27-从MOVED、ASK看集群节点如何处理命令？

你好，我是蒋德钧。

在上节课一开始我给你介绍了，我们在Redis Cluster这个模块中会学习三部分内容：节点间如何传递信息和运行状态、节点如何处理命令，以及数据如何在节点间迁移。那么通过上节课的学习，现在我们已经了解了Gossip协议的基本实现，也就是支持集群节点间信息和运行状态传递的数据结构、关键函数设计与实现。

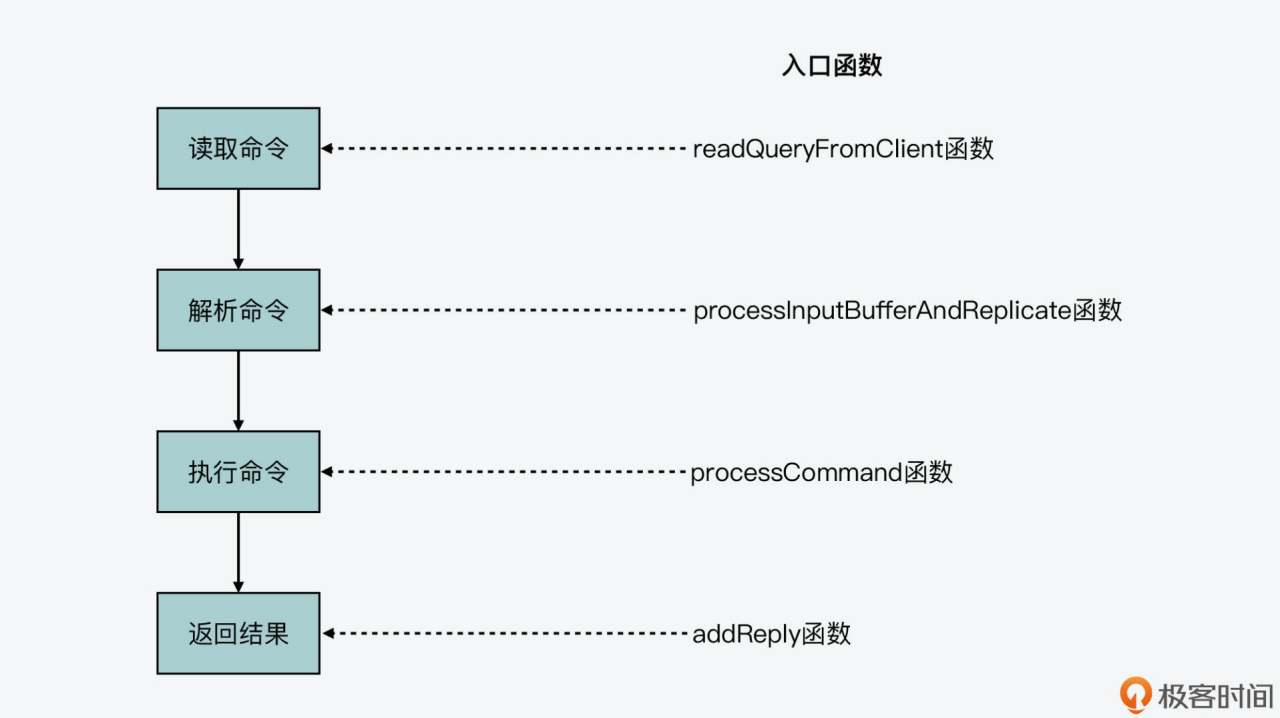
所以在今天这节课，我们就来了解下集群命令处理的实现。这部分内容不仅包括了集群节点处理一个命令的基本流程，更重要的是，我们可以掌握集群特定命令MOVED、ASK是如何实现的。这两个命令对应了Redis Cluster中请求重定向的处理场景，了解了这部分内容之后，我们就可以参考Redis Cluster，来设计和实现分布式系统中的请求重定向。

接下来，我们先来看下集群节点处理一个命令的基本流程，这可以让我们对集群节点的实现有个整体观。

集群节点处理命令的基本流程

我在第14讲中提到过，Redis server处理一条命令的过程可以分成四个阶段，分别是**命令读取、命令解析、命令执行和结果返回**。而和单个Redis server一样，Redis Cluster中的节点，也是按照相同的阶段来处理命令的。

因此，集群节点在各阶段处理命令的入口函数和单个Redis server也是一样的，如下图所示。你也可以再去回顾下第14讲中，我介绍的命令处理详细流程。



但是，在其中的命令执行阶段，如果Redis server是一个集群节点，那么在命令执行的过程中，就会增加额外的处理流程，而这个流程正对应了Redis Cluster中可能遇到的请求重定向问题。

这里所说的**请求重定向**，是指客户端给一个集群节点发送命令后，节点发现客户端请求的数据并不在本地。因此，节点需要让客户端的请求，重新定向发给实际拥有数据的节点，这样客户端的命令才能正常执行。

而你需要注意，请求重定向其实是分布式系统设计过程中需要面对的一个常见问题。尤其对于像Redis

Cluster这样，没有使用中心化的第三方系统来维护数据分布的分布式系统来说，**当集群由于负载均衡或是节点故障而导致数据迁移时，请求重定向是不可避免的**。所以，了解这个设计环节，对于你开发分布式系统同样具有重要的参考价值。

那么，下面我们就先来看下在命令执行阶段中，针对集群节点增加的处理流程，这是在processCommand函数（在server.c文件）中实现的。

processCommand函数在执行过程中，会判断当前节点是否处于集群模式，这是通过全局变量server的**cluster_enable标记**来判断的。如果当前节点处于集群模式，processCommand函数会判断是否需要执行重定向。

当然，如果当前节点收到的命令来自于它在集群中的主节点，或者它收到的命令并没有带key参数，那么在这些情况下，集群节点并不会涉及重定向请求的操作。不过，这里有一个不带key参数的命令是一个例外，就是**EXEC命令**。如果当前节点收到EXEC命令，processCommand函数仍然会判断是否要进行请求重定向。

那么，processCommand函数具体是如何判断是否要执行请求重定向的呢？

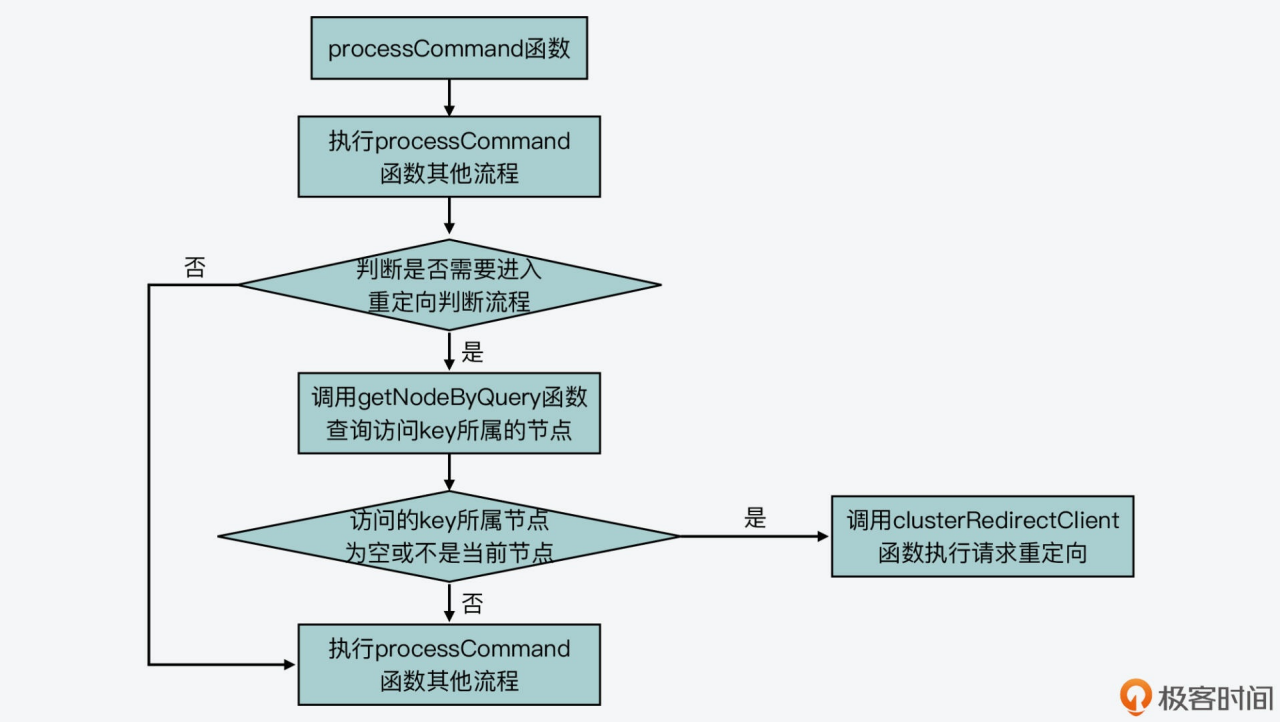
其实，它是调用了**getNodeByQuery函数**（在cluster.c文件中），来查询当前收到的命令能在哪个集群节点上进行处理。如果getNodeByQuery函数返回的结果是空，或者查询到的集群节点不是当前节点，那么，processCommand函数就会调用clusterRedirectClient函数（在cluster.c文件中），来实际执行请求重定向。

下面的代码展示了集群节点处理命令过程中针对请求重定向增加的流程，你可以看下。

```
int processCommand(client *c) {
   …
   //当前Redis server启用了Redis Cluster模式；收到的命令不是来自于当前借的主节点；收到的命令包含了key参数，或者命令是EXEC
   if (server.cluster_enabled && !(c->flags & CLIENT_MASTER)
 && !(c->flags & CLIENT_LUA && server.lua_caller->flags & CLIENT_MASTER)
 && !(c->cmd->getkeys_proc == NULL && c->cmd->firstkey == 0 &&
          c->cmd->proc != execCommand))
 {
    …
    clusterNode *n = getNodeByQuery(c,c->cmd,c->argv,c->argc, &hashslot,&error_code); //查询当前命令可以被哪个集
    if (n == NULL || n != server.cluster->myself) {
       …
       clusterRedirectClient(c,n,hashslot,error_code); //实际执行请求重定向
             return C_OK;
         }
     }
```

当然，如果不需要执行请求重定向，那么processCommand函数会继续执行后续的流程，并调用call函数实际运行命令。

下图展示了processCommand函数针对集群节点增加的基本执行逻辑，你可以再回顾下。



好，接下来，我们就来看下getNodeByQuery函数是如何查询能处理一条命令的集群节点的。

如何查询能运行命令的集群节点？

首先，我们来看下getNodeByQuery函数的原型，如下所示：

```
clusterNode *getNodeByQuery(client *c, struct redisCommand *cmd, robj **argv, int argc, int *hashslot, int
```

它的函数参数包括了节点收到的命令及参数。同时，它的参数中还包括了两个指针：hashslot和error_code，这两个指针分别表示命令访问的key所在的slot（哈希槽），以及函数执行后的错误代码。此外，getNodeByQuery函数的返回值是clusterNode类型，表示的是能处理命令的集群节点。

然后，我们来看下getNodeByQuery函数的具体执行过程，这个过程基本可以分成三个步骤来完成。

第一步，使用multiState结构体封装收到的命令

因为集群节点可能收到**MULTI命令**，而MULTI命令表示紧接着它的多条命令是需要作为一个事务来执行的。当Redis server收到客户端发送的MULTI命令后，它会调用MULTI命令的处理函数multiCommand（在multi.c文件中），在表示客户端的结构体变量client中设置**CLIENT_MULTI标记**，如下所示：

```
void multiCommand(client *c) {
   …
   c->flags |= CLIENT_MULTI; //在客户端的标记中设置CLIENT_MULTI
   addReply(c,shared.ok);
}
```

而在刚才介绍的命令执行函数processCommand中，它在处理命令时，会判断客户端变量client中是否有CLIENT_MULTI标记。如果有的话，processCommand会调用**queueMultiCommand函数**，把后续收到的命令缓存在client结构体的mstate成员变量中。mstate成员变量的类型是**multiState结构体**，它记录了MULTI命令后的其他命令以及命令个数。

下面的代码展示了processCommand函数对CLIENT_MULTI标记的处理，你可以看下。你也可以进一步阅读queueMultiCommand函数（在multi.c文件中）和client结构体（在server.h文件中），详细了解MULTI后续命令的记录过程。

```
int processCommand(client *c) {
…
//客户端有CLIENT_MULTI标记，同时当前命令不是EXEC，DISCARD，MULTI和WATCH
if (c->flags & CLIENT_MULTI &&
  c->cmd->proc != execCommand && c->cmd->proc != discardCommand &&
  c->cmd->proc != multiCommand && c->cmd->proc != watchCommand)
  {
        queueMultiCommand(c); //缓存命令
        …
  }
```

其实，刚才介绍的Redis server处理MULTI命令和缓存后续命令的流程，**对于集群节点来说，也是同样适用的**。也就是对于getNodeByQuery函数来说，它在查询命令访问的key时，就需要考虑MULTI命令的情况。

那么，为了使用同样的数据结构，来处理MULTI命令的后续命令和常规的单条命令，getNodeByQuery函数就使用了multiState结构体，来封装当前要查询的命令，如下所示：

```
multiState *ms, _ms; //使用multiState结构体封装要查询的命令
…
if (cmd->proc == execCommand) { //如果收到EXEC命令，那么就要检查MULTI后续命令访问的key情况，所以从客户端变量c中获取mst
   …
   ms = &c->mstate;
} else {
   ms = &_ms;  //如果是其他命令，那么也使用multiState结构体封装命令
   _ms.commands = &mc;
   _ms.count = 1;  //封装的命令个数为1
   mc.argv = argv; //命令的参数
   mc.argc = argc; //命令的参数个数
   mc.cmd = cmd; //命令本身
}
```

这里你需要**注意**，MULTI命令后缓存的其他命令并不会立即执行，而是需要等到EXEC命令执行时才会执行。所以，在刚才的代码中，getNodeByQuery函数也是在收到EXEC命令时，才会从客户端变量c中获取缓存的命令mstate。

好了，到这里，你就可以看到，getNodeByQuery函数使用multiState结构体，封装了当前的命令。而接下来，它就会检查命令访问的key了。

第二步，针对收到的每个命令，逐一检查这些命令访问的key所在的slots

getNodeByQuery函数会根据multiState结构中记录的命令条数，执行一个循环，逐一检查每条命令访问的key。具体来说，它会调用**getKeysFromCommand函数**（在db.c文件中）获取命令中的key位置和key个数。

然后，它会针对每个key，调用**keyHashSlot函数**（在cluster.c文件中）查询这个key所在的slot，并在全局变量server的cluster成员变量中，查找这个slot所属的集群节点，如下所示：

```
for (i = 0; i < ms->count; i++) {
   …
   //获取命令中的key位置和key个数
   keyindex = getKeysFromCommand(mcmd,margv,margc,&numkeys);
   //针对每个key执行
   for (j = 0; j < numkeys; j++) {
 …
 int thisslot = keyHashSlot((char*)thiskey->ptr, //获取key所属的slot                                    sc
 if (firstkey == NULL) {
    …
    slot = thisslot;
    n = server.cluster->slots[slot]; //查找key所属的slot对应的集群节点
 }
 …
    }
}
```

紧接着，getNodeByQuery函数会根据查找的集群节点结果进行判断，主要有以下三种情况。

- 情况一：查找的集群节点为空，此时它会报错，将error_code设置为CLUSTER_REDIR_DOWN_UNBOUND。

```
if (n == NULL) {
   …
    if (error_code)
       *error_code = CLUSTER_REDIR_DOWN_UNBOUND;
    return NULL;
}
```

- 情况二：查找的集群节点就是当前节点，而key所属的slot正在**做数据迁出操作**，此时，getNodeByQuery函数会设置变量migrating_slot为1，表示正在做数据迁出。

- 情况三：key所属的slot正在**做数据迁入操作**，此时，getNodeByQuery函数会设置变量importing_slot为1，表示正在做数据迁入。

情况二和三的代码逻辑如下所示：

```
//如果key所属的slot正在迁出，则设置migrating_slot为1
if (n == myself && server.cluster->migrating_slots_to[slot] != NULL)
{
   migrating_slot = 1;
} //如果key所属的slot正在迁入，则设置importing_slot为1
else if (server.cluster->importing_slots_from[slot] != NULL) {
   importing_slot = 1;
}
```

这里，你需要注意的是，如果命令包含的key不止1个，而且这些keys不在同一个slot，那么getNodeByQuery函数也会报错，并把error_code设置为CLUSTER_REDIR_CROSS_SLOT。

到这里，getNodeByQuery函数就查找到了命令访问的key所在的slot，以及对应的集群节点。而此时，如果节点正在做数据迁出或迁入，那么，getNodeByQuery函数就会调用**lookupKeyRead函数**（在db.c文件中），检查命令访问的key是否在当前节点的数据库中。如果没有的话，它会用一个变量**missing_keys**，记录缺失的key数量，如下所示：

```
//如果key所属slot正在迁出或迁入，并且当前访问的key不在本地数据库，那么增加missing_keys的大小
if ((migrating_slot || importing_slot) && lookupKeyRead(&server.db[0],thiskey) == NULL)
{
    missing_keys++;
}
```

接下来，getNodeByQuery函数就会根据slot的检查情况来返回相应的结果了。

第三步，根据slot的检查结果返回hashslot、error_code和相应的集群节点

在getNodeByQuery函数的返回结果中，我们可以重点关注以下四种情况。

情况一：命令访问key所属的slot没有对应的集群节点，此时，getNodeByQuery函数会返回当前节点。在这种情况下，有可能是集群有故障导致无法查找到slot所对应的节点，而error_code中会有相应的报错信息。

```
if (n == NULL) return myself;
```

情况二：命令访问key所属的slot正在做数据迁出或迁入，而且当前命令就是用来执行数据迁移的MIGRATE命令，那么，getNodeByQuery函数会返回当前节点，如下所示：

```
if ((migrating_slot || importing_slot) && cmd->proc == migrateCommand)
   return myself;
```

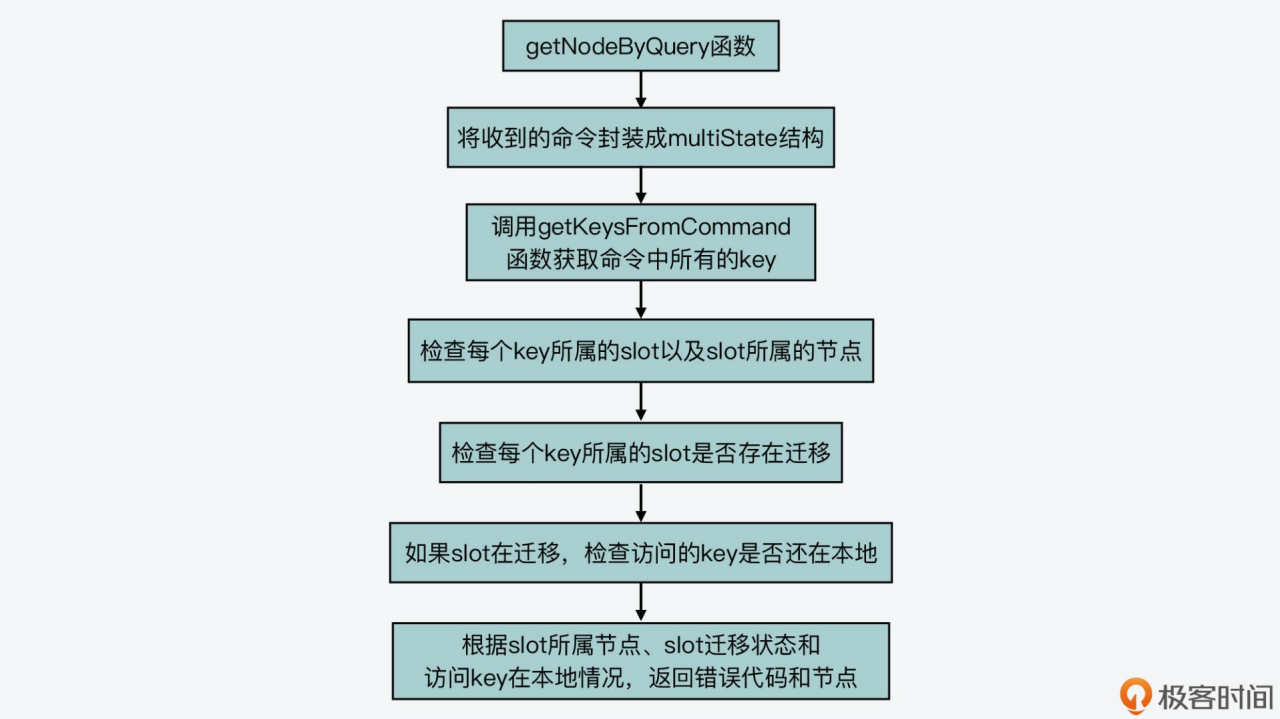
情况三：命令访问key所属的slot正在做数据迁出，并且命令访问的key在当前节点数据库中缺失了，也就是刚才介绍的missing_keys大于0。此时，getNodeByQuery函数会把error_code设置为CLUSTER_REDIR_ASK，并返回数据迁出的目标节点。

```
if (migrating_slot && missing_keys) {
       if (error_code) *error_code = CLUSTER_REDIR_ASK;
       return server.cluster->migrating_slots_to[slot];
}
```

情况四：命令访问key所属的slot对应的节点不是当前节点，而是其他节点，此时，getNodeByQuery函数会把error_code设置为CLUSTER_REDIR_MOVED，并返回key所属slot对应的实际节点。

```
if (n != myself && error_code) *error_code = CLUSTER_REDIR_MOVED;
    return n;
```

好了，到这里，我们就了解了getNodeByQuery函数对命令访问key的查询过程了。我画了张图，展示了getNodeByQuery函数基本执行过程，你可以再回顾下。



那么，有了key所属节点的查询结果后，processCommand函数接下来又会如何进行请求重定向呢？

实际上，这一步是通过执行请求重定向的函数**clusterRedirectClient**来完成的。

请求重定向函数clusterRedirectClient的执行

当getNodeByQuery函数查到的集群节点为空或者不是当前节点时，clusterRedirectClient函数就会被调用。

而clusterRedirectClient函数的逻辑比较简单，它就是**根据getNodeByQuery函数返回的error_code的不同值，执行相应的代码分支**，主要是把key所属slot对应集群节点的情况返回给客户端，从而让客户端根据返回的信息作出相应处理。比如：

- 当error_code被设置成**CLUSTER_REDIR_CROSS_SLOT**时，clusterRedirectClient函数就返回给客户端“key不在同一个slot中”的报错信息；
- 当error_code被设置成**CLUSTER_REDIR_MOVED**时，clusterRedirectClient函数会返回MOVED命令，并把key所属的slot、slot实际所属的节点IP和端口号，返回给客户端；
- 当error_code被设置成**CLUSTER_REDIR_ASK**时，clusterRedirectClient函数会返回ASK命令，并把key所属的slot、slot正在迁往的目标节点IP和端口号，返回给客户端。

下面的代码展示了刚才介绍的clusterRedirectClient函数对三种error_code的处理，你可以看下。

```
void clusterRedirectClient(client *c, clusterNode *n, int hashslot, int error_code) {
if (error_code == CLUSTER_REDIR_CROSS_SLOT) {
        addReplySds(c,sdsnew("-CROSSSLOT Keys in request don't hash to the same slot\r\n"));
}
…
else if (error_code == CLUSTER_REDIR_MOVED || error_code == CLUSTER_REDIR_ASK)
    {
        addReplySds(c,sdscatprintf(sdsempty(),
            "-%s %d %s:%d\r\n",
            (error_code == CLUSTER_REDIR_ASK) ? "ASK" : "MOVED",
            hashslot,n->ip,n->port));
 }
 …
}
```

这样，集群节点处理收到的命令的过程就结束了。

最后，我还想提醒你注意一点，就是Redis Cluster的客户端和针对单个Redis server的客户端，在实现上是有差别的。Redis Cluster客户端需要能处理节点返回的报错信息，比如说，如果集群节点返回MOVED命令，客户端就需要根据这个命令，以及其中包含的实际节点IP和端口号，来访问实际有数据的节点。

小结

今天这节课，我给你介绍了集群节点对客户端命令的处理过程。和单个Redis server处理命令的过程相似，集群节点也会经历命令读取、解析、执行和返回结果四个阶段，并且集群节点也使用了和单Redis server相同的入口处理函数。

不过你要知道的是，Redis Cluster会因为负载均衡或节点故障等原因而执行数据迁移，而这就会导致客户端访问的key并不在接收到命令的集群节点上。因此，集群节点在命令执行函数processCommand中，针对集群模式，就增加了额外的处理逻辑。这主要是包括调用**getNodeByQuery函数**查询访问的key实际所属的节点，以及根据查询结果调用**clusterRedirectClient函数**执行请求重定向。

事实上，对于分布式集群来说，Redis Cluster设计实现的请求重定向机制是一个不错的参考示例。其中，MOVED和ASK两种重定向情况，就充分考虑了数据正在迁移的场景，这种设计值得我们学习。而且，

getNodeByQuery函数在查询key所属的slot和节点时，也充分考虑了Redis的事务操作，在对命令访问key进行查询时，巧妙地使用了**同一个数据结构multiState**，来封装事务涉及的多条命令和常规的单条命令，增加了代码的复用程度，这一点也非常值得学习。

当然，在这节课里我们也多次提到了数据迁移，那么在下节课，我就会给你介绍Redis Cluster中数据迁移的具体实现。

每课一问

processCommand函数在调用完getNodeByQuery函数后，实际调用clusterRedirectClient函数进行请求重定向前，会根据当前命令是否是EXEC，分别调用discardTransaction和flagTransaction两个函数。那么，你能通过阅读源码，知道这里调用discardTransaction和flagTransaction的目的是什么吗?

```
int processCommand(client *c) {
…
clusterNode *n = getNodeByQuery(c,c->cmd,c->argv,c->argc,
                                        &hashslot,&error_code);
if (n == NULL || n != server.cluster->myself) {
   if (c->cmd->proc == execCommand) {
      discardTransaction(c);
   } else {
      flagTransaction (c);
   }
   clusterRedirectClient(c,n,hashslot,error_code);
   return C_OK;
 }
 …
 }
```

精选留言：

- Kaito 2021-10-13 01:14:36

  1、cluster 模式的 Redis，在执行命令阶段，需要判断 key 是否属于本实例，不属于会给客户端返回请求重定向的信息

  2、判断 key 是否属于本实例，会先计算 key 所属的 slot，再根据 slot 定位属于哪个实例

  3、找不到 key 所属的实例，或者操作的多个 key 不在同一个 slot，则会给客户端返回错误；key 正在做数据迁出，并且访问的这个 key 不在本实例中，会给客户端返回 ASK，让客户端去目标节点再次查询一次（临时重定向）；key 所属的 slot 不是本实例，而是其它节点，会给客户端返回 MOVED，告知客户端 key 不在本实例，以后都去目标节点查询（永久重定向）

  课后题：processCommand 函数在调用完 getNodeByQuery 函数后，实际调用 clusterRedirectClient 函数进行请求重定向前，会根据当前命令是否是 EXEC，分别调用 discardTransaction 和 flagTransaction 两个函数。这 2 个函数的目的是什么?

  看代码逻辑，只有当 n == NULL || n != server.cluster->myself 时，才会调用这 2 个方法。

  其中，如果当前执行的是 EXEC 命令，则调用 discardTransaction。这个函数表示放弃整个事务，它会清空这个 client 之前缓存的命令队列，放弃事务中 watch 的 key，重置 client 的事务标记。

如果当前命令不是 EXEC，而是一个普通命令，则调用 flagTransaction。这个函数会给当前 client 打上一个标记 CLIENT_DIRTY_EXEC，如果后面执行了 EXEC，就会判断这个标记，随即也会放弃执行事务，给客户端返回错误。

也就是说，当集群不可用、key 找不到对应的 slot、key 不在本实例中、操作的 keys 不在同一个 slot、key 正在迁移中，发生这几种情况时，都会放弃整个事务的执行。 [1赞]

- 曾轼麟 2021-10-12 14:16:58

回答老师的问题：
按照我个人理解，不知道是否准确。 我们先了解一下Redis事务的实现方式，命令在multiState中是以队列的形式保存着的，只有当执行EXEC的时候，才会按照队列顺序依次执行里面的命令，否则会调用queueMultiCommand将命令保存到这个队列中，而事务在Redis中是以client的维度开启的，如果一个client开启了事务，那么它结构体中的flags会被设置为CLIENT_MULTI（在事务中），那么问题中的两个函数的作用是什么？

1、【discardTransaction】：直接丢弃当前的事务，清空multiState队列里面的命令，并且会对事务中的key unWatch。
2、【flagTransaction】：将client的flages设置为CLIENT_DIRTY_EXEC（事务最终将在EXEC的时候也会失败）。

两个方法刚好对应了client在事务中，执行EXEC命令和普通命令的两种情况。Redis是发现当getNodeByQuery返回的clusterNode节点不是自己的时候才会执行这两个方法，并且当Redis以集群模式运行的时候，跨节点是不支持事务，如果发现当前client有事务开启的情况，可能是之前开启的，那么当getNodeByQuery发现不是自己的时候需要把之前的事务废弃。如果命令直接就是EXEC了那么直接调用discardTransaction丢弃事务，如果是事务中的某个命令出现这种情况(例如：开启事务后发生迁移)，则调用flagTransaction，等到EXEC的时候一样丢弃。

补充：
集群中涉及MULTI/EXEC的操作需要让key都在同一节点上面，如果不在会返回 MOVED 信息或者直接返回error信息。

- 可怜大灰狼 2021-10-12 11:48:34

只要能够进入n == NULL || n != server.cluster->myself，都表示需要重定向客户端了。如果当前是execCommand，discardTransaction就释放整个multi阶段缓存下来的命令。否则就打一个脏标识CLIENT_DIRTY_EXEC