Literature Review on Data Vaults – What is the State of the Art of Literature on Data Vaults?

Seminar paper

Gribova, Svetlana, FH Wedel, Wedel, Germany, winf104887@fh-wedel.de

Abstract

The Data Warehouse is a critical component of information management in an organization. The choice of an appropriate data model to address the requirements of a modern Data Warehouse has been intensively discussed. Two modeling approaches opposed each other initially, namely the 3NF and dimensional approach. Although, they have limitations regarding their flexibility and scalability. The Data Vault approach has emerged as an alternative modeling technique that aims to address the shortcomings of traditional models. This paper reviews the literature on Data Vault modeling and examines the state of current research. It aims to provide a state-of-the-art overview of dimensions connected to the topic.

Keywords: Data Vault, data warehousing, data modeling.

Table of Contents

1 Introduction

The Data Warehouse (DW) has become a key component in an organization's information management since it supports decision-making and serves a variety of tasks, from routine work to complex planning (Shin, 2003). According to Golfarelli (2009), the DW must fulfill specific criteria in order to enable high-quality reporting. First, the scalability of the hardware and software components of the DW is vital since data volume and user requirements constantly increase or change (Golfarelli, 2009). Second, the DW should be extensible and flexible and be able to integrate new data sources and changes without redesigning the initial system (Golfarelli, 2009). Finally, the DW should preserve the history of data and make it accessible for analysis (Golfarelli, 2009).

The choice of an appropriate data model to address the requirements of a modern DW has been intensively discussed (Gluchowski, 2021). Although, there is no standardized data model for DW development (Bojicic et al., 2016). The two conventional data models, namely the dimensional and the relational model, opposed each other initially (Gluchowski, 2021). The 3NF, or the Entity-Relationship approach, was introduced by Bill Inmon (Breslin, 2004). It bases on the theory of relational databases, which are highly normalized (Breslin, 2004). Its counterpart, the dimensional model, was created by Ralph Kimball (Breslin, 2004). The critical difference is the denormalized structure of the data (Breslin, 2004). The dimensional model views data as facts linked to several dimensions (Breslin, 2004). A fact represents a focus of analysis and typically includes attributes called measures (Breslin, 2004). Measures are usually numeric values of central interest for analysis and reporting (e.g., the amount of sold products or transportation costs) (Breslin, 2004). However, these models have some limitations regarding their flexibility and scalability (Gluchowski, 2021). In recent years, the Data Vault (DV) model has been considered a valid alternative for building a DW (Gluchowski, 2021).

In this paper, I will systematically examine the state of the art of academic literature on the topic of Data Vaults and its relevant dimensions. Moreover, I will outline the commonalities and contradictions of the current academical research, analyze gaps and depict some insights into the management value.

In the next part of the paper, I provide some background information for understanding the topic of DV modeling. Afterward, I describe the process of identifying and finding the literature for this paper. In the section after that, I summarize the most relevant contents. In the discussion chapter, I emphasize commonalities and contradictions noticed in the literature and identified gaps for future research. Furthermore, I will provide some practical suggestions. In conclusion, I will give a concise summary of my findings.

2 Background

In this chapter, I define and describe concepts necessary for understanding the topic in advance. It is essential to understand the main aim of a DV, its general structure, and its characteristics. However, it is not a part of the literature review itself.

The Data Vault approach proposed by Linstedt aims to address the weaknesses of the conventional methods and to meet the needs of the modern DW (Linstedt, 2002). It promises to support the agility, flexibility, and scalability of the DW (Linstedt, 2002). The fundamental principle of the DV is providing the “single version of facts” (Linstedt and Olschimki, 2015). While the goal of the “single version of truth” is to provide an integrated, cleansed version of the organizational information, “the single version of facts” goal is to provide and process all data at all times, even data containing errors (Linstedt and Olschimki, 2015).

The core entity of a DV model is a Hub, which represents a business object, e.g., an invoice, customer, or employee (Linstedt, 2002). This business object is identified by a unique business key that does not change over time (Linstedt, 2002). In contrast, all the descriptive information around a Hub entity, which has a volatile character and may frequently change over time, is stored in a Satellite (Linstedt, 2002). Two or more Hubs can be connected by a Link entity, which corresponds to a many-to-many 3NF relationship (Linstedt, 2002). A Link contains the foreign keys referencing the connected Hubs (Linstedt, 2002). It describes the interactions and relationships between the business keys (Linstedt, 2002). Every DV component contains metadata, such as a load timestamp and a record source (Linstedt, 2002). None of the obsolete data is ever removed from a Satellite; instead, a Satellite keeps all the historical data, including the metadata that defines the validity period, such as the load date and the load end date (Linstedt, 2002). Figure 1 shows an example of a DV model.

The DV methodology does not transform data before loading it into the DW (Naamane and Jovanovic, 2016). Its ETL process in the DV approach usually runs in two steps: in the first step, all Hubs are loaded in parallel; in the second step, all Links and Satellites are loaded in parallel (Naamane and Jovanovic, 2016). The DV model is characterized by strong normalization of the data and a high number of components (Linstedt, 2005). One of the most important characteristics is data separation since the data are organized around the business objects (Naamane and Jovanovic, 2016). The descriptive data and relationships are separated from the business key and stored in a physically independent concept (Bojicic et al., 2016). The changes in business requirements or data sources are implemented through additions, which do not influence the existing structure (Naamane and Jovanovic, 2016).

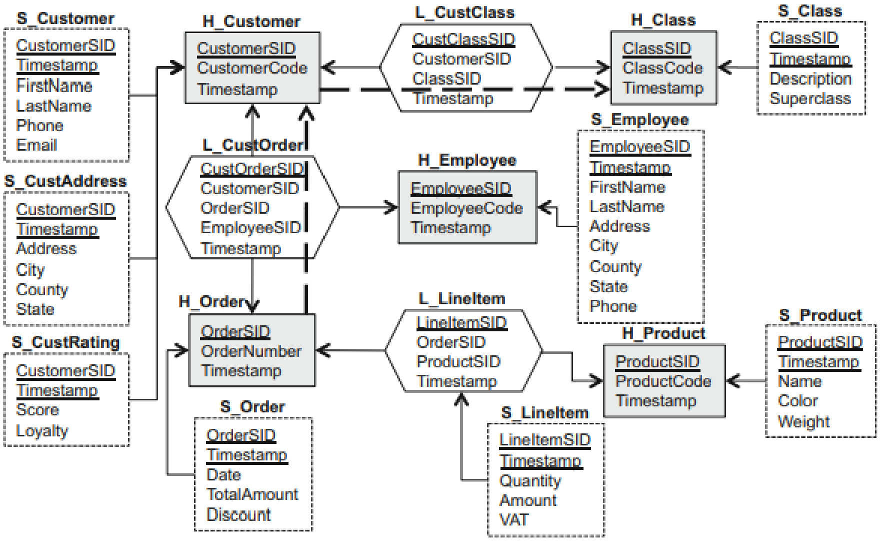

Figure 1. *An exemplary structure of a DV model. The model contains five business objects (customer, class, employee, order, and product). The Hubs are connected via Links and have at least one Satellite with descriptive information. Adapted from Hultgren (2012).*

3 Literature Review

In this chapter, I describe my steps to provide transparency over my reviewing process. In order to conduct a rigorous literature review, a systematic approach was required. I used the method Wolfswinkel et al. (2013) proposed, which contains five stages. In the following, I describe these stages and my proceeding.

Define: In the first stage, I defined the inclusion and exclusion criteria for relevant research areas, databases, and keywords. Since no synonyms or other terms describe the DV topic precisely, I did not use any different keywords. Moreover, I found many papers using the keyword, so including other keywords seemed futile. To ensure that the found literature includes the term “data vault” in consecutive order, I used quotation marks. Afterward, I identified the research fields since this is important for the choice of databases. Since the DV approach relates to DW, it correlates to the management and information systems fields. For these research fields, the following databases were qualified: Business Source Premier, IEEEXplore, Journal Citation Reports, IOPScience, Scopus, SIAM Journals Online, Web of Science Core Collection, and Google Scholar.

Search: In the second phase, I searched for potentially relevant literature in the selected databases. I started my search considering only literature, including the keyword in its title and paper keywords. Since I did not identify other missing keywords or research areas, I did not revisit the first defining phase and continued with the selection phase.

Select: In the third phase, I filtered out the doubles first. Then I read the abstracts of the papers and evaluated if the paper was relevant to the topic. The paper should relate to data modeling and data warehousing using the DV approach. To ensure I gathered a relatively complete set of pertinent literature, I used the forward and backward citation tracking proposed by Webster and Watson (2002). If a new article came up, I included it in the sample and repeated the iteration until there was no relevant literature to consider. After the completion of this stage, there were 14 papers to review.

Analyze: In the fourth phase, I used the approach proposed by Wolfswinkel et al. (2013), which bears on the Grounded Theory, and includes three stages: open coding, axial coding, and selective coding. First, I conducted open coding by reading some papers and highlighting the critical insights. After reading eight papers, common concepts appeared. In the axial coding, I identified the interrelations between the concepts. In the selective coding, I refined the categories. The found dimensions were the foundation for the targeted reading of the remaining literature. To complete a well-structured review, I compiled a concept matrix while analyzing the papers. The concept matrix allowed me to group the key concepts and systematically describe them to present the insights to the reader clearly.

Present: In the last fifth phase, I presented the insights and described the found dimensions using the conducted concept matrix.

4 Findings and Results

4.1 Overview

In the final analysis, I examined an amount of 14 papers. I created a concept matrix to provide a visualized overview of the findings (Table 1). In the following sections, I describe each dimension with its sub-categories in detail.

Authors	*Evaluation*	*Conceptual Data Vault*	*Design Automation*	*Metadata Vault Approach*
Bojicic et al., 2016	•			
Gluchowski, 2021	•			
Golfarelli et al., 2016			•	
Grigoriev et al., 2021	•			
Jovanovic and Bojicic, 2012		•		
Jovanovic et al., 2012	•	•		
Krneta et al., 2014			•	
Krneta et al., 2016			•	
Krneta and Krneta, 2020	•		•	
Naamane and Jovanovic, 2016	•			
Naamane and Jovanovic, 2017				•
Schnider et al., 2014	•			
Subotic, 2015				•
Subotic et al., 2014	•			•

Table 1. Concept matrix.

4.2 Evaluation

In this dimension, I describe the evaluation of the model. First, I highlight the advantages and, afterward, the limitations of the DV found in the analyzed literature.

Rapid parallel data loads: The DV approach enables parallel loading due to minimized dependencies and data separation. This means the tables can be loaded simultaneously (Naamane and Jovanovic, 2016). Subotic et al. (2014) claim that parallel loading enables faster loading time and reduction of time costs, as well as other computational resources. Another advantage is that parallel loadings allow the scalability of the ETL process and the expansion of integrated data sources (Naamane and Jovanovic, 2016). Gluchowski (2021) also recognizes parallel loading as an advantage and notes that the data loading follows simple and consistent patterns. For this reason, the loading process could be automated (Gluchowski, 2021). The automation of ETL processes would reduce effort (Gluchowski, 2021).

Easy integration of multiple sources: Naamane and Jovanovic (2016) claim that the DV model guarantees an easy integration of various data sources. Since a business object can have multiple Satellites, one Satellite can represent one data source (Naamane and Jovanovic, 2016). Krneta and Krneta (2020), who conducted a practical study designing a DW in the energy supply industry, highlight this advantage since the easy integration of multiple sources and clear separation of the sources are especially crucial in their case.

Auditability: According to Naamane and Jovanovic (2016), the DV method helps businesses to answer auditability-related questions due to metadata in the components. The DV stores the information, where, when, and by which process a particular data asset was extracted (Naamane and Jovanovic, 2016). This way, the approach provides detailed audit information and enables data traceability (Schnider et al., 2014). Moreover, preserving all historical data allows a permanent system of record (Naamane and Jovanovic, 2016). The detailed audit information makes the DV model especially interesting for high-security and safety environments, including financial, government, and extensive scientific data warehouses (Jovanovic et al., 2012).

Flexibility: Gluchowski (2021) describes the DV model as flexible and adaptable. Due to the data separation, the data model can be easily adapted to business or data source changes (Schnider et al., 2014). Bojicic et al. (2016) and Schnider et al. (2014) explain that this advantage is based on the fact that any changes are implemented through the addition of further components. Hence, changes expand the model and do not require reconstructions, guaranteeing architectural stability and flexibility (Bojicic et al., 2016). Schnider et al. (2014) and Bojicic et al. (2016) illustrate some possible change scenarios and describe the according reactions of the model, which are listed in the table below. As one can notice, these scenarios show that changes are implemented through the expansion of the model and do not impact the core model (Schnider et al., 2014).

Change Scenario	*The DV reaction*
A business object should store a new addition attribute	Expansion of the model with a new Satellite, which stores the new attributes (Schnider et al., 2014)
A new business object or a relationship should be added to the model	Adding a Hub or a Link component into the model (Schnider et al., 2014)
Changing the cardinality of a relationship in a data source	No change of the model is required; a Link component represents a many-to-many relationship (Bojicic et al., 2016)

Table 2. The table lists possible change scenarios in a DW and specifies how the DV model would react to this change.

Data integrity: The DV model can ensure data integrity in the DW (Subotic et al., 2014). Naamane and Jovanovic (2016) claim, for example, that the data of different systems, which are placed in different Satellites, could be analyzed and compared. A simple report comparing the values of the Satellites would show the differences in the values and uncover data quality issues (Naamane and Jovanovic, 2016). Schnider et al. (2014) propose adding a Satellite that represents data quality as another method of quality issues solutions. Furthermore, the DV method does not distinguish between "bad" and "good" data because all the data is stored at all times, regardless of whether they are adaptable to the business rules (Subotic et al., 2014). The completeness of the data avoids loss of information and provides business users with "a single version of the fact" as well as "a single version of the truth" (Subotic et al., 2014).

Support of agile project management: According to Schnider et al. (2014), the possibility of independent implementation of changes makes the model suitable for agile environments. Naamane and Jovanovic (2016) also highlight this advantage, claiming that DV projects have short release sprints that can result in a production release every 2 or 3 weeks.

Although, the model also has disadvantages, which I recognized by analyzing the papers.

Complexity: Schnider highlights the high complexity of the model, which requires a deep understanding of the modeling and the business context (Schnider et al., 2014). As described in the Background, the DV consists of many components due to the model denormalization. The constant additions due to business requirement changes lead to an even higher number of components and increase the complexity of the model (Schnider et al., 2014).

Inefficient querying: Furthermore, a high number of tables lead to inefficient direct querying of the core DW (Schnider et al., 2014). The reason lies in the computationally expensive join operations that must be executed to connect multiple tables (Schnider et al., 2014). This was shown in a case study conducted by Grigoriev et al. (2021), who compared two DW models, namely the DV and a dimensional model, implemented in the Hadoop architecture. In their case study, they performed multiple queries on both models and showed that the performance of the DV model was worse in the complex queries (Grigoriev et al., 2021). As well as Schnider et al. (2014), Grigoriev et al. (2021) lead the worse performance back to the computationally difficult join operations. Naamane and Jovanovic (2016) claim that the querying performance is the reason why the DV model should be seen and implemented in the core DW as a system of record. For the reporting and end-user accessible side of the DW, they recommend an additional layer for which the dimensional model is the most suitable (Naamane and Jovanovic, 2016).

Data Delivery to the representation layer: Schnider et al. (2014) claim that transforming the DV model to a dimensional one to implement a representation layer is highly complex and costly. It is especially challenging in the incremental load of fact tables, which requires joining multiple tables of the DV (Schnider et al., 2014).

All in all, the DV model addresses the main DW requirements mentioned in the introduction of the paper. This approach benefits from the dependencies minimization and rapid load opportunities, enabling simplified and scalable ETL transformations (Naamane and Jovanovic, 2016). Due to the data separation, the DV model can provide flexibility and scalability, enabling easy change integration (Gluchowski, 2021; Bojicic et al., 2016; Schnider et al., 2014). Another beneficial characteristic is the component metadata which provides a reliable audit foundation and traceability (Naamane and Jovanovic, 2016). Changes implemented only through expansions make the model suitable for historization (Naamane and Jovanovic, 2016).

Although, the inefficient querying and the resulting need for an additional presentation layer require effort and time (Schnider et al., 2014). Although according to Gluchowski (2021), there is potential for savings in other areas due to the flexibility and scalability. Naamane and Jovanovic (2016) claim that the DV approach is most suitable in environments with high compliance demands and frequent changes.

4.3 Conceptual Data Vault

Since the DV model was developed in the industry, it mainly focuses on logical and physical design (Jovanovic and Bojicic, 2012). For this reason, there is no formalized conceptual model (Jovanovic and Bojicic, 2012). That’s why they propose an explicit Conceptual Data Vault (C-DV) model, which allows the DW developers to start the modeling with a platform-independent conceptualization (Jovanovic and Bojicic, 2012).

The Conceptual Data Vault (C-DV) extends the original DV approach and creates its abstraction (Jovanovic and Bojicic, 2012). Its purpose is to help the developers to represent and define the data and information requirements early, to develop transparent relationships between the primary data entities, and to analyze existing data sources (Jovanovic and Bojicic, 2012). According to Jovanovic et al. (2012), the key value of the proposed C-DV is to streamline the development of the DW and the architectural and requirement analysis. Moreover, the authors claim that the C-DV could be used for automatic transformation to the logical DV model (Jovanovic and Bojicic, 2012). Additionally, they present some patterns and rules for transforming a conceptual data source model into a conceptual DV model (Jovanovic and Bojicic, 2012).

In their paper, they present a metadata model of the C-DV (Jovanovic and Bojicic, 2012). The presented metamodel is pictured in the Figure below. The metadata model distinguishes between the CDV-Concepts and P-Concepts (Jovanovic and Bojicic, 2012). All CDV-Concepts are situated in a context of Time, Place, and System-Subject, according to a Convention (such as a law, standard, etc.) (Jovanovic and Bojicic, 2012). The P-Concepts are primary CDV-Concepts, representing notions of interest, namely a Hub or a Link (Jovanovic and Bojicic, 2012). A Hub must have precisely one unique identifier, namely a natural key which may be functionally substituted or complemented by a surrogate key in the logical model (Jovanovic and Bojicic, 2012). Links in the C-DV may represent one of the following relationship types: Dependency, Generalization, or Association (Jovanovic and Bojicic, 2012). The modeling rules of a C-DV mainly conform to the original DV modeling rules except for the following differences (Jovanovic and Bojicic, 2012). The first difference is the 6NF restriction, according to which all Satellites contain only one attribute (Jovanovic and Bojicic, 2012). The 6NF compliance eliminates updating dependencies (Jovanovic and Bojicic, 2012). The second difference is the possibility of a Link connecting to another Link (Jovanovic and Bojicic, 2012).

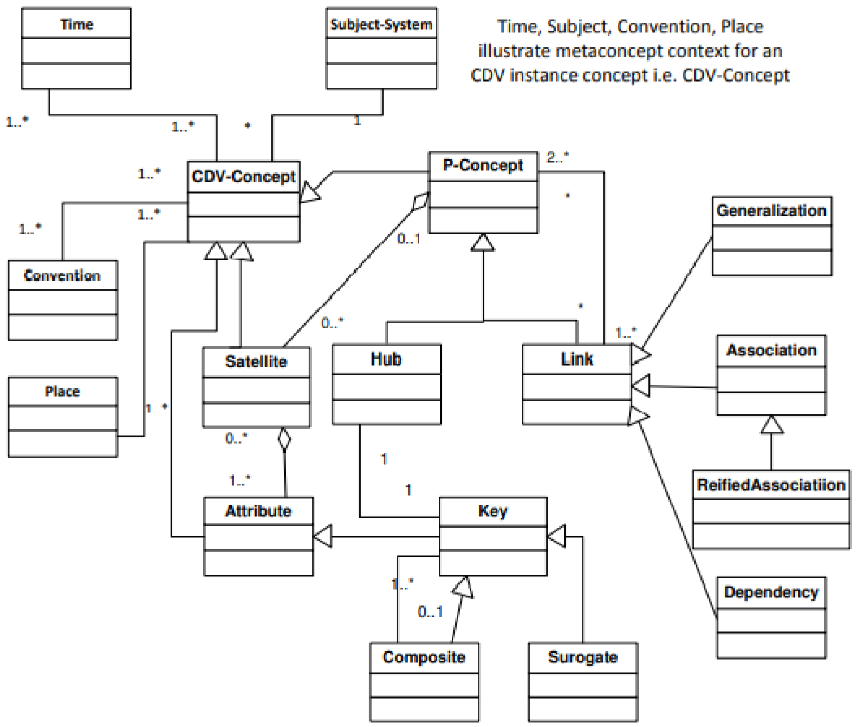

Figure 2. The model of the conceptual Data Vault (Bojicic and Jovanovic, 2012).

To sum up, the C-DV is an abstract representation suitable for analyzing information requirements and available data sources, as well as modeling an integrated DW (Jovanovic et al., 2012). Besides, the described rules and patterns may serve as a guideline for creating a C-DV from conceptual source data models (Jovanovic et al., 2012).

4.4 Design Automation

Some reviewed papers deal with the design automation process in the DV context and depict automation algorithms. Krneta et al. (2014) describe an approach of generating a physical DV model based on metadata of chosen data sources, called the Physical Data Vault approach. Moreover, I examined research papers that depict the automatic transformation of a DV into a dimensional model for creating a representation layer. The approaches regarding this issue were presented by Golfarelli et al. (2016) and Krneta et al. (2016). In this dimension, I describe the mentioned automation approaches.

Krneta et al. (2014) claim that the process of designing a DW, as well as identifying or generating Hubs, Links, and Satellites, is not sufficiently automated, which is a practical problem for businesses. For this reason, they develop an approach to the automatic generation of a physical DV model (Krneta et al., 2014). The fundamental concept of the approach is the conceptualization of the data source metadata and rules for identifying and creating the component tables, namely Hub, Link, and Satellite tables (Krneta et al., 2014). During the process, the user should select potentially essential data sources and identify business keys (Krneta et al., 2014). The metadata, rules and business keys build a foundation for the automatic design of a physical DV model (Krneta et al., 2014). The process results in the generation of Hub, Link, and Satellite tables (Krneta et al., 2014).

The approach benefits from the separation of unchangeable business keys (Hubs) and evolving relationships (Links) as well as descriptive attributes (Satellites) (Krneta et al., 2014). The proposed approach is incremental since the model is expanded by adding data sources in sets or one at a time (Krneta et al., 2014). The successive expansion is possible due to the flexibility of the DV model since additions can easily integrate new components without any reconstructions (Krneta et al., 2014). The authors claim that their automatic design approach supports design performance, scalability, and agility (Krneta et al., 2014). Although, it requires some user intervention (Krneta et al., 2014).

Krneta and Krneta (2020) utilize the described approach on an example in the field of electricity supply. They highlight the agility of the solution (Krneta and Krneta, 2020). As stated by the authors, the proposed solution addresses the challenging generation of a data model for the DW, which is flexible and immune to environmental changes (Krneta and Krneta, 2020). These are some crucial requirements in their practical case (Krneta and Krneta, 2020).

Another practical issue Krneta et al. (2016) mentioned is the lack of automated transformation of a DV into a dimensional model to create a representation layer (Krneta et al., 2016). Examining the literature, I have found two approaches that address this issue.

One of the approaches is The Starry Vault approach described by Golfarelli et al. (2016). It is based on detecting functional dependencies in the initial Data Vault model (Golfarelli et al., 2016). The described method uses DV data modeling specifications to define the components necessary for the dimensional model (Golfarelli et al., 2016). According to the authors, the algorithm consists of the following steps: in the first phase, functional dependencies are determined since they are essential for building dimensional hierarchies (Golfarelli et al., 2016). Golfarelli et al. (2016) state that many-to-many relationships represented by Links hide functional dependencies, especially when they connect more than two Hubs. Hence, this requires a specific approach (Golfarelli et al., 2016). In the second step, the algorithm determines the elements which may come into question as a fact candidate and builds a draft schema for each candidate (Golfarelli et al., 2016). The last step requires user interaction since the user selects one or multiple multidimensional draft schemas (Golfarelli et al., 2016). This way, the user can eliminate unnecessary and uninteresting facts (Golfarelli et al., 2016). After measures and facts are chosen, the process derives the dimensions and generates the final output (Golfarelli et al., 2016).

The second approach is an algorithm presented by Krneta et al. (2016), which uses the metadata model and rules of the initial DV (Krneta et al., 2016). The algorithm benefits from the scalability, flexibility, and rapid data loads and is possible due to data separation (Krneta et al., 2016). The algorithm works in the following steps (Krneta et al., 2016). The first step is the conceptualization of the metadata of the DV (Krneta et al., 2016). The metadata and the rules data are loaded into the created tables (Krneta et al., 2016). This step allows the further process to automate the design of the physical model of the dimensional model with minimal user interaction (Krneta et al., 2016). The second step requires user interaction since the user should identifiy business facts and measures (Krneta et al., 2016). Next, the dimensions are derived according to the user setups (Krneta et al., 2016). In the final step, the physical dimensional model, including fact and dimension tables, is generated according to the business rules defined in the first step (Krneta et al., 2016).

As well as the approach generating a physical DV, the approaches for transformation a DV aim to improve the design performance in the DV context. Although, the processes are not fully automated since they require user intervention.

4.5 Metadata Vault

In this dimension, I describe the Metadata Vault approach proposed by Subotic et al. (2014). Subotic et al. (2014) conducted their research in the field of DW evolution. In their paper, they propose an architecture including a metadata repository designed in the DV technique. Presenting this architecture, they aim to solve the DW evolution problem, which describes the issue of permanent changes and the requirement to preserve the history of data and metadata changes (Subotic et al., 2014). Moreover, they claim their solution could be used as a complete system of records of an enterprise and the basis for data governance (Subotic et al., 2014).

The core enterprise DW of the solution architecture consists of three components: a Raw Data Vault, a Business Data Vault, and a Metadata Vault (Subotic et al., 2014). The architecture is presented in the illustration below.

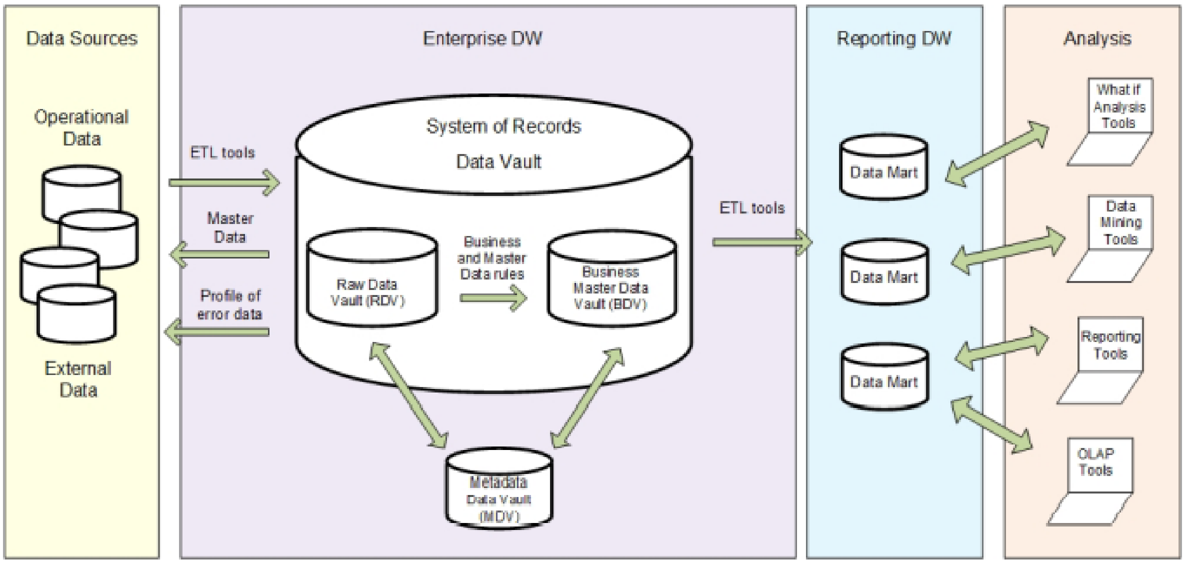

Figure 3. The solution architecture proposed by Subotic (2015).

According to their paper, the components of the enterprise DW are characterized as follows (Subotic et al., 2014). The first component is a Raw Data Vault (RDV), which includes the raw data, i.e., copies of original data that has not been transformed or changed (Subotic et al., 2014). The RDV represents the data source-oriented side of the DW (Subotic et al., 2014). The second component is a Business Data Vault (BDV) (Subotic et al., 2014). The BDV represents, on the contrary, the report-oriented side of the DW (Subotic et al., 2014). It contains the RDV data, on which the business rules have been applied (Subotic et al., 2014). Its purpose is business integration (Subotic et al., 2014). The Metadata Vault is the key component of the solution (Subotic et al., 2014). It represents a standardized metadata repository, which contains the history of changes in data and metadata (Subotic et al., 2014). Its purpose is to integrate the RDV and the BDV as well as to manage schema changes (Subotic et al., 2014).

The authors claim that they chose the DV approach due to its advantages of tracking the data back to its data source and managing the history of changes (Subotic et al., 2014). Additionally, their solution benefits from the DV advantages described in the Evaluation subchapter, such as data separation, preserving history changes, and flexibility (Subotic et al., 2014).

They expect their solution to address the DW evolution problems, such as loss of information, a lack of integration of business rules as well as a lack of mechanisms for tracking the changes in user requirements and the data source model (Subotic, 2015). Additionally, they expect that the proposed architecture will prolong the lifetime of the DW and increase the data integrity and accuracy (Subotic, 2015). Due to the modularity of the DV model, it is suitable for agile development environments (Subotic, 2015).

Although the proposed solution is not validated in the same paper, there is another paper where Naamane and Jovanovic (2017) use the Metadata Vault approach for the DW, which should contain a large amount of scientific data. Their research aims to answer the question of whether the proposed Metadata Vault can be a standard solution that can track changes, integrate big data, and ensure great scalability and flexibility (Naamane and Jovanovic, 2017). As stated in the paper, the storage and management of scientific data are challenging due to complex relationships and data types, evolving data schemas and large data volumes (Naamane and Jovanovic, 2017). Hence, the required solution must integrate large amounts of unstructured data and preserve the data origins and changes (Naamane and Jovanovic, 2017). In the paper, they deploy a prototype and test it with a developed set of changes (Naamane and Jovanovic, 2017). With their experiment, they claim to demonstrate the advantages of resolving issues and tracking the data back to its sources, preserving the history of metadata changes, and avoiding loss of information (Naamane and Jovanovic, 2017).

5 Discussion

In this chapter, I want to discuss the elaborated insights. First, I outline the commonalities and differences in the findings. In the second part, I briefly describe how businesses could benefit from this literature review. In the third section, I outline gaps for future research. Finally, I emphasize the inherent limitations of this literature research.

5.1 Identified commonalities and differences

Throughout the evaluation chapter, one can notice that there is a common understanding of the advantages and disadvantages of the DV model. Namely, the authors agree that historization and auditability ensured through component metadata can establish data quality and completeness (Subotic et al., 2014). Besides, the advantage of rapid data loads was noticed by many authors; these benefits enable the scalability of ETL processes and reduction of computational power (Gluchowski, 2021). Additionally, there is consensus about the advantage of flexibility and scalability due to the minimized dependencies and normalization of the model. Although, the flexibility is a tread-off with the complexity and performance of the model (Schnider et al., 2014; Naamane and Jovanovic, 2016). The authors notice both as shortcomings of the DV approach. All in all, I have not seen any contradictions in the current research regarding the evaluation of the model.

Examining the dimensions described after the Evaluation subchapter, I noticed that the developments and use of the DV model, such as the Conceptual DV, design automation, or the Metadata Vault approach, benefit from the portrayed advantages, i.e., flexibility and scalability. They all have in common that they aim to facilitate using the DV modeling technique or improve the DW performance.

The conceptualization and abstraction of the DV may help with the information requirement and data source analysis (Jovanovic et al., 2012). On the contrary, automatic generation may improve design performance (Krneta et al., 2014; Krneta et al., 2016).; One can notice that the described automation techniques rely on metadata or formal rules for generating a physical model (Krneta et al., 2014, Golfarelli et al., 2016; Krneta et al., 2016). They also have in common that they require user interaction for defining crucial concepts, e.g., business keys in the Physical DV approach or facts and measures in the transformation to the dimensional model. The two outlined approaches for generating a representation layer have similar steps but different foundations: functional dependencies (Golfarelli et al., 2016) or the DV metadata (Krneta et al., 2016). Compared to other dimensions, the described Metadata Vault approach not only aims to improve the performance of the DW based on the DV model (Subotic, 2015). This approach also seeks to solve the evolution problem, preserving the entire schema and metadata history of the DW (Subotic, 2015).

To sum up, the described dimensions are built upon the advantages of the model. They have various purposes but address the same issue, namely the performance of the DW and the fulfillment of its requirements. I have not identified any contradictions describing the developments of the DV model. Moreover, I did not identify the use of these concepts as mutually exclusive. They could rather be used complementary.

5.2 Management Value

This paper may offer businesses valuable information by providing an overview of DV modeling. The Background of the paper highlights the basics which are essential for the overall understanding of the model and how the DV components relate to the business concept. Besides, the actual literature review emphasizes the advantages and shortcomings of the model, which can help businesses consider implementing the DV approach. Furthermore, the introduction of further developments of the model, such as conceptualization and design automation potentials, may also be helpful. The conceptual model may streamline the DV design and data source analysis. The automation practices could help businesses to gain a design performance advantage. Finally, companies could also benefit from the Metadata Vault repository, which aims to address the evolution problem and preserve metadata history.

5.3 Research Gaps

Conducting this literature review, I have recognized the research gaps, which I describe in this subchapter. Outlining the gaps, I want to provide a clear overview and inspire future scholars to examine the presented issues.

In this literature review, I described two approaches that transform the DV into a dimensional model. Although, there are no practical studies that evaluate the approaches. Hence, there is no evidence of which model is more favorable. One could compare the described methods and evaluate them. This could be the foundation of the further development of the approaches or the invention of new automatic techniques.

Additionally, the authors of the Metadata Vault approach claim that their solution may improve Data Governance. However, this was not provided with evidence either. I want to motivate other scholars to examine this issue and how this could be achieved.

Finally, there needs to be research on the data delivery to the representational layer. For this reason, questions about the implementation, efficiency, and issues of this process arise. Future researchers could study the challenges and performance of this procedure. In the Evaluation chapter, I mentioned that data delivery was highlighted as a limitation of a Data Vault. Thus, future researchers could examine how the data delivery to the dimensional model impacts the overall performance of the DW architecture based on the DV model.

5.4 Limitations of the Review

It is important to highlight that this literature review faces certain limitations. First, I aimed to provide a broad overview of different dimensions connected to the topic, so this literature review does not offer a deep, detailed analysis of the technicalities of each dimension. Second, I chose not to review any books and to concentrate on the current research. Hence, I chose academic papers published in journals or peer-reviewed conferences. Since the DV was developed in the industry, there is a limited amount of research papers and, thus, perspectives on this topic. Finally, I considered only papers in English and German, which may have eliminated other relevant research written in different languages.

6 Conclusion

In this literature review, I provided an overview of the current research on the topic of Data Vault modeling. This industry-developed modeling technique offers multiple advantages for the DW, providing flexibility, scalability, and agility. However, this does not represent an ideal solution and has its limitations, such as querying inefficiency and high complexity. The DV was developed further in the research, providing a conceptual model that could guide DW developers. Moreover, there are design automation opportunities, which may facilitate the DV design. Finally, using the DV for the Metadata Repository in the DW may also solve the DW evolution problem and provide data governance.

Writing this paper, I aimed to contribute to the bigger picture of a rather broad topic and to provide a state-of-the-art overview that may be helpful for further research, as well as business managers and Data Warehouse developers interested in the Data Vault modeling approach.

References

Bojičić, I., Marjanović, Z., Turajlić, N., Petrović, M., Vučković, M. and Jovanović, V., 2016, May. A comparative analysis of data warehouse data models. In *2016 6th International Conference on Computers Communications and Control (ICCCC)* (pp. 151-159). IEEE.

Breslin, M., 2004. Data warehousing battle of the giants. *Business intelligence journal*, 7, pp.6-20.

Gluchowski, P., 2021. Data Vault as a Modeling Concept for the Data Warehouse. In *Engineering the Transformation of the Enterprise* (pp. 277-286). Springer, Cham.

Golfarelli, M., Graziani, S. and Rizzi, S., 2016, August. Starry vault: Automating multidimensional modeling from data vaults. In *East European Conference on Advances in Databases and Information Systems* (pp. 137-151). Springer, Cham.

Golfarelli, M. and Rizzi, S., 2009. *Data warehouse design: Modern principles and methodologies*. McGraw-Hill, Inc. (pp. 3-7).

Grigoriev, Y., Ermakov, E. and Ermakov, O., 2017, November. Hadoop/Hive Data Query Performance Comparison Between Data Warehouses Designed by Data Vault and Snowflake Methodologies. In *International Conference on Modern Information Technology and IT Education* (pp. 147-156). Springer, Cham.

Hultgren, H., 2012. Data Vault modeling guide: Introductory guide to data vault modeling. *Genessee Academy*, USA. Available at: https://hanshultgren.files.wordpress.com/2012/09/data-vault-modeling-guide.pdf. (Accessed: 10.12.2022)

Jovanovic, V. and Bojicic, I., 2012. Conceptual data vault model. In *SAIS Conference, Atlanta, Georgia: March* (Vol. 23, pp. 1-6).

Jovanovic, V., Bojicic, I., Knowles, C., Pavlic, M. and Informatike, O., 2012. Persistent staging area models for data warehouses. *Issues in Information Systems*, 13(1), pp.121-132.

Krneta, D., Jovanović, V. and Marjanović, Z., 2014. A direct approach to physical Data Vault design. *Computer Science and Information Systems*, 11(2), pp.569-599.

Krneta, D., Jovanovic, V. and Marjanovic, Z., 2016. An approach to data mart design from a data vault. *INFOTEH-Jahorina BiH*, 15.

Krneta, D. and Krneta. S., 2020. Data Vault as a Decision Support Platform for an Electricity Supplier in the Open Electricity Market. *Journal of Mechatronics, Automation and Identification Technology*, 5(1), pp. 28-31.

Linstedt, D., 2002. Data Vault Series 1 – Data Vault Overview. Available at: https://tdan.com/data-vault-series-1-data-vault-overview/5054. (Accessed: 10.12.2022).

Linstedt, D., 2005. Data Vault Series 5 – Loading Practices. Available at: https://tdan.com/data-vault-series-5-loading-practices/5285. (Accessed: 10.12.2022)

Linstedt, D. and Olschimke, M., 2015. *Building a scalable data warehouse with data vault 2.0*. Morgan Kaufmann.

Naamane, Z. and Jovanovic, V., 2016. Effectiveness of Data Vault compared to Dimensional Data Marts on Overall Performance of a Data Warehouse System. *International Journal of Computer Science Issues (IJCSI)*, 13(4), p.16.

Naamane, Z. and Jovanovic, V. 2017. A Meta Data Vault Approach for Evolutionary Integration of Big Data Sets: Case Study Using the NCBI Database for Genetic Variation. *International Journal of Computer Science and Information Technology*, 9, pp. 79-96.

Shin, B., 2003. An exploratory investigation of system success factors in data warehousing. *Journal of the association for information systems*, 4(1), p.6.

Schnider, D., Martino, A. and Eschermann, M., 2014. Comparison of data modeling methods for a core data warehouse. *Trivadis, Basel.*

Subotic, D., 2015. Data warehouse schema evolution perspectives. In *New Trends in Database and Information Systems II* (pp. 333-338). Springer, Cham.

Subotic, D., Jovanovic, V. and Poscic, P., 2014. Data Warehouse and Master Data Management Evolution - A Meta-Data-Vault Approach. *Issues in Information Systems*, 15(2), pp.14-23.

Webster, J., and Watson, R.T., 2002. Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, pp. xiii-xxiii.

Wolfswinkel, J.F., Furtmueller, E. and Wilderom, C.P., 2013. Using grounded theory as a method for rigorously reviewing literature. *European journal of information systems, 22*(1), pp.45-55.