

大数据分析在用电采集数据分析与智能监测系统的设计与实现

为了加强对配电网的设备工况、计量装置和用户用电特征的监测与分析,用电采集数据分析与智能监测系统以用电信息采集系统为基础,整合相关外部数据及应用,基于Hadoop分布式框架和Mapreduce并行计算架构,将RDBMS与HBASE、HDFS相结合,针对不同的业务场景和数据结构,研究合适的数据挖掘工具,通过设计不同的业务分析模型,实现对配电网运行工况的在线监测和异常智能诊断功能,及时发现配电网运行中的薄弱环节,制定降损节能措施,提升供电质量,提高配电网运行的经济性和可靠性。

李珏¹ 马晓东¹ 张果谋² 吴少雄² 王军²/1.国网苏州供电公司
2. 国网电力科学研究院

0 引言

随着电力用户需求的持续提升以及城市范围不断扩大,部分地区配电网建设与经济社会发展不相适应。同时配电网存在点多面广,用电环境比较复杂等情况,需要采取现代化的监测、监控手段。用电采集数据分析与智能监测系统(以下简称用采分析与监测系统)以用电信息采集系统在配电网采集的海量数据为基础,整合营销业务系统、GIS系统、PMS等外部系统及利用营配集成应用成果,针对配电网的设备运行状态、负荷特征、电能质量、供电可靠性、线损分析、装置运行工况进行在线监测及对用户用电行为智能分析。

用电信息采集系统作为用采分析与监测系统的主要数据源,是集现代数字通信技术、计算机软硬件技术、电能计量技术和电力营销技术为一体的综合的准实时信息采集与分析处理系统。它以移动通信网络、230 MHz无线专网、光纤

专网为主要通信载体,通过RS 485、载波、微功率无线等多种通信方式实现系统主站和现场终端之间的数据通信,具有数据采集、远程抄表、费控管理、线损分析和负荷监控管理等功能。支持对专用变压器终端、集中器以及带通信功能电能表的直接采集^[1-4],采集对象包括A~F共6类用户和变电站^[5],覆盖了配电网所有线路、台区和各类用户。

1 技术架构

1.1 大数据分析架构

基于用采分析与监测系统的大数据分析,从其系统架构包括以下几个重要部分:数据采集、数据预处理(数据准备)、数据存储及管理、数据计算与分析、数据展现和应用(包括数据检索、数据可视化、数据应用、数据安全等)^[6],系统应用中的大数据分析架构,如图1所示。

从上面的架构图可知,系统中的数据主要来自于用电采集系统、营销业务系统和PMS及GIS等,这些数据在准备阶段经过辨析与清洗操作,根据数据的格式、用途、特性存储在传统的RDBMS或非结构化分布式环境中,然后依据处理数据的类型与分析的目的,研究对应的算法模型,进而提高对数据处理的准确性,并以直观的方式将分析结果呈现给用户,同时应用于公司经营的辅助决策中。

在数据分析过程中,对于基于MapReduce的数据挖掘可采用Mahout算法库,而基于RDBMS的数据挖掘可使用weka算法库(Waikato Environment for Knowledge Analysis); Mahout 是机



李珏/工程师

关键词/Keywords

大数据分析 ·
用电信息采集 ·
Hadoop ·
Mapreduce ·
RDBMS ·

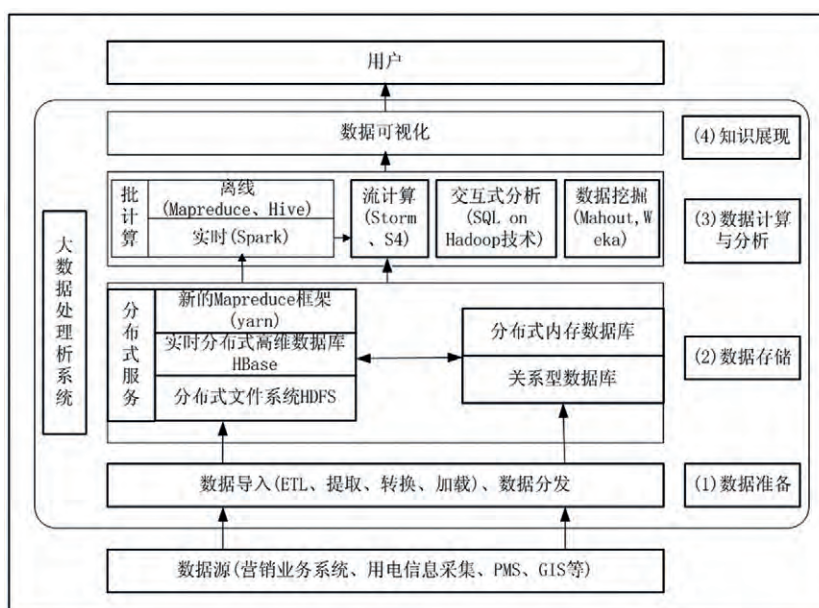


图1 大数据分析架构图

器学习和数据挖掘的一个分布式框架，它是基于hadoop之上并利用MapReduce实现了部分数据挖掘算法，解决了并行挖掘的问题。而weka基于JAVA环境下开源的机器学习以及数据挖掘软件，weka提供一组通用API，可以嵌入到应用程序以完成诸如服务器端自动数据挖掘这样的任务。在本架构中，将weka与Mahout相结合，根据数据的特征、存储的方式以及用户需求，共同对用电采集及其整合后的数据进行分析挖掘。

1.2 关系数据库与非关系数据库特性分析

关系型数据库作为主流数据库被广泛应用。其优点如二维表结构的存储方式容易理解、SQL语言使得操作极为简便等，其丰富的完整性大大降低了数据冗余和数据不一致的概率。而RDBMS拥有完整的约束及丰富的索引，因此更适合处理有关结构化数据复杂查询及在线分析的等任务。面对着大数据时代的到来，由于受到硬盘I/O的制约，以及扩展困难、多表查询复杂耗时、对大表操作效率低下等因素的影响，关系型数据库很难在海量非结构化数据的存储管理上得心应手。

非关系型数据库（HBASE）是一个非关系型、高可靠性、高性能、面向列、可伸缩的分布式存储系统，且一般不保证遵循ACID原则的数据存储系统^[7]。非关系型数据库提出另一种理念，例如，以键值对存储，且结构不固定，这使得存储时不会局限于固定的结构，同时也可以减少一些

时间和空间的开销。

Mapreduce是一种通过大规模廉价服务器集群实现大数据并行处理的分布式计算模型，它将扩展性和系统可用性放在了优先考虑的位置，其核心就是将任务分解成小任务由不同的计算者同时参与计算，并将各个计算者的计算结果合并得出最终的结果^[8]。模型本身非常简单，一般只需要实现两个接口即可（Map和Reduce），即Map阶段和Reduce阶段。Map函数处理Key/Value对，产生一系列的中间Key/Value对，经过混洗阶段将Map阶段产生的键值对进行处理整合后，把同一键值的组合交由同一Reduce阶段处理，Reduce函数用来合并所有具有相同Key值的中间键值对，计算最终结果。

1.3 数据的特性分析

用采分析与监测系统分析的数据种类包括结构化和非结构化数据，结构化数据包括存储在ORACLE或其他RDBMS中的数据，包括档案、采集数据、统计数据和报表数据等；非结构数据存在HBASE、HDFS中，包括现场计量装置和采集装置接线图、SVG图、GIS信息及通信日志文件等方面的数据。运行监控需要的数据是实时数据，需要快速而准确地判断和定位处理；智能分析与诊断除了需要原始数据、模型信息、统计数据外，还需日志、图形化信息等非结构化的数据可以作为准实时数据处理。

系统在分析台区和线路停电事件时，针对台

区停电监测根据终端停电事件,分析配变台区的负荷特征是否满足停电事件判断规则;针对线路停电事件,则是在台区停电的基础上,结合线路的配变台区的负荷与用电监测,分析线路下台区停电事件的分布情况。系统将台区和线路停电事件提供给供电部门作为抢修依据。由于台区和线路停电监测所涉及到的数据在实时性与操作复杂度上都存在差异,因此台区停电用HBASE来对数据进行存储,用Mapreduce进行快速分析比较合理;而线路停电因为涉及设备档案关系,因此采用RDBMS进行存储,通过库内分析比较合适。

2 业务设计

2.1 业务模型设计

用采分析与监测系统建立业务分析模型库,根据用户和配电网运行监测的需求设计七类模型,分别是市场开拓与业扩报装辅助分析、多维度线损分析、电能表整体运行状态分析、采集运维质量预控与工作效能分析、用户电费回收风险分析、用户用电行为分析。各个业务模型包括数据源、算法库、分析流程和信息发布与知识集成。

2.2 应用场景分析流程

本节以市场开拓与业扩报装辅助分析为设定场景,结合本文提到的大数据分析的思路,对其进行分析流程的设计:①获取新装或增容用户的变压器容量、用电负荷特征、工作班次、电压等级、供电地址等信息;②在营配集成的基础上,通过GIS定位合理的供电点位置,并定位对应的线路、开闭站(或开闭所)、厂站;③根据按厂站、线路统计的各维度负荷、电量统计数据,结合营销业务系统销户、报停和减容业务流程,及PMS的电网模型信息,通过引入大数据分析的方法,评估新增供电点所在的线路、厂站的负荷和电量变化趋势、负荷特征及供电质量是否满足用户用电的需求。具体流程如图2所示。

2.3 应用场景分析算法

业务报装辅助分析流程在实际应用中选择朴素贝叶斯分类、决策树、负荷趋势预测算法应用到实际业务中。分类算法用于按电压等级分析负荷特征,负荷趋势算法用于分析厂站、线路及大用户负荷变化趋势。通过决策树分析确定新装用户合适的供电点位置 and 用户增容的可行性。

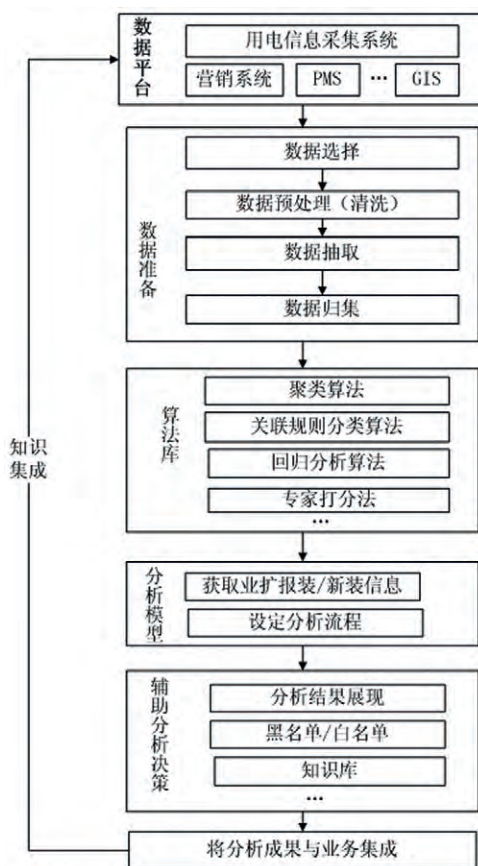


图2 业扩报装分析流程图

针对分类算法,本文采用较为直观的决策树方法,决策树(Decision Tree)是数据挖掘分类方法中最常用的方法之一,能够以图形化的形式表现挖掘的结果,从而方便使用者快速作出决定或预测。决策树是一个类似流程图的树形结构,每个内部节点表示在某个属性上的测试,每个分支表示一个测试输出,每个叶节点代表类或类分布。

决策树中经典的算法包括ID3与C4.5算法,其中C4.5算法是由J. R. Quinlan提出的^[9],该算法针对ID3算法执行过程中每次分裂都会倾向于选择有较多取值的属性和只能处理离散型数据这两个缺点提出的,和ID3算法相比,在选择分裂属性上没有采用信息增益,而是采用信息增益率作为标准。C4.5算法首先定义了分裂信息,其定义可以表示成

$$\text{split_inf}_{o_A}(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|}$$

信息的增益率被定义为

$$\text{gain_ratio}(A) = \frac{\text{gain}(A)}{\text{split_inf}_{o_A}(A)}$$

另外,针对ID3算法不能处理连续型数据的

弱点, C4.5算法通过一定的算法将连续型属性的值分成不同的区间, 即完成“离散化”操作。离散化时, 首先使用快速排序算法对属性值进行排序, 然后计算信息增益, 并确定局部阈值, 再用顺序查找的方法找出最合适的实际分割阈值。具体步骤如下^[10]:

1) 将样本集合 T 按照连续属性 A 的值按升序排列, 假设排序后属性 A 的取值序列为 $\{v_1, v_2, \dots, v_m\}$, 其中 $m=|VA|$

2) 对所有的 $v_i, 1 \leq i \leq m-1$, 按照值 $v=(v_i+v_{i+1})/2$ 将 T 划分为两个子集: $T_1=\{v_j | v_j \leq v\}$ 和 $T_2=\{v_j | v_j > v\}$, 然后计算按 v 划分所得到的信息增益 $\text{Gain } v$ 。

3) 找出使得 $\text{Gain } v$ 最大的 v , 记为 $v_{\max}$, 称之为局部阈值。

4) 在序列 $\{v_1, v_2, \dots, v_m\}$ 中顺序查找, 找出最接近 $v_{\max}$ 但又不超过它的 v_i , 将 v_i 作为 A 的分割阈值, 完成对 A 的离散化过程。

利用上述算法, 按供电单位、统计时间对用户或用户群体及用户所在配网厂站、线路的年月日负荷特性和用电特征等进行分类分析; 在此统计数据的基础上, 针对新装用户的申报容量, 及需增减容用户在过去持续一段时间内负荷变化及运行容量的关系, 分析配网线路和增减容用户短期和中期负荷和电量的增长趋势。为业扩报装提供辅助分析。

3 系统实现

用采分析与监测系统实现了以下主要功能:

①通过建立异常指标库和监测模型实现对配电网和用户异常点或故障点分析和判断功能; ②通过大数据技术, 结合关联异常判断, 实现对用户违约用电窃电的异常分析、线损与用电异常分析、停电事件智能分析, 为现场快速消缺、状态检修和用电检查提供参考依据; ③通过不同异常之间的关联规则分析, 加强了电力公司对配变台区和配网线路的轻载、重载、过载运行工况及电能质量和供电可靠性监测与分析能力, 提升了电力公司对配电网运行管理的水平。

4 结束语

本文通过分析RDBMS与HBASE在用采分析

与监测系统的应用场景, 以及电力行业大数据的特点, 提出了利用RDBMS与HBASE+Mapreduce相结合的方式来构建适合用采分析与监测系统的大数据分析架构, 该架构设计清晰, 处理流程合理, 能够根据模型实现在线分析, 为大数据技术在电力行业的应用提供了新的思路。

参考文献

- [1] 刘振亚. 智能电网技术[M]. 北京: 中国电力出版社, 2010.
- [2] 范洁, 陈霄, 周玉. 基于用电信息采集系统的电能计量装置异常智能分析方法研究[J]. 电测与仪表, 2013(11):4-9.
- [3] 电力行业标准. DL/T 645—2007, 多功能电能表通信协议[S].
- [4] 肖坚红, 严小文, 周永真, 等. 基于数据挖掘的计量装置在线监测与智能诊断系统的设计与实现[J]. 电测与仪表, 2014, 51(14):1-5.
- [5] 国家电网公司企业标准. 电力用户用电信息采集系统功能规范[S]. 北京: 中国电力出版社, 2013.
- [6] 中国电机工程学会电力信息化专业委员会. 中国电力大数据发展白皮书[M]. 北京: 中国电力出版社, 2013.
- [7] 覃雄派, 王会举, 杜小勇, 等. 大数据分析——RDBMS与MapReduce的竞争与共生[J]. 软件学报, 2012, 23(1):32-45.
- [8] Gunarathne T, Wu T, Qiu J, et al. MapReduce in the clouds for science[C]. 2010 IEEE Second International Conference on Cloud Computing Technology and Science(CloudCom), IEEE, 2010:565-572.
- [9] 谢金梅, 王艳妮. 决策树算法综述[J]. 软件导报, 2008, 7(11): 83-85.
- [10] Quinlan J R. C4.5: programs for machine learning[M]. San Francisco: Morgan Kaufmann, 1993.

(收稿日期: 2015-03-30) EA

作者简介

李珏(1981-), 男, 工程师, 硕士, 从事运营监测研究。

马晓东(1971-), 男, 高级工程师, 硕士, 从事运营监测与分析研究。