Big Data Integration (ICDE 2013 Seminar)

Xin Luna Dong, Divesh Srivastava

AT&T Labs–Research, Florham Park, NJ, USA
{lunadong,divesh}@research.att.com

***Abstract*—The Big Data era is upon us: data is being generated, collected and analyzed at an unprecedented scale, and data-driven decision making is sweeping through all aspects of society. Since the value of data explodes when it can be linked and fused with other data, addressing the big data integration (BDI) challenge is critical to realizing the promise of Big Data.**

BDI differs from traditional data integration in many dimensions: (i) the number of data sources, even for a single domain, has grown to be in the tens of thousands, (ii) many of the data sources are very dynamic, as a huge amount of newly collected data are continuously made available, (iii) the data sources are extremely heterogeneous in their structure, with considerable variety even for substantially similar entities, and (iv) the data sources are of widely differing qualities, with significant differences in the coverage, accuracy and timeliness of data provided. This seminar explores the progress that has been made by the data integration community on the topics of schema mapping, record linkage and data fusion in addressing these novel challenges faced by big data integration, and identifies a range of open problems for the community.

I. Introduction

The Big Data era is the inevitable consequence of our ability to generate, collect and store digital data at an unprecedented scale, and our concomitant desire to analyze and extract value from this data in making data-driven decisions to alter all aspects of society. Big Data comes with a lot of promises – as a Dilbert cartoon would have it, "It comes from everywhere. It knows all."[1]

This data is being collected today in a large variety of domains. Examples include Web text and documents, Web logs, large-scale e-commerce, social networks, sensor networks, astronomy, genomics, medical records, surveillance, etc.[2] Since the value of data explodes when it can be linked and fused with other data to create a unified representation, big data integration (BDI) is critical to realizing the promise of Big Data. For example, to understand habitat utilization and animal behavior in reaction to external forces such as weather, marine animal researchers need to combine animal tracking data with bathymetric, meteorological, sea surface temperature and animal habitat data.[3]

BDI differs from traditional data integration (which includes virtual integration and materialized warehousing) in many dimensions.

- *Volume:* Not only can each data source contain a huge volume of data, but also the number of data sources, even for a single domain, has grown to be in the tens of thousands.
  For example, in the recent work by Dalvi et al. [6], where they analyze the nature and distribution of structured data on the Web, they studied many domains (e.g., restaurants, automotive, libraries, schools, hotels) and showed that each domain has tens of thousands of sources on the Web. This is much higher than the number of data sources considered in traditional data integration.
- *Velocity:* As a direct consequence of the rate at which data is being collected and continuously made available, many of the data sources are very dynamic.
  For example, there are many data sources that provide near real time, continuously changing information about the stock market, including bid and ask prices, volume of shares traded, etc. Providing an integrated view of stock market data across all these data sources is beyond the ability of traditional methods for data integration.
- *Variety:* Data sources (even in the same domain) are extremely heterogeneous both at the schema level regarding how they structure their data and at the instance level regarding how they describe the same real-world entity, exhibiting considerable variety even for substantially similar entities.
  For schema-level variety, in the recent work by Li et al. [32], a study of 55 sources in the stock market domain identified 153 *global* attributes that are manually matched from 333 *local* attributes. The number of providers for each global attribute observes Zipf's law, with 13.7% of the attributes provided by at least one third of the sources and over 86% of attributes provided by fewer than 25% of the sources. For instance-level variety, Guo et al. [20] showed that for business listings, the number of distinct business names is typically twice as many as the number of distinct businesses in a zipcode. Similarly, an early study [34] on Google Base showed that just for vehicle color, there are over 250 different colors provided as values, including very specific ones such as "polished pewter" and "light almond pearl metallic."
- *Veracity:* Data sources (even in the same domain) are of widely differing qualities, with significant differences in the coverage, accuracy and timeliness of data provided.
  For example, the work by Dalvi et al. [6] showed that with strong head aggregators such as yelp.com, collecting homepage URLs for 70% restaurants that are mentioned by some websites required only 10 sources; however, collecting URLs for 90% restaurants required 1000 sources,

[1] http://dilbert.com/strips/comic/2012-07-29/
[2] http://en.wikipedia.org/wiki/Big_data
[3] http://en.wikipedia.org/wiki/Data_fusion

and collecting URLs for 95% restaurants required 5000 sources. Similarly, the work by Li et al. [32] showed that even in the stock market domain, inconsistent values were provided by different sources for over 80% of the data items whose values should be fairly stable (such as daily closing price). This is consistent with the belief that "1 in 3 business leaders do not trust the information they use to make decisions." [4]

This seminar explores the progress that has been made by the data integration community on the topics of *schema mapping*, *record linkage* and *data fusion* (discussed in more detail below) in addressing these novel challenges faced by BDI. We do this using illustrative examples that would be of interest to data management researchers and practitioners. We also identify a range of open problems for the community in integrating a galaxy of data sources.

II. Target Audience

The target audience for this seminar is anyone with an interest in understanding data integration in the Big Data environment. In particular, this includes the attendees at database conferences like ICDE. The assumed level of mathematical sophistication will be that of the typical conference attendees.

III. Seminar Outline

The importance of big data integration has led to a substantial amount of research over the past few years on the topics of schema mapping, record linkage and data fusion to deal with the novel challenges faced by big data integration. Table I shows a summary of these techniques. Our seminar is example driven, and organized as follows.

A. BDI: Motivation (10 minutes)

The seminar will start with a variety of real-world examples illustrating the importance of big data integration, building on recent work by Dalvi et al. [6] and Li et al. [32].

B. BDI: Schema Mapping (25 minutes)

Schema mapping in a data integration system refers to (i) creating a mediated (global) schema, and (ii) identifying the mappings between the mediated (global) schema and the local schemas of the data sources to determine which (sets of) attributes contain the same information [41], [3], [1].

Early efforts in integrating a large *number* of sources involved integrating data from the Deep Web. Two types of solutions were proposed. The first is to build mappings between Web forms (interfaces to query the Deep Web) as a means to answer a Web query over all Deep Web sources [5]. The second is to crawl and index the Deep Web data [34], [35]. More recent efforts include extracting and integrating structured data from Web tables [4], [40] and Web lists [21], [15].

The number of sources also increases the *variety* of the data. Traditional data integration systems require a significant schema mapping effort before the system can be used, so is obviously infeasible when the heterogeneity is at the BDI scale. The basic idea of *dataspace systems* is to provide best-effort services such as simple keyword search over the available data sources at the beginning, and gradually evolve schema mappings and improve search quality over time [16], [22], [7], [8], [42], [26], [17], [25], [43].

A related notion becoming popular in the Hadoop community is "schema on read" which, in contrast to the traditional approach of defining the schema before loading data (i.e., schema on write), gives one the freedom to define the schema after the data has been stored.[5]

C. BDI: Record Linkage (25 minutes)

Record linkage refers to the task of identifying records that refer to the same logical entity across different data sources, especially when they may or may not share a common identifier across the data sources [24], [19], [30], [46], [14].

Record linkage has traditionally focused on linking a static set of structured records that have the same schema. In BDI, (i) data sources tend to be heterogeneous in their structure and many sources (e.g., tweets, blog posts) provide unstructured text data, and (ii) data sources are dynamic and continuously evolving. These characteristics make record linkage particularly challenging in BDI.

When there are a large number of sources and a large volume of data, traditional record linkage approaches become inefficient and ineffective in practice. To address the *volume* dimension, new techniques have been proposed to enable parallel record linkage using MapReduce. These include techniques for adaptive blocking [12], [44], [37] and techniques that balance load among different nodes [29], [28].

When the data sources are dynamic and continuously evolving, applying record linkage from scratch for each update becomes unaffordable. To address the *velocity* aspect, incremental clustering techniques have been proposed to address this problem [45], [36].

Record linkage between structured and unstructured data sources arises, e.g., when linking shopping transactions of people with tweets or blog posts about their shopping experience. To address the *variety* aspect, techniques have been proposed that tag and match free text to structured data [27].

Finally, in the BDI environment, information is typically more imprecise and noisy. To address this *veracity* aspect, a variety of clustering and linkage techniques that are robust to noise or evolving values have been proposed [31], [20], [23].

D. BDI: Data Fusion (20 minutes)

Data fusion refers to resolving conflicts from different sources and finding the truth that reflects the real world [2], [11]. Unlike schema mapping and record linkage, data fusion is a new field that has emerged only recently. Its motivation is exactly the *veracity* of data: the Web has made it easy to publish and spread false information across multiple sources

[4] http://www-01.ibm.com/software/data/bigdata/

[5] http://howsoftwareisbuilt.com/2010/01/06/interview-with-amr-awadallah-cloudera-cto/

TABLE I
SUMMARY OF STATE-OF-THE-ART DATA INTEGRATION TECHNIQUES MEETING CHALLENGES OF BIG DATA.

	Schema mapping	Record linkage	Data fusion
Volume	Integrating Deep Web, Web tables/lists	Adaptive blocking, MapReduce-based linkage	Online fusion
Velocity		Incremental linkage	Fusion in a dynamic world
Variety	Dataspace systems	Linking text to structured data	Combining fusion with linkage
Veracity		Value-variety tolerant linkage	Truth discovery

and so it is critical to separate the wheat from the chaff for presenting high quality data.

To address such *veracity* related challenges, techniques have been proposed to find the single truth from conflicting values [49], [32], [48], [39], [38], [18], [9], [47] and to find multiple truths [50]. Such techniques have also been extended to handle the *volume* of data (online data fusion [33]), *velocity* of data (truth discovery for dynamic data [10]), and *variety* of data (combining record linkage and data fusion [20]).

E. BDI: Open Problems (10 minutes)

Finally, we discuss cutting-edge open problems for big data integration, such as integrating crowdsourcing data, integrating data from data markets, providing an exploration tool for data sources, and so on.

IV. BIOGRAPHIES

A. Xin Luna Dong

Xin Luna Dong is a researcher at AT&T Labs-Research. She received her Ph.D. from University of Washington in 2007, received a Master's Degree from Peking University in China in 2001, and received a Bachelor's Degree from Nankai University in China in 1998. Her research interests include databases, information retrieval and machine learning, with an emphasis on data integration, data cleaning, personal information management, and Web search. She has led the Solomon project, whose goal is to detect copying between structured sources and to leverage the results in various aspects of data integration, and the Semex personal information management system, which got the Best Demo award (one of top-3) in Sigmod'05. She co-chaired Sigmod/PODS PhD Symposium'12, QDB'12, WebDB'10 and has served in the program committees of ICDE'13, PVLDB'13, Sigmod'12, VLDB'12, Sigmod'11, VLDB'11, PVLDB'10, WWW'10, ICDE'10, VLDB'09, etc. She has presented tutorials on "Data Fusion: Resolving Data Conflicts for Integration" (with Felix Naumann) at VLDB 2009, and on "Detecting Clones, Copying and Reuse on the Web" (with Divesh Srivastava) at SIGMOD 2011 and ICDE 2012.

B. Divesh Srivastava

Divesh Srivastava is the head of the Database Research Department at AT&T Labs-Research. He received his Ph.D. from the University of Wisconsin, Madison, and his B.Tech. from the Indian Institute of Technology, Bombay. He is an ACM fellow, on the board of trustees of the VLDB Endowment and an associate editor of the ACM Transactions on Database Systems. He has served as the program committee co-chair of many conferences, including VLDB 2007. His research interests and publications span a variety of topics in data management. He has presented tutorials on "Data Stream Query Processing" (with Nick Koudas) at VLDB 2003 and ICDE 2005, on "Record Linkage: Similarity Measures and Algorithms" (with Nick Koudas and Sunita Sarawagi) at VLDB 2005 and SIGMOD 2006, on "Anonymized Data: Generation, Models, Usage" (with Graham Cormode) at SIGMOD 2009 and ICDE 2010, on "Information Theory for Data Management" (with Suresh Venkatasubramanian) at VLDB 2009 and SIGMOD 2010, and on "Detecting Clones, Copying and Reuse on the Web" (with Xin Luna Dong) at SIGMOD 2011 and ICDE 2012.

V. CONCLUSIONS

This seminar reviews the state-of-the-art techniques for data integration in addressing the new challenges raised by Big Data, including *volume* and number of sources, *velocity, variety,* and *veracity.* We discuss how close we are to meeting these challenges and identify many open problems for future research.

REFERENCES

[1] Z. Bellahsene, A. Bonifati, and E. Rahm, editors. *Schema Matching and Mapping*. Springer, 2011.

[2] J. Bleiholder and F. Naumann. Data fusion. *ACM Computing Surveys*, 41(1):1–41, 2008.

[3] M. J. Cafarella and A. Y. Halevy. Web data management. In *Sigmod*, pages 1199–1200, 2011.

[4] M. J. Cafarella, A. Y. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. In *PVLDB*, pages 538–549, 2008.

[5] K. C.-C. Chang, B. He, and Z. Zhang. Toward large scale integration: Building a metaquerier over databases on the web. In *CIDR*, pages 44–55, 2005.

[6] N. N. Dalvi, A. Machanavajjhala, and B. Pang. An analysis of structured data on the web. *PVLDB*, 5(7):680–691, 2012.

[7] X. Dong and A. Y. Halevy. Indexing dataspaces. In C. Y. Chan, B. C. Ooi, and A. Zhou, editors, *SIGMOD Conference*, pages 43–54. ACM, 2007.

[8] X. Dong, A. Y. Halevy, and C. Yu. Data integration with uncertainties. In *VLDB*, 2007.

[9] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *PVLDB*, 2(1), 2009.

[10] X. L. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. *PVLDB*, 2(1), 2009.

[11] X. L. Dong and F. Naumann. Data fusion–resolving data conflicts for integration. *PVLDB*, 2009.

[12] U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg. Adaptive windows for duplicate detection. In *ICDE*, pages 1073–1083, 2012.

[13] A. K. Elmagarmid and D. Agrawal, editors. *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010*. ACM, 2010.

[14] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16, 2007.

[15] H. Elmeleegy, J. Madhavan, and A. Y. Halevy. Harvesting relational tables from lists on the Web. *VLDB J.*, 20:209–226, 2011.

[16] M. Franklin, A. Y. Halevy, and D. Maier. From databases to dataspaces: A new abstraction for information management. *Sigmod Record*, 34(4):27–33, 2005.

[17] M. J. Franklin, A. Y. Halevy, and D. Maier. A first tutorial on dataspaces. *PVLDB*, 1(2):1516–1517, 2008.

[18] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *WSDM*, 2010.

[19] L. Getoor and A. Machanavajjhala. Entity resolution: Theory, practice, and open challenges. In *VLDB*, pages 2018–2019, 2012.

[20] S. Guo, X. Dong, D. Srivastava, and R. Zajac. Record linkage with uniqueness constraints and erroneous values. *PVLDB*, 3(1):417–428, 2010.

[21] R. Gupta and S. Sarawagi. Answering table augmentation queries from unstructured lists on the web. *PVLDB*, 2:289–300, 2009.

[22] A. Y. Halevy, M. J. Franklin, and D. Maier. Principles of dataspace systems. In S. Vansummeren, editor, *PODS*, pages 1–9. ACM, 2006.

[23] O. Hassanzadeh, F. Chiang, R. J. Miller, and H. C. Lee. Framework for evaluating clustering algorithms in duplicate detection. *PVLDB*, 2(1):1282–1293, 2009.

[24] T. N. Herzog, F. J. Scheuren, and W. E. Winkler. *Data quality and record linkage techniques*. Springer, 2007.

[25] B. Howe, D. Maier, N. Rayner, and J. Rucker. Quarrying dataspaces: Schemaless profiling of unfamiliar information sources. In *ICDE Workshops*, pages 270–277. IEEE Computer Society, 2008.

[26] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy. Pay-as-you-go user feedback for dataspace systems. In *Sigmod*, 2008.

[27] A. Kannan, I. Givoni, R. Agrawal, and A. Fuxman. Matching unstructured product offers to structured product specifications. In *SigKDD*, 2011.

[28] L. Kolb, A. Thor, and E. Rahm. Dedoop: Efficient deduplication with hadoop. In *VLDB*, pages 1878–1881, 2012.

[29] L. Kolb, A. Thor, and E. Rahm. Load balancing for mapreduce-based entity resolution. In *ICDE*, pages 618–629, 2012.

[30] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: similarity measures and algorithms. In *SIGMOD*, 2006.

[31] P. Li, X. L. Dong, A. Maurino, and D. Srivastava. Linking temporal records. *PVLDB*, 4(11):956–967, 2011.

[32] X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? *PVLDB*, 6(2), 2013.

[33] X. Liu, X. L. Dong, B. C. Ooi, and D. Srivastava. Online data fusion. *PVLDB*, 4(12), 2011.

[34] J. Madhavan, S. R. Jeffery, S. Cohen, X. L. Dong, D. Ko, C. Yu, and A. Halevy. Web-scale data integration: You can only afford to pay as you go. In *CIDR*, 2007.

[35] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Y. Halevy. Google's deep web crawl. In *PVLDB*, pages 1241–1252, 2008.

[36] C. Mathieu, O. Sankur, and W. Schudy. Online correlation clustering. In J.-Y. Marion and T. Schwentick, editors, *STACS*, volume 5 of *LIPIcs*, pages 573–584. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2010.

[37] B. McNeill, H. Kardes, and A. Borthwick. Dynamic record blocking: Efficient linking of massive databases in mapreduce. In *QDB*, 2012.

[38] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *COLING*, pages 877–885, 2010.

[39] J. Pasternack and D. Roth. Making better informed trust decisions with generalized fact-finding. In *IJCAI*, pages 2324–2329, 2011.

[40] R. Pimplikar and S. Sarawagi. Answering table queries on the web using column keywords. *PVLDB*, 5(10):908–919, 2012.

[41] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *VLDBJ*, 10(4):334–350, 2001.

[42] A. D. Sarma, X. L. Dong, and A. Y. Halevy. Data modeling in dataspace support platforms. In A. Borgida, V. K. Chaudhri, P. Giorgini, and E. S. K. Yu, editors, *Conceptual Modeling: Foundations and Applications*, volume 5600 of *Lecture Notes in Computer Science*, pages 122–138. Springer, 2009.

[43] P. P. Talukdar, Z. G. Ives, and F. Pereira. Automatically incorporating new sources in keyword search-based data integration. In Elmagarmid and Agrawal [13], pages 387–398.

[44] T. Vogel and F. Naumann. Automatic blocking key selection for duplicate detection based on unigram combinations. In *QDB*, 2012.

[45] S. E. Whang and H. Garcia-Molina. Developments in generic entity resolution. *Bulletin of the Technical Committee on Data Engineering*, 34(3):52–60, 2011.

[46] W. E. Winkler. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., 1999.

[47] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. Knowl. Data Eng.*, 20:796–808, 2008.

[48] X. Yin and W. Tan. Semi-supervised truth discovery. In *WWW*, pages 217–226, 2011.

[49] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *QDB*, 2012.

[50] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *PVLDB*, 5(6):550–561, 2012.