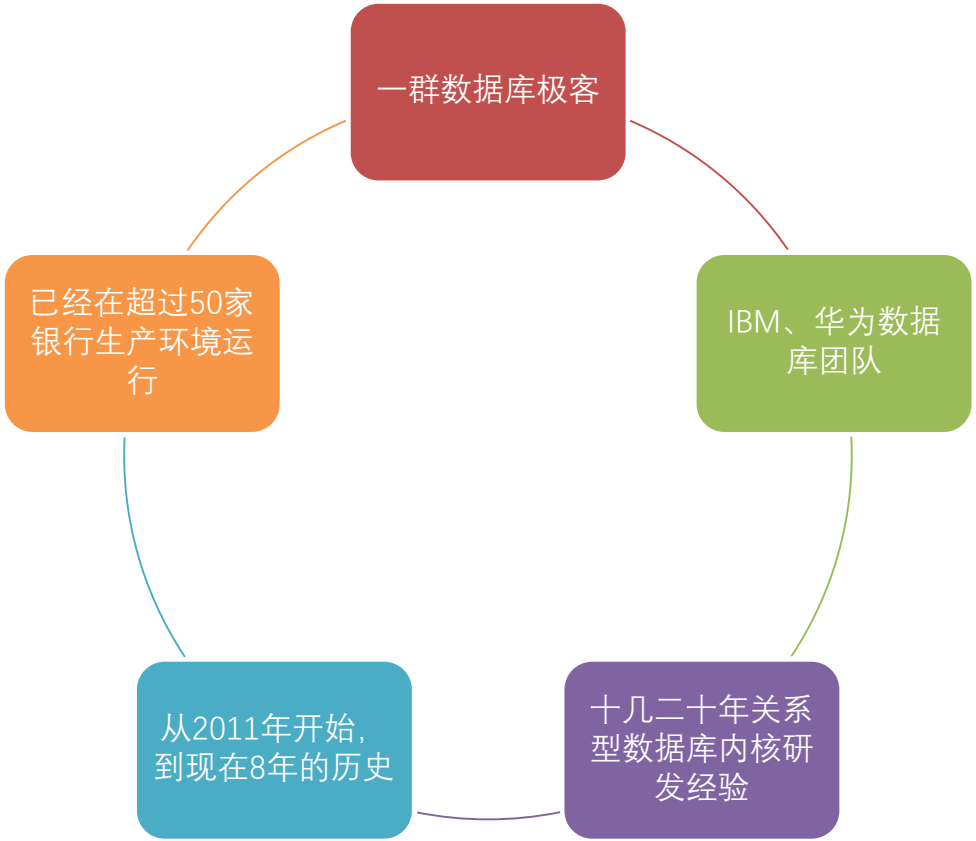
分布式数据库架构实践

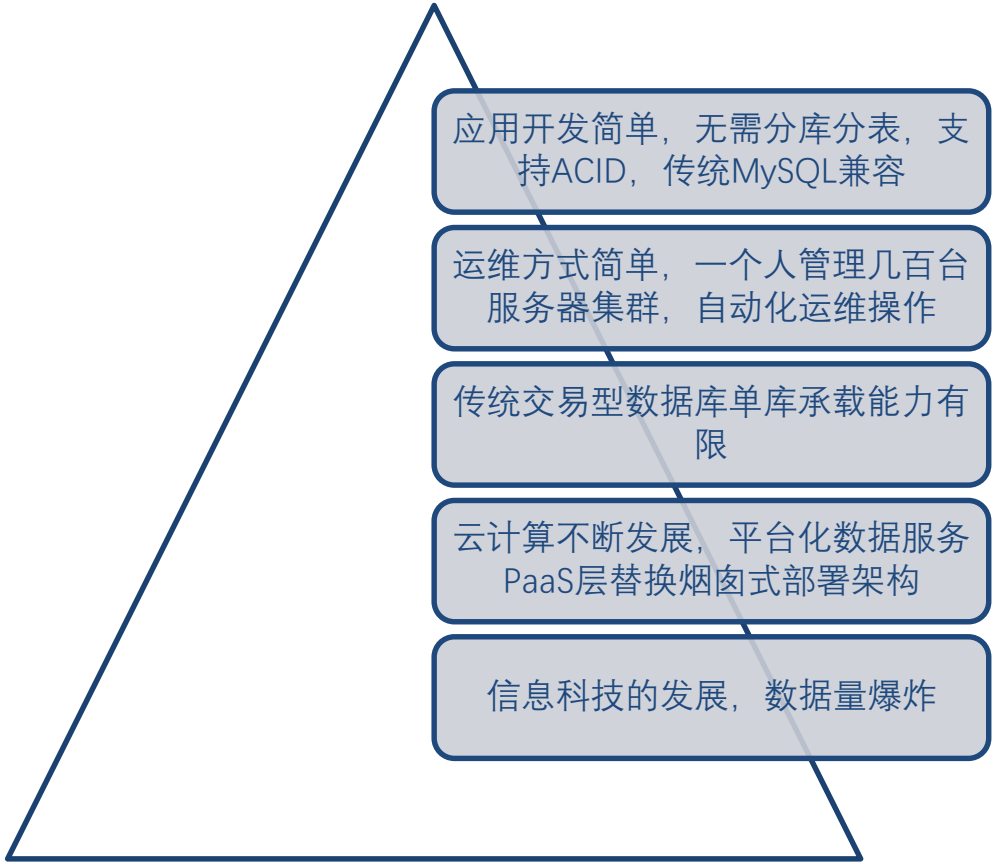
巨杉数据库 王涛

SequoiaDB
巨杉数据库

我们是谁


一群数据库极客
IBM、华为数据库团队
十几二十年关系型数据库内核研发经验
从2011年开始，到现在8年的历史
已经在超过50家银行生产环境运行

我们在解决什么问题


应用开发简单，无需分库分表，支持ACID，传统MySQL兼容
运维方式简单，一个人管理几百台服务器集群，自动化运维操作
传统交易型数据库单库承载能力有限
云计算不断发展，平台化数据服务PaaS层替换烟囱式部署架构
信息科技的发展，数据量爆炸

巨杉数据库：Gartner认可的中国金融级分布式开源数据库技术

- 受到行业客户广泛认可，拥有多家500世界强客户及超过50家大型银行客户，在金融行业树立标杆地位
- 巨杉是国内唯一商业化的新型分布式数据库，在银行核心系统大规模取代北美同类产品，功能性能全面领先国际竞品
- 中国唯一金融级自主分布式数据库，受到顶级基金支持，在自主可控要求下在中国存在巨大市场空间

- 多模分布式数据库，支持结构化交易、半结构化以及非结构化数据
- 完全开源，代码托管在github上 github.com/sequoiadb/sequoiadb
- 数据库开发顶尖团队，中国第一批数据库内核开发者，具有丰富行业经验

巨杉数据库——Gartner认可的中国金融级数据库公司

2017，2018年连续两年入选Gartner数据库系列报告

与阿里云数据库，是中国仅有的两家入选报告的产品

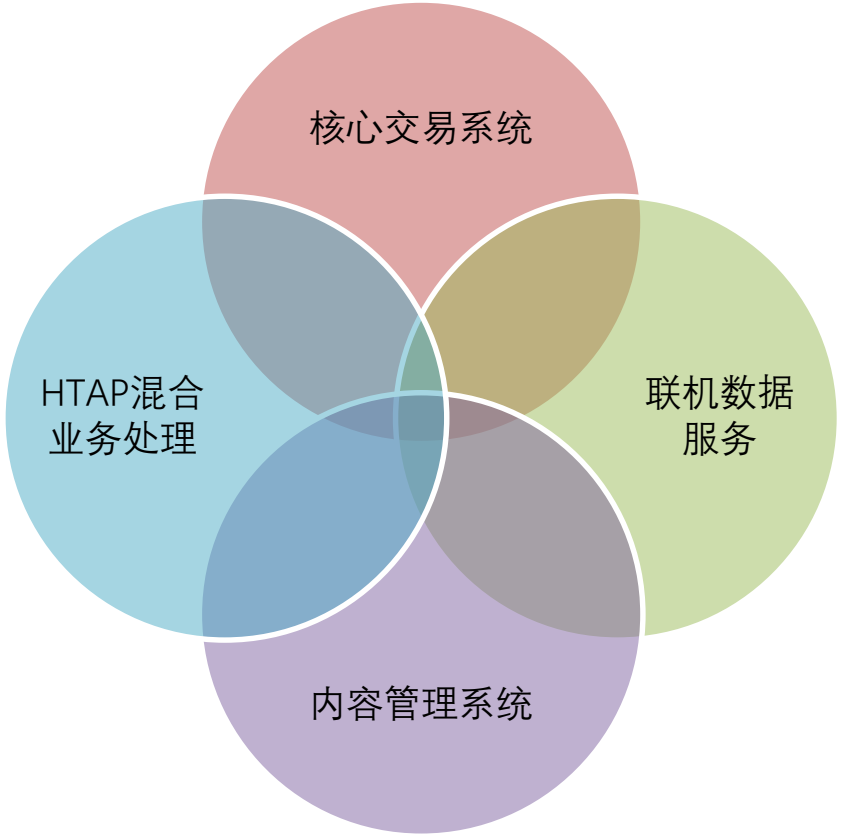
金融级客户的认可

巨杉数据库付费企业级客户与社区用户总数超过1000家

已在超过50家500强级别的银行、保险、证券等大型金融机构核心生产业务上线

案例分享

巨杉数据库典型使用场景分类



众多银行的联机交易系统

在线交易系统是银行的最重要的核心系统之一。随着技术的演进和监管政策的更新，目前银行核心交易系统面临的主要痛点是：

数据量和性能的扩展	分布式架构转型	事务和一致性	高可靠性	自主可控与数据安全要求
• 随着互联网业务的发展，核心交易系统在数据量和并发性等性能要求逐渐增多。	• 分布式架构在扩展性、高可用等方面带来了诸多好处，因此银行在线交易系统也存在分布式架构转型的需求，以应对性能、成本、跨地域管理和数据安全的多种业务要求。	• 事务和数据一致性是核心系统数据库必须要求的特性，保证事务和一致性是在线系统的重要要求。	• 支持两地三中心部署，灾难中零数据丢失，与同城双活架构	• 产品的逐步国产自主可控以及“两地三中心”等数据安全要去越来越重要。

某大型保险公司生产系统迁移案例

统计分析类应用直接访问生产库

统计分析类应用直接以生产库作为数据源，同一份数据甚至被重复访问，占用大量生产库的批处理时间窗口。

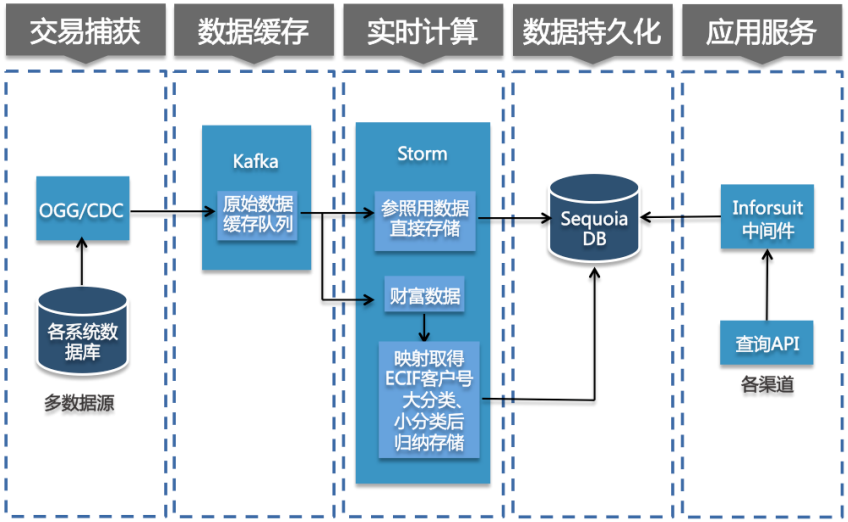
缺少统一的全量数据平台

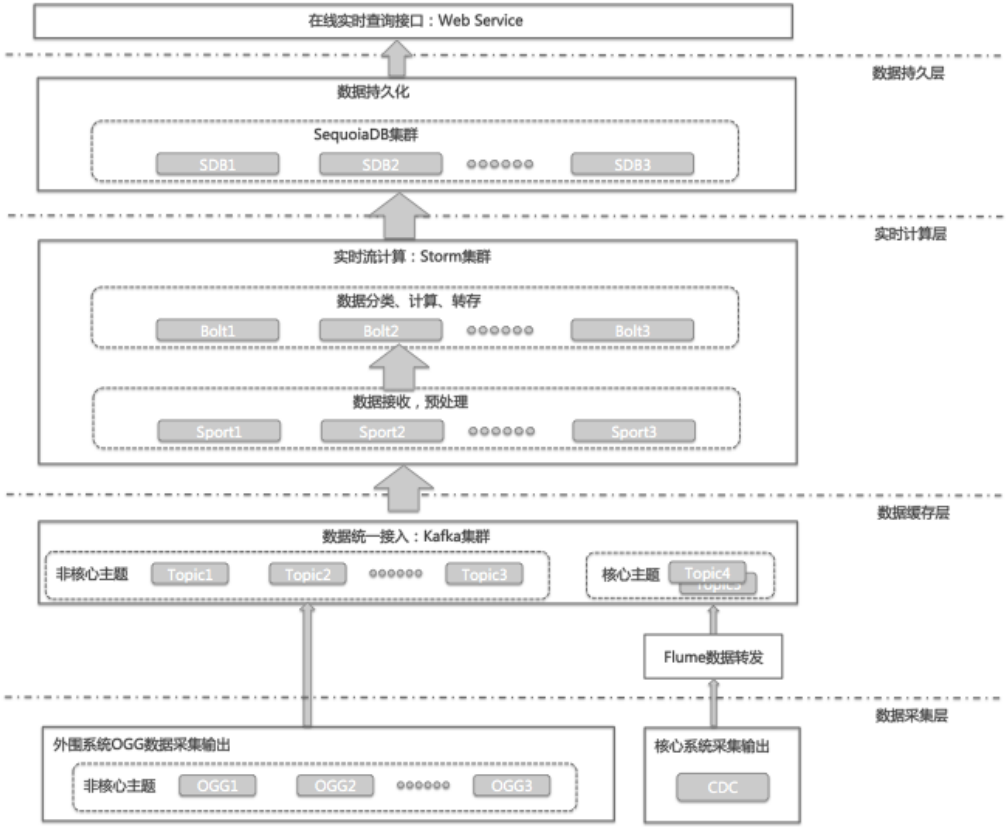
大量近线数据保存在生产库中，近线数据的查询需要访问生产库，加重生产库的负担。

缺乏有效的数据迁移方案

当前的近线数据迁移方案效率不高，导致生产库的规模迟迟不能得到控制，已经影响到个别省份的生产效率。

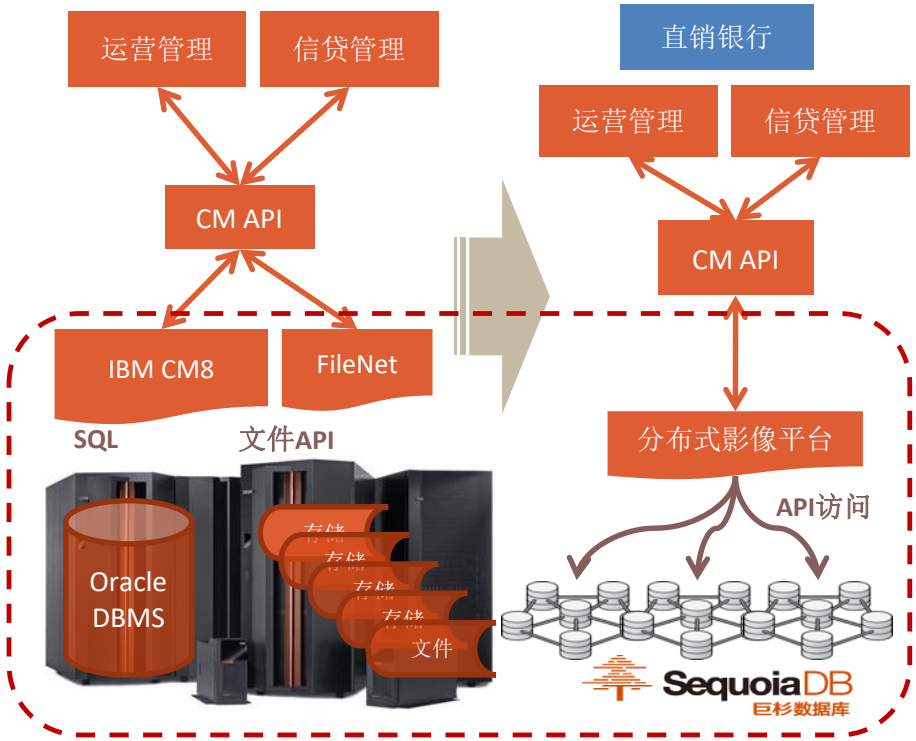
某股份制银行在线服务平台

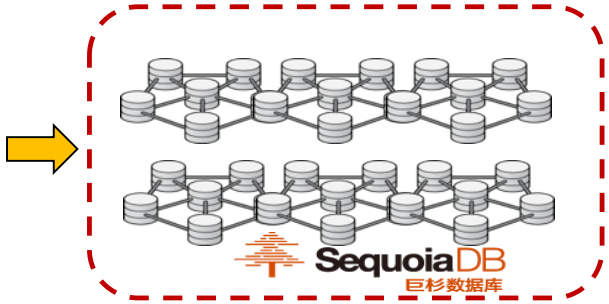




某股份制银行影像数据平台

- 全行影像系统迁移至巨杉数据库，近PB级存储





问题与需求

- 数据量超过2亿条后，原ECM系统性能急速下降
- 新业务要求更高的容量，但扩容成本高，扩容操作麻烦。
- 依赖原系统的工作流。
- 历史数据在光盘库，查阅不便。

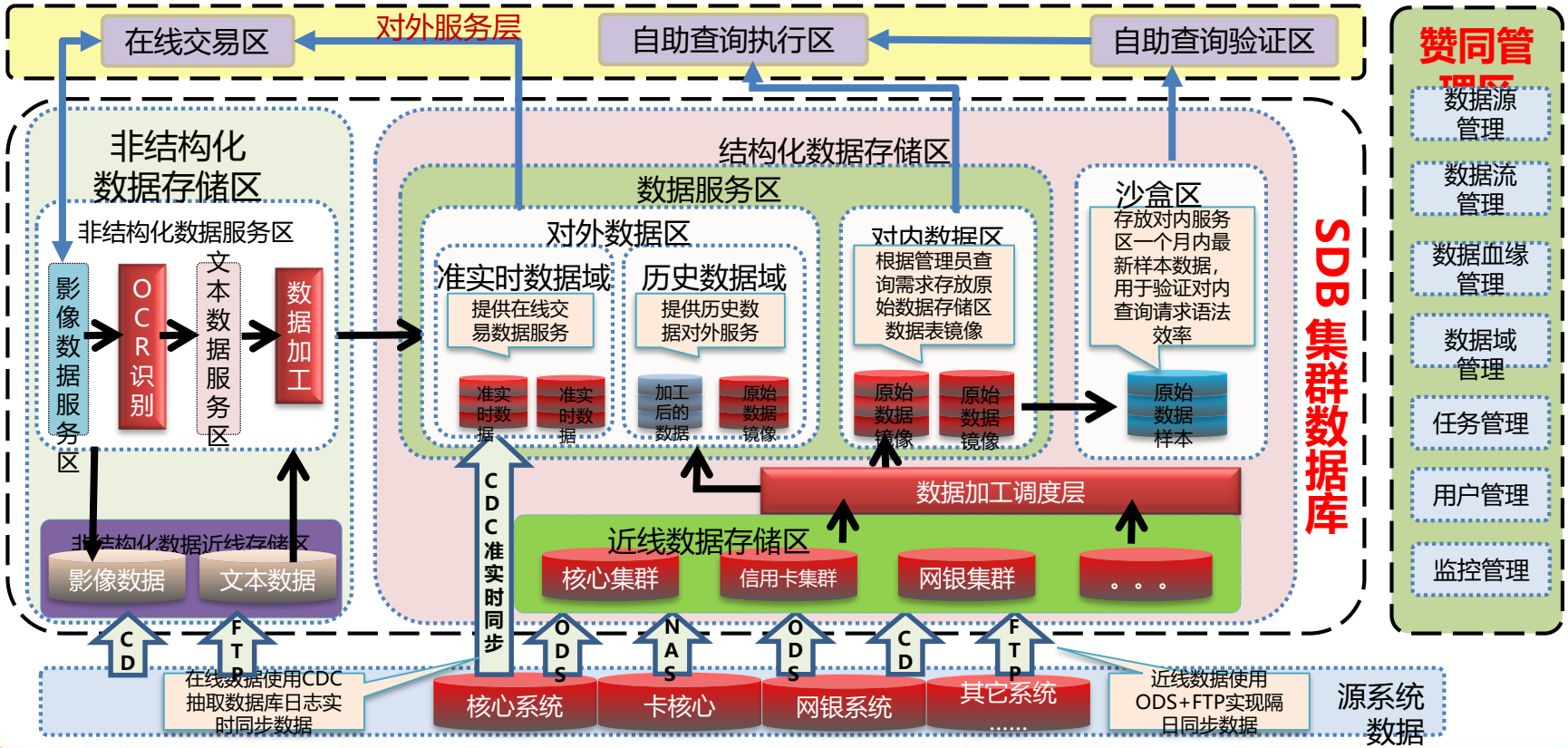
解决方案

- 巨杉平台作为数据归档平台 – 历史数据在线化
- 借助巨杉平台，构建全量数据的同城容灾方案

实现效果

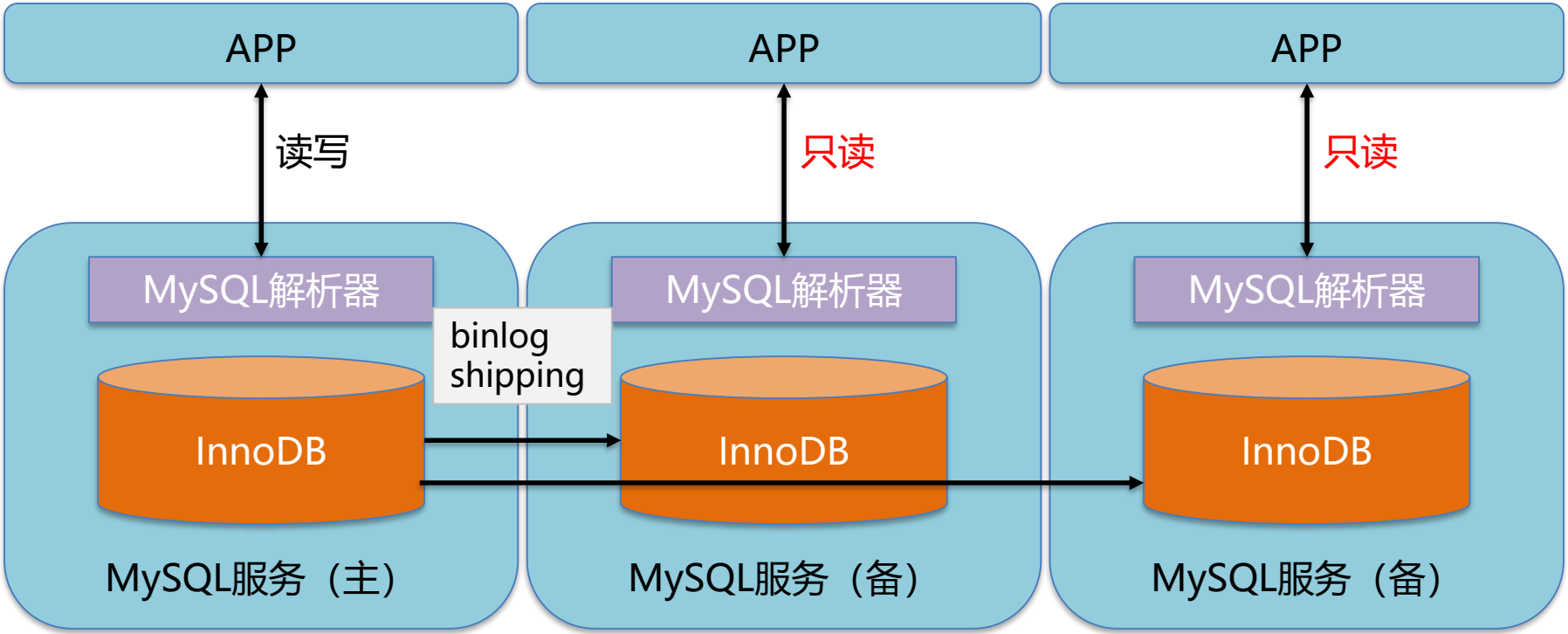
- 超过60个接入业务系统
- 超过600TB数据存储量
- 平均每日千万笔交易

某股份制银行近线数据服务平台

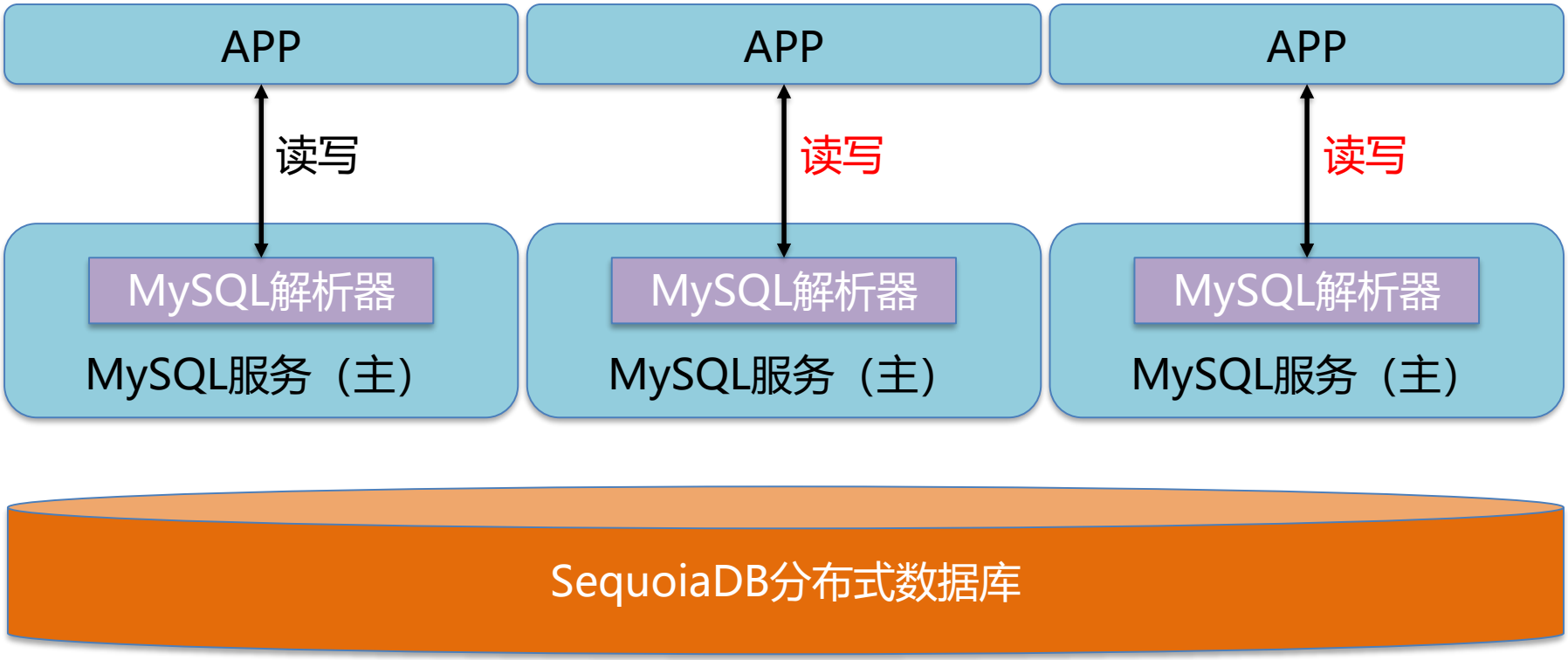


技术特性

MySQL主从复制架构



SequoiaDB – MySQL 100%兼容，弹性扩张，多活架构

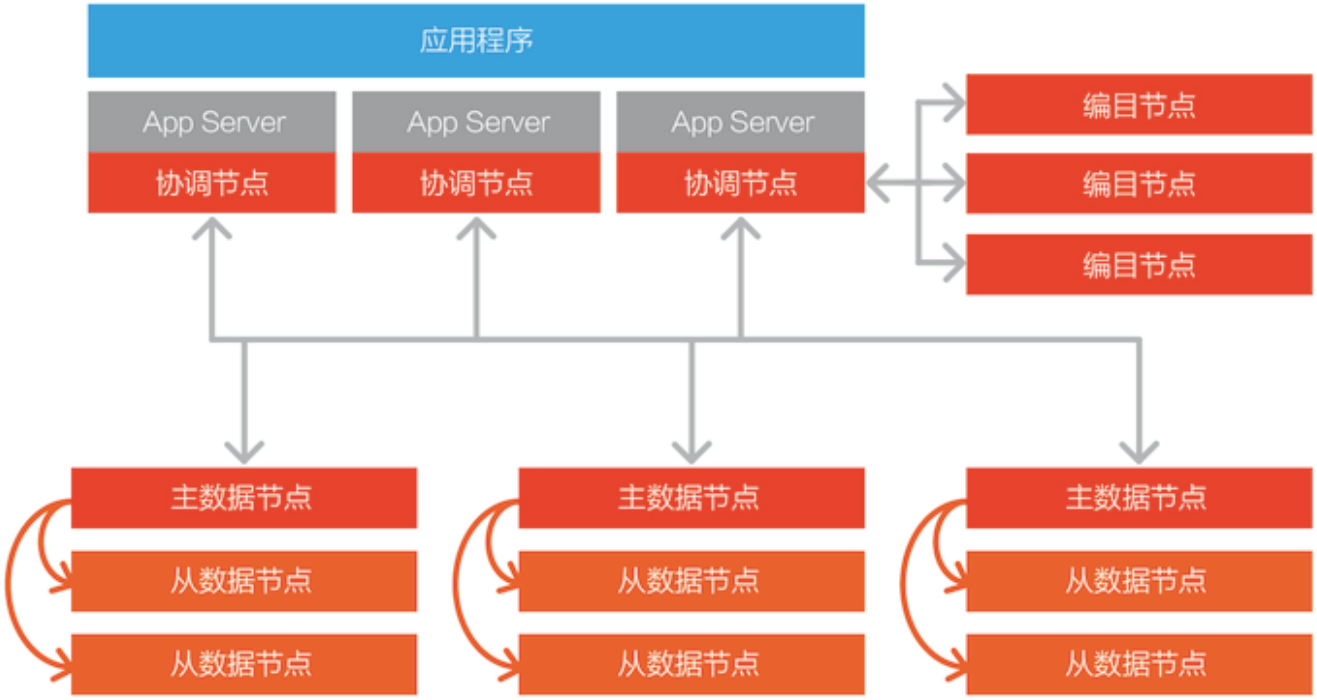


SequoiaDB 分布式数据库架构

计算分布

+

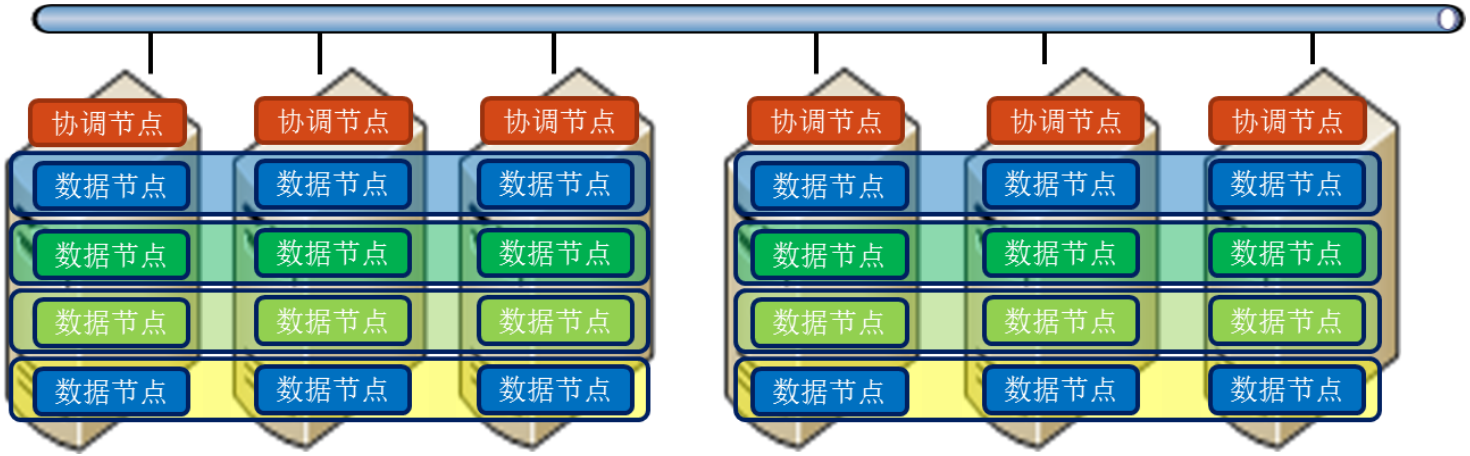
存储分布



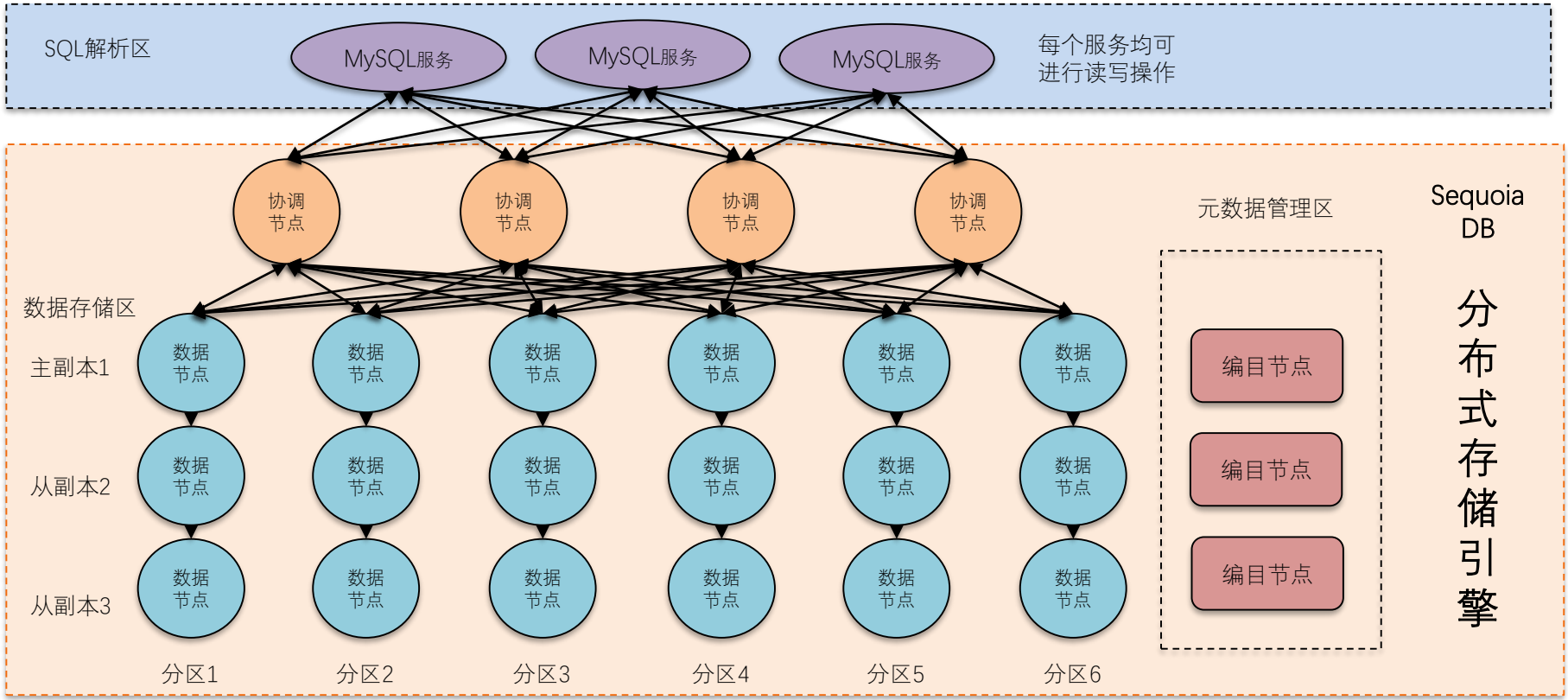
SequoiaDB 分布式数据库架构

SequoiaDB数据存储层采用分布式架构，实现了弹性水平扩展以及高性能和高可用，灵活适应不同规模企业及不同作业方式的需要。

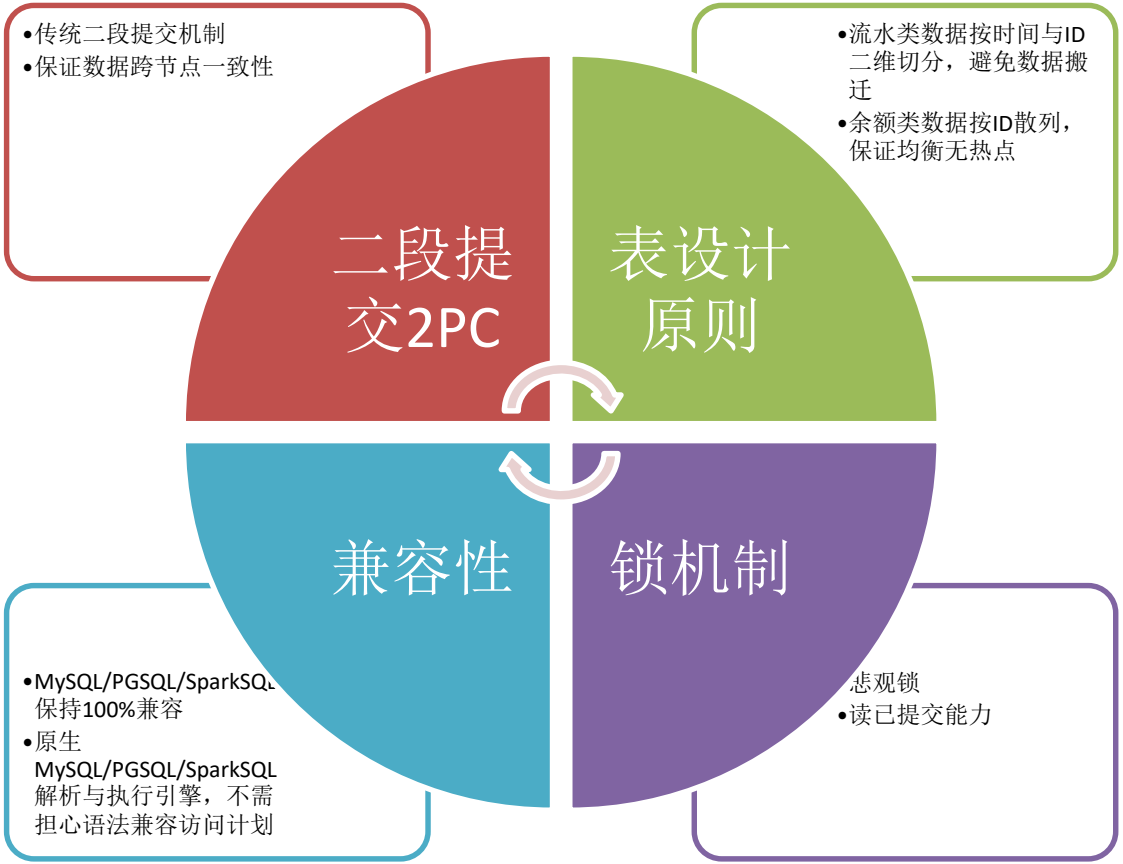
- 数据切分：系统提供垂直切分和水平切分的多维分片管理方式，数据可以按多种条件切分，均匀分布到集群中的各个数据节点。
- 高可用：数据在系统中至少保留三个副本，高可用机制，保证了数据的持续安全使用。
- 弹性扩容：SequoiaDB的存储节点可按需弹性扩展，系统支持在线扩容
- 硬件成本降低：分布式架构均采用通用x86服务器+高密度硬盘，相比传统的“小机+高端存储”的配置，大大节省了用户投资费用。



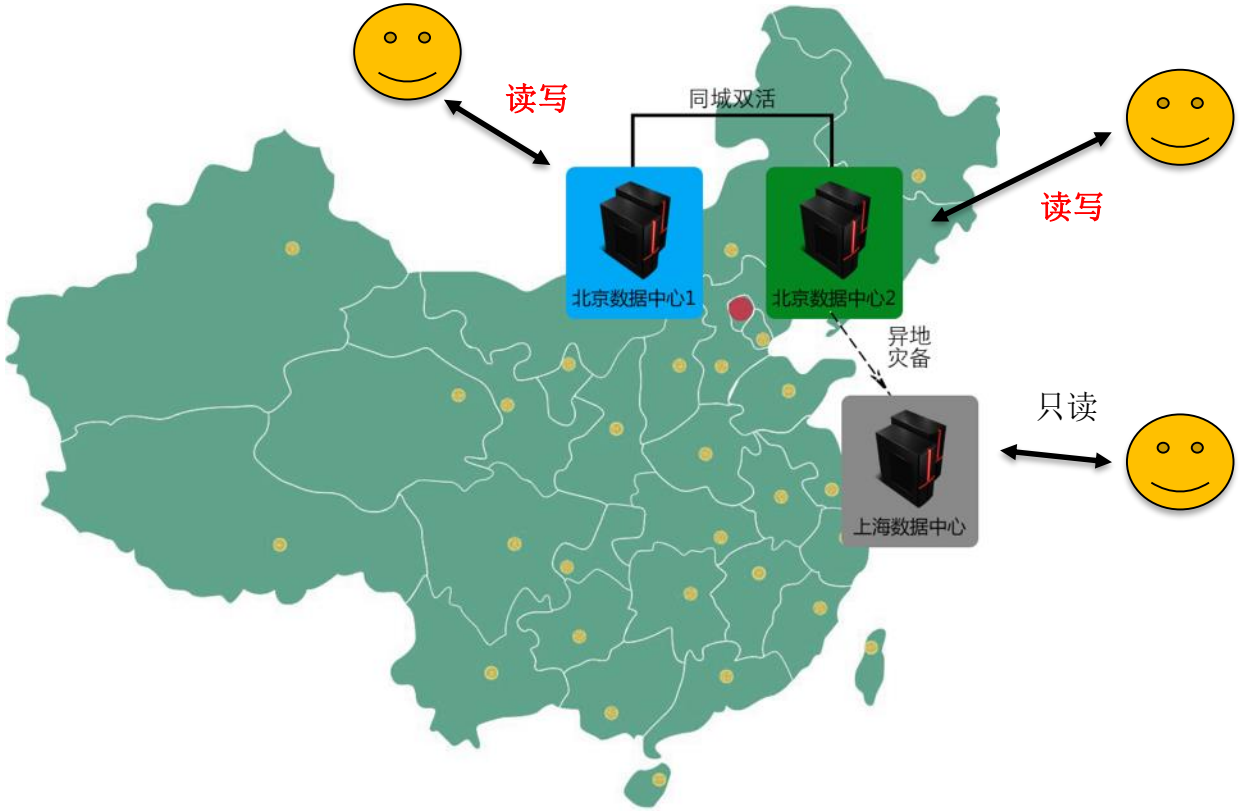
SequoiaDB – MySQL 分布式数据库架构



分布式交易设计


•传统二段提交机制
•保证数据跨节点一致性
二段提交2PC
•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
表设计原则
•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
兼容性
悲观锁
•读已提交能力
锁机制

支持多活架构



容灾和多活方案

同城方案

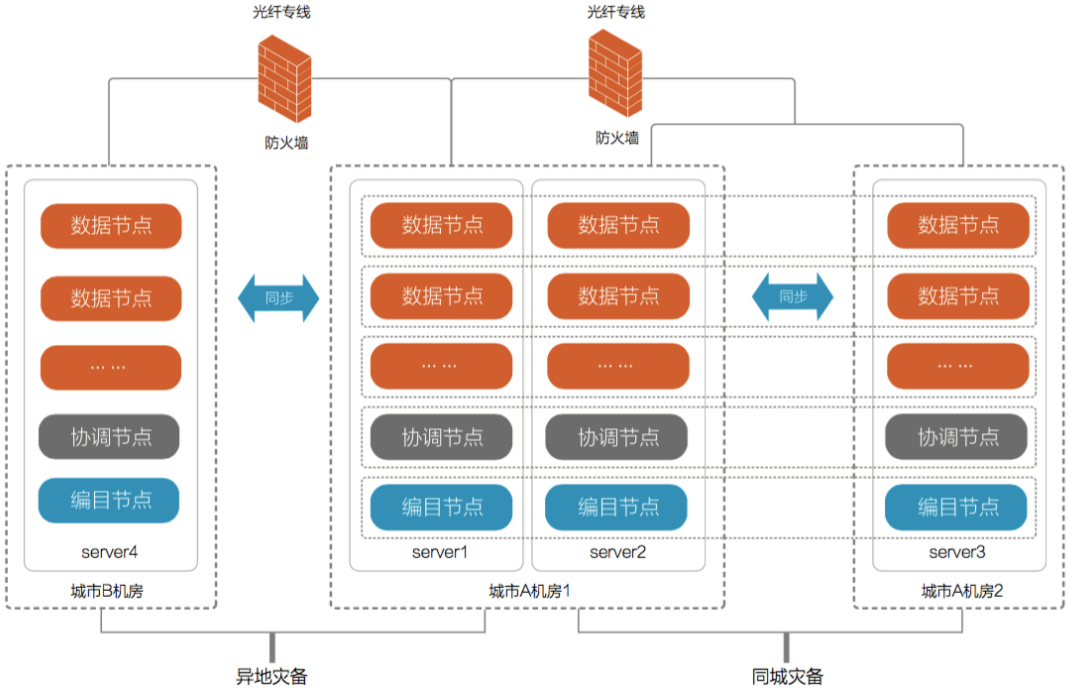
1、主备机房使用可靠高速光纤直连
2、每个分区主节点在主中心
3、平时使用强一致同步策略保障数据不丢
4、故障发生时使用takeover工具进行集群分离，备集群独立运行
5、故障恢复后使用merge工具进行集群合并

双活方案

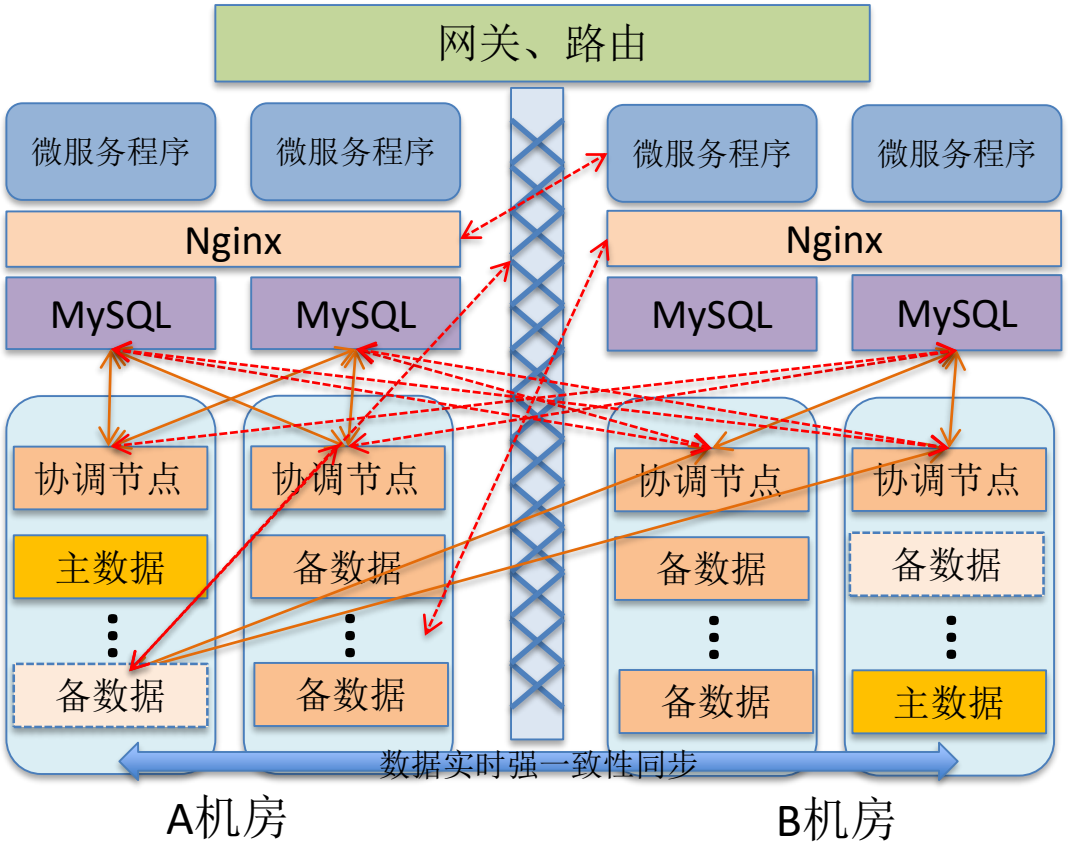
1、应用程序直连本地数据中心数据库协调节点
2、应用程序不需要关注底层数据存储主备中心复制和通讯策略

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



某省级农信核心生产环境中的多活数据中心方案



读写分离原则

- 微服务程序优先连接本机房Nginx服务，只有当本机房Nginx无法提供服务后，再去连接同城机房Nginx服务
- Nginx服务只连接本机房MySQL服务，一个机房拥有多个MySQL服务，确保SQL服务高可用
- 如果一个机房所有MySQL服务均停止服务，则该机房Nginx服务也会停止，微服务程序自动选择同城机房Nginx服务进行连接
- MySQL优先连接本机房协调节点，避免请求在同城机房中交叉访问
- 协调节点在执行写入操作时，自动路由数据库分区的主节点，执行操作
- 协调节点在执行查询操作时，优先选择本机房数据节点进行访问，避免请求在同城机房中交叉访问

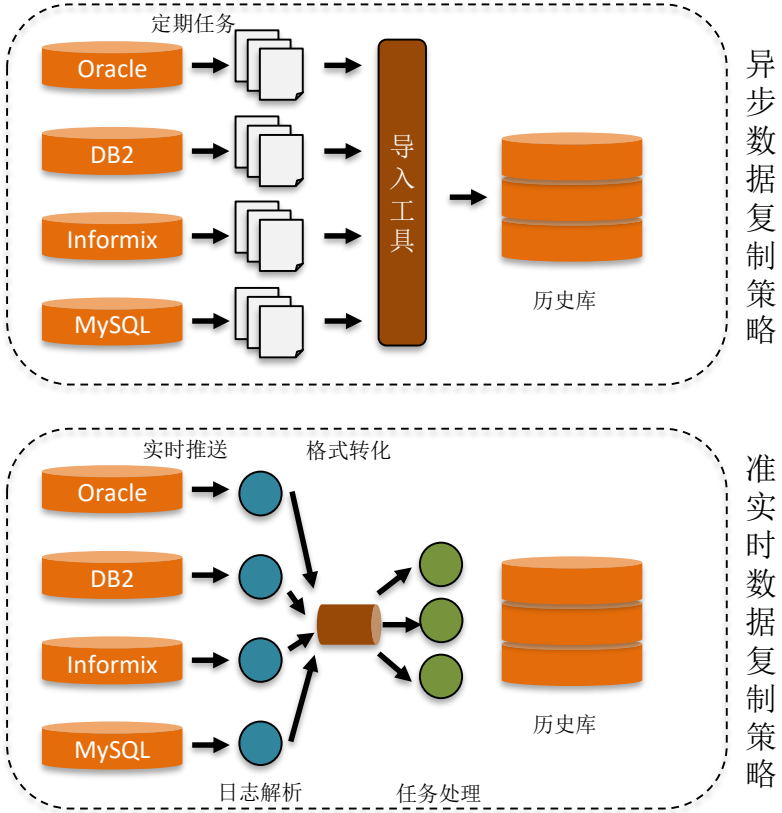
异构数据源复制策略

异步数据复制策略

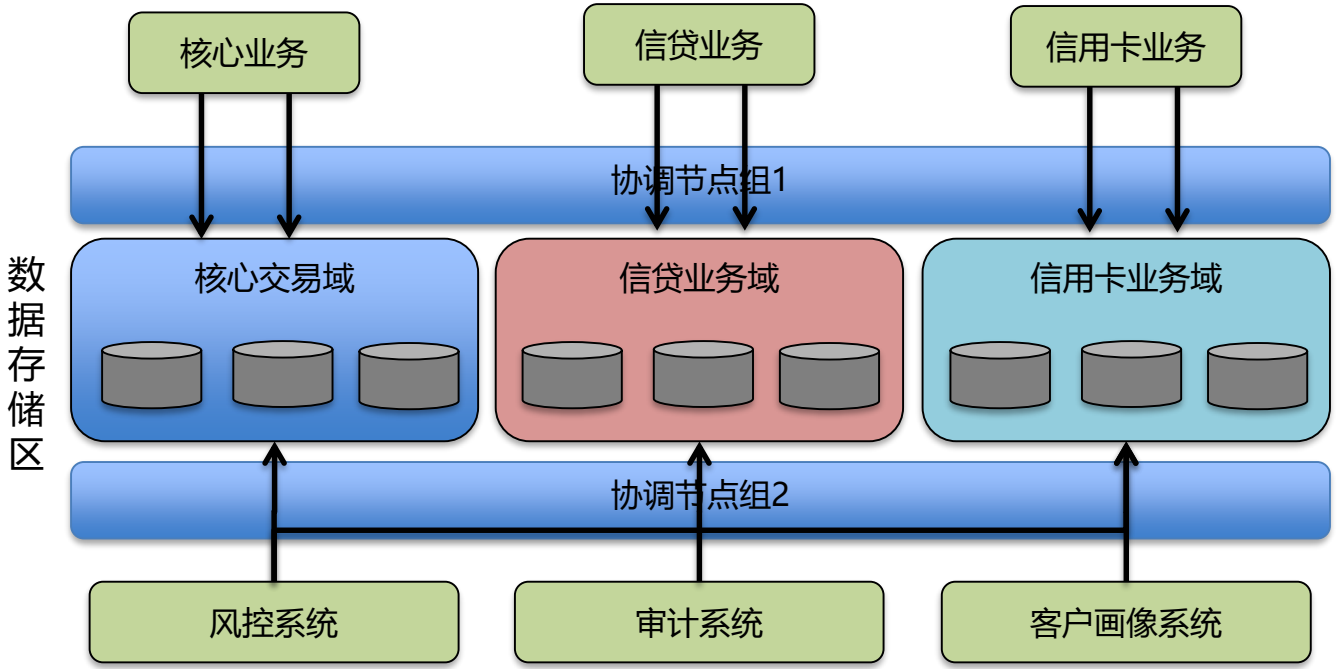
1、使用JSON或CSV格式定期将异构数据源的原始数据进行导出为文本文件
2、通过FTP等方式将文件传输至巨杉数据库的客户端
3、通过sdbimprt工具将文本文件导入巨杉数据库
4、满足异构数据源T+1的数据复制策略，简单可靠

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

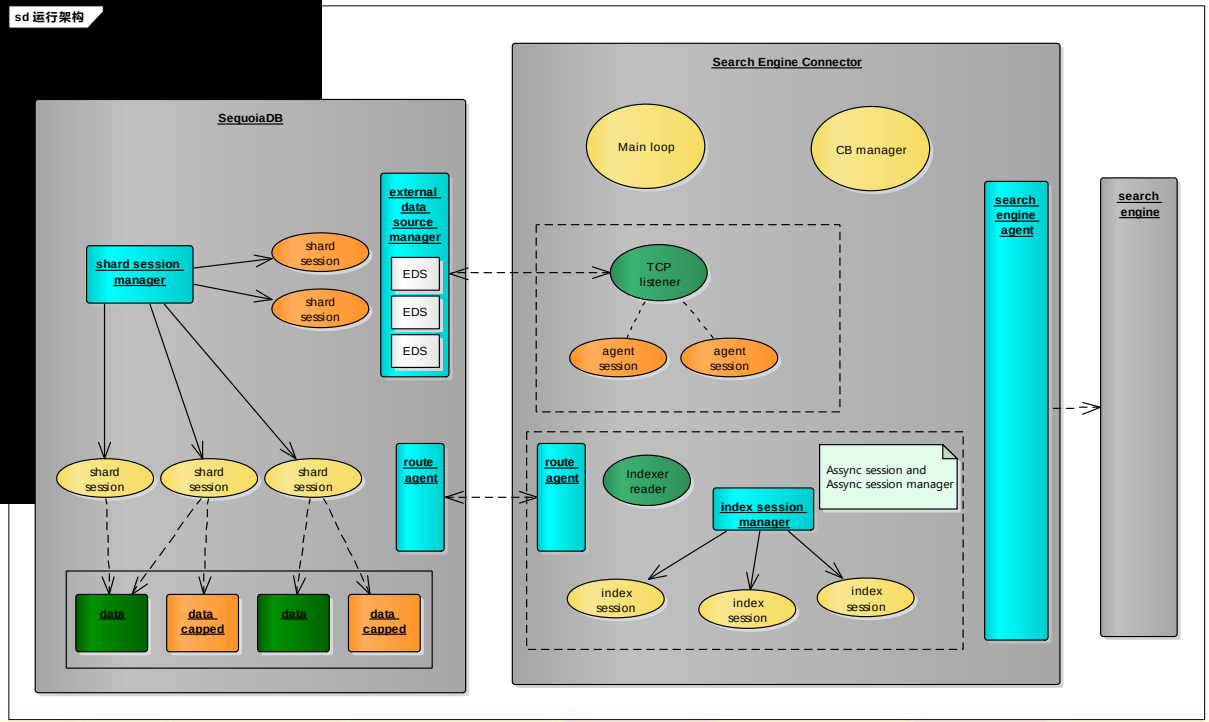


多租户与业务隔离

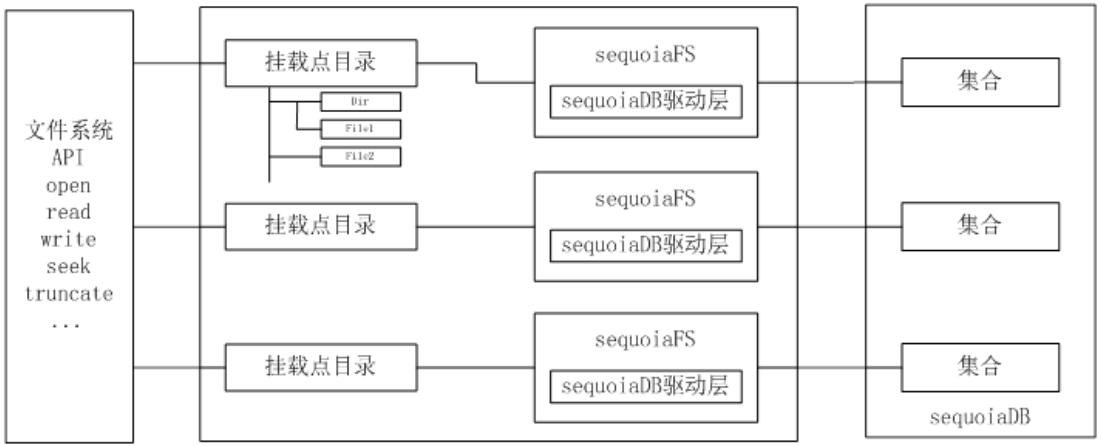

核心业务
信贷业务
信用卡业务
协调节点组1
数据存储区
核心交易域
信贷业务域
信用卡业务域
协调节点组2
风控系统
审计系统
客户画像系统

全文检索索引支持

```
> db.megacorp.employee.find({"":{"$Text":{"query":{"match":{"about" : "rock climbing"}}}}})
{
  "_id": {
    "$oid": "5af000ce7c8c33276e000000"
  },
  "first_name": "Jane",
  "last_name": "Smith",
  "age": 32,
  "about": "I like to collect rock albums",
  "interests": [
    "music"
  ]
}
{
  "_id": {
    "$oid": "5af000d57c8c33276e000002"
  },
  "first_name": "Jim",
  "last_name": "Green",
  "age": 20,
  "about": "I like to go rock climbing very much",
  "interests": [
    "walk"
  ]
}
Return 2 row(s).
Takes 0.111795s.
```

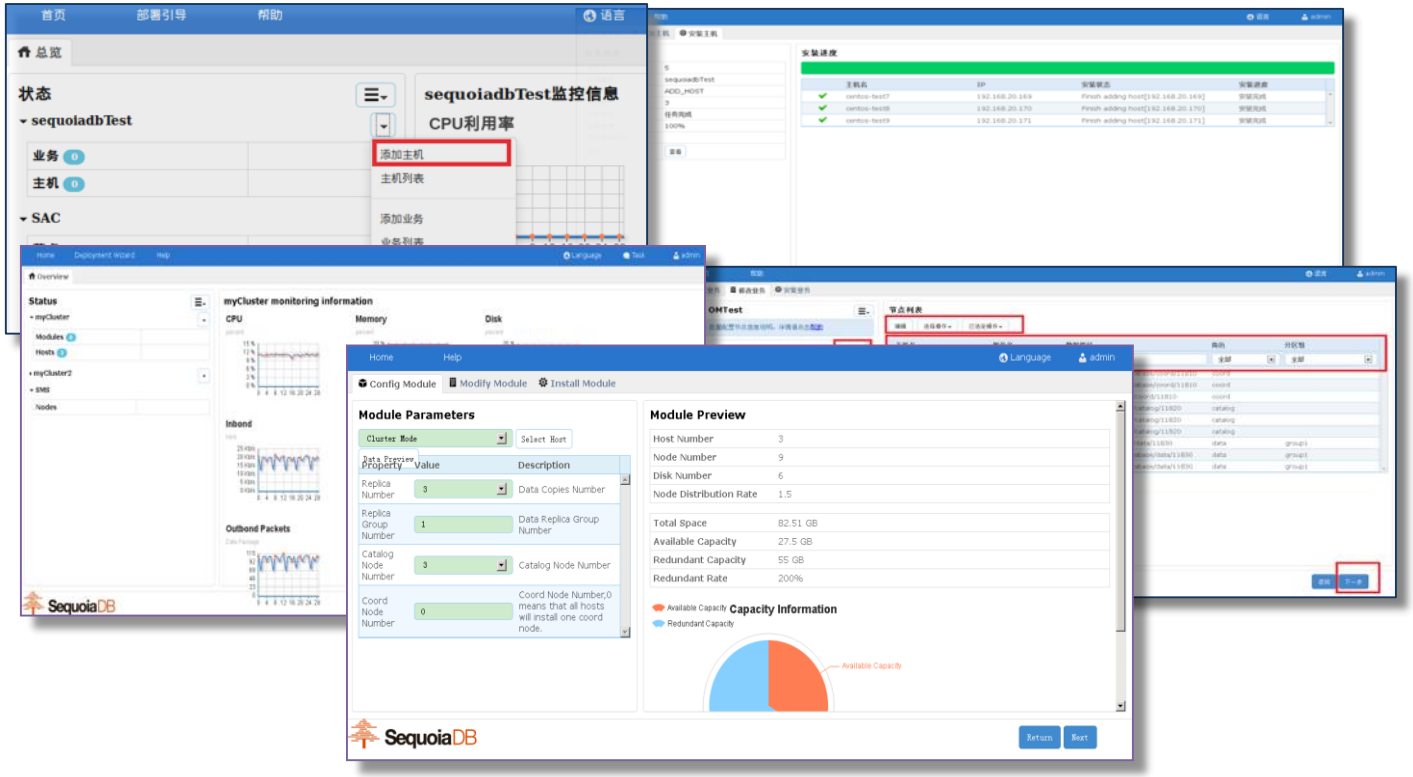


对象存储S3+文件系统支持




```
yuweixing@ubuntu-devywx:/opt/sequoiadb$ cd mountpoint/
yuweixing@ubuntu-devywx:/opt/sequoiadb/mountpoint$ ls
yuweixing@ubuntu-devywx:/opt/sequoiadb/mountpoint$ touch testfile
yuweixing@ubuntu-devywx:/opt/sequoiadb/mountpoint$ mkdir testdir
yuweixing@ubuntu-devywx:/opt/sequoiadb/mountpoint$ ls
testdir  testfile
yuweixing@ubuntu-devywx:/opt/sequoiadb/mountpoint$ echo "this is a testfile" >> testfile
yuweixing@ubuntu-devywx:/opt/sequoiadb/mountpoint$ cat testfile
this is a testfile
yuweixing@ubuntu-devywx:/opt/sequoiadb/mountpoint$
```

平台统一管理监控方案



- SequoiaDB提供全面的图形化集群监控功能
  - 集群健康状态
  - 资源使用状态
- SequoiaDB提供图形化操作界面
  - 集群部署、扩容
  - 集群管理
  - 数据CURD操作
  - 数据管理，切分、均衡等

性能测试

TPCC测试性能

测试环境

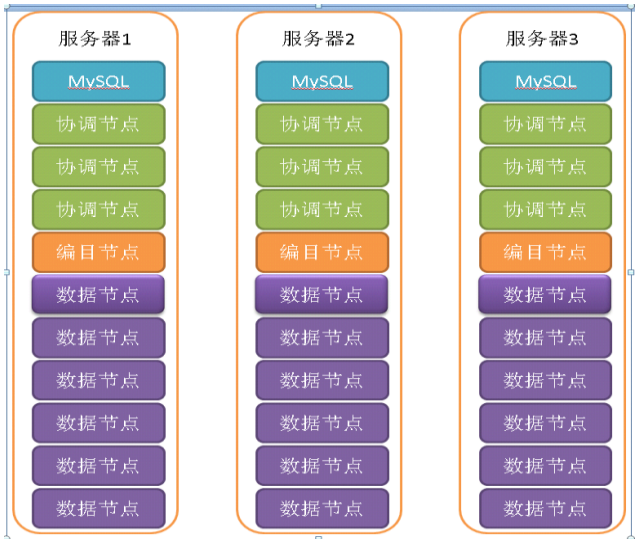
服务器数量

- 数据库服务器（3台）
- 应用压力服务器（1台）

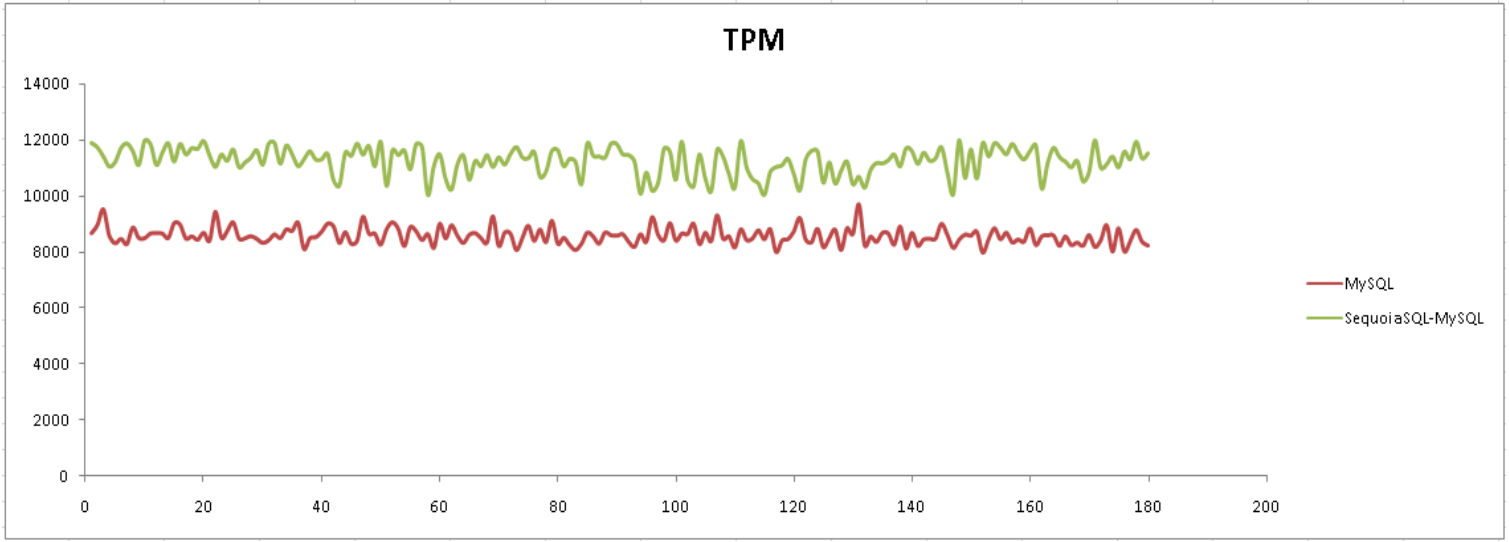
服务器配置

- CPU
  2 CORE * 24
- Memory
  256GB
- Disk
  6 * 3.6TB

名称	TPS
MySQL	8,558.33
SequoiaDB	11,163.00



TPCC测试性能


TPM
14000
12000
10000
8000
6000
4000
2000
0
0
20
40
60
80
100
120
140
160
180
200
MySQL
SequoiaSQL-MySQL

SysBench测试性能

测试环境

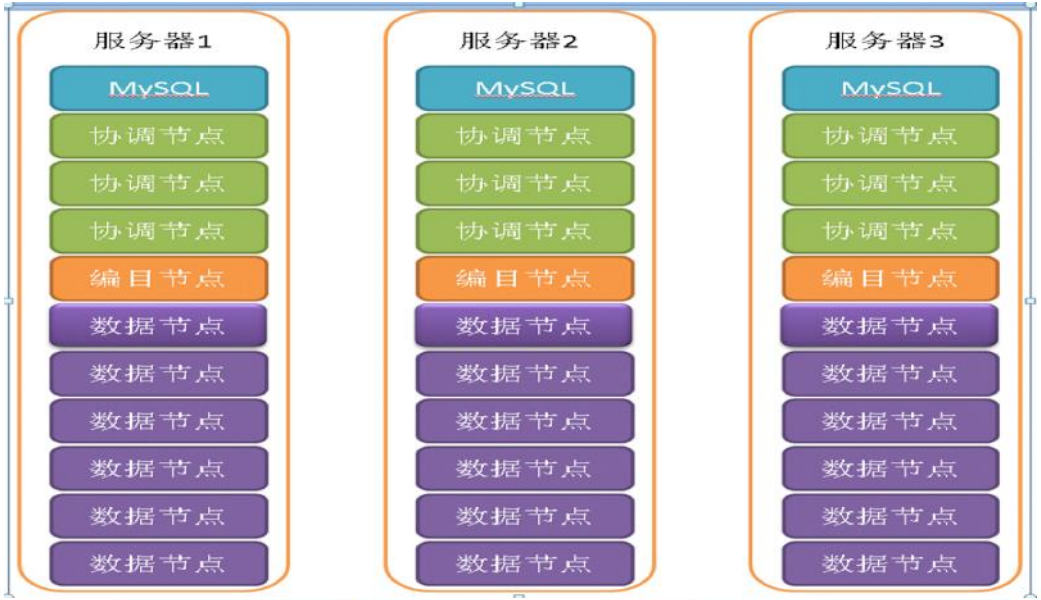
服务器数量

- 数据库服务器（3台）
- 应用压力服务器（1台）

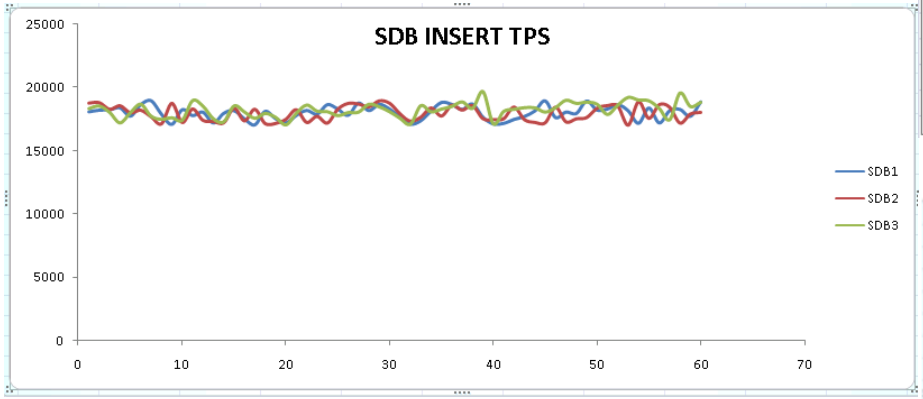
服务器配置

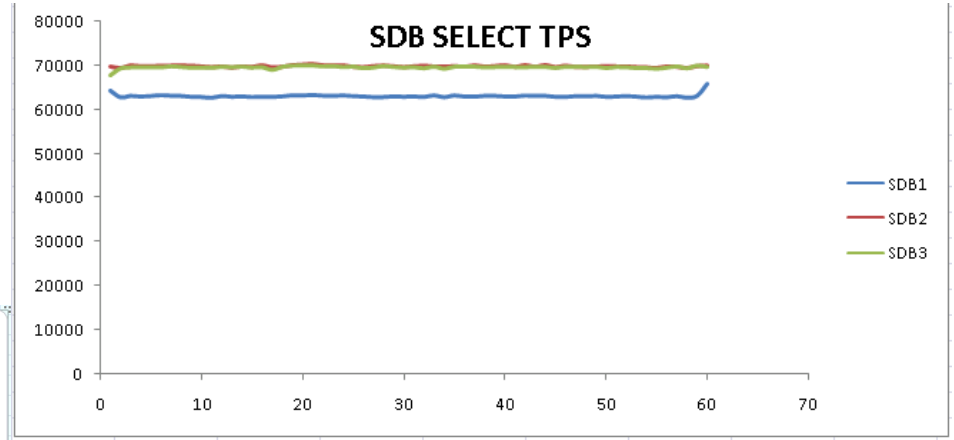
- CPU
  2 CORE * 24
- Memory
  256GB
- Disk
  6 * 3.6TB

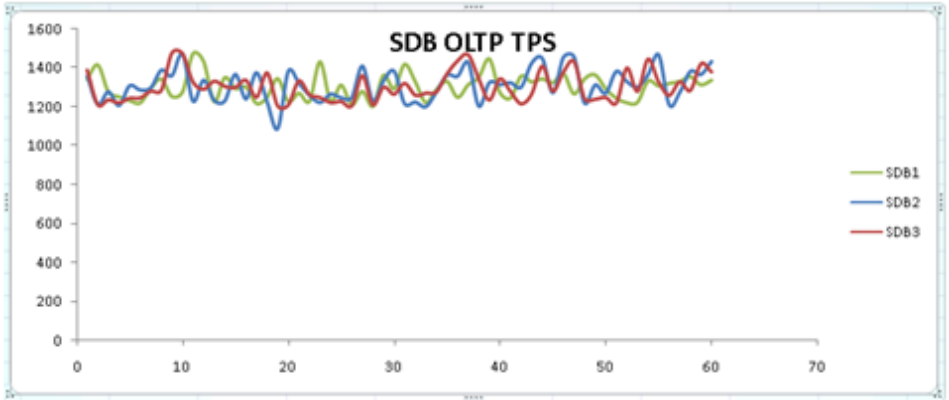
名称	事务响应(ms)	TPS	成功率
Insert	5.28	54,513.58	100%
Select	1.42	202,886.00	100%
OLTP	68.74	4,198.29	100%



SysBench测试性能







MySQL兼容样例

SequoiaDB – MySQL特性：MySQL语法兼容

使用原生MySQL解析引擎
存储引擎从InnoDB设置为SequoiaDB

```
+--------------------+---------+-----------------------------------------------------------------+--------------+------+------------+
| Engine             | Support | Comment                                                         | Transactions | XA   | Savepoints |
+--------------------+---------+-----------------------------------------------------------------+--------------+------+------------+
| InnoDB             | DEFAULT | Supports transactions, row-level locking, and foreign keys      | YES          | YES  | YES        |
| MRG_MYISAM         | YES     | Collection of identical MyISAM tables                           | NO           | NO   | NO         |
| MEMORY             | YES     | Hash based, stored in memory, useful for temporary tables       | NO           | NO   | NO         |
| BLACKHOLE          | YES     | /dev/null storage engine (anything you write to it disappears)  | NO           | NO   | NO         |
| MyISAM             | YES     | MyISAM storage engine                                           | NO           | NO   | NO         |
| CSV                | YES     | CSV storage engine                                              | NO           | NO   | NO         |
| SequoiaDB          | YES     | SequoiaDB storage engine. Sequoiadb: 2.9, Plugin:123456 Debug   | YES          | NO   | NO         |
| PERFORMANCE_SCHEMA | YES     | Performance Schema                                              | NO           | NO   | NO         |
| FEDERATED          | NO      | Federated MySQL storage engine                                  | NULL         | NULL | NULL       |
| ARCHIVE            | YES     | Archive storage engine                                          | NO           | NO   | NO         |
+--------------------+---------+-----------------------------------------------------------------+--------------+------+------------+
10 rows in set (0.00 sec)
```

```
+------+-----------+---------+------------+------+----------------+--------------+-----------------+-----------+----------------+---------------------+-------------+------------+--------------------+----------+----------+
| Name | Engine    | Version | Row_format | Rows | Avg_row_length | Data_length  | Max_data_length | Index_length | Data_free | Auto_increment | Create_time         | Update_time | Check_time | Collation          | Checksum | Create_options | Comment |
+------+-----------+---------+------------+------+----------------+--------------+-----------------+--------------+-----------+----------------+---------------------+-------------+------------+--------------------+----------+----------------+---------+
| c11  | SequoiaDB |      10 | Fixed      | 10000 |           1024 | 107374182400 |  1099511627776 |   1073741824 |         0 |           NULL | NULL                | NULL        | NULL       | latin1_swedish_ci  |     NULL |                |         |
| c12  | SequoiaDB |      10 | Dynamic    | 10000 |           1024 | 107374182400 |  1099511627776 |   1073741824 |         0 |           NULL | NULL                | NULL        | NULL       | latin1_swedish_ci  |     NULL |                |         |
| c13  | SequoiaDB |      10 | Dynamic    | 10000 |           1024 | 107374182400 |  1099511627776 |   1073741824 |         0 |           NULL | NULL                | NULL        | NULL       | latin1_swedish_ci  |     NULL |                |         |
| c14  | SequoiaDB |      10 | Fixed      | 10000 |           1024 | 107374182400 |  1099511627776 |   1073741824 |         0 |           NULL | NULL                | NULL        | NULL       | latin1_swedish_ci  |     NULL |                |         |
| c15  | SequoiaDB |      10 | Fixed      | 10000 |           1024 | 107374182400 |  1099511627776 |   1073741824 |         0 |           NULL | NULL                | NULL        | NULL       | latin1_swedish_ci  |     NULL |                |         |
| c16  | SequoiaDB |      10 | Fixed      | 10000 |           1024 | 107374182400 |  1099511627776 |   1073741824 |         0 |           NULL | NULL                | NULL        | NULL       | latin1_swedish_ci  |     NULL |                |         |
| c17  | SequoiaDB |      10 | Dynamic    | 10000 |           1024 | 107374182400 |  1099511627776 |   1073741824 |         0 |           NULL | NULL                | NULL        | NULL       | latin1_swedish_ci  |     NULL | {cl_options:{ShardingKey:{a:1,b:-1},ShardingType:"range"}} |         |
| c18  | SequoiaDB |      10 | Fixed      | 10000 |           1024 | 107374182400 |  1099511627776 |   1073741824 |         0 |           NULL | NULL                | NULL        | NULL       | latin1_swedish_ci  |     NULL |                |         |
| t1   | InnoDB    |      10 | Dynamic    |     8 |           2048 |        16384 |               0 |            0 |         0 |           NULL | 2018-01-30 15:32:42 | NULL        | NULL       | latin1_swedish_ci  |     NULL |                |         |
+------+-----------+---------+------------+------+----------------+--------------+-----------------+--------------+-----------+----------------+---------------------+-------------+------------+--------------------+----------+----------------+---------+
```

SequoiaDB – MySQL特性：MySQL语法兼容

- 使用原生MySQL解析引擎
  - 100%支持MySQL语法
  - CRUD操作完美支持

```
sequoiadb — root@iz2ze1gznj6m27tp7xuisrz:~ — ssh root@39.107.73.37 — 86×31

mysql> insert into cl values (3,103,"SequoiaDB test1");
Query OK, 1 row affected (0.51 sec)

mysql> select * from cl;
+---+-----+-----------------+
| a | b   | c               |
+---+-----+-----------------+
| 1 | 101 | SequoiaDB test  |
| 2 | 102 | SequoiaDB test  |
| 3 | 103 | SequoiaDB test1 |
+---+-----+-----------------+
3 rows in set (0.00 sec)

mysql> update cl set b=104 where a=3;
Query OK, 1 row affected (0.50 sec)
Rows matched: 1  Changed: 1  Warnings: 0

mysql> delete from cl where a=2;
Query OK, 1 row affected (0.00 sec)

mysql> select * from cl;
+---+-----+-----------------+
| a | b   | c               |
+---+-----+-----------------+
| 1 | 101 | SequoiaDB test  |
| 3 | 104 | SequoiaDB test1 |
+---+-----+-----------------+
2 rows in set (0.00 sec)

mysql>
```

SequoiaDB – MySQL特性：MySQL语法兼容

- 使用原生MySQL解析引擎
  - 支持多表关联
  - 支持跨表跨节点事务操作

```
sequoiadb — root@iz2ze1gznj6m27tp7xuisrz:~ — ssh root@39.107.73.37 — 88×37
mysql> begin;
Query OK, 0 rows affected (0.00 sec)

mysql> insert into cl1 values (2,202);
Query OK, 1 row affected (0.51 sec)

mysql> insert into cl values(2,102,"SequoiaDB test");
Query OK, 1 row affected (0.51 sec)

mysql> select * from cl,cl1 where cl.a=cl1.a;
+---+-----+----------------+------+------+
| a | b   | c              | a    | b    |
+---+-----+----------------+------+------+
| 1 | 101 | SequoiaDB test |    1 |  101 |
| 2 | 102 | SequoiaDB test |    2 |  202 |
+---+-----+----------------+------+------+
2 rows in set (1.11 sec)

mysql> rollback;
Query OK, 0 rows affected (0.41 sec)

mysql> select * from cl,cl1 where cl.a=cl1.a;
+---+-----+----------------+------+------+
| a | b   | c              | a    | b    |
+---+-----+----------------+------+------+
| 1 | 101 | SequoiaDB test |    1 |  101 |
+---+-----+----------------+------+------+
1 row in set (0.41 sec)

mysql>
```

SequoiaDB – MySQL特性：MySQL语法兼容

- 使用原生MySQL解析引擎
  - 支持创建视图
  - 支持存储过程

sequoiadb — root@iz2ze1gznj6m27tp7xuisrz:~ — ssh root@39.107.73.37 — 8...

```
mysql> create view v1 as select cl.a,cl.c,cl1.b from cl, cl1 where cl.a=cl1.a;
Query OK, 0 rows affected (0.01 sec)

mysql> select * from v1;
+---+----------------+------+
| a | c              | b    |
+---+----------------+------+
| 1 | SequoiaDB test |  101 |
+---+----------------+------+
1 row in set (1.11 sec)

mysql> delimiter //
mysql> create procedure delete_match() begin delete from cl where a=1; end//
Query OK, 0 rows affected (0.11 sec)

mysql> delimiter ;
mysql> call delete_match();
Query OK, 1 row affected (1.00 sec)

mysql> select * from v1;
Empty set (1.01 sec)

mysql> select * from cl;
+---+-----+-----------------+
| a | b   | c               |
+---+-----+-----------------+
| 3 | 104 | SequoiaDB test1 |
+---+-----+-----------------+
1 row in set (0.00 sec)

mysql>
```

SequoiaDB – MySQL特性：MySQL语法兼容

- 使用原生MySQL解析引擎
  - 支持索引
  - 支持访问计划

sequoiadb — root@iz2ze1gznj6m27tp7xuisrz:~ — ssh root@39.107.73.37 — 171×44

```
mysql>
mysql>
mysql>
mysql> select * from cl;
+---+-----+-----------------+
| a | b   | c               |
+---+-----+-----------------+
| 3 | 104 | SequoiaDB test1 |
| 1 | 101 | SequoiaDB test  |
| 2 | 102 | SequoiaDB test  |
| 4 | 104 | SequoiaDB test  |
+---+-----+-----------------+
4 rows in set (0.01 sec)

mysql> explain select * from cl where b=101;
+----+-------------+-------+------------+------+---------------+------+---------+------+------+----------+-----------------------------------------------------------------+
| id | select_type | table | partitions | type | possible_keys | key  | key_len | ref  | rows | filtered | Extra                                                           |
+----+-------------+-------+------------+------+---------------+------+---------+------+------+----------+-----------------------------------------------------------------+
|  1 | SIMPLE      | cl    | NULL       | ALL  | NULL          | NULL | NULL    | NULL |    4 |    25.00 | Using where with pushed condition (`cs`.`cl`.`b` = 101) |
+----+-------------+-------+------------+------+---------------+------+---------+------+------+----------+-----------------------------------------------------------------+
1 row in set, 1 warning (0.00 sec)

mysql> alter table cl add index idx_b(b);
Query OK, 0 rows affected (0.01 sec)
Records: 0  Duplicates: 0  Warnings: 0

mysql> explain select * from cl where b=101;
+----+-------------+-------+------------+------+---------------+-------+---------+-------+------+----------+-------+
| id | select_type | table | partitions | type | possible_keys | key   | key_len | ref   | rows | filtered | Extra |
+----+-------------+-------+------------+------+---------------+-------+---------+-------+------+----------+-------+
|  1 | SIMPLE      | cl    | NULL       | ref  | idx_b         | idx_b | 4       | const |    1 |   100.00 | NULL  |
+----+-------------+-------+------------+------+---------------+-------+---------+-------+------+----------+-------+
1 row in set, 1 warning (0.00 sec)

mysql> explain select * from cl,cl1 where cl.a=cl1.a and cl.b=101;
+----+-------------+-------+------------+------+---------------+-------+---------+----------+------+----------+-------+
| id | select_type | table | partitions | type | possible_keys | key   | key_len | ref      | rows | filtered | Extra |
+----+-------------+-------+------------+------+---------------+-------+---------+----------+------+----------+-------+
|  1 | SIMPLE      | cl    | NULL       | ref  | PRIMARY,idx_b | idx_b | 4       | const    |    1 |   100.00 | NULL  |
|  1 | SIMPLE      | cl1   | NULL       | ref  | idx_a         | idx_a | 5       | cs.cl.a  |    2 |   100.00 | NULL  |
+----+-------------+-------+------------+------+---------------+-------+---------+----------+------+----------+-------+
2 rows in set, 1 warning (0.02 sec)

mysql>
```

谢谢

SequoiaDB 官网：
www.sequoiadb.com

Github项目地址：
SequoiaDB/SequoiaDB
SequoiaDB/sequoiasql-mysql

加入巨杉社区