如何限制 zookeeper 的 transaction log 大小

在 Zookeeper 中限制 transaction log 总大小主要有两种方法。

限制 Zookeeper Transaction Log 里面的事务条数

默认情况下，在写入 snapCount(100000) 事务后，Zookeeper 事务日志将会切换。如果 Zookeeper 的数据目录的空间不足与存储三个版本的 Zookeeper Transaction Logs，那么我们需要限制每个 Transaction Logs 日志的大小。

比如，如果我们把 zoo.cfg 文件里面的 snapCount 参数设置为20，那么在我的测试环境下，将产生以下的事务日志：

```
[root@v7 version-2]# ls -altr log*
-rw-r--r-- 1 mapr mapr 67108880 Jan 11 18:40 log.40000000d
-rw-r--r-- 1 mapr mapr 67108880 Jan 11 18:40 log.40000001f
-rw-r--r-- 1 mapr mapr 67108880 Jan 11 18:40 log.400000030

[root@v7 version-2]#  java -cp /opt/mapr/lib/zookeeper-3.4.5-mapr-1503.jar:/opt/mapr/lib/log4j-1.2.17.jar:/opt/mapr/lib/slf4j-log4j12-1.7.12.jar:/opt/mapr/lib/slf4j-api-1.7.12.jar org.apache.zookeeper.server.LogFormatter log.40000000d|tail -1
EOF reached after 18 txns.
[root@v7 version-2]#  java -cp /opt/mapr/lib/zookeeper-3.4.5-mapr-1503.jar:/opt/mapr/lib/log4j-1.2.17.jar:/opt/mapr/lib/slf4j-log4j12-1.7.12.jar:/opt/mapr/lib/slf4j-api-1.7.12.jar org.apache.zookeeper.server.LogFormatter log.40000001f|tail -1
EOF reached after 17 txns.
[root@v7 version-2]#  java -cp /opt/mapr/lib/zookeeper-3.4.5-mapr-1503.jar:/opt/mapr/lib/log4j-1.2.17.jar:/opt/mapr/lib/slf4j-log4j12-1.7.12.jar:/opt/mapr/lib/slf4j-api-1.7.12.jar org.apache.zookeeper.server.LogFormatter log.400000030|tail -1
EOF reached after 8 txns.
```

限制 Zookeeper Transaction Log 的块大小

尽管我们进行了上面的设置，从上面的输出我们可以看到每个日志仍然是64MB。很显然，这些文件里面的很多空间是让费的，因为最小块大小被设置为 preAllocSize 参数配置的值。

比如，我们将 zoo.cfg 文件里面的 preAllocSize 参数设置为 1MB，然后重启 Zookeeper：

```
[root@v5 conf]# cat zoo.cfg |grep pre
preAllocSize=1000
```

重启之后，Zookeeper Transaction Log 最新的块大小变成了 1MB：

```
[root@v7 version-2]# ls -altr log*
-rw-r--r-- 1 mapr mapr 67108880 Jan 11 18:38 log.400000001
-rw-r--r-- 1 mapr mapr 67108880 Jan 11 18:40 log.40000000d
-rw-r--r-- 1 mapr mapr 67108880 Jan 11 18:40 log.40000001f
-rw-r--r-- 1 mapr mapr 67108880 Jan 11 18:40 log.400000030
-rw-r--r-- 1 mapr mapr  1024016 Jan 11 19:43 log.500000001
-rw-r--r-- 1 mapr mapr  1024016 Jan 11 19:43 log.500000013
-rw-r--r-- 1 mapr mapr  1024016 Jan 11 19:43 log.500000022
-rw-r--r-- 1 mapr mapr  1024016 Jan 11 19:43 log.50000002e
-rw-r--r-- 1 mapr mapr  1024016 Jan 11 19:43 log.50000003c
```

关于 Zookeeper 的磁盘空间预分配策略可以参见过往记忆大数据这篇文章：Apache Zookeeper 磁盘空间预分配策略。

preAllocSize 和 snapCount 两个参数解释

Zookeeper 官方文档里面有对这两个参数的详细介绍，参见 https://zookeeper.apache.org/doc/current/zookeeperAdmin.html。

preAllocSize

(Java system property: zookeeper.preAllocSize) To avoid seeks ZooKeeper allocates space in the transaction log file in blocks of preAllocSize kilobytes. The default block size is 64M. One reason for changing the size of the blocks is to reduce the block size if snapshots are taken more often. (Also, see snapCount and snapSizeLimitInKb).

snapCount

(Java system property: zookeeper.snapCount) ZooKeeper records its transactions using snapshots and a transaction log (think write-ahead log).The number of transactions recorded in the transaction log before a snapshot can be taken (and the transaction log rolled) is determined by snapCount. In order to prevent all of the machines in the quorum from taking a snapshot at the same time, each ZooKeeper server will take a snapshot when the number of transactions in the transaction log reaches a runtime generated random value in the [snapCount/2+1, snapCount] range.The default snapCount is 100,000.