

基于大数据技术构建预警平台的方法研究

王 芳¹, 赵俊峰²

(1.苏州供电公司, 江苏 苏州 215004; 2.江苏省电力公司信通分公司, 江苏 南京 210024)

摘 要:随着信息化建设的不断深入推进, 苏州供电公司信通分公司所运维的数据达到 TB 级别, 以后还会越来越多。随着大数据概念的盛行, 海量的数据能为我们带来什么样的好处, 什么样的冲击, 是我们必须提前谋划并解决的问题。本文介绍了信息设备和信息系统预警平台的实施背景、大数据的建模、分析方法, 为实践大数据应用建设进行了有益的探索。

关键词:大数据; 设备; 信息系统; 预警

0 引言

大数据又称海量数据, 之所以产生, 是因为今天无处不在的移动终端、传感器、微处理器, 以及发达的无线网络。所有的电子或机械设备使用后都可以留下状态信息, 或者地理位置信息、性能参数信息。这些设备和它的使用者, 通过互联网的交互, 就形成了一个海量的数据源。

大数据的概念如汹涌浪潮来袭, 势不可挡。大数据与互联网的出现一样, 不仅仅是信息技术领域的革命, 更是在全球范围内加速企业创新、政府阳光、引导社会变革的利器。大数据及其分析, 将会在未来 10 年改变几乎每一个行业的业务功能。任何一个组织, 谁能早一步着手开始大数据的工作, 谁就可以获得明显的竞争优势。中国电机工程学会电力信息化专委会提出 2013 年为大数据元年。

在江苏省电力行业内部, 随着企业信息化建设的不断推进, 信息化水平的不断提升, 信息通信分公司运维的数据总量已达到 300 多 TB, 服务器上千台, 终端数十万台。随着一流配电网、灾备系统等应用的建设, 公司信息化程度的进一步深入, 信息化软硬件资源仍将进一步扩展。

面对如此众多的软硬件资源, 通过人为地分析系统软硬件日志数据等进行隐患排查、异常情况检测分析, 已经很难满足实际运维工作的需要。因此, 利用大数据技术, 综合大量日志数据进行多角度自动异常检测分析, 提前发现网络中存在的异常行为、软件设备中可能存在的异常隐患、硬件设备中可能存在的安全隐患, 将人为地隐患排查、被动地异常

检测分析等传统运维方式, 变成自动地异常告警, 将信息化运维工作变被动为主动, 把可能出现的系统故障扼杀在萌芽状态, 把可能的信息安全高危事件提前预测出来, 并提供优化解决方案。这将极大地提升信息化运维水平, 保障信息系统安全可靠稳定地运行, 助力公司信息化的持续健康发展。

1 研究内容

本文的主要目的是利用大数据技术, 构建设备和信息系统运行预警平台。通过对系统或设备基础数据的搜集, 来为系统或设备安全稳定运行、优化提供决策指导和建议。

1.1 国内外相关技术研究

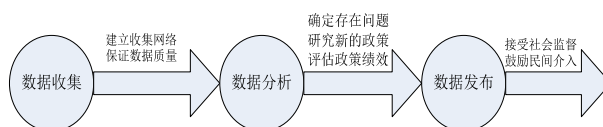


图 1 循“数”管理三部曲

图 1 是美国交通安全管理局循“数”管理三部曲。

美国是拥有机动车数量最早最多的国家, 随着人口和车辆的数量、密度都成倍增长, 车辆的使用频率也大幅增加, 但交通事故的死亡人数却不升反降, 而且幅度显著: 由 1966 年的每年 5 万人下降到现在的 3 万人。美国交通安全管理局把他们的经验概括为: 循“数”管理。他们从 1966 年起就开始收集交通事故死亡记录, 建立“交通事故死亡分析报告系统”。通过数据分析, 发现交通事故确实与时间段、地段、天气、路况等环境因素有关。以数据为导向, 查询事故的原因, 调整交警的配备、汽车的安全性、

路况的修复等等, 调整后再与之前的数据相比较, 确定最有效的措施, 把最好的做法在全国推广, 从而不断降低事故发生率。

1.2 电力企业数据现状

随着信息系统的深化应用, 积累了大量的数据和日志, 如北信源桌面管理系统中的所有终端设备的基本信息、网络接入情况; 5186 报修服务系统中的所有设备、软件故障报修记录、报修人员资料; 人力资源系统的所有员工基本信息、培训信息、履职情况; 网络设备的服务日志; 服务器的系统日志等等。数据量大且不断在增长, 但其中所蕴含的价值并没有被挖掘出来。

1.3 问题分析

信息安全是信息运维人员关注的永恒的重点。为了保证信息安全, 从国家电网公司层级开始就做了很多政策性的部署: 内外网分离、所有设备和系统禁止使用弱口令、内网设备禁止接入外网等等, 还有主干网和主要软件服务可用性达 100% 的考核指标要求。

江苏省电力公司在信息内网使用的终端数量超过数十万台, 用户水平参差不齐, 在全面宣传教育, 重点教育后, 弱口令、违规外联等信息安全事件仍时有发生。由于信息安全人员数量有限, 专业管理手段一般受限于事前宣传和事后考核两类, 但是事后处理也难以挽回影响。

服务器硬件设备故障导致的系统宕机, 影响 100% 可用性的考核指标, 高危漏洞影响了信息的安全性。

受天气影响, 或者市政工程施工, 光缆受到外力破坏, 影响网络可用性。网络设备故障导致的网络不可用。

这些看似是不可避免的问题, 如能早一步发现告警信息, 及时处理, 将可以大大提高信息系统的安全性以及可用性的指标。

1.4 解决方案的设想

通常情况下只在系统异常, 设备发生故障的情况下管理员才会去查看相应的日志, 查找故障原因并解决问题。如果能从这些纷繁复杂的日志中发现一些有价值的预警信息, 提前发送警示通知给管理员, 修复告警信息。将故障处理变革为缺陷处理, 将事后处理变为提前预警, 从现有的数据中挖掘出巨大的价值, 将是信息系统运维方式的一次重大变

革。

2 建立模型

运用保存在电力企业基础信息系统中的设备原始海量运行数据, 通过大数据技术中的多元回归、主分量分析等技术, 在相似性理论支持下, 转化成动态的设备在线模型。将动态设备模型计算生成的实时预估值和设备测点的实测值进行比较, 并根据比较结果发布设备早期故障状态预警。提前预知性检修, 使专业人员有充足的时间进行状态检修, 设备恢复正常状态, 避免了设备损坏和非计划停机。

一年的数据, 一个地区的数据可能看不出太多的规律, 但随着跨年度、跨地区的数据越来越多, 群体的行为特点就会在数据上呈现出一种“秩序、关联和稳定”, 更多的规律将浮出水面。随着时间的积累, 机器动态学习, 不断修正模型, 修正分析方法, 最终对未知设备提前估算, 实时发出预警通知。

3 分析方法

大数据技术就是以有效的信息技术手段和计算方法, 获取、处理和分析各种应用行业的大数据, 发现和提取数据的内在价值, 为行业提供高附加值的应用和服务。

在开展大数据分析工作之前有几个要点必须明确: 1) 明确数据来源及收集数据的工具; 以往信息系统中的数据均存在于数据库中, 各个设备的日志散落在各地, 须明确收集及管理海量数据的工具。2) 明确分析方法, 用它来挖掘海量数据中的“大科学”、“大价值”。3) 要有管理和分析大数据的人才。

3.1 大数据收集工具

大数据是指一个超大的、难以用现有常规的数据库管理技术和工具处理的数据集。

大数据技术描述了一种新一代技术和构架, 用于以更经济的方式、高速的捕获、发现和分析技术, 从各种超大规模的数据中提取价值。大数据使得传统的关系数据库已经难以胜任, 在存储能力和查询性能上都难以满足大数据存储和查询管理的需求。Exadata 是 Oracle 公司 2009 年提出的一个全新的架构。Exadata 实际上是 Oracle 的数据库云服务器。在 I/O 通信、服务器及存储性能方面都有优势。它是一种大容量并行的存储网格, 增加存储单元就可以增加存储管道的数目。利用 Oracle Exadata 的

Infiniband 技术和闪存技术,可以大大提高系统处理性能,从而实现对海量数据的高效处理。

3.2 数据来源

针对电力企业的现状,若要提高设备的可用性和安全性,需要建立设备生命周期数据库。可使用射频卡技术管理终端设备,自动采集相关数据,如:cpu大小、硬盘型号、内存型号、投运日期、生产厂商、生产日期、序列号、价格、采购日期、责任人;以及该机器对应的检修日期、检修原因、故障前状态、故障后状态、系统日志等运行状态数据库。通过大数据技术,分析信息设备使用情况数据。分析电源品牌与维修记录之间的关系,分析电源品牌与运行周期之间的关系,分析硬盘型号与设备宕机之间的关系,找到关联关系后,在发生故障警示信号前发送相关预警信息给管理员,更换相关设备,提高设备的可用性和安全性,对日后的产品采购也起到一定的决策指导作用。分析设备的维修记录,如果多次报修口令重置,则有可能机器是弱口令,提前对该用户进行安全教育或进行设备安全检查。所有的数据都应该是真实的历史记录,保证了数据质量,再在有质量的数据基础上进行分析,才能有真实的效果。

针对光缆外力破坏事件的分析,需记录光缆外力破坏事件时的所有相关系统信息,如:系统负荷、故障状况、运行状态等信息,以及生产系统中操作票、检修计划、工程项目等生产性数据、GIS 平台、TMS、气象平台等辅助数据,通过对系统中的报警、出错、操作等信息进行大数据分析,分析出天气与光缆外力破坏的关系,工程检修与外力破坏的关系,提供实时预警,尽量减少光缆因人为因素导致的外力破坏故障,保障通道安全。

3.3 分析处理工具

大数据在进行处理时要先解决两大技术挑战,见表 1。

表 1 大数据技术必须解决的两大挑战

| 技术挑战 | 数据存储 | 计算性能 |
|------|-----------------|---------------------------------|
| 问题 | 数据规模导致难以应对的存储量。 | 数据规模导致传统算法失效,复杂的数据关联性导致高复杂度的计算。 |
| 解决方法 | 分布存储 | 并行计算 |

MapReduce 是 2004 年由 Google 提出的面向大数据处理的编程模型,起初主要用作互联网数据的处理,例如文档抓取、倒排索引的建立等。但由于

其简单而强大的数据处理接口和对大规模并行执行、容错及负载均衡等实现细节的隐藏,该技术一经推出便迅速在机器学习、数据挖掘、数据分析等领域得到广泛的应用。

MapReduce 将数据处理任务抽象为一系列的 Map (影射)-Reduce (化简)操作对,Map 主要完成数据的过滤操作,Reduce 主要完成数据的聚焦操作。MapReduce 框架自动对任务进行划分以做到并行执行。

Hadoop 是基于 MapReduce 的开源实现。Hadoop 平台运用了传统的数据库索引技术,并通过分区数据并置的方式来提升性能。基于 MapReduce 实现了以流水线方式在各个操作符间传递数据,从而缩短任务执行时间。完美解决两大技术挑战。

3.4 分析处理方法

针对电力企业信息系统环境,通过搭建基于 Hadoop 的数据分布式存储与计算平台,实时收集服务器操作系统、weblogic、网络设备日志数据,形成企业信息系统运行状态日志信息数据库;构建基于每个日志数据的事件描述序列,设计基于日志信息的信息系统运行事件信息提取和整合算法,形成基于日志信息的系统状态预测模型;对不同类型的系统日志进行准实时自动检测,发现其中存在的异常状态序列,为信息系统的隐患排查和故障分析提供有力的支撑。

| | | | | |
|----|----------|---------|----------------------|------|
| 错误 | 2014-7-1 | 8:28:09 | Srv | 无 |
| 错误 | 2014-7-1 | 8:27:09 | Srv | 无 |
| 错误 | 2014-7-1 | 8:21:04 | Application Popup | 无 |
| 错误 | 2014-7-1 | 8:20:59 | Service Control M... | 无 |
| 警告 | 2014-7-1 | 8:12:53 | Foundation Agents | Host |
| 警告 | 2014-7-1 | 8:12:53 | Foundation Agents | Host |

图 2 错误日志

预警平台通过分析系统的错误日志,进行日常的异常检测。见图 2。以短信和或邮件告警的方式,通过每小时对日志数据中存在的异常事件进行检测,为服务器运维人员、weblogic 中间件运维人员、网络运维人员提供异常告警信息,帮助运维人员更好地排查隐患。也可以进行故障相关的异常检测。通过人为地启动特定日志片段的异常检测,为运维人员更快地定位故障原因,发现问题源头,快速的恢复系统对外的正常服务。

4 结论

美国交通安全管理局在 1980 年开始实施驾驶

人员必须佩戴安全带的规定,但随后收集到的数据却表明,实施同样规定的州,死亡率下降幅度却各不相同。后来发现,效果明显的州警察有权力随时截停车辆、检查司乘人员是否佩戴了安全带,而效果不明显的州,警察是没有权利随时检查的。这个发现,说明相同的政策由于执行方法不同,效果可能大不相同。根据大数据分析的结果,制定一些优化的方案,并评估效果的好坏,重新分析,重新制定方案,不断优化,建成高效的设备及信息系统预警平台,才能助力公司信息化企业的建设。

参考文献:

[1] 维克托·迈尔-舍恩伯格,肯尼思·库克耶. 大数据时代

[M].杭州:浙江人民出版社,2012.

[2] 城田真琴.大数据的冲击[M].北京:人民邮电出版社,2013.

[3] 中国电力大数据发展白皮书[R].2013.

[4] 涂子沛.大数据[M].桂林:广西师范大学出版社,2013.

作者简介:

王 芳(1977-),女,江苏泰兴人,工程师,主要从事信息通信运行、应用、安全及技术管理相关工作,E-mail:wfsz@js.sgcc.com.cn;

赵俊峰(1974-),男,江苏常州人,高级工程师,主要从事电网系统信息管理。