神彩云梦生态环境大数据平台 环境“数据金矿”的挖掘利器

本报讯 “大数据是一场革命,它将改变我们的生活、工作和思维方式。大数据技术的发展与应用,将对社会的组织结构、国家的治理模式、企业的决策架构、商业的业务策略以及个人的生活方式等产生深远的影响。”在前不久发布的《大数据标准化白皮书》(2016版)中这样写道。

随着互联网、云计算以及大数据技术的发展,信息数据逐渐成为与物质和能源同等重要的资源。这些数据仿佛一座巨大的金矿,只是现在社会上缺少挖掘金矿、提炼金矿的有效工具。为此,我国也将发展大数据提升到国家战略的高度,并制定了《促进大数据发展行动纲要》;作为信息化起步较早程度较高的环保系统,环保部也在今年3月出台了《生态环境大数据建设总体方案》。

作为江苏省环境信息系统工程技术研究中心的依托单位,江苏梦兰神彩科技股份有限公司早在2012年起就开始投入大量资金和优秀研发团队开始研究大数据技术,并成功研发出神彩云梦生态环境大数据平台。

据悉,这是国内首个通过专业认证的生态环境大数据平台。神彩云梦生态环境大数据平台的落地,使环境监管部门深入挖掘环境“数据金矿”,进而实现用数据说话、用数据决策、用数据管理、用数据创新成为可能。

时代利器——大数据

大数据时代的来临为政府更高效更智慧的管理带来了可能,同时也为企业带来了全新的发展机遇。

梦兰神彩以“让环境管理更智慧,让环境保护更科学”为使命,基于Apache Hadoop开发了一站式生态环境大数据平台——云梦(简称SCHDP)。整体而言,平台通过提供从数据存储、分布式计算、数据分析挖掘以及数据可视化的整套支持,帮助政府、企业建立统一的数据存储和计算平台。用户可以在云梦SCHDP大数据平台上采集、存储、搜索、分析海量数据,挖掘其内在价值。

从技术层面讲,神彩云梦大数据平台通过整合基于国家标准的多种通信协议,实现对环境监测仪器设备的直接接入,同时将接入的实时数据、分钟数据、小时数据、日数据通过消息队列和流式处理实时传输到SCHDP数据存储平台,并且做到对各类采集的数据进行实时分析,以前端报表的形式展现出来。

SCHDP通过内存计算技术、高效索引、执行计划优化和高度容错等技术,使得一个平台能够处理从GB到PB级的数据,并且在每个数量级上都能提供比现有技术更快的性能;用户不再需要混合架构,不需要孤立的多个集群。SCHDP可以伴随用户的数据增长而动态不停机扩容,避免Mpp或传统架构数据迁移的棘手问题。就运算速度而言,SCHDP比主流MPP数据库快5到10倍。当数据全部装载到内存后,能在3秒内多维度的分析10亿条记录。

SCHDP在统一存储上建立资源管理层,提供用户统一的计算资源管理、动态资源分配、多部门之间的资源配置和动态共享等功能,使多部门多应用可以灵活地在统一平台上平滑运行。这一点,与环保部发布的《生态环境大数据建设总体方案》中提到的一个机制、两套体系、三个平台的布局思路完全吻合。

大道至简——与数据分析生态系统的强力整合

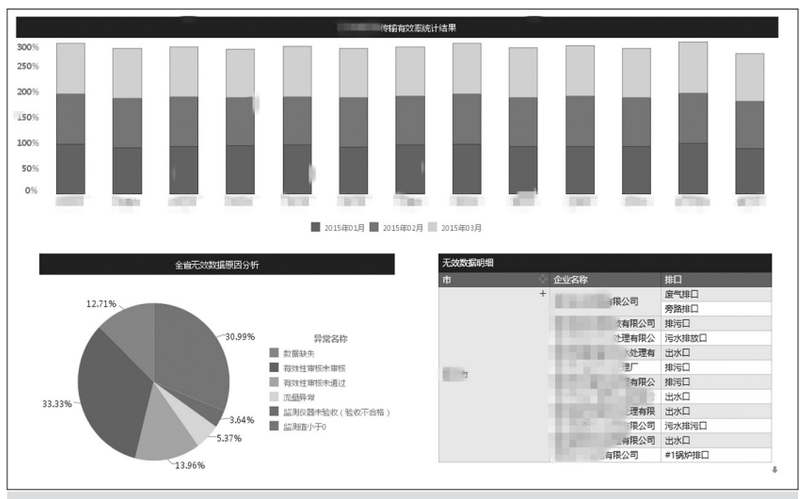
环境信息化建设弥补了原始监管方式的许多不足,但同时也带来新的问题和挑战。如何在海量数据中高效检索分析业务需要的信息?如何辨别数据的有效性?如何让每天源源不断汇集的数据产生更大的价值?……大数据技术的突破和应用,将是解答这些问题的关键。

借助SCHDP,用户能够与环境数据进行“对话”,交互式地探索数据,从而洞察并发现模式和趋势,以便开展进一步的分析及决策。SCHDP通过分布式列式存储和优化的高速执行引擎来支持交互式SQL查询,使得实时和交互式分析变成可能。此外,SCHDP还集成了机器学习算法库Mahout,包含了聚类分析、分类算法、频度关联分析和推荐系统在内的常用机器学习算法。

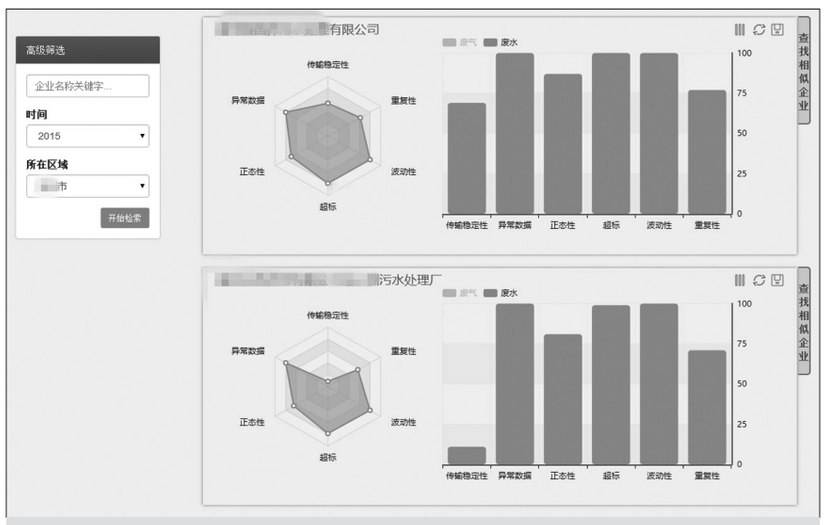
SCHDP非常重视与数据分析生态系统的整合以提高系统的易用性。SCHDP与现有成熟系统的无缝整合涉及了数据获取、数据分析以及数据可视化端。数据分析层与R语言的整合带来了R的数千种统计算法的同时,也可以充分利用R语言中的绘图工具绘制专业的统计报表;数据可视化不仅可将最终分析结果展示给用户,还可以帮助数据分析师来进行数据探索、发现和解决新问题。SCHDP支持多种可视化报表生成工具,使得基于大数据分析的商业决策更易被理解和接受,从而将大数据的潜在价值最大化。

神彩云梦生态环境大数据平台在具体的应用场景中已逐渐崭露锋芒。针对环保监管部门每天汇集的大量数据,通过SCHDP可以快速进行数据分析挖掘,筛选出异常数据、这为环保部门有针对地监管提供科学参考。此外,SCHDP还可以在数十亿条数据中,以秒级速度做出结果响应。

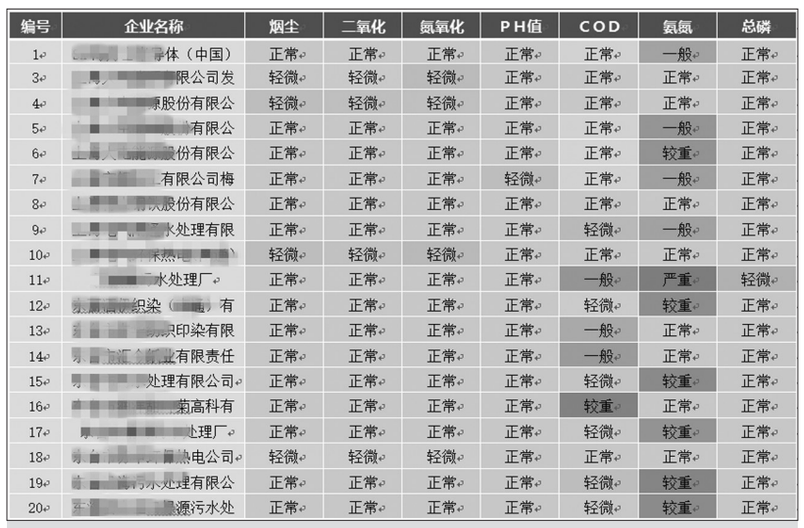
相较其他现有的环境大数据应用而言,梦兰神彩研发设计的环境大数据平台,显得更务实,更贴近当下环境管理的现实需求。随着国家大数据战略的逐步推进,各部门数据共享的范围边界和使用方式也将更清晰明确,有了这些条件的支持和更全的外部数据(如:水利、气象、工商及水电煤供给数据)的接入,届时神彩云梦生态环境大数据平台的价值将在更高层面得以体现。 **刘志勇**



图为数据传输有效率综合分析结果,分析数据分布模式,判断无效数据成因。



图为对企业数据从异常性、传输稳定性、重复性、波动性等多维度分析,排查异常数据。



编号	企业名称	烟尘	二氧化	氮氧化	PH值	COD	氨氮	总磷
1	体(中国)	正常	正常	正常	正常	正常	一般	正常
3	限公司发	轻微	轻微	轻微	正常	正常	正常	正常
4	股份有限公	轻微	轻微	轻微	正常	正常	正常	正常
5	有限公	正常	正常	正常	正常	正常	一般	正常
6	份有限公	正常	正常	正常	正常	正常	较重	正常
7	有限公司梅	正常	正常	正常	轻微	正常	一般	正常
8	股份有限公	正常	正常	正常	正常	正常	正常	正常
9	水处理有限	正常	正常	正常	正常	轻微	一般	正常
10		轻微	轻微	轻微	正常	正常	正常	正常
11	水处理厂	正常	正常	正常	正常	一般	严重	轻微
12	织染(有	正常	正常	正常	正常	轻微	较重	正常
13	织印染有限	正常	正常	正常	正常	一般	正常	正常
14	有限责任	正常	正常	正常	正常	一般	正常	正常
15	处理有限公司	正常	正常	正常	正常	轻微	较重	正常
16	菊高科有	正常	正常	正常	正常	较重	正常	正常
17	处理厂	正常	正常	正常	正常	轻微	较重	正常
18	电公司	轻微	轻微	轻微	正常	正常	正常	正常
19	污水处理有限公	正常	正常	正常	正常	轻微	较重	正常
20	源污水处	正常	正常	正常	正常	轻微	较重	正常

图为监测点非常态分析结果。