

金融级 PostgreSQL监控与优化

梁海安

平安科技（深圳）有限公司



平安科技
PING AN TECHNOLOGY

受托于平安集团，向集团公司和集团所有下属子公司提供IT规划、开发和运营服务的IT服务提供商。

- ✓ 2013年开始引入MySQL等开源数据库
- ✓ 2015年开始正式推广PostgreSQL
- ✓ 已有PostgreSQL实例**1000+**
- ✓ 上线两年零故障



平安科技
PING AN TECHNOLOGY



content



监控实现



性能快照



运维优化

1 监控实现

监控作用

及时发现
主动预防



确定影响范围



定位root cause



快速恢复

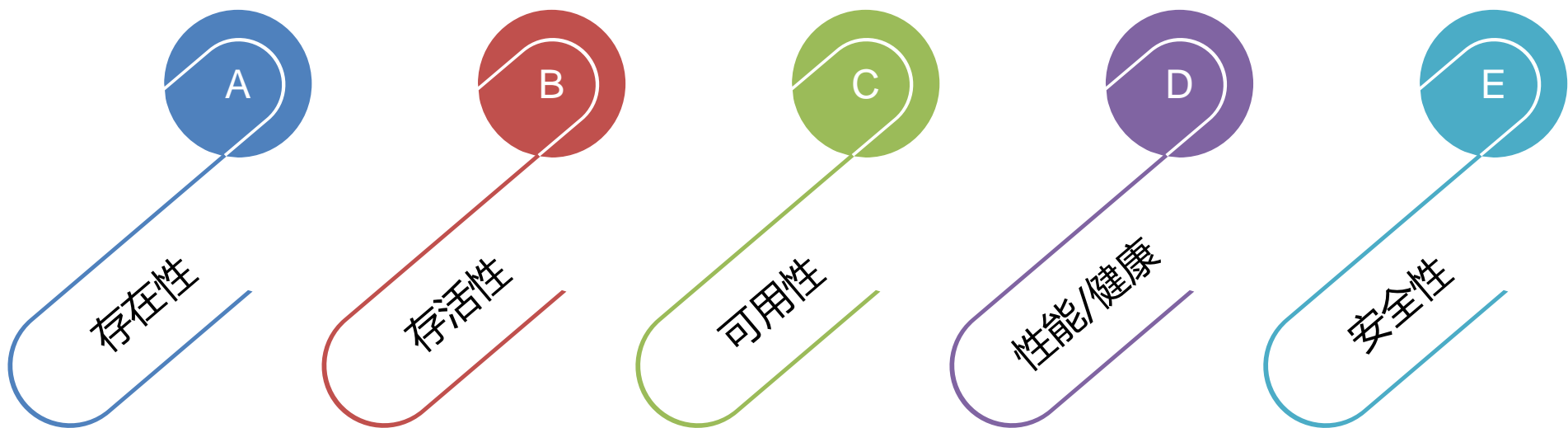


监控方法论



监控原则

应用系统按组件和流程两个维度分解监控节点，每个节点按5个监控原则分析，形成全面的监控体系。



PG监控原则

存在性

已配置未监控
已监控未配置
完整性/准确性

PART 1

存活性

监听
进程
NoDATA

PART 2

可用性

连通性
可读写
主从lag

PART 3

性能/健康

慢SQL
锁队列
XID剩余
错误日志
归档异常

PART 4

安全性

用户登录限制
用户过期策略
用户过期清理
审计策略
异常操作

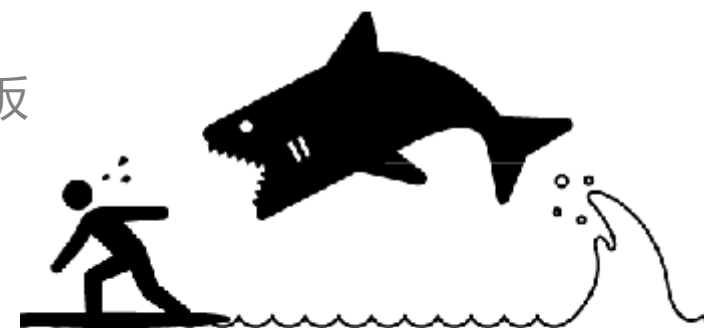
PART 5



PG监控展示

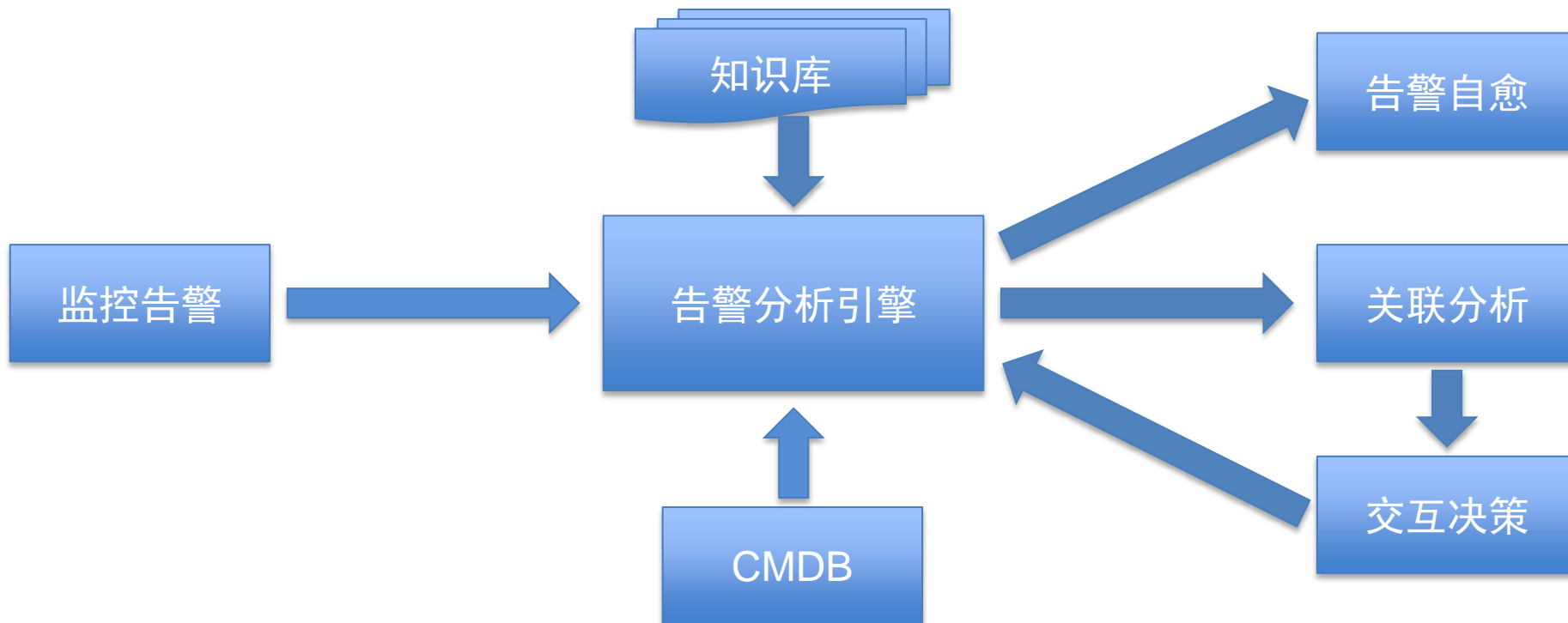


监控自动发现



监控分析

从知识库中智能筛选出有用信息，执行既定方案的自动异常恢复、或推送关联信息给处理人决策，快速恢复异常



监控数据分析

多种分析算法进行多层次的数据分析，产生不同级别的预警，帮助发现异常和隐患。



阈值分析

- 区间值告警
- 匹配值告警
- 命中次数告警
- 自定义算法



趋势分析

- 正态分布分析
- 四分位分析
- 移动平均分析
- 自定义算法



关联分析

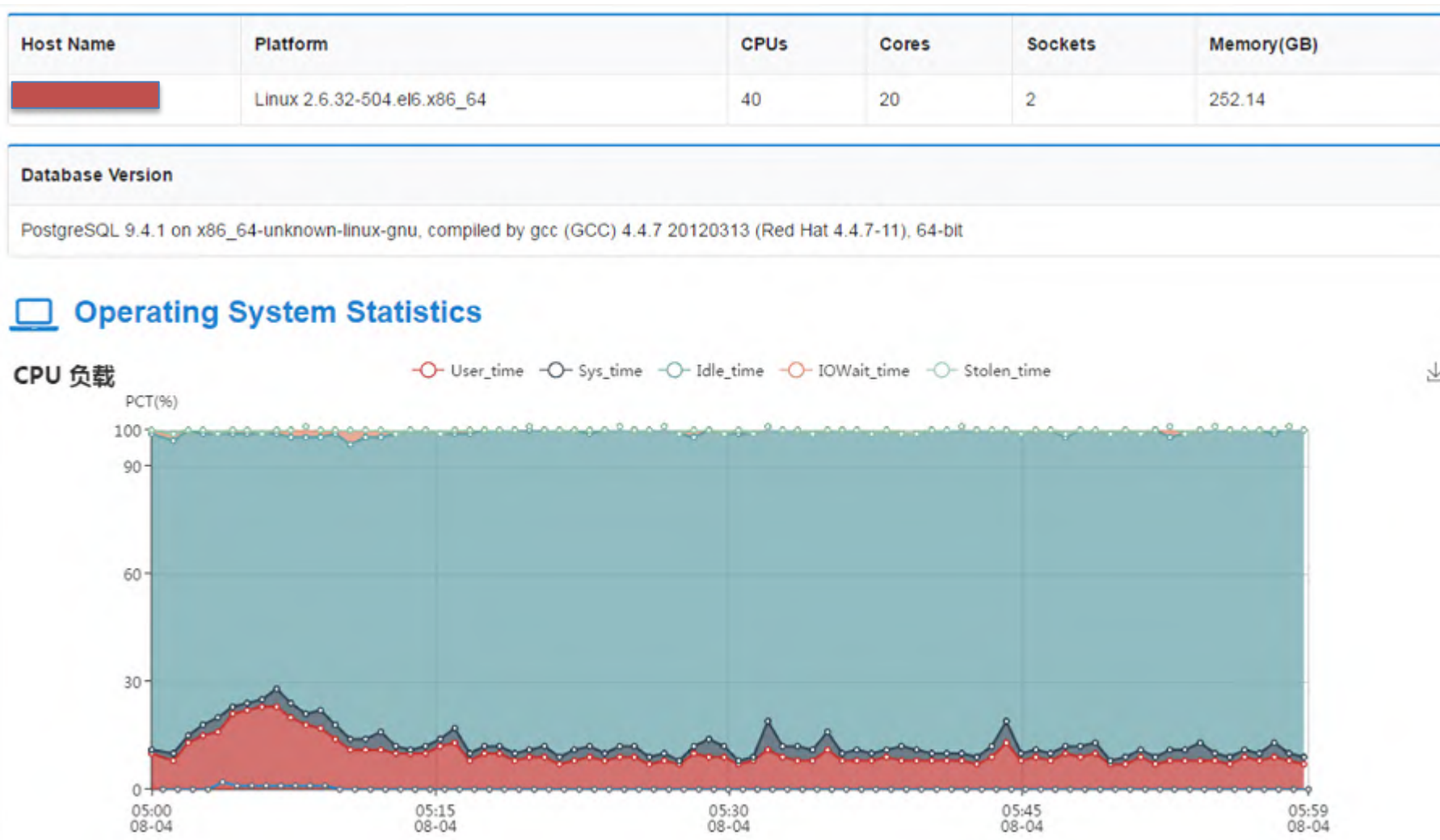
- 相似度分析
- 偏离比分析
- 自定义算法



2 性能快照

主机信息

概览中包含主机的一些基本信息。同时结合Oracle的OSW采集的数据，生成CPU的负载情况，实例自身的瞬时数据来自cgroup。也可以根据其他需求增加负载图，如内存，网络，IO等。



数据库统计

根据DB的读写情况，可以分析内存设置，也可以进一步分析慢SQL。

通过checkpoint情况可以分析checkpoint相关参数，写相关参数是否合理。

Database Statistics

Database	Tps	Hit Rate	Logical IO/s	Physical IO/s	Rollback/s	Deadlocks	Total read time(s)	Total write time(s)	DB Size	Inc Size
users	0.54	97.00	16.65	0.38	0.00	0	103	0	1033 MB	208 kB
	0.07	99.00	23247.27	29.42	0.00	0	20492	1	263 GB	21 MB

Background writer stats

Checkpoints Timed	Checkpoints Req	Buffers Checkpoint	Buffers Clean	Maxwritten Clean	Buffers Backend	Buffers Alloc
12	0	76982	0	0	2720	4227

Checkpoints Timed	Minutes between Checkpoint	Buffers Checkpoint	Buffers Clean	Buffers Backend	Total Writes	Avg Checkpoint Write
100%	5	96%	0%	3%	0.173 MB/s	50.000 MB



SQL统计

通过多个维度对SQL进行分析，定位问题SQL。

Top 20 SQL statements ordered by Elapsed time

Queryid	Calls	Total time(ms)	%Total	Total time/call(ms)	Rows	User	Query
109007142	75654	1468439.983	49.51	19.410	75654		
459705577	75654	1461546.109	49.28	19.319	75654		

Top 20 SQL statements ordered by Physical read

Queryid	Calls	Shared Read	%Total	Shared Read/call(ms)	Rows	User	Query
2716299919	4	1420.000	19.40	355.000	4		
109007142	75654	1330.000	18.17	0.018	75654		

Top 20 SQL statements ordered by IO read time

Queryid	Calls	Read time(ms)	%Total	Read time/call(ms)	Rows	User	Query
1098006768	335	228.919	15.71	0.683	335		
3389415136	335	228.919	15.71	0.683	335		



对象统计

Top 20 tables ordered by high table to index read ratio

Table Name	% Total Database	% Table Read	% Index Read
	99	100	0

Top 20 tables ordered by percentage of tables scanned

Table Name	% Rows Read	% Table Hit	% Index Hit	Table Read	Table Hit	Index Read	Index Hit
	99	99	99	1120	81009494	416	152648

Top 20 tables ordered by high table to index read ratio

Table Name	Live Tuples	Dead Tuples	% Total Database	% Table Read	% Index Read
	535	0	20	99	0

Top 20 Table order by bloat ratio

Table Name	Table Size	Estimate Size	Live Tuples	Avg Row Len	Last Analyze	Bloat Ratio(%)
	264 MB	1338 kB	4241	323	2016-10-23 20:06:17.583762+08	100

通过多维度的对象分析，定位关键对象，合理调整对象参数，如seq的cache size。

对象分析也可以反映出表对象上是否需要增加索引，或者存在多余的索引。

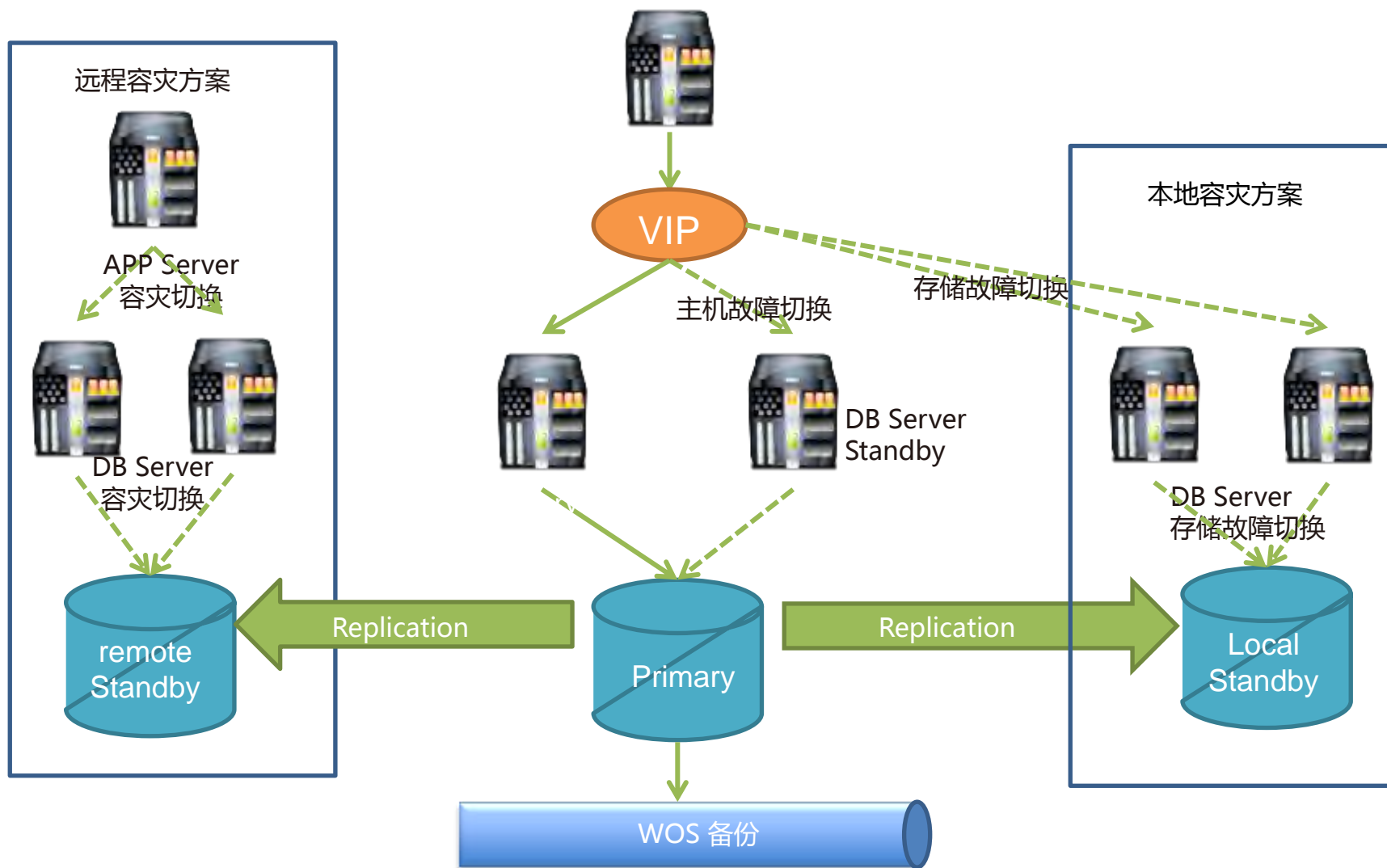
区别于pgtuplestat，通过对一些统计值的计算，大致评估表膨胀和索引膨胀的系统合理安排运维维护。



3 运维优化

架构优化

- 使用传统高可用架构
 - ✓ 计算与存储分离
 - ✓ 更高的可用性
 - ✓ 更快速的迁移扩容
- 每个子网段端口唯一
- 使用廉价对象存储做归档与备份
- 增大编译参数segsize , 以减少数据库文件数
- 增大编译参数walsize , 以减少归档文件数



远城归档传输优化



recovery.conf

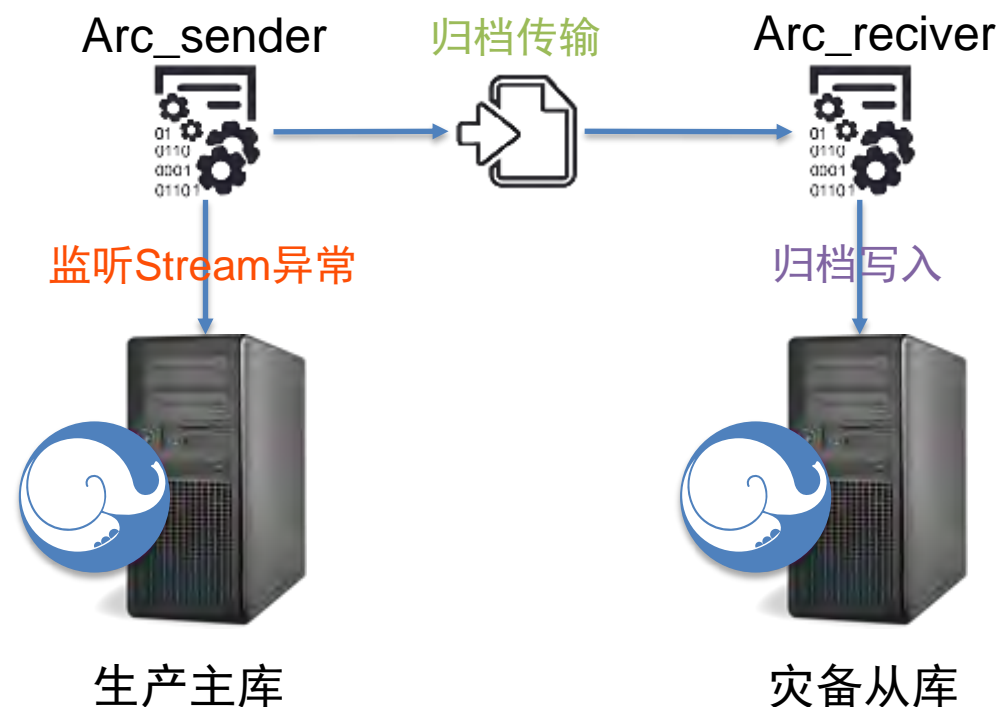
```
recovery_target_timeline = 'latest'
```

```
standby_mode = 'on'
```

```
primary_conninfo = 'XXXX'
```

```
restore_command='cp /paic/xxxx/archive/%f %p '
```

```
archive_cleanup_command = 'pg_archivecleanup  
/paic/xxxx/archive %r'
```



备份优化



pg_rman 进行日常增备和全备

问题：

单线程，备份效率低，4T数据库需要3天（每天wal日志1T）
使用CRC校验备份文件，需要真实从wos读取文件

优化：

多线程改造
不备份归档wal日志，仅记录文件信息
只比较文件size，后续加强使用MD5进行校验



大表与Vacuum优化

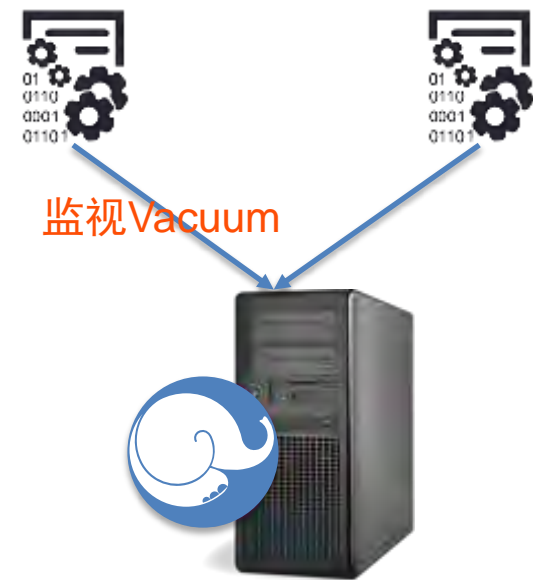
Daily Vacuum方案：

1. 利用业务空闲时间窗加大vacuum数量
2. 并发度根据主机资源动态调整
3. vacuum内存根据表大小动态调整
4. 任务队列根据表大小，age，dead tuple PCT，失败次数等条件全局排序，避开Auto Vacuum表
5. 记录所有Vacuum执行情况供事后分析
6. 增加Vacuum Monitor进程，
 - 中断阻塞DDL的Vacuum进程
 - 维护时间窗口外终止任务队列，取消非auto的vacuum进程。

大表优化：

- 大表必须做分区
- 在线表和历史表分离
- 使用trigger时，保持trigger 函数简单、高效

Vacuum_monitor Vacuum_dispatch

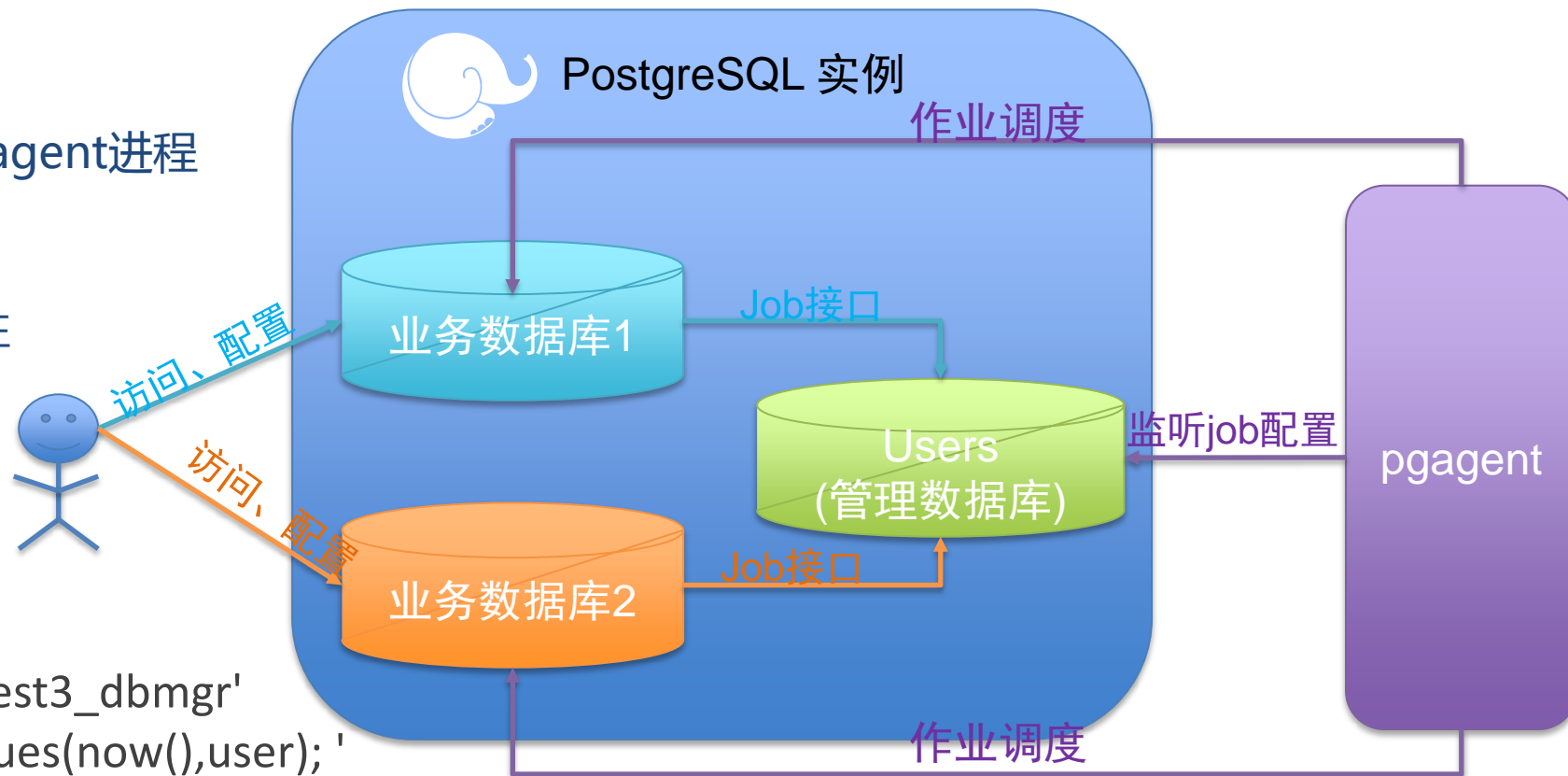



生产主库



pgagent优化

- ✓ 每个cluster对应一个pgagent进程
- ✓ 统一的接口函数与视图
- ✓ 更好的可操作性和可读性



 `select dbms_job.submit('test3_dbmgr'
, 'insert into t2(t,u) values(now(),user); '
, '2/5 5-18 * * *') jobid;`



运维优化

基于pg_stat_activity视图高
频采样

ASH

Flashback query

闪回查询

多线程备份

备份

分表

更好更快的分区实现

分库

并行计算，OLAP仓库



4 隐藏内容

不仅仅是监控



规范制定

规范是灯塔
也是牢笼



配置管理

关联系统组件
关联流程规范



DBaaS

优化人力配置
规范的最好实施者



运维监控

主动预防
及时处理



规范制定



架构规范

机器选型
操作系统版本
操作系统用户配置
存储卷的规划



开发规范

规范SQL写法
避免重复踩坑



用户规范

角色分工
最小权限原则
保障数据安全
Schema使用



安全基线

角色密码复杂度
禁用默认用户
禁用默认端口
审计DDL
下线审计





规划CMDB

CMDB管理着所有IT资源，
是监控和运维自动化的基础

- ✓ 规划层级，理清关系
- ✓ 自动发现和主动配置结合
- ✓ Local配置能加速运维操作
- ✓ CMDB包含的信息能应对灾难恢复



平安运维管理经验

PATCH

Health Check

Safety Compliance

SQL Audit

Upgrade

Impact Analysis

Capacity Expansion

... ..

平安数据库云平台

构建DBaaS



释放人力

把人力从低价值，重复的劳动中抽出来。

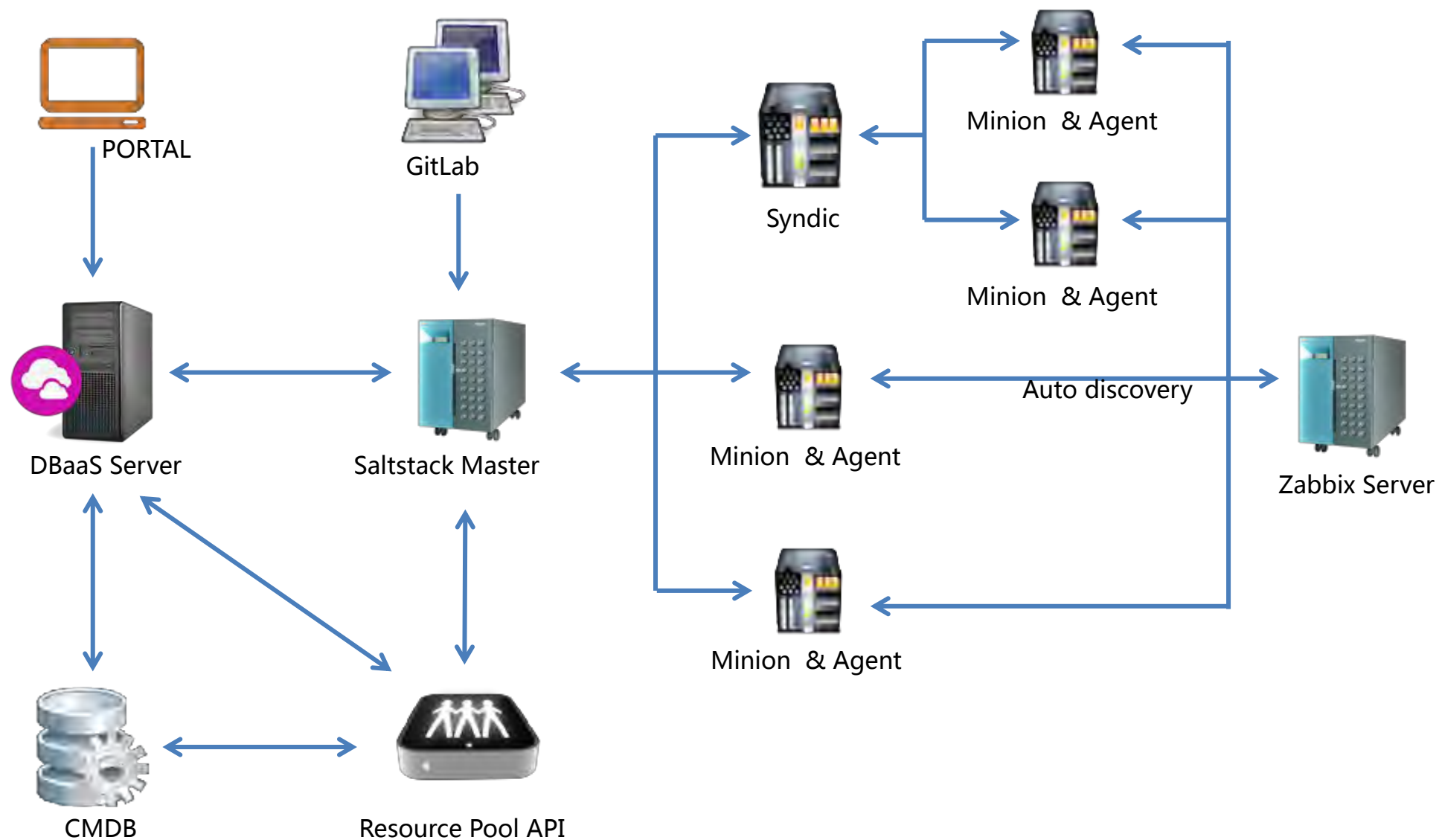


自动化一切

创建各环境数据库实例
数据库用户自助申请
数据库版本审核
影响分析
安全合规检查
故障自愈
自助报表

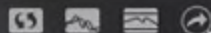


平台架构



端到端监控

EPCIS-NBA



用户行为

用户体验

业务逻辑

业务层



WEB

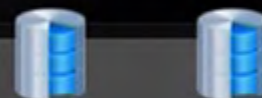
NAS卷

APP

DB

SAN卷

应用层



WEB服务器

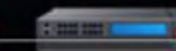
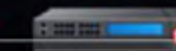
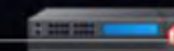
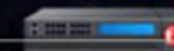
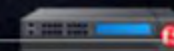
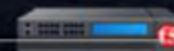
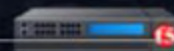
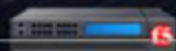
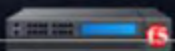
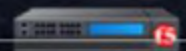
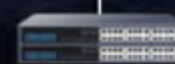
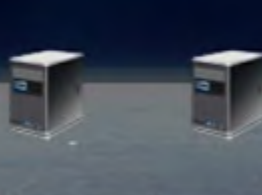
NAS

APP服务器

DB服务器

SAN

物理层



Postgres Conference China 2016 中国用户大会



平安科技
PING AN TECHNOLOGY

Thanks!

Q & A