International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 4, Issue 3, Apr -May 2015, 1-8

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

ANALYSIS OF DATA QUALITY ISSUES IN HETEROGENEOUS ENVIRONMENT

SAPNA PUJARA[1] & KANWAL GARG[2]

[1]Research Scholar, DCSA, Kurukshetra University, Kurukshetra, India

[2]Assistant Professor, DCSA, Kurukshetra University, Kurukshetra, India

ABSTRACT

Amid the growth of distributed computing, the collection and maintenance of data from diverse and heterogeneous sources is the major necessity of each organization. Organizations need to exchange and share the information across the globe. In such a scenario, the databases are intrinsically scattered and managed by different people with different objectives, which intensifies the diversity between the technologies and standards being used at each site. Such Heterogeneity is actually attributed to syntactic and semantic differences while describing same real world entity in various data sources. Moreover, such heterogeneity leads to some security & privacy issues. In this keynote paper an in-depth investigation of primary issues (syntactic, semantic), succeeding issues (preprocessing, integration & transformation) and consequential issues of privacy in such heterogeneous environment.

KEYWORDS: Data Quality, Heterogeneous Environment

INTRODUCTION

A legacy database growing over time contains enormous heterogeneity in data. Usually, this happens due to poor database design and an endeavor of maintaining the structure of original source data while integration. This actually degrades the quality of data and places the organization into risk-prone zone. Legacy data must be cleaned up prior to conversion, integration and finally its usage for some decision making or an organization may undoubtedly have to face serious data problems later. The matter of dirty data also escorts the measures for regularly auditing the quality of information used which advances the cost [Müller, H., and Freytag, J.C., 2005]. Heterogeneous environment like a legacy database actually encourages an organization for exploiting the decentralized nature of upcoming web-based technologies like data-centre, clouds etc. But the essence of high quality data lies in rapid and early detection of quality problems in data and employing proactive measures to eliminate those issues. Furthermore, once cleaned, data have to be monitored regularly for maintaining its standard through some constraints etc.

Data contaminations have a multifaceted effect; by nature, they have a tendency to concentrate around high volume data users [Agrawal, D., Bernstein, P., et al., 2011].

To really facilitate the utilization of assorted data sources, the data have to be accurate, fresh, complete and interpretable.

Data contaminations have a multifaceted effect; by nature, they have a tendency to concentrate around high volume data users [Agrawal, D., Bernstein, P., et al., 2011]. To really facilitate the utilization of assorted data sources, the data have to be accurate, fresh, complete and interpretable.

HETEROGENEOUS ENVIRONMENT

The heterogeneous database environment contains a mix of unstructured (web), semi-structured (XML) or fully-structured (RDBMS) data sources [Carlo, B., Daniele, B., et al., 2011] which may impose more than a few limitations over the data. Data from different sources keep on to be supplemented for more information. The data may be non-fresh, incomplete and may also have schematic differences since the update frequency of database may be irregular at different sites [Liu, H., and Dou, D., 2008]. The heterogeneity of data-sources leads to non-uniform semantics of data since different databases are designed by different people for diverse applications in different context using probably dissimilar technology. The price of going without any measures for coping up with this heterogeneity is, lack of definite relationship between data, problem in integration of data, type conflicts due to non-uniformity and non-standards, structural differences and heavier query processing are the foremost among all.

DATA QUALITY ISSUES OF HETEROGENEOUS DATABASES ENVIRONMENT

From technology viewpoint, Data are the only eminent foundation of organization and need to be maintained well using well established quality improvement techniques. The process of raising and maintaining the efficient and high-quality data has to face a number of problems. We took an example case study with four geographically scattered sites to figure out these issues.

Schema at Site 1: EMPLOYEE (Structured Database)

S.No.	ATTRIBUTE	WIDTH	DATATYPE	CHECK
1	SSN	5	ALPHANUMERIC	"$$0000" pattern
2	FNAME	10	CHARACTER	
3	LNAME	10	CHARACTER	
4	D.O.B	8	DATE	"mm.dd.yy" pattern
5	GENDER	1	INTEGER	"M / F" pattern
6	QUALIFICATION	10	CHARACTER	
7	PH.NO.	10	INTEGER	
8	MARITAL STATUS	1	INTEGER	
9	ADDRESS	50	ALPHANUMERIC	
10	DEPARTMENT	15	CHARACTER	"_ _ _ _ _ _ _ _" letters(8) pattern

Schema at Site 2: EMP+DEPENDENTS (Unstructured Database)

S.No.	ATTRIBUTE	WIDTH	DATATYPE	CHECK
1	EMP_ID	8	INTEGER	"00000" pattern
2	Date of Birth	8	DATE	"dd/mm/yy" pattern
3	GENDER	6	CHARACTER	"Male / Female" pattern
4	DEPNDENT NAME	20	CHARACTER	
5	RELATIONSHIP	20	CHARACTER	

Schema at Site 3: EMP+DEPARTMENT (Semi-Structured Database)

S.No.	ATTRIBUTE	WIDTH	DATATYPE	CHECK
1	EMPNO	8	ALPHANUMERIC	"$0000" pattern
2	NAME	20	ALPHANUMERIC	
3	SEX	1	INTEGER	"0 / 1 " pattern
4	ADDRESS	40	CHARACTER	
5	JOINING_DATE	10	DATE	"dd-mm-yyyy" pattern
6	DEPTT_NAME	5	CHARACTER	"_ _ _ _." Letters(4)+ period pattern
7	DEPTT_NUMBER	5	INTEGER	

8	MGR_ID	8	ALPHANUMERIC	"$0000" pattern
9	DEPTT_LOCATION	15	CHARCTER	

Schema at Site 4: EMP+SALARY (Structured Database)

S.No.	ATTRIBUTE	WIDTH	DATATYPE	CHECK
1	EMP_ID	4	INTEGER	"0000" pattern
2	NAME	20	CHARACTER	
3	WORKING_HOURS	2	INTEGER	
4	DEPARTMENT	6	CHARACTER	"_ _ _ _ _ _" letters(6) pattern
5	WAGES_PER_HOUR	4	INTEGER	

The leading concerns are presented here using these four sites under the assumptions like use of dissimilar technology and attributes' specifications for all of them:

Lack of Synchronization Support

The most inevitable and contrast-prone feature presented in above case study of heterogeneous environment is the use of divergent technology at various sites according to the ease, available support and other facilities. The multiple sources virtually having same information cannot share it amongst each-other since each source may be using dissimilar technology which may not support the collaboration features like import or export data. Further, direct synchronization requires harmonized syntactic structure of the data which may not be always the case [Naiman, Channah F., and Arison M. Ouksel, 1995].

Poor Reliability

The deployment of non-versatile technology at various sources of case study also points to peculiarity in hardware being used for storage of data. The dissimilarity in maintaining RAID levels and using different access pattern disks actually demotes the overall reliability of the entire system where some sites are under-managed and some are over-optimized. The absence of sufficient data provenance techniques also drops the curtain over reliability standards [Simmhan, Yogesh L., Plale, B., and Gannon, D., 2005].

Difficult To Establish Relationship among Entities

Since the process of integration of data from multiple heterogeneous sources requires identification of similar entities. It is quite thorny to figure out parallel attributes of entities so that a relationship may be established between respective entities [Song, Dezhao, 2012]. For example, schema at site 1 stores the SSN and FNAME with LNAME as the details of the employee, whereas schema at site 4 contains the same information in EMP_ID and a single compounded column named NAME. So, establishing the relationship of the same employee becomes further complex when heterogeneity lies in semantics and abstraction level too.

Non-Uniformity Conflicts across Data Sources

The sharing and harmony of data across multiple sources is achievable only if uniform and standard data patterns are being used at each source. As per our case study example schema, the non-uniformity presented in the fields at various sites regarding data type, field width, field patterns etc. [Dai, B. T., Koudas, N., et al., 2006] may present many hurdles for incorporating the feature of quality in data[March, S., Hevner, A. and Ram, S., 2000].

Uncertain to Delete/Modify a Source

Each source in such an environment contains its own designed schemas, where it cannot be deduced that either any particular source is an exact replica or subset of another source until data-integration process. According to exemplary case study, since schema at each site either contains homonyms or synonyms which makes the matching process quite ambiguous, hence to delete a source entirely because of replication or even modify it also becomes skeptical.

Abstraction Level

Even though the data are maintained for the same purpose, delivering the same organization; do not match with other sources in the terms of abstraction level. Since each source is designed and maintained by different people with its own technology, the schematic variations are present, hence, information coverage of each instance may be different or incomplete when compared to similar entity [Liu, H., and Dou, D., 2008]. Case study makes is totally evident that employee information is being stored at all locations with employee and salary information collectively at one site and employee and department information at other site and employee and dependents information at another site. So, the similarity and quality of data is not maintained throughout each source.

Data Integration Concern

The integration of data from various sources is easy and smooth provided each source has semantic homogeneity, which is not the case here. Semantic heterogeneity [George, D., 2005] [Pincus, Z., and Musen, M.A., 2003] presented in case study exists in various forms like homonyms, synonyms, type formats, scale of representation of data, constraints implementation [Goldsmith, D. L., Thuraisingham, B.M., and Bedford, M. A.]. The data integration, which in itself is a crucial operation for any organization when combined with such diversity [Wang, X., Huang, L.P., et al., 2011] becomes quite complex and error-prone.

Structural Differences

Heterogeneity in various sources is primarily attributed to the structural difference in them. These variations may exist in many forms, like an entity is stored in two or more sources but with different attributes with dissimilar specifications, different constraints implemented hence conveying a different presentation about the same information of the same entity. Given case study also presents the structural differences existing in the attributes at various sites. For example, the EMP_ID (Primary Key) itself is stored with different data-type and different storage format at various sites. These differences commonly enhance the complexity of the mining process.

Non-Standard Measures of Security, Privacy and Authentication

The non-uniformity presented in the above case study due to technology, support and API etc. is a natural phenomenon in a heterogeneous environment. Such non-consistency is also present in terms of implementation of diverse schemes of security, privacy and authentication measures in databases which are although maintained by same organization but are poles apart [Sheth, A. P., and Larson, J.A., 1990]. These numerous methods of authentication and security at discrete sources actually hinder the path of collaboration and harmonizing the sharing of data across distinct sources.

Preprocessing Issues

The preprocessing of heterogeneous data-sources is fully asymmetric because each phase of preprocessing involves non-uniform hard work for the cleaning and integration of the data located at all sites. The preprocessing phases for heterogeneous databases are actually enclosed with a super set of the problems with homogeneous environment. As presented in the example case, the preprocessing of the data present at each location would actually involve repeating the same technique (like transformation etc.) for each site. Application of preprocessing methods over different data-sources require the deployment of different mechanisms for cleaning etc. which leads to dissimilar type of resultant data at each site which actually turns the process of integration and other steps quite complex to perform.

Dispossess any Automated Tool

The homogeneity in structure and semantics of the schema (as presented in the case) is the prerequisite for the existence and success of any automated tool for recording, maintaining and mining the data [Thion-Goasdoue, V., Nugier, S., et al., 2007]. This is obviously missing in such a legacy database environment hence automation for entire process of input to output and automated exchange of data between [Eckert, Roland, and Specht, G., 2004] sources is nearly unachievable.

Transformation Apprehensions

While integration of the data from various sources is underway, there is need to slightly alter the presentation or structure of the data in order to make it alike. These purge and merge process essentially present the concern for successfully transforming the shape of the data like changing the scale of the attributes, data-type of fields etc. From the above mentioned schemas in case-study, it is quite obvious that since the data is being maintained at different sites by different people hence have different representations besides having dissimilar checks, patterns and default values for alike fields. So, it becomes very complex operation to converge all the alike fields on the same scale.

CONCLUSIONS

Data quality improvement is one of the foremost issues for every organization and this becomes more crucial when it comes to raise and maintain the quality of data over heterogeneous sources like legacy system. Although many researchers have repeatedly drawn attention to various problems in the context but still there are some obscure areas which must be paid rational attention to incorporate the data quality features in heterogeneous environment [Zhu, H., Madnick, S., et al., 2012]. The major issues highlighted by various researchers circulate around data-preprocessing, data-profiling [Rahm, E., and Do, H. H., 2000] and maintenance of cleansed data [Müller, H., and Freytag, J.C., 2005]. This paper takes further step and explored all thin-line hindrances of heterogeneous sources which include security & privacy measures, data scattered over different sources with different structure, pre-processing techniques, difficulty in transformation and lack of synchronization that leads to reduced reliability. All of the above mentioned matter of concerns should be resolved/handled using well established methods or techniques to enhance the data quality. The discussed area is a blazing research issue and if these concerns are attended thoroughly then the outcomes will be advantageous in the field of maintaining data quality over heterogeneous sources and will endow with a significant ease for maintaining data quality in heterogeneous environment.

REFRENCES

1. Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., et al., 2011, "*Challenges and Opportunities with Big Data 2011-1*."
2. Carlo, B., Daniele, B., Federico, C., and Simone, G., 2011, "*A data quality methodology for heterogeneous data*." International Journal of Database Management Systems (IJDMS) 3, no. 1.
3. Dai, Bing Tian, Nick Koudas, Beng Chin Ooi, Divesh Srivastava, and Suresh Venkatasubramanian,2006, "*Rapid identification of column heterogeneity*." In Data Mining, 2006. ICDM'06. Sixth International Conference on, pp. 159-170. IEEE.
4. Dai, Bing Tian, Nick Koudas, Beng Chin Ooi, Divesh Srivastava, and Suresh Venkatasubramanian, 2006, "*Rapid identification of column heterogeneity*." In Data Mining, ICDM'06. Sixth International Conference on, pp. 159-170. IEEE.
5. Eckert, Roland, and Günther Specht., 2004, "*Challenge of Design Data Exchange between heterogeneous Database Schema*.", pp. 125-132.
6. George, David, 2005, "*Understanding structural and semantic heterogeneity in the context of database schema integration*." Journal of the Department of Computing, UCLAN 4: 29-44.
7. Goldsmith, Deborah L., Bhavani M. Thuraisingham, and M. A. Bedford. "*Maintaining integrity in distributed and heterogeneous database systems*."
8. Liu, Haishan, and Dejing Dou, 2008, "*An Exploration of Understanding Heterogeneity through Data Mining*." In Proceedings of KDD 2008 Workshop on Mining Multiple Information Sources, pp. 18-25.
9. March, Salvatore, Alan Hevner, and Sudha Ram, 2000, "*Research commentary: an agenda for information technology research in heterogeneous and distributed environments*." Information Systems Research 11, no. 4: 327-341.
10. Müller, Heiko, and Johann-Christph Freytag, 2005, "Problems, methods, and challenges in comprehensive data cleansing.", Professoren des Inst. Für Informatik.
11. Naiman, Channah F., and Arison M. Ouksel, 1995, "A classification of semantic conflicts in heterogeneous database systems." Journal of Organizational Computing and Electronic Commerce 5, no. 2: 167-193.
12. Pincus, Zachary, and Mark A. Musen, 2003, "*Contextualizing heterogeneous data for integration and inference*." In AMIA Annual Symposium Proceedings, vol. 2003, p. 514. American Medical Informatics Association.
13. Rahm, Erhard, and Hong Hai Do., 2000, *"Data cleaning: Problems and current approaches."* IEEE Data Eng. Bull. 23, no. 4: 3-13.
14. Sheth, Amit P., and James A. Larson, 1990, "Federated database systems for managing distributed, heterogeneous, and autonomous databases." ACM Computing Surveys (CSUR) 22, no. 3: 183-236.
15. Simmhan, Yogesh L., Beth Plale, and Dennis Gannon, 2005, "A survey of data provenance techniques." Computer Science Department, Indiana University, Bloomington IN 47405.

16. Song, Dezhao, 2012, "Scalable and domain-independent entity coreference: establishing high quality data linkages across heterogeneous data sources." In The Semantic Web–ISWC, pp. 424-432. Springer Berlin Heidelberg.
17. Thion-Goasdoué, Virginie, Sylvaine Nugier, Dominique Duquennoy, and Brigitte Laboisse, 2007, "An Evaluation Framework for Data Quality Tools."
18. Wang, Xin, Linpeng Huang, Xiaohui Xu, Yi Zhang, and Jun-Qing Chen, 2011, "A Solution for Data Inconsistency in Data Integration." J. Inf. Sci. Eng. 27, no. 2: 681-695.
19. Zhu, Hongwei, Stuart Madnick, Yang Lee, and Richard Y. Wang, 2012, "Data and Information Quality Research: Its Evolution and Future." Working Paper, MIT, USA.