Alibaba Cloud | Worldwide Cloud Services Partner

PolarDB In-Memory Column Index
HTAP解决方案

蔡畅
阿里云数据库内核研发

自我介绍

Alibaba Cloud | Worldwide Cloud Services Partner

- 达梦数据库公司
- 淘宝数据库团队

  AliSQL分支, RocksDB引擎，X-Engine引擎
- 阿里云数据库团队

  PolarDB，X-Engine引擎，列存引擎

目录

Alibaba Cloud | Worldwide Cloud Services Partner

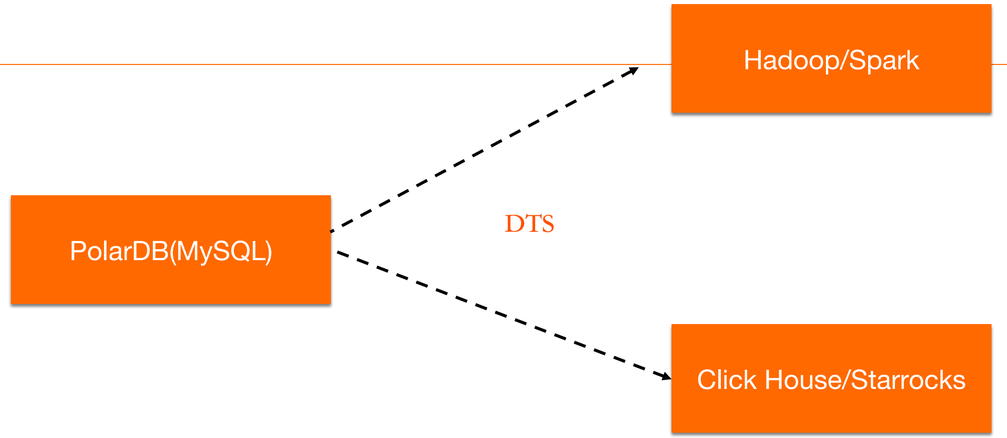
目录

Alibaba Cloud | Worldwide Cloud Services Partner

PolarDB MySQL 的客户诉求：HTAP

Alibaba Cloud | Worldwide Cloud Services Partner

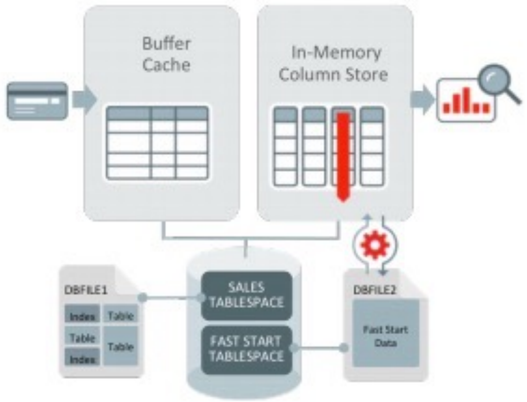
- PolarDB MySQL 被大量在线业务所采用。
- PolarDB MySQL主要面向TP场景优化， 分析型查询由于执行器及存储引擎限制，性能不佳。
- PolarDB 数据导出至外部专用OLAP系统支持大数据的复杂查询，成本高，架构复杂，运行维护麻烦。
- 更易用的一体化数据库，一套DB同时满足OLTP及OLAP需求。



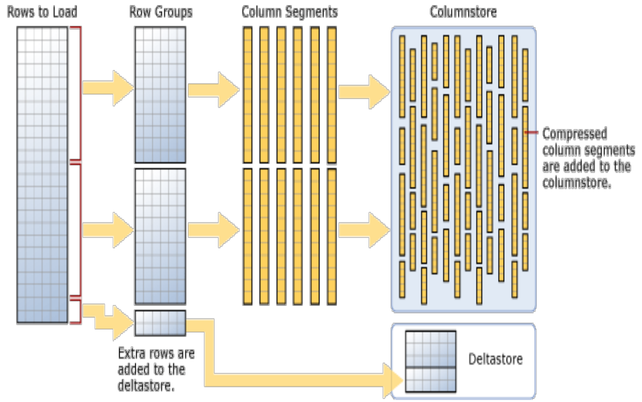
传统流行数据库产品

Alibaba Cloud | Worldwide Cloud Services Partner

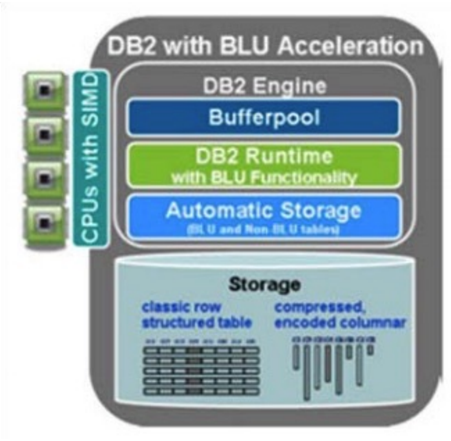
- OLTP与OLAP需求在一套系统里面被满足
- 云原生数据库追赶
  - 容量&弹性
  - AP+TP技术融合



Oracle In-Memory Database



SQL Server Column store Index



IBM DB2 BLU

PolarDB MySQL HTAP技术挑战

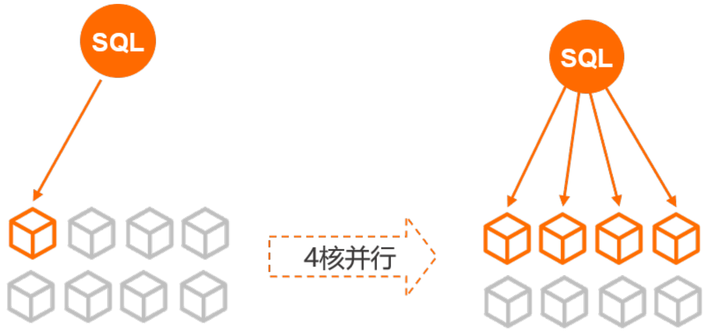
Alibaba Cloud | Worldwide Cloud Services Partner

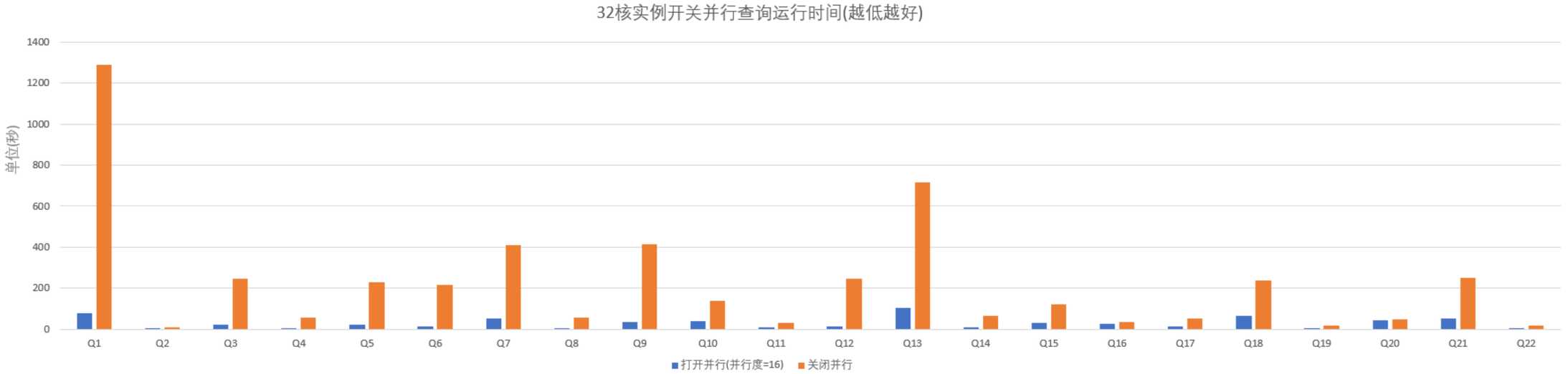
- 性能
  - 处理复杂SQL能力
- 数据实时性
  - TP更新负载下，AP请求实时性保证
- AP&TP隔离性
  - TP吞吐和响应时间不受影响
  - 混合负载自动分流
- 弹性扩缩容
- 100%的MySQL兼容性

PolarDB MySQL: 并行执行加速

Alibaba Cloud | Worldwide Cloud Services Partner

- MySQL-分析场景缺陷
  - 行存
  - 单核串行执行
- PolarDB的改进
  - 并行查询，利用多核CPU并行处理
- 需要配套的列存PK专用OLAP系统

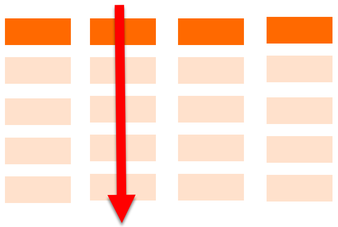




PolarDB MySQL： Why PolarDB Need Column Store

Alibaba Cloud | Worldwide Cloud Services Partner

记录按列拆分存储



- 只访问必要数据

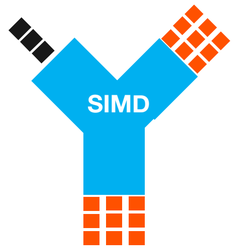
压缩存储

- 压缩数据及减少IO量
- 压缩数据计算

列存稀疏索引

- 大块数据过滤

SIMD向量执行



- 单指令多数据

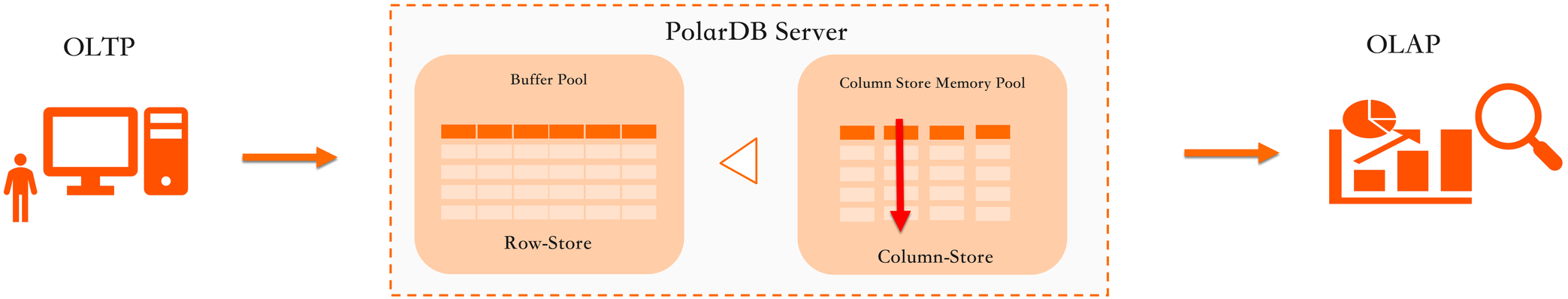
IO加速

CPU加速

PolarDB IMCI：一体化HTAP数据库

Alibaba Cloud | Worldwide Cloud Services Partner

- PolarDB IMCI 在一张表上同时支持行存和列存存储
- OLTP负载访问基于Page存储的行格式，OLAP分析型负载访问列存储格式
- PolarDB内部实时维持行列数据一致性，对用户透明

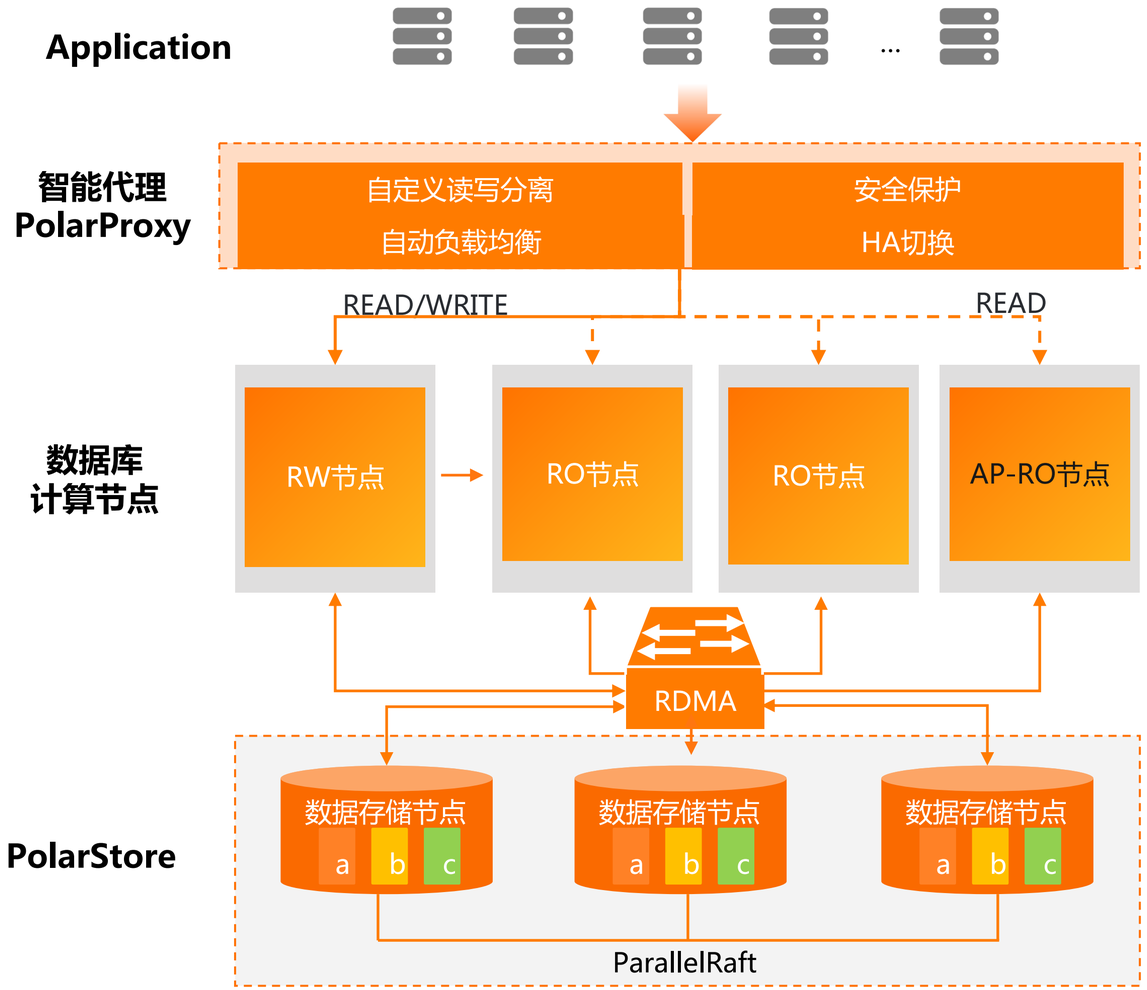


目录

Alibaba Cloud | Worldwide Cloud Services Partner

PolarDB IMCI： 云原生一体化HTAP架构

Alibaba Cloud | Worldwide Cloud Services Partner



高可靠高可用

- 基于分布式共享存储，数据3副本，多可用区部署
- 海量存储，支持100TB
- 秒级快照，可按时间点快速恢复数据
- 主节点故障后快速切换

高弹性高性能

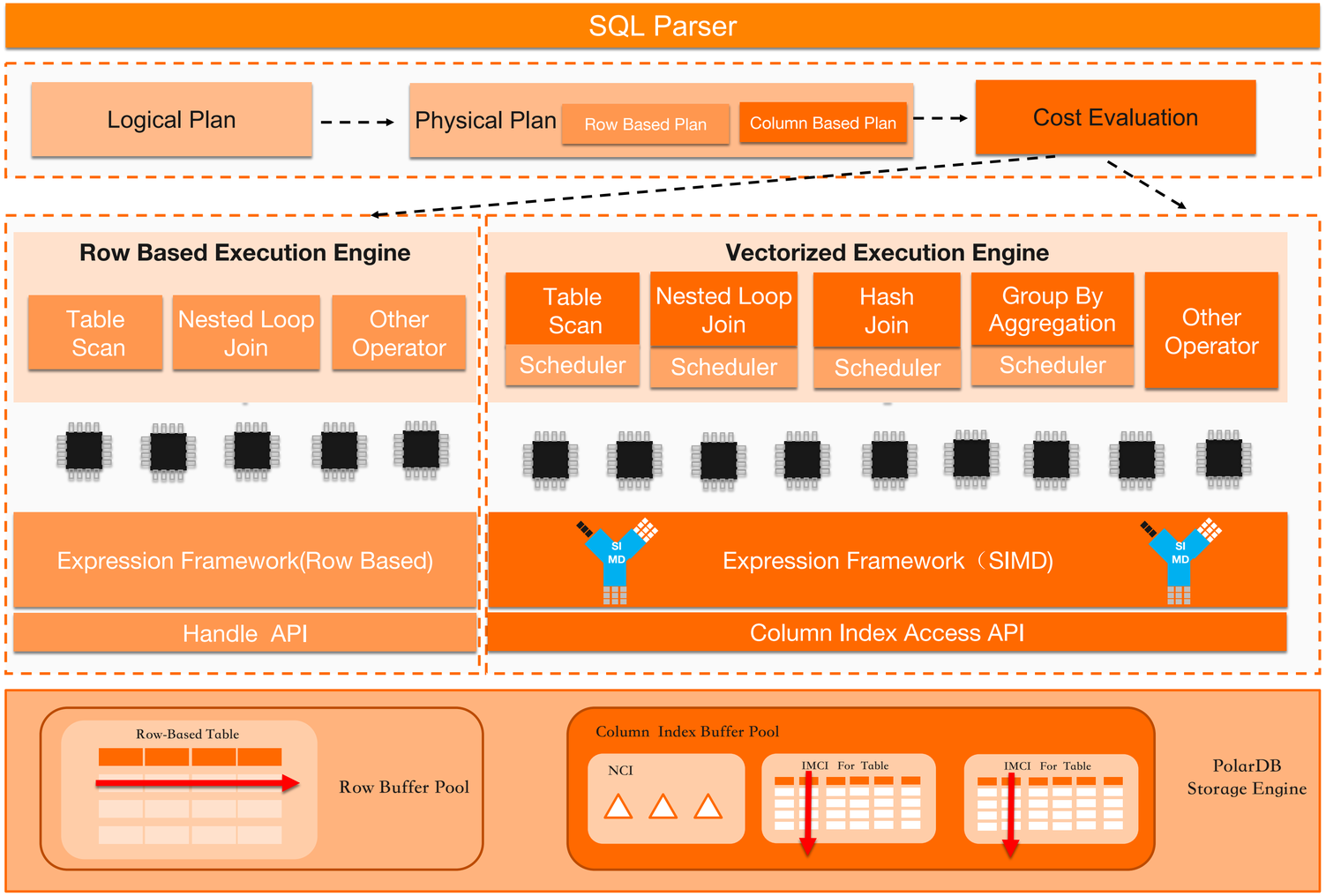
- 计算节点可扩展到**16个计算节点，1000核**
- 计算节点弹性扩展(ScaleUp, ScaleOut)
- 并行查询，充分利用多节点多核CPU
- **AP-RO节点**，并行和向量化技术，复杂查询提供**百倍的加速比**

易用性好

- 多个计算节点自动读写分离
- 多个计算节点**AP&TP负载自动分流**
- AP-RO节点单独部署，**AP&TP完全隔离**

PolarDB In-Memory Column Index 整体架构

Alibaba Cloud | Worldwide Cloud Services Partner



Parser/Optimizer

- 100% MySQL兼容
- 串行/PQ/列存混合调度
- Cost-based Optimizer

Execution Engine

- 向量化处理架构
- 并行执行算子
- SIMD表达式加速

Storage Engine

- 行列混合存储
- 事务级别的行列一致性
- 数据压缩 & 粗糙集索引
- 列存物理复制及一写多读

目录

Alibaba Cloud | Worldwide Cloud Services Partner

IMCI Query Engine: Hybrid Execution

Alibaba Cloud | Worldwide Cloud Services Partner

100% MySQL兼容性如何实现

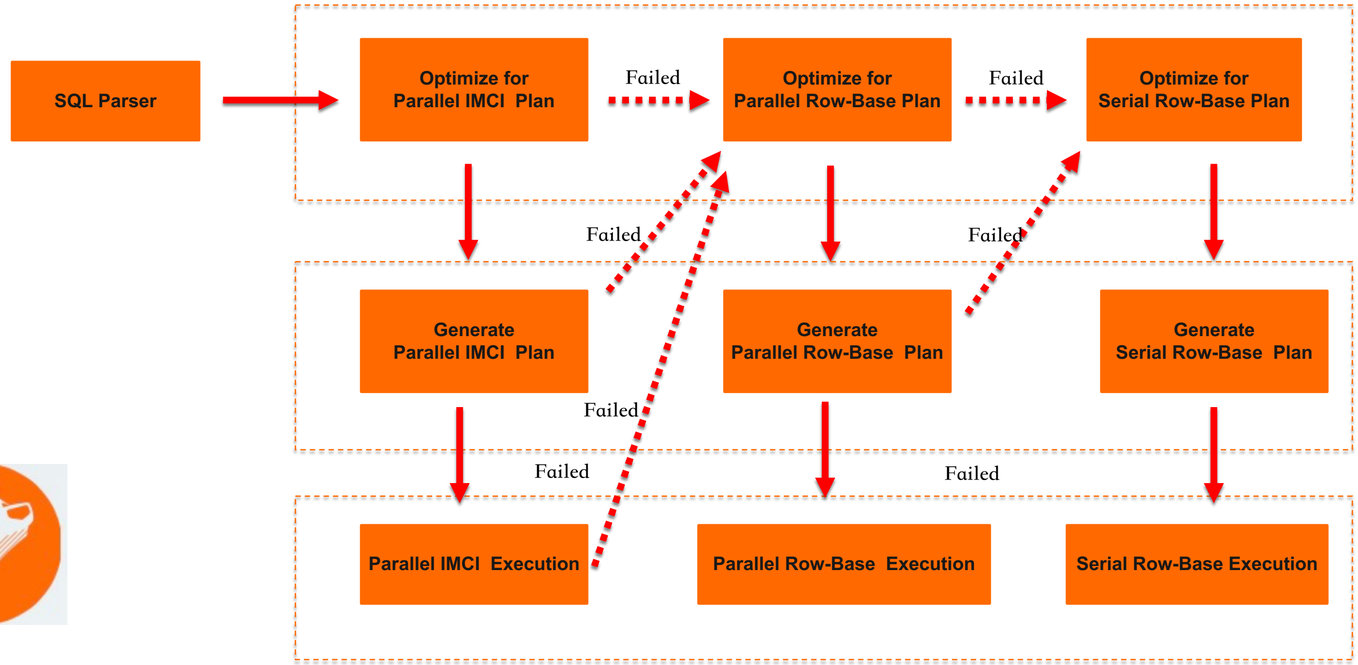
- 行存串行 100%兼容
- 行存并行
- 列存并行

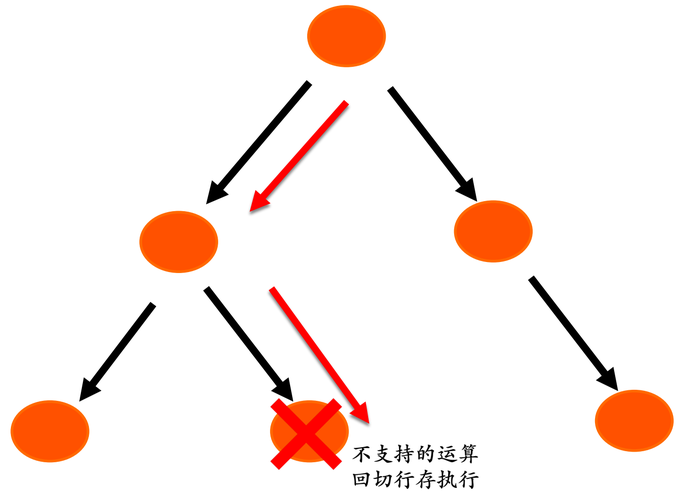
基于白名单的执行加速

- 遍历执行计划树以探测IMCI兼容性
- 不兼容的Plan回退行存执行

基于代价的执行计划选择

- 行存Cost VS PQ COST VS 列存Cost
- 整体代价较低时选择行存





IMCI Query Engine: Parallel Execution

Alibaba Cloud | Worldwide Cloud Services Partner

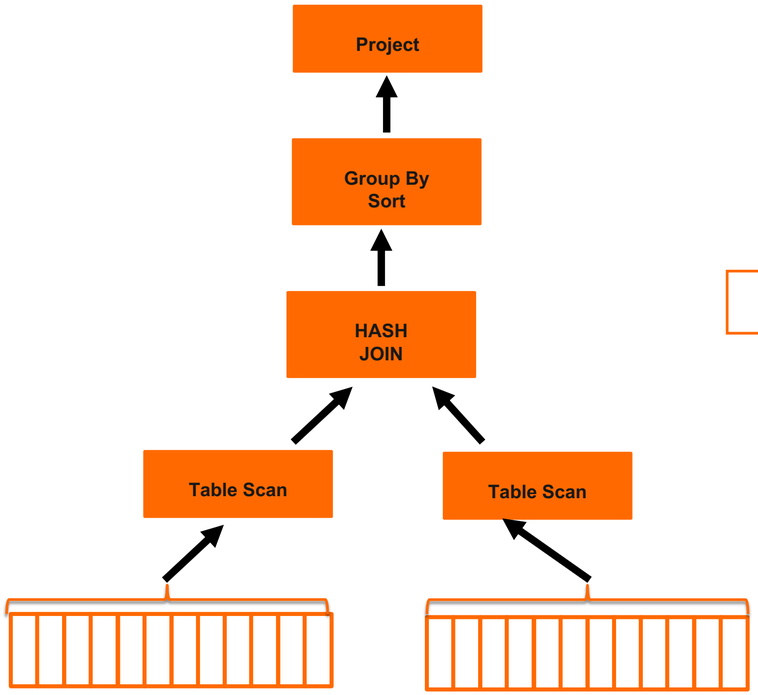
行存执行器

- 火山模型执行模式
- 按行迭代
- 算子串行

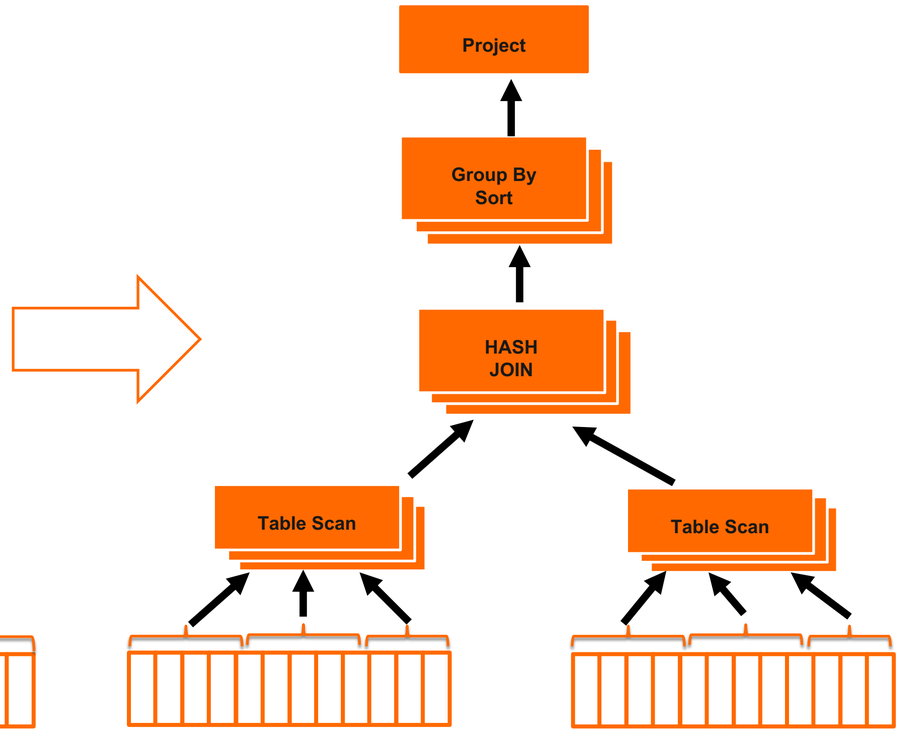
列存执行器

- 火山模型执行模式
- 批处理
- 算子并行

Serial Execution Plan

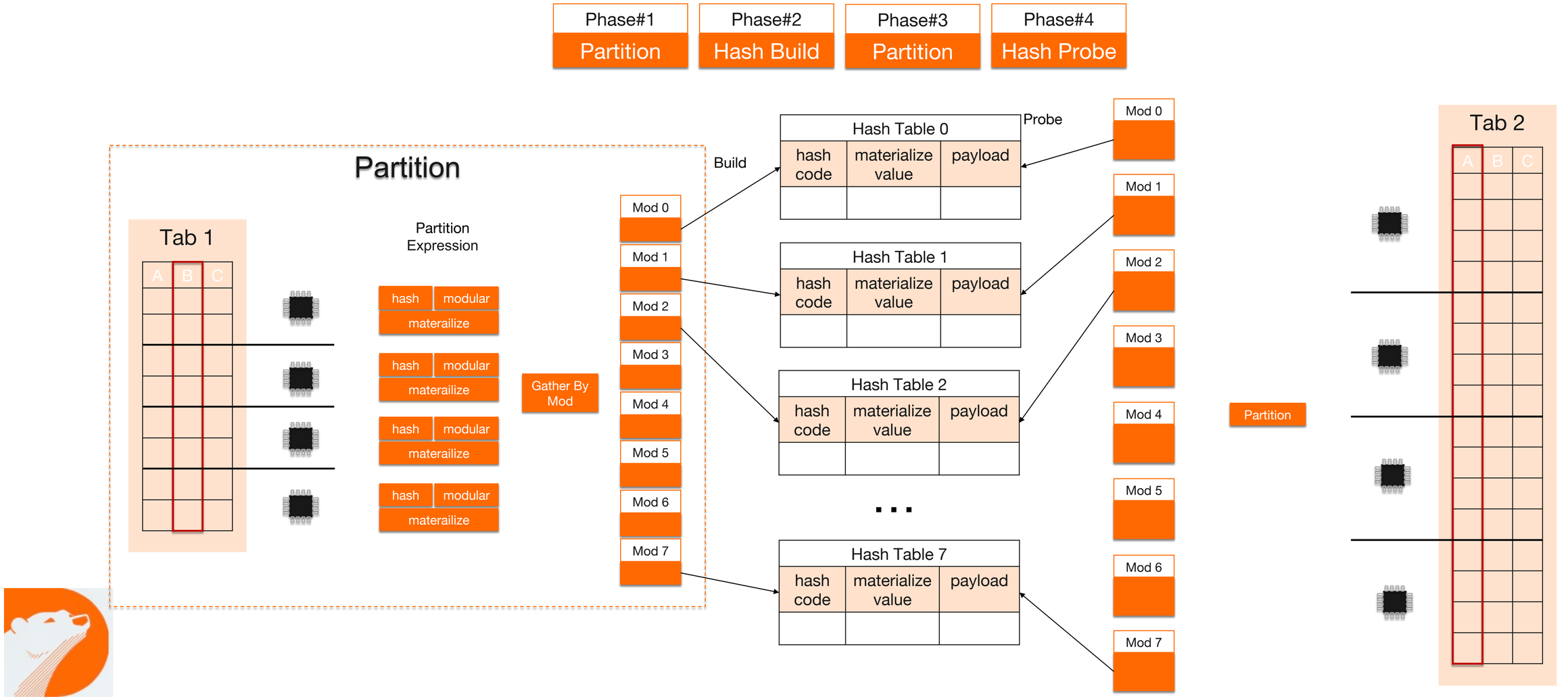


Parallel Execution Plan




```
SELECT   customers.cust_first_name, customers.cust_last_name, MAX(QUANTITY_SOLD), AVG(QUANTITY_SOLD)
FROM     sales, customers
WHERE    sales.cust_id=customers.cust_id
GROUP BY customers.cust_first_name, customers.cust_last_name;
```

IMCI Query Engine: Parallel Hash Join

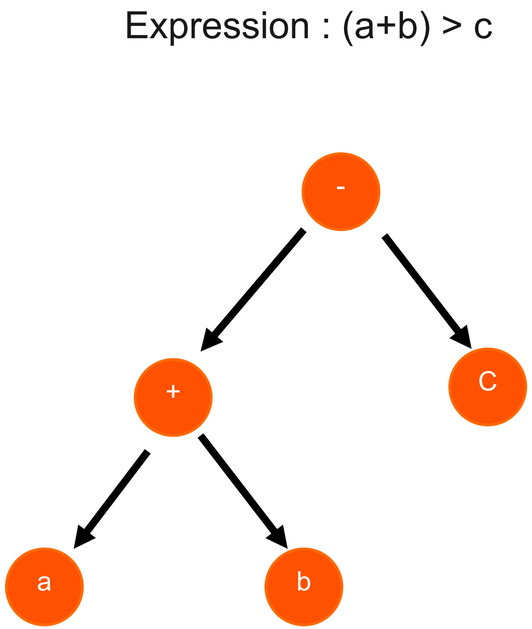


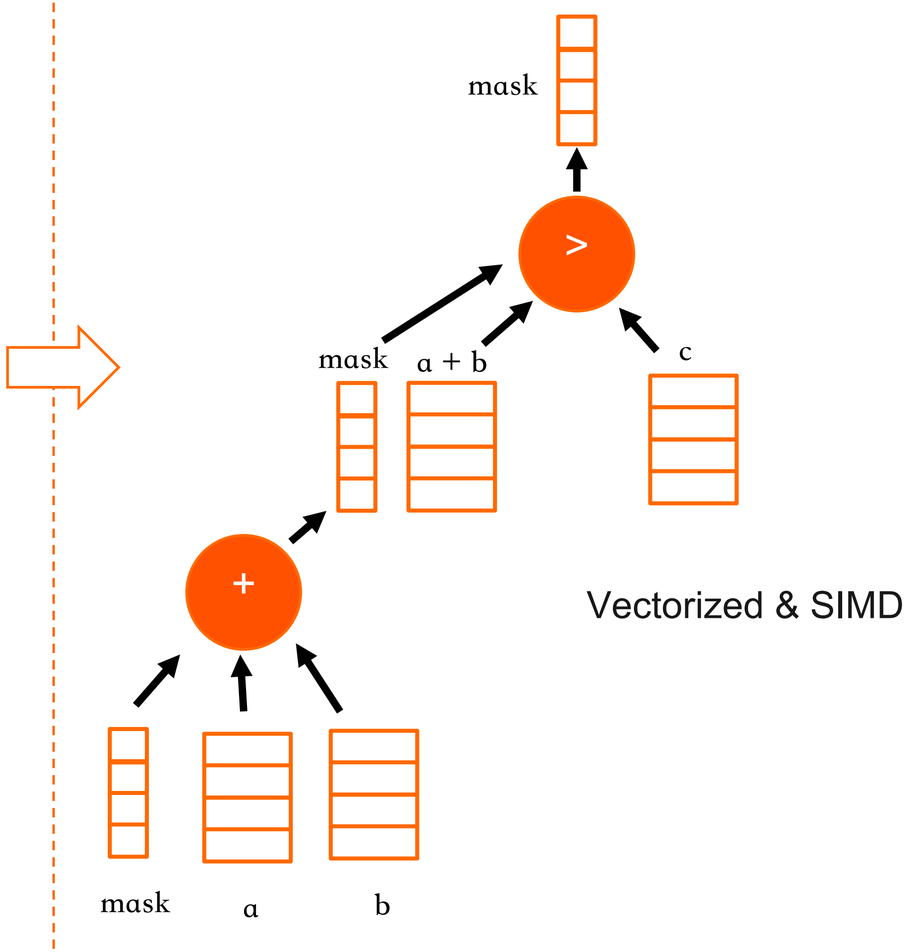
IMCI Query Engine: Vectorized Expression

Alibaba Cloud | Worldwide Cloud Services Partner

向量化表达式

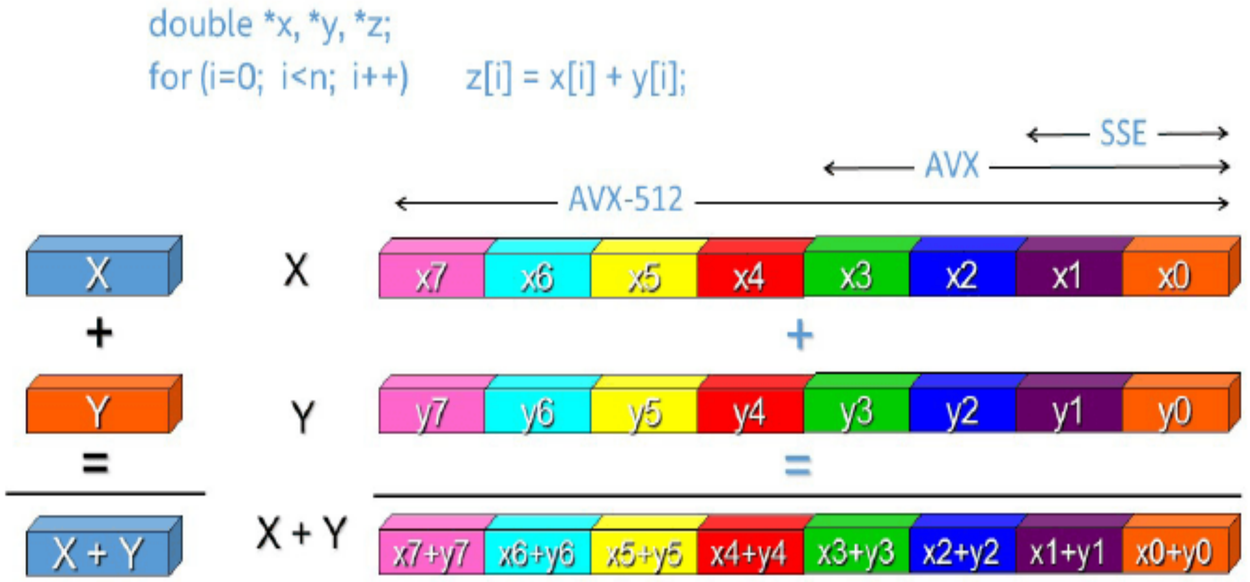
- 批量处理带Mask运算
- AVX 512 SIMD指令加速
- 表达式递归消除提升CPU计算效率

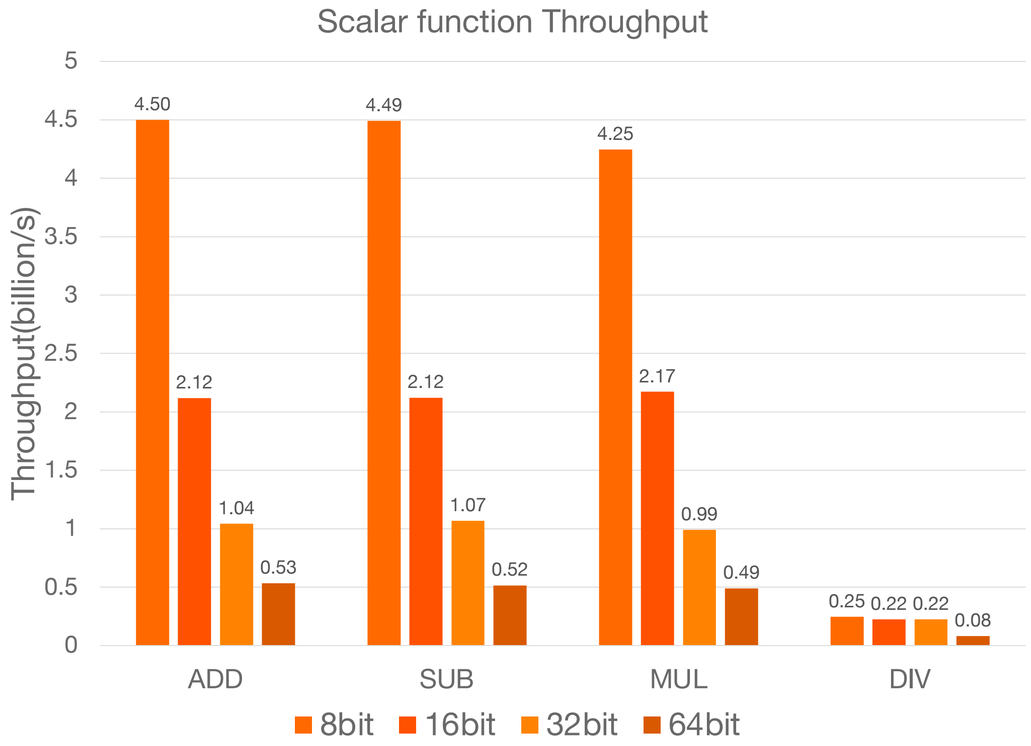




IMCI Query Engine: SIMD based Expression

Alibaba Cloud | Worldwide Cloud Services Partner





使用 SIMD(AVX512) 指令可以获得近10倍的加速效果

目录

Alibaba Cloud | Worldwide Cloud Services Partner

IMCI Storage：行列混合存储

Alibaba Cloud | Worldwide Cloud Services Partner

存储设计，按访问模式聚集数据

- TP场景：点查，查明细，并发实时更新(行存)
- AP场景：大量行扫，少数列聚合，复杂运算
- 行列存储融合

行列混存优势

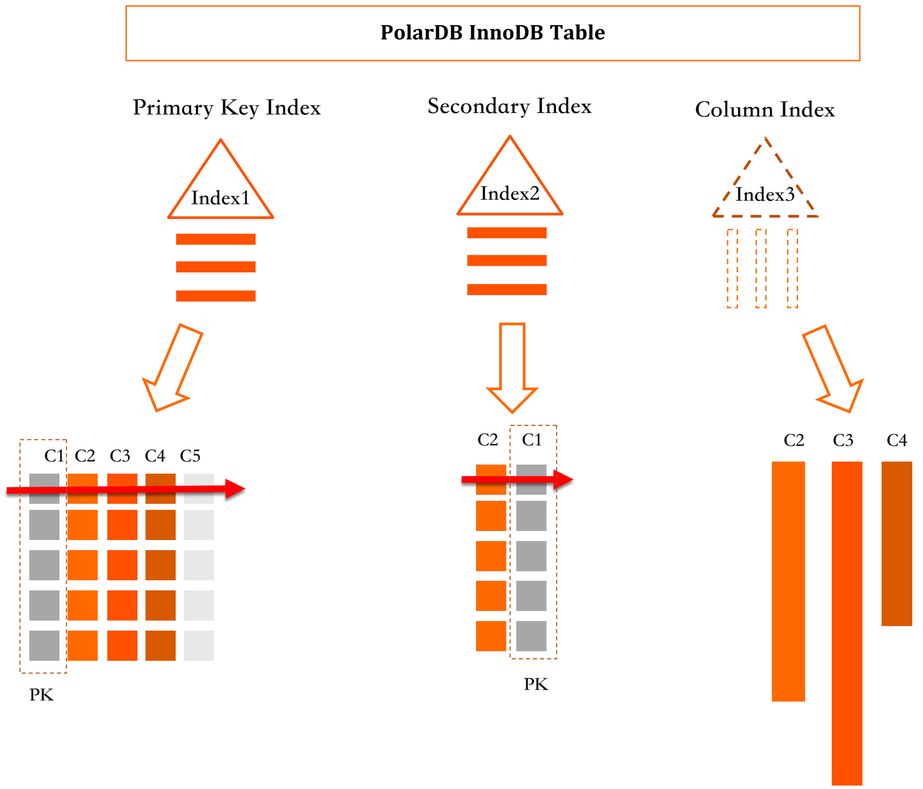
- 数据一致性保证
- 轻量级数据同步，实时性更好
- 行列混合执行
- 对用户透明，管理运维方便

IMCI Storage： 实现为索引的列存

Alibaba Cloud | Worldwide Cloud Services Partner

列存形态

- 列存实现为InnoDB表虚拟列索引(victual column index)
- 按表维度动态新增/删除列索引，与二级索引一样
- Column index 覆盖列上的同步更新Apply到列存索引.
- 按需使用行存索引和列存索引对查询加速

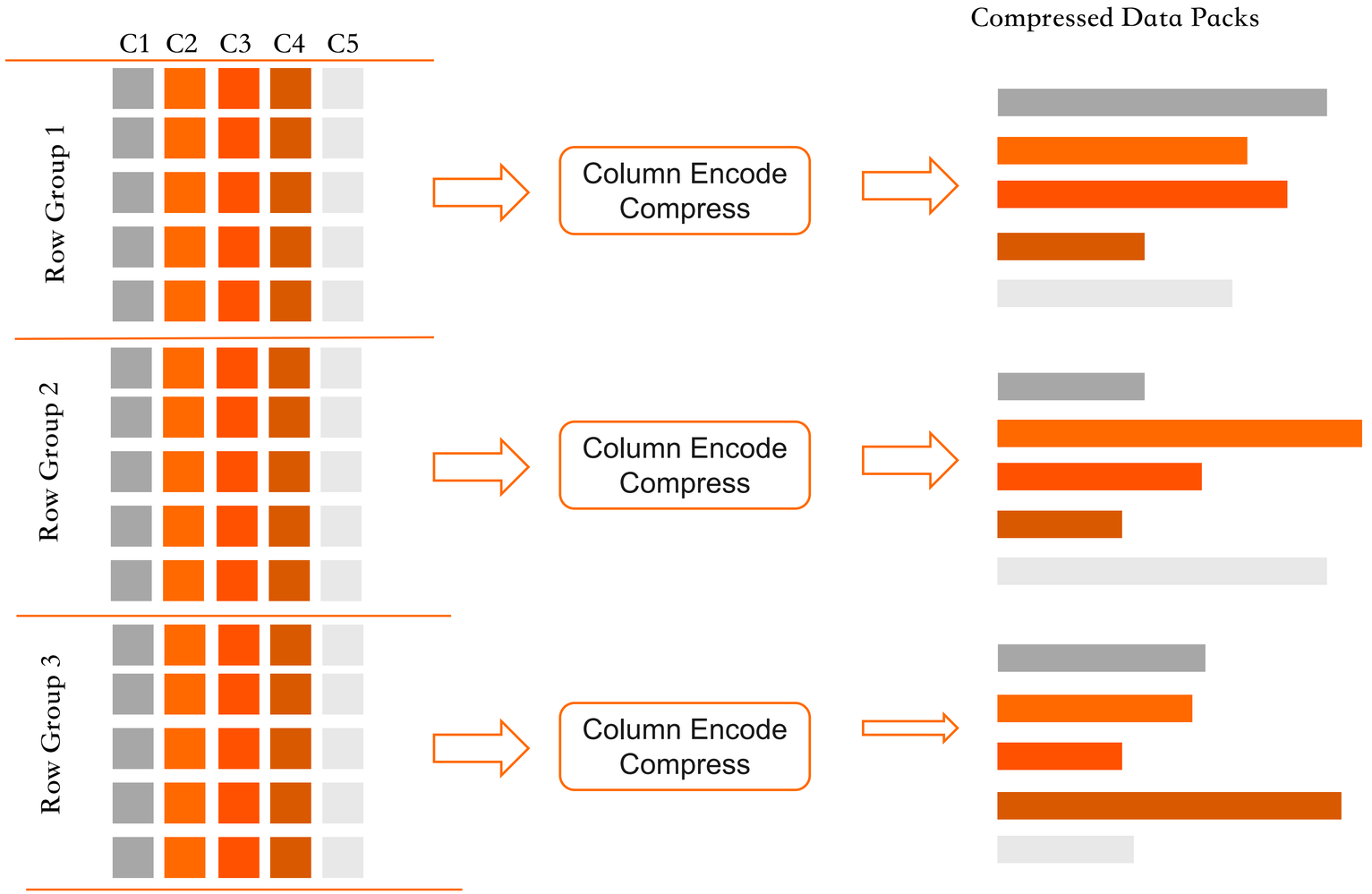


IMCI Storage: 列存数据组织及管理

Alibaba Cloud | Worldwide Cloud Services Partner

IMCI 列存组织

- 堆表，64K行称为一个Row Group
- 逻辑RowID标识唯一行
- Row Group内拆分按列存储
- 列Data Pack压缩(Encoding,Compress)

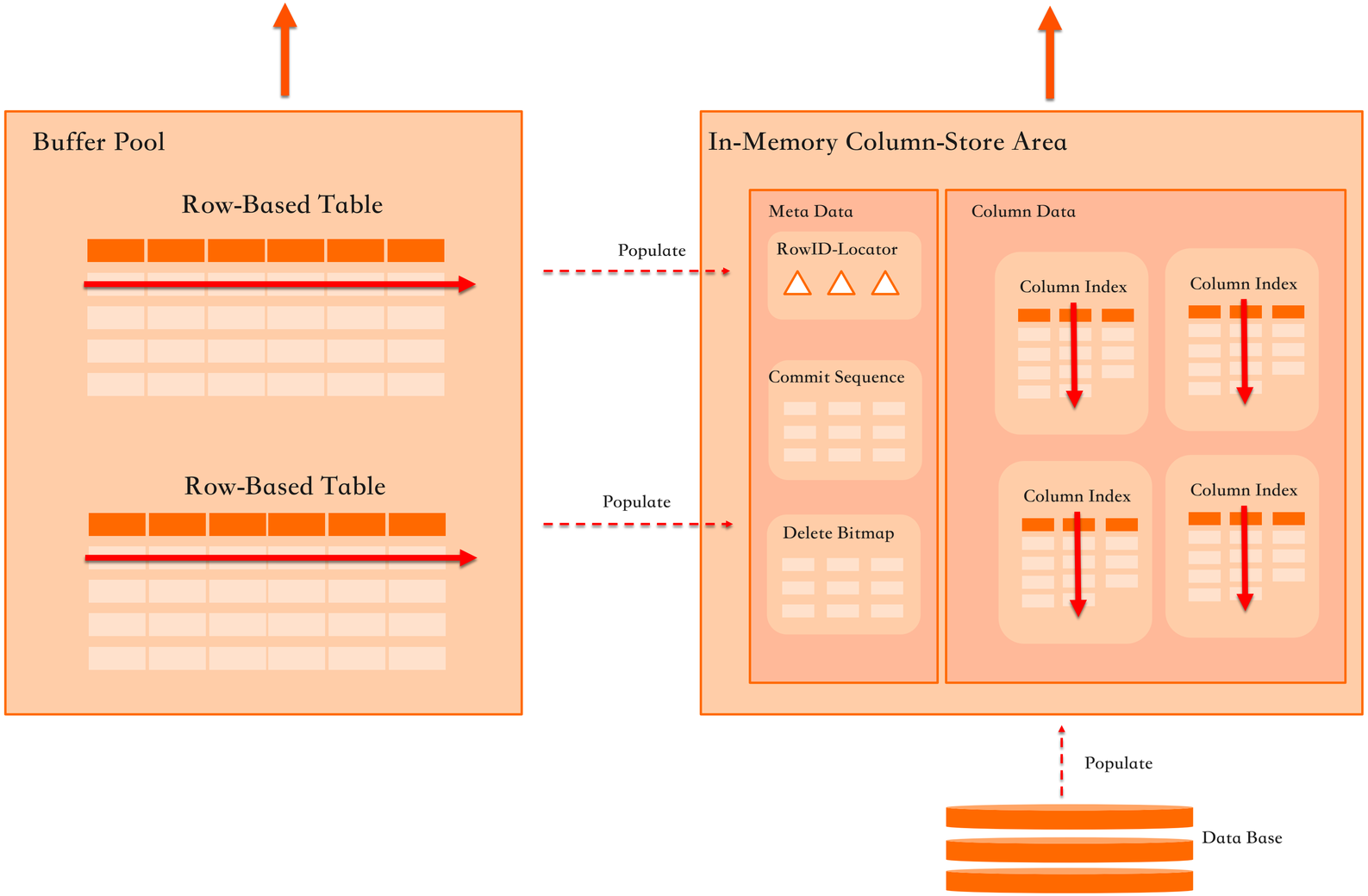


IMCI Storage：列存管理

Alibaba Cloud | Worldwide Cloud Services Partner

列存管理

- insert追加写
- delete bitmap， commit-version
- RowID-locator，快速定位待更新行
- Compaction机制，回收空洞
- Checkpoint机制，fast-recovery

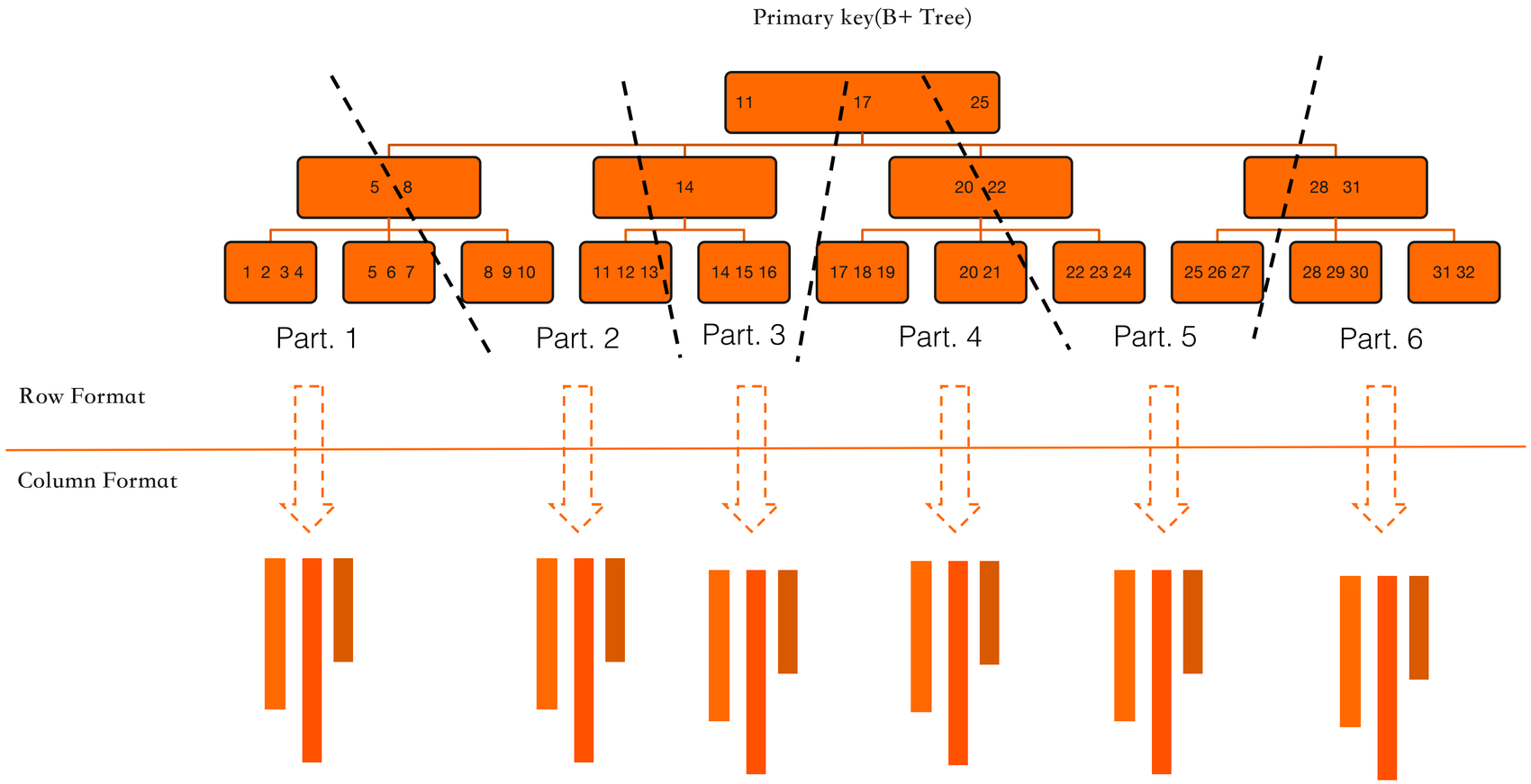


IMCI Storage: 存量行存表转列存

Alibaba Cloud | Worldwide Cloud Services Partner

全量构建列索引

- DDL对存量数据新增列索引
- Parallel Scan主表
- OnlineDDL，不阻塞DML操作

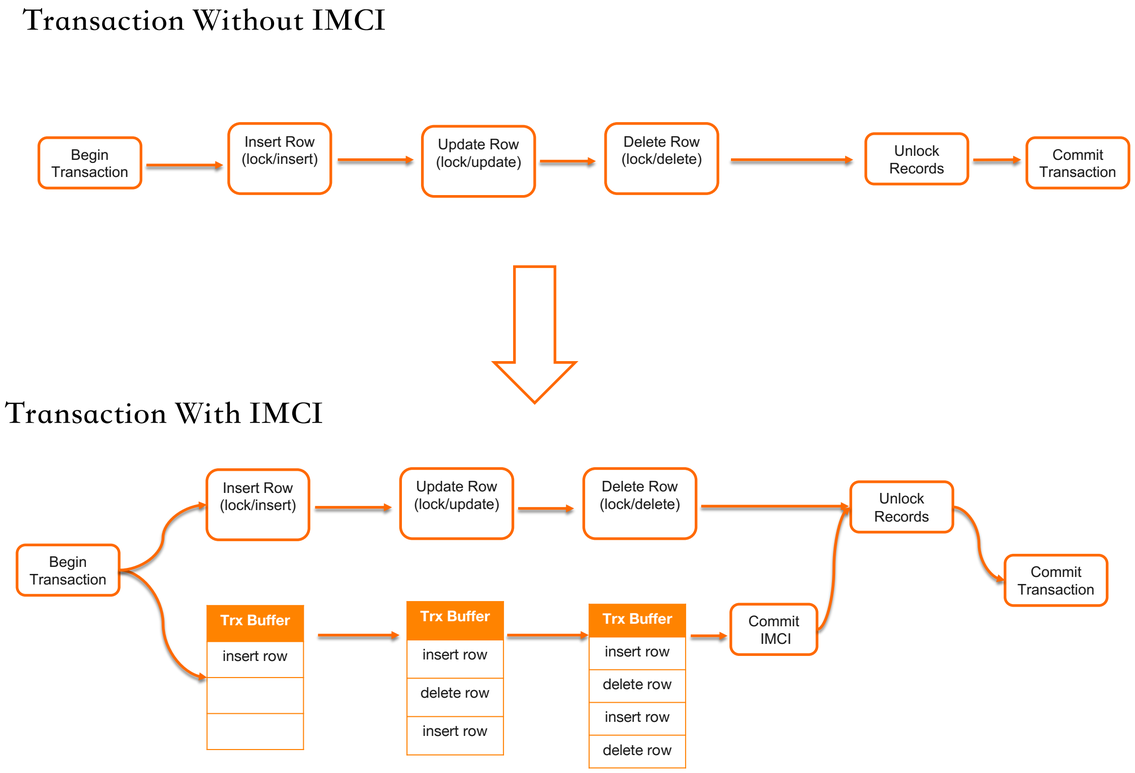


IMCI Storage: 行列数据的同步更新

Alibaba Cloud | Worldwide Cloud Services Partner

列索引增量构建

- 更新操作同时更新行存和列存缓存
- 与行存事务共享相同的并发控制机制
- 轻量级事务提交操作
- 大事务的处理



IMCI Storage：基于Redo的行列混合物理复制

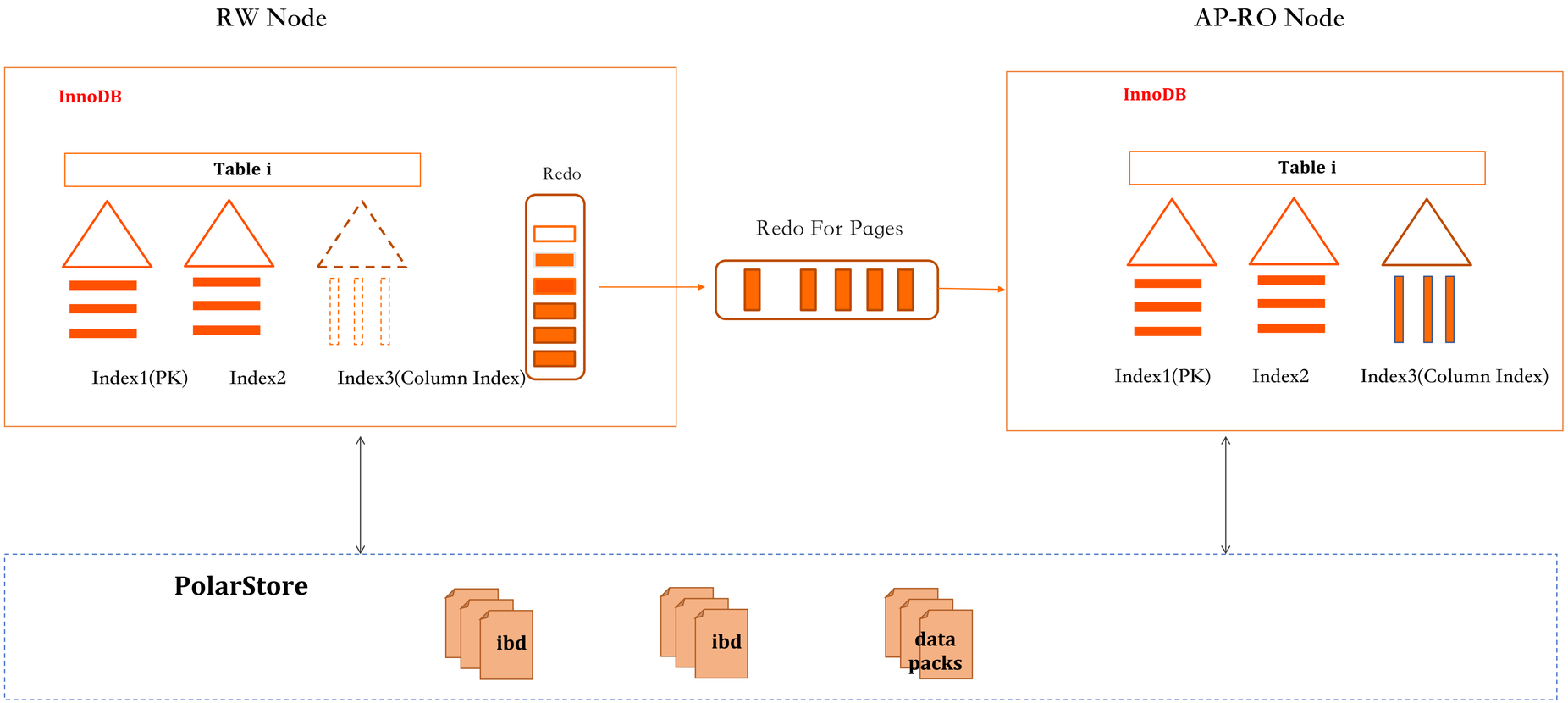
Alibaba Cloud | Worldwide Cloud Services Partner

数据复制技术

- binlog逻辑复制
- binlog-in-redo
- reuse-redo物理复制

实时同步技术

- 异步复制一致性(lsn位点)
- 多级并行回放



IMCI Storage：列存数据压缩

Alibaba Cloud | Worldwide Cloud Services Partner

数字类型编码

22， 22， 23， 68， 63， 63

- 原始数据
  Min value：22

0， 0， 1， 46， 41， 41

- FOR
  存储与区间最小值差值

22， 0， 1， 45， -3， 0

- DELTA
  存储相邻数据的差值

(22, 2), (23, 1), (68, 1), (63, 2)

- RLE
  将重复值用重复数字代替

0， 0， 1， 2， 3， 3

Dict
0 : 22
1 : 23
2 : 68
3 : 63

- DICT
  字典编码

字符串编码压缩

- Prefix Encoding
  前缀压缩
- RLE
  有序/重复字符串
- Huffman coding
  将重复值用重复数字代替
- FSST coding
  字典编码
- DICT
  字典编码

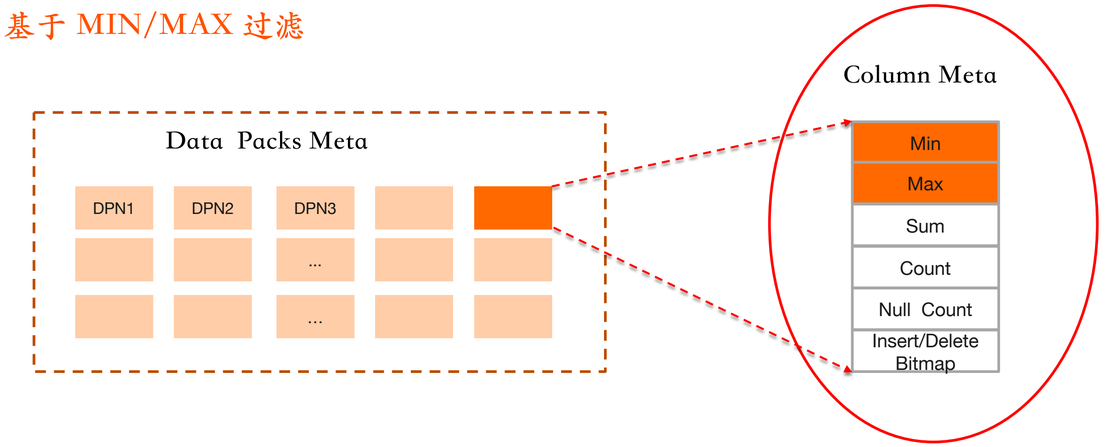
通用块压缩算法

- LZ4
- ZSTD

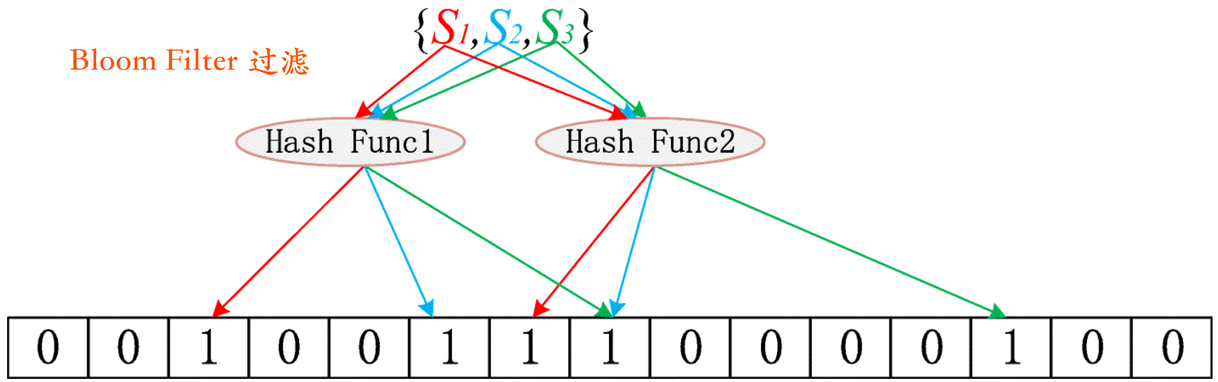
IMCI Storage： 列存辅助索引

Alibaba Cloud | Worldwide Cloud Services Partner

基于 MIN/MAX 过滤



Bloom Filter 过滤



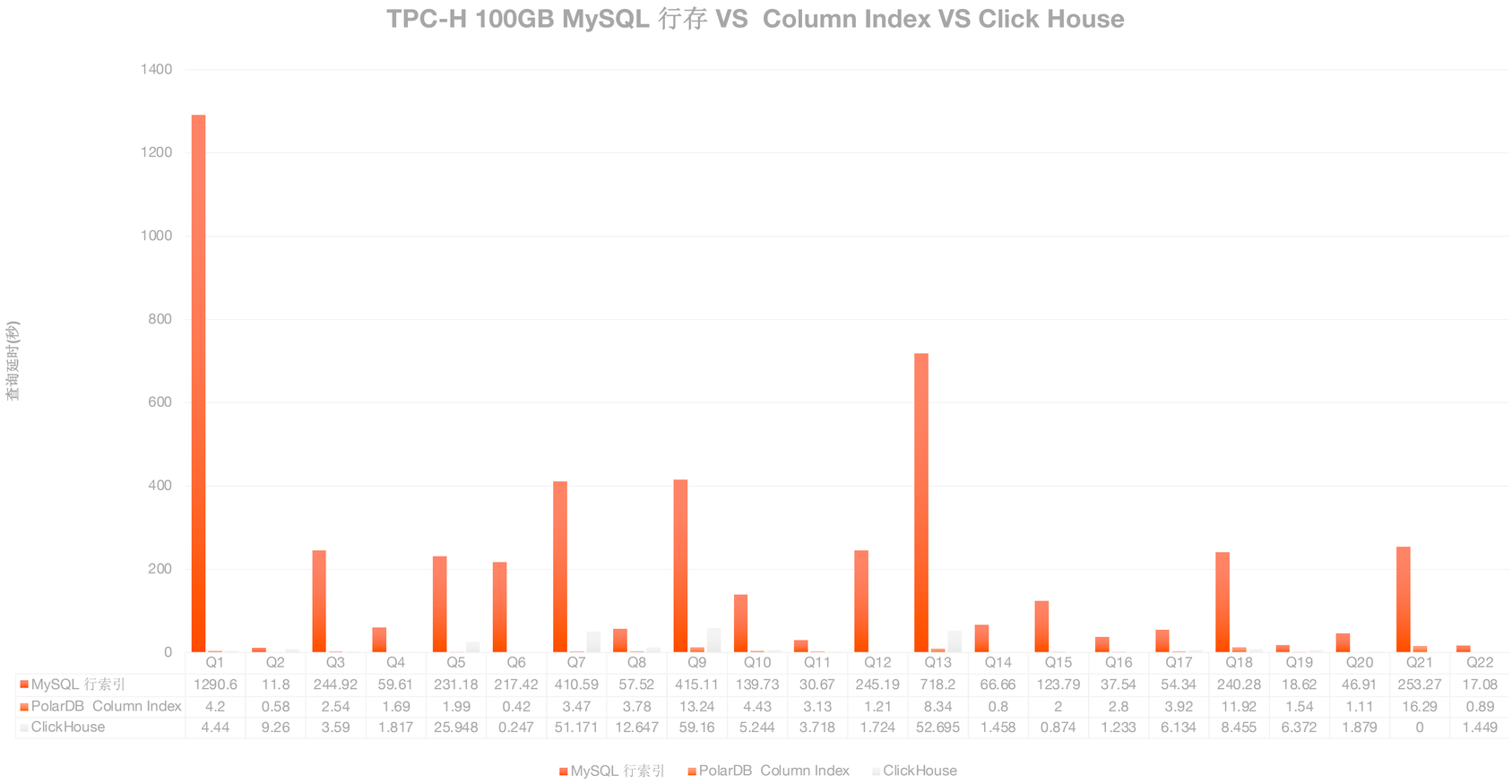
Column Store索引加速

- **Min/Max索引**
  - 列存数据块维护MIN/MAX/SUM等统计信息
  - 基于统计信息实现大块数据的Pruning
- **Histograms**
  - 单个pack按数据min/max等分若干分区
  - 使用bitmap记录对应分区是否有数据命中
- **Bloom Filter**
  - 单个pack中借助多个hash函数构建bitmap
  - 优化等值/in等精确命中查询
- **Adaptive Radix Tree**
  - 单独列再次基于ART 构建有序索引结构
  - 优化字符前缀，范围查询

PolarDB IMCI Performance: TPCH 100G PK Click House

Alibaba Cloud | Worldwide Cloud Services Partner

- TPC-H对比行存加速超100倍
- 80%的SQL比Click House更快



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22
MySQL 行索引	1290.6	11.8	244.92	59.61	231.18	217.42	410.59	57.52	415.11	139.73	30.67	245.19	718.2	66.66	123.79	37.54	54.34	240.28	18.62	46.91	253.27	17.08
PolarDB Column Index	4.2	0.58	2.54	1.69	1.99	0.42	3.47	3.78	13.24	4.43	3.13	1.21	8.34	0.8	2	2.8	3.92	11.92	1.54	1.11	16.29	0.89
ClickHouse	4.44	9.26	3.59	1.817	25.948	0.247	51.171	12.647	59.16	5.244	3.718	1.724	52.695	1.458	0.874	1.233	6.134	8.455	6.372	1.879	0	1.449

Max DOP = 32 Data All in Memory

目录

Alibaba Cloud | Worldwide Cloud Services Partner

跨境电商：易仓科技

Alibaba Cloud | Worldwide Cloud Services Partner

分析业务

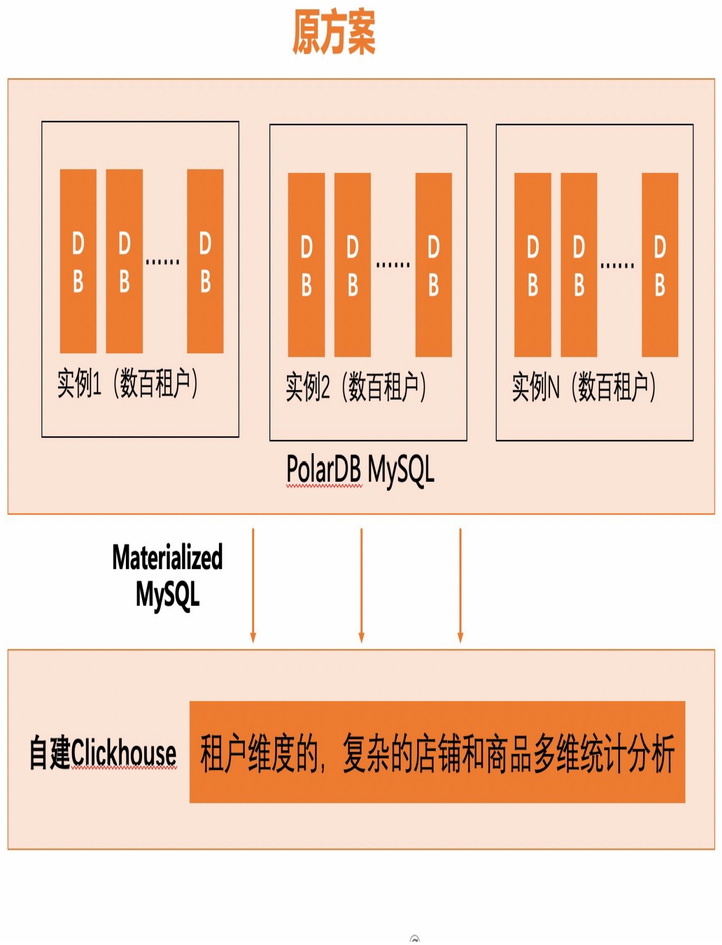
- 店铺仓储管理
- 商品推广分析

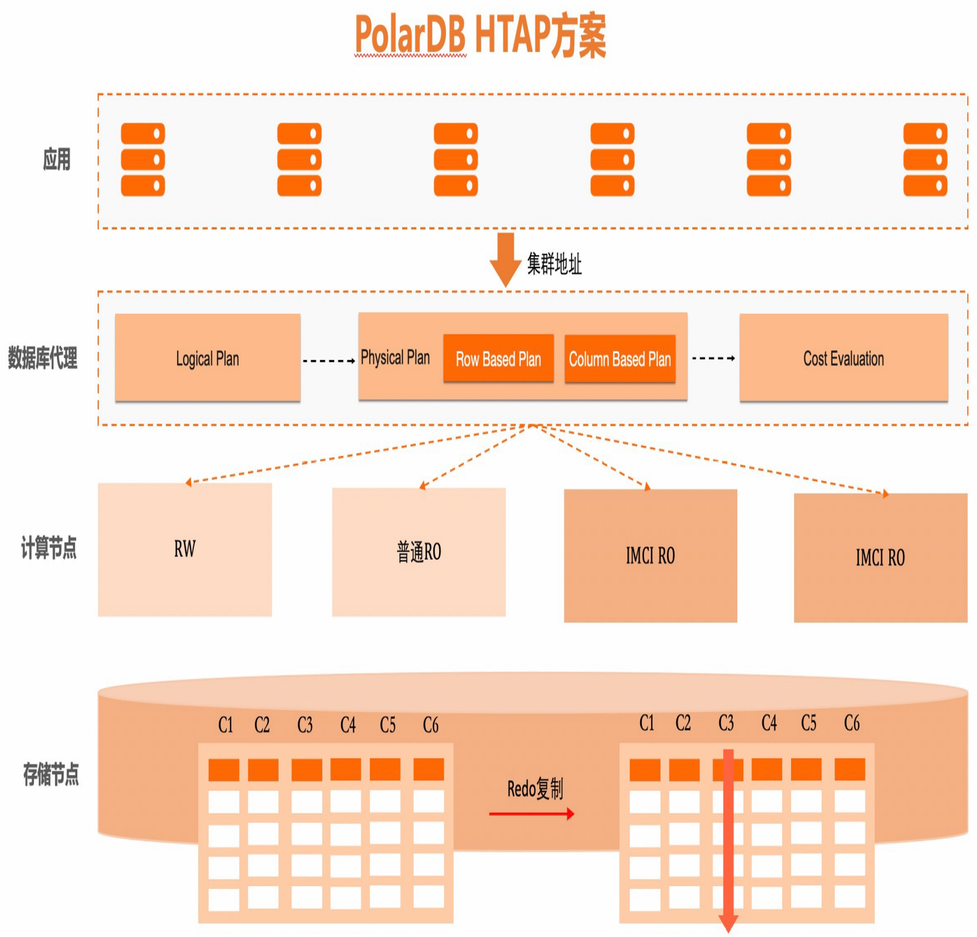
原有架构问题

- 复制延迟大
- 性能瓶颈
- 管理复杂

IMCI核心优势

- 简单易用，100%兼容MySQL
- 实时性好
- 成本低





数字营销：云想科技

Alibaba Cloud | Worldwide Cloud Services Partner

分析业务

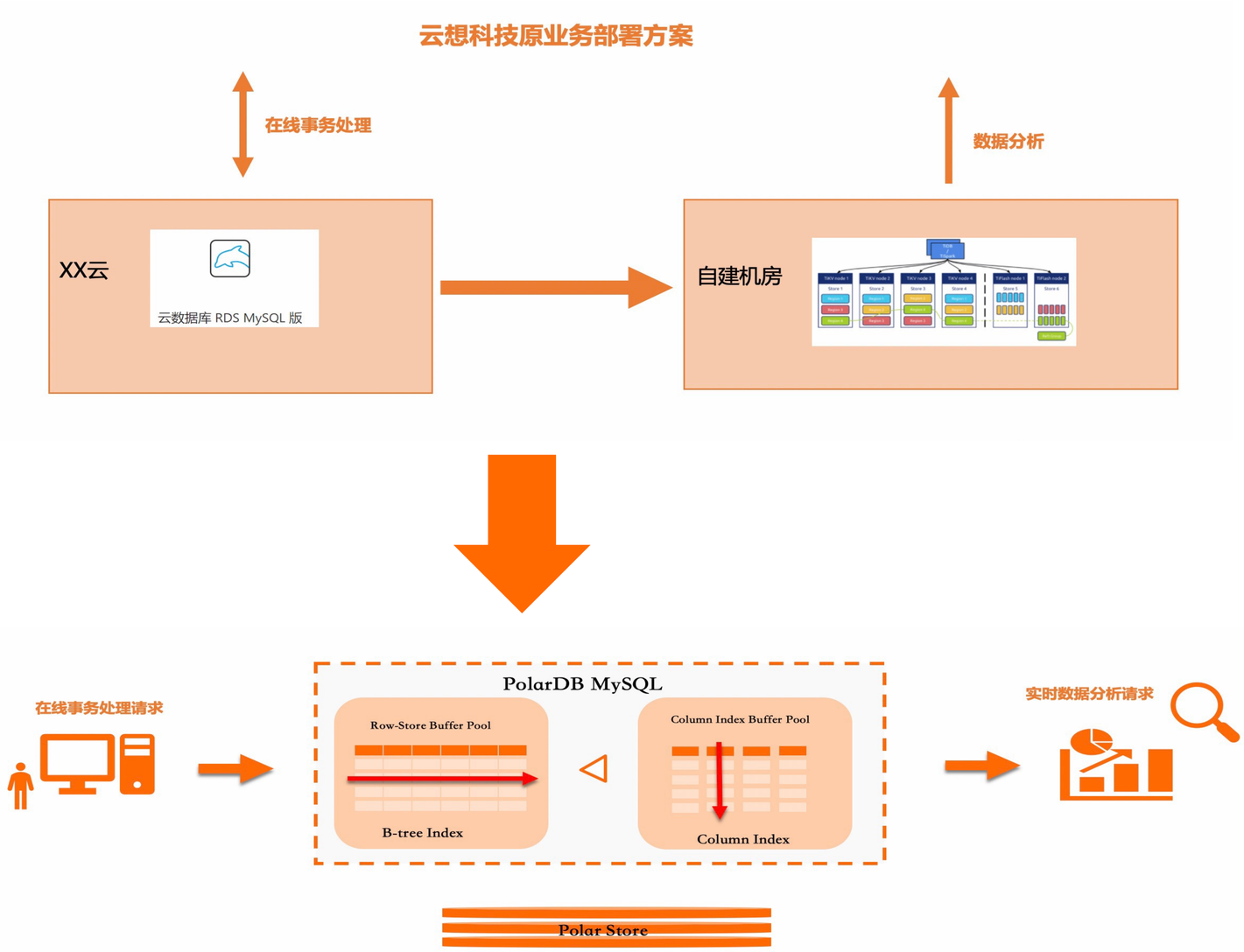
- 短视频精准营销
- 广告智能投放

原有架构问题

- 公司自建网络稳定性差
- TiDB 串行DDL变更卡顿
- 运维TiDB人力成本
- 两套系统一致性问题

IMCI核心优势

- 一套系统同时解决TP&AP
- Instant-DDL，并行DDL



目录

Alibaba Cloud | Worldwide Cloud Services Partner

总结与展望

Alibaba Cloud | Worldwide Cloud Services Partner

PolarDB(HTAP)不是简单的TP+AP

- AP性能
- 数据实时性
- AP&TP隔离性
- 弹性扩缩容
- MySQL兼容性

下一步发展

- MPP 集群，更大容量数据分析
- 行列混合执行
- 列存表，冷热分离，降低分析成本

Alibaba Cloud | Worldwide Cloud Services Partner

Thank You