Framework of Integrated Big Data: A Review

Zhikui Chen[1], Fangming Zhong[2], Xu Yuan[3]
School of Software Technology
Dalian University of Technology
Dalian, China
e-mail: {[1]zkchen, [3]avid}@dlut.edu.cn,
e-mail: [2]fmzhong@mail.dlut.edu.cn

Yueming Hu
College of Natural Resources and Environment
South China Agricultural University
Guangzhou, China
e-mail: ymhu163@163.com

Abstract—**Currently, how to deeply distill potential attributes of big data has become a great challenge for structured, semi-structured and unstructured data (SSU data) with a unified model. Structured data refers to any data that resides in a fixed field within a record or file including data contained in relational databases and spreadsheets. Unstructured data refers to data from text, pictures, audio, video, and other sources that do not fit into a relational database. Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze, such as XML, and HTML documents. In this paper, we present a literature survey and a framework, namely integrated big data (IBD), which aims at exploring the approaches for constructing a universal IBD model, including representation, storage and management, computation, and visual analysis. Firstly, we present a systematic framework to decompose big data analytics into four modules. Next, we present a detailed survey of numerous approaches for these four modules. The main contributions of this paper are summarized in two dimensions. First, we propose a novel integrated big data framework for unified big data representation, storage, computation, and visual analysis. Second, we present the possible future methods in realizing the framework by reviewing methods. Through this paper, we would like to point out a promising research direction in unified investigation and application of big data.**

Keywords-big data; data analytics; data storage; unified representation

I. Introduction

New challenges and opportunities are faced for big data analysis as a result of the characteristics which refer to the volume, variety, velocity, and value [1-3]. Volume means the enormous scale of big data which outstrips the traditional store and analysis techniques. Hence, big data processing needs the large storage space and computing power. Variety indicates big data comes from a great many of realms with various types, which is the major obstacle for data representation, so as to the visualization. Velocity denotes the rapid generation speed of big data, which implies the high requirement on algorithms for big data real-time processing. Value says the huge usefulness of big data, while the low density increases the complexity of mining meaningful value hidden in big data. The big impact and big challenges of big data are discussed in a special issue of Nature on Big Data [4]. Afterwards, a paper entitled Detecting Novel Associations in Large Data Sets was published in Science [5], which revealed the inherent relationship between complicated colossal data sets. Obviously, big data has drastically attracted much attention from researchers.

A large number of studies have been done to address the above big data challenges. Researchers have made a lot of efforts to improve the efficiency of representation, storage and management, data analyzing and mining, and visualization of big data [6], [7]. Big data representation has been discussed in literatures [8], [9]. Numerous models about big data management and storage are presented in literatures [10], [11]. A great many studies focus on the big data analysis [12], [13]. Big data visualization and visual analysis are investigated in literatures [14-16].

However, the researches described above are in isolation, which failed to consider the comprehensive aspects of big data. In order to capture the nature of big data and deeply distill the hidden value, it is critical for the paradigm shift that we should view all phases as a holistic entirety as well as the SSU data. In this paper, we focus on the integrated big data framework and its four components i.e., integrated data representation, data storage and management, data analyzing and mining, and data visualization of big data. In order to quickly and effectively distill knowledge and insights from big data, we examine the representative methodologies and models of these four aspects, respectively. A detailed literature survey and system tutorial for big data analytics platforms are provided. The authors present a systematic framework to decompose big data systems into four sequential modules, including data generation, data acquisition, data storage, and data analytics, which form a big data value chain.

The remainder of this paper is organized as follows. The components and basic architecture of IBD framework are discussed in detail in Section II. Section III reviews the existing methods of integrated representation. The models for unified storage and management of big data are examined in Section IV. In Section V, we survey the deep computation of big data. The techniques for integrated visual analysis are summarized in Section VI. Finally, Section VII concludes the paper.

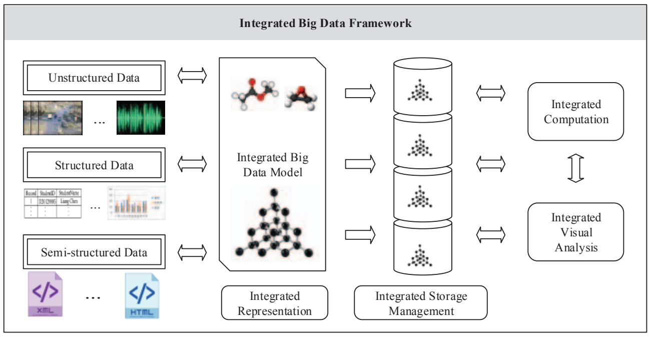
Integrated Big Data Framework
Unstructured Data
...
Structured Data
...
Semi-structured Data
...
Integrated Big Data Model
Integrated Representation
Integrated Storage Management
Integrated Computation
Integrated Visual Analysis

Figure 1. Framework for integrated big data

II. INTEGRATED BIG DATA FRAMEWORK

The integrated big data model includes integrated representation, integrated storage and management, integrated computation, integrated visual analysis.

- Integrated representation unifies the SSU data, and maps big data onto an identical space.
- Integrated storage and management views the SSU data as a “same type”, storing big data, with a unified storage architecture, which aims at achieving effective management, including quick insert, delete, update, query, and joint operation on big data.
- Integrated computation deals with the SSU data at the same time based on integrated representation and storage through deep computing methods, which can learn the intrinsic characteristics and distill buried information in big data. The integrated computation concerns feature learning, association analysis, clustering, classification, and prediction.
- Integrated visual analysis shows SSU data for unified visualization, which can give a multivariate and associated presentation for the characteristics and association in big data.

Fig.1 demonstrates the designed IBD framework. Firstly, a unified model can be used to describe the SSU data. Then, the unified represented heterogeneous data are stored by a compatible and consistency storage model. Subsequently, big data computing, which is based on the former two process heterogeneous data on a same platform. Finally, visual analysis can display the inherent information of big data, in addition the analyzing and mining results also provide feedback for integrated computation.

III. INTEGRATED REPRESENTATION

There are a number of tools for constructing the integrated representation model. Graph, tensor, and other tools, which are used for knowledge representation, with powerful ability to represent multi-source, heterogeneous and multidimensional data, are practicable methods for integrated representation. The followings are two methods based on tensor and graph.

A. *Integrated Representation Based on Tensor*

Science the variety of big data, tensor is applied to the representation, and analysis of big data. In [17], tensor network is used for big data to analyze and extract the core data sets. In literature [8], a novel tensor model is proposed for big data representation; moreover an incremental high order singular value decomposition method (IHOSVD) is presented for the dimensionality reduction of tensor big data.

In the tensor model, data characteristics are represented as tensor orders. Firstly, heterogeneous data is described by various tensors, then these tensors can be joined to an identical tensor through the tensor extension operator. Brief introduction of tensor-based approach for big data representation is as follows.

The SSU data can first be tensorized as low order sub-tensors and then extended to a high order unified tensor. The extension operator merges the identical orders while keeping the diverse orders. After the extension, elements of the identical order are accumulated together. The tensorized unstructured data, structured data, and semi-structured data consist the IBD representation model based on tensor, which integrates the sub-tensors to a unified tensor.

After a large amount of tensor extension operation, the unified high order tensor would contain much redundancy, uncertainty, inconsistency and incompleteness, thus it is essential to extract valuable core data through dimensionality reduction. In tensorized big data, the association between tensor orders, namely data features, is worthy of further exploration and excavation.

B. *Integrated Representation Based on Graph*

A graph is a collection of vertices, edges, and their properties [18]. It is advantageous for the associated data description such as web data, social networks, because of the connectivity and the correlation within them. Due to the powerful representation capacity of graph, Aliaksei Sandryhaila et al. [19] apply it to the big data analysis in signal processing. Zhou et al. [9] propose a graph-based text representation model, which could represent texts with weighted directed graphs. Maugey et al. [20] use graph-based representation to describe the geometric and texture information of multi-view image in image processing, which performed better in image compression and reconstruction compared to the depth-based method. In the geographical space modeling, Domingo et al. [21] utilize graph to indicate the geographic information. The land space data are denoted by graph model, in which way, the relationships between constructions of reality can be revealed. In that way, more information and supports for the analysis of land space structure could be provided. It can be seen from that graph-based representation method is widely used for unstructured data such as text and image.

TABLE I. COMPARISON OF DATA STORAGE MODELS

Data Model	Data Type Support			Scalability		MapReduce Support	Products
	Structured	Unstructured	Semi-Structured	Horizontal	Vertical		
Key-Value	√	√	√	√	√	√	Voldemort, Redis
Column-Oriented	√	√	√	√	√	√	Hypertable, HBase
Document Databases	√	√	√	√	√	√	MongoDB, CouchDB
Graph Databases	√	√	√	×	√	×	Neo4j, Infinite Graph

√-support natively, ×-not support

Property Graph Model has attracted much attention on heterogeneous representation. In the property graph model, the vertices are used to represent knowledge, the relationships are denoted by the edges with start and end nodes, and the properties are contained by vertices and relationships. Property graph model is suitable for the representation of relational data and semi-structured data such as XML [18]. It can get a good performance if the entities are associated.

Moreover, hypergraph would be a powerful representation tool for big data [18]. Although graph can be used to represent some heterogeneous data, further research is needed on the time series and dynamic data such as video.

The tensor-based and graph-based representation models are only two of the many solutions. Currently, most of the researches mainly focus on big data representation, which is the basic and essential part of IBD framework. Obviously, that's not enough for the whole IBD blueprint.

IV. INTEGRATED STORAGE MODEL

A number of works have been presented which are based on the relational model, and they aim to increase the scalability of two-dimensional table. The sparse wide table is considered for heterogeneous data storage. For instance, Zheng et al. [22] propose an open scalable relational data model which is oriented to big data. Yang et al. [23] summarize the wide table models for massive and heterogeneous web data. Four logical representations are discussed namely, n-ary horizontal representation, 2-ary binary representation, 3-ary vertical representation, and the hybrid representation. The most successful application of wide table model might be Google's BigTable, but it does not provide an SQL style query language. Another implementation of hybrid wide table representation is HBase in Hadoop. In contrast to BigTable, HBase provides an SQL style query language. In terms of semi-structured data, Chen et al. [24] propose a novel mapping of XML into a wide sparse table that gives an idea for transforming semi-structured data to structured data and the implementation of unified storage of heterogeneous data.

Compared to SQL of structured and relational data, NoSQL technique is advantageous for unstructured and semi-structured data, which would be more appropriate for the integrated storage and management of big data. Several representative NoSQL models are reviewed, including Key-Value, Column family, Document Database, and Graph Database [23].

A. Key-Value Storage Model

The key-value model stores data as key-value pairs or maps, in which a data object is represented as a value, while key is the only keyword for mapping data to value by the hash function. Typical characteristics of the key-value model include real-time processing of big data, the horizontal scalability across nodes in a cluster or a data center, the reliability and the high availability. The key-value model can quickly query, but the defect is querying a continuous block of key. In order to address this limitation, an ordered key-value model and some improved models such as the key-column and key-document are proposed [25].

B. Column Storage Model

Compared to the row database, column database is column-centric. Data are organized as a one dimensional array in physical memory, therefore the same column data could be saved continuously, which is useful for slight data compression. As a result, column store could process query quickly. In addition, the scalability of column databases is grateful for distributed extension [26].

C. Document Database Model

Data is stored in a document in the formats of XML, YAML, and JSON etc. in a document database. Documents are organized into different collections for data management. Each document is an individual entity. Every field in a document is a unique key, and the content indicates the value. MongoDB, CouchDB etc. are the representative document databases. The document database is mostly used for the semi-structured data storage and management. The advantages of this model are the scalability and ability of evolution, while the defect is the low performance of querying [27].

D. Graph Database Model

Graph database model manages data in a network structure, in which vertices describe the data and the edges represent for relationships between vertices. Compared with Entity-Relationship Model, the vertices correspond to the entities, the properties to attributes, and the edges to relationships between entities. Graph database can quickly process the JOIN operation in massive data, without any predefined models. The features are the strong dynamic adaptability and the combination of graphic algorithms such as entity relationship query and shortest path. The graph database model can store and manage the semi-structured data.

It can be seen from Table 1 that these models are suitable for the storage and management of the SSU data. Many products have been presented based on the above models, which attempt to store and manage data as many types as possible, such as HBase, CouchDB, etc. But they mainly consider that filling different types data into a single database. However, in IBD framework, data storage and management should be based on the IBD representation. Due to the immature model for heterogeneous data, integrated big data storage and management would be one of the key points for efficient management in big data.

V. Integrated Computation

Integrated computation on big data seeks for the deep computing methodologies, which could unify the heterogeneous data. Hinton proposes an unsupervised greedy deep learning algorithm based on Deep Belief Networks (DBNs), which makes it available for the unified feature learning from multi-modal data [28], [29]. In additional, Convolutional Neural Networks (CNNs) and stacked auto-encoders (SAEs) are widely used for big data feature learning, analysis, and prediction. The deep learning algorithms provide basically support for IBD deep computation. Based on integrated representation, deep computation model could uniformly process multi-source heterogeneous data, greatly improve the efficiency and accuracy of data analyzing, and deeply mine the intrinsic characteristics. Here, deep computation algorithm of integrated computation is introduced.

A. Deep Computation

Deep learning algorithms such as deep belief networks, convolutional neural networks, and stacked auto-encoders etc. can be applied to integrated big data computation [30]. The former one is constituted by restricted Boltzmann machine with explicit and implicit layers. CNNs restrict the network architecture with local connectivity and weight sharing. It has been widely used to image and document recognition [31]. The SAE model is constructed by stacking auto-encoders as building blocks which is effectively utilized for multi-modal feature learning [32]. How to apply these deep learning methodologies on IBD with the new designed integrated deep computation algorithms is one of the current critical topics. To a certain extent, tensor-based approach can agree with the requirement of integrated representation [8]. Consequently, it is worth exploring tensor-based deep learning algorithms for unified big data processing.

For the sake of learning from tensor-based integrated represented big data, it needs to extend the inputs, outputs, and progress of Deep Belief Networks or Convolutional Neural Networks to high order tensor in data format, by which the SSU data could be learned simultaneously.

However, the existing works are mainly about the multi-modal data computation. The IBD framework aims at catching the hidden associations through the integrated representation model. Hence, the IBD computation is a synthesis of a variety of big data computing methodologies.

VI. Integrated Visual Analysis

The heterogeneity and dimensionality of big data characteristics bring new opportunities as well as technical challenges for the visualization analysis research [14]. The implementation of integrated visual analysis could be helpful for extracting more complete, intuitive, and hidden information from integrated SSU data [15].

In a clinical environment, the integrated visualization of multi-modal data could provide more comprehensive, intuitive, and targeted data presentation to help clinicians make better diagnosis and treatment plans, in which the data include helical computerized tomography, magnetic resonance imaging and color Doppler ultrasound [33]. J. An et al. [6] propose integrated visualization of multi-modal electronic health record (EHR) data, in which a unified structure was constructed to represent multi-modal EHR data, including structured relational data, unstructured text and image, etc. A top-down method is used to organize data into three categories: numeric data, texts, and binary waveforms and images. After the categorization, they visualize the data by the combination of the data table, trend chart, timeline, thumbnails and keywords. Visual analysis helps the complicate and monotonic data open its mouth. More buried features and trends in data are presented by the visualized tools such as image and visual languages (e.g., CoDe [34]), and some other approaches such as the geometry graph, pixels-based, and graph-based methods [35].

Integrated visualization of multi-source heterogeneous data is still a challenge due to the lack of completeness, consistency, and accuracy of big data. The recent works have mainly focused on the presentation of a wide variety of data types, but they neglected the visualization of computing process and the feedback to computation. The IBD framework would pay much more attention to the visualization and analysis of big data computing process. Integrated visual analysis aims at presenting more intuitive and valuable information to users, and giving a feedback to the big data analyzing. The integrated visual analysis would be a powerful tool for valuable information showing and big data analyzing with further research.

VII. Conclusion

In this paper, we have presented the framework of integrated big data, which aims at deeply mining the intrinsic characteristics and features in big data, by reviewing and summarizing the theories and methodologies. The integrated big data framework consists of for modules: integrated representation, storage and management, computation and visual analysis. Integrated representation describes the structured, semi-structured, and unstructured data with an individual mode. Then through the unified storage and management, system could improve the performance because of the reduction of data accessing and the decrease in communication cost. The integrated computation desires to drill out the mysterious and complex correlations of heterogeneous data effectively and accurately. Ultimately, more complete, intuitive, valuable information could be presented for big data by integrated visual analysis. A comprehensive literature survey on the approaches of the aforementioned four modules has been provided. This leads to conclusions with respect to promising research directions, for instance, to pursue new solutions for big data feature extraction as well as techniques that support deep computation.

Acknowledgment

This paper was supported in part by the State Key Program of National Natural Science of China under Grant U1301253, in part by the Science and Technology Planning Key Project of Guangdong Province under Grant 2015B010110006.

References

[1] A. H. B. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, 2011. [Online]. Available: http://scholar.google.com/scholar.bib?q=info:kkCtazs1Q6wJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,47&ct=citation&cd=0.

[2] V. Marx, “Biology: The big challenges of big data,” Nature, vol. 498, no. 7453, pp. 255–260, 2013.

[3] S. Sagiroglu and D. Sinanc, “Big data: A review,” 2013 Int. Conf. Collab. Technol. Syst., pp. 42–47, May 2013.

[4] “Specials : Nature,” Nature, 2008. [Online]. Available: http://www.nature.com/news/specials/scipublishing/index.html.

[5] D. N. Reshef, Y. a. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, “Detecting Novel Associations in Large Data Sets,” Science (80-.)., vol. 334, no. 2011, pp. 1518–1524, 2011.

[6] J. An, X. Lu, and H. Duan, “Integrated Visualization of Multi-Modal Electronic Health Record Data,” 2008 2nd Int. Conf. Bioinforma. Biomed. Eng., pp. 640–643, May 2008.

[7] X. Wu, X. Zhu, G. Wu, and W. Ding, “Data mining with big data,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, 2014.

[8] L. Kuang, F. E. I. Hao, L. T. Yang, M. A. N. Lin, C. Luo, and G. Min, “A Tensor-Based Approach for Big Data Representation and Dimensionality,” IEEE Trans. Emerg. Top. Comput., vol. 2, no. 3, pp. 280–291, 2014.

[9] F. Zhou and F. Zhang, “Graph-based Text Representation Model and its Realization,” 2010.

[10] V. N. Gudivada, D. Rao, and V. V. Raghavan, “NoSQL Systems for Big Data Management,” 2014 IEEE World Congr. Serv., pp. 190–197, Jun. 2014.

[11] D.-R. Shen, G. Yu, X.-T. Wang, T.-Z. Nie, and Y. Kou, “Survey on NoSQL for Management of Big Data,” J. Softw., vol. 24, no. 8, pp. 1786–1803, Jan. 2014.

[12] F. Frankel and R. Reid, “Big data: Distilling meaning from data,” Nature, vol. 455, no. 7209, pp. 30–30, 2008.

[13] X. Han, J. Li, D. Yang, and J. Wang, “Efficient skyline computation on big data,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 11, pp. 2521–2535, 2013.

[14] J. Kehrer and H. Hauser, “Visualization and visual analysis of multifaceted scientific data: A survey,” IEEE Trans. Vis. Comput. Graph., vol. 19, no. 3, pp. 495–513, 2013.

[15] L. Ren, Y. Du, S. Ma, X. Zhang, and G. Dai, “Visual Analytics Towards Big Data,” J. Softw., vol. 25, no. 9, pp. 1909–1936, 2014.

[16] S. J. Rysavy, D. Bromley, and V. Daggett, “DIVE : A Graph-Based Visual- Analytics Framework for Big Data,” IEEE Comput. Graph. Appl., vol. 34, no. 2, pp. 50–61, 2014.

[17] A. Cichocki, “Era of Big Data Processing: A New Approach via Tensor Networks and Tensor Decompositions,” arXiv Prepr. arXiv1403.2048, 2014., pp. 1–30, Mar. 2014.

[18] I. Robinson, J. Webber, and E. Eifrem, Graph Databases, First Edit. 2013.

[19] A. Sandryhaila and J. M.F. Moura, “Big Data Analysis with Signal Processing on Graphs,” IEEE Signal Process. Mag., vol. 31, no. 5, pp. 80–90, 2014.

[20] T. Maugey, A. Ortega, and P. Frossard, “Graph-based vs depth-based data representation for multiview images,” 2013 Asilomar Conf. Signals, Syst. Comput., pp. 704–708, Nov. 2013.

[21] M. Domingo, R. Thibaud, and C. Claramunt, “A graph-based model for the representation of land spaces,” Proc. 21st ACM SIGSPATIAL Int. Conf. Adv. Geogr., pp. 506–509, 2013.

[22] Z. Zheng, Z. Du, L. Li, and Y. Guo, “BigData Oriented Open Scalable Relational Data Model,” 2014 IEEE Int. Congr. Big Data, pp. 398–405, Jun. 2014.

[23] B. Yang, W. Qian, and A. Zhou, “Using Wide Table to manage web data: a survey,” Front. Comput. Sci. China, vol. 2, no. 3, pp. 211–223, Aug. 2008.

[24] L. J. Chen, P. a. Bernstein, P. Carlin, D. Filipovic, M. Rys, N. Shamgunov, J. F. Terwilliger, M. Todic, S. Tomasevic, and D. Tomic, “Mapping XML to a Wide Sparse Table,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 6, pp. 1400–1414, Jun. 2014.

[25] Y. Wang, R. Xiong, L. Shen, K. Sun, J. Zhang, and L. Qi, “Towards learning from demonstration system for parts assembly: A graph based representation for knowledge,” 4th Annu. IEEE Int. Conf. Cyber Technol. Autom. Control Intell., pp. 174–179, Jun. 2014.

[26] H. Liu, Z. Liu, T. Yuan, and Y. Yao, “Adaptively Incremental Dictionary Compression Method for Column-Oriented Database,” pp. 628–632, 2014.

[27] M. S. Kim, K. Y. Whang, and Y. S. Moon, “Horizontal reduction: Instance-level dimensionality reduction for similarity search in large document databases,” Proc. - Int. Conf. Data Eng., pp. 1061–1072, 2012.

[28] G. E. Hinton, “A Fast Learning Algorithm for Deep Belief Nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.

[29] R. Sarikaya, G. E. Hinton, and B. Ramabhadran, “Deep belief nets for natural language call-routing,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2011, pp. 5680–5683.

[30] J. Gu and Z. Lin, “Implementation and Evaluation of Deep Neural Networks (DNN) on Mainstream Heterogeneous Systems,” Proc. 5th Asia-Pacific Work. Syst., 2014.

[31] L. Xu, J. S. Ren, C. Liu, and J. Jia, “Deep Convolutional Neural Network for Image Deconvolution,” Adv. Neural Inf. Process. Syst. 27 (NIPS 2014), no. 413113, pp. 1–9, 2014.

[32] W. Wang, C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, “Effective Multi-Modal Retrieval based on Stacked,” Proceeding VLDB Endow., vol. 7, no. 8, pp. 649–660, 2014.

[33] Z. Fan, J. Liu, Z. Yin, and H. Duan, “An optimized framework for integrated visualization of distributed medical images,” 2012 5th Int. Conf. Biomed. Eng. Informatics, pp. 1049–1053, Oct. 2012.

[34] M. Risi, M. I. Sessa, M. Tucci, and G. Tortora, “CoDe modeling of graph composition for data warehouse report visualization,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 3, pp. 563–576, 2014.

[35] D. a Keim and H.-P. Kriegel, “Visualization techniques for mining large databases: a comparison,” IEEE Trans. Knowl. Data Eng., vol. 8, no. 6, pp. 923–938, 1996.