

Preface

Data constitute a key resource in the modern world. Big data is a popular term which has been recently used to describe the exponential growth and availability of data. In particular, big data is a new phenomenon which represents the outcome of the development and the convergence of a range of technological advances in communication and computing sciences. In particular, the radical expansion and integration of computation, networking, digital devices, and data storage has provided a robust platform for the explosion in big data as well as being the means by which big data is generated, processed, shared, and analyzed. Big Data has commonly been characterized by 3V properties which refer to huge in *Volume*, consisting of terabytes or petabytes of data; high in *Velocity*, being created in or near real time; and diverse in *Variety* of type, being structured and unstructured in nature. IDC predicts that the worldwide volume of data will reach 40 zettabytes by 2020 where 85% of all of this data will be of new data types and formats including server logs and other machine generated data, data from sensors, social media data, and many more data sources. All these varieties of data types need to be harnessed to provide a more complete picture of what is happening in various application domains.

In general, data are not useful in and of themselves. They only have utility if meaning and value can be extracted from them. Therefore, given their utility and value, there are always continuous increasing efforts devoted to producing and analyzing them. In principle, big data discovery enables data scientists and other analysts to uncover patterns and correlations through analysis of large volumes of data of diverse types. Insights gleaned from big data discovery can provide businesses with significant competitive advantages, such as more successful marketing campaigns, decreased customer churn, and reduced loss from fraud. In practice, the growing demand for large-scale data processing and data analysis applications spurred the development of novel solutions from both industry and academia.

This TLDKS Special Issue presents a representative selection of articles covering a wide range of important topics in the domain of advanced techniques for big data management. The first article “A Proxy Service for Multi-tenant Elastic Extension Tables” by Haitham Yaish et al. proposes a multi-tenant database proxy service called Elastic Extension Tables Proxy Service (EETPS) that combines each tenant relational tables and virtual relational tables and makes them act and operate virtually as one single database schema for each tenant. In particular, the service allows data to be accessed by calling functions and avoids efforts associated with writing SQL queries and backend data management code. In addition, the proposed scheme allows the service provider tenants to focus on their core business and easily create their SaaS, mobile, web, and desktop software applications.

In recent years, consumption of video streams has risen sharply. This phenomenon has played a role in shaping Internet traffic. In their article “Boosting Streaming Video Delivery with WiseReplica” Guthemberg Silvestre et al. introduce WiseReplica as an

adaptive replication scheme for peer-assisted VoD systems that enforces the average bitrate for Internet videos. WiseReplica relies on machine-learned ranking in order to save storage and bandwidth from the vast majority of non-popular contents for the most watched videos.

In the past decade, the Web has been evolving to a sink of disparate information sources which are totally isolated from each other. The technology of Linked Data promises to connect such information sources in order to enable their better exploitation by humans or automated programs. The article “A Cloud-Based, Geospatial Linked Data Management System” by Kyriakos Kritikos et al. proposes a novel, cloud-based geospatial LD management system which can scale out or scale in according to the incoming load in order to serve the respective user requests with the appropriate service level. On top of this system lies an LD-as-a-service offering which abstracts away the user from any LD publishing complexities and provides all the appropriate functionality for enabling a full LD management.

The Random Prism classifier has recently been proposed as an alternative to the popular Random Forests classifier, which is based on decision trees. In principle, Random Prism is based on the Prism family of algorithms, which is more robust to noise. The article “A Scalable Expressive Ensemble Learning Using Random Prism: A MapReduce Approach” by Frederic Stahl et al. provides a detailed and exhaustive description of Random Prism and Parallel Random Prism approaches. Additionally, the article also provides a formal theoretical scalability analysis of Random Prism and Parallel Random Prism, which examines the scalability to much larger computer clusters. This examination provides a theoretical underpinning that can be used for scalability of the MapReduce framework. It also presents a thorough experimental study of Parallel Random Prism’s scalability.

In practice, popular frameworks which are supporting the MapReduce programming model for Big Data applications do not flexibly adapt to these environments. Instead, these frameworks, including Hadoop, typically divide data evenly among worker nodes which induces the well-known problem of stragglers on slower nodes. The first invited article of this special issue “Performance Analysis of Adapting a MapReduce Framework to Dynamically Accommodate Heterogeneity” by Jessica Hartog et al. presents an alternative MapReduce framework, called MARLA, which divides each worker’s labor into subtasks, delays the binding of data to worker processes, and thereby enables applications to run faster in performance-heterogeneous environments. In addition, the article explores and characterizes the opportunity for performance gains, and identifies when the benefits outweigh the costs of the proposed approach.

In general, a Content Distribution Network (CDN) is a distributed network of servers and file storage devices that replicates content/services (e.g., files, video, audio, etc.) on a large number of surrogate systems placed at various locations, distributed across the globe. In practice, CDNs that are using cloud resources such as storage and compute have started to emerge. Unlike traditional CDNs hosted on private data centers, cloud-based CDNs take advantage of the geographical availability and the pay-as-you-go model of cloud platforms. Therefore, the Cloud-based CDNs (CCDNs) promote the content-delivery-as-a-service cloud model. The second invited article of this special issue “An overview of Cloud-Based Content Delivery Networks: Research Dimensions and State of the Art” by Meisong Wang et al. presents a comprehensive

study of Cloud CDNs. In particular, the article presents a state-of-the-art survey on current commercial and research-driven Cloud CDNs and presents an analysis of current Cloud CDN based on a comprehensive taxonomy. In addition, the article identifies some of the promising research opportunities in the Cloud CDN area.

We would like to note that the publication of this TLDKS Special Issue would not have been possible without the help of many people. First, we would like to thank all the authors who submitted their articles to this special issue. We are grateful to all the reviewers for their very valuable efforts to ensure the high quality of the selected articles for this special issue. We also acknowledge the work of Abdelkader Hameurlain, Josef Küng, and Roland Wagner, Editors-in-chief of the TLDKS journal, for their confidence and help. Finally, we are particularly grateful to Gabriela Wagner for her valuable guidance and administrative assistance during the whole process of preparing this special issue.

January 2015

Sherif Sakr
Lizhe Wang
Albert Zomaya

Transactions on Large-Scale Data- and

Knowledge-Centered Systems XX

Special Issue on Advanced Techniques for Big Data

Management

Hameurlain, A.; Küng, J.; Wagner, R.; Sakr, S.; Wang, L.;

Zomaya, A.Y. (Eds.)

2015, XI, 159 p. 64 illus., Softcover

ISBN: 978-3-662-46702-2