SACC

2022 中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2022

激发架构性能 点亮业务活力

云上会议 网络直播 | 2022年10月20-22日

ITT168.com ChinaUnix.net ITPUB

从 Apache Doris 到 SelectDB
下一代云原生实时数仓的演进之路

SelectDB 联合创始人& 产品VP 杨勇强

个人介绍

杨勇强

Apache Doris Committer

SelectDB 产品 VP

曾任职百度智能云存储部主任架构师

具有 10 年云存储产品与架构经验

本科毕业于南开大学 计算机科学与技术专业

硕士毕业于中国科学院计算技术研究所

目录
CONTENTS

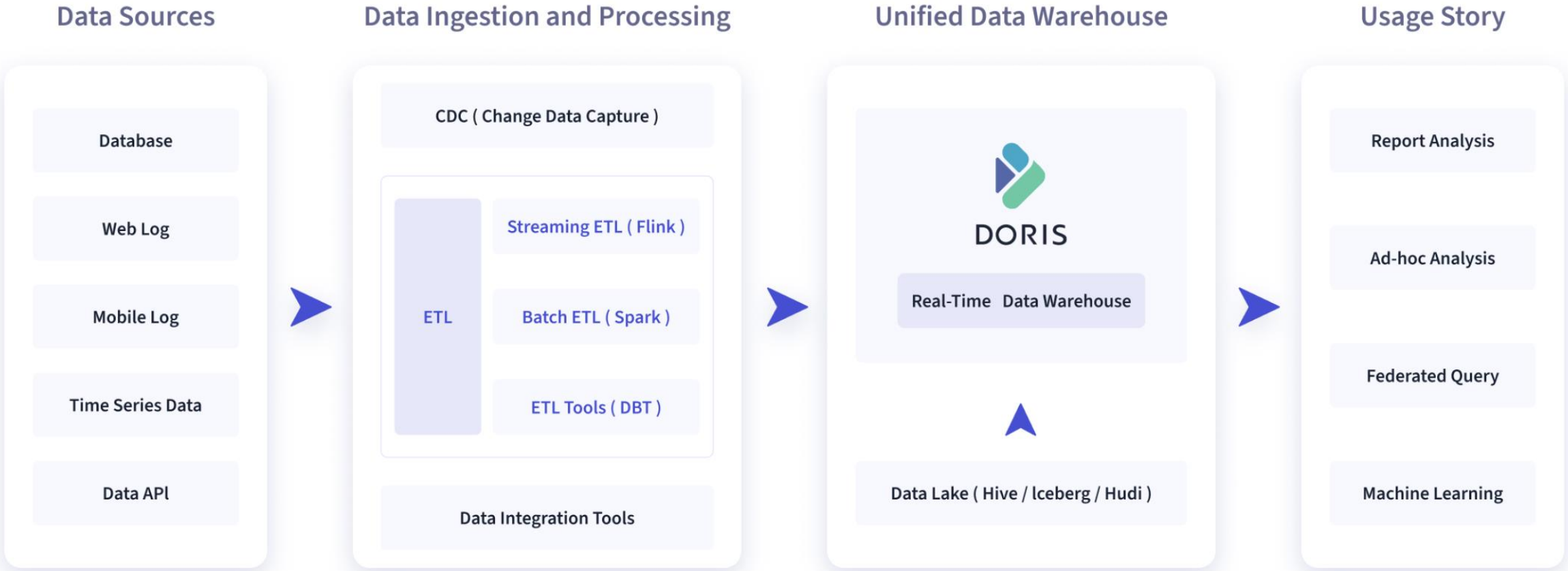
Apache Doris 介绍

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近百家企业的**近 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 **已经在全球接近 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。

易用、极速、实时、统一的湖仓分析引擎



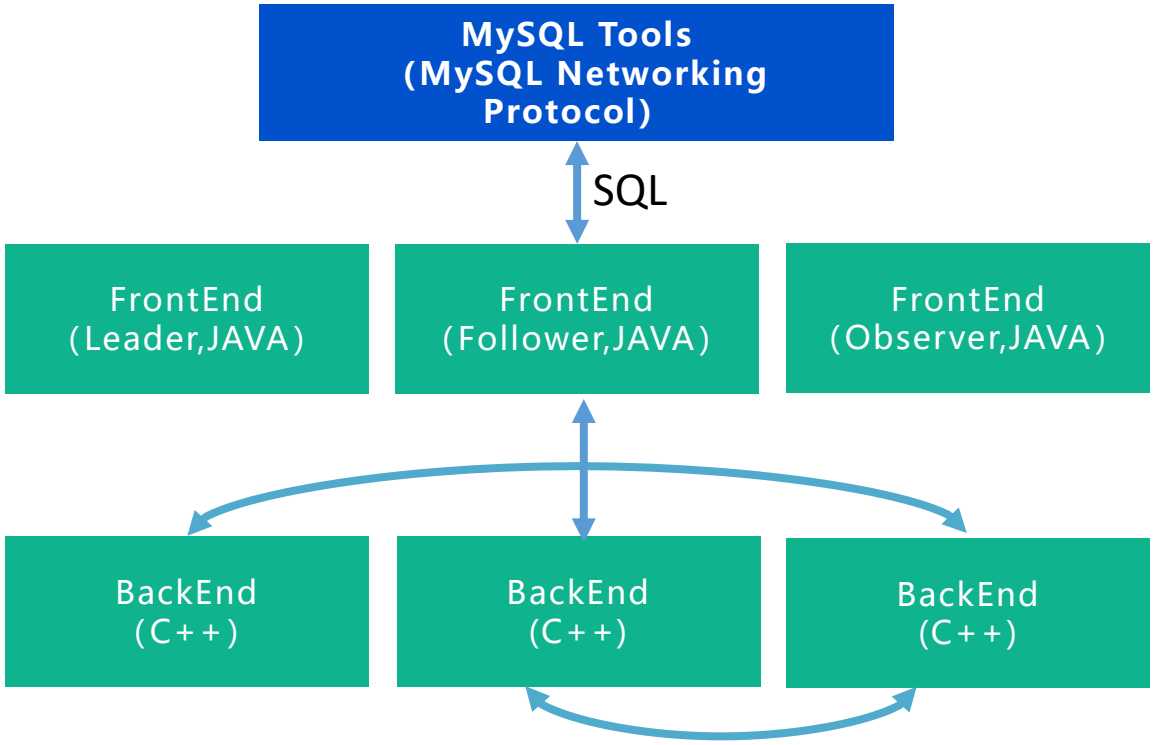
极简易用

业务侧

- 采用MySQL协议和标准SQL，各类客户端工具和BI可以无缝对接
- 支持多表大表join，不同场景有多种优化方案
- 支持子查询、窗口函数、udf/udaf，Grouping Set等高级功能
- 丰富的连接器，flink kafka等生态无缝对接，支持离线高效批量导入以及流式实时导入

运维侧

- 架构简洁，部署简单，只有FE和BE两类进程，无外部依赖
- 在线扩缩容，自动副本恢复
- 在线schema change



系统架构

极致性能（存储引擎）

列式存储

- 数据按照列存储，编码压缩高效
- 丰富的索引结构，减少数据扫描
  - ✓ sorted short key
  - ✓ min/max（等值、范围查询过滤）
  - ✓ bloom filter（高基数列等值过滤）
  - ✓ invert index（快速检索）

场景优化的存储模型

- aggregate key模型：相同key列value列合并，通过提前聚合大幅提升性能
- unique key模型：key唯一，相同key的数据覆盖，实现行级别数据更新
- duplicate key模型：明细模型
- 强一致物化视图，智能选择，加速查询

aggregate key模型

基础数据

时间	门店	销售额
2022/02/14	1	2000
2022/02/15	2	1500
2022/02/15	3	3000

新增数据

时间	门店	销售额
2022/02/14	1	200
2022/02/15	2	100
2022/02/15	4	4000

时间	门店	销售额
2022/02/14	1	2200
2022/02/15	2	1600
2022/02/15	3	3000
2022/02/15	4	4000

unique key模型

基础数据

订单	时间	状态
1	2022/02/14-22	待支付
2	2022/02/14-22	支付完

新增数据

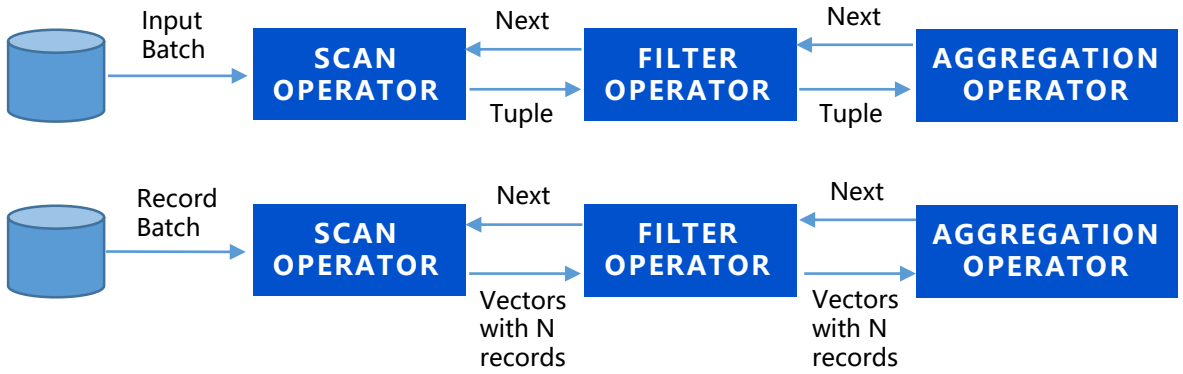
订单	时间	状态
1	2022/02/14-23	支付完

订单	时间	状态
1	2022/02/14-23	支付完
2	2022/02/14-22	支付完

极致性能（查询引擎）

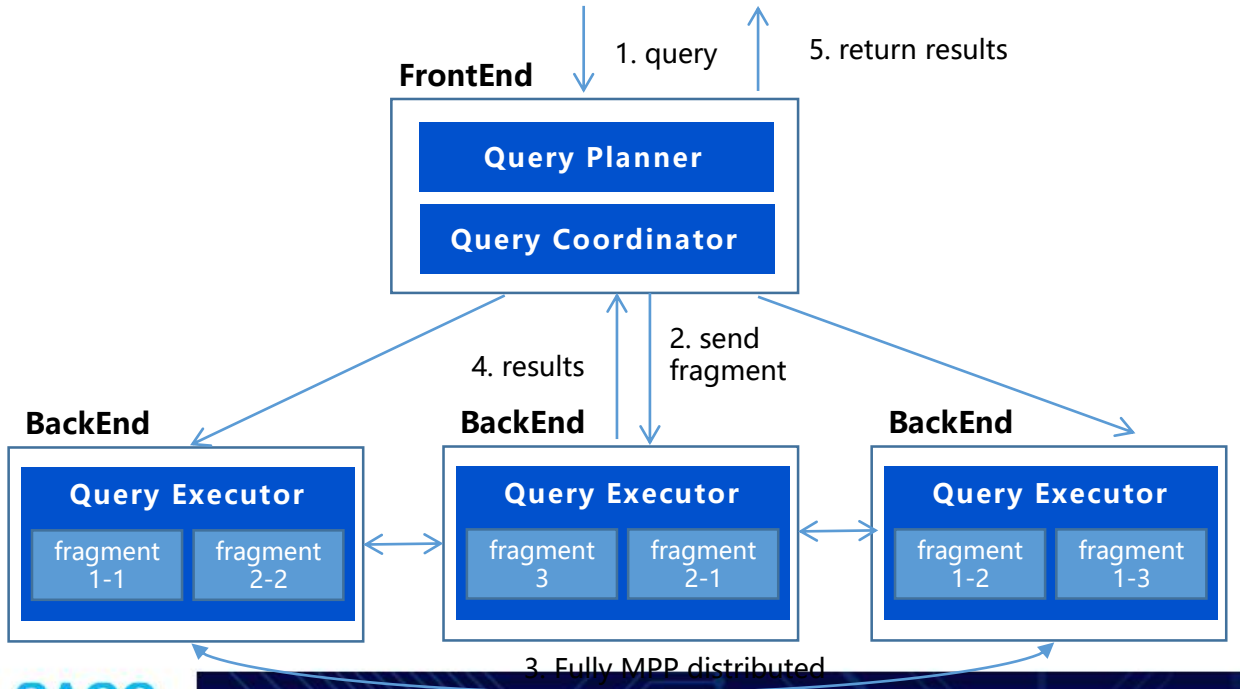
向量化

- 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

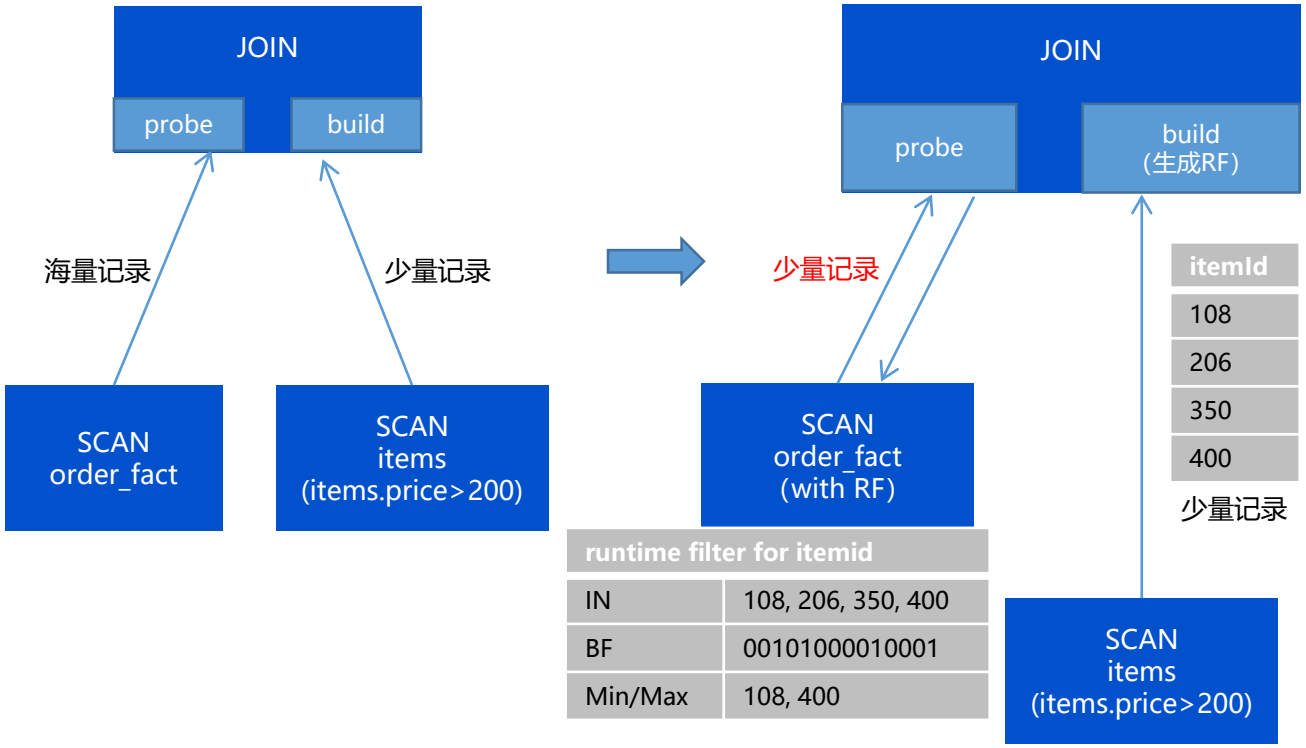
- 分布式MPP的查询框架，节点间和节点内都并行执行，大幅提升效率
- 支持大表的shuffle 分布式join



极致性能（查询优化器）

算子优化

- 自适应两阶段聚合算子优化
- Join的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



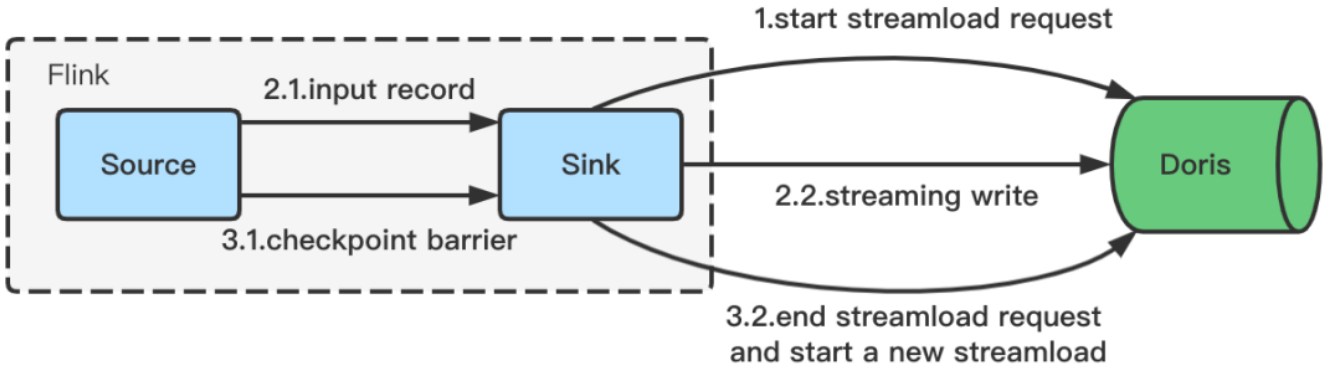
优化器

- 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）

实时性保证

导入

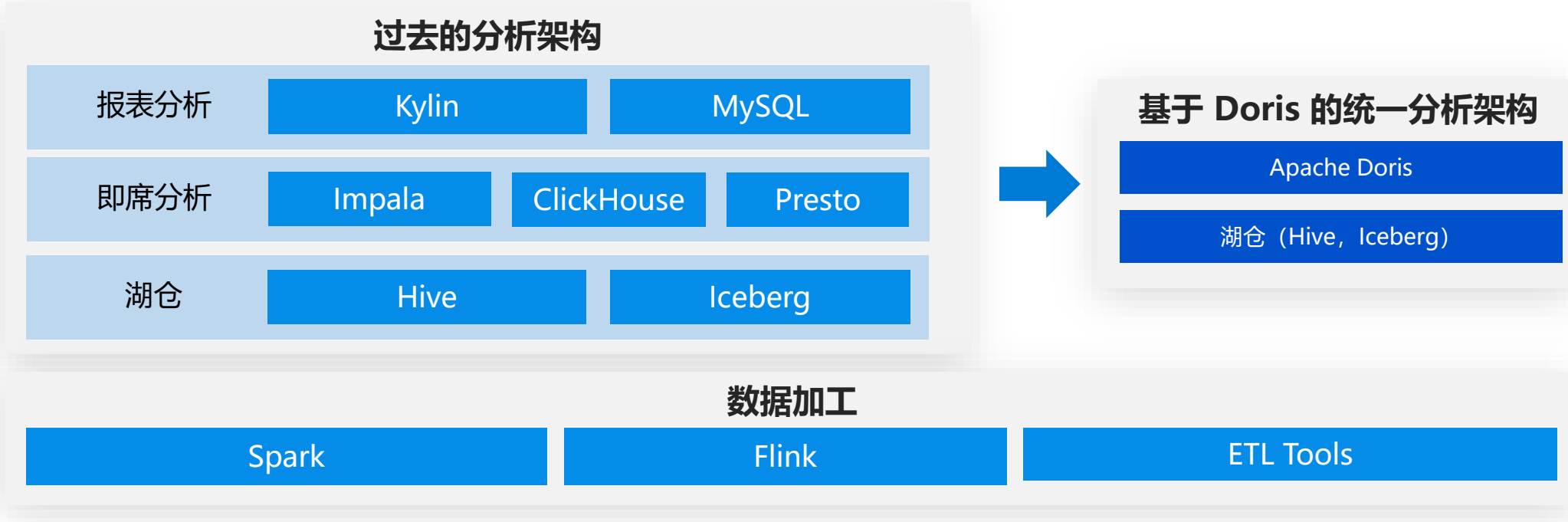
- 两阶段提交，数据不重不丢
- 高并发微批，数据快速聚合
- flink / kafka 无缝集成
- online schema change



更新删除

- 强一致
- 条件更新（sequence column）
- 无冲突删除（batch delete）

统一分析架构



统一 OLAP 分析引擎，从过去每个场景使用不同组件，到最终使用 Apache Doris 构建统一的分析架构，降低复杂架构带来的运维压力与技术融合承办。

使用 Apache Doris 作为Hive和Iceberg的**湖仓查询加速引擎**，性能相比Presto有 **3倍** 提升，相比Hive有 **数10倍** 以上提升。

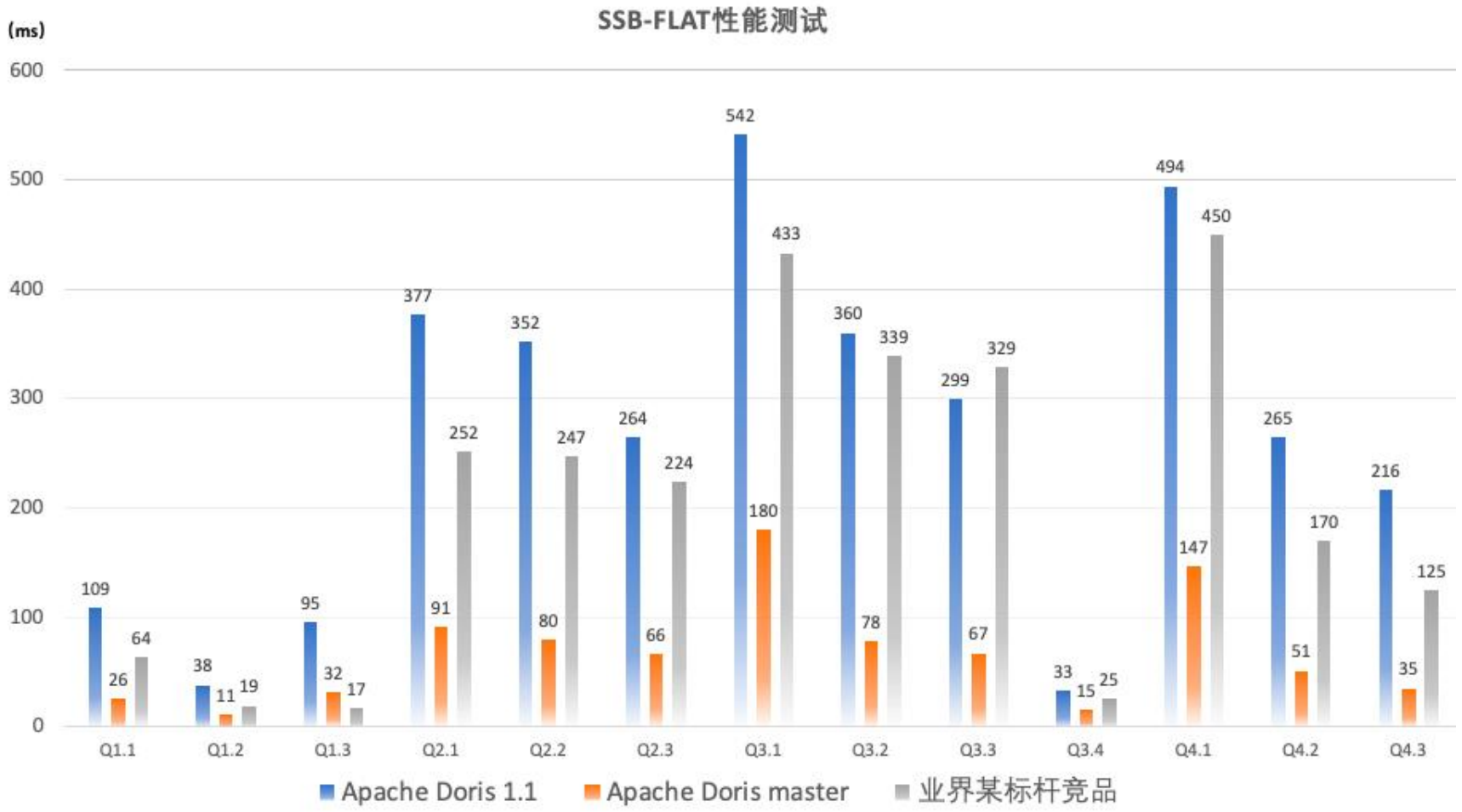
Apache Doris 更多新功能

■ 1.2 新版本功能

- **极速数据湖分析**：增加Multi Catalog，无缝对接多种数据源，性能超 Trino **3倍**
- **Light Schema Change**：毫秒级在线元数据更新，结合 Flink CDC 可实现 DDL 实时同步
- **支持实时更新的主键模型**：全新 Merge on Write 主键模型，实时更新场景性能提升 **10倍** 以上
- **JDBC数据源**：提供便捷的外部数据源访问方式
- **Array 类型**：支持嵌套、支持行列转换
- **New Decimal**：更大精度、更高计算效率、更准确的精度信息
- **New Date/Datetime**：更高计算效率、支持微妙存储
- **Java UDF**：兼容 Hive UDF
- ……

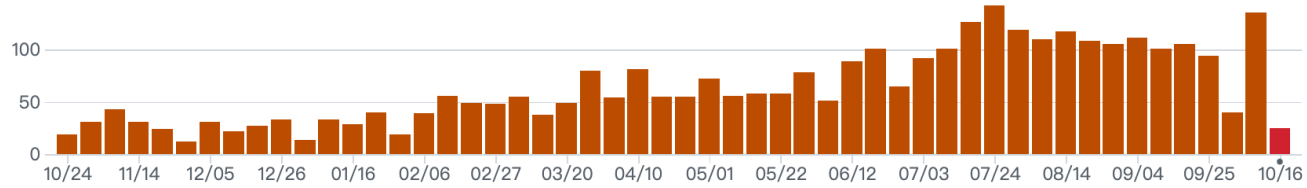
■ 发布时间：2022 年 10 月底

Apache Doris 性能提升

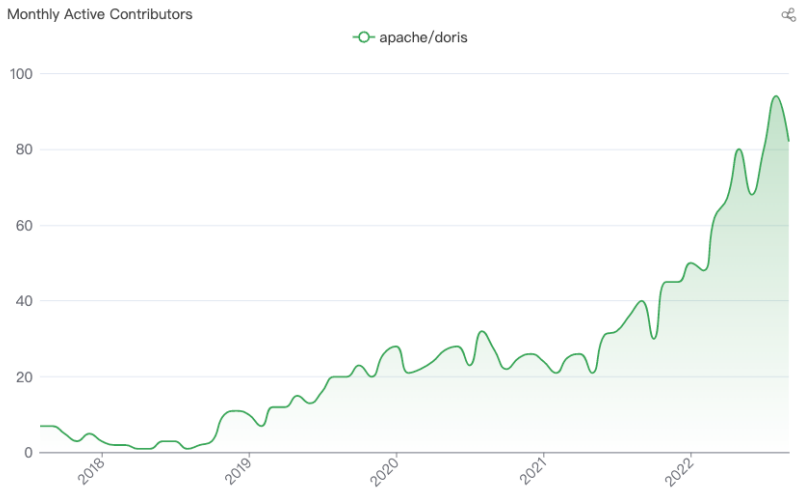


最新 1.2 版本较 1.1 版本提升近 4 倍，是业内标杆竞品 3 倍以上

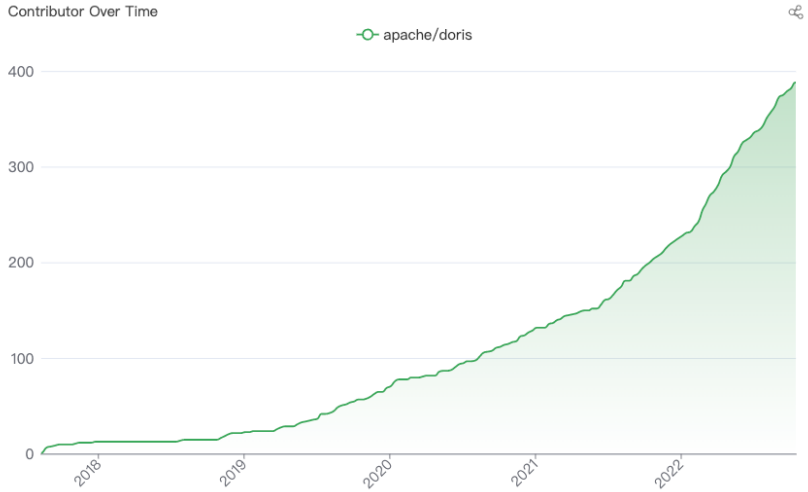
Apache Doris 社区发展 – 期待更多开发者的加入



每周超过100个Commits



活跃贡献者



累计贡献者

目录
CONTENTS

大数据技术发展

- 宏观经营分析、业务决策
- 固定报表、批处理报告、BI

- 互联网业务大规模、多场景的数据计算和分析
- 批量计算、流式计算、报表和自助式分析

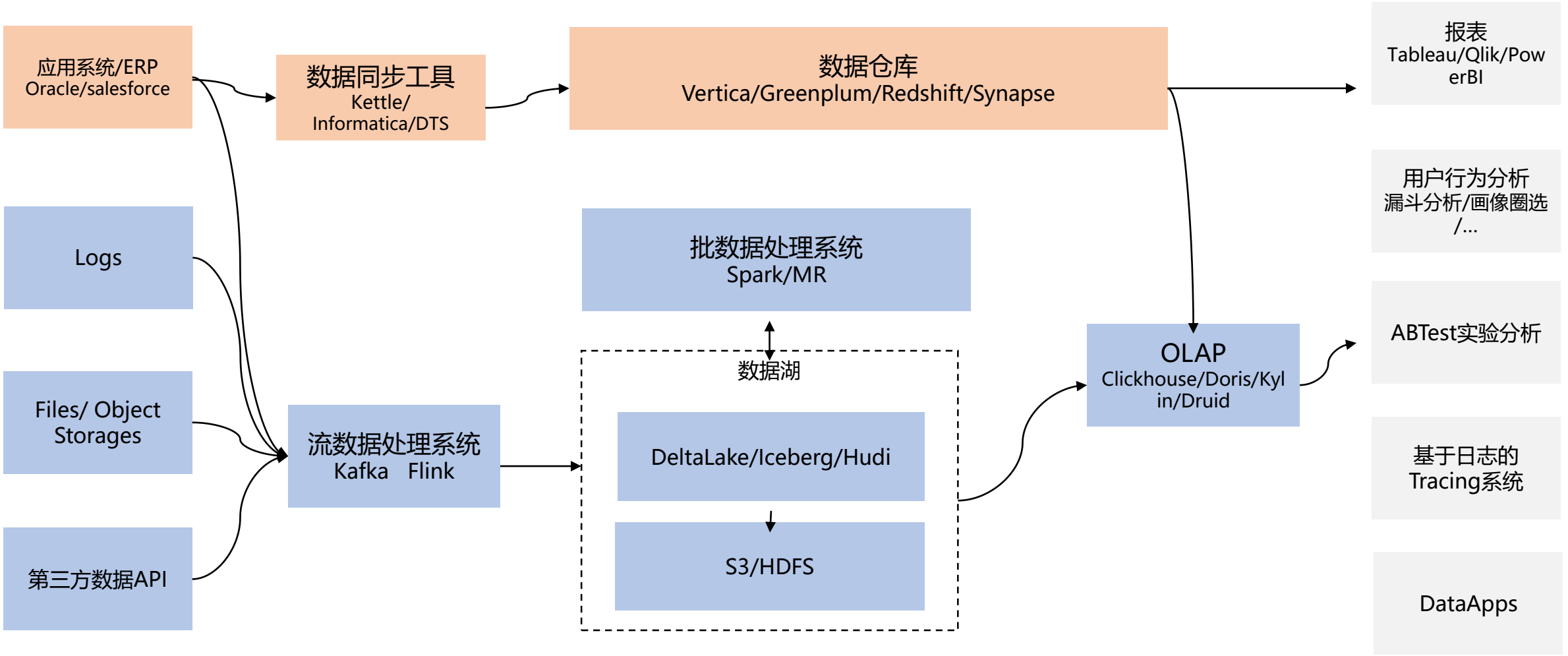
- 数据共享与联动分析、业务的数据驱动与敏捷创新
- BI应用、实时分析、湖仓联邦查询、云服务化

- Teradata
- Greenplum
- Vertica

- Hadoop
- Spark/Flink
- Hive
- Presto/Impala
- Doris/Clickhouse

- Snowflake
- Databricks
- AWS Redshift
- SelectDB

当前数据分析技术栈


应用系统/ERP
Oracle/salesforce
数据同步工具
Kettle/
Informatica/DTS
数据仓库
Vertica/Greenplum/Redshift/Synapse
报表
Tableau/Qlik/PowerBI
Logs
批数据处理系统
Spark/MR
用户行为分析
漏斗分析/画像圈选
/...
Files/ Object
Storages
数据湖
OLAP
Clickhouse/Doris/Kylin/Druid
ABTest实验分析
流数据处理系统
Kafka Flink
DeltaLake/Iceberg/Hudi
基于日志的
Tracing系统
第三方数据API
S3/HDFS
DataApps

云原生时代数据分析需求

日益增长的实时数据分析诉求

- Increasing Real-time Requirement -

复杂性高	性价比低	灵活性差	开放性弱
■ 软件维护的复杂性 ■ 使用的复杂性 ■ 业务系统维护的复杂性	■ 从on-premise到cloud-native 需要新设计； ■ 新硬件的发展突飞猛进，需要适配；	■ 分析的数据类型与分析work-load的变化没有被很好满足； ■ 对半结构化和非结构化数据没有native高效支持 ■ 数据科学、机器学习等深度分析场景满足	■ 客户出于防止锁定、容灾和降本的多云诉求 ■ 客户期待有第三方厂商推出的多云中立的，统一接口和体验的，开源开放的PaaS层产品

用户满意度痛点

云原生时代数据分析技术

性能 - 实时 realtime

- 数据服务实时化（HTAP、HSAP）
- 数据分析实时化（向量化、代码生成、CBO）
- 数据处理实时化（流计算Flink、表格式Iceberg）

功能 - 统一 unified

- 湖仓一体
- 在离线一体
- 流批一体
- 结构化、半结构化、日志等多数据类型支持

开源技术

全新数据分析基础设施

云端服务

云原生 cloud-native

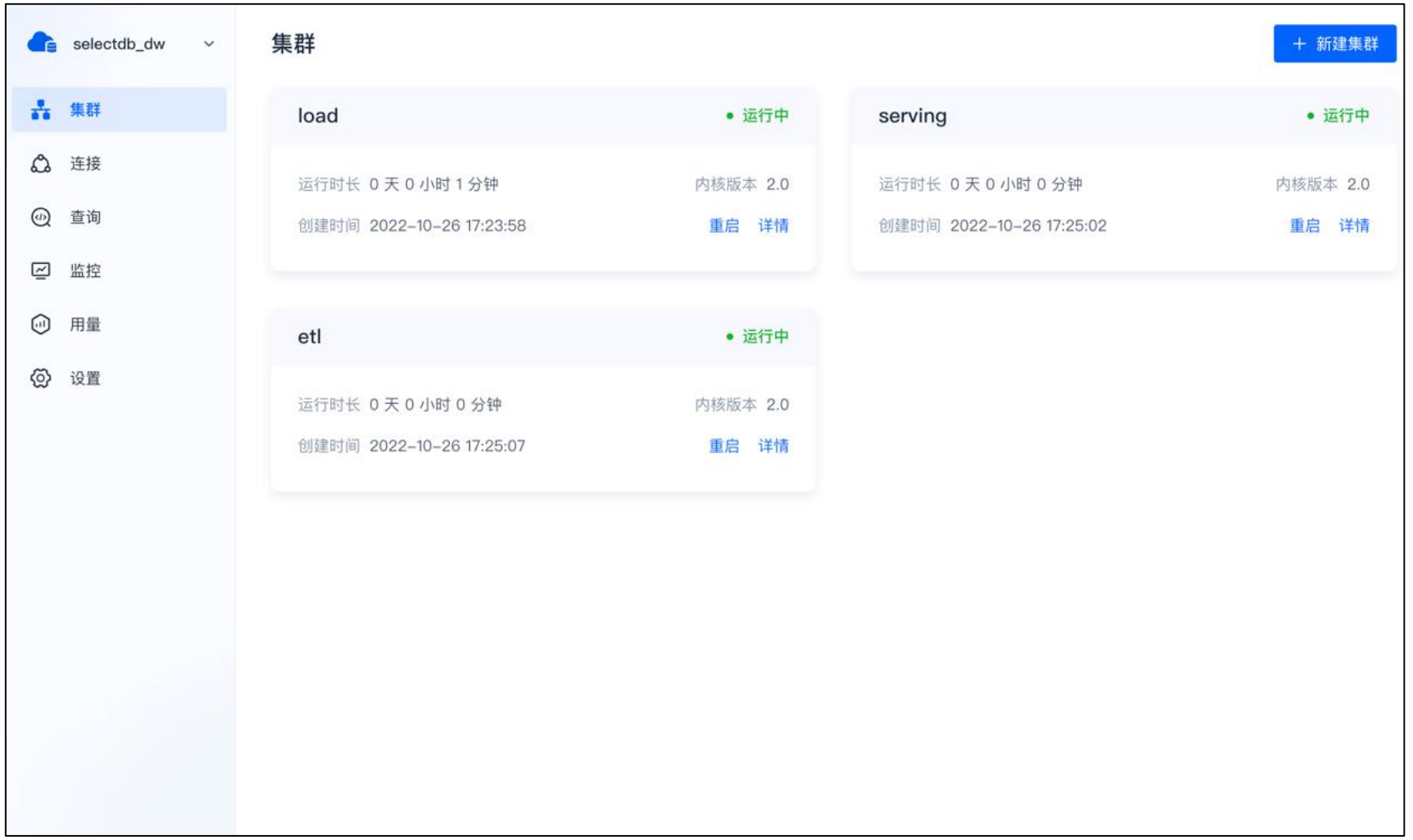
- 存算分离、弹性使用
- 计算隔离，多租户
- 极致性价比
- 极简使用和运维

多云 multi-cloud

- 多云中立
- 统一接口
- 多云复制

目录
CONTENTS

SELECTDB Cloud 概览



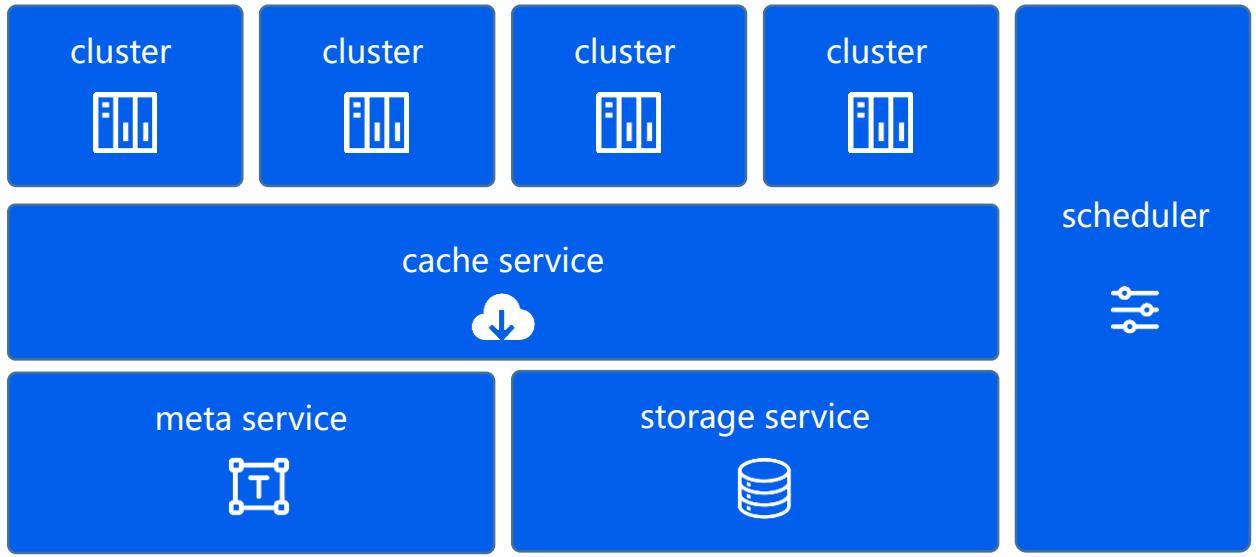
SELECTDB Core 架构

存算分离，以对象存储为主存储

共享缓存，确保在线高可用

timetravel 与 snapshot

数据共享与跨区域复制



SELECTDB 核心功能

极致性能

· 高效列式存储与索引
· 现代 MPP 计算架构
· 适配 X64、ARM64 的向量化执行引擎

融合统一

· 实时/交互/批量数据处理
· 结构化/半结构化数据
· 对数据湖和其他数据库进行联合查询

简单易用

· 兼容 MySQL 连接协议
· 强大的管理控制台
· Spark/Flink/DBT/Kafka 等丰富的连接器

高性价比

· 存算分离
· 按需自动扩缩容
· 冷热数据分级存储

开源开放

· 基于开源 Apache Doris
· 与 Doris 数据自由迁移
· 全云可用

企业特性

· 用户管理与访问控制
· 数据保护与备份
· 数据治理促进数据共享使用

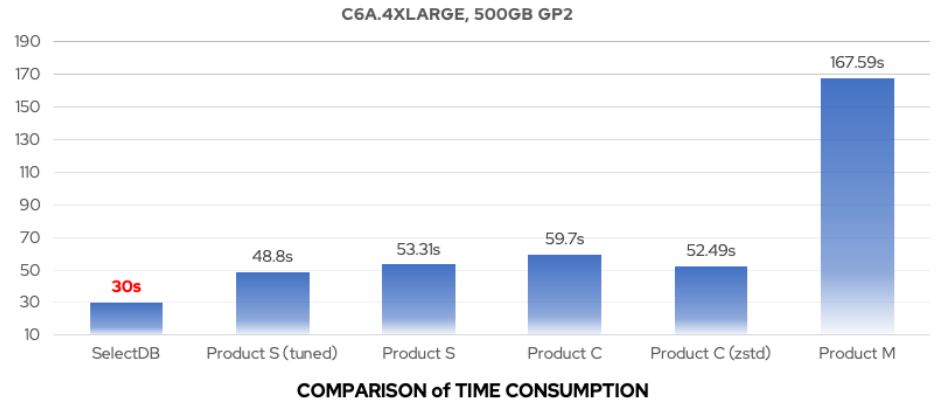
性能业界领先，多项指标登顶ClickBench

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)

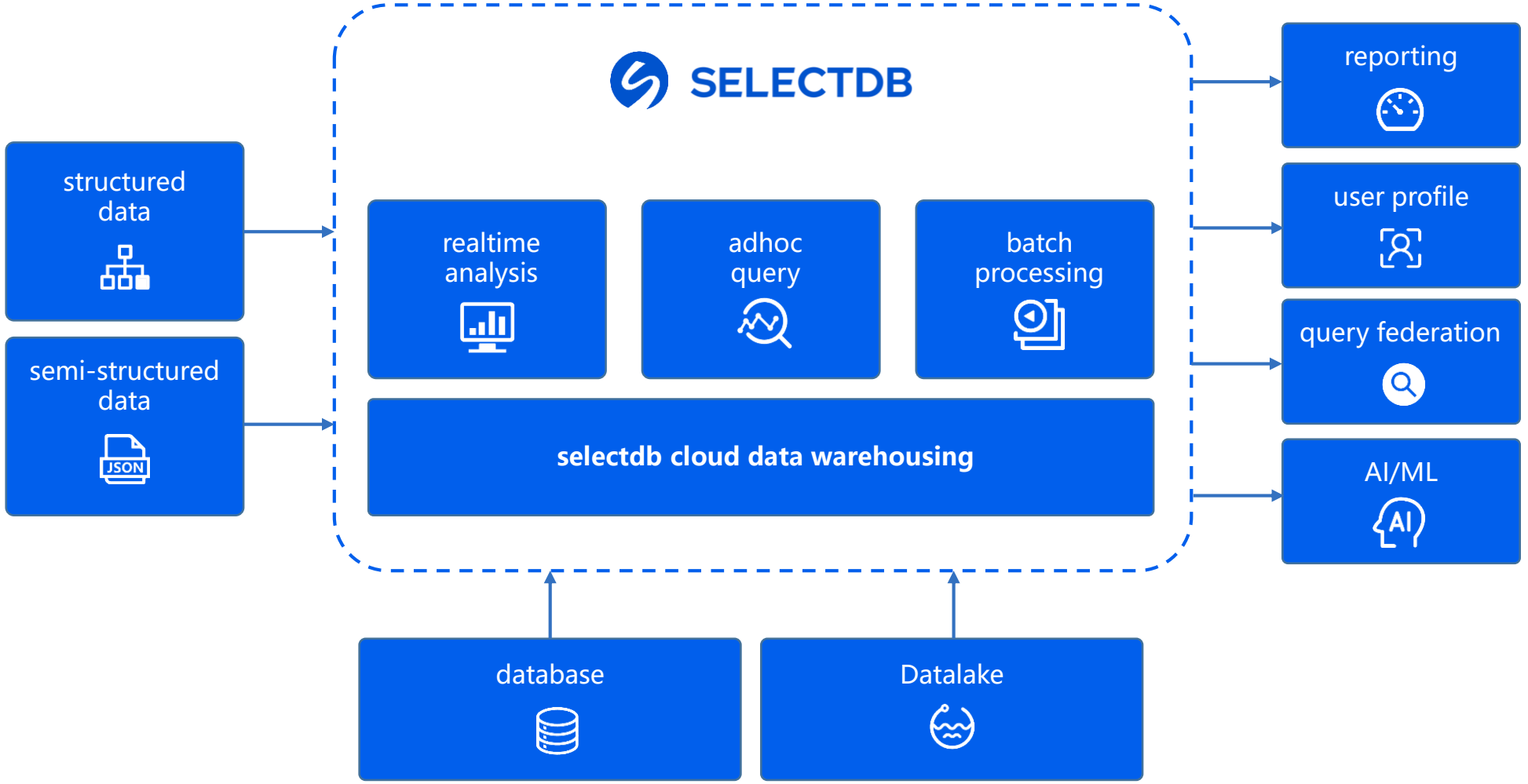
近半数 SQL 在所有产品中最佳

通用机型下 SelectDB 性能第一



查询总耗时远低于行业竞品

融合统一


SELECTDB
structured data
semi-structured data
JSON
realtime analysis
adhoc query
batch processing
selectdb cloud data warehousing
database
Datalake
reporting
user profile
query federation
AI/ML
AI

融合统一 | 混合负载

高并发报表

- 支持数千 ~ 数十万的QPS
  - ✓ 分区裁剪
  - ✓ sorted short key index
  - ✓ 物化视图
  - ✓ SQL, Partition Cache

即席分析

- 支持即席查询，秒级响应
  - ✓ MPP和向量化
  - ✓ 智能索引skip
  - ✓ 智能优化器

ETL

- DAG执行模型增强
- Native支持半结构化和非结构化数据

湖仓联邦分析

- 用SelectDB作为统一SQL查询层查询已经存在于离线湖仓中的数据
- 相对presto和hive性能提升数倍
  - ✓ MPP和向量化
  - ✓ 智能优化器 （下推和Join）
  - ✓ 数据和元数据cache

统一数仓

- Data Engineering： ETL
- Data Analytics： 即席分析 + 湖仓联邦分析
- Data Serving： 报表分析

混合负载管理

- 资源隔离 （硬限, 软限）
- 落盘与容错模型
- WorkLoad Manager (WLM)

融合统一 | 半结构化分析

JSON数据

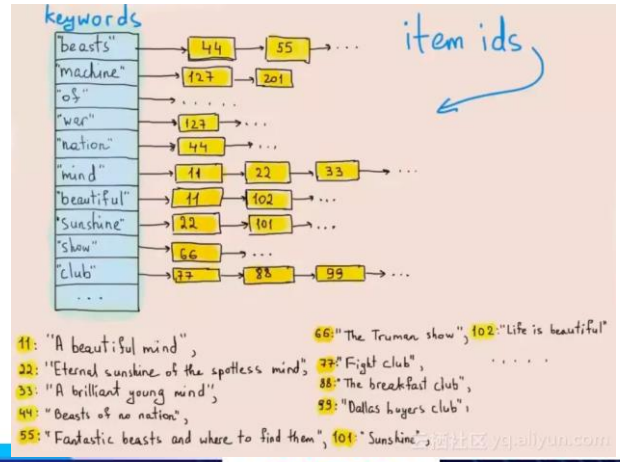
高效点查索引

Free Text

对非数值型字段提供Inverted Index索引

对数值型字段提供BKD Index索引

```
-- 建表
CREATE TABLE IF NOT EXISTS t_dynamic
(
    name string,
    age  int,
    ...
)
DUPLICATE KEY(`name`)
DISTRIBUTED BY HASH(`name`) BUCKETS 10
```



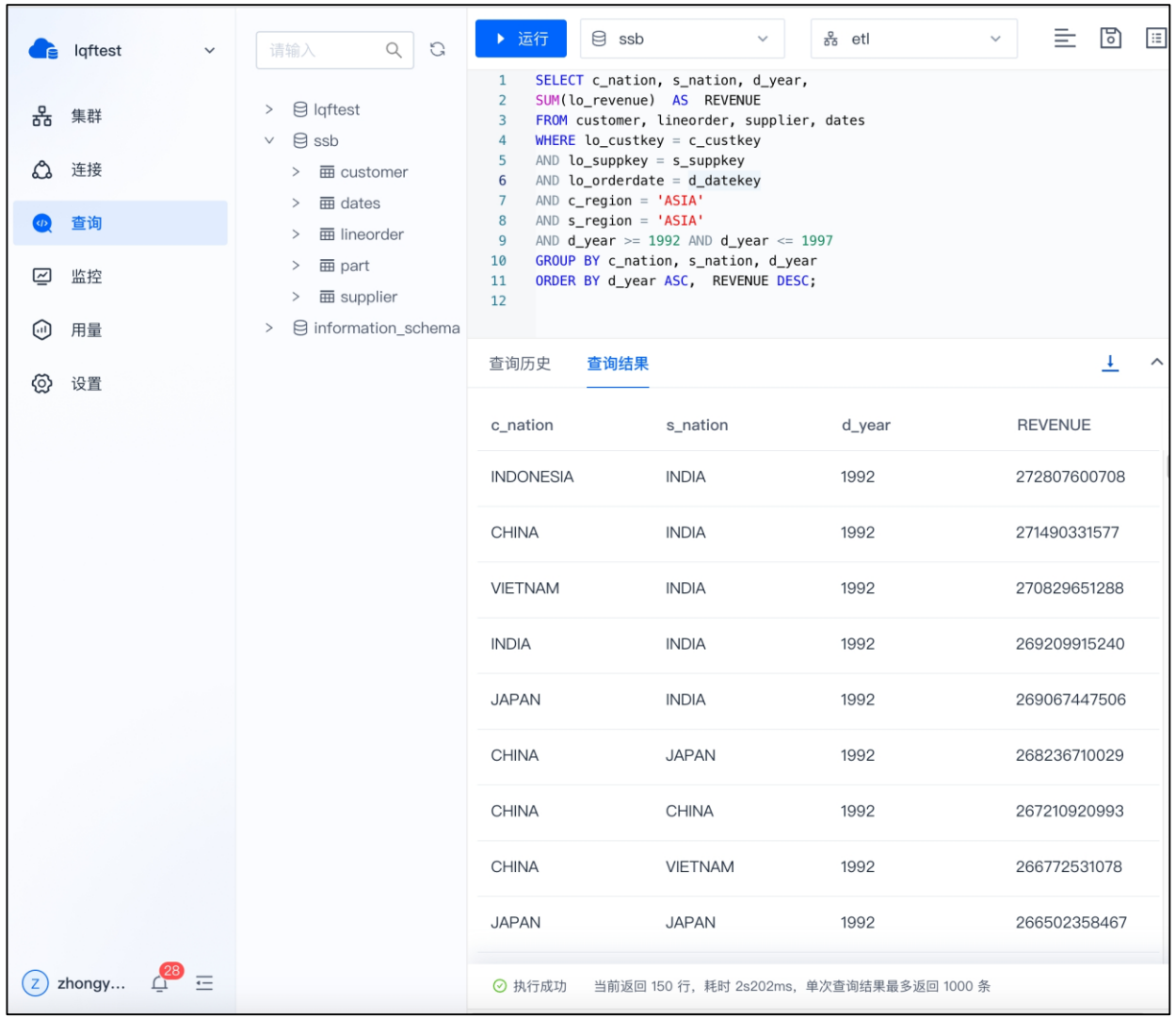

```
SELECT UNIX_TIMESTAMP(log_time) as time, host, type, level, msg
FROM dorislog
WHERE log_time >= $__timeFrom() AND log_time <= $__timeTo()
  AND type = '$log_type'
  AND msg MATCH_ANY '$keywords'
ORDER BY time DESC
LIMIT $limit
```

简单易用

兼容 MySQL 链接协议，标准SQL语法

基于 WebUI 的数据库管理和开发工具

Spark/Flink/Kafka/DBT 等丰富的连接器



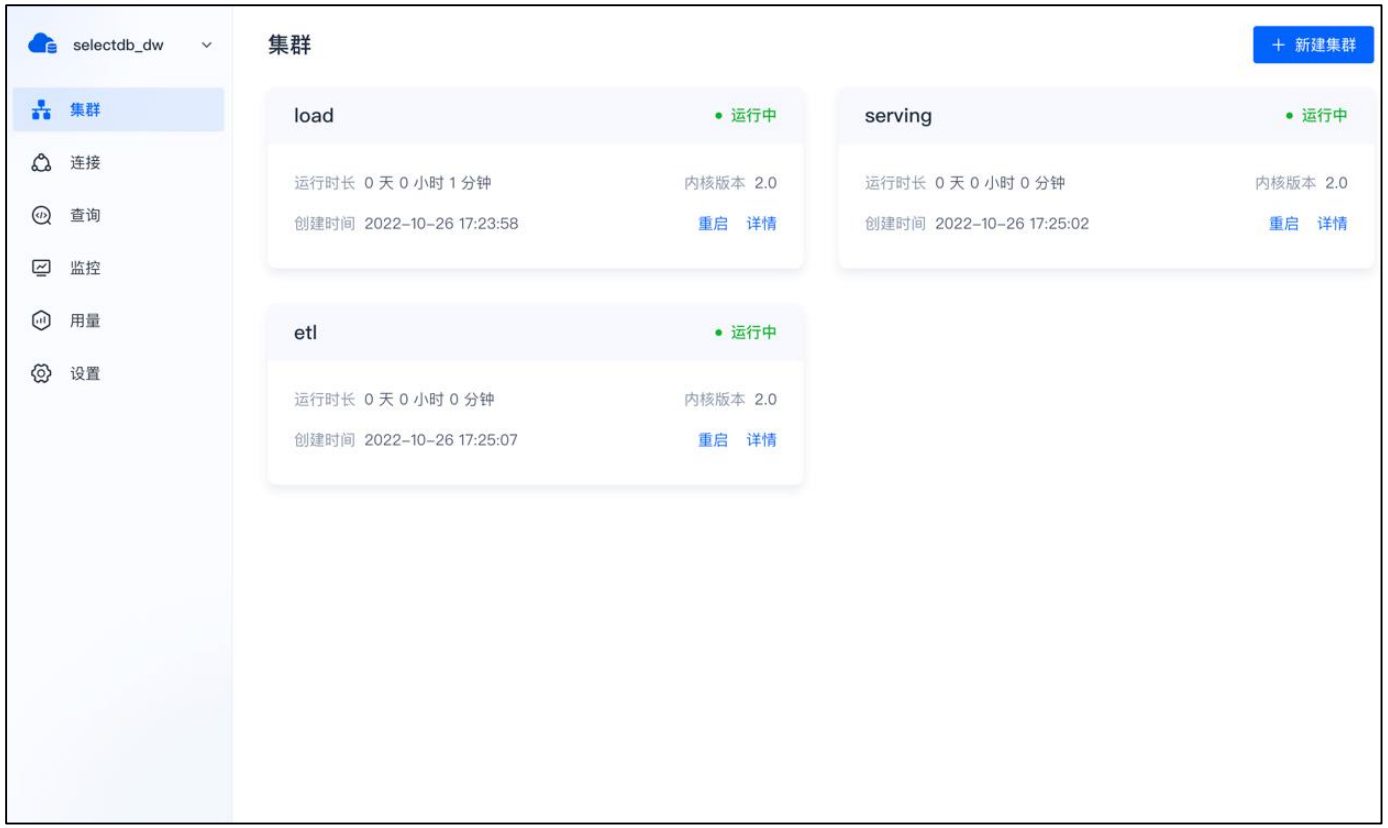
高性价比

存储按使用付费，计算按需弹性

数据共享，消除冗余数据

负载隔离，避免负载互相影响需要的额外资源

突发或者后台任务，云上竞价资源灵活调度



开源开放

- 基于Apache Doris研发，积极贡献Doris，充分利用社区力量和应用场景打磨引擎

- 多云中立，运行在多云之上，提供统一的产品体验，防止被锁定

- 兼容Doris数据格式，方便开源用户到商业产品的迁移

企业特性 | 安全

数据保护：落盘和传输加密，密钥管理，数据脱敏

访问控制：RBAC，细粒度访问控制（行级别和列级别）

认证：多因子认证，IAM，ID Federation

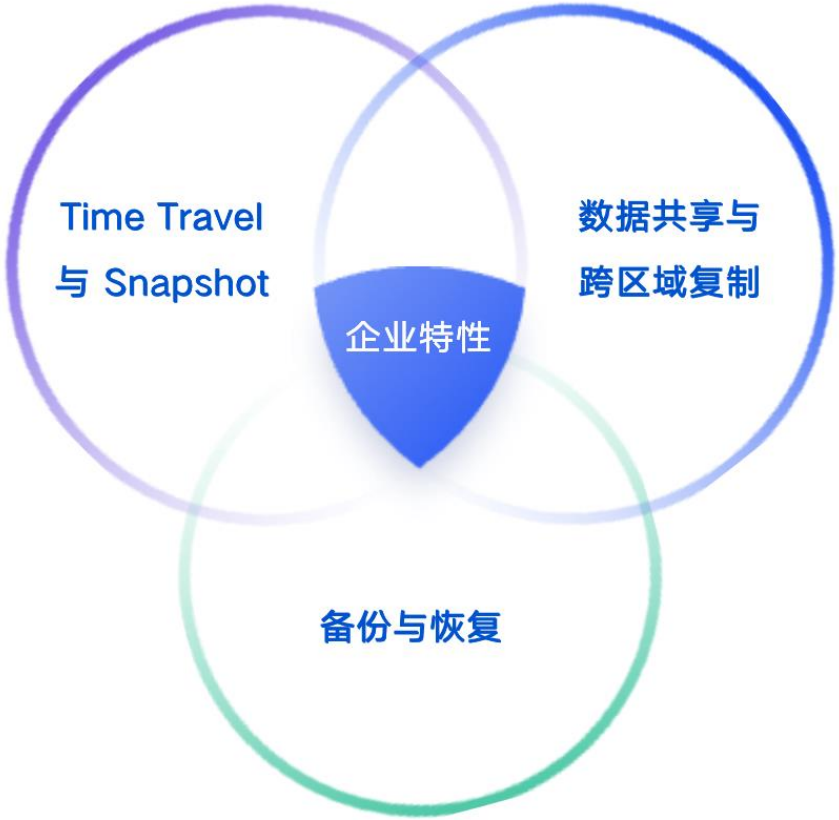
网络安全：网络隔离

企业特性

- 存储表或者数据库的各个导入版本，并且支持对表或者数据库建立快照

- 同一地域多个数据库实例支持数据共享，支持跨地域数据实时复制（CCR）

- 依托对象存储的增量备份及数据恢复机制



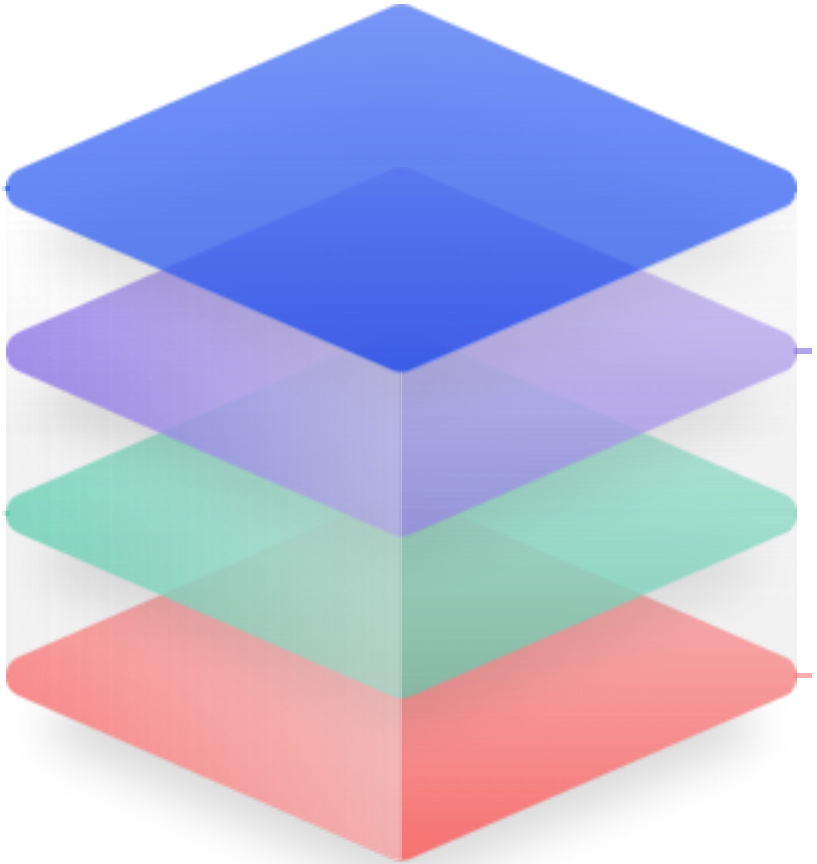
SELECTDB Cloud 应用场景

面向客户的报表与分析

面向管理者的驾驶舱Dashboard；
面向业务分析人员的Reporting；
面向C端/B端用户的高并发报表分析

用户画像与行为分析

收集用户属性与行为数据，构建用户数据平台，进行用户参与、留存和转化等行为分析，以及人群洞察和人群圈选等画像分析



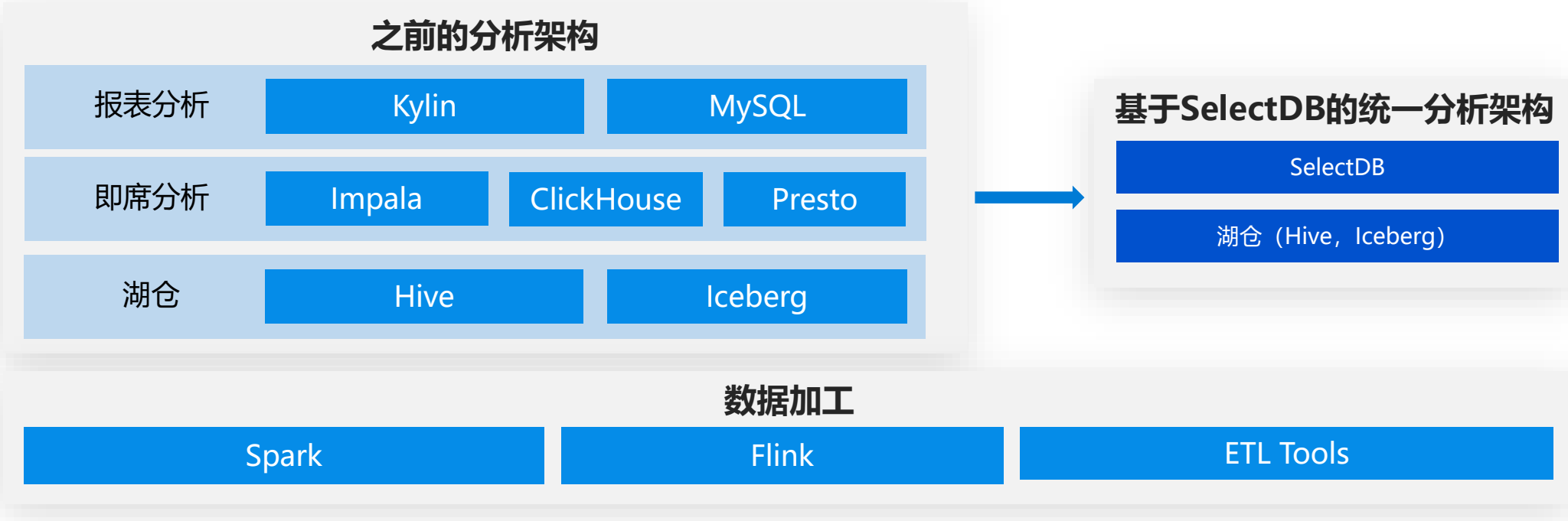
湖仓一体的现代数据平台

统一数据仓库和数据湖到统一平台，提供面向企业内部的Bi报表和adhoc分析，批量以及增量ELT

日志存储与分析

将业务、系统或者物联网相关的日志数据存储为结构化、半结构化或者原始文本，构建统一的日志存储与分析平台。

典型应用场景 – 湖仓一体的现代化分析平台



1. 公司每天新增明细数据100亿，10TB单副本，大数据平台有数千台机器。
2. 旧的架构报表分析使用Kylin和MySQL，即席分析使用Impala, ClickHouse, Presto，将**报表分析**和**即席分析**场景用**SelectDB**，大大降低架构的复杂性，性价比提升**3倍**。
3. 使用SelectDB作为Hive和Iceberg的**湖仓查询加速引擎**，性能相比Presto有**3倍**提升，相比Hive有**10倍**以上提升。

典型应用场景 – 半结构化数据分析

	日志存储分析	可观测性	安全分析
业务应用场景			
	用户行为分析 CND流量分析 服务故障分析	系统资源监控 服务质量监测 APM性能监测	网络流量安全监测 网络安全态势感知 安全事件追踪溯源
分析需求特点	写入吞吐大 存储周期长 快速检索	时效性要求高 历史数据聚合 监控指标多且扩展	写入吞吐大 快速检索 多种数据关联分析

业务价值

- 性能提升**2倍**以上
- 资源成本节省**60%+**
- 融合ES能力在统一数仓，**简化系统架构、减少维护成本**

典型应用场景 – 半结构化数据分析

SELECTDB

业务价值

性能提升**2倍**以上

资源成本节省**60%+**

融合ES能力在统一数仓，**简化系统架构、减少维护成本**

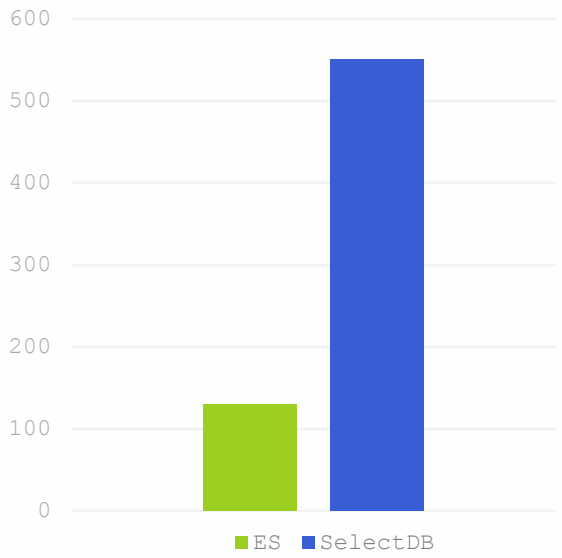
	快速检索	复杂分析	高性价比
半结构化增强	倒排索引	动态表	冷热分离
Doris引擎	执行优化器	标准SQL	场景化存储模型
	MPP架构	向量化执行	列式存储

典型应用场景 – 半结构化数据分析

SELECTDB

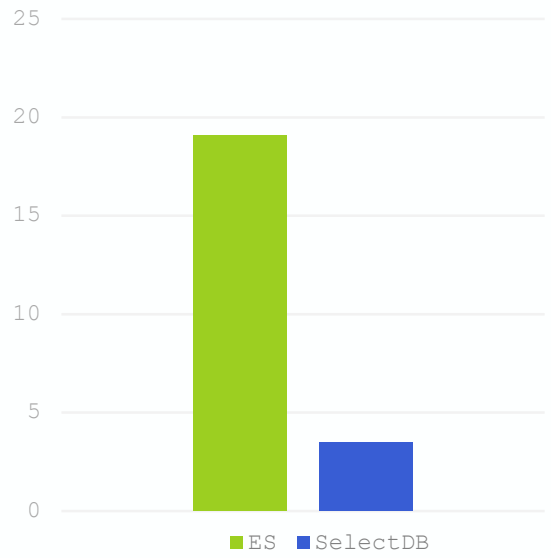
4.2倍

写入速度(MB/s)



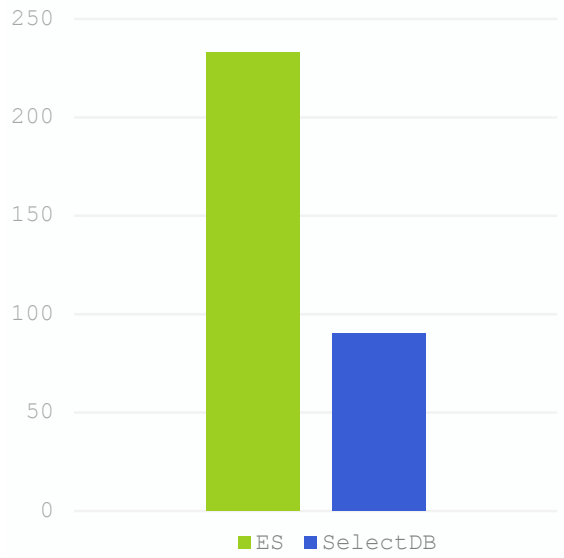
5.4倍

磁盘空间(GB)



2.3倍

查询时间(秒)



说明：

1. 以上数据为ES官方性能benchmark中http_logs场景测试结果
2. 写入速度越高越好，磁盘空间越低越好，查询时间越低越好

欢迎关注

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

THANKS