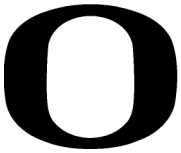
Presented to the Interdisciplinary Studies Program:



UNIVERSITY OF OREGON
APPLIED INFORMATION MANAGEMENT

Applied Information Management and the Graduate School of the University of Oregon in partial fulfillment of the requirement for the degree of Master of Science

Best Practices in Using Semantic Transformation in Data Integration to Address Data Integrity Issues

CAPSTONE REPORT

Saketh Balijepalli
Data Integration Strategist
Information Technology Services
Oregon State University Foundation

University of Oregon
Applied Information Management Program

Spring 2019

Continuing and Professional Education
1277 University of Oregon
Eugene, OR 97403-1277
(800) 824-2714

Approved by

_____

Dr. Kara McFall
Director, AIM Program

Best Practices in Using Semantic Transformation in Data Integration to Address Data Integrity Issues

Saketh Balijepalli

Oregon State University Foundation

Abstract

Data integration is a digital technology used to combine data from multiple sources and provide users with a unified view of data assets (Chen, Hu, & Xu, 2015; Davidovski, 2018). Data is typically loaded into a data warehouse via Execute-Transform-Load (ETL) operations (Hose et al. 2015); sometimes the transformed data lacks data integrity. This paper explores best practices in using semantic transformation to address data integrity issues caused by data integration.

Keywords: data integration, ETL, syntactic transformation, semantic transformation, business intelligence, data analytics, semantic ETL, data warehouse, semantic information, data integrity.

Table of Contents

Introduction to the Annotated Bibliography

Problem Description

Data integration (DI) is a digital technology used to combine data from various heterogeneous sources and provide users with a unified view of an organization's data assets (Chen, Hu, & Xu, 2015; Davidovski, 2018). Cabrera et al. (2018) explain that DI consists of three main steps: (a) parsing/cleaning, (b) transformation, and (c) aggregation. Parsing or cleansing involves identifying the records, fields, and/or other components of the input data and performing checks to ensure that the data is well-formed, whereas transformation involves translating input data into the form expected by the primary computation, typically converting from a file-oriented format to a memory-oriented format (Cabrera et al., 2018). Aggregation is conducted on any pre-analytics computations that result in aggregate information about the input (Cabrera et al., 2018). When DI is performed effectively, an organization is presented with a unified view of data from different sources that can be used to run queries, develop reports, and enable analyses (Cabrera et al., 2018; Hose, Nath, & Pedersen, 2015). The ultimate goal of DI is to produce actionable business intelligence (BI) (Hose et al., 2015), defined as the collection of "concepts and methods to improve business decision making by using fact-based support systems. BI also includes the underlying architecture, tools, databases, applications and methodologies" (Chen, Chen, & Lim, 2013, p. 1).

Data integration is commonly implemented to increase efficiency when performing analyses on ever-growing, prolific data (Cabrera et al., 2018). During the process of DI, data from heterogeneous sources is unified and often stored in a data warehouse (DW), which is "a repository used to store large data volumes from various operational databases in enterprises" (Hose et al., 2015, p. 15). Similar to the process of parsing/cleaning, transformation, and

aggregation that Cabrera et al. (2018) describe for data integration, data is loaded into a DW via Execute-Transform-Load (ETL) operations (Hose et al. 2015). The ETL process is described as "the process in data warehousing that extracts data from outside sources, transforms it to fit operational needs, which can include quality checks, and loads it into the end target database, typically a data warehouse" (Bansal, Chakraborty, & Padki, 2017, p. 414). Data warehouses serve as data sources for numerous data visualization activities, where data visualization is defined as "the representation and presentation of data that exploits our visual perception abilities in order to amplify cognition" (Kirk, 2012). The ultimate goal of DWs is to help in creating better decisions by enabling business analytics (Abedjan, Ilyas, Morcos, Ouzzani, Papotti, & Stonebraker, 2015; Hose et al., 2015), defined as any data-driven process that provides insight and achieves a business outcome (Stubbs, 2011).

One common issue that arises with DWs and other sources of integrated data is lack of data integrity, defined as "the degree to which a collection of data are complete, consistent and accurate" (McDowall, 2019, p. 11). Lack of data integrity has been a key issue with DI even after the introduction of numerous techniques for maintaining data reliability and consistency issues (Chen et al., 2015). One root cause of data integrity issues with DI occurs when heterogeneous data sources must be integrated (Abedjan et al., 2015). Abedjan et al. (2015) note that "several data analytics tasks require integrating heterogeneous data sources, where the same or highly related information might be expressed in different forms" (p. 883). This type of integration causes challenges when important relationship linkages cannot be established among heterogeneous data sources and records therefore cannot be linked together, risking duplication of data across all systems (El Hajji, Lebdaoui, & Orhanou, 2013). Other root causes of data integrity issues occur when the data is deliberately altered through willful falsification, document

adulteration, forgery, and the provision of misleading information, including data elimination and enrichment acts (Snee, 2015). Human error, poor measurements, and computer databases errors also cause data integrity issues (Snee, 2015).

One approach to address the data integrity concerns with DI that are caused when heterogeneous data sources must be integrated is syntactic transformation, defined as the translation of input data into an output form, where the schema of both the data models are completely matched and mapped (Abedjan et al., 2015; Cabrera et al., 2018). Syntactic transformation is a data transformation task typically used for mapping data where there is a clearly defined relationship between the two data models, such as converting date format from MMDDYY to MM-DD-YYYY or YYYY-MM-DD, changing city names to their respective country names, or converting liters to gallons (Abedjan et al., 2015). Syntactic transformation works well when the relationships between the data models are clearly defined, but does not work well when the data translation relies on the use of semantic information to associate meaning between the two data models (Abedjan et al., 2015). Semantic information is defined as "the information that a physical system has about its environment that is causally necessary for the system to maintain its own existence over time" (Kolchinsky & Wolpert, 2018, p. 1).

For cases where data must be translated through the use of semantic information to associate two data models, semantic transformation is required (Abedjan et al., 2015). Semantic transformation is defined as a data translation "requiring lookup for meaningful mapping between input and output values in a repository of reference data" (Abedjan et al., 2015, p. 813). A specific type of semantic transformation is semantic ETL, defined as "the framework that uses semantic technologies to integrate and publish data from multiple sources as open linked data" (Bansal, 2014, p. 522). Semantic technologies are "the technical approaches that facilitate or

make use of interpretation of meaning by machines" (Fürber, 2015, p. 8). Fürber (2015) notes that the collection and storage of relevant knowledge in a form that the machines can understand is necessary for the machines to interpret the knowledge.

Semantic transformations, such as company name to stock ticker symbol or event to date, require a search for mappings between the input and output values in reference tables maintained internally or by a third-party supplier (Abedjan et al., 2015). While syntactic transformations are supported by many famous tools such as Microsoft Excel and Google Spreadsheets, less research has been conducted on semantic transformations because they cannot be directly computed by accepting input values and applying a formula or an operation to complete the translation (Abedjan et al., 2015). Abedjan et al. (2015) also note that there is an immense need for tools that can perform some of the basic semantic transformations, such as "US Dollars to EUR, genome to coordinates, location to temperature" (p. 883). Abedjan et al. (2015) note that this type of data changes over time and requires a look up by the users to find the latest values in continuously updated repositories, sometimes making the previously retrieved datasets invalid. Utilizing millions of reference tables that are available online and maintained by third-party providers not only reduces the workload of data professionals in a company, but also keeps the dynamic data current, thereby ensuring high quality DI (Abedjan et al., 2015; Chen et al., 2015).

Purpose Statement

The purpose of this annotated bibliography is to present selected literature that explores various means organizations can use to address data integrity issues caused during data integration. Sources are presented on data integration processes using ETL methodologies, as these are the most common approaches for data integration. Scholarly sources that describe, compare, and contrast syntactic and semantic transformations are also included. The focus of this

research is to explore the best practices in semantic ETL in the context of DI to address data integrity issues.

Research Questions

Main question. What are the best practices to use semantic transformation in data integration to address data integrity issues?

Sub question. How does semantic transformation help in addressing the most common data integrity issues?

Audience

The core audience members for this study are the leaders and managers of Information Technology departments such as chief information officers (CIOs), defined as the position that "oversees people, processes and technologies within a company's IT organization to ensure they deliver outcomes that support the goals of the business" (Gartner, Inc., 2019) and IT directors, who are responsible for "overseeing the infrastructure of technical operations, managing a team of IT employees, tracking technology in order to achieve business goals, eliminating security risks, increasing user satisfaction, and maintaining operations and systems" (Kidd, 2019). These stakeholders have the ability to facilitate the approval of funding and required resources for initiatives such as the implementation of a semantic transformation project.

Other audience members are data professionals in positions such as data analysts, whose duties are to assess data using statistical methods in order to find something out or assist with decision-making ("data analyst," 2019); data architects, who are responsible for managing data as it transmits through the enterprise and coordinating the efforts of the various teams involved (Friedman & James, 2013); or business intelligence analysts, who are responsible for helping companies utilize already existing data to increase their efficiency and maximize profits. They

work with large volumes of data by querying databases effectively in order to create reports and establish trends to generate actionable business insights (Discover Data Science, 2019). Finally, data consumers across organizations who perform activities ranging from data visualization to making strategic decisions based on the quality of data available in a data warehouse will benefit from, and have an interest in, semantic transformations that provide them with reliable data to perform their business functions.

Search Report

Search strategy. My search strategy was quite straightforward. I started off with keywords related to data integration and ETL. Upon reviewing the reference sources in my search results, I was able to narrow down my topic to exploring ways to utilize semantic transformations in data integration to tackle some of the data integrity issues caused by data integration. Since technology evolves more quickly than other areas of science, I applied a filter to explore articles published more recently, with articles published no later than 2014.

Search engines and databases. I used the following databases in the UO Libraries:

- ACM Digital Library,
- IEEE Xplore,
- Academic Search Premier, and
- UO LibrarySearch.

Keywords. I used the following keywords to search for reference sources for my study:

- Data integration,
- ETL,
- Syntactic transformation,
- Semantic transformation,

- Business intelligence,
- Data analytics,
- Semantic ETL,
- Data warehouse,
- Semantic information,
- Data integrity,
- Issues, and
- Duplication.

Documentation method. The first step I took after I identified an article as a result from a search was to skim through the abstract section to gauge the relevance to my research topic. If I approximated the relevant content in the source to be higher than 60%, I saved the article to Zotero via the Zotero add-on in the Google Chrome browser. Zotero has the ability to organize articles by title, creator, and date and assists in creating references in the American Psychological Association (APA) format. I cross-checked all citations generated by Zotero with the publication manual of the American Psychological Association.

I used Microsoft Excel to list the references shortlisted from my abstract screening and rated the relevance of each source on a scale of 1 to 10. I copied references I cited in the paper to Microsoft Word and formatted them to match the APA style citations. I also downloaded the articles in the Adobe portable document format (PDF) and saved these files in Zotero for future reference.

Evaluation criteria. I used the guide provided by the Center for Public Issues Education (n.d.) at the University of Florida as the framework for evaluating references. I considered five major areas before finalizing the references:

- *Authority.* I selected resources that are from peer-reviewed journals, and where the information was available, considered the authors' education and place of employment.
- *Timeliness.* The timeliness of the articles used in this study is of vital significance because of the data integration is an emerging technology. Data integration solutions or strategies which are older than five years may not be suitable to the current integration scenarios. I used sources older than five years only for definitions.
- *Quality.* I assessed the quality of literature for proficiency in grammar and proper structure. I avoided sources demonstrating grammatical inaccuracies for this research study, with the exception of authors whose native language is not English.
- *Relevancy.* I ensured the relevancy of the literature by selecting sources that are strongly related to data integration, data integrity, data warehouses, and data transformation.
- *Bias.* I determined the bias of a source by checking whether the author had any intention to sell a product or service or was looking to persuade the reader to a specific viewpoint; in these cases, I avoided selection of the source.

Annotated Bibliography

Introduction

The following annotated bibliography presents fifteen references that explore best practices in semantic transformation in data integration to address data integrity issues. References are selected to help organizations understand semantic transformations and realize the potential for semantic transformations to maintain data integrity when performing data integration between various sources of data. References are presented in four categories: (a) data integration and resulting data integrity issues, (b) ETL and syntactic and semantic transformations, and (c) uses of data integration. Each annotation consists of three elements: (a) the full bibliographic citation, (b) abstract, and (c) summary.

Data Integration and Resulting Data Integrity Issues

Chen, Q., Hu, H., & Xu, J. (2015). Authenticated online data integration services. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD'15* (pp. 167-181). New York, NY: Association for Computing Machinery. doi: 10.1145/2723372.2747649

Abstract. Data integration involves combining data from multiple sources and providing users with a unified query interface. Data integrity has been a key problem in online data integration. Although a variety of techniques have been proposed to address the data consistency and reliability issues, there is little work on assuring the integrity of integrated data and the correctness of query results. In this paper, we take the first step to propose authenticated data integration services to ensure data and query integrity even in the presence of an untrusted integration server. We develop a novel authentication code called homomorphic secret sharing seal that can aggregate the inputs from individual

sources faithfully by the untrusted server for future query authentication. Based on this, we design two authenticated index structures and authentication schemes for queries on multi-dimensional data. We further study the freshness problem in multisource query authentication and propose several advanced update strategies. Analytical models and empirical results show that our seal design and authentication schemes are efficient and robust under various system settings.

Summary. Three researchers from the department of computer science at Hong Kong Baptist University published this article as part of the Proceedings of the 2015 Association for Computing Machinery (ACM) Special Interest Group on Management of Data (SIGMOD) International Conference on the Management of Data in Melbourne, Australia. The article is mainly about authenticating online data integration services to enable secure passage for data flow between disparate systems that are integrated. This article was chosen because of the solid introduction to data integration and the occurrence of data integrity issue as a result of DI, which is one of the key aspects of this study. According to the authors, “data integration involves combining data from multiple sources and providing users with a unified query interface” (p. 167). The authors state that “all existing work assumes that the data integrator can always be trusted and, hence, address data inconsistency and unreliability issues arising from the data source side” (p. 167). But as of 2015, the date of publication, the authors assert that this assumption was no longer valid for many real-life applications because the integrator might be altered voluntarily or involuntarily, to forge or selectively exclude the integrated data and query results.

Although this article has been referenced for the definitions of data integration and data integrity, the main focus of the paper is about an online authentication methodology called Homomorphic Secret Sharing Seal (HS3), which is "a novel distributed authentication code for integrity verification in the multi-source environment" (p. 170). The HS3 design principle consists of two parts: "(1) individual codes can be gathered collectively by an untrusted integration server (IS); (2) to prove a set of values is not yielded by the query results does not need the authentication codes of all these values, especially those whose values are far away from the query range" (p. 170). Hence, the authentication code of all the values is divided into pieces and only code related to specific queries is used in the authentication process.

El Hajji, S., Lebdaoui, I., & Orhanou, G. (2013). Data integrity in real-time data warehousing. In *Proceedings of the World Congress on Engineering – WCE'13, Vol 3* (pp. 1 - 4). London, UK: World Congress on Engineering. Retrieved from http://www.iaeng.org/publication/WCE2013/WCE2013_pp1516-1519.pdf

Abstract. Information freshness and integrity are the main pillar of making sound decision. Real-time datawarehousing is a trend of delivering fresh information to decision making processes in "Real-time". However, to flow up real-time data into the datawarehouse, systems may skip some necessary treatments. Thus, data integrity may be threatened. This paper discusses data integrity issues towards the need of accessing to Real-time datawarehouses and introduces an IA-RTDWg model that preserves both integrity and availability.

Summary. This article was published by three Moroccan research students in 2013 and was presented at the World Congress on Engineering in London, UK. They explore the

concept of real-time data warehousing (RTDWg), which is "a trend of delivering fresh information to decision-making processes in real-time" (p. 1). The authors describe different types of data, including static data, defined as "data that might receive no important and frequent changes since their first loading into the DW" (p. 1); and dynamic data, which is defined as "living data which change frequently or continuously" (p. 1). The authors note that the difference between real-time and non-real-time data warehouses is that real-time data warehouses (RTDWs) are updated continuously in real-time, whereas non-real-time DWs are updated according to a fixed frequency, which can vary from every six hours to once or twice every day, depending on the business requirements. Finally, the authors provide examples of data integrity issues, including duplication of data and data of poor quality.

In relation to this study, this article provides information about one of the most common data integrity issues, which is duplication of data. The authors note that duplication of data occurs when the data across heterogeneous sources are missing important linkages and related records cannot be linked together, resulting in duplication across all systems. The authors describe data quality as relying upon data integrity, accuracy, conformity, validity, consistency, and completeness. The authors note that data quality is extremely challenging for RTDWg because this type of data warehouse has to deal with the issue of updating the data in a timely manner. The DW needs to be available for query at the same time as the updates are being made, otherwise the results may be inaccurate when the query is performed before the changes can be applied. The authors stress the importance of guaranteed data integrity for avoiding incorrect decisions. The authors conclude by stressing that "when data integrity is first established and implemented through

constraints, it must be respected along all processes involved" (p. 4), which is important for sustaining the data integrity.

El Hajji, S., Lebdaoui, I., & Orhanou, G. (2016). Managing big data integrity. In *2016 International Conference on Engineering & MIS (ICEMIS)*. Agadir, Morocco: ICEMIS. doi: 10.1109/ICEMIS.2016.7745332

Abstract. Big data becomes a real opportunity and a serious worry for data managers, data scientists, researchers and even for business managers. An opportunity as a set of powerful technologies and a bunch of interesting concepts that aim resolving business problems of an organization. A worry because big data implies big challenges and big resources that make harder the task of mining value from all the available data. The data, even in context of big data, need to have a value to help on making good business decisions. In fact, sound decision requires confidence on data that must draw the real world, at any stage of data processing. In other words, the data have to keep their integrity properties from origination until the final destination in analysis reports or in a business indicator.

In this paper, we will throw light on integrity issues for big data and we introduce a new model about how preserving integrity in the context of big data.

Summary. This article was published by three Moroccan research students in 2016 and was presented at the International Conference on Engineering & MIS (ICEMIS) in Agadir, Morocco. They explore the vital significance of data integrity in the context of big data, which is a serious concern for both commercial and sociotechnical researchers. Data warehouses and big data are an essential mines of information for making business

decisions and deserve special attention in every stage of their treatment, especially when they have important value.

The authors assert that data integration is extremely useful in enabling the flow of data from heterogeneous sources into a central repository. At first, the data is captured and extracted, then integrated or aggregated prior to any analyses or interpretation. According to the authors, integrity is an essential pillar which interacts with notions of *trust*, *fitness for use*, and *consensual understanding*. Integrity challenges can occur at the beginning of data processing or at other times during the actual processing.

The authors introduce two concepts to support data integrity: end-point filtering and real-time security monitoring. Based on the evidence that integrity concerns begin from the time the data starts to exist, which is usually the first input of data, input validation is necessary to purify and validate data for the next steps of data processing. It is important to receive protected data with a great level of integrity assurance. Since the data must go through many stages of processing and manipulations, the integrity must be well-maintained at each point of the process. Real-time monitoring serves to monitor the data infrastructure while using the same infrastructure for data analytics.

This article was chosen because it elaborates on data integrity issues and provides a model for managing data integrity. This article is useful for this study because it provides insight into how and when data integrity issues arise when there is a flow of data from various sources.

McDowall, R. D. (2019). Data integrity and data governance. *What is data integrity?* (pp. 12-13). Croydon, UK: CPI Group (UK) Ltd.

Cover Summary. The aim of this book is to provide practical and detailed advice on how to implement data integrity and data governance for regulated analytical laboratories working in the pharmaceutical and allied industries. Although the main thrust of the book is for chemical laboratories, some microbiological analysis is also discussed. This book is written for analytical scientists, laboratory managers and supervisors and quality assurance personnel working in regulated laboratories in and for the pharmaceutical industry who are involved with data integrity and data governance programs. Where networked systems are discussed, IT professionals may also find useful information.

Summary. This book is mainly published for pharmaceutical and allied industries, but the core concepts are closely related to data integrity and data governance, which are applicable to any industry dealing with data and analytics. The book is chosen for this study because it has clear and concise definitions for data integrity based on various sources and also introduces a framework called ALCOA+ criteria for maintaining the integrity of laboratory data. ALCOA+ stands for:

(a) *Attributable*, which means information is captured in the record so that it is uniquely identified as executed by the originator of the data,

(b) *Legible*, also known as traceable or permanent, refers to the requirements that data are readable, understandable, and allow a clear picture of the sequencing of steps or events in the record so that all activities conducted can be reconstructed by the people reviewing these records at any point during the records retention period,

(c) *Contemporaneous* data are data recorded at the time they are generated or observed,

(d) *Original* data include the first or source capture of data or information and all subsequent data required to fully reconstruct the conduct of the activity,

(e) *Accurate* means data are correct, truthful, complete, valid and reliable, and

(f) *Complete* requires the inclusion of all data from an analysis, including any data generated before a problem is observed and data generated after repeating part or all of the work or reanalysis on the sample. (p. 13)

This source is useful for this study because it provides a framework for the attributes that are important when maintaining the integrity of a specific type of data; the lessons can be applied to other data sources as well.

Snee, R. D. (2015). Data integrity – How to detect lack of integrity. *Institute of Validation Technology – IVT Validation Week'15*. Philadelphia, PA: IVT Validation Week. Retrieved from https://www.researchgate.net/publication/282879414_Data_Integrity_Validation

Abstract. Data are central to the development, manufacture and marketing of pharmaceuticals of all types. The renewed interest in data integrity raises questions regarding what is data integrity and how to assess it. Lack of data integrity comes in two forms: purposeful manipulation of the data to deceive and the inadvertent problems that occur in the production and analysis of data. Humans, equipment or both can be the source of the problem. This session discusses on both types on data integrity and introduces the assessment of "data pedigree" as a concept that puts focus on the types of data integrity issues and analytical and statistical methods for detecting data problems. Pharma and biotech case studies are used throughout the presentation to illustrate how the various approaches together.

Summary. Snee presented this article in 2015 at the Institute of Validation Technology (IVT) week in Philadelphia, PA. He provided a glance into data realities in 2015 in terms of data integrity, the severity of data integrity issues, and the causes of data integrity issues in an enterprise. The author outlines certain deviant human behaviors that lead to a lack of data integrity, including: (a) not recording activities contemporaneously, which results in incomplete datasets leading to inaccuracies in reports used for decision-making, ultimately effecting important business decisions; (b) backdating; (c) fabricating data; (d) copying existing data as new data; (e) rerunning samples, which leads to major effects on conclusions and actions, jeopardizing the company's future in some cases; and (f) discarding data.

The author also provides step-by-step procedures to detect a lack of data integrity in organizations; these steps are: (a) failure modes and effect analysis, which is used to identify where and how data issues occur by following the lifecycle of the data, (b) understanding the data pedigree involving science and engineering data; the structure of the process, product, or service from which the data were collected; the collection process used to obtain the data; and how the measurements were made; (c) identifying where human intervention can occur; and (d) understanding cultural and behavioral issues. Overall, this paper is very helpful in understanding the vital significance of data integrity and appropriate measures to maintain data integrity.

ETL and Syntactic and Semantic Transformations

Abedjan, Z., Ilyas, I. F., Morcos, J., Ouzzani, M., Papotti, P., & Stonebraker, M. (2015). DataXFormer: An interactive data transformation tool. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD'15* (pp. 883-

888). New York, NY: Association for Computing Machinery. doi: 10.1145/2723372.2735366

Abstract. While syntactic transformations require the application of a formula on the input values, such as unit conversion or date format conversions, semantic transformations, such as zip code to city, require a look-up in some reference data. We recently presented DataXFormer, a system that leverages Web tables, Web forms, and expert sourcing to cover a wide range of transformations. In this demonstration, we present the user-interaction with DataXFormer and show scenarios on how it can be used to transform data and explore the effectiveness and efficiency of several approaches for transformation discovery, leveraging about 112 million tables and online sources.

Summary. This article was presented in 2015 at the annual ACM SIGMOD/PODS conference. The six authors are from reputable universities in North America and the Middle East. The authors introduce an interactive data transformation tool named DataXFormer.

The authors provide details about the most common type of data transformation, syntactic transformation, along with some examples of syntactic transformations, such as converting date format from MMDDYY to MM-DD-YYYY or YYYY-MM-DD, changing city names to their respective country names, or applying formula to convert liters to gallons. The authors also note the limitations of syntactic transformation, including the inability to deal with transformations such as company name to a stock ticker symbol or event to a date, which require a search of the repositories of reference data that contain mappings between the input and output values.

The authors introduce the concept of semantic transformation as the methodology to address the limitations of syntactic transformation by stating that semantic transformation makes use of around 112 million reference tables and online resources to look up meaningful information and transform the data accordingly, leading to more current and less obsolete data, resulting in improved data integrity.

The authors present the DataXFormer tool, which is completely based on semantic transformation and mainly uses two types of subsystems, web tables and web forms. Web table and web form subsystems work off of two separate transformation engines, which are based on locally stored static web tables and dynamic web forms. Whenever a user submits a query, DataXFormer converts the user query into the corresponding internal forms for the retrieval components; candidate web tables and web forms are then returned. DataXFormer sends the transformation query to both transformation engines. "Solution integration is one of the components that receives results from both subsystems and presents the best effort results from both sources for a given query" (p. 884). The authors also include examples demonstrating sematic transformations by DataXFormer, including currency conversion, genome to coordinates, and place to temperature, which change over time.

This article is useful for this study because the semantic transformations specified in this article are very helpful in handling different facets of data integration that syntactic transformations are not equipped to handle, thereby addressing the lack of data integrity.

Bansal, S. K. (2014). Towards a semantic extract-transform-load (ETL) framework for big data integration. In Kesselman, C. (Ed.), *2014 IEEE International Congress on Big Data* (pp.

522-529). New York, NY: IEEE Computer Society. doi: 10.1109/BigData.Congress.2014.82

Abstract. Big Data has become the new ubiquitous term used to describe massive collection of datasets that are difficult to process using traditional database and software techniques. Most of this data is inaccessible to users, as we need technology and tools to find, transform, analyze, and visualize data in order to make it consumable for decision-making. One aspect of Big Data research is dealing with the Variety of data that includes various formats such as structured, numeric, unstructured text data, email, video, audio, stock ticker, etc. Managing, merging, and governing a variety of data is the focus of this paper. This paper proposes a semantic Extract-Transform-Load (ETL) framework that uses semantic technologies to integrate and publish data from multiple sources as open linked data. This includes - creation of a semantic data model to provide a basis for integration and understanding of knowledge from multiple sources; creation of a distributed Web of data using Resource Description Framework (RDF) as the graph data model; extraction of useful knowledge and information from the combined data using SPARQL as the semantic query language.

Summary. Srividya Bansal is the author of the article; she holds a Ph. D. and is an associate professor of software engineering at Arizona State University. The author explores various contextual data concepts such as Big Data, data integration, and semantic technologies. Bansal defines Big Data as “the new ubiquitous term used to describe massive collections of datasets so large they cannot easily be processed using traditional database and software techniques” (p. 522). The author also notes that semantic data integration still poses a challenge for Big Data. Bansal also asserts that data

needs to be organized so that similar items are grouped together and distinct items are grouped separately to enhance the use of analytics for SETL frameworks and Big Data. Bansal describes the data integration approach of using semantic technologies to "enhance definitions of ETL activities involved in the process rather than the data itself" (p. 523). In general, semantic mappings are often used in the extraction phase, but the author describes an approach that relies on ontologies, which are useful in maintaining a common terminology and vocabulary for the integrated data. This approach results in the generation of semantic data as the output in the transformation phase. The author also notes an important difference between traditional ETL and semantic ETL frameworks, that occurs in the Transform phase. With the approach using ontologies, the output generated by semantic ETL is the semantic linked data, which is then stored in a data warehouse or a designated repository; this step is not present in the traditional ETL Transform phase.

This article is useful for this study because it focuses on a proper introduction of semantic ETL and the need for it in data integration.

Bansal, S. K., Charkraborty, J., & Padki, A. (2017). Semantic ETL – state-of-the-art and open research challenges. In *2017 IEEE 11th International Conference on Semantic Computing* (pp. 413-418). New York, NY: IEEE Computer Society. doi: 10.1109/ICSC.2017.94

Abstract. There has been an exponential growth and availability of data, both structured and unstructured. Massive amounts of data are available to be harvested for competitive business advantage, sound government policies, and new insights in a broad array of applications (including healthcare, biomedicine, energy, smart cities, genomics, transportation, etc.). Yet, most of this data is inaccessible for users, as we need

technology and tools to find, transform, analyze, and visualize data in order to make it consumable for decision-making. Meaningful data integration in a schema-less, and complex Big Data world of databases is a big open challenge. This survey paper presents a holistic view of literature in data integration and Extract-Transform-Load (ETL) techniques. Limitations and gaps in existing approaches are identified and open research challenges are discussed.

Summary. The authors provide information about the traditional ETL framework, including its capabilities and limitations. They define ETL by elaborating all three components separately: (a) *Extract* is the first phase of the ETL process that mainly deals with data retrieval from various data sources, depending on the requirements; (b) the *Transform* phase involves data cleansing activities to make sure the source data complies with the target schema; and (c) *Load* involves the flow of data into the destination such as a data warehouse or a data mart that serves Big Data.

They also analyzed several traditional open-source ETL tools such as Clover, Talend, and Pentaho by comparing their engine designs, transformation graphs, data support, and semantic web support. Key findings are Clover's engine design has each component run in a separate thread and acts as a consumer or producer, whereas in Talend's engine design, all components are run on a single thread unless a multi-threaded environment is enabled. Pentaho's engine design is multi-threaded with a meta-data driven approach. The transformation graphs are represented as XML files which can be dynamically generated in Clover; Talend's data transformation scripts are generated by its open studio, which acts like a code generator; and Pentaho's transformation graphs are saved as XML files. Data support is absent in Clover but present in both Talend and Pentaho.

Semantic web support via custom standalone plugins is available for Talend but not for Clover or Pentaho.

The authors devote a section to semantic ETL, including semantic data integration techniques, semantic ETL techniques, and the limitations posed by ETL tools in general. Semantic data integration techniques they address include the *Datalift* platform, which is one of the earlier works attempting a Semantic ETL framework. The process falls under a non-traditional ETL framework where there is lifting of raw data sources to semantic interlinked sources. The user provides structured data sources as an input and to maintain a uniform format, all data sources are converted into a raw resource description framework (RDF) format. After a unique data format is ready, vocabularies are selected to assign meaning to the lifted data, after which the ontology is prepared. This ontology is used to map the elements in the RDF file. The final step in the process aims to provide links from the newly published dataset to the datasets which are already published on the Web. These techniques are effective in addressing limitations posed by the traditional ETL tools because traditional ETL tools only work when there is a direct schema match between source and destination and does not associate meaning to the data. The authors address semantic ETL techniques including "python-based programmable semantic ETL framework (SETL). SETL builds on semantic Web (SW) standards and tools and supports developers by offering a number of powerful modules, classes and methods for data warehouse constructs and tasks" (p. 417). Hence, SETL supports semantic integration based on semantic-aware data and semantic DW that are comprised of an ontology and its instances.

The authors conclude by noting open challenges that require further innovation to address; these challenges include ETL job automation, configurations requiring human intervention, lack of support for query optimization, limited ability to extract data from multiple sources in parallel, and the fact that very few relational joins are supported in standard ETL tools.

This article is useful for this study because it provides information about semantic ETL techniques, which is the central focus of this study.

Cabrera, A. M., Cepeda, K., Chamberlain, R.D., Cytron, R. K., Derber, R., Epstein, C., … Zheng, J. (2018). DIBS: A data integration benchmark suite. In *Companion of the 2018 ACM/SPEC International Conference on Performance Engineering - ICPE '18* (pp. 25-28). New York, NY: Association for Computing Machinery. doi: 10.1145/3185768.3186307

Abstract. As the generation of data becomes more prolific, the amount of time and resources necessary to perform analyses on these data increases. What is less understood, however, is the data preprocessing steps that must be applied before any meaningful analysis can begin. This problem of taking data in some initial form and transforming it into a desired one is known as data integration. Here, we introduce the Data Integration Benchmarking Suite (DIBS), a suite of applications that are representative of data integration workloads across many disciplines. We apply a comprehensive characterization to these applications to better understand the general behavior of data integration tasks. As a result of our benchmark suite and characterization methods, we offer insight regarding data integration tasks that will guide other researchers designing solutions in this area.

Summary. Eight researchers at Washington University at St. Louis have introduced a suite of applications called the data integration benchmark suite (DIBS) that are representative of data integration workloads across many disciplines. The authors provide ways to organize data integration applications by placing them into three task categories: parsing/cleaning, transformation, and aggregation. Parsing/cleaning involves computation associated with identifying the records, fields, and other components of the input data, along with checks to see if the data is well-formed. Transformation involves translation of input data into the form expected by the primary computation. Aggregation involves any pre-analytics computations that result in aggregate information about the input.

As the main focus of the Capstone paper is data integration, the article provides insight into what is involved in achieving integration between disparate systems across various domains such as enterprise, Internet of Things (IoT), graph processing, image processing, and computational biology. Key findings are: (a) for enterprise domain, the parsing/cleaning task involves adjusting non-ASCII characters, transformation involves changing input into either text or comma delimiter (csv) format, aggregation involves counting the number of elements, and (b) for the IoT domain, parsing/cleaning requires tokenizing input, transformation involves converting input into csv format, and aggregation involves a running total of file size. The researchers also explore the general qualities and idiosyncrasies of the DIBS tool by applying a comprehensive and architecturally-independent characterization to each application, which results in data integration tasks having a consistent level of both spatial and temporal locality, usually exhibiting higher spatial locality. The applications also exhibit high degrees of control flow regularity and data movement.

The purpose of this article is to demonstrate that the insights gained from their characterizations can guide both software and hardware research in exploring and exploiting the qualities of data integration tasks to improve performance.

Erturkmen, G. B. L., Gonul, S., Pacaci, A., Sinaci, A. A., & Yuksel, M. (2018). A semantic transformation methodology for the secondary use of observational healthcare data in postmarketing safety studies. *Frontiers in Pharmacology,* 9, 435. doi: 10.3389/fphar.2018.00435.

Abstract. Utilization of the available observational healthcare datasets is key to complement and strengthen the postmarketing safety studies. Use of common data models (CDM) is the predominant approach in order to enable large scale systematic analyses on disparate data models and vocabularies. Current CDM transformation practices depend on proprietarily developed Extract—Transform—Load (ETL) procedures, which require knowledge both on the semantics and technical characteristics of the source datasets and target CDM. In this study, our aim is to develop a modular but coordinated transformation approach in order to separate semantic and technical steps of transformation processes, which do not have a strict separation in traditional ETL approaches. Such an approach would discretize the operations to extract data from source electronic health record systems, alignment of the source, and target models on the semantic level and the operations to populate target common data repositories. In order to separate the activities that are required to transform heterogeneous data sources to a target CDM, we introduce a semantic transformation approach composed of three steps: (1) transformation of source datasets to Resource Description Framework (RDF) format, (2) application of semantic conversion rules to get the data as instances of ontological

model of the target CDM, and (3) population of repositories, which comply with the specifications of the CDM, by processing the RDF instances from step 2. The proposed approach has been implemented on real healthcare settings where Observational Medical Outcomes Partnership (OMOP) CDM has been chosen as the common data model and a comprehensive comparative analysis between the native and transformed data has been conducted. Health records of ~1 million patients have been successfully transformed to an OMOP CDM based database from the source database. Descriptive statistics obtained from the source and target databases present analogous and consistent results. Our method goes beyond the traditional ETL approaches by being more declarative and rigorous. Declarative because the use of RDF based mapping rules makes each mapping more transparent and understandable to humans while retaining logic-based computability. Rigorous because the mappings would be based on computer readable semantics which are amenable to validation through logic-based inference methods.

Summary. The authors discuss a semantic transformation methodology in the healthcare sector focused on post-marketing safety studies. The researchers provide a three step approach: (a) conversion of source data into a Resource Description Framework (RDF) format, (b) implementation of semantic conversion rules on the data in RDF format to match the common data model (CDM) specifications, and (c) loading the data resulting from steps a and b into the destination repositories. The approach defined by the authors was tested in a real healthcare setting, where the CDM was from an observational medical outcomes partnership (OMOP).

Electronic Health Records (EHRs) contain patient medical history and complete information about a patient's risk factors. In order to Retrieve EHR data in RDF format,

two main EHR databases were used: LISPA, which is a regional data warehouse in Italy, and TUD, an EHR database at a university hospital in Germany. In the LISPA system, the medical data is either in one of two formats: (a) Health Line 7 (HL7)/ASTM Continuity of Care Document (CCD) or (b) IHE Patient Care Coordination (PCC). A tool called *Ontmalizer* converts these templates into extensible markup language (XML) format. TUD does not require any prior transformations because it is already in the RDF format. Data transformed into the RDF format can be parsed as input to the semantic framework where heterogeneous RDF data are transformed into a common format via semantic mappings to perform a set of analytic routines on the source data.

Once LISPA and TUD are extracted to the RDF format, they are transformed into OMOP CDM format. The ontology of OMOP CDM is represented in Web Ontology Language (OWL) by the researchers because semantic conversion rules are developed using the RDF technology. For each construct in the OMOP CDM, an OWL construct is created with the following mappings: (a) the entity in the ER construct is mapped to Class in OWL, (b) the attribute in the ER construct is mapped to DatatypeProperty in OWL, and (c) the relationship in the ER construct is mapped to ObjectProperty. The mapping rules are evaluated using Notation 3 (N3) logic; the authors note that Euler Yet another proof Engine (EYE) can be used to execute mappings in an N3 reasoner.

Through filtering rules, the semantic rule-based approach makes the transformation process easier to validate compared to the traditional ETL approach. Individual rules that are defined at the concept code level can be extended and combined in a bottom-up manner to define mapping rules for more complex entities such as conditions or persons. In addition, these building blocks can be re-used across various versions and models.

This approach not only makes the transformation process easier to maintain compared to the traditional ETL approach, it also saves the experts from re-developing entire sets of transformation rules for every data model. In the process of exploiting the modularity characteristic, the authors developed unit tests defined as a CONSTRUCT query, which attempts to re-generate all the expected data elements from the outcome described above. These tests are very helpful in verifying and proving the transformation process.

With the EYE reasoning engine, logic-based proofs are generated to log data extractions or any interferences that help in reaching any conclusions in the process. This is helpful in building trust for the reasoning process and maintaining high data integrity.

This article is useful because it provides a proven semantic transformation methodology in the healthcare industry in great detail.

Fürber, C. (2015). Data quality management with semantic technologies. In *Research design* (pp. 8–19). Wiesbaden, Germany: Springer Fachmedien. https://doi.org/10.1007/978-3-658-12225-6_2

Introduction. In this chapter, we provide a brief introduction into the thesis topic, clarify our understanding of the term “data” and its dependency to business processes and decisions, and discuss the economic relevance of the systematic management of data quality. Moreover, we give a short overview of the thesis structure.

Summary. The Research Design chapter in this book provides information about semantic technologies and history related to these technologies. The author states that “in computer science, the term “semantic” has been used in the context of programming languages since the 1960s. With the advent of artificial intelligence as a field, the notion

of semantics in computer science got broader, including the representation of terminological and factual knowledge by data structures" (Fürber, 2015, p. 8).

The area of focus of this annotated bibliography is a specific type of semantic transformation: semantic ETL. According to Fürber, semantic technologies are "the technical approaches that facilitate or make use of interpretation of meaning by machines" (p. 8). This chapter is used to provide history related to semantic technologies and to obtain the definition of semantic technologies so that semantic ETL can be better understood by the audience.

Hose, K., Nath, R. P. D., & Pedersen, T. B. (2015). Towards a programmable semantic extract-transform-load framework for semantic data warehouses. In *Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP - DOLAP'15* (pp. 15-24). New York, NY: Association for Computing Machinery. doi: 10.1145/2811222/2811229

Abstract. In order to create better decisions for business analytics, organizations increasingly use external data, structured, semistructured and unstructured, in addition to the (mostly structured) internal data. Current Extract-Transform-Load (ETL) tools are not suitable for this "open world scenario" because they do not consider semantic issues in the integration process. Also, current ETL tools neither support processing semantic-aware data nor create a Semantic Data Warehouse (DW) as a semantic repository of semantically integrated data. This paper describes SETL: a (Pythonbased) programmable Semantic ETL framework. SETL builds on Semantic Web (SW) standards and tools and supports developers by offering a number of powerful modules, classes and methods for (dimensional and semantic) DW constructs and tasks. Thus it supports semantic-aware

data sources, semantic integration, and creating a semantic DW, composed of an ontology and its instances. A comprehensive experimental evaluation comparing SETL to a solution made with traditional tools (requiring much more handcoding) on a concrete use case, shows that SETL provides better performance, knowledge base quality and programmer productivity.

Summary. Danish researchers Hose, Nath, and Pedersen developed a paper to elaborate on a programmable semantic ETL (SETL) framework for semantic data warehouses. The authors focus on specific limitations of traditional ETL tools available in 2015, the time of publication, including their inability to process meaningful relationships within the external data because they:

(1) do not support semantic-aware data,

(2) are entirely schema-dependent,

(3) do not focus on meaningful semantic relationships to integrate data from disparate sources, and

(4) do not support the deriving new information by active inference and reasoning on the data. (p. 15).

The authors make a case for the increased need for semantic-aware data warehouses for better performance, knowledge base quality, and programmer productivity; they define a semantic-aware data warehouse as “the semantic repository of semantically integrated data” (p. 15).

The authors provide a use case considering a Danish agricultural dataset and a business dataset; these datasets are integrated using the SETL framework, with the extraction and traditional transformation phases taking a relatively long time because they involve two

spatial join operations for the agriculture datasets. Semantic transformations are generally time-intensive because the source data needs to be processed and matched to the schema of the target ontology. The total time required for the linking process is approximately 3,979 seconds and results in the linking of the SETL knowledge base (SETLKB) to 14,153 external resources.

The authors used three loading methods, *TrickleLoad*, *BulkTDBLoader*, and *BulkSPARQLInsert*, to test the performance; *TrickleLoad*'s performance was slightly better than *BulkSPARQLInsert*'s performance in all cases, while *BulkTDBLoader* was 3.5 times slower when the batch file was small but for a batch size of 1 million to 32 million, its performance became two times faster than the other two methods.

This article is useful for this study because it provides technical aspects of Semantic ETL and how it improves performance when used in data integration.

Voegeli, D. (2018). Introduction. In *ETL from RDF to property graph* (pp. ix-x). Bedford, MA: The MITRE Corporation. Retrieved from https://www.mitre.org/sites/default/files/publications/pr-15-2949-ETL-from%20-RDF-to-property-graph_0.pdf

Abstract. Within this introduction is a brief description of the extract, transform, and load process, as well as a general description of both the resource description framework and the property graph concepts. Concluding this section is a side-by-side comparison between the resource description framework and the property graph databases. This information intends to give the reader a sense of how the resource description framework and the property graph databases relate in respect to the extract, transform, and load process.

Summary. The author provides an introduction to ETL, where: (a) *Extract* is the first phase of the ETL process that mainly deals with data retrieval from various data sources depending on the requirements, (b) the *Transform* phase involves data cleansing activities to make sure the source data complies with the target schema, and (c) *Load* involves the flow of data into the destination such as a data warehouse or a data mart. The author introduces a very important term in the world of semantic ETL, Resource Description Framework (RDF), defined as "the framework which specifies a language for defining data items and relationships, by using a graph representation, with the intent to scale up and work across the entire internet" (p. x).

The World Wide Web Consortium (W3C) is a standardization committee for web technologies, which maintains the RDF specification, along with numerous other standards such as: (a) SPARQL – a query language used to query data from RDF and other datasets; (b) schema standards such as Resource Description Framework Schema (RDFS), which "provides the ability to organize data items as sets and to define relationships between those sets" (p. ix); and (c) Web Ontology Language (OWL), which "intends to represent rich and complex knowledge about things, groups, group of things, and relation between things" (p. x).

Both RDFS and OWL add formal definitions which can then be utilized by the inference engines for reasoning. For example, if a reasoning engine facilitated by either RDFS or OWL has two assertions, "A veterinarian examines Coco" and "Animals are checked by veterinarians," but "Coco is an animal" is not specified, then the query "Show me all of the animals" will return "Coco" when an inference engine is used. RDF breaks down knowledge into statements and a statement combines a resource, a property, and a

property value, which is also called a subject-predicate-object assertion. Ultimately, RDF describes subjects in relation to their objects.

For this Capstone paper, RDF is a very important piece of semantic transformation. This article provides the information necessary to describe how RDF helps with the unification of data from various sources and the relationship between a subject in relation to its object, thereby increasing the speed at which information can be retrieved.

Uses of Data Integration

Chen, G., Chen, H., & Lim, E. (2013). Business intelligence and analytics: Research directories. *ACM Transactions on Management Information Systems*, *3*(4), 1–10. https://doi.org/10.1145/2407740.2407741

Abstract. Business intelligence and analytics (BIA) is about the development of technologies, systems, practices, and applications to analyze critical business data so as to gain new insights about business and markets. The new insights can be used for improving products and services, achieving better operational efficiency, and fostering customer relationships. In this article, we will categorize BIA research activities into three broad research directions: (a) big data analytics, (b) text analytics, and (c) network analytics. The article aims to review the state-of-the-art techniques and models and to summarize their use in BIA applications. For each research direction, we will also determine a few important questions to be addressed in future research.

Summary. This article is a result of a collaboration between three researchers from three different universities across the globe and was supported by the National Research Foundation under its International Research Centre in Singapore. The main focus of this paper is to introduce the concept of business intelligence and analytics and demonstrate

their roles in decision-making in the business world. The authors define business intelligence as “an umbrella term to describe concepts and methods to improve business decision making by using fact-based support systems. BI also includes the underlying architectures, tools, databases, applications, and methodologies” (p. 1).

The authors also provide an overview of emerging industry, data, and platform technology trends; key trends include:

- Industry trends include a continuing increase in BI revenue. According to Chen, Chen, and Lim note that businesses are leveraging business intelligence initiatives to gain insights from the growing volumes of transaction, product, inventory, customer, competitor, and industry data generated by enterprise-wide applications. Common enterprise-wide applications that supply the data include enterprise resource planning (ERP), customer relationship management (CRM), supply chain management (SCM), and knowledge management applications.
- Data trends include an enormous increase in mobile, social media, web, and data generated by sensors. The data is generated at terabyte and even petabyte scale and businesses have been accumulating and processing such information for years via relational database management systems (RDBMSs). In some large corporations, there has also been an exponential increase in unstructured content and user log information from e-commerce websites.
- Platform technology trends include the expansion of cloud computing and mobile computing.

This article is useful for this study because it provides a key definition of business intelligence, and trends in industry, data, and platform technology trends within the field of business intelligence.

Kirk, A. (2012). Data visualization: a successful design process. In *The context of data visualization: Defining data visualization*. Birmingham, UK: Packt Publishing Ltd.

Cover Summary. Welcome to the craft of data visualization-a multidisciplinary recipe of art, science, math, technology, and many other interesting ingredients. Not too long ago we might have associated charting or graphing data as a specialist or fringe activity-it was something that scientists, engineers, and statisticians did. Nowadays, the analysis and presentation of data is a mainstream pursuit. Yet, very few of us have been taught how to do these types of tasks well. Taste and instinct normally prove to be reliable guiding principles, but they aren't sufficient alone to effectively and efficiently navigate through all the different challenges we face and the choices we have to make. This book offers a handy strategy guide to help you approach your data visualization work with greater know-how and increased confidence. It is a practical book structured around a proven methodology that will equip you with the knowledge, skills, and resources required to make sense of data, to find stories, and to tell stories from your data. It will provide you with a comprehensive framework of concerns, presenting, step-by-step all the things you have to think about, advising you when to think about them and guiding you through how to decide what to do about them. Once you have worked through this book, you will be able to tackle any project-big, small, simple, complex, individual, collaborative, one-off, or regular-with an assurance that you have all the tactics and guidance needed to deliver the best results possible.

Summary. This book has been chosen because there is a chapter that provides an apt definition of data visualization. The author defines data visualization as "the representation and presentation of data that exploits our visual perception abilities in order to amplify cognition" (p. 16). Data is represented based on how one decides to depict it by choosing physical forms. According to Kirk, the presentation of the data "goes beyond the representation of data and concerns how data representation is integrated into the overall communicated work, exploiting the visual perception abilities related to the scientific understanding of how human eyes and brains process information more effectively" (p. 17).

In order to explain the concept of data visualization, the author considers three main agents involved in the transaction: the messenger, the receiver, and the message. On the source side there is a designer who acts as a messenger attempting to impart results, analysis, and stories. On the destination side there are receivers of the message who read the messenger's message. The message in between is the communication channel. The task of the designers is to look through the readers' perspective and attempt to imagine and determine what the reader is going to be learning from the message. According to Kirk, "this type of appreciation shapes the best practices in visualization design: considering and respecting the needs of the reader" (p. 17). In the context of this paper, data visualization occurs after semantic ETL has been applied to integrate data from multiple sources and transport it to a data warehouse. Data visualization activities will be performed on the data warehouse to provide decision makers with visual representations of various metrics so that important strategic decisions can be made.

Conclusion

This annotated bibliography explores the topic of data integration between disparate systems and the lack of data integrity that sometimes results. Using comprehensive themes from scholarly articles and publications from scholarly sources and various technological conferences across the globe has enabled the identification of the emerging trends in the field of data integration. The themes explored in this annotated bibliography are: (a) data integration and resulting data integrity issues, (b) ETL and syntactic and semantic transformations, and (c) uses of data integration.

Data Integration and Resulting Data Integrity Issues

According to Chen et al. (2015), "data integration involves combining data from multiple sources and providing users with a unified query interface" (p. 167). Data integration is extremely useful in enabling the flow of data from heterogeneous sources into a central repository (El Hajji et al., 2016). Two key components of data integration are data warehouses and big data (Bansal, 2014; Hose et al., 2015). Data warehouses, repositories used to store large amounts of data from various operational databases, are often used to store the data that results from data integration (Hose et al., 2015). Big data refers to very large collections of datasets that pose special challenges in data integration due to the massive size of the datasets and the heterogeneity of the data (Bansal, S., 2014; El Hajji et al., 2016). Data warehouses and big data are essential mines of information for making business decisions and deserve special attention in every stage of their treatment, especially when they have important value (El Hajji et al., 2016).

Data integrity, defined as "the degree to which a collection of data are complete, consistent and accurate" (McDowall, 2019, p. 11), is an essential pillar of data integration tied to the concepts of *trust*, *fitness for use*, and *consensual understanding* (El Hajji et al., 2016). Data

integrity is high when the integrated data is attributable to the original source, legible, contemporaneous, original, accurate, and complete (McDowall, 2019). Lack of data integrity has been a key issue with data integration; Chen et al. (2015) note the lack of effort put into ensuring the integrity of integrated data and the issues with the accuracy of the results of queries performed on this data. One common root cause of data integrity issues with data integration occurs when heterogeneous data sources must be integrated (Abedjan et al., 2015). This type of integration causes challenges when important relationship linkages cannot be established among heterogeneous data sources and records therefore cannot be linked together, risking duplication of data across all systems (El Hajji et al., 2013). Snee (2015) outlines certain deviant human behaviors that lead to a lack of data integrity, including: (a) not recording activities contemporaneously, which results in incomplete datasets leading to inaccuracies in reports used for decision-making, ultimately effecting important business decisions; (b) backdating; (c) fabricating data; (d) copying existing data as new data; (e) rerunning samples, which leads to major effects on conclusions and actions, jeopardizing the company's future in some cases; and (f) discarding data.

According to El Hajji et al. (2013), in order to sustain data integrity, "when data integrity is first established and implemented through constraints, it must be respected along all processes involved" (p. 4). El Hajji et al. (2016) introduce two concepts to support data integrity: end-point filtering and real-time security monitoring. Based on the evidence that integrity concerns begin from the time the data starts to exist, which is usually the first input of data, input validation is necessary to purify and validate data for the next steps of data processing (El Hajji et al., 2016). El Hajji et al. (2016) note the importance of receiving protected data with a great level of integrity assurance. Since the data must go through many stages of processing and

manipulations, the integrity must be well-maintained at each point of the process (El Hajji et al., 2016). Real-time monitoring serves to monitor the data infrastructure while using the same infrastructure for data analytics (El Hajji et al., 2016).

ETL and Syntactic and Semantic Transformations

The process of integrating data between applications can be divided into three task categories: parsing/cleaning, transformation, and aggregation (Cabrera et al., 2018). Parsing/cleaning involves computation associated with identifying the records, fields, and other components of the input data, along with checks to see if the data is well-formed (Cabrera et al., 2018). Transformation involves translation of input data into the form expected by the primary computation. Aggregation involves any pre-analytics computations that result in aggregate information about the input (Cabrera et al., 2018). Collectively, the above steps are also known as execute-transform-load (ETL), where *extract* is the first phase of the ETL process that mainly deals with data retrieval from various data sources, depending on the requirements; the *transform* phase involves data cleansing activities to make sure the source data complies with the target schema; and l*oad* involves the flow of data into the destination such as a data warehouse or a data mart (Bansal et al., 2017). The ETL processes facilitate the extraction of data from multiple sources and load them into a central repository such as a data warehouse, resulting in the integration of data (Hose et al. 2015).

The most common type of data transformation is syntactic transformation, which is often used to perform transformations such as changing date format from MMDDYY to MM-DD-YYYY or YYYY-MM-DD, changing city names to their respective countries, or applying a formula to convert liters to gallons (Abedjan et al., 2015). Syntactic transformation does not work for transformations that require searching for reference data that contain mappings between

the input and output values, such as company name to stock ticker symbol or event to date (Abedjan et al., 2015). In order to address the limitations of syntactic transformation, *semantic transformation* was developed (Abedjan et al., 2015). According to Fürber (2015):

> the term *semantic* has been used in the context of programming languages since the 1960s; with the advent of artificial intelligence as a field, the notion of semantics in computer science got broader, including the representation of terminological and factual knowledge by data structures. (p. 8)

Semantic transformation makes use of over 100 million reference tables and online resources to look up meaningful information and transform the data accordingly, leading to more current and less obsolete data, ultimately resulting in improved data integrity (Abedjan et al., 2015).

Hose et al. (2015) focused on specific limitations of traditional ETL tools available in 2015, the time of publication of their article, including the inability to process meaningful relationships within the external data because they:

> (1) do not support semantic-aware data,
>
> (2) are entirely schema-dependent,
>
> (3) do not focus on meaningful semantic relationships to integrate data from disparate sources, and
>
> (4) do not support the deriving new information by active inference and reasoning on the data. (p. 15).

Semantic ETL (SETL) is the approach of using "semantic technologies to enhance definitions of ETL activities involved in the data integration process rather than the data itself" (Bansal, 2014, p. 523). According to Bansal (2014), semantics are used to enable the extraction process workflow to be generated with semantic mappings. Bansal (2014) used a different approach of

adopting ontologies to provide a common vocabulary for the integrated data, generating semantic data as part of the transformation phase of ETL. The common vocabulary helps in maintaining uniformity, which in turn helps to avoid a certain degree of duplication of the data in the whole process, thereby sustaining data integrity to an extent (Bansal, 2014).

Two of the articles in this annotated bibliography describe successful experiments related to semantic transformation. Danish researchers Hose, Nath, and Pedersen (2015) integrated a Danish agricultural dataset and a business dataset using the SETL framework and found that the extraction and traditional transformation phases took a relatively long time because they involved two spatial join operations for the agriculture datasets (Hose et al., 2015). Hose et al. (2015) recorded a total time of 3,979 seconds required for the process to link the SETL knowledge base (SETLKB) to 14,153 external resources (Hose et al., 2015). The constant linkage to the external sources enabled the retrieval of up-to-date information, leading to a reduction in the lack of data integrity (Hose et al., 2015).

In the second experiment conducted by Erturkmen, Gonul, Pacaci, Sinaci, and Yuksel (2018), LISPA, a regional data warehouse in Italy, and TUD, an EHR database at a university hospital in Germany, were integrated using a three step approach: (a) conversion of source data into a Resource Description Framework (RDF) format, (b) implementation of semantic conversion rules on the data in RDF format to match the common data model (CDM) specifications, and (c) loading the data resulting from steps a and b into the destination repositories. Erturkmen et al. (2018) found that by standardizing both of the databases to an RDF format, defined as "the framework which specifies a language for defining data items and relationships, by using a graph representation, with the intent to scale up and work across the entire internet" (Voegeli, 2018, p. x), and then semantically transforming the databases into a

CDM eliminated discrepancies and duplicates to a large extent, addressing the lack of data integrity caused by data integration.

Uses of Data Integration

One of the main goals of data integration is to produce actionable business intelligence (BI) (Hose et al., 2015). Business intelligence plays a key role in decision-making in the business world (Chen et al., 2013). According to Chen, Chen, and Lim (2013), BI initiatives are enabling businesses to gain "insights from the growing volumes of transaction, product, inventory, customer, competitor, and industry data generated by enterprise-wide applications such as enterprise resource planning (ERP), customer relationship management (CRM), supply chain management (SCM), and knowledge management" (pp. 2-3). The data obtained via data integration is also used in data-visualization activities, mainly by displaying the data in creative ways for easier understanding of a company's past performance (Kirk, 2012). According to Kirk (2012), the presentation of data extends beyond the mere representation of the data and includes "how data representation is integrated into the overall communicated work, exploiting the visual perception abilities related to the scientific understanding of how human eyes and brains process information more effectively" (p. 17). Amplifying cognition is about maximizing the efficiency and effectiveness in processing information into thoughts, understanding, and knowledge. (Kirk, 2012).

Final Thoughts

As data keeps growing exponentially from mobile devices, social media, websites, and sensors, there will be an immense need for data integration in order to place the data in a central repository (Chen et al., 2013; El Hajji et al., 2016). Though data integration is extremely helpful

in consolidating the data, there are some challenges that lead to a lack of data integrity (El Hajji et al., 2013).

Semantic transformations are very useful in filling the gap created by syntactic transformations and address some of the most common data integrity issues such as duplication of data, data inaccuracy, or lack of data currency (Abedjan et al., 2015; El Hajji et al., 2013). Using online reference tables to obtain data dynamically is one of the most efficient ways to avoid obsolete data and helps with keeping the data current (Abedjan et al., 2015). This approach offers promise to organizations that embrace and implement the concept of semantic transformation for maintaining high data integrity, increasing confidence levels when making important decisions based on the data.

References

Abedjan, Z., Ilyas, I. F., Morcos, J., Ouzzani, M., Papotti, P., & Stonebraker, M. (2015). DataXFormer: An interactive data transformation tool. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD'15* (pp. 883-888). New York, NY: Association for Computing Machinery. doi: 10.1145/2723372.2735366

Bansal, S. K. (2014). Towards a semantic extract-transform-load (ETL) framework for big data integration. In Kesselman, C. (Ed.), *2014 IEEE International Congress on Big Data* (pp. 522-529) New York, NY: IEEE Computer Society. doi: 10.1109/BigData.Congress.2014.82

Bansal, S. K., Charkraborty, J., & Padki, A. (2017). Semantic ETL – state-of-the-art and open research challenges. In *2017 IEEE 11th International conference on semantic computing* (pp. 413-418). New York, NY: IEEE Computer Society. doi: 10.1109/ICSC.2017.94

Cabrera, A. M., Cepeda, K., Chamberlain, R.D., Cytron, R. K., Derber, R., Epstein, C., … Zheng, J. (2018). DIBS: A data integration benchmark suite. In *Companion of the 2018 ACM/SPEC International Conference on Performance Engineering - ICPE '18* (pp. 25-28). New York, NY: Association for Computing Machinery. doi: 10.1145/3185768.3186307

Center for Public Issues Education. (n.d.). *Evaluating Information Sources.* University of Florida. Retrieved from https://canvas.uoregon.edu/courses/132553/pages/wk-4-how-to-qualify-references-before-final-selection?module_item_id=2210906

Chen, G., Chen, H., & Lim, E. (2013). Business intelligence and analytics: Research directories. *ACM Transactions on Management Information Systems*, 3(4), 1–10. https://doi.org/10.1145/2407740.2407741

Chen, Q., Hu, H., & Xu, J. (2015). Authenticated online data integration services. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD'15* (pp. 167-181). New York, NY: Association for Computing Machinery. doi: 10.1145/2723372.2747649

Data analyst. (2019). In *Cambridge English Dictionary online*. Retrieved from https://dictionary.cambridge.org/us/dictionary/english/data-analyst

Davidovski, V. (2018). Exponential innovation though digital transformation. In *Proceedings of the 3rd International Conference on Applications in Information Technology - ICAIT'18* (pp. 3-5). New York, NY: Association for Computing Machinery. doi: 10.1145/3274856.3274858

Discover Data Science (2019). *How to become a business intelligence analyst – A complete career guide.* Retrieved from https://www.discoverdatascience.org/career-information/business-intelligence-analyst/

El Hajji, S., Lebdaoui, I., & Orhanou, G. (2013). Data integrity in real-time data warehousing. In *Proceedings of the World Congress on Engineering – WCE'13, Vol 3*. London, UK: World Congress on Engineering. Retrieved from http://www.iaeng.org/publication/WCE2013/WCE2013_pp1516-1519.pdf

El Hajji, S., Lebdaoui, I., & Orhanou, G. (2016). Managing big data integrity. In *2016 International Conference on Engineering & MIS (ICEMIS)*. Agadir, Morocco: ICEMIS. doi: 10.1109/ICEMIS.2016.7745332

Erturkmen, G. B. L., Gonul, S., Pacaci, A., Sinaci, A. A., & Yuksel, M. (2018). A semantic transformation methodology for the secondary use of observational healthcare data in postmarketing safety studies. *Frontiers in Pharmacology, 9*, 435. doi: 10.3389/fphar.2018.00435. Retrieved from https://www.frontiersin.org/articles/10.3389/fphar.2018.00435/full

Friedman, T., & James, G. (2003). *The responsibilities of the data architect.* Retrieved from https://www.bus.umich.edu/kresgepublic/journals/gartner/research/112900/112963/112963.html

Fürber, C. (2015). Data quality management with semantic technologies. In *Research design* (pp. 8–19). Wiesbaden, Germany: Springer Fachmedien. https://doi.org/10.1007/978-3-658-12225-6_2

Gartner, Inc. (2019). CIO (Chief Information Officer). (2019). Retrieved from https://www.gartner.com/it-glossary/cio-chief-information-officer

Hose, K., Nath, R. P. D., & Pedersen, T. B. (2015). Towards a programmable semantic extract-transform-load framework for semantic data warehouses. In *Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP - DOLAP'15* (pp. 15-24). New York, NY: Association for Computing Machinery. doi: 10.1145/2811222/2811229

Kidd, C. (2019, May 31). IT director role and responsibilities: What does a director of technology do? [Blog post]. Retrieved from https://www.bmc.com/blogs/it-director-role-and-responsibilities-what-does-a-director-of-technology-do/

Kirk, A. (2012). Data visualization: A successful design process. In *The context of data visualization: Defining data visualization*. Birmingham, UK: Packt Publishing Ltd.

Kolchinsky, A., & Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus, 8*(6), 41.

McDowall, R. D. (2019). Data integrity and data governance. In *What is data integrity?* (pp. 12-13). Croydon, UK: CPI Group (UK) Ltd.

Snee, R. D. (2015). Data integrity – How to detect lack of integrity. *Institute of validation technology – IVT validation week'15*. Philadelphia, PA: IVT Validation Week. Retrieved from https://www.researchgate.net/publication/282879414_Data_Integrity_Validation

Srividya Bansal - Home. (n.d.). Retrieved June 24, 2019, from http://www.public.asu.edu/%7Eskbansa2/

Stubbs, E. (2011). The value of business analytics: Identifying the path to profitability. *The importance of business analytics*. Hoboken, N.J: John Wiley.

Voegeli, D. (2018). Introduction. In *ETL from RDF to property graph* (pp. ix-x). Bedford, MA: The MITRE Corporation. Retrieved from https://www.mitre.org/sites/default/files/publications/pr-15-2949-ETL-from%20-RDF-to-property-graph_0.pdf