【工程与技术研究】

基于 Spark 技术的大数据智能分析平台构建

贾淑滟

(山西旅游职业学院 计算机科学系,山西 太原 030031)

摘 要:针对大数据分析过程易受数据维度的影响,造成大数据分析平台运行时间长、数据分析平均绝对误差高的问题,构建了基于 Spark 技术的大数据智能分析平台。先采用局部约束学习方法降低大数据维度,再利用 Spark 技术建立兼具应用服务层、权限管理层、中间服务层和基础资源层的平台分层架构,参考映射-归约数据网络质量分析系统,结合数据分析编排器设计平台后台分析服务器,实现了基于 Spark 技术的大数据智能分析平台的构建。所构建平台加速比参数大于 9,运行速度较快,运行效率在 99%左右,数据分析平均绝对误差为 0.5%~0.8%。

关键词:局部约束学习方法;数据维度;Spark 技术;平台分层架构

中图分类号: TP 399 **文献标识码**: A DOI:10.13486/j.cnki.1673-2618.2023.06.012

0 引言

用户在云端输入检索信令获取目标信息时,往往由于云端数据量过于庞大而导致目标信息检索超时[1]。考虑到这种负面因素影响,以云环境下大数据为支撑的企业开始致力于开发适用于自身的大数据智能分析平台。通过分散业务流程,实现大数据资源多层级管理,不仅能梳理杂乱无序的网络数据,还能缩短目标信息的检索时长,达到改善用户操作体验、拉近用户合作关系、增加企业用户流量的目的。汪杰等[2]通过 B/S 结构设计数据分析平台的基本框架,但该方法存在平台运行时间过长的问题。孟光伟[3]通过 Kafka 分布式函数将待处理数据划分成权重系数不同的多组 Kafka 集群,并将多组 Kafka 集群依次输入由 Spark Structured Streaming 网络引擎组成的后台运算程序,实现大数据智能分析平台的构建,该方法存在平台运行效率低、平均绝对误差较大的问题。张波等[4]采用云环境下大数据开源工具 Docker 建立基于业务管控系统的数据分析平台,并在实际应用中结合运营统计装置,实现大数据智能分析平台的构建。Ogiela 等[5]提出了基于人本分析的数据分析技术,为了充分解释所有可能发生的偏好,并对其在产品评估、促销阶段的意义和有用性进行了分析,但这两种方法存在平均绝对误差较大的问题。Abukmeil 等[6]从盲源分离、流形学习和神经网络架构了用于数据分析的无监督生成学习模型,但是该方法存在平台运行效率低的问题。李娟等[7]利用映射-归约(MapReduce,MR)和 Hadoop 构建了 Hadoop 云平台,在云平台中实现了分布式计算、数据挖掘、业务响应以及用户交互。虽然,MR 技术具有高效性、可扩展性、容错性和灵活性优势,但是 MR 需要进行数据划分、映射、归约等多个步骤,存在复杂性、数据倾斜、数据运算成本增加等问题。刘仁芬等[8]在筛选分布空间高维数据特征并进行降维的基础上,利用改进 Spark 技术,设计了高维数据增量式聚类算法,该方法降低了存储空间的占用率,可完成高维数据的有效、可靠聚

收稿日期:2023-04-21

作者简介:贾淑滟(1989—),女,山西岢岚人,讲师,硕士,主要从事计算机大数据应用研究。

E-mail:xiyoujsy@163.com

类。

在构建大数据智能分析平台时，需要综合考虑性能、可扩展性、易用性等因素。Spark 技术和 MR 技术都是用于大数据处理的分布式计算框架，也是构建时选择的主要方法。其中，Spark 技术是兼具 Scala 语义开发模式和分级式数据处理系统的新型平台开发技术，其对数据的存储和运行迭代均以云端为主，这使得最终构建的平台运行空间较大，不易出现由于数据量过大而导致的平台运行卡顿的问题。MR 技术[9]对数据的存储和运行迭代均以本地磁盘为主，这使得最终构建的平台运行空间较小，除运行卡顿外，还易发生平台崩溃和数据丢失的情况。在数据审核的过程中，Spark 技术采用循环审核的方式，最大限度过滤干扰数据，使平台接收到的数据信息优化效果明显，且平台负载率明显下降。MR 技术采用非循环审核的方式，易出现干扰数据扰乱平台数据库的现象，使平台负载率上升，数据运算成本增加。因此，本文提出基于 Spark 技术的大数据智能分析平台构建方法，以优化数据分析质量。

1　基于局部约束学习方法的大数据降维

局部约束学习方法作为一种数据降维方法，主要通过低维空间嵌入技术解决大数据的高维问题[10]。以高维数据集 S 为例，局部约束学习方法想要获取基于高维数据集 S 的低维映射指征，需通过组合相关函数[11]将数据集转化成流形函数曲线上的一组分布式高维数据标志点，达到约束标志点领域内映射的目的。组合相关函数为

$$C'_x=\begin{cases}\int_{i=1}^{k}[\theta_{ij}^{r}K-(R_i+R_j)]\mathrm{d}i, K=[k_1,k_2,\cdots,k_n],\\ \int_{j=1}^{l}\theta_{ij}^{r}(f(x_i)-\sin L)^2\mathrm{d}j, L=[l_1,l_2,\cdots,l_m]。\end{cases}$$

式中，θ_{ij}^{r} 表示组合相关常数，K 表示流形函数曲线斜率，$f(x_i)$ 表示流形函数曲线标志点附着率，$\sin L$ 表示高维数据集转化系数，R_i+R_j 表示高维数据转化误差，x 表示高维数据标志点总数。

流形函数曲线为

$$F=\chi\sum_{i,j\neq 0}^{k}[k_{tr}^{2}(1-e)+C_{ij}C']+\frac{1}{tr\times(1-e)}。$$

式中，χ 表示流形函数的常数，$1-e$ 表示流形函数的取值范围，$C_{ij}C'$ 表示流形函数在 x 轴上的最大值，k_{tr}^{2} 表示流形函数在 y 轴上的最大值，tr 表示函数曲线斜率对高维数据标志点附着情况的影响程度。

距离差分原理公式为

$$H=\min_{T,I}(Y-T_n)+\varphi\int_{n}^{T}\int_{m}^{I}d_{n\cdot m}^{2}(Y-T_n)^2\mathrm{d}n\mathrm{d}m。$$

式中，φ 表示距离差分常数，Y 表示高维数据标志点稳定系数，T_n 表示固定领域内高维数据标志点映射准确率，$d_{n\cdot m}^{2}$ 表示参与约束的高维数据标志点总数，I 表示跨领域映射的高维数据标志点过滤系数[12]。

微分同胚映射原理公式为

$$M=\cos R^{\mathrm{dic}}+\sum_{i\neq 0}^{x}p_y^{r}(c_o+c_i)-\lambda^2\sum_{j\neq 0}^{y}z_x^{r}(b_0+b_j)。$$

式中，$\cos R^{\mathrm{dic}}$ 表示微分同胚常数，p_y^{r} 表示高维数据标志点在低维空间的嵌入指数，z_x^{r} 表示高维数据的低维映射率，c_o+c_i 表示高维数据的低维映射误差，r 表示参与低维空间嵌入的高维数据标志点总数，b_0+b_j 表示空间维度对高维数据映射结果的影响程度。

2　大数据智能分析平台构建

2.1　大数据智能分析平台的架构

要建立非专业用户数据便捷检索的大数据智能分析平台，需从研发体系、数据分析流程、用户接口服

务等多个角度分析，同时考虑应用服务层、权限管理层、中间服务层和基础资源层，共同组建基于繁杂数据业务的平台分层架构图（图 1）。各层框架的功能：(1)应用服务层。应用服务层主要采用 Web Service 系统设计，该系统操作界面简便，方便用户下载数据、查看数据展示图、修改数据参数设定，对外接口服务内部信息系统不限制用户登录地址，即用户可以在任意地点登录并使用该平台。(2)权限管理层。权限管理层加入 Java 服务器约束数据集操作对象[13]，即每次平台操作任务仅能登记在一位用户名下，并保存该用户近三天内全部检索数据，方便用户重复审阅。(3)中间服务层。中间服务层包括主题数据服务模块和数据自动化汇聚模块，其中，主题数据服务模块又被细分为数据监控状态、数据资源目录、数据链接和数据主题库，这些分支将主题数据服务模块进一步细化，不但增加了数据信息配置的精确性，还为开发人员调试服务器提供可靠依据。(4)基础资源层。基础资源层是指企业上传的相关数据。以大数据智能分析平台为例，该平台需要上传的资源信息是经过降维处理的大数据信息[14-15]，在上传过程中，平台会根据 Map 对应映射表检测数据维度，达到整体数据维度无误的目的。

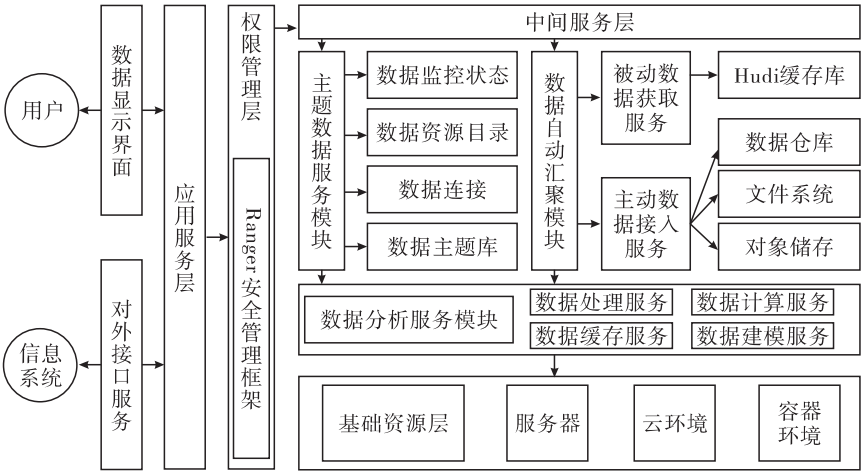

图 1 平台分层架构

2.2 后台数据分析服务设计

在利用 Spark 技术成功建立平台框架的基础上，为了进一步提高任务执行效率，参考 MR 数据网络质量分析系统，结合数据分析编排器，对 MR 数据网络进行优化，实现大数据智能分析平台的后台分析服务设计。MR 数据网络质量分析系统是由 RESTful API 开源包，通过对网络覆盖范围内多角度信息解码而获取的以 JSON 为标准格式的多节点分析系统。MR 数据网络质量分析系统的节点功能如图 2 所示。

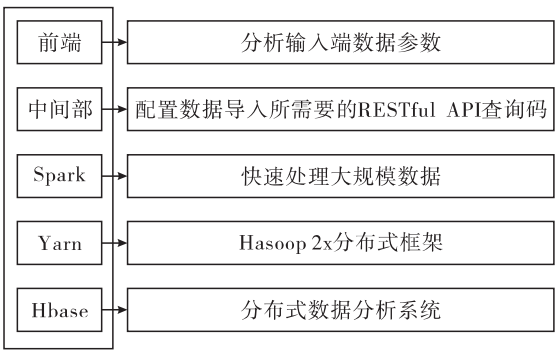

图 2 MR 数据网络质量分析系统的节点功能

数据分析编排器的整体框架借助上述 MR 数据网络质量分析系统，在节点功能不变的前提下，加入 Parquet 运算公式，获取基于可视化组件的数据源算子，为后续数据分析工作做好充足准备。Parquet 运算公式为

$$X_r=\frac{\cos\varepsilon\times[f(x_2)+f(x_{\text{est}})]}{D_w t}。$$

式中，$\cos\varepsilon$ 表示 Parquet 运算常数，$f(x_2)$ 表示 Parquet 运算公式与 MR 数据网络质量分析系统的结合紧密度[16]，$f(x_{\text{est}})$ 表示数据源算子的获取率，$D_w t$ 表示数据源算子的获取误差。数据分析编排器的整体框架如图 3 所示。数据源算子不仅能够实现多主题数据交叉编排分析，还能将输入数据与执行计划直接挂

钩，为用户提供数据驱动检索服务的同时，达到后台数据定位追踪的效果。

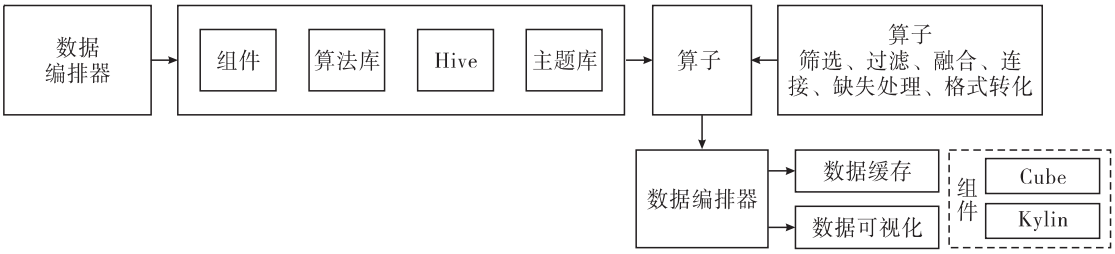

图 3　数据分析编排器的整体框架

3　实验与结果

3.1　实验方法

（1）平台运行时间计算方法。平台运行时间指的是不同方法在处理大规模数据时所消耗的时间。计算所得加速比参数越大，说明该方法的平台运行时间越短，即该方法构建的平台数据处理性能越强。加速比参数 $I^{\text{Speedup}}=\frac{q_1}{q_t}$，$t=1,2,3,\cdots$。式中，$q_1$ 表示大规模数据基于不同方法的数据处理系数，q_t 表示参与处理的数据总量。

（2）数据分析平均绝对误差计算方法。数据分析平均绝对误差 $V^{\text{MAE}}=\sum_{k\in[0,1]}^{m}p^2(x_0+x_k)/M$。式中，$p^2$ 表示平台数据总量对数据分析平均绝对误差的影响，M 表示绝对误差系数，x_0+x_k 表示数据分析平均绝对误差对平台输出结果的干扰程度。

3.2　结果与分析

为了验证基于 Spark 技术的大数据智能分析平台构建的整体有效性，需要对其进行测试。选择规模不同的三组数据库，a 组数据库内存量为 1×10^6 bit，b 组数据库内存量为 1×10^{11} bit，c 组数据库内存量为 1×10^{16} bit；分别采用不同方法建立基于三组实验数据库的智能分析平台，根据不同方法的平台运行时间、平台运行效率和数据分析平均绝对误差，推测不同方法的平台分析性能。

（1）平台运行时间。分别采用 Spark 技术、文献[2]方法和文献[3]方法建立基于三组实验数据库的智能分析平台，并计算各平台的加速比参数（图 4），进而判断不同方法的平台运行时间。由图 4 可知，采用 Spark 技术基于三组规模不同的数据库所建立的智能分析平台的加速比参数均不低 9，相较文献[2]方法提升了 4，相较于文献[3]方法提升了 2，说明 Spark 技术针对任意规模的数据库所建立的智能分析平台，其运行时间均较短，即 Spark 技术构建的平台数据处理性能较强。这是因为 Spark 技术在建立大数据智能分析平台前，首先对平台所需要的大数据降维，即将高维数据标志点嵌入低维空间，实现高维数据的低维映射，使最终构建的大数据智能分析平台运算时间下降。

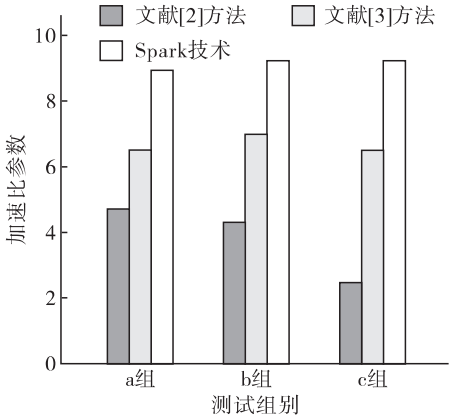

图 4　不同方法的加速比参数

（2）平台运行效率。以三组数据库为例，规定平台运行时间不得超过 30 min，观察固定时间向量时，平台内的数据处理情况。不同方法在固定时间向量下的数据处理情况如图 5 所示。由图 5 可知，Spark 技术在时间向量固定的情况下，基于 a 组数据库的智能分析平台数据处理率接近 100%，这说明 Spark 技术所建立的大数据智能分析平台数据处理效率较高。文献[2]方法和文献[3]方法在时间向量固定的情况

下，基于 a 组数据库的智能分析平台数据处理率分别不超过 80％和 70％，Spark 技术比文献[2]方法和文献[3]方法所建立的大数据智能分析平台数据处理效率高出 20％～30％。综上所述，Spark 技术具有更高的数据处理效率。

（3）数据分析平均绝对误差。采用 Spark 技术、文献[2]方法和文献[3]方法建立基于三组实验数据库的智能分析平台，并计算各平台的数据分析平均绝对误差。不同方法的数据分析平均绝对误差如图 6 所示。采用 Spark 技术基于三组规模不同的数据库所建立的智能分析平台的数据分析平均绝对误差最大值为 0.8％，说明 Spark 技术构建的智能分析平台在数据处理过程中发生错误的概率较小。文献[2]方法和文献[3]方法基于三组规模不同的数据库所建立的智能分析平台的数据分析平均绝对误差最小值分别为 1.6％和 2.2％，Spark 技术比文献[2]方法和文献[3]方法构建的智能分析平台在数据处理过程中发生错误的概率低，由此证明了基于 Spark 技术的大数据智能分析平台具有更低的数据分析平均绝对误差。

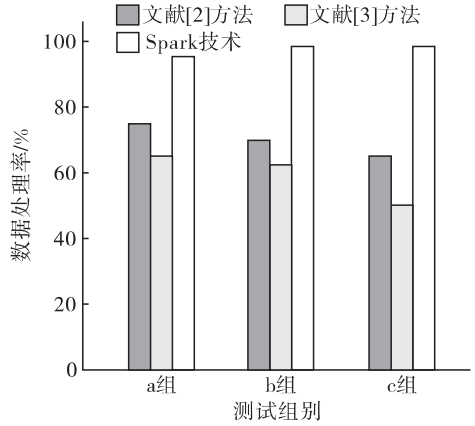

图 5 固定时间向量时的数据处理情况

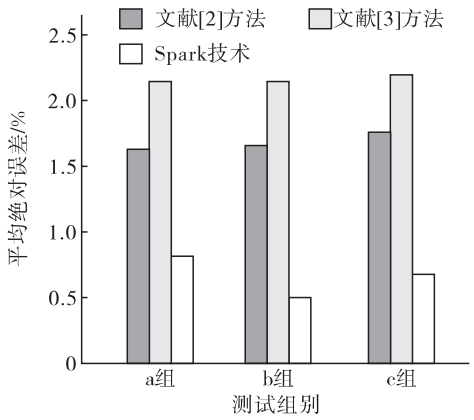

图 6 不同方法的数据分析平均绝对误差

4 结论

为了解决数据分析过程中平台运行时间较长、平台运行效率较低和数据分析平均绝对误差高的问题，提出基于 Spark 技术的大数据智能分析平台构建的方法。结果表明，所设计平台的运行时间短、平台运行效率高、平台数据分析平均绝对误差低。如何在保证大数据智能分析平台高效性的同时，对数据智能分析过程实施全程监控，是下一步研究人员需要努力的重点。

参考文献：

[1] 刘东亮，王军光，张洁，等. 基于知识单元挖掘的网络文库信息存储模型研究[J]. 情报学报，2020，39(2)：171-177.

[2] 汪杰，王春华，李晓华，等. 煤炭行业大数据分析云平台设计研究[J]. 煤炭工程，2021，53(9)：187-192.

[3] 孟光伟. 基于大数据技术的区域煤矿监管数据服务平台设计[J]. 工矿自动化，2021，47(10)：97-102.

[4] 张波，赵耀忠，刘跃，等. 基于大数据的煤矿综合业务管控平台开发[J]. 热力发电，2021，50(9)：72-79.

[5] OGIELA L，SNÁŠEL V. Towards human-oriented solutions for deep semantic data analysis[J]. Concurrency and computation practice and experience，2021，33(19)：e6252.

[6] ABUKMEIL M，FERRARI S，GENOVESE A，et al. A survey of unsupervised generative models for exploratory data analysis and representation learning[J]. ACM computing surveys，2021，54(5)：1-40.

[7] 李娟. 基于 Hadoop 云平台的空间属性数据挖掘技术研究[J]. 南京理工大学学报(自然科学版)，

2022,46(4):419-426.
[8] 刘仁芬,杨凤丽,王霞.基于改进 Spark 技术的高维数据增量式聚类算法[J].计算机仿真,2022,39(12):383-386.
[9] 杨世通,蔡燕霞,鲁国瑞,等.基于 MapReduce 的 CME 参数识别模型并行计算技术[J].空间科学学报,2020,40(2):169-175.
[10] 冷迪.基于"双态"业务的自动化 IT 构架关键技术的研究[J].微型电脑应用,2020,36(9):133-135.
[11] 袁志鑫,周艳玲.基于组合相关函数的 CosBOC(10,5)信号无模糊跟踪方法[J].计算机应用,2020,40(1):207-211.
[12] 卢思安,侯国庆.基于大数据分析技术的云计算资源预测研究[J].计算机仿真,2022,39(10):502-505.
[13] GAO H,LV C,ZHANG T,et al. A structure constraint matrix factorization framework for human behavior segmentation[J]. IEEE transactions on cybernetics,2021,52(12):12978-12988.
[14] LIU P,ZHANG F J. Pricing strategies of dual-channel green supply chain considering big data information inputs[J]. Soft computing,2022,26:2981-2999.
[15] MILLER M. Big data,information asymmetry,and food supply chain management for resilience [J]. Journal of agriculture,food systems,and community development,2021,11(1):171-182.
[16] ABBAS R,MUNOZ A. Designing antifragile social-technical information systems in an era of big data[J]. Information technology and people,2021,34(6):1639-1663.

Construction of Big Data Intelligent Analysis Platform Based on Spark Technology

JIA Shuyan
(*Computer Science Department*,*Shanxi Vocational College of Tourism*,*Taiyuan* 030031, *China*)

Abstract: To solve the problem that the big data analysis platform runs for a long time and the average absolute error of data analysis is high due to the impact of data dimensions in the process of big data analysis,a smart big data analysis platform is built based on Spark technology. The local constraint learning method is adopted to reduce the dimension of big data,and Spark technology is used to establish a platform layered architecture that combines application service layer,authority management layer,intermediate service layer and basic resource layer. Referring to the MapReduce (MR) data network quality analysis system and the design of the background analysis server of the platform in combination with the data analysis choreographer,a big data intelligent analysis platform is built based on Spark technology. The experimental results show that the acceleration ratio parameter of the built platform is greater than 9,the running speed is faster,the running efficiency is about 99%,and the average absolute error of data analysis is between 0.5% and 0.8%,which is relatively low.

Keywords: local constraint learning methods;data dimensionality;Spark technology;platform layered architecture

(责任编辑:王新亮)