Computing (2018) 100:1–2
https://doi.org/10.1007/s00607-018-0586-9

EDITORIAL

A note on resource management techniques and systems for big data workflow processing

Rajiv Ranjan[1] · **Prem Prakash Jayaraman**[2] · **Massimo Villari**[3] · **Dimitrios Georgakopoulos**[2]

Published online: 13 February 2018

The continuous shift towards data-driven enterprises and the necessity of getting real-time insights from streaming data (e.g. tweets, web clicks) has expedited the development of dozens of big data analytics workflows (e.g. click-stream analytics). Resource management of such analytics workflows is vital, since it enables cost-effective usage of cloud services against unpredictable time-varying workloads (characterised by 3Vs—volume, velocity and veracity). A typical streaming data analytics workflow consists of three layers: data ingestion, analytics, and storage, each of which is backed by different data processing platforms (e.g. Amazon Kinesis, Apache Storm, DynamoDB, respectively) and is served by different cloud services (e.g. VM, Queues). Moreover, the application workloads processed by the data analytics workflows are heterogeneous and demand different performance and quality of service measures. Hence, elasticity management of various resources for such big data analytics workflow is both difficult and challenging.

✉ Rajiv Ranjan
raj.ranjan@ncl.ac.uk

Prem Prakash Jayaraman
pjayaraman@swin.edu.au

Massimo Villari
mvillari@unime.it

Dimitrios Georgakopoulos
dgeorgakopoulos@swin.edu.au

1 School of Computing, Urban Sciences Building, Newcastle University, 1 Science Square, Science Central, Newcastle upon Tyne NE4 5TG, UK

2 School of Software and Electrical Engineering, Swinburne University of Technology, John Street, Hawthorn, VIC, Australia

3 School of Engineering, University of Messina, Piazza Pugliatti, 1, 98122 Messina, ME, Italy

A number of inquiries have been made into the elasticity management of data-intensive workflow systems on public clouds. However, we lack effective procedures for navigating the trade-off between costs and performance of the multiple data processing tiers across workflow including ingestion, analytics, and storage all at once as a single unit. Moreover, existing research works in cloud resource management consider neither uncertainty associated with data-analytics driven workflow applications nor heterogeneous performance management requirements of workflow activities mapped to different types of hardware and software resources in datacenters. From resource management perspective, modelling and implementing algorithmic methods to cope with: (i) inherent variation of data flow behaviours across workflow activities (ingestion, analytics, and storage) and (ii) run-time performance uncertainties of datacenter environments remains a very challenging problem.

This special issue solicited papers that provided practical solutions in addressing the aforementioned challenges in big data software systems. The call for special issues received a number of submissions. After a two-phase peer review process, we have accepted three high quality papers.

The first paper titled "A parallel online trajectory compression approach for supporting big data workflow" by authors W. Han, Z. Deng, J. Chu, J. Zhu, P. Gao and T. Shah propose a novel mechanism to efficiently process real-time location big data such as trajectory streams. They propose an online parallel trajectory compression algorithm named PSQUISH-E. In order to compress a trajectory and achieve parallel processing on multiple CPU-cores, PSQUISH-E introduces a small error without introducing any additional overhead. An improvement to PSQUISH-E algorithm called as G- PSQUISH-E is proposed which takes advantage of GPU processing producing significant improvement in compression performance.

The second paper titled "GEODIS: towards the optimization of data locality-aware job scheduling in geo-distributed data centers" by authors M.W. Convolbo, J. Chou, C-H. Hsu, Y.C. Chung propose a solution for solving the makespan optimisation problem to enable data-intensive job scheduling on geo-distributed data centres. They propose a low heuristic scheduling algorithm called GeoDis that allows data locality to cope with data transfer requirements to achieve a greater performance on the makespan. Through experimental evaluations over real-time and synthetic data, the authors validate the feasibility of GeoDis and demonstrate 44% reduction in makespan of processing jobs.

Finally, the paper titled "The BIM-enabled geotechnical information management of a costruction project" by authors J. Zhang, C. Wu, Y. Wang, Y. Ma, Y. Wu, X. Mao propose a management strategy of geotechnical data that can help to integrate geotechnical information into the Building Information Modelling (BIM) of a construction project in order to realize the full life cycle management of geotechnical information. The authors tackle the issues and barriers in integrating BIM and geotechnical data that is in the form of investigation reports and geological sections. They propose, and design a workflow for centralised management of BIM and geotechnical information achieved through the development of a centralized geotechnical database and an informative geotechnical model. Evaluation outcomes demonstrate the advantages in managing the geotechnical data using the proposed management strategy.