Data Quality: The other Face of Big Data

Barna Saha, Divesh Srivastava

AT&T Labs-Research
barna, divesh@research.att.com
Preferred Duration: 1.5 *hours*

***Abstract*—In our Big Data era, data is being generated, collected and analyzed at an unprecedented scale, and data-driven decision making is sweeping through all aspects of society. Recent studies have shown that poor quality data is prevalent in large databases and on the Web. Since poor quality data can have serious consequences on the results of data analyses, the importance of veracity, the fourth 'V' of big data is increasingly being recognized. In this tutorial, we highlight the substantial challenges that the first three 'V's, volume, velocity and variety, bring to dealing with veracity in big data. Due to the sheer volume and velocity of data, one needs to understand and (possibly) repair erroneous data in a scalable and timely manner. With the variety of data, often from a diversity of sources, data quality rules cannot be specified a priori; one needs to let the "data to speak for itself" in order to discover the semantics of the data. This tutorial presents recent results that are relevant to big data quality management, focusing on the two major dimensions of (i) discovering quality issues from the data itself and (ii) trading-off accuracy vs efficiency, and identifies a range of open problems for the community.**

I. Scope and Depth of the Tutorial

With the huge volume of generated data, the fast velocity of arriving data, and the large variety of heterogeneous data, the quality of data is far from perfect. It has been estimated that erroneous data costs US businesses 600 billion dollars annually [18]. Enterprises typically find data error rate of approximately $1-5\%$, and for some companies, it is above 30% [24, 49]. In most data warehousing projects, data cleaning accounts for $30-80\%$ of the development time and budget for improving the quality of the data rather than building the system. On the web, 58% of the available documents are XML, among which only one third of XML documents with accompanying XSD/DTD are valid [42]. 14% of the documents lack well-formedness, a simple error of mismatching tags and missing tags that renders the entire XML-technology useless over these documents. These all highlight the pressing need of data quality management to ensure data in our databases consistently, accurately, completely, timely and uniquely represent the real world entities to which it refers. There has been increasing demand in industries for developing data quality management systems, aiming to effectively detect and correct errors in the data, and thus to add accuracy and value to business processes. Indeed the market for data quality tools is growing at 16% annually, way over the 7% average forecast for other IT segments [34].

With the advent of big data, data quality management has become more important than ever. Typically, *volume, velocity* and *variety* are used to characterize the key properties of big data. But to extract value and make big data operational, the importance of the fourth 'V' of big data, *veracity*, is increasingly being recognized. Veracity directly refers to inconsistency and data quality problem. As [54] states, one of the biggest problems with big data is the tendency for errors to snowball. User entry errors, redundancy and corruption all affect the value of data. Without proper data quality management, even minor errors can accumulate resulting in revenue loss, process inefficiency and failure to comply with industry and government regulations (the butterfly effect [52]).

The Big Data era comes with new challenges for data quality management. Due to the sheer volume and velocity of some data (like stock trades, or machine/sensor generated events), one needs to understand or get rid of the erroneous data extremely fast. And as multi-structured data, often from many different sources, is brought together, determining the semantics of data and understanding correlations between attributes becomes a daunting task. In contrast to traditional data quality management, it is impossible to specify all the data semantics beforehand and no global semantics may fit the entire data. We need context-aware data quality rules to detect semantic errors in our data, and better still fix those errors by using the rules. We need to learn interesting and informative rules from the dirty data itself [15, 28, 38–40, 46, 60]. We therefore go from the close-world assumption of database systems to an open world view where rules are *learned* from data, validated and updated incrementally as more data is gathered and based on the most recent data. Due to the variety of data sources, these rules may apply only to certain subsets of data [3, 27, 38, 40]. Such *conditioning* requires that proper metrics be ascertained to find statistically robust rules as opposed to outliers since rules are inferred from dirty data itself [38, 40]. Violation to rules indicate data inconsistency. Based on the applications, either one deals with these inconsistencies without repairing them, or finds ways to repair them [2, 10, 13, 20, 30–32, 35, 41, 45, 47, 50, 51, 55, 59]. Due to inherent noise in discovering rules, repairing must adjust between

inconsistent data and inaccurate constraints [6, 16].

These various stages of data quality management: discovering rules, checking for inconsistencies, repairing, need to be done in a very scalable manner. This brings in the *efficiency vs accuracy trade-off*. We may have to live with some approximations since generating optimal results are costly [40, 46, 47]. We need either *centralized near-linear time* procedures or *distributed map-reduce processing* to deal with the volume [29, 40, 47]. To tackle the velocity, we need *incremental and streaming processing* [33, 48, 56, 57].

In this tutorial we explore the challenges of data quality management that arise due to volume, velocity and variety of data in the Big Data era. Specifically, our goal is to cover two major dimensions of big data quality management, (i) discovering/learning based on data and (ii) accuracy vs efficiency trade-off under various computing models. We will present the state of the art research in these dimensions for relational, structured and semi-structured data and identify many open problems for future research.

II. Intended Audience

The target audience is anyone with interest in learning data quality challenges in Big Data environment. We expect the tutorial to appeal to a large portion of the ICDE community:

- Researchers in the fields of data cleansing, data consolidation, data extraction, data mining, and Web information management.
- Practitioners developing and distributing products in the data cleansing, ETL & data warehousing, and master data management areas.

The assumed level of mathematical sophistication will be that of the typical conference audience.

III. Tutorial Outline

Our 1.5 hours tutorial is organized as follows.

A. *Introduction (*15 *minutes)*

- Motivating Examples for Big Data Quality
- Different Aspects of Data Quality: consistency, accuracy, completeness and timeliness
- Statistical vs Logical Data Quality Management
- Logical Data Quality Management
  - Dependency Theory
    - Extension with conditions and similarity
    - Static Analysis
    - Aspects of Big Data: volume, velocity, variety
- Overview of various Data Quality Tools

We start the tutorial by providing a variety of real world cases.The principles for data quality management can be broadly classified as statistical/quantitative vs logical/constraint-based. In the former, data is corrected based on statistics over value distribution [19, 44], whereas the latter intends to develop models and inconsistencies are reported as violations of the model [3, 10]. In the tutorial, we mainly focus on logical/constraint-based data quality management. Integrity constraints (ICs) have been recently repurposed towards improving the quality of data. Traditional types of ICs such as key constraints, check constraints, functional dependencies (FD) and their extension conditional functional dependencies (CFD), conditional inclusion dependencies, matching dependencies etc. [4, 27, 60] have been proposed for data quality management. We review the notion of dependency theory and static analysis [4, 27, 31]. Our main goal is to explore the developments in dependency theory pertaining to the big data challenges of volume, velocity and variety. We provide an overview of available data quality management tools and platforms and outline the additional requirements to tackle these challenges [1, 17, 22, 25, 32, 37].

B. *Discovering/Learning Data Quality Semantics (*30 *minutes)*

- Discovering Logical Model
  - Conditional vs Full [Ex. Learning CFD vs FD]
  - Approximate/Soft vs Exact
    - Measures of robustness
  - Template based learning: learning pattern tableau
    - Hold Tableaux for summarization vs Fail Tableaux for outliers
  - Learning keys, FDs, DTD and XSD for semi-structured documents
  - Efficiency vs Accuracy
    - Centralized near linear time algorithms with approximation guarantees
    - Incrementally generating semantic rules
    - Incrementally maintaining pattern tableaux

Due to the large variety of sources from which data is collected and integrated, for its sheer volume and changing nature, it is impossible to manually specify data quality rules. Big data comes with a major promise: having more data allows the "data to speak for itself," instead of relying on unproven assumptions and weak correlations. The key is to learn the rules from the data itself [14, 15, 28, 46, 60]. The rules need to be learnt efficiently, in near linear time centralized or distributed and since the data itself is dirty, to make the rules robust against outliers, approximations need to be allowed [37, 38, 40, 46]. Rules may apply only to part of the data [28, 39, 40]. For learning rules for semi-structured data, both value and structure need to be taken care of. In this part of the tutorial, we focus on efficient learning of integrity constraints [28], schema [9], DTD [8, 36], whether they apply to the entire database or partially [15, 28], proper measures to obtain statistically robust rules [39], generating them incrementally to cope with the dynamic nature of big data [12]. We also focus on template based learning such as learning pattern tableaux for CFD, sequential dependency, conservation

dependency etc. [37, 38, 40]. Approximation plays a crucial role both for finding statistically robust rules and also for designing efficient algorithms.

C. Detecting/ Repairing Inconsistencies (30 minutes)

- Repairing Inconsistencies in Relational Databases
  - Minimal Repairs, Sample of Possible Repairs, Chase based Repairs
- Repairing Structural Problems in Semi-structured Data
  - Well-formedness, Validity: XML, Web data etc.
- Detecting Inconsistencies in Distributed Data and Streaming Data
- Validation of streaming XML/ Web documents

Once rules have been specified, inconsistencies in data show up as violations to the rules. There are two main approaches to deal with inconsistencies, either they are removed [13, 30, 35, 59] or queries are answered in a consistent/approximate manner without repairing the database [20, 45, 50]. In this tutorial, we focus only on the first approach. We discuss how inconsistencies in relational and semi-structured data can be detected incrementally in streaming and distributed fashion [29, 33, 48]. Since rules are discovered based on dirty data, inconsistencies may appear as an effect of faulty rules. Therefore, it is required to detect whether the data is inconsistent or the model is incorrect [6, 16]. Computation of all possible repairs requires to explore a space of solutions of exponential size with respect to the size of the database which is infeasible in practice. Hence conditions are imposed on the computed repairs to restrict the search space. These conditions include, e.g., various notions of cost-based minimality, maximum likelihood [2, 10, 13, 58], repairing using record linkage and master data [31, 32, 41], sampling based repairs [5] and chase-based algorithm to fine-tune the trade-off between quality and scalability of the repair process [7, 11, 43]. We will contrast among these approaches in terms of efficiency, repair strategies, value and solution preferences. For semi-structured data, we highlight the recent progress on finding top-k repairs and validating them in a scalable manner [47, 51, 55]. Again to ensure efficiency, many of these algorithms require approximation and are computed in a distributed or streaming model [29, 48, 56, 57].

D. Open Problems (15 minutes)

Distributed and streaming discovery of data quality semantics as well as detection and repairing inconsistencies is a fledgling topic and there remain many open problems in this area. Apart from this, there are several new directions for data quality research to consider such as using master data for repairing partially-closed databases [23], crowdsourced data cleaning, using value and structure for discovering interesting rules in semi-structured documents and resolving conflicts.

IV. Relationship to Prior Seminars

This is the first time that this seminar will be presented. There are some available recent books, tutorials and invited presentations that cover static analysis of CFD and matching dependency [21, 24, 26]. In addition, [21, 24] cover certain topics that are included in this tutorial, such as discovering CFD [28] and repairing violations to CFD [31, 32]. We are the first to discuss how the state-of-the-art techniques on data quality for relational, structured and semi-structured data address the challenges raised by the Big Data environment. Previous seminars and books did not focus on approximation vs efficiency trade-off, while for discovering/repairing rules, our presentation will cover much beyond CFD on relational data.

V. Biographies

A. Barna Saha

Barna Saha is a researcher at AT&T Labs-Research, where she joined after receiving her Ph.D. from University of Maryland, College Park in 2011. Previously she received a Masters Degree from Indian Institute of Technology, Kanpur, and received a Bachelors Degree from Jadavpur University in India. Her research interests include algorithms, optimization and various aspects of databases with emphasis on data quality, data integration and distributed/uncertain data management. She is the recipient of Deans Dissertation Fellowship Award, University of Maryland, 2010, for her PhD work on scalable approximation algorithm design for distributed workload management. She also received the Best Paper award for her work on ranking under uncertainty at VLDB 2009.

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. 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, on "Detecting Clones, Copying and Reuse on the Web" (with Xin Luna Dong) at SIGMOD 2011 and ICDE 2012, and on "Big Data Integration" (with Xin Luna Dong) at ICDE 2012 and VLDB 2013.

REFERENCES

[1] U. Boobna and M. de Rougemont: Correctors for XML Data. XSym 2004: 97-111.

[2] P. Bohannon, M. Flaster, W. Fan and R. Rastogi: A Cost-Based Model and Effective Heuristic for Repairing Constraints by Value Modification. *SIGMOD 2005*: 143-154

[3] P. Bohannon, W. Fan, F. Geerts, X. Jia, A. Kementsietsidis: Conditional Functional Dependencies for Data Cleaning. *ICDE 2007*: 746-755

[4] L. Bravo, W. Fan and S. Ma, Extending dependencies with conditions, *VLDB 2007*:243–254.

[5] G. Beskales, I. F. Ilyas and L. Golab: Sampling the Repairs of Functional Dependency Violations under Hard Constraints. *PVLDB 3(1)*: 197-207 (2010)

[6] G. Beskales, I. F. Ilyas, L. Golab, A. Galiullin, On the relative trust between inconsistent data and inaccurate constraints, *ICDE 2013*: 541-552.

[7] L. Bertossi, S. Kolahi and Laks VS Lakshmanan: Data cleaning and query answering with matching dependencies and matching functions. *ICDT 2011*:268:279.

[8] G. J. Bex, F. Neven, T. Schwentick, K. Tuyls: Inference of Concise DTDs from XML Data, *VLDB 2006*: 115-126

[9] G. J. Bex, F. Neven, S. Vansummeren: Inferring XML Schema Definitions from XML Data, *VLDB 2007*: 998-1009.

[10] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma: Improving data quality: Consistency and accuracy. *VLDB 2007*:315326.

[11] Y. Cao, W. Fan and W. Yu, Determining the relative accuracy of attributes, *SIGMOD 2013*: 565-576.

[12] G. Cormode, L. Golab, F. Korn, A. McGregor, D. Srivastava, X. Zhang: Estimating the confidence of conditional functional dependencies. *SIGMOD Conference 2009*: 469-482.

[13] X. Chu, I. F. Ilyas and P. Papotti, Holistic data cleaning: Putting violations into context, *ICDE 2013*: 458–469.

[14] X. Chu, I. Ilyas and P. Papotti, Discovering Denial Constraints, 2013

[15] F. Chiang, R. J. Miller: Discovering data quality rules. *PVLDB 2008*: 1166-1177.

[16] F. Chiang, R. J. Miller: A unified model for data and constraint repair.*ICDE 2011*: 446-457.

[17] M. Dallachiesa, A. Ebaid, A. Eldawy, A. K. Elmagarmid, I. F. Ilyas, M. Ouzzani, N. Tang, NADEEF: a commodity data cleaning system, *SIGMOD 2013*: 541-552.

[18] W. W. Eckerson. Data quality and the bottom line: Achieving business success through a commitment to high quality data. *Data Warehousing Institute, 2002*.

[19] L. Berti-Equille, T. Dasu, D. Srivastava: Discovery of complex glitch patterns: A novel approach to Quantitative Data Cleaning. *ICDE 2011*: 733-744.

[20] T. Eiter, M. Fink, G. Greco and D. Lembo: Repair localization for query answering from inconsistent databases.*TODS 2008*.

[21] W. Fan: Data Quality: Theory and Practice. *WAIM 2012*: 1-16.

[22] A. Fuxman, E. Fazli, R. J. Miller: ConQuer: Efficient Management of Inconsistent Databases. *SIGMOD 2005*: 155-166

[23] W. Fan, F. Geerts: Capturing missing tuples and missing values. *PODS 2010*: 169-178.

[24] W. Fan and F. Geerts: Foundations of Data Quality Management. *Morgan & Claypool 2012*.

[25] W. Fan, F. Geerts and X. Jia, Semandaq: a data quality system based on conditional functional dependencies, *PVLDB 2008*:1460–1463.

[26] W. Fan, F. Geerts and X. Jia, A revival of integrity constraints for data cleaning, *PVLDB 2008*, Tutorial: 1522–1523.

[27] W. Fan, F. Geerts, X. Jia and A. Kementsietsidis, Conditional functional dependencies for capturing data inconsistencies, *ACM TODS 2008*:6:1–6:48.

[28] W. Fan, F. Geerts, J. Li, M. Xiong: Discovering Conditional Functional Dependencies. *IEEE Trans. Knowl. Data Eng. 2011*: 683-698.

[29] W. Fan, F. Geerts, S. Ma, H. Mller: Detecting inconsistencies in distributed data: *ICDE 2010*: 64-75.

[30] W. Fan, F. Geerts, N. Tang and W. Yu, Inferring data currency and consistency for conflict resolution, *ICDE 2013*: 470–481.

[31] W. Fan, J. Li, S. Ma, N. Tang and W. Yu: Interaction between record matching and data repairing. *SIGMOD 2011*: 469–480.

[32] W. Fan, J. Li, S. Ma, N. Tang, W. Yu: CerFix: A System for Cleaning Data with Certain Fixes. *PVLDB 2011*: 1375-1378.

[33] W. Fan, J. Li, N. Tang, W. Yu, Incremental Detection of Inconsistencies in Distributed Data, *ICDE 2012*: 318-329.

[34] Gartner. Forecast: Data quality tools, worldwide, 2006- 2011. *Technical report, Gartner, 2007*.

[35] H. Galhardas, D. Florescu, D. Shasha, E. Simon and C. Saita, Declarative Data Cleaning: Language, Model, and Algorithms, *VLDB 2001*: 371–380.

[36] M. N. Garofalakis, A. Gionis, R. Rastogi, S. Seshadri and K. Shim: DTD Inference from XML Documents: The XTRACT Approach. *IEEE Data Eng. Bull.*: 19-25 (2003)

[37] L. Golab, H. Karloff, F. Korn, Flip and D. Srivastava, Data Auditor: exploring data quality and semantics using pattern tableaux, *PVLDB 2010*: 1641-1644.

[38] L. Golab, H. J. Karloff, F. Korn, A. Saha, D. Srivastava: Sequential Dependencies. *PVLDB 2009*: 574-585.

[39] L. Golab, H. J. Karloff, F. Korn, D. Srivastava, B. Yu: On generating near-optimal tableaux for conditional functional dependencies. *PVLDB 2008*: 376-390.

[40] L. Golab, H. J. Karloff, F. Korn, B. Saha, D. Srivastava: Discovering Conservation Rules. *ICDE 2012*: 738-749.

[41] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu: Towards certain xes with editing rules and master data. *PVLDB, 2010*:173184.

[42] S. Grijzenhout and M. Marx: The quality of the XML web. *CIKM 2011*: 1719-1724.

[43] F. Geerts, G. Mecca, P. Papotti, and D. Santoro, The LLUNATIC Data-Cleaning Framework, *PVLDB 2013*.

[44] J. Hellerstein, Quantitative data cleaning for large databases, *UNECE 2008*.

[45] Marios Hadjieleftheriou, Divesh Srivastava: Approximate String Processing. *Foundations and Trends in Databases 2011*: 267-402.

[46] I. F. Ilyas, V. Markl, P. J. Haas, P. Brown, A. Aboulnaga: CORDS: Automatic Discovery of Correlations and Soft Functional Dependencies.*SIGMOD Conference 2004*: 647-658.

[47] F. Korn, B. Saha, D. Srivastava, S. Ying, On Repairing Structural Problems in Semi-structured Data. *PVLDB 2013*.

[48] F. Magniez, C. Mathieu and A. Nayak, Recognizing well-parenthesized expressions in the streaming model, *STOC 2010*: 261-270

[49] T. Redman. The impact of poor data quality on the typical enterprise. *Commun. ACM 1998*.

[50] S. Razniewski, W. Nutt: Completeness of Queries over Incomplete Databases. *PVLDB 2011*: 749-760.

[51] N. Suzuki: Finding an optimum edit script between an XML document and a DTD. *SAC 2005*: 647-653

[52] S. Sarsfield, The Butterfly Effect of Data Quality, *The Fifth MIT Information Quality Industry Symposium, 2011*.

[53] S. Song and L. Chen, Discovering matching dependencies, *CIKM 2009*: 1421–1424.

[54] J. Tee, Handling the four 'V's of big data: volume, velocity, variety, and veracity, *TheServerSide.com 2013*.

[55] H. Samimi, M. Schaefer, S. Artzi, T. Millstein, F. Tip and L. Hendren: Automated Repair of HTML Generation Errors in PHP Applications Using String Constraint Solving. *Int'l Conf. Software Engineering 2012*

[56] L. Segoufin and C. Sirangelo: Constant-Memory Validation of Streaming XML Documents Against DTDs. *ICDT 2007*: 299-313

[57] L. Segoufin and V. Vianu: Validating Streaming XML Documents. *PODS 2002*: 53-64

[58] M. Yakout, L. Berti-Equille and A. K. Elmagarmid, Don't be SCAREd: use SCalable Automatic REpairing with maximal likelihood and bounded changes, *SIGMOD 2013*: 553-564.

[59] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, I. F. Ilyas: Guided data repair. *PVLDB 2011*: 279-289.

[60] M. Zhang, M. Hadjieleftheriou, B. Ooi, C. M. Procopiuc, D. Srivastava: On Multi-Column Foreign Key Discovery. *PVLDB 2010*: 805-814.