doi:10.1631/FITEE.1500015

题目：Dr. Hadoop: Hadoop 的一种无限可扩展元数据管理机制——小象如何不老？

目的：在这个“兆兆兆字节”（Exa byte）时代，数据量随时间指数率增长。剧增的数据在文件系统中制造了大量的元数据（metadata）。虽然 Hadoop 是处理大数据时最广泛采用的软件架构，其效率仍被研究者们广泛质疑。有必要为 Hadoop 创建一个有效且可扩展的元数据管理机制。

创新点：基于哈希的映射和子树分区适用于分布式元数据管理方案。基于哈希的映射在 NameNode（Hadoop 中存储元数据的服务器）间均衡地分配负载，但受到元数据空间局部性的限制；子树分区不需为保持负载均衡而迁移元数据，但也不能在服务器间均衡任务负载。本文提出一种称为 DCMS（dynamic circular metadata splitting，动态环形元数据分割）的环形元数据管理机制（图 3），并依此构建了 Hadoop 的改进框架——Dr. Hadoop（“Dr.”来自于本文作者名字首字母 Dipayan DEV, Ripon PATGIRI）。NameNode 是 Hadoop 的核心，其对所有文件路径树的保存失败将导致单点故障（single point of failure, SPoF）。DCMS 能够移除 Hadoop 中的单点故障，从而提供一种有效且可扩展的元数据管理机制。

方法：通过使用局部保持哈希（locality-preserving hashing，LpH）保持元数据的空间局部性，通过使用一致性哈希（consistent hashing）保持服务器间的负载均衡，通过保留复制后的元数据实现高可靠性。

结论：理论分析表明，Dr. Hadoop 架构在 99.99%的时间能够可靠使用。通过衡量数据吞吐率、容错性和 NameNode 负载等性能，DCMS 在大规模文件系统上较传统方法更具效力。

关键词：Hadoop；NameNode；元数据；局部保持哈希；一致性哈希