营赛洞察

数据驱动营销新纪元

东信云利用 Apache Doris 释放内容分析潜能

梁爱民

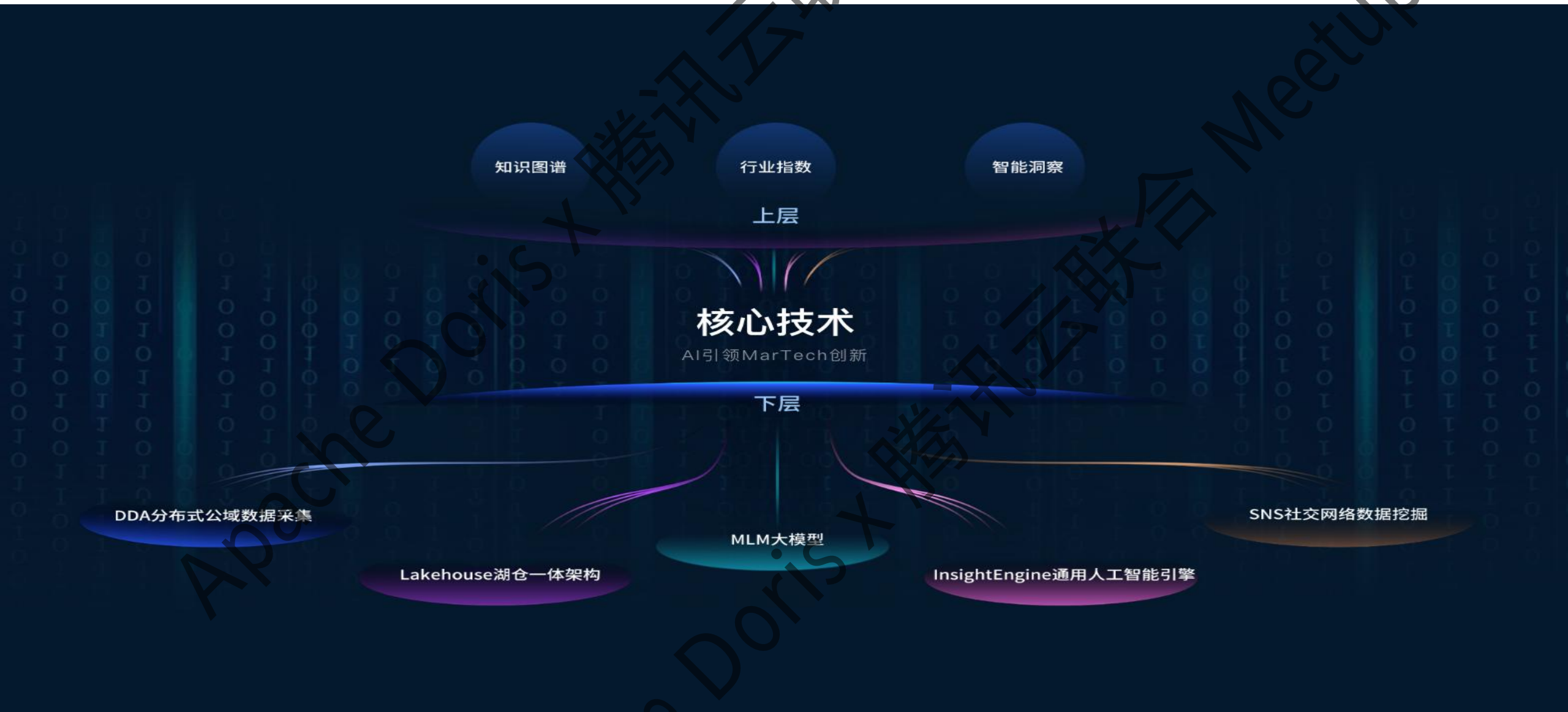
深圳市东信云科技有限公司 高级大数据开发工程师

CONTENTS

目录

东信简介

东信，中国领先的以AI+大模型+大数据驱动的营销科技企业，自2004年成立以来，始终引领中国MarTech技术的研究与应用，赋能千行百业的数智营销和数智商业场景。
东信构建了强大的技术壁垒和业务韧性，客户遍布大消费及大互联网行业的137个垂类。凭借杰出创新成就，荣登“中国互联网综合实力百强企业”第61位，广东企业500强、深圳创新企业100强等荣誉榜单。



数智营销全流程洞察平台

营赛洞察

一站式智能营销内容投放管理平台

鹊挢

更多产品

敬请期待

引流数据聚合平台

云知数

营销效果洞察平台

MarketingDesk（标准版）

更多产品

敬请期待

CONTENTS

目录

营赛洞察全流程数智洞察平台

通过数字化分析数据，优化决策过程、精细目标受众定位，助力企业实现内容营销全流程的数字化转型

产品优势

基于实时的营销内容大数据的洞察与分析，助力品牌内容营销决策

基于认知智能的数据建模

利用全场景认知智能能力，对海量行业数据进行建模，洞察行业竞争力和品牌认知度。

自然语言处理与情感分析

运用深度学习模型和语义关系挖掘，深入分析行业热点话题和品牌口碑，优化品牌策略。

强化学习和智能决策优化

引入强化学习算法和智能决策优化框架，完成营销决策的自动化优化，实现最佳品牌决策。

大数据处理与实时分析

强大的大数据处理和实时分析能力，实现对行业趋势、竞争对手和活动效果的决策支持。

营赛洞察全流程数智洞察平台

营赛自研 50+针对不同的营销场景的自研算法&模型，涵盖了用户行为预测、内容推荐、市场分析和客户细分等多个领域，旨在通过数据驱动的洞察提升营销效果和优化用户体验。

产品优势

CONTENTS

目录

需求与挑战

大规模热数据存储

PB 级别的数据存储与 TB 级别热数据检索的需求

高性能NLP检索

通过多维度的 NLP 检索与内容分析

低延时的数据写入

更新时效延要求高

可拓展性

能够根据数据量规模进行动态扩缩

CONTENTS

目录

整体方案

挑战： 高斯DB/ES不能满足要求 -> 改Doris

写入延迟高

每日更新热数据有 30 亿，同步时间过长

存储成本量高

ES 的存储成本与日俱增
使用中的GaussDB 2.0 不支持冷热分离

并发要求和横向拓展

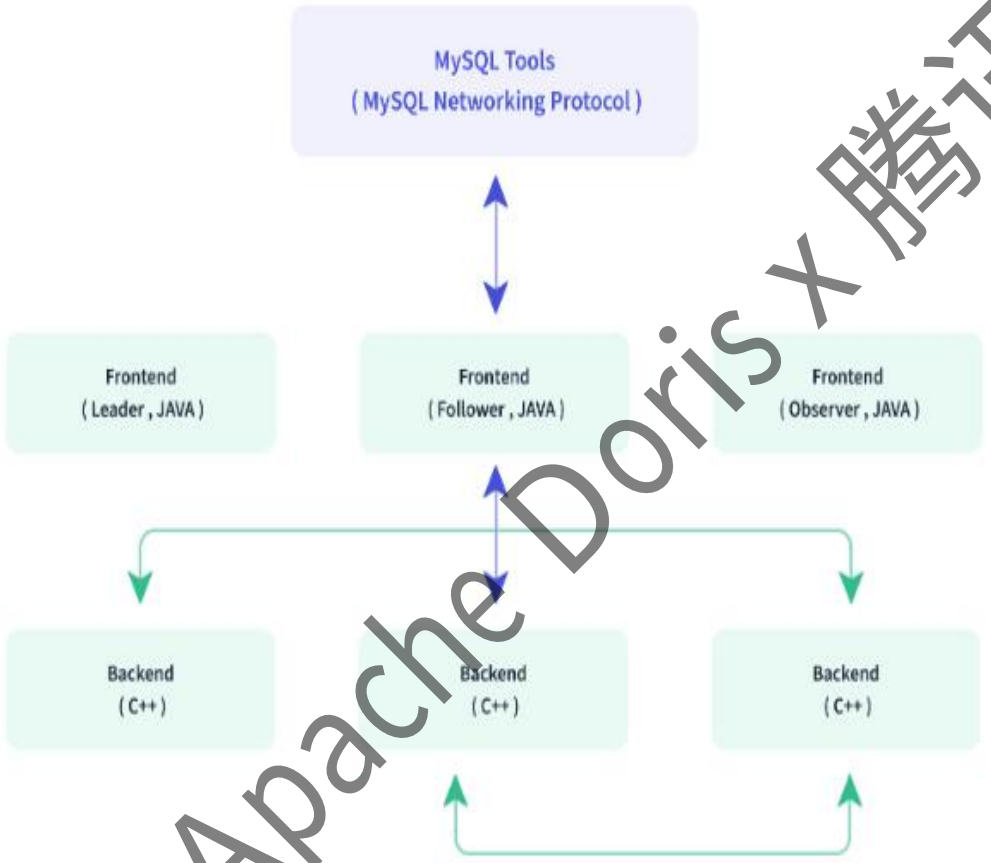
业务并发要求对于读写分离的要求较高

云原生开发的趋势之下，对于资源的弹性扩缩容要求激增

ES/CK vs. Doris

性能指标	Elasticsearch	Doris	ClickHouse
NLP检索	支持	支持（4.15）	不支持
写入速度	131 MB/s	550 MB/s	
存储成本（1亿行数据）	330G	60G	
扩展能力	分布式架构易于水平扩展	分布式架构易于水平扩展	
社区活跃度	高	高	

MDS 产品下Doris的优势



高效的词语距离检索

极致的数据压缩比

高效的数据写入流程

横向拓展能力与优质的社区服务

整体方案

基于 Doris 的数据架构方案

写入延迟

数据写入从 2H 降低到 18min

存储成本量高

在冷热分离的基础上，数据量存储降低至原来的 1/5

并发要求和横向拓展

横向节点扩展简易方便

社区支撑度好

基于 Doris 的方案：存储效能提升

缩减索引数据量

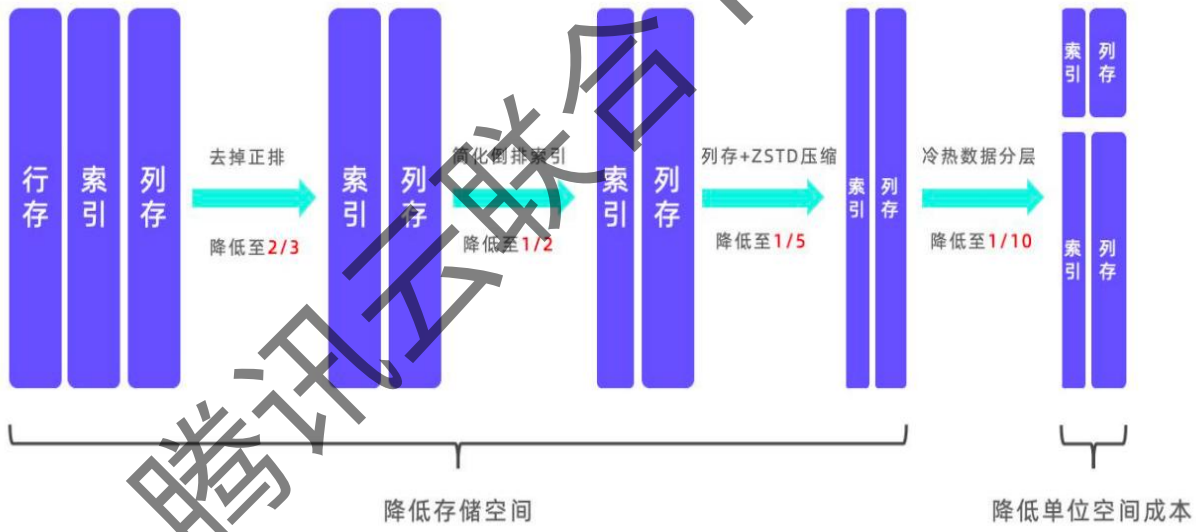
在存储上去掉正排，缩减了 30% 的索引数据量

zstd
压缩算法

采用列式存储和 zstd 压缩算法，压缩比可达到 5-10 倍，远高于 Elasticsearch 的 1.5 倍；

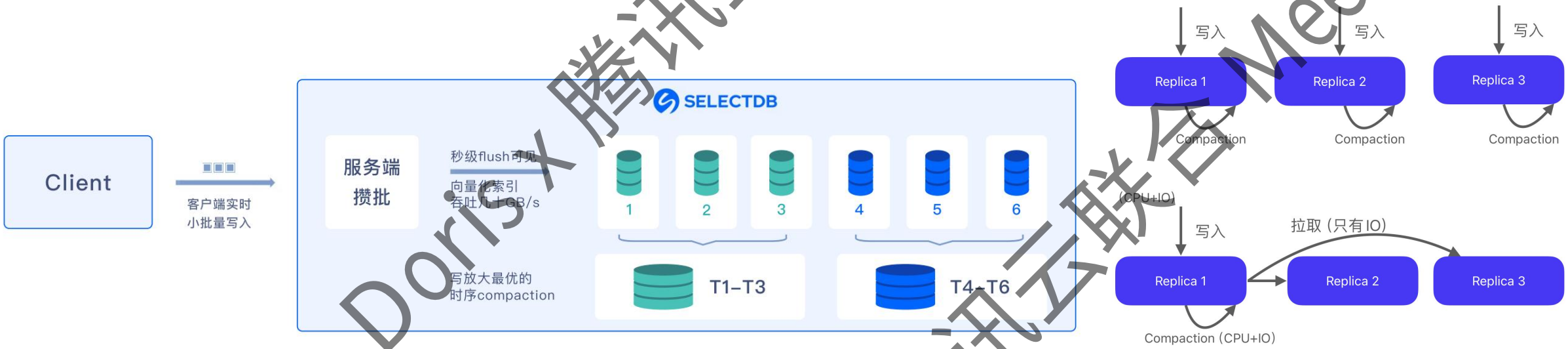
冷热分层

针对日志数据中访问频率很低的冷数据，SelectDB 冷热分层功能可以将超过定义时间段的日志自动存储到更低的对 象存储中，冷数据的存储成本可降低 70% 以上，在后续的使用中，会考虑使用对象存储承载超过一段时间的冷数据。



应用成效：提高写入吞吐量

在相同的资源配置情况下，写入性能 SelectDB 相比于 ES 提升 4 倍



- 向量化指令，提升数据解析，索引构建的性能
- 简化去掉正排等索引结构，降低索引构建开销
- 单副本导入，单副本 compaction

查询优化：短语检索支持词距（slop）

需求：检索同时包含‘立白’和‘洗衣液’两个词的行，且它们之间间隔的词个数（词距）不超过 3 个

正例：‘立白洗衣液’，‘立白去污洗衣液’都能匹配，因为词距分别是 0 和 1

反例：‘立白最新研发的一种新型洗衣液’不能匹配，因为词距是 4

已有全文检索功能不完全满足：

- content MATCH_ALL ‘立白 洗衣液’要求两个词同时出现，但是不限制词距
- content MATCH_PHRASE‘立白 洗衣液’要求两个词同时出现而且连着，也就是限制词距为 0

增强全文检索支持用户自定义词距：

- content MATCH_PHRASE ‘立白 洗衣液 ~3’

查询优化：短语检索支持词距（slop +）

增强全文检索支持用户自定义词距：
content MATCH_PHRASE ‘立白 洗衣液 ~3’

‘洗衣液品牌立白’能否匹配呢？ 结果是能匹配，和大多数人的直觉不一样

增强全文检索支持用户自定义正向词距：
content MATCH_PHRASE ‘立白 洗衣液 ~3+’

语义更符合直觉，性能提升 30%，因为不用做两个方向的词查找，而且查询词越多效果越明显

查询优化：上千个过滤条件性能优化

```
SELECT source, COUNT(*) as content_cnt
FROM `mds_content_detail` WHERE (dt BETWEEN '2024-06-26' AND '2024-07-26') AND ((((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '李叙白'
    AND content MATCH_PHRASE '立白 李叙白 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '李叙白' AND title MATCH_PHRASE '立白 李叙白 ~15')))
OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '陈佩斯' AND content MATCH_PHRASE '立白 陈佩斯 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PH
    AND title MATCH_PHRASE '立白 陈佩斯 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '欧阳娜娜' AND content MATCH_PHRASE '立白 欧阳娜娜 ~15
OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '欧阳娜娜' AND title MATCH_PHRASE '立白 欧阳娜娜 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATC
AND content MATCH_PHRASE '>立白 种地吧少年 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '种地吧少年'
AND title MATCH_PHRASE '立白 种地吧少年 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '十个勤天'
AND content MATCH_PHRASE '立白 十个勤天 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '十个勤天'
AND title MATCH_PHRASE '立白 十个勤天 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '集团'
AND content MATCH_PHRASE '立白 集团 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '集团' AND title MATCH_PHRASE '立白 集团 ~15')))
OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '洗护' AND content MATCH_PHRASE '立白 洗护 ~15') OR (title MATCH_PHRASE '立白'
AND title MATCH_PHRASE '洗护' AND title MATCH_PHRASE '立白 洗护 ~15'))) OR (((content MATCH_PHRASE '立白'
AND content MATCH_PHRASE '家清' AND content MATCH_PHRASE '立白 家清 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '家清'
AND title MATCH_PHRASE '立白 家清 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '官旗' AND content MATCH_PHRASE '立白 官旗 ~15')
OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '官旗' AND title MATCH_PHRASE '立白 官旗 ~15'))) OR (((content MATCH_PHRASE '立
白' AND content MATCH_PHRASE '旗舰店' AND content MATCH_PHRASE '立白 旗舰店 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '旗舰店'
AND title MATCH_PHRASE '立白 旗舰店 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '天猫' AND content MATCH_PHRASE '立白 天猫 ~15')
OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '天猫' AND title MATCH_PHRASE '立白 天猫 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRAS
AND content MATCH_PHRASE '立白 京东 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '京东' AND title MATCH_PHRASE '立白 京东 ~15'))) OR (((content MA
AND content MATCH_PHRASE '除菌' AND content MATCH_PHRASE '立白 除菌 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '除菌' AND title MATCH_PHRASE '立
OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '心心珠' AND content MATCH_PHRASE '立白 心心珠 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PH
AND title MATCH_PHRASE '立白 心心珠 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '洗衣珠' AND content MATCH_PHRASE '立白 洗衣珠 ~15')
OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '洗衣珠' AND title MATCH_PHRASE '立白 洗衣珠 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PH
AND content MATCH_PHRASE '立白 皂液 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '皂液' AND title MATCH_PHRASE '立白 皂液 ~15'))) OR (((content MA
AND content MATCH_PHRASE '白衣净' AND content MATCH_PHRASE '立白 白衣净 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '白衣净' AND title MATCH_PHRA
' AND title MATCH_PHRASE '立白 皂粉 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '去渍粉' AND content MATCH_PHRASE '立白 去渍粉 ~15') OR (
AND title MATCH_PHRASE '去渍粉' AND title MATCH_PHRASE '立白 去渍粉 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '除菌粉' AND content MATC
OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '除菌粉' AND title MATCH_PHRASE '立白 除菌粉 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PH
' AND title MATCH_PHRASE '立白 内衣皂 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '专用皂' AND content MATCH_PHRASE '立白 专用皂 ~15') OR
粉' AND title MATCH_PHRASE '立白 洗碗粉 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '柔顺剂' AND content MATCH_PHRASE '立白 柔顺剂 ~15') O
' AND content MATCH_PHRASE '不伤手' AND content MATCH_PHRASE '立白 不伤手 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '不伤手'
AND title MATCH_PHRASE '立白 不伤手 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '感恩欢购节'
AND content MATCH_PHRASE '立白 感恩欢购节 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '感恩欢购节'
AND title MATCH_PHRASE '立白 感恩欢购节 ~15'))) OR (((content MATCH_PHRASE '立白' AND content MATCH_PHRASE '直播间' AND content MATCH_PHRASE '立白 直播间 ~15')
OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '直播间' AND title MATCH_PHRASE '立白 直播间 ~15'))) OR (((content MATCH_PHRASE '立白'
AND content MATCH_PHRASE '小白鞋' AND content MATCH_PHRASE '立白 小白鞋 ~15') OR (title MATCH_PHRASE '立白' AND title MATCH_PHRASE '小白鞋' AND title MATCH_PHRA
阳娜娜 ~15'))) OR (((content MATCH_PHRASE 'liby' AND content MATCH_PHRASE '种地吧少年' AND content MATCH_PHRASE 'liby 种地吧少年 ~15')
OR (title MATCH_PHRASE 'liby' AND title MATCH_PHRASE '种地吧少年' AND title MATCH_PHRASE 'liby 种地吧少年 ~15'))) OR (((content MATCH_PHRASE 'liby' AND content
 ~15') OR (title MATCH_PHRASE 'liby' AND title MATCH_PHRASE '洗洁精' AND title MATCH_PHRASE 'liby 洗洁精 ~15'))) OR (((content MATCH_PHRASE 'liby' AND content
```

1. 倒排索引查询路径优化
2. IN 利用倒排索引加速降低 CPU 消耗，查询时间 2分钟 => 10s

查询优化：cache warmup

```
76    PROPERTIES (
77    "file_cache_ttl_seconds" = "31536000",
78    "bloom_filter_columns" = "dwd_content_id",
79    "dynamic_partition.enable" = "true",
80    "dynamic_partition.time_unit" = "MONTH",
81    "dynamic_partition.time_zone" = "Asia/Shanghai",
82    "dynamic_partition.start" = "-24",
83    "dynamic_partition.end" = "12",
84    "dynamic_partition.prefix" = "p",
85    "dynamic_partition.buckets" = "32",
```

现象：

设置了file_cache_ttl_seconds 缓存保　存时间。但是仍然有数据会走 obs 查询，查询速度较慢

原因：

历史缓存数据未淘汰

解决方法：

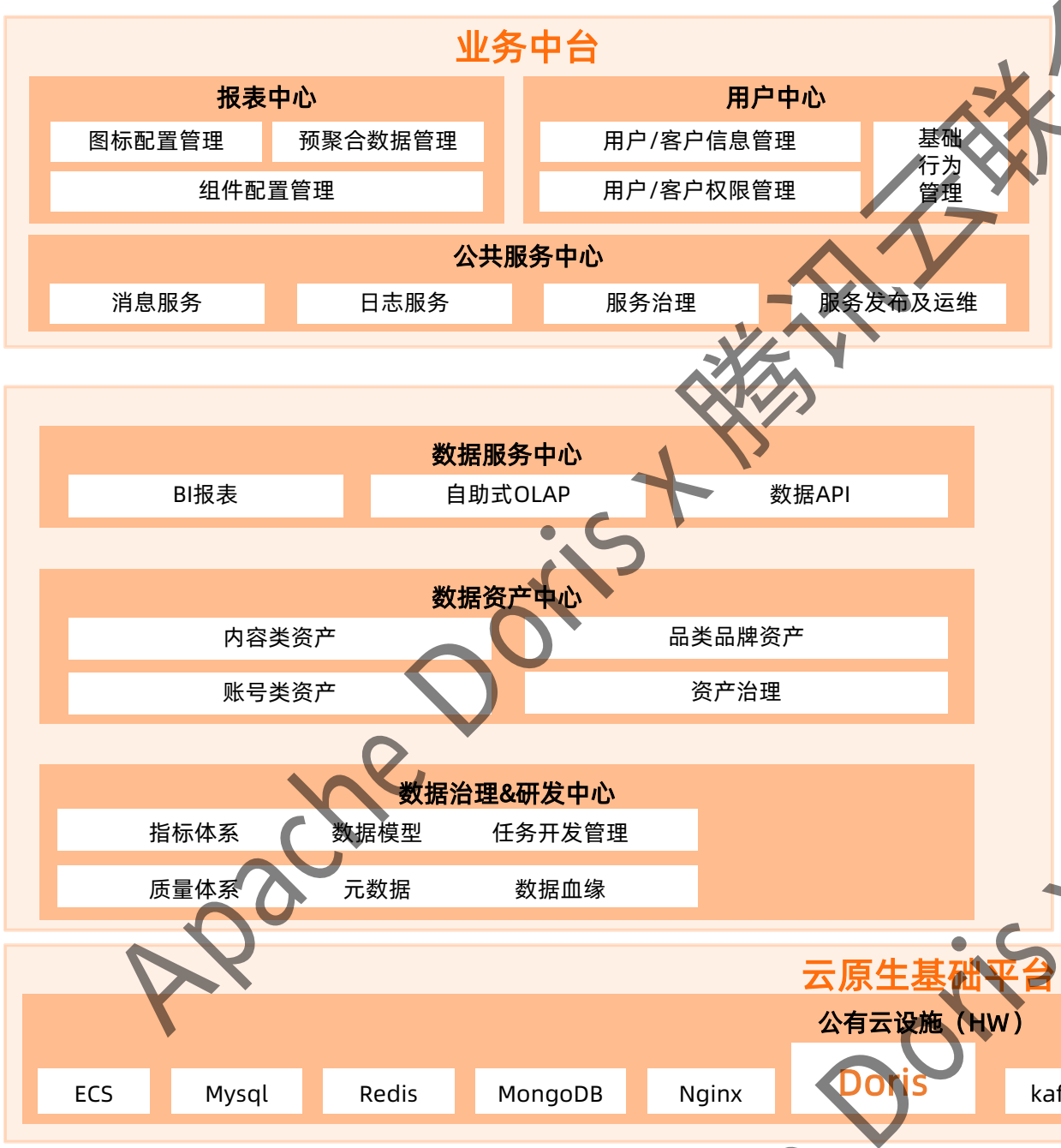
手动warm up cluster 操作&代码优化

使用经验分享

基于 Doris 的方案：一些优化的小技巧与经验之谈

CONTENTS

目录



在未来，结合当下营赛数据当前的业务场景，我们将在数据湖的基础上结合 Doris，进一步完善流批一体架构。

这一战略升级将使我们能够处理更大规模的数据集，提高 OLAP 查询的效率和检索能力。

通过这种方式，不仅能够支持更复杂的数据分析任务，还能确保在各种营销场景下的数据实时性和准确性，从而帮助公司更好地捕捉市场机会和优化客户体验。

Thanks