电力设备状态高速采样数据的云存储技术研究

宋亚奇[1],刘树仁[1],朱永利[1,2],王德文[1],李 莉[1]
(1. 华北电力大学 控制与计算机工程学院,河北 保定 071003;
2. 华北电力大学 新能源电力系统国家重点实验室,北京 102206)

摘要:提出基于 Hadoop 和 HBase 的电力设备状态高速采样数据的存储方案,基于 MapReduce 设计实现了设备状态高速采样数据的并行查询方法。创建了 1 个包含 20 个节点(每个节点配置 4 核 CPU)的 Hadoop 集群,并对集群进行了基准测试,测试结果表明所建集群适合进行大量数据的读写。以绝缘子泄漏电流数据为例,使用 YCSB 对所建存储系统进行了性能测试,测试结果表明,Hadoop 和 Hbase 在存储容量、吞吐量以及查询延迟上提供了足够高的性能,能够满足智能电网状态监测数据可靠性及实时性要求。

关键词:电力设备;监测;采样;数据处理;云计算;Hadoop;HBase
中图分类号:TM 73 **文献标识码**:A DOI:10.3969/j.issn.1006-6047.2013.10.026

0 引言

智能电网的安全可靠,离不开对组成电网的设备的健康状况进行在线监测、数据搜集和评估。智能电网需要监测的设备众多,甚至包括线路的每串绝缘子的泄漏电流等动态信号。为了能对绝缘子放电等状态进行诊断,信号的采样频率必须在 200 kHz 以上。变电站内设备状态监测,要求数据采样率可达 MHz 级,变压器超高频局部放电信号的频率均在 300 MHz 以上,甚至超过 1 GHz。因此,采集的数据量非常巨大[1]。

目前,受存储容量以及网络带宽等限制,对电网状态监测数据的处理方式大多采用就地计算的方式,原始采样数据经过分析后,表征设备状态的相关数据接入到状态监测系统中,原始采样数据并未保存,这种就地处理的方式会导致如放电波形等重要信息丢失,影响电力设备状态评估的准确率。例如,在利用变压器局部放电信号进行故障诊断和状态评估时,已有方法大都利用波形宏观特征(熟数据)进行评估,而非常重要的放电过程波形(微观特征)被丢弃,影响诊断或评估的结果准确率。伴随设备硬件(存储容量和网络带宽)的改善,采集、传输并保存完整电力设备状态高速采样数据成为可能,因此,有必要研究电力设备状态高速采样数据的高效存储方法,为下一代数据中心存储电网设备的动态信号提供理论支持和技术储备。

收稿日期:2012-10-16;**修回日期**:2013-08-12
基金项目:国家自然科学基金资助项目(61074078);新能源电力系统国家重点实验室资助项目;中央高校基本科研业务费专项资金资助项目(13MS88)
Project supported by the National Natural Science Foundation of China(61074078), the Program of State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources and the Fundamental Research Funds for the Central Universities(13MS88)

常规的数据存储与管理方法,基础架构大多采用价格昂贵的大型服务器,存储硬件采用磁盘阵列,数据库管理软件采用关系数据库,紧密耦合类业务应用采用套装软件,因此系统扩展性较差、成本较高。传统的超级计算机[2-3]主要用于“计算密集型”的应用,如量子力学、天气预报等。超级计算机拥有多个处理器,通过精良的设计,达到高度并行的目的,实现快速计算,但是其计算需要的数据通常采用磁盘阵列进行集中式存储(RAID)。在 Yahoo! 集群[4]上的性能测试表明,Hadoop 分布式文件系统(HDFS)读写吞吐量在 Girdmix 测试中显示比 RAID 快 30%[5]。另外,超级计算机交互性较差,所采用并行编程方法(MPI 等)也难以掌握,对用户要求很高[6];系统的扩展性差,且成本高,不适用于智能电网环境下信息平台的建设。

云计算是分布式计算、并行计算和网格计算发展的结果,目前主要应用于“数据密集型”应用[7],通过虚拟技术、海量分布式数据存储技术、MapReduce 并行编程模型等技术,为用户提供高可靠性、高安全性的海量数据存储平台,这为智能电网信息平台的建设提供了全新的解决思路。

本文提出使用面向列的数据库 HBase 在开源的云计算平台 Hadoop 集群上实现海量电力设备状态高速采样数据的云存储方案,是采用云计算技术搭建智能电网信息平台的一次有益尝试。使用 TestDFSIO 和 YCSB 对集群整体输入、输出性能以及读取、插入、数据更新进行了性能测试,实验结果表明,Hadoop 和 HBase 在存储容量、吞吐量以及查询延迟上提供了足够高的性能,能够满足智能电网环境下电力设备状态高速采样数据可靠性及实时性要求。

1　相关工作

1.1　采样数据存储

电力设备状态高速采样数据是一种典型的时间序列数据。已有对时间序列数据的研究多基于内存数据库,主要关注分析和提取时间序列数据的模式,用于匹配或预测[8]。在时间序列数据的存储方面,传统的基于单机关系数据库管理系统受硬件的限制,存储性能无法满足高频率的采样数据存储速度要求,文献[9]提出建立 3 层文件系统,以特定格式文件的形式存储高速采样数据,在存储速度上达到了要求,但并未解决存储容量的问题。文献[10]针对光传输网管系统数据量急剧增长导致的网管数据库更新查询效率极低,甚至出现系统崩溃的问题,提出了一种分布式数据库存储方案,但其分布式系统采用的是 MYSQL 集群,其可扩展性较差,也没有涉及系统可靠性问题。有些文献采用压缩的方法实现时间序列数据的存储。文献[11]研究了时域和频域下时间序列数据的压缩方法,用于大规模数据存储,并支持对压缩数据的关联关系查询,但在查询性能方面无法满足在线监测系统实时性要求;文献[12]提出绝缘子泄漏电流的自适应 SPIHT 数据压缩,允许采样完一个工频周期的数据后就进行压缩,更适合实时或在线的场合,但其压缩目的主要是降低网络传输数据量,且无法对压缩数据直接进行查询;文献[13]研究了天文望远镜采集的 TB 级天文数据分布式存储方法以及在该数据集上实现的特征监测算法,但并未讨论数据存储的细节以及查询性能;文献[14]研究了大规模电能质量时间序列数据的存储与处理方法,但其采用的方法是通过降低采样率的方式实现的,只适合对采样率要求较低的系统。针对电力设备采样数据采样率高、数据量巨大,要求可靠性高、快速数据查询等特点,已有存储方案无法满足存储容量、数据写入速度、查询效率以及系统扩展性方面的要求。本文提出将 Hadoop 平台和 HBase 数据库用于电力设备采样数据的云存储方案以及基于 MapReduce 的并行查询方法,并通过一系列实验,验证了方法的可行性。

1.2　Hadoop

Hadoop[15]是 Apache 开源组织的一个分布式计算框架,支持在大量廉价的硬件设备组成的集群上运行数据密集型应用,具有高可靠性和良好的可扩展性。Hadoop 的系统架构如图 1 所示。

HBase[16]建立在 HDFS 之上,提供高可靠性、高性能、列存储、可伸缩、实时读写的数据库系统。它介于 NoSQL 和 RDBMS 之间,仅能通过主键(RowKey)和主键的 Range 来检索数据,仅支持单行事务,主要用来存储非结构化和半结构化的松散数据。

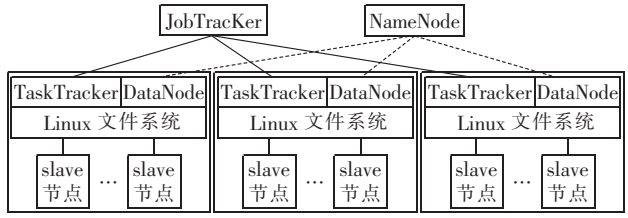

图 1 Hadoop 系统架构

Fig.1 System architecture of Hadoop

2　电力设备状态高速采样数据的云存储架构和实现方法

2.1　系统架构

本文参照云计算技术的体系结构,并结合电力设备采样数据的存储与业务应用需求,提出了如图 2 所示的基于云计算的电力设备采样数据存储系统。

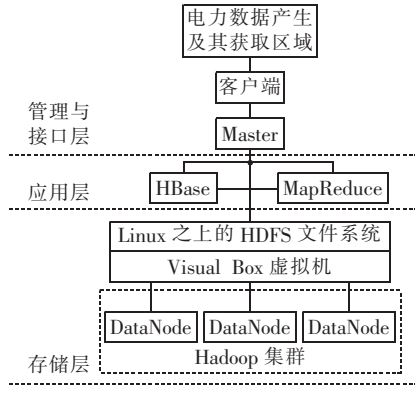

图 2 基于云计算的电力设备采样数据存储系统

Fig.2 Storage system based on cloud computing for sampled data of power equipment

存储系统分为以下 3 层。

a. 存储层为 Master 管理下的 Hadoop 集群,用于数据的物理存储。集群中的普通 PC 通过 Visual Box 虚拟化技术建立同质的 Linux 系统,并使用 Hadoop 平台建立 HDFS 文件系统。

b. 应用层包括基于 HDFS 文件系统的 HBase 和 MapReduce 编程模型,依据所提供的存储接口和并行编程接口完成数据存储以及应用开发。

c. 管理与接口层由数据产生区域、客户端和 Master 组成,可以通过客户端完成电力数据的云存储和实时访问。

2.2　电力设备采样数据描述

电力设备采样数据具有类似的格式,如图 3 所示。包含设备节点物理地址(唯一)、初始时标、产生通道、微气候记录(包括环境温度、湿度等)以及若干个周期长度的数据(默认值,在采样率固定的情况下每个采样点的时间都可计算)。根据设备物理地址可映射为具体的物理设备。

信息格式	物理地址,采集时刻,产生通道,微气候记录,固定间隔N个连续采样点数据(34528,31824,33248,32560,31856,…)

图3 电力设备采样数据信息格式

Fig.3 Message format of sampled data

2.3 HBase表结构设计

设计的存储系统可以对来自多台电力设备的采样数据进行同步存储(要求各采集设备的系统时钟进行同步),所设计的HBase表整体逻辑视图如表1所示。RowKey表示行关键字,用于采样数据检索,由Mac地址与路号的字符串连接构成。一个采集设备有可能采集多个通道,Mac地址唯一表示了采集设备,路号则表示了通道号。HBase表中每行数据都带有时标,表明了该数据的采集时间。与关系数据库表结构不同,HBase表的列定义由多个列族构成,每个列族又可以包含多个列,且列可以动态增加。设计了2个列族,分别描述采集时刻的微气候值(温度、相对湿度)以及采样数据采样点的值,分别对应表2中的Climate和leakage currents。

表1 采样数据HBase存储逻辑视图

Tab.1 Logical diagram of HBase storage for sampled data

RowKey	Time stamp	Climate		leakage currents			
		temperature	humidity	V_1	V_2	…	V_n
Mac1+id	201110090-201237000	20℃	65% rh	15	16	…	20
Mac2+id	201110090-201237001	21℃	64% rh	15	16	…	20
Mac3+id	201110090-201237002	20℃	63% rh	15	16	…	20

注:电流单位为mA。

2.4 基于MapReduce的并行关联查询设计

HBase索引只支持主索引(即RowKey),而电力设备状态监测系统很多应用场景中,采样数据的查询条件通常为多条件关联查询(如根据线路、杆塔、绝缘子ID查询绝缘子泄漏电流数据),这需要自行设计复合RowKey来满足多条件查询。本文设计了基于MapReduce的复合RowKey并行查询方法。

设Term1、Term2、…、Term$N$表示N个查询条件,将这N个条件连接在一起作为RowKey,用于唯一标识采样数据来源的Mac地址(采样数据存储表中的RowKey)。Mac地址为infor列族下的唯一一列,构建HBase表如表2所示。根据这些信息映射出采集设备的Mac地址以及路号id,进行采样数据表的查询。

表2 Mac地址映射表

Tab.2 Mac address mapping table

RowKey	Time stamp	Infor
		Mac
Term1+Term2+…+Term$N$	201204090201237000	Mac1+id
Term1+Term2+…+Term$N$	201204090201237001	Mac2+id
Term1+Term2+…+Term$N$	201204090201237002	Mac3+id

查询过程分为2个步骤,如图4所示。

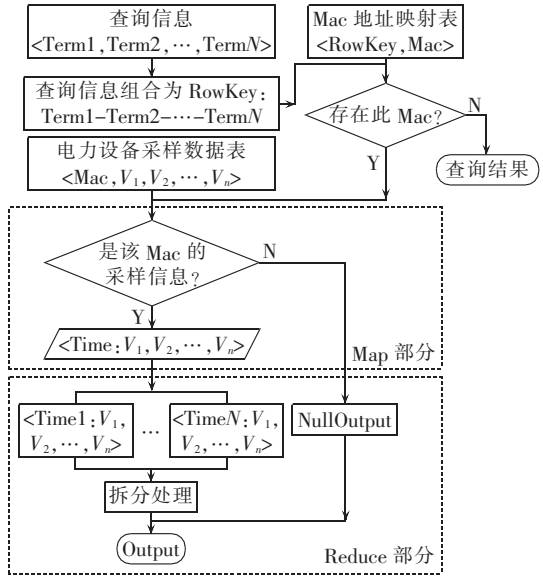

图4 基于MapReduce的泄漏电流并行关联查询

Fig.4 Parallel relational leakage current query based on MapReduce

步骤1:Mac地址查询。首先根据查询信息组合成RowKey查找出在HBase表中对应的Mac,考虑到该部分的数据属于静态信息,数据量较少,因此查询过程直接使用HBase的API进行,未进行并行化处理。

步骤2:采样数据查询。使用hbase.mapreduce包中的类,接收HBase表(电力设备采样数据表)和步骤1中查找到的Mac地址作为MapReduce作业的输入。HBase表在行方向上分成了多个Region,每个Region包含了一定范围内的数据。使用TableInputFormat类完成在Region边界的分割,Splitter(MapReduce框架的分割器)会给HBase表的每个Region分配一个Map task,完成RowKey在所属Region内的查询。在Reduce阶段,多个Map task查询的结果交由Reduce任务进行汇总,并根据设定的格式(TableOutputFormat类)对数据进行拆分,将结果写入Output里(可以是HBase表或者是文件等)。

3 Hadoop云计算平台搭建与基准测试

3.1 平台搭建

所搭建的Hadoop集群由20个节点组成,每个节点的配置为4核CPU(Intel Core i5),主频2.60 GHz,4 GB RAM内存,1 TB SATA 7200 r/min硬盘(64 MB缓存),配备千兆以太网用于集群节点的互联。虚拟机采用ORACLE VirtualBox(Version 4.1.8),配备操作系统Ubuntu(Version 10.04 LTS)。在集群上安装Apache Hadoop(Version 0.20.2)云计算平台,使用TestDFSIO[5]对集群的整体I/O性能进行了基准测试。测试程序用一个MapReduce作业对HDFS进

行高强度的 I/O 操作,测试集群整体的并行写入以及读取数据的性能。

3.2　集群基准测试

为验证数据总体规模、读写文件数量以及文件大小对集群 I/O 性能的影响,进行了 2 组实验,分别改变数据规模、文件大小、文件数量,获取运行时间,进行比较。

实验 1 逐渐增大数据规模(固定单个文件大小为 1000 MB,文件数量由 5 个增至 40 个),进行读、写操作,系统执行时间增长趋缓,平均访问时间有效降低,测试结果如表 3 所示。图 5、6 分别描述了文件数量变化对读、写操作运行时间的影响。由于数据处理规模增大,系统 Map task 数量有效增加,系统并行程度明显提高,节点间通信延迟所占比重降低,使得数据平均访问时间有效降低。因此,Hadoop 集群适合处理大数据量的读写。

实验 2 控制文件总量大小为 5 GB,文件数量从 5 个到 5000 个(文件大小从 1000 MB 到 1 MB)进行变化,图 7 描述了文件总量大小不变时文件数量与运行时间关系。当文件规模大于 1 块数据块(64 MB)时,总体访问时间随文件数量增加而减小;访问性能在文件大小与数据块(64 MB)相当时达到峰值;当文件数量继续增多时,访问时间随文件数量增长而增长。造成这种情况的主要原因是 NameNode 把文件系统的元数据放置在内存中,文件系统所能容纳的文件数目由 NameNode 的内存大小决定。每一个文件、文件夹和 Block 需要占据 150 B 左右的空间,当文件数量增多时,会消耗较多的 NameNode 内存,系统性能下降。另外,因为 Map task 的数量是由 Splits 来决定的,所以用 MapReduce 处理数量较多的文件时,就会产生过多的 Map task,线程管理开销增大,作业运行时间增长。因此,HDFS 不适合存储大量的小文件。在设计数据存储时,在考虑具体计算需求的基础上,应尽量使单个文件的大小和 HDFS 设置的块大小相近,减少整体文件数量,能够有效提升系统性能。

表 3　节点数为 20 时 HDFS 读写时间

Tab.3　Reading/writing time for 20 nodes

测试	文件数量	文件大小/MB	HDFS 字节数/MB	Map task 数量	运行时间/s
Write	5	1000	5000	5	215.992
Write	10	1000	10000	10	310.863
Write	15	1000	15000	15	466.489
Write	20	1000	20000	20	555.490
Write	25	1000	25000	25	643.196
Write	30	1000	30000	30	730.281
Write	40	1000	40000	40	1033.013
Read	5	1000	5000	5	139.618
Read	10	1000	10000	10	187.417
Read	15	1000	15000	15	206.387
Read	20	1000	20000	20	280.787
Read	25	1000	25000	25	341.929
Read	30	1000	30000	30	356.339
Read	40	1000	40000	40	435.794

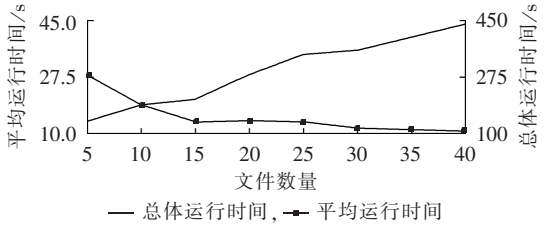

图 5　文件数量变化对运行时间的影响(读操作)

Fig.5　Influence of file number on running time(reading)

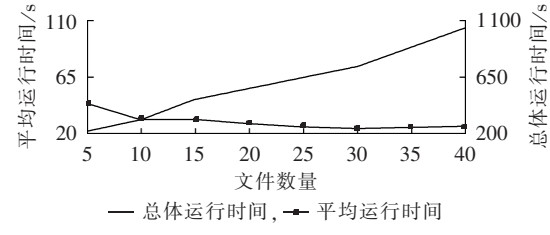

图 6　文件数量变化对运行时间的影响(写操作)

Fig.6　Influence of file number on running time(writing)

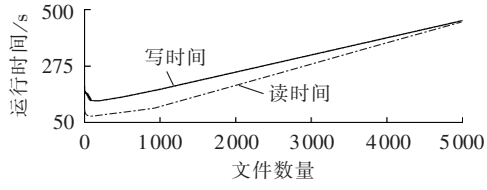

图 7　文件大小不变时文件数量与运行时间关系图

Fig.7　Relationship between file number and running time when total file volume is fixed

4　基于 YSCB 的数据存储性能测试

本文以输电线路上采集的绝缘子泄漏电流数据为例,在所搭建的 Hadoop 集群平台上,使用 YCSB[17] 对所提存储系统进行写入数据、读取数据的性能测试,验证所提存储方案的存储性能、查询性能。用于存储与查询测试的绝缘子泄漏电流数据采集自河北省某 110 kV 输电线路,线路包含 60 基监测杆塔。使用采集设备 246 个,对 246 个绝缘子以 200 kHz 采样率进行数据采集,并使用 16 bit 模数转换保存为物理量。共采集绝缘子泄漏电流数据 163 840 条,累计 5 120 MB。为验证大规模数据查询,对所采集数据进行了复制,总体规模达到 100 万条。

实验 **a**:数据导入测试。根据所设计的 HBase 存储模式(表 1),将 100 万条绝缘子泄漏电流数据(82.38 GB)导入 HBase 数据库。总体运行时间为 5 004.155 s、系统吞吐量达到每秒操作数 199.83,单条记录导入平均时延 49.311 ms。

实验 **b**:数据读、写测试。对已存储的绝缘子泄漏电流数据,定制 YCSB 客户端对 HBase 执行 10000 次读写操作,包括 50% 的读指令和 50% 的写指令,

用于验证当客户端产生非常频繁的数据更新请求时系统的性能。

实验**b**应用场景:在恶劣天气条件下,短时间内需要快速写入大量绝缘子泄漏电流数据,同时,对每条数据进行快速并行读取分析。重复进行5次实验,读写平均延迟变化情况如图8所示。5次实验的运行时间及吞吐量变化平缓,系统运行较稳定。5次实验平均运行时间为2851.167 ms,吞吐量为每秒操作数351.5176,单条数据平均插入时间小于0.3 ms,可以满足所提应用场景对存储系统数据读、写性能的要求。

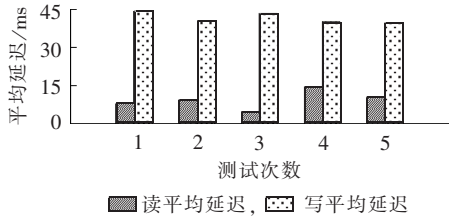

图8 5次实验读、写平均延迟(数据读写测试)
Fig.8 Average data reading/writing delay of five data reading and writing tests

实验**c**:数据读取测试。输电线路监测系统中,大多数应用场景都是对数据的多次读取操作,而写入操作通常只进行一次,很少进行数据更新。因此在电力设备状态监测系统中,存储系统的并行读取性能比写入性能更加重要。

对已存储的100万条绝缘子泄漏电流数据,定制YCSB客户端执行10000次读写操作,包括95%的读指令和5%的写指令。重复进行5次实验,图9描述了5次实验的单条指令执行的平均延迟。各次实验结果略有差异,主要原因是读、写指令操作的数据是随机选取的,如果随机读取数据的分布性较强,则读取时并行化程度较高,读取延迟较短;否则,若数据较集中,影响并行程度,则读取延迟较长。5次实验的运行时间及吞吐量变化平缓,系统运行较稳定。5次实验平均运行时间为1145 ms,吞吐量为每秒操作数836.9911,数据读取操作平均延迟远小于写入操作平均延迟。

目前的电力设备的状态监测尚处于起步阶段,

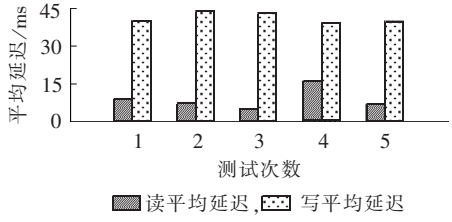

图9 数据读测试5次实验读、写平均延迟
Fig.9 Average data reading/writing delay of five data reading tests

系统性能要求方面尚未形成统一标准,在衡量系统性能时,可将SCADA系统作为参照。但SCADA系统实时性要求更高,在正常状态下窗口功能调用响应时间小于3 ms,数据加载时间小于5 ms[18]。实验中,读取10000条历史数据在2 s内完成,可以满足所提应用场景对存储系统数据读取性能的要求。

实验**d**:客户端并发线程数量对存储系统吞吐量以及访问时延的影响测试。定制YCSB客户端包含多个线程,线程数量由10个增至100个,对已存储的100万条绝缘子泄漏电流数据执行10000次的读取、更新请求,包括50%的读指令和50%的写指令。实验结果如图10所示。实验数据表明,所提存储系统吞吐量随客户端并发请求的线程数量增加而增长,当线程数量增至30时,系统吞吐量达到每秒操作数为415.09的峰值,之后,便随之下降,下降趋势逐渐趋于平缓,线程数量为100时,系统吞吐量下降为每秒操作数为362.18。这反映出所搭建的存储系统所能承受的并发访问数量。据此可知所提存储系统可以同时满足的在线用户数量。

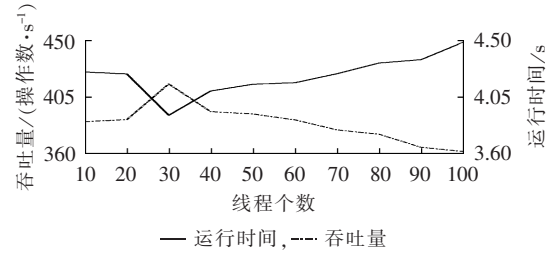

图10 客户端线程数量变化对存储系统吞吐量及访问时间的影响
Fig.10 Influence of client thread number on throughput and access time of storage system

在实验过程中,也遇到一些问题。如在首次数据导入过程中,当执行至60万条数据插入时(3831 s时),出现单个数据节点崩溃,于是终止导入操作;同时发现HBase负载并不均衡,该节点被装入的数据在崩溃前接近14 GB,且在崩溃前,该节点CPU利用率达到100%,而有的节点不到1 GB。造成这种现象的主要原因是没有对Java虚拟机进行垃圾收集(GC)的优化处理,以及Zookeeper(用于负载均衡)的配置问题。在测试10000条数据的读写性能的几个实验里,发现实验结果与2次实验间的时间间隔有关。比如实验**b**中,频繁实验(间隔很短)的平均运行时间为5493 ms,吞吐量为每秒操作数182;而有效增加2次实验的时间间隔后,平均运行时间为2966 ms,吞吐量为每秒操作数337.1。可以看出,频繁实验的过程中,集群完成数据读写后还有后续的内部操作,性能并不是最佳,一段时间后,集群的性能才达到最佳。

5　结语

本文提出了基于 Hadoop 和 HBase 的电力设备状态高速采样数据的云存储架构和实现方法,有效地实现了海量数据存储和快速并行查询,是采用云计算技术搭建智能电网信息平台的一次有益尝试。搭建具有 20 个节点的 Hadoop 集群,对集群的基准测试表明,随着数据规模逐渐增大,系统吞吐量有效提升,读、写操作平均访问时间有效降低,系统适合处理大数据量的读写。应用 HBase 设计实现了电力设备状态高速采样数据的分布式存储方法,以绝缘子泄漏电流数据为例,使用 YCSB 对所建存储系统进行了性能测试,测试结果表明,数据分布均匀,集群运行状况良好,没有出现数据丢失现象;系统在导入数据、读写测试、读测试中均提供了较高的吞吐量和较低的查询延迟。应用 MapReduce 编程模型设计实现了复合 RowKey 并行关联查询方法,解决了 HBase 目前的版本未实现多表关联查询的问题。本文所提存储方案在存储容量、查询延迟、数据可靠性和可扩展性等方面均显示了较高的性能,能够满足智能电网状态监测数据可靠性及实时性要求。

参考文献:

[1] 王德文,宋亚奇,朱永利. 基于云计算的智能电网信息平台[J]. 电力系统自动化,2010,34(22):7-12.
WANG Dewen,SONG Yaqi,ZHU Yongli. Information platform of smart grid based on cloud computing[J]. Automation of Electric Power Systems,2010,34(22):7-12.

[2] ONION A. Number-crunching GRAPE 6[M/OL]. (2000-06-02). http://grape.mtk.nao.ac.jp/grape/news/ABC/ABC-cuttingedge000602.html.

[3] San Diego Supercomputer Center. SDSC's 'Gordon' supercomputer:ready for researchers[M/OL]. (2012-03-05). http://www.sdsc.edu/News%20Items/PR030512_gordon.html.

[4] QI Runping. RAID vs. JBOD[M/OL]. (2009-01-14). http://markmail.org/message/xmzc45zi25htr7ry.

[5] WBITE T. Hadoop 权威指南[M]. 2 版. 曾大聃,周傲英,译. 北京:清华大学出版社,2011:260.

[6] 王鹏. 云计算的关键技术与应用实例[M]. 北京:人民邮电出版社,2010:8-9

[7] 于戈,谷峪,鲍玉斌,等. 云计算环境下的大规模图数据处理技术[J]. 计算机学报,2011,10(34):1753-1767.
YU Ge,GU Yu,BAO Yubin,et al. Large scale graph data processing on cloud computing environments[J]. Chinese Journal of Computers,2011,10(34):1753-1767.

[8] 张天成,岳德君,于戈,等. 数据流挖掘研究及其进展[J]. 小型微型计算机系统,2008,12(12):2241-2246.
ZHANG Tiancheng,YUE Dejun,YU Ge,et al. Research and development on data stream mining[J]. Mini-micro Systems,2008,12(12):2241-2246.

[9] 陈志诚,魏军,曾斌. 基于文件的高速采样数据存储系统设计[J]. 武汉理工大学学报,2006,28(8):72-74,90.
CHEN Zhicheng,WEI Jun,ZENG Bin. Design of a file-based high-speed sample data storage system[J]. Journal of Wuhan University of Technology,2006,28(8):72-74,90.

[10] 朱红霞,黄晓. 光传输网管海量数据存储访问研究[J]. 光通信研究,2011(6):19-21,51.
ZHU Hongxia,HUANG Xiao. Research on mass data storage and accession in optical transmission network management systems[J]. Study on Optical Communications,2011(6):19-21,51.

[11] REEVES G,LIU J,NATH S,et al. Managing massive time series streams with multi-scale compressed trickles[C]//The 35th International Conference on Very Large Databases. Lyon,France:VLDB,2009:97-108.

[12] 朱永利,翟学明,姜小磊. 绝缘子泄漏电流的自适应 SPIHT 数据压缩[J]. 电工技术学报,2011,26(12):190-196.
ZHU Yongli,ZHAI Xueming,JIANG Xiaolei. Adaptive SPIHT algorithm for data compression of insulator leakage currents[J]. Transactions of China Electrotechnical Society,2011,26(12):190-196.

[13] DAS K,BHADURI K,ARORA S. Scalable distributed change detection from astronomy data streams using local,asynchronous eigen monitoring algorithms[C]//Proceedings of the SIAM International Conference on Data Mining. Sparks,Nevada,USA:[s.n.],2009:247-258.

[14] 王学伟,王琳,苗桂君,等. 暂态和短时电能质量扰动信号压缩采样与重构方法[J]. 电网技术,2012,36(3):191-196.
WANG Xuewei,WANG Lin,MIAO Guijun,et al. An approach for compressive sampling and reconstruction of transient and short-time power quality disturbance signals[J]. Power System Technology,2012,36(3):191-196.

[15] Apache. What is Apache Hadoop[M/OL]. (2013-04-08). http://hadoop.apache.org.

[16] DougMeil. Hbase architecture[M/OL]. (2011-04-02). http://wiki.apache.org/hadoop/Hbase/HbaseArchitecture.

[17] COOPER B F,SILBERSTEIN A,TAM E,et al. Benchmarking cloud serving systems with YCSB[C]//Proceedings of the 1st ACM symposium on Cloud computing. New York,USA:ACM,2010:143-154.

[18] 刘东,闫红漫,丁振华,等. SCADA 主站系统集成测试技术研究[J]. 电网技术,2005,29(2):62-67.
LIU Dong,YAN Hongman,DING Zhenhua,et al. Research on integration testing technology for main station SCADA system[J]. Power System Technology,2005,29(2):62-67.

[19] KAUSHIK R T,BHANDARKAR M. GreenHDFS:towards an energy-conserving storage-efficient,hybrid Hadoop compute cluster[C]//Proceedings of the USENIX Annual Technical Conference. Boston,MA,USA:USENIX,2010:556-561.

[20] HUSAIN M F,DOSHI P,KHAN L,et al. Storage and retrieval of large RDF graph using Hadoop and MapReduce[M]. [S.l.]:Springer,2009:680-686.

[21] MYUNG J,YEON J,LEE S. SPARQL basic graph pattern processing with iterative MapReduce[C]//Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud. Raleigh,North Carolina,USA:ACM,2010:6-12.

[22] SHAFER J,RIXNER S,COX A L. The Hadoop distributed

filesystem:balancing portability and performance [C]//2010 IEEE International Symposium on Performance Analysis of Systems and Software. White Plains,NY,USA:IEEE,2010:122-133.

作者简介:

宋亚奇(1979-),男,河北保定人,讲师,博士研究生,研究方向为电力设备状态监测、分布式系统(**E-mail**:bdsyq@163.com);

刘树仁(1988-),男,甘肃白银人,硕士研究生,研究方向为计算机应用;

朱永利(1963-),男,河北冀州人,教授,博士研究生导师,主要研究方向为网络化监控及智能信息处理。

Cloud storage of power equipment state data sampled with high speed

SONG Yaqi[1],LIU Shuren[1],ZHU Yongli[1,2],WANG Dewen[1],LI Li[1]

(1. School of Control and Computer Engineering,North China Electric Power University,Baoding 071003,China;
2. State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources,
North China Electric Power University,Beijing 102206,China)

Abstract: A storage scheme based on Hadoop and HBase is proposed for the power equipment state data sampled with high speed. An approach is designed and implemented based on MapReduce to query in parallel the sampled data. A Hadoop cluster containing 20 nodes,each equipped with a 4-core CPU,is created and benchmarked. Results show it is suitable for massive data reading/writing. A storage system is built for the insulator leakage current data as an example and tested by YCSB,results indicate that,Hadoop and Hbase provide sufficiently high performance in storage capacity,throughput and latency,meeting the requirements of smart grid for the reliability and real-time performance of state monitoring data.

Key words: power equipment; monitoring; sampling; data processing; cloud computing; Hadoop; HBase

(上接第149页 continued from page 149)

Summary of integrated topology of EV traction system and battery charging system

LIU Ying,WANG Hui,QI Wenlong

(School of Electrical Engineering,Shandong University,Ji'nan 250061,China)

Abstract: Several EV(Electric Vehicle) integrated topologies proposed by foreign scholars are summarized, which make use of the hardware structures of drive system to reconstruct the charging system. Based on different types of drive system and their application conditions,the operational principles,advantages and disadvantages of different EV integrated topologies are analyzed,the corresponding control methods are discussed,and their basic problems and key common problems are further summarized. The EV integrated topology with high power & high power factor correction,harmonic compensation and security isolation will become the emphases of future research.

Key words: electric vehicles; motor driver; electric batteries; power factor correction; integrated topology