文章编号：2095-2163(2019)06-0201-05 中图分类号：TP393.06 文献标志码：A

基于 Spark Sreaming 网站流量实时分析系统的设计与实现

刘 珍，方 明
（西安石油大学 计算机学院，西安 710065）

摘 要：针对互联网技术快速发展，用户对各种网站访问量急剧加大，日志数据急剧增加的现状，采用 Hbase 数据库，Flume、Kafka 分布式发布订阅消息系统和 Spark Streaming 流计算框架，设计实现基于 Spark Sreaming 的网站流量实时分析系统，深入探讨了网站流量的分析角度和指标，展示了网站的运营情况，从而引导网站开发、运营人员作出相关决策来改进网站的服务，为网站维护、制定网站营销策略提供有力的依据。

关键词：Spark Streaming；网站流量分析；HBase；Kafka

Design and implementation of Website traffic real-time analysis system based on Spark Sreaming

LIU Zhen, FANG Ming
(School of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 In response to the rapid development of Internet technology, users have greatly increased the number of visits to various websites and the rapid increase of log data. The Hbase database, Flume, Kafka distributed publish and subscribe message system and Spark Streaming flow computing framework are designed and implemented based on Spark Sreaming. Website traffic real-time analysis system, in-depth discussion of the analysis angle and indicators of website traffic analysis, showing the operation of the website, thereby guiding the website development, the operators make relevant decisions to improve the website´s services, and provide website marketing strategies for website maintenance. A strong basis.

【Key words】 Spark Streaming; Website traffic analysis; HBase; Kafka

0 引 言

随着互联网技术的发展，用户对各类网站的访问量急剧加大，导致日志数据急速增加，数据类型也纷繁复杂。因此日志数据的产生、规模、存储、处理方式也悄然发生变化。大数据时代，网站运营管理方应及时地对网站流量和用户访问情况进行统计分析，通过数据来分析用户的浏览习惯，可对优化网站运营架构、调整推广策略起到积极的作用[1]。网站流量统计是改进网站服务的重要手段之一，通过获取用户在网站的行为，对有关数据进行统计、分析，从而发现用户访问网站的规律[2]。通过对网站进行流量分析，可以刻画出网站近期的运营情况，从而引导网站开发、运营人员作出相关决策来改进网站的服务，为网站维护、制定网站营销策略提供有力的依据，促进网站整体的改进[3]。

本文采用大数据的理论和方法，采用 Hbase 数据库[4]、Flume、Kafka 分布式发布订阅消息系统和 Spark Streaming 流计算框架，设计实现了基于 Spark Sreaming 的网站流量实时分析系统。

1 基于 Spark Streaming 网站流量实时分析系统的分析维度和指标

目前，常用的网站流量统计指标一般包括以下情况分析：

(1)在线情况。在线情况分别记录了在线用户的活动信息，包括：来访时间、访客地域路页面、当前停留页面等，这些功能对企业实时掌握自身网站流量有很大地帮助。

(2)时段分析。时段提供网站任意时间内的流量变化情况。或某一时间段的流量变化。如小时段分布，日访问量分布、对于企业了解用户浏览网页的时间段有一个很好地分析。

(3)来源分析。来源提供来路域名带来的来访次数、IP、独立访客、新访客、新访客浏览次数、站内总浏览次数等数据。这些数据可以直接让企业了解推广成效的来路，从而分析出哪些网站投放的广告效果更明显。

系统统计的指标说明：

(1)PV：Page View 页面访问量。本项目以天

作者简介：刘 珍(1992-)，女，硕士研究生，主要研究方向：管理信息系统；方 明(1963-)，男，教授，主要研究方向：管理信息系统。

收稿日期：2019-08-27

为单位,统计一天内总的 PV。用户访问一次网页,就算一次 PV,刷新操作也算 PV。

(2)UV:独立访客数。按人头算,统计一天内有多少不同的用户。处理思路:为每个用户生成一个 uvid,然后存到用户浏览器的 cookie 里,所以统计独立用户数=统计有多少不同的 uvid。

(3)VV:独立会话数。关闭浏览器再打开浏览器算做一个新的会话。实现思路:当用户通过浏览器访问产生一个新会话时,服务端会为这个会话生成一个 ssid。所以独立会话数=不同的 ssid 个数。此外,当一个会话超过 30min,再次访问,会算作一个新会话。

(4)BR:页面跳出率= 跳出会话数/总的独立会话数 。这个指标用于衡量网站优良性的高低。调出率越低,说明网站对于用户的粘度越大。

(5)newCust:新增用户数。新增用户指的是用户的 uvid 在历史 uvid 没有出现过。比如统计今天的 newCust 数:

①统计出今天的所有的 uvid;

②和之前的数据做比对;

③取出历史数据没有出现的 uvid。

(6)newIp:新增 Ip 数。统计一天内,有哪些 ip 是在历史数据中没出现过。

(7)avgDeep:平均的会话访问深度。一个会话的访问深度=一个会话浏览过哪些不同的 url 地址。

(8)avgTime:平均的会话访问时长。

2 基于 Spark Streaming 的网站流量实时分析系统总体结构

基于 Spark Streaming 的网站流量实时分析系统采用了 Flume、Kafka、SparkStreaming、Hbase、MySQL、Echars 等技术,系统总体结构如图 1 所示。

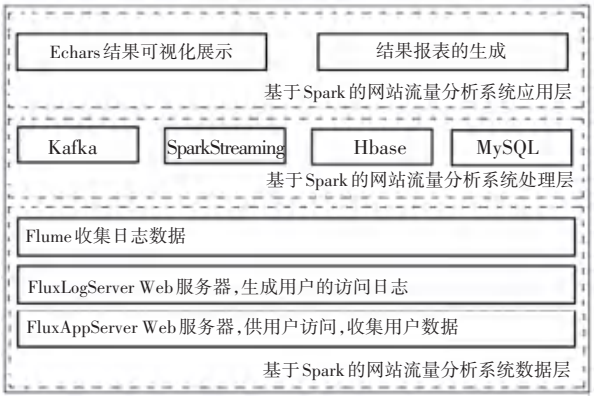

图 1 基于 Spark 的网站流量实时分析系统的总体结构

Fig. 1 Overall structure of the Spark-based Website traffic real-time analysis system

本系统分为日志收集模块、实时数据分析模块和结果展示模块。其中实时数据分析模块又划分为数据采集子模块、数据接入子模块、流式计算子模块、数据输出子模块、结果子模块展示。系统模块如图 2 所示。

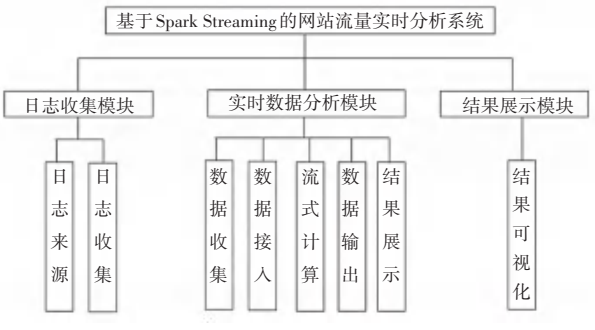

图 2 系统模块

Fig. 2 System module diagram

2.1 日志收集模块

通过 Js 埋点获取网页上的信息作为日志信息,通过反向代理技术 Ngnix 将日志发送到日志服务器。Flume 集群利用 Agent 将日志服务器中日志信息扇入到 Flume 中,而后 Flume 将日志信息通过 Agent 将日志信息扇出到 Kafka,为 Spark Streaming 实时分析提供日志信息。Flume 发送日志信息结构如图 3 所示。

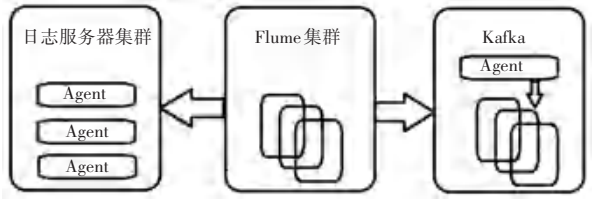

图 3 Flume 发送日志信息结构

Fig. 3 Flume send log information structure

2.2 实时数据分析模块

本模块主要部分:数据采集、数据接入、流式计算、数据输出、结果展示。

(1)数据采集。负责从各个节点进行实时采集数据,选用 cloudera 的 Flume 实现。

(2)数据接入。因为采集数据与数据处理的速度不一定是同步的,由此需要添加一个中间件作为缓冲,这里选用的是 apache 的 kafka。

(3)流式计算。对采集到的数据进行实时分析,选用 Spark Streaming。

(4)数据输出。采用 Hbase 对分析后的结果进行持久化。

(5)结果展示。采用 MySQL 和 Echars 进行前段结果展示。

2.3　结果展示模块

本文利用 MVC 框架实现数据可视化的数据展示,分为数据层、服务层和 Web 层。由 JSP+Echars+Servlet+JavaBean+Dao 构成 MVC 模式:JSP+Echars 模块化单文件引入,组中将分析结果展示到页面上。Servlet 用于验证数据、实例化 JavaBean、调用 DAO 连接数据库、控制页面跳转。DAO 用于连接数据库及进行数据库的操作。JavaBean 用于数据的封装,方便将查询结果在 Servlet 与 JSP 页面之间进行传递等。以上部分共同构成了 MVC 模式,数据可视化框架如图 4 所示。

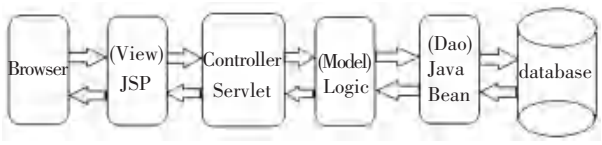

图 4　数据可视化框架

Fig. 4　Data visualization framework

3　基于 Spark Streaming 的网站流量实时分析系统实现框架

系统采用 Hbase 数据库,Flume、Kafka 分布式发布订阅消息系统、Spark Streaming 流计算框架、Echars 结果可视化插件。网站流量实时分析系统的具体实现框架如图 5 所示。

图 5　基于 Spark Streaming 的网站流量分析系统的实现框架

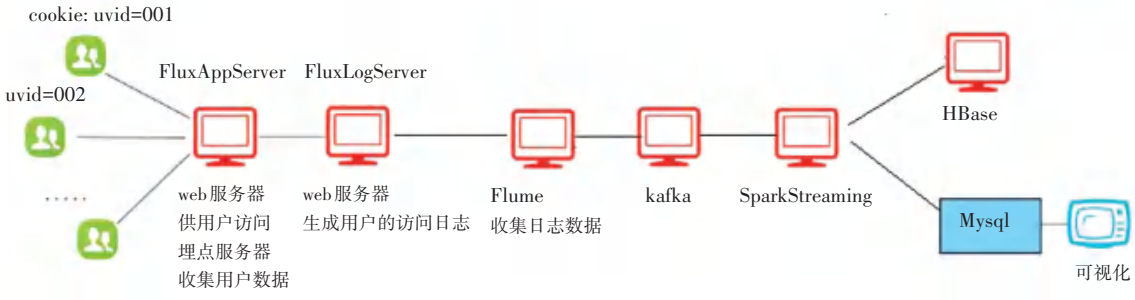

Fig. 5　Implementation framework of Website traffic analysis system based on Spark Streaming

3.1　Flume 简述

Flume 是一个分布式、可靠、高可用的海量日志采集、聚合和传输的系统。在日志系统中定制各类数据发送方,用于收集数据(source)。Flume 提供对数据进行简单处理,并写到各种数据接受方(可定制)的能力(sink)。Flume 的总体架构如图 6 所示。

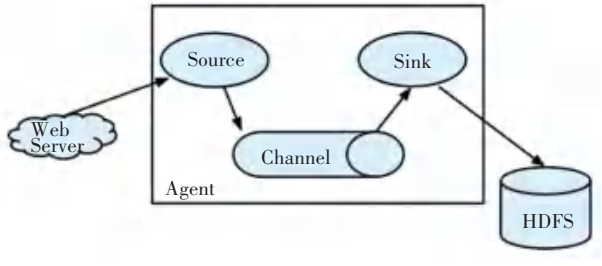

图 6　Flume 的总体架构

Fig. 6　Overall architecture of Flume

Flume 运行的核心是 agent,其本身是一个 Java 进程,里面包含 3 个核心组件:source、channel、sink,类似生产者、仓库、消费者的架构。source 专门用来收集数据,可以处理各种类型、各种格式的日志数据。如 avro、thrift、exec、jms、spooling directory、netcat、sequence generator、syslog、http、legacy、自定义等;channel 把数据收集后,临时存放在 channel 中,即 channel 组件在 agent 中是专门用来存放临时数据的—对采集到的数据进行简单的缓存,可以存放在 memory、jdbc、file 等等;sink 是用于把数据发送到目的地的组件。目的地包括 hdfs、logger、avro、thrift、ipc、file、null、hbase、solr、自定义。完整的工作流程为:source 不断地接收数据,将数据封装成一个一个的 event,然后将 event 发送给 channel,chanel 作为一个缓冲区会临时存放这些 event 数据,随后 sink 会将 channel 中的 event 数据发送到指定的地方。

3.2　Kafka 简述

Kafka 是一个分布式的流式处理平台,主要包含 3 个功能:

(1)发布和订阅数据,类似于消息队列或者企业中的消息传递系统。

(2)存储数据时有容错(分布式+复本机制)和持久化机制。

(3)数据产生时处理记录(数据)。

Kafka 使用 Scala 编写,以可水平扩展和高吞吐率而被广泛使用。目前越来越多的开源分布式处理系统如 Cloudera、Apache Storm、Spark 都支持与 Kafka 集成。Kafka 之间传输数据使用零拷贝技术。

3.3 Spark Streaming 简述

Spark Streaming 是 Spark 的流式处理框架,是面向海量数据实现高吞吐量、高可用的分布式实时计算。Spark Streaming 并非像 Storm 那样是真正的流式计算,二者的处理模型在根本上有很大不同:Storm 每次处理一条消息,而 spark streaming 每次处理的是一个时间窗口的数据流,类似于在一个短暂的时间间隔里处理一批数据。其数据处理框架如图 7 所示。

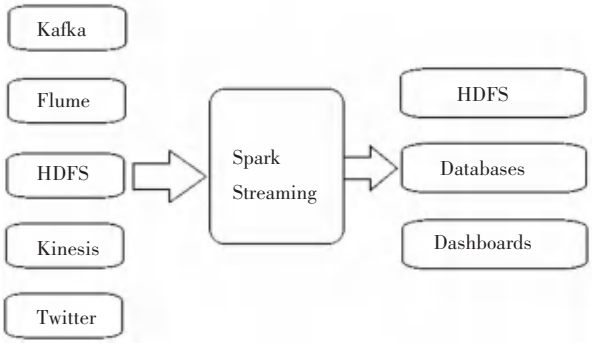

图 7 Spark Streaming 数据处理架构

Fig. 7 Spark Streaming data processing architecture

Spark Streaming 内部工作原理如图 8 所示,其接收实时输入数据流并将数据切分成 batch(批)数据,由 Spark 引擎处理以生成最终的分批流结果。

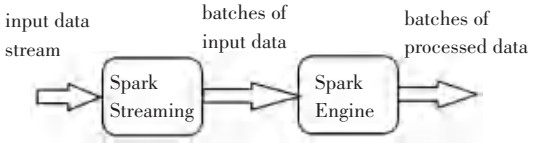

图 8 Spark Streaming 工作原理

Fig. 8 How Spark Streaming works

3.4 Hbase 简述

Hbase 是一个分布式开源数据库,基于 Hadoop 分布式文件系统,其原型是 Google 的 BigTable 分布式数据库。Hbase 的设计目标是处理非常庞大的表,可以使用普通计算机处理超过 10 亿行数据,并且有百万列元素组成的数据表。因此在文件的百万行或者上千万行时不需要使用 Hbase[5]。HDFS 为 Hbase 提供了高可用的底层存储支持,同时 MapReduce 为其提供了高可用性的计算能力。Zookeeper 保证了分布式数据库的一致性要求,Hive、Pig 提供了操作数据库的语言,Sqoop 提供了传统关系数据库的导入功能。Hbase 与 Hadoop 无缝连接有以下几个显著的优点:廉价的节点、高可用性、可扩展性[6]。Hbase 由于流式存储的特性,也存在相应的缺点:Hbase 的实时性差、不善于处理即时业务[7]。

3.5 Echars 简述

Echars(Enterprise Charts)是一个商业级的数据图表,一个纯 JavaScript 的图标库,只是其能够流畅地在 PC 端和移动设备之上运行,兼容当前绝大多数的浏览器。Echars 在底层依赖轻量级的 Canvas 类库 ZRender,能够支持直观的、可交互的、生动的、可以高度个性化定制的数据可视化图表。极大地增强了用户体验的创新特性、有拖拽重计算、数据视图、值域漫游等,同时也赋予了用户对数据进行挖掘及整合的能力。

Echars 支持折线图、柱状图(条状图)、区域图、K 线图、散点图(气泡图)、雷达图(填充雷达图)、和弦图、力导向布局图、地图、饼图(环形图)等 12 类图表,还提供了标题、图例、时间轴、详情气泡等 7 个可交互组件,同时还支持多图表、组件的联动以及混搭展现。

4 基于 Spark Streaming 的网站流量实时分析系统的实现

本系统利用分析的指标:信息浏览量(PV)、独立访客(UV)、会话数(VV)、新增独立 IP(Newip)、新增访客(Newcust)、跳出率(Br)、平均访问深度(Avgdeep)进行在线情况分析、在线时段分析、访客来源分析。

4.1 在线情况分析

在线情况分析分别记录在线用户的活动信息,包括:来访时间、访客地域路页面、当前停留页面等,这些功能对企业实时掌握自身网站流量有很大的帮助。图 9 通过对一天中不同来访时间的访客数进行统计,展示了用户在一天中登录浏览的时间分布。

图 9 用户日在线时间分布图

Fig. 9 User day online time distribution

4.2 在线时段分析

时段分析提供网站任意时间内的流量变化情况,或者某一段时间到某一段时间的流量变化。如

小时段分布、日访问量分布，对于企业了解用户浏览网页的时间段有一个很好地分析。图 10 对用户在一周内的访问次数做以统计，展示了工作日和休息日对用户访问情况的影响。

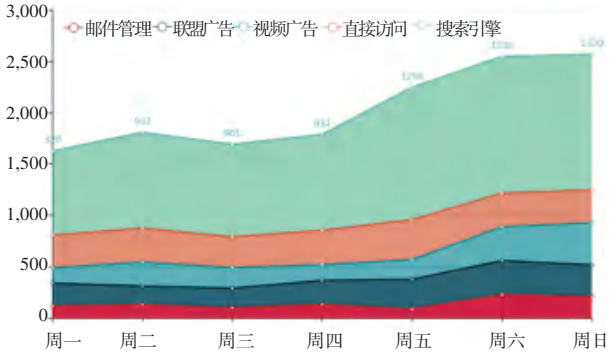

图 10　用户周在线时间分布图

Fig. 10　User week online time distribution

4.3　访客来源分析

来源分析提供来路域名带来的来访次数、IP、独立访客、新访客、新访客浏览次数、站内总浏览次数等数据。这些数据可以直接让企业了解推广成效的来路，从而分析出哪些网站投放的广告效果更明显。图 11 为某产品的访客来源分布图，展示了用户访问产品的途径。

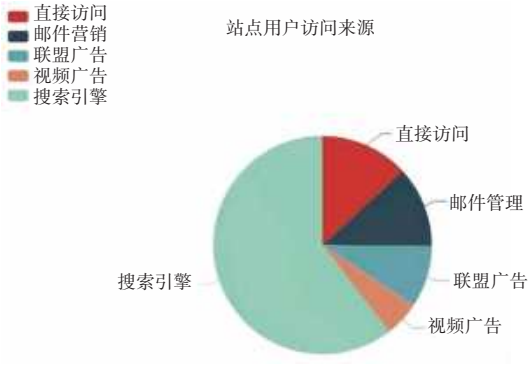

图 11　用户访问来源分布图

Fig. 11　User access source distribution map

5　结束语

本文论述了基于 Spark Streaming 的网站流量实时分析系统的设计与实现，利用实时分析技术对某电子商城网站日志数据进行收集、数据分析、结果持久化、结果可视化展示。实时分析是流处理大数据系统，可对最新实时数据进行高效预设分析处理模型的查询操作，同时数据迟滞低，但是却受限于内存容量。因此，之后的工作重点应放在研发出具有快速高效、智能且自主可控特点的流式大数据实时处理技术与平台。

网站流量分析系统虽然已经发展成为一个相对成熟的体系，但是对于用户的特殊需求还不能满足。因此，流量分析还有待于和数据挖掘技术相融合，从网站访问记录中发掘更有价值的信息。

参考文献

[1] 杜晓春. 基于 Web 的网站数据分析软件 Wysistat 的设计与实现[D]. 西安：西安电子科技大学，2010.

[2] 焦蓉梅. 浅谈网站流量统计分析法[J]. 科技信息(科学教研)，2007(16)：518，512.

[3] 郑腾霄. 基于 Hadoop-Streaming+LNMP 的网站流量分析系统的设计与实现[J]. 现代计算机(专业版)，2018.01：73-77.

[4] 康毅. HBase 大对象存储方案的设计与实现[D].南京大学，2013.

[5] LIU B L，YUAN Minghai，CHEN Guorong，et al. Wireless Body Area Network Data Storage Method Based on HBase[J]. Applied Mechanics and Materials，2013，2748(427).

[6] RAO Lei，YANG Fande，LI Xinming，et al. A Storage Model of Equipment Data Based on HBase[J]. Applied Mechanics and Materials，2015，3744(713). 2015，713-715：2418-2422.

[7] Garg，Nishant.Apache Kafka[M].Packt Publishing，2013

(上接第 200 页)

[2] 周映虹. 多节点 CAN 总线网络的通信设计与测试[J]. 环境技术，2018. 36(3)：78-81.

[3] 朱恒军，于泓博，王发智. 基于 CAN 总线的大棚温度测控系统设计[J]. 微电子学与计算机，2012. 29(5)：183-187.

[4] 王莎. 基于 FPGA 的 CAN 总线通信系统[J]. 工业控制计算机，2018. 31(8)：1-2.5.

[5] 张鹏，吴晓东，崔海青. 一种 ARINC825 总线通信接口可靠性设计方法研究[J]. 计算机测量与控制，2018. 26(9)：210-214.

[6] 王海明. 基于 F28335 的 PROFIBUS-DP 从站与 CAN 通信接口设计[J]. 电子设计工程，2018. 6(19)：123-127，133.

[7] 宋薇，刘晓洁，韩润萍. 基于 C8051F040 CAN 总线的节点通信研究[J]. 计算机系统应用，2009. 18(5)：190-193.

[8] 王东敏，李宏毅. 基于 CAN 总线的风电机组传动系统温度监测节点设计[J]. 电子技术与软件工程，2017(24)：251.

[9] 林雪燕. 基于 CAN 总线的球杆系统的控制系统设计[J]. 自动化与仪表，2017. 32(12)：27-32.

[10] 尚捷. 基于 CAN 总线的旋转导向仪器控制节点设计[J]. 电子测量技术，2017. 40(11)：230-234.

[11] 郭丽萍，张艳荣，林思苗. 嵌入式设备电源控制系统的 CAN 通信软硬件设计[J]. 中国测试，2017. 43(10)：109-113.