中国传媒大学 网络存储实验室 www.nstl.org.cn
COMMUNICATION UNIVERSITY OF CHINA
修身治学 技术报国

一种MongoDB集群数据布局优化方法研究

冯东煜[1,2] 朱立谷[1,2] 肖子达[1,2] 刘 迪[1,2]
[1] (中国传媒大学计算机学院 北京 100024)
[2](安防大数据处理与应用北京市重点实验室 北京 100094)
(fengdy1225@163.com)

传统关系型数据库在处理大规模数据应用时暴露出许多难以克服的问题，NoSQL以独有的特点在大数据背景下得到广泛应用。本文选择快递业寄递大数据应用为背景，研究MongoDB分片集群的数据布局优化方法。

快递运单数据库根据寄件日期按月份划分集合，将每一条运单记录以文档形式存储在基于MongoDB分片集群中，运单存储结构如图1所示。

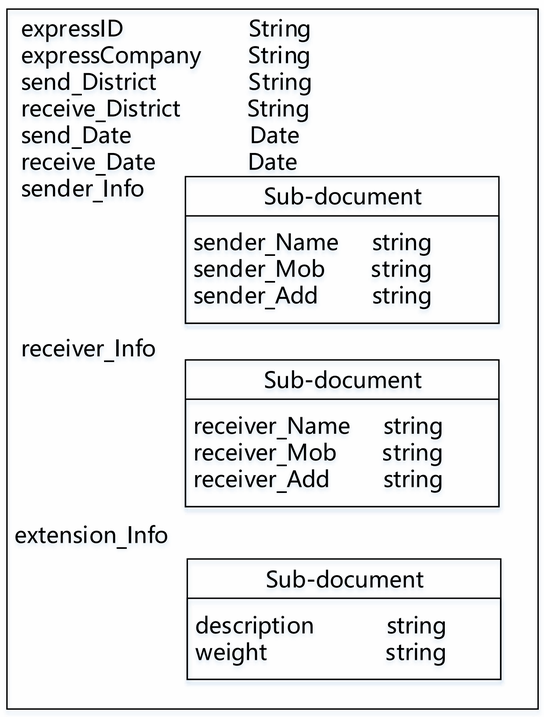

图1 运单存储结构

为对寄递用户信息进行整理，采用去重算法选择用户实体（姓名+手机号码）作为分组的唯一依据，利用MongoDB内建的聚合管道实现用户去重统计，用户整理流程如图2所示。

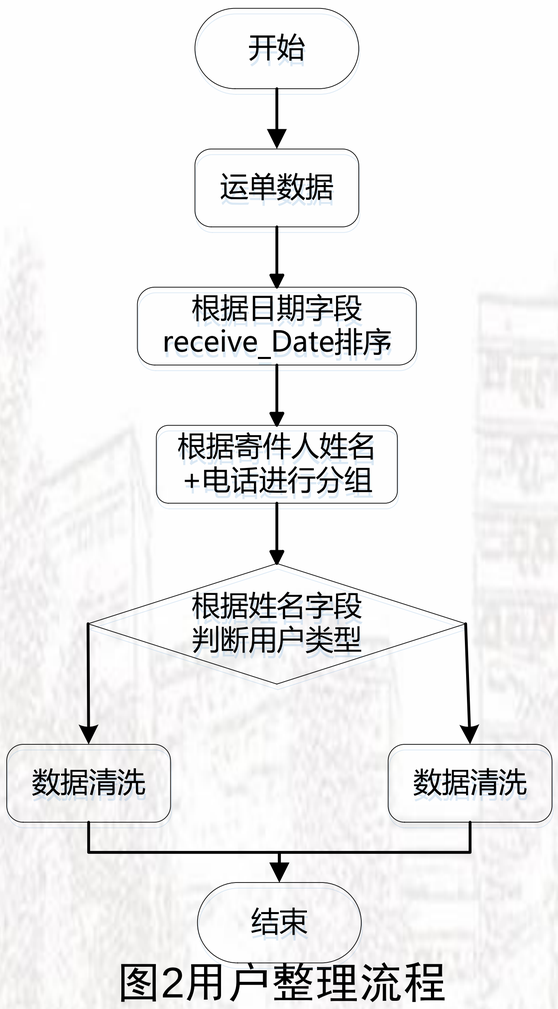

图2用户整理流程

然后，研究MongoDB片键策略，根据快递运单字段特点，设计“升序键+搜索键”形式的“寄件时间+用户姓名”复合片键，提出基于分片标签实现连续均匀数据条带化数据布局方法，然后对该方法对数据均匀分布和统计性能的提升进行测试。片键策略设计如图3所示。

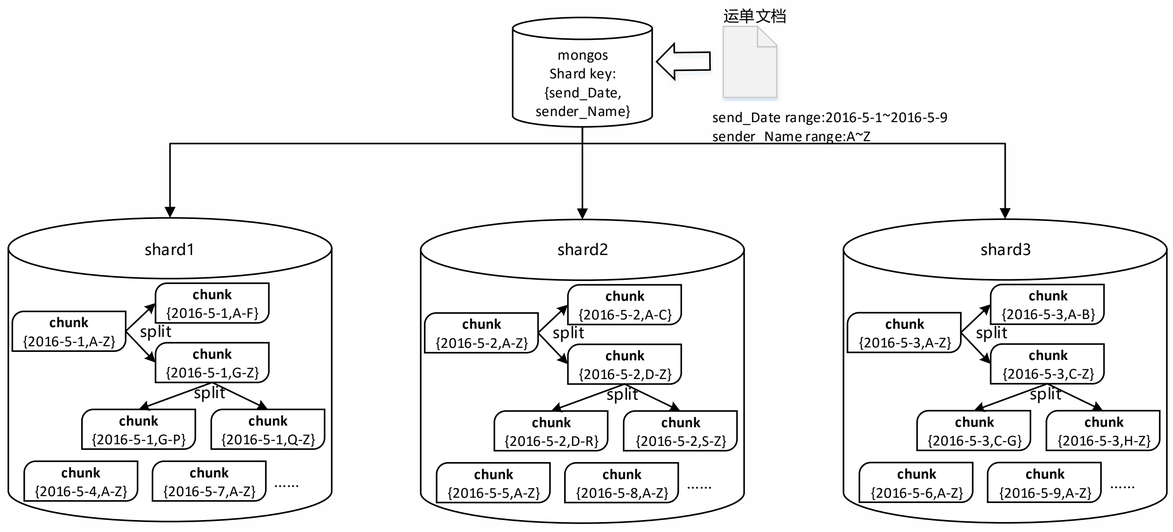

图3 基于分片标签的复合片键策略

在12节点的刀片服务器环境上部署MongoDB集群，模拟生成3亿条快递运单记录。本文提出的“寄件时间+用户姓名”片键策略在文档数量和数据块数量均匀分布情况明显优于其他片键组合。由于良好的数据分布使得MongoDB集群分布式处理的优势被充分发挥，该片键策略下离线分析系统的统计分析性能也优于其他片键组合，统计分析性能如图4所示。

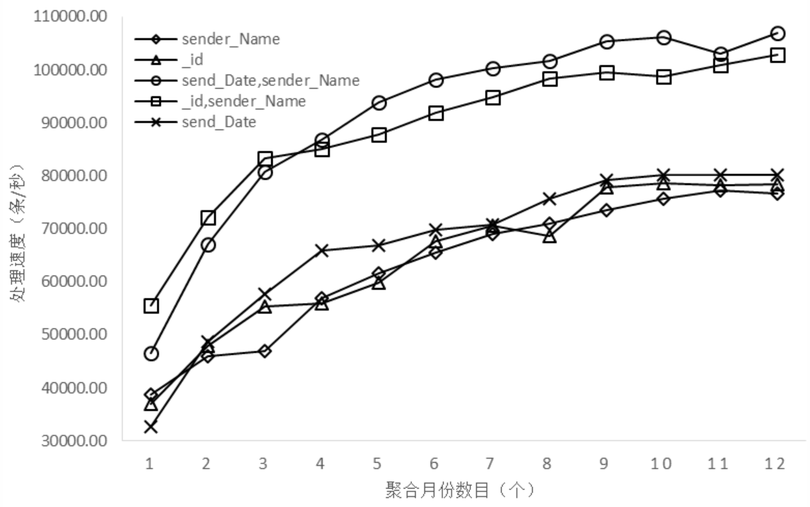

图4 统计性能分析测试结果

集群扩展性能测试结果如图5所示，验证了集群节点数目的扩展对系统整体统计分析性能的提升。

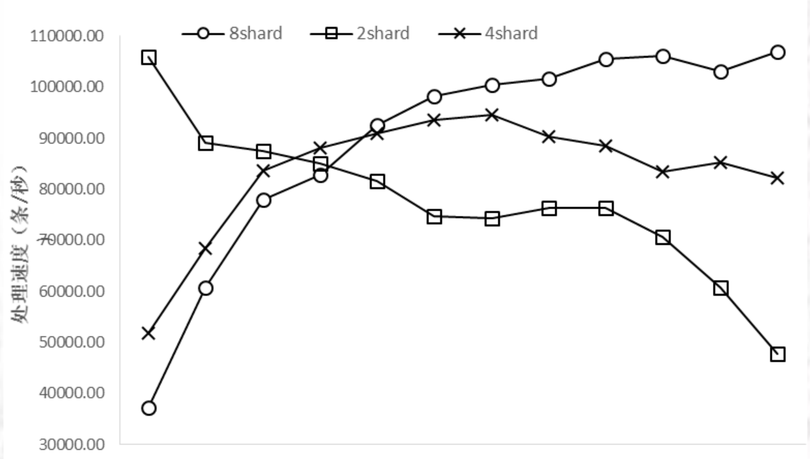

图5 分片扩展性能测试结果

因此本文提出的MongoDB集群数据布局优化方法，选择适合快递运单数据存储的片键策略，能够满足大规模数据存储的扩展性和均匀分布，并且基于分布式的MongoDB集群具有较强的统计分析性能。