big data and cognitive computing

MDPI

Editorial

Perspectives on Big Data, Cloud-Based Data Analysis and Machine Learning Systems

Fabrizio Marozzo * and Domenico Talia

Department of Informatics, Modeling, Electronics and Systems (DIMES), University of Calabria, 87036 Rende, Italy; talia@dimes.unical.it

* Correspondence: fmarozzo@dimes.unical.it

1. Introduction

Huge amounts of digital data are continuously generated and collected from different sources, such as sensors, cameras, in-vehicle infotainment, smart meters, mobile devices, social media platforms, and web applications and services [1]. Those data volumes, commonly referred to as big data, hold immense potential for extracting valuable information and generating useful knowledge in the fields of science, industry, and public services [1,2]. Extracting useful knowledge from huge digital datasets requires smart and scalable analytics algorithms, services, programming tools, and applications. Advanced data analysis techniques and tools are helping to extract patterns, trends, and hidden knowledge from big, complex datasets. Progress in this area is very useful for enabling businesses and research collaborators alike to make informed decisions.

The combination of big data analytics and knowledge discovery techniques with scalable computing systems is an effective strategy for producing new insights in a shorter period of time. Novel technologies, architectures, and algorithms have been developed to manage and analyze big data [3], enabling researchers and data scientists to extract useful information and knowledge to make new discoveries and support decision-making processes [4]. Many researchers have focused on the development of applications for big data analysis in various application fields, including trend discovery, social media analytics, pattern mining, sentiment analysis, and opinion mining. For example, from the analysis of large amounts of user data we can understand human dynamics and behaviors including (i) the main tourist attractions and mobility patterns within a city [5]; (ii) the areas of a city where it is necessary to improve the means of transport [6] or where it is more suitable to open new businesses [7]; (iii) the purchase behavior of users while browsing an ecommerce site [8]; (iv) the behavior of fans following important sporting events [9]; and (v) the political orientation of citizens and estimating the outcome of a political event [10]. To this end, the use of advanced and scalable algorithms, along with parallel programming frameworks and high-performance computers, is commonly used to solve big data problems and obtain valuable information and learning processes in a reasonable time.

From this perspective, this Special Issue aims to contribute to the field by presenting review/survey papers and original research articles in the fields of big data, cloud-based data analysis, and learning systems. Ten papers have been accepted for publication to this Special Issue, which focus on different topics.

The first paper [11] proposes the analysis of urban data to discover multi-density hotspots in metropolitan areas. It examines the limitations of traditional density-based clustering algorithms in handling multi-density data and highlights the need for more suitable techniques. By comparing four approaches (DBSCAN, OPTICS-xi, HDBSCAN, and CHD) for clustering urban data, the study evaluates their performance on state-of-the-art and real-world datasets. The findings demonstrate that multi-density clustering algorithms outperform classic density-based algorithms, providing more accurate results for urban data analysis.

check for updates

Citation: Marozzo, F.; Talia, D. Perspectives on Big Data, Cloud-Based Data Analysis and Machine Learning Systems. *Big Data Cogn. Comput.* **2023**, *7*, 104. https://doi.org/10.3390/bdcc7020104

Received: 23 May 2023
Accepted: 24 May 2023
Published: 30 May 2023

The second paper [12] proposes a novel approach for analyzing customer data in large retail companies. Traditionally, customer behavior is analyzed using simple parameters such as average and variance, which fail to capture the increasing heterogeneity among customers. To address this limitation, the paper suggests representing customer survey samples as discrete probability distributions and assessing their similarities using different models. The study focuses on the Wasserstein distance, a well-defined and interpretable metric for comparing distributions, and on multiple Key Performance Indicators per store. Experimental results using real customer data validate the effectiveness of this approach in providing meaningful global performance measures.

The third paper [13] addresses the challenges and issues associated with the widespread adoption of cloud computing in healthcare corporations for analyzing big data. Technological advancements have facilitated the storage of massive amounts of data, and cloud computing has offered an ideal solution for handling such large datasets by ensuring effective data analysis, sharing, and access; however, the security and privacy of data pose significant concerns, especially when dealing with patients' data. The objective of this study is to highlight the security challenges that hinder the widespread adoption of cloud computing in healthcare corporations.

The fourth paper [14] explores the application of big data in the field of digital health, considering the vast amount of imaging data generated in different medical contexts. The study focuses on recent research efforts in big data analysis within the health domain, along with technical and organizational challenges. Furthermore, a general strategy is proposed for medical organizations seeking to adopt or leverage big data analytics. The study aims to provide healthcare organizations and institutions, both considering and utilizing big data analytics, with a comprehensive understanding of its potential applications, effective targeting, and expected impact.

The fifth paper [15] examines the integration of artificial intelligence technologies to help curb the spread of the COVID-19 pandemic. It assesses various applications and deployments of modern technology, including image processing, disease tracking, outcome prediction, and computational medicine. A comprehensive search of COVID-19-related technology databases was conducted, and the findings were reviewed to explore the potential of technology in addressing the pandemic. While there is existing research on the use of technology for COVID-19, the full extent of its application is still being explored. The study also identifies open research issues and challenges in deploying AI technology to combat the global pandemic.

The sixth paper [16] explores the utilization of big data in healthcare and drug detection sectors. Here the challenge lies in managing and extracting valuable insights from the enormous amount of data generated by patients, hospitals, sensors, and healthcare organizations. Big data has the potential to transform drug development and safety testing in pharmacology, toxicology, and pharmaceutics by providing deeper insights into drug effects on human health; however, challenges include specialized skills and infrastructure requirements. The survey highlights the current applications, challenges, and solutions in using big data in these fields, emphasizing the need for further research.

The seventh paper [17] reviews the literature on big data applications and analytics, highlighting their importance in making strategic decisions, particularly during the COVID-19 pandemic. It compares the use of big data applications in different industry fields (healthcare, education, transportation, and banking) before and during the pandemic. The paper emphasizes the significance of aligning big data applications with relevant analytics models in the COVID-19 era, as they can address the limitations faced by organizations. Additionally, the critical challenges of big data analytics and applications during the pandemic have been investigated.

The eighth paper [18] offers a comprehensive overview of two popular data management platforms in the area of big data analytics: data warehouses and data lakes. It covers the definitions and features of these platforms and existing research related to them, along

with architecture and design considerations. The paper concludes by discussing challenges and suggesting promising research directions for the future.

The ninth paper [19] examines the role of big data in the construction industry, exploring trends and identifying opportunities for improvement. Despite the availability of data and digital technologies such as CAD and BIM, the construction industry has been slow in utilizing big data effectively. The paper analyzes the existing literature to highlight gaps and explore ways that big data analysis and storage can be applied in the construction sector, and suggests future opportunities in areas such as construction safety, site management, heritage conservation, and project waste minimization and quality improvements.

Finally, the tenth paper [20] focuses on the emerging paradigm of Edge Intelligence (EI) as a solution to overcome the limitations of cloud computing in the development and provision of IoT services. It conducts a systematic analysis of the state-of-the-art literature on EI, including literature reviews, surveys, and mapping studies, following the PRISMA methodology. The paper provides a comparison framework and identifies research questions to explore the past, present, and future directions of the EI paradigm and its relationship with IoT and cloud computing. The analysis aims to benefit both experts and beginners in understanding and advancing the field of EI.

2. Future Research Directions

Solving problems in science and engineering was the first motivation for inventing computers capable of calculating complex formulas and equations. Today, science and engineering are still the main areas in which innovative solutions and technologies are being developed and applied, although business and industry play a key role in the exploitation of advanced computing solutions. As the data scale increases, we must address new challenges and attack ever-larger problems. New discoveries will be achieved and more accurate investigations can be carried out due to the increasingly widespread availability of large amounts of data and of high-performance computer systems.

Within the scope of this Special Issue, there are several promising research directions that warrant further exploration, particularly in the area of big data and cloud-based data analysis. One significant area is the effective management and extraction of insights from vast-scale data archives. For instance, a pertinent challenge involves designing and optimizing data-intensive computing platforms capable of accommodating an extensive number of CPU cores, such as those of exascale systems [21,22]. These systems demand the management of millions of threads across an extensive array of cores to ensure optimal performance. To achieve this, it becomes imperative for data-intensive applications to minimize synchronization, reduce communication and remote memory usage, and adeptly handle potential software and hardware faults. Presently, no existing programming languages, frameworks, or infrastructures offer comprehensive solutions to tackle exascale complex issues, especially when it comes to data-intensive applications. For these reasons, in the coming years there will be a pressing need to develop new tools and technologies to unlock the full potential of exascale systems and realize their potential in pushing forward the boundaries of scientific research, big data analytics, and computational simulations.

Another significant area of attention lies in the convergence of high-performance computing (HPC), data analytics (DA), and artificial intelligence (AI) [23] for the analysis of large volumes of data. The expansion of traditional HPC applications to encompass DA and AI tasks raises challenges due to the lack of suitable programming models, environments, and deployment tools for seamless integration. To address these new challenges, ongoing efforts are focused on developing new platforms that leverage specialized software stacks capable of effectively managing big data applications and workflows. Embracing these advancements will facilitate the seamless integration of HPC, DA, and AI, resulting in the improved efficiency and scalability of big data application execution in large-scale computing environments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Belcastro, L.; Marozzo, F.; Talia, D. Programming Models and Systems for Big Data Analysis. *Int. J. Parallel Emergent Distrib. Syst.* **2019**, *34*, 632–652. [CrossRef]
2. Sagiroglu, S.; Sinanc, D. Big data: A review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 42–47.
3. Belcastro, L.; Cantini, R.; Marozzo, F.; Orsino, A.; Talia, D.; Trunfio, P. Programming Big Data Analysis: Principles and Solutions. *J. Big Data* **2022**, *9*, 4. [CrossRef]
4. Talia, D.; Trunfio, P.; Marozzo, F. *Data Analysis in the Cloud: Models, Techniques and Applications*, 1st ed.; Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 2015.
5. Belcastro, L.; Marozzo, F.; Talia, D.; Trunfio, P. G-RoI: Automatic Region-of-Interest detection driven by geotagged social media data. *ACM Trans. Knowl. Discov. Data* **2018**, *12*, 27. [CrossRef]
6. You, L.; Motta, G.; Sacco, D.; Ma, T. Social data analysis framework in cloud and Mobility Analyzer for Smarter Cities. In Proceedings of the 2014 IEEE International Conference on Service Operations and Logistics, and Informatics, Qingdao, China, 8–10 October 2014; pp. 96–101.
7. Ancillai, C.; Terho, H.; Cardinali, S.; Pascucci, F. Advancing Social Media Driven Sales Research: Establishing Conceptual Foundations for B-to-B Social Selling. *Ind. Mark. Manag.* **2019**, *82*, 293–308. [CrossRef]
8. Branda, F.; Marozzo, F.; Talia, D. Ticket Sales Prediction and Dynamic Pricing Strategies in Public Transport. *Big Data Cogn. Comput.* **2020**, *4*, 36. [CrossRef]
9. Cesario, E.; Marozzo, F.; Talia, D.; Trunfio, P. SMA4TD: A Social Media Analysis Methodology for Trajectory Discovery in Large-Scale Events. *Online Soc. Netw. Media* **2017**, *3–4*, 49–62. [CrossRef]
10. Marozzo, F.; Bessi, A. Analyzing Polarization of Social Media Users and News Sites during Political Campaigns. *Soc. Netw. Anal. Min.* **2018**, *8*, 1–13. [CrossRef]
11. Cesario, E.; Lindia, P.; Vinci, A. Detecting Multi-Density Urban Hotspots in a Smart City: Approaches, Challenges and Applications. *Big Data Cogn. Comput.* **2023**, *7*, 29. [CrossRef]
12. Ponti, A.; Giordani, I.; Mistri, M.; Candelieri, A.; Archetti, F. The "Unreasonable" Effectiveness of the Wasserstein Distance in Analyzing Key Performance Indicators of a Network of Stores. *Big Data Cogn. Comput.* **2022**, *6*, 138. [CrossRef]
13. Agapito, G.; Cannataro, M. An Overview on the Challenges and Limitations Using Cloud Computing in Healthcare Corporations. *Big Data Cogn. Comput.* **2023**, *7*, 68. [CrossRef]
14. Berros, N.; El Mendili, F.; Filaly, Y.; El Bouzekri El Idrissi, Y. Enhancing Digital Health Services with Big Data Analytics. *Big Data Cogn. Comput.* **2023**, *7*, 64. [CrossRef]
15. Almotairi, K.H.; Hussein, A.M.; Abualigah, L.; Abujayyab, S.K.M.; Mahmoud, E.H.; Ghanem, B.O.; Gandomi, A.H. Impact of Artificial Intelligence on COVID-19 Pandemic: A Survey of Image Processing, Tracking of Disease, Prediction of Outcomes, and Computational Medicine. *Big Data Cogn. Comput.* **2023**, *7*, 11. [CrossRef]
16. Latha Bhaskaran, K.; Osei, R.S.; Kotei, E.; Agbezuge, E.Y.; Ankora, C.; Ganaa, E.D. A Survey on Big Data in Pharmacology, Toxicology and Pharmaceutics. *Big Data Cogn. Comput.* **2022**, *6*, 161. [CrossRef]
17. Al-Sai, Z.A.; Husin, M.H.; Syed-Mohamad, S.M.; Abdin, R.M.S.; Damer, N.; Abualigah, L.; Gandomi, A.H. Explore Big Data Analytics Applications and Opportunities: A Review. *Big Data Cogn. Comput.* **2022**, *6*, 157. [CrossRef]
18. Nambiar, A.; Mundra, D. An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big Data Cogn. Comput.* **2022**, *6*, 132. [CrossRef]
19. Munawar, H.S.; Ullah, F.; Qayyum, S.; Shahzad, D. Big Data in Construction: Current Applications and Future Opportunities. *Big Data Cogn. Comput.* **2022**, *6*, 18. [CrossRef]
20. Barbuto, V.; Savaglio, C.; Chen, M.; Fortino, G. Disclosing Edge Intelligence: A Systematic Meta-Survey. *Big Data Cogn. Comput.* **2023**, *7*, 44. [CrossRef]
21. Da Costa, G.; Fahringer, T.; Rico-Gallego, J.A.; Grasso, I.; Hristov, A.; Karatza, H.D.; Lastovetsky, A.; Marozzo, F.; Petcu, D.; Stavrinides, G.L.; et al. Exascale machines require new programming paradigms and runtimes. *Supercomput. Front. Innov.* **2015**, *2*, 6–27.
22. Talia, D.; Trunfio, P.; Marozzo, F.; Belcastro, L.; Garcia-Blas, J.; del Rio, D.; Couvée, P.; Goret, G.; Vincent, L.; Fernández-Pena, A.; et al. A Novel Data-Centric Programming Model for Large-Scale Parallel Systems. In Proceedings of the Euro-Par 2019: Parallel Processing Workshops, Göttingen, Germany, 26–30 August 2019; pp. 452–463.
23. Ejarque, J.; Badia, R.M.; Albertin, L.; Aloisio, G.; Baglione, E.; Becerra, Y.; Boschert, S.; Berlin, J.R.; D'Anca, A.; Elia, D.; et al. Enabling dynamic and intelligent workflows for HPC, data analytics, and AI convergence. *Future Gener. Comput. Syst.* **2022**, *134*, 414–429. [CrossRef]