sciendo

Data Vault Modeling: Insights from Industry Interviews

Andreea VINES
The Bucharest University of Economic Studies, Bucharest, Romania
andreea.vines@csie.ase.ro

Laura TANASESCU
The Bucharest University of Economic Studies, Bucharest, Romania
laura.tanasescu@csie.ase.ro

Abstract. *As organizations face an increasingly expending data landscape, the selection and implementation of effective data modeling techniques plays a crucial role in shaping successful data warehousing strategies. The purpose of this paper is to investigate the details of data modeling methods, with a special focus on the Data Vault approach. The case study was conducted by the interview method with data professionals, where the objective is to gather some valuable insights into design considerations, challenges, scalability and performance and adaptability. In conclusion, this paper offers a comprehensive perspective on optimizing data warehousing through a synergy of advanced data modeling techniques, specifically Data Vault, and the judicious application of interview methodology. The insights presented aim to guide researchers, practitioners, and organizations towards more effective and user-oriented data modeling practices in the constantly evolving landscape of data management.*

Keywords: Data Modeling, Data Vault, Data Solutions, Industry Interviews.

Introduction

In the digitalization period, data can be found anywhere and most of the companies depend on it and want to use it in different ways to gain insights (Margulies, 2015). This trend to use data has led to an increase in volume, speed, and diversity of the past decades, hence data modeling methodologies require to be adjusted to keep track of all those changes. For example, considering the high volume and the variety of data, it became harder to only use structured data, hence companies tend to be encouraged to use semi-structured or unstructured data. Nevertheless, traditional data warehouses are only designed for structured data, so they will require various changes to adapt to the new trends. Hence, to address all those challenges, new storage technologies are developed around cloud storing options. Moreover, the traditional data modeling technologies, including dimensional modeling, tend to be adjusted to focus on data diversity and agility. Companies and data projects are now focused on agile solutions and development, so it is required to find new methodologies to adapt to the new way of developing. Another important aspect along the amount of data is the data processing power, which plays a crucial role especially if real time or near real time data needs to be integrated. One such technique is Data vault, which was developed in the early 2000s by Dan Linstedt. Its main objective is to address all challenges of dimensional modeling and to come up with a new flexible data warehouse (Vines & Samoila, 2023). Most of the big companies are embracing data projects, which depend on a data modeling technique as well, to adapt to the digital era and to also position themselves for a sustained growth and competitiveness in this increasingly data-centric business environment.

The purpose of this paper is to gain a comprehensive understanding of Data Vault methodology and explore its practical application in data projects development, as more and more

DOI: 10.2478/picbe-2024-0292

companies are investing their data in analytical projects. The current paper debuts with data modeling evolution, starting with dimensional modeling, including Inmon and Kimball methodologies, and extending to Linstedt's methodology, known as Data Vault. To facilitate an accurate and insightful analysis, this research uses the interview methodology.

To execute the analysis effectively, a questionnaire was designed and disseminated to a diverse group of professionals that have experience and were involved in different data development projects. These professionals were selected to provide diverse perspectives and valuable insights in choosing a data modeling technique that corresponds to business requirements.

The interviews were structured as online discussions, each participant responding to a set of predefined questions tailored to the research objectives. The questionnaire is organized into various sections. It begins with a series of control questions to gain information about the respondents' background and experience. Subsequent questions were divided into areas such as design configurations, data quality and consistency, scalability and performance, adaptability, and integration with reporting layers. The final section is dedicated to understanding any lessons learned or recommendations from the data professionals.

The current paper aims to respond to the following research question: *How does Data Vault modeling differ from Dimensional Modeling in real life projects and how does it respond to various challenges?* In terms of the overall structure, the current paper is organized into the following sections: an introduction, a comprehensive literature review that covers data modeling methodologies and the evolution of Data Vault, a detailed methodology to describe the interview method used and the questionnaire designed for this research purpose, an analysis of the results obtained from the interview, and finally the conclusions with key findings and potential directions for future research.

Literature review

Data evolution over time

The increase in data volume, velocity, and variety over the last decades has brought big data to the forefront of contemporary discussions. Companies are eager to invest their data in analytical projects that will help them to gain insights regarding their business evolutions; it can include multiple types of analysis such as descriptive or predictive. This transformation comes with different challenges to traditional data warehouses, originally designed for structured data, as they try to adapt to the influx of semi-structured data. Although these warehouses can be modified to accommodate these changes, they often lack support for evolving capabilities and analytical demands.

In response to these challenges, innovative data warehouse techniques have emerged. One of these techniques is Data Vault, specifically designed to handle diverse data sources. Unlike traditional warehouses, Data Vault excels in maintaining the integrity of the original design while addressing the complexities brought about by the expanding data landscape.

This section will focus on a comprehensive comparison between dimensional data modeling and Data Vault techniques and will reveal their respective strengths and limitations in the face of today's dynamic data environment.

Dimensional Modeling

Inmon is considered the father of data warehousing (Dennis, 2007). Data warehouse is defined as a subject-oriented, integrated, non-volatile and a time-variant collection of data of which goal is decision-making support (Song, 2009):

- Subject oriented as it is organized around subjected rather than the functions or processes of the organization; some examples are customer, products, time which are known as dimensions.
- Integrated as it can include data from multiple origins into a cohesive and unified dataset.
- Non-volatile as it keeps a consistent and unchanging perspective of its data.
- Time-variant as every piece of information retrieved from the data warehouse is invariably perceived within a distinct time context, offering a temporal dimension to analytical insights.

On the other hand, according to Kimball et al. (1998), who is known as the creator of data, a data mart is defined as a "restriction of the data warehouse to a single business process or to a group of related business processes targeted towards a particular business group". He also defines a data warehouse as a union of all the constituent data mart.

Ralph Kimball is the pioneer of dimensional modeling, which was considered an enhancement over 3NF models in numerous aspects, primarily praised for its advantages in terms of ease of comprehension and reduced need for joins between relational tables. This, consequently, led to enhanced query performance (Kimball et al., 2008).

Dimensional modeling is centered on two key components – facts and dimension tables. The model contains a fact table, which houses facts, and it is surrounded by different tables which houses attributes (as unique identifiers for different subjects such as customers, products, time). Fact tables are used to define the performance measurements of the organization (as sales, stocks, revenue).

According to Kimball et al. (2008), dimensional modeling is designed in four steps – which requires to start with choosing the data mart and declaring the grain; after these two steps are completed, the next ones require to identify the dimensions and facts. Both Kimball & Inmon methodology is based on the dimensional modeling approach, the difference between them being the implementation structure – Inmon's approach is a top-down approach, which considered a centralized data warehouse as a single source of truth which integrated data from multiple areas (Inmon et al., 2008); where Kimball's approach is a bottom-up one which emphasizes building data marts firsts (which are smaller and more focused subsets of data for specific business areas) and integrate all those data marts into a comprehensive data warehouse (Kimball et al., 2008).

Data Vault Modeling

Data vault architecture was released by Dan Linstedt as an alternative to standard dimensional modeling to handle the necessity of a new model that is scalable in the context of big data.

It is designed as a hub and spoke model, where data flows from a source system to a central repository called "hub" without any transformations or modifications and then it is distributed to smaller data structures, called "spokes" which are like data marts, used for reporting and analysis purposes. This modeling type involves replacing the traditional fact & dimension tables with hubs, satellites, and links.

A hub serves as a representation of a core business concept (as vendors, customers, products), embodying essential information such as the business key and metadata related to the

load timestamp within the table. The entrances of an entity are deliberately stored in a satellite table, which is intricately connected to the hub table through a key mechanism. This key, typically generated as a hash key, facilitates the linkage by joining the primary key from the hub with the corresponding foreign key present in the satellite tables. This approach ensures a structured and cohesive data model, allowing for efficient navigation and retrieval of comprehensive details associated with the business concept (Vines & Samoila, 2023).

However, this architecture can be easily adjusted to perform business needs – for example, bridge tables were defined in data vault architecture to address performance issues caused by many joins that might result from the model. Linstedt considers that bridge tables are not part of the data vault architecture, but they can be used if it is needed to improve performance – the mechanism involves in migrating several hub entity foreign keys from two or more link entities into a single table, required to be used in the reporting layer (Linstedt et al., 2009).

Data Vault main characteristic is flexibility as it allows for the integration of new data sources with only minor changes to the existing tables. Another important characteristic can be considered the loading efficiency using parallel processing as there are not that many dependencies with auditability as it keeps track of all changes made to a source system and allows keeping the history of the data (Giebler et al., 2019).

According to Linstedt (2003), data vaults can be modelled using a four-step process, which requires modeling the hubs first, followed by links and satellites. The latest step is to normalize the satellites if it is necessary.

To summarize the distinction between data modeling methodologies, Smith and Elshnoudy (2003) and Giebler et al. (2019) identified some key differences between them which can be observed in Table 1.

Table 1. Data modeling methodologies

Methodology / Criteria	Inmon	Kimball	Data Vault
Data models	Dimensional modeling (Facts & Dimensions)	Dimensional modeling (Facts & Dimensions)	Hubs, Links & Satellites
Scalability	Large upfront investment	Iterative agile (scalable through modular data marts)	Flexible adaptive (easy to implement agile projects).
Time and Delivery	Longer development timeline	Quicker delivery through data marts	Balanced approach, scalable and adaptable
Flexibility and agility	More rigid due to centralized design	More flexible, responsive to business needs	Agile, adaptable to changing business needs
Storage platforms	Relational databases	Relational databases	Relational & non-relational databases
Architecture style	Centralized EDW	Distributed Data marts	Hub and spoke

Source: Authors' own research adapted from (Smith & Elshnoudy, 2003) (Giebler et al., 2019).

Methodology

To respond to the research question addressed in the Introduction section, the interview method was used to collect information from various professionals with experience in implementing data driven projects. The interview focuses on analyzing the behavior of Data Vault methodology on

various topics, such as data handling, challenges, data visualization and so on, and how this technique is being applied in real projects. The current paper uses the interview as a comprehensive research method to collect information that will help evaluate data vault methodology design in different data projects. It is a qualitative methodology to collect data and accumulate knowledge on a specific topic (Alshenqeeti, 2014).

The interviewers are employees that have experience with data modeling, and they were involved in different data vault implementation projects. They all have different levels of experience and seniority, which will lead to a variety of perspectives on data modeling.

The interview is designed as a one-to-one discussion via Teams ranged from 20 to 40 minutes where different questions regarding data modeling are discussed. All discussions were recorded to extract information required for this research paper. The details for the interview structure are presented in Table 2.

Table 2. Interview details

Details	Description
Interview method	Online via Teams
Number of participants	8
Average interview duration	26
Number of questions addressed	18
Period	January 2024

Source: Authors' own research.

The interview debuts with some introductory questions to gain some details regarding the respondents – within as age, gender, role withing the organization and data-related experience. The interview structure consists of nine sections where each section contains two questions, and the last section was designed for collecting different recommendations and suggestions. All the questions discussed in the interview are presented in the Appendix.

Results and discussions

This section will present the results of the responses that were collected in the interview questionnaire presented above.

There were collected 8 responses from data professionals with various experience in modeling, implementing, and supporting data warehousing solutions, who also worked with both modeling techniques, and they were able to compare them. Those data professionals work as data architects, data engineers and data modelers and they all have more than ten years' experience each.

The first section of the interview was focused on the reasoning behind choosing to use Data Vault methodology in implementing a data project. Among all the responses collected, the participants mentioned that this methodology was used for the necessity of integrating data from multiple sources, being mentioned that best use case for it is to have around ten data sources or systems to integrate data. All the data is stored in Cloud systems, using Snowflake for data storage. Another reason mentioned was also the necessity of keeping the history of the data. Regarding the design considerations, the participants started with Data Vault 2.0 architecture and adapted it accordingly to the project requirements. Some examples would be to combine Raw & Business Vault into a single layer or to apply different techniques to correct potential data issues.

The following sections were based on diverse topics, such as data variety, scalability and performance, data quality and adaptability to changes. For data variety, most of the projects were using structured data, but there were also some instances where semi-structured data was used. However, those sources were transformed into a relational format using data ingestion tools. Some challenges faced were also caused by data formats – for example, source systems can store date columns using different formats, hence it is required to accommodate all of them and use a standard format. Data performance represents a bottle neck in Data Vault implementation, as the logic behind all the tables and the possible joins can run slow. There are also some workarounds that were applied, such as bridge tables or some views that get data from multiple sources to speed the performance.

Quality and consistency represent an important aspect in developing data projects. The participants mentioned that they were using different data modeling tools, such as Wherescape 3D, Wherescape Red or ERStudio which can create data lineage and traceability. Part of the professionals mentioned that they were using different quality test to understand data lineage and to verify the correctitude of the implementation, making sure that the business logic was respected as per the requirements received.

For the adaptability to change section, all the participants consider that Data Vault can easily adapt to changes, without making many changes to the existing flow. For example, it was mentioned that in most cases was requested to include new data sources, which was done by creating the new satellite tables and including them in the existing flow. The preferred design that was suggested by the data modelers was to include a source system with all its capabilities end to end, and after that to start adding new sources – which will highlight the adaptability and agility of implementing Data Vault projects.

The current paper aims to respond to the research question that is centered on the challenges faced in Data Vault implementations. Based on the discussions with the professionals, it was mentioned that Data Vault projects can also have various challenges, such as any data projects. They have encountered different challenges, starting with data issues and the incapacity to model some entities or to find their business keys. These difficulties were addressed to the business, and they engaged in various discussions with the team to solve them. Some data issues were also corrected by applying transformations in the ETL logic. Additionally, numerous challenges appeared due to the underestimated period for testing, which led to project delays. Furthermore, the participants observed that the lack of knowledge in implementing Data Vault which caused some challenges.

The final section, lessons learned, was added for participants to express their opinions regarding Data Vault implementations. They consider that it represents a very comprehensive modeling technique, but it is important to have previous knowledge of how to implement it and its characteristics. Moreover, the implementation team must be prepared and have consolidated know-how to successfully deliver the project. Furthermore, it is an implementation that takes time and requires solid effort, which is why it is more suitable for clients or companies that have a large volume of data, needing to retrieve it from many systems. Overall, they consider that Data Vault brings benefits in project implementations, and it is being used more and more. Obviously, this cannot completely replace dimensional modeling, but certain layers in the project design are more suitable for Data Vault. For the reporting layer, it is necessary to switch from Data Vault to dimensional modelling as it can cause performance bottlenecks in loading data into reports which can affect end-user experience.

Conclusion

The current paper was designed to gain some insights into data modeling techniques and approaches, focusing on Dimensional and Data Vault modeling techniques. It was observed that the Data Vault is often used more in real data projects, but it also has different advantages and disadvantages.

For example, according to the collected responses, technical expertise represents a gap in the implementation because it requires knowledge-based professionals to implement this solution. However, this challenge can be expected considering that the technology is relatively new and not so often implemented as Dimensional modeling. On the other hand, even in Data Vault implementation, Dimensional Modeling is still necessary for the reporting layer – where Linstedt technique might be faced with performance problems. Our findings using the interview method clarified that Data Vault represents a useful method to model the data and gain insights, being designed for systems that have multiple sources of data.

Therefore, the selection of the appropriate methodology to implement a data project depends on the specific needs of the project, such as budget, timeline, amount of data required to be integrated and any specific needs in terms of technologies required to be used. However, as in any project implemented, there can be multiple methodologies or modeling techniques aggregated to respond to the requirements and needs.

The current work has some limitations that include the sample data, as it has been collected from less than ten responses. Furthermore, another possible limitation can be the bias, considering that all the responses were collected from employees that work in multinational companies in Romania. However, these two limitations should not have a significant impact on the results, since all professionals have worked on different project implementations, and they have a lot of years of data experience.

The current research can also be extended to do a comparative analysis across different domains to check if there are any industry-specific challenges or best practices which can be presented. Moreover, another area that can be explored would be to investigate the integration of various cloud technologies and compare by different criteria to see which one offers the most advantages in using it for Data Vault implementations. Another opportunity would be to extend the interview questionnaire to a wider group of participants and analyze whether the results will be significantly different. All these extensions aim to provide deeper insights into the aspects of Data Vault implementations, informing strategic decision-making process and driving innovation in the data management practices.

This paper is based on the responses provided by data professionals who have presented their experience regarding the use of Data Vault or Dimensional modeling projects. This study can serve as a foundation for other research works, providing a validation and an extension of knowledge through the interpretation of the results obtained from the interview. Additionally, the study can be used as an input for identifying current implementation problems or challenges and finding solutions and strategies to optimize and improve existing processes.

References

Alshenqeeti, H. (2014). Interviewing as a Data Collection Method: A Critical Review. *English Linguistics Research*, 39-45.

Dennis, A. L. (2007, January 4). *The Data Warehouse: From the Past to the Present*. Retrieved from Dataversity: https://www.dataversity.net/data-warehouse-past-present/

Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2019). Modeling Data Lakes with Data Vault: Practical Experiences, Assessment, and Lessons Learned. *Proceedings of the 38th Conference on Conceptual Modeling (ER 2019)*, (pp. 90-99).

Inmon, W., Strauss, D., & Neushloss, G. (2008). *DW2.0: The Architecture for the Next Generation of Data Warehousing*. New York: Elsevier.

Kimball, R., Reeves, R., M., R., & Thornthwaite, W. (1998). *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*. New York: Wiley.

Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2008). *Data Warehouse Lifecycle Toolkit: Practical Techniques for Building Data Warehouse and Business Intelligence Systems (2nd ed.)*. New York: Wiley.

Linstedt, D. (2003, January 1). *Data Vault Series 2: Data Vault Components*. Retrieved from The Data Administration Newsletter: http://www.tdan.com/viewarticles/5155/

Linstedt, D., Graziano, K., & Hultgreen, H. (2009). The Business of Data Vault Modeling, 2nd Edition. *Proceedings of the 2009 International Conference on Information and Knowledge Engineering (IKE)*, Las Vegas, NV, USA, (pp. 232-238). Las Vegas, NV, USA.

Margulies, J. (2015, October). *Data as Competitive Advantage.* Retrieved from Winterberry Group: https://www.iab.com/wp-content/uploads/2015/10/Winterberry-Group-White-Paper-Data-as-Competitive-Advantage-October-2015v2.pdf

Smith, J., & Elshnoudy, I. A. (2003). *A Comparative Analysis of Data Warehouse Design Methodologies for Enterprise Big Data and Analytics. Emerging Trends in Machine Intelligence and Big Data*, 16–29.

Song, I. (2009). *Data Warehouse. Encyclopedia of Database Systems*, 657–658.

Vines, A., & Samoila, R. (2023). A*n Overview of Data Vault Methodology and Its Benefits. Informatica Economica*, 15-24.

Appendix

Questionnaire format:

Area	Question
Reasons for Choosing Data Vault	What factors influenced the decision to use Data Vault in your data warehouse implementation?
	Were there specific business use cases or data characteristics that favored this decision?
Design Considerations	How did the original design of your data warehouse progress during the implementation process?
	What were the key considerations in designing the architecture for Data Vault?
Handling Data Variety	How did address the challenges posed by diverse data sources and formats?
	Can you share examples of how the chosen technique facilitated the integration of semi-structured or unstructured data?
Scalability and Performance	How did the scalability and performance of your data warehouse differ between the Data Vault and Dimensional Data Modeling approaches?
	Were there specific performance considerations or bottlenecks encountered with either technique?
Data Quality and Consistency	How did each approach ensure data quality and consistency within the data warehouse?
	Were there differences in how data lineage and traceability were maintained?
Adaptability to Changing Requirements	How well did each approach adapt to changes in business requirements or data sources over time?
	Were there instances where one technique proved to be more flexible or agile than the other?
Challenges Faced	What were the main challenges encountered during the implementation process?
	How were these challenges addressed or mitigated?
Integration with Analytics and Reporting	How did each approach impact the ease of creating analytics and reports for end-users?
	Were there notable differences in the reporting capabilities between Data Vault and Dimensional Data Modeling?
Lessons Learned and Recommendations	Based on your experience, what lessons have you learned from implementing both Data Vault and Dimensional Data Modeling?
	Do you have any recommendations for organizations considering these approaches for their data warehouse projects?