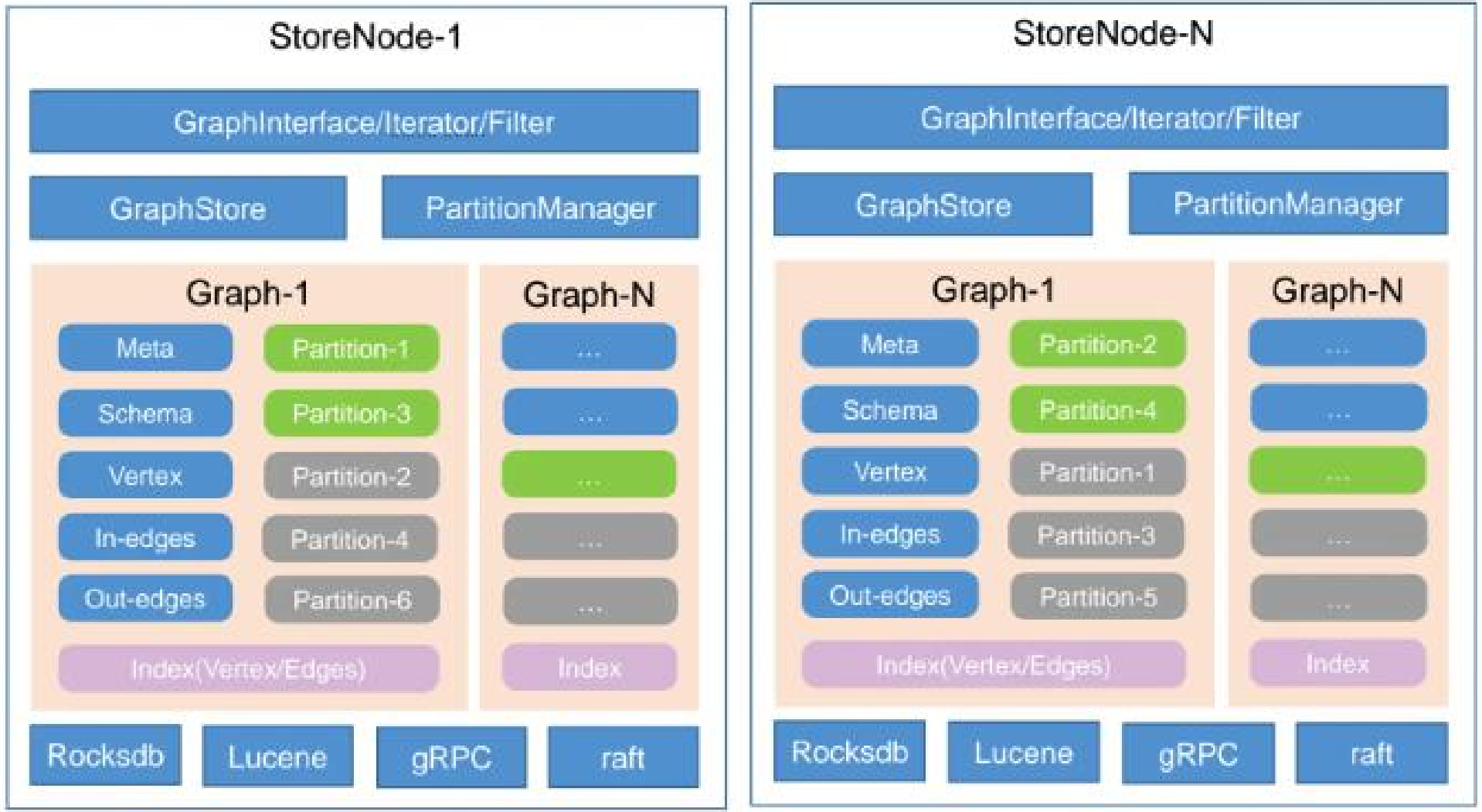


- 集群相关：角色切换、分区迁移到其他服务器
- 图相关：分区分裂、图数据迁移

- Raft 日志立即返回，快照延后导入 SST
- 批量写单leader、异步同步

- Raft 瓶颈：日志发送队列和 raft 状态信息
- 控制心跳粒度

Store 存储结构



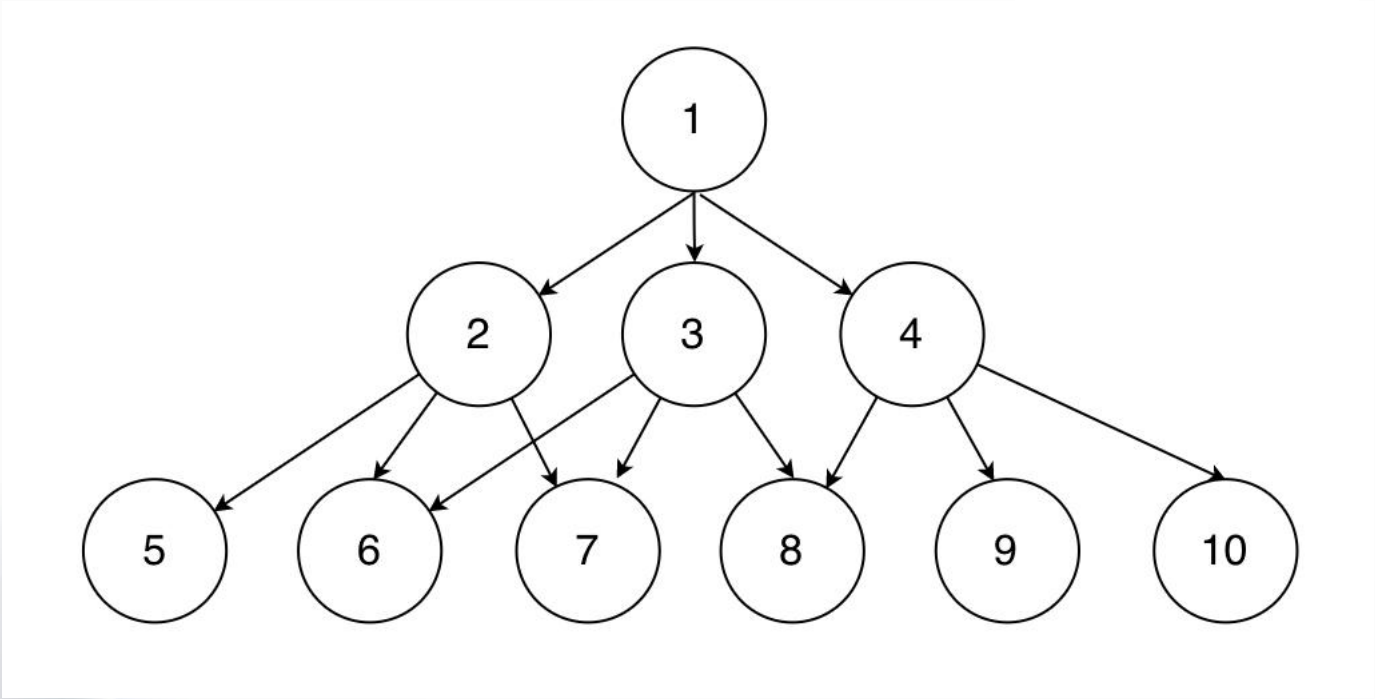
切边法

- 均匀，写入性能好
- 点和边集中存储，方便下推计算（基于边过滤点）

- 顺序遍历、范围查询或前缀查询，需要遍历所有分区
- 超级节点

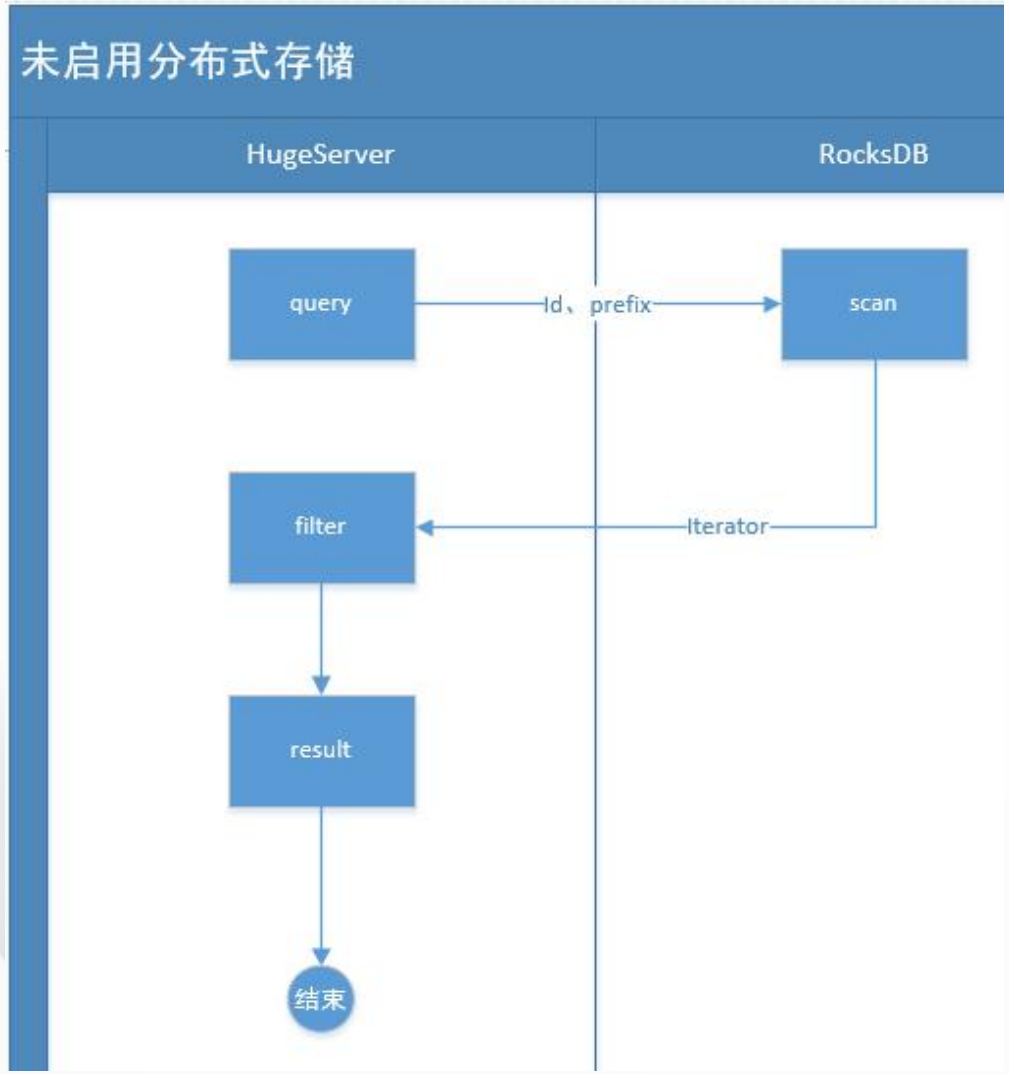
Gremlin 并行化

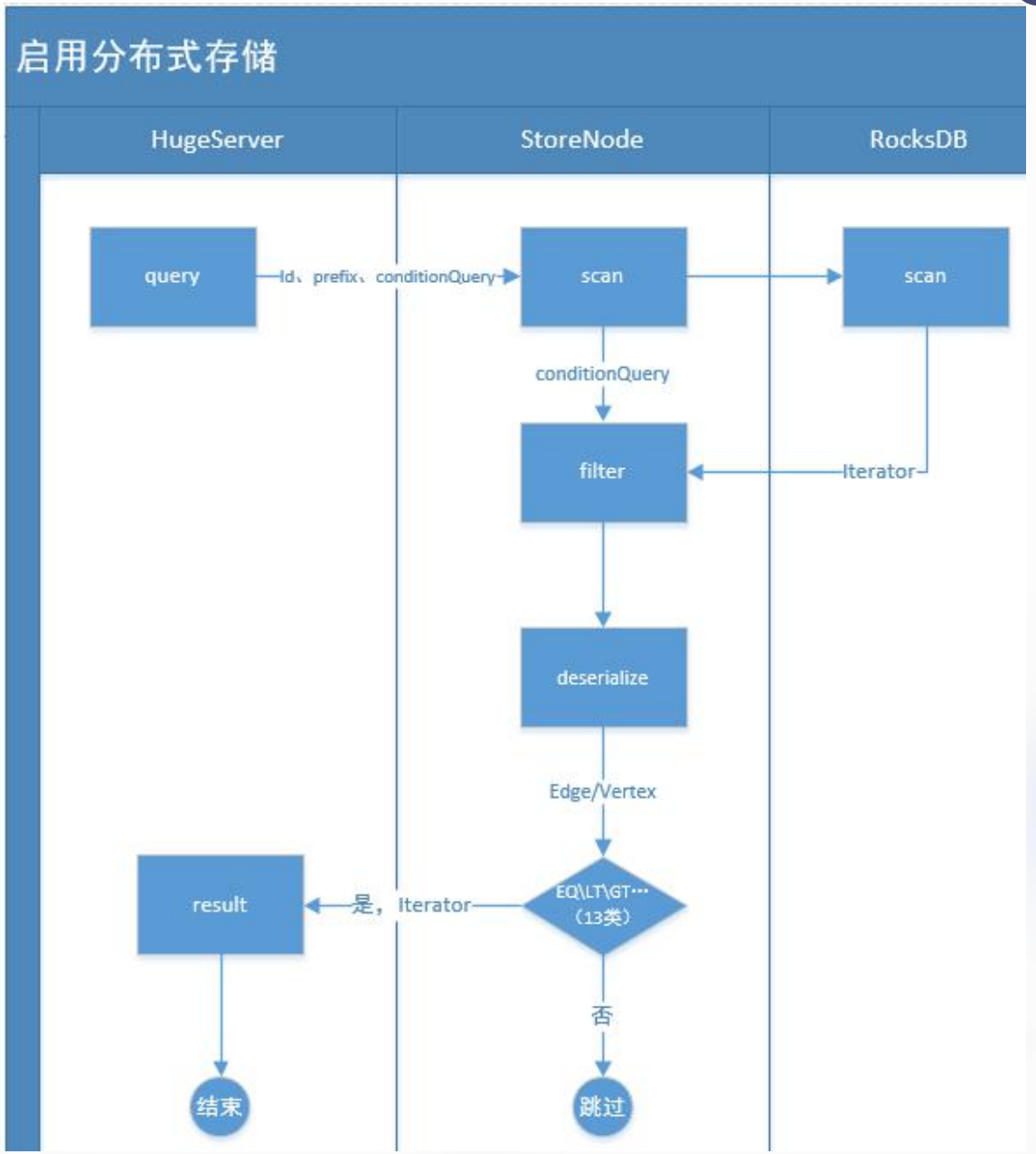
g.V(1).out().out().count()



1. flatmap并行化方案
   - 一次性获取所有starts，针对所有的starts并行化执行flatmap，获取所有starts的iterator，再将所有iterator合并。
   - 定义 flatMapBatch 方法，提取出所有的id之后，将所有id传到store端，在store端依靠rocksdb批量获取数据的能力，获取到所有邻节点的迭代器。
2. 双层迭代器并行化
   - 多个查询，每个查询有多个结果

下沉算子（sink operator）





Aggregate

ConditionQuery

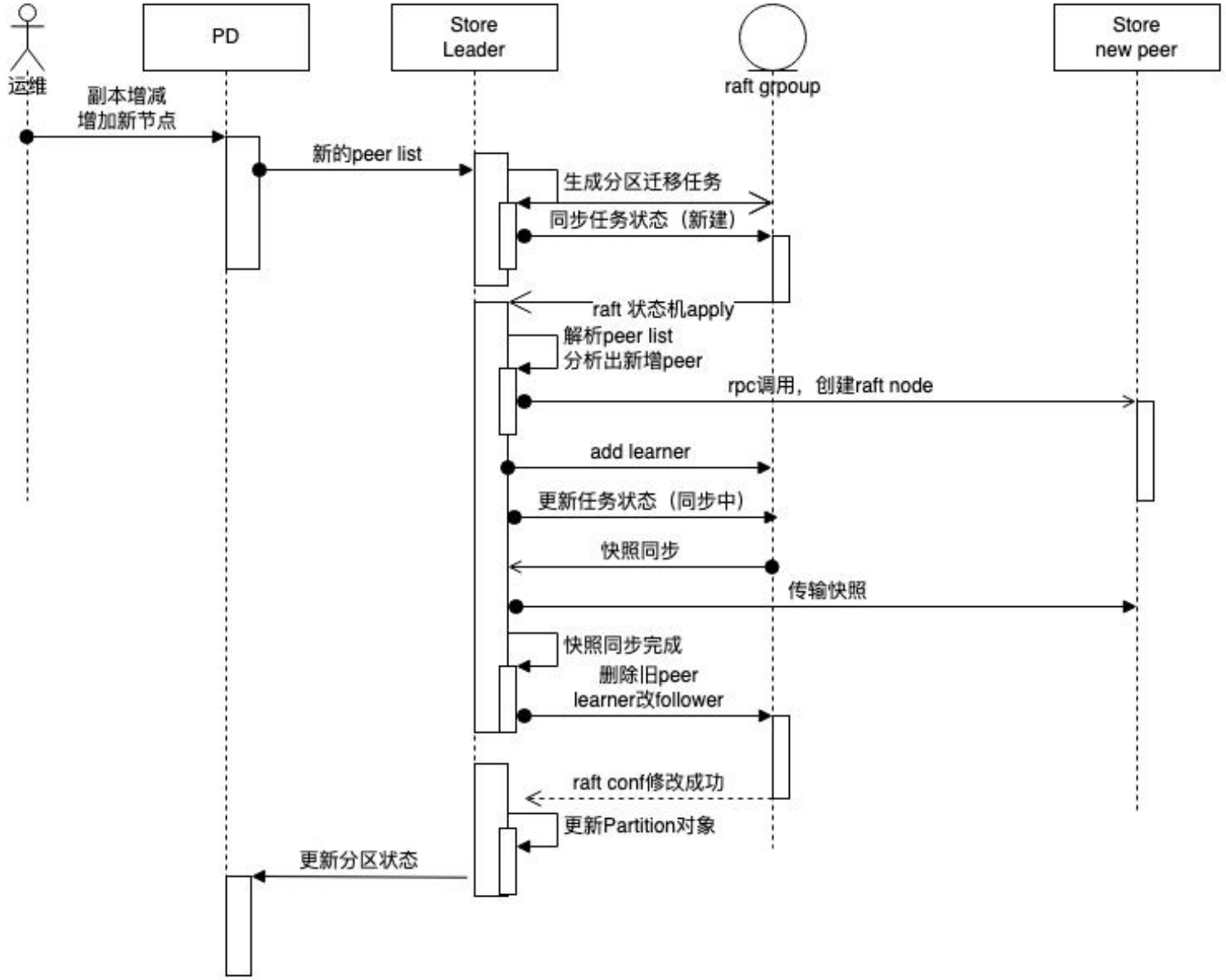
IdQuery

分布式集群管理

增加新的节点

- 增加新的 store 节点，做负载均衡，将一些 raftnode 迁移到新的节点上，那么就涉及 raftnode 的增加和删除

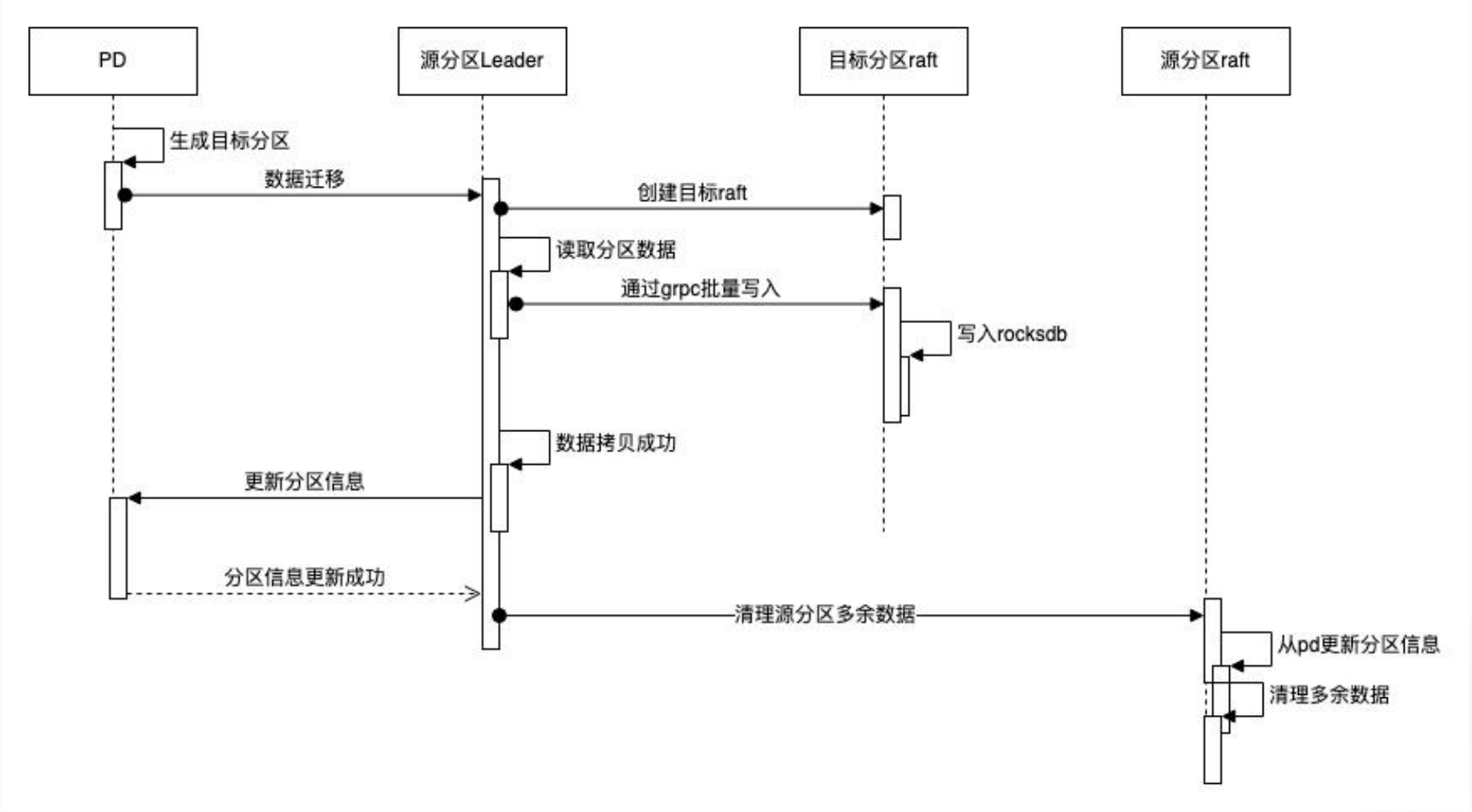
（raftGroup内部）



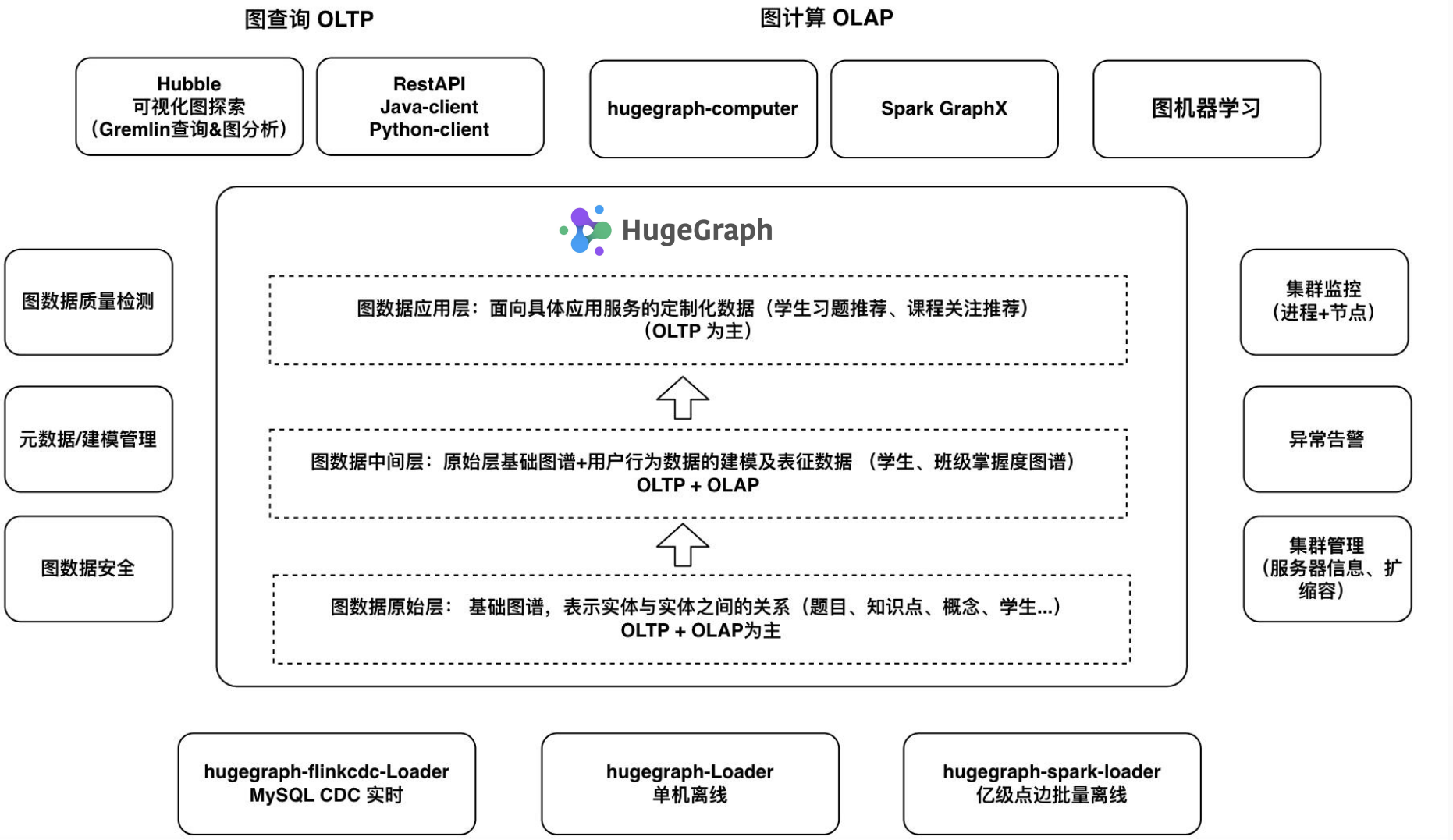
分布式集群管理

分区分裂与合并

- 单个分区的数据大到一定程度，需要分裂成多个分区，或者单个分区的数据小到一定程度，需要与其他的合并（跨raftGroup）



基于HugeGraph的图平台建设



HugeGraph 在 CVTE 视源股份 的应用

搜索推荐

个性化推题、推荐关注、薄弱知识点智能复习、教师资源推荐....

风险控制

备损预测，财务风控.....

数据血缘

基于 Hive、SparkSQL、FlinkSQL 等大数据组件的全局数仓数据血缘实现

知识图谱

教育知识概念图谱，产业链图谱，人才图谱.....

知识库

教育知识库，人才知识库（知识库包括规则，过程性知识（知识抽取）等）

图机器学习

作为图机器学习（图表示学习/图神经网络）的数据输入端

Apache HugeGraph 目前的痛点

- 单次查询数据量大时内存占用多, 线程无法提前终止, 没有细粒度的内存管控
- Hbase 方案体系重, 依赖多, 难上云, 虽能支持千亿数据但是难以同时保证高性能
- RocksDB 方案维护难, 观测难, 调优难, 对普通用户使用门槛高
- 新用户仍然需要学习较为复杂的图查询语言, 并有较高的迁移门槛
- 缺少一体化的业务解决方案, 单纯图数据/图计算系统离用户距离远

Apache HugeGraph 后续规划

- 基于分布式版本持续做更多的优化，保证集群稳定性
- 持续优化图查询, 实现内存管控的完整体系
- 更多的图分析算法支持, 尤其是提供图产品化的解决方案, 大幅降低使用门槛
- 更完整的生态建设，包括与大数据生态的集成、与图计算、图可视化框架的集成等等
- 查询辅助功能，通过类 GPT 模型将自然语言转化为图查询语言
- 探索优化/增强 RocksDB, 改善 RocksDB 难维护, 易用性弱, 配置复杂等问题

欢迎扫码关注

GOTC

THANKS

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE