Front. Comput. Sci., 2013, 7(2): 157–164
DOI 10.1007/s11704-013-3903-7

RESEARCH ARTICLE

Big data challenge: a data management perspective

Jinchuan CHEN, Yueguo CHEN, Xiaoyong DU, Cuiping LI, Jiaheng LU (✉), Suyun ZHAO, Xuan ZHOU

Key Laboratory of Data Engineering and Knowledge Engineering, School of Information, Renmin University of China, Beijing 100872, China

Abstract There is a trend that, virtually everyone, ranging from big Web companies to traditional enterprisers to physical science researchers to social scientists, is either already experiencing or anticipating unprecedented growth in the amount of data available in their world, as well as new opportunities and great untapped value. This paper reviews big data challenges from a data management respective. In particular, we discuss big data diversity, big data reduction, big data integration and cleaning, big data indexing and query, and finally big data analysis and mining. Our survey gives a brief overview about big-data-oriented research and problems.

Keywords big data, performance, databases

1 Introduction

Since information technology is innovating on the way we live, our collection of digital data has started to grow rapidly. Today, there is tremendous amount of data generated everyday in the sectors of manufacturing, business, science and our personal lives. Proper processing of the data could reveal new knowledge about our market, society and environment, and enable us to react to emerging opportunities and changes in a timely manner. However, the growth of the data volume in our digital world seems to outspeed the advance of our computing infrastructure. Conventional data processing technologies, such as database and data warehouse, are becoming inadequate to the amount of data we want to deal with. This new challenge is known as big data. Due to its importance and commonness, it has gained enormous attention in recent years.

There has not been a commonly accepted definition about big data, though people usually believe that "big data should include data sets with sizes beyond the ability of commonly-used software tools to capture, manage, and process the data within a tolerable elapsed time". Based on this concept, researchers have summarized three important aspects of big data that go beyond the ability of our current data processing technology. They are Volume, Velocity and Variety, also known as 3Vs.

The challenge posed by data volume is most noticeable. In science, such as biology, meteorology, astronomy, etc., scientists encounter computing limitation constantly due to the increasing data volume. On the Web, applications such as Google and Facebook are dealing with the numbers of customers that have never been considered by local applications. The sizes of the data sets consumed by today's Web applications can be extraordinarily big. Such big-volume issues can also be found in the areas of finance, communication and business informatics, due to the wide application of information technology and the increasing intensity of online transactions. Despite of the various big-volume issues, there is still no agreement on the quantification of big data. Such quantification depends various factors. First, the complexity of the data structure is an important factor. A relational dataset of several petabytes may not be called big data, since it can be readily handled by today's DBMSs. In contrast, a graph dataset of several terabytes is commonly regarded as big data, as graph processing is very challenging

Received January 5, 2013; accepted February 22, 2013

E-mail: jiahenglu@ruc.edu.cn; duyong@ruc.edu.cn

to our technologies. Second, the requirements of target applications should be considered as a factor too. In scientific research, a waiting time of several hours is usually acceptable for a biologist. Most automated trading systems, in contrast, require sub-second response time regardless of how big the data is. In reality, big data sizes are a constantly moving target, ranging from a few Terabytes to several Zettabytes of a single data set, depending on the particular context in which the data is used.

The challenge of velocity comes with the need to handle the speed with which new data is created or existing data is updated. This issue particularly applies to machine generated data, such as that generated by sensing or mobile devices, which are being deployed everywhere. In those applications, large amount of new and updated data flies into the systems relentlessly, while we require the systems to make sense of the data immediately upon its creation. Data velocity brings challenges to every stack of a data management platform. Both the storage layer and the query processing layer need to be extremely fast and scalable. The technology of data streaming has been investigated for several years to handle high velocity. However, the capacity of the existing streaming systems is still limited, especially when dealing with the increasing volume of incoming data in today's sensor networks, telecommunication system, etc.

In real-world applications, data often does not come from a single source. Big data implementations require handling data from various sources, in which data can be of different formats and models. This bring forth the challenge of data variety. The variety of data provides more information to solve problems or to provide better service. The question is how to capture the different types of data in a way that makes it possible to correlate their meanings. Typically, data can be classified into three general types—structured data, semi-structured data and unstructured data. There have been sophisticated technologies to deal with each of these data types, such as those of database and information retrieval. However, a seamless integration of these technologies remains as a challenge.

While the aforementioned 3Vs draw a big picture about the big data challenge, there are other issues besides the 3Vs we may encounter when dealing with big data. For instance, several 4th Vs have been proposed recently, including Variability, Value and Virtual, which refer to other aspects of data management. Moreover, the challenges vary in different application scenarios. In this paper, we do not discuss the challenges in more details. Instead, we view big data from the perspective of data management system (see Fig. 1). We aim to summarize the various emerging systems and technologies in processing big data. We also discuss the technical gaps and point out the opportunities for system researchers.

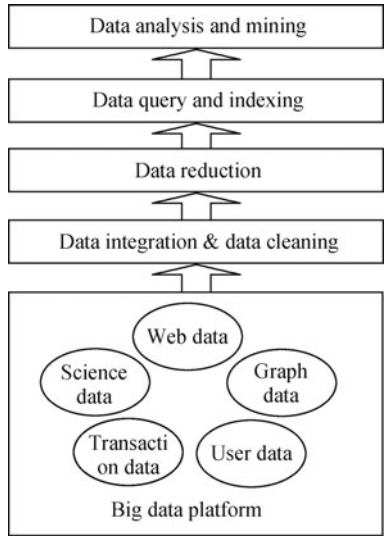

Fig. 1 Big data processing steps covered in this paper

The rest of the paper is organized as follows. In Section 2, we categorize big data in several types, and review the existing data processing technologies for each type. In Section 3, we discuss whether traditional data integration methods can be applied to big data and where the gaps are. In Section 4, we summarize the techniques for big data reduction. In Section 5, we review the various systems and indexing mechanisms for processing big data. In Section 6, we enumerate some research opportunities for data mining in the context of big data. Finally, Section 7 concludes the paper.

2 Big data variety

Nowadays, many applications in various domains are able to generate different types of big data, with many of them are unstructured and semi-structured. Heterogeneity is a natural property of the big data, considering their broad range of sources [1]. We classify the typical sources of big data according to the way they are generated:

- **User generated contents** (UGCs) from applications with massive users. Examples are tweets, blogs, discussions, photos/videos posted and shared by users of many Web 2.0 applications. The data of these applications are directly contributed by users, and therefore, they are typically unstructured for user convenience. As for unstructured data, accompanying metadata such as tags and user names are very important for under-

standing the data. Information extraction is necessary to structuralize the raw data so that they can be easily digested by analysis algorithms [2, 3].

- **Transactional data** that are generated by a large scale system due to massive operations/transactions processed by the system. Examples of big transactional data are Web logs, business transactions, feeds of moving objects, reports of sensor networks, reads of radio-frequency identifications. These data are typically structured with predefined schemas. They are often accumulated in a streaming manner.
- **Scientific data** that are collected from data-intensive experiments or applications. Examples are celestial data, high-energy physics data, genome data, healthcare data. Types of scientific data are very application-dependent, ranging from structured data (e.g., time series data) to semi-structured data (e.g., XML data) and unstructured data (e.g., images). In addition to the original data, provenance data (recording how data are generated and transformed) are very important for scientific data management [4].
- **Web data** that are crawled and processed to support applications such as Web search and mining. As the World Wide Web contains billions of pages, it is quite easy to generate a huge Web corpus of numerous unstructured Web pages. Behind the Web pages, there are also a huge amount of deep web data which are even more important than the surface contents. They can also be crawled and integrated as important sources of Web data [5, 6].
- **Graph data** that are formed by a very huge number of information nodes, and the links between the nodes. Examples are social networks and RDF knowledge bases [7]. Although structured, graph data are more expensive to process than relational data because ad-hoc local topology (pattern) in graphs complicates the processing of graph data.

In many cases, big data do not have neat relational structures. They often contain information of diverse types such as texts, images, tags, metadata, provenance data etc. It causes that big data cannot be simply abstracted using single data model. Variety is an important factor of big data, that has to be addressed by big data management systems. Different strategies can be applied for different types of big data.

For big relational data, the performance of most existing commercial DBMSs running on single node drops significantly when a relation is larger than hundreds of gigabytes. The strong requirements of ACID properties constrain the performance of traditional RDBMs. Distributed and parallel data management solutions such as OceanBase [8] have been tried to address the scalability problem for online processing large scale relational data. In-memory databases such as VoltDB and HANA [9] have been recently commercialized to response the performance challenges for OLTP and OLAP applications of big data.

For big graph data, many solutions on efficient processing large graphs (e.g., Neo4j [10] and Pregel [11]) that cannot be held in memory have been proposed recently. Pregel [11] is a famous one of them. In addition, numerous algorithms [7] have been proposed for various graph mining and management applications, due to the nature of high computational complexity of graph algorithms.

For big unstructured data, they cannot be effectively understood and efficiently processed when in raw format. As such, information extraction techniques have been widely applied to extract important and manageable structured data from the raw unstructured ones [12]. Finally, the big unstructured data are processed through solutions on the extracted structured data. The extracted data are summaries and sketches of the original unstructured data. There must be information loss after the transformation and reduction. However, in many applications, the data are so large that a small summary of them may be accurate enough to support the needs of analyzing and processing big data. Solutions on how to effectively reduce and transform big data are therefore very important.

3 Big data integration and cleaning

Recently big data becomes the major challenge for data integration and cleaning. To highlight its impact, we conduct an experiment and want to see the change of research interests during the past ten years. For this purpose, we select 161 papers focusing on data integration and cleaning which are published after 2002 and are from top database conferences like SIGMOD and VLDB. We partition these papers into two sets according to their publish year. The first set contains all the papers appearing after 2008, and the second one contains those published between 2002 and 2007. The titles and abstracts of these papers are used as the corpus. We then compare the frequent keywords from these two sets. The results are listed in Fig. 2. As shown in the figure, these frequent keywords are split into three groups. The first group, emerging keywords, includes the terms that become popular in the last five years but attract little interests before 2008, such as

feedback, probabilistic etc. These emerging keywords could help us to understand the influence of big data on data integration and cleaning since big data is the dominant change in recent years. We will explain this in more details soon. The second group, i.e., fading keywords, contains those words which were quite popular before 2008 and become cold recently, e.g., the issues related to XML, DTD. Finally, the last group, evergreen keywords, consists of the ones keeping popular during the past ten years. These words, like integrate, mapping and schema are exactly the ones which could depict the basic process of data integration and cleaning.

As implied by the results shown in Fig. 2, there are some new coming research interests in the field of data integration and cleaning. Now we briefly summarize them in the below-ing.

Emerging keywords	feedback, human, user, probabilistic, uncertain, provenance, linkage, entity, pay-as-you-go
Fading keywords	XML, DTD metadata, wrapper
Evergreen keywords	constraint, demain, integrate, mapping, match, schema, semantic

Fig. 2 Keywords about data integration and cleaning

• **User feedback and crowdsourcing** Due to the limit of artificial intelligence, there are always some errors in schema mappings generated by automatical processes. Traditionally these errors are expected to be fixed by domain experts. But this approach cannot work in the era of big data. Many researchers then suggest to utilize users and/or crowds for improving the quality of integrated data [13–17]. In [13], the authors propose a method to determine the order to confirm user feedbacks by evaluating the utilities of candidate matches. In [14, 16], several algorithms are proposed to automatically incorporate user feedbacks and update existing data integration programs. A recent work [17] tries to make use of crowd in data integration tasks. Note that crowd may not be the users of data integration system, and crowdsourcing usually means to initiatively pull knowledge from lots of people.

• **Uncertainty and provenance** The increase of interests in uncertainty and provenance somehow confirms the prediction by Alon Halevy in 2006 [18] that uncertainty and lineage will become a major challenge for data integration. Data uncertainty has become an intrinsic property in many data integration applications. We have to conduct integration and cleaning based on the imprecise data [19–22]. For this purpose, usually a probabilistic model would be constructed to represent the data uncertainty and to make imprecise decisions. Specifically, in [23, 24], the authors try to build imprecise and probabilistic schemas and matchings on the basis of uncertain data. Furthermore, we also need to maintain the provenance of data to track the linkage [25, 26].

• **Pay-as-you-go** It is impossible to build a perfect integration for big data, because of the large volume and the high velocity of data accumulation. Hence a reasonable way is to construct an imperfect system which could provide necessary service, and to incrementally improve this system when there are more resources available like time and money [13, 20, 21, 23]. This is exactly what the idea pay-as-you-go suggests. Note that a pay-as-you-go data integration system usually also concerns about the data uncertainty and provenance [20, 21, 23], and tends to utilize human beings for improving data quality [13].

• **Entity matching and resolution** The goal of entity resolution is to identify which records (entities) refer to the same real-world entity, which is a fundamental task in data integration. In recent years, this topic receives more and more attention because there are increased interests in extracting and integrating tables from web pages. Lots of approaches are proposed to improve the quality of entity resolution, e.g. combining different methods, iterative approach, and using functional dependencies etc. [27–29].

4 Big data reduction

Data reduction is the reduction of multitudinous amounts of data down to the meaningful parts. Further speaking, data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. On considering big data, making a profound transformation in computing, such as different sampling methods, aggregation (computing descriptive statistics), dimensionality reduction techniques, etc., is a feasible and effective approach before big data analysis and management. Instead of operating on complex and large raw data directly, big data reduction tools enable the execution of various data analytic and manage tasks. Therefore big data open new chances and challenges to these techniques which have been well-studied in the past. Furthermore, big data open doors for interesting new approaches of data reduction. Currently, there are roughly two main challenges for big data reduction as follows.

Machine learning is a possibly feasible way to improve traditional data reduction techniques to process or even preprocess big data. Big data requires exceptional technologies

to efficiently process large quantities of data within tolerable elapsed times, because the traditional data reduction techniques may not glance over all records, not even to have records. As a result, some techniques of machine learning may help to understand the trends of data, classify big data into categories, detect similarities and predict the future based on the past. Leveraging a fully parallel machine learning solution on big data will help to identify fraud, bring products to market faster, and become more competitive.

Massively parallel processing is another possibly feasible way to reduce big data. Some massively parallel processing technologies including massively parallel-processing (MPP) databases, data-mining grids, distributed file systems, distributed databases, cloud computing platforms, and scalable storage systems, have been proposed. Among them, cloud computing is one of the successfully gathering momentum [30]. Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. By improving the techniques of cloud computing platform, it may be used to reduce data with the help of massively parallel processing.

5 Big data query and indexing

In the era of big data, various forms of data from all kinds of fields have walked into every corner of our life. When it comes to query and indexing of big data, some challenges arise inevitably. First, the size of digital information in the age of big data is too huge for most softwares and people to manage and process. Also, a single machine cannot hold the sea-like big data, which should be stored in a distributed system. Therefore, the index of big data, distinguishing from the traditional index structure, should be built based on distributed system and corresponding new query theory should be encouraged. Second, big data not only refers to data sets that are very large in size, but also covers data sets that are complex in structure, high dimensional and heterogamous. It is all of these factures that make big data become a real challenge for us. Consequently, traditional methods for indexing and query with small structured data sets are not adequate any more. Third, tree-like structure enjoys magnificent popularity in traditional indexing and query field, while in the realm of big data, the tree-like index does not work that well, as it is hard to avoid bottleneck in the tree-like structure when providing very good concurrent reading and writing operations. Besides, fault tolerance is an important factor that cannot be neglected in big data query and indexing.

Confronted with all these challenges above, the researchers have put forward several trial methods. To some extent, a constructive step has been made. Among these methods, distributed B-tree [31] is a practicable one, whose goal is to perform consistent concurrent updates while allowing high concurrency reading operations. The distributed B-tree index can provide transactional access, by use of optimistic concurrency control strategy. Moreover, it can supply online migration of tree nodes and dynamic addition and removal of severs. Another proper way to meet the challenge is using BATON overlay [32] to support range queries. BATON, short for Balanced Tree Overlay Network, a balanced tree structure overlay for peer-to-peer networks, is capable of both exact query and range query. In an N nodes network, BATON can guarantee that both exact query and range query can be answered in $O(\log N)$ steps and update operations have amortized cost of $O(\log N)$. Further more, CG-index [33], short for Cloud Global index, is a scalable indexing scheme for cloud data management systems. An indexing framework based on CG-index can reduce the amount of data transferred inside the Cloud and facilitate the deployment of database back-end applications.

On account of these characteristics of big data, an effective method to improve query on big data is through summarization, such as different sampling methods, histograms, sketches and synapses, low-rank subspace approximation, dimensionality reduction techniques, etc. When summarizing big data in parallel and distributed environment, we should also qualify the accuracy and efficiency trade-off, without ignoring scalability.

6 Big data analysis and mining

As we all believe, once we own the good way to analyze and mine the big data, it can bring us the big value. However, due to its noisy, dynamic, heterogeneous, inter-related and untrustworthy properties, the analysis and mining of the big data is very challenging.

In order to make use of the big data to guide people's decision, "deep analysis", instead of just generating simple report forms, is needed. This kind of complex analysis must depend on complex analysis models and is difficult to be expressed as SQL. In order to make some active preparations, people not only need to know what is happening currently, but also need to predict what will happen in the future by analyzing data.

For example, if the risk of losing customers can be forecasted, effective actions could be taken to retain them. In this respect, typical OLAP data analysis operations (aggregating, slicing, dicing, rotating, etc.) are not enough. Other types of complex analysis such as path analysis, time series analysis, graph analysis, what-if analysis should be tried. .

Due to its limited expansibility, typical relational database technologies meet unprecedented difficulties when they are used to address these challenges of deep analysis on big data. In 2004, Google first proposed the MapReduce framework. As a paralleled computing model aiming to do analysis and mining on big data, it caused lots of attention from academic as well as industrial communities. Up to now, it has been widely explored in many fields, including data analysis and mining, machine learning, information retrieval, computer simulation, and so on.

In order to endow the typical analysis software with good expansibility, people make a lot of efforts to integrate MapReduce with the typical analysis software. For example, researchers at the IBM Corporation are committed to integrate R and Hadoop [34]. R is one kind of open source statistical analysis software. With this deep integration, R gets the ability to push the computation to the data and to do the analysis in parallel. Meanwhile, Hadoop obtains the powerful deep analysis capabilities through the deep integration with Wegener et al. have integrated Weka with MapReduce [35]. Similar to R, Weka is another kind of open source machine learning and data mining software. Some researchers launched the Apache Mahout project, which aims to develop large-scale machine learning and data mining algorithm libraries based on Hadoop platform [36]. Since these algorithms are open source, they provide application developers with rich data analysis capabilities.

7 Conclusion and future work

We close by sharing our opinions on what some of the important open questions are in this area as well as our thoughts on how the data management community might best seek out answers.

• **Data integration**

In order to utilize the information in big data, it is an important research topic about data integration in big data. Recently, researchers suggest to utilize users and/or crowds for improving the quality of integrated data. So crowd sourcing in data integration is a promising topic.

• **Data reduction**

On considering big data, making a profound transformation in computing, such as different sampling methods, aggregation (computing descriptive statistics), dimensionality reduction techniques, etc., is a feasible and effective approach before big data analysis and management. Here the main challenge is how to run the traditional machine learning and statistics algorithms on big data.

• **Data querying and indexing**

Big data querying and indexing need to modify the existing query optimization and indexing strategy in distributed system. Some traditional concerns to reduce I/O cost may not be useful in big data scenarios.

• **Data analysis and mining**

In order to make use of the big data to guide people's decision, "deep analysis" is needed. Due to its limited expansibility, typical relational database technologies meet unprecedented difficulties when they are used to address these challenges of deep analysis on big data. Therefore, deep analysis on big data is also an important challenge.

Finally, generally speaking, data is increasing with an exponential speed nowadays. However, corresponding information technology falls behind comparatively. Hence there is much remaining work for us to do about the data so that we could face the challenges brought by big data.

Acknowledgements This work was partially done when the authors worked in SA Center for Big Data Research in Renmin University of China. This Center is funded by a Chinese National "111" Project "Attracting International Talents in Data Engineering Research". This paper was also partially supported by Beijing Natural Science Foundation (Grant No. 4112030) and National Natural Science Foundation (Grant No. 61170011) and China National Social Security Foundation (Grant No: 12&ZD220).

References

1. Labrinidis A, Jagadish H. Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 2012, 5(12): 2032–2033
2. Chang C, Kayed M, Girgis M R, Shaalan K F, others . A survey of web information extraction systems. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1411–1428
3. Lu J, Lu Y, Cong G. Reverse spatial and textual K nearest neighbor search. In: Proceedings of the 2011 International Conference on Management of Data. 2011, 349–360
4. Simmhan Y L, Plale B, Gannon D. A survey of data provenance in e-science. ACM Sigmod Record, 2005, 34(3): 31–36
5. He B, Patel M, Zhang Z, Chang K C C. Accessing the deep web. Communications of the ACM, 2007, 50(5): 94–101
6. Lu J, Senellart P, Lin C, Du X, Wang S, Chen X. Optimal top-*k* generation of attribute combinations based on ranked lists. In: Proceedings of the 2012 International Conference on Management of Data. 2012, 409–420
7. Aggarwal C C, Wang H. Managing and mining graph data. Springer

Publishing Company, Incorporated, 2010
8. Oceanbase. http://oceanbase.taobao.org
9. Sikka V, Färber F, Lehner W, Cha S K, Peh T, Bornhövd C. Efficient transaction processing in SAP HANA database: the end of a column store myth. In: Proceedings of the 2012 International Conference on Management of Data. 2012, 731–742
10. Neo4j. http://neo4j.org
11. Malewicz G, Austern M H, Bik A J, Dehnert J C, Horn I, Leiser N, Czajkowski G. Pregel: a system for large-scale graph processing. In: Proceedings of the 2010 International Conference on Management of data. 2010, 135–146
12. Doan A, Naughton J F, Baid A, Chai X, Chen F, Chen T, Chu E, DeRose P, Gao B J, Gokhale C, Huang J, Shen W, Vuong B Q. The case for a structured approach to managing unstructured data. In: Proceedings of the 4th Biennial Conference on Innovative Data Systems Research. 2009
13. Jeffery S R, Franklin M J, Halevy A Y. Pay-as-you-go user feedback for dataspace systems. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008, 847–860
14. Chai X, Vuong B Q, Doan A, Naughton J F. Efficiently incorporating user feedback into information extraction and integration programs. In: Proceedings of the 35th SIGMOD International Conference on Management of Data. 2009, 87–100
15. Talukdar P P, Ives Z G, Pereira F. Automatically incorporating new sources in keyword search-based data integration. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. 2010, 387–398
16. Yakout M, Elmagarmid A K, Neville J, Ouzzani M, Ilyas I F. Guided data repair. Proceedings of the VLDB Endowment, 2011, 4(5): 279–289
17. Wang J, Kraska T, Franklin M J, Feng J. CrowdER: crowdsourcing entity resolution. Proceedings of the VLDB Endowment, 2012, 5(11): 1483–1494
18. Halevy A, Rajaraman A, Ordille J. Data integration: the teenage years. In: Proceedings of the 32nd International Conference on Very Large Data Bases. 2006, 9–16
19. Chen H, Ku W S, Wang H, Sun M T. Leveraging spatio-temporal redundancy for RFID data cleansing. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. 2010, 51–62
20. Mahmoud H A, Aboulnaga A. Schema clustering and retrieval for multi-domain pay-as-you-go data integration systems. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. 2010, 411–422
21. Morton K, Bunker R, Mackinlay J, Morton R, Stolte C. Dynamic workload driven data integration in tableau. In: Proceedings of the 2012 International Conference on Management of Data. 2012, 807–816
22. Agrawal P, Sarma A D, Ullman J, Widom J. Foundations of uncertain-data integration. Proceedings of the VLDB Endowment, 2010, 3(1-2): 1080–1090
23. Das Sarma A, Dong X, Halevy A. Bootstrapping pay-as-you-go data integration systems. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008, 861–874
24. Suchanek F M, Abiteboul S, Senellart P. PARIS: probabilistic alignment of relations, instances, and schema. Proceedings of the VLDB Endowment, 2011, 5(3): 157–168
25. Huang J, Chen T, Doan A, Naughton J F. On the provenance of non-answers to queries over extracted data. Proceedings of the VLDB Endowment, 2008, 1(1): 736–747
26. Ioannou E, Nejdl W, Niederée C, Velegrakis Y. On-the-fly entity-aware query processing in the presence of linkage. Proceedings of the VLDB Endowment, 2010, 3(1-2): 429–438
27. Chen Z, Kalashnikov D V, Mehrotra S. Exploiting context analysis for combining multiple entity resolution systems. In: Proceedings of the 35th SIGMOD International Conference on Management of Data. 2009, 207–218
28. Whang S E, Menestrina D, Koutrika G, Theobald M, Garcia-Molina H. Entity resolution with iterative blocking. In: Proceedings of the 35th SIGMOD International Conference on Management of Data. 2009, 219–232
29. Fan W, Jia X, Li J, Ma S. Reasoning about record matching rules. Proceedings of the VLDB Endowment, 2009, 2(1): 407–418
30. Rimal B P, Choi E, Lumb I. A taxonomy and survey of cloud computing systems. In: Proceedings of the 5th International Joint Conference on INC, IMS and IDC. 2009, 44–51
31. Aguilera M K, Golab W, Shah M A. A practical scalable distributed b-tree. Proceedings of the VLDB Endowment, 2008, 1(1): 598–609
32. Jagadish H V, Ooi B C, Vu Q H. BATON: a balanced tree structure for peer-to-peer networks. In: Proceedings of the 31st International Conference on Very Large Data Bases. 2005, 661–672
33. Wu S, Wu K L. An indexing framework for efficient retrieval on the cloud. In: Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 2009, 1–8
34. Das S, Sismanis Y, Beyer K S, Gemulla R, Haas P J, McPherson J. Ricardo: integrating R and Hadoop. In: Proceedings of the 2010 International Conference on Management of Data. 2010, 987–998
35. Wegener D, Mock M, Adranale D, Wrobel S. Toolkit-based high-performance data mining of large data on MapReduce clusters. In: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops. 2009, 296–301
36. Chu C T, Kim S K, Lin Y A, Yu Y Y, Bradski G, Ng A Y, Olukotun K. Map-reduce for machine learning on multicore. In: Proceedings of the 2006 Conference Advances in Neural Information Processing Systems. 2007, 281–288

Jinchuan CHEN is currently a lecturer of the Key Laboratory of Data Engineering and Knowledge Engineering, Ministry of Education (Renmin University of China). He received his BS from Department of Computer Science and Technology of Beijing Normal University in 2001, and his MS from Institute of Software, Chinese Academy of Sciences in 2004. He then obtained his PhD from COMP (HKPolyU) in 2009. His research interests mainly focus on uncertain data management and unstructured data management.

Yueguo CHEN received the BS and MS from Tsinghua University, Beijing, in 2001 and 2004. He earned his PhD in Computer Science from National University of Singapore in 2009. He is currently an associate professor of Renmin University of China. His recent research interests include interactive analysis of big data, large-scale RDF knowledge base management.

Xiaoyong DU received his BS of Computational Mathematics from Hangzhou University in 1983 and ME of Computer Science from Renmin University of China in 1988. He obtained his PhD of Computer Science from Nagoya Institute of Technology, Japan in 1997. He is currently a professor and Dean of School of Information in Renmin University of China. His current research interests include high-performance database systems, intelligent information retrieval, semantic web and knowledge engineering, and digital library technology.

Cuiping LI received BE from Xi'an Jiao Tong University, China, in 1994 and ME from Xi'an Jiao Tong University, China, in 1997. In 2003, she received her PhD from the Institute of Computing Technology, Chinese Academy of Sciences. She is currently an associate professor of Renmin University of China. Her current research interests include database systems, data warehouse, and data mining.

Jiaheng LU received MS in Computer Science from Shanghai Jiao Tong University in 2001 and PhD in Computer Science at National University of Singapore (NUS). He did his Postdoc research with Prof. Chen Li in the Department of Computer Science, University of California, Irvine, during 2006 and 2008. He is currently a professor of Renmin University of China. His current research interests are database and information systems, including XML query processing, data mining, XML keyword suggestion, approximate string matching, cloud data management.

Suyun ZHAO received BS and MS in School of Mathematics and Computer Science, Hebei University, Baoding, China in 2001 and 2004, respectively. She received her PhD in the Department of Computing, the Hong Kong Polytechnic University. Now she is working with Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China). Her research interests are in the areas of machine learning, pattern recognition, uncertain information processing, especially fuzzy sets and rough sets.

Xuan ZHOU obtained his PhD from the National University of Singapore in 2005. He was a researcher at the L3S Research Centre, Germany, from 2005 to 2008, and a researcher at CSIRO, Australia, from 2008 to 2010. Since 2010, he has been an associate professor at the Renmin University of China. His search interests include database system and information management. He has contributed to a number of research and industrial projects in European Union, Australia, and China.