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

厚积薄发，平安自研数据库演进之路

平安科技　熊灿灿

01 PostgreSQL之源

02 PostgreSQL之势

03 PostgreSQL之痒

04 厚积薄发，破局之刃

一专多长

PIPELINEDB　PostGIS　TIMESCALE　AGENS Graph　Greenplum

荣誉

- 1999年荣获Linux World杂志的该年度"最佳数据库产品"称号
- 2000年荣获Linux Journal杂志编辑选择的"最佳数据库"奖。
- 2002年荣获Linux New Media杂志编辑评选的"最佳数据库"奖
- 2003年再次荣获Linux Journal杂志编辑评选的" 最佳数据库"奖。
- 2004年荣获ArsTechnica最佳服务器应用奖。

- 2008 获得Developer.com编辑选择的数据库工具方向的年度产品。
- 2017、2018年连续两年赢得了"全球年度数据库"冠军称号。
- 2019年获O'Reilly终身成就奖，这是继Linux之后第二个获得该奖的开源产品。
- 2020年再次赢得了"全球年度数据库"冠军称号。

主要企业

Pivotal　Google　vmware　去哪儿?　中国平安 PINGAN　中国工商银行　Tencent 腾讯　高德地图

citusdata　NTT Group　IBM　探探　H3C　中国邮政储蓄银行　阿里巴巴

80年代 起源 → 1994 Postgres95 → 1996 PostgreSQL诞生 → 2005 不断发展 → 2021 v14发布

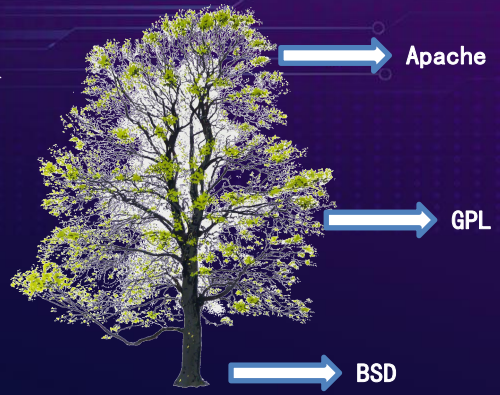
01 PostgreSQL之源

02 PostgreSQL之势

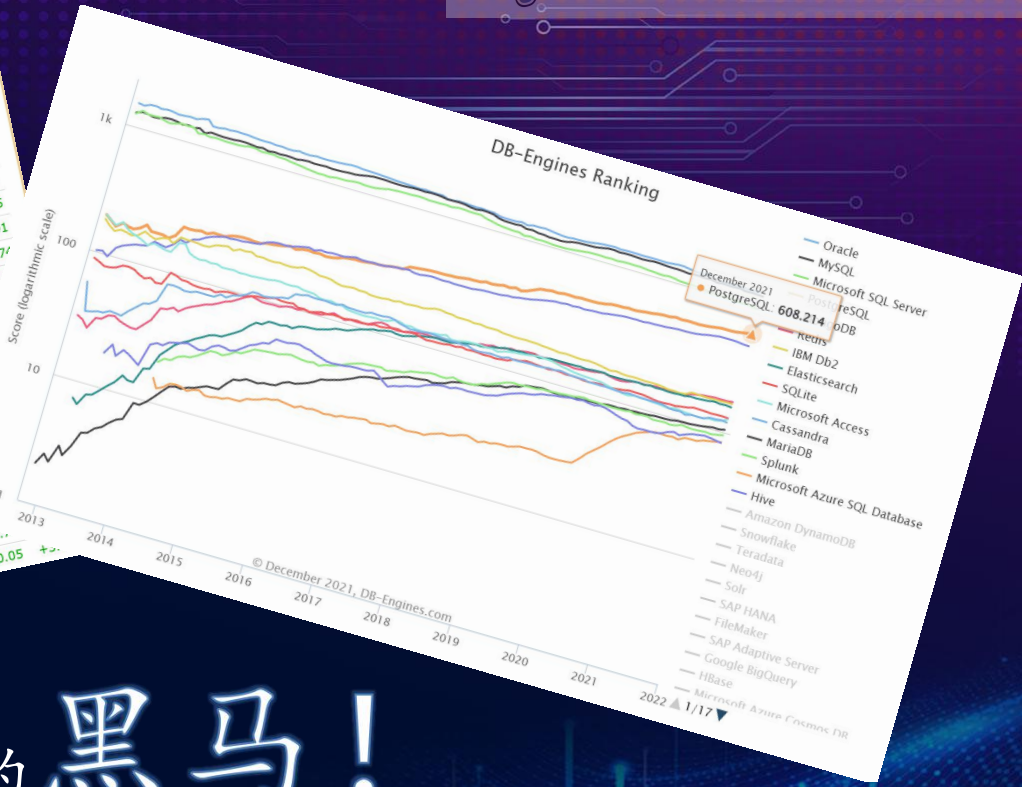
03 PostgreSQL之痒

04 厚积薄发，破局之刃

1. 支持SQL2016大部分特性，至少实现了SQL2011标准中要求的179项主要功能中的160项
2. 功能丰富，利用内核代码中的Hook，可以在不修改数据库内核代码的情况下，自主添加任意功能
3. 社区活跃，生态友好，国内外基于PostgreSQL有很多优秀的产品，这些商业主体会集成和发扬，并反哺社区
4. FDW联邦查询可以在同一个PostgreSQL中像操作本地表一样访问其他数据源
5. 全栈数据库，流式处理Pipelinedb、时序数据库TimescaleDB、空间数据PostGIS、分布式Citus、Greenplum，图数据AgensGraph、NoSQL JSON/JsonB、Hstore等
6. 强大的并行能力，如并行query、seqscan、nestloop join、aggregate、merje join、hash join
7. 强大的物理复制，支持金融级的多副本同步配置，不怕大事务，秒级延迟
8. 协议友好，采用类BSD协议，在使用和二次开发上基本没有限制，企业不会因为分发遇到商业风险，不会因为需要开源核心代码导致辛苦构建的技术壁垒被打破。
9. 扩展接口丰富，与应用深度结合，比如用户画像插件pg_roaringbitmap、虚拟索引插件 hypopg、机器学习插件 madlib等
10. 版本迭代稳定，每年第三季度发布大版本，每个大版本都有重量级特性
11. 真正的全栈数据库，One handle them all，一个打十个，媲美Oracle



Rank Dec 2021	Rank Nov 2021	Rank Dec 2020	DBMS	Database Model	Score Dec 2021	Score Nov 2021	Score Dec 2020
1.	1.	1.	Oracle	Relational, Multi-model	1281.74	+9.01	-43.86
2.	2.	2.	MySQL	Relational, Multi-model	1206.04	-5.48	-49.41
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	954.02	-0.27	-84.07
4.	4.	4.	PostgreSQL	Relational, Multi-model	608.21	+10.94	+60.64
5.	5.	5.	MongoDB	Document, Multi-model	484.67	-2.67	+26.95
6.	6.	↑ 7.	Redis	Key-value, Multi-model	173.54	+2.04	+19.91
7.	7.	↓ 6.	IBM Db2	Relational, Multi-model	167.18	-0.34	+6.7
8.	8.	8.	Elasticsearch	Search engine, Multi-model	157.72	-1.36	+5
9.	9.	9.	SQLite	Relational	128.68	-1.12	+
10.	↑ 11.	↑ 11.	Microsoft Access	Relational	125.99	+6.75	
11.	↓ 10.	↓ 10.	Cassandra	Wide column	119.20	-1.6	
12.	12.	12.	MariaDB	Relational, Multi-model	104.36	+2	
13.	13.	13.	Splunk	Search engine	94.32		
14.	↑ 15.	↑ 16.	Microsoft Azure SQL Database	Relational, Multi-model	83.25		
15.	↓ 14.	15.	Hive	Relational	81.9		
16.	16.	↑ 17.	Amazon DynamoDB	Multi-model	77.		
17.	↑ 18.	↑ 41.	Snowflake	Relational	71.0		
18.	↓ 17.	↓ 14.	Teradata	Relational, Multi-model	70.29	+0.	
19.	19.	19.	Neo4j	Graph	58.03	+0.05	+



发力的黑马！

开源论道 × 数据驱动 × 共建数字化未来

01 PostgreSQL之源

02 PostgreSQL之势

03 PostgreSQL之痒

04 厚积薄发，破局之刃

1. 缺少成熟的ASH、AWR功能：出了故障后能够溯源的手段匮乏，开源的诸如pg_awr、pg_profile等功能单一，缺少关键诊断信息比如wait_event、LWLock、操作系统层的信息等
2. 不支持Failover Slot，意味着假如发生了Failover，消费信息会丢失，对于严格的金融场景，这个比较头疼，需要手动拷贝或者通过技术手段定时记录位点信息
3. 原生流复制若写入过大、网络带宽拥堵等情况下，会造成主从复制延迟，假如WAL日志被归档后会造成主从断掉
4. 32位的事务ID，对于写负载较高的库，经常要面临年龄用完的尴尬
5. 海量连接情况下，TPS随着连接数的上涨线性下降，v14有了一定的性能提升（release不久），缺少原生的进程池
6. 没有好用的列式存储引擎，对AP分析场景稍许吃力
7. 没有原生成熟的TDE（Transparent Data Encryption）
8. 少有的不支持数据压缩的主流数据库（只有TOAST，适用场景有限，Postgre Pro提供了数据压缩，但依赖于商业支持）
9. 独特的MVCC机制，目前来看弊大于利，vacuum经常会因为各种各样的问题，罢工、懈怠等，到了一定的水位，还需要一个昂贵的AccessExclusiveLock来使表变得紧实
10. 没有Hints机制，PostgreSQL并且也不打算支持Hints，原本可以通过一条小小的hints来解决问题，可能需要花费大量时间周期去修改、优化

01 PostgreSQL之源

02 PostgreSQL之势

03 PostgreSQL之痒

04 厚积薄发，破局之刃

可以看到，社区版的PostgreSQL天空中依旧飘着几朵小乌云，总感觉缺了点什么。

基于此，平安科技厚积薄发，充分吸收多年金融场景的实战经验，同时借鉴主流数据库产品的优点，自主研发了一款金融级集中式数据库，RaseSQL，基于PostgreSQL13研发。

平安数据库产品团队

01
2009年
成立专业化团队
引入开发DBA
DBA参与开发全流程

02

03
2013年
自动化运维HiDBA上线
引入MySQL
筹建内部云Dbaas

04
2014年
引入MongoDB
引入Redis

05
2015年
引入PostgreSQL
建立数据库选型原则

06
2016年
引入Oracle+PCServer
运维自动化建设

07
2017年
公有云投产RDS
Oracle迁移PG方案

09
2018年
私有云投产RDS
引入分布式数据库
引入图数据库

10
2019年
金融云投产RDS
政务云投产RDS
行业私有云投产

11
2021年
自研集中数据库
自研分布式数据库

性能分析

性能分析利器

1. 基于AWR/ASH，定时采集获取数据快照，性能问题一览无遗
2. 维度丰富，包括网络、存储、IO、DB等
3. 支持多种风格展示：text、json、html
4. 支持历史性能数据比对

类Oracle风格的html格式报告示例：

RaseSQL Statsinfo Period Report

Summary

- CPU and Memory values are from the end snapshot; averaged across all instances
- Other values are averages for all instances

Database System ID	Host	Port	RaseSQL Version	Snapshot ID Begin	Snapshot ID End	Snapshot Begin	Snapshot End	Snapshot Duratio	Total Database Size	Total Commits	Total Rollbacks	CPUs	Cores	Sockets	Memory
6997962290211990490	SZD-L0167202	5432	1.3	1	3	2021-08-19 14:40:00	2021-08-19 14:50:00	00:10:00	18 MiB	1161	3	4	4	1	7.29 GiB

Database Statistics

Database Name	Database Size	Database Size Increase	Commit/s	Rollback/s	Block Read/s	Block Read/s(disk+cache)	Block Read/s(disk)	Rows Read/s	Temporary Files	Temporary Bytes	Deadlocks	Block Read Time	Block Write Time
postgres	10	0	1.886	0.000	99.1	120.761	1.076	307.606	0	0	0	0.000	0.000
rasesql	7	0	0.048	0.005	93.7	10.373	0.650	58.193	0	0	0	0.000	0.000

Transaction Statistics

- Ordered by date time

DateTime	Database	Commit/s	Rollback/s
2021-08-19 14:44	postgres	3.957	0.000
2021-08-19 14:44	rasesql	0.053	0.008
2021-08-19 14:50	postgres	0.436	0.000
2021-08-19 14:50	rasesql	0.045	0.003

IO Usage

- Ordered by date time

Device	Including TableSpaces	Total Read	Total Write	Total Read Time	Total Write Time	Current IO Queue	Total IO Time
sdb	{pg_default,pg_global}	2924 MiB	25154 MiB	567833 ms	5892217 ms	0.000	6459889 ms
sdc	{""}	0 MiB	70900 MiB	0 ms	198122 ms	0.000	196023 ms
sdf	{pgindexes}	0 MiB	0 MiB	0 ms	0 ms	0.000	0 ms

DateTime	Device	Read Size/s (Peak)	Write Size/s (Peak)	Read Time Rate	Write Time Rate
2021-08-20 15:30	sdb	0.00 KiB (0.00 KiB)	10148.61 KiB (58006.40 KiB)	0.0 %	2.6 %
2021-08-20 15:30	sdc	0.00 KiB (0.00 KiB)	42069.63 KiB (56870.40 KiB)	0.0 %	9.8 %
2021-08-20 15:30	sdf	0.00 KiB (0.00 KiB)	0.00 KiB (0.00 KiB)	0.0 %	0.0 %
2021-08-20 15:40	sdb	4988.84 KiB (140137.60 KiB)	10969.79 KiB (58289.60 KiB)	93.4 %	2.3 %
2021-08-20 15:40	sdc	0.00 KiB (0.00 KiB)	41354.77 KiB (60736.40 KiB)	0.0 %	9.6 %
2021-08-20 15:40	sdf	0.00 KiB (0.00 KiB)	0.00 KiB (0.00 KiB)	0.0 %	0.0 %
2021-08-20 15:50	sdb	2.51 KiB (296.00 KiB)	21812.28 KiB (190054.40 KiB)	1.2 %	977.1 %
2021-08-20 15:50	sdc	0.00 KiB (0.00 KiB)	37582.30 KiB (53334.40 KiB)	0.0 %	13.6 %
2021-08-20 15:50	sdf	0.00 KiB (0.00 KiB)	0.00 KiB (0.00 KiB)	0.0 %	0.0 %

备份恢复

备份恢复工具

1. 支持全量/增量备份、一致性检查
2. 支持远程备份
3. 支持备份管理，统筹管理多实例备份
4. 支持基于时间点的数据恢复
5. 支持部分数据恢复，只还原指定的单个或多个database和tables
6. 支持备份限速
7. 支持备份加密、压缩等
8. 支持S3和Azure兼容对象存储

备份示例：

```
[rasebackup@[redacted] ~]$ rasebackup --stanza=demo --log-level-console=detail backup
--type=full
2021-05-21 10:26:00.319 P00   INFO: backup command begin 1.02: --exec-id=95544-a854c5ca
--log-level-console=detail --log-level-file=detail --pg1-host=[redacted]
--pg1-path=/home/postgres/pg_data --pg1-port=5738 --process-max=4
--repo2-cipher-pass=<redacted> --repo2-cipher-type=aes-256-cbc --repo1-path=/demo
--repo2-path=/phibackup_repo --repo1-retention-full=7 --repo2-retention-full=7
--repo1-s3-bucket=demo --repo1-s3-endpoint=[redacted] --repo1-s3-key=<redacted>
--repo1-s3-key-secret=<redacted> --repo1-s3-region=us-east-1 --repo1-s3-uri-style=path
--repo1-storage-ca-file=/root/public.crt --repo1-storage-host=[redacted]
--repo1-storage-port=9000 --no-repo1-storage-verify-tls --repo1-type=s3 --stanza=demo
--type=full
2021-05-21 10:26:00.320 P00   INFO: repo option not specified, defaulting to repo1
2021-05-21 10:26:06.188 P00   INFO: execute non-exclusive pg_start_backup(): backup begins
after the next regular checkpoint completes
2021-05-21 10:27:36.497 P00   INFO: get_table_infos, table_name:pgbench_history, oid: 17207,
relfilenode:17207, schemaname: public,  db_name: db1, db_id: 17204
```

backup_set	type	start_time	stop_time	archive_start	archive_stop	lsn_start
20210518-095601F	full	21-05-18 09:56:01	21-05-18 09:56:18	000000040000000100	000000040000000100	1/90000028
20210518-095601F_	incr	21-05-18 10:52:30	21-05-18 10:52:59	000000040000000100	000000040000000100	1/B0000028
20210518-095601F_	incr	21-05-18 15:16:04	21-05-18 15:16:13	000000050000000200	000000050000000200	2/38000028
20210518-095601F_	incr	21-05-18 15:35:09	21-05-18 15:35:17	000000050000000200	000000050000000200	2/58000060
20210518-095805F	full	21-05-18 09:58:05	21-05-18 09:58:14	000000040000000100	000000040000000100	1/A0000028
20210518-095805F_	incr	21-05-18 10:54:32	21-05-18 10:54:40	000000040000000100	000000040000000100	1/C0000028
20210518-095805F_	incr	21-05-18 11:21:28	21-05-18 11:21:57	000000040000000200	000000040000000200	2/28
20210518-095805F_	incr	21-05-18 15:13:05	21-05-18 15:13:14	000000050000000200	000000050000000200	2/28000028

数据同步

流式传输归档日志/指定时间点的复制槽

1. 原生PostgreSQL在主库更新量过大，网络带宽拥堵等情况下容易造成主从复制延迟，延迟过大时主库WAL日志被归档或移除后会造成主从复制断开，RaseSQL支持自动从归档目录查找所需日志并进行流式传输。
2. 支持创建指定位点的复制槽，并断点续传

日志分析

日志挖掘剖析

支持强大的日志分析工具，针对数据库日志生成多种维度的分析报表，性能问题一网打尽。

闪回查询

闪回查询

1. select 语句： select * from test as of timestamp '2017-7-14 16:24:19';
2. select into 语句：select * into test_inn from test as of timestamp '2021-11-25 09:15:52.935404+08' where id2 > 100;
3. update from 语句：update test set id= t_user.id from t_user as of timestamp '2021-11-25 15:28:55.914841+08' where test.id != t_user.id;

闪回表

闪回表

1. 闪回表：flashback table rel_flashquery to before drop rename to rel_flashquery_result;
2. 查询闪回结果： select * from rel_flashquery_result;

审计模块

审计模块

1. 多级别审计：实例审计、用户级审计、表级审计
2. 多种操作类型审计：DML、DDL、DQL等
3. 支持审计报表
4. 支持SQL防火墙

数据加密

数据加密

1. 支持透明加密TDE
2. 支持主流加密算法
3. 支持国密SM4

RaseSQL已通过信创测评，完全满足信创要求。在平安核心大BU率先试点信创系统迁移数据库到RaseSQL。

高效率透明数据加密和解密，
对于应用程序完全无感知

将管理数据库对象的权限、管理用户的权限和管理审计日志的权限分离，避免单一管理员权限过度集中

事中通过审计完成告警、记录、防御、阻断
事后通过审计完成安全事件的定位分析、追查取证

数据库层的壁垒，有效杜绝高危SQL

未来已来，将至已至

2020.12

自研集中式数据库立项

2021.6

RaseSQL1.0发布

支持AWR\ASH

企业版备份恢复功能

企业级日志分析工具

数据同步增强功能

2021.11

信创测评

产品功能测评

代码溯源分析

2022.6

1. 安全增强：支持表-回收站功能，数据闪回查询
2. Oracle语法兼容性增强：如insert all，merge into等
3. 功能增强：支持约束的禁用、启用
4. 读写分离实现等

2023.6

兼容PostgreSQL功能增强

兼容Oracle功能增强

性能诊断工具增强

SQL优化工具增强

THANKS

谢谢观看