J Grid Computing (2018) 16:161–163
https://doi.org/10.1007/s10723-018-9439-1

Guest Editors' Introduction: Special Issue on Storage for the Big Data Era

Vlado Stankovski · Radu Prodan

Received: 25 February 2018 / Accepted: 25 February 2018 / Published online: 6 March 2018

1 Introduction

Following decades of progress in computer science, we are at the beginning of a new Big Data Era that poses significant requirements for the management and storage of data. With the emergence of the Internet of Things (IoT) it is expected that the amount of produced data in the Internet will radically grow. The generated data themselves have specific properties such as volume, velocity, variety and veracity, which all require different approaches for their management. In many situations one single storage system and solution with single policy and administration cannot be used to address the varied requirements of the end-users and their applications. Hence, various exascale storage approaches are emerging aiming at high Quality of Service, data locality, availability, distribution, energy efficiency, low operational cost and similar aspects. Undoubtedly, immense storage and data management capacity cannot be achieved without thoughtful storage designs, which take care of many important aspects.

Nowadays, in distributed computing environments such as Cloud federations, the functional and non-functional requirements for storage play an increasingly important role. Therefore, new methods, approaches and technologies are necessary to address these requirements. Innovative approaches usually include various forms of Cloud storage federations, distribution and content-delivery networks, predictive data usage algorithms, low-cost replica management strategies, multi-objective optimization of storage and so on.

The aim of this Special Issue of the Journal of Grid Computing is to investigate new approaches for Big Data management and innovative storage paradigms connected with the runtime execution of the applications on Cloud infrastructures, not yet sufficiently addressed in the literature. We looked in this Special Issue for new approaches, methods, and technologies, addressing areas such as hierarchical approaches for storing data for the IoT, Edge/Fog computing concepts for storage, data fusion technologies, software defined storage, storage-related Quality of Service models, Service Level Agreements (SLAs) for storage, measurement methods for storage properties, storage approaches focusing on virtual machine, container and other disk images, new federation forms for storage, security, privacy and other non-functional aspects of storage, and storage approaches for Open Data.

V. Stankovski (✉)
Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova cesta 2, 1000 Ljubljana, Slovenia
e-mail: Vlado.Stankovski@fgg.uni-lj.si

R. Prodan
Institute of Information Technology, AlpenAdria Universität Klagenfurt, Universitätsstraße 6567, 9020 Klagenfurt, Austria

2 Overview of Contents

The interest in this Special Issue was great and we received a number of contributions world-wide that followed a rigorous peer-review process arriving to a set of accepted publications, briefly described in the following.

The study of Ansar Rafique, Dimitri Van Landuyt and Wouter Joosen "PERSIST: Policy-Based Data Management Middleware for Multi-Tenant SaaS Leveraging Federated Cloud Storage" aims at supporting users with diverse storage and privacy requirements. It delivers several benefits to users, such as an ability to make fine-grained storage- and privacy-related policies, and dynamic reconfigurability of the underlying federated Cloud storage architecture. The multi-tenant federated Cloud storage solution is based on widely known NoSQL data stores.

The authors Binqi Zhang, Chen Wang, Bing Bing Zhou, Dong Yuan and Albert Y. Zomaya design and implement a distributed storage system called DCDedupe that efficiently and intelligently uses delta compression or deduplication to improve storage efficiency based on characteristics of data. Their study entitled "DCDedupe: Selective Deduplication and Delta Compression with Effective Routing for Distributed Storage" concludes that the storage space saving with the newly introduced methods of DCDedupe outweighs the performance penalties.

The study of Marco Meoni, Raffaele Perego and Nicola Tonellotto deals with dataset popularity predictions. Their study entitled "Dataset Popularity Prediction for Caching of Compact Muon Solenoid (CMS) Big Data" is centered around the data collections, simulation and analysis activities of the CMS experiment involving 70 sites worldwide. It is a valuable experience study, which essentially concludes that a new method called Popularity Prediction Caching (PPC) outperforms the Least Recently Used (LRU) popular cashing policy by reducing the number of cache misses up to 20% in some sites.

Privacy-preservation when dealing with large data sets is an important problem tackled by the study "A Privacy-preserving Compression Storage Method for Large Trajectory Data in Road Network" of Peipei Sui and Xiaoyu Yang. The study uses large amounts of real-world trajectory data collected from GPS applications and other mobile devices that can be used for intelligent transportation, route planning and similar applications. The designed system called PP-TrajStore provides efficient storage based on a road segment compression scheme, while preserving privacy by employing a sensitive segment generalization technology at the same time.

A group of authors Akos Hajnal, Gabor Kecskemeti, Attila Csaba Marosi, Jozsef Kovacs, Peter Kacsuk and Robert Lovas present a study entitled "ENTICE VM Image Analysis and Optimised Fragmentation", which focuses on the specific requirements for storing and retrieving Virtual Machine images. The study argues that existing approaches such as deduplication techniques for virtual machine images are not directly applicable due to various difficulties. The study proposes splitting the images into shared parts called fragments stored only once. The solution requires a relatively small set of base images and stores only the additional increments without the contents of the base images, providing significant storage space savings.

The study of Carlos Guerrero, Isaac Lera and Carlos Juiz entitled "Migration-aware Genetic Optimization for MapReduce Scheduling and Replica Placement in Hadoop" aims at optimizing file locality, file availability and replica migration cost in a Hadoop architecture. Their optimization algorithm is based on the Non-dominated Sorting Genetic Algorithm-II and simultaneously determines file block placement with a variable replicating factor and MapReduce job scheduling.

The study entitled "A New Data Layout Scheme for Energy-Efficient MapReduce Processing Tasks" of the authors Xuan T. Tran, Tien Van Do, Csaba Rotter and Dosam Hwang addresses the problem of energy efficiency when using the Yet Another Resource Negotiator (YARN) in relation to Big Data processing in the Hadoop Distributed File System (HDFS). Since available data layout schemes are in general not energy-efficient, the authors present a new data layout scheme that reduces the energy consumption at the slight expense of the mean response time of jobs.

The authors Jianwei Liao, Dong Yin and Xiaoning Peng in their study entitled "Block I/O Scheduling on Storage Servers of Distributed File Systems" present a new scheme of I/O scheduling on storage servers of distributed/parallel file systems aiming at better I/O performance. In addition, they introduce an algorithm that uses priorities for each block of I/O requests. The result is a prototype that achieves better I/O bandwidth

and less I/O time compared to the First Come First Served (FCFS) strategy.

The article "A Novel Graph-based Approach for the Management of Health Data on Cloud-based WSANs" of the authors Yacine Djemaiel, Sarra Berrahal and Noureddine Boudriga advocates the use of Temporal Conceptual Graphs (TCGs) for the storage and management of health data in Wireless Storage Area Networks (WSANs). TCGs represent the collected data and their dependencies thus providing efficient possibilities for their storage and management.

The article "iHOME: Index-based JOIN Query Optimization for Limited Big Data Storage" of the authors Radhya Sahal, Marwah Nihad, Mohamed H. Khafagy and Fatma A. Omara proposes an index-based system for reusing data called indexing HiveQL Optimization for JOIN over Multi-session Big Data Environment (iHOME). The proposed iHOME system addresses eight cases of JOIN queries which classify into three groups: Similar-to-iHOME, Compute-on-iHOME, and Filter-of-iHOME. According to the experimental results of the iHOME system using a benchmark, it is found that the execution time of eight JOIN queries using iHOME on Hive has been reduced.

3 Conclusions

In summary, after a thorough review procedure, we compiled in this Special Issue a set of valuable contributions to the state-of-the-art of "Storage for the Big Data Era" that illustrate the current progress and the pending issues in this topic. Finally, we would like to cordially thank the Editor-in-Chief of the Journal of Grid Computing, Prof. Péter Kacsuk, for allowing us to organize this Special Issue and for his continuous support during the process.