
PostgreSQL中文社区
2021
PostgreSQL China Conference
主办：PostgreSQL 中文社区
第 11 届 PostgreSQL 中国技术大会
开源论道 × 数据驱动 × 共建数字化未来
TDSQL PostgreSQL版企业级分布式数据库技术创新实践
腾讯云高级工程师 谢灿扬

TDSQL-PG 简介

开源论道 × 数据驱动 × 共建数字化未来

TDSQL-PG(原TBase)简史

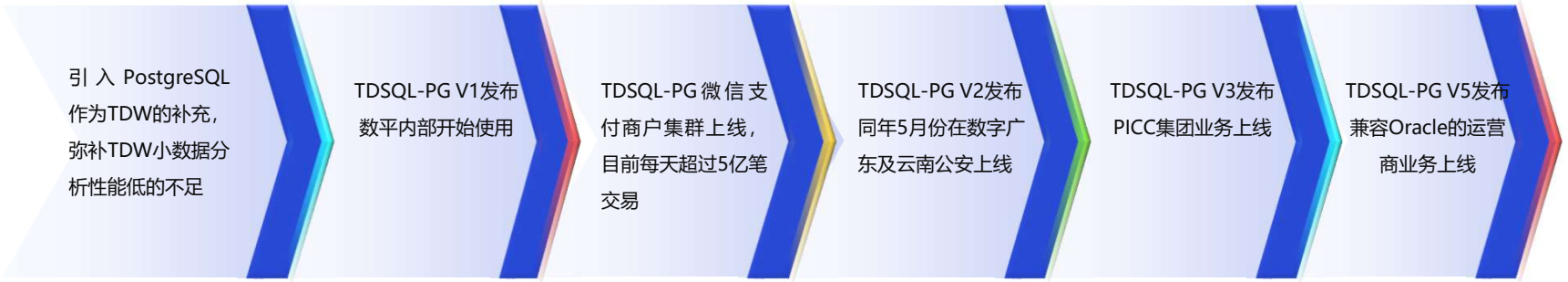
TDSQL-PG是基于PostgreSQL研发的分布式数据库：

V1：具备完整的分布式事务处理能力，具有良好SQL兼容性及在线扩展能力

V2：数据更安全，具备三权分立安全体系，内核独有支持透明数据脱敏

V3：支持OLAP在线分析业务处理，更完备的并行处理能力，提供**一站式整体解决方案**

V5：支持Oracle语法兼容，读写分离功能

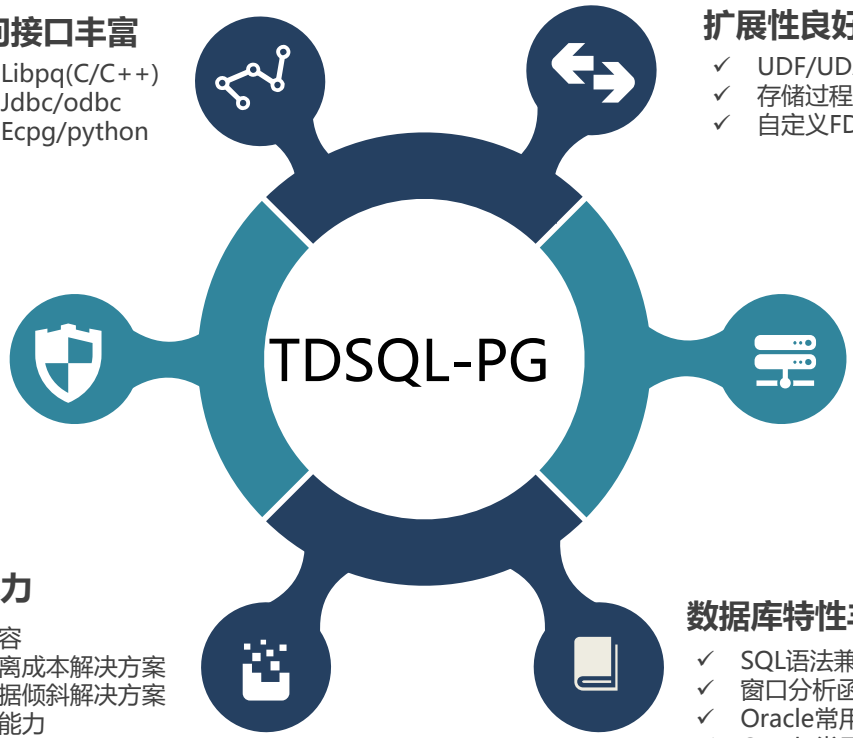


TDSQL-PG的定位

TDSQL-PG 是腾讯**自主研发**的新一代**分布式国产数据库**，其具备业界领先的**HTAP能力**，在提供大型数据仓库处理能力的同时还能完整支持事务。

- **无共享 MPP**
- **兼容 SQL2003**
- **完整分布式事务**
- **强悍数据分析能力**

TDSQL-PG整体能力



访问接口丰富

- ✓ Libpq(C/C++)
- ✓ Jdbc/odbc
- ✓ Ecpg/python

扩展性良好

- ✓ UDF/UDAF/UDWF
- ✓ 存储过程，自定义操作符
- ✓ 自定义FDW(访问外部数据源)

MLS安全保护能力

- ✓ 三权分立能力
- ✓ 强制安全规则
- ✓ 数据审计能力
- ✓ 数据透明脱敏
- ✓ 数据透明加密

高效HTAP能力

- ✓ 千万级TPS事务处理能力
- ✓ 全局sequence支持
- ✓ 支持全并行分布式计算架构，高效完成OLAP计算

强大数据治理能力

- ✓ 高效弹性在线扩容
- ✓ 长尾数据冷热分离成本解决方案
- ✓ 用户无感知的数据倾斜解决方案
- ✓ 两地三中心容灾能力

数据库特性丰富

- ✓ SQL语法兼容SQL2003标准
- ✓ 窗口分析函数
- ✓ Oracle常用函数支持
- ✓ Oracle常用语法支持

TDSQL-PG适用场景

数据量
交易数据量大于1T以上，或分析数据量大于5T以上

并发能力
并发连接数量达到2000以上，业务要求每秒峰值100万笔业务交易

在线水平扩展
替代业务原有需要分库分表的场景

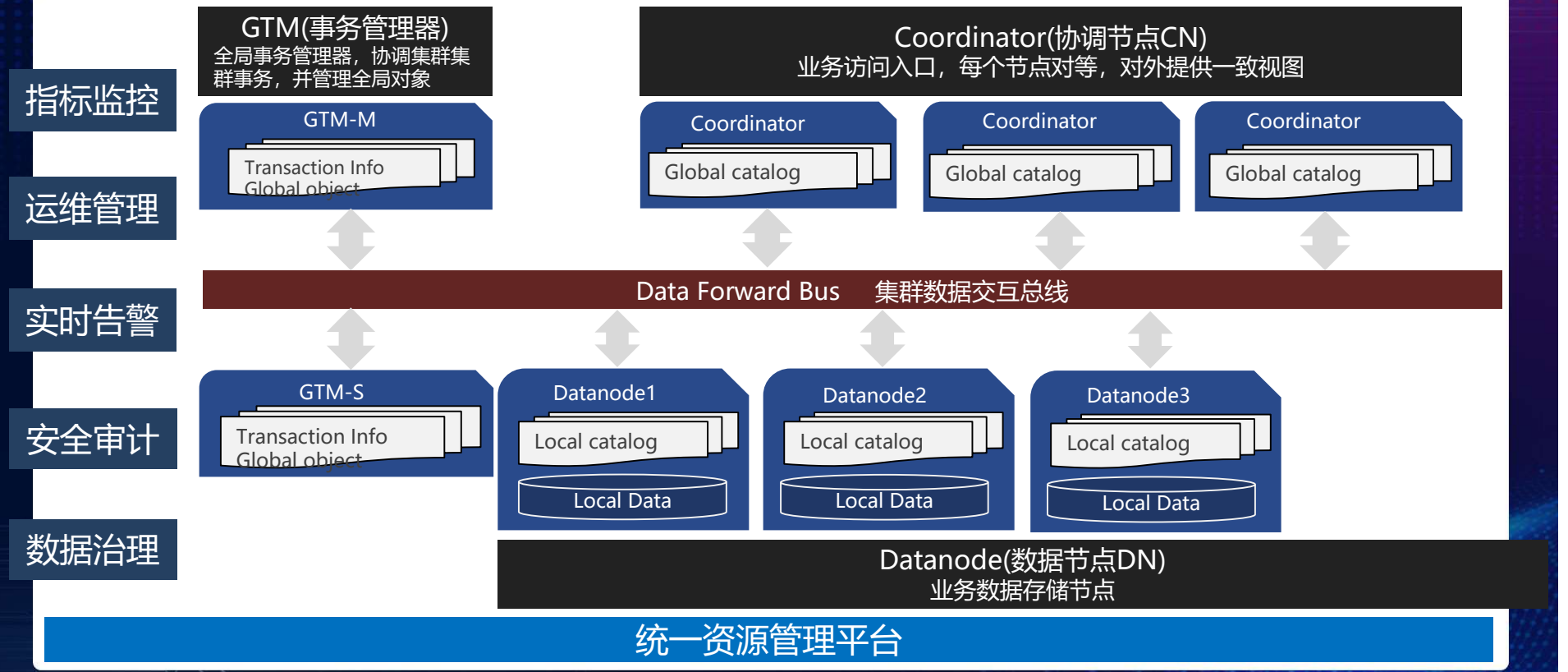
HTAP能力
具备高并发的OLTP处理能力的同时，兼顾相当量级的OLAP分析能力，支持一站式解决业务对数据库的诉求

分布式事务
将事务机制融入到数据库内，解决分库分表模式的痛点

业务场景

- HTAP业务
- 地理信息系统
- 实时高并发系统
- 数据库国产化

TDSQL-PG总体架构

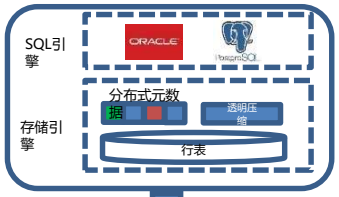


TDSQL-PG 能力介绍

开源论道 × 数据驱动 × 共建数字化未来

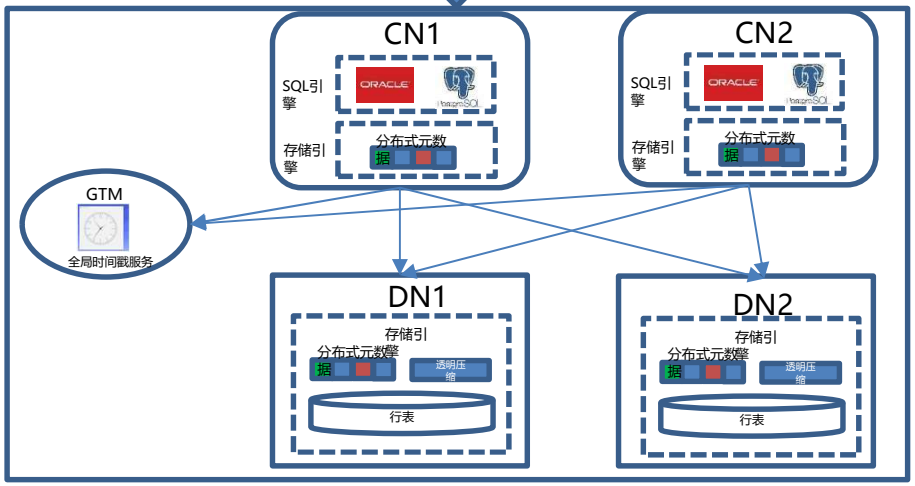
多引擎：集中式分布式一体化（HTAP）

集中式
- 高度兼容ORACLE语法
- 无分布式开销



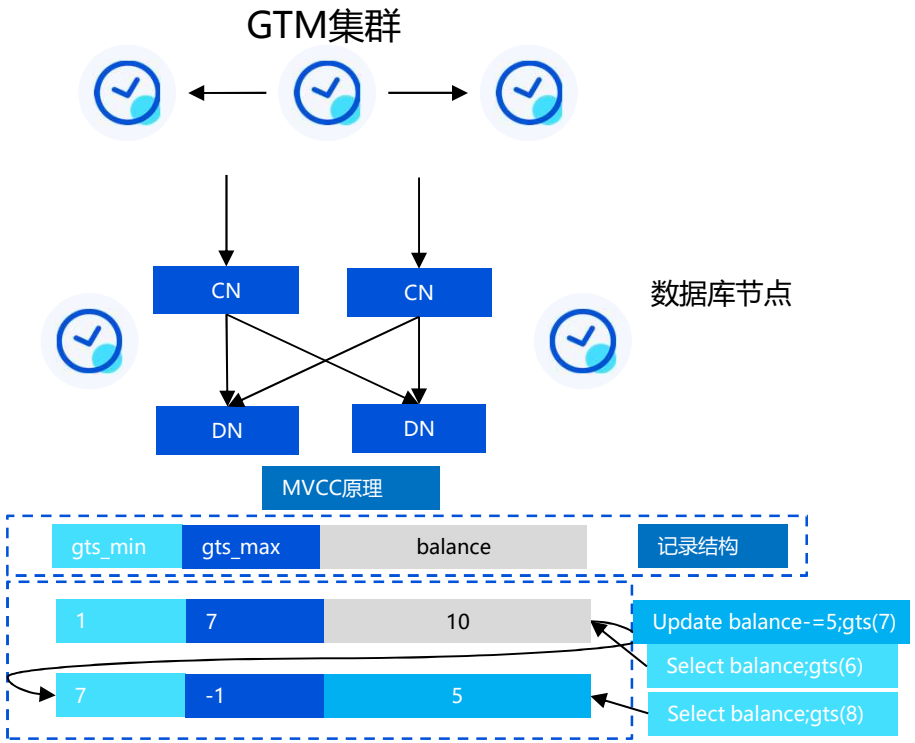
分布式
- 高度完整兼容ORACLE语法
- 全局一致

无缝扩展成分布式



核心能力	内容
分布式事务能力	分布式事务ACID能力，支持分布式一致性读（RR，RC两个隔离级别）
分布式核心能力	分布键更新，全局索引能力，高性能OLAP能力
ORACLE兼容能力（金融/运营商场景98%兼容性）	数据库对象支持，数据类型支持，特有语法支持，PL/SQL支持，系统函数支持，高级包支持，Package，自治事务，查询计划绑定，GBK，GB18030，UTF8

基于GTS的MVCC并发控制



GTS核心要点

01 MVCC能力

段页式存储的MVCC是整个并发控制的基础；同时约定：事务的gts_start > gts_min并且gts_max没有提交或者gts_start < gts_max才能看到对应的事务

02 GTS从哪里来

逻辑时钟从零开始内部单向递增且唯一，由GTM维护，定时和服务器硬件计数器对齐；硬件保证时钟源稳定度

03 GTM单点可靠性问题

多个GTM节点构成集群，主节点对外提供服务；主备之间通过日志同步时间戳状态，保证GTS核心服务可靠性

04 GTM单点瓶颈问题

根据测试推算，TS85服务器每秒能够处理1200万QPS，几乎能满足所有场景需要

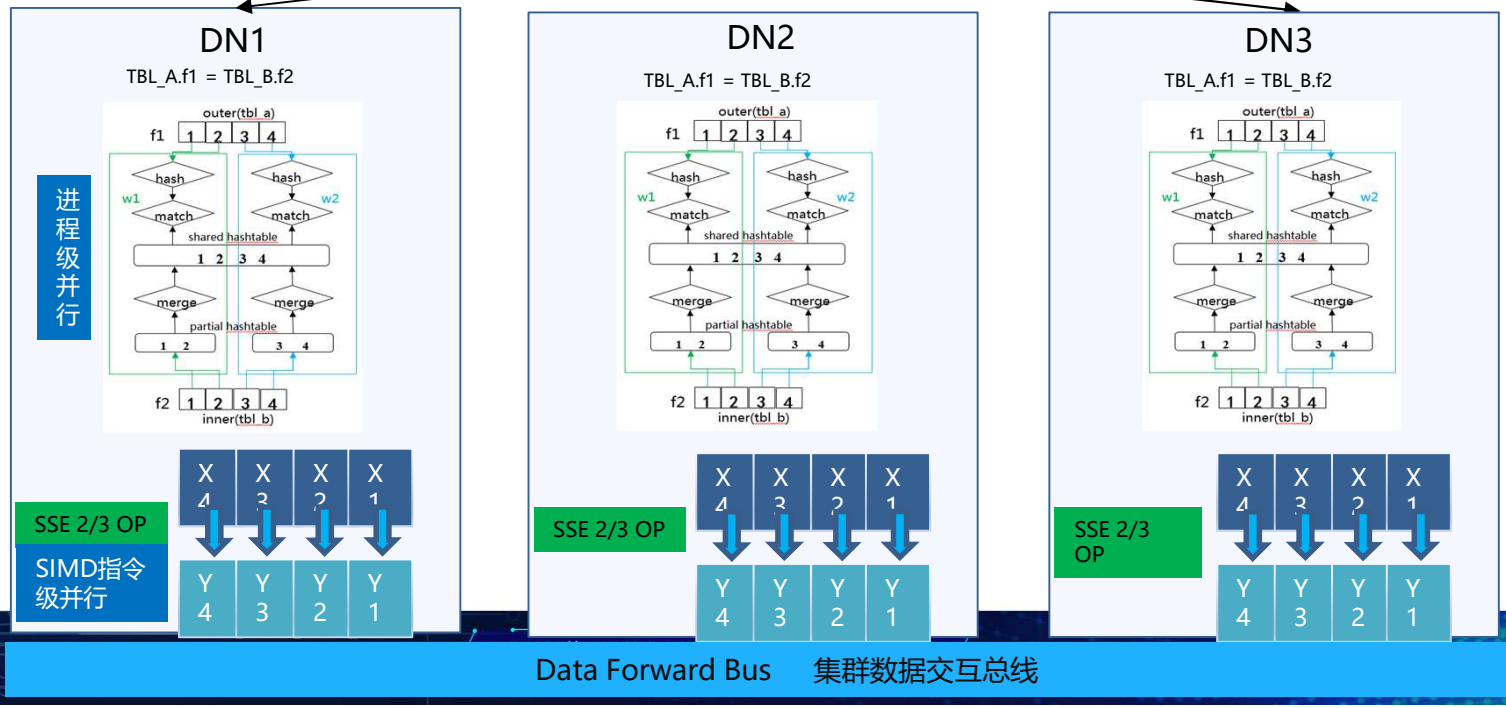
全并行计算能力

```
select * from tbl_a, tbl_b
where tbl_a.f1 = tbl_b.f2;
```

TBL_A(f1--分布列, f2)　　TBL_B(f1--分布列, f2)

节点级并行　CN　TBL_A.f1 = TBL_B.f2

节点级并行
节点内进程级并行
SIMD指令级并行



全局索引支持

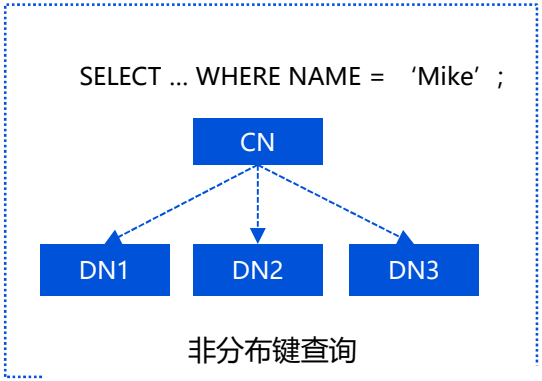
特性支持

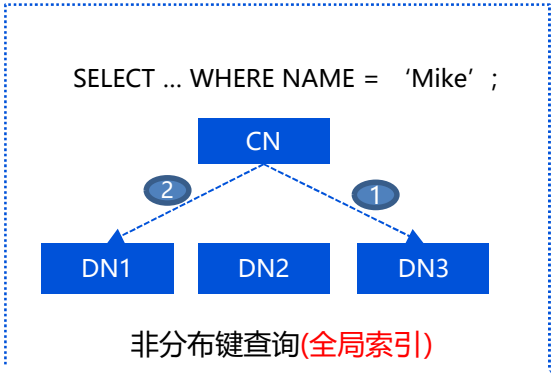
01 **支持非分布键约束**

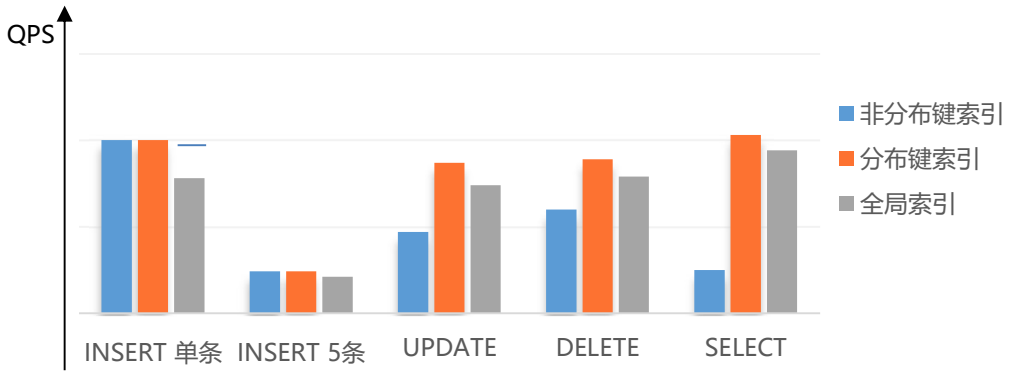
唯一索引，外键等约束之前都是需要在包含分布键，局限性比较大，全局索引可以在保证性能的同时放开约束，更贴近集中式系统。

02 **提升非分布键查询性能**

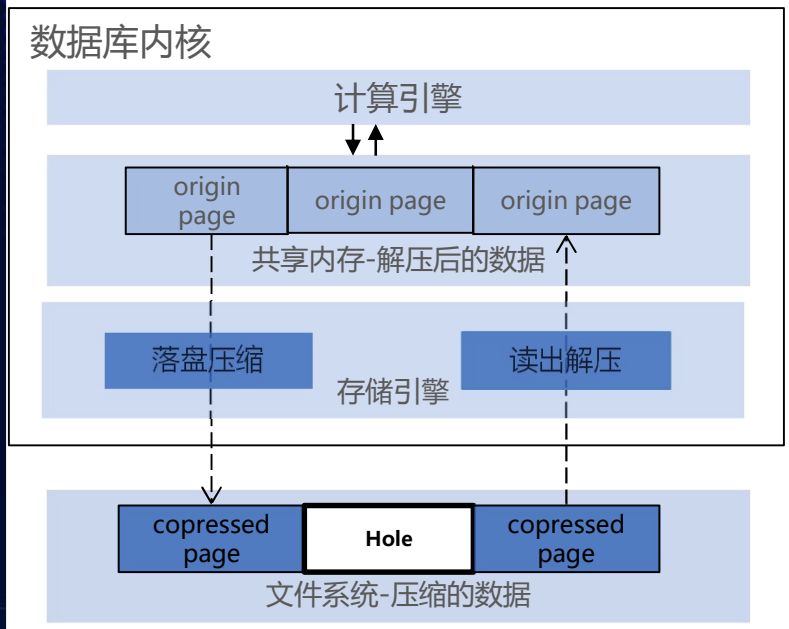
在非分布键查询的场景下性能大幅度提高，是原有的**4倍**，接近分布键查询性能。



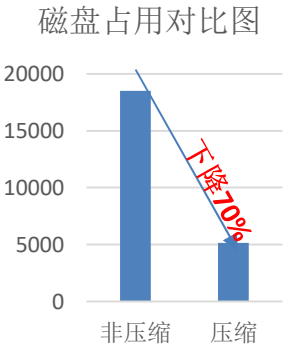


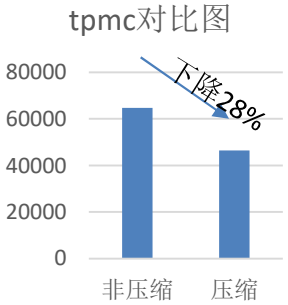


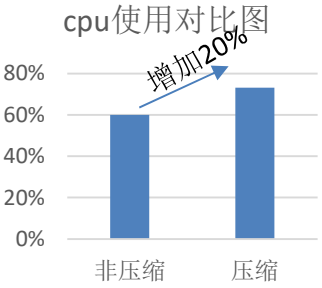
透明压缩



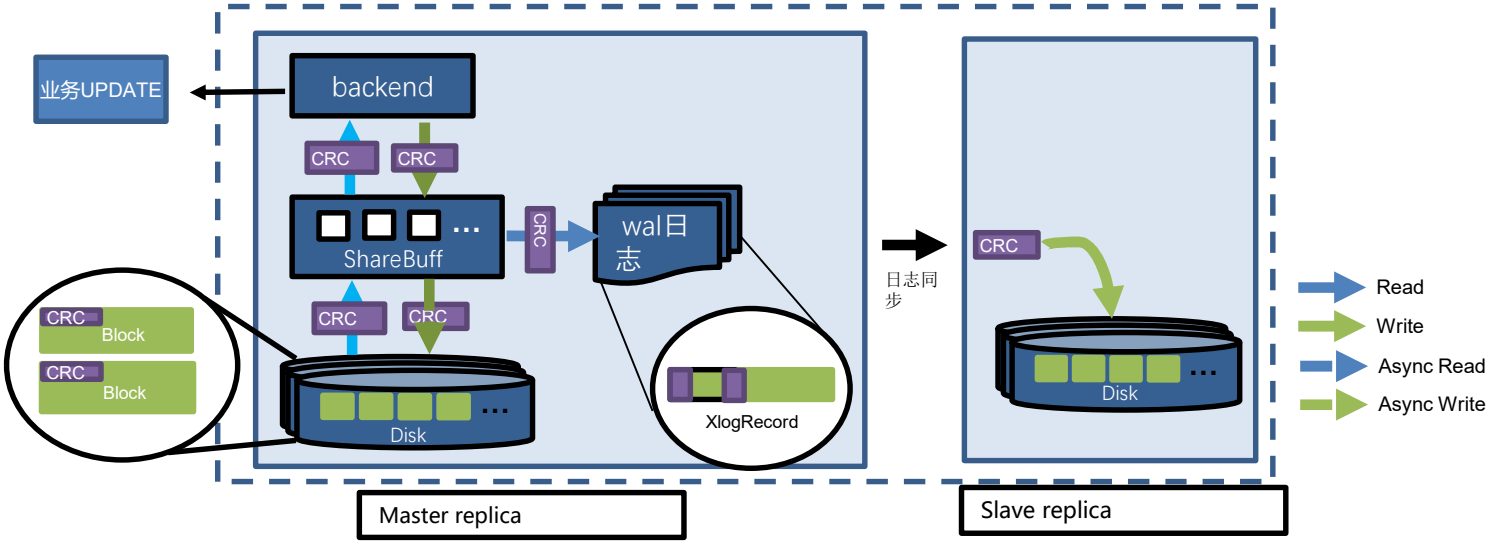
压缩结果--时间换空间：
1、数据文件磁盘占用率下降70%，压缩率达到30%
2、cpu使用率增加20%，tpcc性能下降28%







支持CRC循环校验，避免不可预知的数据损坏



全面检测机制

实时故障探测；
定期全量检测；
定期采样检测。

全路径保护

从数据写入缓存、写入WAL日志、数据刷盘、数据同步，再到数据读取，每个可能扩散环节都有阻断措施，避免故障扩散。

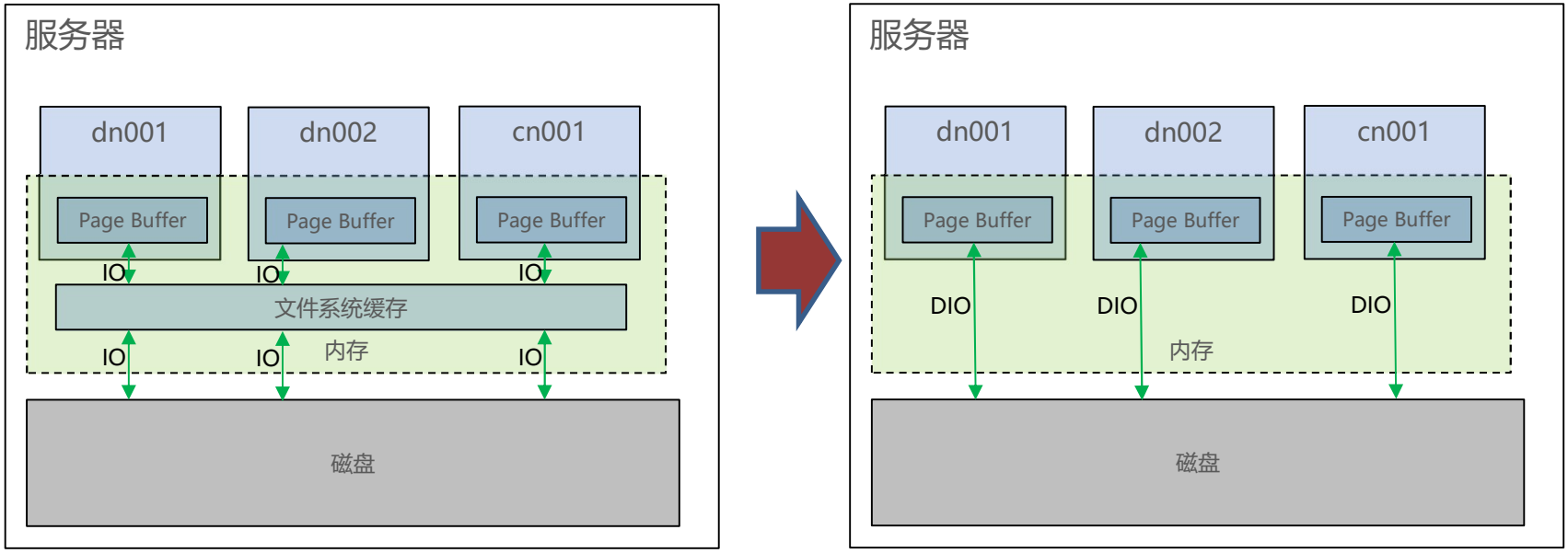
第一时间恢复业务

检测到故障后，自动主备切换，保证数据正确的情况第一时间恢复业务，在启动异步修复任务；
支持从副本或者冷备中恢复。

块级别增量修复

当数据错误/故障被发现后，系统自动发起以块为单位的增量修复，防止副本级修复带来的高时延和高资源消耗。

Direct IO



1）默认IO会使用文件系统缓存，因为DN/CN本身已经有数据页缓存（Page Buffer），所以两个缓存中可能有大量重复数据
2）虽然文件系统缓存占用的内存可以被应用抢占，但回收时可能需要刷脏页等，引起性能波动
3）支持Direct IO，可以满足部分时延敏感的长稳测试要求

易用性提升

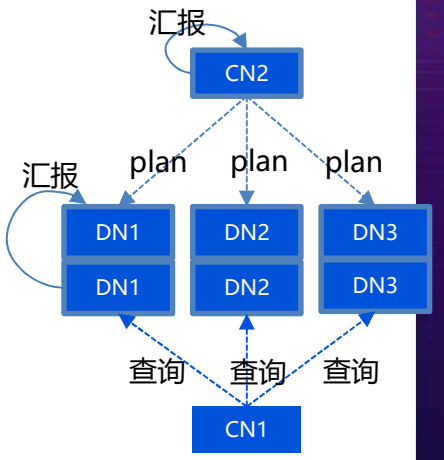
特性支持

01 全局事务视图

由CN发出的事务在多个DN下存在多个进程，同一个事务用独有的ID表示，用一个视图展示所有CN、DN上的进程状态。方便管理。提供给前端一个杀死一个会话下所有进程的接口。

02 内存占用视图

内存占用至关重要，用函数返回当前节点的内存总览以及各类内存的使用情况。




```
             contextname             | totalsize | freesize
-------------------------------------+-----------+----------
 TopMemoryContext                    |    536984 |        0
 pgstat TabStatusArray lookup hash table |    1584 |        0
 TopTransactionContext               |      7024 |        0
 CFuncHash                           |       688 |        0
 Record information cache            |     14976 |        0
 Node Handles Hash                   |      2608 |        0
 AuditContext                        |         0 |        0
 MessageContext                      |      1216 |        0
 Operator class cache                |       688 |        0
 smgr relation table                 |       688 |        0
```

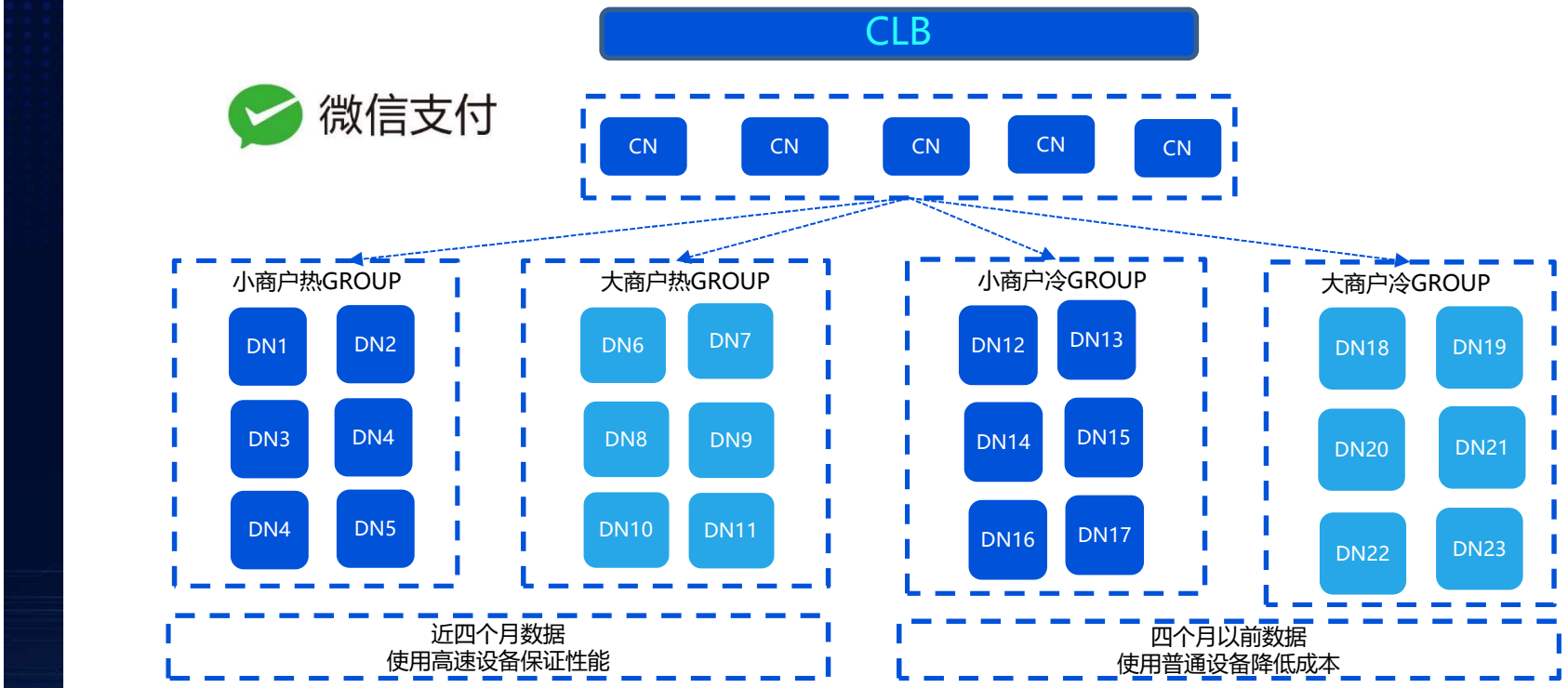
会话ID	节点名	进程ID	状态	运行时间
会话1	CN001	100	active	138
会话1	DN001	100	active	138
会话1	DN002	100	active	138
会话1	DN001	101	active	138
会话1	DN002	101	idle	-
会话2	CN002	100	active	55
会话2	DN001	102	active	55
会话3	CN001	101	idle	-

2021 PostgreSQL China Conference
第 11 届 PostgreSQL 中国技术大会

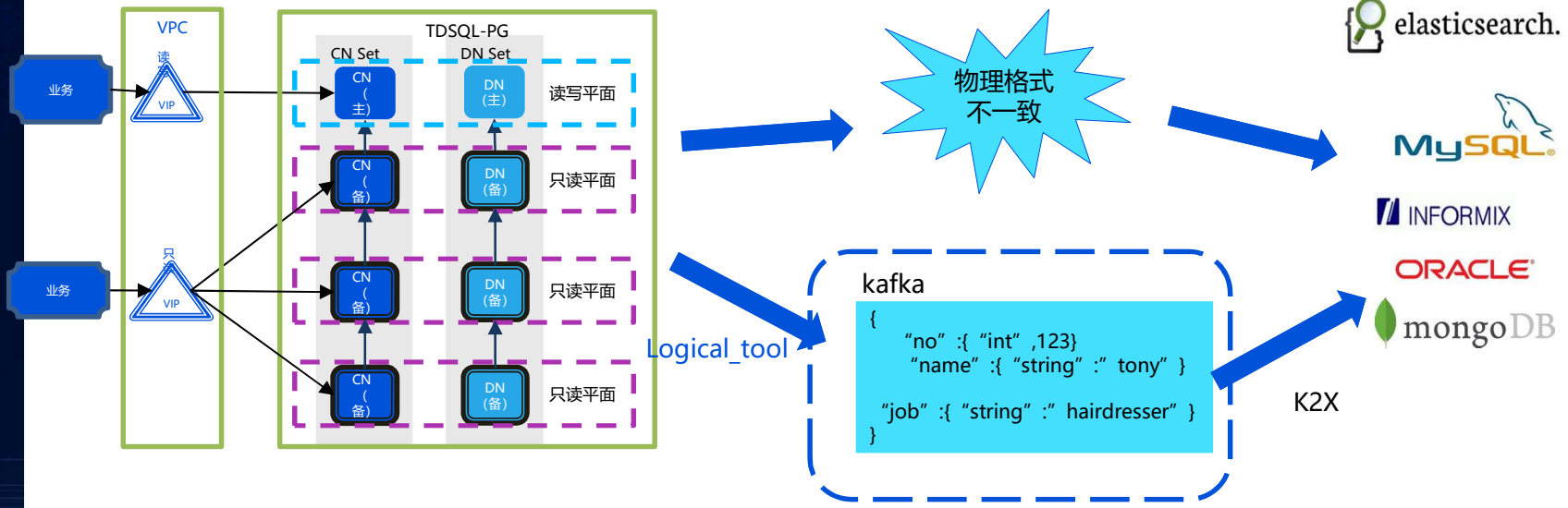
TDSQL-PG经典用户案例

开源论道 × 数据驱动 × 共建数字化未来

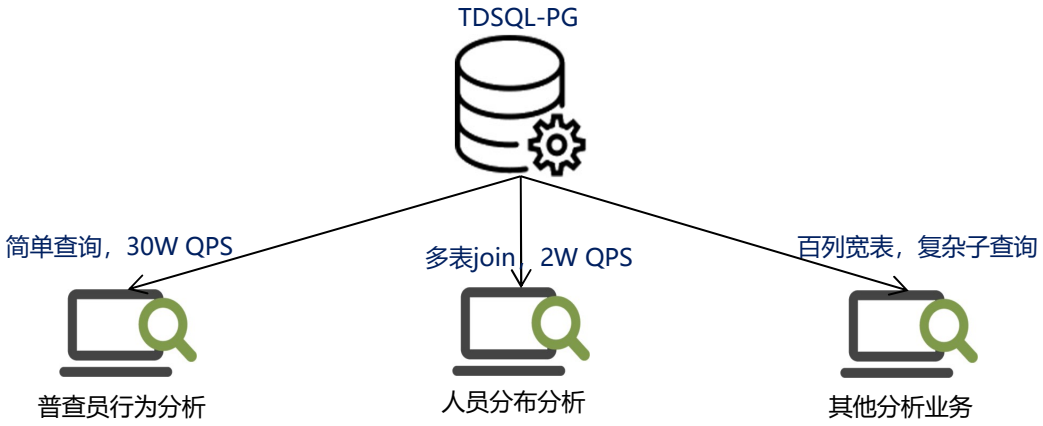
微信支付商户系统案例



外部某保险案例


VPC
读
VIP
业务
只
VIP
业务
TDSQL-PG
CN Set
DN Set
CN
(
主)
DN
(主)
读写平面
CN
(
备)
DN
(备)
只读平面
CN
(
备)
DN
(备)
只读平面
CN
(
备)
DN
(备)
只读平面
Logical_tool
物理格式
不一致
kafka
{
"no" :{ "int" ,123}
"name" :{ "string" :" tony" }
"job" :{ "string" :" hairdresser" }
}
K2X
elasticsearch.
MySQL
INFORMIX
ORACLE
mongoDB

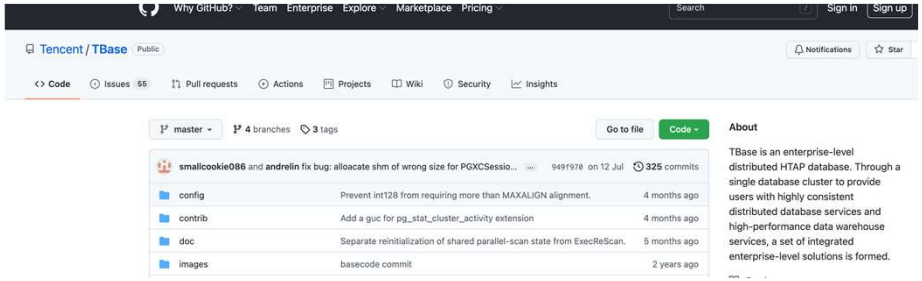
第七次全国人口普查系统案例



- 互联网用户：面向全国，1亿+用户，通过微信小程序自主填报；
- 高并发：700万+的普查员上班时间同时工作，通过企微小程序进行数据采集；
- 海量数据：15天内完成全国短表数据采集，数据库单表记录20亿+；
- 实时同步：异构库海量数据同步延迟达到分钟级
- 离线模式：在弱网的楼道或无网的山区等地都要能正常使用；
- 业务复杂：每天平台端的统计汇总任务十分繁重且多，一刻不能延缓

THANK YOU

https://github.com/Tencent/TBase



TBase开源群