International Journal of Parallel, Emergent and Distributed Systems

ISSN: 1744-5760 (Print) 1744-5779 (Online) Journal homepage: https://www.tandfonline.com/loi/gpaa20

Clouds for scalable Big Data processing

Paolo Trunfio & Vladimir Vlassov

To cite this article: Paolo Trunfio & Vladimir Vlassov (2019): Clouds for scalable Big Data processing, International Journal of Parallel, Emergent and Distributed Systems, DOI: 10.1080/17445760.2019.1580709

To link to this article: https://doi.org/10.1080/17445760.2019.1580709

INTERNATIONAL JOURNAL OF PARALLEL, EMERGENT AND DISTRIBUTED SYSTEMS
https://doi.org/10.1080/17445760.2019.1580709

Clouds for scalable Big Data processing

Paolo Trunfio[a] and Vladimir Vlassov[b]

[a]DIMES, University of Calabria, Rende, Italy; [b]KTH - Royal Institute of Technology, Stockholm, Sweden

ARTICLE HISTORY
Received 4 February 2019; Accepted 4 February 2019

The last decade has been characterised by an exponential growth of digital data production. This trend is particularly strong in scientific computing. For example, in the biological, medical, astronomic and earth science fields, very large data sets are produced every day from the observation or simulation of complex phenomena. At the same time, new massive sources of digital data have emerged. These include social media platforms such as Facebook, Instagram, and Twitter which are credited among the most important sources of data production in Internet. This *Big Data* is hard to process on conventional computing technologies and demands for parallel and distributed processing, which can be effectively provided by Cloud computing systems and services. This special issue focuses on the use and modelling of Clouds as scalable platforms for addressing the computational and data storage needs of the Big Data applications that are being developed nowadays.

In the first paper [1], Belcastro et al. address the main issues in the area of programming models and systems for Big Data analysis, which are extensively used in Cloud environments. As a first contribution, the most popular programming models for Big Data analysis (MapReduce, Directed Acyclic Graph, Message Passing, Bulk Synchronous Parallel, Workflow and SQL-like) are presented and discussed. Then, the paper analyses and compares the features of the main systems implementing these models, with the aim of helping developers identifying and selecting the best solution according to their skills, hardware availability, and application needs. Specifically, the systems are compared according to four criteria: (i) level of abstraction, which refers the programming capabilities of hiding low-level details of a system; (ii) type of parallelism, which describes the way in which a system allows to express parallel operations; (iii) infrastructure scale, which refers to the capability of a system to efficiently execute applications taking advantage from the infrastructure size; and (iv) classes of applications, which describes the most common application domain of a system.

The second paper [2], by Ristov et al., focuses on the accurate scalability modelling of Cloud elastic services. The speedup and efficiency parameters provide important information about performance of a computer system with scaled resources compared with a computer system with a single processor. However, as Cloud elastic services' load is variable, it is also vital to analyse the load in order to determine which system is more effective and efficient. The paper argues that both the speedup and efficiency are not sufficient enough for proper modelling of Cloud elastic services, as the assumptions for both the speedup and efficiency are that the system's resources are scaled, while the load is constant. Accordingly, the paper defines two additional scaled systems by (i) scaling the load and (ii) scaling both the load and resources. A model is introduced to determine the efficiency for each scaled system, which can be used to compare the efficiencies of all scaled systems, regardless if they are scaled in terms of load or resources. An evaluation of the model by using Microsoft Azure is presented to confirm experimentally the theoretical analysis.

CONTACT Paolo Trunfio paolo.trunfio@unical.it; trunfio@dimes.unical.it

In the third paper [3], Altomare et al. present a data mining approach to improve consolidation of virtual machines in Cloud systems. Consolidation of virtual machines is one of the most used and well-studied strategies to reduce the energy consumption in large data centres. It has the goal of allocating virtual machines on as few physical servers as possible, while satisfying the Service Level Agreement established with users. Nevertheless, the effectiveness of a consolidation strategy strongly depends on forecasting the resource needs of virtual machines, which can be made using predictive data mining models. According to this approach, the paper presents the design and development of a system for energy-aware allocation of virtual machines, driven by predictive data mining models. In particular, migrations are driven by the forecast of the future computational needs (CPU, RAM) of each virtual machine, in order to efficiently allocate those on the available servers. An experimental evaluation, based on real-world Cloud data traces, demonstrates the benefit deriving from the use of a predictive data mining approach in terms of energy saving.

The last paper [4], by Bendechache et al., presents a parallel and distributed clustering approach to analyze spatial datasets, which is designed to run on Cloud platforms using the MapReduce model. The application of clustering techniques to very large spatial datasets presents numerous challenges such as high-dimensionality, heterogeneity, and high complexity of some algorithms. The paper describes the design and implementation of a Dynamic Parallel and Distributed Clustering (DPDC) approach that can analyse Big Data within a reasonable response time and produce accurate results, by using high-performance computing and storage infrastructure, such as that provided by Cloud systems. The DPDC approach consists of two phases: a fully parallel phase that generates local clusters, and a phase that aggregates the local results to obtain global clusters. The aggregation phase is designed in such a way that the final clusters are compact and accurate while the overall process is efficient in time and memory allocation. DPDC was thoroughly tested and compared to existing clustering algorithms. The experiments show that the approach produces high-quality results and scales up very well by taking advantage of the MapReduce paradigm.

We would like to thank all the authors for their valuable contribution to the special issue. We are also grateful to the Editor-in-Chief and the editorial office for their continuous support, and to all the reviewers for their time and dedication to the review process.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Paolo Trunfio is an associate professor of computer engineering at DIMES Department, University of Calabria, Italy. He is also co-founder and managing director of DtoK Lab S.r.l., an Italian company that provides cloud solutions for Big Data analysis. He was visiting researcher at the Swedish Institute of Computer Science in Stockholm (2007) and a research collaborator at the Italian National Research Council (2001–2002). His current research focuses on cloud computing, distributed data mining, social data analysis and peer-to-peer networks. He co-authored two books: Service-Oriented Distributed Knowledge Discovery (CRC, 2012) and Data Analysis in the Cloud (Elsevier, 2015). He is currently serving in the editorial boards of the IEEE Transactions on Cloud Computing and Future Generation Computer Systems journals. He is a Senior Member and a Distinguished Speaker of the ACM.

Vladimir Vlassov is a professor of computer systems at the Department of Computer Science, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology in Stockholm, Sweden. Before joining KTH in 1993, during the period 1985–1993, he was an assistant and associate professor at S:t Petersburg's Electrotechnical University (LETI), Russia. He was a visiting scientist at the Massachusetts Institute of Technology (1998), and the University of Massachusetts Amherst (2004). He has participated in a number of research projects funded by the European Commission, and projects funded by Swedish funding agencies and NSF USA. He is one of the coordinators of the Erasmus Mundus Joint Doctorate in distributed computing. His current research focus is on autonomic and cloud computing, data-intensive computing and stream processing.

References

[1] Belcastro L, Marozzo F, Talia D. Programming models and systems for Big Data analysis. Int J Parallel Emergent Distrib Syst. 2018;1–21. doi:10.1080/17445760.2017.1422501.
[2] Ristov S, Mathá R, Kimovski D, et al. A new model for cloud elastic services efficiency. Int J Parallel Emergent Distrib Syst. 2018;1–18. doi:10.1080/17445760.2018.1434174.
[3] Altomare A, Cesario E, Vinci A. Data analytics for energy-efficient clouds: design, implementation and evaluation. Int J Parallel Emergent Distrib Syst. 2018;1–16. doi:10.1080/17445760.2018.1448931.
[4] Bendechache M, Tari A-K, Kechadi M-T. Parallel and distributed clustering framework for big spatial data mining. Int J Parallel Emergent Distrib Syst. 2018;1–19. doi:10.1080/17445760.2018.1446210.