Clickhouse玩转每天千亿数据

趣头条 王海胜

提纲

- 业务背景
- 集群现状
- 我们遇到的问题

业务背景

基于storm的实时指标的计算存在的问题

1：指标口径(SQL) -> 实时任务

2：数据的回溯

3：稳定性

业务背景

什么是我们需要的?

1：实时指标SQL化

2：数据方便回溯，数据有问题，方便恢复

3：运维需要简单

4：计算要快，在一个周期内，要完成所有的指标的计算

集群现状

100+台32核128G

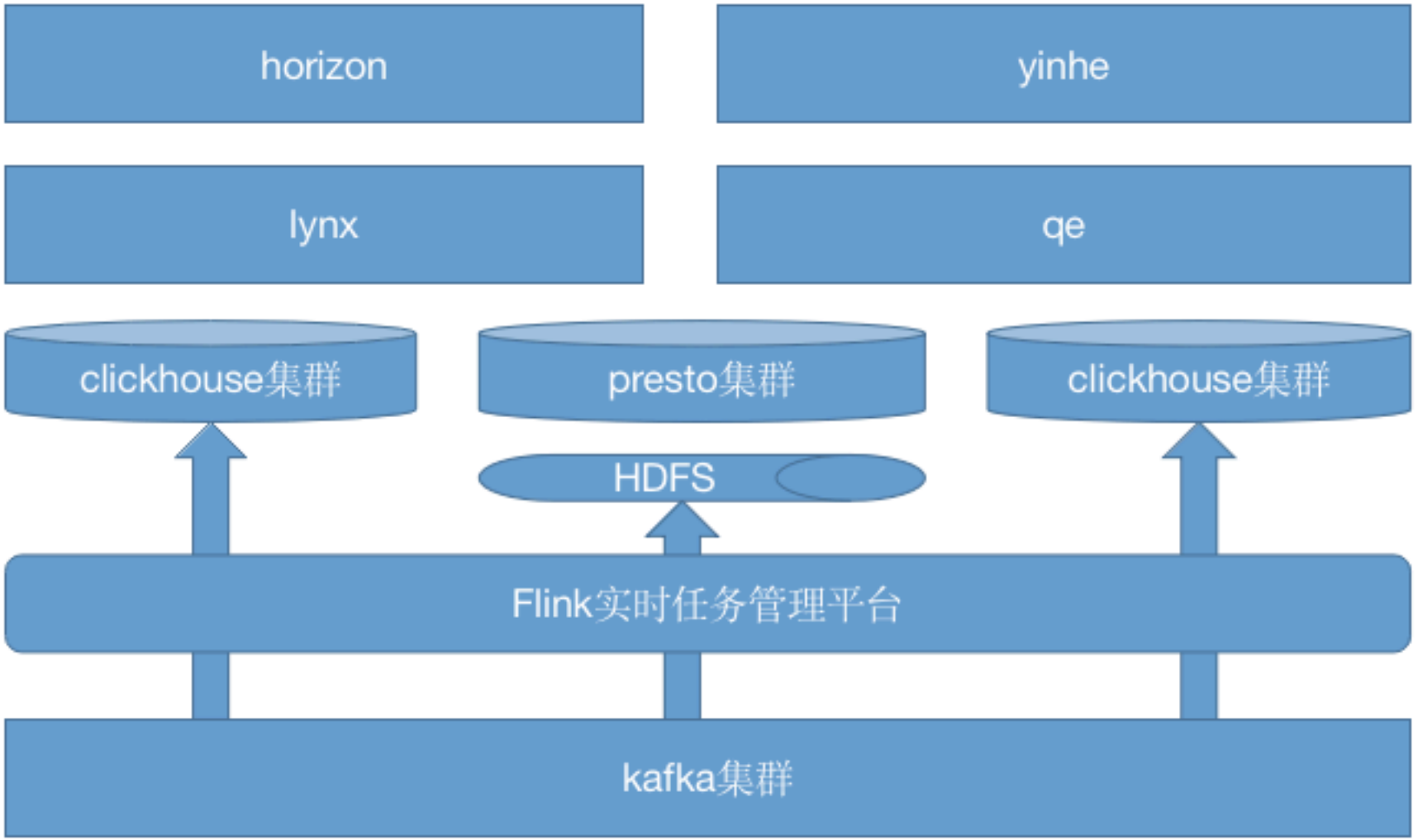
1000亿/天

21万次查询/天

80%查询1S内完成

部分复杂累时查询30S内完成

集群现状


horizon
yinhe
lynx
qe
clickhouse集群
presto集群
clickhouse集群
HDFS
Flink实时任务管理平台
kafka集群

我们遇到的问题

关于机器的配置

早期集群机器配置16核64G 一块1.7T本地SSD

问题：
1：内存限制，对于一些大的查询会出现内存不够问题
2：存储限制，随着表越来多，磁盘报警不断
3：cpu限制
64G对于一些大表(每天600亿+)的处理，很容易报错，虽然有基于磁盘解决方案，但是会影响速度
clickhouse的数据目录还不支持多个数据盘，单块盘的大小限制太大
cpu需要根据实际情况而定

解决：
1：机器的内存推荐128G+
2：采用软连接的方式，把不同的表分布到不同的盘上面，这样一台机器可以挂载更多的盘

最新版本的”冷热数据分离”特性，曲线救国?

我们遇到的问题

order by (timestamp, eventType) or order by (eventType, timestamp)

业务场景
1：趣头条和米读的上报数据是按照”事件类型”(eventType)进行区分
2：指标系统分”分时”和”累时”指标
3：指标的一般都是会按照eventType进行区分
select count(1) from table where dt='' and timestamp>='' and timestamp<='' and eventType=''

建表的时候缺乏深度思考，由于分时指标的特性，我们的表是order by (timestamp, eventType)进行索引的，这样在计算累时指标的时候出现非常耗时(600亿+数据量)

分析：
对于累时数据，时间索引基本就失效了，由于timestamp”基数”比较高，对于排在第二位eventType索引，这个时候对数据的过滤就非常有限了，这个时候几乎就要对当天的数据进行全部扫描

解决：
1：调整索引的顺序，推荐索引列的基数不要太高.

我们遇到的问题

Too many parts(304). Merges are processing significantly slower than inserts.

分析：

1：直接落盘，异步merge - background_pool_size

2：一个Insert Request，涉及N个分区的数据，在磁盘上就会生成N个数据目录，merge跟不上

3：一个目录，一个zxid，zookeeper集群的压力大，插入速度严重变慢

解决：

1：增大background_pool_size治标不治本

2：设置分区的时候需要思考，数据的特性需要了解

我们遇到的问题

查询过程中clickhouse-server进程挂掉

分析：
clickhouse裸奔时max_memory_usage_for_all_queries默认值为0，即不限制clickhouse内存使用

解决：
clickhouse安装完成以后，在users.xml文件中配置一下max_memory_usage_for_all_queries，控制clickhouse-server最大占用内存，避免被OS kill

我们遇到的问题

Memory limit (for query) exceeded:would use 9.37 GiB (attempt to allocate chunk of 301989888 bytes), maximum: 9.31 GiB

分析：

1：max_memory_usage指定单个SQL查询在该机器上面最大内存使用量

2：除了些简单的SQL，空间复杂度是O(1) 如:

select count(1) from table where column=value

select column1, column2 from table where column=value

凡是涉及group by, order by, distinct, join这样的SQL内存占用不再是O(1)

解决：

1：max_bytes_before_external_group_by

2：max_bytes_before_external_sort

3：uniq / uniqCombined / uniqHLL12

4：Join时小表放到右边，“右表广播”^v^

我们遇到的问题

zookeeper相关的问题

问题一：zookeeper的snapshot文件太大，follower从leader同步文件时超时
问题二：zookeeper压力太大，clickhouse表处于”read only mode”，插入失败

分析：
clickhouse对zookeeper的依赖还是很重的，有大量的数据需要写到zookeeper上面，数据Part都在zookeeper上面有个节点与之对应以及表的元数据信息等等.

解决：
1：zookeeper机器的snapshot文件和log文件最好分盘存储(推荐SSD)提高ZK的响应
2：zookeeper的snapshot文件存储盘不低于1T
3：做好zookeeper集群和clickhouse集群的规划，可以多套zookeeper集群服务一套clickhouse集群
3.1：zookeeper集群的znode最好能在400w以下(这个时候snapshot文件到达2G+)
3.2：注意监控zookeeper的指标(排队请求?处理延迟?等等)，排队请求太多可能会导致插入失败

我们遇到的问题

关于引擎选择

推荐Replicated*MergeTree引擎

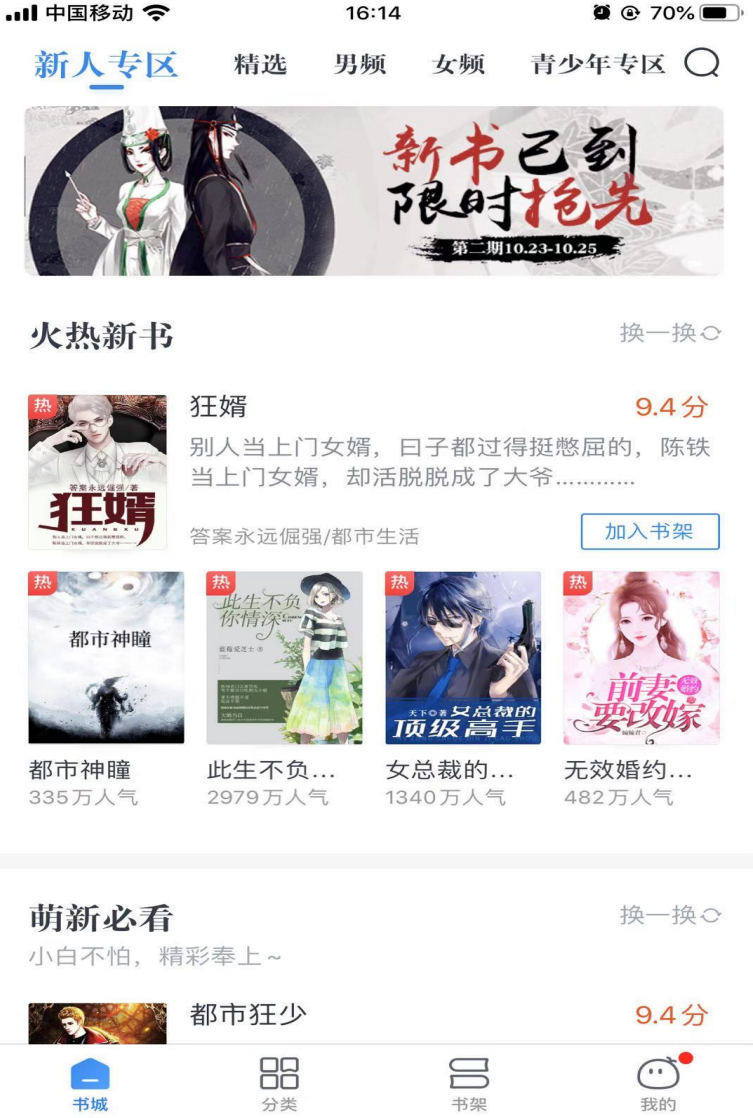
1：安全，数据安全，业务安全

2：升级的时候可以做到业务无感知

3：提升查询的并发度

广告时间


中国移动
16:12
70%
搜索你感兴趣的内容
推荐
视频
上海
热点
小说
娱乐
习近平：加快推动区块链产业创新发展
置顶
新华社 8评
20小时前
全年粮食生产再获丰收
置顶
人民网 4评
5小时前
第四届全国财经院校创新创业大赛在浙江财经大学举行
冠军
中国青年报
刚刚
今日阅读奖励，恭喜获得350金币奖励！
今日奖励
+350
立即领取
刷新
视频
小视频
我的


中国移动
16:14
70%
新人专区
精选
男频
女频
青少年专区
新书已到
限时抢先
第二期10.23-10.25
火热新书
换一换
狂婿
9.4分
别人当上门女婿，曰子都过得挺憋屈的，陈铁当上门女婿，却活脱脱成了大爷.....
答案永远倔强/都市生活
加入书架
都市神瞳
335万人气
此生不负...
2979万人气
女总裁的...
1340万人气
无效婚约...
482万人气
萌新必看
换一换
小白不怕，精彩奉上～
都市狂少
9.4分
书城
分类
书架
我的