

文章编号:1005-0523(2013)05-0037-05

基于Hadoop的电网监控信息流分布式处理研究

屈志坚,郭亮,陈阁

(华东交通大学电气工程学院,江西 南昌 330013)

摘要:针对智能电网调度系统中大数据集监控信息流存取困难问题,提出一种基于Hadoop云计算框架的分布式集群处理新方法。通过分析电网监控系统中信息流特性,提取3类关键信息流;综合利用分布式文件系统HDFS和映射聚合模型Map/Reduce,建立云集群分布式处理平台,实现监控信息的高效并行处理。以某配电网断面量测记录数据集为例进行测试,结果表明:该方法处理效率相比传统客户/服务器数据库方法提高了约1.4倍,集群规模越大,效果越好,有效避免了大数据集信息处理响应延迟问题。

关键词:智能电网;信息流;Hadoop云计算;映射聚合;并行处理

中图分类号:TP274

文献标志码:A

智能电网是未来电网的发展方向,其与分布式电源、微电网等的互动调度^[1-5]都依赖于双向、实时、集成的通信系统。随着电网扩容和大量智能电气设备的接入,以数据应用为核心的电力调度监控系统^[6-7]长期运行将面临大数据集量测信息的存取瓶颈问题。而基于客户/服务器模式的传统开放式数据库存取调用方法容易造成服务器信息量太大,查询效率低^[8]。因此,高效的信息处理已成为制约电网自动化监控的关键技术。

Hadoop是一种专为大数据集对象分析处理提出的分解/聚合云计算框架^[9]。通过分布式文件系统HDFS和并行编程模型Map/Reduce两大核心的协调管理,对大数据集进行有效分割和合理调度,以实现大数据集高效并行处理。框架易部署于价格低廉的通用PC计算机设备上,具有高扩展性和高容错性等优点。目前云计算技术在电信、电子商务和web网页搜索等领域的应用已展示出其在大数据集信息查询处理问题的优势^[10-13],给工业应用系统信息存取方面的研究提供了新思路。

首先分析调度监控系统中关键信息流特性,然后构建Hadoop分布式集群平台。利用分布式文件系统HDFS和并行计算模型Map/Reduce与监控信息存取应用相结合,建立起大数据集信息的分布式云处理平台,实现大数据集监控信息的高效并行处理。实验证明该方法提高了信息处理效率,满足电网调度大数据集监控信息的存取要求。

1 监控信息流和传统数据库调用模式

1.1 监控信息流

在电网监控系统中,通过FTU(feeder terminal unit)、RTU(remote terminal unit)和STU(signal terminal unit)采集电网高压侧和低压侧供电线路的各类实时数据,利用远程通信接口,传输至调度监控主站,实现数据处理以及监视控制功能,其总体功能流程如图1所示。

由图1得出,电网监控业务处理有如下3类关键数据流^[14]:

收稿日期:2013-06-16

基金项目:国家自然科学基金项目(51267005);江西省教育厅科技研究项目(GJJ13350)

作者简介:屈志坚(1978—),男,副教授,博士,研究方向为铁道电网智能调度的监控理论与技术。

1) 下行方向的实时控制命令流,包括远程控制命令和远程参数设置命令(电流、电压阈值、延时限值、报警限值等)处理流;

2) 上行方向的实时数据处理流,包括将实时数据(状态量、模拟量、越限报警等)信息传送到主站监控画面实时显示的数据流和处理后存储进入历史数据库的处理流;

3) 历史性数据显示处理流,包括从关系数据库中取出历史数据(操作记录、事件记录、状态量的遥信历史、模拟量的遥测历史、报警记录等)信息。

这3类数据流在电网监控系统中同时运行,共同完成整个系统的数据处理功能,将涉及到大量高负荷的数据存储和调用操作。

1.2 传统数据调用模式

传统的数据调用模式变现成为一种客户/服务器模式。首先在主程序中建立数据库连接;然后进行SQL操作调取数据;最后断开数据库连接。在每发生一次调用请求时都必须重新与数据库建立连接。由于电网监控实时性要求非常高,即使在很短时间段内,其调用请求数可以达到成百上千,当系统运行出现异常时将产生大量突发性实时数据,这将引起大量数据调用请求,导致非常频繁的数据库连接与断开操作,带来非常庞大的系统开销。因此,当出现大数据集信息存储调用操作时,传统数据库调用模式就不能满足实际应用需求。

利用Hadoop框架将大数据集信息进行分割成若干个小数据块,再合理分配给集群内部从节点并行处理。可有效避免传统数据库模式多次请求调用问题,大大减少了系统开销应用。

2 基于Hadoop的监控信息存储调用系统

2.1 HDFS构架

HDFS采用主/从节点模式。主节点部署名称节点(Namenode)和作业监视节点(Jobtracker),对内维护元数据,对外管理分布式文件系统。从节点部署数据节点(Datanode)和任务监视节点(Tasktracker)。按照Hadoop内部资源调度算法,将大数据集文件分块处理,由主节点合理分配于各个数据节点,系统默认分块大小为64 MB。主节点通过Datanode周期发送的脉搏通信来监控数据节点。Jobtracker负责统一调度子任务于Tasktracker上运行,Tasktracker是运行于多个Datanode上的集群从节点服务,负责子任务执行。数据调用类通过Jobclient函数提交到Jobtracker,创建Map/Reduce任务,分发到各Tasktracker执行。

Hadoop分布式文件系统设计与实现,体现了框架对海量数据进行高可扩展和高可靠的存储和管理,其体系架构如图2所示。

结合HDFS文件系统特点和电网调度监控信息特点,构建一个存储调用系统如图3所示。数据来源为包含终端采集设备和工业以太网。处于中间的Hadoop集群为系统的核心部分。客户通过Namenode访问

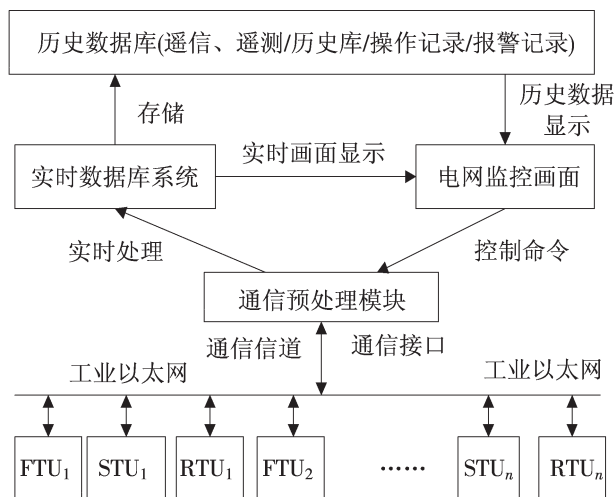


图1 电网监控业务流程

Fig.1 Power grid monitored business information flow

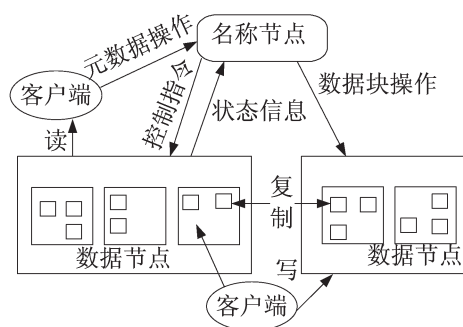


图2 分布式文件系统系统构架

Fig.2 Distribution file system frame

文件系统,电力调度监控数据存储于多个计算机节点组成的分布式文件系统中。存储客户端主要完成数据实时存储功能,放置在调度中心站。查询客户端完成海量数据查询处理操作,放置在工区或者数据中心。

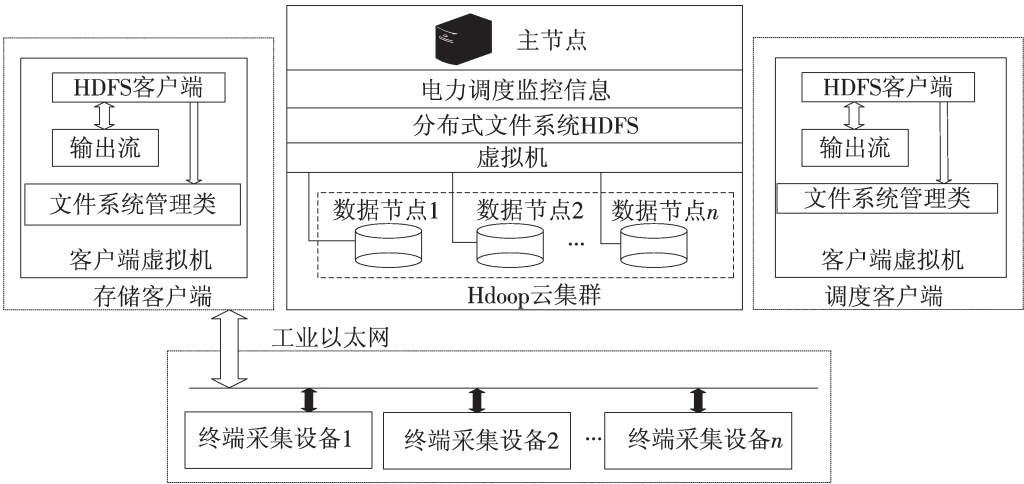


图3 分布式存储调用系统
Fig.3 Distributed storage access system

2.2 Map/Reduce并行处理引擎

映射/聚合 Map/Reduce 模型作为 Hadoop 框架的另一个核心,由映射 Map 和聚合 Reduce 两个大规模数据分析引擎构成,Map 任务用于从 HDFS 中获取键值对 key/value 模式的大数据集输入,将其切分为小片段,交由数据处理进程执行,而 Reduce 用于整理输出映射收集的各种操作结果,同样也是以键值对模式输出,最终存储在 HDFS 中。其处理流程如图 4 所示。

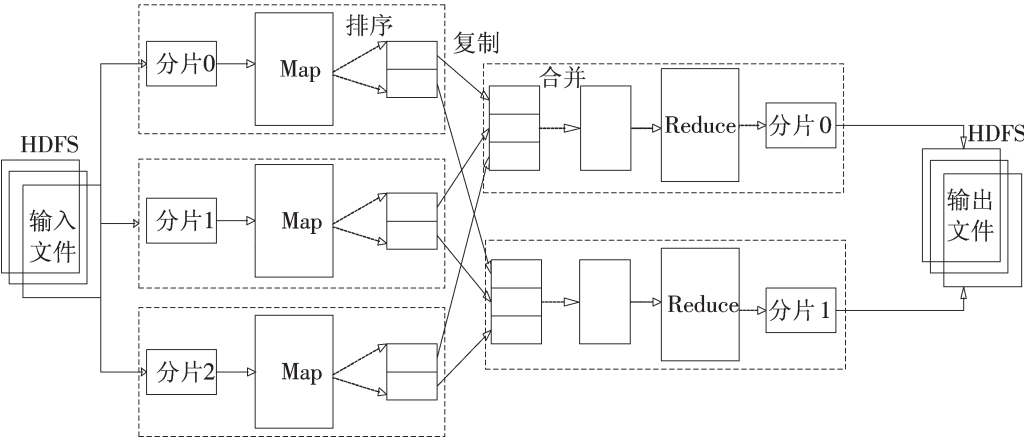


图4 Map/Reduce处理流程
Fig.4 Map/Reduce processing flow

2.3 监控信息的映射/聚合处理实现

首先启动 Hadoop 服务平台,为映射/聚合提供云主节点和云从节点。其次建立与监控系统数据库的连接,获取作业 ID。保证云计算节点能够动态从监控数据库发送数据调用请求。然后,将数据调用任务进行分解、复制、提交至各个数据节点。在数据节点中通过作业 ID 获取作业资源,由数据节点执行数据库查询输出操作。

最后,在各个数据节点执行完毕后聚合查询的数据信息结果输出至 HDFS。整个作业实现流程图如图 5 所示。

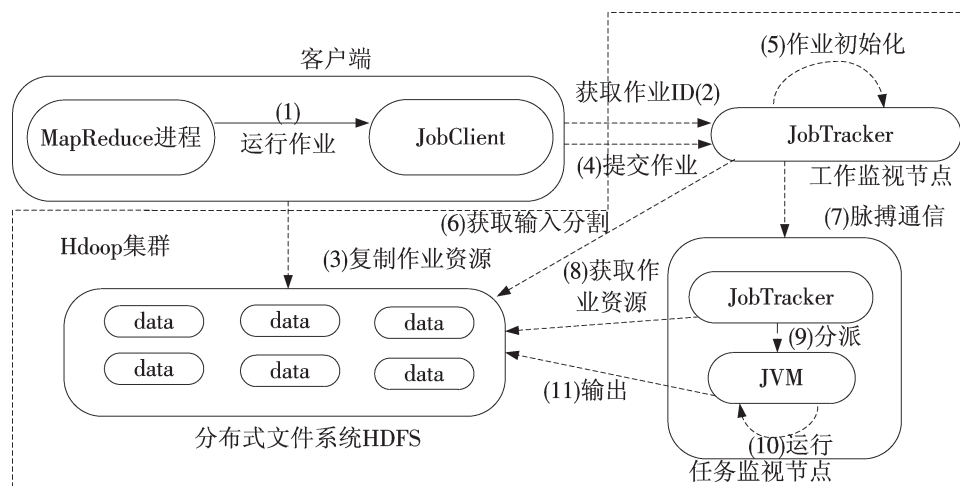


图5 映射/聚合作业实现过程

Fig.5 Implementation procedure of Map/Reduce job implementation procedure

3 实验环境

为了验证 Hadoop 云计算框架建立的分布式存储调用平台在系统工程中的应用,结合调度监控中大数据集量测信息应用需求,建立由 2 台调度机(主调度机和备调度机均采用双核处理器)和 2 台监控数据服务器(采用 64 位处理器)构成的 HDFS 分布式集群实验环境,如表 1 所示。

表 1 分布式调度监控的集群实验环境

Tab.1 Distributed dispatching monitor cluster environment

云集群	计算机配置	操作系统	数量/台
主节点(主调度机)	Intel(R)Core(TM)2 Duo CPU E7200 2.5 3 GHz, 2.0 GB	Ubuntu 11.04	1
从节点(监控服务器)	AMD Athlon 64 Processor 4300+ 1.87 GHz, 2.0 GB	Ubuntu 11.04	2
从节点(备调度机)	AMD Athlon(TM) X2 DualCore Ql- 66 2.20 Hz 2.0 GB	Ubuntu 11.04	1

首先在主调度机中启动 Hadoop 云计算服务平台,包含由 Namenode, Secondnamenode, Datanode, JobTracker 和 Tasktracker 组成的 5 个节点。其中在主节点中运行名称节点和工作监视节点,各从节点中运行数据节点和任务监视节点。主调度机完成对下面的备调度机和监控服务器的监控和管理任务。核心任务由 Map/Reduce 并行模型执行,其程序代码利用 Eclipse(Java 可视化环境)编程。

然后启动任务的调用请求数据处理进程,云主节点通过 Map 任务按照分解规则将信息处理任务进行分割处理。再交给数据节点并行计算处理,然后交给 Reduce 任务进行最终归并、聚合操作,并存储至分布式文件系统 HDFS。

4 响应时间测试

Hadoop 的 Map/Reduce 计算依托 HDFS 分布式集群框架,启动集群环境后,计算引擎以并行方式处理大数据集。以不同的量测断面数为输入,将批量的断面实验数据集映射为结构化对象格式,从配电网调度监控系统量测数据库中导出至 HDFS 中,使断面数据集在云集群环境中被分布存储、分散处理。

对不同数量批量断面数据集进行 Map/Reduce 处理,记录写入 HDFS 的字节数和所需测试时间。分别

在传统数据库模式、单机模式和云集群模式下,选择记录数据分别为(100 000, 500 000, 1 000 000, 3 000 000, 5 000 000)的量测数据集进行调用访问测试,系统执行响应时间曲线如图6所示。

从图6可知,在执行调用100 000量测记录时,3种模式执行效率水平相当;当量测数据集记录从10万逐渐递增时,云集群模式下的处理效率开始凸显优势,在访问记录数达到500万时,响应时间约为141 652 ms,处理效率相比传统模式提高了约1.4倍。与文献[15]提出的方法相比,处理效率更快。同时整个响应时间曲线呈现出趋柔性态,验证了云计算技术非常适合大规模数据量的并行处理。

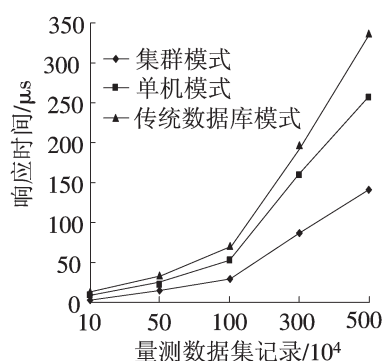


图6 不同量测数据集记录响应时间特性曲线
Fig.6 Time contrast for compression with different measurement data sets

5 结束语

针对电网监控系统中传统客户服务器模式下大数据集处理效率受限制的问题,结合监控信息流特性,提出一种基于云计算Hadoop的分布式处理新方法。以不同的量测侧面数据集记录进行实验测试,结果表明相比传统数据库模式和已有方法,其信息处理效率更加高效和可靠。

参考文献:

- [1] 赵江河,陈新,林涛,等. 基于智能电网的配电自动化建设[J]. 电力系统自动化,2012,36(18):33-36.
- [2] 彭春华,齐彦伟,陈首昆. 基于分层分区模型的分布式电源分布优化[J]. 华东交通大学学报,2013,30(2):85-89.
- [3] HUSAM SULEIMAN, KHAJA ALTAF AHMED, NAUMAN ZAFAR, et al. Inter-domain analysis of smart grid domain dependencies using domain-link matrices[J]. IEEE Transactions on Smart Grid, 2012, 3(2): 692-709.
- [4] 陶爽,袁越,傅质馨. 基于发电负荷需求系统评估方法的智能微网风险评估[J]. 现代电力,2013,30(2):7-12.
- [5] 陈星莺,陈楷,刘健,等. 配电网智能调度模式及其关键技术[J]. 电力系统自动化,2012,36(18):22-26.
- [6] 刘举平,余为清,叶江陵. 铁路作业库进出路自动信息提示系统设计[J]. 华东交通大学学报,2012,29(6):61-64.
- [7] 罗世民. 铁路发电车供电装置微机监控的实现方法[J]. 华东交通大学学报,2012,29(2):83-87.
- [8] 刘莉,屈志坚,王健. 基于SVG和Web Service的监控 图库关联组件柔性重构[J]. 电力系统保护与控制,2010,38(22):179-182,219.
- [9] 罗军舟,金嘉晖,宋爱波,等. 云计算:体系架构与关键技术[J]. 通信学报,2011,32(7):3-21.
- [10] 刘勇,乔秀全,李晓峰. 云计算环境下基于Mashup的一种电信网络能力服务提供模式[J]. 计算机科学,2012,39(1):32-36.
- [11] 吴建国. 基于电子商务协同平台的云计算模式分析[J]. 云南师范大学学报:自然科学版,2012,32(4):43-46.
- [12] 程苗. 基于云计算的Web数据挖掘[J]. 计算机科学,2011,38(10A):146-149.
- [13] 翟岩龙,罗壮,杨凯,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.
- [14] 屈志坚,陈剑云,陈秋林. 电气化铁道远动数据库系统业务流程的UML建模[J]. 电力自动化设备,2008,28(4):95-98.
- [15] 刘莉,屈志坚. 基于GridGain云计算平台的监控管理信息分解聚合研究[J]. 计算机测量与控制,2012,20(12):3235-3238.

(下转第91页)

- [6] 叶晓峰,胡媛媛. 非齐次空间上几类积分算子的有界性[J]. 华东交通大学学报,2012,29(4):12-18.
- [7] HU G, MENG Y, YANG D C. Multilinear commutators of singular integrals with non-doubling measures [J]. Integral Equation and Operator, 2005, 51: 235-255.
- [8] SAWANO Y. Generalized Morrey spaces for non-doubling measures. Nonlinear differ[J]. Equ Appl, 2008, 15(4): 412-425.
- [9] ZHENG T T, ZHANG S Y. Boundedness of a class of operators over generalized morrey spaces with non-doubling measures [J]. Journal of Zhejiang University of Science and Technology, 2011, 23: 01-05.
- [10] HU G, LIN H, YANG D C. Marcinkiewicz integrals with non-doubling measures [J]. Integral Equation and Operator, 2007, 58: 205-238.
- [11] NAKAI E. Hardy-Littlewood maximal operator, singular integral operators and the Riesz potentials on generalized Morrey spaces[J]. Math Nachr, 1994, 166: 95-103.

Boundedness of Marcinkiewicz Integral and Commutator on Non-homogeneous Spaces

Ye Xiaofeng, Wang Meng, Hu Yuanyuan

(School of Information and Computer Science, East China Jiaotong University, Nanchang 330013, China)

Abstract: Under the assumption that μ is a non-doubling measure on R^d , this paper studies the boundedness of Marcinkiewicz integrals and commutators generated by Marcinkiewicz integrals with $RBMO(\mu)$ functions on generalized Morrey spaces of non-homogeneous spaces. The result extends the conclusion of Y.Sawano.

Key words: generalized Morrey spaces; Marcinkiewicz integral; $RBMO(\mu)$ functions; commutators

(上接第41页)

Distributed Processing of Grid Monitoring Information Flow Based on Hadoop

Qu Zhijian, Guo Liang, Chen Ge

(School of Electrical and Electronic Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: In order to solve problems of large data sets access for monitoring information flow in smart grid dispatching system, this paper proposes a distributed cluster processing method based on Hadoop cloud computation framework for monitoring information flow. By analyzing information flow characteristics of power grid monitoring system, the three kinds of key processing stream were extracted; By comprehensively using distributed file system HDFS and the parallel programming model Map/Reduce, the cluster distributed information processing platform was established, and information efficient parallel processing was achieved. Taking power grid dispatching measurement logging data set as an example, the test results show that the new method processing efficiency is improved 1.4 times compared with traditional client/server database method, and is related to cluster node scale, which can avoid response delay problem of large data sets information processing.

Key words: smart grid; information flow; Hadoop cloud computation; Map/Reduce; parallel processing