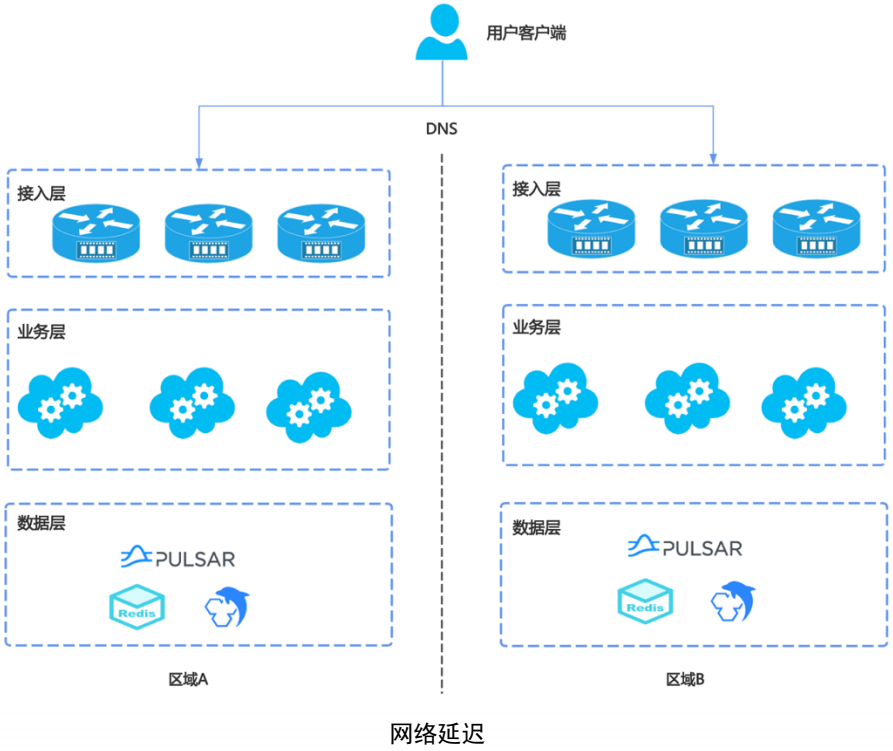
Pulsar如何满足金融级的容灾场景

刘德志

- AscentStream 技术合伙人
- 前腾讯计费平台 TDMQ初创负责人
- 腾讯专家工程师

容灾核心诉求是什么?

当出现灾难时，如何快速恢复业务，减少损失。



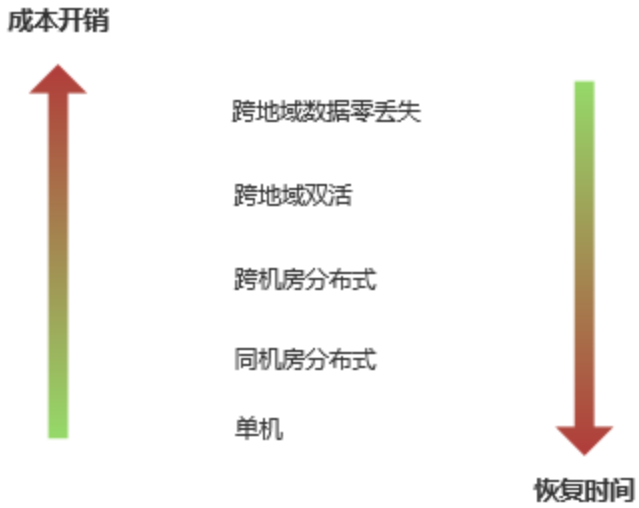
灾

- 单机问题
- 单机房
- 单AZ
- 单区域

冗

?

容灾的成本开销与恢复时间相互制约

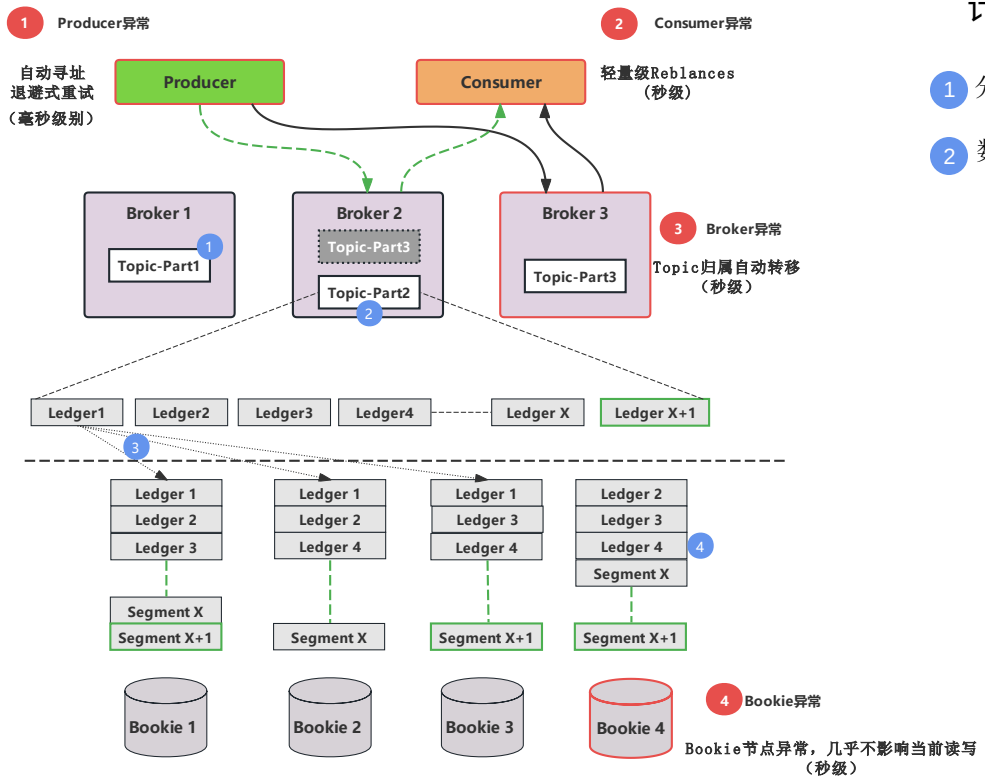


离钱越近，要求越高

AscentStream Platform 基于Apache Pulsar
如何满足金融级的消息中间件容灾场景？

一、Pulsar单集群容错能力

分布式、故障转移、故障自愈

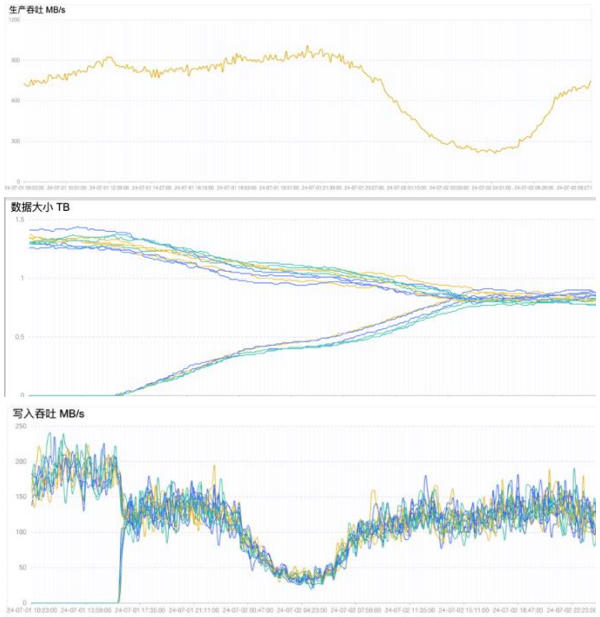


计算层：

1. 分区级**动态均衡**到不同broker
2. 数据**分片条带化**写入bookie

存储层：

3. 数据多**副本均衡**到不同bookie
4. 分片存储，新增节点**快速加入**

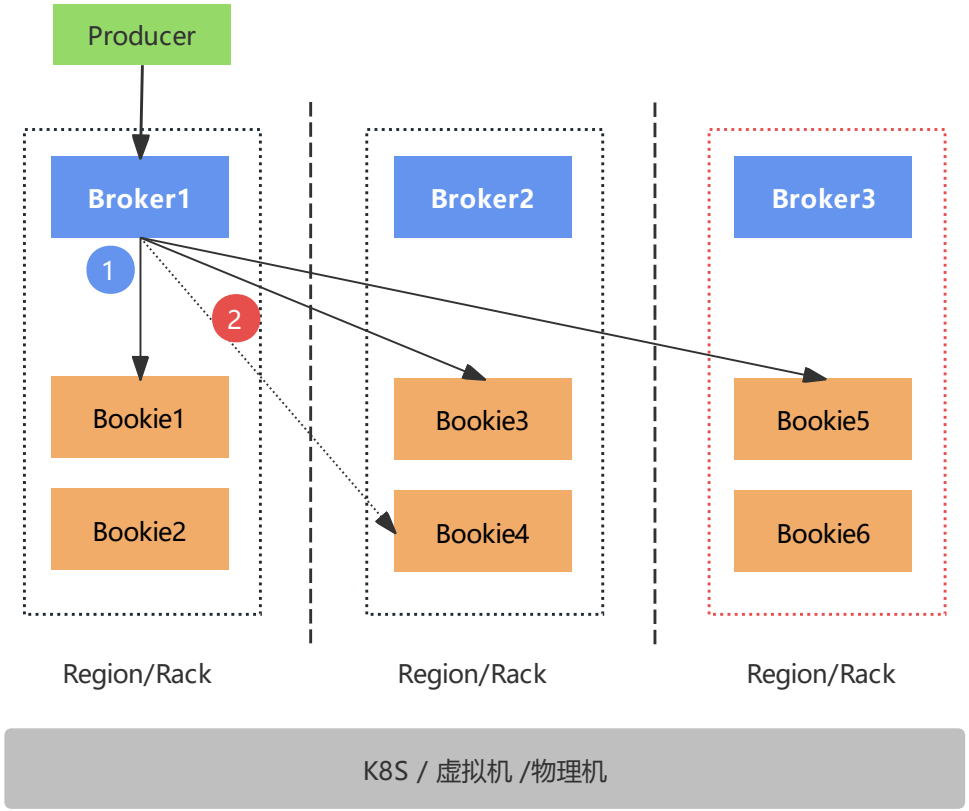


扩容双向奔赴

Pulsar特性：云原生架构

二、Pulsar单集群跨机房能力

机架感知、故障降级



同城容灾：

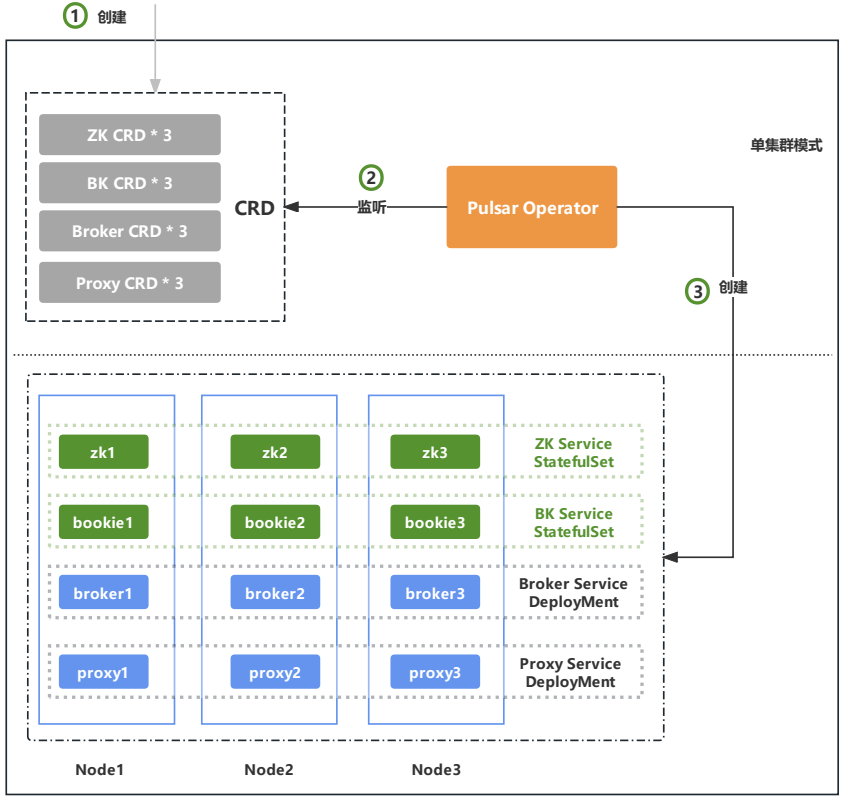
1. 支持配置机架感知配置
2. 支持降级处理，保可用性

虚拟机 / 物理机可轻松支持跨机房容灾。

但K8S环境下，将提出更高的要求：

- 有些情况下，k8s集群不允许跨机房。
- 不希望因为k8s集群异常导致整个服务不可用。

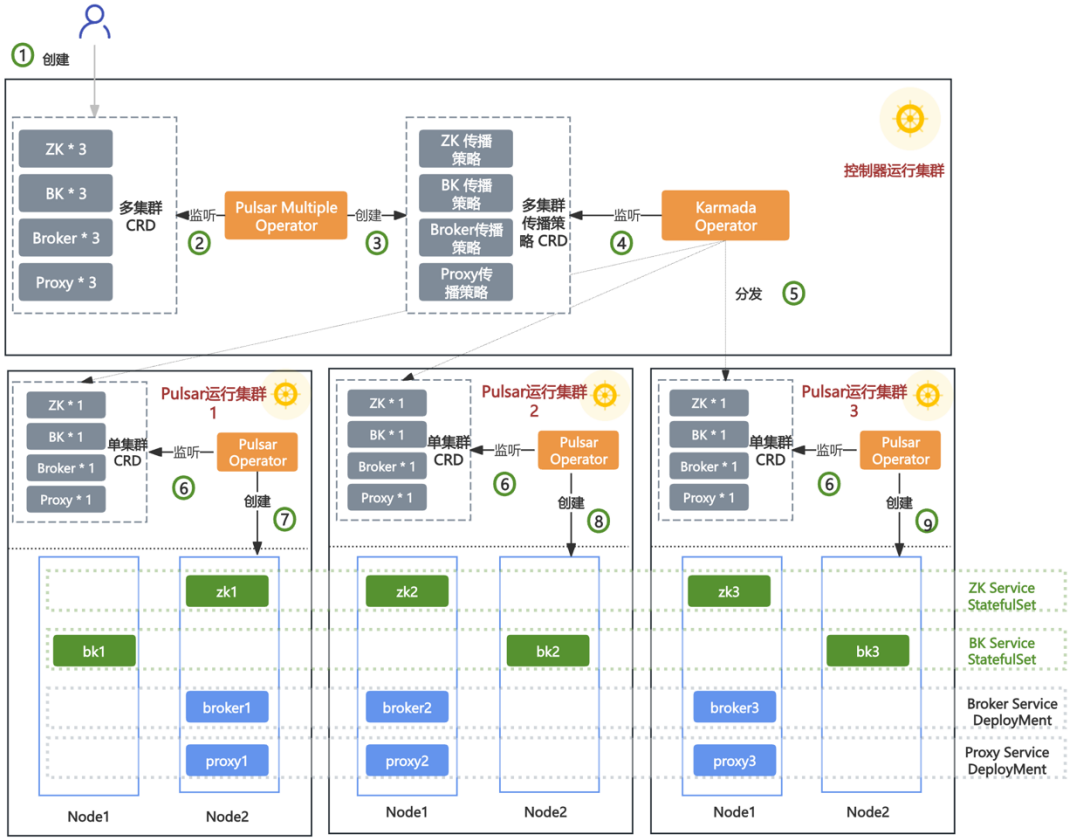
三、Pulsar单集群K8S部署方案



Pulsar Operator 是基础

- 负责单K8S集群下 Pulsar相关资源的创建和管理
- 实现自定义流程部署

四、Pulsar单集群跨K8S部署方案



三大关键组件

- Pulsar Multiple Operator

  负责定义多集群部署模式下资源（CRD)

- Karmada Operator

  负责K8S多集群统一控制

  负责K8S多集群统一调度和策略传播

- Pulsar Operator

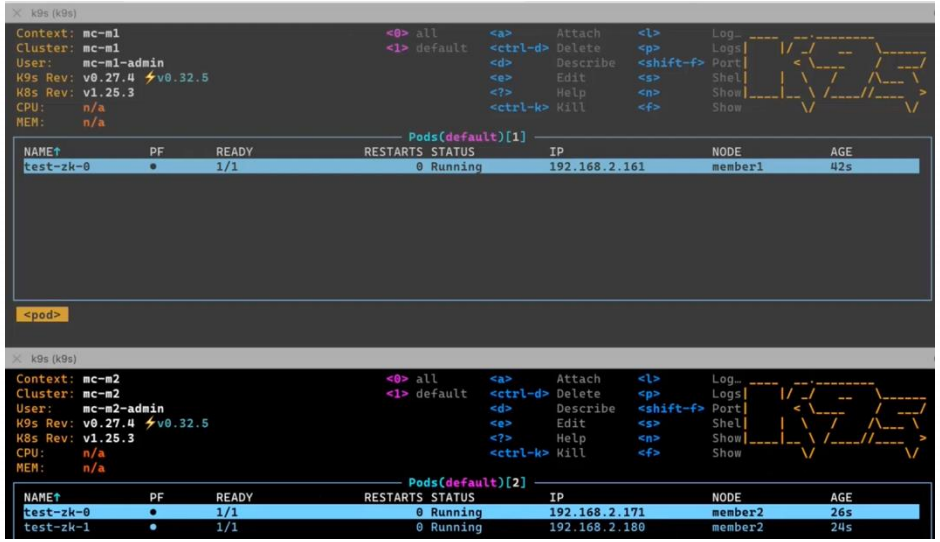
  联动完成单集群资源创建和管理

主要挑战

- 跨集群网络互通
- 全局服务发现
- 跨K8S Statefulset 资源管理

四、Pulsar单集群跨K8S部署能力

Statefulset 顺序控制



可用性能力

故障级别	ZK 服务	Bookie 服务	Broker 服务	Proxy 服务
Pod 级别	无影响	无影响	无影响	无影响
Node 级别	副本数不少于半数无影响	副本数不少于半数无影响	无影响	无影响
K8S 集群级别	副本数不少于半数无影响	副本数不少于半数无影响	无影响	无影响
Operator 级别	无影响	无影响	无影响	无影响

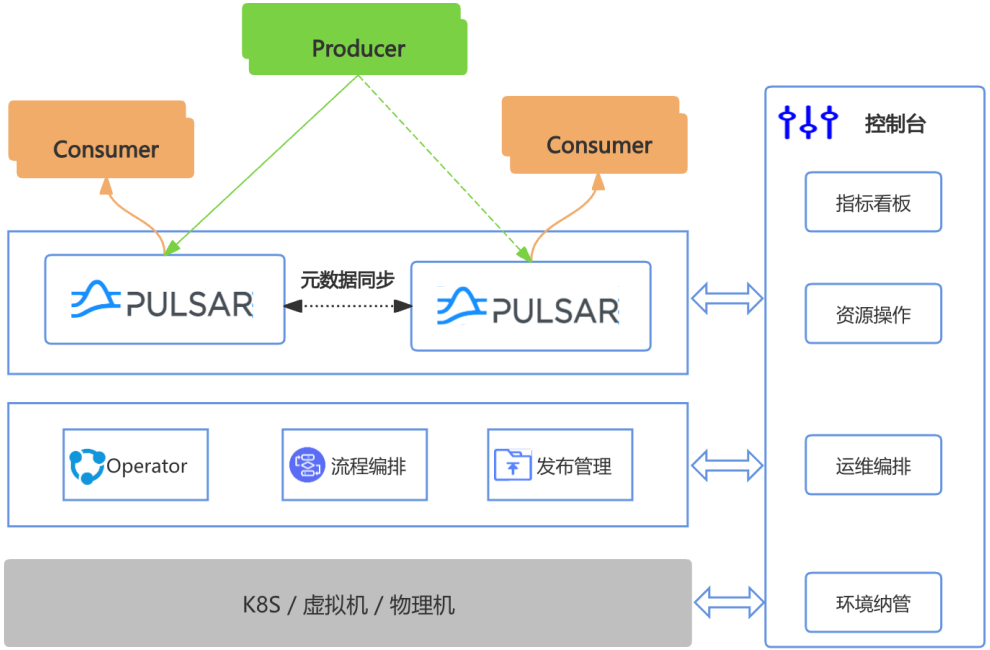
Pulsar 在单集群下的容灾能力得益于存算分离的云原生架构

Pulsar将是一个不错的选择

多集群容灾如何呢？

五、Pulsar多集群灾备方案一

保证业务的连续性



集群级别高可用

- 多集群配置保持一致，需要一套控制台将这个复杂的过程管理起来。
- 生产和消费在多个集群中同时处理，保证业务的连续性。

优势

- 架构简单
- 服务可用性高
- 故障切换快

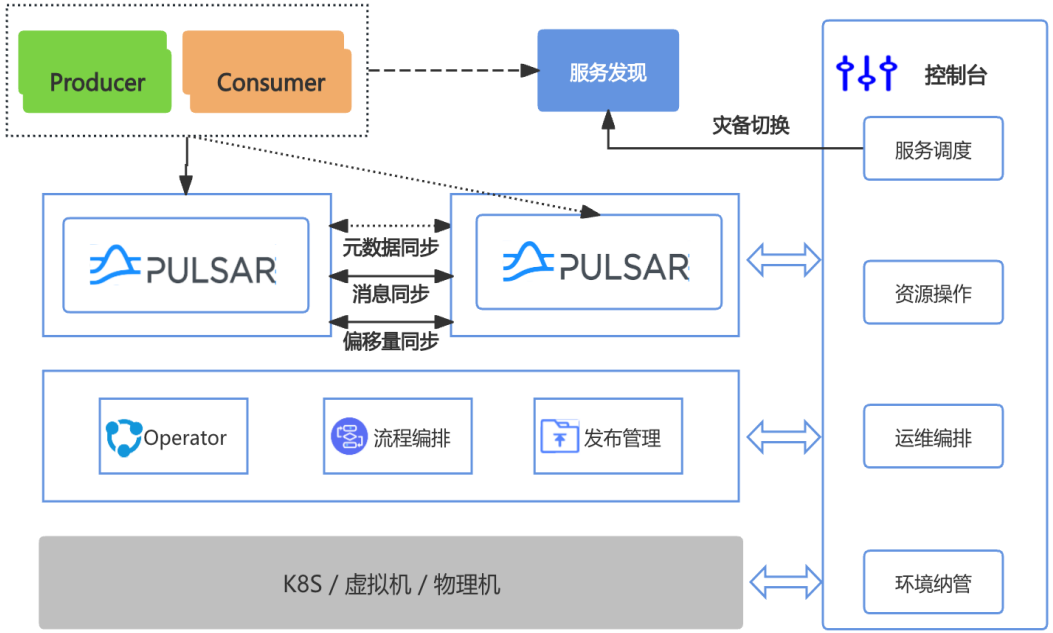
劣势

- 集群故障时积压数据恢复时间需等待集群恢复的时间

想减少数据恢复时间，该怎么办？

六、Pulsar多集群灾备方案二

减少数据恢复时间

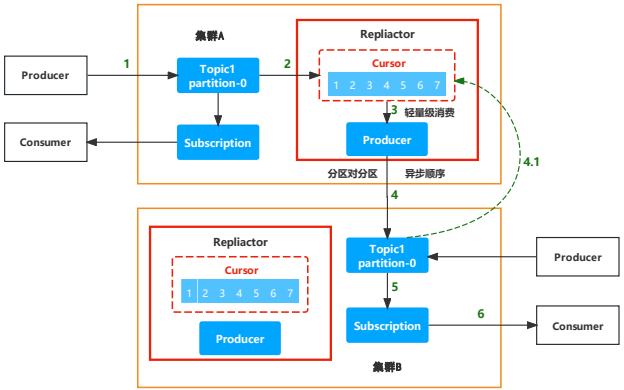


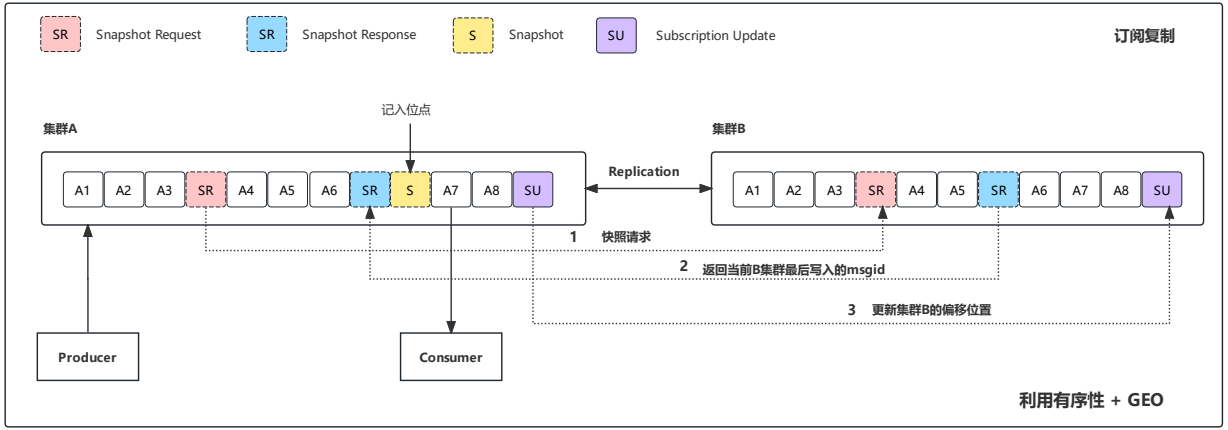
关键问题

- 消息数据灾备集群之间双向复制。
- 订阅偏移量主备同步。
- 灾备切换控制粒度。

GEO + Cluster-Level Failover + Discovery

六、Pulsar多集群灾备方案二





数据恢复时间，取决于GEO复制效率

消息复制-GEO

- Broker原生：减少数据拷贝
- Pub/Sub一体：异步，实时推送
- 可靠数据复制： 消息不丢
- 灵活复制： 主从、汇聚、全联通。

订阅偏移量同步

- 充分利用Pulsar的生产有序性
- 无需为每条消息设置映射关系
- 只需通过远端集群最近位点（定时同步）和当前集群的消费位点进行对比，即可判断是否要更新远端消费位点。

利

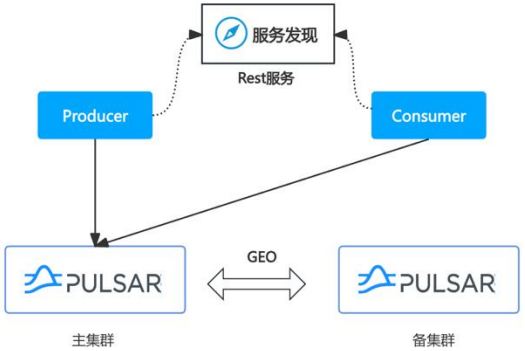
- 使用非常小的代价实现了订阅偏移量同步。

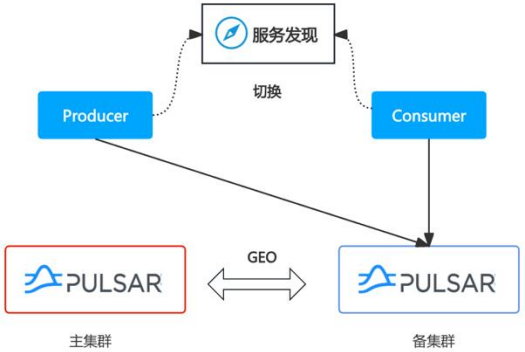
弊

- 存在少量重复消费。 **幂等 OR 事务**
- **相同订阅不能长时间同时在两个集群存活**

六、Pulsar多集群灾备方案二

灾备切换





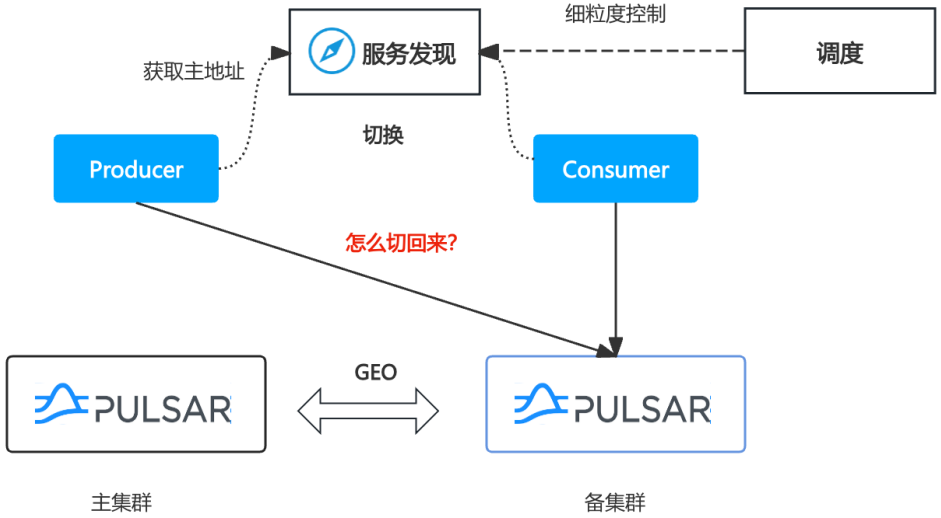

```java
ServiceUrlProvider provider =
        ControlledClusterFailover.builder()
                .defaultServiceUrl(primaryUrl)
                .checkInterval(time, TimeUnit.SECONDS)
                .urlProvider(failoverDiscoveryUrl)
                .build();

PulsarClient pulsarClient = PulsarClient.builder()
        .serviceUrl(primaryUrl)

provider.initialize(pulsarClient);

// 创建 Producer
Producer<byte[]> producer = pulsarClient.newProducer().topic("TOPIC").create();
// 创建 Consumer
Consumer<byte[]> consumer = pulsarClient.newConsumer().topic("TOPIC")
.subscriptionType(SubscriptionType.Shared).subscriptionName("SUB").subscribe();
```

自动切 OR 手动切

建议手动切，自主可控

何时切回主？

六、Pulsar多集群灾备方案二

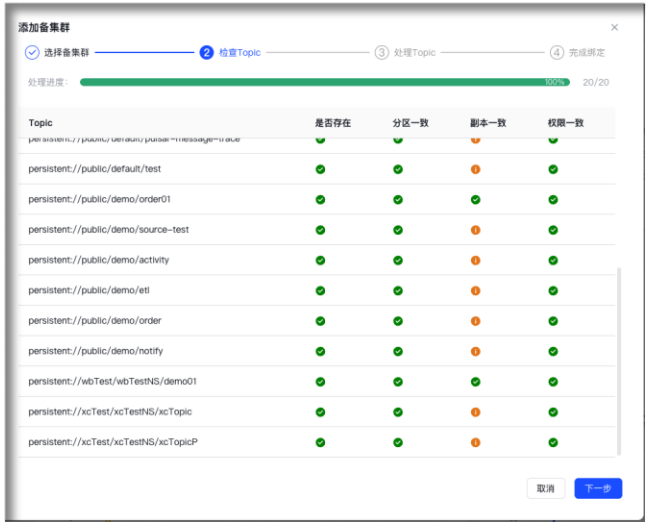


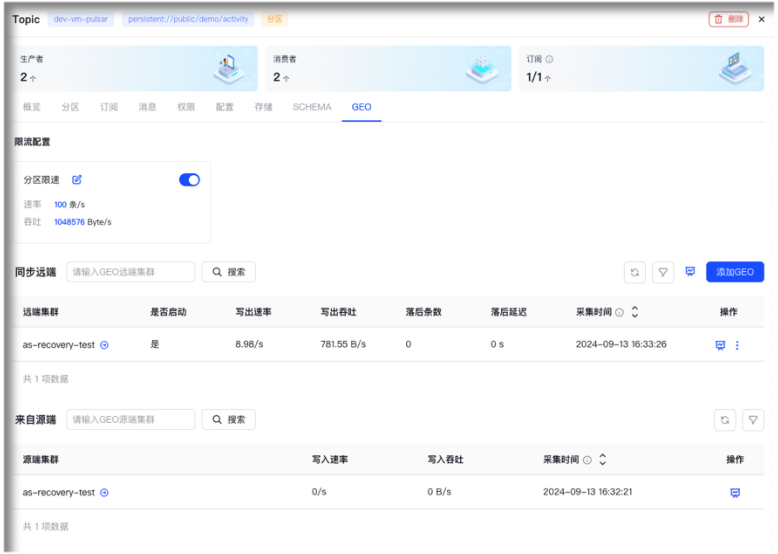
理想情况下，备与主的差距都非常小进行切换

但实际情况从Topic维度存在不统一情况，有的业务需要先切回来。

- 生产和消费的服务发现地址进行区分。
- 消费提供按订阅级别切换粒度。
- 切换时提供运行快照，帮助运维人员决策。
- 巡检、监控告警。

六、Pulsar多集群灾备方案二



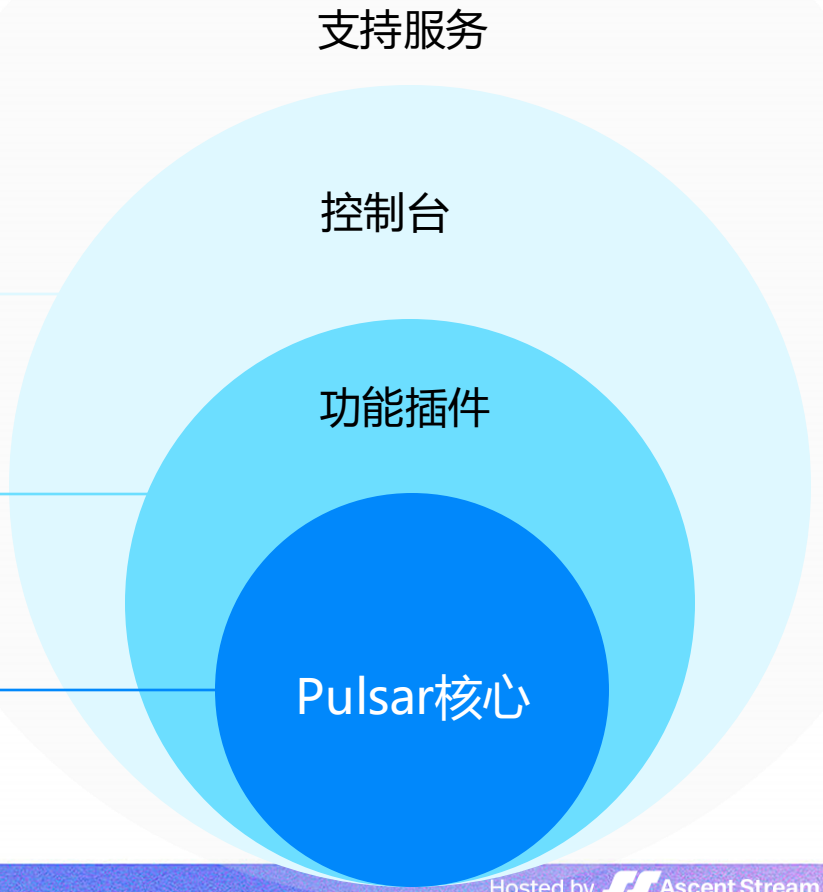


AscentStream Platform

容灾方案		抗风险级别	服务恢复时间	数据恢复时间	数据可靠性
Pulsar 单集群	单机房	机器级别	慢	慢	低
	跨机房	机房级别	快	快	中
	跨K8S	K8S集群级别	快	快	中
Pulsar 灾备集群	数据不同步	地域级别	快	慢	中
	数据同步	地域级别	较快	快	最高

商业支持

开源版	商业版
• 开源社区	• 7 * 24 专家支持 • 架构评审/上线评估/业务咨询/... • 巡检/实施/... • 解决方案
• 简版页面	• 集群管控/Function/Geo/... • 运维部署/扩缩容/... • 灵活对接企业环境
• 无	• 消息轨迹/KoP/限流/... • K8S环境支持/... • 支持灵活接入
• **核心功能保持不变** • 随开源周期发布	• **核心功能保持不变** • 灵活对接企业环境



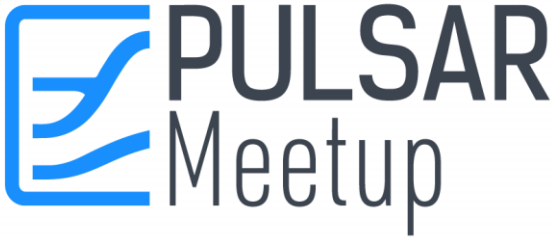
搜索：谙流科技(an)

https://ascentstream.com/

管控台后续将部分开放

中文文档社区

Hosted by AscentStream



Thanks