1. 集群
2. 概述

Redis集群是Redis提供的分布式数据库方案，集群通过分片来进行数据共享，并提供复制和故障转移功能。

1. 节点
2. 连接

一个Redis集群通常由多个节点组成，这些节点需要连接起来，才能够构成一个包含多个节点的集群。连接各个节点的工作通过CLUSTER MEET命令完成，格式如下：

cluster meet <ip> <port>

向一个节点node发送cluster meet命令，可以让node节点与ip和port指定的节点进行握手，当握手成功时，node节点就会将ip和port所指定的节点添加到node节点当前所在的集群中。

1. 启动节点

一个节点就是一个运行在集群模式下的Redis服务器 ，在启动时会根据cluster-enabled配置项来决定是否开启服务器的集群模式。节点会继续使用在单机模式下使用的服务器组件，也会使用新的结构保存在集群模式下使用的数据。

1. 集群数据结构

clusterNode结构保存了一个节点当前的状态，如节点的创建时间，名字，配置纪元等。每个节点都会为自己和集群中的其他节点都创建一个相应的clusterNode结构。同时，每个节点还保存了一个clusterState结构，记录了当前节点的视角下，集群目前所处的状态，如集群是否下线，节点数量等等。

1. 连接的过程

其实就是一个三次握手的过程。当向节点A发送cluster meet命令时，A节点就要将另一个节点B加入到A所在的集群中。

收到命令的节点A会与B进行握手，以此来确认彼此的存在，然后A会为B创建一个clusterNode结构，添加到自己的clusterState.nodes字典中。之后，节点A向B发送一条meet消息；

B收到该消息，为A创建一个clusterNode结构，同样是放到自己的clusterState.nodes字典中，然后向A返回一条PONG消息；

节点A收到PONG消息，直到了B已经收到了自己发送的meet消息，然后再向B发送一条PING消息；

B收到PING消息，直到A已经收到了自己返回的PONG消息，握手完成。

之后，A会将B的信息通过Gossip写协议传播给集群中的其他节点，让其他节点与B握手，经过一段时间，B会被集群中的所有节点认识。

1. 槽的指派
2. 概述

Redis集群通过分片的方式来保存数据库中的键值对，集群的整个数据库被分为了16384个槽，每个键都属于这个些槽中的一个，集群中的每个节点可以处理0或最多16384个槽。当所有的槽都有节点在处理的时候，集群处于上线状态，如果有任何一个槽没有被处理，集群就会处于下线状态。

1. 槽的指派

通过向节点发送cluster addslots命令，可以将一个或多个槽指派给节点负责；

cluster addslots <…>

1. 记录节点的槽指派信息

每个节点的clusterNode结构的slots属性和numslot属性记录了节点负责处理哪些槽以及槽的数量。slots属性是一个二进制位数组，这个数组的长度是2048个字节，共包含16384个位，因此每个索引位i只要是1就表示该节点负责处理槽i，那么当节点想要查找它是否处理某个槽时，只需要O(1)的时间复杂度。

1. 传播节点的槽指派信息

一个节点除了将自己负责处理的槽记录在clusterNode结构中以外，还会将自己的slots数组通过消息发送给集群中的其他节点，其他的节点在收到消息后会更新自己的clusterState.nodes字典中的对应节点，并对节点中的slots数组进行更新或保存。

1. 记录集群中的所有槽的指派信息

clusterState结构中的slots数组记录了集群中所有槽的指派信息，每个数组项都是一个指向clusterNode结构的指针，表示每个槽的指派信息。程序要想知道某个槽i是否被指派了，只需要访问该数组的索引i即可，O(1)的时间复杂度。

1. 集群中执行命令
2. 发送命令

集群上线之后，客户端就可以向集群中的节点发送数据命令了，当客户端向节点发送与数据库键有关的命令时，接受命令的节点会计算出命令要处理的数据库键属于哪个槽，并检查这个槽是否指派给了自己。如果这个槽指派给了自己，就直接执行命令；否则该节点会向客户端返回一个MOVED错误，指引客户端转向正确的节点，并再次发送之前的命令。

1. 键的槽分配算法

节点使用以下算法来计算给定键key属于哪个槽：

CRC16(key) & 16383

CRC-16是键的校验和， & 16383是计算出一个介于0-16383之间的整数作为key的槽号。使用cluster keyslot <key>命令可以查看一个键属于哪个槽。

1. MOVED错误的格式

MOVED <slot> <ip>:<port>

slot为键所在的槽，ip和port是负责处理槽slot的节点的IP和端口号。转向操作就是换了一个套接字来发送命令，一个集群客户端通常会与集群中的多个节点创建套接字连接。

注意，集群模式的redis-cli客户端接受到moved错误时，不会打印出moved错误，而是直接自动转向，而单机模式的客户端不清楚moved错误的作用，会直接打印出来。

1. 节点数据库的实现

节点只能使用0号数据库，单机的Redis服务器就没有这个限制。

1. 重新分片
2. 概述

Redis集群的重新分片操作可以将已经指派给某个节点的槽指派给另一个节点，并且相关槽所属的键值对也会从源节点移动到目标节点。重新分片操作可以在线进行，这两个节点依然可以处理命令请求。

1. 原理

Redis集群的重新分片操作是由Redis的集群管理软件redis-trib负责执行的，Redis提供了进行重新分片所需的所有命令，而redis-trib则通过向源节点和目标节点发送命令来进行重新分片操作。步骤如下：

1. 向目标节点发送命令，准备好从源节点导入属于槽slot的键值对；
2. 向源节点发送命令，准备好将属于槽slot的键值对迁移到目标节点；
3. 向源节点发送命令，获得最多count个属于槽slot的键值对的键名；
4. 对于步骤3中的每个键名，都向源节点发送一个迁移命令，迁移至目标节点；
5. 重复执行（3），（4），直到源节点中属于槽slot的所有键值对都被迁移到目标节点为止；
6. 向集群中的任意一个节点发送命令，将槽slot指派给目标节点，然后发送至整个集群；
7. ASK错误
8. 概述

在进行重新分片期间，源节点向目标节点迁移一个槽的过程中，可能会出现这种情况：属于该槽的一部分键值对保存在源节点中，一部分键值对保存在目标节点中。这时，如果客户端发送了一个与该槽有关的命令，源节点会先在自己的数据库中查找对应的键，如果没有找到，则向客户端返回一个ASK错误，指引客户端转向正在导入槽的目标节点，接到ASK错误的客户端，会转向目标节点，然后发送一个ASKING命令，并再次发送之前的命令。

1. ASKING命令

ASKING命令唯一做的事就是打开发送该命令的客户端的redis\_asking标识，然后再携带着这个标识访问目标节点中关于槽i的命令，如果该节点正在导入槽i，那么该节点将会破例执行这个命令一次，之后则会拒绝执行这个命令，并返回MOVED错误。

1. ASK错误和MOVED错误的区别

MOVED错误代表着槽的负责权已经从一个节点转移到了另一个节点，而ASK错误只是两个节点在迁移槽的过程中使用的一种临时措施，在客户端接受到关于槽i的ASK错误之后，客户端只会在接下来的一次命令请求中将关于槽i的命令请求发送至ASK错误指定的节点中，但是不会对客户端今后发送关于槽i的命令请求产生影响。

1. 复制与故障转移
2. 概述

Redis集群中的节点分为主节点和从节点，主节点用于处理槽，从节点用于复制某个主节点，并且在被复制的主节点下线时，代替下线主节点继续处理命令请求。

1. 设置从节点

向一个节点发送命令：

cluster replicate <id>

可以让接受命令的节点称为id所指定节点的从节点，并开始对主节点进行复制，相当于向从节点发送slaveof命令。一个节点成为从节点，并开始复制某个主节点这一信息会通过消息发送给集群中的其他节点。

1. 故障检测

集群中的每个节点都会定期向集群中的其他节点发送PING消息，以此来检测对方是否在线，如果接受PING消息的节点没有在规定时间内返回PONG消息，那么发送PING消息的节点就会将接受PING消息的节点标记为疑似下线。集群中的各个节点会通过互相发送消息的方式来交换集群中各个节点的状态信息，当一个主节点A通过消息得知主节点B认为主节点C疑似下线时，A会将B的下线报告添加到C的结构中，如果一个集群中，半数以上负责处理槽的主节点都将某个主节点报告为疑似下线，那么这个主节点将被标记为已下线，并广播到集群中。

1. 故障转移

如果一个从节点发现自己正在复制的主节点进入了已下线状态时，从节点将开始对下线主节点进行故障转移，步骤如下：

1. 复制所有从节点，选中一个从节点；
2. 被选中的从节点执行slaveof no one命令，成为新的主节点；
3. 新的主节点会撤销所有对已下线主节点的槽指派，并全部指派给自己；
4. 新的主节点向集群广播一条PONG消息，让其他节点知道；
5. 新的主节点开始接受和自己负责处理的槽有关的命令请求，故障转移完成；
6. 选举新的主节点

也是使用基于Raft算法的领头选举方法，与选举领头Sentinel一样。

1. 消息
2. 概述

集群中的各个节点通过发送和接受消息来进行通信，节点发送的消息主要有5中，MEET消息，PING消息，PONG消息，FAIL消息，PUBLISH消息。

1. MEET消息

当发送者接收到客户端发送的cluster meet命令时，发送者会向接受者发送meet消息，请求接受者加入到集群中。

1. ping消息

集群中的每个节点默认每隔1s就从已知节点中随机选5个节点，对最长时间没有发送过PING命令的节点发送PING命令。

1. PONG消息

为了向发送者确认这条MEET消息或PING消息已经到达，接受者会发送PONG消息，另外，一个节点也可以通过向集群发送自己的PONG消息让其他节点刷新认知。

1. FAIL消息

当一个主节点判断另一个主节点进入FAIL状态，就会向集群广播关于B的FAIL消息。

1. PUBLISH消息

当节点收到一条PUBLISH命令，就会执行这个命令，并向集群广播一条PUBLISH消息，所有接受到该消息的节点都会执行相同的PUBLISH命令。

1. gossip协议

Redis集群中的各个节点通过Gossip协议来交换各自关于不同节点的状态信息，Gossip协议由MEET，PING，PONG三种消息实现。