PGC POSTGRESCONF CN 2021 | PGConf.Asia 12.14-17

从无到有，如何解决传统PostgreSQL下的运维痛点

熊灿灿　平安科技DBA

https://2021.postgresconf.cn

CONTENT

PostgreSQL之我见

世界上最先进的开源数据库

为何选型PostgreSQL

一专多长的全栈数据库

PostgreSQL中的运维痛点

天空中依旧飘着的几朵小乌云

从无到有，从有到优

站在巨人的肩膀上

PART 01

PostgreSQL之我见

最先进的开源数据库

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 PING AN　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	1994	1996	2005	2021
起源	Postgres95	PostgreSQL诞生	不断发展	v14发布

PART 02

为何选型PostgreSQL

身兼数职的全栈数据库

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
1. 强大的并行能力，如并行query、seqscan、nestloop join、aggregate、merje join、hash join
2. 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
3. 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。
4. 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
5. 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

实际案例，经验沉淀

Suning 苏宁

3000+

去哪儿旅行
总有你要的低价

1000+

中国移动
China Mobile

聘 智联招聘

中国平安
PING AN

7000+

中国邮政储蓄银行
POSTAL SAVINGS BANK OF CHINA

HUAWEI

哈啰出行

500+

探探
WWW.DOWNCC.COM

300+

高德地图
amap.com

阿里巴巴
Alibaba.com

PART 03

PostgreSQL天空中的小乌云

飘忽忽的小乌云

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）

PART 04

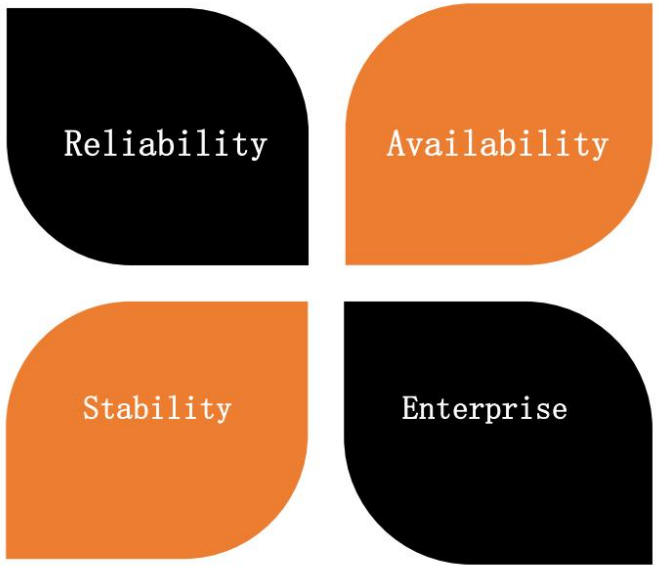
从无到有，从有到优

从无到有，从有到优

那么，有没有一款可以准确痛击这些痛点的数据库呢？

高可靠性

高可用性



高稳定性

企业级支持

从无到有，从有到优

性能分析

支持AWR性能快照\
ASH高频会话信息快照

1. 依赖自研的**awr**，**ash**数据采集插件，实时针对RASESQL进行性能数据采集和定时的**snapshot**快照收集，并对采集数据和**snapshot**的远程数据库进行集中存储。
2. 支持多种格式的报告格式：**text**、**json**、**html**；根据标准的json格式，可定制多种风格的前端展示类型
3. AWR报告支持不同时间段**diff**报告

类Oracle风格的html格式报告示例：

OS Resource Usage

CPU Usage + Load Average

DateTime		User				System				Idle				IOwait				Loadavg1				Loadavg5				Loadavg15			
1st	2nd	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff
2021-08-20 15:30	2021-08-20 15:50	9.3 %	10.9 %	1.60 %	17.20	5.8 %	6.1 %	0.30 %	5.17	84.8 %	82.6 %	-2.20 %	-2.59	0.1 %	0.4 %	0.30 %	300.00	6.220	5.680	-0.54	-8.68	5.810	6.710	0.90	15.49	5.660	6.460	0.80	14.13

Diff average max/min for CPU Usage and Load Average

- AVG means average(MAX means maximum, MIN means minimum) of the specified value

AVG																													
User(%)				System(%)				Idle(%)				IOwait(%)				Loadavg1				Loadavg5				Loadavg15			
1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff
9.3	10.9	1.6	17.20	5.8	6.1	0.3	5.17	84.8	82.6	-2.2	-2.59	0.1	0.4	0.3	300.00	6.220	5.680	-0.540	-8.68	5.810	6.710	0.900	15.49	5.660	6.460	0.800	14.13

MAX																													
User(%)				System(%)				Idle(%)				IOwait(%)				Loadavg1				Loadavg5				Loadavg15			
1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff
9.3	10.9	1.6	17.20	5.8	6.1	0.3	5.17	84.8	82.6	-2.2	-2.59	0.1	0.4	0.3	5.17	6.220	5.680	-0.540	-540.00	5.810	6.710	0.900	15.49	5.660	6.460	0.800	14.13

MIN																													
User(%)				System(%)				Idle(%)				IOwait(%)				Loadavg1				Loadavg5				Loadavg15			
1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff	1st	2nd	diff	%diff
9.3	10.9	1.6	17.20	5.8	6.1	0.3	5.17	84.8	82.6	-2.2	-2.59	0.1	0.4	0.3	5.17	6.220	5.680	-0.540	-540.00	5.810	6.710	0.900	15.49	5.660	6.460	0.800	14.13

Alert

- The 'Message Type' is classified by alert source

DateTime	Message
	First(1st)
2021-08-20 15:20:00	too many transactions in snapshots between '2021-08-20 15:10:00' and '2021-08-20 15:20:00' --- 2566.447 Transactions/sec (threshold = 1000 Transactions/sec)
2021-08-20 15:20:00	load average 1min exceeds threshold in snapshot '2021-08-20 15:20:00' --- 7.24 (threshold = 7)
2021-08-20 15:20:00	load average 5min exceeds threshold in snapshot '2021-08-20 15:20:00' --- 7.34 (threshold = 6)
2021-08-20 15:20:00	load average 15min exceeds threshold in snapshot '2021-08-20 15:20:00' --- 6 (threshold = 5)
2021-08-20 15:20:00	memory swap size exceeds threshold in snapshot '2021-08-20 15:20:00' --- 3048020 KiB (threshold = 1000000 KiB)
2021-08-20 15:30:00	too many transactions in snapshots between '2021-08-20 15:20:00' and '2021-08-20 15:30:00' --- 2557.62 Transactions/sec (threshold = 1000 Transactions/sec)
2021-08-20 15:30:00	load average 15min exceeds threshold in snapshot '2021-08-20 15:30:00' --- 5.66 (threshold = 5)
2021-08-20 15:30:00	memory swap size exceeds threshold in snapshot '2021-08-20 15:30:00' --- 3048020 KiB (threshold = 1000000 KiB)
	Second(2nd)
2021-08-20 15:40:00	too many transactions in snapshots between '2021-08-20 15:30:00' and '2021-08-20 15:40:00' --- 2512.97 Transactions/sec (threshold = 1000 Transactions/sec)
2021-08-20 15:40:00	load average 15min exceeds threshold in snapshot '2021-08-20 15:40:00' --- 5.97 (threshold = 5)
2021-08-20 15:40:00	memory swap size exceeds threshold in snapshot '2021-08-20 15:40:00' --- 3048020 KiB (threshold = 1000000 KiB)
2021-08-20 15:50:00	too many transactions in snapshots between '2021-08-20 15:40:00' and '2021-08-20 15:50:00' --- 2285.17 Transactions/sec (threshold = 1000 Transactions/sec)
2021-08-20 15:50:00	load average 5min exceeds threshold in snapshot '2021-08-20 15:50:00' --- 6.71 (threshold = 6)
2021-08-20 15:50:00	load average 15min exceeds threshold in snapshot '2021-08-20 15:50:00' --- 6.46 (threshold = 5)
2021-08-20 15:50:00	memory swap size exceeds threshold in snapshot '2021-08-20 15:50:00' --- 3048020 KiB (threshold = 1000000 KiB)

Top Event By Pid

pid	event	num
51989	LogicalLauncherMain	21200
51982	AutoVacuumMain	21192
51986	Extension	21163
51985	Extension	20977
53349	ClientRead	20771
53348	ClientRead	20769
51981	WalWriterMain	20097
51980	BgWriterMain	9418
53600	DataFileRead	7098
53603	DataFileRead	7082
53601	DataFileRead	7080
53602	DataFileRead	7031
51979	CheckpointerMain	6978
51980	DataFileWrite	6888
92565	ClientRead	5123
53838	ClientRead	5069

从无到有，从有到优

备份恢复

备份恢复工具

1、支持全量备份、增量备份、一致性检查
2、支持远程备份，从库备份
3、支持备份管理，可将多个实例的备份集中管理
4、支持基于时间点的数据恢复
5、支持部分数据恢复，可以指定N个database级别，或者指定的N个table级别的数据恢复；在部分数据发生异常或丢失时，极大限度的快速恢复数据
6、支持备份限速
7、支持备份加密、多种压缩算法

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

从无到有，从有到优

数据同步

流复制支持归档wal
允许指定LSN的复制槽

1. 原生的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日志被归档后造成的主从断连，RASESQL的WAL Sender支持从归档路径中获取缺失WAL file并发送到standby，从而避免主从断连的现象
2. 创建逻辑复制槽支持指定LSN点位，可以从指定的LSN开始同步数据，并且也支持从归档wal开始同步数据

日志分析

日志文件图形化分析

RASESQL集成了强大的数据库日志分析工具，可针对数据库日志生成多种维度的text、html报表；例如对日志connection、session、checkpoint、lock、query、event等维度进行分析并图表，一目了然。

从无到有，从有到优

2021.6

RASESQL第一个发行版

1、支持AWR\ASH

2、企业版备份恢复功能

3、企业级日志分析报表工具

4、数据同步增强功能

2021.7~8

内部云投产RASESQL数据库产品

2021.Q3~Q4

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

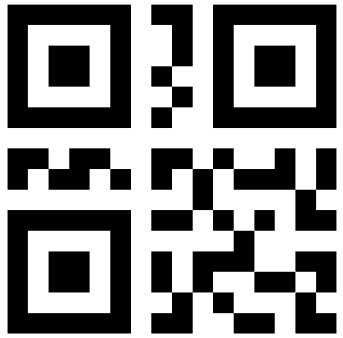
4、审计模块

5、透明加密（支持国密算法）

6、读写分离实现等

THANK YOU

CONTACT INFORMATION



CHINA POSTGRESQL ASSOCIATION

PCC POSTGRESCONF CN 2021 | PGConf.Asia 12.14-17