ClickHouse万亿数据双中心的设计与实践

百分点 赵群（微信:5832842)

目录 CONTENTS

场景与挑战

数据量：2000亿+/日

高峰：500WRow/s

延时：<30秒

熔断/限流

存储与处理

2地双中心

查询/分析透明访问

查询

1TB常规查询<10s

1TB聚合查询(排序/分组)

<5m

离线任务

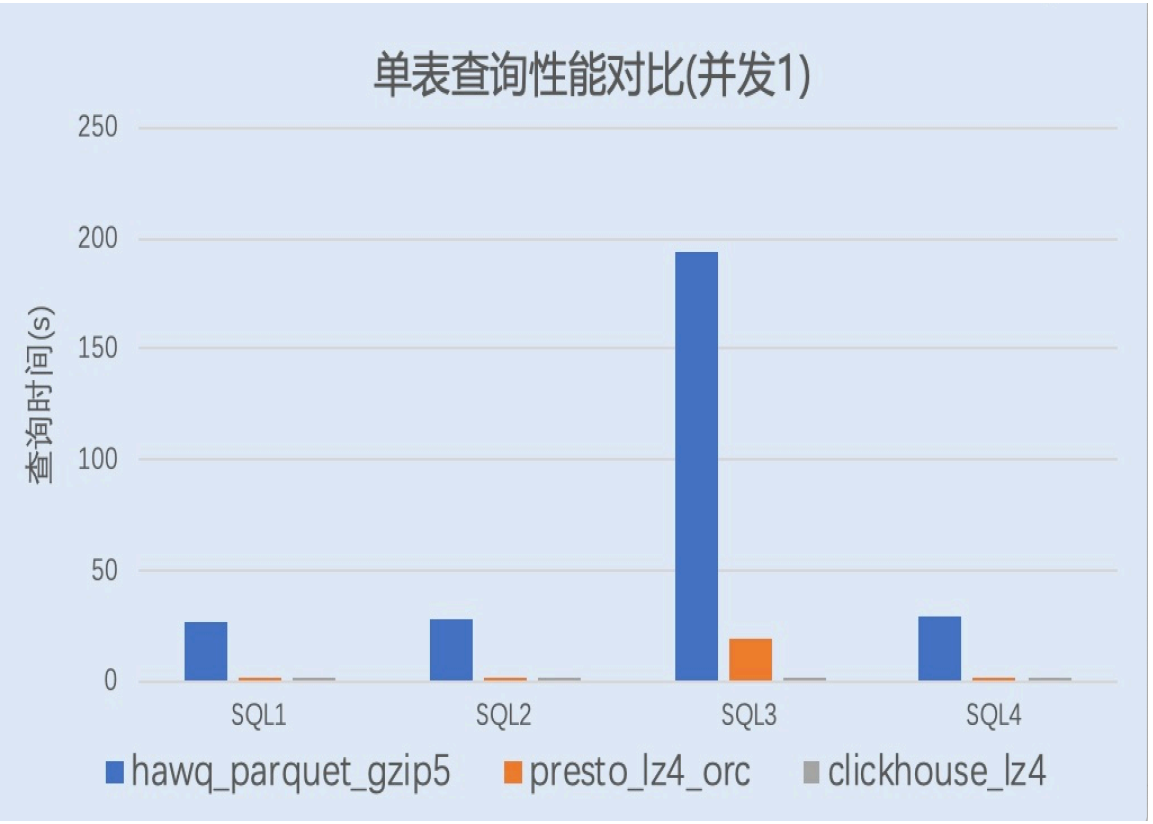
OLAP引擎评估

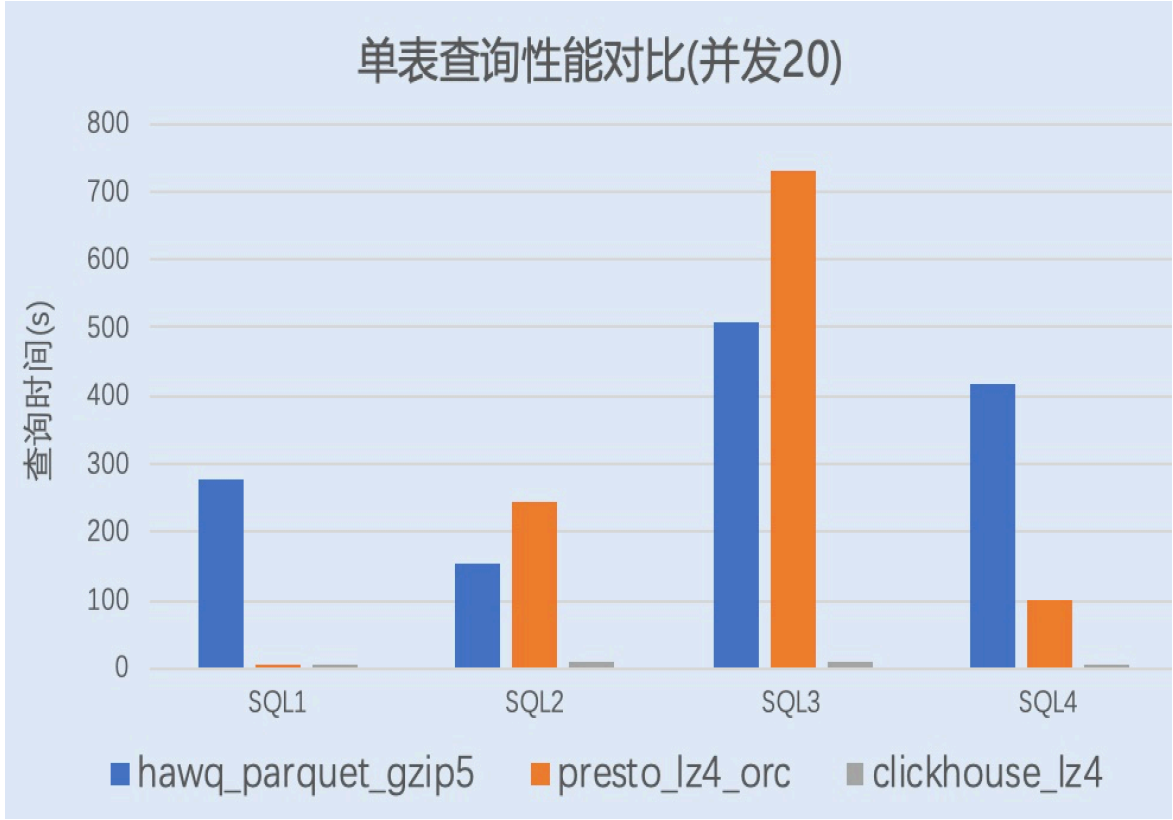
业 务：
1、超大规模的单表查询/分析；
2、有一定的并发要求；
3、实时性要求；

1. PB级的数据存储
2. 高性能的查询/分析能力
3. 低延时写入及吞吐能力
4. 数据压缩
5. 跨中心能力

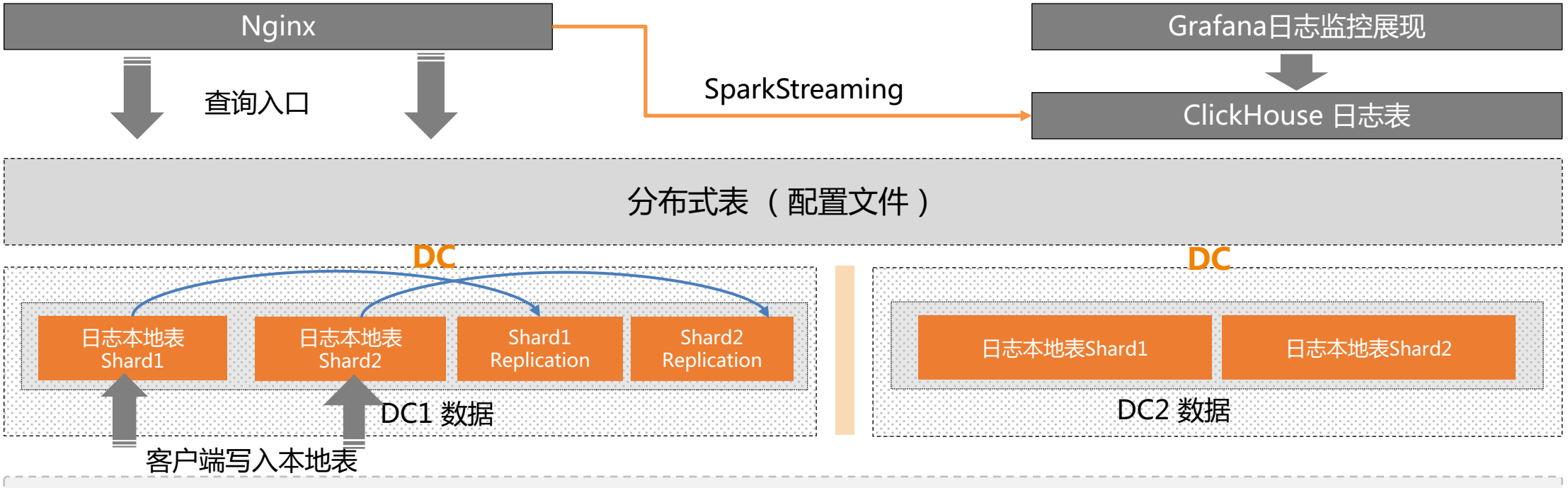
ClickHouse
Presto
HAWQ
Druid
Elastic Search

OLAP引擎评估


单表查询性能对比(并发1)
查询时间(s)
250
200
150
100
50
0
SQL1
SQL2
SQL3
SQL4
hawq_parquet_gzip5
presto_lz4_orc
clickhouse_lz4


单表查询性能对比(并发20)
查询时间(s)
800
700
600
500
400
300
200
100
0
SQL1
SQL2
SQL3
SQL4
hawq_parquet_gzip5
presto_lz4_orc
clickhouse_lz4

ClickHouse的2地双中心设计



1. ClickHouse跨中心透明访问。性能影响：1/4 ～1/3
2. 禁止分布式写。
3. 经过设计Replication是有稳定保障的
4. Nginx负载均衡，路由分发，安全加固
5. 日志采集、展现、分析

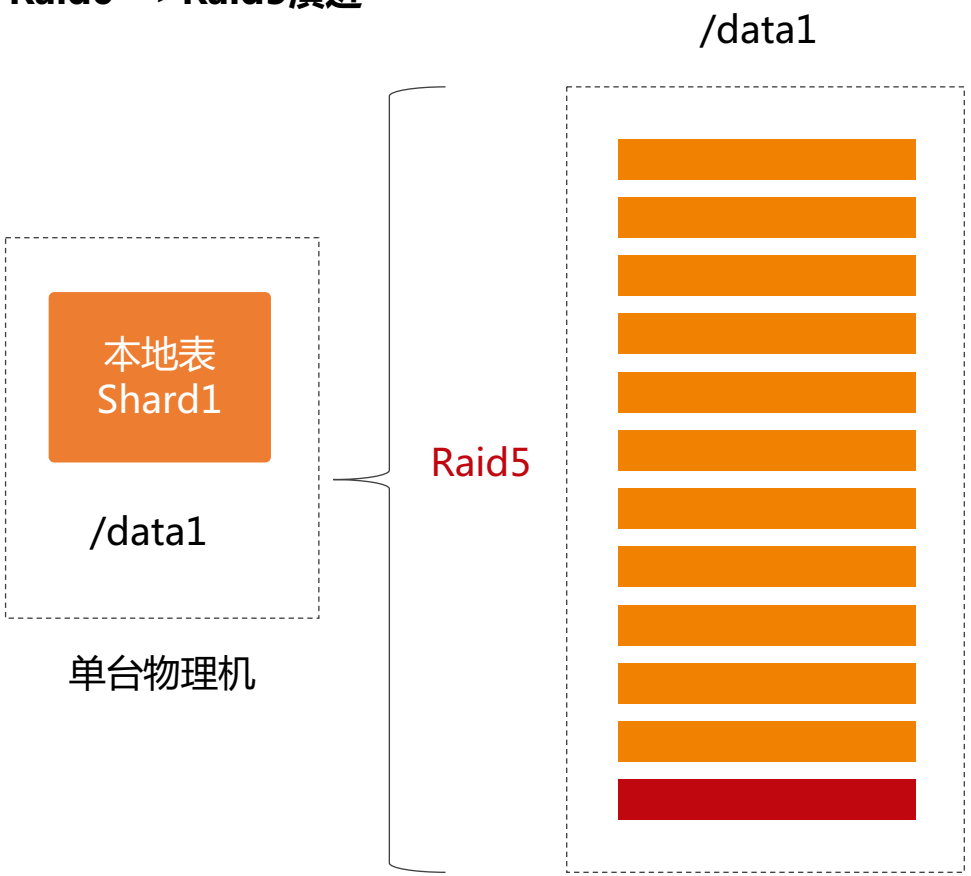
ClickHouse磁盘Raid的选择

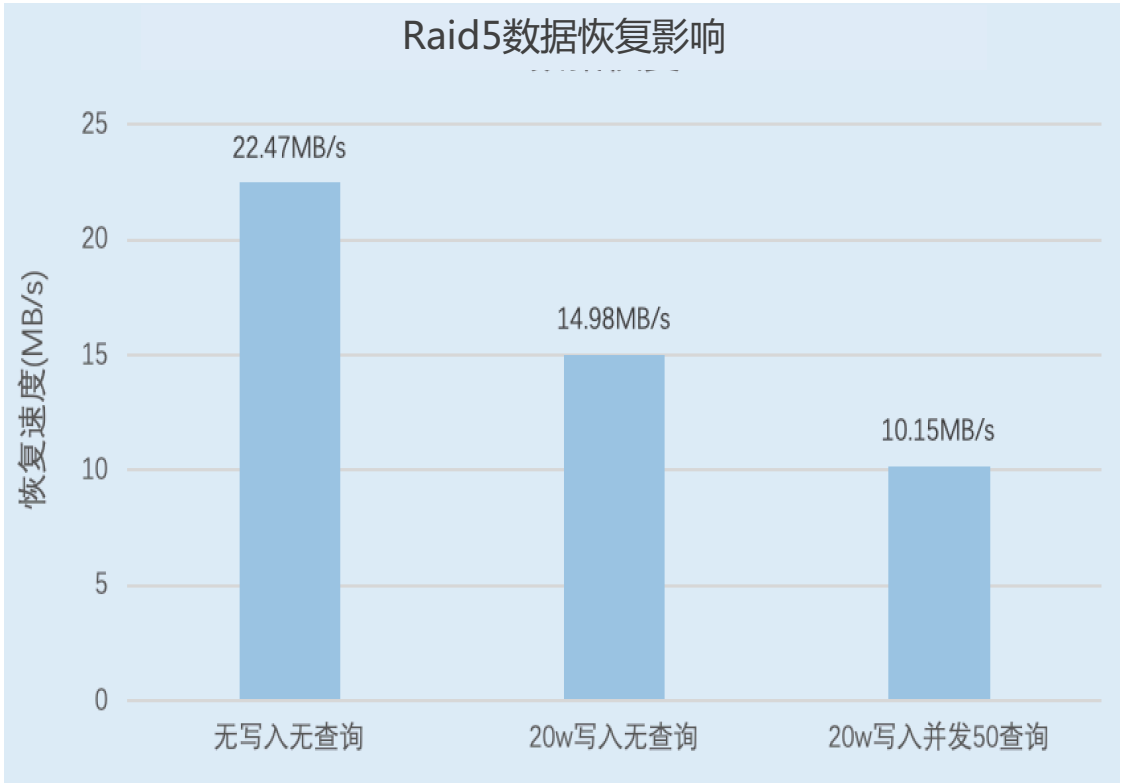
1、Raid5增加磁盘数据可靠性和读取能力

2、热备盘减少运维压力

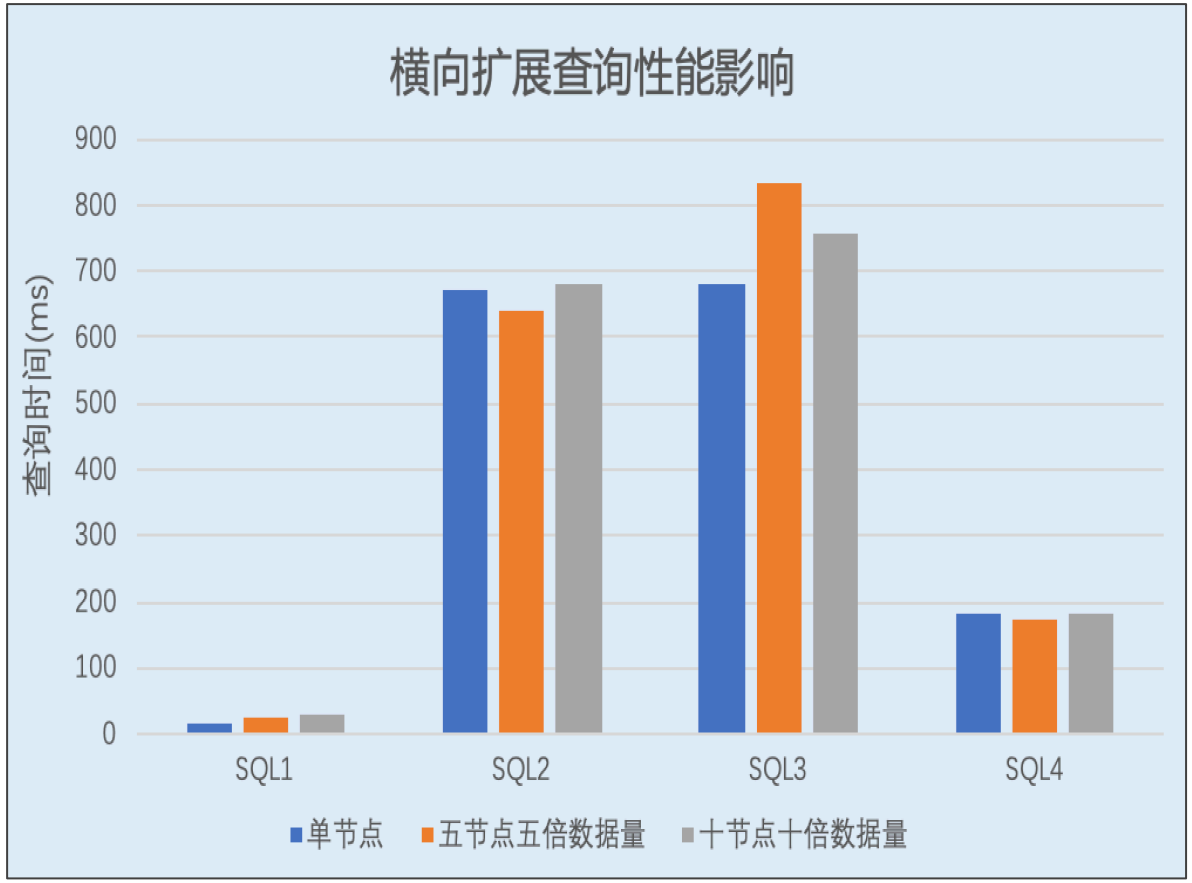
3、控制写入，保障查询性能

Raid0 - >Raid5演进

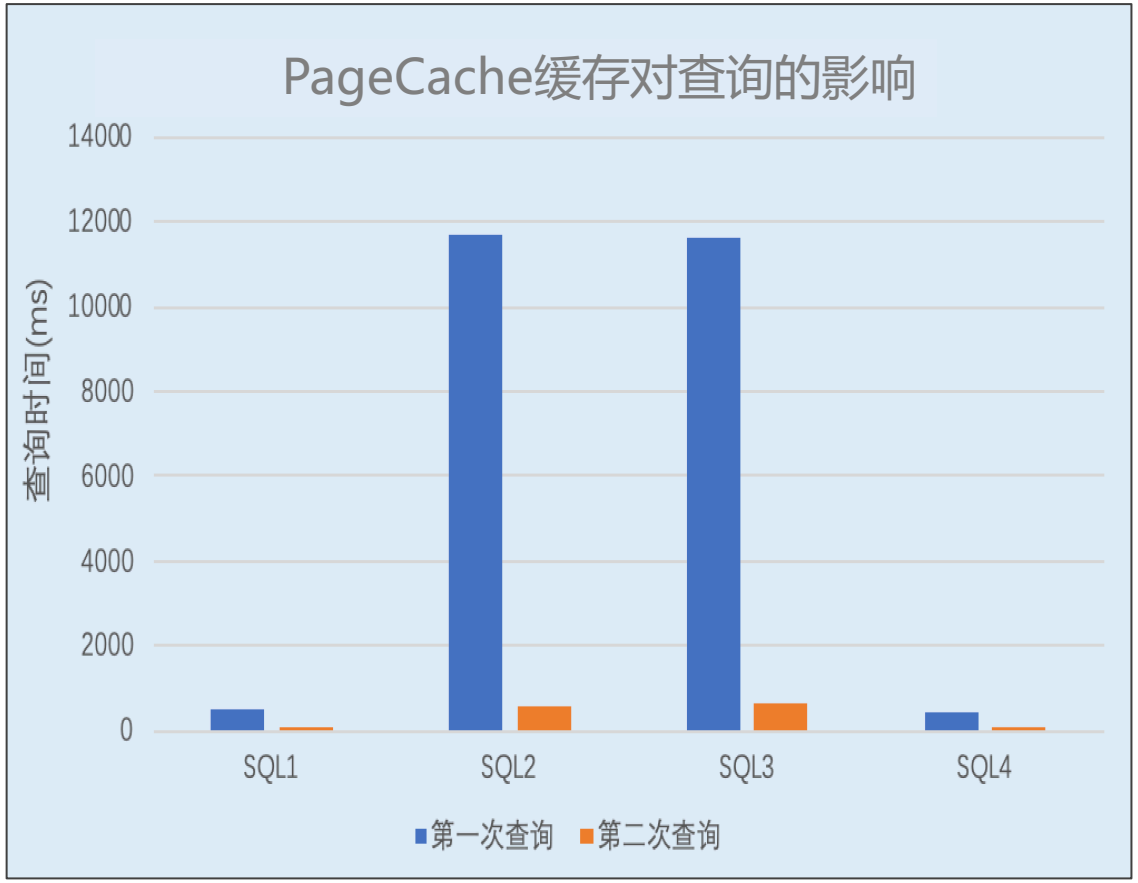




相关测试分析



横向扩展对查询性能几乎无影响
可以基于单节点/分区评估查询性能



数据预热对查询有数量级提升
针对缓存更换条件同样生效

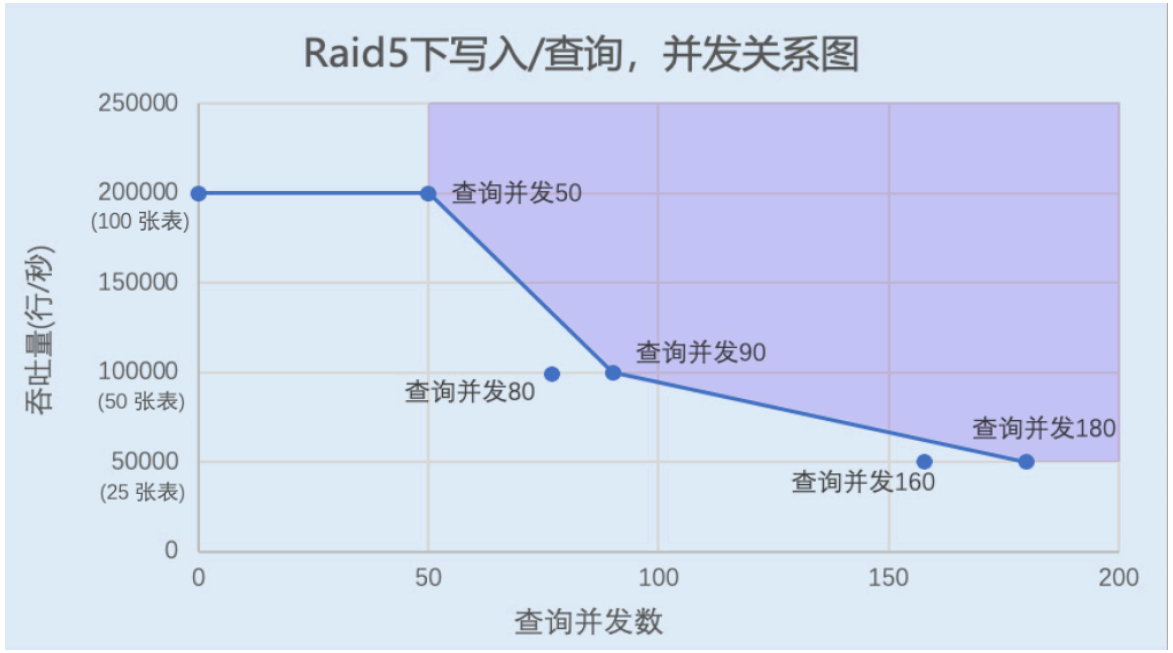
ClickHouse写入的稳定性设计

1、20W/s （35次）提交，并发50

2、10W/s(17次)提交，并发90

3、5W/s(8次)提交，并发90

确保业务命中在安全区域



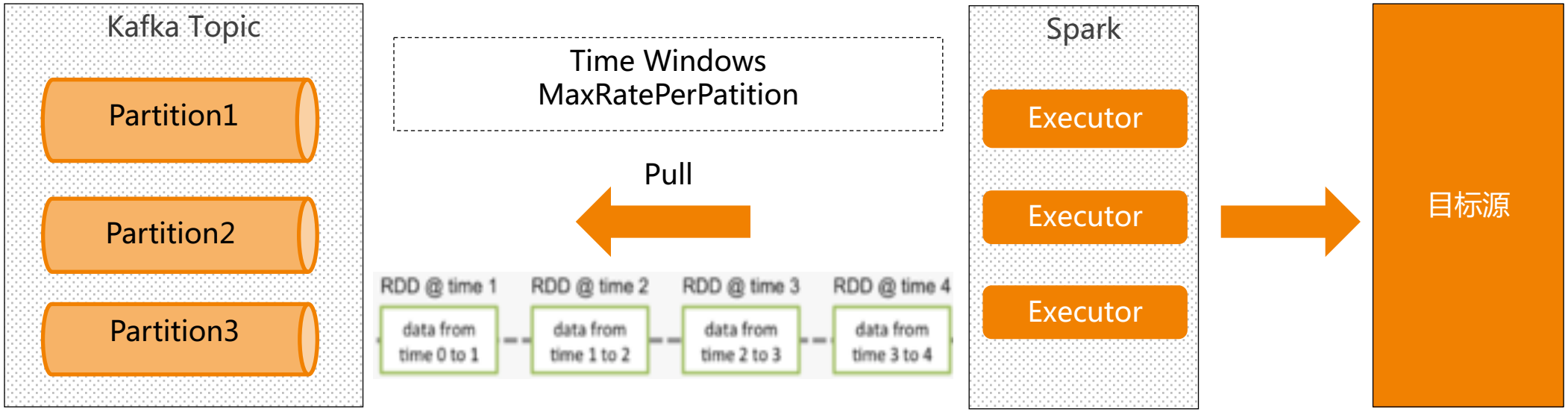
1、平衡好合并速度和Part数量的关系，一定是需要相对均衡的。

2、Part数量，实际代表着提交频率，一定是稳定，且经过估算的。

3、ClickHouse的查询和写入共同受限于Query数限制，需要分配好配额。

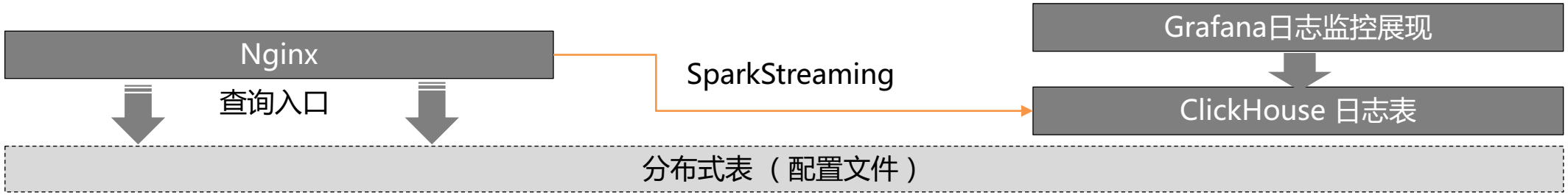
4、禁止直接写入分布式表。

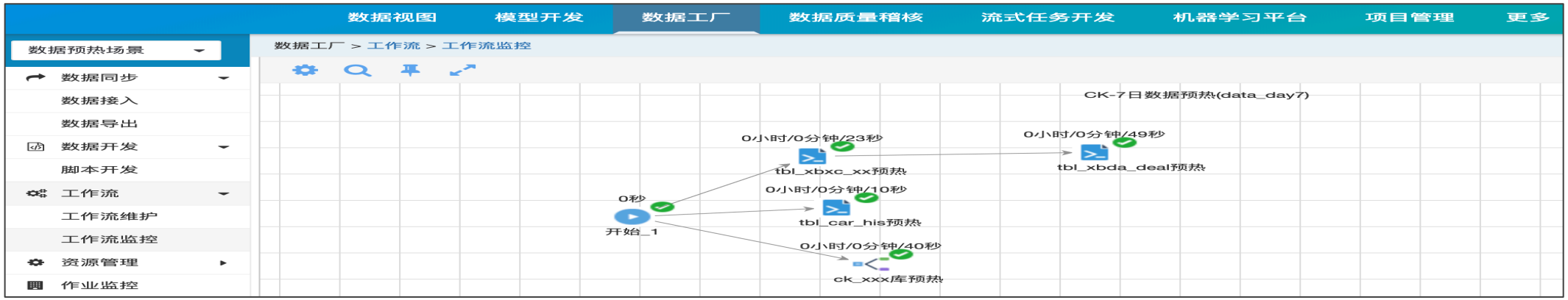
ClickHouse写入的稳定性设计



1. 时间窗口保障持续稳定提交频率。（保障对ClickHouse写入的稳定）
2. SparkStreaming 微批处理（控制处理上限），利用反压机制，实现处理能力动态平衡。
3. Spark on Yarn 资源可控。
4. 以写入ClickHouse为例，目前一个Executor处理在30000/s 左右。
5. 假设我们需要一个满足300W/s的处理能力。 在源读取没有瓶颈的情况下，可以 Executor数： 300 /3 = 100（个）.

ClickHouse的查询优化设计





1、限制单条查询内存使用量和单节点查询内存使用量，预防节点Down机。

2、Query数量限制异常：控制好配额/连接池。

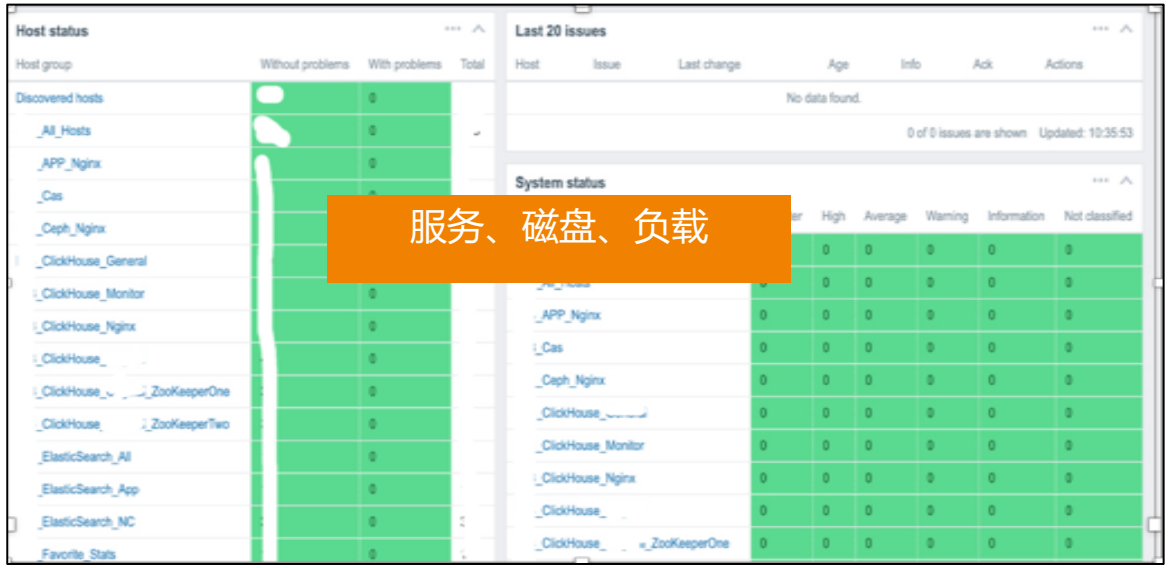
3、集群的Query日志，找出慢查询。我们直接通过Nginx收集了原始日志。

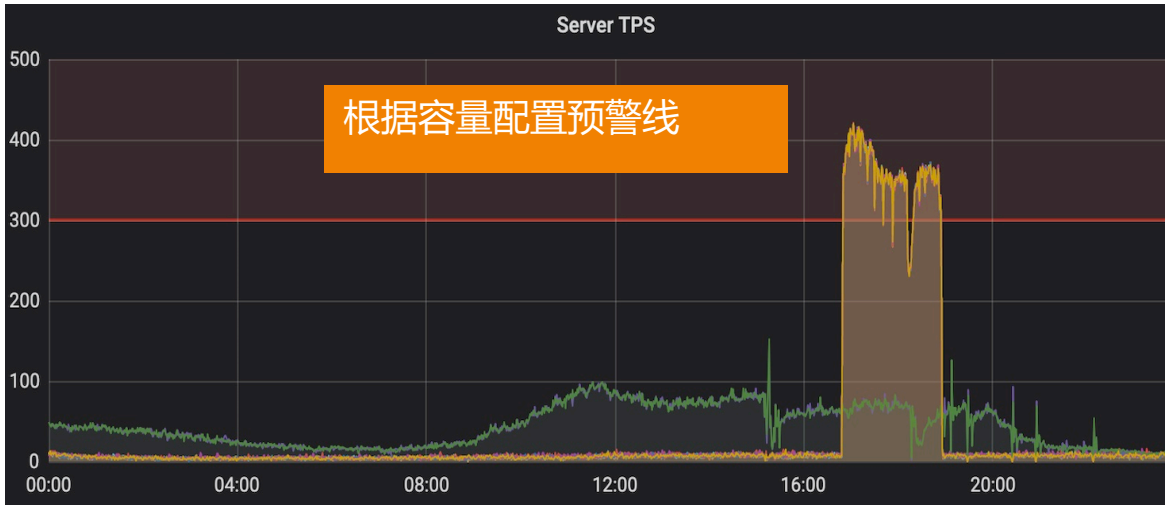
4、针对热数据进行查询预热。

最佳实践配置

参数名称	默认值	调整后的值	参数说明	参数所在配置文件
max_memory_usage_for_all_queries	0	200G	单台服务器上所有查询的内存使用量，默认没有限制	users.xml
max_memory_usage	10G	100G	一个查询在单台服务器的最大内存使用量，默认是10GB	users.xml
max_execution_time	0	300	单次查询耗时的最长时间，单位为秒。默认没有限制	users.xml
distributed_product_mode	deny	local	默认SQL中的子查询不允许使用分布式表，修改为local表示将子查询中对分布式表的查询转换为对应的本地表	users.xml
background_pool_size	16	32	后台用于merge的线程池大小	users.xml
log_queries	0	1	system.query_log表的开关。默认值为0，不存在该表。修改为1，系统会自动创建system.query_log表，并记录每次query的日志信息	users.xml
skip_unavailable_shards	0	1	当通过分布式表查询时，遇到无效的shard是否跳过。默认值为0表示不跳过，抛异常。设置值为1表示跳过无效shard	users.xml
keep_alive_timeout	10	600	服务端与客户端保持长连接的时长，单位为秒	config.xml
max_concurrent_queries	100	150	最大支持的Query数量	config.xml
session_timeout_ms	3000	120000	ClickHouse服务和Zookeeper保持的会话时长，超过该时间Zookeeper还收到不ClickHouse的心跳信息，会将与ClickHouse的Session断开	metrika.xml

ClikHouse的监控





用数据智能推动社会进步