大规模校园网日志的用户行为分析及画像建模

摘 要:传统的日志分析技术在处理海量数据时存在计算瓶颈。针对该问题，研究了基于大数据技术的日志分析方 案，即由多台计算机完成日志文件的存储、分析和挖掘工作;建立了一个网络日志分析系统:Flume完成日志采集，Hadoop 负责分布式存储和并行计算，在 MapReduce 框架下重新实现了 IP 统计算法。实验结果表明，在数据密集型计算中使用 大数据技术可以明显提高算法的执行效率和系统的可扩展性。

 1 引 言

经过多年的信息化建设，大型企业在内部网络中积累了 大量的软硬件资源，这些计算机及网络设备在运行过程中产 生了大量的日志数据。作为软件系统、硬件设备和用户行为 的记录工具，日志文件在监控网络情况、检查硬件故障、保护 软件安全等方面起着重要的作用。通过分析日志文件，能及 时发现用户异常行为和软硬件故障，保证网络运行的稳定性 和安全性。

网络日志的分析、挖掘技术普遍应用于信息安全领域，可 以 进 行 计 算 机 取 证 工 作 [1 ] 、发 现 异 常 的 网 络 访 问 [2 ] 、检 测 泛 洪 攻击[3]、进行防火墙的安全测评。当前，在大型局域网内部， 日 志 文 件 的 种 类 众 多 、格 式 不 一 、体 量 庞 大 (达 到 TB、PB 级 别)，传统的基于关系数据库的日志分析方法性能急剧下降， 无法满足海量数据的处理需求。在大数据领域，日志是广泛 使用的数据采集方法之一[4]，具有4V 特征:数据体量巨大 (Volume)、数 据 类 型 繁 多 (Variety)、有 价 值 但 密 度 低 (Va-

lue)、处 理 速 度 快 (Velocity)。 因 此 ，大 数 据 技 术 是 解 决 日 志 分析的更有效手段。本文提出基于 Hadoop框架的日志分析 引擎，在 MapReduce模型下重新实现了IP统计算法和异常 检测算法，实验证明，基于大数据的日志分析技术具有更高的

执行效率和良好的可扩展性。

2 大数据技术在日志分析中的应用

2 .1 大 数 据 处 理 技 术 与 工 具 根据应用类型的不同，大数据的处理模式可以分为流处

理和批处理两种，实时计算领域使用流处理技术，其他大部分 应 用 ( 如 数 据 挖 掘 、 推 荐 系 统 ) 都 依 赖 批 处 理 技 术 [4 ] 。 H a d o o p 是由 Apache开源的大数据处理框架，它是基于批处理技术 的，默认 MapReduce是其并行计算引擎，并由 HDFS[5]负责 数据存储。由于具备数据海量存储、数据并行处理及资源调 度 、负 载 平 衡 、容 错 处 理 等 底 层 管 理 功 能 ，Hadoop 极 大 地 降 低 了分布式程序开发的难度，得到了工业界的青睐，已经成为大 数 据 领 域 事 实 上 的 标 准 [6 ] 。

Hadop的核心是 MapReduce。MapReduce[7]依托于无共 享大规模集群系统，将计算工作分布到集群中的众多节点并行 运 行 ，它 的 计 算 依 靠 用 户 定 义 的 Map 函 数 和 Reduce 函 数 实 现 ， Map 函 数 负 责 分 块 数 据 的 处 理 ，Reduce 函 数 对 中 间 结 果 进 行 归 约 并 得 到 最 终 结 果 。 MapReduce 编 程 模 型 易 于 理 解 、易 于 使 用，结合 Hadoop平台合适的查询优化和索引技术，在大数据 环 境 下 ，仍 能 保 持 良 好 的 数 据 处 理 性 能 。 因 此 ，MapReduce 被广泛应用于海量数据的搜索、分析、挖掘和机器学习。