Zookeeper数据与存储

zookeeper的数据模型是树结构。在内存数据库中，存储了整棵树的内容，包括所有的节点路径、节点数据、ACL信息，Zookeeper会定时将这个数据存储到磁盘上。

DataTree

DataTree是内存数据存储的核心，是一个树结构，代表了内存中一份完整的数据。DataTree不包含任何与网络、客户端连接及请求处理相关的业务逻辑，是一个独立的组件。

DataNode

DataNode是数据存储的最小单元，其内部保存了节点的数据内容、ACL列表、节点状态之外，还记录了父节点的引用和子节点列表两个属性，其也提供了对子节点列表进行操作的数据接口。

ZKDatabase

Zookeeper的内存数据库，管理Zookeeper的所有会话、DataTree存储和事务日志。ZKDatabase会定时向磁盘dump快照数据。同时在Zookeeper启动的时候，会通过磁盘事务日志和快照文件恢复成一个完整的内存数据库。

事务日志

文件存储

在配置Zookeeper集群时需要配置dataDir目录，其用来存储事务日志文件。也可以为事务日志单独分配一个文件存储目录：dataLogDir。若配置dataLogDir为/home/admin/zkData/zk_log,那么Zookeeper在运行过程中会在该目录下建立一个名字为version-2的子目录，该目录确定了当前Zookeeper使用事务日志格式版本号，当下次某个Zookeeper版本对事务日志格式进行变更时，此目录也会变更，即version-2目录下回生成一系列大小一致的文件。

日志写入

1. 确定是否有事务日志可写
2. 确定事务日志文件是否需要扩容
3. 事件序列化
4. 生成Checksum
5. 写入事务日志文件流
6. 事务日志刷入磁盘

日志截断

在Zookeeper运行过程中，可能出现非Leader记录的事务ID比Leader上大，这是非法运行状态。此时，需要保证所有机器必须与该Leader的数据保持同步，即Leader会发送TRUNC命令给该机器，要求进行日志截断，Learner收到该命令之后，就会删除所有包含或大于该事务的日志文件。

snapshot-数据快照

数据快照是Zookeeper数据存储中非常核心的运行机制，数据快照用来记录Zookeeper服务器上某一时刻的全量内存数据，并将其写入指定的磁盘文件中。

文件存储

与事务文件类似，Zookeeper快照文件也可以指定特定的磁盘目录，通过dataDir属性来配置。若指定dataDir为/home/admin/zkData/zk_data，则在运行过程中会在改目录下创建version-2目录，该目录确定了当前Zookeeper使用的快照数据格式的版本号。

数据快照：

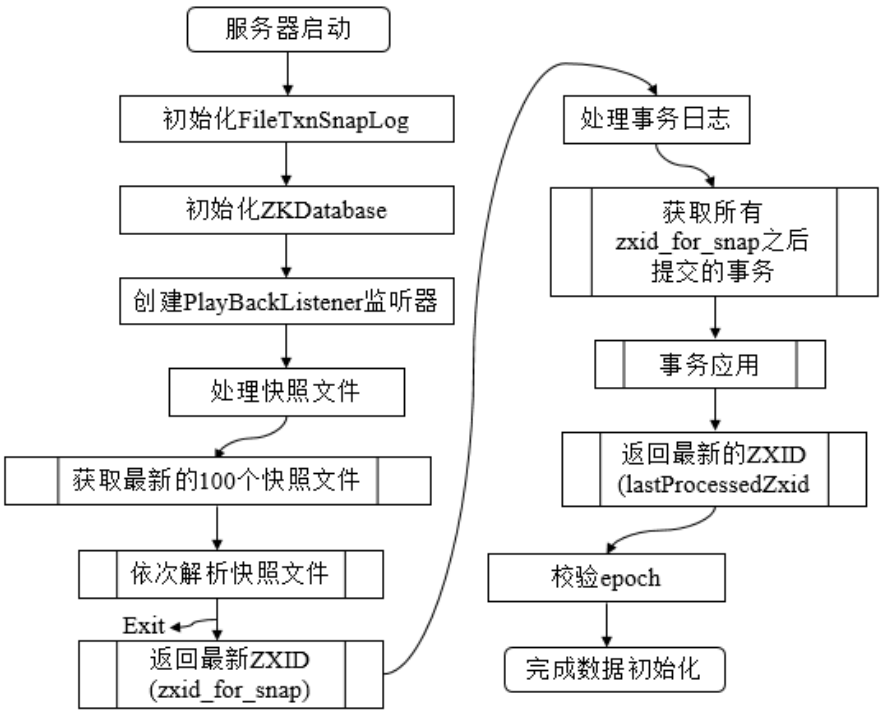
FileSnap负责维护快照数据对外的接口，包括快照数据的写入和读取等，将内存数据库写入快照数据文件其实是一个序列化的过程，针对客户端的每一次事务操作，Zookeeper都会将他们记录到事务日志文件中，同时也会将数据变更应用到内存数据库中，Zookeeper在执行若干次事务日志记录后，将内存数据库的全量数据Dump到本地文件中，这就是数据快照。

步骤：

1. 确定是否需要进行数据快照。
2. 创建事务日志文件
3. 创建数据快照异步线程。
4. 获取全量数据和会话信息
5. 生成快照数据和文件名。
6. 数据序列化。

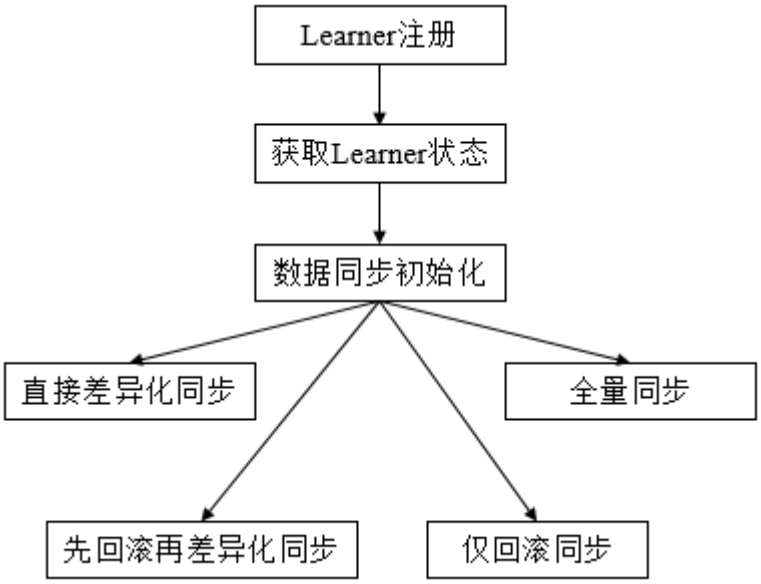
初始化：

在Zookeeper服务器启动期间，首先进行数据初始化工作，用于存储在磁盘上的数据文件加载到Zookeeper服务器内存中。



数据同步：

在整个集群完成Leader选举后，Learner会向Leader进行注册，当Learner向Leader完成注册之后，就进入数据同步环节，同步过程就是Leader将那些没有在Learner服务器上提交过的事务请求同步给Learner服务器。



1. 获取Learner状态。在注册Learner的最后阶段，Learner服务器会发送给Leader服务器一个ACKEPOCH数据包，Leader会从这个数据包中解析出该Learner的currentEpoch和lastZxid
2. 数据同步初始化。首先从Zookeeper内存数据库中提取出事务请求对应的题意缓存队列proposals，同时完成peerLastZxid（该Learner最后处理的ZXID）、minCommittedLog(Leader提议缓存队列commitedLog中最小的ZXID)、maxCommittedLog(Leader提议缓存队列commitedLog中最大ZXID)三个ZXID值的初始化。
3.