计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(4):205−211 [doi: 10.15888/j.cnki.csa.006739]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

环境评估大数据管理平台初探及技术综述①

张达刚[1], 陈海宁[2], 陈　华[2], 张光怡[2]

[1](北京恒辉信达技术有限公司, 北京 100045)
[2](北京市环保局, 北京 100048)
通讯作者: 张达刚, E-mail: zhangdagang@hhdatabase.com.cn

摘　要: 随着软硬件成本的下降和相关技术的日趋成熟, 传统的数据分析和管理方式已不能适应业务发展需求. 由于环境保护监测和评估数据符合大数据的特征, 因此利用大数据技术构建新的环境评估管理平台已成为可能. 本文在通过对环境评估大数据管理平台三层架构以及物理实现方案的阐述, 结合弹性扩展、流处理、数据湖、大规模并行处理、机器学习等技术的研究, 尝试解决环境评估数据的高速、深入、智能处理问题, 从而为业务管理提供更高效率的数据支撑.

关键词: 环境评估; 大数据; 管理平台

引用格式: 张达刚,陈海宁,陈华,张光怡.环境评估大数据管理平台初探及技术综述.计算机系统应用,2019,28(4):205–211. http://www.c-s-a.org.cn/1003-3254/6739.html

Proposal and Technology Research of Big Data Management Platform for Environmental Evaluation

ZHANG Da-Gang[1], CHEN Hai-Ning[2], CHEN Hua[2], ZHANG Guang-Yi[2]

[1](H2 Technology Co. Ltd., Beijing 100045, China)
[2](Beijing Municipal Environmental Protection Bureau, Beijing 100048, China)

Abstract: As the cost decrease of hardware and software with proven technologies, the traditional way to analyze and manage data has obvious shortages to meet the needs of business development. Because the monitoring and evaluation data of environmental protection matches the characteristics of big data, it is possible to build a new data management platform using big data technology. In this study, we design the three-layer architecture and physical implementation proposal of big data management platform for environmental evaluation, with the research of elastic expansion, stream processing, data lake, massive parallel processing, machine learning, and other technologies. We try to resolve the data processing issues of high speed, deep insight, and intelligence so that to provide high efficient support to business management.

Key words: environmental evaluation; big data; management platform

1　引言

近年来, 国家高度重视大数据在推进生态文明建设中的地位和作用, 各区域各行业都在努力贯彻落实加强生态环境大数据综合应用和集成分析的策略要求, 环保部门也希望借助新的技术解决新的数据问题, 为生态环境保护科学决策提供有力支撑[1]. 通过研究发现, 环境监测数据符合大数据的容量大、种类多、增长速度快、价值大的特征, 通过传统的数据处理方法获取其中有用数据, 难以满足业务的发展需求. 本文通过对相关技术及方案的探讨, 为深度挖掘环评监测数据的业务价值提供支撑.

根据北京市环境保护局发布的《建设项目环境影

① 收稿时间: 2018-06-20; 修改时间: 2018-06-22, 2018-08-13; 采用时间: 2018-08-24; csa 在线出版时间: 2019-03-28

响评价分类管理名录》, 环境影响分类从 A 类 (水利)、B 类 (农、林、牧、渔、海洋) 一直到 W 类 (核与辐射), 共计 23 个大类, 199 个子类. 每个子类具有多种不同监测指标, 代表着总数量大概几千个评估维度. 从对污染物监测数据采样数量看, 例如《环境空气质量标准》(GB3095-2012) 列出的大气主要污染物 (二氧化硫、一氧化碳、颗粒物等), 每一种的采样频度如果是 10 分钟, 以每个监测点一次 1 KB(1024 字节) 的采样数据为例, 一年的数据大小是 1 KB×7(污染物)×6(10 分钟采样周期)×24(一天小时数)×365(一年天数)=360 MB, 保守估计京津冀地区 10 万个重点监测企业, 假设每个企业的监测点为 100 个, 大气监测一年的数据量为: 360 MB×100 000×100=3.35 PB. 如此多的维度和数据数量, 以及对数据传输速度、存储和提取速度等方面的实际要求, 远远超出传统数据管理和分析方法所能达到的限度, 对超海量数据处理的多维度分析、性能优化、弹性扩展等方面提出技术挑战.

通过对数据湖、弹性扩展、大规模并行处理、流处理、机器学习等技术的研究, 应对环境评估大数据需求, 我们设计环境评估业务的数据管理平台, 实现适用的数据资源传输交换、存储管理和分析处理功能, 为环境评估业务应用提供统一的数据支撑服务. 经过前期调研分析, 我们利用基础关系型数据库、分析型数据库以及 Hadoop 平台的部分组件搭建了 NoSQL 和 SQL 集成一体的环评文件数据提取系统, 通过较为简单的数据建模, 初步验证了大数据技术平台的能力, 包括能够实现数据传输交换、管理监控、共享开放、分析挖掘等基本功能, 支撑分布式计算、流式数据处理、大数据关联分析、趋势分析、空间分析, 支撑大数据产品研发和应用等, 这些为后续付诸实用的环境评估大数据管理平台, 做出相应的初步验证.

2 关键技术

2.1 弹性扩展

环保监测与评估的数据分析维度众多, 而且数据量日益增长, 造成历史数据需要压缩保存, 部分数据需要定期清空以回收资源, 另外, 不同维度的数据如大气、土壤、水质等数据需要分库分路径保管, 这些对存储和计算资源提出了弹性扩展、回收重用的重要需求.

弹性扩展指的是云应用本身的一种动态的扩展, 也就是在云应用运行期间实现支撑云应用的虚拟机实例个数的动态增加或者减少[2]. 弹性扩展并不是简单的资源复制, 而是通过计算能力、存储能力的调配以及配套的集群、安全管控等功能形成的完整的资源按需分配, 可以在不改变平台部署架构的情况下实现环保海量数据动态增容功能.

2.2 流处理

流式数据是大数据环境下的一种数据形态, 与静态、批处理和持久化的数据库处理相比, 流式计算以连续、无边界和瞬时性为特征, 适合高速并发和大规模数据实时处理的场景[3]. 当前很多环境评估数据, 例如噪声数据, 具有多源并发、瞬间发生、快速失效的特点, 采用流处理技术就实时采集和处理瞬时数据的相关指标, 从而解决环保监测的实时性问题.

大数据环境下, 流式数据作为一种新型的数据类型, 是实时数据处理所面向的数据类型, 其相关研究发展迅速. 这种实时的流式数据, 存在如下几个特征:

1) 实时、高速: 数据能以高并发的方式迅速到达, 业务计算要求快速连续相应. 数据处理的速度至少能够匹配数据到达的速度.

2) 无边界: 数据到达、处理和向后传递均是持续不断的.

3) 瞬时性和有限持久性: 通常情况下, 原始数据在扫描处理后丢弃, 并不进行保存; 只有计算结果和部分中间数据在有限时间内被保存和向后传递.

4) 价值的时间偏倚性: 随着时间的流逝, 数据中所蕴含的知识价值往往也在衰减, 也即流中数据项的重要程度是不同的, 最近到达的数据往往比早先到达的数据更有价值.

2.3 数据湖

数据湖是一种在系统或存储库中以自然格式存储数据的方法, 它有助于以各种模式和结构形式配置数据. 数据湖的主要思想是对企业中的所有数据进行统一存储, 从原始数据 (这意味着源系统数据的精确副本) 转换为用于报告、可视化、分析和机器学习等各种任务的转换数据. 湖中的数据包括结构化数据 (行和列数据)、半结构化数据 (CSV、XML、JSON 的日志)、非结构化数据 (电子邮件, 文档, PDF) 和二进制数据 (图像、音频、视频) 等. 数据湖能够形成一个集中式数据存储, 容纳所有形式的数据[4].

源于数据仓库概念的数据湖理论, 更好地解决了数据仓库和大数据处理技术表现出来的部分弊端, 即

能够接收来自多种数据源的输入, 同时保留原始数据的真实性和数据传输状态, 并满足实时分析的需要, 也能够作为数据仓库满足批处理和数据挖掘的需要, 从而满足环保监测与评估数据需要多种数据源的集成、不限制数据对象集合、保留数据精确度的处理要求, 例如大气污染和地域、气象等等多种因素相关, 只有通过数据湖进行数据融合才能进行有针对性的后续处理.

数据湖具有如下作用.

1) 数据的集中存放管理: 数据湖是平台用于存放所有所需数据的地方, 这些数据包括来自传统数据库的结构化数据和非结构化的文本数据, 包括企业内部生成的数据, 外部数据以及服务数据, 也包括媒体数据, 传感器采集数据和很多企业正在学习使用的遥测数据.

2) 强大的交叉分析平台: 数据湖可以看作是一个大数据分析平台, 不仅仅可以实现所有种类数据的存放, 也可以用于数据分析, 以及找到数据新的关联性. 许多商业分析中的突破并不是来源于数据的多少和分析的熟练程度, 而是来源于能显示出商业表现的数据新式组合.

3) 为商业个体提供所需数据的最优解: 数据湖也同样协调了商业个体真正需要的数据和企业经常使用标准数据的不匹配问题, 它是一种共享资源, 不仅包含了精心管理的数据, 也提供了一个商业个体搜寻真正需要的数据组合的平台.

2.4 大规模并行处理

大规模并行处理 (MPP) 系统由众多松耦合处理单元组成, 每个单元内的处理器都有自己私有的资源, 如总线、内存、硬盘等, 在每个单元内都有操作系统和管理数据库的实例副本, 这种结构最大的特点是不共享资源[5]. MPP 是将任务并行分散到多个服务器和存储节点上, 在每个节点上计算完成后, 将各自部分的结果汇总在一起得到最终的结果.

随着对环境评估时效性要求的提高, 大量环境监测采集数据需要得到快速处理, 以便及时为决策和执行提供依据, 因此, 我们有必要采用大规模并行处理技术来加速海量数据的处理, 其中主要使用到 MPP 架构的数据库.

2.5 机器学习

机器学习技术包括数据存储和模式设计、不同组件的模块化、单独架构每个独立的可扩展组件、系统和性能测试, 以及数据可视化等. 典型的机器学习工作流包括, 使用数据流处理技术读取不同来源的数据, 使用 SQL 过滤、聚合, 并执行数据集上的其他初始化处理, 然后, 使用计算引擎将处理过的数据转换以创建特征向量, 对模型进行训练和评估, 并使机器学习与 SQL 解析和流处理技术达到深度集成[6]. 环境监测数据具有数据量大, 数据维度复杂的特点, 并且常用查询维度的集中度很大, 所以深度学习环境监测的数据访问规律可以大大提升数据访问速度.

在机器学习技术的实现中, 我们采用深度学习技术. 深度学习是机器学习中表征学习算法, 使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象计算, 将用于监督式或半监督式的特征学习和分层特征提取的高效算法来替代手工获取. 基于数据的深度学习过程是数据库系统掌控应用系统的访问规律, 动态调整系统资源, 找到最快速、最高效的访问路径, 给用户带来越用越快的用户体验的过程[7].

3 大数据管理平台

3.1 平台逻辑架构

环境评估大数据管理平台采用云计算环境作为基础设施, 即以云计算基础架构即服务 (IAAS) 层作为物理支撑, 从中得到可弹性扩展的计算服务、存储服务、数据传输服务、安全管控服务等基础服务. 环境评估大数据管理平台的主要作用是大数据管理, 是整个环境评估服务系统的核心, 分为数据层、分析层和业务层, 通过对各类数据的收集、抽取、存储、清洗、标准化、关联、标记、深度加工、可视化等处理, 形成数据资源中心, 并为上层应用提供统一数据服务.

平台的数据层负责大数据存储, 将各种类型和特点的数据统一存储管理, 为分析层提供海量数据和快速提取的服务功能, 分析层负责大数据融合, 通过流处理、并行计算、深度学习等技术将数据进行融合处理, 为业务层提供可用原始数据和整合数据的灵活访问服务功能, 业务层负责业务应用的大数据接入, 对数据进行综合提取和展现, 提供数据的增值服务功能, 供给不同的业务应用进行接入和使用, 参见图 1.

3.1.1 数据层

数据层主要是通过数据湖技术和弹性扩展技术对数据进行接收、存储和初步处理, 主要解决了海量数据和多元数据问题, 包括来自环保数据采集系统和业务系统的结构化数据, 和来自采集端点的实时数据、

业务系统采集生成的多种格式非结构化数据等.

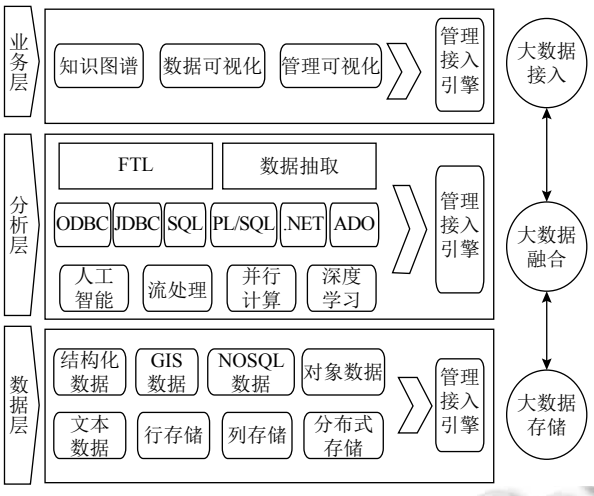

图 1 大数据管理平台架构

(1) 结构化数据的行式存储和列式存储

使用最广的数据存储方式是行式存储, 把一行数据作为一个整体来存储, 但行式存储在维护大量的索引和物化视图场景下, 在处理时间和存储空间方面成本过高. 列式存储数据库以列为单位进行数据存储, 每一列单独存放, 并由一个线程来处理, 这样既可以充分利用处理器的多核心特性, 又能够大大降低系统 I/O 开销, 因此我们采用擅长随机读操作的行式数据库与擅长条件查询的列式数据库相结合的方式, 来管理结构化数据.

(2) 非结构化数据的分布式存储和弹性扩展

非结构化数据需要分布式存储, 并且保证按需的弹性扩展功能. 平台的分布式存储充分利用 HDFS 的低成本、高容错、高吞吐特性来管理数据, 经由并行数据路径完成与 MPP 数据库服务器的数据交换, 通过弹性控制管理模块联动数据协调分发模块提供数据的弹性扩展管理, 参见图 2.

对于弹性扩展在弹性控制管理模块中采用特定语言进行描述, 通过描述中的内容进行灵活的扩展, 例如, 描述一个扩展节点, 包括硬件、软件特征和配置必须明确规定, 并以特定的方式进行表述, 再使用自动化任务解析、执行这些相关的描述文档, 从而实现相应扩展功能.

(3) 支持处理的数据类型

平台支持对常用的所有数据类型进行处理, 包括:

1) 关系数据: 支持关系数据的各种数值类型、字符类型、二进制数据类型、日期时间类型、布尔类型等.

2) 空间数据: 支持几何特征和离散特点的地理要素, 即空间对象数据, 如点、线、面、体等对象的数据组件, 以及 GIS 栅格、图层、坐标等数据存取.

3) NoSQL 数据: 支持 NoSQL 数据类型、位串类型、数组类型、复合类型等.

4) 文本数据: 支持常用的文本数据, 包括日志数据、文章数据、网页数据等.

平台对数据的管理都采用图形化界面进行操作, 例如对 NoSQL 数据的管理已实现如图 3 的界面.

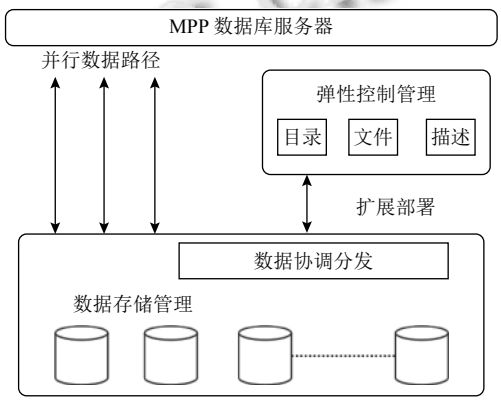

图 2 弹性扩展

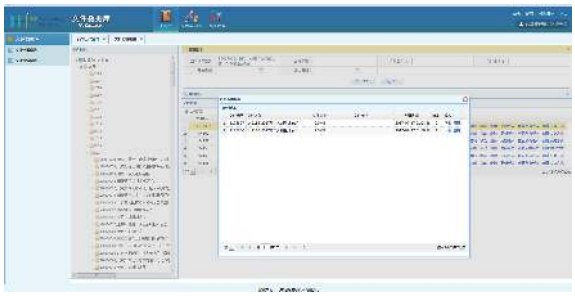

图 3 NoSQL 数据管理界面

3.1.2 分析层

分析层对数据进行融合处理, 是一种针对环保检测和评估数据的容量大、种类多、增长速度快、价值大等特征的集成技术, 包括: 流处理技术、大规模并行处理技术、机器学习技术、并行算法等.

平台通过增加并行度确保使用整个集群的资源, 而不是把任务集中在几个特定的节点上. 对于包含 Apache Spark Shuffle 的操作, 增加其并行度以确保更为充分地使用集群资源; 同时, 流处理默认将接收到的数据序列化后存储, 以减少内存的使用, 但是序列化和反序列化需要更多的处理器资源, 因此优化的序列化

方式和自定义的序列化接口可以更高效地使用处理器资源, 参见图 4.

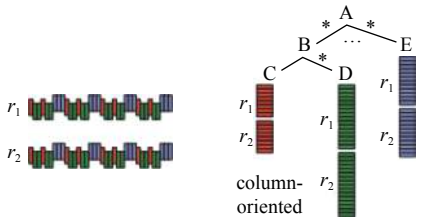

图 4 流处理序列和反序列

在流处理中, 任务之间有可能存在依赖关系, 后面的任务必须确保前面的作业执行结束后才能提交, 通常情况下分析型数据库框架能够高效地确保任务及时分发. 但是, 如果前面的任务执行的时间超出了批处理时间间隔, 那么后面的任务就无法按时提交, 这样就会进一步拖延接下来的任务, 造成后续任务的阻塞, 因此分析层会设置一个合理的批处理间隔以确保作业能够在这个批处理间隔内结束; 同样, 当批处理间隔非常小 (小于 500 毫秒) 时, 提交和分发任务的延迟就变得不可接受了, 通过经验对比, 我们采用 Spark 的 Standalone 和 Coarse-grained Mesos 模式减少因任务提交和分发所带来的延迟.

对于数据的底层模型设计, 因需要进行基于多维模型的交叉分析来有效发现问题, 所以数据的维度越丰富所能实现的交叉也越丰富和灵活; 但相应的, 如果要尽可能地丰富各维度的交叉分析, 对基层模型的要求也就越高. 因此, 我们引用数据立方体来实现模型设计, 参见图 5.

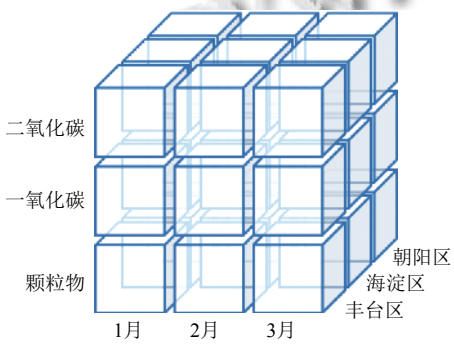

图 5 数据立方体示例

用数据立方体来拓展数据细节有两种方向, 一类是纵深拓展, 也就是基于一个维度的细分, 例如一个月细分到每一天, 一条记录将会被拓展成 30 条; 另一类是横向拓展、多个维度的交叉, 就像立方体中添加了空气污染物维和区域维. 这样存储的数据就从原本单一的时间维度扩展成了时间、污染物和区域三个维度, 也就是三维立方体所能展现的形式, 而且维度可以继续扩展, 四个、五个直到数十个, 理论上都是可行的. 以三个维度进行举例: 对于数据存储而言, 横向的拓展与纵深拓展的影响是一样的, 记录数都是以倍乘的方式增长, 假设有 20 个污染物大类, 再加上十六个区, 那么经过纵深和横向拓展之后, 原先每月的 1 条记录就变成了: $1 \times 30 \times 20 \times 16 = 9600$ (条).

在功能实现方面, 经过数据的多维分析后, 平台在数据准备区进行 ETL 处理, 数据经过抽取、转换后加载到数据仓库中, 分析完主题和数据元后建立数据模型 (概念模型、逻辑模型、物理模型) 并形成事实表和纬度表, 然后通过粒度分析将历史记录先抽取整合, 最后再根据决策者可能用到的数据集合分解成若干记录, 同时利用 OLAP 工具技术进行数据的分析导出, 以供给业务层进行数据可视化处理.

3.1.3 业务层

在业务层, 系统关注将分析层提供的数据进行可视化展现, 其中的重点就是使用知识图谱. 知识图谱基于图的数据结构, 由节点和边组成, 每个节点表示现实世界中存在的具有多种属性的“实体”, 每条边为实体与实体之间的“关系”. 知识图谱把所有不同种类的信息连接在一起而得到一个关系网络, 提供了从“关系”的角度去分析问题的能力, 是关系的最有效的表示方式[8].

基于知识图谱, 我们也尝试提供数据智能搜索服务. 智能搜索的功能类似于知识图谱在互联网搜索引擎上的应用, 也就是说, 对于每一个搜索的关键词, 我们可以通过知识图谱来返回更丰富, 更全面的信息. 比如搜索某个监测点的污染情况, 我们的智能搜索引擎可以返回与这个监测点相关的所有类型的污染记录, 包括水污染、大气污染、土壤污染等, 并同时返回区域涉及的建设项目信息、污染物排放标准等环境保护相关信息, 参见图 6.

另外, 通过可视化技术把复杂的信息以非常直观的方式呈现出来, 参见图 7, 使得我们对隐藏信息的情况也一目了然. 数据可视化是指以柱状图、饼状图、线型图等图形方式展示数据, 让决策者更高效地了解

业务的重要信息和细节层次. 大量实践表明, 人通过图形获取信息的速度比通过阅读文字获取信息的速度要快很多, 因此通过可视化展现配合门户服务, 帮助环保局管理人员实现高效、系统的数据识别和决策.

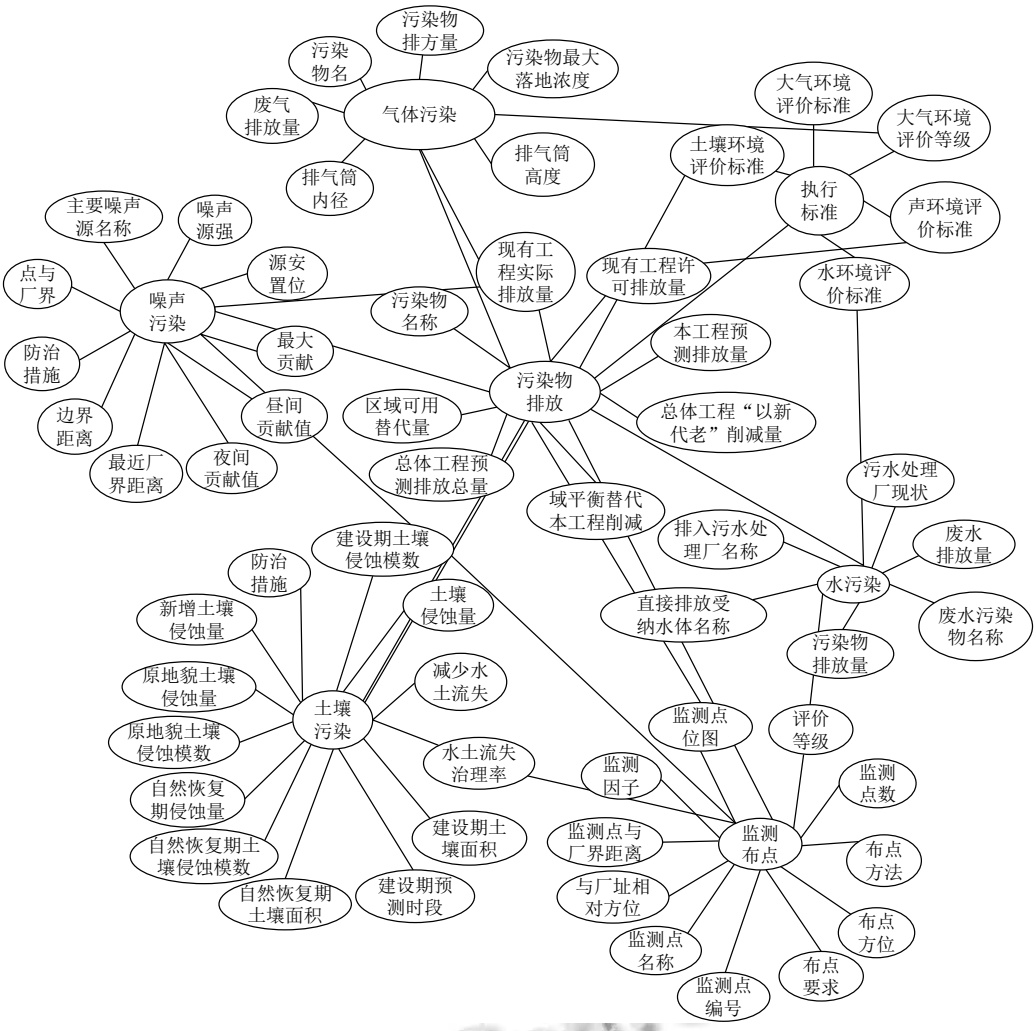

图 6 知识图谱关联

图 7 数据可视化展现示例

3.2 平台物理架构

云基础架构使得计算、存储、网络等可以通过资源池化而按需获得, 我们重点关注的是这些资源的整合以及基于此的动态变化管理策略, 形成一个有机的、可灵活调度和扩展的资源池, 面向大数据管理平台实现自动化的部署、监控、管理和运维.

参见图 8, 我们采用典型的云基础架构融合部署方案. 例如, 通过虚拟防火墙与虚拟机之间的融合, 可以实现虚拟防火墙对虚拟机的感知、关联, 确保虚拟机迁移、新增或减少时, 防火墙策略也能够自动关联. 此外, 虚拟机与负载均衡设备形成联动, 即在业务突发时, 自动按需增加相应数量的虚拟机, 与负载均衡联动实现业务负载分担; 同时, 当业务量减小时, 可以自动减少相应数量的虚拟机, 节省资源. 不仅有效解决虚拟化

环境中面临的负载突变问题, 而且大大提升了业务响应的效率和智能化. 再有, 云基础架构通过虚拟化技术与管理层的融合, 提升了 IT 系统的可靠性. 例如, 虚拟化平台可与网络管理、计算管理、存储管理联动, 当设备出现故障影响虚拟机业务时, 可自动迁移虚拟机, 保障业务正常访问; 对于设备正常、操作系统正常、但某个业务系统无法访问的情况, 虚拟化平台还可以与应用管理联动, 探测应用系统的状态, 例如 Web、应用、数据库等响应速度, 当某个应用无法正常提供访问时, 自动重启虚拟机, 恢复业务正常访问.

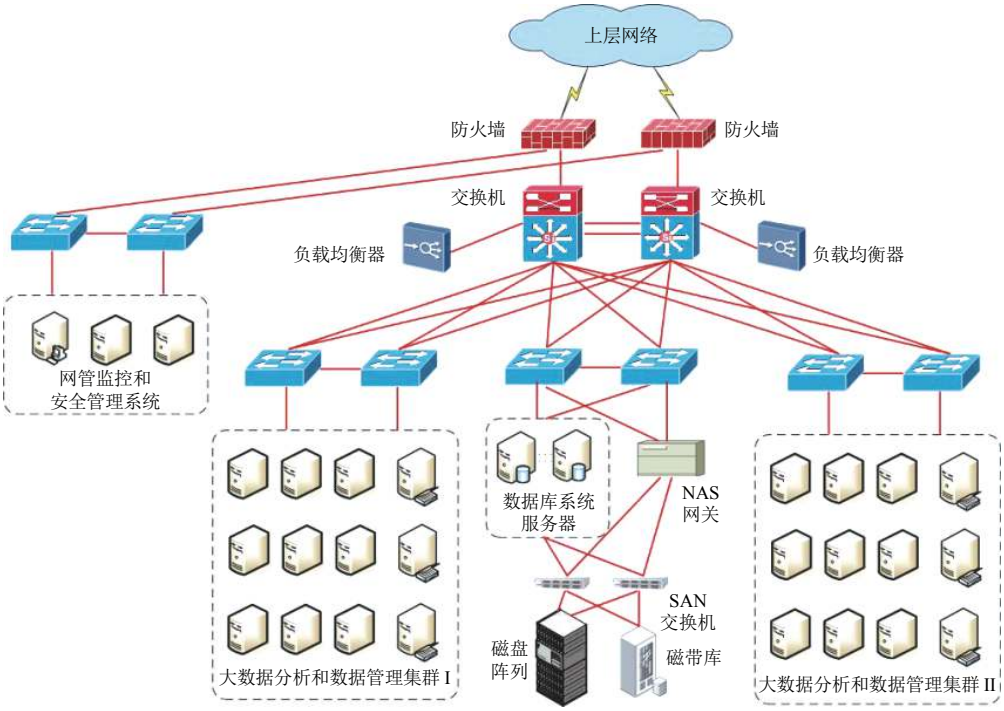

图 8 云基础架构融合部署

4 结语

本文对环境评估大数据管理平台涉及的关键技术和平台逻辑架构、物理架构设计进行阐述, 该平台是行业数据和数据库技术相结合的系统工程, 以大数据技术为支撑, 通过弹性扩展、流处理、数据湖、并行处理、机器学习等技术为手段, 不断结合环境监测与评估数据的需求分析调整技术方法, 实现环境监测和软件工程的软着陆, 为开展生态环境综合决策、环境监管和公共服务提供基础数据支撑, 为生态环境管理和决策提供服务.

参考文献

1 环境保护部办公厅. 关于印发《生态环境大数据建设总体方案》的通知 http://www.cac.gov.cn/2016-03/18/c_1118376330.htm. [2016-03-08]

2 Yang CW, Huang QY, Li ZL, *et al*. Big data and cloud computing: Innovation opportunities and challenges. International Journal of Digital Earth, 2017, 10(1): 13–53. [doi: 10.1080/17538947.2016.1239771]

3 陈付梅, 韩德志, 毕坤, 等. 大数据环境下的分布式数据流处理关键技术探析. 计算机应用, 2017, 37(3): 620–627.

4 Wikimedia Foundation, Inc. Data lake. https://en.wikipedia.org/wiki/Data_lake. [2018-07-16]

5 林荣智. 并行数据库技术分析与展望. 信息通信, 2016, (12): 200–201. [doi: 10.3969/j.issn.1673-1131.2016.12.095]

6 林志, 茆云霞. 深度学习技术在环保督查工作中的应用研究. 信息通信, 2017, (11): 80–82. [doi: 10.3969/j.issn.1673-1131.2017.11.035]

7 林伟声. 深度学习技术在信息系统数据分析中的应用. 电脑与电信, 2017, (6): 51–53.

8 Sivarajah U, Kamal MM, Irani Z, *et al*. Critical analysis of big data challenges and analytical methods. Journal of Business Research, 2017, 70: 263–286. [doi: 10.1016/j.jbusres.2016.08.001]