Cassandra内部运作

Cassandra内部运作——Gossip和Hinted Handoff

Cassandra内部运作 – 技巧  
By Mike Perham Translated By Jametong

在前面的文章中,我介绍了Cassandra如何进行数据[读](http://www.dbthink.com/?p=432)/[写](http://www.dbthink.com/?p=420).在此文中,我想要解释Cassandra中的一些技巧,Cassandra利用它们来提供一个可伸缩的分布式系统.

**闲话协议(Gossip)**

Cassandra是一个有单个节点组成的集群 – 其中没有“主”节点或单点故障-因此，每个节点都必须积极地确认集群中其他节点的状态。它们使用一个称为[闲话](http://wiki.apache.org/cassandra/ArchitectureGossip)（Gossip）的机制来做此事.每个节点每秒中都会将集群中每个节点的状态“以闲话的方式传播”到1-3个其他节点.系统为闲话数据添加了版本,因此一个节点的任何变更都会快速地传播遍整个集群.通过这种方式,每个节点都能知道任一其他节点的当前状态:是在正在自举呢, 还是正常运行呢,等等.

**提示移交(Hinted Handoff)**

在关于写操作的文章中,我提到Cassandra会存储数据的拷贝到N个节点.客户端可以根据数据的重要性选择一个一致性级别(Consistency level),例如, ConsistencyLevel.QUORUM表示,只有这N个节点中的多数返回成功才表示这个写操作成功.

如果这些节点中的一个宕机了,会发生什么呢?写操作稍后将如何传递到此节点呢?Cassandra使用了一种称为[提示移交](http://wiki.apache.org/cassandra/HintedHandoff)(Hinted Handoff)的技术来解决此问题,其中数据会被写入并保存到另一个随机节点X,并提示这些数据需要被保存到节点Y,并在节点重新在线时进行重放(记住,当节点Y变成在线时,闲话机制会快速通知X节点).提示移交可以确保节点Y可以快速的匹配上集群中的其他节点.注意,如果提示移交由于某种原因没有起作用,读修复最终仍然会“修复”这些过期数据，不过只有当客户端访问这些数据时才会进行读修复.

提示的写是不可读的(因为节点X并不是这N份拷贝的其中一个正式节点),因此,它们并不会记入写一致性.如果Cassandra的配置了3份拷贝,而其中的两个节点不可用,就不可能实现一个ConsistencyLevel.QUORUM的写操作.

**逆熵(Anti-Entropy)**

Cassandra的最后一个众所周知的秘密武器是[逆熵](http://wiki.apache.org/cassandra/ArchitectureAntiEntropy)(Anti-entropy).逆熵明确保证集群中的节点一致认可当前数据.如果由于默认情况,读修复(read repair)与提示移交(hinted handoff)都没有生效,逆熵会确保节点达到最终一致性.逆熵服务是在“主压缩”(等价与关系数据库中的重建表)时运行的,因此，它是一个相对重量级但运行不频繁的进程.逆熵使用[Merkle树](http://en.wikipedia.org/wiki/Hash_tree)(也称为散列树)来确定节点在列族(column family)数据树内的什么位置不能一致认可,接着修复该位置的每一个分支.