

数字转型 架构演进

SACC

2019 中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2019



2019年10月31-11月2日



北京海淀永泰福朋喜来登酒店

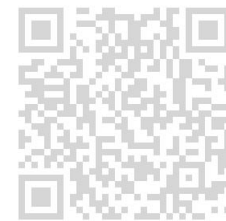
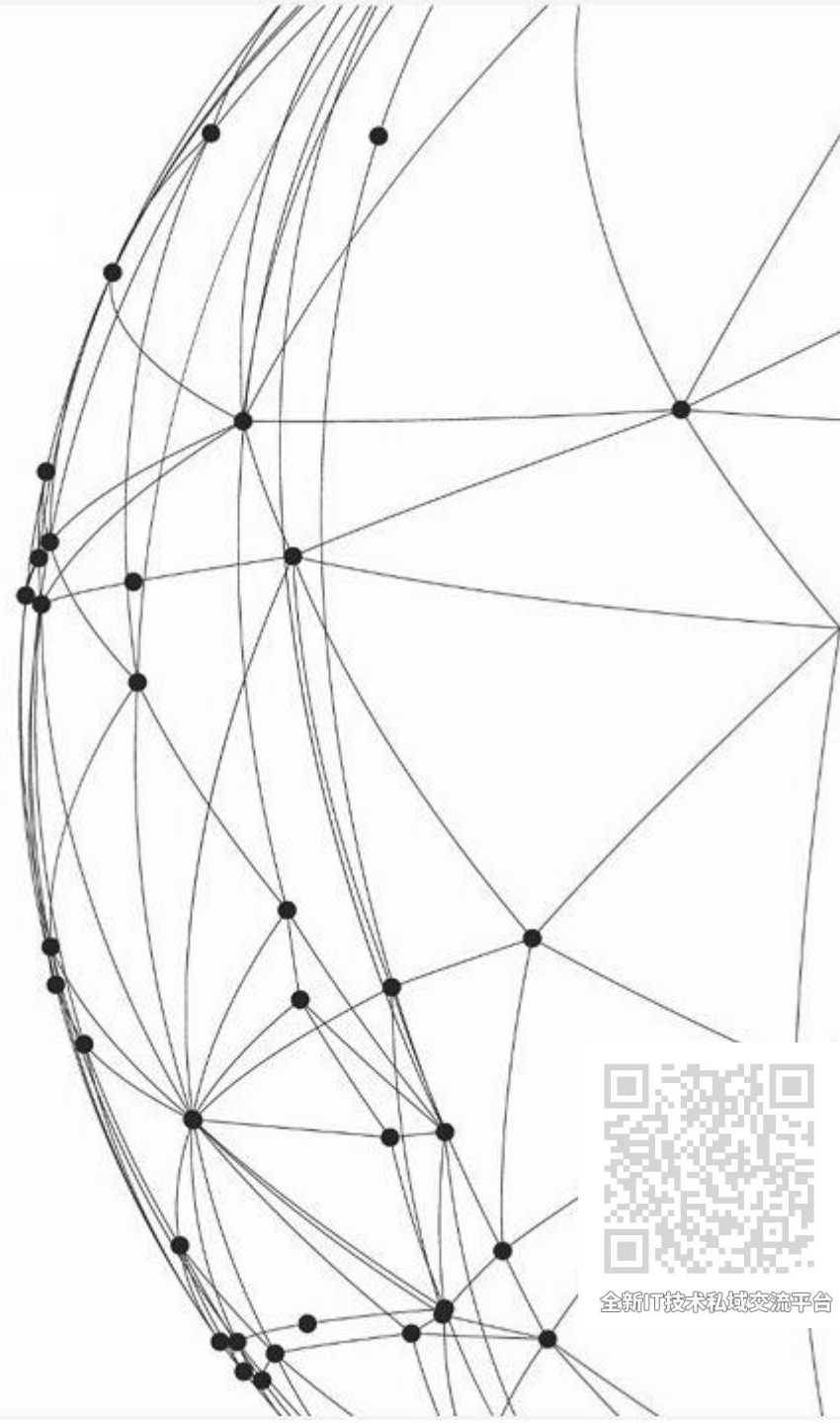


全新IT技术私域交流平台



达梦数据库 国产化推进实践与思考

武汉达梦数据库有限公司 郭一兵

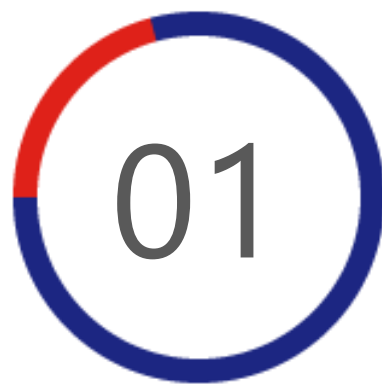




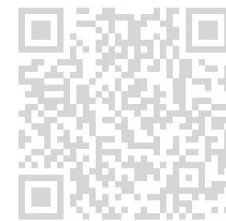
目录 CONTENTS

- 01 简介
- 02 技术路线发展
- 03 **DM8 增强与改进**
- 04 **DM8 架构新趋势**
- 05 总结





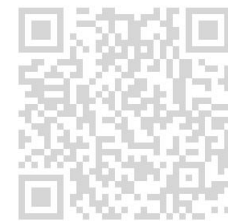
简介



达梦数据库

国产数据库软件厂商

愿景——坚持原始创新、实现产业报国



1978-
1988

高校及科研机构
的研究、探索

理论探索、
原型研究

1989-
2000

原型研究，产
品开发

产品研究

2001-
2012

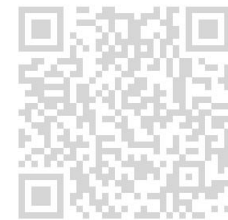
公司化运作，
产学研结合，
示范推广

成果转化期

2013-
至今

全面推广应用，
市场化竞争

市场竞争期



全新IT技术私域交流平台

达梦简介——产品发布历史

1992年数据库与
多媒体研究所

2000年武汉
达梦数据库公司

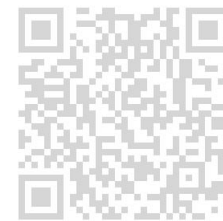
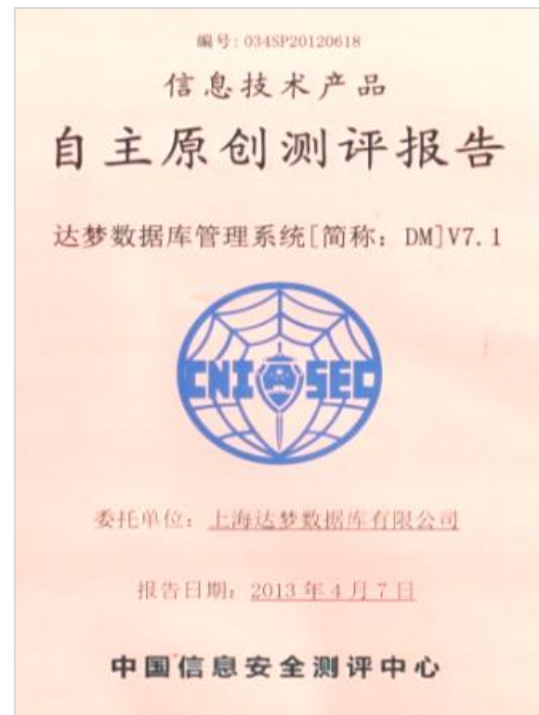
2001年上海
达梦数据库公司

2002年北京
达梦数据库公司

成都、石家庄、
广州、海口等



- 早期基于DOS Pascal/ Vax-11 Ada编写
- X用MDB, 知识库KDB, ADB等成果
- 91年开始基于C/XENIX 多用户
- 92年成立研究所，后推出DM1

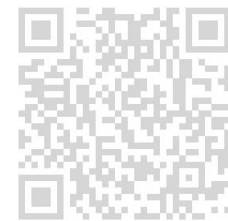
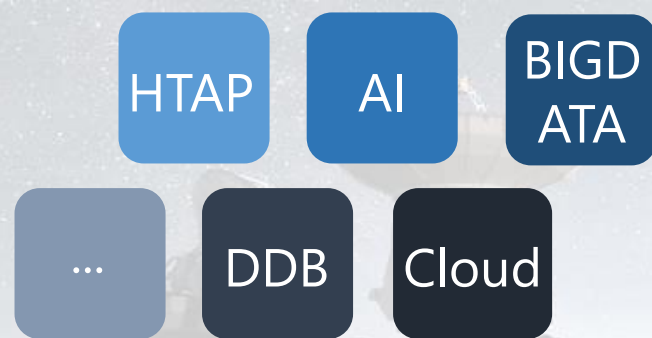


全新IT技术私域交流平台

重视细节、简单实用



功能完备、持续创新



提供丰富的数据处理产品和解决方案：

- 数据交换
- 数据管理
- 数据分析

以关系数据库(DM8)为基础



交易、办公、OA、
网站等类型应用

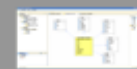


数据分析、报表、决
策支持等类型应用

DM OLAP



数据仓库建模



数据挖掘



数据比对



数据分析

元数据管理



数据质量管理



数据资源管理



数据服务管理



数据管理支撑

实时数据同步工具DMHS



数据交换平台DMETL



数据交换

共享存储
集群



读写分离
集群



大规模并行
集群MPP



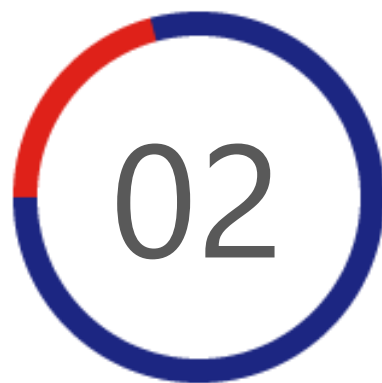
达梦非结构化
数据库MGBASE



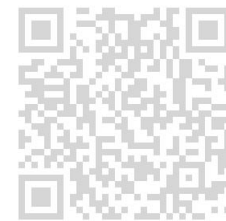
达梦数据库DMDBMS



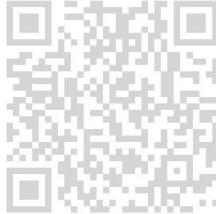
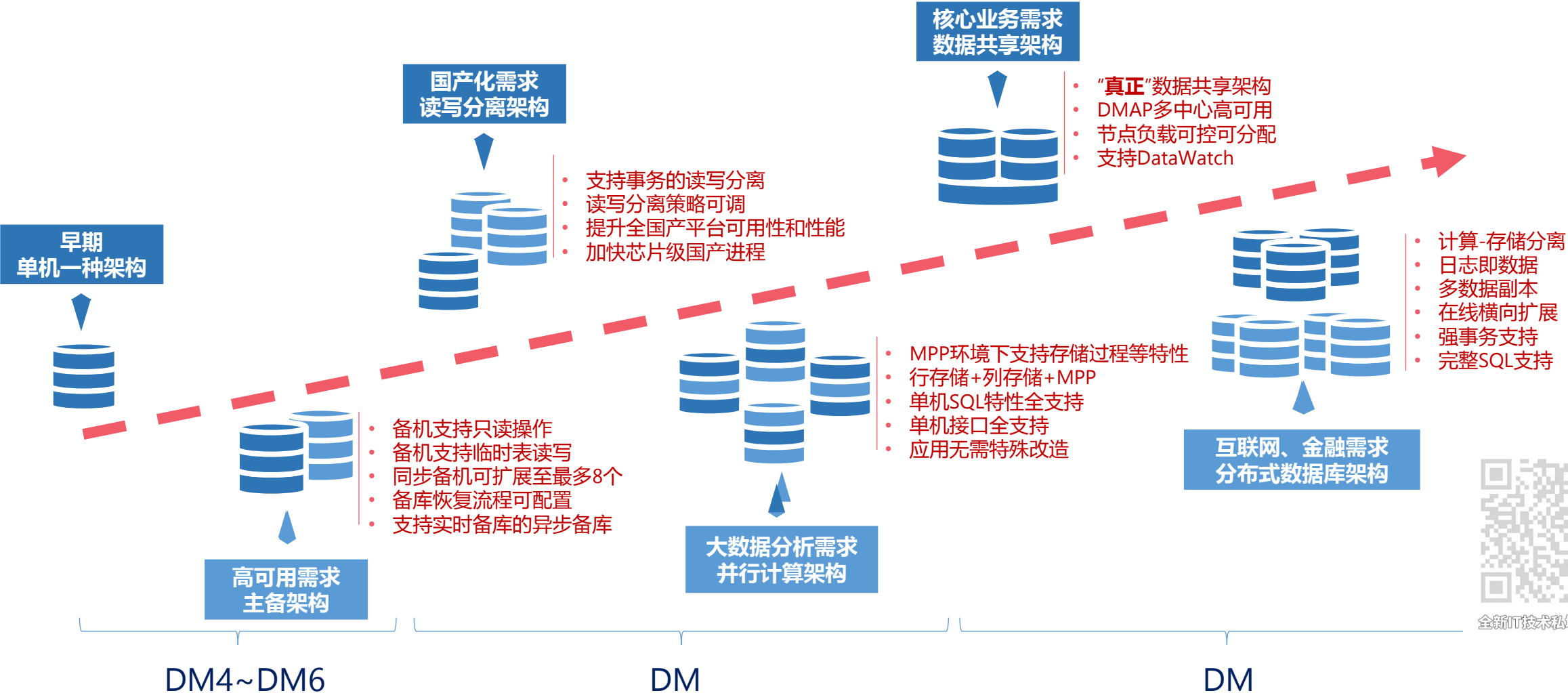
全新IT技术私域交流平台



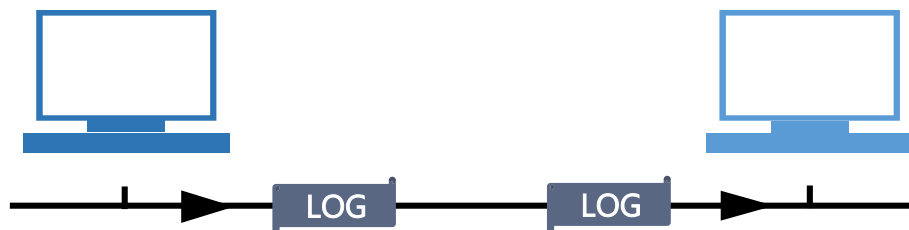
技术路线发展



国产数据库的发展应源于市场需求



2011



国家电网，高可用

中国铁建，备机资源利用率

实时复制+异步复制可配

支持1主8备扩展，支持只读事务

自动/手动切换

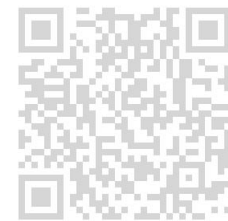
强备机：临时表读写等增强特性

系统整体可用性达99.99%

备机支撑报表业务

实现异地（上百公里级）热备

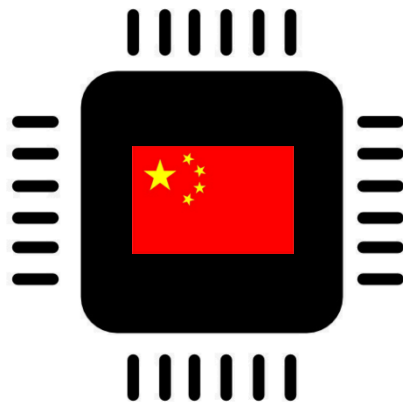
提升了资源利用率与投资效益



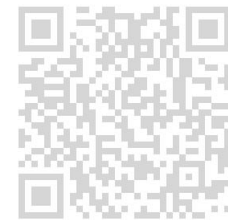
2012

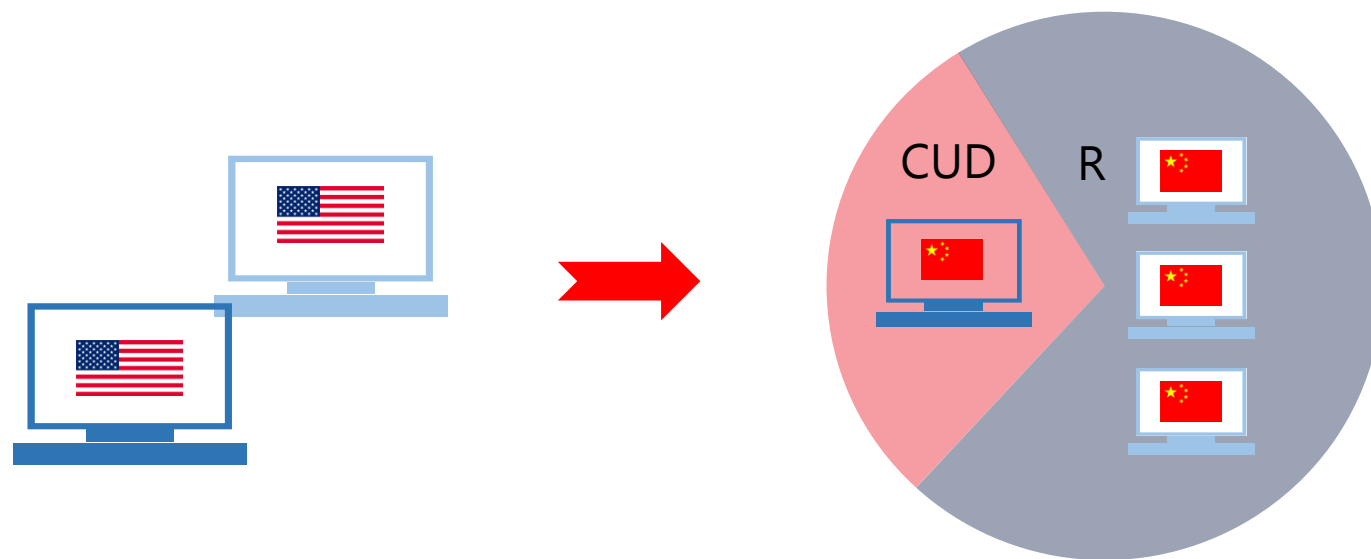
作为国产数据库厂商，有义务促进全国产生态的发展

基于国产自主CPU整机环境的性能、高可用如何保证



主频 内核数 内存容量 存储规格 ...





**大多数业务场景
符合读多写少特征**

读写分离+负载均衡
自动故障切换
读提交事务隔离级



架构一脉相承

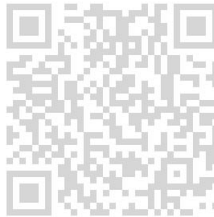
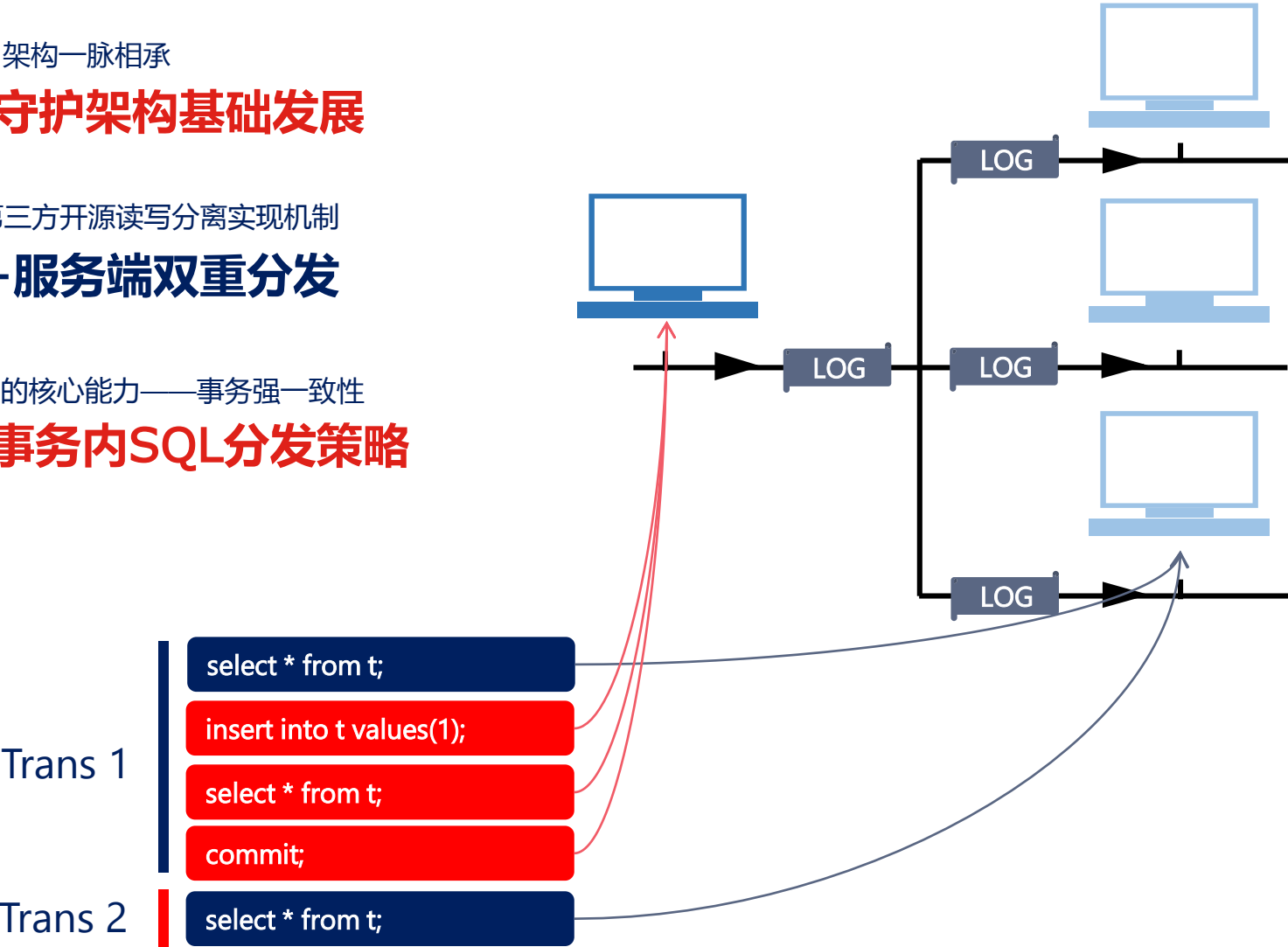
继承数据守护架构基础发展

不同于大量第三方开源读写分离实现机制

驱动端+服务端双重分发

提供企业级产品的核心能力——事务强一致性

基于严格的事务内SQL分发策略



场景	响应时间 (单节点, 200Vuser)	响应时间 (1+2集群, 200Vuser)	响应时间 (1+5集群, 500Vuser)
登录	5.881S	3.413S	4.730S
新建	3.620S	2.161S	4.169S
发文	4.076S	2.258S	3.715S
退出	3.819S	0.049S	1.721S

某国产CPU平台环境下，OA典型操作响应时间

部委

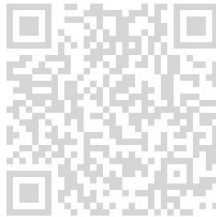
10+

省市

30+

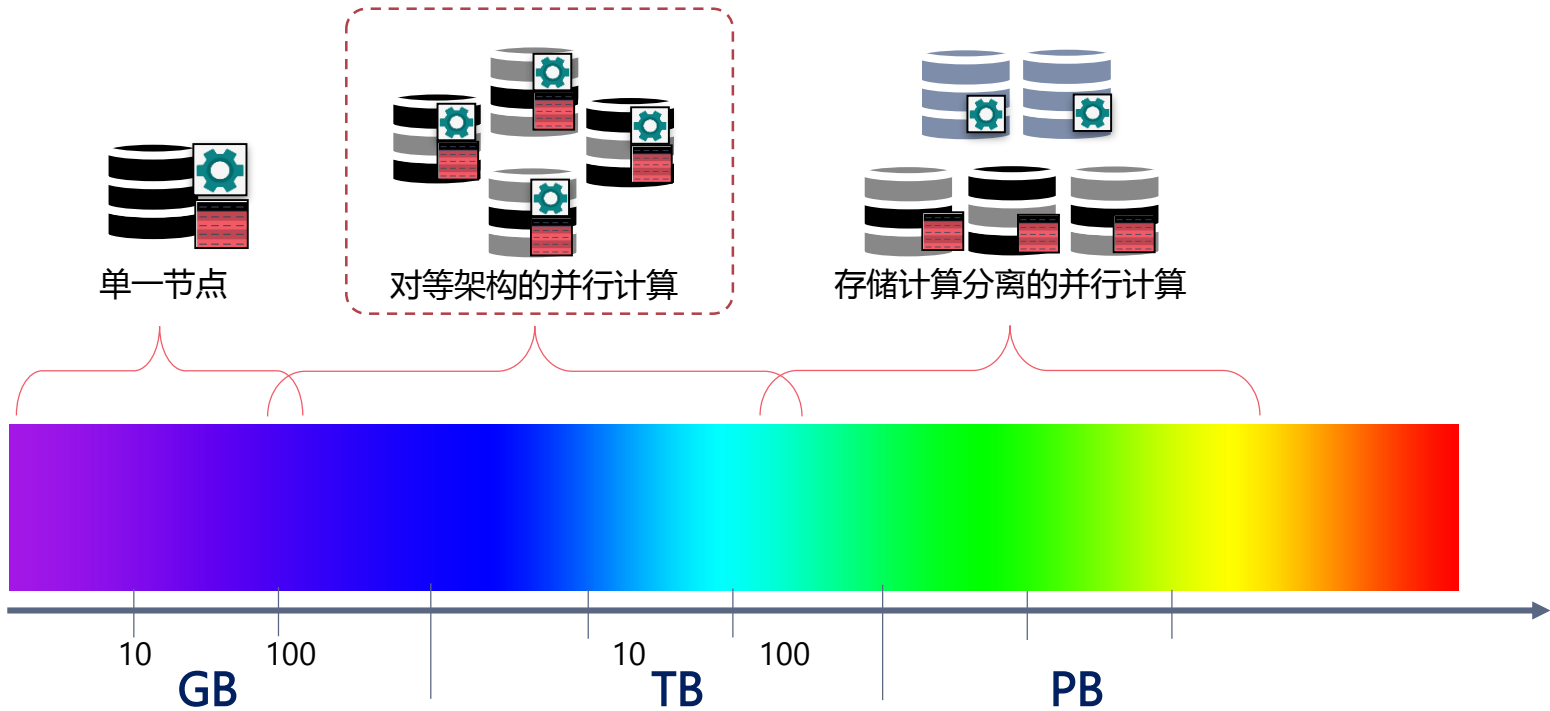
央企

10+



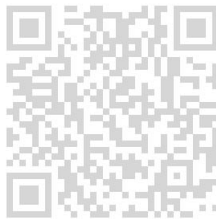
2011

某部门分析业务，TB级，单一节点无法支撑



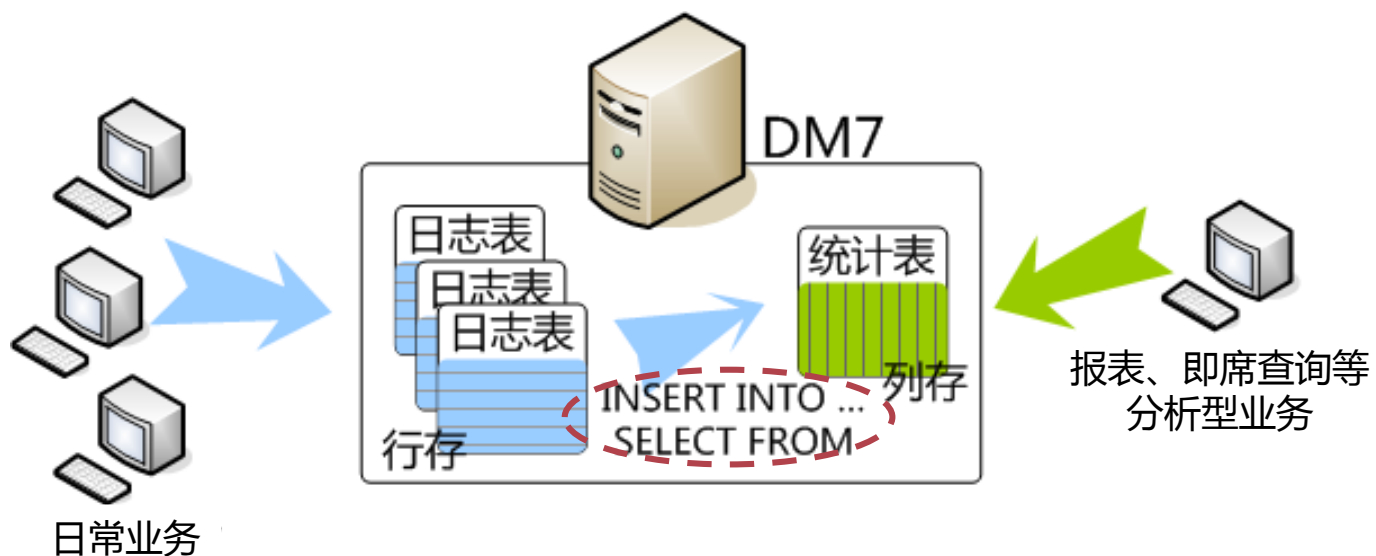
纯分析：MPP+列存储

混合负载：MPP+列存储+行存储



2015

MPP上的行列融合——高速转换

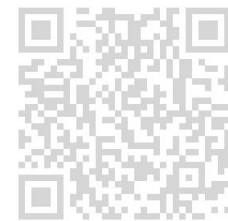


■ 高效行存转入列存，DB内核级优化

■ 节点内高速总线>>网络传输

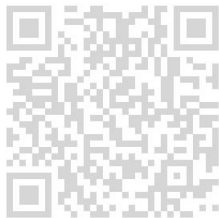
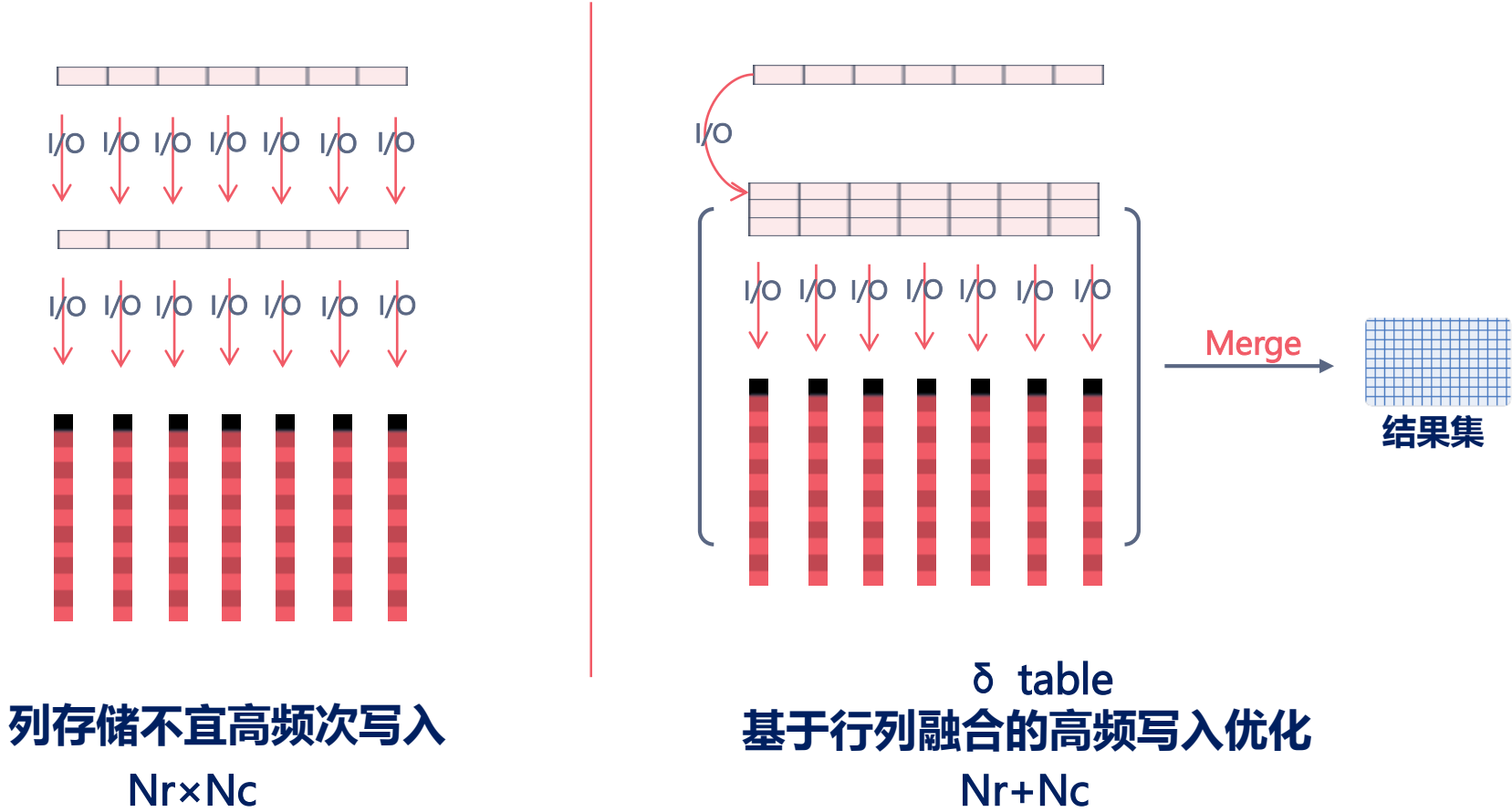
■ 并行读取、计算

■ 内部采用并行化FastLoader装载路径



2017

MPP上的行列融合——列表的快速并发写入优化



一套
MPP

高频插入+并发精确查询

大规模数据集上的统计分析

河北公安云

吉林公安云

湖北公安云

南京警务平台

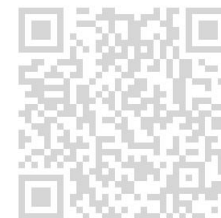
黑龙江公安

国家工商总局

国家电网

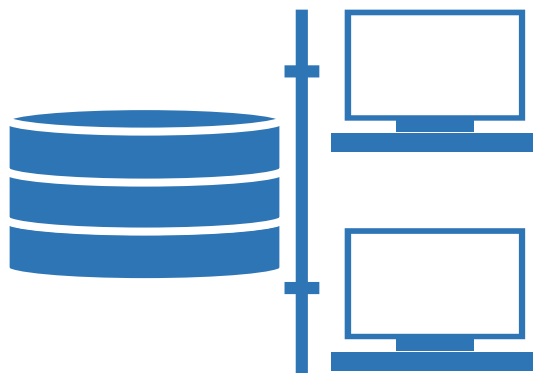
广州政务云

工信部



2014

国产数据库生态被提问最多的话题



不是唯一选择，但仍是好的选择



2018



高可用性

- 故障节点的连接自动切换到活动节点
- 故障节点恢复后自动重加入



高吞吐量

- 多节点同时提供数据库读写服务
- 通过缓存交换提升共享数据访问速度



负载均衡

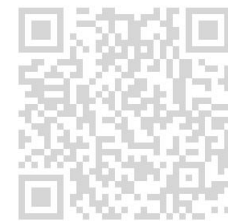
- 并发连接请求被自动、平均分配到各节点



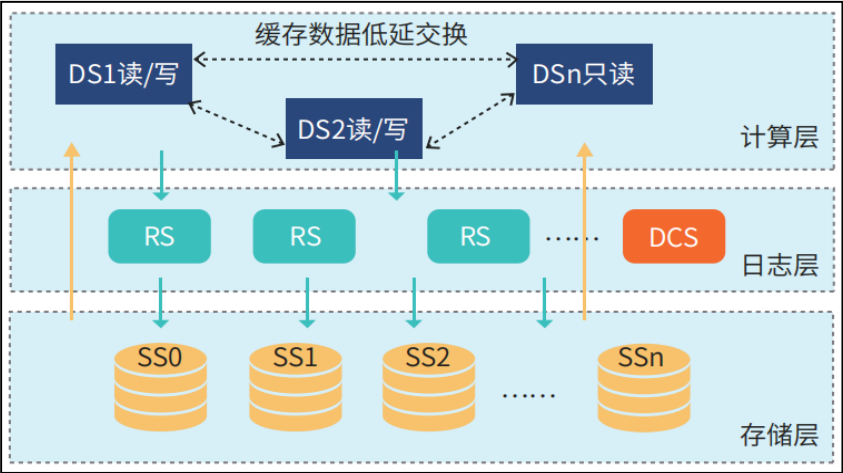
2015 合作伙伴提出分库分表的方案要求，希望改善国产平台上的性能表现



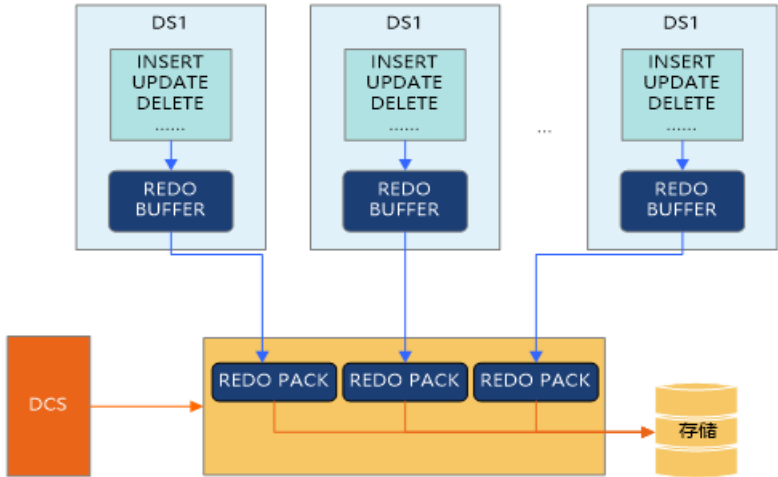
分库分表方案在数据库自身的能力方面妥协过多，DM更认可分布式数据库方案



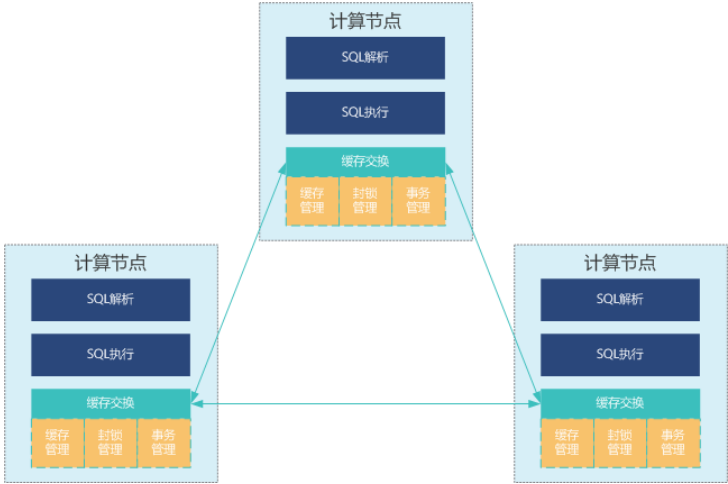
- 2019
- 支持完整SQL特性
 - 多点写入
 - 多副本容灾能力
 - 在线扩/缩容
 - 完整的安全功能特性



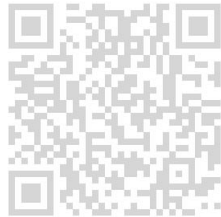
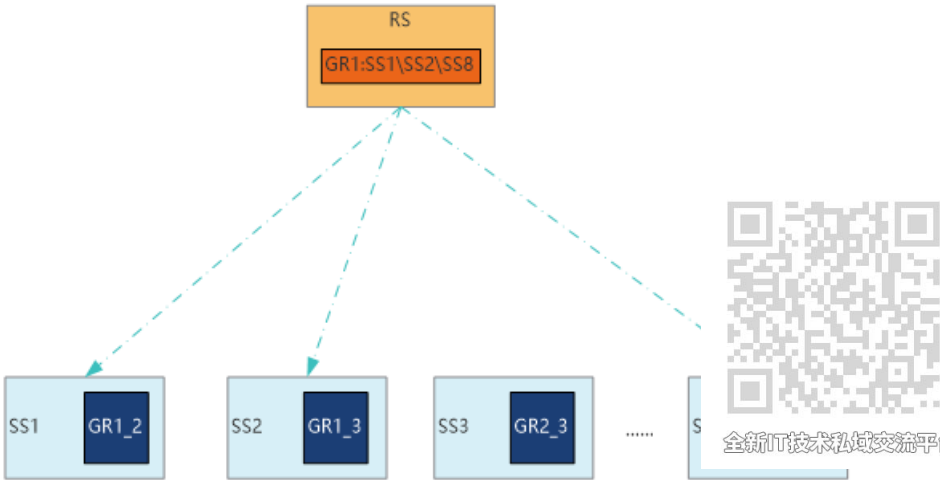
日志

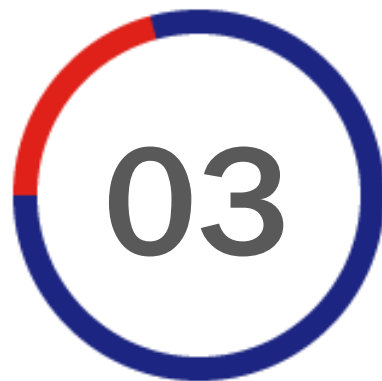


计算

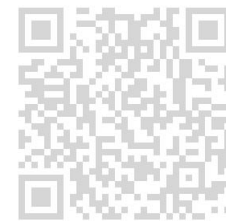


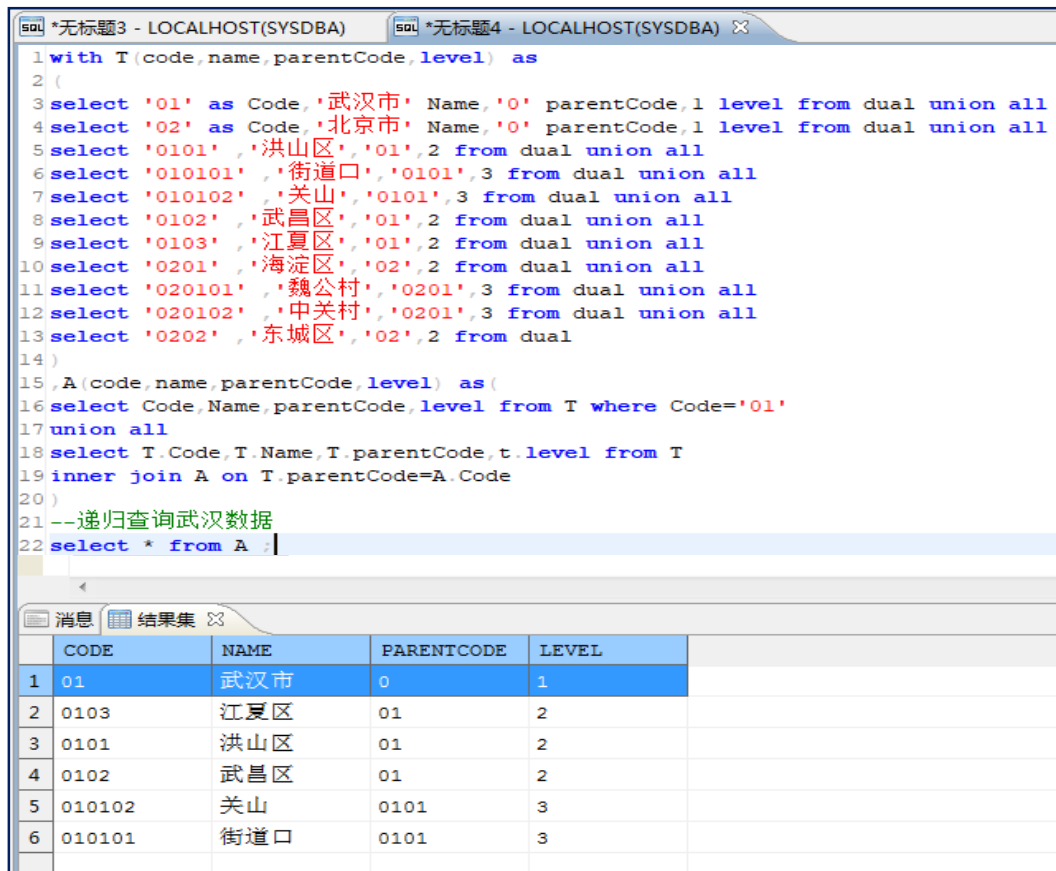
存储





DM8 增强与改进





```
1 with T (code, name, parentCode, level) as
2 (
3 select '01' as Code, '武汉市' Name, '0' parentCode, 1 level from dual union all
4 select '02' as Code, '北京市' Name, '0' parentCode, 1 level from dual union all
5 select '0101' , '洪山区' , '01' , 2 from dual union all
6 select '010101' , '街道口' , '0101' , 3 from dual union all
7 select '010102' , '关山' , '0101' , 3 from dual union all
8 select '0102' , '武昌区' , '01' , 2 from dual union all
9 select '0103' , '江夏区' , '01' , 2 from dual union all
10 select '0201' , '海淀区' , '02' , 2 from dual union all
11 select '020101' , '魏公村' , '0201' , 3 from dual union all
12 select '020102' , '中关村' , '0201' , 3 from dual union all
13 select '0202' , '东城区' , '02' , 2 from dual
14 )
15 , A (code, name, parentCode, level) as (
16 select Code, Name, parentCode, level from T where Code='01'
17 union all
18 select T.Code, T.Name, T.parentCode, t.level from T
19 inner join A on T.parentCode=A.Code
20 )
21 --递归查询武汉数据
22 select * from A ;
```

	CODE	NAME	PARENTCODE	LEVEL
1	01	武汉市	0	1
2	0103	江夏区	01	2
3	0101	洪山区	01	2
4	0102	武昌区	01	2
5	010102	关山	0101	3
6	010101	街道口	0101	3

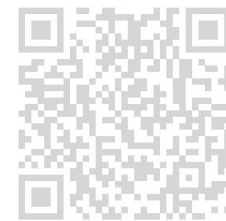
符合新旧国际标准规范

- JDBC4.2 , PHP7.1等规范更新
- OCI7.X等旧版规范支持

兼容业界事实标准

- 新增22个系统Package , 兼容Pack总数达到33个
- Partition outer join、递归的CTE表达式、正则表达式的反向引用等大型细节

■ 保护用户原有投资 , 力图完美兼容现有应用、开发模式



存储过程

- 金融、财政、ERP等方向大量应用
- 调试难、调优难

调优调试改进

- 通过断点，查看SP中DML的执行计划
- 通过V\$DMSQL_EXEC_TIME视图，跟踪SP内多个嵌套调用的执行效率
- 通过V\$SESSIONS视图，查看会话当前SP执行进度

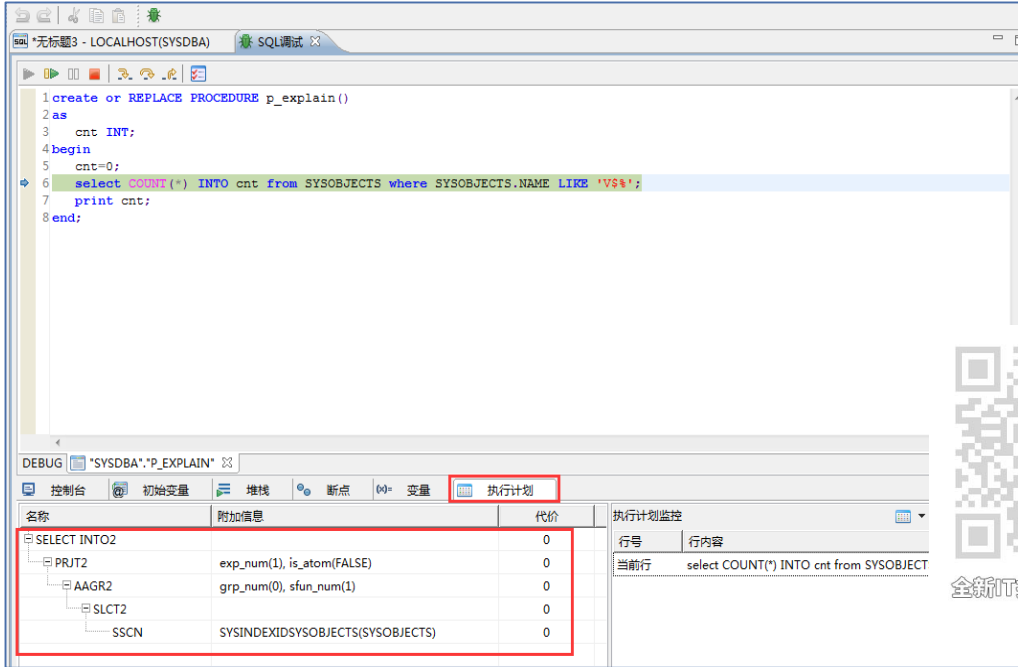
怎样评价一个几千行的 SQL 存储过程？

4 人赞同了该回答

曾经在国内某知名做进销存的公司呆了一年，逻辑全在存储过程里面，遇到过最大的一个存储过程的sql文件是三百多Kb的样子。。。调试只能print，然后脑补这个三层游标里面还是一大串的sql最后会出现什么结果。。。然后一路print或者select。最后实在是觉得狗，年底双薪都没要果断跑路了。。。

发布于 2019-02-14

赞同 4 6 条评论 分享 收藏 感谢 ...

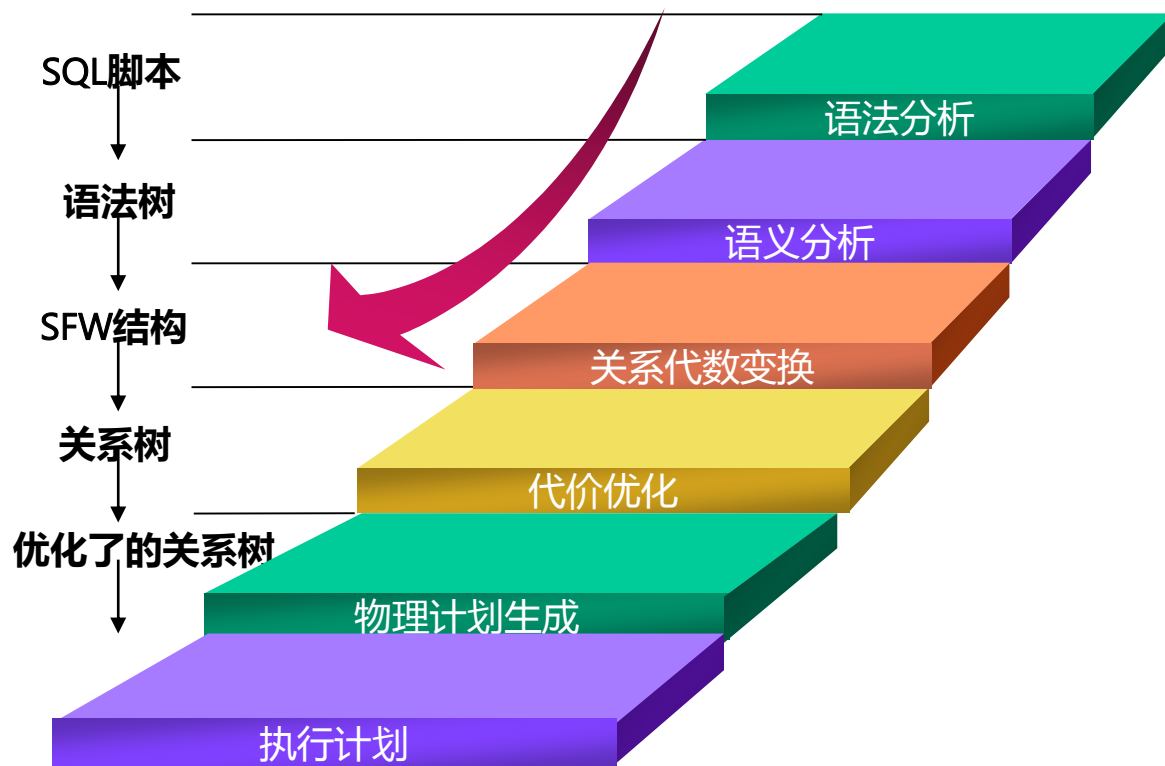


截图展示了达梦数据库的 SQL 调试界面。上方是 SQL 代码编辑区，显示了一个名为 p_explain 的存储过程，其中包含一个 select 语句，用于统计 SYSOBJECTS 表中名称以 'V\$' 开头的对象数量。下方是调试器窗口，显示了当前正在执行的 SQL 语句及其执行计划。执行计划窗口中，SELECT INTO 语句的执行计划被高亮显示，其子计划包括 PRJT2、AAGR2、SLCT2 和 SSCN。右侧的执行计划监控窗口显示了当前行的执行内容。

名称	附加信息	代价
SELECT INTO2		0
PRJT2	exp_num(1), is_atom(FALSE)	0
AAGR2	grp_num(0), sfun_num(1)	0
SLCT2		0
SSCN	SYSINDEXIDSYSOBJECTS(SYSOBJECTS)	0

行号	行内容
当前行	select COUNT(*) INTO cnt from SYSOBJECT

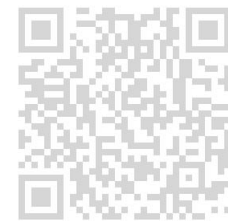


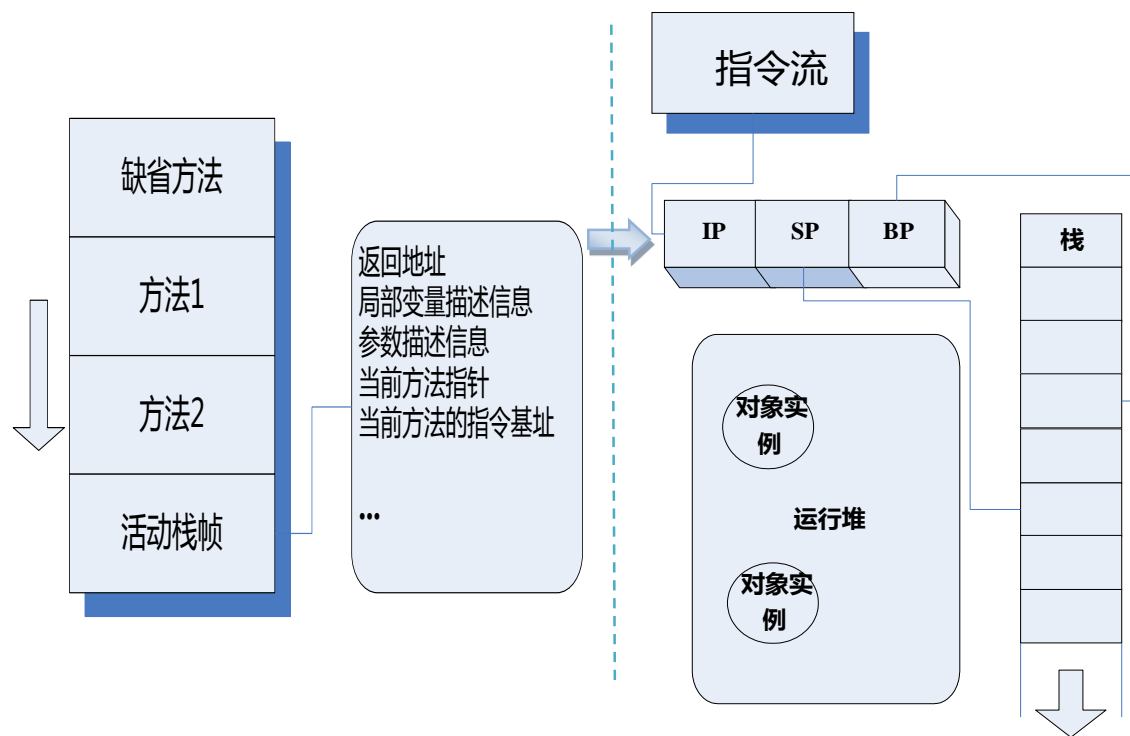


大量增强特性

- 多维统计信息
- 优化参数的自适应支持
- 子查询合并优化
- 复杂表达式优化
- 过滤表消除优化
- 子查询优化扩展
- 视图条件下放优化增强
- 语句块独立HINT支持

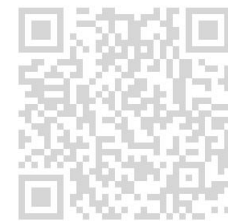
■ 支持机器生成的复杂SQL; 分析型即席复杂查询





- Redo日志包
- 并行Redo日志
- 并行Purge
- 独立的回滚段分片缓存
- 多级分区表扩展
- 物化视图功能增强
- DBlink 增强
-

■ 创新的虚拟机执行内核，完美支持SQL和面向对象的过程语言



TPC-C基准测试

99.4W tpmC

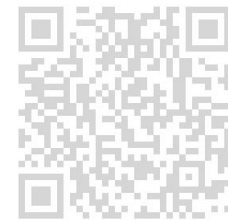
2X Intel Xeon E5-2697v4 Centos 7.3
24X 16GB DDR4 BenchmarkSQL5.0
2X 300GB 15KRPM SAS
1X 10Gbps

单机成本
10W级别

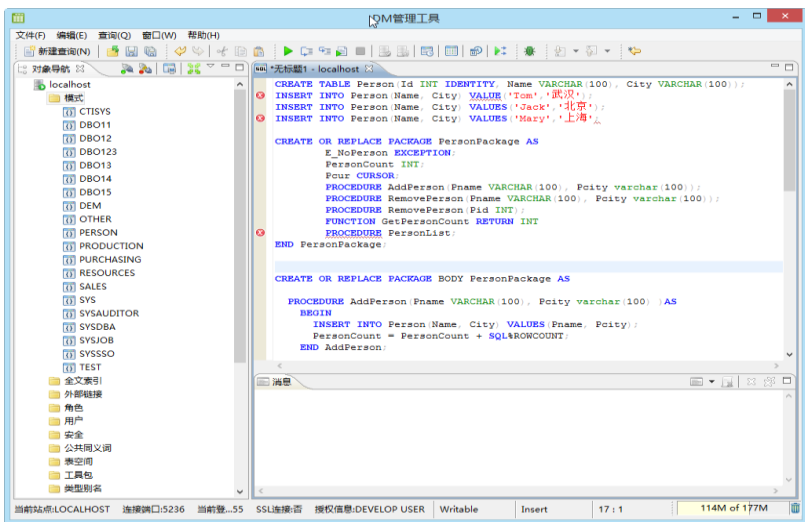
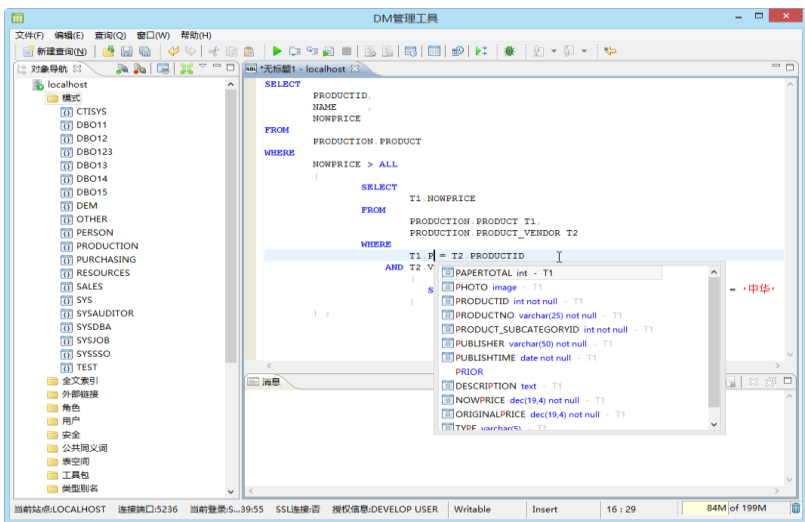
TPC-H@1TB基准测试

17 Min

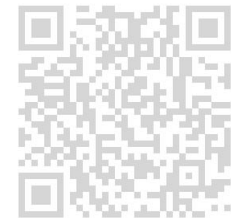
4节点MPP
2X Intel Xeon E5-2650v2 1X 3.2TB PCI-E
16X 16GB DDR4 1X 1Gbps



DM8——更加便捷的管理和运维工具



- SQL助手2.0
 - 联机 and 脱机运行环境
 - SQL语法正确性提示
 - 包、类、自定义类型、过程、函数的语法树展现及快速定位功能；
- DEM部署支持DSC集群
- DTS提供Web版本



■ 运维管理更直观

服务器活动会话数异常

- >分析检查业务有无异动
- >分析中间件参数是否异常
- >分析SQL状态是否异常

.....

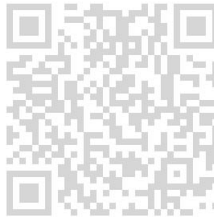
不同的工具、不同的手段

■ NOW

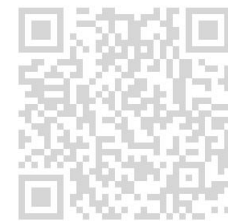
异常状态时间点的点选

+

按时间点的SQL请求状态监控

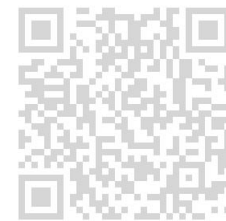


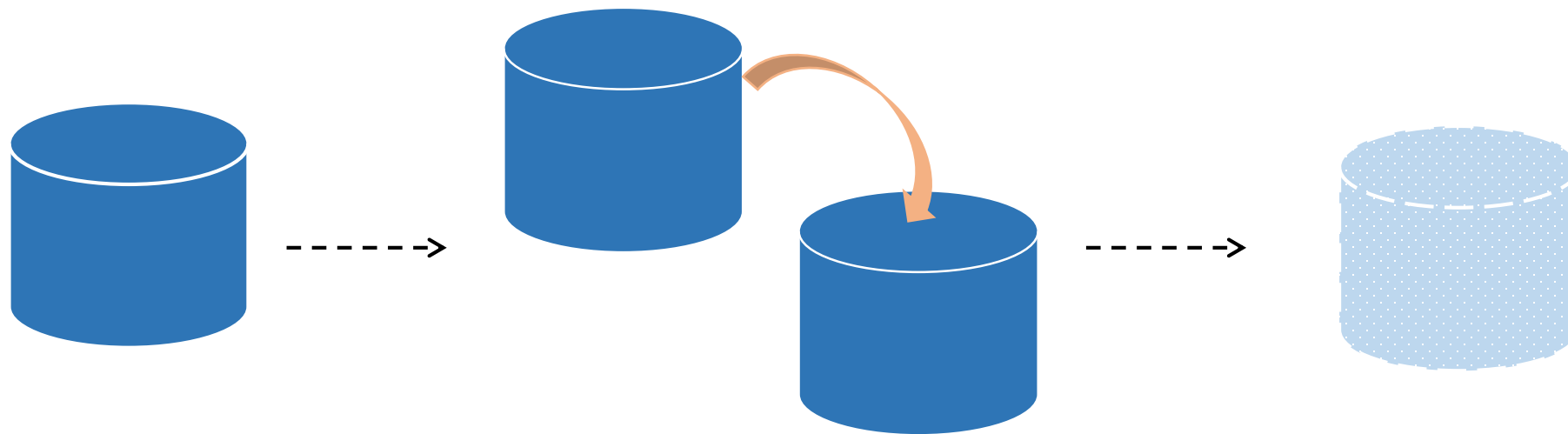
- 服务器性能空前强大
 - 单片**CPU**集成数十个核心
 - 多路扩展
 - **TB**级内存
 - 大容量高速固态存储（**pci-e NVMe**）
 - 高速网络，**5G**
 - 国产硬件持续进步
- 充分利用单机性能，为用户节约硬件投资和运维成本





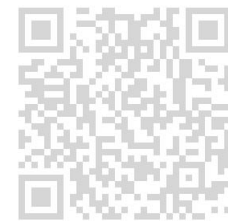
DM8 架构新趋势





HTAP=OLTP+OLAP

?



高级日志

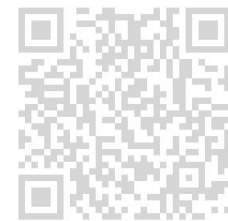
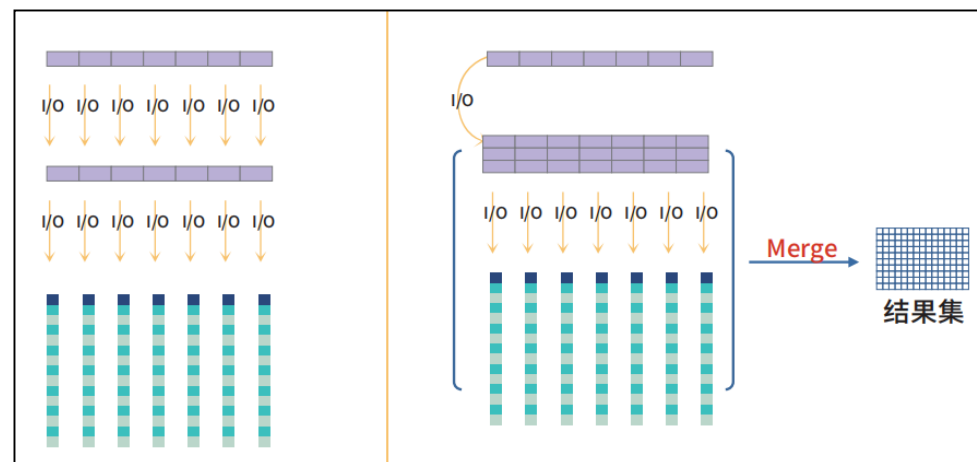
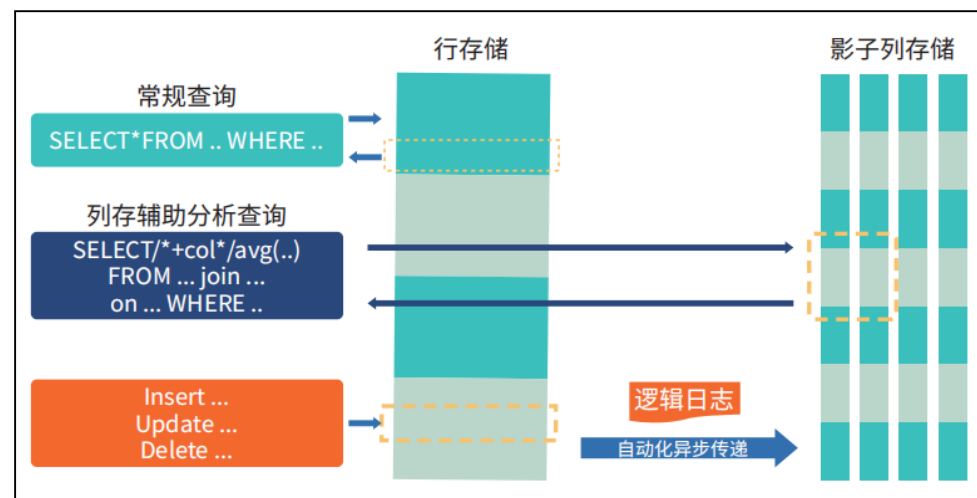
- 行列引擎融合
- 专利号：201810827900.1

SQL引擎改进

- 自动合理优化
- HINT辅助选择存储引擎

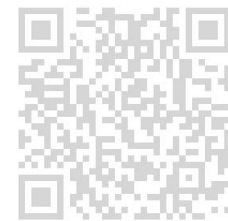
变更缓存机制

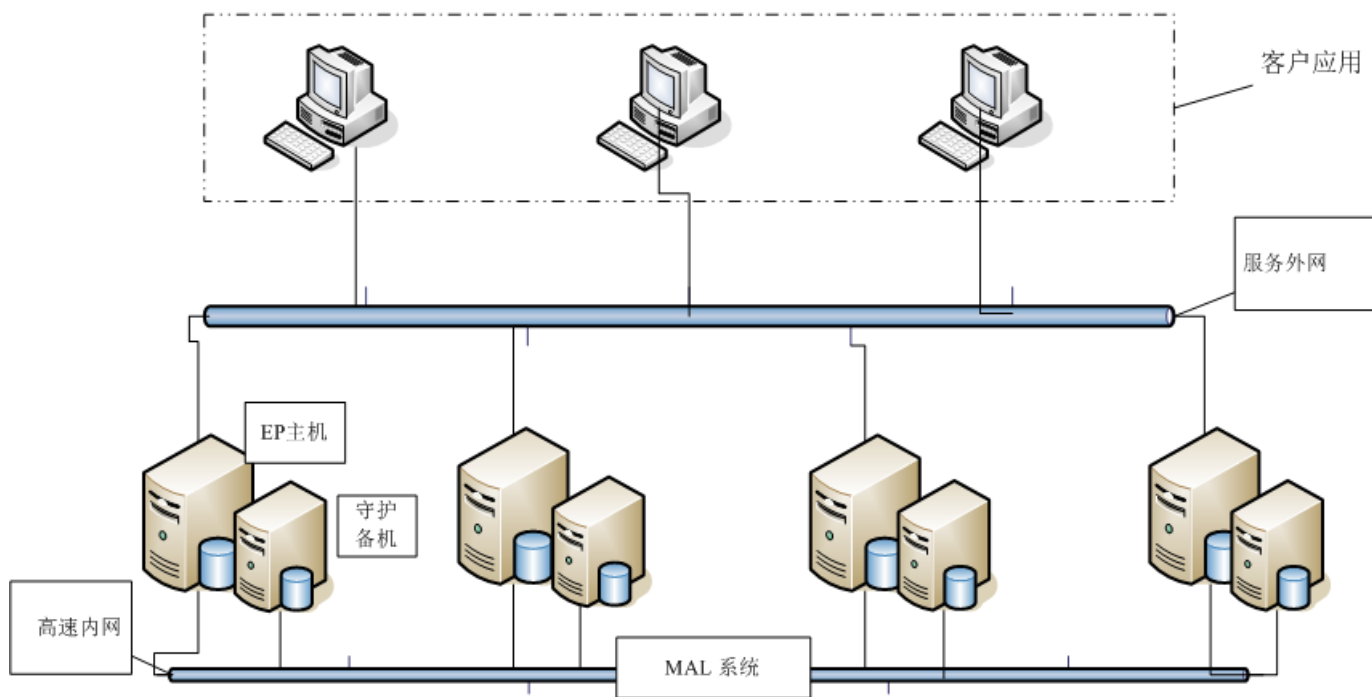
- 使列存引擎支持OLTP特征并发写访问



- 集群与单机系统总体表现相同
 - SQL92标准
 - 各种标准接口
 - 全功能支持：复杂查询，支持视图、存储过程、触发器、序列
- 集群基于单机系统构建
- 重用超过**95%**的代码路径

■ 一套代码、一份介质、按需搭建





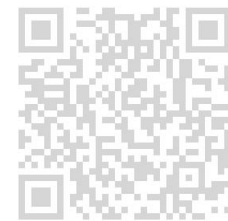
- ➡ **并行处理**
- 节点之间完全无共享，每个节点只处理部分数据
 - 所有节点同时并行处理

- 🔍 **可扩展性**
- 支持达梦透明分布式存储

- 📈 **高性能**
- 继承达梦数据库的功能和性能优势
 - 支持查询内并行；

- 🗄️ **优化的数据存储**
- 行列混合存储
 - 多级数据压缩

■ 优化器综合考虑通讯代价生成MPP计划

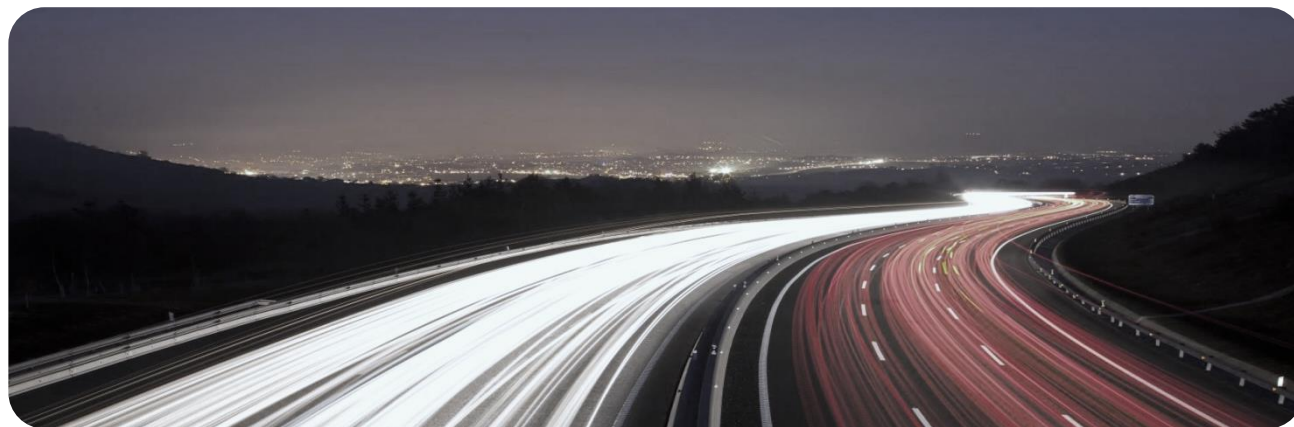


■ 软硬件技术发展的新机遇

NVMe SSD

RDMA

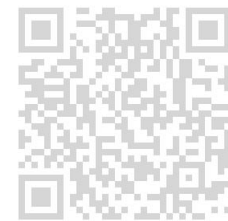
大内存



■ 显著的优势

多点读写

完整的数据库特性

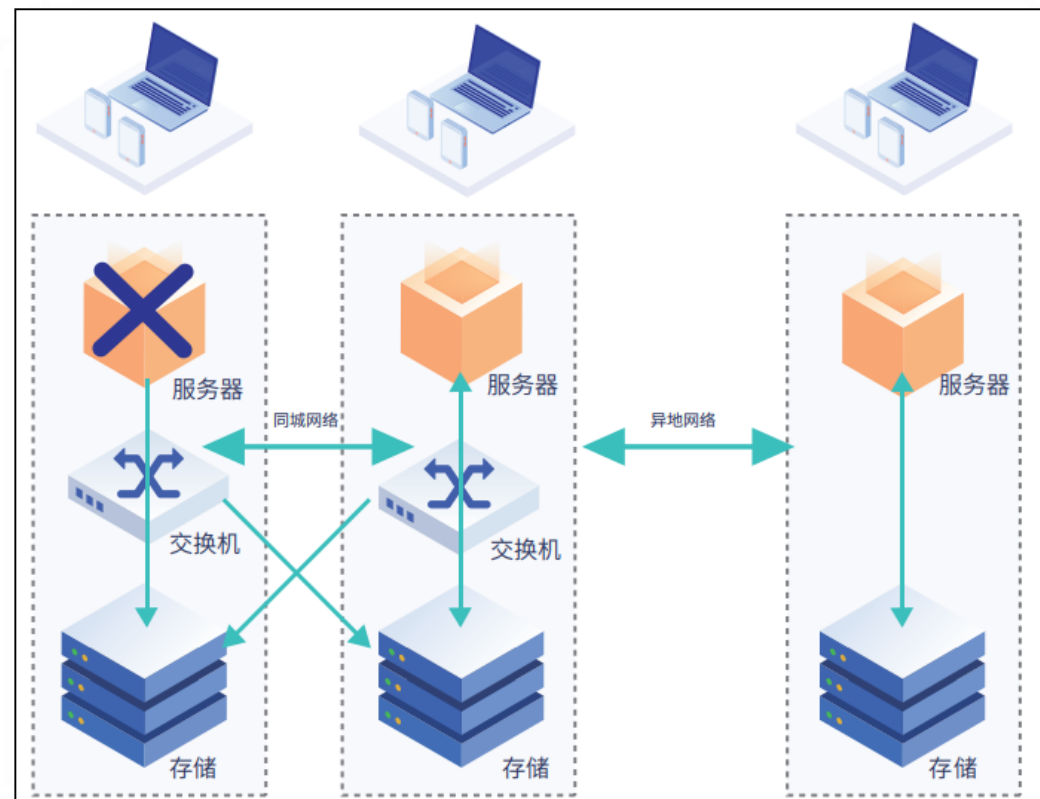


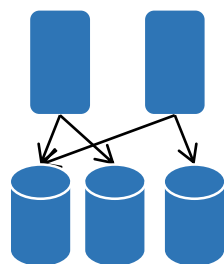
更大规模

■ 2节点增加至8个节点，甚至更多

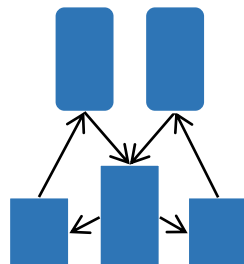
DSC Plus

■ 同城跨机房多活部署

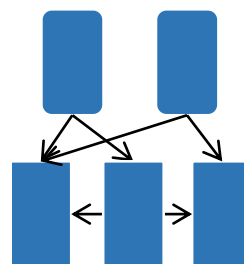




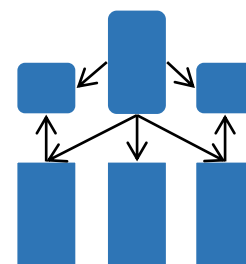
分库分表



集中写入



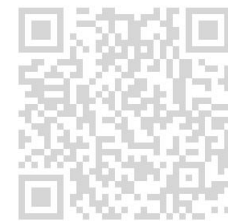
强存储分布式

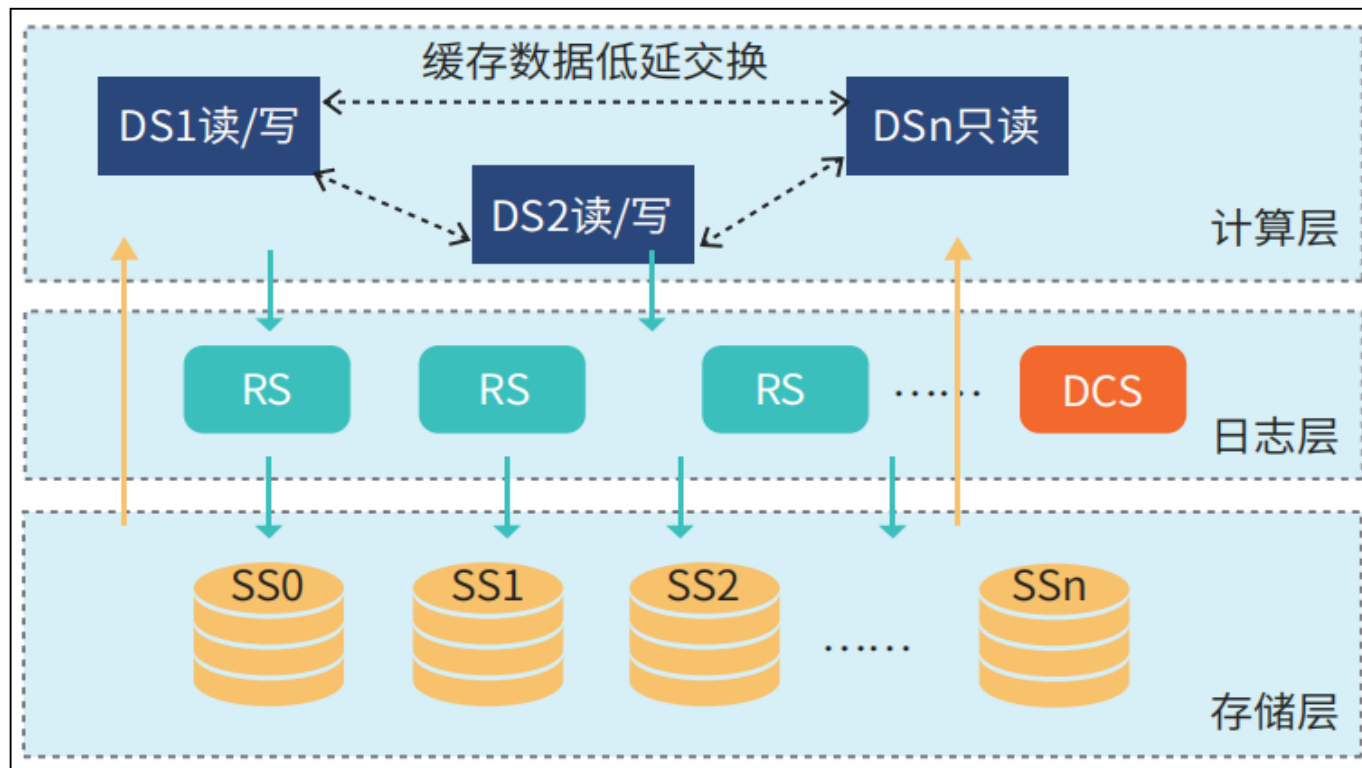


日志即数据

&

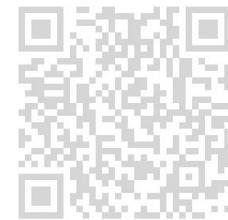
达梦8 透明分布式数据库 (TDD) 架构





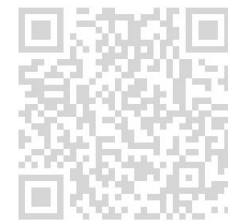
- 支持达梦DSC技术
- 多点写入
- 独立的日志服务
- 日志即数据
- 区为单位多副本容灾

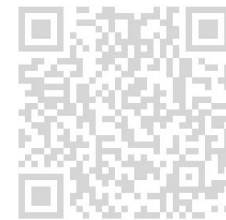
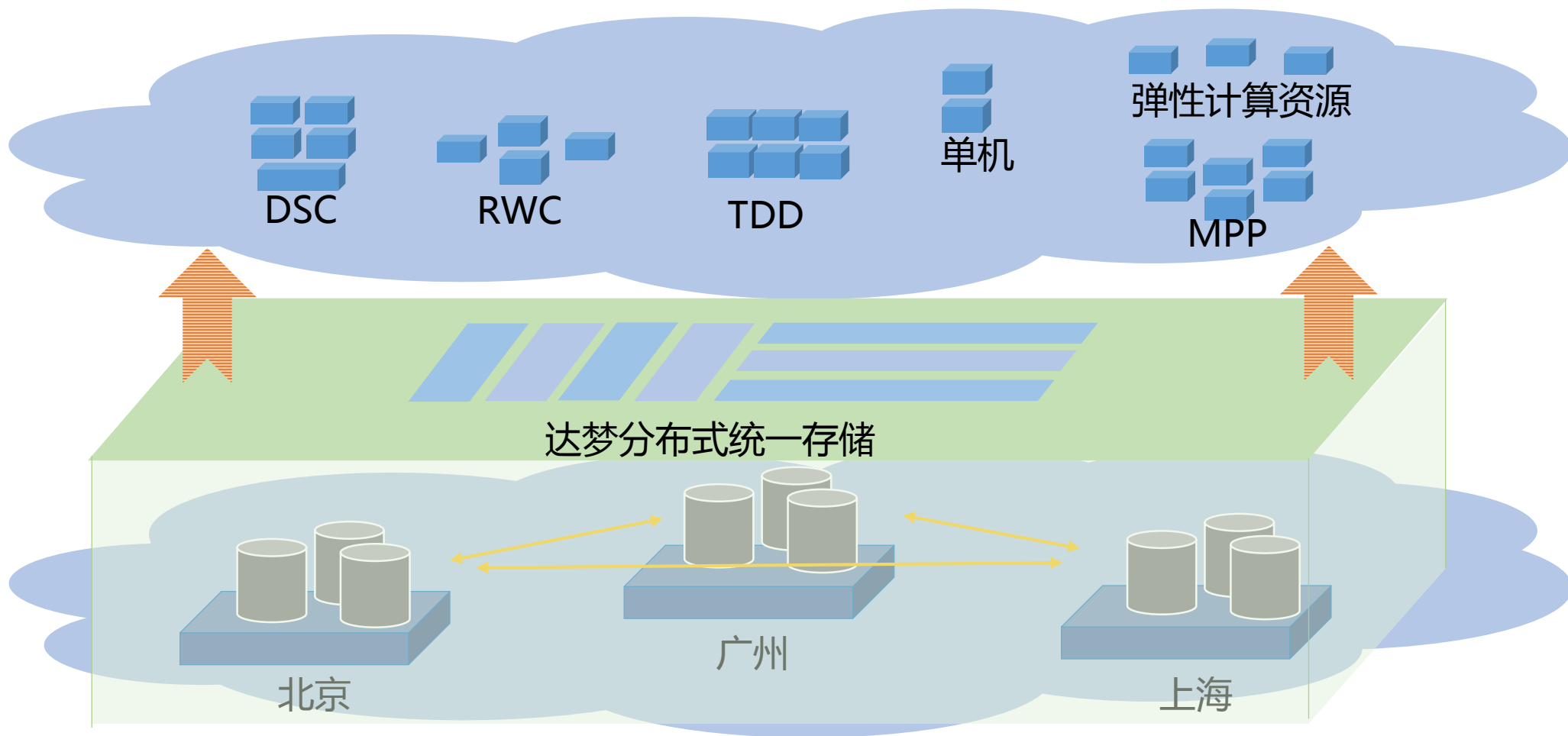
■ 计算与存储分离



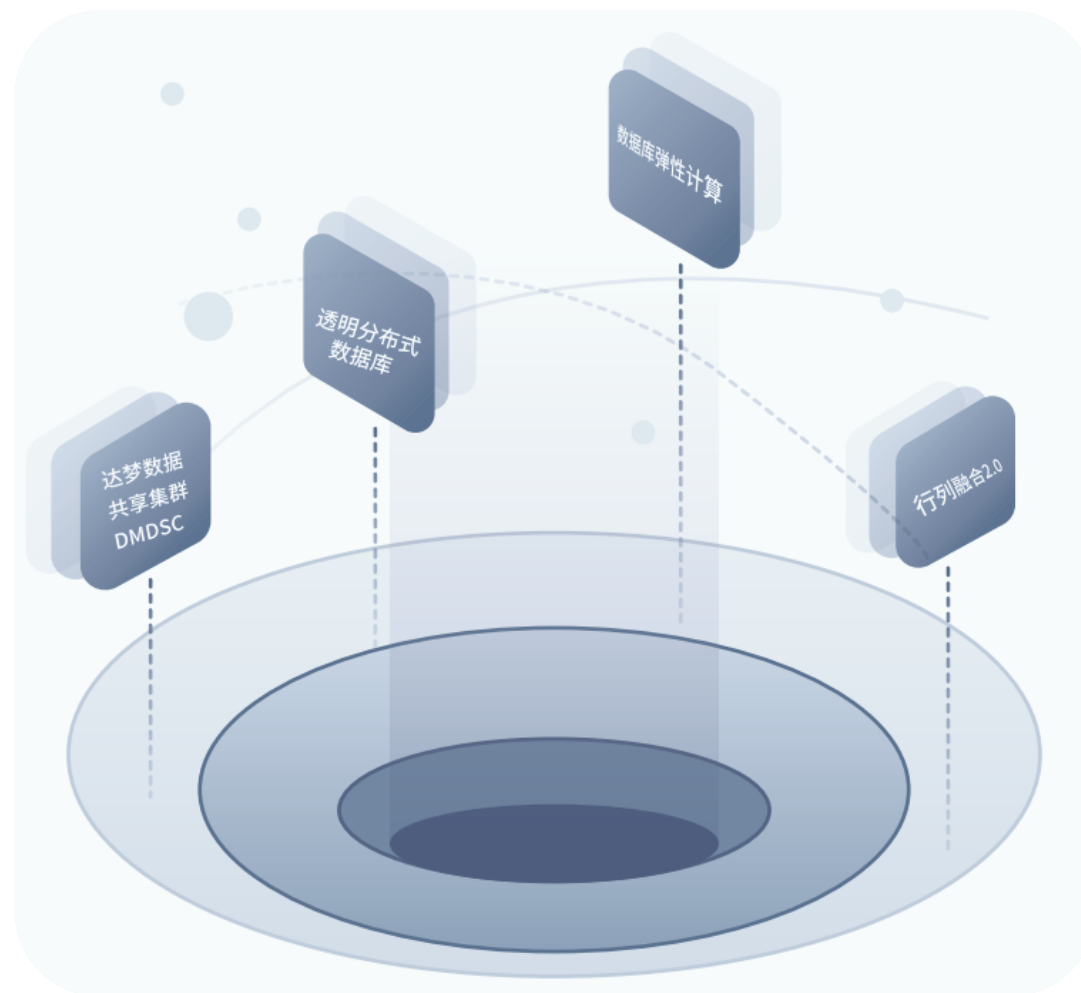
- 计算层、存储层都可横向扩展
- 计算层支持多点读写
- 完整的数据库特性
- 日志即数据，规避写放大问题
- 实现多副本高可用
- 存储层可支持MPP

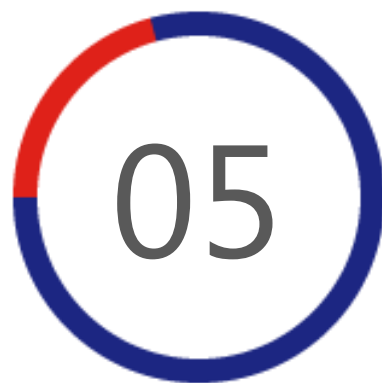
■ 鱼与熊掌兼得，像使用传统单机一样使用分布式



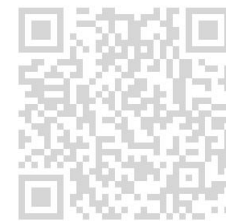


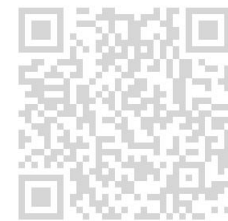
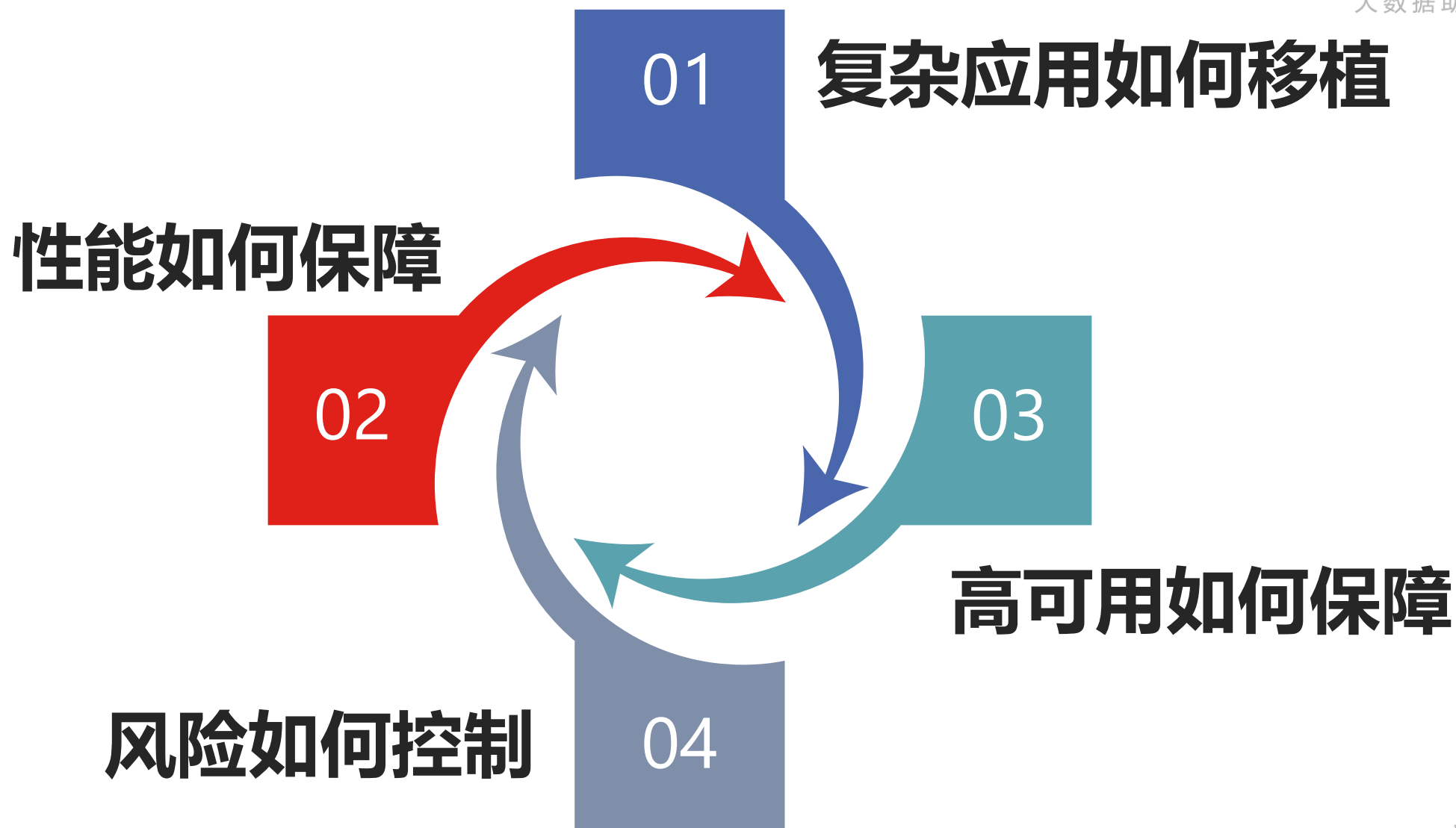
化繁为简
合而为一

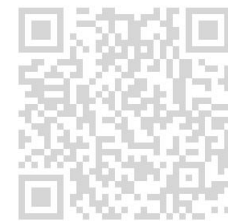




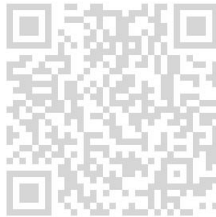
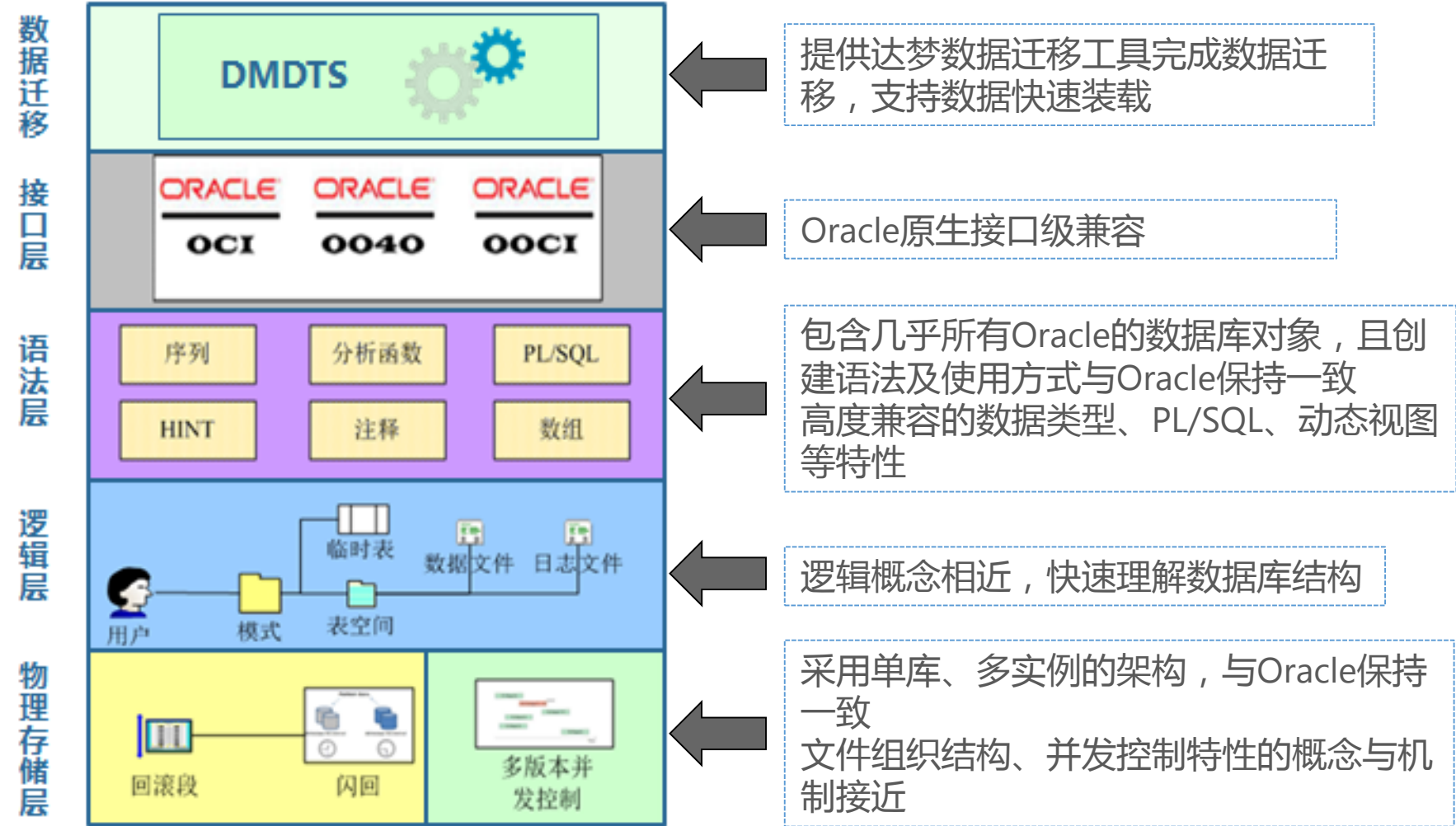
总结







从体系架构、并发机制、语法、接口、运维等方面与Oracle全面兼容，基于Oracle的系统可**轻松**移植。



采用**实时数据同步工具**保持国外主流数据库与达梦数据库的数据实时一致，可互为备份、交替运行，可实现**柔性切换**，提供**科学有序**的替换方案。

部署方式：

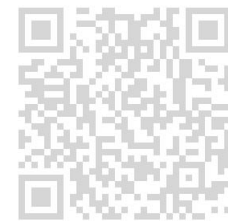
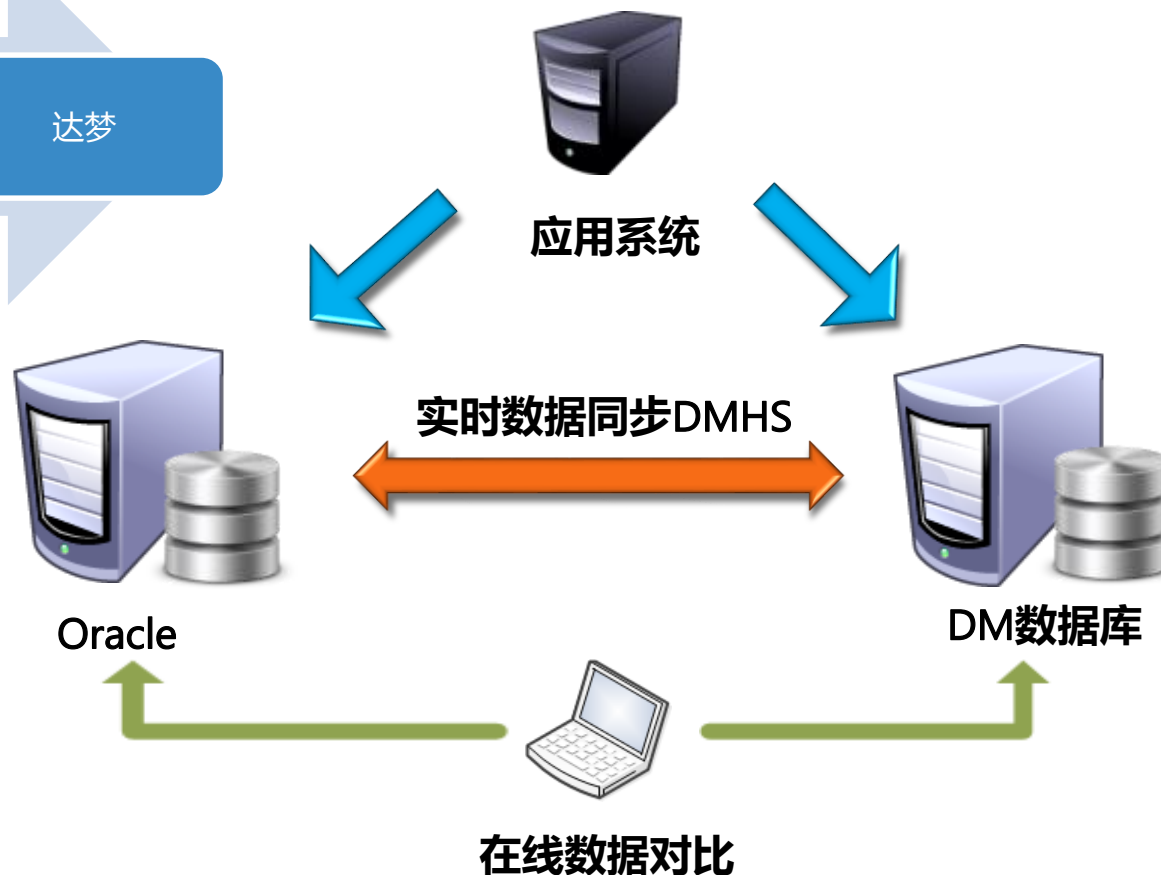


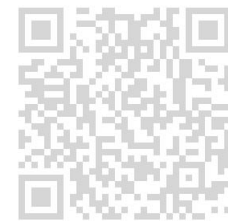
核心技术：

实时同步技术，同时支持国外主流系统和国产系统，并支持双向切换

达梦与ORACLE高度兼容，客户只需要维护**一套应用**

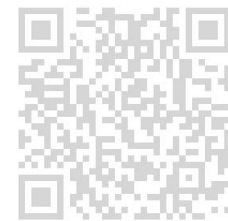
在线数据对比技术，不停止在线生产系统情况下，确保两边数据完全一致





无论是架构演进，还是产品化策略

达梦数据库的产品发展首先受用户需求的引导



脚踏实地、聚焦技术
面对用户永远谦逊!



Thanks



全新IT技术私域交流平台