Semantic Metadata Requirements for Data Warehousing from a Dimensional Modeling Perspective

Susanna E. S. Campher[a]
School of Computer Science and Information Systems, North-West University, Potchefstroom, South-Africa

Keywords: Data Warehousing, Semantic Data Management, Metadata, Dimensional Modeling.

Abstract: The era of big data has brought on new challenges to data warehousing. Emerging architectural paradigms such as *data fabric*, *data mesh*, *lakehouse* and *logical data warehouse* are promoted as solutions to big data analytics challenges. However, such hybrid environments, aimed at offering universal data platforms for analytics, have schemas that tend to grow in size and complexity and become more dynamic and decentralized, having a drastic impact on data management. Data integrity, consistency and clear meaning are compromised in large architectures where traditional (relational) database principles do not apply. This paper proposes an investigation into semantic metadata solutions in modern data warehousing from a (logical) dimensional modeling perspective. The primary goal is to determine which metadata and types of semantics are required to support automated dimensionalization as it is assumed to be a good approach to integrate data with different modalities. A secondary goal is finding a suitable model to represent such metadata and semantics for both human and computer interpretability and use. The proposal includes a description of the research problem, an outline of the objectives, the state of the art, the methdology and assumptions, the exepected outcome and current stage of the research.

1 INTRODUCTION

Since the first concepts of a data warehouse (DW) in support of decision-making emerged in the late 1970s, these information systems have been heavily reliant on relational data theories and architectures appropriate for collecting and integrating mainly structured data from internal operational information systems in well-controlled "closed-world" scenarios (Abelló et al., 2015, p. 571; Krishnan, 2013, p. 128). The era of big data and distributed computing has brought on new challenges to data warehousing as well as data management in general. Decentralized data generation, management, and polyglot persistence result in large volumes of fast-moving, dissociated, heterogeneous data. Big data use cases generally involve collecting and integrating a variety of data from various internal and external sources to repurpose it for value-add analytics (Freitas & Curry, 2016, p. 90). Relational database management systems and Structured Query Language (SQL) are not suitable for storing and analyzing big data. Conventional data warehousing approaches, from data provisioning to schema design, have to be reconsidered. In general, a paradigm shift regarding data assets is required – where they are collected, how they are analyzed, and how insights from analyses are monetized (Kimball & Ross, 2013a, p. 527).

Emerging architectural paradigms such as *data fabric*, *data mesh*, *lakehouse* and *logical data warehouse* are promoted in industry as solutions to big data analytics challenges (Gacitua et al., 2019, p. 15; IBM, 2021; Jägare, 2020; Welch, 2021). However, such hybrid environments, aimed at offering universal data platforms for analytics, have schemas that tend to grow in size and complexity and become more dynamic and decentralized, having a drastic impact on data management (Freitas, 2015, p. 26). Data integrity, consistency and clear meaning are compromised in large architectures where traditional (relational) database principles do not apply (Helland, 2011, p. 40). Moreover, as data are typically generated in formats for specific use cases at the source, attempting to integrate and reuse it outside its original context requires methodologies to better describe and contextualize it for meaningful use

[a] https://orcid.org/0000-0002-3676-9327

Campher, S.
Semantic Metadata Requirements for Data Warehousing from a Dimensional Modeling Perspective.
DOI: 10.5220/0011018200003179
In *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022) - Volume 1*, pages 129-136
ISBN: 978-989-758-569-2; ISSN: 2184-4992

elsewhere. In consequence, approaches to maximize quality and usefulness within multiple contexts have to be investigated (Freitas & Curry, 2016, p. 91). New theory and taxonomy of data are also needed to address data curation issues, such as data (object) identification, versioning, lineage or provenance, and accuracy measurement (Helland, 2011, p. 47), typically of concern in the field of data warehousing.

Modern approaches to data warehousing include the *Kimball Lifecycle* (Kimball et al., 2008), *DW 2.0* (Inmon et al., 2010), and *Data Vault 2.0* (Linstedt & Olschimke, 2016). The Kimball Lifecycle is a well-known industry standard prescribing a bus architecture with a business process focus in which individual logical dimensional models are iteratively designed to fit into the overall architecture. The methodology prescribes how these models may be implemented in relational database management systems (RDBMs) as a queryable presentation layer that is easy to understand and delivers fast query performance. An extended relational DBMS architecture to accommodate big data is briefly presented in Kimball and Ross (2013a). DW 2.0 has a focus on the lifecycle of data and information as well as different types and structures of data and how they relate, with metadata forming the backbone of the infrastructure. The Data Vault 2.0 approach addresses some of the issues associated with traditional data warehousing, such as high rates of production failures, business rule complexity and slow load times over big data volumes. According to Linstedt (2019), data acquisition into a raw data vault is faster and highly automatable. However, the compromise is that business rules are decoupled from a data vault, and only about 60% of the process to create dimensionalized information marts for consumption can be automated, as raw data still need to be cleaned and contextualized. Furthermore, the need to define metadata for additional external data sources requires human effort and places a growing demand on data acquisition teams.

2 RESEARCH PROBLEM

Having more flexible and scalable data and DW architectures and models to accommodate big data does not address the challenges of data management, per se. Issues that have to be addressed include controlled data creation, integration, discovery and retrieval; data contextualisation and resolution; quality assurance, reliability and lineage requirements; change data capture; and value-added interpretability for both humans and machines. Automating processes to address these issues is necessary in order to scale with big data (Stonebraker et al., 2013). As such, the inclusion of suitable metadata about the meaning of the data is essential, something that is often overlooked and underspecified in source schemas (Krishnan, 2013, p. 181; Miller et al., 2001, p. 78). In that regard, semantic data management and semantification of (granular) data have to be considered.

In some of the Kimball Group’s last writings before their retirement in 2015, best practices for various aspects of big data were specified. The overall recommendations include thinking dimensionally – dividing the world into facts and dimensions – and integrating data based on conformed dimensions. In terms of architecture, Kimball and Ross (2013a, p. 533) advise building comprehensive ecosystems around a logical data highway with various data cashes of increased latency and data quality. In such a highway, most of the data are kept in non-relational form and big data analytics may be employed to perform extract, transform and load (ETL) from one cache to the next. ETL includes data filtering, cleaning, dimensionalization and extracting value-added information and business measures. From a modeling perspective, emphasis is placed on automated dimensionalization of data early on the data highway as the best approach to integrate data with different modalities in order to add value to data as soon as possible. In one of their final design tips, Kimball asserts that modern storage architecture should be open to various types of tools and analytic clients, with access to the data being provided through a universal metadata layer. As such, metadata descriptions of data of all types have to be extensible, customizable and powerful enough to include enough semantics for new complex data sources (Kimball, 2016, p. 848). Suitable semantic metadata descriptions would enable seamless integration of heterogeneous data into various presentational formats, including dimensional models. For example, social media sentiments such as “This product is great!” with a link to a product page, may be dimensionalized with information about the product, customer, location, promotion, weather, etc. In a healthcare environment, patient profiles may be combined with image data such as sonograms and electrocardiograms, and text data such as physician reports or medical treatments.

In general, semantification of big data would prove valuable to address data curation and integration challenges in innovative ways, as traditional schema matching methods (for instance-level data integration) do not work well with heterogeneous and dynamic data schemas (Miller, 2014). Platforms where "specified (and unspecified) schema information, constraints and relationships can be learned, reasoned about and verified" are desirable for integrating heterogeneous data (Miller et al., 2001, p. 78). Machine-assisted data integration from various data sources, including relational databases, data lakes, distributed files and web data feeds, is envisioned with active metadata management and supporting technologies, such as semantic knowledge graphs and embedded machine learning (Gupta, 2021). For this to realize, it is crucial that computer systems and machines understand the semantics of data and are able to connect technical, business, operational and social types of metadata. Syntactic representations of schemas, metadata and data need to convey enough semantics on a fine-grained data (attribute) level for the automatic creation of mappings between data elements and other data integration tasks. Incorporating Semantic Web (SW) technologies, such as taxonomies, ontologies and the Resource Description Framework (RDF) into data solutions enables intelligent links to be created within the data, resulting in a powerful semantic layer for data integration and analysis (Krishnan, 2013, pp. 193, 204). Semantification should also be done in a manner that is easy for business users to understand; for example, with the use of knowledge graphs (Den Hamer et al., 2021; Gupta, 2021).

The focus of this study will be on semantic metadata solutions in modern data warehousing from a (logical) dimensional modeling perspective, in particular. The primary goal is to determine which metadata and types of semantics are required to enable automated dimensionalization while meeting fundamental data warehousing principles and practices (such as having subject-oriented, integrated, time variant, and non-volatile collections of data). Dimensionalization of data is assumed to be a good approach to integrate data with different modalities. A secondary goal is finding a suitable model to represent such metadata and semantics for both human and computer interpretability and use. The following research questions are proposed:

Q1) "*What are the metadata and semantic requirements that will enable automated dimensionalization for integration?";* and

Q2) "*How can heterogeneous data be semantified to enable both human and machine interpretability in data warehousing?*"

3 OUTLINE OF OBJECTIVES

This research will endeavor to answer the questions by systematically studying design principles and best practices (or patterns) in data warehousing, specifically related to dimensionalization for the purpose of data integration and management. The aim is to analyze and understand these design patterns, and to document their rationale and function into a suitable knowledge base (KB) from which requirements for semantic data enrichment can be extracted. The KB will be used to derive a conceptual model for metadata and semantified data representation in support of machine-enabled data warehousing.

The objectives of the study are as follows:

O1. Conduct a literature study regarding Semantic Web technologies and the use of these technologies in data architectures and data warehousing.

O2. Conduct a literature study regarding knowledge representation as well as data, metadata and semantic modeling.

O3. Perform pattern mining on key data warehousing practices and principles, specifically related to dimensional modeling for integration and data management in data warehousing. This objective includes the creation of an ontology for representing the design pattern concepts and categories. The patterns are to capture and include the rationale behind the specific practice as well as metadata and semantic data requirements.

O4. Design and develop a KB to document the design patterns (Artifact 1). The KB should represent a theoretical model of the observed patterns in a suitable representation.

O5. Evaluate the validity of the KB in terms of fit, relevance, workability, and modifiability by applying it within the context of the second research question (Q2). A comparison to a "design theory nexus", as described by Pries-Heje and Baskerville (2008), may be included.

O6. Derive a conceptual model (Artifact 2) to semantically enrich heterogeneous data based on the metadata and semantic data requirements embedded in the design pattern KB.

O7. Demonstrate and evaluate the utility, quality, and efficacy of the conceptual model using a real-world problem scenario. Evaluation could also include comparison to the Common Warehouse Metamodel (CWM) multidimensional metamodel (OMG, 2003), the RDF data cube vocabulary (W3C, 2014), and the HANDLE metadata model for data lakes (Eichler et al., 2021).

Although some of these objectives may seem loosely formulated, a large aspect of this study will involve determining the specific objectives of a solution to the research problem. The research problem is considered a “wicked problem” as described by Rittel and Webber (1973) since the solution objectives are not clear, the problem itself can only be formulated in terms of a solution, and the various consequences of a particular solution is unbounded over application domain and time. (More information regarding the characteristics of wicked problems are provided in assumption A5 in section 5.3.) It is also the reason why a problem-centred Design Science Research (DSR) approach is considered appropriate. In a problem-centred DSR cycle, determining specific solution objectives is an essential part of the research process (Peffers et al., 2007, p. 55). Furthermore, the search for suitable metadata and semantics in support of automation in data warehousing, in this study, will focus specifically on those required for automated dimensionalization as a means to both instance-level and conceptual level data integration and management. Considering all possible data warehousing domains and purposes is beyond the scope of this investigation. However, creating an extensible KB for data warehousing design patters is an effort towards assimilating a design theory nexus that would be helpful in future when more design patterns regarding alternative data warehousing practices and approaches are added. A design theory nexus is helpful to understand, evaluate and compare available alternatives from many competing, and often highly dissimilar approaches with different underlying assumptions and theories; it assists in selecting an ideal solution to a complex problem, sometimes by combining dissimilar design theories (Pries-Heje & Baskerville, 2008, p. 750).

4 STATE OF THE ART

Surveys regarding the use of SW technologies in data warehousing include Abelló et al. (2015), and Gacitua et al. (2019). Results indicate that these technologies offer promising solutions to challenges associated with data provisioning and schema management in DWs. Specific areas mentioned as opportunities for further research include automatic (or semi-automatic) derivation of ontological mappings of heterogeneous data, ways to enable semantic-aware data integration and DW schema management, automatic data transformation for semantically traceable models, and semantically enriching and annotating multidimensional models.

Ptiček et al. (2019) describe the potential of a semantic paradigm in warehousing of big data. They emphasise the need to address the integration of (schemaless) big data into DWs on a higher level of abstraction, i.e., the data modeling phase, and describe various promising solutions, such as ontology-based schema integration. Selma et al. (2012) propose a methodology for designing an ontology-based Web DW given ontology-based databases as sources. Mami et al. (2016) developed a method of big data semantic enrichment using RDF vocabularies, but without considering an attached source schema (which requires source schema extraction). McCusker et al. (2009) demonstrated how semantically well-annotated, distributed data can be integrated using semantic extraction, transformation and loading (SETL) with the Web Ontology Language (OWL) into an ontologically-driven data store or knowledge base. Bansal and Kagemann (2015) also proposed a SETL framework for integrating linked Web data which involves steps of data preparation, creating a semantic data model, and generating RDF triples.

Although all of these studies show promising results and suggest further research into SW technologies for data warehousing, these approaches still require either sources that are already semantified well enough, or big data schema extraction methods. Schema extraction is not trivial and an underdeveloped field that could prove to be unsustainable as a semi-automatic process having to be continually developed for new data representations and having to scale proportionally with data size (Mami et al., 2016, p. 389; Ptiček et al., 2019, p. 402). Furthermore, McCusker et al. (2009) and Bansal and Kagemann (2015) reported the need for manual intervention even though their experimental data were annotated with semantic metadata.

5 METHODOLOGY

5.1 Research Approach

The study will be conducted using a mix of methods from the interpretive paradigm as well as DSR. Firstly, the study is concerned with analyzing and understanding (existing) knowledge about data warehousing principles and practices from a dimensional modeling perspective. The aim is not to determine which practices are the best, but to critically examine the context that led to different practices, and compare and summarize them in order to incorporate such perspectives into a proposed solution to the research problem. Secondly, conceptualizations of this knowledge need to be presented in a manner that is open, extensible and interpretable by humans and machines. Understanding how this can be done will be investigated through the creation of artifacts – a KB and derived conceptual model for semantification of heterogeneous data.

The overall process of the study will follow a DSR approach similar to the one conducted by Berndt et al. (2003) and described by Peffers et al. (2007) as an example of a problem-centered approach (as opposed to an objective-centered solution for which the specific solution objectives are clear from the start). It is also acknowledged that DSR can take on other types of methodologies, such as one to support context-specific research; Peffers et al. (2007) provide the example of employing custom methods for requirements analysis in information systems.

The research plan is depicted in Figure 1. The process flow on the left outlines the general activities of DSR, while detailed steps are depicted on the right.

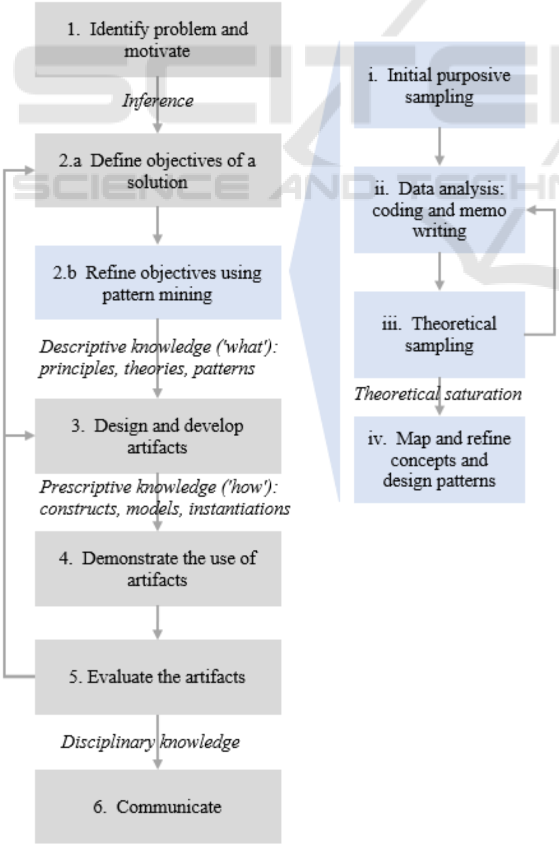
1. Identify problem and motivate
Inference
2.a Define objectives of a solution
2.b Refine objectives using pattern mining
Descriptive knowledge ('what'): principles, theories, patterns
3. Design and develop artifacts
Prescriptive knowledge ('how'): constructs, models, instantiations
4. Demonstrate the use of artifacts
5. Evaluate the artifacts
Disciplinary knowledge
6. Communicate
i. Initial purposive sampling
ii. Data analysis: coding and memo writing
iii. Theoretical sampling
Theoretical saturation
iv. Map and refine concepts and design patterns

Figure 1: Research plan.

During activity 1, the research problem is identified and the value of a solution is justified. Some of the metarequirements (or objectives of the solution) may also be obtained during this activity by investigating the state of the problem and existing knowledge (Peffers et al., 2007, p. 55). Consequently, some general objectives for the solution can be inferred in activity 2.a (defining the objectives of a solution), but may require additional investigation. More specific objectives, related to the first research question (Q1), will be obtained through data collection and analysis during activity 2.b, which will involve pattern mining of data warehousing principles and practices. (This activity is described in more detail in the following section about data collection and analysis.)

Knowledge of theory emerging from the previous steps may then be applied during activity 3 to design and develop DSR artifacts. The artifacts will be demonstrated and evaluated during activities 4 and 5, respectively. Both artifacts will be evaluated by comparing the various design objectives to actual observed results, and possibly to other industry standards. Finally, the disciplinary knowledge obtained from the process will be communicated in a doctoral thesis (activity 6). The activities of design, demonstration and evaluation are iterative, and may also require revisiting the solution objectives.

5.2 Data Collection and Analysis

Data collection and analysis will involve model-based qualitative methods similar to those described by Hentrich et al. (2015) and also conducted by Zdun et al. (2018). It is a combination of methods based on Grounded Theory (GT), and those for collecting and analyzing best practices, such as pattern mining. The empirical data about data warehousing design patterns will undergo coding and comparison, resulting in concepts and categories (and a pattern taxonomy) that can be organized and formulated into a suitable KB (see Gorton et al. (2015) for similar work). Memo-writing and diagramming will be performed throughout to record comparisons and the developing thinking during analysis.

The process of data collection and analysis (the pattern mining) is depicted in the detail steps of activity 2.b of the research plan Figure 1. Sampling will begin with initial, purposive sampling, e.g., best practices for big data in data warehousing provided by Kimball and Ross (2013a), and the latest collection of best practices and data warehousing design tips provided by Kimball and Ross (2016). Theoretical sampling will follow the analysis of data from this first source, informed by coding, comparison and memo-writing, in an attempt to fill gaps, clarify uncertainties, and test interpretations in the developing theory. Data analysis followed by theoretical sampling will be repeated until theoretical saturation is reached, i.e., when the design patterns (the theory) are "dense and logical and there are no gaps in the explanations" (Corbin & Strauss, 2014, p. 139) or all the concepts in the developing theory can be substantiated from the data and are well understood (Sbaraini et al., 2011, p. 3). The design and development of the DSR artifacts will be based on theory obtained from this process.

5.3 Assumptions

The following assumptions are currently relevant within the context of this study.

A1. A specific paradigm (and associated methodology) is a construct of human thought – a constellation of assumptions, theories and methods – which remains a heuristic device that should not be singled out as an absolute view to which the world could actually conform (Mingers, 2001). Declaring assumptions and critically selecting and using suitable methods throughout (different phases of) a project are more important than committing to a single customary paradigm or methodology.

A2. Software design patterns can be considered sociological phenomena – they represent successful problem-solving behavior and involve significant evolutionary and human aspects. Grounded Theory methods are consequently suitable to discover such patterns (Hentrich et al., 2015). The same reasoning can be applied to design patterns in data warehousing.

A3. Dimensional models present data in a predictable, intuitive framework, which simplifies both human and computer processing of the schemas (Kimball & Ross, 2016, pp. 149,153); "This process of dimensionalization is a good application for big data analytics" (Kimball & Ross, 2013b, p. 538). Dimensionalization is therefore a good approach to integrate data with different modalities.

A4. "We must be systematic, but we should keep our systems open" (Whitehead, 1956, p. 8). It is important to be sensitive to the limitations in specific systems and to realize there is always more detail beyond. Such an "open system" approach could be applicable to this study in terms of two aspects. Firstly, any model or mechanism to represent semantics of data, ought to be void of global consistency requirements. This is in line with the thinking of Berners-Lee (1998) about a Semantic Web language being "complete" in the sense that it should be able to represent all kinds of data and information about the world with arbitrary complexity, including paradoxes and tautologies. Berners-Lee (1998) further supposed that knowledge representation systems will be "webized" (become part of the Semantic Web) once "centralized concepts of absolute truth, total knowledge, and total provability" are eliminated. Secondly, in terms of qualitative research conducted, a researcher ought to remain open to theoretical frameworks.

A5. The problem of finding a model for semantic content management is a "wicked problem". March and Hevner (2007, pp. 1036-1037) state that "data integration has been studied extensively in the context of heterogeneous databases; however, its solution remains elusive". Bizer et al. (2011) challenge the scientific community to "demonstrate the benefit of semantics for data integration", and that "[w]e have talked extensively about smarter databases but the actual requirement remains vague". According to Rittel and Webber (1973), a wicket problem does not have a definitive formulation; including sufficient detail about the problem would require an exhaustive list of all its conceivable solutions up front. A wicked problem has neither clear criteria that determines when a solution has been found, nor criteria to prove that all solutions have been considered. It also has a good-or-bad solution rather than a true-or-false one, has no ultimate test for a solution, and no one solution that fit all variants of the problem. A wicket problem is often a symptom of another problem and the resolution of the problem depends on how the underlying problem is explained. For example, it is possible to claim that effective data management is a problem because (a) modern data architectures have schemas that tend to grow in size and complexity, and (b) data integrity, consistency and clear meaning are compromised in large architectures where traditional database principles do not apply. Because of explanations like these, one might argue for stricter schema management, or for finding ways to translate and apply relational principles in big data environments. Regardless, in literature there are explanations of the data management problem in terms of lack of suitable data semantics. Consequently, it is considered worth investigating ways in which adding suitable semantics may solve the problem.

6 EXPECTED OUTCOME

The contribution of this study, in general, is towards solving problems related to data warehousing in the era of big data. Specific contributions foreseen are C1) a collection of design patterns regarding modelling, management and data curation practices and principles for dimensionalization and integration in data warehousing; C2) a prototype of an extensible KB for DW design patters; and C3) a conceptual model for semantification of heterogeneous data in support of automated dimensionalization and integration in data warehousing.

7 STAGE OF THE RESEARCH

This study is currently in phase 2.a (define objectives of a solution) of the research plan depicted in Figure 1. It involves a literature study and investigation into existing knowledge and solutions regarding the research problem and questions.

REFERENCES

Abelló, A., Romero, O., Pedersen, T. B., Berlanga, R., Nebot, V., Aramburu, M. J., & Simitsis, A. (2015). Using Semantic Web technologies for exploratory OLAP: A survey. *IEEE transactions on knowledge and data engineering*, *27*(2), 571-588.

Bansal, S. K., & Kagemann, S. (2015). Integrating big data: a semantic extract-transform-load framework. *Computer*, *48*(3), 42-50.

Berndt, D. J., Hevner, A. R., & Studnicki, J. (2003). The Catch data warehouse: Support for community health care decision-making. *Decision Support Systems*, *35*(3), 367-384.

Berners-Lee, T. (1998). *What the Semantic Web can represent*. Retrieved September 9, 2021, from https://www.w3.org/DesignIssues/RDFnot.html

Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. (2011). The meaningful use of big data: Four perspectives - four challenges. *ACM SIGMod Record*, *40*(4), 56-60.

Corbin, J. M., & Strauss, A. L. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.

Den Hamer, P., Hare, J., Jones, L. C., Choudhary, F., Sallam, R., & Vashisth, S. (2021). *Top trends in data and analytics for 2021: From big to small and wide data*. Retrieved October 15, 2021, from https://www.gartner.com/doc/reprints?id=1-27GN5DZK&ct=210917

Eichler, R., Giebler, C., Gröger, C., Schwarz, H., & Mitschang, B. (2021). Modeling metadata in data lakes – a generic model. *Data & knowledge engineering*, *136*, 101931.

Freitas, A. (2015). *Schema-agnostic queries for large-schema databases: A distributional semantics approach* [Thesis – PhD, National University of Ireland]. Galway.

Freitas, A., & Curry, E. (2016). Big data curation. In J. M. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe* (pp. 87-118). Springer.

Gacitua, R., Mazon, J. N., & Cravero, A. (2019). Using Semantic Web technologies in the development of data warehouses: A systematic mapping. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1293.

Gorton, I., Klein, J., & Nurgaliev, A. (2015). Architecture knowledge for evaluating scalable databases. In L. Bass, P. Lago, & P. Kruchten (Eds.), *Proceedings* (pp. 95-104). 12th Working IEEE/IFIP Conference on Software Architecture (WICSA 2015), Montreal, Canada. IEEE.

Gupta, A. (2021, May 11). *Data fabric architecture is key to modernizing data management and integration*. Gartner. Retrieved October 13, 2021, from https://www.gartner.com/smarterwithgartner/data-fabric-architecture-is-key-to-modernizing-data-management-and-integration

Helland, P. (2011). If you have too much data, then 'good enough' is good enough. *Communications of the ACM*, *54*(6), 40-47.

Hentrich, C., Zdun, U., Hlupic, V., & Dotsika, F. (2015). An approach for pattern mining through grounded theory techniques and its applications to process-driven SOA patterns. In U. Van Heesch & C. Kohls (Eds.), *EuroPLoP 2013: Proceedings of the 18th European Conference on Pattern Languages of Program* (pp. 1-16). Irsee, Germany. ACM.

IBM. (2021). *Data fabric architecture delivers instant benefits*. Retrieved November 22, 2021, from https://www.ibm.com/downloads/cas/V4QYOAPR

Inmon, W. H., Strauss, D., & Neushloss, G. (2010). *DW 2.0: The architecture for the next generation of data warehousing*. Morgan Kaufmann.

Jägare, U. (2020). *The modern cloud data platform for dummies: Databricks special edition*. Wiley.

Kimball, R. (2016). The future is bright. In R. Kimball & M. Ross (Eds.), *The Kimball Group reader: Relentlessly practical tools for data warehousing and business intelligence* (2nd ed. pp. 847-851). Wiley. (Design Tip #180, December 1, 2015)

Kimball, R., & Ross, M. (2013a). Big data analytics. In *The data warehouse toolkit: the definitive guide to dimensional modeling* (3rd ed. pp. 527-542). Wiley.

Kimball, R., & Ross, M. (2013b). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). Wiley.

Kimball, R., & Ross, M. (2016). *The Kimball Group reader: Relentlessly practical tools for data warehousing and business intelligence* (2nd ed.). Wiley.

Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2008). *The data warehouse lifecycle toolkit* (2nd ed.). Wiley.

Krishnan, K. (2013). *Data warehousing in the age of big data*. Morgan Kaufmann.

Linstedt, D. (2019). *Understanding Data Vault 2.0* [Video]. https://datavaultalliance.com/news/dv/understanding-data-vault-2-0/

Linstedt, D., & Olschimke, M. (2016). *Building a scalable data warehouse with data vault 2.0*. Morgan Kaufmann.

Mami, M. N., Scerri, S., Auer, S., & Vidal, M.-E. (2016). Towards semantification of big data technology. In T. Hara & S. Madria (Eds.), *Big Data Analytics and Knowledge Discovery* (pp. 376-390). 18th International Conference (DaWaK 2016), Porto, Portugal. Springer.

March, S. T., & Hevner, A. R. (2007). Integrated decision support systems: A data warehousing perspective. *Decision Support Systems*, *43*(3), 1031-1043.

McCusker, J. P., Phillips, J. A., Beltrán, A. G., Finkelstein, A., & Krauthammer, M. (2009). Semantic web data warehousing for caGrid. *BMC bioinformatics*, *10*(10), 1-10.

Miller, R. (2014). *Big data curation* [Keynote abstract]. Paper delivered at the 20th International Conference on Management of Data, Hyderabad. http://comad.in/comad2014/Proceedings/Keynote2.pdf

Miller, R. J., Hernández, M. A., Haas, L. M., Yan, L., Howard Ho, C., Fagin, R., & Popa, L. (2001). The Clio project: Managing heterogeneity. *ACM SIGMod Record*, *30*(1), 78-83.

Mingers, J. (2001). Combining IS research methods: Towards a pluralist methodology. *Information systems research*, *12*(3), 240-259.

OMG. (2003). *Common Warehouse Metamodel (CWM) Specification*. Retrieved November 11, 2021, from https://www.omg.org/spec/CWM/1.1/PDF

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.

Pries-Heje, J., & Baskerville, R. (2008). The design theory nexus. *MIS quarterly*, *32*(4), 731-755.

Ptiček, M., Vrdoljak, B., & Gulić, M. (2019). The potential of semantic paradigm in warehousing of big data. *Automatika*, *60*(4), 393-403.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy sciences*, *4*, 155-169.

Sbaraini, A., Carter, S. M., Evans, R. W., & Blinkhorn, A. (2011). How to do a grounded theory study: A worked example of a study of dental practices. *BMC Medical Research Methodology*, *11*, 128.

Selma, K., Ilyès, B., Ladjel, B., Eric, S., Stéphane, J., & Michael, B. (2012). Ontology-based structured web data warehouses for sustainable interoperability: requirement modeling, design methodology and tool. *Computers in industry*, *63*(8), 799-812.

Stonebraker, M., Bruckner, D., Ilyas, I. F., Beskales, G., Cherniack, M., Zdonik, S. B., Pagan, A., & Xu, S. (2013). *Data curation at scale: The data tamer system*. Paper delivered at the 6th Biennial Conference on Innovative Data Systems Research (CIDR 2013), Asilomar, CA.

W3C. (2014). *The RDF data cube vocabulary*. Retrieved October 27, 2021, from https://www.w3.org/TR/vocab-data-cube/

Welch, S. (2021). *External data platforms as part of the modern data stack (DSC webinar series)* [Video]. https://www.datasciencecentral.com/video/dsc-webinar-series-external-data-platforms-as-part-of-the-modern

Whitehead, A. N. (1956). *Modes of thought*. Cambridge University Press.

Zdun, U., Stocker, M., Zimmermann, O., Pautasso, C., & Lübke, D. (2018). Guiding architectural decision making on quality aspects in microservice APIs. In C. Pahl, M. Vukovic, J. Yin, & Q. Yu (Eds.), *Service-oriented computing* (pp. 73-89). 16th International Conference on Service-Oriented Computing (ICSOC 2018), Hangzhou. Springer.