Oracle Sharding 可扩展的多模型分布式数据库

公益讲座11：00准时开始，请大家先浏览云技术微信公众号技术文章。资料会在各群同步发布，已入群客户请勿重复入群！

20-21

数据库和云讲座群

甲骨文云技术公众号

B站专家系列课程

基于 Oracle 数据库
免费企业数据健康检查

- 及时了解数据库健康状况，发现并解决潜在问题
- 维护数据库系统良好状态，保护数据资产的安全
- 提升数据库性能、稳定性和安全性，降低业务风险

免费咨询热线：

400-699-8888

* 活动最终解释权归甲骨文公司所有

Oracle Sharding
可扩展的多模型分布式数据库

甲骨文技术公益课 - 数据库专场

2023 年 7 月 21 日 11:00

线上直播

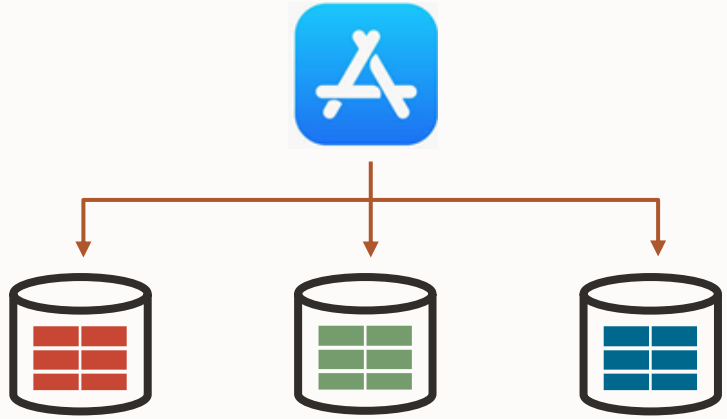
范宏伟

议程

1. Oracle Sharding概览
2. Sharding 生命周期管理
3. Oracel Sharding的演进及新特性

Oracle Database Sharding-数据库的分布式部署

- 跨独立数据库的数据水平分区(shards)
  - 每个shard持有数据的一个子集
  - 可以是单节点/RAC/PDB
  - 高可用复制
- Shared-nothing 架构:
  - Shard间不共享任何硬件(CPU、内存、磁盘)
  - 或软件(集群软件)
- 海量并发和并行度
  - 百万并发交易
  - 并行的多分片查询



表Table	1	1	1
Shard	1	2	3
服务器 Server	1	2	3

2017年3月，发布了GA 12.2版本，推出了Oracle Sharding功能，适用于OLTP应用

←- - - - Single Logical Database - - - - →

将单个逻辑数据库分片到N个物理数据库中

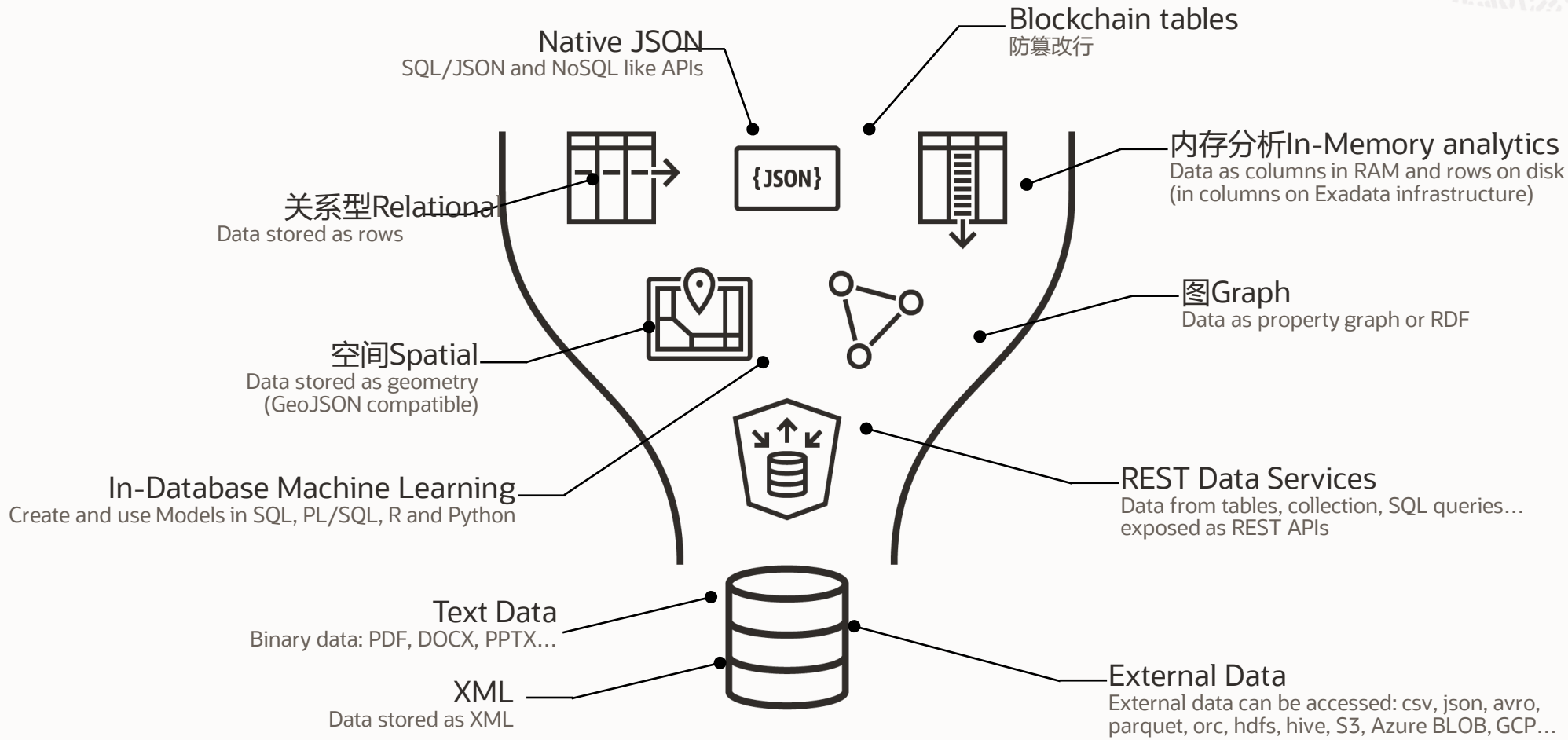
关键客户使用案例

客户使用案例	选择Oracle Sharding前评估的产品
互联网规模OLTP	Cassandra, MongoDB, MariaDB, Couchbase, Aerospike, ScyllaDB
全球数据库/数据自治	Google Spanner, Azure Cosmos DB, AWS Aurora, CockroachDB
日志 / 文本存储	Apache Lucene, Elastic Search, Solr
Metric/Time Series 存储, IoT, 基础设施健康	AWS Redshift/EMR, Druid, Cassandra, Graphite, InfluxDB
机器学习	Apache Spark, HDFS, NoSQL and SQL Sharded DBs
大数据分析	Apache Spark, SingleStore

客户之所以选择Oracle sharding，是因为客户评估的许多产品都是分片系统，缺乏企业级功能，如支持严格的数据一致性、事务、ACID属性、复杂join联接、完全SQL支持、高级安全性、跨区域复制、性能优化器、备份和恢复、触发器、存储过程、常规安全补丁，可管理性等

Oracle 融合数据库是多模数据库

支持任何工作负载，并且可以针对超大规模和地理分布进行分片



Oracle Sharding – 架构及组件

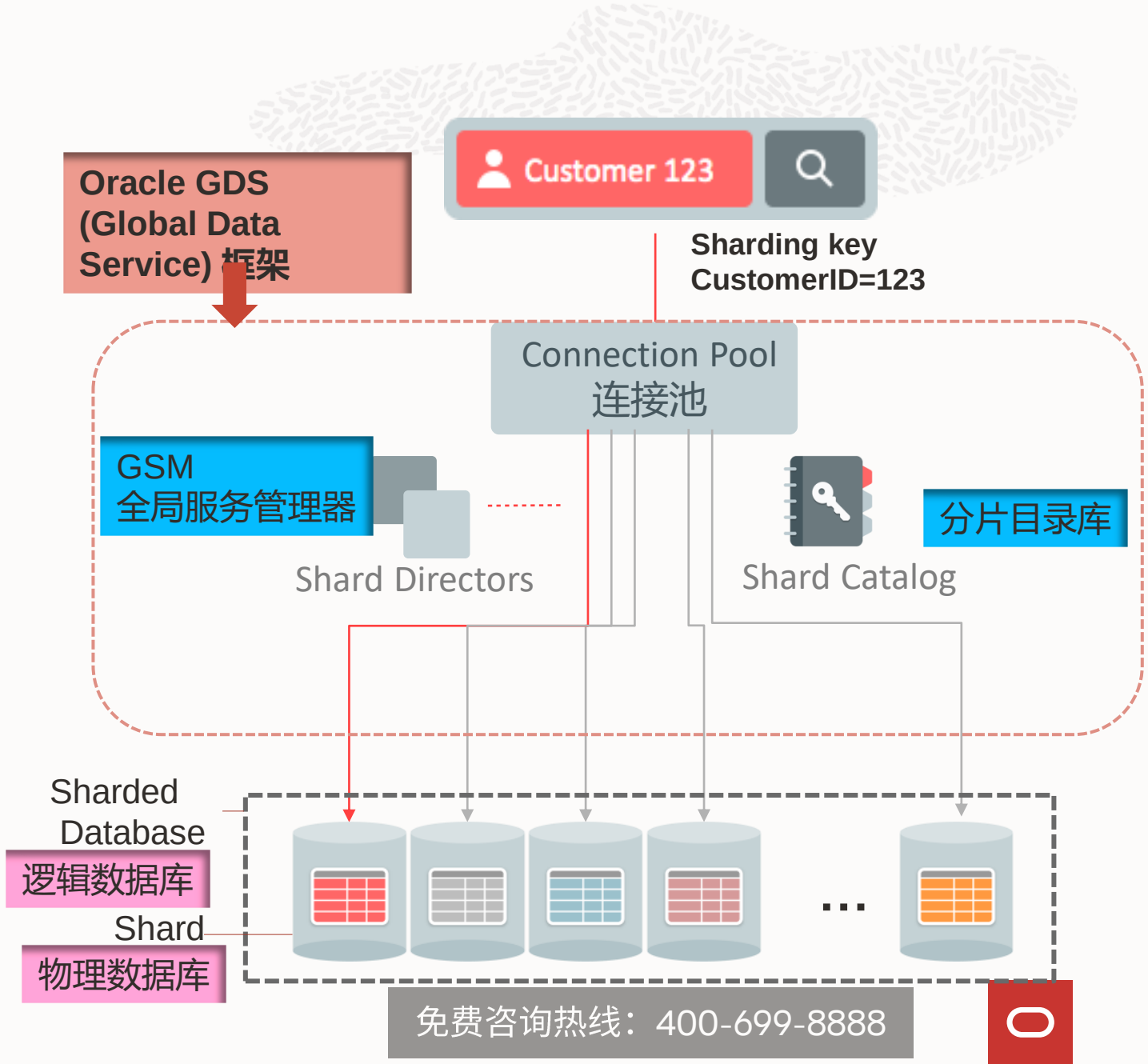
Oracle Sharding架构主要由两部分组成：

Oracle GDS（Global Data Services）框架来实现自动部署和sharding的管理以及拓扑复制

- ✓ GDS提供了对整个sharded database访问的负载均衡和基于位置的路由功能
- ✓ 在GDS框架中，Global Service Manager(GSM)负责将应用请求转发到合适的shard上
- ✓ shard catalog分片目录库，支持跨shard的查询功能，同时存储了sharded database的配置数据

底层使用Oracle表分区技术，将数据水平分片、存储到不同的物理数据库

- ✓ 每个物理数据库称为shard，位于不同的服务器，这些shard组成一个逻辑数据库，称为sharded database (SDB)



Sharding组件

Shard Director

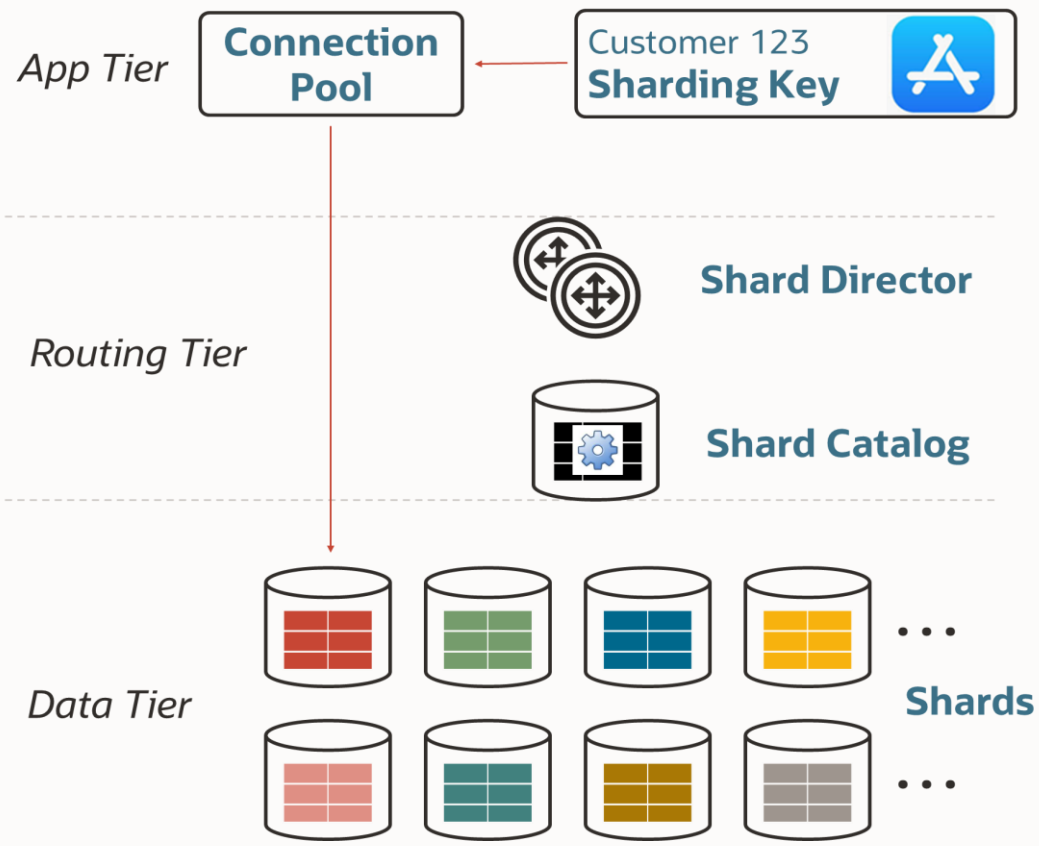
- 用于将连接请求路由到分片的全局侦听器
- 向客户端发布运行时 SDB 拓扑更新、负载平衡建议和 FAN 事件

Client-side Connection Pool

- 缓存SDB拓扑图
- 将请求直接路由到分片

Shard Catalog

- 集中存储和管理SDB配置信息
- 应用跨多个shard查询，由Shard catalog统一协调
- 添加/删除shard等配置变化都记录在Shard catalog



Schema创建 - 分片表和复制表

表家族Table Family

Customers

Customer	Name
123	Mary
456	John
999	Peter

Orders

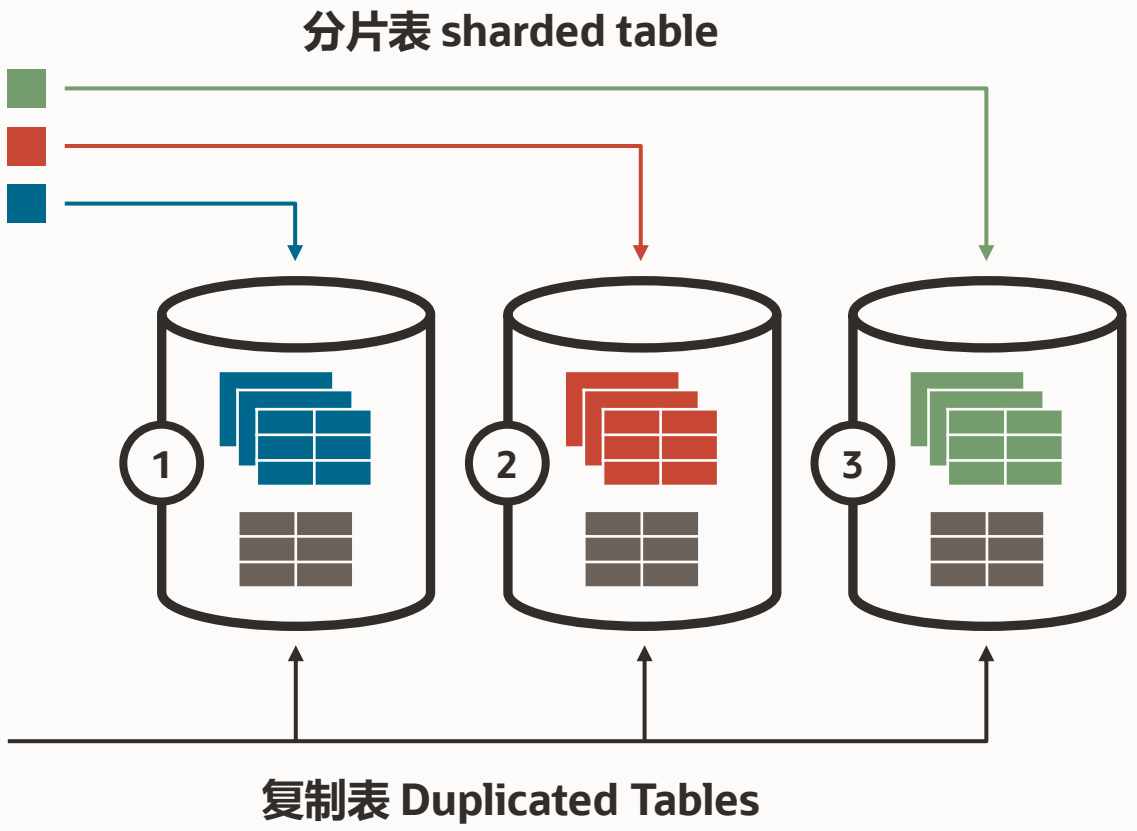
Order	Customer
4001	123
4002	456
4003	999
4004	456
4005	456

Line Items

customer	order	Line
123	4001	40011
999	4003	40012
123	4001	40013
456	4004	40014
999	4003	40015
999	4003	40016

Products

SKU	Product
100	Coil
101	Piston
102	Belt



分片中的组成

Chunk 集， 数据来自分片表 + 复制表

复制表的功能改进

Duplicated tables是物化视图

- master table 存在shard catalog中
- duplicated table 也可以在各分片更新（18c）
- 在23c之前, 所有duplicated tables以相同的频率刷新

每个表的刷新率

- 可以在表创建时指定或稍后更改
- 覆盖全局参数：shrd_dupletable_refresh_rate

On-demand 刷新

Duplicated Table同步

- catalog上的DML commit时被刷新
- 分片和catalog的内容保持一致

Oracle Sharding 分片方法

系统管理分片（System Managed Sharding）
- 使用一致性哈稀 （Consistent Hash）
  - ✓ 分配给每个块的哈希值范围

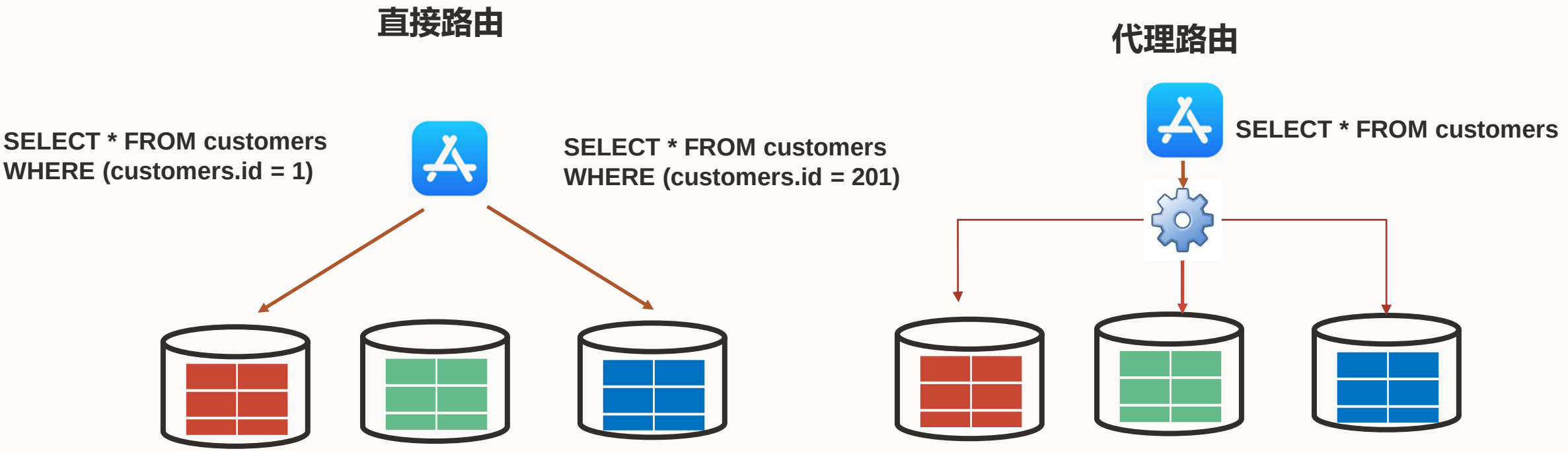
用户定义的分片 （User-defined Sharding）
- 按范围（ Range）
  - ✓ 分配给每个Chunk的分片键值的范围
- 按列表（ List）
  - ✓ 每个chunk与分片键值列表相关联

复合分片 (Composite Sharding)
- 按照 Range - Consistent Hash 或者按照 List - Consistent Hash
  - ✓ 两级分片，使用两个键

New In 23c

细粒度的自定义分片（Fine grained custom sharding）
- 各个键值到分区（分片）的自定义映射
  - ✓ 映射在运行时指定：在插入期间或之前

客户端请求流程

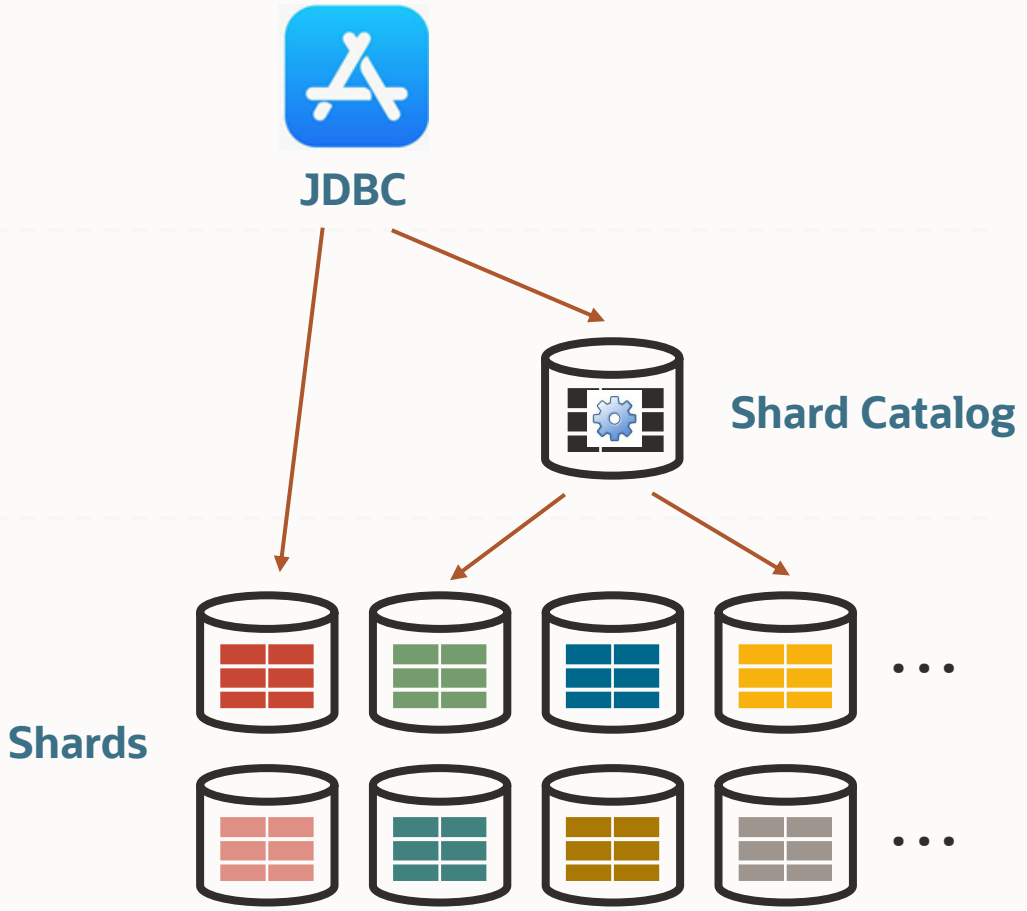


Java应用程序的自动路由

Java应用程序不需要在连接时传递分片的sharding key

JDBC Driver:

- 标识给定SQL语句的哪个绑定变量是分片键
- 提取 sharding key的值
- 将单分片查询或DML路由到适当的分片
- 跨分片查询和DML路由到 shard catalog (coordinator)



Oracle Database 19c (19.3) for Linux x86-64

Download	Description
LINUX.X64_193000_db_home.zip	(3,059,705,302 bytes) (sha256sum - ba8329c757133da313ed3b6d7f86c5ac42cd9970a28bf2e6233f3

Directions
Installation guides and general Oracle Database 19c documentation are here.

Oracle Database 19c Grid Infrastructure (19.3) for Linux x86-64

Download	Description
LINUX.X64_193000_grid_home.zip	(2,889,184,573 bytes) (sha256sum - d668002664d9399cf61eb03c0d1e3687121fc890b1ddd50b35dc

Contains the Grid Infrastructure Software including Oracle Clusterware, Automated Storage Management (ASM), and ASM Cluster File System. Download and install prior to installing Oracle Real Application Clusters, Oracle Real Application Clusters One Node, or other application software in a Grid Environment

Oracle Database 19c Global Service Manager (GSM/GDS) (19.3) for Linux x86-64

Download	Description
LINUX.X64_193000_gsm.zip	(959,891,519 bytes) (sha256sum - 9e2ebf7bdc10ad91c9e400dd721ed67707eb93800085267b681

Contains the Global Service Manager Software. Download and install as part of Global Data Services (GDS) deployment.

部署Oracle Sharding所需软件

Oracle Database +

Oracle Global Service Manager

议程

1. Oracle Sharding概览
2. Sharding 生命周期管理
3. Oracel Sharding的演进以及新特性

自动化部署

- **分片顾问（Shard Advisor）**
- ✓ 用于建议架构从非分片数据库迁移到分片的工具
- ✓ 主要目标是最大程度地提高并行度（在所有分片上扩展查询执行），最小化跨分片操作并最小化重复数据
- ✓ 分析现有的数据库架构，用户工作量并提出建议，例如要对哪些表进行分片，将哪个列用作分片键，要使用的分片方法，要复制的表
- **使用Terraform，Kubernetes和Ansible进行部署自动化**
- ✓ 简单的输入文件，描述了部署拓扑
- ✓ 从其中的一台主机运行以进行分布式设置
- ✓ 出现错误时可重启/恢复/清理
- ✓ 独立扩展分片组件

Shard Advisor Sample Output

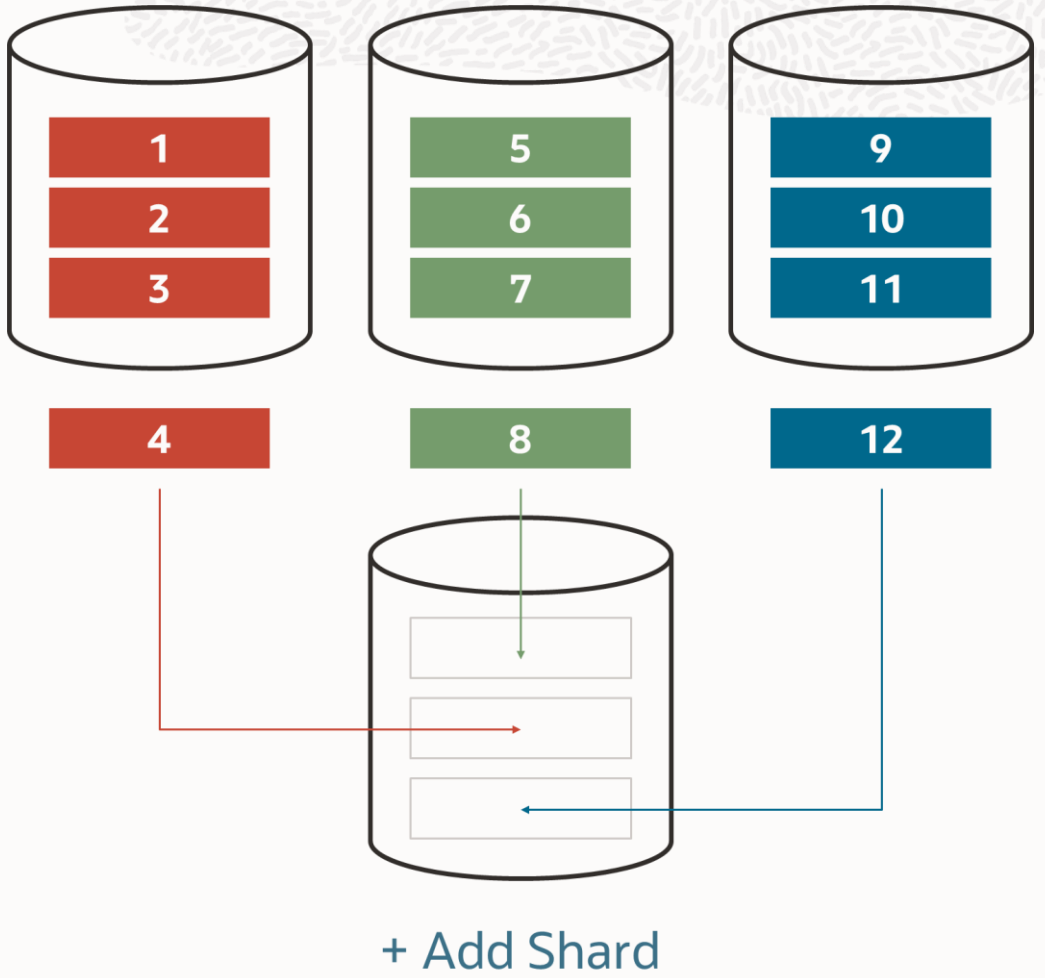
rank	tname	type	tlevel	parent	shardBy	cols	size	unenforced
1	CUSTOMER	S	1		HASH	C_CUSTKEY	44	CUSTOMERFK
1	ORDERS	S	2	CUSTOMER	REFERENCE	ORDERSFK	289	
1	LINEITEM	S	3	ORDERS	REFERENCE	LINEITEMFK1	1472	LINEITEMFK2
1	NATION	D			NONE		1	
1	PART	D			NONE		43945	
1	PARTSUPP	D			NONE		23340	
1	REGION	D			NONE		1	
1	SUPPLIER	D			NONE		260	

Terraform Script Input File

```
shards = {
  "shard-1" = {
       host = "den02ffv"
       port = "1521"
       sid = "sh1"
       globalDBName = "sh1"
       shard_group = "primary_shardgroup"
  },
  "shard-2" = {
       host = "den02ffw"
       port = "1521"
       sid = "sh2"
       globalDBName = "sh2"
       shard_group = "primary_shardgroup"
  }
```

在线添加和重新平衡分片

- chunk是Re-sharding的单位
- 移动chunk可以自动或DBA手动启动
- 使用RMAN增量备份和传输表空间

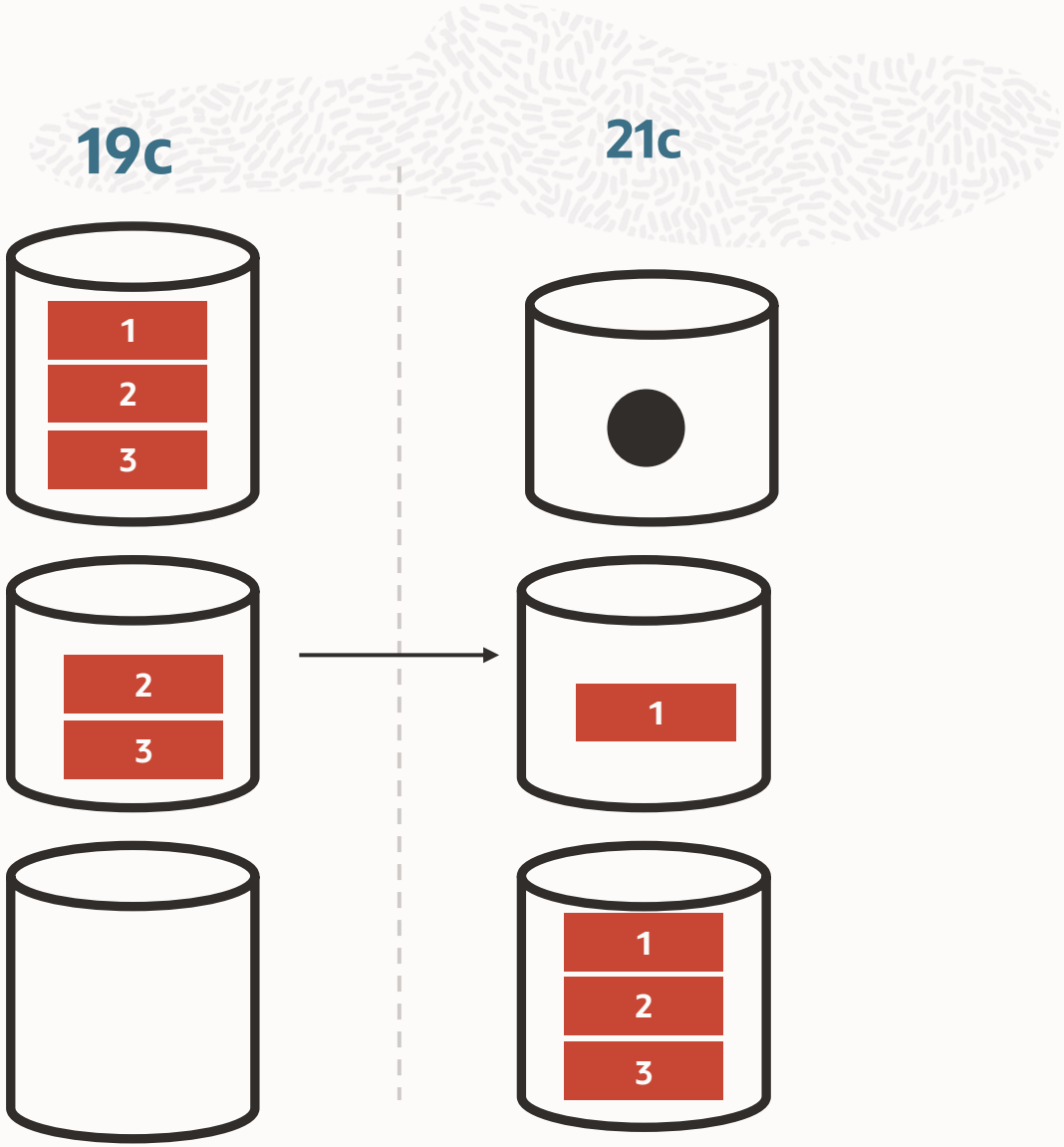


使用分片替换进行升级

添加一个空shard

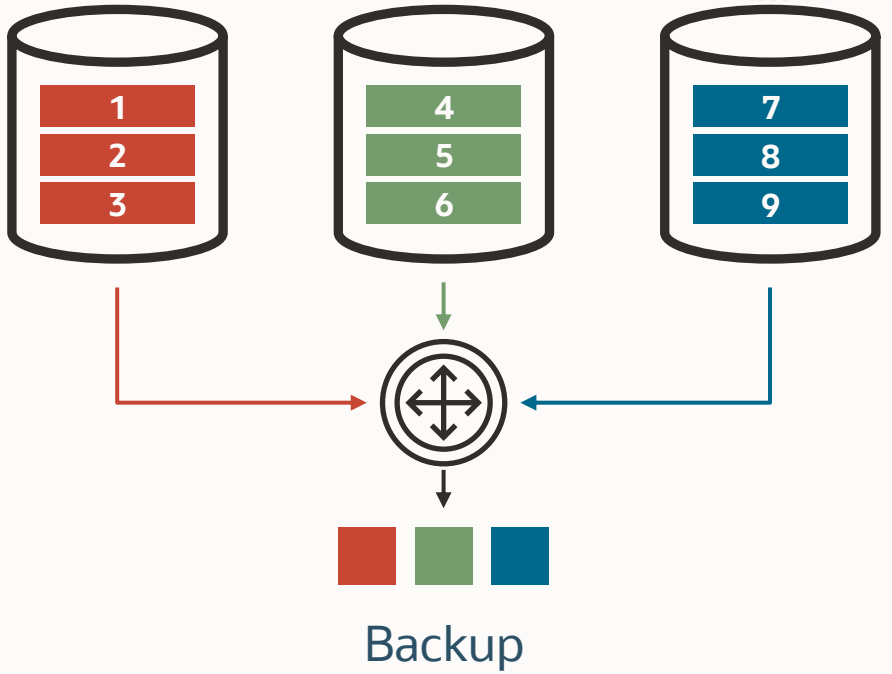
在线将chunks从旧版本分片移动到新分片

停用旧版本分片



集中备份恢复

- 分片数据库备份/恢复自动化
  - 整个分片库的一致性备份
  - 恢复粒度：chunck、单个分片或整个分片
- 感知结构变化（例如块移动）的操作
- GDSCTL管理



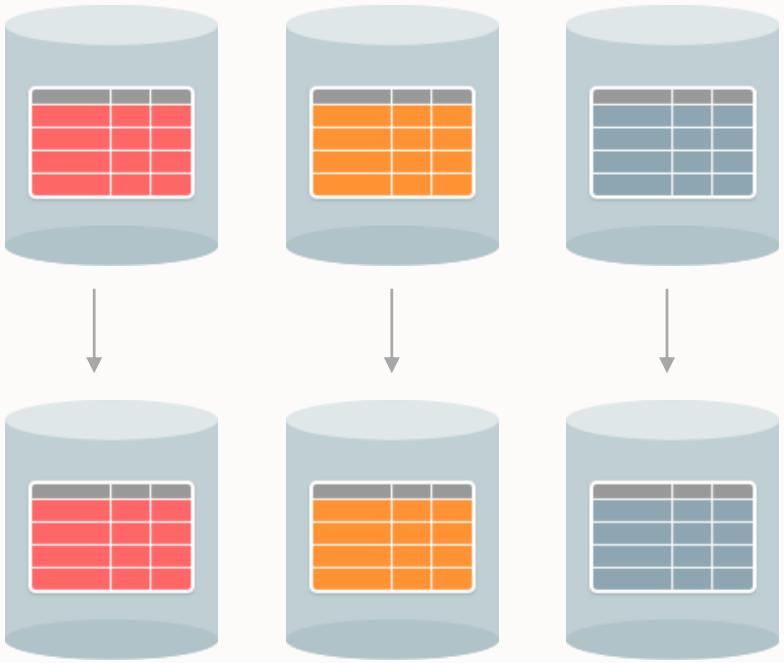

```
GDSCTL> RUN BACKUP -sync -shard ALL;

GDSCTL> RESTORE BACKUP -restorepoint
grp010119000130 -controlfile -shard shard1 -sync;
```

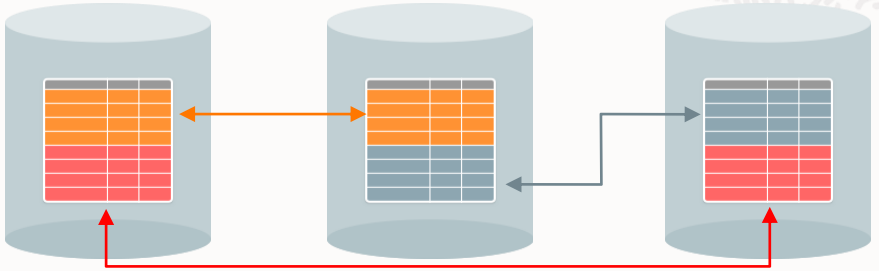
迁移至Sharding

- 转换schema 支持sharding，Sharding Advisor 可以给出建议:
    - 那些表分片，哪些表复制
    - sharding key和分片方法
- 从non-sharded database 迁移数据至sharded database
    - Data Pump，Golden Gate
- 应用侧的改动
    - 为了获得更高的性能和可扩展性，尽可能多地使用直接路由
    - 从连接池请求连接时，传递需要传递sharding key（Oracle 21c及更高版本+Java应用程序一起使用除外）

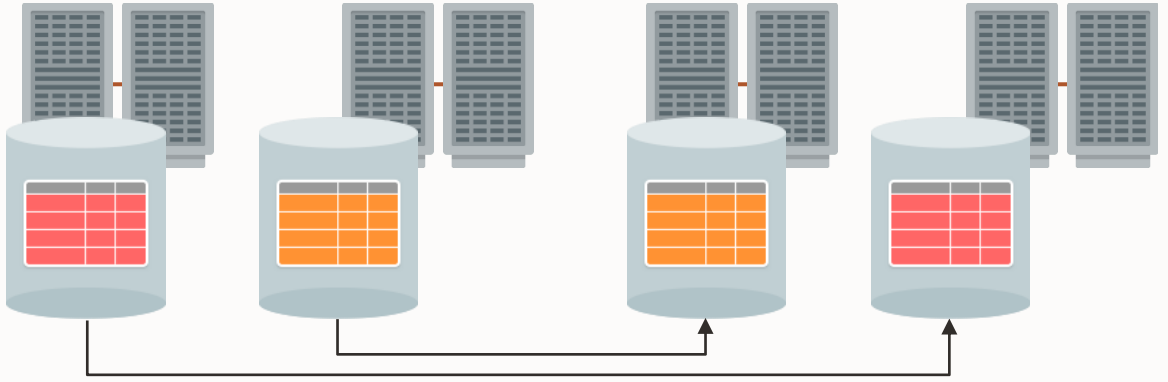
高可用/容灾 配置



Active Data Guard + Fast-Start Failover



GoldenGate ‘chunk-level’ active-active replication
with automatic conflict detection/resolution (OGG 12.3)



Optionally – complement replication with Oracle RAC for server HA

shardcat

Page Refreshed Sep 17, 2016 12:12:05 AM GMT

Summary

Sharded Database Name	orasdb
Configuration Name	oradbcloud
Catalog Database	shardcat
Catalog Version	12.2.0.1.0
Sharding Type	System-managed
Replication Type	Data Guard
Shard Directors	1 (1)
Master Shard Director	sharddirector1

Shard Load Map

Total Active Sessions : 0.05

Instance: shardcat
Total Active Load: 0.020 active sessions

Load Summary
CPU: 0.013
IO: 0.000
WAIT: 0.007

View Level : Database Instance

Members

Shardspaces | Shardgroups | Shard Directors | Shards

Name	Shardspace	Shardgroup	Data Guard Role	Region
sh1	shardspaceora	shgrp1	Primary	availability_domain1
sh1s1	shardspaceora	shgrp2	Active Standby	availability_domain2
sh2	shardspaceora	shgrp1	Primary	availability_domain1
sh2s1	shardspaceora	shgrp2	Active Standby	availability_domain2

Services

Name	Status	Data Guard Role
No services found.		

Incidents

View ▾ Target Local target and Related targets ▾ Category All ▾ 0 5 »

Summary	Target	Severity	Status	Escalation Level	Type	Time Since Last Update
Problem: KUP 600			New	-	Problem	0 days 0 hours
Problem: KUP 600			New	-	Problem	0 days 2 hours
The Data Guard fast-start failover observer status is Error Fast-Start Failover observer is no longer observing this database.			New	-	Incident	0 days 2 hours
Checker run found 1 new persistent data failures.			New	-	Incident	0 days 4 hours

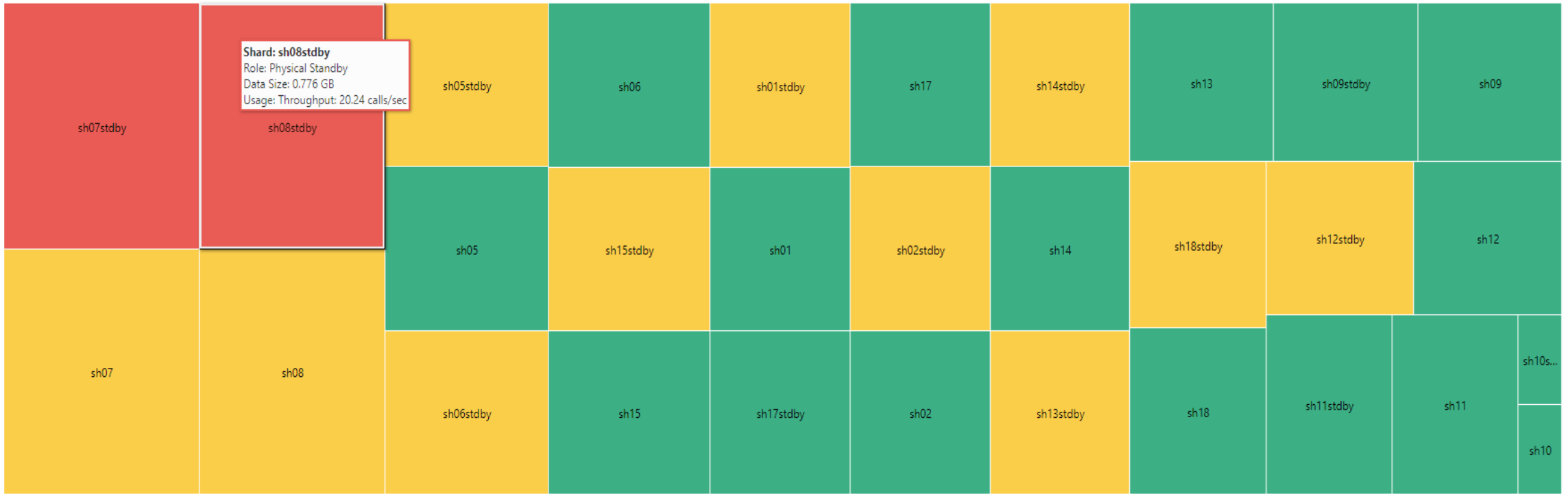
Data Distribution and Performance

Overview

Last Collection: 28-May-20 23:21:55 UTC

Regions	Shardspaces	Shardgroups	Shards	Chunks	Services
2	1	2	30 (Primary 15, Standby 15)	100	2

Use Live Performance Data: Off | View Size By: Data Size | View Color By: Usage: Throughput (calls/sec)



议程

1. Oracle Sharding概览
2. Sharding 生命周期管理
3. Oracel Sharding的演进以及新特性

Oracle Sharding 的演进

Oracle Sharding | 19c 新特性

多个PDB的分片以允许整合和故障隔离

- CDB现在可以支持多个PDB分片，方便整合
- 来自相同分片数据库的不同PDB位于不同的CDB上，以提供故障隔离

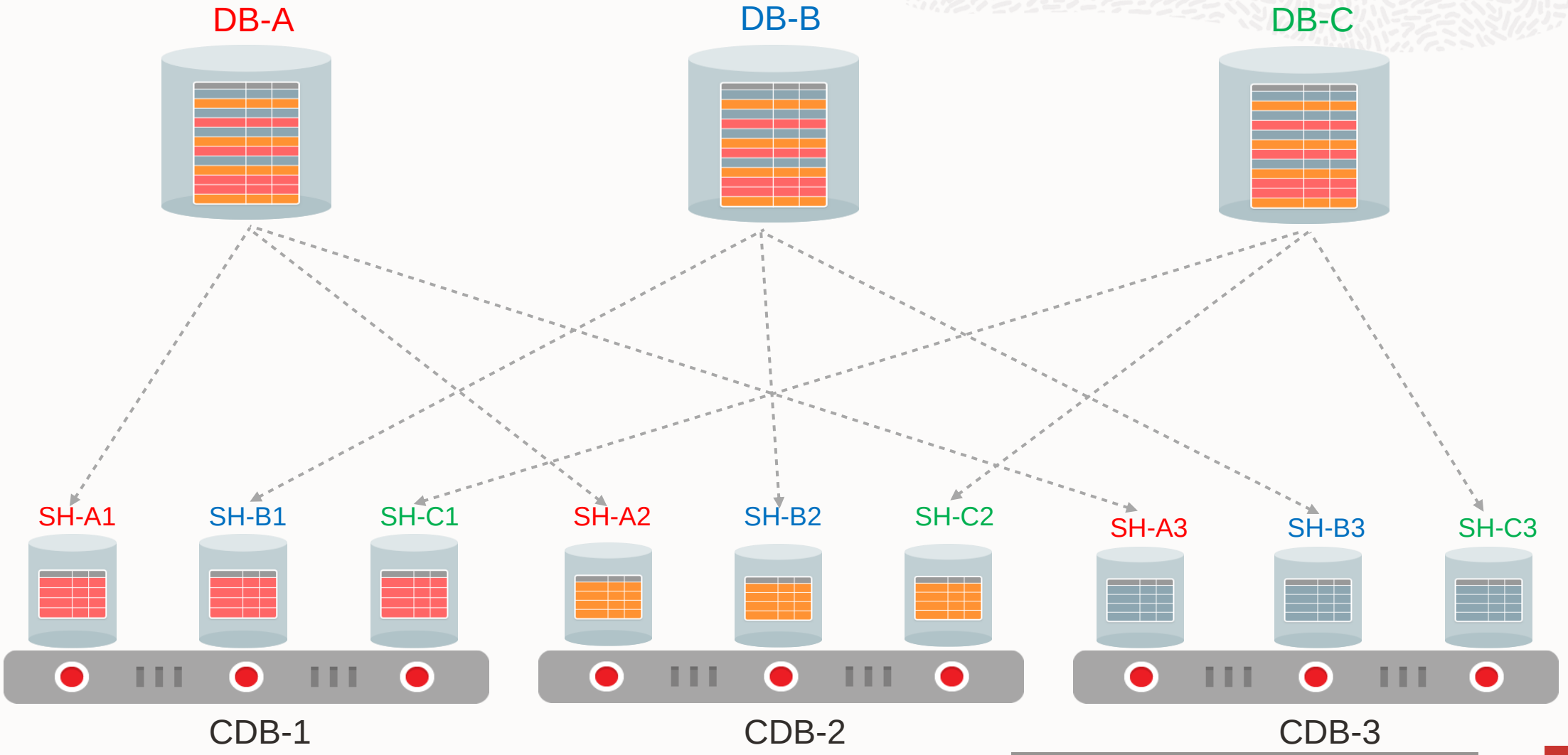
可扩展的跨分片查询协调器（coordinator），用于报告类和分析类的工作负载

- Shard catalog 的Active Data Guard备库可以充当多分片查询协调器。

通过允许在同一数据库中通过用不同的键对不同的表进行分片来提高资源利用率

- 允许分片数据库支持多个表族，每个表族都可以使用不同的分片键进行分片

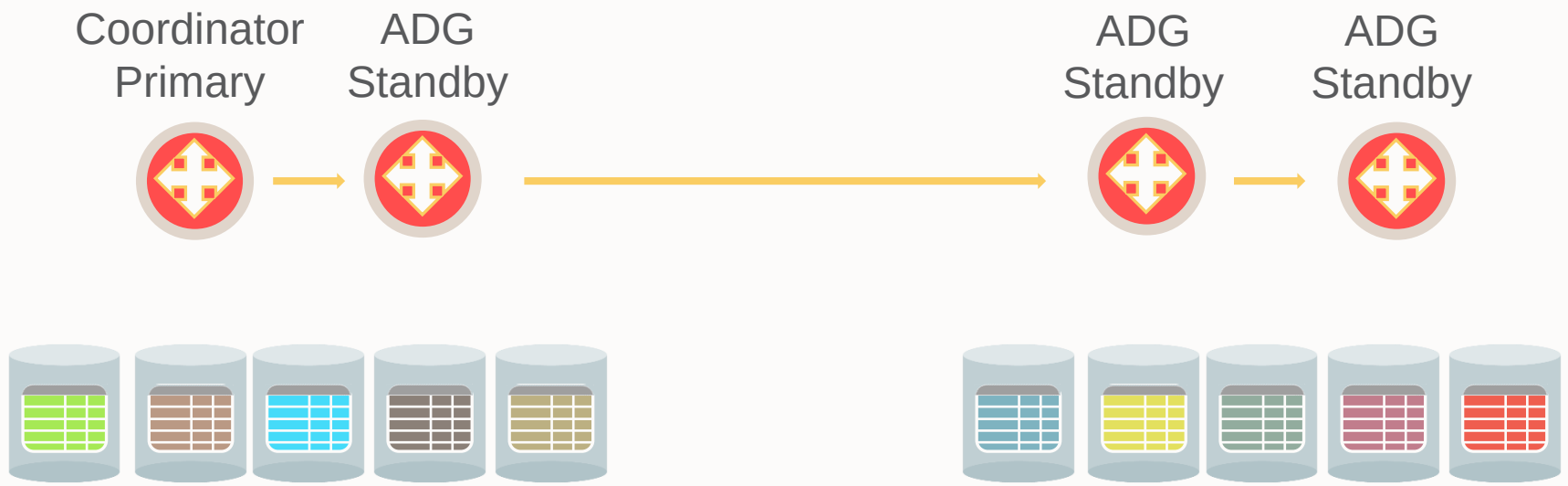
整合分片数据库



跨分片并行查询的可伸缩性

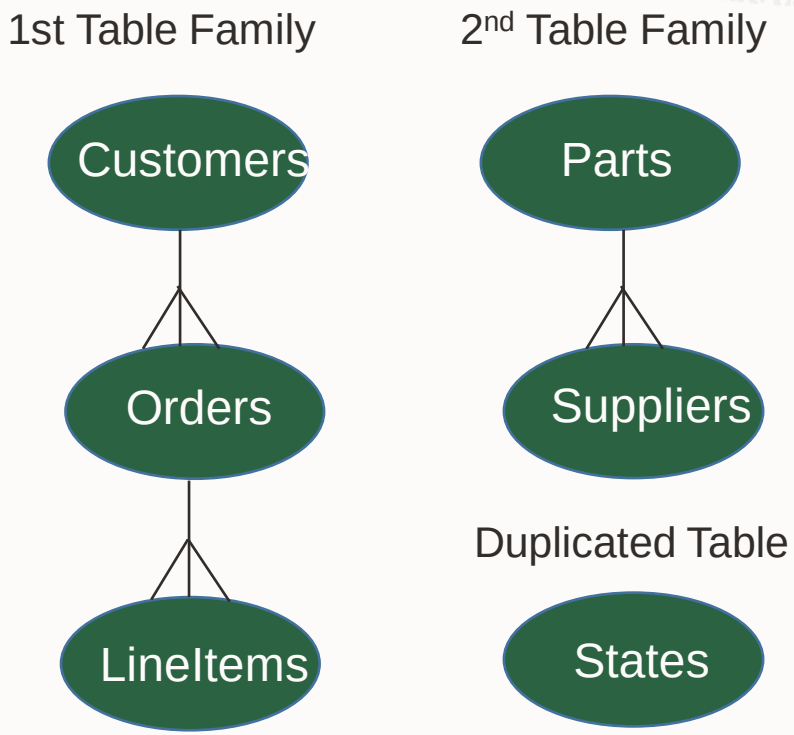
可以使用Active Data Guard备库配置多个跨分片并行查询协调器

- 改善了跨分片并行查询执行的可扩展性，可用于报告类和分析类工作负载
- 可以将协调器（Coordinators）放置在不同的可用域和地理区域中



多个表族（Table Families）

- 允许您进行权衡：
  - 减少重复数据量
  - 可能会增加执行连接的时间
- 支持系统管理的分片（一致性哈希）
- 所有表族共享同一组chunk



Oracle Sharding | 21c 新特性

联合分片（Federated Sharding）

- 允许在多个地理区域中的现有相似数据库中进行查询

跨分片DML和查询增强

- 跨所有分片并行更新
- 在分片故障转移（shard failover）的情况下，跨分片查询执行继续

自动识别分片键

- 简化应用程序设计和维护

部署自动化增强

- 使用Sharding Advisor迁移到分片部署的架构设计建议
- 使用Terraform脚本进行部署自动化

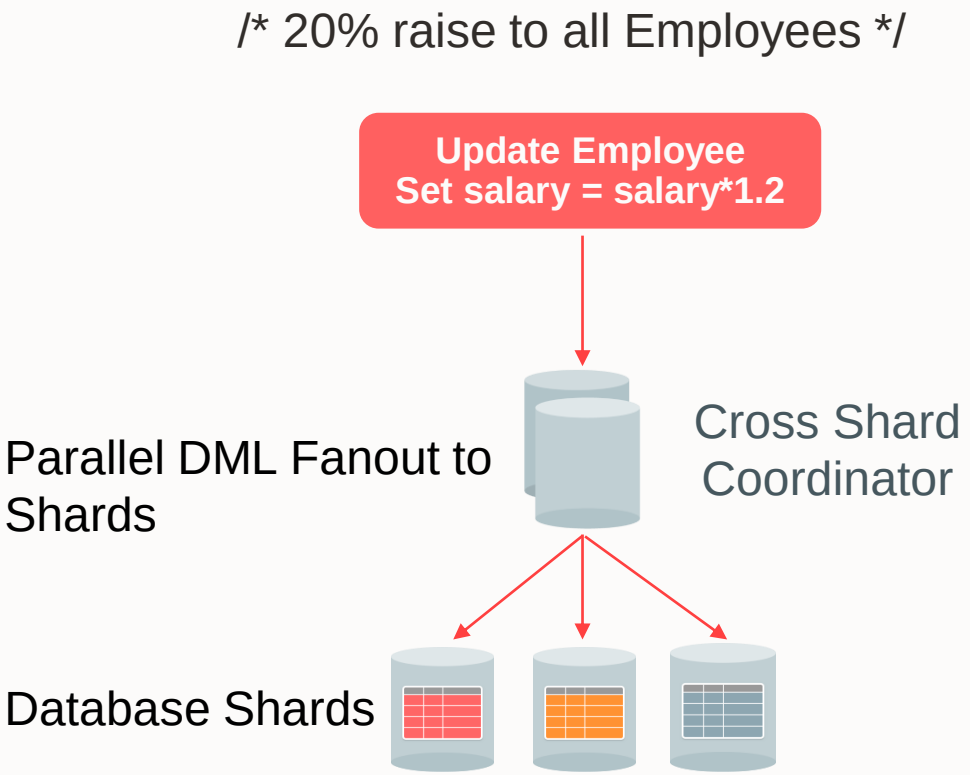
Reactive Streams Ingestion Library

- 使用直接路径写入将数据拆分并直接并行加载到分片

跨分片DML和弹性查询

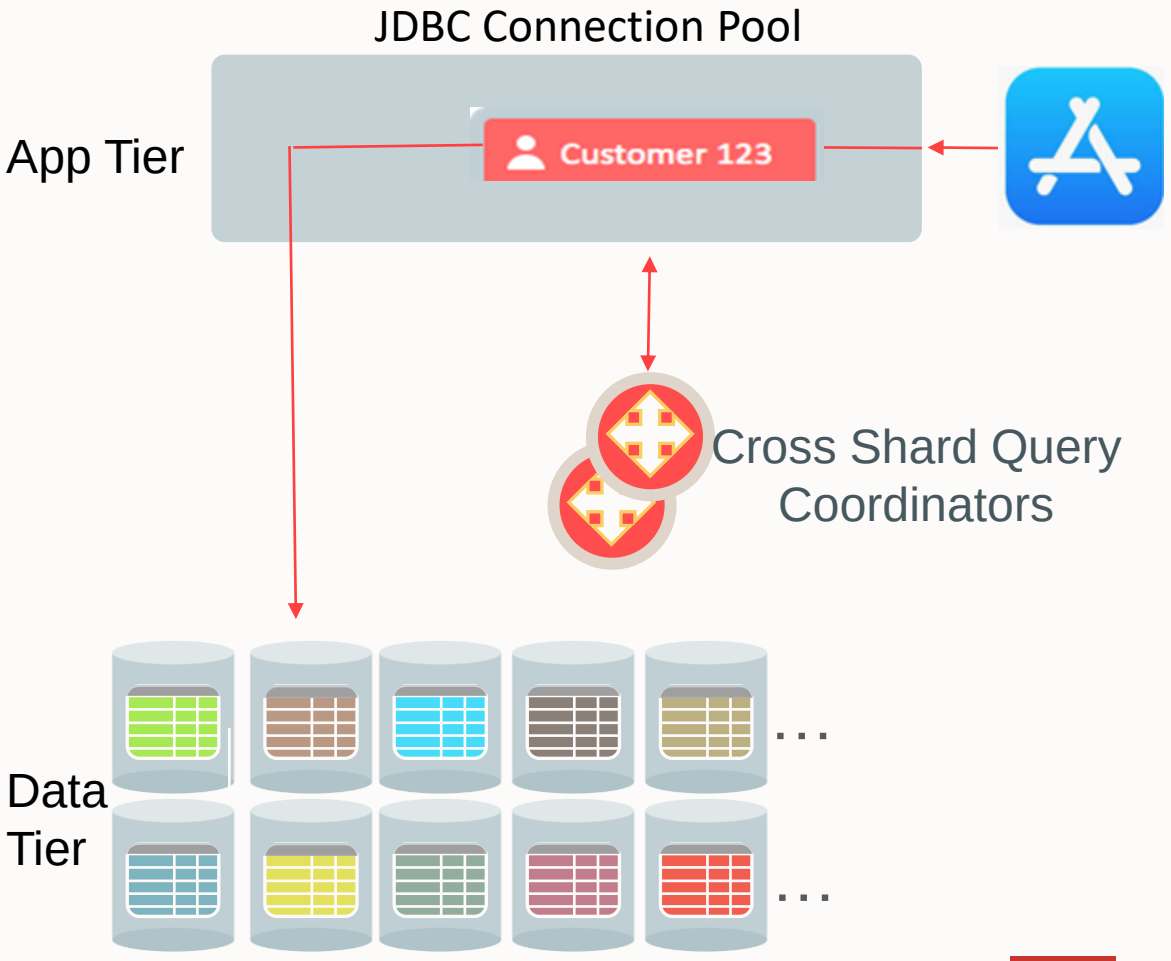
- 多分片DML支持
  - 协调器（coordinator）将DML与所有分片并行散出
  - 保留ACID交易属性

- 多分片查询弹性
  - 如果分片发生故障，Data Guard会自动将发生故障的分片故障转移到备用分片
  - 查询协调器在故障转移的分片上恢复查询执行



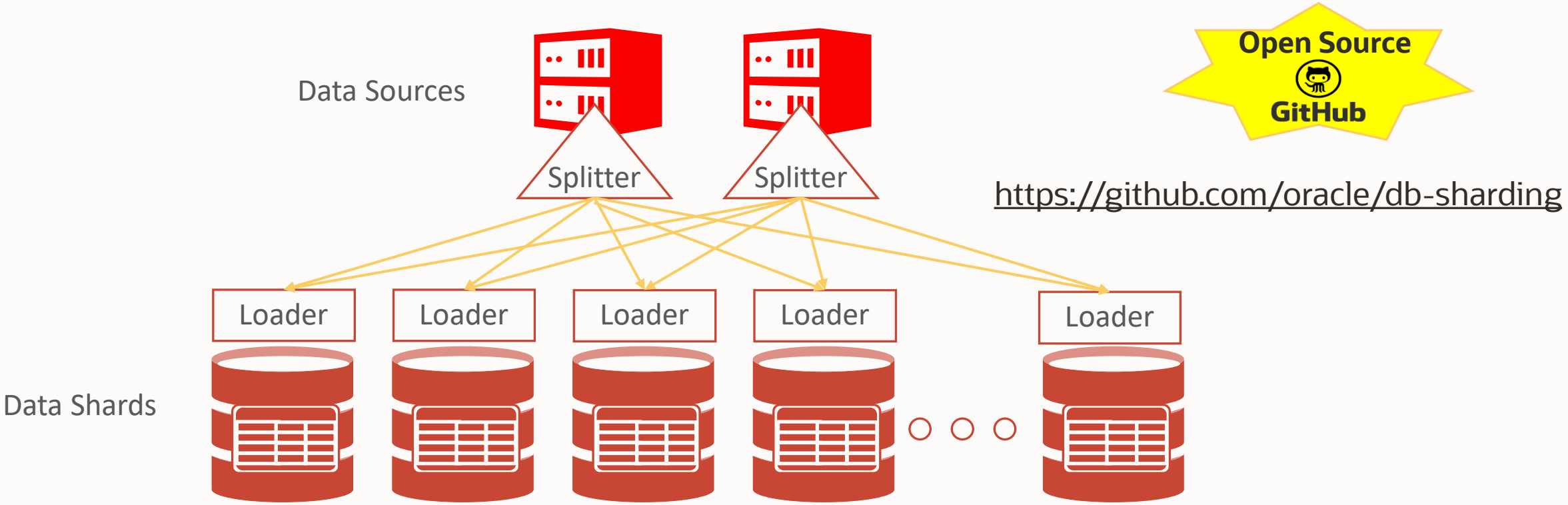
自动识别 Sharding Key

- 新的JDBC驱动程序，它使Java驱动连接到分片的数据库，而无需应用程序提供shrding key。
  - ✓ 自动识别给定SQL语句的绑定变量是分片键
- 自动查询路由
  - ✓ 单个分片查询到适当的分片
  - ✓ 通过协调器跨分片查询



高速数据加载

- 利用完全并行的直接分片数据加载器
- 适用于IIoT、IoT和边缘计算场景



Open Source
GitHub

https://github.com/oracle/db-sharding

Oracle Sharding 应用

1. RAC和Data Guard对于超过99%的应用而言都能够满足要求并且是应用透明的
2. 一些超大型的应用程序需要独立的数据库群——数据库分片
   - **避免特大型的单一数据库的可扩展性和可用性问**题
   - **愿意修改应用，把负载路由到数据库群中的特定的数据库**上

> *"Sharding should be used only when all other options for optimization are inadequate."*
>
> --维基百科

Oracle BlueKai 使用Oracle sharding案例

1 **million/second** Transactions 事务	**2.5** **petabytes** total database size 数据量	**30** **billion/day** API calls	**125** **Kilobytes/API call (Average)** API payload size
1.6 **milliseconds/API call** average read time	**2.5** **milliseconds/API call** average write time	**22** **billion** rows in largest table	**180** **terabytes/hour** redo generation rate
52 **Oracle compute instances** total machines 服务器	**2,704** **cores** total CPU	**38,740** **gigabytes** total memory	**1** **terabit/second** network traffic

Oracle Sharding 应用场景

应用场景	应用举例
应用业务流程间无依赖或耦合度低	大型计费系统
高并发访问量（秒杀类）	民航票务系统
跨国家/区域业务，海量数据	媒体类应用
数据间关系较为单一或简单	搜索/社交类应用
无数据交叉或能忍受少量数据交叉带来的性能问题	网站内容服务
无批量处理	在线游戏公司
无状态或弱状态类应用	互联网金融服务

Oracle Sharding | 资源

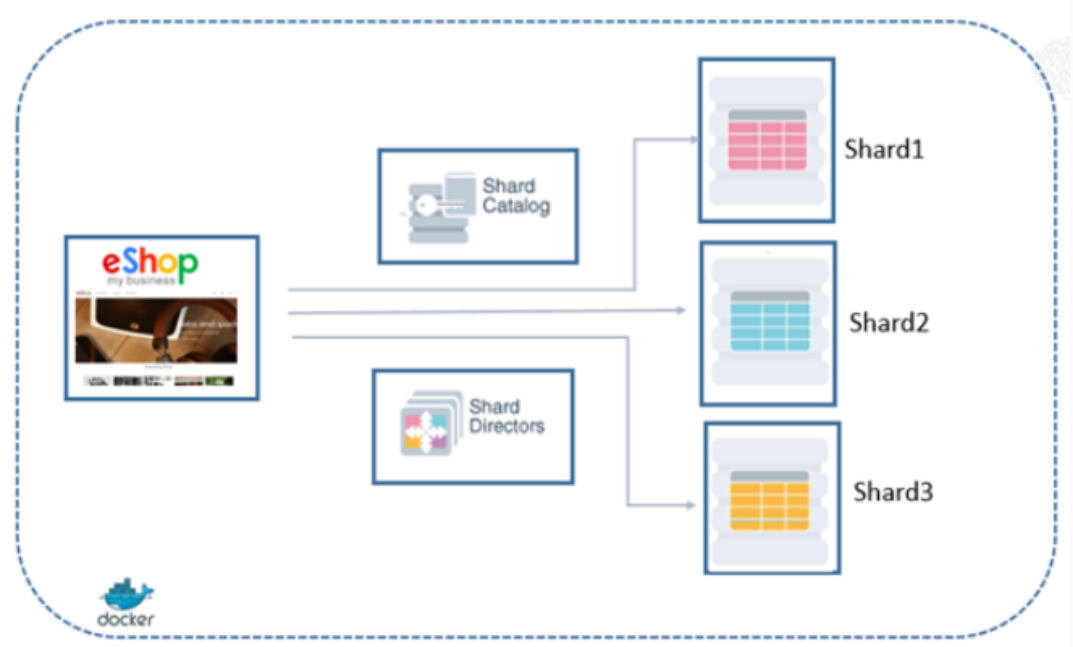
https://www.oracle.com/goto/oraclesharding

https://docs.oracle.com/en/database/oracle/oracle-database/19/shard/index.html

https://github.com/oracle/db-sharding

Oracle Sharding LiveLab

BlueKai Case Study - Video



基于 Oracle 数据库
免费企业数据健康检查

- 及时了解数据库健康状况，发现并解决潜在问题
- 维护数据库系统良好状态，保护数据资产的安全
- 提升数据库性能、稳定性和安全性，降低业务风险

免费咨询热线：

400-699-8888

* 活动最终解释权归甲骨文公司所有

数据库和云讲座群

20-21

甲骨文云技术公众号

技术专家1V1深入交流


ORACLE