Editorial

Alfredo Cuzzocrea

University of Trieste and ICAR-CNR,
Piazzale Europa, 1,
Trieste, 34127, Italy
Email: alfredo.cuzzocrea@dia.units.it

Big data engineering is gaining the momentum in actual research topics, mostly stirred-up by emerging technologies and applicative settings. Among others, social networks, cloud computing systems, blockchain platforms, and so forth play a leading role, and put emphasis on methods, methodologies and techniques that overcome the limitations of traditional data management paradigms and completely marry the well-understood *4V model* (i.e., volume, velocity, variety, veracity) of big data.

Starting from this critical evidence, recently a lot of research efforts have been devoted to the big data context, with particular regards to design methodologies, formal models, parallel and distributed management algorithms, querying algorithms, indexing techniques, security protocols, and so forth. Indeed, these are still embryonic studies that must be further investigated and developed, and, most importantly, assessed by real-life (big data engineering) systems and applications.

Big data engineering embraces several research perspectives that deserve relevant attention, such as:

- *Distributed big data management:* the main research challenge here consists in devising innovative models, techniques and algorithms for managing big data in distributed environments effectively and efficiently, by considering the specific characteristics of such data.
- *Big data repository modelling:* while design methodologies are well-established in related contexts, such as relational databases, these methodologies are still not well-developed in big data research – therefore, there is an emerging need for supporting all the design phases of big data repositories, perhaps by devising ad-hoc design life-cycle paradigms that are reminiscent of well-known relational database modelling experiences.
- *Flexible integration paradigms for big data repositories:* while data heterogeneity is a specific characteristic of big data applications, big data integration is still useful for a plethora of *big data analytics* scenarios – integrating big data repositories via flexible paradigms plays a critical role in many real-life systems and applications, such as social inter-networks, where the same users share different profiles on different social networks.
- *Formal languages and models for big data repositories:* as soon as big data processing will become a dominant paradigm, then the emerging need for formal languages and models for describing and characterising big data repositories will

> become a mandatory requirement – this research area is still in its embryonic phase, hence it is easy to foresee its growth during next years, with several proposals that put roots in traditional database and software engineering methodologies.

This special issue on 'Big data engineering: recent advances in intelligent methods, methodologies and techniques' of the *International Journal of Data Mining, Modelling and Management* focuses on latest research results and open research challenges in the context of intelligent methods, methodologies and techniques for big data engineering, according to principles and guidelines provided above.

With the aim of adequately fulfilling both theoretical and practical issues deriving from intelligent methods, methodologies and techniques for big data engineering, this special issue contains four papers, which have gone through two rigorous review rounds before being accepted for the final inclusion.

In the first paper, titled 'Allegories for database modelling', by Bartosz Zieliński, Paweł Maślanka and Ścibor Sobieski, authors address the traditional *database modelling problem* in an innovative manner. Indeed, as authors recognise, allegories abstract and generalise (in the categorical framework) the algebra of binary relations. Arrows in an allegory enjoy a lot of properties and structure available for plain binary relations. At the same time, allegories are sufficiently general to allow the description within the same uniform framework the *lattice-valued* (e.g., fuzzy) relations and some more general structures. The paper presents a conceptual data modelling formalism which uses the language of allegories. Authors provide examples demonstrating expressiveness of this formalism. While most of the examples are meant to be interpreted in the allegory of sets and binary relations, authors also show the usefulness of using other allegories, such as the allegory of sets and lattice valued relations, with which one can model replicated data or data stored in a valid time temporal database.

The second paper, titled 'A grammar-based approach for XML schema extraction and heterogeneous document integration', by Prudhvi Janga and Karen C. Davis, proposes an *approach for supporting XML schema extraction and heterogeneous document integration via formal grammars*. Main motivations of authors start from recognising that the availability of vast amounts of heterogeneous XML web data motivates finding efficient methods to search, integrate, query, and present such data repositories. The structure of XML documents is useful for achieving these tasks; however, not every XML document on the web includes a schema. Therefore, authors discuss challenges and solutions in the area of generation and integration of XML schemas. They propose and implement a framework for efficient schema extraction and integration from heterogeneous XML document collections collected from the web. The proposed approach introduces the so-called *schema extended context free grammar* (SECFG) to model XML schemas, including detection of attributes, data types, and element occurrences. Unlike other implementations, the approach supports the generation of XML schemas in any XML schema language, e.g., DTD or XSD. Also, in addition to this conceptual contribution, authors compare the proposed approach with other approaches and conclude that SECFG offers the same or better functionality more efficiently and with greater flexibility, and also it is flexible enough to facilitate integration of and translation to relational data.

The third paper, titled 'Towards a comparative evaluation of text-based specification formalisms and diagrammatic notations', by Kobamelo Moremedi and John Andrew van der Poll, focuses the attention on the problem of *analysing and assessing text-based* *specification formalisms and diagrammatic notations in software engineering systems*. Specification plays a pivotal role in software engineering to facilitate the development of *highly-dependable software*. Various techniques for specification work have been developed to provide for precise and unambiguous specifications. *Z* is a formal specification language that is based on a strongly-typed fragment of Zermelo-Fraenkel set theory and first-order logic to provide for provably correct specifications. While diagrammatic specification languages may lack precision, they may, owing to their visual characteristics be a lucrative option for advocates of semi-formal specification techniques. On the basis of this main intuition, authors investigate the extent to which diagrammatic notations may capture the essence of, e.g., a *Z* specification. Several diagrammatic notations are considered and combined for this purpose. A case study is employed towards the end to evaluate the utility of the diagrammatic notation developed in this research. Finally, comparisons on the merits of a diagrammatic notation are presented in order to further determine their feasibility.

Finally, in the fourth paper, titled ‘Effective and efficient distributed management of big clinical data: a framework’, by Alfredo Cuzzocrea, Giorgio Mario Grasso and Massimiliano Nolich, authors move the attention on the *management of big data in distributed environments*, which is recognised as a critical research challenge that has driven the attention from the community. In this context, there are several issues to be faced-off, including:

1 dealing with massive and heterogeneous data

2 inconsistency problems

3 query optimisation bottlenecks, and so forth.

Clinical data represent a vibrant case of big data, due to both practical as well as methodological challenges exposed by such data, also dictated by tight requirements of applications that manage them. Following these considerations, authors present an architecture for the storage, exchange and use of health data for administrative and epidemiological purposes, which focuses on the patient, who in a safe and easy way can make use of their data for therapeutic and research purposes. The patient is the real owner and holder of her/his data, and she/he is provided with a smart card containing all the clinical reports, prescriptions or medical imagery. On the other hand, each clinical centre does not have to invest in expensive data centres, because the proposed solution takes advantage of a distributed architecture, which includes the already operating local servers. In conclusion, the proposed architecture would bring benefits both to patients, giving them the desired centrality in the care process, and to health administration, that could exploit the same infrastructure for better addressing health policies, for example implementing epidemiological and drug safety research. The proposed research has been conducted as part of a real-life project, in order to experience a new kind of storage architecture and data exchange within the field of clinical oncology.

The editor would like to thank very much the Editor-In-Chief of the *International Journal of Data Mining, Modelling and Management*, Prof. John Wang, for accepting his proposal of a special issue focused on recent advances in intelligent methods, methodologies and techniques for big data engineering, and for assisting him whenever required. The editor 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 final papers. Last but not least, the editor is grateful to the authors who have submitted papers to this special issue. The editor truly appreciates their patience and understanding throughout the review process.