构建生态化

分布式数据库架构体系

Alibaba.com

About 自我介绍

陶勇

brave.taoy@alibaba-inc.com

Twitter & Sina: @bravetao

Index 内容概要

- 分布式数据库解决方案 @ Alibaba B2B
  - 分布式数据存储与访问
    - Cobar
  - 准实时增量数据获取与消费
    - Erosa/Eromanga
  - 多维度数据同步与网站镜像
    - Otter
- 构建分布式数据库生态架构 @ Alibaba B2B
  - 全站数据架构
  - 思考与展望

Keywords 关键词

MySQL protocol

Schema垂直拆分

Table水平拆分

Global Failover

实时日志解析

Global ID

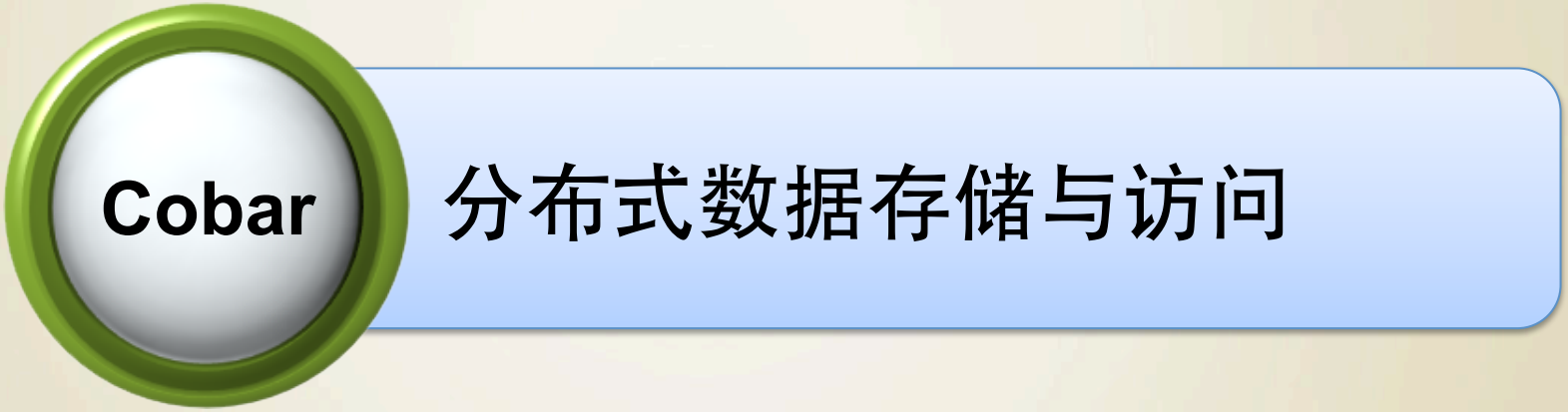
事务顺序

实时镜像

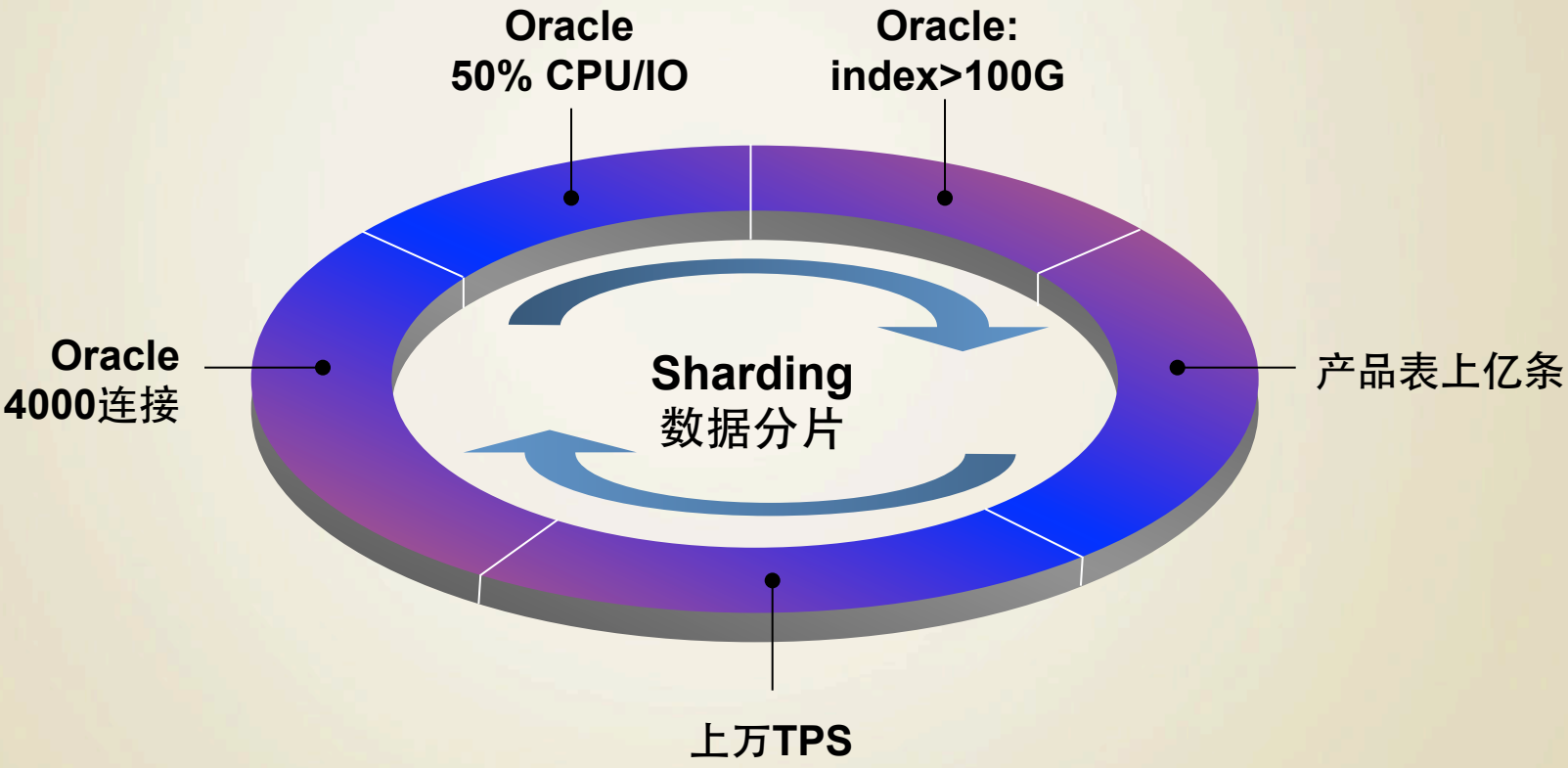
双向同步

同步事务支持

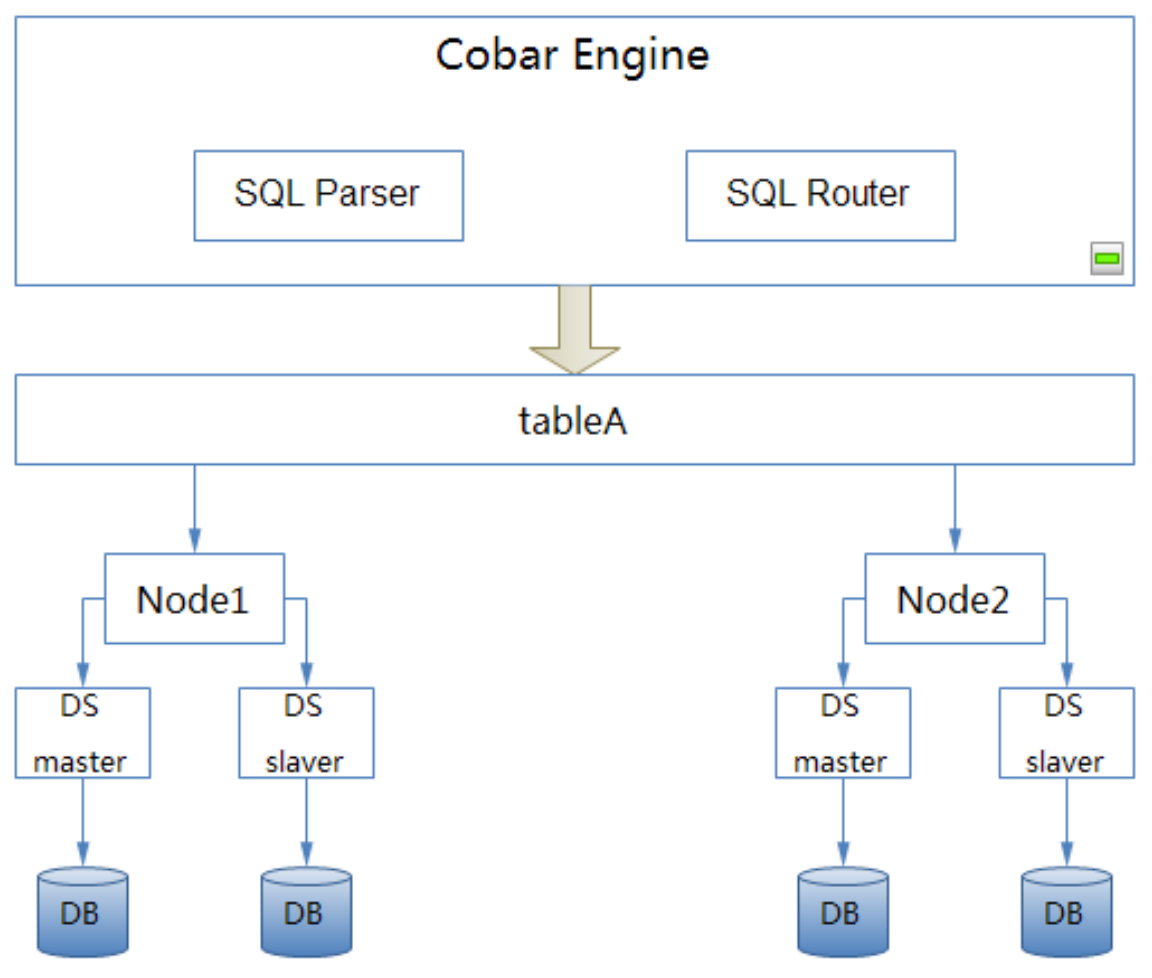
分布式数据库解决方案



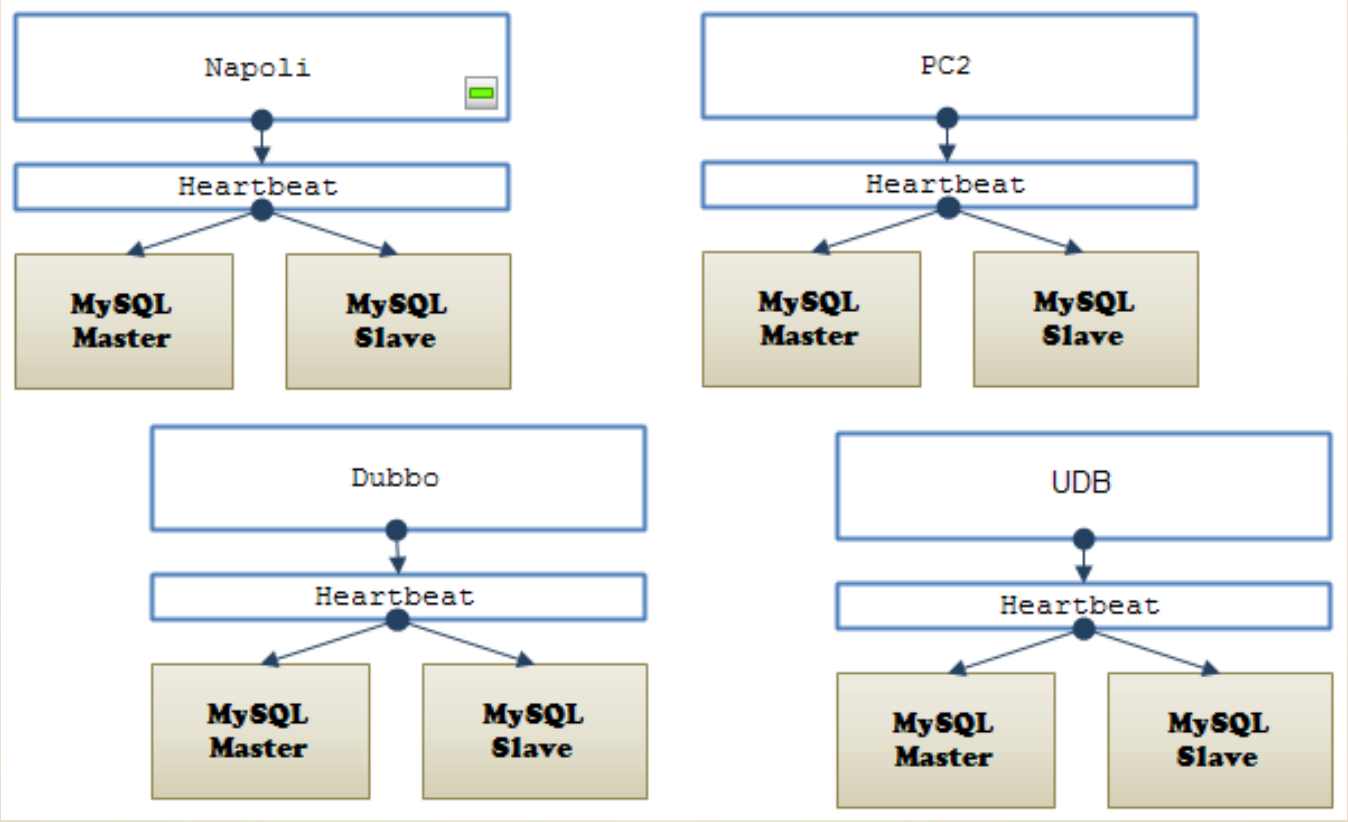
遇到的问题 - 2008


Oracle
50% CPU/IO
Oracle:
index>100G
Oracle
4000连接
Sharding
数据分片
产品表上亿条
上万TPS

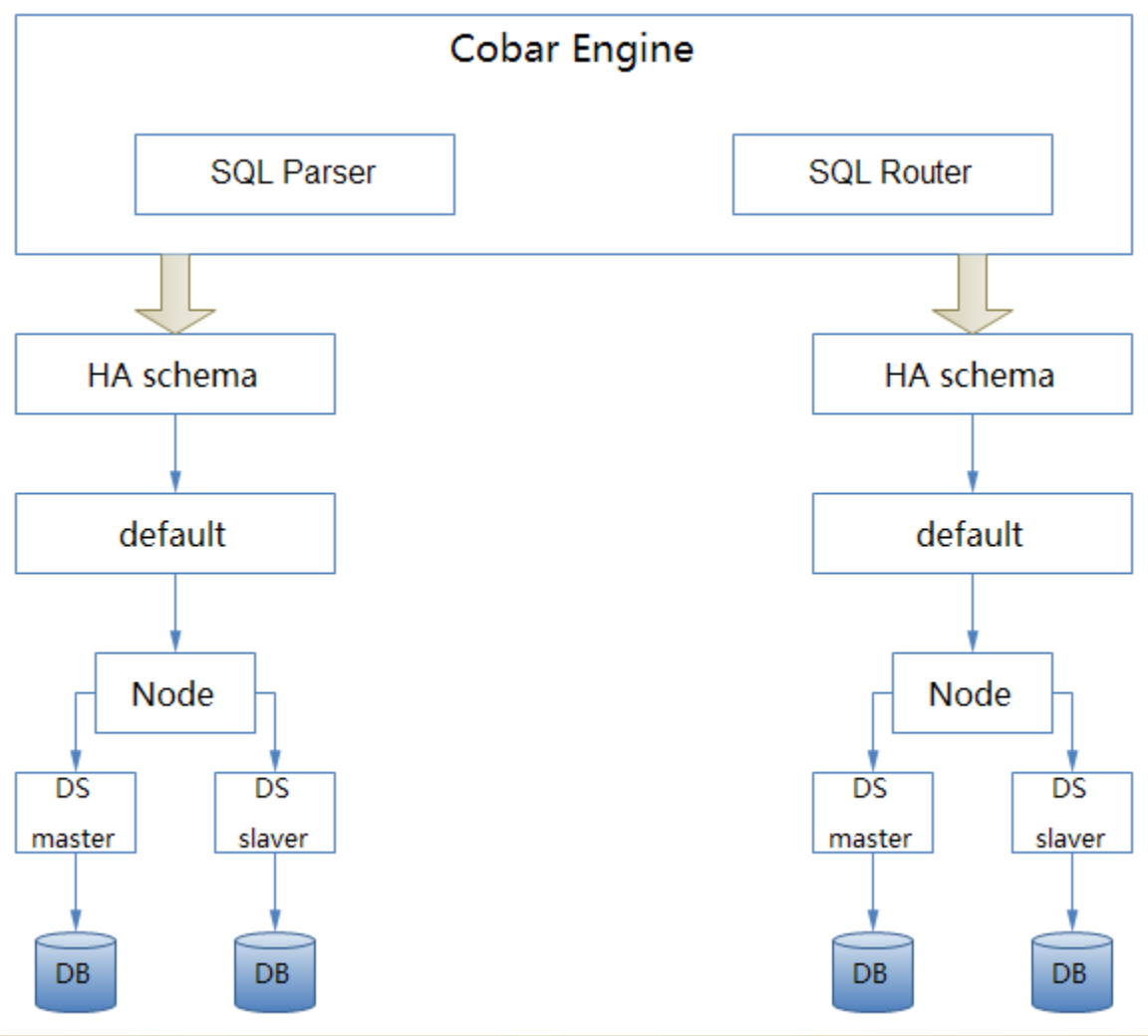
v0.6-1.0.x (08-10)


Cobar Engine
SQL Parser
SQL Router
tableA
Node1
Node2
DS
master
DS
slaver
DS
master
DS
slaver
DB
DB
DB
DB

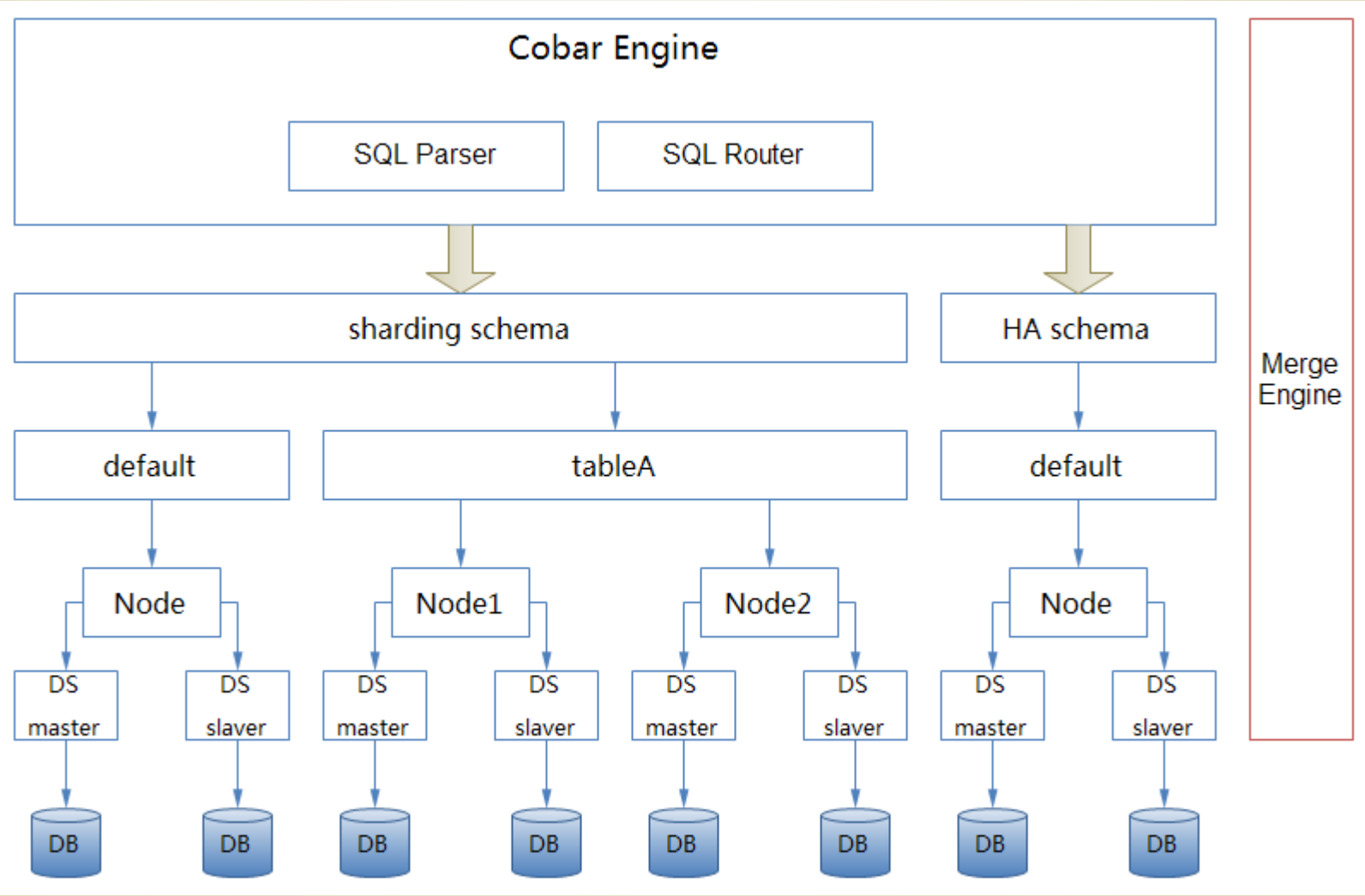
遇到的问题 - 2010


Napoli
Heartbeat
MySQL Master
MySQL Slave
PC2
Heartbeat
MySQL Master
MySQL Slave
Dubbo
Heartbeat
MySQL Master
MySQL Slave
UDB
Heartbeat
MySQL Master
MySQL Slave

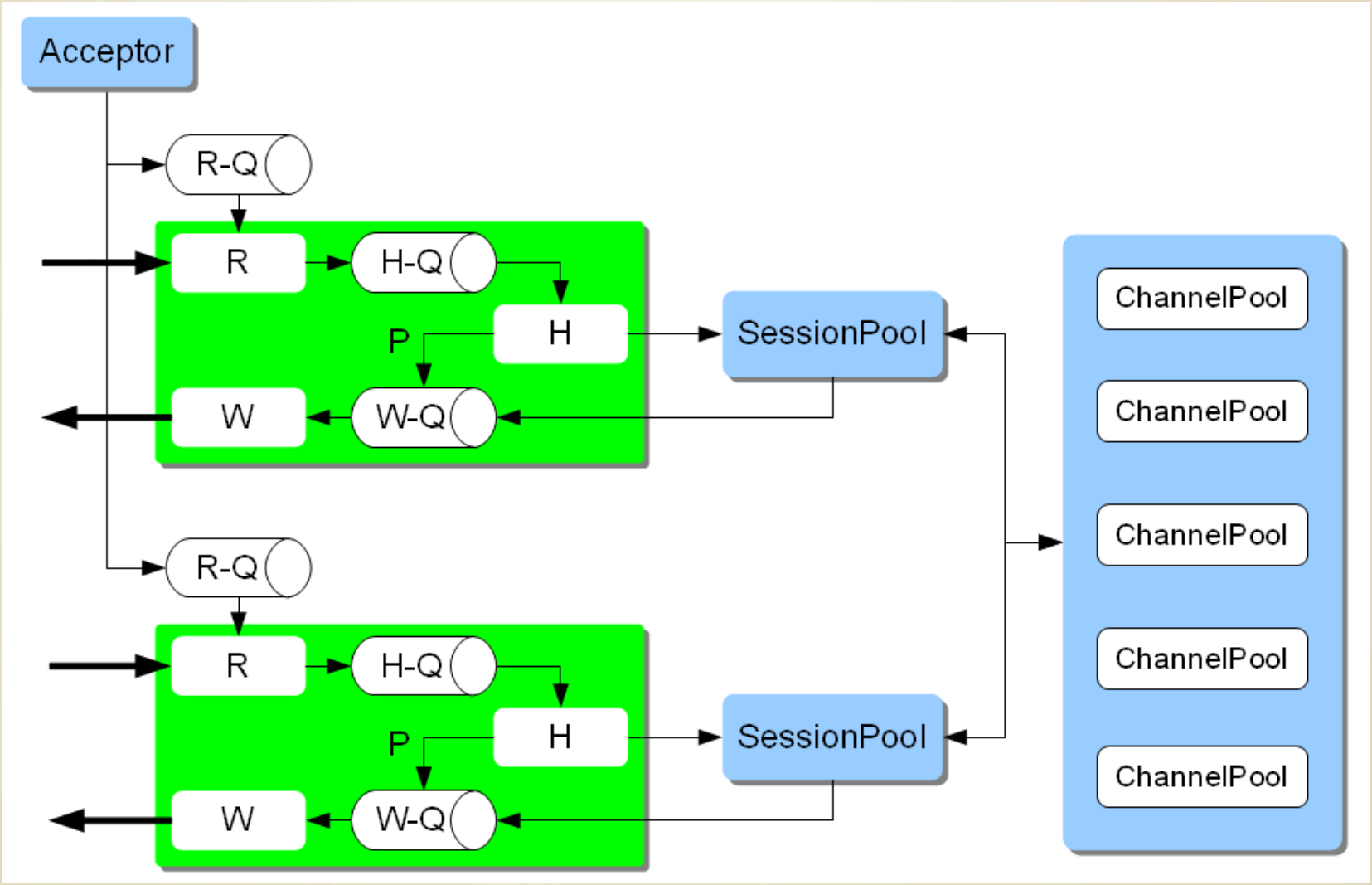
v1.1.x (10-11)


Cobar Engine
SQL Parser
SQL Router
HA schema
HA schema
default
default
Node
Node
DS
master
DS
slaver
DS
master
DS
slaver
DB
DB
DB
DB

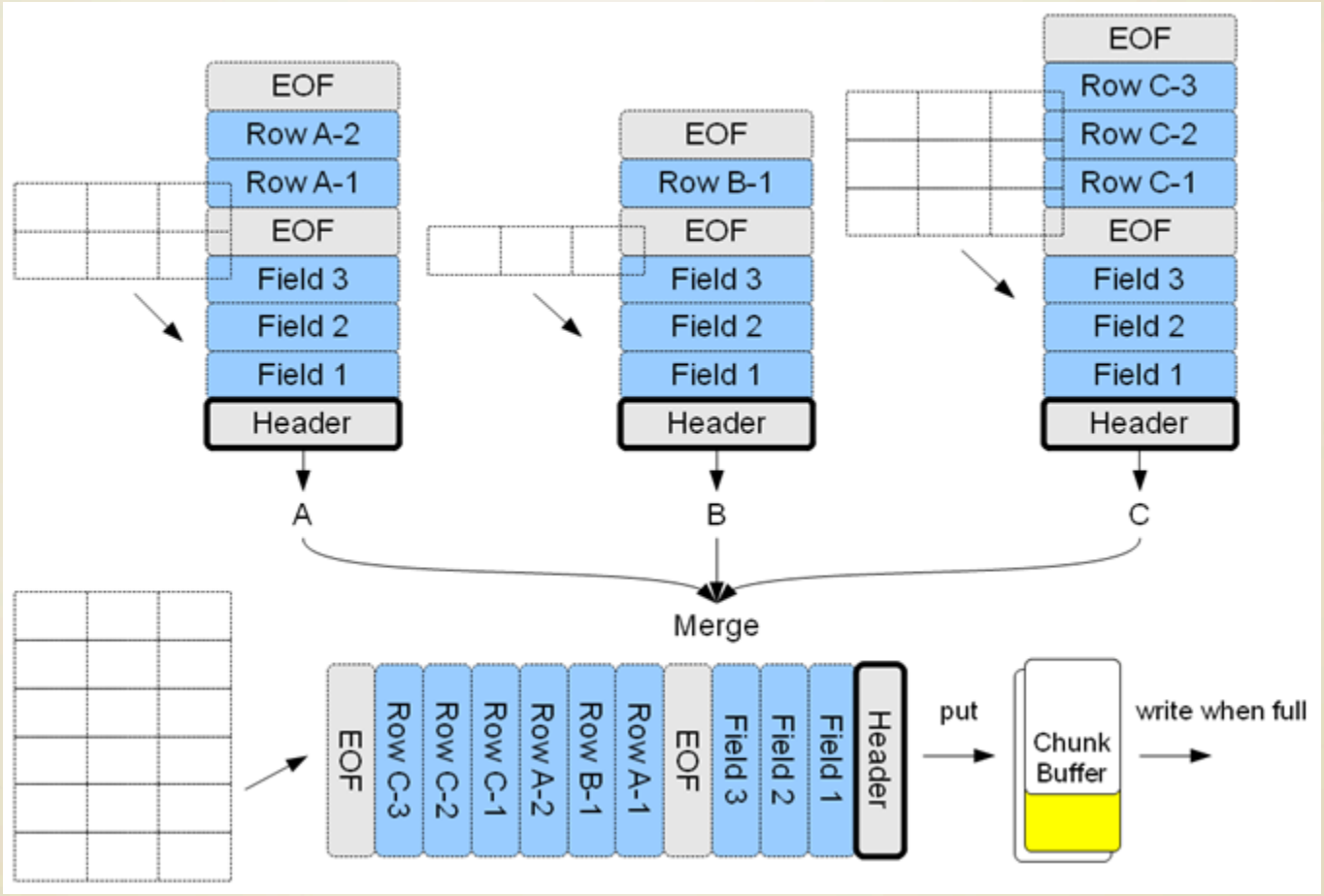
现在 (10-2012.12.23)



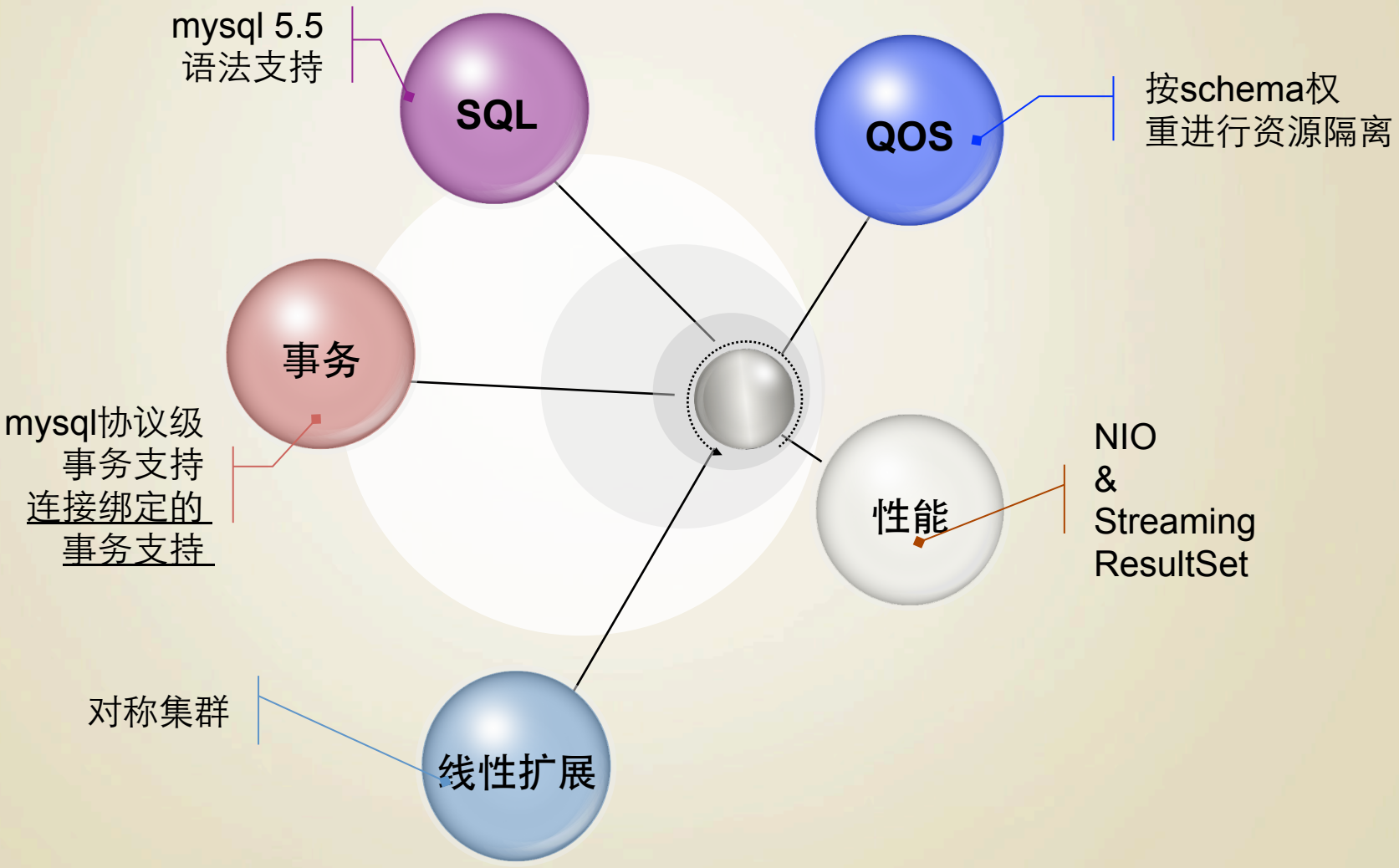
细节：线程复用模型


Acceptor
R-Q
R
H-Q
H
P
W
W-Q
SessionPool
R-Q
R
H-Q
H
P
W
W-Q
SessionPool
ChannelPool
ChannelPool
ChannelPool
ChannelPool
ChannelPool

细节：事务、ResultSet



特性


mysql 5.5
语法支持
SQL
QOS
按schema权
重进行资源隔离
事务
mysql协议级
事务支持
连接绑定的
事务支持
性能
NIO
&
Streaming
ResultSet
对称集群
线性扩展

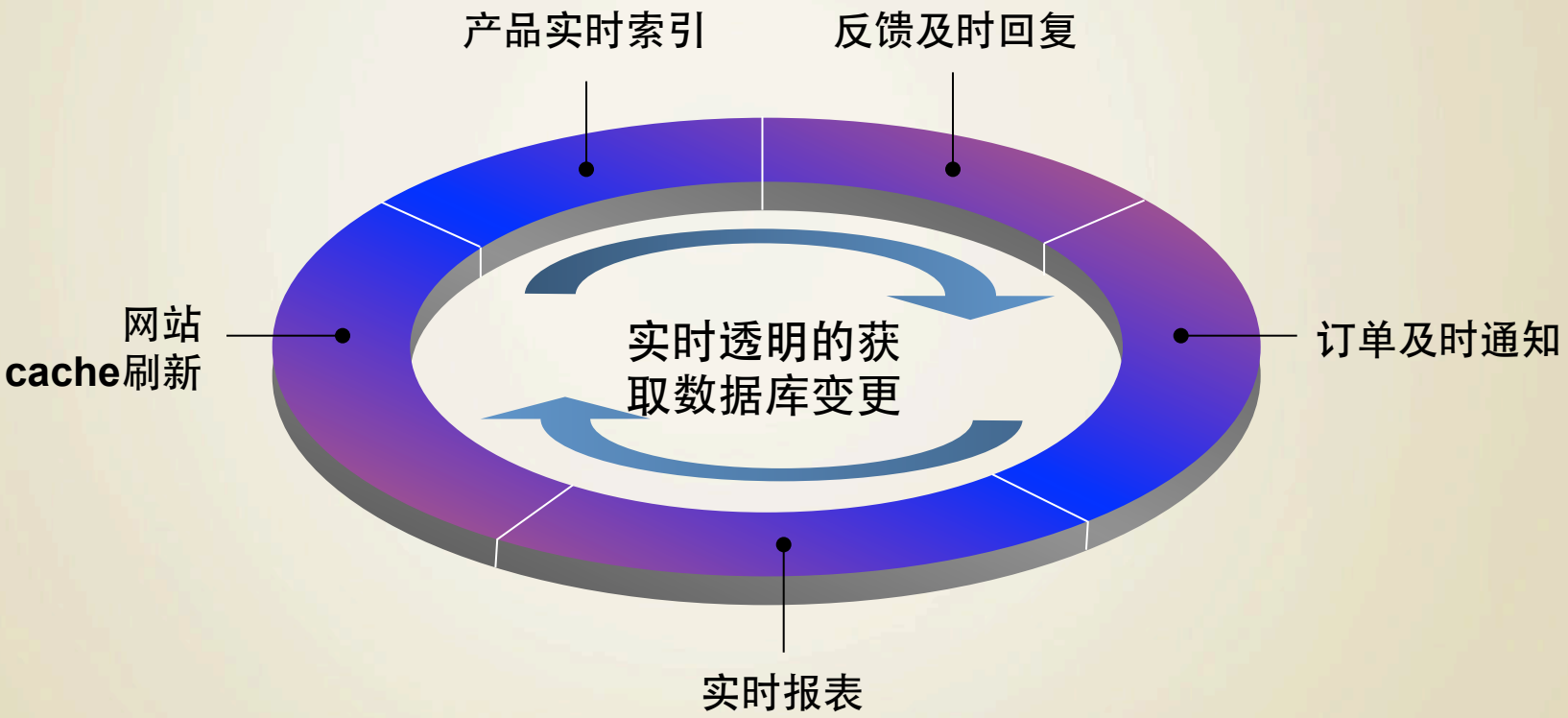
一些数据

- 20+ Cobar
- 线上高峰期 4-5w TPS，单机性能 10w TPS
- 15并发以上，整体性能超过直接访问mysql
- 支持mysql 5.5 全部DML和部分DDL语法

分布式数据库解决方案

准实时增量数据获取与消费

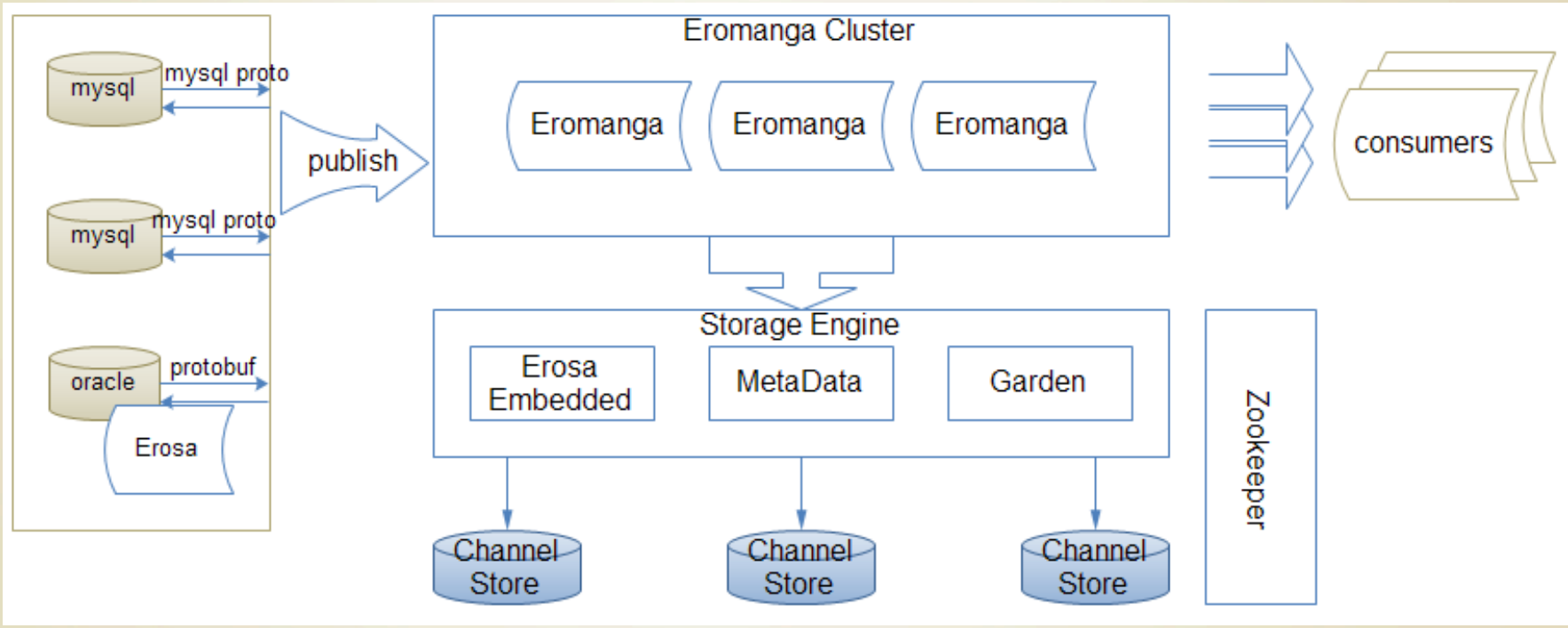
遇到的问题



准实时增量数据获取与消费

- 以前的做法
  - DB Trigger
  - Dump table
  - Application MQ
- 问题
  - 运维困难
  - 数据库、网络瞬时压力大
  - 业务侵入性强

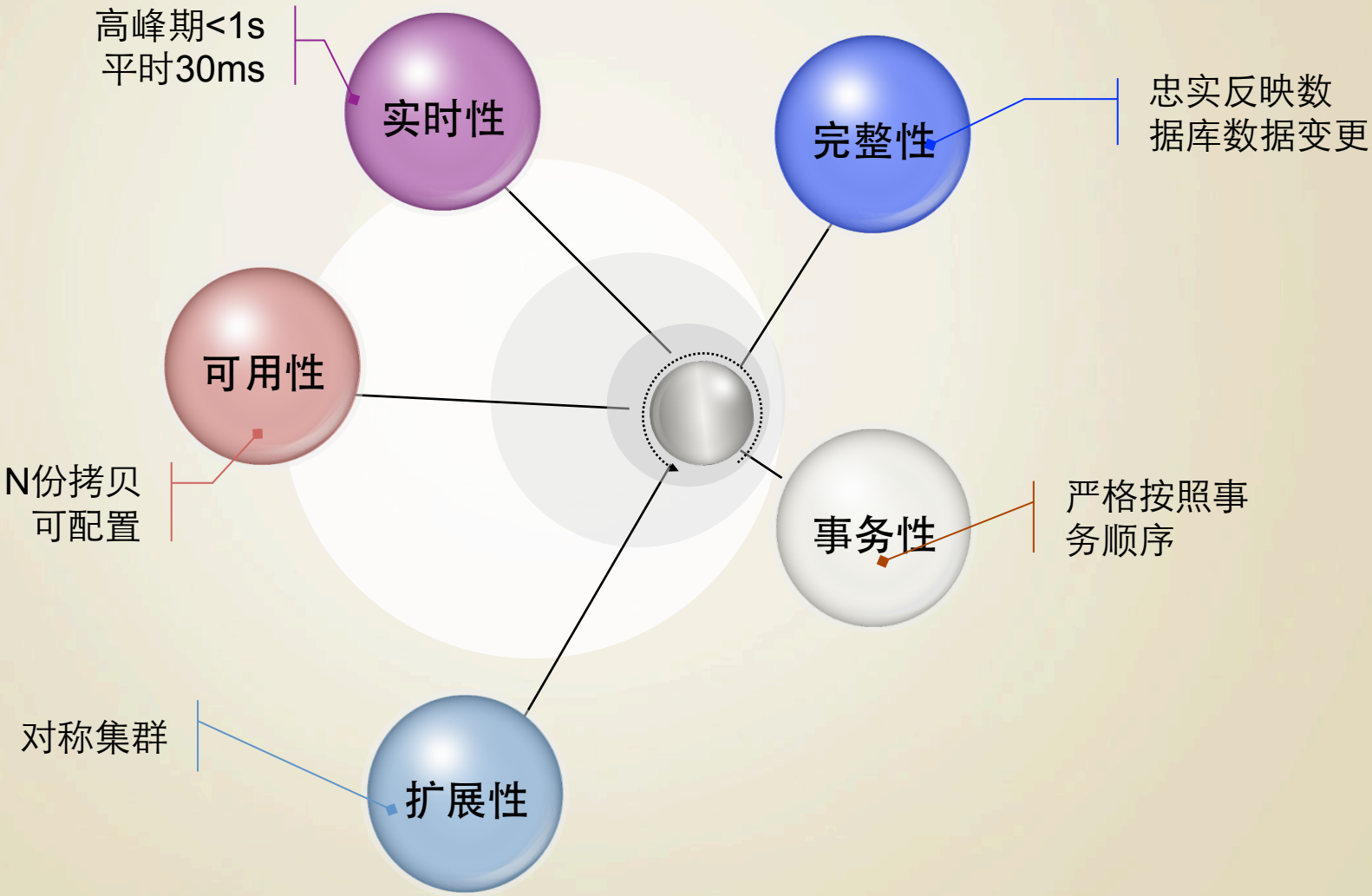
整体架构


mysql
mysql proto
mysql
mysql proto
oracle
protobuf
Erosa
publish
Eromanga Cluster
Eromanga
Eromanga
Eromanga
consumers
Storage Engine
Erosa Embedded
MetaData
Garden
Zookeeper
Channel Store
Channel Store
Channel Store

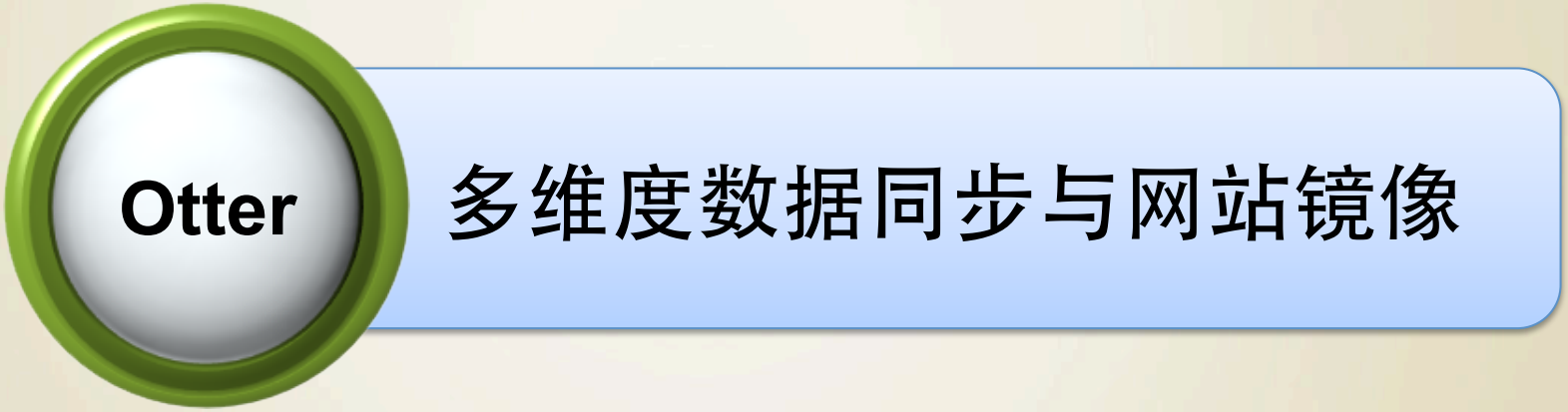
细节

- 数据解析
  - oracle : redolog parser
    - Archive log ~ 2min
    - Online log < 10ms
    - Open column supplemental log
  - mysql : binlog parser
    - COM_BINLOG_DUMP
- 数据消费
  - 对称集群
  - Data cursor : ZooKeeper

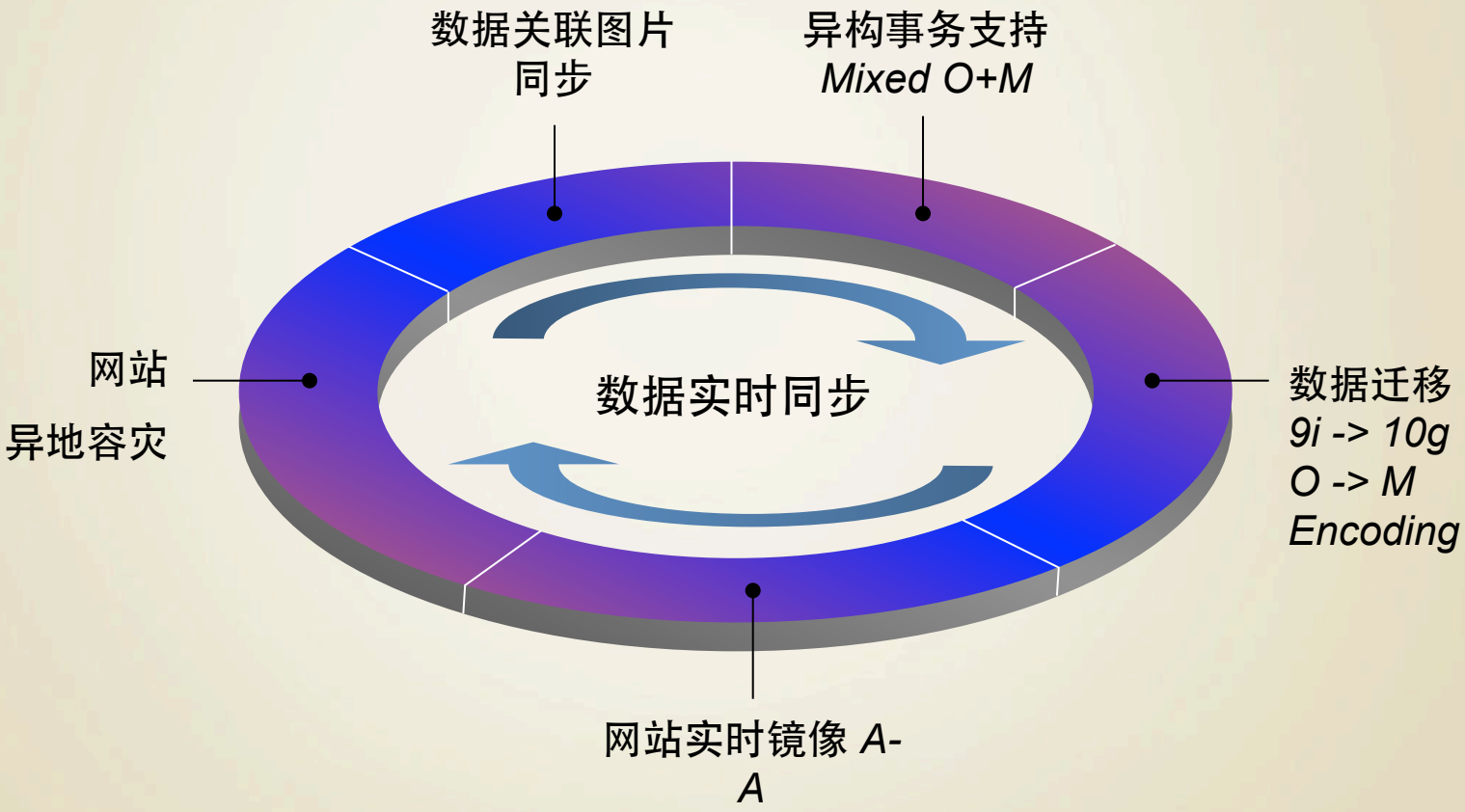
特性


高峰期<1s
平时30ms
实时性
完整性
忠实反映数
据库数据变更
可用性
N份拷贝
可配置
事务性
严格按照事
务顺序
对称集群
扩展性

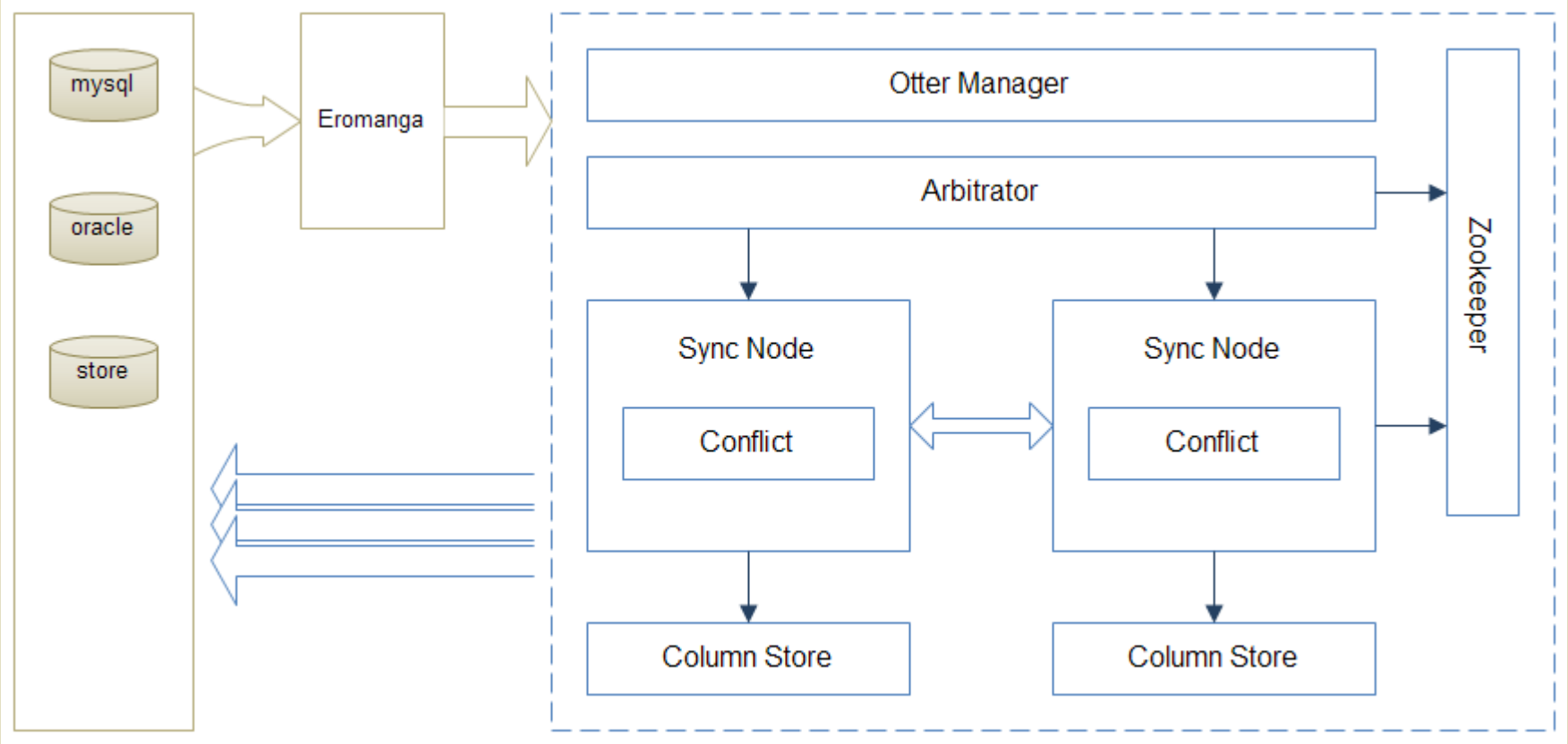
分布式数据库解决方案



遇到的问题


数据关联图片
同步
异构事务支持
Mixed O+M
网站
异地容灾
数据实时同步
数据迁移
9i -> 10g
O -> M
Encoding
网站实时镜像 A-
A

整体架构


mysql
oracle
store
Eromanga
Otter Manager
Arbitrator
Zookeeper
Sync Node
Conflict
Sync Node
Conflict
Column Store
Column Store

细节

- 数据同步
  - 自定义字段过滤
  - 自定义文件同步逻辑
  - 按事务表并发加载
  - 按表PK hash并发加载

- 数据冲突
  - 实时字段级冲突合并
  - 冲突判断及解决

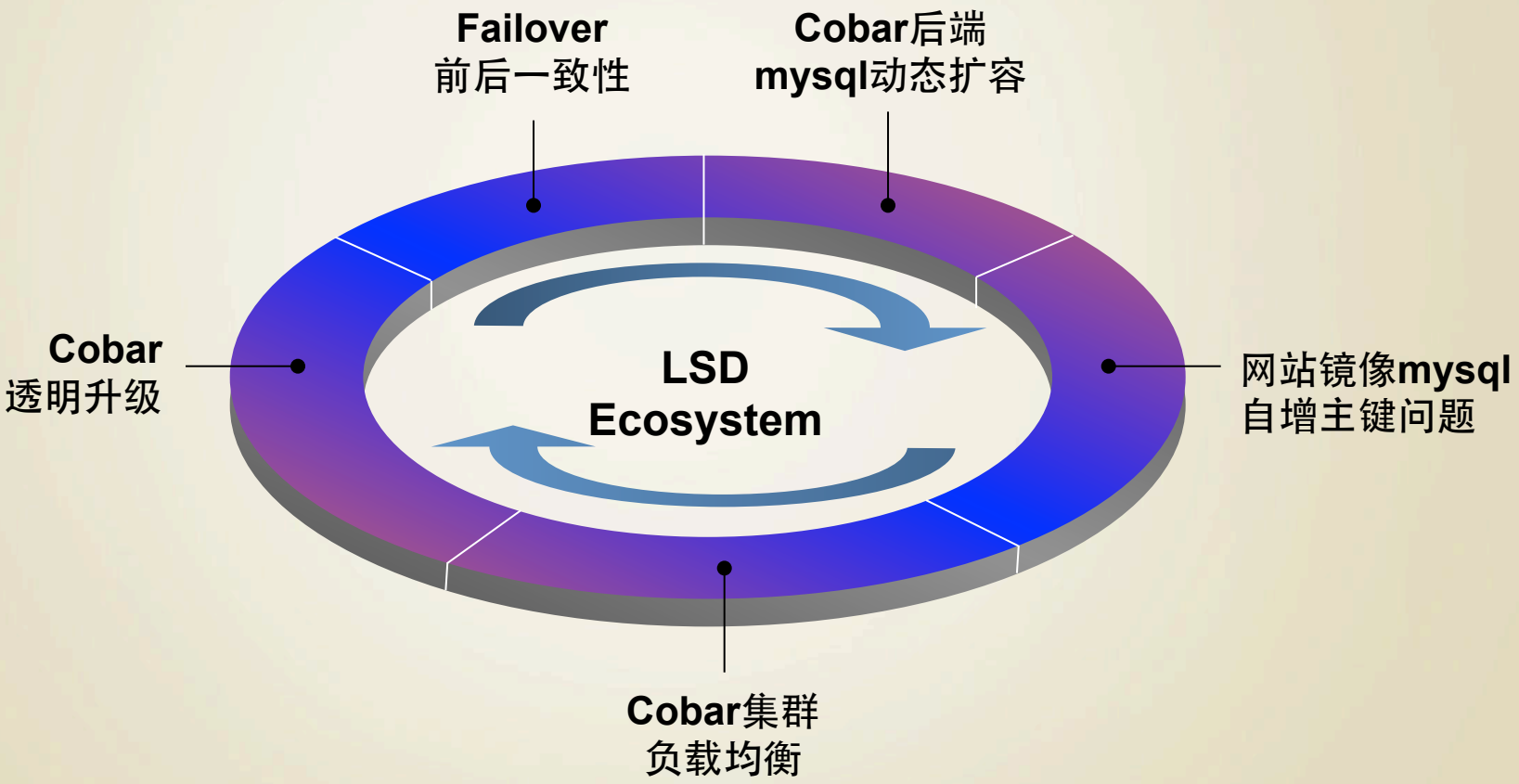
现有的应用场景

- 数据库
  - 备份：多master -> 单slaver
  - 异构迁移、跨版本迁移
  - *Oracle Active-Active*
- 网站容灾
  - 容灾备份
  - 读写分离
- 网站镜像
  - 双向读写
  - 按字段同步
  - 按事务并发同步

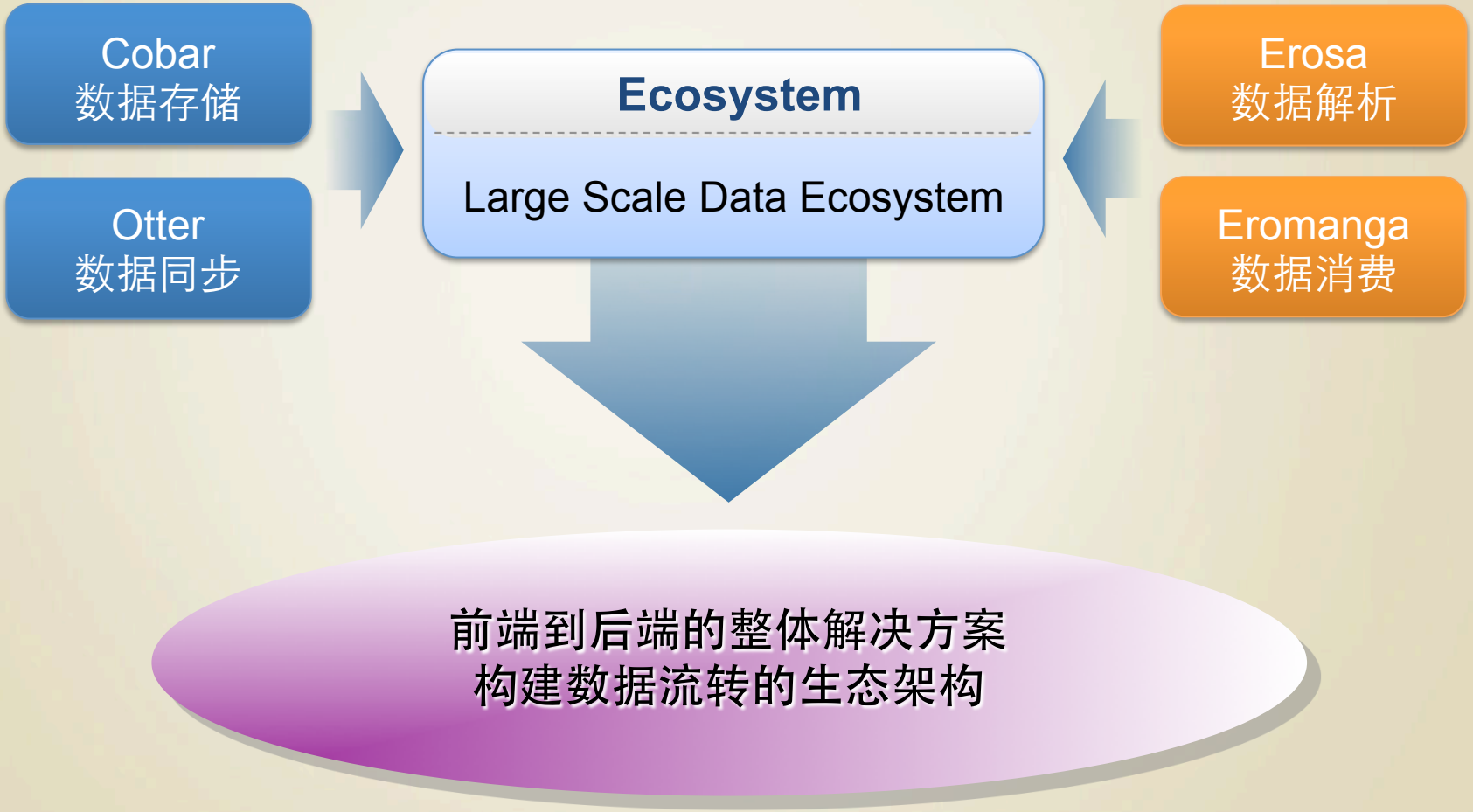
分布式数据库解决方案

构建分布式数据库生态架构

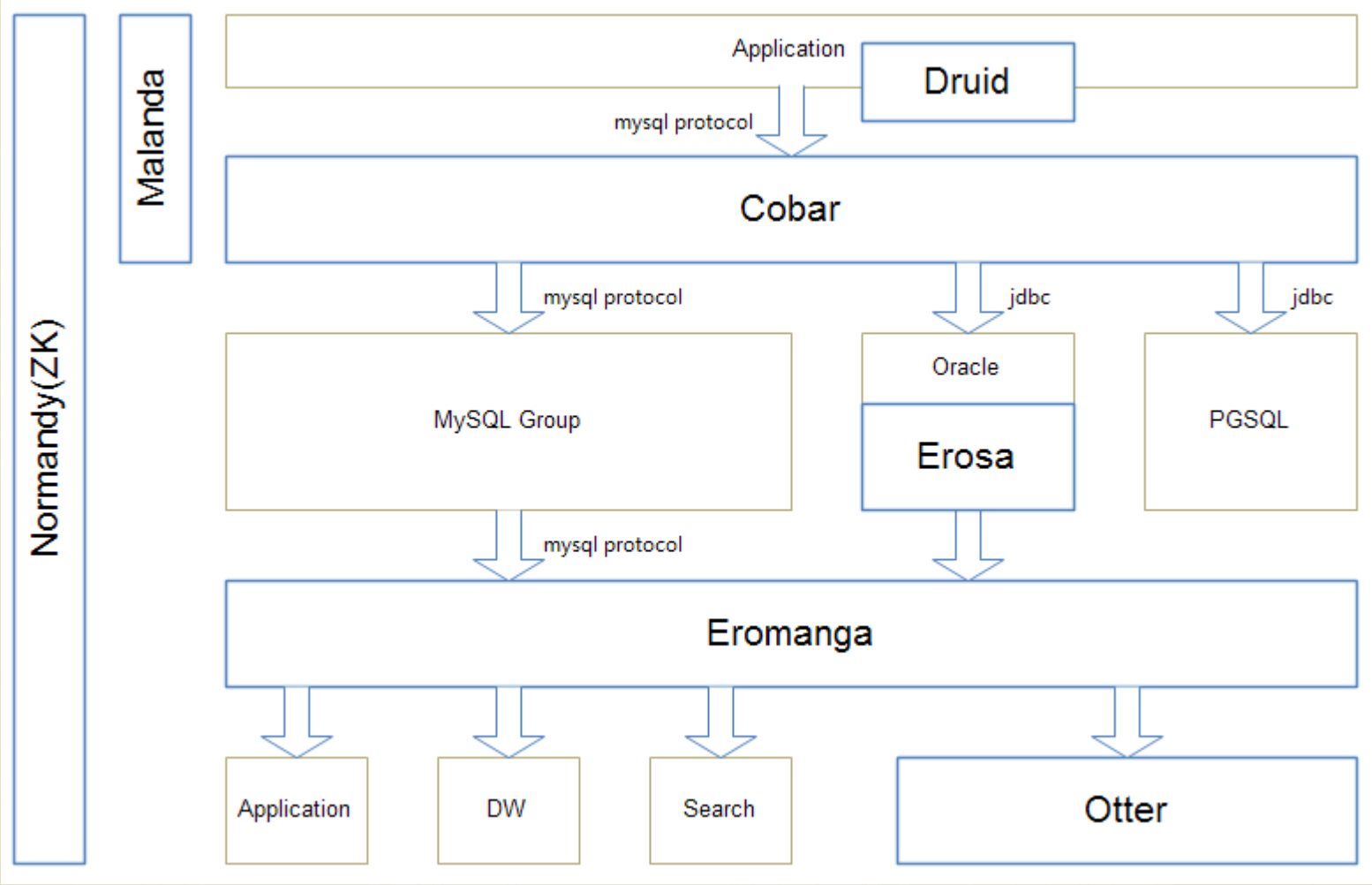
遇到的问题


Failover
前后一致性
Cobar后端
mysql动态扩容
Cobar
透明升级
LSD
Ecosystem
网站镜像mysql
自增主键问题
Cobar集群
负载均衡

分布式数据库架构生态体系



整体架构


Normandy(ZK)
Malanda
Application
Druid
mysql protocol
Cobar
mysql protocol
jdbc
jdbc
MySQL Group
Oracle
Erosa
PGSQL
mysql protocol
Eromanga
Application
DW
Search
Otter

思考与展望

- 产品架构闭环
  - Under constructing
    - OLTP & OLAP
    - NoSQL数据库
- 软硬结合
  - 单机多实例
  - 单机高性能

MySQL优化

单机多实例

- 存储技术飞速发展，IO不再是瓶颈
- MySQL对多核CPU利用率低

单机高性能

- RAID：BBWC/Fastpath
- Fusionio
- Flashcache
- Semi-Sync

Q&A

Thanks!

No best, only the most suitable


QCon
北京站·2012年4月18~20日
www.qconbeijing.com（11月启动）
QCon杭州站官网和资料
www.qconhangzhou.com
全球企业开发大会
INTERNATIONAL
SOFTWARE DEVELOPMENT
CONFERENCE