

Apache Doris: 现代化数据仓库的创新之路

马如悦

Apache Doris 创始人 & PMC 成员

十年磨一剑，打造全球最知名开源软件基金会 Apache 的顶级项目



Open-Source Real-Time Data Warehouse

Apache 所有项目中 官网每日浏览量 No.1

开源大数据、数据库项目中 月活开发者 No.1

2013 Project Creation

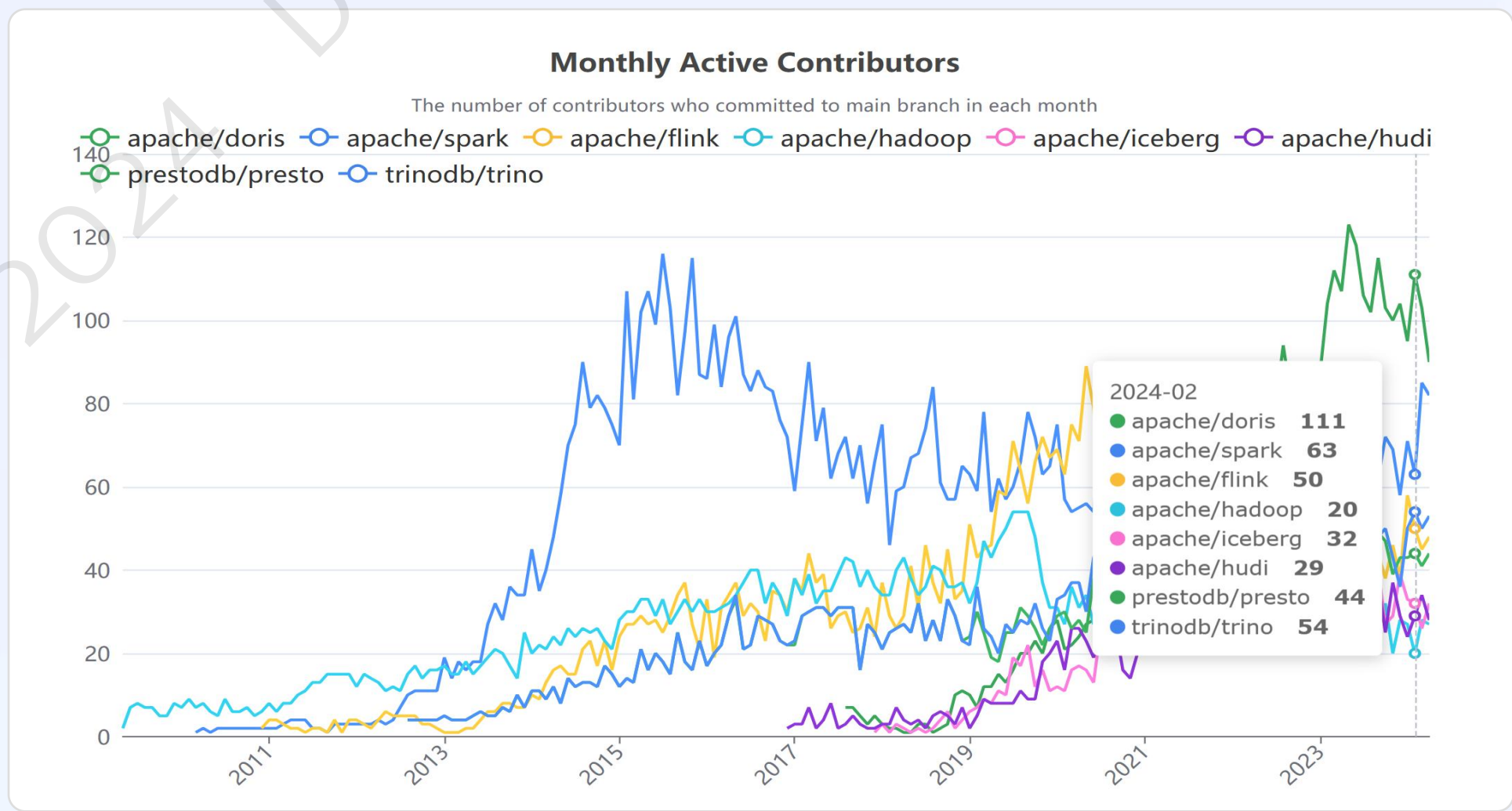
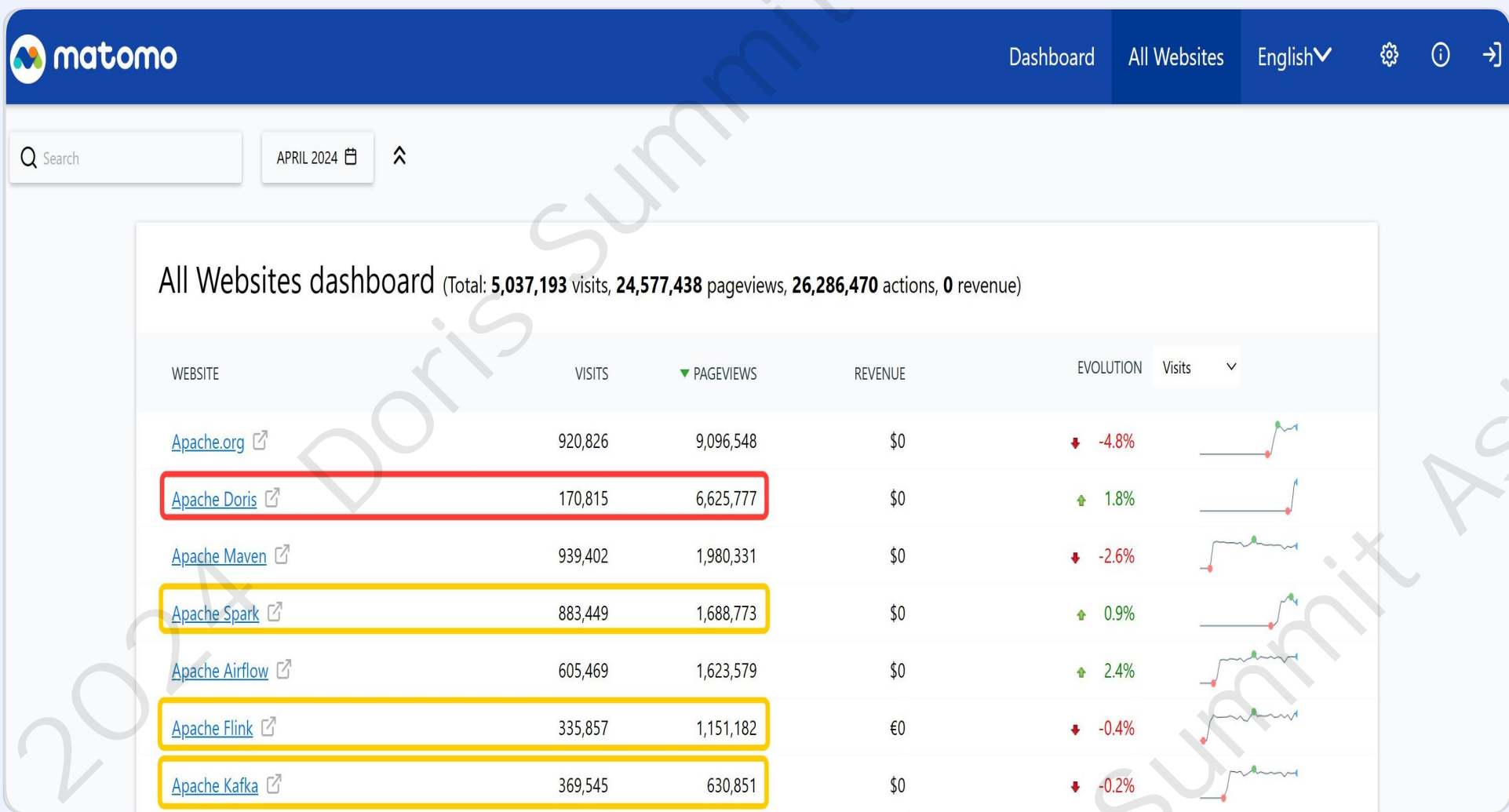
2017 Open Source

2022 ASF Top Project

13k GitHub Stars

600+ Contributors

5000+ Enterprises



超 5000+ 中大型企业使用

金融



互联网



电信



交通物流



能源制造



游戏



零售快消



Apache Doris 3.0，现代化数据仓库创新之路的重大里程碑

实时之路
Real-Time

统一之路
Unified

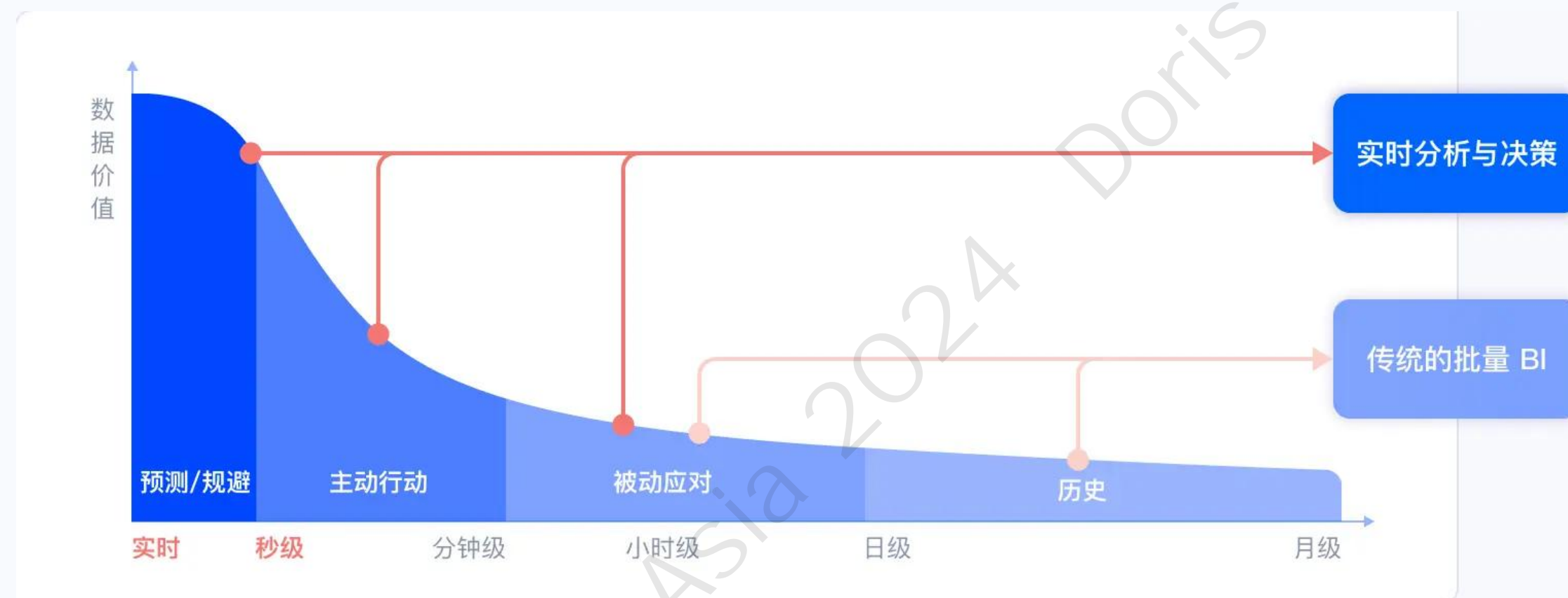
弹性之路
Elastic

Real-Time

实时之路

Apache Doris 实时分析特性

企业正全面进入实时分析的时代



来源: Perishable insights, Mike Gualtieri, Forrester

- 从跑批报表到实时仪表盘
- 从面向内部的分析到面向外部客户的分析
- 从预置报表到交互式即席查询
- 从面向人的分析到算法自动决策

实时数据 Fresh

秒级实时数据写入
确保数据新鲜度

极速分析 Fast

极速的交互式
分析性能

高并发查询 High-Concurrency

支撑超大规模用户的高
并发查询

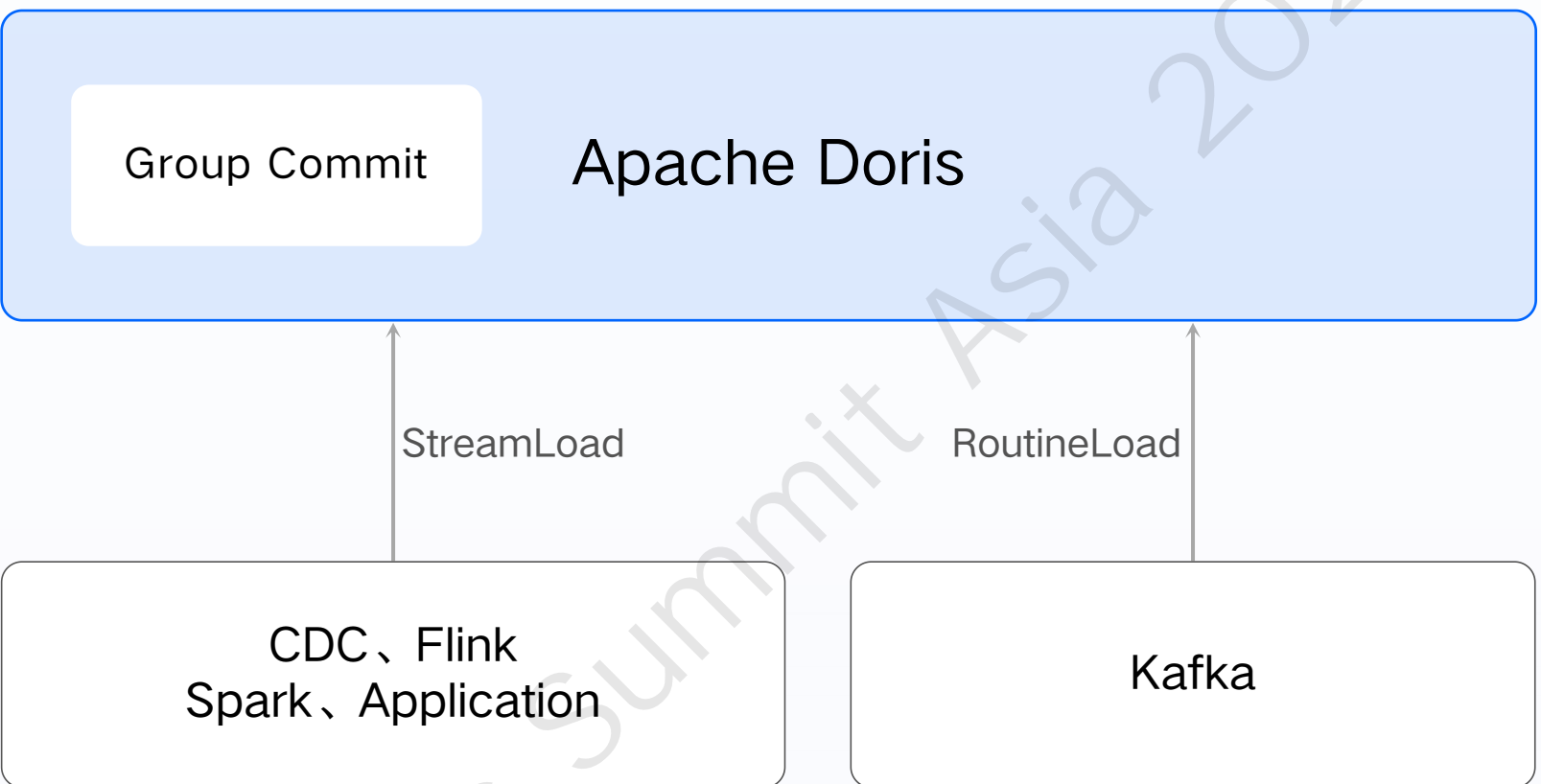
在线高可用 Online

高可用架构设计
满足在线服务要求

实时数据 Fresh Data

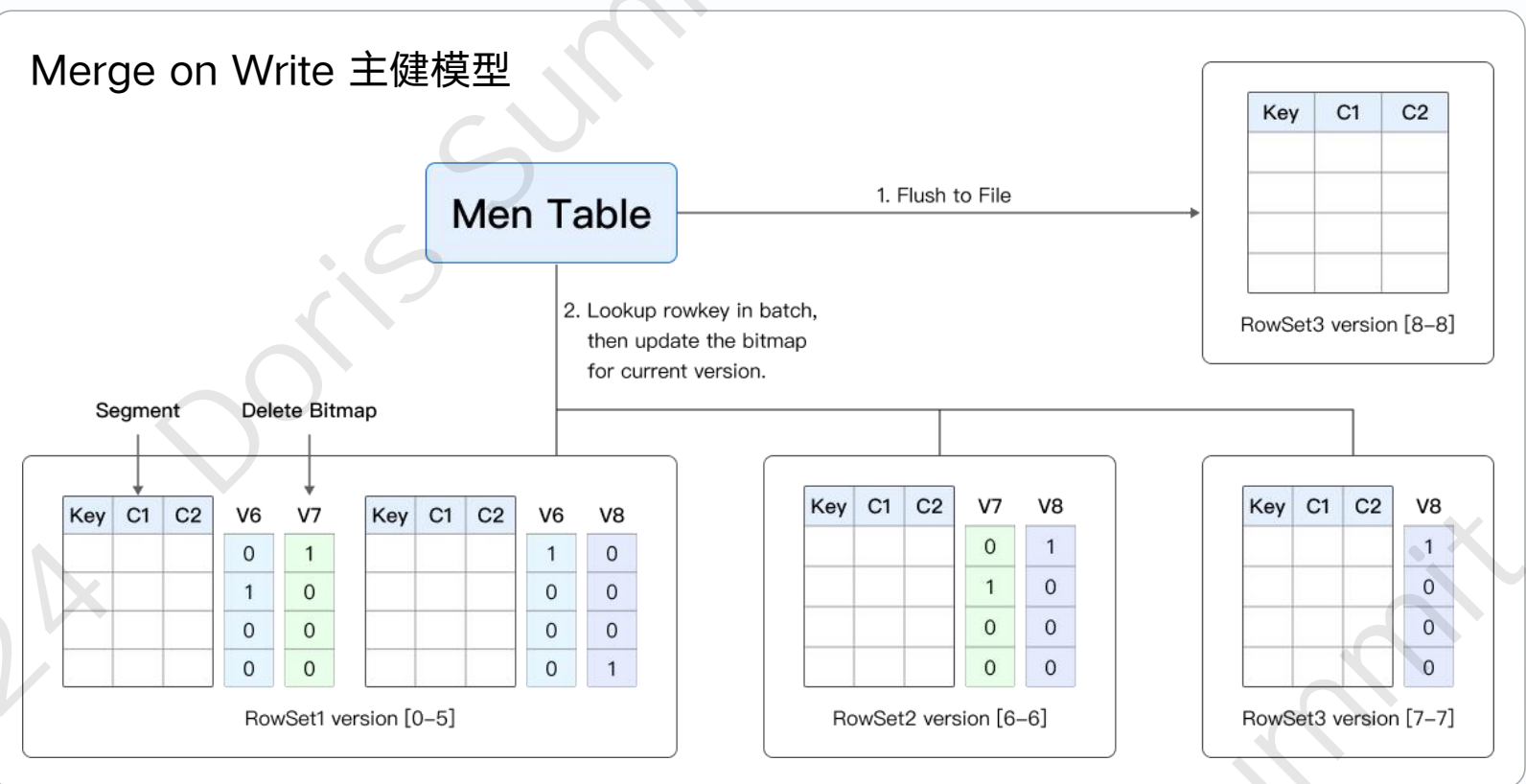
秒级实时数据写入

- 秒级别的 StreamLoad/Insert into 数据导入
- 支持服务端攒批 Group Commit
- 支持从 Kafka 自动拉取导入 RoutineLoad
- 支持数据库 CDC、Flink、Spark 等实时导入



高性能数据更新

- 提供 MoW 的主键模型
- 支持 UPSERT、条件更新、条件删除、部分列更新
- 基于 Sequence 列的并发更新事务控制



轻量级元数据变更

- 更改列名
- 增加和删除 Value 列
- 增加 VARCHAR 列（Value列）的长度

```
mysql> alter table lineitem add column L_NEW_COLUMN INTEGER NOT NULL DEFAULT '0';
Query OK, 0 rows affected (0.00 sec)

mysql> SHOW ALTER TABLE COLUMN WHERE TableName='lineitem' ORDER BY createtime DESC LIMIT 1;
```

JobId	TableName	CreateTime	FinishTime	IndexName	IndexId	OriginIndexId	SchemaVersion	Transaction
10824	lineitem	2022-11-03 12:19:48.808	2022-11-03 12:19:48.811	lineitem	10303	10303	7:1656234387	-1

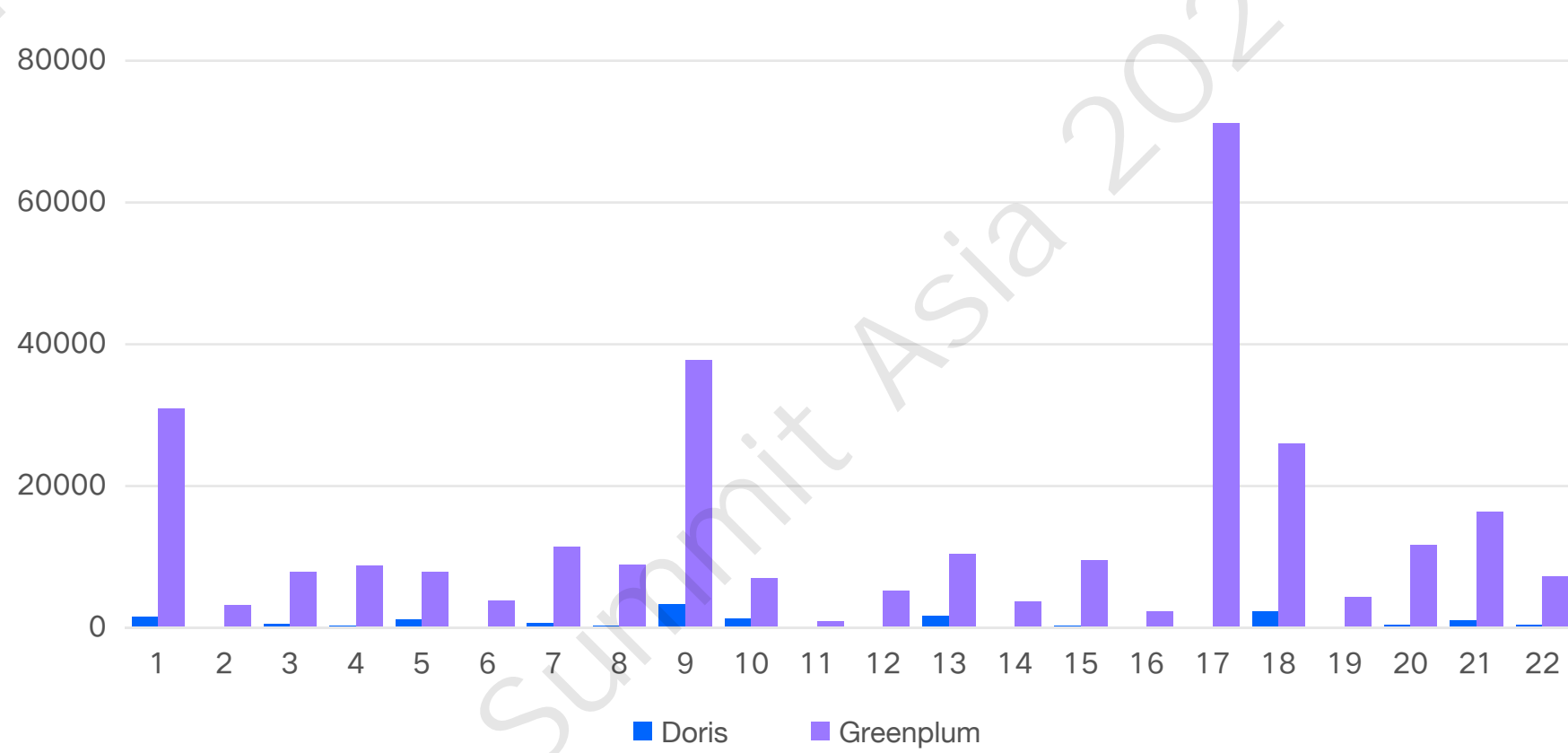
1 row in set (0.00 sec)

极速分析 Fast Analytics

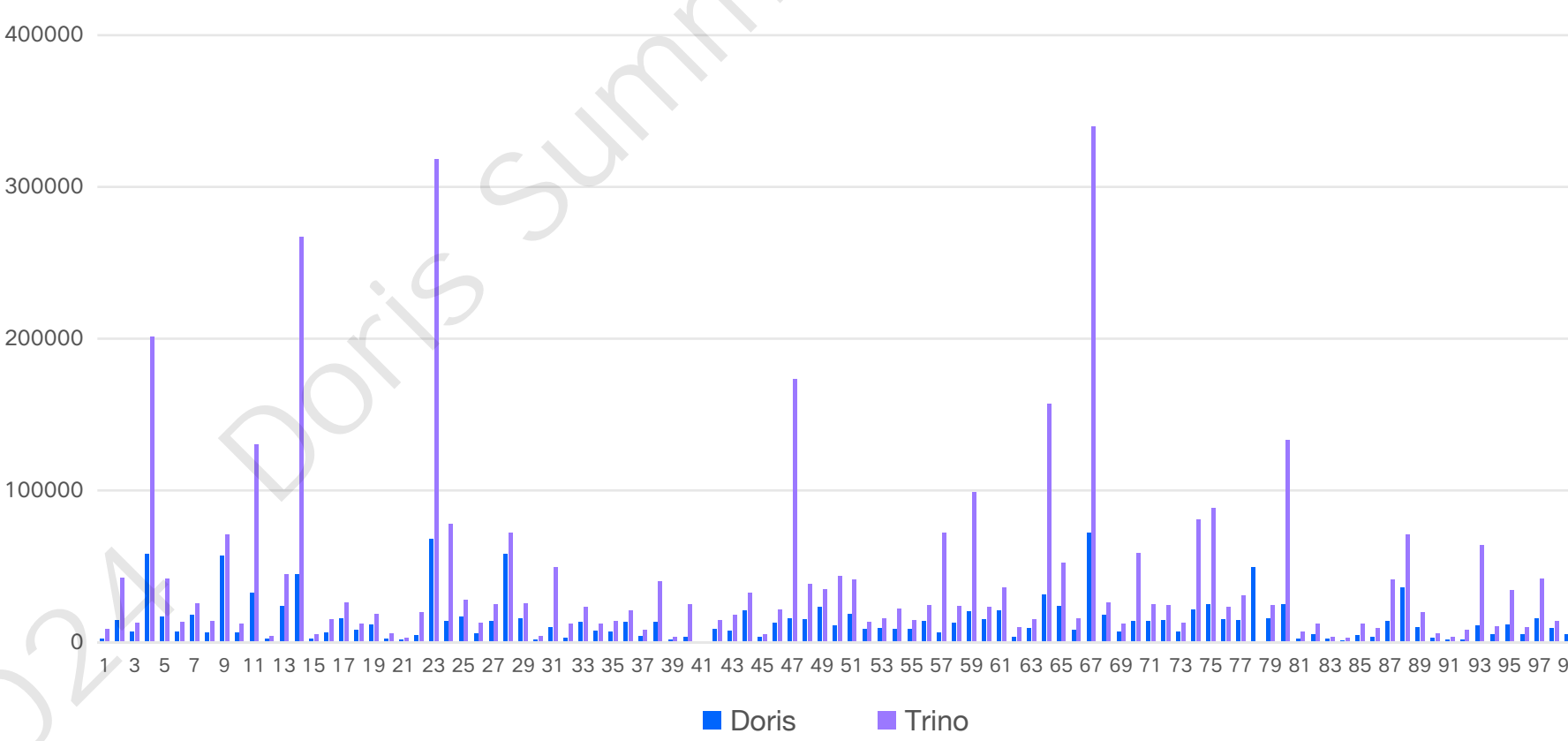
ClickBench

System & Machine	Relative time (lower is better)
Umbra (c6a.metal, 500gb gp2)	x 1.61
	x 1.95
ClickHouse (tuned) (c6a.metal, 500gb gp2)	x 2.04
Apache Doris (c6a.metal, 500gb gp2)	x 2.15
ClickHouse (c6a.metal, 500gb gp2)	x 2.21
	x 2.38
Umbra (c6a.4xlarge 500gb gp2)	x 2.40

TPC-H



TPC-DS



向量化

- 向量化处理，减少虚函数调用和 cache miss
- 高效利用 SIMD 指令，同时支持 x86 和 arm

CBO 优化器

- 基于代价的 join reorder, pushdown
- Runtime Filter

丰富的索引

- 跳数索引: bloom filter, min/ max/ sum
- 点查索引: prefix sorted index
- 任意维度检索: 倒排索引

物化视图

- 强一致的单表物化视图，支持通用聚合函数
- 单表和多表异步物化视图

ARM 架构优化

Doris v2.1 针对 ARM 架构做了深度优化，性能相比 V2.0 提升近一倍，是当前 ARM 架构下性能最好的开源数仓

高并发查询 High-Concurrency Queries

分区分桶

Partition[2019-01-01,2020-01-01)

Partition[2020-01-01,2021-01-01)

Partition[2021-01-01,2022-01-01)

Tablet A

Tablet B

Tablet C

Tablet D

- create_time 作为分区键、ID 作为分桶键
- select * from user_table where id = 5122 and create_time = '2022-01-01'

主键索引 & 倒排索引



SST

Dictionary entry

f(x):4

Inverted list

1	3	4	5
1	2	1	1
1	2	1	1
4	2	7	2
2	2	2	2

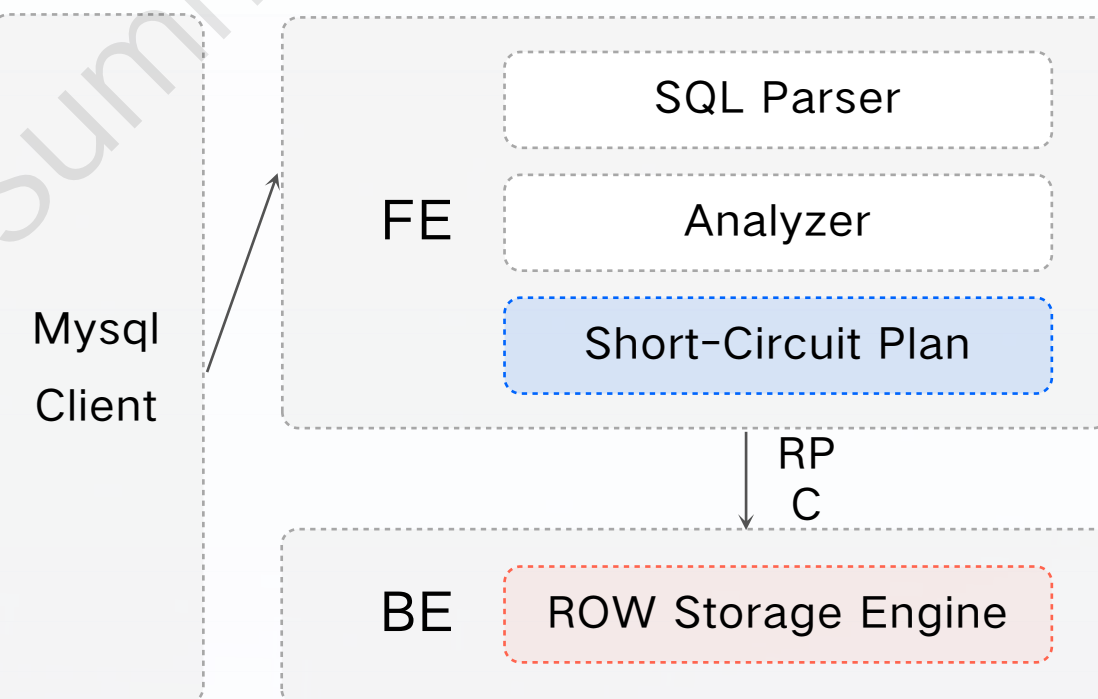
```
CREATE TABLE sales_order (  
  orderid BIGINT,  
  status TINYINT,  
  username VARCHAR (32),  
  amount BIGINT DEFAULT '0',  
  INDEX idx_user_name (username) USING  
  INVERTED  
)  
UNIQUE KEY orderid)  
DISTRIBUTED BY HASH(orderid) BUCKETS 10
```

行列混存

Row-based					Column-based				
Row ID	DateTime	Material	Customer	Qty	Row ID	DateTime	Material	Customer	Qty
1	845	2	3	1	1	845	2	3	1
2	851	5	2	2	2	851	5	2	2
3	872	4	4	1	3	872	4	4	1
4	878	1	5	2	4	878	1	5	2
5	888	2	3	3	5	888	2	3	3
6	895	3	4	1	6	895	3	4	1
7	901	4	1	1	7	901	4	1	1

- 支持行存和列存
- 列存用于高吞吐分析
- 行存用于高并发点查

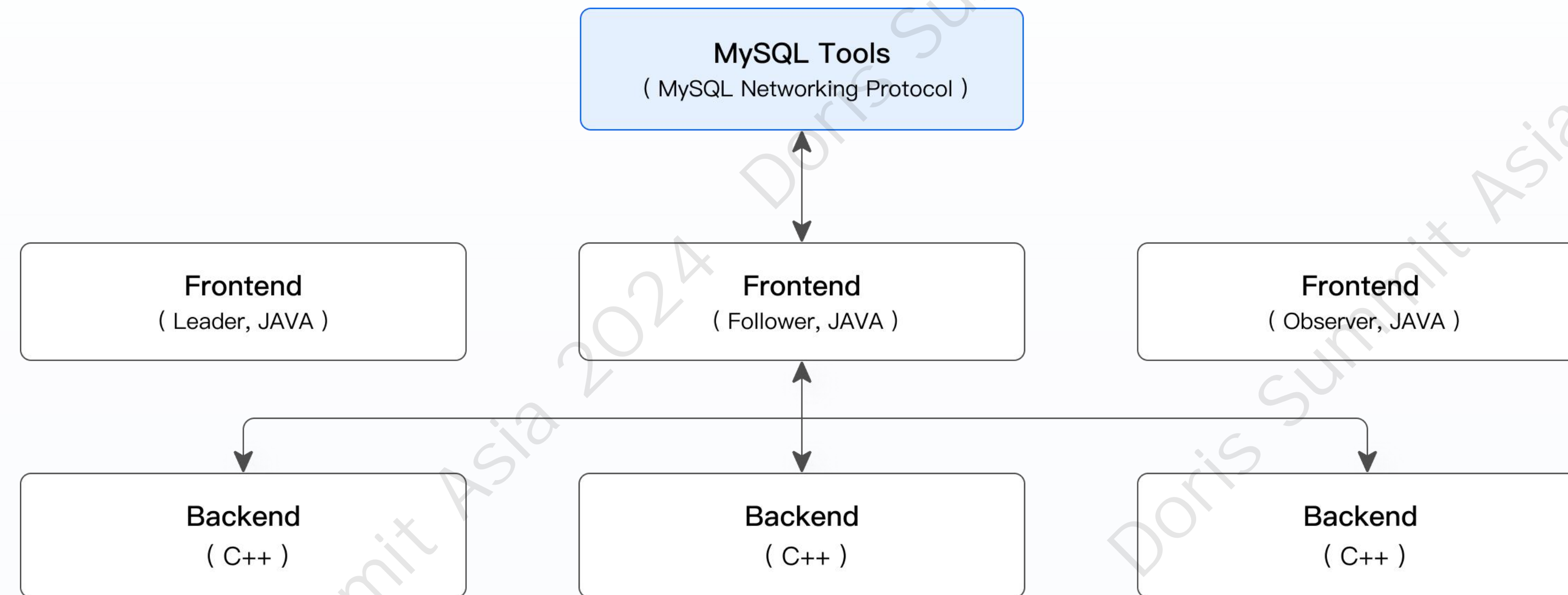
点查优化



- 点查询短路径优化 (Short-Circuit)
- 预处理语句优化 (Prepared Statement)

单节点最高上万 QPS

在线高可用 High Availability Online System

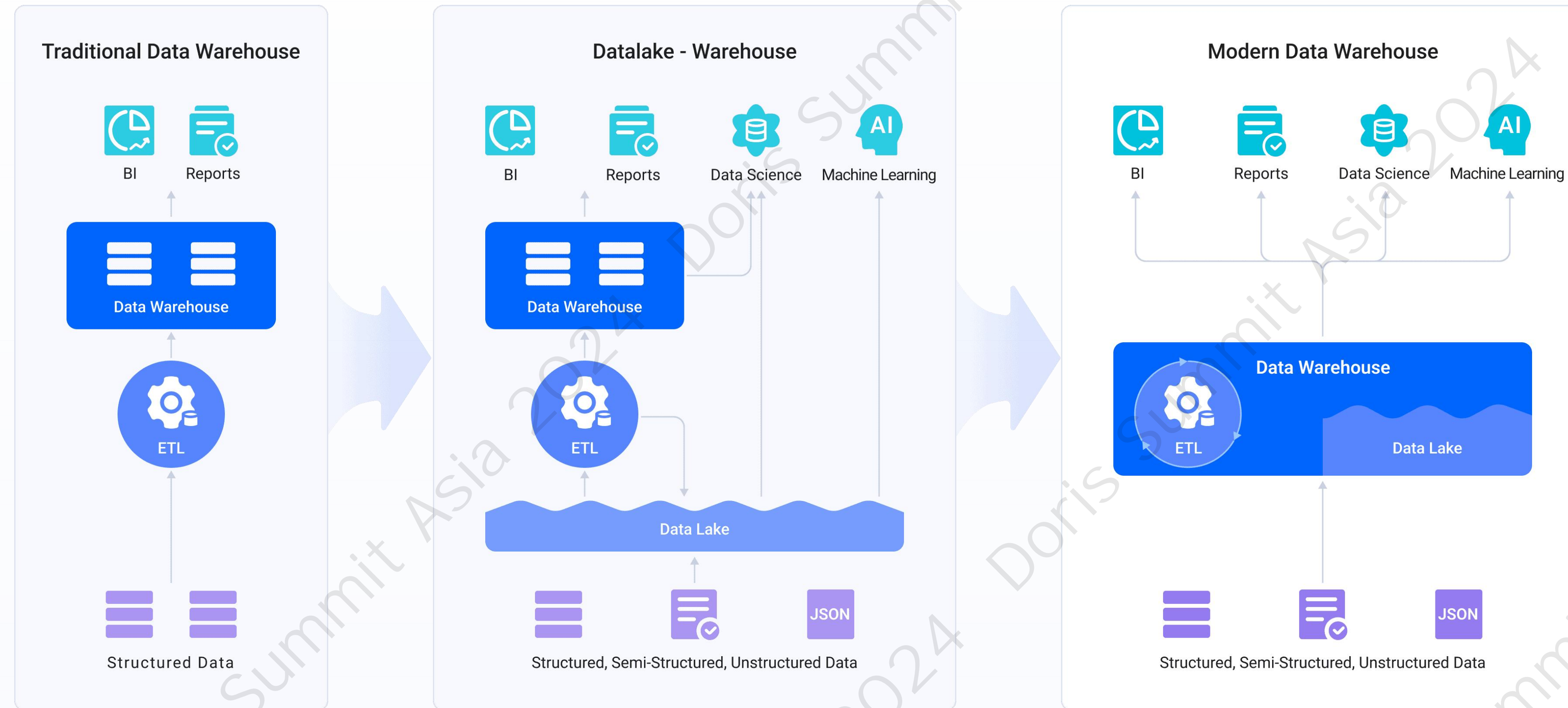


- 元数据节点 FE 和数据节点均无单点
- 支持在线扩容、滚动升级，数据自动均衡
- 支持 Online Schema Change
- 支持动态增删索引
- 3.0 存算分离支持多计算组

Unified

统一之路

单一系统支持各种分析负载



统一实时和批量计算

不仅支持实时分析，也支持批量计算

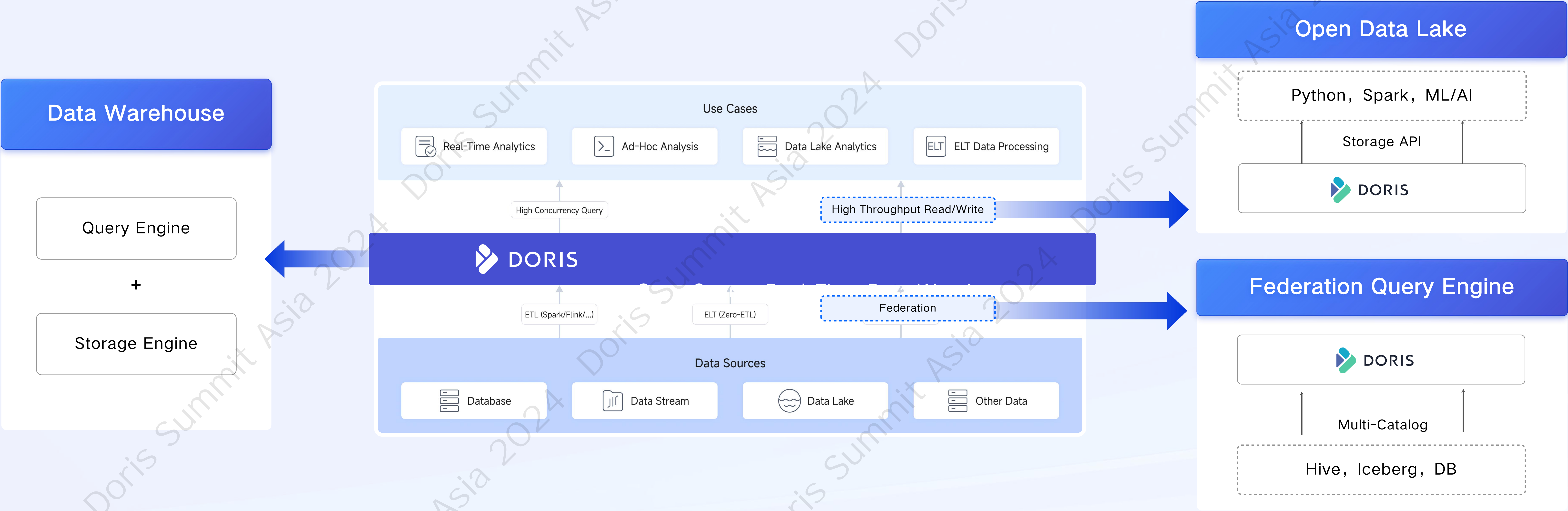
统一数据仓库和数据湖仓

既可以作为Warehouse，也可以作为Lakehouse

半结构化数据分析

不仅支持结构化数据分析，也支持半结构化数据分析

湖仓无界：统一数据仓库和数据湖仓，避免过早复杂性



半结构化数据分析

复合数据类型

- Array
- Map
- Struct
- JSON

```
CREATE TABLE `semi_table` (  
  `id` int(11) NULL,  
  `c_array` ARRAY<int(11)> NULL,  
  `c_map` Map<STRING, INT> NULL,  
  `c_struct` STRUCT<s_id:int(11), s_name:string, s_address:string> NULL,  
  `c_json` JSON  
)
```

Variant

- 可以支持任意类型、任意形状的 JSON 格式数据
- 自动推断类型进行列式存储，自动处理列增加、类型变更
- 不需要繁琐的 DDL 操作以及 Schema Change 操作

```
CREATE TABLE IF NOT EXISTS ${table_name} (  
  k BIGINT,  
  v VARIANT  
)  
  
SELECT v["properties"]["title"] from ${table_name}
```

	Storage Space		
Pre-Defined Static Columns	12.618 GB		
Variant Type	12.718 GB		
JSON Type	35.711 GB		
	First Time (Cold Run)	Second Time (Hot Run)	Third Time (Hot Run)
Pre-Defined Static Columns	233.79s	86.02s	83.03s
Variant Type	248.66s	94.82s	92.29s
JSON Type	Mostly OOM	789.24s	743.69s

String 类型

- 高效文本匹配（Like）：NGram BF，高性能正则匹配和子串匹配算法
- 文本分词，全文检索（Match）：倒排索引实现快速检索

str LIKE 'abc'	equals(str, 'abc')
str LIKE 'abc%'	starts_with(str, 'abc')
str LIKE '%abc'	ends_with(str, 'abc')
str LIKE '%abc%'	sub_string(str, 'abc')
str LIKE '%a%b%c'	regex(str, '.*a.*b.*c')

```
-- search for request contains word 'login'  
SELECT * FROM httplogs WHERE request MATCH 'login';  
  
-- search for request contains word 'login' or 'error'  
SELECT * FROM httplogs WHERE request MATCH 'login error';  
  
-- search for request contains word 'login' and 'error'  
SELECT * FROM httplogs WHERE request MATCH_ALL 'login error';
```

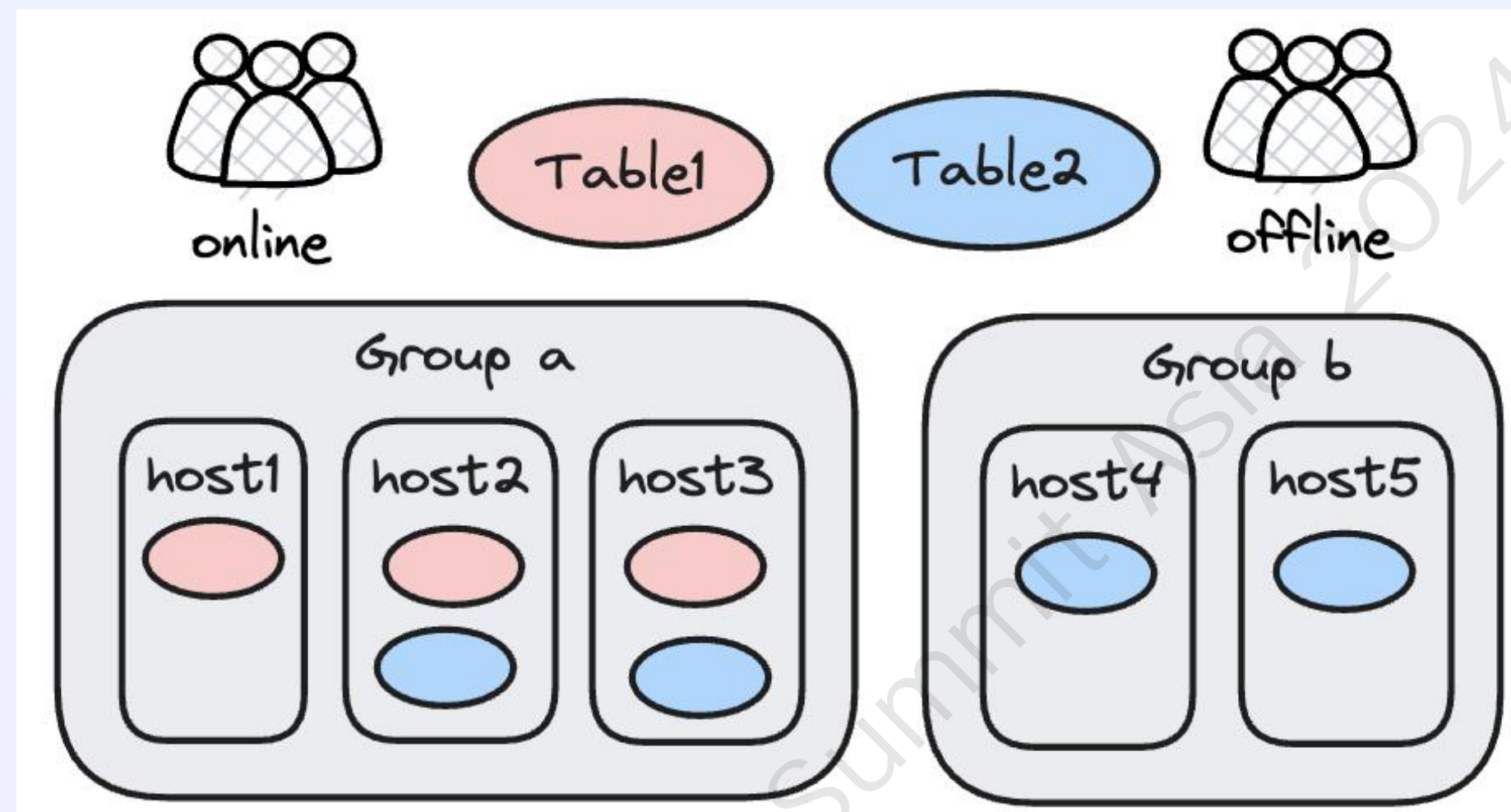

Elastic

弹性之路

存算一体时代的弹性资源管理

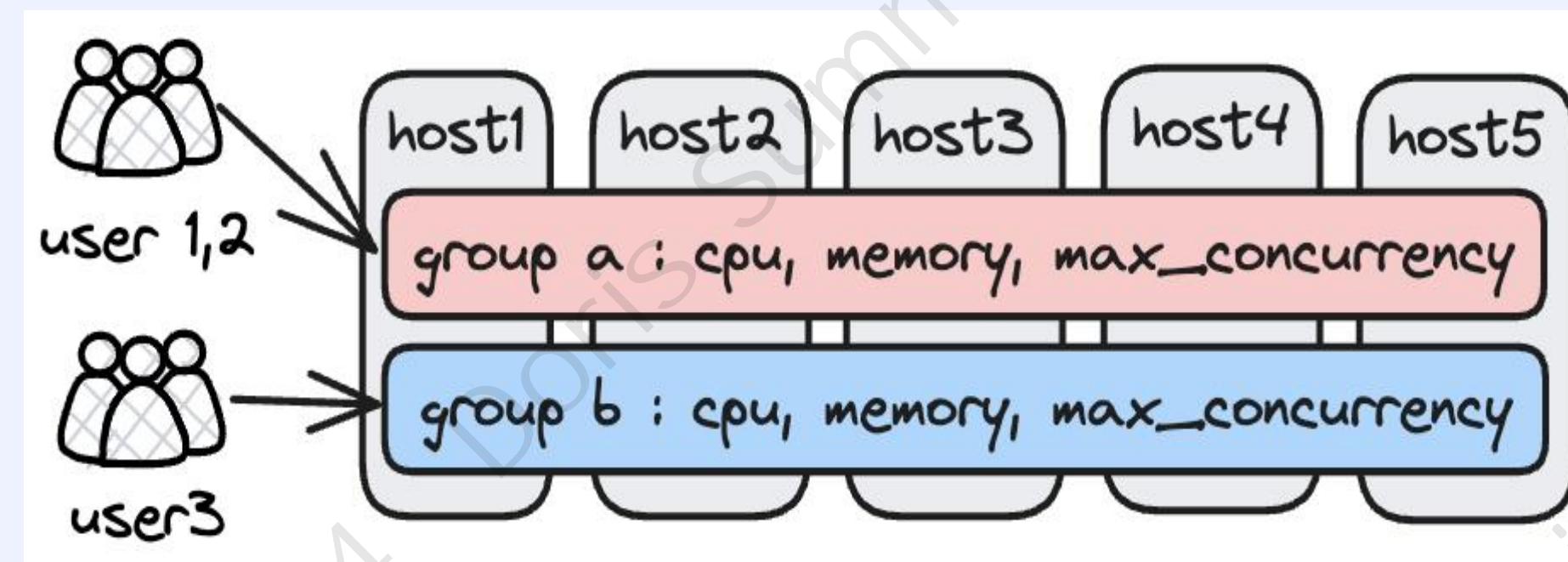
Resource Group

机器分组，副本放置到分组，user 绑定分组



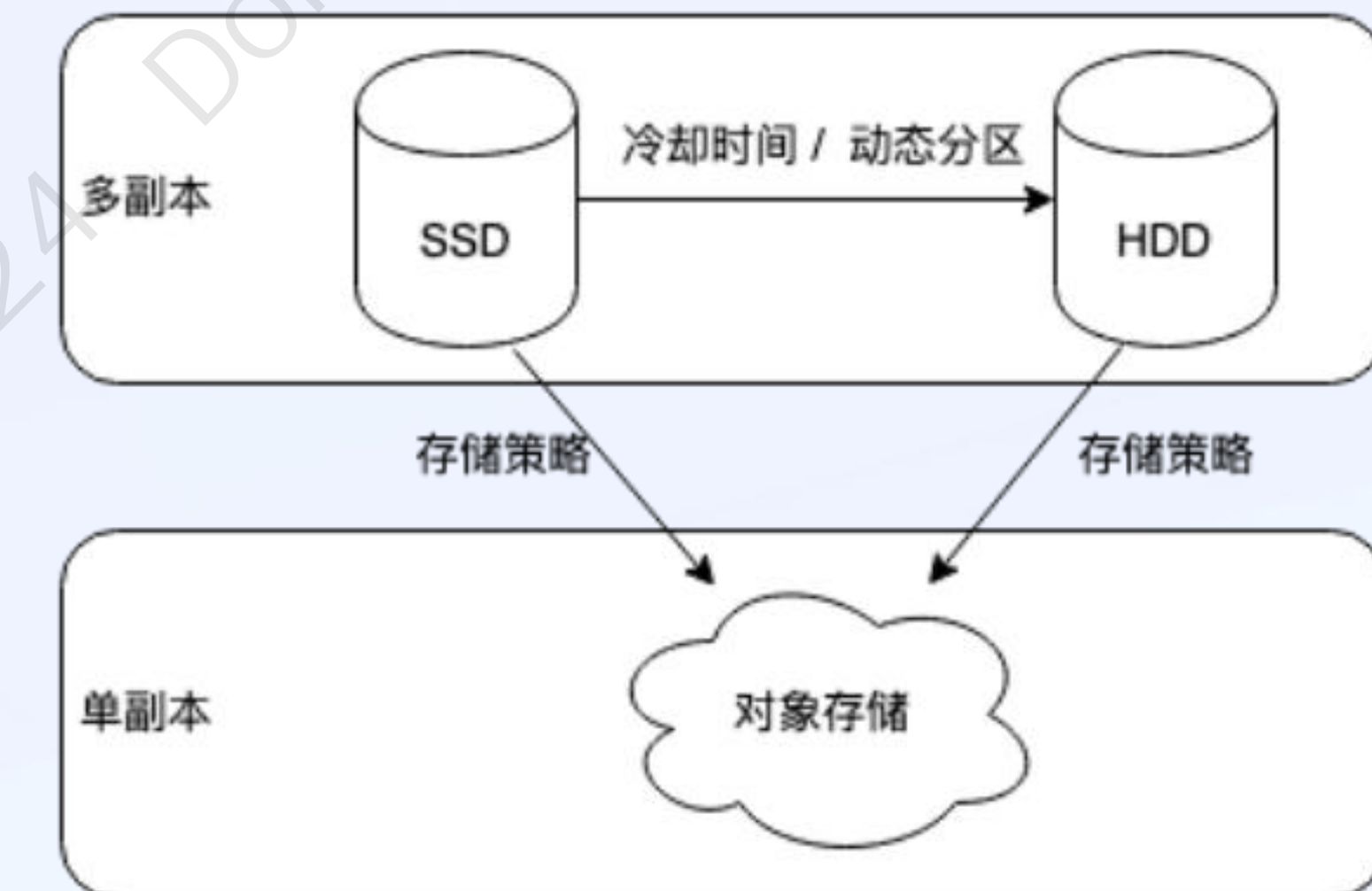
Workload Group

支持资源软硬限制



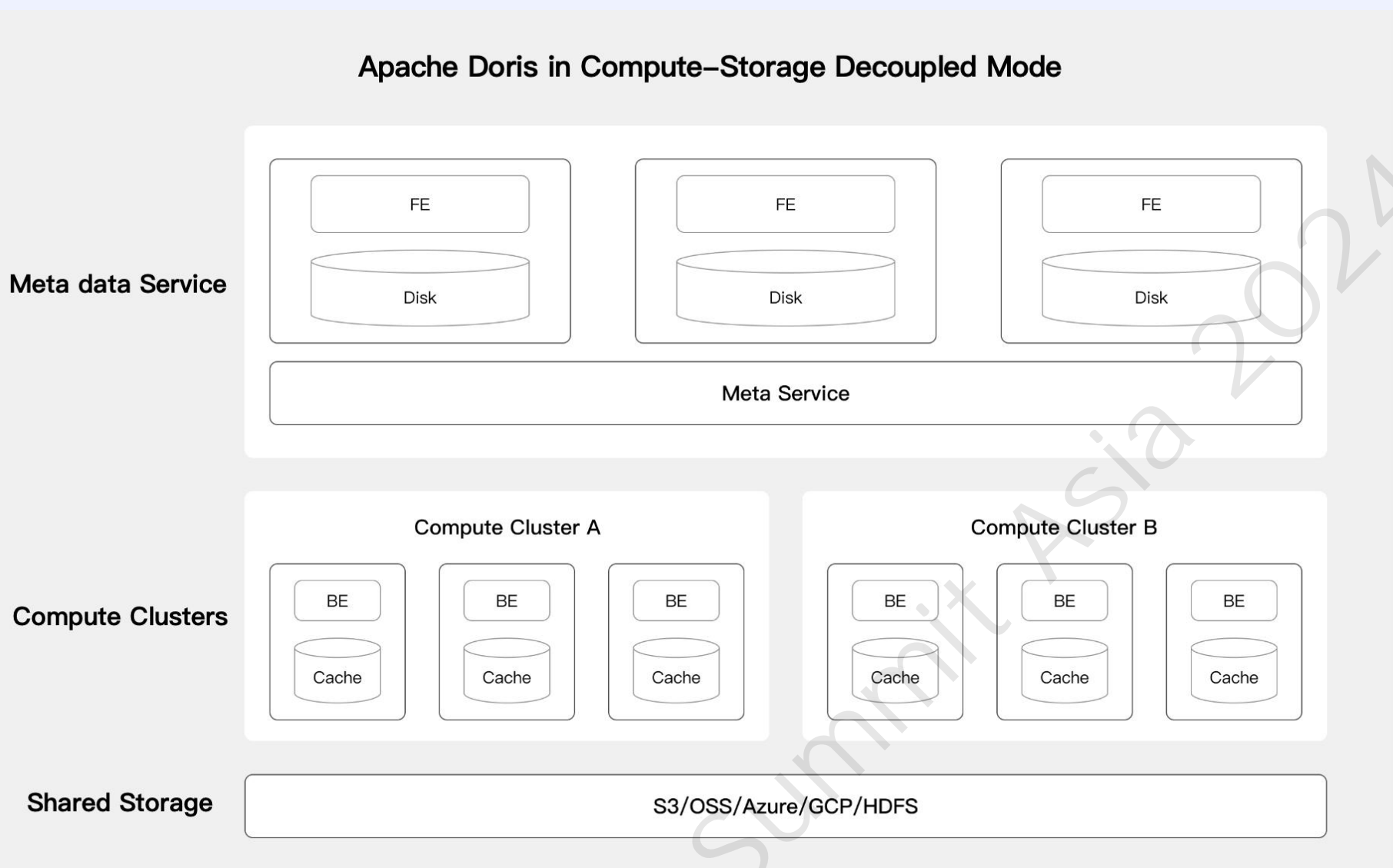
冷热分层 (Tiered Storage)

冷热分层，将冷数据存储在更低成本的存储上

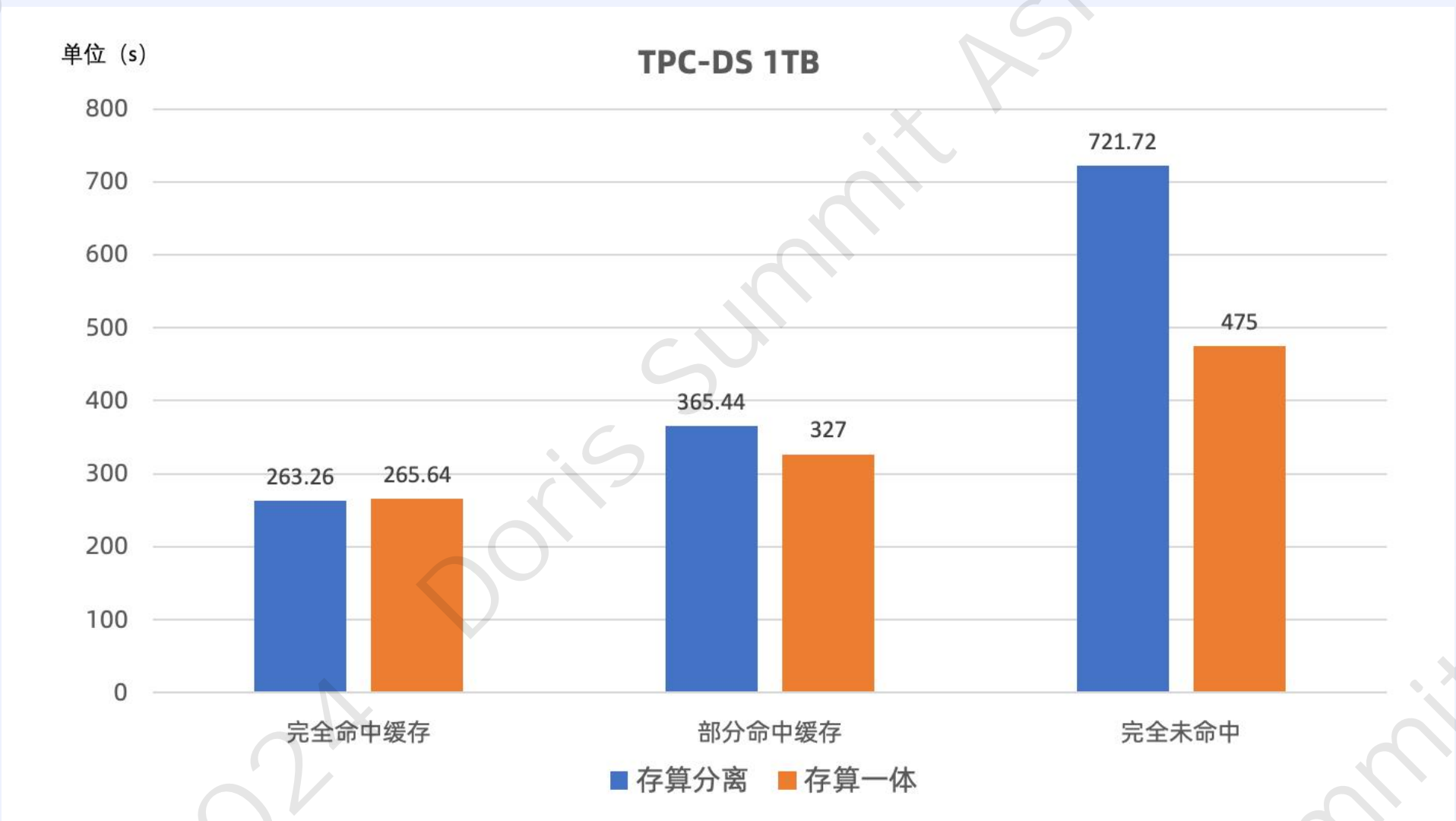


Apache Doris 3.0 部署新形态 - 存算分离

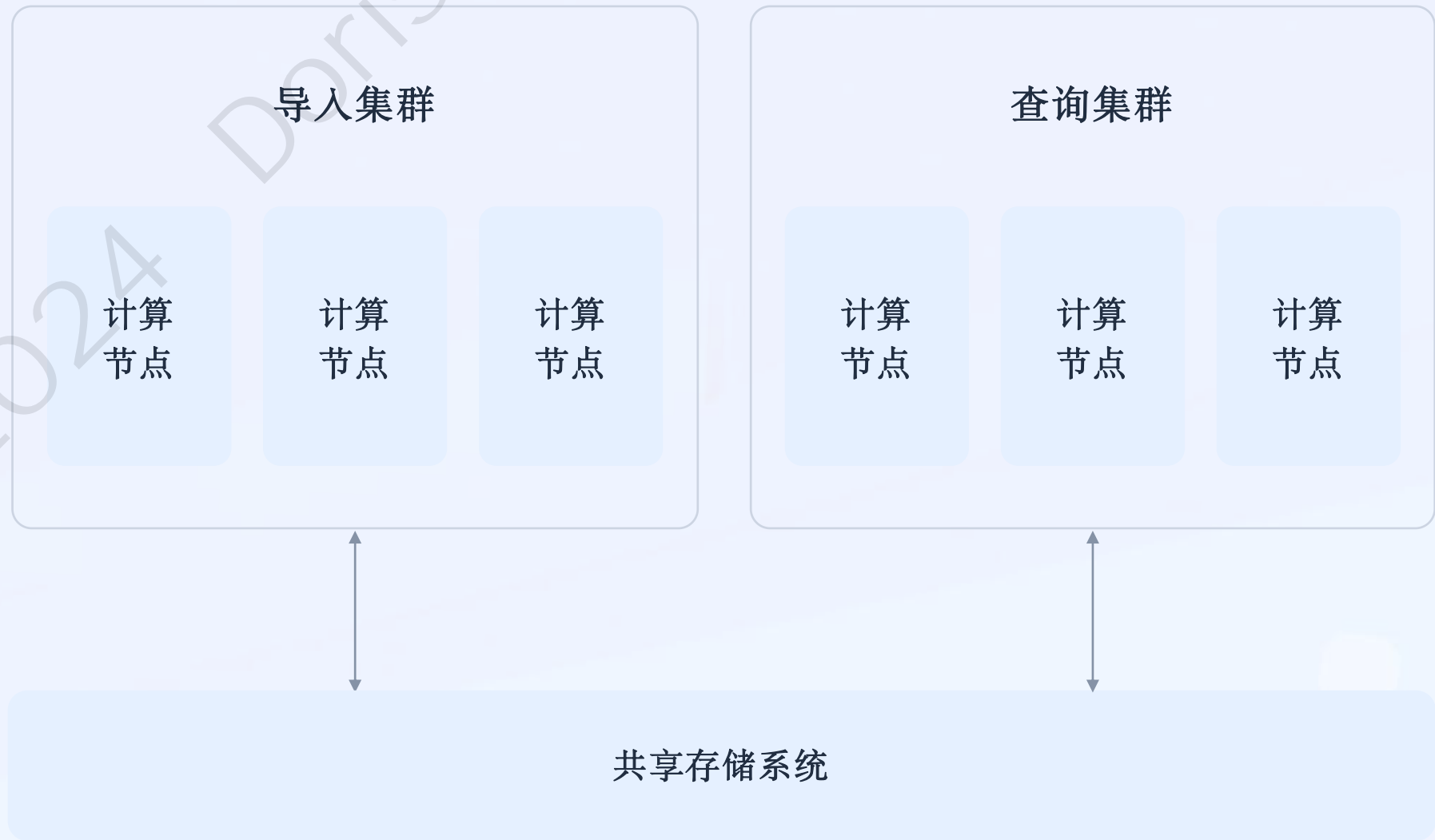
存算分离架构



共享存储与本地高速缓存



多计算集群

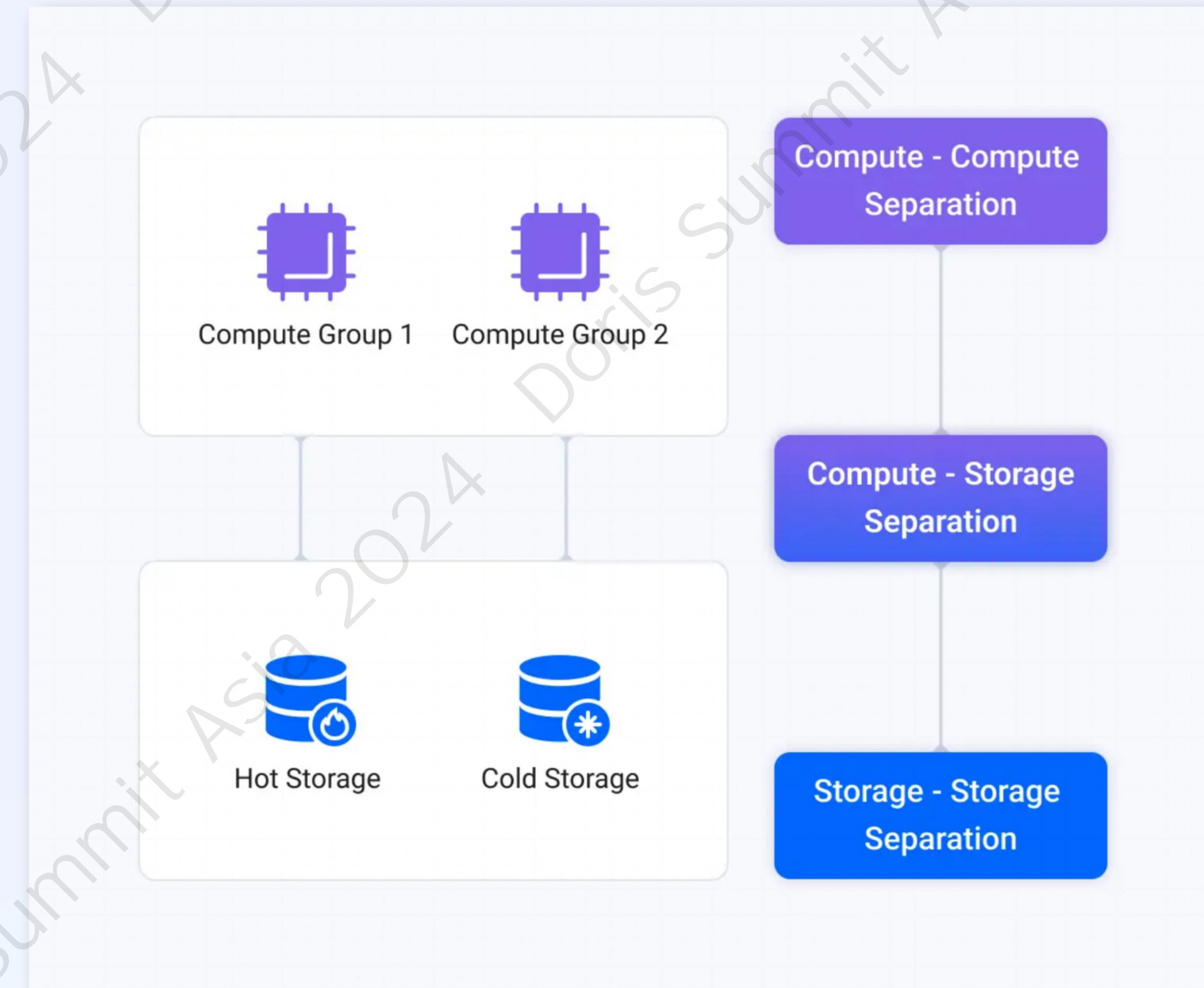


弹性的资源管理是数据分析基础设施永恒的追求

存算一体与存算分离各有优势

- 存算一体部署简单、性能较优
- 存算分离，独立扩缩容，部署需要依托高性能存储和网络
- 存算分离未必云原生、云原生也未必是存算分离，但是存算分离+云原生是绝配

两种部署形态的融合



Future Work

未来工作

新功能

内置 CDC 同步

可以不依赖外部工具，支持从众多TP数据库直接CDC导入数据，打造 HTAP Solution。

支持增量批量处理

统一实时和批量处理。增量处理，需要 Doris支持增量读取表的更新数据。

完善湖仓一体

插件化体系，兼容Trino/Presto Connector 框架；完善高吞吐读写的Data API。

存算一体和存算分离部署形态融合

不再需要两种部署形态，用户可以在使用过程中，无缝切换。避免过早复杂性。

新功能

非功能性能力的加强

更优

提升性能

导入更实时，查询更极速

更稳

提升稳定性

版本迭代机制改进，加强测试覆盖，内核插件化

更易用

易用易运维

简化参数配置，提升产品文档，完善周边生态工具

Thanks for Watching!