文章编号:2095-7386(2015)04-0035-04
DOI:10.3969/j.issn.2095-7386.2015.04.010

基于Hadoop的超市数据分析系统的设计

李 博,范丽丽
(武汉轻工大学 数学与计算机学院,湖北 武汉 430023)

摘 要:大数据的利器——Hadoop,使得对海量数据处理和分析变得更加便宜和快速,吸引了众多急需优化行业模式的企业。构建一个数据分析系统,该系统立足于超市数据的海量、复杂、有利用价值的特点,利用分布式的架构——Hadoop数据分析处理平台对超市小票数据及从微信平台采集的会员信息进行处理和分析,使用优化的关联规则算法,快速准确的分析出超市商品销售的关联性及对顾客消费行为进行预测,从而提高超市商品的销售额以及超市的利润。

关键词:大数据;Hadoop;关联规则算法;数据挖掘
中图分类号:TP 311.13 **文献标识码**:A

Design of data analysis system of the supermarket based on Hadoop

LI Bo, FAN Li-li
(School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan 430023, China)

Abstract: Big data weapon - Hadoop, makes the mass data processing and analysis become cheaper and faster, attracting many enterprises in urgent need to optimize the industry model. This paper presents a data analysis system, the system based on the magnanimity of the small supermarket ticket data, complexity, the use value, uses distributed architecture, Hadoop data analysis platform to process and analyze the small supermarket ticket data processing, using the optimized association rules algorithm to analyse the relevance of the supermarket sales of goods and prediction of consumer behavior, so as to improve the supermarket merchandise sales and profits of a supermarket.

Key words: big data; Hadoop; association rule algorithm; data mining

1 引言

Hadoop[1]作为大数据[2]中最耀眼的技术之一,以其建立于MapReduce基础之上,能够快速,高效,经济地处理和分析数据,因而获得了很多关注。国内外很多企业基于对大数据的追捧及肯定,对Hadoop进行了大量的研究和探讨,均得到一定的突破。重量级的应用有常见的气象数据[3]的处理,热门的微博信息挖掘,高回报的电子商务[4]的商品推荐等。轻量级的应用,如文献[5]提出的基于Hadoop实现的实验数据管理及文献[6]提出的基于Hadoop的电费数据处理等等。

超市作为一个传统零售业,由以前的分散经营

收稿日期:2015-09-18.
作者简介:李博(1991-),男,硕士研究生,E-mail:304237607@qq.com.
通信作者:范丽丽(1981-),女,副教授,E-mail:fll810@live.cn.

到现在的连锁经营,店内的商品也由常见日用消耗品扩为商品仓库。超市面临着如何去管理这些种类和数量繁多的商品的问题。于此同时,网上商城不断涌现也在冲击传统的超市,基于消费者足不出户就可以在网上商城买到心仪的商品,那么传统的门店超市如何去和这些网上商城竞争呢?

超市要想提高与网上商城的竞争力,尽量满足各色各样的消费者需求,同时将具有巨大价值的超市销售海量数据利用起来,其中一个合理的解决方案就是利用数据处理平台或者工具去进行数据管理,分析和挖掘。由 apache 基金会开发一个分布式的架构 Hadoop 数据分析处理平台契合时机的解决了这一问题。用户可以不必在意分布式底层的细节,就能进行海量的数据可靠性存储和高效的数据处理和分析。同时,超市的软硬件环境,也满足构建基于 hadoop 的数据分析系统的要求。

2 基于 Hadooop 数据分析系统环境

本系统由多台 PC 机组成的硬件环境,硬件上搭载 Hadoop 集群所需的软件包,利用网络和 Hadooop 分布式的特点,将数据储存在集群的各个节点上,利用集群的并行计算能力进行数据处理和分析

2.1 硬件环境

超市内部的 11 台 PC 机,企业级路由器一台。

内存:2GB。

硬盘 500GB。

内部局域网具有 100MB 宽带速度。

2.2 软件环境

Linux Ubuntu11.0。

Hadoop 0.20.0 包。

Sun-java7-jdk 包。

SSH 包。

2.3 数据采集环境

利用微信公众平台收集顾客信息,将超市会员信息从微信平台接口转存储在超市数据分析系统的数据库中。将超市会员的消费记录也存储在数据库中。利用积分活动的方式,鼓励会员在线上对其购买的商品进行评分。数据分析平台对这些数据进行处理分析从而预测顾客购买行为。

相对于非会员客户,超市会员消费数据具有准确性和稳定性,这有利于对数据进行快速处理和分析。因此,对于非会员顾客,一方面要主动引导其加入会员,另一方面对收银台的非会员顾客消费小票数据进行存储,拟定一个周期,对其进行处理分析。

3 系统涉及相关技术

本文系统以 hadoop 构成的集群作为数据分析的平台,文件存储系统基于 HDFS,采用 pig 作为用户数据分析的工具,利用 MapReduce 对数据进行分布式计算。

3.1 分布式文件系统 HDFS

Hadoop 分布式处理平台的存储是基于分布式文件系统 HDFS,它采用的是只写一次,读取多次的文件访问模型。文件被创建写入关闭后就不允许被修改。这个假设简化了数据一致性问题因而保证了高数据访问吞吐量。超市小票明细数据属于历史分类数据,不需要对其进行增删改查的操作,只要批量导入和查询,符合 HDFS 数据访问模型。

HDFS 框架图如图一:

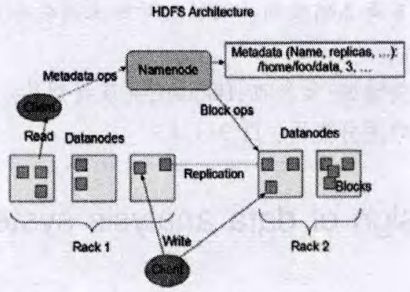

图 1 HDFS 框架图

3.2 MapReduce 分布式计算框架

基于分布式的数据存储模型,mapreduce 也是一个分布式的数据计算框架。并行计算是处理海量数据的常规手法,但是并行计算并不是大多数程序员具备的能力。Mapreduce 提供一种有两个过程的简单计算模型,一个叫做 map,一个叫 reduce。Map 过程负责通过 split 将数据一分为二,而 reduce 则合并 map 处理的结果。通过这两个步骤隐藏了很多数据并行计算的复杂性。Hadoop 也正是依靠这种化繁为简的框架成为了分布式计算中的行业标准,得到广泛应用。

数据处理模型如图二:

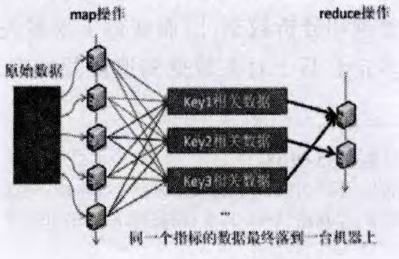

图 2 数据处理过程模型

3.3 数据分析工具 Pig

Pig 是一个基于 Hadoop 的大数据处理分析工具,它提供了一种类 SQL 语言——Pig Latin,它通过编译器把 SQL 数据请求转换为 Mapreeduce 运算,并且提供操作简单的操作和多种编程语言的接口。

即便是简单的操作,例如将一个文件中数据提取出来,对数据进行一次处理操作,再将处理后数据存储在另一个文件中。利用传统的 MapReduce 模型去处理,得单独写一个预处理复杂的应用程序。用 pig 来写脚本的话,只需要三步。例如在文件 file 中查找 pig 字符的脚本:

```
File = LOAD'file';
files = FILTER file BY $0 MATCHES'. * pig +. *';
STORE files INTO 'file1';
```

三步操作分别是读取 file 文件数据,筛选数据,存储数据到 file1 中。

3.4 数据收集平台

微信上亿的用户,海量的用户数据和可捕捉的用户行为现在已经牵动着众多商户的注意力。商家可以通过微信公众平台采集用户的用户行为数据及反馈数据以期快速的做出相应对策来应对大数据时代。本系统利用微信公众平台的 access_token 接口,通过 OAUTH2.0 方式弹出信息录入页面,经过用户授权后将获取的数据存储。将微信获取的用户基本信息与数据库中会员信息进行匹配,同时进行对应的推送信息。超市就可以利用现有的微信来获取用户数据,辅助和验证本文提到的设计的可行性和有效性。

4 系统结果分析

基于 Hadoop 框架的数据处理系统是分布式的处理数据,最重要的就是 Map - Reduce 思想及 Hadoop 集群的搭建,再通过关联规则算法找出数据处理的结果。

4.1 Hadoop 平台集群的实现

1)利用交换机将主机连接在一起,组成局域网,安装 JDK,安装 SSH 以方便远程无密码连接

2)为主机设置固定 IP,并起一个方便识别的别名

3)安装 Hadoop 客户端,对主机的相关文件进行配置,配置成一个 namenode 节点,多个 datanode 节点。节点拓扑图如图三:

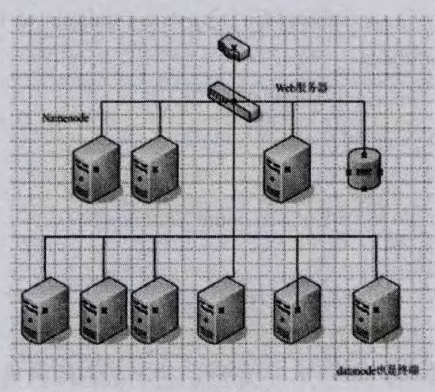

图 3　节点拓扑图

4)打开 Hadoop 进程,利用 JPS 测试 Hadoop 是否安装成功。安装成功后,利用 Hadoop 自带的集群状态页面查看集群状态。

4.2 MapReduce 数据处理的实现

基于 HDFS 的文件是分布式存储的。每一个节点都分布式的存储着数据。在需要数据的时候,节点都是按照就近原则,对自己节点所存储的数据进行处理,处理完成以后再统一交给某一节点汇总处理。

具体实现:程序将数据文件的地址告诉 Namenode 节点,数据文件的地址作为 Mapreduce 的输入。在 map 阶段,利用相关函数对数据进行清理,划分,得到 key 值,在利用 reduce 程序对 < key,value > 进行整合,将整合的文件作为输出。

处理数据的时候,数据量比较大,所以不能简单的把数据放在内存中进行处理,而是要把数据保存在文件里,然后反复的读取。利用内存映射或者是分块处理的办法提高文件数据读取速度。数据保存为二进制,这样载入内存更快,占用内存空间更小。

4.3 Apriori 算法的实现

Apriori 算法[7]通过两个重要的性质实现压缩搜索空间的,这利于提高频繁项目集逐层产生的效率。性质如下

性质一:如 A 是频繁项集,则 A 的所有子集是频繁项集。

性质二:如 B 为非频繁项集,则 B 的所有超集是非频繁项集。

算法:Apriori 算法

输入:数据库 DB;最小支持度值 min。

输出:DB 的频繁项集 L。

```
1) L1 = find(DB);
2) for (n = 2; Ln-1 ≠? ; n++) {
3) Cn = aproiri_gen(Ln-1,min);
4) for each transaction t D{ //scan D for count
```

5) Ct = subset(Cn,t); //get subsets of t that are candidates

6) for each candidate c Ct

7) c.count + +;

8) }

9) Ln = {c Cn | c.count ≥ min }

10) }

11) return L = ∪nLn;

具体过程如下:

第一步,算法第一次迭代,进行数据库扫描后计算出 DB 中每个项目出现的次数,生成 1 - 项集 C1。

第二步,通过最小支持度筛选出频繁 1 - 项集 L1。

第三步,由 L1 产生 C2 并且扫描 DB 中 C2 项目出现的次数,由最小支持度生成 L2

第四步,由项集 L2 生成 C3,以此类推直至找不到高层项集。

假设有如下表的数据,A,B,C 表示三种商品,支持度表示购买一次销售小票中此商品出现的概率,支持度如下表 1 所示:

表 1 商品支持度

项	支持度
A	0.4
B	0.3
C	0.2
A and B	0.2
A and C	0.15
B and C	0.1
A,B and C	0.04

可得到以下规则,如表 2 所示:

表 2 商品规则

规则	置信度
If B and C then A	0.04/0.1 * 100% =40%
If A and C then B	0.04/0.15 * 100% =26.7%
If A and B then C	0.04/0.2 * 100% =20%

对于规则" If B and C then A",同时购买 B 和 C 的人中,购买 A 的顾客有 40%。A 本身的支持度有 0.4,也就有 40% 的人购买 A 商品。

引入一个度量此规则有效性的量,即提升度(Lift)。此度量通过量化数据来验证使用规则和不适用规则的差距。暂定大于 1 的提示度是有用的规则的。

计算规则如下公式:

Lift(A = = > B) = Confidence(A = = > B)/Support(B) = Support(A = = > B)/(Support(A) * Support(B))。在上例中,Lift(If B and C then A) = 0.04/(0.1 * 0.4) = 1。而 Lift(If A then B) = 0.2/(0.4 * 0.3) = 1.33。由此,买了 A 的人进行推荐 B,购买概率是随机推荐 B 的 1.33 倍。

5 系统演示

本节将探讨基于 Hadoop 的数据处理系统和传统的数据处理工具 SQL Server 和 Excel 的性能比较。现取某一超市不同日期 3 天的所有销售数据,113 万条数据记录,51 万条数据记录和 9 万条数据记录。由于 Excel 和 SQL Sercer 都是单机运行,故比较他们的数据载入时间。三个平台分别计算记录中销量最高的前 10 位,所需要时间如表 3 所示,Hadoopt Server 耗时截图如图 4 所示。

表 3 简单数据处理对比

平台\量	113w	51w	9w
Excel			1753"
SQL server	529"	314"	29"
Hadoop	133"	59"	57"

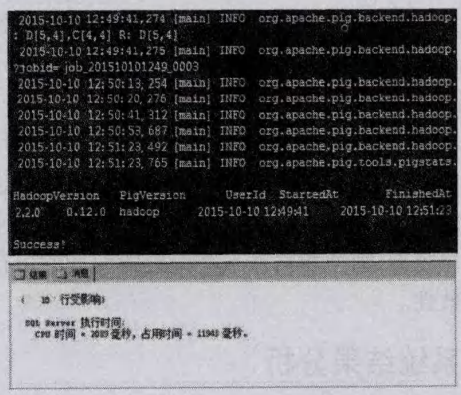

图 4 Hadoop 和 SQL Server 耗时截图

再分别在三个平台上测试找出记录中两种商品出现在同一个小票的的前 10 种,结果如表 4 所示。

表 4 关联数据处理对比

平台\量	113 万	51 万	9 万
Excel			
SQL Server	1329"	714"	2'29"
Hadoop	343"	226"	142"

由表 3 和表 4 所示,Excel 对于数据量大,缺乏快速处理的能力。SQL Server 在处理数量级小和简单的数据时耗时比 Hadoop 平台少,但是处理关联数

(下转第 46 页)

喜好商铺、购物时间等信息,实现更人性化的购物导航。

参考文献:

[1] 新浪科技. 谷歌开始在美国加州测试 WiFi 气球[EB/OL]. (2015-05-20) http://tech.sina.com.cn/i/2013-08-10/13448625401.shtml.

[2] 网易新闻. Facebook 造大型太阳能无人机建全球 WiFi 网[EB/OL]. 2015-05-20. http://news.163.com/air/15/0328/10/ALpn0g6k00014P42.html.

[3] 张运超, 陈靖, 王涌天. 基于移动增强现实的智慧城市导览[J]. 计算机研究与发展, 2014, 51(2): 302-310.

[4] 王瑞峰. 基于 WLAN 构建无线城市的规划设计分析[J]. 电信科学, 2011, 27(6): 121-126.

[5] 吴雷. 移动互联网领域的商业模式创新趋势[J]. 中国传媒科技, 2015, (1): 63-66.

[6] 张桂玲. 提升 Wi-Fi 商业价值的四条路径[J]. 通信世界, 2014, (17): 36.

[7] 陈永光, 李修和. 基于型号强度的室内定位技术[J]. 电子学报, 2004, 32(9): 1456-1458.

[8] 董梅, 杨曾, 张健,等. 基于信号强度的无线局域网定位技术[J]. 计算机应用, 2004, 24(12): 49-52.

[9] 孙佩刚, 赵海, 罗玎玎,等. 智能空间中 RSSI 定位问题研究[J]. 电子学报, 2007, 35(7), 1240-1245.

[10] 朱剑, 赵海, 孙佩刚,等. 基于 RSSI 均值的等边三角定位算法[J]. 东北大学学报, 2007, 28(8): 1094-1097.

(上接第 38 页)

据或者百万级别的数据记录耗时比较长。相比之下,Hadoop 数据分析处理平台处理这种关系型数据,耗时短,速度快,可以提高掌握实时销售情况的能力,让超市更机动灵活的制定销售策略,提高销售额。同时,系统可以根据当前任务,调节 Hadoop 的配置,如文件复制数,blocksize 等,也能够提高系统的数据处理速度。

6 结语

结合超市现有的软硬件资源,构建基于 Hadoop 的数据分析系统,通过数据挖掘技术,利用 Hadoop 集群的强大的数据处理能力,实现对超市海量数据的处理分析,发现出商品 - 商品,商品 - 顾客之间的联系,从而提高了超市的销量和利润。大数据利器—Hadoop 对以后的超市所代表的零售业的商业模式有着潜在的巨大影响,合理的运用能提高企业的竞争力。

参考文献:

[1] Tom White. Hadoop: The Definitive Guide[M]. California, USA: O'Reily Media, Inc., 2012.

[2] Viktor Mayer-Schönberger, Big Data: A Revolution That Will Transform How We Live, Work, and Think [M]. Oxford, UK: Eamon Dolan, 2012, 12.

[3] 薛胜军,基于 Hadoop 的气象信息数据仓库建立与测试,计算机测量与控制,2012,20(4): 925-928.

[4] 李文海,基于 hadoop 的电子商务推荐系统的设计与实现,计算机工程与设计,2014,17(4):71-75.

[5] 管莹,基于 hadoop 的实验数据管理系统的实现,电脑编程技巧与维护,2014,11(4):42-45.

[6] 郭朝鹏, HaoLap: 基于 Hadoop 的海量数据 OLAP 系统,计算机研究与发展,2013,50(z1):380-383.

[7] 李唐平,基于矩阵的关联规则算法与 Apriori 算法的研究及改进,计算机学报,2013, 31(5):120-126.