SequoiaDB
巨杉数据库

云架构下的分布式数据库设计与实践

巨杉数据库 许建辉

SequoiaDB巨杉数据库简介

Gartner

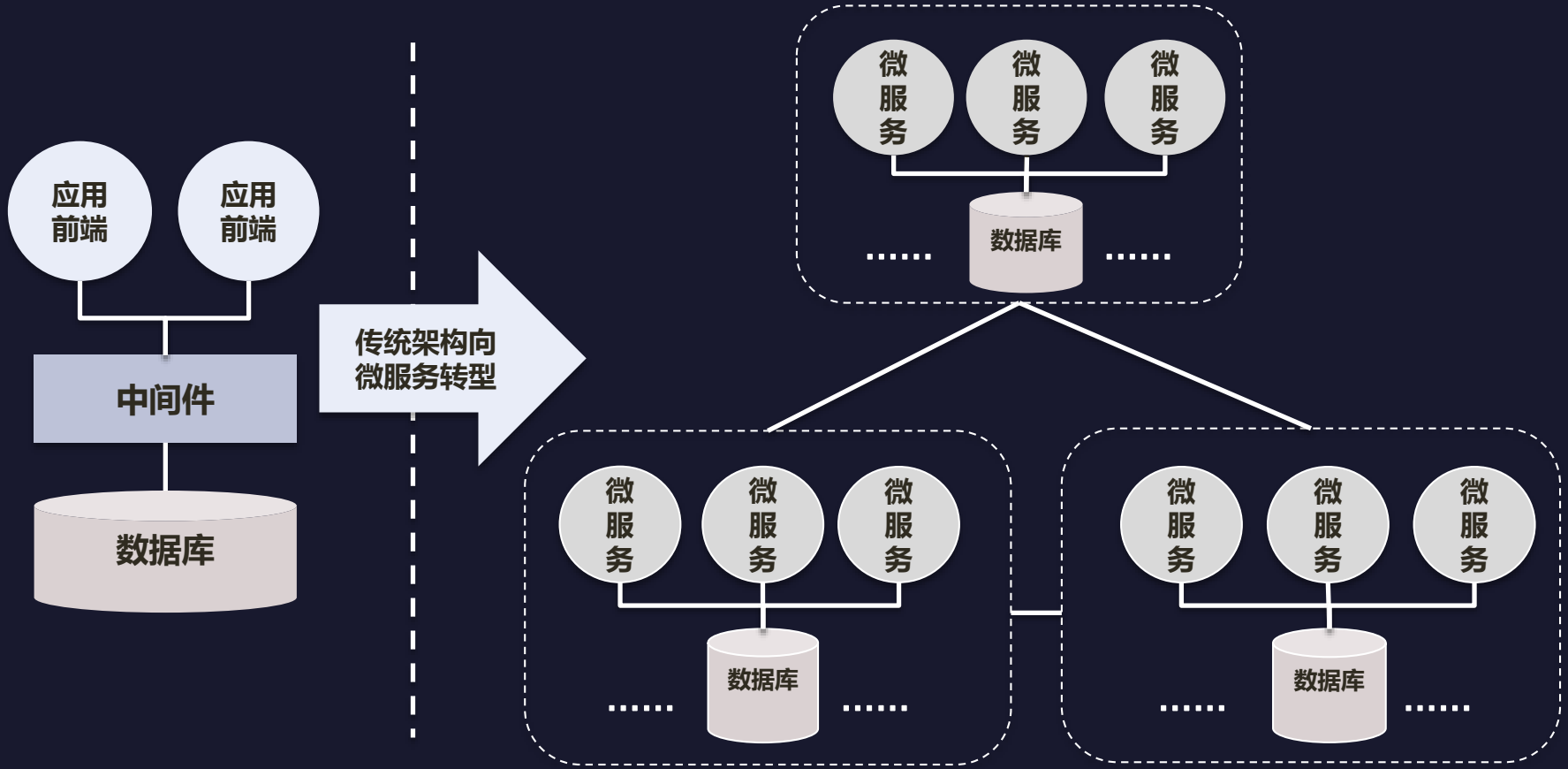
首款入选Gartner数据库推荐报告的国产分布式数据库产品

超过100家大型金融企业核心业务系统上线使用

应用程序开发
面临怎样的趋势

应用程序开发从烟囱式架构向分布式的转型

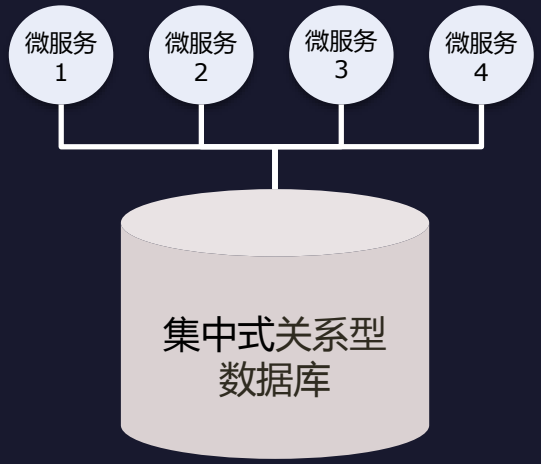
SequoiaDB 巨杉数据库



数据库该如何
应对微服务应用框架

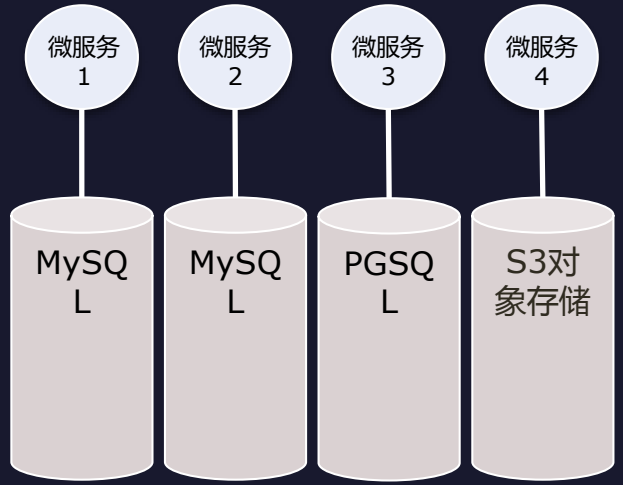
数据库如何应对微服务应用框架

SequoiaDB 巨杉数据库



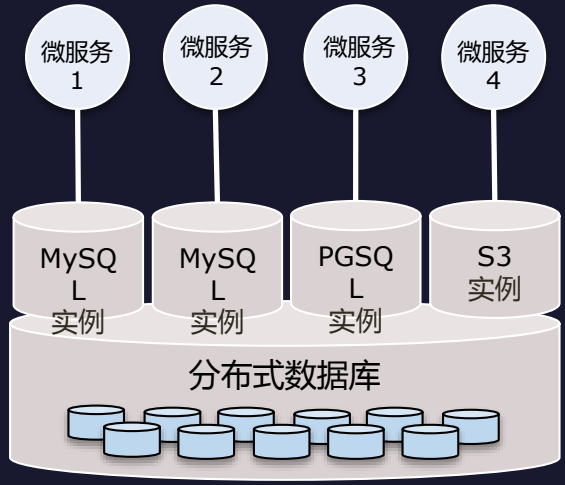
集中式存储

- 数据紧耦合
- 无法弹性扩张
- 单点故障



碎片化存储

- 数据碎片化
- 数据无共享
- 运维成本高



分布式存储

- 微服务对应独立实例
- 物理分散存储
- 逻辑集中管理

联机交易业务需要什么样的分布式数据库

联机交易业务需要什么样的分布式数据库

SequoiaDB 巨杉数据库

新技术前瞻性

传统技术兼容性

分布式与扩展性

分布式是新一代架构的基础，扩展性能应对变化的数据量

ACID的支持

事务、一致性等，处理OLTP

HTAP

混合事务和分析场景，适应更多数据应用需求

SQL完整支持

MySQL/PostgreSQL语法的完整兼容

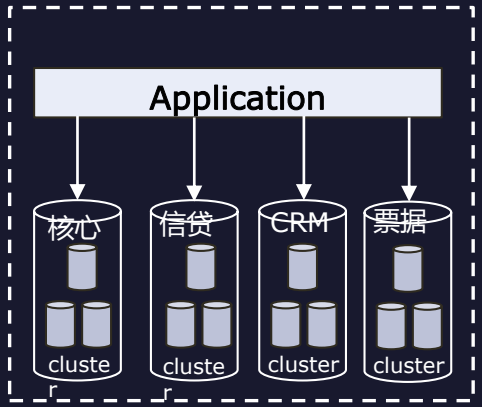
Multi-model与多租户

multi-model多模数据库引擎，同一引擎处理多种数据应用场景，符合微服务和云数据库的架构理念

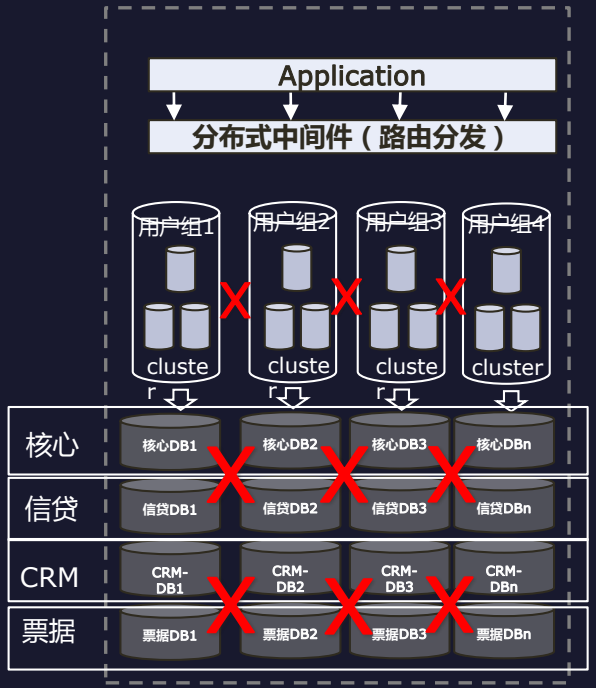
分布式交易型数据库技术发展体系

SequoiaDB 巨杉数据库

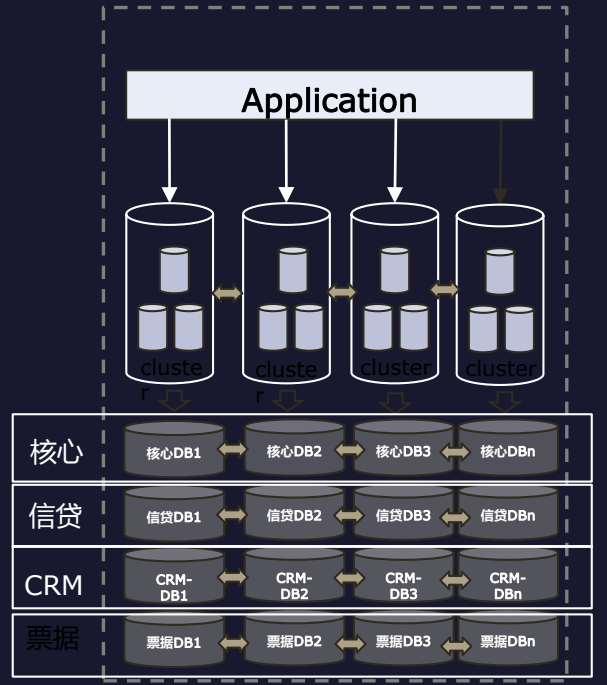
应用垂直分库



分库分表



原生分布式数据库



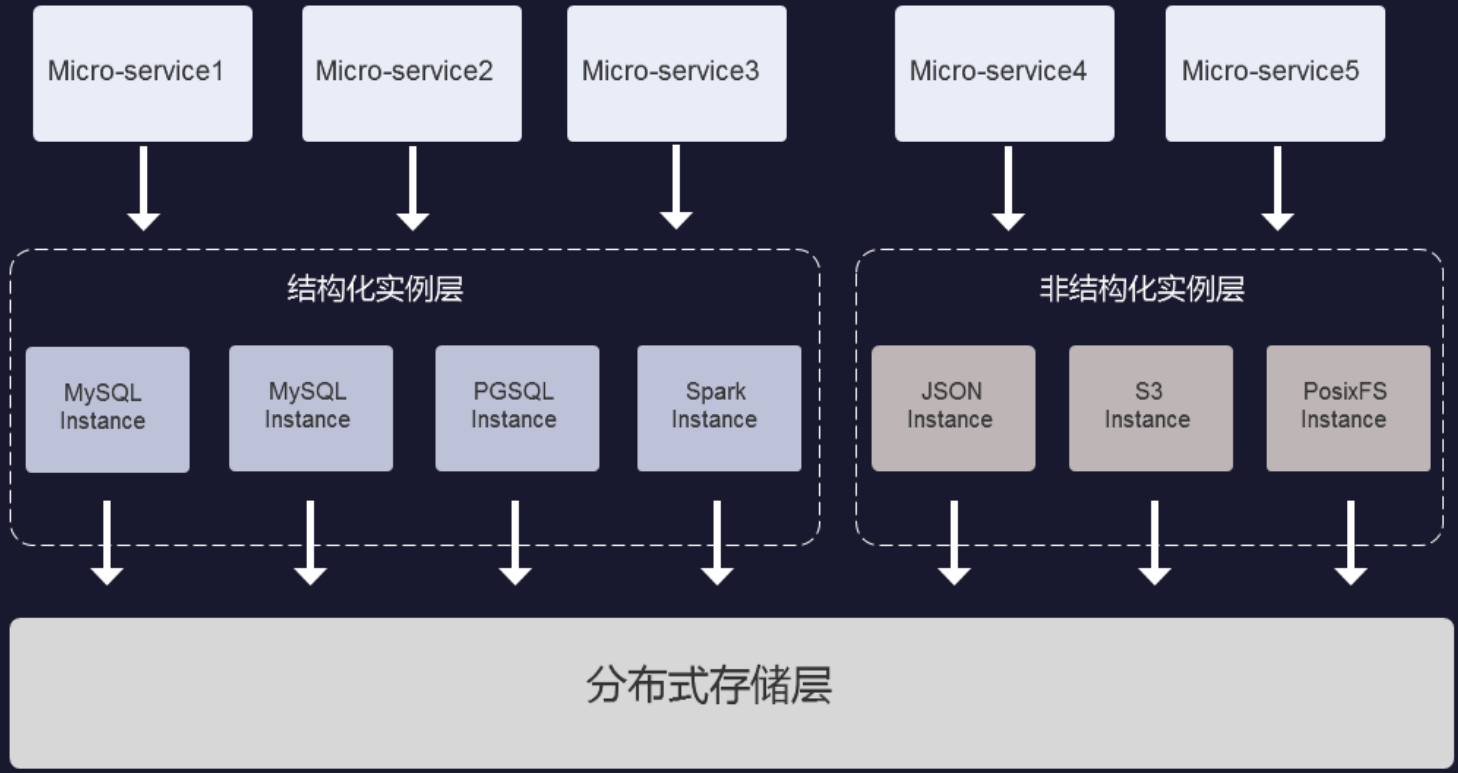
分布式数据库技术发展体系对比

SequoiaDB 巨杉数据库

	垂直分库	分库分表	原生分布式数据库
优势	• 起点比较早，应用控制能力强，可进行深度定制化 • 对于底层数据库没有任何特殊要求，完全在应用程序内部进行分库	• 构建中间SQL解析层，尽可能将标准SQL拆分成多个子查询下压到下层数据库，在SQL层进行结果拼装 • 对于底层数据库无特殊要求，在中间件进行SQL切分（支持XA即可） • 部分兼容传统SQL，应用程序开发难度小于垂直分库	• 数据库内部处理分布式事务与数据切分逻辑，对于应用程序完全透明，不需感知底层数据分布 • 数据库内部原生支持分布式事务，性能远远高于分库分表 • 高可用与容灾能力由数据库内核原生支持，不需额外辅助工具
劣势	• 应用程序逻辑侵入性极强，应用程序需要进行复杂逻辑才能进行合理数据分布 • 拓扑结构调整或扩容时非常痛苦，几乎不可能完成在线扩容 • 很难支持跨库事务	• 应用程序逻辑侵入性较强，应用程序需感知底层数据分布结构，才能设计出优化后的查询逻辑 • 中间件实现分布式事务，跨库事务使用XA机制，性能大幅度下降 • 作为单点向新型分布式数据库转型的过渡阶段，技术延续性堪忧	• 技术较新，业界成熟案例相对较少 • 辅助工具相对较少，生态环境有待完善

巨杉分布式数据库

SequoiaDB 巨杉数据库



“计算存储分离”架构

SequoiaDB 巨杉数据库

三大核心应用场景

SequoiaDB 巨杉数据库

联机交易

- 交易型业务场景
- 替换 MySQL、PGSQL 等传统关系型数据库

数据中台

- 数据服务与高频只读类业务
- 提供比 Hbase 更加友好的开发接口以及更加简便的运维能力

内容管理

- 音视频、图片、文件等对象存储类业务
- 提供比 Ceph 更优的实时容灾能力以及更加丰富的内容管理特性

新一代分布式数据库
如何适应微服务云化架构需求

数据分片能力

SequoiaDB 巨杉数据库

CategoriesID	MovieName	MovieDirector	MovieActor	MovieDesc	MovieData	MovieTime
6	何以笙箫默	无名	摩尼教	23额角o	2012-02-22 0...	20
1	速度与激情	张艺谋	不可谓	科ouee	2012-02-18 0...	120
6	阿科未婚夫	无名	ndk	jkduw	2012-02-22 0...	20
1	井底蛙电脑	张艺谋	马拉	看到	2012-02-18 0...	120
6	洛帝古镇	无名	农户	鸡窝	2012-02-22 0...	20
1	测试名字	张艺谋	潘长	摩尼教	2012-02-18 0...	120
6	洛帝古镇	无名	无名	，罚款	2012-02-22 0...	20

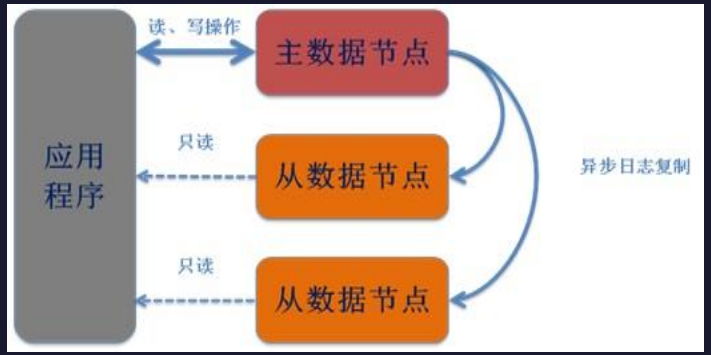
TargetPartition = DHT (Row->PartitionKey)

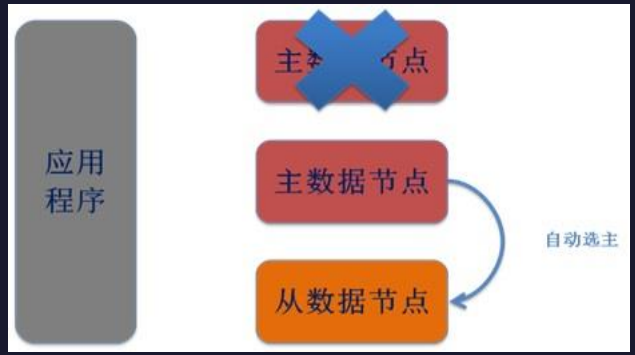
datagroup1

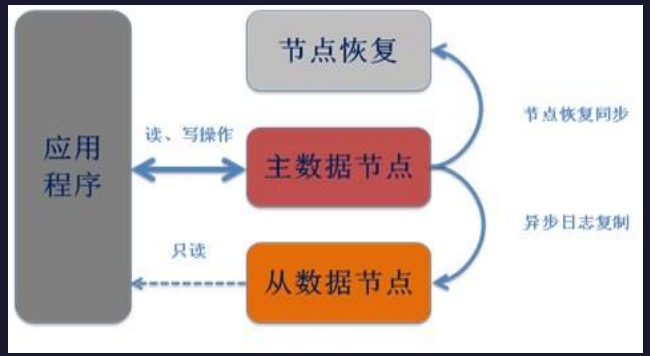
datagroup2

datagroup3

高可用能力



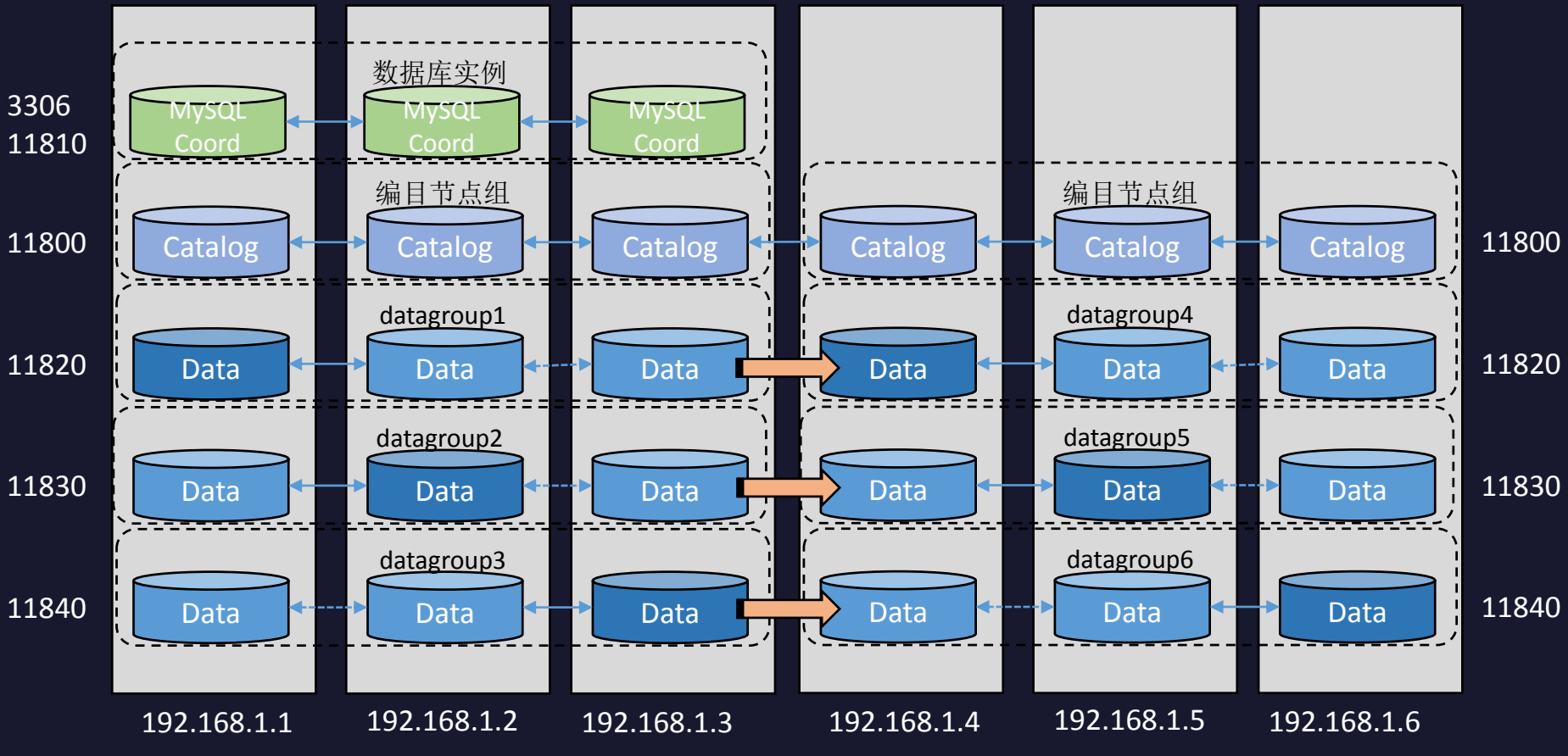




- 同分区内数据节点之间通过心跳保持连接
- 主节点2轮接收不到超半数节点心跳会自动降备
- 备节点2轮接收不到主节点心跳会发起选举投票
- 超半数节点同意后备节点当选新的主节点

水平扩展能力

SequoiaDB 巨杉数据库



强大的分布式事务能力

SequoiaDB 巨杉数据库

- 传统二段提交机制
- 保证数据跨节点一致性

- 流水类数据按时间与ID二维切分，避免数据搬迁
- 余额类数据按ID散列，保证均衡无热点

- MySQL/PGSQL/SparkSQL保持100%兼容
- 原生MySQL/PGSQL/SparkSQL解析与执行引擎，不需担心语法兼容访问计划

- 悲观锁
- MVCC读已提交能力

MySQL兼容能力

SequoiaDB 巨杉数据库

语法

- 增删改查语法（SQL、DML）
- 视图、存储过程、触发器、自增字段（DDL、DCL）
- 跨节点跨表事务、四种隔离级别、读已提交能力

通讯协议

- 协议级兼容MySQL客户端
- 协议级兼容JDBC驱动
- 支持所有MySQL开发框架

访问计划

- 访问计划计算方式兼容MySQL
- 统计信息收集策略兼容MySQL

准实时数据复制

SequoiaDB 巨杉数据库

准实时数据复制策略

1、异构数据源使用相关的工具将日志文件实时解析并写入管道

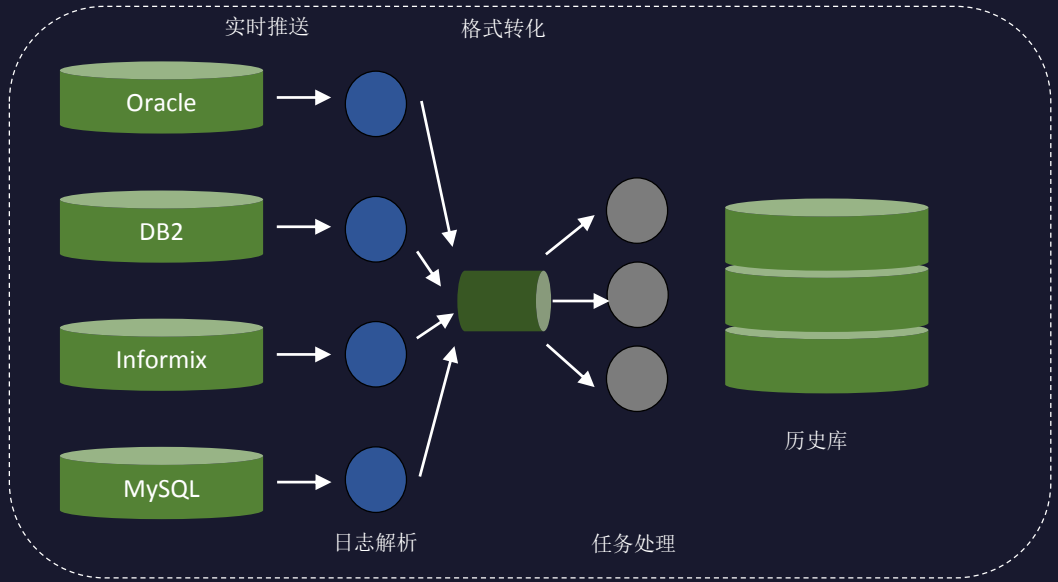
2、通过Apache Storm对管道信息监听并转换为标准DML/DDL命令

3、指令分发至多线程处理服务进行巨杉历史数据库的增删改查

4、满足异构数据源T+0的数据复制策略，秒级延时

5、当前支持Oracle Golden Gate（对应Oracle数据源）、IBM CDC（对应IBM DB2）、IIE（对应IBM Informix）、以及Cannel（对应MySQL）

6、对于当前不支持的数据库需要寻找开源的日志解析工具或进行独立开发



准实时数据复制策略

异步数据复制

SequoiaDB 巨杉数据库

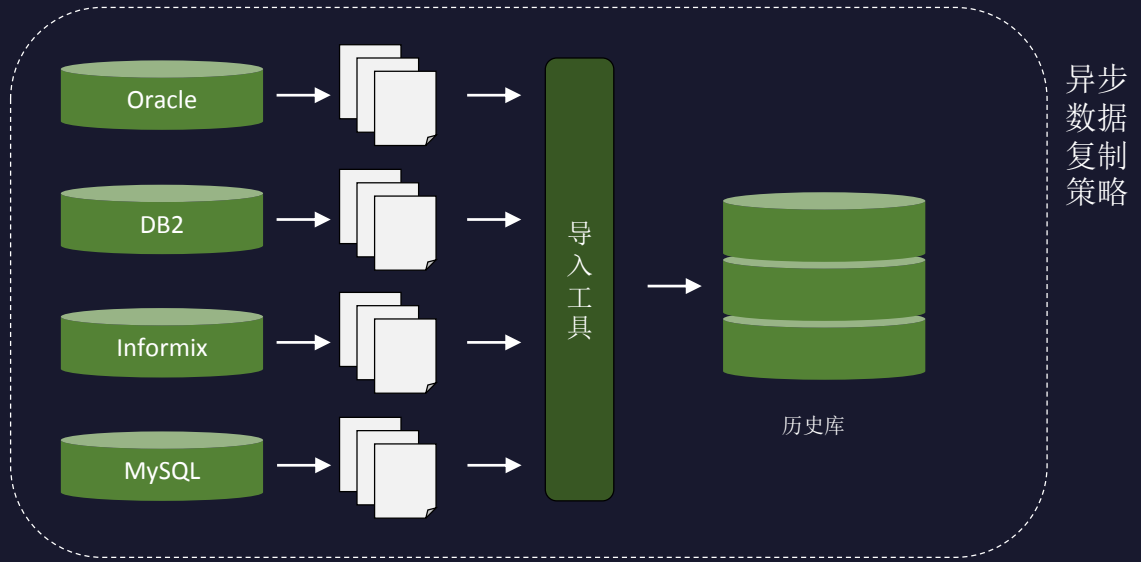
异步数据复制策略

1、使用JSON或CSV格式定期将异构数据源的原始数据进行导出为文本文件

2、通过FTP等方式将文件传输至巨杉数据库的客户端

3、通过sdbimprt工具将文本文件导入巨杉数据库

4、满足异构数据源T+1的数据复制策略，简单可靠



数据备份恢复

SequoiaDB 巨杉数据库

数据库实例级备份

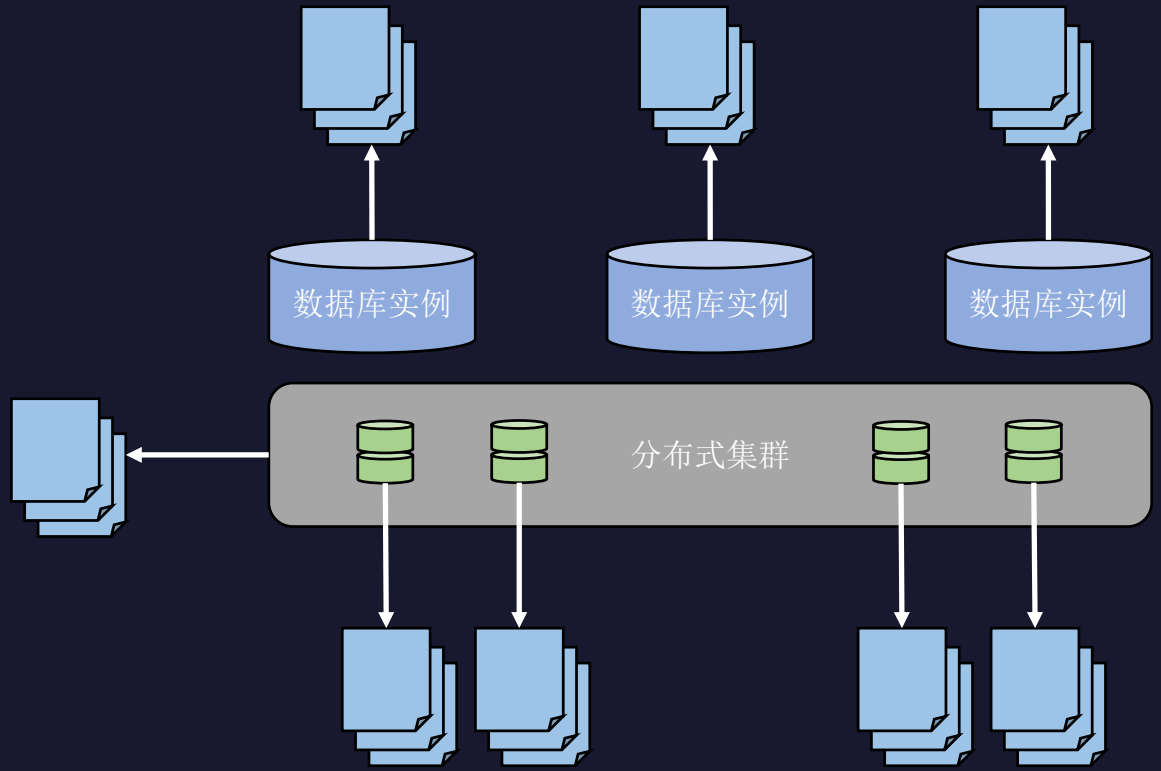
- MySQL/PostgreSQL原生记录备份策略

集群级备份

- 全量离线备份
- 全量在线备份
- 增量在线备份

文件系统级备份

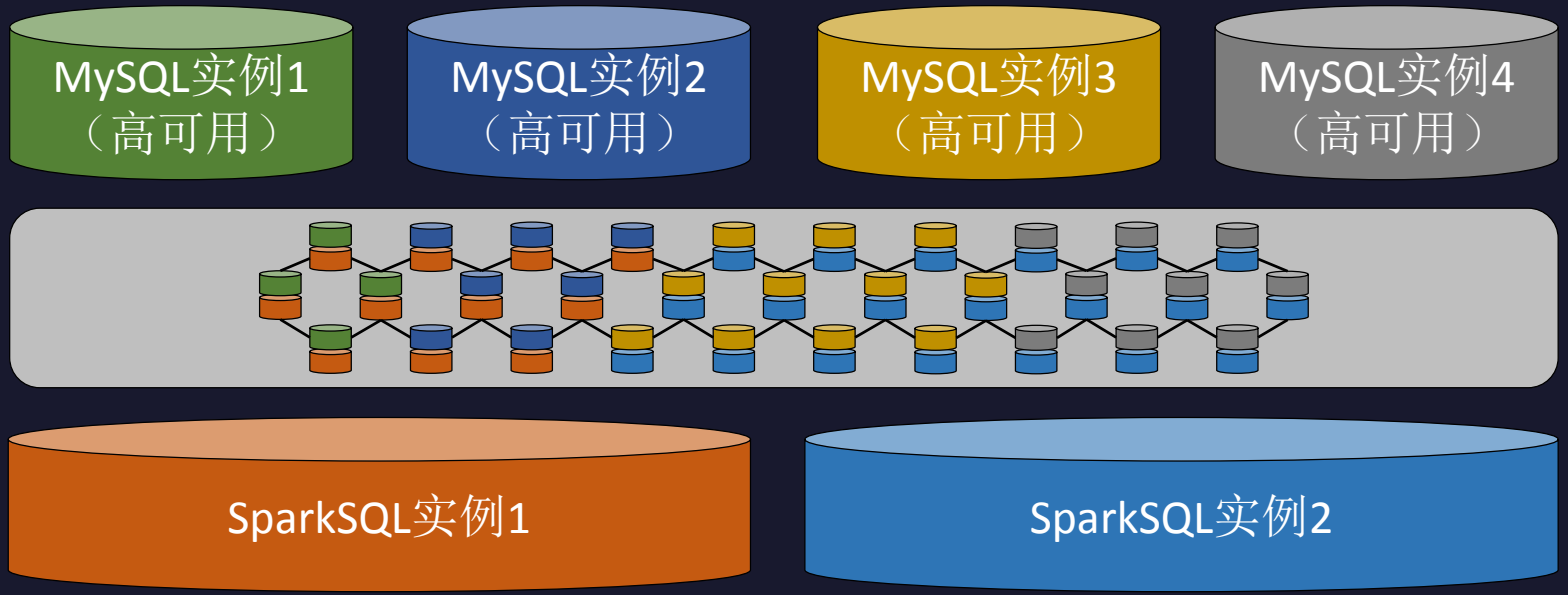
- 读节点文件系统全量备份
- 静态文件增量备份



HTAP读写分离能力

SequoiaDB 巨杉数据库

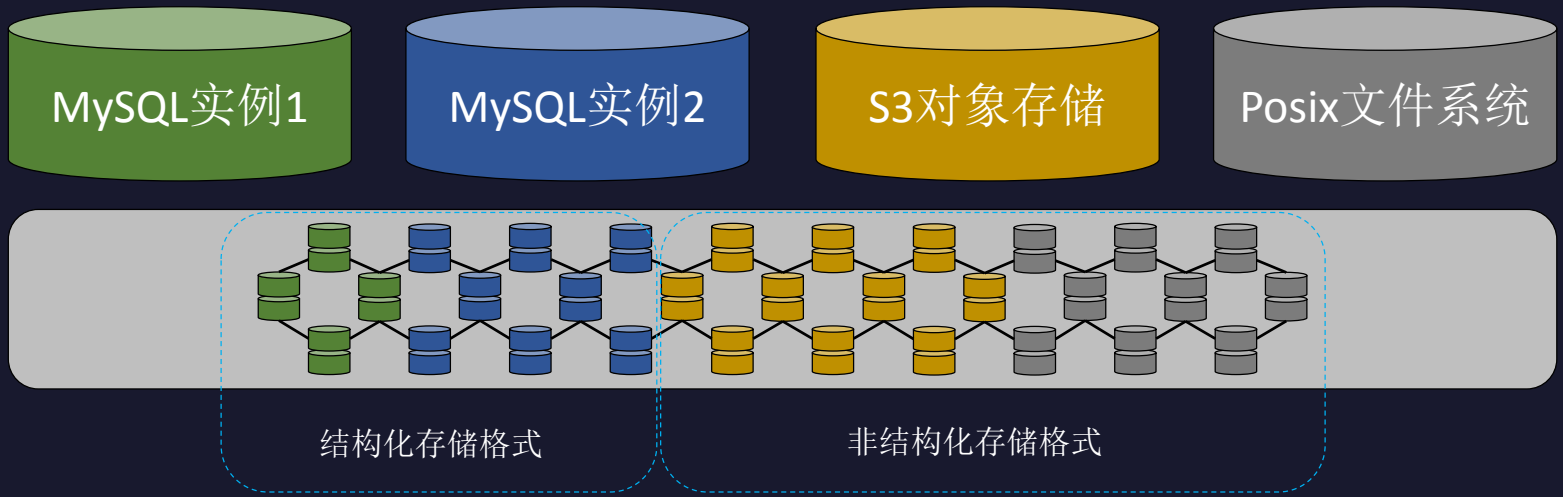
微服务框架下，对成千上万个MySQL数据库实例做到统一化管理，防止数据碎片化，并对来自不同实例和服务的数据统一实时分析，避免联机交易与分析业务相互干扰



多租户物理隔离能力

SequoiaDB 巨杉数据库

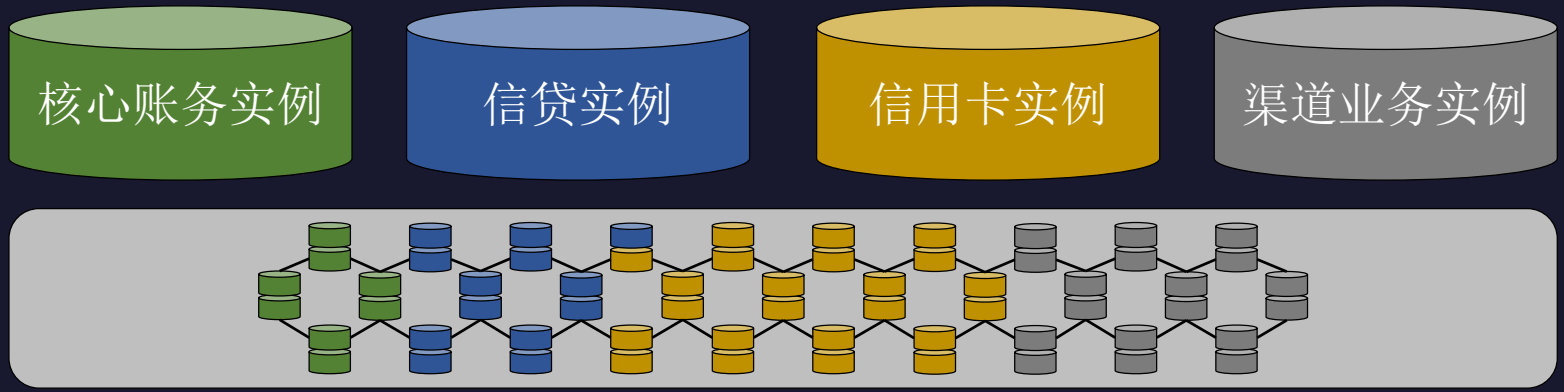
在一个集群内同时提供关系型数据库以及对象存储实例
尽可能减少用户对于异构产品的学习与运维成本



多租户权限管理能力

SequoiaDB 巨杉数据库

每个实例提供完全隔离的权限控制与数据可视范围
确保不会管理员不会有意无意使实例访问被隔离的其他信息

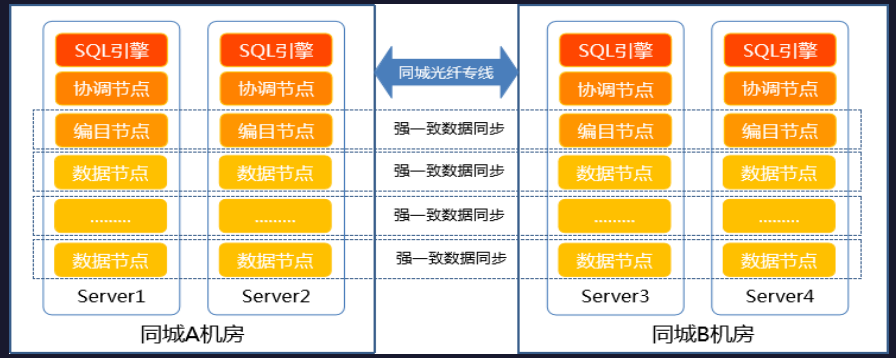


多中心容灾能力

SequoiaDB 巨杉数据库

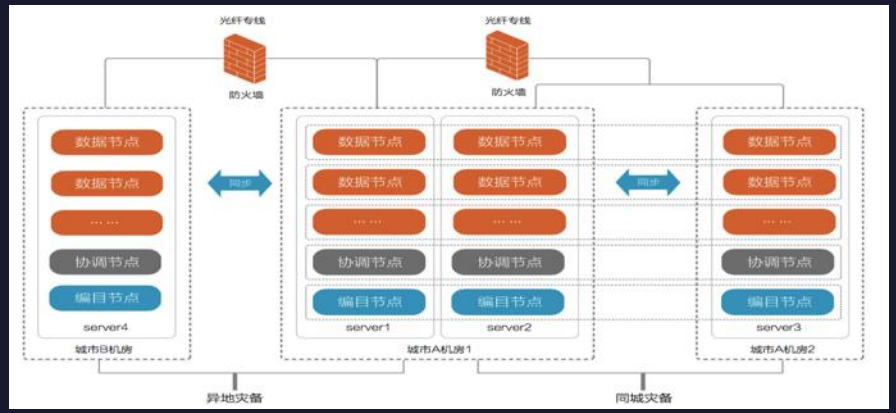
同城方案

1、主备机房使用可靠高速光纤直连
2、每个分区主节点在主中心
3、平时使用强一致同步策略保障数据不丢
4、故障发生时使用takeover工具进行集群分离，备集群独立运行
5、故障恢复后使用merge工具进行集群合并



双活方案

1、应用程序直连本地数据中心数据库协调节点
2、应用程序不需要关注底层数据存储主备中心复制和通讯策略



两地三中心

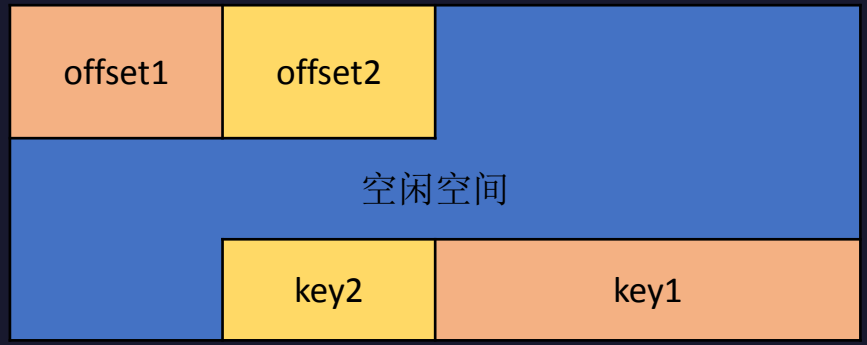
1、远程数据中心使用异步机制进行数据复制
2、数据中心之间可进行流量控制保证不会占用过多带宽

分布式数据库存储引擎机制

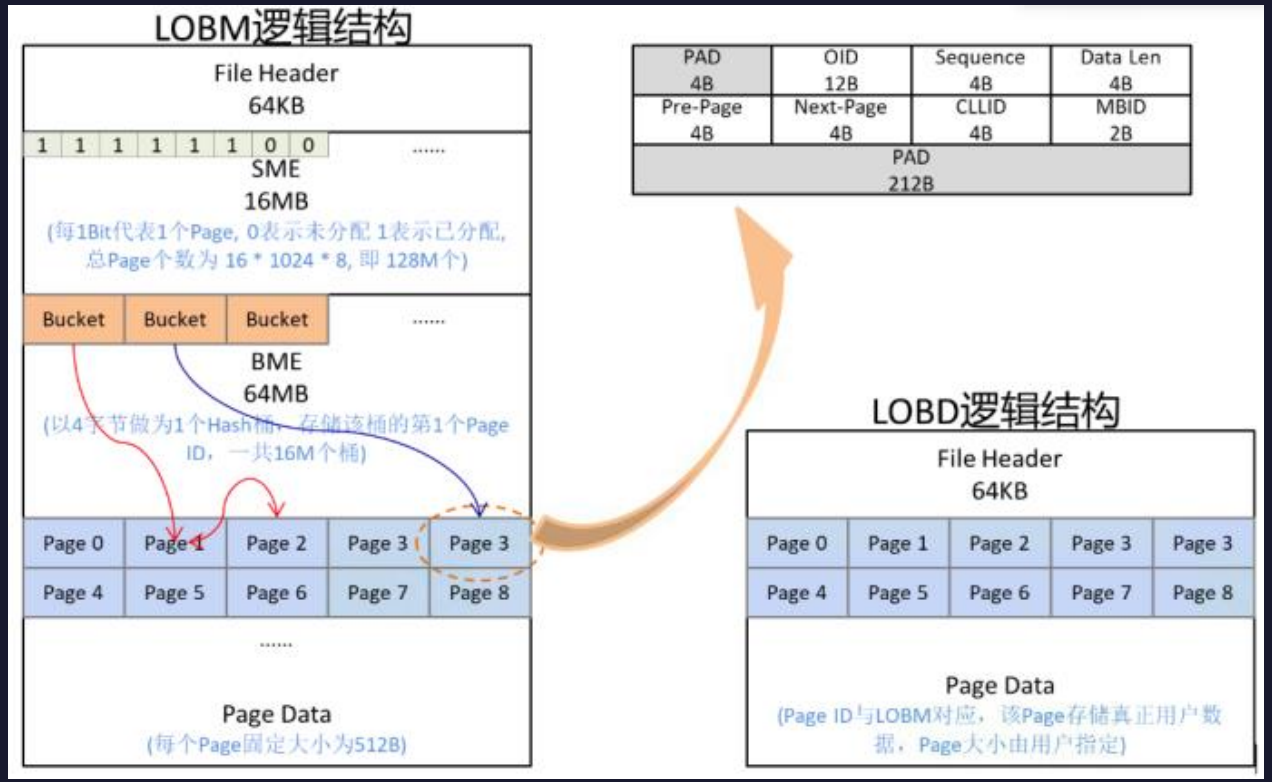
记录存储格式

SequoiaDB 巨杉数据库

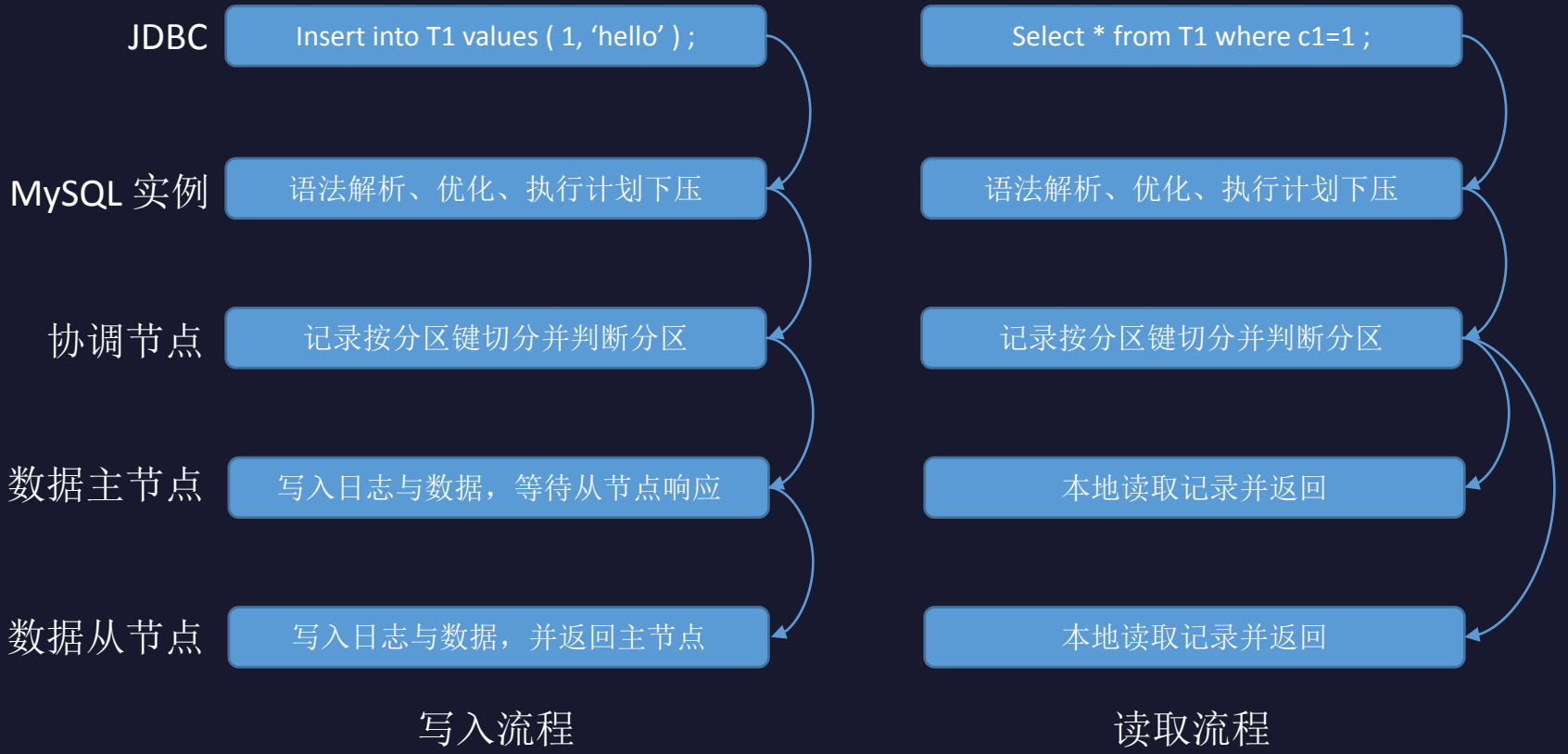
File Header					
Space Management Extent					
Data Segment Metadata					
CS Metadata					
page1	page2	...			



对象文件存储格式


LOBM逻辑结构
File Header
64KB
1 1 1 1 1 1 0 0
.....
SME
16MB
(每1Bit代表1个Page, 0表示未分配 1表示已分配,
总Page个数为 16 * 1024 * 8, 即 128M个)
Bucket
Bucket
Bucket
.....
BME
64MB
(以4字节做为1个Hash桶，存储该桶的第1个Page
ID，一共16M个桶)
Page 0
Page 1
Page 2
Page 3
Page 3
Page 4
Page 5
Page 6
Page 7
Page 8
.....
Page Data
(每个Page固定大小为512B)
PAD
4B
OID
12B
Sequence
4B
Data Len
4B
Pre-Page
4B
Next-Page
4B
CLLID
4B
MBID
2B
PAD
212B
LOBD逻辑结构
File Header
64KB
Page 0
Page 1
Page 2
Page 3
Page 3
Page 4
Page 5
Page 6
Page 7
Page 8
Page Data
(Page ID与LOBM对应，该Page存储真正用户数
据，Page大小由用户指定)

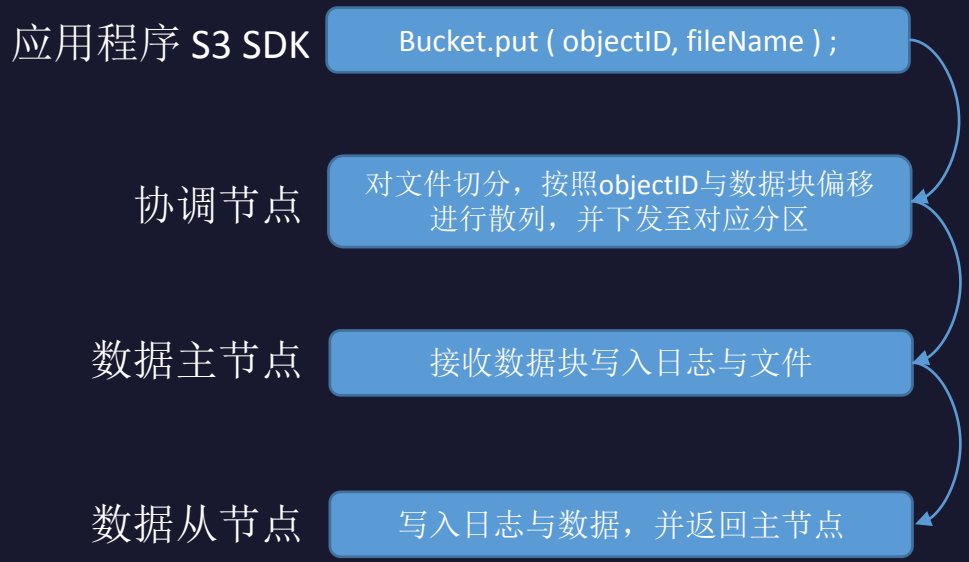
记录读写流程



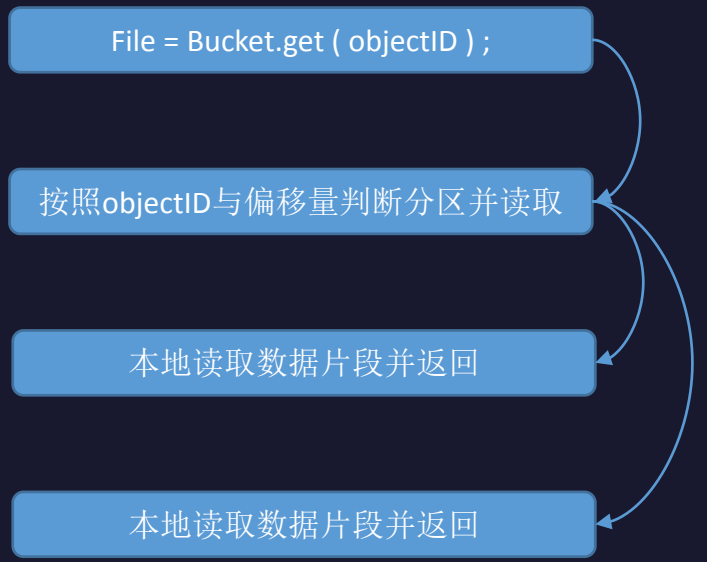
写入流程

读取流程

对象读写流程



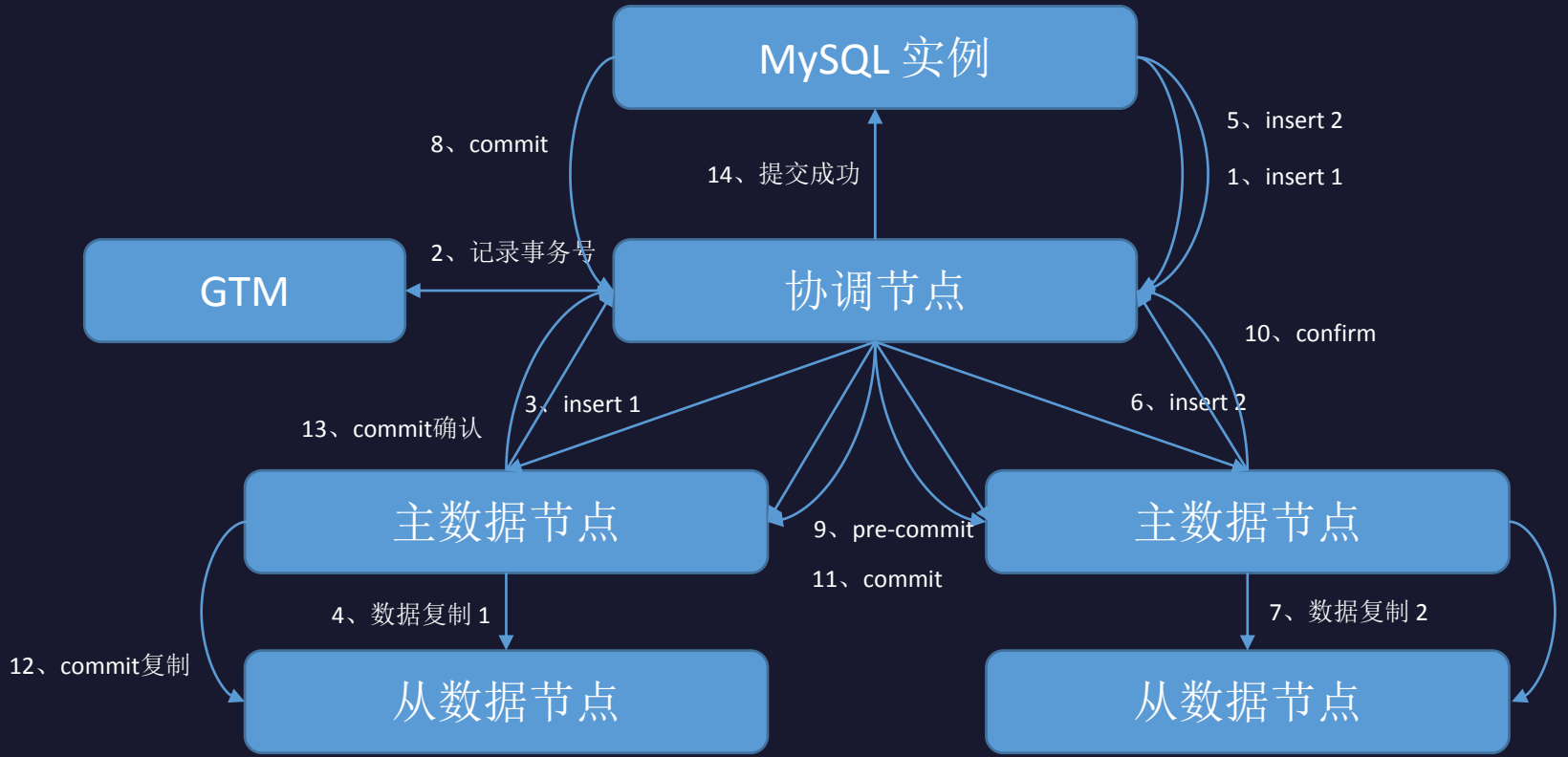
写入流程



读取流程

两阶段提交过程

SequoiaDB 巨杉数据库



金融级分布式关系型数据库

立即开启全新体验：

http://download.sequoiadb.com/cn/