Raft 算法（2）：如何复制日志

jask

2024-08-11

Raft 日志

格式：主要包含用户指定的数据，也就是指令（Command），还包含一些附加信息，比如索引值（Log index）、任期编号（Term）。那你该怎么理解这些信息呢？

指令：一条由客户端请求指定的、状态机需要执行的指令。你可以将指令理解成客户端指定的数据。
索引值：日志项对应的整数索引值。它其实就是用来标识日志项的，是一个连续的、单调递增的整数号码。任期编号：创建这条日志项的领导者的任期编号。

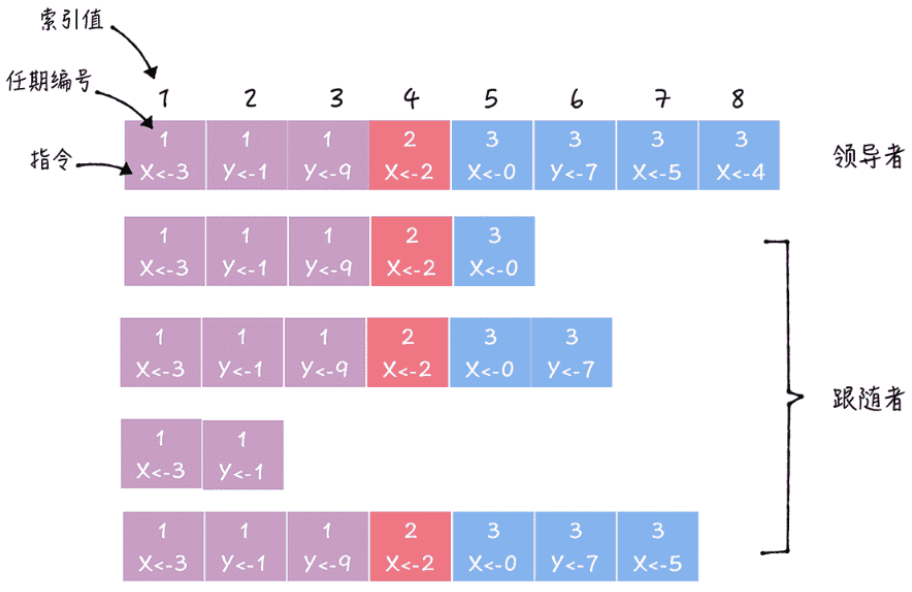

Figure 1: 日志结构

如何复制日志？

可以理解为一个优化的二阶段提交，减少了一半的往返消息，也就是降低了一半的消息延迟。

过程

首先，领导者进入第一阶段，通过日志复制（AppendEntries）RPC 消息，将日志项复制到集群其他节点上。

接着，如果领导者接收到大多数的“复制成功”响应后，它将日志项提交到它的状态机，并返回成功给客户端。如果领导者没有接收到大多数的“复制成功”响应，那么就返回错误给客户端。

问题：领导和将日志项提交到状态机，怎么没通知跟随者提交日志项呢？

这是 Raft 中的一个优化，领导者不直接发送消息通知其他节点提交指定日志项。因为领导者的日志复制 RPC 消息或心跳消息，包含了当前最大的，将会被提交的日志项索引值。所以通过日志复制 RPC 消息或心跳消息，跟随者就可以知道领导者的日志提交位置信息。

因此，当其他节点接受领导者的心跳消息，或者新的日志复制 RPC 消息后，就会将这条日志项提交到它的状态机。而这个优化，降低了处理客户端请求的延迟，将二阶段提交优化为了一段提交，降低了一半的消息延迟。

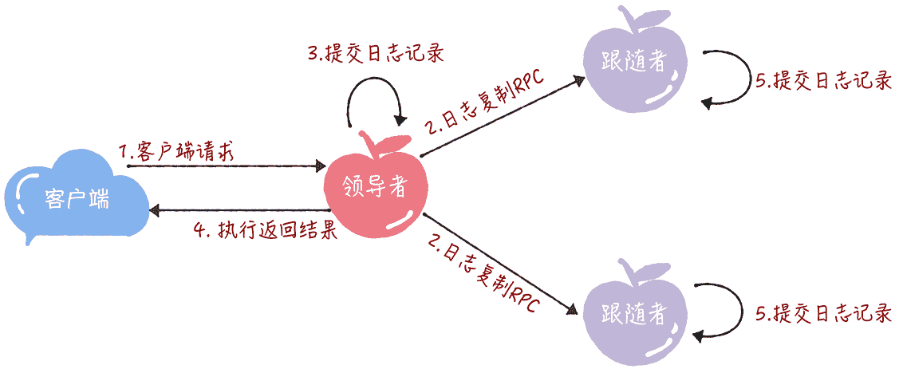

Figure 2: 流程

1. 接收到客户端请求后，领导者基于客户端请求中的指令，创建一个新日志项，并附加到本地日志中。

2. 领导者通过日志复制 RPC，将新的日志项复制到其他的服务器。

3. 当领导者将日志项，成功复制到大多数的服务器上的时候，领导者会将这条日志项提交到它的状态机中。

4. 领导者将执行的结果返回给客户端。

5. 当跟随者接收到心跳信息，或者新的日志复制 RPC 消息后，如果跟随者发现领导者已经提交了某条日志项，而它还没提交，那么跟随者就将这条日志项提交到本地的状态机中。

如何保证日志的一致？

在 Raft 算法中，领导者通过强制跟随者直接复制自己的日志项，处理不一致日志。也就是说，Raft 是通过以领导者的日志为准，来实现各节点日志的一致的。

首先，领导者通过日志复制 RPC 的一致性检查，找到跟随者节点上，与自己相同日志项的最大索引值。也就是说，这个索引值之前的日志，领导者和跟随者是一致的，之后的日志是不一致的了。

然后，领导者强制跟随者更新覆盖的不一致日志项，实现日志的一致。

领导者通过日志复制 RPC 一致性检查，找到跟随者节点上与自己相同日志项的最大索引值，然后复制并更新覆盖该索引值之后的日志项，实现了各节点日志的一致。需要你注意的是，跟随者中的不一致日志项会被领导者的日志覆盖，而且领导者从来不会覆盖或者删除自己的日志。

总结

在 Raft 中，副本数据是以日志的形式存在的，其中日志项中的指令表示用户指定的数据。

兰伯特的 Multi-Paxos 不要求日志是连续的，但在 Raft 中日志必须是连续的。而且在 Raft 中，日志不仅是数据的载体，日志的完整性还影响领导者选举的结果。也就是说，日志完整性最高的节点才能当选领导者。

Raft 是通过以领导者的日志为准，来实现日志的一致的。