DTCC 2010-2022

第十三届中国数据库技术大会

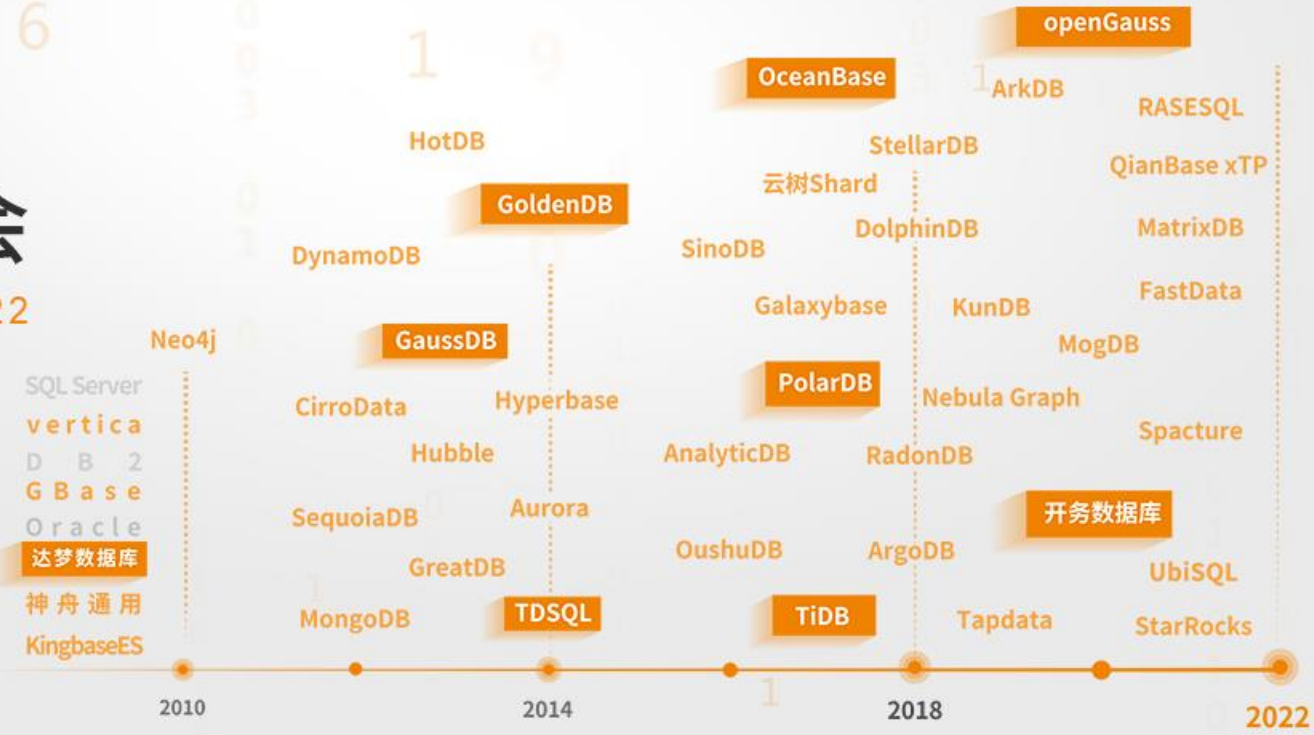
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



Apache Doris
在日志存储与分析场景的实践

肖康 SelectDB 联合创始人

01

Apache Doris 基本介绍

Introduction

关于 Apache Doris

Apache Doris 是一个基于 MPP 架构的高性能、实时的分析型数据库，以极速易用的特点被人们所熟知，仅需亚秒级响应时间即可返回海量数据下的查询结果，不仅可以支持高并发的点查询场景，也能支持高吞吐的复杂分析场景。基于此，Apache Doris 在多维报表、即席查询、用户画像、实时大屏、日志分析、数据湖查询加速等诸多业务领域都能得到很好应用。

Apache Doris 于 2022 年 6 月成功从 Apache 孵化器毕业，正式**成为 Apache 顶级项目**，截止目前 Apache Doris 社区已经聚集了来自不同行业百余家企业的超 400 位贡献者，每月活跃贡献者人数也接近 100 位。

Apache Doris 如今在中国乃至全球范围内都拥有着广泛的用户群体，截止目前， Apache Doris 已经在全球范围**内 1000 家企业的生产环境中得到应用**，在中国市值或估值排行前 50 的互联网公司中，有超过 80% 长期使用 Apache Doris，包括百度、美团、小米、京东、字节跳动、腾讯、快手、网易、微博、新浪、360 等，同时在一些传统行业如金融、能源、制造、电信等领域也有着丰富的应用。

定位：极速易用实时统一的湖仓分析引擎

Data Sources

- Database
- Web Log
- Mobile Log
- Time Series Data
- Data APl

Data Ingestion and Processing

- CDC (Change Data Capture)
- ETL
  - Streaming ETL (Flink)
  - Batch ETL (Spark)
  - ETL Tools (DBT)
- Data Integration Tools

Unified Data Warehouse

- Real-Time Data Warehouse
- Data Lake (Hive / Iceberg / Hudi)

Usage Story

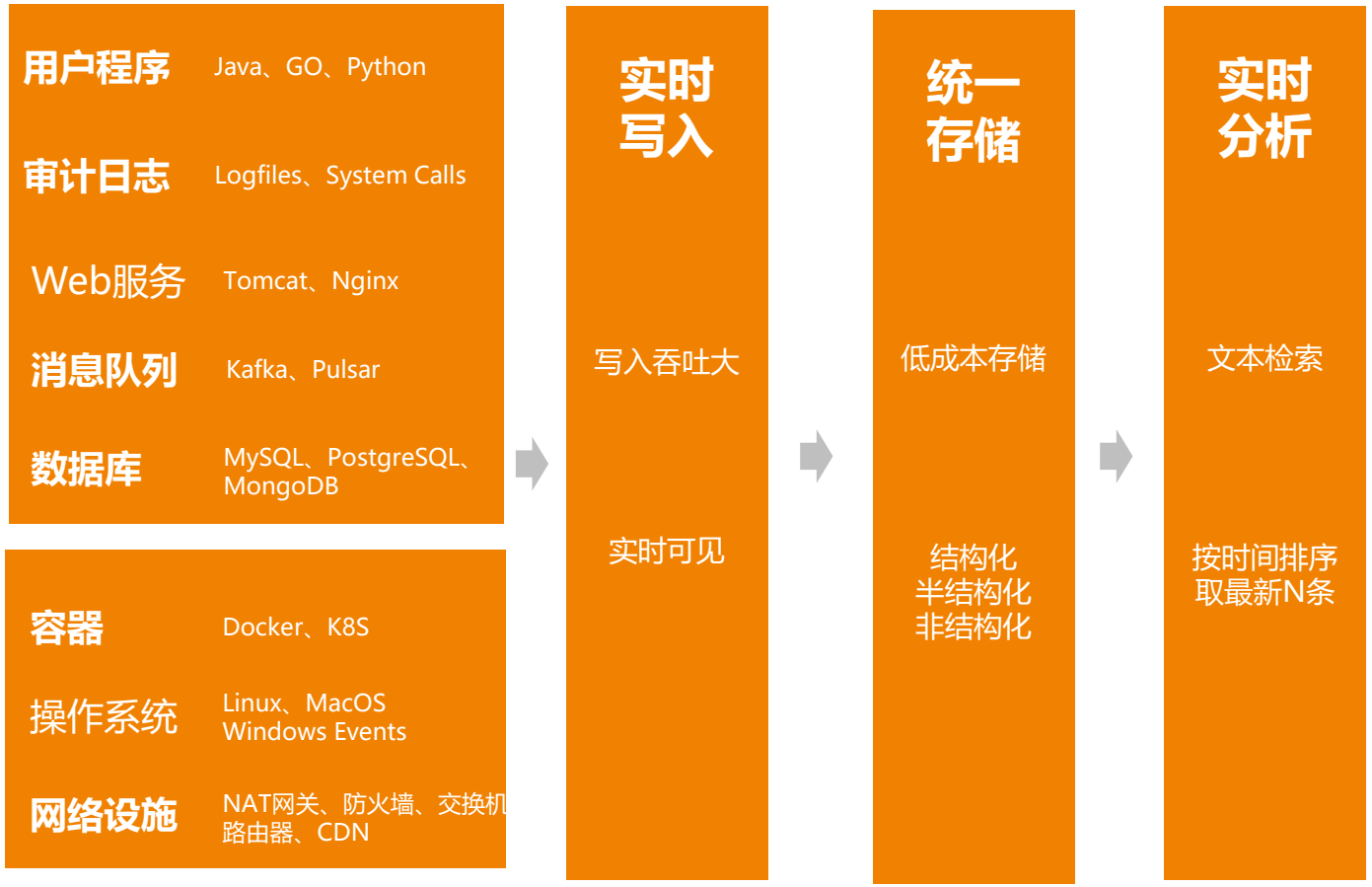
- Report Analysis
- Ad-hoc Analysis
- Federated Query
- Machine Learning

02

典型日志存储与分析场景

Log Storage and Analysis

日志存储与分析场景



日志对于保障系统、业务稳定性至关重要，
常用于故障排查、监控告警等。

特点：

1. 数据写入吞吐量大，还要实时可见
2. 数据存储量大，还要成本低
3. 交互式查询速度快，且支持文本检索、时间排序

典型方案与不足

	ES为代表的 倒排索引 检索架构	Loki为代表的 元数据索引/无索引架构
实时写入吞吐	低（建倒排索引慢）	高
存储规模	中（本地存储）	大（存算分离）
存储成本	高（多份数据存储）	低
semi & free text 交互式查询性能	快	慢（数据无索引，字符串匹配）
总结	优化查询性能 牺牲写入性能和存储空间	优化写入性能和存储空间 牺牲查询性能

典型方案与不足

优化这个牺牲那个，是不是头痛医头脚痛医脚了？

倒排索引是0-1选择吗，它是问题的全部吗？

向量化计算成熟前，是不是都认为OLAP加速要靠预计算？

03 日志场景解决方案

Solution

SelectDB日志场景解决方案

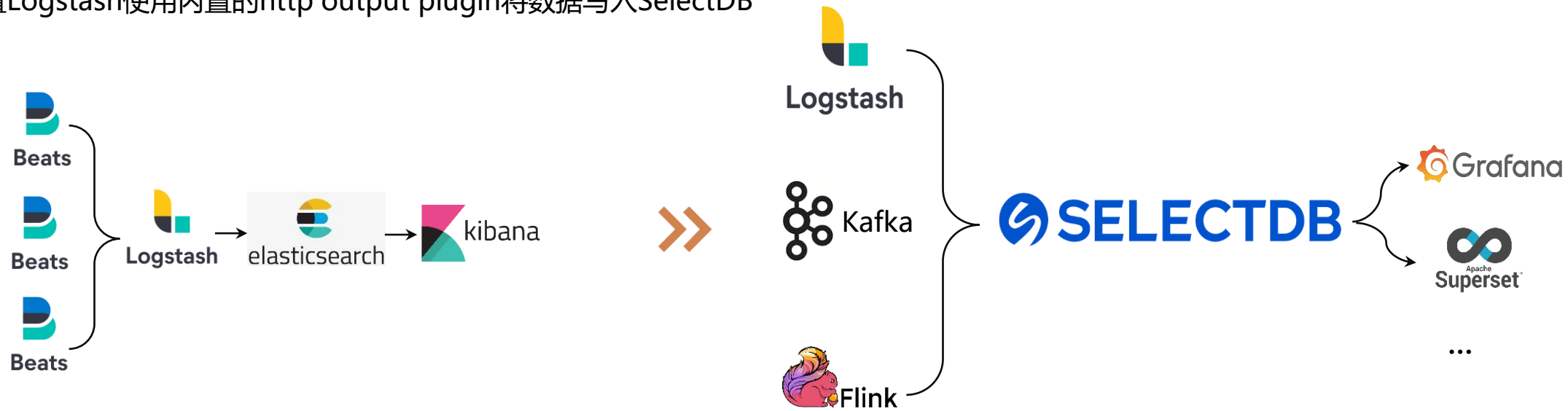
Doris 高性能向量化引擎底座 + SelectDB 存算分离架构 轻量级倒排索引 时序数据管理

	高吞吐 实时写入	低成本 存储	交互式 查询
日志场景增强	服务端攒批	存算分离	轻量倒排索引
	时序compaction	对象存储	dynamic table
Doris核心能力	向量化引擎	列式存储	MPP架构

SelectDB日志场景解决方案

上游写入

配置Logstash使用内置的http output plugin将数据写入SelectDB



下游查询

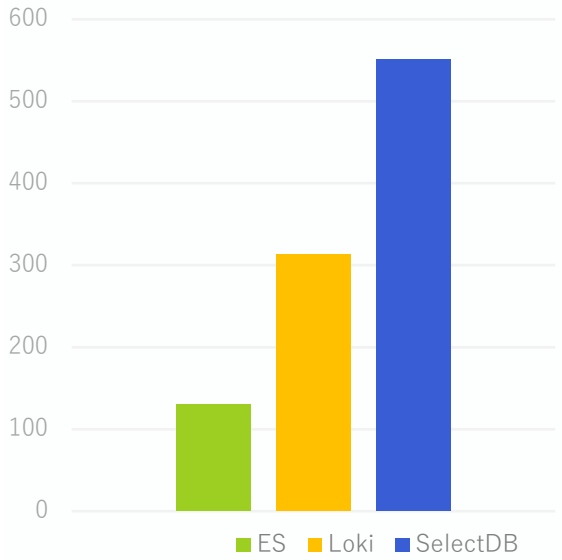
可观测性：Grafana中使用内置MySQL数据源，导入已有模板配置可视化日志看板、检索界面

商业智能：Superset等BI工具通过MySQL协议，即可开箱即用访问SelectDB进行可视化BI分析

性能测试

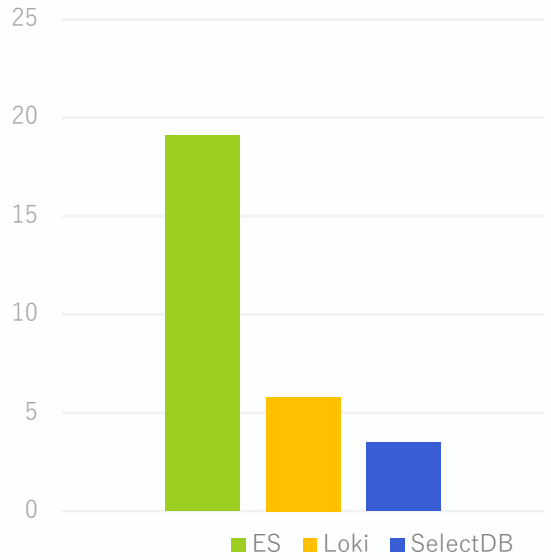
4.2倍

写入速度(MB/s)



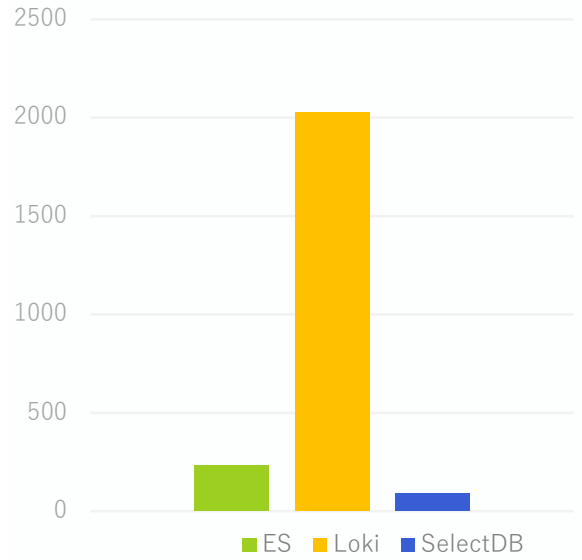
5.4倍

存储空间(GB)



2.3倍

查询时间(秒)



测试说明:

1. 测试环境是3台16c 64g云主机组成的集群
2. 测试数据和测试case来源于ES官方性能benchmark中http_logs，数据总量32GB，2.47亿行
3. 查询时间是ES官方性能benchmark中的11个query，每个串行执行100次的总时间
4. 写入速度越高越好，磁盘空间越低越好，查询时间越低越好

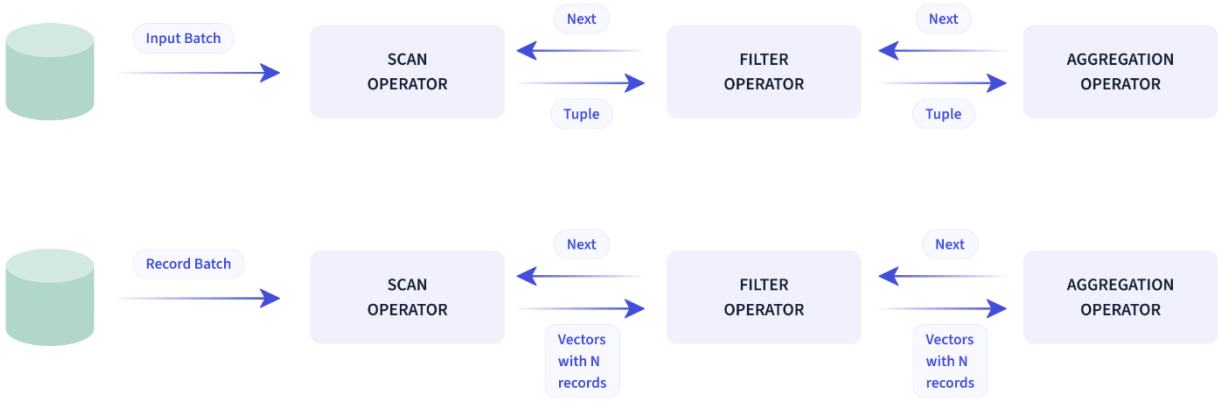
04 关键技术解析

Technology

关键技术：MPP查询与向量化引擎

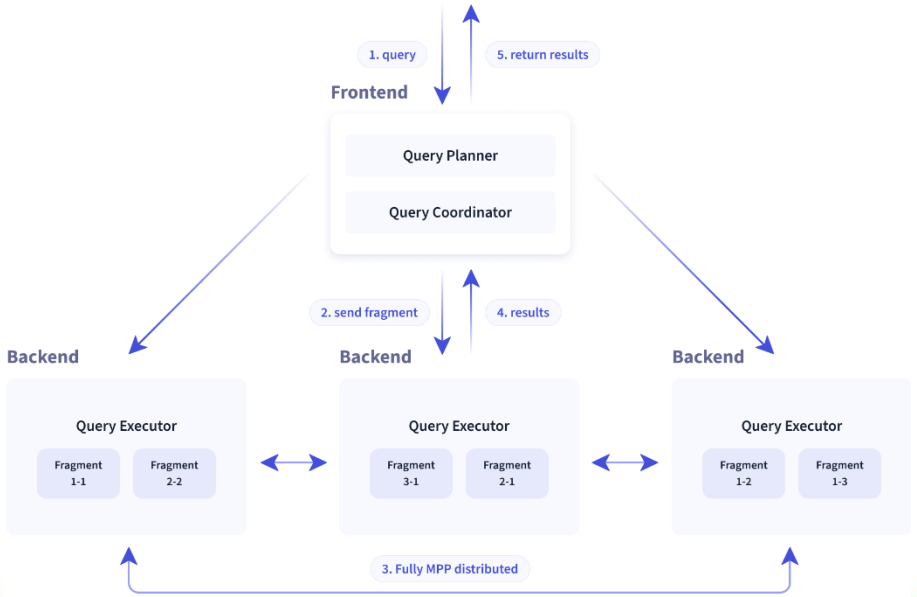
向量化

- 列式内存布局，向量化计算框架
  - ✓ 大幅减少虚函数调用
  - ✓ 大幅提升cache命中率
  - ✓ 高效利用SIMD指令
- 在宽表聚合场景下性能提升5-10倍



MPP查询

- 分布式MPP的查询框架，节点间和节点内都并行执行，大幅提升效率
- 支持大表的shuffle 分布式join



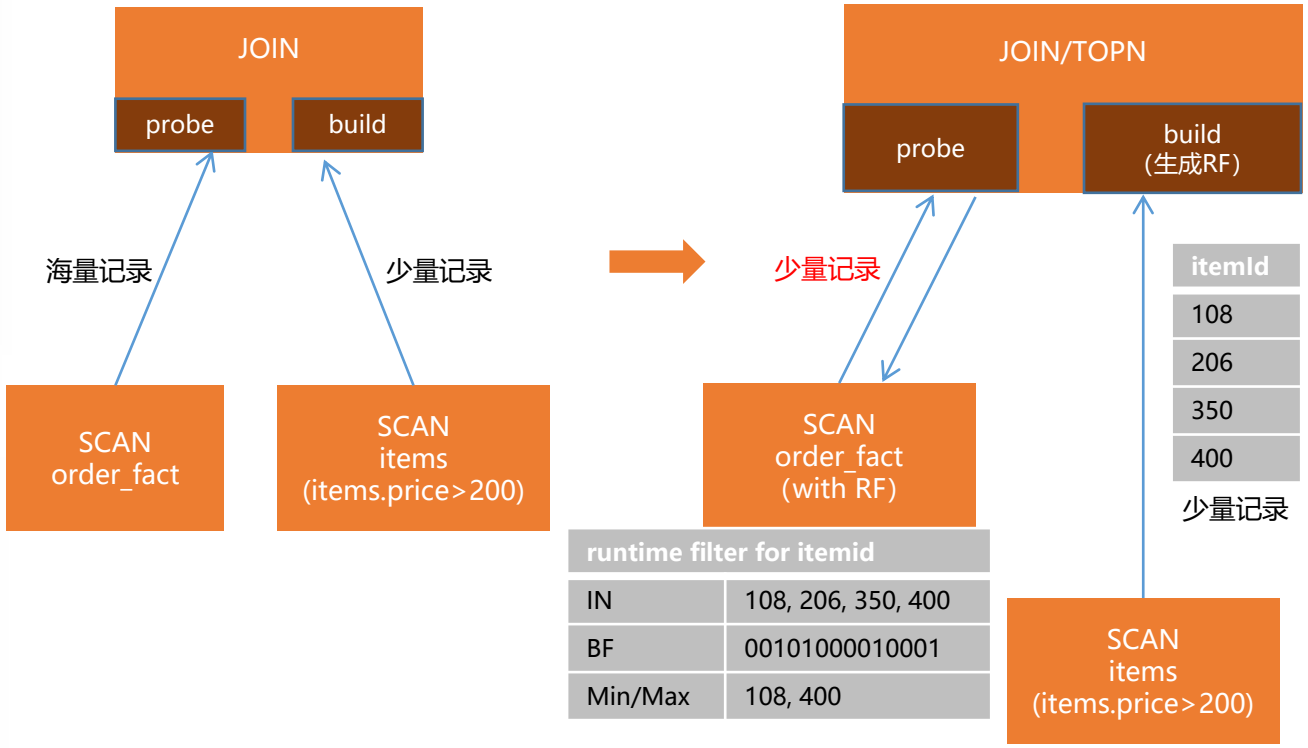
关键技术：多重算子优化与查询优化器

算子优化

- 自适应两阶段聚合算子优化
- JOIN/TOPN runtime filter优化
  - ✓ 为连接列生成filter推到左表
  - ✓ 支持in/min/max/bf等filter
  - ✓ filter自动穿透到最底层
- SSB部分查询依赖RF有2-10倍提升

优化器

- CBO和RBO结合的优化器
- RBO常见规则常量折叠、子查询改写、谓词下推等
- CBO支持Join Reorder
- 新一代智能优化器（Nereids）



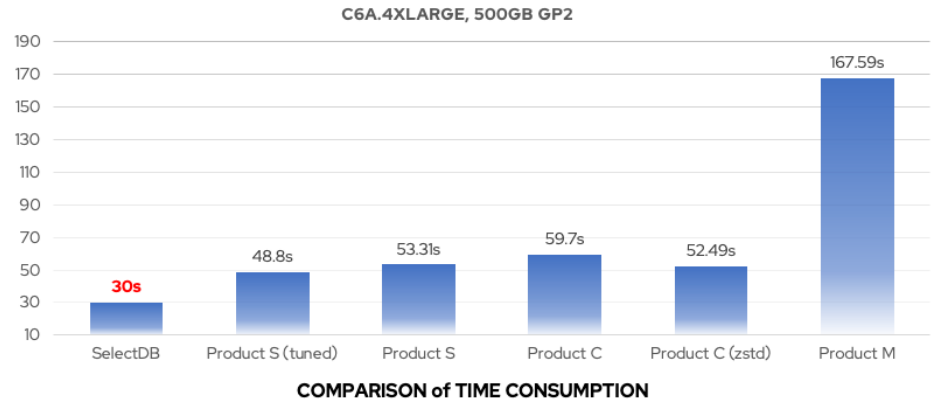
性能世界领先，多项指标登顶ClickBenck

Detailed Comparison

	SelectDB (c6a.4xlarge, 500gb gp2)	ClickHouse (tuned) (c6a.4xlarge, 500gb gp2)	StarRocks (c6a.4xlarge, 500gb gp2)	ClickHouse (c6a.4xlarge, 500gb gp2)	MonetDB (c6a.4xlarge, 500gb gp2)
Load time:	482s (×1.00)	**480s (×1.00)**	737s (×1.53)	484s (×1.01)	939s (×1.96)
Data size:	15.95 GiB (×1.24)	13.58 GiB (×1.06)	16.51 GiB (×1.29)	13.36 GiB (×1.04)	46.28 GiB (×3.61)
Q0.	0.02s (×2.95)	0.00s (×1.08)	0.03s (×3.94)	0.00s (×1.08)	**0.00s (×1.00)**
Q1.	0.03s (×2.00)	0.01s (×1.10)	**0.01s (×1.00)**	0.01s (×1.15)	0.02s (×1.42)
Q2.	0.04s (×1.19)	0.03s (×1.02)	0.06s (×1.67)	**0.03s (×1.00)**	0.03s (×1.07)
Q3.	0.05s (×1.53)	0.04s (×1.38)	0.08s (×2.30)	0.04s (×1.38)	**0.03s (×1.00)**
Q4.	**0.30s (×1.00)**	0.57s (×1.88)	0.47s (×1.55)	1.49s (×4.83)	4.51s (×14.58)
Q5.	**0.62s (×1.00)**	0.70s (×1.13)	0.90s (×1.44)	1.09s (×1.74)	4.27s (×6.79)
Q6.	**0.01s (×1.00)**	0.02s (×1.55)	0.08s (×4.50)	0.02s (×1.65)	0.06s (×3.66)
Q7.	0.03s (×1.67)	0.02s (×1.08)	0.02s (×1.25)	**0.01s (×1.00)**	0.02s (×1.29)
Q8.	**0.51s (×1.00)**	0.96s (×1.86)	0.75s (×1.46)	0.58s (×1.13)	4.41s (×8.50)
Q9.	**0.56s (×1.00)**	0.69s (×1.22)	0.86s (×1.53)	0.65s (×1.16)	6.55s (×11.51)
Q10.	**0.11s (×1.00)**	0.29s (×2.48)	0.12s (×1.08)	0.23s (×2.03)	0.23s (×2.00)
Q11.	0.14s (×1.07)	0.31s (×2.26)	**0.13s (×1.00)**	0.27s (×1.98)	0.27s (×1.97)
Q12.	0.54s (×1.81)	0.76s (×2.52)	0.66s (×2.20)	0.67s (×2.23)	3.07s (×10.14)
Q13.	0.96s (×1.00)	1.07s (×1.11)	1.48s (×1.54)	**0.96s (×1.00)**	7.76s (×8.02)
Q14.	**0.80s (×1.00)**	0.94s (×1.18)	0.91s (×1.14)	0.89s (×1.11)	3.29s (×4.08)
Q15.	**0.34s (×1.00)**	0.76s (×2.21)	0.54s (×1.57)	0.72s (×2.08)	3.46s (×9.93)
Q16.	**1.20s (×1.00)**	2.71s (×2.25)	2.24s (×1.86)	2.53s (×2.10)	9.14s (×7.56)
Q17.	**0.23s (×1.00)**	0.92s (×3.89)	0.47s (×2.00)	1.65s (×6.93)	9.42s (×39.29)
Q18.	**2.18s (×1.00)**	4.54s (×2.08)	4.39s (×2.01)	4.53s (×2.07)	19.61s (×8.96)
Q19.	0.00s (×0.98)	0.02s (×2.66)	0.00s (×0.98)	0.02s (×2.66)	0.00s (×1.05)
Q20.	1.12s (×14.13)	1.05s (×13.31)	1.05s (×13.25)	1.04s (×13.14)	1.57s (×19.75)
Q21.	0.91s (×2.27)	0.94s (×2.35)	0.94s (×2.34)	0.95s (×2.36)	**0.40s (×1.00)**
Q22.	0.99s (×2.47)	1.86s (×4.62)	2.14s (×5.30)	1.84s (×4.56)	**0.40s (×1.00)**
Q23.	0.46s (×1.36)	3.69s (×10.74)	2.65s (×7.71)	3.60s (×10.47)	1.80s (×5.26)
Q24.	0.04s (×3.33)	0.27s (×18.60)	0.23s (×16.00)	0.27s (×18.93)	0.16s (×11.10)
Q25.	0.14s (×1.07)	0.23s (×1.75)	0.16s (×1.21)	0.24s (×1.78)	0.17s (×1.27)
Q26.	**0.03s (×1.00)**	0.27s (×6.98)	0.16s (×4.25)	0.51s (×12.93)	0.17s (×4.53)
Q27.	1.20s (×1.56)	0.81s (×1.05)	1.45s (×1.88)	**0.77s (×1.00)**	3.11s (×4.02)
Q28.	2.87s (×2.23)	**1.28s (×1.00)**	6.21s (×4.81)	17.51s (×13.56)	☠
Q29.	**0.56s (×1.00)**	2.33s (×4.10)	1.38s (×2.44)	2.18s (×3.84)	2.68s (×4.71)
Q30.	**0.39s (×1.00)**	0.53s (×1.35)	0.50s (×1.27)	0.48s (×1.22)	2.45s (×6.14)
Q31.	**0.47s (×1.00)**	0.66s (×1.40)	0.71s (×1.50)	0.65s (×1.38)	3.47s (×7.24)
Q32.	**2.93s (×1.00)**	4.39s (×1.50)	5.13s (×1.75)	4.31s (×1.47)	22.57s (×7.68)
Q33.	4.13s (×8.21)	3.76s (×7.48)	6.82s (×13.55)	3.69s (×7.34)	18.89s (×37.50)
Q34.	4.30s (×1.11)	3.86s (×1.00)	6.82s (×1.76)	**3.86s (×1.00)**	19.05s (×4.92)
Q35.	**0.46s (×1.00)**	1.08s (×2.32)	2.49s (×5.32)	1.05s (×2.26)	12.98s (×27.65)
Q36.	0.06s (×1.40)	0.09s (×1.90)	**0.04s (×1.00)**	0.08s (×1.76)	0.32s (×6.64)
Q37.	0.03s (×1.35)	0.04s (×1.52)	0.03s (×1.35)	0.03s (×1.48)	0.18s (×6.27)
Q38.	0.03s (×1.94)	0.03s (×1.90)	0.02s (×1.46)	0.03s (×1.94)	0.16s (×8.24)
Q39.	0.13s (×1.47)	0.16s (×1.80)	0.14s (×1.57)	0.18s (×2.02)	0.56s (×5.98)
Q40.	0.03s (×2.27)	0.02s (×1.47)	0.02s (×1.70)	0.02s (×1.53)	0.10s (×6.39)
Q41.	0.02s (×1.68)	0.01s (×1.23)	0.01s (×1.12)	0.01s (×1.29)	0.15s (×9.17)
Q42.	0.03s (×2.00)	0.01s (×1.10)	**0.01s (×1.00)**	0.01s (×1.05)	0.10s (×5.28)

近半数 SQL 在所有产品中最佳

通用机型下 SelectDB 性能第一



查询总耗时远低于行业竞品

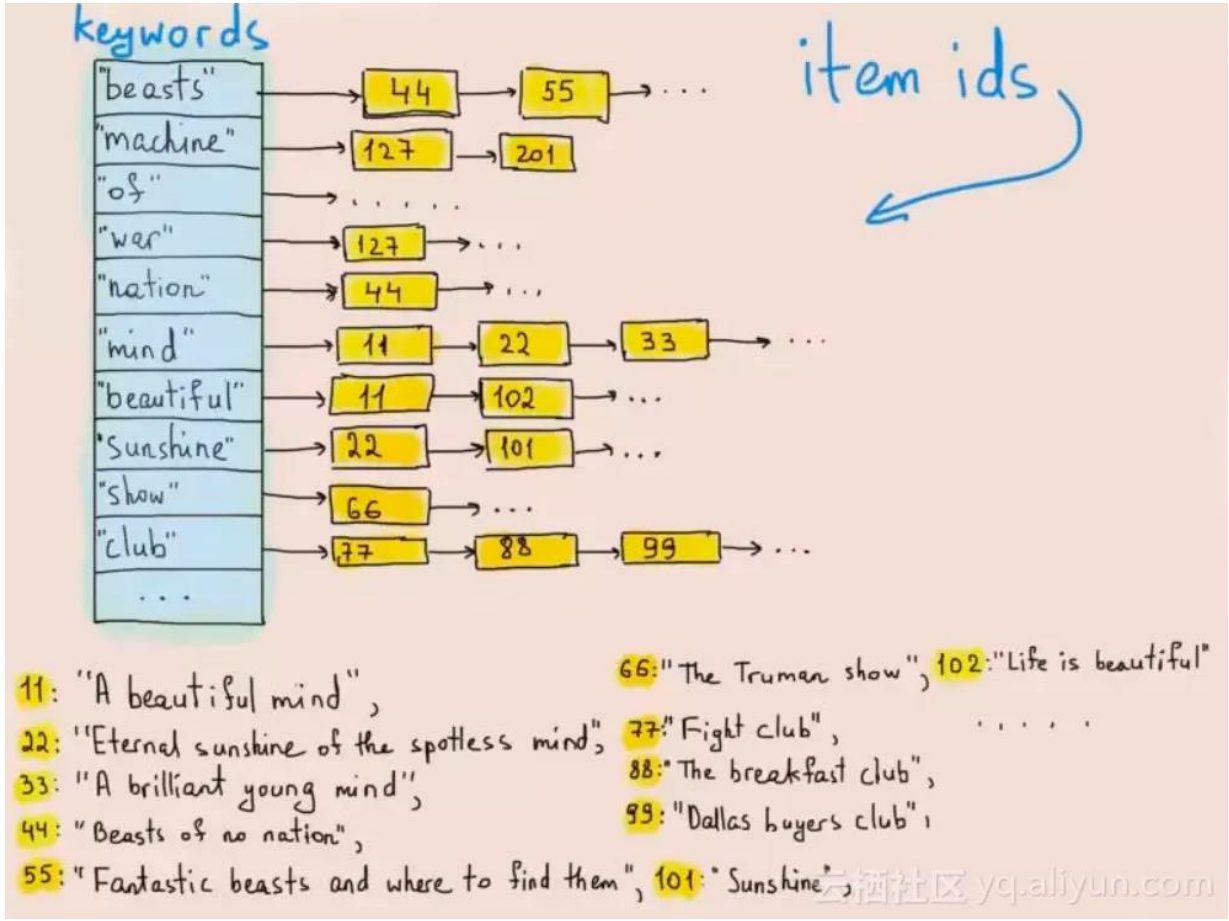
关键技术：轻量级倒排索引

支持快速检索

- 支持文本检索、普通数值/日期查找
- 支持多条件AND OR组合

扩展数据库引擎，内置倒排索引

- 避免了外挂式的跨系统通信、冗余存储



关键技术：轻量级倒排索引

为日志场景精简优化索引结构

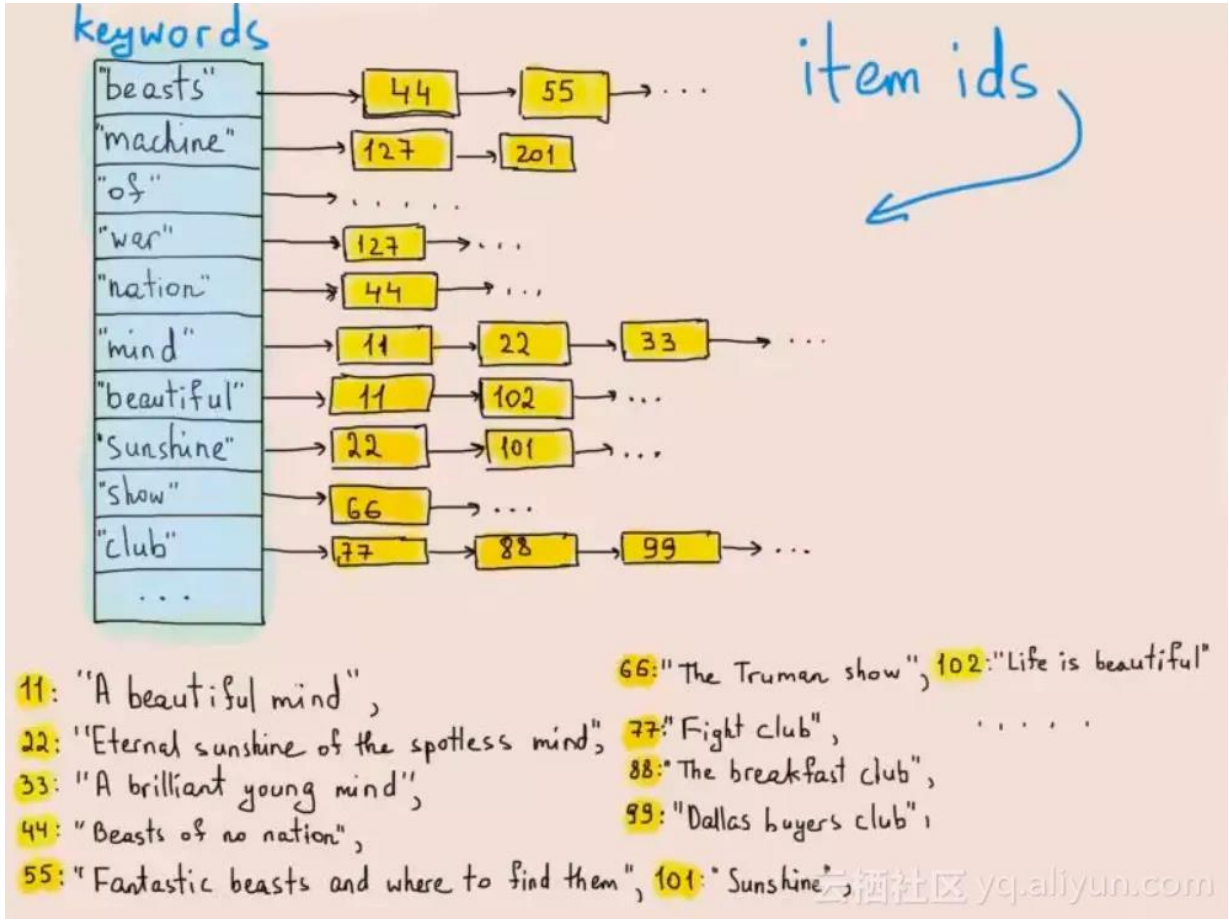
- 简化norm、score等
- bitmap等优化倒排表

列式存储 + ZSTD高效压缩算法

- 列式存储压缩率高
- zstd比gzip快5倍，压缩率更高
- 数据和索引都采用

采用C++和向量化的高性能实现

- 单核吞吐>20MB/s vs ES 5MB/s



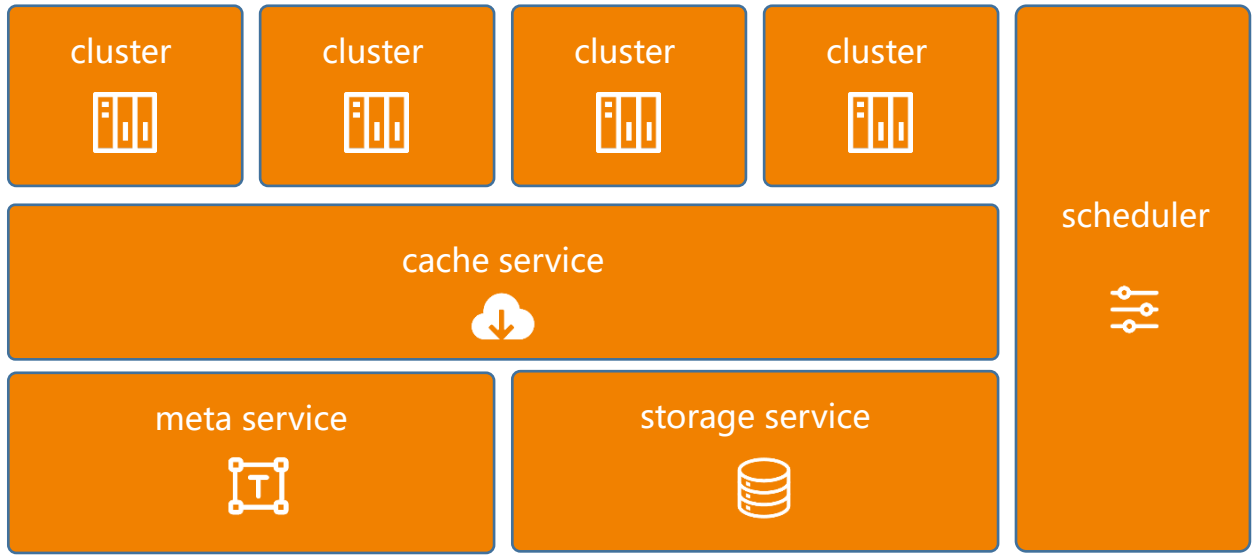
关键技术：存算分离云原生架构

存算分离，以对象存储为主存储

共享缓存，写入即缓存提高性能

弹性扩展，利用云的弹性加速查询

负载隔离，避免业务互相影响



存储成本大幅降低

存储成本 = 存储空间 x 单位成本

存储空间：
- 原始数据行存 => 列式存储
- 复杂倒排索引 => 专为日志简化索引
- 普通压缩算法 => 日志高压缩比

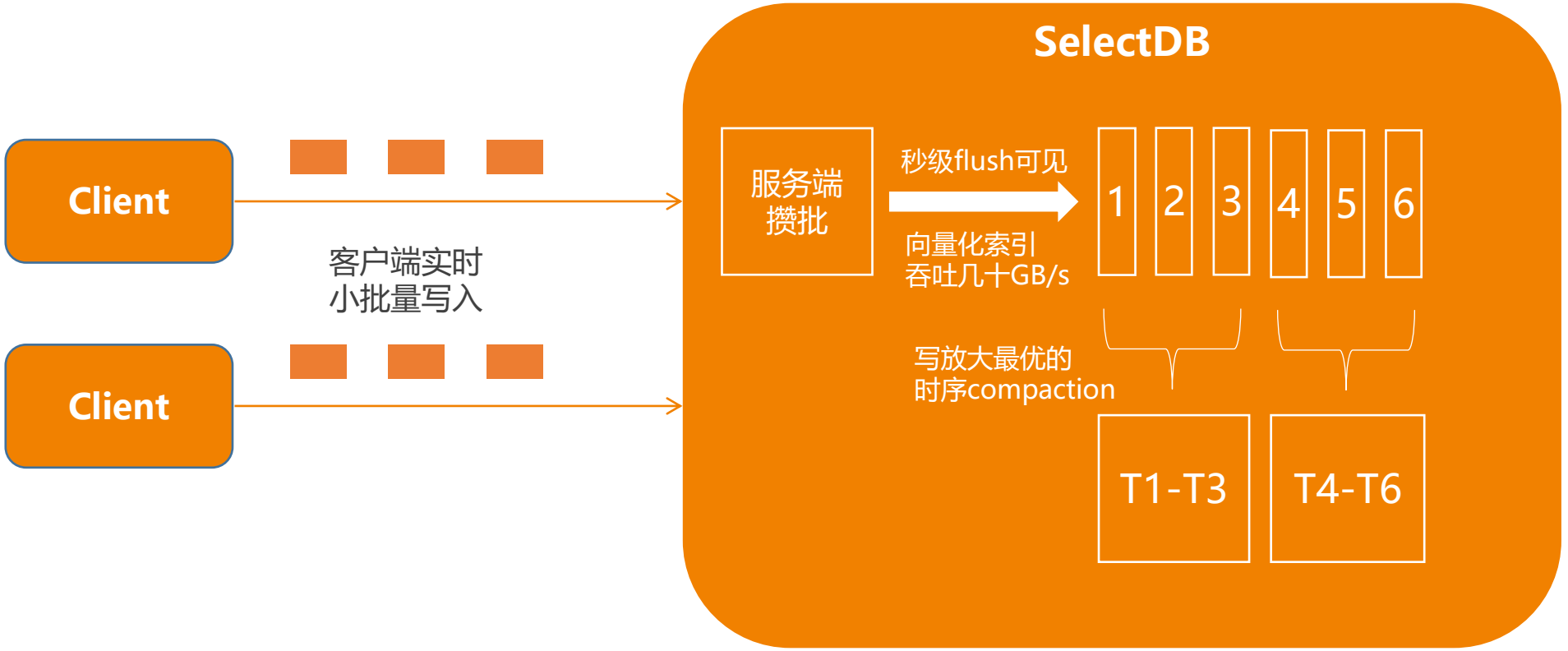
单位成本：
- 冷热数据全本地盘 => 冷热分离
- 存储计算同步扩容 => 存储分离独立扩容

降至 1/15

降至 1/5

降至 1/3

关键技术：高吞吐实时写入



关键技术：快速交互式查询

```
SELECT * FROM log
WHERE ts >= t1 AND ts <=t2 AND message MATCH 'error'
ORDER BY ts DESC LIMIT 100
```

挑战：从海量日志中全文检索关键词

基于分区、主键的时间范围快速跳过

基于倒排索引的全文检索精确定位

挑战：从时间排序取满足条件的最新N条日志

按时间排序的时序存储模型

动态剪枝的TopN查询算法

百亿日志检索秒级响应

05

关于我们

Contact US

飞轮科技：专注于开源技术创新的云原生实时数据库厂商

开源数据仓库技术创新

秉持开源开放的核心理念，大力投入研发力量，加强Apache Doris 在数据分析技术上的持续创新力，使其成为世界领先的开源分析数据库。

云端数据仓库商业服务

基于 Apache Doris，构建运行于多云之上的新一代云原生实时数仓 SelectDB，为客户提供极简运维和极致性价比的数仓服务。

联系我们

欢迎关注SelectDB微信公众号

公司邮箱：support@selectdb.com

SelectDB 官网：www.selectdb.com

Apache Doris 官网：https://doris.apache.org/

Apache Doris GitHub：https://github.com/apache/doris


openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云树Shard
GoldenDB
DolphinDB
MatrixDB
DynamoDB
SinoDB
FastData
Galaxybase
KunDB
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
开务数据库
UbiSQL
神舟通用
TDSQL
TiDB
Tapdata
StarRocks
KingbaseES
2010
2014
2018

THANKS