



PB级多源异构数据 管理实践

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

中国民生银行
CHINA MINSHENG BANK

广发银行 | CGB

中国银行
BANK OF CHINA

恒丰银行
HENG FENG BANK

渤海银行
CHINA BOHAI BANK

PICC
中国人保财险

广东农信
GDRC

广州农商银行
GUANGZHOU RURAL COMMERCIAL BANK

bsb 包商银行

锦州银行
BANK OF JINZHOU

贵州农信
GZRC

吉林省农村信用社
JILIN PROVINCE RURAL CREDIT BANK

四川省农村信用社
SICHUAN RURAL CREDIT UNION

东莞农村商业银行
DRC Bank

东莞银行
BANK OF DONGGUAN

广发证券
GF SECURITIES

CSDC
中国结算

广州银行
BANK OF GUANGZHOU

嘉实基金
Harvest Fund

张家口银行
BANK OF ZHANGJIAKOU

朝阳银行
BANK OF CHAOYANG

鞍山银行
BANK OF ANSHAN

华商银行

营口沿海银行

承德银行
BANK OF CHENGDE

阳泉市商业银行
YangQuan City Commercial Bank

CEC
中国电子
CHINA ELECTRONICS

中国税务
国家税务总局

中国移动
China Mobile

中国电信
CHINA TELECOM

广州市人民政府

爱同科技
AGREE TECHNOLOGY

中国税务

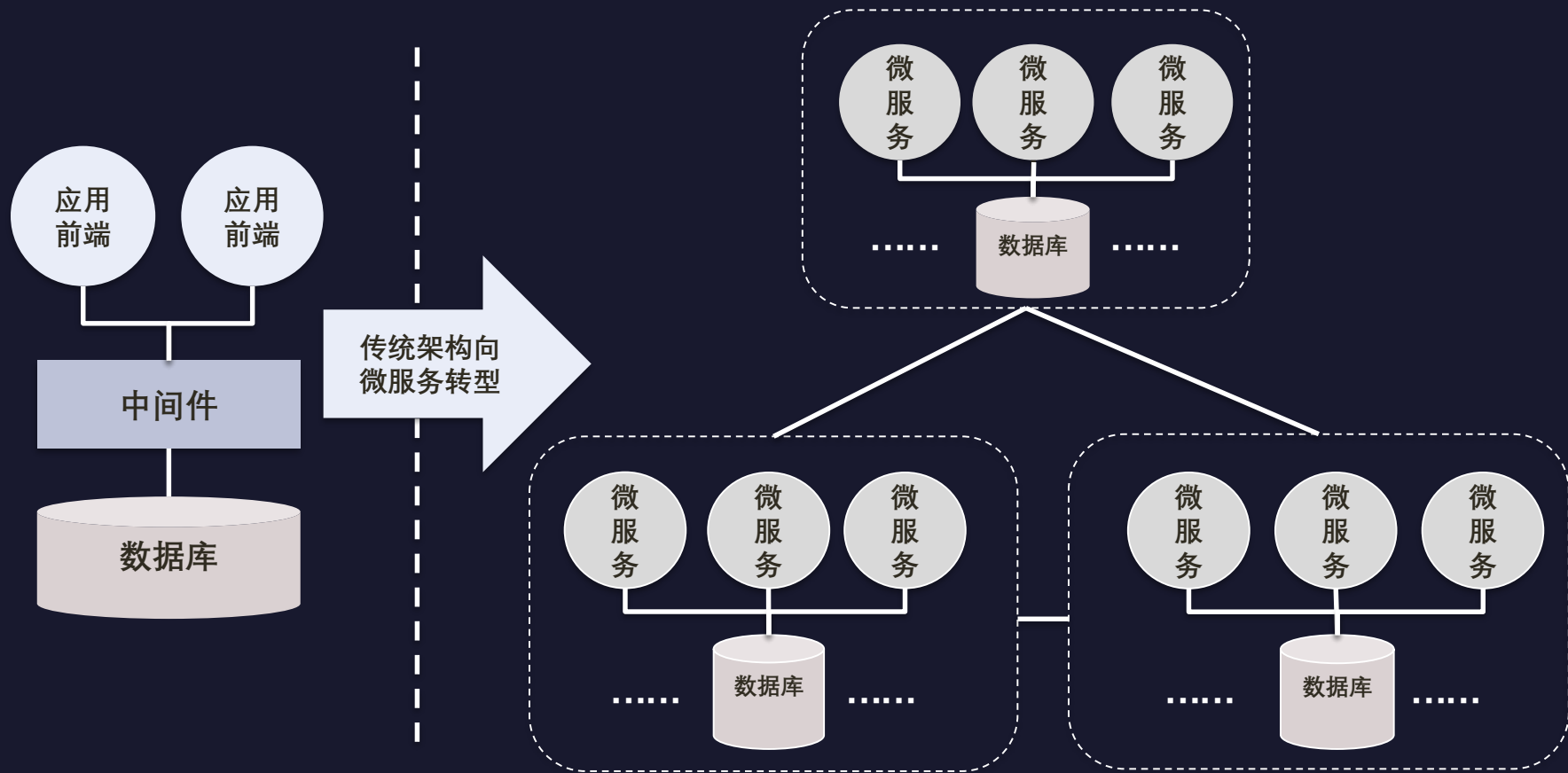
去哪儿
Qunar.Com
聪明你的旅行

途牛
tuniu.com

应用程序开发 面临怎样的趋势



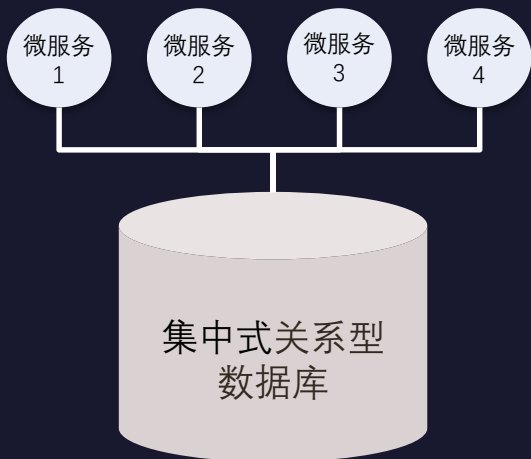
应用程序开发从烟囱式架构向分布式的转型



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



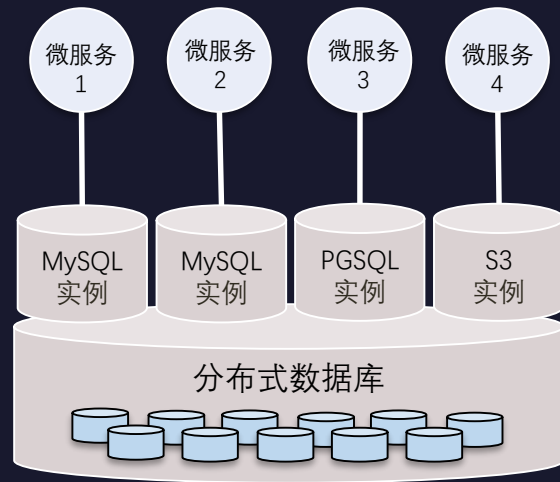
集中式存储

- 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

新技术前瞻性

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

HTAP

混合事务和分析场景，适应更多数据应用需求

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

ACID的支持

事务、一致性等，处理OLTP

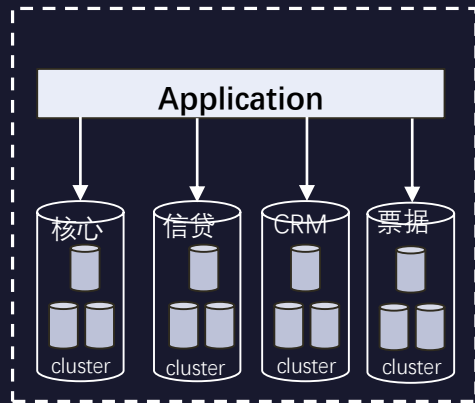
SQL完整支持

MySQL/PostgreSQL语法的完整兼容

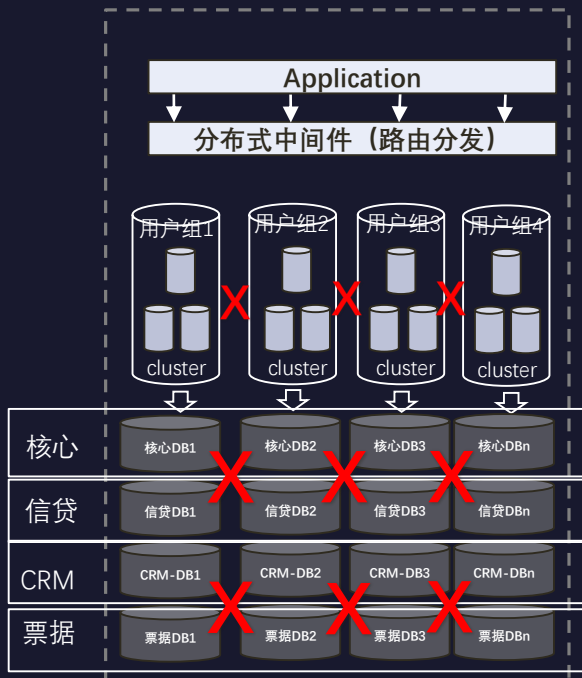
传统技术兼容性

分布式交易型数据库技术发展体系

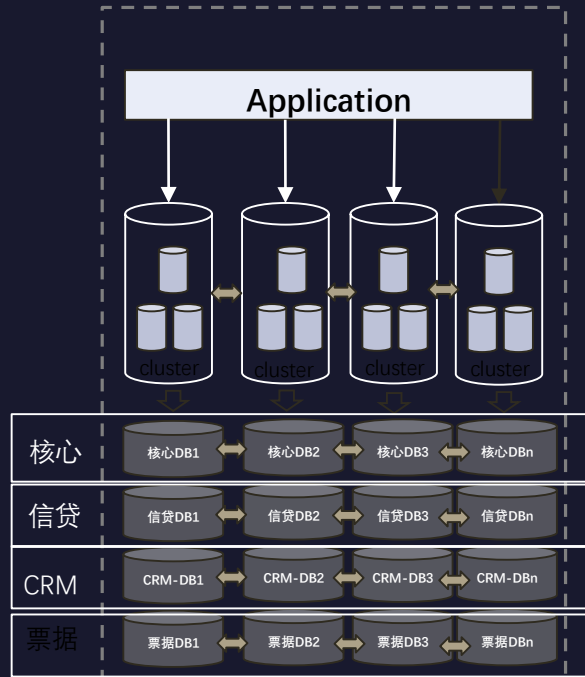
应用垂直分库



分库分表



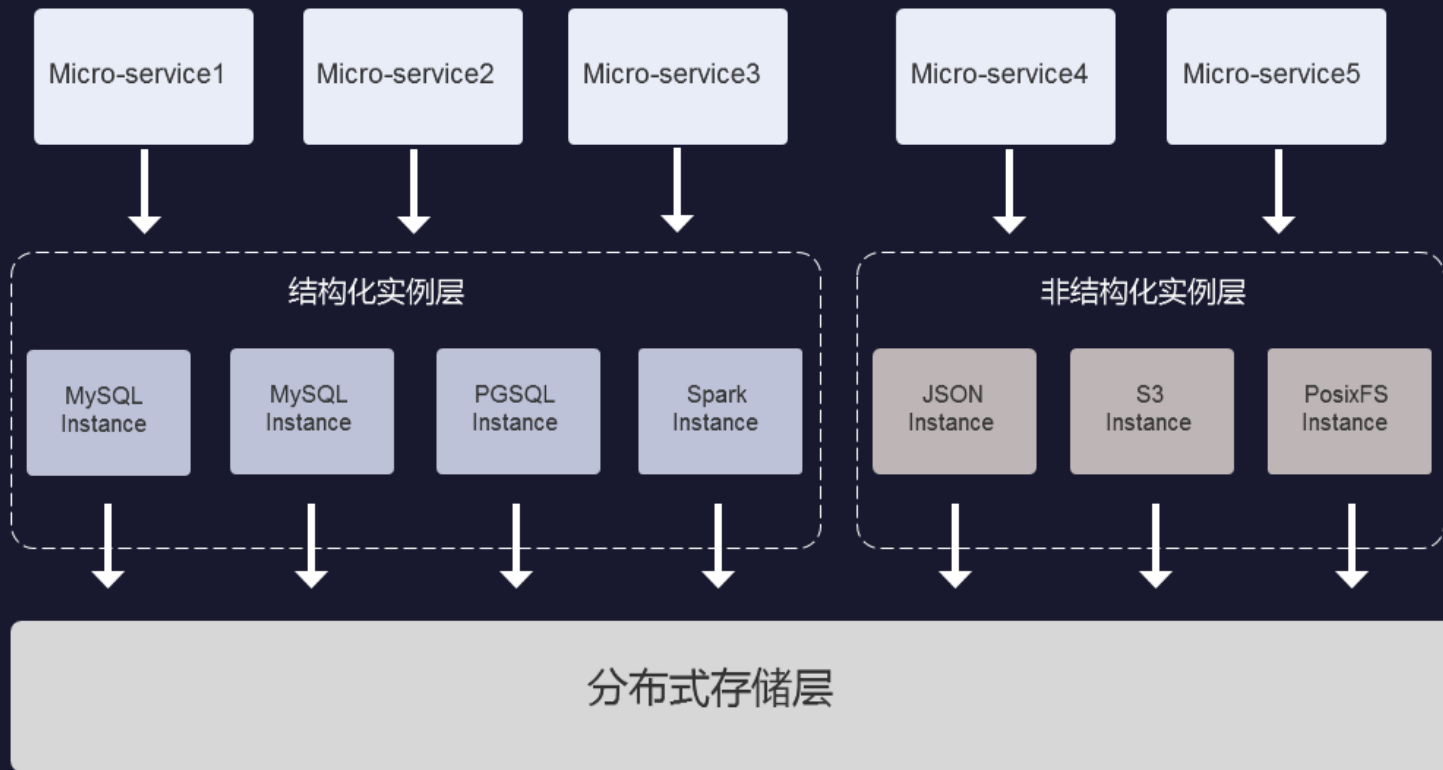
原生分布式数据库



优势

劣势

垂直分库	分库分表	原生分布式数据库
<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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
<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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



类型	计算引擎	用途
结构化数据访问	SequoiaSQL-MySQL	交易型应用场景，精准查询
	SequoiaSQL-PostgreSQL	交易型应用场景，数据中台应用场景，中等数据量关联聚合查询
	SparkSQL	离线统计分析应用场景，大数据量关联聚合查询
	SequoiaDB JSON API	交易型应用场景，单表增删改查
半结构化数据访问	SequoiaDB JSON API	偏互联网的新型应用场景，半结构化数据功能优先，速度优先
非结构化数据访问	SequoiaDB JSON API	非结构化数据最高速增删改查，适用于影像平台、内容管理、非结构化数据存储
	SequoiaS3	兼容 Amazon S3 对象存储接口，把巨杉数据库当做对象存储使用
	SequoiaFS	兼容 POSIX fuse 文件系统接口，把巨杉数据库当做网络文件系统使用



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

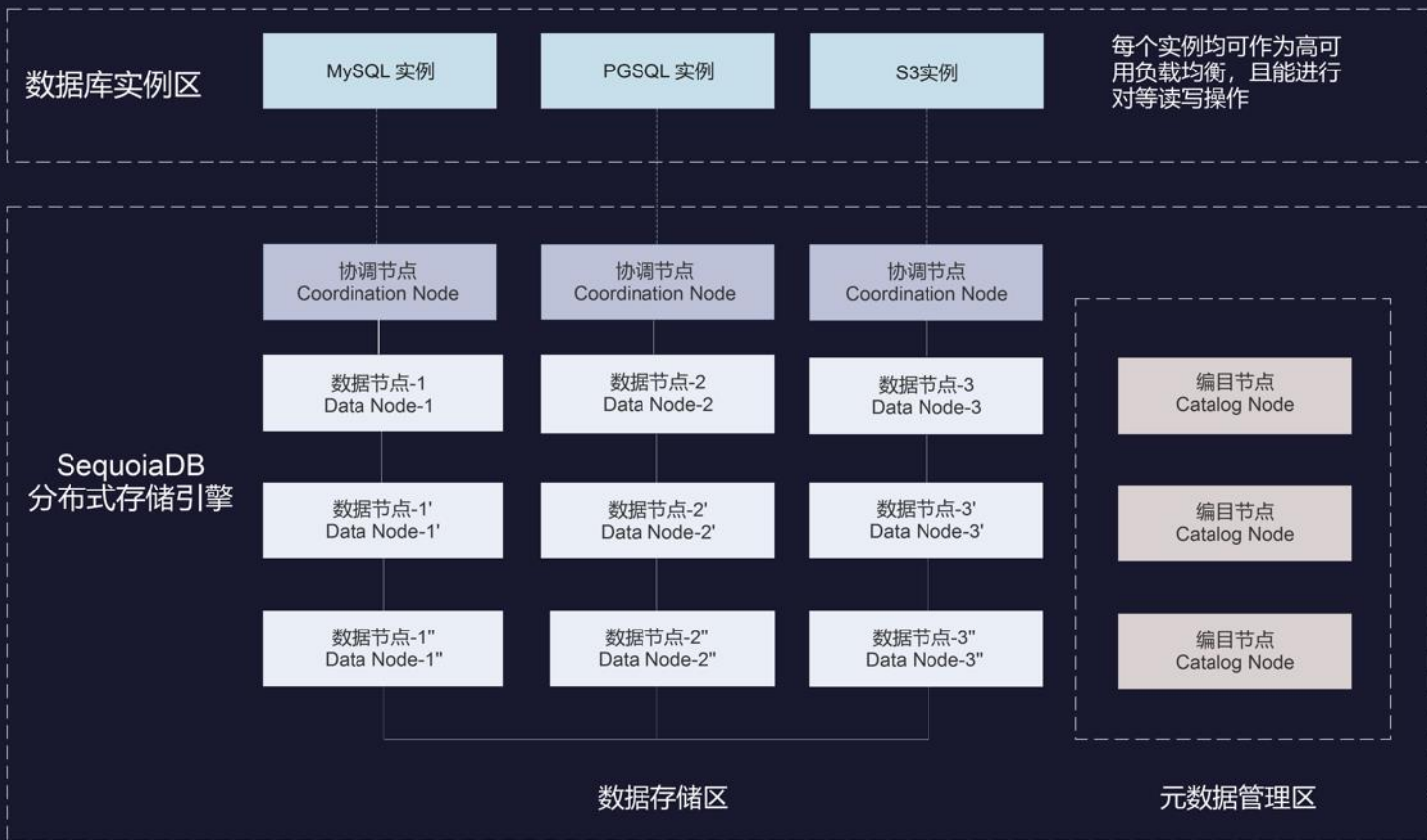
- 数据服务与高频只读类业务
- 提供比 Hbase 更加友好的开发接口以及更加简便的运维能力



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

“计算存储分离”架构



关系型

MySQL

PostgreSQL

SparkSQL

文档型

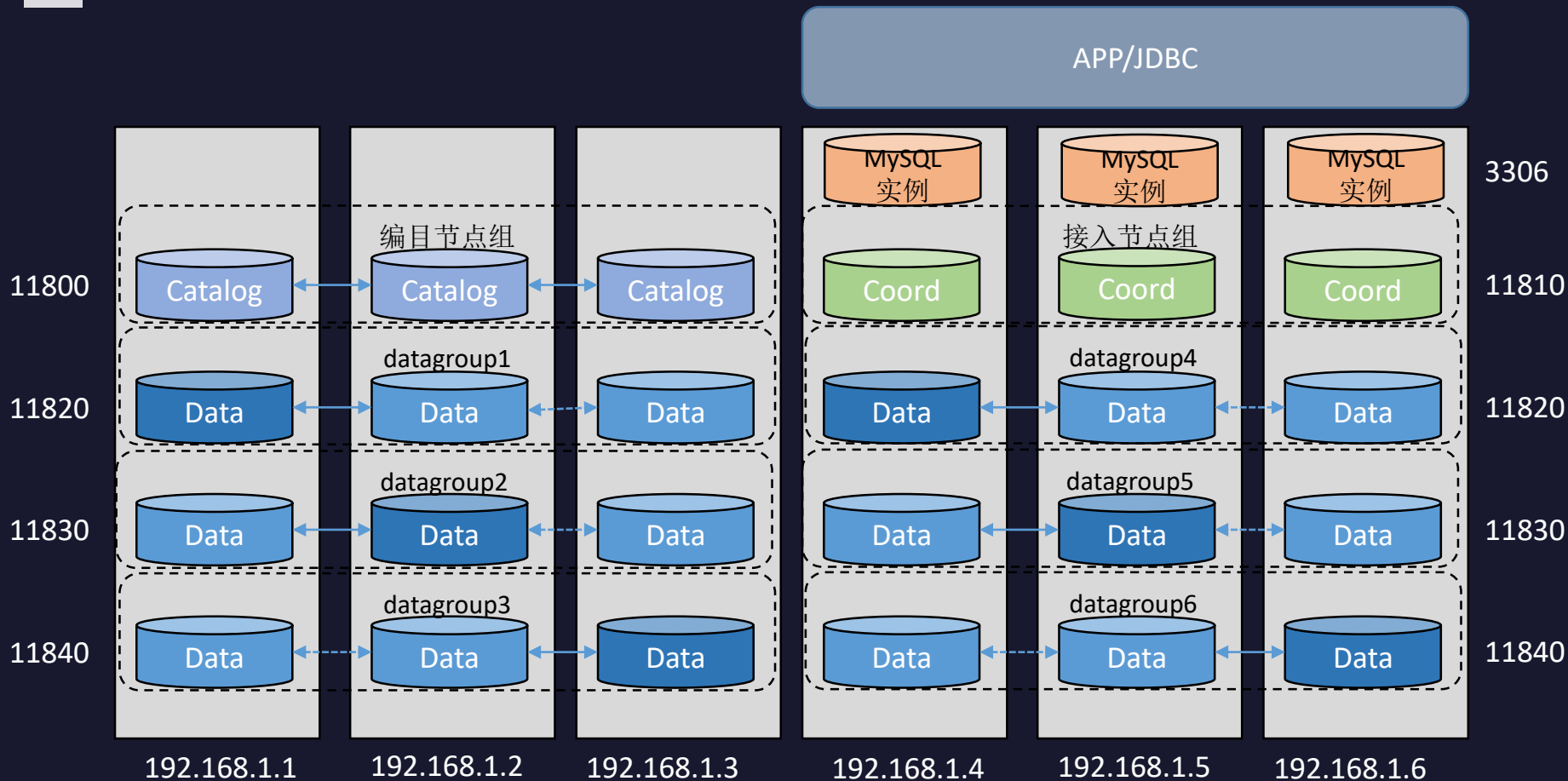
JSON

对象型

S3

Posix
文件系统

典型部署方式



3台服务器或以上	联机交易	数据中台服务	影像平台
CPU	2路48C	2路32C	2路24C
内存	256GB	256GB	128GB
磁盘	6 x 512GB SSD	12 x 2TB SAS	12 x 4TB SATA
网络	万兆网	万兆网	万兆网
操作系统	CentOS 7.4	CentOS 7.4	CentOS 7.4

Key Features



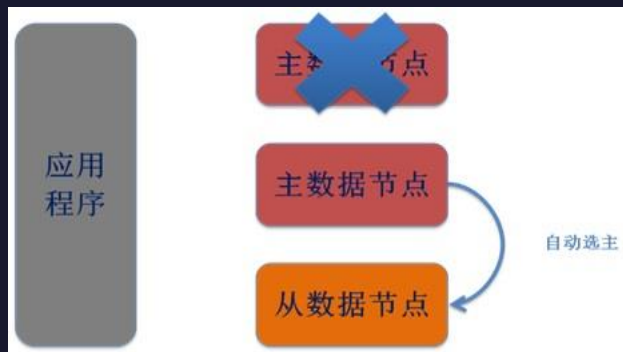
CategoriesID	MovieName	MovieDirector	MovieActor	MovieDesc	MovieData	MovieTime
6	何以笙箫默	无名	摩尼教	23颗角o	2012-02-22 0...	20
1	速度与激情	张艺谋	不可谓	科ouee	2012-02-18 0...	120
6	阿科未婚夫	无名	ndk	jkduw	2012-02-22 0...	20
1	井底蛙电脑	张艺谋	马拉	看到	2012-02-18 0...	120
6	洛带古镇	无名	农户	鸡窝	2012-02-22 0...	20
1	测试名字	张艺谋	潘长	摩尼教	2012-02-18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

TargetPartition = DHT (Row->PartitionKey)

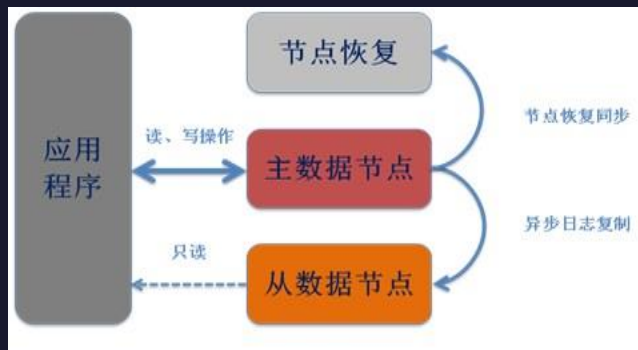
datagroup1

datagroup2

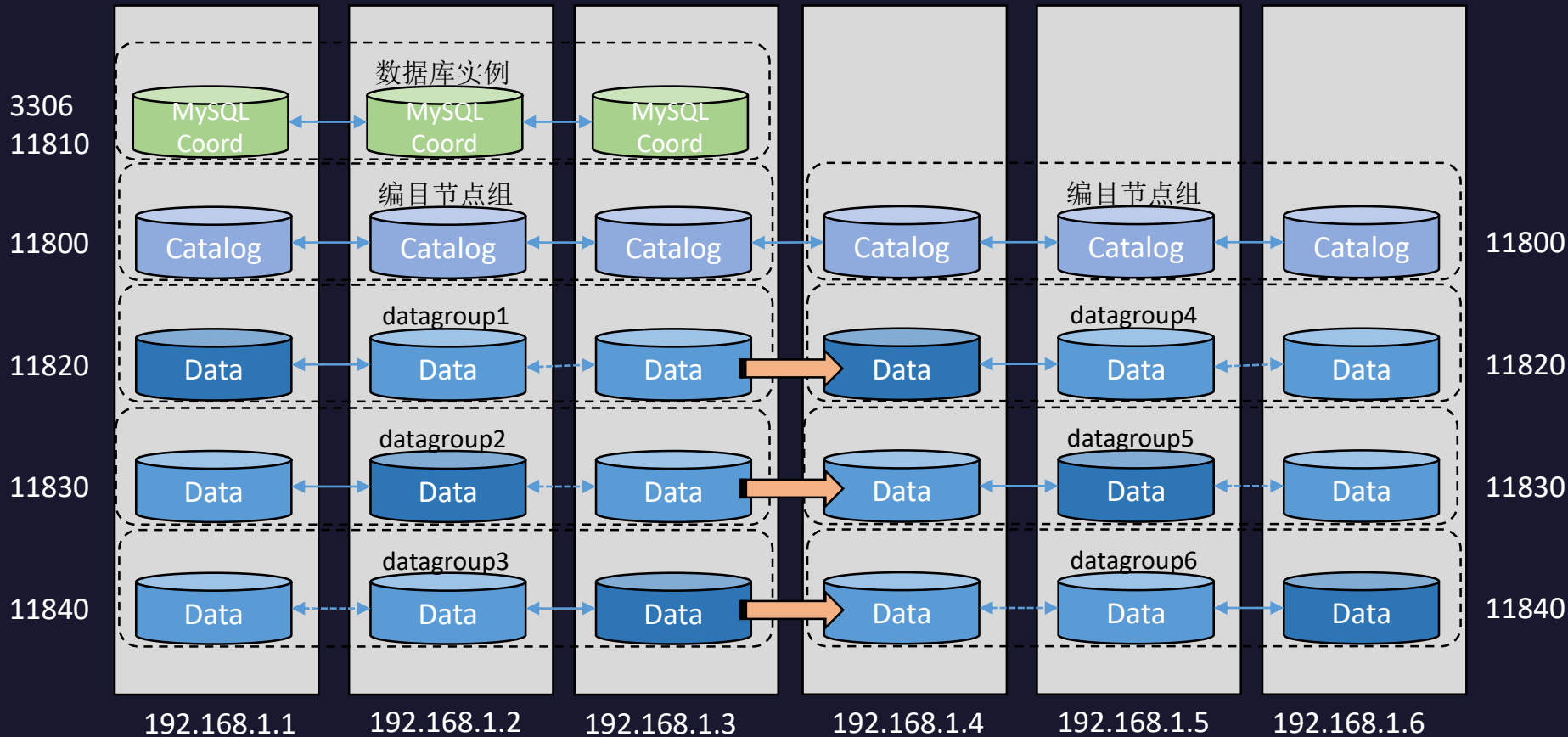
datagroup3



- 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



- 传统二段提交机制
- 保证数据跨节点一致性



二段提交
2PC



表设计原则

- 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



兼容性



锁机制

- 悲观锁
- MVCC读已提交能力

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



数据库实例 3

分布式存储 1

主机 3

创建实例

删除实例

重启实例

配置实例存储



鉴权 | 关联信息 | MySQLInstance
MySQL | 创建时间: 2019-06-07-08.49.05

MySQL | 创建时间: 2019-06-07-08.49.05

● 正常运行...

一共 1 台主机

CPU

0%

内存

21.91%

磁盘

18.82%

存储集群

MySQLInstance → SequoiaDB1



鉴权 | 关联信息 | MySQLInstance1
MySQL | 创建时间: 2019-06-07-08.50.07

● 正常运行...

一共 1 台主机

存储集群

集群

sdbDemo
Demo cluster

操作

一键部署

创建集群

删除集群

SequoiaDB SAC
admin

数据

监控

部署

策略

配置

MySQLInstance

◆ #	◆ 表名		◆ 存储引擎	◆ 行数	表大小	◆ 创建时间	◆ 描述
1	customer		SequoiaDB	60000	70.25 MB		
2	district		SequoiaDB	20	960 KB		
3	history		SequoiaDB	60038	14 MB		
4	item		SequoiaDB	100000	25.88 MB		
5	new_order		SequoiaDB	16940	3.75 MB		
6	oorder		SequoiaDB	60000	23.63 MB		
7	order_line		SequoiaDB	600152	218.5 MB		
8	stock		SequoiaDB	200000	143.5 MB		
9	warehouse		SequoiaDB	2	896 KB		

1 / 1

显示 9 条记录，一共 9 条

数据库列表:

hsdb

test

wordpress

information_schema

mysql

performance_schema

sys

操作

创建数据库

删除数据库

创建数据表

删除数据表

©2019 SequoiaDB. 版本: 系统时间: 17:44:00 系统状态: 良好

数据

监控

部署

策略

配置

节点

资源

主机

节点列表

分机组列表

图表

分机组快照

节点快照

节点数据同步

启动节点

停止节点

同步服务

状态	节点名	主机名	分机组	主节点	角色	集合数	记录数	Lob数
全部				全部	全部	=	=	=
正常	sdbserver002:11820	sdbserver002	SYSCatalogGroup	true	catalog	-	-	-
正常	sdbserver003:11820	sdbserver003	SYSCatalogGroup	false	catalog	-	-	-
正常	sdbserver001:11820	sdbserver001	SYSCatalogGroup	false	catalog	-	-	-
正常	sdbserver001:11810	sdbserver001	SYSCoord	-	coord	-	-	-
正常	sdbserver002:11810	sdbserver002	SYSCoord	-	coord	-	-	-
正常	sdbserver003:11810	sdbserver003	SYSCoord	-	coord	-	-	-
正常	sdbserver001:11840	sdbserver001	group2	false	data	23	345080	0
正常	sdbserver002:11840	sdbserver002	group2	true	data	23	345080	0

1 / 1

显示 15 条记录，一共 15 条

SequoiaDB SAC

admin

数据

监控

部署

策略

配置

SequoiaDB1

节点列表

刷新

修改配置

重启

NodeName	catalogaddr
sdbserver001:11810	sdbserver001:11810,sdbserver003,sdbserver023
sdbserver001:11820	sdbserver001:11820,sdbserver003,sdbserver023
sdbserver001:11830	sdbserver001:11830,sdbserver003,sdbserver023
sdbserver001:11840	sdbserver001:11840,sdbserver003,sdbserver023
sdbserver001:11850	sdbserver001:11850,sdbserver003,sdbserver023
sdbserver001:11823	

修改配置项

普通

高级

自定义

numpreload

页面预加载代理数据

maxprefpool

数据预取代理池最大数量

maxreplsync

日志同步最大并发数量

logbuffsize

复制日志内存页面数

sortbuf

排序缓存大小

hjbuf

哈希连接缓存大小

syncstrategy

副本组之间数据同步控制

确定

取消

transisolation	weight
1	
1	
1	30
1	10
1	20

©2019 SequoiaDB. 版本: 系统时间: 17:45:35 系统状态: 良好

命令行可视化管理

```
sequoiadb — sdbadmin@c7401:~ — ssh 218.17.39.139 — 111x32
refresh= 3 secs          version 3.2.2          snapshotMode: GLOBAL
displayMode: ABSOLUTE    snapshotModeInput: NULL
hostname: localhost      filtering Number: 0
servicename: 11810       sortingWay: NULL sortingField: NULL
usrName: NULL
```

```
##### For help type sdbtop -h: u
#####

SOB Interactive Snapshot Monitor V2.0
Use these keys to ENTER:

window options(choose to enter window):
  m : return to main window      s : show
  c : show collection spaces     t : show
  d : show database state        h : help

options(use under window above):
  G : show options only          g : filter
  n : filter by specified node   r : sort
  A : sort column by ascending order D : skip
  C : specify filter condition   Q : move
  N : skip specified lines ahead W : switch
Tab : switch display Model      <- : move
-> : move right                 Enter : to
ESC : cancel current operation  F5 : refresh
q : quit
```

```
sequoiadb — sdbadmin@c7401:~ — ssh 218.17.39.139 — 111x32
refresh= 3 secs          version 3.2.2          snapshotMode: GLOBAL
displayMode: ABSOLUTE    snapshotModeInput: NULL
hostname: localhost      filtering Number: 0
servicename: 11810       sortingWay: NULL sortingField: NULL
usrName: NULL
Refresh: F5, Quit: q, Help: h
```

SessionID	TID	Type	Name	QueueSize
1	1	88922 Task	DATASYNC-JOB-D	0
2	1	21445 Task	DATASYNC-JOB-D	0
3	1	193724 Task	DATASYNC-JOB-D	0
4	1	89012 Task	DATASYNC-JOB-D	0
5	1	21501 Task	DATASYNC-JOB-D	0
6	1	193755 Task	DATASYNC-JOB-D	0
7	1	21500 Task	DATASYNC-JOB-D	0
8	1	193754 Task	DATASYNC-JOB-D	0
9	1	89013 Task	DATASYNC-JOB-D	0
10	1	193756 Task	DATASYNC-JOB-D	0
11	1	89014 Task	DATASYNC-JOB-D	0
12	1	21502 Task	DATASYNC-JOB-D	0
13	2	88958 LogWriter	**	0
14	2	21455 LogWriter	**	0
15	2	193735 LogWriter	**	0
16	2	89015 LogWriter	**	0
17	2	21510 LogWriter	**	0
18	2	193784 LogWriter	**	0
19	2	21503 LogWriter	**	0
20	2	193757 LogWriter	**	0
21	2	89045 LogWriter	**	0
22	2	193811 LogWriter	**	0
23	2	89088 LogWriter	**	0
24	2	21555 LogWriter	**	0

```
~ — ssh 218.17.39.139 — 111x32
3.2.2          snapshotMode: GLOBAL
snapshotModeInput: NULL
filtering Number: 0
sortingWay: NULL sortingField: NULL
Refresh: F5, Quit: q, Help: h

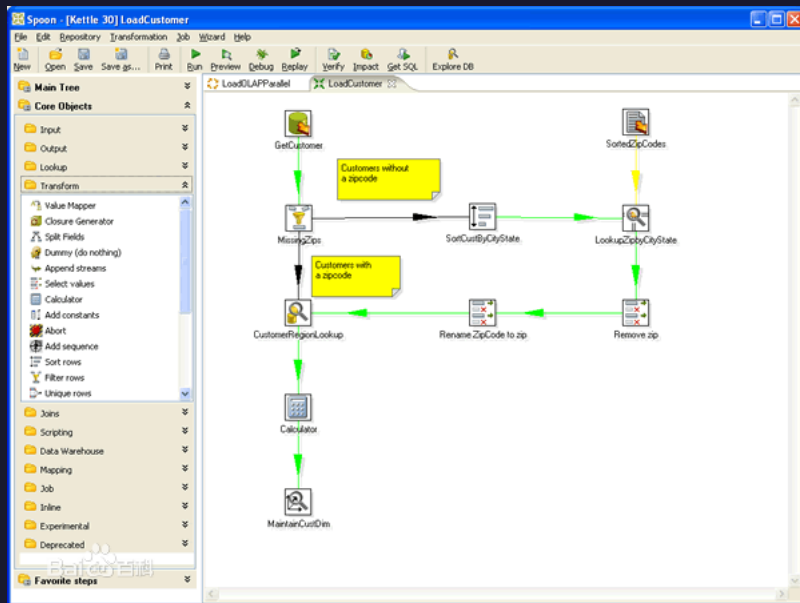
IndexRead      DataWrite      IndexWrite
0.667          98894.000      98893.667

Insert          ReplUpdate      ReplDelete
98894.000      0.000          0.000

Read            ReadTime        WriteTime
2.000          0.000          0.000
```

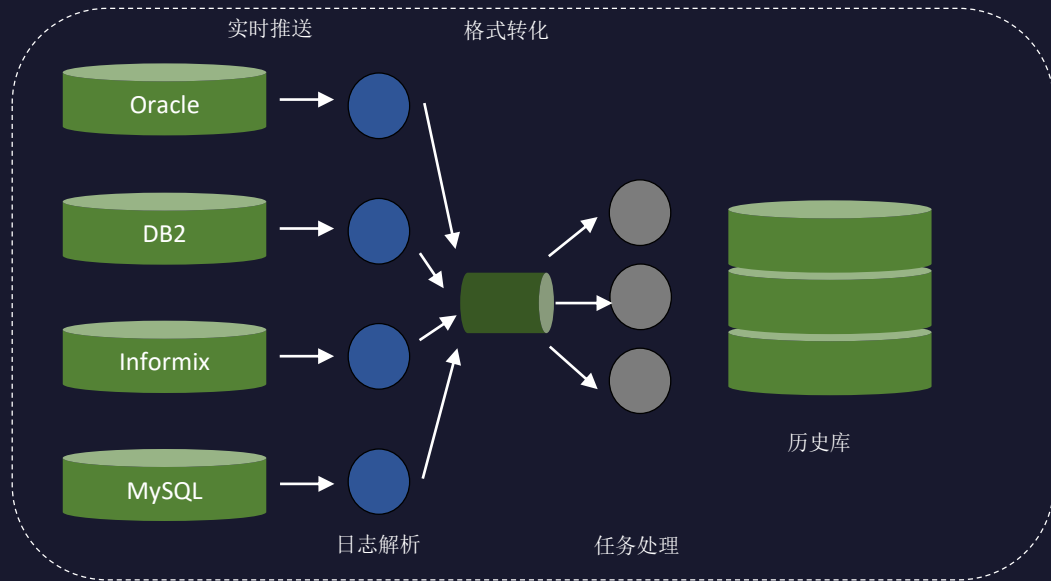
支持多种数据导入导出机制

- 标准开源工具mydumper/myloader
- 标准开源图形化工具Kettle
- 巨杉数据库自带工具sdbimpirt/sdbexpirt
- Oracle实时数据复制工具 Golden Gate
- IBM实时数据复制工具CDC
- MySQL实时数据复制Binlog Replication



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/DDI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

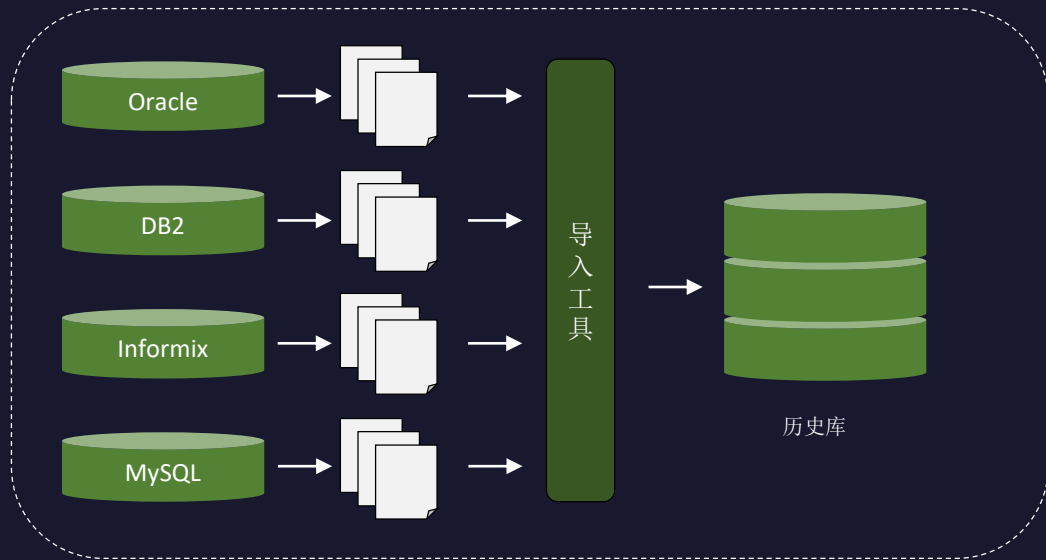


准实时数据复制策略

定期任务

异步数据复制策略

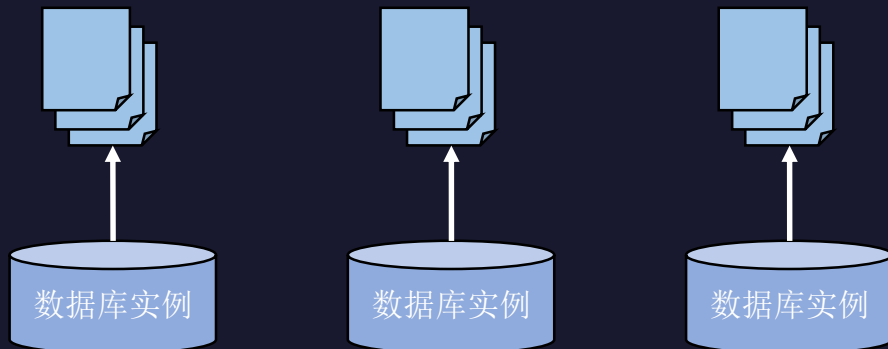
- 1、使用JSON或CSV格式定期将异构数据源的原始数据进行导出为文本文件
- 2、通过FTP等方式将文件传输至巨杉数据库的客户端
- 3、通过sdbimprt工具将文本文件导入巨杉数据库
- 4、满足异构数据源T+1的数据复制策略，简单可靠



异步
数据
复制
策略

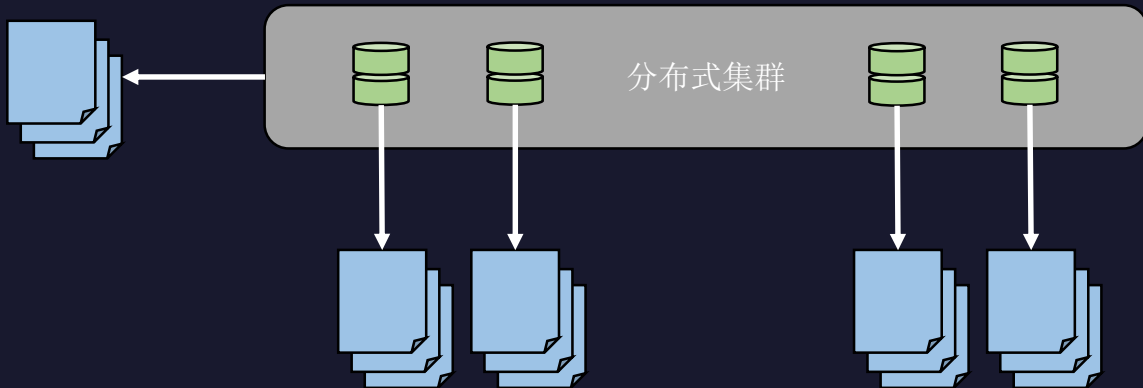
数据库实例级备份

- 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

MySQL 实例优势

- 1、通用型最强，使用范围广
- 2、与 MySQL 协议级兼容，100% 支持增删改查、存储过程、视图、触发器、自增字段、临时表、自定义函数等全部功能
- 3、针对 OLTP 场景优化，支持分布式事务
- 4、可插拔存储引擎，与 InnoDB 进行替换对应用无感知

```
sequoiadb — sdbadmin@c7401:~ — ssh 218.17.39.139 — 88x55
Copyright (c) 2000, 2019, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

[mysql> use hsd; ;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
[mysql> show tables;
+-----+
| Tables_in_hsd |
+-----+
| customer      |
| district      |
| history       |
| item          |
| new_order     |
| oorder        |
| order_line    |
| stock         |
| warehouse     |
+-----+
9 rows in set (0.00 sec)

[mysql> select count(*) from order_line;
+-----+
| count(*) |
+-----+
| 29719817 |
+-----+
1 row in set (1.13 sec)

[mysql> select * from order_line limit 3;
+-----+
| ol_w_id | ol_d_id | ol_o_id | ol_number | ol_i_id | ol_delivery_d | ol_amount |
| ol_supply_w_id | ol_quantity | ol_dist_info |
+-----+
| 16 | 39 | 1 | 1 | 1 | 63845 | 2019-06-12 18:03:47 | 0.00 |
| 16 | 39 | 5 | abcdefghigklmno71110337 | 5 | 25464 | 2019-06-12 18:03:47 | 0.00 |
| 16 | 34 | 1 | 1 | 2 | 25464 | 2019-06-12 18:03:47 | 0.00 |
| 16 | 34 | 5 | abcdefghigklmno71110559 | 5 | 43616 | 2019-06-12 18:03:47 | 0.00 |
| 16 | 10 | 1 | 1 | 3 | 43616 | 2019-06-12 18:03:47 | 0.00 |
| 16 | 10 | 5 | abcdefghigklmno71110579 | 5 |
+-----+
3 rows in set (0.00 sec)
```

PostgreSQL 实例优势

- 1、相比 MySQL, PGSQL 对复杂查询支持相对较好
- 2、PGSQL 支持 HSJN、MSJN 等多种关联机制, 对于统计类场景性能较高
- 3、支持分布式事务能力

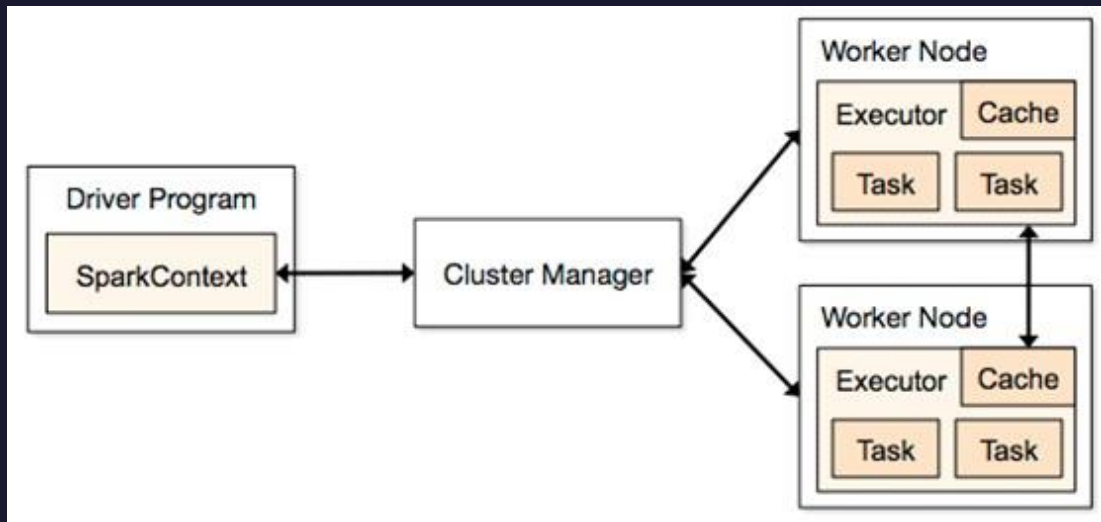
```
sequoiadb — sdbadmin@c7401:/opt/sequoiasql/postgresql — ssh 218.17.39.13...
[[sdbadmin@c7401 postgresql]$ bin/psql -p 5432 sample
psql (9.3.4)
Type "help" for help.

sample=# create foreign table test (name text, id numeric) server sdb_server opti
ions ( collectionspace 'sample', collection 'sample', decimal 'on' );
CREATE FOREIGN TABLE
sample=# analyze test;
ANALYZE
sample=# insert into test values ( 'tom', 1 );
INSERT 0 1
sample=# insert into test values ( 'bob', 2 );
INSERT 0 1
sample=# select * from test ;
 name | id 
-----+----
  tom  |  1 
  bob  |  2 
(2 rows)

sample=# update test set id=9 where name='tom' ;
UPDATE 1
sample=# select * from test ;
 name | id 
-----+----
  tom  |  9 
  bob  |  2 
(2 rows)
```

SparkSQL 实例优势

- 1、专门针对统计分析审计等场景使用
- 2、多个不同联机交易库中的表可以被直接映射到同一个 SparkSQL 实例中，避免ETL迁移流程
- 3、支持数据分区的读写分离，确保联机交易业务与统计分析任务在不同物理机中执行



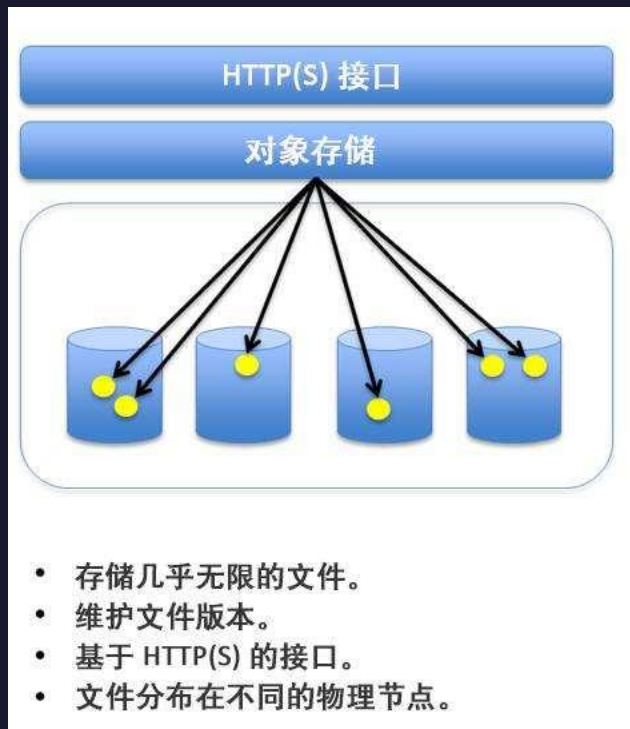
JSON (MongoDB) 实例优势

- 1、与 MongoDB 语法高度兼容
- 2、支持嵌套半结构化数据
- 3、支持数组索引与子对象索引
- 4、支持 JSON 事务能力

```
sequoiadb — sdbadmin@c7401:~ — ssh 218.17.39.139 — 106x32
[sdbadmin@c7401 ~]$ /opt/sequoiadb/bin/sdb
Welcome to SequoiaDB shell!
help() for help, Ctrl+c or quit to exit
> db=new Sdb()
localhost:11810
Takes 0.004531s.
> db.hsd.order_line.count();
29719817
Takes 0.003193s.
> db.hsd.order_line.find().limit(1);
{
  "_id": {
    "$oid": "5d00ce03ad84487c1a8f2fda"
  },
  "ol_w_id": 16,
  "ol_d_id": 1,
  "ol_o_id": 1,
  "ol_number": 1,
  "ol_i_id": 63845,
  "ol_delivery_d": "2019-06-12 18:03:47",
  "ol_amount": {
    "$decimal": "0.00"
  },
  "ol_supply_w_id": 39,
  "ol_quantity": {
    "$decimal": "5"
  },
  "ol_dist_info": "abcdefghijklmno71110337"
}
Return 1 row(s).
Takes 0.001216s.
```

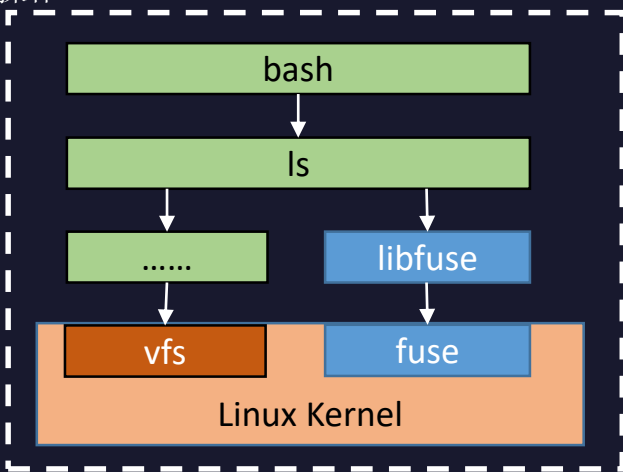
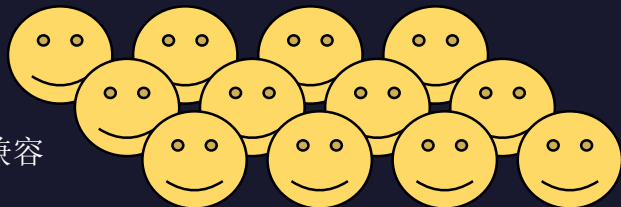
S3 对象存储实例优势

- 1、与 AWS S3 协议级兼容
- 2、支持多版本批次上传
- 3、支持大文件分段上传（断点续传）
- 4、支持元数据标签化管理
- 5、支持元数据标签模糊检索



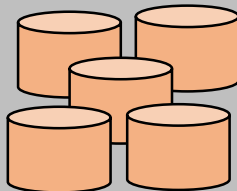
文件系统实例优势

- 1、与 Posix 文件系统完整兼容
- 2、支持全部 Linux 文件系统操作
- 3、应用程序透明无感知
- 4、弹性水平扩张

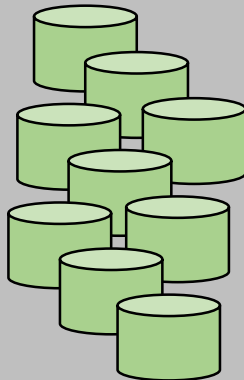


巨杉数据库集群

元数据区



对象存储区



HTAP读写分离能力

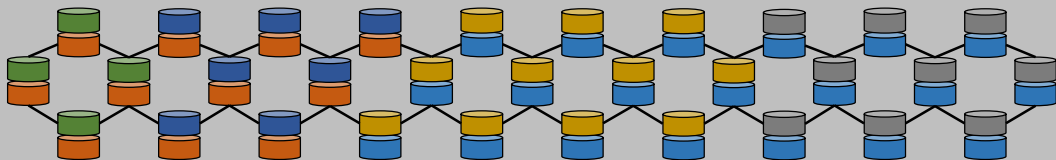
微服务框架下，对成千上万个MySQL数据库实例做到统一化管理，防止数据碎片化，并对来自不同实例和服务的数据统一实时分析，避免联机交易与分析业务相互干扰

MySQL实例1
(高可用)

MySQL实例2
(高可用)

MySQL实例3
(高可用)

MySQL实例4
(高可用)



SparkSQL实例1

SparkSQL实例2

多租户物理隔离能力

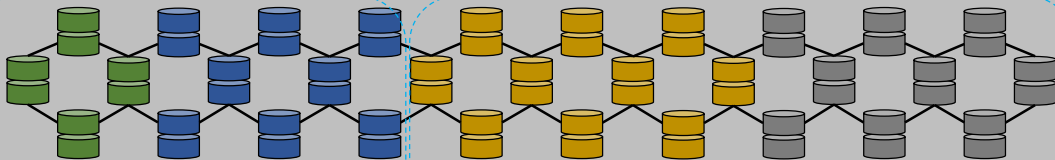
在一个集群内同时提供关系型数据库以及对象存储实例
尽可能减少用户对于异构产品的学习与运维成本

MySQL实例1

MySQL实例2

S3对象存储

Posix文件系统



结构化存储格式

非结构化存储格式

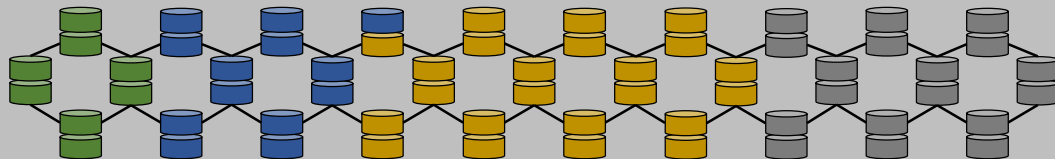
每个实例提供完全隔离的权限控制与数据可视范围
确保不会管理员不会有意无意使实例访问被隔离的其他信息

核心账务实例

信贷实例

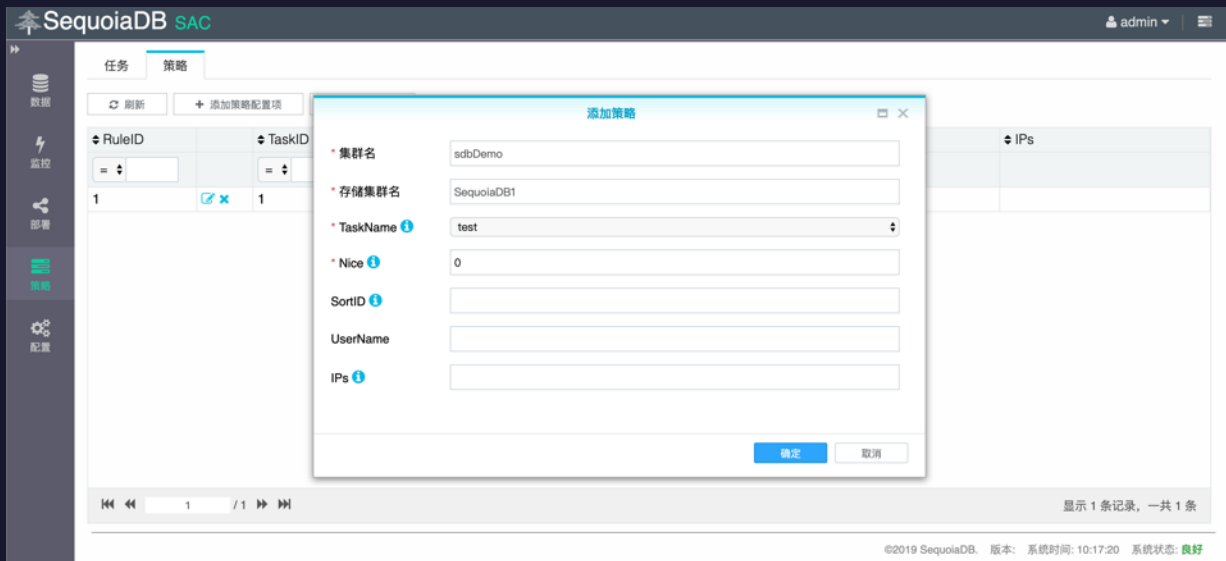
信用卡实例

渠道业务实例



优先级策略管理

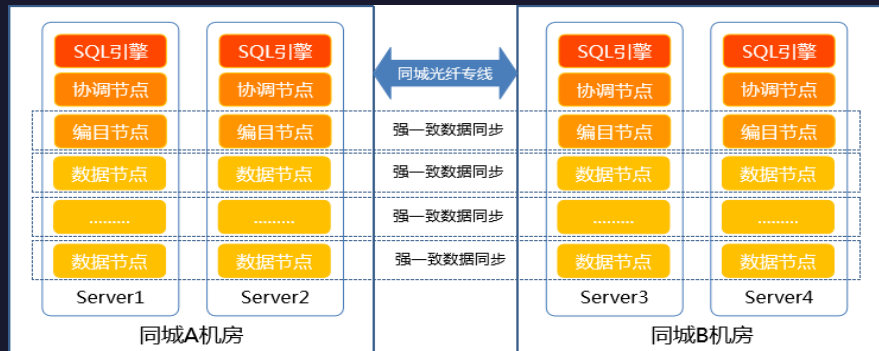
- 1、针对用户名与 IP 地址（数据库实例）定义业务优先级
- 2、硬件资源发生竞争时基于优先级进行调度
- 3、低负载情况下不造成任何额外负担与开销
- 4、高负载情况下优先调度重要应用



多中心容灾能力

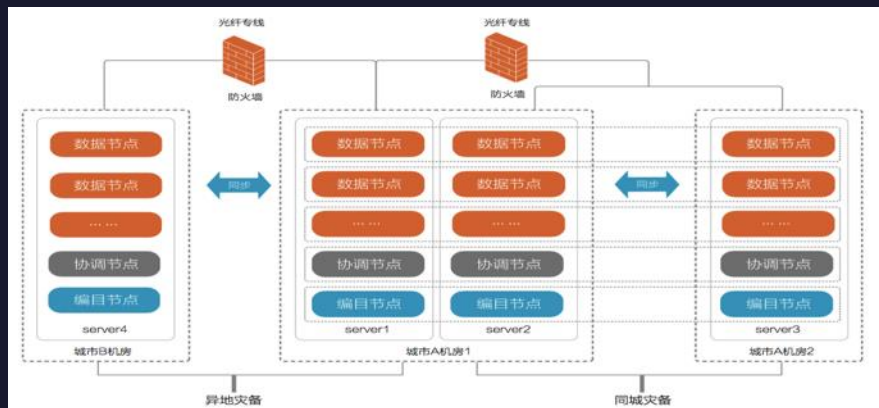
同城方案

- 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



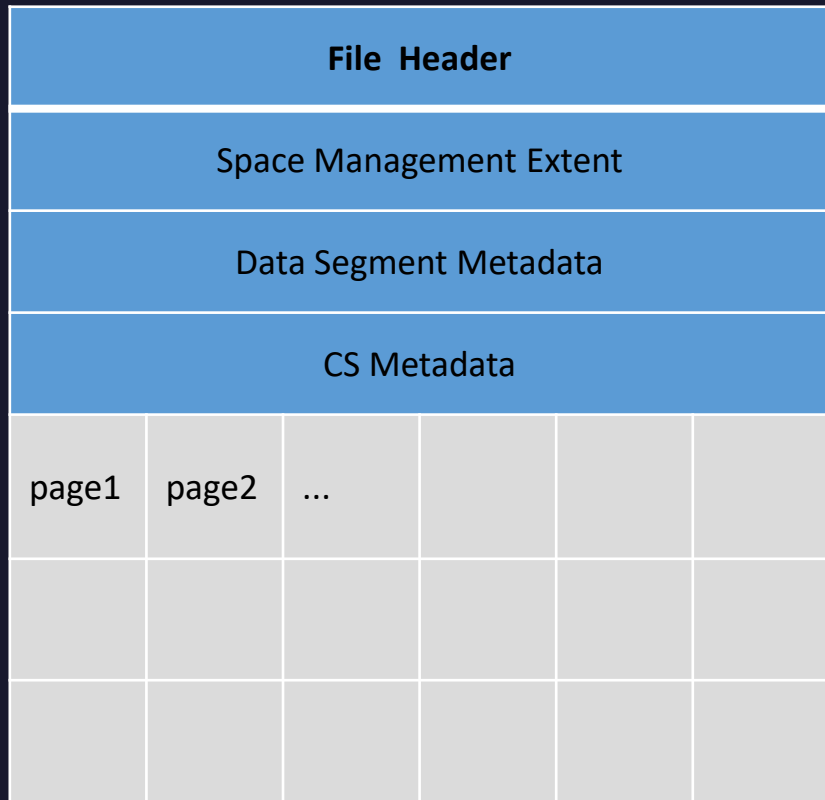
两地三中心

- 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

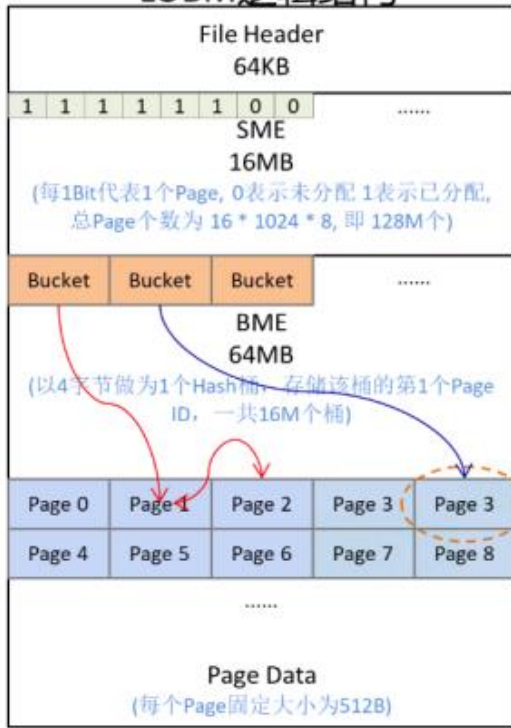
Deep Dive



记录存储格式



LOBM逻辑结构



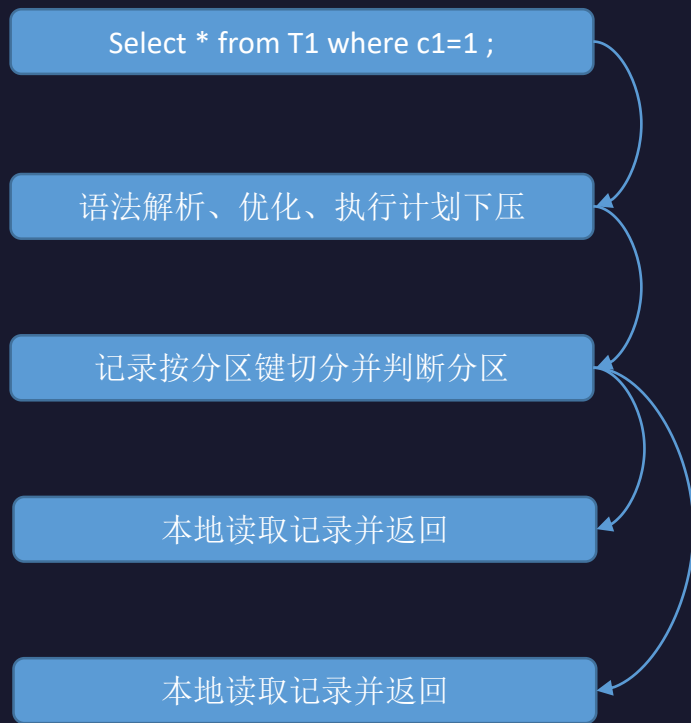
PAD 4B	OID 12B	Sequence 4B	Data Len 4B
Pre-Page 4B	Next-Page 4B	CLLID 4B	MBID 2B
PAD 212B			

LOBD逻辑结构





写入流程



读取流程

应用程序 S3 SDK

Bucket.put (objectID, fileName) ;

协调节点

对文件切分，按照objectID与数据块偏移进行散列，并下发至对应分区

数据主节点

接收数据块写入日志与文件

数据从节点

写入日志与数据，并返回主节点

写入流程

File = Bucket.get (objectID) ;

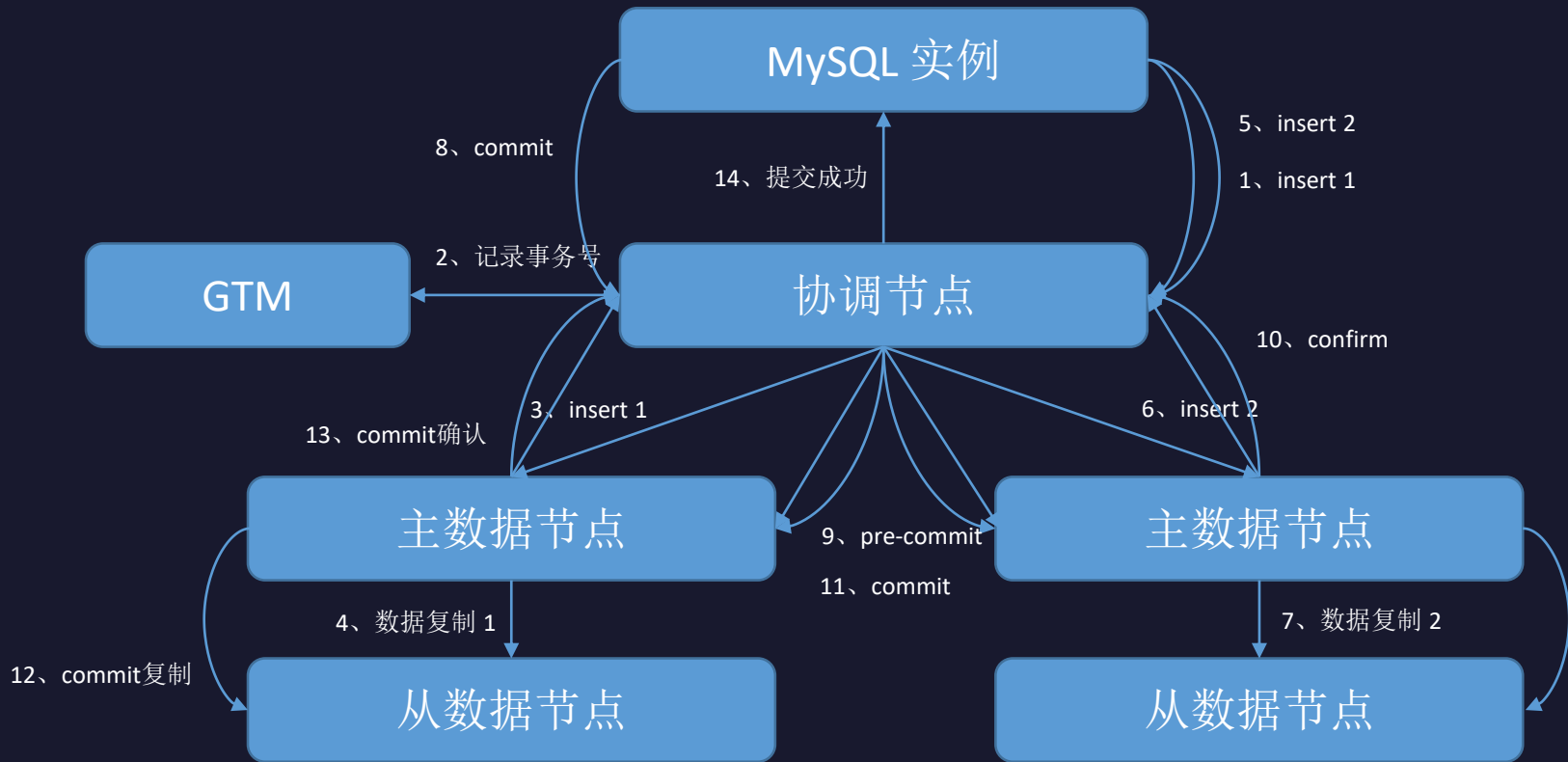
按照objectID与偏移量判断分区并读取

本地读取数据片段并返回

本地读取数据片段并返回

读取流程

两阶段提交过程



CPU、网络与磁盘资源压不满

- 1、top 看到有大量空闲资源
- 2、应用并发能力不足 - 提升应用并发量
- 3、应用存在大量锁等待 - 检查应用逻辑，调整隔离级别
- 4、应用程序为主要瓶颈 - 检查应用程序日志，判断主要耗时模块

网络占用高

- 1、网络使用率超过 90%
- 2、考虑使用万兆网或增加服务器数量，提升整体集群的网络吞吐能力
- 3、应用程序是否存在大量无效的数据请求

CPU占用高

- 1、top 发现CPU 90% 以上占用
- 2、usr CPU 过高 - 检查是否存在大量表扫描，增加集群容量增加计算节点
- 3、sys CPU 过高 - 考虑使用 tcmalloc 等内存分配机制

I/O 占用高

- 1、iowait% 超过 50%
- 2、热数据数据量是否过大，是否存在数据倾斜导致某节点数据堆积，考虑集群扩容
- 3、考虑使用 SSD 磁盘替换 SAS 或 SATA 盘
- 4、检查是否存在大量表扫描，计算数据读与索引读比例

MySQL实例配置	默认值	描述
sequoiadb_use_partition	on	是否默认创建分区表
sequoiadb_use_bulk_insert	on	批量插入时是否开启批插功能
sequoiadb_bulk_insert_size	100	开启批插功能时默认批次大小
sequoiadb_use_autocommit	on	是否开启自动提交功能
sequoiadb_replica_size	-1	表同步复制份数（-1为强一致，三副本全写入）

配置文件位置：实例目录下auto.cnf

分布式集群性能相关配置	默认值	描述
maxpool	50	最大线程池数量
numpreload	0	是否开启数据预读功能
sortbuf	256MB	排序缓存大小
preferedinstance	M	会话优先访问的副本
plancachelevel	0	访问计划缓存级别

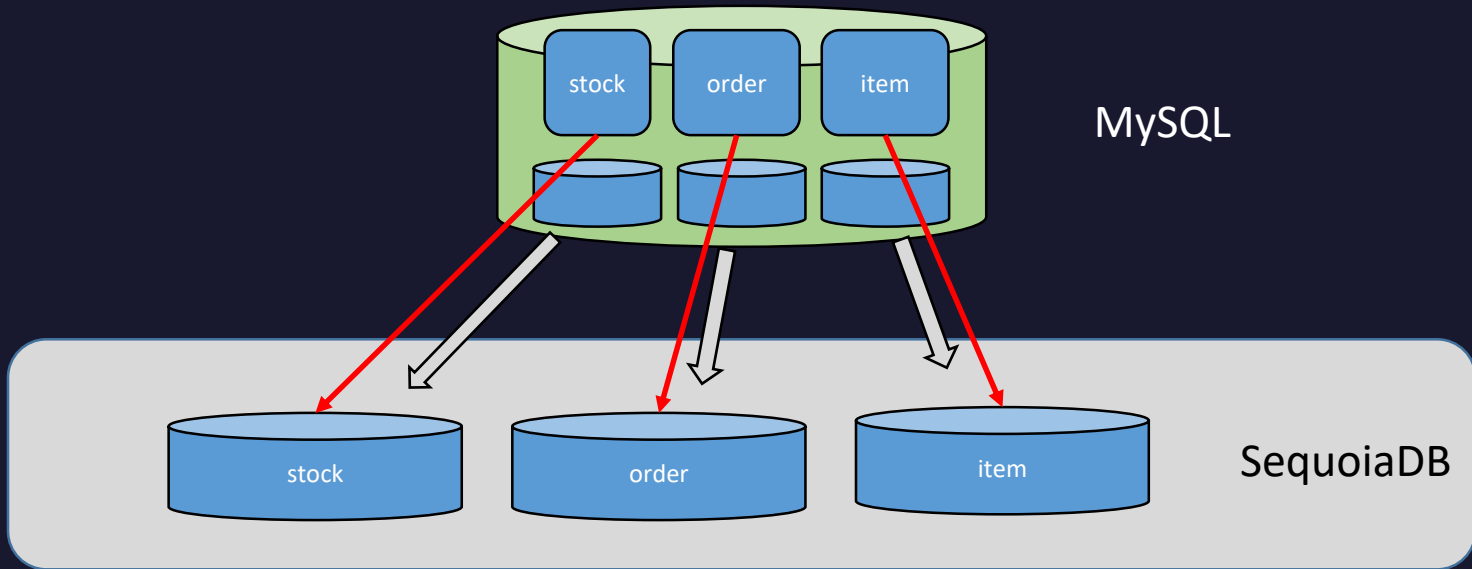
配置文件位置：安装目录/conf/local/端口号/sdb.conf

分布式集群事务相关配置	默认值	描述
transactionon	true	是否开启事务
transactiontimeout	60s	事务超时时间
transisolation	0	隔离级别，默认为UR
translockwait	false	读记录发现锁时是否等待，还是直接读取之前的已提交版本
transautocommit	false	事务是否自动提交
transautorollback	true	操作失败后该事务是否自动回滚
transuserbs	true	是否使用回滚段记录已提交版本记录

配置文件位置：安装目录/conf/local/端口号/sdb.conf

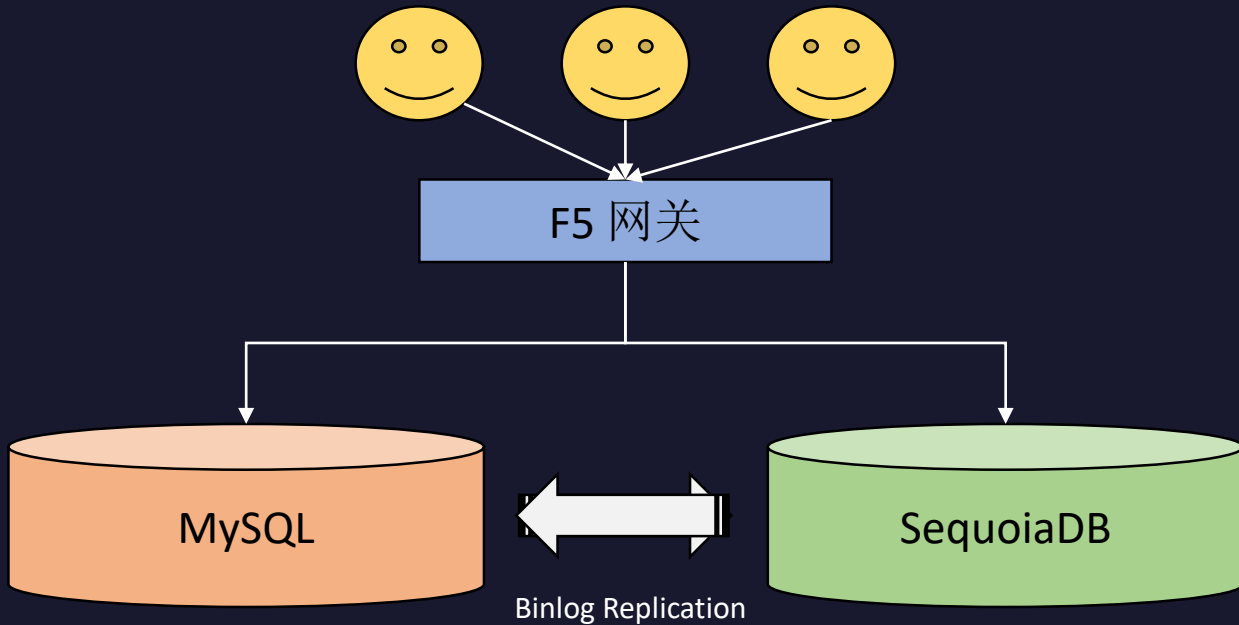
InnoDB/SequoiaDB混布模式

已有 MySQL 应用在进行数据库迁移时，可以分阶段针对不同表进行分批迁移。
SequoiaDB 支持 MySQL 实例中对 InnoDB 引擎与 SequoiaDB 引擎混合部署，减少用户迁移风险与时间



MySQL / SDB binlog Replication双向复制

支持与标准 MySQL 的binlog replication实时双向复制
结合 100% MySQL 兼容特性，关键时刻应用可在 SequoiaDB 与 MySQL 之间无缝自由切换



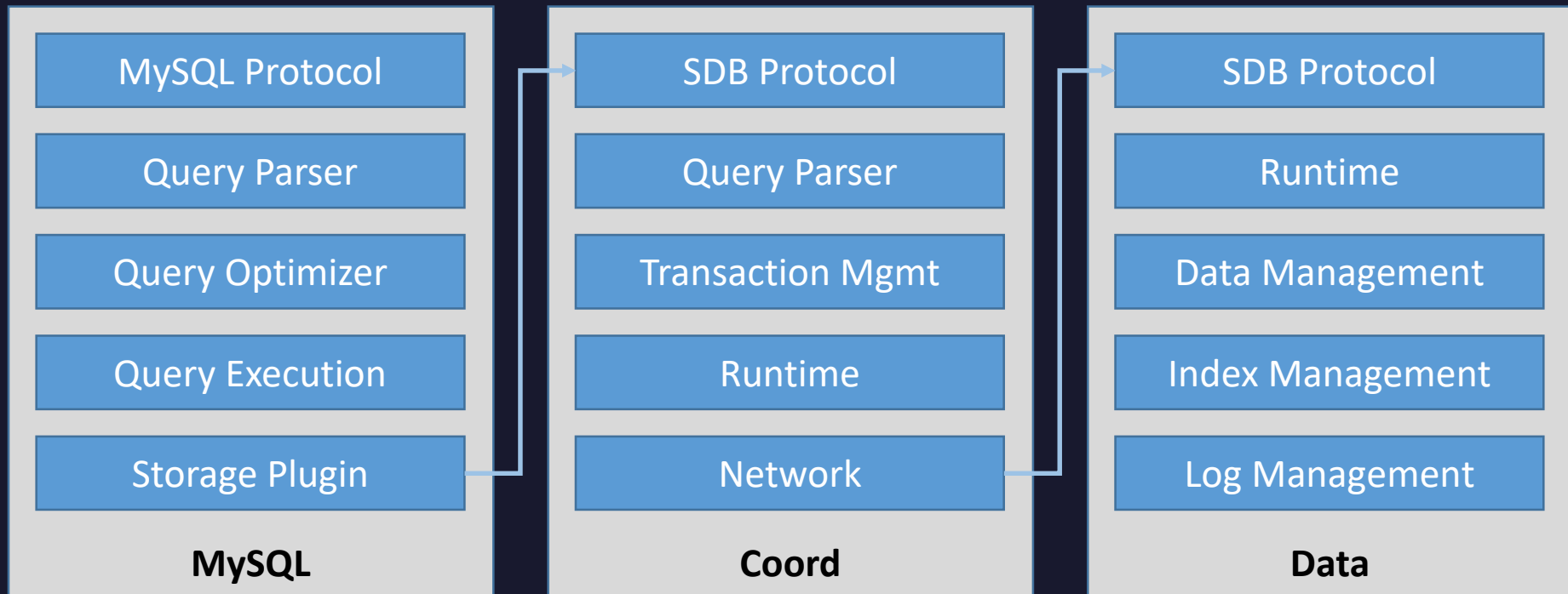
标准 S3 协议	自定义 API
PUT Bucket	Create User
DELETE Bucket	Create AccessKey
GET Service	DELETE User
GET Bucket location	GET AccessKey
HEAD Bucket	LIST Regions
PUT Object	PUT Region
GET Bucket (List Objects) Version 2	DELETE Region
GET Bucket Object versions	GET Region
GET Object	HEAD Region
HEAD Object	
DELETE Object	
PUT Bucket versioning	
GET Bucket versioning	

统计信息（表）

字段名	数据类型	默认值	说明
CollectionSpace	String	--	统计的collection所在collection space的名称
Collection	String	--	统计的collection的名称（不带collection space名字）
CreateTime	NumberLong	0	统计收集的时间戳
SampleRecords	NumberLong	0	统计收集时抽样的文档个数
TotalRecords	NumberLong	10	统计收集时的文档个数 对应dmsMBStatInfo的_totalRecords，用于对比统计信息是否过期
TotalDataPages	NumberInt	1	统计收集时的数据页个数
TotalDataSize	NumberLong		对应dmsMBStatInfo的_totalDataLen
AvgNumFields	NumberInt	10	每个文档中平均字段数

统计信息（索引）

字段名	数据类型	默认值	说明
CollectionSpace	String	--	统计的collection所在collection space的名称
Collection	String	--	统计的collection的名称（不带collection space名字）
CreateTime	NumberLong	0	统计收集的时间戳
Index	String	--	统计Index的名称
KeyPattern	BSONObj	--	统计索引的字段定义，例如：{a:1, b:-1}
SampleRecords	NumberLong	0	统计时抽样的索引项个数
TotalRecords	NumberLong	10	统计收集时的文档个数（收集时间不同，因此可能与SYSSTAT.SYSCOLLECTIONSTAT的TotalRecords不相等） 对应dmsMBStatInfo的_totalRecords，用于对比统计信息是否过期
IndexPages	NumberInt	1	统计收集时索引的页个数
IndexLevels	NumberInt	1	统计收集时索引的层数
IsUnique	BOOL	FALSE	Index是否唯一索引
DistinctValues	NumberLong	0	统计字段的唯一值个数，如果没有统计，取对应Collection的TotalRecords
NullFrac	NumberInt	0	null值在字段中的比例，最终比例 NullFrac / 10000
UndefinedFrac	NumberInt	0	\$undefined在字段中比例，最终比例 UndefinedFrac / 10000
MCV	Object	undefined	Most Common Values MCV: { Values: [{a:1,b:1}, {a:2, b:2}, ...], Frac: [1000, 1000, ...] }
MCV.Values	Array		MCV的值
MCV.Frac	Array		MCV的比例，每个值的取值 0 ~ 10000，最终比例 Frac / 10000
Histogram	Object	undefined	直方图
Histogram.Frac	NumberDouble	0	直方图的比例，每个值的取值 0 ~ 10000，最终比例 Frac / 10000
Histogram.Bounds	Array		直方图的边界值
TypeSet	Object	undefined	类型比例
TypeSet.Types	Array		字段的类型
TypeSet.Frac	Array		字段的各个类型占比例，每个值的取值 0 ~ 10000，最终比例 Frac / 10000



数据均衡分布

- 1、在给定Domain中确保数据分布均匀
- 2、表分区键选择合理

数据域规划合理

- 1、针对不同业务类型（如交易型与对象存储）建议使用不同的数据域进行硬件隔离
- 2、针对业务压力可能同时暴增的业务尽可能拆分到不同的数据域与硬件，合理分配资源
- 3、多维分区中可以根据数据归档周期（例如以年为单位）划分数据域，每次清空维度可以快速释放硬件资源
- 4、异构硬件应尽可能使用不同数据域（例如不同批次采购的配置不同的设备，或SSD、SAS盘混布的集群）

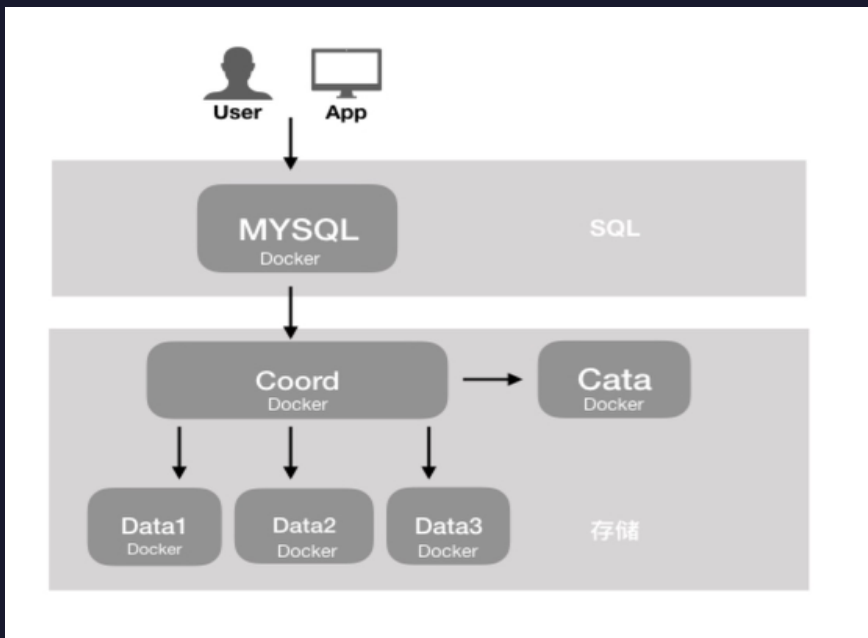
合理使用多维分区

- 1、针对历史流水记录，或非交易型递增为主的表考虑使用多维分区表
- 2、针对银行业务流水记录，最佳实践平均一月一个分区表
- 3、访问每个分区会有额外overhead，避免主要查询每次访问超过10个分区表

合理进行读写分离规划

- 1、非交易类只读业务，在不需要强一致保障的情况下可以优先读取备节点
- 2、后督审计类业务可以优先访问备节点
- 3、所有强一致交易类业务必须从主节点进行读写

支持容器化部署方式，支持通过k8s进行容器编排
通过 sequoiadb 镜像生成 Coord、Data 和 Catalog 节点，通过 sequoysql-mysql 镜像生成 MySQL 实例





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/>