链家大数据多维分析引擎实践

邓钫元
链家大数据部

关于我

链家网大数据集群架构组负责人

15年至今负责链家大数据集群建设，打造公司级计算存储平台。

专注于hadoop生态组件的定制开发及应用，

深入hadoop,hbase等源码，成为contributor回馈社区，

擅长底层性能调优。

目录

OLAP vs OLTP

Online Analytical Processing / Online Transactional Processing

	OLTP（事务）	OLAP （分析）
面向应用	日常交易处理	明细查询，分析决策
访问模式	简单小事务，操作少量数据	复杂聚合查询，可以过大量数据
数据	当前最新数据	历史数据
数据规模	GB	TB ~ PB
数据更新	实时更新	批量更新
数据组织	满足3NF	反范式，星型模型 ,雪花模型

example

给定时间范围，按 “returnflag” 和 “orderstatus” 报告营收。

```sql
select
    l_returnflag, o_orderstatus,
    sum(l_quantity) as sum_qty,
    sum(l_extendedprice) as sum_base_price
    ...
from
    v_lineitem
    inner join v_orders on l_orderkey = o_orderkey
where
    l_shipdate <= '1998-09-16'
group by
    l_returnflag, o_orderstatus
order by
    l_returnflag, o_orderstatus;
```

OLAP引擎分类

ROLAP（Relational OLAP）

基于关系模型，实时进行聚合计算

实现：传统数据库引擎/spark sql/presto

MOLAP（Multi-dimension OLAP）

基于预定义模型，预先进行聚合计算，存储汇总结果

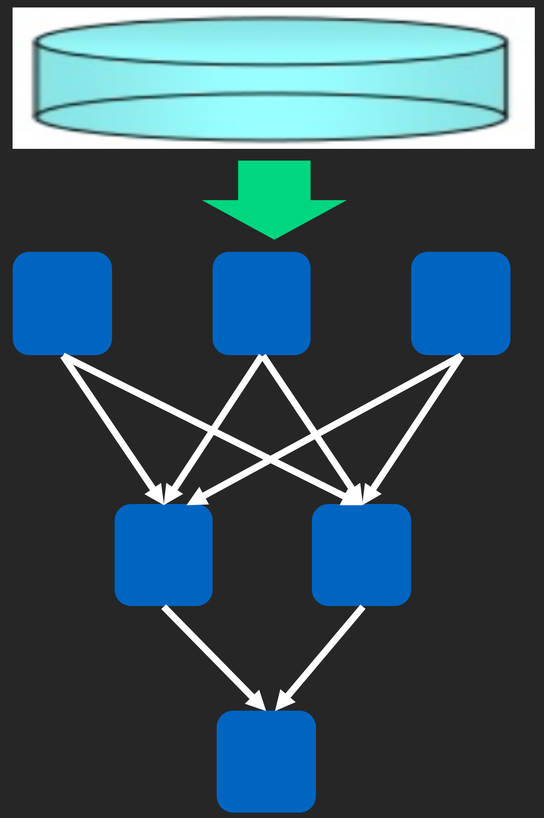
实现：Kylin/Druid

HOLAP（Hybrid OLAP）

混合多引擎，不同场景路由到不同引擎

ROLAP

warehouse



扫描

聚合

汇总

优势

支持任意的sql查询

无数据冗余，一致性好

缺点

大数据量及复杂查询返回慢

并发较差

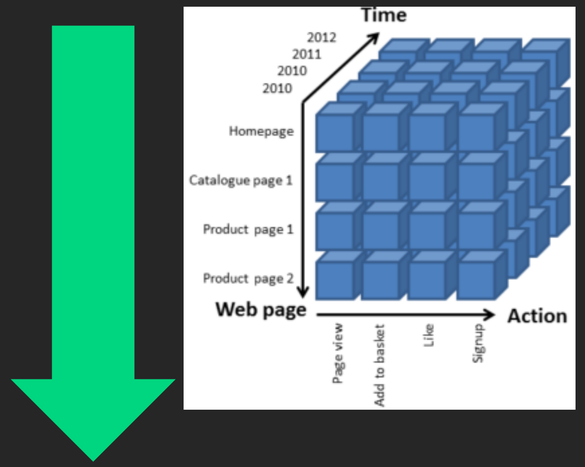
场景

灵活性很高的分析

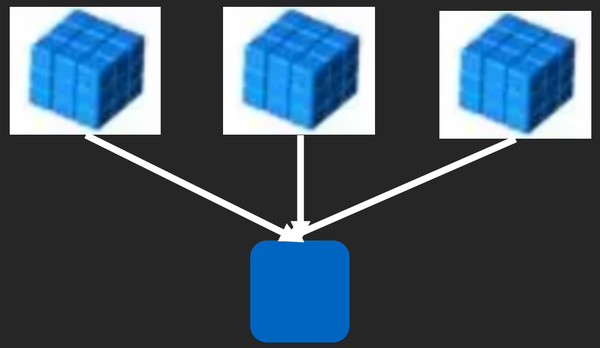
MOLAP

warehouse

定义模型（cube）



预聚合存储



少量计算汇总

优势

支持超大原始数据集

快速返回，并发高

缺点

不支持明细

需要预先定义维度和指标

场景

能预知查询模式，并发有要求的场景

目录

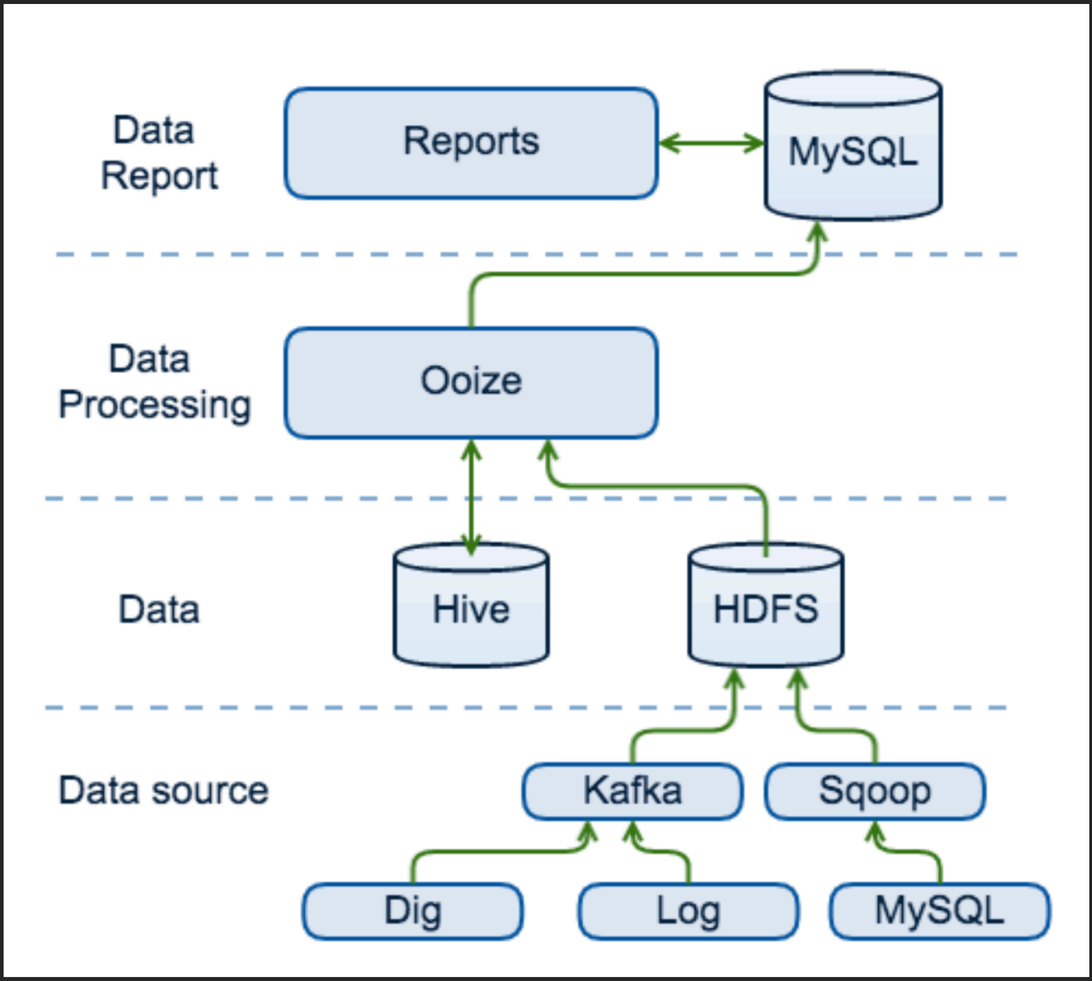
OLAP背景简介

链家多维分析演进和展望

OLAP平台全链路优化实践

Q&A

早期数据分析实现



数据量迅速膨胀，mysql无法存储

查询速度变慢

查询维度多，需求定制开发时间较长

技术选型

KYLIN简介

Apache Kylin™是一个开源的分布式分析引擎，提供Hadoop之上的SQL查询接口及多维分析（OLAP）能力以支持超大规模数据，最初由eBay Inc. 开发并贡献至开源社区。它能在亚秒内查询巨大的Hive表。

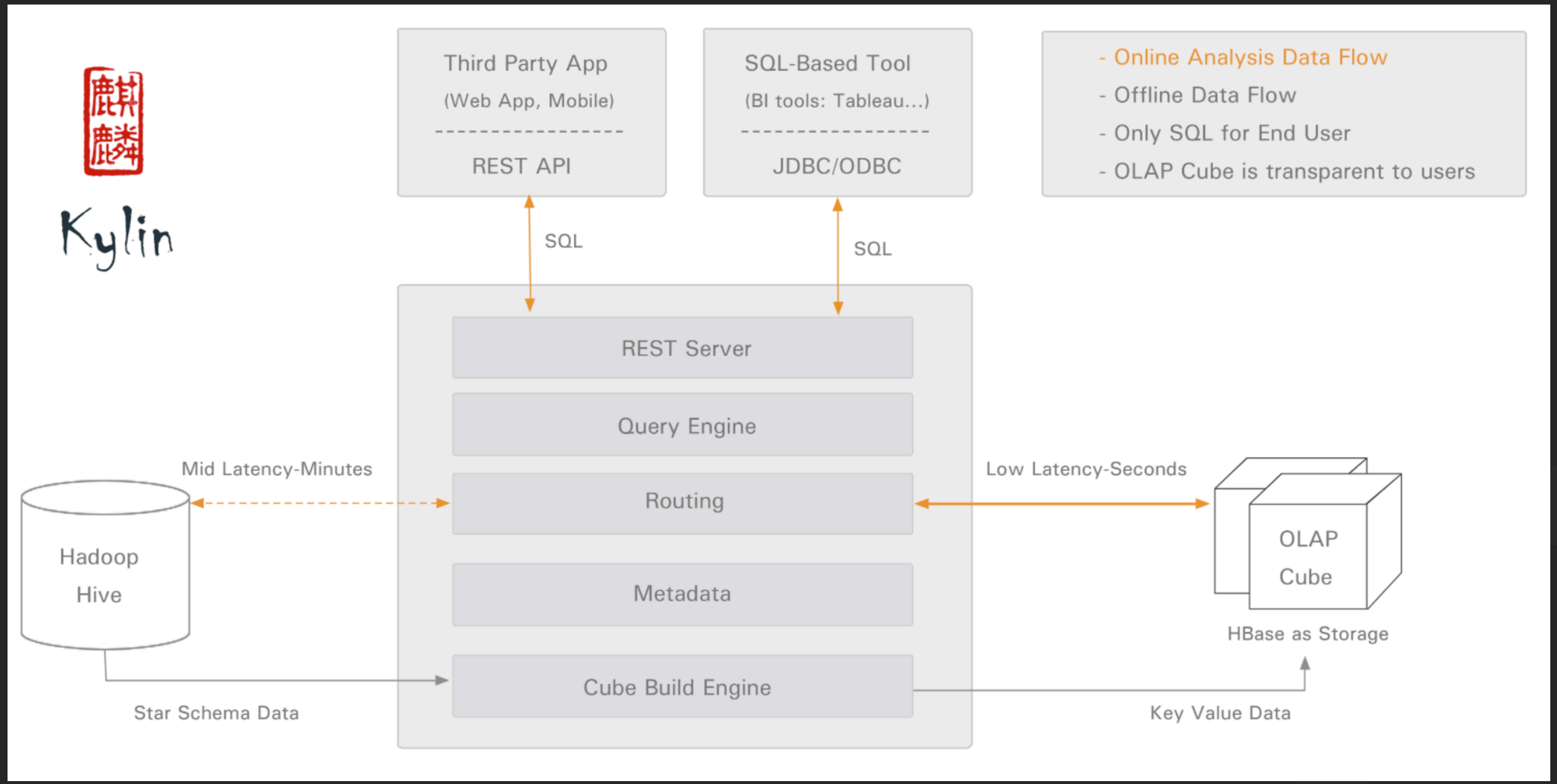
MOLAP解决方案

预先定义维度和指标

预计算cube，存储到hbase中

查询时解析sql路由到hbase中获取结果

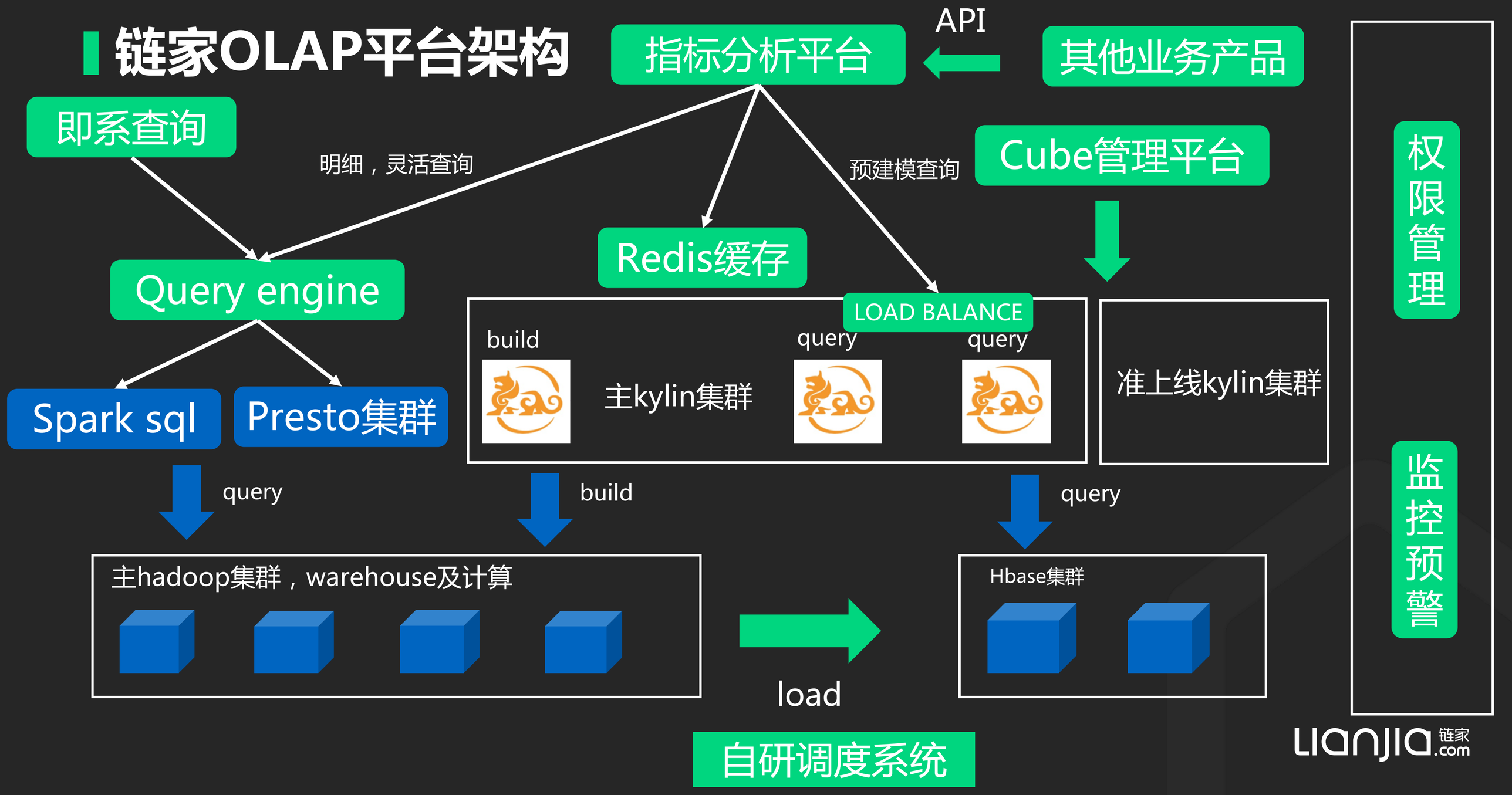
KYLIN架构



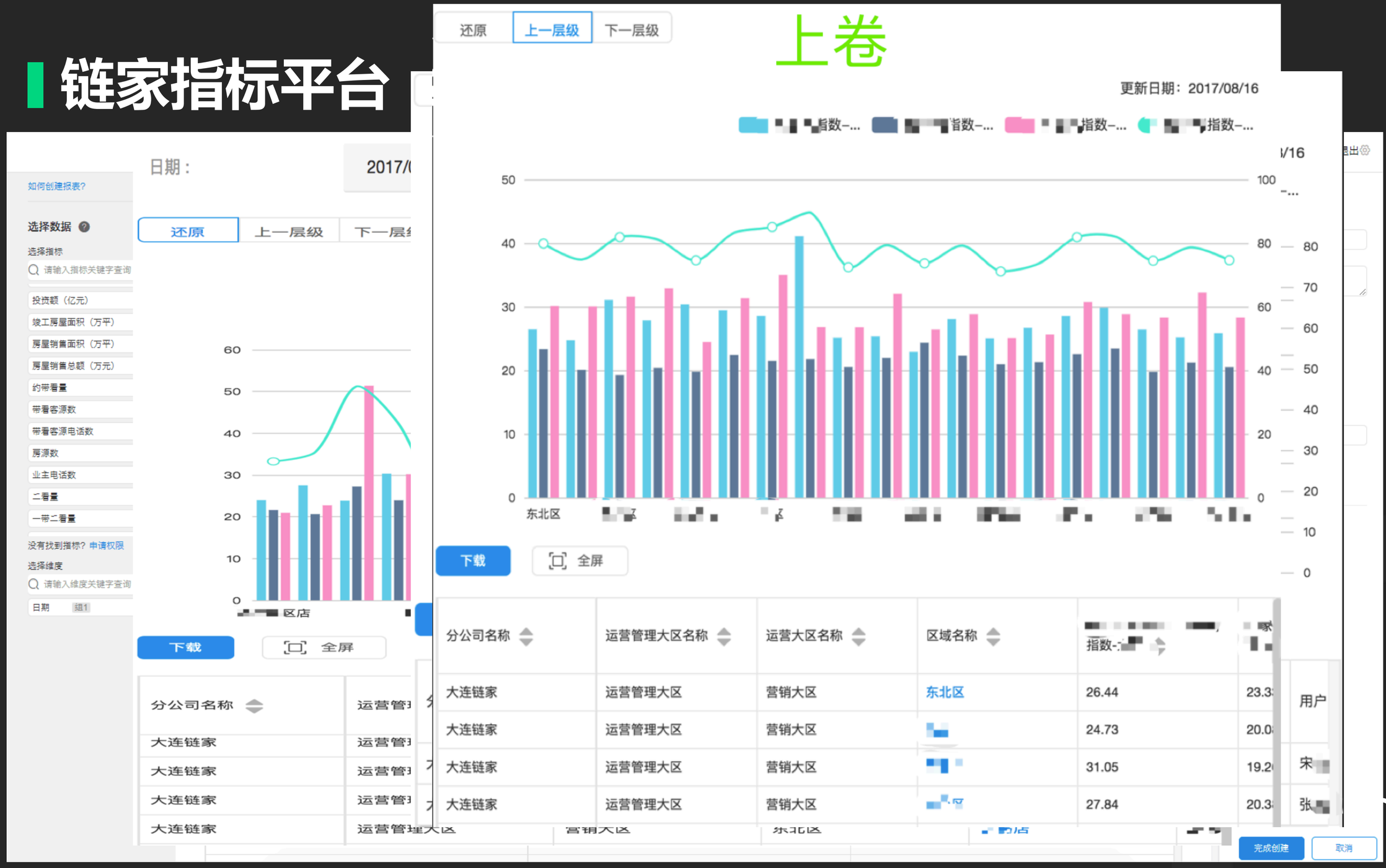
链家kylin使用统计

定位：离线OLAP引擎
100+ cube，覆盖公司8个业务线
Cube存储总量30T+, 数据行数800+亿行,单cube最大40+亿行
日查询量10万+，时延<500ms(95%), <1s(99%)

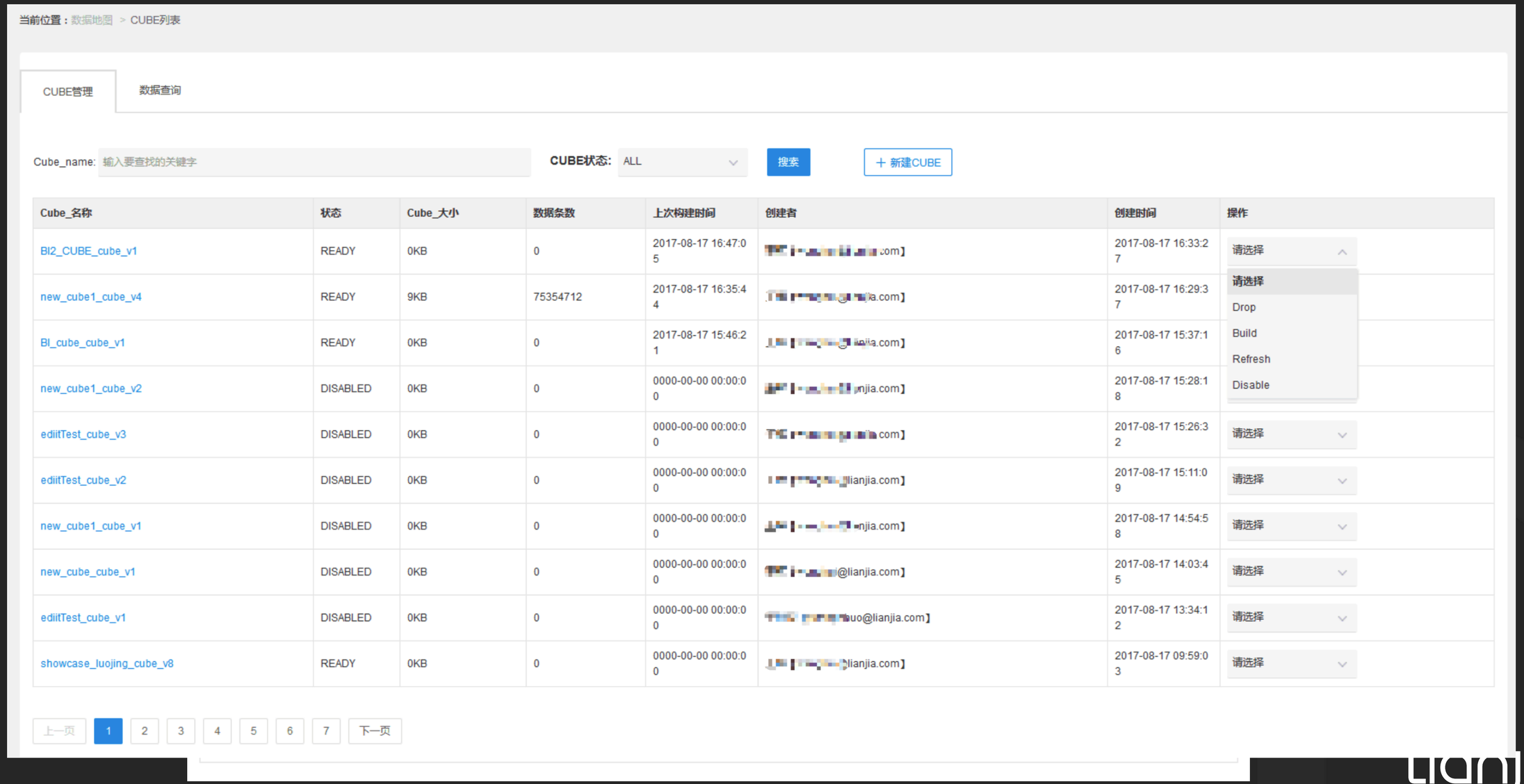
链家OLAP平台架构


指标分析平台
API
其他业务产品
即系查询
明细，灵活查询
预建模查询
Cube管理平台
Redis缓存
Query engine
LOAD BALANCE
build
query
query
主kylin集群
准上线kylin集群
Spark sql
Presto集群
query
build
query
主hadoop集群，warehouse及计算
Hbase集群
load
自研调度系统
权限管理
监控预警

链家指标平台


上卷
还原
上一层级
下一层级
更新日期：2017/08/16
50
40
30
20
10
0
100
80
60
40
20
0
东北区
下载
全屏
分公司名称
运营管理大区名称
运营大区名称
区域名称
大连链家
运营管理大区
营销大区
东北区
26.44
24.73
31.05
27.84
用户
日期：
如何创建报表?
选择数据
选择指标
投资额（亿元）
竣工房屋面积（万平）
房屋销售面积（万平）
房屋销售总额（万元）
约带看量
带看客源数
带看客源电话数
房源数
业主电话数
二看量
一带二看量
没有找到指标? 申请权限
选择维度
日期
组1
60
50
40
30
20
10
0
区店
完成创建
取消

自研CUBE管理


当前位置：数据地图 > CUBE列表
CUBE管理
数据查询
Cube_name: 输入要查找的关键字
CUBE状态: ALL
搜索
+ 新建CUBE
Cube_名称
状态
Cube_大小
数据条数
上次构建时间
创建者
创建时间
操作
BI2_CUBE_cube_v1 READY 0KB 0 2017-08-17 16:47:05 2017-08-17 16:33:27 请选择
new_cube1_cube_v4 READY 9KB 75354712 2017-08-17 16:35:44 2017-08-17 16:29:37
BI_cube_cube_v1 READY 0KB 0 2017-08-17 15:46:21 2017-08-17 15:37:16
new_cube1_cube_v2 DISABLED 0KB 0 0000-00-00 00:00:00 2017-08-17 15:28:18
ediitTest_cube_v3 DISABLED 0KB 0 0000-00-00 00:00:00 2017-08-17 15:26:32 请选择
ediitTest_cube_v2 DISABLED 0KB 0 0000-00-00 00:00:00 2017-08-17 15:11:09 请选择
new_cube1_cube_v1 DISABLED 0KB 0 0000-00-00 00:00:00 2017-08-17 14:54:58 请选择
new_cube_cube_v1 DISABLED 0KB 0 0000-00-00 00:00:00 2017-08-17 14:03:45 请选择
ediitTest_cube_v1 DISABLED 0KB 0 0000-00-00 00:00:00 2017-08-17 13:34:12 请选择
showcase_luojing_cube_v8 READY 0KB 0 0000-00-00 00:00:00 2017-08-17 09:59:03 请选择
请选择
Drop
Build
Refresh
Disable
上一页 1 2 3 4 5 6 7 下一页

链家OLAP特色

自研可视化平台

支持上卷下钻，维度对比，可视化报表创建，指标管理

相比开源saiku, UI美观，贴合业务，灵活定制

引擎能力

超越MOLAP, HOLAP

支持跨cube查询

监控，高可用

自研CUBE管理

简化配置，提升管理效率

优化前端页面

对接链家权限体系

OLAP新展望

扩展能力

多源数据查询

优化路由选择

元数据同步增加pull模式

KYLIN 2.0

雪花模型

spark构建引擎

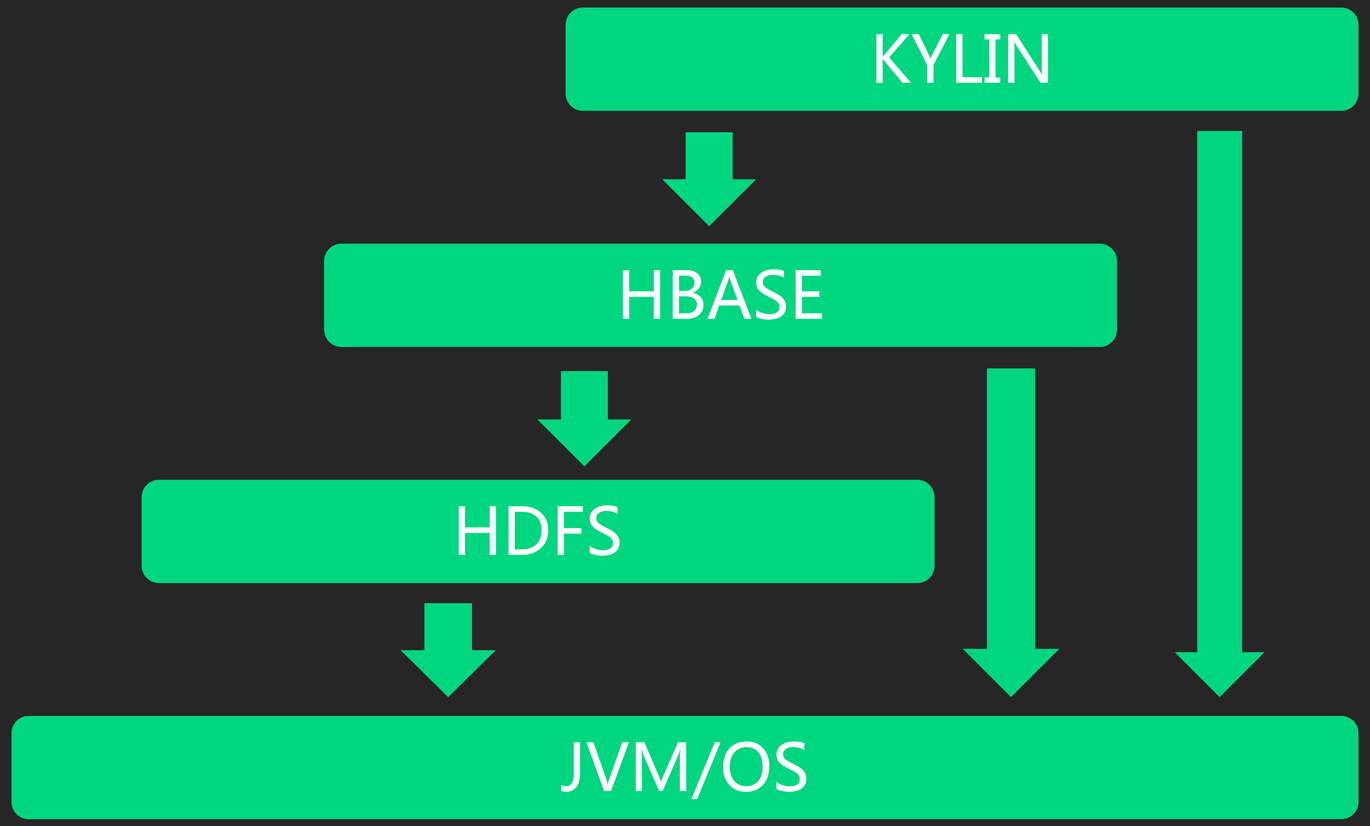
实时

kylin streaming cubing, 小批量近实时

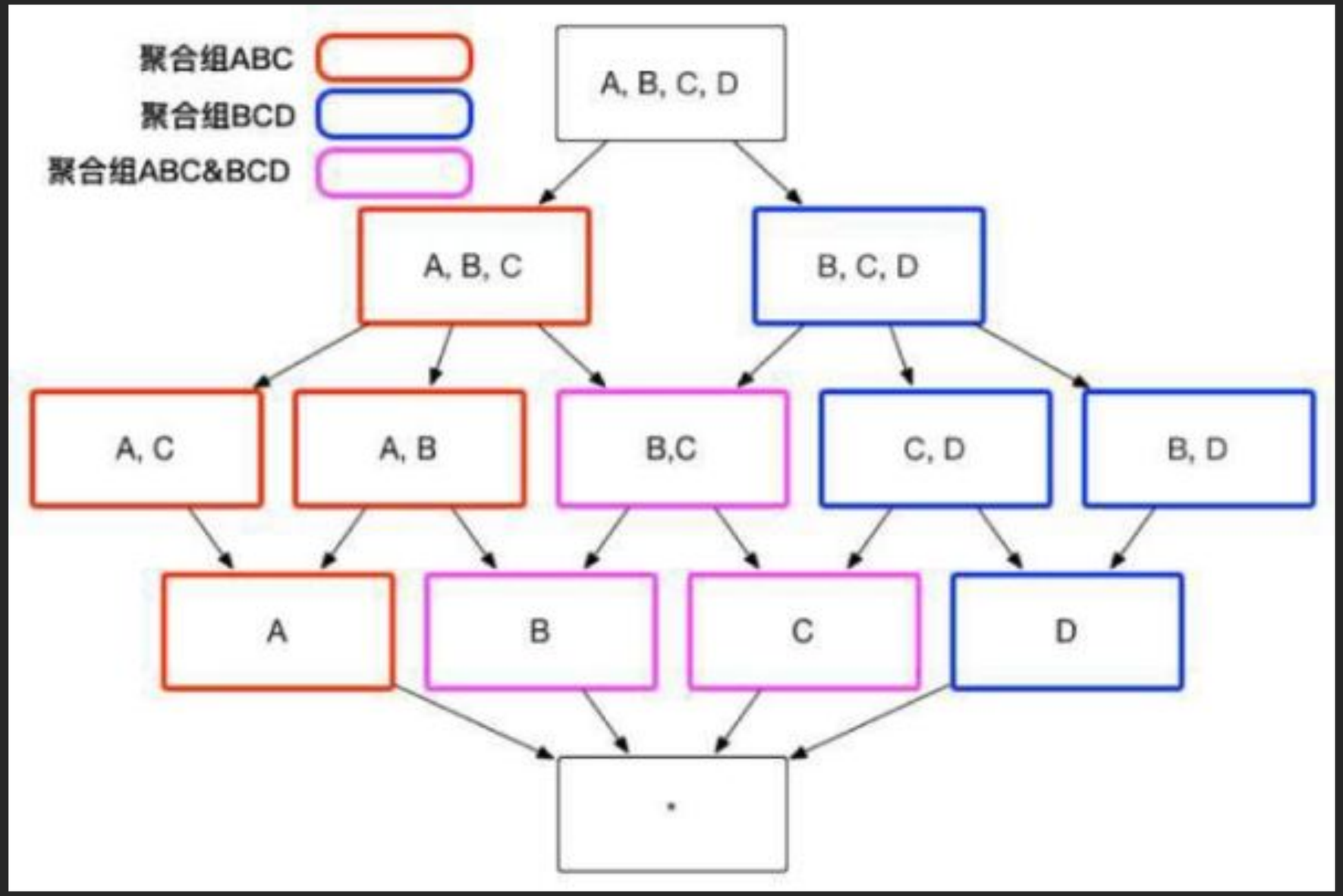
druid/palo, lamada架构，纯实时

目录

KYLIN全链路架构



KYLIN降维方式



聚合组

衍生维度

强制维度

层次维度

联合维度

KYLIN改造及优化

KYLIN的HBASE表支持配置前缀，隔离多套环境

修复更新缓存的死锁bug(KYLIN-2780)

开启mapjoin,snappy压缩提升一倍构建速度

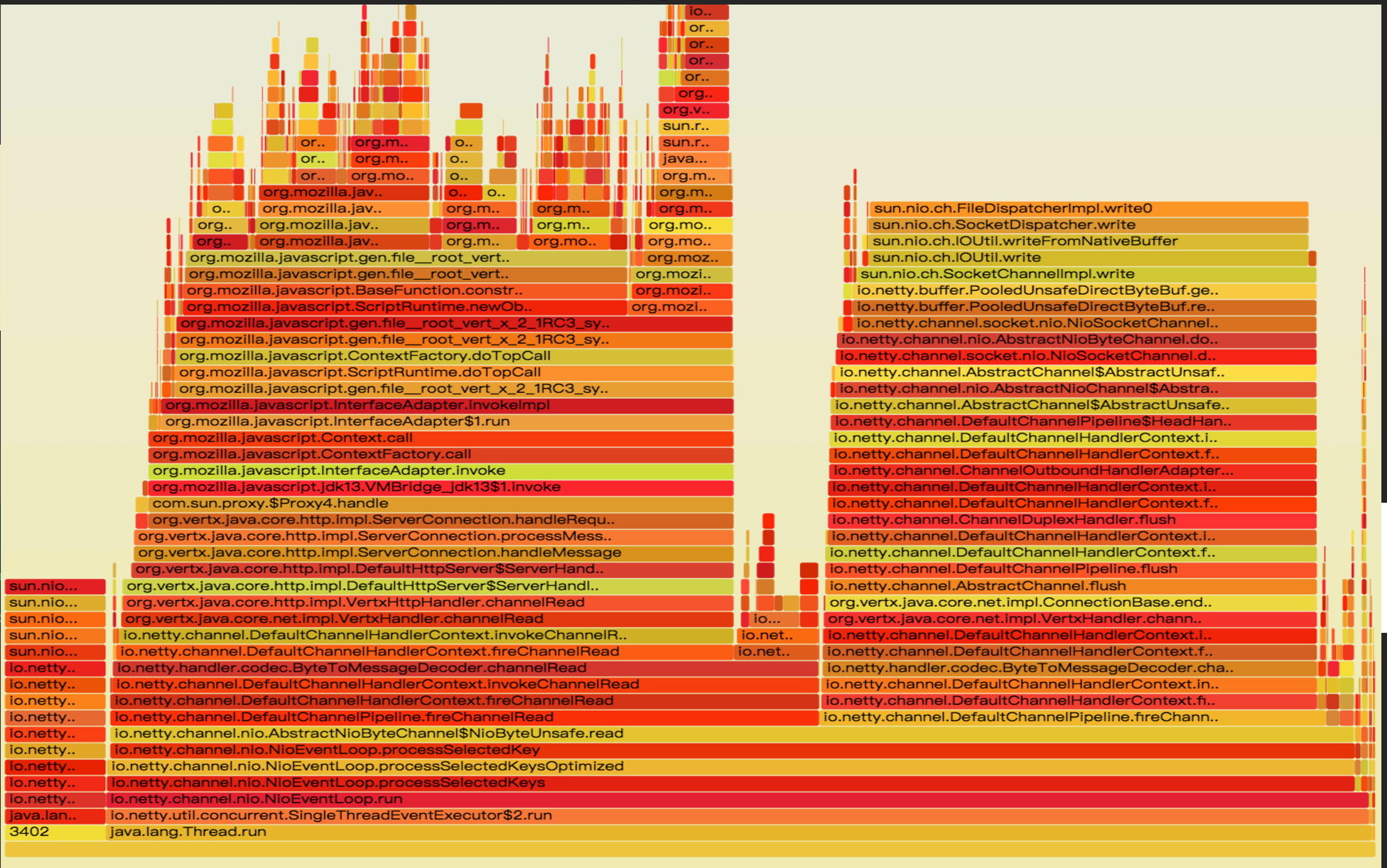
GC调优，缩短暂停时间

全局字典调大MAP内存

定期merge segment/清理无效数据，减少hbase region数

Bcypt算法性能问题

性能调优神器-火焰图



Special thx:

Brendan Gregg

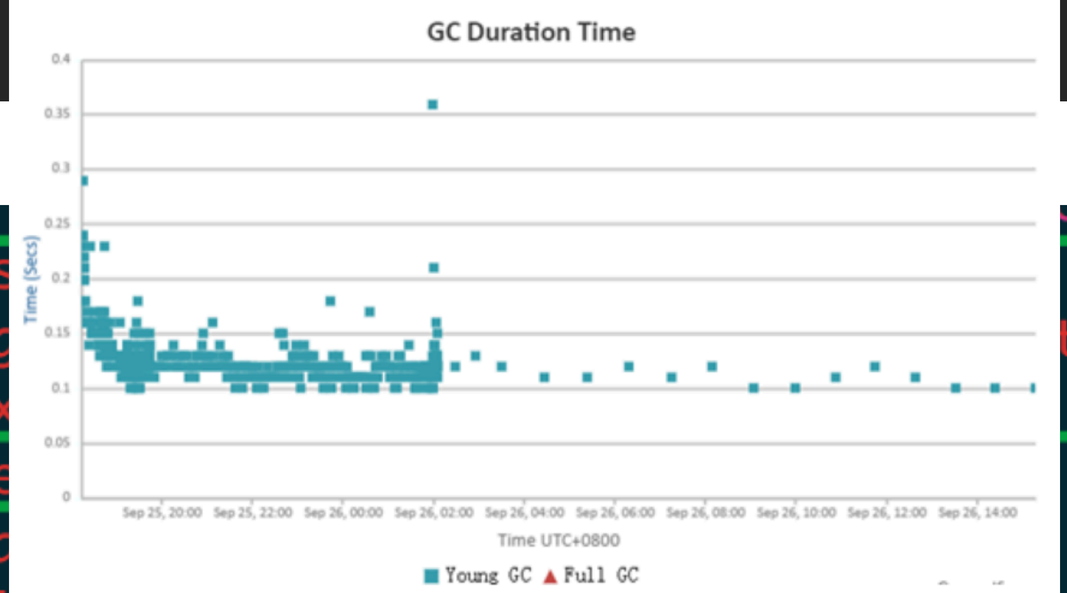
章亦春

java GC

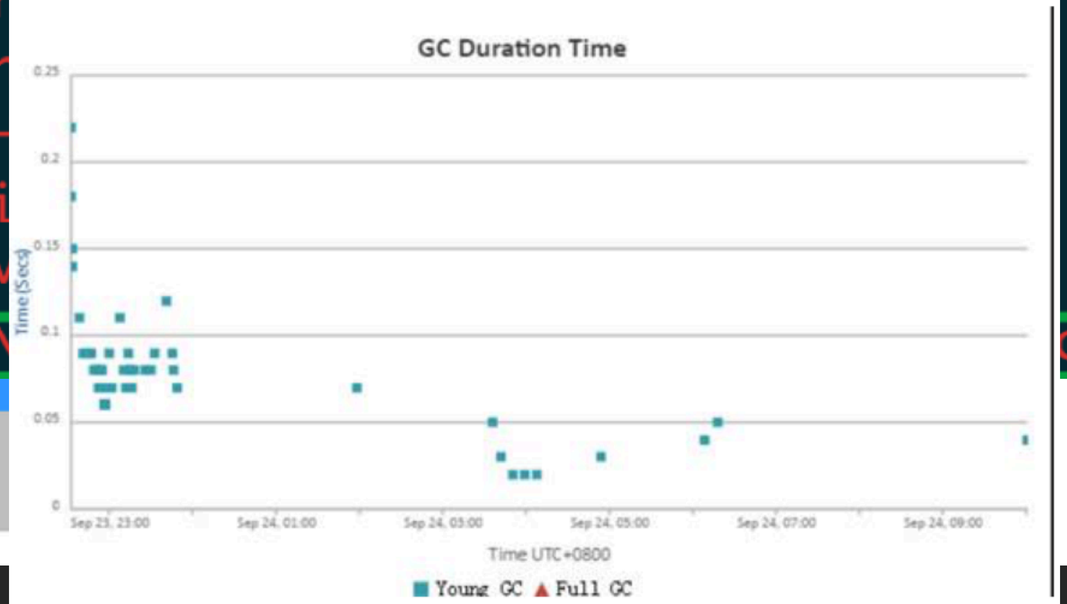
CMS

G1

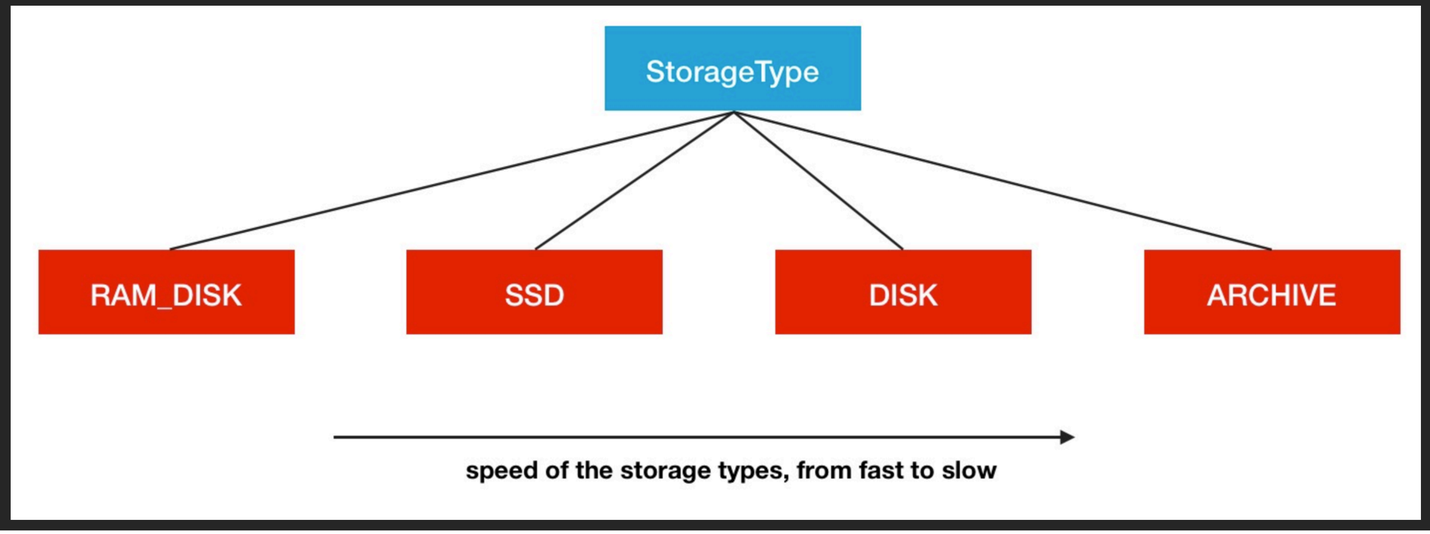
CMS GC停顿时间图



G1 GC停顿时间图



HBASE ON HDFS混合存储



Policy Name	Block Placement (n replicas)
Lazy_Persist	RAM_DISK: 1, DISK: *n*-1
All_SSD	SSD: *n*
One_SSD	SSD: 1, DISK: *n*-1
Hot (default)	DISK: *n*
Warm	DISK: 1, ARCHIVE: *n*-1
Cold	ARCHIVE: *n*

ALL_SSD

WAL/核心业务

ONE_SSD

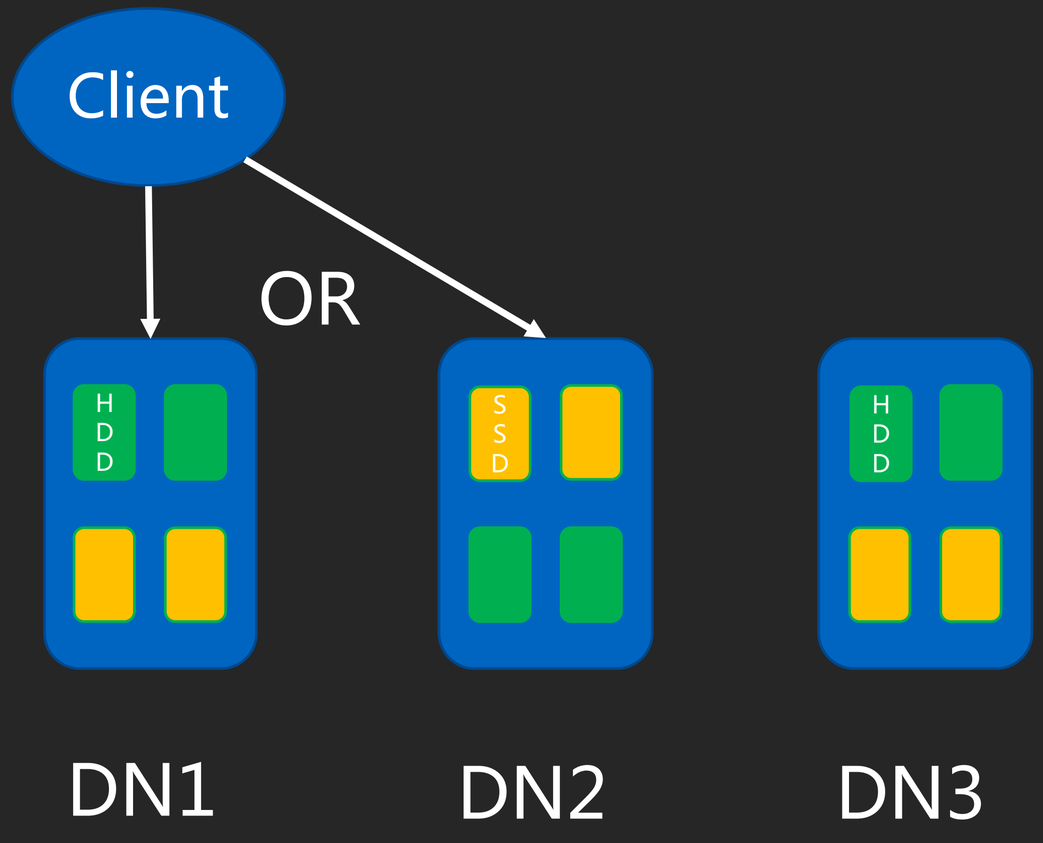
重要业务

HOT

普通业务

SSD-FIRST策略

/pathtofile/file.dat block1 (dn1:hdd,dn2:ssd,dn3:hdd)

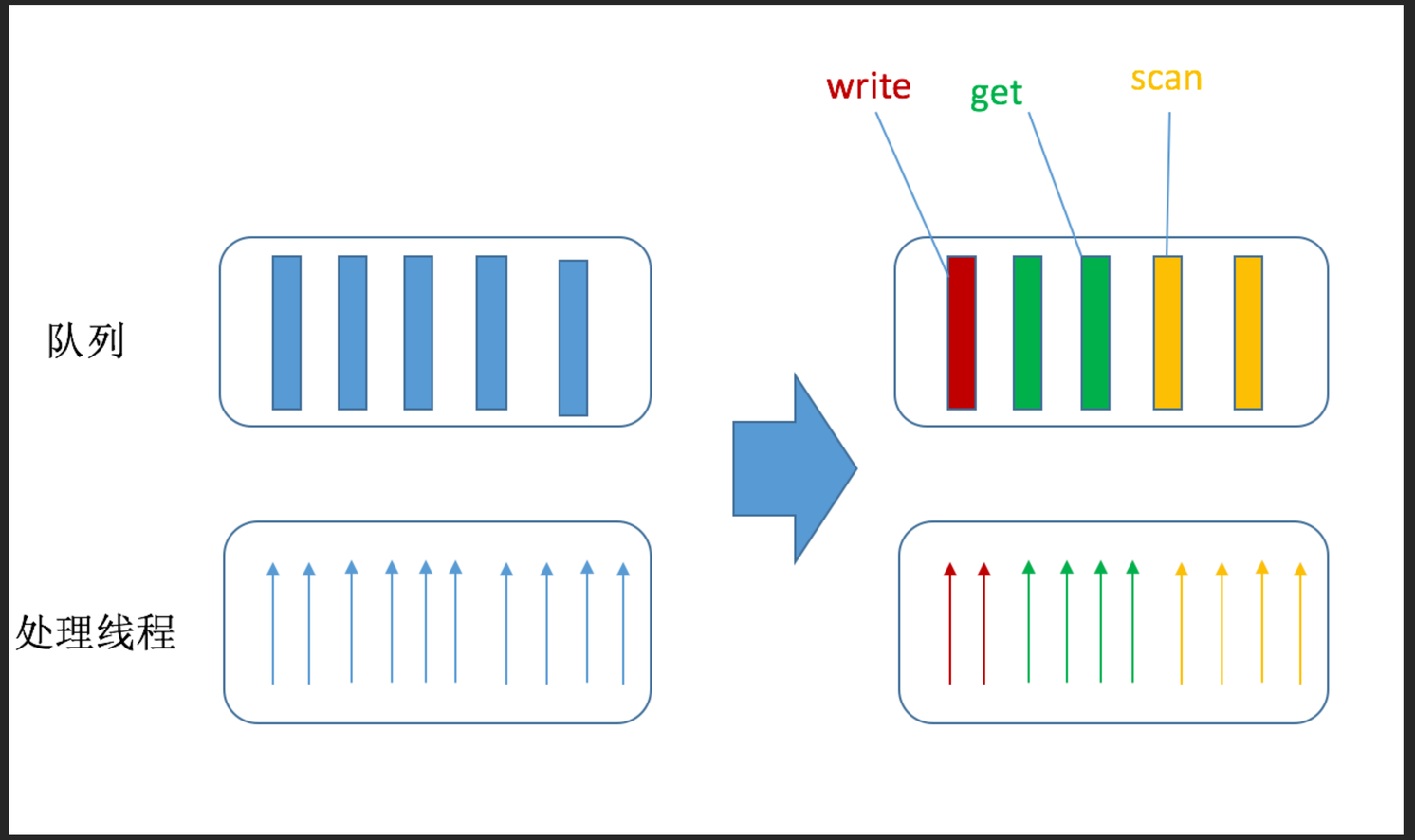


Default:优先读取本地

SSD-FIRST策略：优先读取远程SSD

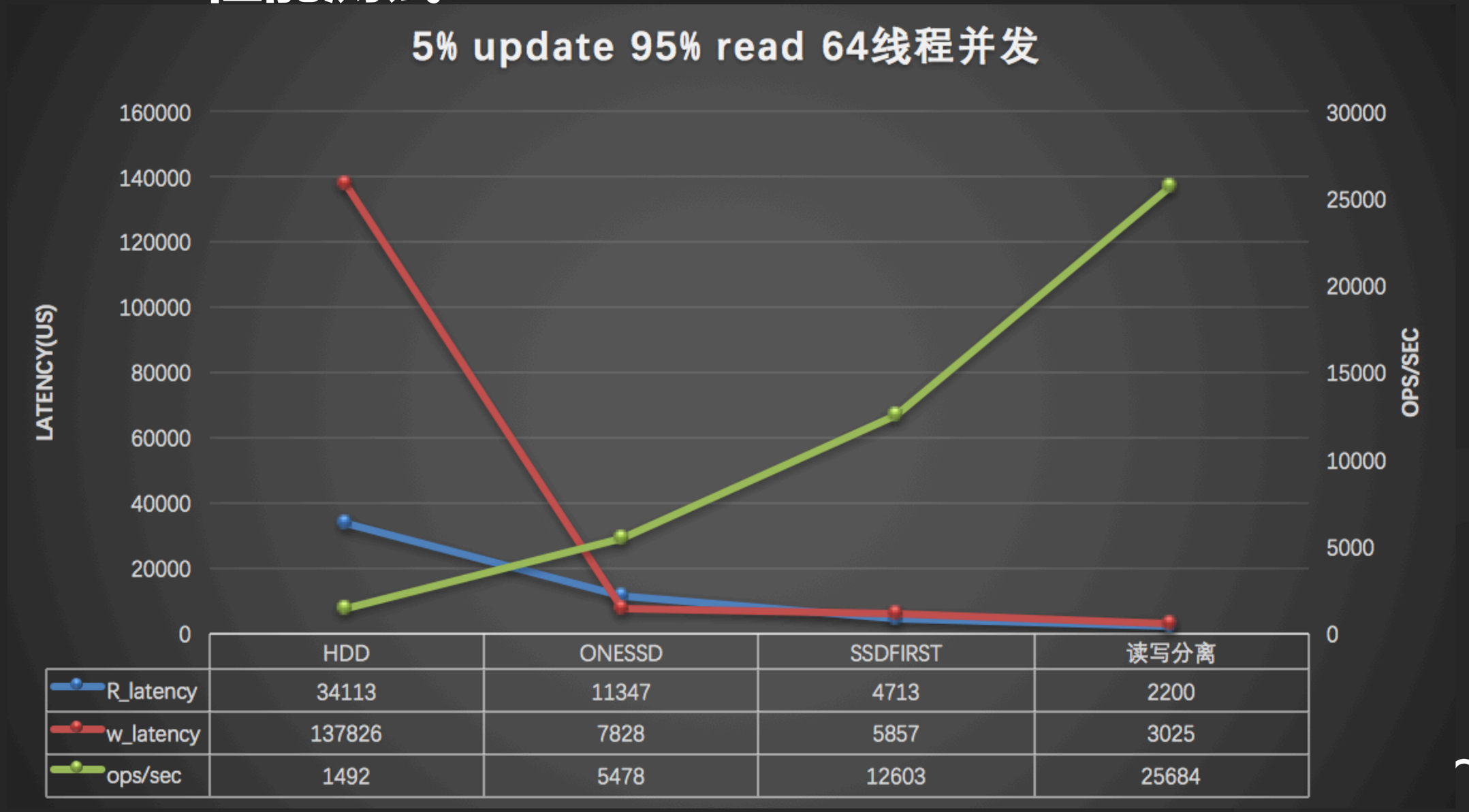
提交社区HDFS-11942

HBASE读写分离



避免scan阻塞小请求

Hbase性能测试



	HDD	ONESSD	SSDFIRST	读写分离
R_latency	34113	11347	4713	2200
w_latency	137826	7828	5857	3025
ops/sec	1492	5478	12603	25684

HBASE&HDFS优化杂项

Hdfs短路读（short-circuit）

并行多线程读取datanode（Hedged Reads）

检测stale DataNodes, 避免读取慢速datanode

关闭hadoop均衡器, 避免影响hbase本地性

Multiwal，提升hbase写入性能

神秘的cache & 内核分析工具systemtap

```
global mark_page_accessed,mark_buffer_dirty,add_to_page_ca
probe   kernel.function("mark_page_accessed")
{
```

cached
37

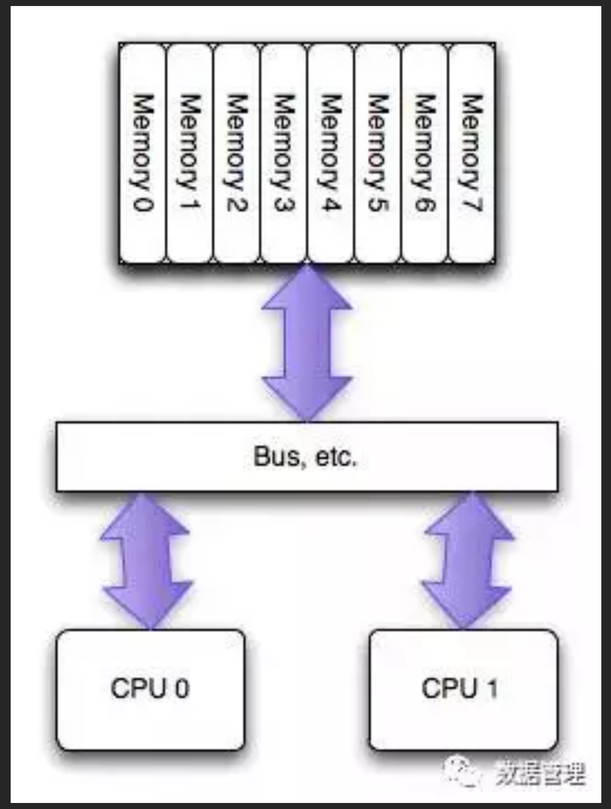
	host	date	hit
除	jx-bd-hadoop24.zeus.lianjia.com	2016-10-01	79.23045
除	jx-bd-hadoop24.zeus.lianjia.com	2016-10-02	89.20050
除	jx-bd-hadoop24.zeus.lianjia.com	2016-10-03	89.16474
除	jx-bd-hadoop24.zeus.lianjia.com	2016-10-04	87.54686
除	jx-bd-hadoop24.zeus.lianjia.com	2016-10-05	88.97900

```
probe timer.ms(10000) {
    total = mark_page_accessed - mark_buffer_dirty
    misses = add_to_page_cache_lru - account_page_dirtied
    hit = 1 - misses*1.0/total
}
```

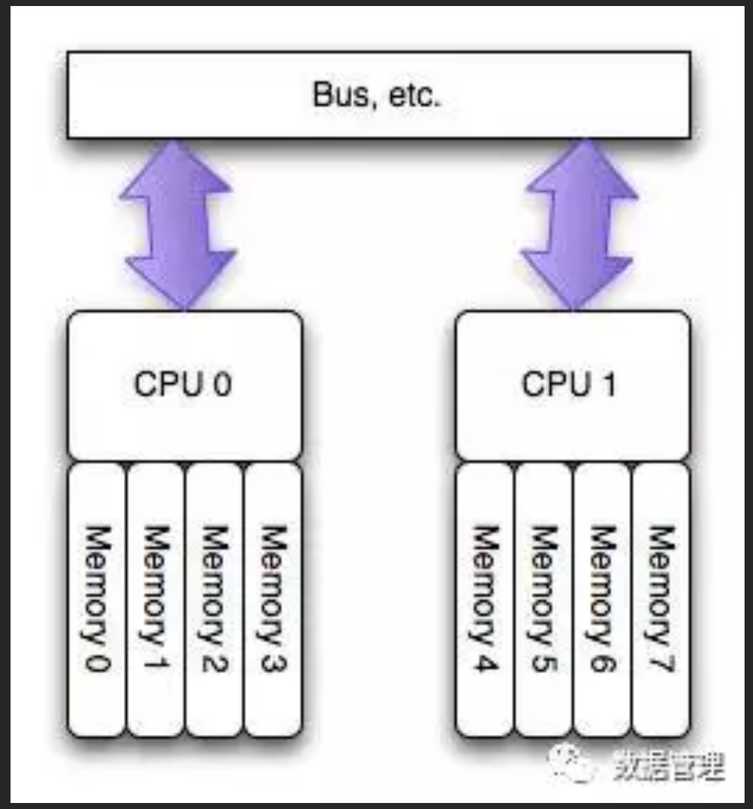
Cache命中率=

$$100\% - \frac{\text{添加page缓存次数}}{\text{page访问次数}}$$

多处理器之坑-NUMA



SMP(对称多处理器)



NUMA：非一致性内存访问

默认亲和模式

优先分配/淘汰专属内存

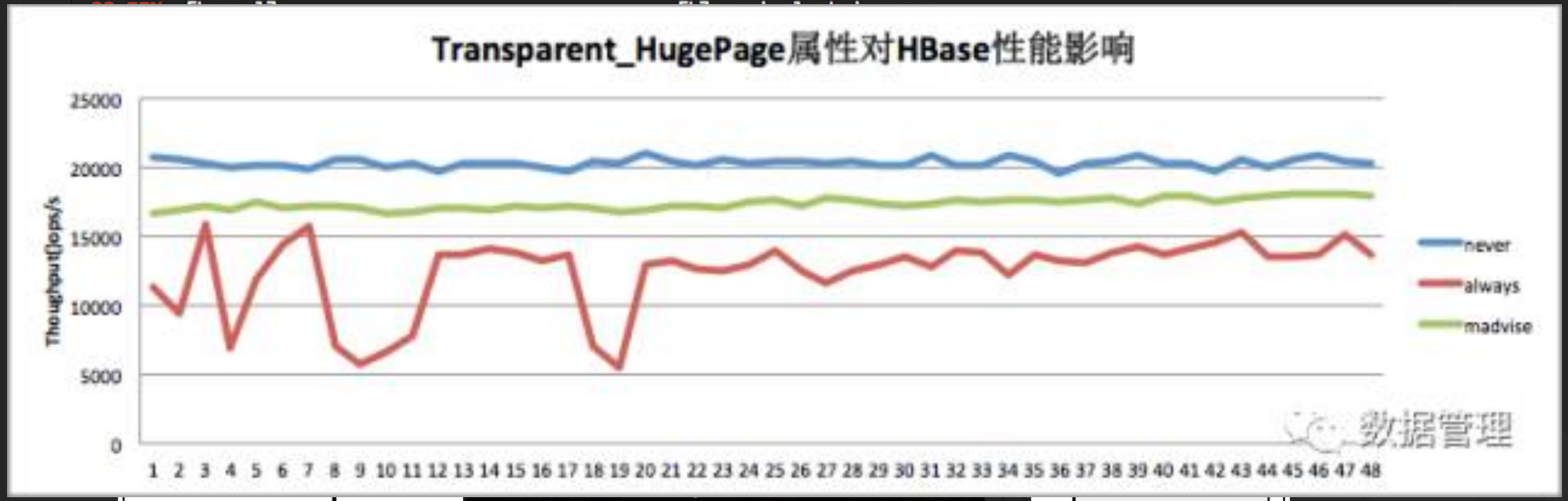
对小内存应用友好

设置zone_reclaim_mode=0

专属不够时允许访问远程内存

大内存之坑-透明大页 & perf工具

```
Samples: 17M of event 'cycles', Event count (approx.): 787773799469
```




```
try_to_compact_pages
__alloc_pages_direct_compact
__alloc_pages_nodemask
```

```
echo never > /sys/kernel/mm/transparent_hugepage/enabled
echo never > /sys/kernel/mm/transparent_hugepage/defrag
```

```
do_page_fault
+ page_fault
+ 5.03% compact_zone
```

memory

OS优化杂项

增加 打开文件数，进程数限制 (nproc，nofile）

关闭swap（ swappiness=0)

TCP快速回收重用，优化短连接（ tcp_tw_reuse = 1）

Question & Answering

Thanks!