一种基于大数据的网络日志分析模型构建研究

邓小盾
（西安外事学院 陕西 西安 710077）

摘要：针对海量web日志数据在存储和计算方面存在的问题，结合当前的大数据技术，提出一种基于Hadoop与聚类分析的网络日志分析模型。利用Hadoop中的MapReduce编程模型对海量Web日志进行处理；利用HDFS结合的方式对数据进行存储；利用聚类分析算法构建web日志分析模型，对用户行为进行分析。最后通过搭建Hadoop测试环境对日志分析系统功能进行测试，并与单机系统比较，验证了该设计方案的优势。

关键词：大数据；web日志；MapReduce编程模型；HDFS；聚类分析
中图分类号：TN0 **文献标识码：**A **文章编号：**1674-6236(2017)23-0097-04

Research on the construction of a network log analysis model based on large data

DENG Xiao-dun
(*Xi'an International University*, *Xi'an* 710077, *China*)

Abstract: In view of the problems existing in the storage and computation of massive web log data, a new network log analysis model based on Hadoop and cluster analysis is proposed. To deal with the massive Web log using MapReduce programming model in Hadoop; using the combination of HDFS for data storage; model Web log analysis algorithm based on clustering, user behavior analysis. Finally, the function of the log analysis system is tested by building the Hadoop test environment, and the advantages of the design scheme are verified by comparing with the single machine system.

Key words: big data; web log; mapReduce programming model; HDFS; cluster analysis

DOI:10.14022/j.cnki.dzsjgc.2017.23.023

随着我国信息技术的不断发展，在各个企业、公共部门等结构内部网络中，积累了大量的软件和硬件资源，如交换机、路由器、防火墙PC服务器、Unix小型机、各类业务应用系统、中间件、数据库等。这些设备每天持续不断产生大量日志，并对这些日志进行记录。对于日志文件，作为不同软硬件资源运行中对故障问题和用户行为记录的一个重要工具，受到广泛的关注。通过日志可监控系统运行，查看不同硬件的故障，并保护系统的安全等，从而发现其中的异常行为，为及时处置网络安全事件提供参考。而随着大量日志的产生，日志的存储开始由原来的GB开始往TB或PB级别发展。同时，传统日志分析中采用的单机技术对海量数据进行处理，这给数据存储和分析带来很大的技术瓶颈。对此，针对这些问题，大数据技术开始被人们关注，并应用到对海量数据的处理中，典型的代表则是谷歌Hadoop平台下的Google MapReduce和GFS等存储工具。在Hadoop平台下，集成了对数据存储、数据处理和数据管理等全部功能，并成为了当前大数据应用的标准，特别是在对海量数据的搜索、挖掘和分析。

对此，本文针对传统日志分析的缺陷，从web日志分析角度出发，在Hadoop平台下，提出一种基于MapReduce编程模型和聚类分析的行为日志模型，并对其进行了详细的实现。

1 系统功能分析

本文提出的基于Hadoop平台下的web日志分析

收稿日期：2016-10-11 **稿件编号：**201610039
基金项目：2015—2016年度陕西省高教学会高水平民办大学建设研究项目(15GJ044)；2016年度西安市社会科学规划基金项目(16IN13)；2016年度陕西省教育厅科学研究项目(16JK2178)
作者简介：邓小盾(1979—)，女，陕西泾阳人，硕士，讲师。研究方向：大数据、人工智能。

系统主要包含以下几个功能模块：日志预处理模块、日志存储模块、日志挖掘模块。其中，日志预处理是数据挖掘的前提，对数据处理的好坏直接决定系统的运行；数据存储包含系统数据存储架构，是系统运行的保障；日志挖掘主要负责对预处理的文件进行挖掘分析，从而挖掘出有用的信息。具体功能见如图1所示。

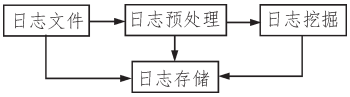

图1　系统功能需求分析

2　日志分析系统整体架构设计

针对图1所示的功能，设计出一款主要针对海量日志数据的用户行为系统，从而旨在改变传统单机技术在海量数据分析方面存在的局限，提高数据分析和挖掘的效率。架构具体见如图2所示。

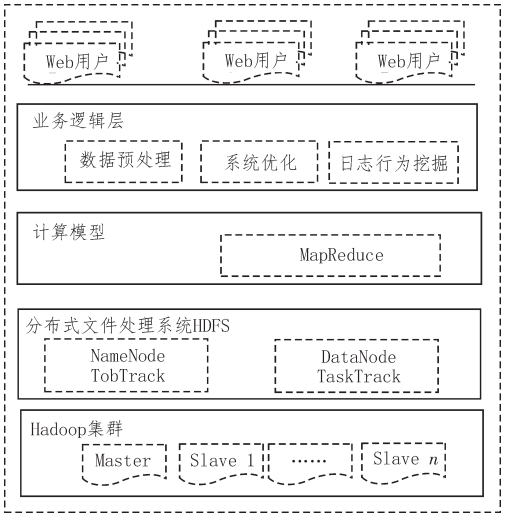

图2　系统整体架构设计

通过图2看出，Hadoop集群层是整个日志行为分析系统的基础，为整个提供软硬件和网络支撑。在该层中由1个主服务器和N个从服务器组成；HDFS分布式文件处理系统主要负责对数据进行存储，并提供与MapReduce层调用的接口；MapReduce计算层主要负责对海量数据采用映射/规约等方式进行并行运算，从而将任务分配给下属的各个节点，在通过处理后，将信息整合进行整合得到最终结果；业务逻辑层主要利用Hadoop的集群环境对日志数据进行存储，并运用MapReduce的计算来分析日志用户的行为偏好。

3　系统详细设计

3.1　数据预处理

数据预处理作为该系统的一个重要模块，是数据挖掘的基础。本系统则将数据直接存入到HDFS文件系统中，默认切分为每块为64 MB大小的文件，保存在各个DataNode上。而为了加快对原始数据的清理和识别，引入MapReduce编程模型，其具体原理如图3所示。

1）Map阶段

在Map阶段，主要输入web日志中的每条记录，输出格式为<key，value>键值对，其中key表示用户ID，value表示<链接地址、上一跳链接地址、时间>。对于Map阶段来看，其主要实现对数据的清理。具体步骤为：

①对web日志进行切分处理，并分割不同数据的属性；

②判断数据请求方式是否为“GET”，如不是，删除该数据，如是，继续处理；

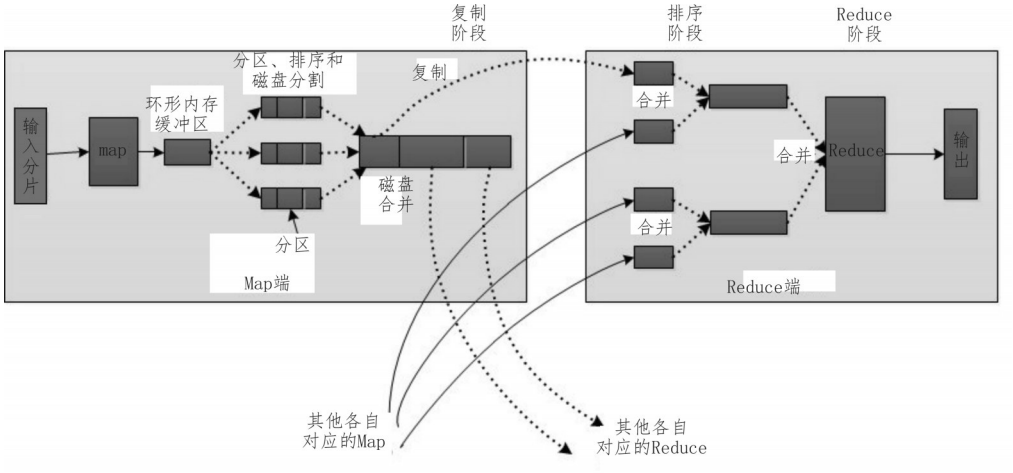

图3　MapReduce工作原理

③判断文件格式是否在要求的文件格式中，如没有，删除；

④判断该日志返回的标识码是否在200~299之间，如不在，删除，如在则按照标准格式输出。

2）Reduce阶段

Reduce阶段主要完成对对应的Map处理数据的合并。在该阶段引入Combiner函数。具体步骤为：

①使用迭代器遍历每个记录的value，取出其中的ip，并将其放入到USIt容器中。

②遍历USIt容器，判断ip是否相同，如不同则视作新用户。

3.2　数据存储设计

对存储模块的设计则采用HDFS文件处理系统与Mysql组合的方式，其中HDFS主要负责存储全部数据，Mysql负责存储从HDFS系统中导出的数据，从而为数据挖掘奠定基础。为了将HDFS系统中的文件导入到Mysql中，引入sqoop工具，从而将HDFS中的数导入到Mysql中。具体实现架构如图4所示。

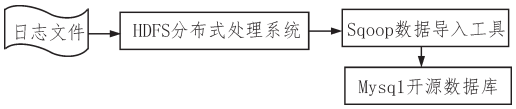

图4　存储架构

3.3　挖掘模块设计

在对大数据的挖掘中，常用的方法包括关联规则、回归分析、聚类等，其中K-means聚类是最为简单，也是最为高效的算法。其核心思想是以样本空间中的个点作为中心点，通过迭代，对其周围的样本进行归类并重新计算中心点的值，直到收敛。虽然该方法简单，但是在该算法中最为关键的K值往往是根据经验来确定。为解决该问题，通常会通过反复的实验，从而取一个较好的聚类K值。对此本文结合以往的研究，提出一种最小最大距离算法对K值进行确定。该方法的核心思想是在对web日志进行聚类的过程中，选取原点作为初始点的参照点，选择一个最远和最近的点作为初始点，计算数据集合中点到这两点的距离。将K-means放到MapReduce中，具体算法则可以描述为如图5所示。

4　实验验证

搭建基于Hadoop的平台，总共5台计算机，其中一台为主服务器，其他的4台为从服务器。安装Java6版本，Mysql数据库采用5.5版本，Hadoop采用

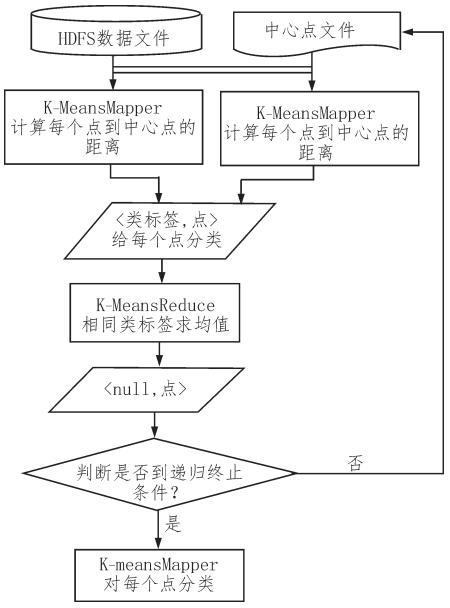

图5　基于K-means的MapReduce并行化处理

2.3版本。

通过上述环境的搭建，可以得到如图6所示的数据处理对比结果。

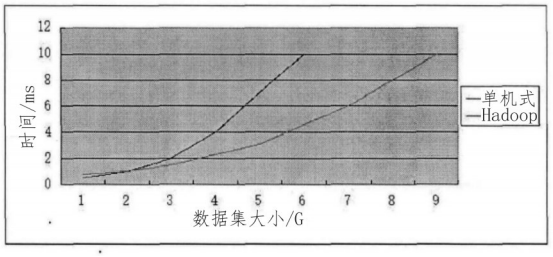

图6　分布式和单机式数据处理结果对比图

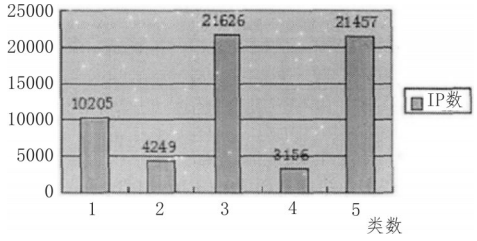

图7　日志聚类结果

通过图7可以看出，采用基于Hadoop的集群其数据处理的大小要远远大于传统的单机处理方式。同时通过引入K-means算法对web文本日志进行分析，得到总共5类不同的活跃用户，其中每一类的数据都是通过IP解析得到，并得到第三类、第五类的访问最多，最值得我们关注。

5　结束语

针对传统单机处理基础存在的弊端，引入基于

Hadoop的平台，从而大大提提高了数据处理的效率，同时引入K-means算法对日志进行挖掘分析，取得良好的效果，可有效地根据用户IP得到其行为特征，从而为大数据技术的进一步利用提供了参考，也奠定了未来应用的基础。

参考文献：

[1] 罂岩龙，罗壮，杨說，等. 基于hadoop的高性能海量数据处理平台研究[J].计算机科学，2013，40(3):100-103.

[2] 周诗慧，殷建. Hadoop平台下的并行Web日志挖掘算法[J]. 计算机工程，2013(6):43-46.

[3] 马汉达，郝晓宇，马仁庆. 基于Hadoop的并行PSO-kmeans算法实现Web日志挖掘[J]. 计算机科学，2015(S1):470-473.

[4] 任凯，邓武，俞琰. 基于大数据技术的网络日志分析系统研究[J]. 现代电子技术，2016(2):39-41，44.

[5] 于兆良，张文涛，葛慧，等. 基于Hadoop平台的日志分析模型[J]. 计算机工程与设计，2016(2):338-344，428.

[6] 张春生，郭长杰，尹兆涛. 基于大数据技术的IT基础设施日志分析系统设计与实现[J]. 微型电脑应用，2016(6):49-52.

[7] 陈洁，于永刚，刘明恒，等. 安全管理平台中基于云计算的日志分析系统设计[J]. 计算机工程，2015(2):21-25.

[8] 江小平，李成华，向文，等. k-means聚类算法的MapReduce并行化实现[J]. 华中科技大学学报：自然科学版，2011(S1):120-124.

[9] 周婷，张君瑛，罗成. 基于Hadoop的K-means聚类算法的实现[J]. 计算机技术与发展，2013(7):18-21.

[10]李洪成，吴晓平，陈燕. MapReduce框架下支持差分隐私保护的k-means聚类方法[J]. 通信学报，2016(2):124-130.

[11]李欢，刘锋，朱二周. 基于改进K-means算法的海量数据分析技术研究[J]. 微电子学与计算机，2016(5):52-57.

[12]杨勇，任淑霞，冉娟，李春青. 基于粒子群优化的k-means改进算法实现Web日志挖掘[J]. 计算机应用，2016(S1):29-32，36.

[13]衣治安，王月. 基于MapReduce的K_means并行算法及改进[J]. 计算机系统应用，2015(6):188-192.

[14]谢雪莲，李兰友. 基于云计算的并行K-means聚类算法研究[J]. 计算机测量与控制，2014(5):1510-1512.

[15]何佩佩，谢颖华. 云环境下K-means算法的并行化[J]. 微型机与应用，2015(24):25-27，31.

（上接第96页）

[11]杨帆. 基于FPGA的SDI接口的研究与开发[D].天津：天津理工大学，2010.

[12]黄隶凡，郑学仁. 基于FPGA的三速SDI设计[J].电视技术，2011，35(3)：13-18.

[13]岳元，彭量节. SDI红外图像在火炮光电跟踪系统中的应用[J].激光与红外2016，46(8)：1024-1027.

[14]杨洋. 基于FPGA的4路HD-SDI光纤传输系统[J].光通信技术，2015，39(5)：97-103.

[15]苏建，林水生. 基于FPGA的SDI接口设计[J].中国有线电视，2015(24)：62-68.

[16]李彦迪，金伟正，王丹.基于FPGA的HD-SDI编解码技术的研究与开发[J].电子技术应用，2012(12)：16-21.

[17]党俊博，李 哲，李雅俊.基于FPGA的串口通信电路设计与实现[J].电子科技，2016(7):106-109.

[18]罗帅，徐进，夏杰，等.一种光纤数据采集系统的设计[J].西安工程大学学报，2016(3)：312-315.

[19]严明，李斌康，郭明安，等.高速光电探测器阵列实时信号处理系统[J].现代应用物理，2014(4)：316-321.