

Redis

原作者github: <https://github.com/CyC2018/CS-Notes>

PDF制作github: <https://github.com/sjsdfg/CS-Notes-PDF>

一、概述

Redis 是速度非常快的非关系型（NoSQL）内存键值数据库，可以存储键和五种不同类型的值之间的映射。

键的类型只能为字符串，值支持五种数据类型：字符串、列表、集合、散列表、有序集合。

Redis 支持很多特性，例如将内存中的数据持久化到硬盘中，使用复制来扩展读性能，使用分片来扩展写性能。

二、数据类型

数据类型	可以存储的值	操作
STRING	字符串、整数或者浮点数	对整个字符串或者字符串的其中一部分执行操作 对整数和浮点数执行自增或者自减操作
LIST	列表	从两端压入或者弹出元素 对单个或者多个元素 进行修剪，只保留一个范围内的元素
SET	无序集合	添加、获取、移除单个元素 检查一个元素是否存在于集合中 计算交集、并集、差集 从集合里面随机获取元素
HASH	包含键值对的无序散列表	添加、获取、移除单个键值对 获取所有键值对 检查某个键是否存在

数据类型	可以存储的值	操作
ZSET	有序集合	添加、获取、删除元素 根据分值范围或者成员来获取元素 计算一个键的排名

What Redis data structures look like

STRING

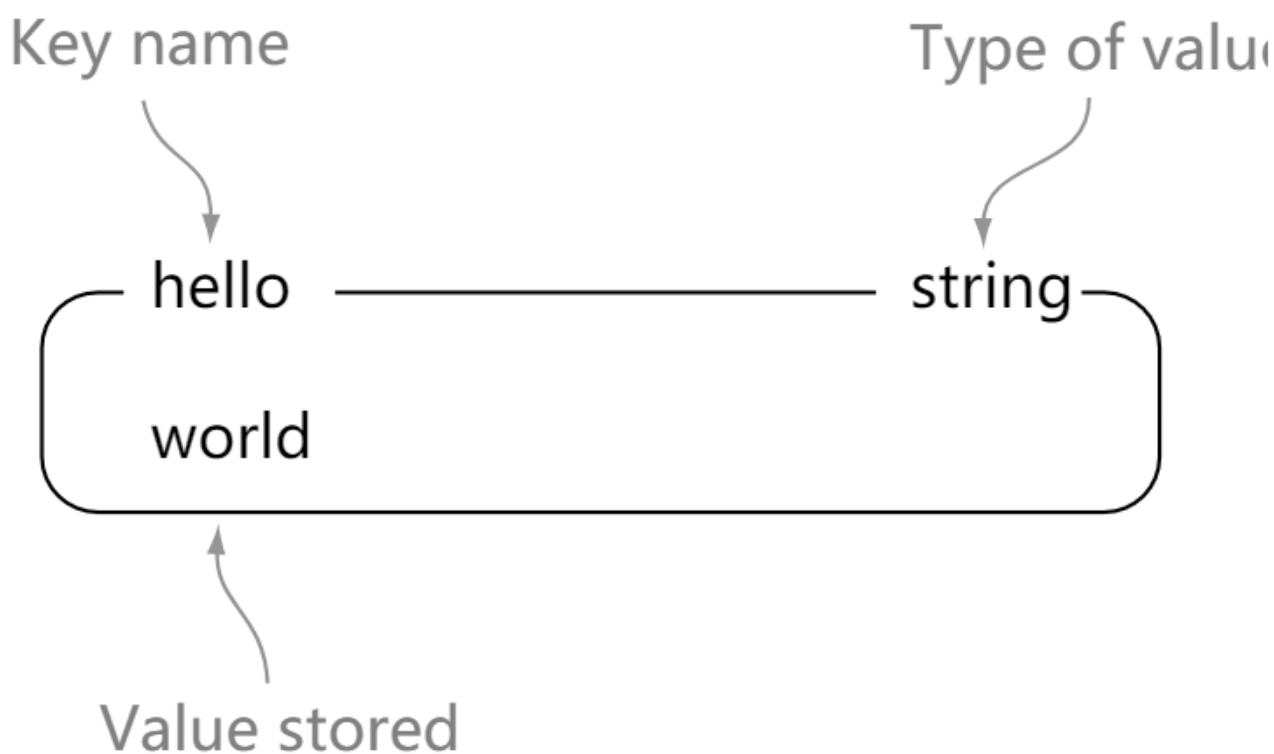


Figure 1.1 An example of a *STRING*, *world*, stored under a key, *hello*

```

1. > set hello world
2. OK
3. > get hello
4. "world"
5. > del hello
6. (integer) 1
7. > get hello
8. (nil)

```

LIST

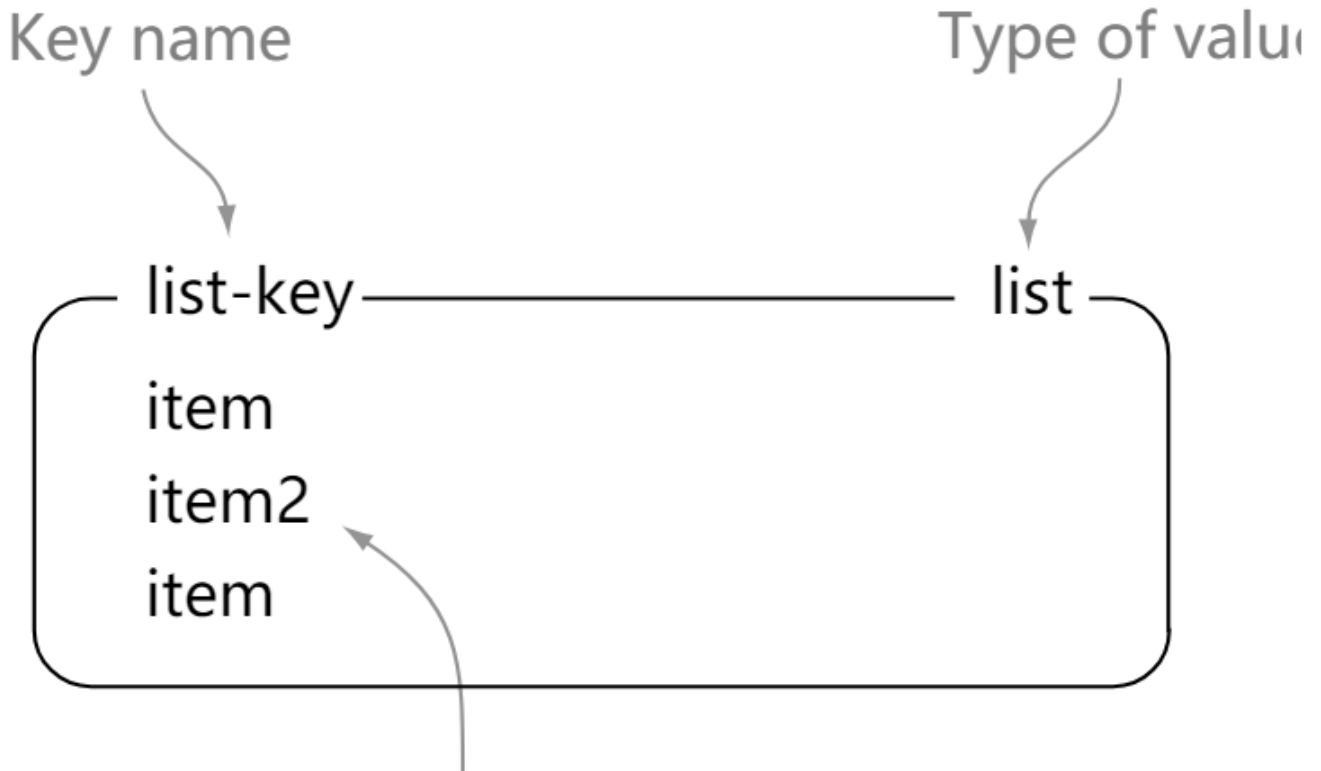
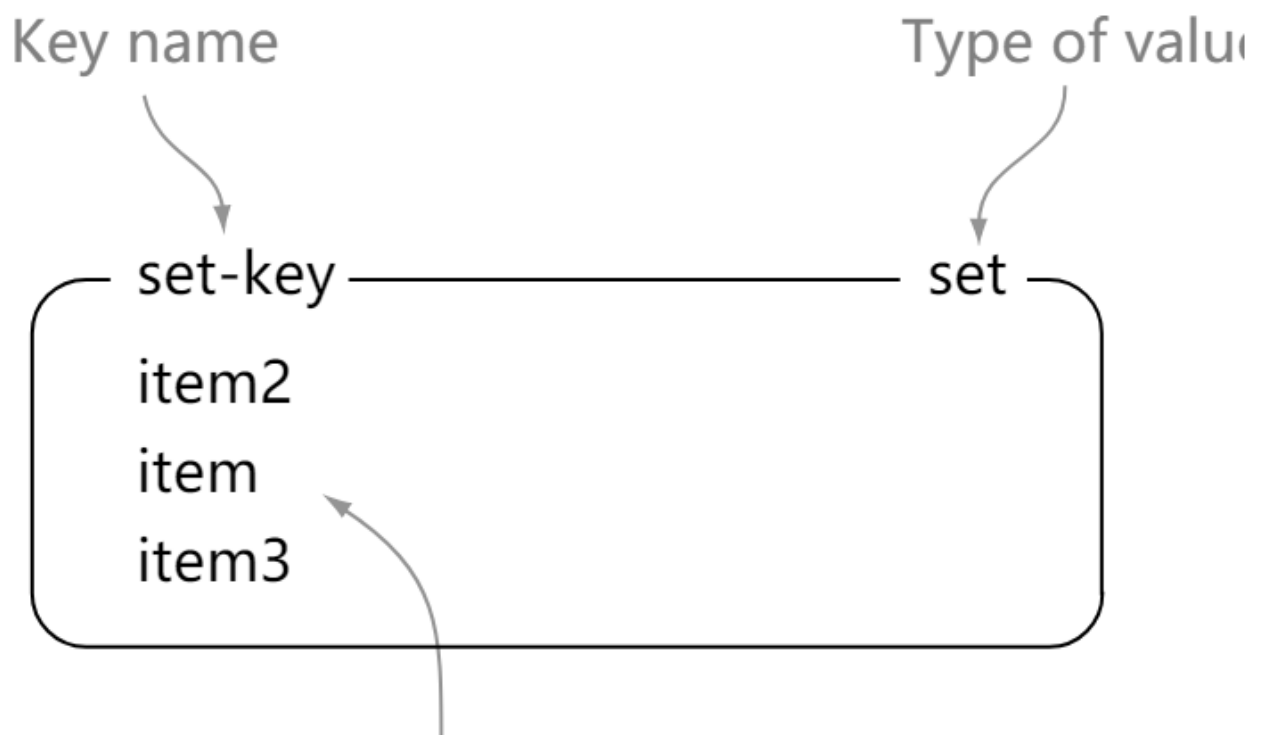


Figure 1.2 An example of a *LIST* with three items under the key, *list-key*. Note that *item* can be in the list more than once.

```
1. > rpush list-key item
2. (integer) 1
3. > rpush list-key item2
4. (integer) 2
5. > rpush list-key item
6. (integer) 3
7.
8. > lrange list-key 0 -1
9. 1) "item"
10. 2) "item2"
11. 3) "item"
12.
13. > lindex list-key 1
14. "item2"
15.
16. > lpop list-key
17. "item"
```

```
18.  
19. > lrange list-key 0 -1  
20. 1) "item2"  
21. 2) "item"
```

SET



Set of distinct values, undefined order

Figure 1.3 An example of a **SET** with three items under the key, **set-key**

```
1. > sadd set-key item  
2. (integer) 1  
3. > sadd set-key item2  
4. (integer) 1  
5. > sadd set-key item3  
6. (integer) 1  
7. > sadd set-key item  
8. (integer) 0  
9.  
10. > smembers set-key  
11. 1) "item"  
12. 2) "item2"  
13. 3) "item3"
```

```
14.
15. > sismember set-key item4
16. (integer) 0
17. > sismember set-key item
18. (integer) 1
19.
20. > srem set-key item2
21. (integer) 1
22. > srem set-key item2
23. (integer) 0
24.
25. > smembers set-key
26. 1) "item"
27. 2) "item3"
```

HASH

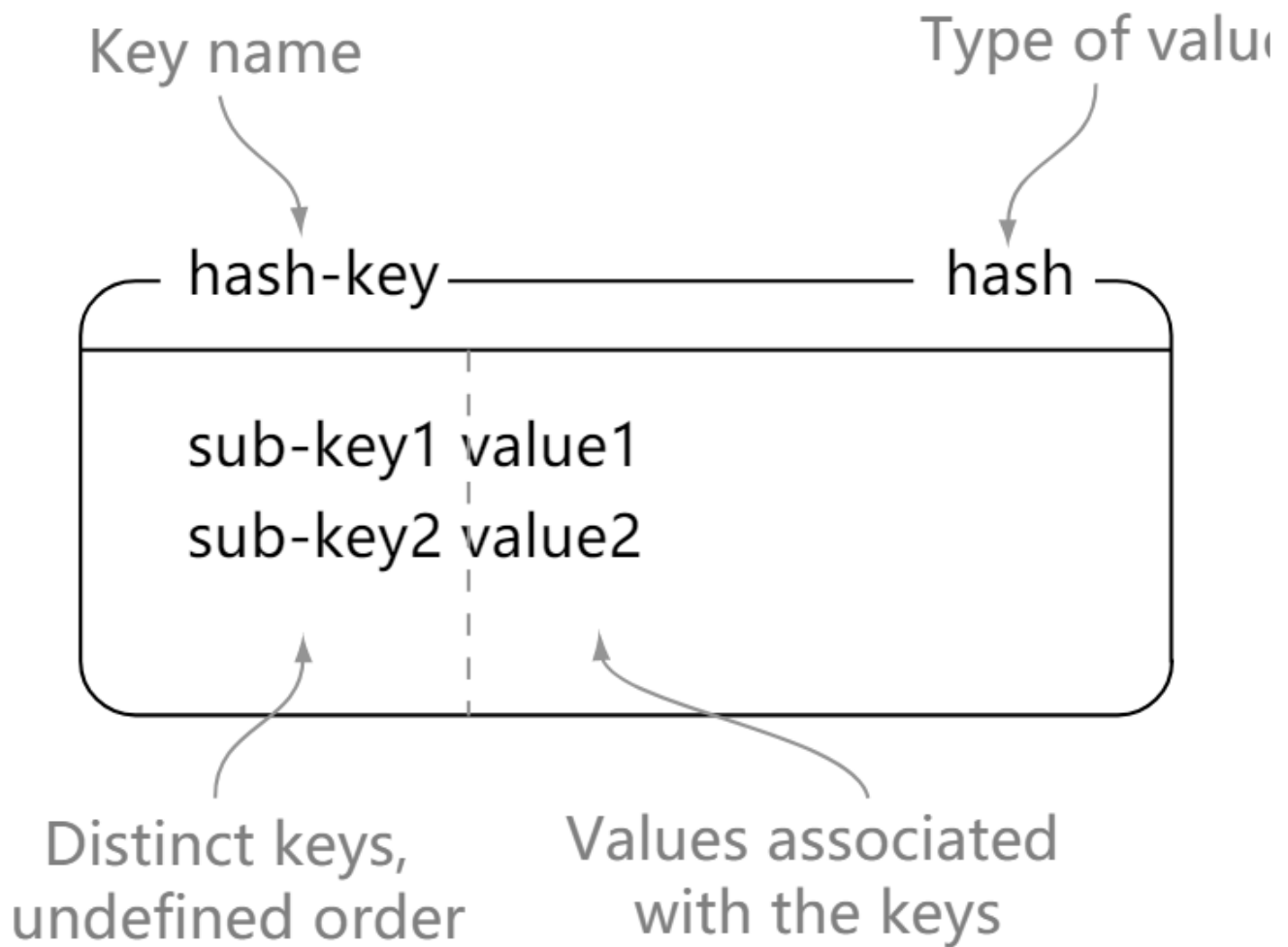


Figure 1.4 An example of a `HASH` with two keys/values under the key `hash-key`

```
1. > hset hash-key sub-key1 value1
2. (integer) 1
3. > hset hash-key sub-key2 value2
4. (integer) 1
5. > hset hash-key sub-key1 value1
6. (integer) 0
7.
8. > hgetall hash-key
9. 1) "sub-key1"
10. 2) "value1"
11. 3) "sub-key2"
12. 4) "value2"
13.
14. > hdel hash-key sub-key2
15. (integer) 1
16. > hdel hash-key sub-key2
17. (integer) 0
18.
```

```
19. > hget hash-key sub-key1
20. "value1"
21.
22. > hgetall hash-key
23. 1) "sub-key1"
24. 2) "value1"
```

ZSET

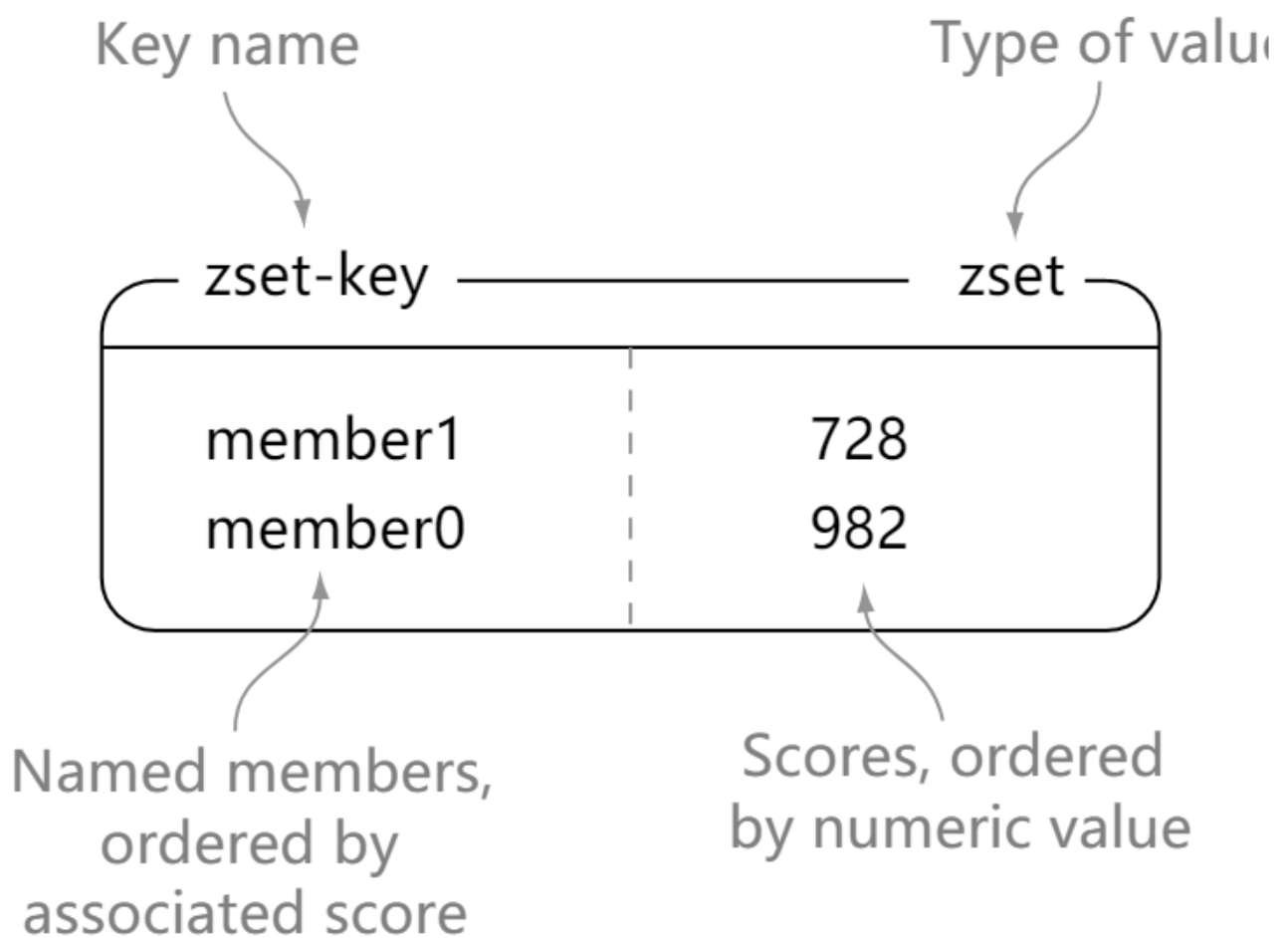


Figure 1.5 An example of a **ZSET** with two members/scores under the key **zset-key**

```
1. > zadd zset-key 728 member1
2. (integer) 1
3. > zadd zset-key 982 member0
4. (integer) 1
5. > zadd zset-key 982 member0
6. (integer) 0
7.
```

```

8. > zrange zset-key 0 -1 withscores
9. 1) "member1"
10. 2) "728"
11. 3) "member0"
12. 4) "982"
13.
14. > zrangebyscore zset-key 0 800 withscores
15. 1) "member1"
16. 2) "728"
17.
18. > zrem zset-key member1
19. (integer) 1
20. > zrem zset-key member1
21. (integer) 0
22.
23. > zrange zset-key 0 -1 withscores
24. 1) "member0"
25. 2) "982"

```

三、数据结构

字典

dictht 是一个散列表结构，使用拉链法保存哈希冲突的 dictEntry。

```

1.  /* This is our hash table structure. Every dictionary has two of this
2.   * implement incremental rehashing, for the old to the new table. */
3.  typedef struct dictht {
4.      dictEntry **table;
5.      unsigned long size;
6.      unsigned long sizemask;
7.      unsigned long used;
8.  } dictht;

```

```

1.  typedef struct dictEntry {
2.      void *key;
3.      union {
4.          void *val;

```



```

5.         uint64_t u64;
6.         int64_t s64;
7.         double d;
8.     } v;
9.     struct dictEntry *next;
10. } dictEntry;

```

Redis 的字典 dict 中包含两个哈希表 dictht，这是为了方便进行 rehash 操作。在扩容时，将其中一个 dictht 上的键值对 rehash 到另一个 dictht 上面，完成之后释放空间并交换两个 dictht 的角色。

```

1.     typedef struct dict {
2.         dictType *type;
3.         void *privdata;
4.         dictht ht[2];
5.         long rehashidx; /* rehashing not in progress if rehashidx == -1 */
6.         unsigned long iterators; /* number of iterators currently running */
7.     } dict;

```

rehash 操作不是一次性完成，而是采用渐进方式，这是为了避免一次性执行过多的 rehash 操作给服务器带来过大的负担。

渐进式 rehash 通过记录 dict 的 rehashidx 完成，它从 0 开始，然后每执行一次 rehash 都会递增。例如在一次 rehash 中，要把 dict[0] rehash 到 dict[1]，这一次会把 dict[0] 上 table[rehashidx] 的键值对 rehash 到 dict[1] 上，dict[0] 的 table[rehashidx] 指向 null，并令 rehashidx++。

在 rehash 期间，每次对字典执行添加、删除、查找或者更新操作时，都会执行一次渐进式 rehash。

采用渐进式 rehash 会导致字典中的数据分散在两个 dictht 上，因此对字典的操作也需要到对应的 dictht 去执行。

```

1.     /* Performs N steps of incremental rehashing. Returns 1 if there are s
2.        * keys to move from the old to the new hash table, otherwise 0 is ret
3.        *

```

```

4.      * Note that a rehashing step consists in moving a bucket (that may ha
ve more
5.      * than one key as we use chaining) from the old to the new hash table
, however
6.      * since part of the hash table may be composed of empty spaces, it is
not
7.      * guaranteed that this function will rehash even a single bucket, sin
ce it
8.      * will visit at max N*10 empty buckets in total, otherwise the amount
of
9.      * work it does would be unbound and the function may block for a long
time. */
10.     int dictRehash(dict *d, int n) {
11.         int empty_visits = n * 10; /* Max number of empty buckets to visit.
*/
12.         if (!dictIsRehashing(d)) return 0;
13.
14.         while (n-- && d->ht[0].used != 0) {
15.             dictEntry *de, *nextde;
16.
17.             /* Note that rehashidx can't overflow as we are sure there are
more
18.             * elements because ht[0].used != 0 */
19.             assert(d->ht[0].size > (unsigned long) d->rehashidx);
20.             while (d->ht[0].table[d->rehashidx] == NULL) {
21.                 d->rehashidx++;
22.                 if (--empty_visits == 0) return 1;
23.             }
24.             de = d->ht[0].table[d->rehashidx];
25.             /* Move all the keys in this bucket from the old to the new ha
sh HT */
26.             while (de) {
27.                 uint64_t h;
28.
29.                 nextde = de->next;
30.                 /* Get the index in the new hash table */
31.                 h = dictHashKey(d, de->key) & d->ht[1].sizemask;
32.                 de->next = d->ht[1].table[h];
33.                 d->ht[1].table[h] = de;
34.                 d->ht[0].used--;
35.                 d->ht[1].used++;
36.                 de = nextde;
37.             }
38.             d->ht[0].table[d->rehashidx] = NULL;
39.             d->rehashidx++;

```

```

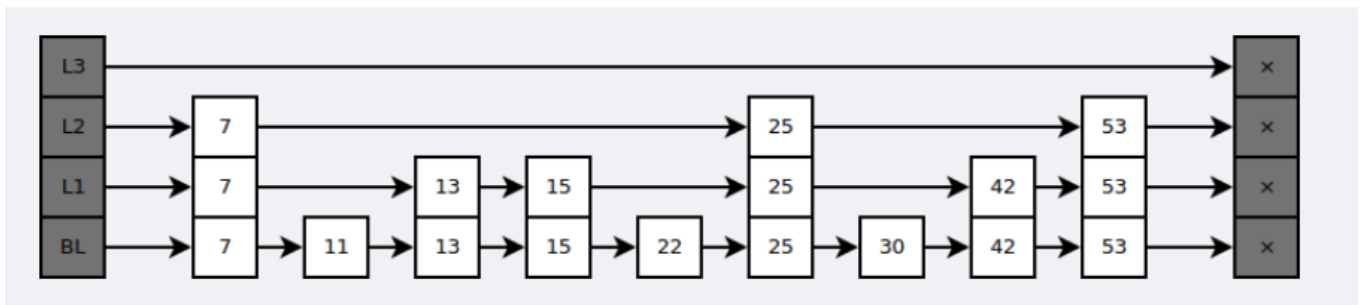
40.     }
41.
42.     /* Check if we already rehashed the whole table... */
43.     if (d->ht[0].used == 0) {
44.         zfree(d->ht[0].table);
45.         d->ht[0] = d->ht[1];
46.         _dictReset(&d->ht[1]);
47.         d->rehashidx = -1;
48.         return 0;
49.     }
50.
51.     /* More to rehash... */
52.     return 1;
53. }

```

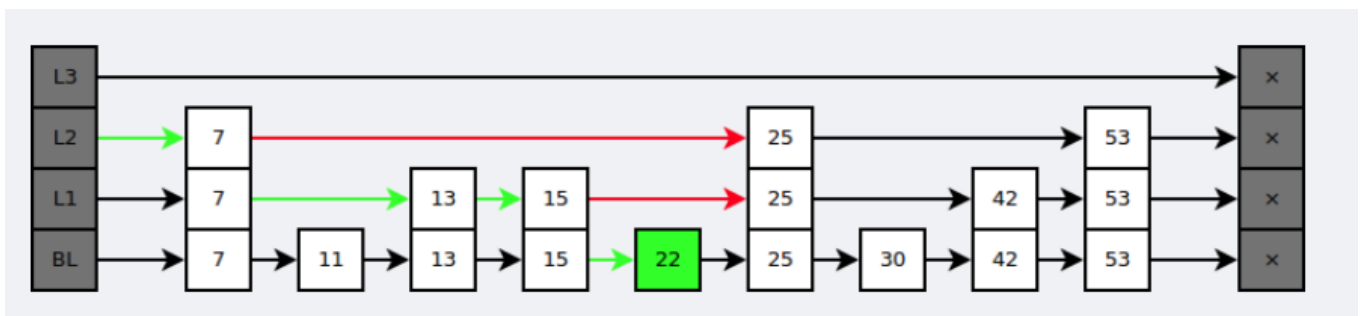
跳跃表

是有序集合的底层实现之一。

跳跃表是基于多指针有序链表实现的，可以看成多个有序链表。



在查找时，从上层指针开始查找，找到对应的区间之后再到下一层去查找。下图演示了查找 22 的过程。



与红黑树等平衡树相比，跳跃表具有以下优点：

- 插入速度非常快，因为不需要进行旋转等操作来维护平衡性；
- 更容易实现；
- 支持无锁操作。

四、使用场景

计数器

可以对 String 进行自增自减运算，从而实现计数器功能。

Redis 这种内存型数据库的读写性能非常高，很适合存储频繁读写的计数量。

缓存

将热点数据放到内存中，设置内存的最大使用量以及淘汰策略来保证缓存的命中率。

查找表

例如 DNS 记录就很适合使用 Redis 进行存储。

查找表和缓存类似，也是利用了 Redis 快速的查找特性。但是查找表的内容不能失效，而缓存的内容可以失效，因为缓存不作为可靠的数据来源。

消息队列

List 是一个双向链表，可以通过 lpop 和 lpush 写入和读取消息。

不过最好使用 Kafka、RabbitMQ 等消息中间件。

会话缓存

在分布式场景下具有多个应用服务器，可以使用 Redis 来统一存储这些应用服务器的会话信息。

当应用服务器不再存储用户的会话信息，也就不再具有状态，一个用户可以请求任意一个应用服务器，从而更容易实现高可用性以及可伸缩性。

分布式锁实现

在分布式场景下，无法使用单机环境下的锁来对多个节点上的进程进行同步。

可以使用 Redis 自带的 SETNX 命令实现分布式锁，除此之外，还可以使用官方提供的 RedLock 分布式锁实现。

其它

Set 可以实现交集、并集等操作，从而实现共同好友等功能。

ZSet 可以实现有序性操作，从而实现排行榜等功能。

五、Redis 与 Memcached

两者都是非关系型内存键值数据库，主要有以下不同：

数据类型

Memcached 仅支持字符串类型，而 Redis 支持五种不同的数据类型，可以更灵活地解决问题。

数据持久化

Redis 支持两种持久化策略：RDB 快照和 AOF 日志，而 Memcached 不支持持久化。

分布式

Memcached 不支持分布式，只能通过客户端使用一致性哈希来实现分布式存储，这种方式在存储和查询时都需要先在客户端计算一次数据所在的节点。

Redis Cluster 实现了分布式的支持。

内存管理机制

- 在 Redis 中，并不是所有数据都一直存储在内存中，可以将一些很久没用的 value 交换到磁盘，而 Memcached 的数据则会一直在内存中。
- Memcached 将内存分割成特定长度的块来存储数据，以完全解决内存碎片的问题。但是这种方式会使得内存的利用率不高，例如块的大小为 128 bytes，只存储 100 bytes 的数据，那么剩下的 28 bytes 就浪费掉了。

六、键的过期时间

Redis 可以为每个键设置过期时间，当键过期时，会自动删除该键。

对于散列表这种容器，只能为整个键设置过期时间（整个散列表），而不能为键里面的单个元素设置过期时间。

七、数据淘汰策略

可以设置内存最大使用量，当内存使用量超出时，会施行数据淘汰策略。

Redis 具体有 6 种淘汰策略：

策略	描述
volatile-lru	从已设置过期时间的数据集中挑选最近最少使用的数据淘汰

策略	描述
volatile-ttl	从已设置过期时间的数据集中挑选将要过期的数据淘汰
volatile-random	从已设置过期时间的数据集中任意选择数据淘汰
allkeys-lru	从所有数据集中挑选最近最少使用的数据淘汰
allkeys-random	从所有数据集中任意选择数据进行淘汰
noeviction	禁止驱逐数据

作为内存数据库，出于对性能和内存消耗的考虑，Redis 的淘汰算法实际实现上并非针对所有 key，而是抽样一小部分并且从中选出被淘汰的 key。

使用 Redis 缓存数据时，为了提高缓存命中率，需要保证缓存数据都是热点数据。可以将内存最大使用量设置为热点数据占用的内存量，然后启用 allkeys-lru 淘汰策略，将最近最少使用的数据淘汰。

Redis 4.0 引入了 volatile-lfu 和 allkeys-lfu 淘汰策略，LFU 策略通过统计访问频率，将访问频率最少的键值对淘汰。

八、持久化

Redis 是内存型数据库，为了保证数据在断电后不会丢失，需要将内存中的数据持久化到硬盘上。

RDB 持久化

将某个时间点的所有数据都存放到硬盘上。

可以将快照复制到其它服务器从而创建具有相同数据的服务器副本。

如果系统发生故障，将会丢失最后一次创建快照之后的数据。

如果数据量很大，保存快照的时间会很长。

AOF 持久化

将写命令添加到 AOF 文件 (Append Only File) 的末尾。

使用 AOF 持久化需要设置同步选项，从而确保写命令什么时候会同步到磁盘文件上。这是因为对文件进行写入并不会马上将内容同步到磁盘上，而是先存储到缓冲区，然后由操作系统决定什么时候同步到磁盘。有以下同步选项：

选项	同步频率
always	每个写命令都同步
everysec	每秒同步一次
no	让操作系统来决定何时同步

- always 选项会严重减低服务器的性能；
- everysec 选项比较合适，可以保证系统崩溃时只会丢失一秒左右的数据，并且 Redis 每秒执行一次同步对服务器性能几乎没有任何影响；
- no 选项并不能给服务器性能带来多大的提升，而且也会增加系统崩溃时数据丢失的数量。

随着服务器写请求的增多，AOF 文件会越来越大。Redis 提供了一种将 AOF 重写的特性，能够去除 AOF 文件中的冗余写命令。

九、事务

一个事务包含了多个命令，服务器在执行事务期间，不会改去执行其它客户端的命令请求。

事务中的多个命令被一次性发送给服务器，而不是一条一条发送，这种方式被称为流水线，它可以减少客户端与服务器之间的网络通信次数从而提升性能。

Redis 最简单的事务实现方式是使用 MULTI 和 EXEC 命令将事务操作包围起来。

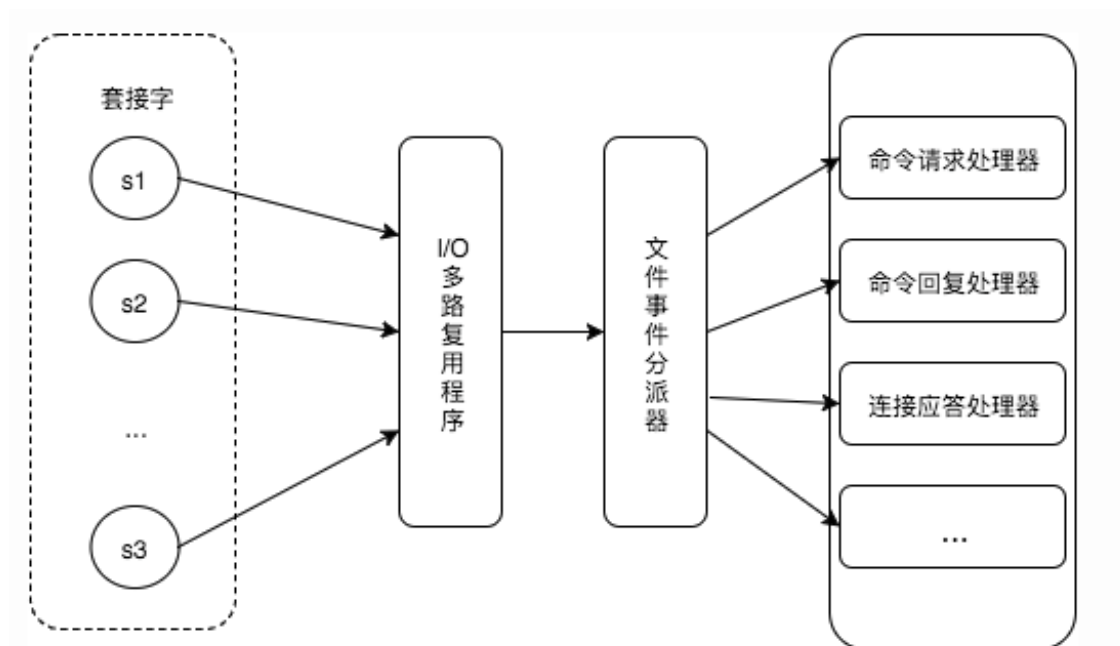
十、事件

Redis 服务器是一个事件驱动程序。

文件事件

服务器通过套接字与客户端或者其它服务器进行通信，文件事件就是对套接字操作的抽象。

Redis 基于 Reactor 模式开发了自己的网络事件处理器，使用 I/O 多路复用程序来同时监听多个套接字，并将到达的事件传送给文件事件分派器，分派器会根据套接字产生的事件类型调用相应的事件处理器。



时间事件

服务器有一些操作需要在给定的时间点执行，时间事件是对这类定时操作的抽象。

时间事件又分为：

- 定时事件：是让一段程序在指定的时间之内执行一次；
- 周期性事件：是让一段程序每隔指定时间就执行一次。

Redis 将所有时间事件都放在一个无序链表中，通过遍历整个链表查找出已到达的时间事件，并调用相应的事件处理器。

事件的调度与执行

服务器需要不断监听文件事件的套接字才能得到待处理的文件事件，但是不能一直监听，否则时间事件无法在规定的时间内执行，因此监听时间应该根据距离现在最近的时间事件来决定。

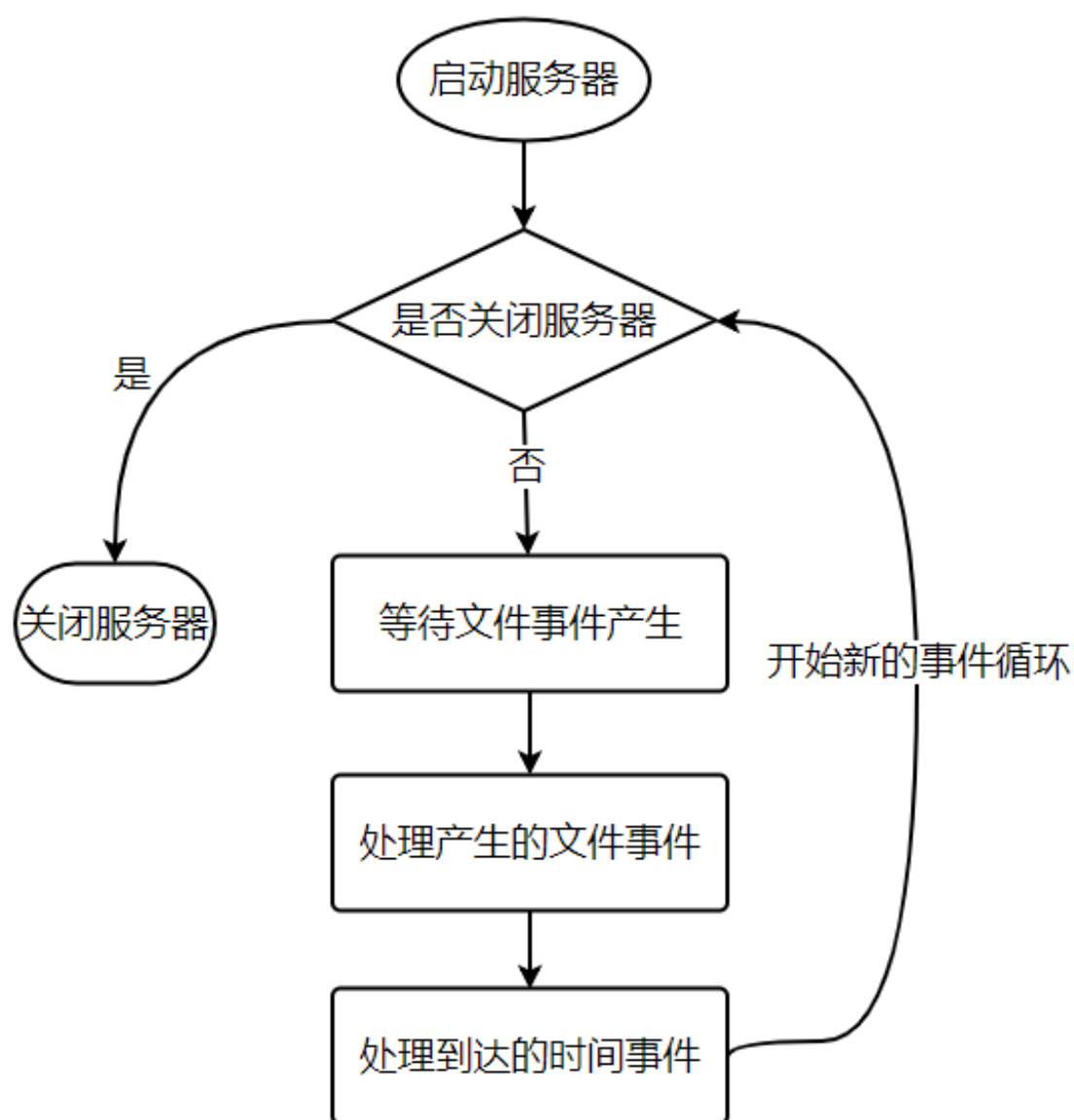
事件调度与执行由 `aeProcessEvents` 函数负责，伪代码如下：

```
1.  def aeProcessEvents():
2.      # 获取到达时间离当前时间最近的时间事件
3.      time_event = aeSearchNearestTimer()
4.      # 计算最近的时间事件距离到达还有多少毫秒
5.      remaind_ms = time_event.when - unix_ts_now()
6.      # 如果事件已到达，那么 remaind_ms 的值可能为负数，将它设为 0
7.      if remaind_ms < 0:
8.          remaind_ms = 0
9.      # 根据 remaind_ms 的值，创建 timeval
10.     timeval = create_timeval_with_ms(remaind_ms)
11.     # 阻塞并等待文件事件产生，最大阻塞时间由传入的 timeval 决定
12.     aeApiPoll(timeval)
13.     # 处理所有已产生的文件事件
14.     procesFileEvents()
15.     # 处理所有已到达的时间事件
16.     processTimeEvents()
```

将 `aeProcessEvents` 函数置于一个循环里面，加上初始化和清理函数，就构成了 Redis 服务器的主函数，伪代码如下：

```
1.  def main():
2.      # 初始化服务器
3.      init_server()
4.      # 一直处理事件，直到服务器关闭为止
5.      while server_is_not_shutdown():
6.          aeProcessEvents()
7.      # 服务器关闭，执行清理操作
8.      clean_server()
```

从事件处理的角度来看，服务器运行流程如下：



十一、复制

通过使用 `slaveof host port` 命令来让一个服务器成为另一个服务器的从服务器。

一个从服务器只能有一个主服务器，并且不支持主主复制。

连接过程

1. 主服务器创建快照文件，发送给从服务器，并在发送期间使用缓冲区记录执行的写命令。快照文件发送完毕之后，开始向从服务器发送存储在缓冲区中的写命令；
2. 从服务器丢弃所有旧数据，载入主服务器发来的快照文件，之后从服务器开始接受主服务器发来的写命令；
3. 主服务器每执行一次写命令，就向从服务器发送相同的写命令。

主从链

随着负载不断上升，主服务器可能无法很快地更新所有从服务器，或者重新连接和重新同步从服务器将导致系统超载。为了解决这个问题，可以创建一个中间层来分担主服务器的复制工作。中间层的服务器是最上层服务器的从服务器，又是最下层服务器的主服务器。

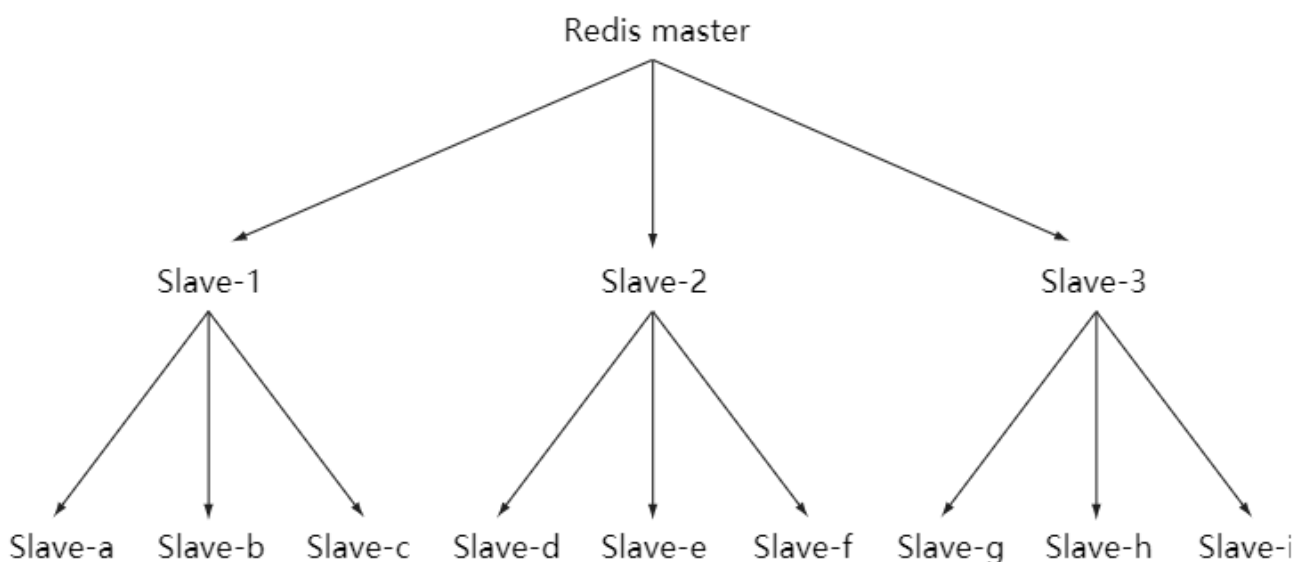


Figure 4.1 An example Redis master/slave replica tree with nine lowest-level slaves and three intermediate replication helper servers

十二、Sentinel

Sentinel（哨兵）可以监听集群中的服务器，并在主服务器进入下线状态时，自动从从服务器中选举出新的主服务器。

十三、分片

分片是将数据划分为多个部分的方法，可以将数据存储到多台机器里面，这种方法在解决某些问题时可以获得线性级别的性能提升。

假设有 4 个 Reids 实例 R0, R1, R2, R3, 还有很多表示用户的键 user:1, user:2, ... , 有不同的方式来选择指定的键存储在哪个实例中。

- 最简单的方式是范围分片，例如用户 id 从 0~1000 的存储到实例 R0 中，用户 id 从 1001~2000 的存储到实例 R1 中，等等。但是这样需要维护一张映射范围表，维护操作代价很高。
- 还有一种方式是哈希分片，使用 CRC32 哈希函数将键转换为一个数字，再对实例数量求模就能知道应该存储的实例。

根据执行分片的位置，可以分为三种分片方式：

- 客户端分片：客户端使用一致性哈希等算法决定键应当分布到哪个节点。
- 代理分片：将客户端请求发送到代理上，由代理转发请求到正确的节点上。
- 服务器分片：Redis Cluster。

十四、一个简单的论坛系统分析

该论坛系统功能如下：

- 可以发布文章；
- 可以对文章进行点赞；
- 在首页可以按文章的发布时间或者文章的点赞数进行排序显示。

文章信息

文章包括标题、作者、赞数等信息，在关系型数据库中很容易构建一张表来存储这些信息，在 Redis 中可以使用 HASH 来存储每种信息以及其对应的值的映射。

Redis 没有关系型数据库中的表这一概念来将同种类型的数据存放在一起，而是使用命名空间的方式来实现这一功能。键名的前面部分存储命名空间，后面部分的内容存储 ID，通常使用：来进行分隔。例如下面的 HASH 的键名为 article:92617，其中 article 为命名空间，ID 为 92617。

article:92617 — hash	
title	Go to statement considered harmful
link	http://goo.gl/kZUSu
poster	user:83271
time	1331382699.33
votes	528

Figure 1.8 An example article stored as a *HASH* for our article voting system

点赞功能

当有用户为一篇文章点赞时，除了要对该文章的 votes 字段进行加 1 操作，还必须记录该用户已经对该文章进行了点赞，防止用户点赞次数超过 1。可以建立文章的已投票用户集合来进行记录。

为了节约内存，规定一篇文章发布满一周之后，就不能再对它进行投票，而文章的已投票集合也会被删除，可以为文章的已投票集合设置一个一周的过期时间就能实现这个规定。

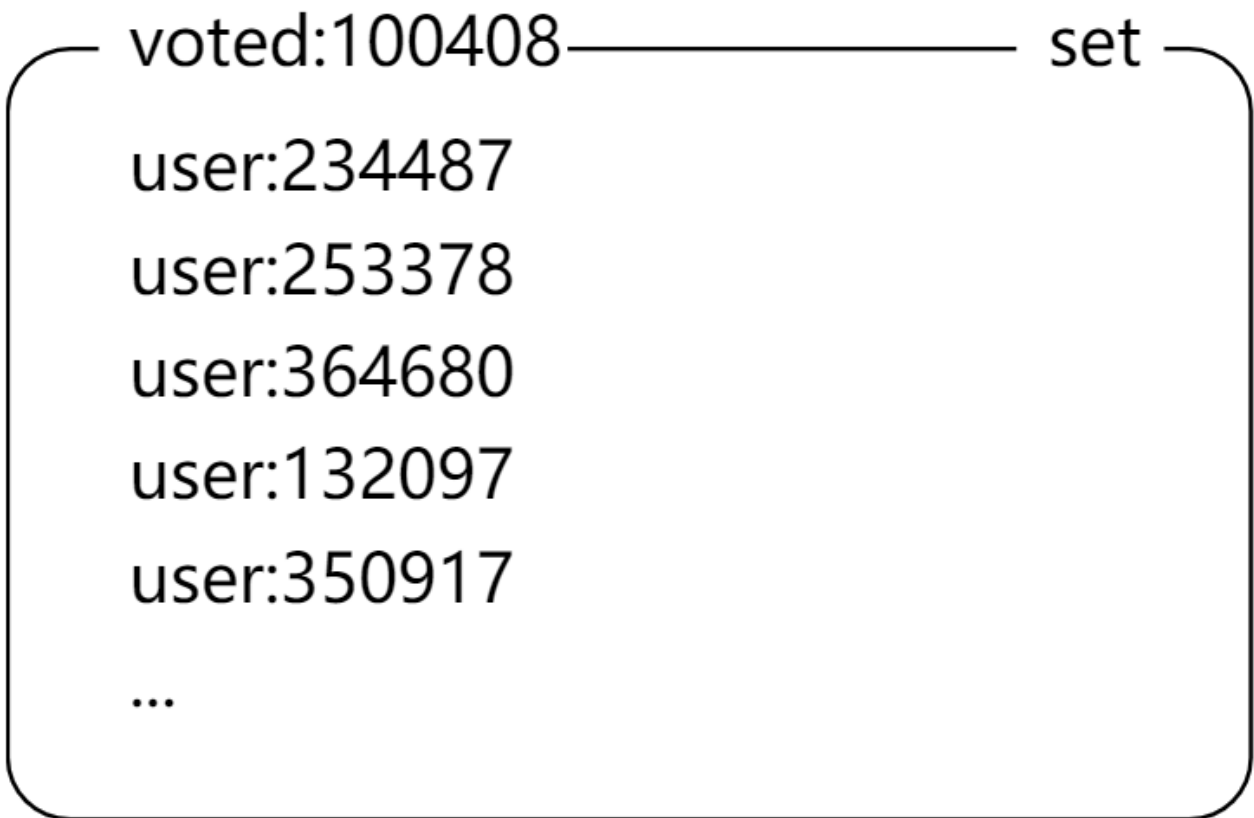


Figure 1.10 Some users who have voted for article 100408

对文章进行排序

为了按发布时间和点赞数进行排序，可以建立一个文章发布时间的有序集合和一个文章点赞数的有序集合。（下图中的 `score` 就是这里所说的点赞数；下面所示的有序集合分值并不直接是时间和点赞数，而是根据时间和点赞数间接计算出来的）



Figure 1.9 Two sorted sets representing time-ordered and score-ordered article indexes

参考资料

- Carlson J L. Redis in Action[J]. Media.johnwiley.com.au, 2013.
- 黄健宏. Redis 设计与实现 [M]. 机械工业出版社, 2014.
- REDIS IN ACTION
- Skip Lists: Done Right
- 论述 Redis 和 Memcached 的差异
- Redis 3.0 中文版- 分片
- Redis 应用场景
- Using Redis as an LRU cache

github: <https://github.com/sjsdfg/CS-Notes-PDF>