开源社
kaiyuanshe

StarRocks
湖仓一体新范式

王欢明 - StarRocks Committer

StarRocks

目录

CONTENTS

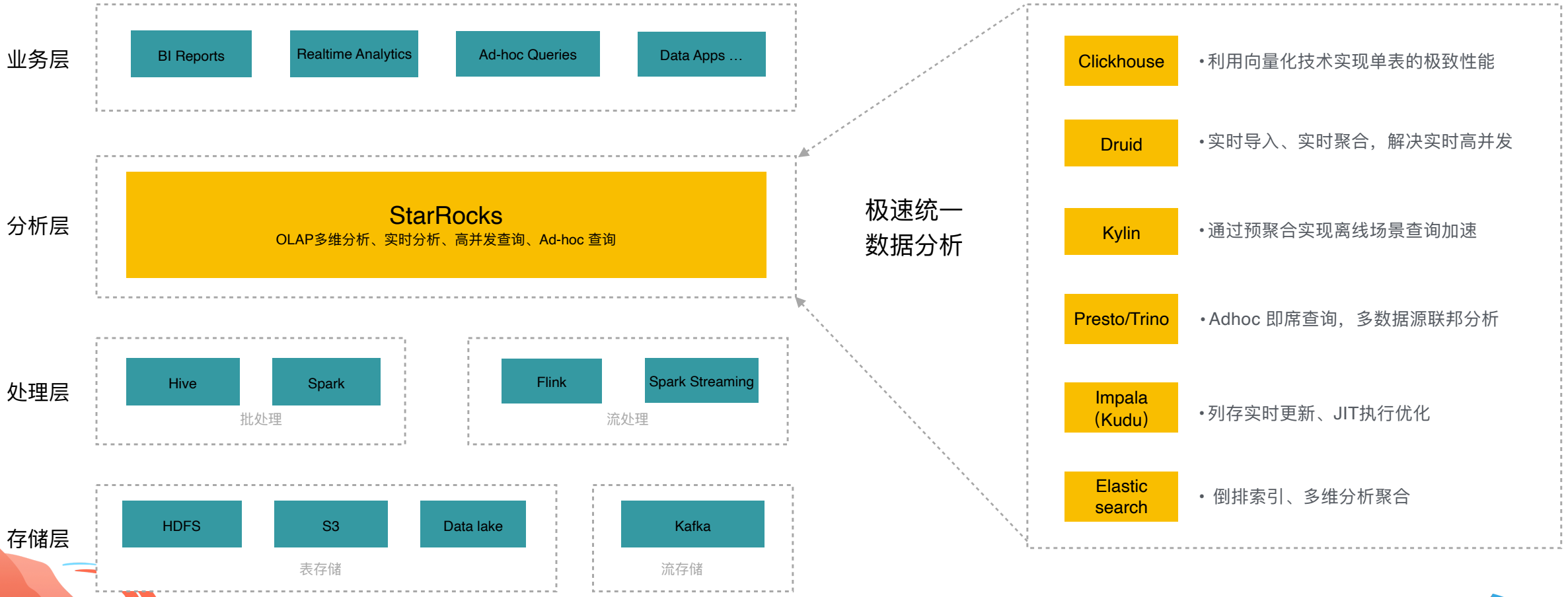
开源社
kaiyuanshe

01

StarRocks LakeHouse

StarRocks

StarRocks 产品定位



StarRocks 社区发展

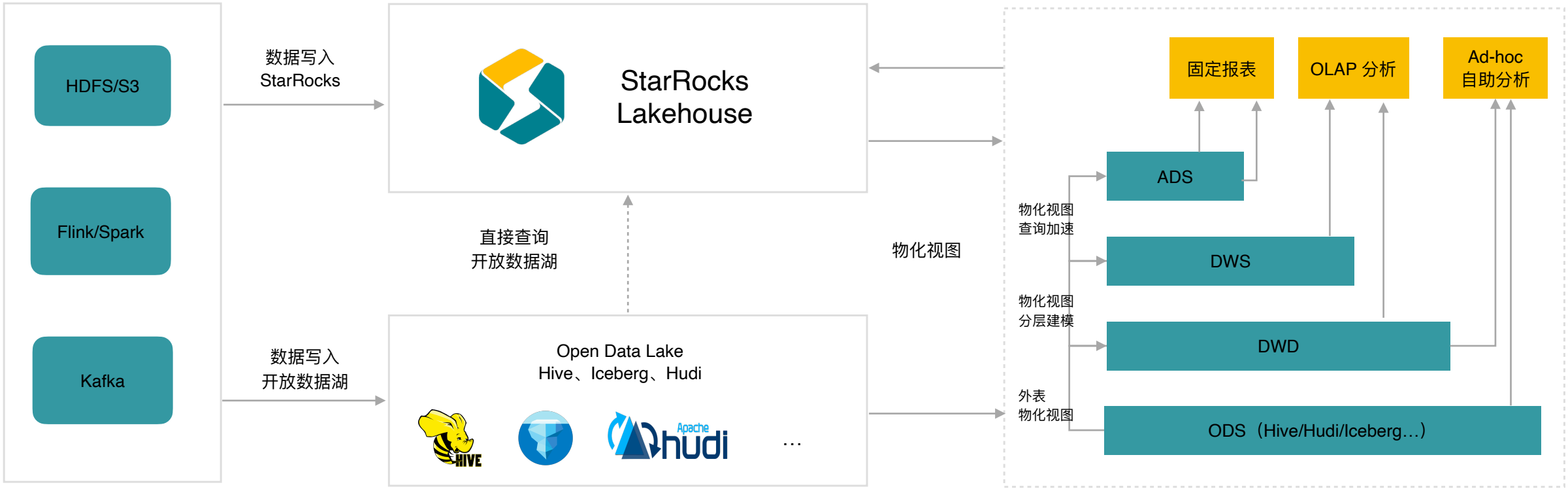
Github star	Pull request	Contributor	社区参与者	大型企业
5200+	20000+	250+	10000+	300+

- 向量化执行引擎
- CBO 优化器
- Runtime filter
- Lateral join、FastDecimal
- 大宽表读写优化
- Bitmap 性能优化
- 主键模型
- Pipeline 执行引擎
- Query cache
- 低基数全局字典
- 直方图统计信息
- 多表物化视图
- 半/非结构化数据查询优化
- 算子落盘（Disk spill）
- Trino 兼容性
- ……

StarRocks 1.x | **StarRocks 2.x** | **StarRocks 3.x**

- Hive/ES/MySQL 外表
- StarRocks 外表
- Open source
- Hudi/Iceberge/Delta lake Catalog
- JSON/Map/Struct 类型
- 资源隔离 Resource group
- 存算分离架构
- Multi-wareshouse
- ……

湖仓融合新范式



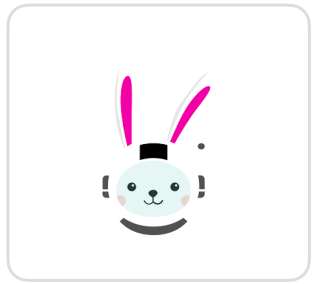
1 **StarRocks 作为统一 Lakehouse**
数据写入 StarRocks 提供极速分析

2 数据写入开放数据湖
使用 **StarRocks 直接分析数据湖**

3 通过**物化视图简化湖仓建模**
无需 ETL，按需加速数据湖查询

湖仓融合价值

1x base



Trino/Presto
查询数据湖

Data lake
Hive/Iceberg/Hudi

3x

- CBO
- Vector engine
- IO Coalesce
- Late Materialization

StarRocks
查询数据湖

Data lake
Hive/Iceberg/Hudi

6x

- CBO
- Vector engine
- IO Coalesce
- Late Materialization
- **Local block cache**

StarRocks
开启 local cache

Data lake
Hive/Iceberg/Hudi

10x

- ……
- **Partition/Bucketing**
- **Colocate join**
- **Index**
- **Statistics**
- **Realtime update**

数据导入 StarRocks 内表

30x

- **Materialized View**

物化视图预计算

开源社 kaiyuanshe

02

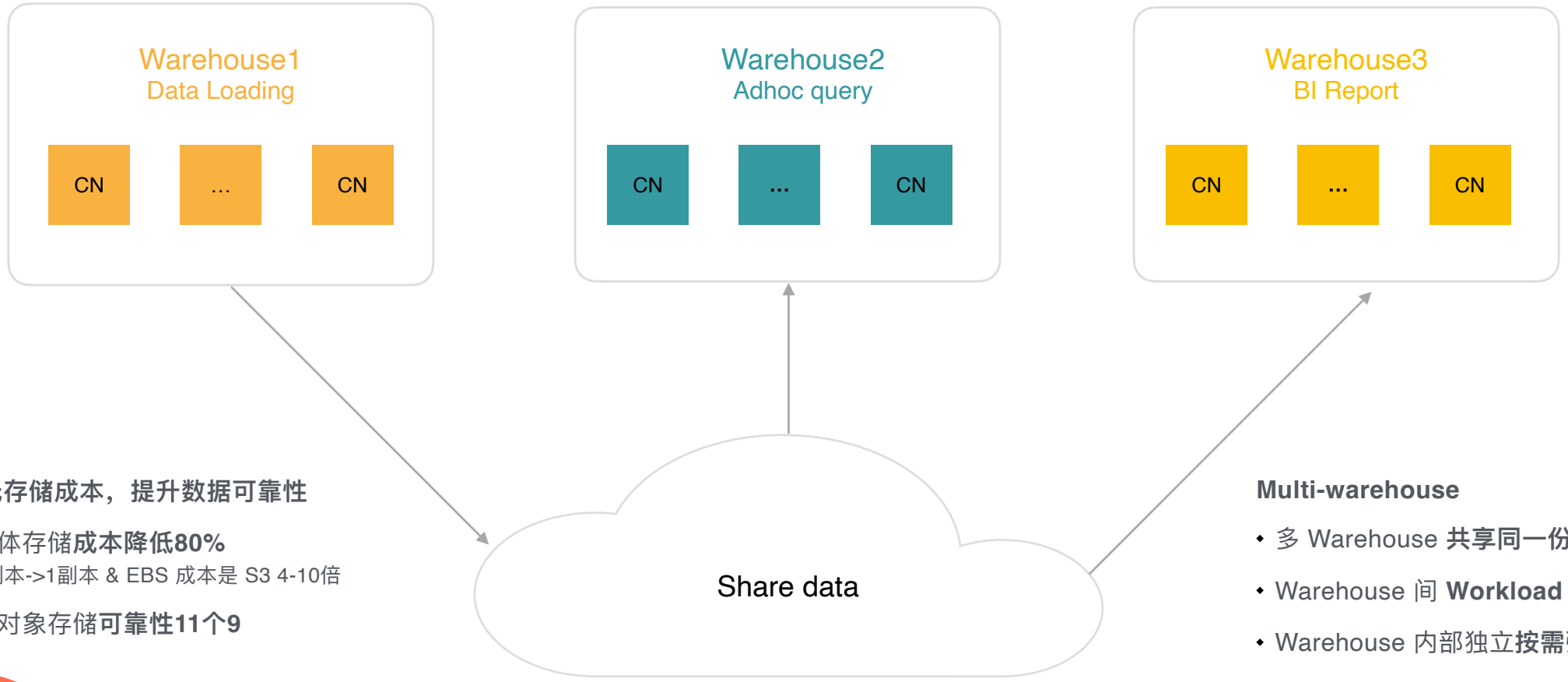
LakeHouse 产品技术演进

StarRocks

版本发布

StarRocks 3.0-3.1	StarRocks 3.2-3.3	StarRocks 3.x
• 2023-04 & 2023-08	• 2023-11 & 2024-02	• 2024
• 发布存算分离架构 • 存算分离支持主键模型表	• 存算分离支持 block 级别数据缓存 • 存算分离支持 Multi-Warehouse	
• 支持数据写回Iceberg • Trino 语法兼容、函数兼容	• 支持数据写回 Hive • Parquet Reader 性能优化 • 支持更多数据类型的JNI Connector	• 支持增量更新的 Materialized view • 支持物化视图自动推荐 • 支持行存，面向更高并发场景 • 支持 Lake Compaction，全面托管 Lake Data • ……
• 支持 Disk Spill • 物化视图查询改写、易用性优化 • 物化视图支持 Hive 分区刷新	• 物化视图支持资源组 • 物化视图支持 Iceberg 分区刷新 • Table Function 导入支持更多文件格式 • 支持 Table Random Distribution	

存算分离架构



降低存储成本，提升数据可靠性

- 整体存储**成本降低80%**
  3副本->1副本 & EBS 成本是 S3 4-10倍
- 云对象存储**可靠性11个9**

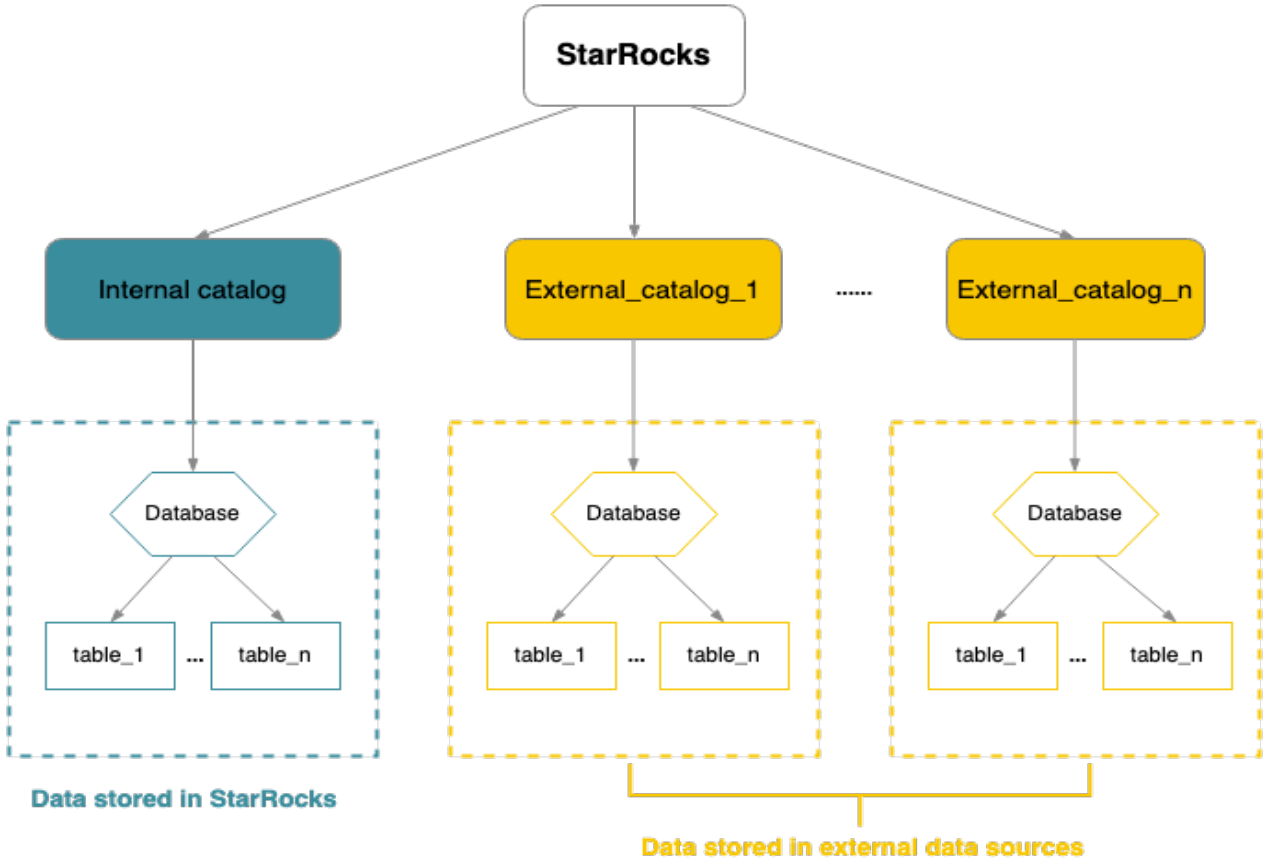
Multi-warehouse

- 多 Warehouse **共享同一份数据**
- Warehouse 间 **Workload 物理隔离**
- Warehouse 内部独立**按需弹性扩展**

LakeHouse Catalog

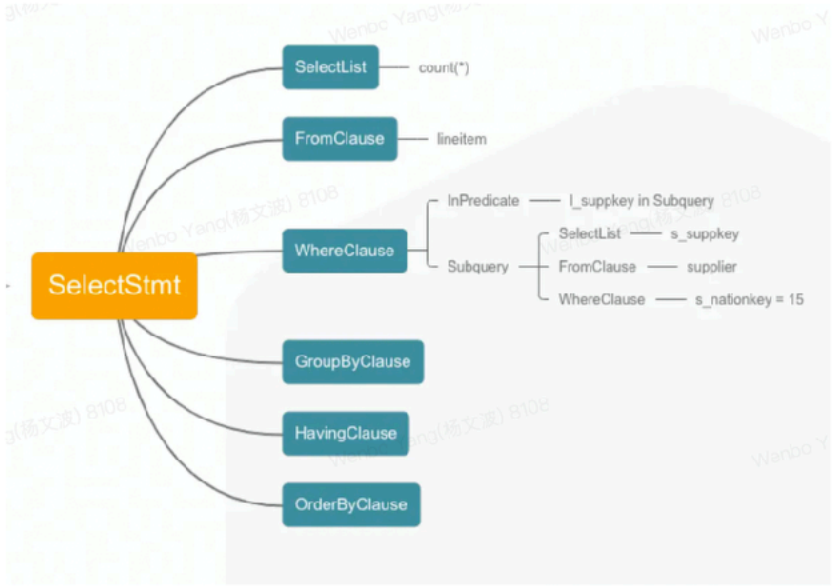
- 支持 Hive/Iceberg/Hudi，直接分析湖上数据
- 跨数据源（Catalog）联邦分析
- 内外表数据访问统一管理

```
CREATE EXTERNAL CATALOG
    <catalog_name>
PROPERTIES
(
    "type" = “iceberg”,
    MetastoreParams,
    StorageCredentialParams,
    MetadataUpdateParams
)
```

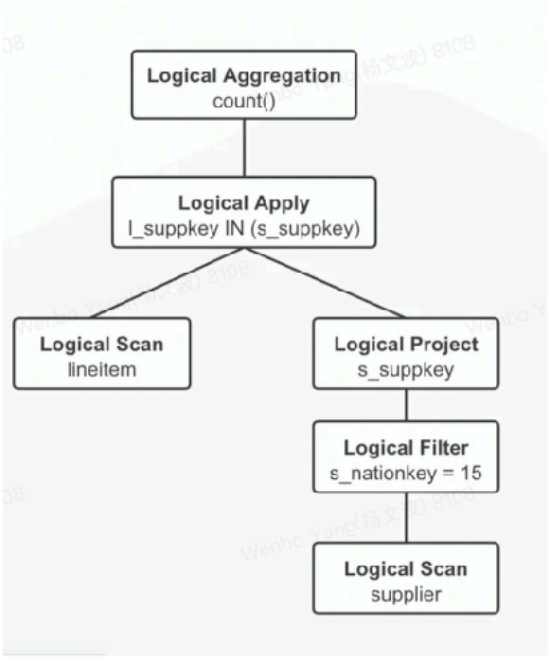


LakeHouse - Trino 兼容

Parse



Plan



SQL Dialect

StarRocks AST

StarRocks Logic Plan

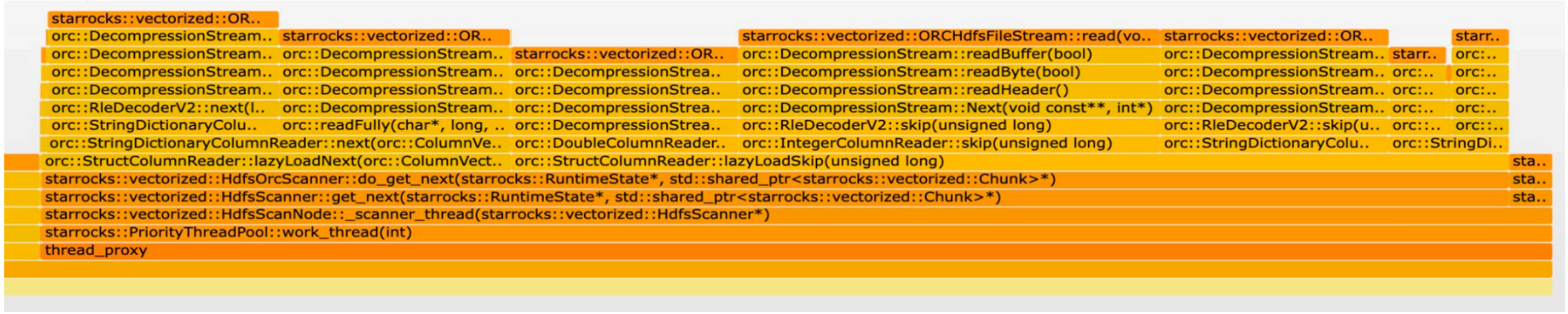
- 支持 Trino SQL 的关键字、语法、函数转义等
- Trino SQL 转换为 StarRocks AST，并生成执行计划
- set sql_dialect = “trino” 启用

LakeHouse - 极速查询性能

极速数据湖分析挑战

- 不同文件格式、不同存储系统 IO 特征不同
- 数据组织，文件小、Row group 设置不合理
- IO 延迟高，无法利用 Page cache 加速

	FileSize	IOCounter	IOBytes	HDFS IOTimer	S3 IOTimer
ORC/int	36GB~50GB	~7000	1.7GB	9101ms	138029ms
ORC/s-string	36GB~50GB	~11000	4.3GB	51775ms	217000ms
Parquet/int	90GB~100GB	~800	2.3GB	19316ms	74835ms
Parquet/s-string	90GB~100GB	~1600	10.6GB	647000ms(带宽受限)	380000ms



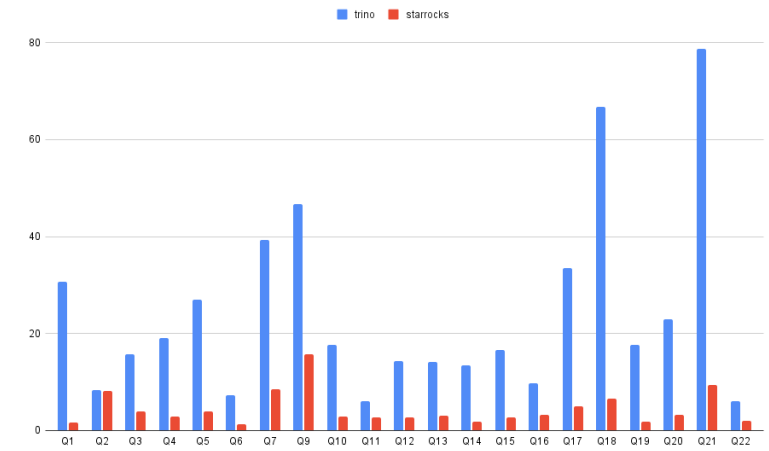
LakeHouse - 极速查询性能

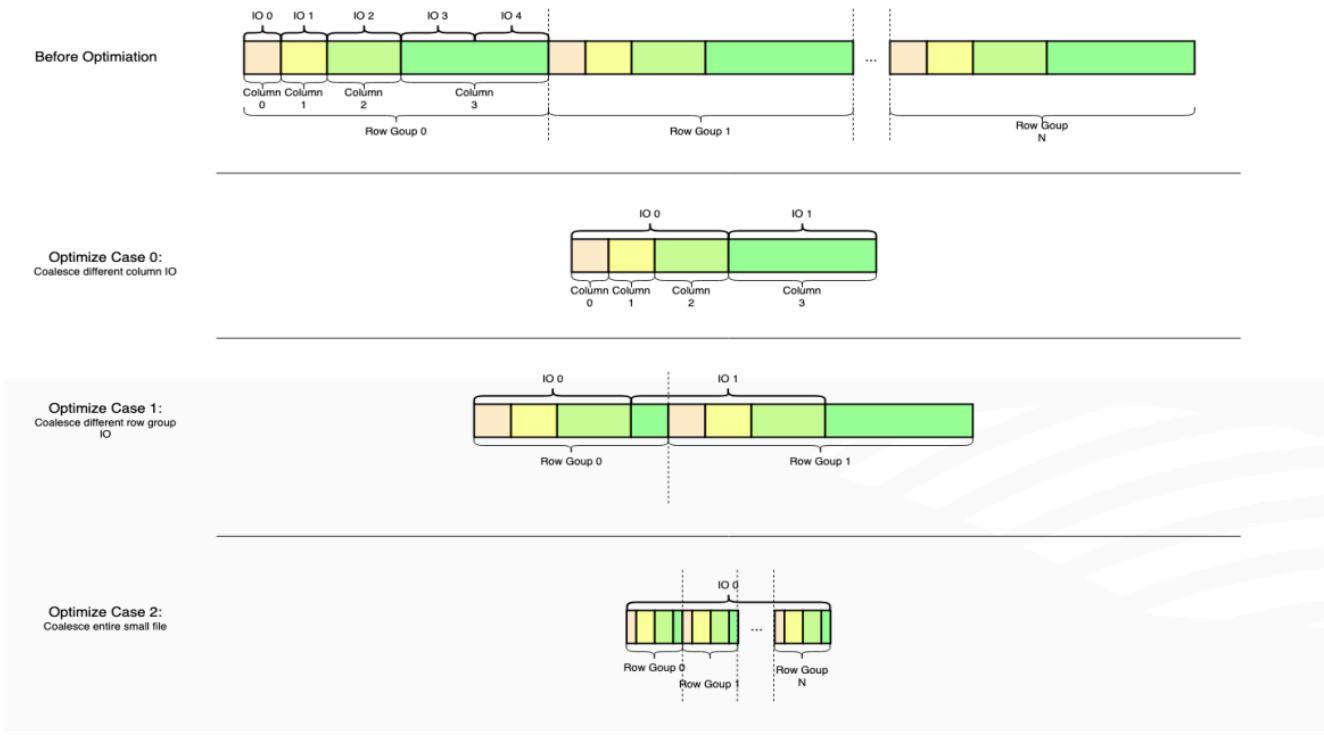
极速数据湖分析挑战

- 不同文件格式、不同存储系统 IO 特征不同
- 数据组织，文件小、Row group 设置不合理
- IO 延迟高，无法利用 Page cache 加速

关键技术

- Local Block Cache：Disk + Memory
- IO 合并，减少 IO 次数；Column 读取合并 -> Row group 合并-> 整个小文件合并读取
- 延迟物化，根据带查询条件的部分列过滤结果，读取其他需要访问的列，减少 IO 总量





- Iceberg 、TPC-H 测试集、4x`16c128g` 计算节点
- StarRocks 直接查询数据湖 比 Trino 快3~5倍

开源社 kaiyuanshe

03

StarRocks 湖仓融合

StarRocks

物化视图助力湖仓融合

数据建模

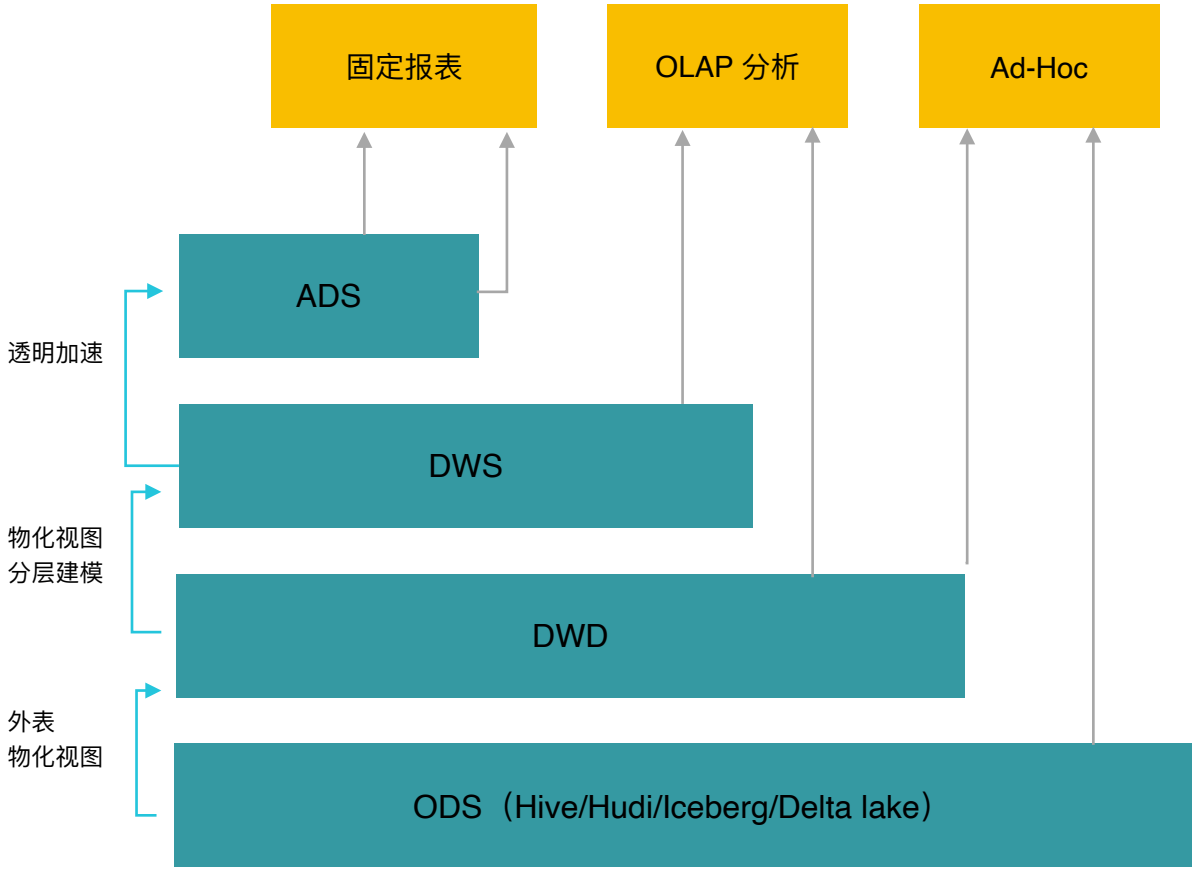
- 简化数据加工，无需维护外部组件
- 分层数据建模，分区增量刷新

湖仓融合

- 冷热数据融合
- 混合负载融合

透明加速

- 优化器自动改写查询，透明加速
- 后置建模：先查询、再建模优化



MV - 基础概念

功能特性

- **MATERIALIZED**: 预计算
- **PARTITION BY**: 按时间分区，缩小刷新粒度
- **REFRESH**: 定时刷新/手动刷新/自动刷新
- RESOURCE GROUP：弹性调度，隔离工作负载
- QUERY: Aggregation/Join/Union
- REWRITE: 优化器自动查询改写

```
CREATE MATERIALIZED VIEW mv1
PARTITION BY dt
REFRESH EVERY (INTERVAL 1 HOUR)
PROPERTIES( “resource_group” = “rg1”)
AS
  SELECT dt, c_city, count(*)
  FROM lineorder t1
  JOIN customer t2
  ON t1.lo_custkey = t2.c_custkey
  GROUP BY dt, c_city
```

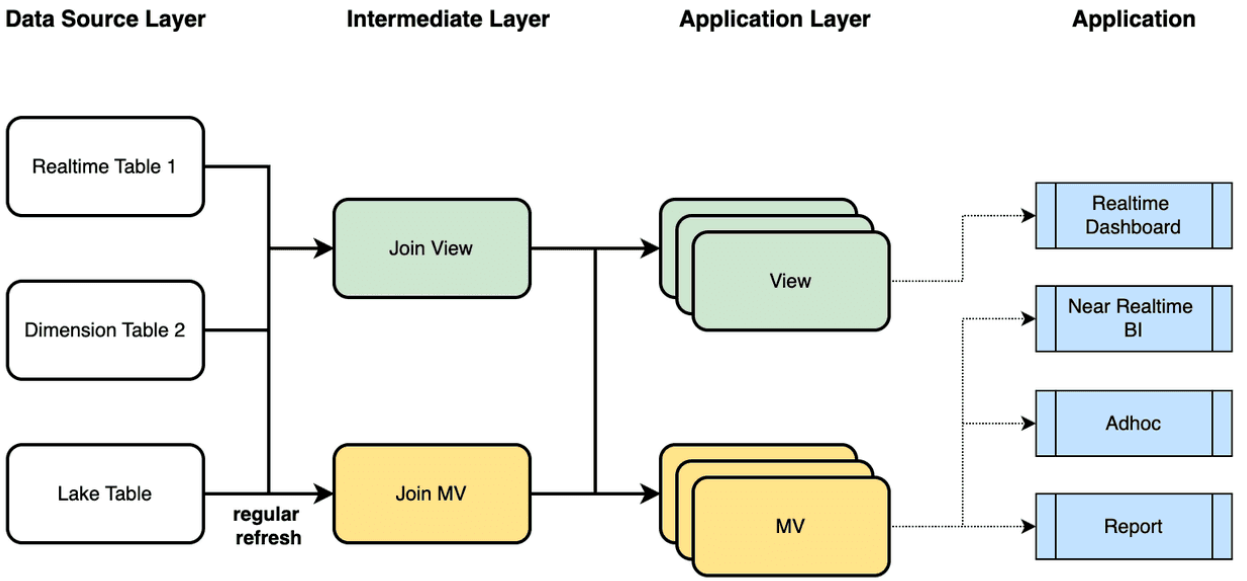
MV - 数据建模

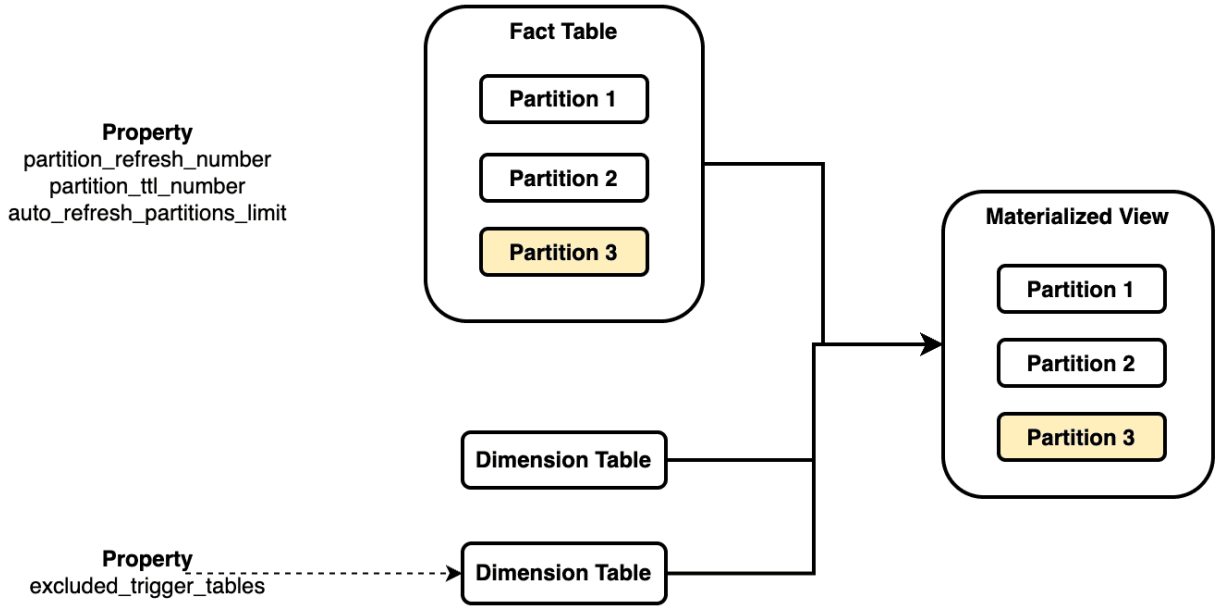
分层建模

- 使用 View & Materialized View 建模
- 自动刷新：数据变更自动触发刷新
- Schema Change：自动变更

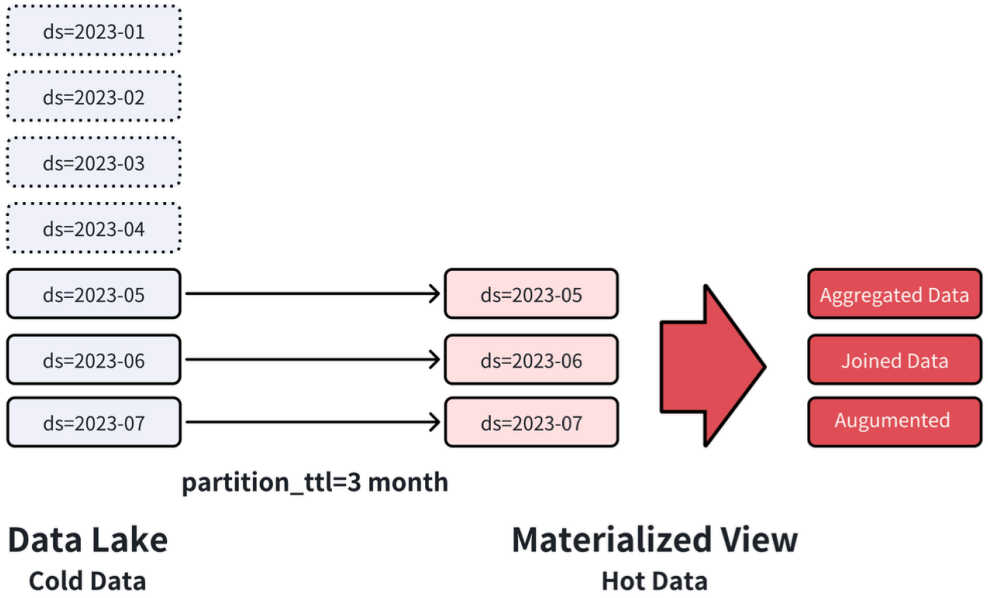
分区建模

- 分区关联：外表增加分区，MV 自动刷新
- 增量刷新：仅刷新变更的分区
- 事实表 & 维度表刷新
- 数据源支持：Hive/Iceberg/Hudi





MV - 冷热数据融合

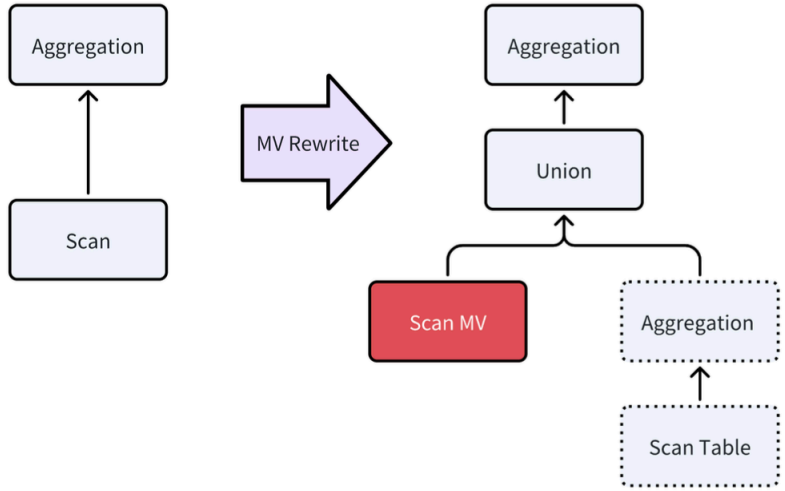


业务场景

- 数据具有冷热特性，想要加速近期查询
- 物化视图指定 TTL：partition_ttl=3 month

冷热分区合并

- 数据物化加速：MV 物化近 3 个月数据，历史数据存储在 Lake 中
- 自动查询改写：无需修改查询，优化器自动改写



MV - 弹性负载融合

使用场景

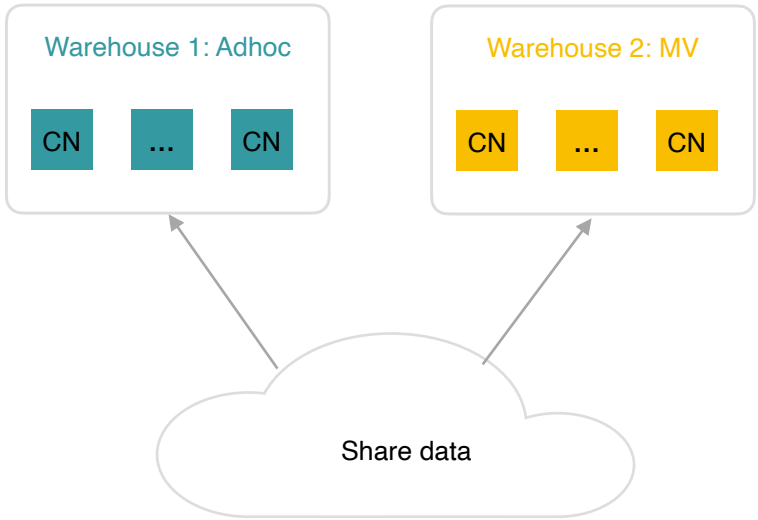
- 多种业务负载在同一集群运行，充分利用资源
- 负载如何实现资源隔离?

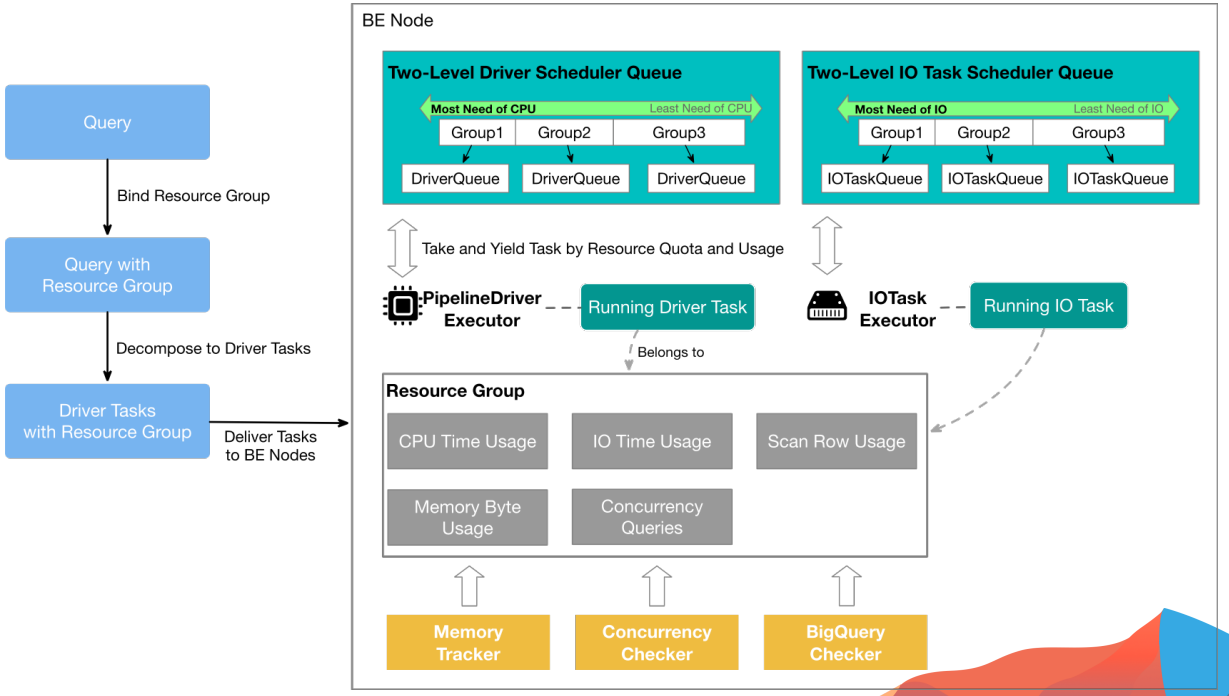
软性隔离：Resource Group

- default query：默认查询资源组，高优先级调度
- default mv: 默认MV资源组，低优先级调度
- CPU 按时间片调度，Memory Tracker + Disk Spill，IO 线程池隔离

硬性隔离：Multi Warehouse

- 每种负载指定 Warehouse：导入、查询、MV
- 存算分离：数据共享，计算隔离





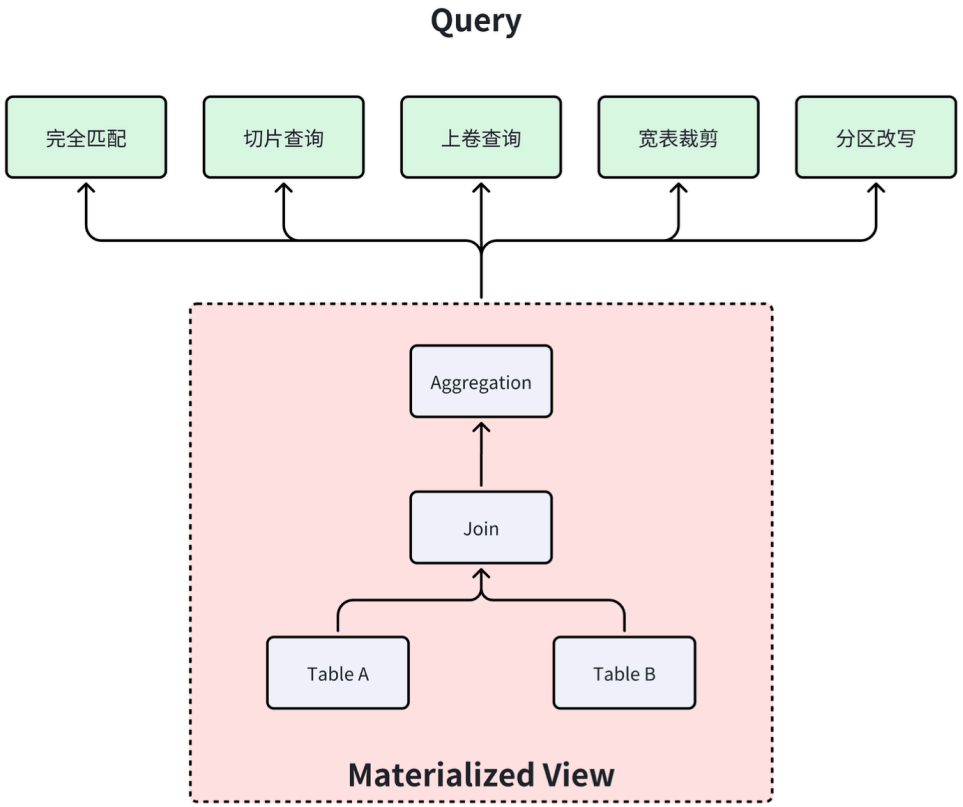
MV - 透明加速

应用场景

- 痛点：BI 报表不易修改 SQL，不易调优
- 方案：声明物化视图，优化器自动改写

关键技术

- 聚合改写
- 分区改写
- 宽表改写



MV - 透明加速

改写示例1：聚合上卷改写

示例1：

```
SELECT lo_orderdate, count(*)
FROM lineorder_flat
GROUP BY lo_orderdate
```

示例2：

```
SELECT c_city, count(*)
FROM lineorder_flat
WHERE
    lo_orderdate >= ‘2023-09-00’ AND
    lo_orderdate < ‘2023-09-31’
GROUP BY c_city
```

示例3:

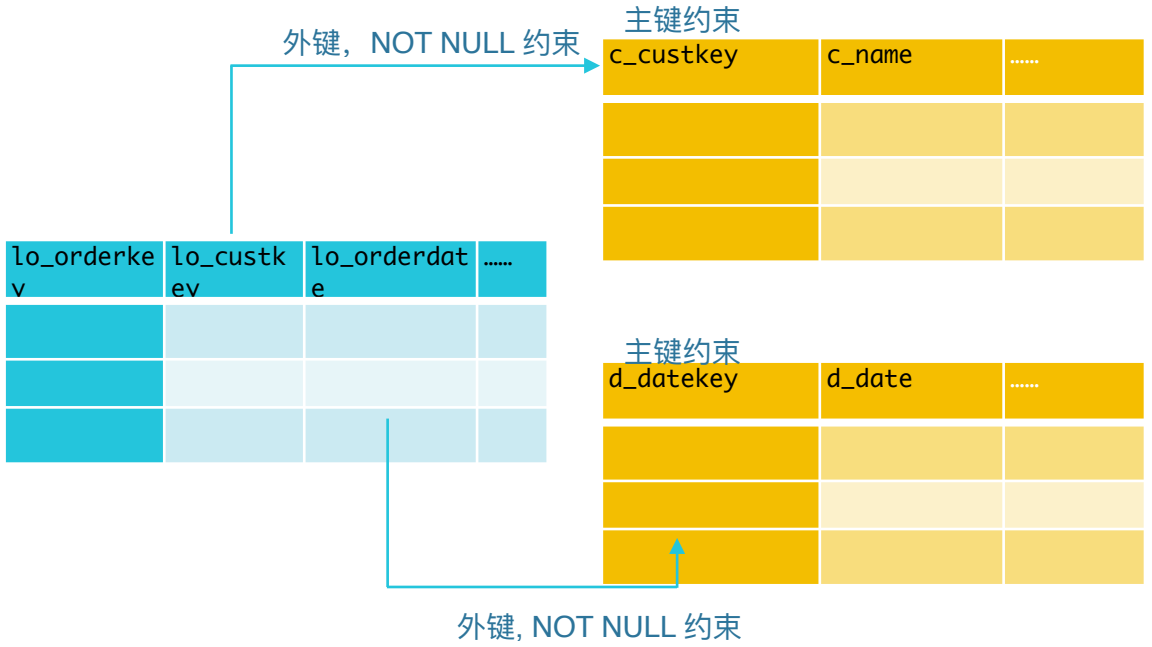
```
SELECT count(*)
FROM lineorder_flat
WHERE lo_orderdate = ‘2023-09-28’
```

```
CREATE MATERIALIZED VIEW mv1 AS
  SELECT lo_orderdate, c_city, count(*)
  FROM lineorder_flat
  GROUP BY lo_orderdate, c_city
```

lo_orderdate	c_city	……

MV - 透明加速

改写示例2：宽表 Join 改写

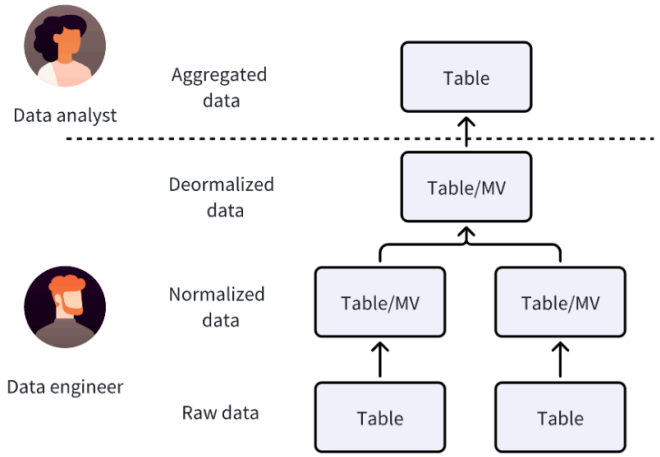


Query：

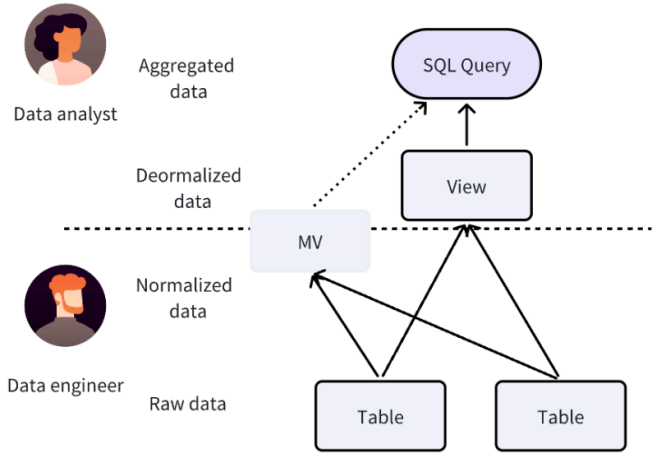
```
SELECT xxxx
FROM lineorder t1
LEFT JOIN customer t2
ON t1.lo_custkey = t2.c_custkey
GROUP BY c_city
```

```
CREATE MATERIALIZED VIEW mv1 AS
  SELECT lo_orderdate, c_city, count(*)
  FROM lineorder t1
  LEFT JOIN customer t2
  ON t1.lo_custkey = t2.c_custkey
  LEFT JOIN supplier t3
  ON t1.lo_suppkey = t3.s_suppkey
```

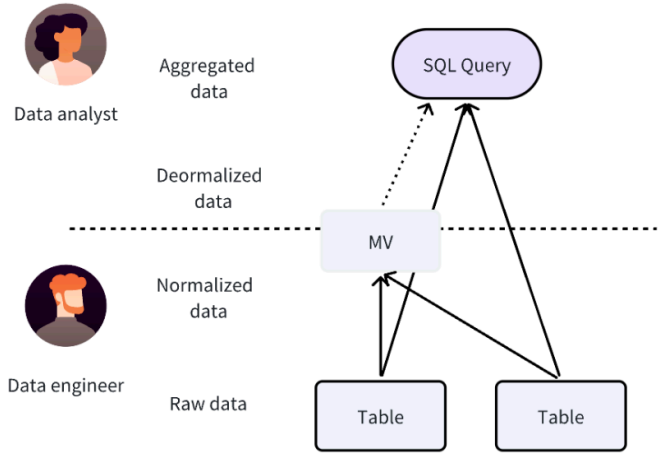
MV - 透明加速



layer modeling:
Build model in advance



View modeling:
View as logical model, MV for accleration



Lazy modeling:
Transparent acceleration using MV

Proactive Governance ⟷ Post-Hoc Governance

从前置建模，到后置建模
从技术驱动，到业务驱动

THANK YOU

QUESTIONS?

微信公众号：开源社KAIYUANSHE

视频号：开源社KAIYUANSHE

新浪微博：开源社

B站：开源社KAIYUANSHE

简书：开源社

头条：开源社

Facebook：KaiyuansheChina

Twitter：开源社KAIYUANSHE

开源社 kaiyuanshe

欢迎扫码打卡
积分可兑换对应礼品哟！

扫码关注开源社公众号

公众号：StarRocks