

# Preface

Today, a great deal of attention is devoted to the issue of managing and mining big data, whose main goal consists in efficiently representing and extracting useful knowledge from such kind of data that encompass the well-known 3V characteristics, i.e., volume, velocity, and variety. This has occurred after it was recognized that traditional approaches developed during several years of data management and mining research are not suitable to comply with such novel characteristics. Another relevant property of big data to be considered is represented by their strict coupling with emerging cloud computing environments, which try to deal with research challenges deriving from managing and mining big data via specialized architectures, platforms, and paradigms based on the principles of high performance, high availability, and resource virtualization.

Within the broad scope of big data management and mining, data warehousing and knowledge discovery from big data plays a leading role and collects a wide family of models and methodologies for devising advanced data models (e.g., multidimensional models), warehousing, OLAPing, and extracting useful knowledge from big data, via a wide spectrum of specialized warehousing/mining “predicates,” such as ETL processing, aggregation, data mart indexing, frequent pattern mining, machine learning techniques, emerging pattern mining, association rule discovery, etc. All these initiatives have a common denominator, i.e., starting from the limitations of traditional data warehousing and knowledge discovery approaches in dealing with big data, not being scalability issues is the only drawback to face-off.

Last but not least, data warehousing and knowledge discovery from big data also animates a very wide family of modern applications that, without doubt, are inspiring a plethora of novel models, techniques, and algorithms in this scientific context. Among others, relevant applications are: Web advertisement, scientific computing, social network data management, energy management systems, smart city applications, etc.

In order to fulfill the innovative requirements posed by the issue of realizing data warehousing and knowledge discovery in the big data era, this special issue on *Data Warehousing and Knowledge Discovery from Big Data* of *LNCS Transactions on Large-Scale Data- and Knowledge-Centered Systems* collects a selection of the best papers presented at 14<sup>th</sup> International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2012), held in Vienna, Austria, during September 3–6, 2012. Following its successful tradition, DaWaK 2012 attracted a large number of submissions, and, after a rigorous selection process among the accepted conference papers, only 10 papers were invited for submission to the *LNCS Transactions on Large-Scale Data- and Knowledge-Centered Systems* special issue on *Advances in Data Warehousing and Knowledge Discovery*. After two thorough review rounds, only eight papers were accepted for final publication in this special issue.

The aim of the special issue is to offer an innovative, modern research perspective on the topic of data warehousing and knowledge discovery from big data, with particular emphasis on models, methods, and algorithms, by highlighting recent top-quality contributions and results in this scientific context, and, at the same, stimulating further investigation in the research field. In the following, we provide a summary of the papers included in the special issue.

The first paper, titled “Efficient Level-Based Top-Down Data Cube Computation Using MapReduce,” by Suan Lee, Jinho Kim, Yang-Sae Moon, and Wookey Lee, identifies data cubes as essential parts of OLAP to support efficient multi-dimensional analysis over large-size data. The computation of data cube takes relevant amounts of time, because a data cube with  $d$  dimensions consists of  $2^d$  (i.e., exponential order of  $d$ ) cuboids. To build ROLAP (Relational OLAP) data cubes efficiently, many algorithms (e.g., GBLP, Pipe-Sort, Pipe-Hash, BUC, etc.) have been developed, which share sort cost and input data scans in order to reduce data computation time. Several parallel processing algorithms have also been proposed. On the other hand, MapReduce is recently emerging as an authoritative framework for processing huge volumes of data, such as Web-scale data, in a distributed/parallel manner via using a large number of computers (e.g., several hundred or thousands). In the MapReduce framework, the degree of parallel processing is more important than elaborate strategies (e.g., short-share and computation-reduction) in order to reduce total execution time. Following these main considerations, the authors propose two distributed parallel processing algorithms. The first one, called MRLevel, heavily borrows from the inherent features of the MapReduce framework. The second one, called MRPipeLevel, is based on the existing Pipe-Sort algorithm that is one of the most efficient for supporting top-down cube computation. The MRLevel algorithm tries to parallelize cube computation and to reduce the number of data scans by level at the same time. The MRPipeLevel algorithm is based on the functionalities and benefits of MRLevel, and aims at further reducing the number of data scans by pipelining at the same time. Finally, the authors also identify factors for enhancing the performance of MapReduce in order to process very huge data.

The second paper, titled “Differentiated Multiple Aggregations in Multi-dimensional Databases,” by Ali Hassan, Frank Ravat, Olivier Teste, Ronan Tournier, and Gilles Zurfluh, focuses on multidimensional databases (MDBs), which support efficient querying and analysis of data stored in a data warehouse. In this context, classical MDBs support only the calculation of a measure made by the same aggregation function while performing drilling or rotating operations (i.e., changing the analyzed slice of the underlying data cube). For instance, if we consider sales amounts, these can be calculated as the sum of the products sold by cities and years. When drilling from cities to countries, the new amounts are calculated using the same aggregation function. When the user wishes to change the aggregation function between two slices of the manipulated cube, classical MDBs no longer guarantee the validity of the calculated data, or even worse: They do not support this type of manipulation. In order to provide solutions to this limitation, the authors provide a novel conceptual model that supports (1) multiple aggregations, which associate to the same measure a different aggregation function according to analysis hierarchies, and (2) differentiated aggregation, which allows for specific aggregations at each detail level. The proposed model is based on a

graphical formalism that allows one to control the validity of aggregation functions (distributive, algebraic, or holistic). Finally, the authors also show how conceptual modeling can be used in a ROLAP environment in order to build lattices of pre-computed aggregates.

The third paper, titled “MIRABEL DW: Managing Complex Energy Data in a Smart Grid,” by Laurynas Šikšnys, Christian Thomsen, and Torben Bach Pedersen, presents research and practical results from the MIRABEL project, which focuses on the definition and development of a data management system for smart grids targeted at achieving smarter scheduling of energy consumption such that, for instance, charging of car batteries is done during the night when there is an overcapacity of green energy from windmills etc. Energy can then be requested by means of flex-offers which define flexibility with respect to time, amount, and/or price. The authors describe MIRABEL DW, a data warehouse (DW) for the management of the large amounts of complex energy data in the MIRABEL system. In more detail, they present a unified schema that can manage data both at the level of the entire electricity network and the level of individual nodes, such as a single consumer node. The schema has a number of complexities compared with typical DW schemas. These include facts about facts and composed non-atomic facts and unified handling of different kinds of flex-offers and time series. The authors also discuss alternative data modeling strategies and how specialized variants of the generic schema can be used by different node types while maintaining compatibility and consistency between them. Finally, the authors complement their analytical contributions by presenting typical queries from the energy domain, and a related performance study.

The fourth paper, titled “Modular Neural Networks for Extending OLAP to Prediction,” by Wiem Abdelbaki, Sadok Ben Yahia, and Riadh Ben Messaoud, takes into consideration limitations of classical OLAP analysis that, as the authors recognize, offers a good applications package to explore and navigate data cubes, but, unfortunately, it is limited to exploratory tasks. As a consequence, OLAP does not assist the decision maker in performing information investigation. Thus, various studies have been trying to extend OLAP to new capabilities by coupling it with data-mining algorithms. The paper stands within this trend, and introduces two major contributions. First, a multi-perspectives cube exploration framework (MCEF) is introduced. MCEF is a generalized framework designed to assist the application of classical data-mining algorithms on OLAP cubes. Second, a neural approach for prediction over high-dimensional cubes (NAP-HC) is also introduced, which extends modular neural networks (MNN) architecture to the multidimensional context of OLAP cubes, to predict non-existent measures. A pre-processing stage is embedded in NAP-HC to assist it in facing the challenges arising from the particularity of OLAP cubes. This phase consists of an OLAP-oriented cube exploration strategy coupled with a dimensionality reduction step that relies on principal component analysis (PCA). The experiments described highlight the efficiency of MCEF in assisting the application of MNN on OLAP cubes and the high predictive capabilities of NAP-HC.

The fifth paper, titled “Cut-and-Rewind: Extending Query Engine for Continuous Stream Analytics,” by Qiming Chen and Meichun Hsu, focuses on combining data warehousing and stream processing technologies, which has proved to have great potential in offering low-latency data-intensive analytics. Unfortunately, such

convergence has not been properly addressed so far. The current generation of stream-processing systems is in general built separately from the data warehouse and query engine, which can cause significant overhead in data access and data movement, and is unable to take advantage of the functionalities already offered by the existing data warehouse systems. Starting from this evidence, the authors tackle some hard problems in integrating stream analytics capability into the existing query engine. They introduce an extended SQL query model that unifies queries over both static relations and dynamic streaming data, and they develop techniques to extend query engines to support the unified model. Also, they propose the cut-and-rewind query execution model to allow a query with full SQL expressive power to be applied to stream data by converting the latter into a sequence of “chunks,” and executing the query over each chunk sequentially, but without shutting the query instance down between chunks for continuously maintaining the application context across the execution cycles as required by sliding-window operators. They also propose the cycle-based transaction model to support continuous querying with continuous persisting (CQCP) with cycle-based isolation and visibility. In order to support their framework, the authors finalize the implementation of their approach by extending the PostgreSQL, thus resulting in a new kind of tightly integrated, highly efficient system with advanced stream-processing capability as well as full DBMS functionality. The authors demonstrate the system with the popular linear road benchmark, and report on the performance. By leveraging the matured code base of a query engine to the maximal extent, the proposed approach can significantly reduce the engineering investment needed for developing the streaming technology.

The sixth paper, titled “Mining Popular Patterns: A Novel Mining Problem and Its Application to Static Transactional Databases and Dynamic Data Streams,” by Alfredo Cuzzocrea, Fan Jiang, Carson K. Leung, Dacheng Liu, Aaron Peddle and Syed K. Tanbeer, recognizes that, since the introduction of the frequent pattern mining problem, researchers have extended frequent patterns to different useful patterns such as cyclic, emerging, periodic, and regular patterns. In line with this trend, the paper introduces popular patterns, which captures the popularity of individuals, items, or events among their peers or groups. Moreover, they also propose the Pop-tree structure for capturing the essential information from transactional databases, and the Pop-growth algorithm for mining popular patterns from the Pop-tree. The authors illustrate how the proposed algorithm mines popular friends from social networks, as a relevant case study of the proposed framework. Because the framework is not confined to mining popular patterns from static transactional databases, they extend the work to mining popular patterns from dynamic data streams. Specifically, the Pop-stream structure to capture the popular patterns in batches of data streams is proposed, as well as the Pop-streaming algorithm for mining popular patterns from the Pop-stream structure. Finally, the experimental results show that (a) the proposed tree structure is compact and space efficient and (b) the proposed algorithm is time efficient in mining popular patterns from static transactional databases and dynamic data streams.

The seventh paper, titled “Rare Pattern Mining from Data Streams Using SRP-Tree and Its Variants,” by David Tse Jung Huang, Yun Sing Koh, and Gillian Dobbie, addresses research in the area of rare pattern mining where the researchers try to capture patterns involving events that are unusual in a data set. These patterns are

considered more useful than frequent patterns in some domains, including detection of computer attacks or fraudulent credit transactions. To date, most of the research in this area has concentrated only on finding rare rules in a static data set. Nevertheless, there is a proliferation of applications that generate data streams, such as network logs and banking transactions, and applying techniques that mine static data sets is not practical for data streams. In order to fill this gap, the authors propose a novel approach called streaming rare pattern tree (SRP-Tree) and its variations, which finds rare rules in a data stream environment using a sliding window, and show that it both finds the complete set of item sets and runs with fast execution time.

Finally, the eight paper, titled “Improving Cross-Document Knowledge Discovery Through Content and Link Analysis of Wikipedia Knowledge,” by Peng Yan and Wei Jin, focuses on the research context of the vector space model (VSM), which has been widely used in natural language processing (NLP) for representing text documents as a bag of words (BOW). However, according to this model, only document-level statistical information is recorded (e.g., document frequency, inverse document frequency) and word semantics cannot be captured. Improvement in understanding the meaning of words in texts is a challenging task and sufficient background knowledge may need to be incorporated to provide a better semantic representation of texts. Following this main trend, the authors present a text-mining model that can automatically discover semantic relationships between concepts across multiple documents, where the traditional search paradigm such as search engines cannot help much, and effectively integrate various evidence mined from Wikipedia knowledge. The authors argue that this integration may effectively complement existing information contained in text corpus and facilitate the construction of a more comprehensive representation and retrieval framework. Experimental results demonstrate that the search performance has been significantly enhanced when compared with two competitive baseline methods.

The editors would like to express their sincere gratitude to the Editors-In-Chief of LNCS Transactions on Large-Scale Data- and Knowledge-Centered Systems, Prof. Abdelkader Hameurlain, Prof. Josef Küng, and Prof. Roland Wagner, for accepting their proposal of a special issue focused on data warehousing and knowledge discovery from big data, and for assisting them whenever required. The editors would also like to thank all the reviewers who have worked within a tight schedule and whose detailed and constructive feedbacks to authors have contributed to substantial improvement in the quality of the final papers.

June 2015

Alfredo Cuzzocrea  
Umeshwar Dayal

Transactions on Large-Scale Data- and  
Knowledge-Centered Systems XXI

Selected Papers from DaWaK 2012

Hameurlain, A.; Küng, J.; Wagner, R.; Cuzzocrea, A.;  
Dayal, U. (Eds.)

2015, XIII, 185 p. 84 illus., Softcover

ISBN: 978-3-662-47803-5