文章编号：2096-1472(2019)-02-11-02
DOI:10.19644/j.cnki.issn2096-1472.2019.02.003

基于Hadoop的Web日志分析系统的设计

何 璇，马佳琳
(沈阳师范大学软件学院，辽宁 沈阳 110000)

摘 要：在大数据时代，数据成为推动各个行业发展的动力，有效的分析数据不仅对社会经济效应有巨大影响，而且对政府，企业的管理也有深远影响。于是，怎样高效且快速地从Web日志中挖掘出有用的价值并且转化为分析依据是系统设计的重点。本文主要采用Hadoop为开源框架，利用HDFS进行数据的存储，Hive为开源数据仓库工具，设计并实现一个Web日志分析系统。文章主要阐述了系统的结构、设计思想和实现方法。

关键词：Hadoop；Web；Hive
中图分类号：TP399 **文献标识码**：A

The Design of the Web Log Analysis System Based on Hadoop

HE Xuan,MA Jialin
(*Software Institute*,*Shenyang Normal University*,*Shenyang* 110000,*China*)

Abstract: In the era of big data,data has become a driving force for the development of various industries.Effective analysis of data not only has a huge impact on social and economic effects,but also has a profound impact on the management of governments and enterprises.Therefore,how to efficiently and quickly extract useful value from Web logs and turn it into analysis basis is the key point of system design.In this paper,by means of adopting Hadoop as the open source framework,HDFS for data storage and Hive as the open source data warehouse tool,a Web log analysis system is designed and implemented.This article mainly elaborates the system structure,the design thought as well as the realization method.

Keywords: Hadoop;Web;Hive

1 引言(Introduction)

随着互联网技术的迅速发展，每天在Web服务器上都会产生大量的访问日志，如何挖掘出销售量较好的商品，网站最受欢迎的版块，网站点击量最高的广告，并提供针对性的产品与服务，是传统的数据解决方案和方法所不能企及的。本文打破传统的数据处理方式，使海量数据的处理变得更加高效[1]。利用Hadoop技术的开源性与并行处理的高效性，利用普通的计算机就能搭建出性能优越的集群，充分利用各个计算机节点的资源，成本低廉，技术的成熟稳定，使批量数据得到及时的处理，有效提高工作效率。

2 主要相关技术(Main correlation technique)

2.1 Hadoop介绍

Hadoop是Apache软件基金会下的开源式分布框架，主要应用于大规模数据的处理[2,3]，其主要核心项目是HDFS与Map/Reduce。HDFS主要利用分布式存储对大规模数据进行管理[4]，Map/Reduce主要对分布式文件系统上的数据进行整合，保证分析与数据处理的高效性。

HDFS主要支持流数据读取和处理超大规模文件，结构模型为主从结构(Master/Slave)，一个HDFS集群由一个名称结点和若干个数据结点组成。名称结点作为中心服务器，主要工作是管理整个文件系统。数据节点作为工作结点，主要任务为处理文件系统客户端的读/写请求，并向名称结点传输存储信息。实现原理是当如果有文件提交给Master节点，Master会将其切分成N个块，并为每个块拷贝多个副本，然后将这些块分散地存储在不同的Slave节点上。Master负责维护整个NameSpace，NameSpace中记录着输入文件的切割情况，以及每个块的存储信息。Slave是实际的工作站，负责存储块。

Mapreduce主要处理存储在分布式文件系统中的数据，它的核心是Map函数与Reduce函数，输入输出方式都以中间键/值的形式。在Map函数中按照一定的映射规则转换为中间键/值对集合，通过Reduce函数将临时形成的中间键/值对集

合进行处理和输出结果[5]。

2.2 Hive介绍

Hive是Hadoop下的数据仓库工具，将结构化的数据文件进行整理，特殊查询和分析存储[6]。Hive提供了类似于关系数据库SQL语言的查询语言—HIVE QL，先对数据进行管理，通过解析和编译，最终生成基于Hadoop的Mapreduce任务，最后，Hadoop通过执行这些任务完成查询任务和数据处理。

3 系统框架的设计(System framework design)

本文阐述Web日志分析系统的设计方案，以某搜索引擎网站日志为分析对象，利用HDFS为存储平台，Map/Reduce对原始数据进行清洗，利用Hive对数据进行统计分析，通过Sqoop把统计分析后结果导出到MYSQL，最后利用Web网页进行数据展示，如图1所示。

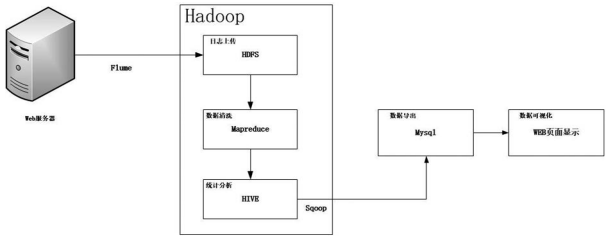

图1 系统框架设计图

Fig.1 System framework design

4 主要功能模块设计(Main function module design)

4.1 日志上传模块

(1)在集群上搭建所需要的数据框架，比如HBASE，首先启动Hadoop分布式集群，然后启动Zookeeper集群，最后在Master(元数据结点)上启动HBASE集群。

(2)在以上四种结点的Linux系统的根目录下创建日志文件夹(Apache_Logs)，用于存放日志文件执行命令，启动集群。

(3)在HDFS文件系统中的HDFS根目录下创建Web_Logs(网页日志)文件夹，通过日志收集模块(Flume)与集群通过RPC(远程过程调用协议)通信交互，让日志收集任务让后台程序进行，监控Apache-Logs文件夹，一旦文件夹收集到日志文件，就同步到HDFS中Web—Logs文件夹下。

4.2 数据清洗模块

根据需求，对日志文件进行数据清洗处理，删除与挖掘出与任务不相关的数据并且合并某些记录，对HDFS中的日志文件数据进行清洗转换，清洗转换后的数据放进HDFS中[7,8]，数据清洗的内容为检查数据的一致性，处理无效值和缺失值等。过滤掉不符合要求的数据，如不完整的数据，错误的数据，重复的数据三种。数据清洗完毕后，可以通过网页的形式在浏览器端访问查看文件系统，查看到所需要数据。

4.3 统计分析模块

当数据清洗完成之后，利用Hive构建日志数据的数据仓库，对数据进行统计分析，首先创建外部表，以及内容更为详细的分区表，最后实现数据的分析需求。

实验利用搜索网站的日志实现五大分析需求，条数统计、关键词分析、UID分析、用户行为分析，以及实时数据分析。

4.4 数据导出模块

将得到的各个统计量分别存放到相应的表中。然后把各个表中的数据汇聚到一张表中。使用数据导出模块(Sqoop)把汇总的数据导出到外面的关系型数据库Mysql中，也可以使用HBASE实现数据的快速查询。

4.5 数据可视化模

在获得分析结果后，使用JSP技术和Struts2框架，完成不同统计分析业务页面设计，为用户提供查询结果的页面展示窗口，页面的功能模块设计可以根据不同的业务需求，进行个性化的业务拓展[9]。

5 实验结果与分析(Experimental results and analysis)

5.1 实验环境

由四台计算机搭建Hadoop集群，一台为Master结点，另外三台为Slave结点，实验数据基于某搜索网站的搜索日志为基础数据。

5.2 实验结果

为了验证Hadoop对于日志分析处理的高效性，在单机和HDFS集群进行了实验对比。在实验中对于不同文件大小的Web日志分别进行处理，并计算执行的时间，结果如图2所示。

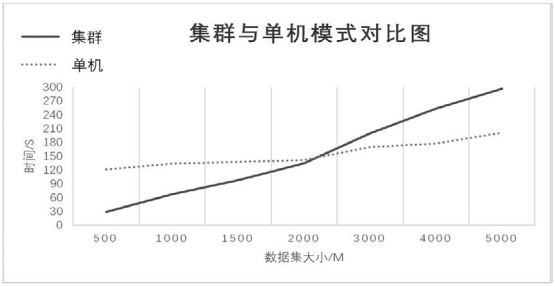

图2 集群与单机模式对比图

Fig.2 Comparison of the cluster mode with the stand-alone mode

5.3 实验分析

图2给出了相同数据集在Hadoop分布式平台和单机式平台处理时间的比较。由图可以看出当数据集较小时单机式Web日志分析处理的效率更高，但随着数据集的不断增大Hadoop平台逐渐表现出了优势。这是由于，在Hadoop平台下每次迭代都需要重新启动一个MapReduce任务，所以在数据集较小的情况下，系统启动MapReduce任务所消耗的时间占用的比例较大，从而导致了计算效率的下降。而在大数据集的情况下系统启动MapReduce任务所消耗的时间可以忽略不计，因此这时的计算效率较高。

(下转第4页)