Research on Big Data Management, Analysis and Visualization of Smart Meter

CAO Wenzhong*, PENG Haoyue
(State Grid Chongqing Electric Power Company, Chongqing 400014, China)

Abstract: With the continuous expansion of power grid construction scale and the continuous advancement of power grid intelligence, smart meters are widely used in power grids at all levels, and the energy consumption data generated by smart meters bring great pressure to power grid data management. Therefore, a distributed data file storage system is constructed from three data dimensions of smart meter data, i.e., time, type of electric equipment and name of electric household, and a data query model is proposed according to data characteristics. Based on the power big data in the open source database, the big data of one million smart meters are visualized, and finally the data query delay and data transmission speed of memory and disk data access methods are compared. The research results show that the method proposed can effectively realize the query and visualization of smart meter big data, and the comparison results show that memory has better data access performance than disk.

Key words: smart meter; big data; data visualization; data management
EEACC: 7220; 8110B **doi**: 10.3969/j.issn.1005-9490.2023.02.041

智能电表大数据管理、分析和可视化方法研究

曹文忠*, 彭皓月
(国网重庆市电力公司, 重庆 400014)

摘 要: 随着电网建设规模的不断扩大, 以及电网智能化的不断推进, 智能电表正大量应用于各级电网, 其产生的能耗大数据为电网数据管理带来了较大压力。为此, 从智能电表数据的时间、用电设备类型、用电户名三个数据维度出发, 构建了一套分布式数据文件存储系统, 针对数据特征提出了一种数据查询模型, 基于开源数据库中的电力大数据, 对一百万个智能电表的大数据进行了的可视化, 最后对比了内存和磁盘两种数据存取方式的数据查询时延、数据传输速度。研究结果表明所提出的方法可有效实现智能电表大数据的查询、可视化, 对比结果表明相比磁盘, 内存具有更好的数据存取性能。

关键词: 智能电表; 大数据; 数据可视化; 数据管理
中图分类号: TP274+.2 **文献标识码**: A **文章编号**: 1005-9490(2023)02-0567-06

智能电表在不断发展的能源管理系统中发挥着重要作用。基于物联网的智能电表从住宅区家用电器读取能耗, 产生的数据通常表现出大数据的 3 V 特性, 即体积大、时变性和速度快[1-4]。现有数据表明在 15 min 的时间间隔内, 一百万个智能电表每年可以产生 400 TB 的数据量。智能电表在短时间间隔内读取产生的大量、不同格式的智能能源数据对电力公司的信息技术资源造成了压力[5-7]。因此智能电表大数据管理、查询和可视化一直是研究人员与运行部门关注的焦点。通过利用家庭的能源消耗数据, 电网公司可以获得与其能源计划密切相关的能源消耗生活方式和能源消费行为的重要信息。对于家庭用户来说, 设备级能耗的可视化将使业主更好地操作和管理家庭用电设备, 降低能源费用。为此研究人员在智能电表数据分析领域已经开展了大量工作, 包括负荷预测、异常数据检测、负荷调整策略和动态定价等[8-11]。然而目前消费者只能通过账单信息或独立的应用程序接收能耗信息, 这些数据具有离线和不可在线实时访问的特点。虽然电网公司为个人客户提供了基本的能耗分析服务, 但这些数据很难反映用户实时能耗变化, 也难以辅助用户调整用电设备的使用, 可见目前仍然缺少可扩展的、基于商用硬件的分布式基础设施来监控和管理智能电表数据。能源管理系统的主要作用是从智能电表收集数据, 并在智能家居的分布式系统中实施控制策略[12-15]。但现有的智能电表数据管理方式仅适用于处理有限的数千个用电户数据, 对于数万级别的大数据量仍然缺乏研究[13-14]。同时目前对于家庭、区县、市区、省级电网的实时用电量可视化和分析的研究较为有限, 还缺乏对于各种不同用户级别

收稿日期: 2021-08-30 **修改日期**: 2022-02-18

的数据结构模型和数据查询效率的相关研究与数据。为此本文基于 MIT 数据库建立了 1 百万个智能电表在一年内的用电数据集,设计了一套分布式数据文件存储系统,提出了一种数据查询模型,对这一百万个智能电表的大数据进行了的可视化操作,最后对比了内存和磁盘两种数据存取方式的数据查询时延、数据传输速度。研究结果可为电力大数据管理、分析和可视化提供参考。

1 数学模型

1.1 计算数据集

智能电表实时数据集通过 MIT 数据库网络在线获得,该数据集包含有 2019—2020 年间每 30 min 记录的 10 台家用电器的能耗。根据这些数据应用自回归移动平均(Autoregressive Moving Average, ARIMA)方法生成了一百万套房屋的时域能耗数据集。ARIMA 模型中假设变量的预测值是一些先前观测值和随机误差的线性函数,如式(1)所示:

$$Y_t=c+\phi_1 y_d+\phi_p y_d+\cdots+\theta_1 e_{t-1}\theta_q e_{t-q}+e_t \quad (1)$$

式中:ϕ_1、ϕ_2、θ_1、θ_2 是模型参数,p 和 q 分别是自回归数和移动平均项数的参数,y_d 是当前值和前一个值之间的 d 倍差,e_t 是时间 t 的随机误差,c 是常数。变量 d 表示用于给定时间序列数据集平稳化的差异程度。为了建立最佳 ARIMA 模型,通过选择优化模型拟合准则的一组(p,d,q)值进行搜索,拟合服从三个标准,即:Akaike 信息准则(Akaike Information Criterion,AIC)、Akaike 信息准则校正(Akaike Information Criterion Correction,AICc)和贝叶斯信息准则(Bayesian Information Criterion, BIC)。其中 AIC、AICc 和 BIC 的数值就是给定数据集的拟合优度。表 1 所示是本文得到的不同 ARIMA 拟合模型的统计结果,可见在各种 ARIMA 模型参数组合中,$(p,d,q)=(2,1,3)$具有最佳效果。

表 1 不同 ARIMA 拟合模型的统计结果

ARIMA 参数(p,d,q)	AIC	AICc	BIC
0,1,0	1 596.29	1 596.65	1 598.23
1,1,0	1 585.27	1 585.39	1 593.89
1,1,1	1 569.34	1 569.56	1 578.25
2,1,1	1 581.56	1 581.99	1 590.78
2,1,3	1 536.51	1 535.71	1 548.42

1.2 分布式文件系统集群设计

图 1 所示是智能电表的房屋大数据模型示意图,该模型用三个维度表示,即位置电表编号、用电时间周期和设备类型。图中 MH1、MH2、DH1 代表电表 id,时间主要是指一周中的具体天数,用电设备主要有不同类型的家电:例如加热器、冰箱、洗衣机、地窖、炉子、照明设备、其他等。如图 1 所示,立方体模型中的单元包含了每个维度的功耗值。

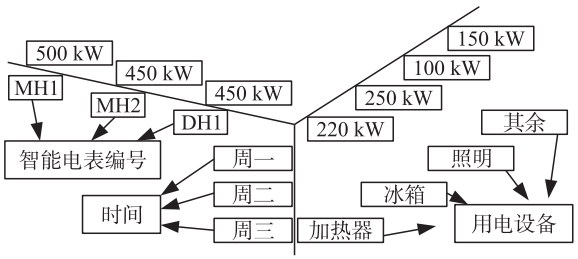


图 1 智能电表数据集的 OLAP 立方体

根据模型采用的数据形式,本文构建了一个 4 节点数据群,以分布式的方式存储生成的一百万个智能电表的数据。分别使用一个节点、两个节点、三个节点和四个节点进行数据处理。其中各个节点的功能如下所示:①主节点,主节点用于存储和管理集群的元数据,其本质是跨数据节点存储的文件块的目录,通过跟踪程序,将任务分配给从属节点。②从属节点,数据集群中的从节点称为数据节点,这些数据节点主要用于存储数据集,并在客户端应用程序执行查询时执行读/写操作。

本文使用四台基于英特尔处理器的计算机来建立由一个主节点和四个从节点组成的数据计算集群,该集群采用了以下配置参数:①每个节点的复制因子设置为 2,每个数据块大小为 64 MB。②将一百万个智能电表的数据集(1.5 TB),分割为 23 438 个数据块(1.5 TB/64 MB),将其分配给数据节点进行存储和处理。③每个数据节点存储 5 860 个数据块,其整体存储体积为 375 GB。

1.3 评估模型

本文计算模型采用的输入、输出和控制变量如表 2 所示。

表 2 模型采用的变量统计表

输入变量	控制变量	输出变量
节点数	核心	时延
数据文件	内存大小	传输速度
数据用户级别	网络贷款	

通过改变输入变量,由模型计算得到模型时延和数据传输速度。其中时延是跨数据节点呈现数据查询结果所需要的完成时间。数据传输速度表示单位时间内完成的读/写次数。对于输入变量,本文的智能电表数据文件采用 CSV 格式,以 10、100、1 000、10 000、100 000 和 1 000 000 的对数比例导入模型中。本文设计了两个实验来研究内存和磁盘缓存两个数据查询引擎对输出变量的影响。

通过测量每批数据的运行时间(时延)和数据传输速度,来确定智能电表在数据查询期间,数据量大小的影响。在多个阶段中执行数据查询,其中每个阶段包含由数个数据映射,如式(2)所示:

$$J=\{St_i:0<i<M\},St=\{T_sk_j,j:0<j<N\} \quad (2)$$

式中:M 是查询阶段数,N 是一个阶段中的任务数。T_s 是单个任务花费的时间;k_j 是阶段数。其中主节点的资源管理器将这些查询阶段分布在集群节点上。每个阶段内所有映射任务所花费的总执行时间如式(3)所示:

$$L=\sum_{i=0}^{M}St_i\sum_{j=0}^{N}T_sk_j \quad (3)$$

数据传输速度是指每次查询执行时每秒执行的数据量。数据传输速度取决于输入文件的大小和时延,如式(4)所示:

$$R=\frac{NF}{L}\times S \quad (4)$$

式中:R 是数据传输速度(MB/s),NF 是智能电表中的文件数量,L 是时延(s),S 是单个智能电表中数据文件的大小(MB)

2　结果与分析

2.1　大数据可视化

为了验证本文数据可视化效果,从生成的数据集中采集了 10 个智能电表的数据。对用户家庭中的十种能源消耗和不同时间周期的能源消耗情况进行了查询,得到的结果如图 2~图 9 所示。

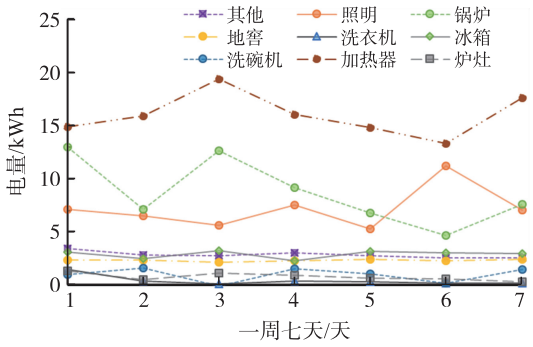


图 2　一周内每天消费者家中每个电器的总消耗量

本文考虑的数据用户级别共有四种:房屋业主、区县公司、市公司以及省公司。能源数据结果有助于消费者有效地监控家用电器能源;同时也有助于电力公司了解电网的状态。图 2~图 9 所示是不同时间周期、电网区域级别的数据结果,根据四个数据用户级别进行了分类。根据数据结果,房屋业主和各级电网运行部门就可以通过监控这些特定时间周期的可视化图形和图表,获得对不同家庭用户、区县电网、市级电网和省级电网的周期性能源消费趋势。图 2~图 4 是家庭用户在一周和一年内的能耗数据结果;图 5~图 6 是区县电网在一周和一年内的能耗数据结果;图 7~图 8 所示是市级电网下属两个区县的能源年消耗及总消耗量数据图表;图 9 所示是省内两个市在一周内的用电量数据图。

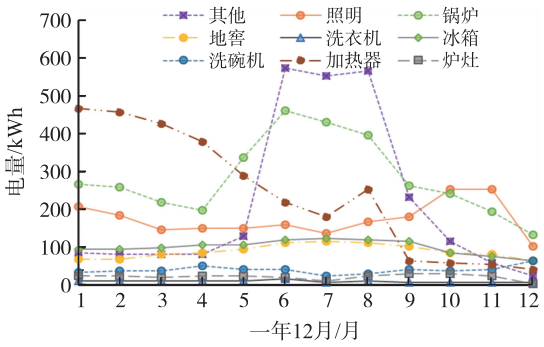


图 3　每月消费者家中每个电器的总消耗量

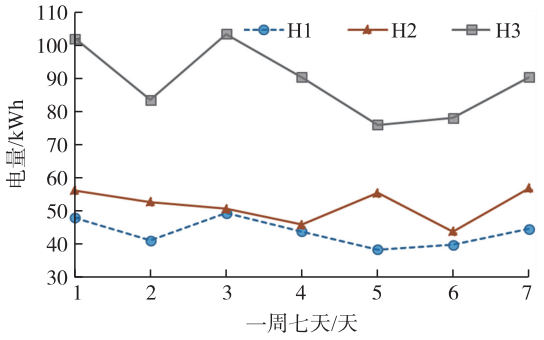


图 4　一周内一个社区中每栋房屋每天的总消耗量

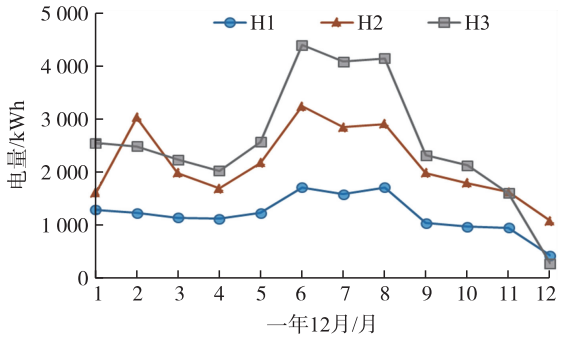


图 5　一年内区县电网每个月每套房子的总消费

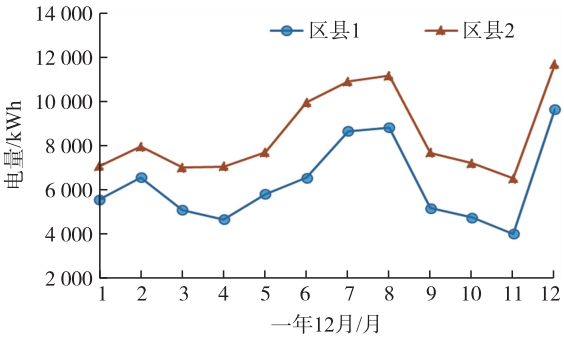


图 6　市级电网每个区县每月的总消费量

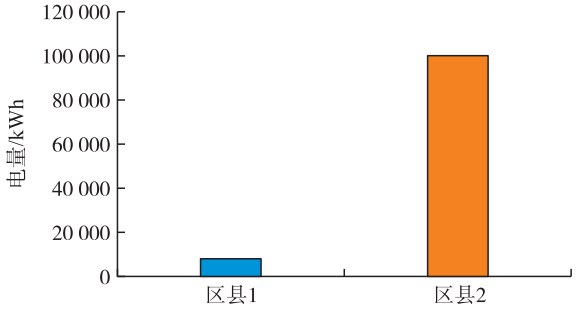


图 7 每个区县的年总消费量

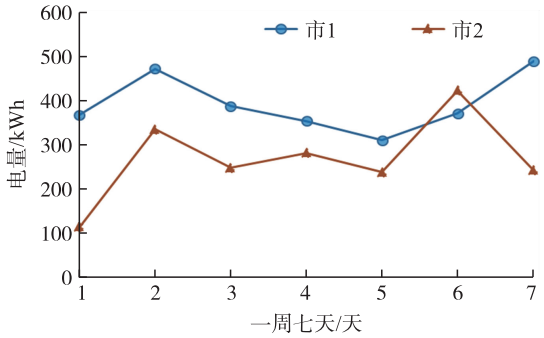


图 8 一周内省公司每个市每天的总消耗量

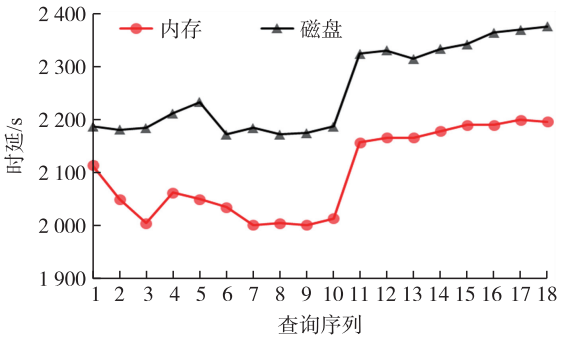


图 9 1 个节点的 100 万个智能电表的平均每次查询延迟

根据图 2~图 9 所示的结果可见,本文建立的数据结构和查询模型能够有效针对各个用户级别、各个时间周期内的能源消费情况进行查询显示。

2.2 数据查询系统

本节针对数据的查询时延以及数据传输速度问题,针对两种主要数据存取介质进行了的对比分析和研究:一种是采用磁盘缓存(主存储器),另一种是采用内存储存(辅助存储器)。图 10 中列出的结果是执行一百次查询得到的平均值。

2.2.1 查询时延

图 10 和图 11 所示是 2019—2020 年内,10 种家用电器每 30 min 的能耗,数据共包含一百万个智能电表的数据。

根据图中结果可见查询时间很大程度上取决于查询序列的选择性特征。随查询序列的增加时延增加。例如查询序列 11 花费了较长的执行时间,这是因为在该查询中使用了多个 JOIN 子句来查找总的家庭能源消费及其各自的区县能源消费。由于多 JOINs 的存在,文件之间有很强的相互依赖性,随着智能电表数据集之间的相互依赖性越来越大,将使得数据查询执行时间急剧增加,由图可见查询序列 15 到查询序列 18 花费的时间最长。很明显基于 4 节点的解决方案可以在不到 15 min 的时间内运行一百万个电表的复杂查询,通过添加更多的节点肯定会相应地减少这一时间。图 10 所示是利用四个节点实现了一百万个电表的最佳处理性能。因此对于处理更大批量的数据文件时,添加节点对减少执行时间有显著影响。图 11 所示是 1 至 4 个节点,不同电表数量执行数据查询的时延,可见从一个节点扩展到四个节点时,执行数据查询的时间减少了三分之一。由图 11 可见不管节点的数量如何,延迟都是线性增加的。例如对于两个节点的情况,将智能电表的数量从 100 000 增加到 1 000 000 时平均延迟仅仅从 1 400 s 增加到 1 800 s。此外在图 11 中,可见与内存相比,磁盘存储的斜率明显更高。这意味着随着查询数据量从 10 个智能电表文件增加到 1 000 000 个,采用磁盘进行数据查询的速度相比内存更慢。

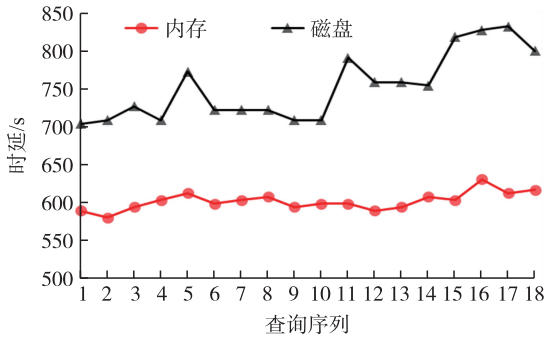
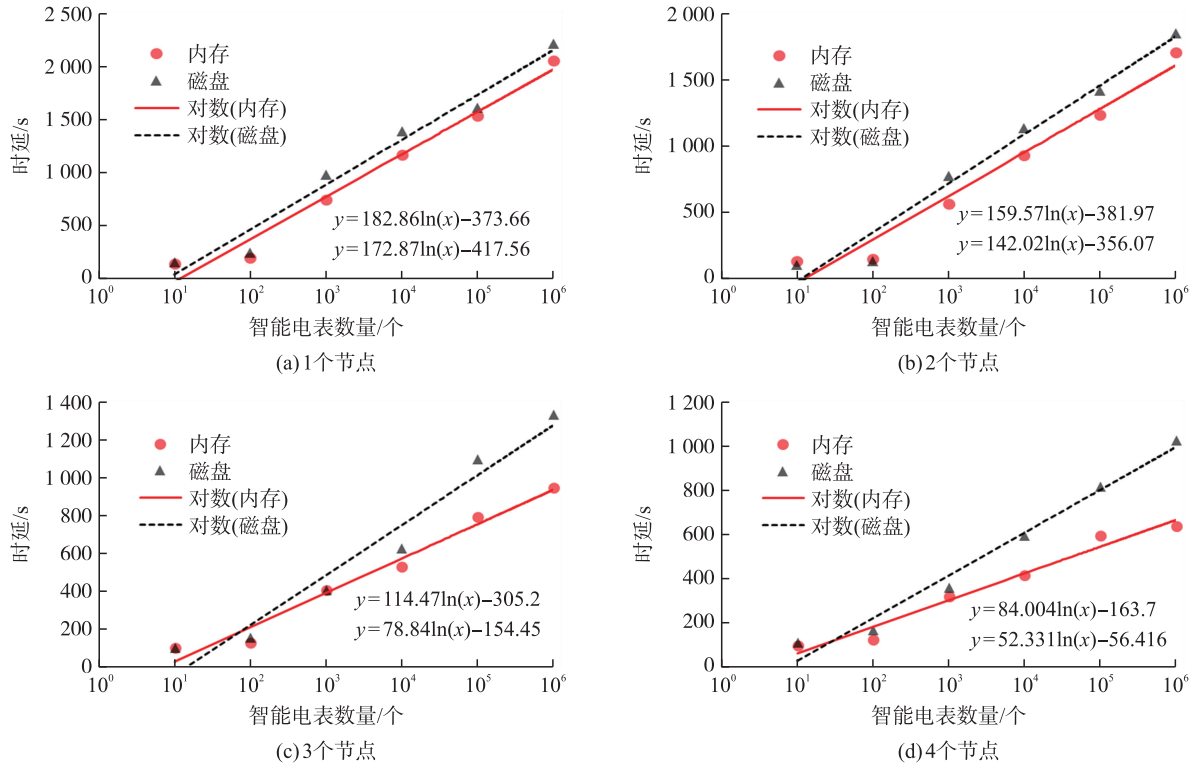


图 10 4 个节点的 100 万个智能电表的平均每次查询延迟

2.2.2 数据传输速度

数据传输速度是根据时延计算的,其中数据传输速度越快,时延也就越小。如图 12 显示了 10、100、1 000 和 10 000 个智能电表数据集在不同节点数条件下的数据传输速度。如图 13 所示是 100 000 和 1 000 000 个智能电表数据集的平均数据传输速度。

从图 12 和图 13 可以看出,与单节点相比,4 节点的数据传输速度更高。从图中可得,1 000 000 个智能电表的大数据集的斜率比 10 个智能电表的数据集的斜率大。因此与较小的数据集相比,较大数据集的



(a)1个节点　(b)2个节点　(c)3个节点　(d)4个节点

图 11　1 节点、2 节点、3 节点和 4 节点的平均延迟

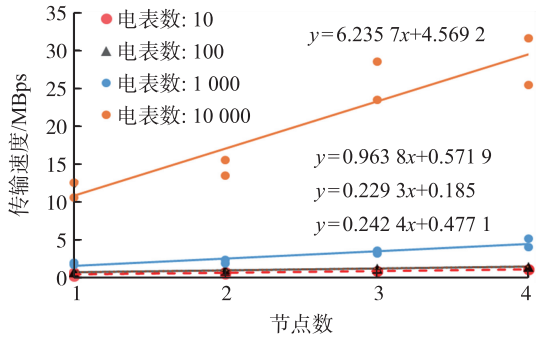


图 12　1、2、3 和 4 个节点上 10、100、1 000、10 000 个智能电表数据的平均吞吐量

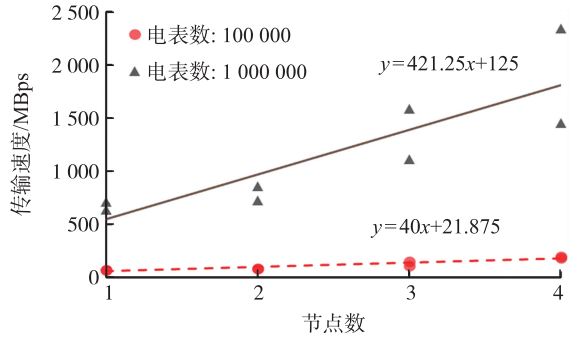


图 13　1、2、3 和 4 个节点上 100 000 和 1 000 000 个智能电表数据的平均吞吐量

数据传输速度随着节点数量的增加而增加得更快。此外对于在 4 节点中查询的一组一百万个智能电表数据,内存处理的最大数据传输速度约为 23 Mbit/s,而基于磁盘的处理的数据传输速度为 1 400 Mbit/s。这是由于增加了四个节点,在相同数量的智能电表数据集上执行查询的处理时间大大减少。此外对于相同的节点集和相同的处理引擎类型,随着数据集大小的增加,数据传输速度以单调递增的方式增长。这一结果与理论预期相符,因为与一个节点相比,四个节点的延迟更小,因此数据传输速度更高。

总体而言,内存缓存的数据传输速度优于磁盘缓存,因为前者在查询一百万个智能电表数据集时提供了性能提升。当输入数据存储在内存中时,数据传输在很大程度上取决于迭代查询方法。因此与磁盘缓存相比,内存缓存更具优势,然而在非迭代查询中,内存和磁盘缓存处理的查询性能是相当的。

图 14 所示是内存和磁盘两者的查询时延与迭代次数的关系,可见二者第一次查询运行的执行时间彼此相当。然而在迭代查询过程中,内存比磁盘

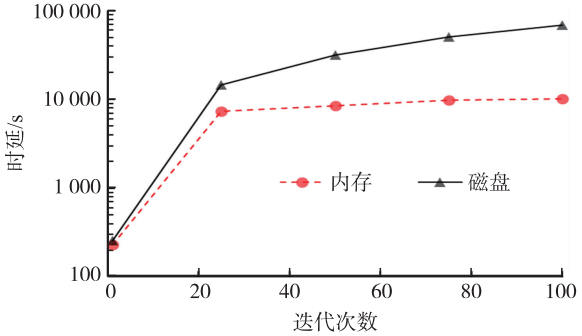


图 14　4 节点一百万个智能电表的内存和磁盘迭代查询性能

缓存的迭代性能好近 10 倍。这是因为在第一次迭代中,内存引擎将数据提取到内存中,并在后续迭代中利用缓存的内存数据,从而缩短了响应时间。另一方面,磁盘在每次迭代中将数据连续溢出到磁盘,因此不会显著减少后续查询迭代的响应时间。

实际工况中选择合适的大数据处理引擎至关重要。由本文分析结果可见,尽管内存缓存在迭代查询中提供了内存计算,有助于降低数据查询延迟,但应该考虑内存限制和网络带宽限制。如果可用的硬件资源有限,建议使用磁盘进行处理和分析。对于在小集群上处理的小体量数据集,磁盘是获得稳定查询响应而没有任何内存不足问题的好选择。此外在磁盘或内存中写入几个 MBs 的执行时间没有太大差异,因此建议小数据集查询时使用磁盘缓存,在处理大中型数据集时,则建议使用内存。

3 结论

针对智能电表应用背景下,大数据的查询、管理、分析和可视化的问题,建立了一套数据结构查询系统,提出了数据查询模型,实现了能源消费大数据查询和可视化,得到了以下几个主要结论:①采用本文建立的数据文件框架和模型,可有效针对家庭用户、区县电网、市区电网以及省级电网,在不同时间周期内的用电数据进行可视化展示和查询。②数据查询时延的对比分析结果表明,这意味着随着查询数据量从 10 个智能电表文件增加到 1 000 000 个,采用磁盘进行数据查询的速度相比内存更快。③数据传输速度的对比研究结果表明,内存的数据传输速度优于磁盘,原因是内存缓存在迭代查询中提供了内存计算,有助于降低数据查询延迟。实际工况中建议小数据集查询时使用磁盘缓存,在处理大中型数据集时则建议使用内存。

参考文献:

[1] Gao L C,Yao L M,Yang Z W,et al. Research on Hybrid Index Method of Double-LayerB+Tree for Power Big Data Considering Knowledge Graph[J]. Journal of Physics: Conference Series,2021,1771(1):012004.

[2] 徐长辉,吴永云,昌云松,等. 电表智能读写嵌入式终端设计与实现[J]. 企业科技与发展,2021(1):38-40.

[3] 刘晓华,高俊成,王少林,等. 智能电网异常用电检测框架[J]. 自动化技术与应用,2020,39(12):124-128,160.

[4] 徐晓东,吕干云,鲁涛,等. 基于智能电表数据与模糊 C 均值算法的台区识别[J]. 南京工程学院学报(自然科学版),2020,18(4):1-7.

[5] 孙毅,毛烨华,李泽坤,等. 面向电力大数据的用户负荷特性和可调节潜力综合聚类方法[J]. 中国电机工程学报:2021,41(18):6259-6271.

[6] 袁智勇,肖泽坤,于力,等. 智能电网大数据研究综述[J]. 广东电力,2021,34(1):1-12.

[7] 杨春波,凌松. 基于大数据框架的智能电网分析和可视化应用[J]. 电子器件,2020,43(5):1004-1009.

[8] Liu S, Ji X, Zhao X L, et al. Equipment Operation Analysis and Application Research Based on Big Data Electric Power Distribution Network[J]. Journal of physics. Conference Series,2020,1631(1):012133.

[9] 毛先胤,文屹,马晓红,等. 基于电力大数据的多源异构数据融合技术研究与应用[J]. 电力大数据,2020,23(8):33-39.

[10] 叶海峰,陈存林,张炜,等. 用于智能电表数据分析的大数据平台[J]. 信息技术,2020,44(1):134-139.

[11] Li J,Zhang S C,Zhang Z J,et al. The Application of Big Data in Electrical Power System[J]. International Journal of Power and Energy Engineering,2019,1(12):46-48.

[12] 李石,赵苏虹. 大数据在智能电网中的应用研究[J]. 电工技术,2019(24):126-127,134.

[13] 胡志朝. 基于电力物联网的数据可视化分析[D]. 温州:温州大学,2019.

[14] 高欣,刁新平,刘婧,等. 基于模型自适应选择融合的智能电表故障多分类方法[J]. 电网技术,2019,43(6):1955-1961.

[15] 张令涛,赵林,张亮,等. 配用电大数据分布式计算集群负载均衡框架[J]. 电网技术,2019,43(1):259-265.

曹文忠(1975—),男,重庆市丰都人,学士学位,高级工程师,主要研究方向:数据管理,数据接入、数据质量、数据应用等研究。