Computing (2015) 97:333–335
DOI 10.1007/s00607-015-0449-6

EDITORIAL

Guest editorial: Special issue on data intensive cloud computing

Jinjun Chen · Surya Nepal

Published online: 12 March 2015

Big data is an emerging paradigm applied to datasets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Such datasets are often from various sources (Variety) yet unstructured such as social media, sensors, scientific applications, surveillance, video and image archives, Internet texts and documents, Internet search indexing, medical records, business transactions and web logs; and are of large size (Volume) with fast data in/out (Velocity). Various technologies are being discussed to support the handling of big data such as massively parallel processing databases, scalable storage systems, and cloud computing platforms.

Big data is an emerging research area which requires new technologies to efficiently process large quantities of data within tolerable elapsed times. Cloud computing which promises to accommodate a huge volume of data and its processing is in a position as a promising technology to deal with big data issues. This special issue is focusing on this new strategic research area to address how to use cloud computing to process big data intensive applications. This special issue has selected five papers. The first three papers are related to workflows: service selection, monitoring and scheduling. The fourth paper is related to the emerging new applications on location selection, whereas the fifth paper is related to data security. We next present the summary of the papers presented in this special issue.

The modern workflow management systems are developed using the service-oriented approach, where independent services provided by different service providers

J. Chen (✉)
University of Technology, Sydney, Australia
e-mail: jinjun.chen@gmail.com

S. Nepal
CSIRO, Clayton South, Australia
e-mail: surya.nepal@csiro.au

are selected and composed together to build a workflow. It is important to note that, in the service-oriented world, there are many services providing the same functionality that form the pool of candidate services for selection. The question is then what are the good strategies to be used to select the services for individual tasks in the workflows so that the execution of the workflows can be completed successfully. There are many alternative mechanisms that can be used to select the appropriate services. The current approaches in the literature use the trustworthiness of the services or the probability of success. None of these approaches consider the structure of the workflow such as the failure of the last task in the workflow is much more expensive than the earlier tasks. The paper titled "*Reasoning task dependencies for robust service selection in data intensive workflows*" by Wang et al. presents a risk evaluation model considering the task dependencies and impact factors of each task in the overall workflow structure. They have also shown through experiments and analysis that the application of the proposed risk evaluation model into service selection can generally improve the probability of successful completion of the workflow.

The selection of services, no matter what approaches are taken, requires continuous monitoring of services. In addition to the selection, the monitoring tool in the cloud environment supports providers or application developers in: (i) keeping their resources and applications operating at peak efficiency, (ii) detecting variations in resource and application performance, (iii) accounting the service level agreement violations of certain QoS parameters, and (iv) tracking the leave and join operations. There are many commercially available cloud monitoring tool. In order to provide the overview of such tools, Alhamazani et al. present a survey in the paper titled "*An overview of the commercial cloud monitoring tools: research dimensions, design issues, and state-of-the-art*".

In addition to service selection, the monitoring tools help to schedule the tasks in the cloud and provide higher utilization of resources. The best-effort model to allocate as many as possible resources from clouds is not always cost effective or feasible for cloud users to compute their workflow applications. Wang et al. in the paper titled "*ACS: an effective admission control scheme with deadlock resolutions for workflow scheduling in clouds*" address this problem by presenting an effective admission control scheme (ACS) that integrates a set of deadlock resolution algorithms to admit workflow instances to the system based on the available storage capacities. The proposed technique reduces the competitiveness on the finite storage and minimize the adverse impact of deadlock. This has been demonstrated via intensive simulation studies on the performance changes of a set of selected benchmark workflows.

The classical problem of location selection in operational research has a wide range of applications in decision support systems including data intensive applications such as urban planning, and outsourcing. The basic idea is how to find the optimal locations for establishing services by increasing the utility of the services. Traditional algorithms suffer from the scalability problem. Sun et al. in paper titled "*MapReduce based location selection algorithm for utility maximization with capacity constraints*" present a way of using MapReduce to overcome the scalability problem. They have conducted extensive experiments using both real and synthetic data sets of large sizes and demonstrated the efficiency and scalability of the proposed MapReduce based algorithm.

Big data technologies provide a support for data intensive computation. However, the distributed nature of the big data technologies brings the problems of data security. There is no absolute way to secure the data and data transformations in large scale distributed systems. Existing techniques rely on human experiences in analyzing and detecting anomalies and intrusions. Network visualization has gained the popularity in recent times to enhance the human perception and understanding of different types of network intrusions and attacks, and has been proven to increase the efficiency and effectiveness of network intrusion detection significantly by the reduction of human cognition process. Huang et al., in paper titled "*Using arced axes in parallel coordinates geometry for high dimensional BigData visual analytics in cloud computing*" present an arc based parallel coordinates visualization method, termed as arc coordinate plots (ACP), which is an parallel coordinate plots (PCP). They develop a prototype system for network scan detection, and show the effectiveness in visualizing multivariate datasets and detecting attacks from a variety of networking patterns, such as the features of DDoS attacks.