PostgreSQL中文社区

2021 PostgreSQL China Conference

主办：PostgreSQL 中文社区

第 11 届 PostgreSQL 中国技术大会

开源论道 × 数据驱动 × 共建数字化未来

华为云DDS数据库容灾关键技术

党李飞

个人介绍

- 华为云高级工程师，2011年加入华为，一直从事存储和大数据领域相关开发工作。
- 2017年开始从事MongoDB数据库内核开发，有丰富的数据库内核开发，性能调优，运维经验。

目录

- **DDS简介**
- DDS（文档数据库）灾备技术原理
- 原生Change Streams分析
- DDS对于Change Streams的优化
- 总结

概述

文档数据库 DDS（Document Database Service）**完全兼容 MongoDB 协议**，在华为云高性能、高可用、高安全、可弹性伸缩的基础上，提供了一键部署，弹性扩容，容灾，备份，恢复，监控等服务能力。目前支持**分片集群（Sharding）、副本集（ReplicaSet）、单节点（Single）**三种部署架构。

MongoDB的数据结构

```
{
  "_id": 1,
  "name" : { "first" : "John", "last" : "Backus" },
  "contribs" : [ "Fortran", "ALGOL", "Backus-Naur Form"],
  "awards" : [
    {
      "award" : "W.W. McDowell Award",
      "year" : 1967,
      "by" : "IEEE Computer Society"
    }, {
      "award" : "Draper Prize",
      "year" : 1993,
      "by" : "National Academy of Engineering"
    }
  ]
}
```

MongoDB存储结构

- **文档（Document）**：MongoDB中最基本的单元，由BSON键值对（key-value）组成。**相当于关系型数据库中的行（Row）。**
- **集合（Collection）**：一个集合可以包含多个文档，**相当于关系型数据库中的表（Table）。**
- **数据库（Database）**：等同于关系型数据库中的**数据库概念**，一个数据库中可以包含多个集合。您可以在MongoDB中创建多个数据库。

DDS 的产品优势

MongoDB

100% 兼容 MongoDB

- 具备无需业务改造，直接迁移上云的能力
- 支持社区3.4/4.0版本

3种架构

集群、副本集、单节点

- 集群： nTB存储、在线扩容
- 副本集：2TB存储，3副本
- 单节点：高性价比

高可用

架构高可用、跨AZ部署

- 支持副本集，Shard高可用架构（集群）
- 副本集多节点（三、五、七）
- 集群、副本集支持跨AZ部署

高可靠

自动/手动备份，数据恢复

- 每天自动备份，保留 732 天
- 手动备份，永久保存
- 备份恢复

高安全

- 具备多层安全防护
- 网络：VPC 网络隔离
- 传输：SSL 安全连接
- 访问：安全组出、入限制

管理、监控

- 可视化监控：CPU、内存、IO、网络等
- 实例一键扩容、规格变更
- 错误日志、慢日志管理
- 参数组配置

DDS服务部署形态——单节点（Single）

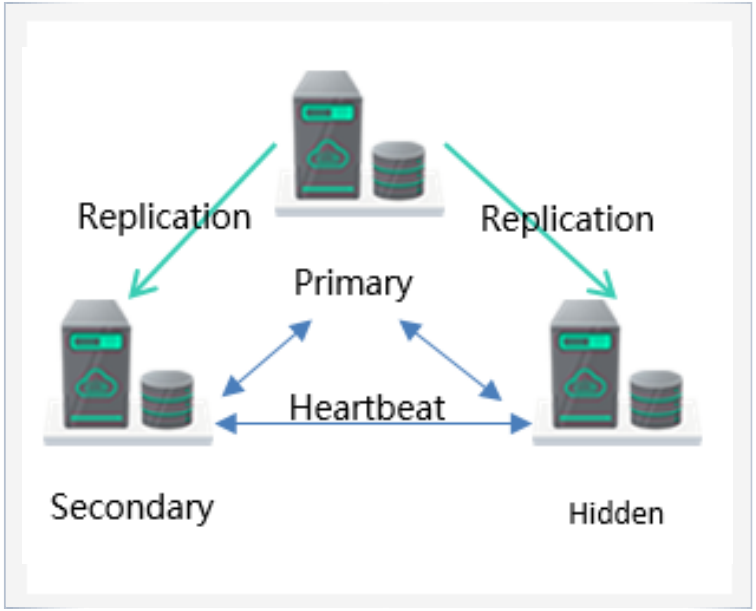
架构特点

1. 超低成本，仅需支付一个节点的费用；
2. 支持10GB-1000GB 的数据存储；
3. 较副本集/集群可用性不高：当节点故障，业务不可用；

适用场景

- 非核心数据存储
- 学习实践；
- 测试环境的业务；

DDS部署形态——副本集（Replica Set）



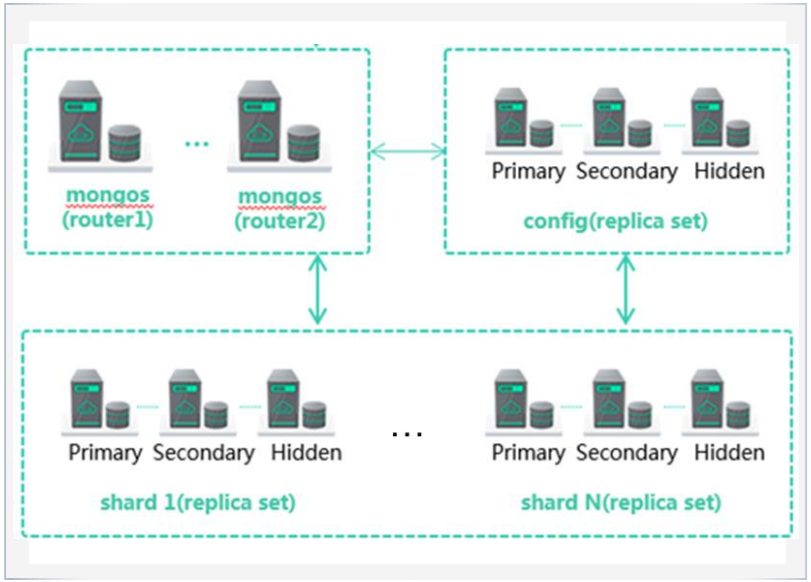
架构特点

1. 三节点高可用架构：当主节点故障时，系统自动选出新的主节点
2. 支持10GB-3000GB 数据存储；
3. 具备扩展到5节点，7节点副本集的能力。

适用场景

- 有高可用需求，数据存储<3T

DDS服务部署形态——集群（Sharding）



架构特点

1. 组件构成：由 mongos（路由）、config（配置）、shard（分片）三种类型的节点构成
2. Shard 分片：每个 shard 都是一个副本集架构，负责存储业务数据。可创建2-16个分片，每个分片10GB-2000GB。因此，集群空间范围（2-16）*（10GB-2000GB）
3. 扩展能力：在线规格变更、在线横向扩展

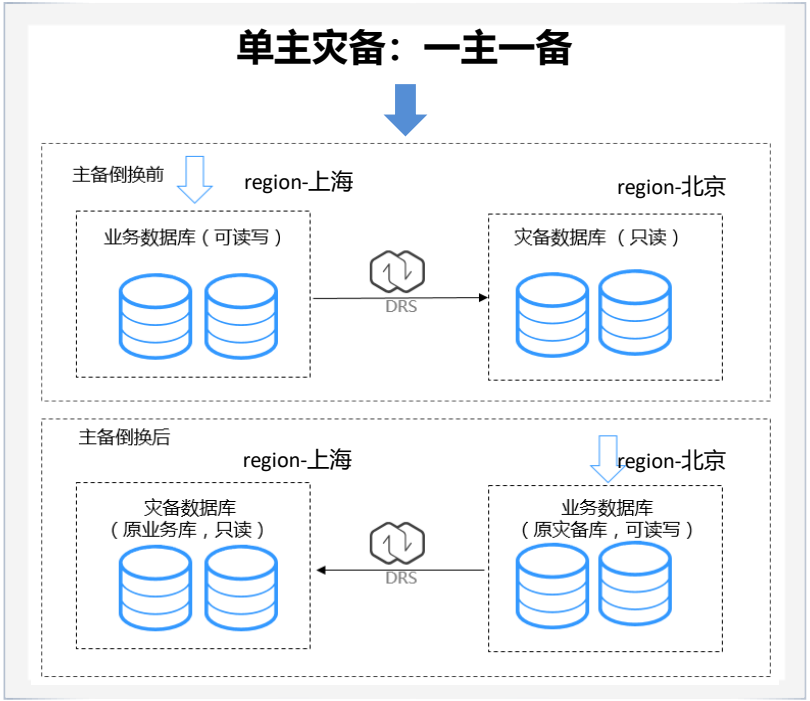
适用场景

- 要求高可用，数据量大且未来横向扩展要求

目录

- DDS简介
- **DDS（文档数据库）灾备技术原理**
- 原生Change Streams分析
- DDS对于Change Streams的优化
- 总结

灾备技术原理



- 可以进行主备倒换
- 手动一键倒换，提供批量主备倒换API

约束：当前仅支持华为云-华为云同版本灾备

主备倒换流程

1.业务停写region-上海原DDS+数据库
2.操作DRS任务主备倒换， region-北京的DDS+数据库由只读变为可读写
3.业务写region-北京的数据库，DRS将数据同步回region-上海数据库

灾备技术原理

- RPO（Recovery Point Objective），为业务数据库与DRS实例数据差的一种度量方式
- RPO=0时，意味着**业务数据库的数据已经全部到达DRS实例。**

- RTO（Recovery Time Objective），处在传输中数据量的一种度量方式
- RTO=0时，意味着**DRS实例上的事务已经全部在灾备数据库上执行完毕**

灾备技术原理

灾备初始化

全量同步

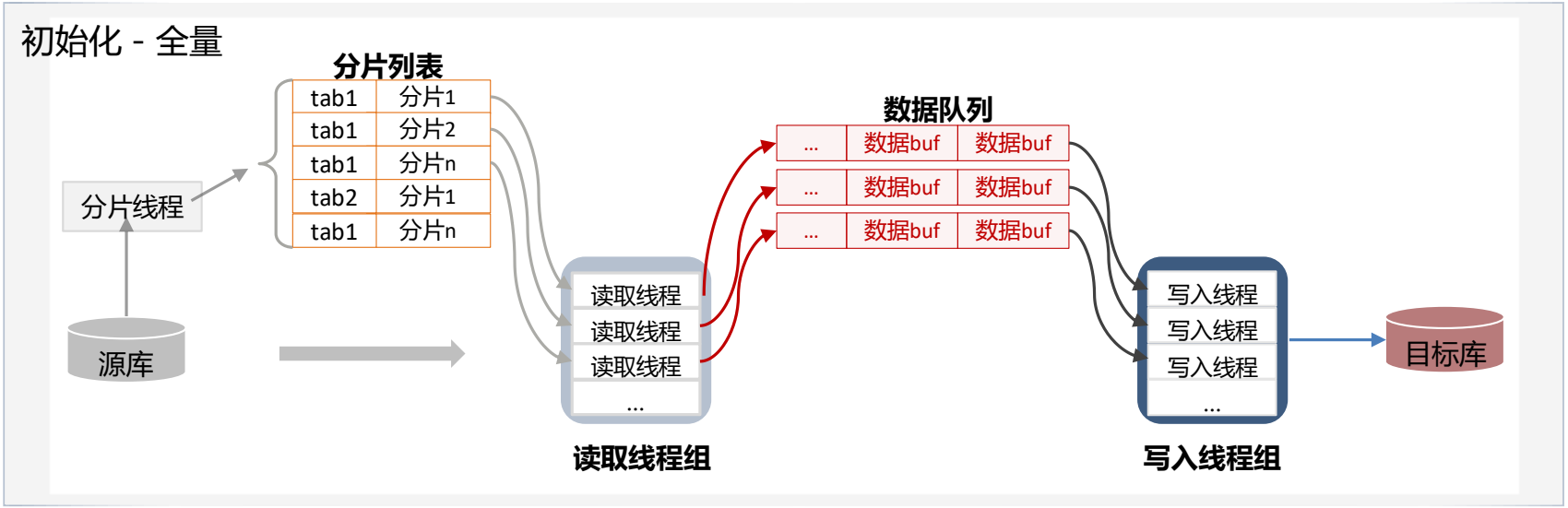
- 将源业务库的全量数据一次性同步到目的灾备库
- 支持对分片键、索引、视图、用户、角色的同步
- 支持对表进行分分片同步

灾备中

增量同步

- 实时解析源库日志，并将解析到的变更数据转换为DRS内存存储格式
- DRS将抓取到的数据落盘存储
- 读取落盘数据，重构成对应的语句在目标库回放

灾备技术原理

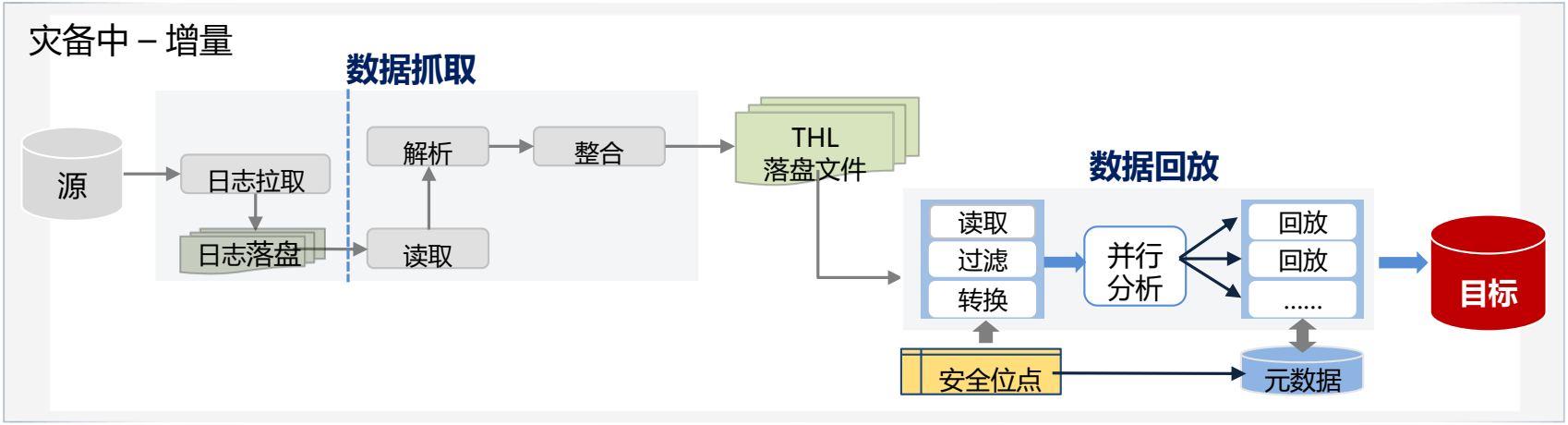


分片线程：从源库将要同步的表信息导出，并对每张表进行分片，生产分片列表

读取线程组：从分片列表获取未同步的分片，根据分片范围从源库获取该分片的全部数据，并将数据写入数据队列中，

写入线程组：从数据队列获取数据，构建sql语句将数据写入到目标库

灾备技术原理



数据抓取

日志拉取：通过Oracle Logminer读取原库日志，并存储到本地磁盘

读取->解析->整合：从磁盘读取日志数据，解析日志中有效变化记录，并按原库发生的先后顺序整合成事务

THL落盘文件：整合后的数据被转换为DRS内部存储格式，并写入磁盘

数据回放

读取->过滤->转换：读取THL文件数据，并对数据进行过滤和转换，形成目标库可应用的数据。

并行分析：评估各记录之间的依赖关系，生产可并行回放的数据队列

回放：多线程并发的将数据写入到目标库

目录

- DDS简介
- DDS（文档数据库）灾备技术原理
- **原生Change Streams分析**
- DDS对于Change Streams的优化
- 总结

原生Change Streams分析--1

什么是Change Streams

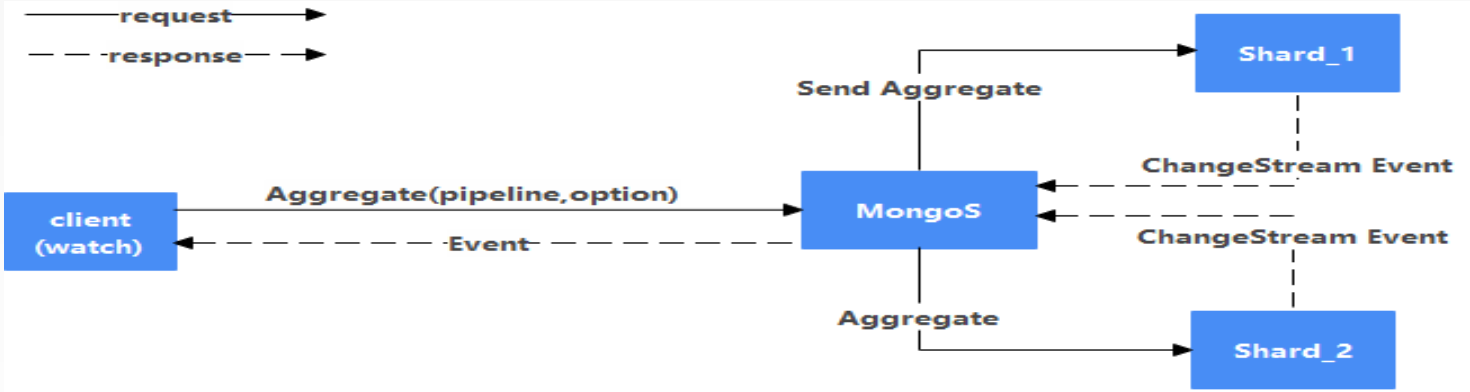
Change Streams可以直译为"**变更流**"，也就是说会将数据库中的所有变更以流式的方式呈现出来。用户可以很方便地对数据库建立一个监听（订阅）进程，一旦数据库发生变更，使用change stream的客户端都可以收到相应的通知。使用场景可以包括但不限于以下几种：

1）多个MongoDB集群之间的增量数据同步；

2）高风险操作的审计（删库删表）；

3）将MongoDB的变更订阅到其他关联系统实现离线分析/计算等等

原生Change Streams分析--2



集群场景下，灾备实例之间的日志拉取依靠Change Steams完成

1 客户端向MongoS发起了一个Aggregate命令,把该命令发给对应的Shard节点,同时服务端向客户端返回一个游标。

2 Shard Server端收到Aggregate命令后，扫描oplog集合，读取数据并返回给Mongos.

3 客户端通过Mongos节点拿到游标ID，在该游标上不断的执行getMore请求，来获取整个集群的oplog（操作日志）信息。

原生Change Streams分析--3

1 事件拉取性能有待提升：

如之前分析，当前的Change Streams请求发到Mongos节点后，通过单线程的方式向每个Shard节点发送异步请求命令来完成数据的拉取，并做数据归并，如果将该方式替换为多线程并发拉取，对于分片表来说，性能会有提升。

2 支持DDL事件不完善，对于集合和DB删除事件导致事件监听中断，需要重新开始。
Change Stream目前支持的事件如下：

Insert Event：数据插入
Update Event： 数据更新
Replace Event：数据替换
Delete Event：删除数据
Drop Event：删除集合
Rename Event：重命名集合
DropDatabase Event：删除DB
invalidate Event:非法事件

目录

- DDS简介
- DDS（文档数据库）灾备技术原理
- 原生Change Streams分析
- **DDS对于Change Streams的优化**
- 总结

并发Change Streams架构

- **Change Streams Buffer:**

与Shard是一对一的关系。每个Change Streams Buffer 默认1GB，在Buffer满之前，该Buffer无条件的向对应的Shard(secondary节点)拉取Change Streams数据

- **Merged Queue**

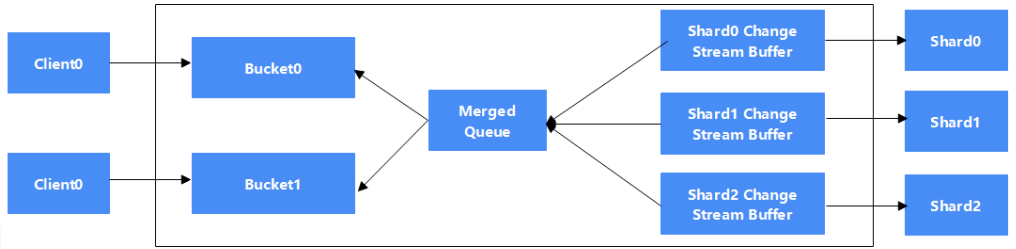
Merged Queue是一个内存队列，是Change Streams Buffer的消费者，是 Bucket的生产者。Merged Queue 归并所有Shard的Change Streams Buffer，并等待合适的时机按照规则放入对应Client的Bucket。

- **Bucket**

Bucket 是一个内存队列，是MergedQueue的消费者，是Client的生产者。每个Client对应一个Bucket。每个Bucket维护该Bucket内所有文档的的集合。

- **Merged Queue 与Bucket的交互过程**

Merged Queue不停的从头部拿出尽可能多的数据，并从前往后的按照hash(document.ns)%n的规则放入对应的Bucket, document.ns是指这个文档的NameSpace， 所以同一个集合的数据一定在一个Bucket里面。



DDL事件的增强

并发Change Stream除了支持原生的Change Stream外，还新增支持如下事件：

- CreateCollection Event：创建集合
- CollMod Event： 修改集合属性
- CreateIndex Event： 创建索引
- Drop Index Event： 删除索引
- CreateView Event： 创建视图
- DropView Event： 删除视图
- ShardCollection Event： 对集合分片

目录

- DDS简介
- DDS（文档数据库）灾备技术原理
- 原生Change Streams分析
- DDS对于Change Streams的优化
- **总结**

使用华为云DRS做灾备的优势

- **DRS灾备与Mongoshake灾备的对比**

	华为云DRS灾备	Mongoshake
倒换方式	一键倒换	手动配置参数倒换
倒换位点	不需要指定增量位点	手动配置位点
数据一致性	平台展示，可以进行对象、行数、内容三种对比粒度，内容对比可以展示所有不一致数据	脚本对比，有对象、行数和抽样内容，抽样内容无法展示所有不一致数据，可能会漏掉不一致数据
同步进度	平台展示时延监控，时延为0代表追平，可设置阈值告警，随时关注灾备状态	api调用，参数多且繁杂，对比位点是否相同来判断是否追平
主键_id支持	支持所有类型_id混合且不影响并发	源库集合只支持单一类型_id
断点续传	全量、增量都可以断点续传	增量断点续传
项目投入	极小人力成本，不需要客户申请部署ecs，创建DRS灾备任务即可完成灾备	前期基础设施部署和网络需要人力成本

华为云DDS对社区版的优势

	华为云DDS	社区版MongoDB
ChangeStream内DDL操作	支持丰富的DDL：CollMod/CreateIndex/DropIndex/CreateView/DropView, ChangeStream将对应的Oplog封装成事件，解决DDL操作不全的问题	DDL操作不完全，缺少对ShardCollection/CollMod/CreateIndex/DropIndex/CreateView/DropView的支持
	对于集合删除，数据库删除事件，日志拉取不中断	碰到集合删除和数据库删除事件会导致Change Streams中断，需要重新开始监听

华为云DDS对社区版的优势

	华为云DDS	社区版MongoDB
性能	单个ChangeStream能被多个客户端消费	客户端只能单线程拉取数据(单线程意味着CPU最高到100%，对于大表容易追不上)
	mongos上load数据和客户端消费数据独立，各自streaming处理	客户端拉取数据时，数据在mongos上并未准备好(因为没有预读)，要当场从mongod上查
ops	50000左右	10000左右

THANKS

欢迎关注GaussDB数据库公众号

开源论道 × 数据驱动 × 共建数字化未来