基于 Hadoop 集群的多表并行关联算法及应用

郑晓薇,马 琳
(辽宁师范大学 计算机与信息技术学院,辽宁 大连 116081)

摘 要:针对因特网环境下并行数据库实现多个大数据表关联存在的计算瓶颈,基于 Hadoop 集群设计了一个并行关联多个大数据表的简便算法 MR_Join。以商业网站凡客诚品的销售数据为例进行实验,验证算法的可行性并做出应用实例。实验结果表明,MR_Join 算法可以有效地实现大数据表的快速关联,具有显著的并行效率。

关键词:Hadoop 集群;Mapreduce 编程模式;MR_Join 算法;数据表并行关联
中图分类号:TP393 **文献标识码**:A **文章编号**:1674-7720(2013)04-0091-03

Multi-chart parallel correlation algorithm and application based on the Hadoop cluster

Zheng Xiaowei, Ma Lin
(Computer and Information Technology Institute, Liaoning Normal University, Dalian 116081, China)

Abstract: In the environment of Internet, using parallel database to realize the data table connection has calculation bottlenecks. Based on the Hadoop cluster, design a parallel correlation multiple large data table MR_Join simple algorithm. In order to further verified the feasibility of the algorithm, in the Hadoop cluster to commercial websites where sales data as an example for the experiment. The experimental results show that MR_Join algorithm can effectively achieve the large data table fast connection, has the remarkable parallel efficiency.

Key words: Hadoop cluster; Mapreduce programming model; MR_Join algorithm; data table parallel association

近年来,随着网络技术的飞速发展,具有价位合理、购买便捷等优势的电子商务迎来了崭新的春天。商品交易市场正从卖家市场转向买家市场,消费者面对种类繁多的商业网站及产品有更多的选择性,商家只有把握顾客才能达到企业盈利的目的。深层挖掘网站交易数据信息有利于营销决策的制定,多个大数据表关联并转换成适合挖掘的形式是必要步骤。常见的方法是使用并行数据库[1],但由于其架设和调优难度大、对异构硬件的支持有限、成本高以及需要对数据的存储进行格式定义等缺陷导致了其处理因特网中多个大数据表关联时使用不便。

Mapreduce 是一种分布式并行编程模型,Apache 开源社区的 Hadoop 项目是一个使用 Java 语言实现 Mapreduce 模型的开源平台。近年来,基于 Hadoop 平台在 Web 日志挖掘[2]、微博信息挖掘[3]、搜索引擎用户行为分析[4]、城市交通碳排放数据挖掘研究[5]等方面都有很多应用。与并行数据库相比,Hadoop 集群不需要对数据的存储进行格式定义,可将大数据表分解到各个计算节点,由各节点并行执行,集群监控各个计算节点的任务状态,具有高容错性和高扩展性。同时,Hadoop 满足数据的多级计算和处理,可有效解决“一个程序的输出是另外一个程序的输入”这类复杂的数据挖掘。

本文基于 Hadoop 集群,设计了一个适合多个大数据表并行关联的简便算法 MR_Join。以商业网站凡客诚品的销售数据表为例进行实验,实验结果显示 Hadoop 具有在处理网络下大数据表关联的优势,也验证了 MR_Join 算法的可行性。集群中各个计算节点并行处理,处理速度快、延迟低且易于操作。

1 MR_Join 关联算法设计

Hadoop 的两个核心部分是分布式文件系统 HDFS 和 Mapreduce 模型。HDFS 为分布式计算提供了底层存储支持,易于读取大规模的数据文件。Hadoop 集群由一个 NameNode 和一组 DataNode 组成,采取 Master/Slave 的架构。Mapreduce 分布式并行编程模型的基本思想源于函

数式编程语言[6]，Map 和 Reduce 是该模型的两大基本操作。Map 函数指定各分块数据的处理过程并映射出中间结果，Reduce 函数指定如何对中间结果进行归约并生成最终的处理结果。其概念可以表达为：Map<k1，v1>list<k2，v2>；Reduce<k2，list(v2)>list(v2)；

1.1 MR_Join 大数据表并行关联算法

MR_Join 算法是一种关联有相同表主键的表的关联方法，其计算流程如图 1 所示。算法思路为：被处理的多个表应具有相同的表主键，算法首先判断输入的数据块归属于哪个表，同一个表的数据块由同一个 Map 处理。Map 操作把表数据拆分成以表主键为 Key 值的键值对，Reduce 操作把具有相同 Key 值的键值对聚合在一起。最后从各个表中分别选取一个属性（选取的属性之间应存在直接联系），将属性值进行运算得到一个新属性的属性值，在运算过程中实现数据表关联。在 Hadoop 集群中，NameNode 节点分配计算任务给各个 DataNode，多个 DataNode 并行执行计算任务，达到短时间内快速准确地完成两个大数据表关联的效果，生成一个可以有效分析的大数据表。本文采用 Java 语言实现 MR_Join 算法。

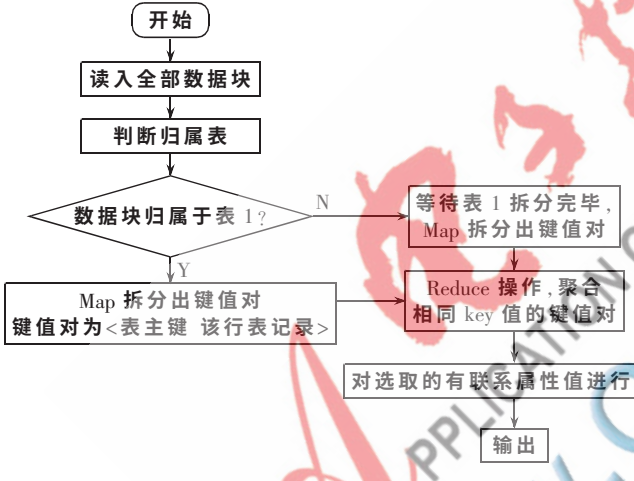

图 1 MR_Join 大数据表关联计算流程

1.2 MR_Join 算法应用实例

本文以消费网站凡客诚品的商品信息和销售数据为例，数据获取的方式采用 Web 爬虫方式。Web 爬虫通过访问凡客诚品网站的 Web 页面，解析页面相关内容和页面源代码，获取所需的数据信息，生成商品属性 Product 表和购买商品的用户信息 Sale 表，并以文本文件的形式输出数据表。两个数据表的“商品编号”字段作为主键，并根据 Product 表的“单价”和 Sale 表的“数量”计算出“总价格”，计算结果按照“商品编号”升序排序。两个表相连能够显现出购买每种商品的用户信息和该用户的消费总价格，有利于商家对客户进行归类及后续对应的商品营销。Reduce 操作按“商品编号”对计算结果升序排序，同类商品销售情况一目了然，为下一步的市场营销策划提供客观依据。

MR_Join 算法处理数据基本流程如图 2 所示。

(1)定义 Product ID 为商品表的“商品编号”字段，Sale ID 为用户表的“商品编号”字段。num 为用户表的“数量”字段，在 Product 表中 num 值为 0。“总价格”是通过“单价”和“数量”之间的运算得到的新字段。

(2)算法默认逐行读取表记录并将记录偏移量及该行记录内容映射为初始键值对。Map 操作对初始键值对进行处理，提取出“商品编号”形成中间键值对，生成<(Product ID，num)记录内容>或<(Sale ID，num)记录内容>。

(3)Shuffle 和 Sort 操作把具有相同键值的键值对合并分组，其结果作为 Reduce 操作的输入键值对。

(4)Reduce 操作将相同键值的键值对聚集，由于 Web 爬虫在爬取信息记录时，消费数量是生成相同数量的相同消费记录，所以 Reduce 过程先对 Sale 表的键值对进行 For 循环计算“数量”，再用 For 循环作相乘运算计算“总价格”。总价格=(Sale 表)num×(Product 表)单价。

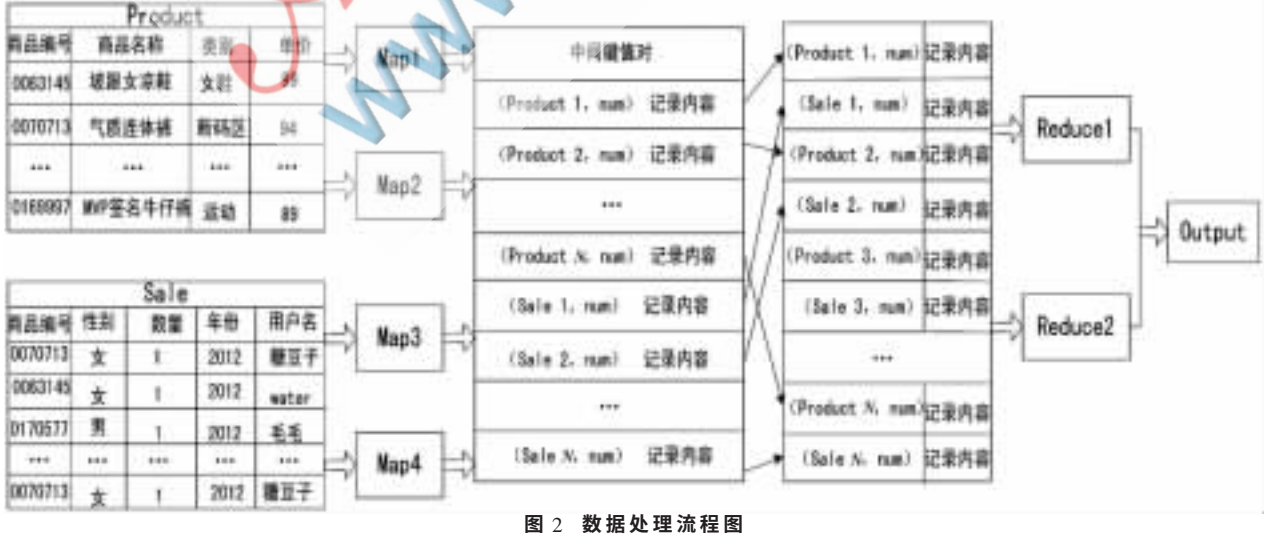

图 2 数据处理流程图

2 实验结果及分析

2.1 实验环境

本实验运行在4台PC机搭建的Hadoop集群上，均为同等配置。各节点名分别为Master、Slave01、Slave02和Salve03。各个节点均安装了Hadoop-0.20.2和JDK。在Hadoop搭建的集群系统上运行了本文开发的MR_Join算法。本实验分别在单机、多个节点集群中运行。实验数据是利用Web爬虫在凡客诚品网站获取的2012年4月份的商品信息和销售数据。

2.2 实验内容及结果分析

(1)单节点测试。分别执行了200 MB、500 MB、1.1 GB和1.5 GB 4种大小不同的商品属性表和用户信息表的关联。通过图3可以发现，200 MB表文件的执行时间比500 MB表文件的执行时间长，之后随着表文件规模的增大，算法执行的时间也相应地增加。说明当数据规模较低时，由于Hadoop框架启动时间长、开销较大，执行效率低。当数据规模增大时，执行效率逐渐增大。

(2)多节点并行测试。图3为单节点环境与双节点环境下MR_Join算法执行时间对比，表1为MR_Join算法多节点并行执行效果分析表。P为集群节点数目。

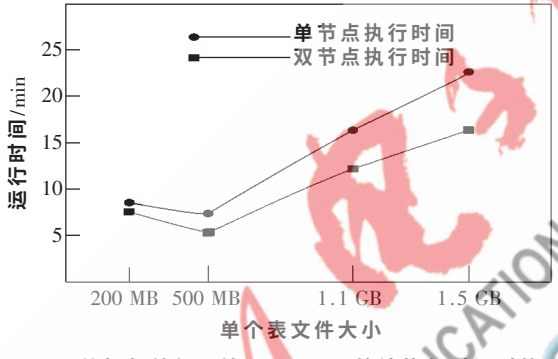

图3 单机与并行环境下MR_Join算法执行时间对比

实验结果显示出3对不同规模的大数据表关联时，并行效率与集群的节点数目成正比，且随着表数据规模的增大而增大。当表数据规模较低时，利用Hadoop集群执行大数据表关联效率低。当表数据规模较大时，利用Hadoop集群并行执行计算，在节点数为4时可达到78.2%的并行效率。说明在表数据规模和Hadoop集群节点数目选择适当的情况下，本文设计的MR_Join算法可以取得良好的并行效率。

表1 MR_Join算法多节点并行执行效果分析表

任务规模	计算时间/s			加速比			并行效率/%		
	P=2	P=3	P=4	P=2	P=3	P=4	P=2	P=3	P=4
500 MB	5.09	3.36	2.46	1.39	2.11	2.89	69.8	70.3	72.2
1.1 GB	11.25	7.24	5.27	1.43	2.22	3.05	71.4	73.9	76.2
1.5 GB	15.26	10.05	7.16	1.47	2.23	3.13	73.5	74.4	78.2

大数据表关联后产生的数据表是企业数据挖掘和营销决策制定的重要数据基础。本文设计的MR_Join算法以商业网站凡客诚品的商品销售数据为例进行实验，成功实现了大数据表关联。实验结果表明，Hadoop集群对于并行执行计算任务的高容错性及高度可扩展性使得大数据表关联的结果准确且并行效率显著，避免了并行数据库在因特网环境下实现表关联的弊端，可以在电子商务的商业应用中体现更高的价值。现今Hadoop已经被广泛应用于海量数据存储和分析、互联网服务、搜索引擎等。本文针对Hadoop下的大数据表关联作了有益的尝试，本文的算法也可以用于其他相关应用中。

参考文献

[1] 王珊，王会举，覃雄派，等.架构大数据：挑战、现状与展望[J].计算机学报，2011，34(10)：1741-1751.

[2] 程苗，陈华平.基于Hadoop的Web日志挖掘[J].计算机工程，2011，37(11)：37-39.

[3] 林大云.基于Hadoop的微博信息挖掘[J].计算机光盘软件与应用，2012(1)：7-8.

[4] 王振宇，郭力.基于Hadoop搜索引擎用户行为分析[J].计算机工程与科学，2011，33(4)：115-120.

[5] 朱钥，贾思奇，张俊魁，等.基于Hadoop的城市交通碳排放数据挖掘研究[J].计算机应用研究，2011，28(11)：4213-4215.

[6] 谢桂兰，罗省贤.基于Hadoop Mapreduce模型的应用研究[J].微型机与应用，2010，29(8)：4-7.

(收稿日期：2012-10-30)

作者简介：

郑晓薇，女，1957年生，教授，主要研究方向：并行计算、数据库与计算机决策支持。

马琳，女，1987年生，硕士，主要研究方向：并行计算。