Apache Doris 3.0 云原生存算分离架构的设计与实现

陈明雨 | 飞轮科技技术副总裁 & Apache Doris PMC Chair

关于 Apache Doris 城市行

Apache Doris 城市行是由飞轮科技发起、面向全国各地大数据、数据库以及实时数据仓库技术爱好者的线下技术交流活动。

Apache Doris 城市行旨在更好的为各地区技术爱好者分享一手的技术知识和实践经验，并搭建一个可以帮助技术爱好者与 Apache Doris 社区技术大咖线下交流、讨论的平台。同时，我们也希望通过线下面对面交互的形式，能够与各地区的社区成员产生紧密的连接，倾听社区成员们的想法，共同建设 Apache Doris 社区。

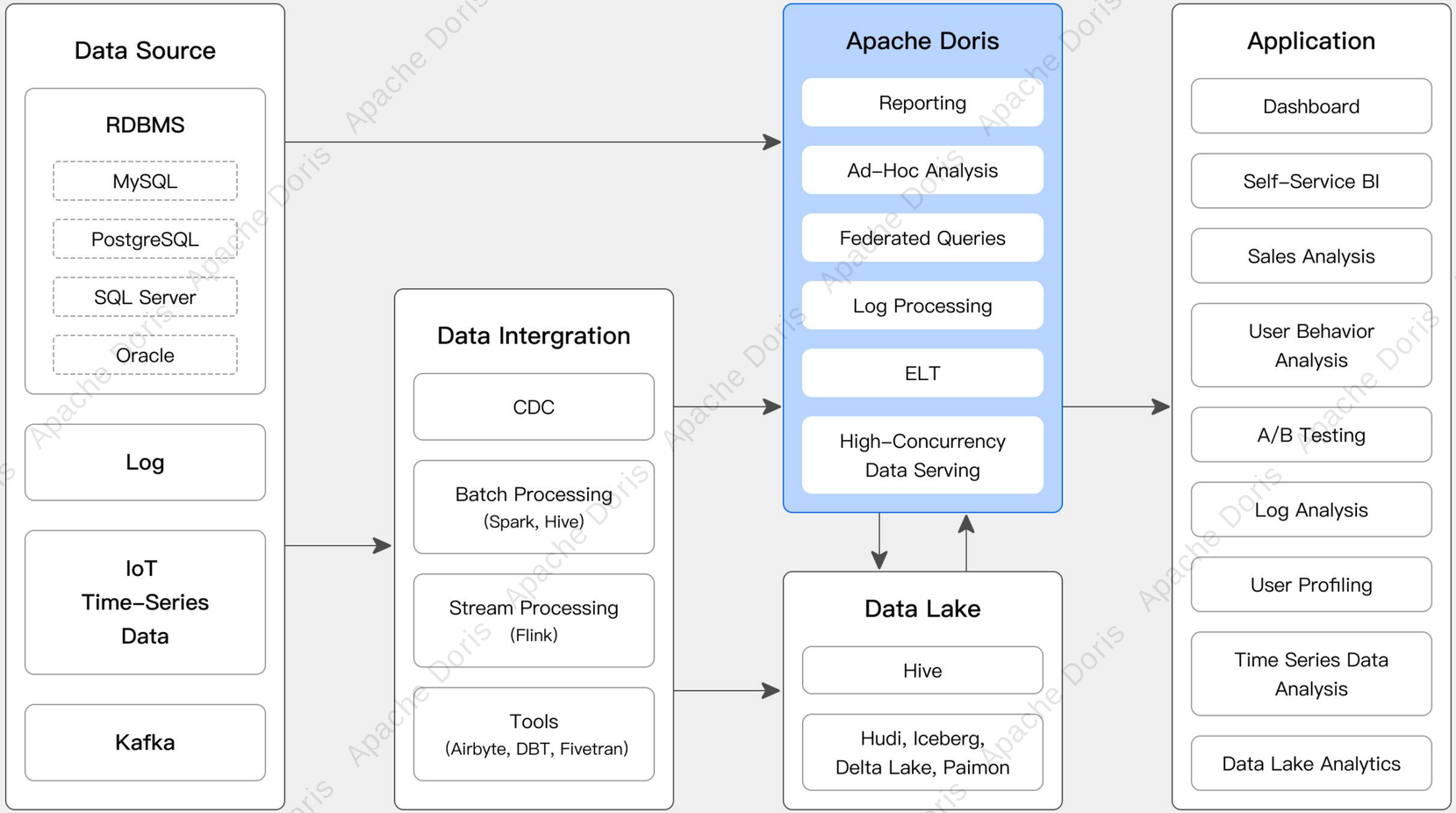
截止目前，已分别在北京、上海、深圳、杭州、武汉、成都、西安、广州等国内连续成功举办近 10 场次，吸引线上线下近 10w+ 国内技术爱好者的参与、活动曝光量超 100w+ ，是国内大数据领域参与人数最多、氛围最佳、内容最为丰富的技术交流活动之一。

个人介绍

陈明雨

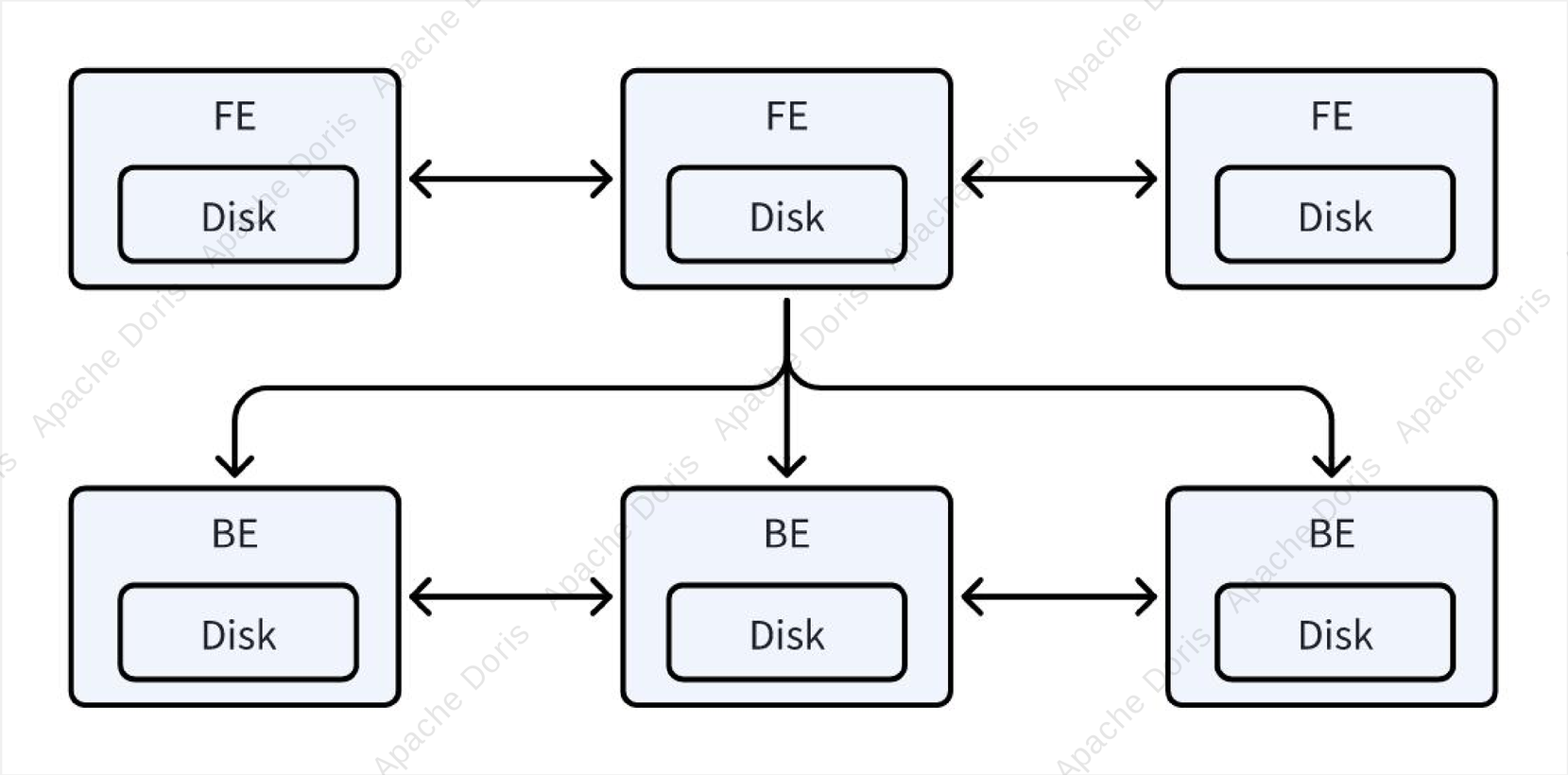
- 飞轮科技技术副总裁、Apache Doris PMC Chair、Apache Member
- 曾担任百度 Doris 团队技术负责人，主导了 Apache Doris 从毕业成为 Apache 基金会顶级项目的全过程
- 有近 10 年分布式数据库架构设计、研发与团队管理经验

What is Apache Doris



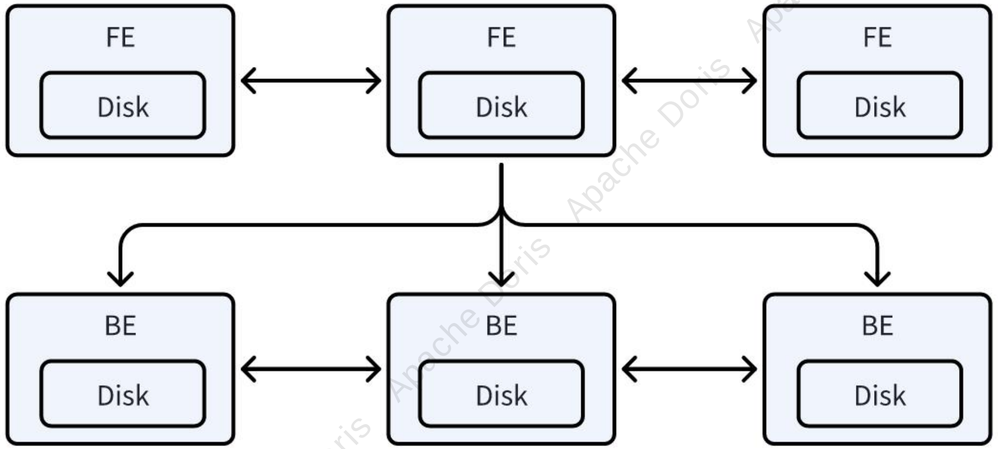
目录

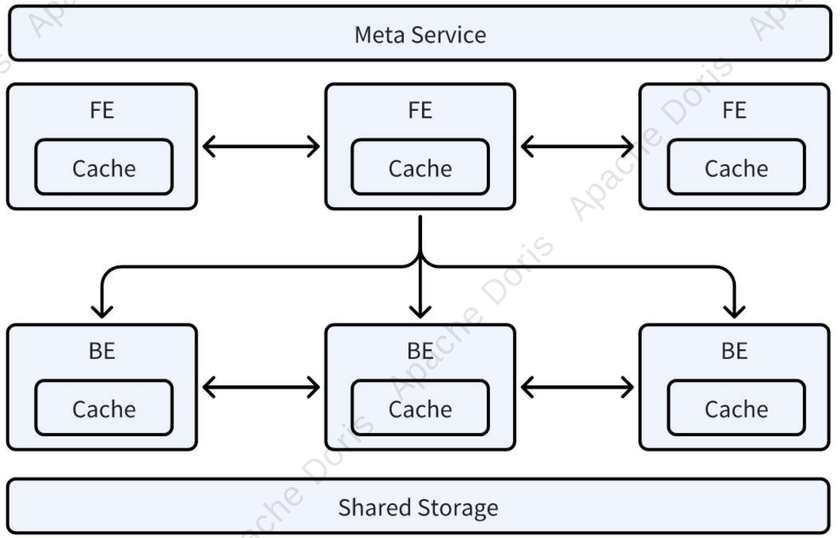
Apache Doris 存算一体模式



在存算一体架构下，BE 节点上存储与计算紧密耦合，数据主要存储在 BE 节点上，多 BE 节点采用 MPP 分布式计算架构。

Apache Doris 存算一体模式





- **部署简单**：仅FE与BE进程，BE和FE都可以单独扩容
- **稳定可靠**：不依赖共享存储系统
- **性能优异**：计算节点访问本地存储

- 为什么要存算分离？

为什么需要存算分离

低成本与资源弹性

- 计算和存储解绑，单独扩缩容
- 计算资源波谷波峰，灵活弹性
- 数据存储冷热效应明显

负载隔离

- 读写任务分离
- 更彻底的业务隔离，解决不同业务间的相互影响以及资源抢占问题

数据共享

- 单一数据面向不同的分析负载使用
- 数据快速移动、快速备份恢复
- Single Source of Truth

云基础设施的成熟

- 云上基础设施逐步完善，提供可靠的共享存储
- 完全按量付费，灵活可控

目录

设计出发点 - 性价比与架构稳定性

如何降低成本

- 引入对象存储节省冷数据资源
- 增加弹性计算能力，按需使用计算资源

不同架构如何迭代

- 绝大多数用户已采取存算一体架构
- 升级过程中需要保证对已有架构的兼容

设计目标

负载隔离

读写分离

业务隔离

内部负载隔离

低成本

存储成本大幅下降

计算和存储可以独立弹性

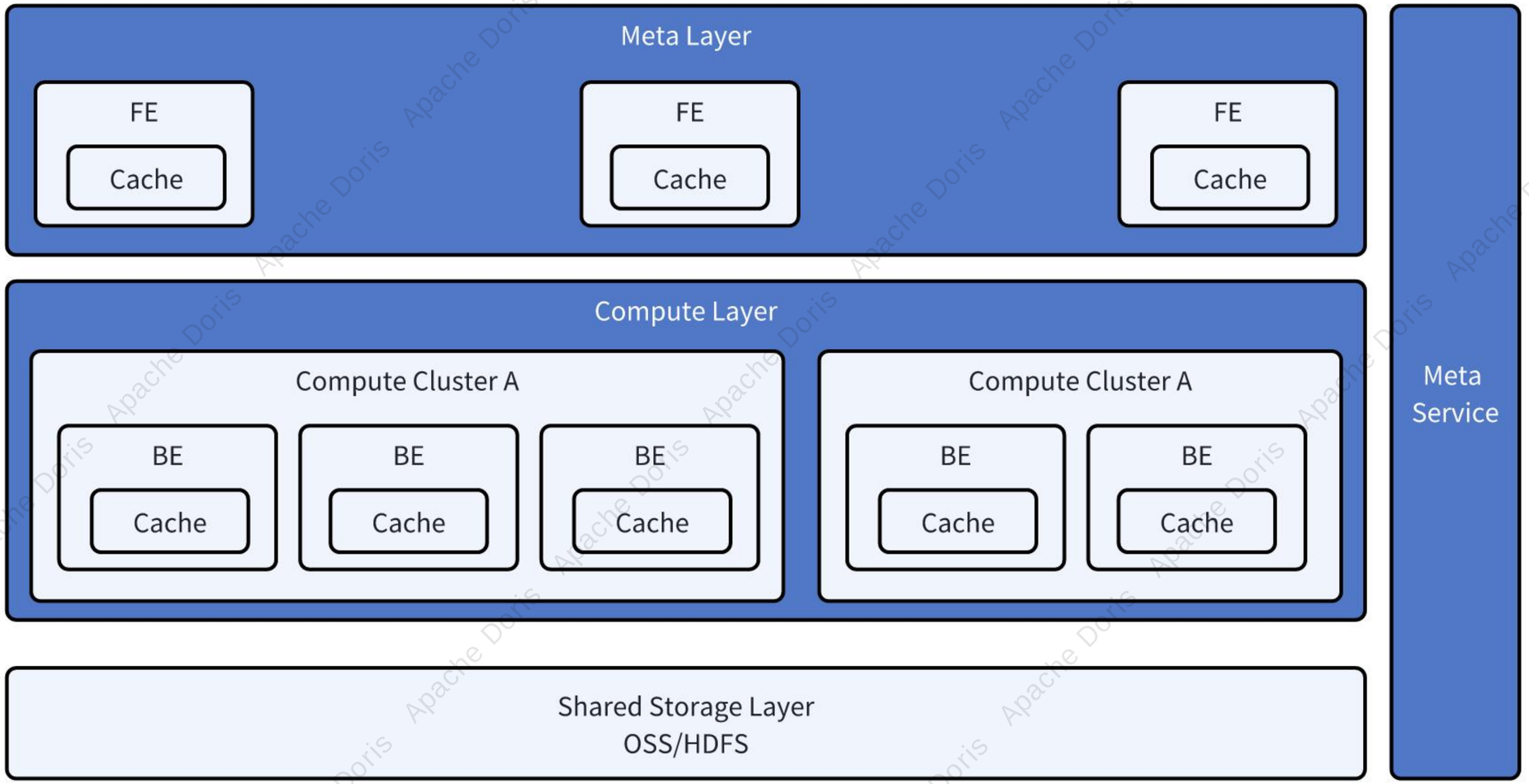
使用业务的波峰波谷调整计算资源

数据共享

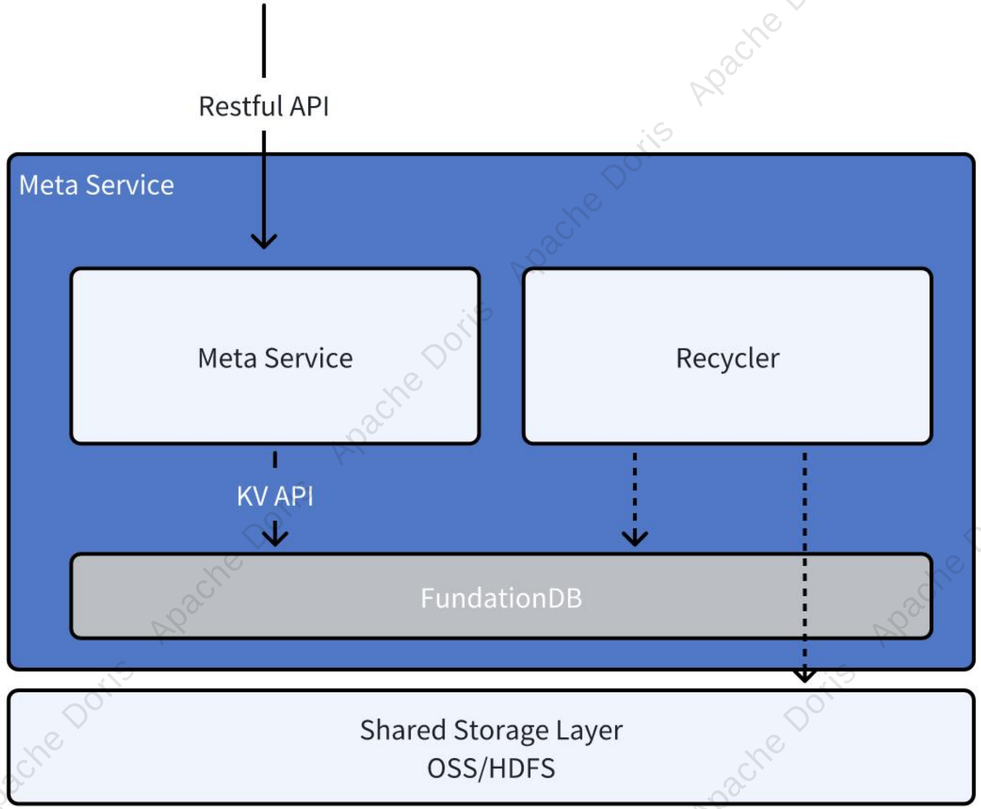
统一元数据服务

统一存储

存算分离整体架构


Meta Layer
FE
Cache
FE
Cache
FE
Cache
Compute Layer
Compute Cluster A
BE
Cache
BE
Cache
BE
Cache
Compute Cluster A
BE
Cache
BE
Cache
Shared Storage Layer
OSS/HDFS
Meta
Service

元数据服务层



- 统一语义层：Restful API
- 高性能分布式KV存储：FoundationDB
- 正向垃圾回收：Recycler

数据存储层

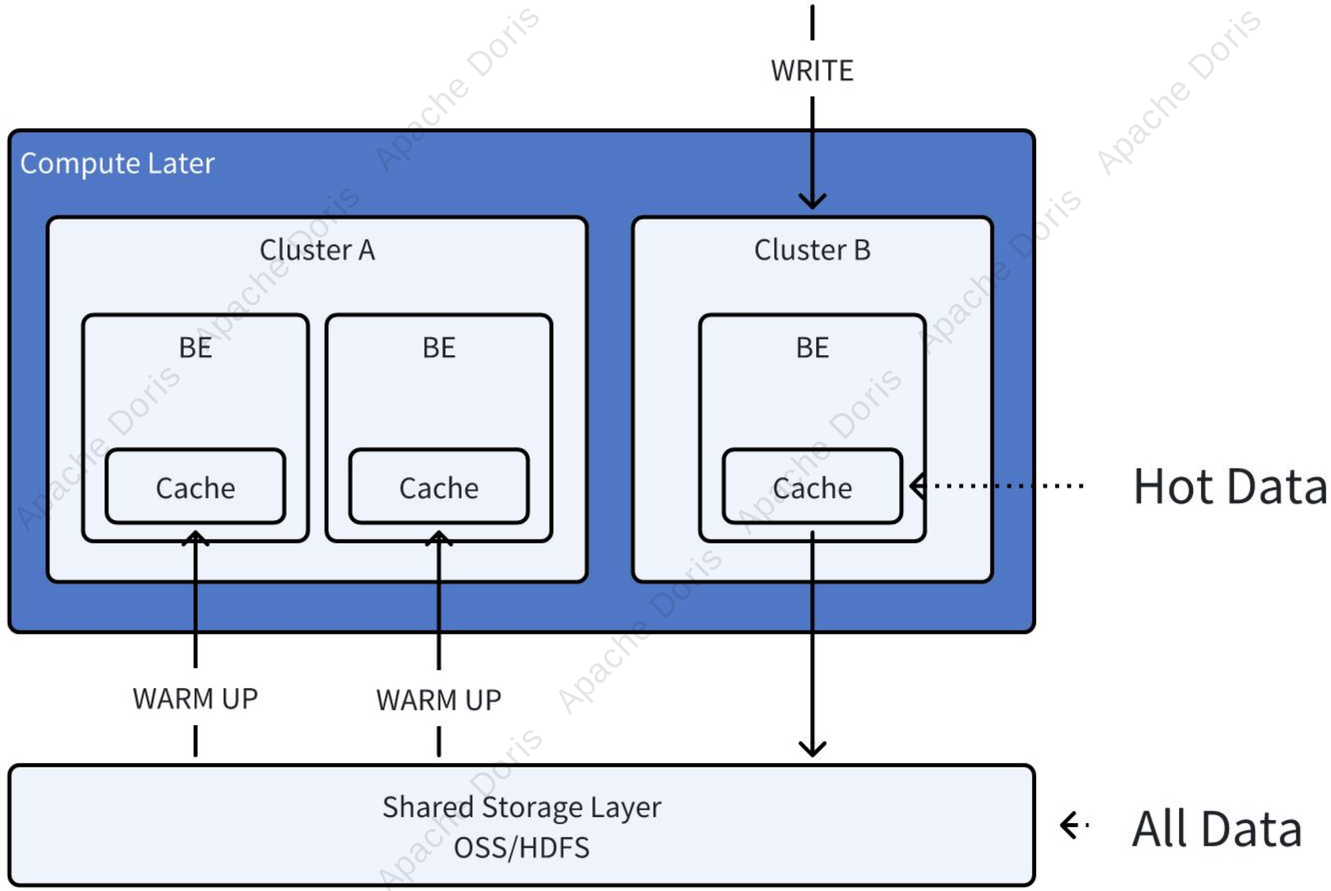
成本最高降低90%

- **存算一体**
- 全量数据 * 3 * 块存储价格
- **存算分离**
- 热数据 * 1 * 块存储价格 + 全量数据 * 对象存储价格

- 最高可以节省90%以上

灵活的Cache管理

- LRU、TTL
- 缓存预热

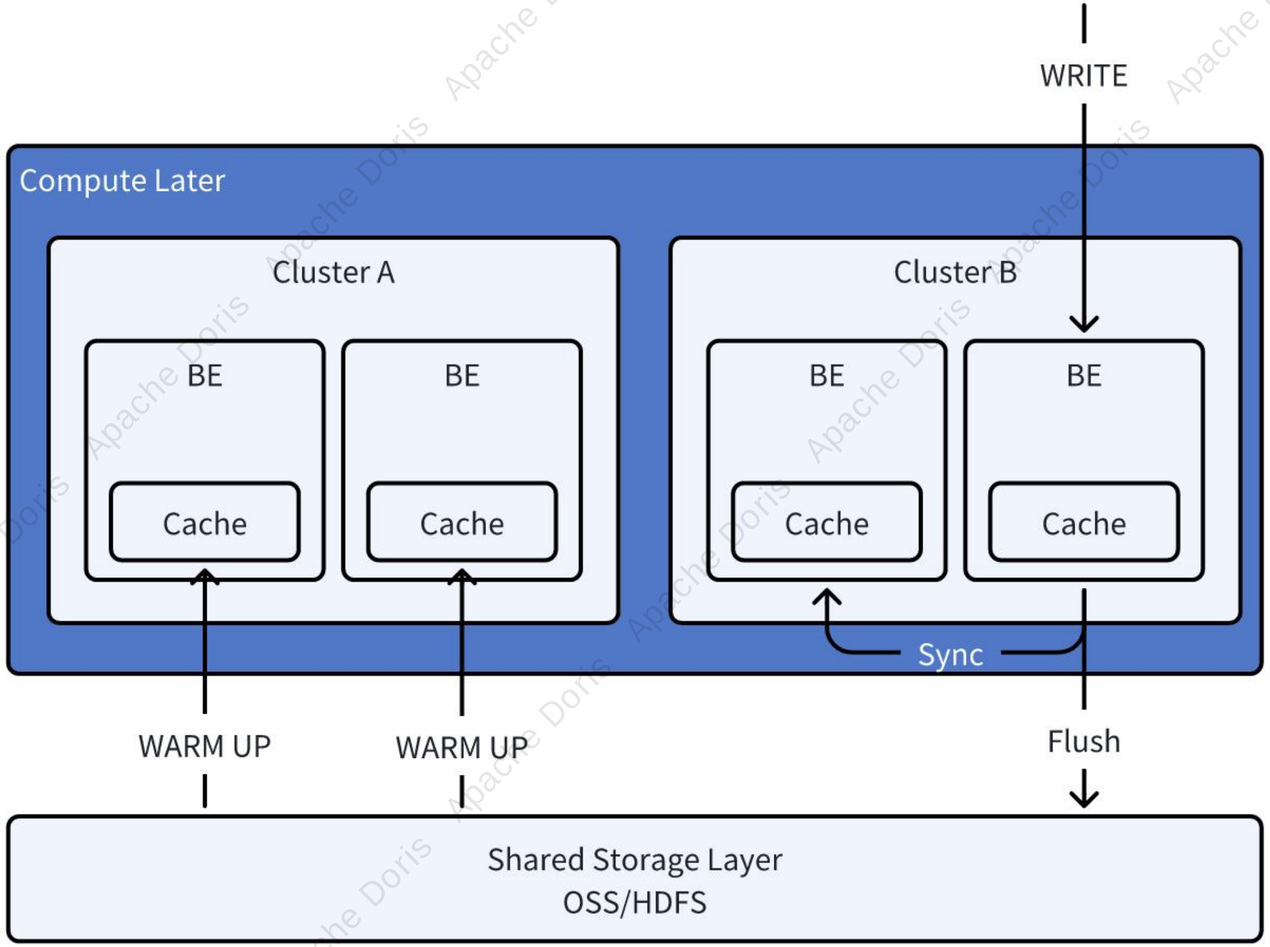


数据计算层-数据导入

1. 数据进入协调者 BE
2. 数据分发到多个 BE
3. 数据写入 Cache
4. 数据写入 S3
5. 读写分离 Cluster 预热 Cache

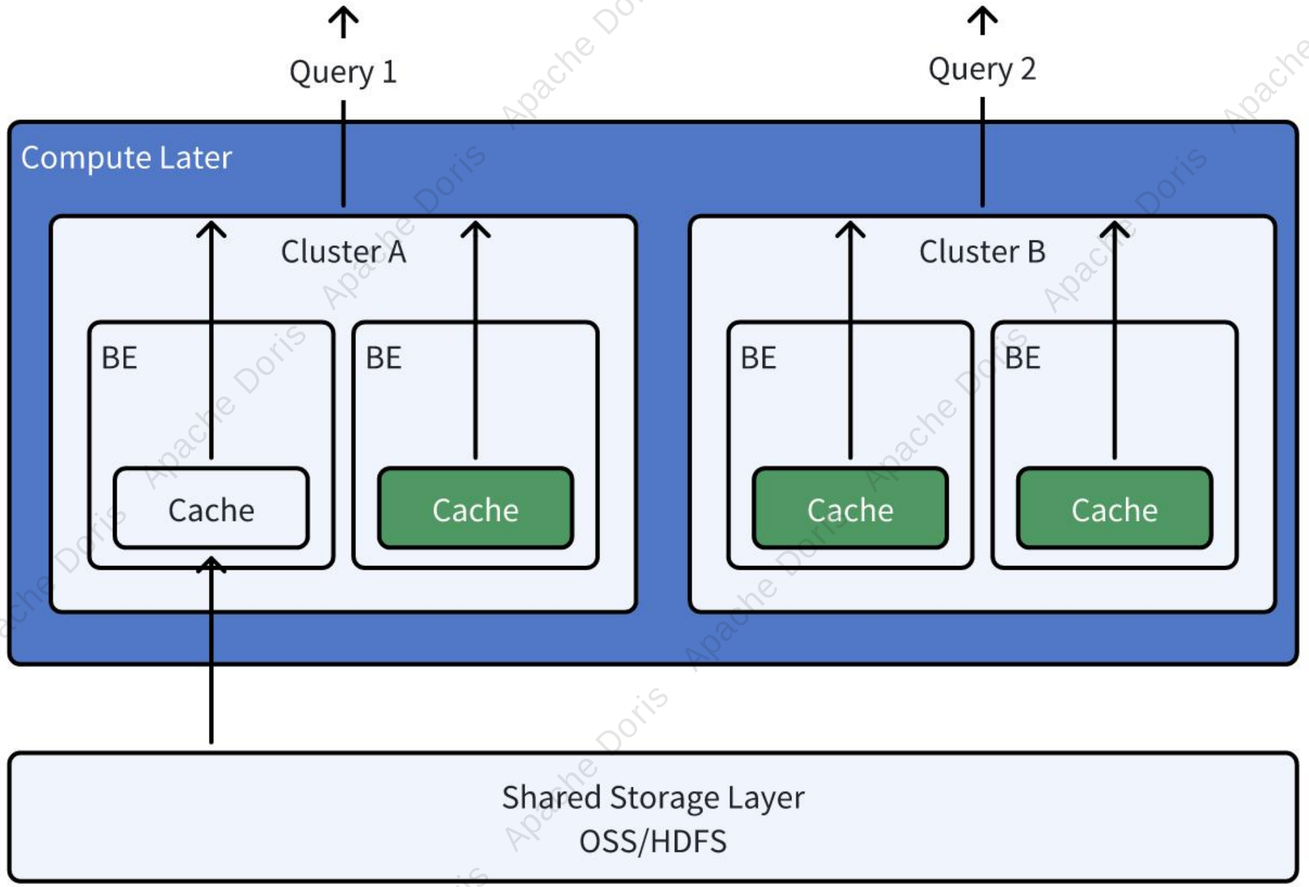
数据导入效率更高

- 只需处理单副本数据
- 数据和BE没有固定的关系
- 没有 Publish 阶段，写入流程更短

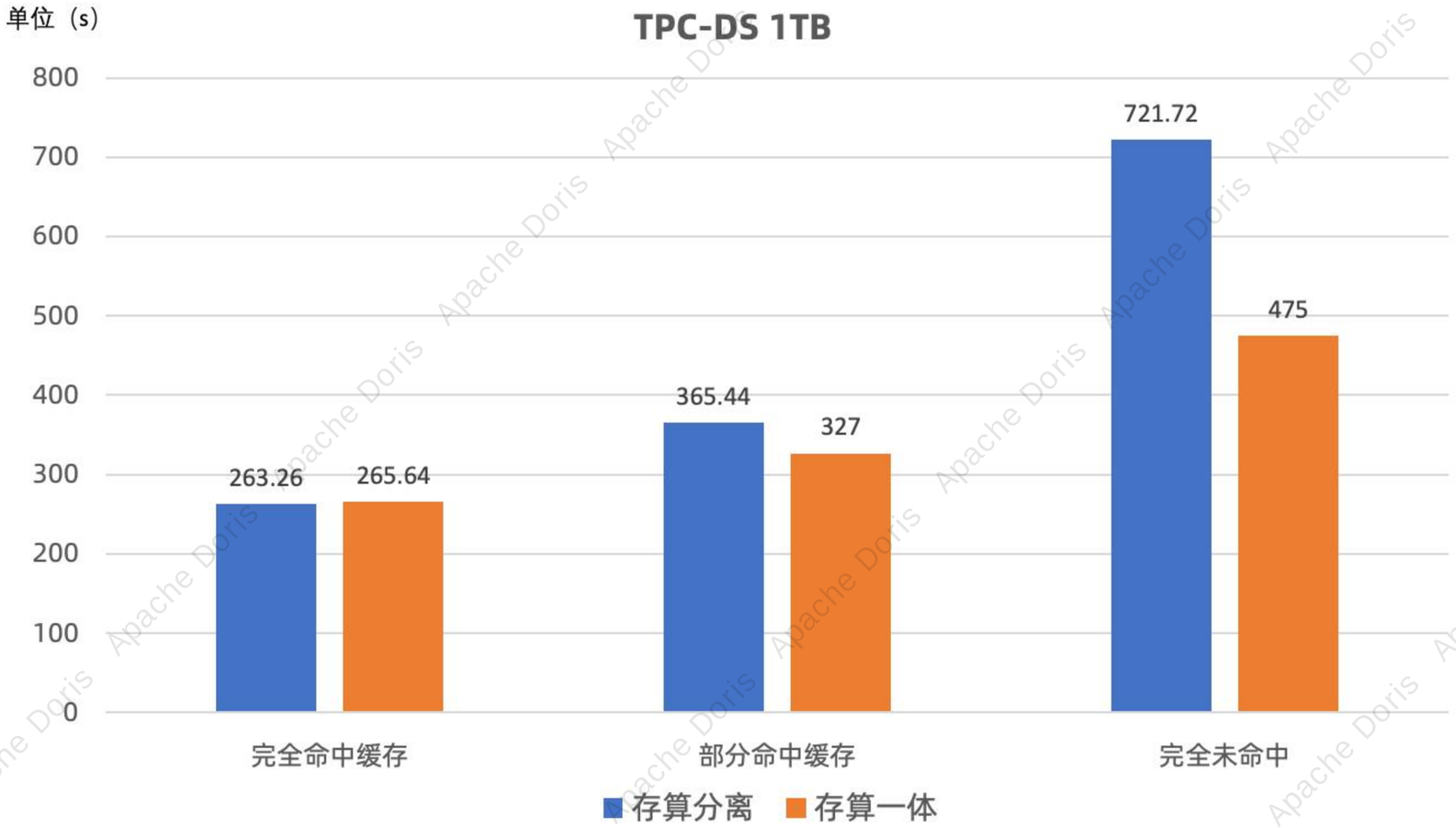


数据计算层-数据查询

- 多个 cluster 独立
- 不命中 Cache 时从 S3 读数据
- 命中时从本地 Cache 读数据
- 弹性资源大幅降低成本



查询性能对比

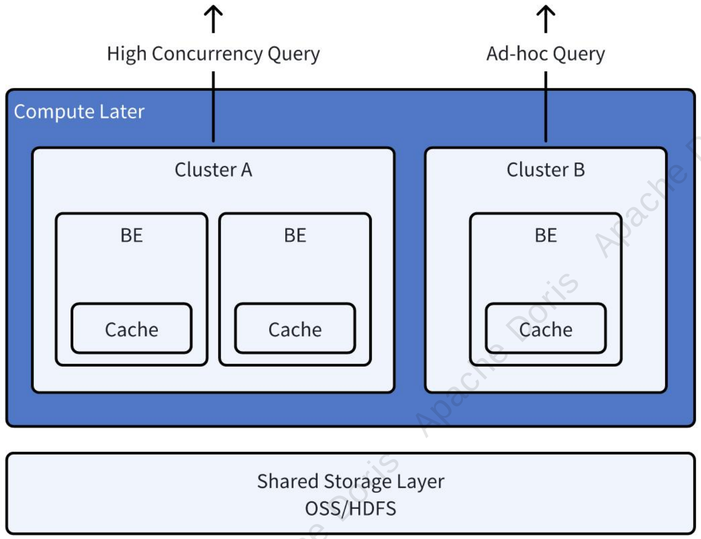


缓存层级

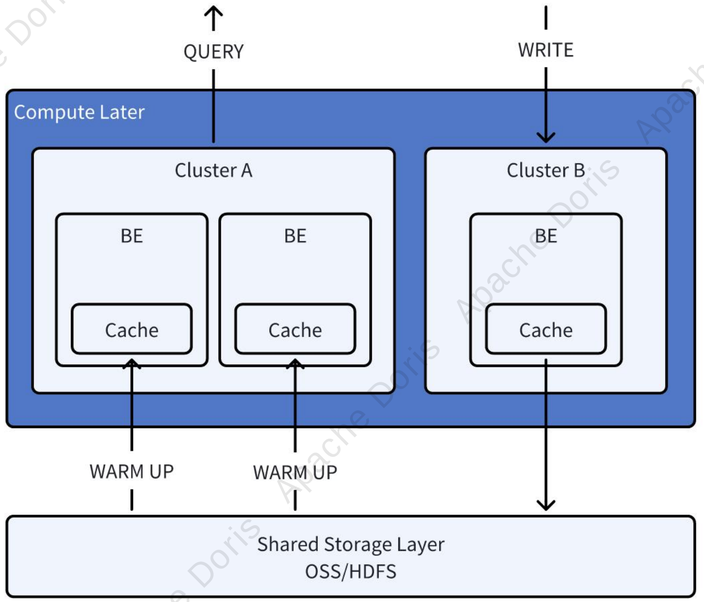
- Remote Storage
- Local File Cache
- System Page Cache
- Doris Page Cache

完全命中缓存时查询性能完全持平，部分命中缓存时有10%的性能损耗、随测试进行数据逐渐加载进缓存，性能随之提升；极端情况下（完全未命中任何缓存）性能损耗约 30%。

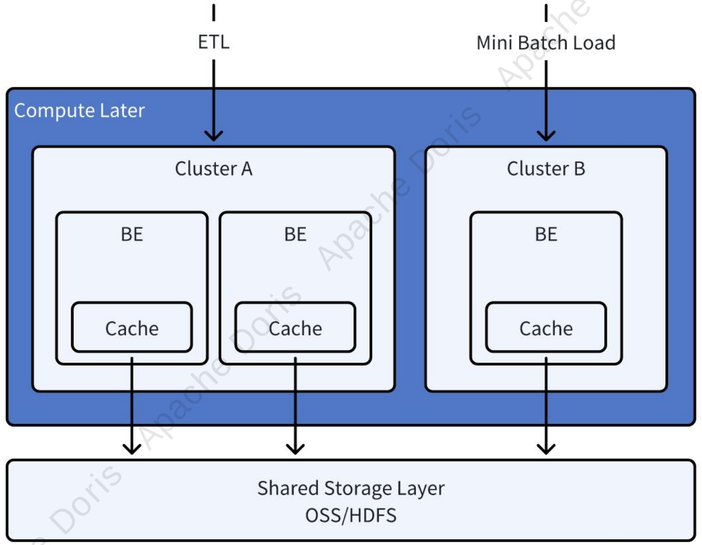
最佳实践



- **读读隔离**
- 高优查询和普通查询
- 高并发点差和即席分析



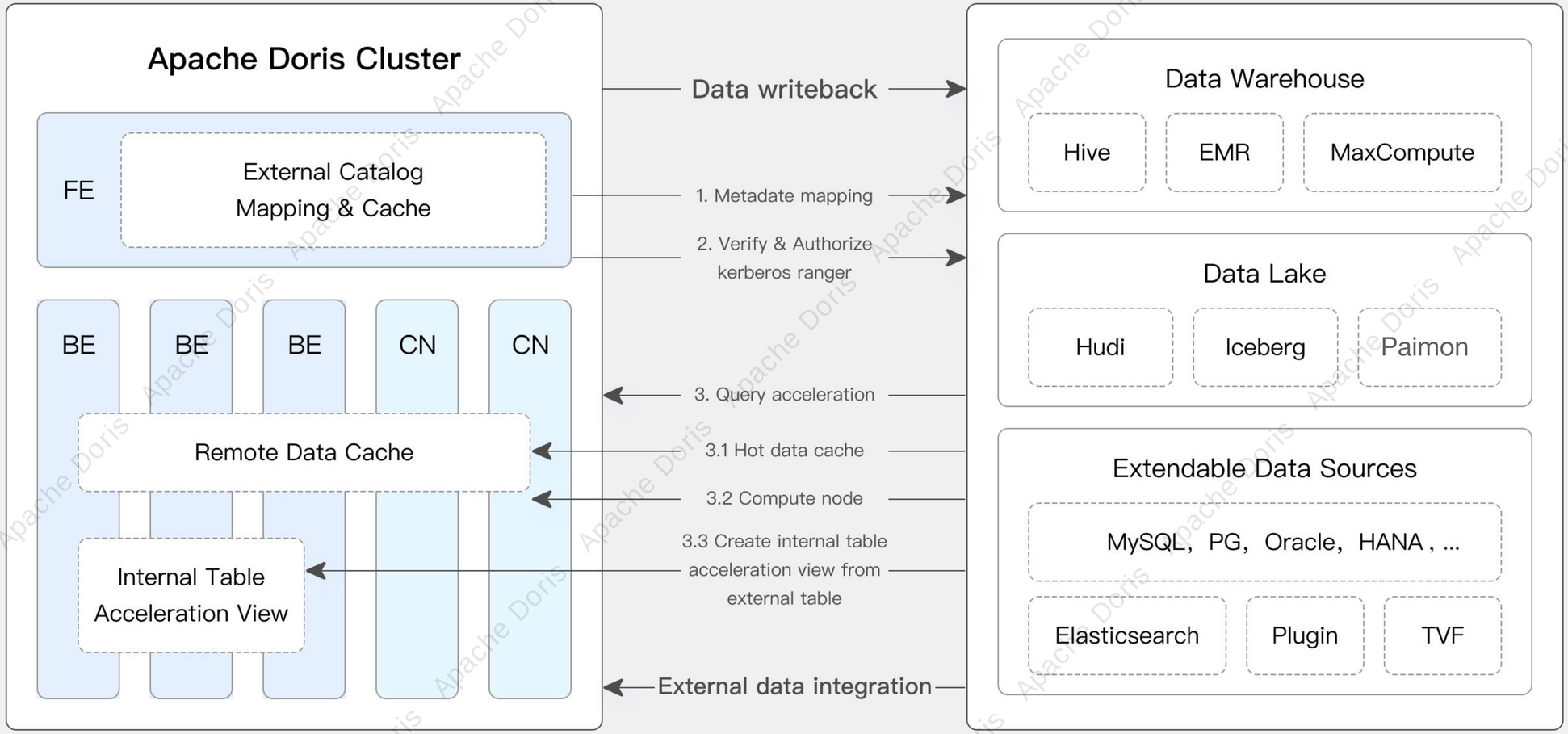
- **读写隔离**
- 实时同步
- 自动预热



- **写写隔离**
- 高频导入和ETL

目录

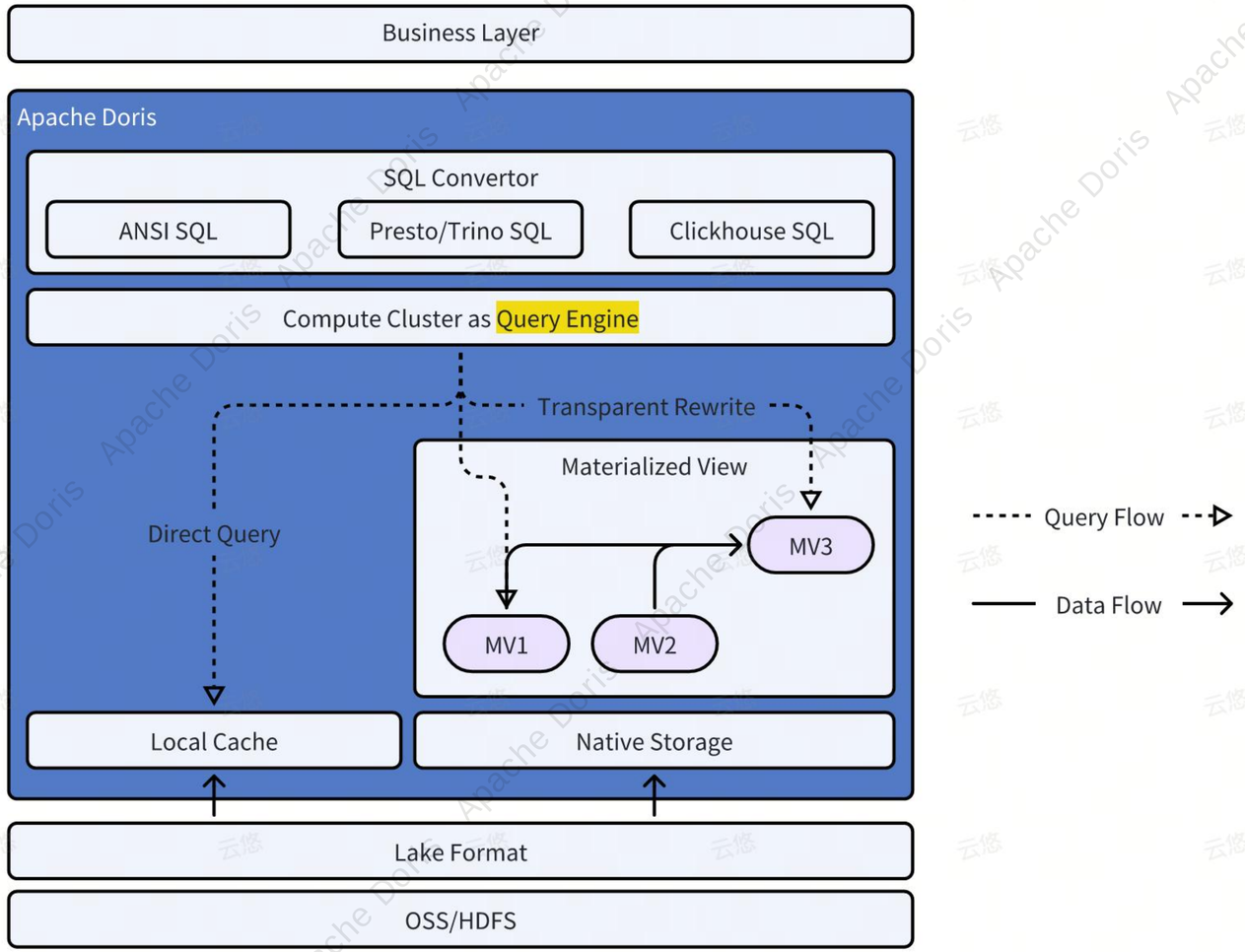
3.0 特性 - Lakehouse 再进化



3.0 特性 - Lakehouse 再进化

阶段1：湖仓加速

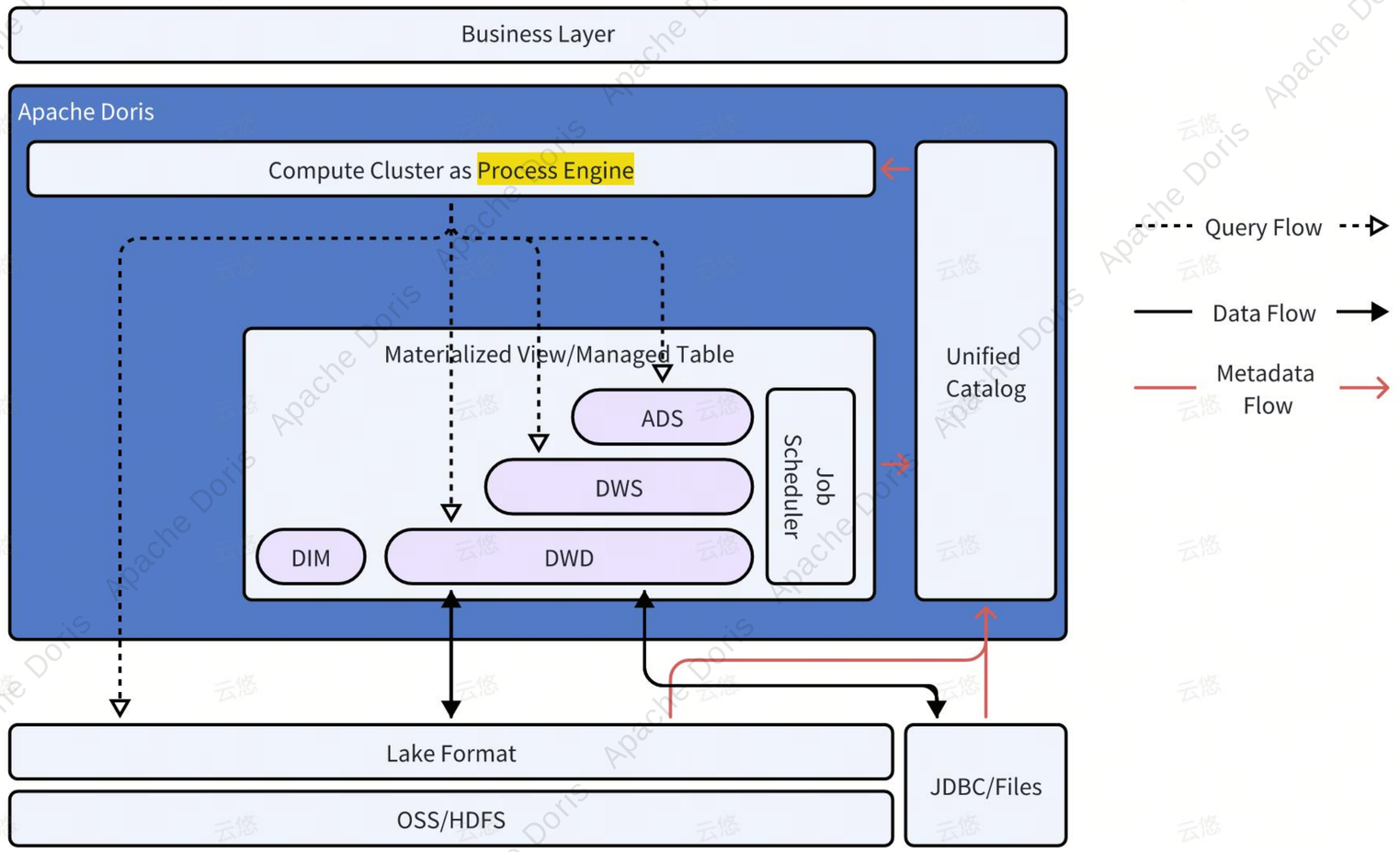
- 丰富的数据源连接
- 联邦分析能力和SQL方言兼容
- 物化视图和透明改写



3.0 特性 - Lakehouse 再进化

阶段2：湖仓数据处理

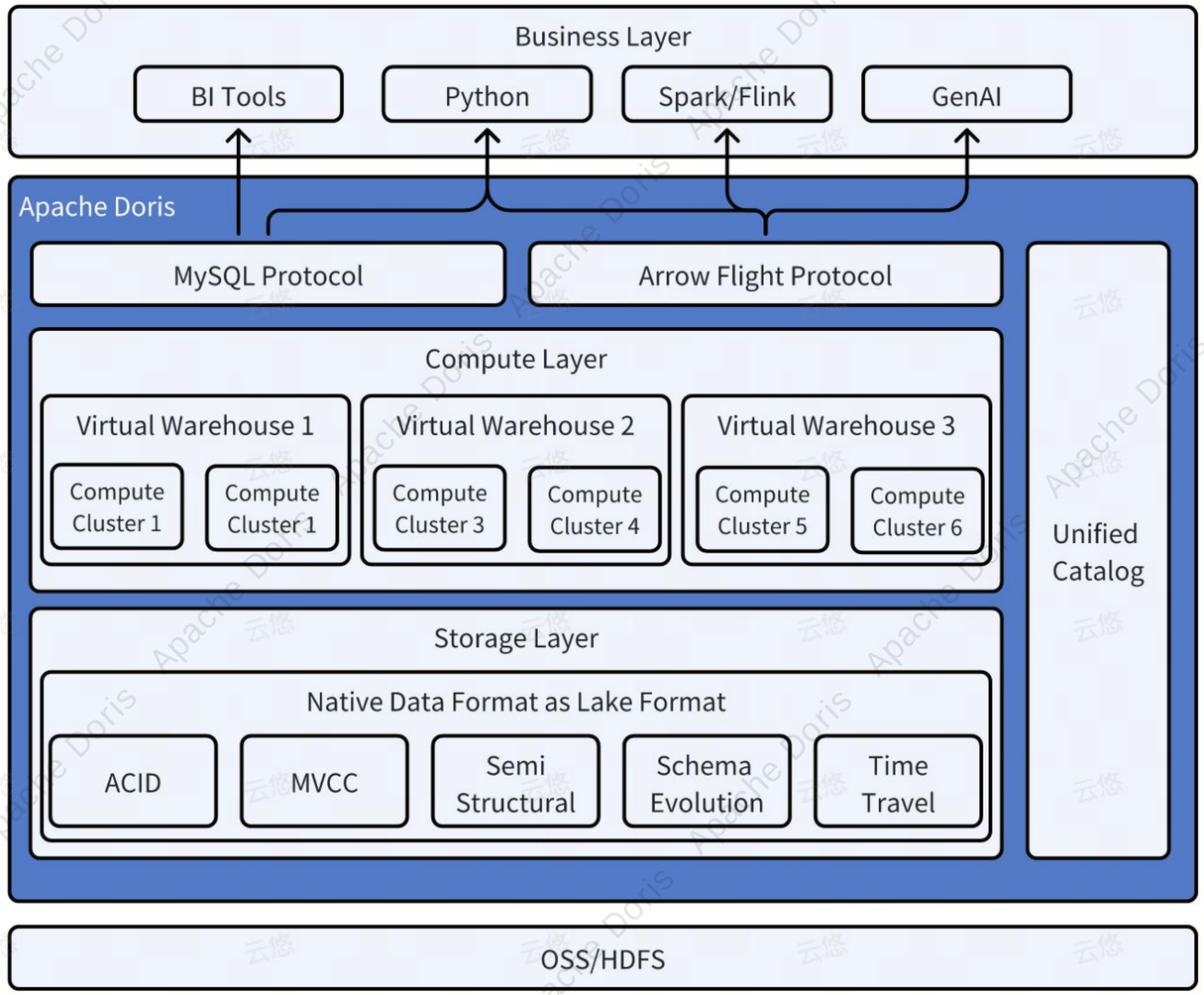
- 统一元数据管理
- 数据湖写回
- 物化视图和数据分层加工



3.0 特性 - Lakehouse 再进化

阶段3：湖仓一体

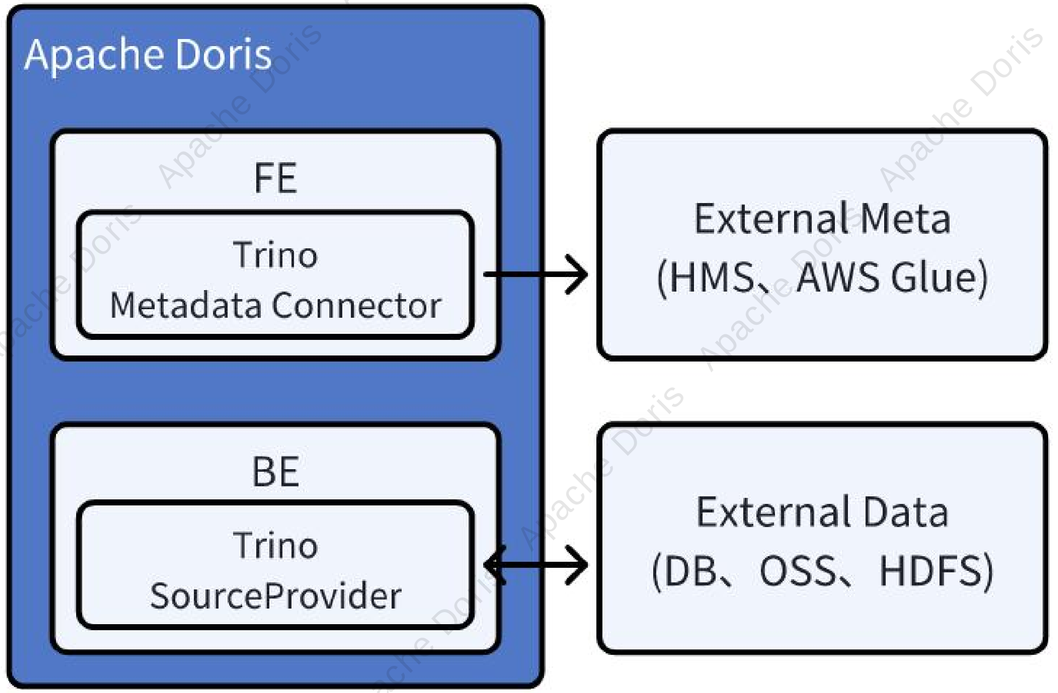
- 内外表特性的统一
- 存算分离架构



3.0 特性 - Trino Connector 兼容

即插即用、快速对接

- DeltaLake
- BigQuery
- Kudu
- Redis
- Kafka
- ...



参考文档：https://doris.apache.org/community/how-to-contribute/trino-connector-developer-guide

3.0 更多特性

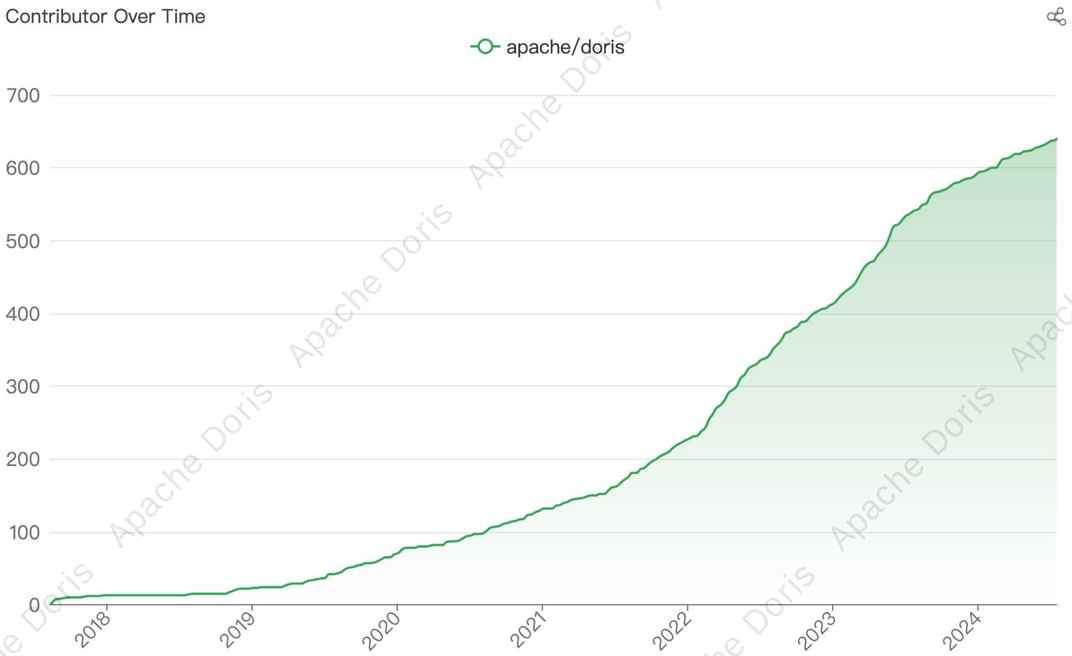
- Runtime Filter 自适应能力增强
- 查询算子落盘
- 显式事务支持
- Routine Load 写入优化
- Variant 性能和易用性增强
- 倒排索引性能进一步提升

目录

全球大数据和数据库领域最活跃的开源社区之一

*统计时间：截止2024年3月

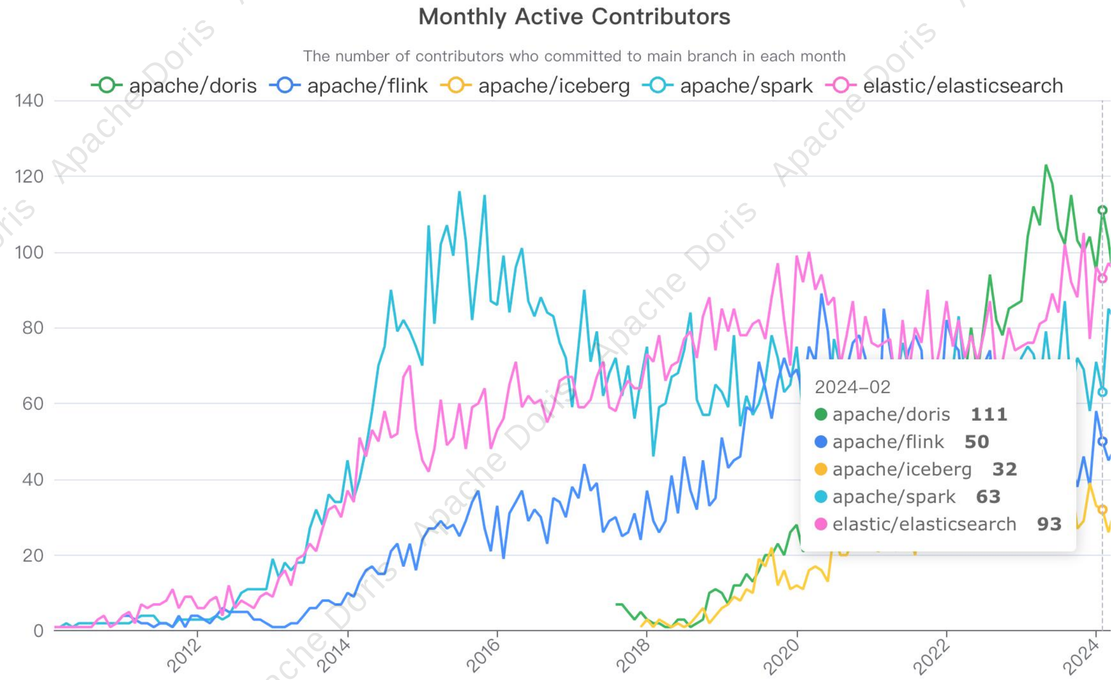
累计贡献者



650+

累计贡献者**已经超过650人**，距上一年度新增贡献者**超过100**，并仍处于持续上升的态势。

活跃贡献者



Top1

自2022年7月起至今，一直稳居在全球大数据开源项目排行中**活跃贡献者数Top1**

获得全球超过5000家中大型企业的信赖，广泛应用于核心线上分析场景

金融

中国银行 BANK OF CHINA
中国邮政储蓄银行 POSTAL SAVINGS BANK OF CHINA
中国平安 PING AN
中国光大银行 CHINA EVERBRIGHT BANK
兴业银行 INDUSTRIAL BANK CO.,LTD.
银联商务 UMS
中信证券 CITIC SECURITIES
中金财富 CICC Wealth Management
国金证券 SINOLINK SECURITIES
财通证券 CAITONG SECURITIES
无锡锡商银行 WUXI XISHANG BANK
无锡农村商业银行 WUXI RURAL COMMERCIAL BANK
金融壹账通 ONECONNECT
中国平安人寿保险 PING AN LIFE INSURANCE OF CHINA
招商信诺 Cigna & CMB
泰康 Taikang
众安保险
嘉实基金 Harvest Fund
博时基金 BOSERA FUNDS
汇添富基金
拉卡拉
易生支付 Easy Pay
招联
马上消费 WWW.MSXF.COM
杭银消费金融公司
HCFC 幸福消费金融
FUSE
度小满
奇富科技 QIFU TECHNOLOGY
同程数科
WeLab
腾棱科技 TECHSIGHT
趣米网络 InRice.com
百融云创 Bairong.Inc.

互联网

Baidu 百度
美团
xiaomi
Tencent 腾讯
JD.COM
Alibaba.com
360
HUAWEI
kwai
网易 NETEASE
ByteDance
sina
搜狐 SOHU.com
拼多多
知乎
汽车之家
有道 youdao
一点资讯
叮咚买菜
Soul APP
探探
TT语音
huya虎牙
Moka
斗鱼 DOUYU.COM
keep
奇安信
BOSS直聘
观测云 GUANCE.COM
AISPEECH
新东方在线
ADVANCE INTELLIGENCE GROUP
途虎养车
LeetCode
BOTON 宝通
中仑网络 ZHONG LUN NETWORK
Linkedcare 领健
十方融海
lifewit
曙光云
度言软件 DUYANSOFT
慧策
Letv
天眼查 TianYanCha.com
华数 wasu
蛟龙云
尚硅谷 www.atguigu.com

电信

中国电信 CHINA TELECOM
中国移动 China Mobile
China unicom 中国联通
HUAWEI
ZTE 中兴
cisco
天翼云
翼支付
浩瀚深度 HAOHAN Data

游戏

miHoYo
网易游戏
IGG
电魂 electronic soul
三七互娱 37 Interactive Entertainment
竞技世界
玩心不止 WAN XIN BU ZHI
冰川网络 WWW.QT.COM
动游科技

交通物流

航旅纵横 Umetrip.com
ZTO 中通快递
圆通速递
CAINIAO 菜鸟
JDL 京东物流
uda
德邦快递 DEPPON EXPRESS
满帮
aftership
货拉拉 HuoLaLa
ORANGE 橙 CONNEX 联
纵腾集团
物昂云通
Bondex 海程邦达
哪吒科技

零售快消

海底捞
luckin coffee
澳优 Ausnutria
雨润集团 YURUN GROUP
永辉超市
百胜 YumChina
ANTA
特步
Cabbeen
Semir
SEPTWOLVES
PEACEBIRD

能源制造

Lenovo
中国石油
Goldwind
Midea 美的
ENVISION 远景能源
地平线 Horizon Robotics
Ford
长安汽车 CHANGAN AUTO
VOLVO
长城汽车 Great Wall
CHERY
LOTUS NYO
极越
WULING
爱玛
SUZUKI

加入 Apache Doris 社区任何时候都不晚

- **订阅开发者邮件组**

  订阅社区开发者邮件组 dev@doris.apache.org 并参与讨论

- **Doris 小助手**

  如想进入社区用户社群请备注 **加群**

Doris 小助手

扫一扫上面的二维码图案，加我为朋友。

Thanks !