

25 | RocketMQ与Kafka中如何实现事务？

2019-09-21 李玥

消息队列高手课

[进入课程 >](#)



讲述：李玥

时长 15:01 大小 13.77M



你好，我是李玥。

在之前《[04 | 如何利用事务消息实现分布式事务？](#)》这节课中，我通过一个小例子来和大家讲解了如何来使用事务消息。在这节课的评论区，很多同学都提出来，非常想了解一下事务消息到底是怎么实现的。不仅要会使用，还要掌握实现原理，这种学习态度，一直是我们非常提倡的，这节课，我们就一起来学习一下，在 RocketMQ 和 Kafka 中，事务消息分别是如何来实现的？

RocketMQ 的事务是如何实现的？

首先我们来看 RocketMQ 的事务。我在之前的课程中，已经给大家讲解过 RocketMQ 事务的大致流程，这里我们再一起通过代码，重温一下这个流程。

```

1 public class CreateOrderService {
2
3     @Inject
4     private OrderDao orderDao; // 注入订单表的 DAO
5     @Inject
6     private ExecutorService executorService; // 注入一个 ExecutorService
7
8     private TransactionMQProducer producer;
9
10    // 初始化 transactionListener 和 producer
11    @Init
12    public void init() throws MQClientException {
13        TransactionListener transactionListener = createTransactionListener();
14        producer = new TransactionMQProducer("myGroup");
15        producer.setExecutorService(executorService);
16        producer.setTransactionListener(transactionListener);
17        producer.start();
18    }
19
20    // 创建订单服务的请求入口
21    @PUT
22    @RequestMapping(...)
23    public boolean createOrder(@RequestBody CreateOrderRequest request) {
24        // 根据创建订单请求创建一条消息
25        Message msg = createMessage(request);
26        // 发送事务消息
27        SendResult sendResult = producer.sendMessageInTransaction(msg, request);
28        // 返回：事务是否成功
29        return sendResult.getSendStatus() == SendStatus.SEND_OK;
30    }
31
32    private TransactionListener createTransactionListener() {
33        return new TransactionListener() {
34            @Override
35            public LocalTransactionState executeLocalTransaction(Message msg, Object arg) {
36                CreateOrderRequest request = (CreateOrderRequest ) arg;
37                try {
38                    // 执行本地事务创建订单
39                    orderDao.createOrderInDB(request);
40                    // 如果没抛异常说明执行成功，提交事务消息
41                    return LocalTransactionState.COMMIT_MESSAGE;
42                } catch (Throwable t) {
43                    // 失败则直接回滚事务消息
44                    return LocalTransactionState.ROLLBACK_MESSAGE;
45                }
46            }
47            // 反查本地事务
48            @Override
49            public LocalTransactionState checkLocalTransaction(MessageExt msg) {
50                // 从消息中获得订单 ID

```



```

3      throws MQClientException {
4      TransactionListener transactionListener = getCheckListener();
5      if (null == localTransactionExecuter && null == transactionListener) {
6          throw new MQClientException("tranExecuter is null", null);
7      }
8      Validators.checkMessage(msg, this.defaultMQProducer);
9
10     SendResult sendResult = null;
11
12     // 这里给消息添加了属性, 标明这是一个事务消息, 也就是半消息
13     MessageAccessor.putProperty(msg, MessageConst.PROPERTY_TRANSACTION_PREPARED, "true");
14     MessageAccessor.putProperty(msg, MessageConst.PROPERTY_PRODUCER_GROUP, this.defaultMQProducer.getGroup());
15
16     // 调用发送普通消息的方法, 发送这条半消息
17     try {
18         sendResult = this.send(msg);
19     } catch (Exception e) {
20         throw new MQClientException("send message Exception", e);
21     }
22
23     LocalTransactionState localTransactionState = LocalTransactionState.UNKNOW;
24     Throwable localException = null;
25     switch (sendResult.getSendStatus()) {
26         case SEND_OK: {
27             try {
28                 if (sendResult.getTransactionId() != null) {
29                     msg.putUserProperty("__transactionId__", sendResult.getTransactionId());
30                 }
31                 String transactionId = msg.getProperty(MessageConst.PROPERTY_UNIQ_CLIENT_MESSAGE);
32                 if (null != transactionId && !"".equals(transactionId)) {
33                     msg.setTransactionId(transactionId);
34                 }
35
36                 // 执行本地事务
37                 if (null != localTransactionExecuter) {
38                     localTransactionState = localTransactionExecuter.executeLocalTransactionBranch(msg);
39                 } else if (transactionListener != null) {
40                     log.debug("Used new transaction API");
41                     localTransactionState = transactionListener.executeLocalTransactionBranch(msg);
42                 }
43                 if (null == localTransactionState) {
44                     localTransactionState = LocalTransactionState.UNKNOW;
45                 }
46
47                 if (localTransactionState != LocalTransactionState.COMMIT_MESSAGE) {
48                     log.info("executeLocalTransactionBranch return {}", localTransactionState);
49                     log.info(msg.toString());
50                 }
51             } catch (Throwable e) {
52                 log.info("executeLocalTransactionBranch exception", e);
53                 log.info(msg.toString());
54                 localException = e;

```

```

55         }
56     }
57     break;
58     case FLUSH_DISK_TIMEOUT:
59     case FLUSH_SLAVE_TIMEOUT:
60     case SLAVE_NOT_AVAILABLE:
61         localTransactionState = LocalTransactionState.ROLLBACK_MESSAGE;
62         break;
63     default:
64         break;
65 }
66
67 // 根据事务消息和本地事务的执行结果 localTransactionState，决定提交或回滚事务消息
68 // 这里给 Broker 发送提交或回滚事务的 RPC 请求。
69 try {
70     this.endTransaction(sendResult, localTransactionState, localException);
71 } catch (Exception e) {
72     log.warn("local transaction execute " + localTransactionState + ", but end broke
73 }
74
75 TransactionSendResult transactionSendResult = new TransactionSendResult();
76 transactionSendResult.setSendStatus(sendResult.getSendStatus());
77 transactionSendResult.setMessageQueue(sendResult.getMessageQueue());
78 transactionSendResult.setMsgId(sendResult.getMsgId());
79 transactionSendResult.setQueueOffset(sendResult.getQueueOffset());
80 transactionSendResult.setTransactionId(sendResult.getTransactionId());
81 transactionSendResult.setLocalTransactionState(localTransactionState);
82 return transactionSendResult;
83 }


```

这段代码的实现逻辑是这样的：首先给待发送消息添加了一个属性 `PROPERTY_TRANSACTION_PREPARED`，表明这是一个事务消息，也就是半消息，然后会像发送普通消息一样去把这条消息发送到 Broker 上。如果发送成功了，就开始调用我们之前提供的接口 `TransactionListener` 的实现类中，执行本地事务的方法 `executeLocalTransaction()` 来执行本地事务，在我们的例子中就是在数据库中插入一条订单记录。

最后，根据半消息发送的结果和本地事务执行的结果，来决定提交或者回滚事务。在实现方法 `endTransaction()` 中，producer 就是给 Broker 发送了一个单向的 RPC 请求，告知 Broker 完成事务的提交或者回滚。由于有事务反查的机制来兜底，这个 RPC 请求即使失败或者丢失，也都不会影响事务最终的结果。最后构建事务消息的发送结果，并返回。

以上，就是 RocketMQ 在 Producer 这一端事务消息的实现，然后我们再看一下 Broker 这一端，它是怎么样来处理事务消息和进行事务反查的。

Broker 在处理 Producer 发送消息的请求时，会根据消息中的属性判断一下，这条消息是普通消息还是半消息：

 复制代码

```
1 // ...
2 if (traFlag != null && Boolean.parseBoolean(traFlag)) {
3     // ...
4     putMessageResult = this.brokerController.getTransactionMessageService().prepareMe
5 } else {
6     putMessageResult = this.brokerController.getMessageStore().putMessage(msgInner);
7 }
8 // ...
```

这段代码在

`org.apache.rocketmq.broker.processor.SendMessageProcessor#sendMessage` 方法中，然后我们跟进去看看真正处理半消息的业务逻辑，这段处理逻辑在类 `org.apache.rocketmq.broker.transaction.queue.TransactionalMessageBridge` 中：

 复制代码

```
1 public PutMessageResult putHalfMessage(MessageExtBrokerInner messageInner) {
2     return store.putMessage(parseHalfMessageInner(messageInner));
3 }
4
5 private MessageExtBrokerInner parseHalfMessageInner(MessageExtBrokerInner msgInner) {
6
7     // 记录消息的主题和队列，到新的属性中
8     MessageAccessor.putProperty(msgInner, MessageConst.PROPERTY_REAL_TOPIC, msgInner.get
9     MessageAccessor.putProperty(msgInner, MessageConst.PROPERTY_REAL_QUEUE_ID,
10         String.valueOf(msgInner.getQueueId()));
11     msgInner.setSysFlag(
12         MessageSysFlag.resetTransactionValue(msgInner.getSysFlag(), MessageSysFlag.TRAN
13     // 替换消息的主题和队列为：RMQ_SYS_TRANS_HALF_TOPIC, 0
14     msgInner.setTopic(TransactionalMessageUtil.buildHalfTopic());
15     msgInner.setQueueId(0);
16     msgInner.setPropertiesString(MessageDecoder.messageProperties2String(msgInner.getPro
17     return msgInner;
18 }
```

我们可以看到，在这段代码中，RocketMQ 并没有把半消息保存到消息中客户端指定的那个队列中，而是记录了原始的主题队列后，把这个半消息保存在了一个特殊的内部主题 `RMQ_SYS_TRANS_HALF_TOPIC` 中，使用的队列号固定为 0。这个主题和队列对消费者是不可见的，所以里面的消息永远不会被消费。这样，就保证了在事务提交成功之前，这个半消息对消费者来说是消费不到的。

然后我们再看一下，RocketMQ 是如何进行事务反查的：在 Broker 的 `TransactionalMessageCheckService` 服务中启动了一个定时器，定时从半消息队列中读出所有待反查的半消息，针对每个需要反查的半消息，Broker 会给对应的 Producer 发一个要求执行事务状态反查的 RPC 请求，这部分的逻辑在方法 `org.apache.rocketmq.broker.transaction.AbstractTransactionalMessageCheckListener#sendCheckMessage` 中，根据 RPC 返回响应中的反查结果，来决定这个半消息是需要提交还是回滚，或者后续继续来反查。

最后，提交或者回滚事务实现的逻辑是差不多的，首先把半消息标记为已处理，如果是提交事务，那就把半消息从半消息队列中复制到这个消息真正的主题和队列中去，如果要回滚事务，这一步什么都不需要做，最后结束这个事务。这部分逻辑的实现在 `org.apache.rocketmq.broker.processor.EndTransactionProcessor` 这个类中。

Kafka 的事务和 Exactly Once 可以解决什么问题？

接下来我们再说一下 Kafka 的事务。之前我们讲事务的时候说过，Kafka 的事务解决的问题和 RocketMQ 是不太一样的。RocketMQ 中的事务，它解决的问题是，确保执行本地事务和发消息这两个操作，要么都成功，要么都失败。并且，RocketMQ 增加了一个事务反查的机制，来尽量提高事务执行的成功率和数据一致性。

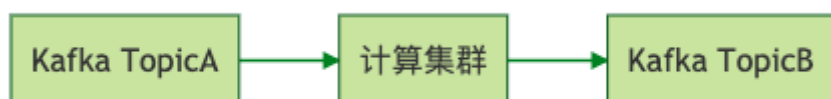
而 Kafka 中的事务，它解决的问题是，确保在一个事务中发送的多条消息，要么都成功，要么都失败。注意，这里面的多条消息不一定要在同一个主题和分区中，可以是发往多个主题和分区的消息。当然，你可以在 Kafka 的事务执行过程中，加入本地事务，来实现和 RocketMQ 中事务类似的效果，但是 Kafka 是没有事务反查机制的。

Kafka 的这种事务机制，单独来使用的场景不多。更多的情况下被用来配合 Kafka 的幂等机制来实现 Kafka 的 Exactly Once 语义。我在之前的课程中也强调过，这里面的 Exactly Once，和我们通常理解的消息队列的服务水平中的 Exactly Once 是不一样的。

我们通常理解消息队列的服务水平中的 Exactly Once，它指的是，消息从生产者发送到 Broker，然后消费者再从 Broker 拉取消息，然后进行消费。这个过程中，确保每一条消息恰好传输一次，不重不丢。我们之前说过，包括 Kafka 在内的几个常见的开源消息队列，都只能做到 At Least Once，也就是至少一次，保证消息不丢，但有可能会重复。做不到 Exactly Once。



那 Kafka 中的 Exactly Once 又是解决的什么问题呢？它解决的是，在流计算中，用 Kafka 作为数据源，并且将计算结果保存到 Kafka 这种场景下，数据从 Kafka 的某个主题中消费，在计算集群中计算，再把计算结果保存在 Kafka 的其他主题中。这样的过程中，保证每条消息都被恰好计算一次，确保计算结果正确。



举个例子，比如，我们把所有订单消息保存在一个 Kafka 的主题 Order 中，在 Flink 集群中运行一个计算任务，统计每分钟的订单收入，然后把结果保存在另一个 Kafka 的主题 Income 里面。要保证计算结果准确，就要确保，无论是 Kafka 集群还是 Flink 集群中任何节点发生故障，每条消息都只能被计算一次，不能重复计算，否则计算结果就错了。这里面有一个很重要的限制条件，就是数据必须来自 Kafka 并且计算结果都必须保存到 Kafka 中，才可以享受到 Kafka 的 Exactly Once 机制。

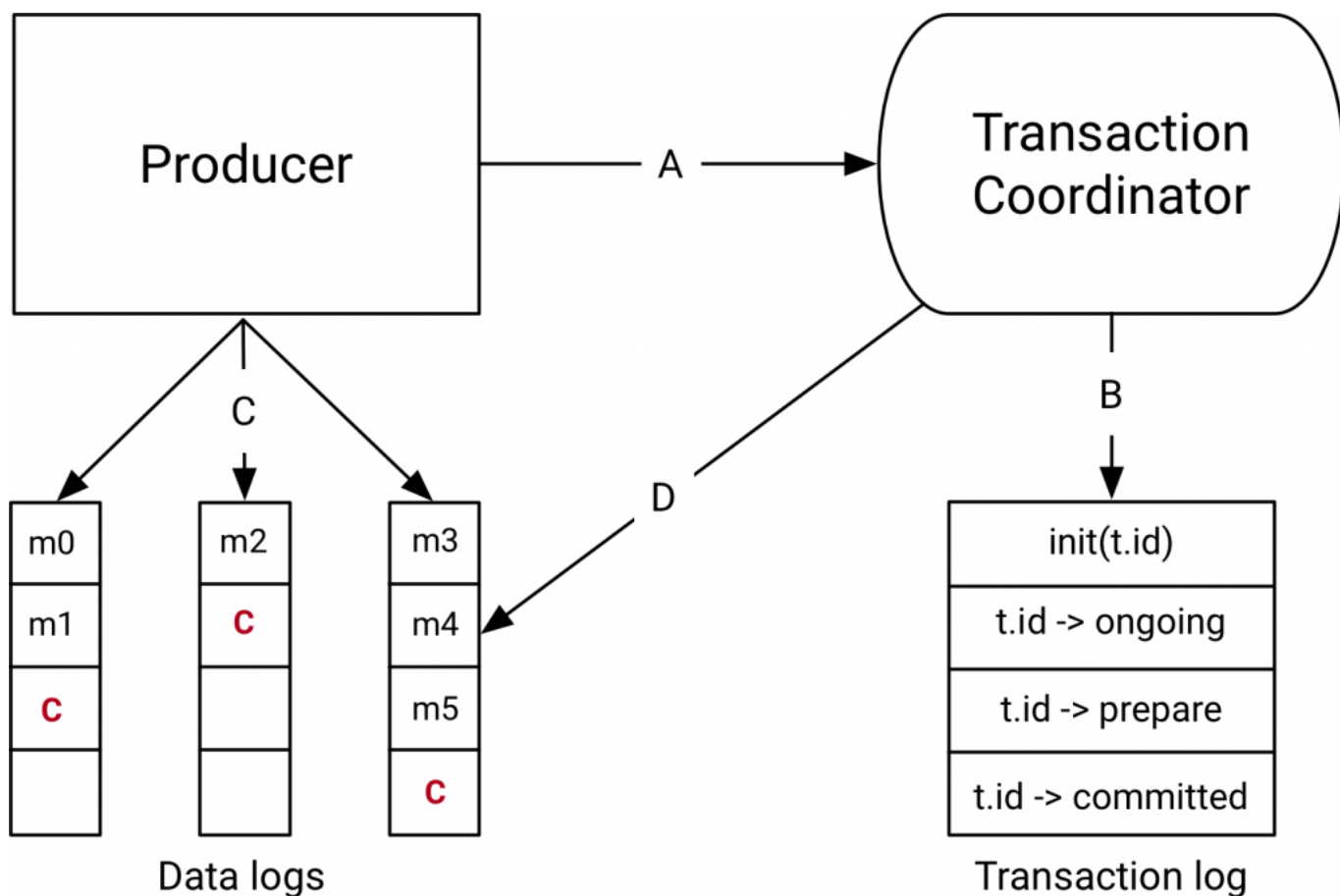
可以看到，Kafka 的 Exactly Once 机制，是为了解决在“读数据 - 计算 - 保存结果”这样的计算过程中数据不重不丢，而不是我们通常理解的使用消息队列进行消息生产消费过程中的 Exactly Once。

Kafka 的事务是如何实现的？

那 Kafka 的事务又是怎么实现的呢？它的实现原理和 RocketMQ 的事务是差不多的，都是基于两阶段提交来实现的，但是实现的过程更加复杂。

首先说一下，参与 Kafka 事务的几个角色，或者说是模块。为了解决分布式事务问题，Kafka 引入了事务协调者这个角色，负责在服务端协调整个事务。这个协调者并不是一个独立的进程，而是 Broker 进程的一部分，协调者和分区一样通过选举来保证自身的可用性。

和 RocketMQ 类似，Kafka 集群中也有一个特殊的用于记录事务日志的主题，这个事务日志主题的实现和普通的主题是一样的，里面记录的数据就是类似于“开启事务”“提交事务”这样的事务日志。日志主题同样也包含了很多的分区。在 Kafka 集群中，可以存在多个协调者，每个协调者负责管理和使用事务日志中的几个分区。这样设计，其实就是为了能并行执行多个事务，提升性能。



(图片来源: [Kafka 官方](#))

下面说一下 Kafka 事务的实现流程。

首先，当我们开启事务的时候，生产者会给协调者发一个请求来开启事务，协调者在事务日志中记录下事务 ID。

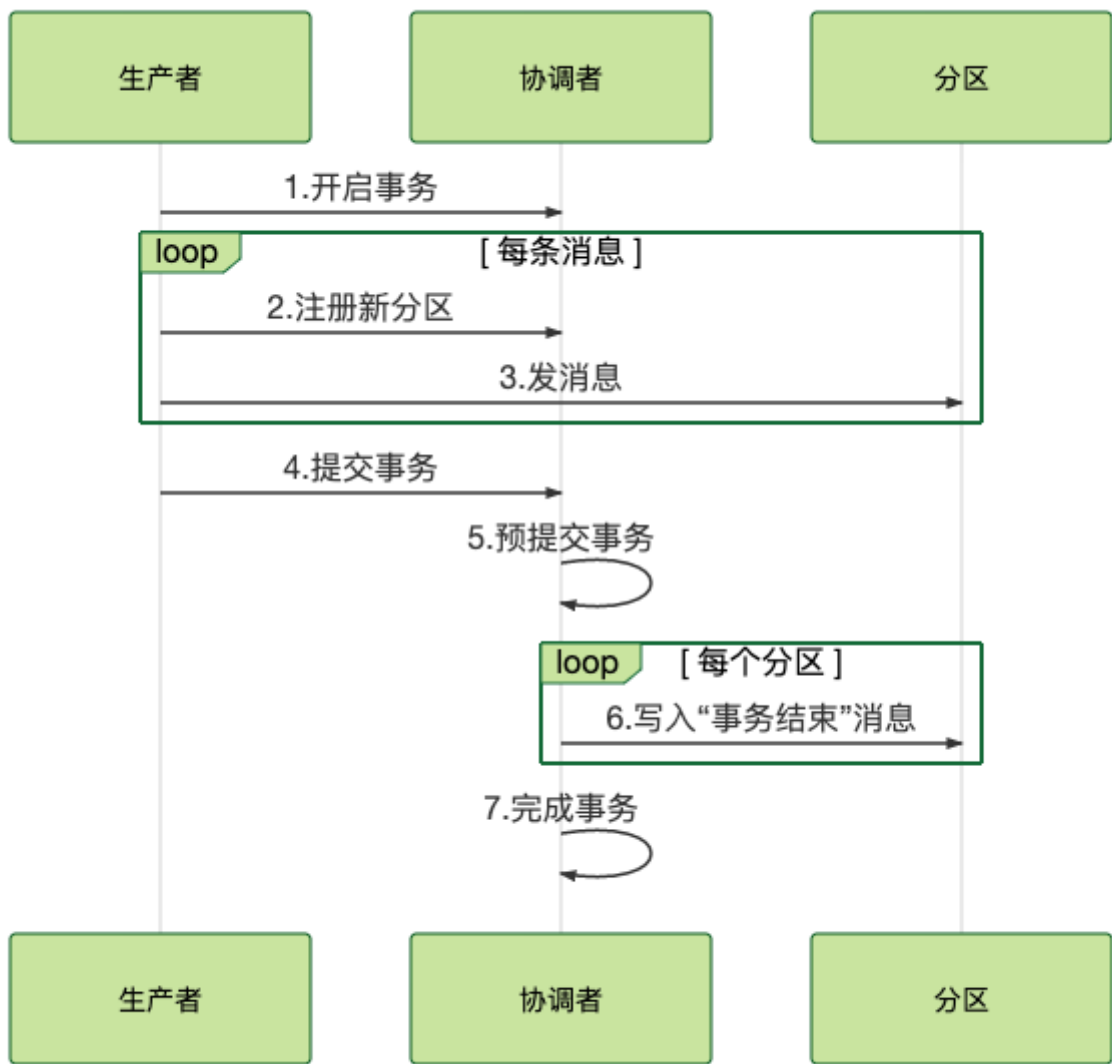
然后，生产者在发送消息之前，还要给协调者发送请求，告知发送的消息属于哪个主题和分区，这个信息也会被协调者记录在事务日志中。接下来，生产者就可以像发送普通消息一样

来发送事务消息，这里和 RocketMQ 不同的是，RocketMQ 选择把未提交的事务消息保存在特殊的队列中，而 Kafka 在处理未提交的事务消息时，和普通消息是一样的，直接发给 Broker，保存在这些消息对应的分区中，Kafka 会在客户端的消费者中，暂时过滤未提交的事务消息。

消息发送完成后，生产者给协调者发送提交或回滚事务的请求，由协调者来开始两阶段提交，完成事务。第一阶段，协调者把事务的状态设置为“预提交”，并写入事务日志。到这里，实际上事务已经成功了，无论接下来发生什么情况，事务最终都会被提交。

之后便开始第二阶段，协调者在事务相关的所有分区中，都会写一条“事务结束”的特殊消息，当 Kafka 的消费者，也就是客户端，读到这个事务结束的特殊消息之后，它就可以把之前暂时过滤的那些未提交的事务消息，放行给业务代码进行消费了。最后，协调者记录最后一条事务日志，标识这个事务已经结束了。

我把整个事务的实现流程，绘制成一个简单的时序图放在这里，便于你理解。



总结一下 Kafka 这个两阶段的流程，准备阶段，生产者发消息给协调者开启事务，然后消息发送到每个分区上。提交阶段，生产者发消息给协调者提交事务，协调者给每个分区发一条“事务结束”的消息，完成分布式事务提交。

小结

这节课我分别讲解了 Kafka 和 RocketMQ 是如何来实现事务的。你可以看到，它们在实现事务过程中的一些共同的地方，它们都是基于两阶段提交来实现的事务，都利用了特殊的主题中的队列和分区来记录事务日志。

不同之处在于对处于事务中的消息的处理方式，RocketMQ 是把这些消息暂存在一个特殊的队列中，待事务提交后再移动到业务队列中；而 Kafka 直接把消息放到对应的业务分区中，配合客户端过滤来暂时屏蔽进行中的事务消息。

同时你需要了解，RocketMQ 和 Kafka 的事务，它们的适用场景是不一样的，RocketMQ 的事务适用于解决本地事务和发消息的数据一致性问题，而 Kafka 的事务则是用于实现它的 Exactly Once 机制，应用于实时计算的场景中。

思考题

课后，请你根据我们课程中讲到的 Kafka 事务的实现流程，去 Kafka 的源代码中把这个事务的实现流程分析出来，将我们上面这个时序图进一步细化，绘制一个粒度到类和方法调用的时序图。然后请你想一下，如果事务进行过程中，协调者宕机了，那事务又是如何恢复的呢？欢迎你在评论区留言，写下你的想法。

感谢阅读，如果你觉得这篇文章对你有一些启发，也欢迎把它分享给你的朋友。

消息队列高手课

从源码角度全面解析 MQ 的设计与实现

李玥

京东零售技术架构部资深架构师



新版升级：点击「 请朋友读」，20位好友免费读，邀请订阅更有**现金**奖励。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 24 | Kafka的协调服务ZooKeeper：实现分布式系统的“瑞士军刀”

精选留言 (2)

写留言



miniluo

2019-09-21

老师，有个疑问：文中说到rocketmq#checkLocalTransaction这个方法反查到可能本地事务还在提交中就返回了unknown，那后续呢？还会通过定时轮询检查？求解，谢谢

作者回复：会一直定时轮询，直到有结果或者超时。

2

1



Imtoo

2019-09-23

kafka的第二阶段，事务协调者发送给每个分区的事务结束的消息，每个分区是怎么处理这个事务结束的消息的？这个事务结束的消息保存到哪儿了？是不是消费者挂机重启之后，事务结束的消息就没了？

