The ISC Int'l Journal of Information Security

August 2019, Volume 11, Number 3 (pp. 131–137)
http://www.isecure-journal.org

Selected Paper at the ICCMIT'19 in Vienna, Austria

Critical Success Factors for Data Virtualization: A Literature Review☆

Matthias Gottlieb [1,*], Marwin Shraideh [1], Isabel Fuhrmann [1], Markus Böhm [1], and Helmut Krcmar [1]

[1] *Technical University of Munich, Chair for Information Systems (i17), Boltzmannstr. 3, 85748 Garching by Munich*

ARTICLE INFO.

Keywords:
Data Virtualization, ETL, Critical Success Factors, Data Integration, Business Intelligence, Literature Review.

Abstract

Data Virtualization (DV) has become an important method to store and handle data cost-efficiently. However, it is unclear what kind of data and when data should be virtualized or not. We applied a design science approach in the first stage to get a state of the art of DV regarding data integration and to present a concept matrix. We extend the knowledge base with a systematic literature review resulting in 15 critical success factors for DV. Practitioners can use these critical success factors to decide between DV and Extract, Transform, Load (ETL) as data integration approach.

1 Introduction

Data Virtualization (DV) has become an important method to store and handle data in a cost-efficient way. However, for practice, it is unclear what data should be virtualized or not. Therefore, we search for existing critical success factors to decide between DV and Extract, Transform, Load (ETL) data integration approaches. This study focuses on DV. The overall goal is developing an IT artifact that supports in deciding which data can or cannot be virtualized. In this study, we present our findings from our first stepâĂŤbuilding the foundation for this IT artifact by conducting a literature review according to [1]. Therefore, we follow the research question: "What are critical success factors for DV?".

2 Method

We applied the well-established framework from [2] to develop our IT artifact. First, we defined the scope of the literature review. Second, we synthesized the literature to a concept matrix. Third, we deduced critical success factors from the concept matrix. Figure 1 shows the overall research approach including our focus of this study, illustrated by the numbers 1 to 4.

We followed the approach from [1] to conceptualize relevant literature. Hereby, we focused the literature review on capabilities of DV, the application of DV, and differences between DV and ETL as data integration approach. To identify relevant publications, we used the key term "Data Virtualization" for searching through *title, abstract* and *keywords* of literature listed in the databases *AISeL, Scopus, EBSCOhost, ACM digital library, IEEE Xplore* and *Science Direct* to cover most of the possible search results related to the information systems, computer science, and math-

☆ The ICCMIT'19 program committee effort is highly acknowledged for reviewing this paper.

* Corresponding author.
Email addresses: matthias.gottlieb@tum.de, marwin.shraideh@in.tum.de, i.fuhrmann@tum.de, markus.boehm@in.tum.de, krcmar@in.tum.de

ISSN: 2008-2045

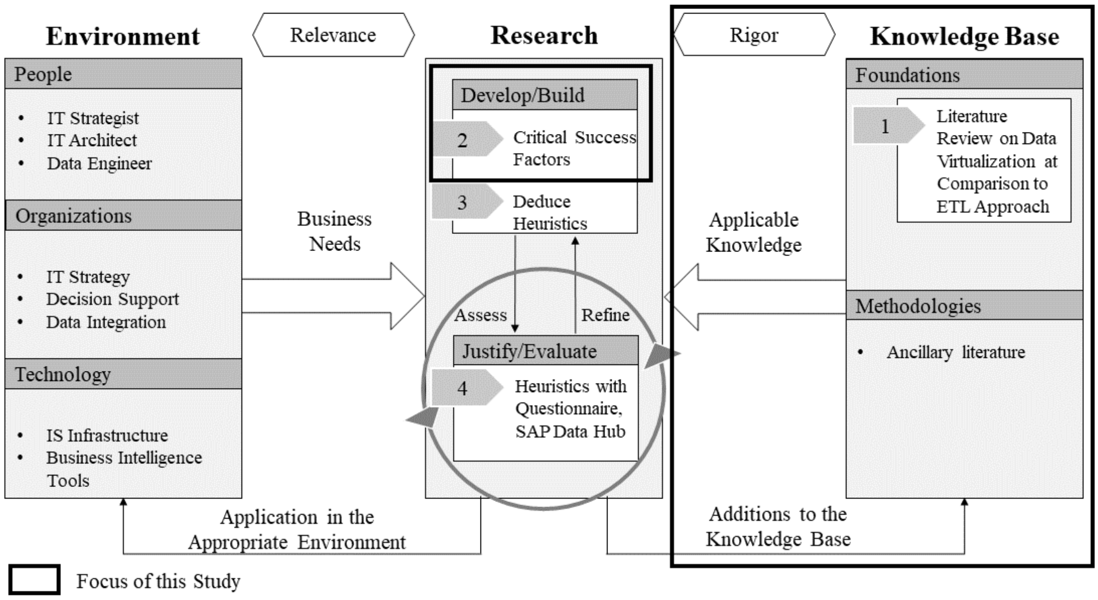

Figure 1. Overall research approach: Design science according to [2]

ematics fields of science. To enhance the knowledge base by practical insights, we included information from market leaders in DV: Informatica, TIBCO, Denodo, IBM and SAP, as identified by [3]. Afterwards, we limited the results based on title and abstract.

Next, we coded each paper to come from an author-centric to the concept-centric approach [1]. To build a concept matrix, we screened the remaining literature for generating categories from keywords mentioned in the source texts. In an iterative step, we derived the concept matrix. A publication belongs to a specific category if it deals specifically with the topic or a category thereof.

3 Results

The literature review indicated that DV is a practice-driven approach and is less addressed in scientific research.

The combination of the term DV with related concepts such as "Data Integration", "Logical Data Warehouse", or "Business Intelligence" returned no results. In total, we discovered 14 relevant publications out of 26,116 overall hits in the databases used. The review of articles from practitioners added 30 relevant blog articles, whitepapers (WhP), and case studies. In the following, we outline the identified critical success factors:

Source data quality (SDQ) determines the estimated amount of effort needed for data cleansing steps such as matching or conflict resolution. DV tools do not solve human interventions efficiently such as complicated cleansing steps [4]. or even bad data quality such as redundant data favors choosing a physical consolidation approach [5].

Transformation need (TrN) explains the issue that complicates the work-flow with multiple transformation steps decreasing DV performance massively [4, 6]. With the increasing complexity of transformation steps, DV toolsâĂŹ performance slows down.

Extend of historization (ExH) describes the versioning of data changes. DV tools map existing data records from source systems to a target schema [7]. However, physical replication is required to track changes with great extent [8, 9].

Source system availability (SSA) is the stability and the reliability of the system. It is necessary because virtualization always requires data to be stored in the source system [7].

Computing capacity (CoC) is the remaining computing power of the source system that can be utilized without without performance losses. Source system utilization is a significant criterion of computing capacity needed for effective and efficient implementation of DV [7]. DV makes additional computing capacity available [10].

Budget (Bud) is the cost framework for the project, which influences the possible actions such as developing data integration solutions [11]. Therefore, it

influences the decision on DV. DV requires reduced investment in IT infrastructure, can cause fewer implementation steps in comparison to ETL [4] and has the potential to reduce operational costs in the long-term [12].
Replication constraints (ReC) means any constraints when replicating the data is forbidden or limited due to regulations by law or the owner. In cases of any compliance or policy restrictions, where replicating data is not allowed, DV is the approach of choice [7].
Data model stability (DMS) Describes how often changes in the data model of the source system are made. A DV solution enables users to integrate changes more quickly because of the flexibility [13]. With an ETL data integration approach, it is more complicated to integrate changes, due to interdependencies of processing steps [7].
Time-to-market (T2M) is the time until a solution is ready. There are significant differences regarding the time until data is available, depending on the data integration approach and its complexity [14]. The benefits of DV is that DV supports fast development cycles to speed up the time to market of new reports and new forms of analytics in comparison to ETL [15]. DV is a considerable option for quick data access [16].
Technology freedom (TeF) describes the required flexibility to choose from many solutions of different vendors independently. DV offers the required freedom to use needed BI tools instead of physical data consolidation [17].
Agility (Agi) is the possibility to react on changes in fast-paced business environments [5] with adapting the structure of underlying source systems. DV promotes an agile business culture and provides the capability to adapt to new requirements [18]. Physical data integration restricts an overall agile BI approach [15].
Target data format (TDF) is the availability of the needed data formats in the source system. The efficiency of data integration depends on the chosen data format. DV tools can handle standard relational or hierarchical (XML) structured data [4].
Data volume (DaV) is the amount of accessed data. DV tools read and transform data on demand and process them while reading [6]. When the accumulation of large amounts of data is expected, [7] the usage of a data warehouse (DWH) approach is adviced.
Refresh intervals (ReI) is the frequency of data updates in the source system. DV enables a view on the actual source data and thereby avoids latency caused by physically replicating data [14]. DV can deliver near real-time data and can thus include intraday changes. A DWH approach with physical replication in comparison to DV works with less frequent updates, like batch jobs at the end of the day [19].
Application area (ApA) describes the analytical workload necessary to get the expected results such as for data mining or predictions. A DWH solution compared to DV is preferred for a large amount of data [4]. Table 1 presents the derived critical success factors concerning the literature in a concept matrix.

4 Dicussion

In this section, we discuss potential reasons for the identified difference between research and practice for DV. We found a small number of publications in information systems concerning DV in comparison to ETL. However, the usage of the terminology is present because of the enormous number of hits.

Practice focuses mainly on implementing applications rather than on underlying concepts. Research tends to explore solutions, test limits, optimize the development of existing technologies, or deal with the transfer of solutions from one area to another. Here, research has the potential to advance knowledge in DV. Data replication is still relevant since DV cannot deal with higher amounts of data or complex transformations with efficient performance. However, the numerous white-papers and case studies from practitioners do not offer an objective point of view. Case studies from each vendor focus on specific strengths of their tool with individual vendors being more likely to promote their market position. Therefore, we extend these findings with a neutral representation of a comparison between DV and ETL approaches.

Further, the results present 15 critical factors of success derived from literature, to distinguish between DV and ETL approaches. These factors can enhance other research areas in the computer science field and accelerate results as well as technical progress. Nowadays, integration and providing data promptly is becoming essential for organizations to stay competitive. The need for further research identified in this study can support broader knowledge in DV.

5 Conclusion and Future Work

After critically evaluating the topic, we derived 15 critical success factors for deciding between DV and ETL. These factors result from the literature review and build a basis for a future planned IT artifact to automatically give a decision support for the usage of DV. The IT artifact should supports the question of when and for what kind of data to apply DV. For practitioners, an answer to this question is essential for successful application. With the IT artifact, practitioners can exploit the potential of data integration, build their strategy and support operations adequately.

Table 1. Concept Matrix with the derived Critical Success Factors

Author(s)	Year	SDQ	TrN	ExH	SSA	CoC	Bud	ReC	DMS	T2M	TeF	Agi	TDF	DaV	ReI	ApA	WhP
Bhatti [5]	2013	x	x		x	x			x	x		x		x	x		
Bologa & Bologa [6]	2011	x	x				x			x	x	x		x	x		
Chandramouly [20]	2013											x					x
Data Virtuality [21]	2014	x	x	x			x		x	x	x	x	x	x	x	x	x
Denodo & IBM [22]	2014		x							x		x					x
Denodo [4]	2014	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Denodo [23]	2014							x			x	x					x
Denodo [24]	2016						x	x	x	x					x	x	x
Denodo [25]	2016	x	x							x		x					x
Denodo [26]	2016						x	x		x					x		x
Denodo [18]	2017						x			x							x
Denodo [27]	2017														x		x
Denodo [28]	2018								x		x	x			x	x	x
Earley [13]	2016								x			x			x		
Farooq [14]	2013									x					x	x	
Ferguson [29]	2011									x		x		x		x	x
Ferguson [16]	2011			x			x			x	x	x		x		x	x
Goetz & Yuhanna [30]	2015									x							
Grosser & Janoschek [10]	2014			x	x	x			x	x	x	x		x	x	x	x
Guo et al. [31]	2015						x			x	x						
Hopkins [32]	2011	x					x			x		x				x	
Kimball & Ross [19]	2013								x	x	x	x			x	x	
Loshin [33]	2010	x	x					x									x
Matzer & Kurze [34]	2017							x		x		x					x
Mousa & Shiratuddin [35]	2015				x						x	x			x		
Moxon [9]	2015			x											x		x
Powell [12]	2011	x					x		x	x		x					
Russom [36]	2010									x					x		x
Schroeck [37]	2012						x			x					x	x	x
Shankar [17]	2017						x			x	x	x					
TIBCO [38]	2017		x	x				x								x	x
TIBCO [39]	2017						x	x		x	x	x			x	x	x
TIBCO [40]	2017				x	x				x		x			x		x
TIBCO [41]	2018			x	x	x		x		x				x			x
Van der Lans [7]	2012	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Van der Lans [42]	2016				x	x										x	x
Van der Lans [43]	2016			x	x	x				x						x	x
Van der Lans [15]	2016							x		x			x	x	x		x
Van der Lans [44]	2017									x						x	x
Van der Lans [45]	2018			x			x	x			x	x					x
Vinay [8]	2012	x	x	x	x	x	x			x							
Voet [11]	2018						x				x				x		x
Yuhanna [3]	2017														x		
Yuhanna, Giplin [46]	2012						x								x		
Sum		10	10	11	10	9	18	11	11	30	15	23	4	10	23	17	30

We will evaluate the critical success factors by conducting expert interviews to derive a validated IT artifact. Future research can provide more in-depth knowledge in DV to support the handling of large amounts of distributed data and thereby address data integration challenges resulting from trends in big data. In addition, practitioners can exploit the potential of data integration with the IT artifact, build their strategy and support operations effectively.

References

[1] Jane Webster and Richard T. Watson. Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2):xiii–xxiii, 2002.

[2] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004.

[3] N. Yuhanna. The forrester wave: Enterprise data virtualization, q4 2017: The 13 vendors that matter most and how they stack up. *Forrester*, 2017.

[4] Denodo. Data virtualization and etl. 2014.

[5] N. D. Bhatti. Overcoming data challenges with virtualization. *Business Intelligence Journal*, (Vol. 48, No. 4), 2013.

[6] A. R. Bologa and R. Bologa. A perspective on the benefits of data virtualization technology. *Informatica Economica*, (Vol. 15, No. 4):110âĂŞ118, 2011.

[7] R. F. Van der Lans. *Data virtualization for business intelligence systems: Revolutionizing data integration for data warehouses*. The Morgan Kaufmann Series on Business Intelligence. Elsevier/Morgan Kaufmann, Amsterdam, 2012.

[8] S. Vinay. Logical data warehousing for big data: Extracting value from the data! *Gartner*, 2012.

[9] P. Moxon. Data integration alternatives. 2015.

[10] T. Grosser and N. Janoschek. Datenmanagement im wandel: Data warehousing und datenintegration im zeitalter von self service und big data. 2014.

[11] M. Voet. Data virtualization is a revenue generator, 2018.

[12] J. E. Powell. Enabling bi agility with data virtualization. *Business Intelligence Journal*, (Vol. 16, No. 4):53âĂŞ55, 2011.

[13] S. Earley. Data virtualization and digital agility. *IT Professional*, 18(5):70âĂŞ72, 2016.

[14] F. Farooq. The data warehouse virtualization framework for operational business intelligence. *Expert Systems*, 30(5):451âĂŞ472, 2013.

[15] R. F. Van der Lans. Developing a bi-modal logical data warehouse architecture using data virtualization: A whitepaper. *R20/Consultancy*, 2016.

[16] M. Ferguson. Succeeding with data virtualization: High value use cases for operational and data management data services. 2011.

[17] R. Shankar. Enabling self-service bi with a logical data warehouse. *Business Intelligence Journal*, (Vol. 22, No. 3), 2017.

[18] Denodo. Deploying data virtualization at an enterprise scale âĂŞ a journey towards an agile, data-driven infrastructure. 2017.

[19] R. Kimball. *The data warehouse toolkit: The definitive guide to dimensional modeling*. J. Wiley & Sons, [Erscheinungsort nicht ermittelbar], 3rd ed. edition, 2013.

[20] A. Chandramouly, N. Patil, R. Ramamurthy, S. R. Krishnan, and J. Story. Integrating data warehouses with data virtualization for bi agility. 2013.

[21] Data Virtuality. Der komplette guide zur datenintegration: Einfache datenintegration im digitalen zeitaler e-book. 2014.

[22] Denodo and IBM. Achieve value and insight with ibm big data analytics and denodo data virtualization. 2014.

[23] Denodo. Data virutalization goes mainstream: Solving key data integration challenges with more agility than traditional technologies for structured, unstructured, web, cloud and big data sources. 2014.

[24] Denodo. Data virtualization usage patterns for business intelligence/data warehouse architectures. 2016.

[25] Denodo. Denodo platform 6.0. 2016.

[26] Denodo. Die 10 hÃďufigsten fragen: Datenvirtualisierung. 2016.

[27] Denodo. Realizing the promise of self-service analytics. 2017.

[28] Denodo. Overcoming telecommunications challenges with data virtualization. 2018.

[29] M. Ferguson. Succeeding with data virtualization: High value use cases for analytical data services. 2011.

[30] M. Goetz and N. Yuhanna. Create a road map for a real-time, agile, self-service data platform: Road map: The data management playbook. *Forrester*, 2015.

[31] S. S. Guo, Z. M. Yuan, A. B. Sun, and Q. Yue. A new etl approach based on data virtualization. *Journal of Computer Science and Technology*, 30(2):311âĂŞ323, 2015.

[32] B. Hopkins. Data virtualization reaches critical mass: Technology advancements, new patterns, and customer successes make this enterprise technology both a short- and long-term solution. *Forrester*, 2011.

[33] D. Loshin. Effecting data quality improvement

through data virtualization. 2010.

[34] M. Matzer and C. Kurze. Datenvirtualisierung: Bindeglied zwischen verteilten datensilos zum aufbau flexibler analytischer ÃŰkosysteme. 2017.

[35] A. H. Mousa and N. Shiratuddin. Data warehouse and data virtualization: Comparative study. page 369âĂŞ372, 2015.

[36] P. Russom. Data integration for real-time data warehousing and data virtualization. *TDWI*, 2010.

[37] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano. Analytics: Big data in der praxis: Wie innovative unternehmen ihre datenbestÃďnde effektiv nutzen. 2012.

[38] TIBCO. Applying data virtualization: 13 use cases that matter. 2017.

[39] TIBCO. Ten things you need to know about data virtualization. 2017.

[40] TIBCO. Tibco data virtualization technical overview: Tibco data virtualization deployment, development, run-time, and management capabilities. 2017.

[41] TIBCO. Data virtualization: Achieve better business outcomes more quickly. 2018.

[42] R. F. Van der Lans. Designing a data virtualization environment a step-by-step- approach: A technical whitepaper. 2016.

[43] R. F. Van der Lans. Designing a logical data warehouse: A technical whitepaper. 2016.

[44] R. F. Van der Lans. Data virtualization in the time of big data: A technical whitepaper. 2017.

[45] R. F. Van der Lans. Architecting the multipurpose data lake with data virtualization. 2018.

[46] N. Yuhanna and M. Gilpin. The forrester wave: Data virtualization, q1 2012. *Forrester*, 2012.

Matthias Gottlieb holds doctoral degree in natural sciences at the Technical University of Munich. Currently, he is working as a scientific research assistant at the chair of information systems. As part of his position, he made several projects on Big Data, employer attractiveness, and driving simulator. He is an expert in experimental design and supervised multiple student theses. Besides, he has been the course coordinator for the bachelorâĂŹs degree in information systems of the department of informatics at the TUM for more than five years. In addition, he was joining the local arrangement chair of the International Conference on Information Systems (ICIS) 2019 in Munich. He is the Deputy Editor-in-Chief of the international Journal of Engineering Pedagogy (iJEP). Matthias has been a visiting researcher at the Department of Information Technology and Management at the Illinois Institute of Technology. He served as a member of the TUM Senate and TUM Board of Trustees from 10/2009 until 09/2011.

Marwin Shraideh is PhD student and research assistant at the Chair for Information Systems at the Department of Informatics of the Technical University of Munich. He is also working for the SAP University Competence Center in parallel. Marwin attained his B.S. and M.S. degree in Information Systems at the Hochschule Pforzheim. During his activity as a working student in the Business Intelligence department of MHP Management- und IT-Beratung GmbH for two and a half years and several other activities, he acquired knowledge in service engineering, software development and project management. Further areas of interest are big data analytics and internet of things. His urge for getting in touch with new topics lead him to his current research project “Bioinformatics-as-a-Service” trying to pave the way for the future of health: personalized medicine and its requirements from service systems perspective.

Isabel Fuhrmann received her B.S. degree in information systems from University of Mannheim in 2015 and her M.S. in information systems from Technical University of Munich in 2018. She now works as a Data Scientist for a software company and advises clients on process optimization topics. Her focus is the implementation of data extraction and transformation methods to support the roll out of the software.

Markus Böhm is Research Group Leader at the Chair for Information Systems at TUM. He has a diploma in Information Systems from the University Erlangen-Nürnberg and holds a PhD in Information Systems from TUM. Markus has a profound industry experience as project manager, analyst and software developer at among others forties, Siemens, Bosch and BMW. His research interests are the Role of IT in Mergers & Acquisitions (M &A) and Divestitures (Carve-Out), Business Model Innovation and IT-enabled Business Models. Markus is teaching Information and Knowledge Management at TUM, in the Executive MBA program at the University of Fribourg, Switzerland and at the Steinbeis School of Management and Technology.

Helmut Krcmar (born December 16, 1954 in Hanau, Germany) is a German IS and Management scholar. Since 2002 he holds the Chair for Information Systems, Department of Informatics at the Technische Universität München (TUM) in Germany with a joint appointment to TUM School of Management. He is member of the TUM Senate and TUM Board of Trustees. Helmut served as Dean, Faculty of Informatics from 10/2010 until 09/2013. In July 2018, he was elected Vice Dean TUM School of Management and Founding Dean TUM Campus Heilbronn. Since 2003, he is Academic Director of the SAP University Competence Center @ TUM and member of the board of the Center for Digital Technology and Management (CDTM). From 2004-2007, he was founding director of TUM Executive Education and today is Academic Director of TUM EEC EMBA “Business and IT”.