DBAplus

数据库年终盘点大会-上海站

Oracle12c多租户体系下的容灾设计

杨欣捷

01

什么是资源池

- 共享

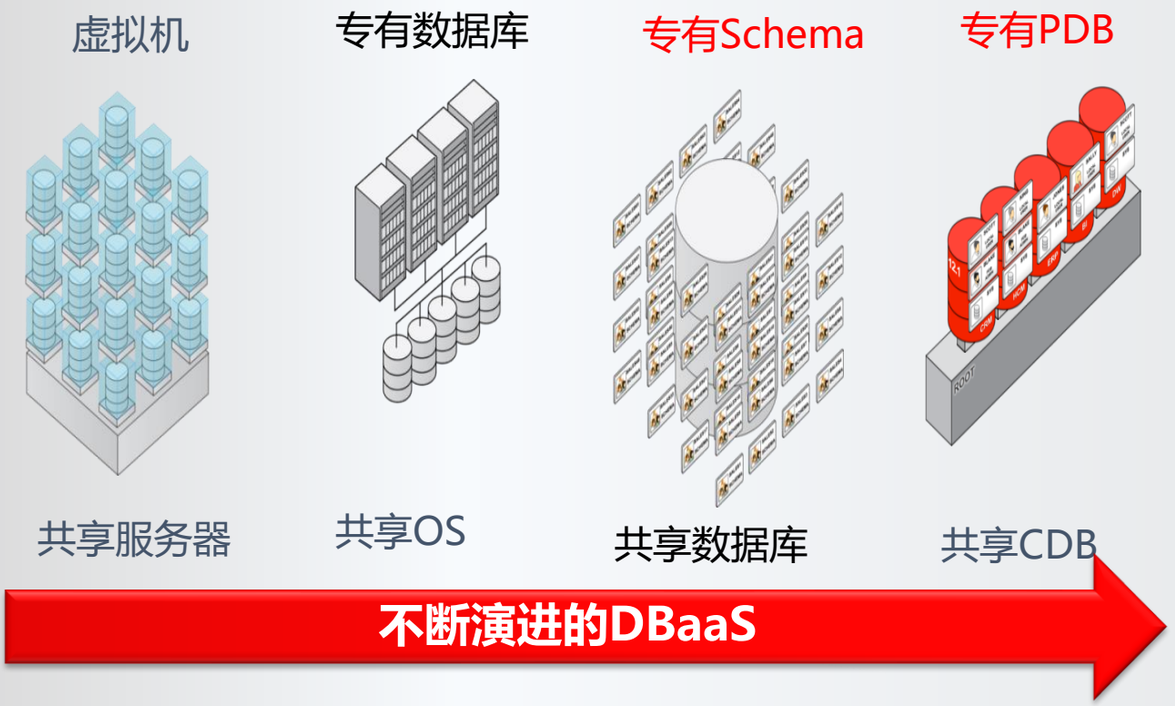
- 标准化

- 按需、快速获取

- BaaS (Bike as a Service）

DBaaS

- 为什么要基于Oracle12c构建数据库资源池？



- 虚拟机&共享OS：整合度太低、无横向扩展能力
- Schema方式：隔离性有限，应用相关、无原生的快速发布接口，自服务能力差

- Oracle12c
  - 命名空间隔离
  - 权限控制
  - 更好的资源隔离
  - 灵活的service
  - ……

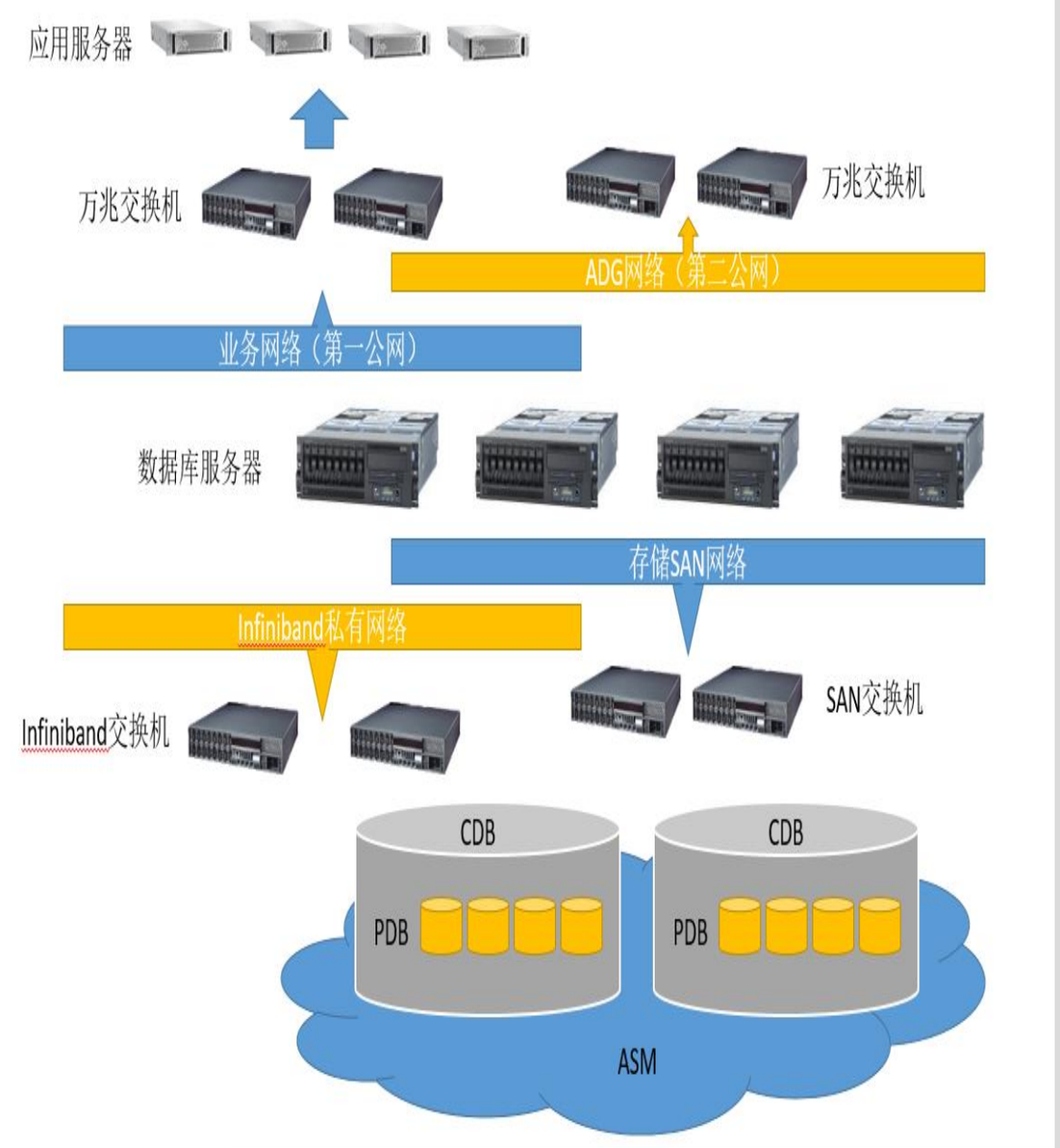
- 为什么共享单车2016年火了起来？

全面DBaaS的时机已经成熟

- 时机已经成熟，合理的规划及技术使用，是资源池成败的关键

基础架构

- 4节点集群
- X86架构，4路服务器，56C1T
- Infiniband (EDR 100GB) 实现私网
- 每节点配置4块16GB HBA访问存储
- 配置NVme SSD卡用于DB二级缓存
- 操作系统使用OEL 7.3
- 集群存储使用ASM方式管理



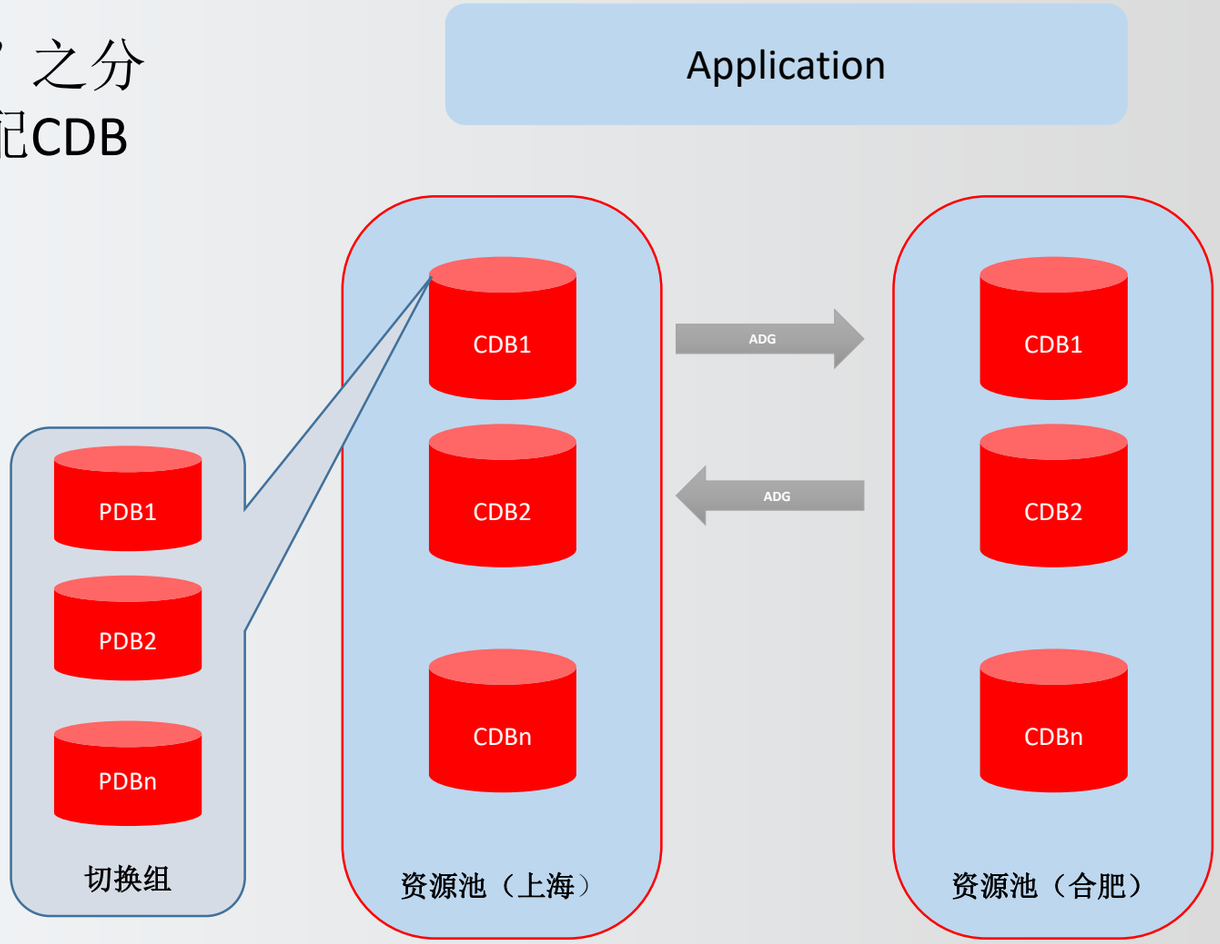
网络设计

	第一路公网-业务网段		第二路公网-ADG网段	
节点1	业务网段IP	4个 业务网段VIP	ADG网段IP	4个 ADG网段VIP
节点2	业务网段IP		ADG网段IP	
节点3	业务网段IP		ADG网段IP	
节点4	业务网段IP		ADG网段IP	
SCANIP	1个或3个业务网段IP		不提供	

- 集群私网
  - 2个Infiniband交换机实现内联
  - 目前使用IPoIB，未来使用RDS传输协议。
  - HAIP实现网络冗余
- 对外网络
  - 三个SCANIP供DNS轮询。应用通过DNS+服务名方式访问数据库
- 日志传输网络
  - ADG网络为第二公网
  - 配置ADG专用的VIP地址和监听器
  - 使用VIP保障ADG传输在节点故障时的连续性

资源池容灾架构

- ADG的保护单位是CDB
- CDB建立ADG关系自动作用于其中所有PDB
- 双中心的资源池分别有作为主库和ADG库的CDB
- 没有“主站点”和“备站点”之分
- 根据容灾一致性切换组来分配CDB



优点：

- “逻辑”故障的保护能力
- 备库可读，资源利用率高
- 更快的切换速度

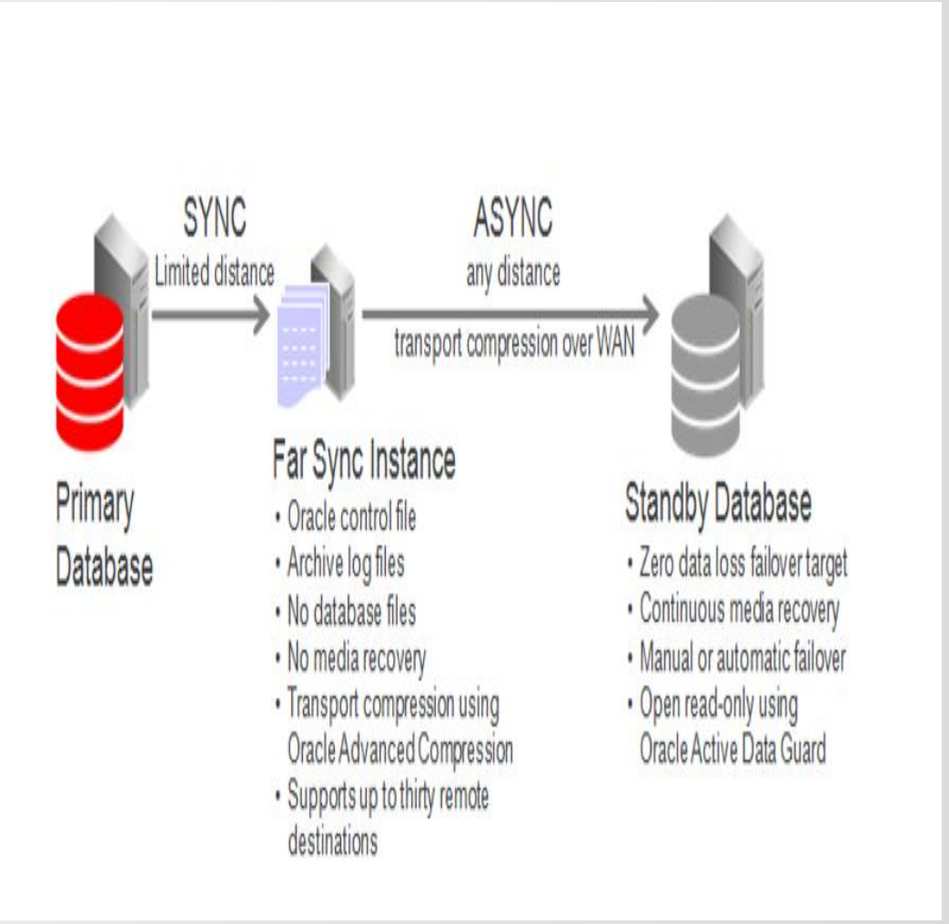
运维的怒吼


我们是谁！
运维！
资源池容灾满意吗？
不满意！！
哪里不满意？
RPO=0的需求怎么办！！！


我们是谁！
还是运维！
资源池灾备演练方便吗？
不方便！！
哪里不方便？
切换A应用为啥要带B应用？

Far Sync

- Oracle12c引入
- 位于同城的轻量级实例
- 不提供应用访问
- 只有控制文件和日志文件
- 只需要少量计算和存储资源
- 同城同步传输（fast sync）确保灾难情况下0数据丢失



低成本解决RPO=0的容灾需求

部署实践

- 实践中Far Sync节点部署2个，作为高可用
- 物理设备上可以建立多个Far Sync实例应对多个CDB的容灾需求
- Far Sync实例分别作为log_archive_dest_2和log_archive_dest_3，ADG备库为log_archive_dest_4
- 通过配置不同的优先级参数让数据库选择相应的dest传输日志
- Sync noaffirm表示使用fast sync
- 实测far sync节点实例故障切换时间5s左右

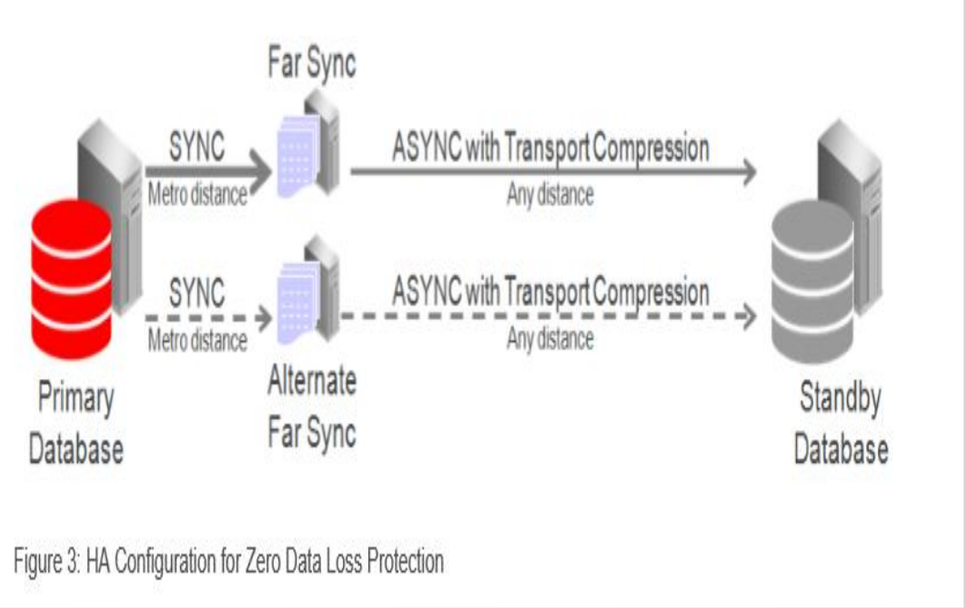


Figure 3: HA Configuration for Zero Data Loss Protection

Example:

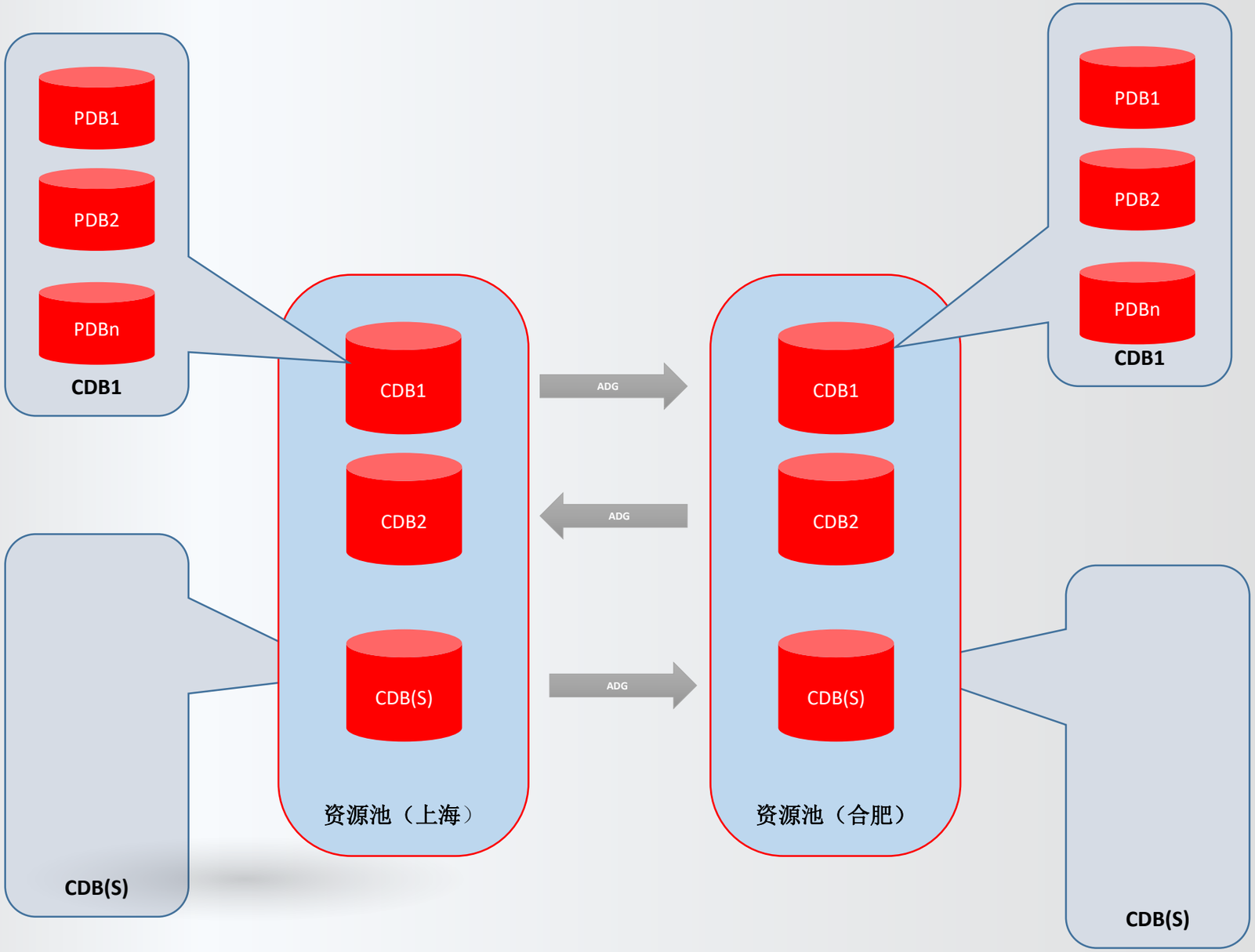
```
log_archive_dest_3='SERVICE=FARSYNC2 SYNC
NOAFFIRM net_timeout=10
valid_for=(online_logfile,primary_role)
db_unique_name=farsync2 GROUP=1 PRIORITY=2';
alter system set
log_archive_dest_state_3=ALTERNATE;
```

单个PDB切换

- 场景特点分析
  - 站点故障所有CDB均需要切换
  - 本地服务器故障由集群高可用解决
  - 数据逻辑错误ADG无法应对
  - 不需要数据回传的演练直接用snapshot standby方式演练
  - 需求来自于“演练”专用
  - Oracle官方将单个PDB切换作为18c的feature

- 应对方法
  - 建立一个平时同步着的灾备“演练专用”CDB
  - 多租户带来的问题用多租户的方法解决
  - 将需要切换演练的PDB插拔到“演练专用”的CDB中
  - 演练时对“演练专用”CDB做switchover
  - 演练完毕再将PDB插拔回原来的CDB
  - 整个过程不需要重建ADG关系，额外增加的停机时间很少

单个PDB切换


PDB1
PDB2
PDBn
CDB1
CDB(S)
CDB1
CDB2
CDB(S)
资源池（上海）
ADG
ADG
ADG
CDB1
CDB2
CDB(S)
资源池（合肥）
PDB1
PDB2
PDBn
CDB1
CDB(S)

单个PDB切换

- 实践中的各种坑
  - ENABLED_PDBS_ON_STANDBY参数

这个参数默认为*，意思是所有在CDB上建立的pdb都会同步到ADG库。但是在实际使用中出现，同步过去的数据文件不能正常创建，而是创建成了$ORACLE_HOME/dbs/UNAME000这类文件。alter system set ENABLED_PDBS_ON_STANDBY='*'（是的，重复赋同样的值）就能后再创建就没问题了。

  - 如果仅仅是在mount的recovery manage状态，此时主库创建pdb，备库的MRP进程会立即crash，后续无法启动MRP进程。

  - 需要将备库的pdb1的数据文件转移至自动能“认出”的位置，这个位置经过测试目前基本来说是“Hard Coding“。格式为：<db_file_create_dest>/<db_unique_name>/<pdb1guid>/DATAFILE/。为此必须创建alias

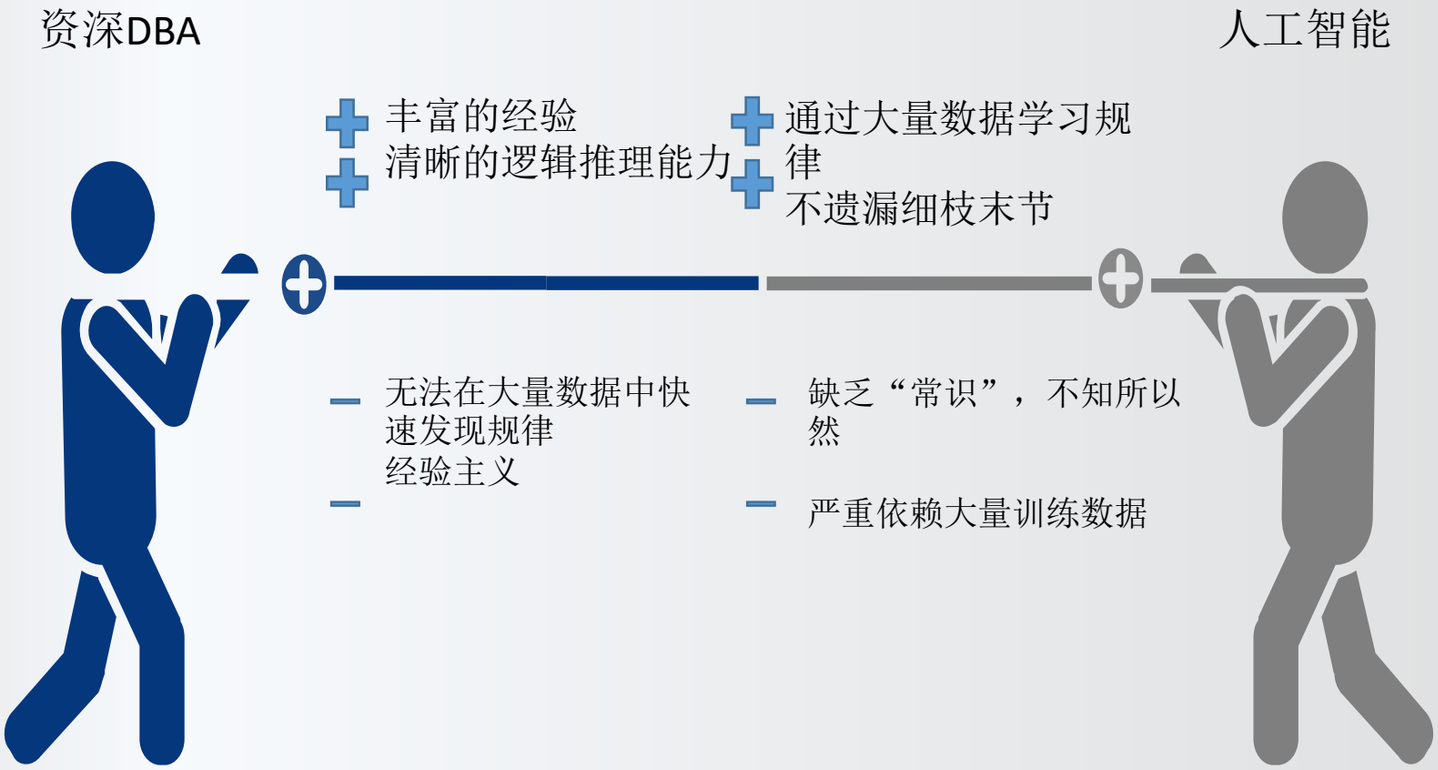
由于Alias不能跨ASM磁盘组指向，这意味着”目标库”的db_file_create_dest必须和源库所有文件所在的ASM磁盘组一致。也就是说，这种方式不支持源库的单个PDB数据文件不分布在2个或以上的ASM磁盘组中。

总结&下一步工作

- DBaaS提供了整合和弹性优势
- 配套运维架构和模式必须调整

- 资源调配
  - 如何在生产上有效监控资源池内的PDB实际资源使用情况
  - 基于监控数据预测资源需求，及时调度
  - 结合云管平台实现DBaaS自服务门户
- 版本升级
  - 资源池环境下针对单个PDB实现平滑的版本升级
- 压测评估
  - 资源池环境下如何解读压力测试数据
  - 在干扰环境下获得应用的真实容量负载

优缺点互补


资深DBA
人工智能
丰富的经验
清晰的逻辑推理能力
通过大量数据学习规律
不遗漏细枝末节
无法在大量数据中快速发现规律
经验主义
缺乏“常识”，不知所以然
严重依赖大量训练数据

- 故障现象：
  - 开发环境12c资源池中一个节点ORA-4031，所有PDB无法访问，sqlplus无法登录
  - Shared_pool中一个component “gc index split transactio”增长到了80G
  - 网上查不到任何gcindex split transactio的资料
  - 重启后可以正常使用，但“gc index split transactio”仍然以每天6G的速度增长

- 分析思路：

  - 首先肯定是开SR，但是SR效率较低
  - 不管是不是BUG，需要找到触发原因，规避问题
  - Oracle动态视图和AWR中存有大量数据
  - 设法找出gc index split transactio增长和其他“事件”的关联

  选取V$SYSSTAT中所有STAT_NAME，然后把每个快照期间他们的增量和“gc index split transactio”的增长量求一个相关系数。DBA_HIST_SGASTAT是累积量，需要使用分析函数获取增量

```
select a.snap_id,a.stat_name,a.value-lag(a.value)
over(order by a.snap_id) logon_delta,(b.bytes-lag(b.bytes)
over (order by b.snap_id))/1024/1024 sga_delta
from dba_hist_sysstat a, dba_hist_sgastat b
where a.snap_id=b.snap_id and a.instance_number=1 and
b.instance_number=1 and a.stat_name=:name and b.name
like '%index split%' order by 1;
```

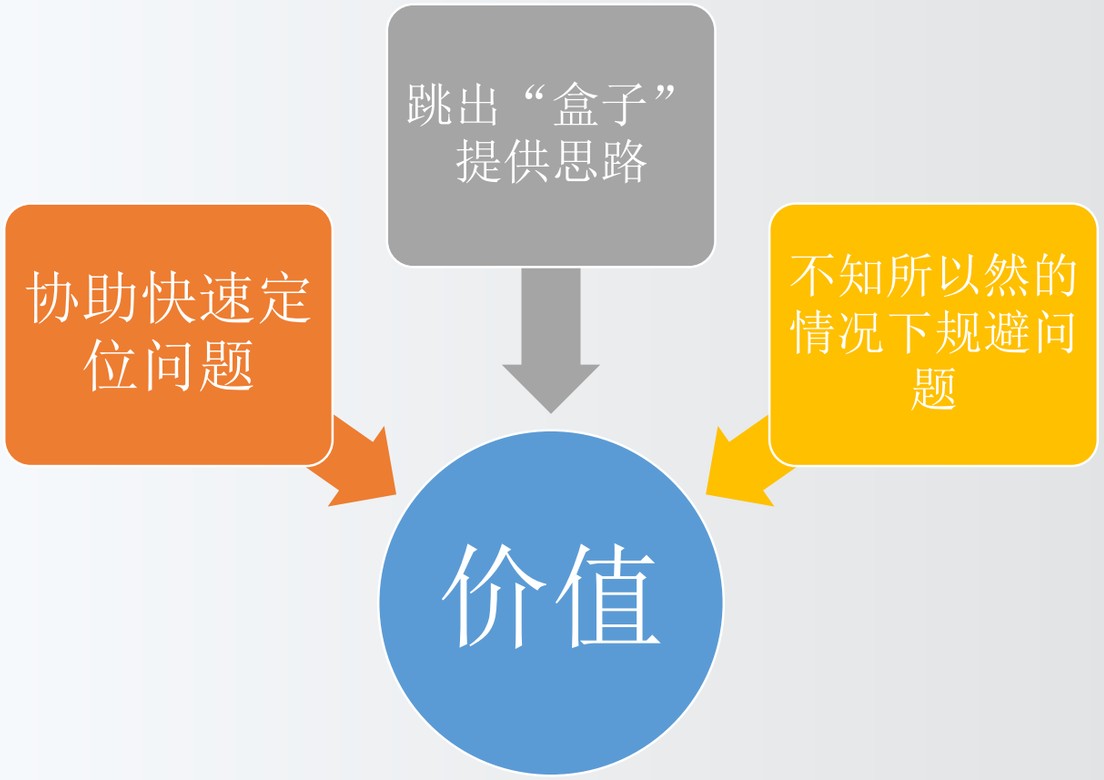
- 分析结果：

将每一个stat_name和gc index split transactio计算相关系数（Python）。
排名靠前的指标和相关性：
('redo synch time overhead count (8ms)', 0.2911595045810626), ('redo synch long waits', 0.4429637320736422), ('non-idle wait count', 0.5188710364742095 1), ('logons cumulative', 0.99999502850291333), ('user logons cumulative', 0.9999999693048103)]

User logons cumulative的相关系数达到了8个9

这个数据提供了后续分析的方向，详细查看AWR和监听日志后发现，有个应用平均每小时失败登录200万次。解决了这个问题后，SGA停止增长。问题得到规避。

彩蛋—总结




DBAplus
www.dbaplus.cn
THANK YOU