对云环境中改进Mongodb自动分片技术的研究

摘要

随着网际网路2.0技术的快速发展，大规模分布式服务的需求，给存储在云计算带来的巨大的挑战，动摇了传统的的关系型数据库。NoSQL数据库从RDBMS的桎梏中解放出来成为人们关注的焦点。在这篇文章中，首先本着自动的原则和实施机制，提出了MongoDB数据库分片技术，而后为了解决分布不均的问题提出一个在数据操作的频率的基础上的改进的数据自动分片算法. 改进的均衡策略能有效平衡之间的数据碎片，提高集群的并发读写性能。

关键词 NoSQL，MongoDB，自动分片，平衡策略

一、绪论

近年来，随着互联网数据量的快速增长和互联网的Web 2.0技术的发展，如何有效地存储，处理和提取的大量数据，成为一个迫切需要解决的问题，云计算在这种情境下应运而生。云计算通过网络（通常是因特网）[1]以计算而不是一个产品的形式作为服务的交付，从而共享资源，软件和信息提供给计算机和其他作为计量的服务设备。许多大学，供应商和政府组织各地投资于发展云计算的项目，例如：亚马逊推出简单存储服务（S3）和弹性计算云（EC2），谷歌推动的BigTable和MapReduce，GFS，已经全部成功地应用在生产环境中。分布式文件

系统可以组织大量的数据在云端，并且也能够有效地读取云数据，但我们为更好的结构化数据的管理，仍然需要专门的数据管理工具。

云数据管理是一个随云计算的发展产生的新的数据管理概念，它能够在云中的海量数据集有效地管理大型数据集，这使得快速找到特定的数据，以下常见的云数据管理的特点：

（1）高并发读写性能，

（2）有效地存储和访问大量的数据

（3）高可扩展性和高可用性要求

面对这些需求，传统的关系数据管理系统（RDBMS）遇到了难以逾越的障碍。因此，NoSQL数据库系统使用量大幅上涨。主要的互联网公司，如谷歌，亚马逊，Twitter和Facebook上有显着不同的挑战，在处理数据的RDBMS解决方案无法应对。这些公司意识到，性能和实时性比一致性更重要，而传统的关系数据库花费高额的处理时间来实现一致性。因此，NoSQL数据库通常高度优化检索和追加操作，而且往往以小功能超出记录存储。减少运行时的灵活性相比， RDBMS系统的显着收益补偿可扩展性和性能。通常情况下，NoSQL数据库根据他们的数据存储和下降的方式分类根据类别，如键 - 值存储（如迪纳摩[2]），BigTableimplementations[3] anddocumentstore数据库（例如MongoDB[4]）。然而，由于未成熟的技术云数据MANGEMENT，在实际生产环境中还是有很多需要解决的问题。

本文讨论MongoDB数据库的设计原则和，实现机制专注于自动分片的原则。自动分片的目标就是分裂整个机器的数据和实现重新平衡，使得能够存储更多的数据和处理而无需大或强大的机器。但平衡算法没有那么智能，数据是不是均匀均衡地分布在服务器之间。为了解决这个问题，提出了一个改善FODO（即频率的数据操作）算法。FODO算法是在数据操作的频率的基础上的，考虑服务器负载实现的。基于FODO算法的数据均衡策略能够有效地平衡服务器之间的数据，提高集群的并发读写性能。

二、 Mongodb中的自动分片机制