

云上的分布式数据库

DRDS原理与实践

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

大数据技术探索和价值发现



自我介绍

- 花名 沈询
- 新浪微博：搜 淘宝沈询
- 阿里分布式数据库 (DRDS/TDDL)、阿里分布式消息服务 (ONS) 负责人
- 参与过阿里集团大部分的Oracle到MySQL的迁移工作（商品、用户、交易、评价 etc..）
- 在分布式存储领域经验比较丰富



提纲

- DRDS 简介
- DRDS 云上的实践
- DRDS 功能特性与原理
- 小结



DRDS 简介

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

大数据技术探索和价值发现



DRDS简介

- DRDS是什么
 - 分布式数据库整体解决方案
- DRDS能做到什么
 - 尽可能保留数据库的优势 (RDBMS--)
 - MySQL兼容
 - 复杂SQL解决方案、分布式事务解决方案
 - 线性扩缩 (RDBMS++)
 - 读写分离
 - 动态伸缩 (解决读写瓶颈)
 - 业务在线正常运行



DRDS简介

- DRDS的应用场景

- 您是DRDS的用户

- 您的企业有高速增长预期
 - 您的业务的目标客户是70亿人
 - 您希望您的目标客户有优质的软件体验
 - 您希望寻找高容错的数据库解决方案
 - 您觉得NoSQL太麻烦

- 您不是DRDS的用户

- 您的企业已经不再增长
 - 您服务的目标客户在一千人以下
 - 您认为您的IT系统挂掉稀松平常，影响不大



DRDS简介

- 历史悠久
 - Cobar 的协议层（from 08年）
 - TDDL 的执行引擎和运维生态（from 08年）
- 为云打造
 - 更好的自主运维体验
 - 更强大的功能与兼容性
 - 与世界最新潮流并驾齐驱



DRDS 云上实践

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

大数据技术探索和价值发现



DRDS云上实践

- 内部使用情况
 - 内部几千应用使用
 - 至今6年，多次双11
 - 无p1故障
- 2014-12-4 正式公测
 - 五个月时间内
 - x000+公测申请
 - 正式运行的活跃应用超过x00
 - 保险（众安保险, etc..）
 - 互联网应用（为知笔记，虾米，华甫达, etc..）
 - 通信（阿里通信, etc..）
 - 金融（xx贷, etc..）
 - 医疗健康（阿里医药, etc..）
 - GIS（高德, etc..）
 - Etc ...



DRDS云上实践

- 分布式数据库核心功能产品化
 - 动态读写分离
 - 小表复制
 - 弹性扩容缩容
 - 监控、告警、性能调优



DRDS云上实践

- 易用性功能补强
 - MySQL协议兼容性提升
 - GUI工具兼容性提升
 - 复杂SQL和聚合兼容性
 - 智能下推
 - 分布式Join
- 系统稳定性提升
 - 网络结构大调整，解决Cobar不稳定问题
 - Cobar读写分离数据不一致问题改进



DRDS云上实践

- 外部故障
 - 4月x日
 - 运维系统变更导致P3 影响了10+客户不能正常登陆
- 后续action
 - 百倍赔付
 - 管理工具是最高优先级，尽可能减少变更
 - 鸡蛋放在多个篮子里，集群拆分力度进一步划小



DRDS 功能特性与原理

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

大数据技术探索和价值发现



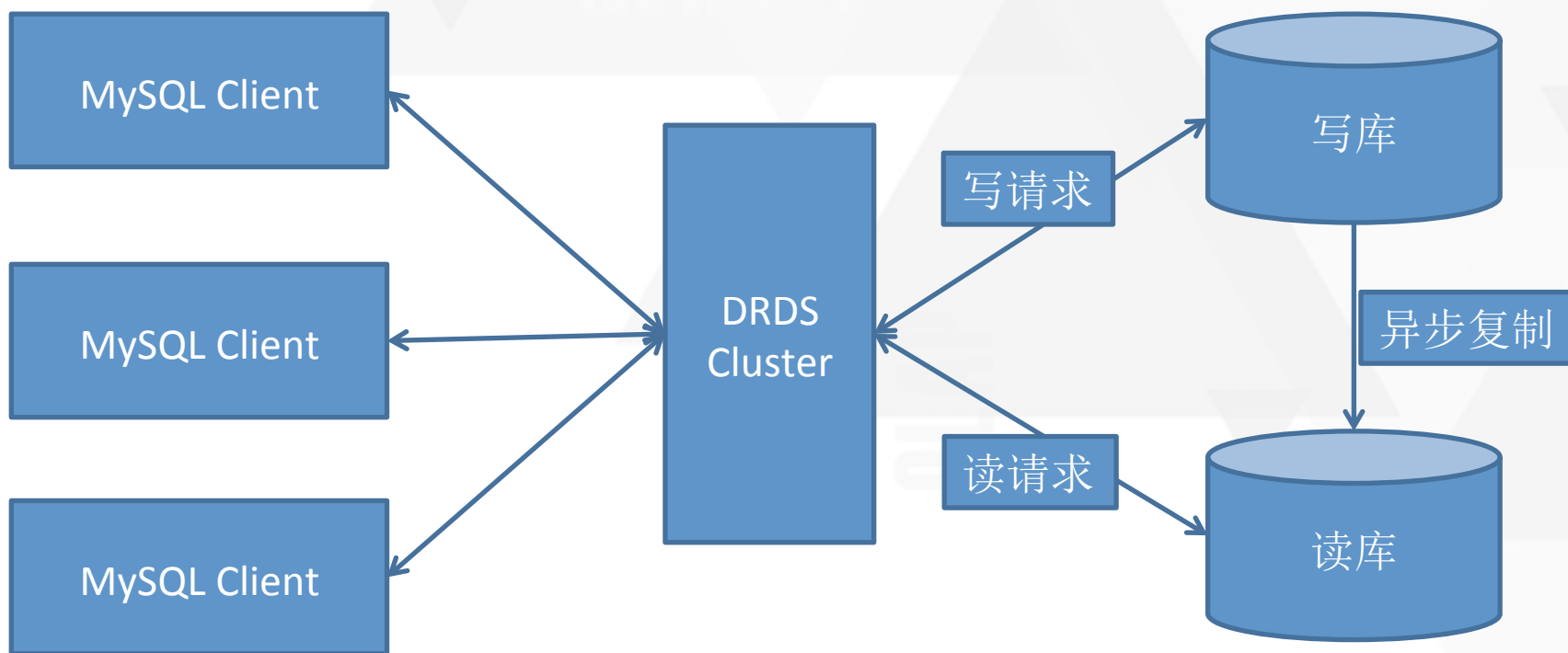
DRDS 功能特性与原理

- 社区版提供
 - 分库支持(cobar)
 - 读写分离(TDDL)
- DRDS云数据库版本
 - 在线动态读写分离
 - 分布式MySQL执行引擎
 - 小表异步广播
 - 弹性扩展
 - 事务方案套件



在线动态读写分离

- 按需“动态”读写分离



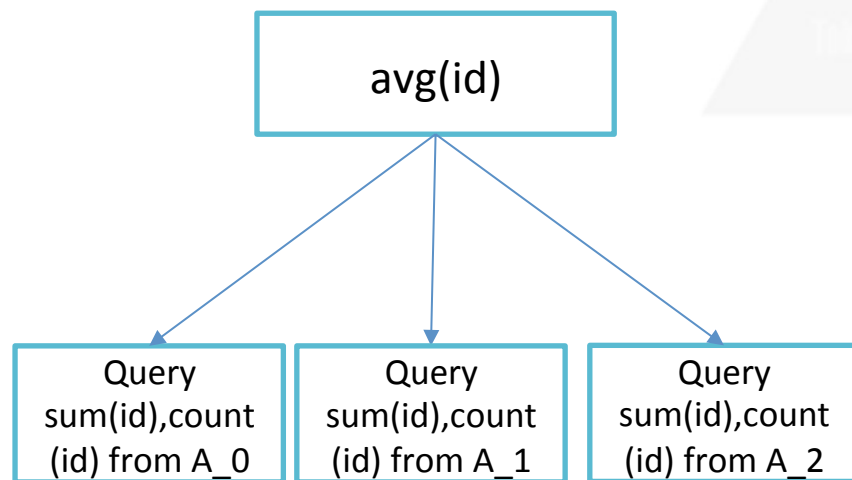
DRDS功能介绍-执行引擎

- 高兼容分布式SQL引擎
 - MySQL 5.5 的各类复杂查询
 - Join
 - 嵌套
 - 函数
- 智能下推
 - 减少网络传输
 - 减少计算量
 - 充分发挥下层存储的全部能力



DRDS功能介绍-执行引擎

- 智能下推
 - 表A 分库分表3个
 - `select avg(id) from A`



Merge
avg (id)
subQuery

Q1:select count(id),sum(id) A_0
Q2:select count(id),sum(id) A_1
Q3:select count(id),sum(id) A_2

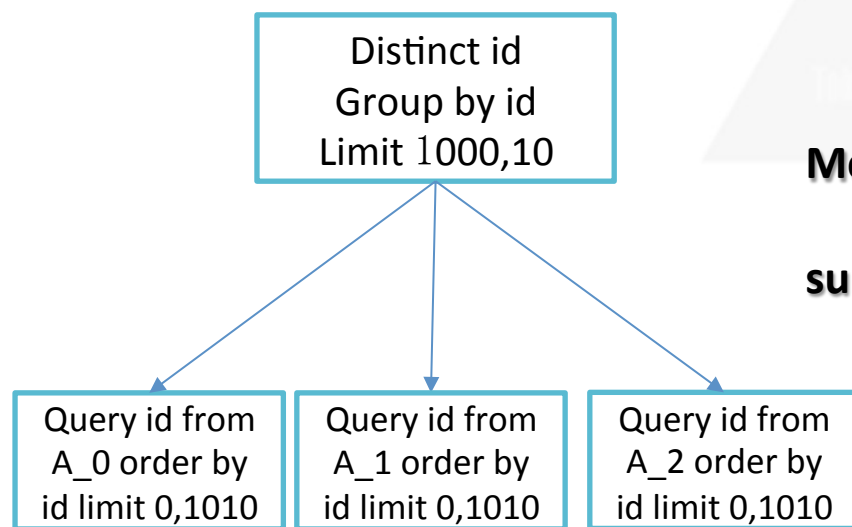


DRDS功能介绍-执行引擎

- 智能下推

- 全表distinct groupby的执行计划

- Select distinct(id) from A order by id limit 1000, 10



Merge

distinct id , group by id, Limit 1000,10

subQuery

Q1:select id from A_0 order by id limit 0,1010

Q2:select id from A_1 order by id limit 0,1010

Q2:select id from A_2 order by id limit 0,1010



DRDS功能介绍-小表异步广播

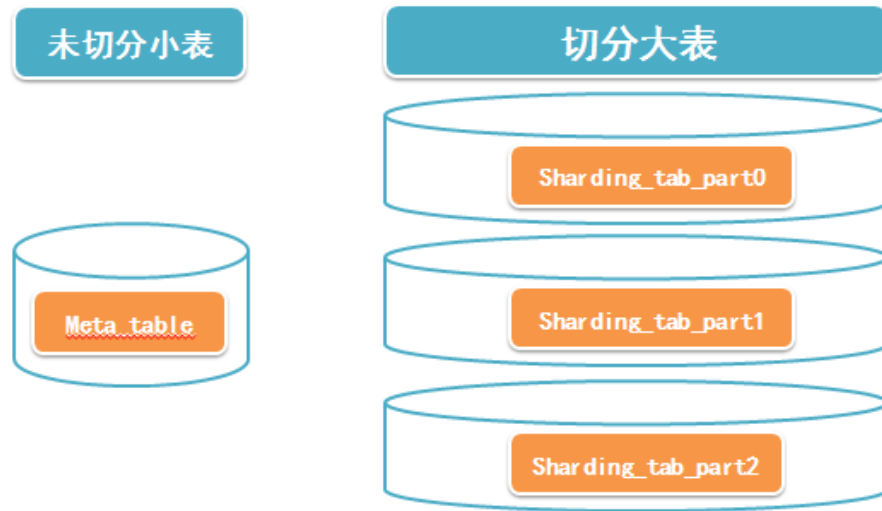
- 跨机JOIN

- 优势:

- 一致性
 - 空间比较节省

- 劣势

- 网络消耗
 - 延迟增加



DRDS功能介绍-小表异步广播

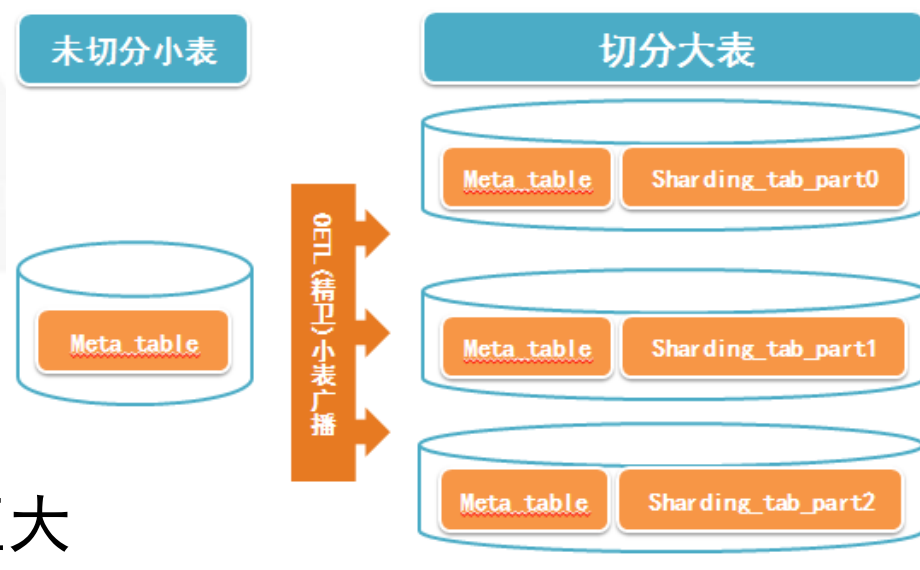
- 小表广播JOIN

- 优势

- 性能高
 - 延迟低
 - 网络消耗小

- 劣势

- 最终一致性
 - 小表更新量不能太巨大



DRDS 实践-分布式查询优化

- 让请求可以水平扩展
 - 原则1：选择的切分条件要能够让所有存储节点均衡的负载读写请求
 - 系统可以简单加机器来扩展
 - 没有系统瓶颈
 - 原则2：尽可能的让查询发生在一台物理数据库上，查询尽可能带上切分条件
 - 将跨网络请求尽可能减少
 - 减少并行查询时的机器消耗



DRDS 实践-分布式查询优化

- CASE1 :

- 应该选择哪个列作为切分条件？

- 按照买家ID的查询（买家查看自己买了哪些商品）

bizOrderID	buyerID	sellerID	content
0	0	1	床上用品
1	0	2	路上用品
2	0	3	销售路由器
3	0	4	中文书籍
4	0	5	电脑
5	1	0	ipad
6	2	0	笔记本
7	3	0	铅笔
8	4	0	桌面



DRDS 实践-分布式查询优化

- CASE2:

- 应该选择哪个列作为切分条件?

- 按照买家ID的查询
(买家查看自己买了哪些商品)
 - 按照卖家ID的查询
(卖家查看自己卖了哪些商品)

Table_bid
buyerID % 4

bizOrderID	buyerID	sellerID	content
0	0	1	床上用品
1	0	2	路上用品
2	0	3	销售路由器
3	0	4	中文书籍
4	0	5	电脑
8	4	0	桌面

bizOrderID	buyerID	sellerID	content
5	1	0	ipad

bizOrderID	buyerID	sellerID	content
6	2	0	笔记本

bizOrderID	buyerID	sellerID	content
7	3	0	铅笔



DRDS 实践-分布式查询优化

- 异构复制

Table_bid
buyerID % 4

bizOrderID	buyerID	sellerID	content
0	0	1	床上用品
1	0	2	路上用品
2	0	3	销售路由器
3	0	4	中文书籍
4	0	5	电脑
8	4	0	桌面

bizOrderID	buyerID	sellerID	content
5	1	0	ipad

bizOrderID	buyerID	sellerID	content
6	2	0	笔记本

bizOrderID	buyerID	sellerID	content
7	3	0	铅笔

异构
复制

Table_sid
sellerID % 4

bizOrderID	buyerID	sellerID	content
5	1	0	ipad
6	2	0	笔记本
7	3	0	铅笔
8	4	0	桌面
3	0	4	中文书籍

bizOrderID	buyerID	sellerID	content
0	0	1	床上用品
4	0	5	电脑

bizOrderID	buyerID	sellerID	content
1	0	2	路上用品

bizOrderID	buyerID	sellerID	content
2	0	3	销售路由器



DRDS 实践-分布式查询优化

- CASE3:
 - 卖家在商城销售的所有商品

type	平台名
0	商城
1	专卖店

Table_bid
buyerID % 4

bizOrderID	buyerID	sellerID	type	content
0	0	1	0	床上用品
1	0	2	1	路上用品
2	0	3	0	销售路由器
3	0	4	1	中文书籍
4	0	5	0	电脑
8	4	0	0	桌面

bizOrderID	buyerID	sellerID	type	content
5	1	0	1	ipad

bizOrderID	buyerID	sellerID	type	content
6	2	0	0	笔记本

bizOrderID	buyerID	sellerID	type	content
7	3	0	1	铅笔



DRDS 实践-分布式查询优化

- 小表异步广播

Table_bid
buyerID % 4

type	平台名
0	商城
1	专卖店

bizOrderID	buyerID	sellerID	type	content
0	0	1	0	床上用品
1	0	2	1	路上用品
2	0	3	0	销售路由器
3	0	4	1	中文书籍
4	0	5	0	电脑
8	4	0	0	桌面

type	平台名
0	商城
1	专卖店

bizOrderID	buyerID	sellerID	type	content
5	1	0	1	ipad

type	平台名
0	商城
1	专卖店

bizOrderID	buyerID	sellerID	type	content
6	2	0	0	笔记本

type	平台名
0	商城
1	专卖店

bizOrderID	buyerID	sellerID	type	content
7	3	0	1	铅笔



DRDS 实践-分布式查询优化

- CASE4:

- 应该选择哪个列作为切分条件？

- 最近1周内所有卖家销售的商品量？

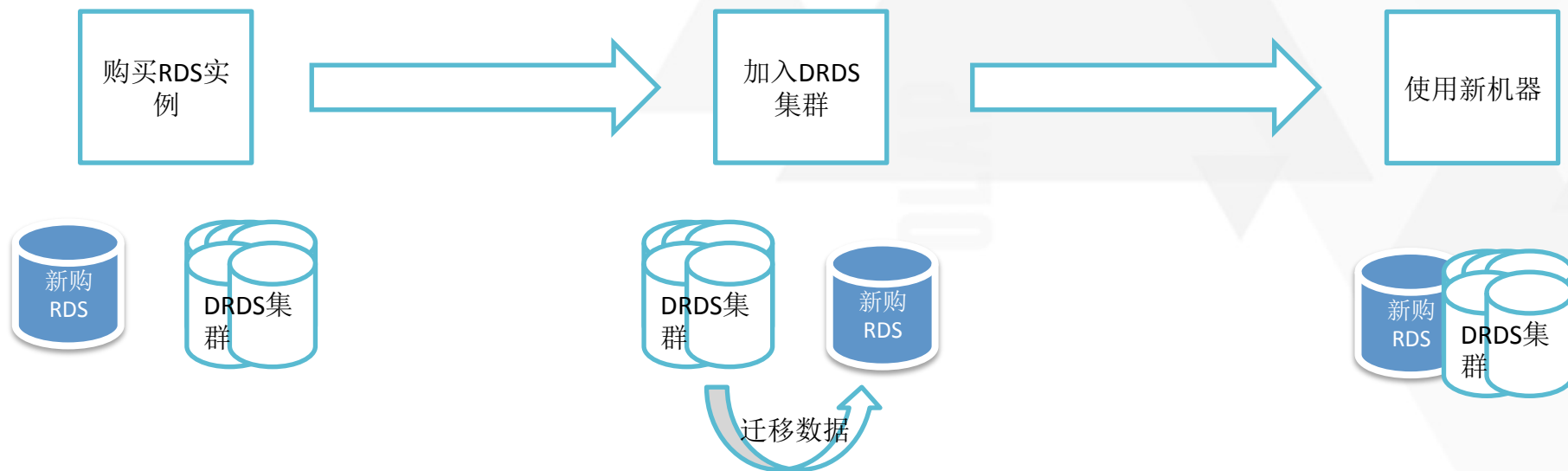
bizOrderID	buyerID	sellerID	content	GMT_MODIFIED
0	0	1	床上用品	2014-09-01
1	0	2	路上用品	2014-09-01
2	0	3	销售路由器	2014-09-01
3	0	4	中文书籍	2014-09-01
4	0	5	电脑	2014-09-02
5	1	0	ipad	2014-09-02
6	2	0	笔记本	2014-09-04
7	3	0	铅笔	2014-09-03
8	4	0	桌面	2014-09-05



DRDS功能介绍-弹性扩展

- 自动扩容、缩容

平滑扩容				
连接串	端口	读流量份额(%)	数据库类型	操作
rds3qm3eamubniq.mysql.rds.aliyuncs...	3306	100	MySQL	添加只读实例



DRDS 实践-事务的分布式优化

- 目标：

- 完整的事务支持

- 像传统单机事务一样的操作方式
 - 可按需无限扩展

- 快醒醒~~别做梦了

- 容易理解的模型往往性能都不好，性能好的模型往往不容易理解

这就是生活

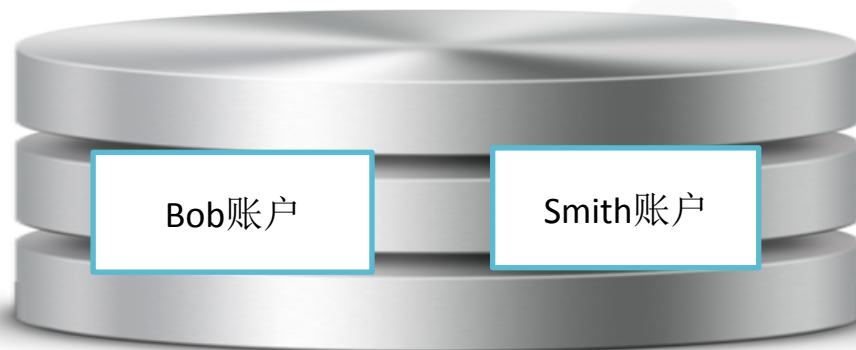


DRDS实践 - 事务的分布式优化

事务单元

操作指令	耗时	总耗时
锁定Bob账户	0.001ms	5.004ms
锁定Smith账户	0.001ms	
查看Bob是否有100元	1ms	
从Bob账号中减少100元	2ms	
给Smith账户中增加100元	2ms	
解锁Bob账户	0.001ms	
解锁Smith账户	0.001ms	

事务时间序



DRDS实践 - 事务的分布式优化

操作指令	耗时	总耗时
锁定Bob账户	0.001ms	11.004ms
通过网络锁定Smith账户	2ms+0.001ms	
查看Bob是否有100元	1ms	
从Bob账号中减少100元	2ms	
通过网络给Smith账户中增加100元	2ms+2ms	
解锁Bob账户	0.001ms	
通过网络解锁Smith账户	2ms+0.001ms	

事务时间序



DRDS实践 - 事务的分布式优化

事务单元

操作指令	耗时
锁定Bob账户	0.001ms
查看Bob是否有100元	1ms
从Bob账号中减少100元	2ms
解锁Bob账户	0.001ms

异步事务单元

操作指令	耗时
锁定Smith账户	0.001ms
给Smith账户中增加100元	2ms
解锁Smith账户	0.001ms

异步并行消息

Bob账户

Smith账户

事务时间序



DRDS实践 - 事务的分布式优化



IT168... ChinaUnix ITpub

THANKS