J Grid Computing (2014) 12:187–189
DOI 10.1007/s10723-014-9305-8

Guest Editors' Introduction: Special Issue on Service and Cloud Based Data Integration

Yanbo Han · Jianwu Wang

Published online: 7 June 2014

1 Introduction

In recent years, the trend towards publishing data on the Web is gaining momentum. People have witnessed a sharp growth of diversities and amounts of available data. Integration and synthesis of heterogeneous, autonomous and distributed data sources have been an essential and hard issue in enterprise computing, scientific computing and social computing. It is not always feasible to achieve effective data integration around definite schemas when there are difficulties in getting them cross organizational borders and when such issues as compatibility, scalability, timeliness, and user manipulation are concerned. Service Oriented Architecture (SOA) and Cloud computing have brought light to dealing with these hard issues. While Cloud computing helps to establish a scalable and better optimized infrastructure for storing, retrieving and managing such distributed and large-scale data, SOA provides a loosely-coupled and standard-friendly way to realize data integration. SOA is being employed by large projects like Data Conservancy[1] to deal with data collection, integration and discovery challenges across data repositories. Some "Data as a Service" applications, such as Google Public Data Explorer[2] and WCF Data Service,[3] target rich, on-demand and latest data provisioning.

This special issue on "Service and Cloud Based Data Integration" provides the community a dedicated forum for presenting new research and application advances in service and cloud based data integration. It focuses on the use of service and/or cloud based technologies to meet the new data integration challenges that are not well served by the current approaches. We believe this issue will be an excellent place to help the community define the current state, determine future goals, and present architectures and services for future data integration.

The data integration challenges include various aspects throughout its lifecycle. Before integration, we first need new data service models with good usability to handle the increasing volume, velocity and variety of data sources, also known as big data challenges. During integration, different integration techniques, such as business process management, workflow or mashup, need to address how to work with Service and Cloud environments and achieve secure, efficient, scalable and reliable integration. After integration,

Y. Han (✉)
Cloud Computing Research Center, North China University of Technology, No. 5 Jinyuanzhuang Road, Beijing, 100144, China
e-mail: hanyanbo@ncut.edu.cn

J. Wang
San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, MC 0505, La Jolla, CA 92093-0505, USA
e-mail: jianwu@sdsc.edu

[1] http://dataconservancy.org

[2] http://www.google.com/publicdata/directory/

[3] http://msdn.microsoft.com/en-us/data/odata.aspx

we need approaches or techniques to verify the correctness and freshness of distributed data integration results. This special issue attracted ten high quality submissions from around the world, out of which four papers were accepted for publication. These papers cover a variety of issues related to the topics of interest of this special issue, including Data Service, Semantic Matching, Mashup, Business Process Management, Workflow and so on. The remainder of this introduction serves as a guide to the content of this issue.

2 In this Issue

"A Transformation-Based Approach to Business Process Management in the Cloud" by Evert Ferdinand Duipmans et al. focuses on sensitive data protection when applying Cloud computing technologies to business process management. They propose a transformation-based approach that allows companies to control the parts of their business processes that should be allocated to their own premises and the cloud. The approach can profit from the high performance of cloud environments without exposing unwanted confidential data. With the approach, the user can annotate which activities and data should be placed in the cloud and which should be placed on-premise. Through an automated transformation, the process fragments for cloud and on-premise deployment can be generated. Finally, they discuss the challenges of developing the transformation and use a case study to demonstrate the applicability of the approach.

In "Mashroom+: An Interactive Data Mashup Approach with Uncertainty Handling", Chen Liu et al. define and analyze the uncertainty problem emerging in the on-demand data integration on the Internet. An approach called *Mashroom+* is proposed to support human-machine interactive data mashup in order to better handle uncertainties during the semantic matching process. The approach can synthesize matching results from automatic matchers as well as user feedbacks to ensure better correctness. They ran experiments with six large data sets from different sources. Their results show that the proposed approach can achieve good balance between high correctness of matching results and low user burden with real data.

"Science in the Cloud: Allocation and Execution of Data-Intensive Scientific Workflows" by Claudia Szabo et al. investigates the problem about the efficient allocation and execution of data-intensive scientific workflows to reduce execution time and the size of transferred data. It is a realistic challenge for the adoption of Cloud computing in the scientific community. They propose an evolutionary approach for task allocation on public clouds considering both data transfer and execution. They also make experimental study to compare the proposed approach and related approaches using synthetic and real-life workflows. Their experimental results show that the proposed algorithm performs similarly to existing heuristics for small workflows and achieves up to eighty percent improvements for larger synthetic workflows. To make further validation, they also compare between the scheduling results obtained by the proposed approach with that gotten by popular scientific workflow managers. The results show a ten percent improvement in runtime over existing schedulers, caused by an eighty percent reduction in transferred data and optimized allocation and ordering of tasks.

"Mining Integration Patterns of Programmable Ecosystem with Social Tags" by Yuanbin Han et al. aims to make a complete analysis of the integration patterns of application programming interfaces (APIs) and composite services based the APIs that also known as Mashups in Web-based applications. They introduce various network models by considering social tags as crucial components in exploiting the integration or usage patterns for both API and Mashup applications. Through network analysis, they present a comprehensive analysis of the programmable ecosystem in which all the Web APIs and Mashups can be covered. Their experiments verify that their approach can deliver more comprehensive analyses of the programmable ecosystem than current related studies.

3 About the Editors

Yanbo Han is a Full Professor and the Director at Research Center for Cloud Computing, North China University of Technology. He holds a Ph.D. from Technical University of Berlin. His current research interests include Internet Computing, Services

Interoperability and Composition, Dependable Distributed systems, Business Process Collaboration and Management. He has authored or coauthored over 140 papers and four books. His team has acquired over 50 intellectual properties, and five of them have been transferred to the industry. Dr. Han has organized over 20 academic events as general chairs or program chairs including eight journal special issues.

Jianwu Wang is the Assistant Director for Research at Workflows for Data Science (WorDS) Center of Excellence at San Diego Supercomputer Center (SDSC), University of California, San Diego (UCSD), U.S., and an Assistant Project Scientist at SDSC, UCSD. He is also an Adjunct Professor at North China University of Technology, China. He got his Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences in 2007. His research interests include Service-Oriented Computing, End User Programming, Scientific Workflow, Distributed Computing and Big Data. He has published over 40 papers with more than 300 citations. He is associate editor or editorial board member of four international journals, co-chair of three related workshops. He is also program committee member for over 20 conferences/workshops, and reviewer of over ten journals or books.