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

阿里云RDS PostgreSQL内核技术分享

阿里云智能数据库高级技术专家 –王欢

阿里云蝉联Gartner全球数据库领导者象限

作为中国公司代表，过去 4 年，阿里云在 Gartner 全球数据库魔力象限评估*中连连跃升，现稳居全球第一阵营

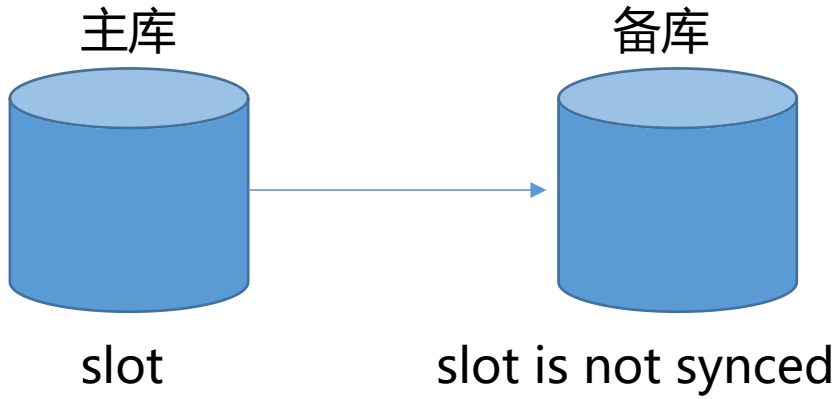
总纲

介绍RDS PostgreSQL 解决用户痛点问题的方案思考和技术细节

1. Logical Replication Slot Failover
2. 高并发场景审计日志（log_statement=all）优化
3. 智能索引推荐
4. 一键大版本升级
5. SGX全加密数据库

Logical Replication Slot Failover - 问题背景

- Replication Slot 不会被复制到备库
- HA后，Slot丢失，逻辑订阅中断
- 重新创建Logical Slot 导致部分增量丢失
  - Create Logical Slot 无法指定lsn



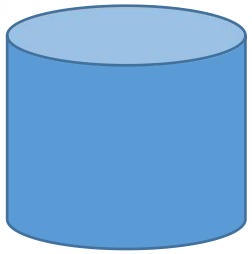
解决方案

- slot create/drop/update 通过复制协议同步到备库

- 难点 – 兼容性：
- WAL日志类型选用RM_XLOG_ID，新增1种 XLOG info
- RM_XLOG_ID xlog_redo() 对新增info处理保守
  - 不认识的info，直接丢弃，不会FATAL。
- 对用户自建replica、wal生态工具全兼容

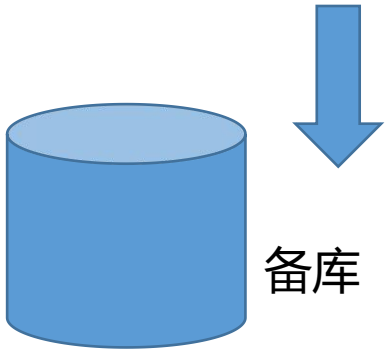
```
/* XLOG info values for XLOG rmgr */
#define XLOG_CHECKPOINT_SHUTDOWN        0x00
#define XLOG_CHECKPOINT_ONLINE          0x10
#define XLOG_NOOP                       0x20
#define XLOG_NEXTOID                    0x30
#define XLOG_SWITCH                     0x40
#define XLOG_BACKUP_END                 0x50
#define XLOG_PARAMETER_CHANGE           0x60
#define XLOG_RESTORE_POINT              0x70
#define XLOG_FPW_CHANGE                 0x80
#define XLOG_END_OF_RECOVERY            0x90
#define XLOG_FPI_FOR_HINT               0xA0
#define XLOG_FPI                        0xB0
#define XLOG_FPI_MULTI                  0xC0
```

改进效果 - slot信息从主库实时同步到备库



主库

```
postgres=# select * from pg_replication_slots;
   slot_name    |  plugin  | slot_type | datoid |     database     | temporary | active | active_pid | xmin | catalog_xmin |  restart_lsn  | confirmed_flush_lsn
----------------+----------+-----------+--------+------------------+-----------+--------+------------+------+--------------+---------------+---------------------
 debezium_prod | pgoutput | logical   |  74342 | prod_data...     | f         | t      |     107204 |      |   3571451039 | 2801/E5F36F60 | 2801/E7D51068
(1 row)
```



备库

```
postgres=# select * from pg_replication_slots;
   slot_name    |  plugin  | slot_type | datoid |     database     | temporary | active | active_pid | xmin | catalog_xmin |  restart_lsn  | confirmed_flush_lsn
----------------+----------+-----------+--------+------------------+-----------+--------+------------+------+--------------+---------------+---------------------
 debezium_prod | pgoutput | logical   |  74342 | prod_data...     | f         | f      |            |      |   3571455599 | 2801/E7FB88E8 | 2801/E98FC498
(1 row)
```

总纲

介绍RDS PostgreSQL 解决用户痛点问题的方案思考和技术细节

1. Logical Replication Slot Failover
2. **高并发场景审计日志（log_statement=all）优化**
3. 智能索引推荐
4. 一键大版本升级
5. SGX全加密数据库

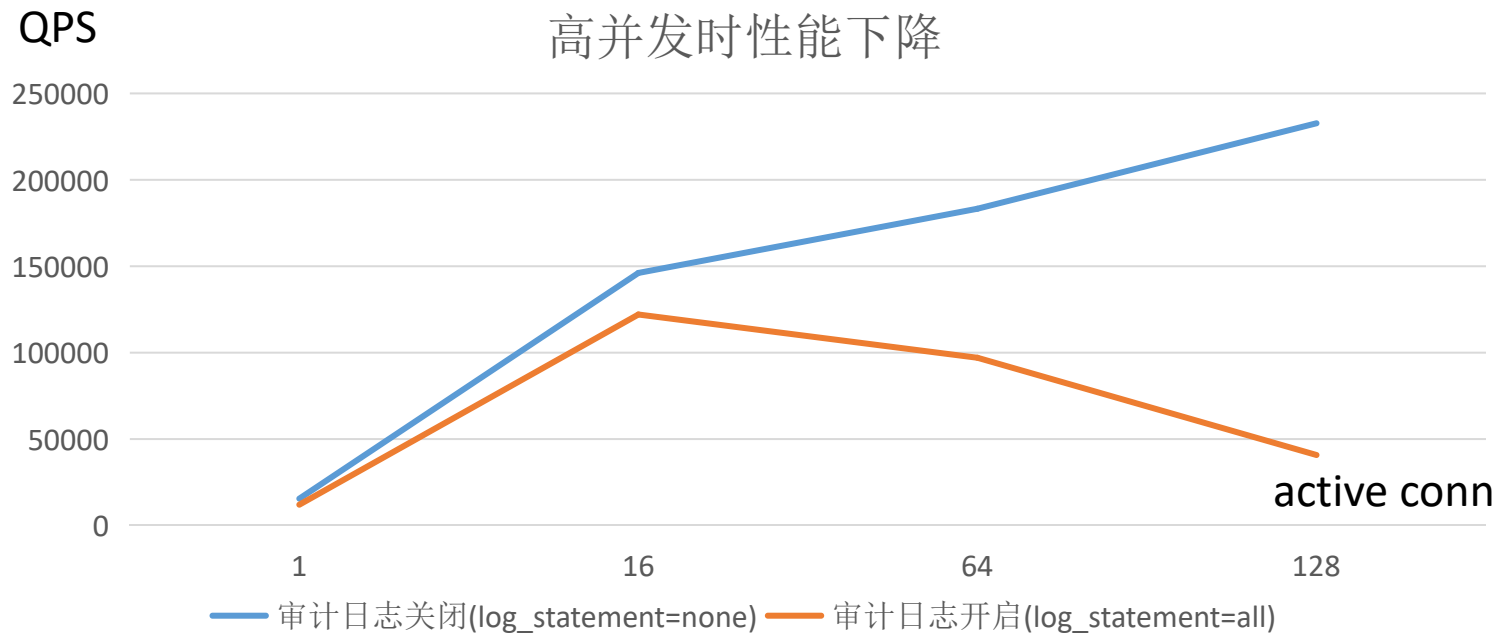
高并发场景审计日志（log_statement=all）优化

背景：

- log_statement=all 所有执行过的SQL写入日志
- 高安全要求业务：对所有SQL进行安全审计是强需求

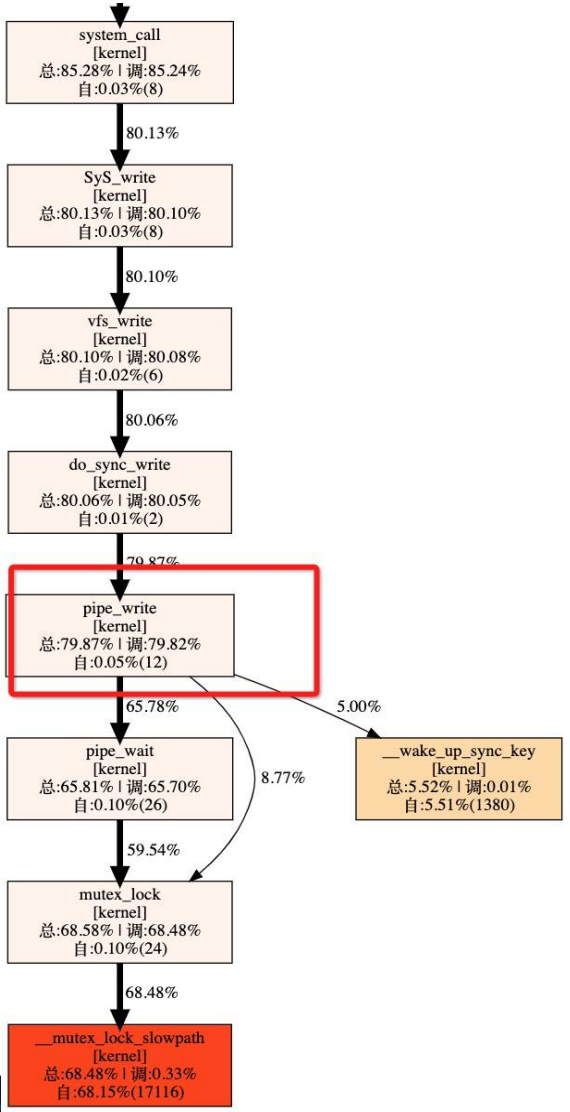
问题1 – 高并发时性能下降

- 16C64GB实例
- TPC-B只读 scale=100
- 审计日志开启，在高并发时性能严重下降



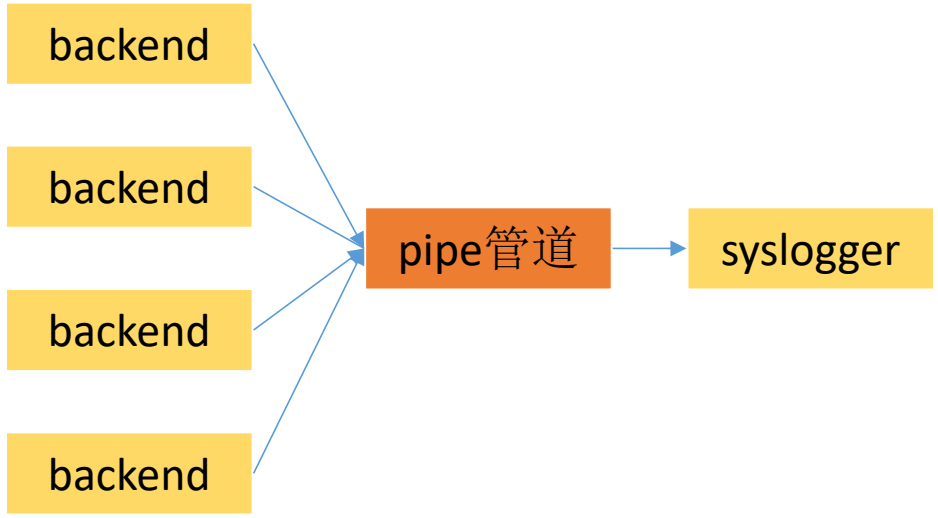
问题2 - 高并发时SysCpu高

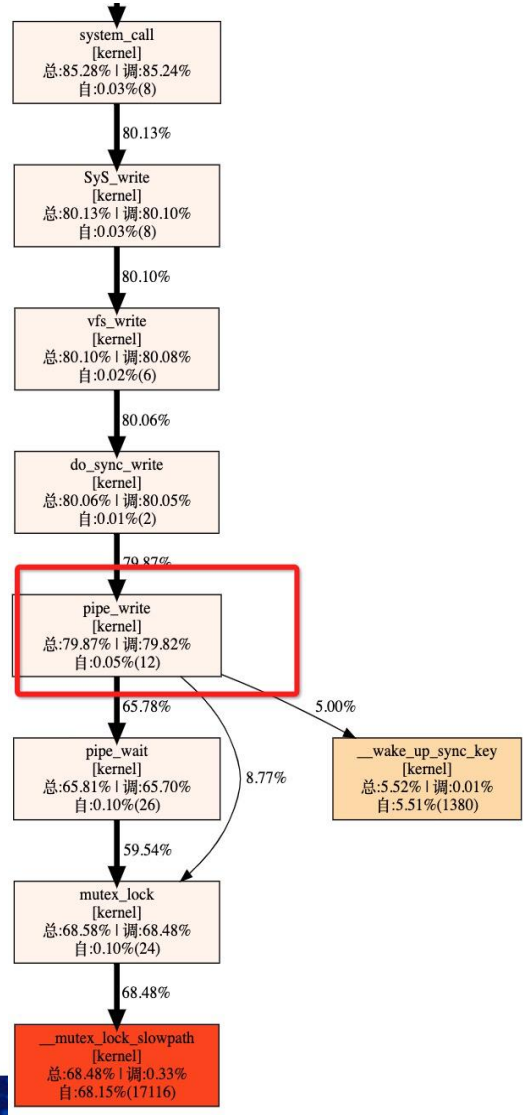
- SysCpu高：高并发时，SysCpu 飙升到实例Cpu阈值的80+%。
- 实例雪崩风险。



问题原因分析

- PG日志模型：backend进程向同一个pipe管道写入日志
- SysCpu高：高并发pipe_write() 带来的严重锁争抢

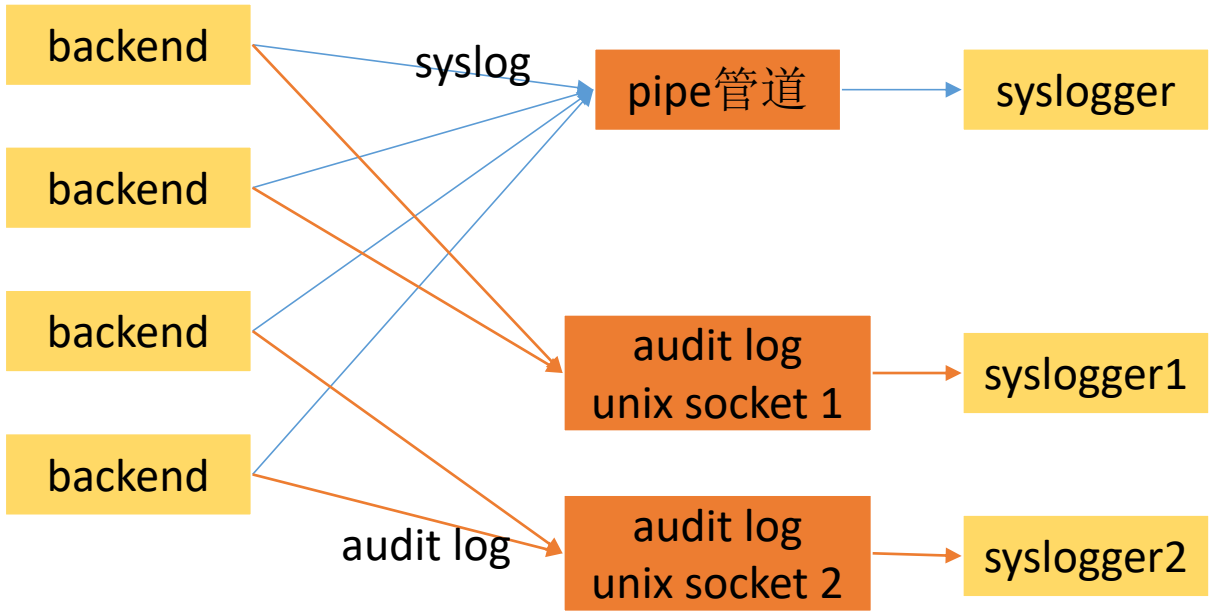




优化 – 高并发pipe_write() 带来的严重锁争抢

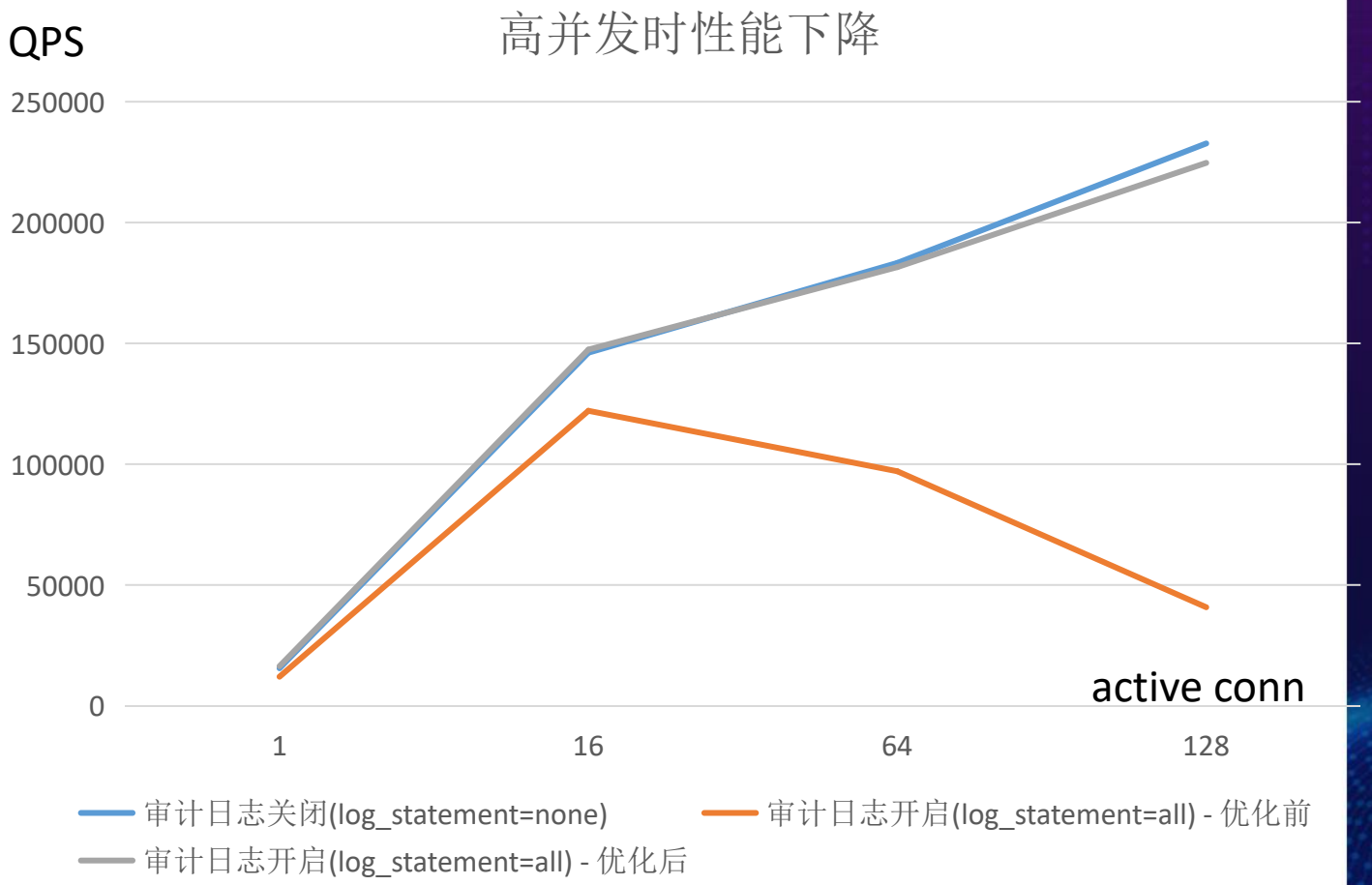
方案：auditlog 从pipe管道旁路出去

- 拆分单个日志文件到2个日志文件：syslog、auditlog；
- 原有通过pipe传输syslog保持不变，新增通过 socketpair 来传输 auditlog；
- 单个syslogger进程改为多个syslogger进程；
- 加大缓冲区，采用全缓冲+定时的方式进行日志下刷，减少磁盘IO次数；



优化效果

- 审计日志优化后，性能与审计日志关闭时性能持平
- 同时消除了SysCpu高问题



总纲

介绍RDS PostgreSQL 解决用户痛点问题的方案思考和技术细节

1. Logical Replication Slot Failover
2. 高并发场景审计日志（log_statement=all）优化
3. **智能索引推荐**
4. 一键大版本升级
5. SGX全加密数据库

PostgreSQL 索引推荐现状

- 社区 PostgreSQL 没有集成索引推荐能力
- 开源插件：索引推荐 **pg_idx_advisor**，作者cohenjo，2014年废弃
- 开源插件：索引推荐 **pg_adviser**，作者gurjeet，2010年废弃
- 开源插件：虚拟索引 **hypopg** 如火如荼，与索引推荐还有距离
- **EDB Postgres Advanced Server**： **Index Advisor**，比较成熟，闭源

RDS PostgreSQL 智能索引推荐 - 方案

- PostgreSQL 提供的基础设施
  - 基于代价的优化器
  - What-If能力，优化器具有假设某些索引存在，并估算出SQL执行代价的能力。
  - 扩展能力：planner_hook

- 方案：通过分析SQL，枚举可能的索引组合，并通过优化器What-If的能力，选出其中收益最高的索引组合推荐给用户

RDS PostgreSQL智能索引推荐 - 步骤分解

分析 Indexable Column

- 分析出SQL中哪些列可以利用索引，例如:
- Where条件中的 =, >, <, between, in等列
- Order By的排序列
- Group By的聚合列，MIN， MAX函数列
- Join的Condition等值条件列

构建 Candidate Index

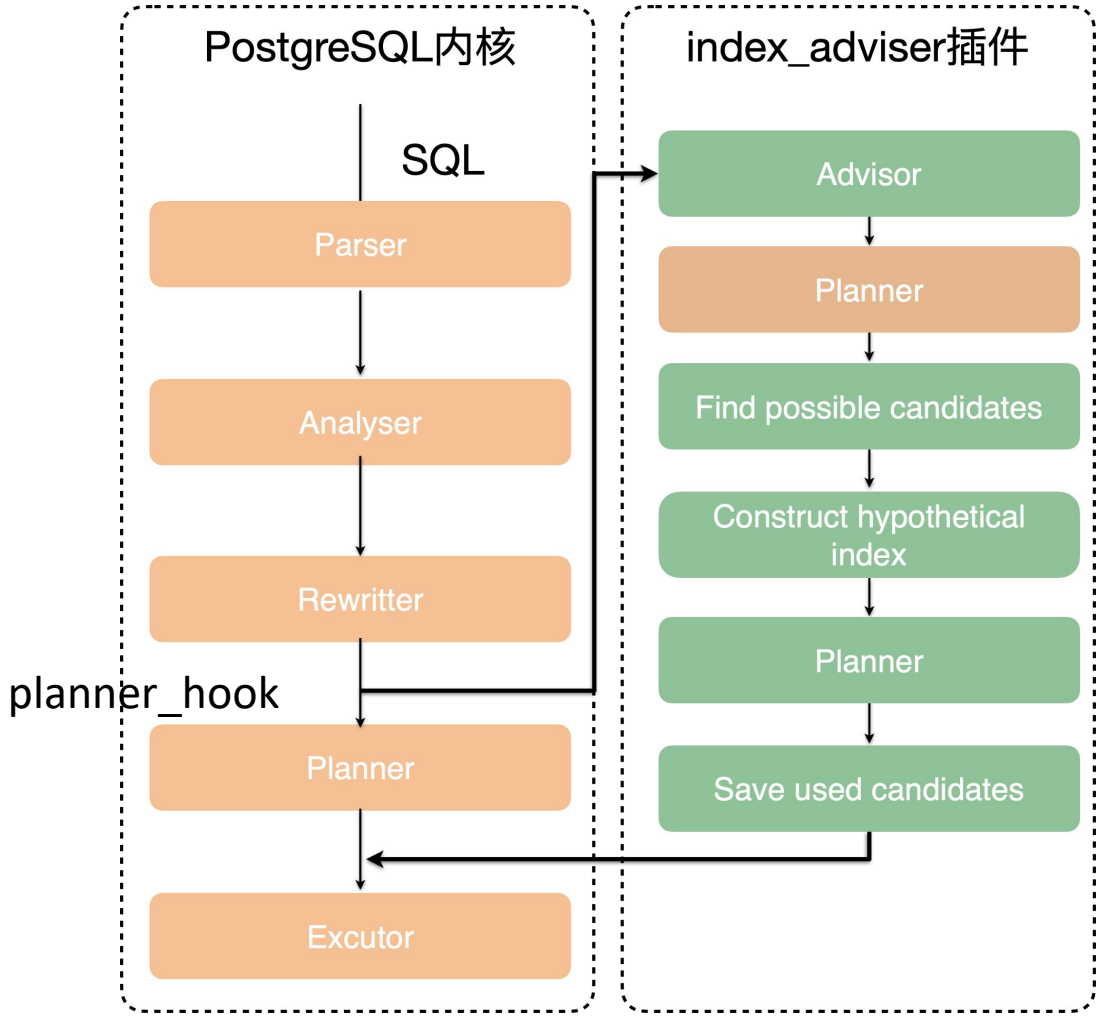
- 从Indexable Column中构建出所有可能的**Candidate Index**
- **Candidate Index**包含组合索引，会根据一些规则裁剪掉部分组合索引

选择最优，优化器What-If能力

- 优化器通过What-If能力将index_adviser中枚举到的**Candidate Index**逐一评估并获取SQL的执行代价。最终选择出使得SQL执行代价最低的**Candidate Index**

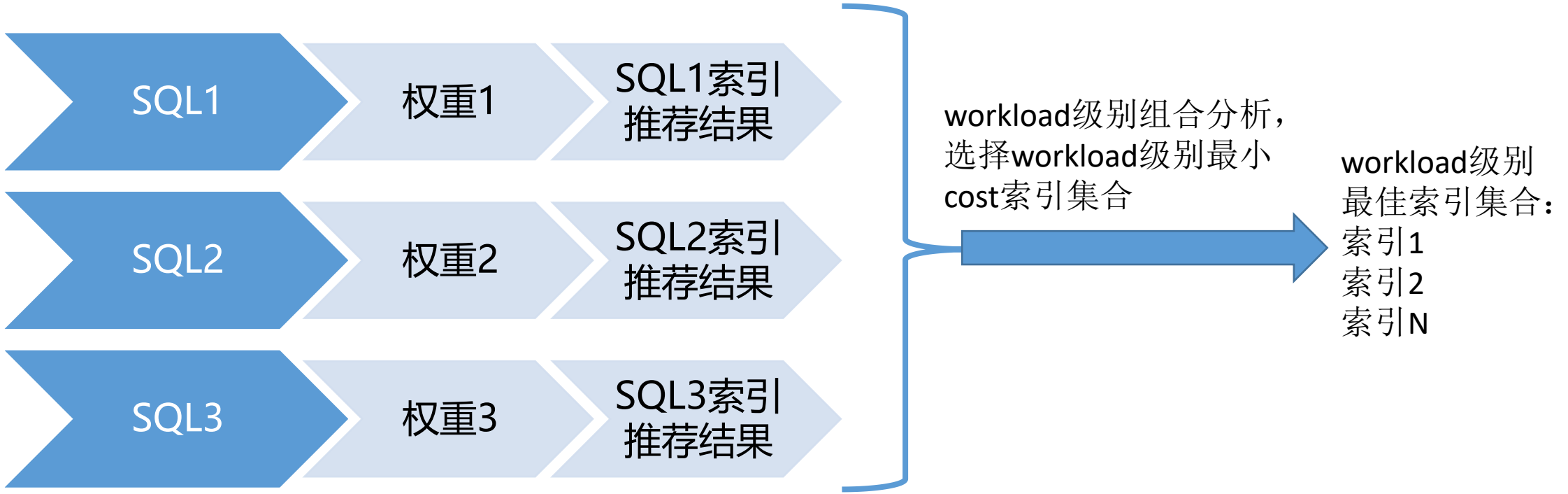
RDS PostgreSQL智能索引推荐 - 具体实现

- 插件化
- planner_hook



RDS PostgreSQL智能索引推荐 - workload负载

- 针对负载SQL集合，推荐全局级别的最佳索引集合



RDS PostgreSQL智能索引推荐 - 业界领先

索引推荐能力	阿里云 RDS PostgreSQL	EDB Postgres Advanced Server	国内/国际友商 PostgreSQL
where条件/join	✓	✓	✗
group by/order by组合索引	✓	✗	✗
join组合索引	✓	✗	✗
子查询组合索引	✓	✗	✗
最左匹配	✓	✗	✗
表达式组合索引	✓	✗	✗

RDS PostgreSQL智能索引推荐 - 产品化

- 2022年1月中旬，发布上线。
- 当前只支持btree索引， hash、brin、gin、gist等索引还在研发中。

总纲

介绍RDS PostgreSQL 解决用户痛点问题的方案思考和技术细节

1. Logical Replication Slot Failover
2. 高并发场景审计日志（log_statement=all）优化
3. 智能索引推荐
4. **一键大版本升级**
5. SGX全加密数据库

大版本升级 - 设计原则

PG9.4 ~ PG14，每个大版本都提升巨大。

可验证、可回滚

- 版本回滚：大版本回滚
- DNS地址：连接串回滚
- 可验证: 高版本可验证能力

场景全覆盖

- DDL限制
- 表结构限制
- 数据类型限制
- 版本全系覆盖

一键升级产品化

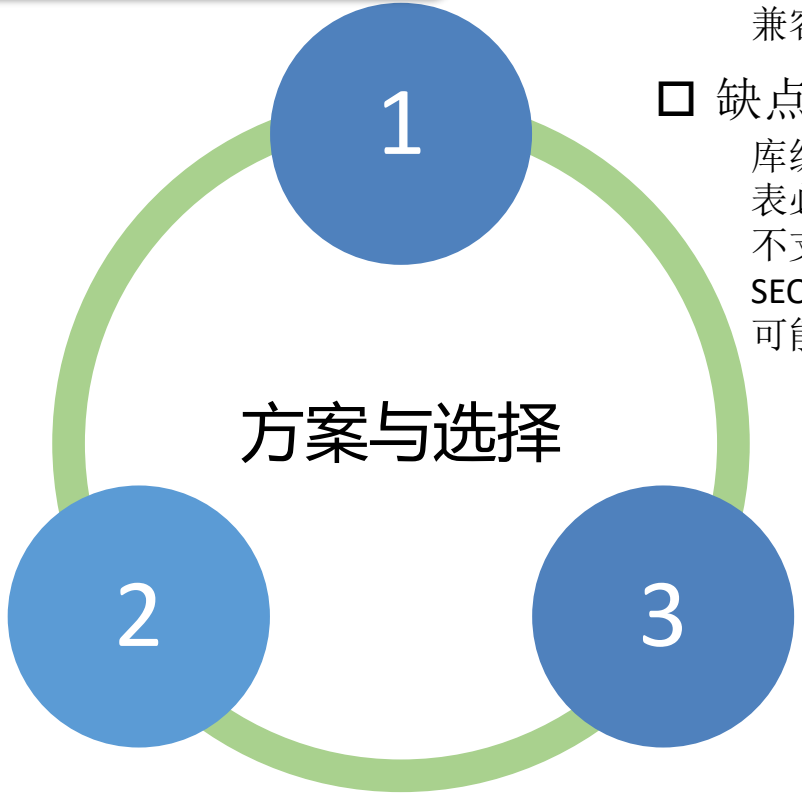
- 拒绝升级手册
- 一键升级：一键产品化能力
- 插件兼容性适配

应用不停服零宕机

- 升级过程应用不停服
- 升级过程速度快
- 连接地址平滑割接

大版本升级 - 方案选择

最终方案：RDS PG最终选择限制少、兼容性好、效率高、平滑割接的pg_upgrade方案。



1. 逻辑复制

- 优点：
  - 兼容性好、平滑割接
- 缺点：
  - 库级别的发布、订阅
  - 表必须有PK / UK
  - 不支持DDL、大对象
  - SEQUENCE 序列
  - 可能导致WAL日志堆积

2. pg_upgrade

- 优点：
  - 不拷贝数据, 仅元数据升级
  - 效率高, 2TB数据,升级 < 10s
- 缺点：
  - 升级预检查
  - 回滚验证策略
  - 参数、插件兼容性
  - 复杂度高、工作量大、挑战大

3. pg_dump

- 优点：
  - 兼容性好
  - 实现简单、工作量小
- 缺点：
  - 仅适用全量迁移
  - 效率低下
  - 应用停服时间长

大版本升级 - 应用不停服零宕机

如何做到?

01 克隆目标实例

目标实例采用类克隆实例方案, 源端实例一直可用。

02 可验证、可回滚

非割接模式, 提供验证能力

连接地址切换之前, 均可回滚。

03 DNS地址切换

切换用户连接DNS地址到目标实例上, 避免应用改动。

04 pg_upgrade元数据升级

pg_upgrade仅元数据升级,耗时与数据量大小无关, 实测2TB数据, 少于10秒。

一键平滑

总纲

介绍RDS PostgreSQL 解决用户痛点问题的方案思考和技术细节

1. Logical Replication Slot Failover
2. 高并发场景审计日志（log_statement=all）优化
3. 内置进程池
4. 一键大版本升级
5. **SGX全加密数据库**

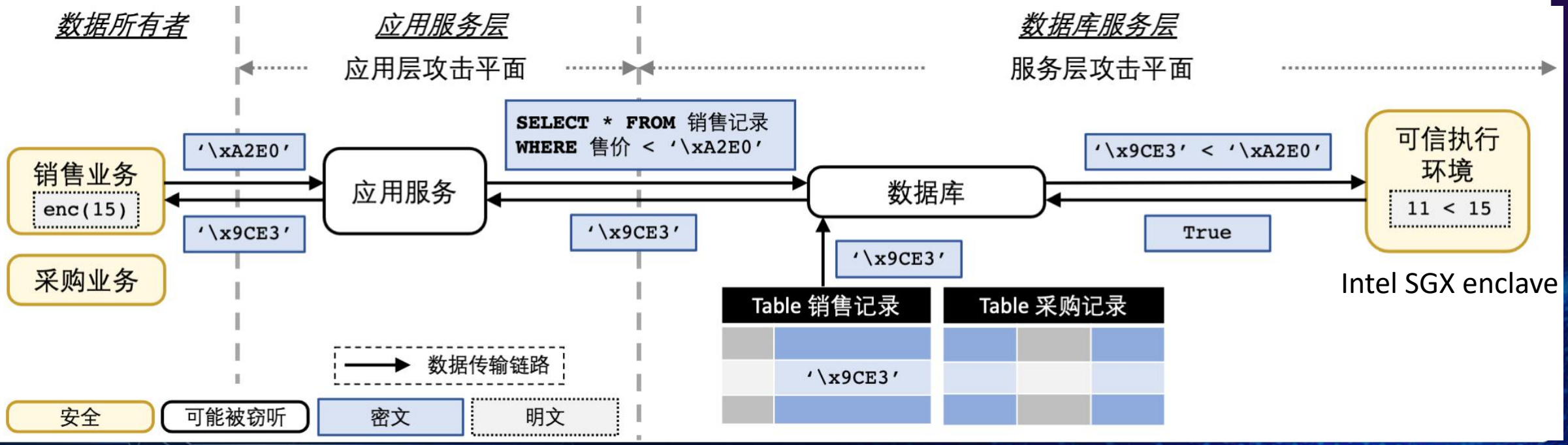
Intel SGX

- 提供物理级的加密计算环境
- **可信根仅包括硬件**，避免了基于软件的可信根可能自身存在安全漏洞的缺陷
- 利用Intel SGX构建最小可信计算基
  - **密钥、涉密代码** 保存在 Intel SGX enclave
  - 明文仅仅存在于 Intel SGX enclave（可信执行环境）的计算过程中
  - Intel SGX enclave 对外的输入、输出均是密文

外部无法获取SGX enclave中的任何信息

SGX全加密数据库 —— 基于可信硬件的全加密数据库

- 利用Intel SGX构建最小可信计算基
- 密文：网络、应用服务内存、PGServer内存、PGServer存储
- 明文：客户端、SGX enclave可信执行环境



SGX全加密数据库 -基于可信硬件的全加密数据库

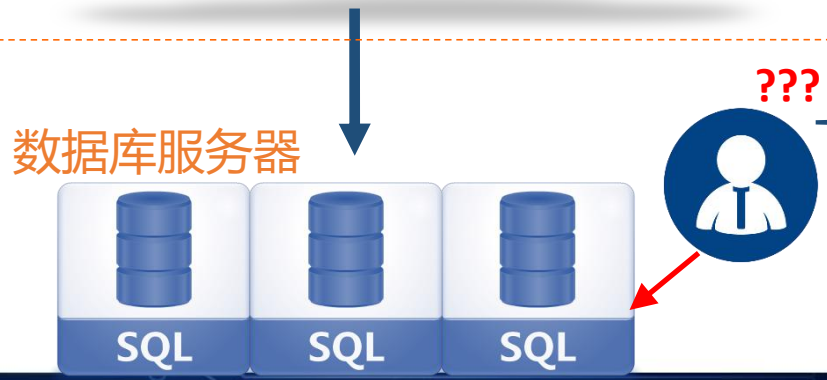

```
SELECT Name, Balance, Addr
FROM Account
WHERE Balance BETWEEN
180,000.00 AND 190,000.00
```

```
Name: Alice
Balance: 183,746.00
Address: 969 West Wen Yi
         Road, Hangzhou
```


```
SELECT Name, Balance, Addr
FROM Account
WHERE Balance BETWEEN
DA6859D786 AND 80EC4071D9
```

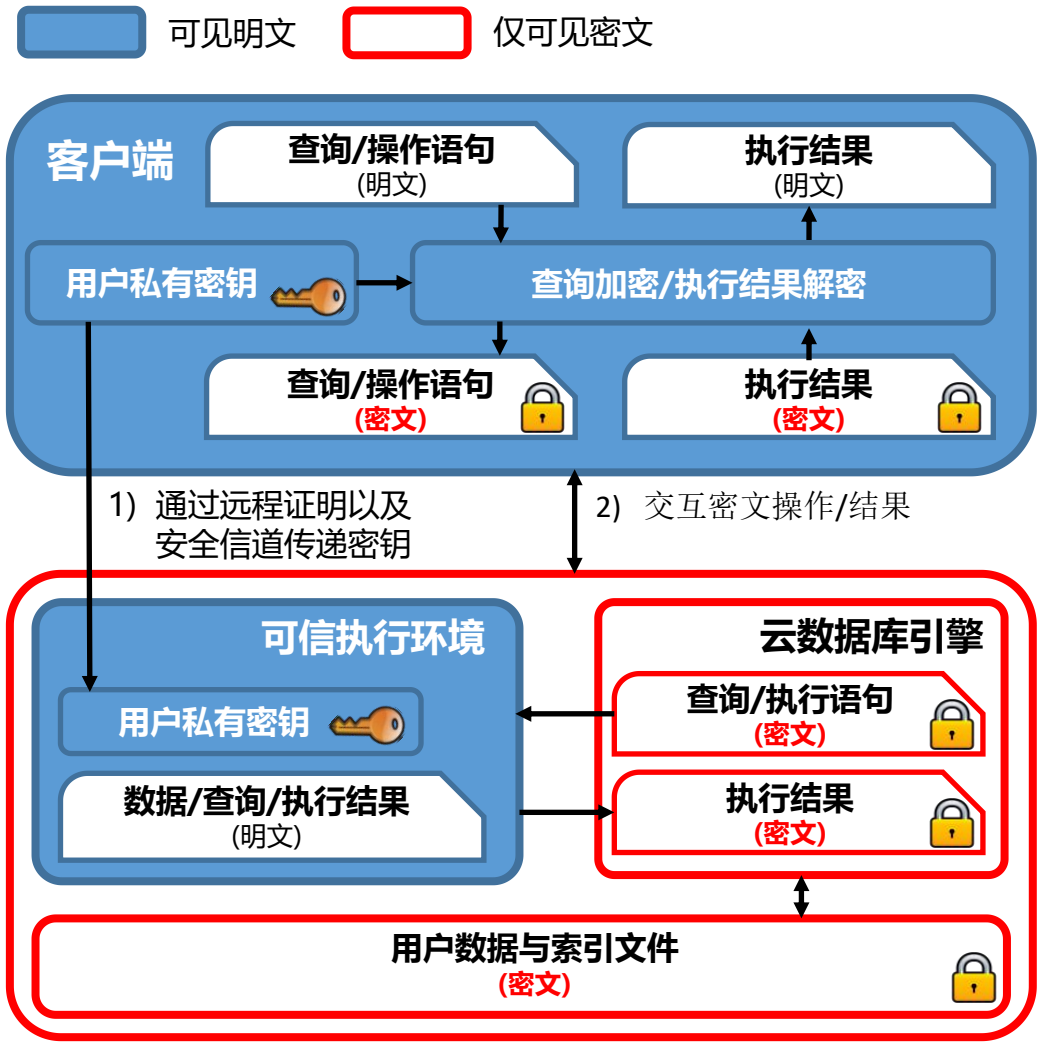
```
Name: FF01AC
Balance: BA695A798B
Address: ACCA00CFE0DAA0
         4AEBB1312624C6
```

用户终端 & 应用服务端



AccountID	Name	Balance	ResidentialAddress
6C26AB	73B725	7F632FD0172	00CFE0DAA0A5AEBB1312AC2D
4AC703	FDDDE9	C1F5680E4AB	B033A0FD9547B425A79BA695
DCDF1D	EFE6A3	2AE0A0490C5	26DA3885FC1942D2EDAFC46B
A14D22	F48BC1	5354DF60B93	04C6314D38B2DEF5745DA7D7
BED9DA	9DD0CE	A4392C6C0CA	25DDBFF01A41AD8EBA183EA1

全加密数据库 —— 基于可信硬件的全加密数据库



❑ 数据不可见（安全性高）

- 客户端加密数据，**全程处理密文**

❑ 数据可处理（功能全）

- 密文上支持数据库**原生SQL能力**
- **索引、范围查询与原生一致**

❑ 数据库生态兼容性

- 【独立组件】与分布式架构、跨域高可用、存储计算分离等技术高度兼容
- 【生态工具】支持数据迁移 (DTS)、备份 (DBS) 等，可支持“应用0改造”迁移

❑ 跨平台密态计算能力

- 支持Intel SGX（x86架构）
- 支持自研密态计算硬件（x86/ARM架构）

SGX全加密数据库 – 支持的数据类型

- ❑ **数据不可见（安全性高）**
  - 客户端加密数据，**全程处理密文**
- ❑ **数据可处理（功能全）**
  - 密文上支持数据库**原生SQL能力**
  - **索引、范围查询与原生一致**
- ❑ **支持单列加密**

数据类型	说明	支持计算
enc_int4	加密后的整形数，对应的明文数据为4字节整形数。	+、-、*、/、%、>、=、<、>=、<=、!=
enc_int8	加密后的整形数，对应的明文数据为8字节整形数。	
enc_float4	加密后的浮点数，对应的明文数据为4字节单精度浮点数。	+、-、*、/、>、=、<、>=、<=、!=
enc_float8	加密后的浮点数，对应的明文数据为8字节双精度浮点数。	+、-、*、/、>、=、<、>=、<=、!=、pow
enc_decimal	加密后的十进制数，对应的明文数据类型为decimal。	+、-、*、/、>、=、<、>=、<=、!=、pow、%
enc_text	加密后的可变长度字符串，对应的明文数据类型为text。	substr/substring、||、like、~~、!~~、>、=、<、>=、<=、!=
enc_timestamp	加密后的时间戳记，对应的明文数据类型为timestamp without time zone。	extract year、>、=、<、>=、<=、!=

RDS PostgreSQL 其他特性介绍

功能增强

- PG14 全网发布
- Ganos时空数据引擎插件
- PASE插件，支持512维的高维向量搜索

安全加固

- 基于Intel SGX的加密数据库
- SSL链路加密
- AD域控、LADP访问控制

易用性

- 基于物理复制的一键上云
- 秒级全量备份
- 一键告警模板设置
- 实例监控新增80个秒级指标

THANKS

谢谢观看