applied sciences

MDPI

Editorial

Cloud Computing for Big Data Analysis

Fabrizio Marozzo * and Loris Belcastro

Department of Informatics, Modeling, Electronics and Systems (DIMES), University of Calabria, 87036 Rende, Italy

* Correspondence: fmarozzo@dimes.unical.it

1. Introduction

With the spread of the Internet of Things, large amounts of digital data are generated and collected from different sources, such as sensors, cameras, in-vehicle infotainment, smart meters, mobile devices, applications, and web services. The large volume of data produced daily, coupled with the speed with which such data are generated and its heterogeneity, have led to interesting new technological challenges in the collection, storage, and analysis of this data. Those data volumes, commonly referred to as big data, can be exploited to extract useful information and produce helpful knowledge for science, industry, and public services [1,2]. Novel technologies, architectures, and algorithms have been developing to capture and analyze big data [3]. For example, in scientific and business fields, researchers and data scientists are analyzing big data to extract information and knowledge useful for making new discoveries and supporting decision processes [4].

Many researchers focused their studies 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 the following: (i) the main tourist attractions and also the 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 behavior purchase of users while browsing an ecommerce [8]; (iv) the behavior of fans following important sporting events [9]; and (v) the political orientation of citizens and then estimates the outcome of a political event [10].

In this context, cloud computing is a valid and cost-effective solution for supporting big data storage and executing data analytic applications. Cloud computing can be defined as a distributed computing paradigm in which all resources, dynamically scalable and often virtualized, are provided as services over the Internet. As defined by NIST (National Institute of Standards and Technology) [11], cloud computing can be described as follows: "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction". From the NIST definition, we can identify five essential characteristics of cloud computing systems, which are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. Due to elastic resource allocation and high computing power, cloud computing represents a compelling solution for big data analytics, allowing faster data analysis and resulting in more timely results and greater data value.

2. Cloud Computing for Big Data Analysis

From this perspective, this Special Issue aimed to contribute to the field by presenting the most relevant advances in this research area. Specifically, key scientific fields discussed in the papers that have been selected for this Special Issue include the following:

check for updates

Citation: Marozzo, F.; Belcastro, L. Cloud Computing for Big Data Analysis. *Appl. Sci.* **2022**, *12*, 10567. https://doi.org/10.3390/app122010567

Received: 19 September 2022
Accepted: 18 October 2022
Published: 19 October 2022

- Programming models and algorithms for distributed computing environments;
- Systems for data processing on cloud platforms;
- Data analysis workflows for distributed environments;
- Scalable data mining algorithms;
- Programming models and scalable algorithms for big data;
- Big data analytics and applications;
- Applications of machine learning in big data;
- Cloud-based data mining applications;
- Libraries, algorithms, and applications for big social data analysis.

There were five papers accepted for publication to this Special Issue, which focus on different topics. The first paper [12] proposes a machine learning approach to predict energy demands in the household, which exploits the Random Forest algorithm in a horizontal and vertical scaling cloud environment. Specifically, to ensure scalability and availability for large data volumes, the application has been designed to be executed on a Spark cluster, using the machine learning algorithms included in the native MLlib library.

The second and third papers focus on problems related to the efficiency and throughput of virtual machines and containers in the cloud environment, including performance overload and adaptive resource management issues for big data and scientific workflows. In particular, the second paper [13] investigates the benefits of using the vertical scalability of Docker for implementing an adaptive resource management scheme for big data workloads in a container-based cloud environment. During the execution, the adaptive resource manager scheme periodically monitors the resource usage of running containers and dynamically adjusts allocated computing resources, which results in substantial improvements in the overall system throughput. Instead, the third paper [14] focuses on addressing throughput and efficiency problems of virtual machines and containers in the cloud, exploiting different efficient approaches for resource provisioning that combine four CPU technologies and methods: hyperthreading, vCPU cores selection, vCPU affinity, and the isolation of vCPUs.

The fourth paper [15] presents three social big data analysis applications, defined and executed in parallel on a cloud platform by using ParSoDA [16], a programming library written in Java that enables developers to create cloud-based parallel applications for analyzing large volumes of social media data. Such applications focused on analyzing data from three different perspectives: (i) discovering the main tourist attractions and also the mobility patterns (i.e., trajectories) from geotagged posts [17]; (ii) understanding the political orientation of social media users so as to predict the outcome of political events [18]; (iii) analyzing the hashtags used by social media users to discover the main topics underlying social media conversation and how users refer to them in publishing online content [19].

Finally, the latest paper [20] investigates the use of two supervised classification algorithms (i.e., Random Forest and K-Nearest Neighbor) to predict the behavior of criminal networks and turn it into useful information using natural language processing (NLP). Specifically, the authors extracted an unstructured database containing data on the crimes committed. Then, to estimate the criminals' next actions, the authors performed a hotspot-based spatial analysis, for which its results are sent to two different classifiers for classification and prediction.

3. Future Directions for Research

Although the Special Issue has been closed, substantially more research can be conducted in the context of big data and cloud-based analyses in which many issues need to be addressed, particularly regarding the management and mining of large-scale data archives. As an example, an open issue is the design and optimization of data-intensive computing platforms with a very large number of CPU cores, such as the recent exascale systems. Exascale systems refer to highly parallel computing systems that are capable of at least one exaFLOPS. Therefore, their implementation represents a major challenge from a

technological and research point of view [21]. The design and development of Exascale systems is currently under investigation. Programming paradigms traditionally used in HPC systems (e.g., MPI, OpenMP, OpenCL, Map-Reduce, and HPF) are not sufficient/appropriate for programming software designed to run on systems composed of a very large set of computing elements. To reach Exascale size, it is required to define new programming models and languages that combine abstraction with both scalability and performance [22]. Hybrid models (shared/distributed memory) and communication mechanisms based on locality and grouping are currently investigated as promising approaches.

Data-intensive applications running on Exascale systems need to control millions of threads running on a very large set of cores. Such applications will need to avoid or limit synchronization, use less communication and remote memory, and handle software and hardware faults that could occur. Currently, no available programming languages provide solutions to these issues, especially when data-intensive applications are targeted. From a software point of view, these new IT platforms open great problems and challenges for software tools and runtime systems, which must be able to handle an extremely high degree of parallelism, communication, and data locality. Porting existing data analysis algorithms (or developing new ones) and designing novel fine-grained runtime models to exploit the exascale hardware will be a focus of research in the coming years.

Author Contributions: All the authors contributed equally to the structuring, writing and review of this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. 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.
2. 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]
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. [CrossRef]
11. Mell, P.; Grance, T. The NIST Definition of Cloud Computing. NIST Special Publication. 800-145. 2011. Available online: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf (accessed on 17 October 2022).
12. Cáceres, L.; Merino, J.I.; Díaz-Díaz, N. A Computational Intelligence Approach to Predict Energy Demand Using Random Forest in a Cloudera Cluster. *Appl. Sci.* **2021**, *11*, 8635. [CrossRef]
13. Choi, J.Y.; Cho, M.; Kim, J.S. Employing Vertical Elasticity for Efficient Big Data Processing in Container-Based Cloud Environments. *Appl. Sci.* **2021**, *11*, 6200. [CrossRef]
14. Shah, S.A.R.; Waqas, A.; Kim, M.H.; Kim, T.H.; Yoon, H.; Noh, S.Y. Benchmarking and Performance Evaluations on Various Configurations of Virtual Machine and Containers for Cloud-Based Scientific Workloads. *Appl. Sci.* **2021**, *11*, 993. [CrossRef]
15. Belcastro, L.; Cantini, R.; Marozzo, F. Knowledge Discovery from Large Amounts of Social Media Data. *Appl. Sci.* **2022**, *12*, 1209. [CrossRef]

16. Belcastro, L.; Marozzo, F.; Talia, D.; Trunfio, P. ParSoDA: High-Level Parallel Programming for Social Data Mining. *Soc. Netw. Anal. Min.* **2019**, *9*, 4. [CrossRef]
17. Belcastro, L.; Marozzo, F.; Perrella, E. Automatic detection of user trajectories from social media posts. *Expert Syst. Appl.* **2021**, *186*, 115733. [CrossRef]
18. Belcastro, L.; Cantini, R.; Marozzo, F.; Talia, D.; Trunfio, P. Learning political polarization on social media using neural networks. *IEEE Access* **2020**, *8*, 47177–47187. [CrossRef]
19. Cantini, R.; Marozzo, F.; Bruno, G.; Trunfio, P. Learning sentence-to-hashtags semantic mapping for hashtag recommendation on microblogs. *ACM Trans. Knowl. Discov. Data* **2021**, *16*, 32. [CrossRef]
20. Umair, A.; Sarfraz, M.S.; Ahmad, M.; Habib, U.; Ullah, M.H.; Mazzara, M. Spatiotemporal Analysis of Web News Archives for Crime Prediction. *Appl. Sci.* **2020**, *10*, 8220. [CrossRef]
21. Da Costa, G.; Fahringer, T.; 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.; FernÃ ndez-Pena, A.; et al. A Novel Data-Centric Programming Model for Large-Scale Parallel Systems. In *Lecture Notes in Computer Science: Proceedings of the Euro-Par 2019: Parallel Processing Workshops*; Springer: Cham, Switzerland, 2020; pp. 452–463.